B9824 Foundations of Optimization

Lecture 1: Introduction

Fall 2011

Outline

1. Administrative matters
2. Introduction
3. Existence of optima
4. Local theory of unconstrained optimization
5. Constrained local optimality

Administrative Matters

Instructor

Prof. Ciamac Moallemi
Uris 416
email: ciamac@gsb.columbia.edu

Office Hours

Drop by anytime M–F for quick questions, otherwise email for an appointment. (I will schedule formal office hours if there is demand.)

Teaching Assistant

Mehmet Sağlam
Uris Cubicle 4O
email: msaglam13@gsb.columbia.edu

Administrative Matters

Calendar

- Class meets every Thursday, 4:00pm–7:15pm.
- No class on Thursday 10/20 (midterm period).
- No class on Thursday 10/27. Rescheduled for Friday 11/4.
- Last class on 12/8.
- Take home final examination: Saturday 12/17–Sunday 12/18.

Coursework and Grading

- Homework (50%)
- Final (50%)

Course Website

http://angel.gsb.columbia.edu

Mathematical Prerequisites

Linear Algebra

- Vectors and matrices over $\mathbb{R}$
- Null space, range
- Transpose, inner product, norm
- Eigenvalues of symmetric matrices; Spectral theorem
- Positive definite and semidefinite matrices

Mathematical Prerequisites

Calculus / Real Analysis

- Open, closed, compact sets
- Convergent sequences and subsequences
- Continuity
- Differentiability
- Taylor series expansion
- Mean value theorem
- Implicit function theorem

Background Reading

- Bertsekas NLP, Appendix A.0-A.5
- Boyd & Vandenberghe, Appendix A

Required Texts

- D. P. Bertsekas, *Nonlinear Programming*, 2nd Edition. Athena Scientific, 1999.
- S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004. Available online at `http://www.stanford.edu/~boyd/cvxbook`.
- D. G. Luenberger, *Optimization by Vector Space Methods*. Wiley, 1969.

Selected References

Real Analysis:

- W. Rudin, *Principles of Mathematical Analysis*, 3rd Edition. McGraw-Hill, 1976.

Linear Algebra:

- G. Strang, *Linear Algebra and Its Applications*, 3rd Edition. Brooks Cole, 1988.

Optimization:

- D. P. Bertsekas, *Convex Optimization Theory*. Athena Scientific, 2009.
- D. G. Luenberger and Y. Ye, *Linear and Nonlinear Programming*, 3rd Edition. Springer, 2008.

Introduction

An optimization problem (program):

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in \mathcal{C} \end{array}$$

x is the collection of decision variables

The real-valued function $f(\cdot)$ is the objective

$\mathcal{C}$ is the constraint set (feasible set, search space), it is a subset of the domain of f

x^* is an optimal solution (global minimizer) if

$$f(x^*) \leq f(x), \quad \forall\ x \in \mathcal{C}$$

The optimal value is $f(x^*)$

Problem Classification

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in \mathcal{C} \end{array}$$

- Maximization also falls within this framework

  maximize $f(x)$ $\Leftrightarrow$ minimize $-f(x)$

- Problem classifications
  - continuous vs. discrete
  - deterministic vs. stochastic
  - static vs. dynamic

Constraint Sets

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in \mathcal{C} \end{array}$$

The feasible set will usually be constructed as the intersection of a number of constraints, e.g.

$$\begin{aligned} \mathcal{C} = & \{x \ : \ h_1(x) = 0, \ldots, h_m(x) = 0\} && \text{(equality constraints)} \\ & \cap \{x \ : \ g_1(x) \leq 0, \ldots, g_r(x) \leq 0\} && \text{(inequality constraints)} \\ & \cap \Omega && \text{(set constraint)} \end{aligned}$$

Motivating Example: Resource Allocation

- Activities $1, \ldots, m$ (e.g., divisions of a firm)
- Resources $1, \ldots, n$ (e.g., capital, labor, etc.)
- Each activity consumes resources, and generates a benefit (utility, profit, etc.)
- Decision variables:

  $x_{ij} =$ quantity of resource j allocated to activity i

  $x_{ij} \geq 0$
- The ith activity generates utility according to

$$U_i(x_{i1}, \ldots, x_{in})$$

- The supply of the resources is limited, so we require that

$$\sum_{i=1}^{m} x_{ij} \leq C_j, \quad \forall j$$

Motivating Example: Resource Allocation

- Objective: maximize total utility

$$\begin{array}{ll} \text{maximize} & \displaystyle\sum_{i=1}^{m} U_i(x_{i1}, \ldots, x_{in}) \\ \text{subject to} & \displaystyle\sum_{i=1}^{m} x_{ij} \le C_j, \qquad \forall\, 1 \le j \le n \\ & x \ge 0, \\ & x \in \mathbb{R}^{m \times n} \end{array}$$

Motivating Example: Portfolio Optimization

- Securities $1, \ldots, n$ are available for purchase
- Security i has return r_i, which is a random variable

$$\mathsf{E}[r_i] = \mu_i, \qquad \mathsf{Cov}(r_i, r_j) = \Gamma_{ij}, \qquad \Gamma \succeq 0$$

- Decision variables: $x_i =$ fraction of wealth to invest in security i

$$x \ge 0, \qquad \mathbf{1}^\top x = \sum_{i=1}^{n} x_i = 1$$

- Given a portfolio x,

$$\mathsf{E}[\text{return}] = \sum_{i=1}^{n} \mu_i x_i = \mu^\top x, \quad \mathsf{Var}(\text{return}) = \sum_{i=1}^{n}\sum_{j=1}^{n} \Gamma_{ij} x_i x_j = x^\top \Gamma x$$

- The investor requires return $\bar{\mu}$, so

$$\mu^\top x = \bar{\mu}$$

Motivating Example: Portfolio Optimization

- Objective: minimize the variance of the portfolio return (risk)

$$\begin{array}{ll} \text{minimize} & x^\top \Gamma x \\ \text{subject to} & \mathbf{1}^\top x = 1, \\ & \mu^\top x = \bar{\mu}, \\ & x \geq 0, \\ & x \in \mathbb{R}^n \end{array}$$

Motivating Example: Production Planning

- A manufacturer is planning the production goods $1, \ldots, n$ over the time horizon $[0, T]$
- Goods are produced at the rate $r(t) \in \mathbb{R}^n$ at time t, $r(t) \geq 0$
- $d(t) \in \mathbb{R}^n$ is the rate of demand at time t, manufacturer must meet this demand
- The inventory at time t is $x(t) \in \mathbb{R}^n$
- Given a fixed initial inventory $x(0)$,

$$x(t) = x(0) + \int_0^t [r(\tau) - d(\tau)] \, d\tau \geq 0$$

- Production cost rate $c(r(t))$, for some function $c(\cdot) \geq 0$
- Holding/inventory cost rate $h^\top x(t)$, for some vector $h \geq 0$ (linear)

Motivating Example: Production Planning

- Objective: minimize the total cost of the production plan

$$\begin{aligned}
\text{minimize} \quad & \int_0^T \left[c(r(t)) + h^\top x(t) \right] \, dt \\
\text{subject to} \quad & x(t) = x(0) + \int_0^t [r(\tau) - d(\tau)] \, d\tau \geq 0, \quad \forall \, t \in [0, T] \\
& r(t) \geq 0, \qquad \forall \, t \in [0, T] \\
& r(\cdot) \in C([0, T], \mathbb{R}^n)
\end{aligned}$$

Here,

$$C([0, T], \mathbb{R}^n) = \text{continuous functions from } [0, T] \text{ to } \mathbb{R}^n$$

Other Examples

- Data analysis: fitting, statistical estimation, machine learning
- Solution of equilibrium models
- Game theory
- Communications: scheduling, routing
- Computer-aided design

The Basic Questions

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in \mathcal{C} \end{array}$$

Does an optimal solution exist?

Can we characterize the set of optimal solutions? (Necessary & sufficient conditions)

Is the optimal solution unique?

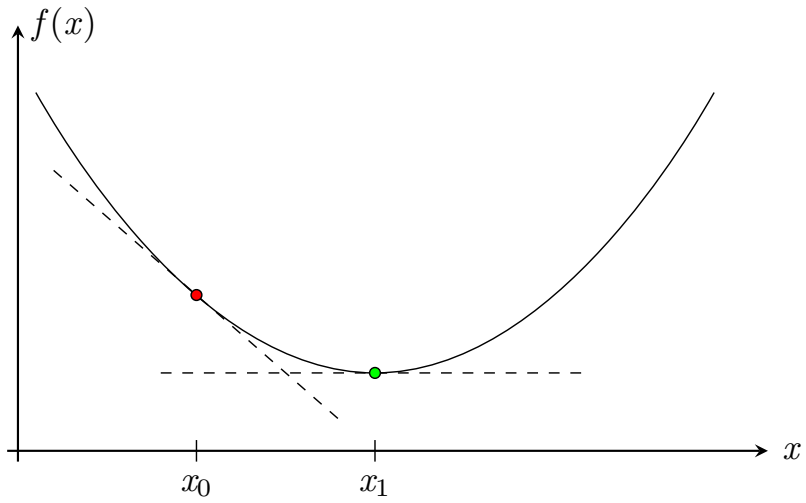
How sensitive is the solution to changes in the objective function or constraint set?

How can an optimal solution/the optimal value be efficiently computed?

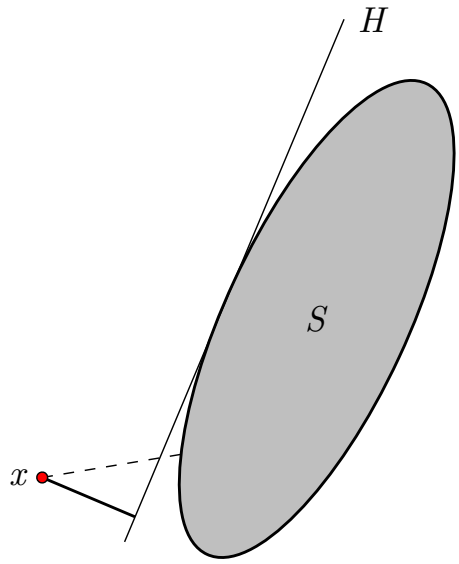
Big Idea I: Differentials

The behavior of a "smooth" function $f(\cdot)$ close to a point x can be approximated with derivatives, e.g., if y is close to x,

$$f(y) \approx f(x) + f'(x)(y - x)$$



Big Idea II: Duality



Problem 1. Find the smallest distance between a vector x and any point in the convex set S

Problem 2. Find the largest distance between a vector x and any hyperplane separating x from S

These problems are equivalent!

minimize over points $\Leftrightarrow$ maximize over hyperplanes

A Taxonomy of Mathematical Programming

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & h_i(x) = 0, \quad \forall\ 1 \le i \le m \\ & g_i(x) \le 0, \quad \forall\ 1 \le i \le r \\ & x \in \mathbb{R}^n \end{array}$$

- **Linear programming:** $f(\cdot), \{h_i(\cdot)\}, \{g_i(\cdot)\}$ are linear
- **Convex programming:** $f(\cdot), \{g_i(\cdot)\}$ are convex, $\{h_i(\cdot)\}$ linear
- **Non-convex programming:** anything goes

$$\overbrace{\underbrace{\text{linear} \quad \le \quad \text{convex}}_{\text{global theory}} \quad \le \quad \text{non-convex}}^{\text{local theory}}$$

Rough Syllabus

1. Introduction
2. Local theory of optimization (differentials, Lagrangians)
3. Global theory of optimization (convexity, duality)
4. Applications & problem formulation
5. Vector spaces: a unifying theory

Existence of Solutions

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in \mathcal{C} \end{array}$$

When does an optimal solution exist?

Example.

$$\begin{array}{ll} \text{minimize} & x \\ \text{subject to} & x \in \mathbb{R} \end{array} \quad \Rightarrow \text{Unbounded!}$$

Example.

$$\begin{array}{ll} \text{minimize} & 1/x \\ \text{subject to} & x > 0 \end{array} \quad \Rightarrow \text{Bounded, but no optima!}$$

Background: Basic Topology

Definition. An open ball (or, "neighborhood") around a point $x \in \mathbb{R}^n$ with radius $r > 0$ is the set

$$N_r(x) \triangleq \{y \in \mathbb{R}^n \ : \ \|x - y\| < r\}$$

Here,

$$\|x\| \triangleq \left(\sum_{i=1}^{n} x_i^2 \right)^{\frac{1}{2}}$$

Background: Open and Closed Sets

Consider a set $\mathcal{E} \subset \mathbb{R}^n$.

Definition. A point $x \in \mathcal{E}$ is an interior point if there exists an open ball $N_r(x)$ such that $N_r(x) \subset \mathcal{E}$. The interior **int** $\mathcal{E}$ is defined to be the set of all interior points of $\mathcal{E}$.

Definition. $\mathcal{E}$ is open if $\mathcal{E} =$ **int** $\mathcal{E}$.

Definition. A point $x \in \mathbb{R}^n$ is a closure point of $\mathcal{E}$ if, for every open ball $N_r(x)$, there exists $y \in \mathcal{E}$ with $y \in N_r(x)$. The closure **cl** $\mathcal{E}$ is defined to be the set of all closure points of $\mathcal{E}$.

Definition. $\mathcal{E}$ is closed if every closure point if $\mathcal{E} =$ **cl** $\mathcal{E}$.

Background: Open and Closed Sets

Theorem.

(a) The union of any collection of open sets is open.

(b) The intersection of any collection of closed sets is closed.

(c) The intersection of any finite collection of open sets is open.

(d) The union of any finite collection of closed sets is closed.

Background: Convergence

Definition. A sequence of vectors $\{x_k\} \subset \mathbb{R}^n$ converges to a limit $x \in \mathbb{R}^n$ if

$$\lim_{k \to \infty} \|x - x_k\| = 0$$

We say $x_k \to x$.

Background: Compactness

Consider a set $\mathcal{E} \subset \mathbb{R}^n$.

Definition. $\mathcal{E}$ is (sequentially) compact if, given a sequence $\{x_k\} \subset \mathcal{E}$, there is a subsequence $\{x_{k_i}\}$ converging to an element $x \in \mathcal{E}$.

Definition. $\mathcal{E}$ is bounded if there exists a neighborhood $N_r(x)$ such that $\mathcal{E} \subset N_r(x)$.

Theorem. (Heine-Borel) A set $\mathcal{E} \subset \mathbb{R}^n$ is compact if and only if it is closed and bounded.

Theorem. A closed subset of a compact set is compact.

Theorem. Suppose $\{\mathcal{E}_n\}$ are is a sequence of non-empty, compact sets that are nested, i.e., $\mathcal{E}_{n+1} \subset \mathcal{E}_n$. Then, their intersection is non-empty.

Background: Continuity

Consider a real-valued function $f(\cdot)$ defined on a domain $\mathcal{X} \subset \mathbb{R}^n$.

Definition. $f(\cdot)$ is continuous at the point $x \in \mathcal{X}$ if, for every sequence $\{x_k\} \subset \mathcal{X}$ with $x_k \to x$,

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

We say $f(\cdot)$ is continuous if it is continuous at all points of $\mathcal{X}$.

Consider a set $\mathcal{A} \subset \mathbb{R}$, define the inverse image

$$f^{-1}(\mathcal{A}) \triangleq \{x \in \mathcal{X} \ : \ f(x) \in \mathcal{A}\}$$

Theorem. Assume that $f(\cdot)$ is continuous. If the domain $\mathcal{X}$ is open (respectively, closed) and $\mathcal{A}$ is open (closed), then $f^{-1}(\mathcal{A})$ is also open (closed).

Note: This is usually the way to prove that a set is open/closed.

Existence Theorem I

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in \mathcal{C} \subset \mathbb{R}^n \end{array}$$

When does an optimal solution exist?

Theorem. (Weierstrass) Assume that $\mathcal{C}$ is non-empty and that $f(\cdot)$ is continuous over $\mathcal{C}$.

If $\mathcal{C}$ is compact, then the set of optimal solutions of $f(\cdot)$ is non-empty and compact.

Proof of Existence Theorem I

Define

$$f^* = \inf_{x \in \mathcal{C}} f(x) \in \mathbb{R} \cup \{-\infty\}$$

Note: f^* always exists!

Given $\gamma > f^*$, the sub-level set

$$\mathcal{C}(\gamma) \triangleq \{x \in \mathcal{C} \ : \ f(x) \leq \gamma\}$$

must be non-empty and, by the continuity of $f(\cdot)$, is closed. Then, since $\mathcal{C}$ is compact, $\mathcal{C}(\gamma)$ also compact. Given a sequence of real numbers $\{\gamma_k\}$ with $\gamma_k \downarrow f^*$, the set of optimal solutions is

$$\mathcal{X}^* = \bigcap_{k=1}^{\infty} \mathcal{C}(\gamma_k)$$

The intersection of a collection of nested, non-empty compact sets is non-empty and also compact.

Application: Portfolio Optimization

$$\begin{aligned} \text{minimize} \quad & x^\top \Gamma x \\ \text{subject to} \quad & \mathbf{1}^\top x = 1, \\ & \mu^\top x = \bar{\mu}, \\ & x \geq 0, \\ & x \in \mathbb{R}^n \end{aligned}$$

Check that:

- The objective is continuous
  $\Rightarrow$ true since it is a polynomial
- The feasible set is non-empty
  $\Rightarrow$ true iff $\min_i \mu_i \leq \bar{\mu} \leq \max_i \mu_i$
- The feasible set is compact
  $\Rightarrow$ clearly bounded
  $\Rightarrow$ closed since it is the inverse image of a closed set under a continuous function

Coerciveness

Example. Consider the program

$$\begin{aligned} \text{minimize} \quad & x^2 \\ \text{subject to} \quad & x \in \mathbb{R} \end{aligned}$$

This has a unique optimal solution $x^* = 0$, but the Weierstrass Theorem does not apply since the feasible set is not compact.

Consider a real-valued function $f(\cdot)$.

Definition. $f(\cdot)$ is coercive over a set $\mathcal{C} \subset \mathbb{R}^n$, if, for every sequence $\{x_k\} \subset \mathcal{C}$ with $\|x_k\| \to \infty$,

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

Existence Theorem II

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in \mathcal{C} \subset \mathbb{R}^n \end{array}$$

When does an optimal solution exist?

Theorem. Assume that $\mathcal{C}$ be non-empty and that $f(\cdot)$ is continuous over $\mathcal{C}$.

If $\mathcal{C}$ is closed and $f(\cdot)$ is coercive over $\mathcal{C}$, then the set of optimal solutions of $f(\cdot)$ is non-empty and compact.

Proof. Since $f(\cdot)$ is coercive, that $\mathcal{C}(\gamma)$ is non-empty and bounded for any $\gamma > f^*$. Since the domain $\mathcal{C}$ is closed and $\mathcal{C}(\gamma) \triangleq f^{-1}((-\infty, \gamma])$, then $\mathcal{C}(\gamma)$ is closed. Thus, $\mathcal{C}(\gamma)$ is compact. The proof proceeds as before. □

Application: Unconstrained Quadratic Optimization

Given a symmetric matrix $\Gamma \in \mathbb{R}^{n \times n}$ and a vector $b \in \mathbb{R}^n$, consider:

$$\begin{array}{ll} \text{minimize} & \frac{1}{2} x^\top \Gamma x - b^\top x \\ \text{subject to} & x \in \mathbb{R}^n \end{array}$$

What are sufficient conditions to guarantee the existence of an optimal solution?

Answer: If λ is the smallest eigenvalue of Γ, we have

$$\frac{1}{2} x^\top \Gamma x - b^\top x \geq \frac{\lambda}{2} \|x\|^2 - \|b\| \|x\|$$

Thus, if $\lambda > 0$, the objective is coercive. This is equivalent to $\Gamma \succ 0$, i.e., Γ is positive definite.

Remarks on Existence

The key condition for both theorems is that there exists some γ^* such that the sub-level set

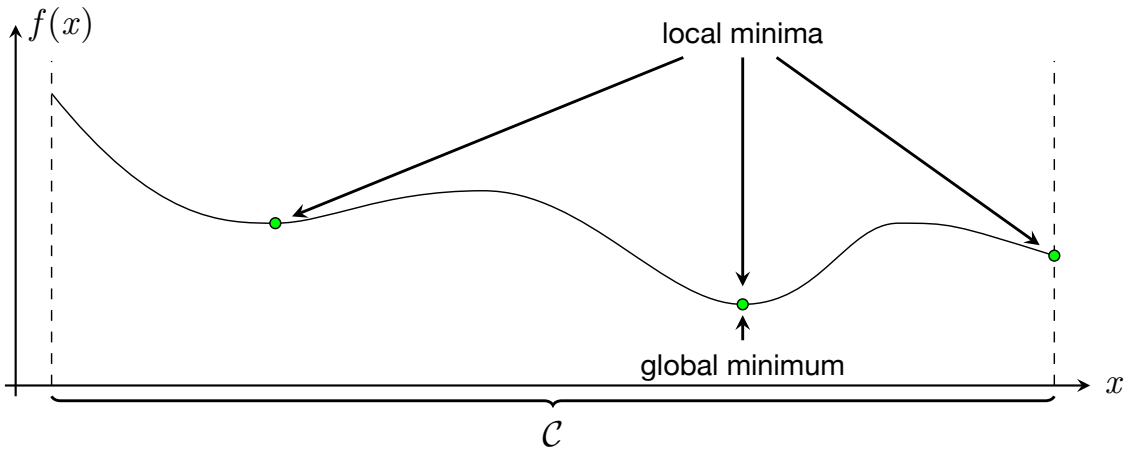
$$\mathcal{C}(\gamma^*) = \{x \in \mathcal{C} \ : \ f(x) \leq \gamma^*\}$$

is non-empty and compact.

Continuity of $f(\cdot)$ was only used to establish that $\mathcal{C}(\gamma)$ is closed for $\gamma \leq \gamma^*$. This is also implied by a weaker condition known as lower semi-continuity.

Local Optimality

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in \mathcal{C} \subset \mathbb{R}^n \end{array}$$

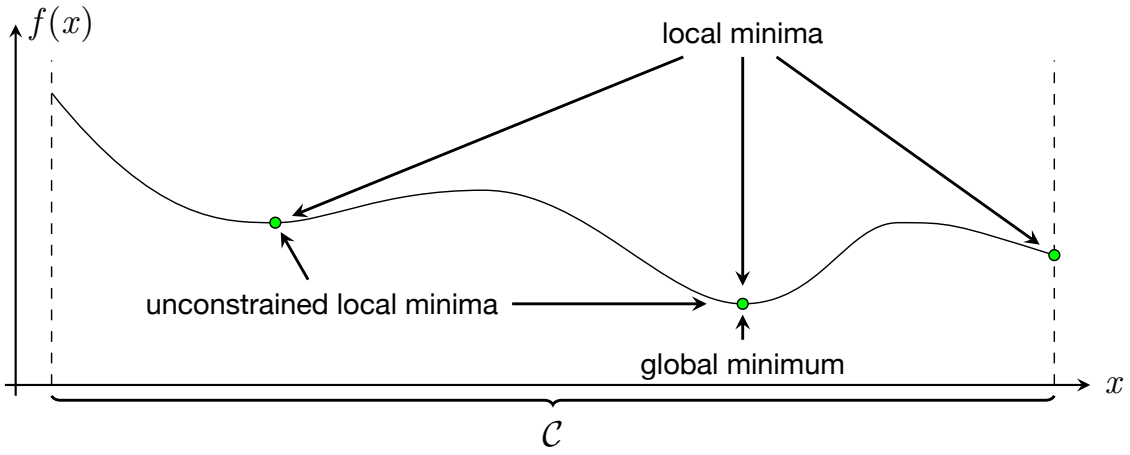


Definition. A point $x \in \mathcal{C}$ is a local minimum if there exists a neighborhood $N_r(x)$ such that

$$f(x) \leq f(y), \quad \forall \ y \in \mathcal{C} \cap N_r(x).$$

Local Optimality

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in \mathcal{C} \subset \mathbb{R}^n \end{array}$$



Definition. A point $x \in \mathcal{C}$ is an unconstrained local minimum if there exists a neighborhood $N_r(x) \subset \mathcal{C}$ such that

$$f(x) \leq f(y), \quad \forall\ y \in N_r(x).$$

Strict Local Optimality

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in \mathcal{C} \subset \mathbb{R}^n \end{array}$$

Definition. A point $x \in \mathcal{C}$ is a strict local minimum if there exists a neighborhood $N_r(x)$ such that

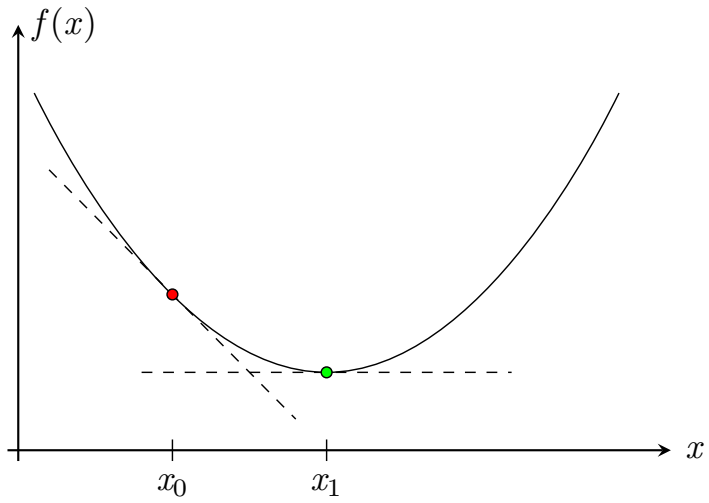
$$f(x) < f(y), \quad \forall\ y \in \mathcal{C} \cap N_r(x),\ y \neq x.$$

Definition. A point $x \in \mathcal{C}$ is an strict unconstrained local minimum if there exists a neighborhood $N_r(x) \subset \mathcal{C}$ such that

$$f(x) < f(y), \quad \forall\ y \in N_r(x),\ y \neq x.$$

Local Optimality

$$f(x+h) \approx f(x) + f'(x)h + \tfrac{1}{2}f''(x)h^2$$



Necessary conditions:

$$f'(x) = 0$$

$$f''(x) \geq 0$$

Sufficient conditions:

$$f'(x) = 0$$

$$f''(x) > 0$$

Background: Differentiation

Consider a real-valued function $f\colon \mathcal{X} \to \mathbb{R}$ with $\mathcal{X} \subset \mathbb{R}^n$.

Definition. $f(\cdot)$ is differentiable at the point $x \in$ **int** $\mathcal{X}$ if there exists a vector $\nabla f(x) \in \mathbb{R}^n$, such that

$$\lim_{d \to 0} \frac{f(x+d) - f(x) - \nabla f(x)^\top d}{\|d\|} = 0.$$

$\nabla f(x)$ is known as the gradient.

Definition. $f(\cdot)$ is differentiable over an open set $\mathcal{U} \subset \mathcal{X}$ if it is differentiable at every point $x \in \mathcal{U}$. If, in addition, the components of the gradient $\nabla f(x)$ are continuous over $\mathcal{U}$, we say $f(\cdot)$ is continuously differentiable over $\mathcal{U}$.

Background: Differentiation

Note: If $f(\cdot)$ is differentiable at x, then

$$\nabla f(x) = \left[\frac{\partial f(x)}{\partial x_1}, \ldots, \frac{\partial f(x)}{\partial x_n}\right]^\top \in \mathbb{R}^n,$$

where

$$\frac{\partial f(x)}{\partial x_i} = \lim_{h \to 0} \frac{f(x + he_i) - f(x)}{h}.$$

Here, $e_i \in \mathbb{R}^n$ is the ith coordinate vector.

Background: Mean Value Theorem

Theorem. Suppose that $g\colon \mathbb{R} \to \mathbb{R}$ is continuously differentiable over an open interval $\mathcal{I} \subset \mathbb{R}$. Then, for all $a, b \in \mathcal{I}$, there exists some $\zeta \in [a, b]$ such that

$$g(b) - g(a) = g'(\zeta)(b - a)$$

Note: This can be applied in to a multi-variable function $f\colon \mathbb{R}^n \to \mathbb{R}$, so show that if $x, y \in \mathbb{R}^n$, there exists $\tilde{x} \in \mathbb{R}^n$ on the line segment between x and y so that

$$f(y) - f(x) = \nabla f(\tilde{x})^\top (y - x)$$

if the single-variable function $g(t) = f\big(ty + (1 - t)x\big)$ is continuously differentiable on an interval containing $[0, 1]$.

Background: The Hessian Matrix

Consider a real-valued function $f\colon \mathcal{X} \to \mathbb{R}$ with $\mathcal{X} \subset \mathbb{R}^n$.

Definition. Consider a point $x \in$ **int** $\mathcal{X}$, and suppose that each component of the vector $\nabla f(\cdot)$ is differentiable at x. We say that $f(\cdot)$ is twice differentiable at x, and define the Hessian to be the matrix $\nabla^2 f(x) \in \mathbb{R}^{n\times n}$ by

$$\nabla^2 f(x) = \left[\frac{\partial^2 f(x)}{\partial x_i \partial x_j}\right]_{ij}$$

Note: If $f(\cdot)$ is twice continuously differentiable in a neighborhood of x, then the Hessian is symmetric, i.e.,

$$\nabla^2 f(x) = \nabla^2 f(x)^\top$$

Background: Second Order Taylor Expansion

Consider a real-valued function $f\colon \mathcal{X} \to \mathbb{R}$ with $\mathcal{X} \subset \mathbb{R}^n$, and a point $x \in$ **int** $\mathcal{X}$.

Theorem. Suppose that $f(\cdot)$ is twice continuously differentiable over a neighborhood $N_r(x)$. Then, for all $d \in N_r(0)$,

$$f(x+d) = f(x) + \nabla f(x)^\top d + \tfrac{1}{2} d^\top \nabla^2 f(x) d + o(\|d\|^2)$$

Formally, this means that if

$$R(d) = f(x+d) - \left(f(x) + \nabla f(x)^\top d + \tfrac{1}{2} d^\top \nabla^2 f(x) d\right),$$

then, for every $C > 0$, there exists a neighborhood $N_\epsilon(0)$ and such that

$$|R(d)| < C\|d\|^2, \quad \forall\, d \in N_\epsilon(0)$$

Necessary Conditions for Optimality

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in \mathcal{C} \end{array}$$

Theorem. Let $x^* \in$ **int** $\mathcal{C}$ be an unconstrained local minimum.

(i) Suppose that $f(\cdot)$ is continuously differentiable in a neighborhood of x^*. Then,

$$\nabla f(x^*) = 0 \qquad \text{(first order necessary condition)}$$

(ii) If $f(\cdot)$ is twice continuously differentiable in a neighborhood of x^*, then $\nabla^2 f(x^*)$ is positive semidefinite, i.e.

$$\nabla^2 f(x^*) \succeq 0 \qquad \text{(second order necessary condition)}$$

Necessary Conditions: Proof

(i) Fix $d \in \mathbb{R}^n \setminus \{0\}$. Note that, by the definition of the gradient,

$$0 = \lim_{\alpha \to 0} \frac{f(x^* + \alpha d) - f(x^*) - \alpha \nabla f(x^*)^\top d}{\alpha \| d \|}$$

Thus,

$$\lim_{\alpha \to 0} \frac{f(x^* + \alpha d) - f(x^*)}{\alpha} = \nabla f(x^*)^\top d$$

For α sufficiently small, however,

$$f(x^* + \alpha d) - f(x^*) \geq 0$$

Then,

$$\nabla f(x^*)^\top d \geq 0$$

Since d is arbitrary, it follows that $\nabla f(x^*) = 0$.

Necessary Conditions: Proof

(ii) Fix $d \in \mathbb{R}^n$. For α sufficiently small, using a second order expansion,

$$\begin{aligned} f(x^* + \alpha d) - f(x^*) &= \alpha \nabla f(x^*)^\top d + \tfrac{1}{2}\alpha^2 d^\top \nabla^2 f(x^*) d + o(\alpha^2) \\ &= \tfrac{1}{2}\alpha^2 d^\top \nabla^2 f(x^*) d + o(\alpha^2) \end{aligned}$$

Then

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

Taking the limit as $\alpha \to 0$, it follows that

$$0 \leq d^\top \nabla^2 f(x^*) d$$

Sufficient Conditions for Optimality

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in \mathcal{C} \end{array}$$

Theorem. Consider a point $x^* \in$ **int** $\mathcal{C}$. Suppose that $f(\cdot)$ is twice continuously differentiable in a neighborhood $N_r(x^*) \subset \mathcal{C}$ of x^*, and that

$$\nabla f(x^*) = 0, \qquad \nabla^2 f(x^*) \succ 0.$$

Then, x^* is a strict unconstrained local minimum.

Sufficient Conditions: Proof

Let $\lambda > 0$ be the smallest eigenvalue of $\nabla^2 f(x^*)$.

Fix $d \in N_r(0) \setminus \{0\}$,

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

For any $\gamma \in (0, \lambda)$, there exists Pick $\epsilon \in (0, r]$ and so that

$$\frac{\lambda}{2} + \frac{o(\|d\|^2)}{\|d\|^2} \geq \frac{\gamma}{2}, \quad \forall\ d \text{ with } 0 < \|d\| < \epsilon$$

Then, for all $d \in N_\epsilon(0) \setminus \{0\}$,

$$f(x^* + d) \geq f(x^*) + \frac{\gamma}{2}\|d\|^2 > f(x^*)$$

Application of Necessary Conditions

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in \mathcal{C} \end{array}$$

How can we find the optimal solution(s)?
(Assuming $f(\cdot)$ is continuously differentiable on **int** $\mathcal{C}$)

(i) Check that a global minima exists (e.g., Weierstrass' Theorem, coerciveness)

(ii) Find a set of possible unconstrained local minima via the necessary condition

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

(iii) Find the set of "boundary" points $\mathcal{C} \setminus$ **int** $\mathcal{C}$

(iv) The global minima must be among the points in (ii) and (iii), so evaluate $f(\cdot)$ at each of these points and find those with the smallest value

Application of Necessary Conditions

Example. $f\colon [a,b] \to \mathbb{R}$ continuous on $[a,b]$ and continuously differentiable on (a,b)

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in [a,b] \end{array}$$

Global minima must exist and are contained in the set

$$\{a,b\} \cup \{x \in (a,b) \ : \ f'(x) = 0\}$$

Example. $\Gamma \in \mathbb{R}^{n\times n}$, $\Gamma \succ 0$, $b \in \mathbb{R}^n$

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}x^\top \Gamma x - b^\top x \\ \text{subject to} & x \in \mathbb{R}^n \end{array}$$

Global minima must exist, and $\mathcal{C} \setminus \textbf{int}\ \mathcal{C}$ is empty, so global minima must be unconstrained local minima. First order necessary conditions:

$$\Gamma x - b = 0 \quad \Rightarrow \quad x^* = \Gamma^{-1} b$$

Application of Necessary Conditions

The necessary conditions for unconstrained local optima are only useful if the boundary $\mathcal{C} \setminus \textbf{int}\ \mathcal{C}$ is empty or "small", or if the (for other reasons) we know the global minima will not occur on the boundary.

Example. (Portfolio optimization)

$$\begin{array}{ll} \text{minimize} & x^\top \Gamma x \\ \text{subject to} & \mathbf{1}^\top x = 1, \\ & \mu^\top x = \bar{\mu}, \\ & x \geq 0, \\ & x \in \mathbb{R}^n \end{array}$$

The interior of the constraint set is empty, no unconstrained local optima!

Sensitivity Analysis

Consider a function $f\colon \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$, and, for a fixed $a \in \mathbb{R}^m$, the unconstrained optimization problem

$$\begin{array}{ll} \text{minimize} & f(x, a) \\ \text{subject to} & x \in \mathbb{R}^n \end{array}$$

Denote by $x^*(a)$ a local minimizer, assuming it exists, and define $f^*(a) \triangleq f(x^*(a), a)$.

a is a parameter vector. We would like to understand how the local minimum $x^*(a)$ and the associated value $f^*(a)$ are sensitive to changes in a.

Differentiation of Vector-Valued Functions

Consider a vector-valued function $F\colon \mathcal{X} \to \mathbb{R}^m$, $\mathcal{X} \subset \mathbb{R}^n$, and a point $x \in \textbf{int}\ \mathcal{X}$. We can analyze $F(\cdot)$ in terms of component functions $F_i\colon \mathcal{X} \to \mathbb{R}$ by

$$F(x) = (F_1(x), \ldots, F_m(x))\,.$$

Definition. $F(\cdot)$ is differentiable at the point x each component function $F_i(\cdot)$ is differentiable at x. We define the gradient to be the matrix $\nabla F(x) \in \mathbb{R}^{n \times m}$ with

$$\nabla F(x) = [\nabla F_1(x), \ldots, \nabla F_m(x)]$$

If $F(\cdot)$ is differentiable at x, then

$$\nabla F(x)_{ij} = \frac{\partial F_j(x)}{\partial x_i}$$

Theorem. (Chain Rule) If $f\colon \mathcal{X} \to \mathbb{R}^m$, $\mathcal{X} \subset \mathbb{R}^n$, is differentiable at $x \in$ **int** $\mathcal{X}$ and $g\colon \mathcal{Y} \to \mathbb{R}^p$, $\mathcal{Y} \subset \mathbb{R}^m$, is differentiable at $f(x) \in$ **int** $\mathcal{Y}$, then the composition

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

is differentiable at x, and

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

Non-rigorous Sensitivity Analysis

minimize	$f(x, a)$	$x^*(a) =$ local minimum
subject to	$x \in \mathbb{R}^n$	$f^*(a) = f(x^*(a), a)$

First order conditions: $\nabla_x f(x^*(a), a) = 0$

Taking derivatives: $\nabla x^*(a) \nabla^2_{xx} f(x^*(a), a) + \nabla^2_{xa} f(x^*(a), a) = 0$

Sensitivity of local minimum:

$$\nabla x^*(a) = -\nabla^2_{xa} f(x^*(a), a) \left(\nabla^2_{xx} f(x^*(a), a) \right)^{-1}$$

Sensitivity of value:

$$\begin{aligned} \nabla f^*(a) &= \nabla x^*(a) \nabla_x f(x^*(a), a) + \nabla_a f(x^*(a), a) \\ &= \nabla_a f(x^*(a), a) \end{aligned}$$

Implicit Function Theorem

Theorem. Let $F\colon \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ be a function with

(i) $F(\bar{x}, \bar{y}) = 0$

(ii) $F(\cdot, \cdot)$ is continuous and has a continuous and invertible gradient matrix $\nabla_x F(x, y)$ in an open set containing $(\bar{x}, \bar{y})$

Then, there exist open sets $\mathcal{U}_x \subset \mathbb{R}^n$ and $\mathcal{U}_y \subset \mathbb{R}^m$ with $x \in \mathcal{U}_x$ and $y \in \mathcal{U}_y$ and a continuous function $\phi\colon \mathcal{U}_y \to \mathcal{U}_x$ such that

(i) $\bar{x} = \phi(\bar{y})$ and $F(\phi(y), y) = 0$ for all $y \in \mathcal{U}_y$

(ii) The function $\phi(\cdot)$ is unique in the sense that, if $x \in \mathcal{U}_x$, $y \in \mathcal{U}_y$, and $F(x, y) = 0$, then $x = \phi(y)$

(iii) If $F(\cdot, \cdot)$ is k times continuously differentiable the same is true for $\phi(\cdot)$, and

$$\nabla \phi(y) = -\nabla_y F(\phi(y), y) \left(\nabla_x F(\phi(y), y)\right)^{-1}, \quad \forall\ y \in \mathcal{U}_y$$

Constrained Local Optimality

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in \mathcal{C} \subset \mathbb{R}^n \end{array}$$

Definition. A point $x \in \mathcal{C}$ is a local minimum if there exists a neighborhood $N_r(x)$ such that

$$f(x) \leq f(y), \quad \forall\ y \in \mathcal{C} \cap N_r(x).$$

We are interested in characterizing local minima that are not in **int** $\mathcal{C}$. The constraint set plays an fundamental role in this case.

Constrained Local Optimality

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in \mathcal{C} \subset \mathbb{R}^n \end{array}$$

If $f(\cdot)$ is differentiable at a point $x^* \in \mathcal{C}$, and $d \in \mathbb{R}^n$, then

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

If x^* is a local minimum and d is a "feasible" direction, then

$$0 \leq \nabla f(x^*)^\top d$$

The Descent and Tangent Cones

minimize	$f(x)$	$x^* \in \mathcal{C}$, $f(\cdot)$ continuously differentiable in a
subject to	$x \in \mathcal{C} \subset \mathbb{R}^n$	neighborhood of x^*

Definition. The set of descent directions of the objective function $f(\cdot)$ at x^* is the set

$$\mathcal{D}(x^*) = \{d \in \mathbb{R}^n : \nabla f(x^*)^\top d < 0\}$$

Definition. The tangent cone $\mathcal{T}(x^*)$ of the constraint set $\mathcal{C}$ at x^* is the set of directions $d \in \mathbb{R}^n$ such that either

(i) $d = 0$

(ii) there exists a sequence $\{x_k\} \subset \mathcal{C}$ with $x_k \to x^*$ and

$$\frac{x_k - x^*}{\|x_k - x^*\|} \to \frac{d}{\|d\|}$$

Local Optimality Necessary Condition

minimize	$f(x)$	$x^* \in \mathcal{C}$, $f(\cdot)$ continuously differentiable in a
subject to	$x \in \mathcal{C} \subset \mathbb{R}^n$	neighborhood of x^*

Theorem. If x^* is a local minimum, then there is no descent direction in the tangent cone. That is,

$$\mathcal{D}(x^*) \cap \mathcal{T}(x^*) = \emptyset$$

Local Optimality Necessary Condition: Proof

Consider $d \in \mathcal{T}(x^*) \setminus \{0\}$. There exists $\{x_k\} \subset \mathcal{C}$, $x_k \neq x^*$, $x_k \to x^*$ such that

$$\frac{x_k - x^*}{\|x_k - x^*\|} \to \frac{d}{\|d\|}$$

Define

$$\zeta_k \triangleq \frac{x_k - x^*}{\|x_k - x^*\|} - \frac{d}{\|d\|}, \quad d_k \triangleq d + \|d\|\zeta_k$$

Then, $\zeta_k \to 0$ and $d_k \to d$.

By the mean value theorem,

$$f(x_k) = f(x^*) + \nabla f(\tilde{x}_k)^\top (x_k - x^*)$$

where $\tilde{x}_k$ is a point on the line segment between x_k and x^*.

Local Optimality Necessary Condition: Proof

Equivalently,

$$f(x_k) = f(x^*) + \frac{\|x_k - x^*\|}{\|d\|} \nabla f(\tilde{x}_k)^\top d_k$$

If $d \in \mathcal{D}(x^*)$, then $\nabla f(x^*)^\top d < 0$. Thus, for k sufficiently large, $\nabla f(\tilde{x}_k)^\top d_k < 0$. Then,

$$f(x_k) < f(x^*),$$

which contradicts local minimality of x^*.

B9824 Foundations of Optimization

Lecture 2: Local Theory of Optimization

Fall 2011

Outline

1. Necessary conditions for equality constraints
2. Lagrangian cookbook recipe
3. Sufficient conditions for equality constraints, sensitivity analysis
4. Necessary conditions for inequality constraints
5. KKT cookbook recipe
6. Sufficient conditions for inequality constraints, sensitivity analysis

The Descent and Tangent Cones

minimize	$f(x)$	$x^* \in \mathcal{C}$, $f(\cdot)$ continuously differentiable in a
subject to	$x \in \mathcal{C} \subset \mathbb{R}^n$	neighborhood of x^*

Definition. The set of descent directions of the objective function $f(\cdot)$ at x^* is the set

$$\mathcal{D}(x^*) = \{d \in \mathbb{R}^n : \nabla f(x^*)^\top d < 0\}$$

Definition. The tangent cone $\mathcal{T}(x^*)$ of the constraint set $\mathcal{C}$ at x^* is the set of directions $d \in \mathbb{R}^n$ such that either

(i) $d = 0$

(ii) there exists a sequence $\{x_k\} \subset \mathcal{C}$ with $x_k \to x^*$ and

$$\frac{x_k - x^*}{\|x_k - x^*\|} \to \frac{d}{\|d\|}$$

Local Optimality Necessary Condition

minimize	$f(x)$	$x^* \in \mathcal{C}$, $f(\cdot)$ continuously differentiable in a
subject to	$x \in \mathcal{C} \subset \mathbb{R}^n$	neighborhood of x^*

Theorem. If x^* is a local minimum, then there is no descent direction in the tangent cone. That is,

$$\mathcal{D}(x^*) \cap \mathcal{T}(x^*) = \emptyset$$

Equality Constrained Optimization

Consider

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & h_1(x) = 0, \ldots, h_m(x) = 0, \\ & x \in \mathbb{R}^n \end{array}$$

where

$$f : \mathbb{R}^n \to \mathbb{R}, \quad h_i : \mathbb{R}^n \to \mathbb{R}, \ \forall\, 1 \leq i \leq m$$

- Assume that $f(\cdot)$ and $\{h_i(\cdot)\}$ are continuously differentiable on $\mathbb{R}^n$
- The necessary and sufficient conditions are also true if these functions are just defined and continuously differentiable in a neighborhood of the local minimum

Equality Constrained Optimization

$$\begin{array}{l} f : \ \mathbb{R}^n \to \mathbb{R} \\ h : \ \mathbb{R}^n \to \mathbb{R}^m \end{array}$$

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & h(x) = 0, \\ & x \in \mathbb{R}^n \end{array}$$

Assume that x^* is a feasible point and $d \in \mathbb{R}^n$ is a direction. For small $\alpha > 0$,

$$h(x^* + \alpha d) \approx h(x^*) + \nabla h(x^*)^\top (\alpha d) = \alpha \nabla h(x^*)^\top d$$

Definition. The cone of first order feasible variations at a point $x^* \in \mathbb{R}^n$ is the set

$$\mathcal{V}(x^*) = \{d \in \mathbb{R}^n : \nabla h(x^*)^\top d = 0\}$$

Note: If $d \in \mathcal{V}(x^*)$, then $-d \in \mathcal{V}(x^*)$. Indeed, $\mathcal{V}(x^*)$ is a subspace of $\mathbb{R}^n$.

Regularity

$$f : \mathbb{R}^n \to \mathbb{R}$$
$$h : \mathbb{R}^n \to \mathbb{R}^m$$

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & h(x) = 0, \\ & x \in \mathbb{R}^n \end{array}$$

Definition. A point $x^* \in \mathbb{R}^n$ is a regular point if it is feasible and if the constraint gradients

$$\nabla h_1(x^*), \ldots, \nabla h_m(x^*)$$

are linearly independent.

Note: If $m > n$, no regular points exist.

If $m = 1$, regularity is equivalent to $\nabla h_1(x^*) \neq 0$.

Regularity Lemma

$$f : \mathbb{R}^n \to \mathbb{R}$$
$$h : \mathbb{R}^n \to \mathbb{R}^m$$

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & h(x) = 0, \\ & x \in \mathbb{R}^n \end{array}$$

Lemma. Let x^* be a regular point. Then,

(i) For each $d \in \mathcal{V}(x^*)$, there exists $\tau > 0$ and a curve $x : (-\tau, \tau) \to \mathbb{R}^n$ such that

  (a) $x(0) = x^*$, $h(x(t)) = 0$ for $t \in (-\tau, \tau)$
  (b) $x(\cdot)$ is continuously differentiable, and $\dot{x}(0) = d$
  (c) if $h(\cdot)$ is twice continuously differentiable, then so is $x(\cdot)$

(ii) $\mathcal{T}(x^*) = \mathcal{V}(x^*)$

Regularity Lemma: Proof

(i) Fix $d \in \mathcal{V}(x^*)$. Given a scalar t, consider the m equations

$$h\big(x^* + td + \nabla h(x^*)u(t)\big) = 0$$

for an unknown $u(t) \in \mathbb{R}^m$. For $t = 0$, this has solution $u(0) = 0$. The gradient w.r.t. u is at $t = 0$, $u = 0$ is

$$\nabla h(x^*)^\top \nabla h(x^*)$$

This is invertible since the columns of $\nabla h(x^*)$ are linearly independent. By the implicit function theorem, for some $\tau > 0$, a solution $u(t)$ exists for $t \in (-\tau, \tau)$.

Define $x(t) \triangleq x^* + td + \nabla h(x^*)u(t)$. Differentiating $h(x(t)) = 0$ at with respect to t at $t = 0$,

$$0 = \left(d^\top + \dot{u}(0)^\top \nabla h(x^*)^\top\right) \nabla h(x^*)$$

Since $d \in \mathcal{V}(x^*)$, $d^\top \nabla h(x^*) = 0$. Then, $\dot{u}(0) = 0$, thus $\dot{x}(0) = d$.

Regularity Lemma: Proof

(ii) $\mathcal{V}(x^*) \subset \mathcal{T}(x^*)$: If $d \in \mathcal{V}(x^*) \setminus \{0\}$, define $x(t)$ from (i), and there exists a sequence $t_k \subset (0, \tau)$, $t_k \to 0$, with $x_k \triangleq x(t_k) \neq x^*$. Then,

$$\frac{x_k - x^*}{\|x_k - x^*\|} \to \frac{\dot{x}(0)}{\|\dot{x}(0)\|} = \frac{d}{\|d\|}$$

by the mean value theorem applied to $x(t)$.

$\mathcal{T}(x^*) \subset \mathcal{V}(x^*)$: Consider $d \in \mathcal{T}(x^*) \setminus \{0\}$, and an associated $\{x_k\}$. By the mean value theorem,

$$0 = h(x_k) = h(x^*) + \nabla h(\tilde{x}_k)^\top (x_k - x^*)$$

Thus,

$$\nabla h(\tilde{x}_k)^\top \frac{x_k - x^*}{\|x_k - x^*\|} = 0$$

Take the limit as $k \to \infty$.

Local Optimality Necessary Condition

$$f : \mathbb{R}^n \to \mathbb{R}$$
$$h : \mathbb{R}^n \to \mathbb{R}^m$$

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & h(x) = 0, \\ & x \in \mathbb{R}^n \end{array}$$

Theorem. If x^* is a local minimum that is a regular point, then

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

for all directions $d \in \mathcal{V}(x^*)$. In particular, there is no descent direction that is a first order feasible variation.

Proof. Since x^* is a local minimum, $\mathcal{D}(x^*) \cap \mathcal{T}(x^*) = \emptyset$. Since x^* is regular, $\mathcal{T}(x^*) = \mathcal{V}(x^*)$, thus we have $\mathcal{D}(x^*) \cap \mathcal{V}(x^*) = \emptyset$. Assume $d \in \mathcal{V}(x^*)$. Then $\nabla f(x^*)^\top d \geq 0$. Since $-d \in \mathcal{V}(x^*)$, the result follows. □

A Linear Algebra Lemma

Definition. Consider a matrix $A \in \mathbb{R}^{m \times n}$. The kernel (nullspace) is the set

$$\ker A \triangleq \{x \in \mathbb{R}^n \ : \ Ax = 0\}$$

The image (range) is the set

$$\text{im } A \triangleq \{y \in \mathbb{R}^m \ : \ y = Ax, \ x \in \mathbb{R}^n\}$$

Definition. Given a set $\mathcal{S} \subset \mathbb{R}^n$, the orthogonal complement is defined to be the set

$$\mathcal{S}^\perp \triangleq \{x \in \mathbb{R}^n \ : \ x^\top y = 0, \ \forall \ y \in \mathcal{S}\}$$

Lemma. If $A \in \mathbb{R}^{m \times n}$ is a matrix, then $\text{im } A = [\ker (A^\top)]^\perp$.

In other words, given $z \in \mathbb{R}^m$,

$$z = Ax \text{ for some } x \in \mathbb{R}^n \quad \Leftrightarrow \quad z^\top y = 0 \text{ for all } y \text{ with } A^\top y = 0$$

Linear Algebra Lemma: Proof

Note that if $\mathcal{S} \subset \mathbb{R}^k$ is a subspace, then $(\mathcal{S}^\perp)^\perp = \mathcal{S}$. So, we will prove $\ker(A^\top) = [\text{im } A]^\perp$.

$\ker(A^\top) \subset [\text{im } A]^\perp$:
If $z \in \ker(A^\top)$ and $y \in \text{im } A$. Then, $y = Ax$ for some $x \in \mathbb{R}^n$, and

$$z^\top y = z^\top Ax = 0 \quad \Rightarrow \quad z \in [\text{im } A]^\perp$$

$[\text{im } A]^\perp \subset \ker(A^\top)$:
If $z \in [\text{im } A]^\perp$, then

$$z^\top Ax = 0,\ \forall\ x \in \mathbb{R}^n \quad \Rightarrow \quad (A^\top z)^\top x = 0,\ \forall\ x \in \mathbb{R}^n \quad \Rightarrow \quad A^\top z = 0$$
$$\Rightarrow \quad z \in \ker(A^\top)$$

Linear Algebra Lemma & Local Optimality

Our local optimality necessary condition was, for a regular point x^*,

$$\nabla f(x^*)^\top d = 0, \quad \forall\ d \in \mathcal{V}(x^*)$$

In other words,

$$\nabla f(x^*) \in \mathcal{V}(x^*)^\perp = [\ker \nabla h(x^*)^\top]^\perp \quad \Leftrightarrow \quad \nabla f(x^*) \in \text{im } \nabla h(x^*)$$

Equivalently, there exists $\lambda \in \mathbb{R}^m$ such that

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

Lagrange Multiplier Theorem: Necessary Conditions

$$f:\ \mathbb{R}^n \to \mathbb{R}$$
$$h:\ \mathbb{R}^n \to \mathbb{R}^m$$

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & h(x) = 0, \\ & x \in \mathbb{R}^n \end{array}$$

Theorem. If x^* is a local minimum that is a regular point, then there exists a unique vector $\lambda^* \in \mathbb{R}^m$ called a Lagrange multiplier, such that

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

If, in addition, $f(\cdot)$ and $h(\cdot)$ are twice continuously differentiable,

$$d^\top \left(\nabla^2 f(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla^2 h_i(x^*) \right) d \geq 0, \quad \forall\ d \in \mathcal{V}(x^*).$$

Necessary Conditions: Proof

First order conditions: existence of λ^* follows from earlier discussion, λ^* is unique since the columns of $\nabla h(x^*)$ are linearly independent.

Second order conditions: consider $d \in \mathcal{V}(x^*)$. Define the path $x(t)$ by the regularity lemma with $h(x(t)) = 0$, $x(0) = x^*$, $\dot{x}(0) = d$. If $g(t) \triangleq f(x(t))$, $t = 0$ must be an unconstrained local minimum of $g(t)$. Then,

$$\begin{aligned} 0 \leq \ddot{g}(0) &= \dot{x}(0)^\top \nabla^2 f(x^*) \dot{x}(0) + \ddot{x}(0)^\top \nabla f(x^*) \\ &= d^\top \nabla^2 f(x^*) d + \ddot{x}(0)^\top \nabla f(x^*) \end{aligned}$$

Differentiate $\ell(t) \triangleq \lambda^{*\top} h(x(t)) = 0$ twice at $t = 0$ to obtain

$$0 = \ddot{\ell}(0) = d^\top \left(\sum_{i=1}^{m} \lambda_i \nabla^2 h_i(x^*) \right) d + \ddot{x}(0)^\top \nabla h(x^*) \lambda^*$$

Lagrange Multiplier Theorem: Interpretation

Definition. The Lagrangian function $L: \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is defined by

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

The necessary conditions can written as

$$\nabla_x L(x^*, \lambda^*) = 0, \qquad \nabla_\lambda L(x^*, \lambda^*) = 0,$$
$$d^\top \nabla^2_{xx} L(x^*, \lambda^*) d \geq 0, \qquad \forall\ d \in \mathcal{V}(x^*).$$

- Geometric interpretation
- Penalty function interpretation

The Lagrangian Cookbook Recipe

$$f: \mathbb{R}^n \to \mathbb{R}$$
$$h: \mathbb{R}^n \to \mathbb{R}^m$$

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & h(x) = 0, \\ & x \in \mathbb{R}^n \end{array}$$

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

(i) Check that a global minima exists

(ii) Find the set of (x^*, λ^*) satisfying the necessary conditions

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

(iii) Find the set of non-regular points

(iv) The global minima must be among the points in (ii) and (iii)

(Assuming $f(\cdot)$, $h(\cdot)$ continuously differentiable on $\mathbb{R}^n$)

Example

$$\begin{array}{ll} \text{minimize} & f(x) = x_1 + x_2 \\ \text{subject to} & h_1(x) = x_1^2 + x_2^2 - 2 = 0, \\ & x \in \mathbb{R}^2 \end{array}$$

Objective & constraints continuously differentiable

Global minima exist (Weierstrass)

First order conditions:

$$\begin{aligned} 1 + 2\lambda^* x_1^* &= 0 \\ 1 + 2\lambda^* x_2^* &= 0 \\ (x_1^*)^2 + (x_2^*)^2 &= 2 \end{aligned} \qquad (x_1^*, x_2^*, \lambda^*) = \begin{cases} (-1, -1, \frac{1}{2}) \\ (1, 1, -\frac{1}{2}) \end{cases}$$

Regularity: $\nabla h_1(x) = (2x_1, 2x_2) \neq 0$ Global minimum: $x^* = (-1, -1)$

Example: Maximum Volume Box

$$\begin{array}{ll} \text{maximize} & f(x) = x_1 x_2 x_3 \\ \text{subject to} & h_1(x) = x_1 x_2 + x_2 x_3 + x_1 x_3 - c/2 = 0, \quad (c > 0) \\ & x \geq 0, \ x \in \mathbb{R}^3 \end{array}$$

Objective & equality constraints continuously differentiable

Global maxima exist (why?), must have $x^* > 0$

First order conditions: (when $x^* > 0$)

$$\begin{aligned} x_2^* x_3^* + \lambda^*(x_2^* + x_3^*) &= 0 \\ x_1^* x_3^* + \lambda^*(x_1^* + x_3^*) &= 0 \\ x_1^* x_2^* + \lambda^*(x_1^* + x_2^*) &= 0 \\ x_1^* x_2^* + x_2^* x_3^* + x_1^* x_3^* &= c/2 \end{aligned} \qquad \begin{aligned} x_1^* = x_2^* = x_3^* &= \sqrt{c/6} \\ \lambda^* &= -\tfrac{1}{2}\sqrt{c/6} \end{aligned}$$

Regularity: $\nabla h_1(x) = (x_2 + x_3, x_1 + x_3, x_1 + x_2) \neq 0$ if $x > 0$

Unique global maximum: $x_1^* = x_2^* = x_3^* = \sqrt{c/6}$

Example

$$\begin{array}{ll} \text{minimize} & f(x) = x_1 + x_2 \\ \text{subject to} & h_1(x) = (x_1 - 1)^2 + x_2^2 - 1 = 0 \\ & h_2(x) = (x_1 - 2)^2 + x_2^2 - 4 = 0 \\ & x \in \mathbb{R}^2 \end{array}$$

$x^* = (0, 0)$ is the only feasible point, thus global minimum

First order conditions:

$$\begin{aligned} 1 + 2\lambda_1^*(x_1^* - 1) + 2\lambda_2^*(x_1^* - 2) &= 0 \\ 1 + 2\lambda_1^* x_2^* + 2\lambda_2^* x_2^* &= 0 \\ (x_1^* - 1)^2 + (x_2^*)^2 &= 1 \\ (x_1^* - 2)^2 + (x_2^*)^2 &= 4 \end{aligned}$$

No solution to necessary conditions!

Regularity:

$$\nabla h_1(x) = (2x_1 - 2, 2x_2) \qquad \nabla h_2(x) = (2x_1 - 4, 2x_2)$$

$x^* = (0, 0)$ is not regular!

Constraint Qualification

Let x^* be a local minimum.

Constraint qualification refers to conditions on the constraints that guarantee the existence of Lagrange multipliers satisfying necessary conditions at x^*. Examples:

- Regularity
- Linear constraints [homework!]

More generally:

- $\mathcal{D}(x^*) \cap \mathcal{T}(x^*) = \emptyset$ (since x^* is a local minimum)
- $\mathcal{D}(x^*) \cap \mathcal{V}(x^*) = \emptyset$ implies the existence of Lagrange multipliers

Definition. A feasible point x^* is quasiregular if $\mathcal{V}(x^*) = \mathcal{T}(x^*)$.

Regularity

Note: Regularity is a property of the representation of constraints, not the constraint set.

Example. $h(x) = x_1 \;\Rightarrow\; \mathcal{C} = \{x \in \mathbb{R}^2 \;:\; h(x) = 0\} = \{(x_1, x_2) \;:\; x_1 = 0\}$
All points in $\mathcal{C}$ are regular.

Example.
$h(x) = x_1^2 \;\Rightarrow\; \mathcal{C} = \{x \in \mathbb{R}^2 \;:\; h(x) = 0\} = \{(x_1, x_2) \;:\; x_1 = 0\}$
No points in $\mathcal{C}$ are regular.

Lagrange Multiplier Theorem: Sufficient Conditions

$$f:\ \mathbb{R}^n \to \mathbb{R}$$
$$h:\ \mathbb{R}^n \to \mathbb{R}^m$$

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & h(x) = 0, \\ & x \in \mathbb{R}^n \end{array}$$

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

Theorem. Assume that $f(\cdot)$ and $h(\cdot)$ are twice continuously differentiable, and that $x^* \in \mathbb{R}^n$ and $\lambda^* \in \mathbb{R}^m$ satisfy

$$\nabla_x L(x^*, \lambda^*) = 0, \qquad \nabla_\lambda L(x^*, \lambda^*) = 0,$$
$$d^\top \nabla^2_{xx} L(x^*, \lambda^*) d > 0, \quad \forall\ d \in \mathcal{V}(x^*) \setminus \{0\}.$$

Then, x^* is a strict local minimum.

Sufficient Conditions: Proof

Clearly x^* is feasible. Suppose x^* is not a strict local minimum. Then, there exists $\{x_k\} \subset \mathbb{R}^n$, $h(x_k) = 0$, $x_k \neq x^*$, $x_k \to x^*$, with $f(x_k) \leq f(x^*)$. Define

$$d_k = \frac{x_k - x^*}{\|x_k - x^*\|}, \quad \delta_k = \|x_k - x^*\|$$

Then $\delta_k \to 0$, and $\{d_k\}$ must have a subsequence converging to some d with $\|d\| = 1$. WLOG, assume $d_k \to d$.

By the mean value theorem, there exists $\tilde{x}_k$ between x^* and x_k with

$$h(x_k) = h(x^*) + \nabla h(\tilde{x}_k)^\top (\delta_k d_k) \quad \Rightarrow \quad \nabla h(\tilde{x}_k)^\top d_k = 0$$

Taking a limit as $k \to \infty$, $\nabla h(x^*)^\top d = 0$. Thus, $d \in \mathcal{V}(x^*)$.

Sufficient Conditions: Proof

By the second order Taylor expansion (with remainder),

$$0 = h_i(x_k) = h_i(x^*) + \delta_k \nabla h_i(x^*)^\top d_k + \tfrac{1}{2}\delta_k^2 d_k^\top \nabla^2 h_i(\hat{x}_{i,k}) d_k$$

$$0 \geq f(x_k) - f(x^*) = \delta_k \nabla f(x^*)^\top d_k + \tfrac{1}{2}\delta_k^2 d_k^\top \nabla^2 f(\hat{x}_{0,k}) d_k$$

(Each $\hat{x}_{i,k}$ is a point on the line segment between x_k and x^*.)

Adding these two equations and using the fact that $\nabla_x L(x^*, \lambda^*) = 0$,

$$0 \geq \tfrac{1}{2}\delta_k^2 d_k^\top \left(\nabla^2 f(\hat{x}_{0,k}) + \sum_{i=1}^m \lambda_i \nabla^2 h_i(\hat{x}_{i,k}) \right) d_k$$

Dividing by $\frac{1}{2}\delta_k^2$ and taking a limit as $k \to \infty$, we have

$$0 \geq d^\top \left(\nabla^2 f(x^*) + \sum_{i=1}^m \lambda_i \nabla^2 h_i(x^*) \right) d$$

Since $d \in \mathcal{V}(x^*) \setminus \{0\}$, this contradicts the assumed second order condition.

Lagrange Multiplier Theorem: Sensitivity Analysis

Consider a family of problems, parameterized by $u \in \mathbb{R}^m$:

$$f : \mathbb{R}^n \to \mathbb{R}$$
$$h : \mathbb{R}^n \to \mathbb{R}^m$$

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & h(x) = u, \\ & x \in \mathbb{R}^n \end{array}$$

Theorem. Suppose there exists a local minimum-Lagrange multiplier pair (x^*, λ^*) satisfying the second order sufficient conditions when $u = 0$, with x^* regular. Then, there exists a neighborhood $N_\epsilon(0) \subset \mathbb{R}^m$ of $u = 0$ and a function $x^*(\cdot)$ defined on $N_\epsilon(0)$ such that

(i) $x^*(0) = x^*$, and for each $u \in N_\epsilon(0)$, $x^*(u)$ is a strict local minimum

(ii) $x^*(\cdot)$ is continuously differentiable

(iii) If $p(u) = f(x^*(u))$, then

$$\nabla p(0) = -\lambda^*$$

Sensitivity Analysis: Proof

(i) & (ii): Consider, for $u \in \mathbb{R}^m$ the following system of equations in (x, λ):

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

At $u = 0$, this has the gradient

$$\begin{bmatrix} \nabla^2_{xx} L(x^*, \lambda^*) & \nabla h(x^*) \\ \nabla h(x^*)^\top & 0 \end{bmatrix}$$

which is non-singular by the second order sufficient conditions. By the implicit function theorem, we can define $(x^*(u), \lambda^*(u))$ satisfying the first order conditions for all u in some $N_\epsilon(0)$. Second order sufficient conditions follow for u sufficiently close to zero from continuity assumptions.

Sensitivity Analysis: Proof

(iii): Note that, for all $u \in N_\epsilon(0)$,

$$\nabla x^*(u) \left(\nabla f(x^*(u)) + \nabla h(x^*(u))\lambda^*(u)\right) = 0$$

Differentiating $h(x^*(u)) = u$,

$$I = \nabla_u\{h(x^*(u))\} = \nabla x^*(u)\nabla h(x^*(u))$$

Then,

$$\nabla p(u) = \nabla_u\{f(x^*(u))\} = \nabla x^*(u)\nabla f(x^*(u)) = -\lambda^*(u)$$

Application: Portfolio Optimization

Consider the portfolio optimization problem without short sale constraints:

$$\begin{array}{rll} \sigma^2 = & \text{minimize} & x^\top \Gamma x \\ & \text{subject to} & \mathbf{1}^\top x = 1, \\ & & \mu^\top x = \bar{\mu}, \\ & & x \in \mathbb{R}^n \end{array}$$

Here, we assume that $\Gamma \succ 0$ and $\mathbf{1}$ and μ are linearly independent. First order conditions:

$$2\Gamma x^* + \lambda_1^* \mathbf{1} + \lambda_2^* \mu = 0,\ \mathbf{1}^\top x^* = 1,\ \mu^\top x^* = \bar{\mu}$$

Then,

$$x^* = -\tfrac{1}{2}\Gamma^{-1}\mathbf{1}\lambda_1^* - \tfrac{1}{2}\Gamma^{-1}\mu\lambda_2^*$$

$$1 = \mathbf{1}^\top x^* = -\tfrac{1}{2}\mathbf{1}^\top\Gamma^{-1}\mathbf{1}\lambda_1^* - \tfrac{1}{2}\mathbf{1}^\top\Gamma^{-1}\mu\lambda_2^*$$

$$\bar{\mu} = \mu^\top x^* = -\tfrac{1}{2}\mu^\top\Gamma^{-1}\mathbf{1}\lambda_1^* - \tfrac{1}{2}\mu^\top\Gamma^{-1}\mu\lambda_2^*$$

Application: Portfolio Optimization

The system of equations for $(\lambda_1^*, \lambda_2^*)$ is non-singular if $\mathbf{1}$ and μ are linearly independent and $\Gamma \succ 0$, so

$$\lambda_1^* = \eta_1 + \zeta_1 \bar{\mu}, \quad \lambda_2^* = \eta_2 + \zeta_2 \bar{\mu}$$

for some scalars η_1, η_2, ζ_1, ζ_2 (depending on Γ and μ)

$$\Rightarrow \quad x^* = \bar{\mu} v + w$$

for some vectors v, w (depending on Γ and μ)

$$\Rightarrow \quad \sigma^2 = (\bar{\mu} v + w)^\top \Gamma (\bar{\mu} v + w) = (\alpha \bar{\mu} + \beta)^2 + \gamma$$

for some scalars α, β, γ (depending on Γ and μ)

Inequality Constrained Optimization

Consider

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & h_1(x) = 0, \ldots, h_m(x) = 0, \\ & g_1(x) \leq 0, \ldots, g_r(x) \leq 0, \\ & x \in \mathbb{R}^n \end{array}$$

where

$$f : \mathbb{R}^n \to \mathbb{R}, \quad h_i : \mathbb{R}^n \to \mathbb{R}, \quad g_j : \mathbb{R}^n \to \mathbb{R}$$

- Assume that $f(\cdot)$, $\{h_i(\cdot)\}$, $\{g_j(\cdot)\}$ are continuously differentiable on $\mathbb{R}^n$
- The necessary and sufficient conditions are also true if these functions are just defined and continuously differentiable in a neighborhood of the local minimum

Reduction to Equality Constraints

$$f : \mathbb{R}^n \to \mathbb{R}$$
$$h : \mathbb{R}^n \to \mathbb{R}^m$$
$$g : \mathbb{R}^n \to \mathbb{R}^r$$

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & h(x) = 0, \\ & g(x) \leq 0, \\ & x \in \mathbb{R}^n \end{array}$$

Definition. Given a feasible point $x^* \in \mathbb{R}^n$, the set of active inequality constraints $\mathcal{A}(x^*)$ is defined by

$$\mathcal{A}(x^*) \triangleq \{j \ : \ g_j(x^*) = 0\} \subset \{1, \ldots, r\}$$

Reduction to Equality Constraints

$$f : \mathbb{R}^n \to \mathbb{R}$$
$$h : \mathbb{R}^n \to \mathbb{R}^m$$
$$g : \mathbb{R}^n \to \mathbb{R}^r$$

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & h(x) = 0, \\ & g(x) \leq 0, \\ & x \in \mathbb{R}^n \end{array}$$

Lemma. Let x^* be a local minimum for the inequality constrained program (ICP). Then, it is also a local minimum for the equality constrained program (ECP)

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & h(x) = 0, \\ & g_j(x) = 0, \quad j \in \mathcal{A}(x^*) \\ & x \in \mathbb{R}^n \end{array}$$

Reduction to Equality Constraints: Proof

Suppose x^* is not a local minimum for (ECP). Then, there is a sequence of points $\{x_k\}$ feasible for (ECP), such that $x_k \to x$, and $f(x_k) < f(x^*)$.

Since $g(\cdot)$ is continuous, we have $g(x_k) \to g(x^*)$. In particular, if $j \notin \mathcal{A}(x^*)$,

$$g_j(x_k) \to g_j(x^*) < 0$$

Thus, for k sufficiently large, $g_j(x_k) < 0$ and x_k is feasible for (ICP). This contradicts the local optimality of x^* for (ICP).

Regularity

$$\begin{aligned} f &: \mathbb{R}^n \to \mathbb{R} \\ h &: \mathbb{R}^n \to \mathbb{R}^m \\ g &: \mathbb{R}^n \to \mathbb{R}^r \end{aligned}$$

$$\begin{aligned} \text{minimize} \quad & f(x) \\ \text{subject to} \quad & h(x) = 0, \\ & g(x) \leq 0, \\ & x \in \mathbb{R}^n \end{aligned}$$

Definition. A point $x^* \in \mathbb{R}^n$ is a regular point if it is feasible and if the set of constraint gradients

$$\{\nabla h_i(x^*) \ : \ 1 \leq i \leq m\} \cup \{\nabla g_j(x^*) \ : \ j \in \mathcal{A}(x^*)\}$$

are linearly independent.

Definition. The cone $\mathcal{V}^{\text{EQ}}(x^*)$ at a point $x^* \in \mathbb{R}^n$ is the set of vectors $d \in \mathbb{R}^n$ such that

$$\nabla h_i(x^*)^\top d = 0, \ \forall \ 1 \leq i \leq m, \quad \nabla g_j(x^*)^\top d = 0, \ \forall \ j \in \mathcal{A}(x^*).$$

Karush-Kuhn-Tucker Theorem: Necessary Conditions

$$f:\ \mathbb{R}^n \to \mathbb{R}$$
$$h:\ \mathbb{R}^n \to \mathbb{R}^m$$
$$g:\ \mathbb{R}^n \to \mathbb{R}^r$$

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & h(x) = 0, \\ & g(x) \leq 0, \\ & x \in \mathbb{R}^n \end{array}$$

Theorem. If x^* is a local minimum that is a regular point, then there exists a unique Lagrange multiplier vectors $\lambda^* \in \mathbb{R}^m$ and $\mu^* \in \mathbb{R}^r$, such that

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

$$\mu_j^* \geq 0, \quad \forall\ 1 \leq j \leq r, \qquad \mu_j^* = 0, \quad \forall\ j \notin \mathcal{A}(x^*).$$

If, in addition, $f(\cdot)$, $h(\cdot)$, and $g(\cdot)$ are twice continuously differentiable, for all $d \in \mathcal{V}^{\mathsf{EQ}}(x^*)$,

$$d^\top \left(\nabla^2 f(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla^2 h_i(x^*) + \sum_{j=1}^{r} \mu_j^* \nabla^2 g_j(x^*) \right) d \geq 0$$

Necessary Conditions: Proof

Everything follows by apply the necessary conditions from the Lagrange multiplier theorem to the equality constrained program defined by the active constraints, except the assertion that $\mu_j^* \geq 0$, for $j \in \mathcal{A}(x^*)$.

Suppose this does not hold for some j. Then, let $\mathcal{C}_j \subset \mathbb{R}^n$ be the set of points feasible for all other active constraints,

$$\mathcal{C}_j \triangleq \left\{ x\ :\ h(x) = 0,\ g_k(x) = 0,\ \forall\ k \in \mathcal{A}(x^*) \setminus \{j\} \right\}$$

and $\mathcal{V}_j^{\mathsf{EQ}}(x^*)$ the corresponding cone of first order feasible directions,

$$\mathcal{V}_j^{\mathsf{EQ}}(x^*) \triangleq \left\{ d\ :\ \nabla h(x^*)^\top d = 0,\ \nabla g_k(x^*)^\top d = 0,\ \forall\ k \in \mathcal{A}(x^*) \setminus \{j\} \right\}$$

Necessary Conditions: Proof

Note that:

(i) There exists $d \in \mathcal{V}_j^{\mathsf{EQ}}(x^*)$ with $\nabla g_j(x^*)^\top d < 0$.
Otherwise, $\nabla g_j(x^*) \in [\mathcal{V}_j^{\mathsf{EQ}}(x^*)]^\perp$. Then, $\nabla g_j(x^*)$ is in the span of

$$\{\nabla h_i(x^*) \ : \ 1 \le i \le m\} \cup \{\nabla g_k(x^*) \ : \ k \in \mathcal{A}(x^*) \setminus \{j\}\}.$$

This contradicts regularity.

(ii) By the regularity theorem applied to the constraint set $\mathcal{C}_j$, there exists a curve $x(\cdot) \in \mathcal{C}_j$ with $x(0) = x^*$, $\dot{x}(0) = d$.

(iii) For sufficiently small $t \ge 0$, $g(x(t)) \le 0$ since $\nabla g_j(x^*)^\top d < 0$, hence $x(t)$ is also feasible for the original problem.

Then, if $\ell(t) \triangleq f(x(t))$,

$$\begin{aligned} \dot{\ell}(0) &= d^\top \nabla f(x^*) = -d^\top (\nabla h(x^*)\lambda^* - \nabla g(x^*)\mu^*) \\ &= -d^\top \nabla g_j(x^*)\mu_j^* < 0, \end{aligned}$$

contradicting the local optimality of x^* for the original problem.

KKT Theorem: Interpretation

Definition. The Lagrangian function $L : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^r \to \mathbb{R}$ is defined by

$$\begin{aligned} L(x, \lambda, \mu) &= f(x) + \sum_{i=1}^m \lambda_i h_i(x) + \sum_{j=1}^r \mu_j g_j(x) \\ &= f(x) + \lambda^\top h(x) + \mu^\top g(x) \end{aligned}$$

The first order necessary conditions can be written as:

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

$$\mu^* \ge 0,$$

$$\mu_j^* g_j(x^*) = 0, \quad \forall \ 1 \le j \le r$$

The second order necessary conditions can be written as:

$$d^\top \nabla_{xx}^2 L(x^*, \lambda^*, \mu^*) d \ge 0, \quad \forall \ d \in \mathcal{V}^{\mathsf{EQ}}(x^*)$$

KKT Cookbook Recipe

$$f:\ \mathbb{R}^n \to \mathbb{R}$$
$$h:\ \mathbb{R}^n \to \mathbb{R}^m$$
$$g:\ \mathbb{R}^n \to \mathbb{R}^r$$

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & h(x) = 0, \\ & g(x) \leq 0, \\ & x \in \mathbb{R}^n \end{array}$$

$$L(x, \lambda, \mu) = f(x) + \lambda^\top h(x) + \mu^\top g(x)$$

(i) Check that a global minima exists

(ii) Find the set of (x^*, λ^*, μ^*) satisfying the necessary conditions

$$\nabla_x L(x^*, \lambda^*, \mu^*) = 0,\ h(x^*) = 0,\ g(x^*) \leq 0,$$
$$\mu^* \geq 0, \quad \mu_j^* g_j(x^*) = 0,\ \forall\ 1 \leq j \leq r$$

(iii) Find the set of non-regular points

(iv) The global minima must be among the points in (ii) and (iii)

(Assuming $f(\cdot)$, $h(\cdot)$, $g(\cdot)$ continuously differentiable on $\mathbb{R}^n$)

Example

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}(x_1^2 + x_2^2 + x_3^2) \\ \text{subject to} & x_1 + x_2 + x_2 \leq -3, \\ & x \in \mathbb{R}^3 \end{array}$$

Objective & constraints continuously differentiable

Global minima exist (coerciveness)

First order conditions:

$$\begin{array}{r} x_1^* + \mu^* = 0 \\ x_2^* + \mu^* = 0 \\ x_3^* + \mu^* = 0 \\ x_1^* + x_2^* + x_2^* \leq -3 \\ \mu^*(x_1^* + x_2^* + x_2^* + 3) = 0 \end{array}$$

$$x^* = (-1, -1, -1)$$
$$\mu^* = 1$$

All points are regular

Global minimum: $x^* = (-1, -1, -1)$

KKT Theorem: Sufficient Conditions

$$f:\ \mathbb{R}^n \to \mathbb{R}$$
$$h:\ \mathbb{R}^n \to \mathbb{R}^m$$
$$g:\ \mathbb{R}^n \to \mathbb{R}^r$$

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & h(x) = 0, \\ & g(x) \leq 0, \\ & x \in \mathbb{R}^n \end{array}$$

Theorem. Assume that $f(\cdot)$, $h(\cdot)$ and $g(\cdot)$ are twice continuously differentiable, and that $x^* \in \mathbb{R}^n$, $\lambda^* \in \mathbb{R}^m$, $\mu^* \in \mathbb{R}^r$ satisfy

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

$$\mu^* \geq 0, \quad \mu_j^* = 0,\ \forall\ j \notin \mathcal{A}(x^*),$$

$$d^\top \nabla^2_{xx} L(x^*, \lambda^*) d > 0,\ \forall\ d \in \mathcal{V}^{\text{EQ}}(x^*) \setminus \{0\}$$

Assume also that

$$\mu_j^* > 0,\ \forall\ j \in \mathcal{A}(x^*)$$

Then, x^* is a strict local minimum.

Sufficient Conditions: Proof

Following the equality case: Suppose x^* is not a strict local minimum. Then, there exists $\{x_k\} \subset \mathbb{R}^n$, $h(x_k) = 0$, $g(x_k) \leq 0$, $x_k \neq x^*$, $x_k \to x^*$, with $f(x_k) \leq f(x^*)$. Define

$$d_k = \frac{x_k - x^*}{\|x_k - x^*\|}, \quad \delta_k = \|x_k - x^*\|$$

Then $\delta_k \to 0$, and $\{d_k\}$ must have a subsequence converging to some d with $\|d\| = 1$. WLOG, assume $d_k \to d$.

As in the equality case, $\nabla h(x^*)^\top d = 0$. If $j \in \mathcal{A}(x^*)$, by the mean value theorem,

$$g_j(x_k) - g_j(x^*) \leq 0 \quad \Rightarrow \quad \nabla g_j(x^*)^\top d \leq 0$$

If $\nabla g_j(x^*)^\top d = 0$ for all $j \in \mathcal{A}(x^*)$, then $d \in \mathcal{V}^{\text{EQ}}(x^*)$, and we proceed as before.

Sufficient Conditions: Proof

Suppose for some $j \in \mathcal{A}(x^*)$, $\nabla g_j(x^*)^\top d < 0$. Then,

$$d^\top \nabla f(x^*) = -d^\top (\nabla h(x^*)\lambda^* + \nabla g(x^*)\mu^*) > 0$$

However, by the mean value theorem,

$$f(x_k) - f(x^*) \leq 0 \quad \Rightarrow \quad \nabla f(x^*)^\top d \leq 0$$

KKT Theorem: Sensitivity Analysis

Consider a family of problems, parameterized by $u \in \mathbb{R}^m$ and $v \in \mathbb{R}^r$:

$$\begin{aligned} f &: \mathbb{R}^n \to \mathbb{R} \\ h &: \mathbb{R}^n \to \mathbb{R}^m \\ g &: \mathbb{R}^n \to \mathbb{R}^r \end{aligned}$$

$$\begin{aligned} \text{minimize} \quad & f(x) \\ \text{subject to} \quad & h(x) = u, \\ & g(x) \leq v, \\ & x \in \mathbb{R}^n \end{aligned}$$

Theorem. Suppose there exists a triple (x^*, λ^*, μ^*) satisfying the second order sufficient conditions when $(u, v) = (0, 0)$, with x^* regular. Then, there exists a neighborhood N of $(u, v) = (0, 0)$ and a function $x^*(\cdot, \cdot)$ defined on N such that

(i) $x^*(0, 0) = x^*$, and for each $(u, v) \in N$, $x^*(u, v)$ is a strict local minimum

(ii) $x^*(\cdot, \cdot)$ is continuously differentiable

(iii) If $p(u, v) = f(x^*(u, v))$, then

$$\nabla_u p(0, 0) = -\lambda^*, \quad \nabla_v p(0, 0) = -\mu^*$$

Geometric Interpretation

$$f:\ \mathbb{R}^n \to \mathbb{R}$$
$$h:\ \mathbb{R}^n \to \mathbb{R}^m$$
$$g:\ \mathbb{R}^n \to \mathbb{R}^r$$

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & h(x) = 0, \\ & g(x) \leq 0, \\ & x \in \mathbb{R}^n \end{array}$$

Assume that x^* is a feasible point and $d \in \mathbb{R}^n$ is a direction. For small $\alpha > 0$,

$$h(x^* + \alpha d) \approx h(x^*) + \nabla h(x^*)^\top (\alpha d) = \alpha \nabla h(x^*)^\top d$$

If $j \in \mathcal{A}(x^*)$,

$$g(x^* + \alpha d) \approx g(x^*) + \nabla g(x^*)^\top (\alpha d) = \alpha \nabla g(x^*)^\top d$$

Definition. The cone of first order feasible variations at a point $x^* \in \mathbb{R}^n$ is the set

$$\mathcal{V}(x^*) = \{d \in \mathbb{R}^n : \nabla h(x^*)^\top d = 0,\ \nabla g_j(x^*)^\top d \leq 0,\ \forall\, j \in \mathcal{A}(x^*)\}$$

Farkas' Lemma

Lemma. (Farkas) Consider a matrix $A \in \mathbb{R}^{m \times n}$. Then, a vector $z \in \mathbb{R}^m$ satisfies

$$z^\top y \leq 0 \text{ for all } y \in \mathbb{R}^m \text{ with } A^\top y \leq 0$$

if and only if

$$z = Ax \text{ for some } x \in \mathbb{R}^n \text{ with } x \geq 0$$

Existence of Lagrange Multipliers

$$f: \mathbb{R}^n \to \mathbb{R}$$
$$h: \mathbb{R}^n \to \mathbb{R}^m$$
$$g: \mathbb{R}^n \to \mathbb{R}^r$$

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & h(x) = 0, \\ & g(x) \leq 0, \\ & x \in \mathbb{R}^n \end{array}$$

Theorem. Let x^* be a local minimum. Then there exist Lagrange multipliers (λ^*, μ^*) satisfying

$$\nabla f(x^*) + \nabla h(x^*)\lambda^* + \nabla g(x^*)\mu^* = 0,$$
$$\mu^* \geq 0, \quad \mu_j^* g_j(x^*) = 0, \ \forall\ 1 \leq j \leq r,$$

if and only if

$$\mathcal{D}(x^*) \cap \mathcal{V}(x^*) = \emptyset.$$

Existence of Lagrange Multipliers: Proof

Suppose there are no equality constraints. Then,

$$\mathcal{D}(x^*) \cap \mathcal{V}(x^*) = \emptyset$$

is equivalent to

$\nabla f(x^*)^\top d \geq 0$ for all d such that $\nabla g_j(x^*)^\top d \leq 0, \ \forall\ j \in \mathcal{A}(x^*)$

By Farkas' Lemma, this is equivalent to

$$\nabla f(x^*) + \nabla g(x^*)\mu = 0$$

for some μ with $\mu \geq 0$ and $\mu_j = 0$ if $j \notin \mathcal{A}(x^*)$.

If there are equality constraints $h(x) = 0$, we can add inequality constraints $h(x) \leq 0$ and $-h(x) \leq 0$.

Constraint Qualification

Let x^* be a local minimum.

- $\mathcal{D}(x^*) \cap \mathcal{T}(x^*) = \emptyset$ (since x^* is a local minimum)
- $\mathcal{D}(x^*) \cap \mathcal{V}(x^*) = \emptyset$ implies the existence of Lagrange multipliers
- Quasiregularity: $\mathcal{V}(x^*) = \mathcal{T}(x^*)$

Corollary. Let x^* be a local minimum that is quasiregular. Then, Lagrange multipliers exist satisfying the KKT first order necessary conditions.

Linear Constraint Qualification

$$f : \mathbb{R}^n \to \mathbb{R}$$
$$A \in \mathbb{R}^{m \times n}, \; b \in \mathbb{R}^m$$

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & Ax \leq b, \\ & x \in \mathbb{R}^n \end{array}$$

Theorem. Suppose that x^* is a local minimum. Then, x^* is quasiregular and Lagrange multipliers exist.

Note: Trivially applies to linear equality constraints also. Can be extended to linear equality constraints and concave inequality constraints.

A General Sufficiency Condition

$f : \mathbb{R}^n \to \mathbb{R}$
$g : \mathbb{R}^n \to \mathbb{R}^r$
$\Omega \subset \mathbb{R}^n$

minimize $f(x)$
subject to $g(x) \leq 0,$
$x \in \Omega$

Theorem. Let $x^* \in \mathbb{R}^n$ be a feasible point and $\mu^* \in \mathbb{R}^r$ be a vector such that

$$\mu^* \geq 0,$$
$$\mu_j^* = 0, \quad \forall\, j \notin \mathcal{A}(x^*),$$
$$x^* \in \operatorname*{argmin}_{x \in \Omega} L(x, \mu^*)$$

Then, x^* is a global minimum.

Note: No differentiability or continuity assumptions made!

A General Sufficiency Condition: Proof

$$\begin{aligned} f(x^*) &= f(x^*) + (\mu^*)^\top g(x^*) \\ &= \min_{x \in \Omega} \; f(x) + (\mu^*)^\top g(x) \\ &\leq \min_{x \in \Omega, g(x) \leq 0} \; f(x) + (\mu^*)^\top g(x) \\ &\leq \min_{x \in \Omega, g(x) \leq 0} \; f(x) \end{aligned}$$

B9824 Foundations of Optimization

Lecture 3: Convexity

Fall 2011

Outline

1. Convex sets
2. Convex functions
3. Projection theorem

Affine Sets

Consider a set $\mathcal{C} \subset \mathbb{R}^n$.

Definition. The set $\mathcal{C}$ is affine if, for all points $x_1, x_2 \in \mathcal{C}$, and a scalar $\lambda \in \mathbb{R}$,

$$\lambda x_1 + (1 - \lambda) x_2 \in \mathcal{C}$$

Example. The empty set is affine. Any line is affine. Any subspace is affine.

Example. If $\mathcal{C}$ is the solution to a set of linear equations, e.g.

$$\mathcal{C} = \{x \in \mathbb{R}^n \ : \ Ax = b\},$$

for some matrix $A \in \mathbb{R}^{m \times n}$ and vector $b \in \mathbb{R}^m$, then $\mathcal{C}$ is an affine set.

Affine Sets

Definition. Given a set of points $\mathcal{X} \subset \mathbb{R}^n$, the affine hull **aff** $\mathcal{X}$ is the set of points

$$\lambda_1 x_1 + \cdots + \lambda_k x_k,$$

where $k \geq 1$, $\{x_i\} \subset \mathcal{X}$, and

$$\lambda_1 + \cdots + \lambda_k = 1.$$

The affine hull **aff** $\mathcal{X}$ is affine, and is the smallest affine set containing $\mathcal{X}$.

Convex Sets

Definition. The set $\mathcal{C}$ is convex if, for all points $x_1, x_2 \in \mathcal{C}$, and scalars $0 \leq \lambda \leq 1$,

$$\lambda x_1 + (1-\lambda)x_2 \in \mathcal{C}$$

Remark. Clearly affine sets are also convex.

Definition. Given a set of points $\mathcal{X} \subset \mathbb{R}^n$, the convex hull **conv** $\mathcal{X}$ is the set of points

$$\lambda_1 x_1 + \cdots + \lambda_k x_k,$$

where $k \geq 1$, $\{x_i\} \subset \mathcal{X}$, $\lambda \geq 0$, and

$$\lambda_1 + \cdots + \lambda_k = 1.$$

The convex hull **conv** $\mathcal{X}$ is convex, and is the smallest convex set containing $\mathcal{X}$.

Hyperplanes and Halfspaces

Definition. A hyperplane is a set of the form

$$\{x \in \mathbb{R}^n \ : \ a^\top x = b\},$$

where $a \in \mathbb{R}^n \setminus \{0\}$ is a non-zero vector called the normal vector and $b \in \mathbb{R}$ is a scalar.

Hyperplanes are affine and thus convex.

Definition. A halfspace is a set of the form

$$\{x \in \mathbb{R}^n \ : \ a^\top x \leq b\},$$

where $a \in \mathbb{R}^n \setminus \{0\}$ is a non-zero vector and $b \in \mathbb{R}$ is a scalar.

Halfspaces are not affine but are convex.

Norms

Definition. A norm is a real-valued function $\|\cdot\|$ on $\mathbb{R}^n$ such that

- $\|x\| = 0$ if and ony if $x = 0$
- For all $x \in \mathbb{R}^n$ and $\lambda \in \mathbb{R}$, $\|\lambda x\| = |\lambda|\|x\|$
- For all $x_1, x_2 \in \mathbb{R}^n$, $\|x_1 + x_2\| \leq \|x_1\| + \|x_2\|$

Example.

$$\|x\|_2 = \left(\sum_{i=1}^n x_i^2\right)^{\frac{1}{2}} = \sqrt{x^\top x}$$

$$\|x\|_\Gamma = \sqrt{x^\top \Gamma x}, \quad \Gamma \text{ symmetric positive definite}$$

$$\|x\|_p = \left(\sum_{i=1}^n |x_i|^p\right)^{\frac{1}{p}}, \quad p \geq 1 \qquad \|x\|_\infty = \max\left(|x_1|, \ldots, |x_n|\right)$$

Norm Balls

Given a norm $\|\cdot\|$, the (closed) ball with center $x_0 \in \mathbb{R}^n$ and radius $r \geq 0$,

$$\{x \in \mathbb{R}^n \ : \ \|x - x_0\| \leq r\}$$

is convex.

Example. $\|\cdot\|_2 \quad \Rightarrow \quad$ spheres are convex

Example. $\|\cdot\|_\Gamma \quad \Rightarrow \quad$ ellipsoids are convex

Example. $\|\cdot\|_\infty, \|\cdot\|_1 \quad \Rightarrow \quad$ 'boxes' are convex

Elementary Properties of Convex Sets

Theorem. (Scalar Multiplication) If $\mathcal{C} \subset \mathbb{R}^n$ is a convex set, and $\alpha \in \mathbb{R}$ is a scalar, then the set

$$\alpha\mathcal{C} \triangleq \{\alpha x \ : \ x \in \mathcal{C}\}$$

is also convex.

Theorem. (Vector Sum) If $\mathcal{C}, \mathcal{D} \subset \mathbb{R}^n$ are convex sets, then the set

$$\mathcal{C} + \mathcal{D} \triangleq \{x + y \ : \ x \in \mathcal{C}, \ y \in \mathcal{D}\}$$

is also convex.

Theorem. (Affine Transformations) If $\mathcal{C} \subset \mathbb{R}^n$ is a convex set, $A \in \mathbb{R}^{m \times n}$ a matrix, and $b \in \mathbb{R}^m$ a vector, then the set

$$\{Ax + b \ : \ x \in \mathcal{C}\}$$

is a convex subset of $\mathbb{R}^m$.

Elementary Properties of Convex Sets

Theorem. (Intersection) If $\mathcal{K}$ is an arbitrary collection of convex sets, then the intersection

$$\bigcap_{\mathcal{C} \in \mathcal{K}} \mathcal{C}$$

is also convex.

Polyhedra

Definition. A set $\mathcal{P}$ is a polyhedron if it is of the form

$$\mathcal{P} = \{x \in \mathbb{R}^n \ : \ Ax \leq b\},$$

for a matrix $A \in \mathbb{R}^{m \times n}$ and a vector $b \in \mathbb{R}^m$.

Remark. Linear equality constraints are also trivially allowed.

Polyhedra are convex.

Example. The non-negative orthant $\{x \in \mathbb{R}^n \ : \ x \geq 0\}$.

Example. The unit simplex $\{x \in \mathbb{R}^n \ : \ x \geq 0, \ \mathbf{1}^\top x \leq 1\}$.

Example. The probability simplex $\{x \in \mathbb{R}^n \ : \ x \geq 0, \ \mathbf{1}^\top x = 1\}$.

Cones

Consider a set $\mathcal{C} \subset \mathbb{R}^n$.

Definition. The set $\mathcal{C}$ is a cone if, for all points $x \in \mathcal{C}$, $\lambda \geq 0$,

$$\lambda x \in \mathcal{C}$$

Definition. The set $\mathcal{C}$ is a convex cone if it is convex and a cone, i.e., for all $x_1, x_2 \in \mathcal{C}$ and $\lambda_1, \lambda_2 \geq 0$,

$$\lambda_1 x_1 + \lambda_2 x_2 \in \mathcal{C}$$

Cones

Example. Given a collection of points $\mathcal{X} \subset \mathbb{R}^n$, the conic hull consists of the points

$$\lambda_1 x_1 + \cdots + \lambda_k x_k,$$

where $k \geq 1$, $\{x_i\} \subset \mathcal{X}$, and $\lambda \geq 0$. This is a convex cone.

Example. Given a norm $\|\cdot\|$ on $\mathbb{R}^n$, the norm cone

$$\{(x, t) \in \mathbb{R}^{n+1} \ : \ \|x\| \leq t\}$$

is a convex cone in $\mathbb{R}^{n+1}$. When $\|\cdot\| = \|\cdot\|_2$, this is known as a second-order cone.

Example. The set of symmetric positive semidefinite matrices

$$S_+^n \triangleq \{X \in R^{n \times n} \ : \ X^\top = X, \ X \succeq 0\}.$$

is a convex cone in $\mathbb{R}^{n \times n}$ known as the positive semidefinite cone.

Generalized Inequalities

Definition. The set $\mathcal{K}$ is a proper cone if it is a closed convex cone that is:

- solid — non-empty interior, i.e., **int** $\mathcal{K} \neq \emptyset$
- pointed — contains no lines, i.e., if $x \in \mathcal{K}$ and $-x \in \mathcal{K}$, then $x = 0$

Definition. A proper cone $\mathcal{K} \subset \mathbb{R}^n$ defines a partial ordering $\preceq_\mathcal{K}$ and a strict partial ordering $\prec_\mathcal{K}$ on $\mathbb{R}^n$ via

$$x \preceq_\mathcal{K} y \iff y - x \in \mathcal{K}$$

$$x \prec_\mathcal{K} y \iff y - x \in \textbf{int } \mathcal{K}$$

Example. If $\mathcal{K} = \mathbb{R}^n_+$ is the positive orthant, then $\preceq_\mathcal{K}$ is the usual component-wise inequality $\leq$.

Example. If $\mathcal{K} = S^n_+$ is the semidefinite cone, then $\preceq_\mathcal{K}$ is the matrix inequality $\preceq$.

Optimization Over Convex Sets: Necessary Condition

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in \mathcal{C} \subset \mathbb{R}^n \end{array}$$

Theorem. Suppose that $\mathcal{C}$ is a convex set, x^* is a local minimum, and $f(\cdot)$ is continuously differentiable in a neighborhood of x^*. Then, for all $x \in \mathcal{C}$,

$$\nabla f(x^*)^\top (x - x^*) \geq 0$$

Necessary Condition: Proof

Suppose there exists $x \in \mathcal{C}$ with

$$\nabla f(x^*)^\top (x - x^*) < 0$$

By the mean value theorem, for a given $\epsilon > 0$ sufficiently small,

$$f\big(x^* + \epsilon(x - x^*)\big) = f(x^*) + \epsilon \nabla f\big(x^* + s\epsilon(x - x^*)\big)^\top (x - x^*),$$

for some $s \in [0, 1]$. Since ∇f is continuous, for ϵ sufficiently small,

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

Then,

$$f\big(x^* + \epsilon(x - x^*)\big) < f(x^*)$$

Since $\mathcal{C}$ is convex, this contradicts local optimality of x^*.

Convex Functions

Definition. Let $\mathcal{X} \subset \mathbb{R}^n$ be a convex set. A real-valued function $f\colon \mathcal{X} \to \mathbb{R}$ is convex if, for all $x_1, x_2 \in \mathcal{X}$ and $\lambda \in [0,1]$,

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2)$$

$f(\cdot)$ is strictly convex if, in addition,

$$f(\lambda x_1 + (1-\lambda)x_2) < \lambda f(x_1) + (1-\lambda) f(x_2)$$

when $x_1 \neq x_2$ and $\lambda \in (0,1)$.

Remark. If $f : \mathbb{R}^n \to \mathbb{R}$ is convex when restricted to a (convex) subset $\mathcal{X} \subset \mathbb{R}$, we will say $f(\cdot)$ is convex over $\mathcal{X}$.

Definition. A function $f(\cdot)$ is concave if $-f(\cdot)$ is convex, it is strictly concave if $-f(\cdot)$ is strictly convex.

Extended-Value Functions

Given a convex function $f(\cdot)$ with (convex) domain $\mathcal{X} \subset \mathbb{R}^n$, we can define the extended-value extension $\tilde{f}\colon \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ by

$$\tilde{f}(x) = \begin{cases} f(x) & \text{if } x \in \mathcal{X} \\ \infty & \text{otherwise} \end{cases}$$

We define

$$\textbf{dom}\ \tilde{f} \triangleq \{x \in \mathbb{R}^n \ :\ \tilde{f}(x) < \infty\}$$

Definition. An extended-value function $g\colon \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is convex if

- the set **dom** g is convex
- for all $x_1, x_2 \in \mathbb{R}^n$ and $\lambda \in [0,1]$,

$$g(\lambda x_1 + (1-\lambda)x_2) \leq \lambda g(x_1) + (1-\lambda) g(x_2)$$

We will sometimes implicitly identify convex functions with their extended-value extensions.

Indicator Functions

Definition. Given a set $\mathcal{C} \subset \mathbb{R}^n$, the indicator function $I_{\mathcal{C}} : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is defined by

$$I_{\mathcal{C}}(x) = \begin{cases} 0 & \text{if } x \in \mathcal{C} \\ \infty & \text{otherwise} \end{cases}$$

If $\mathcal{C}$ is a convex set, then $I_{\mathcal{C}}(\cdot)$ is a convex function.

First-Order Conditions for Convexity

Theorem. Let $\mathcal{C} \subset \mathbb{R}^n$ be a convex set and $f : \mathbb{R}^n \to \mathbb{R}$ a differentiable function. Then,

(i) $f(\cdot)$ is convex over $\mathcal{C}$ if and only if

$$f(x_1) \geq f(x_0) + \nabla f(x_0)^\top (x_1 - x_0), \quad \forall\, x_0, x_1 \in \mathcal{C}$$

(ii) $f(\cdot)$ is strictly convex over $\mathcal{C}$ if and only if the inequality is strict when $x_0 \neq x_1$

First-Order Conditions: Proof

Suppose that the inequality in (i) holds. Suppose $x, y \in \mathcal{C}$, $\lambda \in [0,1]$, and $z = \lambda x + (1-\lambda)y$. Then,

$$f(x) \geq f(z) + \nabla f(z)^\top (x - z), \qquad f(y) \geq f(z) + \nabla f(z)^\top (y - z),$$

Thus,

$$\lambda f(x) + (1-\lambda) f(y) \geq f(z) + \nabla f(z)^\top (\lambda x + (1-\lambda)y - z) = f(z)$$

First-Order Conditions: Proof

Conversely, suppose $f(\cdot)$ is convex. Define, for $x, z \in \mathcal{C}$, $x \neq z$, $\lambda \in (0,1)$

$$g(\lambda) \triangleq \frac{f(x + \lambda(z-x)) - f(x)}{\lambda}$$

Note that, by convexity, $g(\cdot)$ is monotonically increasing.

Then,

$$\nabla f(x)^\top (z - x) = \lim_{\lambda \downarrow 0} g(\lambda) \leq g(1) = f(z) - f(x)$$

(ii) is proved the same way.

Second-Order Conditions

Theorem. Let $\mathcal{C} \subset \mathbb{R}^n$ be a convex set and $f\colon \mathbb{R}^n \to \mathbb{R}$ a twice continuously differentiable function. Then,

(i) if $\nabla^2 f(x) \succeq 0$ for all $x \in \mathcal{C}$, $f(\cdot)$ is convex over $\mathcal{C}$

(ii) if $\nabla^2 f(x) \succ 0$ for all $x \in \mathcal{C}$, $f(\cdot)$ is strictly convex over $\mathcal{C}$

Remark. Previous theorems can be applied if $f\colon \mathcal{C} \to \mathbb{R}$, $\mathcal{C}$ is convex, and differentiability assumptions hold on an open set containing $\mathcal{C}$ (e.g., if $\mathcal{C}$ itself is open).

Second-Order Conditions: Proof

(i) By the second order Taylor expansion, if $x, y \in \mathcal{C}$,

$$f(y) = f(x) + \nabla f(x)^\top (y - x) + \tfrac{1}{2}(y - x)^\top \nabla^2 f(x + \epsilon(y - x))(y - x),$$

for some $\epsilon \in [0, 1]$. If $\nabla^2 f \succeq 0$,

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x)$$

Convexity follows from the first order conditions.

(ii) is proved the same way.

Examples

- *Exponential.* e^{ax} is convex on $\mathbb{R}$, for any $a \in \mathbb{R}$.
- *Powers.* x^a is convex on $(0, \infty)$, for any $a \geq 1$ or $a \leq 0$, and concave for $0 \leq a \leq 1$.
- *Powers of absolute value.* $|x|^p$ is convex on $\mathbb{R}$, for any $p \geq 1$.
- *Logarithm.* $\log x$ is concave on $(0, \infty)$
- *Negative entropy.* $x \log x$ is convex on $(0, \infty)$, or convex on $[0, \infty)$ if we set $0 \log 0 = 0$.

Examples

- *Affine functions.* Any affine function
$$f(x) \triangleq a^\top x + b$$
is concave and convex on $\mathbb{R}^n$.
- *Norms.* Any norm $\|\cdot\|$ on $\mathbb{R}^n$ is convex.
- *Log-sum-exp.* The function
$$f(x) \triangleq \log\left(e^{x_1} + \ldots + e^{x_n}\right)$$
is convex on $\mathbb{R}^n$.
- *Geometric mean.* The function
$$f(x) \triangleq \left(\prod_{i=1}^n x_i\right)^{\frac{1}{n}}$$
is concave on $(0, \infty)^n$.

Operations That Perserve Convexity

- *Nonnegative multiples.* If $f(\cdot)$ is convex and $w \geq 0$, then
$$g(x) \triangleq wf(x)$$
is convex.

- *Sums.* If $f_1(\cdot)$ and $f_2(\cdot)$ are convex, then
$$g(x) \triangleq f_1(x) + f_2(x)$$
is convex.

- *Nonnegative weighted sums.* If $f_1(\cdot), \ldots, f_k(\cdot)$ are convex and $w_i \geq 0$, then
$$g(x) \triangleq w_1 f_1(x) + \cdots + w_k f_k(x)$$
is convex.

Operations That Perserve Convexity

- *Perspective.* If $f(\cdot)$ is convex, then the perspective function of $f(\cdot)$,
$$g(x,t) \triangleq tf(x/t)$$
with domain
$$\textbf{dom } g \triangleq \{(x,t) \in \mathbb{R}^{n+1} \ : \ x/t \in \textbf{dom } f, \ t > 0\}$$
is convex.

Operations That Perserve Convexity

- *Pointwise maxima.* If $f(\cdot, y)$ is convex for each $y \in \mathcal{Y}$, then
$$g(x) \triangleq \sup_{y \in \mathcal{Y}} f(x, y)$$
is an extended-value convex function.

- *Minimization.* If $f(x, y)$ is convex over $(x, y) \in \mathbb{R}^n \times \mathbb{R}^m$, and $\mathcal{C} \subset \mathbb{R}^m$ is convex, then
$$g(x) \triangleq \inf_{y \in \mathcal{C}} f(x, y)$$
is convex provided $g(x) > -\infty$.

- *Composition with affine functions.* If $f(\cdot)$ is convex, then
$$g(x) \triangleq f(Ax + b)$$
is convex.

General Composition

Definition. A function $h\colon \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is non-decreasing iff, for all $x, y \in \mathbb{R}^n$ with $x \leq y$,
$$h(x) \leq h(y)$$
A function $h(\cdot)$ is non-increasing iff $-h(\cdot)$ is non-decreasing.

Theorem. Consider $h\colon \mathbb{R}^k \to \mathbb{R} \cup \{\infty\}$ and $g_i\colon \mathbb{R}^n \Rightarrow \mathbb{R}$, $i = \{1, \ldots, k\}$. Define
$$f(x) \triangleq h\left(g_1(x), \ldots, g_k(x)\right)$$
Then:

- If h is convex and non-decreasing, and $\{g_i\}$ are convex, then f is convex.
- If h is convex and non-increasing, and $\{g_i\}$ are concave, then f is convex.

Examples

Example. Consider a non-empty set $\mathcal{C} \subset \mathbb{R}^n$, define the support function

$$S_{\mathcal{C}}(x) \triangleq \sup_{y \in \mathcal{C}} \; x^\top y$$

Then, $S_{\mathcal{C}}(\cdot)$ is convex.

Example. Consider a non-empty convex set $\mathcal{C} \subset \mathbb{R}^n$, with $0 \in$ **int** $\mathcal{C}$. Define the Minkowski functional

$$p_{\mathcal{C}}(x) \triangleq \inf \{t > 0 \; : \; x/t \in \mathcal{C}\}$$

Then, $p_{\mathcal{C}}(\cdot)$ is convex.

Strategies to Verify Convexity

- Construct from known convex functions using operations that preserve convexity
- Use first- or second-order differentiability properties of convex functions
- Restrict to a line, e.g. $f(\cdot)$ is convex over $\mathcal{C}$ if and only if, for every $x_1, x_2 \in \mathcal{C}$,

  $$g(t) \triangleq f(x_1 + t(x_2 - x_1))$$

  is convex over $[0, 1]$
- Directly verify using the definition of convexity

Sublevel Sets

If $f(\cdot)$ is convex over a convex set $\mathcal{C} \subset \mathbb{R}^n$, every sublevel set

$$\{x \in \mathcal{C} \ : \ f(x) \leq \gamma\}$$

is a convex subset of $\mathbb{R}^n$.

Remark. The converse is not true! For example, $\log x$ is not convex on $(0, \infty)$, however every sublevel set is convex.

Definition. An extended real-valued function $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is quasiconvex if, for all $\gamma \in \mathbb{R}$, the sublevel set

$$\{x \in \mathbb{R}^n : \ f(x) \leq \gamma\}$$

is convex.

Optimality for Convex Optimization

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in \mathcal{C} \subset \mathbb{R}^n \end{array}$$

Theorem. Suppose that $\mathcal{C} \subset \mathbb{R}^n$ is convex, $f : \mathbb{R}^n \to \mathbb{R}$ is convex over $\mathcal{C}$.

(i) any local minimum of $f(\cdot)$ is also a global minimum

(ii) if $f(\cdot)$ is strictly convex, then there exists at most one global minimum

Optimality: Proof

(i) Suppose x^* is a local minimum, and there exists some $x \neq x^*$ with $f(x) < f(x^*)$. Then, if $\lambda \in [0, 1)$,

$$f(\lambda x^* + (1-\lambda)x) \leq \lambda f(x^*) + (1-\lambda) f(x) < f(x^*)$$

This contradicts local optimality.

(ii) Suppose $x_0 \neq x_1$ are two global minima. Then,

$$f\left(\tfrac{1}{2}(x_0 + x_1)\right) < \frac{f(x_0) + f(x_1)}{2}$$

This contradicts the global optimality of x_0 and x_1.

Necessary & Sufficient Optimality Condition

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in \mathcal{C} \subset \mathbb{R}^n \end{array}$$

Theorem. Suppose that $\mathcal{C} \subset \mathbb{R}^n$ is convex, $f\colon \mathbb{R}^n \to \mathbb{R}$ is convex over $\mathcal{C}$ and differentiable, and $x^* \in \mathcal{C}$ is a feasible point. Then, x^* is a global minimum if and only if

$$\nabla f(x^*)^\top (x - x^*) \geq 0, \quad \forall\, x \in \mathcal{C}$$

Remark. Differentiable functions that satisfy this property are called pseudoconvex.

Proof. Necessity follows from earlier theorem, since $\mathcal{C}$ is convex. For sufficiency, note that

$$f(x) \geq f(x^*) + \nabla f(x^*)^\top (x - x^*) \geq f(x^*), \quad \forall\, x \in \mathcal{C}$$

□

Projection Theorem

Let $\mathcal{C} \subset \mathbb{R}^n$ be a closed and non-empty convex set, and $\|\cdot\| = \|\cdot\|_2$ the Euclidean norm. Fix the vector $x \in \mathbb{R}^n$.

$$\begin{array}{ll} \text{minimize} & \|z - x\| \\ \text{subject to} & z \in \mathcal{C} \subset \mathbb{R}^n \end{array}$$

Theorem. For every $x \in \mathbb{R}^n$, the optimization problem has a unique global minimum x^* called the projection of x onto $\mathcal{C}$. A vector $x' \in \mathcal{C}$ is equal to x^* if and only if

$$(x - x')^\top (z - x') \leq 0, \quad \forall\ z \in \mathcal{C}$$

Projection Theorem: Proof

Existence follows from the fact that $\|z - x\|$ is coercive, $\mathcal{C}$ is closed.

Uniqueness follows since we can equivalently minimize $f(z) \triangleq \|z - x\|^2$, and

$$f(x) = (z - x)^\top (z - x) = z^\top z - 2z^\top x + x^\top x$$

is strictly convex.

Necessary and sufficent conditions follow from the fact that

$$\nabla f(x^*) = 2(x^* - x)$$

Example: Function Approximation

Suppose we are given a function $f : \mathbb{R}^n \to \mathbb{R}$. We wish to approximate $f(\cdot)$ over a set of points $\{x_1, x_2, \ldots, x_m\} \subset \mathbb{R}^n$ with a function

$$g(x) \triangleq \sum_{\ell=1}^{k} r_\ell \phi_\ell(x),$$

where $\{\phi_1(\cdot), \ldots, \phi_k(\cdot)\}$ are a set of basis functions and r is a vector of weights.

Consider the least squares optimization problem

$$\begin{array}{ll} \text{minimize} & \displaystyle\sum_{i=1}^{m} \left(f(x_i) - g(x_i)\right)^2 \\ \text{subject to} & g(\cdot) \text{ is a linear combination of } \{\phi_\ell(\cdot)\} \end{array}$$

Example: Function Approximation

Define the matrix $\Phi \in \mathbb{R}^{m \times k}$ and the vector $y \in \mathbb{R}^m$ by

$$\Phi_{i,\ell} \triangleq \phi_\ell(x_i), \quad y_i \triangleq f(x_i)$$

We have the equivalent projection problem

$$\begin{array}{ll} \text{minimize} & \|y - z\| \\ \text{subject to} & z \in \{\Phi r \ : \ r \in \mathbb{R}^k\} \end{array}$$

This is a projection problem, hence an unique optimizer z^* exists.

B9824 Foundations of Optimization

Lecture 4: Duality I

Fall 2011

Outline

1. Separating/supporting hyperplanes
2. Farkas' Lemma
3. Geometric multipliers
4. Weak duality

Supporting Hyperplane Theorem

Definition. The hyperplane $\{x \in \mathbb{R}^n \ : \ \mu^\top x = b\}$ with normal vector $\mu \in \mathbb{R}^n \setminus \{0\}$ and $b \in \mathbb{R}$ supports the convex set $\mathcal{C} \subset \mathbb{R}^n$ at the point $\bar{x}$ if

$$\mu^\top x \geq \mu^\top \bar{x} = b, \quad \forall \ x \in \mathcal{C}$$

Equivalently,

$$\inf_{x \in \mathcal{C}} \ \mu^\top x \geq \mu^\top \bar{x} = b$$

Theorem. Let $\mathcal{C} \subset \mathbb{R}^n$ be a convex set and $\bar{x} \in \mathbb{R}^n$ be a point that is not in the interior of $\mathcal{C}$. Then, there exists a supporting hyperplane at $\bar{x}$, that is, a vector $\mu \in \mathbb{R}^n$, $\mu \neq 0$, such that

$$\mu^\top x \geq \mu^\top \bar{x}, \quad \forall \ x \in \mathcal{C}$$

Supporting Hyperplane Theorem: Proof

Define the $\bar{\mathcal{C}} = \textbf{cl}\ \mathcal{C}$, and note that $\bar{\mathcal{C}}$ is convex.

Let $\{x_k\}$ be a sequence of points such that $x_k \notin \bar{\mathcal{C}}$, $x_k \neq x$, and $x_k \to \bar{x}$. This sequence exists since $\bar{x}$ is not an interior point of $\mathcal{C}$.

For each x_k, let $\hat{x}_k$ be the projection of x_k onto $\bar{\mathcal{C}}$ (note that $\bar{\mathcal{C}}$ is closed). Then,

$$(\hat{x}_k - x_k)^\top (x - \hat{x}_k) \geq 0, \forall \ x \in \bar{\mathcal{C}}$$

Then, for all k and $x \in \bar{\mathcal{C}}$,

$$\begin{aligned}(\hat{x}_k - x_k)^\top x &\geq (\hat{x}_k - x_k)^\top \hat{x}_k = (\hat{x}_k - x_k)^\top (\hat{x}_k - x_k) + (\hat{x}_k - x_k)^\top x_k \\ &\geq (\hat{x}_k - x_k)^\top x_k\end{aligned}$$

Supporting Hyperplane Theorem: Proof

Set

$$\mu_k \triangleq \frac{\hat{x}_k - x_k}{\|\hat{x}_k - x_k\|},$$

then

$$\mu_k^\top x \geq \mu_k^\top x_k, \quad \forall\ k,\ x \in \bar{\mathcal{C}}$$

Since $\|\mu_k\| = 1$, the sequence $\{\mu_k\}$ has a non-zero subsequential limit μ, and

$$\mu^\top x \geq \mu^\top \bar{x}, \quad \forall\ k,\ x \in \bar{\mathcal{C}}$$

Separating Hyperplane Theorem

Theorem. Let $\mathcal{C}_1,\ \mathcal{C}_2 \subset \mathbb{R}^n$ be two disjoint non-empty convex sets. There exists a hyperplane that separates them, that is a vector $\mu \in \mathbb{R}^n$, $\mu \neq 0$, and a scalar $b \in \mathbb{R}$ with

$$\mu^\top x_1 \leq b \leq \mu^\top x_2, \quad \forall\ x_1 \in \mathcal{C}_1, x_2 \in \mathcal{C}_2$$

Separating Hyperplane Theorem: Proof

Consider the convex set

$$\mathcal{D} \triangleq \mathcal{C}_1 - \mathcal{C}_2 = \{x_1 - x_2 \ : \ x_1 \in \mathcal{C}_1, \ x_2 \in \mathcal{C}_2\}$$

Since the sets are disjoint, $0 \notin \mathcal{D}$, thus there exists a vector $\mu \neq 0$ with

$$0 \leq \mu^\top (x_1 - x_2), \quad \forall \ x_1 \in \mathcal{C}_1, \ x_2 \in \mathcal{C}_2$$

Set

$$b \triangleq \sup_{x_2 \in \mathcal{C}_2} \ \mu^\top x_2$$

Then,

$$\mu^\top x_2 \leq b \leq \mu^\top x_1, \quad \forall \ x_1 \in \mathcal{C}_1, \ x_2 \in \mathcal{C}_2$$

Strictly Separating Hyperplanes

Theorem. Let $\mathcal{C} \subset \mathbb{R}^n$ be a closed convex set and $\bar{x} \notin \mathcal{C}$ a point. Then, there exists a hyperplane that strictly separates $\bar{x}$ and $\mathcal{C}$. In other words, there exists a vector $\mu \in \mathbb{R}^n \setminus \{0\}$ and a scalar $b \in \mathbb{R}$ such that

$$\mu^\top \bar{x} < b < \inf_{x \in \mathcal{C}} \ \mu^\top x$$

Strictly Separating Hyperplanes: Proof

Define

$$r \triangleq \min_{x \in \mathcal{C}} \|x - \bar{x}\|.$$

Since $\mathcal{C}$ is closed and $x \notin \mathcal{C}$, $r > 0$. Then, define

$$\bar{\mathcal{C}} = \{x \in \mathbb{R}^n \ : \ \|x - \bar{x}\| \leq r/2\}$$

Clearly $\mathcal{C}$ and $\bar{\mathcal{C}}$ are disjoint, so we can apply the separating hyperplane theorem.

Halfspace Characterization

Corollary. If $\mathcal{C} \subsetneq \mathbb{R}^n$ is a closed convex set, then $\mathcal{C}$ is the intersection of all the closed halfspaces that contain it.

Proof. Let $\mathcal{H}$ be the collection of all of all closed halfspaces containing $\mathcal{C}$. Since $\mathcal{C} \neq \mathbb{R}^n$, by the strictly separating hyperplane theorem, $\mathcal{H}$ is non-empty. Clearly

$$\mathcal{C} \subset \bar{H} \triangleq \bigcap_{H \in \mathcal{H}} H$$

Suppose there exists $x \in \bar{H}$ with $x \notin \mathcal{C}$. Then, there exists a hyperplane that strictly separates x and $\mathcal{C}$, and x does not lie in the closed halfspace containing $\mathcal{C}$. Thus, $x \notin \bar{H}$. By contradiction, $\mathcal{C} = \bar{H}$. □

Farkas' Lemma

Lemma. (Farkas) Consider a matrix $A \in \mathbb{R}^{m \times n}$. Given a vector $z \in \mathbb{R}^m$, the following conditions are equivalent:

(i) $z^\top y \leq 0$ for all $y \in \mathbb{R}^m$ with $A^\top y \leq 0$

(ii) $z = Ax$ for some $x \in \mathbb{R}^n$ with $x \geq 0$

Farkas' Lemma: Proof

(ii) $\Rightarrow$ (i): $z^\top y = xA^\top y \leq 0$, since $A^\top y \leq 0$ and $x \geq 0$

(i) $\Rightarrow$ (ii): Suppose z satisfies (i), and there is no $x \geq 0$ with $z = Ax$. Define $\mathcal{C} \triangleq \{Ax \in \mathbb{R}^m \ : \ x \geq 0\}$. $\mathcal{C}$ is a closed convex set, and $z \notin \mathcal{C}$. By the strictly separating hyperplane theorem, there exists $y \in \mathbb{R}^m$, $y \neq 0$, with

$$y^\top z > y^\top z', \quad \forall \ z' \in \mathcal{C}$$

Since $0 \in \mathcal{C}$,

$$y^\top z > 0$$

Now, if A_i is a column of A and $\lambda > 0$, $\lambda A_i \in \mathcal{C}$. Thus,

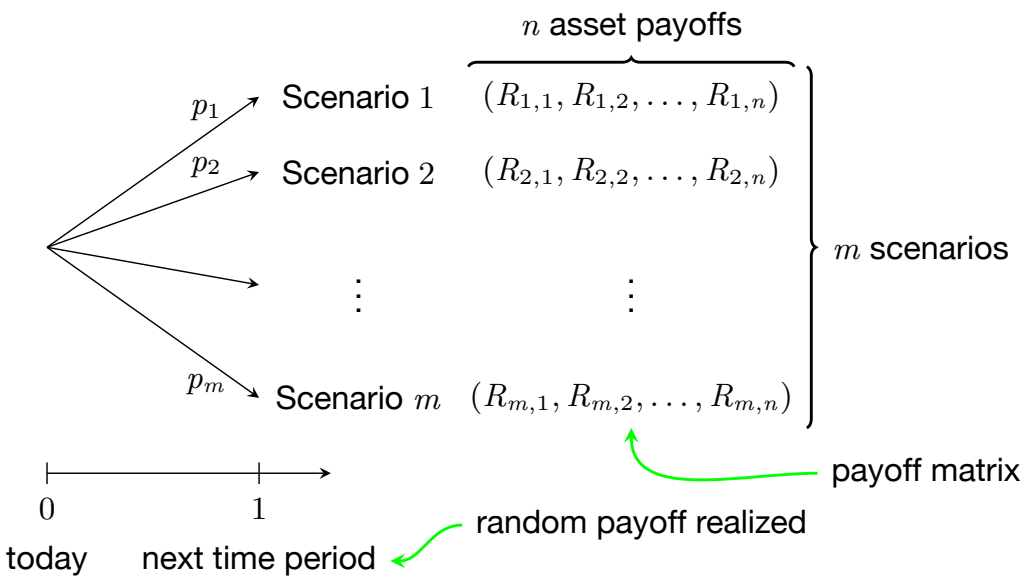
$$y^\top z > \lambda y^\top A_i$$

Dividing by λ and taking $\lambda \to \infty$,

$$0 \geq y^\top A_i$$

Then, $A^\top y \leq 0$. Contradiction.

Application: Arbitrage



$$\text{prices today} = (v_1, v_2, \ldots, v_n)$$

Application: Arbitrage

A portfolio is described by a vector $x \in \mathbb{R}^n$, specifying a quantity x_i of each ith asset.

$$\text{price today} = v^\top x, \qquad \text{future payoffs} = Rx$$

Definition. An arbitrage opportunity is a portfolio x such that

$$v^\top x < 0, \quad Rx \geq 0$$

A consistent market has no arbitrage opportunities.

How can we determine if a market is consistent?

A market is consistent if and only if there exists a vector $q \geq 0$, with $v = R^\top q$.

Application: Risk-Neutral Pricing

Suppose we have $q \geq 0$, $v = R^\top q$, $\mathbf{1}^\top q \neq 0$. Define

$$r \triangleq \frac{1}{\mathbf{1}^\top q} - 1, \quad \pi \triangleq (1+r)q$$

Then, π is a probability distribution ($\pi \geq 0$, $\mathbf{1}^\top \pi = 1$).

Further,

$$v_i = \frac{1}{1+r} \sum_{j=1}^{m} \pi_j R_{ji}$$

Thus, the security prices are the expected discounted value under the distribution π, which is known as a risk-neutral distribution.

The Primal Problem

Consider the primal optimization problem

$$\begin{array}{ll} f\colon \Omega \to \mathbb{R} & \text{minimize} \quad f(x) \\ g\colon \Omega \to \mathbb{R}^r & \text{subject to} \quad g(x) \leq 0 \\ \Omega \subset \mathbb{R}^n & \qquad\quad x \in \Omega \end{array}$$

Define f^* to be the value

$$f^* = \inf_{x \in \Omega,\ g(x) \leq 0} f(x)$$

Assumption. Assume that the feasible set is non-empty and that the optimal cost is bounded below. In other words,

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

Note: We are not making any other assumptions about $f(\cdot)$, $g(\cdot)$, or Ω for the moment! For example, we are not assuming an optimal solution exists.

Geometric Multipliers

$f\colon \Omega \to \mathbb{R}$
$g\colon \Omega \to \mathbb{R}^r$
$\Omega \subset \mathbb{R}^n$

minimize $f(x)$
subject to $g(x) \leq 0$
$x \in \Omega$

Definition. For $x \in \Omega$ and $\mu \in \mathbb{R}^r$, define the Lagrangian function

$$L(x, \mu) \triangleq f(x) + \mu^\top g(x) = f(x) + \sum_{j=1}^{r} \mu_j g_j(x)$$

A vector $\mu^* \in \mathbb{R}^r$ is a geometric multiplier if

(i) $\mu^* \geq 0$

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

Visualization

Definition. The set $\mathcal{S} \subset \mathbb{R}^{r+1}$ of constraint-cost pairs is defined by

$$\mathcal{S} \triangleq \{(g(x), f(x)) \in \mathbb{R}^{r+1} \ : \ x \in \Omega\}$$

Definition. Given a normal $(\mu, \mu_0) \in \mathbb{R}^{r+1} \setminus \{0\}$, define the hyperplane passing through $(\bar{z}, \bar{w}) \in \mathbb{R}^{r+1}$ by

$$\{(z, w) \in \mathbb{R}^{r+1} \ : \ \mu^\top z + \mu_0 w = \mu^\top \bar{z} + \mu_0 \bar{w}\}$$

Define the positive halfspace

$$\{(z, w) \in \mathbb{R}^{r+1} \ : \ \mu^\top z + \mu_0 w \geq \mu^\top \bar{z} + \mu_0 \bar{w}\}$$

and the negative halfspace

$$\{(z, w) \in \mathbb{R}^{r+1} \ : \ \mu^\top z + \mu_0 w \leq \mu^\top \bar{z} + \mu_0 \bar{w}\}$$

The hyperplane is non-vertical if $\mu_0 \neq 0$.

Note: Any non-vertical hyperplane with normal (μ, μ_0) can be normalized so that $\mu_0 = 1$.

Visualization Lemma

Lemma.

(i) The hyperplane with normal $(\mu, 1)$ that passes through the vector $(g(x), f(x))$ intercepts the vertical axis

$$\{(0, w) \in \mathbb{R}^{r+1} \ : \ x \in \mathbb{R}\}$$

at the level $L(x, \mu)$

(ii) Among all hyperplanes with normal $(\mu, 1)$ that contain $\mathcal{S}$ in the positive halfspace, the highest interception of the vertical axis is attained at

$$\inf_{x \in \Omega} L(x, \mu)$$

(iii) μ^* is a geometric multiplier if and only if $\mu^* \geq 0$ and, among all hyperplanes with normal $(\mu^*, 1)$ that contain $\mathcal{S}$ in the positive halfspace, the highest interception of the vertical axis is attained at f^*

Visualization Lemma: Proof

(i): The hyperplane is the set of (z, w) satisfying

$$\mu^\top z + w = \mu^\top g(x) + f(x)$$

If $z = 0$, then we must have $w = L(x, \mu)$.

(ii): The hyperplane with normal $(\mu, 1)$ that intercepts the axis at level c is the set of (z, w) with

$$\mu^\top z + w = c$$

If $\mathcal{S}$ lies in the positive halfspace, then

$$L(x, \mu) = f(x) + \mu^\top g(x) \geq c, \quad \forall \ x \in \Omega$$

Thus, the maximum intercept is $c^* = \inf_{x \in \Omega} L(x, \mu)$.

(iii): Follows from (ii) and the definition of a geometric multiplier.

Geometric Multipliers and Optimality

$f\colon \Omega \to \mathbb{R}$
$g\colon \Omega \to \mathbb{R}^r$
$\Omega \subset \mathbb{R}^n$

minimize $f(x)$
subject to $g(x) \leq 0$
$x \in \Omega$

Theorem. Let μ^* be a geometric multiplier. Then, x^* is a global minimum if and only if x^* is feasible and

$$x^* \in \operatorname*{argmin}_{x\in\Omega} L(x, \mu^*), \qquad \mu_j^* g_j(x^*) = 0, \ \forall\ 1 \leq j \leq r$$

Geometric Multipliers and Optimality: Proof

Assume x^* is a global minimum. Then,

$$f^* = f(x^*) \geq f(x^*) + (\mu^*)^\top g(x^*) = L(x^*, \mu^*) \geq \inf_{x\in\Omega} L(x, \mu^*) = f^*$$

Then, $(\mu^*)^\top g(x^*) = 0$.

Conversely,

$$f(x^*) = f(x^*) + (\mu^*)^\top g(x^*) = L(x^*, \mu^*) = \min_{x\in\Omega} L(x, \mu^*) = f^*$$

The Dual Function

$$\begin{array}{ll} f\colon \Omega \to \mathbb{R} & \text{minimize} \quad f(x) \\ g\colon \Omega \to \mathbb{R}^r & \text{subject to} \quad g(x) \leq 0 \\ \Omega \subset \mathbb{R}^n & \qquad\qquad x \in \Omega \end{array}$$

Definition. The dual function $q\colon \mathbb{R}^r \to \mathbb{R} \cup \{-\infty\}$ is defined by

$$q(\mu) \triangleq \inf_{x \in \Omega} L(x, \mu)$$

Note: $q(\mu) < \infty$ since Ω is non-empty, by assumption. However, $q(\mu)$ may be $-\infty$ for some μ. We define the domain

$$\textbf{dom}\ q = \{\mu \in \mathbb{R}^r \ : \ q(\mu) > -\infty\}$$

The Dual Problem

$$\begin{array}{ll} f\colon \Omega \to \mathbb{R} & \text{minimize} \quad f(x) \\ g\colon \Omega \to \mathbb{R}^r & \text{subject to} \quad g(x) \leq 0 \\ \Omega \subset \mathbb{R}^n & \qquad\qquad x \in \Omega \end{array}$$

Definition. The dual problem is defined by

$$\begin{array}{ll} \text{maximize} & q(\mu) \\ \text{subject to} & \mu \geq 0 \end{array}$$

The dual optimal value is given by

$$q^* \triangleq \sup_{\mu \geq 0} q(\mu)$$

The dual problem corresponds to finding the maximum point of interception of the vertical axis, over all hyperplanes with normal $(\mu, 1)$, where $\mu \geq 0$.

Note: It is possible that $q(\mu) = -\infty$ for all $\mu \geq 0$. In this case, we say that the dual problem is infeasible, and set $q^* = -\infty$.

The Dual Function

$f\colon \Omega \to \mathbb{R}$
$g\colon \Omega \to \mathbb{R}^r$
$\Omega \subset \mathbb{R}^n$

minimize $f(x)$
subject to $g(x) \leq 0$
$x \in \Omega$

$$q(\mu) \triangleq \inf_{x \in \Omega} L(x, \mu)$$

Theorem. The domain **dom** q is convex, and $q(\cdot)$ is concave over its domain.

Proof. Follows since $q(\cdot)$ is a pointwise minimum of concave (in fact, linear) functions. □

Weak Duality Theorem

Theorem. (Weak Duality) $q^* \leq f^*$

Proof. If $\mu \geq 0$, $x \in \Omega$, and $g(x) \leq 0$,

$$q(\mu) = \inf_{z \in \Omega} L(z, \mu) \leq f(x) + \mu^\top g(x) \leq f(x)$$

Thus,

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

□

Duality Gap

Definition. If $q^* = f^*$ we say there is no duality gap. If $q^* < f^*$, there is a duality gap.

Theorem. If there is no duality gap, the set of geometric multipliers is equal to the set of dual optimal solutions.

If there is a duality gap, the set of geometric multipliers is empty.

Proof. By definition, $\mu^* \geq 0$ is a geometric multiplier iff $f^* = q(\mu^*) \leq q^*$. By the weak duality theorem, this holds iff there is no duality gap. □

Examples

$$\begin{array}{ll} \text{minimize} & f(x) = x_1 - x_2 \\ \text{subject to} & g(x) = x_1 + x_2 - 1 \leq 0 \\ & x \in \Omega = \{(x_1, x_2) \ : \ x_1, x_2 \geq 0\} \end{array}$$

$$f^* = -1, \qquad x^* = (0, 1)$$

$$q(\mu) = \min_{x \geq 0} \ x_1 - x_2 + \mu(x_1 + x_2 - 1) = \begin{cases} -\mu & \text{if } \mu \geq 1 \\ -\infty & \text{if } \mu < 1 \end{cases}$$

$$q^* = -1, \qquad \mu^* = 1$$

Examples

$$\begin{array}{ll} \text{minimize} & f(x) = |x_1| + x_2 \\ \text{subject to} & g(x) = x_1 \leq 0 \\ & x \in \Omega = \{(x_1, x_2) \ : \ x_2 \geq 0\} \end{array}$$

$$f^* = 0, \qquad x^* = (0, 0)$$

$$q(\mu) = \min_{(x_1, x_2) \ : \ x_2 \geq 0} |x_1| + x_2 + \mu x_1 = \begin{cases} 0 & \text{if } |\mu| \leq 1 \\ -\infty & \text{if } |\mu| > 1 \end{cases}$$

$$q^* = 0, \qquad \mu^* \in [0, 1]$$

Examples

$$\begin{array}{ll} \text{minimize} & f(x) = x \\ \text{subject to} & g(x) = x^2 \leq 0 \\ & x \in \Omega = \mathbb{R} \end{array}$$

$$f^* = 0, \qquad x^* = 0$$

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

$q^* = 0$, no duality gap, but no optimal dual solution

Examples

$$\begin{array}{ll} \text{minimize} & f(x) = -x \\ \text{subject to} & g(x) = x - 1/2 \leq 0 \\ & x \in \Omega = \{0, 1\} \end{array}$$

$$f^* = 0, \qquad x^* = 0$$

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

$$q^* = -1/2, \qquad \mu^* = -1$$

Primal and Dual Optimality Conditions

Theorem. (x^*, μ^*) is an optimal solution/geometric multiplier pair if and only if

(i) $x^* \in \Omega$, $g(x^*) \leq 0$ (primal feasibility)

(ii) $\mu^* \geq 0$ (dual feasibility)

(iii) $x^* \in \operatorname{argmin}_{x \in \Omega} L(x, \mu^*)$ (Lagrangian optimality)

(iv) $\mu_j^* g_j^*(x^*) = 0, \forall\ 1 \leq j \leq r$ (complementary slackness)

Note: This is only useful if there is no duality gap!

Primal and Dual Optimality Conditions: Proof

If (x^*, μ^*) is an optimal solution/geometric multiplier pair, then clearly (i) and (ii) hold. (iii) and (iv) follow from the earlier theorem.

Conversely, if (i)–(iv) hold,

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

By weak duality, equality must hold and hence x^* is primal optimal and μ^* is dual optimal.

Saddle Point Theorem

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

$$L(x^*, \mu) \leq L(x^*, \mu^*) \leq L(x, \mu^*), \quad \forall\ x \in \Omega,\ \mu \geq 0$$

Note: This is only useful if there is no duality gap!

Saddle Point Theorem: Proof

If (x^*, μ^*) is an optimal solution/geometric multiplier pair, from optimality condition (iii),

$$L(x^*, \mu^*) \leq L(x, \mu^*), \quad \forall\, x \in \Omega$$

Further, if $\mu \geq 0$, using optimality conditions (iii) and (iv),

$$L(x^*, \mu) \leq f(x^*) = L(x^*, \mu^*)$$

Conversely, assume the $x^* \in \Omega$, $\mu^* \geq 0$, and (x^*, μ^*) is a saddle point. Then,

$$\sup_{\mu \geq 0} L(x^*, \mu) = \sup_{\mu \geq 0} f(x^*) + \mu^\top g(x^*) = \begin{cases} f(x^*) & \text{if } g(x^*) \leq 0 \\ +\infty & \text{otherwise} \end{cases}$$

Thus, $g(x^*) \leq 0$, $L(x^*, \mu^*) = f(x^*)$, and $\mu_j^* g_j(x^*) = 0$, $\forall\, j$. Thus, optimality conditions (i)–(iv) hold.

Infeasibility and Unboundedness

Suppose the primal problem is unbounded, that is, $f^* = -\infty$. Then, the proof of the weak duality theorem applies, and

$$q(\mu) = -\infty, \quad \forall\, \mu \geq 0$$

Thus, the dual problem is infeasible. Similarly, if the dual problem is feasible, the primal problem is bounded.

Alternatively, assume that the primal problem is infeasible. In general, nothing can be said about the dual problem.

Equality Constraints

The theory developed thus far can be extended to equality constaints by introducing a pair of inequality constraints for each equality constraint. Equivalently, we can eliminate non-negativity constraints on the multipliers for equality constraints.

$$\begin{array}{ll} f\colon \Omega \to \mathbb{R} & \text{minimize} \quad f(x) \\ h\colon \Omega \to \mathbb{R}^m & \text{subject to} \quad h(x) = 0 \\ g\colon \Omega \to \mathbb{R}^r & \qquad\qquad\ \ g(x) \leq 0 \\ \Omega \subset \mathbb{R}^n & \qquad\qquad\ \ x \in \Omega \end{array}$$

Definition. For $x \in \Omega$, $\lambda \in \mathbb{R}^m$, $\mu \in \mathbb{R}^r$, define the Lagrangian function

$$L(x, \lambda, \mu) \triangleq f(x) + \lambda^\top h(x) + \mu^\top g(x) = f(x) + \sum_{i=1}^{m} \lambda_i h_i(x) + \sum_{j=1}^{r} \mu_j g_j(x)$$

Equality Constraints

$$\begin{array}{ll} f\colon \Omega \to \mathbb{R} & \text{minimize} \quad f(x) \\ h\colon \Omega \to \mathbb{R}^m & \text{subject to} \quad h(x) = 0 \\ g\colon \Omega \to \mathbb{R}^r & \qquad\qquad\ \ g(x) \leq 0 \\ \Omega \subset \mathbb{R}^n & \qquad\qquad\ \ x \in \Omega \end{array}$$

Definition. $(\lambda^*, \mu^*) \in \mathbb{R}^m \times \mathbb{R}^r$ is a geometric multiplier if

(i) $\mu^* \geq 0$

(ii) $f^* = \inf_{x \in \Omega} L(x, \lambda^*, \mu^*)$

Equality Constraints

$$f\colon \Omega \to \mathbb{R}$$
$$h\colon \Omega \to \mathbb{R}^m$$
$$g\colon \Omega \to \mathbb{R}^r$$
$$\Omega \subset \mathbb{R}^n$$

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & h(x) = 0 \\ & g(x) \leq 0 \\ & x \in \Omega \end{array}$$

Definition. The dual function $q\colon \mathbb{R}^m \times \mathbb{R}^r \to \mathbb{R} \cup \{-\infty\}$ is defined by

$$q(\lambda, \mu) \triangleq \inf_{x \in \Omega} L(x, \lambda, \mu)$$

The dual problem is defined by

$$\begin{array}{ll} \text{maximize} & q(\lambda, \mu) \\ \text{subject to} & \mu \geq 0 \\ & \lambda \in \mathbb{R}^m \end{array}$$

Issues

This theory is most useful when there is no duality gap, or, when geometric multipliers exist.

We will develop necessary conditions for this that will require:

- Convexity of the objective and constraints
- Technical conditions, similar to constraint qualification

B9824 Foundations of Optimization

Lecture 5: Duality II

Fall 2011

Outline

1. Duality and decentralization
2. Duality and combinatorial optimization
3. Strong duality
4. Duality for linear programs, quadratic programs

Duality and Decentralization

Suppose that the decision variables x decompose according to

$$x = (x_1, x_2, \ldots, x_k) \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \times \cdots \times \mathbb{R}^{n_k}$$

where $n = n_1 + \cdots + n_k$.

Consider the separable optimization problem

$$\begin{array}{lll} \text{minimize} & \displaystyle\sum_{i=1}^{k} f_i(x_i) & \\ \text{subject to} & \displaystyle\sum_{i=1}^{k} g_{ij}(x_i) \leq 0, & \forall\ 1 \leq j \leq r \\ & x_i \in \Omega_i, & \forall\ 1 \leq i \leq k \end{array}$$

Here,

$$f_i : \ \Omega_i \to \mathbb{R}, \quad g_{ij} : \ \Omega_i \to \mathbb{R}, \quad \Omega_i \subset \mathbb{R}^{n_i}$$

Separablility and Duality

The Lagrangian is

$$L(x, \mu) = \sum_{i=1}^{k} f_i(x_i) + \sum_{j=1}^{r} \mu_j \sum_{i=1}^{k} g_{ij}(x_i)$$

The dual function is

$$q(\mu) = \inf_{x \in \Omega_1 \times \cdots \times \Omega_k} \ \sum_{i=1}^{k} \left(f_i(x_i) + \sum_{j=1}^{r} \mu_j g_{ij}(x_i) \right)$$

Note that if

$$q_i(\mu) \triangleq \inf_{x_i \in \Omega_i} \ f_i(x_i) + \sum_{j=1}^{r} \mu_j g_{ij}(x_i)$$

then the dual problem becomes

$$\begin{array}{ll} \text{maximize} & \displaystyle\sum_{i=1}^{k} q_i(\mu) \\ \text{subject to} & \mu \geq 0 \\ & \mu \in \mathbb{R}^r \end{array}$$

Example: Resource Allocation

- Activities $1, \ldots, k$ (e.g., divisions of a firm)
- Resources $1, \ldots, r$ (e.g., capital, labor, etc.)
- Each activity consumes resources, and generates a benefit (utility, profit, etc.)
- Decision variables:

  $x_{ij} =$ quantity of resource j allocated to activity i

  $x_{ij} \geq 0$
- The ith activity generates utility according to

$$U_i(x_i) \triangleq U_i(x_{i1}, \ldots, x_{ir})$$

- The supply of the resources is limited, so we require that

$$\sum_{i=1}^{k} x_{ij} \leq C_j, \quad \forall\, 1 \leq j \leq r \qquad [C_j > 0]$$

Example: Resource Allocation

- Objective: maximize total utility

$$\begin{aligned} \text{maximize} \quad & \sum_{i=1}^{k} U_i(x_i) \\ \text{subject to} \quad & \sum_{i=1}^{k} x_{ij} \leq C_j, \quad \forall\, 1 \leq j \leq r \\ & x \geq 0, \\ & x \in \mathbb{R}^{k \times r} \end{aligned}$$

Example: Resource Allocation

Lagrangian:

$$L(x,\mu) = \sum_{i=1}^{k} U_i(x_i) - \sum_{j=1}^{r} \mu_j \left(\sum_{i=1}^{k} x_{ij} - C_j \right)$$

Dual function:

$$q(\mu) = \sup_{x \geq 0} \; L(x,\mu) = \sum_{i=1}^{k} q_i(\mu) + \sum_{j=1}^{r} \mu_j C_j$$

where

$$q_i(\mu) = \sup_{x_i \geq 0} U_i(x_i) - \sum_{j=1}^{r} \mu_j x_{ij}$$

Example: Resource Allocation

If μ^* is a geometric multiplier, then a feasible allocation x^* is a global maximum if and only if

$$x^* \in \operatorname*{argmax}_{x \geq 0} \; L(x,\mu^*) = \operatorname*{argmax}_{x \geq 0} \; \sum_{i=1}^{k} \left(U_i(x_i) - \sum_{j=1}^{r} \mu_j^* x_{ij} \right)$$

$$\Leftrightarrow \quad x_i^* \in \operatorname*{argmax}_{x_i \geq 0} \; U_i(x_i) - \sum_{j=1}^{r} \mu_j^* x_{ij}$$

The dual variables μ^* can be interpreted as prices that serve as a coordination mechanism. Given the proper selection of prices, the optimal solution can be constructed by solving independent subproblems for each activity.

Prices are proxies for decentralization.

Example: Resource Allocation

Conditions for optimal solution-geometric multiplier pair (x^*, μ^*):

(i) Primal feasibility

$$x^* \geq 0, \quad \sum_{i=1}^{k} x_{ij}^* \leq C_j, \ \forall\, 1 \leq j \leq r$$

(ii) Dual feasibility $\quad \mu^* \geq 0$

(iii) Lagrangian optimality

$$x_i^* \in \underset{x_i \geq 0}{\text{argmax}} \ U_i(x_i) - \sum_{j=1}^{r} \mu_j^* x_{ij}, \forall\, 1 \leq i \leq k$$

(iv) Complementary slackness

$$\mu_j^* \left(\sum_{i=1}^{k} x_{ij}^* - C_j \right) = 0, \ \forall\, 1 \leq j \leq r$$

Tâtonnement Procedure

1. Pick a set of initial prices $\mu \geq 0$.
2. Compute the allocation x by

$$x_i \in \underset{x_i' \geq 0}{\text{argmax}} \ U_i(x_i') - \sum_{j=1}^{r} \mu_j x_{ij}'$$

3. For each resource j,

$$\text{if } \sum_{i=1}^{k} x_{ij} > C_j \quad \Rightarrow \quad \text{raise } \mu_j \text{ slightly}$$

$$\text{if } \sum_{i=1}^{k} x_{ij} < C_j \quad \Rightarrow \quad \text{lower the } \mu_j \text{ slightly, keeping } \mu_j \geq 0$$

4. Repeat.

Under proper technical conditions (concavity of utility functions, existence of geometric multipliers, good choice of step-sizes, etc.), this decentralized procedure with find the global minimum.

The knapsack problem consists of determining which of n items to place in a knapsack of limited capacity.

- Decision variables: $(x_1, \ldots, x_n)$ where $x_i = 1$ if the item i is placed in the knapsack, $x_i = 0$ otherwise
- Knapsack constraint: the total capacity of the knapsack is $C > 0$, item i has weight $0 < w_i \leq C$, so
$$\sum_{i=1}^{n} w_i x_i \leq C$$
- Objective: the value of item i is $v_i > 0$, we would like to maximize the total value of items in the knapsack
$$\sum_{i=1}^{n} v_i x_i$$

Example: The Knapsack Problem

$$\begin{aligned} f^* = \quad & \text{maximize} \quad f(x) = \sum_{i=1}^{n} v_i x_i \\ & \text{subject to} \quad \sum_{i=1}^{n} w_i x_i \leq C \\ & \qquad\qquad x_i \in \{0,1\}, \qquad \forall\, 1 \leq i \leq n \end{aligned}$$

Integer problem, "hard"

Dual function:

$$\begin{aligned} q(\mu) &= \sup_{x \in \{0,1\}^n} L(x,\mu) = \sup_{x \in \{0,1\}^n} \sum_{i=1}^{n} v_i x_i - \mu \left(\sum_{i=1}^{n} w_i x_i - C \right) \\ &= \sup_{x \in \{0,1\}^n} \sum_{i=1}^{n} (v_i - \mu w_i) x_i + \mu C \\ &= \sum_{i=1}^{n} \max\left(v_i - \mu w_i, 0\right) + \mu C \end{aligned}$$

Example: The Knapsack Problem

Weak duality:

$$f^* \leq q^* = \inf_{\mu \geq 0} \; q(\mu) = \inf_{\mu \geq 0} \; \sum_{i=1}^{n} w_i \max\left(v_i/w_i - \mu, 0\right) + \mu C$$

Without loss of generality, assume

$$\frac{v_1}{w_1} \geq \frac{v_2}{w_2} \geq \cdots \geq \frac{v_n}{w_n}$$

Define

$$I^*(\mu) \triangleq \max\left\{i : \; v_i/w_i \geq \mu\right\}$$

Then,

$$q(\mu) = \sum_{i=1}^{I^*(\mu)} \left(v_i - w_i\mu\right) + \mu C = \sum_{i=1}^{I^*(\mu)} v_i + \mu \left(C - \sum_{i=1}^{I^*(\mu)} w_i\right)$$

$$\Rightarrow \text{piecewise linear}$$

Example: The Knapsack Problem

Define

$$I^* \triangleq \min\left\{I : \; \sum_{i=1}^{I} w_i > C\right\} \quad \Rightarrow \quad \mu^* = v_{I^*}/w_{I^*}$$

Then,

$$\begin{aligned} f^* \quad &\leq q^* = \inf_{\mu \geq 0} \; q(\mu) = \sum_{i=1}^{I^*} v_i + \frac{v_{I^*}}{w_{I^*}} \left(C - \sum_{i=1}^{I^*} w_i\right) \\ &= \sum_{i=1}^{I^*-1} v_i + \frac{v_{I^*}}{w_{I^*}} \left(C - \sum_{i=1}^{I^*-1} w_i\right) \end{aligned}$$

We have an upper bound on the optimal value. How about a “good” solution? A bound on the duality gap?

Example: The Knapsack Problem

Consider the optimization

$$\max_{x\in\{0,1\}^n} L(x,\mu^*) = \max_{x\in\{0,1\}^n} \sum_{i=1}^{n} \left(v_i - \frac{v_{I^*}}{w_{I^*}} w_i\right) x_i + \frac{v_{I^*}}{w_{I^*}} C$$

$$\Rightarrow \quad \tilde{x}_i = \begin{cases} 1 & \text{if } i < I^* \\ 0 & \text{if } i \geq I^* \end{cases}$$

Note that the trial solution $\tilde{x}$ is clearly feasible. It is constructed with a greedy algorithm: sort the items by "bang per buck" v_i/w_i, greedily add items to knapsack until it is full.

$$f(\tilde{x}) = \sum_{i=1}^{I^*-1} v_i \qquad q^* - f(\tilde{x}) = \frac{v_{I^*}}{w_{I^*}} \left(C - \sum_{i=1}^{I^*-1} w_i\right) \leq v_{I^*}$$

Example: The Knapsack Problem

Putting it all together,

$$f(\tilde{x}) \leq f^* \leq f(\tilde{x}) + v_{I^*}$$

Can we do better?

Set $\hat{x}$ to be the the better of

1. the trial solution $\tilde{x}$
2. a solution consisting of only item I^*

$$f(\hat{x}) \leq f^* \leq q^* \leq f(\tilde{x}) + v_{I^*} \leq 2f(\hat{x})$$

The revised algorithm is a 2-approximation to the knapsack problem.

Strong Duality

We would like to understand two related questions:

- When is there no duality gap?
- When do optimal solutions exist for the dual problem?

Strong Duality: Linear Constraints

Primal problem:

$$f : \mathbb{R}^n \to \mathbb{R}$$
$$A \in \mathbb{R}^{m \times n}, \ b \in \mathbb{R}^m$$

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & Ax \leq b, \\ & x \in \mathbb{R}^n \end{array}$$

Dual problem:

$$L(x, \mu) = f(x) + \mu^\top (Ax - b)$$
$$q(\mu) = \inf_{x \in \mathbb{R}^n} L(x, \mu)$$

$$\begin{array}{ll} \text{maximize} & q(\mu) \\ \text{subject to} & \mu \geq 0, \\ & \mu \in \mathbb{R}^m \end{array}$$

Theorem. Suppose that $f(\cdot)$ is convex over $\mathbb{R}^n$ and continuously differentiable. If the primal problem has an optimal solution, then there is no duality gap, and at least one geometric multiplier exists.

Linear Constraints: Proof

Let x^* be an optimal solution for the primal. Then, there exists $\mu^* \in \mathbb{R}^m$ such that

$$\mu^* \geq 0, \quad (\mu^*)^\top (Ax^* - b) = 0, \quad \nabla f(x^*) + A\mu^* = 0$$

Since $L(x, \mu)$ is convex in x, we have

$$x^* \in \operatorname*{argmin}_{x \in \mathbb{R}^n} L(x, \mu^*)$$

Thus,

$$f(x^*) = \min_{x \in \mathbb{R}^n} \; f(x) + (\mu^*)^\top (Ax - b) = q(\mu^*)$$

Then,

$$q(\mu^*) \leq q^* \leq f^* \leq f(x^*) \quad \Rightarrow \quad q^* = f^*$$

Linear Constraints: Remarks

Trivally applies to linear equality constraints also. More generally,

$$\begin{array}{l} f : \mathbb{R}^n \to \mathbb{R} \\ A \in \mathbb{R}^{m \times n}, \; b \in \mathbb{R}^m \\ g : \mathbb{R}^n \to \mathbb{R}^r \end{array} \qquad \begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & Ax = b, \\ & g(x) \leq 0, \\ & x \in \mathbb{R}^n \end{array}$$

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

If:

(a) There exists an optimal solution x^*

(b) $f(\cdot)$, $g(\cdot)$ are continuously differentiable

(c) There exist multipliers (λ^*, μ^*) satisfying the KKT conditions (e.g., some type of regularity)

(d) $f(\cdot)$, $g(\cdot)$ are convex over $\mathbb{R}^n$

Then, there is no duality gap, and geometric multipliers exist.

Strong Duality: Remarks

We would like a more general theory, which

- Does not require differentiablility
- Allows for set constraints

Slater's Condition

$$f: \Omega \to \mathbb{R}$$
$$g: \Omega \to \mathbb{R}^r$$
$$\Omega \subset \mathbb{R}^n$$

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & g(x) \leq 0 \\ & x \in \Omega \end{array}$$

Theorem. (Slater's Condition) Suppose that:

(i) The problem is bounded, i.e.

$$-\infty < f^* = \inf_{x \in \Omega,\ g(x) \leq 0} f(x)$$

(ii) The set Ω is convex, and $f(\cdot)$, $g(\cdot)$ are convex over Ω

(iii) There exists a vector $\bar{x} \in \Omega$ with $g(\bar{x}) < 0$

Then, there is no duality gap and there exists at least one geometric multiplier.

Slater's Condition: Proof

Define

$$\mathcal{A} = \{(z, w) \subset \mathbb{R}^{r+1} \ : \ \exists \ x \in \Omega \text{ with } g(x) \leq z, \ f(x) \leq w\}$$

Observe that $\mathcal{A}$ is convex, by the convexity of Ω, $f(\cdot)$, and $g(\cdot)$.

Next, observe that $(0, f^*)$ is not in the interior of $\mathcal{A}$. Otherwise, for some $\epsilon > 0$, $(0, f^* - \epsilon) \in \mathcal{A}$, contradicting the definition of f^*.

By the supporting hyperplane theorem, there exists a normal vector $(\mu, \beta) \neq (0, 0)$ such that

$$\beta f^* \leq \beta w + \mu^\top z, \quad \forall \ (z, w) \in \mathcal{A}$$

If $(z, w) \in \mathcal{A}$, then $(z, w + \gamma) \in \mathcal{A}$ for all $\gamma \geq 0$. Then, we must have $\beta \geq 0$. Similarly, $\mu \geq 0$.

Slater's Condition: Proof

We would like to show that $\beta > 0$. Suppose not. Then,

$$0 \leq \mu^\top z, \quad \forall \ (z, w) \in \mathcal{A}$$

Since $(g(\bar{x}), f(\bar{x})) \in \mathcal{A}$,

$$0 \leq \mu^\top g(\bar{x})$$

Then, we must have $\mu = 0$ and $(\mu, \beta) = (0, 0)$ which is a contradiction.

Since $\beta > 0$, we can divide by β and assume that $\beta = 1$. Thus,

$$f^* \leq w + \mu^\top z, \quad \forall \ (z, w) \in \mathcal{A}$$

$$\Rightarrow \quad f^* \leq f(x) + \mu^\top g(x), \quad \forall \ x \in \Omega$$

Minimizing over $x \in \Omega$,

$$f^* \leq \inf_{x \in \Omega} \ f(x) + \mu^\top g(x) = q(\mu) \leq q^*$$

Then, by weak duality, $f^* = q^*$ and μ is a geometric multiplier.

Slater's Condition: Remarks

Existence of the interior point in condition (iii) is important!

Does not apply to equality constraints! These are harder to deal with.

Definition. Suppose $\mathcal{C} \subset \mathbb{R}^n$ is a convex set. The relative interior of $\mathcal{C}$ is the set **relint** $\mathcal{C}$ of all $x \in \mathbb{R}^n$ for which there exists an $\epsilon > 0$ such that if $z \in$ **aff** $\mathcal{C}$ with $\|z - x\| < \epsilon$, then $z \in \mathcal{C}$.

In other words, **relint** $\mathcal{C}$ is the interior of $\mathcal{C}$ relative to the affine hull **aff** $\mathcal{C}$.

Mixed Constraints

$$
\begin{array}{ll}
f: \Omega \to \mathbb{R} & \text{minimize} \quad f(x) \\
g: \Omega \to \mathbb{R}^r & \text{subject to} \quad g(x) \leq 0 \\
A \in \mathbb{R}^{m \times n}, \ b \in \mathbb{R}^m & \qquad\qquad Ax \leq b \\
E \in \mathbb{R}^{m \times k}, \ d \in \mathbb{R}^k & \qquad\qquad Ex = d \\
\Omega \subset \mathbb{R}^n & \qquad\qquad x \in \Omega
\end{array}
$$

Theorem. Suppose that the optimal value f^* is finite, and:

(i) Ω is the intersection of a convex set $\mathcal{C}$ and a polyhedron

(ii) The functions $f(\cdot)$, $g(\cdot)$ are convex over Ω

(iii) There is a feasible vector $\bar{x}$ with $g(\bar{x}) < 0$

(iv) There is a vector x with $Ax \leq b$, $Ex = d$, $x \in$ **relint** $\mathcal{C}$, and $x \in \Omega$

Then, there is no duality gap and there exists at least one geometric multiplier.

Duality for Linear Programs

$c \in \mathbb{R}^n$
$A \in \mathbb{R}^{r \times n}$
$b \in \mathbb{R}^r$

minimize	$c^\top x$
subject to	$Ax \geq b$
	$x \in \Omega = \mathbb{R}^n$

The Lagrangian is

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

The dual objective function is

$$q(\mu) = \inf_x \; c^\top x + \mu^\top (b - Ax) = \begin{cases} b^\top \mu & \text{if } c = A^\top \mu \\ -\infty & \text{otherwise} \end{cases}$$

The dual problem is

maximize	$q(\mu)$		maximize	$b^\top \mu$
subject to	$\mu \geq 0$	$\Rightarrow$	subject to	$A^\top \mu = c$
	$\mu \in \mathbb{R}^r$			$\mu \geq 0$
				$\mu \in \mathbb{R}^r$

Duality for Linear Programs

More generally, let A be a matrix with rows $a_i^\top \in \mathbb{R}^n$ and columns $A_j \in \mathbb{R}^r$. Then:

	Primal			**Dual**	
minimize	$c^\top x$		maximize	$b^\top y$	
subject to	$a_i^\top x \geq b_i,$	$\forall\, i \in M_1$	subject to	$y_j \geq 0,$	$\forall\, j \in M_1$
	$a_i^\top x \leq b_i,$	$\forall\, i \in M_2$		$y_j \leq 0,$	$\forall\, j \in M_2$
	$a_i^\top x = b_i,$	$\forall\, i \in M_3$		y_j free,	$\forall\, j \in M_3$
	$x_j \geq 0,$	$\forall\, j \in N_1$		$A_j^\top y \leq c_j,$	$\forall\, j \in N_1$
	$x_j \leq 0,$	$\forall\, j \in N_2$		$A_j^\top y \geq c_j,$	$\forall\, j \in N_2$
	x_j free,	$\forall\, j \in N_3$		$A_j^\top y = c_j,$	$\forall\, j \in N_3$

Note: The dual is being taken with respect to the set constraint

$$\Omega \triangleq \{x \in \mathbb{R}^n \;:\; x_j \geq 0,\ \forall\, j \in N_1,\ x_j \leq 0,\ \forall\, j \in N_2\}$$

Duality for Linear Programs

Primal	minimize	maximize	**Dual**
constraints	$\geq b_i$ $\leq b_i$ $= b_i$	≥ 0 ≤ 0 free	variables
variables	≥ 0 ≤ 0 free	$\leq c_j$ $\geq c_j$ $= c_j$	constraints

Duality for Quadratic Programs

$Q \in \mathbb{R}^{n \times n}$, $Q \succ 0$
$c \in \mathbb{R}^n$
$A \in \mathbb{R}^{r \times n}$
$b \in \mathbb{R}^r$

$$\begin{array}{ll} \text{minimize} & \frac{1}{2} x^\top Q x + c^\top x \\ \text{subject to} & Ax \leq b \\ & x \in \Omega = \mathbb{R}^n \end{array}$$

The Lagrangian is

$$L(x, \mu) = \tfrac{1}{2} x^\top Q x + c^\top x + \mu^\top (Ax - b)$$

The dual objective function is

$$q(\mu) = \inf_x \ \tfrac{1}{2} x^\top Q x + c^\top x + \mu^\top (Ax - b)$$

This is minimized when

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

Thus,

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

Duality for Quadratic Programs

$Q \in \mathbb{R}^{n \times n}$, $Q \succ 0$
$c \in \mathbb{R}^n$
$A \in \mathbb{R}^{r \times n}$
$b \in \mathbb{R}^r$

$$\begin{array}{ll} \text{minimize} & \frac{1}{2} x^\top Q x + c^\top x \\ \text{subject to} & Ax \leq b \\ & x \in \Omega = \mathbb{R}^n \end{array}$$

The dual problem is

$$\begin{array}{ll} \text{maximize} & q(\mu) \\ \text{subject to} & \mu \geq 0 \\ & \mu \in \mathbb{R}^r \end{array} \quad \Rightarrow \quad \begin{array}{ll} \text{maximize} & -\frac{1}{2} \mu^\top P \mu - t^\top \mu - d \\ \text{subject to} & \mu \geq 0 \\ & \mu \in \mathbb{R}^r \end{array}$$

Here,

$$P \triangleq A Q^{-1} A^\top, \quad t \triangleq b + A Q^{-1} c, \quad d \triangleq \frac{1}{2} c^\top Q^{-1} c$$

Note: The dual has simpler contraints and possibly smaller dimension. However, dual may be dense if primal is sparse.

B9824 Foundations of Optimization

Lecture 6: Duality III

Fall 2011

Outline

1. Conjugate functions
2. Application: Chernoff bounds & large deviations
3. Conjugacy & duality
4. Sensitivity
5. Subgradients

Conjugate Functions

Definition. A proper function is an extended-real valued function $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$, with **dom** $f \neq \emptyset$.

Definition. A closed function is an extended-real valued function $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$, such that for every $\alpha \in \mathbb{R}$, the sublevel set

$$\{x \in \mathbb{R}^n \ : \ f(x) \leq \alpha\}$$

is closed.

Definition. The convex conjugate (Fenchel-Legendre transformation) of a proper function $f(\cdot)$ is the extended-real valued function $f^* : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ defined by

$$f^*(y) \triangleq \sup_{x \in \textbf{dom } f} x^\top y - f(x)$$

Conjugate Functions: Examples

Example.

$$f(x) = a^\top x + b \quad \Rightarrow \quad f^*(y) = \begin{cases} -b & \text{if } y = a \\ \infty & \text{otherwise} \end{cases}$$

Example.

$$f(x) = \tfrac{1}{2} x^\top Q x, \ Q \succ 0 \quad \Rightarrow \quad f^*(y) = \tfrac{1}{2} y^\top Q^{-1} y$$

Example.

$$f(x) = \log \left(\sum_{i=1}^n e^{x_i} \right) \quad \Rightarrow \quad f^*(y) = \begin{cases} \sum_{i=1}^n y_i \log y_i & \text{if } y \geq 0, \ \mathbf{1}^\top y = 1 \\ \infty & \text{otherwise} \end{cases}$$

Conjugate Functions: Examples

Example. Given a set $\mathcal{C} \subset \mathbb{R}^n$, the indicator function $I_{\mathcal{C}} : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is defined by

$$I_{\mathcal{C}}(x) = \begin{cases} 0 & \text{if } x \in \mathcal{C} \\ \infty & \text{otherwise} \end{cases}$$

The conjugate of the indicator function is called the support function

$$S_{\mathcal{C}}(x) \triangleq \sup_{y \in \mathcal{C}} \; x^\top y$$

Dual Norms

Definition. Given a norm $\|\cdot\|$ on $\mathbb{R}^n$, define the dual norm $\|\cdot\|_*$ on $\mathbb{R}^n$ by

$$\|y\|_* \triangleq \sup_{\|x\| \le 1} \; x^\top y$$

Note that the dual norm is the support function of the unit ball of the norm.

Example. $\|\cdot\| = \|\cdot\|_1 \quad \Rightarrow \quad \|\cdot\|_* = \|\cdot\|_\infty$

Example. $\|\cdot\| = \|\cdot\|_2 \quad \Rightarrow \quad \|\cdot\|_* = \|\cdot\|_2$

Example. $\|\cdot\| = \|\cdot\|_p, \quad p \in (1, \infty)$

$\Rightarrow \quad \|\cdot\|_* = \|\cdot\|_q$ where $p^{-1} + q^{-1} = 1$

Conjugate Functions: Examples

Example. Given a norm $\|\cdot\|$ on $\mathbb{R}^n$, suppose $f(x) \triangleq \|x\|$. Then,

$$f^*(y) = \begin{cases} 0 & \text{if } \|y\|_* \leq 1 \\ \infty & \text{otherwise} \end{cases}$$

In other words, the conjugate of a norm is the indicator function of the unit ball of the dual norm.

Elementary Properties of Conjugate Functions

Theorem. $f^*(\cdot)$ is convex.
Proof. Follows since it is a pointwise supremum of convex (linear) functions in y. This is true even if $f(\cdot)$ is not convex. □

Theorem. (Fenchel's Inequality) For all x, $y \in \mathbb{R}^n$,

$$x^\top y \leq f(x) + f^*(y)$$

Proof. Follows from the definition, since

$$f^*(y) \geq x^\top y - f(x)$$

□

Theorem. If $f(u, v) = f_1(u) + f_2(v)$, then

$$f^*(w, z) = f_1^*(w) + f_2^*(z)$$

Theorem. Suppose that $A \in \mathbb{R}^{n\times n}$ is invertible and $b \in \mathbb{R}^n$, and $g(x) \triangleq f(Ax + b)$. Then,

$$g^*(y) = f^*(A^{-\top} y) - b^\top A^{-\top} y$$

Conjugate of the Conjugate

Theorem. If $f(\cdot)$ is a proper function, then $f^{**} \leq f$.

If $f(\cdot)$ is a proper, closed convex function, then $f^{**} = f$.

Differentiablility

Theorem. Suppose $f(\cdot)$ is a convex function differentiable at $x \in$ **int** (**dom** f). Then, if $y = \nabla f(x)$, then

$$f^*(y) = x^\top \nabla f(x) - f(x)$$

Proof. If $y = \nabla f(x)$, then $z = x$ maximizes the function $z^\top y - f(z)$ (note that this function is concave and differentiable, so first order conditions are sufficient). □

Note: This allows us to determine $f^*(y)$ for any y where we can solve $y = \nabla f(x)$ for x.

Example: Chernoff Bounds

Suppose that X is a real-valued random variable. If $X \geq 0$, and $t > 0$ is a constant,

$$\mathbf{1}_{\{X \geq t\}} \leq \frac{1}{t} X \;\Rightarrow\; \mathsf{E}\left[\mathbf{1}_{\{X \geq t\}}\right] \leq \frac{1}{t}\mathsf{E}[X] \;\Rightarrow\; \mathsf{P}(X \geq t) \leq \frac{1}{t}\mathsf{E}[X]$$

The last inequality is known as Markov's inequality.

Now, if X is any real-valued random variable and $\lambda > 0$ is a constant,

$$\mathbf{1}_{\{X \geq t\}} \leq e^{\lambda(X-t)} \;\Rightarrow\; \mathsf{E}\left[\mathbf{1}_{\{X \geq t\}}\right] \leq \mathsf{E}\left[e^{\lambda(X-t)}\right]$$

$$\Rightarrow\; \mathsf{P}(X \geq t) \leq e^{-\lambda t}\mathsf{E}\left[e^{\lambda X}\right]$$

This is known as a Chernoff bound. The inequality trivially holds for $\lambda = 0$, so we have

$$\mathsf{P}(X \geq t) \leq e^{-\lambda t}\mathsf{E}\left[e^{\lambda X}\right], \quad \forall\, \lambda \geq 0$$

Example: Chernoff Bounds

Define the cumulant generating function

$$f(\lambda) \triangleq \log \mathsf{E}\left[e^{\lambda X}\right], \quad f_+(\lambda) \triangleq \begin{cases} f(\lambda) & \text{if } \lambda \geq 0, \\ \infty & \text{otherwise.} \end{cases}$$

Note that these are always convex functions!

Optimizing over the choice of $\lambda \geq 0$ in the Chernoff bound,

$$\mathsf{P}(X \geq t) \leq \inf_{\lambda \geq 0}\; e^{-\lambda t}\mathsf{E}\left[e^{\lambda X}\right] = \exp\left(-\sup_{\lambda \geq 0}\; \lambda t - f(\lambda)\right)$$

$$\Rightarrow\; \mathsf{P}(X \geq t) \leq e^{-f_+^*(t)}$$

The tightest possible bound is closely related to the conjugate of $f_+(\cdot)$!

Example: Large Deviations

Consider a collection of random vectors $\{X_1, X_2, \ldots\}$, where each $X_i \in \{0, 1\}$ is independently and identically distributed coin flip, and $\mathsf{E}[X_i] = 1/2$.

Consider the cumulative 'number of heads'

$$Y_k = \sum_{i=1}^{k} X_i$$

By the law of large numbers, we expect that $Y_k/k \approx 1/2$. But how fast? To be precise, if $\alpha > 1/2$, how fast does

$$\mathsf{P}(\, Y_k/k \geq \alpha)$$

go to zero? The study of this question is known as large deviations.

Example: Large Deviations

Note that, if $\lambda \geq 0$,

$$\mathsf{P}(\, Y_k \geq \alpha k) \leq e^{-\alpha\lambda k}\mathsf{E}\left[e^{\lambda Y_k}\right] = e^{-\alpha\lambda k}\left(\mathsf{E}\left[e^{\lambda X_1}\right]\right)^k$$

Define

$$f(\lambda) \triangleq \log \mathsf{E}\left[e^{\lambda X_1}\right] = \log\left(1 + e^{\lambda}\right) - \log 2$$

Then,

$$\mathsf{P}(\, Y_k \geq \alpha k) \leq \exp\left\{-k\left(\lambda\alpha - f(\lambda)\right)\right\}$$

When $1/2 < \alpha < 1$,

$$\mathsf{P}(\, Y_k \geq \alpha k) \leq \exp\left(-k \sup_{\lambda \geq 0} \lambda\alpha - f(\lambda)\right) = \exp\left(-k f_+^*(\alpha)\right)$$

and the probability goes to zero exponentially fast at the rate

$$\begin{aligned} f_+^*(\alpha) &= \alpha \log \alpha + (1-\alpha)\log(1-\alpha) + \log 2 \\ &= \alpha \log \frac{\alpha}{1/2} + (1-\alpha)\log\frac{1-\alpha}{1/2} \end{aligned}$$

Example: Multivariate Chernoff Bounds

More generally, given a random variable X taking values in $\mathbb{R}^n$ and a set $\mathcal{C} \subset \mathbb{R}^n$, how can we estimate or bound $\mathsf{P}(X \in \mathcal{C})$?

Suppose we have a function $f\colon \mathbb{R}^n \to \mathbb{R}$, so that

$$\mathbf{1}_{\{z \in \mathcal{C}\}} \leq f(z),\ \forall\ z \in \mathbb{R}^n$$

Then,

$$\mathsf{P}(X \in \mathcal{C}) \leq \mathsf{E}\left[f(X)\right]$$

Consider functions of the form

$$f(z) = e^{\lambda^\top z + \mu}$$

with parameters $\lambda \in \mathbb{R}^n$, $\mu \in \mathbb{R}$.

$$\mathbf{1}_{\{z \in \mathcal{C}\}} \leq f(z),\ \forall\ z \in \mathbb{R}^n \quad \Leftrightarrow \quad \lambda^\top z + \mu \geq 0,\ \forall\ z \in \mathcal{C}$$

Example: Multivariate Chernoff Bounds

$$\mathsf{P}(X \in \mathcal{C}) \leq \mathsf{E}\left[e^{\lambda^\top X + \mu}\right] \quad \text{if } \lambda^\top z + \mu \geq 0,\ \forall\ z \in \mathcal{C}$$

Define the cumulant generating function

$$f(\lambda) \triangleq \log \mathsf{E}\left[e^{\lambda^\top X}\right]$$

Then,

$$\begin{aligned}
\log \mathsf{P}(X \in \mathcal{C}) &\leq \inf_{\lambda,\mu} \left\{\mu + f(\lambda)\ :\ -\lambda^\top z \leq \mu,\ \forall\ z \in \mathcal{C}\right\} \\
&= \inf_{\lambda} \left\{\sup_{z \in \mathcal{C}} (-\lambda^\top z) + f(\lambda)\right\} \\
&= \inf_{\lambda} \left\{S_{\mathcal{C}}(-\lambda) + f(\lambda)\right\} \\
&= -f_C^*(0)
\end{aligned}$$

with

$$f_C(\lambda) \triangleq S_{\mathcal{C}}(-\lambda) + f(\lambda)$$

Example: Gaussian on a Polyhedron

Suppose $X \sim N(0, I)$ is a multivariate Gaussian, i.e.,

$$\mathsf{P}(X \in \mathcal{C}) = \int_{\mathcal{C}} \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}x^\top x}\, dx$$

Then,

$$f(\lambda) \triangleq \log \mathsf{E}\left[e^{\lambda^\top X}\right] = \tfrac{1}{2}\lambda^\top \lambda$$

Given a polyhedron $\mathcal{C} = \{x \ : \ Ax \leq b\}$, how can we bound $\mathsf{P}(X \in \mathcal{C})$?

$$S_{\mathcal{C}}(y) = \sup_{Ax \leq b} x^\top y = \inf_{u \geq 0,\ A^\top u = y} b^\top u \qquad \text{(LP Duality)}$$

$$\log \mathsf{P}(X \in \mathcal{C}) \leq \inf_{\lambda,\ u \geq 0, A^\top u + \lambda = 0} b^\top u + \tfrac{1}{2}\lambda^\top \lambda$$

Example: Gaussian on a Polyhedron

$$\log \mathsf{P}(X \in \mathcal{C}) \leq \inf_{u \geq 0} b^\top u + \tfrac{1}{2} u^\top A A^\top u$$

$$\log \mathsf{P}(X \in \mathcal{C}) \leq \sup_{Ax \leq b} -\tfrac{1}{2}\|x\|_2^2 \qquad \text{(QP Duality)}$$

$$\mathsf{P}(X \in \mathcal{C}) \leq \exp\left(-\tfrac{1}{2}\mathsf{dist}(0, \mathcal{C})^2\right)$$

where $\mathsf{dist}(0, \mathcal{C})$ is the shortest Euclidean distance between 0 and a point in $\mathcal{C}$

Conjugacy & Duality

Consider a proper convex function $f:\ \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$. Consider the optimization problem

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \leq 0 \\ & [x \in \textbf{dom}\ f] \end{array}$$

The Lagrangian is

$$L(x, \lambda) = f(x) + \lambda^\top x$$

The dual objective function is

$$q(\lambda) = \inf_x\ f(x) + \lambda^\top x = -\sup_x\ (-\lambda)^\top x - f(x) = -f^*(-\lambda)$$

Conjugacy & Duality

$f:\ \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$
proper, convex
$A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$
$C \in \mathbb{R}^{r \times n}$, $d \in \mathbb{R}^r$

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & Ax = b \\ & Cx \leq d \\ & [x \in \textbf{dom}\ f] \end{array}$$

The Lagrangian is

$$L(x, \lambda, \mu) = f(x) + \lambda^\top (Ax - b) + \mu^\top (Cx - d)$$

The dual objective function is

$$\begin{aligned} q(\lambda, \mu) &= \inf_x\ f(x) + \lambda^\top (Ax - b) + \mu^\top (Cx - d) \\ &= -b^\top \lambda - d^\top \mu + \inf_x\ f(x) + \lambda^\top Ax + \mu^\top Cx \\ &= -b^\top \lambda - d^\top \mu - f^*(-A^\top \lambda - C^\top \mu) \end{aligned}$$

How to Take a Dual

In general, in order of difficulty:

- Lookup a standard form, e.g., linear program
- Use conjugacy
- Directly from the definition

In general, the dual is sensitive to problem formulation. For example, the problems

$$\begin{array}{lll} \text{minimize} & \|x\|_2 \\ \text{subject to} & Ax = b \\ & x \in \mathbb{R}^n \end{array} \qquad \begin{array}{lll} \text{minimize} & \frac{1}{2}\|x\|_2^2 \\ \text{subject to} & Ax = b \\ & x \in \mathbb{R}^n \end{array}$$

are mathematically equivalent, but have very different duals!

The Primal Problem

$$\begin{array}{l} f : \Omega \to \mathbb{R} \\ g : \Omega \to \mathbb{R}^r \\ \Omega \subset \mathbb{R}^n \end{array} \qquad \begin{array}{lll} \text{minimize} & f(x) \\ \text{subject to} & g(x) \leq u \\ & x \in \Omega \end{array}$$

Definition. The primal function is given by

$$p(u) \triangleq \inf_{x \in \Omega,\ g(x) \leq u} f(x)$$

with domain

$$\textbf{dom}\ p \triangleq \{u \in \mathbb{R}^r \ : \ \exists\, x \in \Omega \text{ with } g(x) \leq u\}$$

In order to keep things simple, we will make the assumption that the primal problem is always bounded.

Assumption. Assume that $p(u) > -\infty$ for all $u \in$ **dom** p. Then, $p : \mathbb{R}^r \to \mathbb{R} \cup \{\infty\}$ is an extended-real valued function.

The Primal Problem

$$
\begin{array}{ll}
f: \Omega \to \mathbb{R} & \text{minimize} \quad f(x) \\
g: \Omega \to \mathbb{R}^r & \text{subject to} \quad g(x) \leq u \\
\Omega \subset \mathbb{R}^n & \qquad\qquad x \in \Omega
\end{array}
$$

Theorem. Suppose that Ω is convex, and that $f(\cdot)$ and $g_j(\cdot)$, $1 \leq j \leq r$, are convex over Ω. Then, $p(\cdot)$ is convex.

Theorem. If $u_1 \geq u_2$, then $p(u_1) \leq p(u_2)$.

The Dual Function

$$
\begin{array}{ll}
f: \Omega \to \mathbb{R} & \text{minimize} \quad f(x) \\
g: \Omega \to \mathbb{R}^r & \text{subject to} \quad g(x) \leq 0 \\
\Omega \subset \mathbb{R}^n & \qquad\qquad x \in \Omega
\end{array}
$$

Then, for $\mu \geq 0$,

$$
\begin{aligned}
q(\mu) &\triangleq \inf_{x \in \Omega} \; f(x) + \mu^\top g(x) \\
&= \inf_{x,\; u} \left\{ f(x) + \mu^\top g(x) : x \in \Omega,\; u \in \mathbb{R}^r,\; g(x) \leq u \right\} \\
&= \inf_{x,\; u} \left\{ f(x) + \mu^\top u : x \in \Omega,\; u \in \mathbb{R}^r,\; g(x) \leq u \right\} \\
&= \inf_{u \in \mathbb{R}^r} \; \inf_{x \in \Omega,\; g(x) \leq u} f(x) + \mu^\top u \\
&= \inf_{u \in \mathbb{R}^r} \; \mu^\top u + \inf_{x \in \Omega,\; g(x) \leq u} f(x) \\
&= -p^*(-\mu)
\end{aligned}
$$

Sensitivity

$$f:\ \Omega \to \mathbb{R} \qquad \text{minimize} \quad f(x)$$
$$g:\ \Omega \to \mathbb{R}^r \qquad \text{subject to} \quad g(x) \leq u$$
$$\Omega \subset \mathbb{R}^n \qquad x \in \Omega$$

Theorem. Suppose that strong duality holds, and the dual optimum is attained when $u = 0$, with μ^* being a geometric multiplier. Then, for all $u \in \mathbb{R}^n$,

$$p(u) \geq p(0) - u^\top \mu^*$$

Proof. For all $x \in \Omega$ with $g(x) \leq u$,

$$p(0) = q(\mu^*) \leq f(x) + g(x)^\top \mu^* \leq f(x) + u^\top \mu^*$$

The result follows. □

Note: If $p(\cdot)$ is convex and differentiable, this result implies that

$$\nabla p(0) = -\mu^*$$

Subgradients

Definition. Consider a proper convex function $f:\ \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$. A vector g is a subgradient of f at $x \in$ **dom** f if

$$f(z) \geq f(x) + g^\top (z - x), \quad \forall\ z \in \mathbb{R}^n$$

The subdifferential $\partial f(x)$ is defined to be the set of all subgradients at x.

Note: $\partial f(x) \triangleq \emptyset$ if $x \notin$ **dom** f

Note: $\partial f(x)$ is a closed and convex set

Note: x minimizes f iff $0 \in \partial f(x)$

Subgradients

Theorem. Suppose that f is a proper convex function. Then, if $x \in \textbf{int}\,(\textbf{dom}\ f)$, $\partial f(x)$ is non-empty.

Proof. Use supporting hyperplane theorem! □

Theorem. Suppose that f is differentiable in a neighborhood of $x \in \textbf{int}\,(\textbf{dom}\ f)$. Then, $\partial f(x) = \{\nabla f(x)\}$.

Proof. Clearly $\nabla f(x) \in \partial f(x)$. To see uniqueness, suppose that $g \in \partial f(x)$. Then, for any $d \in \mathbb{R}^n$ and $\alpha \in \mathbb{R}$ sufficiently small,

$$f(x) + \alpha g^\top d \leq f(x + \alpha d) = f(x) + \alpha \nabla f(x)^\top d + o(|\alpha|)$$

Taking $d = \nabla f(x) - g$,

$$0 \leq \alpha(\nabla f(x) - g)^\top d + o(|\alpha|) = \alpha \|\nabla f(x) - g\|^2 + o(|\alpha|)$$

Thus, if $\alpha < 0$,

$$\|\nabla f(x) - g\|^2 \leq -\frac{o(|\alpha|)}{\alpha} \to 0$$

□

Subgradients

Theorem. Suppose f is a proper convex function. Then,

$$x^\top y = f(x) + f^*(y) \text{ iff } y \in \partial f(x)$$

If, in addition, f is closed, these are equivalent also to $x \in \partial f^*(y)$.

Proof. If $x^\top y = f(x) + f^*(y)$, then x attains the supremum in

$$f^*(y) = \sup_z\ y^\top z - f(z)$$

This is the case if and only if y is a subgradient.

The last part follows since, if f is closed $f^{**} = f$, so the first part can be applied with the roles of f^* and f reversed. □

Theorem. Suppose f is a proper closed convex function. Then,

(a) f^* is differentiable at $y \in \textbf{int}\,(\textbf{dom}\ f^*)$ if and only if the supremum of $x^\top y - f(x)$ is uniquely attained over $x \in \mathbb{R}^n$

(b)

$$\operatorname*{argmin}_{x \in \mathbb{R}^n} f(x) = \partial f^*(0)$$

Proof. (Sketch) Both statements follow from previous theorem, which implies that

$$\operatorname*{argmax}_{x \in \mathbb{R}^n} x^\top y - f(x) = \partial f^*(y)$$

□

B9824 Foundations of Optimization

Lecture 7: Problem Formulation I

Fall 2011

Outline

1. Standard forms of convex programs
2. Equivalent formulations
3. Geometric programming
4. Robust optimization

Why is Convexity Important?

- A convex function has no local minima that are not global minima
- A convex set is connected and has feasible directions at every point
- A convex set can be characterized by extreme points or supporting hyperplanes
- Efficient algorithms are available for solving convex optimization problems

Equivalent Formulations

Two optimization problems are equivalent if the solution of one can be readily obtained from the other, and vice versa.

The same optimization problem can admit multiple, different formulations. Some formulations may be more advantageous than others:

- More or fewer variables/constraints
- Simpler or more complicated objective function/constraint set
- Sparse vs. not-sparse
- Convex vs. non-convex
- Differentiable vs. non-differentiable
- Decentralized vs. centralized
- Different dual problems

Standard Forms

A convex optimization problem in standard form is:

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & Ax = b \\ & g_i(x) \leq 0, \quad \forall\, 1 \leq i \leq r \\ & x \in \mathbb{R}^n \end{array}$$

$f\colon \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$, convex
$g_i\colon \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$, convex
$A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$

- Convex objective
- Convex inequality constraints
- Linear equality constraints

Standard Forms

If an optimization problem can be converted to more specialized standard form, it can be solved efficiently by off-the-shelf software.

Some common standard forms for convex problems are:

Linear Program (LP)

$$\begin{array}{ll} \text{minimize} & c^\top x \\ \text{subject to} & Ax = b \\ & Gx \leq h \\ & x \in \mathbb{R}^n \end{array}$$

$c \in \mathbb{R}^n$
$A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$
$G \in \mathbb{R}^{r\times n}$, $h \in \mathbb{R}^r$

linear objective, linear equality/inequality constraints

Standard Forms

Quadratic Program (QP)

minimize	$\frac{1}{2}x^\top Qx + c^\top x$	$Q \in \mathbb{R}^{n\times n}$ symmetric, positive semidefinite
subject to	$Ax = b$	$c \in \mathbb{R}^n$
	$Gx \leq h$	$A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$
	$x \in \mathbb{R}^n$	$G \in \mathbb{R}^{r\times n}$, $h \in \mathbb{R}^r$

convex quadratic objective, linear equality/inequality constraints

Examples: portfolio optimization, linear regression

LP $\subset$ QP

Standard Forms

Quadratically Constrained Quadratic Program (QCQP)

minimize	$\frac{1}{2}x^\top Qx + c^\top x$	$Q,\ P_i \in \mathbb{R}^{n\times n}$ symmetric, positive semidefinite
subject to	$Ax = b$	$c \in \mathbb{R}^n$
	$\frac{1}{2}x^\top P_i x + g_i^\top x + h_i \leq 0,$	$A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$
	$\forall\ 1 \leq i \leq r$	$g_i \in \mathbb{R}^n$, $h_i \in \mathbb{R}$
	$x \in \mathbb{R}^n$	

convex quadratic objective, convex quadratic inequality constraints, linear equality constraints

LP $\subset$ QP $\subset$ QCQP

Standard Forms

Second Order Cone Program (SOCP)

$$\begin{array}{ll} \text{minimize} & c^\top x \\ \text{subject to} & Ax = b \\ & \|F_i x + q_i\|_2 \leq g_i^\top x + h_i, \\ & \forall\ 1 \leq i \leq r \\ & x \in \mathbb{R}^n \end{array}$$

$c \in \mathbb{R}^n$
$A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$
$F_i \in \mathbb{R}^{n_i \times n}$, $q_i \in \mathbb{R}^{n_i}$
$g_i \in \mathbb{R}^n$, $h_i \in \mathbb{R}$

linear objective and equality constraints, second-order cone constraints

LP $\subset$ QP $\subset$ QCQP $\subset$ SOCP

Standard Forms

Semidefinite Program (SDP)

$$\begin{array}{ll} \text{minimize} & c^\top x \\ \text{subject to} & Ax = b \\ & x_1 F_1 + \cdots + x_n F_n + H \preceq 0 \\ & x \in \mathbb{R}^n \end{array}$$

$F_i, H \in \mathbb{R}^{k \times k}$, symmetric
$c \in \mathbb{R}^n$
$A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$

linear objective and equality constraints, linear matrix inequalities

LP $\subset$ QP $\subset$ QCQP $\subset$ SOCP $\subset$ SDP

Schur Complements

Symmetric block matrix:

$$X = X^\top = \begin{bmatrix} A & B \\ B^\top & C \end{bmatrix}$$

Here, A and C are symmetric matrices.

Definition. The Schur complement of A in X is the symmetric matrix

$$S \triangleq C - B^\top A^{-1} B$$

assuming $\det A \neq 0$.

Lemma.

- $X \succ 0 \Leftrightarrow A \succ 0$ and $S \succ 0$
- if $A \succ 0$, then $X \succeq 0 \Leftrightarrow S \succeq 0$

Schur Complements

Schur complements can be used to trade-off dimension with non-linearity and express nonlinear convex constraints as LMIs

Example. Convex quadratic constraint ($P \succeq 0$)

$$\tfrac{1}{2} x^\top P x + g^\top x + h \leq 0$$

$$\Leftrightarrow \quad \begin{bmatrix} I & P^{1/2} x \\ (P^{1/2} x)^\top & -g^\top x - h \end{bmatrix} \succeq 0$$

Example. Convex second order cone constraint

$$\|Fx + q\|_2 \leq g^\top x + h$$

$$\Leftrightarrow \quad \begin{bmatrix} (g^\top x + h) I & Fx + q \\ (Fx + q)^\top & g^\top x + h \end{bmatrix} \succeq 0$$

Example: Eigenvalue Minimization

$A_0, A_1, \ldots, A_n \in \mathbb{R}^{k\times k}$ symmetric matrices

$$A(x) \triangleq A_0 + x_1 A_1 + \cdots + x_n A_n$$

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \ \lambda_{\max}(A(x)) \triangleq \text{largest eigenvalue of } A(x)$$

Note that

$$\lambda_{\max}(A(x)) \le t \quad \Leftrightarrow \quad A(x) \preceq tI$$

Equivalent SDP formulation:

$$\begin{array}{ll} \text{minimize} & t \\ \text{subject to} & A(x) \preceq tI \\ & x \in \mathbb{R}^n, t \in \mathbb{R} \end{array}$$

Transformation of Objective/Constraints

$$\begin{array}{l} f\colon \Omega \to \mathbb{R} \\ h\colon \Omega \to \mathbb{R}^m \\ g\colon \Omega \to \mathbb{R}^r \\ \Omega \subset \mathbb{R}^n \end{array} \qquad \begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & h(x) = 0 \\ & g(x) \le 0 \\ & x \in \Omega \end{array}$$

Suppose:

- $\phi : \mathbb{R} \to \mathbb{R}$ is a strictly increasing function
- $\psi : \mathbb{R}^m \to \mathbb{R}^m$ satisfies $\psi(u) = 0$ if and only if $u = 0$
- $\chi : \mathbb{R}^r \to \mathbb{R}^r$ satisfies $\chi(v) \le 0$ if and only if $v \le 0$

Then, an equivalent problem is:

$$\begin{array}{l} \tilde{f}(x) = \phi(f(x)) \\ \tilde{h}(x) = \psi(h(x)) \\ \tilde{g}(x) = \chi(g(x)) \end{array} \qquad \begin{array}{ll} \text{minimize} & \tilde{f}(x) \\ \text{subject to} & \tilde{h}(x) = 0 \\ & \tilde{g}(x) \le 0 \\ & x \in \Omega \end{array}$$

Transformation of Objective/Constraints: Example

Consider the least-norm approximation problems:

$$\begin{array}{ll} \text{minimize} & \|Ax-b\|_2 \\ \text{subject to} & x \in \mathbb{R}^n \end{array} \qquad \begin{array}{ll} \text{minimize} & \|Ax-b\|_2^2 \\ \text{subject to} & x \in \mathbb{R}^n \end{array}$$

These problems have the same global optimum, and both are convex. However:

- the objective in first problem is not differentiable for x with $Ax - b = 0$, it is an SOCP
- the second objective is differentiable for all x, and, in fact, is a QP

Change of Variables

$$\begin{array}{l} f\colon \Omega \to \mathbb{R} \\ h\colon \Omega \to \mathbb{R}^m \\ g\colon \Omega \to \mathbb{R}^r \\ \Omega \subset \mathbb{R}^n \end{array} \qquad \begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & h(x) = 0 \\ & g(x) \leq 0 \\ & x \in \Omega \end{array}$$

Suppose $\phi\colon \tilde{\Omega} \to \Omega$ is surjective. Then, an equivalent problem is:

$$\begin{array}{l} \tilde{f}(z) = f(\phi(z)) \\ \tilde{h}(z) = h(\phi(z)) \\ \tilde{g}(z) = g(\phi(z)) \end{array} \qquad \begin{array}{ll} \text{minimize} & \tilde{f}(z) \\ \text{subject to} & \tilde{h}(z) = 0 \\ & \tilde{g}(z) \leq 0 \\ & z \in \tilde{\Omega} \end{array}$$

Example: Geometric Programming

Definition. A monomial is a function $f\colon (0,\infty)^n \to \mathbb{R}$ of the form

$$f(x) = c x_1^{a_1} x_2^{a_2} \cdots x_n^{a_n}$$

where $c > 0$ and $a_i \in \mathbb{R}$.

Definition. A posynomial is a function $f\colon (0,\infty)^n \to \mathbb{R}$ that is a sum of monomials, that is, of the form

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

where $c_k > 0$ and $a_{ik} \in \mathbb{R}$.

Example: Geometric Programming

Definition. A geometric program (GP) is an optimization program of the form

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & h_i(x) = 1, \quad \forall\ 1 \le i \le m \\ & g_j(x) \le 1, \quad \forall\ 1 \le j \le r \\ & x > 0 \\ & x \in \mathbb{R}^n \end{array}$$

where

- $f\colon (0,\infty)^n \to \mathbb{R}$ is a posynomial
- each $h_i\colon (0,\infty)^n \to \mathbb{R}$ is a monomial
- each $g_j\colon (0,\infty)^n \to \mathbb{R}$ is a posynomial

Example: Maximum Volume Box

$$\begin{array}{ll} \text{maximize} & x_1 x_2 x_3 \\ \text{subject to} & x_1 x_2 + x_2 x_3 + x_1 x_3 \leq c/2 \quad (c > 0) \\ & x > 0 \\ & x \in \mathbb{R}^3 \end{array}$$

This is equivalent to the standard form GP

$$\begin{array}{ll} \text{minimize} & x_1^{-1} x_2^{-1} x_3^{-1} \\ \text{subject to} & \frac{2}{c} x_1 x_2 + \frac{2}{c} x_2 x_3 + \frac{2}{c} x_1 x_3 \leq 1 \\ & x > 0 \\ & x \in \mathbb{R}^3 \end{array}$$

Geometric Programming: Convex Form

GPs are not convex. However, they can easily be converted to a convex form. Consider the posynomial

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

Apply change of variables $y_i = \log x_i$ (since $x_i > 0$),

$$f(y) = \sum_{k=1}^{K} \exp\left(\sum_{i=1}^{n} a_{ik} y_i + b_k \right) = \sum_{k=1}^{K} \exp(a_k^\top y + b_k)$$

where $a_k = (a_{1k}, \ldots, a_{nk})$ and $b_k = \log c_k$.

Taking a logarithm,

$$\log f(y) = \log \left(\sum_{k=1}^{K} \exp(a_k^\top y + b_k) \right)$$

This is a convex function.

Geometric Programming: Convex Form

Definition. A convex form GP is an optimization program of the form

$$\begin{array}{lll} \text{minimize} & \log\left(\displaystyle\sum_{k=1}^{K_0} \exp(a_k^\top y + b_k)\right) & \\ \text{subject to} & c_i^\top y + d_i = 0, & \forall\ 1 \le i \le m \\ & \log\left(\displaystyle\sum_{k=1}^{K_j} \exp(e_{kj}^\top y + f_{kj})\right) \le 0, & \forall\ 1 \le j \le r \\ & y \in \mathbb{R}^n & \end{array}$$

Note that if $K_j = 1$ for all $0 \le j \le r$, this reduces to an LP.

Eliminating Equality Constraints

$$\begin{aligned} f &: \mathbb{R}^n \to \mathbb{R} \\ h &: \mathbb{R}^n \to \mathbb{R}^m \\ g &: \mathbb{R}^n \to \mathbb{R}^r \end{aligned}$$

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & h(x) = 0 \\ & g(x) \le 0 \\ & x \in \mathbb{R}^n \end{array}$$

Suppose $\phi : \mathbb{R}^k \to \mathbb{R}^n$ is such that $h(x) = 0$ if and only if $x = \phi(z)$, for some $z \in \mathbb{R}^k$. Then, an equivalent problem is:

$$\begin{aligned} \tilde{f}(z) &= f(\phi(z)) \\ \tilde{g}(z) &= g(\phi(z)) \end{aligned}$$

$$\begin{array}{ll} \text{minimize} & \tilde{f}(z) \\ \text{subject to} & \tilde{g}(z) \le 0 \\ & z \in \mathbb{R}^k \end{array}$$

Eliminating Equality Constraints

$$f\colon \mathbb{R}^n \to \mathbb{R}$$
$$A \in \mathbb{R}^{m\times n},\ b \in \mathbb{R}^m$$
$$g\colon \mathbb{R}^n \to \mathbb{R}^r$$

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & Ax = b \\ & g(x) \le 0 \\ & x \in \mathbb{R}^n \end{array}$$

If $Ax = b$ has a solution, then all solutions are of the form $x = Fz + x_0$ for $z \in \mathbb{R}^k$, where $F \in \mathbb{R}^{n\times k}$, and x_0 is any solution. We can pick $k = n - \text{rank}(A)$. Then,

$$\begin{array}{ll} \text{minimize} & f(Fz + x_0) \\ \text{subject to} & g(Fz + x_0) \le 0 \\ & z \in \mathbb{R}^k \end{array}$$

Note that this problem is of smaller dimension and has no equality constraints. Convexity is also preserved.

Adding Equality Constraints

Adding equality constraints can be helpful for decomposing and optimization problem into independent subproblems. Consider:

$$A_i \in \mathbb{R}^{m_i\times n},\ b \in \mathbb{R}^{m_i}$$
$$f\colon \mathbb{R}^{m_0} \to \mathbb{R}$$
$$g_i\colon \mathbb{R}^{m_i} \to \mathbb{R}$$

$$\begin{array}{ll} \text{minimize} & f(A_0x + b_0) \\ \text{subject to} & g_i(A_ix + b_i) \le 0,\ \forall 1 \le i \le r \\ & x \in \mathbb{R}^n \end{array}$$

This is equivalent to:

$$\begin{array}{lll} \text{minimize} & f(y_0) & \\ \text{subject to} & g_i(y_i) \le 0, & \forall\ 1 \le i \le r \\ & y_i = A_ix + b_i, & \forall\ 0 \le i \le r \\ & x \in \mathbb{R}^n & \\ & y_i \in \mathbb{R}^{m_i}, & \forall\ 0 \le i \le r \end{array}$$

This problem has an objective and inequality constraints which are independent.

Partitioning

It is always true that

$$\inf_{x,y} f(x,y) = \inf_{x} \tilde{f}(x), \quad \text{where } \tilde{f}(x) = \inf_{y} f(x,y)$$

For example:

$$\begin{array}{l} f\colon \mathbb{R}^{n_1} \to \mathbb{R} \\ g\colon \mathbb{R}^{n_1 \times n_2} \to \mathbb{R} \\ A_i \in \mathbb{R}^{m \times n_i},\ b \in \mathbb{R}^m \\ \Omega_i \subset \mathbb{R}^{n_i} \end{array}$$

$$\begin{array}{ll} \text{minimize} & f(x_1) + g(x_1, x_2) \\ \text{subject to} & A_1 x_1 + A_2 x_2 = b \\ & x_i \in \Omega_i,\ i = 1, 2 \end{array}$$

Equivalent problem:

$$\tilde{g}(x_1) = \inf_{\substack{A_2 x_2 = c - A_1 x_1, \\ x_2 \in \Omega_2}} g(x_1, x_2)$$

$$\begin{array}{ll} \text{minimize} & f_1(x_1) + \tilde{g}_2(x_1) \\ \text{subject to} & x_1 \in \Omega_1 \end{array}$$

This is useful when $\tilde{g}(\cdot)$ is easy to compute.

Epigraph Formulation

$$\begin{array}{l} f\colon \Omega \to \mathbb{R} \\ h\colon \Omega \to \mathbb{R}^m \\ g\colon \Omega \to \mathbb{R}^r \\ \Omega \subset \mathbb{R}^n \end{array}$$

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & h(x) = 0 \\ & g(x) \leq 0 \\ & x \in \Omega \end{array}$$

Equivalent problem:

$$\begin{array}{ll} \text{minimize} & t \\ \text{subject to} & f(x) - t \leq 0 \\ & h(x) = 0 \\ & g(x) \leq 0 \\ & x \in \Omega \\ & t \in \mathbb{R} \end{array}$$

This is useful often when $f(\cdot)$ is a 'worst-case' objective.

Example: Min-Max Facility Location

$y_i \in \mathbb{R}^n$

$$\begin{array}{ll} \text{minimize} & \max\limits_{1 \le i \le r} \|x - y_i\|_2 \\ \text{subject to} & x \in \mathbb{R}^n \end{array}$$

Equivalent problem:

$$\begin{array}{ll} \text{minimize} & t \\ \text{subject to} & \|x - y_i\|_2^2 \le t, \quad \forall\, 1 \le i \le n \\ & x \in \mathbb{R}^n \\ & t \in \mathbb{R} \end{array}$$

This is a QCQP.

Linear-Fractional Programming

$$f(x) = \frac{c^\top x + d}{e^\top x + f}$$

dom $f \triangleq \{x \in \mathbb{R}^n \ : \ e^\top x + f > 0\}$

Note: $f(\cdot)$ is not convex!

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & Ax = b \\ & Gx \le h \\ & x \in \mathbb{R}^n \end{array}$$

$c, e \in \mathbb{R}^n$
$d, f \in \mathbb{R}$
$A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$
$G \in \mathbb{R}^{r \times n}$, $h \in \mathbb{R}^r$

Equivalent problem: (if the LFP is feasible)

$$\begin{array}{ll} \text{minimize} & c^\top y + dz \\ \text{subject to} & Az - bz = 0 \\ & Gy - hz \le 0 \\ & e^\top y + fz = 1 \\ & y \in \mathbb{R}^n, \ z \ge 0 \end{array}$$

This is a LP.

Robust Optimization

The parameters in an optimization problem are often uncertain. For example, consider the LP:

$$\begin{array}{ll} \text{minimize} & c^\top x \\ \text{subject to} & a_i^\top x \leq b_i, \quad \forall\, 1 \leq i \leq r \\ & x \in \mathbb{R}^n \end{array}$$

Suppose there is some uncertainty in $\{a_i\}$. We would like "robust" solutions which do not require knowledge of the precise value of a_i.

Robust Optimization

One approach is to require the constraints to hold for all a_i in a set $\mathcal{E}_i \subset \mathbb{R}^n$:

$$\begin{array}{ll} \text{minimize} & c^\top x \\ \text{subject to} & a_i^\top x \leq b_i, \quad \forall\, a_i \in \mathcal{E}_i,\ 1 \leq i \leq r \\ & x \in \mathbb{R}^n \end{array}$$

Equivalent problem with worst-case constraints:

$$\begin{aligned} S_{\mathcal{E}_i}(x) &\triangleq \sup_{a_i \in \mathcal{E}_i} a_i^\top x \\ &= I^*_{\mathcal{E}_i}(x) \end{aligned}$$

$$\begin{array}{ll} \text{minimize} & c^\top x \\ \text{subject to} & S_{\mathcal{E}_i}(x) \leq b_i, \quad \forall\, 1 \leq i \leq r \\ & x \in \mathbb{R}^n \end{array}$$

Robust Optimization

Suppose we take $\mathcal{E}_i$ to be an ellipsoid:

$$\mathcal{E}_i = \{\bar{a}_i + P_i u \ : \ \|u\|_2 \leq 1\}$$

Here, $\bar{a}_i \in \mathbb{R}^n$, $P_i \in \mathbb{R}^{n \times n}$. The axes of the ellipse are determined by the eigenvalues/eigenvectors of $P_i^\top P_i$.

Then,

$$S_{\mathcal{E}_i}(x) = \sup_{\|u\|_2 \leq 1} (\bar{a}_i + P_i u)^\top x = \bar{a}_i^\top x + \|P_i^\top x\|_2$$

Thus, the robust LP is a SOCP:

$$\begin{array}{ll} \text{minimize} & c^\top x \\ \text{subject to} & \bar{a}_i^\top x + \|P_i^\top x\|_2 \leq b_i, \quad \forall\, 1 \leq i \leq r \\ & x \in \mathbb{R}^n \end{array}$$

Robust Optimization

Suppose we take $\mathcal{E}_i$ to be a bounded polyhedron,

$$\mathcal{E}_i = \{a_i \in \mathbb{R}^n \ : \ E_i a_i \leq f_i\}$$

Here, $E_i \in \mathbb{R}^{m_i \times n}$, $f_i \in \mathbb{R}^{m_i}$.

Then,

$$S_{\mathcal{E}_i}(x) = \sup_{E_i y \leq f_i} x^\top y$$

Since $\mathcal{E}_i$ is bounded, the optimum must occur at a vertex $\{\bar{a}_{i,1}, \ldots, \bar{a}_{i,k_i}\}$.

Thus, the robust LP is also an LP (with more constraints):

$$\begin{array}{ll} \text{minimize} & c^\top x \\ \text{subject to} & \bar{a}_{i,j}^\top x \leq b_i, \quad \forall\, 1 \leq i \leq r,\ 1 \leq j \leq k_i \\ & x \in \mathbb{R}^n \end{array}$$

Alternatively, by duality of LPs,

$$S_{\mathcal{E}_i}(x) = \sup_{E_i y \le f_i} x^\top y = \inf_{E_i^\top z_i = x,\ z_i \ge 0} f_i^\top z_i$$

Thus, the robust LP can be expressed as

$$\begin{array}{lll} \text{minimize} & c^\top x & \\ \text{subject to} & f_i^\top z_i \le b_i, & \forall\ 1 \le i \le r \\ & E_i^\top z_i - x = 0, & \forall\ 1 \le i \le r \\ & z_i \ge 0, & \forall\ 1 \le i \le r \\ & z_i \in \mathbb{R}^{m_i} & \\ & x \in \mathbb{R}^n & \end{array}$$

B9824 Foundations of Optimization

Lecture 8: Problem Formulation II

Fall 2011

Outline

1. Approximation
2. Statistical estimation
3. Classification

Norm Approximation

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \ \|Ax-b\| \qquad A\in\mathbb{R}^{m\times n},\ b\in\mathbb{R}^m,\ m\geq n,\ \|\cdot\| \text{ a norm on } \mathbb{R}^n$$

Interpretation: Suppose x^* is an optimal solution

- **geometric**: Ax^* is the point in im A closest to b
- **estimation**: Consider a linear measurement model

  $y = Ax+v$ — y are observations/measurements; x is unknown; v is measurement error

  Given $y=b$, then the best guess of x is x^*
- **optimal design**:
  x are design variables (input), Ax is the result (output)

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

Norm Approximation: Examples

Example. Least-squares approximation ($\|\cdot\|_2$):
Solve the normal equations $A^\top Ax = A^\top b$
If rank $A=n$, then $x^* = (A^\top A)^{-1}A^\top b$
(Note: Computationally, use something numerically robust like SVD.)

Example. Chebyshev approximation ($\|\cdot\|_\infty$):
Solve the LP

$$\begin{array}{ll}\text{minimize} & t\\ \text{subject to} & -t\mathbf{1}\leq Ax-b\leq t\mathbf{1}\\ & x\in\mathbb{R}^n,\ t\in\mathbb{R}\end{array}$$

Example. Sum of absolute residuals ($\|\cdot\|_1$):
Solve the LP

$$\begin{array}{ll}\text{minimize} & \mathbf{1}^\top y\\ \text{subject to} & -y\leq Ax-b\leq y\\ & x\in\mathbb{R}^n,\ y\in\mathbb{R}^n\end{array}$$

Penalty Functions

$$\begin{array}{ll} \text{minimize} & \phi(r_1) + \ldots + \phi(r_m) \\ \text{subject to} & r = Ax - b \\ & r \in \mathbb{R}^m, x \in \mathbb{R}^n \end{array} \qquad A \in \mathbb{R}^{m \times n},\ b \in \mathbb{R}^m,\ m \geq n$$

Here, $\phi \colon \mathbb{R} \to \mathbb{R}$ is a convex penalty function.

Examples:

- ℓ_p-norm: $\phi(u) = |u|^p$, $p \in [1, \infty)$
- Deadzone-linear with width a: $\phi(u) = \max\,(0, |u| - a)$
- Log-barrier with limit a:

$$\phi(u) = \begin{cases} -a^2 \log\left(1 - (u/a)^2\right) & \text{if } |u| < a \\ \infty & \text{otherwise} \end{cases}$$

Least-Norm Problems

$$\begin{array}{ll} \text{minimize} & \|x\| \\ \text{subject to} & Ax = b \\ & x \in \mathbb{R}^n \end{array} \qquad \begin{array}{l} A \in \mathbb{R}^{m \times n},\ b \in \mathbb{R}^m,\ m \leq n \\ \|\cdot\| \text{ a norm on } \mathbb{R}^n \end{array}$$

Interpretation: Suppose x^* is an optimal solution

- **geometric**: x^* is the point in the affine set $\{x \ : \ Ax = b\}$ with minimum distance to 0
- **estimation**: $b = Ax$ are (perfect) measurements of x

  x^* is the smallest ('most plausible') estimate consistent with the measurements
- **optimal design**:
  x are design variables (input), b is the required result (output)

  x^* is smallest ('most efficient') design that meets the requirements

Least-Norm Problems: Examples

Example. Least-squares solution of linear equations ($\|\cdot\|_2$):
Solve the optimality conditions

$$2x + A^\top v = 0, \quad Ax = b.$$

Example. Minimum sum of absolute values ($\|\cdot\|_1$):
Solve the LP

$$\begin{array}{ll} \text{minimize} & \mathbf{1}^\top y \\ \text{subject to} & -y \leq x \leq y \\ & Ax = b \\ & x \in \mathbb{R}^n,\ y \in \mathbb{R}^n \end{array}$$

This tends to produce a sparse solution x^*

Extension: Least-penalty problem

$$\begin{array}{ll} \text{minimize} & \phi(x_1) + \ldots + \phi(x_n) \\ \text{subject to} & Ax = b,\ x \in \mathbb{R}^n \end{array}$$

$\phi\colon \mathbb{R} \to \mathbb{R}$ a convex penalty function

Regularized Approximation

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \ (\|Ax - b\|, \|x\|)$$

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

Norms on $\mathbb{R}^n$ and $\mathbb{R}^m$ can be different

Interpretation: Find a good approximation $Ax \approx b$ with small x

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

Scalarized Regularization

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \ \|Ax-b\| + \gamma\|x\|, \qquad \gamma > 0$$

Solution for various values of γ traces out optimal trade-off curve

- **Tikhonov Regularization**:

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \ \|Ax-b\|_2^2 + \gamma\|x\|_2^2, \qquad \gamma > 0$$

$$\Rightarrow \quad \underset{x\in\mathbb{R}^n}{\text{minimize}} \ \left\| \begin{bmatrix} A \\ \sqrt{\gamma} I \end{bmatrix} x - \begin{bmatrix} b \\ 0 \end{bmatrix} \right\|_2^2 \quad \Rightarrow \quad x^* = (A^\top A + \gamma I)^{-1} A^\top b$$

Scalarized Regularization

- **Smoothness Regularization**:

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \ \|Ax-b\|_2^2 + \gamma\|Dx\|_2^2, \qquad \gamma > 0$$

for some 'differentiation' operator $D \in \mathbb{R}^{k\times n}$

$$\Rightarrow \quad x^* = (A^\top A + \gamma D^\top D)^{-1} A^\top b$$

- **Lasso**:

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \ \|Ax-b\|_2^2 + \gamma\|x\|_1, \qquad \gamma > 0$$

Solve the QP

$$\begin{array}{ll} \text{minimize} & \|Ax-b\|_2^2 + \gamma \mathbf{1}^\top y \\ \text{subject to} & -y \le x \le y \\ & x \in \mathbb{R}^n, \ y \in \mathbb{R}^n \end{array}$$

Heuristic for 'regressor selection'

$$\begin{array}{ll} \text{minimize} & \|Ax-b\|_2 \\ \text{subject to} & \mathbf{card}(x) \le k \\ & x \in \mathbb{R}^n \end{array}$$

Robust Approximation

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \ \|Ax-b\| \quad \text{with uncertain } A$$

Two approaches:

- **stochastic**: assume A is random, optimize the expected error

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \ \mathsf{E}\big[\|Ax-b\|\big] \qquad \text{(Always convex!)}$$

- **worst-case**: assume A comes from an uncertainty set $\mathcal{A}$, optimize the worst-case error

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \ \sup_{A\in\mathcal{A}} \|Ax-b\| \qquad \text{(Always convex!)}$$

Generally need some structure for these problems to be tractable (certain norms, distributions, uncertainty sets, etc.)

Stochastic Robust Approximation

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \ \mathsf{E}\big[\|Ax-b\|\big]$$

Assume that A takes values in the set $\{A_1,\ldots,A_k\}\subset\mathbb{R}^{m\times n}$ with

$$\mathsf{P}\left(A=A_i\right)=p_i, \quad i=1,\ldots,k$$

This sum-of-norms problem can be written as

$$\begin{array}{ll}\text{minimize} & p^\top t\\ \text{subject to} & \|A_ix-b\|\le t_i,\\ & \quad i=1,\ldots,k\\ & x\in\mathbb{R}^n,\ t\in\mathbb{R}^k\end{array}$$

- $\|\cdot\|_2 \Rightarrow$ SOCP
- $\|\cdot\|_1, \|\cdot\|_\infty \Rightarrow$ LP

Stochastic Robust Least-Squares

$$\underset{x\in\mathbb{R}^n}{\text{minimize}}\ \mathsf{E}\big[\|Ax-b\|_2^2\big]$$

We can write $A=\bar{A}+U$, where

- $\bar{A}\triangleq\mathsf{E}[A]$
- $U\triangleq A-\mathsf{E}[A]$ is zero-mean

$$\begin{aligned}\mathsf{E}\big[\|Ax-b\|_2^2\big] &= \|\bar{A}x-b\|_2^2+x^\top Px,\ \text{where } P\triangleq\mathsf{E}\left[U^\top U\right]\\ &= \|\bar{A}x-b\|_2^2+\|P^{1/2}x\|_2^2\end{aligned}$$

Example. If A has i.i.d. entries, $P=\delta I\Rightarrow$ Tikhonov regularization

Worst-Case Robust Approximation

$$\underset{x\in\mathbb{R}^n}{\text{minimize}}\ \sup_{A\in\mathcal{A}}\ \|Ax-b\|$$

$\mathcal{A}\subset\mathbb{R}^{m\times n}$ bounded, non-empty

- **finite set**: $\mathcal{A}\triangleq\{A_1,\ldots,A_k\}$

  Epigraph form:

  $$\begin{array}{ll}\text{minimize} & t\\ \text{subject to} & \|A_ix-b\|\le t,\\ & \quad i=1,\ldots,k\\ & x\in\mathbb{R}^n,\ t\in\mathbb{R}\end{array}$$

  $\|\cdot\|_2\Rightarrow$ SOCP

  $\|\cdot\|_1,\|\cdot\|_\infty\Rightarrow$ LP

- **polyhedron**:
  $\mathcal{A}=$ **conv** $\{A_1,\ldots,A_k\}$, same as finite case

  Tractable if there are not too many vertices of $\mathcal{A}$

Worst-Case Least Squares

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \ \sup_{A\in\mathcal{A}} \|Ax-b\|_2 \qquad \mathcal{A}\subset\mathbb{R}^{m\times n} \text{ bounded, non-empty}$$

- **uncertainty ellipsoid**:

$$\mathcal{A} \triangleq \left\{ [a_1 \cdots a_m]^\top \ : \ a_i \in \mathcal{E}_i, \ i=1,\ldots,m \right\}$$

with

$$\mathcal{E}_i \triangleq \{\bar{a}_i + P_i u \ : \ \|u\|_2 \le 1\}$$

Equivalent to the SOCP

$$\begin{array}{ll} \text{minimize} & \delta \\ \text{subject to} & \bar{a}_i^\top x - b_i + \|P_i^\top x\|_2 \le t_i, \\ & \quad i=1,\ldots,m \\ & -\bar{a}_i^\top x + b_i + \|P_i^\top x\|_2 \le t_i, \\ & \quad i=1,\ldots,m \\ & \|t\|_2 \le \delta, \\ & x\in\mathbb{R}^n, \ t\in\mathbb{R}^m, \delta\in\mathbb{R} \end{array}$$

Parametric Estimation

Suppose we wish to estimate the density $p(y)$ of a random variable given an observation.

In a parametric estimation problem, we would like to find the best guess of $p(y)$ from a family of densities $p_x(y)$ indexed by $x\in\Omega$.

The maximum likelihood (ML) estimation problem is given an observation y is:

$$\begin{array}{ll} \text{maximize} & p_x(y) \\ & x\in\Omega \end{array} \iff \begin{array}{ll} \text{maximize} & \log p_x(y) \\ & x\in\Omega \end{array}$$

If $\Omega\subset\mathbb{R}^n$ is a convex set and the log-likelihood

$$\ell(x) = \log p_x(y)$$

is concave (for fixed y), this is a convex optimization problem.

Example: Linear Measurements

$$y_i = a_i^\top x + v_i, \quad i = 1, \ldots, m$$

- $x \in \Omega \subset \mathbb{R}^n$ is a vector of unknown parameters
- v_i is IID measurement noise, with density $p(z)$
- y_i is a measurement, $y \in \mathbb{R}^m$ has density

$$p_x(y) = \prod_{i=1}^m p(y_i - a_i^\top x)$$

The ML estimation problem is

$$\begin{array}{ll} \text{maximize} & \ell(x) = \sum_{i=1}^m \log p(y_i - a_i^\top x) \\ \text{subject to} & x \in \Omega \end{array}$$

This also has a least penalty approximation interpretation.

Example: Linear Measurements

Example. Gaussian noise: $p(z) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-z^2/(2\sigma^2)}$, $\sigma > 0$

$$\ell(x) = -\frac{m}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^m (a_i^\top x - y_i)^2$$

The ML estimate is a least-squares approximation.

Example. Laplacian noise: $p(z) = \frac{1}{2\alpha} e^{-|z|/\alpha}$, $\alpha > 0$

$$\ell(x) = -m\log 2\alpha - \frac{1}{\alpha}\sum_{i=1}^m |a_i^\top x - y_i|$$

The ML estimate is an ℓ_1-norm approximation.

Example: Linear Measurements

Example. Uniform noise: $p(z) = \begin{cases} 1/(2\alpha) & \text{if } z \in [-\alpha, \alpha] \\ 0 & \text{otherwise} \end{cases}$

$$\ell(x) = \begin{cases} -m \log 2\alpha & \text{if } |y_i - a_i^\top x| \leq \alpha, \; \forall \; i \\ -\infty & \text{otherwise} \end{cases}$$

The ML estimate is any x with $|y_i - a_i^\top x| \leq \alpha, \; \forall \; i$.

Example: Logistic Regression

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

$$p = \mathsf{P}(y = 1) = \frac{\exp(a^\top x)}{1 + \exp(a^\top x)}$$

- $x \in \mathbb{R}^n$ are unknown parameters; $a \in \mathbb{R}^n$ are observable explanatory variables
- estimation problem: estimate x from m observations (a_i, y_i)

Assume that $y_1 = \cdots = y_k = 1$, $y_{k+1} = \cdots = y_m = 0$.

$$\begin{aligned} \ell(x) &= \log \left(\prod_{i=1}^{k} \frac{\exp(a_i^\top x)}{1 + \exp(a_i^\top x)} \prod_{i=k+1}^{m} \frac{1}{1 + \exp(a_i^\top x)} \right) \\ &= \sum_{i=1}^{k} a_i^\top x - \sum_{i=1}^{m} \log \left(1 + \exp(a_i^\top x) \right) \end{aligned}$$

This is concave.

Linear Discrimination

Consider two sets of points

$$\{x_1, x_2, \ldots, x_N\} \subset \mathbb{R}^n \qquad \{y_1, y_2, \ldots, y_M\} \subset \mathbb{R}^n$$

We wish to separate these two points with a hyperplane, or, find $a \in \mathbb{R}^n$ and $b \in \mathbb{R}$ so that:

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

By normalizing, this is equivalent to

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

This is a set of linear equalities in a and b.

Robust Linear Discrimination

Consider the two hyperplanes:

$$\mathcal{H}_1 = \{z \in \mathbb{R}^n \ : \ a^\top z + b = 1\} \qquad \mathcal{H}_2 = \{z \in \mathbb{R}^n \ : \ a^\top z + b = -1\}$$

What is the distance (or, margin) between $\mathcal{H}_1$ and $\mathcal{H}_2$?

$$\frac{2}{\|a\|_2}$$

To find the maximum margin classifier, solve:

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}\|a\|_2 \\ \text{subject to} & a^\top x_i + b \geq 1, \quad i = 1, \ldots, N \\ & a^\top y_i + b \leq -1, \quad i = 1, \ldots, M \\ & a \in \mathbb{R}^n, \ b \in \mathbb{R} \end{array}$$

Robust Linear Discrimination

Dual:

$$\begin{array}{ll} \text{maximize} & \mathbf{1}^\top \lambda + \mathbf{1}^\top \mu \\ \text{subject to} & \left\| \sum_{i=1}^N \lambda_i x_i - \sum_{i=1}^M \mu_i y_i \right\|_2 \leq \frac{1}{2} \\ & \mathbf{1}^\top \lambda = \mathbf{1}^\top \mu \\ & \lambda \geq 0,\ \mu \geq 0,\ \lambda \in \mathbb{R}^N,\ \mu \in \mathbb{R}^M \end{array}$$

Interpretation:

- Change variables: $\theta \triangleq \lambda / \mathbf{1}^\top \lambda$, $\gamma \triangleq \mu / \mathbf{1}^\top \mu$, $t \triangleq 1/(\mathbf{1}^\top \lambda + \mathbf{1}^\top \mu)$

$$\begin{array}{ll} \text{minimize} & t \\ \text{subject to} & \left\| \sum_{i=1}^N \theta_i x_i - \sum_{i=1}^M \gamma_i y_i \right\|_2 \leq t \\ & \theta \geq 0,\ \mathbf{1}^\top \theta = 1,\ \gamma \geq 0,\ \mathbf{1}^\top \gamma = 1 \\ & \theta \in \mathbb{R}^N,\ \gamma \in \mathbb{R}^M \end{array}$$

The optimal value is the distance between convex hulls!

Robust Linear Discrimination

QP Form:

$$\begin{array}{lll} \text{minimize} & \frac{1}{2}\|a\|_2^2 & \\ \text{subject to} & a^\top x_i + b \geq 1, & i = 1, \ldots, N \\ & a^\top y_i + b \leq -1, & i = 1, \ldots, M \\ & a \in \mathbb{R}^n,\ b \in \mathbb{R} & \end{array}$$

Dual:

$$\begin{array}{ll} \text{maximize} & \mathbf{1}^\top \lambda + \mathbf{1}^\top \mu - \frac{1}{2} \left\| \sum_{i=1}^N \lambda_i x_i - \sum_{i=1}^M \mu_i y_i \right\|_2^2 \\ \text{subject to} & \mathbf{1}^\top \lambda = \mathbf{1}^\top \mu \\ & \lambda \geq 0,\ \mu \geq 0,\ \lambda \in \mathbb{R}^N,\ \mu \in \mathbb{R}^M \end{array}$$

Robust Linear Discrimination

Dual:

$$\begin{array}{ll} \text{maximize} & \mathbf{1}^\top\lambda + \mathbf{1}^\top\mu - \frac{1}{2}\left\|\sum_{i=1}^N \lambda_i x_i - \sum_{i=1}^M \mu_i y_i\right\|_2^2 \\ \text{subject to} & \mathbf{1}^\top\lambda = \mathbf{1}^\top\mu \\ & \mu \geq 0,\ \lambda \geq 0 \\ & \mu \in \mathbb{R}^N,\ \lambda \in \mathbb{R}^M \end{array}$$

- Primal optimum can be constructed from dual optimum by

$$a = \sum_{i=1}^N \lambda_i x_i - \sum_{i=1}^M \mu_i y_i$$

  $\Rightarrow$ Linear combination of support vectors on the margin

- Constructing the dual requires only the inner products

$$\{x_i^\top x_j,\ x_i^\top y_j,\ y_i^\top y_j\}$$

  In many cases, $M, N \ll n$. We can solve the dual so long as $(N+M)^2$ is not too big, even if $n = \infty$!

Approximate Linear Separation

$$\begin{array}{lll} \text{minimize} & \mathbf{1}^\top u + \mathbf{1}^\top v & \\ \text{subject to} & a^\top x_i + b \geq 1 - u_i, & i = 1, \ldots, N \\ & a^\top y_i + b \leq -1 + v_i, & i = 1, \ldots, M \\ & u \geq 0,\ v \geq 0,\ u \in \mathbb{R}^N,\ v \in \mathbb{R}^M & \\ & a \in \mathbb{R}^n,\ b \in \mathbb{R} & \end{array}$$

- Introduced slack variables u, v (soft constraints)
- Linear program in (a, b, u, v)
- At optimum,

$$u_i = \max_i\{1 - a^\top x_i - b, 0\}, \quad v_i = \max_i\{1 + a^\top y_i + b, 0\}$$

- Heuristic to minimize number of misclassifications
- Other penalty functions possible, but be careful to maintain convexity!

Support Vector Classifier

$$
\begin{array}{lll}
\text{minimize} & \frac{1}{2}\|a\|_2^2 + C\left(\mathbf{1}^\top u + \mathbf{1}^\top v\right) & \\
\text{subject to} & a^\top x_i + b \geq 1 - u_i, & i = 1, \ldots, N \\
& a^\top y_i + b \leq -1 + v_i, & i = 1, \ldots, M \\
& u \geq 0,\ v \geq 0,\ u \in \mathbb{R}^N,\ v \in \mathbb{R}^M & \\
& a \in \mathbb{R}^n,\ b \in \mathbb{R} &
\end{array}
$$

- Parameter $C \geq 0$ controls trade-off between maximizing margin and minimizing misclassifications
- How can we this solve efficiently if $n \gg M, N$?

Dual:

$$
\begin{array}{lll}
\text{maximize} & \mathbf{1}^\top\lambda + \mathbf{1}^\top\mu - \frac{1}{2}\left\|\sum_{i=1}^N \lambda_i x_i - \sum_{i=1}^M \mu_i y_i\right\|_2^2 & \\
\text{subject to} & \mathbf{1}^\top\lambda = \mathbf{1}^\top\mu & \\
& 0 \leq \mu \leq C\mathbf{1},\ 0 \leq \lambda \leq C\mathbf{1} & \\
& \mu \in \mathbb{R}^N,\ \lambda \in \mathbb{R}^M &
\end{array}
$$

Non-Linear Features

Separate two sets of points by a non-linear function

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

- Choose a linearly parameterized family of functions

$$f(z) \triangleq \theta^\top F(z)$$

Here,

$$F = (F_1, \ldots, F_k)\colon \mathbb{R}^n \to \mathbb{R}^k$$

are basis functions

- Solve the linear inequalities in θ

$$\theta^\top F(x_i) \geq 1, \quad i = 1, \ldots, N \qquad \theta^\top F(y_i) \leq -1, \quad i = 1, \ldots, M$$

B9824 Foundations of Optimization

Lecture 9: Vector Space Optimization I

Fall 2011

Outline

1. Vector spaces, Banach spaces
2. Weierstrass' theorem
3. Inner product spaces, Hilbert spaces
4. Projection theorem
5. Linear functionals
6. Dual spaces

Vector Spaces

Definition. A vector space is a set $\mathcal{X}$ equipped with the operations of

- **addition:** $x, y \in \mathcal{X} \quad \Rightarrow \quad x + y \in \mathcal{X}$
- **scalar multiplication:** $x \in \mathcal{X}$, $\alpha \in \mathbb{R} \quad \Rightarrow \quad \alpha x \in \mathcal{X}$

that satisfies the axioms

1. $x + y = y + x$ (commutative)
2. $(x + y) + z = x + (y + x)$ (associative)
3. there exists $\mathbf{0} \in \mathcal{X}$ with $x + \mathbf{0} = x$
4. $\alpha(x + y) = \alpha x + \alpha y$ (distributive)
5. $(\alpha + \beta)x = \alpha x + \beta x$ (distributive)
6. $(\alpha\beta)x = \alpha(\beta x)$ (associative)
7. $0x = \mathbf{0}, \quad 1x = x$

Examples

Example. $\mathcal{X} = \mathbb{R}^n$

Example. $\mathcal{X} = \mathbb{R}^\infty$, the set of (countably) infinite sequences of real numbers

Example. $\mathcal{X} = \{x \in \mathbb{R}^\infty \ : \ x \text{ has finitely many non-zero terms}\}$

Example. $\mathcal{X} = c_0 \triangleq \{x \in \mathbb{R}^\infty \ : \ \lim_{n\to\infty} x_n = 0\}$

Example. $\mathcal{X} = C[a, b] \triangleq \{\text{continuous functions from } [a, b] \text{ to } \mathbb{R}\}$

Subspaces, etc.

Definition. The set $\mathcal{C} \subset \mathcal{X}$ is a subspace if, for all points $x, y \in \mathcal{C}$, and scalars $\alpha, \beta \in \mathbb{R}$,

$$\alpha x + \beta y \in \mathcal{C}$$

Definition. The set $\mathcal{C} \subset \mathcal{X}$ is affine (linear variety) if, for all points $x, y \in \mathcal{C}$, and scalars $\lambda \in \mathbb{R}$,

$$\lambda x + (1 - \lambda) y \in \mathcal{C}$$

Definition. The set $\mathcal{C} \subset \mathcal{X}$ is convex if, for all points $x, y \in \mathcal{C}$, and scalars $0 \leq \lambda \leq 1$,

$$\lambda x + (1 - \lambda) y \in \mathcal{C}$$

Definition. The set $\mathcal{C} \subset \mathcal{X}$ is a cone if, for all points $x \in \mathcal{C}$, and scalars $\lambda \geq 0$,

$$\lambda x \in \mathcal{C}$$

Linear Independence, Dimension

Definition. A linear combination of the (finite) collection of vectors $\{x_1, \ldots, x_n\}$ is a sum of the form

$$\alpha_1 x_1 + \cdots + \alpha_n x_n$$

The subspace generated by the set $\mathcal{S} \subset \mathcal{X}$ is the set of all linear combinations of vectors in $\mathcal{S}$.

Definition. A vector $x \in \mathcal{X}$ is linearly independent from a set $\mathcal{S} \subset \mathcal{X}$ if it cannot be expressed as a linear combination of elements of $\mathcal{S}$.

Definition. A set of vectors $\mathcal{S} \subset \mathcal{X}$ is linear independent if, for each $x \in \mathcal{S}$, x is linearly independent from $\mathcal{S} \setminus \{x\}$.

Definition. A linearly independent, finite set $\mathcal{S} \subset \mathcal{X}$ is said to be a (finite) basis if $\mathcal{S}$ generates $\mathcal{X}$. If a vector space has a finite basis, it is called finite dimensional, otherwise it is called infinite dimensional.

Normed Linear Spaces

Definition. A normed vector space is a vector space $\mathcal{X}$ associated with a real-valued function $\|\cdot\|$ on $\mathcal{X}$ such that

1. $\|x\| = 0$ if and only if $x = \mathbf{0}$
2. $\|\alpha x\| = |\alpha|\|x\|$
3. $\|x + y\| \leq \|x\| + \|y\|$

Examples

Example. $C[a, b]$, $\|x\| \triangleq \max_{a \leq t \leq b} |x(t)|$

Example.
$D[a, b] \triangleq \{$continuously differentiable functions from $[a, b]$ to $\mathbb{R}\}$

$$\|x\| \triangleq \max_{a \leq t \leq b} |x(t)| + \max_{a \leq t \leq b} |\dot{x}(t)|$$

Example. finitely non-zero sequences, $\|x\| \triangleq \sum_{i=1}^{\infty} |x_i|$

Example. real-valued continuous functions over $[a, b]$,
$\|x\| \triangleq \int_a^b |x(t)|\, dt$

Classical Normed Spaces

Definition. For $1 \leq p \leq \infty$, the normed linear space ℓ_p is the space of all sequences $x \in \mathbb{R}^\infty$ with

$$\sum_{i=1}^{\infty} |x_i|^p < \infty, \quad \text{if } p < \infty$$

$$\sup_i |x_i| < \infty, \quad \text{if } p = \infty$$

with norm

$$\|x\|_p \triangleq \begin{cases} \left(\sum_{i=1}^{\infty} |x_i|^p \right)^{1/p} & \text{if } p < \infty \\ \sup_i |x_i| & \text{if } p = \infty \end{cases}$$

Classical Normed Spaces

Definition. For $1 \leq p < \infty$, the normed linear space $L_p[a, b]$ is the space of all measurable, real-valued functions $x \colon [a, b] \to \mathbb{R}$ where $|x(t)|^p$ is Lebesgue integrable, with norm

$$\|x\|_p \triangleq \left(\int_a^b |x(t)|^p \, dt \right)^{1/p}$$

Note: Functions in $L_p[a, b]$ are considered equal if the differ only on a set of measure zero. For example, $\|x\|_p = 0$ implies that $x(t) = 0$ except possibly on a set of measure zero, we identify all such functions with the element $\mathbf{0} \in L_p[a, b]$.

Definition. The normed linear space $L_\infty[a, b]$ is the space of all measurable, real-valued functions $x\colon [a, b] \to \mathbb{R}$ that are bounded except possibly on a set of measure zero, with norm

$$\begin{aligned}\|x\|_\infty &\triangleq \inf_{y(t)=x(t) \text{ a.e.}} \sup_{a\le t\le b} |y(t)| \\ &= \operatorname*{ess\,sup}_{a\le t\le b} |x(t)|\end{aligned}$$

Basic Topology

Consider a normed linear space $\mathcal{X}$.

Definition. An open ball (or, "neighborhood") around a point $x \in \mathcal{X}$ with radius $r > 0$ is the set

$$N_r(x) \triangleq \{y \in \mathcal{X} \ : \ \|x - y\| < r\}$$

Open and Closed Sets

Consider a set $\mathcal{E} \subset \mathcal{X}$.

Definition. A point $x \in \mathcal{E}$ is an interior point if there exists an open ball $N_r(x)$ such that $N_r(x) \subset \mathcal{E}$. The interior **int** $\mathcal{E}$ is defined to be the set of all interior points of $\mathcal{E}$.

Definition. $\mathcal{E}$ is open if $\mathcal{E} =$ **int** $\mathcal{E}$.

Definition. A point $x \in \mathcal{X}$ is a closure point of $\mathcal{E}$ if, for every open ball $N_r(x)$, there exists $y \in \mathcal{E}$ with $y \in N_r(x)$. The closure **cl** $\mathcal{E}$ is defined to be the set of all closure points of $\mathcal{E}$.

Definition. $\mathcal{E}$ is closed if every closure point if $\mathcal{E} =$ **cl** $\mathcal{E}$.

Convergence

Definition. A sequence of vectors $\{x_k\} \subset \mathcal{X}$ converges to a limit $x \in \mathcal{X}$ if

$$\lim_{k \to \infty} \|x - x_k\| = 0$$

We say $x_k \to x$.

Compactness

Consider a set $\mathcal{E} \subset \mathcal{X}$.

Definition. $\mathcal{E}$ is compact if, given a sequence $\{x_k\} \subset \mathcal{E}$, there is a subsequence $\{x_{k_i}\}$ converging to an element $x \in \mathcal{E}$.

Note: In infinite dimensional vector spaces, compactness is not equivalent to being closed and bounded!

Transformations

Definition. If $\mathcal{X}$ and $\mathcal{Y}$ are two vector spaces, a transformation is a function $T : \mathcal{X} \to \mathcal{Y}$.

Definition. If $\mathcal{X}$ is a vector space, a functional is a map $f : \mathcal{X} \to \mathbb{R}$.

Definition. A transformation $T : \mathcal{X} \to \mathcal{Y}$ between two normed vector spaces is continuous at the point $x \in \mathcal{X}$ if, for every sequence $\{x_k\} \subset \mathcal{X}$ with $x_k \to x$,

$$T(x_k) \to T(x)$$

We say $T(\cdot)$ is continuous if it is continuous at all points of $\mathcal{X}$.

Weierstrass' Theorem

Theorem. (Weierstrass) Let $\mathcal{X}$ be a normed linear space, and $\mathcal{C} \subset \mathcal{X}$ a non-empty, compact set. If $f : \mathcal{C} \to \mathbb{R}$ is a continuous function, then the optimization program

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in \mathcal{C} \end{array}$$

has a globally optimal solution.

Weierstrass' Theorem: Proof

Since $\mathcal{C}$ is non-empty, set

$$M \triangleq \inf_{x \in \mathcal{C}} f(x) \in [-\infty, \infty)$$

Then, there exists a sequence $x_k \subset \mathcal{C}$ with $f(x_k) \to M$.

By compactness, there must exist a convergent subsequence $\{x_{k_i}\}$, with $x_{k_i} \to x \in \mathcal{C}$. By continuity, we must have $f(x) = M$.

Thus, $M > -\infty$, and the global optimum is achieved by x.

Banach Spaces

Let $\mathcal{X}$ be a normed linear space.

Definition. A sequence $\{x_n\} \subset \mathcal{X}$ is a Cauchy sequence if, given $\epsilon > 0$, there is an integer N such that

$$\|x_n - x_m\| < \epsilon$$

for all $m, n > N$.

Definition. A normed linear space $\mathcal{X}$ is complete if every Cauchy sequence converges. Such a space is known as a Banach space.

Examples

Example. $C[a, b]$ (with sup-norm) is a Banach space

Example. ℓ_p, $1 \leq p \leq \infty$, is a Banach space

Example. $L_p[a, b]$, $1 \leq p \leq \infty$, is a Banach space

Inner Product Spaces

Definition. A (real) inner product space $\mathcal{X}$ is a vector space together with an inner product $\langle \cdot, \cdot \rangle : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ that satisfies the axioms

1. $\langle x, y \rangle = \langle y, x \rangle$ (symmetry)
2. $\langle x + y, z \rangle = \langle x, z \rangle + \langle y, z \rangle$ (linearity)
3. $\langle \alpha x, y \rangle = \alpha \langle x, y \rangle$ (linearity)
4. $\langle x, x \rangle \geq 0$ and $\langle x, x \rangle = 0$ if and only if $x = \mathbf{0}$ (positive definiteness)

Given an inner product space $\mathcal{X}$, the norm induced by the inner product is

$$\|x\| \triangleq \sqrt{\langle x, x \rangle}$$

Inner Product Spaces

Let $\mathcal{X}$ be an inner product space.

Lemma. (Cauchy-Schwartz Inequality) For all $x, y \in \mathcal{X}$,

$$|\langle x, y \rangle| \leq \|x\| \|y\|$$

with equality if and only if $x = \lambda y$ or $y = \mathbf{0}$.

Proof. If $y = \mathbf{0}$, the result is clear. If $y \neq \mathbf{0}$,

$$0 \leq \langle x - \lambda y, x - \lambda y \rangle = \langle x, x \rangle - 2\lambda \langle x, y \rangle + \lambda^2 \langle y, y \rangle$$

Set $\lambda = \langle x, y \rangle / \langle y, y \rangle$, then

$$0 \leq \langle x, x \rangle - \langle x, y \rangle^2 / \langle y, y \rangle$$

□

An inner product can be thought of as defining an 'angle' θ between non-zero $x, y \in \mathcal{X}$ by

$$\cos \theta = \frac{\langle x, y \rangle}{\|x\| \|y\|}$$

Theorem. An inner product space equipped with the induced norm is a normed linear space.

Proof.

$$\begin{aligned}\|x+y\|^2 = \langle x+y, x+y\rangle &= \|x\|^2 + 2\langle x, y\rangle + \|y^2\| \\ &\leq \|x\|^2 + 2\|x\|\|y\| + \|y\|^2 = (\|x\| + \|y\|)^2\end{aligned}$$

□

Examples

Example. $\mathbb{R}^n$, $\langle x, y\rangle = x^\top y$, $\|x\| =$ Euclidean norm

Example. ℓ_2, $\langle x, y\rangle = \sum_{i=1}^{\infty} x_i y_i$, $\|x\| = \ell_2$-norm

Example. $L_2[a, b]$, $\langle x, y\rangle = \int_a^b x(t)y(t)\,dt$, $\|x\| = L_2$-norm

Definition. A complete inner product space is called a Hilbert space.

$\mathbb{R}^n$, ℓ_2, and $L_2[a, b]$ are all Hilbert spaces

Projection Theorem

Let $\mathcal{X}$ be a Hilbert space, and $\mathcal{C} \subset \mathcal{X}$ a closed and non-empty convex set. Fix the vector $x \in \mathcal{X}$.

$$\begin{array}{ll} \text{minimize} & \|z - x\| \\ \text{subject to} & z \in \mathcal{C} \end{array}$$

Theorem. For every $x \in \mathcal{X}$, the optimization problem has a unique global minimum x^* called the projection of x onto $\mathcal{C}$. A vector $x' \in \mathcal{C}$ is equal to x^* if and only if

$$\langle x - x', z - x' \rangle \leq 0, \quad \forall\ z \in \mathcal{C}$$

Projection Theorem: Proof

To prove existence, let $\{z_i\} \subset \mathcal{C}$ be a sequence with

$$\|z_i - x\| \to \delta \triangleq \inf_{z \in \mathcal{C}} \|z - x\|$$

Note that for all $w, v \in \mathcal{X}$,

$$2\|w\|^2 + 2\|v\|^2 = \|w + v\|^2 + \|w - v\|^2 \quad \text{(Parallelogram law)}$$

Then,

$$\|z_i - z_j\|^2 = 2\|z_i - x\|^2 + 2\|z_j - x\|^2 - 4\left\|x - \frac{z_i + z_j}{2}\right\|^2$$

Since $\mathcal{C}$ is convex,

$$\left\|x - \frac{z_i + z_j}{2}\right\| \geq \delta$$

Thus

$$\|z_i - z_j\|^2 \leq 2\|z_i - x\|^2 + 2\|z_j - x\|^2 - 4\delta^2 \to 0$$

Then, $\{z_i\}$ is a Cauchy sequence, thus $z_i \to x^* \in \mathcal{C}$ and $\|x^* - x\| = \delta$ (continuity).

Projection Theorem: Proof

To prove uniqueness, let $\tilde{x}^* \in \mathcal{C}$ be a point with $\|\tilde{x}^* - x\| = \delta$. Define

$$z_i = \begin{cases} x^* & i \text{ odd} \\ \tilde{x}^* & i \text{ even} \end{cases}$$

Clear $\|z_i - x\| \to \delta$, so by the same argument as before, $\{z_i\}$ is a Cauchy sequence and convergent. Then, $\tilde{x}^* = x^*$.

Projection Theorem: Proof

We wish to show that

$$\langle x - x^*, z - x^* \rangle \leq 0, \quad \forall\ z \in \mathcal{C}$$

Suppose there is some z_1 with

$$\langle x - x^*, z_1 - x^* \rangle = \epsilon > 0$$

Define

$$z(\alpha) \triangleq (1-\alpha)x^* + \alpha z_1, \quad 0 \leq \alpha \leq 1$$

Then,

$$\|x - z(\alpha)\|^2 = (1-\alpha)^2\|x - x^*\|^2 + 2\alpha(1-\alpha)\langle x - x^*, x - z_1\rangle + \alpha^2\|x - z_1\|^2$$

This is a differentiable function of α, and

$$\frac{d}{d\alpha}\|x - z(\alpha)\|^2\Big|_{\alpha=0} = -2\langle x - x^*, z_1 - z^*\rangle = -2\epsilon < 0$$

This contradicts the optimality of x^*.

Projection Theorem: Proof

Conversely, suppose that there exists some $x' \in \mathcal{C}$ with

$$\langle x - x', z - x' \rangle \leq 0, \quad \forall\, z \in \mathcal{C}$$

Then, if $z \in \mathcal{C}$, $z \neq x'$,

$$\begin{aligned} \|x - z\|^2 &= \|x - x' + x' - z\|^2 \\ &= \|x - x'\|^2 + 2\langle x - x', x' - z \rangle + \|x' - z\|^2 \\ &> \|x - x'\|^2 \end{aligned}$$

Thus, x' is the unique optimizer.

Linear Functionals

Definition. If $\mathcal{X}$ is a vector space, a linear functional is a functional $f\colon \mathcal{X} \to \mathbb{R}$ such that

$$f(\alpha x + \beta y) = \alpha f(x) + \beta f(y)$$

Example. On $\mathbb{R}^n$, for every $y \in \mathbb{R}^n$, the function

$$f(x) = y^\top x$$

is a linear functional. Moreover, all linear functionals are of this form.

Example. On $C[0,1]$, the functional

$$f(x) = x(1/2)$$

is a linear functional.

Example. On $L_2[0,1]$, for every $y \in L_2[0,1]$, the functional

$$f(x) = \int_0^1 x(t)y(t)\, dt$$

is a linear functional.

Linear Functionals

Let $\mathcal{X}$ be a normed linear space.

Theorem. If a linear functional is continuous at a point in $\mathcal{X}$, it is continuous over all of $\mathcal{X}$.

Proof. Suppose f is linear and continuous at y, and $x_n \to x$. Then, $x_n + y - x \to y$. Thus,

$$|f(x_n) - f(x)| = |f(x_n + y - x) - f(y)| \to 0$$

by the linearity and continuity at y of f. □

Most commonly, we check that a linear functional is continuous just at $\mathbf{0}$.

A Discontinuous Linear Functional

Example.

$$\mathcal{X} = \{x \in \mathbb{R}^\infty \ : \ x \text{ has finitely many non-zero components}\}$$

$$\|x\| = \max_i |x_i|$$

Consider the linear functional

$$f(x) = \sum_{\ell=1}^{\infty} \ell x_\ell$$

Define $x^{(k)}$ to have $1/\sqrt{k}$ in the kth component, zero everywhere else.

$$\|x^{(k)} - \mathbf{0}\| = 1/\sqrt{k} \to 0$$

$$f(x^{(k)}) = \sqrt{k} \to \infty \neq f(\mathbf{0})$$

Linear Functionals

Let $\mathcal{X}$ be a normed linear space.

Definition. A linear functional f is bounded if there is a constant M such that

$$|f(x)| \leq M\|x\|, \quad \forall\, x \in \mathcal{X}$$

Theorem. A linear functional is continuous if and only if it is bounded.

Proof. Suppose f is a bounded linear functional, with $|f(x)| \leq M\|x\|$. Then, if $x_n \to \mathbf{0}$, $|f(x_n)| \leq M\|x_n\| \to 0$. Thus, f is continuous.

Conversely, assume that f is continuous. Then, there exists a $\delta > 0$ such that $|f(x)| < 1$ for $\|x\| \leq \delta$. Thus, if $x \neq 0$,

$$|f(x)| = \left|f\left(\frac{\delta x}{\|x\|}\right)\right| \frac{\|x\|}{\delta} < \frac{\|x\|}{\delta},$$

and $1/\delta$ is a bound for f. □

Dual Spaces

Let $\mathcal{X}$ be a normed linear space. We define the normed dual space $\mathcal{X}^*$ to be the space of bounded linear functionals on $\mathcal{X}$, equipped with

- addition: $(f_1 + f_2)(x) \triangleq f_1(x) + f_2(x)$
- scalar multiplication: $(\alpha f)(x) \triangleq \alpha f(x)$
- zero element: $\mathbf{0}(x) \triangleq 0$
- norm:

$$\begin{aligned}\|f\|_* &\triangleq \inf\{M \;:\; |f(x)| \leq M\|x\|,\ \forall\, x \in \mathcal{X}\} \\ &= \sup_{x \neq \mathbf{0}} \frac{|f(x)|}{\|x\|} = \sup_{\|x\| \leq 1} |f(x)| = \sup_{\|x\| = 1} |f(x)|\end{aligned}$$

Given a bounded linear functional $x^* \in \mathcal{X}^*$, we will abuse notation to write

$$\langle x, x^* \rangle \triangleq x^*(x)$$

Theorem. If $\mathcal{X}$ is a normed linear space, the dual space $\mathcal{X}^*$ is a Banach space.

Proof. Clearly $\mathcal{X}^*$ is a normed linear space. We need to show that it is complete.
Given a Cauchy sequence $\{x_n^*\}$ and $x \in \mathcal{X}$, note that

$$|x_n^*(x) - x_m^*(x)| \leq \|x_n^* - x_m^*\|\|x\|$$

Then, $\{x_n^*(x)\}$ is a Cauchy sequence. Define x^* point-wise by $x^*(x) = \lim\ x_n^*(x)$. It is easy to verify that x^* is a linear operator, it is bounded, and $\|x^* - x_n^*\| \to 0$. □

B9824 Foundations of Optimization

Lecture 10: Vector Space Optimization II

Fall 2011

Outline

1. Dual spaces
2. Hahn-Banach Theorem

Some Common Duals

Example. $\mathcal{X} = \mathbb{R}^n$, $\|x\| = \|x\|_2 = \sqrt{x^\top x}$
$\Rightarrow \quad \mathcal{X}^* = \mathbb{R}^n, \ \|x^*\|_* = \|x^*\|_2$

This space is self-dual.

Theorem. (Riesz-Fréchet) If $\mathcal{X}$ is a Hilbert space, then $\mathcal{X}^* = \mathcal{X}$.

Example. $\mathcal{X} = \ell_p$, $1 \leq p < \infty$
$\Rightarrow \quad \mathcal{X}^* = \ell_q$, where $1/p + 1/q = 1$ (if $p = 1$, $q = \infty$)

Note: The dual to ℓ_∞ is not ℓ_1!

Example. $\mathcal{X} = L_p[a, b]$, $1 \leq p < \infty$
$\Rightarrow \quad \mathcal{X}^* = L_q[a, b]$, where $1/p + 1/q = 1$ (if $p = 1$, $q = \infty$)

Note: The dual to $L_\infty[a, b]$ is not $L_1[a, b]$!

Functions of Bounded Variation

Definition. Given a function $x : [a, b] \to \mathbb{R}$, define the total variation to be

$$\mathsf{TV}(x) \triangleq \sup \sum_{i=1}^{n} |x(t_i) - x(t_{i-1})|$$

where the supremum is taken over all partitions

$$a = t_0 \leq t_1 \leq \cdots \leq t_n = b$$

of $[a, b]$. It is often written as

$$\mathsf{TV}(x) = \int_a^b |dx(t)|$$

Definition. $\mathrm{BV}[a, b]$ is the space of functions on $[a, b]$ of bounded total variation with norm

$$\|x\| = |x(a)| + \mathsf{TV}(x)$$

The Dual of $C[a, b]$

Theorem. (Riesz Representation Theorem) Suppose that f is a bounded linear functional on $C[a, b]$. Then, there is a function $v \in \mathrm{BV}[a, b]$ such that for all $x \in C[a, b]$

$$f(x) = \int_a^b x(t)\, dv(t)$$

Note: The representation v of a linear functional f is not unique. To remove this ambiguity, define $\mathrm{NBV}[a, b]$ to be the set of functions $x \in \mathrm{BV}[a, b]$ with $x(a) = 0$ that are right continuous on (a, b). Then, $C[a, b]^* = \mathrm{NBV}[a, b]$.

Hahn-Banach Theorem: Extension Form

Definition. A sublinear functional is a map $p : \mathcal{X} \to \mathbb{R}$ such that

1. $p(x + y) \leq p(x) + p(y), \quad \forall\, x, y \in \mathcal{X}$
2. $p(\alpha x) = \alpha p(x), \quad \forall\, x \in \mathcal{X},\ \alpha \geq 0$

Theorem. (Hahn-Banach) Let $\mathcal{X}$ be a normed linear space and p a continuous, sublinear functional on $\mathcal{X}$. Suppose $\mathcal{M} \subset \mathcal{X}$ is a subspace, and f is a linear functional on $\mathcal{M}$, with

$$f(m) \leq p(m), \quad \forall\, m \in \mathcal{M}$$

Then, there is a linear functional F which is an extension of f from $\mathcal{M}$ to $\mathcal{X}$ such that

$$F(x) \leq p(x), \quad \forall\, x \in \mathcal{X}$$

Hahn-Banach Theorem: Proof Sketch

We describe the "induction" step: suppose $y \in \mathcal{X} \setminus \mathcal{M}$, we will extend f to the subspace

$$\mathcal{M}' = \{m + \alpha y \ : \ m \in \mathcal{M}, \ \alpha \in \mathbb{R}\}$$

Given $m_1, m_2 \in \mathcal{M}$,

$$f(m_1) + f(m_2) = f(m_1 + m_2) \leq p(m_1 + m_2) \leq p(m_1 - y) + p(m_2 + y)$$

$$\Rightarrow \quad f(m_1) - p(m_1 - y) \leq p(m_2 + y) - f(m_2)$$

$$\Rightarrow \quad c \triangleq \sup_{m \in \mathcal{M}} f(m) - p(m - y) \leq \inf_{m \in \mathcal{M}} p(m + y) - f(m)$$

Hahn-Banach Theorem: Proof Sketch

If $x \in \mathcal{M}'$, the x can be uniquely written as $m + \alpha y$, and we define the extension

$$g(x) = f(m) + \alpha c$$

Clearly g is linear. We would like to show that

$$g(m + \alpha y) \leq p(m + \alpha y)$$

If $\alpha > 0$,

$$\begin{aligned} f(m) + \alpha c &= \alpha \left[c + f\left(\frac{m}{\alpha}\right) \right] \leq \alpha \left[p\left(\frac{m}{\alpha} + y\right) - f\left(\frac{m}{\alpha}\right) + f\left(\frac{m}{\alpha}\right) \right] \\ &= p(m + \alpha y) \end{aligned}$$

$\alpha < 0$ is handled similarly.

Corollary. Let f be a bounded linear functional on a subspace $\mathcal{M}$ of a normed linear space $\mathcal{X}$. There is an extension F which is a bounded linear functional defined on $\mathcal{X}$ with $\|F\|_* = \|f\|_{\mathcal{M},*}$.

Proof. Take $p(x) = \|f\|_{\mathcal{M},*}\|x\|$. □

Corollary. Let x be an element of a normed linear space $\mathcal{X}$. There exists a bounded non-zero linear functional F with $F(x) = \|F\|_*\|x\|$.

Proof. Suppose $x \neq 0$. On the subspace defined by x, define $f(\alpha x) = \alpha\|x\|$. Clearly $\|f\| = 1$. Extend f to the entire space.

If $x = 0$, any non-zero bounded linear functional will do, and we have just proved that one exists. □

Optimization Interpretation

Given a normed linear space $\mathcal{X}$, a subspace $\mathcal{M}$, and a bounded linear functional f on $\mathcal{M}$, consider the optimization problem

$$\begin{array}{ll} \text{minimize} & \|x^*\|_* \\ \text{subject to} & \langle x, x^* \rangle = f(x), \quad \forall\, x \in \mathcal{M} \\ & x^* \in \mathcal{X}^* \end{array}$$

The Hahn-Banach theorem implies that a global optimum exists, and that the optimal value is $\|f\|_{\mathcal{M},*}$.

B9824 Foundations of Optimization

Lecture 11: Vector Space Optimization III

Fall 2011

Outline

1. Minimum norm duality
2. Applications
3. Geometric Hahn-Banach Theorem

Orthogonality and Alignment

Definition. A vector $x \in \mathcal{X}$ and a functional $x^* \in \mathcal{X}^*$ are aligned if

$$\langle x, x^* \rangle = \|x^*\|_* \|x\|$$

Definition. A vector $x \in \mathcal{X}$ and a functional $x^* \in \mathcal{X}^*$ are orthogonal if

$$\langle x, x^* \rangle = 0$$

Definition. The orthogonal complement of a set of vectors $\mathcal{M} \subset \mathcal{X}$ is the set of functionals

$$\mathcal{M}^\perp = \{x^* \in \mathcal{X}^* \ : \ \langle x, x^* \rangle = 0, \ \forall \ x \in \mathcal{M}\}$$

Alignment

Example. $x \in \ell_2$, $y \in \ell_2$

From the Cauchy-Schwartz inequality, x and y are aligned if $x_i = \lambda y_i$ for some $\lambda \geq 0$ and all i.

Example. $x \in \ell_p$, $1 < p < \infty$, $y \in \ell_q$, $1/p + 1/q = 1$

From the Hölder inequality, x and y are aligned if

$$x_i = \lambda (\text{sgn } y_i) |y_i|^{q/p}$$

for some $\lambda \geq 0$ and all i.

Example. $x \in L_p[a, b]$, $1 < p < \infty$, $y \in L_q[a, b]$, $1/p + 1/q = 1$

From the Hölder inequality, x and y are aligned if

$$x(t) = \lambda (\text{sgn } y(t)) |y(t)|^{q/p}$$

for some $\lambda \geq 0$ and (almost) all $t \in [a, b]$.

Alignment

Example. $x \in C[a,b]$, $y \in \mathrm{NBV}[a,b]$

Let $\Gamma \subset [a,b]$ be the set of points t at which $|x(t)| = \|x\|$, this must be non-empty. x and y are aligned if and only if y varies only on Γ, y is non-decreasing at t if $x(t) > 0$, and $y(t)$ is non-increasing when $x(t) < 0$.

Minimum Norm Duality

Theorem. Suppose $\mathcal{X}$ is a normed linear space, $\mathcal{M} \subset \mathcal{X}$ is a subspace, and $x \in \mathcal{X}$. Then,

$$\inf_{m \in \mathcal{M}} \|x - m\| = \max_{\substack{\|x^*\|_* \leq 1 \\ x^* \in \mathcal{M}^\perp}} \langle x, x^* \rangle$$

where the maximum is achieved for some $x_0^* \in \mathcal{M}^\perp$.

If $m_0 \in \mathcal{M}$ is achieves the minimum, then x_0^* is aligned with $x - m_0$.

Further, if $m_0 \in \mathcal{M}$ and there is a non-zero vector $x^* \in \mathcal{M}^\perp$ aligned with $x - m_0$, then m_0 achieves the minimum.

Minimum Norm Duality: Proof

Set $d \triangleq \inf_{m \in \mathcal{M}} \|x - m\|$.

For any $\epsilon > 0$, suppose $m_\epsilon \in \mathcal{M}$ with $\|x - m_\epsilon\| \leq d + \epsilon$. If $x^* \in \mathcal{X}^*$ is feasible for the dual,

$$\langle x, x^* \rangle = \langle x - m_\epsilon, x^* \rangle \leq \|x^*\|_* \|x - m_\epsilon\| \leq d + \epsilon$$

Since ϵ was arbitrary, $\langle x, x^* \rangle \leq d$.

Let $\mathcal{N}$ be the subspace spanned by $\mathcal{M}$ and x. Define f on $\mathcal{N}$ by

$$f(\alpha x + m) = \alpha d, \quad \forall\, \alpha \in \mathbb{R},\ m \in \mathcal{M}$$

Then,

$$\|f\|_{\mathcal{N},*} = \sup_{(\alpha,m)} \frac{|\alpha| d}{\|\alpha x + m\|} = \sup \frac{|\alpha| d}{|\alpha| \|x + m/\alpha\|} = \frac{d}{\inf \|x + m/\alpha\|} = 1$$

By H-B, there exists $x_0^* \in \mathcal{X}$ that is an extension of f, and $\|x_0^*\|_* = 1$. Clearly $x_0^* \in \mathcal{M}^\perp$ and $\langle x, x_0^* \rangle = d$, so x_0^* is optimal for the dual.

Minimum Norm Duality: Proof

Suppose m_0 is a primal optimal. Then,

$$\langle x - m_0, x_0^* \rangle = \langle x, x_0^* \rangle = d = \|x - m_0\|$$

Thus, $x - m_0$ and x_0^* are aligned.

For the last part, suppose $m_0 \in \mathcal{M}$ and assume that $x^* \in \mathcal{M}^\perp$ is non-zero vector aligned with $x - m_0$. WLOG, $\|x^*\|_* = 1$. Then,

$$\langle x, x^* \rangle = \langle x - m_0, x^* \rangle = \|x - m_0\|$$

whereas, for all $m \in \mathcal{M}$,

$$\langle x, x^* \rangle = \langle x - m, x^* \rangle \leq \|x - m\|$$

Thus, $\|x - m_0\| \leq \|x - m\|$.

Example: Chebyshev Approximation

Suppose $f \in C[a, b]$. We see a polynomial p_0 of degree n or less which approximates f in the sense of minimizing

$$\|f - p_0\| \triangleq \max_{a \le t \le b} |f(t) - p_0(t)|$$

Theorem. (Tonelli) If p_0 is the minimizing polynomial, then $|f(t) - p_0(t)|$ achieves its maximum at at least $n + 2$ points in $[a, b]$.

Minimum Norm Duality

Theorem. Suppose $\mathcal{X}$ is a normed linear space, $\mathcal{M} \subset \mathcal{X}$ is a subspace, and $x^* \in \mathcal{X}^*$. Then,

$$\min_{m^* \in \mathcal{M}^\perp} \|x^* - m^*\|_* = \sup_{\substack{\|x\| \le 1 \\ x \in \mathcal{M}}} \langle x, x^* \rangle$$

where the minimum is achieved for some $m_0^* \in \mathcal{M}^\perp$.

If $x_0 \in \mathcal{M}$ achieves the maximum, then x_0 is aligned with $x^* - m_0^*$.

Applications of Minimum Norm Duality

General strategy:

1. In the problem can be formulated as a minimum norm problem in a dual space, this guarantees the existence of an optimal solution

2. To characterize the optimal solution, use the alignment properties

3. Dual may be easier (e.g., finite dimensional)

Minimum Norm Problems with Affine Constraints

Consider fixed $y_i \in \mathcal{X}$, $i = 1, \ldots, n$, and define the affine constraint set

$$\mathcal{D} = \{x^* \in \mathcal{X}^* \ : \ \langle y_i, x^* \rangle = c_i, \ 1 \leq i \leq n\}$$

For $a \in \mathbb{R}^n$, define $Ya \in \mathcal{X}$ by

$$Ya \triangleq \sum_{i=1}^{n} y_i a_i$$

Theorem. Suppose $\mathcal{D}$ is non-empty. Then,

$$\min_{x^* \in \mathcal{D}} \|x^*\|_* = \max_{a: \ \|Ya\| \leq 1} c^\top a$$

Furthermore, the optimal x^* is aligned with the optimal Ya.

Example: Rocket Control

We would like to optimize the fuel consumption of a rocket:

- $x(t) =$ altitude of a rocket at time t
- $u(t) =$ thrust applied at time t
- system dynamics:

$$\ddot{x}(t) = u(t) - 1, \quad x(0) = \dot{x}(0) = 0$$

- constraint: acheive a certain altitude by time T (free variable),

$$x(T) = 1$$

- fuel consumption $= \int_0^T |u(t)|\, dt$

Example: Rocket Control

For now, fix T (we will optimize over T later).

By integration by parts, the dynamics be written as

$$x(T) = \int_0^T (T - t)u(t)\, dt - \tfrac{1}{2}T^2$$

$$\begin{aligned} \text{minimize} \quad & \int_0^T |u(t))|\, dt \\ \text{subject to} \quad & \int_0^T (T - t)u(t)\, dt = \tfrac{1}{2}T^2 + 1 \end{aligned}$$

We would like to formulate this as a minimum norm problem in a dual space.

Example: Rocket Control

$$\begin{aligned} &\text{minimize} && \int_0^T |dv(t))| \\ &\text{subject to} && \int_0^T (T-t)\, dv(t) = \tfrac{1}{2}T^2 + 1 \end{aligned}$$

$$\Leftrightarrow \quad \begin{aligned} &\text{minimize} && \|v\| \\ &\text{subject to} && \langle g, v\rangle = \tfrac{1}{2}T^2 + 1 \quad \text{where} \quad g(t) \triangleq T - t \\ & && v \in \mathrm{NBV}[0,\, T] \end{aligned}$$

$$\Leftrightarrow \quad \begin{aligned} &\text{maximize} && \alpha\left(\tfrac{1}{2}T^2 + 1\right) \\ &\text{subject to} && \|\alpha g\| \le 1 \\ & && \alpha \in \mathbb{R} \end{aligned}$$

$$\Leftrightarrow \quad \begin{aligned} &\text{maximize} && \alpha\left(\tfrac{1}{2}T^2 + 1\right) \\ &\text{subject to} && |\alpha| \le 1/T \qquad \text{since } \|g\| = T \\ & && \alpha \in \mathbb{R} \end{aligned}$$

Example: Rocket Control

$$\max_{\alpha:\ |\alpha| \le 1/T} \alpha\left(\tfrac{1}{2}T^2 + 1\right)$$

The maximum is achieved when $\alpha = 1/T$, thus we have the optimal value

$$\min_{v:\ \langle g, v\rangle = \frac{1}{2}T^2 + 1} \|v\| = \tfrac{1}{2}T + 1/T$$

The optimal v must be aligned with αg, therefore it is a step function at $t = 0$ (impulse control). Now, optimizing over T, we obtain

$$T^* = \sqrt{2}, \quad v^*(t) = \begin{cases} 0 & \text{if } t = 0 \\ \sqrt{2} & \text{if } t \in (0, \sqrt{2}] \end{cases}$$

Hyperplanes

Definition. A hyperplane H in a normed linear space $\mathcal{X}$ is maximal proper affine set. In other words, $H \subsetneq \mathcal{X}$, and if V is an affine set with $H \subsetneq V$, then $V = \mathcal{X}$.

Theorem. A set H is a hyperplane if and only if it is of the form

$$\{x \in \mathcal{X} \ : \ f(x) = c\}$$

where f is a non-zero linear functional and c is a scalar.

Theorem. The hyperplane $H = \{x \ : \ f(x) = c\}$ is closed for all scalars c if and only if f is continuous.

Definition. A halfspace is a set of the form $\{x \ : \ f(x) \leq c\}$. It is closed if f is continuous.

Hyperplanes and Convex Sets

Definition. Suppose $\mathcal{K}$ is a convex subset of a normed linear space $\mathcal{X}$ and $0 \in$ **int** $\mathcal{K}$. The Minkowski functional $p : \ \mathcal{X} \to \mathbb{R}$ is defined by

$$p(x) \triangleq \text{inf } \{r \in \mathbb{R} \ : \ x/r \in \mathcal{K}, \ r > 0\}$$

Lemma. The Minkowski functional satisfies:

1. $0 \leq p(x) < \infty$
2. $p(\alpha x) = \alpha p(x)$, for $\alpha > 0$
3. $p(x_1 + x_2) \leq p(x_1) + p(x_2)$
4. p is continuous
5. **cl** $\mathcal{K} = \{x \ : \ p(x) \leq 1\}$, **int** $\mathcal{K} = \{x \ : \ p(x) < 1\}$

Geometric Hahn-Banach Theorem

Theorem. Let $\mathcal{K}$ be a convex set with a non-empty interior in a normed linear space $\mathcal{X}$. Suppose $\mathcal{V} \subset \mathcal{X}$ is an affine set containing no interior points of $\mathcal{K}$. Then, there exists a closed hyperplane H containing $\mathcal{V}$ but no interior points of $\mathcal{K}$. In other words, there exists $x^* \in \mathcal{X}^*$ and $c \in \mathbb{R}$ such that

$$\langle x, x^* \rangle = c, \quad \forall\, x \in \mathcal{V}$$

$$\langle x, x^* \rangle < c, \quad \forall\, x \in \textbf{int}\ \mathcal{K}$$

Geometric Hahn-Banach Theorem: Proof

WLOG, assume that $\mathbf{0} \in \textbf{int}\ \mathcal{K}$. Let $\mathcal{M}$ be the subspace of $\mathcal{X}$ generated by $\mathcal{V}$. Since $\mathcal{V}$ is a hyperplane of $\mathcal{M}$ which does not contain $\mathbf{0}$, there is exists a functional f on $\mathcal{M}$ with

$$\mathcal{V} = \{x \in \mathcal{M} \ :\ f(x) = 1\}$$

Since $\mathcal{V}$ contains no interior points of $\mathcal{K}$,

$$f(v) = 1 \leq p(v), \quad \forall\, v \in \mathcal{V}$$

By homogeneity,

$$f(\alpha v) = \alpha \leq p(\alpha v), \quad \forall\, v \in \mathcal{V},\ \alpha > 0$$

$$f(\alpha v) \leq 0 \leq p(\alpha v), \quad \forall\, v \in \mathcal{V},\ \alpha < 0$$

Thus,

$$f(x) \leq p(x), \quad \forall\, x \in \mathcal{M}$$

Geometric Hahn-Banach Theorem: Proof

By the Hahn-Banach Theorem, we can extend f to a functional F on $\mathcal{X}$ with

$$F(x) \leq p(x), \quad \forall\, x \in \mathcal{X}$$

Since p is continuous, so is F, and $F(x) < 1$ for all $x \in$ **int** $\mathcal{K}$. Then, the desired hyperplane is

$$H = \{x \in \mathcal{X} \ : \ F(x) = 1\}$$

Minimum Norm Duality Revisited

Suppose $\mathcal{K} \subset \mathcal{X}$ is a convex set in a normed vector space $\mathcal{X}$. Define the support functional $h\colon \mathcal{X}^* \to \mathbb{R} \cup \{\infty\}$ by

$$S_{\mathcal{K}}(x^*) \triangleq \sup_{x \in \mathcal{K}} \langle x, x^* \rangle$$

Theorem. Suppose $x_1 \in \mathcal{X}$. Then,

$$\inf_{x \in \mathcal{K}} \|x - x_1\| = \max_{\|x^*\|_* \leq 1} \langle x_1, x^* \rangle - S_{\mathcal{K}}(x^*)$$

where the maximum is achieved by some $x_0^* \in \mathcal{X}^*$.

If the minimum is acheived by some $x_0 \in \mathcal{K}$, then $-x_0^*$ is aligned with $x_0 - x_1$.