Lecture 3

Optimization Problems and Iterative Algorithms

August 27, 2008

Outline

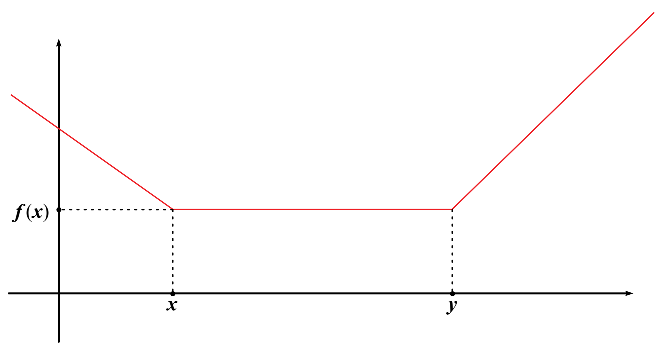
- Special Functions: Linear, Quadratic, *Convex*
- Criteria for Convexity of a Function
- Operations Preserving Convexity
- Unconstrained Optimization
  - First-Order Necessary Optimality Conditions
- Constrained Optimization
  - First-Order Necessary Optimality Conditions
  - KKT Conditions
- Iterative Algorithms

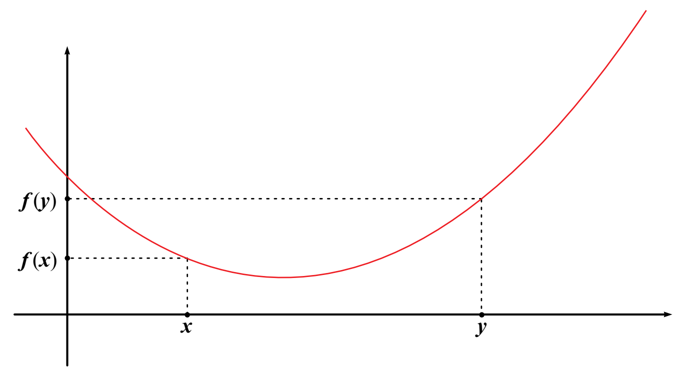
Convex Function

f is **convex** when $\mathrm{dom}(f)$ is convex set and there holds

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

for all $x, y \in \mathrm{dom}(f)$ and $\alpha \in [0, 1]$
strictly convex if the inequality is strict for all $x, y \in \mathrm{dom}(f)$ & $\alpha \in (0, 1)$





f is **concave** when $-f$ is convex
f is *strictly concave* when $-f$ is strictly convex

Examples of Convex/Concave Functions

Examples on $\mathbb{R}$

Convex:

- Affine: $ax + b$ over $\mathbb{R}$ for any $a, b \in \mathbb{R}$
- Exponential: e^{ax} over $\mathbb{R}$ for any $a \in \mathbb{R}$
- Power: x^p over $(0, +\infty)$ for $p \geq 1$ or $p \leq 0$
- Powers of absolute value: $|x|^p$ over $\mathbb{R}$ for $p \geq 1$
- Negative entropy: $x \ln x$ over $(0, +\infty)$

Concave:

- Affine: $ax + b$ over $\mathbb{R}$ for any $a, b \in \mathbb{R}$
- Powers: x^p over $(0, +\infty)$ for $0 \leq p \leq 1$
- Logarithm: $\ln x$ over $(0, +\infty)$

Examples on $\mathbb{R}^n$

- Affine functions are both convex and concave
- Norms $\|x\|, \|x\|_1, \|x\|_\infty$ are convex

Second-Order Conditions for Convexity

Let f be **twice differentiable** and let $\mathrm{dom}(f)$ be the domain of f
[In general, when differentiability is considered, it is required that $\mathrm{dom}(f)$ is open]
The Hessian $\nabla^2 f(x)$ is a symmetric $n \times n$ matrix whose entries are the second-order partial derivatives of f at x:

$$\left[\nabla^2 f(x)\right]_{ij} = \frac{\partial^2 f(x)}{\partial x_i \partial x_j} \qquad \text{for } i, j = 1, \ldots, n$$

2nd-order conditions:

- f is convex if and only if $\mathrm{dom}(f)$ is convex set and
$$\nabla^2 f(x) \succeq 0 \qquad \text{for all } x \in \mathrm{dom}(f)$$
- f is strictly convex if $\mathrm{dom}(f)$ is convex set
$$\nabla^2 f(x) \succ 0 \qquad \text{for all } x \in \mathrm{dom}(f)$$

Examples

- **Quadratic function:** $f(x) = (1/2)x'Qx + q'x + r$ with a symmetric $n \times n$ matrix Q
$$\nabla f(x) = Qx + q, \qquad \nabla^2 f(x) = Q$$
Convex for $Q \geq 0$

- **Least-squares objective**: $f(x) = \|Ax - b\|^2$ with an $m \times n$ matrix A
$$\nabla f(x) = 2A^T(Ax - b), \qquad \nabla^2 f(x) = 2A^T A$$
Convex for any A

- **Quadratic-over-linear:** $f(x, y) = x^2/y$
$$\nabla^2 f(x, y) = \tfrac{2}{y^3} \begin{bmatrix} y \\ -x \end{bmatrix} \begin{bmatrix} y \\ -x \end{bmatrix}^T \succeq 0$$
Convex for $y > 0$

First-Order Condition for Convexity

Let f be *differentiable* and let $\mathrm{dom}(f)$ be its domain. Then, the gradient

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

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

- **1st-order condition:** f is convex if and only if $\mathrm{dom}(f)$ is convex and

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

- Note: *A first order approximation is a global underestimate of f*

- Very important property used in convex optimization for algorithm designs and performance analysis

Operations Preserving Convexity

Let f and g be convex functions over $\mathbb{R}^n$

- **Positive Scaling**: λf is convex for $\lambda > 0$; $(\lambda f)(x) = \lambda f(x)$ for all x
- **Sum**: $f + g$ is convex; $(f + g)(x) = f(x) + g(x)$ for all x
- **Composition with affine function**: for g affine [i.e., $g(x) = Ax + b$], the composition $f \circ g$ is convex, where
$$(f \circ g)(x) = f(Ax + b) \text{ for all } x$$
- **Pointwise maximum**: For convex functions $f_1, \ldots, f_m$, *the pointwise-max function*
$$h(x) = \max\{f_1(x), \ldots, f_m(x)\} \quad \text{is convex}$$
  - *Polyhedral function*: $f(x) = \max_{i=1,\ldots,m}(a_i^T x + b_i)$ is convex
- **Pointwise supremum**: Let $Y \subseteq \mathbb{R}^m$ and $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$. Let $f(x, y)$ be convex in x for each $y \in Y$. Then, *the supremum function over the set Y*
$$h(x) = \sup_{y \in Y} f(x, y) \quad \text{is convex}$$

Optimization Terminology

Let $C \subseteq \mathbb{R}^n$ and $f : C \to \mathbb{R}$. Consider the following *optimization problem*

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

Example: $C = \{x \in \mathbb{R}^n \mid g(x) \leq 0,\ x \in X\}$

Terminology:

- The set C is referred to as *feasible set*
- We say that the problem is *feasible* when C is nonempty
- The problem is **unconstrained** when $C = \mathbb{R}^n$, and it is **constrained** otherwise
- We say that a vector x^* is **optimal** solution or a **global minimum** when x^* is *feasible* and the value $f(x^*)$ is not exceeded at any $x \in C$, i.e.,

$$x^* \in C$$

$$f(x^*) \leq f(x) \quad \text{for all } x \in C$$

Local Minimum

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

- A vector $\widehat{x}$ is a **local minimum** for the problem if $\widehat{x} \in C$ and there is a ball $B(\widehat{x}, r)$ such that

$$f(\widehat{x}) \leq f(x) \qquad \text{for all } x \in C \text{ with } \|x - \widehat{x}\| \leq r$$

- Every global minimum is also a local minimum
- When **the set C is convex and the function f is convex** then **a local minimum is also global**

First-Order Necessary Optimality Condition: Unconstrained Problem

Let f be a differentiable function with $\mathrm{dom}(f) = \mathbb{R}^n$ and let $C = \mathbb{R}^n$.

- If $\widehat{x}$ is a local minimum of f over $\mathbb{R}^n$, then the following holds:

$$\nabla f(\widehat{x}) = 0$$

- The gradient relation can be equivalently given as:

$$(y - \widehat{x})' \nabla f(\widehat{x}) \geq 0 \qquad \text{for all } y \in \mathbb{R}^n$$

  This is a **variational inequality** $VI(K, F)$ with the set K and the mapping F given by

$$K = \mathbb{R}^n, \qquad F(x) = \nabla f(x)$$

- Solving a minimization problem can be reduced to solving a corresponding variational inequality

First-Order Necessary Optimality Condition: Constrained Problem

Let f be a differentiable function with $\mathrm{dom}(f) = \mathbb{R}^n$ and let $C \subseteq \mathbb{R}^n$ be a closed convex set.

- If $\widehat{x}$ is a local minimum of f over C, then the following holds:

$$(y - \widehat{x})' \nabla f(\widehat{x}) \geq 0 \qquad \text{for all } y \in C \tag{1}$$

  Again, this is a **variational inequality** $VI(K, F)$ with the set K and the mapping F given by

$$K = C, \qquad F(x) = \nabla f(x)$$

- Recall that when f is convex, then a local minimum is also global
- **When f is convex**: the preceding relation is also sufficient for $\widehat{x}$ to be a global minimum i.e.,
  if $\widehat{x}$ satisfies relation (1), then $\widehat{x}$ is a (global) minimum

Equality and Inequality Constrained Problem

Consider the following problem

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & h_1(x) = 0, \ldots, h_p(x) = 0 \\ & g_1(x) \leq 0, \ldots, g_m(x) \leq 0 \end{array}$$

where f, h_i and g_j are continuously differentiable over $\mathbb{R}^n$.

Def. For a feasible vector x, an **active set of (inequality) constraints** is the set given by

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

If $j \notin \mathcal{A}(x)$, we say that the j-th constraint is *inactive* at x

Def. We say that a vector x is **regular** if the gradients

$$\nabla h_1(x), \ldots, \nabla h_p(x), \quad \text{and} \quad \nabla g_j(x) \text{ for } j \in \mathcal{A}(x)$$

are *linearly independent*

NOTE: x is regular when there are no equality constraints, and all the inequality constrains are inactive [$p = 0$ and $\mathcal{A}(x) = \emptyset$]

Lagrangian Function

With the problem

$$\begin{aligned} \text{minimize} \quad & f(x) \\ \text{subject to} \quad & h_1(x) = 0, \dots, h_p(x) = 0 \\ & g_1(x) \le 0, \dots, g_m(x) \le 0 \end{aligned} \qquad (2)$$

we associate the **Lagrangian function** $\mathcal{L}(x, \lambda, \mu)$ defined by

$$\mathcal{L}(x, \lambda, \mu) = f(x) + \sum_{i=1}^{p} \lambda_i h_i(x) + \sum_{j=1}^{m} \mu_j g_j(x)$$

where $\lambda_i \in \mathbb{R}$ for all i, and $\mu_j \in \mathbb{R}_+$ for all j

First-Order Karush-Kuhn-Tucker (KKT) Necessary Conditions

Th. Let $\widehat{x}$ be a local minimum of the equality/inequality constrained problem (2). Also, assume that $\widehat{x}$ is regular. Then, there exist unique multipliers $\widehat{\lambda}$ and $\widehat{\mu}$ such that

- $\nabla_x \mathcal{L}(\widehat{x}, \widehat{\lambda}, \widehat{\mu}) = 0$ [$\mathcal{L}$ is the Lagrangian function]
- $\widehat{\mu}_j \geq 0$ for all j
- $\widehat{\mu}_j = 0$ for all $j \notin \mathcal{A}(\widehat{x})$

The last condition is referred to as **complementarity conditions**
We can compactly write them as:

$$g(\widehat{x}) \perp \widehat{\mu}$$

Second-Order KKT Necessary Conditions

Th. Let $\widehat{x}$ be a local minimum of the equality/inequality constrained problem (2). Also, assume that $\widehat{x}$ is regular and that f, h_i, g_j are twice continuously differentiable. Then, there exist unique multipliers $\widehat{\lambda}$ and $\widehat{\mu}$ such that

- $\nabla_x \mathcal{L}(\widehat{x}, \widehat{\lambda}, \widehat{\mu}) = 0$
- $\widehat{\mu}_j \geq 0$ for all j
- $\widehat{\mu}_j = 0$ for all $j \notin \mathcal{A}(\widehat{x})$
- For any vector y such that $\nabla h_i(\widehat{x})' y = 0$ for all i and $\nabla g_j(\widehat{x})' y = 0$ for all $j \in \mathcal{A}(\widehat{x})$, the following relation holds:

$$y' \nabla^2_{xx} \mathcal{L}(\widehat{x}, \widehat{\lambda}, \widehat{\mu}) y \geq 0$$

Solution Procedures: Iterative Algorithms

For solving problems, we will consider **iterative algorithms**

- Given an initial iterate x_0
- We generate a new iterate

$$x_{k+1} = G_k(x_k)$$

  where G_k is a mapping that depends on the optimization problem

Objectives:

- Provide necessary conditions on the mappings G_k that yield a sequence $\{x_k\}$ converging to a solution of the problem of interest
- Study how fast the sequence $\{x_k\}$ converges:
  - Global convergence rate (when far from optimal points)
  - Local convergence rate (when near an optimal point)

Gradient Descent Method

Consider continuously differentiable function f. We want to

$$\text{minimize} \quad f(x) \text{ over } x \in \mathbb{R}^n$$

Gradient descent method

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

- The scalar α_k is a *stepsize*: $\alpha_k > 0$
- The stepsize choices $\alpha_k = \alpha$, or line search, or other stepsize rule so that $f(x_{k+1}) < f(x_k)$

Convergence Rate:

- Looking at the tail of an error $e(x_k) = dist(x_k, X^*)$ sequence: Local convergence is at the best **linear**

$$\limsup_{k\to\infty} \frac{e(x_{k+1})}{e(x_k)} \le q \qquad \text{for some } q \in (0,1)$$

- Global convergence is also at the best **linear**

Newton's Method

Consider twice continuously differentiable function f with Hessian $\nabla^2 f(x) > 0$ for all x. We want to

$$\text{minimize} \quad f(x) \text{ over } x \in \mathbb{R}^n$$

Newton's method

$$x_{k+1} = x_k - \alpha_k \nabla^2 f(x_k)^{-1} \nabla f(x_k)$$

Global Convergence Rate (when far from x^*)

- $f(x)$ decreases by at least γ at each iteration; therefore, there can be at most $(f(x_0) - f^*)/\gamma$ iterations [under some additional conditions on f]
- Method converges in **finite number of iterations**

Local Convergence Rate (near x^*)

- $\|\nabla f(x)\|$ converges to zero **quadratically**:

$$\|\nabla f(x_k)\| \le C\, q^{2^k} \qquad \text{for all large enough } k$$

where $C > 0$ and $q \in (0, 1)$

Penalty Methods

For solving inequality constrained problems:

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & g_j(x) \leq 0, \quad j = 1, \ldots, m \end{array}$$

Penalty Approach: Remove the constraints but penalize their violation

$$\mathcal{P}_c: \qquad \text{minimize} \;\; F(x,c) = f(x) + cP(g_1(x), \ldots, g_m(x)) \quad \text{over } x \in \mathbb{R}^n$$

where $c > 0$ is a penalty parameter and P is some penalty function

Penalty methods operate in two stages for c and x, respectively

- Choose initial value c_0

(1) Having c_k, solve the problem $\mathcal{P}_{c_k}$ to obtain its optimal $x^*(c_k)$

(2) Using $x^*(c_k)$, update c_k to obtain c_{k+1} and go to step 1

Interior-Point Methods

Solve inequality (and more generally) constrained problem:

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

The IPM solves a sequence of problems parametrized by $t > 0$:

$$\text{minimize} \;\; f(x) - \frac{1}{t} \sum_{j=1}^{m} \ln(-g_j(x)) \qquad \text{over } x \in \mathbb{R}^n$$

- Can be viewed as a penalty method with
  - Penalty parameter $c = \frac{1}{t}$
  - Penalty function

$$P(u_1, \ldots, u_m) = -\sum_{j=1}^{m} \ln(-u_j)$$

This function is known as **logarithmic barrier** or **log barrier** function

References for this lecture

The material for this lecture:

- (B) Bertsekas D.P. *Nonlinear Programming*
  - Chapter 1 and Chapter 3 (descent and Newton's methods, KKT conditions)
- (BNO) Bertsekas, Nedić, Ozdaglar *Convex Analysis and Optimization*
  - Chapter 1 (convex functions)
- Lecture slides for Convex Optimization at https://netfiles.uiuc.edu/angelia/www/angelia.html
  - Lectures 14 and 17 (descent and interior point methods)