EECS 545: Machine Learning

Supplementary Materials of Lecture 10: Brief Intro to Convex Optimization

Honglak Lee

02/14/2020

Basics of convex optimization

- general optimization problem
  - very difficult to solve
  - methods involve some compromise, e.g., very long computation time, or not always finding the solution

- exceptions: certain problem classes can be solved efficiently and reliably
  - least-squares problems
  - convex optimization problems

Slide credit: Stephen Boyd

Convex Sets

line segment between x_1 and x_2: all points

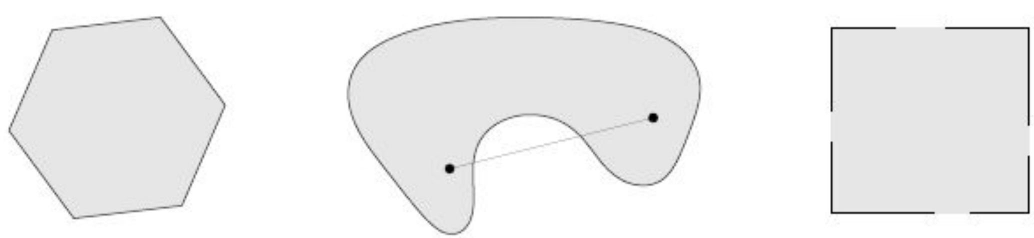
$$x = \theta x_1 + (1 - \theta) x_2$$

with $0 \leq \theta \leq 1$

convex set: contains line segment between any two points in the set

$$x_1, x_2 \in C, \quad 0 \leq \theta \leq 1 \quad \Longrightarrow \quad \theta x_1 + (1 - \theta) x_2 \in C$$

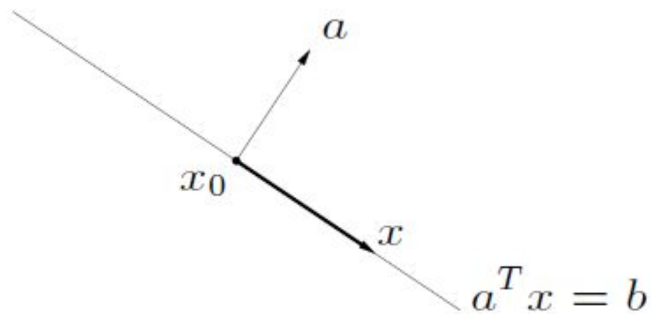
examples (one convex, two nonconvex sets)



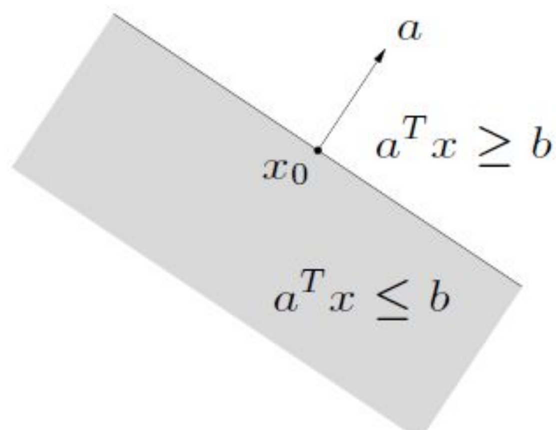
Slide credit: Stephen Boyd

Example: Hyper-planes and half-spaces

hyperplane: set of the form $\{x \mid a^T x = b\}$ $(a \neq 0)$



halfspace: set of the form $\{x \mid a^T x \leq b\}$ $(a \neq 0)$



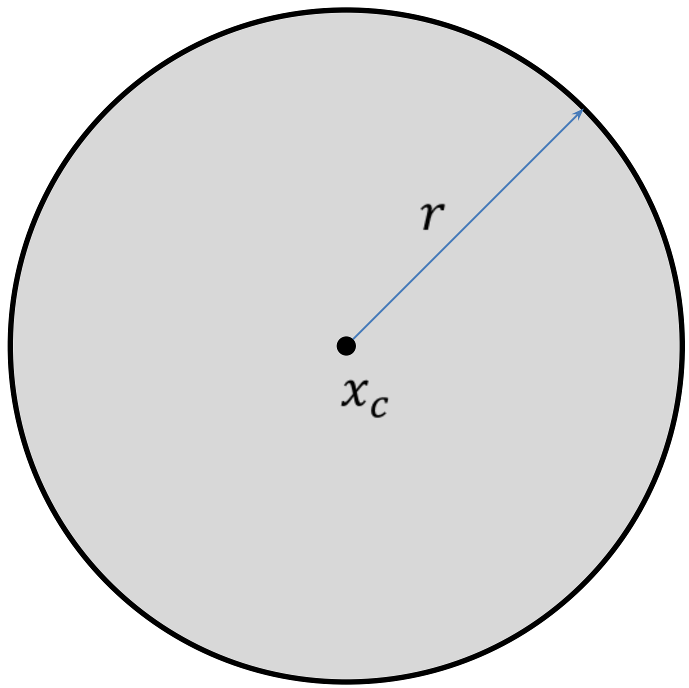
- a is the normal vector

Slide credit: Stephen Boyd

Example: Euclidean balls

(Euclidean) ball with center x_c and radius r:

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



Slide credit: Stephen Boyd

Convex Functions

$f : \mathbf{R}^n \to \mathbf{R}$ is convex if $\mathbf{dom}\, f$ is a convex set and

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

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

- f is concave if $-f$ is convex
- f is strictly convex if $\mathbf{dom}\, f$ is convex and

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

for $x, y \in \mathbf{dom}\, f$, $x \ne y$, $0 < \theta < 1$

Slide credit: Stephen Boyd

Examples of convex functions

convex:

- affine: $ax + b$ on $\mathbf{R}$, for any $a, b \in \mathbf{R}$
- exponential: e^{ax}, for any $a \in \mathbf{R}$
- powers: x^α on $\mathbf{R}_{++}$, for $\alpha \geq 1$ or $\alpha \leq 0$
- powers of absolute value: $|x|^p$ on $\mathbf{R}$, for $p \geq 1$
- negative entropy: $x \log x$ on $\mathbf{R}_{++}$

concave:

- affine: $ax + b$ on $\mathbf{R}$, for any $a, b \in \mathbf{R}$
- powers: x^α on $\mathbf{R}_{++}$, for $0 \leq \alpha \leq 1$
- logarithm: $\log x$ on $\mathbf{R}_{++}$

Slide credit: Stephen Boyd

Examples of convex functions

affine functions are convex and concave; all norms are convex

examples on $\mathbf{R}^n$

- affine function $f(x) = a^T x + b$
- norms: $\|x\|_p = (\sum_{i=1}^n |x_i|^p)^{1/p}$ for $p \geq 1$; $\|x\|_\infty = \max_k |x_k|$

examples on $\mathbf{R}^{m \times n}$ ($m \times n$ matrices)

- affine function

$$f(X) = \mathbf{tr}(A^T X) + b = \sum_{i=1}^{m} \sum_{j=1}^{n} A_{ij} X_{ij} + b$$

Slide credit: Stephen Boyd

Examples

quadratic function: $f(x) = (1/2)x^T Px + q^T x + r$ (with $P \in \mathbf{S}^n$)

$$\nabla f(x) = Px + q, \qquad \nabla^2 f(x) = P$$

convex if $P \succeq 0$

least-squares objective: $f(x) = \|Ax - b\|_2^2$

$$\nabla f(x) = 2A^T(Ax - b), \qquad \nabla^2 f(x) = 2A^T A$$

convex (for any A)

First-order condition for convexity

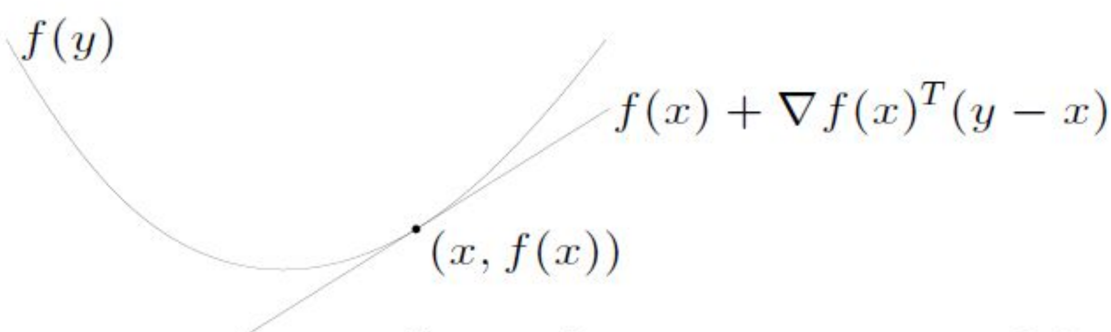
f is **differentiable** if $\mathbf{dom}\, f$ is open and the gradient

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

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

1st-order condition: differentiable f with convex domain is convex iff

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



first-order approximation of f is global underestimator

Second-order condition for convexity

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

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

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

2nd-order conditions: for twice differentiable f with convex domain

- f is convex if and only if

$$\nabla^2 f(x) \succeq 0 \quad \text{for all } x \in \mathbf{dom}\, f$$

- if $\nabla^2 f(x) \succ 0$ for all $x \in \mathbf{dom}\, f$, then f is strictly convex

Jensen's inequality

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

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

extension: if f is convex, then

$$f(\mathbf{E}\, z) \leq \mathbf{E}\, f(z)$$

for any random variable z

basic inequality is special case with discrete distribution

$$\mathbf{prob}(z = x) = \theta, \qquad \mathbf{prob}(z = y) = 1 - \theta$$

Operations that preserve convexity

practical methods for establishing convexity of a function

1. verify definition (often simplified by restricting to a line)

2. for twice differentiable functions, show $\nabla^2 f(x) \succeq 0$

3. show that f is obtained from simple convex functions by operations that preserve convexity

   - nonnegative weighted sum
   - composition with affine function
   - pointwise maximum and supremum
   - composition
   - minimization
   - perspective

Slide credit: Stephen Boyd

Convex Optimization

- Convex optimization is described as follows:

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

f: convex function, C : convex set

- Rewriting C using equality and inequality constraints:

$$\begin{aligned} &\text{minimize} && f(x) && \\ &\text{subject to} && g_i(x) \leq 0, && i = 1, \dots, m \\ & && h_i(x) = 0, && i = 1, \dots, p \end{aligned}$$

f: convex function, g_i: convex function, h_i: affine function.

- Special kinds of convex programming:
  - Linear Programming
  - Quadratic Programming

Linear and Quadratic Programming

- We say a convex optimization problem is **a linear program (LP)** if both f and inequality constraints g_i are affine. That is,

$$\begin{aligned} \text{minimize} \quad & c^T x + d \\ \text{subject to} \quad & Gx \leq h \\ & Ax = b \end{aligned}$$

$$x \in \mathbb{R}^n \; c \in \mathbb{R}^n, d \in \mathbb{R}, G \in \mathbb{R}^{m \times n}, h \in \mathbb{R}^m, A \in \mathbb{R}^{p \times n}, b \in \mathbb{R}^p$$

- We say a convex optimization problem is **a quadratic program (QP)** if f is convex quadratic function, and g_i are affine. That is,

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2} x^T P x + c^T x + d \qquad (P \in \mathbb{S}^n_+) \\ \text{subject to} \quad & Gx \leq h \\ & Ax = b \end{aligned}$$

Solving Constrained Optimization: General Overview and Recipe

Recap of Lecture 10

Constrained Optimization

- General **constrained problem** has the form:

$$\begin{aligned} \min_{\mathbf{x}} \quad & f(\mathbf{x}) \\ \text{subject to} \quad & g_i(\mathbf{x}) \leq 0,\ i = 1, ..., m \\ & h_i(\mathbf{x}) = 0,\ i = 1, ..., p \end{aligned}$$

- If **x** satisfies all the constraints, **x** is called <u>feasible</u>.

Lagrangian Formulation

- The **Lagrangian function** is

$$\mathcal{L}(\mathbf{x}, \lambda, \nu) = f(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i g_i(\mathbf{x}) + \sum_{i=1}^{p} \nu_i h_i(\mathbf{x})$$

  - Here, $\lambda = [\lambda_1, \dots, \lambda_m]$ $(\lambda_i \geq 0, \forall i)$ and $\nu = [\nu_1, \dots, \nu_p]$ are called Lagrange multipliers (or dual variables)

- This leads to **primal optimization problem** (see next slide):

$$\min_{\mathbf{x}} \max_{\nu, \lambda : \lambda_i \geq 0, \forall i} \mathcal{L}(\mathbf{x}, \lambda, \nu)$$

- Difficult to solve directly!

Primal and Feasibility

- Primal optimization problem:

$$p^* = \min_{\mathbf{x}} \max_{\nu,\lambda:\lambda_i \geq 0,\forall i} \mathcal{L}(\mathbf{x},\lambda,\nu)$$

  – where

$$\mathcal{L}(\mathbf{x},\lambda,\nu) = f(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i g_i(\mathbf{x}) + \sum_{i=1}^{p} \nu_i h_i(\mathbf{x})$$

- Notice that:

$$\mathcal{L}_p(\mathbf{x}) = \max_{\nu,\lambda:\lambda_i \geq 0,\forall i} \mathcal{L}(\mathbf{x},\lambda,\nu) = \begin{cases} f(\mathbf{x}) & \text{if x is feasible} \\ \infty & \text{otherwise} \end{cases}$$

Lagrange Dual

- Dual optimization problem:

$$\max_{\nu,\lambda:\lambda_i\geq 0,\forall i} \min_{\mathbf{x}} \mathcal{L}(\mathbf{x},\lambda,\nu)$$

- We can also write as:

$$\begin{aligned} \max_{\lambda,\nu} \min_{\mathbf{x}} \quad & \mathcal{L}(\mathbf{x},\lambda,\nu) \\ \text{subject to} \quad & \lambda_i \geq 0,\ \forall i \end{aligned}$$

Weak Duality

- Claim:

$$\begin{aligned} d^* &= \max_{\lambda,\nu:\lambda_i \geq 0} \min_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \lambda, \nu) \\ &\leq \min_{\mathbf{x}} \max_{\lambda,\nu:\lambda_i \geq 0} \mathcal{L}(\mathbf{x}, \lambda, \nu) \\ &= p^* \end{aligned}$$

- Difference between *p** and *d** is called <u>duality gap</u>.

Weak Duality

- Proof:

Let $\tilde{\mathbf{x}}$ be feasible. Then for any λ, ν with $\lambda_i \geq 0$,

$$\mathcal{L}(\tilde{\mathbf{x}}, \lambda, \nu) = f(\tilde{\mathbf{x}}) + \sum_i \lambda_i g_i(\tilde{\mathbf{x}}) + \sum_i \nu_i h_i(\tilde{\mathbf{x}}) \leq f(\tilde{\mathbf{x}})$$

Thus, $\tilde{\mathcal{L}}(\lambda, \nu) = \min_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \lambda, \nu) \leq \mathcal{L}(\tilde{\mathbf{x}}, \lambda, \nu) \leq f(\tilde{\mathbf{x}})$.
Then,

$$d^* = \max_{\lambda, \nu : \lambda_i \geq 0} \tilde{\mathcal{L}}(\lambda, \nu) \leq f(\tilde{\mathbf{x}}) \text{ for any feasible } \tilde{\mathbf{x}}$$

Finally,

$$d^* = \max_{\lambda, \nu : \lambda_i \geq 0} \tilde{\mathcal{L}}(\lambda, \nu) \leq \min_{\tilde{\mathbf{x}}:\text{feasible}} f(\tilde{\mathbf{x}}) = p^*$$

Strong Duality

- If $p^*= d^*$, we say <u>strong duality</u> holds.

- What are the conditions for strong duality?
  - does not hold in general
  - holds for convex problems (under mild conditions)
  - conditions that guarantee strong duality in convex problems are called constraint qualification.

- Two well-known conditions
  - Slater’s constraint qualification
  - Karush-Kuhn-Tucker (KKT) condition

Conditions for strong duality: Slater's constraint qualification

- Strong duality holds for a convex problem

$$\begin{aligned} \min_{\mathbf{x}} \quad & f(\mathbf{x}) \\ \text{subject to} \quad & g_i(\mathbf{x}) \leq 0,\ i = 1, ..., m \\ & h_i(\mathbf{x}) = 0,\ i = 1, ..., p \end{aligned}$$

(where *f, g_i* are convex, *and h_i* are affine)

– If it is strictly feasible, i.e.,

$$\begin{aligned} \exists x : \quad & g_i(\mathbf{x}) < 0,\ \forall i = 1, ..., m \\ & h_i(\mathbf{x}) = 0,\ \forall i = 1, ..., p \end{aligned}$$

Slater's condition is a sufficient condition for strong duality to hold for a convex problem

Karush-Kuhn-Tucker (KKT) condition

$$\nabla_{\mathbf{x}} f(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i \nabla_{\mathbf{x}} g_i(\mathbf{x}^*) + \sum_{i=1}^{p} \nu_i \nabla_{\mathbf{x}} h_i(\mathbf{x}^*) = 0 \qquad (1)$$

$$h_i(\mathbf{x}^*) = 0,\ i = 1, \ldots, p \qquad (2)$$

$$g_i(\mathbf{x}^*) \leq 0,\ i = 1, \ldots, m \qquad (3)$$

$$\lambda_i^* \geq 0,\ i = 1, \ldots, m \qquad (4)$$

$$\lambda_i^* g_i(\mathbf{x}^*) = 0,\ i = 1, \ldots, m \qquad (5)$$

– The last condition is called complementary slackness.

Conditions for strong duality: KKT Conditions

- Assume *f*, g_i, h_i are differentiable

- If the original problem is **convex** (where *f*, g_i are convex, *and* h_i are affine) and $\mathbf{x}^*, \lambda^*, \nu^*$ satisfy the KKT conditions, then

  - $\mathbf{x}^*$ is primal optimal
  - (λ^*, ν^*) is dual optimal, and
  - the <u>duality gap is zero</u> (i.e., strong duality holds)

Proof for sufficiency

- From (2) and (3), $\mathbf{x}^*$ is primal feasible.
- From (4), (λ^*, ν^*) is dual feasible.

- $\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu})$ is a convex differentiable function. Thus, from (1), $\mathbf{x}^*$ is a minimizer of $\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu})$.
- Then,

$$
\begin{aligned}
d^* = \tilde{\mathcal{L}}(\lambda^*, \nu^*) \quad &= \quad \min_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \lambda^*, \nu^*) \\
&= \quad \mathcal{L}(\mathbf{x}^*, \lambda^*, \nu^*) \\
&= \quad f(\mathbf{x}^*) + \sum_i \lambda_i g_i(\mathbf{x}^*) + \sum_i \nu_i h_i(\mathbf{x}^*) \\
&= \quad f(\mathbf{x}^*)
\end{aligned}
$$

- Due to weak duality, $d^* = f(\mathbf{x}^*) \leq f(\mathbf{x})$ for all feasible $\mathbf{x}$. Therefore, $d^* = p^*$.

KKT conditions: Conclusion

- If a constrained optimization if differentiable and has convex objective function and constraint sets, then the KKT conditions are **(necessary and) sufficient conditions** for **strong duality** (zero duality gap).

- Thus, the KKT conditions can be used to solve such problems.

Recap: General Recipe

- Given an original optimization

$$\begin{aligned} \min_{\mathbf{x}} \quad & f(\mathbf{x}) \\ \text{subject to} \quad & g_i(\mathbf{x}) \le 0,\ i = 1, ..., m \\ & h_i(\mathbf{x}) = 0,\ i = 1, ..., p \end{aligned}$$

- Solve dual optimization with <u>Lagrangian function</u>:

$$\begin{aligned} \max_{\lambda,\nu} \min_{\mathbf{x}} \quad & \mathcal{L}(\mathbf{x}, \lambda, \nu) = f(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i g_i(\mathbf{x}) + \sum_{i=1}^{p} \nu_i h_i(\mathbf{x}) \\ \text{subject to} \quad & \lambda_i \ge 0,\ \forall i \end{aligned}$$

- Alternatively, solve the dual optimization with <u>Lagrange dual</u>:

$$\begin{aligned} \max_{\lambda,\nu} \quad & \tilde{\mathcal{L}}(\lambda, \nu) \qquad \text{where} \ \ \tilde{\mathcal{L}}(\lambda, \nu) = \min_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \lambda, \nu) \\ \text{subject to} \quad & \lambda_i \ge 0,\ \forall i \end{aligned}$$

Recap: KKT Optimality condition

- Karush-Kuhn-Tucker (KKT) condition:

$$\nabla_{\mathbf{x}} f(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i \nabla_{\mathbf{x}} g_i(\mathbf{x}^*) + \sum_{i=1}^{p} \nu_i \nabla_{\mathbf{x}} h_i(\mathbf{x}^*) = 0$$

$$h_i(\mathbf{x}^*) = 0,\ i = 1, ..., p$$

$$g_i(\mathbf{x}^*) \leq 0,\ i = 1, ..., m$$

$$\lambda_i^* \geq 0,\ i = 1, ..., m$$

$$\lambda_i^* g_i(\mathbf{x}^*) = 0,\ i = 1, ..., m$$

  – The last condition is called complementary slackness.

Additional Resource

- Convex Optimization
  - http://www.stanford.edu/~boyd/cvxbook/
  - http://www.stanford.edu/class/ee364a/
  - For materials covered today, see Chapter 5 (and earlier chapters).