# EECS 545: Machine Learning

# Supplementary Materials (Review Session)
# Brief Intro to Convex Optimization

Junghwan Kim

02/04/2022

* Many slides are based on Stephen Boyd’s course:
Convex Optimization (website: http://www.stanford.edu/class/ee364a/)

# Basics of convex optimization

- General optimization problem

$$\begin{array}{ll} \min\limits_{\mathbf{x}} & f(\mathbf{x}) \\ \text{subject to} & g_i(\mathbf{x}) \leq 0,\ i = 1, ..., m \\ & h_i(\mathbf{x}) = 0,\ i = 1, ..., p \end{array}$$

  - 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

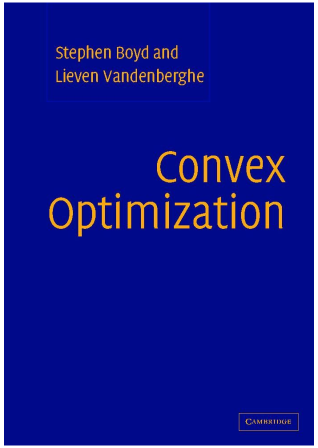
http://www.stanford.edu/class/ee364a/

Slide credit: Stephen Boyd

# Contents

- Review: Convex Set, Convex Function
- Linear Programming, Quadratic Programming
- Constrained Optimization
- Lagrangian and Duality
- KKT Conditions for Strong Duality

  ← the goal of today!




https://web.stanford.edu/class/ee364a/
https://stanford.edu/~boyd/cvxbook/

# Convex Sets

**line segment** between $x_1$ and $x_2$: all points

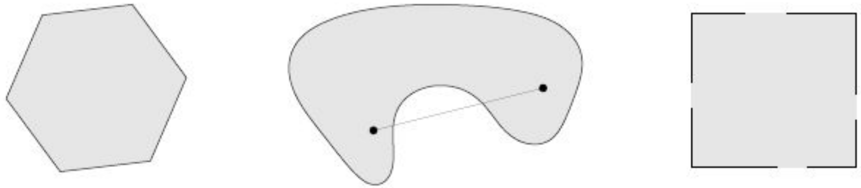
$$x = \theta x_1 + (1 - \theta) x_2$$

with $0 \leq \theta \leq 1$

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

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

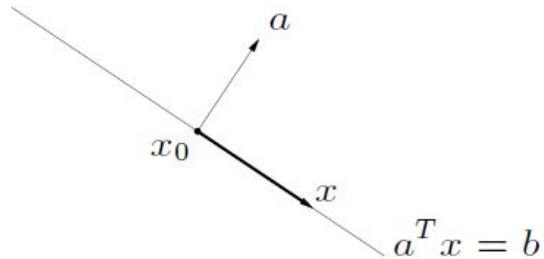
**examples** (one convex, two nonconvex sets)



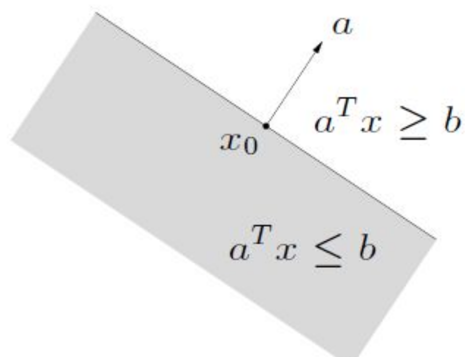
Slide credit: Stephen Boyd

# 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 $\{x \mid a^T x \leq b\}$ $(a \neq 0)$



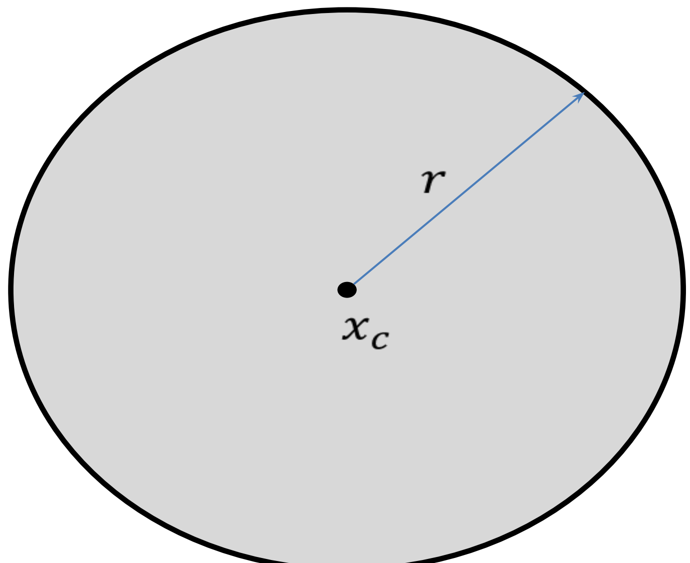

- $a$ is the normal vector

Slide credit: Stephen Boyd

# Example: Euclidean balls

**(Euclidean) ball** with center $x_c$ and radius $r$:

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



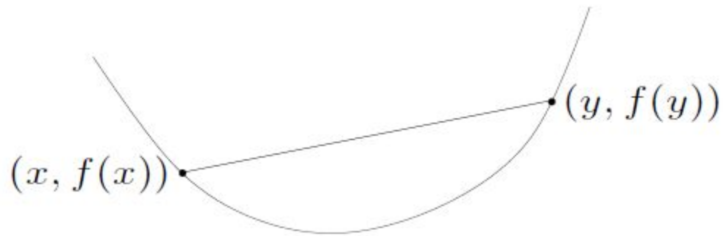

Slide credit: Stephen Boyd

# Convex Functions

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

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

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




- $f$ 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 $x, y \in \mathbf{dom}\, f$, $x \neq y$, $0 < \theta < 1$

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

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

convex if $P \succeq 0$ $\leftarrow$ Definition: P is positive semi-definite.

**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

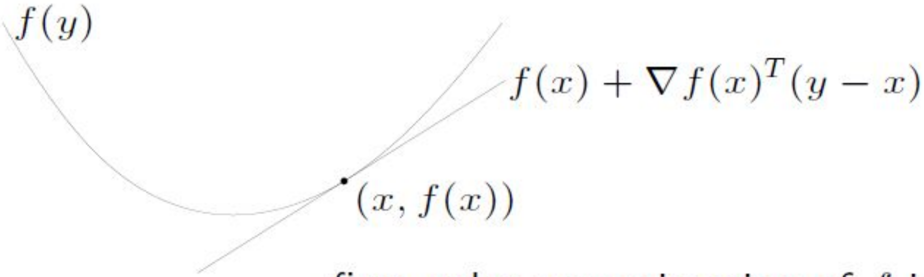
$f$ is **differentiable** if $\mathbf{dom}\, f$ is open and the gradient

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

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

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

(i.e., Hessian matrix at x is positive semi-definite for all x.)

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

(positive definite)

# 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)
   $f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta)f(y)$
2. for twice differentiable functions, show $\nabla^2 f(x) \succeq 0$
   (i.e. Hessian is positive semi-definite)
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

Any questions so far?

# Convex Optimization

- Convex optimization is described as follows:

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

$f$: convex function, $C$ : convex set

- Rewriting $C$ using equality and inequality constraints:

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

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

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

# Linear Programming

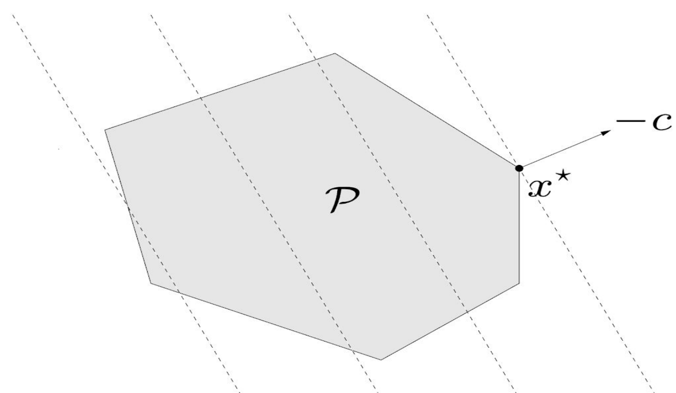
We say a convex optimization problem is **a linear program (LP)** if both $f$ and inequality constraints $g_i$ are affine. That is,

$$\begin{array}{ll} \text{minimize} & c^T x + d \\ \text{subject to} & Gx \preceq h \\ & Ax = b \end{array}$$

← element-wise inequality
($g_i$ is the i-th row of G)

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

- feasible set is a polyhedron




Slide credit: Stephen Boyd

# Linear Programming: applications

**diet problem:** choose quantities $x_1, \ldots, x_n$ of $n$ foods

- one unit of food $j$ costs $c_j$, contains amount $a_{ij}$ of nutrient $i$
- healthy diet requires nutrient $i$ in quantity at least $b_i$

to find cheapest healthy diet,

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & Ax \succeq b, \quad x \succeq 0 \end{array}$$

**piecewise-linear minimization**

$$\text{minimize} \quad \max_{i=1,\ldots,m}(a_i^T x + b_i)$$

equivalent to an LP

$$\begin{array}{ll} \text{minimize} & t \\ \text{subject to} & a_i^T x + b_i \leq t, \quad i = 1, \ldots, m \end{array}$$

Slide credit: Stephen Boyd

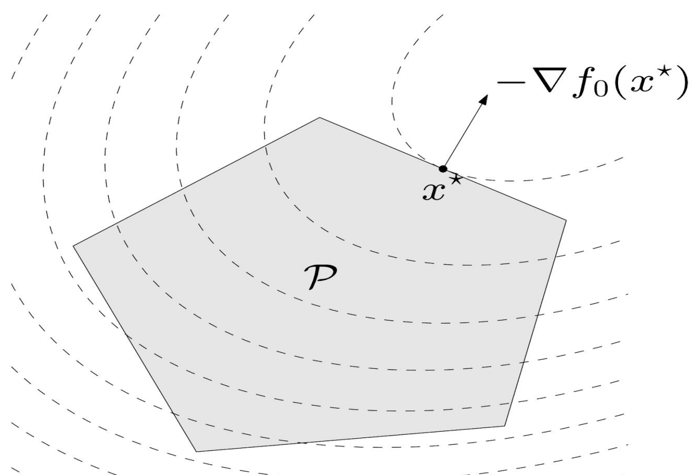
# Quadratic Programming

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

$$\begin{array}{ll} \text{minimize} & (1/2)x^T P x + q^T x + r \\ \text{subject to} & Gx \preceq h \\ & Ax = b \end{array}$$

← element-wise inequality
($g_i$ is the i-th row of G)

- $P \in \mathbf{S}^n_+$, so objective is convex quadratic
- minimize a convex quadratic function over a polyhedron




Slide credit: Stephen Boyd

# Quadratic Programming: applications

**least-squares**

$$\text{minimize} \quad \|Ax - b\|_2^2$$

- analytical solution $x^\star = A^\dagger b$ ($A^\dagger$ is pseudo-inverse)
- can add linear constraints, *e.g.*, $l \preceq x \preceq u$

**linear program with random cost**

$$\begin{array}{ll} \text{minimize} & \bar{c}^T x + \gamma x^T \Sigma x = \mathbf{E}\, c^T x + \gamma\, \mathbf{var}(c^T x) \\ \text{subject to} & Gx \preceq h, \quad Ax = b \end{array}$$

- $c$ is random vector with mean $\bar{c}$ and covariance $\Sigma$
- hence, $c^T x$ is random variable with mean $\bar{c}^T x$ and variance $x^T \Sigma x$
- $\gamma > 0$ is risk aversion parameter; controls the trade-off between expected cost and variance (risk)

Slide credit: Stephen Boyd 

# Solving Constrained Optimization: General Overview and Recipe

# 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>.

# A Big Picture

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

**Constrained Optimization Problem**

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

**Lagrangian**

e.g. convex optimizations, KKT conditions

**strong duality** (if conditions are met 😎)

$$p^* = d^*$$

**Primal Optimization Problem** (min-max)

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

**weak duality**

$$p^* \geq d^*$$

**Dual Optimization Problem** (max-min)

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

# Lagrangian Formulation

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

- 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!

# Example: Lagrangian Derivation

Consider the following problem:

$$\begin{aligned} \text{minimize}_{x\in\mathbb{R}^2} \quad & (2x_1-1)^2+(x_2-2)^2 \\ \text{subject to} \quad & 3x_1+2x_2 \le 4 \\ & x_2 \ge x_1 \end{aligned}$$

Lagrangian:

$$\begin{aligned} \mathcal{L}(x,\lambda) &= (2x_1-1)^2+(x_2-2)^2+\lambda_1(3x_1+2x_2-4)+\lambda_2(x_1-x_2) \\ &= \left(2x_1-\frac{4-3\lambda_1-\lambda_2}{4}\right)^2+\left(x_2-\frac{4-2\lambda_1+\lambda_2}{2}\right)^2 \\ &\quad -\frac{1}{16}\left[25\lambda_1^2+5\lambda_2^2-10\lambda_1\lambda_2-24\lambda_1+24\lambda_2\right] \end{aligned}$$

# Primal and Feasibility

- Primal optimization problem:

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

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

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:

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

This eliminates the constraints on **x**, yielding an equivalent optimization problem.

# Lagrange Dual

- Dual optimization problem:

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

**cf) primal optimization problem**

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

- We can also write it as:

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

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

"dual function" $\tilde{\mathcal{L}}(\lambda,\nu) = \min_{\mathbf{x}} \mathcal{L}(\mathbf{x},\lambda,\nu)$

$g(\lambda,\nu)$

# Example: Dual Problem Derivation

Consider the following problem:

$$\begin{aligned} \text{minimize}_{x \in \mathbb{R}^2} \quad & (2x_1 - 1)^2 + (x_2 - 2)^2 \\ \text{subject to} \quad & 3x_1 + 2x_2 \leq 4 \\ & x_2 \geq x_1 \end{aligned}$$

Lagrangian:

$$\mathcal{L}(x, \lambda) = \left(2x_1 - \frac{4 - 3\lambda_1 - \lambda_2}{4}\right)^2 + \left(x_2 - \frac{4 - 2\lambda_1 + \lambda_2}{2}\right)^2 - \frac{1}{16}\left[25\lambda_1^2 + 5\lambda_2^2 - 10\lambda_1\lambda_2 - 24\lambda_1 + 24\lambda_2\right]$$

Dual Objective: $g(\lambda) = \min_x \mathcal{L}(x, \lambda) = -\frac{1}{16}\left[25\lambda_1^2 + 5\lambda_2^2 - 10\lambda_1\lambda_2 - 24\lambda_1 + 24\lambda_2\right]$

# 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 the **duality gap**.
- In other words, the dual maximization problem (usually easier) gives a “**lower bound**” for the primal minimization problem (usually more difficult).

# Example: Quadratic Programming

**Primal problem:**

$$\begin{aligned} &\text{minimize} \quad x^T P x \qquad\qquad (P \succ 0) \\ &\text{subject to} \quad Ax \preceq b \end{aligned}$$

**Lagrangian:** $\mathcal{L} = x^T P x + \lambda^T (Ax - b)$

**Primal:** $\min\limits_{x} \max\limits_{\lambda \geq 0} \left\{ x^T P x + \lambda^T (Ax - b) \right\}$

**Dual:** $\max\limits_{\lambda \geq 0} \min\limits_{x} \left\{ x^T P x + \lambda^T (Ax - b) \right\}$

$$\tilde{\mathcal{L}}(\lambda) = \min_{x} (x^T P x + \lambda^T (Ax - b)) = -\frac{1}{4} \lambda^T A P^{-1} A^T \lambda - b^T \lambda$$

**Dual:**

$$\begin{aligned} &\text{maximize} \quad -\frac{1}{4} \lambda^T A P^{-1} A^T \lambda - b^T \lambda \\ &\text{subject to} \quad \lambda \succeq 0 \end{aligned}$$

# Weak Duality

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

**Proof:** Let $\tilde{\mathbf{x}}$ be feasible. Then for any $\lambda, \nu$ 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, $\tilde{\mathcal{L}}(\lambda,\nu) = \min_{\mathbf{x}} \mathcal{L}(\mathbf{x},\lambda,\nu) \leq \mathcal{L}(\tilde{\mathbf{x}},\lambda,\nu) \leq f(\tilde{\mathbf{x}})$.
for any $\lambda, \nu$ with $\lambda_i \geq 0$, any feasible $\tilde{\mathbf{x}}$

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 **strong duality** 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

Let $\mathbf{x}^*$ be a primal optimal and $\lambda^*, \nu^*$ be a dual optimal solution.
If the strong duality holds, then we have the following:

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

$$g_i(\mathbf{x}^*) \leq 0, \quad i = 1, \ldots, m, \quad (2)$$

$$h_i(\mathbf{x}^*) = 0, \quad i = 1, \ldots, p, \quad (3)$$

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

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

(called complementary slackness)

$$\begin{array}{ll} \min\limits_{\mathbf{x}} & f(\mathbf{x}) \\ \text{subject to} & g_i(\mathbf{x}) \leq 0,\ i = 1, ..., m \\ & h_i(\mathbf{x}) = 0,\ i = 1, ..., p \end{array}$$

$$\begin{array}{ll} \max\limits_{\lambda,\nu} \min\limits_{\mathbf{x}} & \mathcal{L}(\mathbf{x}, \lambda, \nu) \\ \text{subject to} & \lambda_i \geq 0,\ \forall i \end{array}$$

Note: we do **NOT** assume the optimization problem is necessarily convex.

# KKT condition: complementary slackness

Let $\mathbf{x}^*$ be a primal optimal and $\lambda^*, \nu^*$ be a dual optimal solution.
If the strong duality holds,

$$
\begin{aligned}
f(\mathbf{x}^*) &= g(\lambda^*, \nu^*) \\
&= \min_{\mathbf{x}} \left( f(\mathbf{x}) + \sum_i \lambda_i^* g_i(\mathbf{x}) + \sum_i \nu_i^* h_i(\mathbf{x}) \right) \\
&\leq f(\mathbf{x}^*) + \sum_i \lambda_i^* g_i(\mathbf{x}^*) + \sum_i \nu_i^* h_i(\mathbf{x}^*) \\
&\leq f(\mathbf{x}^*)
\end{aligned}
$$

$$\therefore \sum_i \lambda_i^* g_i(\mathbf{x}^*) = 0$$

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

(called complementary slackness)

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

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

# Example: KKT Conditions

Consider
$$\begin{aligned} \text{minimize}_{x\in\mathbb{R}^2}\ & (2x_1-1)^2+(x_2-2)^2 \\ \text{subject to}\ & 3x_1+2x_2 \le 4 \\ & x_2 \ge x_1 \end{aligned}$$

KKT Condition:
$$\begin{bmatrix} 4(2x_1^*-1)+3\lambda_1^*+\lambda_2^* \\ 2(x_2^*-2)+2\lambda_1^*-\lambda_2^* \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \qquad (1)$$

$$3x_1^*+2x_2^*-4 \le 0,\ x_1^*-x_2^* \le 0 \qquad (2)$$

$$\lambda_1^* \ge 0,\ \lambda_2^* \ge 0 \qquad (4)$$

$$\begin{aligned} \lambda_1^*(3x_1^*+2x_2^*-4) &= 0 \\ \lambda_2^*(x_1^*-x_2^*) &= 0 \end{aligned} \qquad (5)$$

# 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
  - $(\lambda^*, \nu^*)$ is dual optimal, and
  - the <u>duality gap is zero</u> (i.e., strong duality holds)

**For convex optimization problems** (+ differentiable objectives/constraints)**,**
**KKT is a sufficient condition for strong duality.**

# Proof for sufficiency

Claim: When KKT (1)-(5) holds, the strong duality holds.

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

$$
\begin{aligned}
d^* = \widetilde{\mathcal{L}}(\lambda^*, \nu^*) &= \min_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \lambda^*, \nu^*) \\
&= \mathcal{L}(\mathbf{x}^*, \lambda^*, \nu^*) \\
&= f(\mathbf{x}^*) + \underbrace{\sum_i \lambda_i g_i(\mathbf{x}^*)}_{= 0 \ \because \text{(5) complementary slackness}} + \sum_i \nu_i h_i(\mathbf{x}^*) \\
&= f(\mathbf{x}^*)
\end{aligned}
$$

(See also: derivation of complementary slackness)

- Then,

$$
d^* = \widetilde{\mathcal{L}}(\lambda^*, \nu^*) \leq \underbrace{\max_{\lambda, \nu : \lambda_i \geq 0} \widetilde{\mathcal{L}}(\lambda, \nu) \leq \min_{\mathbf{x}} f(\mathbf{x})}_{\text{same proof as in weak duality}} \leq f(\mathbf{x}^*) = d^*
$$

- Then,

$$
\max_{\lambda, \nu : \lambda_i \geq 0} \widetilde{\mathcal{L}}(\lambda, \nu) = \min_{\mathbf{x}} f(\mathbf{x})
$$

which proves that the strong duality holds (i.e., duality gap is zero).

# KKT conditions: Conclusion

- If a constrained optimization is 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.
  - e.g. Support Vector Machines!

# A Big Picture

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

**Constrained Optimization Problem**

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

**Lagrangian**

e.g. convex optimizations,
KKT conditions

**strong duality** (if conditions are met 😎)

$$p^* = d^*$$

**Primal Optimization Problem** (min-max)

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

**weak duality**

$$p^* \ge d^*$$

**Dual Optimization Problem** (max-min)

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

# Recap: General Recipe

- Given an original optimization

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

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

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

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

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

# Recap: KKT Optimality condition

- Karush-Kuhn-Tucker (KKT) condition:

$$
\begin{aligned}
&\nabla_{\mathbf{x}} f(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i \nabla_{\mathbf{x}} g_i(\mathbf{x}^*) + \sum_{i=1}^{p} \nu_i^* \nabla_{\mathbf{x}} h_i(\mathbf{x}^*) = 0, \\
&g_i(\mathbf{x}^*) \leq 0, \quad i = 1, \ldots, m, \\
&h_i(\mathbf{x}^*) = 0, \quad i = 1, \ldots, p, \\
&\lambda_i^* \geq 0, \quad i = 1, \ldots, m, \\
&\lambda_i^* g_i(\mathbf{x}^*) = 0, \quad i = 1, \ldots, m
\end{aligned}
$$

- The last condition is called complementary slackness, and guarantees the strong duality for convex optimization
- In Lecture 10, you'll learn how this condition is used to determine support vectors in SVM

# 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).