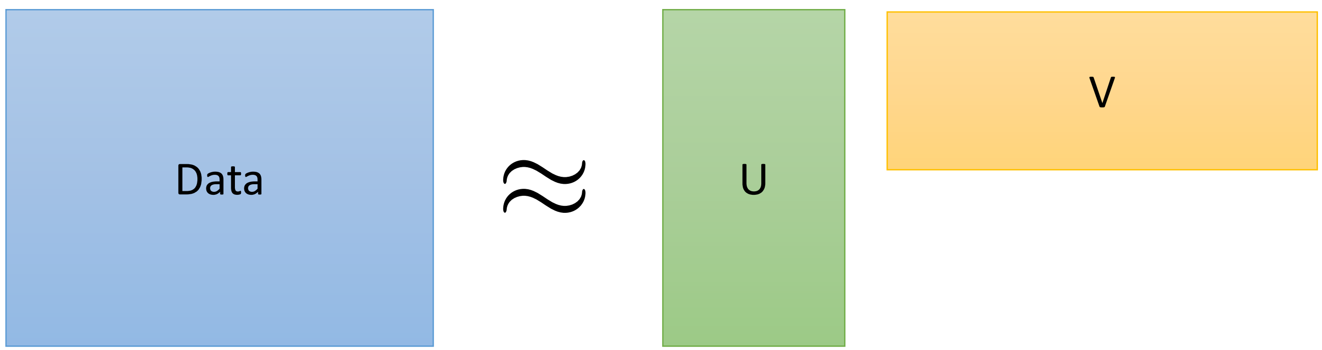
# Lecture 16 Duality and Support Vector Machines

Lei Li, **Yu-Xiang Wang**

(some slides from my convex optimization class,
originally taught by Ryan Tibshirani in CMU)

# Recap: Modeling by writing down an optimization problem

- Unsupervised learning as matrix factorization




  - Example: Principle Component Analysis
  - Example: Topic model with Latent Dirichlet Allocation
  - Example: Gaussian mixture model
  - Example: Movie recommendation
  - Example: Dictionary learning (sparse coding)
  - Example: Robust PCA

Does not have to be unsupervised…

# Recap: Structural inducing regularization and convex relaxation

- Sparsity

$$\|x\|_0 \qquad\qquad \|x\|_1$$

- Low-rank matrix with Nuclear norm regularization

$$\mathrm{rank}(X) \qquad\qquad \|X\|_*$$

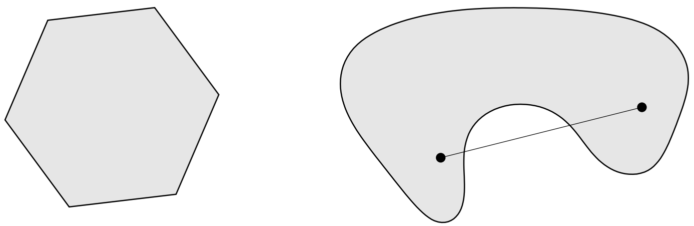
- Piecewise polynomials with a small number of pieces

$$\|D^{(k+1)} f\|_0 \qquad\qquad \|D^{(k+1)} f\|_1$$

# Recap: Convex Set and Functions

Convex set: $C \subseteq \mathbb{R}^n$ such that

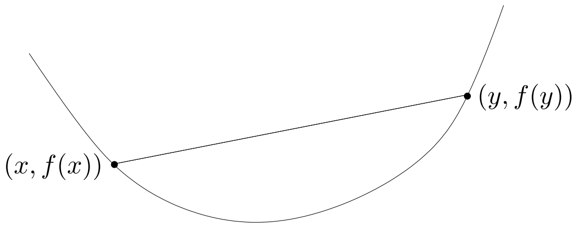
$$x, y \in C \implies tx + (1-t)y \in C \text{ for all } 0 \le t \le 1$$



Convex function: $f : \mathbb{R}^n \to \mathbb{R}$ such that $\mathrm{dom}(f) \subseteq \mathbb{R}^n$ convex, and

$$f(tx + (1-t)y) \le tf(x) + (1-t)f(y) \quad \text{for all } 0 \le t \le 1$$

and all $x, y \in \mathrm{dom}(f)$



# Recap: Convex optimization problem --- the standard form

Optimization problem:

$$
\begin{aligned}
\min_{x \in D} \quad & f(x) \\
\text{subject to} \quad & g_i(x) \leq 0,\ i = 1, \ldots m \\
& h_j(x) = 0,\ j = 1, \ldots r
\end{aligned}
$$

Here $D = \text{dom}(f) \cap \bigcap_{i=1}^{m} \text{dom}(g_i) \cap \bigcap_{j=1}^{p} \text{dom}(h_j)$, common domain of all the functions

This is a convex optimization problem provided the functions $f$ and $g_i, i = 1, \ldots m$ are convex, and $h_j, j = 1, \ldots p$ are affine:

$$h_j(x) = a_j^T x + b_j, \quad j = 1, \ldots p$$

# Recap: High school examples

$$\min_{x \in \mathbb{R}} x^2 - 4x + 9$$

$$\min_{x \in [0,1]} x^2 - 4x + 9$$

$$\min_{x \in \mathbb{R}} |x| - 4x + 9$$

$$\min_{x \in \mathbb{R}} \log(e^{5x+6} + e^{-8x+3})$$

# Why learning convex optimization when deep learning is non-convex?

- A lot of non-convex problems has a convex reformulation or convex relaxation
- Helpful in designing optimization algorithms for non-convex problems too.
- The technical training helps to develop skills that makes you a better researcher and more effective problem solver.

# Example: principal components analysis

Given $X \in \mathbb{R}^{n \times p}$, consider the low rank approximation problem:

$$\min_R \|X - R\|_F^2 \ \text{ subject to } \ \mathrm{rank}(R) = k$$

Here $\|A\|_F^2 = \sum_{i=1}^n \sum_{j=1}^p A_{ij}^2$, the entrywise squared $\ell_2$ norm, and $\mathrm{rank}(A)$ denotes the rank of $A$

Also called principal components analysis or PCA problem. Given $X = UDV^T$, singular value decomposition or SVD, the solution is

$$R = U_k D_k V_k^T$$

where $U_k, V_k$ are the first $k$ columns of $U, V$ and $D_k$ is the first $k$ diagonal elements of $D$. I.e., $R$ is reconstruction of $X$ from its first $k$ principal components

The PCA problem is not convex. Let's recast it. First rewrite as

$$\min_{Z \in \mathbb{S}^p} \|X - XZ\|_F^2 \;\; \text{subject to} \;\; \text{rank}(Z) = k, \;\; Z \text{ is a projection}$$

$$\iff \max_{Z \in \mathbb{S}^p} \text{tr}(SZ) \;\; \text{subject to} \;\; \text{rank}(Z) = k, \;\; Z \text{ is a projection}$$

where $S = X^T X$. Hence constraint set is the nonconvex set

$$C = \left\{ Z \in \mathbb{S}^p : \lambda_i(Z) \in \{0, 1\}, \; i = 1, \ldots p, \; \text{tr}(Z) = k \right\}$$

where $\lambda_i(Z)$, $i = 1, \ldots n$ are the eigenvalues of $Z$. Solution in this formulation is

$$Z = V_k V_k^T$$

where $V_k$ gives first $k$ columns of $V$

Now consider relaxing constraint set to $\mathcal{F}_k = \mathrm{conv}(C)$, its convex hull. Note

$$\begin{aligned}\mathcal{F}_k &= \{Z \in \mathbb{S}^p : \lambda_i(Z) \in [0,1],\ i = 1, \dots p,\ \mathrm{tr}(Z) = k\} \\ &= \{Z \in \mathbb{S}^p : 0 \preceq Z \preceq I,\ \mathrm{tr}(Z) = k\}\end{aligned}$$

This set is called the Fantope of order $k$. It is convex. Hence, the linear maximization over the Fantope, namely

$$\max_{Z \in \mathcal{F}_k} \ \mathrm{tr}(SZ)$$

is a convex problem. Remarkably, this is equivalent to the original nonconvex PCA problem (admits the same solution)!

(Famous result: Fan (1949), "On a theorem of Weyl conerning eigenvalues of linear transformations")

Ky Fan
樊㙨

1914 - 2010
UCSB Math
Professor

# Why is this useful? We already have Singular Value Decomposition!

## Sparse PCA with Fantope Projection and Selection

- Having an optimization formulation allows us to add additional problem specific considerations.
- Suppose we want the recovered principle components to be sparse

$$\max_{Z \in \mathcal{F}_k} \operatorname{tr}(SZ) - \lambda \sum_{i,j} |Z_{i,j}| \text{ subject to } \operatorname{rank}(R) = k$$

- This is the algorithm for the sparse PCA problem that achieves the minimax rate. (Vu and Lei, NIPS 2013).

# This lecture

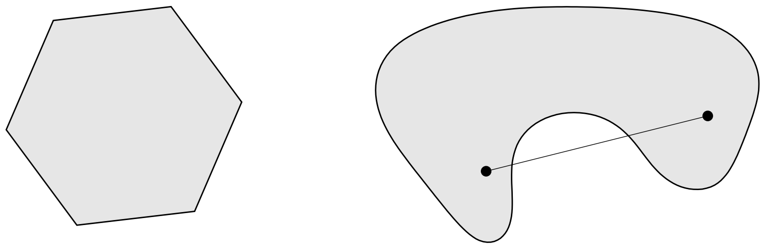
- Examples of convex sets / convex functions
- Duality
- Application to Support Vector Machines

# Convex sets

Convex set: $C \subseteq \mathbb{R}^n$ such that

$$x, y \in C \implies tx + (1-t)y \in C \text{ for all } 0 \leq t \leq 1$$

In words, line segment joining any two elements lies entirely in set



Convex combination of $x_1, \ldots x_k \in \mathbb{R}^n$: any linear combination

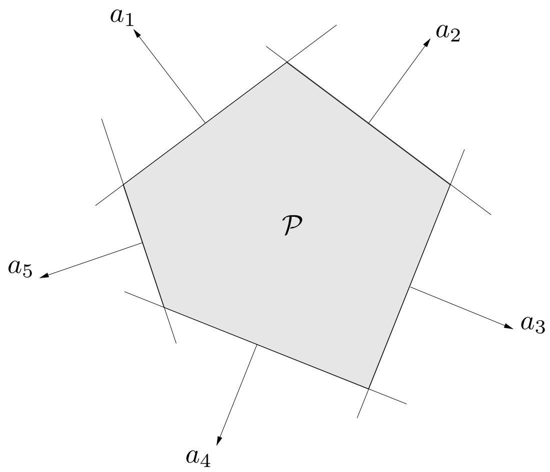
$$\theta_1 x_1 + \ldots + \theta_k x_k$$

with $\theta_i \geq 0$, $i = 1, \ldots k$, and $\sum_{i=1}^k \theta_i = 1$. Convex hull of a set $C$, $\mathrm{conv}(C)$, is all convex combinations of elements. Always convex

# Examples of convex sets

- Trivial ones: empty set, point, line
- Norm ball: $\{x : \|x\| \leq r\}$, for given norm $\|\cdot\|$, radius $r$
- Hyperplane: $\{x : a^T x = b\}$, for given $a, b$
- Halfspace: $\{x : a^T x \leq b\}$
- Affine space: $\{x : Ax = b\}$, for given $A, b$

- Polyhedron: $\{x : Ax \le b\}$, where inequality $\le$ is interpreted componentwise. Note: the set $\{x : Ax \le b, Cx = d\}$ is also a polyhedron (why?)




- Simplex: special case of polyhedra, given by $\mathrm{conv}\{x_0, \dots x_k\}$, where these points are affinely independent. The canonical example is the probability simplex,

$$\mathrm{conv}\{e_1, \dots e_n\} = \{w : w \ge 0,\ 1^T w = 1\}$$

# Operations preserving convexity

- Intersection: the intersection of convex sets is convex
- Scaling and translation: if $C$ is convex, then

$$aC + b = \{ax + b : x \in C\}$$

is convex for any $a, b$

- Affine images and preimages: if $f(x) = Ax + b$ and $C$ is convex then

$$f(C) = \{f(x) : x \in C\}$$

is convex, and if $D$ is convex then

$$f^{-1}(D) = \{x : f(x) \in D\}$$

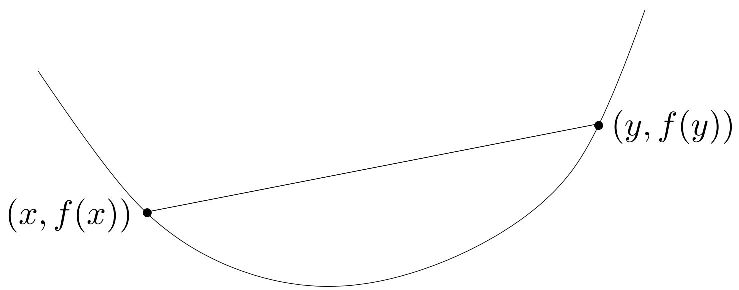
is convex

# Convex functions

Convex function: $f : \mathbb{R}^n \to \mathbb{R}$ such that $\mathrm{dom}(f) \subseteq \mathbb{R}^n$ convex, and

$$f(tx + (1-t)y) \le tf(x) + (1-t)f(y) \quad \text{for } 0 \le t \le 1$$

and all $x, y \in \mathrm{dom}(f)$




In words, function lies below the line segment joining $f(x)$, $f(y)$

Concave function: opposite inequality above, so that

$$f \text{ concave} \iff -f \text{ convex}$$

Important modifiers:

- Strictly convex: $f\big(tx + (1-t)y\big) < tf(x) + (1-t)f(y)$ for $x \neq y$ and $0 < t < 1$. In words, $f$ is convex and has greater curvature than a linear function

- Strongly convex with parameter $m > 0$: $f - \frac{m}{2}\|x\|_2^2$ is convex. In words, $f$ is at least as convex as a quadratic function

Note: strongly convex $\Rightarrow$ strictly convex $\Rightarrow$ convex

(Analogously for concave functions)

# Examples of convex functions

- Univariate functions:
  - Exponential function: $e^{ax}$ is convex for any $a$ over $\mathbb{R}$
  - Power function: $x^a$ is convex for $a \geq 1$ or $a \leq 0$ over $\mathbb{R}_+$ (nonnegative reals)
  - Power function: $x^a$ is concave for $0 \leq a \leq 1$ over $\mathbb{R}_+$
  - Logarithmic function: $\log x$ is concave over $\mathbb{R}_{++}$
- Affine function: $a^T x + b$ is both convex and concave
- Quadratic function: $\frac{1}{2} x^T Q x + b^T x + c$ is convex provided that $Q \succeq 0$ (positive semidefinite)
- Least squares loss: $\|y - Ax\|_2^2$ is always convex (since $A^T A$ is always positive semidefinite)

- Norm: $\|x\|$ is convex for any norm; e.g., $\ell_p$ norms,

$$\|x\|_p = \left(\sum_{i=1}^n x_i^p\right)^{1/p} \quad \text{for } p \geq 1, \quad \|x\|_\infty = \max_{i=1,\ldots n} |x_i|$$

and also operator (spectral) and trace (nuclear) norms,

$$\|X\|_{\text{op}} = \sigma_1(X), \quad \|X\|_{\text{tr}} = \sum_{i=1}^r \sigma_r(X)$$

where $\sigma_1(X) \geq \ldots \geq \sigma_r(X) \geq 0$ are the singular values of the matrix $X$

- Indicator function: if $C$ is convex, then its indicator function

$$I_C(x) = \begin{cases} 0 & x \in C \\ \infty & x \notin C \end{cases}$$

  is convex

- Support function: for any set $C$ (convex or not), its support function

$$I_C^*(x) = \max_{y \in C} \; x^T y$$

  is convex

- Max function: $f(x) = \max\{x_1, \ldots x_n\}$ is convex

# Key properties of convex functions

- A function is convex if and only if its restriction to any line is convex

- Epigraph characterization: a function $f$ is convex if and only if its epigraph

$$\mathrm{epi}(f) = \{(x, t) \in \mathrm{dom}(f) \times \mathbb{R} : f(x) \leq t\}$$

  is a convex set

- Convex sublevel sets: if $f$ is convex, then its sublevel sets

$$\{x \in \mathrm{dom}(f) : f(x) \leq t\}$$

  are convex, for all $t \in \mathbb{R}$. The converse is not true

- First-order characterization: if $f$ is differentiable, then $f$ is convex if and only if $\mathrm{dom}(f)$ is convex, and

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

  for all $x, y \in \mathrm{dom}(f)$. Therefore for a differentiable convex function $\nabla f(x) = 0 \iff x$ minimizes $f$

- Second-order characterization: if $f$ is twice differentiable, then $f$ is convex if and only if $\mathrm{dom}(f)$ is convex, and $\nabla^2 f(x) \succeq 0$ for all $x \in \mathrm{dom}(f)$

- Jensen's inequality: if $f$ is convex, and $X$ is a random variable supported on $\mathrm{dom}(f)$, then $f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)]$

# Operations preserving convexity

- Nonnegative linear combination: $f_1, \dots f_m$ convex implies $a_1 f_1 + \dots + a_m f_m$ convex for any $a_1, \dots a_m \geq 0$
- Pointwise maximization: if $f_s$ is convex for any $s \in S$, then $f(x) = \max_{s \in S} \; f_s(x)$ is convex. Note that the set $S$ here (number of functions $f_s$) can be infinite
- Partial minimization: if $g(x, y)$ is convex in $x, y$, and $C$ is convex, then $f(x) = \min_{y \in C} \; g(x, y)$ is convex

# Example: distances to a set

Let $C$ be an arbitrary set, and consider the maximum distance to $C$ under an arbitrary norm $\|\cdot\|$:

$$f(x) = \max_{y \in C} \|x - y\|$$

Let's check convexity: $f_y(x) = \|x - y\|$ is convex in $x$ for any fixed $y$, so by pointwise maximization rule, $f$ is convex

Now let $C$ be convex, and consider the minimum distance to $C$:

$$f(x) = \min_{y \in C} \|x - y\|$$

Let's check convexity: $g(x, y) = \|x - y\|$ is convex in $x, y$ jointly, and $C$ is assumed convex, so apply partial minimization rule

# More operations preserving convexity

- Affine composition: if $f$ is convex, then $g(x) = f(Ax + b)$ is convex
- General composition: suppose $f = h \circ g$, where $g : \mathbb{R}^n \to \mathbb{R}$, $h : \mathbb{R} \to \mathbb{R}$, $f : \mathbb{R}^n \to \mathbb{R}$. Then:
  - $f$ is convex if $h$ is convex and nondecreasing, $g$ is convex
  - $f$ is convex if $h$ is convex and nonincreasing, $g$ is concave
  - $f$ is concave if $h$ is concave and nondecreasing, $g$ concave
  - $f$ is concave if $h$ is concave and nonincreasing, $g$ convex

  How to remember these? Think of the chain rule when $n = 1$:

$$f''(x) = h''(g(x))g'(x)^2 + h'(g(x))g''(x)$$

- Vector composition: suppose that

$$f(x) = h\big(g(x)\big) = h\big(g_1(x), \ldots g_k(x)\big)$$

where $g : \mathbb{R}^n \to \mathbb{R}^k$, $h : \mathbb{R}^k \to \mathbb{R}$, $f : \mathbb{R}^n \to \mathbb{R}$. Then:

  - $f$ is convex if $h$ is convex and nondecreasing in each argument, $g$ is convex
  - $f$ is convex if $h$ is convex and nonincreasing in each argument, $g$ is concave
  - $f$ is concave if $h$ is concave and nondecreasing in each argument, $g$ is concave
  - $f$ is concave if $h$ is concave and nonincreasing in each argument, $g$ is convex

# Example: log-sum-exp function

Log-sum-exp function: $g(x) = \log(\sum_{i=1}^k e^{a_i^T x + b_i})$, for fixed $a_i, b_i$, $i = 1, \ldots k$. Often called "soft max", as it smoothly approximates $\max_{i=1,\ldots k} (a_i^T x + b_i)$

How to show convexity? First, note it suffices to prove convexity of $f(x) = \log(\sum_{i=1}^n e^{x_i})$ (affine composition rule)

Now use second-order characterization. Calculate

$$\nabla_i f(x) = \frac{e^{x_i}}{\sum_{\ell=1}^n e^{x_\ell}}$$
$$\nabla_{ij}^2 f(x) = \frac{e^{x_i}}{\sum_{\ell=1}^n e^{x_\ell}} 1\{i = j\} - \frac{e^{x_i} e^{x_j}}{(\sum_{\ell=1}^n e^{x_\ell})^2}$$

Write $\nabla^2 f(x) = \mathrm{diag}(z) - zz^T$, where $z_i = e^{x_i} / (\sum_{\ell=1}^n e^{x_\ell})$. This matrix is diagonally dominant, hence positive semidefinite

# Linear program

A linear program or LP is an optimization problem of the form

$$\begin{aligned} \min_x \quad & c^T x \\ \text{subject to} \quad & Dx \le d \\ & Ax = b \end{aligned}$$

Observe that this is always a convex optimization problem

- First introduced by Kantorovich in the late 1930s and Dantzig in the 1940s
- Dantzig's simplex algorithm gives a direct (noniterative) solver for LPs (later in the course we'll see interior point methods)
- Fundamental problem in convex optimization. Many diverse applications, rich history

# Examples of linear programs

## Example: diet problem

Find cheapest combination of foods that satisfies some nutritional requirements (useful for graduate students!)

$$\begin{aligned} \min_{x} \quad & c^T x \\ \text{subject to} \quad & Dx \geq d \\ & x \geq 0 \end{aligned}$$

Interpretation:

- $c_j$ : per-unit cost of food $j$
- $d_i$ : minimum required intake of nutrient $i$
- $D_{ij}$ : content of nutrient $i$ per unit of food $j$
- $x_j$ : units of food $j$ in the diet

## Example: transportation problem

Ship commodities from given sources to destinations at min cost

$$\begin{aligned} \min_{x} \quad & \sum_{i=1}^{m} \sum_{j=1}^{n} c_{ij} x_{ij} \\ \text{subject to} \quad & \sum_{j=1}^{n} x_{ij} \leq s_i, \; i = 1, \ldots, m \\ & \sum_{i=1}^{m} x_{ij} \geq d_j, \; j = 1, \ldots, n, \quad x \geq 0 \end{aligned}$$

Interpretation:

- $s_i$ : supply at source $i$
- $d_j$ : demand at destination $j$
- $c_{ij}$ : per-unit shipping cost from $i$ to $j$
- $x_{ij}$ : units shipped from $i$ to $j$

# Convex quadratic program

A convex quadratic program or QP is an optimization problem of the form

$$\begin{aligned} \min_x \quad & c^T x + \frac{1}{2} x^T Q x \\ \text{subject to} \quad & Dx \le d \\ & Ax = b \end{aligned}$$

where $Q \succeq 0$, i.e., positive semidefinite

Note that this problem is not convex when $Q \not\succeq 0$

From now on, when we say quadratic program or QP, we implicitly assume that $Q \succeq 0$ (so the problem is convex)

# Example: portfolio optimization

Construct a financial portfolio, trading off performance and risk:

$$\begin{array}{ll} \max\limits_{x} & \mu^T x - \frac{\gamma}{2} x^T Q x \\ \text{subject to} & 1^T x = 1 \\ & x \geq 0 \end{array}$$

Interpretation:

- $\mu$ : expected assets' returns
- $Q$ : covariance matrix of assets' returns
- $\gamma$ : risk aversion
- $x$ : portfolio holdings (percentages)

# Example: support vector machines

Given $y \in \{-1, 1\}^n$, $X \in \mathbb{R}^{n \times p}$ having rows $x_1, \ldots x_n$, recall the support vector machine or SVM problem:

$$\begin{aligned} \min_{\beta, \beta_0, \xi} \quad & \frac{1}{2}\|\beta\|_2^2 + C \sum_{i=1}^n \xi_i \\ \text{subject to} \quad & \xi_i \geq 0, \ i = 1, \ldots n \\ & y_i(x_i^T \beta + \beta_0) \geq 1 - \xi_i, \ i = 1, \ldots n \end{aligned}$$

This is a quadratic program

# Hierarchy of Canonical Optimizations

- Linear programs
- Quadratic programs
- Semidefinite programs
- Cone programs

# This lecture

- Examples of convex sets / convex functions
- Duality
- Application to Support Vector Machines

# Lower bounds in linear programs

Suppose we want to find lower bound on the optimal value in our convex problem, $B \le \min_x f(x)$

E.g., consider the following simple LP

$$\begin{aligned} \min_{x,y} \quad & x + y \\ \text{subject to} \quad & x + y \ge 2 \\ & x, y \ge 0 \end{aligned}$$

What's a lower bound? Easy, take $B = 2$

But didn't we get "lucky"?

Try again:

$$\min_{x,y} \quad x + 3y$$
$$\text{subject to} \quad x + y \geq 2$$
$$x, y \geq 0$$

$$\begin{aligned} & x + y \geq 2 \\ + \quad & 2y \geq 0 \\ = \quad & x + 3y \geq 2 \end{aligned}$$

Lower bound $B = 2$

More generally:

$$\min_{x,y} \quad px + qy$$
$$\text{subject to} \quad x + y \geq 2$$
$$x, y \geq 0$$

$$a + b = p$$
$$a + c = q$$
$$a, b, c \geq 0$$

Lower bound $B = 2a$, for any $a, b, c$ satisfying above

What's the best we can do? Maximize our lower bound over all possible $a, b, c$:

$$\begin{aligned} \min_{x,y} \quad & px + qy \\ \text{subject to} \quad & x + y \geq 2 \\ & x, y \geq 0 \end{aligned}$$

Called primal LP

$$\begin{aligned} \max_{a,b,c} \quad & 2a \\ \text{subject to} \quad & a + b = p \\ & a + c = q \\ & a, b, c \geq 0 \end{aligned}$$

Called dual LP

Note: number of dual variables is number of primal constraints

Try another one:

| Primal LP | Dual LP |
| --- | --- |
| $\min_{x,y} \quad px + qy$ | $\max_{a,b,c} \quad 2c - b$ |
| subject to $\quad x \geq 0$ | subject to $\quad a + 3c = p$ |
| $y \leq 1$ | $-b + c = q$ |
| $3x + y = 2$ | $a, b \geq 0$ |

Note: in the dual problem, $c$ is unconstrained

# Duality for general form LP

Given $c \in \mathbb{R}^n$, $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, $G \in \mathbb{R}^{r \times n}$, $h \in \mathbb{R}^r$:

| Primal LP | Dual LP |
|---|---|
| $\min_x \; c^T x$ subject to $Ax = b$, $Gx \le h$ | $\max_{u,v} \; -b^T u - h^T v$ subject to $-A^T u - G^T v = c$, $v \ge 0$ |

Explanation: for any $u$ and $v \ge 0$, and $x$ primal feasible,

$$u^T(Ax - b) + v^T(Gx - h) \le 0, \quad \text{i.e.,}$$
$$(-A^T u - G^T v)^T x \ge -b^T u - h^T v$$

So if $c = -A^T u - G^T v$, we get a bound on primal optimal value

# Another perspective on LP duality

$$\begin{array}{ll} \min\limits_{x} & c^T x \\ \text{subject to} & Ax = b \\ & Gx \leq h \end{array}$$

Primal LP

$$\begin{array}{ll} \max\limits_{u,b} & -b^T u - h^T v \\ \text{subject to} & -A^T u - G^T v = c \\ & v \geq 0 \end{array}$$

Dual LP

Explanation # 2: for any $u$ and $v \geq 0$, and $x$ primal feasible

$$c^T x \geq c^T x + u^T (Ax - b) + v^T (Gx - h) := L(x, u, v)$$

So if $C$ denotes primal feasible set, $f^\star$ primal optimal value, then for any $u$ and $v \geq 0$,

$$f^\star \ \geq \ \min_{x \in C} \ L(x, u, v) \ \geq \ \min_{x} \ L(x, u, v) \ := \ g(u, v)$$

In other words, $g(u,v)$ is a lower bound on $f^\star$ for any $u$ and $v \geq 0$

Note that

$$g(u,v) = \begin{cases} -b^T u - h^T v & \text{if } c = -A^T u - G^T v \\ -\infty & \text{otherwise} \end{cases}$$

Now we can maximize $g(u,v)$ over $u$ and $v \geq 0$ to get the tightest bound, and this gives exactly the dual LP as before

This last perspective is actually completely general and applies to arbitrary optimization problems (even nonconvex ones)

# Lagrangian

Consider general minimization problem

$$\begin{aligned} \min_x \quad & f(x) \\ \text{subject to} \quad & h_i(x) \le 0, \ i = 1, \dots m \\ & \ell_j(x) = 0, \ j = 1, \dots r \end{aligned}$$

Need not be convex, but of course we will pay special attention to convex case

We define the Lagrangian as

$$L(x, u, v) = f(x) + \sum_{i=1}^{m} u_i h_i(x) + \sum_{j=1}^{r} v_j \ell_j(x)$$

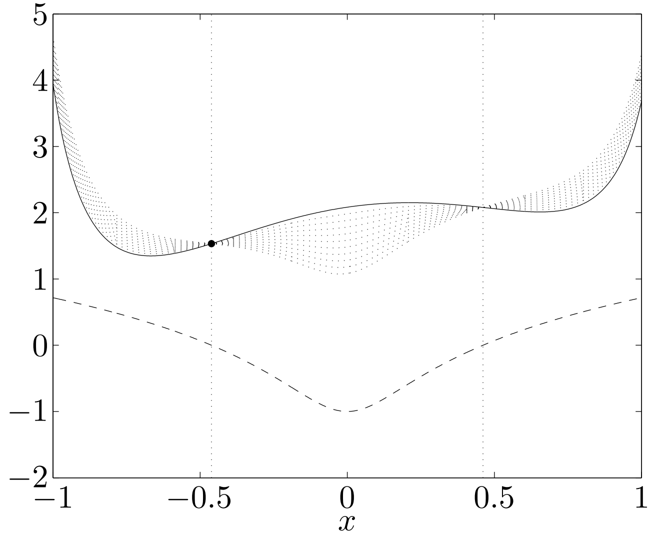
New variables $u \in \mathbb{R}^m, v \in \mathbb{R}^r$, with $u \ge 0$ (implicitly, we define $L(x, u, v) = -\infty$ for $u < 0$)

Important property: for any $u \geq 0$ and $v$,

$$f(x) \geq L(x,u,v) \quad \text{at each feasible } x$$

Why? For feasible $x$,

$$L(x,u,v) = f(x) + \sum_{i=1}^{m} u_i \underbrace{h_i(x)}_{\leq 0} + \sum_{j=1}^{r} v_j \underbrace{\ell_j(x)}_{=0} \leq f(x)$$




- Solid line is $f$
- Dashed line is $h$, hence feasible set $\approx [-0.46, 0.46]$
- Each dotted line shows $L(x,u,v)$ for different choices of $u \geq 0$

(From B & V page 217)

# Lagrange dual function

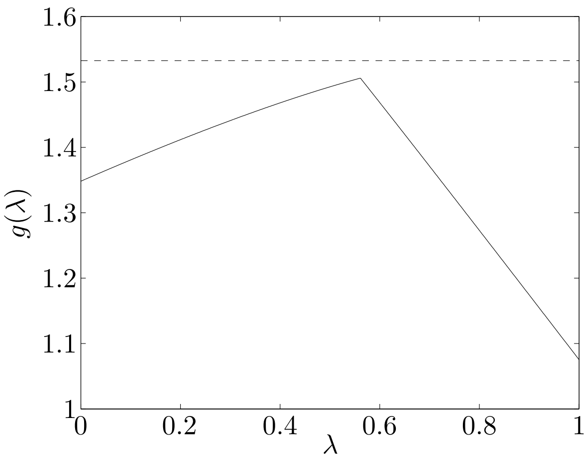
Let $C$ denote primal feasible set, $f^\star$ denote primal optimal value. Minimizing $L(x,u,v)$ over all $x$ gives a lower bound:

$$f^\star \;\geq\; \min_{x\in C}\, L(x,u,v) \;\geq\; \min_{x}\, L(x,u,v) \;:=\; g(u,v)$$

We call $g(u,v)$ the Lagrange dual function, and it gives a lower bound on $f^\star$ for any $u \geq 0$ and $v$, called dual feasible $u,v$

- Dashed horizontal line is $f^\star$
- Dual variable $\lambda$ is (our $u$)
- Solid line shows $g(\lambda)$

(From B & V page 217)



# Lagrange dual problem

Given primal problem

$$\begin{aligned} \min_x \quad & f(x) \\ \text{subject to} \quad & h_i(x) \leq 0, \ i = 1, \ldots m \\ & \ell_j(x) = 0, \ j = 1, \ldots r \end{aligned}$$

Our constructed dual function $g(u, v)$ satisfies $f^\star \geq g(u, v)$ for all $u \geq 0$ and $v$. Hence best lower bound is given by maximizing $g(u, v)$ over all dual feasible $u, v$, yielding Lagrange dual problem:

$$\begin{aligned} \max_{u,v} \quad & g(u, v) \\ \text{subject to} \quad & u \geq 0 \end{aligned}$$

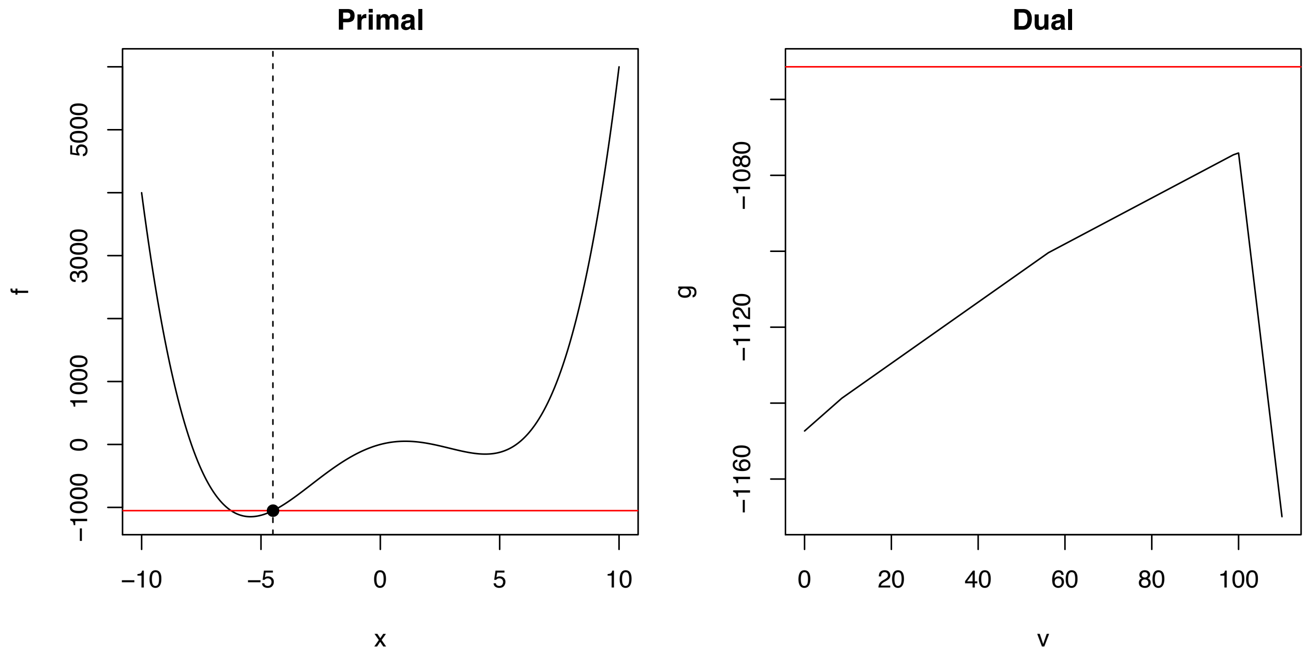
Key property, called weak duality: if dual optimal value is $g^\star$, then

$$f^\star \geq g^\star$$

Note that this always holds (even if primal problem is nonconvex)

# Example: nonconvex quartic minimization

Define $f(x) = x^4 - 50x^2 + 100x$ (nonconvex), minimize subject to constraint $x \geq -4.5$




Dual function $g$ can be derived explicitly, via closed-form equation for roots of a cubic equation

Form of $g$ is rather complicated:

$$g(u) = \min_{i=1,2,3} \left\{ F_i^4(u) - 50F_i^2(u) + 100F_i(u) \right\},$$

where for $i = 1, 2, 3$,

$$F_i(u) = \frac{-a_i}{12 \cdot 2^{1/3}} \left( 432(100-u) - \left(432^2(100-u)^2 - 4 \cdot 1200^3\right)^{1/2} \right)^{1/3}$$

$$-100 \cdot 2^{1/3} \frac{1}{\left( 432(100-u) - \left(432^2(100-u)^2 - 4 \cdot 1200^3\right)^{1/2} \right)^{1/3}},$$

and $a_1 = 1$, $a_2 = (-1 + i\sqrt{3})/2$, $a_3 = (-1 - i\sqrt{3})/2$

Without the context of duality it would be difficult to tell whether or not $g$ is concave ... but we know it must be!

# Strong duality

Recall that we always have $f^\star \geq g^\star$ (weak duality). On the other hand, in some problems we have observed that actually

$$f^\star = g^\star$$

which is called strong duality

Slater's condition: if the primal is a convex problem (i.e., $f$ and $h_1, \ldots h_m$ are convex, $\ell_1, \ldots \ell_r$ are affine), and there exists at least one strictly feasible $x \in \mathbb{R}^n$, meaning

$$h_1(x) < 0, \ldots h_m(x) < 0 \quad \text{and} \quad \ell_1(x) = 0, \ldots \ell_r(x) = 0$$

then strong duality holds

This is a pretty weak condition. An important refinement: strict inequalities only need to hold over functions $h_i$ that are not affine

# This lecture

- Examples of convex sets / convex functions
- Duality
- Application to Support Vector Machines

# Example: support vector machine dual

Given $y \in \{-1, 1\}^n$, $X \in \mathbb{R}^{n \times p}$, rows $x_1, \ldots x_n$, recall the support vector machine problem:

$$\begin{aligned} \min_{\beta, \beta_0, \xi} \quad & \frac{1}{2}\|\beta\|_2^2 + C\sum_{i=1}^n \xi_i \\ \text{subject to} \quad & \xi_i \geq 0, \; i = 1, \ldots n \\ & y_i(x_i^T\beta + \beta_0) \geq 1 - \xi_i, \; i = 1, \ldots n \end{aligned}$$

Introducing dual variables $v, w \geq 0$, we form the Lagrangian:

$$\begin{aligned} L(\beta, \beta_0, \xi, v, w) = \frac{1}{2}\|\beta\|_2^2 + C\sum_{i=1}^n \xi_i - \sum_{i=1}^n v_i\xi_i + \\ \sum_{i=1}^n w_i\big(1 - \xi_i - y_i(x_i^T\beta + \beta_0)\big) \end{aligned}$$

Minimizing over $\beta, \beta_0, \xi$ gives Lagrange dual function:

$$g(v, w) = \begin{cases} -\frac{1}{2} w^T \tilde{X} \tilde{X}^T w + 1^T w & \text{if } w = C1 - v, \ w^T y = 0 \\ -\infty & \text{otherwise} \end{cases}$$

where $\tilde{X} = \mathrm{diag}(y) X$. Thus SVM dual problem, eliminating slack variable $v$, becomes

$$\begin{array}{ll} \max\limits_{w} & -\frac{1}{2} w^T \tilde{X} \tilde{X}^T w + 1^T w \\ \text{subject to} & 0 \leq w \leq C1, \ w^T y = 0 \end{array}$$

Check: Slater's condition is satisfied, and we have strong duality. Further, from study of SVMs, might recall that at optimality

$$\beta = \tilde{X}^T w$$

This is not a coincidence, as we'll later via the KKT conditions

# Next lecture

- KKT conditions (with examples in SVM)

- Online Learning