Optimization Models

EECS 127 / EECS 227AT

Laurent El Ghaoui

EECS department
UC Berkeley

Fall 2018

LECTURE 19

Strong Duality

At once he became an enigma. One side or the other of his nature was perfectly comprehensible; but both sides together were bewildering.

Jack London

Outline

1. Strong duality for convex problems
   - Slater's condition for strong duality
   - Geometry
   - Recovering a primal solution from the dual
   - Duality in unconstrained problems

2. Strong duality in min-max problems
   - Minimax theorem
   - Geometry

3. Examples
   - Square-root LASSO
   - Strong duality in zero-sum games
   - Logistic regression

A convex problem

We now focus on a *convex* problem:

$$p^* = \min_{x \in \mathbb{R}^n} \; f_0(x) \text{ subject to: } \quad f_i(x) \le 0, \quad i = 1, \ldots, m$$
$$Ax = b,$$

where

- $f_0, \ldots, f_m$ are convex functions;
- the equality constraints are affine, and represented via the matrix $A \in \mathbb{R}^{q \times n}$ and vector $b \in \mathbb{R}^q$;
- we denote the problem's domain, *i.e.* the set of points for which all the functions involved are well-defined, by $\mathcal{D}$.

The Lagrangian is the function $\mathcal{L} : \mathcal{D} \times \mathbb{R}^m_+ \times \mathbb{R}^q \to \mathbb{R}$, with values

$$\mathcal{L}(x, \lambda, \nu) = f_0(x) + \sum_{i=1}^m \lambda_i f_i(x) + \nu^\top (Ax - b).$$

The problem's domain and its relative interior

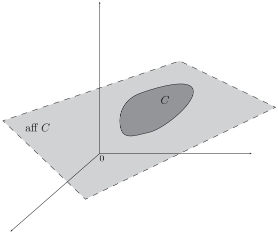
For simplicity, here we assume that domain of the problem is a polytope, described as

$$\mathcal{D} = \{x \in \mathbb{R}^n \ : \ Cx \leq d, \ \ Fx = g\},$$

where C, F, d, g are matrices and vectors of appropriate sizes, with F full row rank. We define the relative interior of the domain as:

$$\operatorname{relint} \mathcal{D} = \{x \in \mathbb{R}^n \ : \ Cx < d, \ \ Fx = g\},$$

We assume that the relative interior is not empty: $\operatorname{relint} \mathcal{D} \neq \emptyset$.



Example: the probability simplex $\mathcal{P} = \{p \ : \ \mathbf{1}^\top p = 1, \ p \geq 0\}$ has relative interior $\operatorname{relint} \mathcal{P} = \{p \ : \ \mathbf{1}^\top p = 1, \ p > 0\}$.

Strong duality for convex problems

We say that the problem is strictly feasible if there exist a point $x_0 \in \mathrm{relint}\,\mathcal{D}$ such that

$$Ax_0 = b, \quad f_i(x_0) < 0, \quad i = 1, \ldots, m.$$

Intuitively: all the inequalities are satisfied strictly, including the ones that are implicit (related to the problem's domain).

Proposition 1 (Slater's conditions for convex programs)

If the problem is strictly feasible, then strong duality holds: $p^* = d^*$.

To illustrate, consider the problem

$$p^* = \min_x \ f_0(x) \ : \ f_1(x) \leq 0.$$

with f_0, f_1 convex, and assume that the problem is strictly feasible (there exist $x_0 \in \mathrm{relint}\,\mathcal{D}$ such that $f_1(x_0) < 0$).

Geometry

Why strong duality holds

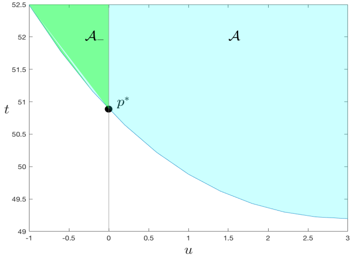
Recall from lecture 18 the 2D set of achievable values:

$$\mathcal{A} = \left\{ (u,t) \in \mathbb{R}^2 \ : \ \exists\, x \in \mathcal{D}, \ \ u \geq f_1(x), \ \ t \geq f_0(x) \right\}.$$

We can visualize the problem as a 2D problem:

$$p^* = \min_{u,t} \ t \ : \ (u,t) \in \mathcal{A}, \ \ u \leq 0.$$

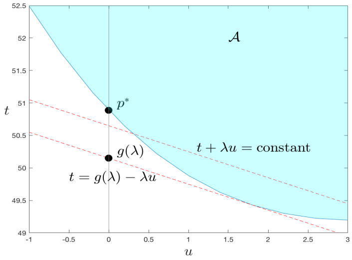
Slater condition implies that the set $\mathcal{A}$ has a non-empty intersection $\mathcal{A}_-$ with the left-half plane defined by $u < 0$.



Sets $\mathcal{A}$, $\mathcal{A}_-$ for a problem of the form

$$p^* = \min_x \ \|Ax - b\|_2 \ : \ \|x\|_2 \leq 1.$$

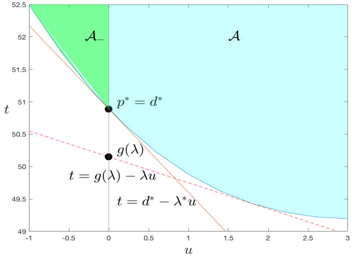
Dual problem



For a given $\lambda > 0$, computing the dual function amounts to solve

$$\begin{aligned} g(\lambda) &= \min_x \|Ax - b\|_2 + \lambda \|x\|_2 \\ &= \min_{(u,t) \in \mathcal{A}} t + \lambda u. \end{aligned}$$

By construction, the intercept $g(\lambda) \leq p^*$.



Optimizing over the slope $-\lambda < 0$ of the line we can close the duality gap $p^* - g(\lambda)$.

Recovering a primal solution from the dual

Assume that strong duality holds. Assume that we have solved the dual and computed (λ^*, ν^*). How can we recover a primal optimal point x^*, or decide the primal problem is not attained?

- A "natural" candidate would be any point in the set of x-minimizers of the Lagrangian $x \to \mathcal{L}(x, \lambda^*, \nu^*)$. In general it is not true that any such candidate is even feasible for the primal problem.
- A particular case arises when $\mathcal{L}(x, \lambda^*, \nu^*)$ has an *unique* minimizer. In this case, it can be shown that the unique minimizer x^* of $\mathcal{L}(\cdot, \lambda^*, \nu^*)$ is either primal feasible, and then it is the primal-optimal solution, or it is not primal-feasible, and then we can conclude that the no primal-optimal solution exists.
- In lecture 18, we saw an example of this (projection on the probability simplex).

Example

Minimum-norm solution to linear equations

Consider the problem (with $A \in \mathbb{R}^{m,n}$ full row rank, $b \in \mathbb{R}^m$ given)

$$p^* = \min_x \; \frac{1}{2}\|x\|_2^2 \; : \; Ax = b,$$

Since $AA^\top \succ 0$, we have $b \in \mathcal{R}(A)$, and the problem satisfies Slater's condition[1].

- Lagrangian: $\mathcal{L}(x,\nu) = \frac{1}{2}\|x\|_2^2 + \nu^\top(b - Ax)$.
- The function $\mathcal{L}(\cdot,\nu$ has the unique minimizer $x^*(\nu) = A^\top \nu$.
- Dual function: $g(\nu) = \min_x \; \mathcal{L}(x,\nu) = \nu^\top b - \frac{1}{2}\|A^\top \nu\|_2^2$.
- Dual problem:

$$d^* = \max_\nu \; \nu^\top b - \frac{1}{2}\|A^\top \nu\|_2^2.$$

[1] If A is full row rank then its column space is the whole space, so that $b \in \mathcal{R}(A)$ for any b.

Example

Minimum-norm solution to linear equations (cont'd)

We can solve the dual problem: (recall $AA^\top \succ 0$)

$$d^* = \max_\nu \; \nu^\top b - \frac{1}{2}\|A^\top \nu\|_2^2 = \frac{1}{2} b^\top (AA^\top)^{-1} b,$$

with unique maximizer $\nu^* = (AA^\top)^{-1} b$. Set

$$x^* = \arg\min_x \; \mathcal{L}(x, \nu^*) = A^\top \nu^* = A^\top (AA^\top)^{-1} b.$$

Since $Ax^* = b$, x^* is feasible. We can check directly what the theory tells us: that the feasible point x^* indeed achieves the lower bound d^*, so it is optimal:

$$\frac{1}{2}\|x^*\|_2^2 = \frac{1}{2} b^\top (AA^\top)^{-1} b.$$

Duality in unconstrained problems

In many applications (*e.g.* machine learning) we want to solve an unconstrained problem, for example LASSO

$$p^* = \min_w \ \frac{1}{2}\|Ax - b\|_2^2 + \lambda\|x\|_1,$$

where $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, $\lambda \geq 0$ are given.

Applying duality we get that the dual function is a constant (there are no dual variables), equal to p^* itself. Here, duality does not bring any information.

One way out is to add variables and constraints:

$$p^* = \min_{x,z} \ \frac{1}{2}\|z\|_2^2 + \lambda\|x\|_1 \ : \ z = Ax - b.$$

The problem satisfies Slater's condition, since the equality constraints are satisfied for some pair (x_0, z_0).

Dual of LASSO (cont'd)

By strong duality:

$$\begin{aligned} p^* &= \min_{x,z} \max_{\nu} \frac{1}{2}\|z\|_2^2 + \lambda\|x\|_1 + \nu^\top(z - Ax + b) \\ &= d^* \doteq \max_{\nu} g(\nu), \end{aligned}$$

where

$$\begin{aligned} g(\nu) &\doteq \min_{x,z} \frac{1}{2}\|z\|_2^2 + \lambda\|x\|_1 + \nu^\top(z - Ax + b) \\ &= \nu^\top b + \min_{x} \lambda\|x\|_1 - x^\top(A^\top\nu) + \min_{z} \frac{1}{2}\|z\|_2^2 + \nu^\top z \\ &= \nu^\top b - \frac{1}{2}\|\nu\|_2^2 \text{ if } \|A^\top\nu\|_\infty \le \lambda, \ +\infty \text{ otherwise.} \end{aligned}$$

We get the dual:

$$d^* = \max_{\nu} \nu^\top b - \frac{1}{2}\|\nu\|_2^2 \ : \ \|A^\top\nu\|_\infty \le \lambda.$$

Min-max problems

Many optimization problems can be written in min-max form:

$$p^* = \min_{x \in X} \max_{y \in Y} F(x, y).$$

for appropriate sets X, Y and function F.

We always have

$$p^* \geq d^* := \max_{y \in Y} \min_{x \in X} F(x, y).$$

Under certain assumptions, detailed later, equality holds.

Sion's minimax theorem

Theorem 1 (Minimax theorem (simplified))

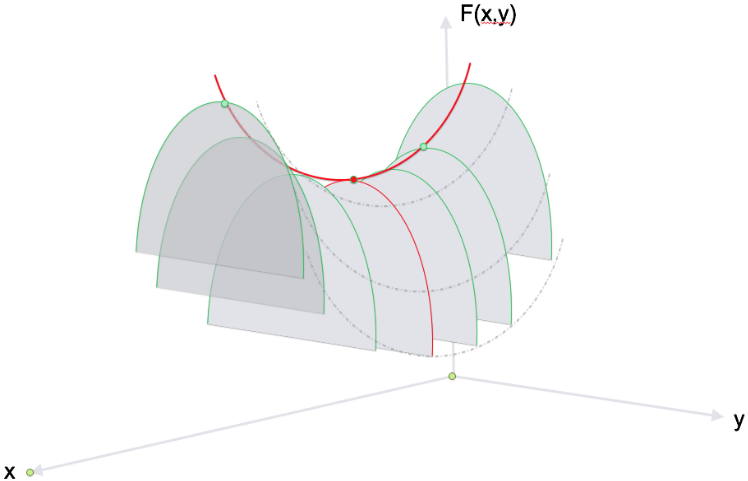
Let $X \subseteq \mathbb{R}^n$ be convex, and let $Y \subseteq \mathbb{R}^m$ be a compact set Let $F : X \times Y \to \mathbb{R}$ be a function such that for every $y \in Y$, $F(\cdot, y)$ is convex and continuous over X, and for every $x \in X$, $F(x, \cdot)$ is concave and continuous over Y. Then*

$$\max_{y \in Y} \min_{x \in X} F(x, y) = \min_{x \in X} \max_{y \in Y} F(x, y).$$

* *Compact set:* a set that is bounded and closed (contains the limit of any converging sequence of points inside it). Example includes norm balls, such as $\{y \in \mathbb{R}^m : \|y\|_p \leq 1\}$, with $p \in \{1, 2, \infty\}$; and the probability simplex $\Delta^m \doteq \{y \in \mathbb{R}^m_+ : \mathbf{1}^\top y = 1\}$.

Geometry

A convex-concave function and its saddle point



Example

Square-root LASSO

$$p^* = \min_x \; \|Ax - b\|_2 + \mu\|x\|_1,$$

where $A \in \mathbb{R}^{m,n}$, $b \in \mathbb{R}^m$, and $\mu > 0$.

Min-max form: from (generalized) Cauchy-Schwartz

$$p^* = \min_x \; \max_{(u,v)\in\mathcal{U}} \; u^\top(b - Ax) + v^\top x$$

where

$$\mathcal{U} := \left\{(u, v) \in \mathbb{R}^m \times \mathbb{R}^n \; : \; \|u\|_2 \le 1, \;\; \|v\|_\infty \le \mu\right\}.$$

Square-root LASSO (cont'd)

Sion's theorem applies, and leads to

$$p^* = \max_{(u,v)\in\mathcal{U}} \min_x \; u^\top(b - Ax) + v^\top x.$$

The inner problem is very simple to solve:

$$g(u,v) \doteq \min_x \; u^\top(b - Ax) + v^\top x = \begin{cases} b^\top u & \text{if } A^\top u = v, \\ -\infty & \text{otherwise.} \end{cases}$$

Hence

$$p^* = \max_{(u,v)\in\mathcal{U}} \; b^\top u \;:\; A^\top u = v.$$

Eliminating v:

$$p^* = \max_u \; b^\top u \;:\; \|u\|_2 \le 1, \;\; \|A^\top u\|_\infty \le \mu.$$

Safe feature elimination (SAFE)

Let a_i denote the i-th column of matrix A; a_i corresponds to the i-th feature (variable x_i) in the original problem.

Primal problem:

$$p^* = \min_x \ \|Ax - b\|_2 + \mu\|x\|_1,$$

Dual problem:

$$p^* = \max_u \ b^\top u \ : \ \|u\|_2 \le 1, \ \ |a_i^\top u| \le \mu, \ \ i = 1, \ldots, n.$$

We observe that if $\|a_i\|_2 < \mu$, then

$$|a_i^\top u| < \mu \text{ for all } u \ : \ \|u\|_2 \le 1.$$

This means that i-constraint is not active in the dual problem: its value remains the same if we remove that constraint.

SAFE

Removing the i-th constraint from the dual really amounts to solving the original problem with that feature removed. The previous statement says that in that process we do not change the value of the problem. This implies that we can safely set $x_i = 0$.

The result is that a quick computation (computing the norms of the colums of A) potentially allows to eliminate features from the data, leading to computational savings.

Strong duality in zero-sum games

Revisiting the game from Lecture 18

7	-8	-7	-8	3	5
9	-5	10	-2	-10	5
-8	1	10	9	7	-2
9	10	0	6	9	3
3	10	6	10	4	-7

$n \times m$ payoff matrix.

If the minimizing (column) player plays first, the value of the game is

$$p^* = \min_j \max_i M_{ij} = \mathbf{3}.$$

Otherwise, it is lower (better for the column, minimizing player):

$$d^* = \max_i \min_j M_{ij} = \mathbf{0}.$$

Can the first player "close the gap"?

Idea

Playing a random strategy

Assume the minimizing (column) player plays first, and that the game is played many times.

In an attempt to confuse the other player, the minimizing player plays according to a randomized strategy, parametrized by a probability distribution, *i.e.* a vector x in the probability simplex

$$\Delta^m \doteq \left\{ x \in \mathbb{R}^n_+ \ : \ \mathbf{1}^\top x = 1 \right\}.$$

For $j \in \{1, \ldots, m\}$, x_j is the probability of selecting column j. We assume that the maximizing player does the same, with a randomized strategy $y \in \Delta^n$. Note that randomized strategies include "pure" ones as special case, so the second player has nothing to loose in using such a strategy. For given $x \in \Delta^n$, $y \in \Delta^m$, the expected payoff is

$$\sum_{i,j} y_i M_{ij} x_j = x^\top M y.$$

If the minimizing (column) player plays first, it should choose its randomized strategy according to

$$p^*_{\mathrm{rand}} = \min_{x \in \Delta^m} \max_{y \in \Delta^n} \ y^\top M x.$$

Strong duality result

Applying Sion's minimax theorem gives us

$$p^*_{\mathrm{rand}} = \min_{x \in \Delta^m} \max_{y \in \Delta^n} y^\top M x = \max_{y \in \Delta^n} \min_{x \in \Delta^m} y^\top M x = d^*_{\mathrm{rand}}.$$

With a randomized strategy, the order of play is irrelevant.

We can also take the dual of the inner problem; strong duality holds via Slater's condition. Hence, for a given x:

$$\max_{y \in \Delta^n} y^\top M x = \min_t \ t \ : \ t \geq (Mx)_i, \ \ i = 1, \ldots, m.$$

Letting x be variable leads to the LP expression for the value of the game:

$$p^*_{\mathrm{rand}} = d^*_{\mathrm{rand}} = \min_{x,t} \ t \ : \ t \cdot \mathbf{1} \geq Mx, \ \ x \geq 0, \ \ \mathbf{1}^\top x = 1.$$

Example

Solving the LP with the game payoff data given, we obtain

$$d^* = 0 < p^*_{\rm rand} = d^*_{\rm rand} = 2.1525 < 3 = p^*,$$

so the minimizing player looses less with the randomized strategy than with a pure one. The second player achieves the best average payoff, given an adversary that plays random.

The solutions for both players are obtained by noticing that the variable y is dual to the constraint[2] $t \cdot \mathbf{1} \geq Mx$. Alternatively we can solve the dual LP to get y^*:

$$p^*_{\rm rand} = d^*_{\rm rand} = \max_{y,t}\ t \ :\ t \cdot \mathbf{1} \leq M^\top y, \ \ y \geq 0, \ \ \mathbf{1}^\top y = 1,$$

which leads to

$$\begin{array}{rcl} y^* & = & \left(\begin{array}{cccc} 0.4517 & 0.1714 & 0.33480 & 0.0420 \end{array} \right)^\top, \\ x^* & = & \left(\begin{array}{cccccc} 0.2568 & 0 & 0.2246 & 0 & 0.3332 & 0.1854 \end{array} \right)^\top. \end{array}$$

[2] Dual variables can be provided as part of the solution in CVX.

Logistic regression

Recall the logistic regression problem in lecture 17:

$$p^* = \min_{w,b} \; -\log L(w,b) \doteq -\sum_{i=1}^{m} \log\left(1 + \exp(-y_i(w^\top x_i + b))\right),$$

where $X = [x_1, \ldots, x_m]$ is the data matrix (each column x_i a data point) and $y \in \{-1, 1\}^m$ is the label vector.

If we try Lagrange duality directly, since there are no constraints, we get that the dual function is a constant, equal to p^* — not a very useful dual!

We can get a meaningful dual by introducing new variables and equality constraints.

Introducing equality variables

Assuming that the bias term b is zero for simplicity, we have

$$p^* = \min_{w,v} \sum_{i=1}^{m} f(v_i) \ : \ v = A^\top w,$$

where $A = [y_1 x_1, \ldots, y_m x_m] \in \mathbb{R}^{n \times m}$, and f is the convex function, with values for $\xi \in \mathbb{R}$ given by

$$f(\xi) \doteq \log(1 + e^{-\xi}).$$

The problem has (feasible) linear equality constraints; thanks to Slater's conditions for strong duality, we can write $p^* = d^* = \max_\nu \ g(\nu)$, where g is the dual function

$$g(\nu) \doteq \min_{w,v} \sum_{i=1}^{m} f(v_i) + \nu^\top (A^\top w - v).$$

The minimization over w results in $-\infty$ if $A\nu \neq 0$, 0 otherwise; the one over v can be done in closed form, as seen next ...

Dual problem

With the convention $0 \log 0 = 0$: for any $\alpha \in \mathbb{R}$,

$$\min_{\xi} \ f(\xi) - \alpha\xi = \begin{cases} -\alpha \log \alpha - (1-\alpha)\log(1-\alpha) & \text{if } 0 \leq \alpha \leq 1, \\ +\infty & \text{otherwise.} \end{cases}$$

We obtain a dual form of the problem:

$$p^* = d^* = \max_{\nu} \ g(\nu) \ : \ \nu \in [0,1]^m, \ \ A\nu = 0,$$

where g is the (concave) "entropy" function, with values for $\nu \in [0,1]^m$ given by

$$g(\nu) \doteq -\sum_{i=1}^{m} \left(\nu_i \log \nu_i + (1-\nu_i)\log(1-\nu_i)\right).$$