Convex Optimization for Statistics and Machine Learning

Part III: Duality and Optimality

Ryan Tibshirani

Depts. of Statistics & Machine Learning
Carnegie Mellon University

http://www.stat.cmu.edu/~ryantibs/talks/cuso-part3-2019.pdf

Outline for Part III

- Part A. Linear program duality
- Part B. Lagrangian duality
- Part C. KKT optimality conditions
- Part D. Duality correspondences

Part III: Duality and optimality

A. Linear program duality

Lower bounds in linear programs

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

For example, consider the following simple LP

$$\begin{array}{ll} \min\limits_{x,y} & x + y \\ \text{subject to} & x + y \geq 2 \\ & x, y \geq 0 \end{array}$$

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

But didn't we get "lucky"?

Try again:

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

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

Lower bound $B = 2$

More generally:

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

$$\begin{aligned} a + b &= p \\ a + c &= q \\ a, b, c &\geq 0 \end{aligned}$$

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:

$$\begin{array}{ll} \min\limits_{x,y} & px + qy \\ \text{subject to} & x \geq 0 \\ & y \leq 1 \\ & 3x + y = 2 \end{array}$$

Primal LP

$$\begin{array}{ll} \max\limits_{a,b,c} & 2c - b \\ \text{subject to} & a + 3c = p \\ & -b + c = q \\ & a, b \geq 0 \end{array}$$

Dual LP

Note: in the dual problem, c is unconstrained

Duality for general form LP

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

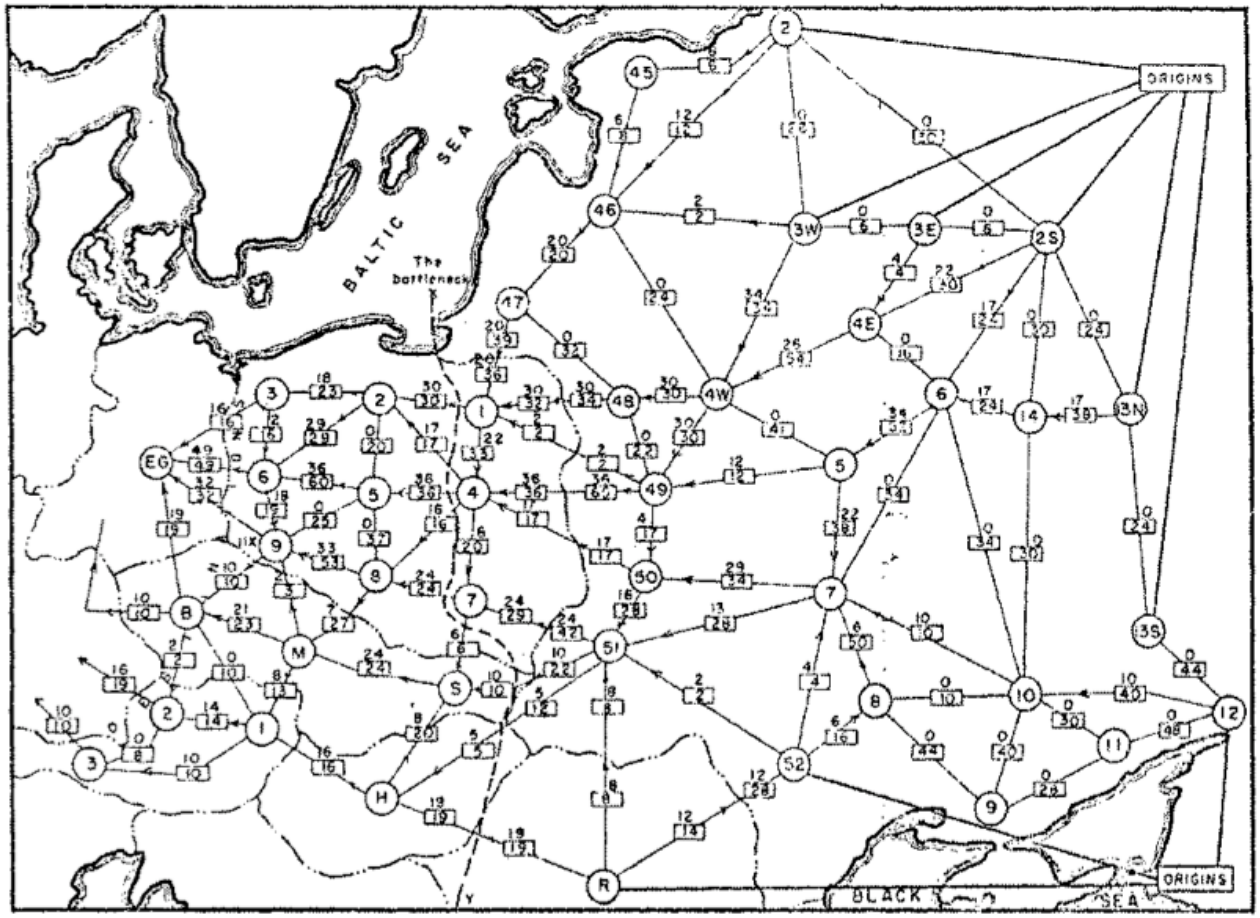
Primal LP	Dual LP
$\min_x \quad c^T x$	$\max_{u,v} \quad -b^T u - h^T v$
subject to $\quad Ax = b$	subject to $\quad -A^T u - G^T v = c$
$\quad Gx \leq h$	$\quad v \geq 0$

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

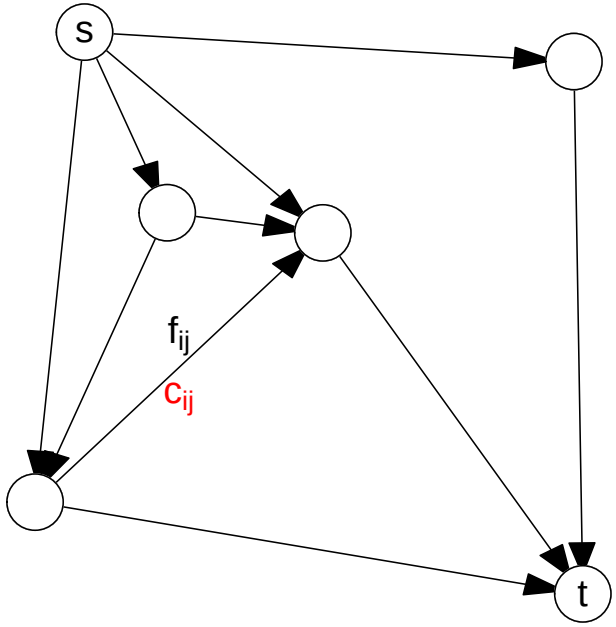
$$u^T(Ax - b) + v^T(Gx - h) \leq 0, \quad \text{i.e.,}$$
$$(-A^T u - G^T v)^T x \geq -b^T u - h^T v$$

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

Example: max flow and min cut



Soviet railway network (from Schrijver (2002), "On the history of transportation and maximum flow problems")



Given graph $G = (V, E)$, define flow f_{ij}, $(i, j) \in E$ to satisfy:

- $f_{ij} \geq 0$, $(i, j) \in E$
- $f_{ij} \leq c_{ij}$, $(i, j) \in E$
- $\sum_{(i,k) \in E} f_{ik} = \sum_{(k,j) \in E} f_{kj}$, $k \in V \setminus \{s, t\}$

Max flow problem: find flow that maximizes total value of the flow from s to t. That is, as an LP:

$$\max_{f \in \mathbb{R}^{|E|}} \quad \sum_{(s,j) \in E} f_{sj}$$

$$\text{subject to} \quad 0 \leq f_{ij} \leq c_{ij} \quad \text{for all } (i, j) \in E$$

$$\sum_{(i,k) \in E} f_{ik} = \sum_{(k,j) \in E} f_{kj} \quad \text{for all } k \in V \setminus \{s, t\}$$

Derive the dual, in steps:

- Note that

$$\sum_{(i,j)\in E} \Big(- a_{ij} f_{ij} + b_{ij}(f_{ij} - c_{ij}) \Big) + \sum_{k \in V \setminus \{s,t\}} x_k \Big(\sum_{(i,k)\in E} f_{ik} - \sum_{(k,j)\in E} f_{kj} \Big) \leq 0$$

  for any $a_{ij}, b_{ij} \geq 0$, $(i,j) \in E$, and x_k, $k \in V \setminus \{s,t\}$

- Rearrange as

$$\sum_{(i,j)\in E} M_{ij}(a,b,x) f_{ij} \leq \sum_{(i,j)\in E} b_{ij} c_{ij}$$

  where $M_{ij}(a,b,x)$ collects terms multiplying f_{ij}

- Want to make LHS in previous inequality equal to primal objective, i.e., $\begin{cases} M_{sj} = b_{sj} - a_{sj} + x_j & \text{want this} = 1 \\ M_{it} = b_{it} - a_{it} - x_i & \text{want this} = 0 \\ M_{ij} = b_{ij} - a_{ij} + x_j - x_i & \text{want this} = 0 \end{cases}$
- We've shown that

$$\text{primal optimal value} \leq \sum_{(i,j)\in E} b_{ij} c_{ij},$$

subject to a, b, x satisfying constraints. Hence dual problem is (minimize over a, b, x to get best upper bound):

$$\begin{aligned} \min_{b\in\mathbb{R}^{|E|},\, x\in\mathbb{R}^{|V|}} \quad & \sum_{(i,j)\in E} b_{ij} c_{ij} \\ \text{subject to} \quad & b_{ij} + x_j - x_i \geq 0 \quad \text{for all } (i,j) \in E \\ & b \geq 0,\ x_s = 1,\ x_t = 0 \end{aligned}$$

Suppose that at the solution, it just so happened that

$$x_i \in \{0, 1\} \quad \text{for all } i \in V$$

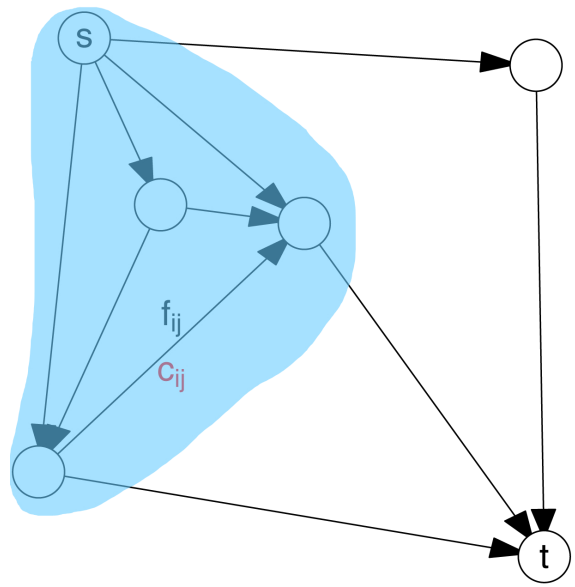
Let $A = \{i : x_i = 1\}$, $B = \{i : x_i = 0\}$; note $s \in A$, $t \in B$. Then

$$b_{ij} \geq x_i - x_j \quad \text{for } (i, j) \in E, \quad b \geq 0$$

imply that $b_{ij} = 1$ if $i \in A$ and $j \in B$, and 0 otherwise. Moreover, the objective $\sum_{(i,j)\in E} b_{ij} c_{ij}$ is the capacity of cut defined by A, B

That is, we've argued that the dual is the LP relaxation of the min cut problem:

$$\begin{aligned} \min_{b \in \mathbb{R}^{|E|},\, x \in \mathbb{R}^{|V|}} \quad & \sum_{(i,j)\in E} b_{ij} c_{ij} \\ \text{subject to} \quad & b_{ij} \geq x_i - x_j \\ & b_{ij}, x_i, x_j \in \{0, 1\} \\ & \text{for all } i, j \end{aligned}$$



Therefore, from what we know so far:

$$\text{value of max flow} \leq \text{optimal value for LP relaxed min cut} \leq \text{capacity of min cut}$$

Famous result, called max flow min cut theorem: value of max flow through a network is exactly the capacity of the min cut

Hence in the above, we get all equalities. In particular, we get that the primal LP and dual LP have exactly the same optimal values, a phenomenon called strong duality

How often does this happen? More on this soon

Part III: Duality and optimality

B. Lagrangian duality

Another perspective on LP duality

Primal LP	Dual LP
$\min_x \quad c^T x$	$\max_{u,b} \quad -b^T u - h^T v$
subject to $\quad Ax = b$	subject to $\quad -A^T u - G^T v = c$
$\quad\quad Gx \leq h$	$\quad\quad v \geq 0$

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 latest perspective is actually completely general and applies to arbitrary optimization problems

Lagrangian

Consider general minimization problem

$$\begin{array}{ll} \min\limits_{x} & f(x) \\ \text{subject to} & h_i(x) \leq 0,\ i = 1, \ldots, m \\ & \ell_j(x) = 0,\ j = 1, \ldots, r \end{array}$$

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

We define the Lagrangian as

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

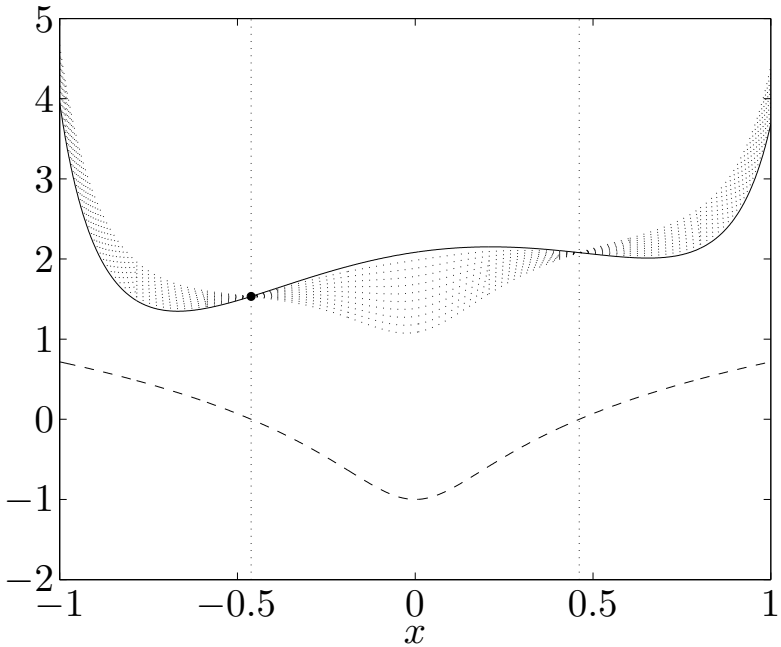
New variables $u \in \mathbb{R}^m, v \in \mathbb{R}^r$, with $u \geq 0$ (else $L(x, u, v) = -\infty$)

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

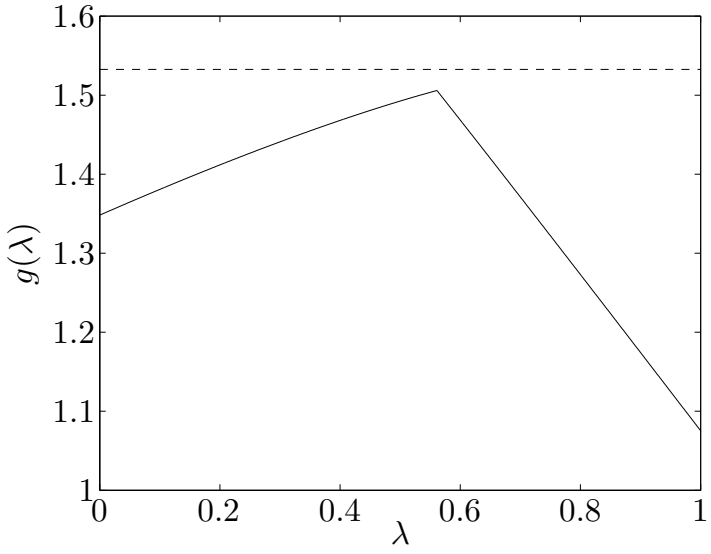
Let C denote primal feasible set, $f^\star$ denote primal optimal value. Minimizing $L(x, u, v)$ over all x gives a lower bound:

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

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 λ is (our u)
- Solid line shows $g(\lambda)$

(From B & V page 217)



Example: quadratic program

Consider quadratic program:

$$\begin{array}{ll} \min\limits_{x} & \frac{1}{2}x^TQx + c^Tx \\ \text{subject to} & Ax = b,\ x \geq 0 \end{array}$$

where $Q \succ 0$. Lagrangian:

$$L(x,u,v) = \frac{1}{2}x^TQx + c^Tx - u^Tx + v^T(Ax-b)$$

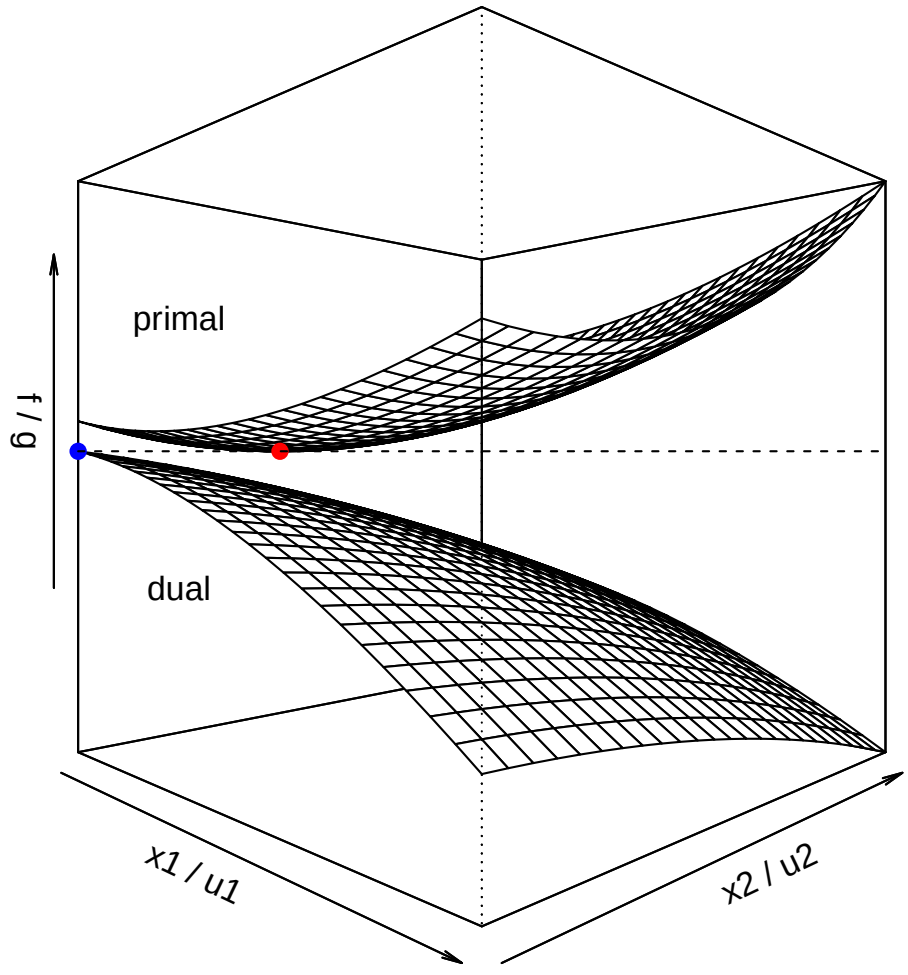
Lagrange dual function:

$$g(u,v) = \min_x\ L(x,u,v) = -\frac{1}{2}(c-u+A^Tv)^TQ^{-1}(c-u+A^Tv) - b^Tv$$

For any $u \geq 0$ and any v, this lower bounds primal optimal value $f^\star$

Example: quadratic program in 2D

We choose $f(x)$ to be quadratic in 2 variables, subject to $x \geq 0$. Dual function $g(u)$ is also quadratic in 2 variables, also subject to $u \geq 0$



Dual function $g(u)$ provides a bound on $f^\star$ for every $u \geq 0$

Largest bound this gives us: turns out to be exactly $f^\star$ ... coincidence?

More on this later, via KKT conditions

Lagrange dual problem

Given primal problem

$$\begin{array}{ll} \min\limits_{x} & f(x) \\ \text{subject to} & h_i(x) \leq 0,\ i = 1, \ldots, m \\ & \ell_j(x) = 0,\ j = 1, \ldots, r \end{array}$$

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

$$\begin{array}{ll} \max\limits_{u,v} & g(u, v) \\ \text{subject to} & u \geq 0 \end{array}$$

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)

Another key property: the dual problem is a convex optimization problem (as written, it is a concave maximization problem)

Again, this is always true (even when primal problem is not convex)

By definition:

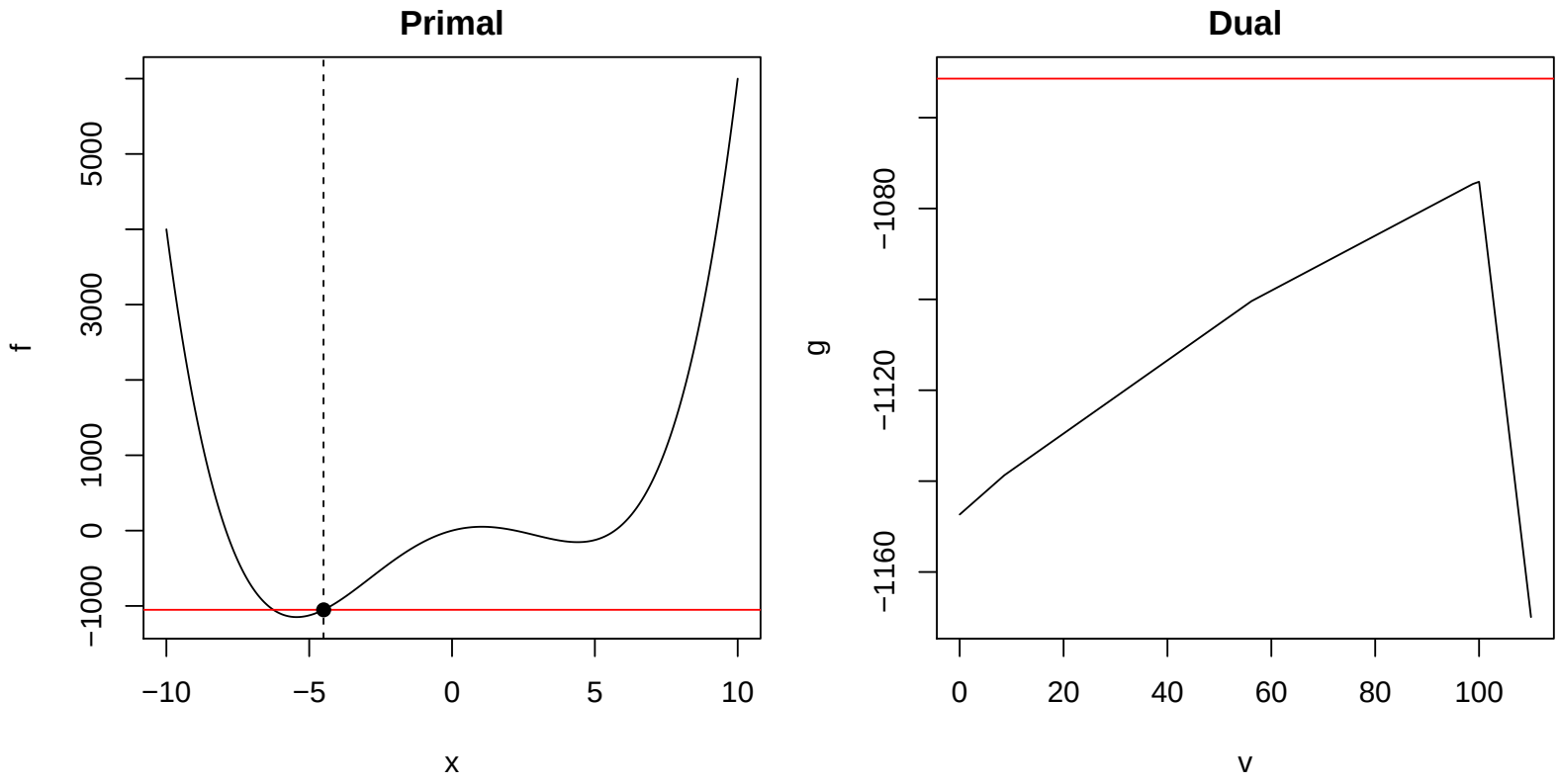
$$g(u,v) = \min_x \left\{ f(x) + \sum_{i=1}^m u_i h_i(x) + \sum_{j=1}^r v_j \ell_j(x) \right\}$$

$$= \underbrace{- \max_x \left\{ -f(x) - \sum_{i=1}^m u_i h_i(x) - \sum_{j=1}^r v_j \ell_j(x) \right\}}_{\text{pointwise maximum of convex functions in } (u,v)}$$

That is, g is concave in (u, v), and $u \geq 0$ is a convex constraint, hence dual problem is a concave maximization problem

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

Refinement: actually only need strict inequalities for non-affine h_i

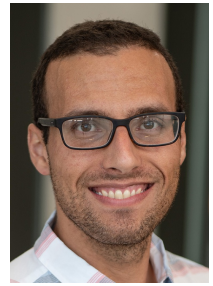
LPs: back to where we started

For linear programs:

- Easy to check that the dual of the dual LP is the primal LP
- Refined version of Slater's condition: strong duality holds for an LP if it is feasible
- Apply same logic to its dual LP: strong duality holds if it is feasible
- Hence strong duality holds for LPs, except when both primal and dual are infeasible

In other words, we nearly always have strong duality for LPs

Example: mixed strategies for matrix games

Setup: two players,  vs.  , and a payout matrix P

		R			
		1	2	$\dots$	n
A	1	P_{11}	P_{12}	$\dots$	P_{1n}
	2	P_{21}	P_{22}	$\dots$	P_{2n}
	$\dots$				
	m	P_{m1}	P_{m2}	$\dots$	P_{mn}

Game: if A chooses i and R chooses j, then A must pay R amount P_{ij} (don't feel bad for A—this can be positive or negative)

They use mixed strategies, i.e., each will first specify a probability distribution, and then

$$x: \quad \mathbb{P}(\text{A chooses } i) = x_i,\ i = 1, \dots, m$$
$$y: \quad \mathbb{P}(\text{R chooses } j) = y_j,\ j = 1, \dots, n$$

The expected payout then, from A to R, is

$$\sum_{i=1}^{m}\sum_{j=1}^{n} x_i y_j P_{ij} = x^T P y$$

Now suppose that, because A is wiser, he will allow R to know his strategy x ahead of time. In this case, R will choose y to maximize $x^T P y$, which results in A paying off

$$\max\left\{x^T P y:\ y \geq 0,\ 1^T y = 1\right\} = \max_{i=1,\ldots,n} (P^T x)_i$$

A's best strategy is then to choose his distribution x according to

$$\begin{array}{ll} \min\limits_{x} & \max\limits_{i=1,\ldots,n} (P^T x)_i \\ \text{subject to} & x \geq 0,\ 1^T x = 1 \end{array}$$

In an alternate universe, if R were somehow wiser than A, then he might allow A to know his strategy y beforehand

By the same logic, R's best strategy is to choose his distribution y according to

$$\begin{array}{ll} \max\limits_{y} & \min\limits_{j=1,\ldots,m} (Py)_j \\ \text{subject to} & y \geq 0,\ 1^T y = 1 \end{array}$$

Call R's expected payout in first scenario $f_1^\star$, and expected payout in second scenario $f_2^\star$. Because it is clearly advantageous to know the other player's strategy, $f_1^\star \geq f_2^\star$

But by Von Neumman's minimax theorem: we know that $f_1^\star = f_2^\star$ ... which may come as a surprise!

Recast first problem as an LP:

$$\begin{aligned} &\min_{x,t} \\ &\text{subject to} \quad x \geq 0,\ 1^T x = 1 \\ &\qquad\qquad\quad P^T x \leq t \end{aligned}$$

Now form the Lagrangian:

$$L(x,t,u,v,y) = t - u^T x + v(1 - 1^T x) + y^T(P^T x - t1)$$

and the Lagrange dual function:

$$\begin{aligned} g(u,v,y) &= \min_{x,t}\ L(x,t,u,v,y) \\ &= \begin{cases} v & \text{if } 1 - 1^T y = 0,\ Py - u - v1 = 0 \\ -\infty & \text{otherwise} \end{cases} \end{aligned}$$

Hence dual problem, after eliminating slack variable u, is

$$\begin{array}{ll} \max\limits_{y,v} & v \\ \text{subject to} & y \geq 0,\ 1^T y = 1 \\ & Py \geq v \end{array}$$

This is exactly the second problem

Strong duality holds because both primal and dual are feasible (we only need one). Thus von Neumann's minimax theorem is a direct consequence of LP duality

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 or SVM problem:

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

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 β, β_0, ξ gives Lagrange dual function:

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

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

$$\begin{aligned} \max_w \quad & -\frac{1}{2} w^T \tilde{X} \tilde{X}^T w + 1^T w \\ \text{subject to} \quad & 0 \le w \le C1,\ w^T y = 0 \end{aligned}$$

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 see via the KKT conditions

Duality gap

Given primal feasible x and dual feasible u, v, the quantity

$$f(x) - g(u, v)$$

is called the duality gap between x and u, v. Note that

$$f(x) - f^\star \leq f(x) - g(u, v)$$

so if the duality gap is zero, then x is primal optimal (and similarly, u, v are dual optimal)

Also from algorithmic viewpoint, provides a stopping criterion: if $f(x) - g(u, v) \leq \epsilon$, then we are guaranteed that $f(x) - f^\star \leq \epsilon$

Part III: Duality and optimality

C. KKT optimality conditions

Karush-Kuhn-Tucker conditions

Given general problem

$$\begin{array}{ll} \min\limits_{x} & f(x) \\ \text{subject to} & h_i(x) \leq 0,\ i = 1, \ldots, m \\ & \ell_j(x) = 0,\ j = 1, \ldots, r \end{array}$$

The Karush-Kuhn-Tucker conditions or KKT conditions are:

- $0 \in \partial\left(f(x) + \sum_{i=1}^{m} u_i h_i(x) + \sum_{j=1}^{r} v_j \ell_j(x) \right)$ (stationarity)
- $u_i \cdot h_i(x) = 0$ for all i (complementary slackness)
- $h_i(x) \leq 0,\ \ell_j(x) = 0$ for all i, j (primal feasibility)
- $u_i \geq 0$ for all i (dual feasibility)

Necessity

Let $x^\star$ and $u^\star, v^\star$ be primal and dual solutions with zero duality gap (strong duality holds, e.g., under Slater's condition). Then

$$\begin{aligned} f(x^\star) &= g(u^\star, v^\star) \\ &= \min_x \; f(x) + \sum_{i=1}^m u_i^\star h_i(x) + \sum_{j=1}^r v_j^\star \ell_j(x) \\ &\leq f(x^\star) + \sum_{i=1}^m u_i^\star h_i(x^\star) + \sum_{j=1}^r v_j^\star \ell_j(x^\star) \\ &\leq f(x^\star) \end{aligned}$$

In other words, all these inequalities are actually equalities

Two things to learn from this:

- The point $x^\star$ minimizes $L(x, u^\star, v^\star)$ over $x \in \mathbb{R}^n$. Hence the subdifferential of $L(x, u^\star, v^\star)$ must contain 0 at $x = x^\star$—this is exactly the stationarity condition
- We must have $\sum_{i=1}^m u_i^\star h_i(x^\star) = 0$, and since each term here is ≤ 0, this implies $u_i^\star h_i(x^\star) = 0$ for every i—this is exactly complementary slackness

Primal and dual feasibility hold by virtue of optimality. Therefore:

> If $x^\star$ and $u^\star, v^\star$ are primal and dual solutions, with zero duality gap, then $x^\star, u^\star, v^\star$ satisfy the KKT conditions

(Note that this statement assumes nothing a priori about convexity of our problem, i.e., of f, h_i, ℓ_j)

Sufficiency

If there exists $x^\star, u^\star, v^\star$ that satisfy the KKT conditions, then

$$\begin{aligned} g(u^\star, v^\star) &= f(x^\star) + \sum_{i=1}^{m} u_i^\star h_i(x^\star) + \sum_{j=1}^{r} v_j^\star \ell_j(x^\star) \\ &= f(x^\star) \end{aligned}$$

where the first equality holds from stationarity, and the second holds from complementary slackness

Therefore the duality gap is zero (and $x^\star$ and $u^\star, v^\star$ are primal and dual feasible) so $x^\star$ and $u^\star, v^\star$ are primal and dual optimal. Hence, we've shown:

> If $x^\star$ and $u^\star, v^\star$ satisfy the KKT conditions, then $x^\star$ and $u^\star, v^\star$ are primal and dual solutions

Putting it together

In summary, KKT conditions are equivalent to zero duality gap:

- always sufficient
- necessary under strong duality

Putting it together:

For a problem with strong duality (e.g., assume Slater's condition: convex problem and there exists x strictly satisfying non-affine inequality contraints),

$$x^\star \text{ and } u^\star, v^\star \text{ are primal and dual solutions}$$
$$\iff x^\star \text{ and } u^\star, v^\star \text{ satisfy the KKT conditions}$$

(Warning, concerning the stationarity condition: for a differentiable function f, we cannot use $\partial f(x) = \{\nabla f(x)\}$ unless f is convex!)

What's in a name?

Older folks will know these as the KT (Kuhn-Tucker) conditions:

- First appeared in publication by Kuhn and Tucker in 1951
- Later people found out that Karush had the conditions in his unpublished master's thesis of 1939

For unconstrained problems, the KKT conditions are nothing more than the subgradient optimality condition

For general convex problems, the KKT conditions could have been derived entirely from studying optimality via subgradients

$$0 \in \partial f(x^\star) + \sum_{i=1}^m \mathcal{N}_{\{h_i \leq 0\}}(x^\star) + \sum_{j=1}^r \mathcal{N}_{\{\ell_j = 0\}}(x^\star)$$

where recall $\mathcal{N}_C(x)$ is the normal cone of C at x

Example: quadratic with equality constraints

Consider for $Q \succeq 0$,

$$\begin{array}{ll} \min\limits_{x} & \frac{1}{2} x^T Q x + c^T x \\ \text{subject to} & Ax = 0 \end{array}$$

(For example, this corresponds to Newton step for the constrained problem $\min_x \ f(x) \ \text{ subject to } \ Ax = b$)

Convex problem, no inequality constraints, so by KKT conditions: x is a solution if and only if

$$\begin{bmatrix} Q & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} x \\ u \end{bmatrix} = \begin{bmatrix} -c \\ 0 \end{bmatrix}$$

for some u. Linear system combines stationarity, primal feasibility (complementary slackness and dual feasibility are vacuous)

Example: support vector machines

Given $y \in \{-1, 1\}^n$, and $X \in \mathbb{R}^{n \times p}$, back to the SVM problem:

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

Introduce dual variables $v, w \geq 0$. KKT stationarity condition:

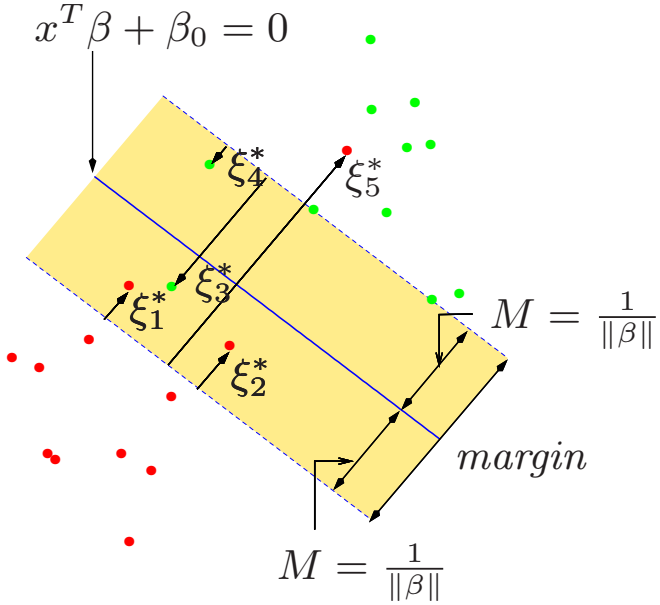
$$0 = \sum_{i=1}^n w_i y_i, \quad \beta = \sum_{i=1}^n w_i y_i x_i, \quad w = C1 - v$$

Complementary slackness:

$$v_i \xi_i = 0, \ w_i\big(1 - \xi_i - y_i(x_i^T\beta + \beta_0)\big) = 0, \quad i = 1, \ldots, n$$

Hence at optimality we have $\beta = \sum_{i=1}^{n} w_i y_i x_i$, and w_i is nonzero only if $y_i(x_i^T \beta + \beta_0) = 1 - \xi_i$. Such points i are called the support points

- For support point i, if $\xi_i = 0$, then x_i lies on edge of margin, and $w_i \in (0, C]$;
- For support point i, if $\xi_i \neq 0$, then x_i lies on wrong side of margin, and $w_i = C$



KKT conditions do not really give us a way to find solution, but gives a better understanding

In fact, we can use this to screen away non-support points before performing optimization

Constrained and Lagrange forms

Often in statistics and machine learning we'll switch back and forth between constrained form, where $t \in \mathbb{R}$ is a tuning parameter,

$$\min_x \ f(x) \ \text{ subject to } \ h(x) \leq t \tag{C}$$

and Lagrange form, where $\lambda \geq 0$ is a tuning parameter,

$$\min_x \ f(x) + \lambda \cdot h(x) \tag{L}$$

and claim these are equivalent. Is this true (assuming convex f, h)?

(C) to (L): if problem (C) is strictly feasible, then strong duality holds, and there exists some $\lambda \geq 0$ (dual solution) such that any solution $x^\star$ in (C) minimizes

$$f(x) + \lambda \cdot (h(x) - t)$$

so $x^\star$ is also a solution in (L)

(L) to (C): if $x^\star$ is a solution in (L), then the KKT conditions for (C) are satisfied by taking $t = h(x^\star)$, so $x^\star$ is a solution in (C)

Conclusion:

$$\bigcup_{\lambda \geq 0} \{\text{solutions in (L)}\} \quad \subseteq \quad \bigcup_{t} \{\text{solutions in (C)}\}$$

$$\bigcup_{\lambda \geq 0} \{\text{solutions in (L)}\} \quad \supseteq \quad \bigcup_{\substack{t \text{ such that (C)} \\ \text{is strictly feasible}}} \{\text{solutions in (C)}\}$$

This is nearly a perfect equivalence. Note: when the only value of t that leads to a feasible but not strictly feasible constraint set is $t = 0$, then we do get perfect equivalence

So, e.g., if $h \geq 0$, and (C), (L) are feasible for all $t, \lambda \geq 0$, then we do get perfect equivalence

Uniqueness in ℓ_1 penalized problems

Using the KKT conditions and simple probability arguments, we have the following (perhaps surprising) result:[1]

Theorem: Let f be differentiable and strictly convex, let $X \in \mathbb{R}^{n \times p}$, $\lambda > 0$. Consider

$$\min_{\beta} \; f(X\beta) + \lambda \|\beta\|_1$$

If the entries of X are drawn from a continuous probability distribution (on $\mathbb{R}^{np}$), then w.p. 1 there is a unique solution and it has at most $\min\{n, p\}$ nonzero components

Remark: here f must be strictly convex, but no restrictions on the dimensions of X (we could have $p \gg n$)

[1] For example, Tibshirani (2013), "The lasso problem and uniqueness"

Part III: Duality and optimality

D. Duality correspondences

Back to duality

A key use of duality: under strong duality, can characterize primal solutions from dual solutions

Recall that under strong duality, the KKT conditions are necessary for optimality. Given dual solutions $u^\star, v^\star$, any primal solution $x^\star$ satisfies the stationarity condition

$$0 \in \partial f(x^\star) + \sum_{i=1}^m u_i^\star \partial h_i(x^\star) + \sum_{j=1}^r v_i^\star \partial \ell_j(x^\star)$$

In other words, $x^\star$ solves $\min_x \; L(x, u^\star, v^\star)$

In particular, if this is satisfied uniquely (above problem has unique minimizer), then corresponding point must be the primal solution ... very useful when dual is easier to solve than primal

When is dual easier?

Key facts about primal-dual relationship:

- Dual has complementary number of variables: recall, number of primal constraints
- Dual involves complementary norms: $\|\cdot\|$ becomes $\|\cdot\|_*$
- Dual has "identical" smoothness: L/m (Lipschitz constant of gradient by strong convexity parameter) is unchanged between f and its conjugate f^*
- Dual can "shift" linear transformations between terms ... this leads to key idea: dual decomposition

Dual norms

Let $\|x\|$ be a norm, e.g.,

- ℓ_p norm: $\|x\|_p = (\sum_{i=1}^n |x_i|^p)^{1/p}$, for $p \geq 1$
- Trace norm: $\|X\|_{\mathsf{tr}} = \sum_{i=1}^r \sigma_i(X)$

We define its dual norm $\|x\|_*$ as

$$\|x\|_* = \max_{\|z\| \leq 1} \; z^T x$$

Gives us the inequality $|z^T x| \leq \|z\| \|x\|_*$ (like generalized Holder).
Back to our examples,

- ℓ_p norm dual: $(\|x\|_p)_* = \|x\|_q$, where $1/p + 1/q = 1$
- Trace norm dual: $(\|X\|_{\mathsf{tr}})_* = \|X\|_{\mathsf{op}} = \sigma_1(X)$

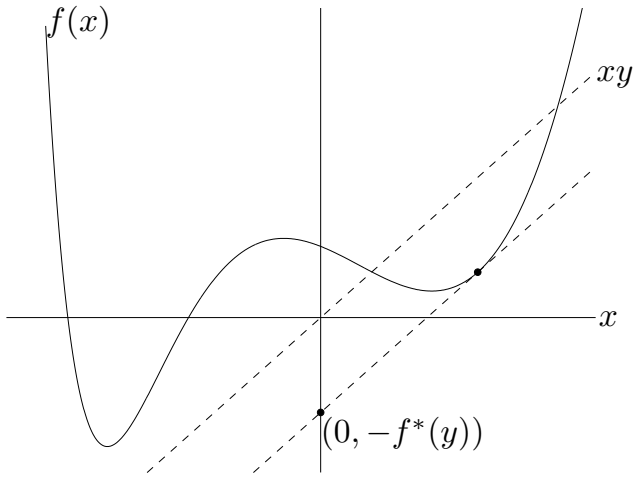
Dual norm of dual norm: can show that $\|x\|_{**} = \|x\|$

Conjugate function

Given a function $f : \mathbb{R}^n \to \mathbb{R}$, define its conjugate $f^* : \mathbb{R}^n \to \mathbb{R}$,

$$f^*(y) = \max_x \; y^T x - f(x)$$

Note that f^* is always convex, since it is the pointwise maximum of convex (affine) functions in y (here f need not be convex)



$f^*(y)$: maximum gap between linear function $y^T x$ and $f(x)$

(From B & V page 91)

For differentiable f, conjugation is called the Legendre transform

Properties:

- Fenchel's inequality: for any x, y,

$$f(x) + f^*(y) \geq x^T y$$

- Conjugate of conjugate f^{**} satisfies $f^{**} \leq f$
- If f is closed and convex, then $f^{**} = f$
- If f is closed and convex, then for any x, y,

$$\begin{aligned} x \in \partial f^*(y) &\iff y \in \partial f(x) \\ &\iff f(x) + f^*(y) = x^T y \end{aligned}$$

- If $f(u, v) = f_1(u) + f_2(v)$, then

$$f^*(w, z) = f_1^*(w) + f_2^*(z)$$

Examples:

- Simple quadratic: let $f(x) = \frac{1}{2}x^T Q x$, where $Q \succ 0$. Then $y^T x - \frac{1}{2}x^T Q x$ is strictly concave in y and is maximized at $y = Q^{-1}x$, so

$$f^*(y) = \frac{1}{2}y^T Q^{-1} y$$

- Indicator function: if $f(x) = I_C(x)$, then its conjugate is

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

  called the support function of C

- Norm: if $f(x) = \|x\|$, then its conjugate is

$$f^*(y) = I_{\{z \,:\, \|z\|_* \leq 1\}}(y)$$

  where $\|\cdot\|_*$ is the dual norm of $\|\cdot\|$

Example: lasso dual

Given $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times p}$, recall the lasso problem:

$$\min_{\beta} \ \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1$$

Its dual function is just a constant (equal to $f^\star$). Therefore we transform the primal to

$$\min_{\beta, z} \ \frac{1}{2}\|y - z\|_2^2 + \lambda\|\beta\|_1 \ \text{ subject to } \ z = X\beta$$

so dual function is now

$$\begin{aligned} g(u) &= \min_{\beta, z} \ \frac{1}{2}\|y - z\|_2^2 + \lambda\|\beta\|_1 + u^T(z - X\beta) \\ &= \frac{1}{2}\|y\|_2^2 - \frac{1}{2}\|y - u\|_2^2 - I_{\{v \,:\, \|v\|_\infty \leq 1\}}(X^T u/\lambda) \end{aligned}$$

Therefore the lasso dual problem is

$$\max_u \frac{1}{2}\Big(\|y\|_2^2 - \|y-u\|_2^2\Big) \text{ subject to } \|X^T u\|_\infty \le \lambda$$
$$\iff \min_u \|y-u\|_2^2 \text{ subject to } \|X^T u\|_\infty \le \lambda$$

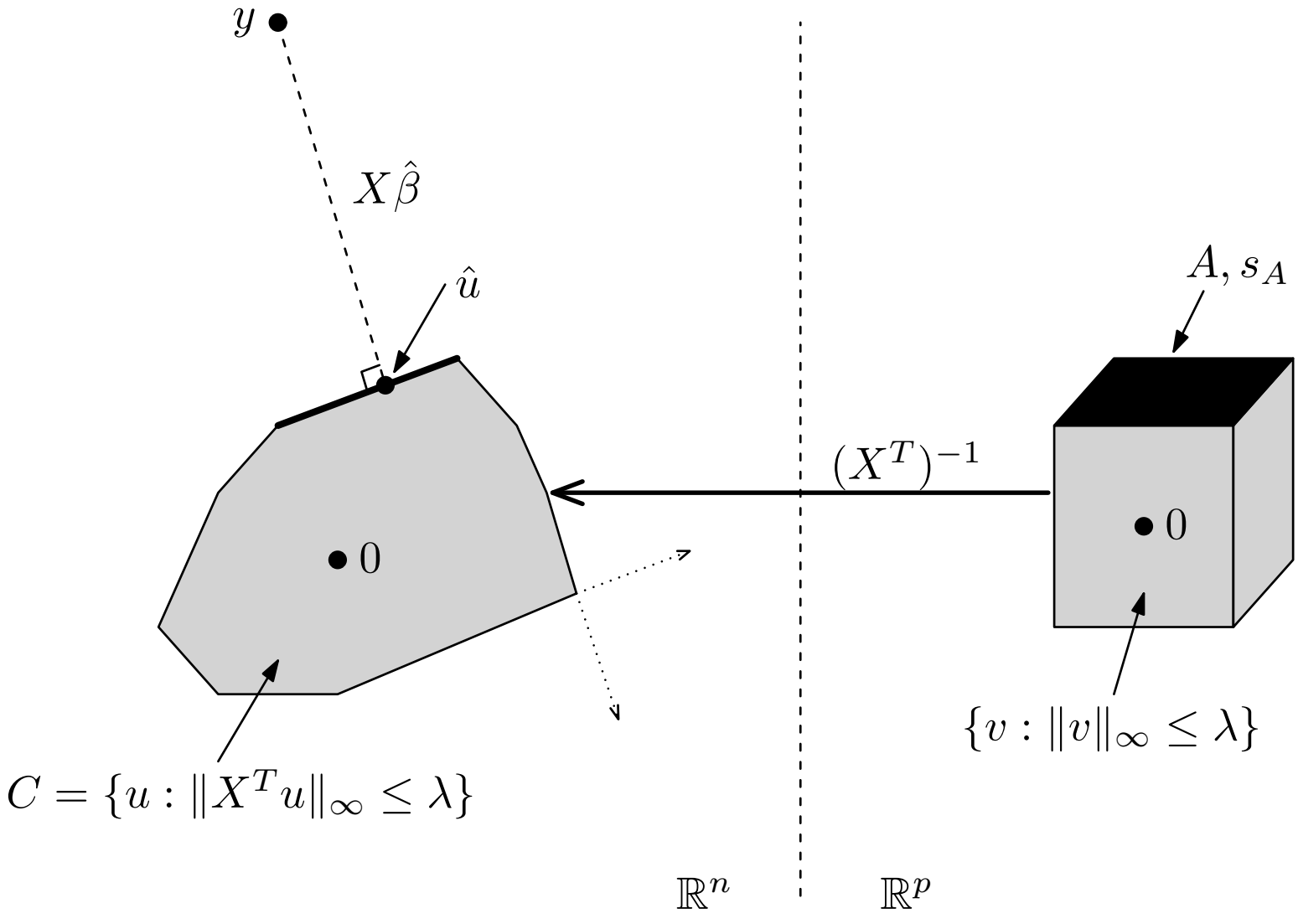
Check: Slater's condition holds, and hence so does strong duality. But note: the optimal value of the last problem is not the optimal lasso objective value

Note that given the dual solution u, any lasso solution β satisfies

$$X\beta = y - u$$

This is from KKT stationarity condition for z (i.e., $z - y + \beta = 0$). So the lasso fit is just the dual residual[2]

[2] See, e.g., Tibshirani and Taylor (2012), "Degrees of freedom in lasso problems", for consequences of dual perpsective


y
$X\hat{\beta}$
$\hat{u}$
A, s_A
$(X^T)^{-1}$
0
0
$\{v : \|v\|_\infty \leq \lambda\}$
$C = \{u : \|X^T u\|_\infty \leq \lambda\}$
$\mathbb{R}^n$
$\mathbb{R}^p$

Conjugates and dual problems

Conjugates appear frequently in derivation of dual problems, via

$$-f^*(u) = \min_x \; f(x) - u^T x$$

in minimization of the Lagrangian. E.g., consider

$$\min_x \; f(x) + g(x)$$

Equivalently: $\min_{x,z} \; f(x) + g(z) \;\; \text{subject to} \;\; x = z$. Dual function:

$$g(u) = \min_x \; f(x) + g(z) + u^T(z - x) = -f^*(u) - g^*(-u)$$

Hence dual problem is

$$\max_u \; -f^*(u) - g^*(-u)$$

Examples of this last calculation:

- Indicator function:

$$\text{Primal}: \quad \min_x \; f(x) + I_C(x)$$
$$\text{Dual}: \quad \max_u \; -f^*(u) - I_C^*(-u)$$

  where I_C^* is the support function of C

- Norms: the dual of

$$\text{Primal}: \quad \min_x \; f(x) + \|x\|$$
$$\text{Dual}: \quad \max_u \; -f^*(u) \;\; \text{subject to} \;\; \|u\|_* \leq 1$$

  where $\|\cdot\|_*$ is the dual norm of $\|\cdot\|$

Shifting linear transformations

Dual formulations can help us by "shifting" a linear transformation between one part of the objective and another. Consider

$$\min_x \; f(x) + g(Ax)$$

Equivalently: $\min_{x,z} \; f(x) + g(z) \;\; \text{subject to} \;\; Ax = z$. Like before, dual is:

$$\max_u \; -f^*(A^T u) - g^*(-u)$$

Example: for a norm and its dual norm, $\|\cdot\|$, $\|\cdot\|_*$:

$$\begin{aligned} \text{Primal}: \quad & \min_x \; f(x) + \|Ax\| \\ \text{Dual}: \quad & \max_u \; -f(A^T u) \;\; \text{subject to} \;\; \|u\|_* \leq 1 \end{aligned}$$

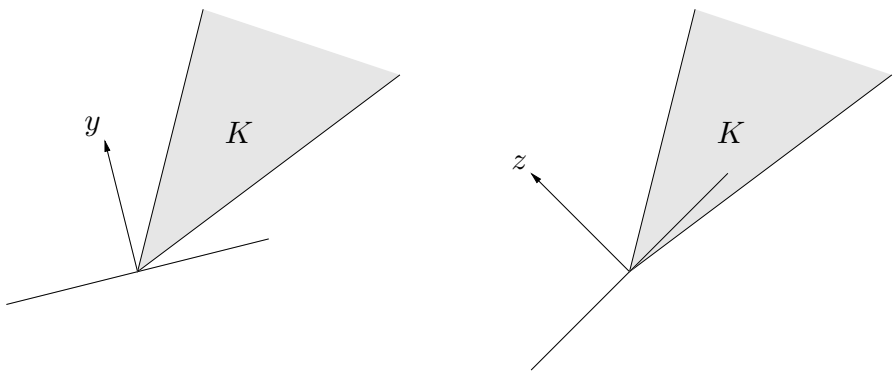
The dual can often be a helpful transformation here

Dual cones

For a cone $K \subseteq \mathbb{R}^n$ (recall this means $x \in K$, $t \geq 0 \Longrightarrow tx \in K$),

$$K^* = \{y : y^T x \geq 0 \text{ for all } x \in K\}$$

is called its dual cone. This is always a convex cone (even if K is not convex)



Notice that $y \in K^*$ $\Longleftrightarrow$ the halfspace $\{x : y^T x \geq 0\}$ contains K

(From B & V page 52)

Important property: if K is a closed convex cone, then $K^{**} = K$

Examples:

- Linear subspace: the dual cone of a linear subspace V is $V^\perp$, its orthogonal complement. E.g., $(\mathrm{row}(A))^* = \mathrm{null}(A)$

- Norm cone: the dual cone of the norm cone

$$K = \{(x,t) \in \mathbb{R}^{n+1} : \|x\| \le t\}$$

  is the norm cone of its dual norm

$$K^* = \{(y,s) \in \mathbb{R}^{n+1} : \|y\|_* \le s\}$$

- Positive semidefinite cone: the convex cone $\mathbb{S}^n_+$ is self-dual, meaning $(\mathbb{S}^n_+)^* = \mathbb{S}^n_+$. Why? Check that

$$Y \succeq 0 \iff \mathrm{tr}(YX) \ge 0 \text{ for all } X \succeq 0$$

  by looking at the eigendecomposition of X

Dual cones and dual problems

Consider the cone constrained problem

$$\min_x \ f(x) \ \text{ subject to } \ Ax \in K$$

Recall that its dual problem is

$$\max_u \ -f^*(A^T u) - I_K^*(-u)$$

where recall $I_K^*(y) = \max_{z \in K} \ z^T y$, the support function of K. If K is a cone, then this is simply

$$\max_u \ -f^*(A^T u) \ \text{ subject to } \ u \in K^*$$

where K^* is the dual cone of K, because $I_K^*(-u) = I_{K^*}(u)$

This is quite a useful observation, because many different types of constraints can be posed as cone constraints

Dual subtleties

- Often, we will transform the dual into an equivalent problem and still call this the dual. Under strong duality, we can use solutions of the (transformed) dual problem to characterize or compute primal solutions

  Warning: the optimal value of this transformed dual problem is not necessarily the optimal primal value

- A common trick in deriving duals for unconstrained problems is to first transform the primal by adding a dummy variable and an equality constraint

  Usually there is ambiguity in how to do this. Different choices can lead to different dual problems!

References

Parts A, B, and C:

- S. Boyd and L. Vandenberghe (2004), "Convex optimization", Chapter 5
- R. T. Rockafellar (1970), "Convex analysis", Chapters 28–30

Part D:

- S. Boyd and L. Vandenberghe (2004), "Convex optimization", Chapters 2, 3, 5
- R. T. Rockafellar (1970), "Convex analysis", Chapters 12, 13, 14, 16, 28–30