Duality Uses and Correspondences

Last time: KKT conditions

Recall that for the problem

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

the KKT conditions are

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

These are necessary for optimality (of a primal-dual pair $x^\star$ and $u^\star, v^\star$) under strong duality, and always sufficient

2

A. Why is duality useful?

Uses of duality

Two key uses of duality:

- For x primal feasible and u, v dual feasible,

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

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

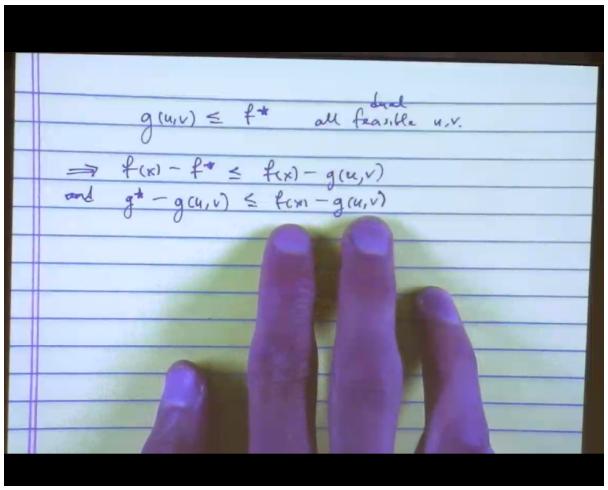
  $$f(x) - f(x^\star) \leq f(x) - g(u, v)$$

  a zero duality gap implies optimality. Also, the duality gap can be used as a stopping criterion in algorithms

- Under strong duality, given dual optimal $u^\star, v^\star$, any primal solution minimizes $L(x, u^\star, v^\star)$ over all x (i.e., it satisfies stationarity condition). This can be used to characterize or compute primal solutions

3

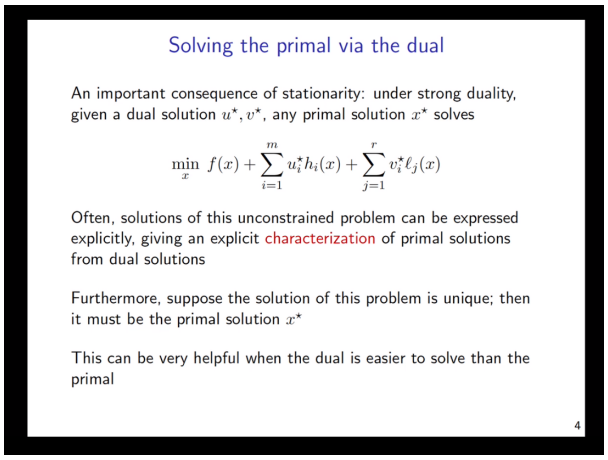
1. Algorithmic: duality gap



Stop an iterative algorithm with this guaranteed upper bound on distance from optimal criterion

2. It is easy sometimes to characterize a primal solution using the dual solution
    - either by solving the dual directly
    - The important piece to remember is that under strong duality, given a dual solution u* and v*, any primal solution x* minimizes the unconstrained Lagrangian!

Solving the primal via the dual



Example from B & V page 249:

$$\min_x \sum_{i=1}^n f_i(x_i) \quad \text{subject to} \quad a^T x = b$$

where each $f_i : \mathbb{R} \to \mathbb{R}$ is smooth, strictly convex. Dual function:

$$\begin{aligned} g(v) &= \min_x \sum_{i=1}^n f_i(x_i) + v(b - a^T x) \\ &= bv + \sum_{i=1}^n \min_{x_i} \{f_i(x_i) - a_i v x_i\} \\ &= bv - \sum_{i=1}^n f_i^*(a_i v) \end{aligned}$$

where f_i^* is the conjugate of f_i, to be defined shortly

5

The primal solution would be easily solved if it were not for the equality constraint, since each piecewise component i of the sum depends on just one component x_i. But the equality onstraint ties all the x_i together in one block.

But the dual can help untangle this block into individual components once agian

The conjugate $f_i^*(x,\ a_i v)\ \equiv\ -\min\{f_i(x_i)\ -\ a_i v x_i\}$
And we can write the dual in the form of a sum of isolated cnojugates.

Therefore the dual problem is

$$\max_v\ bv - \sum_{i=1}^n f_i^*(a_i v) \iff \min_v \sum_{i=1}^n f_i^*(a_i v) - bv$$

This is a convex minimization problem with scalar variable—much easier to solve than primal

Given $v^\star$, the primal solution $x^\star$ solves

$$\min_x \sum_{i=1}^n \left(f_i(x_i) - a_i v^\star x_i\right)$$

Strict convexity of each f_i implies that this has a unique solution, namely $x^\star$, which we compute by solving $\nabla f_i(x_i) = a_i v^\star$ for each i

The dual problem is to then maximize the dual function

$\max_v bv\ -\ \sum f_i^*(x,\ a_i v)\ \Longrightarrow\ \min_v \sum f_i^*(x,\ a_i v)\ -\ bv$
which is a much easier convex optimization problem to solve than the primal since it is univariate!

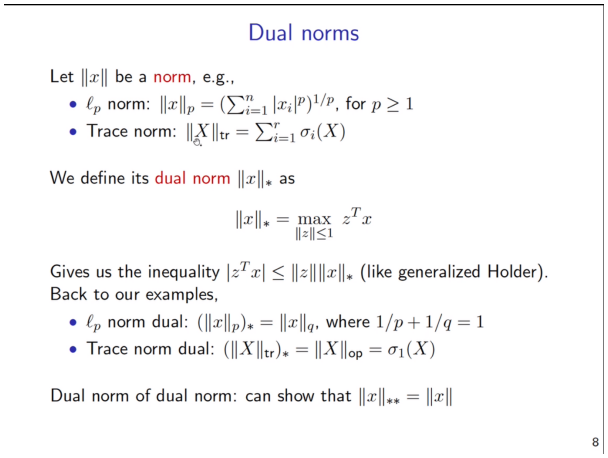
- can do grid search, golden ratio, etc for each f_i^* in general since it is modelled as uivariate

In this case, we know that $f_i^*(x,\ a_i v)\ \equiv\ -\min\{f_i(x_i)\ -\ a_i v x_i\}$
so given v*, the primal solution x^* solves

The dual is easier - so solve the dual - then use stationarity condition to characterize the primal

Outline
A. Dual norms
B. Conjugate functions
C. Dual cones
D. Dual tricks and subtleties

A. Dual norms

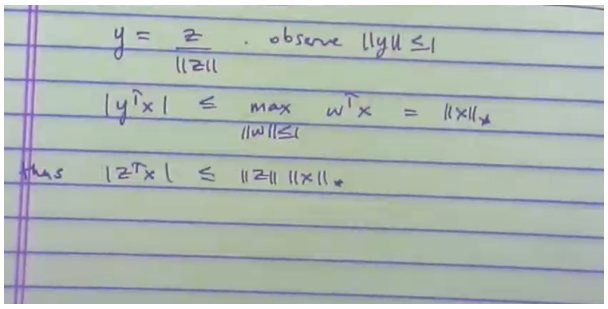


Trace norm = sum of the singular values of a matrix X

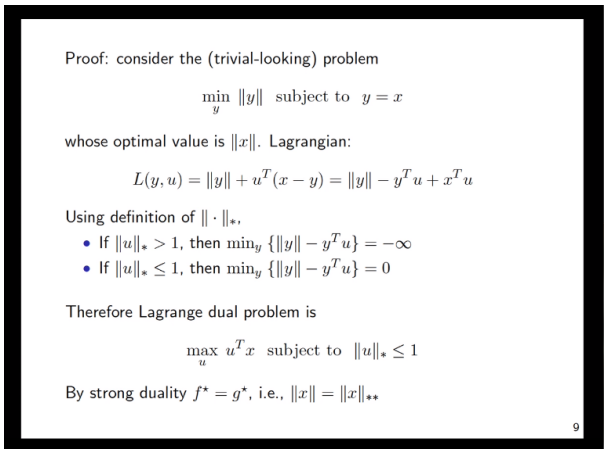
Dual norm is defined by the support function of the unit ball in the primal norm.
Dual norm $||x||_*\ =\ \max_{||z|| \leqslant 1} z^T x$

Dual of the trace norm = operator norm = largest singular value

Holder's inequality



Proof: The dual of the dual norm is the primal norm



B. Conjugate function

The biggest gap between a function linear in x and $f(x)$.

Given function $f: R^n \rightarrow R$, its conjugate $f^*: R^n \rightarrow R$ is

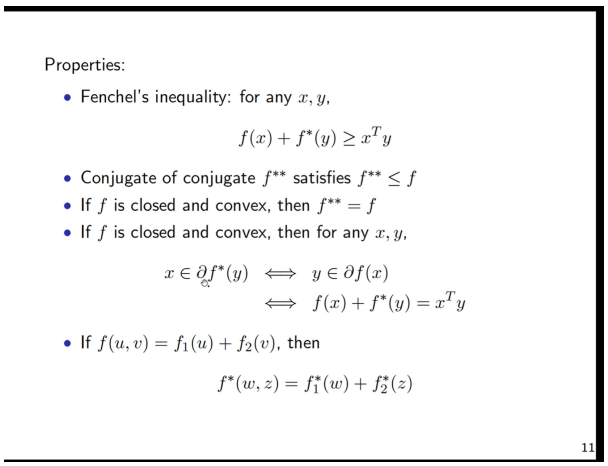
$f^*(y) = \max_x y^T x - f(x)$

Note that f^* **is always convex in y**, as it is the pointwise maximum of convex (affine) functions in y

- Note that f need not be convex here
- This is why this is sometimes called the convex conjugate

For a differentiable function f, conjugation is called the Legendre transform

Conjugates usually make dual calculations easier.



- Fenchel's inequality is key: $f(x) + f^*(y) \geqslant x^T y$
  - Proof?
- If x is a subgradient of the conjugate at y, then y is a subgradient of the primal evaluated at x.
  - This in turn, along with the knowledge that $f^{**} = f$, implies that $f(x) + f^*(y) = x^T y$ if f is closed and convex, for any x and y

B.1. Conjugate gradient

When f *and* f^* are differentiable, we know that

$x = \nabla f^*(y) \implies y = \nabla f(x)$

That is, $(\nabla f)^{-1} = \nabla f^*$!

- the gradient map is the inverse of the conjugate gradient!

B.2. Examples of conjugates

1. Convex quadratic

$f(x) = \frac{1}{2} x^T Q x,\ Q \succ 0$

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

Setting gradient of $y^T x - \frac{1}{2} x^T Q x$ wrt x to zero to find the expression for the max,

$0 = y - Qx \implies x = Q^{-1} y$

plugging this value in, $f^*(y) = y^T Q^{-1} y - \frac{1}{2} y^T Q^{-1} y = \frac{1}{2} y^T Q^{-1} y$

A note: by Fenchel's inequality, this means that $f(x) + f^*(y) \geqslant x^T y$

i.e. $\frac{1}{2} x^T Q x + \frac{1}{2} y^T Q^{-1} y \geqslant x^T y$

Here, we can see easily that the conjugate of this conjugate is indeed the primal.

2. Indicator functions

$f(x) = I_C(x)$

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

$= max_{x \in C} y^T x$

$\equiv I_C^*(y)$

the support function of C.

- the maximum inner product between given y and some element x in the set C

Question: what is the conjugate of the support function?

Assuming C is closed and convex, $I_C^{**}(y) = I_c(x)$

Why?

$f^*(z) = max_x \left\{ z^T x - f(x) \right\} = max_x \left\{ z^T x - f^*(y) \right\} = max_x \left\{ z^T x - max_{x \in C} y^T x \right\} \rightarrow$??

3. Norm

If $f(x) = ||x||$, its conjugate is $f^*(y) = I_{\{z:\, ||z_*|| \leqslant 1\}}(y)$

B.2.1. Lasso dual

Primal problem: $\min_B f(B) = \min_B \left\{ ||y - XB||_2^2 + \lambda ||B|| \right\}$

There are no constraints! So the dual $g(B, \phi) = \min_B f(B) = f_{opt}$

Without constraints, the dual is always the (constant) optimal primal value.

But you can parameterize the primal to introduce constraints, so we have non-null dual variables.

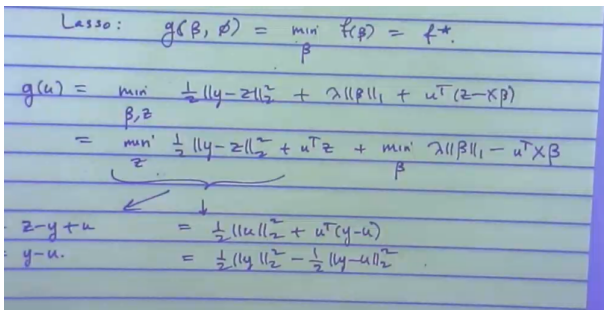
Its dual function is just a constant (equal to f^*). Therefore we transform the primal to

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

so dual function is now

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

How to make this jump in the slide from line 2 to 3?



Separate the dual out into parts dependent on Z and on B, and minimize them separately.

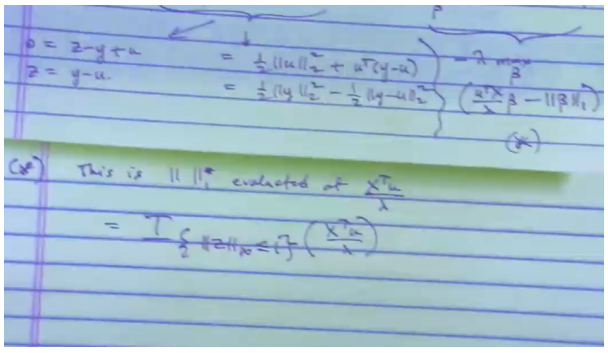
The former is simple - it is a quadratic in Z
The latter is more complicated. We could minimize it using subgradients...but an easier siwtch would be to rewrite it so it fits the form of a dual function: $y^T x - f(x)$

$$\min_B \lambda||B||_1 - u^T XB = -\lambda \max_B \left(\frac{u^T X}{\lambda} B - ||B||_1 \right)$$

This is just $||.||_1^*$ evaluated at $y = \frac{X^T u}{\lambda}$

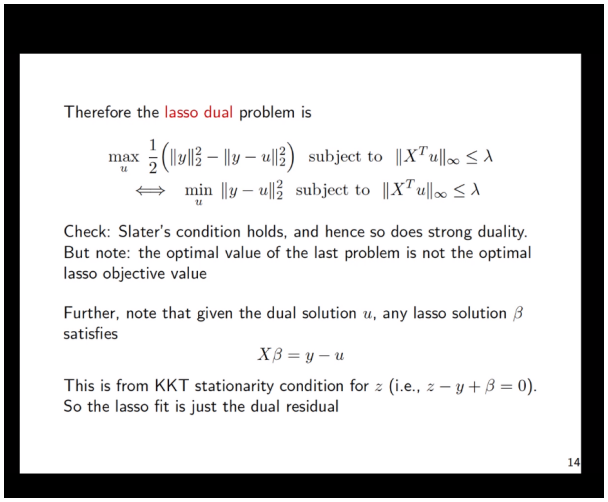
So $\min_B \lambda||B||_1 - u^T XB = \lambda I_{\{||z||_\infty \leq 1\}}\left(\frac{X^T u}{\lambda}\right) = I_{\{||z||_\infty \leq 1\}}\left(\frac{X^T u}{\lambda}\right)$

since the λ doesn't matter - the value of the indicator is either 0 or inf



Now notice that this indicator function effectively forms a constraint on the dual.

If $||X^T u||_\infty > \lambda$, then $I_{\{||z||_\infty \leq 1\}}\left(\frac{X^T u}{\lambda}\right) = \infty$.



The dual problem is exactly a projection problem

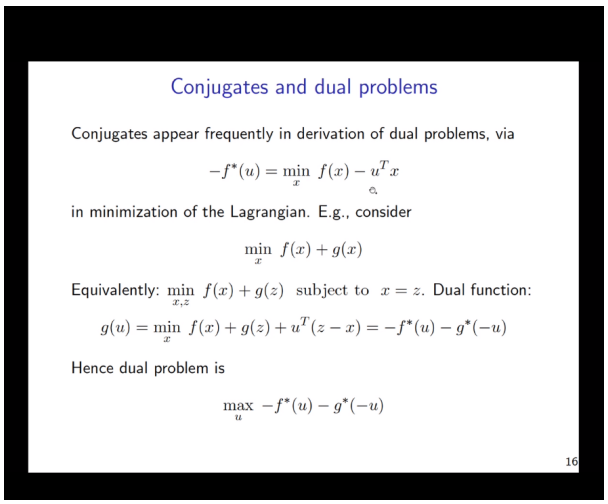
Dual problem is: $\min_{s,t} ||y-u||_2^2 \; st \; ||X^T u||_\infty \leqslant \lambda$

This projects y onto $\left\{u \; : \; ||X^T u||_\infty \leqslant \lambda\right\}$, which is a polyhedron.

This geometry tells us something about the nature of the lasso.

- a property of projections is that the distance between projections onto a convex set $\leqslant$ the dsitance between the original points themselves
- so the distance between different ys will be less than the distance between their correponding projections in the dual problem
- ie LASSO is a non expansive operation

This dual example illustrates how conjugates are useufl - they pop up naturally all the time when dealing with Lagrangians/duals!



Examples of this last calculation:

- Indicator function:

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

  where I_C^* is the support function of C

- Norms: the dual of

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

  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