

## Duality Uses and Correspondences

Last time: KKT conditions

Recall that for the problem

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

the 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)
- $u_i \cdot h_i(\hat{x}) = 0$  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  $x^*$  and  $u^*, v^*$ ) 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

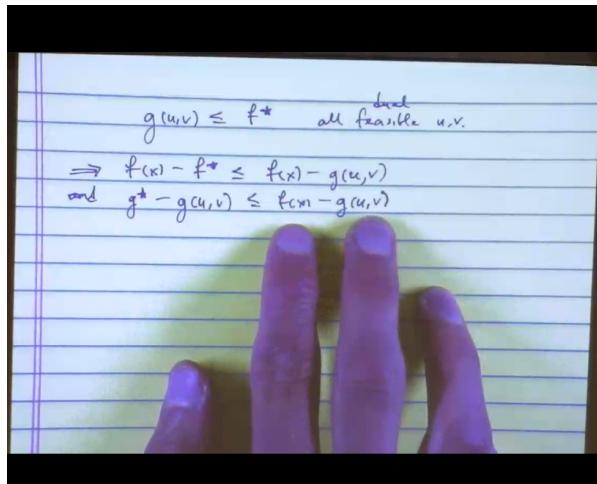
$$f(x) - f(x^*) \leq f(x) - g(u, v)$$

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  $u^*, v^*$ , any primal solution minimizes  $L(x, u^*, v^*)$  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

**Solving the primal via the dual**

An important consequence of stationarity: under strong duality, given a dual solution  $u^*, v^*$ , any primal solution  $x^*$  solves

$$\min_x f(x) + \sum_{i=1}^m u_i^* h_i(x) + \sum_{j=1}^r v_j^* \ell_j(x)$$

Often, solutions of this unconstrained problem can be expressed explicitly, giving an explicit **characterization** of primal solutions from dual solutions

Furthermore, suppose the solution of this problem is unique; then it must be the primal solution  $x^*$

This can be very helpful when the dual is easier to solve than the primal

4

Example from B & V page 249:

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

where each  $f_i : \mathbb{R} \rightarrow \mathbb{R}$  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 constraint ties all the  $x_i$  together in one block.

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

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 conjugates.

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  $v^*$ , the primal solution  $x^*$  solves

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

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

The dual problem is to then maximize the dual function

$$\max_v bv - \sum f_i^*(x, a_i v) \implies \min_v \sum f_i^*(x, a_i v) - bv$$

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 univariate

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

##### Dual norms

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

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

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,

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

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

8

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

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

#### Holder's inequality

$y = \frac{z}{\|z\|}$ . observe  $\|y\| \leq 1$   
 $|y^T x| \leq \max_{\|w\| \leq 1} w^T x = \|x\|_*$   
 thus  $|z^T x| \leq \|z\| \|x\|_*$

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

Proof: consider the (trivial-looking) problem

$$\min_y \|y\| \text{ subject to } y = x$$

whose optimal value is  $\|x\|$ . Lagrangian:

$$L(y, u) = \|y\| + u^T(x - y) = \|y\| - y^T u + x^T u$$

Using definition of  $\|\cdot\|_*$ ,

- If  $\|u\|_* > 1$ , then  $\min_y \{\|y\| - y^T u\} = -\infty$
- If  $\|u\|_* \leq 1$ , then  $\min_y \{\|y\| - y^T u\} = 0$

Therefore Lagrange dual problem is

$$\max_u u^T x \text{ subject to } \|u\|_* \leq 1$$

By strong duality  $f^* = g^*$ , i.e.,  $\|x\| = \|x\|_{**}$

9

## B. Conjugate function

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

Given function  $f: \mathbb{R}^n \rightarrow \mathbb{R}$ , its conjugate  $f^*: \mathbb{R}^n \rightarrow \mathbb{R}$  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.

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

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

11

- Fenchel's inequality is key:  $f(x) + f^*(y) \geq x^T y$ 
  - 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

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

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

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

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

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

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

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

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

## 2. Indicator functions

$$\begin{aligned} f(x) &= I_C(x) \\ f^*(y) &= \max_x y^T x - f(x) \\ &= \max_{x \in C} y^T x \\ &\equiv I_C^*(y) \end{aligned}$$

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?

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

Why?

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

## 3. Norm

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

### B.2.1. Lasso dual

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

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

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

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?

Lasso:  $g(\beta, \rho) = \min_{\beta} f(\beta) = f^*$

$$\begin{aligned} g(u) &= \min_{\beta, z} \frac{1}{2} \|y - z\|_2^2 + \lambda \|\beta\|_1 + u^T(z - X\beta) \\ &= \underbrace{\min_z \frac{1}{2} \|y - z\|_2^2 + u^T z}_{z} + \min_{\beta} \lambda \|\beta\|_1 - u^T X \beta \\ &\quad \downarrow \quad \downarrow \\ z - y + u &= \frac{1}{2} \|u\|_2^2 + u^T(y - u) \\ y - u &= \frac{1}{2} \|y\|_2^2 - \frac{1}{2} \|y - u\|_2^2. \end{aligned}$$

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 switch would be to rewrite it so it fits the form of a dual function:  $y^T x - f(x)$

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

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

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

since the  $\lambda$  doesn't matter - the value of the indicator is either 0 or inf

$$\begin{aligned} \beta &= z - y + u &= \frac{1}{2} \|u\|_2^2 + u^T(y - u) & \xrightarrow{\text{min } \beta} \\ z &= y - u &= \frac{1}{2} \|y\|_2^2 - \frac{1}{2} \|y - u\|_2^2 & (u^T \beta - \|\beta\|_1) \\ \text{(*)} & \text{This is } \|\cdot\|_1^* \text{ evaluated at } \frac{X^T u}{\lambda} \\ &= I_{\{\|z\|_\infty \leq 1\}} \left( \frac{X^T u}{\lambda} \right) \end{aligned}$$

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

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

Therefore the **lasso dual** problem is

$$\begin{aligned} \max_u \frac{1}{2} (\|y\|_2^2 - \|y - u\|_2^2) & \text{ subject to } \|X^T u\|_\infty \leq \lambda \\ \iff \min_u \|y - u\|_2^2 & \text{ subject to } \|X^T u\|_\infty \leq \lambda \end{aligned}$$

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

Further, note that given the dual solution  $u$ , any lasso solution  $\beta$  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

14

The dual problem is exactly a projection problem

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

This projects  $y$  onto  $\{u : \|X^T u\|_\infty \leq \lambda\}$ , 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  $\leq$  the distance between the original points themselves
- so the distance between different  $y$ s will be less than the distance between their corresponding projections in the dual problem
- ie LASSO is a non expansive operation

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

### 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:  $\min_{x,z} f(x) + g(z)$  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

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

16

Examples of this last calculation:

- Indicator function:

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

where  $I_C^*$  is the support function of  $C$

- Norms: the dual of

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

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

17

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

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

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

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

The dual can often be a helpful transformation here

18