

## Recap

Last time: duality

Given a minimization problem

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

we defined the **Lagrangian**:

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

and **Lagrange dual function**:

$$g(u, v) = \min_x L(x, u, v)$$

2

The subsequent **dual problem** is:

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

Important properties:

- Dual problem is always convex, i.e.,  $g$  is always concave (even if primal problem is not convex)
- The primal and dual optimal values,  $f^*$  and  $g^*$ , always satisfy weak duality:  $f^* \geq g^*$
- Slater's condition: for convex primal, if there is an  $x$  such that

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

then **strong duality** holds:  $f^* = g^*$ . Can be further refined to strict inequalities over the nonaffine  $h_i$ ,  $i = 1, \dots, m$

3

Today

1. KKT conditions
2. Examples
3. Use of KKT: regularize as a constraint vs a penalty are equivalent
4. Uniqueness of L1 problems without strict convexity

### A. KKT conditions

**Karush-Kuhn-Tucker conditions**

Given general 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 **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)

5

## 4 conditions

$x$  = primal variable

$u, v$  = dual variables

### 1. Stationarity

At optimal  $x, u, v$ , the subdifferential of the Lagrangian function wrt  $x$  has to contain 0

$$0 \in \partial \left( f(x) + \sum u_i h_i(x) + \sum v_j l_j(x) \right)$$

- in the absence of constraints, this reduces to subgradient optimality - indeed, one way to understand KKT conditions is as an extension of subgradient optimality!

### 2. Complementary slackness

If you look at all the dual variables and all the inequality constraint functions, then for each  $i$ , either one or both must be 0

$$u_i \cdot h_i(x) = 0 \quad \forall i$$

### 3. Primal feasibility

$x$  has to satisfy all the primal constraints

$$h_i(x) \leq 0, l_j(x) = 0 \quad \forall i, j$$

### 4. Dual feasibility

$u$  has to satisfy the dual constraints (for the inequality constraints in the primal problem)

$$u_i \geq 0 \quad \forall i$$

If you have an  $x$  and  $u, v$  that satisfies these problems, that is **necessary** and **sufficient** for these to be solutions to the primal and dual problems.

### A1. Necessity of KKT conditions at optimal solution

Assuming zero duality gap, i.e. that  $f(x^*) = g(u^*, v^*)$

### Necessity

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

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

In other words, all these inequalities are actually equalities

6

First step is by assumption under Slater's condition

Second comes from definition of dual

Third comes from plugging in one value of  $L()$  and claiming it is less than minimum

$$\begin{aligned} f(x^*) &= g(u^*, v^*) \text{ from Slater's condition} \\ &= \min_x L(x, u, v) \text{ by defintion} \\ &= \min_x f(x) + \sum u_i^* h_i(x) + \sum v_j^* l_j(x), \text{ evaluated at optimal } u \text{ and } v \\ &\leq f(x^*) + \sum u_i^* h_i(x^*) + \sum v_j^* l_j(x^*) = L(x^*, u^*, v^*) \\ &\leq f(x^*) \end{aligned}$$

Which means they are all equalities

$$\text{Hence, } \min_x L(x, u, v) = L(x^*, u^*, v^*)$$

By subgradient optimality, we know that at  $\min_x L(x, u, v)$ ,

$$0 \in \partial_x(L(x, u, v))$$

$$\text{so } 0 \in \partial_x(L(x^*, u^*, v^*))$$

$$\implies 0 \in \partial_x \left( f(x^*) + \sum u_i^* h_i(x^*) + \sum v_j^* l_j(x^*) \right)$$

which is the stationarity condition

So the **stationarity** condition must hold at a solution!

Then observe that

$f(x^*) + \sum u_i^* h_i(x^*) + \sum v_j^* l_j(x^*) = f(x^*)$  means that

$\sum u_i^* h_i(x^*) = 0$ , because we know that  $\sum v_j^* l_j(x^*) = 0$  since  $l_j(x^*) = 0 \forall j$ .

We also know that  $u_i \geq 0$ , and  $h_i(x) \leq 0 \forall i$

So every term in the sum has the same sign i.e. every term is  $\leq 0$

So the terms can't cancel out – therefore if they sum to 0, each individual term must be zero, which is the complementary slackness condition

So the **complementary slackness** condition must hold at a solution!

If  $x^*$  and  $u^*, v^*$  are primal and dual solutions with zero duality gap, then they must satisfy the KKT conditions.

→ under strong duality, the KKT conditions must hold

Two things to learn from this:

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

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

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

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

## A2. Sufficiency of KKT conditions for an optimal solution

If  $x^*, u^*, v^*$  fulfill KKT conditions, then

$g(u^*, v^*) = \min_x L(x, u, v)$  by definition

by stationarity, we know that  $0 \in \partial_x \left( f(x^*) + \sum u_i^* h_i(x^*) + \sum v_j^* l_j(x^*) \right)$

and by subgradient optimality, at  $\min_x L(x, u, v)$ , we know that

$0 \in \partial_x \left( f(x) + \sum u_i h_i(x) + \sum v_j l_j(x) \right)$

so  $g(u^*, v^*) = f(x^*) + \sum u_i^* h_i(x^*) + \sum v_j^* l_j(x^*)$

but we know  $\sum v_j^* l_j(x^*)$  as  $l_j(x) = 0 \forall j$  by feasibility conditions

and  $\sum u_i^* h_i(x^*) = 0$  by complementary slackness  
so,  $g(u^*, v^*) = f(x^*)$

Therefore the duality gap is 0, and  $x^*, u^*, v^*$  are primal and dual feasible, so they are also primal and dual optimal.

**If  $x^*$  and  $u^*, v^*$  satisfy the KKT conditions, then they are primal and dual solutions - with zero duality gap (?)**

### Sufficiency

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

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

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

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

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

8

### Putting it together

In summary, KKT conditions:

- 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 constraints).

$x^*$  and  $u^*, v^*$  are primal and dual solutions  
 $\iff$   $x^*$  and  $u^*, v^*$  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!)

9

### 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

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

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

10

1 < 17.1.01 - 0 /

$$\begin{aligned} \text{min} \quad & \frac{1}{2} x^T Q x + c^T x, \quad Q \succeq 0 \\ \text{st.} \quad & Ax = 0. \end{aligned} \quad \left. \begin{array}{l} \text{unconstrained} \\ Qx = c \\ x = -Q^{-1}c. \end{array} \right\}$$

$$L(x, u) = \frac{1}{2} x^T Q x + c^T x + u^T Ax.$$

KKT conditions:

- stat.  $\nabla_x L(x, u) = 0.$
- comp. slack:  $(Qx + c + A^T u = 0)^\top \rightarrow [Q \quad A^T]^\top [x]$
- feasibility:  $Ax = 0 \rightarrow [A \quad 0]^\top [u] = [-c]^\top [u] = [0]$

### Example: quadratic with equality constraints

Consider for  $Q \succeq 0$ ,

$$\begin{aligned} \min_x \quad & \frac{1}{2} x^T Q x + c^T x \\ \text{subject to} \quad & Ax = 0 \end{aligned}$$

E.g., as we will see, this corresponds to Newton step for equality-constrained problem  $\min_x f(x)$  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)

11

$$L(x, u, v) = -\sum \log(2_i + x_i) - \sum u_i x_i + v(1^T x - 1)$$

KKT:

- o stat.  $\nabla_x L(x, u, v) = 0$
- for  $i=1..n$ :  $0 = -\frac{1}{2(1+x_i)} - u_i + v \quad \left. \begin{array}{l} \\ \end{array} \right\}$
- o comp. slack:  $u_i x_i = 0, \quad i=1..n$
- o feasibility:  $x \geq 0, \quad 1^T x = 1$
- $u \geq 0$ .

*lead to an algorithm for computing  $x^*$ .*

## B. Example: SVM, again

Example: support vector machines

Given  $y \in \{-1, 1\}^n$ , and  $X \in \mathbb{R}^{n \times p}$ , the support vector machine problem is:

$$\min_{\beta, \beta_0, \xi} \frac{1}{2} \|\beta\|_2^2 + C \sum_{i=1}^n \xi_i$$

subject to  $\xi_i \geq 0, \quad i = 1, \dots, n$

$$y_i(x_i^T \beta + \beta_0) \geq 1 - \xi_i, \quad i = 1, \dots, n$$

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:

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

14

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

15

The SVM classifier is a hyperplane of the form  $x_i^T B + B_0 = 0$

$x_i$  is said to 'support' the solution, since  $B$  is a linear combination of these  $x_i$ , if the corresponding dual variable  $w_i$  is not zero. If the corresponding dual variable is zero, then  $x_i$  doesn't matter

But by complementary slackness,  $w_i(1 - \xi_i - y_i(x_i^T B + B_0)) = 0$ . So  $w_i$  can only be non-zero if  $(1 - \xi_i - y_i(x_i^T B + B_0)) = 0$ . So support points only occur when the signed predicted value of the svm prediction for the point is exactly  $1 - \xi_i$ .

everything between  $x_i^T B + B_0 = 1$  and  $x_i^T B + B_0 = -1$  is the margin

Also by complimentary slackness, we know  $v_i \xi_i = 0$ . This gives us two cases to study within support points.

Case 1: For a support point  $i$ , if  $\xi_i = 0$  (i.e the point is on the right side of the margin) then  $v_i$  can be any number. ( $\geq 0$ , as per the constraints on the dual problem)

But from KKT stationarity, we know that  $w = c.1 - v$  which means that  $w_i \in (0, C]$

Case 2: For a support point  $i$ , if  $\xi_i \neq 0$  (i.e the point is on the wrong side of the margin) then  $v_i$  must be zero, which means that  $w_i = C$ .

What am I not understanding rn?

- what does  $x_i^T B + B_0$  mean intuitively? And then  $y_i(x_i^T B + B_0)$ ?
- what is the  $\xi_i$  variable? is it distance from the right side?
- <https://www.cs.cornell.edu/courses/cs4780/2018fa/lectures/lecturenote09.html>
  - excellent resource

## Constrained and Lagrange forms

### 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,

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

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

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

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  $x^*$  in (C) minimizes

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

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

16

Is the problem  $\min_x f(x)$  subject to  $h(x) \leq t$ ,  $t \in R$  (constrained form), equivalent to  $\min_x f(x) + \lambda h(x)$ ,  $\lambda \geq 0$  (Lagrange form)?

- This comes from KKT conditions

### C. Constrained $\rightarrow$ Lagrangian form

Constrained problem

$$\min_x f(x)$$

subject to

$$h(x) \leq t$$

So the Lagrangian  $L(x, \lambda) = f(x) + \lambda(h(x) - t)$

If the problem is strictly feasible, then we know strong duality/Slater's condition holds, then by

KKT stationarity, the solution  $x^*$  must also minimize  $L(x, \lambda) = f(x) + \lambda(h(x) - t)$

which in turn means it must minimize  $f(x) + \lambda h(x)$

$$L(x, \lambda) = f(x) + \lambda(h(x) - t)$$

KKT: stationarity:  $x^*$  must solve  $\min_x \{f(x) + \lambda(h(x) - t)\} \implies \min_x \{f(x) + \lambda h(x)\}$

### Lagrangian $\rightarrow$ Constrained form

Idea: look at solution  $x^*$  in Lagrange form - show that KKT conditions for this solution apply to constrained form - therefore it must also solve the constrained form

$\Leftrightarrow \min_x f(x) + \lambda h(x)$

(a) to  
(c) KKT: stat:  $x^*$  must solve  $\min_x f(x) + \lambda(h(x) - t)$

comp slacks:  $\lambda \cdot (h(x^*) - t) = 0$ . note: satisfied by feasibility:  $h(x^*) \leq t$ .  $t = h(x^*)$ .

$\lambda \geq 0$ .

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

**Conclusion:**

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

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

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

17

## D. Uniqueness in L1 penalized problems, from the KKT conditions

eg  $f(u) = \frac{1}{2} \|y - u\|_2^2$  so  $f(x\beta) = \frac{1}{2} \|y - x\beta\|_2^2$

or logistic loss  
or Poisson loss.

Note that we need strict convexity in the loss function  $f(u=\text{fit})$ , not in terms of B

- in this case  $f(X, B)$  is not strictly convex in B when X is 'wide'?
- **From lecture2: Strictly convex:**  $f(tx + (1-t)y) < t(f(x) + (1-t)f(y))$
- When  $X = y$ ,

LHS

$$\begin{aligned}
f(X; tB_1 + (1-t)B_2) &= \frac{1}{2} \|y - X(tB_1 + (1-t)B_2)\|_2^2 \\
&= \frac{1}{2}(y - X(tB_1 + (1-t)B_2))^T(y - X(tB_1 + (1-t)B_2)) \\
&= \frac{1}{2}(y^T - (tB_1 + (1-t)B_2)^T X^T)(y - X(tB_1 + (1-t)B_2)) \\
&= \frac{1}{2}(y^T y - y^T X(tB_1 + (1-t)B_2) - (tB_1 + (1-t)B_2)^T X^T y + \\
&\quad (tB_1 + (1-t)B_2)^T X^T X(tB_1 + (1-t)B_2)) \\
&= \frac{1}{2}(y^T y - ty^T X B_1 + (1-t)y^T X B_2 - t B_1^T X^T y - (1-t) B_2^T X^T y + \\
&\quad (tB_1 + (1-t)B_2)^T X^T X(tB_1 + (1-t)B_2))
\end{aligned}$$

RHS

$$\begin{aligned}
tf(X; B_1) &= \frac{1}{2}t((y - XB_1)^T(y - XB_1)) \\
\bullet &= \frac{1}{2}t(y^T y - y^T X B_1 - B_1^T X^T y + B_1^T X^T X B_1)
\end{aligned}$$

$$(1-t)f(X; B_2) = \frac{1}{2}(1-t)(y^T y - y^T X B_2 - B_2^T X^T y + B_2^T X^T X B_2)$$

$$so \ tf(X; B_1) + (1-t)f(X; B_2)$$

$$\begin{aligned}
&= \frac{1}{2}(y^T y - ty^T X B_1 - (1-t)y^T X B_2 - t B_1^T X^T y - (1-t) B_2^T X^T y + t B_1^T X^T X B_1 + (1-t) B_2^T X^T X B_2)
\end{aligned}$$

Need to compare

$$(tB_1 + (1-t)B_2)^T X^T X(tB_1 + (1-t)B_2) ? t B_1^T X^T X B_1 + (1-t) B_2^T X^T X B_2$$

LHS

$$\begin{aligned}
&t^2 B_1 X^T X B_1 + t(1-t) B_1^T X^T X B_2 + t(1-t) B_2^T X^T X B_1 + (1-t)^2 B_2^T X^T X B_2 \\
&= t^2 B_1 X^T X B_1 + (1-t)^2 B_2^T X^T X B_2 + 2t(1-t) B_1^T X^T X B_2 \\
&= \|t X B_1 + (1-t) X B_2\|_2^2 = L2norm(
\end{aligned}$$

RHS

$$= t \|X B_1\|_2^2 + (1-t) \|X B_2\|_2^2$$

where does this go....? Is the L2 norm squared strictly convex? If so, then this function is strictly convex in  $X B$ .

<https://math.stackexchange.com/questions/2368311/show-x-2-is-strictly-convex>

The answer is yes!

- How can you check strict convexity?
  - Hessian psd or greater than 0
  - Convex function definition

I just wasted 2 hours on a throwaway comment that I heard wrong. I need to be far more ruthless. Wtf?

$$f(x) = x^T x$$

$$f'(x) = 2x$$

$$f''(x) = 2I > 0$$

But

$$f(x) = \sqrt{x^T x}$$

$$f'(x_i) = 2$$

$$f''(x) = 0$$

So the L2 norm is not strictly convex, but L2norm^2 is.

Remember: the derivative is  $\left[ \frac{dy}{dx_1}, \frac{dy}{dx_2}, \dots, \frac{dy}{dx_p} \right]$ . So take the expression of the final value

and get the index-wis

#### Uniqueness in $\ell_1$ penalized problems

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

**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  $p \gg n$ )

18

Saturation of L1 regularized solution

Uniqueness without strict convexity here

- If  $n \ll p$ ,  $X$  will have a null space, so the unconstrained problem has infinitely many solutions
- Q: why does having a null space mean this?
- Because for a given solution  $B$ , you can add  $B+a$  such that  $a$  is in the null space of  $X$ ,

for infinite values of a, with some scaling b

- But under certain guarantees/conditions of randomness, you still get a unique solution!

Proof: the KKT conditions are

$$-X^T \nabla f(X\beta) = \lambda s, \quad s_i \in \begin{cases} \{\text{sign}(\beta_i)\} & \text{if } \beta_i \neq 0 \\ [-1, 1] & \text{if } \beta_i = 0 \end{cases}, \quad i = 1, \dots, n$$

Basic but important observations:

- $X\beta$  is unique by strict convexity of  $f$
- The KKT conditions hence imply  $s$  is unique

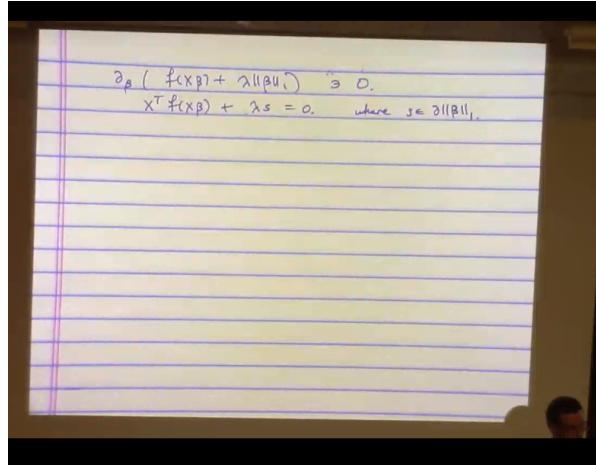
Thus we can define **equicorrelation set**

$$S = \{j : |X_j^T \nabla f(X\beta)| = \lambda\}$$

This is also unique, any solution satisfies  $\beta_i = 0$  for all  $i \notin S$

19

- Use the subgradient optimality/stationarity condition for the criterion - we know what the subgradient looks like for the L1 loss function (see earlier lecture)
- By strict convexity of the loss function wrt to the fit  $XB$  (as done above), we know that  $XB$  must be unique at the solution
- therefore as the LHS is unique, the RHS must also be unique - so  $S$  must also be unique
- Equicorrelation set: set (at the solution) where each feature is correlated with the gradient, with the biggest correlation possible
  - this also depends only on the fit and therefore at the solution must be unique
- Any solution for  $XB$  will therefore have for  $i$  not in  $S$  i.e. not in the equicorrelation set,  $B_{-i} = 0$ 
  - this follows by definition - of the subgradient and the equicorrelation set



- typo in the sheet - that should be  $\nabla f$

Now, we show that the L1 norm regularizer would not pick redundant features ie features that are linearly dependent.

Statement: The submatrix of  $X$  with columns in  $S$  is linearly independent.

Proof by contradiction:

- take a submatrix of the full feature matrix  $X$ , only extract the columns in  $S$
- Assume this submatrix was not linearly independent, i.e. that  $\text{rank}(X_S) < |S|$

First assume that  $\text{rank}(X_S) < |S|$  (here  $X \in \mathbb{R}^{n \times |S|}$ , submatrix of  $X$  corresponding to columns in  $S$ ). Then for some  $i \in S$ ,

$$X_i = \sum_{j \in S \setminus \{i\}} c_j X_j$$

for constants  $c_j \in \mathbb{R}$ , so that

$$s_i X_i = \sum_{j \in S \setminus \{i\}} s_j c_j \lambda(s_j X_j)$$

Hence taking an inner product with  $-\nabla f(X\beta)$ ,

$$\lambda = \sum_{j \in S \setminus \{i\}} (s_i s_j c_j) \lambda, \quad \text{i.e.,} \quad \sum_{j \in S \setminus \{i\}} s_i s_j c_j = 1$$

20

- typo in the second line: see below
- first, multiply both sides by  $s_{-i}$
- then, you can multiply RHS by  $s_{-j}$  twice since it is just a sign

—

$$s_i X_i = \sum_j s_j c_j (X_j \cdot s_j)$$

$a_j$

- Interpretation: the signed feature  $X_i$  = weighted sum of signed features  $X_j$
- And deduce that  $\sum a_j = 1$

Then, from the KKT conditions, we know that taking the inner product of  $s_i X_i$  with  $-\nabla f(XB)$  will yield  $\lambda$ . → why is this true for the subset matrix?

In other words, we've proved that  $\text{rank}(X_S) < |S|$  implies

$$s_i X_i = \sum_{j \in S \setminus \{i\}} a_j (s_j X_j)$$

i.e.,  $s_i X_i$  is in the affine span of  $s_j X_j$ ,  $j \in S \setminus \{i\}$  (subspace of dimension  $< n$ )

It is easy to show that, if the entries of  $X$  have a density over  $\mathbb{R}^{np}$ , then almost surely, this cannot happen

21

Therefore, if entries of  $X$  are drawn from continuous probability distribution, any solution must satisfy  $\text{rank}(X_S) = |S|$

Conclusions:

- Recalling the KKT conditions, we see the number of nonzero components in any solution at most  $|S| \leq \min\{n, p\}$
- Further, we can reduce our optimization problem (by partially solving) to

$$\min_{\beta_S \in \mathbb{R}^{|S|}} f(X_S \beta_S) + \lambda \|\beta_S\|_1$$

- Finally, strict convexity implies uniqueness of the solution in this problem, and hence in our original problem

□

22