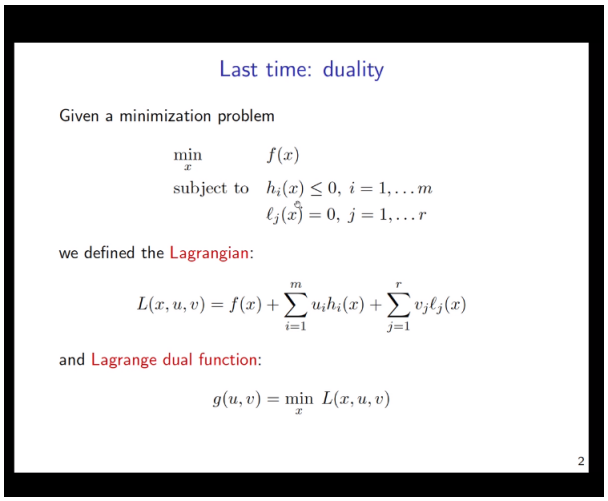
Recap



Last time: duality

Given a minimization problem

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

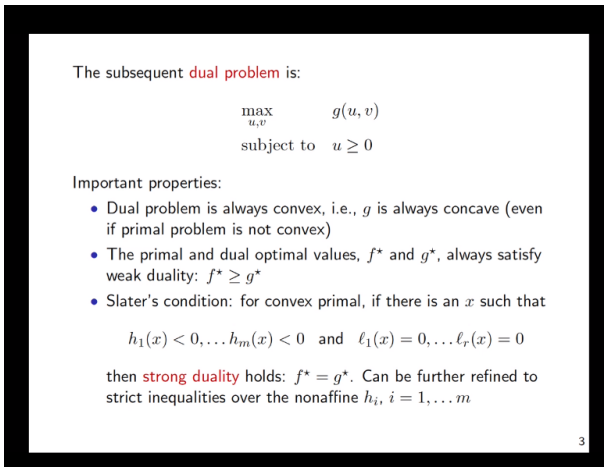
we defined the Lagrangian:

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

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, $f^\star$ and $g^\star$, always satisfy weak duality: $f^\star \geq g^\star$
- Slater's condition: for convex primal, if there is an x such that

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

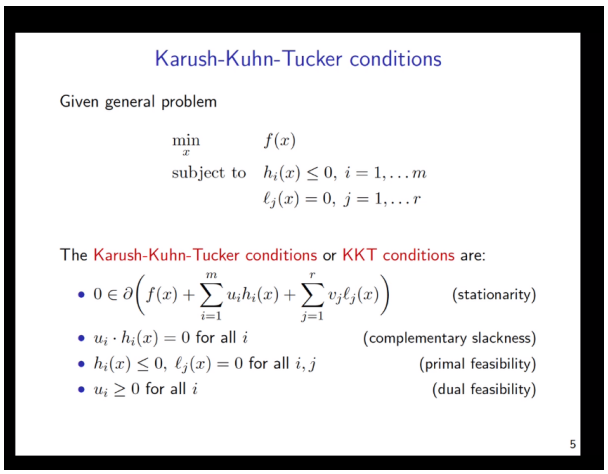
then strong duality holds: $f^\star = g^\star$. Can be further refined to strict inequalities over the nonaffine h_i, $i = 1, \dots m$

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



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

$$u_i . h_i(x) = 0 \ \forall \ i$$

3. Primal feasibility
x has to satisfy all the primal constraints

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

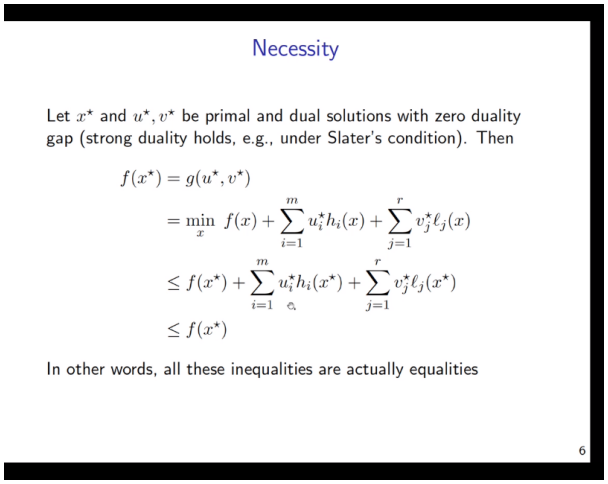
4. Dual feasibility
u has to satisfy the dual constraints (for the inequality constraints in the primal problem)

$$u_i \geqslant 0 \ \forall \ i$$

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

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

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 then minimum

$f(x^*) = g(u^*, v^*)$ *from Slater's condition*
$= min_x\, L(x, u, v)$ *by defintion*
$= min_x\, f(x) + \sum u_i^* h_i(x) + \sum v_j^* l_j(x)$, *evaluated at optimal u and v*
$\leqslant f(x^*) + \sum u_i^* h_i(x^*) + \sum v_j^* l_j(x^*) = L(x^*, u^*, v^*)$
$\leqslant f(x^*)$
Which means they are all equalities
Hence, $min_x\, L(x, u, v) = L(x^*, u^*, v^*)$

By subgradient optimality, we know that at $min_x\, L(x, u, v)$,
$0 \in \partial_x(L(x, u, v))$
so $0 \in \partial_x\left(L(x^*, u^*, v^*)\right)$
$\implies 0 \in \partial_x\left(f(x^*) + \sum u_i^* h_i(x^*) + \sum v_j^* l_j(x^*)\right)$
which is the stationarity condition
So the **stationarity** condition must hold at a solution!

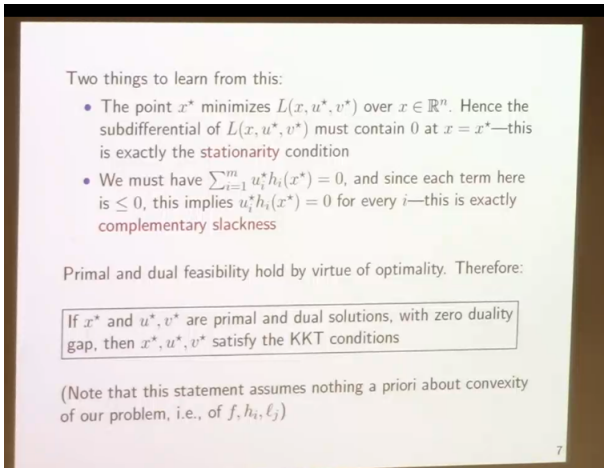
Then observe that

$f(x^*) + \sum u_i^* h_i(x^*) + \sum v_j^* l_j(x^*) = f(x^*)$ *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 $u_i \geqslant 0$, *and* $h_i(x) \leqslant 0 \; \forall \; i$
So every term in the sum has the same sign i.e. every term is $\leqslant 0$
So the terms can't cancel out – therefore if they sum to 0, *each individaul term must be zero, which is the complementaty 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 satisdy the KKT conditions.
→ under strong duality, the KKT conditions must hold



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*

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

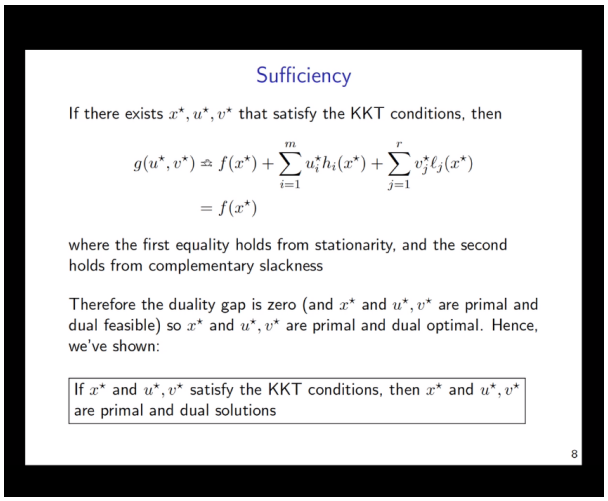
so $g(u^*, v^*) = f(x^*) + \sum u_i^* h_i(x^*) + \sum v_j^* l_j(x^*)$

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

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

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

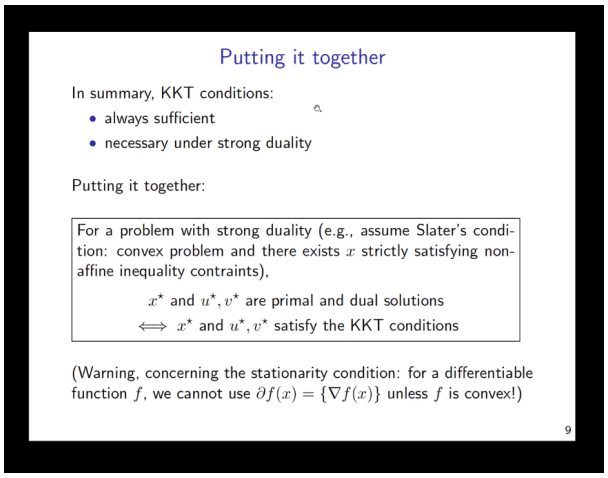
$$g(u^\star, v^\star) = f(x^\star) + \sum_{i=1}^{m} u_i^\star h_i(x^\star) + \sum_{j=1}^{r} v_j^\star \ell_j(x^\star)$$
$$= f(x^\star)$$

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

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 contraints),

$x^\star$ and $u^\star, v^\star$ are primal and dual solutions
$\iff$ $x^\star$ and $u^\star, v^\star$ 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:

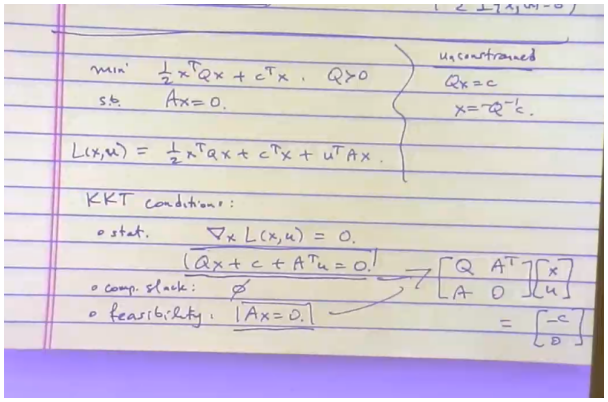
- First appeared in publication by Kuhn and Tucker in 1951
- Later people found out that Karush had the conditions in his unpublished master's thesis of 1939

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

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

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



Example: quadratic with equality constraints

Consider for $Q \succeq 0$,

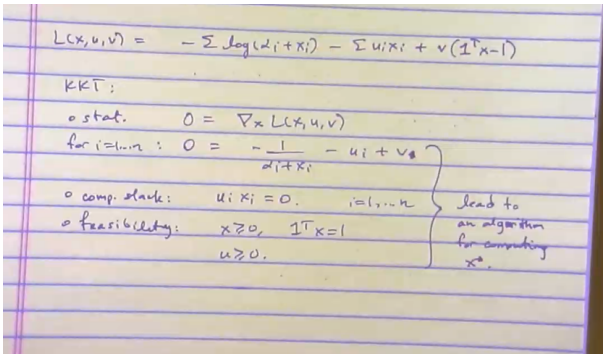
$$\begin{array}{ll} \min\limits_x & \frac{1}{2}x^TQx + c^Tx \\ \text{subject to} & Ax = 0 \end{array}$$

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

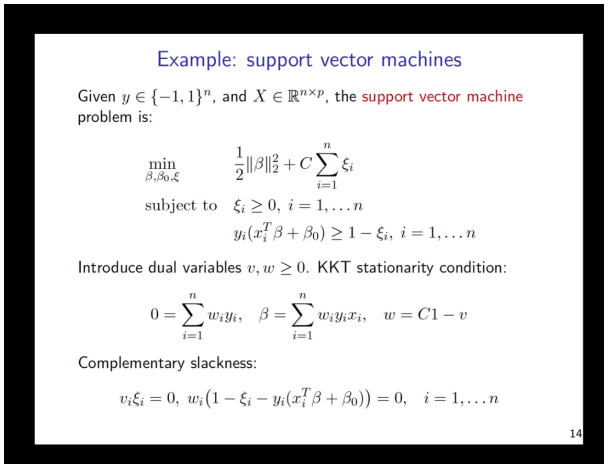
Convex problem, no inequality constraints, so by KKT conditions: x is a solution if and only if

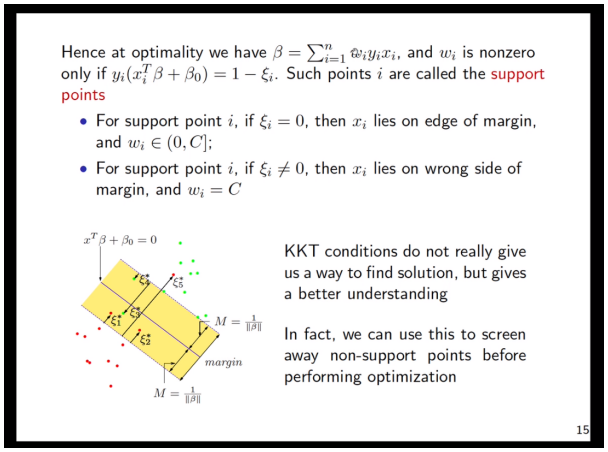
$$\begin{bmatrix} Q & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} x \\ u \end{bmatrix} = \begin{bmatrix} -c \\ 0 \end{bmatrix}$$

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



B. Example: SVM, again





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 doesnt matter

But by compliementary slackness, $w_i\left(1 - \xi_i - y_i\left(x_i^T B + B_0\right)\right) = 0$. So w_i can only be non-zero if $\left(1 - \xi_i - y_i\left(x_i^T B + B_0\right)\right) = 0$. 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.

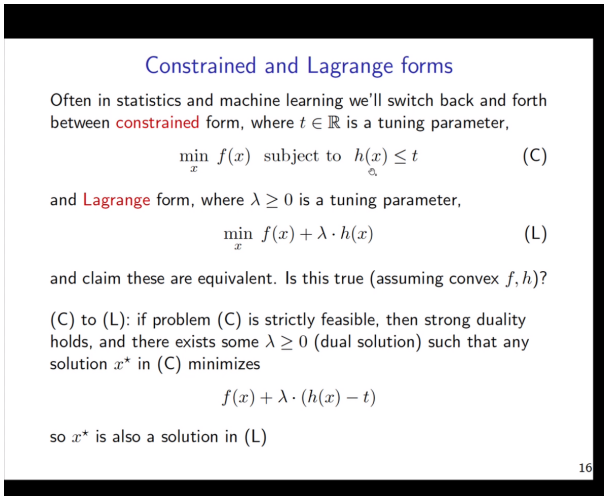
Case 1: For a support point i, if $\xi_i = 0$ (i.e the point is on the right side of the margin) then v_i can be any number. ($\geqslant$ 0, as per the constraints on the dual problem)
But from KKT stationarity, we know that $w = c.1 - v$ which means that $w_i \in (0, C]$

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 $y_i\left(x_i^T B + B_0\right)$?
- 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



Is the problem $min_x\ f(x)\ subject\ to\ h(x)\ \leqslant\ t,\ t \in R$ (constrained form), equivalent to $min_x\ f(x)\ +\ \lambda h(x),\ \lambda\ \geqslant\ 0$ (Lagrange form)?
- This comes from KKT conditions

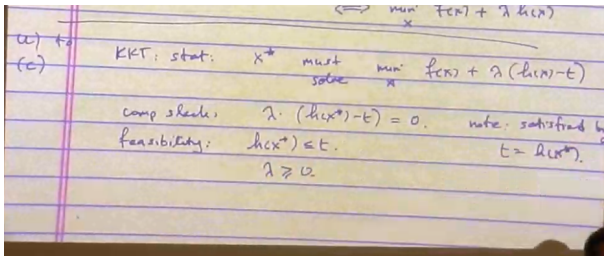
C. Constrained → Lagrangian form

Constrained problem
$min_x\ f(x)$
$subject\ to$
$h(x)\ \leqslant\ t$

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 $min_x\ \{f(x)\ +\ \lambda(h(x)\ -\ t)\}\ \Longrightarrow\ min_x\ \{f(x)\ +\ \lambda h(x)\}$

Lagrangian → 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



(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:

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

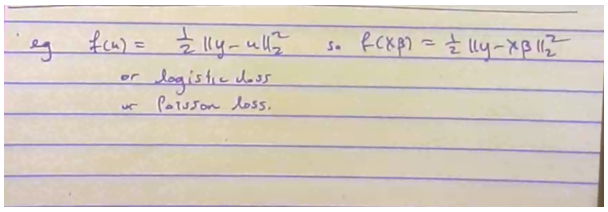
$$\bigcup_{\lambda \geq 0} \{\text{solutions in (L)}\} \quad \supseteq \quad \bigcup_{\substack{t \text{ such that (C)} \\ \text{is strictly feasible}}} \{\text{solutions in (C)}\}$$

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

17

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



Note that we need strict convexity in the loss function f(u=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

$$f(X;\ tB_1 + (1-t)B_2) = \frac{1}{2}||y - X(tB_1 + (1-t)B_2)||_2^2$$

$$= \frac{1}{2}(y - X(tB_1 + (1-t)B_2)^T(y - X(tB_1 + (1-t)B_2)$$

$$= \frac{1}{2}\left(y^T - (tB_1 + (1-t)B_2)^T X^T\right)(y - X(tB_1 + (1-t)B_2))$$

$$= \frac{1}{2}(y^T y - y^T X(tB_1 + (1-t)B_2) - (tB_1 + (1-t)B_2)^T X^T y +$$

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

$$= \frac{1}{2}(y^T y - ty^T XB_1 + (1-t)y^T XB_2 - tB_1^T X^T y - (1-t)B_2{}^T X^T y +$$

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

RHS

$$tf(X;B_1) = \frac{1}{2}t\left((y - XB_1)^T(y - XB_1)\right)$$

- $$= \frac{1}{2}t\left(y^T y - y^T XB_1 - B_1^T X^T y + B_1^T X^T XB_1\right)$$

$$(1-t)f(X;\ B_2) = \frac{1}{2}(1-t)\left(y^T y - y^T XB_2 - B_2^T X^T y + B_2^T X^T XB_2\right)$$

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

$$= \frac{1}{2}\left(y^T y - ty^T XB_1 - (1-t)\, y^T XB_2 - tB_1^T X^T y - (1-t)B_2^T X^T y + tB_1^T X^T XB_1 + (1-t)B_2^T X^T\right.$$

Need to compare

$$(tB_1 + (1-t)B_2)^T X^T X(tB_1 + (1-t)B_2)\ ?\ tB_1^T X^T XB_1 + (1-t)B_2^T X^T XB_2$$

LHS

$$t^2 B_1 X^T XB_1 + t(1-t)B_1^T X^T XB_2 + t(1-t)B_2^T X^T XB_1 + (1-t)^2 B_2^T X^T XB_2$$

$$= t^2 B_1 X^T XB_1 + (1-t)^2 B_2^T X^T XB_2 + 2t(1-t)B_1^T X^T XB_2$$

$$= ||tXB_1 + (1-t)XB_2||_2^2 = L2norm($$

RHS

$$= t||XB_1||_2^2 + (1-t)||XB_2||_2^2$$

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

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?

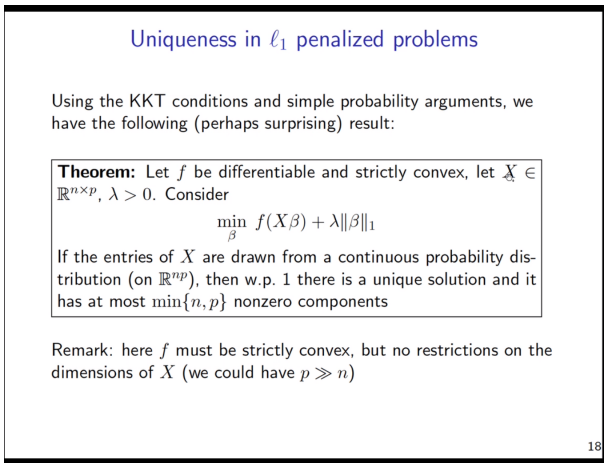
$$f(x) = x^T x$$
$$f'(x) = 2x$$
$$f''(x) = 2I > 0$$

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 $\left[\frac{dy}{dx_1}, \frac{dy}{dx_2}, ..., \frac{dy}{dx_p}\right]$. So take the expression of the final value and get the index-wis

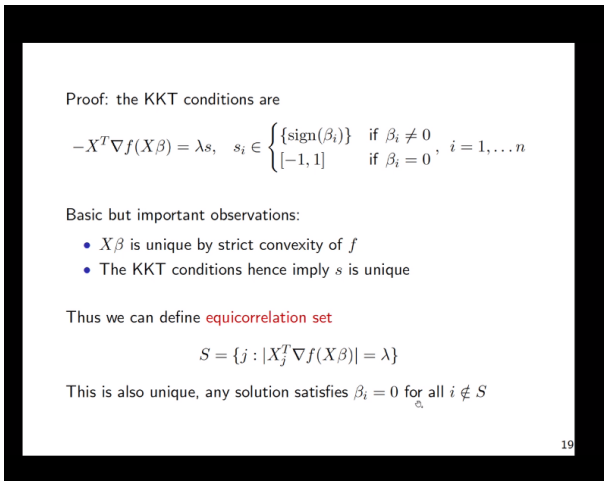


Saturation of L1 regularized solution
Uniqueness without strict convexity here

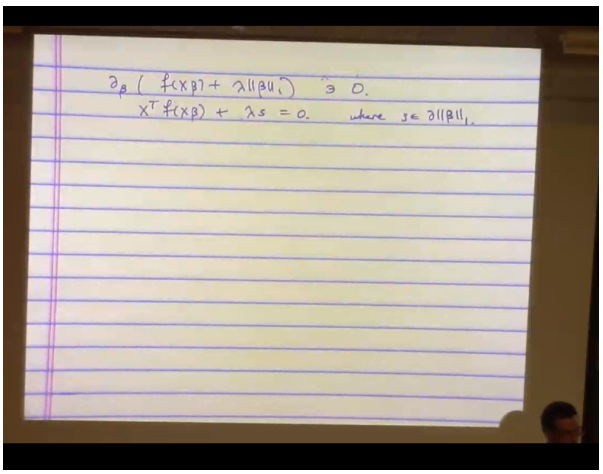
- If n≪ p, X will have a null space, so the unconstrained problem has infintely many solutions
- Q: why does having a null space mean this?
- Because for a given solution B, you can add B+a such that a is in the null space of X,

for infinite values of a, with some scaling b

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



- 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 fuction 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

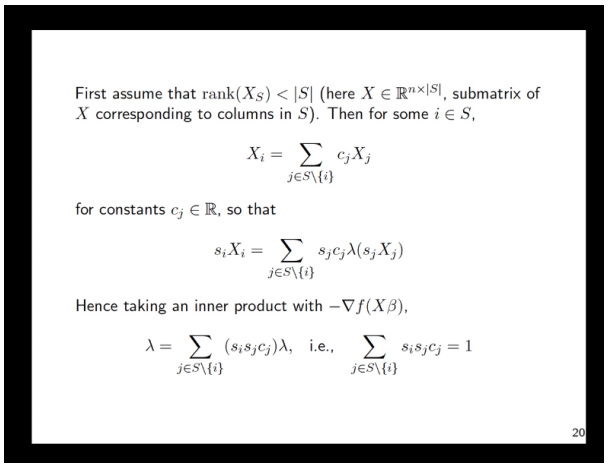


- typo in the sheet - that should be ∇f

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

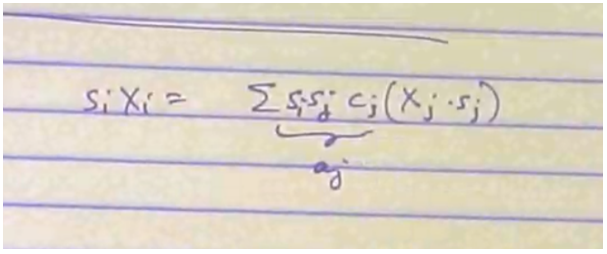
Statement: The submatrix of X with columns in S is linearly independent.
Proof by contradiction:

- 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 rank(X_s) < |S|



- 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

–



- 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 sX_i with $-\nabla f(XB)$ will yield λ. → why is this true for the subset matrix?