446 Section 4

Plans for today!

1. Reminders
2. Gradient Descent
3. Generalized Least Squares
4. Importance of Regularization in Least Squares
5. Convexity
6. Ridge/LASSO (if time)

Reminders

- HW1 was due yesterday
  - Remember that you have 5 late days!
- HW2 was released yesterday; due Wednesday, May 6

Some tips:

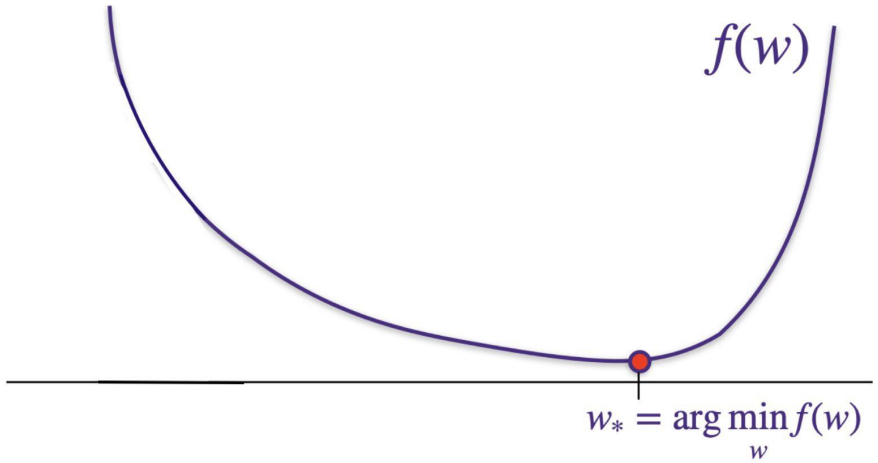
- Use office hours to your advantage
  - Student TA OH for homework questions
  - Professor OH for conceptual questions
  - Motivates you to get things done on time, starting an untouched assignment can be daunting

Gradient Descent

Gradient Descent

Purpose of this exercise: Understanding how gradient descent relates to approximations, and why it works.

Consider some function $f(w)$, which has some w_* for which $w_* = \arg\min_w f(w)$:



Question 1a

Let w_0 be some initial guess for the minimum of $f(w)$. Gradient descent will allow us to improve this solution.

(a) For some w that is very close to w_0, give the Taylor series approximation for $f(w)$ starting at $f(w_0)$.

Remember Taylor expansion?

↳ To approximate a function around a point a

$$f(x) \approx f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \frac{f'''(a)}{3!}(x-a)^3 \cdots$$

Exact at a, close around a

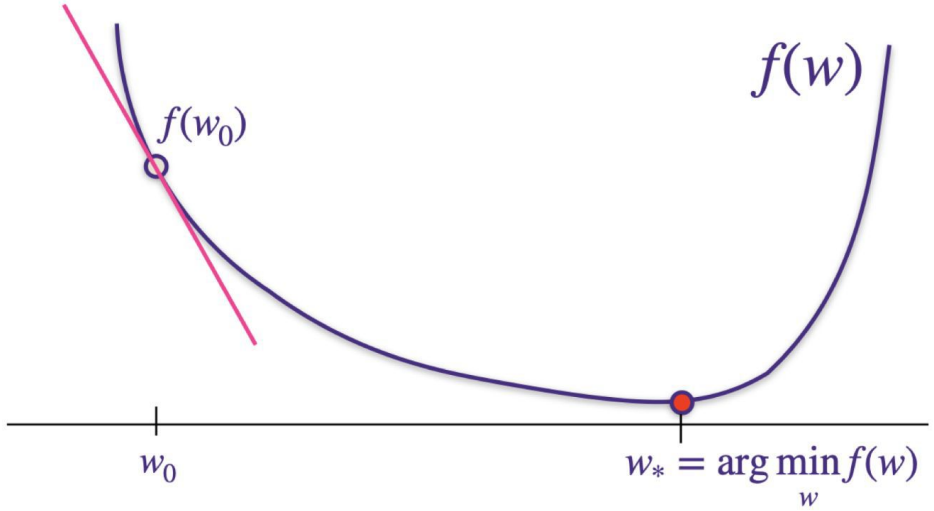
Better and better approximations

Question 1a

Let w_0 be some initial guess for the minimum of $f(w)$. Gradient descent will allow us to improve this solution.

(a) For some w that is very close to w_0, give the Taylor series approximation for $f(w)$ starting at $f(w_0)$.

For w very close to w_0, we see that $f(w) \approx f(w_0) + (w - w_0)\left(\frac{df(w)}{dw}\Big|_{w=w_0}\right)$.



Question 1b

(b) Now, let us choose some $\eta > 0$ that is *very small*. With this very small η, let's assume that $w_1 = w_0 - \eta \left(\frac{df(w)}{dw} \Big| _{w=w_0} \right)$. Using your approximation from part (a), give an expression for $f(w_1)$.

Hint: Plug in here

$$f(w) \approx f(w_0) + (w - w_0) \left(\frac{df(w)}{dw} \Big| _{w=w_0} \right).$$

Fancy way of saying f'(w_0)

(Derivative of f(w) at w_0)

Question 1b

$$w_1 = w_0 - \eta \left(\frac{df(w)}{dw} \,\middle|\, _{w=w_0} \right)$$ ← Given

plug in here

$$f(w_1) \approx f(w_0) + (w_1 - w_0) \left(\frac{df(w)}{dw} \,\middle|\, _{w=w_0} \right)$$

$$= f(w_0) + \left(w_0 - \eta \left(\frac{df(w)}{dw} \,\middle|\, _{w=w_0} \right) - w_0 \right) \left(\frac{df(w)}{dw} \,\middle|\, _{w=w_0} \right)$$

$$= f(w_0) - \eta \left(\frac{df(w)}{dw} \,\middle|\, _{w=w_0} \right)^2$$

Question 1c

(c) Given your expression for $f(w_1)$ from part (b), explain why, if η is small enough and if the function approximation is a good enough approximation, we are guaranteed to move in the "right" direction closer to the minimum w_*.

Remember: We want to minimize this

Hint: Why would this be good?

$$f(w_1) \approx f(w_0) - \eta \left(\frac{df(w)}{dw} \Big|_{w=w_0} \right)^2$$

Question 1c

Note that in part (b), the derivative is squared and will always be a nonnegative value. Therefore, $f(w_1) < f(w_0)$.

$$f(w_1) \approx f(w_0) - \eta \left(\frac{df(w)}{dw} \Big| _{w=w_0} \right)^2$$

In English: The loss function after a weight update will always evaluate to be smaller than before the weight update

- If the step size is small enough
- If the approximation is good enough

Question 1d

(d) Building from your answer in part (c), write a general form for the gradient descent algorithm.

Hint: how could we generalize this equation from part b?

$$w_1 = w_0 - \eta \left(\frac{df(w)}{dw} \bigg|_{w=w_0} \right)$$

Question 1d

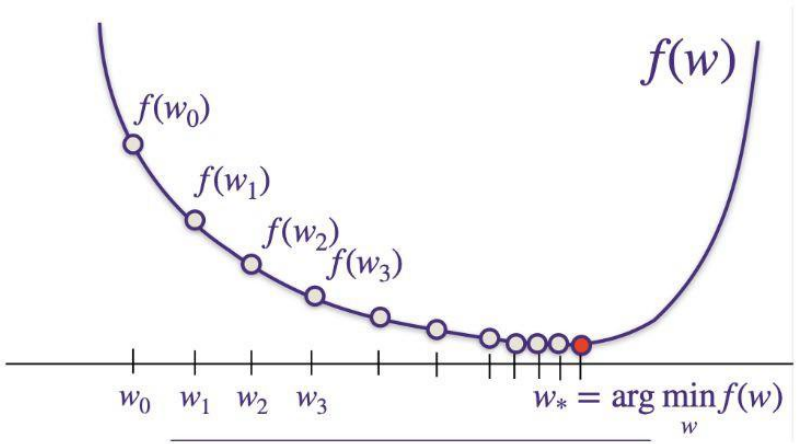
Gradient descent is written as:

For $k = 0, 1, 2, 3, \ldots$, $w_{k+1} = w_k - \eta \left(\frac{df(w)}{dw} \Big|_{w=w_k} \right)$.

Note that as $k \to \infty$, $\left(\frac{df(w)}{dw} \Big|_{w=w_k} \right) \to 0$.

Convergence guarantees iff convex!

We visualize as:



Gradescope Section Participation

Suppose you are guaranteed to find the global minimum when your gradient descent algorithm converges. What assumption can you make about the loss function?

Convexity

Generalized Least Squares

Least Squares Proof(s)

Should look familiar…

$$\widehat{\omega}_{\text{general}} = (X^\top X + \lambda D)^{-1} X^\top y$$

Has shown up…

- In lecture (Lecture 2)
- On your homework (A5 Ridge Regression proof)
- And now here!

$$\widehat{\omega}_{\text{general}} = \left(\sum_{i=1}^{n} x_i x_i^\top + \lambda D \right)^{-1} \left(\sum_{i=1}^{n} x_i y_i \right)$$

You can look at the generalized proof in your own time.

Question 2.2a

$$\widehat{\omega}_{\text{general}} = (X^\top X + \lambda D)^{-1} X^\top y$$

(a) In the simple least squares case ($\lambda = 0$ above), what happens to the resulting $\widehat{\omega}$ if we double all the values of y_i?

Solution:

As can be seen from the formula $\widehat{\omega} = (X^\top X)^{-1} X^\top y$, doubling y doubles ω as well. This makes sense intuitively as well because if the observations are scaled up, the model should also be.

Question 2.2b

$$\widehat{\omega}_{\text{general}} = (X^\top X + \lambda D)^{-1} X^\top y$$

(b) In the simple least squares case ($\lambda = 0$ above), what happens to the resulting $\widehat{\omega}$ if we double the data matrix $X \in \mathbb{R}^{n \times d}$?

Solution:

As can be seen from the formula $\widehat{\omega} = (X^\top X)^{-1} X^\top y$, doubling X halves ω. This also makes sense intuitively because the error we are trying to minimize is $\|X\omega - y\|_2^2$, and if the X has doubled, while y has remained unchanged, then ω must compensate for it by reducing by a factor of 2.

Importance of Regularization in Least Squares

Question 2.2c

$$\widehat{\omega}_{\text{general}} = (X^\top X + \lambda D)^{-1} X^\top y$$

(c) Suppose $D = I$ (that is, it is the identity matrix). That is, this is the *ridge* regression setting. Explain why $\lambda > 0$ ensures that the solution exists and the matrix can be inverted.

2.2c setup

Let's do a linear algebra refresher so that we can show off an interesting and actually useful result about the utility of regularization!

$$A : \mathbb{R}^d \to \mathbb{R}^n, \text{ if } d \gg n$$

d

$$A : \begin{bmatrix} \vdots & \vdots & \vdots \end{bmatrix} n$$

must have a non-empty nullspace
(can show using rank-nullity theorem)

This is bad for invertibility

Note: $Null(A) = Null(A^T A)$

$\hookrightarrow$ proof in Section 02 handout

Invertible means $(A^T A)^{-1}(A^T A) = I$, so $(A^T A)^{-1}(A^T A)x = x$

Way to think about nullspaces

$A \in \mathbb{R}^{n \times d} \quad x \in \mathbb{R}^d$

Nullspace: Subspace of $\mathbb{R}^d$, contains all solutions to $Ax = 0$

In other words, all the vectors are "annihilated" by A

Invertible means $(A^\top A)^{-1}(A^\top A) = I$, so $(A^\top A)^{-1}(A^\top A)x = x$

If $A^\top A$ has a non-empty nullspace, then

$$\exists x \text{ s.t. } (A^\top A)x = 0$$

Makes $(A^\top A)^{-1}(A^\top A)x = x$ impossible!

If this $= 0$, no way to recover x!

Main idea: If $X \in \mathbb{R}^{n \times d}, d \gg n,$ then

$\text{Null}(X)$ and $\text{Null}(X^\top X)$ are non-empty

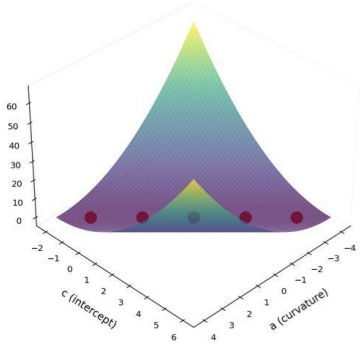
This means $X^\top X$ has no inverse

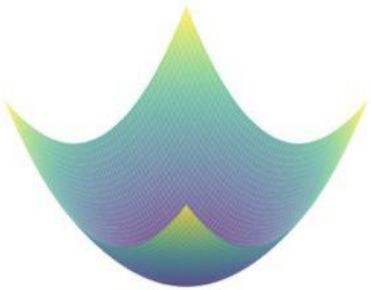
An issue because... $\hat{w} = (X^\top X)^{-1} X^\top y$

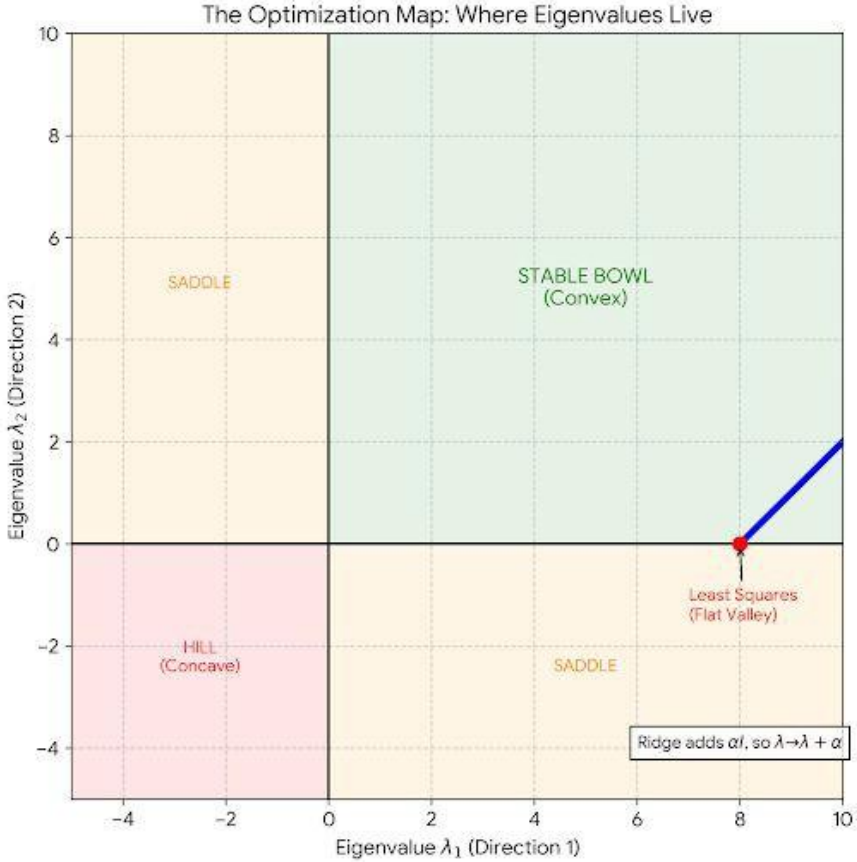
Let's not give up!

Visualized

- If X has a non-empty null space, matrix X^TX has an eigenvalue of 0. This corresponds to a “flat valley” – no unique solution
- By adding I, we shift all the eigenvalues, moving it into a “stable bowl”







Let's add in $\lambda I : \hat{w} = (X^T X + \lambda I)^{-1} X^T Y$

Is $X^T X + \lambda I$ always invertible for $\lambda > 0$?

A matrix A is positive semi-definite if $x^\top Ax \geq 0$ and positive definite if $x^\top Ax > 0$
Positive Definite (PD): All eigenvalues are strictly positive
Positive Semi-Definite (PSD): All eigenvalues are 0 or positive

We want to show that $u^\top (X^\top X + \lambda I) u > 0 \quad \forall u \in \mathbb{R}^d$

$\hookrightarrow$ Show matrix is positive definite, meaning it must have an inverse

We want to show that $u^\top (X^\top X + \lambda I)u > 0 \quad \forall u \in \mathbb{R}^d$

$$u^\top (X^T X + \lambda I)u = \underbrace{u^\top (X^T X)u}_{\text{Is this always } >0?} + u^\top (\lambda I)u$$

L2 norm

$$u^\top (X^\top X)u = u^\top X^\top X u = \|Xu\|_2^2 \geq 0 \quad \text{Yes!}$$

$$u^\top (X^\top X + \lambda I)u = u^\top (X^\top X)u + u^\top (\lambda I)u$$

$$\geq u^\top (\lambda I)u \leftarrow \text{this is PD, } \therefore > 0$$

$$> 0$$

We have shown $X^T X + \lambda I$ is PD and therefore always invertible if $\lambda > 0$!

$\hookrightarrow$ <u>Even if $d \gg n$!</u>

Question 2.2c

$$\widehat{\omega}_{\text{general}} = (X^\top X + \lambda D)^{-1} X^\top y$$

(c) Suppose $D = I$ (that is, it is the identity matrix). That is, this is the *ridge* regression setting. Explain why $\lambda > 0$ ensures that the solution exists and the matrix can be inverted.

Solution:

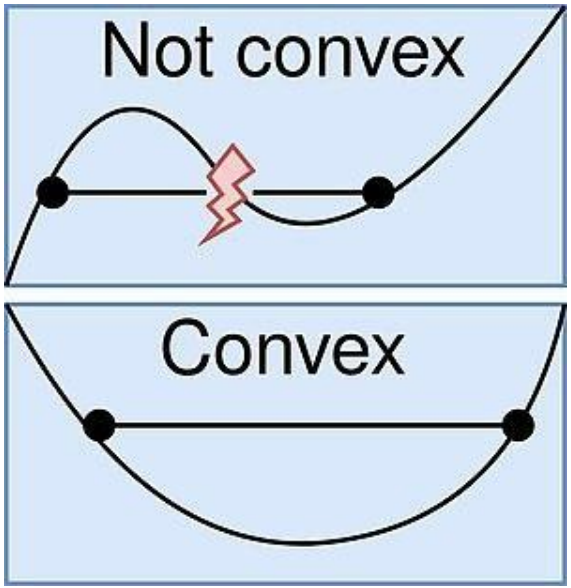
The solution is $\widehat{\omega} = (X^\top X + \lambda I)^{-1} X^\top y$. We already saw in a previous part that $X^\top X$ is always positive semidefinite, that is, its eigenvalues are at least zero. Adding λI, where $\lambda > 0$, ensures that $X^\top X + \lambda I$ is in fact positive *definite*. This helps us have a good condition number.

Convexity

Convexity in functions

Definition 2 (convex functions). *A function* $f : \mathbb{R}^d \to \mathbb{R}$ *is* ***convex*** *on a set* A *if for all* $x, y \in A$ *and* $\lambda \in [0, 1]$*:*

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y)$$



The function between x and y

Must be less than or equal to

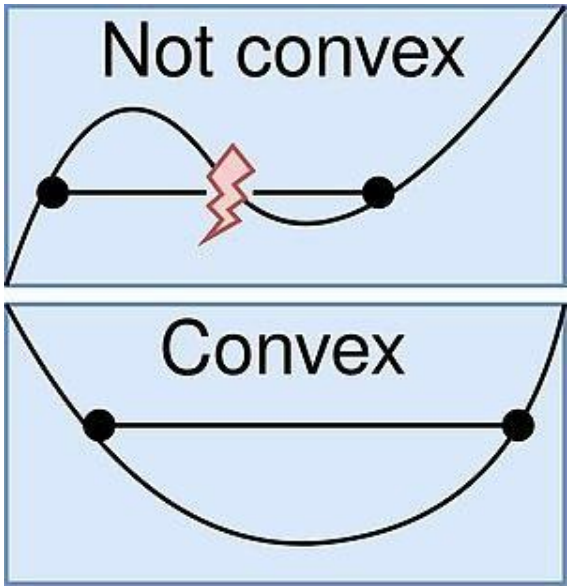
A straight line between x and y

Note: The sum of convex functions is convex

Convexity in functions

Definition 2 (convex functions). *A function* $f : \mathbb{R}^d \to \mathbb{R}$ *is* ***convex*** *on a set* A *if for all* $x, y \in A$ *and* $\lambda \in [0, 1]$*:*

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y)$$



Guarantees that any local minimum we find will be as low as the global minimum

If you perform GD with a small step size on a convex loss function, you **will** reach the best possible performance!

Problem 3b

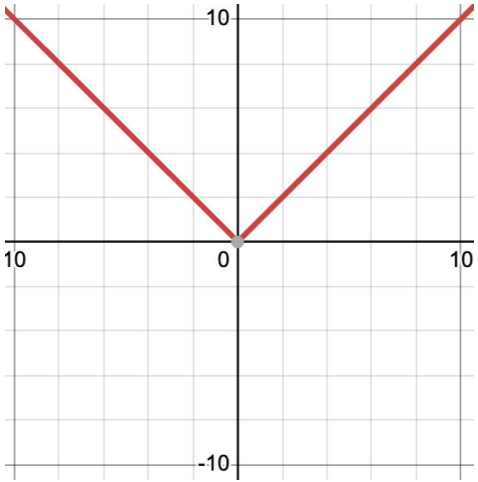
(b) Which of the following functions are convex? (Hint: draw a picture!)

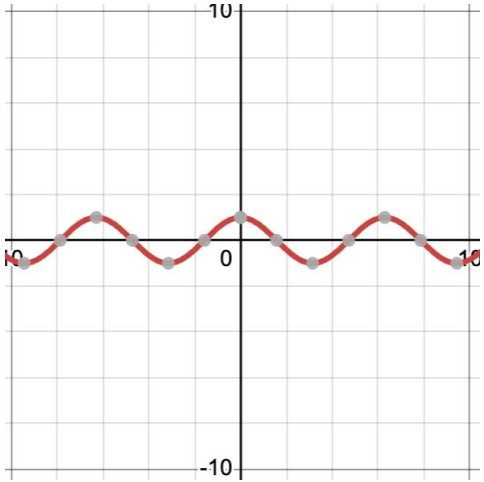
(i) $|x|$

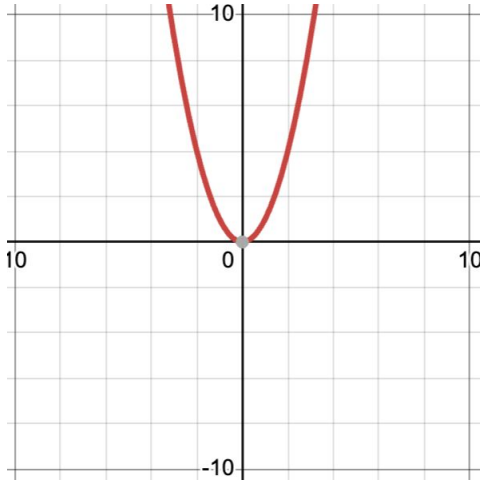
(ii) $\cos(x)$

(iii) $x^T x$

Note: Convex over all real numbers







Problem 3b

$|x|$ and $x^T x$ are both convex. $\cos(x)$ is not convex since we can draw a line at two points (from say $\frac{\pi}{2}$ to $2\pi + \frac{\pi}{2}$) that is not entirely above the function.

Proof that $|x|$ is convex:

$$\begin{aligned} f(\lambda x + (1-\lambda)y) &= |\lambda x + (1-\lambda)y| \\ &\leq \lambda|x| + (1-\lambda)|y| \end{aligned}$$

Proof that $x^T x$ is convex:
We begin by examining the definition: whenever $\lambda \in [0,1]$, we have

$$\begin{aligned} (\lambda x + (1-\lambda)y)^T(\lambda x + (1-\lambda)y) &= \lambda^2 x^T x + (1-\lambda)^2 y^T y + 2\lambda(1-\lambda)x^T y \\ &= \lambda x^T x + (1-\lambda)y^T y - \lambda(1-\lambda)(x^T x - 2x^T y + y^T y) \\ &= \lambda x^T x + (1-\lambda)y^T y - \lambda(1-\lambda)(x-y)^T(x-y) \\ &\leq \lambda x^T x + (1-\lambda)y^T y, \end{aligned}$$

where the inequality holds because $(x-y)^T(x-y) = ||x-y||_2^2 \geq 0$. So our function is convex.

Problem 4b

(b) If f is twice differentiable with convex domain, then f is convex if and only if

$$\nabla^2 f(x) \succeq 0,$$

for any x in the domain of f. Use this method to show that the objective function in linear regression is convex.

$$f(w) = Y^T Y - 2w^T X^T Y + w^T X^T X w \longrightarrow \nabla f(w) = -2X^T Y + 2X^T X w \longrightarrow \nabla^2 f(w) = 2X^T X$$

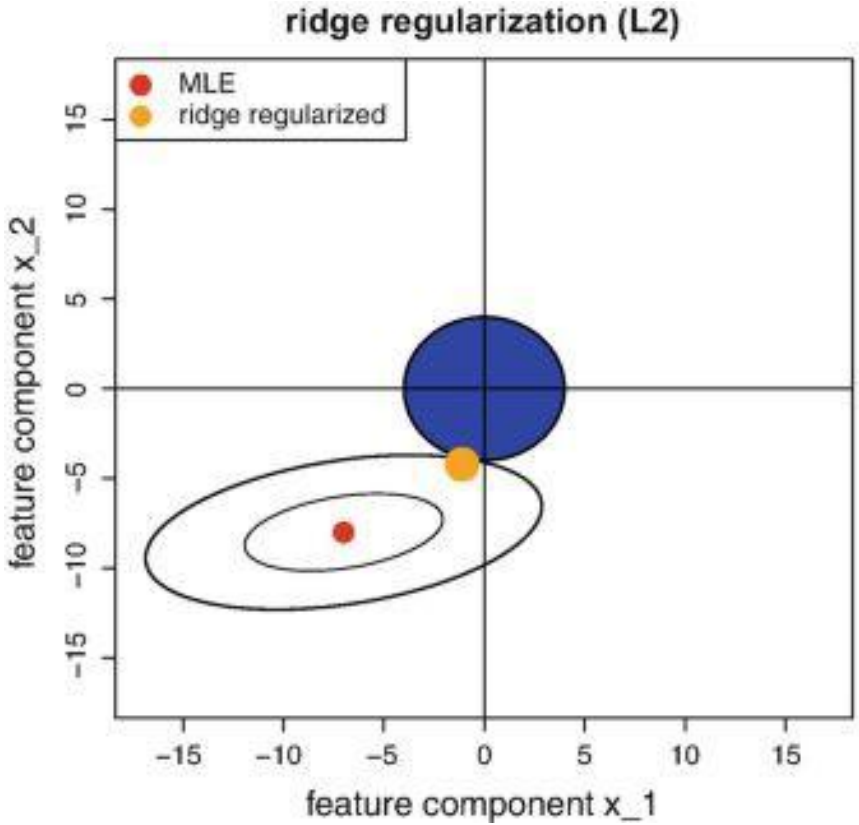
Solution:

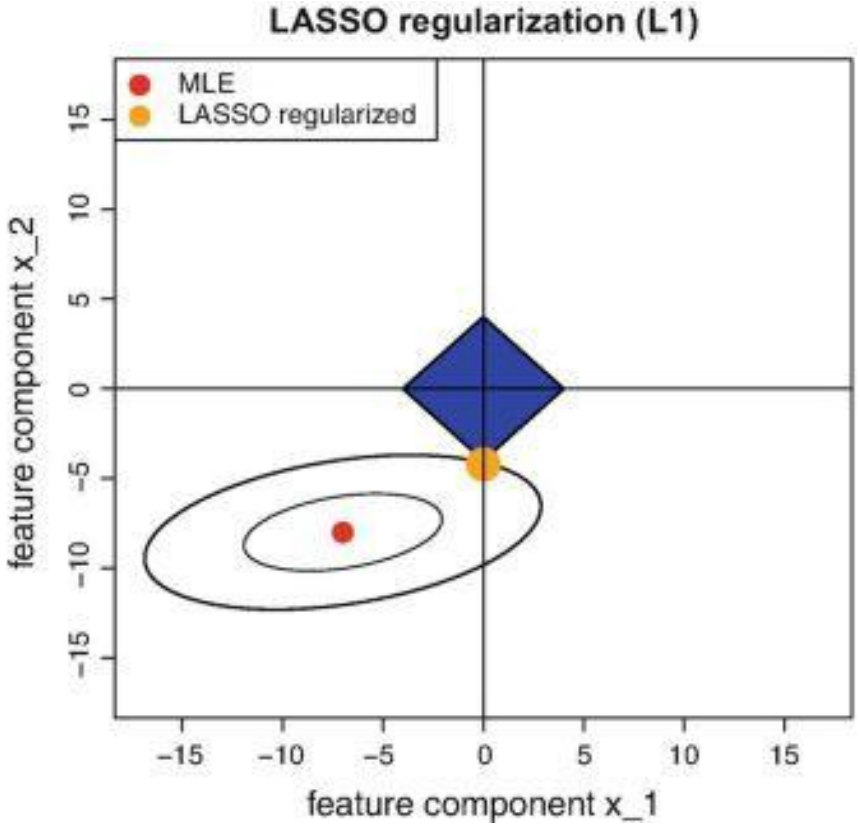
Let $f(w) = (Y - Xw)^T (Y - Xw)$, then

$$\nabla^2 f(w) = 2(X^T X),$$

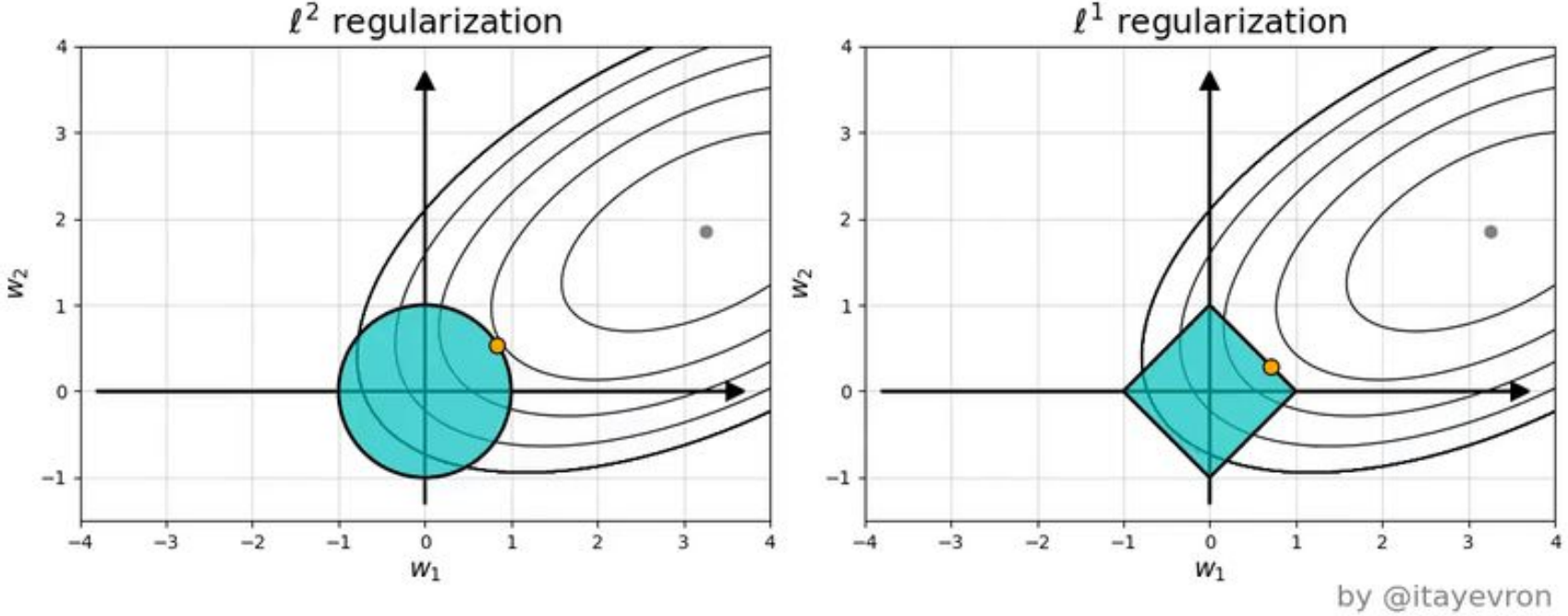
which is clearly a positive semidefinite matrix.

Ridge vs. LASSO


ridge regularization (L2)
MLE
ridge regularized
feature component x_2
feature component x_1
15
10
5
0
−5
−10
−15


LASSO regularization (L1)
MLE
LASSO regularized
feature component x_2
feature component x_1
15
10
5
0
−5
−10
−15

ℓ^1 induces sparse solutions for least squares



Questions/Chat Time!