

Important announcements

Midterm details:

- 25 multiple choice (select all that apply) questions using scantron (Bring a pencil!!)
- Covers up to (and including) today lecture up to (and including) sec 6.2 in the course notes
- Formula sheet will be provided (no cheat sheet) will post tonight
- You can bring a calculator
- Similar to assignment questions and examples in course notes
- we will use lecture on Thursday to do a review session for the midterm exam.

=> Our goal (from supervised learning lecture)

Defining $A(D)$: Empirical Risk Minimization (ERM)

Estimation:

Use D to estimate $L(f)$ for all $f \in \mathcal{F} \subset \{f/f: \mathcal{X} \rightarrow \mathcal{Y}\}$
call the estimate $\hat{L}(f)$

Optimization:

pick $\hat{f}$ to be the $f \in \mathcal{F}$ that minimizes $\hat{L}(f)$

Function
class

Optimization

finding the best solution from a set of possible solutions

Usually this means finding the minimum or maximum value of some function

we will care about:

$$\min_{w \in W} g(w)$$

minimum value of $g(w)$
over all $w \in W$

or $w^* = \operatorname{argmin}_{w \in W} g(w)$

the $w \in W$ that achieves
the minimum value of $g(w)$

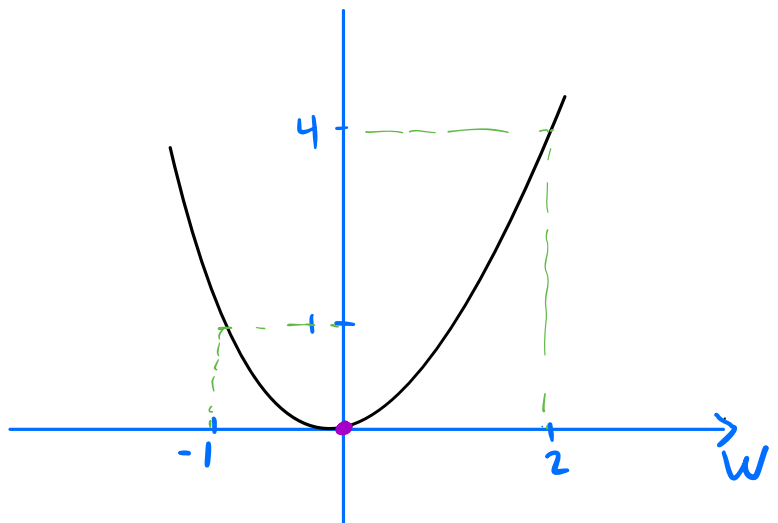
w^* is a "minimizer"

$$\min_{w \in W} g(w) = g(w^*)$$

Ex: $g(w) = w^2 \quad W = \mathbb{R}$

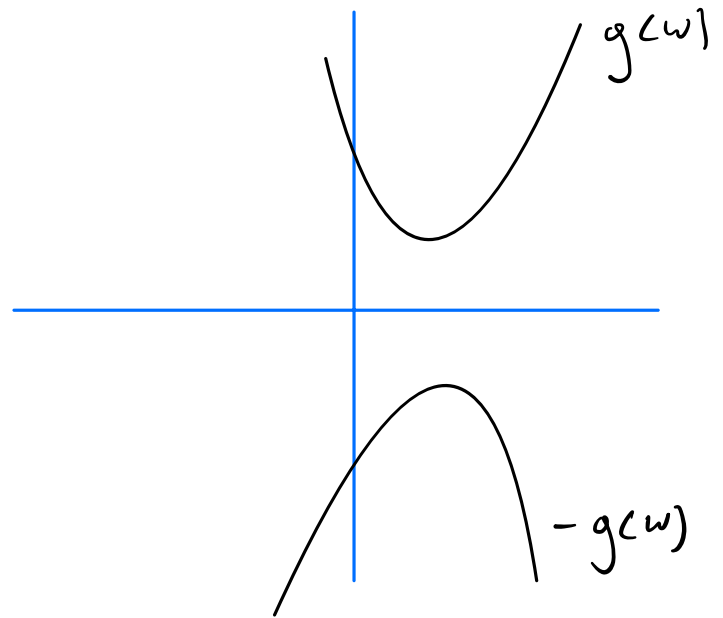
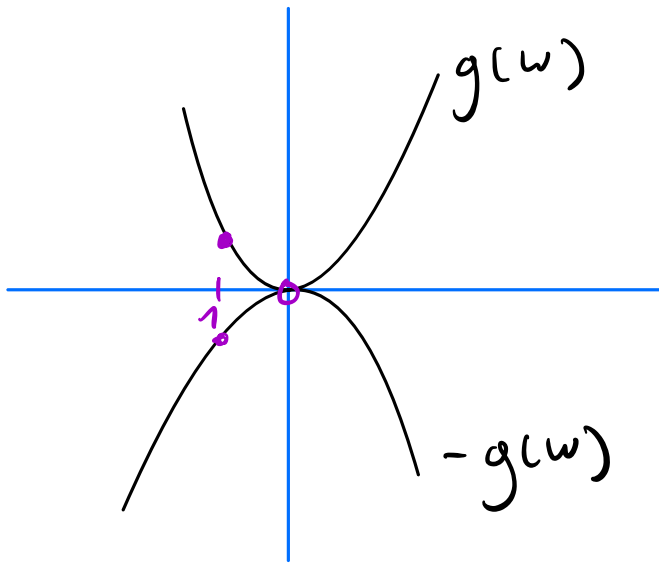
$$w^* = \operatorname{argmin}_{w \in W} g(w) = 0$$

$$\min_{w \in W} g(w) = 0 = g(w^*)$$



$$W = \{-1, 2\} \quad \min_{w \in W} g(w) = 1 \quad \arg\min_{w \in W} g(w) = -1 = w^* \\ = g(w^*)$$

Note: There is a relationship between minimizing and maximizing



$$w^* = \arg\min_{w \in W} g(w) = \arg\max_{w \in W} -g(w)$$

$$g(w^*) = \min_{w \in W} g(w) = - \left(\max_{w \in W} -g(w) \right) = -(-g(w^*)) = g(w^*)$$

$\Rightarrow$ We can choose if we want to maximize a function or minimize its negative (of that function)

How do we solve minimization problems?

Cases:

1. If $\mathcal{W}$ is discrete

we compare $g(w)$ for all $w \in \mathcal{W}$

2. If $\mathcal{W}$ is continuous we can sometimes use derivatives

$\Rightarrow$ we will focus on $\mathcal{W}$ continuous

Additional assumptions:

If $g(w)$ is convex and twice differentiable

then:

Cases: 1. If $\mathcal{W} = \mathbb{R}$ then w^* is the solution to $g'(w) = 0$

2. If $\mathcal{W} = [a, b]$ then w^* is the solution to $g'(w) = 0$ if this solution is in $[a, b]$
Otherwise, w^* is a or b .

Twice differentiable: The second derivative of $g(w)$ written $g''(w)$ exists for all $w \in \mathcal{W}$

Convex: $g(w)$ is convex if $g''(w) \geq 0$ for all $w \in \mathcal{W}$

"usually $g(w)$ is bowl-shaped"

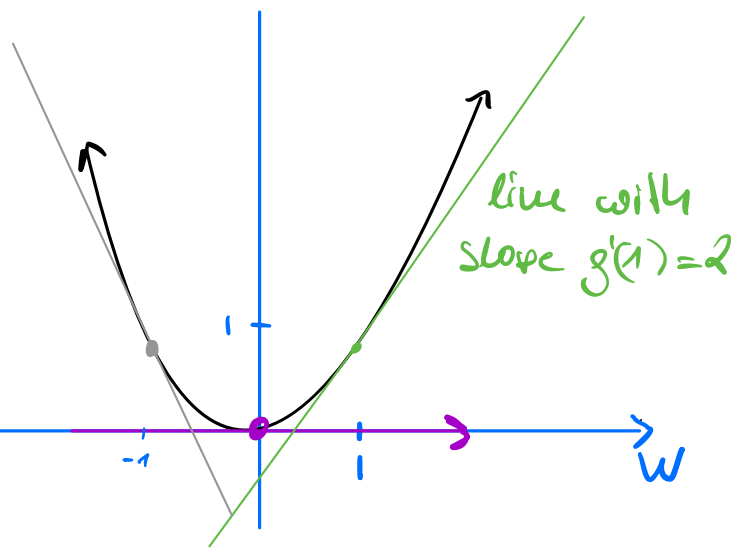
Ex: $g(w) = w^2$, $\mathcal{W} = \mathbb{R}$

$$w^* = \underset{w}{\operatorname{argmin}} g(w)$$

$$g'(w) = 2w, \quad g''(w) = 2$$

$$\Rightarrow g(w) \text{ is convex: } g''(w) = 2 \geq 0$$

$$g'(w) = 2w = 0 \rightarrow w^* = 0$$



$$g'(1) = 2 \cdot 1$$

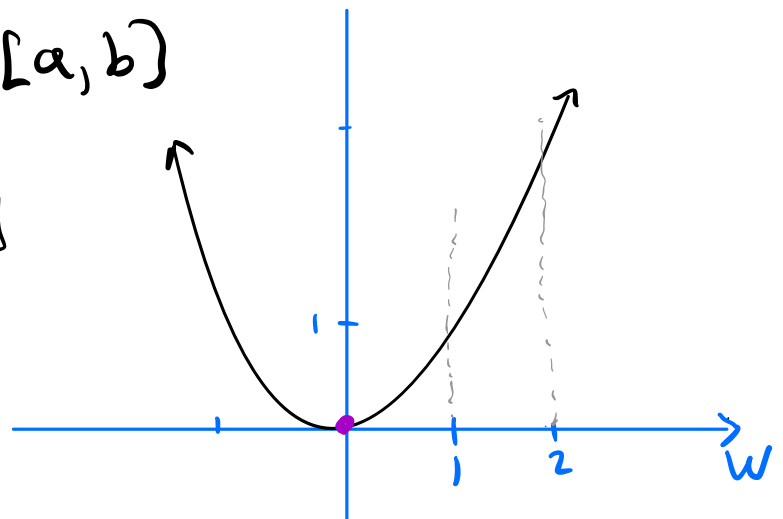
$$g'(-1) = 2 \cdot (-1) = -2$$

$$g'(0) = 0$$

Ex: $g(w) = w^2$, $\mathcal{W} = [1, 2] = [a, b]$

$$g'(w) = 2w = 0 \rightarrow w = 0 \notin [1, 2]$$

$$g(1) = 1, \quad g(2) = 4$$



$$\omega^* = 1 \rightarrow g(\omega^*) = \min_{\omega} g(\omega) =$$

Ex: $g(\omega) = \omega^3, \mathcal{W} = \mathbb{R}$

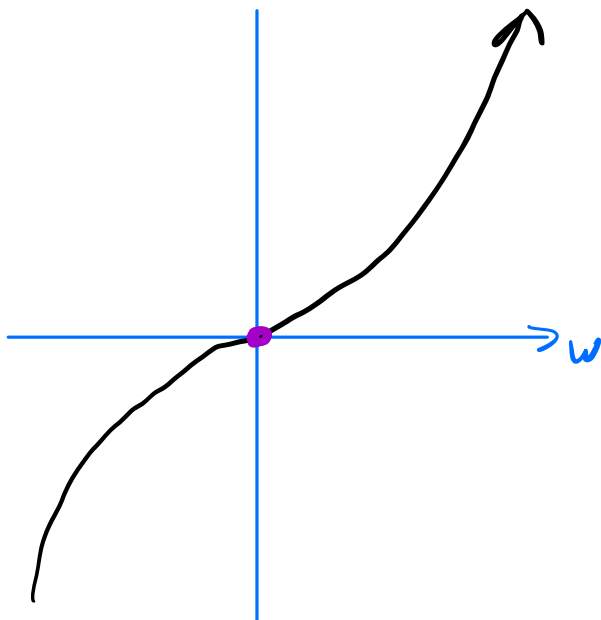
$$g'(\omega) = 3\omega^2, \quad g''(\omega) = 6\omega$$

$$g''(\omega) < 0$$

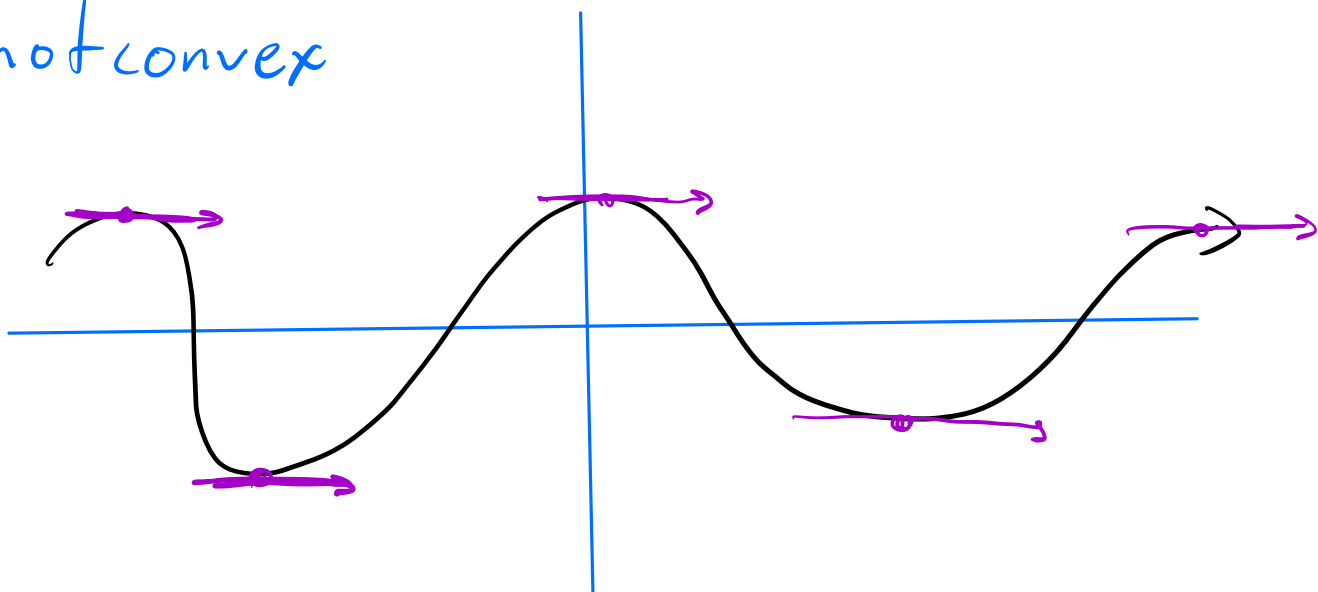
when $\omega < 0$

$$g'(\omega) = 3\omega^2 = 0$$

not the minimum



Ex: not convex



Multidimensional Minimization

If $\mathcal{W} = \mathbb{R}^d$ for $d > 1$, and $g(\vec{w})$ is convex

Note: it is more complex to check if $g(\vec{w})$ is convex if $d > 1$. So, I will tell you.

then we calculate

$$\vec{w}^* = (w_1^*, \dots, w_d^*)^T = \underset{\vec{w} \in \mathcal{W}}{\operatorname{argmin}} g(\vec{w})$$

by setting w_j^* as the solutions to

$$\frac{\partial g(\vec{w})}{\partial w_j} = 0 \quad \text{for all } j \in \{1, \dots, d\}$$

Ex: $g(\vec{w}) = g(w_1, w_2) = w_1^2 + w_2^2,$

$$\mathcal{W} = \mathbb{R}^2$$

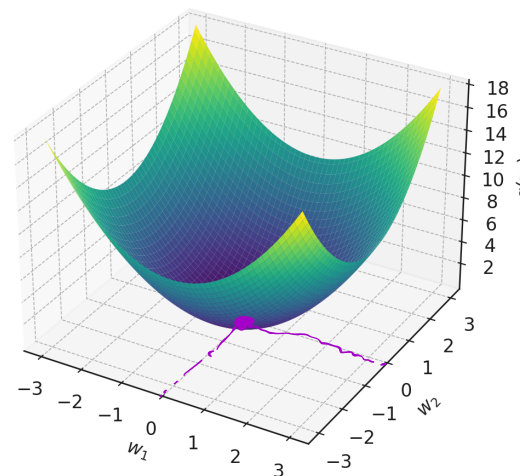
$$\frac{\partial g(\vec{w})}{\partial w_1} = 2w_1 = 0 \Rightarrow w_1^* = 0$$

$$\frac{\partial g(\vec{w})}{\partial w_2} = 2w_2 = 0 \Rightarrow w_2^* = 0$$

$$\vec{w}^* = (w_1^*, w_2^*)^T = (0, 0)^T$$

$$g(\vec{w}^*) = \min_{\vec{w}} g(\vec{w})$$

$$g(\mathbf{w}) = w_1^2 + w_2^2$$



Finding a good predictor (Linear Regression)

Optimization step of ERM

$$\mathcal{X} = \mathbb{R}^{d+1}, y = \mathbb{R} \quad (\text{regression})$$

$$\mathcal{F} \subset \{f \mid f: \mathbb{R}^{d+1} \rightarrow \mathbb{R}\}$$

$$\mathcal{D} = ((\vec{x}_1, y_1), \dots, (\vec{x}_n, y_n)) \quad \text{fixed dataset}$$

$$\hat{L}(f) = \frac{1}{n} \sum_{i=1}^n \ell(f(\vec{x}_i), y_i) \quad \begin{array}{l} \text{estimate of } L(f) \\ \text{for all } f \in \mathcal{F} \end{array}$$

(sample mean)

what we want

$$\hat{f} = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \hat{L}(f)$$

$$\text{pick } \mathcal{F} = \{f \mid f: \mathcal{X} \rightarrow \mathcal{Y} \text{ and } f(\vec{x}) = \vec{x}^T \vec{w} \text{ where } \vec{w} \in \mathbb{R}^{d+1}\}$$

$\rightarrow$ "linear functions"

$$\begin{aligned} f(\vec{x}) &= \vec{x}^T \vec{w} + b = (x_1, \dots, x_d)^T (w_1, \dots, w_d) + b \\ &= \overset{b = w_0}{b} + x_1 w_1 + \dots + x_d w_d \\ &= \overset{x_0 = 1}{x_0} w_0 + x_1 w_1 + \dots + x_d w_d \\ &= (x_0, x_1, \dots, x_d)^T (w_0, w_1, \dots, w_d) \\ &= \vec{x}_0^T \vec{w}_0 \end{aligned}$$

Assume $\vec{x} = (x_0=1, x_1, \dots, x_d) \in \mathbb{R}^{d+1}$

$$\vec{x}_i = (x_{i,0}=1, x_{i,1}, \dots, x_{i,d}) \in \mathbb{R}^{d+1} \text{ (dataset)}$$

Notice $f \in \tilde{F}$ so $f(\vec{x}) = \vec{x}^T \vec{w}$ for some $\vec{w} \in \mathbb{R}^{d+1}$

$$\vec{w}^1 = \underset{\substack{\vec{w} \in \mathbb{R}^{d+1} \\ \downarrow \\ \mathcal{W}}}{\operatorname{argmin}} \quad \underset{\substack{\downarrow \\ \mathcal{G}}}{L}(\vec{w}) \quad \text{where} \quad L(\vec{w}) = \frac{1}{n} \sum_{i=1}^n \ell(\vec{x}_i^T \vec{w}, y_i)$$