Important announcements

Feb 13

- midterm marks will be released on the weekend or early next week
- release solutions tonight
- Thank you for filling out the midterm evaluation survey

=> Our goal (from supervised learning lecture)

Defining $\mathcal{A}(D)$: Empirical Risk Minimization (ERM)

Estimation:

Use D to estimate $L(f)$ for all $f \in \mathcal{F} \subset \{f \mid f: \mathcal{X} \to \mathcal{Y}\}$

call the estimate $\hat{L}(f)$

Optimization:

pick $\hat{f}$ to be the $f \in \underbrace{\mathcal{F}}_{\text{Function class}}$ that minimizes $\hat{L}(f)$

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:

$$\min_{w \in \mathcal{W}} g(w)$$ minimum value of $g(w)$ over all $w \in \mathcal{W}$

or $$w^* = \underset{w \in \mathcal{W}}{\operatorname{argmin}}\, g(w)$$ the $w \in \mathcal{W}$ that achieves the minimum value of $g(w)$

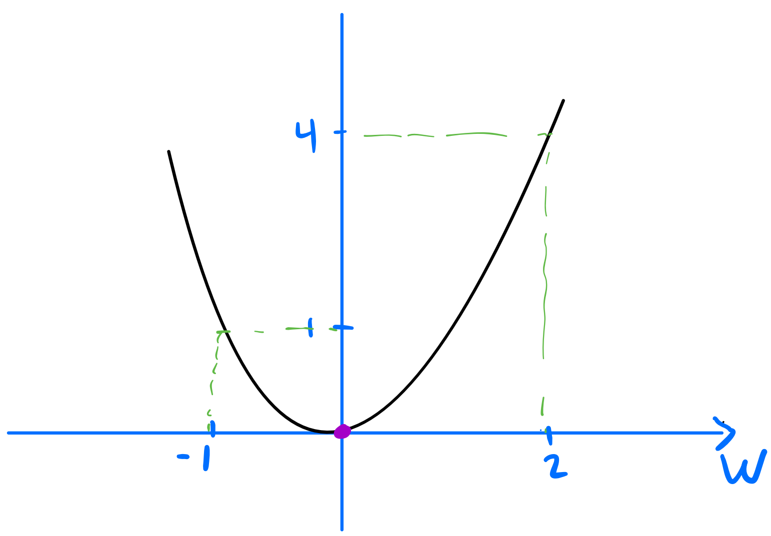
w^* is a "minimizer"

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

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

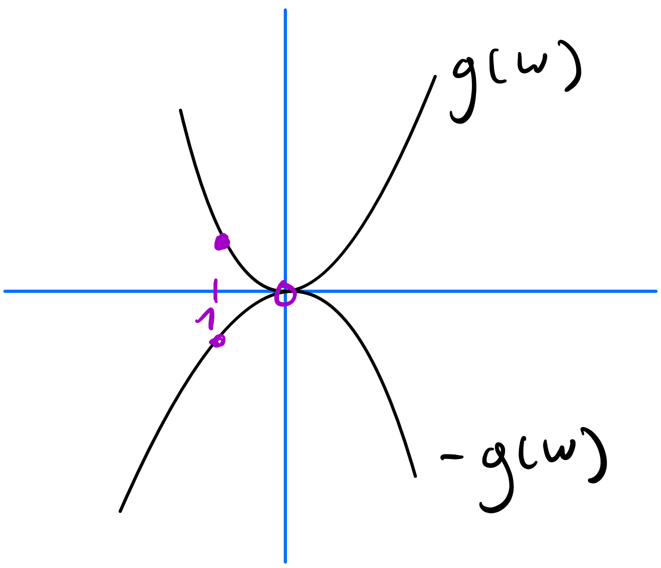
$$w^* = \operatorname{argmin}\, g(w) = 0$$

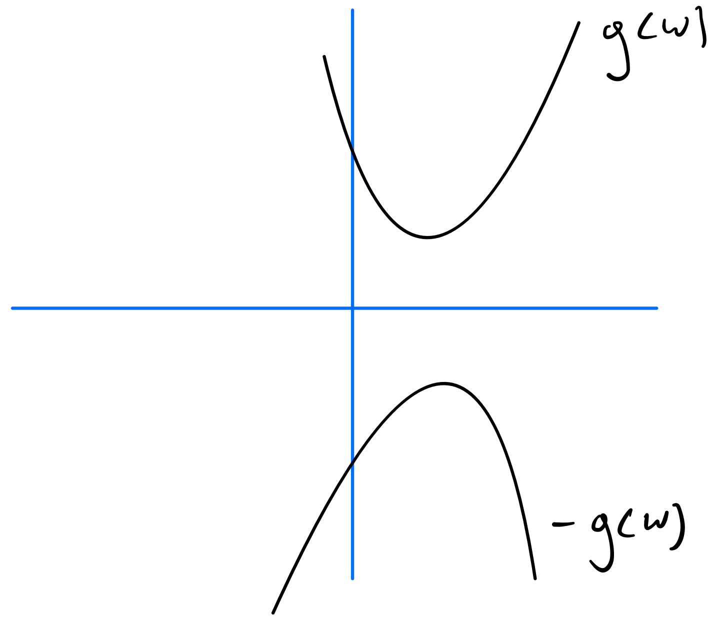
$$\min_{w \in \mathcal{W}} g(w) = 0 = g(w^*)$$



$$W = \{-1, 2\} \qquad \min_{w \in W} g(w) = 1 = g(w^*) \qquad \underset{w \in W}{\operatorname{argmin}}\, g(w) = -1 = w^*$$

Note: There is a relationship between minimizing and maximizing





$$w^* = \underset{w \in W}{\operatorname{arg\,min}}\, g(w) = \underset{w \in W}{\operatorname{arg\,max}} -g(w)$$

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

=> 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 $g'(w) = 0$

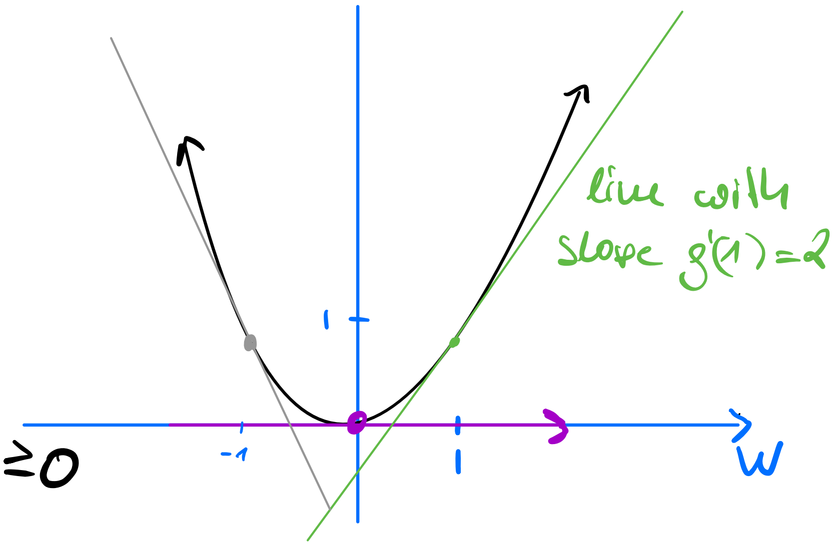
2. If $\mathcal{W} = [a, b]$ then w^* is the solution to $g'(w) = 0$ if this solution is in $[a, b]$
   Otherwise, w^* is a or b.

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



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

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

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

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

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

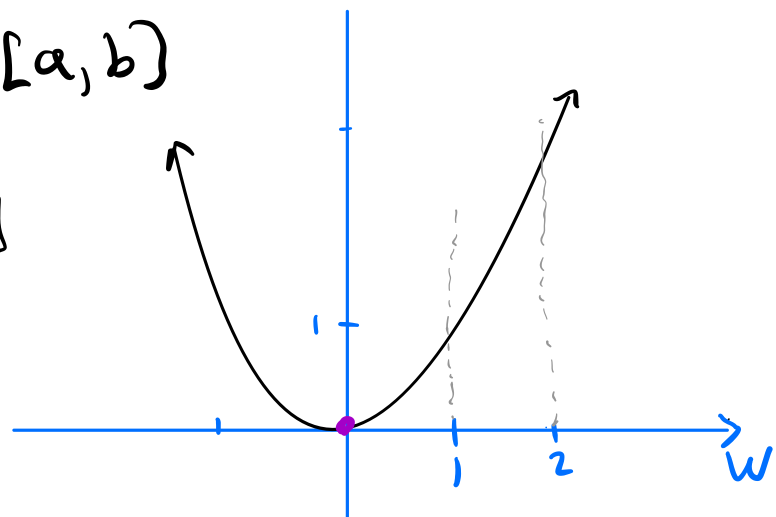
$g'(-1) = 2 \cdot (-1) = -2$

$g'(0) = 0$

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

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

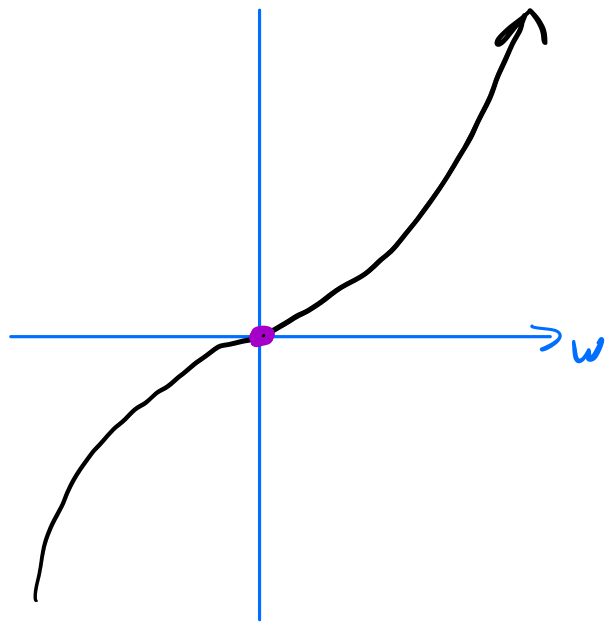
$g(1) = 1$, $g(2) = 4$



$w^* = 1 \to g(w^*) = \min\limits_{w} g(w) =$

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

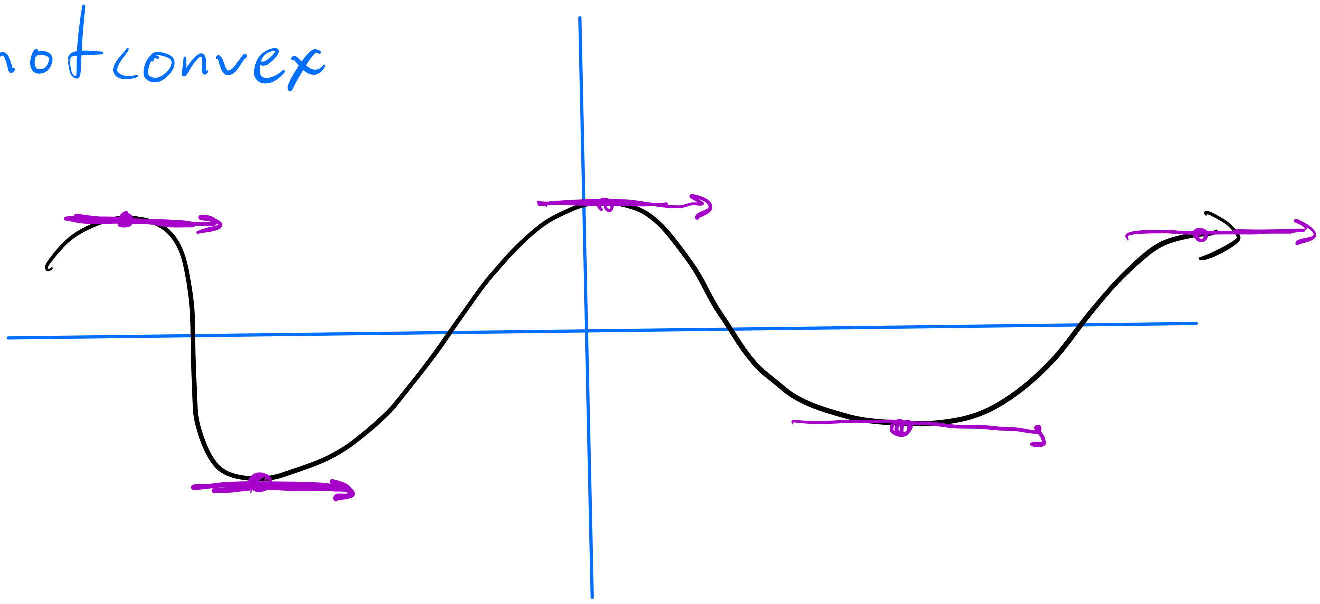
$g'(w) = 3w^2$, $g''(w) = 6w$



$g''(w) < 0$ when $w < 0$

$g'(w) = 3w^2 = 0$ → w of the minimum

Ex: not convex



Multidimensional Minimization

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

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

then we calculate

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

by setting w_j^* as the solution to

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

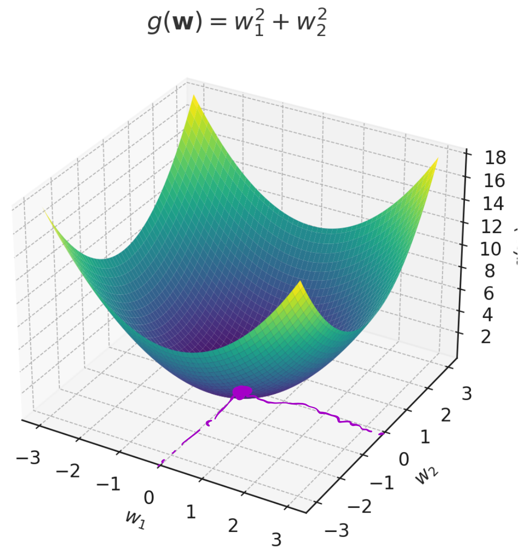
Ex: $g(\vec{w}) = g(w_1, w_2) = w_1^2 + w_2^2$, $\quad \mathcal{W} = \mathbb{R}^2$

$$\frac{\partial g(\vec{\omega})}{\partial w_1} = 2\omega_1 = 0 \Rightarrow \omega_1^* = 0$$

$$\frac{\partial g(\vec{\omega})}{\partial \omega_2} = 2\omega_2 = 0 \Rightarrow \omega_2^* = 0$$

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

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



Finding a good predictor (Linear Regression)

Optimization step of ERM

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

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

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

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

(sample mean)

what we want

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

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

$\to$ "linear functions"

$$f(\vec{x}) = \vec{x}^T \vec{w} + b = (x_1, \dots, x_d)^T (w_1, \dots, w_d) + b$$

$$= \overset{\Rightarrow w_0}{b} + x_1 w_1 + \dots + x_d w_d$$

$b = w_0$

$x_0 = 1$

$$= x_0 w_0 + x_1 w_1 + \dots + x_d w_d$$

$$= (x_0, x_1, \dots, x_d)^T (w_0, w_1, \dots, w_d)$$

$$= \vec{x}_0^T \vec{w}_0$$

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

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

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

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

Solve for $\hat{w}$:

$$\frac{\partial \hat{L}}{\partial w_j}(\vec{w}) =$$

$$\frac{\partial \hat{L}}{\partial w_j}(\vec{w}) =$$