Important announcements

Feb 25

- grades for midterm 1 are out
- example for linear regression

Gradient Descent

Can we always find a closed form expression for $w^* = \underset{w \in \mathcal{W}}{\arg\min}\ g(w)$? $\Rightarrow$ No !

Linear regression is closed form because of the squared loss & linear function class

Ex: $g(\omega) = \omega^2 + e^w$, $g'(\omega) = 2\omega + e^\omega$, $g''(\omega) = 2 + e^w \geq 0$

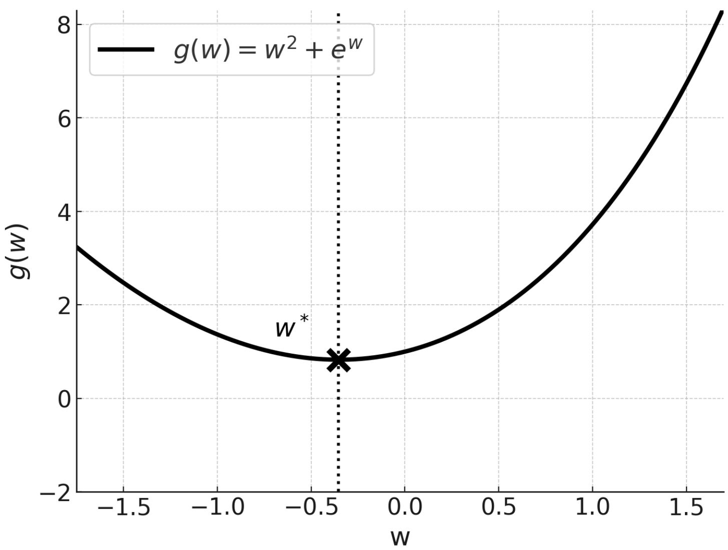
g is convex

$g'(\omega) = 2\omega + e^w = 0$

$\rightarrow 2\omega = -e^w$

No way to solve for w !

-> No closed form solution although g(w) convex.



Gradient descent helps with this problem

Second-Order Gradient Descent (Newton-Raphson)

If $g(w)$ is a degree 5 polynomial or less

→ then there exists a closed form solution for $g(w)=0$

Let's approximate $g(w)$ with a convex low degree polynomial (i.e. 2nd degree polynomial)

In general, the Taylor series at a point $w^{(0)}$ of $g(w)$ is

$$g(w) = \sum_{n=0}^{\infty} \frac{g^{(n)}(w^{(0)})}{n!}(w - w^{(0)})^n$$

We use a 2nd order approximation that takes the first 3 terms

→ not exactly $g(w)$ but good enough around $w^{(0)}$

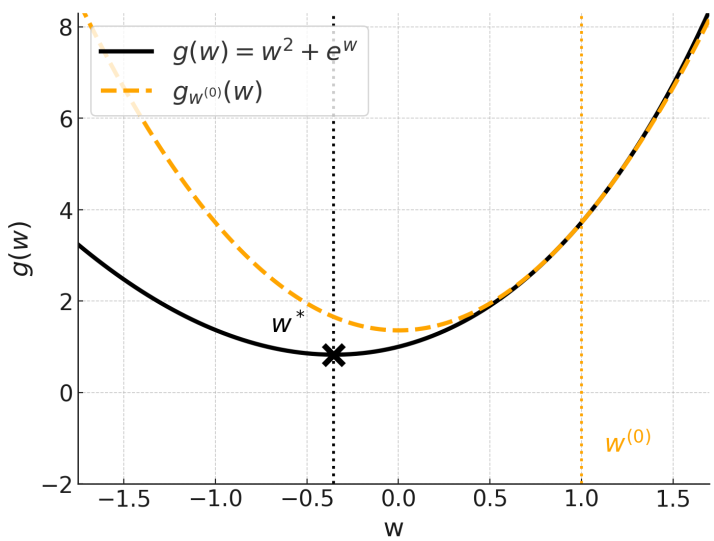
$$g(w) \approx g_{w^{(0)}}(w) = \underbrace{g(w^{(0)})}_{\text{const}} + \underbrace{g'(w^{(0)})}_{\text{const}}(w - \overset{\text{const}}{w^{(0)}}) + \underbrace{\frac{g''(w^{(0)})}{2}}_{\text{const}}(w - \underbrace{w^{(0)}}_{\text{const}})^2$$

→ 2nd order polynomial

Ex: $g(w) = w^2 + e^w$

$w^{(0)} = 1$ ← arbitrary choice

$g_{w^{(0)}}(w) = g_1(w)$



minimize $g_{w^{(0)}}(w)$:

$$\frac{d}{dw} g(w) \approx \frac{d}{dw} g_{w^{(0)}}(w) = g'(w^{(0)}) + g''(w^{(0)})(w - w^{(0)}) = 0$$

$$\Rightarrow g''(w^{(0)})\, w = g''(w^{(0)})\, w^{(0)} - g'(w^{(0)})$$

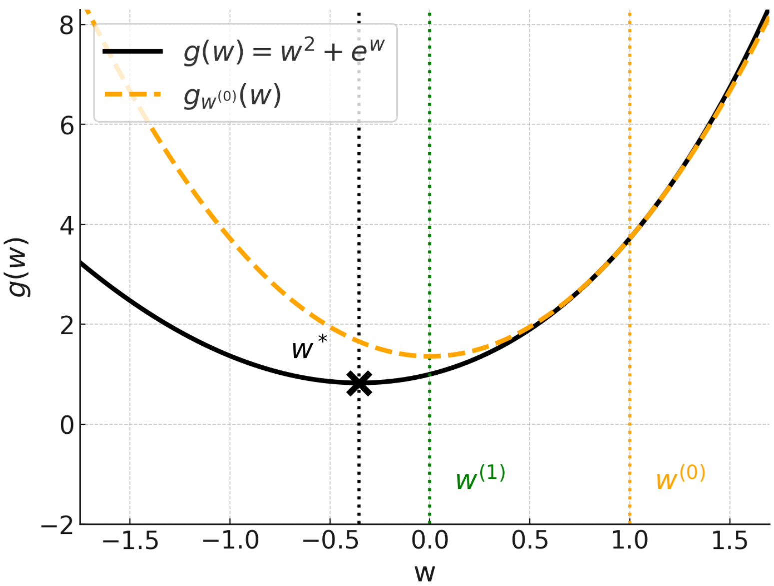
$$\Rightarrow \boxed{w = w^{(0)} - \frac{g'(w^{(0)})}{g''(w^{(0)})}}$$

Ex: $w^{(0)} = 1$

$g'(1) = 2 \cdot 1 + e^1$

$g''(1) = 2 + e^1$

$w^{(1)} = 1 - \frac{2+e}{2+e} = 0$



$\Rightarrow$ We can approximat $g(w)$ again but at $w^{(1)}$

$$g_{w^{(1)}}(w) = g(w^{(1)}) + g'(w^{(1)})(w - w^{(1)}) + \frac{g''(w^{(1)})}{2}(w - w^{(1)})^2$$

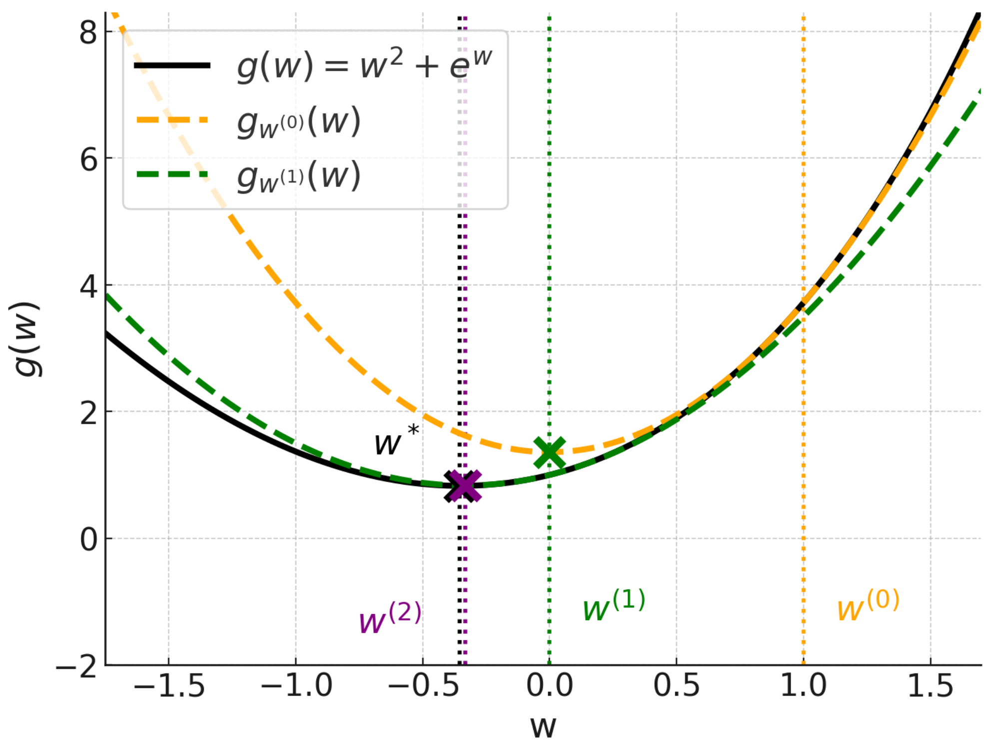
Ex: $w^{(1)} = 0$

$g_{w^{(1)}}(w) = g_0(w)$

And we can minimize $g_{w^{(1)}}(w)$ to get

$$w^{(2)} = w^{(1)} - \frac{g'(w^{(1)})}{g''(w^{(1)})}$$

Ex: $w^{(2)} = -1/3$



In general

$$w^{(t+1)} = w^{(t)} - \frac{g'(w^{(t)})}{g''(w^{(t)})}$$

where $w^{(t+1)}$ approaches w^* as $t \to \infty$

(First-Order) Gradient Descent

Sometimes it is hard to calculate $g''(w)$, especially in high dimension.

Instead we can replace it with $\eta^{(t)}$

$$w^{(t+1)} = w^{(t)} - \eta^{(t)} g'(w^{(t)})$$

$\eta^{(t)}$ is the "step size" or "learning rate"

if we know $g''(w)$, then we can set $\eta^{(t)} = \frac{1}{g''(w^{(t)})}$ and get back the 2nd order gradient descent.

-> in 1st order gradient descent we still approximate the underlying func with a 2nd order Taylor expansion, but remove the necessity to calculate $g''(w)$.

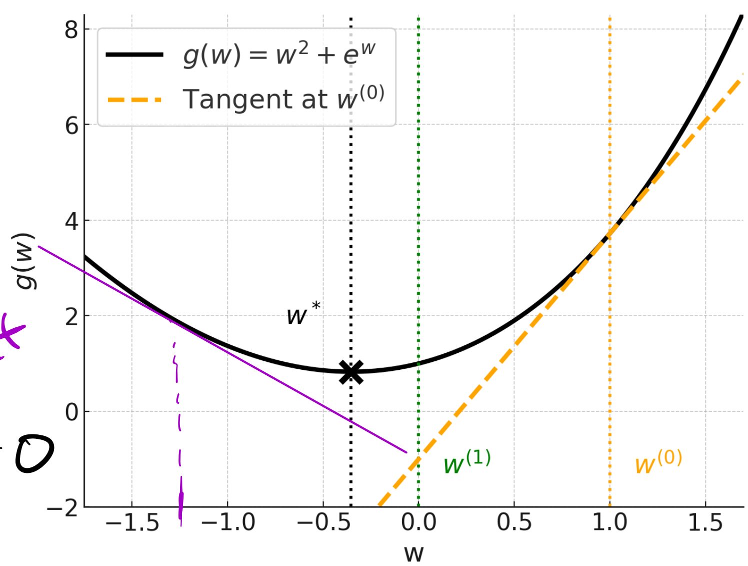
Ex: $\eta^{(t)} = \frac{1}{g''(w^{(t)})}$

$$g(w) = w^2 + e^w$$

$$g'(w) = 2w + e^w$$

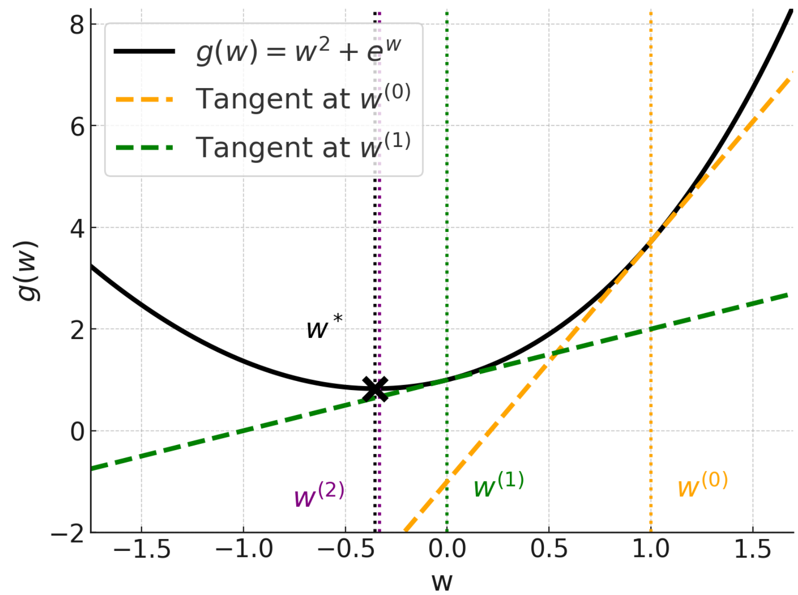
$w^{(0)} = 1$, $\eta^{(t)} = \frac{1}{2+e}$

$$w^{(1)} = w^{(0)} - \underbrace{\eta^{(0)}}_{\text{scale of the step}} \underbrace{g'(w^{(0)})}_{\text{direction of ascent}} = 1 - 1 = 0$$



$$\eta^{(1)} = \frac{1}{2+e^{0}} = \frac{1}{3}$$

$$w^{(2)} = w^{(1)} - \eta^{(1)} g'(w^{(1)})$$



↑ known

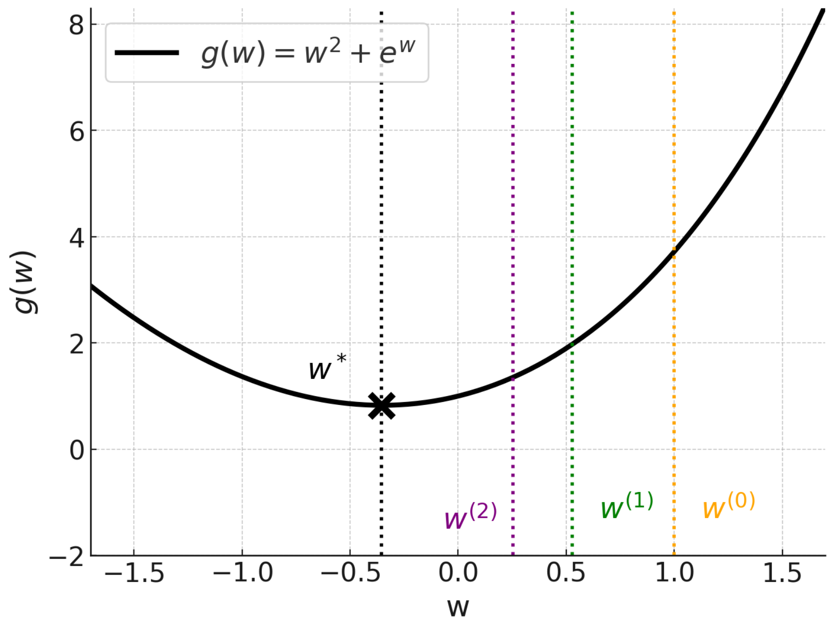
$g''(w)$

↓ unknown

Ex: $\eta^{(t)} = \frac{1}{10} < \frac{1}{2+e}$

small $\eta^{(t)}$

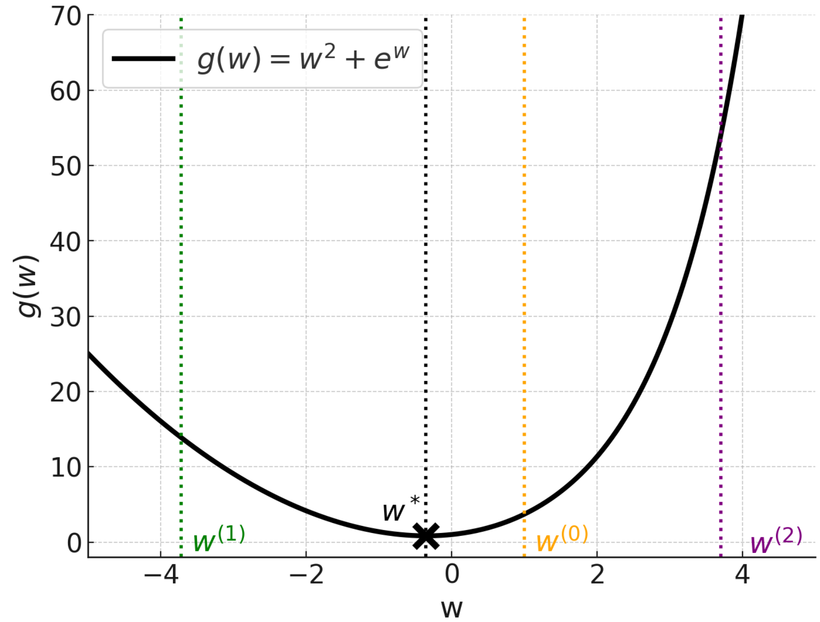
slowly reaches w^*



Ex: $\eta^{(t)} = 1 > \frac{1}{2+e}$

large $\eta^{(t)}$

might never reach w^* (diverge)



-> small step size is safe (probably converge to w^*) however it might take long

-> large step size get you fast to w^* but might diverge

Multivariate Gradient Descent

$\vec{w} \in W = \mathbb{R}^d$, $d > 1$, $g(\vec{w})$

Objective : $\vec{w}^* = (w_1^*, \dots, w_d^*) = \underset{\vec{w} \in W}{\arg\min}\ g(\vec{w})$

multivariat gradient descent update rule

$$\vec{w}^{(t+1)} = \vec{w}^{(t)} - \eta^{(t)} \nabla g(\vec{w}^{(t)})$$

where $\nabla g(\vec{w}) = \left(\frac{\partial}{\partial w_1} g(\vec{w}), \dots, \frac{\partial}{\partial w_d} g(\vec{w})\right)^T \in \mathbb{R}^d$

$\eta^{(t)} \in \mathbb{R}$

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

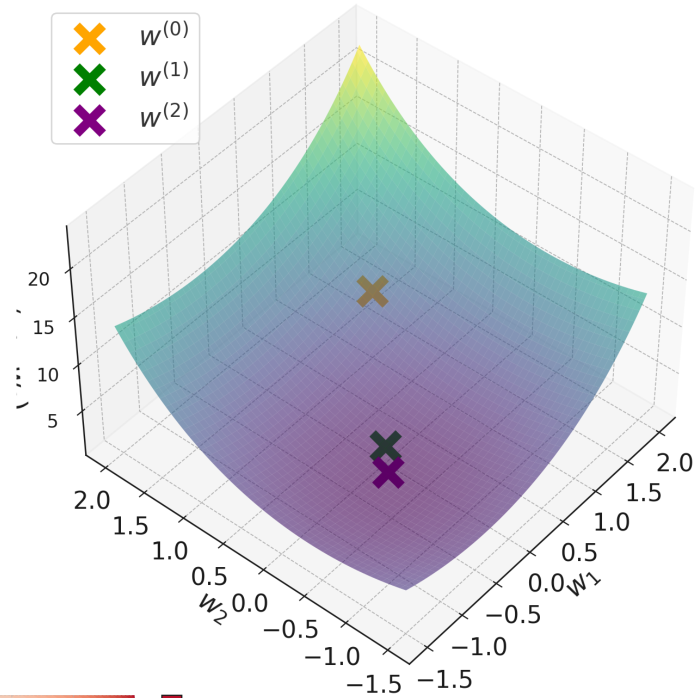
$\frac{\partial}{\partial w_1} g(\vec{w}) = 2w_1 + e^{w_1}$ $\qquad$ $\frac{\partial}{\partial w_2} g(\vec{w}) = 2w_2 + e^{w_2}$

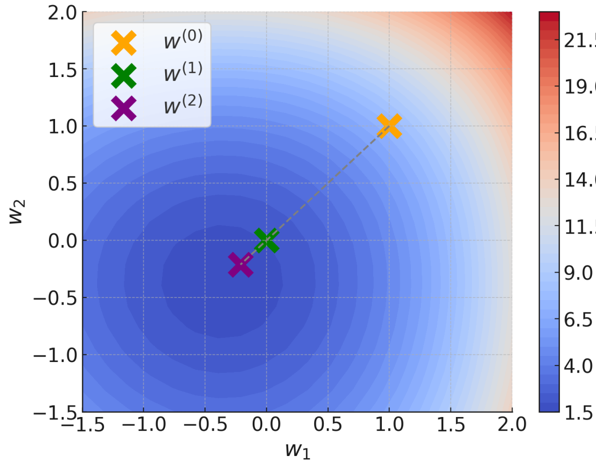
$$\vec{w}^{(t+1)} = \begin{pmatrix} w_1^{(t+1)} \\ w_2^{(t+1)} \end{pmatrix} = \begin{pmatrix} w_1^{(t)} \\ w_2^{(t)} \end{pmatrix} - \eta^{(t)} \begin{pmatrix} 2w_1^{(t)} + e^{w_1^{(t)}} \\ 2w_2^{(t)} + e^{w_2^{(t)}} \end{pmatrix}$$

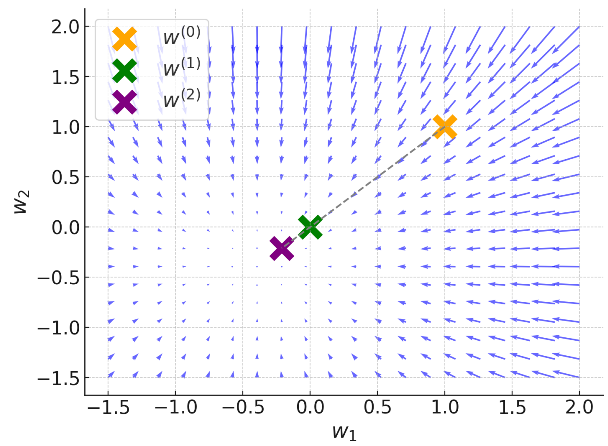
Let $\vec{w}^{(0)} = (1, 1)^T$, $\eta^{(0)} = \frac{1}{2+e}$

$$\vec{w}^{(1)} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} - \frac{1}{2+e}\begin{pmatrix} 2+e \\ 2+e \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} - \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$\eta^{(1)} = \frac{1}{2+e} \rightarrow \vec{w}^{(2)} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} - \frac{1}{2+e}\begin{pmatrix} 1 \\ 1 \end{pmatrix}$







Ex: $g(w_1, w_2) = (1 - w_1)^2 + 100(w_2 - w_1^2)^2$