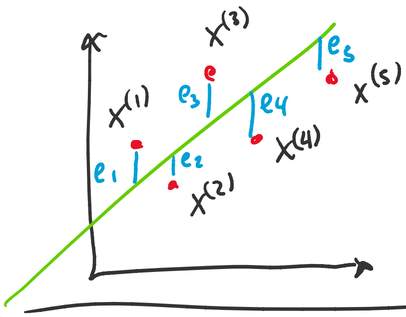
Section B

Monday, February 6, 2023 9:13 AM

$$D = \{(\vec{x}^{(i)}, y^{(i)})\}_{i=1}^{N} \qquad \hat{y}^{(i)} = h(\vec{x}^{(i)})$$

Residuals

Def: a residual is the vertical distance from observed value $y^{(i)}$ to predicted output $\hat{y}^{(i)}$



$$e_i = \left| y^{(i)} - h(\vec{x}^{(i)}) \right|$$
$$= \left| y^{(i)} - (\vec{w}^T \vec{x}^{(i)} + b) \right|$$

Key Idea of Lin. Reg.

Find the linear function h (w/parameters w and b) that minimizes $\boxed{\text{the squares of}}$ the residuals for a training set.

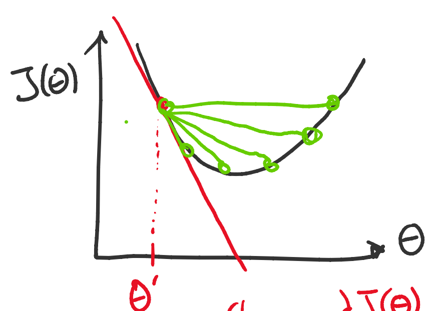
one option

Def: mean squared error (MSE)

$$J_D(\vec{w}, b) = \frac{1}{N} \sum_{i=1}^{N} (e_i)^2 = \frac{1}{N} \sum_{i=1}^{N} \left(y^{(i)} - (\vec{w}^T \vec{x}^{(i)} + b) \right)^2$$

Derivatives

① Deriv. as Slope of Tangent

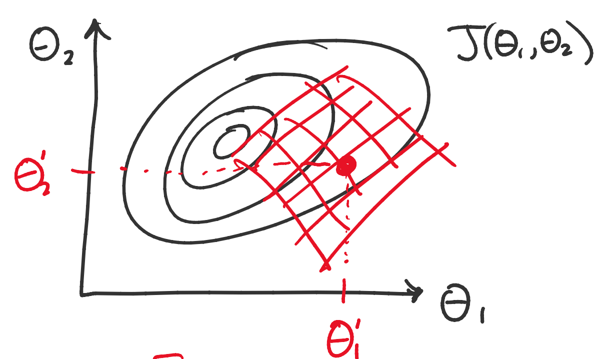


$$\text{Slope} = \frac{\partial J(\theta)}{\partial \theta}$$

② Deriv. as a Limit

$$\frac{\partial J(\theta)}{\partial \theta} = \boxed{\lim_{e \to 0}} \frac{J(\theta + e) - J(\theta)}{e}$$

③ Partial Deriv. as Tangent Planes



$$\begin{bmatrix} \partial J(\theta_1, \theta_2) / \partial \theta_1 \\ \partial J(\theta_1, \theta_2) / \partial \theta_2 \end{bmatrix}$$

$$\frac{\partial J(\theta)}{\partial \theta} = \underbrace{\boxed{\lim_{e \to 0}}}_{\text{limit of secants is the tangent}} \frac{J(\theta + e) - J(\theta)}{e}$$

$$\left[\partial J(\theta_1, \theta_2) / \partial \theta_2 \right]$$

Gradient

Def: the gradient of $J : \mathbb{R}^M \to \mathbb{R}$

$$\nabla J(\vec{\theta}) = \begin{bmatrix} \partial J(\vec{\theta}) / \partial \theta_1 \\ \partial J(\theta) / \partial \theta_2 \\ \vdots \\ \partial J(\vec{\theta}) / \partial \theta_M \end{bmatrix}$$

first order partial derivative

Gradient Descent

Algorithm:

1. Choose an initial point $\vec{\theta}$
2. Repeat $t = 1, 2, 3, \ldots$
   a) Compute gradient $\vec{g} = \nabla J(\vec{\theta})$
   b) Choose a step size γ_t
   c) Update $\vec{\theta} \leftarrow \vec{\theta} - \gamma \vec{g}$
3. Return $\vec{\theta}$ when stopping criterion is reached

Remarks:

Initial Point:
- randomly
- $\vec{\theta}$ = all zeros

Step Size:
- fixed value $\gamma = 0.1$
- schedule $\gamma_t = \frac{\gamma_0}{(t-1)\gamma_0 + 1}$
- line search

Stopping Criterion:
- when $\vec{g} \approx \vec{0}$
- $\| \nabla J(\vec{\theta}) \|_2 < \epsilon$ for $\epsilon = 10^{-8}$

Gradient for Lin. Reg.

MSE: $J(\vec{\theta}) = \frac{1}{N} \sum_{i=1}^{N} J^{(i)}(\vec{\theta})$ where $J^{(i)}(\vec{\theta}) = \overbrace{\frac{1}{2} \Big(\underbrace{y^{(i)} - \vec{\theta}^T \vec{x}^{(i)}}_{f(\theta)} \Big)^2}^{(f(\theta))^2}$

Gradient for Lin. Reg.

MSE: $J(\vec{\theta}) = \frac{1}{N} \sum_{i=1}^{N} J^{(i)}(\vec{\theta})$ where $J^{(i)}(\vec{\theta}) = \frac{1}{2}\left(y^{(i)} - \vec{\theta}^T \vec{x}^{(i)}\right)^2$

does not change the argmin

Partial Derivatives:

$$\frac{\partial J(\vec{\theta})}{\partial \theta_j} = \cancel{\frac{1}{2}} \cancel{2}\left(y^{(i)} - \vec{\theta}^T \vec{x}^{(i)}\right) \frac{\partial}{\partial \theta_j}\left(y^{(i)} - \theta^T \vec{x}^{(i)}\right)$$

$$= \left(y^{(i)} - \vec{\theta}^T \vec{x}^{(i)}\right) \frac{\partial}{\partial \theta_j}\left(y^{(i)} - \sum_{m=1}^{M} \theta_m x_m^{(i)}\right)$$

$$= -\underbrace{\left(y^{(i)} - \vec{\theta}^T \vec{x}^{(i)}\right)}\, x_j^{(i)}$$

Gradient:

$$\nabla J^{(i)}(\vec{\theta}) = \begin{bmatrix} \partial J^{(i)}(\vec{\theta})/\partial \theta_1 \\ \partial J^{(i)}(\theta)/\partial \theta_2 \\ \vdots \\ \partial J^{(i)}(\vec{\theta})/\partial \theta_M \end{bmatrix} = -\underbrace{\left(y^{(i)} - \vec{\theta}^T \vec{x}^{(i)}\right)}_{\text{scalar}} \underbrace{\vec{x}^{(i)}}_{\text{vector}}$$

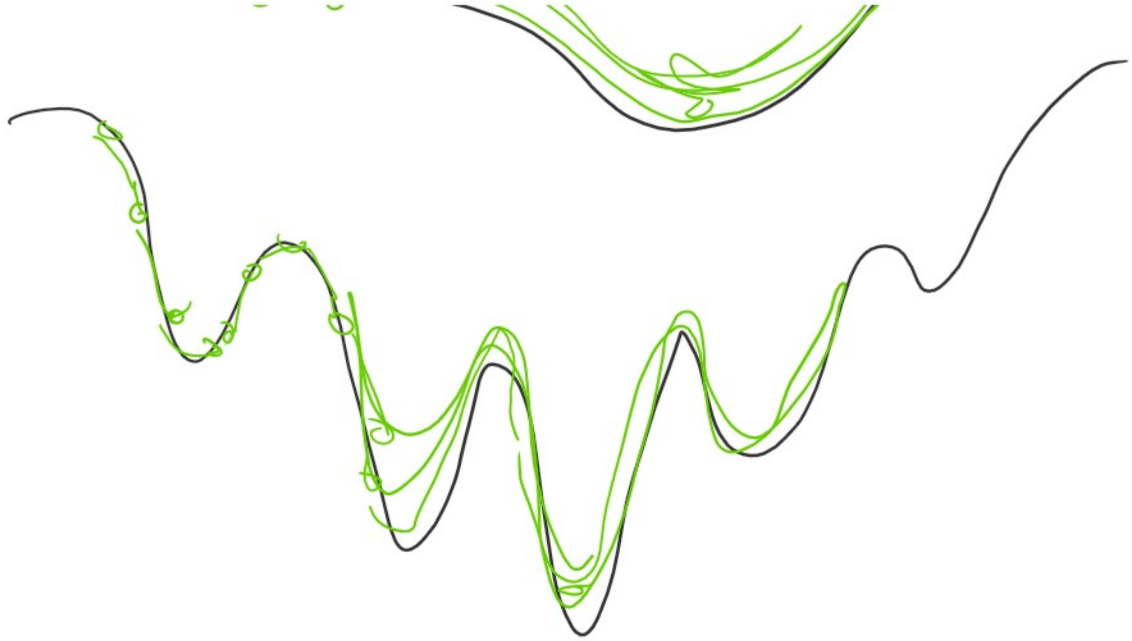
$$\nabla J(\vec{\theta}) = \nabla\left(\frac{1}{N} \sum_{i=1}^{N} J^{(i)}(\vec{\theta})\right) = \frac{1}{N} \sum_{i=1}^{N} \nabla J^{(i)}(\vec{\theta})$$

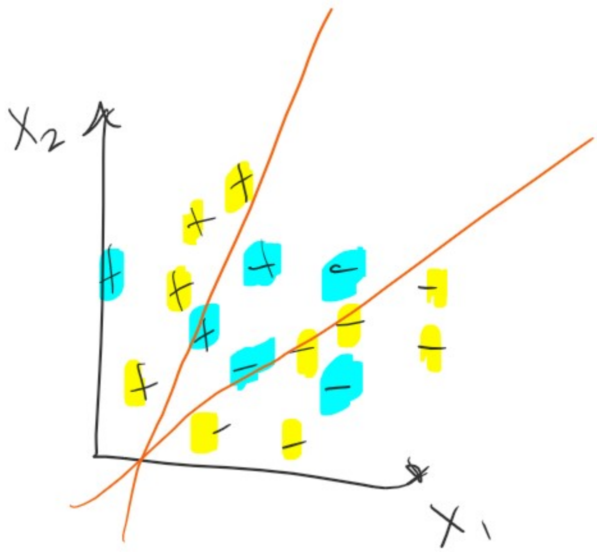
$$= \frac{1}{N} \sum_{i=1}^{N} -\left(y^{(i)} - \vec{\theta}^T \vec{x}^{(i)}\right) \vec{x}^{(i)}$$

known b/c in training

known b/c passed it in

OH




X_2
X_1