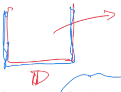
✓ $J(\theta) \rightarrow$ cost / training loss / error $\rightarrow$ $\min^m$

$\theta \rightarrow$ SGD / Newton' / MLE

$\downarrow$

how good is this θ

Test loss



draw new example, $(x, y) \sim D \rightarrow$ test distribution

evaluate :

test loss $\rightarrow L(\theta) = \mathbb{E}_{(x,y)\sim D}[(y - h_\theta(x))^2]$

$(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}) \ldots\ldots$ } avg

training phase

θ / h

testing phase

(x_1^{test}, y_1^{test})

(x_2^t, y_2^t)

$\vdots$

n

⊛ $J(\theta)$, $L(\theta) - J(\theta) \Rightarrow$ Generalization gap

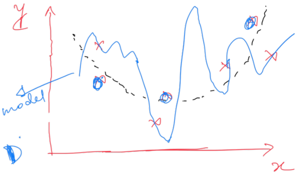
$\downarrow$ low $\quad \downarrow$ low

⊛ $L(\theta)$ is too big ✓

- overfitting : $J(\theta)$ is small, $L(\theta)$ big $\rightarrow$ variance
- underfitting : $J(\theta)$ is big $\rightarrow$ bias

Next class / $L(\theta) = (\text{bias})^2 + \text{variance} + \xi$

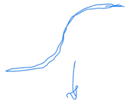
$\xi \downarrow$ irreducible noise



$J(\theta) \approx 0$

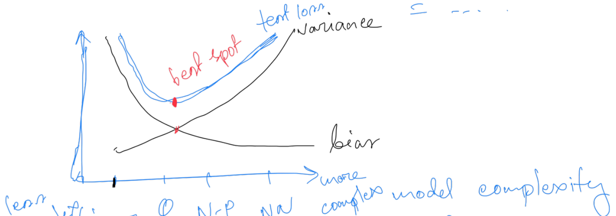
$\rightarrow$ overfitting

- linear
- quad.
- 5th deg. poly
- NN

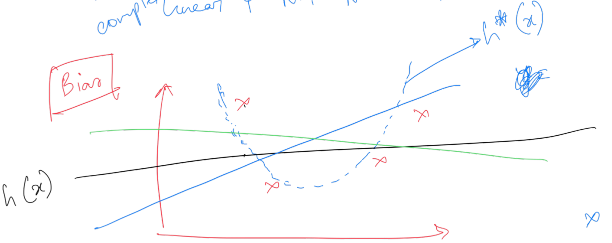


$~~L(\theta) = (h^*(\theta)' - y)^2 = a^2 + 2ab + b^2~~$

$= \ldots\ldots$



Bias



$y = h^*(x) + \xi$

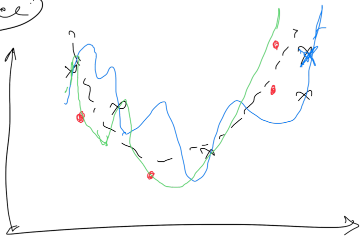
$\Rightarrow y = \text{quad}(x) + \xi$

– no linear model would work here

↓
family of func.

— more data won't help

variance



5th degree poly. → h(x)

$= \theta_5 x^5 + \theta_4 x^4 + \dots \theta_0$

linear in terms of θ? ✓

Soln

- add more data
- simpler model

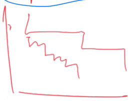
— o —

Q1 * can you express complexity w/ a metric?

- # parameters / weights
  - — NN.
- Norm. of parameters
  - ↓ L1 / L2 norm
  - normalization
- smoothness / Lipschitzness
- how invariant to data augmentation

Q2

* can you control complex. during train?



Q2 Regularization

↳ encouraging low variance / complexity

⇒ Optimize

$$J(\theta) + \underbrace{\lambda}_{\text{Regularization strength}} \cdot \underbrace{R(\theta)}_{\text{Regularizer}}$$

Regularization strength
↓
const [0 — ...]

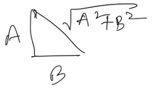
① $R(\theta) = \frac{1}{2} \|\theta\|_2^2$

L2 norm $\longrightarrow \|\theta\|_2$

$\vec{\theta}$ vector w/ n components

$\theta_1, \theta_2, \dots \theta_n$

$\|\theta\|_2 = \sqrt{\theta_1^2 + \theta_2^2 + \dots + \theta_n^2}$



② L0 Norm

$R(\theta) = \|\theta\|_0$ (~ # non-zero in your param)

$\vec{\theta} = (2\ 3.5\ 0\ -4\ 0)$

$\|\theta\|_0 = 3$

- sparsity of your param/weight

= feature selection

✓ $h(x) = \bar{\theta} \cdot \bar{x}$

$= \underset{\overset{\downarrow}{0}}{\theta_1 x_1} + \theta_2 x_2 + \dots + \underset{\overset{\downarrow}{0}}{\theta_3 x_3} + \dots$

$= \theta_2 x_2 + \theta_4 x_4 + \dots$

* Not differentiable

$\theta = [0000 \dots \underset{\uparrow}{1}0000] \sim \theta = R(\theta)$

③ L1 Norm

$\|\theta\|_1 = |\theta_1| + |\theta_2| + \dots$

$= \sum_{i=1}^{d} |\theta_i|$

$\|\theta\|_0 \approx \|\theta\|_1 \longrightarrow \theta = [0, 1]$

—o—

⊛ when?

- regularization introduces structure/ prior belief

- θ is sparse $\longrightarrow$ L0/L1

$J(\underset{R}{\theta}) + \lambda \cdot \|\theta\|_0$

$\theta = [\ .\ .\ .\ .\ .\ \ .\ ]$

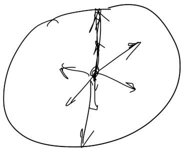
— θ to use all features
w/ $\approx$ equal values

$\bar{\theta} = [24 \quad 2 \quad 3.7 \quad 0 \quad -9 \quad 0 \quad 97]$

$h(x) = 24x_1 + 2x_2 + \dots \; 0 \cdot x_4$

$X^{(i)} \longrightarrow 24 \times 0 \dots$

$\Rightarrow \|\theta\|_0$

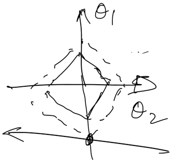


Search Space

— o —

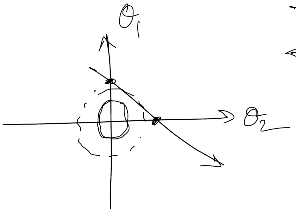
⊛ Linear regression w/

L1 norm → $J(\theta) + \lambda \|\theta\|_1$ → LASSO



L2 norm ⟶

Ridge



⟶ Deep learning / NN ⟶ L2

⊛ $\lambda \longrightarrow \infty \longrightarrow R(\theta) \longrightarrow 0$

— o —

⊛ Hyperparam / set hyperparam

— validation

λ	0.1	0.2	0.3			0.7
val. acc.	97	81				37