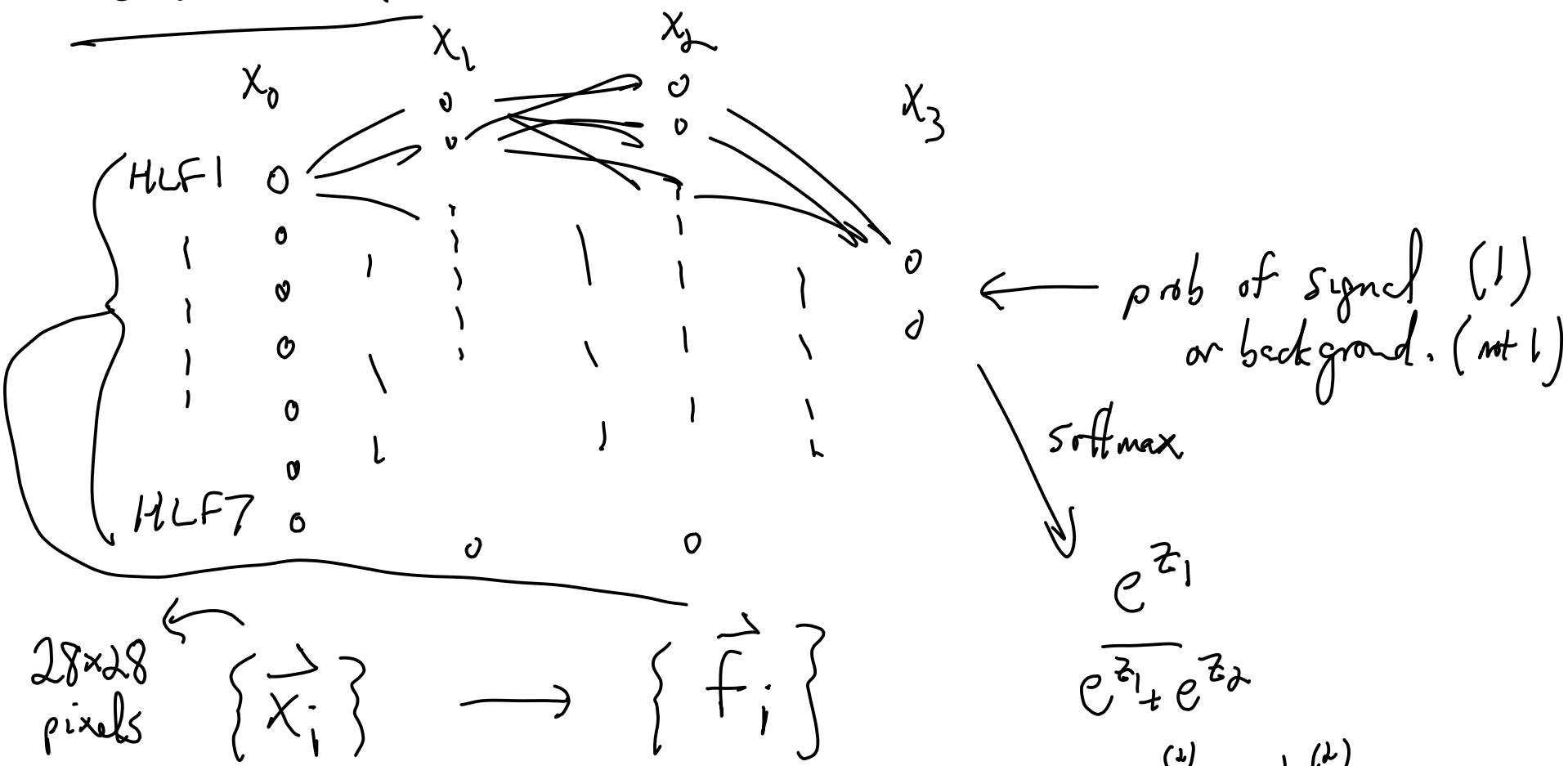
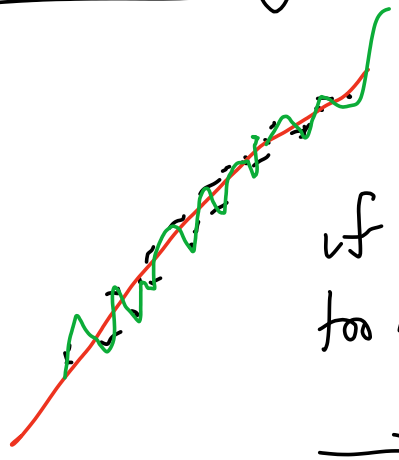
Friday, January 29, 2021 3:05 PM

Lecture 4



Friday, January 29, 2021 3:10 PM

Overfitting

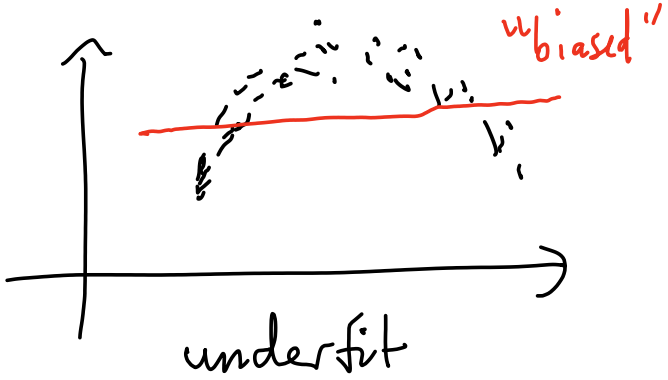


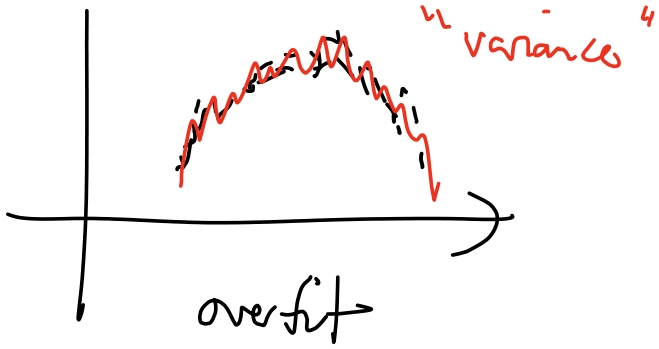
if fit fn has
too many parameters ("overparametrization")

→ fit fn can "memorize" training data

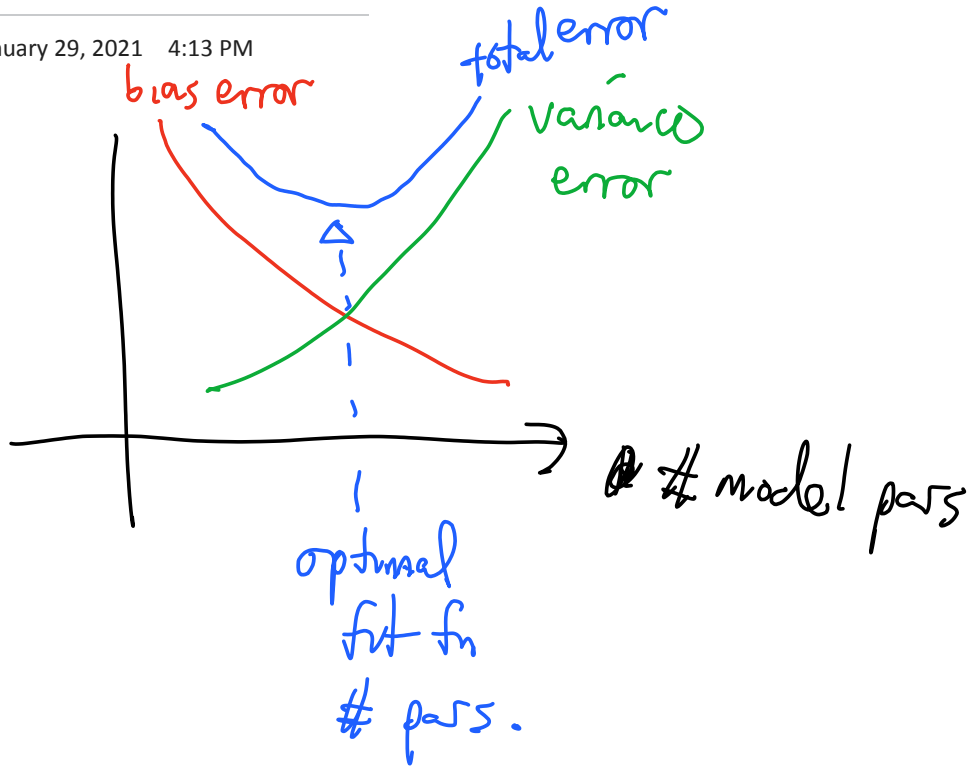
→ overfitting.

"Bias-variance tradeoff"





Friday, January 29, 2021 4:13 PM



How to diagnose & quantify & prevent overfitting?

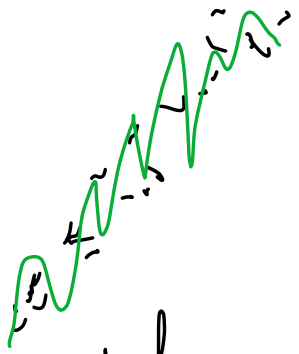
→ validation dataset. (and test dataset)

Friday, January 29, 2021 4:17 PM

val. set: statistically indep. dataset that is not used during training (except to diagnose overfitting)

train

train loss → 0

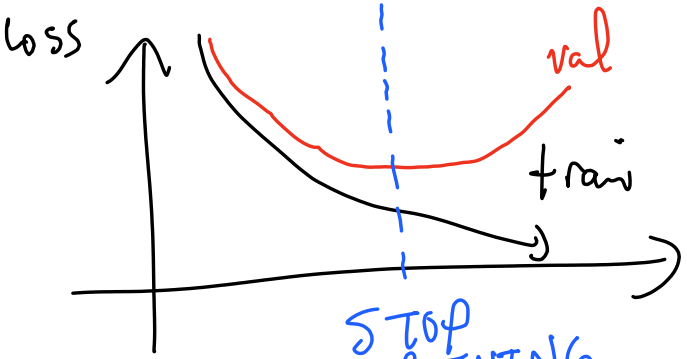


val

val loss ↛ 0

may increase after bad

"poor generalization"



"early stopping"

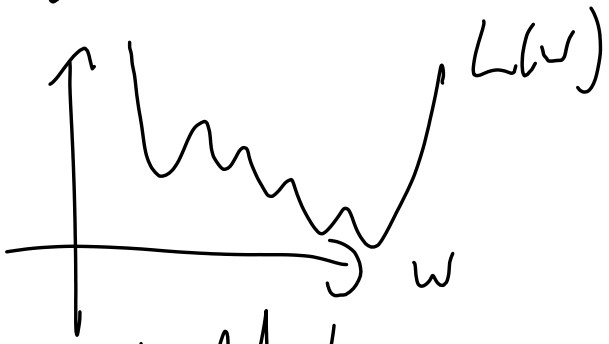
Friday, January 29, 2021 4:24 PM

NN training → (stochastic) gradient descent
→ backpropagation

- What is the goal of training?

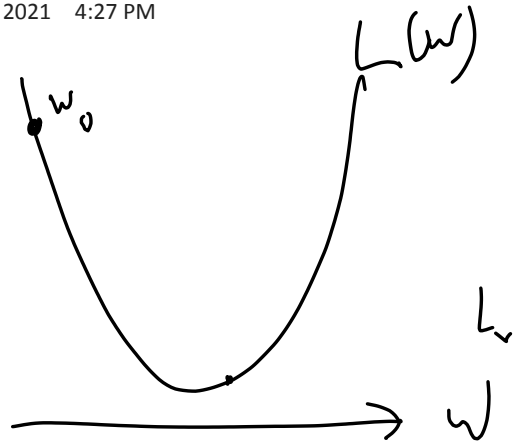
Minimize L w.r.t. weights w

Want: $\frac{\partial L}{\partial w} = 0$.



might be many local minima!
How not to get stuck?

Friday, January 29, 2021 4:27 PM



warm up: how to find min. of this fn numerically?

1. Initial guess

2. $L(w) \simeq L(w_0) + \delta w \left.\frac{\partial L}{\partial w}\right|_0 + \frac{1}{2}\delta w^2 \left.\frac{\partial^2 L}{\partial w^2}\right|_0 + \cdots$

Newton method: 2nd order

$$\delta w \approx -\left.\frac{\partial L}{\partial w}\right|_0 \Big/ \left.\frac{\partial^2 L}{\partial w^2}\right|_0 .$$

3. iterate

Problem: doesn't scale well w/ $w \to \vec{w}$ many parameters
Because $\frac{\partial^2 L}{\partial w^2}$ becomes matrix that we have to invert.

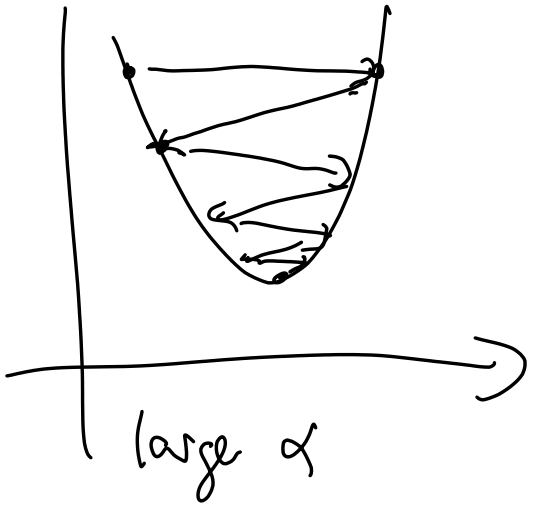
Friday, January 29, 2021 4:33 PM

So for NNs we make do w/ 1st order method

gradient descent: $\delta w = -\alpha \left.\frac{\partial L}{\partial w}\right|_0$

$\delta \vec{w} = -\alpha \left.\vec{\nabla} L\right|_0$

fixed constant parameter "learning rate"



Friday, January 29, 2021 4:38 PM

Problems w/ Vanilla GD:

1. $L = \sum_{x \in \text{Data}} L(x)$

Need to evaluate L on entire data for a single weight update. Expensive & slow!

2. 

easy to get trapped in local minima!

→ "stochastic gradient descent"