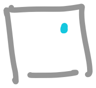
Overfitting

When your model memorizes the training data.
(Your model does not generalize.)

Take MNIST as an example.

60,000 training images → each input image (or subset) is recognized by a handful of parameters

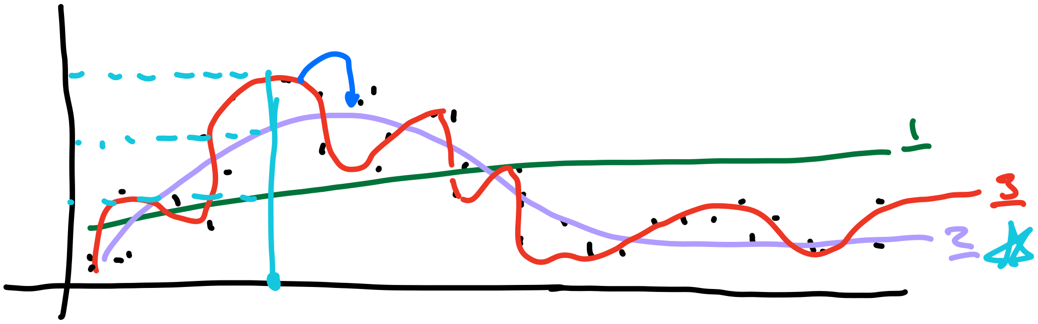


10,000 validation images → we perform poorly, because the network hasn't "memorized" them

Polynomial Regression

degrees

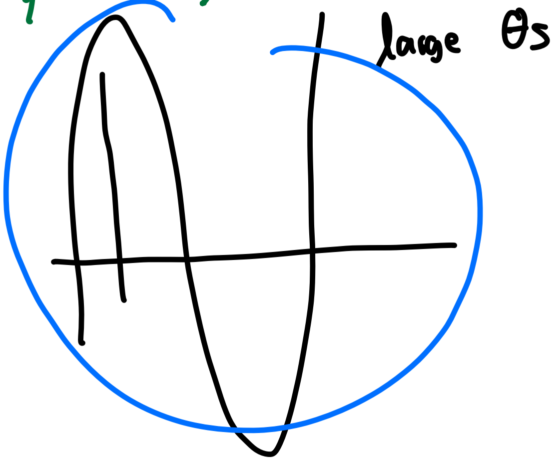
$$y = x\theta_1 + x^2\theta_2 + x^3\theta_3 \ldots$$



What is a symptom of the overfit model?

↳ High values for parameters.

$$y = \underset{\uparrow}{\theta_0} + x\underset{\uparrow}{\theta_1} + x^2\underset{\uparrow}{\theta_2} + x^3\underset{\uparrow}{\theta_3}$$



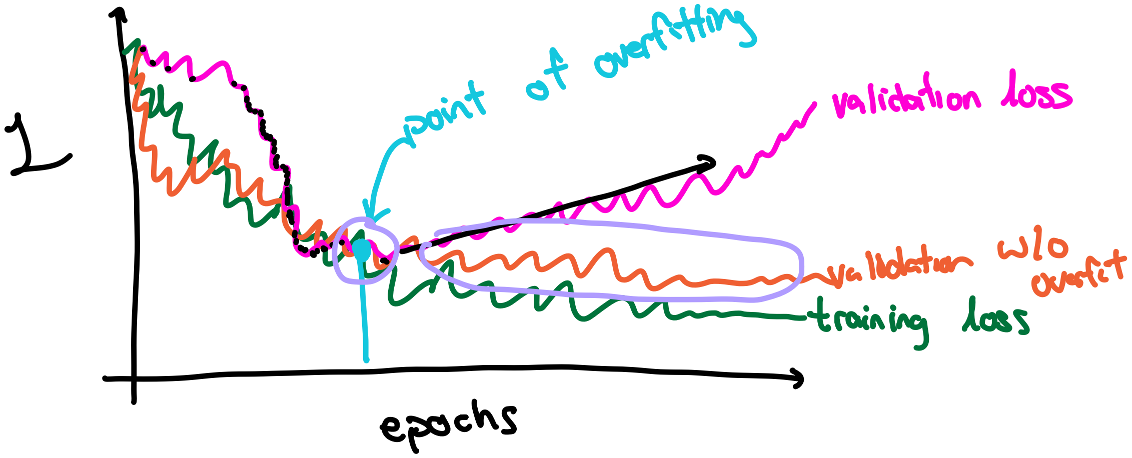
Solutions

- early stopping
- parameter norm penalty (weight decay)
- dropout
- data augmentation

cross validation

ensemble methods

Early Stopping



1. Track validation loss
2. Save model checkpoints
3. Take model just prior to validation loss not beating the best value X updates in a row

Parameter Norm Penalization

- Weight Decay
- Regularization
- L-1, L-2 Norm Penalty
- Ridge regression
- Tikhonov regularization

$$\mathcal{L}(\hat{y}, y)_{HMSE} = \frac{1}{2} \| (\hat{y} - y)^2 \|_1$$

$$\mathcal{L}(\hat{y}, y) = \mathcal{L}(\hat{y}, y)_{HMSE} + \underbrace{\frac{\lambda}{2} \theta^T \theta}$$

What does this do to loss?

- When params are large? → larger loss
- small? → increase loss by smaller amount

$$\frac{\partial \mathcal{L}}{\partial \theta} = \frac{\partial \mathcal{L}_{HMSE}}{\partial \theta} + \frac{\partial}{\partial \theta} \underbrace{\overset{\text{weight decay parameter}}{\frac{\lambda}{2}} \theta^T \theta}_{\sum_i \theta_i^2}$$

$$= \frac{\partial \mathcal{L}_{HMSE}}{\partial \theta} + \frac{\lambda}{\cancel{2}} \cancel{2} \theta$$

$$= \frac{\partial \mathcal{L}_{HMSE}}{\partial \theta} + \lambda \underset{\uparrow}{\theta}$$

How does regularization impact the parameter updates?

What if θ_i is a large # (high positive)

small # (high negative)

$$\theta_{t+1} = \theta_t - \eta \frac{\partial \mathcal{L}}{\partial \theta}$$

$$= \theta_t - \eta \left(\frac{\partial \mathcal{L}_{HMSE}}{\partial \theta} + \lambda \theta \right)$$

$$= \theta_t - \eta \frac{\partial \mathcal{L}_{HMSE}}{\partial \theta} \boxed{- \lambda \eta \theta}$$

1. $1{,}756 - \eta \nabla \mathcal{L} \boxed{- \lambda \eta \, 1{,}756}$ sub

Drive to zero

2. $-845 - \eta \nabla \mathcal{L} - \lambda \eta (-845)$

$-845 - \eta \nabla \mathcal{L} \boxed{+ \lambda \eta \, 845}$ add

Fixing the symptom, not the problem.