Overfitting

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

\[ \text{epochs} \]

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

\[ y = x_1 \theta_1 + x^2 \theta_2 + x^3 \theta_3 \ldots \]

What is a symptom of the overfit model?

→ High values for parameters.

\[ y = \theta_0 + x \theta_1 + x^2 \theta_2 + x^3 \theta_3 \]

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

\[ L(\hat{y}, y)_{\text{MSE}} = \frac{1}{2} ||(\hat{y} - y)^2||_1 \]

\[ L(\hat{y}, y) = L(\hat{y}, y)_{\text{MSE}} + \frac{\lambda}{2} \Theta^T \Theta \]

What does this do to loss?
- When params are large? \( \Rightarrow \text{larger loss} \)
- When small? \( \Rightarrow \text{increase loss by smaller amount} \)

\[ \frac{\partial L}{\partial \Theta} = \frac{\partial L_{\text{MSE}}}{\partial \Theta} + \frac{\partial}{\partial \Theta} \frac{\lambda}{2} \Theta^T \Theta \]

\[ = \frac{\partial L_{\text{MSE}}}{\partial \Theta} + \frac{\lambda}{2} \Theta \]

How does regularization impact the parameter updates?

What if \( \Theta_i \) is a large \# (high positive)?
Small \# (high negative)
\[ \theta_{t+1} = \theta_t - \eta \frac{\partial y}{\partial \theta} \]

\[ = \theta_t - \eta \left( \frac{\partial \text{HMAE}}{\partial \theta} + \lambda \theta \right) \]

1. \( \lambda \theta \) sub

2. \(- \lambda \theta \) add

Fixing the symptom, not the problem.