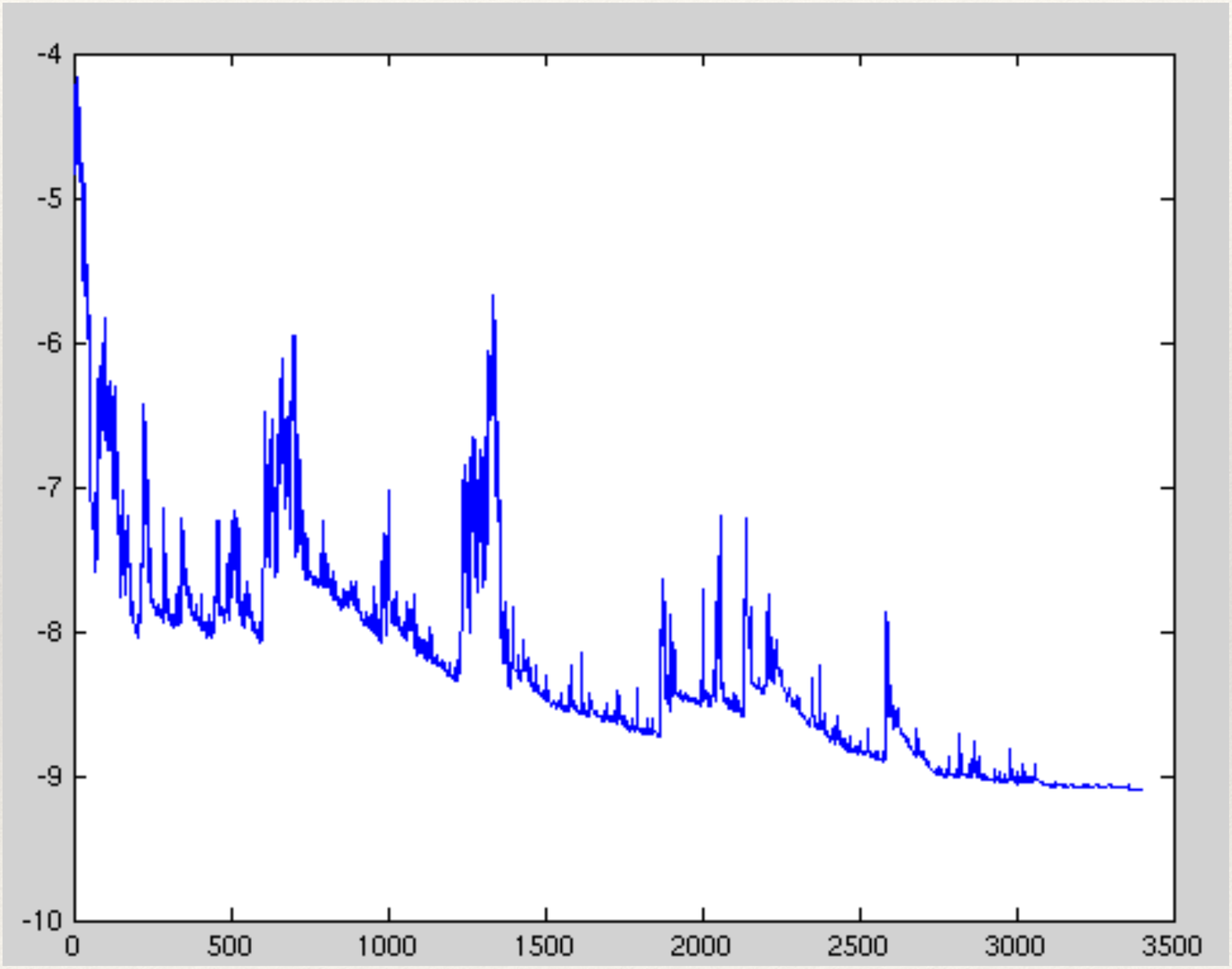
E9 205 Machine Learning for Signal Processing

Neural Networks - Generalization

23-10-2019

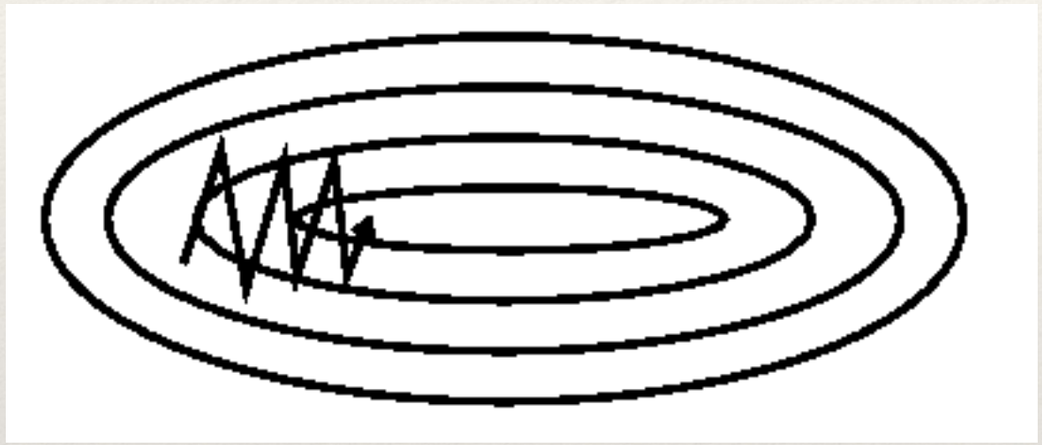
Instructor - Sriram Ganapathy (sriramg@iisc.ac.in.in)

Gradient Descent Analysis

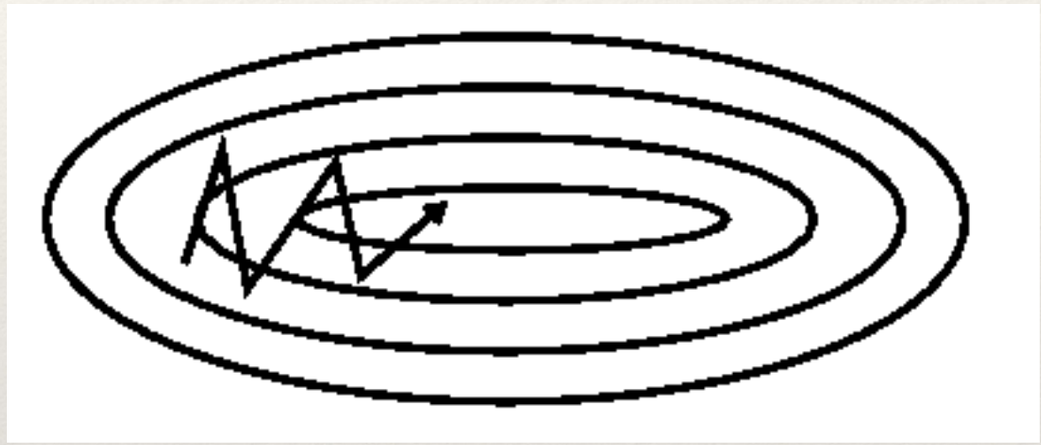


The Training Loss of SGD can fluctuate

Momentum in Learning

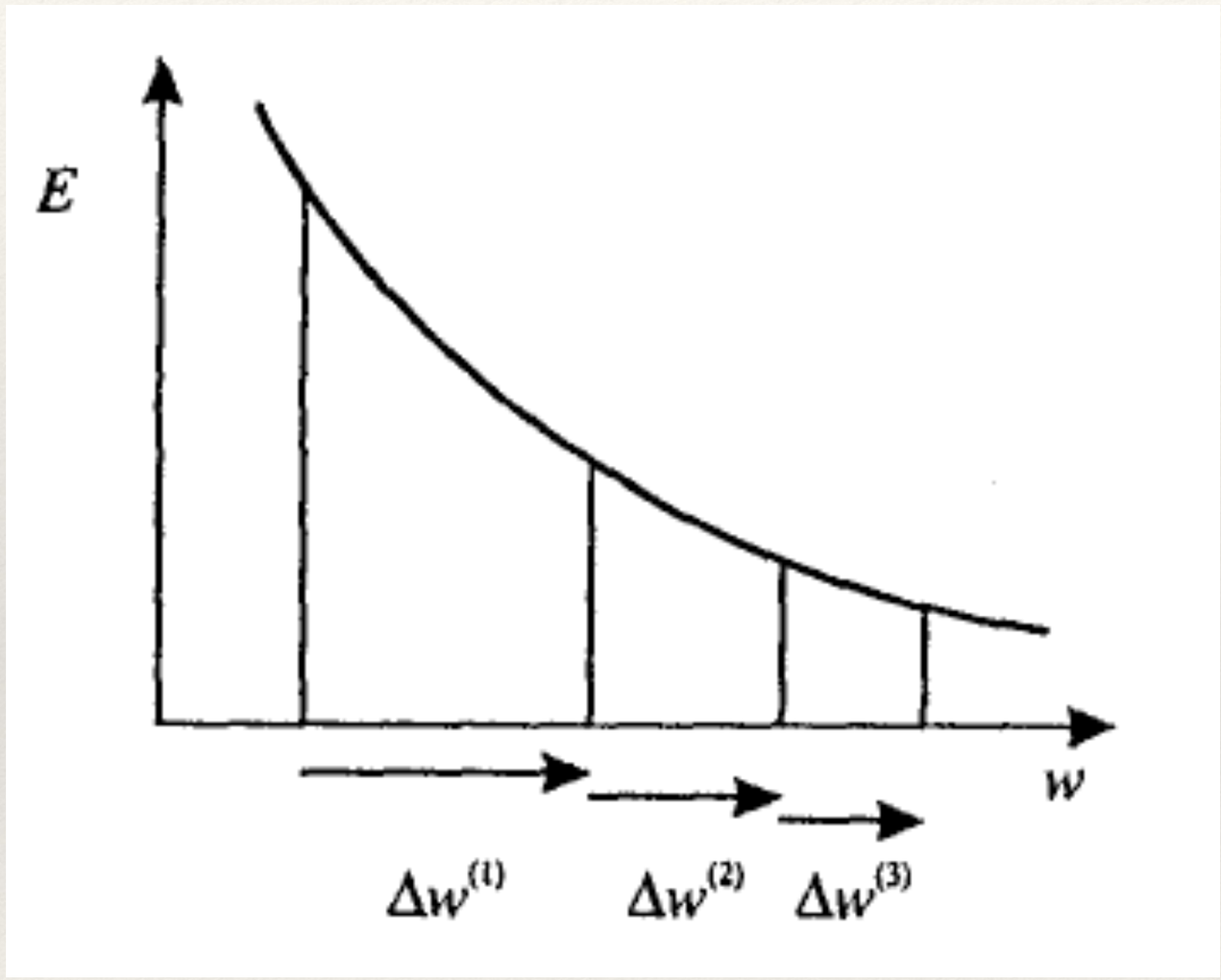


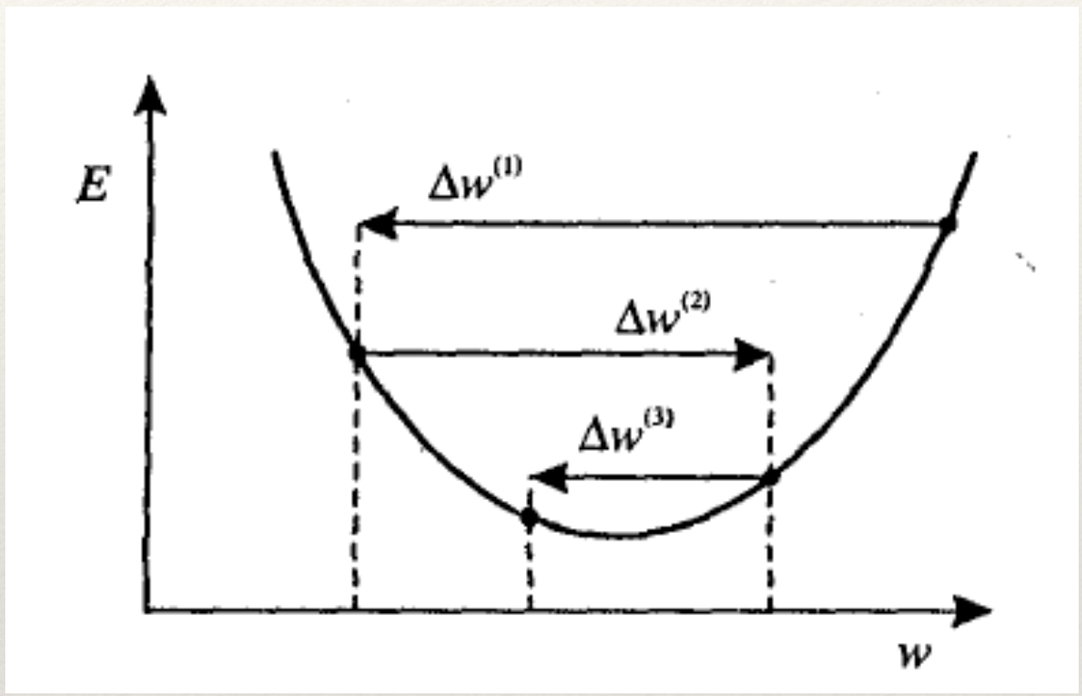
Without Momentum



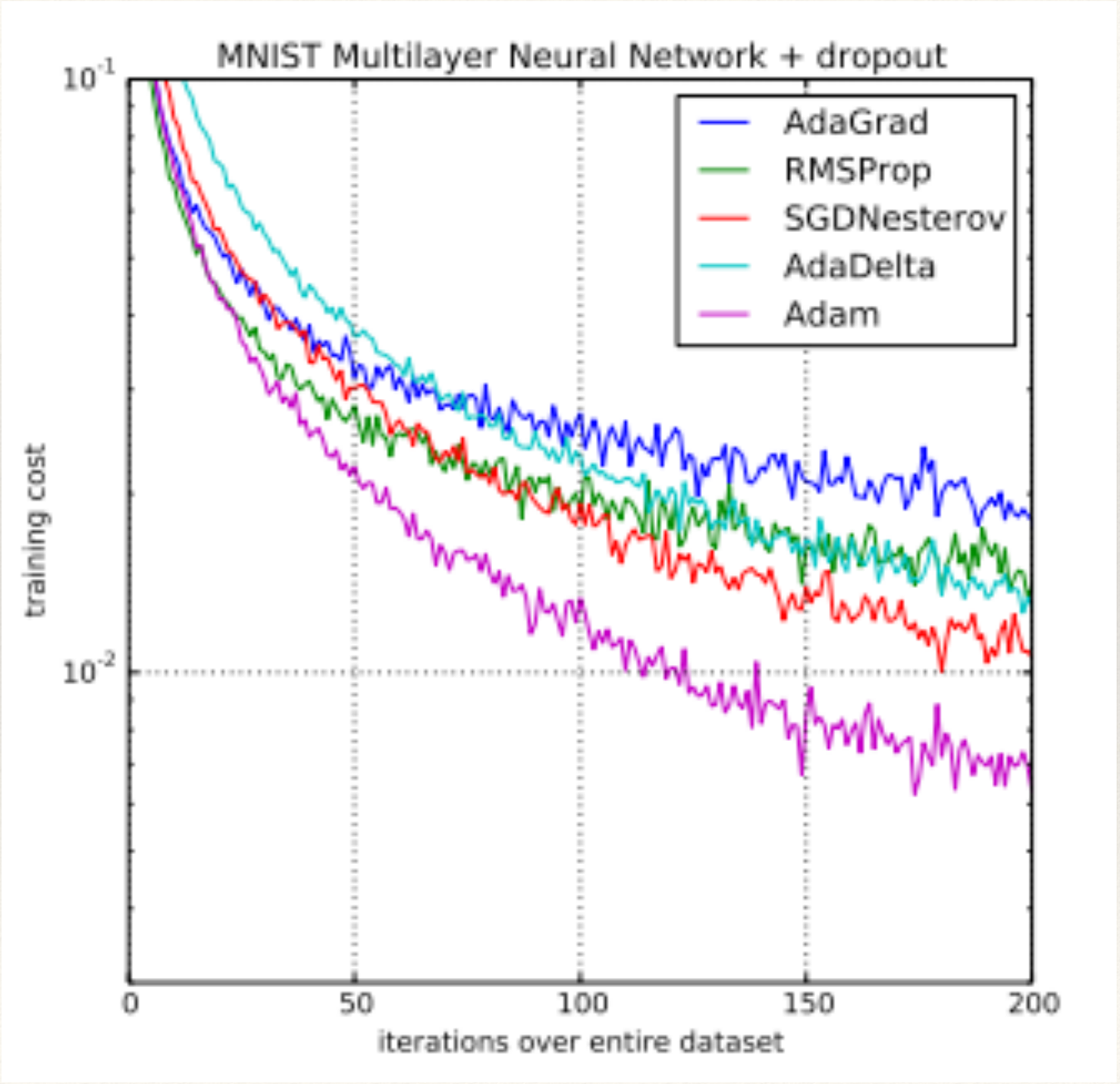
With Momentum

Momentum in Learning





An Overview of Gradient Descent Methods

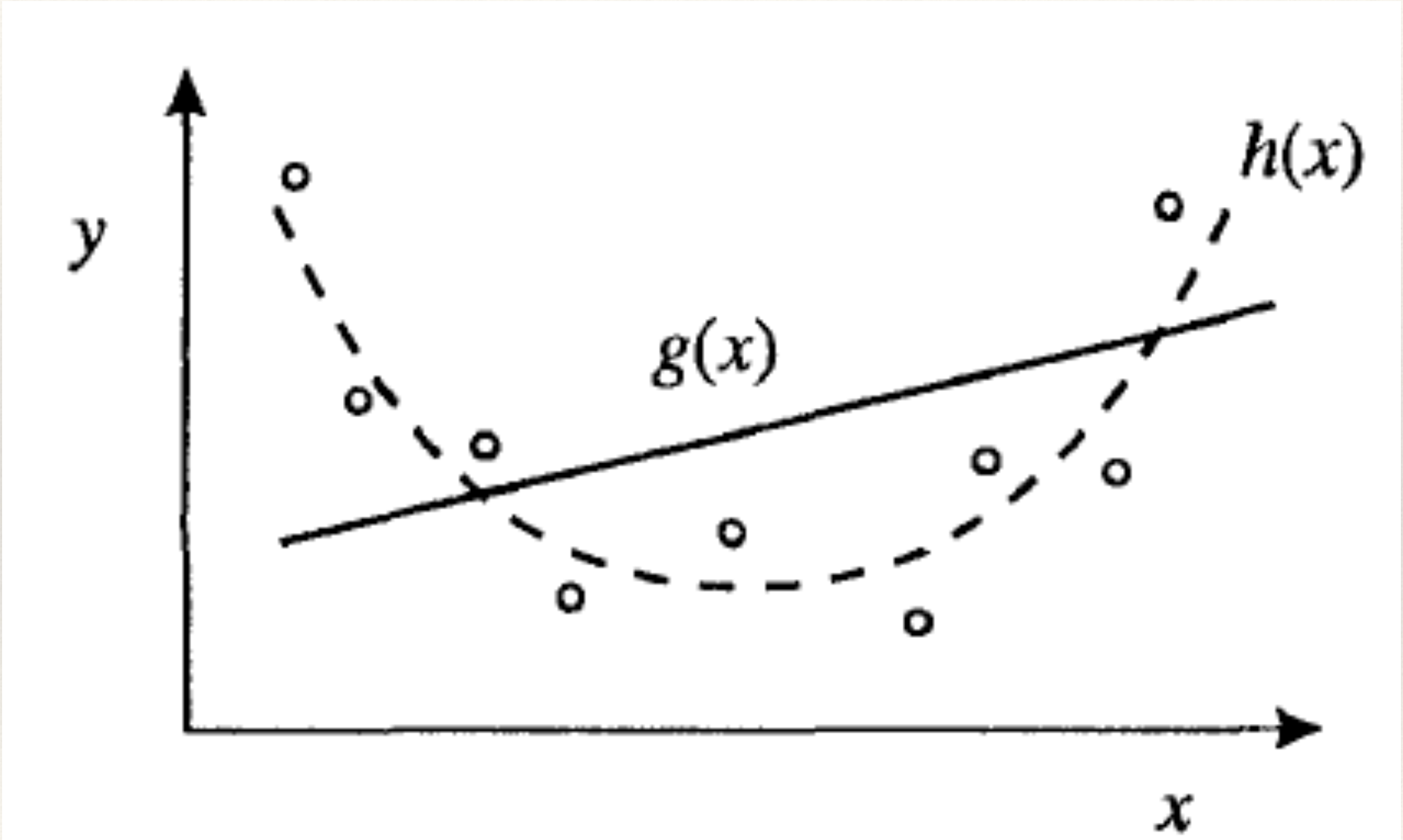


Bias and Variance In Neural Network Training

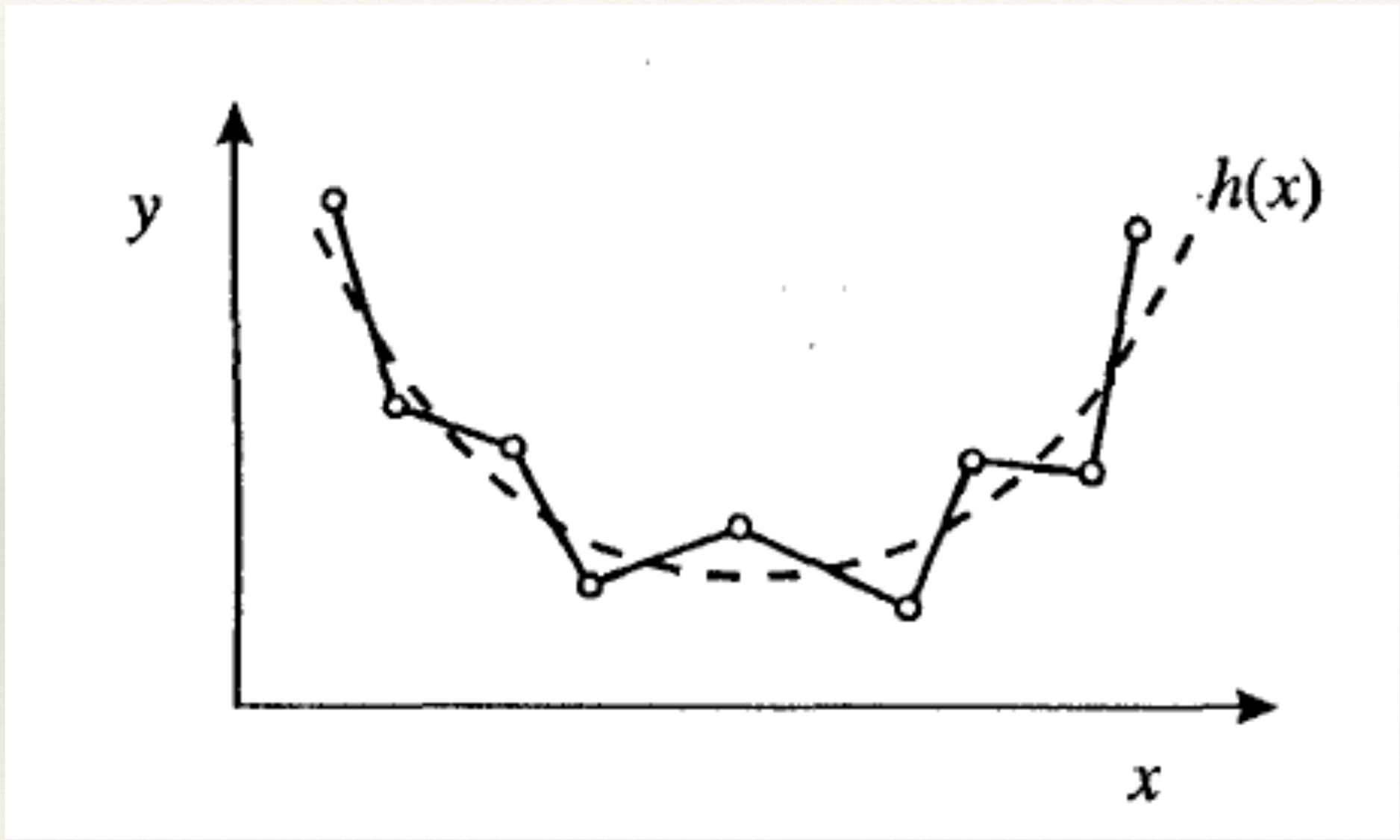
$$(\text{bias})^2 = \frac{1}{2}\int \{\mathcal{E}_D[y(\mathbf{x})] - \langle t|\mathbf{x}\rangle\}^2 p(\mathbf{x})\, d\mathbf{x}$$

$$\text{variance} = \frac{1}{2}\int \mathcal{E}_D[\{y(\mathbf{x}) - \mathcal{E}_D[y(\mathbf{x})]\}^2] p(\mathbf{x})\, d\mathbf{x}.$$

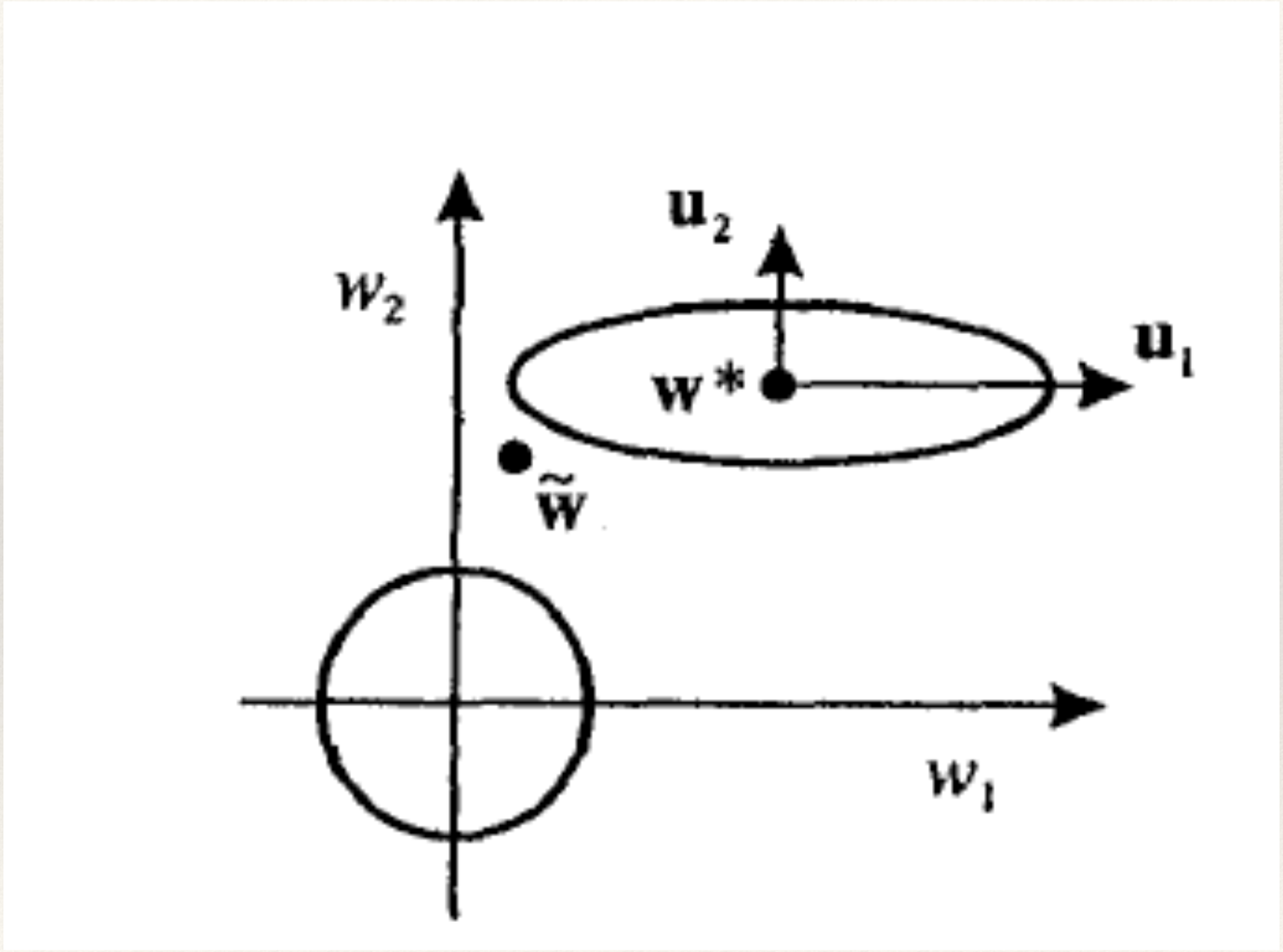
Underfit



Overfit

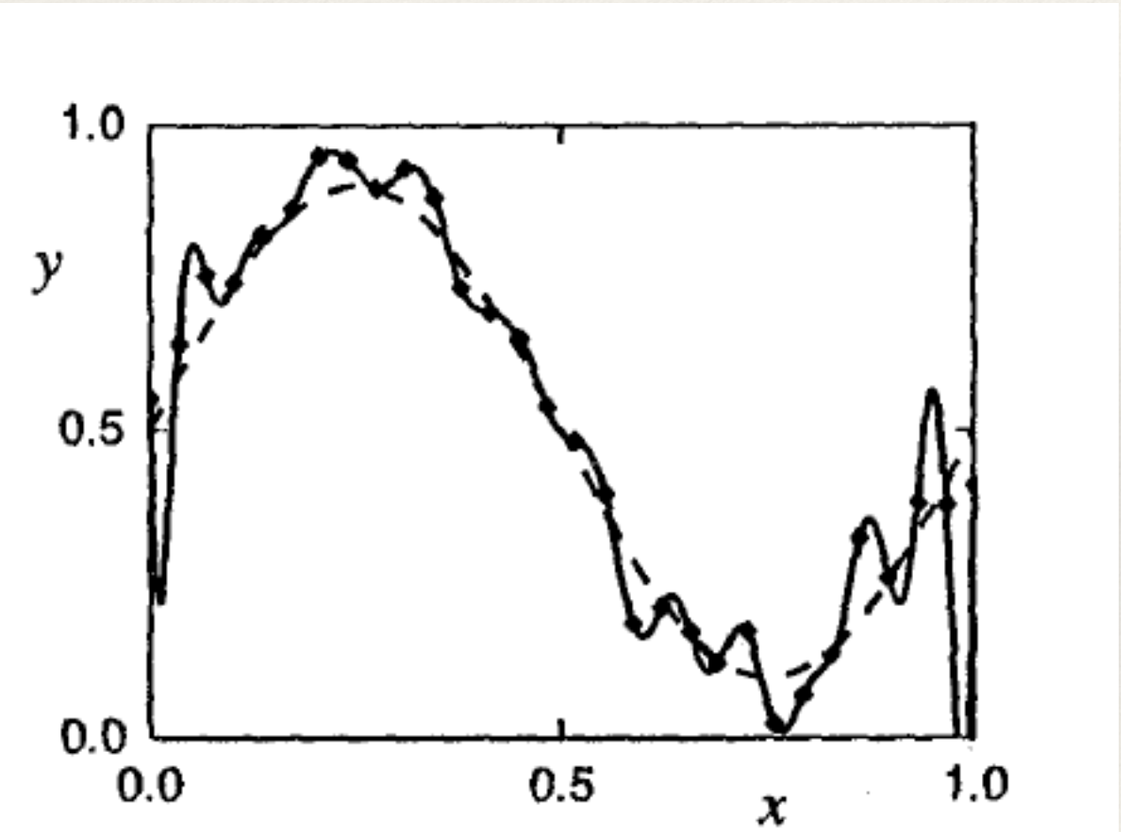


Weight Decay Based Regularization

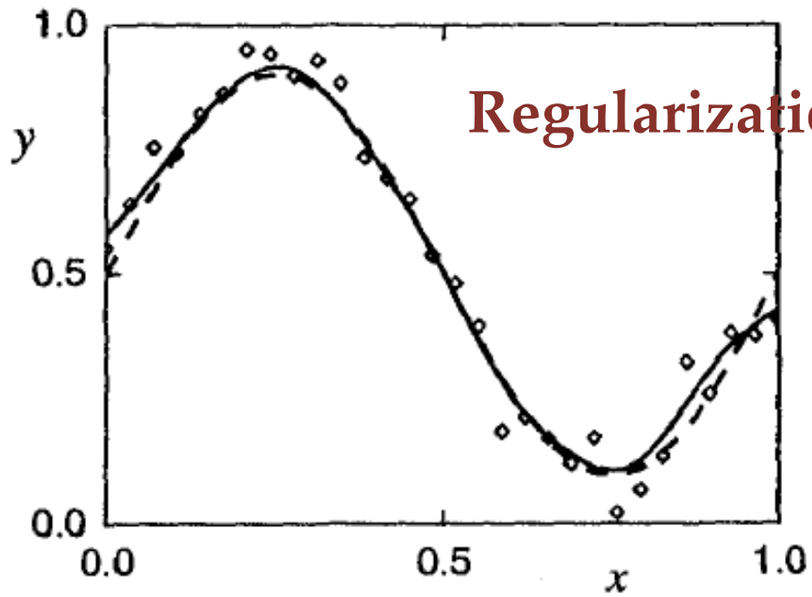


Weight Decay Regularization

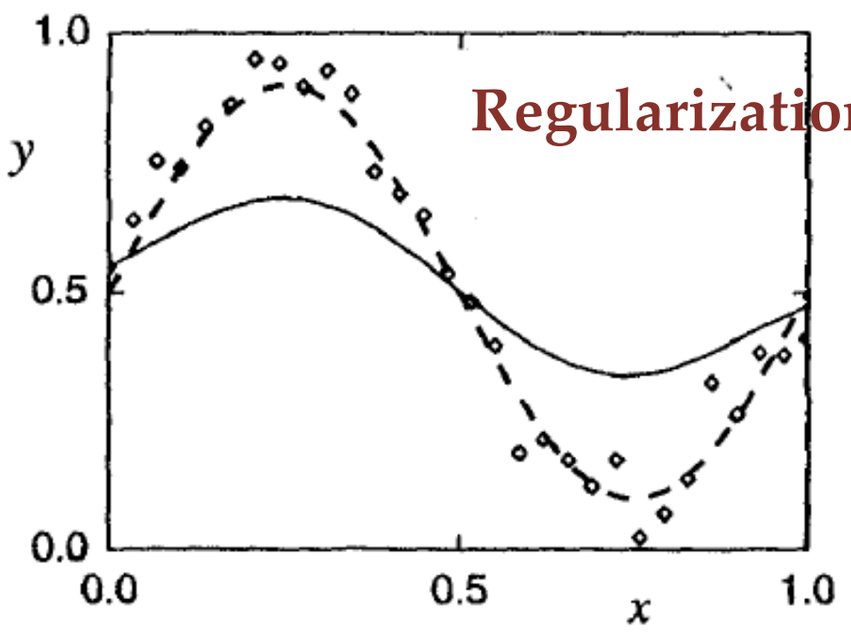
Regularization = 0



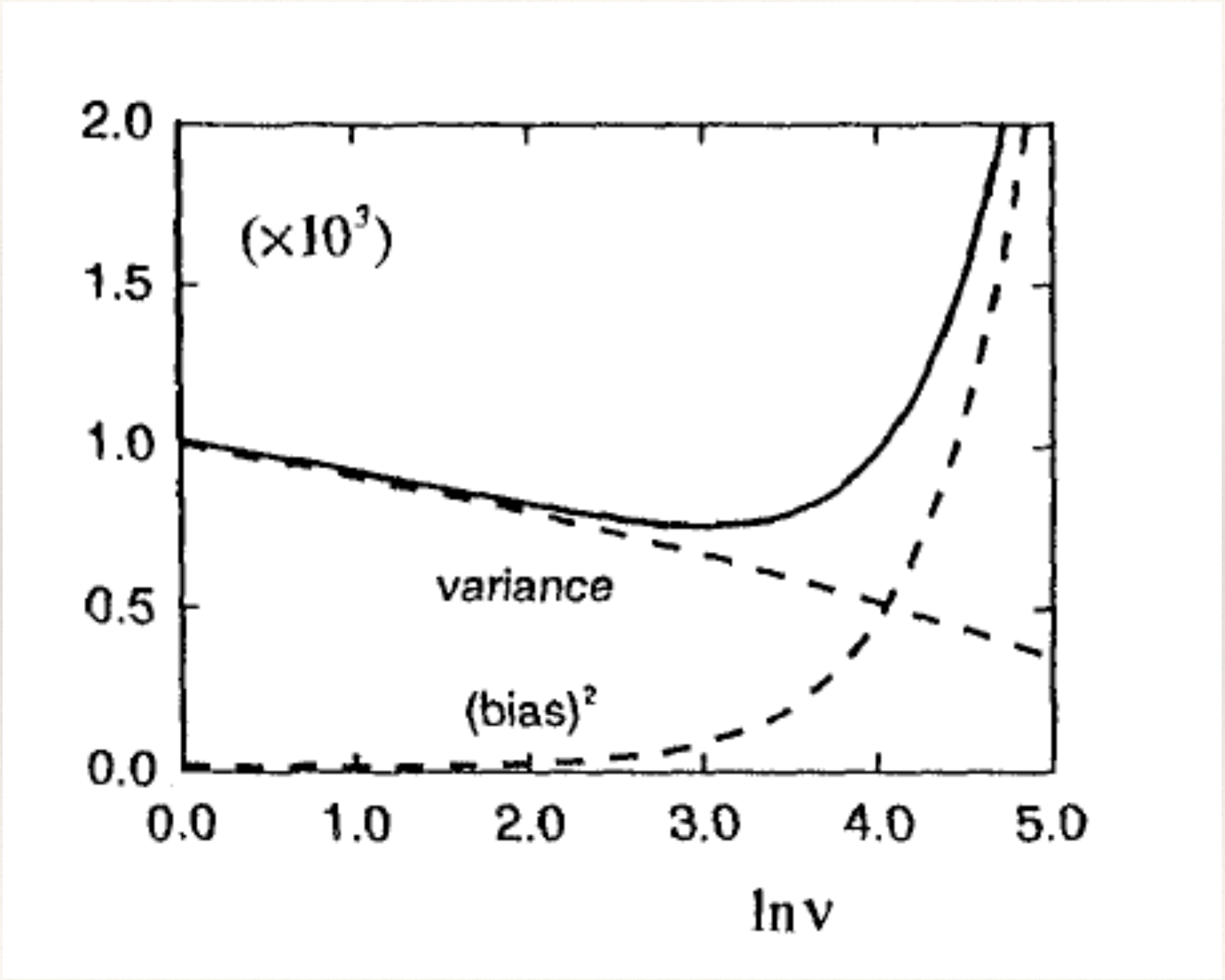
Regularization = 40



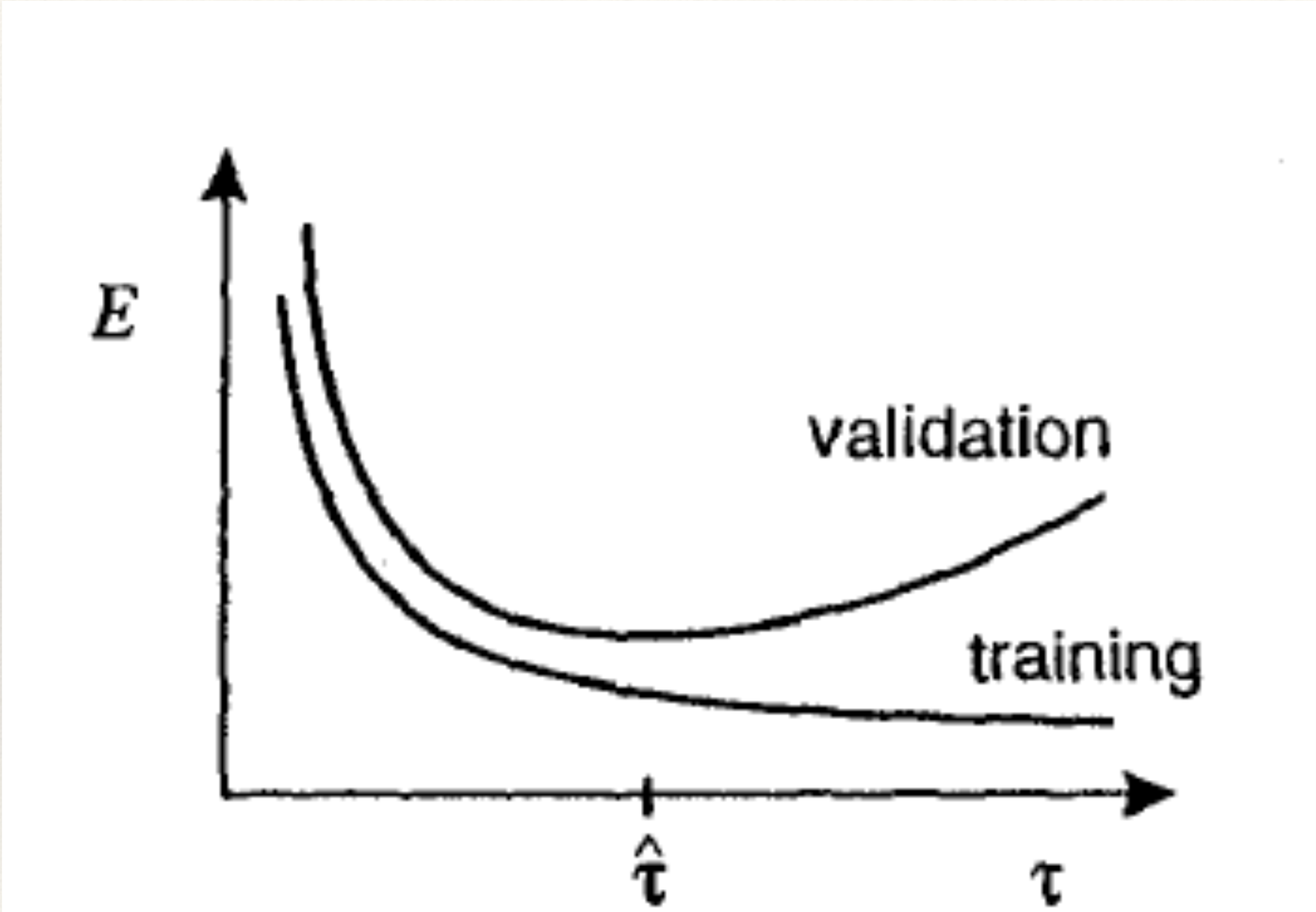
Regularization = 4000



Regularization Effect on Learning



Early Stopping

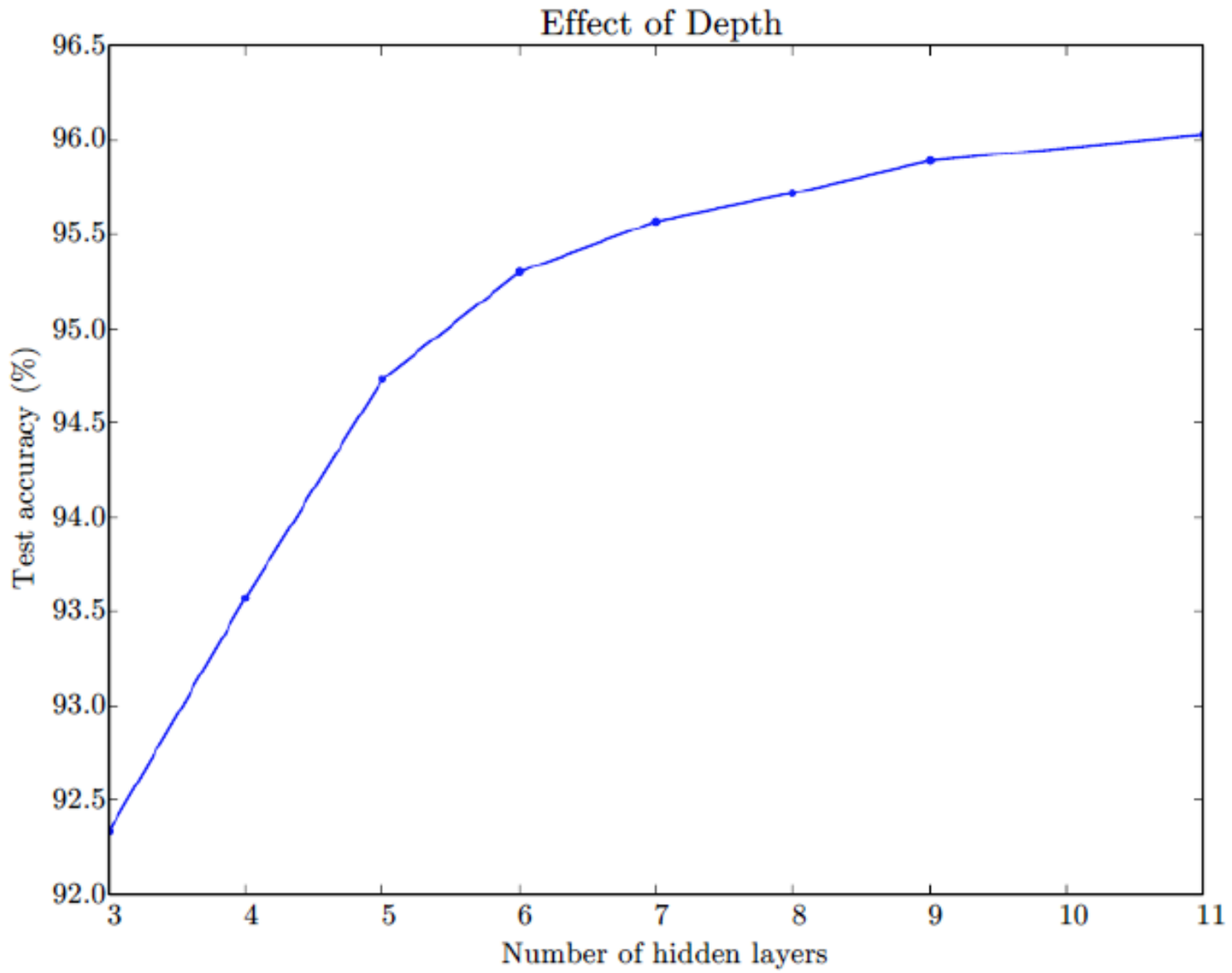


Most Popular in Practice

Neural Networks - Summary

- Details of Architecture
- Computation of gradient using back propagation.
- Error function and output layer activation
  - Neural networks estimate posterior probabilities
- Learning in Neural networks
  - Gradient descent - Properties
- Generalization of Neural Networks

Need for Depth



$$\boldsymbol{h}^{(1)} = g^{(1)}\left(\boldsymbol{W}^{(1)\top}\boldsymbol{x} + \boldsymbol{b}^{(1)}\right)$$

$$\boldsymbol{h}^{(2)} = g^{(2)}\left(\boldsymbol{W}^{(2)\top}\boldsymbol{h}^{(1)} + \boldsymbol{b}^{(2)}\right)$$

Need for Depth