

Notation and Math

Average Training Error

$$E_{\text{in}}(g) = \underbrace{\frac{1}{N} \sum_{i=1}^N}_{\text{Average error on the } N \text{ training examples}} \underbrace{\text{error}(y^{(i)}, \underbrace{g(\mathbf{x}^{(i)})}_{\text{Prediction on input } \mathbf{x}^{(i)}}))}_{\substack{\text{Cost (loss) for} \\ \text{prediction not being} \\ \text{the same as true} \\ \text{label}}}$$

If cost is RSS then

$$E_{\text{in}}(g) = \frac{1}{N} \sum_{i=1}^N \left(y^{(i)} - \underbrace{\mathbf{w}^T \mathbf{x}^{(i)}}_{\substack{\text{Prediction} \\ \text{on input } \mathbf{x}^{(i)}}} \right)^2$$

MSE (mean squared error) over the training data is called the “in sample” error.

**Goal: Mathematical description
of expected errors in the future**

First, a simpler example

Expected Value

i.e. what is the long-term average?

Coin flip game: if heads you pay \$2, tails you get \$1. Do you play?

- What is the **expected** cost/loss?

$$E[A] = -0.5 * 2 + 0.5 * 1 = -0.5$$

- Unfair coin: $p(\text{heads})=0.3$
- Do you play? $E[A] = -0.3 * 2 + 0.7 * 1 = 0.1$
- General formula: $E[A] = \sum_{A=a} a \cdot p(a)$

If the outcome was continuous, then the sum is replaced by an integral, and the distribution is replaced by a density function.

Hypothesis/model depends on which examples it has seen

When we want to emphasize that our hypothesis $g(\mathbf{x})$ depends on which D was used we write $g^{(D)}(\mathbf{x})$

Given $D = \{(\mathbf{x}^{(1)}, y^{(1)}), \dots, (\mathbf{x}^{(N)}, y^{(N)})\}$ drawn i.i.d. from a distribution
Our hypothesis given D is $g(\mathbf{x}) = g^{(D)}(\mathbf{x})$ (example: $g^{(D)}(\mathbf{x}) = \mathbf{w}^T \mathbf{x}$)

Expected Error

In the noise-free case (no ϵ term).

Here we assume that for any \mathbf{x}
there is only one y

The expected test error. What do we *expect* the cost/loss of our hypothesis/model in the future?

$$\underbrace{E_{\text{out}}(g^{(D)})}_{\text{Defining this here}} = E_{\mathbf{x}}[\text{error when using } g^{(D)}] = E_{\mathbf{x}} [\text{error}(y, g(\mathbf{x}))]$$

- If we are using squared error:

$$E_{\text{out}}(g^{(D)}) = E_{\mathbf{x}}[(y - g^{(D)}(\mathbf{x}))^2]$$

If there are only a finite number of values \mathbf{x} can take then

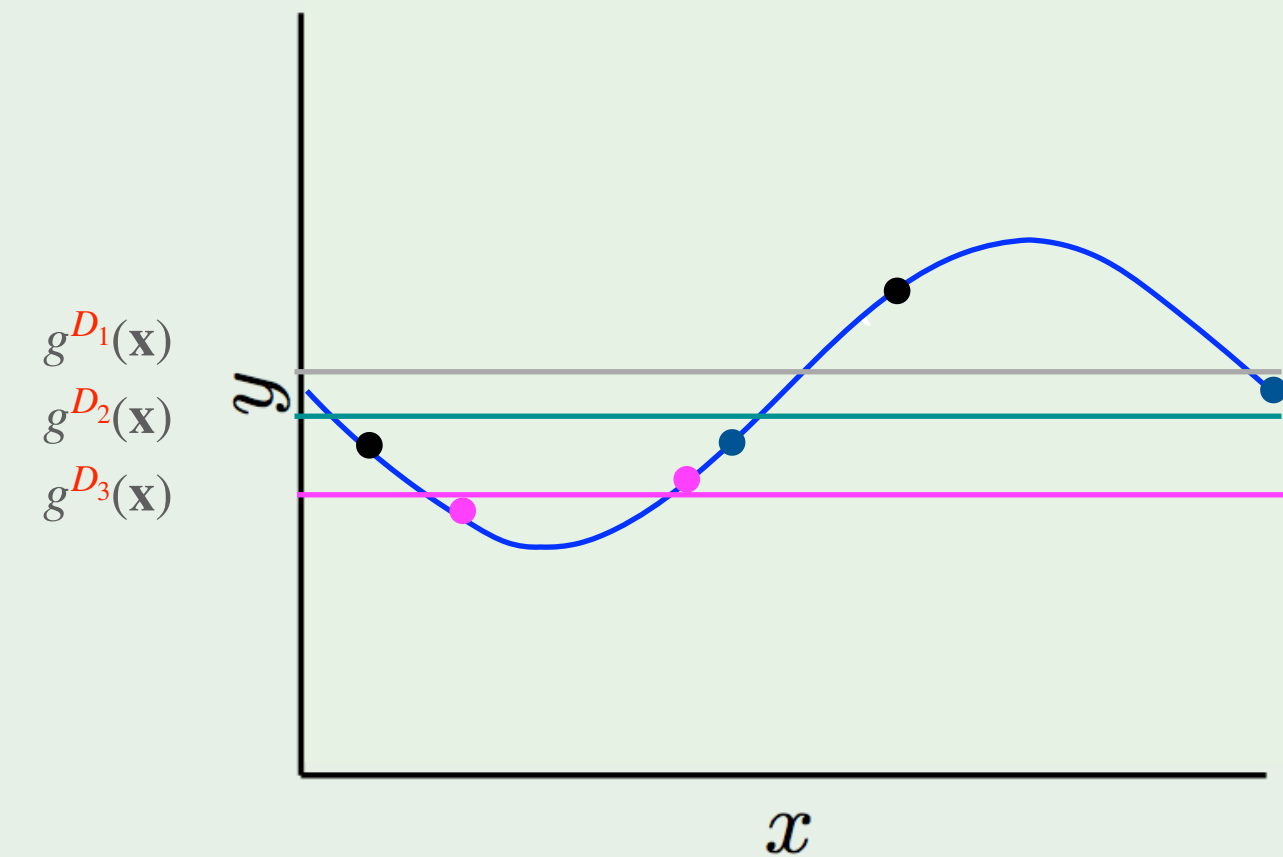
$$E_{\text{out}}(g^{(D)}) = \sum_{\mathbf{x}} (y - g^{(D)}(\mathbf{x}))^2 p(\mathbf{x})$$

- If \mathbf{x} is continuous, then the *sum* is replaced by an *integral*, and the distribution is replaced by a density function.


If there is noise (ϵ term)

$$E_{\text{out}}(g^{(D)}) = E_{\mathbf{x}, y} [\text{error}(y, g^{(D)}(\mathbf{x}))]$$

Average hypothesis $\bar{g}(\mathbf{x})$



Slide modified from

©  Creator: Yaser Abu-Mostafa - LFD Lecture 8

Model class : $g(\mathbf{x}) = w_0$

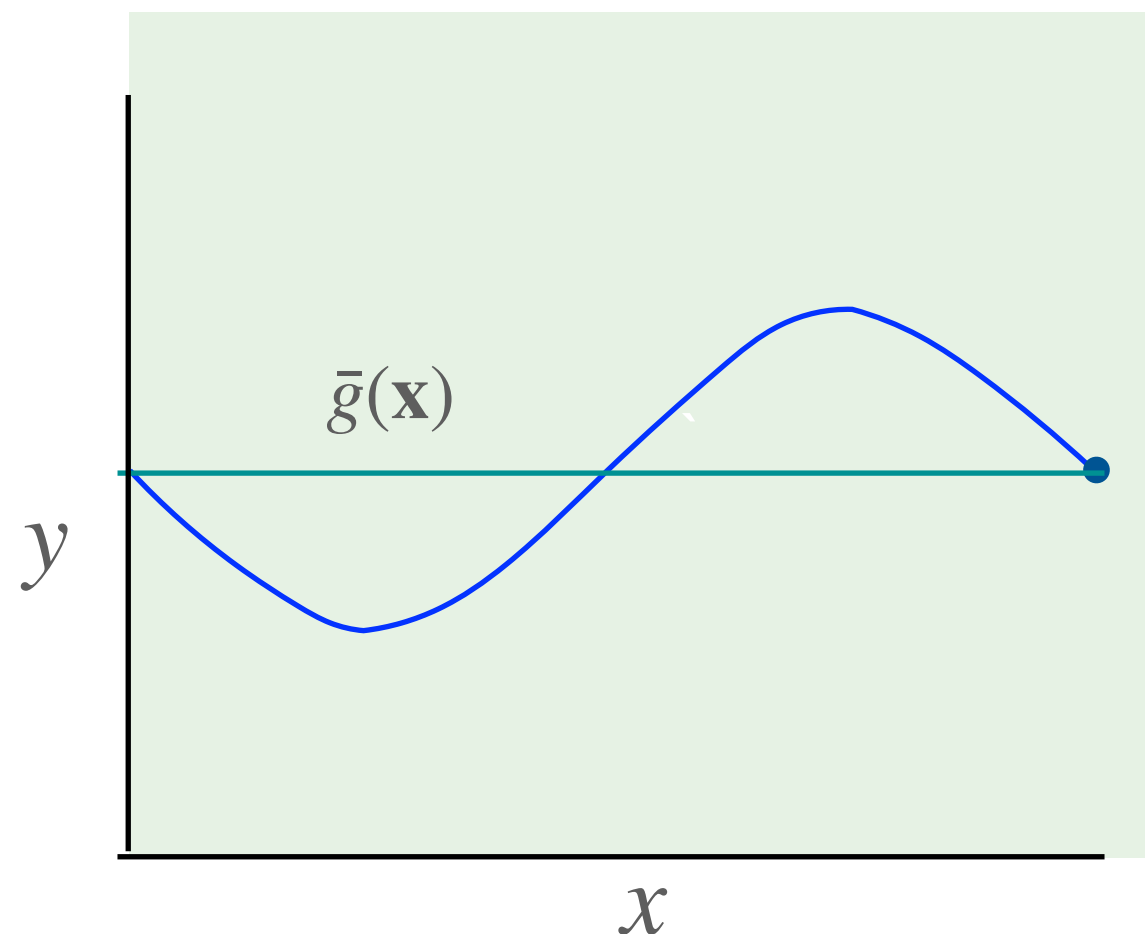
Expected Prediction

$$D = \{(\mathbf{x}^{(1)}, y^{(1)}), \dots, (\mathbf{x}^{(N)}, y^{(N)})\}$$

The average prediction: What do we *expect* our hypothesis will predict after seeing N examples?

$$\bar{g}(\mathbf{x}) = E_D[g^{(D)}(\mathbf{x})] \approx \frac{1}{k} \sum_{i=1}^k g_i^{(D_i)}(\mathbf{x}) \quad D_1, D_2, \dots, D_k$$

Bias (also call “Bias squared”) of a model/hypothesis class



For a fixed \mathbf{x}

$$\text{bias}(\mathbf{x}) = (f(\mathbf{x}) - \bar{g}(\mathbf{x}))^2$$

General case

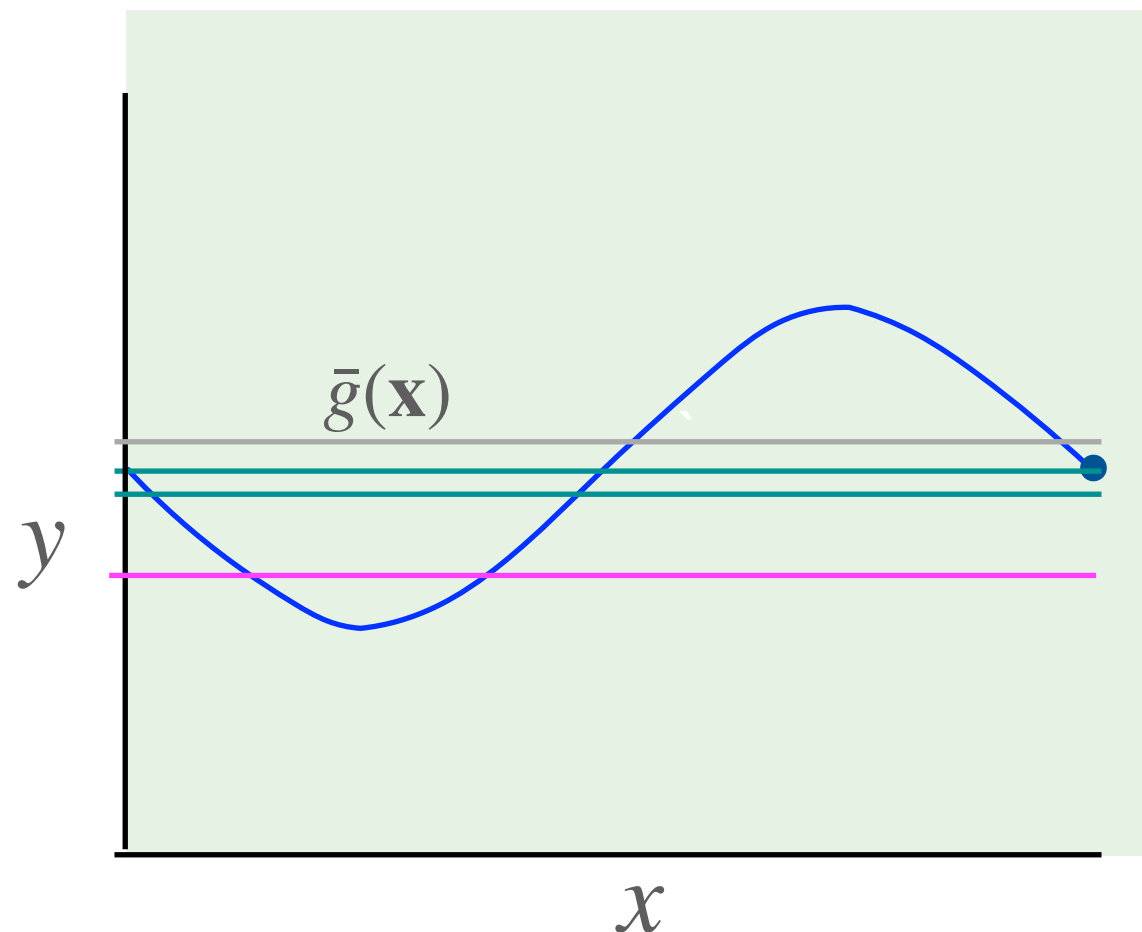
$$\text{bias} = E_{\mathbf{x}}[(f(\mathbf{x}) - \bar{g}(\mathbf{x}))^2]$$

Typo fixed on this slide after class

When using this model class, measures how well you expect the “average prediction” to represent the true solution

We expect the bias to decrease with a more complex model

Variance of a model/hypothesis class



For a fixed \mathbf{x}

$$\text{var}(\mathbf{x}) = E_D[(g^{(D)}(\mathbf{x}) - \bar{g}(\mathbf{x}))^2]$$

General case

$$\text{var} = E_{\mathbf{x}} \left[E_D [(g^{(D)}(\mathbf{x}) - \bar{g}(\mathbf{x}))^2] \right]$$

Typo fixed on this slide after class

Measures how sensitive a hypothesis class (model class) is to a specific dataset
Variance typically decreases with simpler models