Learning From Data
Lecture 7
Approximation Versus Generalization

M. Magdon-Ismail
CSCI 4100/6100

Bias-Variance Analysis

Another way to quantify the tradeoff:

1. How well *can* the learning approximate f.

   ...as opposed to how well *did* the learning approximate f in-sample (E_{in}).
2. How close can you get to that approximation with a finite data set.

   ...as opposed to how close is E_{in} to E_{out}.

Bias-variance analysis applies to squared errors (classification and regression)

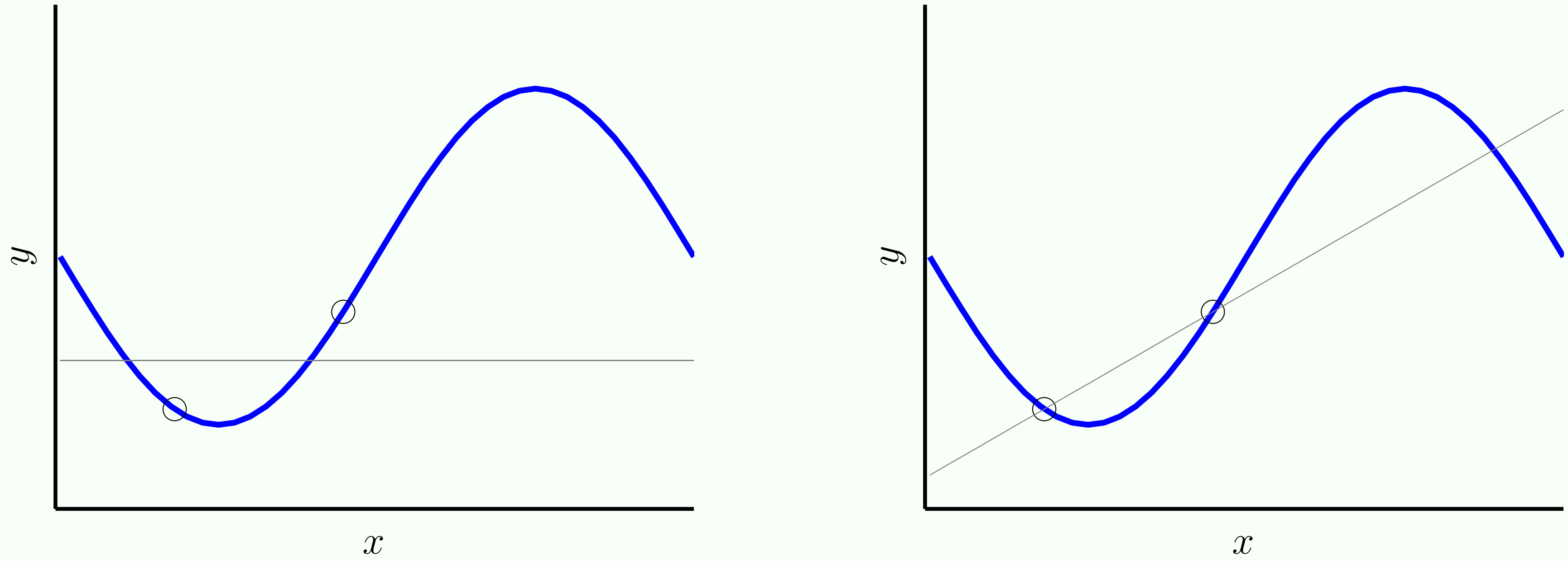
Bias-variance analysis can take into account the *learning algorithm*

Different learning algorithms can have different E_{out} when applied to the same $\mathcal{H}$!

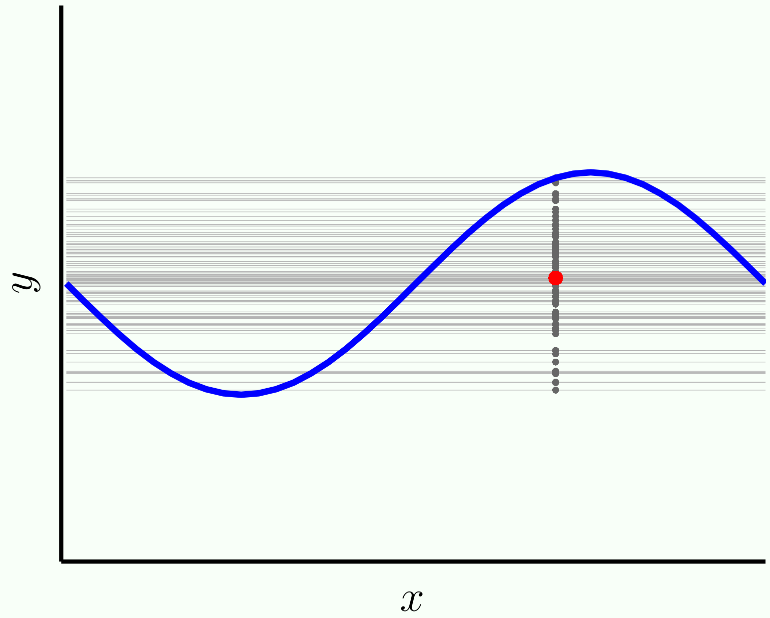
A Simple Learning Problem

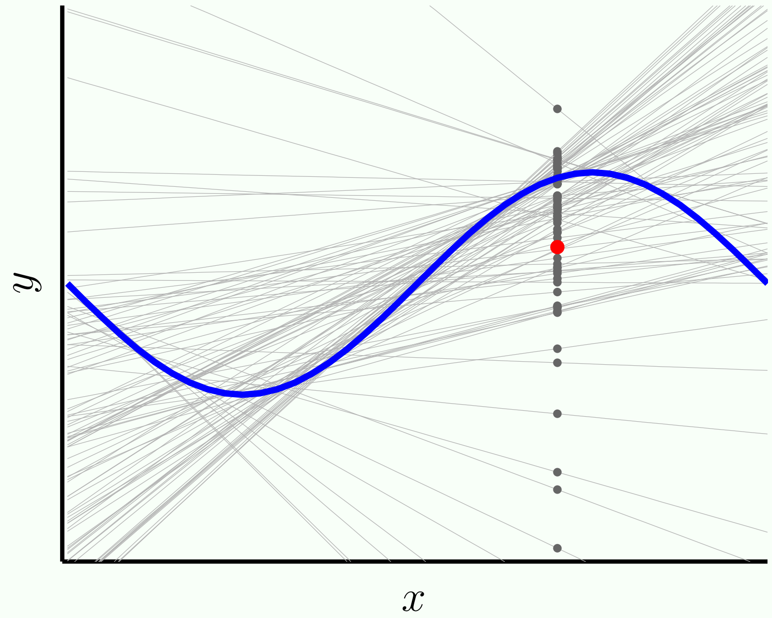
2 Data Points. 2 hypothesis sets:

$$\mathcal{H}_0 : \quad h(x) = b$$
$$\mathcal{H}_1 : \quad h(x) = ax + b$$



Let's Repeat the Experiment Many Times

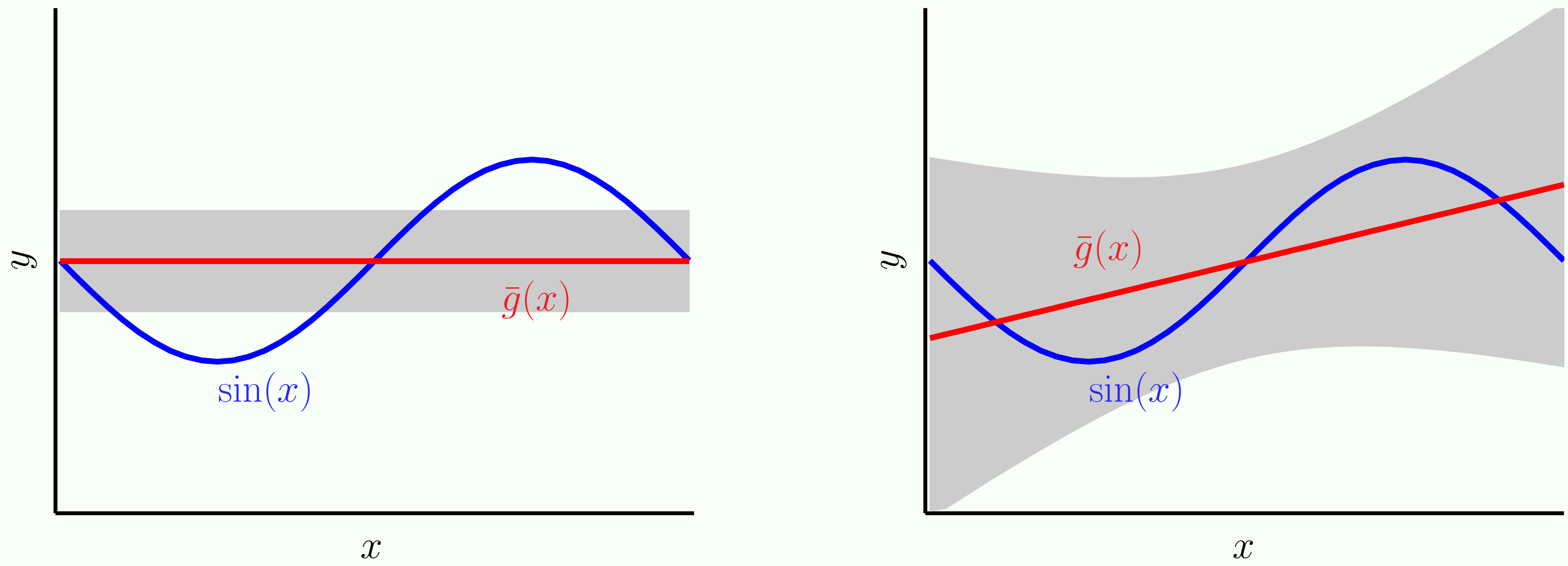




For each data set $\mathcal{D}$, you get a different $g^{\mathcal{D}}$.

So, for a fixed $\mathbf{x}$, $g^{\mathcal{D}}(\mathbf{x})$ is random value, depending on $\mathcal{D}$.

What's Happening on Average



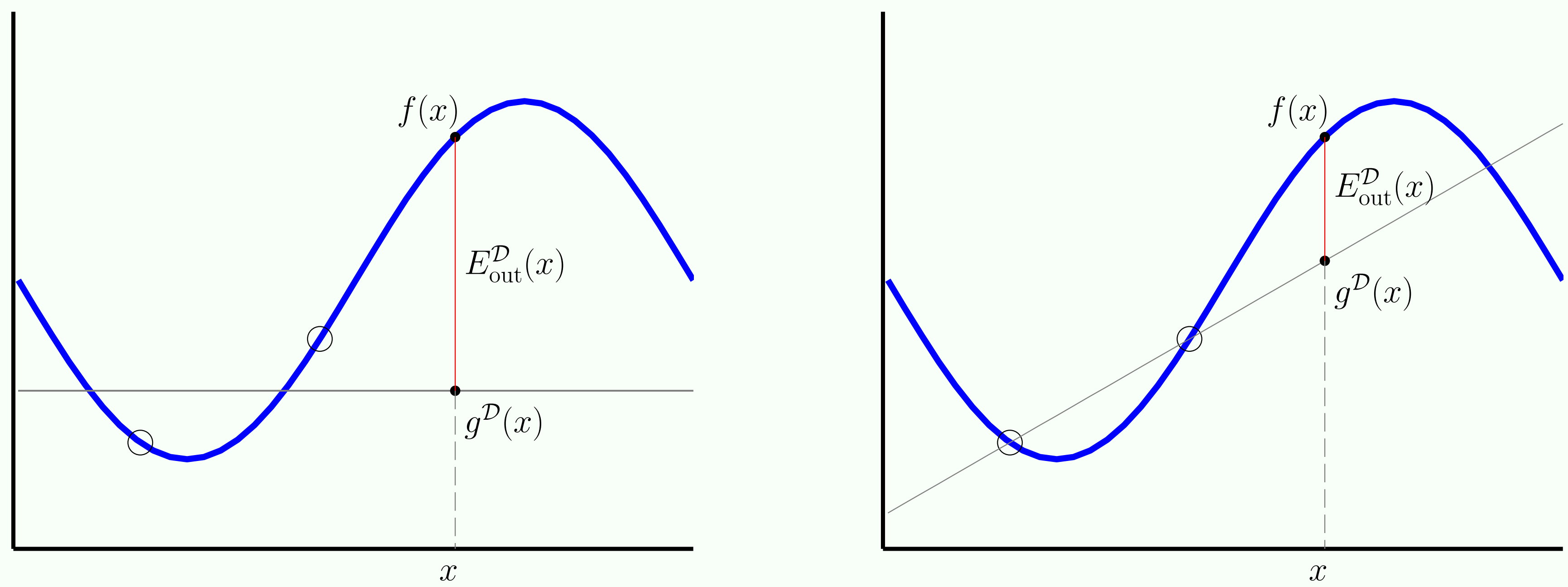
We can define:

$$g^{\mathcal{D}}(\mathbf{x}) \qquad \leftarrow \textbf{random value}\text{, depending on } \mathcal{D}$$

$$\begin{aligned} \bar{g}(\mathbf{x}) &= \mathbb{E}_{\mathcal{D}}\left[g^{\mathcal{D}}(\mathbf{x})\right] \\ &\approx \tfrac{1}{K}(g^{\mathcal{D}_1}(\mathbf{x}) + \cdots + g^{\mathcal{D}_K}(\mathbf{x})) \end{aligned} \qquad \leftarrow \text{your average prediction on } \mathbf{x}$$

$$\begin{aligned} \mathsf{var}(\mathbf{x}) &= \mathbb{E}_{\mathcal{D}}\left[(g^{\mathcal{D}}(\mathbf{x}) - \bar{g}(\mathbf{x}))^2\right] \\ &= \mathbb{E}_{\mathcal{D}}\left[g^{\mathcal{D}}(\mathbf{x})^2\right] - \bar{g}(\mathbf{x})^2 \end{aligned} \qquad \leftarrow \text{how variable is your prediction?}$$

E_{out} on Test Point x for Data $\mathcal{D}$



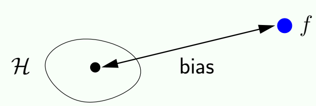
$$E_{\text{out}}^{\mathcal{D}}(\mathbf{x}) = (g^{\mathcal{D}}(\mathbf{x}) - f(\mathbf{x}))^2$$ $\leftarrow$ **squared error**, a random value depending on $\mathcal{D}$

$$E_{\text{out}}(\mathbf{x}) = \mathbb{E}_{\mathcal{D}}\left[E_{\text{out}}^{\mathcal{D}}(\mathbf{x})\right]$$ $\leftarrow$ expected $E_{\text{out}}(\mathbf{x})$ before seeing $\mathcal{D}$

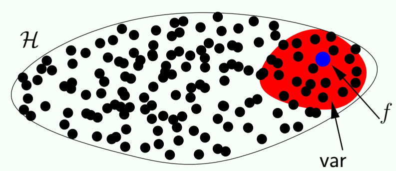
The Bias-Variance Decomposition

$$
\begin{aligned}
E_{\text{out}}(\mathbf{x}) &= \mathbb{E}_{\mathcal{D}}\left[(g^{\mathcal{D}}(\mathbf{x}) - f(\mathbf{x}))^2\right] \\
&= \mathbb{E}_{\mathcal{D}}\left[g^{\mathcal{D}}(\mathbf{x})^2 - 2g^{\mathcal{D}}(\mathbf{x})f(\mathbf{x}) + f(\mathbf{x})^2\right] \\
&= \mathbb{E}_{\mathcal{D}}\left[g^{\mathcal{D}}(\mathbf{x})^2\right] - 2\bar{g}(\mathbf{x})f(\mathbf{x}) + f(\mathbf{x})^2 \qquad \leftarrow \text{understand this; the rest is just algebra} \\
&= \mathbb{E}_{\mathcal{D}}\left[g^{\mathcal{D}}(\mathbf{x})^2\right] - \bar{g}(\mathbf{x})^2 + \bar{g}(\mathbf{x})^2 - 2\bar{g}(\mathbf{x})f(\mathbf{x}) + f(\mathbf{x})^2 \\
&= \underbrace{\mathbb{E}_{\mathcal{D}}\left[g^{\mathcal{D}}(\mathbf{x})^2\right] - \bar{g}(\mathbf{x})^2}_{\mathsf{var}(\mathbf{x})} + \underbrace{(\bar{g}(\mathbf{x}) - f(\mathbf{x}))^2}_{\mathsf{bias}(\mathbf{x})}
\end{aligned}
$$

$$\boxed{E_{\text{out}}(\mathbf{x}) = \mathsf{bias}(\mathbf{x}) + \mathsf{var}(\mathbf{x})}$$



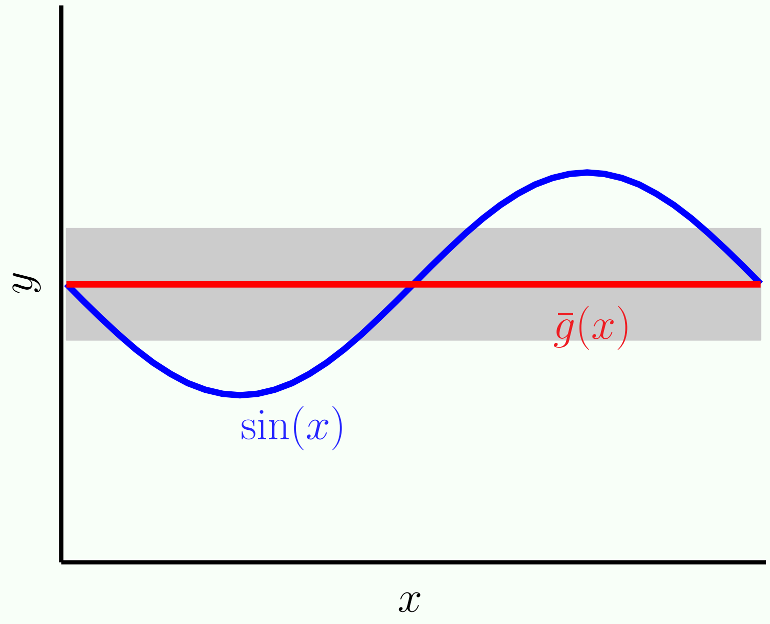
Very small model

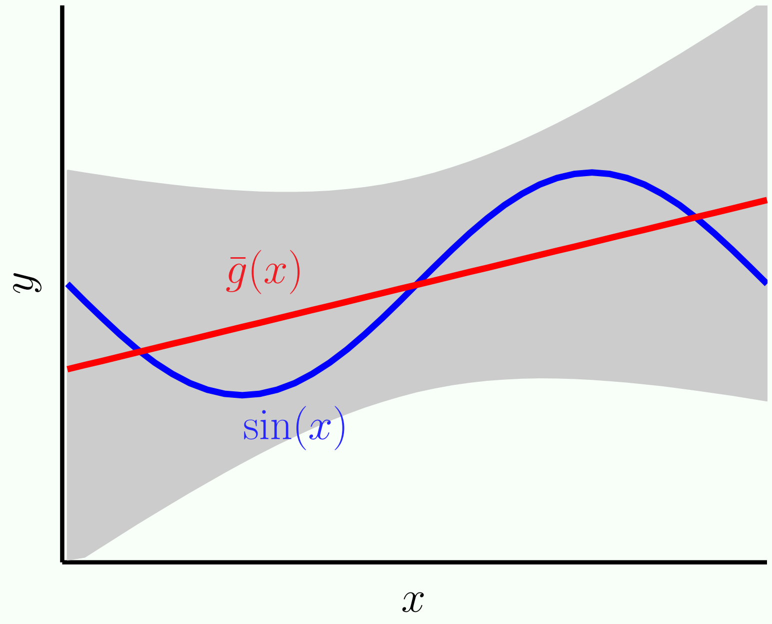


Very large model

If you take average over $\mathbf{x}$: $\quad E_{\text{out}} = \mathsf{bias} + \mathsf{var}$

Back to $\mathcal{H}_0$ and $\mathcal{H}_1$; and, our winner is ...

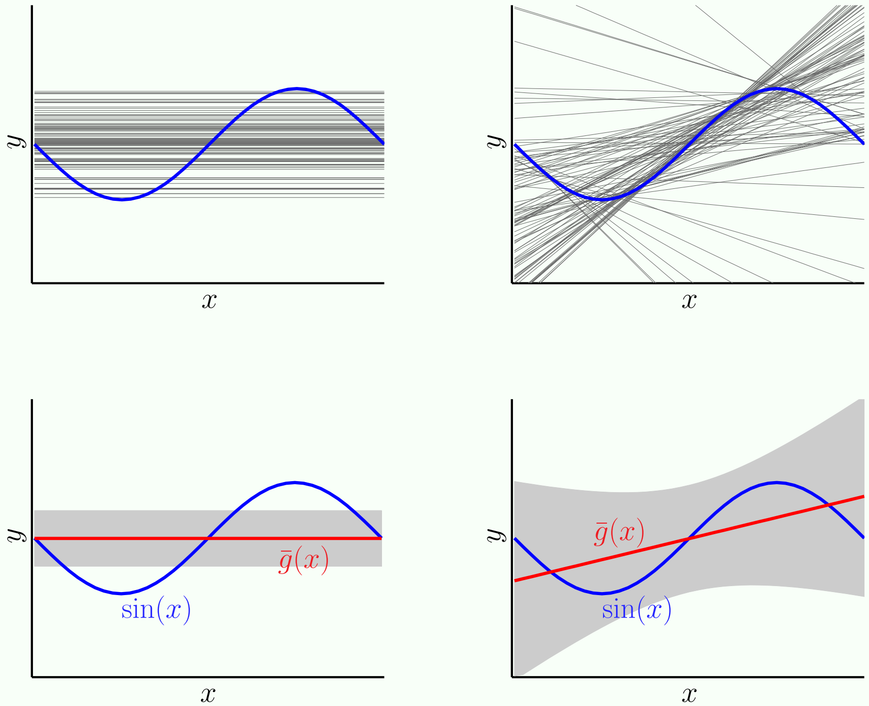




$\mathcal{H}_0$	$\mathcal{H}_1$
bias $= 0.50$	bias $= 0.21$
var $= 0.25$	var $= 1.69$
$E_{\text{out}} = 0.75$ ✓	$E_{\text{out}} = 1.90$

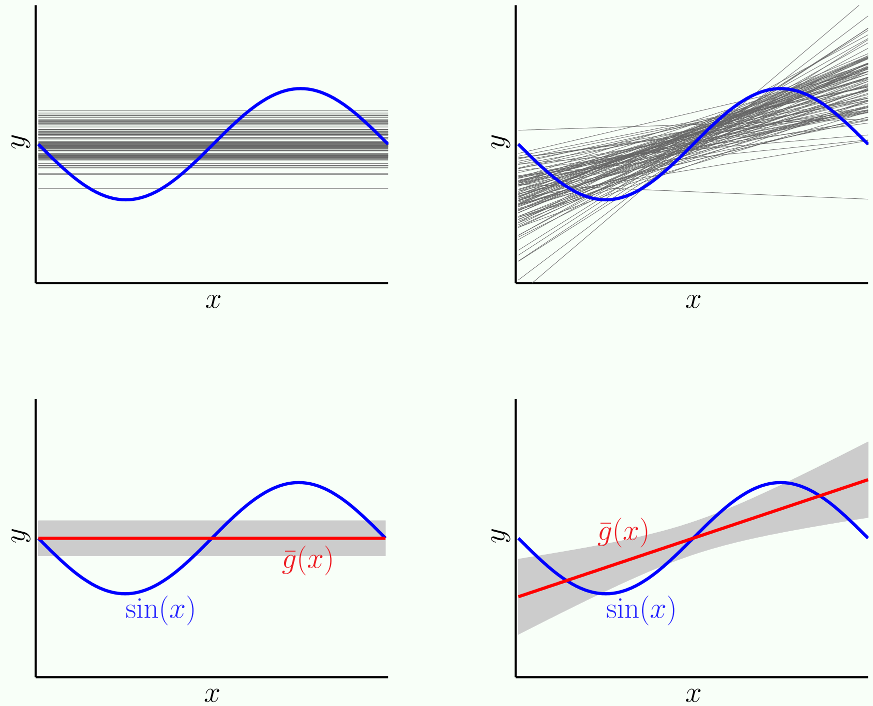
Match Learning Power to Data, . . . Not to f

2 Data Points



$\mathcal{H}_0$	$\mathcal{H}_1$
bias $= 0.50$;	bias $= 0.21$;
var $= 0.25$.	var $= 1.69$.
$E_{\text{out}} = 0.75$ ✓	$E_{\text{out}} = 1.90$

5 Data Points



$\mathcal{H}_0$	$\mathcal{H}_1$
bias $= 0.50$;	bias $= 0.21$;
var $= 0.1$.	var $= 0.21$.
$E_{\text{out}} = 0.6$	$E_{\text{out}} = 0.42$ ✓