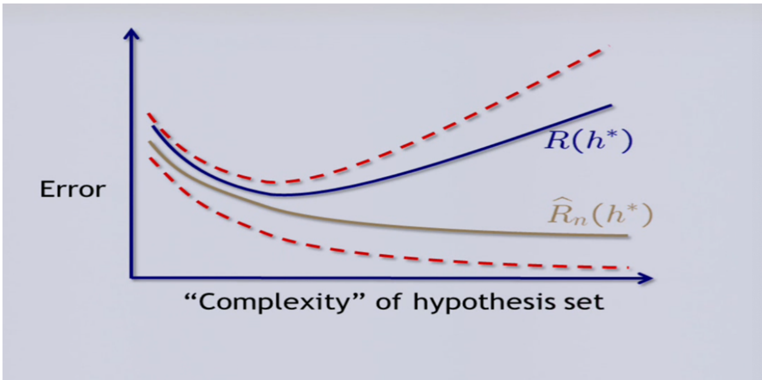
ECE 0402 - Pattern Recognition

Lecture 11

Recap: True performance lives somewhere in-between the red dashed curves:



This graph, at the very least, gives us someway of understanding the tradeoff.

- VC bounds gives us a crude way of handling this tradeoff.

$$R(h) \lessapprox \hat{R}_n(h) + \epsilon(\mathcal{H}, n)$$

- "bias-variance" decomposition is alternative (extra) way of understanding this tradeoff

  In the last lecture, we noted that bias-variance decomposition is especially useful because it easily generalizes to regression.

  - bias: how well can $\mathcal{H}$ approximate f^*
  - variance: how well can we pick a good $h \in \mathcal{H}$

$$\implies R(h) = \text{bias} + \text{variance}$$

    Note that this formulation does not have anything to do with training error (as oppose to VC bound). We will control "overfitting" by trading off between these two terms.

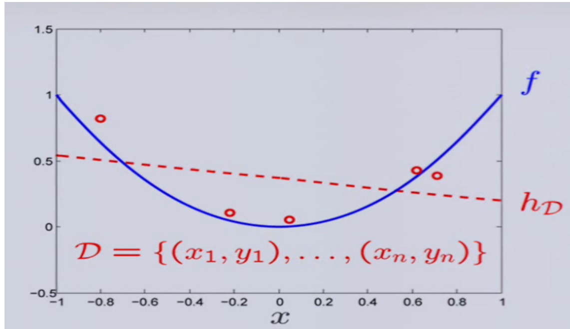
- And next we will talk about practical ways of controlling this...

Notation:

$\mathcal{D} = \{(x_1, y_1), ..., (x_n, y_n)\}$ where$x \in \mathbb{R}^d$ and $y \in \mathbb{R}$

$f : \mathbb{R}^d \to \mathbb{R}$: unknown target function

$h_\mathcal{D} : \mathbb{R}^d \to \mathbb{R}$: function in $\mathcal{H}$ we pick using $\mathcal{D}$



Expected squared error for a given function $h_{\mathcal{D}}$: (mean-squared error)

$$R(h_{\mathcal{D}}) = \mathbb{E}_X \left[(h_{\mathcal{D}}(X) - f(X))^2 \right]$$

Notice here $h_{\mathcal{D}}$ is random which depends on $\mathcal{D}$.

Review the linear fit in the figure below. After we observe some input-output pairs, we come up with a linear-line, and this line depends on this dataset!

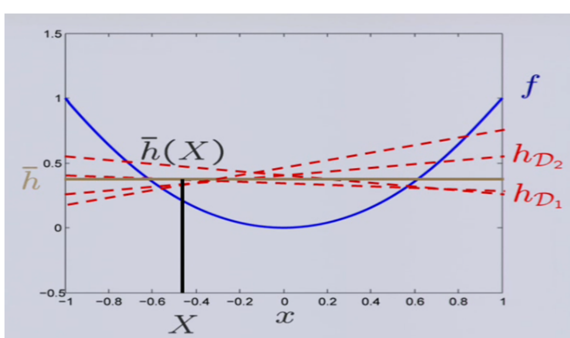
$$\begin{aligned}
\mathbb{E}_{\mathcal{D}} \left[R(h_{\mathcal{D}}) \right] &= \mathbb{E}_{\mathcal{D}} \left[\mathbb{E}_X \left[(h_{\mathcal{D}}(X) - f(X))^2 \right] \right] \\
&= \mathbb{E}_X \left[\mathbb{E}_{\mathcal{D}} \left[(h_{\mathcal{D}}(X) - f(X))^2 \right] \right]
\end{aligned}$$

We said let's fix X for a moment and focus on evaluating $\mathbb{E}_{\mathcal{D}} \left[(h_{\mathcal{D}}(X) - f(X))^2 \right]$.

To evaluate this we will define average hypothesis: $\bar{h}(X) = \mathbb{E}_D \left[h_{\mathcal{D}}(X) \right]$ average over all hypotheses

Interpretation of this could be: imagine drawing many data sets $\mathcal{D}_1, ..., \mathcal{D}_p$, and averaging them.

$$\bar{h}(X) \approx \frac{1}{p} \sum_{i=1}^{p} h_{\mathcal{D}_i}(X)$$



$$
\begin{aligned}
&\mathbb{E}_D\left[(h_{\mathcal{D}}(X) - f(X))^2\right] \\
&\qquad = \mathbb{E}_D\left[(h_{\mathcal{D}}(X) - \bar{h}(X) + \bar{h}(X) - f(X))^2\right] \\
&\qquad = \mathbb{E}_D\left[(h_{\mathcal{D}}(X) - \bar{h}(X))^2 + (\bar{h}(X) - f(X))^2 + 2(h_{\mathcal{D}}(X) - \bar{h}(X))(\bar{h}(X) - f(X))\right] \\
&\qquad = \mathbb{E}_D\left[(h_{\mathcal{D}}(X) - \bar{h}(X))^2\right] + (\bar{h}(X) - f(X))^2
\end{aligned}
$$

Finally plugging back this into the we found,

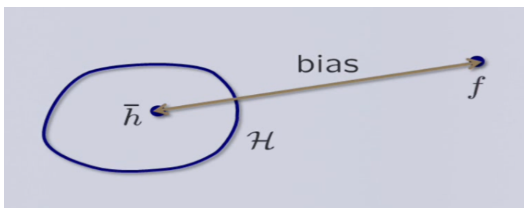
$$
\begin{aligned}
R(h_{\mathcal{D}}) &= \mathbb{E}_X\left[\mathbb{E}_D\left[(h_{\mathcal{D}}(X) - f(X))^2\right]\right] \\
&= \mathbb{E}_X\left[\text{bias}(X) + \text{variance}(X)\right] \\
&= \text{bias} + \text{variance}
\end{aligned}
$$

↳ variance of hypothesis

Visualization the bias:

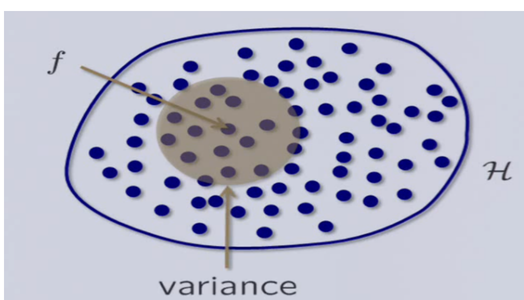
$$
\text{bias} = \mathbb{E}\left[(\bar{h}(X) - f(X))^2\right]
$$

i.e. trying to fit a straight line to a quadratic function will never be perfect



Visualization the variance:

$$
\text{variance} = \mathbb{E}_X\left[\mathbb{E}_D\left[(h_{\mathcal{D}}(X) - \bar{h}(X))^2\right]\right]
$$



Example: Suppose $f(x) = sin(\pi x)$, x are drawn uniformly from $[-1, 1]$, and we get 2 training examples.

We are going to consider two different hypothesis sets:

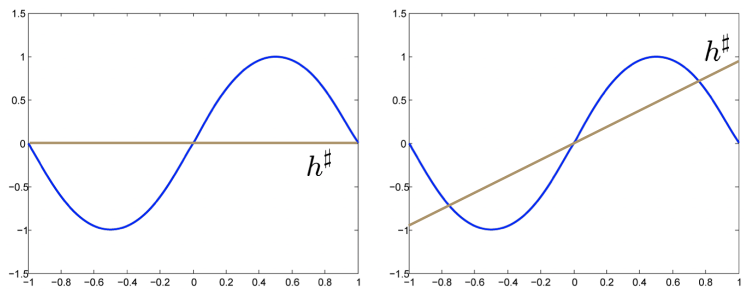
Horizontal line

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

- Which one is better over this interval? Neither

$h^\sharp$ are optimal values for $h(x) = b$ and $h(x) = ax+b$

$\mathcal{H}_0$ $\qquad$ $\mathcal{H}_1$

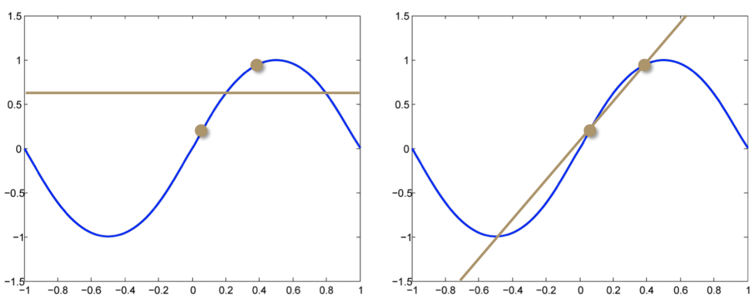


$R(h^\sharp) = \frac{1}{2}$ $\qquad$ $R(h^\sharp) = \frac{1}{2} - \frac{3}{\pi^2} \approx 0.196$

Error/risk

- This is the case if you know f! How about estimating these from data?

$\mathcal{H}_0$ $\qquad$ $\mathcal{H}_1 \to$ offset line



- What is the average hypothesis–so that we can calculate bias and variance of these estimators?
- we are looking at: $\mathbb{E}_{\mathcal{D}}\left[R(h_{\mathcal{D}})\right] =$ bias + variance
  - offset-line has a smaller bias
  - but it has bigger variance
  - hence, the winner is...

Randomly generated 2 data points 10000 times, then fit both lines to it

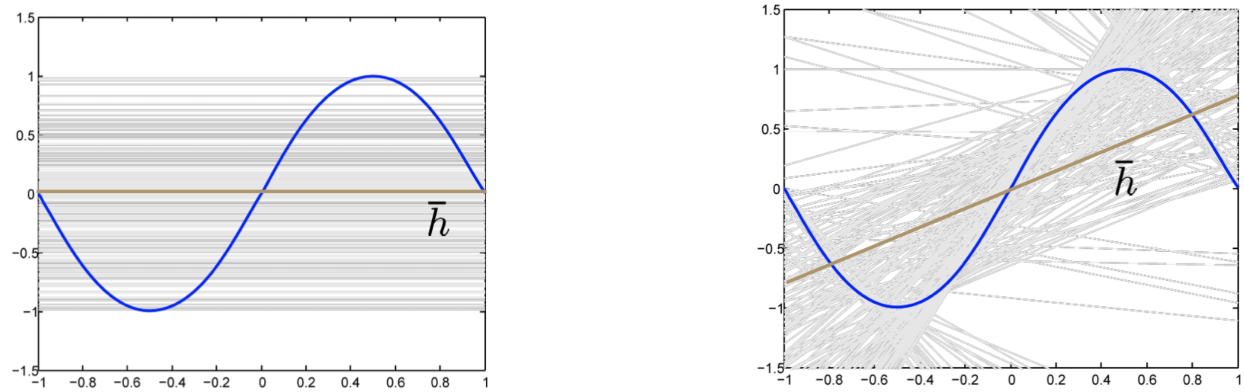
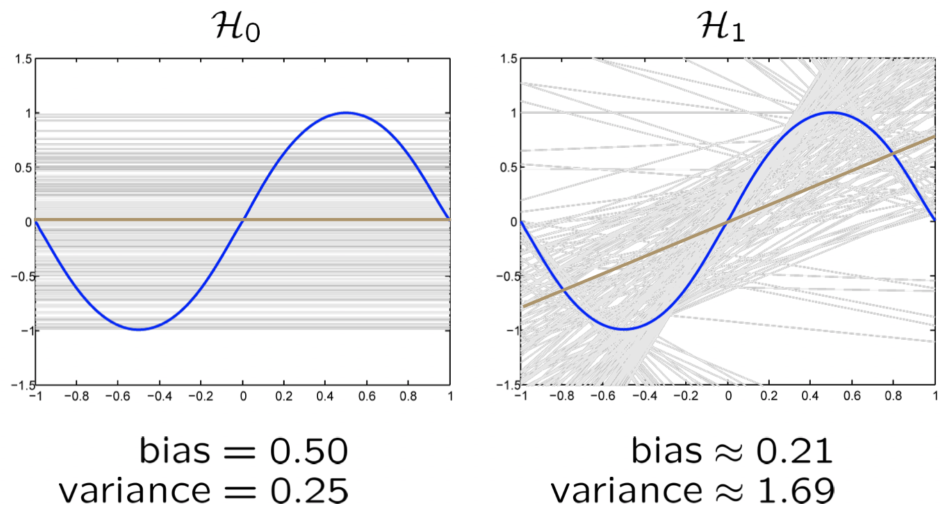


Figure 1: average hypothesis

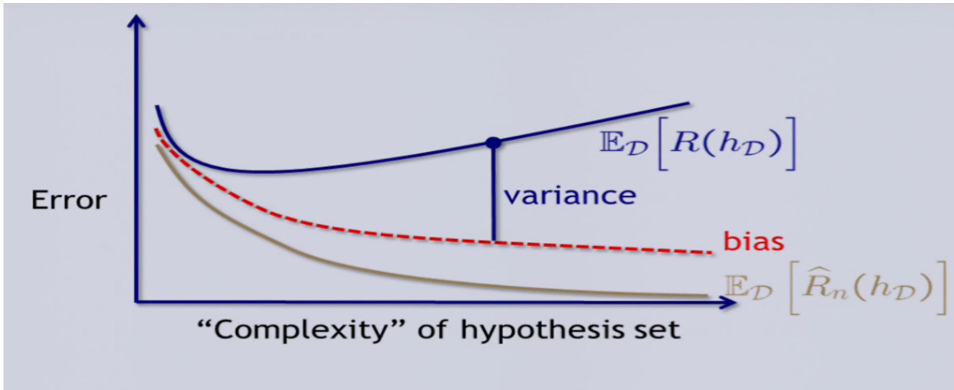


Summary:

- VC bound says: keep the "model complexity" small enough relative to how much data we have n and we can learn **any** f –emprical risk and true risk will align with each other
- Bias-variance decomposition says: suppose we have any particular f, we do best by matching the "model complexity" to the "data resources" –not to f

Moral of this story is basically the same! You need to kind of match "how complicated of a model you're dealing with" to "how much data you have"–not necessarily to "how complicated is the thing that you are trying to estimate".

- increasing the model complexity to reduce bias
- decreasing the model complexity to reduce variance



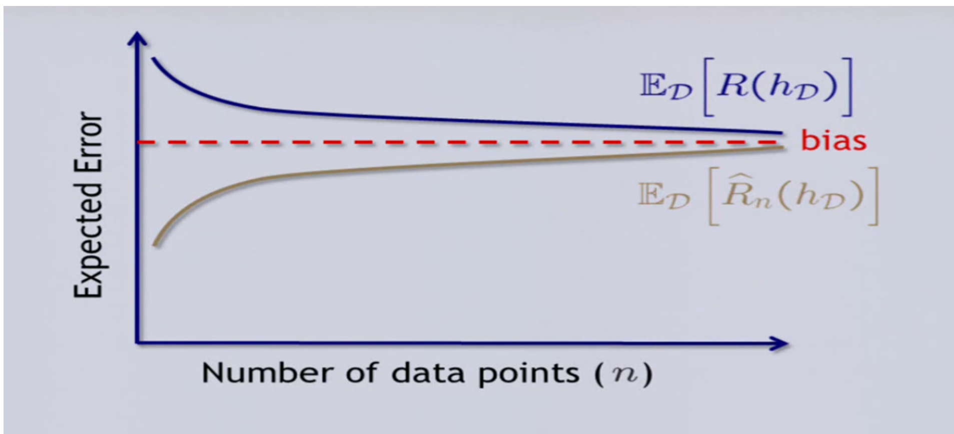


Figure 2: Simple Model

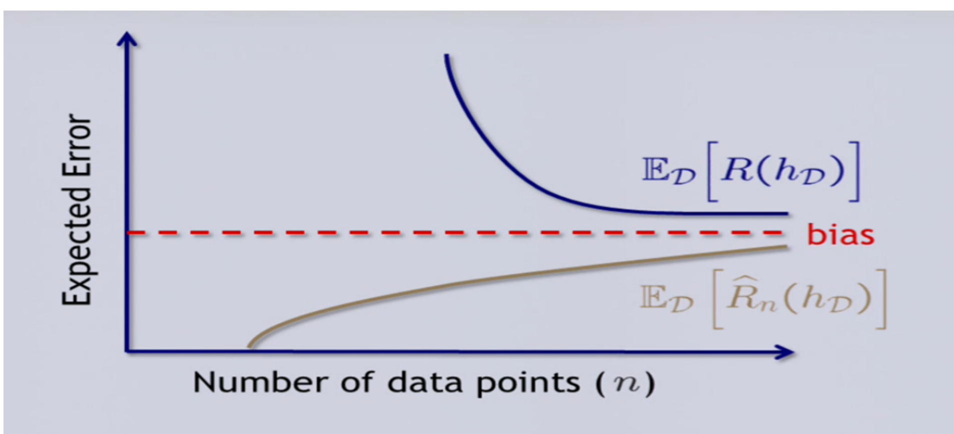


Figure 3: Complex Model