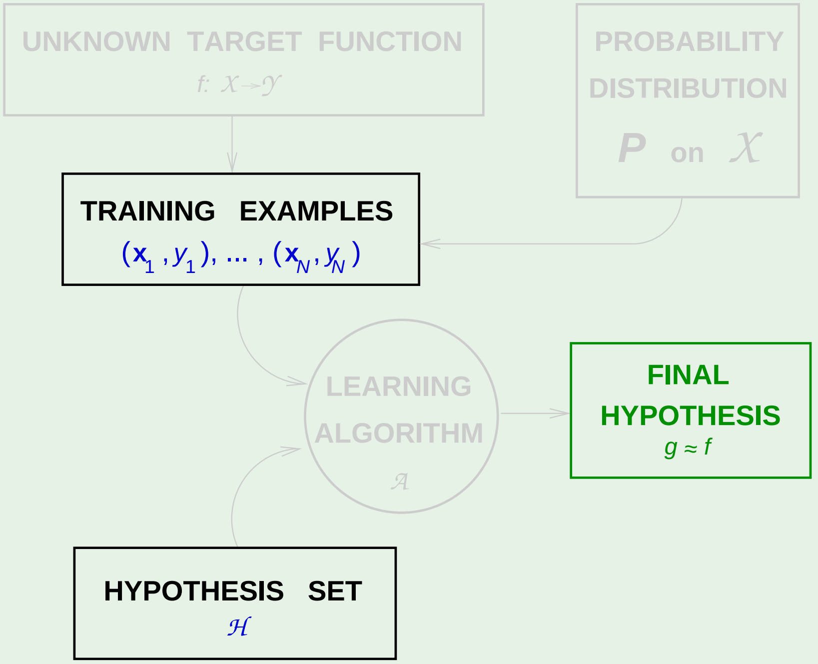
Review of Lecture 7

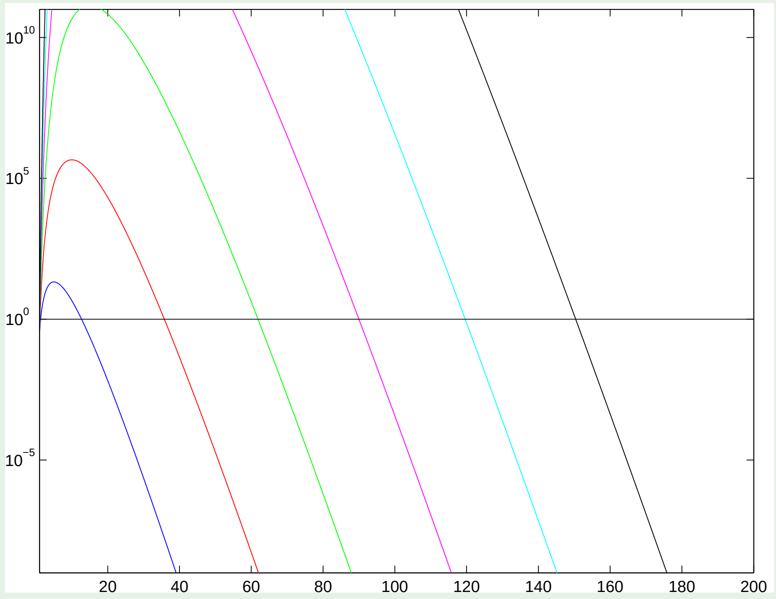
- VC dimension $d_{\mathrm{VC}}(\mathcal{H})$

most points $\mathcal{H}$ can shatter

- Scope of VC analysis



- Utility of VC dimension



$$N \propto d_{\mathrm{VC}}$$

Rule of thumb: $N \geq 10\, d_{\mathrm{VC}}$

- Generalization bound

$$E_{\mathrm{out}} \leq E_{\mathrm{in}} + \Omega$$

In reality, Eout = Ein + omega*, where omega* is something that behaves like omega

Learning From Data

Yaser S. Abu-Mostafa
California Institute of Technology

Lecture 8: Bias-Variance Tradeoff

Sponsored by Caltech's Provost Office, E&AS Division, and IST • Thursday, April 26, 2012

Outline

- Bias and Variance

- Learning Curves

Approximation-generalization tradeoff

Small E_{out}: good approximation of f out of sample.

More complex $\mathcal{H} \Longrightarrow$ better chance of **approximating** f

on the training data

Less complex $\mathcal{H} \Longrightarrow$ better chance of **generalizing** out of sample

Ideal $\mathcal{H} = \{f\}$ winning lottery ticket ☺

With more hypotheses, H is more likely to contain one which can approximate f, however it is more difficult to identify the good hypothesis without sufficient data. We do not know f, so we will have to make H big enough to stand a chance of it containing a good approximate hypothesis, but the learning process/ navigating through H using the training data means that we are less likely to find it/will likely end up with a poor hypothesis and generalization out of sample may be poor.

When you select your hypothesis set, you should balance these two conflicting goals; to have some hypothesis in H that can approximate f, and to enable the data to zoom in on the right hypothesis (i.e. hoping the data will pin down that hypothesis).

Quantifying the tradeoff

VC analysis was one approach: $E_{\text{out}} \leq E_{\text{in}} + \Omega$

Bias-variance analysis is another: decomposing E_{out} into two different error terms:

1. How well $\mathcal{H}$ can approximate f overall/in reality - as if we had access to the target function and we look for which h best describes f, then we quantify how well this performs

2. How well we can zoom in on a good $h \in \mathcal{H}$ We have now found the best hypothesis with a certain approximation ability, we now need to pick it. So we use the training examples to zoom in on H to try pick h - but can we zoom in on it or do we get something that is a poor approximation of the approximation?

Applies to **real-valued targets** and uses **squared error**
(regression)

Start with E_{out}

$$E_{\text{out}}(g^{(\mathcal{D})}) = \mathbb{E}_{\mathbf{x}}\left[\left(g^{(\mathcal{D})}(\mathbf{x}) - f(\mathbf{x})\right)^2\right]$$

We want to decompose this quantity into the two conceptual components of approximation and generalization we saw before

$$\begin{aligned}\mathbb{E}_{\mathcal{D}}\left[E_{\text{out}}(g^{(\mathcal{D})})\right] &= \mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_{\mathbf{x}}\left[\left(g^{(\mathcal{D})}(\mathbf{x}) - f(\mathbf{x})\right)^2\right]\right] \\ &= \mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_{\mathcal{D}}\left[\left(g^{(\mathcal{D})}(\mathbf{x}) - f(\mathbf{x})\right)^2\right]\right]\end{aligned}$$

We want to remove the dependence on the particular data sample that we are given and express Eout for a general N data points

Simply change the order of integration, allowed since non-negative integrand

Now, let us focus on:

$$\mathbb{E}_{\mathcal{D}}\left[\left(g^{(\mathcal{D})}(\mathbf{x}) - f(\mathbf{x})\right)^2\right]$$

The average hypothesis

To evaluate $\mathbb{E}_{\mathcal{D}}\left[\left(g^{(\mathcal{D})}(\mathbf{x}) - f(\mathbf{x})\right)^2\right]$

we define the 'average' hypothesis $\bar{g}(\mathbf{x})$:

$$\bar{g}(\mathbf{x}) = \mathbb{E}_{\mathcal{D}}\left[g^{(\mathcal{D})}(\mathbf{x})\right]$$

Imagine **many** data sets $\mathcal{D}_1, \mathcal{D}_2, \cdots, \mathcal{D}_K$

$$\bar{g}(\mathbf{x}) \approx \frac{1}{K}\sum_{k=1}^{K} g^{(\mathcal{D}_k)}(\mathbf{x})$$

Using $\bar{g}(\mathbf{x})$

$$\mathbb{E}_{\mathcal{D}}\left[\left(g^{(\mathcal{D})}(\mathbf{x}) - f(\mathbf{x})\right)^2\right] = \mathbb{E}_{\mathcal{D}}\left[\left(g^{(\mathcal{D})}(\mathbf{x}) - \bar{g}(\mathbf{x}) + \bar{g}(\mathbf{x}) - f(\mathbf{x})\right)^2\right]$$

$$= \mathbb{E}_{\mathcal{D}}\Big[\left(g^{(\mathcal{D})}(\mathbf{x}) - \bar{g}(\mathbf{x})\right)^2 + \left(\bar{g}(\mathbf{x}) - f(\mathbf{x})\right)^2$$

$$+\ 2\ \left(g^{(\mathcal{D})}(\mathbf{x}) - \bar{g}(\mathbf{x})\right)\ \left(\bar{g}(\mathbf{x}) - f(\mathbf{x})\right)\Big]$$

$$= \mathbb{E}_{\mathcal{D}}\left[\left(g^{(\mathcal{D})}(\mathbf{x}) - \bar{g}(\mathbf{x})\right)^2\right] + \left(\bar{g}(\mathbf{x}) - f(\mathbf{x})\right)^2$$

Bias and variance

gbar acts as an intermediate, loosely considered the best possible (fictitious) hypothesis as it is average over infinite datasets so is created from all of h

$$\mathbb{E}_{\mathcal{D}}\left[\left(g^{(\mathcal{D})}(\mathbf{x}) - f(\mathbf{x})\right)^2\right] = \underbrace{\mathbb{E}_{\mathcal{D}}\left[\left(g^{(\mathcal{D})}(\mathbf{x}) - \bar{g}(\mathbf{x})\right)^2\right]}_{\text{var}(\mathbf{x})} + \underbrace{\left(\bar{g}(\mathbf{x}) - f(\mathbf{x})\right)^2}_{\text{bias}(\mathbf{x})}$$

Therefore, $\mathbb{E}_{\mathcal{D}}\left[E_{\text{out}}(g^{(\mathcal{D})})\right] = \mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_{\mathcal{D}}\left[\left(g^{(\mathcal{D})}(\mathbf{x}) - f(\mathbf{x})\right)^2\right]\right]$

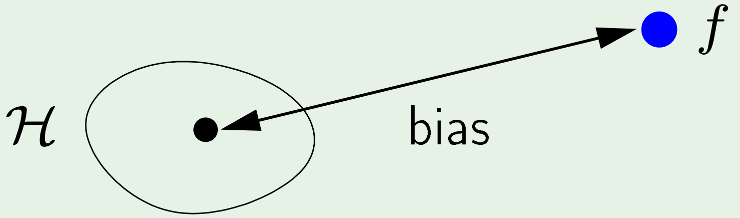
$$= \mathbb{E}_{\mathbf{x}}[\text{bias}(\mathbf{x}) + \text{var}(\mathbf{x})]$$

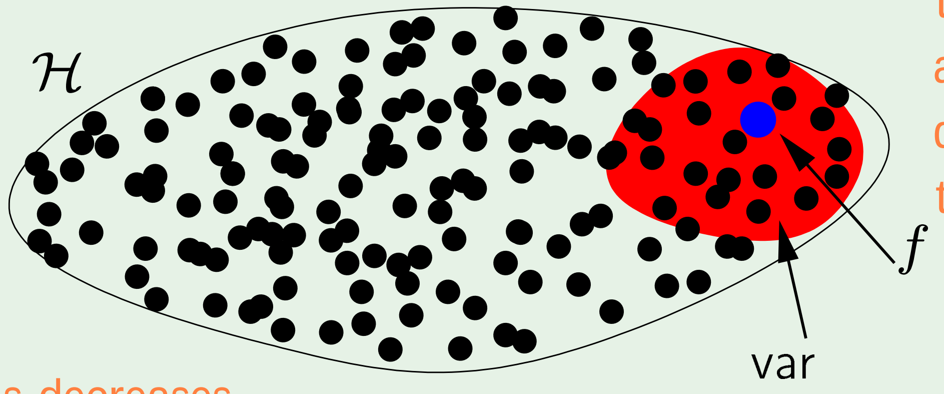
$$= \quad \text{bias} \quad + \quad \text{var}$$

If we do consider gbar the best possible/mean hypothesis, created from averaging over all datasets, the bias can be seen as representing the limitations of the model/hypothesis set itself. From the comment earlier, if we could zoom in perfectly, we would pick the best h (gbar, or close to it). However, due to our finite data set, the first term represents the variance of the final hypothesis we pick from the idealised average (gbar).

The tradeoff

$$\text{bias} = \mathbb{E}_{\mathbf{x}}\left[\left(\bar{g}(\mathbf{x}) - f(\mathbf{x})\right)^2\right] \qquad \text{var} = \mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_{\mathcal{D}}\left[\left(g^{(\mathcal{D})}(\mathbf{x}) - \bar{g}(\mathbf{x})\right)^2\right]\right]$$





red cloud indicates the variety of h that are picked using different datasets - the centroid is gbar

Moving from small H to large H, bias decreases but variance increases, hence a tradeoff

$$\downarrow \qquad \mathcal{H}\uparrow \qquad \uparrow$$

One can also view the variance as a measure of 'instability' in the learning model. Instability manifests in wild reactions to small variations or idiosyncrasies in the data, resulting in vastly different hypotheses - we can see the variation depending on the dataset in the H1 hypothesis set below.

Example: sine target

$f : [-1, 1] \rightarrow \mathbb{R} \qquad f(x) = \sin(\pi x)$

Only two training examples! $\quad N = 2$

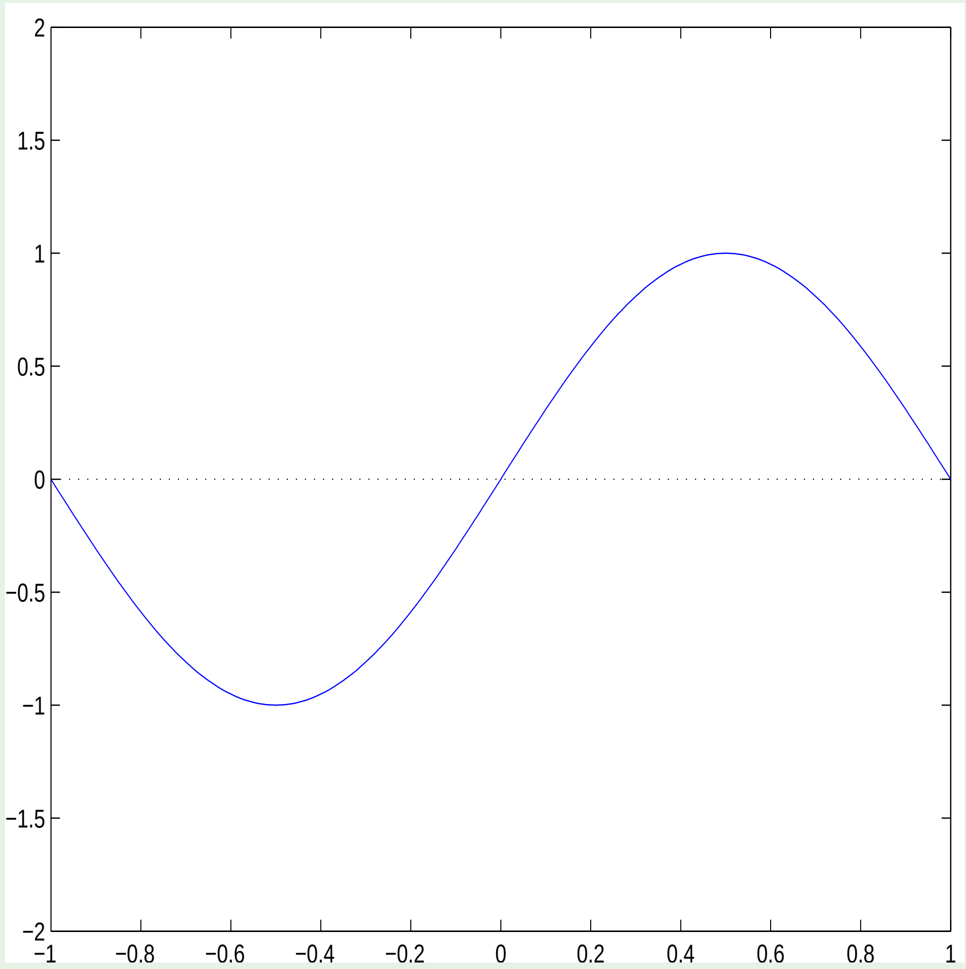
Two models used for learning:

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

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

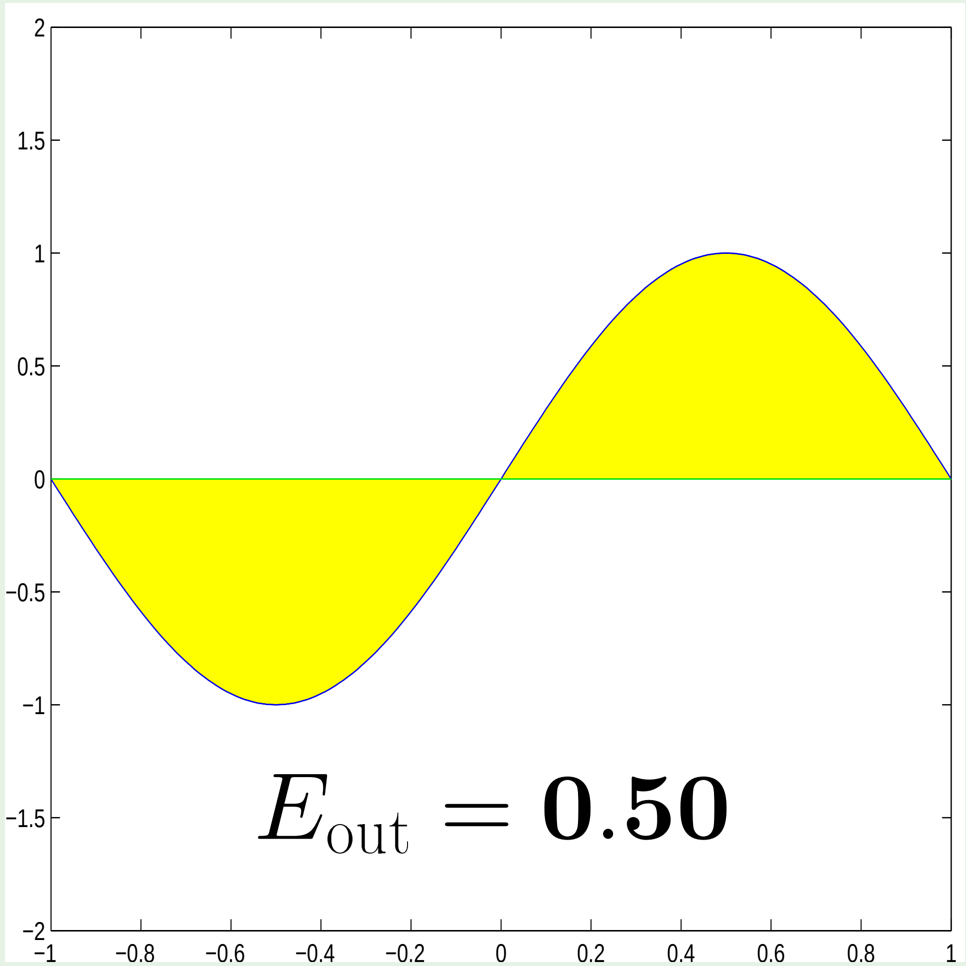
Which is better, $\mathcal{H}_0$ or $\mathcal{H}_1$?

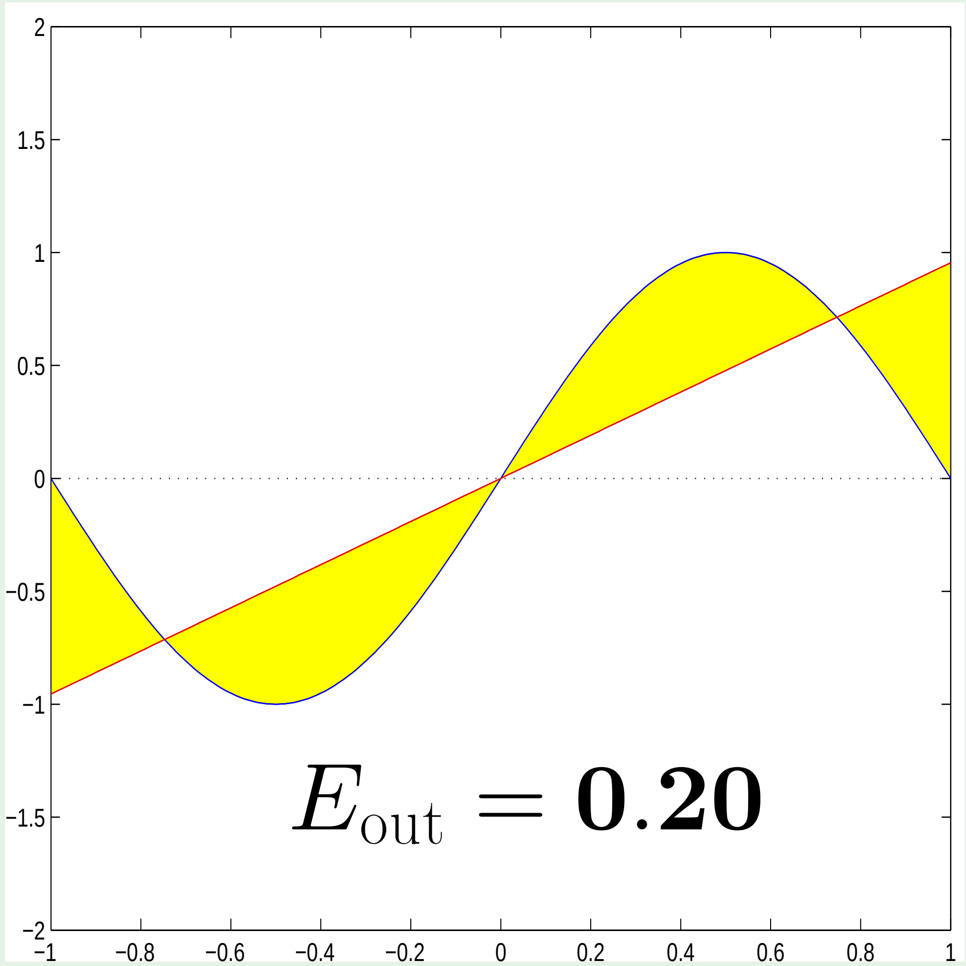
f



Approximation - $\mathcal{H}_0$ versus $\mathcal{H}_1$

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

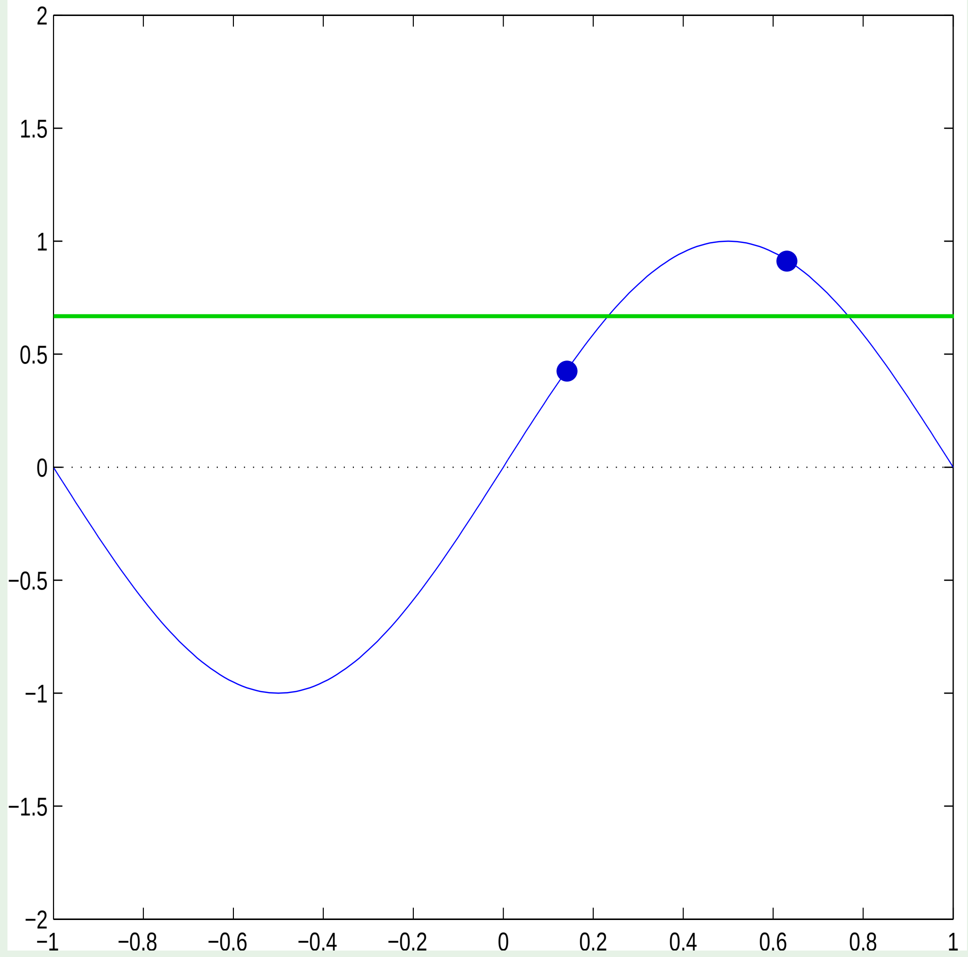




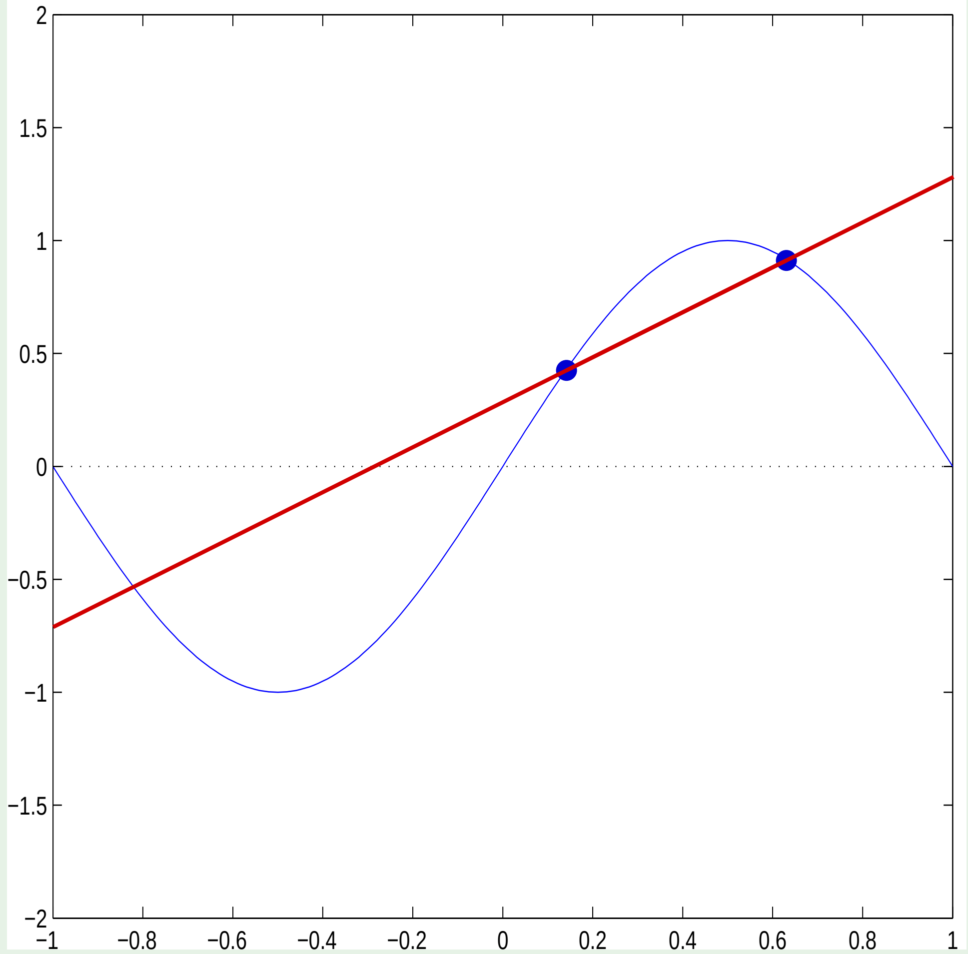
Learning - $\mathcal{H}_0$ versus $\mathcal{H}_1$

The bias-variance analysis uses the expectation of Eout w.r.t. D to grade the model, so we talk about the model learning a target using N=2 regardless of which two points we are talking about (a more general approach).

$\mathcal{H}_0$

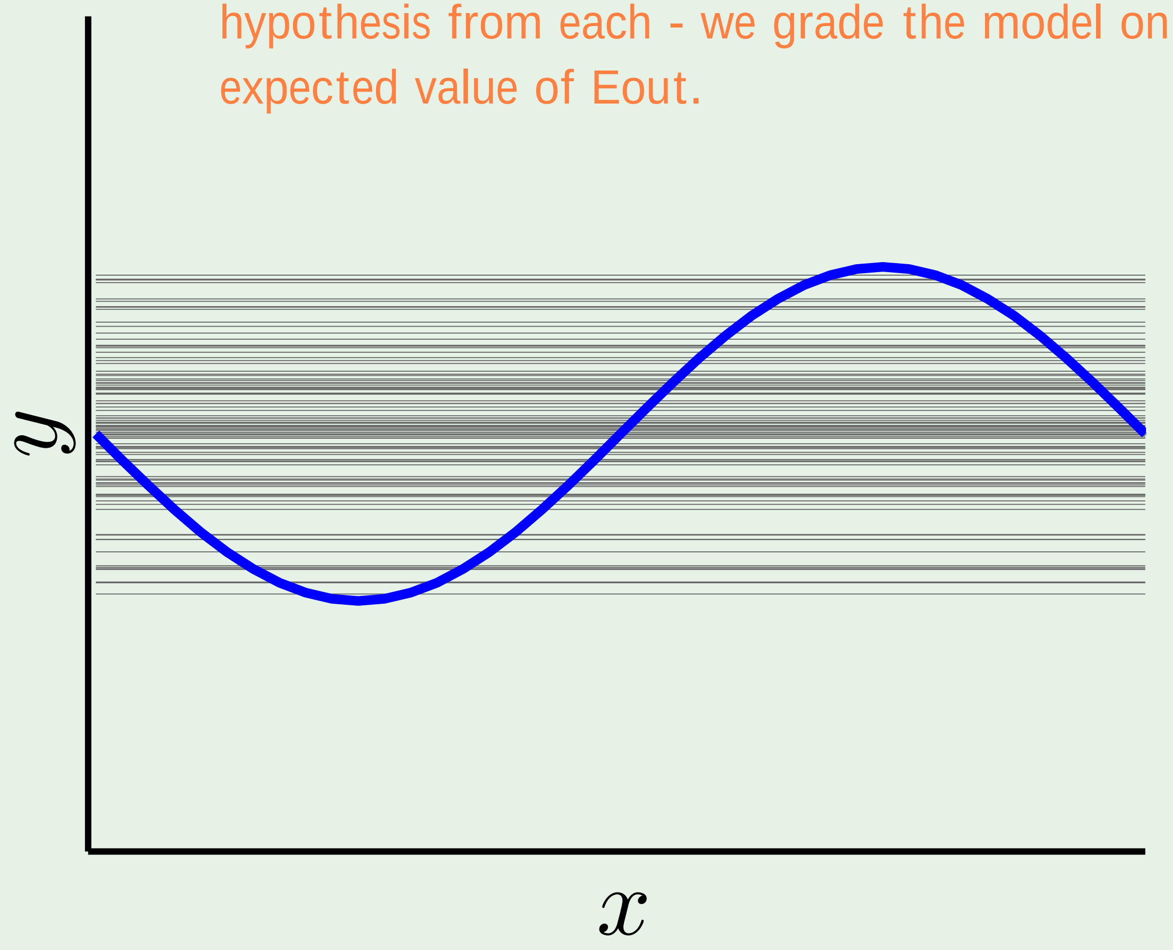


$\mathcal{H}_1$

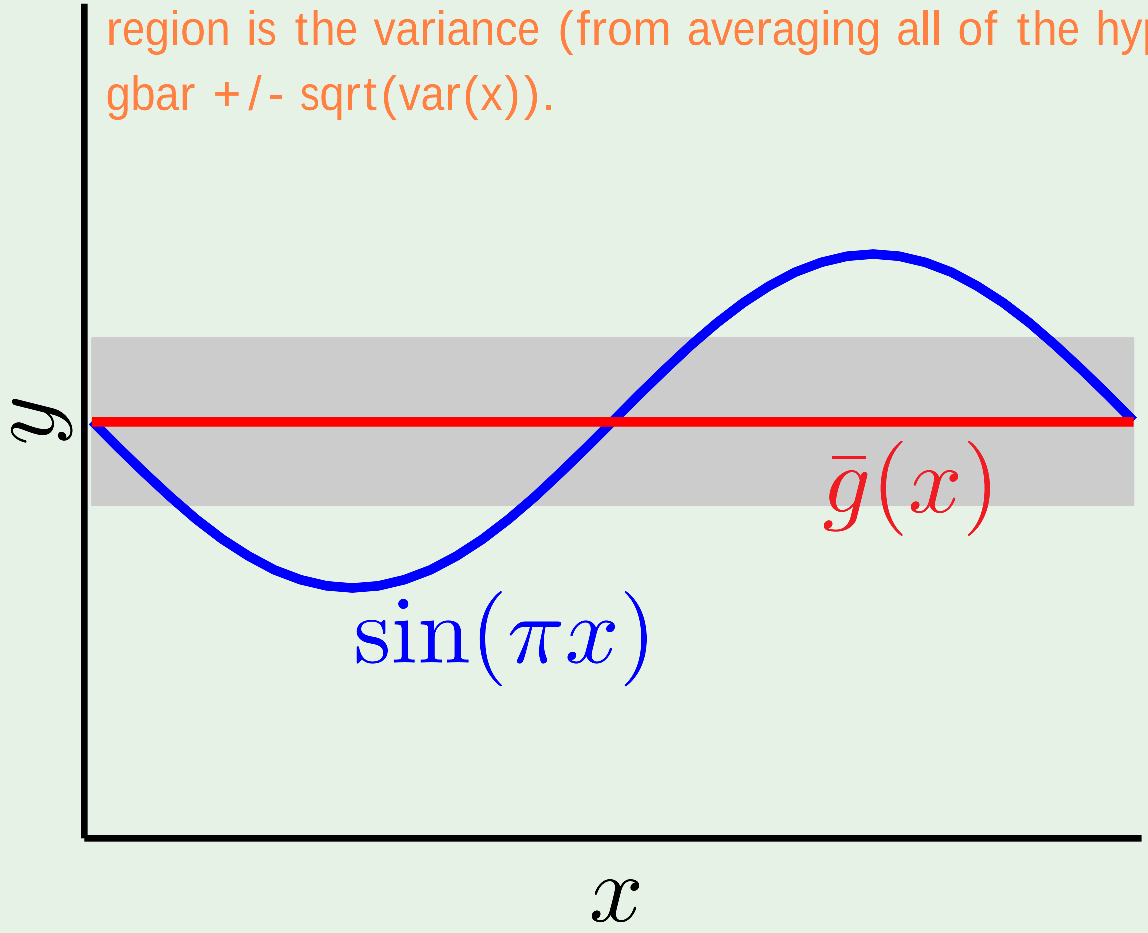


Bias and variance - $\mathcal{H}_0$

From many D of size N=2, fit a line (which will be the midpoint) and repeat for each, showing the hypothesis from each - we grade the model on the expected value of Eout.

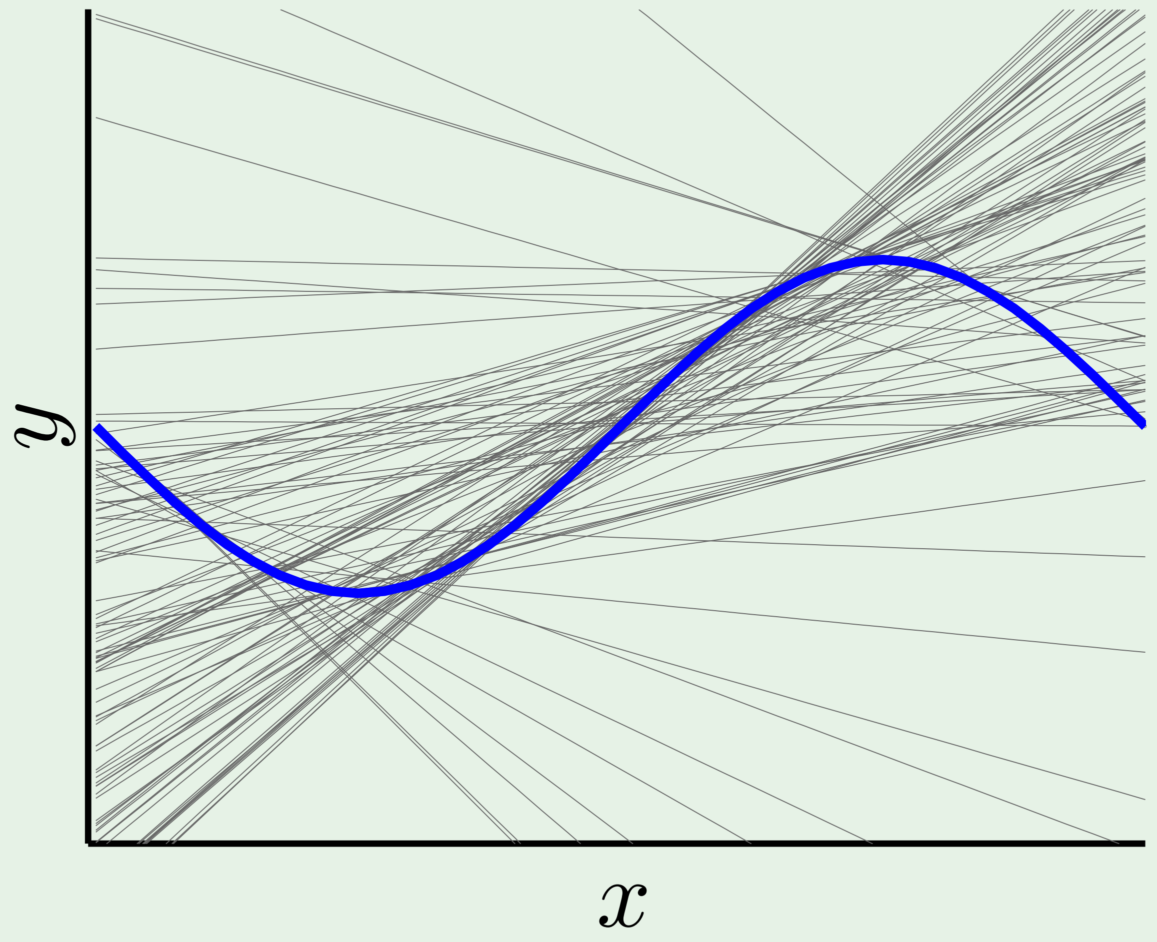


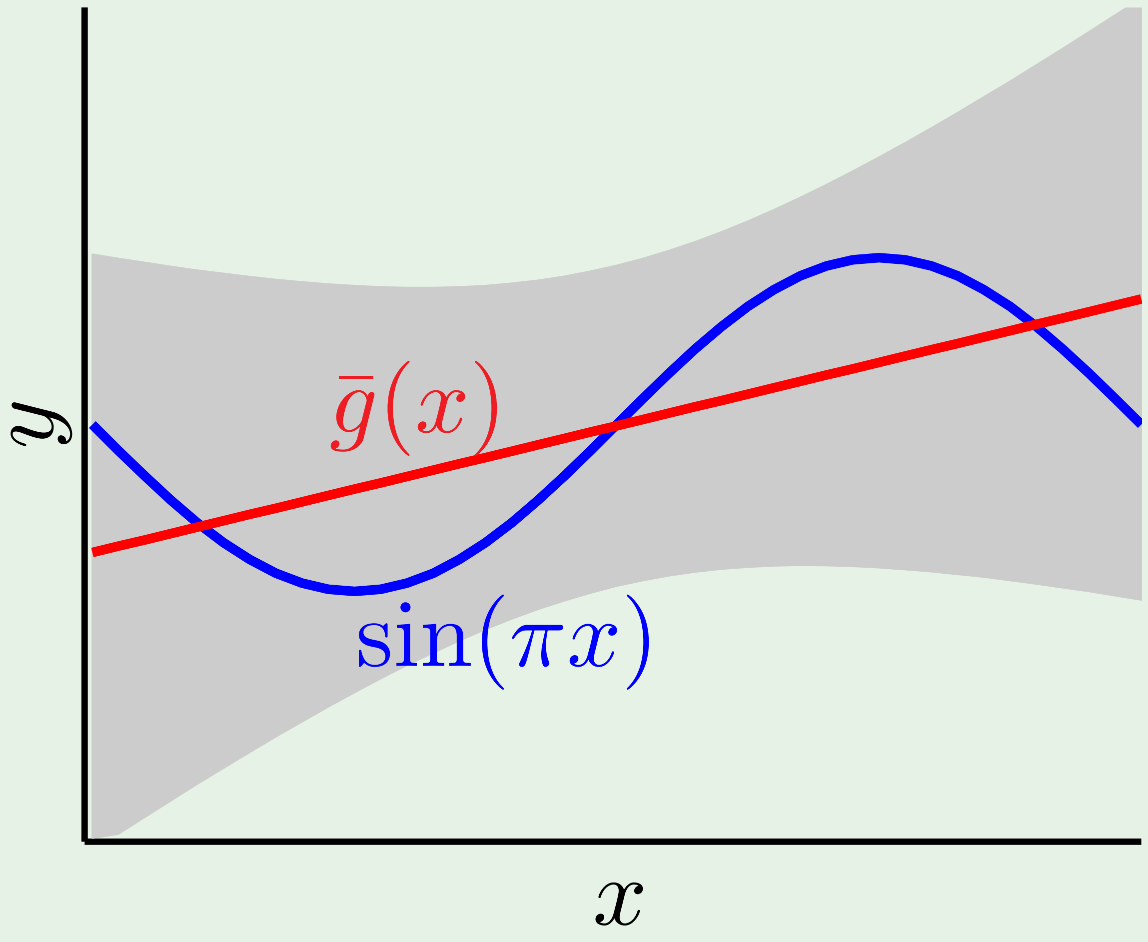
Note we do not use gbar as our final hypothesis, it is just the average of all of the hypotheses of each D. The difference from gbar to the target function is the bias, and the width of the grey region is the variance (from averaging all of the hypotheses): gbar +/- sqrt(var(x)).



Bias and variance - $\mathcal{H}_1$

Much larger average variance (the expectation value of var(x) over the domain) in the hypotheses, but smaller bias

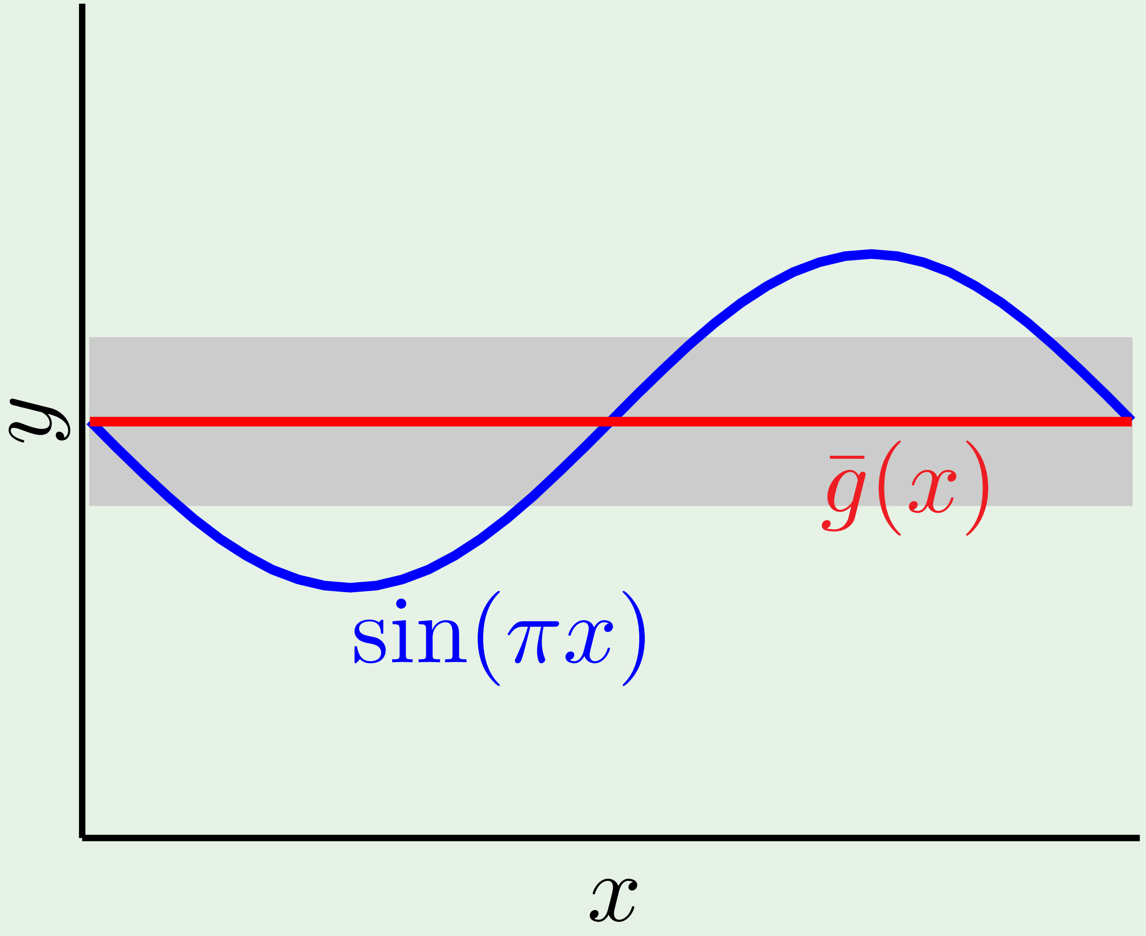




Note we are not asking if a constant or general line is better at approximating a sinusoid, but from a learning scenario: you have two points from a target function we do not know - is it better to use a constant or a line to best learn f?

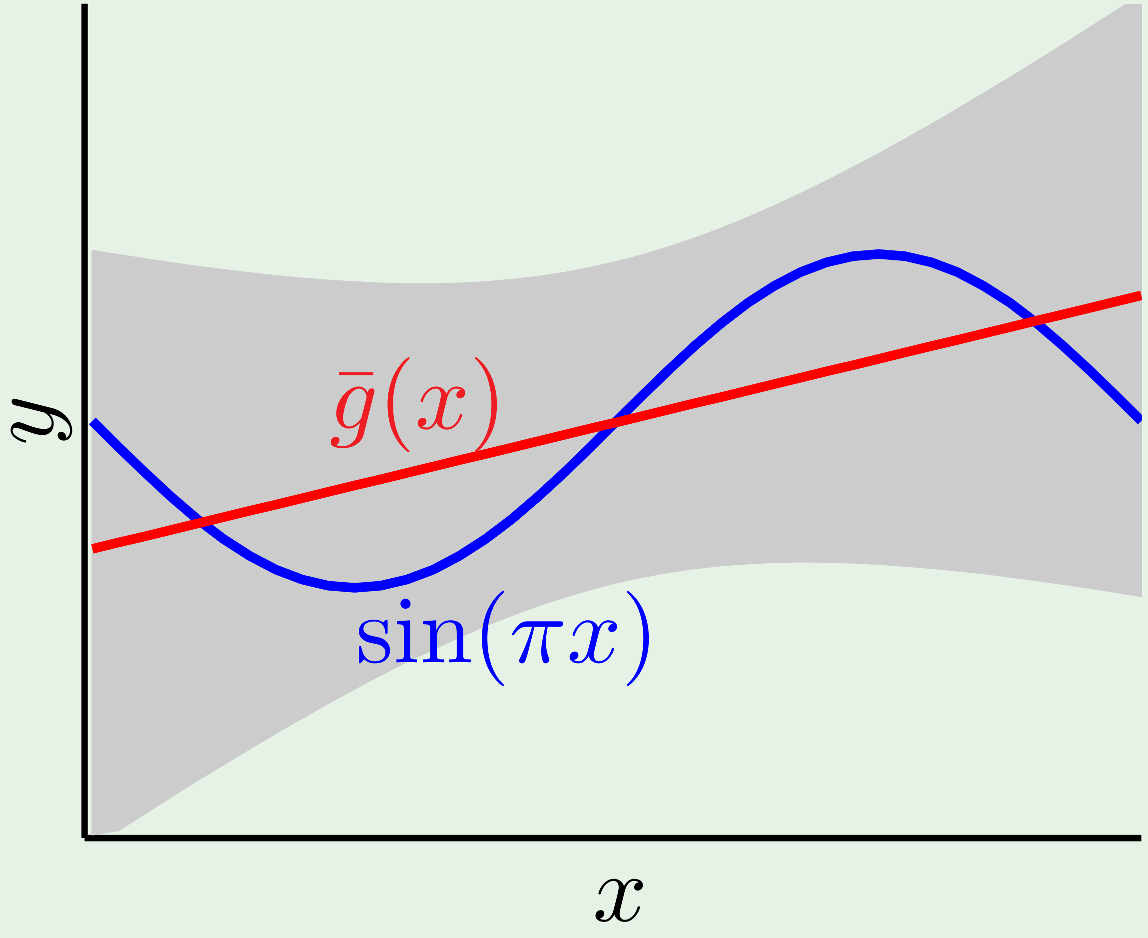
and the winner is ...

$\mathcal{H}_0$



bias $=$ **0.50** var $=$ **0.25**

$\mathcal{H}_1$



bias $=$ **0.21** var $=$ **1.69**

Note the small difference in the H1 bias to Eout in the approximation on slide 11 - so gbar is not exactly the best fit due to the non-linearity of taking two points at a time, making a fit then taking an average (from many trials), so it is concievable that doing this (even for many pairs of points) gives us a different result to having the target function and fitting it outright.

Lesson learned

Match the '**model complexity**'

to the **data resources**, not to the **target complexity**

what do we have to navigate H (how much data, how noisy is it) - from this pick a H we can afford to navigate

We do not know the target, and even if we knew the level of complexity that it has, we do not have the resources to match it, since if we match it, we will have the target in H, but we will never arrive at it/finding it is very unlikely.

Outline

- Bias and Variance

- Learning Curves

Expected E_{out} and E_{in}

Data set $\mathcal{D}$ of size N

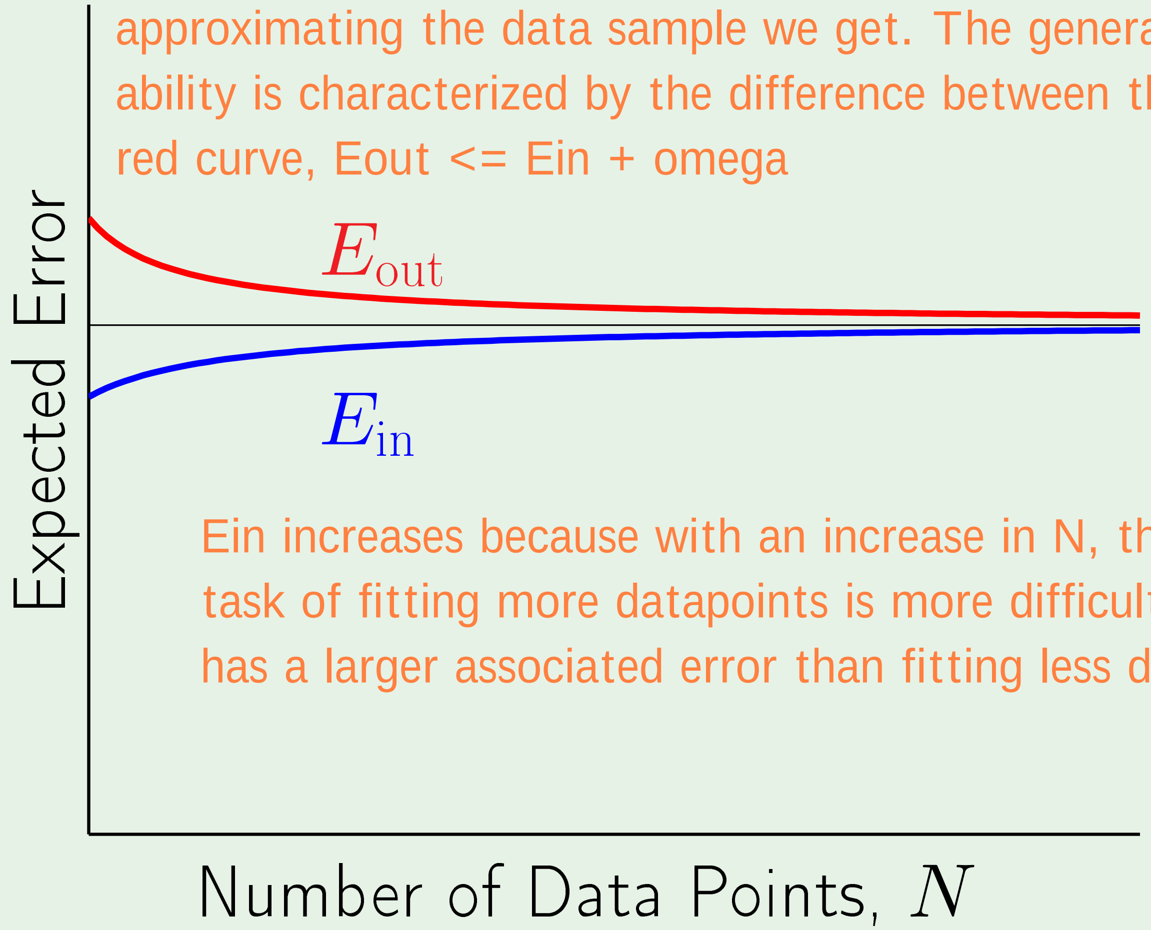
Expected out-of-sample error $\mathbb{E}_{\mathcal{D}}[E_{\text{out}}(g^{(\mathcal{D})})]$ general quantity which only relies on N

Expected in-sample error $\mathbb{E}_{\mathcal{D}}[E_{\text{in}}(g^{(\mathcal{D})})]$

How do they vary with N?

The curves

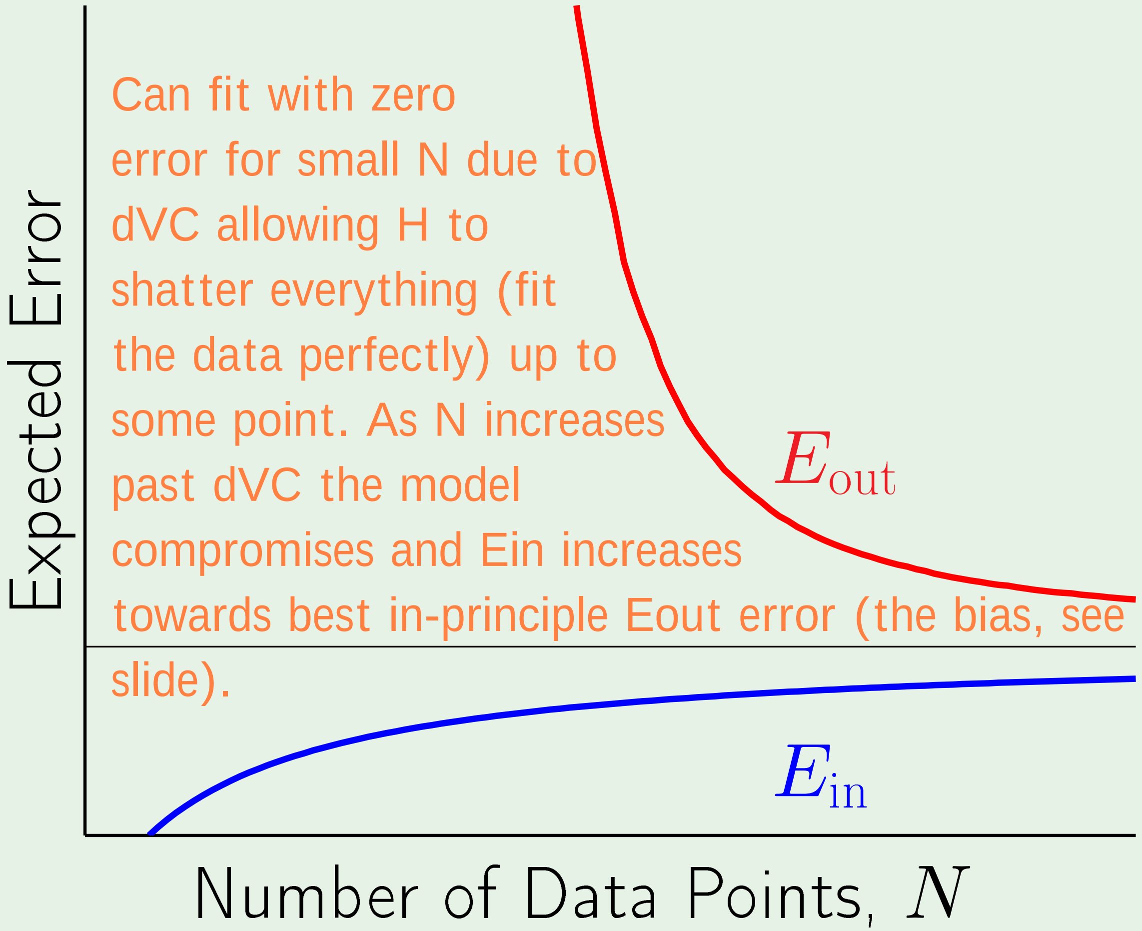
Black horizontal line gives error of approximation, note that it decreases for the more complex model. Blue curve is the error of approximating the data sample we get. The generalization ability is characterized by the difference between the blue and red curve, Eout <= Ein + omega



Ein increases because with an increase in N, the task of fitting more datapoints is more difficult and has a larger associated error than fitting less data.

Simple Model

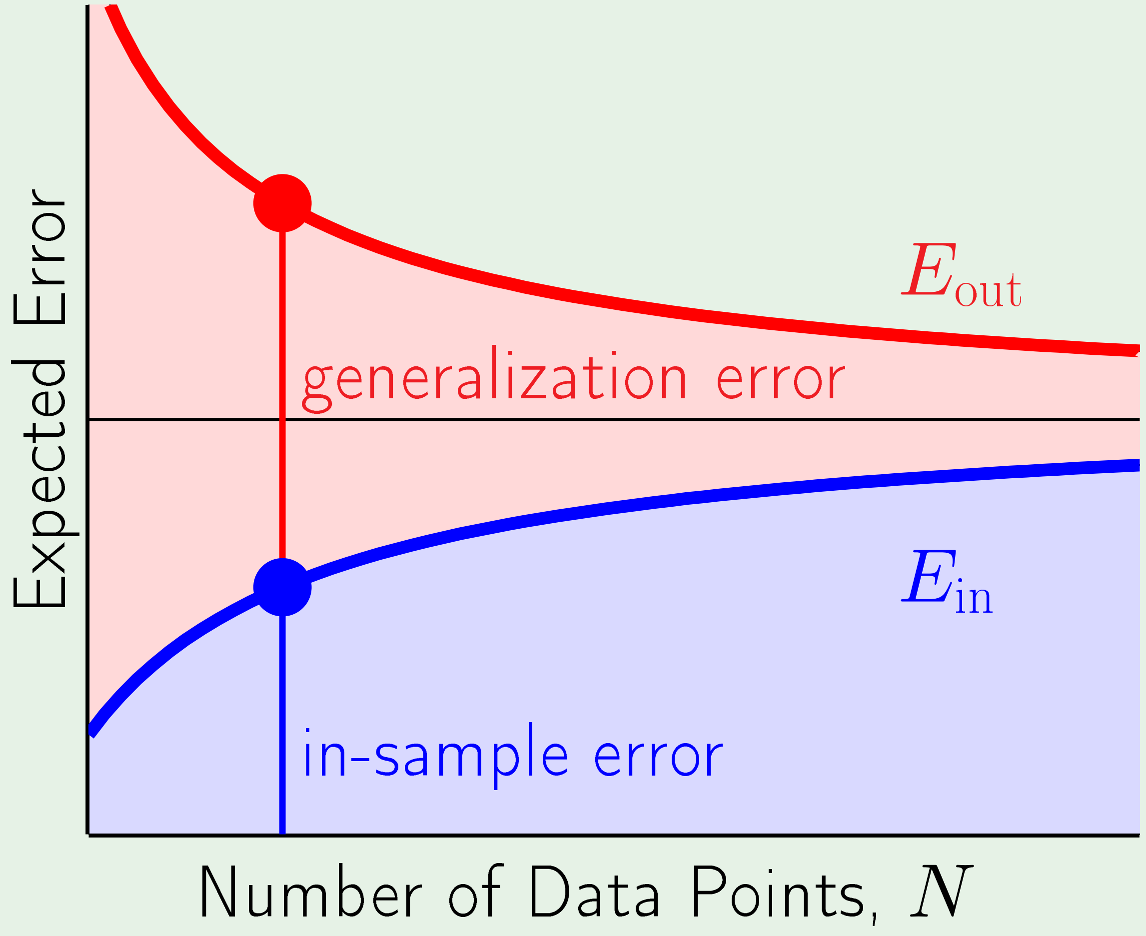
Can fit with zero error for small N due to dVC allowing H to shatter everything (fit the data perfectly) up to some point. As N increases past dVC the model compromises and Ein increases towards best in-principle Eout error (the bias, see next slide).



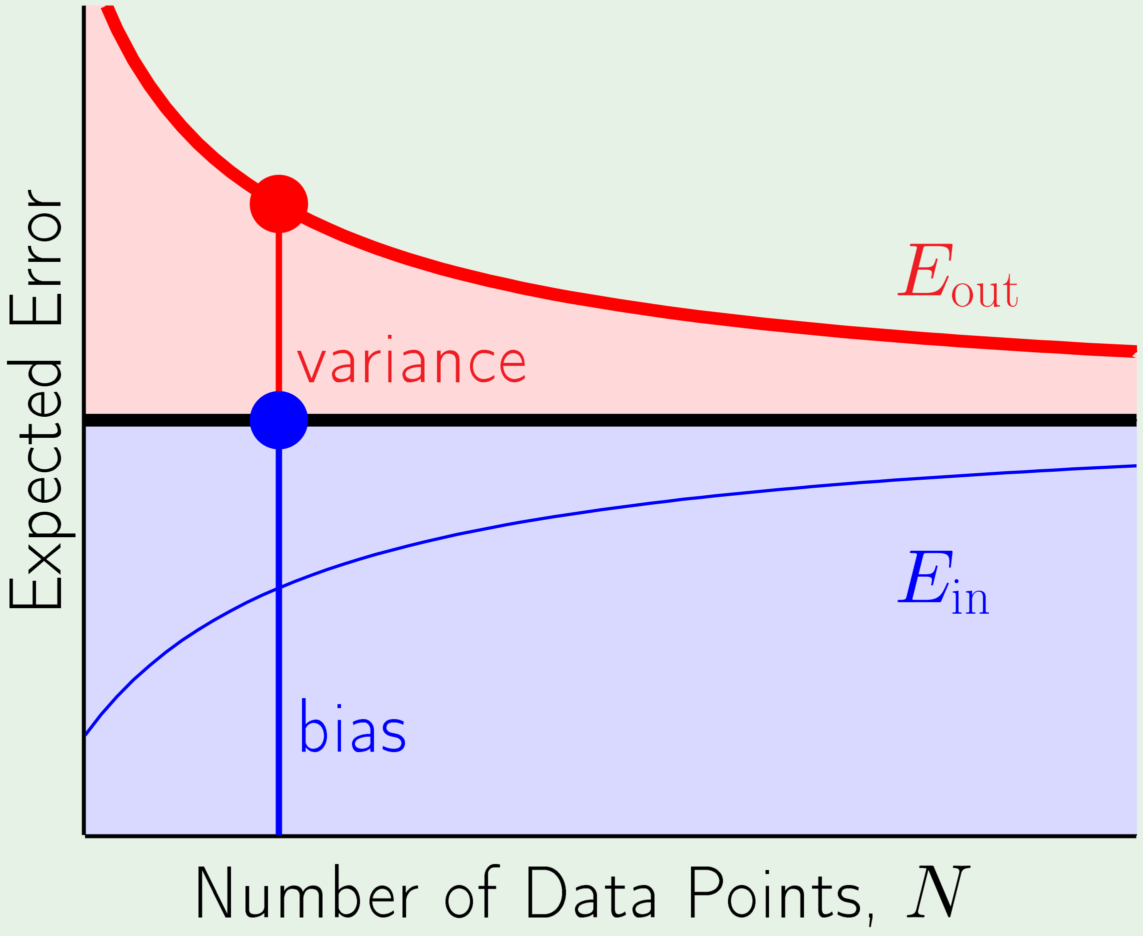
Complex Model

Large Eout for small N as the model simply memorises the data and has not learnt anything. The difference/generalization error between Ein and Eout is larger due to the more complex model (both the bound and actual value is larger)

VC versus bias-variance



VC analysis



bias-variance

Linear regression case

linear + noise

Noisy target $y = \mathbf{w}^{*\intercal}\mathbf{x} + \text{noise}$

Data set $\mathcal{D} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$

Linear regression solution: $\mathbf{w} = (\mathrm{X}^\intercal\mathrm{X})^{-1}\mathrm{X}^\intercal\mathbf{y}$

In-sample error vector $= \mathrm{X}\mathbf{w} - \mathbf{y}$

'Out-of-sample' error vector $= \mathrm{X}\mathbf{w} - \mathbf{y}'$

Xw compared to y', which is the same as y but with a different realisation of the noise

Learning curves for linear regression

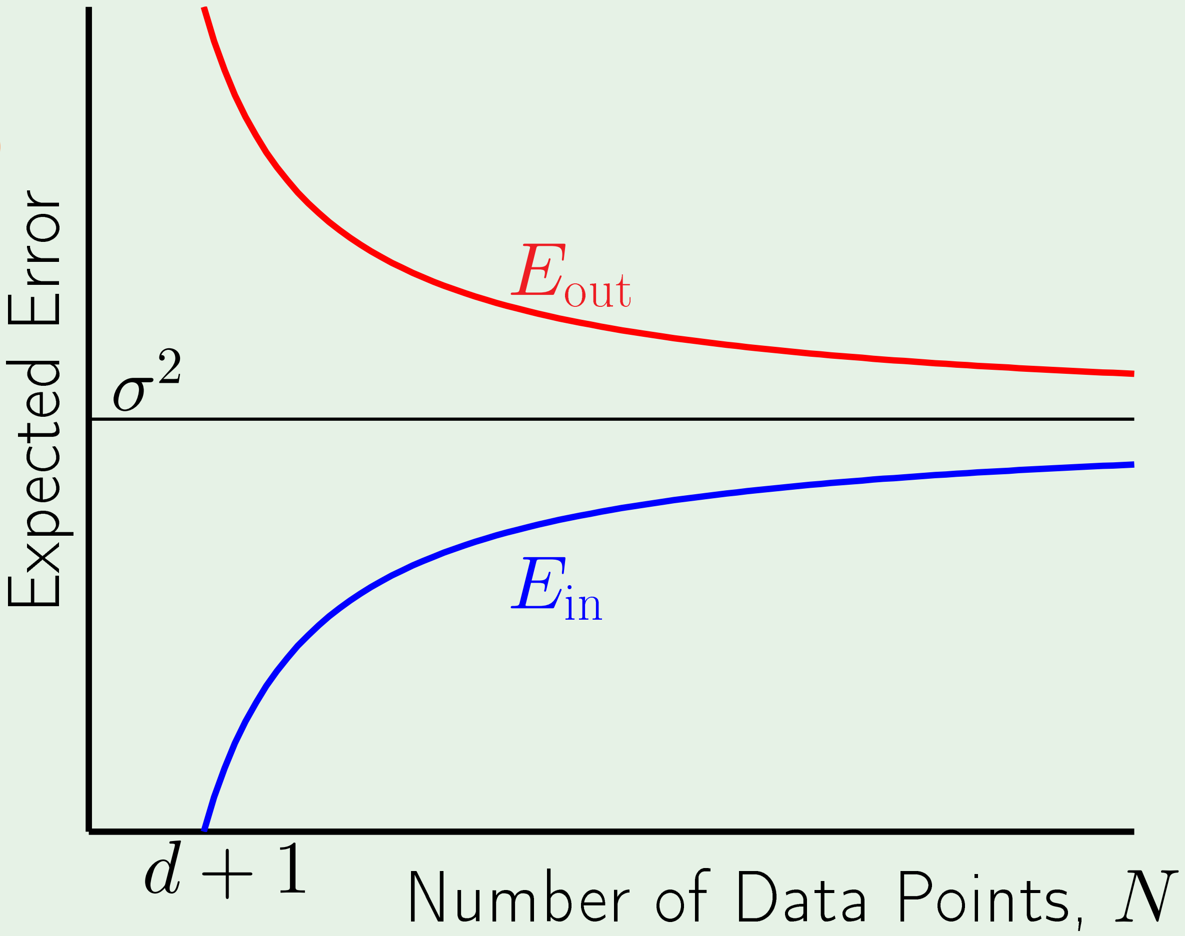
d+1 is a sort of dVC (it is equal to it for perceptron)
but charactersises the degrees of freedom of the model

Best approximation error $= \sigma^2$ = variance of noise (inevitable error due to addition of noise)

Expected in-sample error $= \sigma^2 \left(1 - \frac{d+1}{N}\right)$

Expected out-of-sample error $= \sigma^2 \left(1 + \frac{d+1}{N}\right)$

Expected generalization error $= 2\sigma^2 \left(\frac{d+1}{N}\right)$



Exact generalization error, form of VC dimension / N.
Shows the compromise between the number of d.o.f
(in the case of linear regression) and the size of the dataset.

more examples means that we fit less of the noise since the linear pattern persists and the noise starts to be canceled out in the fitting (not enough d.o.f to fit it) so it becomes as if we are fitting perfectly