Lab 4: Bias-Variance Trade-Off

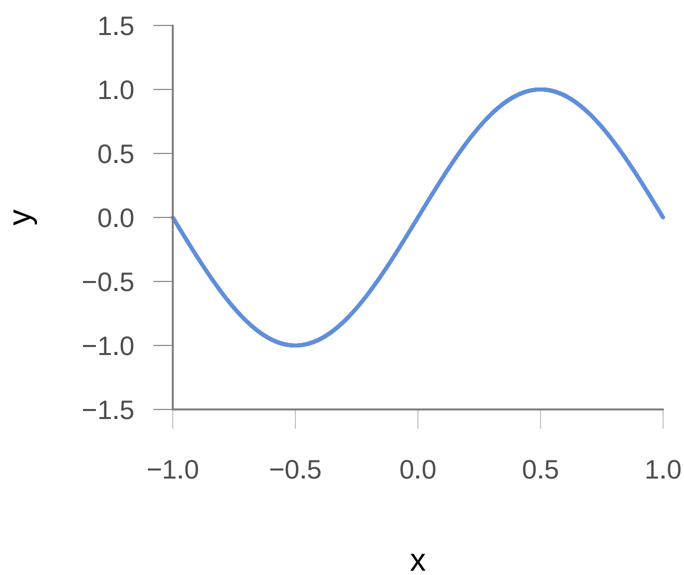
Stat 154, Fall 2019

In this lab assignment you will have the opportunity to further review the case-study discussed in lecture to illustrate the Bias-Variance decomposition. This should allow to have a better grasp of such decomposition:

- understand the notion of **bias**: "How close the (average) functional form of a particular class of models, $\bar{g}(x)$, can get to the *true* target function $f(x)$". Or simply put: *how well our class of $g(x)$ models, represented by $\bar{g}(x)$, can approximate $f(x)$.*
- understand the notion of **variance**: "How variable (or precise) a particular model $g(x)$ is around its average $\bar{g}(x)$". Or simply put: *how close a particular $g(x)$ can get to the average $\bar{g}(x)$.*

Target Function

Consider a noiseless target function $f(x) = sin(\pi x)$, with the input variable x in the interval $[-1, 1]$, like in the following picture:



In R, you can get a similar plot to the one above (although not identical) with the following code:

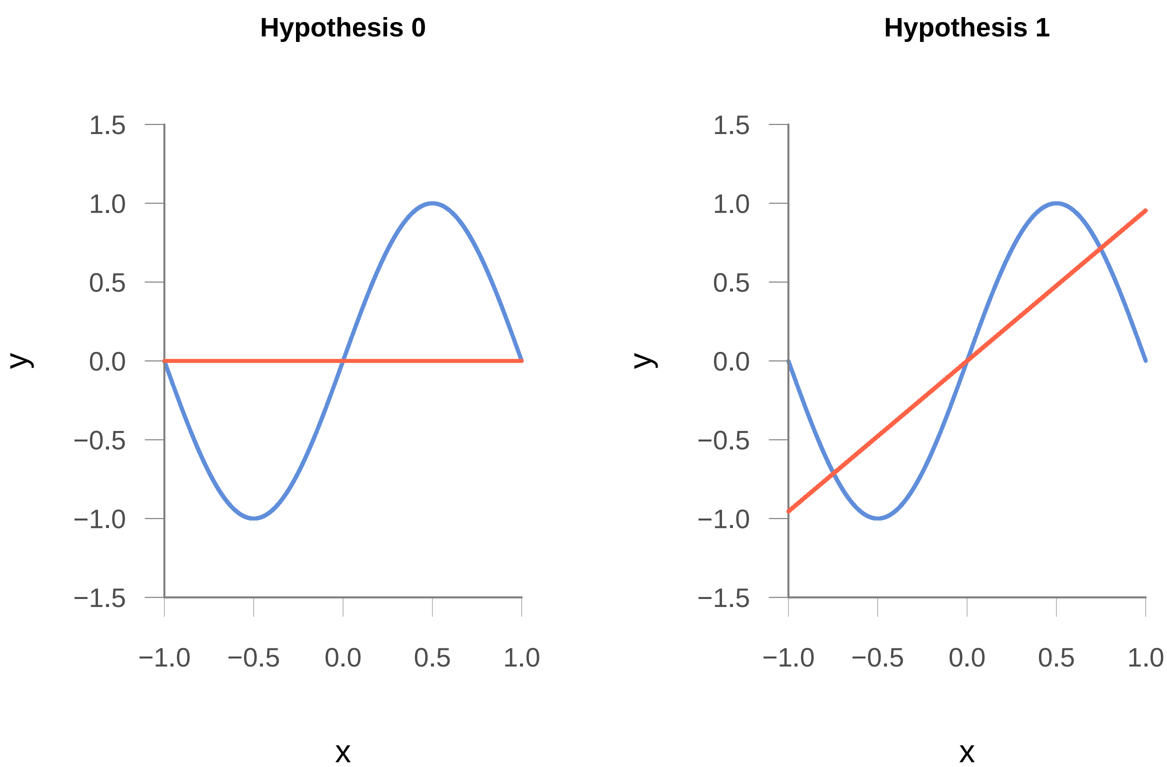
```r
# (figure above NOT produced with this command)
x <- seq(-1, 1, by = 0.0001)
y <- sin(x * pi)

plot(x, y, type = "l", lwd = 2,  col = "#608EDB", las = 1,
     xlim = c(-1, 1), ylim = c(-1.5, 1.5))
```

Two Hypotheses

Let's assume a learning scenario in which, given a data set of n points, we fit the data using one of two models (see the *idealized* figure shown below):

- $\mathcal{H}_0$: Set of all lines of the form $h(x) = b$
- $\mathcal{H}_1$: Set of all lines of the form $h(x) = b_0 + b_1 x$



Learning from two points

In this case study, we will assume a data set of size $n = 2$. That is, we sample x uniformly in $[-1, 1]$ to generate a data set of two points $(x_1, y_1), (x_2, y_2)$; and fit the data using the two models $\mathcal{H}_0$ and $\mathcal{H}_1$.

For $\mathcal{H}_0$, we choose the constant hypothesis that best fits the data (the horizontal line at the midpoint $b = \frac{y_1 + y_2}{2}$).

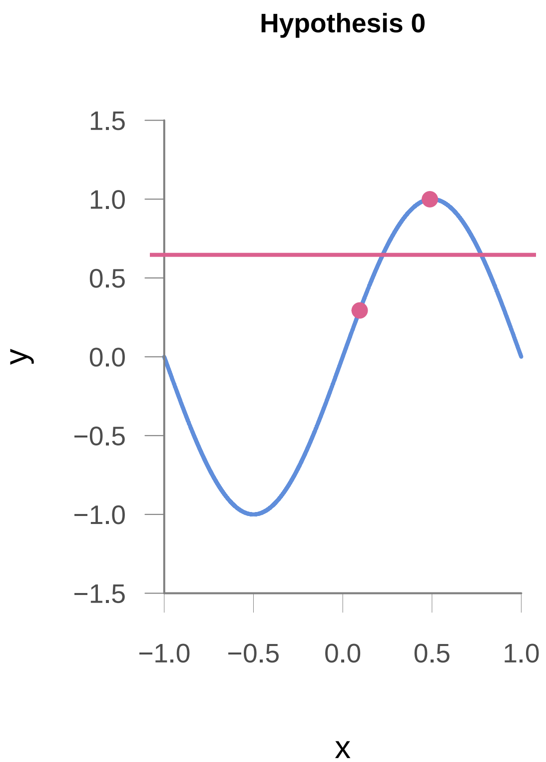
For $\mathcal{H}_1$, we choose the line that passes through the two data points (x_1, y_1) and (x_2, y_2).

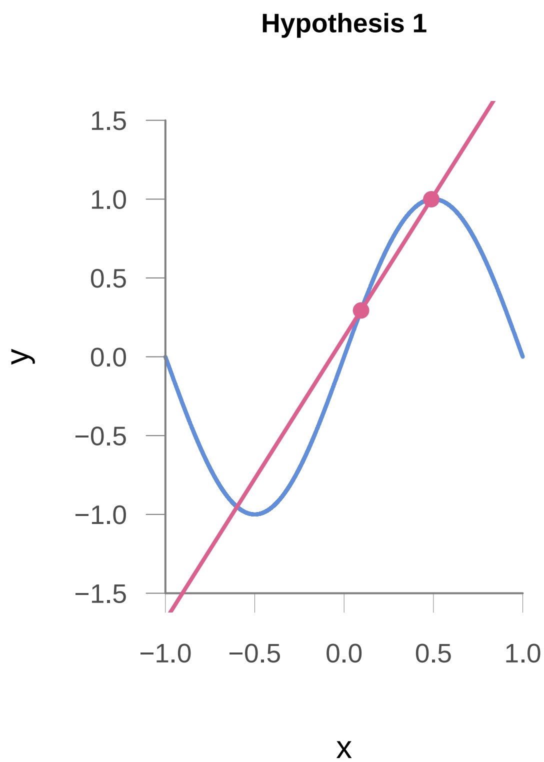
Here's an example in R of two x-points randomly sampled from a uniform distribution in the interval $[-1, 1]$, and their corresponding y-points:

```
# two in-sample data-points (i.e. training set)
set.seed(553)
x_pts <- runif(n = 2, min = -1, max = 1)
y_pts <- sin(pi * x_pts)
```

With the given points above, the two fitted models are:

- $h_0(x) = 0.646544$
- $h_1(x) = 0.123472 + 1.794385\,x$





Bias-Variance Decomposition

Recall that the bias-variance decomposition is a theoretical device of the **expected out-of-sample MSE**, that is, the *average* squared error that we would obtain if we repeatedly estimated f using a large number of training sets $\mathcal{D}$, and tested each $g^{(\mathcal{D})}$ at a single out-of-sample point x_0:

$$\mathbb{E}_{\mathcal{D}} \left[\left(g^{(\mathcal{D})}(x_0) - f(x_0) \right)^2 \right] = \text{var}(g) + \text{bias}^2(g)$$

As we described in lecture, the decomposition becomes:

$$\mathbb{E}_{\mathcal{D}} \left[\left(g^{(\mathcal{D})}(x_0) - f(x_0) \right)^2 \right] = \underbrace{\mathbb{E}_{\mathcal{D}} \left[\left(g^{(\mathcal{D})}(x_0) - \bar{g}(x_0) \right)^2 \right]}_{\text{variance}} + \underbrace{\mathbb{E}_{\mathcal{D}} \left[(\bar{g}(x_0) - f(x_0))^2 \right]}_{\text{bias}^2}$$

where the (noiseless) target function is represented by $f()$, and the average hypothesis is represented by $\bar{g}()$.

Notice that the above equation assumes that the squared error corresponds to just one out-of-sample (i.e. test) point $(x_0, y_0) = (x_0, f(x_0))$.

So, in order to get the **overall expected out-of-sample MSE**, we need to take the expectation over all possible test points $x_0 \in \mathcal{X}$, that is:

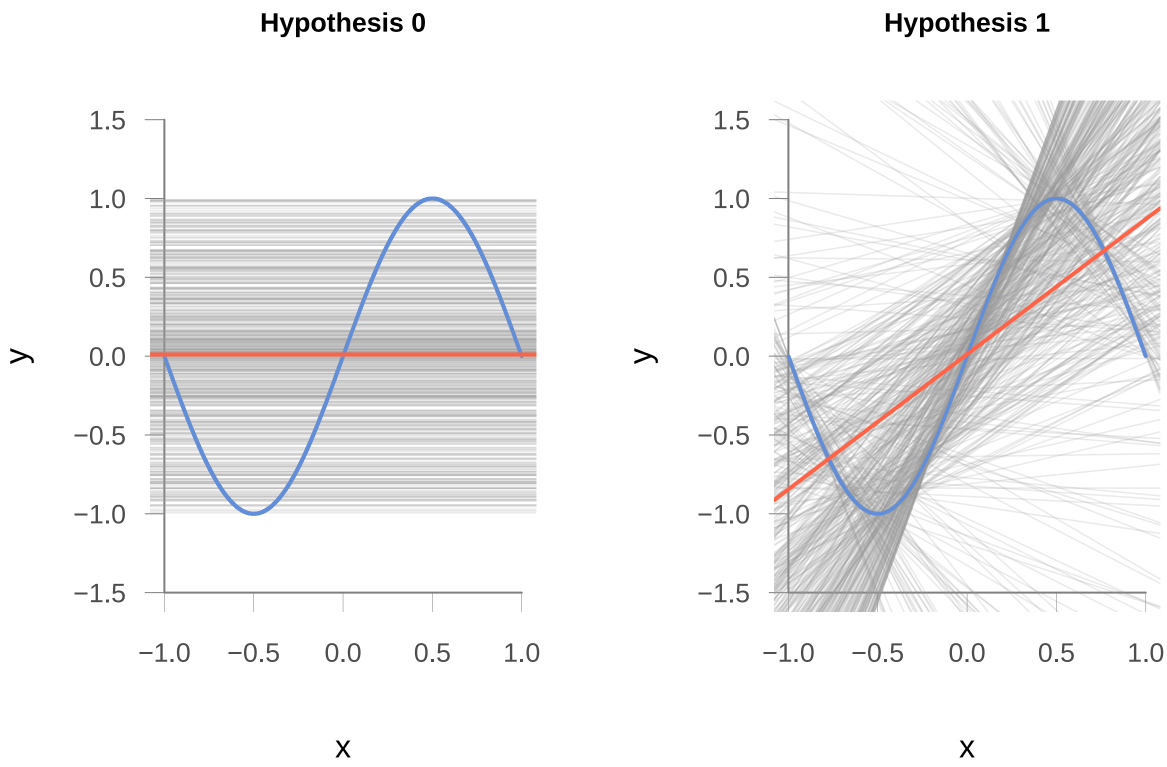
$$\mathbb{E}_{\mathcal{X}} \left\{ \mathbb{E}_{\mathcal{D}} \left[\left(g^{(\mathcal{D})}(x_0) - f(x_0) \right)^2 \right] \right\} = \underbrace{\mathbb{E}_{\mathcal{X}}[\text{var}(g(x))]}_{\text{variance}} + \underbrace{\mathbb{E}_{\mathcal{X}}[\text{bias}^2(g(x))]}_{\text{bias}^2}$$

This *overall expected out-of-sample MSE* is the ultimate theoretical measure of model performance (out-of-sample). But of course, in practice you will never be able to compute it—that's why we are doing the simulation in this lab.

Simulation

Your mission is to carry out a simulation of the sampling process described above (e.g. repeat it at least 500 times): randomly sample two x points in the interval $[-1, 1]$, and fit both models $\mathcal{H}_0$ and $\mathcal{H}_1$.

The ultimate goal is to estimate the bias and the variance under each hypothesis. The figures which follow show the resulting 500 fits on the same (random) data sets for both methods. Notice the display of the average hypotheses $\bar{g}_0$ and $\bar{g}_1$ (colored in red), in each case.



To carry out the simulation, use a simulated *out-of-sample set* of 20000 points:

```
# simulated out-of-sample data points
set.seed(12345)
n_out <- 20000
x_out <- runif(n_out, min = -1, max = 1)
y_out <- sin(pi * x_out)
```

Also, to compute expectations, calculate them with the averages (i.e. means) of your 500 repetitions.

In case you are curious, for the simulated data that produced the fits displayed above, the corresponding bias and variance are (yours will very likely be a bit different):

$\text{bias}^2(g_0) \approx 0.4988$, and $\text{var}(g_0) \approx 0.2502$, -vs- overall expected out-of-sample 0.74904

$\text{bias}^2(g_1) \approx 0.19934$, and $\text{var}(g_1) \approx 1.34875$, -vs- overall expected out-of-sample 1.54202

Include plots for both types of hypothesis: like in the figure above.