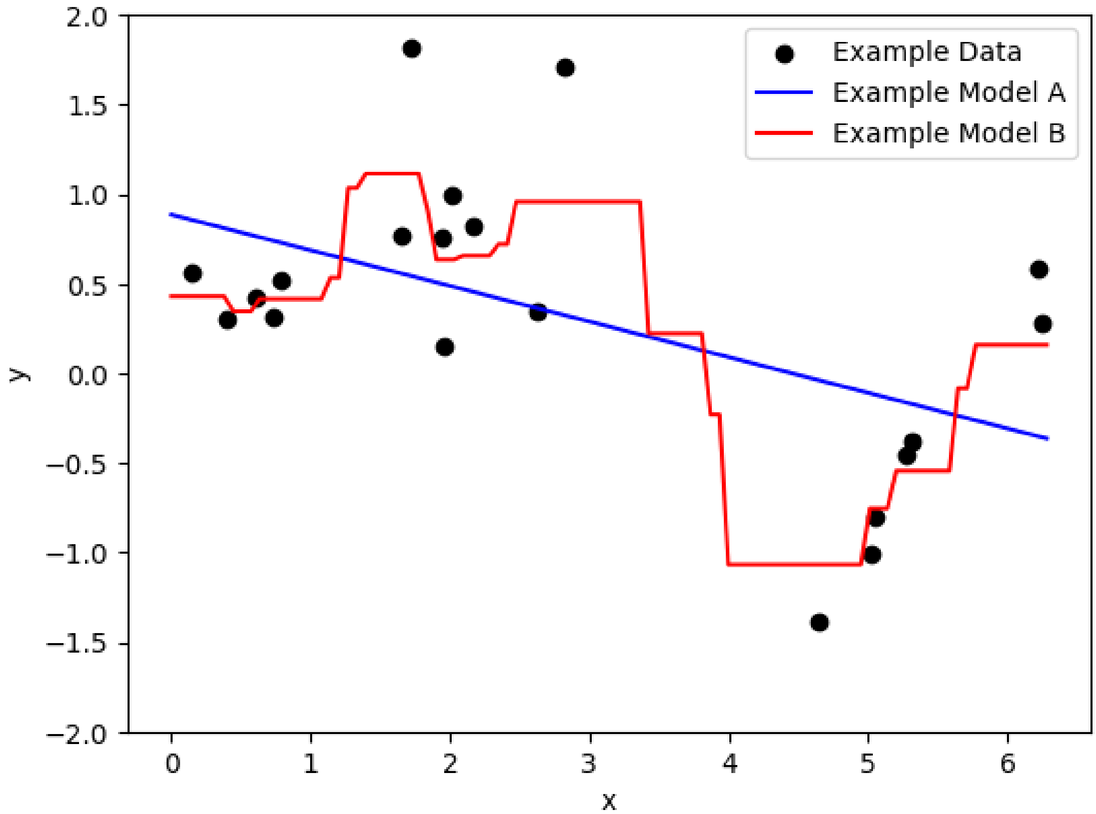
M10-L1 Problem 1

In this problem you will look compare models with lower/higher variance/bias by computing bias and variance at a single point.

```
import numpy as np
import matplotlib.pyplot as plt
from sklearn.linear_model import LinearRegression
from sklearn.neighbors import KNeighborsRegressor

def plot_model(model,color="blue"):
    x = np.linspace(0, np.pi*2, 100)
    y = model.predict(x.reshape(-1,1))
    plt.plot(x, y, color=color)
    plt.xlabel("x")
    plt.ylabel("y")

def plot_data(x, y):
    plt.scatter(x,y,color="black")

def eval_model_at_point(model, x):
    return model.predict(np.array([[x]])).item()

def train_models():
    x = np.random.uniform(0,np.pi*2,20).reshape(-1,1)
    y = np.random.normal(np.sin(x),0.5).flatten()

    modelA = LinearRegression()
    modelB = KNeighborsRegressor(3)
    modelA.fit(x,y)
    modelB.fit(x,y)
    return modelA, modelB, x, y
```

The function `train_models` gets 20 new data points and trains two models on these data points. Model A is a linear regression model, while model B is a 3-nearest neighbor regressor.

```
modelA, modelB, x, y = train_models()
plt.figure()
plot_data(x,y)
plot_model(modelA,"blue")
plot_model(modelB,"red")
plt.legend(["Example Data", "Example Model A", "Example Model B"])
plt.ylim([-2,2])
plt.show()
```



Training models

First, train 50 instances of model A and 50 instances of model B. Store all 100 total models for use in the next few cells. Generate these models with the function: `modelA, modelB, x, y = train_models()`.

```
# YOUR CODE GOES HERE
# train 50 instances of modelA and modelB

modelA_list = []
modelB_list = []

for i in range(50):
    modelA, modelB, x, y = train_models()
    modelA_list.append(modelA)
    modelB_list.append(modelB)
```

Bias and Variance

Now we will use the definitions of bias and variance to compute the bias and variance of each type of model. You will focus on the point x = 1.57 only. First, compute the prediction for each model at x. (You can use the function `eval_model_at_point(model, x)`).

```
x = 1.57

# YOUR CODE GOES HERE
# evaluate the models at x = 1.57
modelA_values = []
modelB_values = []

for i in range(50):
    modelA_values.append(eval_model_at_point(modelA_list[i], x))
    modelB_values.append(eval_model_at_point(modelB_list[i], x))
```

In this cell, use the values you computed above to compute and print the bias and variance of model A at the point x = 1.57. The true function value `y_GT` is given as 1 for x=1.57.

```
yGT = 1

# YOUR CODE GOES HERE
# calculate the bias and variance of modelA and modelB
bias_A = np.mean(np.array(modelA_values)) - yGT
bias_B = np.mean(np.array(modelB_values)) - yGT

var_A = np.var(np.array(modelA_values))
var_B = np.var(np.array(modelB_values))

print(f"Model A:   Bias = {bias_A:.3f},  Variance = {var_A:.3f}")
print(f"Model B:   Bias = {bias_B:.3f},  Variance = {var_B:.3f}")

Model A:   Bias = -0.524,  Variance = 0.035
Model B:   Bias = -0.076,  Variance = 0.096
```

Questions

1. Which model has smaller bias at $x = 1.57$?

   At $x = 1.57$, Model B has a smaller bias at -0.076.

2. Which model has lower variance at $x = 1.57$?

   At $x = 1.57$, Model A has a smaller variance at 0.035.

Plotting models

Now use the `plot_model` function to overlay all Model A predictions on one plot and all Model B predictions on another. Notice the spread of each model.

```
plt.figure(figsize=(9,3))

plt.subplot(1,2,1)
plt.title("Model A")
# YOUR CODE GOES HERE
```

```python
for i in range(50):
    plot_model(modelA_list[i], "blue")


plt.subplot(1,2,2)
plt.title("Model B")
# YOUR CODE GOES HERE
for i in range(50):
    plot_model(modelB_list[i], "red")

plt.show()
```