Linear Regression: SGD solution *(find and work with a partner)*

In linear regression, we seek to minimize the sum of squared errors between the actual response and our prediction. We often call this RSS (residual sum of squares) or SSE (sum of squared errors). As an objective function, we often call it J and include a $\frac{1}{2}$ in front to make the derivatives work out nicely.

$$J(\boldsymbol{w}) = \frac{1}{2}\sum_{i=1}^{n}\left(y_i - h_{\boldsymbol{w}}(\boldsymbol{x_i})\right)^2$$

For linear regression in general, one iteration of stochastic gradient descent includes the following updates (usually with the data points shuffled):

```
for i = 1,2,...,n:
```
$\quad \boldsymbol{w} \leftarrow \boldsymbol{w} - \alpha(\boldsymbol{w}\cdot\boldsymbol{x}_i - y_i)\boldsymbol{x}_i$

We will begin with our same data from the previous two handouts: $(x_1, y_1) = (0, 1)$ and $(x_2, y_2) = (1, 0)$, except we will reverse the order of the points to make the progress of gradient descent a bit clearer. So in this case our matrix/vector formulation is:

$$\boldsymbol{X} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}, \qquad \boldsymbol{y} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

1. Assuming $\alpha = 0.1$ and our initial values are $w_0 = 0$ and $w_1 = 0$, what are w_0 and w_1 after the just the first data point is used to update the gradient?

2. What are w_0 and w_1 after the second data point is used? Since we only have two examples here, your result would be the weight vector after the first iteration of SGD.

3. What is the value of the objective function (cost) after this initial iteration?

4. *(optional)* Sketch the linear model after this initial iteration.