Linear Regression: SGD

Begin with the following data (we will omit the first column of 1's in simple linear regression):

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

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.

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

(a) For this small example (cont. from Handout 4), the stochastic gradient descent updates are:

```
for i = 1,2:
```

$$w_0 \leftarrow w_0 - \alpha(w_0 + w_1 x_i - y_i) \cdot 1$$
$$w_1 \leftarrow w_1 - \alpha(w_0 + w_1 x_i - y_i) \cdot x_i$$

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 first step of gradient descent (i.e. first pass through all the data points)?

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

(c) Use the analytic solution (normal equations) to verify your results from Handout 4.