Some hints for Week 5's programming assignment

Please use this **only** if you are completely stuck!

1. Recall that in ridge regression you need to find $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$ to minimize the loss function

   $$L(w,b) = \sum_{i=1}^{n} (y^{(i)} - (w \cdot x^{(i)} + b))^2 \ + \ \lambda \|w\|^2,$$

   where the $(x^{(i)}, y^{(i)}) \in \mathbb{R}^d \times \mathbb{R}$ are the training data.

2. This loss function is convex and can be minimized using gradient descent. The first step in doing so is to compute the gradient $\nabla L(w,b)$ at any w and b. Try this before moving on to the next hints below.

3. For any parameters w, b and any data point i, we can define the ith residual as

   $$r_i = y^{(i)} - (w \cdot x^{(i)} + b).$$

   This tells us how far off the prediction $w \cdot x^{(i)} + b$ is on this point.

4. The derivative of the loss with respect to b is

   $$\frac{dL}{db} = 2 \sum_{i=1}^{n} r_i.$$

   Do you see why this is the case?

5. The derivative of the loss with respect to w is

   $$\frac{dL}{dw} = -2 \sum_{i=1}^{n} r_i x^{(i)} + 2\lambda w.$$

   Do you see why?

6. At this point, you can write down a gradient descent algorithm.

7. How to set the step size η? We'd like it to be as large as possible, without overshooting the mark. Here's a possible schedule: *for each update*,

   - Start with $\eta = 0.001$, say.
   - Repeatedly half η until you reach a value that leads to a reduction in the loss function. If η ever drops below some very small value (like 2^{-20}), halt and conclude that the algorithm has converged.
   - If the starting value of η already leads to a reduction in the loss function: repeatedly double η as long as this keep yielding bigger reductions in the loss function.