Exercise 2: Perceptron Questions (5 pts)

1. (3 pts) The perceptron algorithm makes an update every time it witnesses a mistake. What if it makes an update on every point, regardless of whether or not that point is correctly classified? For simplicity, consider a setting where where b is fixed to 0. Give an example of an infinite sequence of points (x_i, y_i) with the following properties:
   - I The sequence is strictly linearly separable with $b = 0$ (i.e., the margin is some constant $\gamma > 0$) ,
   - II The sequence has max $\|x_i\|_2 \leq 1$,
   - III The modified perceptron algorithm makes an infinite number of mistakes on this sequence .

   Prove that it has these properties. Note that the perceptron convergence theorem and the first two conditions imply that, at some point, the unmodified perceptron algorithm would only make a finite number of mistakes.

   Ans: **Answer 2.1**
2. (1 pt) Give examples of where the perceptron algorithm converges to a 0 margin halfspace, and a separate example where it converges to a maximum margin halfspace. As pointed out by some on Piazza, tech- nically, if a halfspace has 0 margin, then it would misclassify anything that still lies on the hyperplane, and the perceptron algorithm would not yet have halted. If you came up with an example that ignores these cases and halts with a point on the hyperplane, that's fine. However, the intent was more like the following, so consider solving the problem where it converges to an arbitrarily small margin halfspace. More precisely: for any $0 < \varepsilon < 1/2$, give a dataset (with margin at least 1) and a order in which to process the points such that the perceptron algorithm halts providing a halfspace with margin $\leq \varepsilon$. This problem has to do with the original perceptron, not the modified perceptron from part 1.

   Ans: **Answer 2.2**
3. (1 pt) Suppose that in each iteration of the perceptron algorithm, a point is chosen uniformly at random from the dataset. Show how the perceptron algorithm can be viewed as an instantiation of stochastic gradient descent (SGD). In particular, you must provide a loss function and a learning rate such that SGD with these parameters and perceptron are identical.

   Ans: **Answer 2.3**

Answer 2.2: Margin

a. Example for 0 margin halfspace:
It happens when $\gamma = 0$,
hence, all the data points are concentrated on one point,
and it would make infinity number of mistakes as $\lim_{\gamma \to 0}(\frac{R}{\gamma})^2 = \infty$.
Example: $(x_i, y_i) = ((0,0), (-1)^i)$

b. Example for a maximum margin halfspace:
It happens when $\gamma = R = \max_i ||a_i||_2$,
hence, all the data points are concentrated on two symmetrical points with opposing label.
It would also make one mistake $\lim_{\gamma \to R}(\frac{R}{\gamma})^2 = 1$
Example: $(x_i, y_i) = (((-1)^i, 0), (-1)^i)$, $\gamma = R = \max_i ||a_i||_2 = 1$