Q1. (1) Consider a modified perceptron algorithm that updates on every point, even if it is correctly classified, with $b = 0$. We construct an infinite sequence of points as follows: Let $\gamma > 0$ be the margin and set $\alpha = \sqrt{1-\gamma^2}$. Define

$$x_+ = (\gamma, \alpha), \quad x_- = -x_+ = (-\gamma, \alpha),$$

and let $n \in \mathbb{N}$ such that $n \geq \lceil 1/\gamma^2 \rceil$. Construct the sequence by repeating

$$(x_+, 1), \dots, (x_+, 1) \text{ (n times)}, (x_-, 1), \dots$$

infinitely.

Proving the properties:

i. Linearly separable: The sequence is separable with $b = 0$ and margin γ, because for $w = (0, 1)$, we have $y_i w^T x_i \geq \gamma$.

ii. Bounded norm: $\|x_i\|_2 = \sqrt{\gamma^2 + \alpha^2} = 1$, so $\max_i \|x_i\|_2 \leq 1$.

iii. Infinite mistakes: After k repetitions, let $P = kn$ and $N = k - 1$ be the counts of x_+ and x_- seen. Then

$$w = Px_+ + Nx_-, \quad b = 0.$$

When x_- is processed next, the update is

$$s = w^T x_- = k(\gamma^2 n - 1) + 2 \geq 2 > 0,$$

ensuring a mistake occurs. Since this pattern repeats for all $k \in \mathbb{N}$, the modified perceptron makes infinitely many mistakes.

(2) Examples of perceptron convergence:

(a) Arbitrarily small margin: For any $0 < \epsilon < 1/2$, let $a = (1, 0)$, $b = (\cos\theta, \sin\theta)$ with $\alpha = \epsilon/(1 - \cos\theta)$ and points

$$(\alpha a, 1), (\alpha b, 1).$$

Processing $(\alpha a, 1)$ first, then $(\alpha b, 1)$, the perceptron halts with weight $w = \alpha(a - b)$, giving a margin

$$\|w\| \cdot \min_i y_i x_i^T \hat{w} = \alpha(1 - \cos\theta) = \epsilon.$$

(b) Maximum margin: Take points $(\alpha a, 1)$ and $(-\alpha a, 1)$. Starting from $w_0 = 0$, after processing αa, then $-\alpha a$, the updates stabilize with $w = \alpha a$. The resulting halfspace is parallel to a, achieving the maximum margin, as no further updates occur.

(3) Perceptron as stochastic gradient descent:

Define the hinge-like loss function

$$\ell_w(x, y) = \max\{0, -y\langle w, x\rangle\}.$$

The gradient is

$$\nabla_w \ell_w(x_i, y_i) = \begin{cases} 0, & \text{if } y_i\langle w, x_i\rangle > 0, \\ -y_i x_i, & \text{if } y_i\langle w, x_i\rangle \leq 0. \end{cases}$$

Choosing a learning rate $\eta = 1$ and picking a point uniformly at random at each iteration, the SGD update

$$w_t = w_{t-1} - \eta \nabla_w \ell_w(x_{i_t}, y_{i_t})$$

matches the perceptron update exactly. Hence, the perceptron can be seen as SGD on ℓ_w with $\eta = 1$.