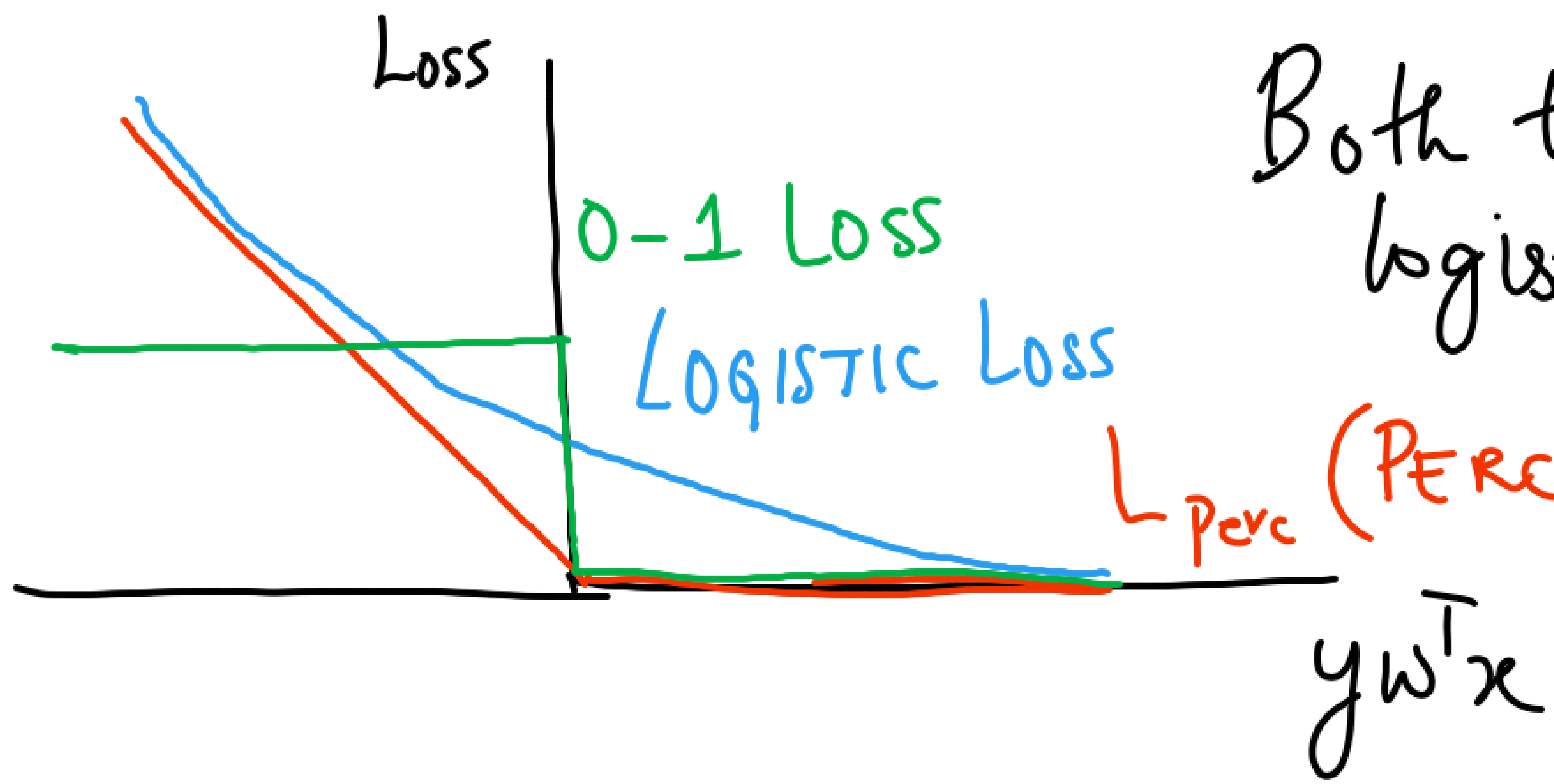
CS725

$$L_{perceptron}(x, y; w) = \max(0, -yw^Tx)$$



Both the perceptron loss & the logistic loss are <u>surrogates</u> of the 0-1 loss

Convergence of the Perceptron Learner

Consider a linearly separable dataset $\mathcal{D}$, i.e. there exists a u weight vector s.t. $y = \text{sign}(u^T x)$ $\forall x, y \in \mathcal{D}$.

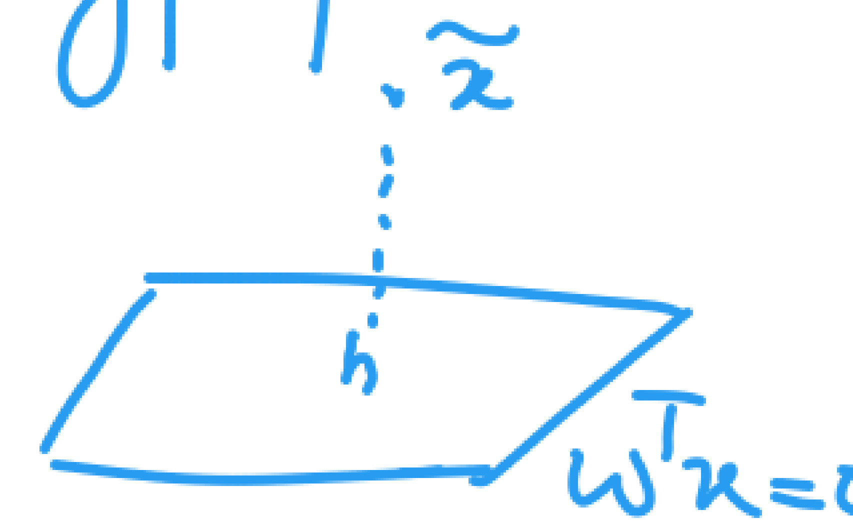
Two assumptions (without loss of generality):

1. Assume u is a unit vector
2. Assume all the x's $\in \mathcal{D}$ lie within a Euclidean ball of radius 1 i.e., $\|x\| \leq 1$

Define a new quantity called "MARGIN OF SEPARATION", γ

$$\gamma = \min_{x \in \mathcal{D}} |u^T x| \quad \left[\text{minimum distance of a point from the hyperplane } u^T x = 0 \right]$$

Distance of a point $\tilde{x}$ from a hyperplane



Pick a point h on the hyperplane

$$\tilde{x} - h = kw \Rightarrow h = \tilde{x} - kw$$

$$w^T(\tilde{x} - kw) = 0 \Rightarrow k = \frac{w^T \tilde{x}}{\|w\|^2}$$

Distance of $\tilde{x}$ from the hyperplane $= \|\tilde{x} - h\| = \dfrac{|w^T \tilde{x}|}{\|w\|_2}$

Theorem of convergence: If there exists a unit vector u such that $yu^T x \geq \gamma \;\; \forall\, x, y \in \mathcal{D}$, then the total number of mistakes made by a perceptron learner is bounded by $\frac{1}{\gamma^2}$ when trained on $\mathcal{D}$.

(Assume w is initialized to 0.)

Proof: Let us monitor the progression of two quantities:

(A) $w^T u$ and (B) $\|w\|_2$.

Why these quantities? We want $w^T u$ to be large but not only due to a large $\|w\|_2$. Keep track of both quantities as we iteratively estimate w.

Ⓐ Monitor $w^T u$

Let w_i be the weight vector at iteration i.

$$w_{i+1}^T u - w_i^T u = (w_i + yx)^T u - w_i^T u$$

$$\Rightarrow \quad w_{i+1}^T u = w_i^T u + yx^T u$$

$$\Rightarrow \quad w_{i+1}^T u \geq w_i^T u + \gamma$$

Assuming $w_0 = 0 \Rightarrow w_k^T u \geq k\gamma \longrightarrow$ Ⓐ

(B) Monitor $\|w\|_2$

$$\|w+yx\|_2^2 = \|w\|^2 + y^2\|x\|^2 + 2yw^Tx$$
$$= \|w\|^2 + \|x\|^2 + 2yw^Tx$$
$$\leq \|w\|^2 + \|x\|^2 \quad \left[\because\ yw^Tx < 0\right]$$
$$\leq \|w\|^2 + 1 \quad \left[\because\ \|x\|_2 < 1\right]$$

After k iterations, $\|w_k\|^2 \leq k \longrightarrow$ (B)

$$w_k^T u \geq k\gamma \longrightarrow ①$$

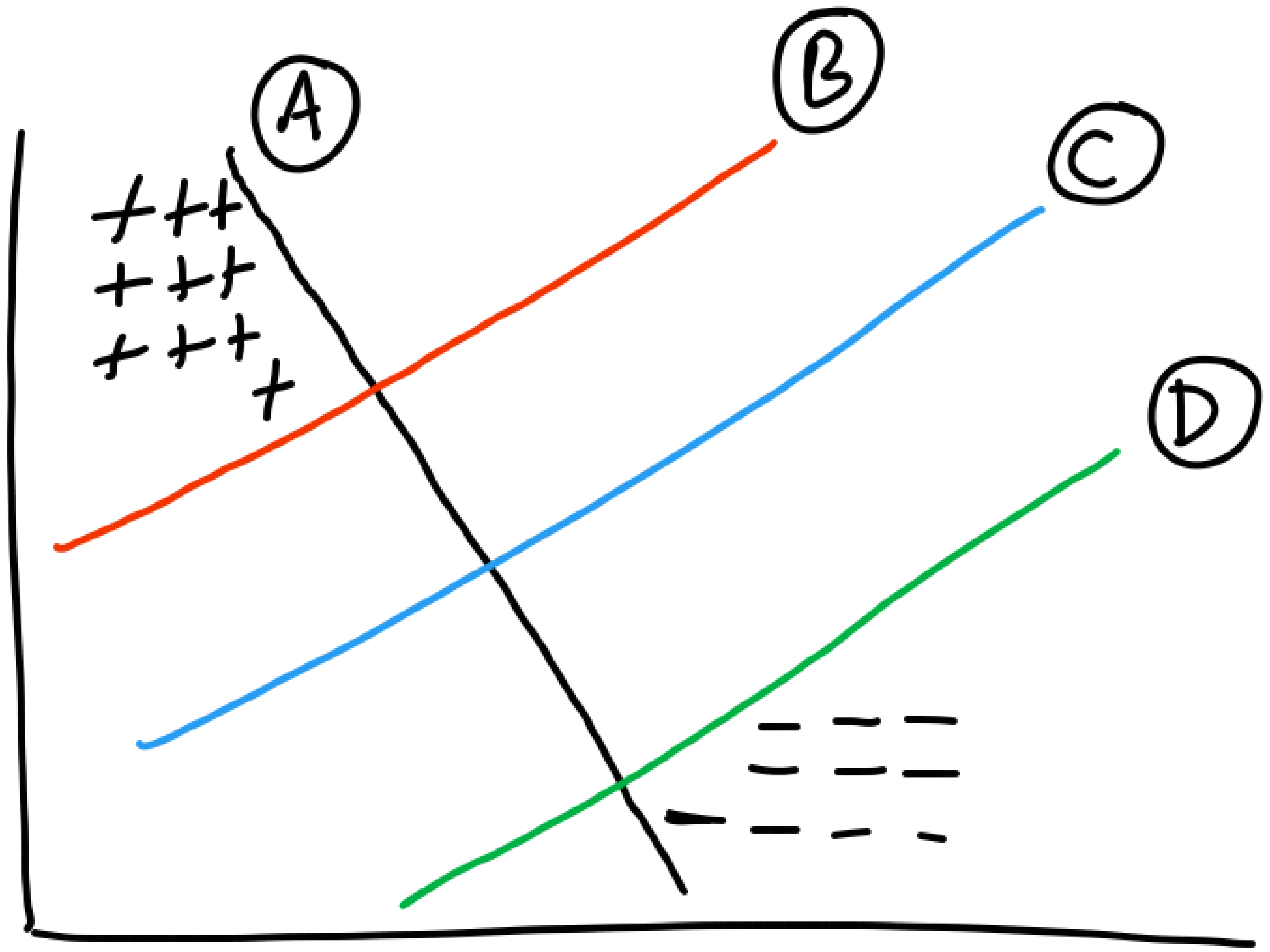
$$\|w_k\|^2 \leq k \longrightarrow ②$$

Putting ① and ② together

$$\sqrt{k} \geq \|w_k\| \geq w_k^T u \geq k\gamma$$

$$\Rightarrow \sqrt{k} \geq k\gamma$$

$$\Rightarrow \boxed{k \leq \frac{1}{\gamma^2}}$$



C is the natural pick

⇒ results in faster convergence ∵ of larger γ

⇒ results in better generalization

⇒ Motivates SVMs (Support Vector Machines)

find the predictor that yields the largest margin [SVMs are called max-margin classifiers]