Generating Adversarial Examples

Given:

$f(x) = \sigma(w \cdot x + b)$ with $\sigma(z) = \frac{1}{1+e^{-z}}$, $w = 0.5, b = -1$

$\text{loss}_t(x) = \left(t - f(x)\right)^2$

$x = 2.5$

Wanted:

η such that $f(x + \eta) \leq 0.5$

Tasks:

- Compute η using the fast gradient sign method with step $\epsilon = 0.5$
- Does $f(x + \eta) \leq 0.5$ hold?

1. Since $f(x) \in (0,1)$ for any x, and we need a perturbation η such that $f(x+\eta) \leq 0.5$, we set the target output of $f(x)$ to 0. This determines the target t loss for our loss function $loss_t$.
2. The target loss function $loss_t$ for target $t = 0$ is:

$$loss_0(x) = \left(0 - \sigma\left(\frac{x}{2} - 1\right)\right)^2 = \left(-\sigma\left(\frac{x}{2} - 1\right)\right)^2 = \sigma(\frac{x}{2} - 1)^2$$

3. Compute $\nabla_x loss_0(x)$:

Recall: $\frac{d}{dz}\sigma(z) = \sigma(z) \cdot (1 - \sigma(z))$

$$\nabla_x loss_0(x) = \frac{d}{dx} loss_0(x) =$$

$$= \frac{d}{dx}\sigma(\frac{x}{2} - 1)^2$$

$$= 2 \cdot \sigma\left(\frac{x}{2} - 1\right) \cdot \sigma\left(\frac{x}{2} - 1\right) \cdot (1 - \sigma\left(\frac{x}{2} - 1\right) \cdot \frac{1}{2}$$

$$= \sigma\left(\frac{x}{2} - 1\right)^2 - \sigma\left(\frac{x}{2} - 1\right)^3$$

4. $\eta = \epsilon \cdot sign\big(-\nabla_x loss_0(x)\big) = 0.5 \cdot sign\big(-\nabla_x loss_0(2.5)\big)$

$$= 0.5 \cdot sign(-0.138371) = 0.5 \cdot (-1) = -0.5$$

5. $f(x+\eta) = f(2.5 - 0.5) = f(2) = \sigma(0) = 0.5$, and therefore $f(x+\eta) \leq 0.5$ holds