Generating Adversarial Examples

Given:

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

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

$x = 2.5$

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?