(A) From Logistic Regression Model
After training & find best w, use \( P(y=1|x) = \frac{1}{1 + \exp(-w^T x)} \)

\[
0.8 < \frac{1}{1 + \exp(-w^T x)} = \text{true} \\
2 = 1 + \exp(-w^T x) \\
\exp(-w^T x) = 1 \\
-3x + 4 = 0 \\
x = \frac{4}{3}
\]

(B) Non-Adversarial Example
Let's start with an example where \( y = 1 \)

\( x \in \mathbb{R} \)

Now, we have the same \( w \), but now with label \( y = 0 \).

1st step: Check if boundary \( x \)

\[ x = \frac{w_1}{w_2} \]

End result (see Figure): Get classified as \( y = 0 \)

(C) Better Adversarial Example
Split with our example \( (x, y) \)
Take with label \( y = 0 \)

\[ w_1 x + w_2 \leq 3 \]

End result (see Figure): Get classified as \( y = 0 \)

(D) Black Box vs. White Box Formulas

1st step: Formulate & pick weights

2nd step: Compute loss

3rd step: Update weights
(c) Targeted vs. Non-targeted Attacks

Targeted: M is used only to speak directly
Non-targeted: M is used to any class other than target

(i) First Gradient Step Method
- (a) may be hard to find gradient steps
- Note: Just let $x_{i+1} = x - \epsilon \cdot \nabla f(x)$ where $L$ is a small honest loss.
- True
- Gradient penalty difference
  e.g. if $f(x)$ is high magnitude, without taking the step $\epsilon$ may be larger
- But the original $x$

- Especially effective once net can make activations at normally
  in the linear regime (for fast training)
  e.g. $y \approx \sigma(f(x))$

(iii) Conclude

- $y_{i+1} = 1$ if step is $x$
- $y_{i+1} = 0$ if $\rho(x) 
- D(b(x))$ and $D(\epsilon)$ are probabilities with $\rho(x)$.

$D$ is a classifier with the BCE loss!

$$
\log \frac{p_i}{1-p_i} = \frac{1}{\epsilon} \sum_{j=1}^{\epsilon} \log \left( 1 + \frac{p_i}{1-p_i} \epsilon \right) - \frac{1}{\epsilon} \sum_{j=1}^{\epsilon} \log \left( 1 + \frac{1-p_i}{p_i} \epsilon \right)
$$

$G_i$ is trying to fool $D$

$$
\log \frac{p_i}{1-p_i} = \frac{1}{\epsilon} \sum_{j=1}^{\epsilon} \log \left( 1 + \frac{p_i}{1-p_i} \epsilon \right) - \frac{1}{\epsilon} \sum_{j=1}^{\epsilon} \log \left( 1 + \frac{1-p_i}{p_i} \epsilon \right)
$$

By the definition, if $y_{i+1}$ is a following term. This is known actually. Discriminates is
damned well (make sense to have direct as opposed to generic).
Instead of measuring the likelihood of a transmission being correct, we measured the likelihood of a transmission being wrong:

\[
\begin{align*}
\text{max} J(\nu) &= \max \left[ \frac{1}{2} \sum_n \log \left( \frac{1}{\text{deg}(x_n)} \right) \right] \\
&= \min \left[ \frac{1}{2} \sum_n \log \left( \frac{1}{\text{deg}(x_n)} \right) \right] \\
&= \text{min} J(\nu)
\end{align*}
\]

One may verify \( \text{deg}(x_0) > 0 \) then

\[
\frac{\text{min} J(\nu)}{\text{deg}(x_0)} = \frac{\frac{1}{2} \sum_n \log \left( \frac{1}{\text{deg}(x_n)} \right)}{\text{deg}(x_0)} \to 0
\]

We call \( J(\nu) \) a non-negative loss function.