(A) True Logistic Regression Classifier
After training, pick a cut-off, e.g. 0.5.
\[ 0.5 = \frac{1}{1 + e^{-x}} \]
\[ 2 = 1 + e^{-x} \]
\[ e^{-x} = 1 \]
\[ -x = 0 \]
\[ x = 0 \]

(B) False Adversarial Example
Let's start with an example where \( y = 1 \).
\[ 2x + 1 \]
Now, we have the same\( \frac{\text{first line}}{\text{last line}} \) with label \( y = 0 \).
Let's adjust \( x \) to balance \( x \).
\[ x = \frac{1}{2} e \]
End Result (See Figure) Get classified as \( y = 0 \) \( \checkmark \)

(C) Better Adversarial Example
Spot with same example \( (x, 1) \)
Take with label \( y = 0 \)
\[ \text{loss} = 0.5 \exp -0.5(1 - y)^2 \]
End Result (See Figure) Get classified as \( y = 0 \) \( \checkmark \)

(D) Black Box vs. White Box Formulas
With Black Box, no knowledge about weights
\[ \text{Loss} = \text{Logit with } y = 0 \]
\text{Learn weights}
\[ \text{Loss} = \text{Logit with } y = 1 \]
(a) Targeted vs. Non-targeted Attacks

Targeted: M is precisely what we want to achieve.
Non-targeted: M is arbitrary to any class other than the one it chooses.

(b) Batch Gradient Descent Method
- Loss: \( \frac{1}{n} \sum_{i=1}^{n} -(y_i \log y + (1-y_i) \log(1-y)) \)
- Max: \( x = x - \alpha \cdot \nabla J(x) \)
- \( \alpha \) is the learning rate.

(c) Catalyst

Diagram:

- \( x \rightarrow y \rightarrow \phi(x) \rightarrow \nabla \phi \rightarrow \nabla f \rightarrow \nabla \phi(x) \rightarrow y \rightarrow x \)

(d) Catalyst

Diagram:

- Catalyst converts the gradient of \( f \) into the gradient of \( g \).
- \( g(y) = \frac{1}{n} \sum_{i=1}^{n} \log(1 + ye_i) \)
- \( \phi(x) = \frac{1}{n} \sum_{i=1}^{n} \log(1 + xe_i) \)

(e) Catalyst

Diagram:

- \( D = \lambda D + (1-\lambda) D(\nabla f(x)) \)
- \( \lambda \in (0,1) \)

(f) Catalyst

Diagram:

- \( J(\theta) = \frac{1}{n} \sum_{i=1}^{n} (\phi(x_i) - y_i) \)
- \( \phi(x) = \frac{1}{n} \sum_{i=1}^{n} \log(1 + xe_i) \)
- \( y_i \) is the target value.

(g) Catalyst

Diagram:

- \( \frac{1}{n} \sum_{i=1}^{n} \phi(x_i) \leq D(\phi) \)
- \( D(\phi) \) is a measure of divergence between the data and the function.

(h) Catalyst

Diagram:

- \( \frac{1}{n} \sum_{i=1}^{n} \phi(x_i) \) decreases as \( \phi(x) \) approaches the target value.
From (14), $\frac{a^{(1)}}{\hat{a}^{(1)}} \leq \frac{\sum \log(1 - D(G(y'|y)))}{\eta x}$

We can use $J(\theta)$ as a non-decreasing loss function to be trained.