Leakage From Model Updates

Model updates from gradient descent:

- Gradient updates reveal h:

$$y = W \cdot h, \frac{\delta L}{\delta W} = \frac{\delta L}{\delta Y} \cdot \frac{\delta y}{\delta W} = \frac{\delta L}{\delta y} \cdot h$$

- h = features of x learned to predict y

Spoiler: h leaks properties of x that are uncorrelated with y.
If the adv has examples of data with these properties, can use supervised learning to infer properties from the updates!

32

Leakage From Model Updates

Model updates from gradient descent:

- Gradient updates reveal h:

$$y = W \cdot h, \frac{\delta L}{\delta W} = \frac{\delta L}{\delta Y} \cdot \frac{\delta y}{\delta W} = \frac{\delta L}{\delta y} \cdot h$$

- h = features of x learned to predict y

Spoiler: h leaks properties of x that are uncorrelated with y.
If the adv has examples of data with these properties, can use supervised learning to infer properties from the updates!

Passive Property Inference Attack