PART 6

L15 Adversarial Attack

$x \Rightarrow [\,] \Rightarrow \cdots \Rightarrow [\,] \Rightarrow y \quad] \quad p(y|x)$

$= \quad p(y|z)$

Classification:

$$C(x) = \arg\max_{y} P(y|x)$$

Benign example: $(x, y) \qquad C(x) = y$

Targeted attack: $x \rightarrow x'$, $D(x, x') \leq \delta$ → L_1, L_2, L_∞

$C(x') = t \neq y$

Untargeted attack:

$C(x') \neq y$

General Principle

$\nabla_\theta L(x, y, \theta)$

Training: $\min_\theta \; E_{x,y}[-\log p(y|x,\theta)]$

Choose θ to max prob of true class y

After training: θ fixed (x, y)

Untargeted attack: $\max_{x'} -\log p(y|x',\theta)$ s.t. $D(x', x) \le \delta$

- ↑ prob of y
- ~~↓ prob of y~~

$t \neq y$

Targeted attack: $\min_{x'} -\log p(t|x',\theta)$ s.t. $D(x', x)$

- ~~↑ prob of t~~
- ↓ prob of t

$\nabla_x L(x, y, \theta)$ $\quad \nabla_x L(x, t, \theta)$

Fast Gradient Sign Method (FGSM)

Target : $y \to t$

$$\min_{x'} \; L(\bar{x}', t)$$
$$D(x, x') \leq \sigma$$

$$\bar{x}' = \bar{x} - \varepsilon \nabla_x L(x, t)$$

$(0.1, 100, -200, \cdots)$

problem: Magnitude of perturbations not controlled

$$\bar{x}' = \bar{x} - \varepsilon \, \text{Sig} \, \nabla_x L(x, t)$$

$(1, 1, -1 \cdots)$

what t to choose?

$x \Rightarrow$ cat dog .. Tower

0.7 0.1 0.00001

✓

Least likely class

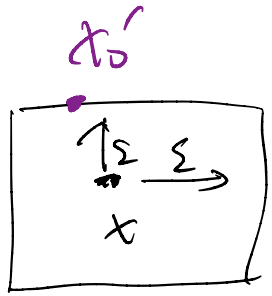
Basic Iterative Method (BIM)

$$x_0' = x$$

$$x_{i+1}' = \text{Clip}_{\bar{x},\varepsilon}\left[x_i' - \varepsilon \, \text{sign}\left(\nabla_x L(x_i', t) \right) \right]$$

If x_0' random point on the L_∞ ε-ball of $\bar{x}$

projected Gradient Descent (PGD)



First Attack algo: Box-constrained L-BFGS (2014)

$$\min_r \|r\|_2^2$$

s.t. $C(x') = t \quad x' = x + r \quad x' \in [0,1]^d$

$\Rightarrow$ Hard to solve. Reformulated as follows

$$\min_r \; c\|r\|_2^2 + L(x+r, t)$$

s.t. $x + r \in [0,1]^d$

$L: -\log p(\,)$

Strong attack : CW or Margin-Based (2017)

$$x \Rightarrow \cdots \Rightarrow \begin{bmatrix} z_1 \\ \vdots \\ z_t \\ \vdots \\ z_C \end{bmatrix} \quad \text{logits}$$

Ideas

① $\max z_t$, $\min \max_{i \neq t} z_i$: $\delta = z_t - \max_{i \neq t} z_i$

$\max \delta$

② $\delta > K$. turn attention to $\min \|x - x'\|_2^2$

CW objective

$$\min_{x'} \; c\|x - x'\|_2^2 + \max\left\{ \max_{i \neq t} z_i - z_t, -k \right\}$$

$-\delta$ $\begin{cases} \leq -k & ② \\ > -k & ① \end{cases}$

Linearity Hypothesis (Goodfellow et al 2015)

Linear classifier: $\bar{w}^T \bar{x} > 0$

$\bar{r} = -\epsilon \bar{w}$

$\bar{x}' = \bar{x} + \bar{r}$

$d = 65{,}000$

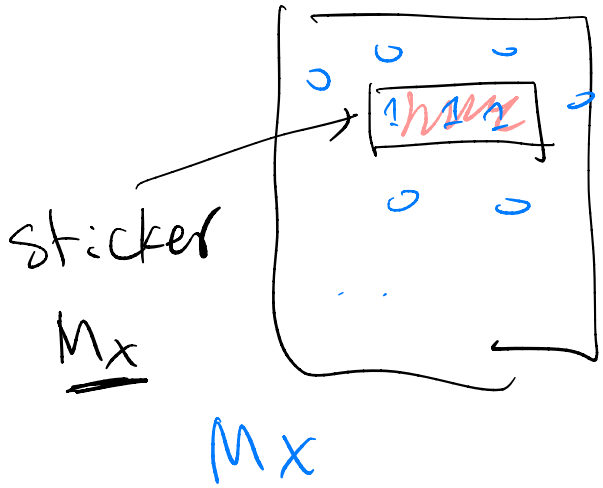
$\bar{w}^T \bar{x}' = \bar{w}^T \bar{x} - \epsilon \bar{w}^T \bar{w}$

Small → $w_1^2 + w_2^2 \cdots + w_d^2$ large

can easily cross the decision boundary

Works even if $\bar{r} = -\epsilon \, \text{sign}(\bar{w})$

sticker attack



$$\min_{r} \; c\,\|r \cdot m_x\|_2^2 + L(x + r \cdot M_x, t)$$

Adversarial Training / Data Augmentation

↳ mixup

crop

Data ⇒ model 1

⇓

A.E.

Continuity

Data + A.E. ⇒ model 2

⇓

More robust against attack

L16: XAI

2022-05-06

Sensitivity vs Attribution

$$z_c = 0.1\, x_1 + 10\, x_2 + 2$$

Sensitivity:

x_1	x_2
0.1	10

$$\frac{\partial z_c}{\partial x_i}$$

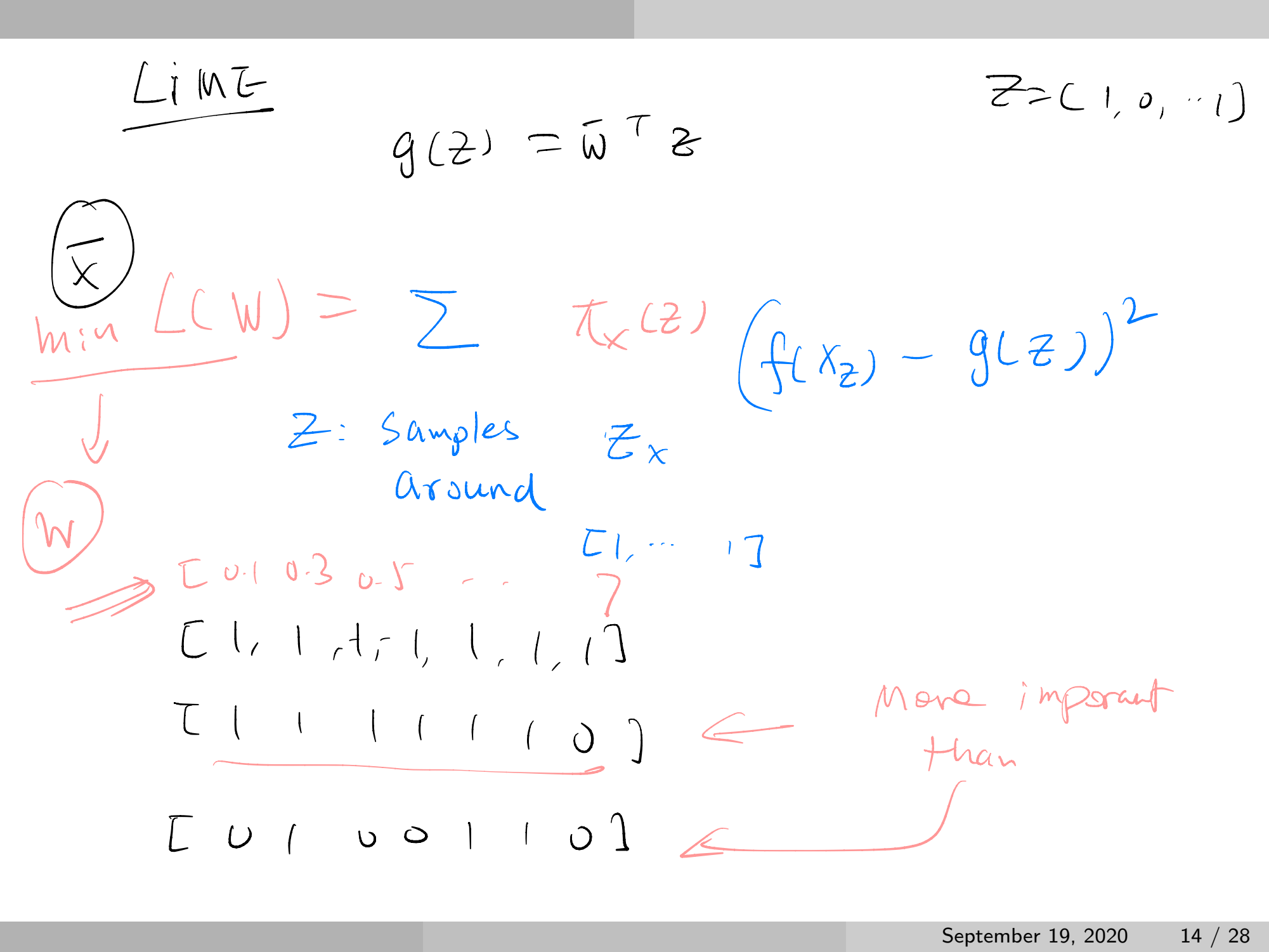
Attribution:

(100, 0.1)	10 ✓	1
(1, 1)	0.1	10 ✓

x_i $\quad$ x_2

input

$$x_i \quad \frac{\partial z_c}{x_i}$$


LIME
$g(z) = \bar{w}^T z$
$z = [1, 0, \cdots 1]$
$\bar{x}$
$\min L(w) = \sum \pi_x(z) \left(f(x_z) - g(z)\right)^2$
z: samples around
z_x
w
[0.1 0.3 0.5 . . .
[1, ··· 1]
?
[1, 1, 1, 1, 1, 1, 1]
[1 1 1 1 1 1 0]
More important than
[0 1 0 0 1 1 0]