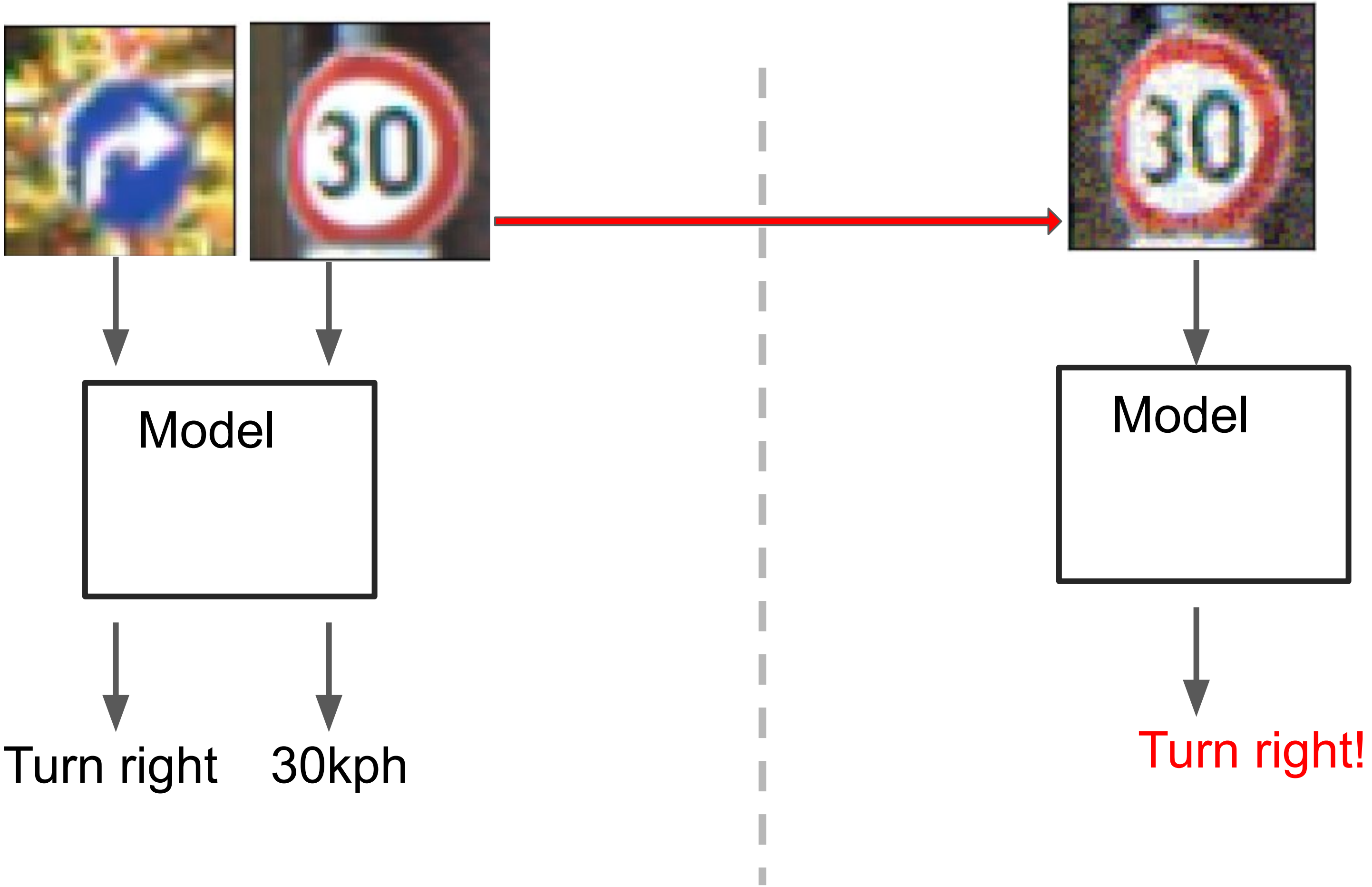
Constrained Gradient Descent: A Powerful and Principled Evasion Attack Against Neural Networks

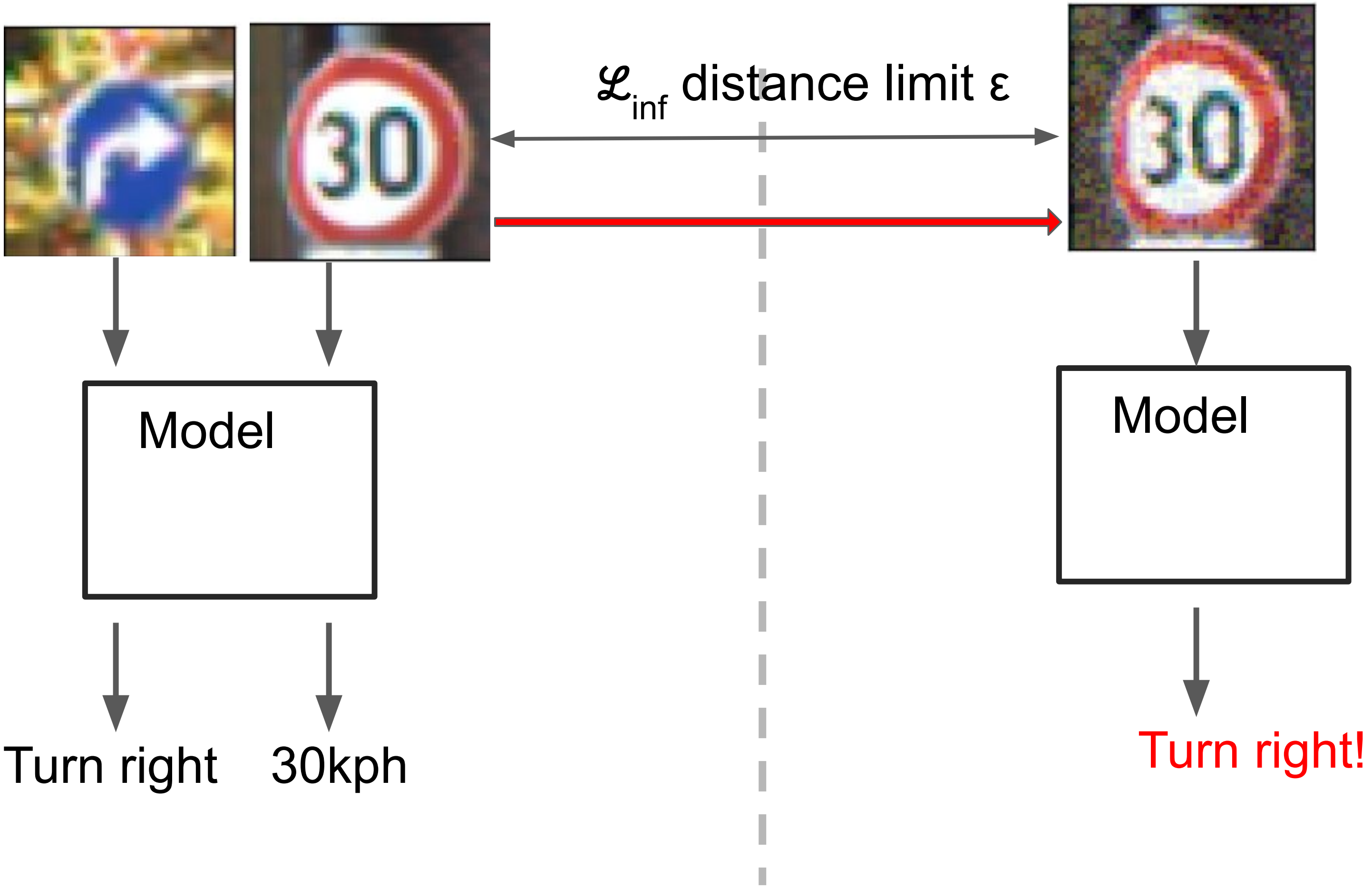
Weiran Lin[1] Keane Lucas[1] Lujo Bauer[1] Michael K. Reiter[2] Mahmood Sharif[3]

[1] Carnegie Mellon University
[2] Duke University
[3] Tel Aviv University

Why is studying targeted attacks important?



Why is studying targeted attacks important?



What are our contributions?

What are our contributions?

- We define a new loss function, MD loss
  - improves the previous best targeted evasion attack

What are our contributions?

- We define a new loss function, MD loss
    - improves the previous best targeted evasion attack
- We propose a new attack, CGD
    - finds *more adversarial examples*
    - and is also *faster*

What are the previous best targeted attacks?

- Previous best targeted evasion attack: auto-PGD

What are the previous best targeted attacks?

- Previous best targeted evasion attack: auto-PGD
- Previous best loss function: CW loss

Is CW loss *always* the best?

$L_{CW} = -Z_t + \max_{i \neq t} Z_i$

Is CW loss *always* the best?

$$L_{CW} = -\underline{Z_t} + \max_{i \neq t} Z_i$$

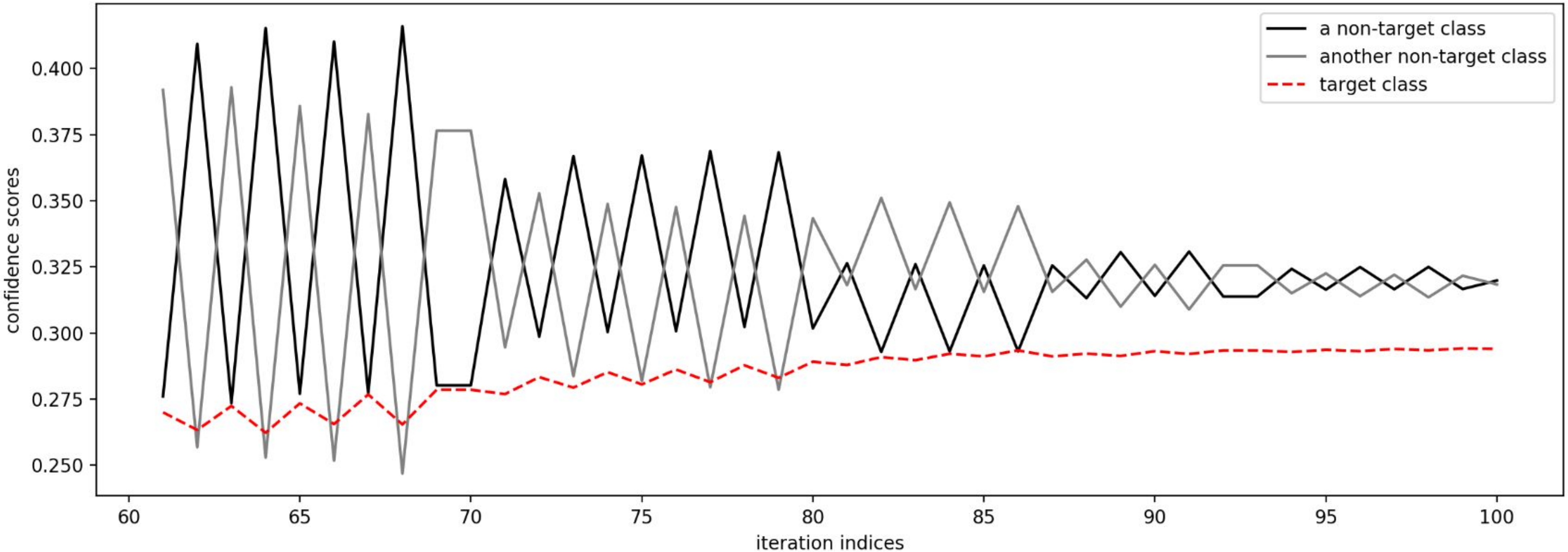
Logit of target class

Is CW loss *always* the best?

$$L_{CW} = -\underbrace{Z_t}_{\text{Logit of target class}} + \underbrace{\max_{i \neq t} Z_i}_{\text{Logit of the most probable non-target class}}$$

Is CW loss *always* the best?

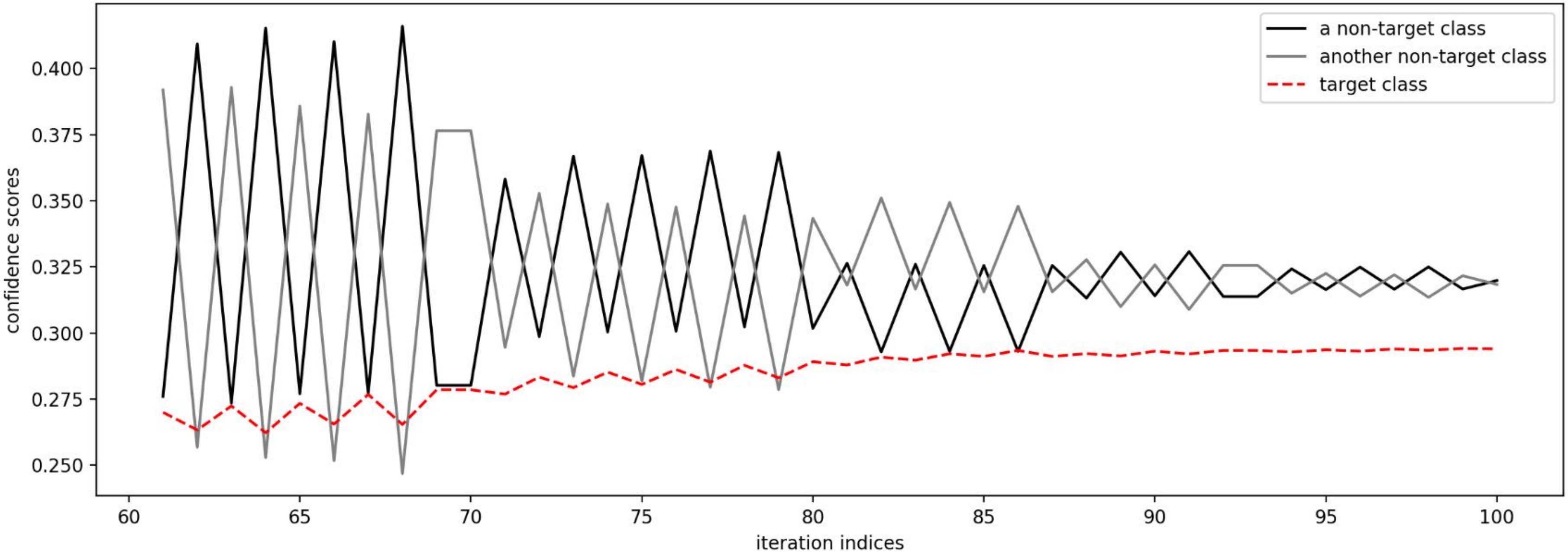
$L_{CW} = -Z_t + \max_{i \neq t} Z_i$



Is CW loss *always* the best?

$L_{CW} = -Z_t + \max_{i \neq t} Z_i$

Our first main contribution:
A loss function that captures ***all*** non-target logits



Is CW loss *always* the best?

$L_{CW} = -Z_t + \max_{i \neq t} Z_i$

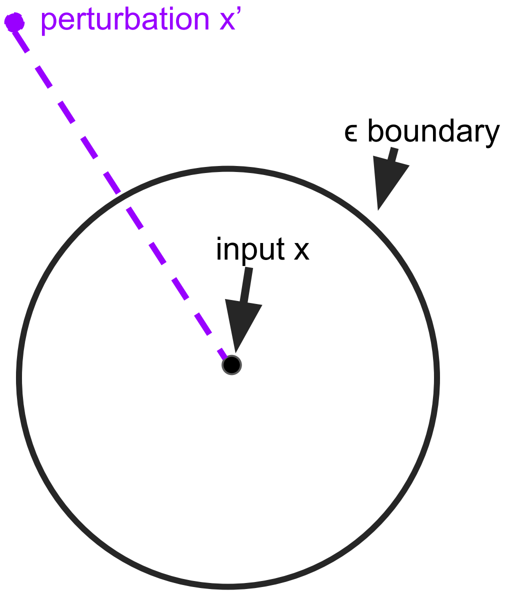
Our first main contribution:
A loss function that captures ***all*** non-target logits

Minimal Difference (MD) loss

$L_{MD} = \sum_{\text{all } i} \text{ReLU}(-Z_t + Z_i + \Delta)$

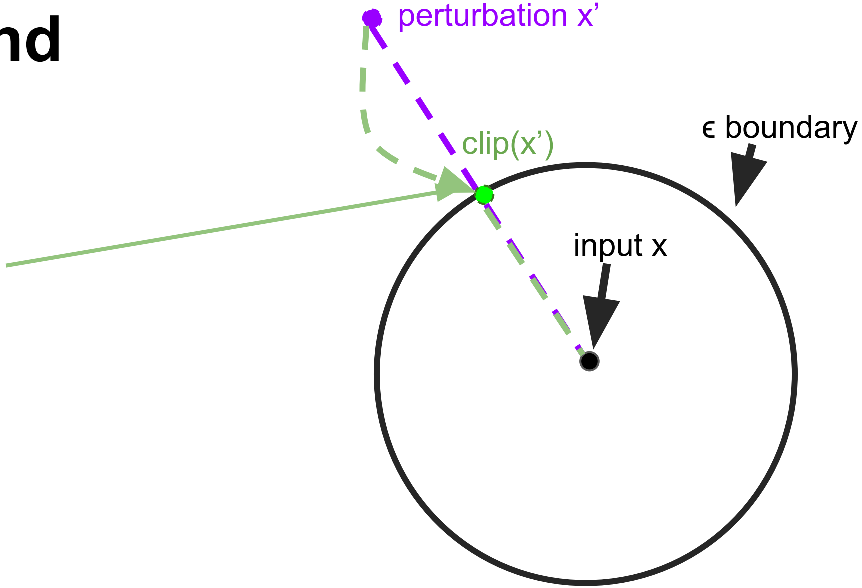
What else can we do to find a stronger attack?

What else can we do to find a stronger attack?



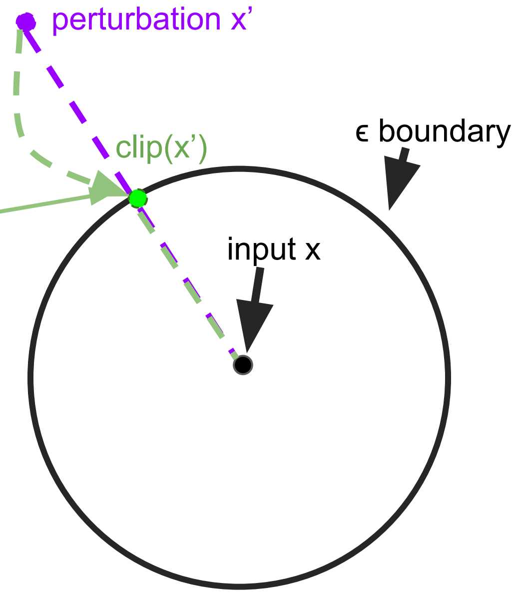
What else can we do to find a stronger attack?

Auto-PGD: projection



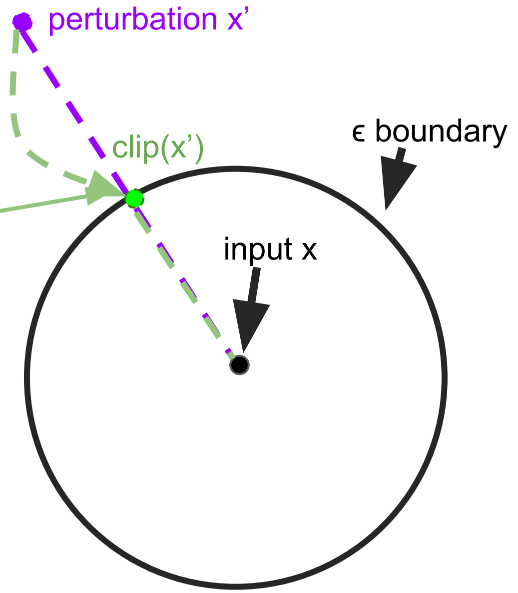
What else can we do to find a stronger attack?

Auto-PGD: projection



Our solution: **Constrained Gradient Descent (CGD)**

What else can we do to find a stronger attack?



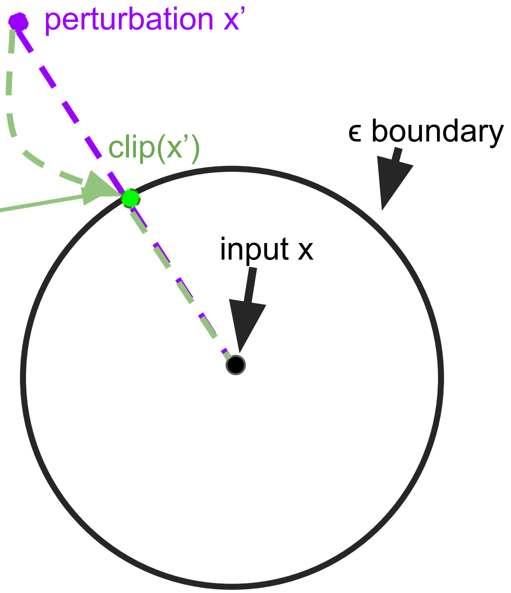
Auto-PGD: projection

Our solution: **Constrained Gradient Descent (CGD)**

- Distance limit as part of the loss function

What else can we do to find a stronger attack?

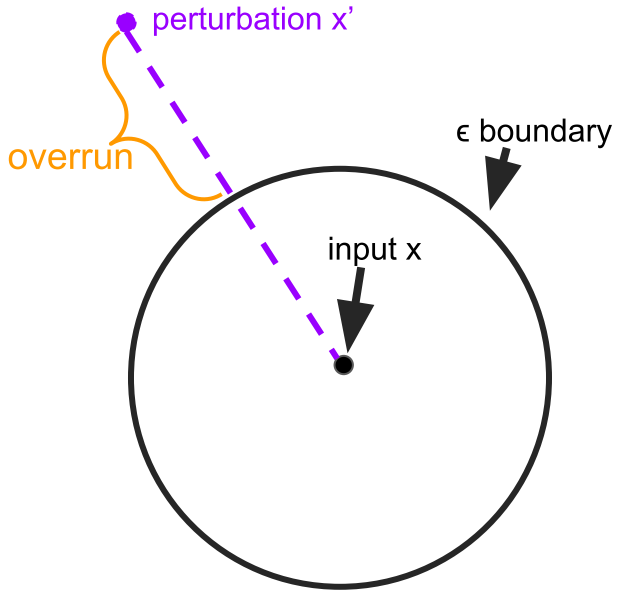
Auto-PGD: projection



Our solution: **Constrained Gradient Descent (CGD)**

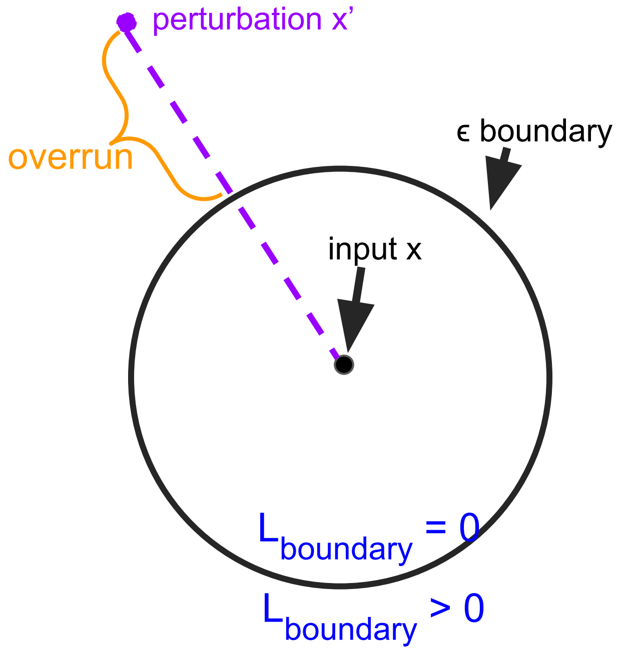
- Distance limit as part of the loss function
- Perturbation gradually encouraged to stay within distance limit

How does CGD work?



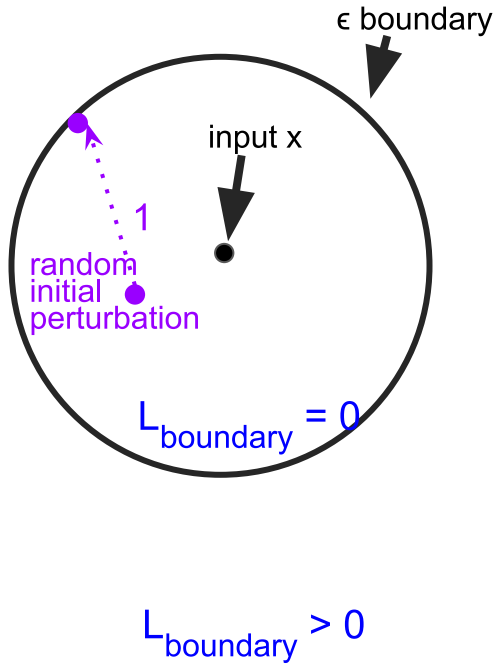
How does CGD work?

$$L = \underbrace{w \times L_{MD}}_{L_{classification}} + \underbrace{(1-w) \times overrun^2}_{L_{boundary}}$$



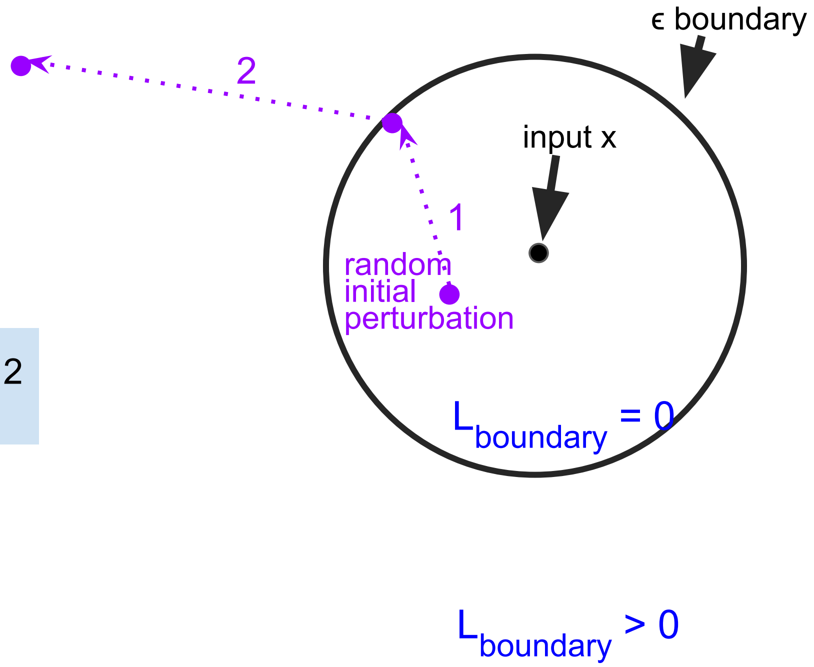
How does CGD work?

$$L = \underbrace{w \times L_{\mathbf{MD}}}_{L_{classification}} + \underbrace{(1-w) \times \text{overrun}^2}_{L_{boundary}}$$



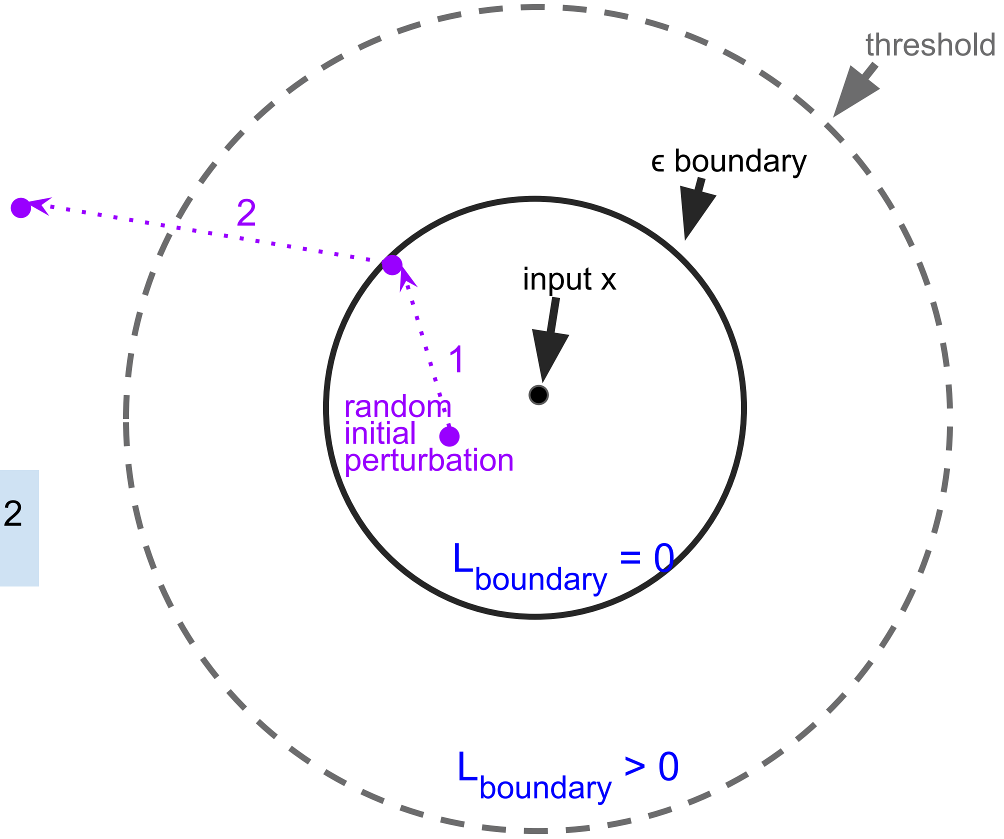
How does CGD work?

$$L = \underbrace{w \times L_{MD}}_{L_{classification}} + \underbrace{(1-w) \times \text{overrun}^2}_{L_{boundary}}$$



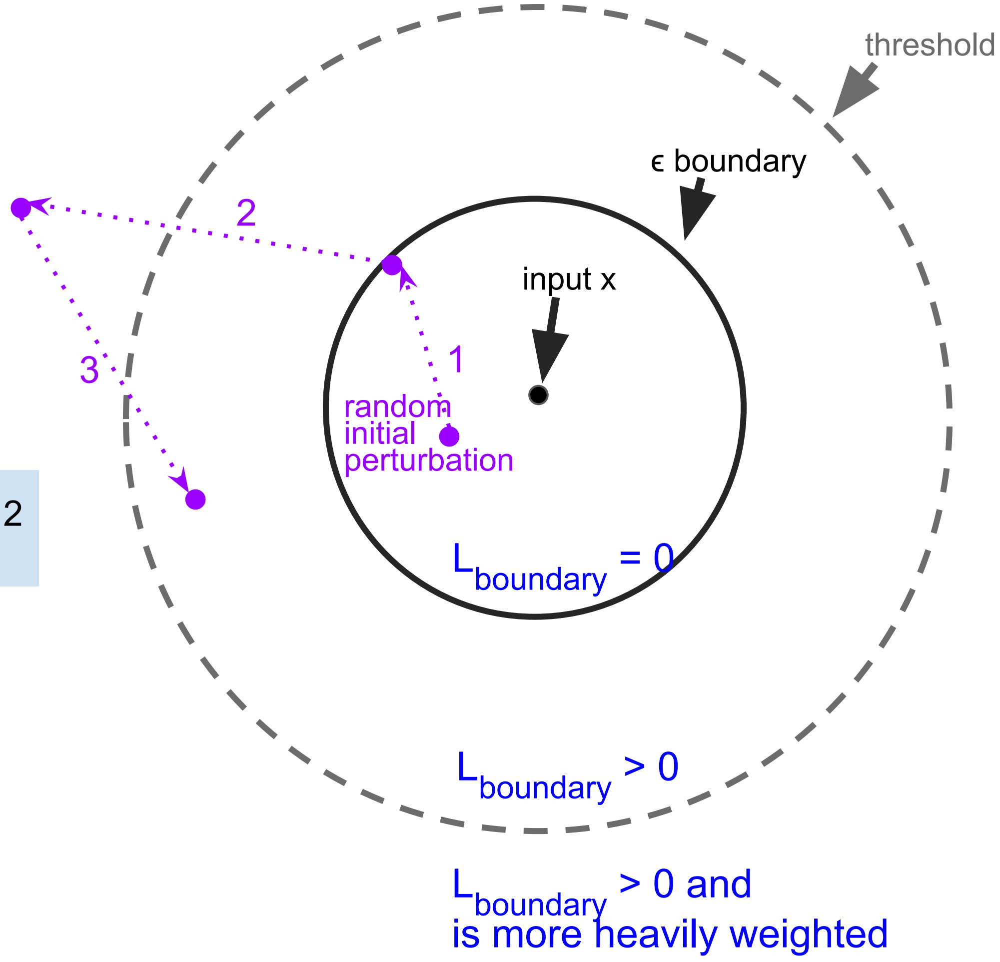
How does CGD work?

$$L = \underbrace{w \times L_{\mathbf{MD}}}_{L_{classification}} + \underbrace{(1-w) \times overrun^2}_{L_{boundary}}$$



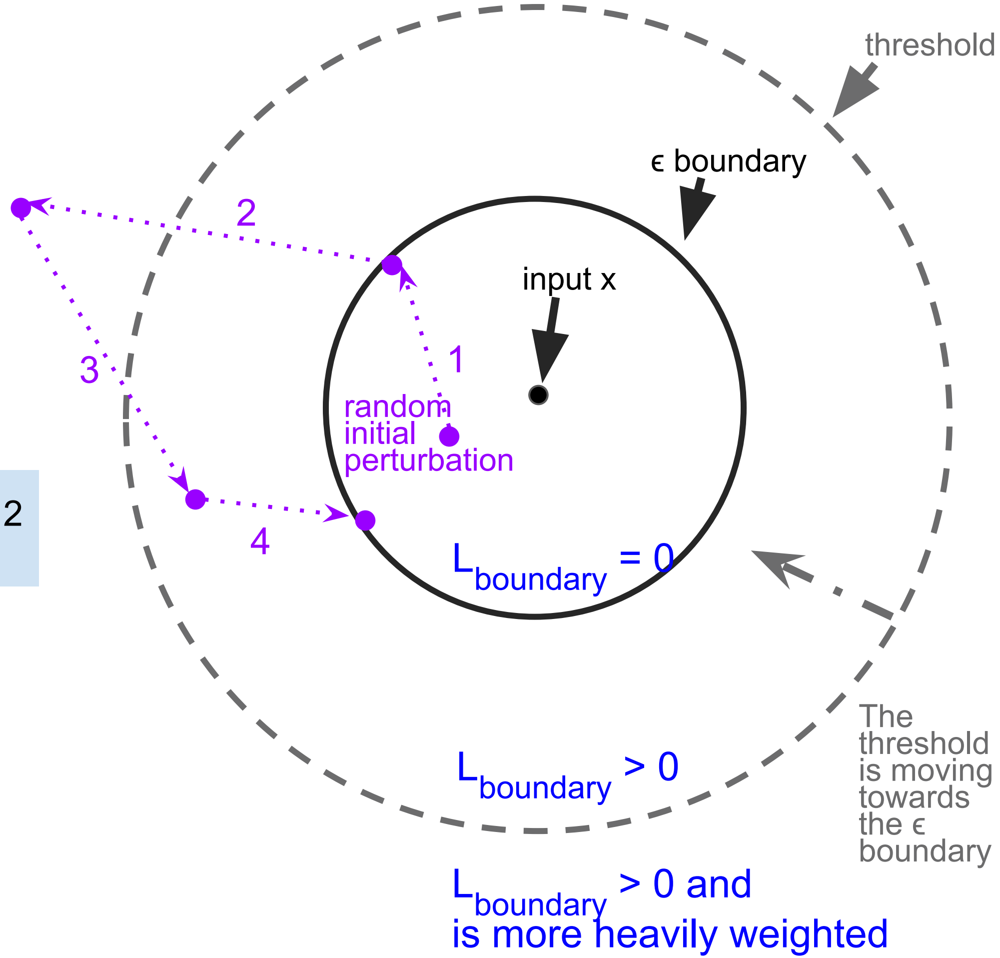
How does CGD work?

$$L = \underbrace{w \times L_{\mathbf{MD}}}_{L_{\text{classification}}} + \underbrace{(1-w) \times \text{overrun}^2}_{L_{\text{boundary}}}$$



How does CGD work?

$$L = \underbrace{w \times L_{MD}}_{L_{classification}} + \underbrace{(1-w) \times \text{overrun}^2}_{L_{boundary}}$$



How well do our attacks work?

How well do our attacks work?

- Auto-PGD with MD loss
  → up to 12% more adversarial examples

How well do our attacks work?

- Auto-PGD with MD loss
  → up to 12% more adversarial examples

- CGD
  → up to additional 1% more adversarial examples

How well do our attacks work?

- Auto-PGD with MD loss
  → up to 12% more adversarial examples

- CGD
  → up to additional 1% more adversarial examples
  → and, up to 19% faster

What are our takeaways?

What are our takeaways?

- We define a new loss function, MD loss
  - improves the previous best targeted evasion attack

Weiran Lin Keane Lucas Lujo Bauer Michael K. Reiter Mahmood Sharif

What are our takeaways?

- We define a new loss function, MD loss
  - improves the previous best targeted evasion attack
- We propose a new attack, CGD
  - finds *more adversarial examples*
  - and is also *faster*

Constrained Gradient Descent: A Powerful and Principled Evasion Attack Against Neural Networks

Weiran Lin Keane Lucas Lujo Bauer Michael K. Reiter Mahmood Sharif

What are our takeaways?

- We define a new loss function, MD loss
  - improves the previous best targeted evasion attack
- We propose a new attack, CGD
  - finds *more adversarial examples*
  - and is also *faster*
- We use CGD as a *framework* for attacks
  - second example use: a stronger untargeted attack

Constrained Gradient Descent: A Powerful and Principled Evasion Attack Against Neural Networks

Weiran Lin Keane Lucas Lujo Bauer Michael K. Reiter Mahmood Sharif