THE UNIVERSITY OF CHICAGO

PennState

Lower Bounds on Cross-Entropy Loss in the Presence of Test-time Adversaries

Arjun Nitin Bhagoji*, Daniel Cullina*,
Vikash Sehwag and Prateek Mittal

PRINCETON UNIVERSITY

Learning with test-time adversaries

Learning with test-time adversaries

Stop

Speed limit 80kmph

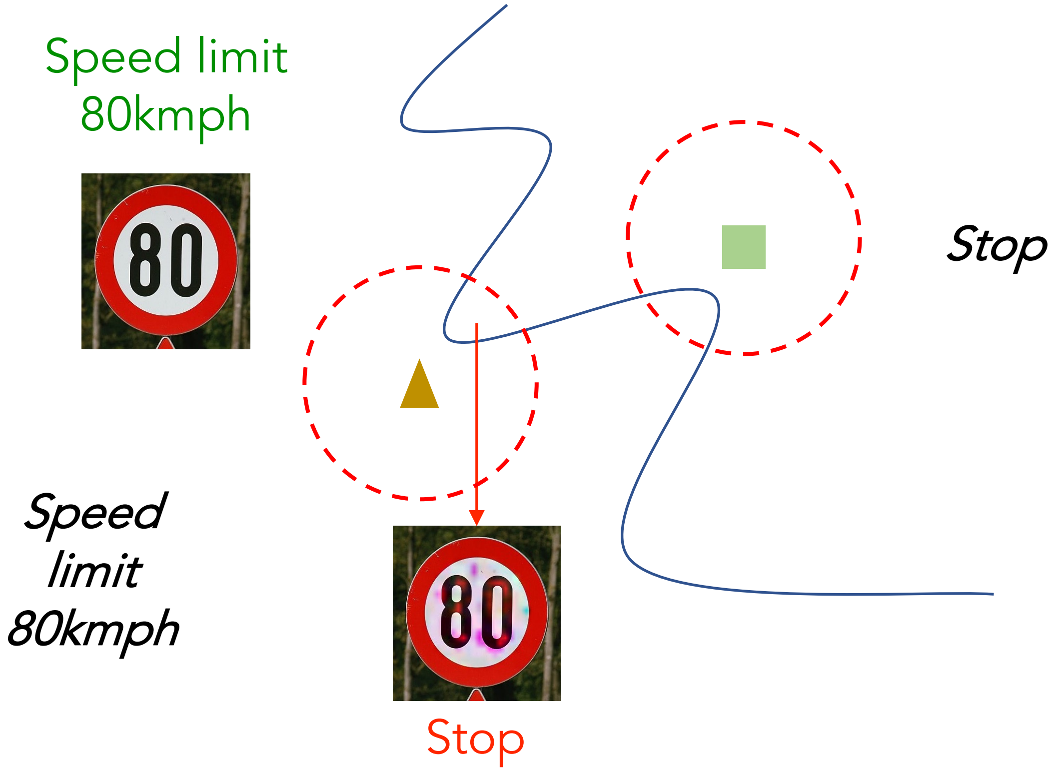
Learning with test-time adversaries

Speed limit
80kmph

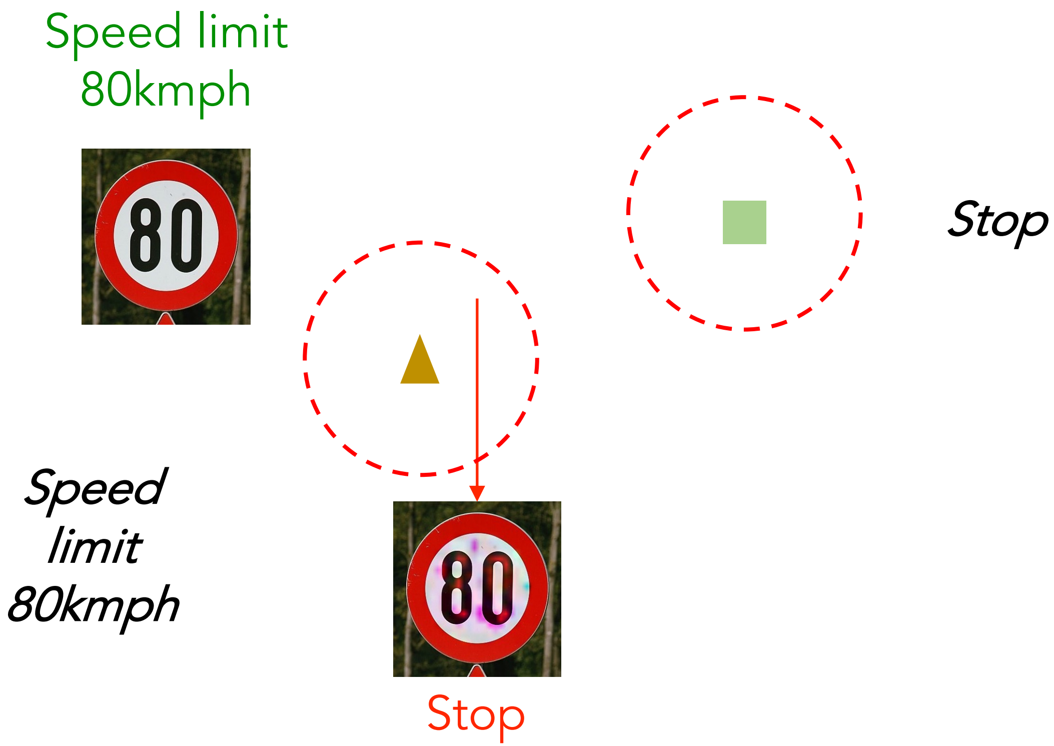
Stop

Speed
limit
80kmph

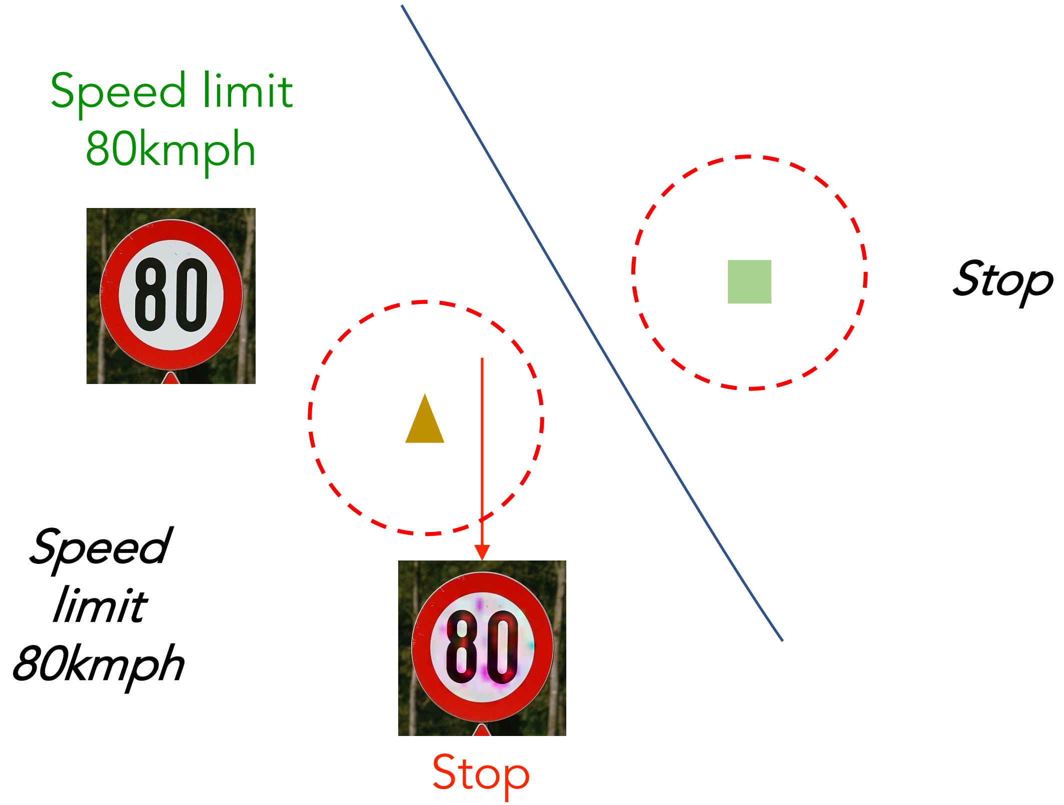
Learning with test-time adversaries



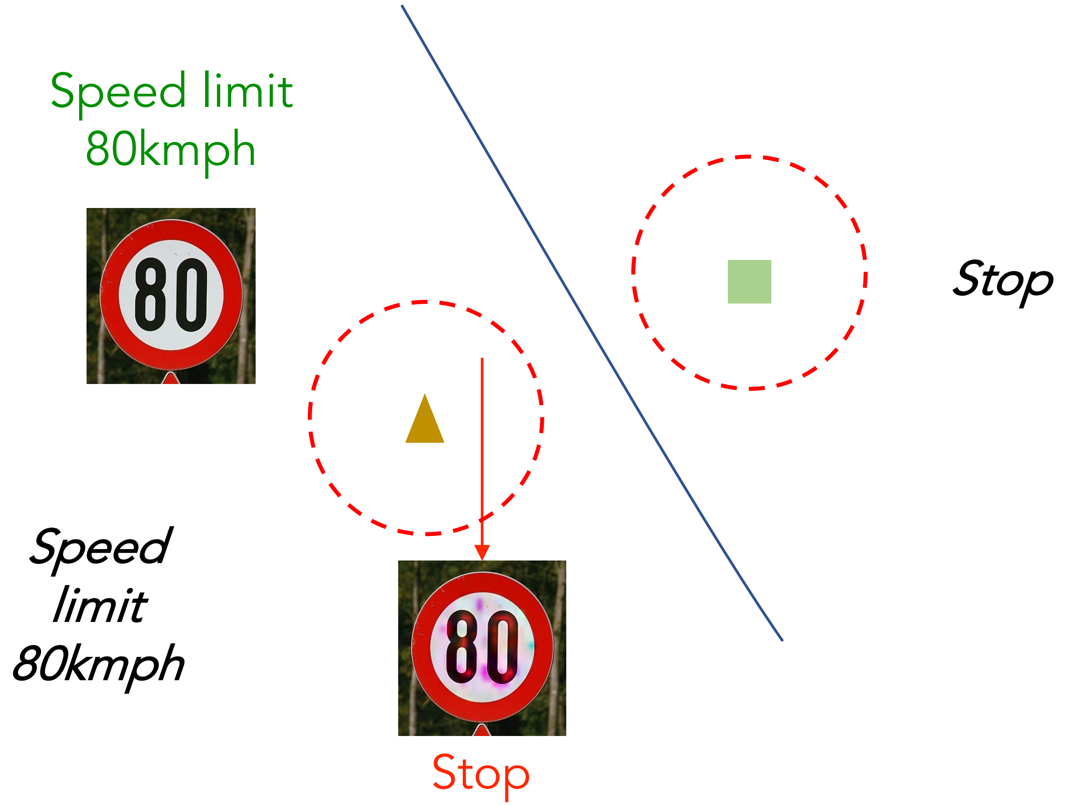
Learning with test-time adversaries



Learning with test-time adversaries



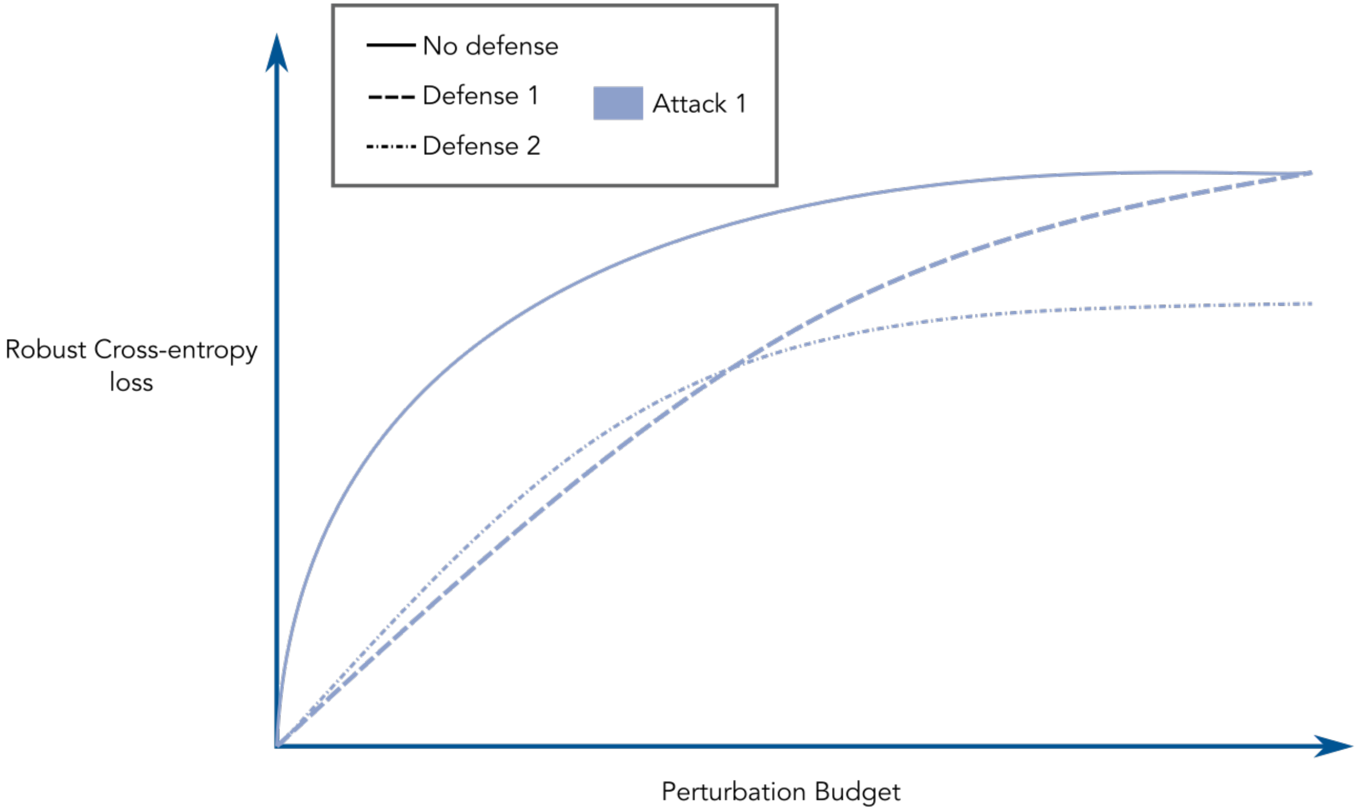
Learning with test-time adversaries



Overarching Question: What is the best performance any classifier can achieve in the presence of a worst-case perturbation?

The importance of lower bounds

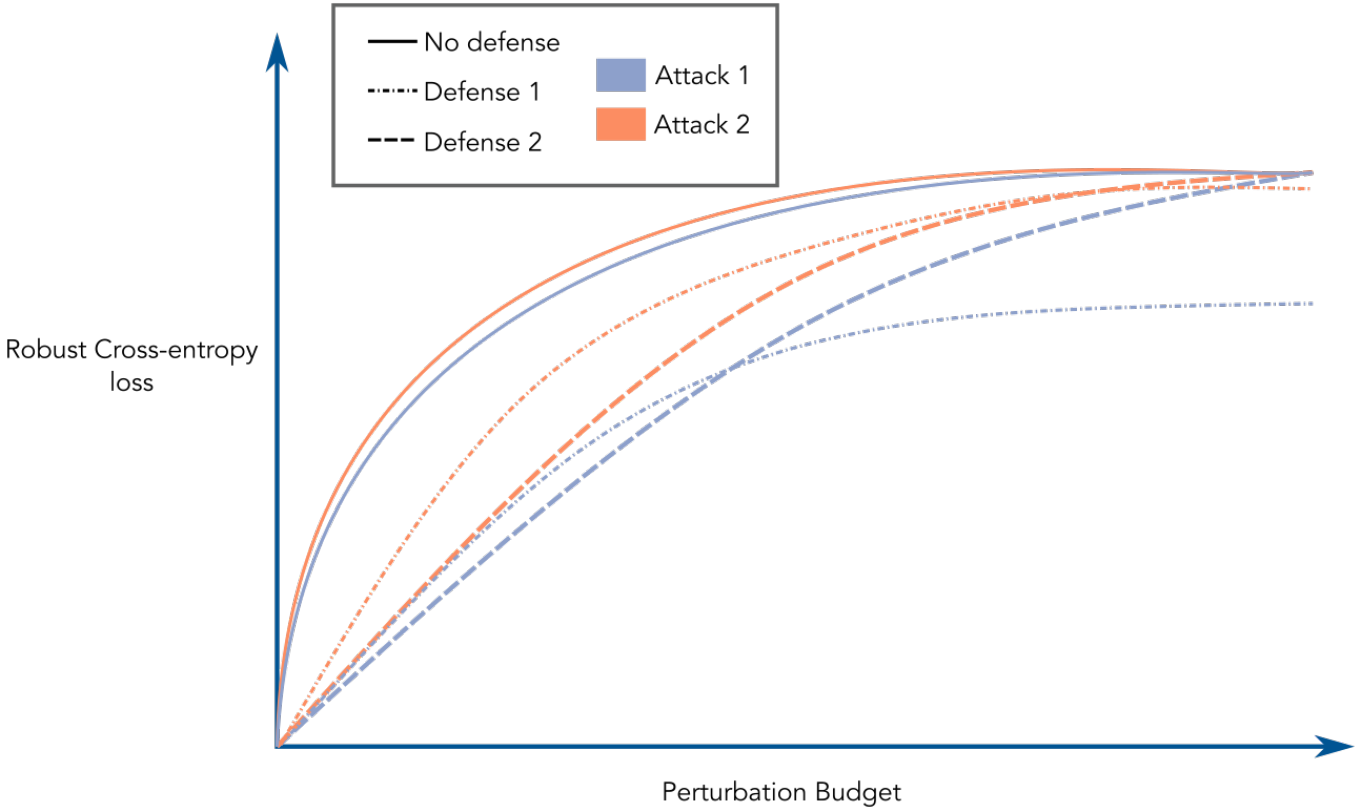
The importance of lower bounds



Cat and mouse game

- Defenses which improve upon regular training found by accounting for the attack

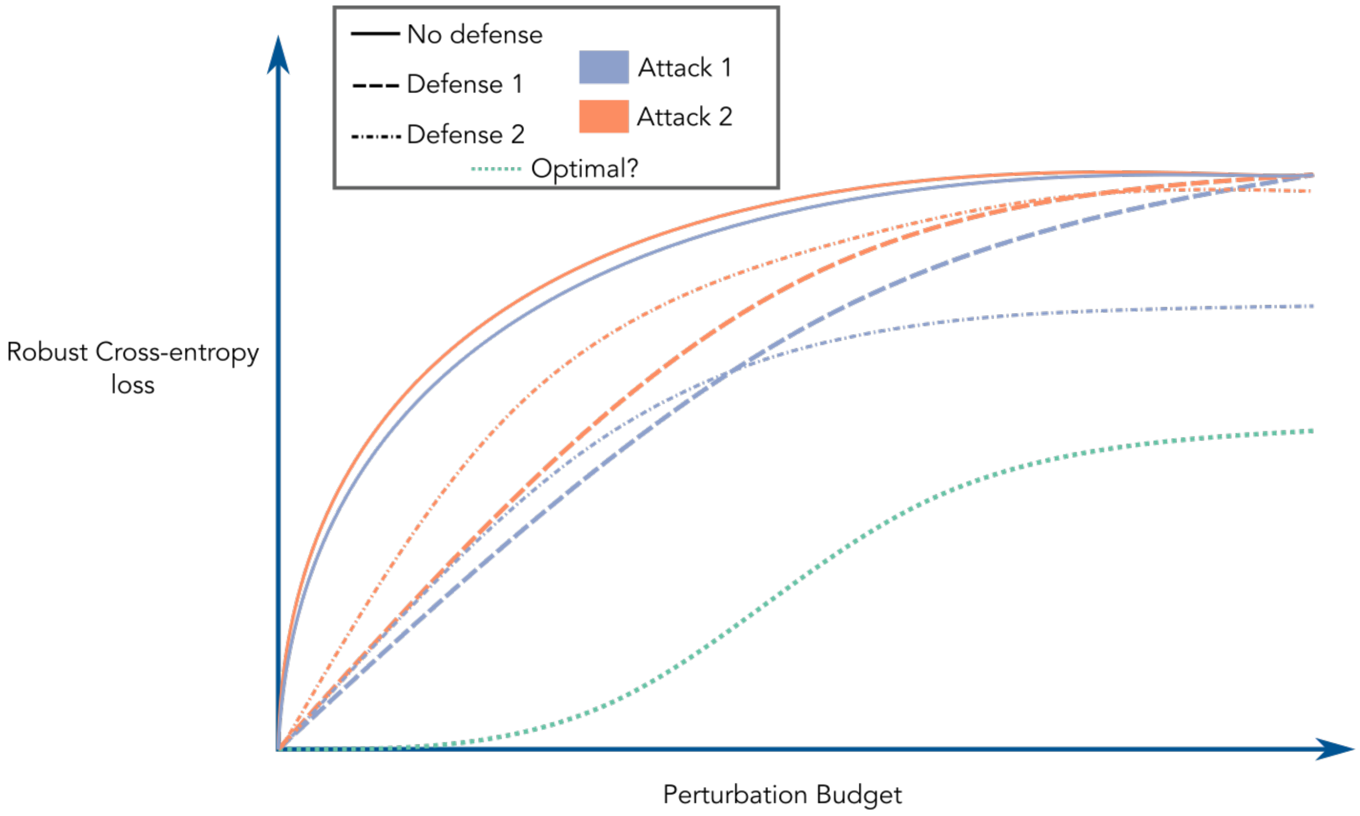
The importance of lower bounds



Cat and mouse game

- Defenses which improve upon regular training found by accounting for the attack
- Stronger (computationally and/or algorithmically) attack found, increasing loss

The importance of lower bounds



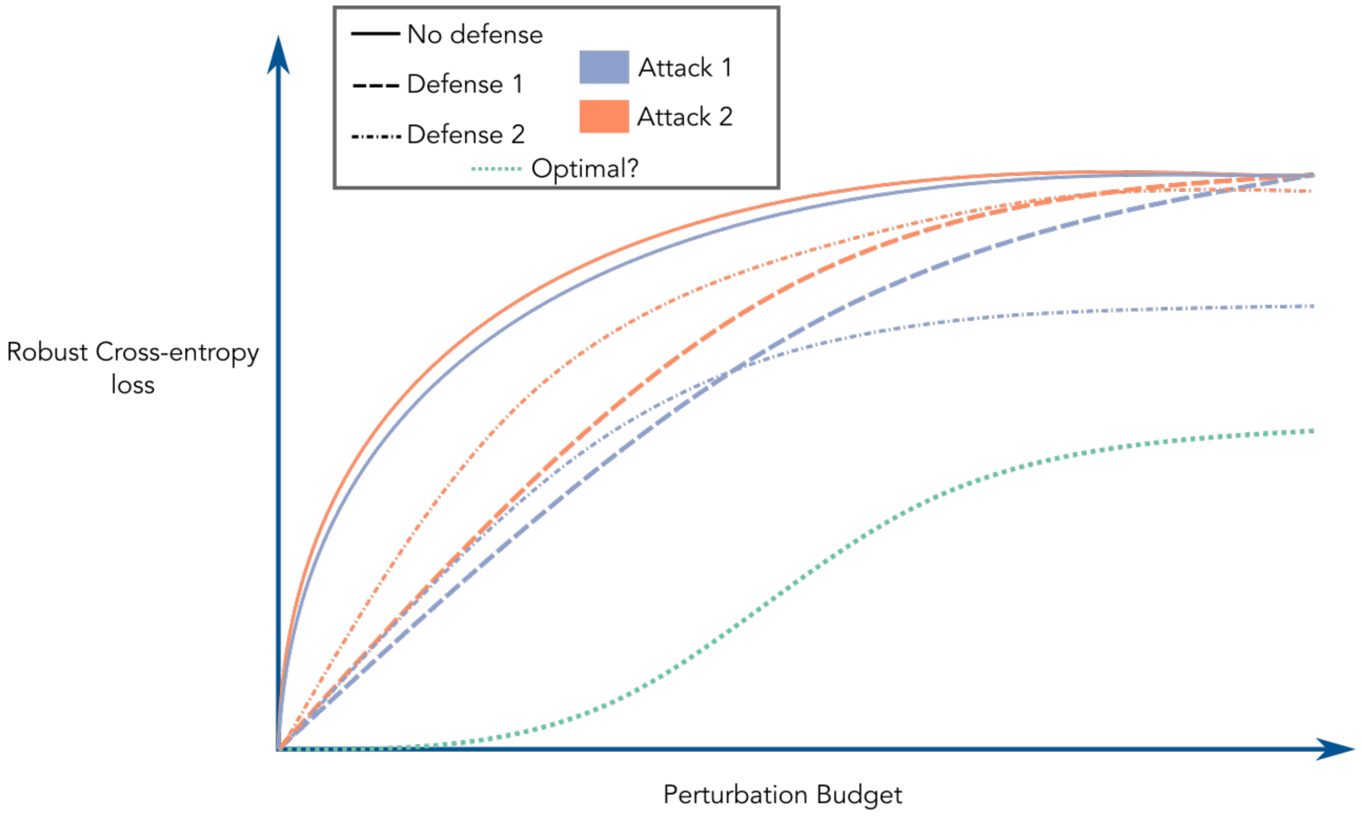
Cat and mouse game

- Defenses which improve upon regular training found by accounting for the attack
- Stronger (computationally and/or algorithmically) attack found, increasing loss

Breaking the cycle

- Lower bound determines lowest loss for the **best defense** against the **best attack**, ending the cat and mouse game!

The importance of lower bounds



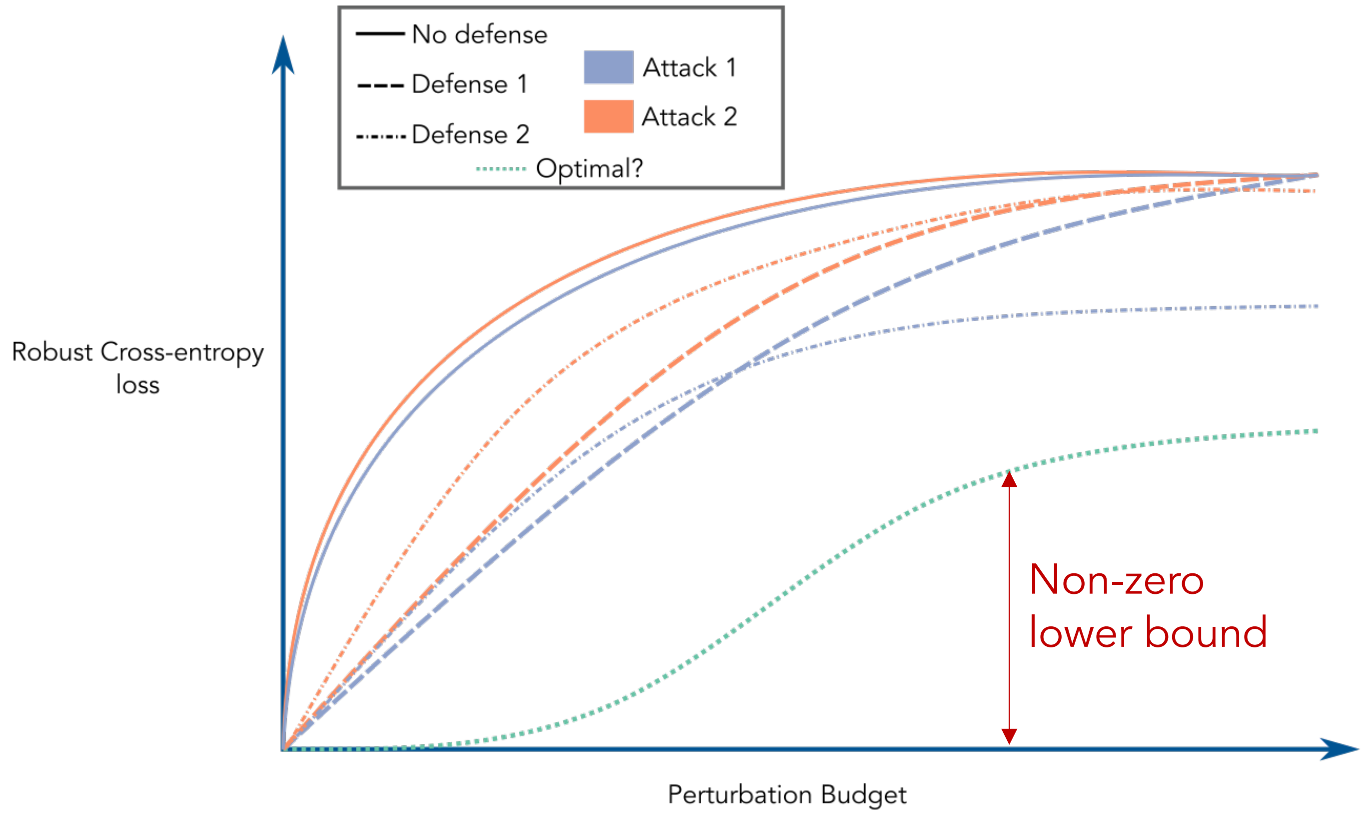
Cat and mouse game

- Defenses which improve upon regular training found by accounting for the attack
- Stronger (computationally and/or algorithmically) attack found, increasing loss

Breaking the cycle

- Lower bound determines lowest loss for the **best defense** against the **best attack**, ending the cat and mouse game!
- Provides essential information on
  - **Regimes where robustness is achievable**

The importance of lower bounds



Cat and mouse game

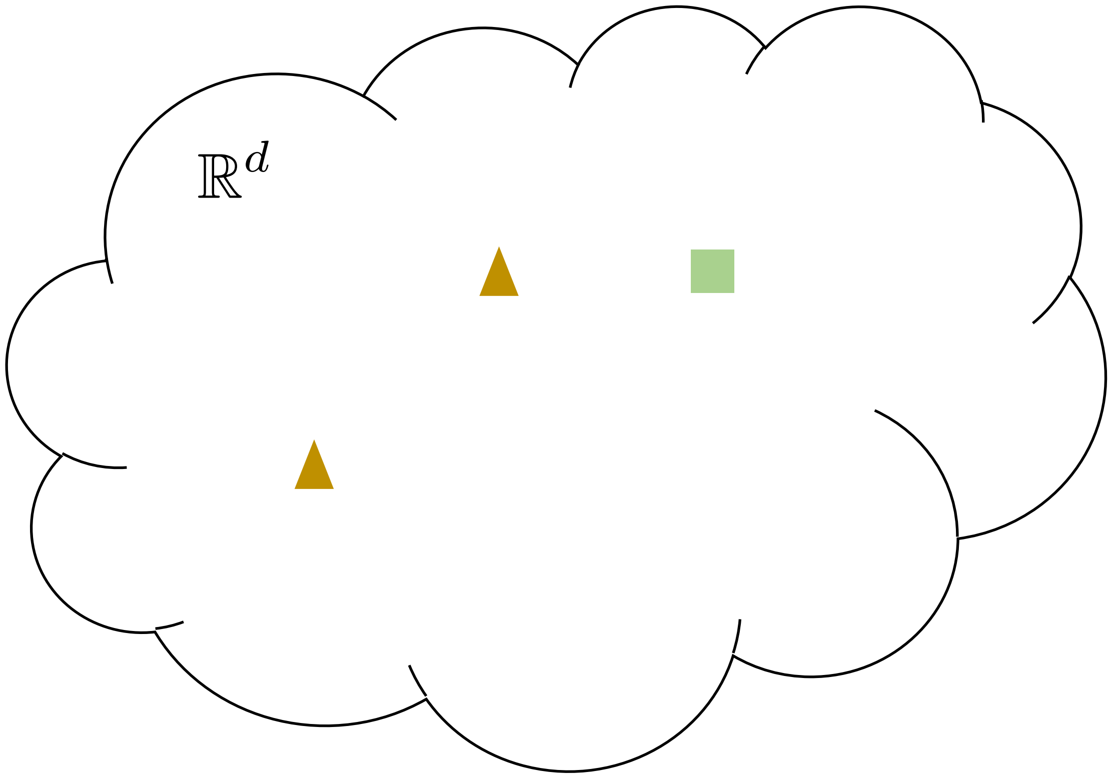
- Defenses which improve upon regular training found by accounting for the attack
- Stronger (computationally and/or algorithmically) attack found, increasing loss

Breaking the cycle

- Lower bound determines lowest loss for the **best defense** against the **best attack**, ending the cat and mouse game!
- Provides essential information on
  - **Regimes where robustness is achievable**
  - **Convergence of training**

Determining lower bounds on cross-entropy

Determining lower bounds on cross-entropy

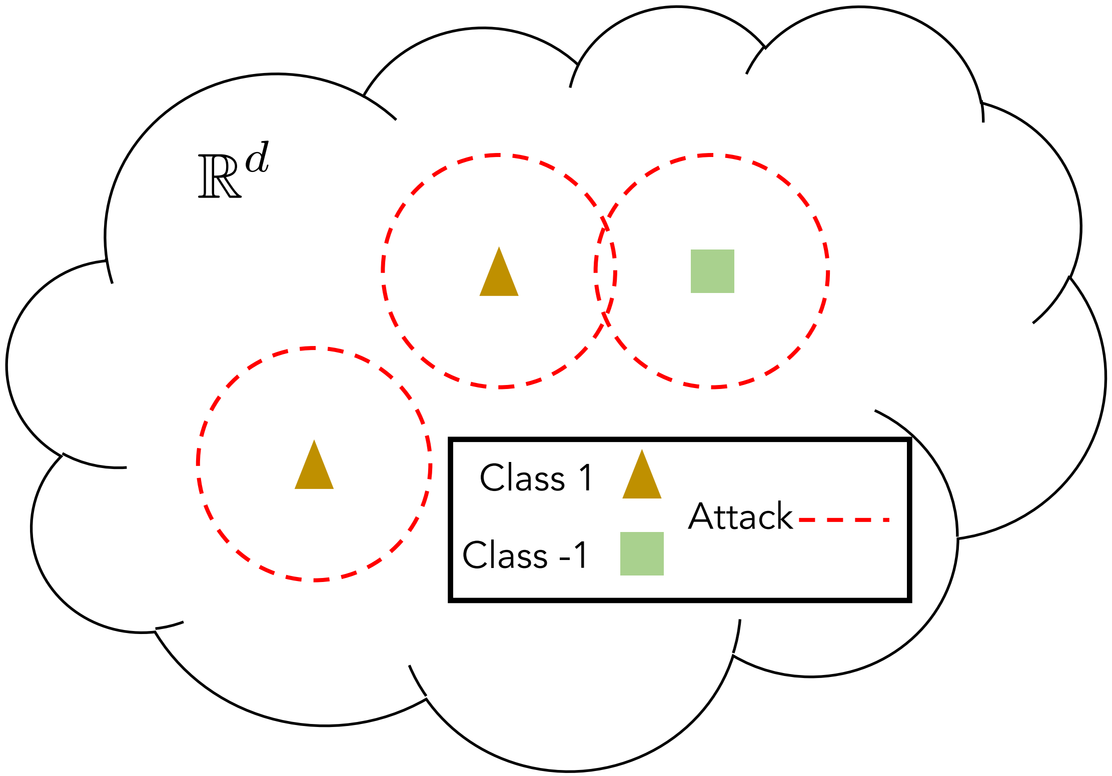


Minimal working example

Data distribution and Attack

- Data (in $\mathbb{R}^d$) is drawn from two classes (1 and -1), with equal sampling probability for each point

Determining lower bounds on cross-entropy

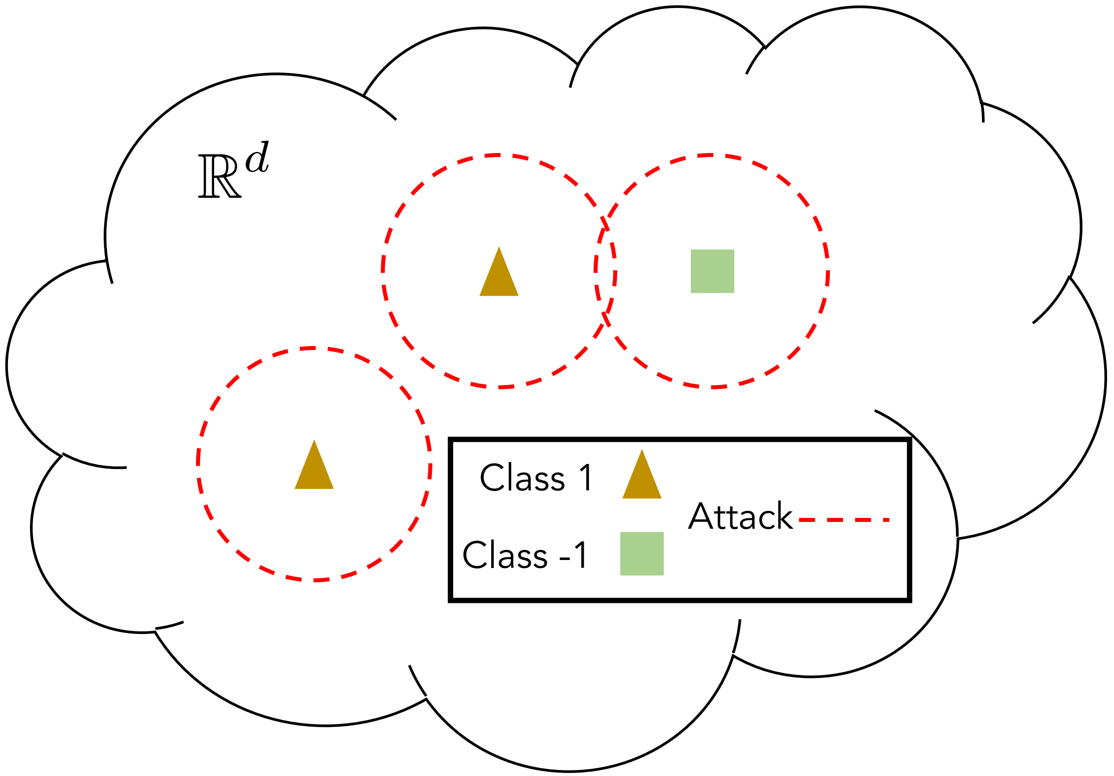


Minimal working example

Data distribution and Attack

- Data (in $\mathbb{R}^d$) is drawn from two classes (1 and -1), with equal sampling probability for each point
- Attacker perturbs data in ℓ_2 ball around each datapoint

Determining lower bounds on cross-entropy

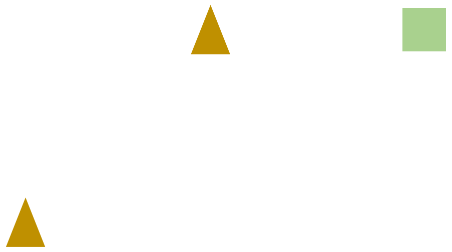


Minimal working example

Data distribution and Attack

- Data (in $\mathbb{R}^d$) is drawn from two classes (1 and -1), with equal sampling probability for each point
- Attacker perturbs data in ℓ_2 ball around each datapoint
- **Goal**: Find the minimum cross-entropy loss achievable by **any classifier**

Determining lower bounds on cross-entropy



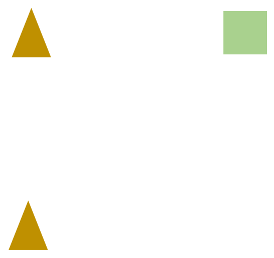
Minimal working example

Data distribution and Attack

- Data (in $\mathbb{R}^d$) is drawn from two classes (1 and -1), with equal sampling probability for each point
- Attacker perturbs data in ℓ_2 ball around each datapoint
- **Goal**: Find the minimum cross-entropy loss achievable by **any classifier**

Determining lower bounds on cross-entropy

Conflict graph $\mathcal{G}$



Minimal working example

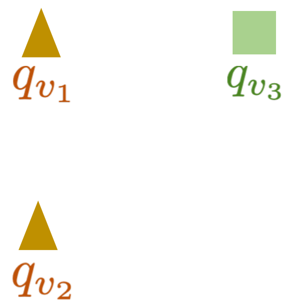
Data distribution and Attack

- Data (in $\mathbb{R}^d$) is drawn from two classes (1 and -1), with equal sampling probability for each point
- Attacker perturbs data in ℓ_2 ball around each datapoint
- **Goal**: Find the minimum cross-entropy loss achievable by **any classifier**

Graph representation and solution

Determining lower bounds on cross-entropy

Conflict graph $\mathcal{G}$



Minimal working example

Data distribution and Attack

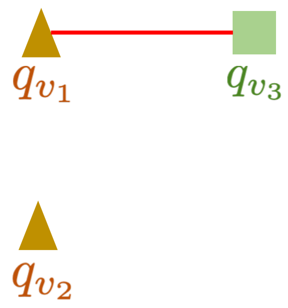
- Data (in $\mathbb{R}^d$) is drawn from two classes (1 and -1), with equal sampling probability for each point
- Attacker perturbs data in ℓ_2 ball around each datapoint
- **Goal**: Find the minimum cross-entropy loss achievable by **any classifier**

Graph representation and solution

-Probability on vertices represents classifier output

Determining lower bounds on cross-entropy

Conflict graph $\mathcal{G}$



Minimal working example

Data distribution and Attack

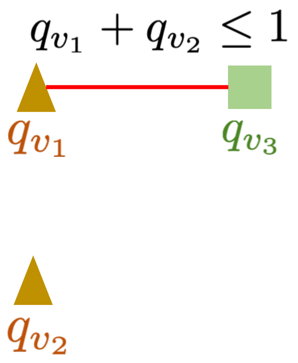
- Data (in $\mathbb{R}^d$) is drawn from two classes (1 and -1), with equal sampling probability for each point
- Attacker perturbs data in ℓ_2 ball around each datapoint
- **Goal**: Find the minimum cross-entropy loss achievable by **any classifier**

Graph representation and solution

-Probability on vertices represents classifier output
-Edges represent overlapping perturbation balls and

Determining lower bounds on cross-entropy

Conflict graph $\mathcal{G}$



Minimal working example

Data distribution and Attack

- Data (in $\mathbb{R}^d$) is drawn from two classes (1 and -1), with equal sampling probability for each point
- Attacker perturbs data in ℓ_2 ball around each datapoint
- **Goal**: Find the minimum cross-entropy loss achievable by **any classifier**

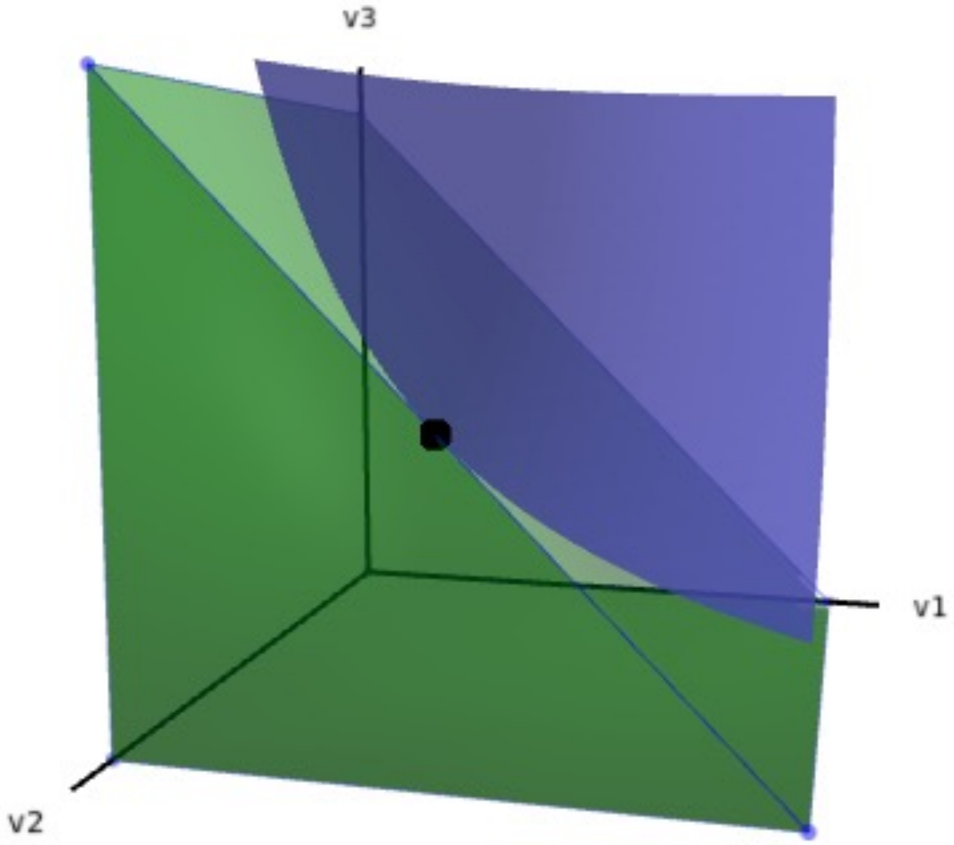
Graph representation and solution

-Probability on vertices represents classifier output

-Edges represent overlapping perturbation balls and

-Enforce constraints on the convex minimization problem

Determining lower bounds on cross-entropy



Data distribution and Attack

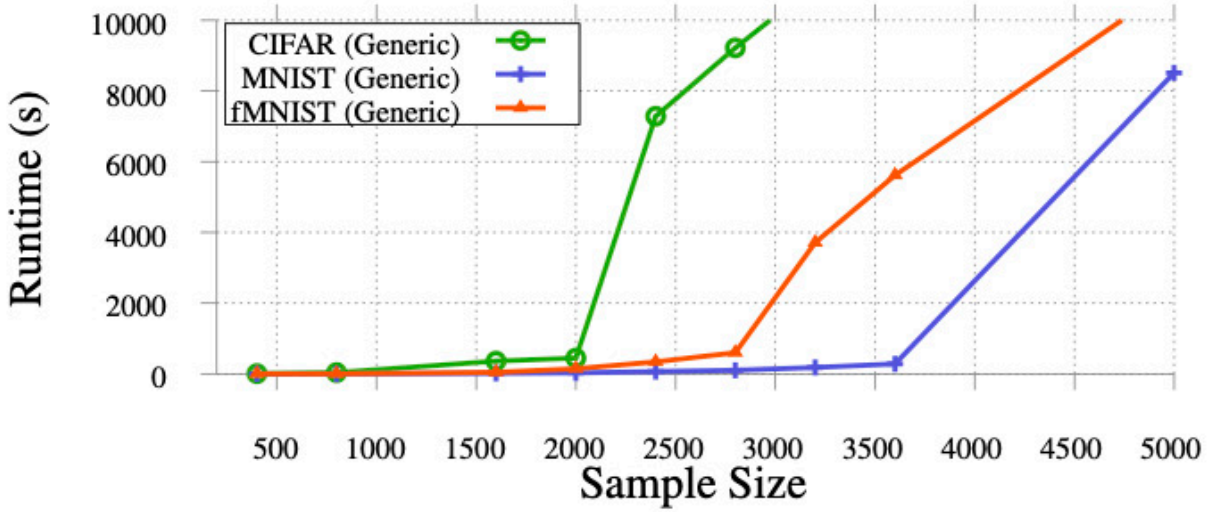
- Data (in $\mathbb{R}^d$) is drawn from two classes (1 and -1), with equal sampling probability for each point
- Attacker perturbs data in ℓ_2 ball around each datapoint
- **Goal**: Find the minimum cross-entropy loss achievable by **any classifier**

Graph representation and solution

-Probability on vertices represents classifier output
-Edges represent overlapping perturbation balls and
-Enforce constraints on the convex minimization problem
-Intersection of polytope and loss surface gives correct classification probs.

Efficiently computing lower bounds

Efficiently computing lower bounds

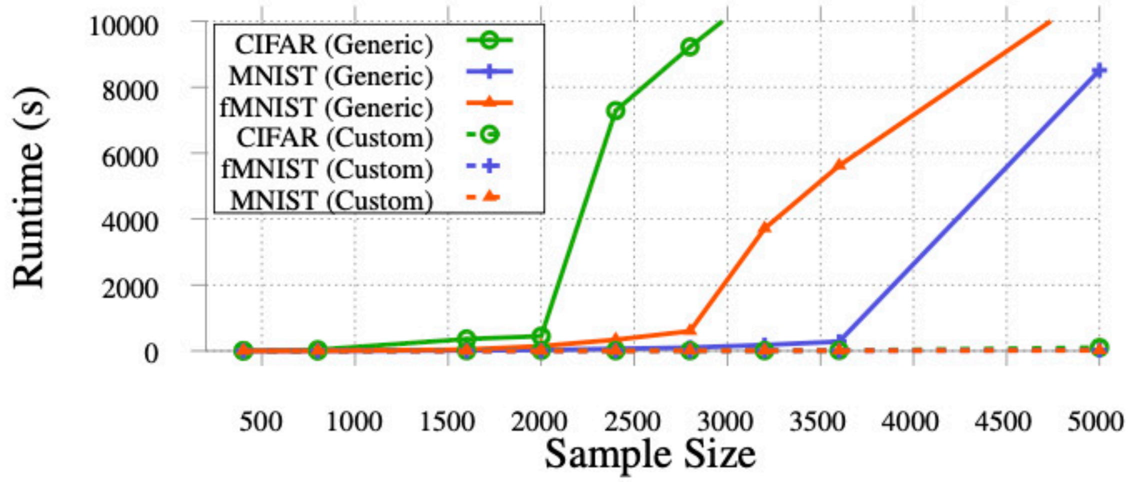


Generic convex solver

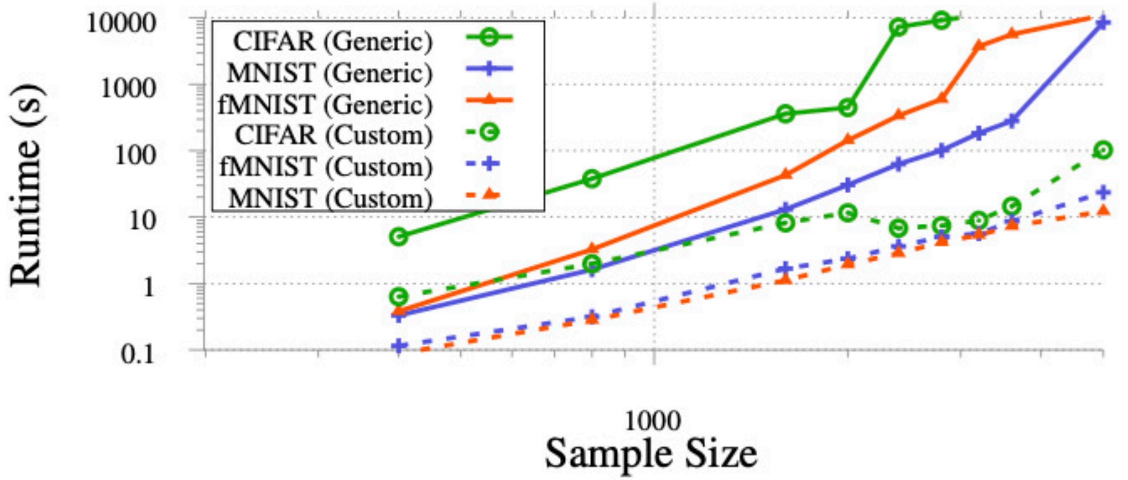
- Tractable in theory, but too slow in practice (~13 hours for complete 2-class CIFAR-10)

Efficiently computing lower bounds

Linear-linear



Log-log



Generic convex solver

- Tractable in theory, but too slow in practice (~13 hours for complete 2-class CIFAR-10)

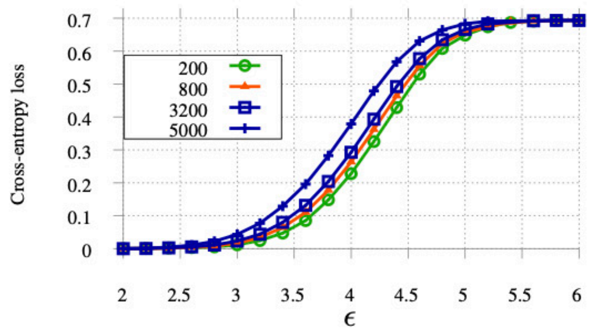
Custom algorithm

- Simultaneously finds both the optimal classifier (primal) and attack (dual)
- Achieves **1000x** speed-up by
  - iteratively splitting graph into portions where probs. are over/under-estimated
  - Utilizing the bipartite graph structure
- Enables the computation of lower bounds in a vast range of settings

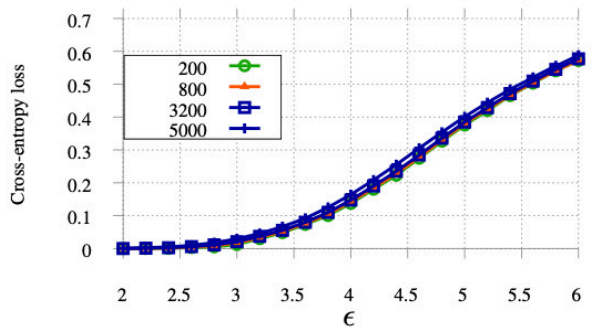
Comparing optimal and empirical CE loss

Comparing optimal and empirical CE loss

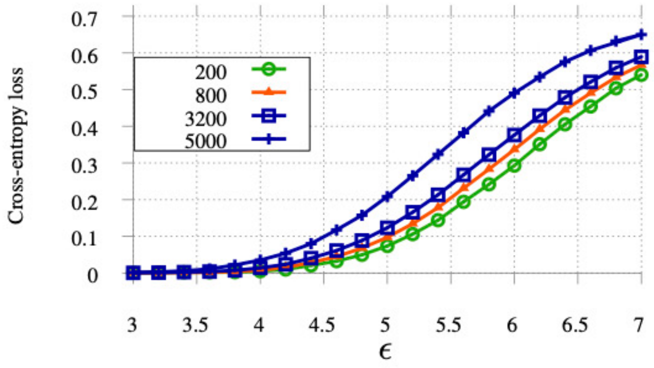
Optimal cross-entropy loss



MNIST

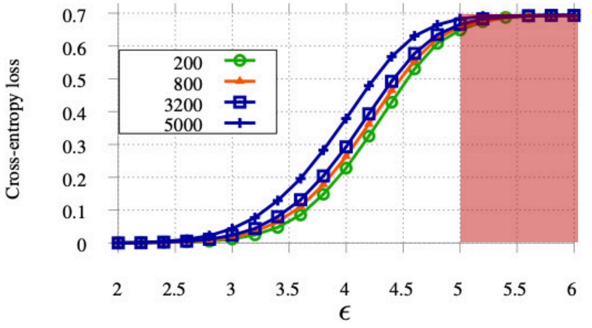


Fashion MNIST

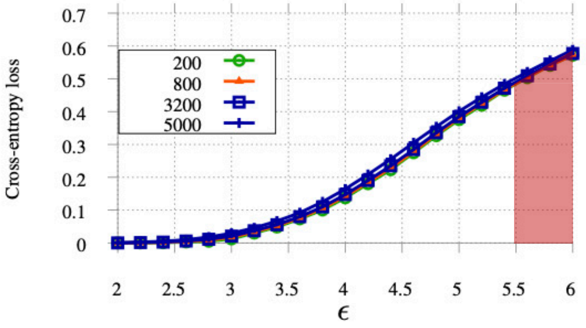


CIFAR-10

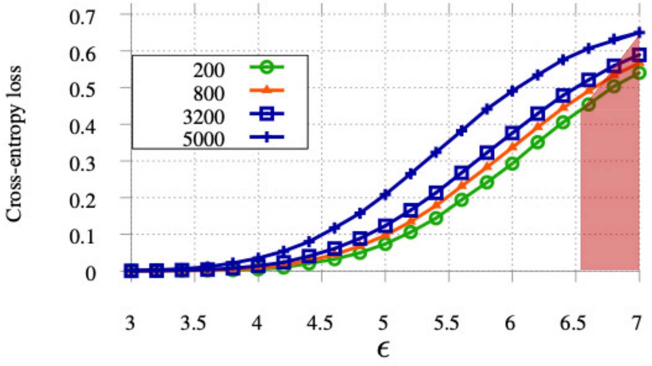
Comparing optimal and empirical CE loss



MNIST



Fashion MNIST

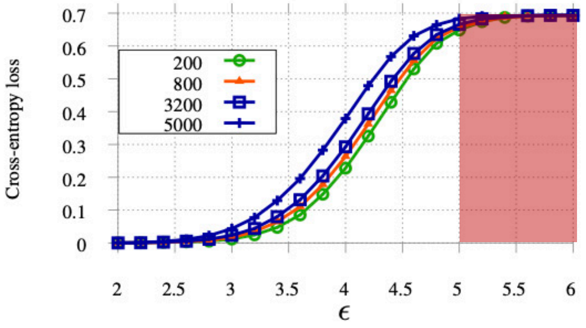


CIFAR-10

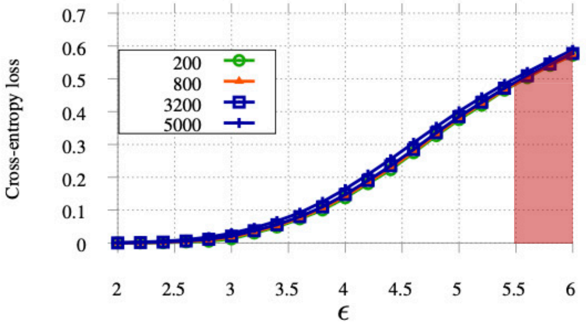
Optimal cross-entropy loss

- Identifies regimes for each dataset where the 2-class robust classification problem is challenging/impossible

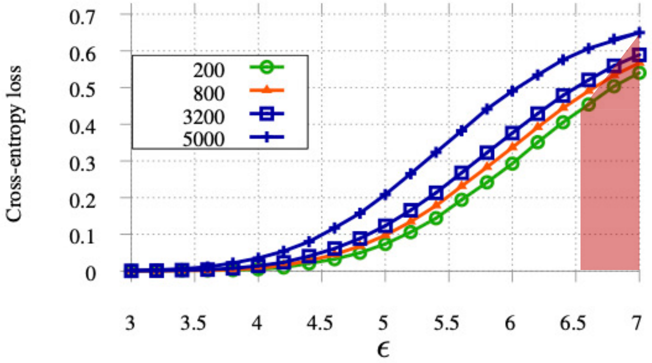
Comparing optimal and empirical CE loss



MNIST



Fashion MNIST

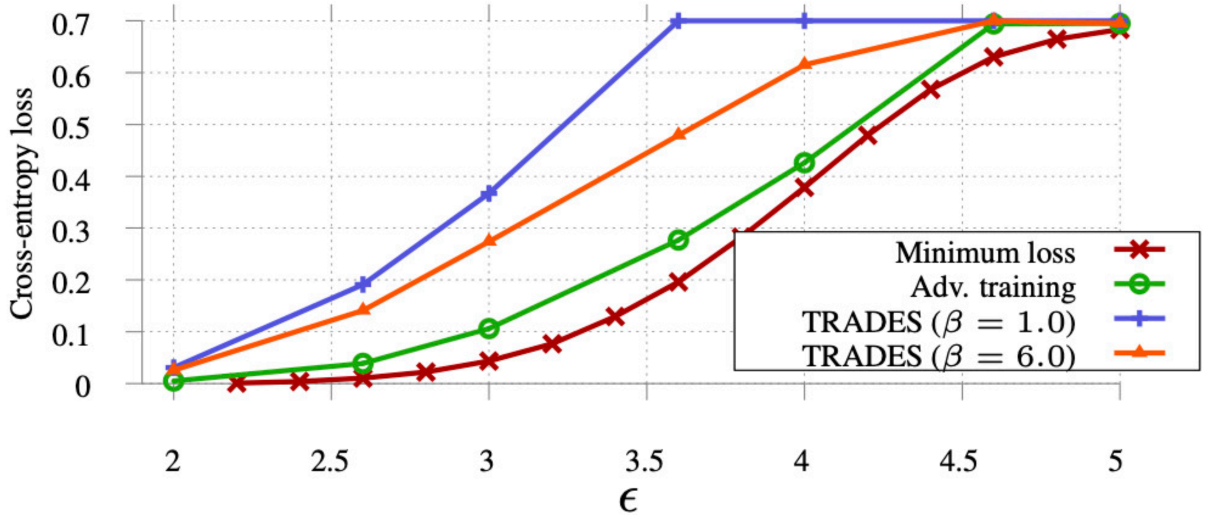


CIFAR-10

Optimal cross-entropy loss

- Identifies regimes for each dataset where the 2-class robust classification problem is challenging/impossible
- Bound increases with number of samples

Comparing optimal and empirical CE loss



MNIST

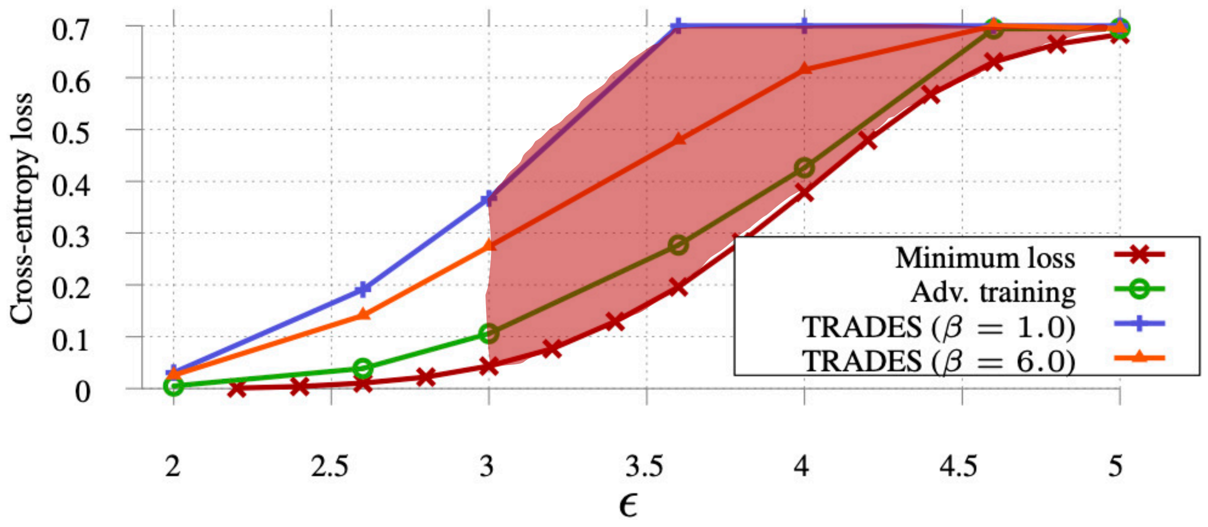
Optimal cross-entropy loss

- Identifies regimes for each dataset where the 2-class robust classification problem is challenging/impossible
- Bound increases with number of samples

Comparing to empirical

- Current robust training is close to optimal (w.r.t strong empirical attack) at lower budgets

Comparing optimal and empirical CE loss



MNIST

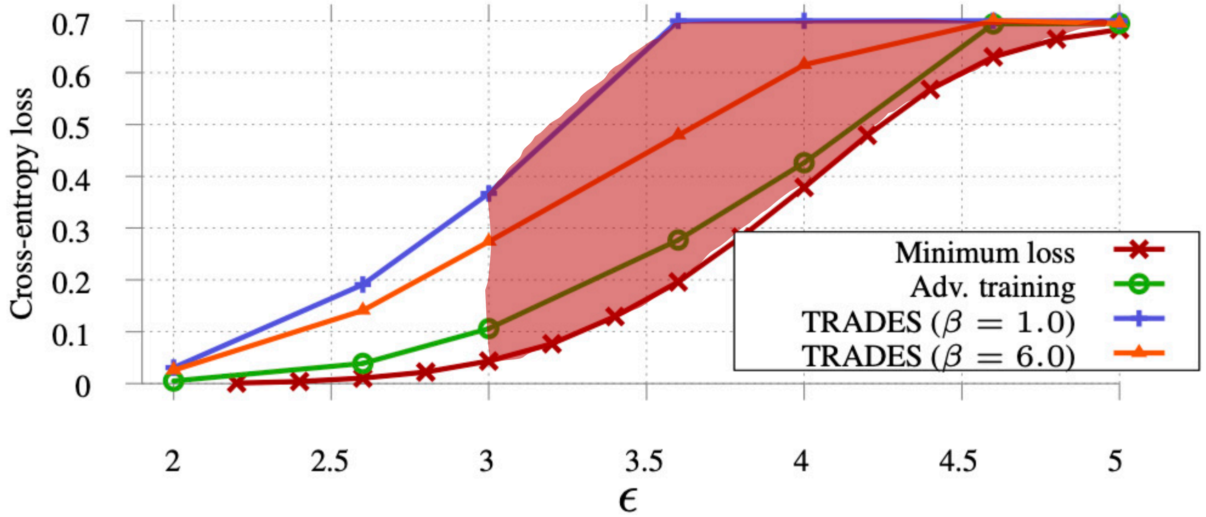
Optimal cross-entropy loss

- Identifies regimes for each dataset where the 2-class robust classification problem is challenging/impossible
- Bound increases with number of samples

Comparing to empirical

- Current robust training is close to optimal (w.r.t strong empirical attack) at lower budgets
- Gap exists between the empirical loss of a robustly trained classifier and optimal one at higher budgets

Comparing optimal and empirical CE loss



MNIST

Optimal cross-entropy loss

- Identifies regimes for each dataset where the 2-class robust classification problem is challenging/impossible
- Bound increases with number of samples

Comparing to empirical

- Current robust training is close to optimal (w.r.t strong empirical attack) at lower budgets
- Gap exists between the empirical loss of a robustly trained classifier and optimal one at higher budgets
- Closing the gap and its impact on generalization is an **open question**

Paper:
https://arxiv.org/abs/2104.08382

Code:
https://github.com/arjunbhagoji/log-loss-lower-bounds