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

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

THE UNIVERSITY OF CHICAGO · PennState · PRINCETON UNIVERSITY

Formulation: Learning with a Test-time Adversary

(x, y)	Labeled natural examples in $\mathcal{X} \times \{-1, 1\}$
$\tilde{x}$	Adversarial example in $\mathcal{X}$
$N(\cdot)$	Neighborhood constraint function for adversary, i.e. $\tilde{x} \in N(x)$
P	Distribution of labeled examples (on $\mathcal{X} \times \{-1, 1\}$)
h	Soft classifier with $h(x) \in [0, 1]^{\{-1,1\}}$

Learner:
Receives i.i.d. labeled training data $((x_1, y_1), \ldots, (x_n, y_n)) \sim P^n$ and selects $\hat{h}$

Test-time adversary:
Receives labeled natural example $(x_{\text{Test}}, y_{\text{Test}}) \sim P$ and selects $\tilde{x} \in N(x_{\text{Test}})$

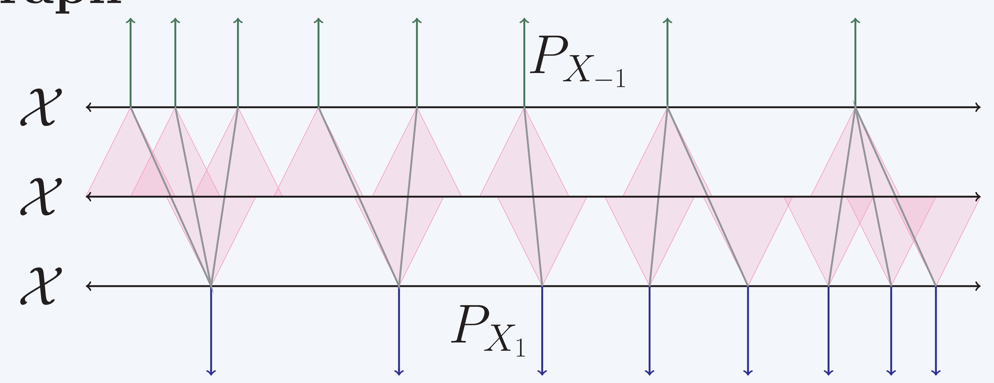
Performance metric:
$L^{\text{CE}}(h, N, P) = \mathbb{E}_P \left[\sup_{\tilde{x} \in N(x)} \ell^{\text{CE}}(h, (\tilde{x}, y_{\text{Test}}))\right]$ (Cross-entropy loss of h on adversarial examples)

Problem Statement:
What is $\inf_h L^{\text{CE}}(h, N, P)$ for all classifiers h, a 2-class discrete distribution P and adversary constrained to $N(\cdot)$?

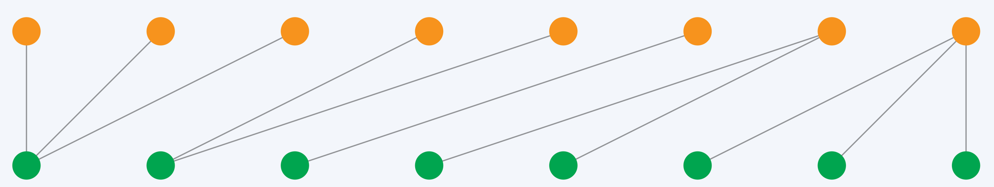
Conflict graph and feasible outputs

Constructing the conflict graph

For discrete distributions of examples P_{X_1} and $P_{X_{-1}}$, construct a bipartite graph recording neighborhood intersection information:



Conflict graph $\mathcal{G}$, with vertex set $\mathcal{V}$ and edge set $\mathcal{E}$



Determining set of feasible output probabilities

Conflict graph $\mathcal{G}$ allows for the determination of polytope of feasible output probabilities

Lemma

Let $q \in \mathbb{R}^{\mathcal{V}}$ be the vector of correct-classification probabilities obtained by a classifier. The feasible set of such probabilities is

$$q \geq \mathbf{0}$$
$$Mq \leq \mathbf{1}.$$

where $M = \begin{pmatrix} E \\ I \end{pmatrix} \in \mathbb{R}^{(\mathcal{E} \sqcup \mathcal{V}) \times \mathcal{V}}$ *and* $E \in \mathbb{R}^{\mathcal{E} \times \mathcal{V}}$ *is the edge incidence matrix of the conflict graph.*

Takeaways:

- This is the fractional vertex packing polytope of the graph
- Larger budgets lead to more intersections and a smaller feasible set
- Construction of the conflict graph and associated polytope can be applied to any possible adversarial constraint, including standard ℓ_p ball constraints

Establishing lower bounds

Theorem

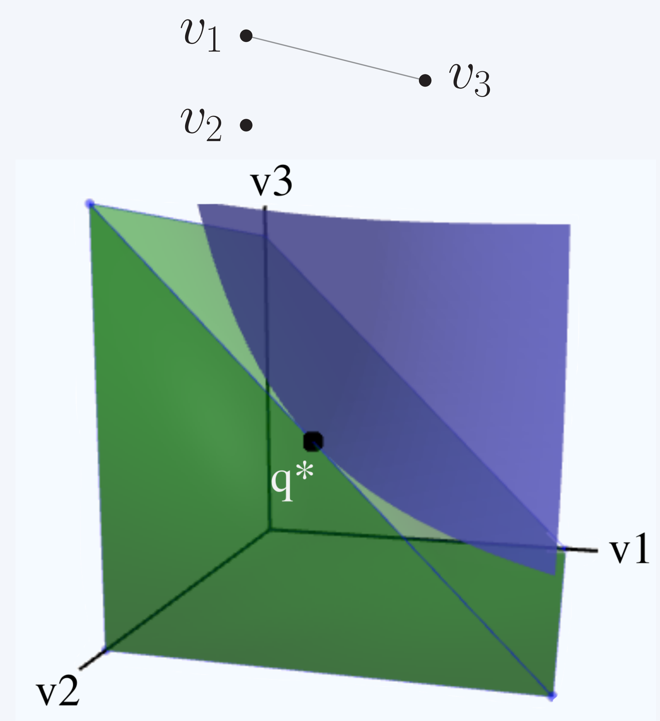
The discrete joint probability distribution P over data from two classes, and the neighborhood function $N(\cdot)$ define a bipartite conflict graph $\mathcal{G}$ with incidence matrix E. Let $p \in \mathbb{R}^{\mathcal{V}}$ with $p_v = P(\{v\})$. Let q^* be the minimizer of the following program:

$$\min_q \sum_{v: p_v > 0} -p_v \log q_v \quad s.t.\ q \geq \mathbf{0}$$
$$Mq \leq \mathbf{1}.$$

Then, there is a classifier h^* that achieves the correct-classification probabilities q^* and for all h, $\mathbb{E}_P[\tilde{\ell}^{CE}(h^*, v)] \leq \mathbb{E}_P[\tilde{\ell}^{CE}(h, v)]$.

Takeaways:

- Lower bound is achievable: the solution provides optimal classification probabilities
- Solution to the dual convex problem is the optimal adversarial strategy
- Applies to all discrete distributions, including popular vision datasets such as CIFAR-10 etc.

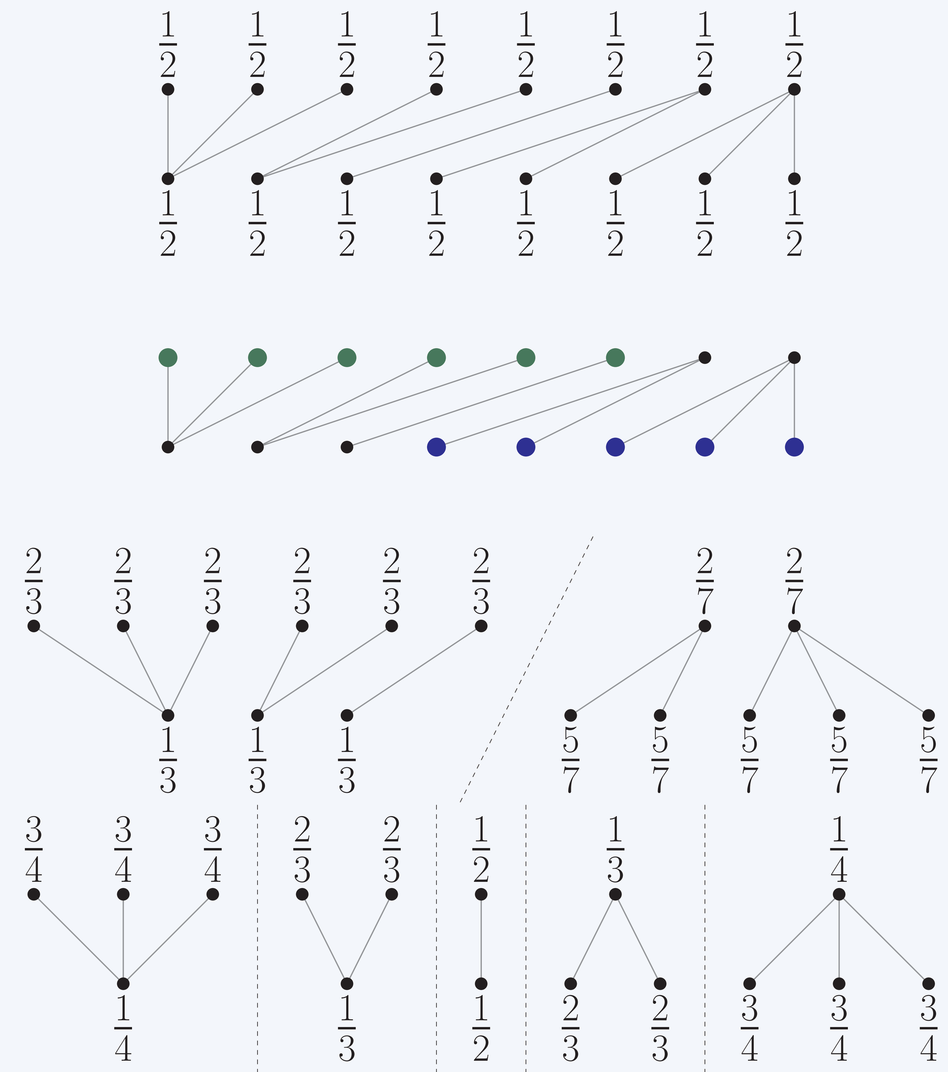


Efficient computation of lower bounds

- Lower bound can be computed using a generic convex solver, but too slow in practice (13 hours for complete 2-class CIFAR-10)
- Custom algorithm achieves $1000\times$ speed-up by exploiting the bipartite graph structure
- Enables the computation of lower bounds in a vast range of settings

Algorithm sketch with an example:

1. Initial guess for correct classification probabilities comes from global class frequencies
2. Maximum weight independent set identifies vertices where current probabilities are too low
3. Split into subproblems and obtain next guess from class frequencies within the subproblems
4. Recursion stops when the max-weight independent sets are the classes

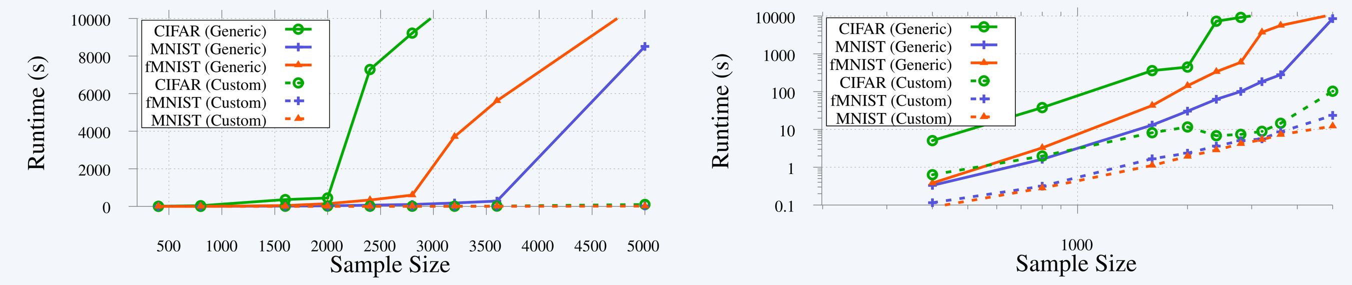


Speedups from custom algorithm

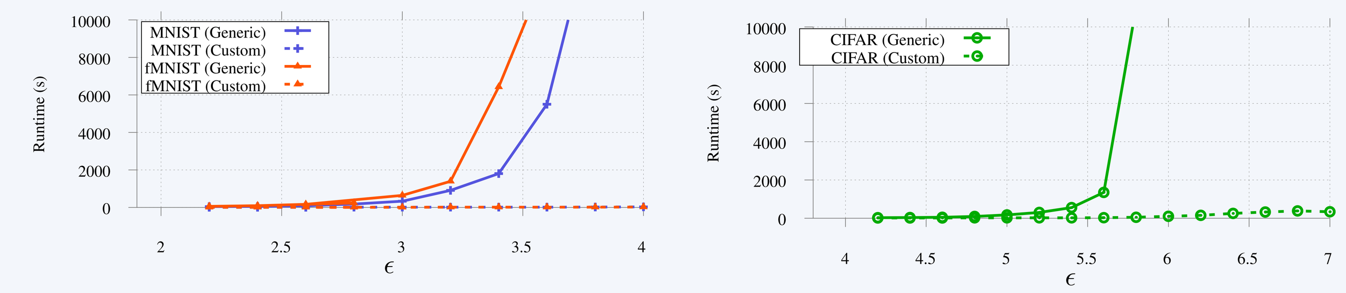
Setup and parameters:

- All results are for an ℓ_2 adversary
- Generic solver is the non-linear general purpose convex solver from CVXOPT

Variation in runtime with number of samples:



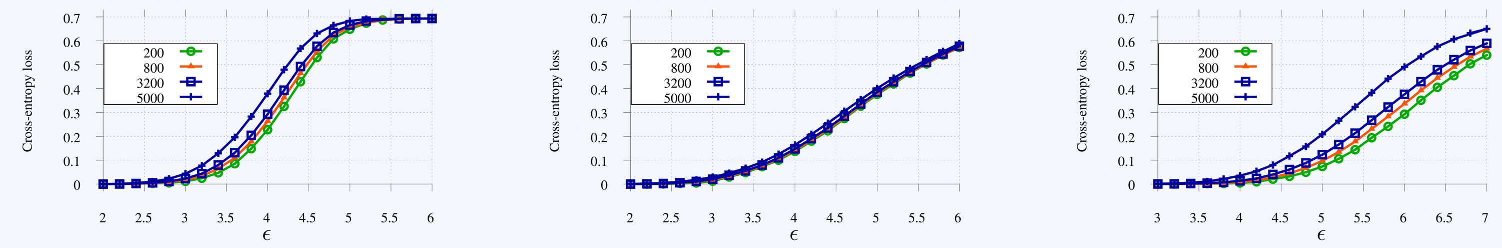
Variation in runtime with adversary's budget:



Comparing empirical and optimal lower bounds

Optimal lower bounds:

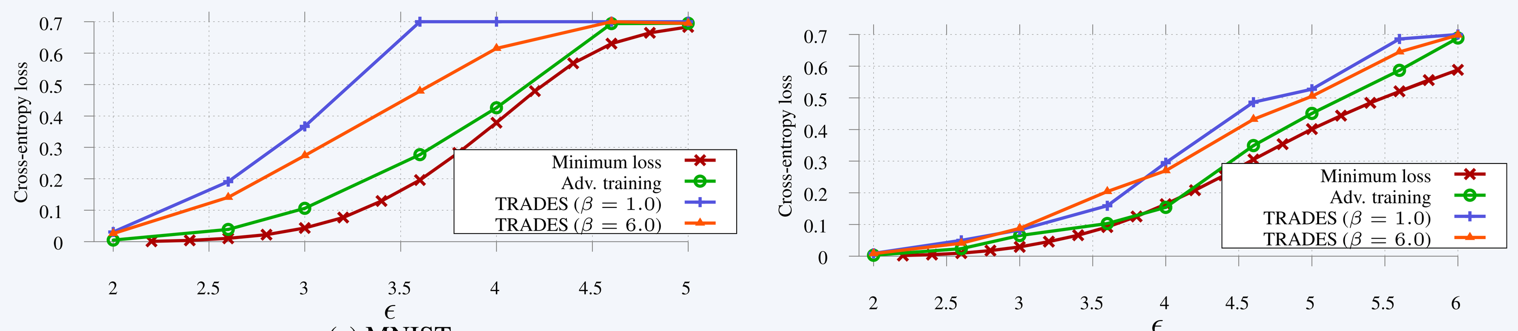
- 3 vs. 7 binary classification problem for each dataset
- Optimal lower bounds obtained using custom algorithm
- Lower bound *increases* with the number of samples



(a) MNIST (b) Fashion MNIST (c) CIFAR-10

How effective is robust training?

- ResNet-18 is trained for each dataset using Projected Gradient Descent-based adversarial training and TRADES
- Attack strength is empirically evaluated using AutoAttack



(a) MNIST (b) Fashion MNIST

References

1. Madry et. al., *Towards deep learning models resistant to adversarial attacks*, ICLR 2018
2. Zhang et. al., *Theoretically principled trade-off between robustness and accuracy*, ICML 2019
3. Croce and Hein, *Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks*, ICML 2020