Toward Efficient Robust Training against Union of Lp Threat Models

Gaurang Sriramanan Maharshi Gor Soheil Feizi

Adversarial Vulnerability

$+ 0.02 \quad *$

$=$

Prediction: **Hamster**

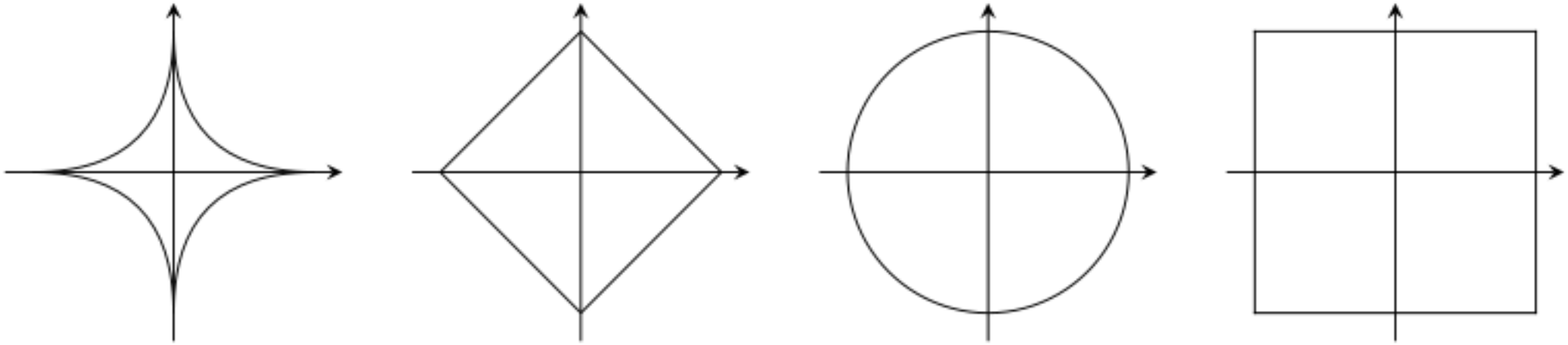
Confidence = 99.99%

50-step PGD targeted attack with $\varepsilon = \frac{8}{255}$ scaled by 50x

Prediction: **Banjo**

Confidence = 100%

Lp Norm Threat Models



$p = \frac{1}{2}$ $p = 1$ $p = 2$ $p = \infty$

$$|\mathbf{x}|_p \equiv \left(\sum_i |x_i|^p \right)^{1/p}$$

L-infinity Adversarial Attack

$+ 0.02 \quad *$

$=$

Prediction: **Hamster**

Confidence = 99.99%

50-step PGD targeted attack with $\varepsilon = \frac{8}{255}$ scaled by 50x

Prediction: **Banjo**

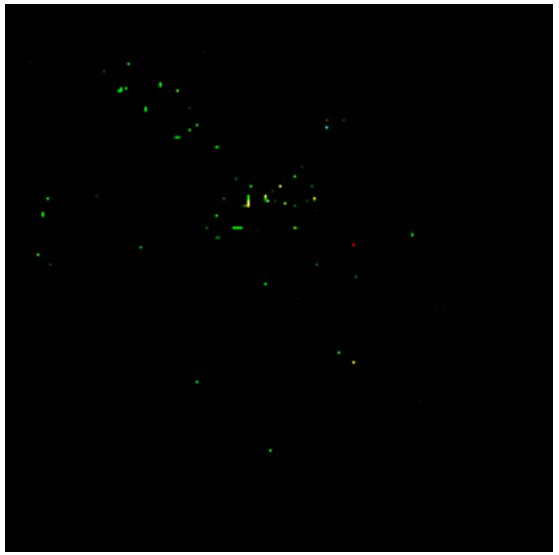
Confidence = 100%

L1 Adversarial Attacks

Prediction: **Hamster**

Confidence = 99.99%

$+ 0.02 \quad *$



50-step PGD targeted attack
with $\varepsilon = 17$ scaled by 50x

$=$

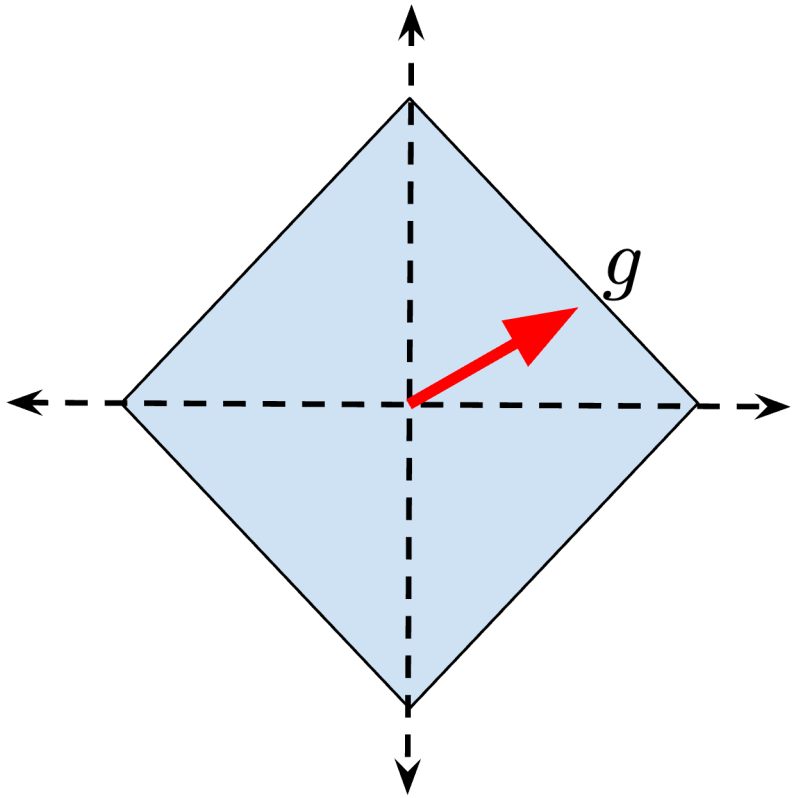
Prediction: **Banjo**

Confidence = 60.94%

Challenges in Robust Training

- L-inf robust models are vulnerable to L1 attacks and vice versa
- To achieve robustness against the union of threat models, prior works either use:
  - Large number of attack steps for different adversaries
  - Fine-tune existing robust models
- For L1 robustness, even certain multi-step adversarial training methods susceptible to catastrophic failure

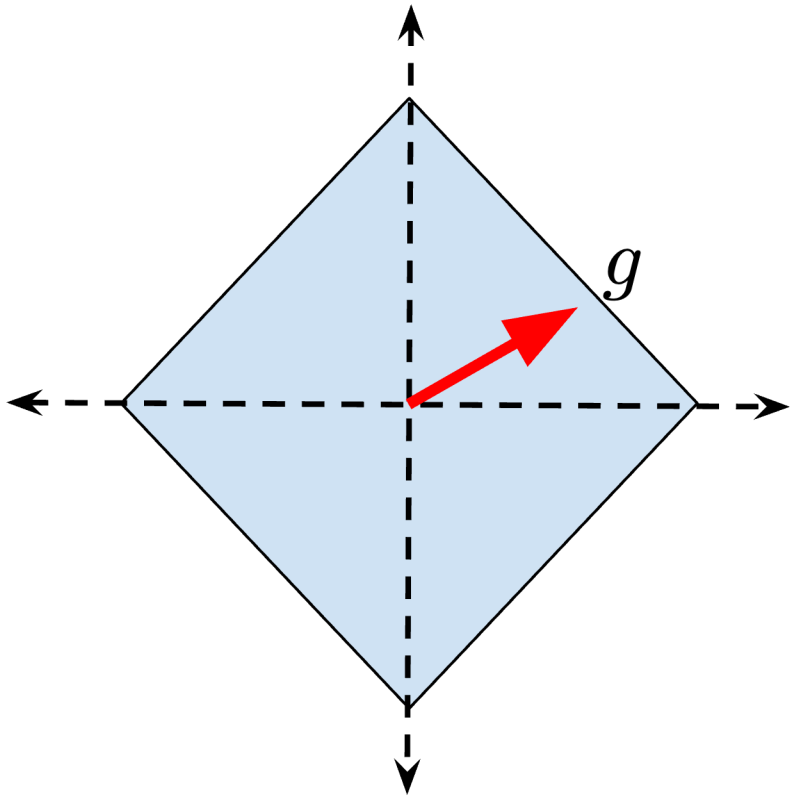
Steepest Ascent in L1 Geometry

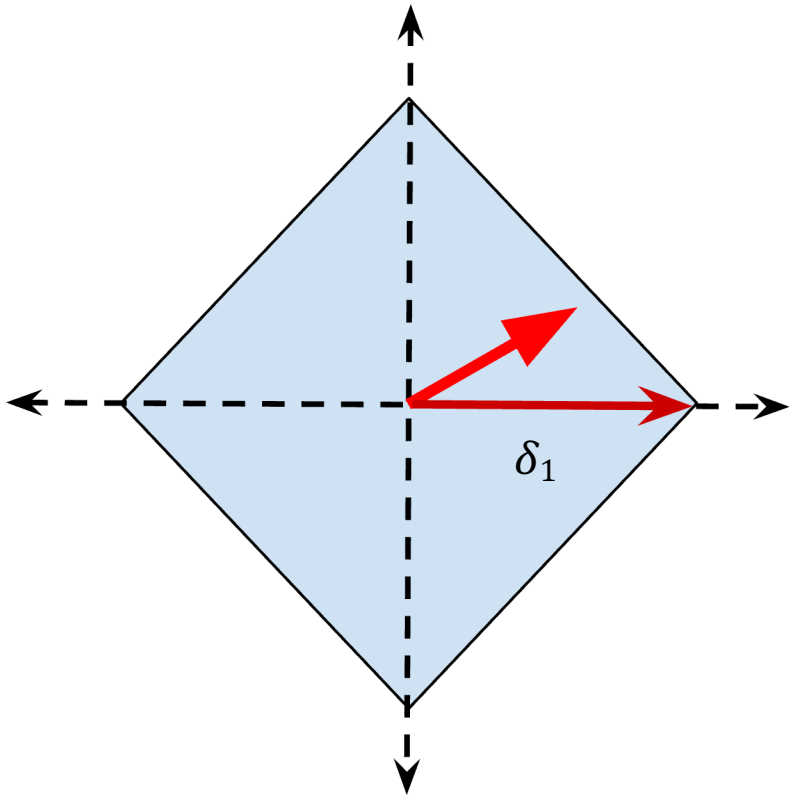


$$\max_{\delta} \left[\sum_{i=1}^{d} g_i \delta_i \right] \quad \text{such that}$$

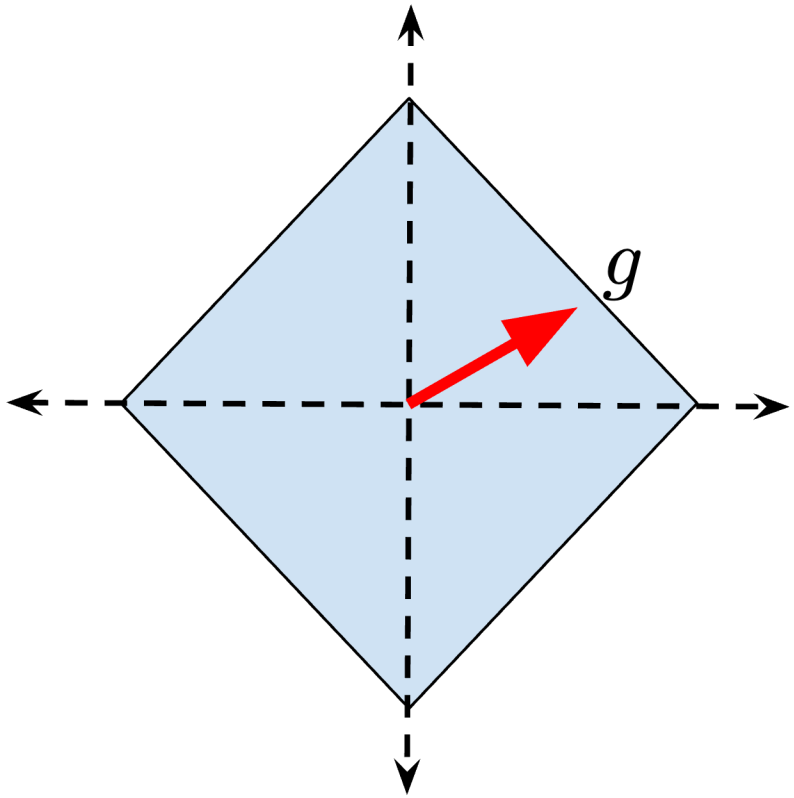
$$\text{(a) } 0 \leq x_i + \delta_i \leq 1 \; \forall \; i, \text{ and (b) } ||\delta||_1 \leq \epsilon_1$$

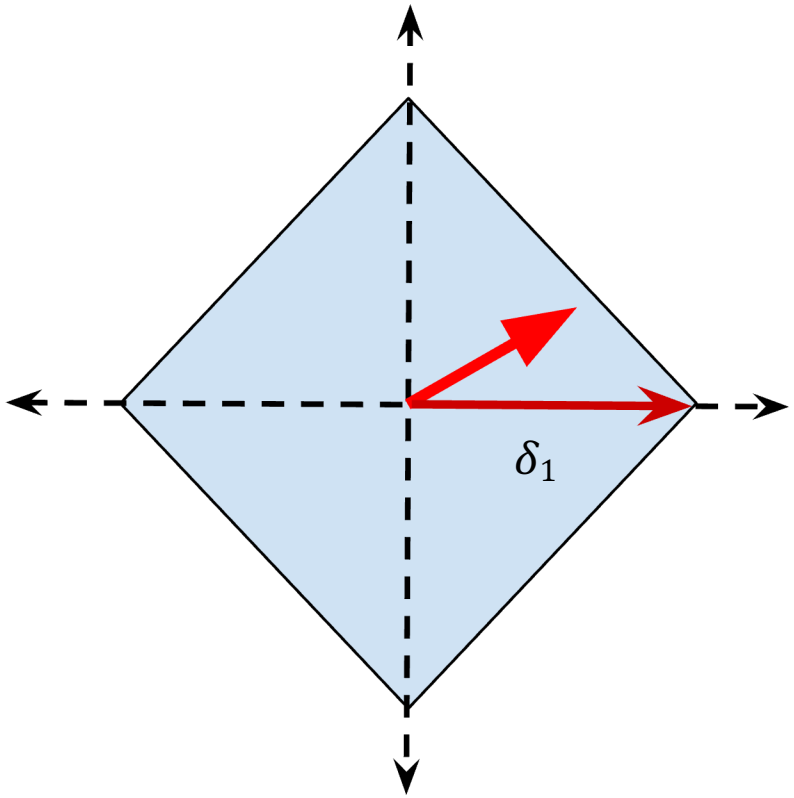
Steepest Ascent in L1 Geometry

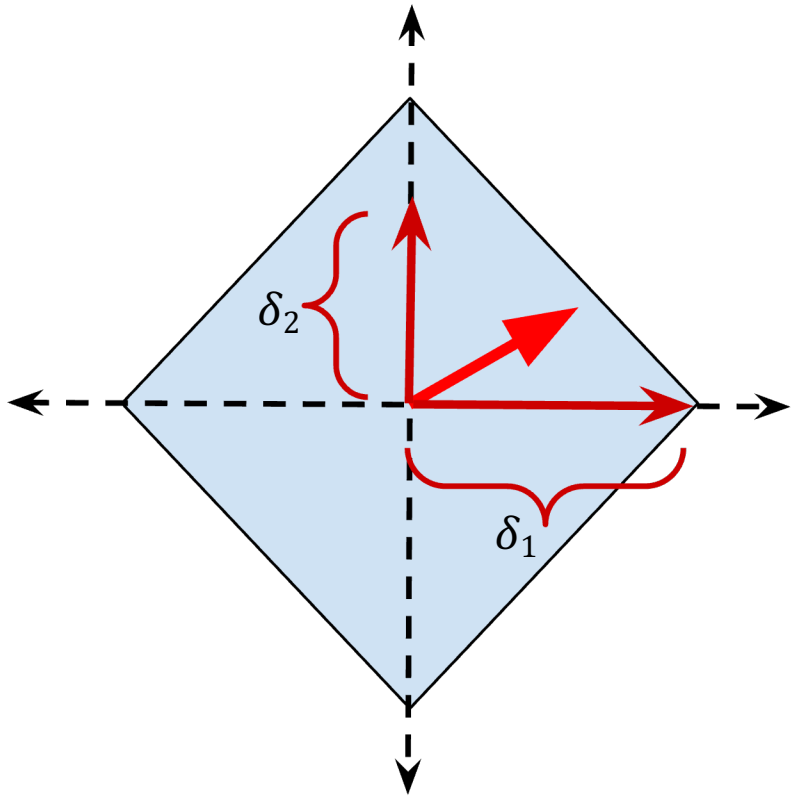




Steepest Ascent in L1 Geometry



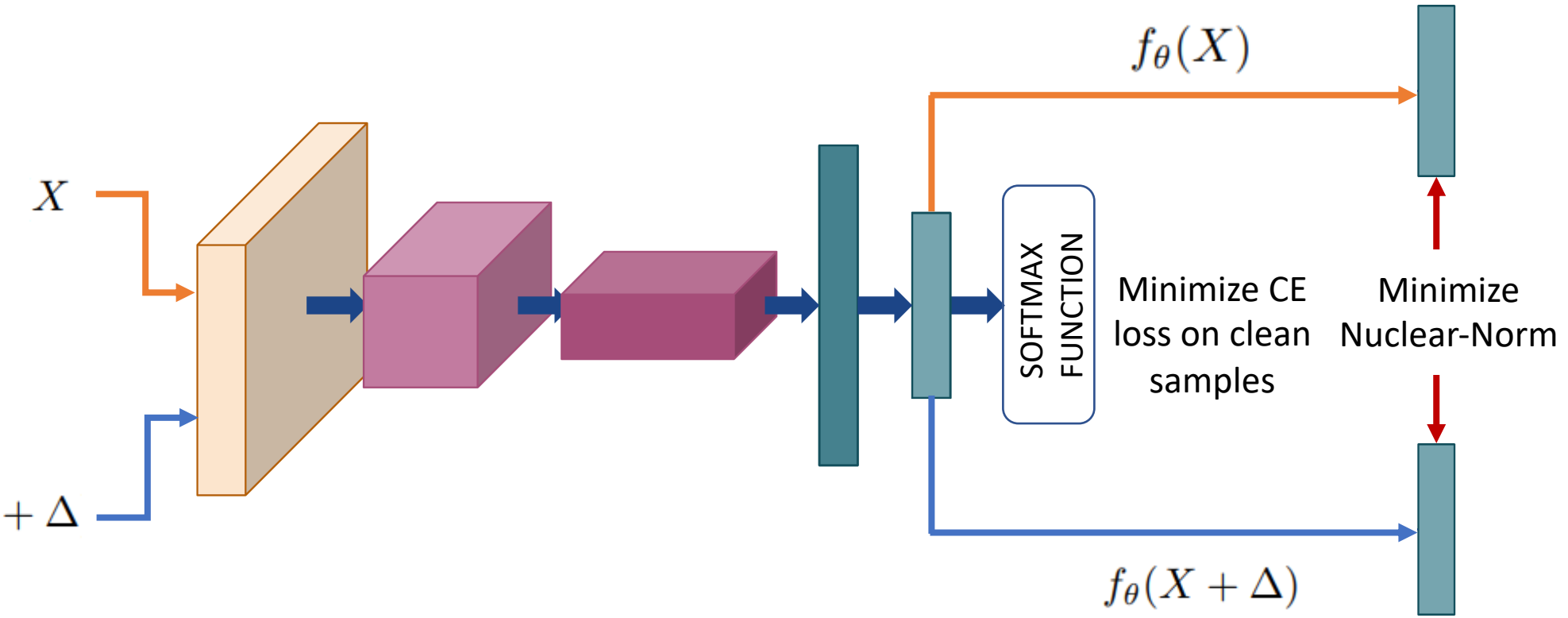




Nuclear Norm Regularization

$$X = \begin{bmatrix} \dots & x_1 & \dots \\ \dots & \vdots & \dots \\ \dots & x_M & \dots \end{bmatrix}$$

$$\Delta = \begin{bmatrix} \dots & \delta_1 & \dots \\ \dots & \vdots & \dots \\ \dots & \delta_M & \dots \end{bmatrix}$$

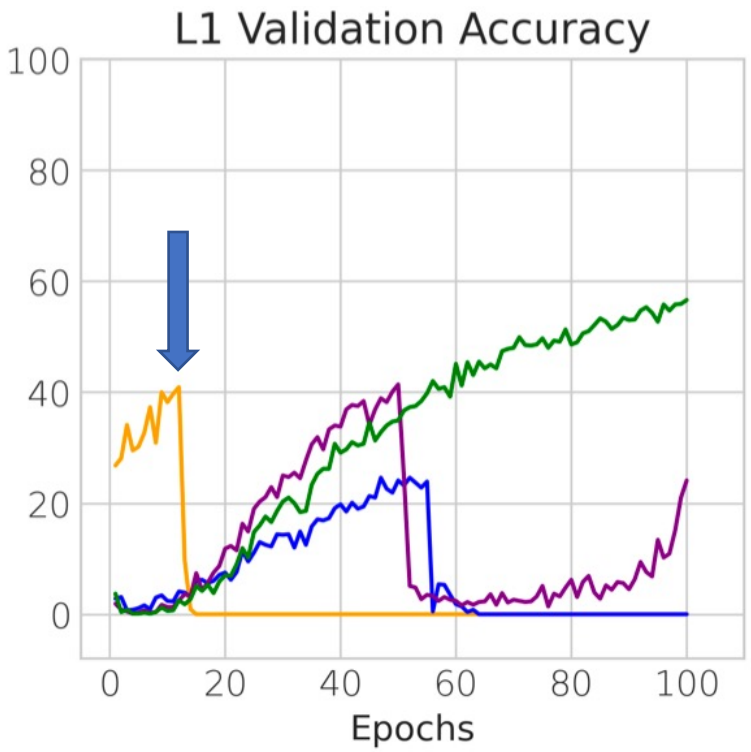


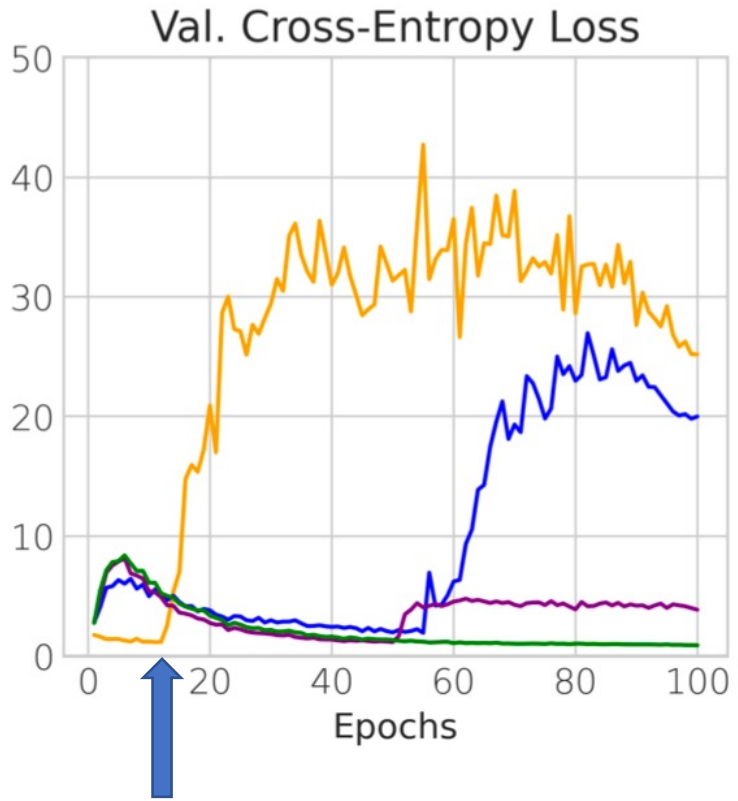
$$L = \ell_{CE}(f_\theta(X), Y) + \lambda \cdot ||f_\theta(\widetilde{X}) - f_\theta(X)||_*$$

Curriculum Scheduling

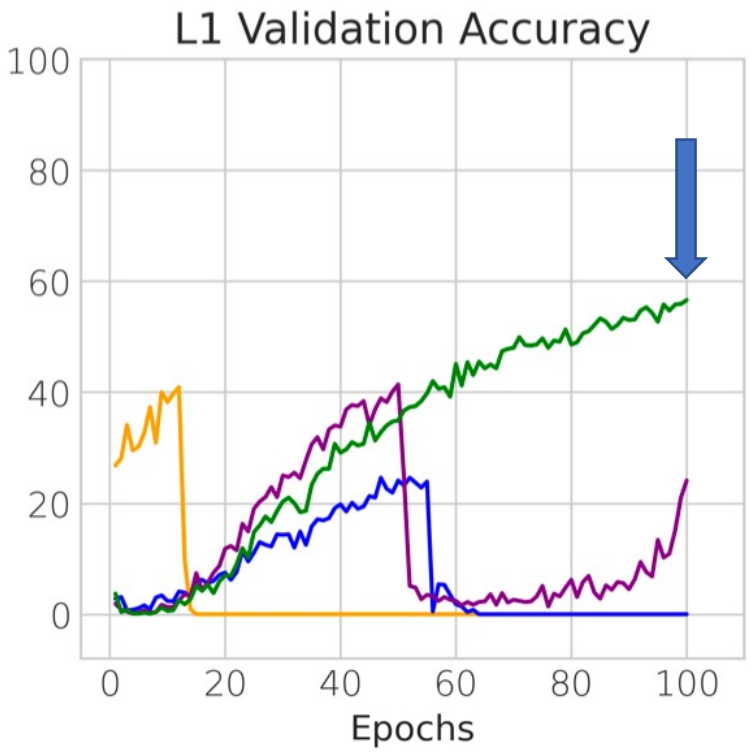
- In practice, NuAT on its own is not stable enough for L1 training
- We propose to use a Curriculum Schedule to select the nature of perturbations during L1 training
- To achieve robustness against the union of threat models, we propose to use a Decision function to select adversary generation
- Maintains low compute requirement: single-step attack per minibatch

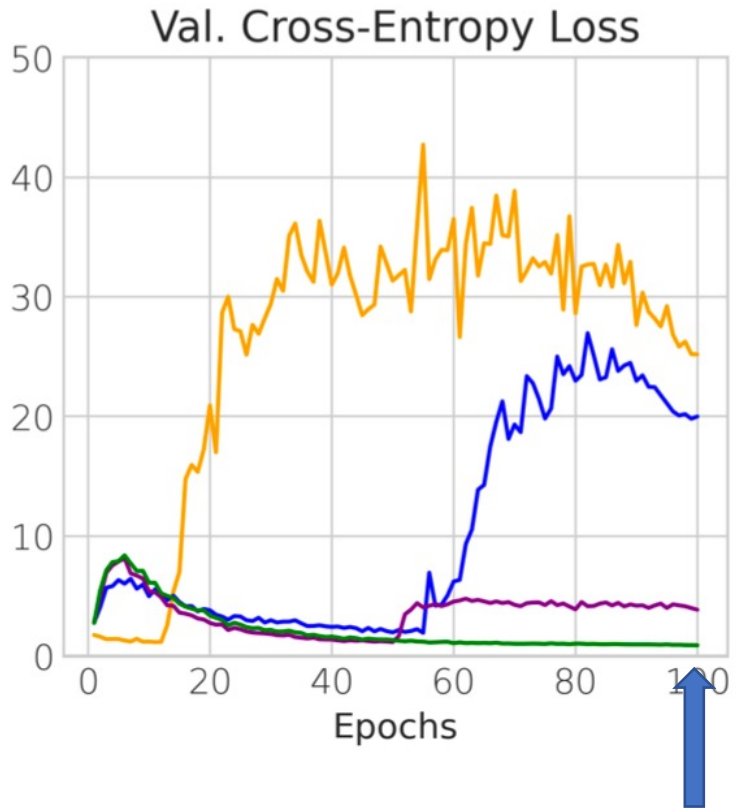
Catastrophic Overfitting in L1 Training





Stabilized L1 Training with NCAT





R-FGSM (Val.)
R-FGSM-C (Val.)
NuAT (Val.)
NCAT (Val.)

ResNet-18 Results on CIFAR-10

Evaluations on Threat models with constraint sets: $\boldsymbol{\varepsilon}_1 = 12$, $\boldsymbol{\varepsilon}_2 = 0.5$ and $\boldsymbol{\varepsilon}_\infty = 8/255$

Method	Number of AT Steps	Clean Acc	Worst-Case Acc	Average Acc	ℓ_1 Acc	ℓ_2 Acc	ℓ_∞ Acc
ℓ_1 Training Alone							
APGD-ℓ_1	10	85.9	22.1	48.8	59.5	64.9	22.1
NCAT-ℓ_1	1	81.1	37.9	53.6	55.9	67.0	38.0
Training under Union of Threat Models							
SAT	13.33†	83.9	40.4	54.2	54.0	68.0	40.7
AVG	30	84.6	40.1	53.8	52.1	68.4	40.8
MAX	30	80.4	44.0	53.4	48.6	66.0	45.7
MSD	50	81.1	43.9	53.4	49.5	65.9	44.9
EAT	10††	82.2	42.4	54.6	53.6	67.5	42.7
NCAT	1	80.3	42.6	53.3	46.9	67.0	46.0
NCAT$^+$	1	77.5	43.7	53.4	48.4	65.7	46.1

Stability on Large Networks - WideResNet

Evaluations on Threat models with constraint sets: $\boldsymbol{\varepsilon}_1 = 12$, $\boldsymbol{\varepsilon}_2 = 0.5$ and $\boldsymbol{\varepsilon}_\infty = 8/255$

Method	**Number of AT Steps**	**Clean Acc**	**Worst-Case Acc**	**Average Acc**	ℓ_1 **Acc**	ℓ_2 **Acc**	ℓ_∞ **Acc**
ℓ_1 Training Alone							
APGD-ℓ_1	10	83.7	30.7	52.5	61.6	65.1	30.7
NCAT-ℓ_1	1	80.7	39.2	54.6	56.1	68.6	39.3
Training under Union of Threat Models							
SAT	13.33†	80.5	45.7	56.2	55.9	66.7	45.9
AVG	30	82.5	45.1	56.1	55.0	68.0	45.4
MAX	30	79.9	47.4	54.6	50.2	65.3	48.4
MSD	50	80.6	46.9	55.1	51.7	65.6	48.0
EAT	10††	79.9	46.4	56.3	56.0	66.2	46.6
NCAT	1	81.5	44.6	54.8	49.9	68.3	46.3

Results on ImageNet-100

Evaluations on Threat models with constraint sets: $\boldsymbol{\varepsilon}_1 = 255$, $\boldsymbol{\varepsilon}_2 = 1200/255$ and $\boldsymbol{\varepsilon}_\infty = 4/255$

Method	**Number of AT Steps**	**Arch**	**Clean Acc**	**Worst-Case Acc**	ℓ_1 **Acc**	ℓ_2 **Acc**	ℓ_∞ **Acc**	**PPGD Acc**
ℓ_∞-AT	10	RN50	81.7	0.8	0.8	3.7	55.7	1.5
PAT	10	RN50	72.6	37.8	41.2	37.7	45.0	29.2
NCAT-ℓ_1	1	RN18	64.9	41.1	48.3	41.4	42.1	26.6
NCAT	1	RN18	63.9	41.5	46.8	41.9	45.7	29.1

Summary

- Successfully achieves robustness against L1 adversaries in an efficient manner
- Extends to robust training against union of threat models
- NCAT requires only a single step attack for multiple threat models
- Generalizes to unseen threat models, even to Perceptual Projected Gradient Descent (PPGD) attack

Thank you!


NEURAL INFORMATION
PROCESSING SYSTEMS


UNIVERSITY OF
18
56
MARYLAND