Cooperation or Competition: Avoiding Player Domination for Multi-target Robustness by Adaptive Budgets

Yimu Wang (University of Waterloo), Dinghuai Zhang (Mila & University of Montreal), Yihan Wu (University of Pittsburgh), Heng Huang (University of Pittsburgh), Hongyang Zhang (University of Waterloo)

June, 2023

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

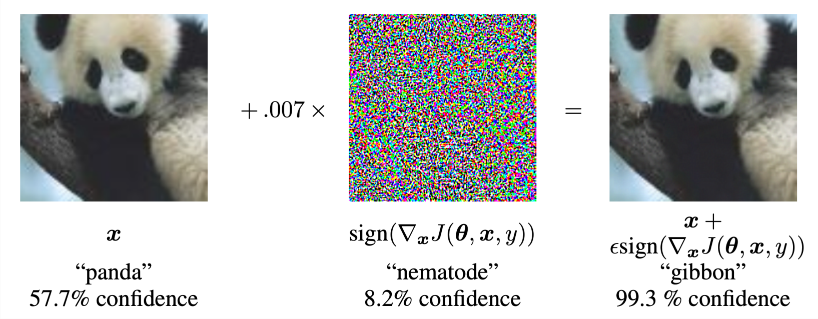
Table of contents

1. Background
2. Our Analysis
3. Our Method
4. Experimental Results
5. Conclusion

Background

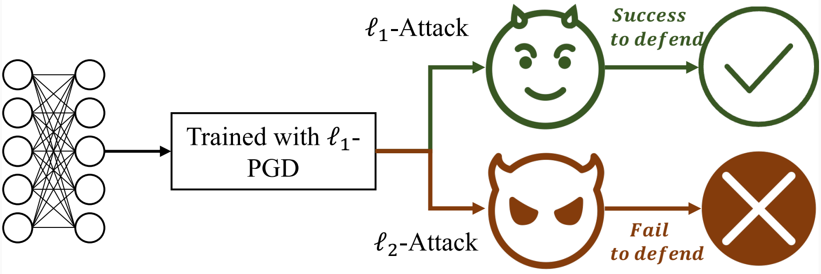
Machine learning models are susceptible to adversarial examples

Figure 1: Example of adversarial examples. Image credit [2].



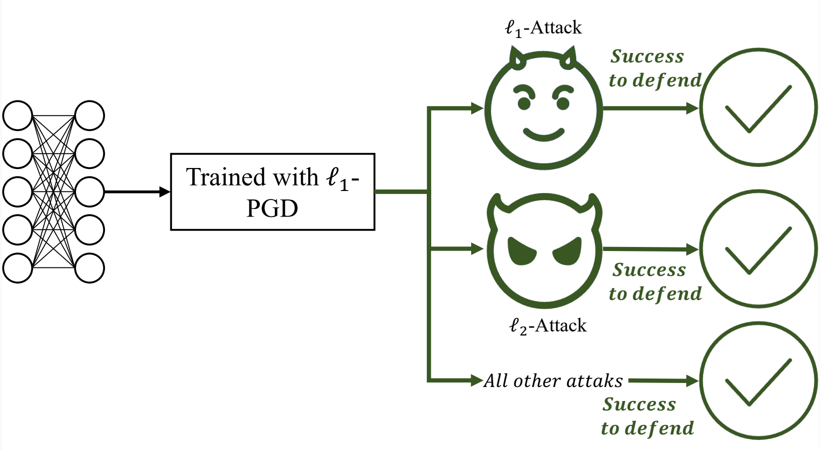
Most of the existing defenses are not universally robust

Figure 2: Most of the existing defenses are not universally robust and fail to defend against other adversaries [3, 4].

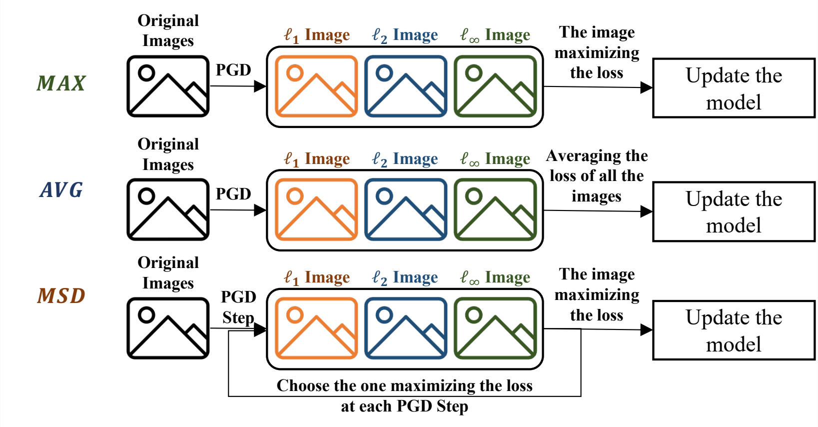


Ultimate goal of robustness

Figure 3: Targeting robustness against multiple adversaries simultaneously [1].



Previous Methods



Our Analysis

Theoretical Analysis on SVM

We first introduce the data distribution and the SVM model.

Data Distribution

Data $\mathbf{x}$ and label y are sampled as

$$y \overset{\text{u.a.r}}{\sim} \{+1, -1\}, \quad x_1 = \begin{cases} +y, & \text{w.p. } p; \\ -y, & \text{w.p. } 1-p, \end{cases} \qquad x_2, \ldots, x_{d+1} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu y, 1).$$

SVM Model

We train a linear SVM model $f_{\mathbf{w}}(\cdot)$ with soft-SVM loss on the data sampled as above:

$$\min_{\mathbf{w}} \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}} \sum_{p\in\{1,2,\infty\}} \gamma_p \max\left(0, 1 - y f_{\mathbf{w}}(\mathbf{x} + \delta(\mathbf{x})_p)\right),$$
$$\text{s.t. } \|\mathbf{w}\|_2 = 1,$$

where $f_{\mathbf{w}}(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$, $\delta_p(\mathbf{x})$ is the p-adversarial example for $\mathbf{x}$, and $\gamma = [\gamma_1, \gamma_2, \gamma_\infty]$ satisfies $\sum_{i\in\{1,2,\infty\}} \gamma_i = 1$.

Theoretical Analysis on SVM

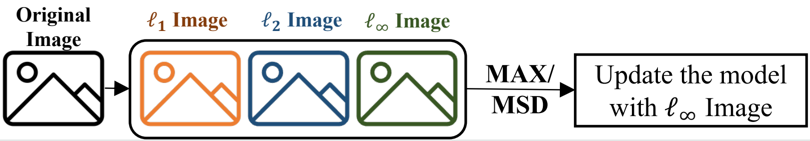
We first found that under the following case, there will be player domination.

Definition (Player dominates the cooperative game)

If $\exists i \in [k]$ such that $\gamma_i^t = 1$ and $\gamma_j^t = 0, \forall j \in [K]/\{i\}, \forall t$, then we call that player dominates the bargaining game.

ℓ_{infty} domination, Informal

Let $\mu \geq 4/\sqrt{d}$, $\epsilon_\infty \geq 2\mu$, $p \leq 0.977$, $\epsilon_\infty \geq \frac{2}{d}\epsilon_1$ and $\epsilon_\infty \geq \sqrt{\frac{2}{d}}\epsilon_2$. With MAX and MSD, ∞-player (∞-adversary) dominates the training procedure as shown below.



Theoretical Analysis on SVM

After analyzing the training dynamics of SVM, we notice that when the ∞-player dominates the bargaining game, and given $\epsilon_\infty > \mu$, the SVM model may not converge.

Theorem [Player domination makes the training procedure not converge, Informal]

With MAX and MSD, if ∞-player dominates and $\epsilon_\infty > \mu$, the **weights for the non-robust features flips over time**, *i.e.*, $\text{sign}(\mathbf{w}_i^t) = -\text{sign}(\mathbf{w}_i^{t-1}), \forall i \geq 2$.

Theoretical Analysis on Linear Model

Assuming the loss function of each player is denoted as $\ell_k, k \in [K]$, which is L-smooth and μ-strongly convex, we have the following theorems.

Theorem [MAX and MSD might not converge, Informal]

If the training is dominated by one player during the whole game, then the loss of all players and the overall loss would **increase** as time t **grows**.

Theorem [AVG's loss decreases, Informal]

Using AVG to train the linear model, the overall loss **decreases** as time t **grows**.

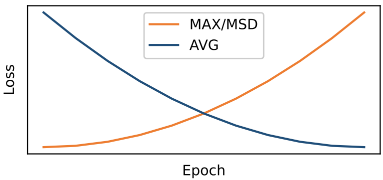


Figure 4: An example under the linear case.

Our Method

AdaptiveBudget

AdaptiveBudget is designed to **avoid the phenomenon of the same player dominating the whole training procedure** as this phenomenon leads to non-convergence under SVM and Linear cases.

Algorithm 1 Framework of Multi-target Adversarial Training with Adaptive Budget

Require: Training Epochs E, Training samples $(\mathcal{X}, \mathcal{Y})$, adversarial budgets $(\epsilon_\infty, \epsilon_1, \epsilon_2)$, model $f(\cdot)$, loss function ℓ.

for $e \in [E]$ **do**
    for $\mathbf{x}, y \in (\mathcal{X}, \mathcal{Y})$ **do**
        $g_p \leftarrow \ell^{'}(\mathbf{x} + \boldsymbol{\delta}_p(\mathbf{x}))$, $\boldsymbol{\delta}_p(\mathbf{x}) \leftarrow \mathrm{PGD}(\mathbf{x}, k, \eta, \ell, \epsilon_p, \ell), \forall p \in \{1, 2, \infty\}$
        Get adaptive budgets $\hat{\epsilon}_1, \hat{\epsilon}_2, \hat{\epsilon}_\infty \leftarrow$ *AdaptiveBudget*$([g_1, g_2, g_\infty], [\epsilon_1, \epsilon_2, \epsilon_\infty])$;
        Adversarial training using MAX, MSD or AVG with budgets $(\hat{\epsilon}_1, \hat{\epsilon}_2, \hat{\epsilon}_\infty)$;
    end for
end for
Return the classifier f.

AdaptiveBudget(Gradients$[g_1, g_2, g_\infty]$, Epsilon$[\epsilon_1, \epsilon_2, \epsilon_\infty]$):
    $p_{\max} \leftarrow \mathrm{argmax}_{p \in \{\infty, 1, 2\}} \|g_p\|$, $p_{\min} \leftarrow \mathrm{argmin}_{p \in \{\infty, 1, 2\}} \|g_p\|$;
    $p_{\mathrm{mid}} \leftarrow \{1, 2, \infty\} / \{p_{\max}, p_{\min}\}$;
    $\epsilon_{p_{\max}} \leftarrow \epsilon_{p_{\max}} \cdot \frac{\|g_{p_{\max}}\|}{\|g_{p_{\mathrm{mid}}}\|}, \quad \epsilon_{p_{\min}} \leftarrow \epsilon_{p_{\min}} \cdot \frac{\|g_{p_{\min}}\|}{\|g_{p_{\mathrm{mid}}}\|}$;
    Return $\epsilon_1, \epsilon_2, \epsilon_\infty$.

Experimental Results

Experimental Results on MNIST

Models w. adaptive budget	ℓ_1	ℓ_2	ℓ_∞	MAX	MAX ℓ_1 (ours)	MAX ℓ_2 (ours)	MSD	MSD ℓ_1 (ours)	MSD ℓ_2 (ours)	AVG	AVG ℓ_1 (ours)	AVG ℓ_2 (ours)
Clean Accuracy (%)	97.2	99.1	99.2	98.6	98.9	98.9	98.2	98.3	98.9	99.1	99.1	99.1
ℓ_1 PGD Robust Acc (%)	47.3*	67.8*	54.6*	67.1*	**71.4**↑	**69.7**↑	67.3*	66.8↓	65.9↓	70.6*	68.2↓	68.9↓
ℓ_2 PGD Robust Acc (%)	24.1*	66.8*	61.8*	67.2*	**69.4**↑	**69.5**↑	68.0*	67.9↓	65.3↓	69.4*	68.3↓	68.3↓
ℓ_∞ PGD Robust Acc (%)	0*	0.1*	88.9*	21.2*	**67.2**↑	**67.6**↑	62.4*	**69.7**↑	**69.7**↑	59.5*	**67.7**↑	**65.6**↑
All PGD Robust Acc (%)	0*	0.1*	52.1*	21.2*	**61.3**↑	**61.4**↑	59.7*	**62.1**↑	**61.0**↑	55.4*	**59.2**↑	**58.2**↑

Experimental Results on CIFAR-10

Models w. adaptive budget	ℓ_1	ℓ_2	ℓ_∞	MAX	ℓ_1 (ours)	ℓ_2 (ours)	MSD	ℓ_1 (ours)	ℓ_2 (ours)	AVG	ℓ_1 (ours)	ℓ_2 (ours)
Clean Accuracy	92.4	87.5	84.2	79.6	76.9	78.7	79.2	77.6	79.0	83.8	81.6	81.5
ℓ_1 PGD Robust Acc (%)	90.8	31.7	17.3	44.0*	**50.7**↑	**51.7**↑	50.8*	**51.2**↑	**52.6**↑	55.7*	**57.3**↑	**56.3**↑
ℓ_2 PGD Robust Acc (%)	0.1	64.0	60.6	55.6*	**63.4**↑	**65.1**↑	64.3*	63.6↓	**65.5**↑	67.0*	66.6↓	67.0
ℓ_∞ PGD Robust Acc (%)	0	27.8	51.2	41.3*	**47.5**↑	**47.6**↑	45.7*	**48.4**↑	**47.2**↑	39.4*	**45.5**↑	**44.2**↑
All PGD Robust Acc (%)	0	23.8	17.3	40.4*	**46.0**↑	**46.8**↑	44.1*	**47.2**↑	**46.4**↑	39.2*	**45.2**↑	**43.6**↑
ℓ_1 AA Robust Acc (%)	0	23.8	6.2	41.4*	**45.7**↑	**45.5**↑	45.5*	**46.4**↑	**46.7**↑	49.7*	**52.7**↑	**50.8**↑
ℓ_2 AA Robust Acc (%)	0	63.0	57.4	53.7*	**60.4**↑	**63.2**↑	61.9*	**62.3**↑	**62.1**↑	65.4*	64.6↓	**65.5**↑
ℓ_∞ AA Robust Acc (%)	0	26.1	48.0	38.4*	**44.7**↑	**44.1**↑	43.1*	**45.2**↑	**44.4**↑	37.0*	**43.1**↑	**42.1**↑
All AA Robust Acc (%)	0	19.5	6.2	37.6*	**42.9**↑	**42.3**↑	41.6*	**43.4**↑	**43.0**↑	36.6*	**42.5**↑	**41.2**↑

Experimental Results on CIFAR-100

Models w. **AdaptiveBudget**	MAX			MSD			AVG		
		ℓ_1 (ours)	ℓ_2 (ours)		ℓ_1 (ours)	ℓ_2 (ours)		ℓ_1 (ours)	ℓ_2 (ours)
Clean Accuracy	55.49*	56.48	55.53	56.09*	55.52	54.94	59.94*	57.78	58.16
ℓ_1 PGD Robust Acc (%)	25.45*	**29.27**↑	**29.78**↑	35.50*	30.31↓	28.87↓	30.35*	**33.16**↑	**32.62**↑
ℓ_2 PGD Robust Acc (%)	39.55*	**40.00**↑	**39.85**↑	40.14*	**40.28**↑	39.28↓	40.26*	**41.03**↑	**40.27**↑
ℓ_∞ PGD Robust Acc (%)	25.03*	**25.34**↑	**25.87**↑	24.83*	**26.19**↑	**25.59**↑	18.92*	**21.81**↑	**21.57**↑
All PGD Robust Acc (%)	21.11*	**24.14**↑	**24.76**↑	25.10*	25.03↓	24.43↓	18.61*	**21.55**↑	**21.16**↑
ℓ_1 AA Robust Acc (%)	13.00*	**23.00**↑	**20.90**↑	25.10*	24.00↓	24.20↓	25.20*	**28.60**↑	**28.00**↑
ℓ_2 AA Robust Acc (%)	36.30*	35.60↓	**36.40**↑	37.60*	35.80↓	36.40↓	37.00*	**37.90**↑	**37.10**↑
ℓ_∞ AA Robust Acc (%)	22.00*	21.50↓	**22.30**↑	21.80*	**22.80**↑	**22.70**↑	16.30*	**19.00**↑	**19.70**↑
All AA Robust Acc (%)	12.20*	**20.60**↑	**18.60**↑	21.00*	**21.30**↑	**21.50**↑	16.10*	**18.90**↑	**19.50**↑

Conclusion

Summary

- We show the first theoretical results on the convergence of MAX, MSD, and AVG on the multi-target robustness.

Summary

- We show the first theoretical results on the convergence of MAX, MSD, and AVG on the multi-target robustness.
- We design a novel algorithm namely AdaptiveBudget which is able to alleviate the player domination phenomenon and thus might avoid the non-convergence of MAX and MSD under SVM and Linear cases.

Summary

- We show the first theoretical results on the convergence of MAX, MSD, and AVG on the multi-target robustness.
- We design a novel algorithm namely AdaptiveBudget which is able to alleviate the player domination phenomenon and thus might avoid the non-convergence of MAX and MSD under SVM and Linear cases.
- Experimental results show that AdaptiveBudget improves the performance of MSD, MAX, and AVG.

Thanks for listening!

References i

F. Croce and M. Hein.
Provable robustness against all adversarial l_p-perturbations for $p \geq 1$.
In *International Conference on Learning Representations*, 2020.

I. J. Goodfellow, J. Shlens, and C. Szegedy.
Explaining and Harnessing Adversarial Examples.
In *International Conference on Learning Representations*, 2015.

D. Kang, Y. Sun, T. Brown, D. Hendrycks, and J. Steinhardt.
Transfer of Adversarial Robustness Between Perturbation Types.
In *arxiv:1905.01034*, 2019.

P. Maini, E. Wong, and J. Z. Kolter.
Adversarial Robustness Against the Union of Multiple Perturbation Models.
In *International Conference on Machine Learning*, pages 6640–6650, 2020.