SAT: Improving Adversarial Training via Curriculum-Based Loss Smoothing

Chawin Sitawarin[†] Supriyo Chakraborty[*] David Wagner[†]

[†]UC Berkeley
[*]IBM T. J. Watson Research Center

AISec 2021

Contact: chawins@berkeley.edu

(Part of the work done while I interned at IBM Research)

Adversarial Examples

- Attacks on machine learning models are becoming real concerns
- Adversarial examples: small perturbation on inputs to mislead a classifier into making a wrong prediction
- Generated by solving an optimization problem:

$$x^{adv} = x + \delta^* \quad \text{where} \quad \delta^* = \operatorname*{arg\,max}_{\delta : \|\delta\|_\infty \le \epsilon} \ \ell(x + \delta; \theta) \tag{1}$$

Defenses against Adversarial Examples

- **Adversarial Training** [Madry et al., 2018] is a popular and effective method for training robust networks against adversarial examples.

$$\arg\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \ell_{\epsilon}(x_i; \theta) \tag{2}$$

$$\text{where} \quad \ell_{\epsilon}(x; \theta) := \max_{\delta : \|\delta\|_{\infty} \leq \epsilon} \ell(x + \delta; \theta) \tag{3}$$

- We call $\ell(x;\theta)$ **normal loss** and $\ell_{\epsilon}(x;\theta)$ **adversarial loss**.

Problems with Adversarial Training

Our work attempts to address the following problems:

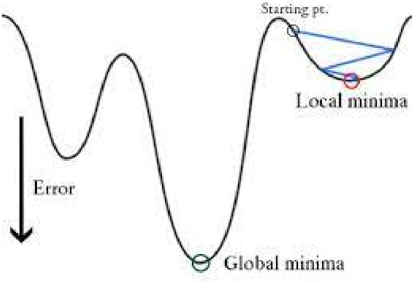
- Large drop on clean accuracy
- Stuck in “poor” local minima, learn a trivial classifier
- Large adversarial generalization gap

Outline

- Brief introduction to curriculum learning
- Adversarial training + curriculum learning
- H-SAT
- P-SAT
- Results

Curriculum Learning: An Introduction

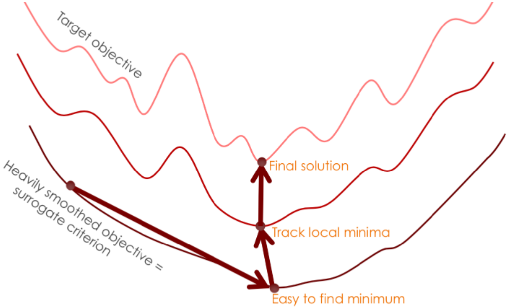
- Curriculum Learning [Bengio et al., 2009] is an idea borrowed from Numerical Continuation Methods [Allgower and Georg, 1990] to solve non-convex problems
- Bad: Solve the non-convex problem directly $\rightarrow$ get stuck in poor local optima



Ref: www.cs.ubc.ca/labs/lci/mlrg/slides/non_convex_optimization.pdf

Curriculum Learning: An Introduction

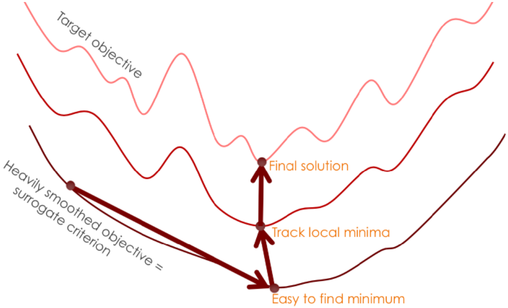
- Good: "Start easy"



Ref: Wang et al. [2020]

Curriculum Learning: An Introduction

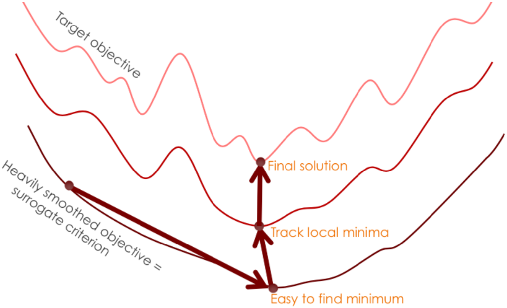
- Good: "Start easy"
  1. Solve a smooth version of the problem



Ref: Wang et al. [2020]

Curriculum Learning: An Introduction

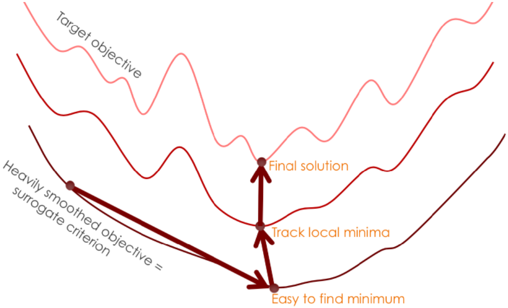
- Good: "Start easy"
  1. Solve a smooth version of the problem
  2. Move closer to global optima



Ref: Wang et al. [2020]

Curriculum Learning: An Introduction

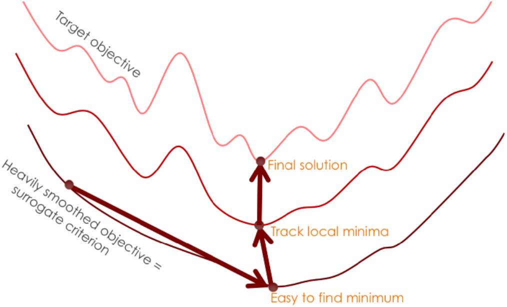
- Good: "Start easy"
  1. Solve a smooth version of the problem
  2. Move closer to global optima
  3. Update the smooth version to be more similar to the real one



Ref: Wang et al. [2020]

Curriculum Learning: An Introduction

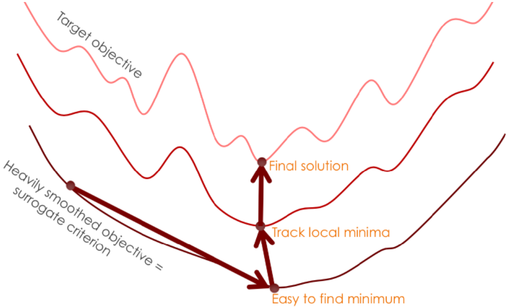
- Good: "Start easy"
  1. Solve a smooth version of the problem
  2. Move closer to global optima
  3. Update the smooth version to be more similar to the real one
  4. Repeat 1-3



Ref: Wang et al. [2020]

Curriculum Learning: An Introduction

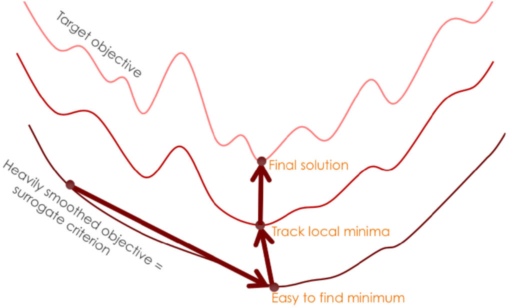
- Good: "Start easy"
  1. Solve a smooth version of the problem
  2. Move closer to global optima
  3. Update the smooth version to be more similar to the real one
  4. Repeat 1-3
  5. Stop when the smooth version is equivalent to the real one



Ref: Wang et al. [2020]

Curriculum Learning: An Introduction

- Good: "Start easy"
  1. Solve a smooth version of the problem
  2. Move closer to global optima
  3. Update the smooth version to be more similar to the real one
  4. Repeat 1-3
  5. Stop when the smooth version is equivalent to the real one



Ref: Wang et al. [2020]

- Increase likelihood of reaching global optima

Curriculum Learning: An Introduction

- For deep learning, this means "train on easy samples first then slowly include harder ones"

Curriculum Learning: An Introduction

- For deep learning, this means "train on easy samples first then slowly include harder ones"
- In adversarial settings, easy = clean samples or weak adversary, and hard = strong adversary

Curriculum Learning: An Introduction

- For deep learning, this means "train on easy samples first then slowly include harder ones"
- In adversarial settings, easy = clean samples or weak adversary, and hard = strong adversary
- Determining notion of difficulty is a crucial part

Curriculum Learning: An Introduction

- For deep learning, this means "train on easy samples first then slowly include harder ones"
- In adversarial settings, easy = clean samples or weak adversary, and hard = strong adversary
- Determining notion of difficulty is a crucial part
- Previous works have considered multiple notion of difficulty (e.g., ϵ, adversarial loss, number of PGD steps)

Curriculum Learning: An Introduction

- For deep learning, this means "train on easy samples first then slowly include harder ones"
- In adversarial settings, easy = clean samples or weak adversary, and hard = strong adversary
- Determining notion of difficulty is a crucial part
- Previous works have considered multiple notion of difficulty (e.g., ϵ, adversarial loss, number of PGD steps)
- We propose two new difficulty metrics based on the Hessian matrix and the softmax probability

Our Contributions

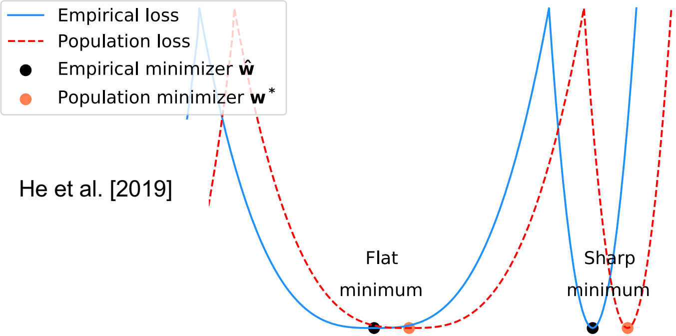
- ~~Large drop on clean accuracy~~ $\rightarrow$ Train on easy (i.e. clean) samples
- Stuck in "poor" local minima, learn a trivial classifier
- Large adversarial generalization gap

Our Contributions

- ~~Large drop on clean accuracy~~ → Train on easy (i.e. clean) samples
- ~~Stuck in "poor" local minima, learn a trivial classifier~~ → Smoothness
- Large generalization gap

Large generalization gap

- Flat (or smooth) local minima are believed to generalize better than sharp (or non-smooth) minima.
- Curriculum learning can lead smoother loss landscapes



He et al. [2019]

Our Contributions

- ~~Large drop on clean accuracy~~ $\rightarrow$ Train on easy (i.e. clean) samples
- ~~Stuck in "poor" local minima, learn a trivial classifier~~ $\rightarrow$ Smoothness
- ~~Large generalization gap~~ $\rightarrow$ Smoothness

Curriculum Learning + Adversarial Training

- Previous works proposed different forms of curriculum learning.

Curriculum Learning + Adversarial Training

- Previous works proposed different forms of curriculum learning.
- We unify all of them under one general formulation: **curriculum constraint** and **curriculum loss** $\ell_{\psi,\epsilon}$:

$$\ell_{\psi,\epsilon}(x, \lambda) = \max_{\delta:\|\delta\|_\infty \leq \epsilon} \ell(x+\delta) \quad \text{s.t.} \quad \psi(x+\delta) \leq \lambda \tag{4}$$

where $\psi : \mathbb{R}^d \to \mathbb{R}$ is a given difficulty metric.

Curriculum Learning + Adversarial Training

- Previous works proposed different forms of curriculum learning.
- We unify all of them under one general formulation: **curriculum constraint** and **curriculum loss** $\ell_{\psi,\epsilon}$:

$$\begin{aligned} \ell_{\psi,\epsilon}(x,\lambda) \quad &= \quad \max_{\delta:\|\delta\|_\infty \le \epsilon} \quad \ell(x+\delta) \\ & \qquad\qquad \text{s.t.} \quad \psi(x+\delta) \le \lambda \end{aligned} \tag{4}$$

  where $\psi : \mathbb{R}^d \to \mathbb{R}$ is a given difficulty metric.
- In general, we try to have $\psi(x) \in [0,1]$. when $\lambda = 1$, it reduces to original adversarial loss.

Curriculum Learning + Adversarial Training

- Previous works proposed different forms of curriculum learning.
- We unify all of them under one general formulation: **curriculum constraint** and **curriculum loss** $\ell_{\psi,\epsilon}$:

$$\ell_{\psi,\epsilon}(x, \lambda) = \max_{\delta:\|\delta\|_\infty \le \epsilon} \ell(x+\delta) \quad \text{s.t.} \quad \psi(x+\delta) \le \lambda \tag{4}$$

  where $\psi : \mathbb{R}^d \to \mathbb{R}$ is a given difficulty metric.
- In general, we try to have $\psi(x) \in [0, 1]$. when $\lambda = 1$, it reduces to original adversarial loss.
- We can start training with $\lambda = 0$ or some small λ (easy) and gradually increase it to 1 (hard).

Our Approaches

- Question: How to design the difficulty metric?

Our Approaches

- Question: How to design the difficulty metric?
- H-SAT: Hessian-based Smooth Adversarial Training

Our Approaches

- Question: How to design the difficulty metric?
- H-SAT: Hessian-based Smooth Adversarial Training
- P-SAT: Probability-based Smooth Adversarial Training

H-SAT: Hessian-based Smooth Adversarial Training

- Can we directly encourage smoothness through the difficulty metric?

H-SAT: Hessian-based Smooth Adversarial Training

- Can we directly encourage smoothness through the difficulty metric?
- Liu et al. [2020] shows that larger ϵ (more difficult) leads to less smoothness through Lipschitz-type bound

H-SAT: Hessian-based Smooth Adversarial Training

- Can we directly encourage smoothness through the difficulty metric?
- Liu et al. [2020] shows that larger ϵ (more difficult) leads to less smoothness through Lipschitz-type bound
- Make sure that adversarial training generates adversarial examples that result in smooth loss landscapes **locally** w.r.t. θ

H-SAT: Hessian-based Smooth Adversarial Training

- Can we directly encourage smoothness through the difficulty metric?
- Liu et al. [2020] shows that larger ϵ (more difficult) leads to less smoothness through Lipschitz-type bound
- Make sure that adversarial training generates adversarial examples that result in smooth loss landscapes **locally** w.r.t. θ
- We follow Liu et al.'s notion of local smoothness:

H-SAT: Hessian-based Smooth Adversarial Training

- Can we directly encourage smoothness through the difficulty metric?
- Liu et al. [2020] shows that larger ϵ (more difficult) leads to less smoothness through Lipschitz-type bound
- Make sure that adversarial training generates adversarial examples that result in smooth loss landscapes **locally** w.r.t. θ
- We follow Liu et al.'s notion of local smoothness:

Definition 1: Local Smoothness of Adversarial Loss

The largest eigenvalue of the Hessian evaluated at the adversarial example ("maximal Hessian eigenvalue" in short): $\|H_\epsilon(x;\theta)\|_{(2)}$.

$$H_\epsilon(x;\theta) \coloneqq \nabla^2_\theta \ell(x^{adv};\theta) \quad \text{for } x^{adv} \in \underset{z:\|z-x\|_p \le \epsilon}{\arg\max} \; \ell(z;\theta) \tag{5}$$

H-SAT: Hessian-based Smooth Adversarial Training

Definition 1: Local Smoothness of Adversarial Loss

The largest eigenvalue of the Hessian evaluated at the adversarial example ("maximal Hessian eigenvalue" in short): $\|H_\epsilon(x;\theta)\|_{(2)}$.

$$H_\epsilon(x;\theta) := \nabla^2_\theta \ell(x^{adv};\theta) \quad \text{for } x^{adv} \in \underset{z:\|z-x\|_p \leq \epsilon}{\arg\max} \; \ell(z;\theta) \tag{6}$$

- H-SAT's difficulty metric:

$$\psi_H(x) \approx \|H_\epsilon(x;\theta)\|_{(2)} \tag{7}$$

H-SAT: Hessian-based Smooth Adversarial Training

Definition 1: Local Smoothness of Adversarial Loss

The largest eigenvalue of the Hessian evaluated at the adversarial example ("maximal Hessian eigenvalue" in short): $\|H_\epsilon(x;\theta)\|_{(2)}$.

$$H_\epsilon(x;\theta) := \nabla^2_\theta \ell(x^{adv};\theta) \text{ for } x^{adv} \in \underset{z:\|z-x\|_p\leq\epsilon}{\arg\max}\ \ell(z;\theta) \tag{6}$$

- H-SAT's difficulty metric:

$$\psi_H(x) \approx \|H_\epsilon(x;\theta)\|_{(2)} \tag{7}$$

- However, it is computationally expensive which leads to multiple approximation and performs slightly worse than P-SAT

P-SAT: Probability-Based Smooth Adversarial Training

- We propose **Probability-Based Smooth Adversarial Training** (P-SAT) with "softmax probability gap" as the difficulty metric:

$$\psi_P(x) \coloneqq \max_{j \neq y} f(x)_j - f(x)_y \tag{8}$$

  where $y \in \{1, ..., c\}$ is the ground-truth label of x, and $f : \mathbb{R}^d \to \mathbb{R}^c$ is the softmax output of a neural network.
- Has stronger connection to notion of difficulty than H-SAT: large $\psi_P(x) =$ wrong prediction with high confidence
- Connection to smoothness in logistic regression
- No computational overhead
- We use early stopping to satisfy this constraint when generating adversarial examples.

Empirical Results

Table: Clean and adversarial accuracy (AutoAttack) of the defenses on **MNIST**. The numbers in red indicate that the network is stuck in a sub-optimal local minimum.

Defenses	$\epsilon = 0.3$		$\epsilon = 0.45$	
	Clean	Adv	Clean	Adv
Madry et al. [2018]	98.07	85.47	11.22	11.22
Zhang et al. [2019]	98.98	90.70	97.36	0.00
Wang et al. [2019]	98.93	**92.24**	97.98	**65.71**
Cheng et al. [2020]	99.46	0.00	99.39	0.00
H-SAT (ours)	99.01	80.71	98.35	54.10
P-SAT (ours)	99.16	92.00	97.87	58.50

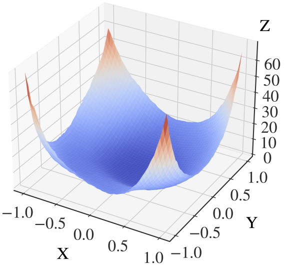
- Only Wang et al. [2019] and ours do not learn trivial classifiers.

Table: Clean and adversarial accuracy on **Imagenette** dataset.

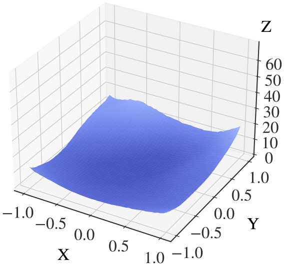
Defenses	$\epsilon = 16/255$		$\epsilon = 24/255$	
	Clean	Adv	Clean	Adv
Madry et al. [2018]	49.10	28.00	42.55	21.05
Zhang et al. [2019]	**78.05**	8.90	**68.50**	1.90
Wang et al. [2019]	66.20	30.30	52.50	24.50
H-SAT (ours)	69.10	**35.45**	47.50	**27.75**
P-SAT (ours)	72.20	31.25	62.15	20.00

- Stabilize adversarial training, especially on non-ResNet models
- Minor but consistent improvement over previous works on CIFAR-10 and CIFAR-100
- 2-5 percentage points improvement on clean accuracy over Madry et al. [2018], or 1-2 for adversarial accuracy
- Larger improvement on Imagenette and larger ϵ

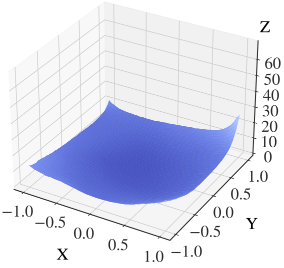
Loss Landscapes



Madry et al. [2018]



H-SAT (ours)



P-SAT (ours)

Summary

In summary, we...

- Propose a general formulation of curriculum-based adversarial training.
- Propose H-SAT and P-SAT which aim at improving smoothness of adversarial training and solving its drawbacks.
- Empirically confirm our intuitions and trains neural networks with higher robustness and clean accuracy compared to the baselines on various datasets.

Thank You!

References I

E. L. Allgower and K. Georg. Numerical Continuation Methods: An Introduction. Springer-Verlag, Berlin, Heidelberg, 1990. ISBN 0-387-12760-7.

Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09, pages 1–8, Montreal, Quebec, Canada, 2009. ACM Press. ISBN 978-1-60558-516-1. doi: 10.1145/1553374.1553380.

M. Cheng, Q. Lei, P.-Y. Chen, I. Dhillon, and C.-J. Hsieh. CAT: Customized adversarial training for improved robustness. arXiv:2002.06789 [cs, stat], Feb. 2020.

C. Liu, M. Salzmann, T. Lin, R. Tomioka, and S. Süsstrunk. On the loss landscape of adversarial training: Identifying challenges and how to overcome them. In Advances in Neural Information Processing Systems, 2020.

References II

A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In International Conference on Learning Representations, 2018.

X. Wang, Y. Chen, and W. Zhu. A comprehensive survey on curriculum learning. ArXiv, abs/2010.13166, 2020.

Y. Wang, X. Ma, J. Bailey, J. Yi, B. Zhou, and Q. Gu. On the convergence and robustness of adversarial training. In K. Chaudhuri and R. Salakhutdinov, editors, Proceedings of the 36th International Conference on Machine Learning, volume 97 of Proceedings of Machine Learning Research, pages 6586–6595, Long Beach, California, USA, June 2019. PMLR.

H. Zhang, Y. Yu, J. Jiao, E. P. Xing, L. E. Ghaoui, and M. I. Jordan. Theoretically principled trade-off between robustness and accuracy. In International Conference on Machine Learning, 2019.