# Provably Robust Adversarial Examples

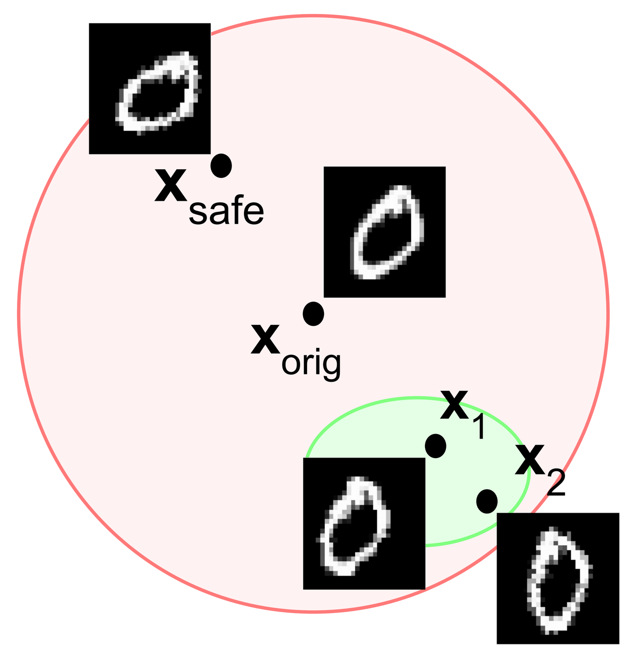
Dimitar I. Dimitrov, Gagandeep Singh, Timon Gehr, Martin Vechev

vmware Research | Department of Computer Science, University of Illinois Urbana-Champaign | Department of Computer Science, ETH zürich

## Single adversarial attacks vs Robust adversarial regions

### Problem setting

- Traditionally, robustness of $\mathbf{x}_{orig}$ is assessed by generating **individual** attacks $\mathbf{x}_1$ and $\mathbf{x}_2$ within a ball around it (in red).
- Description of the **whole adversarial region** (in green) is preferable. The region can contain **trillions** of adversarial images.




### Single adversarial attacks vs Adversarial regions

**Single attacks:**

- Easy to generate
- Less informative

**Adversarial regions:**

- More Informative
- Efficiently summarizes many individual attack
- Computationally expensive

**Key idea:** Use single attacks to generate initial region and **refine** it until provably verifiable.

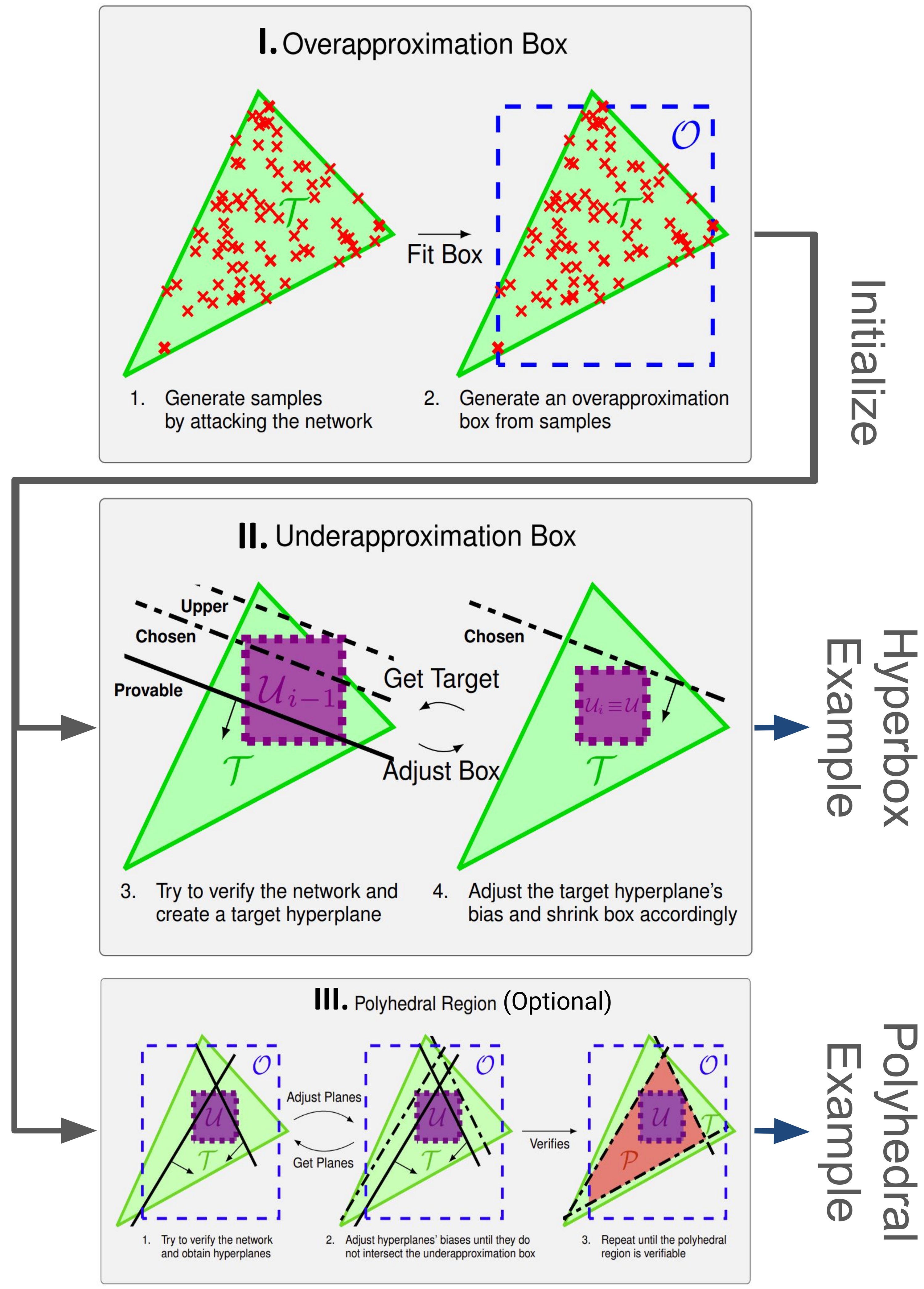
### Algorithm overview

**I.** Use **PGD** to generate many individual attacks. Fit a **hyperbox around them** to restrain search region. The region is shown in blue.

**II.** Use the overapproximation box to initialize. At each step use **black box verification tool** based on convex-relaxations to generate a **half-space constraint** which if added to the current box makes the resulting region **verifiably adversarial.** Adjust the constraints' **bias** such that a part of the box is removed but the constraint is **weaker**. Create **maximal box not intersecting** the adjusted constraint. Repeat until **verification succeeds**. Results in **hyperbox robust example**. The region is shown in purple.

**III.** Initialize with the overapproximation box. At each step use **black box verification tool** to generate half-space constraints that force the ReLU neurons to become **decided** and for the verification objective to become **positive**. **Bias-adjust** them so they **do not intersect** the underapproximation box region. This **enforces** the polyhedral region to be **larger** than the hyperbox example. Repeat until **verification**. Results in **polyhedral robust example**. The region is shown in red.

## PARADE: ProvAbly Robust ADversarial Examples




### Robust adversarial examples to $\ell_\infty$ -attacks

| Network | $\epsilon$ | #Cor | #Img | #Reg | Baseline #Ver | Baseline Time | Baseline Size | PARADE Box #Ver | PARADE Box Time | PARADE Box Size | PARADE Poly #Ver | PARADE Poly Time | PARADE Poly SizeO |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| MNIST 8x200 | 0.045 | 97 | 22 | 53 | 41 | 272 s | $10^{24}$ | 53 | 114 s | $10^{121}$ | 53 | 1556 s | $< 10^{191}$ |
| MNIST ConvSml | 0.12 | 100 | 21 | 32 | 31 | 171 s | $10^{339}$ | 32 | 74 s | $10^{494}$ | 32 | 141 s | $< 10^{561}$ |
| MNIST ConvBig | 0.05 | 98 | 18 | 29 | 15 | 1933s | $10^{9}$ | 28 | 880 s | $10^{137}$ | 28 | 5636 s | $< 10^{173}$ |
| CIFAR10 ConvSml | 0.006 | 59 | 23 | 44 | 28 | 238 s | $10^{360}$ | 44 | 113 s | $10^{486}$ | 44 | 264 s | $< 10^{543}$ |
| CIFAR10 ConvBig | 0.008 | 60 | 25 | 36 | 26 | 479 s | $10^{380}$ | 36 | 404 s | $10^{573}$ | 36 | 610 s | $< 10^{654}$ |

- **PARADE** regions contain up to **$10^{573}$** individual adversarial images.
- **PARADE** produces adversarial regions for all but one adversarial image.
- Regions generated by **PARADE** are much larger than uniform shrinking baseline.
- **PARADE** hyperbox example generation is **2x** faster than the uniform shrinking baseline.

## Experimental evaluation

### Robust adversarial examples to geometric perturbations

| Network | Transform | #Cor | #Img | #Reg | Baseline #Ver | Baseline Time | Baseline #Splits | PARADE #Ver | PARADE Time | PARADE #Splits | PARADE Under | PARADE Over |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| MNIST ConvSml | R(17) Sc(18) Sh(0.03) | 99 | 38 | 54 | 10 | 890 s | 2x5x2 | 51 | 774 s | 1x2x1 | $> 10^{96}$ | $< 10^{195}$ |
| | Sc(20) T(-1.7,1.7,-1.7,1.7) | 99 | 32 | 56 | 5 | 682 s | 4x3x3 | 51 | 521 s | 2x1x1 | $> 10^{71}$ | $< 10^{160}$ |
| | Sc(20) R(13) B(10, 0.05) | 99 | 33 | 48 | 2 | 420 s | 3x2x2x2 | 40 | 370 s | 2x1x1x1 | $> 10^{70}$ | $< 10^{455}$ |
| MNIST ConvBig | R(10) Sc(15) Sh(0.03) | 95 | 40 | 50 | 9 | 812 s | 2x4x2 | 44 | 835 s | 1x2x1 | $> 10^{77}$ | $< 10^{205}$ |
| | Sc(20) T(0,1,0,1) | 95 | 34 | 46 | 2 | 435 s | 4x2x2 | 42 | 441 s | 2x1x1 | $> 10^{64}$ | $< 10^{174}$ |
| | Sc(15) R(9) B(5, 0.05) | 95 | 39 | 52 | 2 | 801 s | 3x2x2x2 | 46 | 537 s | 2x1x1x1 | $> 10^{119}$ | $< 10^{545}$ |
| CIFAR ConvSml | R(2.5) Sc(10) Sh(0.02) | 53 | 24 | 29 | 1 | 1829 s | 5x2x2 | 29 | 1369 s | 2x1x1 | $> 10^{599}$ | $< 10^{1173}$ |
| | Sc(10) T(0,1,0,1) | 53 | 28 | 32 | 1 | 1489 s | 4x3x3 | 32 | 954 s | 2x1x1 | $> 10^{66}$ | $< 10^{174}$ |
| | Sc(5) R(8) B(1, 0.01) | 53 | 21 | 25 | 1 | 2189 s | 5x2x2x2 | 21 | 1481 s | 2x1x1x1 | $> 10^{513}$ | $< 10^{2187}$ |

- **PARADE** can handle diverse combinations of geometric perturbations, as it relies on DeepG in a **black-box** way.
- In similar time, **PARADE** generates **more verifiable regions** containing **more images** compared to baseline based on splitting.

### Robust adversarial examples and Randomized Smoothing

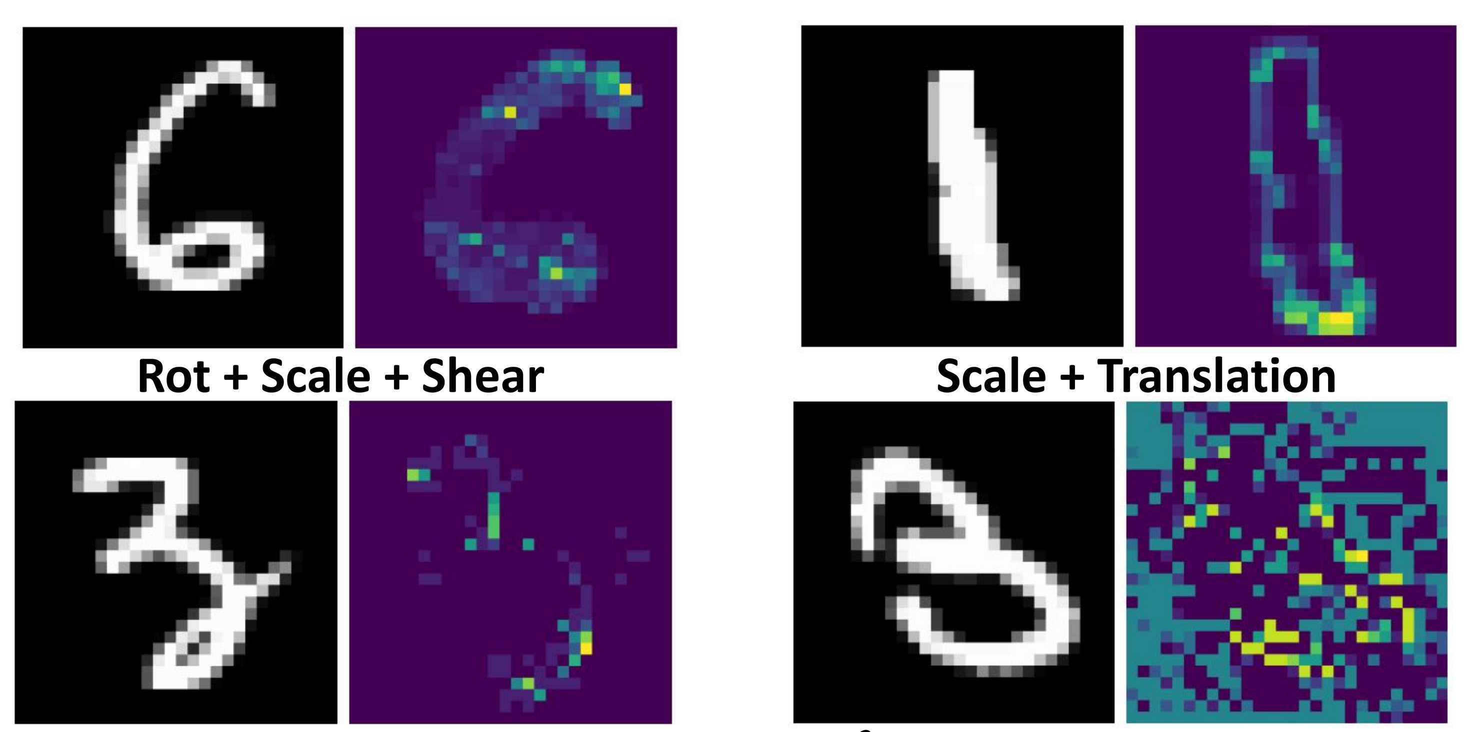
| Method | MNIST 8x200 | MNIST ConvSml | MNIST ConvBig | CIFAR ConvSml | CIFAR ConvBig |
|---|---|---|---|---|---|
| Baseline | 0.55 | 0.38 | 0.59 | 0.53 | 0.26 |
| PARADE | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** |
| Ind att mean | 0.29 | 0.16 | 0.18 | 0.48 | 0.25 |
| Ind att 95% perc | 0.53 | 0.44 | 0.51 | 0.61 | 0.37 |

- **PARADE** produces regions that are more robust (have **bigger robust radius** verified using **smoothing**) compared to uniform shrinking and individual attacks used during **Step I** of the algorithm.

### Empirically vs Provably robust adversarial examples

- Empirical examples can exhibit high Expectation-Over-Transformation (EoT), while their **subregions** close to the original attacked point **incur very low EoT** scores.
- Empirically robust adversarial example techniques recovered less regions: **44 vs 24**.

### Visualisation of Robust Adversarial Examples



Rot + Scale + Shear | Scale + Translation | Scale + Rot + Brightness | $\ell_\infty$-adversarial examples