Certified Training: Small Boxes are All You Need

Mark Müller*

Franziska Eckert*

Marc Fischer

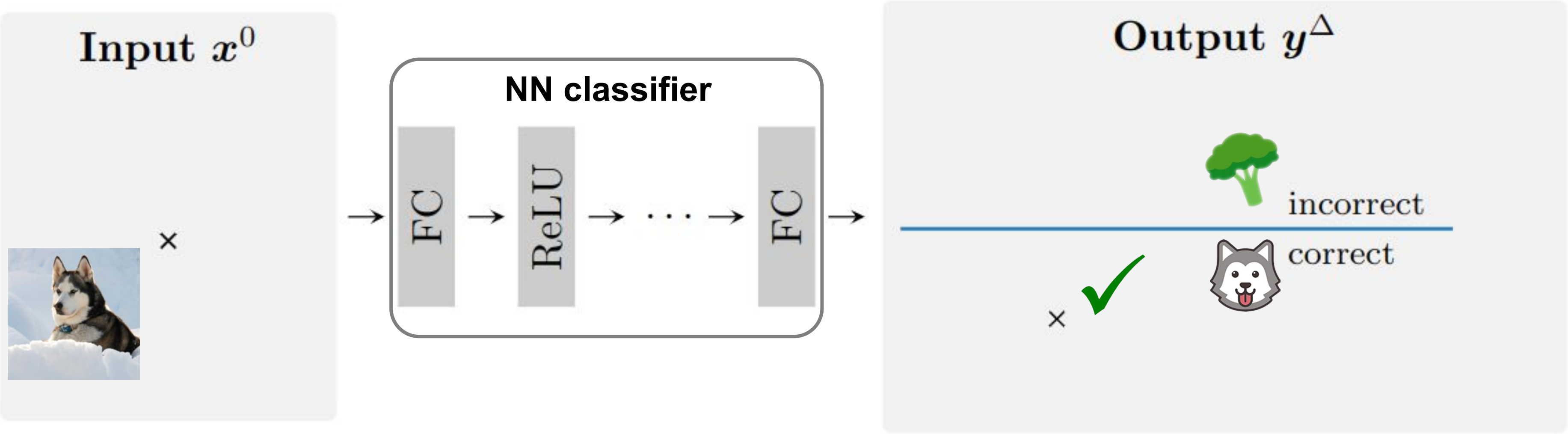
Martin Vechev

ETH zürich

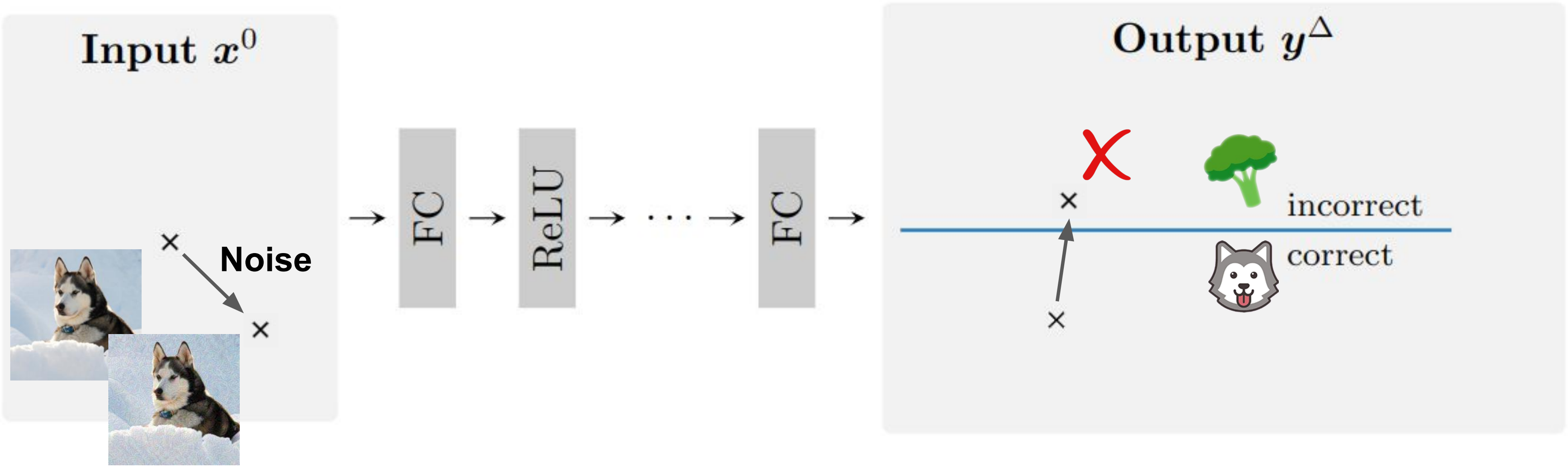
ICLR

SRILAB

Standard Classification

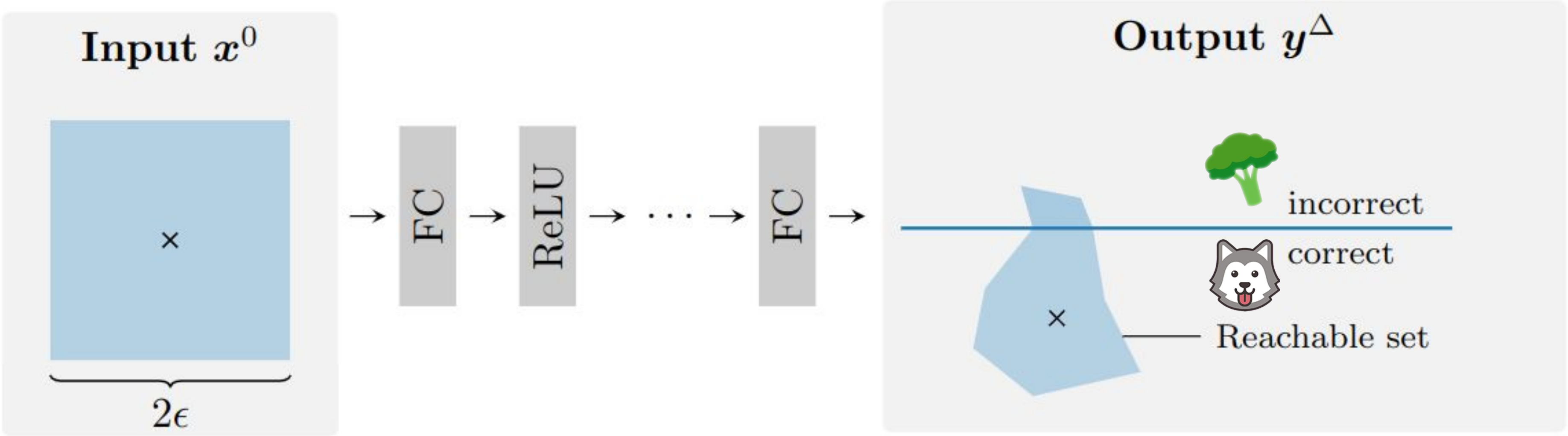


Adversarial Examples

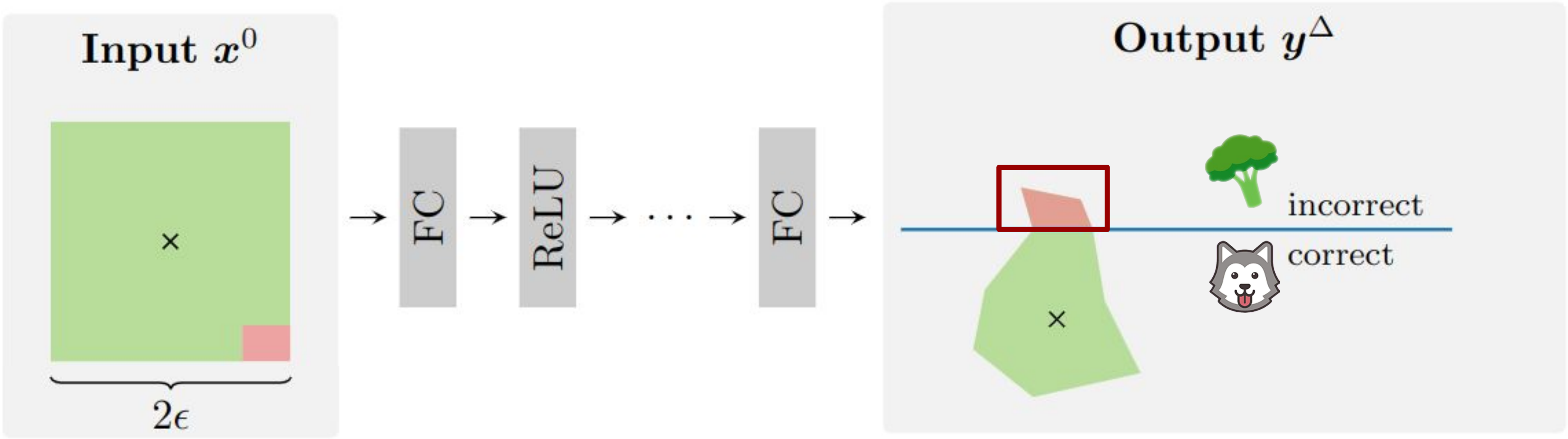


Szegedy et al. “Intriguing properties of neural networks” ICLR 2014
Biggio et al. “Evasion attacks against machine learning at test time” ECML PKDD 2013

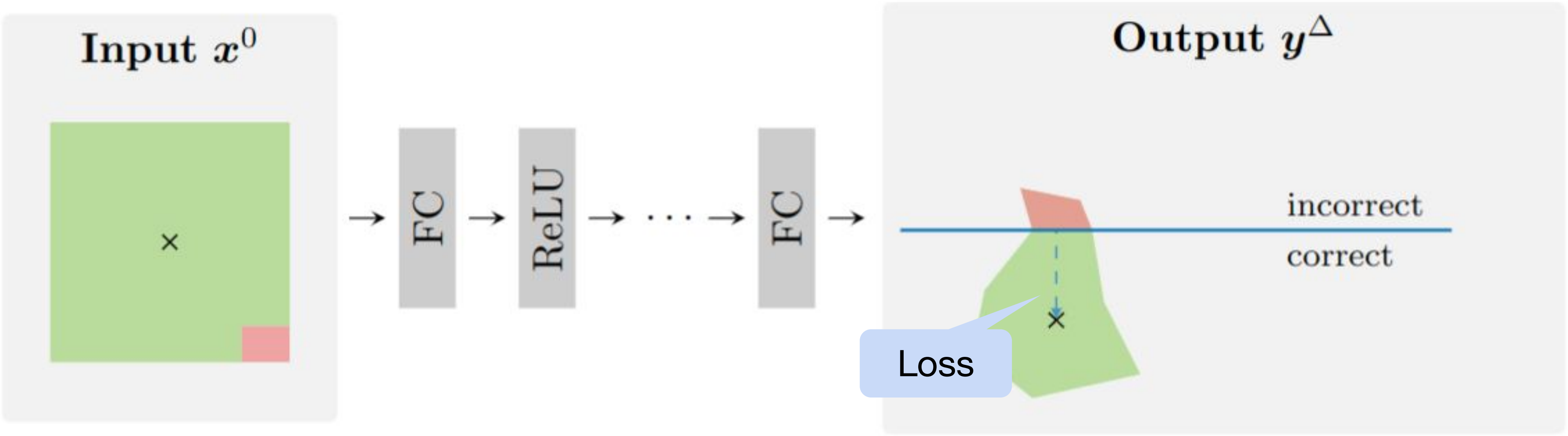
Exact Propagation



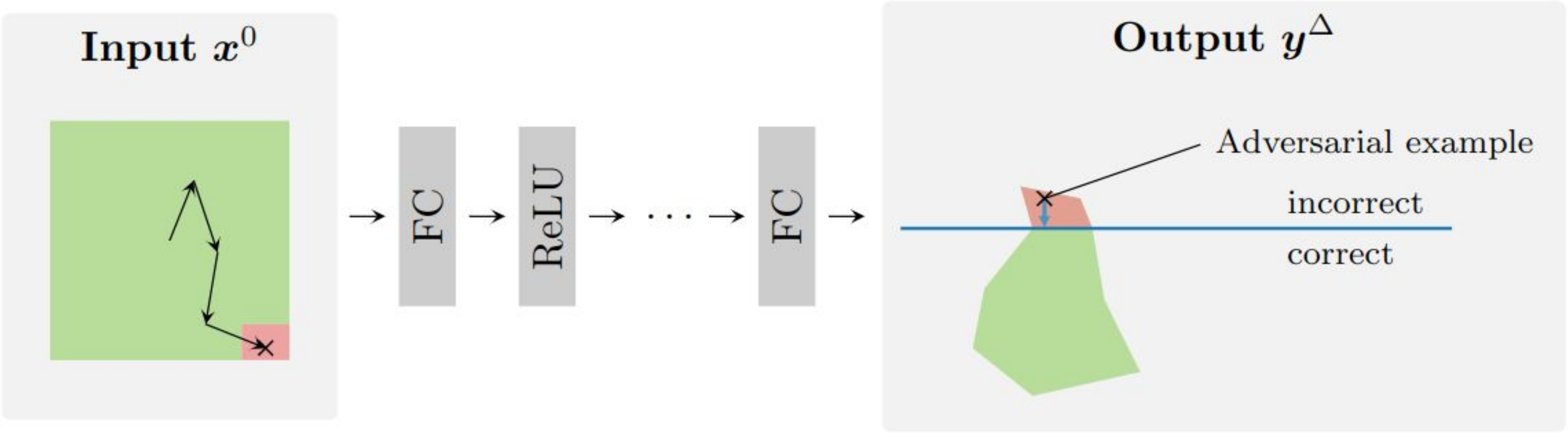
Exact Propagation



Standard Training

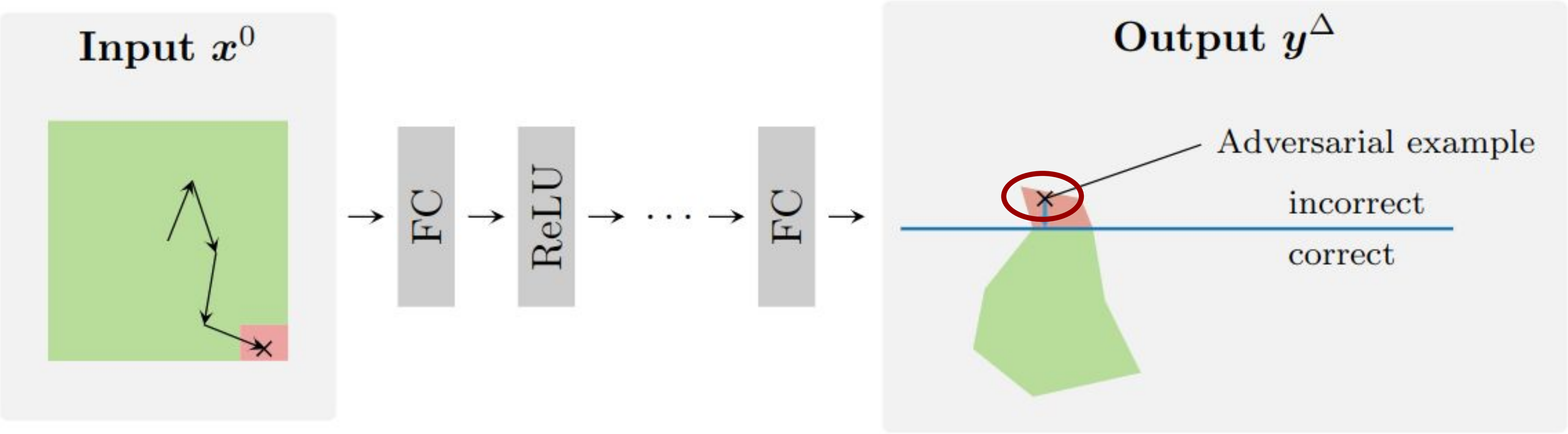


Adversarial Training (PGD)



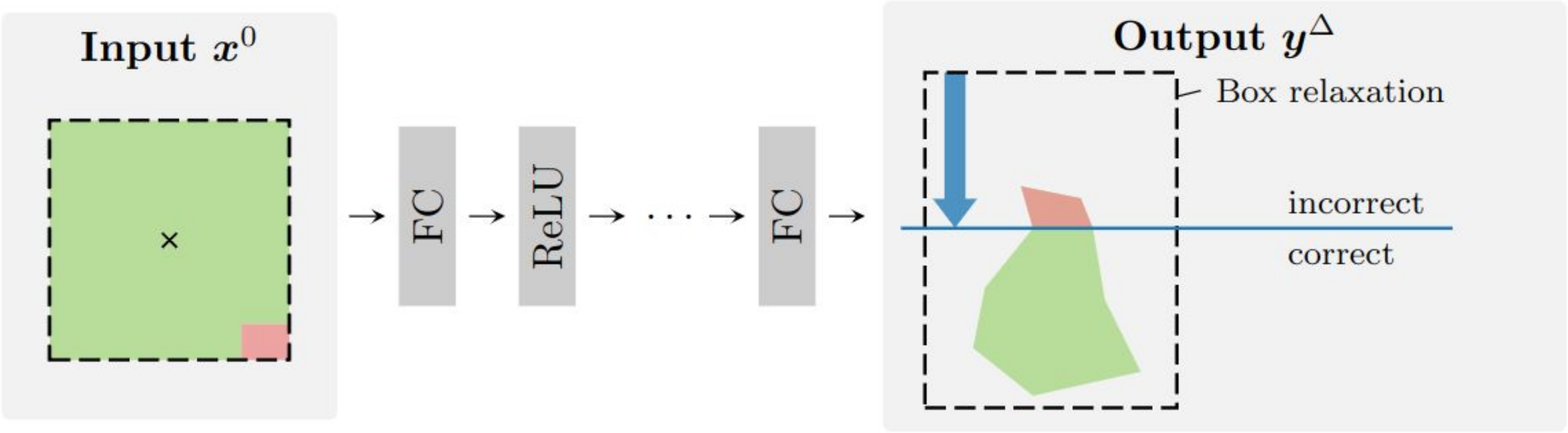
Madry et al. "Towards Deep Learning Models Resistant to Adversarial Attacks." ICLR 2018

Adversarial Training (PGD)



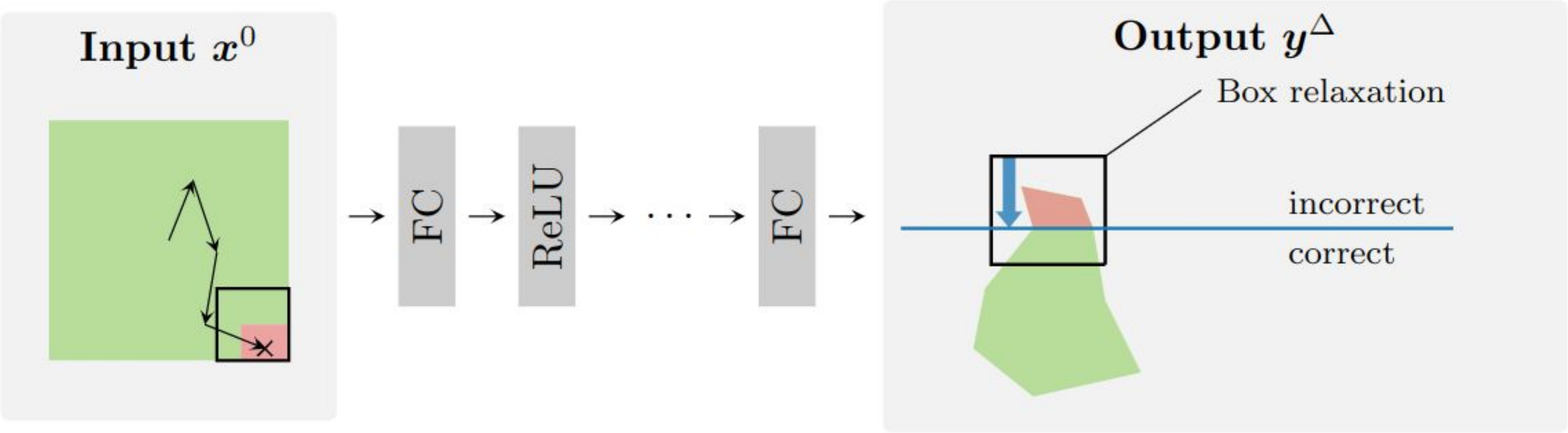
Madry et al. "Towards Deep Learning Models Resistant to Adversarial Attacks." ICLR 2018

Certified Training (IBP)

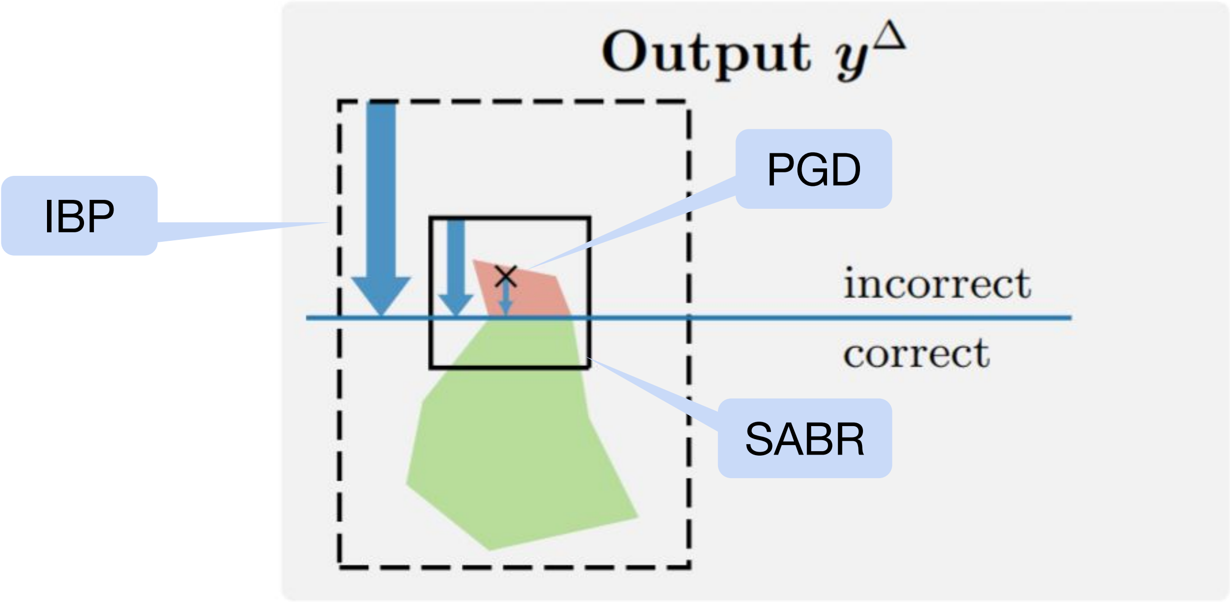


Gowal, et al. "On the effectiveness of interval bound propagation for training verifiably robust models." arXiv 2018
Mirmann et al. "Differentiable abstract interpretation for provably robust neural networks." ICML 2018

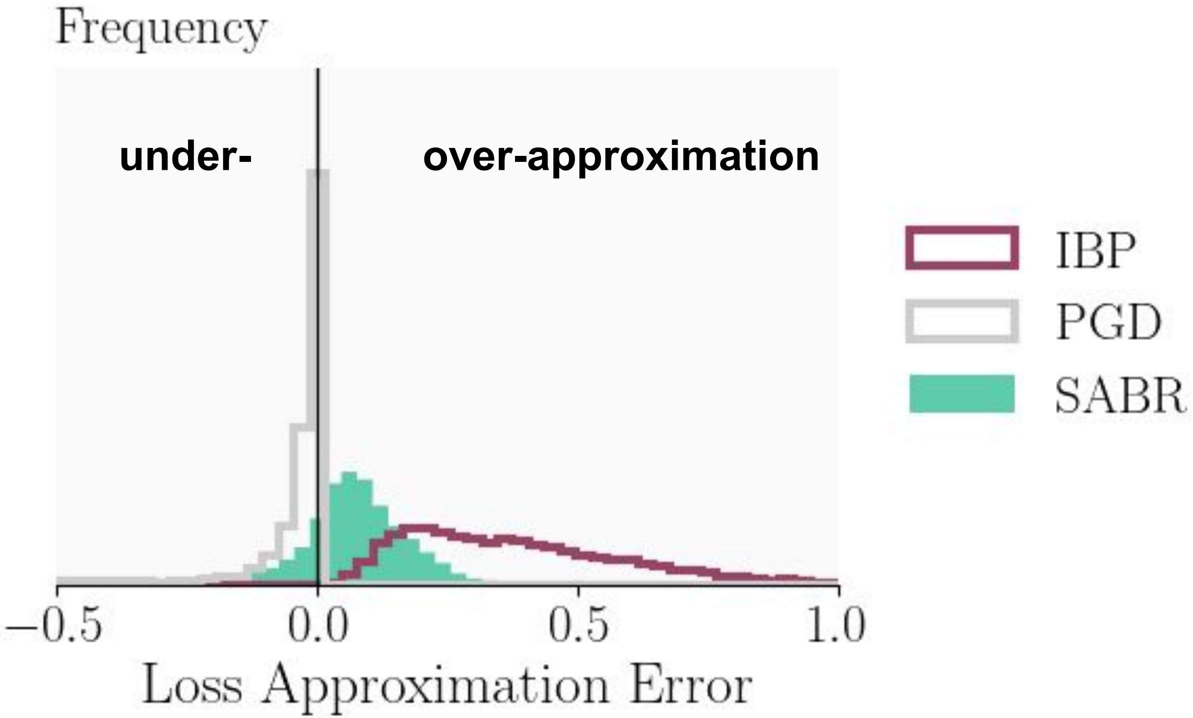
SABR – This Work



Regularisation Comparison

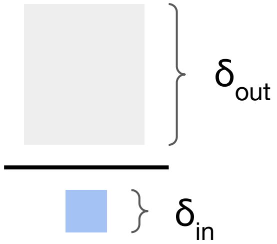


Worst-Case Loss Approximation Precision



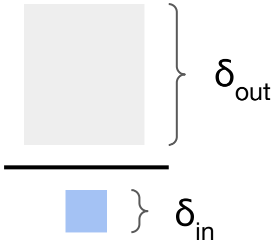
Box Abstraction Size Growth

Growth rate $\kappa = \frac{\mathbb{E}\left[\text{Output Box Size}\right]}{\text{Input Box Size}}$ =

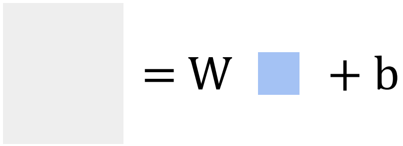


Box Abstraction Size Growth

Growth rate $\kappa = \dfrac{\mathbb{E}\left[\text{Output Box Size}\right]}{\text{Input Box Size}}$ =



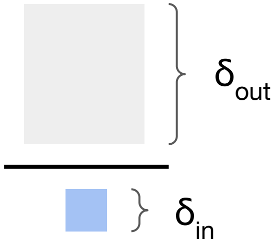
Linear layers: κ is independent of input box scale:



$\kappa \sim [10, 100]$

Box Abstraction Size Growth

Growth rate $\kappa = \frac{\mathbb{E}\left[\text{Output Box Size}\right]}{\text{Input Box Size}}$ =



Linear layers: κ is independent of input box scale: $= W \quad + b$

$\kappa \sim [10, 100]$

ReLU layers: κ depends on box scale and box centre: $= \max(\quad , 0)$

$\kappa \sim [0, 1]$

Box Abstraction Size Growth

Growth rate $\kappa = \dfrac{\mathbb{E}\,[\text{Output Box Size}]}{\text{Input Box Size}}$ $=$ $\dfrac{\delta_{out}}{\delta_{in}}$

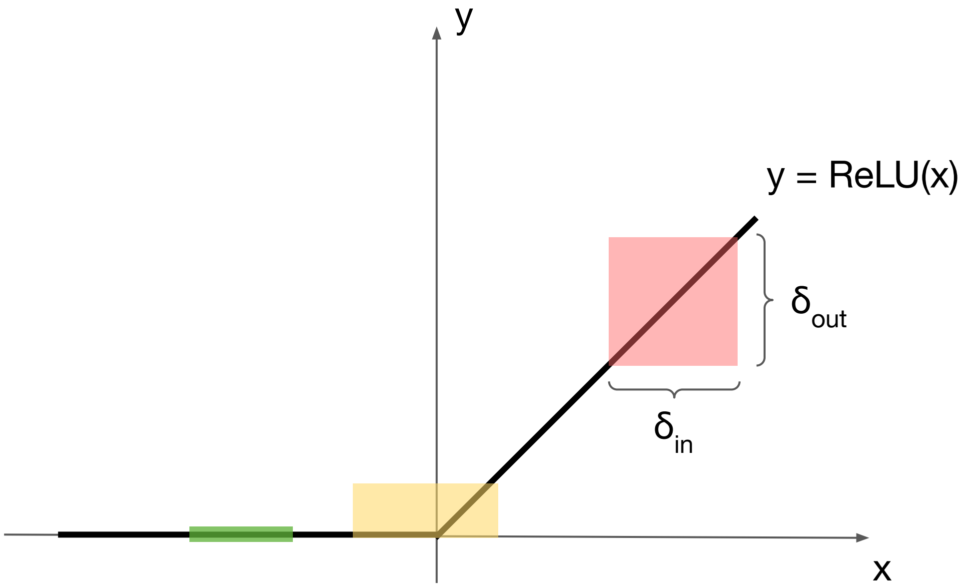
Linear layers: κ is independent of input box scale: $= W \quad + b$

$\kappa \sim [10, 100]$

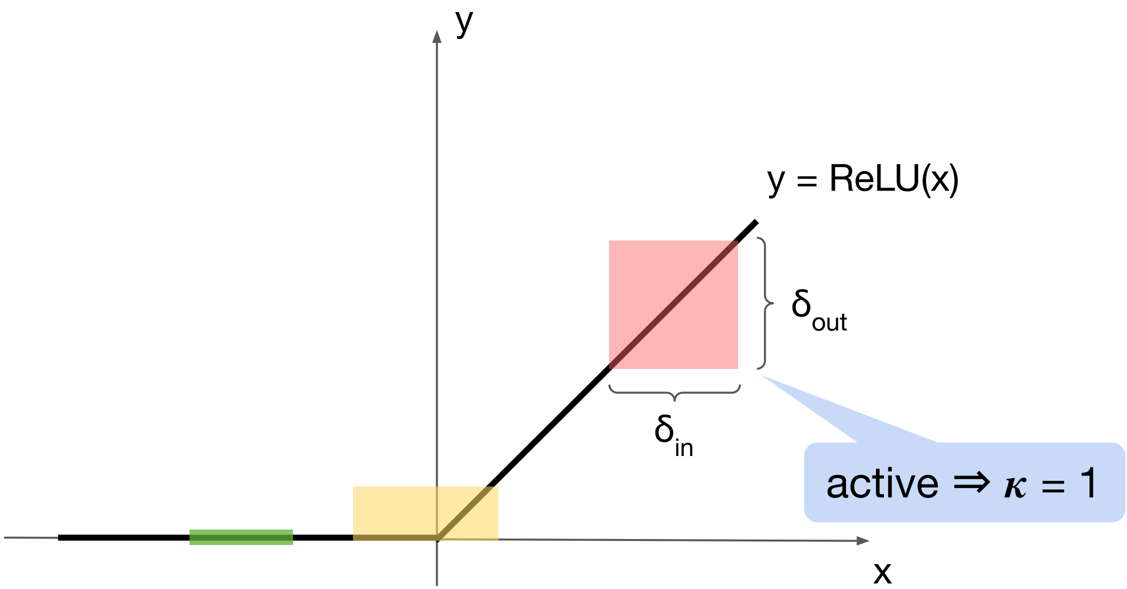
ReLU layers: κ depends on box scale and box centre: $= \max(\quad , 0)$

$\kappa \sim [0, 1]$

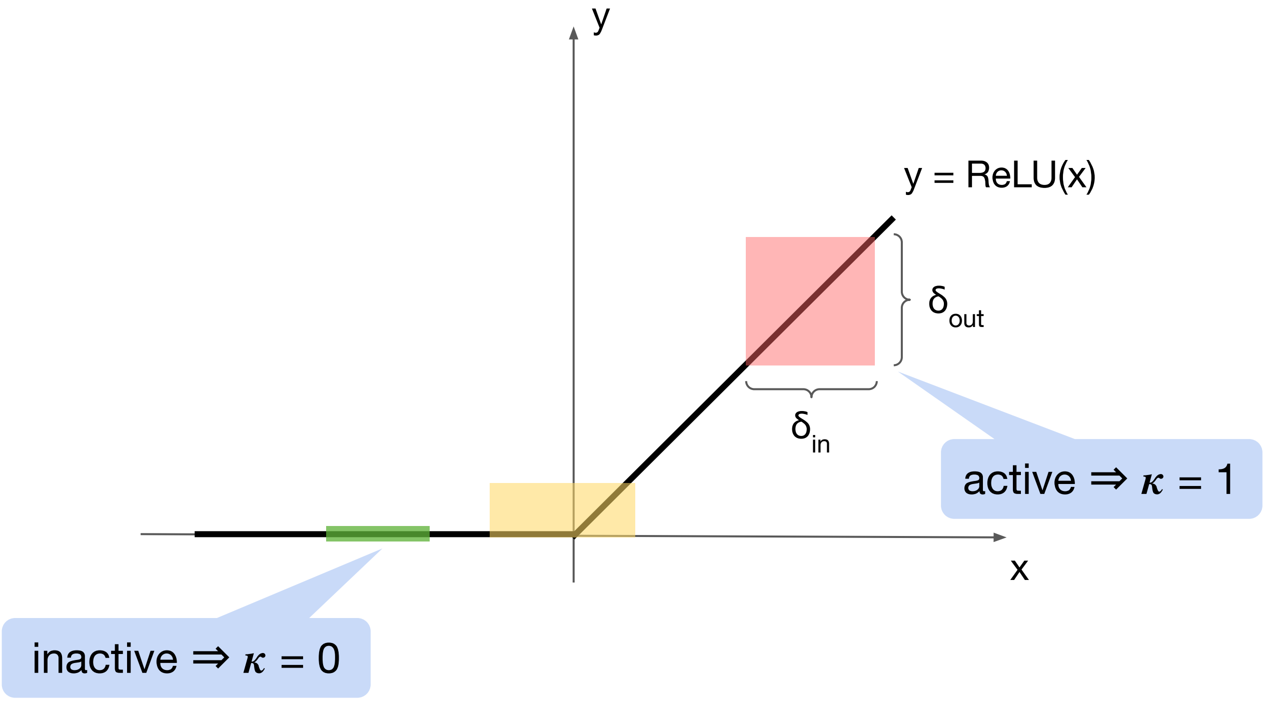
Box Abstraction Size Growth – ReLUs



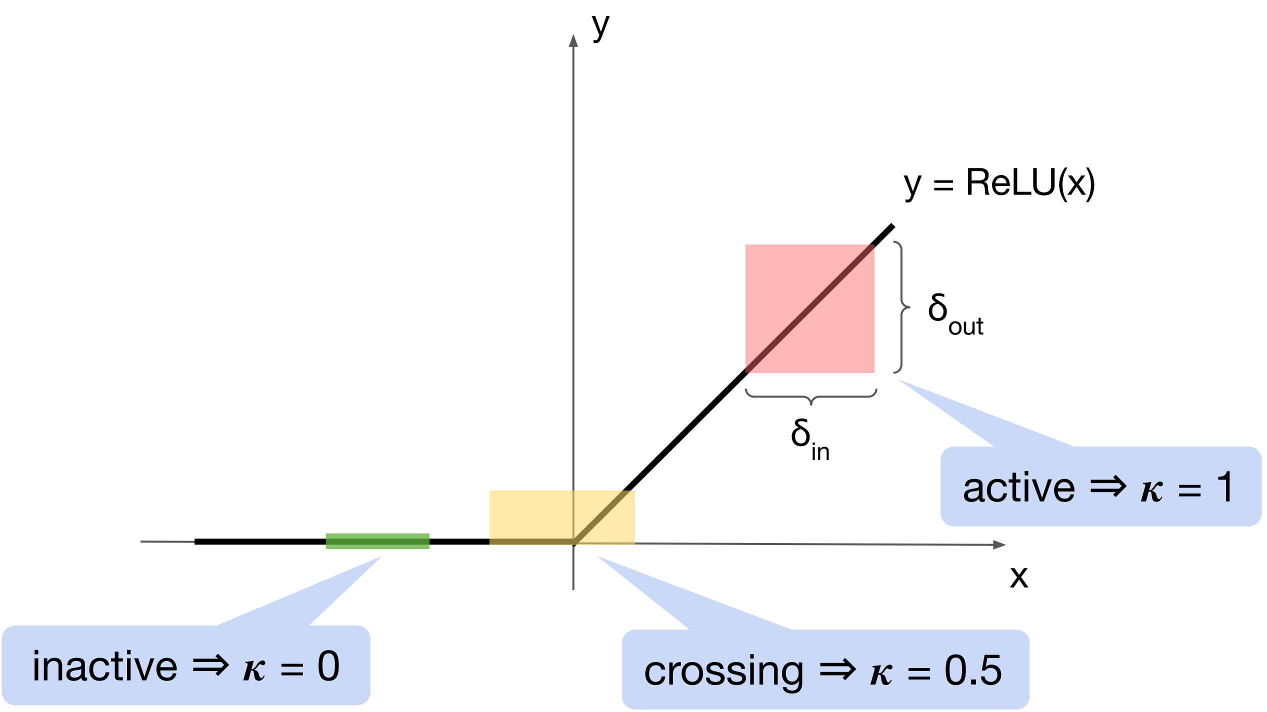
Box Abstraction Size Growth – ReLUs



Box Abstraction Size Growth – ReLUs



Box Abstraction Size Growth – ReLUs



Box Abstraction Size Growth – ReLUs

For input box sizes $\varepsilon \to 0$

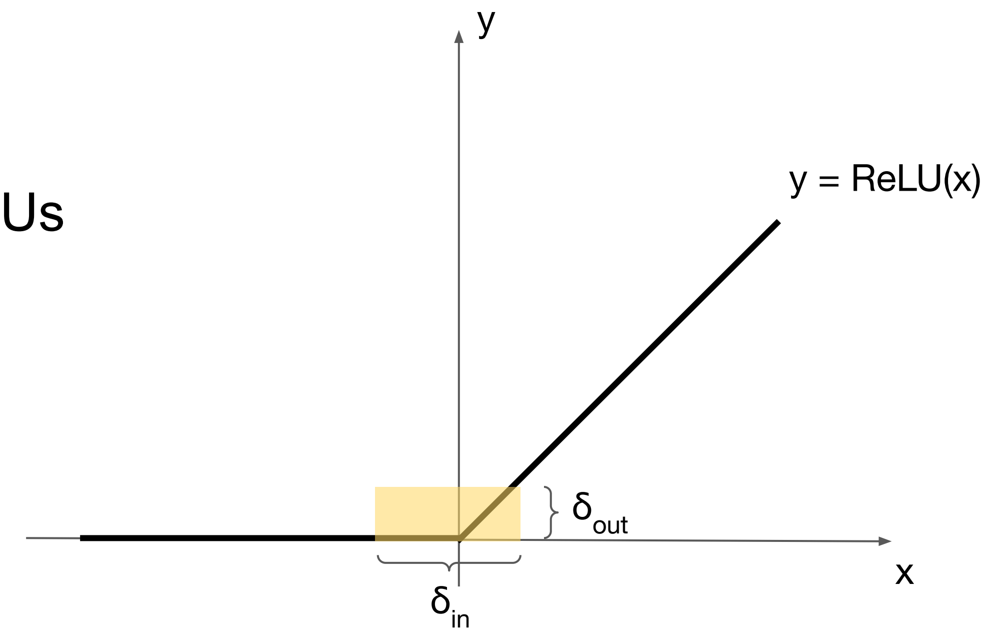
κ = Portion of active ReLUs

For input box sizes $\varepsilon \to \infty$

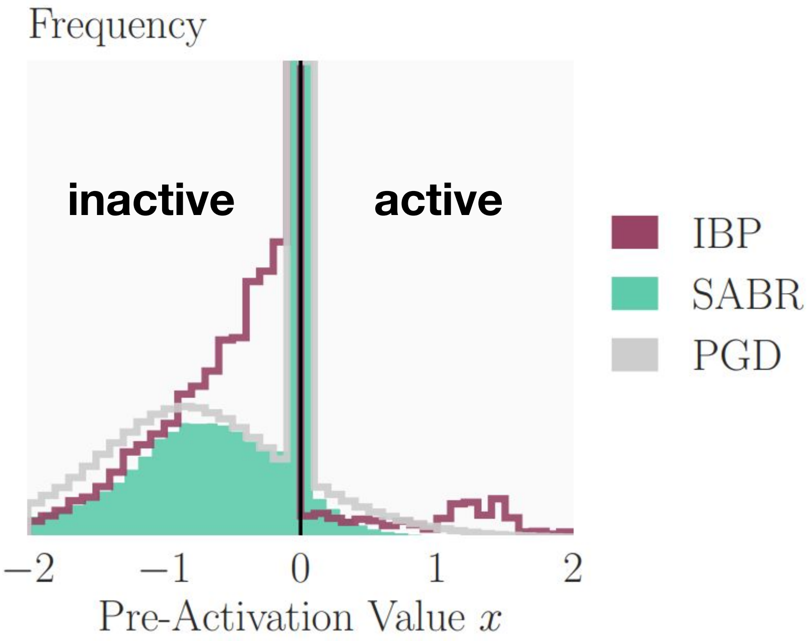
κ = 0.5

In-between:

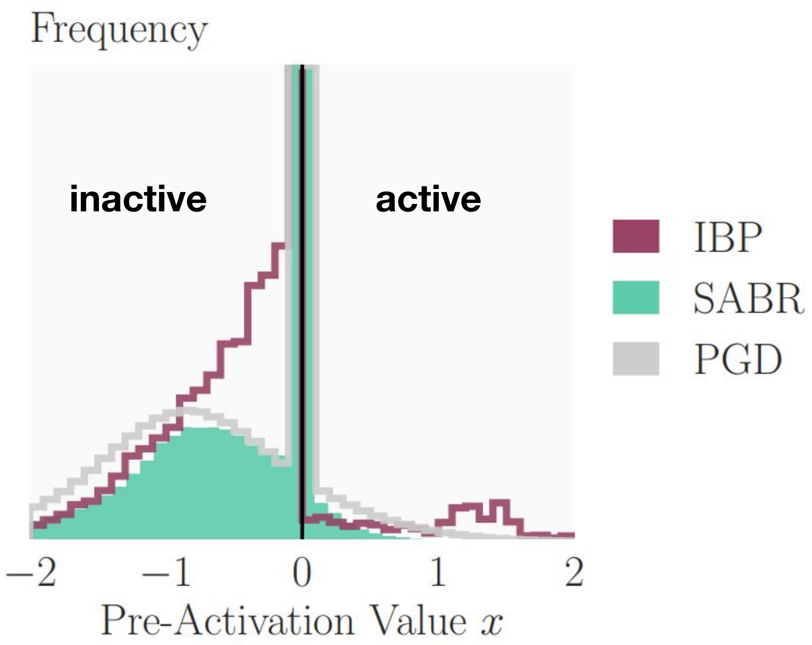
κ depends on box positions

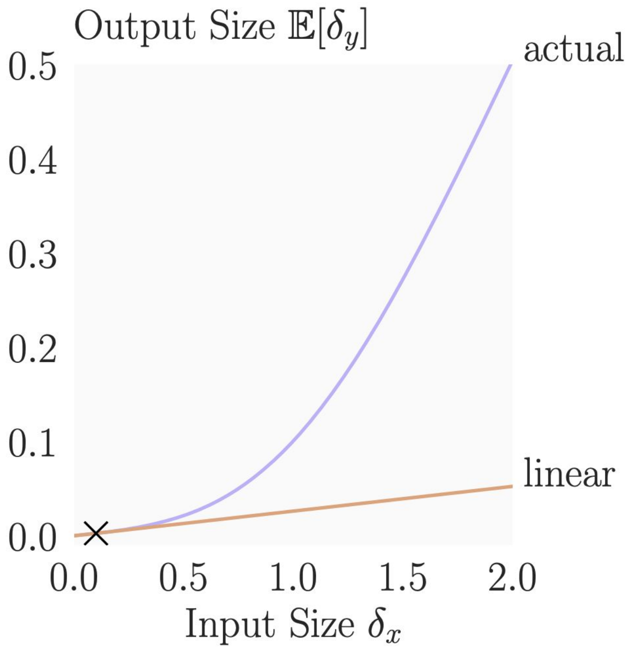


Box Abstraction Size Growth – ReLUs

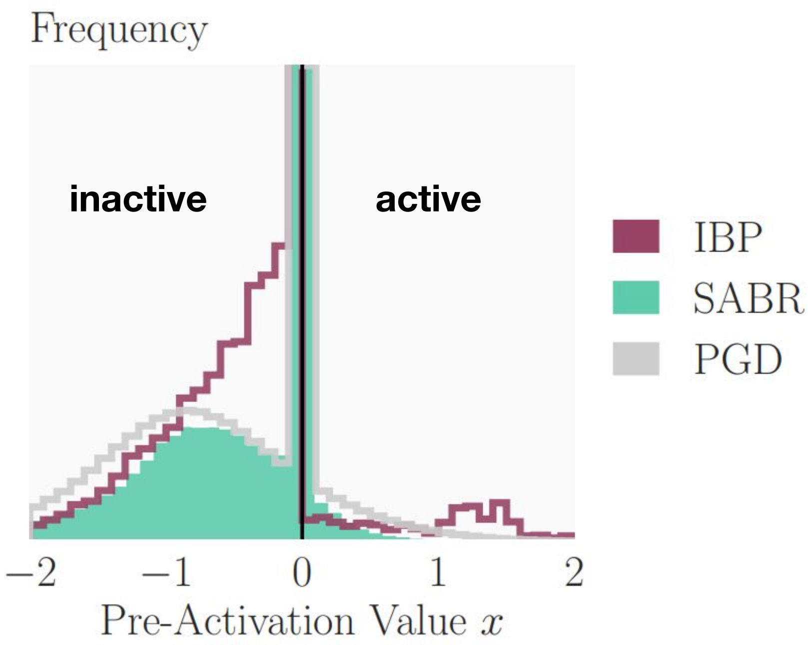


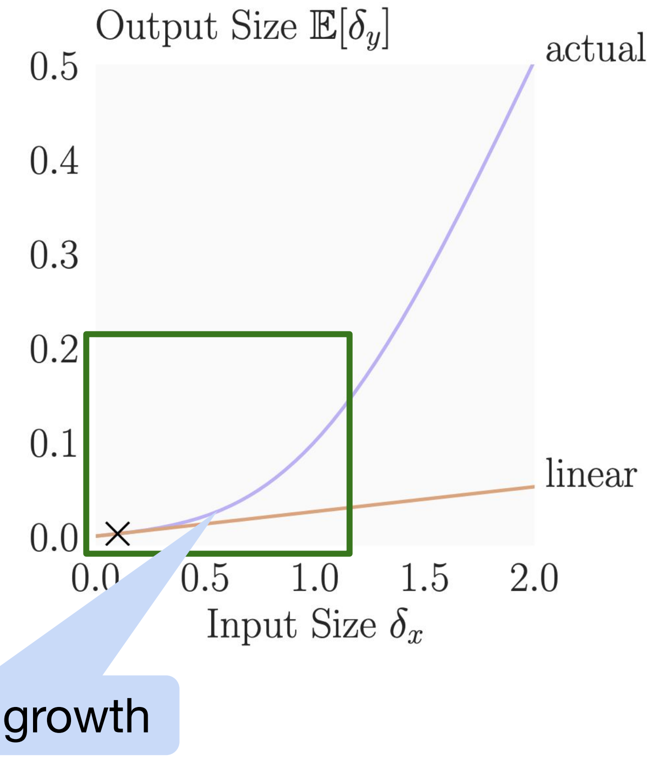
Box Abstraction Size Growth – ReLUs



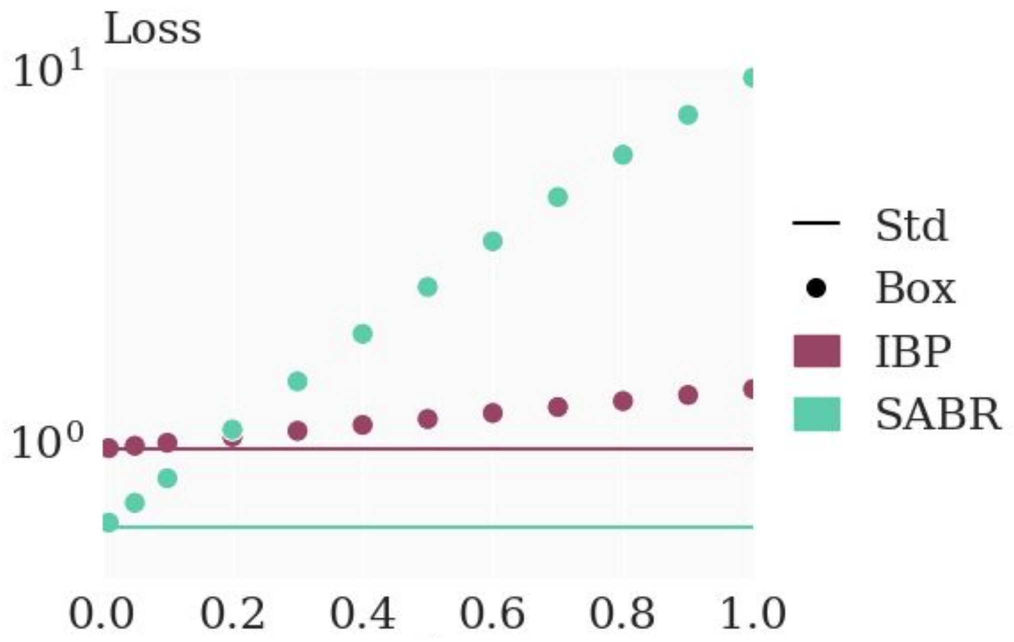


Box Abstraction Size Growth – ReLUs



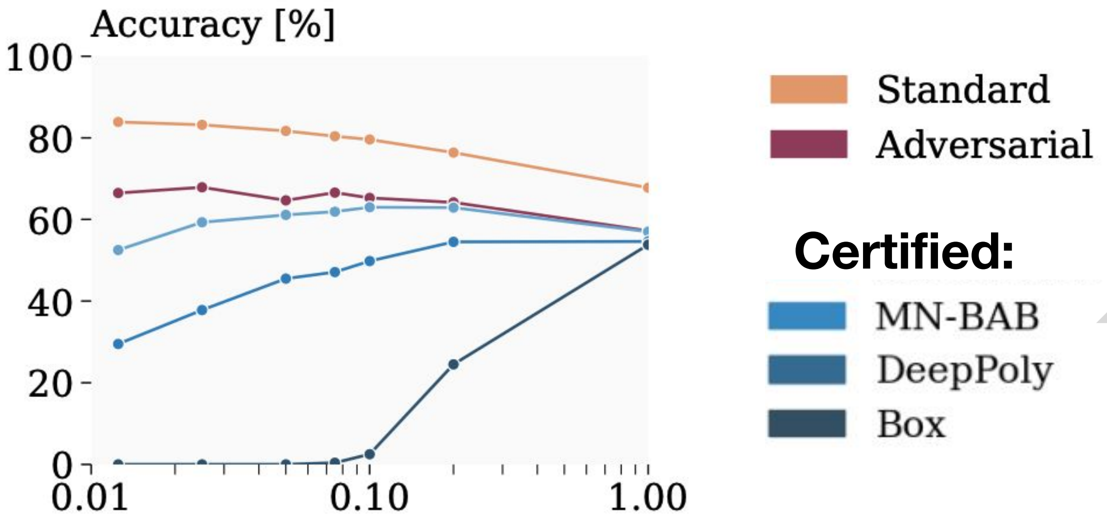


Full Network Loss Growth



Relative Box Size

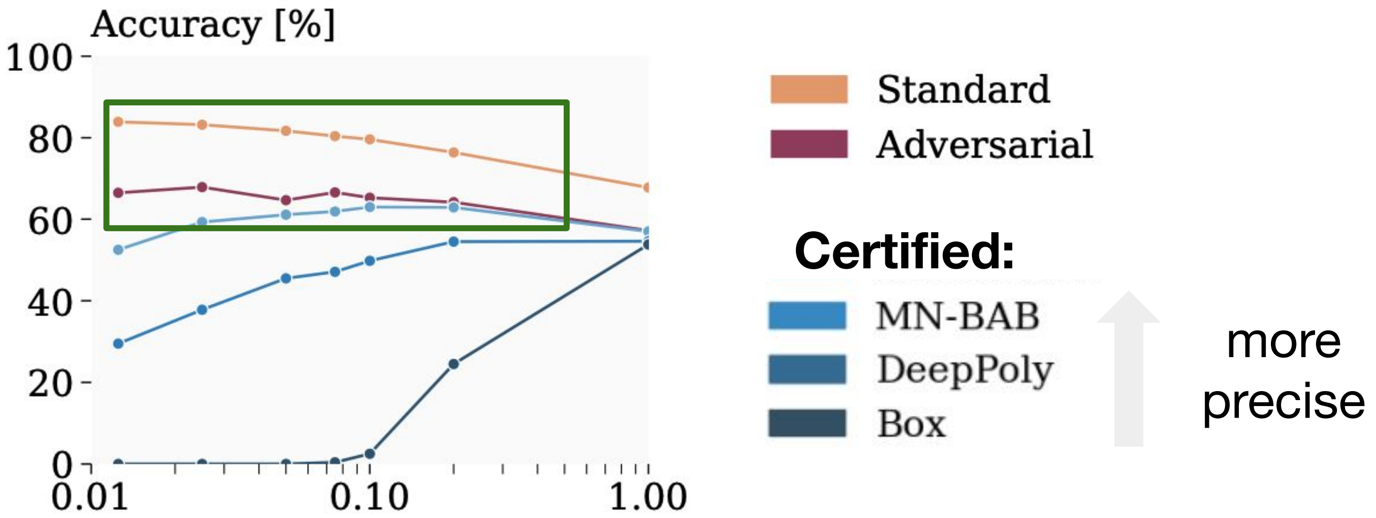
Impact of Verification Method



Training Box Size

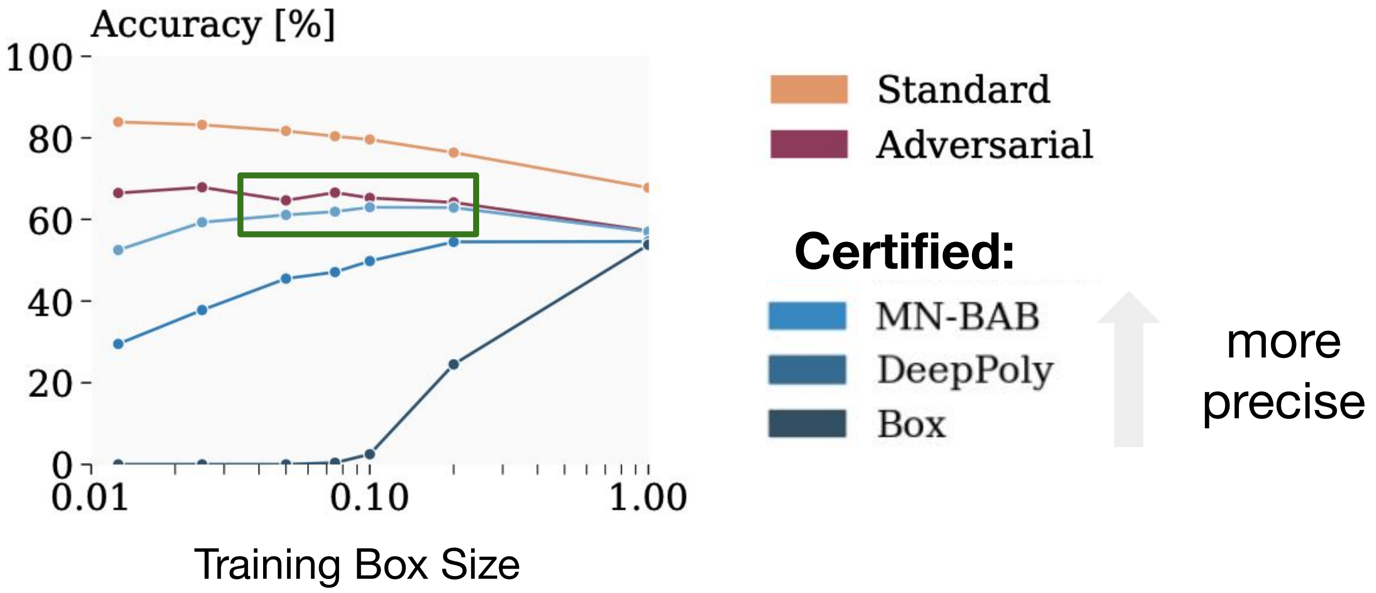
more precise

Impact of Verification Method

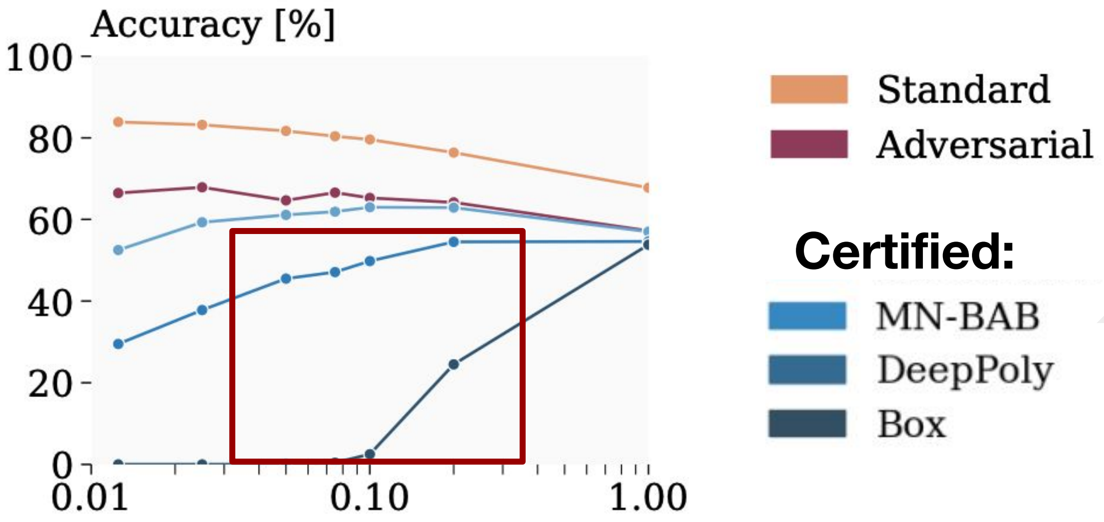


Training Box Size

Impact of Verification Method



Impact of Verification Method

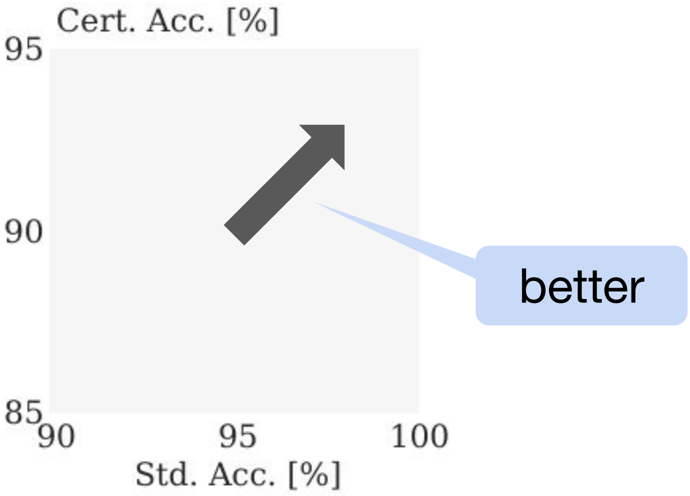


Training Box Size

Certified:

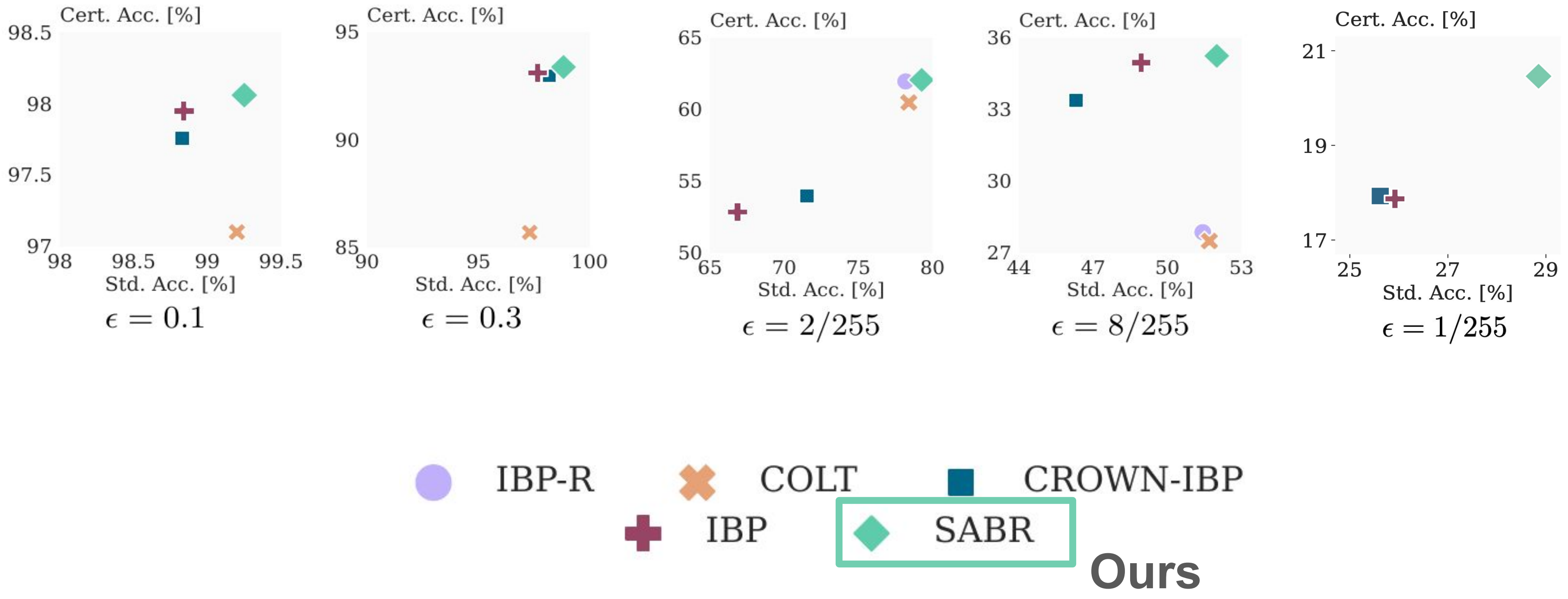
more precise

Empirical Results

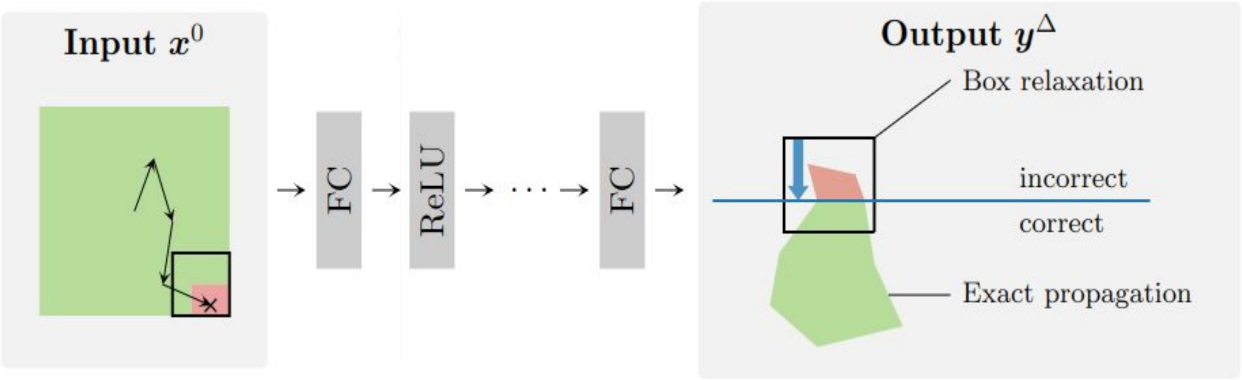


Empirical Results

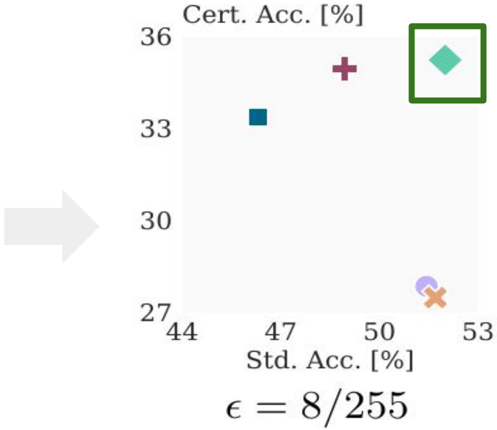
MNIST **CIFAR-10** **TinyImageNet**



Conclusion

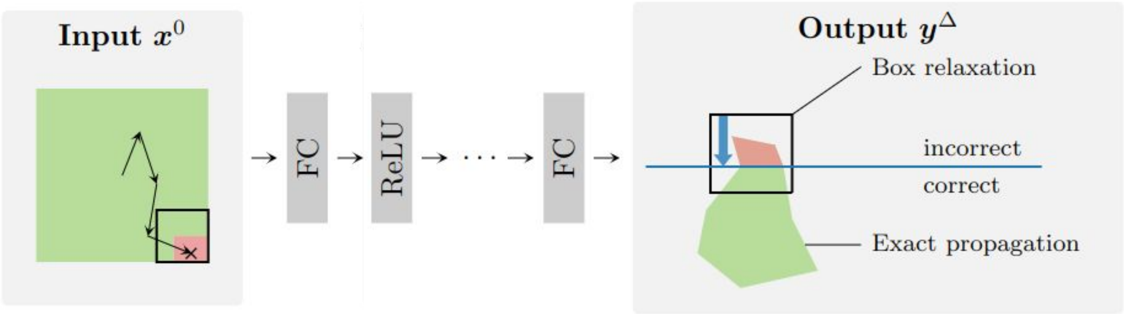


Input x^0
FC
ReLU
FC
Output y^Δ
Box relaxation
incorrect
correct
Exact propagation



Cert. Acc. [%]
36
33
30
27
44
47
50
53
Std. Acc. [%]
$\epsilon = 8/255$

Thank You For Your Attention!



Paper & Code:

https://www.sri.inf.ethz.ch/publications/mueller2022sabr

https://github.com/eth-sri/SABR