Certified Robust Models with Slack Control and Large Lipschitz Constant

Max Losch, David Stutz, Bernt Schiele & Mario Fritz

max planck institut informatik

CISPA HELMHOLTZ CENTER FOR INFORMATION SECURITY

1 Certified Robustness w/ Lipschitzness

Consider differentiable function f.

- f has Lipschitz constant K^f
- Then input distance bounds output distance:

$$\|f(x_1) - f(x_2)\|_p \leq K^f \|x_1 - x_2\|_p$$

- Enables cheap certification for ϵ, e.g.

$$|f_1(x) - f_2(x)| > \epsilon K^f$$

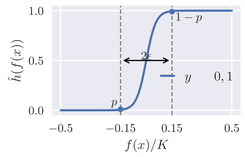
- Generally though, K^f is huge and $|f_1(x) - f_2(x)|$ is small $\implies$ certification infeasible

To achieve certification, existing methods focus on keeping K^f small, e.g. by:

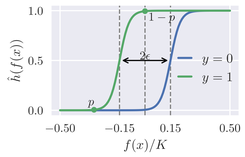
- minimizing it [3]
- constraining the architecture, such that $K^f = 1$ [5, 4, 1, 6]

2 Our Calibrated Lipschitz-Margin Loss (CLL)

- Existing losses not calibrated to input margin width 2ϵ
- Calibration of logistic distribution width σ via K^f
- Results in explicit definition for slack: p
- Slack determines constant of the whole model $K^{(h \circ f)} = 1/\sigma_\epsilon(p)$
- Easy to generalize to Cross-entropy-loss



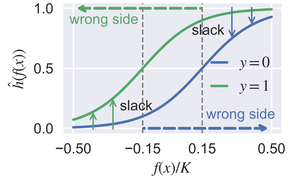
Calibrated logistic $2\epsilon = 0.3$.



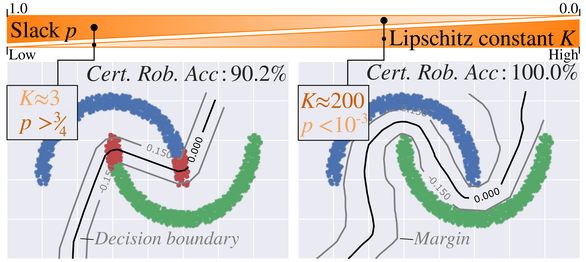
Calibrated with margin offset $\pm y\epsilon$

$$\hat{h}(f(x); y) = h\left(\underbrace{-\frac{y\epsilon}{\sigma_\epsilon(p)}}_{\text{Margin offset}} + \underbrace{\frac{1}{\sigma_\epsilon(p)}\frac{f(x)}{K^f}}_{\text{Output calibration}}\right)$$

$$\sigma_\epsilon(p) = \frac{2\epsilon}{h^{-1}(1-p) - h^{-1}(p)}$$



In a Nutshell



Existing Lipschitz margin methods control the Lipschitz constant K to be low, yet we observe decision functions becoming overly smooth when K is too low (left) – impairing accuracy.

Our Calibrated Lipschitz-Margin loss (CLL) provides slack control, which we show is inversely proportional to K. We can control K to be high and observe improved clean **and** robust accuracies (right). Incorrect or not robust samples marked red.

Contributions

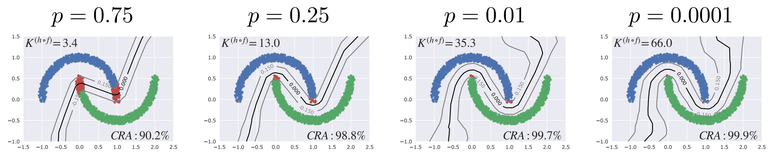
We propose a new loss with explicit control over slack and show:

- slack is tightly linked to the Lipschitz constant of the model
- we can train models with larger Lipschitz constants K
- models with larger K provide better clean and robust accuracies
- we achieve SOTA performance on CIFAR-100 and Tiny-ImageNet

Paper and Code: `github.com/mlosch/CLL`

[1] Alexandre Araujo et al. "A Unified Algebraic Perspective on Lipschitz Neural Networks". In: *ICLR* (2023).
[2] Yujia Huang et al. "Training Certifiably Robust Neural Networks with Efficient Local Lipschitz Bounds". In: *NeurIPS* (2021).
[3] Klas Leino, Zifan Wang, and Matt Fredrikson. "Globally-robust neural networks". In: *ICML* (2021).
[4] Laurent Meunier et al. "A Dynamical System Perspective for Lipschitz Neural Networks". In: *ICML* (2022).
[5] Sahil Singla, Surbhi Singla, and Soheil Feizi. "Improved deterministic l2 robustness on CIFAR-10 and CIFAR-100". In: *ICLR* (2021).
[6] Ruigang Wang and Ian Manchester. "Direct Parameterization of Lipschitz-Bounded Deep Networks". In: *ICML* (2023).

3 Explicit Slack and Lipschitz Constant Control



Slack control on two moons. High $p \Rightarrow$ high slack $\Rightarrow$ smooth decision boundary. To reach perfect certified robust accuracy (CRA), p needs to be small. Note, that $K^{(h \circ f)}$ is increasing with p.

4 Results

- Integrated CLL with existing methods
- Consistent certified robust accuracy (CRA) improvements while maintaining clean acc. thereby recording larger $K^{(h \circ f)}$
- CLL unlocks SOTA performance on lightweight models. 8C2F: $4M$, vs LBDN: $1.1B$ parameters

On Tiny-ImageNet	Model	Clean (%)	CRA $\frac{36}{255}$ (%)	CRA $\frac{72}{255}$ (%)	CRA $\frac{108}{255}$ (%)	$K^{(h \circ f)}$
GloRo[3]	8C2F	39.5	23.9	14.3	9.0	3.9
Local-Lip-B[2]	8C2F	36.9	23.4	12.7	6.1	-
Ours $\epsilon = 0.5, p = 0.01$	8C2F	**39.8 (+0.3)**	**25.9 (+2.0)**	**16.5 (+2.2)**	**10.7 (+1.7)**	10.6
SOC[5]	LipConv-10	32.1	21.5	12.4	7.5	6.4
	LipConv-20	31.7	21.0	12.9	7.5	6.4
Ours	LipConv-20	**32.6 (+0.5)**	**26.0 (+3.5)**	**20.2 (+7.3)**	**15.5 (+8.0)**	11.6
SLL[1]	XL	32.1	23.2	16.8	12.0	-
LBDN[6]	Sandwich	33.4	24.7	18.1	13.4	-
Ours $\epsilon = 1.0, p = 0.025$	8C2F	**33.5 (+0.1)**	**25.3 (+0.6)**	**19.0 (+0.9)**	**13.8 (+0.4)**	4.4

On CIFAR-100	Model	Clean (%)	CRA $\frac{36}{255}$ (%)	CRA $\frac{72}{255}$ (%)	CRA $\frac{108}{255}$ (%)	$K^{(h \circ f)}$
SOC[5]	LipConv-20	47.8	34.8	23.7	15.8	6.4
Ours	LipConv-20	**48.2 (+0.4)**	**35.1 (+0.3)**	**25.3 (+1.6)**	**18.3 (+2.5)**	9.2
CPL[4]	XL	47.8	33.4	20.9	12.6	1.6
Ours	XL	**47.9 (+0.1)**	**36.3 (+2.9)**	**28.1 (+7.2)**	**21.5 (+8.9)**	7.6
SLL[1]	XL	46.5	36.5	29.0	23.3	6.0
Ours	XL	**46.9 (+0.4)**	**36.6 (+0.1)**	29.0	**23.4 (+0.1)**	6.5