

SHIELD: Interval Bound Propagation for Certifiably Robust Continual Learning

Patryk Krukowski^{1,2,3}, Łukasz Gorczyca^{1,2}, Piotr Helm^{1,2},
Kamil Książek^{1,4} and Przemysław Spurek^{1,5}

¹ Jagiellonian University, Faculty of Mathematics and Computer Science
patryk.krukowski@doctoral.uj.edu.pl

² Jagiellonian University, Doctoral School of Exact and Natural Sciences
{lukasz.gorczyca, piotr.helm, kamil.ksiazek,

przemyslaw.spurek}@uj.edu.pl ³ Akces NCBR ⁴ Warsaw University of
Technology ⁵ IDEAS Research Institute

Keywords: Interval Bound Propagation, Interval Arithmetic, Certified Robustness, Neural Networks, Continual Learning, Wrapping Effect

Introduction

Interval Bound Propagation (IBP) [2] applies interval arithmetic [1] layer by layer through a neural network, producing rigorous output enclosures that certify robustness against adversarial perturbations within a given budget.

We present SHIELD [3], a framework that integrates IBP into a hypernetwork-based continual learning architecture, enabling certified adversarial robustness across a sequence of tasks without storing past data or gradients. As a core contribution, we introduce *Interval MixUp*, a training strategy that mitigates the wrapping effect [4] by controlling the radii of virtual interval-bounded inputs.

Interval propagation in neural networks

Consider a feedforward network of K layers, each applying $z_k = W_k z_{k-1} + b_k$, where W_k and b_k are the weights and biases in the $k \in \{1, \dots, K\}$

followed by a monotonic activation g . Given input $x \in \mathbb{R}^d$ ($z_0 = 0$) and perturbation radius $\varepsilon \geq 0$, the input enclosure is

$$I_\varepsilon(x) = \prod_{i=1}^d [x_i - \varepsilon, x_i + \varepsilon].$$

Using the midpoint μ_{k-1} and radius r_{k-1} , propagation gives:

$$\mu_k = W_k \mu_{k-1} + b_k, \quad r_k = |W_k| r_{k-1}, \quad (1)$$

where $|W_k|$ is the element-wise absolute value. For monotonic g : $\underline{z}_k = g(\mu_k - r_k)$, $\bar{z}_k = g(\mu_k + r_k)$. A prediction is *certifiably robust* if $\underline{z}_{K, y_{\text{true}}} > \max_{j \neq y_{\text{true}}} \bar{z}_{K, j}$.

SHIELD architecture

SHIELD consists of a **hypernetwork** $\mathcal{H}$ mapping trainable task embeddings $\mathbf{e}_t \in \mathbb{R}^N$ to weights $\boldsymbol{\theta}_t = \mathcal{H}(\mathbf{e}_t; \Phi)$ of a **target network** that processes interval-bounded inputs via IBP. Each task gets a dedicated classifier generated on demand, without storing past data.

The training loss combines: (i) standard cross-entropy on clean inputs, (ii) worst-case cross-entropy over IBP bounds (blended via an annealed κ), and (iii) a regularisation term penalising changes in the hypernetwork’s output for previous task embeddings, preventing catastrophic forgetting.

Interval MixUp: mitigating the wrapping effect

The wrapping effect [4] causes axis-aligned enclosures to grow overly conservative through nonlinear maps, weakening IBP’s certified guarantees. MixUp [5] trains on convex combinations

$$\tilde{x} = \lambda x_a + (1 - \lambda)x_b,$$

smoothing decision boundaries, but wrapping each $\tilde{x}$ in a full ε -ball exacerbates this effect. *Interval MixUp* scales the radius by proximity

to real data:

$$\varepsilon' = |2\lambda - 1| \cdot \varepsilon. \quad (2)$$

Near real samples ($\lambda \approx 0$ or 1): full radius; at the midpoint ($\lambda = 0.5$): $\varepsilon' = 0$. Smaller initial radii yield tighter enclosures after propagation, directly reducing the wrapping effect and closing the gap between classical and *verified* accuracy.

Robustness preservation across tasks

We prove that certified robustness is preserved across tasks: if the maximum logit change induced by a hypernetwork update $\mathbf{h}$ satisfies

$$\Delta_{\max}^{(s,t)}(x; \mathbf{h}) \leq \frac{1}{2}M(x, y_{\text{true}}; \boldsymbol{\theta}_{s,t}),$$

where M is the certified margin, then robustness guarantees for all previous tasks are maintained. The regularisation loss directly bounds this change.

Experimental results

We evaluate SHIELD on Permuted MNIST, Rotated MNIST, Split CIFAR-100, and Split miniImageNet (10 tasks each) under FGSM, PGD, and AutoAttack. SHIELD achieves state-of-the-art adversarial accuracy across all settings, reaching 57.9% under AutoAttack on Split miniImageNet versus 32.1% for the strongest baseline — nearly $2\times$ improvement — while maintaining competitive clean accuracy. Experiments on TinyImageNet (40 tasks) confirm scalability. Code: <https://github.com/pkrukowskii/SHIELD>.

References

- [1] R.E. Moore, R.B. Kearfott, M.J. Cloud. *Introduction to Interval Analysis*. SIAM, 2009.

- [2] S. Gowal et al. On the effectiveness of interval bound propagation for training verifiably robust models. *arXiv:1810.12715*, 2018.
- [3] P. Krukowski, Ł. Gorczyca, P. Helm, K. Książek, P. Spurek. SHIELD: Secure hypernetworks for incremental expansion learning defense. *arXiv:2506.08255*, 2025.
- [4] A. Neumaier. *The Wrapping Effect, Ellipsoid Arithmetic, Stability and Confidence Regions*. Springer, 1993.
- [5] H. Zhang, M. Cisse, Y.N. Dauphin, D. Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*, 2018.