Understanding Certified Training with Interval Bound Propagation

Yuhao Mao [1] **Mark Müller** [1] **Marc Fischer** [1] **Martin Vechev** [1]

Abstract

As robustness verification methods are becoming more precise, training certifiably robust neural networks is becoming ever more relevant. To this end, certified training methods compute and then optimize an upper bound on the worst-case loss over a robustness specification. Curiously, training methods based on the imprecise interval bound propagation (IBP) consistently outperform those leveraging more precise bounding methods. Still, we lack an understanding of the mechanisms making IBP so successful. In this work, we thoroughly investigate these mechanisms theoretically and empirically by leveraging a novel metric measuring the tightness of IBP bounds.

1. Introduction

As significant progress has been made on certifying neural networks (Zhang et al., 2022; Ferrari et al., 2022) against adversarial examples (Biggio et al., 2013; Szegedy et al., 2014), the focus in the field is shifting to the development of novel training methods that improve certifiable robustness while minimizing the accompanying reduction in accuracy.

Certified training aims to compute and then optimize approximations of the network's worst-case loss over an input region defined by an adversary specification. To this end, most methods compute an over-approximation of the network's reachable set using symbolic bound propagation methods (Singh et al., 2018; 2019; Gowal et al., 2018). Surprisingly, training methods based on the least precise bounds, obtained via interval bound propagation (IBP), empirically yield the best performance (Jovanovic et al., 2021).

This work We take a first step towards building a deeper understanding of the mechanisms underlying this surprising success of IBP training. To this end, we derive necessary and sufficient conditions on a network's weights under which IBP bounds become tight, a property we call *propagation invariance*, and prove that it implies an extreme regularization, agreeing well with the observed trade-off between certifiable robustness and accuracy. To investigate how close real networks are to full propagation invariance, we introduce the metric *propagation tightness* as the ratio of optimal and IBP bounds. This novel metric enables us to theoretically investigate the effects of model architecture, weight initialization, and training methods on IBP bound tightness for deep linear networks (DLNs). Conducting an extensive empirical study, we confirm the predictiveness of our theoretical results for deep ReLU networks and observe that: (i) increasing network width but not depth improves state-of-the-art certified accuracy, (ii) IBP (-based) training increases tightness, while (iii) non-IBP-based (certified) training methods do not increase tightness, leading to higher accuracy but worse robustness.

2. Background

We consider a classifer $\boldsymbol{f}\colon \mathbb{R}^{d_{\text{in}}} \mapsto \mathbb{R}^c$ predicting a numerical score $\boldsymbol{y} := \boldsymbol{f}(\boldsymbol{x})$ per class given an input $\boldsymbol{x} \in \mathcal{X} \subseteq \mathbb{R}^{d_{\text{in}}}$. A classifier $\boldsymbol{f}$ is *adversarially robust* if it predicts the target class t for all perturbed inputs $\boldsymbol{x}'$ in an ℓ_p-norm ball $\mathcal{B}_p^{\epsilon_p}(\boldsymbol{x})$ (we use $p = \infty$ and drop the subscript) of radius ϵ_p:

$$\arg\max_j \boldsymbol{f}(\boldsymbol{x}')_j = t, \quad \forall \boldsymbol{x}' \in \{\boldsymbol{x}' \in \mathcal{X} \mid \|\boldsymbol{x} - \boldsymbol{x}'\|_p \leq \epsilon_p\}.$$

IBP Certification (Gowal et al., 2018; Mirman et al., 2018) can be used to formally prove the robustness of a classifier $\boldsymbol{f}$ for a given input region $\mathcal{B}^{\epsilon}(\boldsymbol{x})$ by propagating it through $\boldsymbol{f}$ and computing BOX over-approximations (each dimension is described as an interval) of the state after every layer until we reach the output space. Consider an L-layer network $\boldsymbol{f} = \boldsymbol{h}_L \circ \boldsymbol{\sigma} \circ \boldsymbol{h}_{L-2} \circ \ldots \circ \boldsymbol{h}_1$, with linear layers $\boldsymbol{h}_i$ and ReLU activation functions $\boldsymbol{\sigma}$. After expressing the $\mathcal{B}^{\epsilon}(\boldsymbol{x})$ as BOX with radius $\boldsymbol{\delta}^0 := \epsilon$ and center $\dot{\boldsymbol{x}}^0 := \boldsymbol{x}$, such that we have $x_i^0 \in [x_i, \bar{x}_i] := [\dot{x}_i^0 - \delta_i^0, \dot{x}_i^0 + \delta_i^0]$ for the i^{th} dimension of the input. Propagating such a BOX through the linear layer $\boldsymbol{h}_i(\boldsymbol{x}^{i-1}) = \boldsymbol{W}\boldsymbol{x}^{i-1} + \boldsymbol{b} =: \boldsymbol{x}^i$, we obtain the output hyperbox with centre $\dot{\boldsymbol{x}}^i = \boldsymbol{W}\dot{\boldsymbol{x}}^{i-1} + \boldsymbol{b}$ and radius $\boldsymbol{\delta}^i = |\boldsymbol{W}|\boldsymbol{\delta}^{i-1}$, where $|\cdot|$ denotes the element-wise absolute value. To propagate a BOX through the ReLU activation

[1]Department of Computer Science, ETH Zürich, Zürich, Switzerland. Correspondence to: Yuhao Mao <yuhmao@ethz.ch>.

Presented at the 2^{nd} Workshop on Formal Verification of Machine Learning, co-located with the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA., 2023.

$\mathrm{ReLU}(\boldsymbol{x}^{i-1}) := \max(0, \boldsymbol{x}^{i-1})$, we propagate the lower and upper bound separately, resulting in an output Box with $\dot{\boldsymbol{x}}^i = \frac{\bar{\boldsymbol{x}}^i + \underline{\boldsymbol{x}}^i}{2}$ and $\boldsymbol{\delta}^i = \frac{\bar{\boldsymbol{x}}^i - \underline{\boldsymbol{x}}^i}{2}$ where $\underline{\boldsymbol{x}}^i = \mathrm{ReLU}(\dot{\boldsymbol{x}}^{i-1} - \boldsymbol{\delta}^{i-1})$ and $\bar{\boldsymbol{x}}^i = \mathrm{ReLU}(\dot{\boldsymbol{x}}^{i-1} + \boldsymbol{\delta}^{i-1})$. Thus obtaining an upper bound $\bar{\boldsymbol{y}}^\Delta$ on the logit difference $y_i^\Delta := y_i - y_t$, we can show robustness on the considered input region if $\bar{y}_i^\Delta < 0,\ \forall i \neq t$.

IBP Training To train for robustness, we, aim to minimize the expected *worst-case loss* for a given robustness specification, leading to a min-max optimization problem:

$$\theta_{\text{rob}} = \arg\min_\theta \mathbb{E}_{\mathcal{D}} \left[\max_{\boldsymbol{x}' \in \mathcal{B}^\epsilon(\boldsymbol{x})} \mathcal{L}_{\text{CE}}(\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}'), t) \right]. \quad (1)$$

As computing the worst-case loss by solving the inner maximization problem is generally intractable, it is commonly under- or over-approximated (e.g., by IBP bounds). Surprisingly, the imprecise IBP bounds consistently yield better performance (Shi et al., 2021a) than methods based on tighter approximations (Zhang et al., 2020; Balunovic & Vechev, 2020b; Wong et al., 2018), even though they severely reduce standard accuracies. Jovanovic et al. (2021) trace this back to the optimization problems induced by the more precise methods becoming intractable to solve. Recent training methods utilize IBP bounds, for regularization (Palma et al., 2022) or to compute more precise but unsound bound approximations (Müller et al., 2022b; Mao et al., 2023) to obtain state-of-the-art results.

3. Understanding IBP Training

We focus our theoretical analysis on deep linear networks (DLNs), *i.e.*, $\boldsymbol{f}(x) = \Pi_{i=1}^L \boldsymbol{W}^{(i)} \boldsymbol{x}$, popular for theoretical discussion of neural networks (Saxe et al., 2014; Ji & Telgarsky, 2019; Wu et al., 2019). While they are linear functions, they perfectly describe ReLU networks for infinitesimal perturbation magnitudes, retaining their layer-wise structure and joint non-convexity in the weights of different layers, making them a popular analysis tool (Ribeiro et al., 2016). We defer all proofs to App. B.

We define the optimal hyper-box $\mathrm{Box}^*(\boldsymbol{f}, \mathcal{B}^\epsilon(\boldsymbol{x}))$ as the smallest hyper-box $[\underline{\boldsymbol{z}}, \bar{\boldsymbol{z}}]$ such that it contains the image $\boldsymbol{f}(\boldsymbol{x}')$ of all points $\boldsymbol{x}'$ in $\mathcal{B}^\epsilon(\boldsymbol{x})$, *i.e.*, $\boldsymbol{f}(\boldsymbol{x}') \in [\underline{\boldsymbol{z}}, \bar{\boldsymbol{z}}], \forall \boldsymbol{x}' \in \mathcal{B}^\epsilon(\boldsymbol{x})$. Similarly, we define the layer-wise box approximation $\mathrm{Box}^\dagger$ as the result of applying the optimal approximation to every layer individually, in a recursive fashion $\mathrm{Box}^\dagger(\boldsymbol{f}, \mathcal{B}^\epsilon(\boldsymbol{x})) := \mathrm{Box}^*(\boldsymbol{W}_L, \mathrm{Box}^*(\cdots, \mathrm{Box}^*(\boldsymbol{W}^{(1)}, \mathcal{B}^\epsilon(\boldsymbol{x}))))$. We write their upper- and lower-bounds as $[\underline{\boldsymbol{z}}^*, \bar{\boldsymbol{z}}^*]$ and $[\underline{\boldsymbol{z}}^\dagger, \bar{\boldsymbol{z}}^\dagger]$, respectively. Optimal box bounds on the logit differences $\boldsymbol{y}^\Delta := \boldsymbol{y} - y_t$ are sufficient for exact verification (see Lemma A.1 in App. A). For DLNs, we can efficiently compute both optimal Box^* and layerwise $\mathrm{Box}^\dagger$ box bounds:

Theorem 3.1. *For an L-layer DLN $\boldsymbol{f} = \Pi_{k=1}^L \boldsymbol{W}^{(k)}$, we obtain the box centres $\dot{\boldsymbol{z}}^* = \dot{\boldsymbol{z}}^\dagger = \boldsymbol{f}(\boldsymbol{x})$ and the radii $\frac{\bar{\boldsymbol{z}}^* - \underline{\boldsymbol{z}}^*}{2} = \left|\Pi_{k=1}^L \boldsymbol{W}^{(k)}\right| \boldsymbol{\epsilon}$, and $\frac{\bar{\boldsymbol{z}}^\dagger - \underline{\boldsymbol{z}}^\dagger}{2} = \left(\Pi_{k=1}^L \left|\boldsymbol{W}^{(k)}\right|\right) \boldsymbol{\epsilon}$.*

Comparing the radius computations of the optimal and layer-wise approximations, we observe that the main difference lies in where the element-wise absolute value $|\cdot|$ of the weight matrix is taken. For the optimal box, we first multiply all weight matrices before taking the absolute value $|\Pi_{k=1}^L \boldsymbol{W}^{(k)}|$, thus allowing for cancellations of terms of opposite signs. For the layer-wise approximation, in contrast, we first take the absolute value of each weight matrix before multiplying them together $\Pi_{k=1}^L |\boldsymbol{W}^{(k)}|$, thereby losing all relational information between variables. Let us now investigate under which conditions layer-wise and optimal bounds become identical.

Propagation Invariance We call a network (globally) *propagation invariant* (PI) if the layer-wise and optimal box over-approximations are identical for every input box. Clearly, non-negative weight matrices lead to propagation invariant networks (Lin et al., 2022), as the absolute value in Theorem 3.1 loses its effect. However, non-negative weights significantly reduce network expressiveness and performance (Chorowski & Zurada, 2014), raising the question of whether they are a necessary condition. Indeed, we show that they are not necessary, by deriving the following sufficient *and* necessary condition for a two-layer DLN:

Lemma 3.2. *A two-layer DLN $\boldsymbol{f} = \boldsymbol{W}^{(2)}\boldsymbol{W}^{(1)}$ is propagation invariant if and only if for every fixed (i, j), we have $\left|\sum_k W_{i,k}^{(2)} \cdot W_{k,j}^{(1)}\right| = \sum_k |W_{i,k}^{(2)} \cdot W_{k,j}^{(1)}|$,* i.e., *either $W_{i,k}^{(2)} \cdot W_{k,j}^{(1)} \geq 0$ for all k or $W_{i,k}^{(2)} \cdot W_{k,j}^{(1)} \leq 0$ for all k.*

Conditions for Propagation Invariance To see how strict the condition described by Lemma 3.2 is, we observe that propagation invariance requires the sign of the last element in any two-by-two block in $\boldsymbol{W}^{(2)}\boldsymbol{W}^{(1)}$ to be determined by the signs of the other three elements:

Theorem 3.3. *Assume $\exists i, i', j, j'$, such that $W_{\cdot,j}^{(1)}$, $W_{\cdot,j'}^{(1)}$, $W_{i,\cdot}^{(2)}$ and $W_{i',\cdot}^{(2)}$ are all non-zero. If $(\boldsymbol{W}^{(2)}\boldsymbol{W}^{(1)})_{i,j} \cdot (\boldsymbol{W}^{(2)}\boldsymbol{W}^{(1)})_{i,j'} \cdot (\boldsymbol{W}^{(2)}\boldsymbol{W}^{(1)})_{i',j} \cdot (\boldsymbol{W}^{(2)}\boldsymbol{W}^{(1)})_{i',j'} < 0$, then $\boldsymbol{f} = \boldsymbol{W}^{(2)}\boldsymbol{W}^{(1)}$ is not propagation invariant.*

To obtain a propagation invariant network with the weights $\boldsymbol{W}^{(2)}, \boldsymbol{W}^{(1)} \in \mathcal{R}^{d \times d}$, we can thus only choose $2d - 1$ (e.g., one row and one column) of the d^2 signs freely (see Corollary A.2 in App. A). The statements of Lemma 3.2 and Theorem 3.3 naturally extend to DLNs with more than two layers $L > 2$. However, the conditions within Theorem 3.3 become increasingly complex and strict as more and more terms need to yield the same sign. Thus, we focus our analysis on $L = 2$ for clarity.

IBP Bound Tightness To analyze the tightness of IBP bounds for networks that do not satisfy the strict conditions for propagation invariance, we relax it to *propagation tightness* as the ratio between the optimal and layer-wise box radius, simply referred to as *tightness* in this paper.

Definition 3.4. *Given a DLN $\boldsymbol{f}$, we define the global propagation tightness $\boldsymbol{\tau}$ as the ratio between optimal* $\mathrm{Box}^*(\boldsymbol{f}, \mathcal{B}^{\boldsymbol{\epsilon}}(\boldsymbol{x}))$ *and layer-wise* $\mathrm{Box}^\dagger(\boldsymbol{f}, \mathcal{B}^{\boldsymbol{\epsilon}}(\boldsymbol{x}))$ *approximation radius,* i.e., $\boldsymbol{\tau} = \frac{\overline{\boldsymbol{z}}^* - \underline{\boldsymbol{z}}^*}{\overline{\boldsymbol{z}}^\dagger - \underline{\boldsymbol{z}}^\dagger}$.

Intuitively, tightness measures how much smaller the exact dimension-wise bounds Box^* are, compared to the layer-wise approximation $\mathrm{Box}^\dagger$, thus quantifying the gap between IBP certified and true adversarial robustness. When tightness equals 1, the network is propagation invariant and can be certified exactly with IBP; when tightness is close to 0, IBP bounds become arbitrarily imprecise.

ReLU Networks The nonlinearity of ReLU networks leads to locally varying tightness and makes the computation of optimal box bounds intractable. However, for infinitesimal perturbation magnitudes, the activation patterns of ReLU networks remain stable, making them locally linear. We thus introduce a local version of tightness around concrete inputs, which we will later use to confirm the applicability of our results on DLNs to ReLU networks.

Definition 3.5. *For an L-layer ReLU network with weight matrices $\boldsymbol{W}^{(k)}$ and activation pattern $\boldsymbol{d}^{(k)}(\boldsymbol{x}) = \mathbb{1}_{\boldsymbol{x}^{(k-1)}>0} \in \{0,1\}^{d_k}$ (1 for active and 0 for inactive) depending on the input $\boldsymbol{x}$, we define its local tightness as*

$$\boldsymbol{\tau} = \frac{\frac{d}{d\epsilon}(\overline{\boldsymbol{z}}^* - \underline{\boldsymbol{z}}^*)\big|_{\epsilon=0}}{\frac{d}{d\epsilon}(\overline{\boldsymbol{z}}^\dagger - \underline{\boldsymbol{z}}^\dagger)\big|_{\epsilon=0}} = \frac{\left|\Pi_{k=1}^{L} \operatorname{diag}(\boldsymbol{d}^{(k)})\boldsymbol{W}^{(k)}\right| \mathbf{1}}{(\Pi_{k=1}^{L} \operatorname{diag}(\boldsymbol{d}^{(k)}) \left|\boldsymbol{W}^{(k)}\right|)\mathbf{1}}.$$

Beyond the results discussed here, this metric enables our study of how architecture impacts propagation tightness at initialization and allows us to show that IBP training increases tightness (see App. A.3 and A.4, respectively).

We, now, study the reconstruction loss of a linear embedding into a low-dimensional subspace as a proxy for the network performance as many high-dimensional computer vision datasets were shown to possess a small intrinsic data dimensionality (Pope et al., 2021). Let us consider a k-dimensional data distribution, linearly embedded into a d dimensional space with $d \gg k$, *i.e.*, the data matrix X has a rank-k eigendecomposition $\mathrm{Var}(X) = U\Lambda U^\top$. We know that in this setting, the optimal reconstruction $\hat{X} = U_k U_k^\top X$ of the original data is exact for rank k matrices and obtained by choosing U_k as the k columns of U associated with the non-zero eigenvalues. Interestingly, this is not the case even for optimal box propagation:

Theorem 3.6. *Consider the linear embedding and reconstruction $\hat{\boldsymbol{x}} = U_k U_k^\top \boldsymbol{x}$ of a d dimensional data distribution $\boldsymbol{x} \sim \mathcal{X}$ into a k dimensional space with $d \gg k$ and eigenmatrices U drawn uniformly at random from the orthogonal group. Propagating the input box $\mathcal{B}^{\boldsymbol{\epsilon}}(\boldsymbol{x})$ layer-wise and optimally, thus, yields $\mathcal{B}^{\boldsymbol{\delta}^\dagger}(\hat{\boldsymbol{x}})$, and $\mathcal{B}^{\boldsymbol{\delta}^*}(\hat{\boldsymbol{x}})$, respectively. Then, we have, (i) $\mathbb{E}(\delta_i/\epsilon) = ck \in \Theta(k)$ for a positive constant c depending solely on d and $c \to \frac{2}{\pi} \approx 0.64$ for large d; and (ii) $\mathbb{E}(\delta_i^*/\epsilon) \to \frac{2}{\sqrt{\pi}} \frac{\Gamma(\frac{1}{2}(k+1))}{\Gamma(\frac{1}{2}k)} \in \Theta(\sqrt{k})$.*

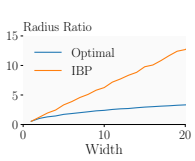

Figure 1: Box reconst. error on bottleneck width w.

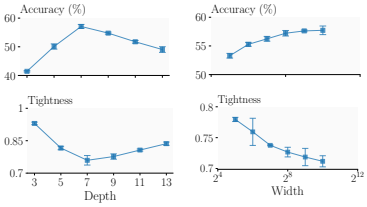

Figure 3: Effect of network depth and width on tightness and training set IBP-certified accuracy.

Intuitively, Theorem 3.6 implies that while input points can be embedded into and reconstructed from a k dimensional space losslessly, box propagation will yield a box growth of $\Theta(\sqrt{k})$ for optimal and $\Theta(k)$ for layer-wise propagation. However, as soon as we have $k = d$, we can choose U_k to be an identity matrix, thus obtaining lossless "reconstruction", even for layer-wise propagation. This highlights that sufficient network width is crucial for box propagation, even in the linear setting.

4. Empirical Evaluation Analysis

Here, we leverage our novel tightness metric and specifically its local variant (see Definition 3.5) to gain a deeper understanding of IBP-based training methods and confirm that our analysis of DLNs carries over to ReLU networks. We defer details of our experimental setup to App. C.

Network Architechture We first confirm our predictions on the inherent hardness of linear reconstruction in Figure 1, where we plot the ratio of recovered and original box radius for optimal and IBP propagation, given a bottleneck layer of width w and data with intrinsic dimensionality $k = w$. As predicted by Theorem 3.6, IBP propagation leads to linear growth while optimal propagation yields sublinear growth.

Next, we study the interaction of network architecture and IBP training. To this end, we train networks with 3 to 13 layers on CIFAR-10 for $\epsilon = 2/255$, visualizing results in Figure 3 (left), reporting the IBP-certified accuracy on the training set as a measure of the goodness of fit. Generally, we would expect that increasing network depth increases capacity, thus reducing the robust training loss and increasing training set IBP-certified accuracy. However, we only observe such an increase in accuracy until a depth of 7 layers before accuracy starts to drop. This maximum in accuracy

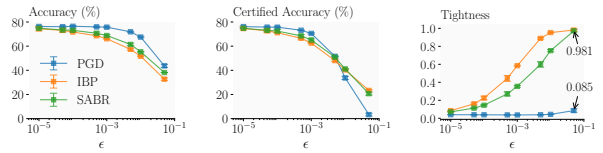

Figure 4: Standard and certified accuracy and tightness for CNN3 on CIFAR-10 depending on training method and perturbation magnitude ϵ used for training and evaluation.

coincides with a minimum of tightness, agreeing well with the popularity of the 7-layer CNN7 in literature (Gowal et al., 2018; Shi et al., 2021b; Müller et al., 2022b).

Varying the width of a standard CNN7 (using IBP training), we observe in Figure 3 (right) that increasing capacity via width yields a monotone although diminishing increase in accuracy as tightness decreases gradually. The different trends for width and depth increases agree well with our theoretical results showing that, at initialization, tightness decreases exponentially with depth (Corollary A.4) and polynomially with width (Lemma A.3) and predicting that sufficient network width is essential for trained networks (see Theorem 3.6). Intuitively, this suggests that less regularization is required to offset the tightness penalty of increasing network width rather than depth. We thus increase the width of a SABR-trained CNN7 and improve upon their SOTA performance on a magnitude comparable to years of progress on certified training methods: On MNIST ($\epsilon = 0.3$) $4\times$ width pushes certified accuracy from 93.38 to **93.85** and on CIFAR-10 ($\epsilon = \frac{2}{255}$) $2\times$ width yields $62.84 \rightarrow$ **63.28**.

Certified Training Increases Tightness Next, we compare IBP, PGD (Madry et al., 2018), and SABR training, on CNN3 for CIFAR-10 across a wide range of perturbation magnitudes ($\epsilon \in [10^{-5}, 5 \cdot 10^{-2}]$), illustrating results in Figure 4. While IBP propagates the whole input region, SABR propagates only a small subregion via IBP. PGD, in contrast, trains with samples that approximately yield the worst-case loss. We observe that IBP-based methods increase tightness with perturbation magnitude until networks become almost propagation invariant for $\epsilon = 0.05$ with a tightness of $\tau = 0.98$. In contrast, PGD barely influences tightness. The regularization required to yield such high tightness also severely reduces standard accuracies. However, while this reduced standard accuracy translates to smaller certified accuracies for very small perturbation magnitudes ($\epsilon \leq 5 \cdot 10^{-3}$), the increased tightness improves certifiability sufficiently to yield higher certified accuracies for larger perturbation magnitudes ($\epsilon \geq 10^{-2}$). We further investigate this dependency between (certified) robustness and tightness by varying the subselection ratio λ when training with SABR, where $\lambda = 1$ recovers IBP and $\lambda = 0$ PGD. In Figure 5, we observe that while decreasing λ severely reduces tightness and thus regularization, it not only leads to increasing natural but also certified accuracies until tightness falls below $\tau < 0.5$ at $\lambda = 0.4$. This highlights that reducing tightness while maintaining sufficient certifiability is a promising path to new certified training methods. In App. D.1 we show simlar trends when varying the regularization level for other training methods.

We now investigate whether non-IBP-based certified training methods affect a similar increase in tightness as IBP-based methods. To this end, we consider COLT (Balunovic & Vechev, 2020a) which combines precise ZONOTOPE bounds with adversarial training. However, as COLT does not scale to the popular CNN7, we compare with it on their 4-layer CNN architecture. In Table 1, we observe that the ordering of tightness and accuracy is exactly inverted, thus highlighting the large accuracy penalty associated with the strong regularization for tightness. While COLT only affects a minimal increase in tightness compared to SABR or IBP, it still yields networks, an order of magnitude tighter than PGD, suggesting that slightly increased tightness might be desirable for certified robustness. This is further corroborated by the observation that while COLT reaches the highest certified accuracies at small perturbation magnitudes, the more heavily regularizing SABR performs better at larger radii.

Table 1: Multiple training methods.

Method	ϵ	Accuracy	Tightness	Certified
PGD	2/255	81.2	0.001	-
	8/255	69.3	0.007	-
COLT	2/255	78.4	0.009	60.7
	8/255	51.7	0.057	26.7
SABR	2/255	75.6	0.182	57.7
	8/255	48.2	0.950	31.2
IBP	2/255	63.0	0.803	51.3
	8/255	42.2	0.977	31.0

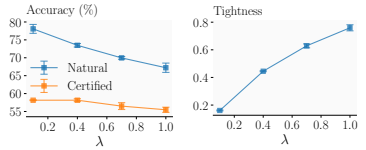

Figure 5: Acc. and tightness of CNN7 for CIFAR-10 $\epsilon = \frac{2}{255}$ on regularization strength λ with SABR.

For a detailed discussion of related work, please see App. E.

5. Conclusion

Motivated by the surprising dominance of IBP-based certified training methods, we investigated its underlying mechanisms. Quantifying the tightness of IBP compared to optimal BOX bounds, we were able to predict the influence of architecture choices on deep linear networks at initialization and after training. Our experimental results confirm the applicability of these results to ReLU networks and show that wider networks improve the performance of state-of-the-art methods, while deeper networks do not. Finally, we show that IBP-based certified training methods, in contrast to non-IBP-based methods, significantly increase propagation tightness at the cost of strong regularization. We believe that this insight and the novel metric of propagation tightness will constitute a key step towards developing novel and more effective certified training methods.

References

Baader, M., Mirman, M., and Vechev, M. T. Universal approximation with certified networks. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=B1gX8kBtPr.

Balunovic, M. and Vechev, M. Adversarial training and provable defenses: Bridging the gap. In *International Conference on Learning Representations*, 2020a. URL https://openreview.net/forum?id=SJxSDxrKDr.

Balunovic, M. and Vechev, M. T. Adversarial training and provable defenses: Bridging the gap. In *Proc. of ICLR*, 2020b.

Biggio, B., Corona, I., Maiorca, D., Nelson, B., Srndic, N., Laskov, P., Giacinto, G., and Roli, F. Evasion attacks against machine learning at test time. In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2013, Prague, Czech Republic, September 23-27, 2013, Proceedings, Part III*, volume 8190, 2013. doi: 10.1007/978-3-642-40994-3_25.

Chorowski, J. and Zurada, J. M. Learning understandable neural networks with nonnegative weight constraints. *IEEE transactions on neural networks and learning systems*, 26(1):62–69, 2014. doi: 10.1109/TNNLS.2014.2310059.

Cook, J. Rational formulae for the production of a spherically symmetric probability distribution. *Mathematics of Computation*, 11(58):81–82, 1957.

Ferrari, C., Müller, M. N., Jovanovic, N., and Vechev, M. T. Complete verification via multi-neuron relaxation guided branch-and-bound. In *ICLR*. OpenReview.net, 2022.

Glorot, X. and Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. In Teh, Y. W. and Titterington, D. M. (eds.), *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2010, Chia Laguna Resort, Sardinia, Italy, May 13-15, 2010*, volume 9 of *JMLR Proceedings*, pp. 249–256. JMLR.org, 2010. URL http://proceedings.mlr.press/v9/glorot10a.html.

Gowal, S., Dvijotham, K., Stanforth, R., Bunel, R., Qin, C., Uesato, J., Arandjelovic, R., Mann, T. A., and Kohli, P. On the effectiveness of interval bound propagation for training verifiably robust models. *ArXiv preprint*, abs/1810.12715, 2018.

He, K., Zhang, X., Ren, S., and Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pp. 1026–1034. IEEE Computer Society, 2015. doi: 10.1109/ICCV.2015.123. URL https://doi.org/10.1109/ICCV.2015.123.

Ji, Z. and Telgarsky, M. Gradient descent aligns the layers of deep linear networks. In *ICLR (Poster)*. OpenReview.net, 2019.

Jovanovic, N., Balunovic, M., Baader, M., and Vechev, M. T. Certified defenses: Why tighter relaxations may hurt training? *ArXiv preprint*, abs/2102.06700, 2021.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

LeCun, Y., Cortes, C., and Burges, C. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

Lin, V., Ivanov, R., Weimer, J., Sokolsky, O., and Lee, I. T4V: exploring neural network architectures that improve the scalability of neural network verification. In Raskin, J., Chatterjee, K., Doyen, L., and Majumdar, R. (eds.), *Principles of Systems Design - Essays Dedicated to Thomas A. Henzinger on the Occasion of His 60th Birthday*, volume 13660 of *Lecture Notes in Computer Science*, pp. 585–603. Springer, 2022. doi: 10.1007/978-3-031-22337-2_28. URL https://doi.org/10.1007/978-3-031-22337-2_28.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. In *ICLR (Poster)*. OpenReview.net, 2018.

Mao, Y., Fu, C., Wang, S., Ji, S., Zhang, X., Liu, Z., Zhou, J., Liu, A. X., Beyah, R., and Wang, T. Transfer attacks revisited: A large-scale empirical study in real computer vision settings. In *IEEE Symposium on Security and Privacy*, pp. 1423–1439. IEEE, 2022.

Mao, Y., Müller, M. N., Fischer, M., and Vechev, M. Taps: Connecting certified and adversarial training, 2023.

Marsaglia, G. Choosing a point from the surface of a sphere. *The Annals of Mathematical Statistics*, 43(2):645–646, 1972.

Mirman, M., Gehr, T., and Vechev, M. T. Differentiable abstract interpretation for provably robust neural networks. In *ICML*, volume 80 of *Proceedings of Machine Learning Research*, pp. 3575–3583. PMLR, 2018.

Mirman, M. B., Baader, M., and Vechev, M. The fundamental limits of neural networks for interval certified robustness. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=fsacLLU35V.

Müller, M. N., Brix, C., Bak, S., Liu, C., and Johnson, T. T. The third international verification of neural networks competition (VNN-COMP 2022): Summary and results. *CoRR*, abs/2212.10376, 2022a. doi: 10.48550/arXiv.2212.10376. URL https://doi.org/10.48550/arXiv.2212.10376.

Müller, M. N., Eckert, F., Fischer, M., and Vechev, M. T. Certified training: Small boxes are all you need. *CoRR*, abs/2210.04871, 2022b.

Palma, A. D., Bunel, R., Dvijotham, K., Kumar, M. P., and Stanforth, R. IBP regularization for verified adversarial robustness via branch-and-bound. *ArXiv preprint*, abs/2206.14772, 2022.

Pinelis, I. and Molzon, R. Optimal-order bounds on the rate of convergence to normality in the multivariate delta method, 2016.

Pope, P., Zhu, C., Abdelkader, A., Goldblum, M., and Goldstein, T. The intrinsic dimension of images and its impact on learning. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=XJk19XzGq2J.

Ribeiro, M. T., Singh, S., and Guestrin, C. "why should I trust you?": Explaining the predictions of any classifier. In Krishnapuram, B., Shah, M., Smola, A. J., Aggarwal, C. C., Shen, D., and Rastogi, R. (eds.), *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pp. 1135–1144. ACM, 2016. doi: 10.1145/2939672.2939778. URL https://doi.org/10.1145/2939672.2939778.

Saxe, A. M., McClelland, J. L., and Ganguli, S. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *ICLR*, 2014.

Shi, Z., Wang, Y., Zhang, H., Yi, J., and Hsieh, C. Fast certified robust training via better initialization and shorter warmup. *ArXiv preprint*, abs/2103.17268, 2021a.

Shi, Z., Wang, Y., Zhang, H., Yi, J., and Hsieh, C. Fast certified robust training with short warmup. In *NeurIPS*, pp. 18335–18349, 2021b.

Singh, G., Gehr, T., Mirman, M., Püschel, M., and Vechev, M. T. Fast and effective robustness certification. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, 2018.

Singh, G., Gehr, T., Püschel, M., and Vechev, M. T. An abstract domain for certifying neural networks. *Proc. ACM Program. Lang.*, 3(POPL):41:1–41:30, 2019.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. J., and Fergus, R. Intriguing properties of neural networks. In *ICLR (Poster)*, 2014.

Wang, Y., Shi, Z., Gu, Q., and Hsieh, C. On the convergence of certified robust training with interval bound propagation. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022a. URL https://openreview.net/forum?id=YeShU5mLfLt.

Wang, Z., Albarghouthi, A., Prakriya, G., and Jha, S. Interval universal approximation for neural networks. *Proc. ACM Program. Lang.*, 6(POPL):1–29, 2022b. doi: 10.1145/3498675. URL https://doi.org/10.1145/3498675.

Wong, E. and Kolter, J. Z. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *ICML*, volume 80 of *Proceedings of Machine Learning Research*, pp. 5283–5292. PMLR, 2018.

Wong, E., Schmidt, F. R., Metzen, J. H., and Kolter, J. Z. Scaling provable adversarial defenses. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, 2018.

Wu, L., Wang, Q., and Ma, C. Global convergence of gradient descent for deep linear residual networks. In *NeurIPS*, pp. 13368–13377, 2019.

Zhang, H., Chen, H., Xiao, C., Gowal, S., Stanforth, R., Li, B., Boning, D. S., and Hsieh, C. Towards stable and efficient training of verifiably robust neural networks. In *ICLR*. OpenReview.net, 2020.

Zhang, H., Wang, S., Xu, K., Li, L., Li, B., Jana, S., Hsieh, C., and Kolter, J. Z. General cutting planes for bound-propagation-based neural network verification. *ArXiv preprint*, abs/2208.05740, 2022.

A. Additional Theoretical Results

A.1. Optimal Box Bounds Allow for Exact Certification

Below we present a Lemma, showing that optimal BOX bounds are indeed sufficient for exact certification, as suggested in Section 3.

Lemma A.1. *Any $\mathcal{C}^0$ continuous classifier $\boldsymbol{f}$, computing the logit difference $y_i^\Delta := y_i - y_t, \forall i \neq t$, is robustly correct on $\mathcal{B}^\epsilon(\boldsymbol{x})$ if and only if* $\text{Box}^*(\boldsymbol{f}, \mathcal{B}^\epsilon(\boldsymbol{x})) \subseteq \mathbb{R}^{c-1}_{<0}$, i.e., $\bar{y}_i^{\Delta^*} < 0, \forall i \neq t$.

Proof. On the one hand, assume $y_i - y_{\text{true}} < 0$ for all i. Then for the i^{th} output dimension, the optimal bounding box is $\max y_i - y_{\text{true}}$. Since the classifier is continuous, $\boldsymbol{f}(\mathcal{B}(\boldsymbol{x}, \boldsymbol{\epsilon}))$ is a closed and bounded set. Therefore, by extreme value theorem, $\exists \eta \in \mathcal{B}(\boldsymbol{x}, \boldsymbol{\epsilon})$ such that $\eta = \arg\max y_i - y_{\text{true}}$, thus $\max y_i - y_{\text{true}} < 0$. Since this holds for every i, $\text{Box}^*(\boldsymbol{f}, \mathcal{B}(\boldsymbol{x}, \boldsymbol{\epsilon})) \subseteq \mathcal{R}^{K-1}_{<0}$.

On the other hand, assume $\text{Box}^*(\boldsymbol{f}, \mathcal{B}(\boldsymbol{x}, \boldsymbol{\epsilon})) \subseteq \mathcal{R}^{K-1}_{<0}$. Since $\boldsymbol{f}(\mathcal{B}(\boldsymbol{x}, \boldsymbol{\epsilon})) \subseteq \text{Box}^*(\boldsymbol{f}, \mathcal{B}(\boldsymbol{x}, \boldsymbol{\epsilon})) \subseteq \mathcal{R}^{K-1}_{<0}$, we get $y_i - y_{\text{true}} < 0$ for all i. □

A.2. Propagation Invariance Regularization Strength

Below we present a corollary, formalizing the intuitions on the regularization strength of propagation invariance we provided in Section 3.

Corollary A.2. *Assume all elements of $\boldsymbol{W}^{(1)}$, $\boldsymbol{W}^{(2)}$ and $\boldsymbol{W}^{(2)}\boldsymbol{W}^{(1)}$ are non-zero and $\boldsymbol{W}^{(2)}\boldsymbol{W}^{(1)}$ is propagation invariant. Then choosing the signs of one row and one column of $\boldsymbol{W}^{(2)}\boldsymbol{W}^{(1)}$ fixes all signs of $\boldsymbol{W}^{(2)}\boldsymbol{W}^{(1)}$.*

Proof. For notational reasons, we define $W := W^{(2)}W^{(1)}$. Without loss of generality, assume we know the signs of the first row and the first column, *i.e.*, $W_{1,\cdot}$ and $W_{\cdot,1}$. We prove via a construction of the signs of all elements. The construction is given by the following: whenever $\exists i, j$, such that we know the sign of $W_{i,j}$, $W_{i,j+1}$ and $W_{i+1,j}$, we fix the sign of $W_{i+1,j+1}$ to be positive if there are an odd number of positive elements among $W_{i,j}$, $W_{i,j+1}$ and $W_{i+1,j}$, otherwise negative.

By Theorem 3.3, propagation invariance requires us to fix the sign of the last element in the $W_{i:i+1,j:j+1}$ block in this way. We only need to prove that when this process terminates, we fix the signs of all elements. We show this via recursion.

When $i = 1$ and $j = 1$, we have known the signs of $W_{i,j}$, $W_{i,j+1}$ and $W_{i+1,j}$, thus the sign of $W_{i+1,j+1}$ is fixed. Continuing towards the right, we gradually fix the sign of $W_{2,j+1}$ for $j = 1, \ldots, d-1$. Continuing downwards, we gradually fix the sign of $W_{i+1,2}$ for $i = 1, \ldots, d-1$. Therefore, all signs of the elements of the second row and the second column are fixed. By recursion, we would finally fix all the rows and the columns, thus the whole matrix. □

A.3. Tightness at Initialization

We first investigate the (expected) tightness $\tau = \frac{\mathbb{E}_{\mathcal{D}_{\boldsymbol{\theta}}}(\underline{\boldsymbol{z}}^* - \overline{\boldsymbol{z}}^*)}{\mathbb{E}_{\mathcal{D}_{\boldsymbol{\theta}}}(\overline{\boldsymbol{z}}^\dagger - \underline{\boldsymbol{z}}^\dagger)}$ (independent of the dimension due to symmetry) at initialization, *i.e.*, w.r.t. a weight distribution $\mathcal{D}_{\boldsymbol{\theta}}$. Let us consider a two-layer DLN at initialization, *i.e.*, with i.i.d. weights following a zero-mean Gaussian distribution $\mathcal{N}(0, \sigma^2)$ with an arbitrary but fixed variance σ^2 (Glorot & Bengio, 2010; He et al., 2015), again deferring a proof to App. B.

Lemma A.3 (Initialization Tightness w.r.t. Width)**.** *Given a 2-layer DLN with weight matrices $\boldsymbol{W}^{(1)} \in \mathcal{R}^{d_1 \times d_0}$, $\boldsymbol{W}^{(2)} \in \mathcal{R}^{d_2 \times d_1}$ with i.i.d. entries from $\mathcal{N}(0, \sigma_1^2)$ and $\mathcal{N}(0, \sigma_2^2)$ (together denoted as $\boldsymbol{\theta}$), we obtain the expected tightness* $\tau(d_1) = \frac{\mathbb{E}_{\boldsymbol{\theta}}(\underline{\boldsymbol{z}}^* - \overline{\boldsymbol{z}}^*)}{\mathbb{E}_{\boldsymbol{\theta}}(\overline{\boldsymbol{z}}^\dagger - \underline{\boldsymbol{z}}^\dagger)} = \frac{\sqrt{\pi}\,\Gamma(\frac{1}{2}(d_1+1))}{d_1\Gamma(\frac{1}{2}d_1)} \in \Theta(\frac{1}{\sqrt{d_1}})$.

Even for just two linear layers, the tightness at initialization decreases quickly with internal width ($\Theta(\frac{1}{\sqrt{d_1}})$), e.g., by a factor of $\tau(500) \approx 0.056$ for the penultimate layer of the popular `CNN7` (Gowal et al., 2018; Zhang et al., 2020). It, further, follows directly that tightness will decrease exponentially w.r.t. network depth.

Corollary A.4 (Initialization Tightness w.r.t. Depth)**.** *The expected tightness of an L-layer DLN $\boldsymbol{f}$ with minimum internal dimension d_{min} is at most $\tau \leq \tau(d_{min})^{\lfloor \frac{L}{2} \rfloor}$ at initialization.*

Note that this result is independent of the variance σ_1^2, σ_2^2, chosen for weight initialization. Thus, tightness at initialization can not be increased by scaling σ^2, as proposed by Shi et al. (2021b) to achieve constant box radius over network depth.

A.4. IBP Training Increases Tightness

As we have seen that networks are initialized with low tightness, we now investigate the effect of IBP training and show that it indeed increases tightness. To this end, we again consider a DLN with layer-wise propagation matrix $\boldsymbol{W}^\dagger = \Pi_{i=1}^L |\boldsymbol{W}^{(i)}|$ and optimal propagation matrix $\boldsymbol{W}^* = |\Pi_{i=1}^L \boldsymbol{W}^{(i)}|$, obtaining the expected risk for IBP training as $R(\boldsymbol{\epsilon}) = \mathbb{E}_{\boldsymbol{x},y}\mathcal{L}(\text{Box}^\dagger(\boldsymbol{f}, \mathcal{B}^{\boldsymbol{\epsilon}}(\boldsymbol{x})), y)$, again deferring a proof to App. B.

Theorem A.5 (IBP Training Increases Tightness)**.** *Assume homogenous tightness,* i.e., $\boldsymbol{W}^* = \tau \boldsymbol{W}^\dagger$, *and* $\frac{\|\nabla_\theta \boldsymbol{W}^*_{ij}\|_2}{\boldsymbol{W}^*_{ij}} \leq \frac{1}{2}\frac{\|\nabla_\theta \boldsymbol{W}^\dagger_{ij}\|_2}{\boldsymbol{W}^\dagger_{ij}}$ *for all i, j, then, the gradient difference between the IBP and standard loss is aligned with an increase in tightness,* i.e., $\langle \nabla_\theta(R(\boldsymbol{\epsilon}) - R(0)), \nabla_\theta \tau \rangle \leq 0$ *for all $\boldsymbol{\epsilon} > 0$.*

B. Deferred Proofs

Proof of Theorem 3.1 We first prove Theorem 3.1 for a 2-layer DLN as Lemma B.1.

Lemma B.1. *For a two-layer DLN* $\boldsymbol{f} = \boldsymbol{W}^{(2)}\boldsymbol{W}^{(1)}$, $(\overline{\boldsymbol{z}}^* - \underline{\boldsymbol{z}}^*)/2 = \left|W^{(2)}W^{(1)}\right|\boldsymbol{\epsilon}$ *and* $(\overline{\boldsymbol{z}}^\dagger - \underline{\boldsymbol{z}}^\dagger)/2 = \left|W^{(2)}\right|\left|W^{(1)}\right|\boldsymbol{\epsilon}$. *In addition,* Box* *and* Box† *have the same center* $\boldsymbol{f}(\boldsymbol{x})$.

Proof. First, assume $W^{(1)} \in \mathcal{R}^{d_1 \times d_0}$, $W^{(2)} \in \mathcal{R}^{d_2 \times d_1}$ and $B_i = [-1, 1]^{d_i}$ for $i = 0, 1, 2$, where $d_i \in \mathcal{Z}_+$ are some positive integers. The input box can be represented as $\mathrm{diag}(\epsilon_0)B_0 + b$ for $\epsilon_0 = \epsilon$.

For a single linear layer, the box propagation yields

$$\begin{aligned}
&\mathrm{Box}(W^{(1)}(\mathrm{diag}(\epsilon_0)B_0 + b)) \\
&= \mathrm{Box}(W^{(1)}\,\mathrm{diag}(\epsilon_0)B_0) + W^{(1)}b \\
&= \mathrm{diag}\left(\sum_{j=1}^{d_0} |W^{(1)}_{i,j}|\epsilon_0[j]\right) B_1 + W^{(1)}b \\
&:= \mathrm{diag}(\epsilon_1)B_1 + W^{(1)}b. \qquad (2)
\end{aligned}$$

Applying Equation (2) iteratively, we get the explicit formula of layer-wise propagation for two-layer linear network:

$$\begin{aligned}
&\mathrm{Box}(W^{(2)}\,\mathrm{Box}(W^{(1)}(\mathrm{diag}(\epsilon_0)B_0 + b))) \\
&= \mathrm{Box}\left(W^{(2)}(\mathrm{diag}(\epsilon_1)B_1 + W^{(1)}b)\right) \\
&= \mathrm{diag}\left(\sum_{k=1}^{d_1} |W^{(2)}_{i,k}|\epsilon_1[k]\right) B_2 + W^{(2)}W^{(1)}b \\
&= \mathrm{diag}\left(\sum_{j=1}^{d_0} \epsilon_0[j]\left(\sum_{k=1}^{d_1} |W^{(2)}_{i,k}W^{(1)}_{k,j}|\right)\right) B_2 \\
&\quad + W^{(2)}W^{(1)}b. \qquad (3)
\end{aligned}$$

Applying Equation (2) on $W := W^{(2)}W^{(1)}$, we get the explicit formula of the tightest box:

$$\begin{aligned}
&\mathrm{Box}(W^{(2)}W^{(1)}(\mathrm{diag}(\epsilon_0)B_0 + b)) \\
&= \mathrm{diag}\left(\sum_{j=1}^{d_0} |(W^{(2)}W^{(1)})_{i,j}|\epsilon_0[j]\right) B_2 + W^{(2)}W^{(1)}b \\
&= \mathrm{diag}\left(\sum_{j=1}^{d_0} \epsilon_0[j]\left|\sum_{k=1}^{d_1} W^{(2)}_{i,k}W^{(1)}_{k,j}\right|\right) B_2 \\
&\quad + W^{(2)}W^{(1)}b. \qquad (4)
\end{aligned}$$

□

Now, we use induction and Lemma B.1 to prove Theorem 3.1, restated below for convenience. The key insight is that a multi-layer DLN is equivalent to a single-layer linear network. Thus, we can group layers together and view general DLNs as two-layer DLNs.

Theorem 3.1. *For an L-layer DLN* $\boldsymbol{f} = \Pi_{k=1}^{L}\boldsymbol{W}^{(k)}$, *we obtain the box centres* $\dot{\boldsymbol{z}}^* = \dot{\boldsymbol{z}}^\dagger = \boldsymbol{f}(\boldsymbol{x})$ *and the radii* $\frac{\overline{\boldsymbol{z}}^* - \underline{\boldsymbol{z}}^*}{2} = \left|\Pi_{k=1}^{L}\boldsymbol{W}^{(k)}\right|\boldsymbol{\epsilon}$, *and* $\frac{\overline{\boldsymbol{z}}^\dagger - \underline{\boldsymbol{z}}^\dagger}{2} = \left(\Pi_{k=1}^{L}\left|\boldsymbol{W}^{(k)}\right|\right)\boldsymbol{\epsilon}$.

Proof. For $L = 2$, by Lemma B.1, the result holds. Assume for $L \leq m$, the result holds. Therefore, for $L = m + 1$, we group the first m layers as a single layer, resulting in a "two" layer equivalent network. Thus, $(\overline{\boldsymbol{z}}^* - \underline{\boldsymbol{z}}^*)/2 = \left|\boldsymbol{W}^{(m+1)}\Pi_{k=1}^{m}\boldsymbol{W}^{(k)}\right|\boldsymbol{\epsilon} = \left|\Pi_{k=1}^{L}\boldsymbol{W}^{(k)}\right|\boldsymbol{\epsilon}$. Similarly, by Equation (2), we can prove $(\overline{\boldsymbol{z}}^* - \underline{\boldsymbol{z}}^*)/2 = \left(\left|\boldsymbol{W}^{(m+1)}\right|\Pi_{k=1}^{m}\left|\boldsymbol{W}^{(k)}\right|\right)\boldsymbol{\epsilon} = \left(\Pi_{k=1}^{L}\left|\boldsymbol{W}^{(k)}\right|\right)\boldsymbol{\epsilon}$. The claim about center follows by induction similarly. □

Proof of Lemma 3.2 Here, we prove Lemma 3.2, restated below for convenience.

Lemma 3.2. *A two-layer DLN* $\boldsymbol{f} = \boldsymbol{W}^{(2)}\boldsymbol{W}^{(1)}$ *is propagation invariant if and only if for every fixed* (i, j), *we have* $\left|\sum_k W^{(2)}_{i,k} \cdot W^{(1)}_{k,j}\right| = \sum_k |W^{(2)}_{i,k} \cdot W^{(1)}_{k,j}|$, i.e., *either* $W^{(2)}_{i,k} \cdot W^{(1)}_{k,j} \geq 0$ *for all* k *or* $W^{(2)}_{i,k} \cdot W^{(1)}_{k,j} \leq 0$ *for all* k.

Proof. We prove the statement via comparing the box bounds. By Lemma B.1, we need $\left|\sum_{k=1}^{d_1} W^{(2)}_{i,k}W^{(1)}_{k,j}\right| = \sum_{k=1}^{d_1} |W^{(2)}_{i,k}W^{(1)}_{k,j}|$. The triangle inequality of absolute function says this holds if and only if $W^{(2)}_{i,k}W^{(1)}_{k,j} \geq 0$ for all k or $W^{(2)}_{i,k}W^{(1)}_{k,j} \leq 0$ for all k. □

Proof of Theorem 3.3 Here, we prove Theorem 3.3, restated below for convenience.

Theorem 3.3. *Assume* $\exists i, i', j, j'$, *such that* $W^{(1)}_{\cdot,j}$, $W^{(1)}_{\cdot,j'}$, $W^{(2)}_{i,\cdot}$ *and* $W^{(2)}_{i',\cdot}$ *are all non-zero. If* $(\boldsymbol{W}^{(2)}\boldsymbol{W}^{(1)})_{i,j} \cdot (\boldsymbol{W}^{(2)}\boldsymbol{W}^{(1)})_{i,j'} \cdot (\boldsymbol{W}^{(2)}\boldsymbol{W}^{(1)})_{i',j} \cdot (\boldsymbol{W}^{(2)}\boldsymbol{W}^{(1)})_{i',j'} < 0$, *then* $\boldsymbol{f} = \boldsymbol{W}^{(2)}\boldsymbol{W}^{(1)}$ *is not propagation invariant.*

Proof. The assumption $(W^{(2)}W^{(1)})_{i,j} \cdot (W^{(2)}W^{(1)})_{i,j'} \cdot (W^{(2)}W^{(1)})_{i',j} \cdot (W^{(2)}W^{(1)})_{i',j'} < 0$ implies three elements are of the same sign while the other element has a different sign. Without loss of generality, assume $(W^{(2)}W^{(1)})_{i',j'} < 0$ and the rest three are all positive.

Assume $W^{(2)}W^{(1)}$ is propagation invariant. By Lemma 3.2, this means $W^{(2)}_{i,\cdot}$.sign $= W^{(1)}_{\cdot,j}$.sign, $W^{(2)}_{i,\cdot}$.sign $= W^{(1)}_{\cdot,j'}$.sign, $W^{(2)}_{i',\cdot}$.sign $= W^{(1)}_{\cdot,j}$.sign and $W^{(2)}_{i',\cdot}$.sign $= -W^{(1)}_{\cdot,j'}$.sign. Therefore, we have $-W^{(1)}_{\cdot,j'}$.sign $= W^{(1)}_{\cdot,j'}$.sign, which implies all elements of $W^{(1)}_{\cdot,j'}$ must be zero. However, this results in $(W^{(2)}W^{(1)})_{i,j'} = 0$, a contradiction. □

Proof of Lemma A.3 Here, we prove Lemma A.3, restated below for convenience.

Lemma A.3 (Initialization Tightness w.r.t. Width)**.** *Given a 2-layer DLN with weight matrices $\boldsymbol{W}^{(1)} \in \mathcal{R}^{d_1 \times d_0}$, $\boldsymbol{W}^{(2)} \in \mathcal{R}^{d_2 \times d_1}$ with i.i.d. entries from $\mathcal{N}(0, \sigma_1^2)$ and $\mathcal{N}(0, \sigma_2^2)$ (together denoted as $\boldsymbol{\theta}$), we obtain the expected tightness $\tau(d_1) = \frac{\mathbb{E}_{\boldsymbol{\theta}}(\underline{\boldsymbol{z}}^* - \overline{\boldsymbol{z}}^*)}{\mathbb{E}_{\boldsymbol{\theta}}(\overline{\boldsymbol{z}}^\dagger - \underline{\boldsymbol{z}}^\dagger)} = \frac{\sqrt{\pi}\,\Gamma(\frac{1}{2}(d_1+1))}{d_1 \Gamma(\frac{1}{2} d_1)} \in \Theta(\frac{1}{\sqrt{d_1}})$.*

Proof. We first compute the size of the layer-wisely propagated box. From Equation (3), we get that for the i-th dimension,

$$\begin{aligned}\mathbb{E}(u_i - l_i) &= \mathbb{E}\left(\sum_{j=1}^{d_0} \epsilon_0[j] \left(\sum_{k=1}^{d_1} |W_{i,k}^{(2)} W_{k,j}^{(1)}|\right)\right) \\ &= \sum_{j=1}^{d_0} \epsilon_0[j] \left(\sum_{k=1}^{d_1} \mathbb{E}(|W_{i,k}^{(2)}|) \cdot \mathbb{E}(|W_{k,j}^{(1)}|)\right) \\ &= \sigma_1\sigma_2 \sum_{j=1}^{d_0} \epsilon_0[j] \left(\sum_{k=1}^{d_1} \mathbb{E}(|\mathcal{N}(0,1)|)^2)\right).\end{aligned}$$

Since $\mathbb{E}(|\mathcal{N}(0,1)|) = \sqrt{\frac{2}{\pi}}$ [1], we have

$$\mathbb{E}(u_i - l_i) = \frac{2}{\pi}\sigma_1\sigma_2 d_1 \|\epsilon_0\|_1. \tag{5}$$

Now we compute the size of the tightest box. From Equation (4), we get that for the i-th dimension,

$$\begin{aligned}\mathbb{E}(u_i^* - l_i^*) &= \mathbb{E}\left(\sum_{j=1}^{d_0} \epsilon_0[j] \left|\sum_{k=1}^{d_1} W_{i,k}^{(2)} W_{k,j}^{(1)}\right|\right) \\ &= \sigma_1\sigma_2 \sum_{j=1}^{d_0} \epsilon_0[j] \mathbb{E}\left(\left|\sum_{k=1}^{d_1} X_k Y_k\right|\right),\end{aligned}$$

where X_k and Y_k are i.i.d. standard Gaussian random variables. Using the law of total expectation, we have

$$\begin{aligned}\mathbb{E}\left(\left|\sum_{k=1}^{d_1} X_k Y_k\right|\right) &= \mathbb{E}\left(\mathbb{E}\left(\left|\sum_{k=1}^{d_1} X_k Y_k\right| \,\middle|\, Y_k\right)\right) \\ &= \mathbb{E}\left(\mathbb{E}\left(\left|\mathcal{N}(0, \sum_{k=1}^{d_1} Y_k^2)\right| \,\middle|\, Y_k\right)\right) \\ &= \sqrt{\frac{2}{\pi}}\mathbb{E}\left(\sqrt{\sum_{k=1}^{d_1} Y_k^2}\right) \\ &= \sqrt{\frac{2}{\pi}}\mathbb{E}(\sqrt{\chi^2(d_1)}).\end{aligned}$$

[1] https://en.wikipedia.org/wiki/Half-normal_distribution

Since $\mathbb{E}(\sqrt{\chi^2(d_1)}) = \sqrt{2}\Gamma(\frac{1}{2}(d_1+1))/\Gamma(\frac{1}{2}d_1)$, [2] we have

$$\mathbb{E}(u_i^* - l_i^*) = \frac{2}{\sqrt{\pi}}\sigma_1\sigma_2\|\epsilon_0\|_1\Gamma(\frac{1}{2}(d_1+1))/\Gamma(\frac{1}{2}d_1). \tag{6}$$

Combining Equation (5) and Equation (6), we have:

$$\frac{\mathbb{E}(u_i - l_i)}{\mathbb{E}(u_i^* - l_i^*)} = \frac{d_1\Gamma(\frac{1}{2}d_1)}{\sqrt{\pi}\Gamma(\frac{1}{2}(d_1+1))}. \tag{7}$$

To see the asymptotic behavior, use $\Gamma(x+\alpha)/\Gamma(x) \sim x^\alpha$, [3] we have

$$\frac{\mathbb{E}(u_i - l_i)}{\mathbb{E}(u_i^* - l_i^*)} \sim \frac{1}{\sqrt{\pi}} d_1^{\frac{1}{2}}. \tag{8}$$

To establish the bounds on the minimum expected slackness, we use Lemma B.2. □

Lemma B.2. *Let $g(n) := \frac{n\Gamma(\frac{1}{2}n)}{\sqrt{\pi}\Gamma(\frac{1}{2}(n+1))}$. $g(n)$ is monotonically increasing for $n \geq 1$. Thus, for $n \geq 2$, $g(n) \geq g(2) > 1.27$.*

Proof. Using polygamma function $\psi^{(0)}(z) = \Gamma'(z)/\Gamma(z)$, [4] we have

$$g'(n) \propto 1 + \frac{1}{2}n\left(\psi^{(0)}\left(\frac{1}{2}n\right) - \psi^{(0)}\left(\frac{1}{2}(n+1)\right)\right).$$

Using the fact that $\psi^{(0)}(z)$ is monotonically increasing for $z > 0$ and $\psi^{(0)}(z+1) = \psi^{(0)}(z) + \frac{1}{z}$, we have

$$\begin{aligned}&1 + \frac{1}{2}n\left(\psi^{(0)}\left(\frac{1}{2}n\right) - \psi^{(0)}\left(\frac{1}{2}(n+1)\right)\right) \\ &> 1 + \frac{1}{2}n\left(\psi^{(0)}\left(\frac{1}{2}n\right) - \psi^{(0)}\left(\frac{1}{2}n+1\right)\right) \\ &= 1 + \frac{1}{2}n\left(-\frac{2}{n}\right) \\ &= 0.\end{aligned}$$

Therefore, $g'(n)$ is strictly positive for $n \geq 1$, and thus $g(n)$ is monotonically increasing for $n \geq 1$. □

As a final comment, we visualize $g(n)$ in Figure 6. As expected, $g(n)$ is monotonically increasing in the order of $O(\sqrt{n})$.

Proof of Corollary A.4 Here, we prove Corollary A.4, restated below for convenience.

Corollary A.4 (Initialization Tightness w.r.t. Depth)**.** *The expected tightness of an L-layer DLN $\boldsymbol{f}$ with minimum internal dimension d_{min} is at most $\tau \leq \tau(d_{min})^{\lfloor \frac{L}{2} \rfloor}$ at initialization.*

[2] https://en.wikipedia.org/wiki/Chi_distribution

[3] https://en.wikipedia.org/wiki/Gamma_function#Stirling's_formula

[4] https://en.wikipedia.org/wiki/Polygamma_function

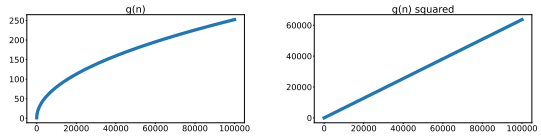

Figure 6: $g(n)$ and $g^2(n)$ visualized.

Proof. This is pretty straightforward and only requires a coarse application of Lemma A.3. Without loss of generality, we assume L is even. If L is odd, then we simply discard the slackness introduced by the last layer, *i.e.*, assume the last layer does not introduce additional slackness.

We group the $2i-1$-th and $2i$-th layer as a new layer. By Lemma A.3, these $L/2$ subnetworks all introduce an additional slackness factor of τ. Note that Equation (5) implies that the size of the output box is proportional to the size of the input box. Therefore, the layer-wisely propagated box of $\Pi_{i=1}^{L} W_i$ is $\tau^{L/2}$ looser than the layer-wisely propagated box of $\Pi_{j=1}^{L/2}(W_{2j-1}W_{2j})$. In addition, the size of the tightest box for $\Pi_{i=1}^{L} W_i$ is upper bounded by layer-wisely propagating $\Pi_{j=1}^{L/2}(W_{2j-1}W_{2j})$. Therefore, the minimum expected slackness is lower bounded by $\tau^{L/2}$. □

Proof of Theorem A.5 Here, we prove Theorem A.5, restated below for convenience.

Theorem A.5 (IBP Training Increases Tightness)**.** *Assume homogenous tightness,* i.e., $\boldsymbol{W}^* = \tau \boldsymbol{W}^\dagger$, *and* $\frac{\|\nabla_\theta \boldsymbol{W}^*_{ij}\|_2}{\boldsymbol{W}^*_{ij}} \leq \frac{1}{2}\frac{\|\nabla_\theta \boldsymbol{W}^\dagger_{ij}\|_2}{\boldsymbol{W}^\dagger_{ij}}$ *for all* i, j*, then, the gradient difference between the IBP and standard loss is aligned with an increase in tightness,* i.e., $\langle \nabla_\theta (R(\boldsymbol{\epsilon}) - R(0)), \nabla_\theta \tau \rangle \leq 0$ *for all* $\boldsymbol{\epsilon} > 0$.

Proof. We prove a stronger claim: $\langle \nabla_\theta (R(\boldsymbol{\epsilon} + \Delta\boldsymbol{\epsilon}) - R(\boldsymbol{\epsilon})), \nabla_\theta \tau \rangle \leq 0$ for all $\boldsymbol{\epsilon} \geq 0$ and $\Delta\boldsymbol{\epsilon} > 0$. Let $\boldsymbol{\epsilon} = \mathbf{0}$ yields the theorem.

We prove the claim for $\Delta\boldsymbol{\epsilon} \to 0$. For large $\Delta\boldsymbol{\epsilon}$, we can break it into $R(\boldsymbol{\epsilon} + \Delta\boldsymbol{\epsilon}) - R(\boldsymbol{\epsilon}) = \sum_{i=1}^{n} R(\boldsymbol{\epsilon} + \frac{i}{n}\Delta\boldsymbol{\epsilon}) - R(\boldsymbol{\epsilon} + \frac{i-1}{n}\Delta\boldsymbol{\epsilon})$, thus proving the claim since each summand satisfies the theorem.

Let $L_1 = R(\boldsymbol{\epsilon})$ and $L_2 = R(\boldsymbol{\epsilon} + \Delta\boldsymbol{\epsilon})$. By Taylor expansion, we have $L_2 = L_1 + \Delta\boldsymbol{\epsilon}^\top \boldsymbol{W}^\dagger \nabla_{\boldsymbol{u}} g = L_1 + \frac{1}{\tau}\Delta\boldsymbol{\epsilon}^\top \boldsymbol{W}^* \nabla_{\boldsymbol{u}} g$, where $\nabla_{\boldsymbol{u}} g = \nabla_{\boldsymbol{u}} g(\boldsymbol{u})$ evaluated at $\boldsymbol{u} = \boldsymbol{W}^\dagger \boldsymbol{\epsilon}$. Note that the increase of $\boldsymbol{\epsilon}$ would increase the risk, thus $\nabla_{\boldsymbol{u}} g \geq \mathbf{0}$.

For the i^{th} parameter θ_i, $\nabla_{\theta_i}(L_2 - L_1)\nabla_{\theta_i}\tau = \frac{1}{\tau^2}\Delta\boldsymbol{\epsilon}^\top(\tau\nabla_{\theta_i}\boldsymbol{W}^* - \boldsymbol{W}^*\nabla_{\theta_i}\tau)\nabla_{\boldsymbol{u}} g \nabla_{\theta_i}\tau$. Thus, $\langle \nabla_\theta(L_2 - L_1), \nabla_\theta \tau \rangle = \frac{1}{\tau^2}\Delta\boldsymbol{\epsilon}^\top(\tau \sum_i \nabla_{\theta_i}\tau \cdot \nabla_{\theta_i}\boldsymbol{W}^* - \boldsymbol{W}^* \|\nabla_\theta \tau\|_2^2)\nabla_{\boldsymbol{u}} g$. Since $\Delta\boldsymbol{\epsilon} > \mathbf{0}$ and $\nabla_{\boldsymbol{u}} g \geq \mathbf{0}$, it sufficies to prove that $\tau \sum_i \nabla_{\theta_i}\tau \cdot \nabla_{\theta_i}\boldsymbol{W}^* - \boldsymbol{W}^* \|\nabla_\theta \tau\|_2^2$ is nonpositive, *i.e.*, $\tau\langle \nabla_\theta \tau, \nabla_\theta \boldsymbol{W}^*_{ij}\rangle - \boldsymbol{W}^*_{ij}\|\nabla_\theta \tau\|_2^2$ is nonpositive for every i, j.

Since $\|\boldsymbol{u}\|_2\|\boldsymbol{v}\|_2 \geq \langle \boldsymbol{u}, \boldsymbol{v}\rangle$, we have

$$\frac{\|\nabla_\theta \boldsymbol{W}^*_{ij}\|_2}{\boldsymbol{W}^*_{ij}} \leq \frac{1}{2}\frac{\|\nabla_\theta \boldsymbol{W}^\dagger_{ij}\|_2}{\boldsymbol{W}^\dagger_{ij}}$$
$$\Rightarrow \|\nabla_\theta \log \boldsymbol{W}^\dagger\|_2 \geq 2\|\nabla_\theta \log \boldsymbol{W}^*\|_2$$
$$\Rightarrow \|\nabla_\theta \log \boldsymbol{W}^\dagger\|_2^2 \geq 2\langle \nabla_\theta \log \boldsymbol{W}^\dagger, \nabla_\theta \log \boldsymbol{W}^*\rangle$$

Therfore, $\|\nabla_\theta \log \tau\|_2^2 = \|\nabla_\theta(\log \boldsymbol{W}^*_{ij} - \log \boldsymbol{W}^\dagger_{ij})\|_2^2 = \|\nabla_\theta \log \boldsymbol{W}^*_{ij}\|_2^2 - 2\langle \nabla_\theta \log \boldsymbol{W}^\dagger, \nabla_\theta \log \boldsymbol{W}^*\rangle + \|\nabla_\theta \log \boldsymbol{W}^\dagger\|_2^2 \geq \|\nabla_\theta \log \boldsymbol{W}^*_{ij}\|_2^2$. This means $\frac{\|\nabla_\theta \boldsymbol{W}^*_{ij}\|_2}{\boldsymbol{W}^*_{ij}} \leq \frac{\|\nabla_\theta \tau\|_2}{\tau}$, thus $\boldsymbol{W}^*_{ij}\|\nabla_\theta \tau\|_2^2 \geq \tau\|\nabla_\theta \tau\|_2 \|\nabla_\theta \boldsymbol{W}^*_{ij}\|_2 \geq \tau\langle \nabla_\theta \tau, \nabla_\theta \boldsymbol{W}^*_{ij}\rangle$, which fulfills our goal. □

Proof of Theorem 3.6 Here, we prove Theorem 3.6, restated below for convenience.

Theorem 3.6. *Consider the linear embedding and reconstruction* $\hat{\boldsymbol{x}} = U_k U_k^\top \boldsymbol{x}$ *of a* d *dimensional data distribution* $\boldsymbol{x} \sim \mathcal{X}$ *into a* k *dimensional space with* $d \gg k$ *and eigenmatrices* U *drawn uniformly at random from the orthogonal group. Propagating the input box* $\mathcal{B}^{\boldsymbol{\epsilon}}(\boldsymbol{x})$ *layer-wise and optimally, thus, yields* $\mathcal{B}^{\boldsymbol{\delta}^\dagger}(\hat{\boldsymbol{x}})$*, and* $\mathcal{B}^{\boldsymbol{\delta}^*}(\hat{\boldsymbol{x}})$*, respectively. Then, we have, (i)* $\mathbb{E}(\delta_i/\epsilon) = ck \in \Theta(k)$ *for a positive constant* c *depending solely on* d *and* $c \to \frac{2}{\pi} \approx 0.64$ *for large* d*; and (ii)* $\mathbb{E}(\delta^*_i/\epsilon) \to \frac{2}{\sqrt{\pi}}\frac{\Gamma(\frac{1}{2}(k+1))}{\Gamma(\frac{1}{2}k)} \in \Theta(\sqrt{k})$.

Proof. Since box propagation for linear functions maps the center of the input box to the center of the output box, the center of the output box is exactly $\hat{X}$. By Lemma B.1, we have $\boldsymbol{\delta} = |U_k||U_k|^\top \epsilon \mathbf{1}$. For notational simplicity, let $V = |U_k|$, thus

$$\begin{aligned}\boldsymbol{\delta}_i &= \sum_{j=1}^{k} V_{ij}\Big(\sum_{p=1}^{d} V_{jp}^\top \epsilon\Big) \\ &= \epsilon \sum_{p=1}^{d}\sum_{j=1}^{k} V_{ij} V_{pj} \\ &= \epsilon \sum_{j=1}^{k} V_{ij}\|V_{:j}\|_1.\end{aligned}$$

Therefore, $\mathbb{E}\boldsymbol{\delta}_i/\epsilon = \sum_{j=1}^{k} \mathbb{E}(V_{ij}\|V_{:j}\|_1) = ck$, where $c = \mathbb{E}(V_{ij}\|V_{:j}\|_1)$. Since $V_{:j}$ is the absolute value of a column of the orthogonal matrix uniformly drawn, $V_{:j}$ itself is the absolute value of a vector drawn uniformly from the unit hyper-ball. By Cook (1957) and Marsaglia (1972), $V_{:j}$ is equivalent in distribution to *i.i.d.* draw samples from the standard Gaussian for each dimension and then normalize it by its L_2 norm. For notational simplicity, let $V_{:j} \overset{d}{=} v = |u|$, where $u = \hat{u}/\|\hat{u}\|_2$ and all dimensions of $\hat{u}$ are *i.i.d.*

drawn from the standard Gaussian distribution, thus $c = \mathbb{E}(v_1\|v\|_1)$.

Expanding $\|v\|_1$, we have $c = \mathbb{E}(v_1^2) + \sum_{i=2}^{d}\mathbb{E}(v_1 v_i) = \frac{1}{d}\mathbb{E}(\|v\|_2^2) + (d-1)\mathbb{E}(v_1 v_2) = \frac{1}{d} + (d-1)\mathbb{E}(v_1 v_2)$. From page 20 of Pinelis & Molzon (2016), we know each entry in u converges to $\mathcal{N}(0, 1/d)$ at $O(1/d)$ speed in Kolmogorov distance. In addition, $\mathbb{E}(v_1 v_2) = \mathbb{E}(\mathbb{E}(v_2 \mid v_1) \cdot v_1) = \mathbb{E}(v_1\sqrt{1-v_1^2})\mathbb{E}(v_2')$, where v' is the absolute value of a random vector uniformly drawn from the $d-1$ dimensional sphere. Therefore, for large d, $c = (d-1)\mathbb{E}(v_1\sqrt{1-v_1^2})\mathbb{E}(v_2') = (d-1)\mathbb{E}(v_1)\mathbb{E}(v_2') = (d-1)\mathbb{E}(|\mathcal{N}(0,1/d)|)\mathbb{E}(|\mathcal{N}(0,1/(d-1))|) = \frac{2}{\pi}$.

To show how good the asymptotic result is, we run Monte-Carlo to get the estimation of c. As shown in the left of Figure 7, the Monte-Carlo result is consistent to this theorem. In addition, it converges very quickly, *e.g.*, stablizing at 0.64 when $d \geq 100$.

Now we start proving (2). By Lemma B.1, we have $\delta^* = |U_k U_k^\top|\epsilon\mathbf{1}$. Thus,

$$\begin{aligned}\mathbb{E}(\boldsymbol{\delta}_i^*/\epsilon) &= \sum_{j=1}^{d}\mathbb{E}\left|\sum_{p=1}^{k} U_{ip}U_{jp}\right| \\ &= \sum_{j\neq i}\mathbb{E}\left|\sum_{p=1}^{k} U_{ip}U_{jp}\right| + \mathbb{E}(\sum_{p=1}^{k} U_{ip}^2) \\ &= (d-1)\mathbb{E}\left|\sum_{p=1}^{k} U_{ip}U_{jp}\right| + \frac{k}{d}.\end{aligned}$$

In addition, we have

$$\begin{aligned}&(d-1)\mathbb{E}\left|\sum_{p=1}^{k} U_{ip}U_{jp}\right| \\ &= (d-1)\mathbb{E}_{U_i}\left(\mathbb{E}_{U_j}\left(\left|\sum_{p=1}^{k} U_{ip}U_{jp}\right| \middle| U_i\right)\right) \\ &\to (d-1)\mathbb{E}_{U_i}\left(\mathbb{E}\left|\mathcal{N}\left(0, \frac{\sum_{p=1}^{k} U_{ip}^2}{d-1}\right)\right|\right) \\ &= (d-1)\sqrt{\frac{2}{\pi(d-1)}}\mathbb{E}\sqrt{\sum_{p=1}^{k} U_{ip}^2} \\ &= \sqrt{\frac{2(d-1)}{\pi}}\mathbb{E}\sqrt{\frac{1}{d}\chi^2(k)} \\ &\to \frac{2}{\sqrt{\pi}}\frac{\Gamma(\frac{1}{2}(k+1))}{\Gamma(\frac{1}{2}k)},\end{aligned}$$

where we use again that for large d, the entries of a column tends to Gaussian. This proves (2). The expected tightness follows by definition, *i.e.*, dividing the result of (1) and (2). □

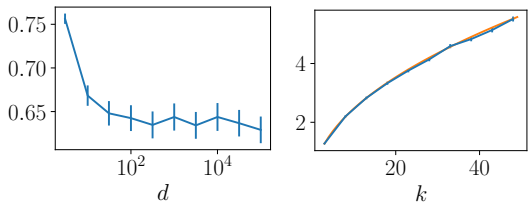

Figure 7: Monte-Carlo estimations of Theorem 3.6. Result bases on 10000 samples for each d. Left: c plotted against d in log scale. Right: $\mathbb{E}(\boldsymbol{\delta}_i^*)$ plotted against k for $d = 2000$ (blue), together with the theoretical predictions (orange).

The right of Figure 7 plots the Monte-Carlo estimations against our theoretical results. Clearly, this confirms our result.

C. Experimental Details

C.1. Dataset

We use the MNIST (LeCun et al., 2010) and CIFAR-10 (Krizhevsky et al., 2009) datasets for our experiments. Both are open-source and freely available. For MNIST, we do not apply any preprocessing or data augmentation. For CIFAR-10, we normalize images with their mean and standard deviation and, during training, first apply 2-pixel zero padding and then random cropping to 32×32.

C.2. Model Architecture

We follow previous works (Shi et al., 2021a; Müller et al., 2022b; Mao et al., 2023) and use a 7-layer convolutional network `CNN7` in most experiments. For the experiments investigating the effect of epsilon on tightness and certifiability depending on the training method, visualized in Figure 4, we use a smaller 3-layer convolutional network `CNN3`. Details about them can be found in the released code.

C.3. Training

Following previous works (Müller et al., 2022b; Mao et al., 2023), we use the initialization, warm-up regularization, and learning schedules introduced by Shi et al. (2021a). Specifically, for MNIST, the first 20 epochs are used for ϵ-scheduling, increasing ϵ smoothly from 0 to the target value. Then, we train an additional 50 epochs with two learning rate decays of 0.2 at epochs 50 and 60, respectively. For CIFAR-10, we use 80 epochs for ϵ-annealing, after training models with standard training for 1 epoch. We continue training for 80 further epochs with two learning rate decays of 0.2 at epochs 120 and 140, respectively. The initial learning rate is 5×10^{-3} and the gradients are clipped to an L_2 norm of at most 10.0 before every step.

C.4. Certification

We apply MN-BAB (Ferrari et al., 2022) to certify all models. MN-BAB is a state-of-the-art complete certification method built on multi-neuron relaxations. When certifying SABR-trained CNN7 to obtain state-of-the-art results, we use the same hyperparameters for MN-BAB as Müller et al. (2022b) and set the timeout to 1000 seconds. For other experiments, we use the same hyperparameters but reduce timeout to 200 seconds for efficiency reasons.

D. Extended Empirical Evaluation

D.1. STAPS-Training and Regularization Level

To confirm our observations on the interaction of regularization level, accuracies, and propagation tightness from Section 4, we extend our experiments to STAPS (Mao et al., 2023), an additional state-of-the-art certified training method beyond SABR (Müller et al., 2022b). Recall that STAPS combines SABR with adversarial training as follows. The model is first (conceptually) split into a feature extractor and classifier. Then, during training IBP is used to propagate the input region through the feature extractor yielding box bounds in the model's latent space. Then, adversarial training with PGD is conducted over the classifier using these box bounds as input region. As IBP leads to an over-approximation while PGD leads to an under-approximation, STAPS induces more regularization as fewer (ReLU) layers are included in the classifier.

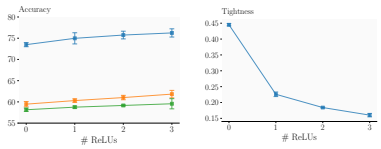

Figure 8: Accuracies and tightness of a CNN7 for CIFAR-10 $\epsilon = \frac{2}{255}$ depending on regularization strength with STAPS

We visualize the result of thus varying regularization levels by changing the number of ReLU layers in the classifier in Figure 8. We observe very similar trends as for SABR in Figure 5, although to a lesser extent, as 0 ReLU layers in the classifier still recovers SABR and not standard IBP. Again, decreasing regularization (increasing the number of ReLU layers in the classifier) leads to reducing tightness and increasing standard and certified accuracies.

E. Related Work

Certified Training Sound certified training methods compute and optimize an over-approximation of the worst-case loss obtained via bound propagation methods (Wong & Kolter, 2018; Wong et al., 2018; Zhang et al., 2020). A particularly efficient and scalable method is IBP (Gowal et al., 2018; Mirman et al., 2018), for which Shi et al. (2021b) propose a custom initialization scheme, significantly shortening training schedules, and Lin et al. (2022) propose a non-negativity regularization, marginally improving certified accuracies. More recent methods use unsound but more precise approximations. COLT (Balunovic & Vechev, 2020a) combines precise ZONOTOPE (Singh et al., 2018) bounds with adversarial training but is severely limited in scalability. IBP-R (Palma et al., 2022) combines an IBP-based regularization with adversarial training at larger perturbation magnitudes. SABR (Müller et al., 2022b) applies IBP to small but carefully selected regions in the adversary specification to reduce regularization. TAPS (Mao et al., 2023), similar to COLT, combines IBP with adversarial training. This recent dominance of IBP-based methods motivates our work to develop a deeper understanding of how IBP training affects network robustness.

Theoretical Analysis of IBP The capability of IBP has been studied theoretically in the past. Baader et al. (2020) first show that continuous functions can be approximated by IBP-certifiable ReLU networks up to arbitrary precision. Wang et al. (2022b) extend this result to more activation functions and prove that constructing such networks is strictly harder than NP-complete problems assuming coNP $\not\subseteq$ NP. Wang et al. (2022a) study the convergence of IBP-training and find that it converges to a global optimum with high probability for infinite width. Mirman et al. (2022) derive a negative result, showing that even optimal box bounds can fail on simple datasets. However, none of these works study the tightness of IBP bounds, *i.e.*, their relationship to optimal interval approximations. Motivated by recent certified training methods identifying this approximation precision as crucial (Müller et al., 2022a; Mao et al., 2022), we bridge this gap by deriving sufficient and necessary conditions for propagation invariance, introducing the relaxed measure of propagation tightness and studying how it interacts with network architecture and IBP training, both theoretically and empirically.

F. Reproducibility

We publish our code, all trained models, and detailed instructions on how to reproduce our results at ANONYMIZED. Further, we provide detailed descriptions of all hyperparameter choices, data sets, and preprocessing steps in App. C.