Improving Adversarial Robustness via Joint Classification and Multiple Explicit Detection Classes

Sina Baharlouei[*] baharlou@usc.edu

Fatemeh Sheikholeslami[†] f.sheikholeslami@gmail.com

Meisam Razaviyayn[*] razaviya@usc.edu

Zico Kolter[‡] zkolter@cs.cmu.edu

Abstract

This work concerns the development of deep networks that are certifiably robust to adversarial attacks. Joint robust classification-detection was recently introduced as a certified defense mechanism, where adversarial examples are either correctly classified or assigned to the "abstain" class. In this work, we show that such a provable framework can be extended to networks with multiple explicit abstain classes, where the adversarial examples are adaptively assigned to those. While naïvely adding multiple abstain classes can lead to "model degeneracy", we propose a regularization approach and a training method to counter this degeneracy by promoting full use of the multiple abstain classes. Our experiments demonstrate that the proposed approach consistently achieves favorable standard vs. robust verified accuracy tradeoffs, outperforming state-of-the-art algorithms for various choices of number of abstain classes.

1 Introduction

Deep Neural Networks (DNNs) have revolutionized many machine learning tasks such as image processing [Krizhevsky et al., 2012, Zhu et al., 2021] and speech recognition [Graves et al., 2013, Nassif et al., 2019]. However, despite their superior performance, DNNs are highly vulnerable to adversarial attacks and perform poorly on out-of-distributions samples [Goodfellow et al., 2014, Liang et al., 2017, Yuan et al., 2019].

To address the vulnerability of DNNs to adversarial attacks, the community have designed various defense mechanisms that are robust against adversarial attacks [Papernot et al., 2016, Jang et al., 2019, Goldblum et al., 2020, Madry et al., 2017, Huang et al., 2021]. These mechanisms provide robustness against certain types of attacks such as the Fast Gradient Sign Method (FGSM) [Szegedy et al., 2013, Goodfellow et al., 2014]. However, the overwhelming majority of these defense mechanisms are highly ineffective against more complex attacks such as adaptive and brute-force methods [Tramer et al., 2020, Carlini and Wagner, 2017]. This ineffectiveness necessitates: 1) the design of rigorous verification approaches that can measure the robustness of a given network; 2) the development of defense mechanisms that are verifiably robust against *any* attack strategy within the class of permissible attack strategies. To verify robustness of a given network against *any* attack in a reasonable set of permissible attacks (e.g. ℓ_p-norm ball around the given input data), one needs to solve a hard non-convex optimization problem (see, e.g., Problem (1) in this paper). Consequently, exact verifiers, such as [Tjeng et al., 2017, Xiao et al., 2018], are not scalable to large networks. To

[*]Industrial and Systems Engineering, University of Southern California

[†]Amazon Alexa AI

[‡]Computer Science Department, Carnegie Mellon University

ML Safety Workshop, 36th Conference on Neural Information Processing Systems (NeurIPS 2022).

develop scalable verifiers, the community turn to "inexact" verifiers. Such methods can only verify a subset of perturbations to the input data that the network can defend against successfully. This is typically achieved by finding tractable lower-bounds for the verification optimization problem. Gowal et al. [2018] finds such a lower-bound by *interval bound propagation (IBP)* which is essentially an efficient convex relaxation of the constraint sets in the verification problem. Despite its simplicity, this approach demonstrates a relatively superior performance compared to prior works. Another line of work for enhancing the performance of certifiably robust neural networks relies on the idea of learning a detector alongside the classifier to capture adversarial and out-of-distribution samples. Instead of trying to classify adversarial images correctly, these works design a *detector* to determine whether a given sample is natural/in-distribution or it is a crafted attack/out-of-distribution. A more resilient approach is to jointly learn the detector and the classifier [Laidlaw and Feizi, 2019, Sheikholeslami et al., 2021, Chen et al., 2021] by adding an auxiliary *abstain* output class capturing adversarial samples.

Building on these prior works, this paper extends the idea of using a single abstain class to using multiple abstain classes. We observe that naïvely adding multiple abstain classes results in a model degeneracy phenomenon where all adversarial examples are assigned to a small fraction of abstain classes (while other abstain classes are not utilized). To resolve this issue, we propose a regularizer that balances the assignment of adversarial examples to abstain classes. Our experiments demonstrate that utilizing multiple abstain classes in conjunction with the proper regularization enhances the robust verified accuracy of joint detectors/classifiers on adversarial examples while maintaining the standard accuracy of the classifier.

2 Background

Consider an L-layer feedforward neural network with $\mathbf{W}_i$ denoting the weight associated with layer i, and $\mathbf{b}_i$ denoting the bias parameter of layer i. Let $\sigma_i(\cdot)$ denote the activation function applied at layer i. Throughout the paper, we assume the activation function is the same for all hidden layers, i.e., $\sigma_i(\cdot) = \text{ReLU}(\cdot),\ \forall i = 1, \ldots, L-1$. Thus, our neural network can be described as

$$\begin{aligned} \mathbf{z}_i &= \sigma(\mathbf{W}_i \mathbf{z}_{i-1} + \mathbf{b}_i), \quad i = 1, 2, \ldots, L-1, \\ \mathbf{z}_L &= \mathbf{W}_L \mathbf{z}_{L-1} + \mathbf{b}_L, \end{aligned}$$

where $\mathbf{z}_0 = \mathbf{x}$ is the input to the neural network and $\mathbf{z}_i$ is the output of layer i. To explicitly show the dependence of $\mathbf{z}_L$ on the input data, we use the notation $\mathbf{z}_L(\mathbf{x})$ to denote logit values when $\mathbf{x}$ is used as the input data point.

Given an input $\mathbf{x}_0$ with the ground-truth label y, and a perturbation set $\mathcal{C}(\mathbf{x}_0, \epsilon)$ (e.g. $\mathcal{C}(\mathbf{x}_0, \epsilon) = \{\mathbf{x} \,|\, \|\mathbf{x} - \mathbf{x}_0\|_\infty \leq \epsilon\}$), the network is provably robust against adversarial attacks on $\mathbf{x}_0$ if

$$0 \leq \min_{\mathbf{x} \in \mathcal{C}(\mathbf{x}_0, \epsilon)} \mathbf{c}_{yk}^T \mathbf{z}_L(\mathbf{x}), \quad \forall k \neq y, \tag{1}$$

where $\mathbf{c}_{yk} = \mathbf{e}_y - \mathbf{e}_k$ with $\mathbf{e}_k$ (resp. $\mathbf{e}_y$) being the standard unit vector whose k-th row (resp. y-th row) is 1 and the other entries are zero. Condition (1) implies that the logit score of the network for the true label y is always greater than that of any other label k for all $\mathbf{x} \in \mathcal{C}(\mathbf{x}_0, \epsilon)$. Thus, the network will classify all the points inside $\mathcal{C}(\mathbf{x}_0, \epsilon)$ correctly. The objective function in Eq. (1) is non-convex when $L \geq 2$. It is customary in many works to move the non-convexity of the problem to the constraint set and reformulate Eq. (1) as

$$0 \leq \min_{\mathbf{z} \in \mathcal{Z}(\mathbf{x}_0, \epsilon)} \mathbf{c}_{yk}^T \mathbf{z}, \qquad \forall k \neq y, \tag{2}$$

where $\mathcal{Z}(\mathbf{x}_0, \epsilon) = \{\mathbf{z} \,|\, \mathbf{z} = \mathbf{z}_L(\mathbf{x}) \quad \text{for some } \mathbf{x} \in \mathcal{C}(\mathbf{x}_0, \epsilon)\}$. Since both problems (1) and (2) are non-convex, existing works proposed efficiently computable lower-bounds for the optimal objective value of them. The dominant approach for estimating Problem (2) is to convexify the constraint set by Interval Bound Propagation (IBP) [Gowal et al., 2018]. After this relaxation, problem (2) can be lower-bounded by the convex problem:

$$\min_{\underline{\mathbf{z}}(\mathbf{x}_0) \leq \mathbf{z} \leq \bar{\mathbf{z}}(\mathbf{x}_0)} \mathbf{c}_{yk}^T \mathbf{z} \tag{3}$$

The upper and lower bounds $\underline{\mathbf{z}}(\mathbf{x}_0)$ and $\bar{\mathbf{z}}(\mathbf{x}_0)$ are obtained by recursively finding the convex relaxation of the image of the set $\mathcal{C}(\mathbf{x}_0, \epsilon)$ at each layer of the network.

3 Verification of neural networks with multiple abstain classes

Motivation: The robust verified accuracy of an L-layer joint classifier and detector can be enhanced by introducing multiple abstain classes instead of a single abstain class for detecting adversarial examples. The set of all adversarial images that can be generated within the ϵ-neighborhood of clean images might not be a connected set that can be detected only by one detection class. This observation is illustrated in a simple example in Appendix K where 2 detection classes can drastically increase the performance of the detector compared to 1 detection class.

Note that a network with multiple detection classes can be equivalently modeled by another network with one more layer and a single abstain class. This added layer can merge all abstain classes and reduce them to a single class. *Thus, any L-layer neural network with multiple abstain classes can be equivalently modeled by an $L+1$-layer neural network with a single abstain class.* However, the performance of verifiers such as IBP reduces as we increase the number of layers. This is due to the fact that increasing the number of layers leads to looser bounds in Equation 3. As we can observe in Figure 1, the number of verified points by a $2-$layer neural network is higher than the number of points verified by an equivalent network with 3 layers. The description of both networks can be found in Appendix L.

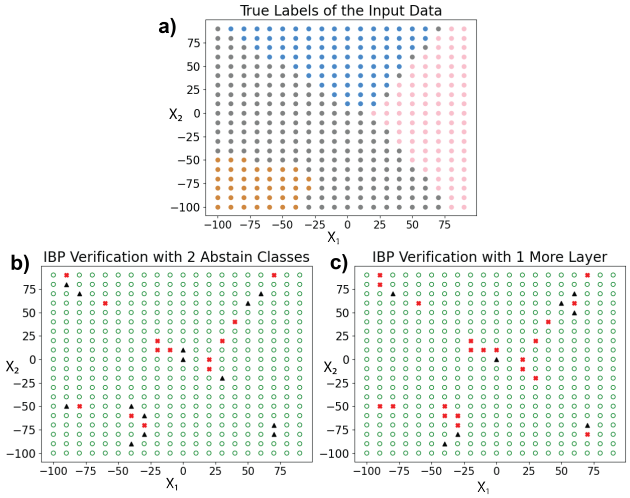

Figure 1: The IBP verification for 400 input data points of 2-layer and 3-layer neural networks. Part **(a)** shows the assigned label to each data point. Part **(b)** demonstrates that IBP can verify 14 points using one of two detection classes (black triangles), while it cannot verify 13 data points (red marks). **c)** On the other hand shows that when IBP applied to a network with one more layer and one detection class, 8 points are verified by the detection class, while it fails to verify 21 points. That means for this simple neural networks, the one with smaller number of layers can be verified more accurately with via IBP.

Thus, it is beneficial to train/verify the original L-layer neural network with multiple abstain classes instead of $L+1$-layer network with a single abstain class. This fact will be illustrated further in the experiments on MNIST and CIFAR-10 datasets depicted in Figure 2. Next, we present how one can verify a network with multiple abstain classes. Let $a_1, a_2, \ldots, a_M$ be M abstain classes detecting adversarial samples. A sample is considered adversarial if the output of the network is any of the M abstain classes. A neural network with K regular classes and M abstain classes outputs the label of a given sample as $\hat{y}(\mathbf{x}) = \text{argmax}_{i\in\{1,\ldots,K,a_1,\ldots,a_M\}}[\mathbf{z}_L(\mathbf{x})]_i$. An input is verified if the network either correctly classifies it or assigns it to any of the explicit M abstain classes. More formally and following equation (7), the neural network is verified for input $\mathbf{x}_0$ against a target class k if

$$0 \leq \min_{\mathbf{z}_L \in \mathcal{Z}(\mathbf{x}_0,\epsilon)} \max\left\{\mathbf{c}_{yk}^T\mathbf{z}_L, \mathbf{c}_{a_1k}^T\mathbf{z}_L, \ldots, \mathbf{c}_{a_Mk}^T\mathbf{z}_L\right\}, \tag{4}$$

Since the set $\mathcal{Z}(\mathbf{x}_0,\epsilon)$ is highly nonconvex, verifying (4) is computationally expensive.

Following the IBP approach to relax the nonconvex set $\mathcal{Z}(\mathbf{x}_0,\epsilon)$ leads to the following result:

Theorem 1 *Condition* (4) *is satisfied if for all $k \neq y$:*

$$0 \geq \min_{\boldsymbol{\eta}\in\mathcal{P}} \max_{\underline{\mathbf{z}}_{L-1}\leq\mathbf{z}_{L-1}\leq\bar{\mathbf{z}}_{L-1}} -c_k(\boldsymbol{\eta})^T(\mathbf{W}_L\mathbf{z}_{L-1}+\mathbf{b}_L), \tag{5}$$

where $\mathcal{P} = \{(\eta_0,\ldots,\eta_M)|\sum_{i=0}^M \eta_i = 1, \eta_i \geq 0, \forall i = 0,1,\ldots,M\}$, and $c_k(\boldsymbol{\eta}) = \eta_0\mathbf{c}_{yk} + \eta_1\mathbf{c}_{a_1k}\cdots+\eta_M\mathbf{c}_{a_Mk}$. Here, the bounds $\underline{\mathbf{z}}_{L-1}$ and $\bar{\mathbf{z}}_{L-1}$ are obtained by IBP.

Unlike (4), the condition in (5) is easy to verify computationally. To understand this, let us define

$$J_k(\boldsymbol{\eta}) = \max_{\underline{\mathbf{z}}\leq\mathbf{z}_{L-1}\leq\bar{\mathbf{z}}} -c_k(\boldsymbol{\eta})^T(W_L\mathbf{z}_{L-1}+\mathbf{b}_L). \tag{6}$$

Then, our aim in (5) is to minimize $J_k(\boldsymbol{\eta})$ over $\mathcal{P}$.

First notice that the maximization problem (6) can be solved in closed form as described in Step 4 of Algorithm 1. Consequently, one can rely on Danskin's Theorem [Danskin, 2012] to compute the subgradient of the function $J_k(\cdot)$. Thus, to minimize $J_k(\cdot)$ in (5), we can rely on the Bregman proximal (sub)gradient method (see [Gutman and Pena, 2018] and the references therein). This algorithm is guaranteed to find $\epsilon-$ accurate solution to (5) in $T = O(1/\sqrt{\epsilon})$ iterations–see [Gutman and Pena, 2018, Corollary 2].

Algorithm 1 IBP verification of networks with multiple abstain classes

1: **Parameters**: Stepsize $\nu > 0$, number of iterations T.
2: Initialize $\eta_0 = 1$ and $\eta_1 = \ldots = \eta_M = 0$.
3: **for** $t = 0, 1, \ldots, T$ **do**
4: Set $[\mathbf{z}_{L-1}^{*t}]_j = \begin{cases} [\underline{\mathbf{z}}_{L-1}]_j & \text{if } [\mathbf{W}_L^T c(\boldsymbol{\eta})]_j \geq 0 \\ [\bar{\mathbf{z}}_{L-1}]_j & \text{otherwise.} \end{cases}$, for every j.
5: Set $\eta_m^{t+1} = \frac{\eta_m^t \exp(-2\nu(\mathbf{z}_{L-1}^{*t})^T \mathbf{W}_L^T \mathbf{c}_{a_m k})}{\sum_{j=0}^{M} \eta_j^t \exp(-2\nu(\mathbf{z}_{L-1}^{*t})^T \mathbf{W}_L^T \mathbf{c}_{a_j k})}$, $\forall m \in \{0, \ldots, M\}$, where a_0 is defined as y.

Based on the verifier developed above, we can **train** provably robust neural networks with multiple detection classes against adversarial attacks (see Appendix B).

4 Numerical Results

We devise a diverse set of experiments on shallow and deep networks to investigate the effectiveness of our proposed joint classifier and detector with multiple abstain classes. To train the neural networks on MNIST and CIFAR-10 datasets, we use Algorithm 2 as a part of an optimizer scheduler. In the first phase, we set $\lambda_1 = \lambda_2 = 0$. Thus, the network is trained without considering any abstain classes initially. In the second phase we optimize the objective function (12), where we linearly increase ϵ from 0 to ϵ_{train}. In the last phase, we further tune the network on the fixed $\epsilon = \epsilon_{\text{train}}$ (see Appendix F for further details).

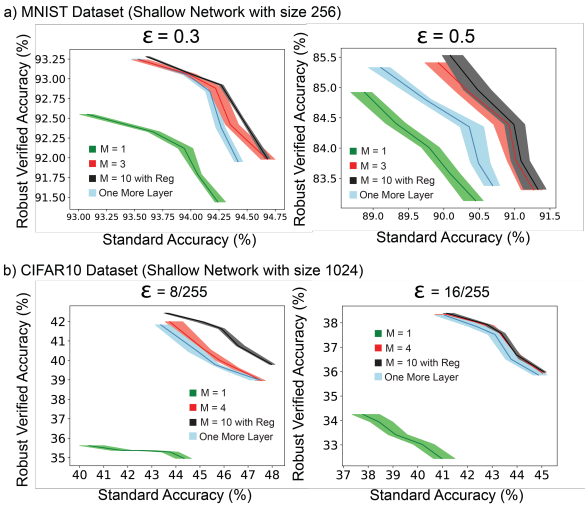

Figure 2: Performance of Multiple-abstain shallow networks on MNIST and CIFAR-10 datasets. We compared multiple abstain neural networks (both regularized and non-regularized version) with the single abstain networks and networks with one more layer. The above and below rows demonstrate the trade-off between standard and robust verified accuracy on MNIST and CIFAR-10 datasets.

In the first set of experiments depicted in Figure 2, we compare the performance of the shallow networks with the optimal number of abstain classes to the single abstain network, the network with an additional layer, and the network regularized to have balance between different abstain classes (see Appendix C). The shallow networks have one convolutional layer with size 256 and 1024 for training on MNIST and CIFAR-10 datasets respectively. This convolutional layer is connected to the second (last) layer consisting of $K + M$ nodes where K is the number of regular classes (10 for both MNIST and CIFAR-10 datasets) and M is the number of abstain classes. The optimal number of abstain classes is obtained by changing the number of them from $M = 1$ to $M = 20$ on both CIFAR-10 and MNIST datasets. The optimal value for the network trained on MNIST is $M = 3$ and $M = 4$ for CIFAR-10 dataset. Moreover, we compare the optimal multi-abstain shallow network to two other baselines: One is the network with the number of abstain classes equal to the number of regular classes ($M = K$) and is trained via the regularizer described in (13). The other baseline is a network with one more layer compared to the shallow

network. Instead of the last layer in the shallow network, this network has $K + M$ nodes in the layer one to the last, and $K + 1$ nodes in the last layer. Ideally the set of models can be supported by such a network is a super-set of the original shallow network. However, due to the training procedure (IBP) which is sensitive to higher number of layers (the higher the number of layers, the looser the lower and upper bounds), we obtain better results with the original network with multiple abstain classes.

Figure 3 shows the percentage of adversarial examples captured by each abstain class ($M = 10$) on CIFAR-10 dataset for both regularized and non-regularized networks. The hyper-parameter γ is set to $\frac{1}{K+M} = \frac{1}{20}$ in (13). For further experiments on deep networks and comparison with other state-of-the-art approaches see Appendix E.

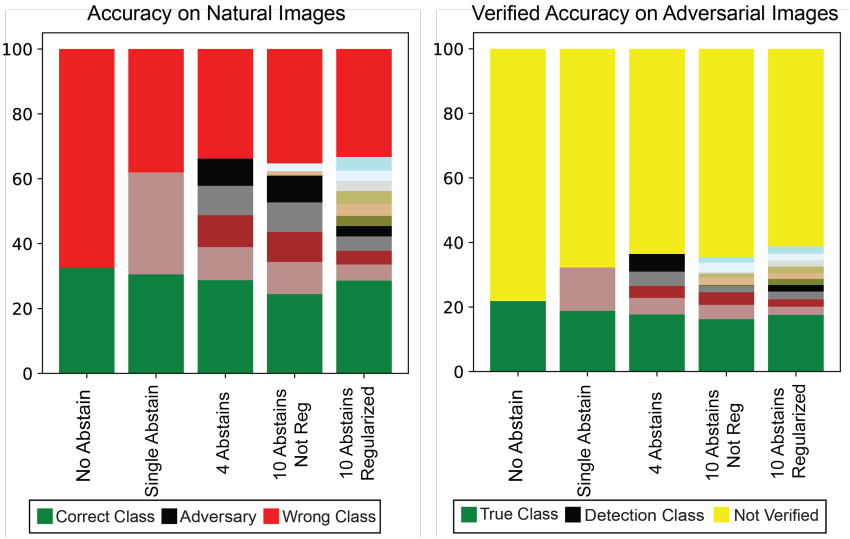

Figure 3: Distribution of natural and adversarial images over different abstain classes on CIFAR-10 dataset. When there are 10 abstain classes, model degeneracy leads to lower performance compared to the baseline. Adding the regularization term (right most column) will utilize all abstain classes and enhance both standard and robust verified accuracy. standard accuracy is the proportion of correctly classified natural images, while robust verified accuracy is the proportion of images that are robust against all adversarial attacks within the ϵ-neighborhood.

Beside the shallow networks, networks with multiple abstain classes show a better trade-off between standard and verified robust accuracy on the deep networks (See Table 1). The structure of the trained deep network is exactly as same as the one described in Sheikholeslami et al. [2021].

ϵ	Method	Standard Error (%)	Robust Verified Error (%)
$\epsilon_{\text{train}} = 8.8/255$ $\epsilon_{\text{test}} = 8/255$	Interval Bound Propagation [Gowal et al., 2018]	50.51	68.44
	IBP-CROWN [Zhang et al., 2019]	54.02	66.94
	[Balunovic and Vechev, 2019]	48.3	72.5
	Single Abstain [Sheikholeslami et al., 2021]	55.60	63.63
	Multiple Abstain Classes (Current Work)	56.72	61.45
	Multiple Abstain Classes (Verified by Beta-crown)	56.72	57.55
$\epsilon_{\text{train}} = 17.8/255$ $\epsilon_{\text{test}} = 16/255$	Interval Bound Propagation [Gowal et al., 2018]	68.97	78.12
	IBP-CROWN [Zhang et al., 2019]	66.06	76.80
	Single Abstain [Sheikholeslami et al., 2021]	66.37	67.92
	Multiple Abstain Classes (verified by IBP)	66.25	64.57
	Multiple Abstain Classes (Verified by Beta-crown)	66.25	62.81

Table 1: Standard and Robust Verified error of state-of-the-art approaches on CIFAR-10 dataset.

References

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25, 2012.

Zheng Zhu, Guan Huang, Jiankang Deng, Yun Ye, Junjie Huang, Xinze Chen, Jiagang Zhu, Tian Yang, Jiwen Lu, Dalong Du, et al. Webface260m: A benchmark unveiling the power of million-scale deep face recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10492–10502, 2021.

Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In 2013 IEEE international conference on acoustics, speech and signal processing, pages 6645–6649. Ieee, 2013.

Ali Bou Nassif, Ismail Shahin, Imtinan Attili, Mohammad Azzeh, and Khaled Shaalan. Speech recognition using deep neural networks: A systematic review. IEEE access, 7:19143–19165, 2019.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572, 2014.

Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. arXiv preprint arXiv:1706.02690, 2017.

Xiaoyong Yuan, Pan He, Qile Zhu, and Xiaolin Li. Adversarial examples: Attacks and defenses for deep learning. IEEE transactions on neural networks and learning systems, 30(9):2805–2824, 2019.

Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In 2016 IEEE symposium on security and privacy (SP), pages 582–597. IEEE, 2016.

Yunseok Jang, Tianchen Zhao, Seunghoon Hong, and Honglak Lee. Adversarial defense via learning to generate diverse attacks. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 2740–2749, 2019.

Micah Goldblum, Liam Fowl, Soheil Feizi, and Tom Goldstein. Adversarially robust distillation. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, pages 3996–4003, 2020.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083, 2017.

Tianjian Huang, Shaunak Halbe, Chinnadhurai Sankar, Pooyan Amini, Satwik Kottur, Alborz Geramifard, Meisam Razaviyayn, and Ahmad Beirami. Dair: Data augmented invariant regularization. arXiv preprint arXiv:2110.11205, 2021.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199, 2013.

Florian Tramer, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. On adaptive attacks to adversarial example defenses. arXiv preprint arXiv:2002.08347, 2020.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In 2017 ieee symposium on security and privacy (sp), pages 39–57. IEEE, 2017.

Vincent Tjeng, Kai Xiao, and Russ Tedrake. Evaluating robustness of neural networks with mixed integer programming. arXiv preprint arXiv:1711.07356, 2017.

Kai Y Xiao, Vincent Tjeng, Nur Muhammad Shafiullah, and Aleksander Madry. Training for faster adversarial robustness verification via inducing relu stability. arXiv preprint arXiv:1809.03008, 2018.

Sven Gowal, Krishnamurthy Dvijotham, Robert Stanforth, Rudy Bunel, Chongli Qin, Jonathan Uesato, Relja Arandjelovic, Timothy Mann, and Pushmeet Kohli. On the effectiveness of interval bound propagation for training verifiably robust models. arXiv preprint arXiv:1810.12715, 2018.

Cassidy Laidlaw and Soheil Feizi. Playing it safe: Adversarial robustness with an abstain option. arXiv preprint arXiv:1911.11253, 2019.

Fatemeh Sheikholeslami, Ali Lotfi Rezaabad, and J Zico Kolter. Provably robust classification of adversarial examples with detection. International Conference on Learning Representations (ICLR), 2021.

Jiefeng Chen, Jayaram Raghuram, Jihye Choi, Xi Wu, Yingyu Liang, and Somesh Jha. Revisiting adversarial robustness of classifiers with a reject option. In The AAAI-22 Workshop on Adversarial Machine Learning and Beyond, 2021.

John M Danskin. The theory of max-min and its application to weapons allocation problems, volume 5. Springer Science & Business Media, 2012.

David H Gutman and Javier F Pena. A unified framework for bregman proximal methods: subgradient, gradient, and accelerated gradient schemes. arXiv preprint arXiv:1812.10198, 2018.

Huan Zhang, Hongge Chen, Chaowei Xiao, Sven Gowal, Robert Stanforth, Bo Li, Duane Boning, and Cho-Jui Hsieh. Towards stable and efficient training of verifiably robust neural networks. arXiv preprint arXiv:1906.06316, 2019.

Mislav Balunovic and Martin Vechev. Adversarial training and provable defenses: Bridging the gap. In International Conference on Learning Representations, 2019.

Fatemeh Sheikholeslami, Swayambhoo Jain, and Georgios B Giannakis. Minimum uncertainty based detection of adversaries in deep neural networks. In 2020 Information Theory and Applications Workshop (ITA), pages 1–16. IEEE, 2020.

Gagandeep Singh, Timon Gehr, Matthew Mirman, Markus Püschel, and Martin T Vechev. Fast and effective robustness certification. NeurIPS, 1(4):6, 2018.

Sumanth Dathathri, Krishnamurthy Dvijotham, Alexey Kurakin, Aditi Raghunathan, Jonathan Uesato, Rudy Bunel, Shreya Shankar, Jacob Steinhardt, Ian Goodfellow, Percy Liang, et al. Enabling certification of verification-agnostic networks via memory-efficient semidefinite programming. arXiv preprint arXiv:2010.11645, 2020.

Shiqi Wang, Huan Zhang, Kaidi Xu, Xue Lin, Suman Jana, Cho-Jui Hsieh, and J Zico Kolter. Beta-crown: Efficient bound propagation with per-neuron split constraints for complete and incomplete neural network verification. arXiv preprint arXiv:2103.06624, 2021.

Maher Nouiehed and Meisam Razaviyayn. Learning deep models: Critical points and local openness. INFORMS Journal on Optimization, 2021.

A Training a Joint Robust Classifier and Detector

Sheikholeslami et al. [2021] improves the performance tradeoff on natural and adversarial examples by introducing an auxiliary class for detecting adversarial examples. If this auxiliary class is selected as the output, the networks "abstains" from declaring any of the original K classes for the given input. Let a be the abstain class. The network classifies performs correctly on an adversarial image if it is classified correctly (similar to robust networks without detectors) or it is classified as the abstain class (detected as an adversarial example). Hence, the network is verified against a certain class k if

$$0 \leq \min_{\mathbf{z} \in \mathcal{Z}(\mathbf{x}_0, \epsilon)} \max(\mathbf{c}_{yk}^T \mathbf{z}, \mathbf{c}_{ak}^T \mathbf{z}), \tag{7}$$

i.e., if the score of the true label y or the score of the abstain class a is larger than the score of class k. To train a neural network that can jointly detect and classify a dataset of images, Sheikholeslami et al. [2021] relies on the loss function of the form:

$$L_{\text{Total}} = L_{\text{Robust}} + \lambda_1 L_{\text{Robust}}^{\text{Abstain}} + \lambda_2 L_{\text{Natural}}, \tag{8}$$

where the term L_{Natural} denotes the natural loss when no adversarial examples are considered. More precisely, $L_{\text{Natural}} = \frac{1}{n} \sum_{i=1}^{n} \ell_{\text{xent}}\big(\mathbf{z}_L(\mathbf{x}_i), y_i\big)$, where ℓ_{xent} is the standard cross-entropy loss. The term L_{Robust} in (8) represents the worst-case adversarial loss used in [Madry et al., 2017], without considering the abstain class. Precisely,

$$\begin{aligned} L_{\text{Robust}} = \max_{\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_n} \; & \frac{1}{n} \sum_{i=1}^{n} \ell_{\text{xent}}\big(\mathbf{z}_L(\mathbf{x}_i + \boldsymbol{\delta}_i), y_i\big) \\ \text{s.t. } & \|\boldsymbol{\delta}_i\|_\infty \leq \epsilon, \;\; \forall i = 1, \ldots, n. \end{aligned}$$

Finally, the Robust-Abstain loss $L_{\text{Robust}}^{\text{Abstain}}$ is the minimum of the detector and the classifier losses:

$$\begin{aligned} L_{\text{Robust}}^{\text{Abstain}} = \max_{\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_n} \; & \frac{1}{n} \sum_{i=1}^{n} \min \Big(\ell_{\text{xent}}\big(\mathbf{z}_L(\mathbf{x}_i + \boldsymbol{\delta}_i), y_i\big), \\ & \qquad \ell_{\text{xent}}\big(\mathbf{z}_L(\mathbf{x}_i + \boldsymbol{\delta}_i), a\big) \Big) \\ & \text{s.t. } \|\boldsymbol{\delta}_i\|_\infty \leq \epsilon, \; \forall i \end{aligned} \tag{9}$$

In (8), tuning λ_1 and λ_2 controls the trade-off between standard and robust accuracy. Furthermore, to obtain non-trivial results, IBP-relaxation should be incorporated during training for the minimization sub-problems in L_{robust} and $L_{\text{robust}}^{\text{abstain}}$ [Sheikholeslami et al., 2021, Gowal et al., 2018].

B Training of Neural Networks with Multiple Abstain Classes

To train a neural network consisting of multiple abstain classes, we follow a similar combination of loss functions as in (8). While the last term (L_{Natural}) can be computed efficiently, the first and second terms cannot be computed efficiently because even evaluating the functions L_{Robust} and $L_{\text{Robust}}^{\text{Abstain}}$ requires maximizing nonconcave functions. Thus, instead of minimizing these two terms, we will minimize their upper-bounds. Particularly, following [Sheikholeslami et al., 2020, Equation (17)], we use $\bar{L}_{\text{Robust}}$ as an upper-bound to L_{Robust}. This upper-bound is obtained by the IBP relaxation procedure described in Gowal et al. [2018]. To obtain an upper-bound for the Robust-Abstain loss term $L_{\text{Robust}}^{\text{Abstain}}$ in (8), let us first start by clarifying its definition in the multi-abstain class scenario:

$$\begin{aligned} L_{\text{Robust}}^{\text{Abstain}} = \max_{\delta_1, \ldots, \delta_n} \; & \frac{1}{n} \sum_{i=1}^{n} \min \Big\{ \ell_{\text{xent}}\big(\mathbf{z}_L(\mathbf{x}_i + \delta_i), y_i\big), \\ & \qquad \min_{m=1,\ldots,M} \ell_{\text{xent}}\big(\mathbf{z}_L(\mathbf{x}_i + \delta_i), a_m\big) \Big\}. \end{aligned} \tag{10}$$

This definition implies that the classification is considered "correct" for a given input if the predicted label is the ground-truth label or if it is assigned to one of the abstain classes. Since the maximization problem w.r.t. $\{\delta_i\}$ is nonconcave, it is hard to even evaluate $L_{\text{Robust}}^{\text{Abstain}}$. Thus, we minimize an efficiently computable upper-bound of this loss function as described in Theorem 2.

Theorem 2 *Let*

$$\ell_{\textit{Robust}}^{\textit{Abstain}}(\mathbf{x}, y) = \max_{\|\delta\| \leq \epsilon} \min \Big\{ \ell_{\textit{xent}}\big(\mathbf{z}_L(\mathbf{x}+\delta), y\big), \min_{m=1,\ldots,M} \ell_{\textit{xent}}\big(\mathbf{z}_L(\mathbf{x}_i+\delta_i), a_m\big) \Big\}$$

Then,

$$\ell_{\textit{Robust}}^{\textit{Abstain}}(\mathbf{x}, y) \leq \bar{\ell}_{\textit{Robust}}^{\textit{Abstain}}(\mathbf{x}, y) = \ell_{\textit{xent}\setminus\mathcal{A}_0}(J(\mathbf{x}), y), \tag{11}$$

where $J(\mathbf{x})$ is a vector whose k-th component equals $J_k(\mathbf{x})$ as defined in (6) *and* $\ell_{\textit{xent}\setminus\mathcal{A}_0}(\mathbf{x}_0, y) := -\log\left(\frac{\exp(\mathbf{e}_y^T \mathbf{z}_L(\mathbf{x}_0))}{\sum_{i\in\mathcal{I}\setminus\mathcal{A}_0}\exp(\mathbf{e}_i^T \mathbf{z}_L(\mathbf{x}_0))}\right)$. *Here, $\mathcal{I} = \{1,\ldots,K,a_1,\ldots,a_M\}$ is the set of all classes (true labels and abstain classes) and $\mathcal{A}_0 = \{a_1,\ldots,a_M\}$ is the set of abstain classes.*

Notice that the definition of $\ell_{\text{xent}\setminus\mathcal{A}_0}(\mathbf{x}_0, y)$ removes the terms corresponding to the abstain classes in the denominator. This definition is less restrictive toward abstain classes compared to incorrect classes. Thus, for a given sample, it is more advantageous for the network to classify it as an abstain class instead of incorrect classification. This mechanism enhances the performance of the network on detecting adversarial examples by abstain classes, while it does not have an adverse effect on the performance of the network on natural samples.Note that during the evaluation/test phase, this loss function does not change the final prediction of the network for a given input, since the winner (the entry with the highest score) remains the same.

Overall, we upper-bound the loss in (8) by replacing L_{Robust} with the IBP relaxation approach utilized in Gowal et al. [2018], Sheikholeslami et al. [2021] and replacing $L_{\text{Robust}}^{\text{Abstain}}$ with $\bar{L}_{\text{Robust}}^{\text{Abstain}} = \frac{1}{n}\sum_{i=1}^{n} \bar{\ell}_{\text{Robust}}^{\text{Abstain}}(\mathbf{x}_i, y_i)$ presented in Theorem 2. Thus our total training loss can be presented as:

$$L_{\text{Total}} = \bar{L}_{\text{Robust}} + \lambda_1 \bar{L}_{\text{Robust}}^{\text{Abstain}} + \lambda_2 L_{\text{Natural}} \tag{12}$$

Algorithm 2 describes the procedure of optimizing (12) on a joint classifier and detector with multiple abstain classes.

Algorithm 2 Train a robust neural network on a training data

1: **Input**: Batches of data $\mathcal{D}_1,\ldots,\mathcal{D}_R$, step-size ν, $\boldsymbol{\theta}(L)$: optimization parameters for loss L.
2: **for** $t = 1,\ldots,R$ **do**
3: Let $(\mathbf{x}_1, y_1),\ldots,(\mathbf{x}_N, y_N) \in \mathcal{D}_t$
4: Compute $J_o(\mathbf{x}) \quad \forall\, \mathbf{x} \in \mathcal{D}_t, \quad \forall o \in \{1,\ldots,K\}$ by Algorithm 1.
5: Compute L_{Robust} as described in Gowal et al. [2018] on Batch $\mathcal{D}_t$.
6: Compute $\bar{L}_{\text{Robust}}^{\text{abstain}}$ on Batch $\mathcal{D}_t$ using Theorem 2.
7: $\boldsymbol{\theta}(L) = \boldsymbol{\theta}(L) - \nu\nabla\big(\boldsymbol{\theta}(\bar{L}_{\text{Robust}}) + \lambda_1\boldsymbol{\theta}(\bar{L}_{\text{Robust}}^{\text{abstain}}) + \lambda_2\boldsymbol{\theta}(L_{\text{Natural}})\big)$

C Balance Between Abstain Classes and Model Degeneracy

Having multiple abstain classes can potentially increase the capacity of our classifier to detect adversarial examples. However, as we will see in Figure 3 (10 abstains, unregularized), several abstain classes collapse together and capture similar adversarial patterns. Such a phenomenon, which we referred to as "model degeneracy" and is illustrated with an example in Appendix K, will prevent us from utilizing all abstain classes fully.

To address this issue, we impose a regularization term to the loss function such that the network utilizes all abstain classes in balance. We aim to make sure the $\boldsymbol{\eta}$ values are distributed nearly uniformly and there are no *idle* abstain classes. Let $\boldsymbol{\eta}^{ik}$, $\mathbf{z}_{L-1}(\mathbf{x}_i)$, and y_i be the abstain vector corresponding to the sample $\mathbf{x}_i$ verifying against the target class k, the output of the layer $L-1$, and the assigned label to the data point $\mathbf{x}_i$ respectively. Therefore, the regularized verification problem over n given samples takes the following form:

$$\min_{\boldsymbol{\eta}^1,\ldots,\boldsymbol{\eta}^n \in \mathcal{P}} \quad \sum_{i=1}^{n} \sum_{k\neq y_i} \max_{\underline{\mathbf{z}}(\mathbf{x}_i) \leq \mathbf{z}_{L-1} \leq \bar{\mathbf{z}}(\mathbf{x}_i)} -c_k(\boldsymbol{\eta}^{ik})$$

$$(\mathbf{W}_L\mathbf{z}_{L-1} + \mathbf{b}_L) + \mu\|[\frac{\gamma\mathbf{1}}{M+1} - \frac{\sum_{j=1}^{n}\sum_{o\neq y_i}\boldsymbol{\eta}^{jo}]_+}{n(K-1)}\|^2, \tag{13}$$

The above regularizer penalizes the objective function if the average value of $\boldsymbol{\eta}$ coefficient corresponding to a given abstain class over all samples of the batch is smaller than a threshold (the threshold is determined by the hyper-parameter γ). In other words, if an abstain class is not contributing enough to the detection of adversarial samples, it will be penalized accordingly. Note that if γ is larger, we penalize an *idle* abstain class more.

Note that in the unregularized case, the optimization of parameters $\boldsymbol{\eta}^{ik}$ are independent of each other. In contrast, by adding the regularizer described in (13) we require to optimize $\boldsymbol{\eta}^{ik}$ parameters of different samples and target classes jointly (they are coupled in the regularization term). Since optimizing (13) over the set of all n samples is infeasible for datasets with large number of samples, we solve the problem over smaller batches of the data to reduce the complexity of problem in each iteration. We utilize the same Bregman divergence procedure used in Algorithm 1, while the gradient with respect to $\boldsymbol{\eta}^{ik}$ takes the regularization term into account as well.

D Verification with β-Crown

Despite its simplicity, IBP-based verification comes with a certain limitation, namely the looseness of its layer-by-layer bounds of the input. To overcome this limitation, tighter verification methods have been proposed in the literature [Singh et al., 2018, Zhang et al., 2019, Dathathri et al., 2020, Wang et al., 2021]. Among these, β-crown [Wang et al., 2021] utilizes the branch-and-bound technique to generalize and improve the IBP-CROWN proposed in Zhang et al. [2019]. Let $\underline{\mathbf{z}}_i$ and $\bar{\mathbf{z}}_i$ be the estimated element-wise lower-bound and upper-bounds for the pre-activation value of $\mathbf{z}_i$, i.e., $\underline{\mathbf{z}}_i \leq \mathbf{z}_i \leq \bar{\mathbf{z}}_i$, where these lower and upper bounds are obtained by the method in Zhang et al. [2019]. Let $\hat{\mathbf{z}}_i$ be the value we obtain by applying ReLU function to $\mathbf{z}_i$. We say a neuron is unstable if its sign after applying ReLU activation cannot be determined based on only knowing the corresponding lower and upper bounds. That is, a neuron is unstable if $\underline{\mathbf{z}}_i < 0 < \bar{\mathbf{z}}_i$. For **stable** neurons, no relaxation is needed to enforce convexity of $\sigma(\mathbf{z})$ (since the neuron operates in a linear regime). On the other hand, given an unstable neuron, they use branch-and-bound (BAB) approach to split the input range of the neuron into two sub-domains $\mathcal{C}_{il} = \{\mathbf{x} \in \mathcal{C}(\mathbf{x}_0, \epsilon) | \ \hat{z}_i \leq 0\}$ and $\mathcal{C}_{iu} = \{\mathbf{x} \in \mathcal{C}(\mathbf{x}_0, \epsilon) | \ \hat{z}_i > 0\}$. Within each subdomain, the neuron operates linearly and hence verification is easy. Thus we can verify for each of these subdomains separately. If we have N unstable nodes, BAB algorithm requires the investigation of 2^N sub-domains in the worst-case. β-Crown proposes a heuristic for traversing all these subdomains: The higher the absolute value of the corresponding lower-bound of a node is, the sooner it is visited by the verifier. For verifying each sub-problem, Wang et al. [2021] proposed a lower-bounded which requires solving a maximization problem over two parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$:

$$\begin{aligned} &\min_{\mathbf{z} \in \mathcal{Z}(\mathbf{x}_0, \epsilon)} \mathbf{c}_{yk}^T \mathbf{z} \geq \max_{\boldsymbol{\alpha}, \boldsymbol{\beta}} g(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \\ &\text{where} \quad g(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = (\mathbf{a} + \mathbf{P}_{\boldsymbol{\alpha}} \boldsymbol{\beta})^T \mathbf{x} + \mathbf{q}_{\boldsymbol{\alpha}}^T \boldsymbol{\beta} + \mathbf{d}_{\boldsymbol{\alpha}}. \end{aligned} \tag{14}$$

Here, the matrix $\mathbf{P}$ and the vectors $\mathbf{q}, \mathbf{a}$ and $\mathbf{d}$ are functions of $\mathbf{W}_i, \mathbf{b}_i, \underline{\mathbf{z}}_i, \bar{\mathbf{z}}_i, \boldsymbol{\alpha}$, and $\boldsymbol{\beta}$ parameters. See Wang et al. [2021] for the precise definition of g. Notice that any choice of $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ provides a valid lower bound for verification. However, optimizing $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ in (14) leads to a tighter bound.

Now, we focus on β–Crown verification of networks with multiple abstain classes. To this end, we will find a sufficient condition for (4) using the lower-bound technique of (14) in β–Crown. In particular, by switching the minimization and maximization in (4) and using the β–Crown lower bound (14), we can find a lower-bound of the form

$$\begin{aligned} &\min_{\mathbf{z}_L \in \mathcal{Z}(\mathbf{x}_0, \epsilon)} \max\{\mathbf{c}_{yk}^T \mathbf{z}_L, \mathbf{c}_{a_1 k}^T \mathbf{z}_L, \ldots, \mathbf{c}_{a_M k}^T \mathbf{z}_L\} \geq \\ &\max_{\boldsymbol{\eta} \in \mathcal{P}, \boldsymbol{\alpha}, \boldsymbol{\beta} \geq 0} G(\mathbf{x}_0, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\eta}). \end{aligned} \tag{15}$$

The details of this inequality and the exact definition of function $G(\cdot)$ is provided in Appendix J. Note that any feasible solution to the right hand side of (15) is a valid lower-bound to the original verification problem (left-hand-side). Thus, in order for (4) to be satisfied, it suffices to find a feasible $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\eta})$ such that $G(\mathbf{x}_0, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\eta}) \geq 0$. To optimize the RHS of (15) in Algorithm 3, we utilize AutoLirpa library of [Zhang et al., 2019] for updating $\boldsymbol{\alpha}$, and use Bregman proximal subgradient method to update $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$ – See appendix G. We use Euclidean norm Bregman divergence for updating $\boldsymbol{\beta}$, and Shannon entropy Bregman divergence for $\boldsymbol{\eta}$ to obtain closed-form updates.

Algorithm 3 β–Crown verification of networks with multiple abstain classes

Input: number of iterations T, number of iterations in the inner-loop T_0, Step-size γ.
for $t = 0, 1, \ldots, T$ **do**
  Update α using AutoLirpa library [Zhang et al., 2019]
  for $k = 0, 1, \ldots, T_0$ **do**
    $\boldsymbol{\beta} = [\boldsymbol{\beta} + \gamma \frac{\partial G(\mathbf{x}_0, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\eta})}{\partial \boldsymbol{\beta}}]_+$, where $[w]_+ = \max\{0, w\}$ is projection to non-negative orthant
    $\eta_m^{\text{new}} = \frac{\eta_m^{\text{old}} \exp(2\gamma \frac{\partial G(\mathbf{x}_0, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\eta})}{\partial \eta_m})}{\sum_{j=0}^{M} \eta_j^{\text{old}} \exp(2\gamma \frac{\partial G(\mathbf{x}_0, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\eta})}{\partial \eta_j})}, \quad \forall m \in \{0, \ldots, M\}$

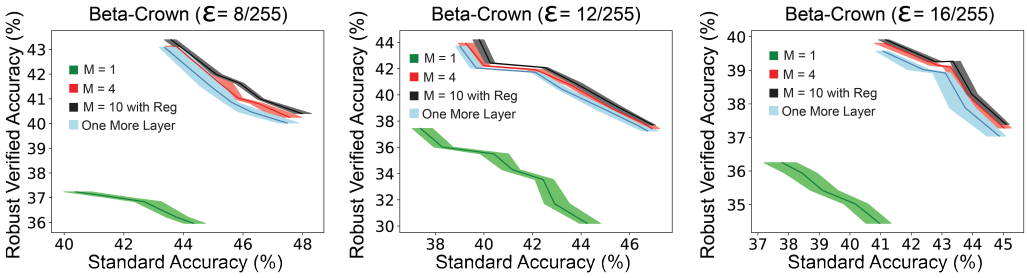

Figure 4: Performance of β-crown on verification of Neural Networks with single abstain, 4 abstain classes, 10 abstain classes with regularized, and a network with one more layer (single abstain) on CIFAR-10 dataset. $M = 1$ coincides with Sheikholeslami et al. [2021] approach.

E Further Experiments

In Figure 5, we investigate the effect of changing the number of abstain classes of the shallow network described above. As we observe, the unregularized network and the network with one more layer is much more sensitive to the change of M than the regularized version. This means, we can use the regularized network with the same performance while it does not require to be tuned for the optimal M. In the unregularized version, by increasing the number of abstain classes from $M = 1$ to $M = 5$ we see improvement. However, after this threshold, the network performance drops gradually such that for $M = 10$ where the number of labels and abstain classes are equal ($M = K = 10$) the performance of the network in this case is even worse than the single-abstain network due to the model degeneracy of the multi-abstain network. However, the network trained on the regularized loss maintains the its performance when M changes from the optimal value to larger values.

Moreover, we illustrate the performance of networks trained in the first set of experiments by β-crown in Figure 4 . The networks whose robust accuracy are verified by Beta-crown has 1% to 2% improvement compared to the same networks verified by IBP on average.

F Implementation Details

In table F, we demonstrate the structure of the deep networks used in experiments of Table 1. The scheduler used for the experiments is the one utilized by Sheikholeslami et al. [2021]. On both MNIST and CIFAR-10 datasets, we have used an Adam optimizer with learning rate 5×10^{-4}. κ is scheduled by a linear ramp-down process, starting at 1, which after a warm-up period is ramped down to value $\kappa_{\text{end}} = 0.5$. Value of ϵ during the training is also simultaneously scheduled by a linear ramp-up, starting at 0 and ϵ_{Train} as the final value. The networks are trained with four NVIDIA V100 GPUs. The trade-off between standard accuracy on clean images, and robust verified accuracy can be tuned by changing λ_2 from 0 to $+\infty$ where the larger values correspond to more robust networks. For the networks with the regularizer addressing the model degeneracy issue, we choose γ by tuning it in the $[\frac{0.1}{K+M}, \frac{1.5}{K+M}]$. Our observations on both MNIST and CIFAR-10 datasets for different ϵ values show that the optimal value for γ is consistently close to $\frac{1}{K+M}$. Thus, we suggest to choose hyper-parameter $\gamma = \frac{1}{K+M}$ where K is the number of labels and M is the number of

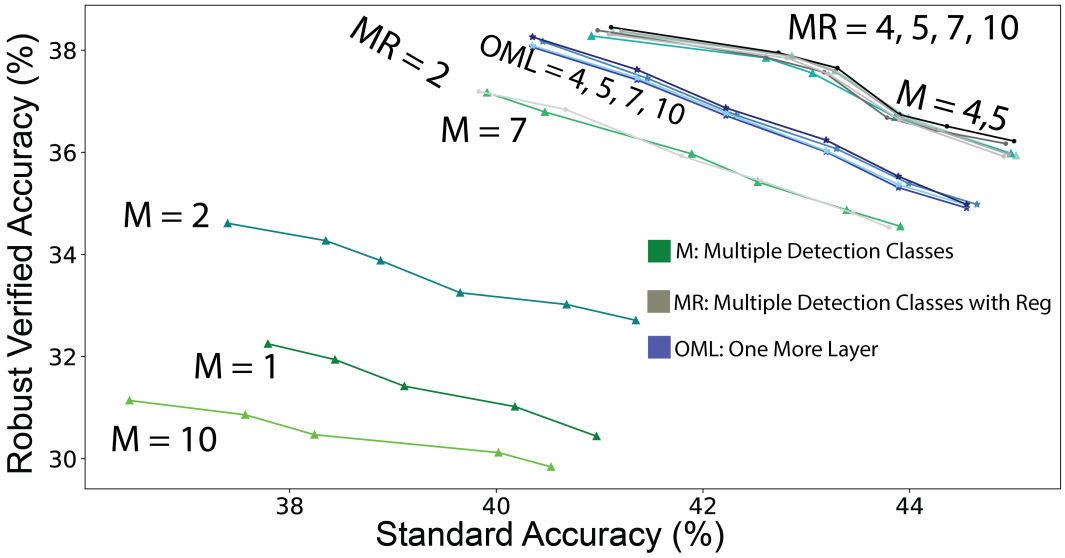

Figure 5: Performance of Multiple-abstain shallow networks on CIFAR-10 datasets.

detection classes. The optimal value for M is 4 for the CIFAR-10 and $M = 3$ for MNIST dataset. By adding the "model degeneracy" regularizer, the obtained network has nearly the same performance for $M \in [4, 2K]$. Overall, we suggest to choose $M = K$ and $\gamma = \frac{1}{K+M}$ as the default values for hyper-parameters M and γ.

Network Layers
Conv 64 3×3
Conv 64 3×3
Conv 128 3×3
Conv 128 3×3
Fully Connected 512
Linear 10

Table 2: Standard and Robust Verified error of state-of-the-art approaches on CIFAR-10 dataset.

1. For MNIST, we train on a single Nvidia V100 GPU for 100 epochs with batch sizes of 100. The total number of training steps is 60K. We decay the learning rate by $10\times$ at steps 15K and 25K. We use warm-up and ramp-up duration of 2K and 10K steps, respectively. We do not use any data augmentation techniques and use full 28×28 images without any normalization.

2. CIFAR-10, we train for 3200 epochs with batch sizes of 1600. The total number of training steps is 100K. We decay the learning rate by 10× at steps 60K and 90K. We use warm-up and ramp-up duration of 5K and 50K steps, respectively. During training, we add random translations and flips, and normalize each image channel (using the channel statistics from the train set).

G Bregman-Divergence Method for Optimizing a Convex Function Over a Probability Simplex

In this section, we show how to optimize a convex optimization problem over a probability simplex by using the Bregman divergence method. Let $\boldsymbol{\eta}$ be a vector of n elements. We aim to minimize the following constrained optimization problem where J is a convex function with respect to $\boldsymbol{\eta}$:

$$\min_{\eta_1,\dots,\eta_n} J(\eta_1,\dots,\eta_n) \quad \text{subject to} \quad \sum_{i=1}^{n}\eta_i = 1, \quad \eta_i \geq 0 \quad \forall i = 1,\dots,n. \tag{16}$$

To solve the above problem, we define the Bregman distance function as:

$$B(\mathbf{x},\mathbf{y}) = \gamma(\mathbf{x}) - \gamma(\mathbf{y}) - \langle \nabla\gamma(\mathbf{x}), \mathbf{x}-\mathbf{y}\rangle$$

where γ is a strictly convex function. For this specific problem where the constrain is over a probability simplex, we choose $\gamma(\mathbf{x}) = \sum_{i=1}^{n} x_i \log(x_i)$. Thus:

$$B(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{n} x_i \log(\frac{x_i}{y_i})$$

One can rewrite problem 16 as:

$$\min_{\eta_1,\dots,\eta_n} J(\eta_1,\dots,\eta_n) + \mathcal{I}_{\mathbb{P}}(\boldsymbol{\eta}) \tag{17}$$

where $\mathbb{P} = $. Applying proximal gradient descent method on the above problem, we have:

$$\boldsymbol{\eta}^{r+1} = \operatorname*{argmin}_{\boldsymbol{\eta}} \mathcal{I}_{\mathbb{P}}(\boldsymbol{\eta}) + \langle \nabla J(\boldsymbol{\eta}), \boldsymbol{\eta} - \boldsymbol{\eta}^i\rangle + \frac{1}{2\nu} B(\boldsymbol{\eta}, \boldsymbol{\eta}^i) \tag{18}$$

$$= \operatorname*{argmin}_{\boldsymbol{\eta}} \sum_{i=1}^{n} \frac{\partial J(\boldsymbol{\eta}^r)}{\partial \eta_i}(\eta_i - \eta_i^r) + \frac{1}{2\nu}\Big(\sum_{i=1}^{n}\eta_i\log(\eta_i) - \sum_{i=1}^{n}\frac{\partial\gamma(\boldsymbol{\eta}_i^r)}{\partial\eta_i}(\eta_i - \eta_i^r)\Big) \tag{19}$$

By simplifying the above problem, it turns to:

$$\boldsymbol{\eta}^{r+1} = \operatorname*{argmin}_{\boldsymbol{\eta}} \sum_{i=1}^{n}\eta_i(\frac{\partial J(\boldsymbol{\eta}^r)}{\partial\eta_i} - \frac{1}{2\nu}\log(\eta_i^r) - \frac{1}{2\nu}) + \frac{1}{2\nu}\sum_{i=1}^{n}\eta_i\log(\eta_i) \tag{20}$$

$$\text{subject to} \quad \sum_{i=1}^{n}\eta_i = 1, \quad \eta_i \geq 0 \quad \forall i = 1,\dots,n. \tag{21}$$

Writing the Lagrangian function of the above problem, we have:

$$\boldsymbol{\eta}^{r+1} = \operatorname*{argmin}_{\boldsymbol{\eta}} \sum_{i=1}^{n}\eta_i(\frac{\partial J(\boldsymbol{\eta}^r)}{\partial\eta_i} - \frac{1}{2\nu}\log(\eta_i^r) - \frac{1}{2\nu}) + \frac{1}{2\nu}\sum_{i=1}^{n}\eta_i\log(\eta_i) + \lambda^*(\sum_{i=1}^{n}\eta_i - 1) \tag{22}$$
$$\text{subject to} \quad \eta_i \geq 0 \quad \forall i = 1,\dots,n.$$

By taking the derivative with respect to η_i and using the constraint $\sum_{i=1}^{n}\eta_i = 1$, it can be shown that:

$$\eta_i^{r+1} = \frac{\eta_i^r \exp(-2\nu\nabla J(\boldsymbol{\eta})_i)}{\sum_{j=1}^{n}\eta_j^r\exp(-2\nu\nabla J(\boldsymbol{\eta})_j)} \tag{23}$$

We use the update rule (23) in Algorithm 1 and Algorithm 3 to obtain the optimal $\boldsymbol{\eta}$ at each iteration.

H Proof of Theorems

In this section, we prove Theorem 1 and Theorem 2.

Proof of Theorem 1: Starting from Equation 4, we can equivalently formulate it as:

$$\min_{\mathbf{z}\in\mathcal{Z}(\mathbf{x}_0,\epsilon)} \max(\mathbf{c}_{yk}^T\mathbf{z}, \mathbf{c}_{a_1k}^T\mathbf{z}, \ldots, \mathbf{c}_{a_Mk}^T\mathbf{z}) = \min_{\mathbf{z}\in\mathcal{Z}(\mathbf{x}_0,\epsilon)} \max_{\{\eta_0,\ldots,\eta_M\}\in\mathcal{P}} \quad c_k(\boldsymbol{\eta})^T\mathbf{z}. \tag{24}$$

Note that the maximum element of the left hand side can be obtained by setting its corresponding η coefficient to 1 on the right hand side. Conversely, any optimal solution to the right hand is exactly equal to the maximum element of the left hand side. According to the min-max equality (duality), when the minimum and the maximum problems are interchanged, the following inequality holds:

$$\begin{aligned} &\min_{\mathbf{z}\in\mathcal{Z}(\mathbf{x}_0,\epsilon)} \max_{\{\eta_0,\ldots,\eta_M\}\in\mathcal{P}} \quad \eta_0\mathbf{c}_{yk}^T\mathbf{z} + \eta_1\mathbf{c}_{a_1k}^T\mathbf{z} + \cdots + \eta_M\mathbf{c}_{a_Mk}^T\mathbf{z} \geq \\ &\max_{\{\eta_0,\ldots,\eta_M\}\in\mathcal{P}} \min_{\mathbf{z}\in\mathcal{Z}(\mathbf{x}_0,\epsilon)} \quad \eta_0\mathbf{c}_{yk}^T\mathbf{z} + \eta_1\mathbf{c}_{a_1k}^T\mathbf{z} + \cdots + \eta_M\mathbf{c}_{a_Mk}^T\mathbf{z}. \end{aligned} \tag{25}$$

Moreover, by the definition of upper-bounds and lower-bounds presented in Gowal et al. [2018], $\mathcal{Z}(\mathbf{x}_0,\epsilon)$ is a subset of $\underline{\mathbf{z}}_L \leq \mathbf{z} \leq \bar{\mathbf{z}}_L$. Thus:

$$\begin{aligned} &\max_{\{\eta_0,\ldots,\eta_M\}\in\mathcal{P}} \min_{\mathbf{z}\in\mathcal{Z}(\mathbf{x}_0,\epsilon)} \quad \eta_0\mathbf{c}_{yk}^T\mathbf{z} + \eta_1\mathbf{c}_{a_1k}^T\mathbf{z} + \cdots + \eta_M\mathbf{c}_{a_Mk}^T\mathbf{z} \geq \\ &\max_{\{\eta_0,\ldots,\eta_M\}\in\mathcal{P}} \min_{\underline{\mathbf{z}}_L\leq\mathbf{z}\leq\bar{\mathbf{z}}_L} \quad \eta_0\mathbf{c}_{yk}^T\mathbf{z} + \eta_1\mathbf{c}_{a_1k}^T\mathbf{z} + \cdots + \eta_M\mathbf{c}_{a_Mk}^T\mathbf{z}. \end{aligned} \tag{26}$$

Combining Equality (24) with (25) and (26), we have:

$$\min_{\mathbf{z}\in\mathcal{Z}(\mathbf{x}_0,\epsilon)} \max(\mathbf{c}_{yk}^T\mathbf{z}, \mathbf{c}_{a_1k}^T\mathbf{z}, \ldots, \mathbf{c}_{a_Mk}^T\mathbf{z}) \geq \max_{\{\eta_0,\ldots,\eta_M\}\in\mathcal{P}} \min_{\underline{\mathbf{z}}_L\leq\mathbf{z}\leq\bar{\mathbf{z}}_L} \quad c_k(\boldsymbol{\eta})^T\mathbf{z}. \tag{27}$$

Since $\mathbf{z}_L = \mathbf{W}_L\mathbf{z}_{L-1} + \mathbf{b}_L$, the right-hand-side of the above inequality can be rewritten as:

$$\min_{\mathbf{z}\in\mathcal{Z}(\mathbf{x}_0,\epsilon)} \max(\mathbf{c}_{yk}^T\mathbf{z}, \mathbf{c}_{a_1k}^T\mathbf{z}, \ldots, \mathbf{c}_{a_Mk}^T\mathbf{z}) \geq \max_{\boldsymbol{\eta}\in\mathcal{P}} \min_{\underline{\mathbf{z}}_{L-1}\leq\mathbf{z}\leq\bar{\mathbf{z}}_{L-1}} \quad c(\boldsymbol{\eta})^T(\mathbf{W}_L\mathbf{z} + \mathbf{b}_L),$$

which is exactly the claim of Theorem 1.

Proof of Theorem 2: For the simplicity of the presentation, assume that $a_0 = y$. Partition the set of possible values of $\mathbf{z}_L$ in the following sets:

$$\hat{\mathcal{Z}}_{a_i} = \{\mathbf{z}_L | [\mathbf{z}_L]_{a_i} \geq [\mathbf{z}_L]_{a_j}\ \forall j \neq i\}$$

If $\mathbf{z}_L \in \hat{\mathcal{Z}}_{a_i}$, then:

$$\begin{aligned} [\mathbf{z}_L]_{a_i} - [\mathbf{z}_L]_k \geq [\mathbf{z}_L]_{a_j} - [\mathbf{z}_L]_k \quad \forall j \neq i \Rightarrow [\mathbf{z}_L]_{a_i} - [\mathbf{z}_L]_k \\ = \max_{i=0,\ldots,M}\{[\mathbf{z}_L]_{a_i} - [\mathbf{z}_L]_k\} = \max_{i\in\{0,\ldots,M\}}\{\mathbf{c}_{a_i,k}^T\mathbf{z}_L\} \end{aligned}$$

Thus:

$$\begin{aligned} [\mathbf{z}_L]_{a_i} - [\mathbf{z}_L]_k = \max_{i=0,\ldots,M}\{\mathbf{c}_{a_i,k}^T\mathbf{z}_L\} &\geq \min_{\mathbf{z}_L\in\hat{\mathcal{Z}}(\mathbf{x}_0,\epsilon)} \max_{i=0,\ldots,M}\{\mathbf{c}_{a_i,k}^T\mathbf{z}_L\} \\ &= \min_{\mathbf{z}_{L-1}\in\mathcal{Z}_{L-1}(\mathbf{x}_0,\epsilon)} \max_{i=0,\ldots,M}\{\mathbf{c}_{a_i,k}^T(\mathbf{W}_L\mathbf{z}_{L-1} + \mathbf{b}_L)\} \\ &\geq \min_{\underline{\mathbf{z}}\leq\mathbf{z}_{L-1}\leq\bar{\mathbf{z}}} \max_{i=0,\ldots,M}\{\mathbf{c}_{a_i,k}^T(\mathbf{W}_L\mathbf{z}_{L-1} + \mathbf{b}_L)\} \\ &= \min_{\underline{\mathbf{z}}\leq\mathbf{z}_{L-1}\leq\bar{\mathbf{z}}} \max_{\boldsymbol{\eta}\in\mathbb{P}} c(\boldsymbol{\eta})^T(\mathbf{W}_L\mathbf{z}_{L-1} + \mathbf{b}_L) \end{aligned} \tag{28}$$

Note that the second inequality holds since the minimum is taken over a larger set in the right hand side of the inequality. Using the min-max inequality:

$$\min_{\underline{\mathbf{z}} \le \mathbf{z}_{L-1} \le \bar{\mathbf{z}}} \max_{\boldsymbol{\eta} \in \mathbb{P}} c(\boldsymbol{\eta})^T (\mathbf{W}_L \mathbf{z}_{L-1} + \mathbf{b}_L) \ge \max_{\boldsymbol{\eta} \in \mathbb{P}} \min_{\underline{\mathbf{z}} \le \mathbf{z}_{L-1} \le \bar{\mathbf{z}}} c(\boldsymbol{\eta})^T (\mathbf{W}_L \mathbf{z}_{L-1} + \mathbf{b}_L) = -J_k(\boldsymbol{\eta}) \tag{29}$$

Combining (28) and (29), and multiplying both sides by -1, we obtain:

$$[\mathbf{z}_L]_k - [\mathbf{z}_L]_{a_i} \le J_k(\boldsymbol{\eta}) \tag{30}$$

On the other hand:

$$\begin{aligned}
&\max_{\|\boldsymbol{\delta}\|_\infty \le \epsilon} \min_{m=0,\dots,M} \ell_{\text{xent}\backslash\mathcal{A}_m}\Big(\mathbf{z}_L(\mathbf{x}+\delta), a_m\Big) \\
&\le \max_{\|\boldsymbol{\delta}\|_\infty \le \epsilon} \ell_{\text{xent}\backslash\mathcal{A}_i}\Big(\mathbf{z}_L(\mathbf{x}+\delta), a_i\Big) \\
&\le \max_{\underline{\mathbf{z}}_{L-1} \le \mathbf{z} \le \bar{\mathbf{z}}_{L-1}} \ell_{\text{xent}\backslash\mathcal{A}_i}(\mathbf{z}_L) \quad \text{s.t.} \quad \mathbf{z}_L = \mathbf{W}_L \mathbf{z}_{L-1} + \mathbf{b}_L.
\end{aligned} \tag{31}$$

Moreover, by the property of the cross-entropy loss, we have:

$$\ell_{\text{xent}\backslash\mathcal{A}_i}(\mathbf{z}_L) = \ell_{\text{xent}\backslash\mathcal{A}_i}(\mathbf{z}_L - [\mathbf{z}_L]_{a_i}\mathbf{1}) \tag{32}$$

Combining (30), (31) and (32), we have:

$$\begin{aligned}
&\max_{\|\boldsymbol{\delta}\|_\infty \le \epsilon} \min_{m=0,\dots,M} \ell_{\text{xent}\backslash\mathcal{A}_m}\Big(\mathbf{z}_L(\mathbf{x}+\delta), a_m\Big) \\
&\le \max_{\underline{\mathbf{z}}_{L-1} \le \mathbf{z}_{L-1} \le \bar{\mathbf{z}}_{L-1}} \ell_{\text{xent}\backslash\mathcal{A}_i}(\mathbf{z}_L) \quad \text{s.t.} \quad \mathbf{z}_L = \mathbf{W}_L \mathbf{z}_{L-1} + \mathbf{b}_L. \\
&= \max_{\underline{\mathbf{z}}_{L-1} \le \mathbf{z}_{L-1} \le \bar{\mathbf{z}}_{L-1}} \ell_{\text{xent}\backslash\mathcal{A}_i}(\mathbf{z}_L - [\mathbf{z}_L]_{a_i}\mathbf{1}) \quad \text{s.t.} \quad \mathbf{z}_L = \mathbf{W}_L \mathbf{z}_{L-1} \\
&\le \max_{\underline{\mathbf{z}}_{L-1} \le \mathbf{z}_{L-1} \le \bar{\mathbf{z}}_{L-1}} \ell_{\text{xent}\backslash\mathcal{A}_i}(J_k(\boldsymbol{\eta}), a_i) \\
&= \max_{\underline{\mathbf{z}}_{L-1} \le \mathbf{z}_{L-1} \le \bar{\mathbf{z}}_{L-1}} \ell_{\text{xent}\backslash\mathcal{A}_0}(J_k(\boldsymbol{\eta}), a_0)
\end{aligned}$$

Summing up over all data points, the desired result is proven.

I Details of β-Crown

In this section, we show how β-crown sub-problems can be obtained for neural networks without abstain classes and with multiple abstain classes respectively. Before proceeding, let us have a few definitions and lemmas.

Lemma 3 ***[Zhang et al., 2019, Theorem 15]*** *Given two vectors* $\mathbf{u}$ *and* $\mathbf{v}$*, the following inequality holds:*

$$\mathbf{v}^\top \mathit{ReLU}(\mathbf{u}) \ge \mathbf{v}^\top \mathbf{D}_{\boldsymbol{\alpha}} \mathbf{u} + \mathbf{b}',$$

where $\mathbf{b}'$ *is a constant vector and* $\mathbf{D}_{\boldsymbol{\alpha}}$ *is a diagonal matrix containing* $\boldsymbol{\alpha}_j$*'s as free parameters:*

$$\mathbf{D}_{j,j}(\boldsymbol{\alpha}) = \begin{cases} 1, & \textit{if } \underline{\mathbf{z}}_j \ge 0 \\ 0, & \textit{if } \bar{\mathbf{z}}_j \le 0 \\ \boldsymbol{\alpha}_j, & \textit{if } \bar{\mathbf{z}}_j > 0 > \underline{\mathbf{z}}_j \textit{ and } \mathbf{v}_j \ge 0 \\ \frac{\bar{\mathbf{z}}_j}{\bar{\mathbf{z}}_j - \underline{\mathbf{z}}_j}, & \textit{if } \bar{\mathbf{z}}_j > 0 > \underline{\mathbf{z}}_j \textit{ and } \mathbf{v}_j < 0, \end{cases} \tag{33}$$

Definition 4 *The recursive function* $\Omega(i,j)$ *is defined as follows [Wang et al., 2021]:*

$$\Omega(i,i) = \boldsymbol{I}, \quad \Omega(i,j) = \boldsymbol{W}_i \boldsymbol{D}_{i-1}(\boldsymbol{\alpha}_{i-1}) \Omega(i-1,j)$$

β-crown defines a matrix $\mathbf{S}$ for handling splits through the branch-and-bound process. The multiplier(s) $\boldsymbol{\beta}$ determines the branching rule.

$$\mathbf{S}_i[j][j] = \begin{cases} -1, & \text{if split } \mathbf{z}_i[j] \geq 0 \\ 1, & \text{if split } \mathbf{z}_i[j] < 0 \\ 0, & \text{if no split } \bar{\mathbf{z}}_j, \end{cases} \tag{34}$$

Thus, the verification problem of β-crown is formulated as:

$$\min_{\mathbf{z} in \mathcal{Z}} \mathbf{c}^T\big(\mathbf{W}^L \text{ReLU}(\mathbf{z}_{L-1}) + \mathbf{b}_{L-1}\big) \geq \min_{\mathbf{z} in \mathcal{Z}} \quad \max_{\boldsymbol{\beta}_{L-1}} \mathbf{c}^T\big(\mathbf{W}^L \mathbf{D}_{L-1}\mathbf{z}_{L-1} + \mathbf{b}_{L-1}\big) + \boldsymbol{\beta}_{L-1}^\top \mathbf{S}_{L-1} \tag{35}$$

Having these definitions, we can write $\mathbf{P}, \mathbf{q}, \mathbf{a}$, and $\mathbf{d}$ explicitly as functions of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. $\mathbf{P} \in \mathbb{R}^{d_0 \times (\sum_{i=1}^{L-1} d_i)}$ is a block matrix $\mathbf{P} := \begin{bmatrix}\mathbf{P}_1^\top \ \mathbf{P}_2^\top \ \cdots \ \mathbf{P}_{L-1}^\top\end{bmatrix}$, $\mathbf{q} \in \mathbb{R}^{\sum_{i=1}^{L-1} d_i}$ is a vector $\mathbf{q} := \begin{bmatrix}\mathbf{q}_1^\top \ \cdots \ \mathbf{q}_{L-1}^\top\end{bmatrix}^\top$. Moreover:

$$\mathbf{a} = [\Omega(L,1)\mathbf{W}_1]^\top \in \mathbb{R}^{d_0 \times 1},$$

$$\mathbf{P}_i = \mathbf{S}_i \Omega(i,1)\mathbf{W}_1 \in \mathbb{R}^{d_i \times d_0}, \quad \forall\, 1 \leq i \leq L-1$$

$$\mathbf{q}_i = \sum_{k=1}^{i} \mathbf{S}_i \Omega(i,k)\mathbf{b}_k + \sum_{k=2}^{i} \mathbf{S}_i \Omega(i,k)\mathbf{W}_k \underline{\mathbf{b}}_{k-1} \in \mathbb{R}^{d_i}, \quad \forall\, 1 \leq i \leq L-1$$

$$\mathbf{d} = \sum_{i=1}^{L} \Omega(L,i)\mathbf{b}_i + \sum_{i=2}^{L} \Omega(L,i)\mathbf{W}_i \underline{b}_{i-1}$$

$$\underline{b}_i = \begin{cases} 1, & \text{if } \underline{\mathbf{z}}_j \geq 0 \\ 0, & \text{if } \bar{\mathbf{z}}_j \leq 0 \\ \boldsymbol{\alpha}_j, & \text{if } \bar{\mathbf{z}}_j > 0 > \underline{\mathbf{z}}_j \text{ and } \mathbf{v}_j \geq 0 \\ \frac{\bar{\mathbf{z}}_j}{\bar{\mathbf{z}}_j - \underline{\mathbf{z}}_j}, & \text{if } \bar{\mathbf{z}}_j > 0 > \underline{\mathbf{z}}_j \text{ and } \mathbf{v}_j < 0, \end{cases}$$

Now we extend the definition of g for the network consisting of multiple abstain classes. Let $\bar{\mathbf{z}}$ be the pre-activation value of vector z before applying ReLU function. We aim to solve the following verification problem:

$$\min_{\mathbf{z}_{L-1} \in \mathcal{Z}_{L-1}(\mathbf{x}_0, \epsilon)} \quad \max_{\boldsymbol{\eta} \in \mathcal{P}} c_k(\boldsymbol{\eta})^T(\mathbf{W}_L \mathbf{z}_{L-1} + \mathbf{b}_L).$$

Applying Lemma 3 to the above problem, we have:

$$\begin{aligned} & \min_{\mathbf{z}_{L-1} \in \hat{\mathcal{Z}}_{L-1}(\mathbf{x}_0, \epsilon)} \quad \max_{\boldsymbol{\eta} \in \mathcal{P}} c_k(\boldsymbol{\eta})^T \Big(\mathbf{W}_L \mathbf{z}_{L-1} + \mathbf{b}_L\Big) \\ \leq & \min_{\mathbf{z}_{L-1} \in \hat{\mathcal{Z}}_{L-1}(\mathbf{x}_0, \epsilon)} \quad \max_{\boldsymbol{\eta} \in \mathcal{P}} c_k(\boldsymbol{\eta})^T \Big(\mathbf{W}_L \mathbf{D}_{L-1}\big(\boldsymbol{\alpha}_{L-1}\big)\hat{\mathbf{z}}_{L-1} + \mathbf{b}_L\Big) \end{aligned}$$

Adding the β-crown Lagrangian multiplier to the above problem, it turns to:

$$\min_{\mathbf{z}_{L-1}\in\mathcal{Z}_{L-1}(\mathbf{x}_0,\epsilon)} \quad \max_{\boldsymbol{\eta}\in\mathcal{P}} \quad c_k(\boldsymbol{\eta})^T\Big(\mathbf{W}_L\mathbf{D}_{L-1}\big(\boldsymbol{\alpha}_{L-1}\big)\hat{\mathbf{z}}_{L-1}+\mathbf{b}_L\Big) \leq$$

$$\min_{\mathbf{z}_{L-1}\in\mathcal{Z}_{L-1}(\mathbf{x}_0,\epsilon)} \quad \max_{\boldsymbol{\eta}\in\mathcal{P},\boldsymbol{\alpha}_{L-1},\boldsymbol{\beta}_{L-1}} \quad c_k(\boldsymbol{\eta})^T\Big(\mathbf{W}_L\mathbf{D}_{L-1}(\boldsymbol{\alpha}_{L-1})\mathbf{z}_{L-1}+\mathbf{b}_L\Big)+\boldsymbol{\beta}_{L-1}^\top\mathbf{S}_{L-1}\mathbf{z}_{L-1}$$

$$\leq \max_{\boldsymbol{\alpha}_{L-1},\boldsymbol{\beta}_{L-1}} \min_{\mathbf{z}_{L-1}\in\mathcal{Z}_{L-1}(\mathbf{x}_0,\epsilon)} \quad \max_{\boldsymbol{\eta}\in\mathcal{P}} \Big(c_k(\boldsymbol{\eta})^T\mathbf{W}_L\mathbf{D}_{L-1}(\boldsymbol{\alpha}_{L-1})+\boldsymbol{\beta}_{L-1}^\top\mathbf{S}_{L-1}\Big)\hat{\mathbf{z}}_{L-1}$$

$$+c_k(\boldsymbol{\eta})^T\mathbf{b}_L = \max_{\boldsymbol{\alpha}_{L-1},\boldsymbol{\beta}_{L-1}} \min_{\mathbf{z}_{L-1}\in\mathcal{Z}_{L-1}(\mathbf{x}_0,\epsilon)} \quad \max_{\boldsymbol{\eta}\in\mathcal{P}} \Big(c_k(\boldsymbol{\eta})^T\mathbf{W}_L\mathbf{D}_{L-1}(\boldsymbol{\alpha}_{L-1})$$

$$+\boldsymbol{\beta}_{L-1}^\top\mathbf{S}_{L-1}\Big)\Big(\mathbf{W}_{L-1}\mathbf{z}_{L-2}+\mathbf{b}_{L-1}\Big)+c_k(\boldsymbol{\eta})^T\mathbf{b}_L$$

Replace the definition of $\mathbf{A}^{(i)}$ in [Wang et al., 2021, Theorem 3.1] with the following matrix and repeat the proof.

$$\mathbf{A}^{(i)} = \begin{cases} c_k(\boldsymbol{\eta})^T\mathbf{W}_L, & \text{if} \quad i = L-1 \\ \Big(\mathbf{A}^{(i+1)}\mathbf{D}_{i+1}(\boldsymbol{\alpha}_{i+1})+\boldsymbol{\beta}_{i+1}^\top\mathbf{S}_{i+1}\Big)\mathbf{W}_{i+1}, & \text{if} \quad 0\leq i\leq L-2 \end{cases} \tag{36}$$

Note that the definition of $\mathbf{d}$ will be changed in the following way:

$$\mathbf{d} = c_k(\boldsymbol{\eta})^T\mathbf{b}_L+\sum_{i=1}^{L}\Omega(L,i)\mathbf{b}_i+\sum_{i=2}^{L}\Omega(L,i)\mathbf{W}_i\underline{b}_{i-1}$$

Moreover, $\Omega(L,j) = c_k(\boldsymbol{\eta})^T\boldsymbol{W}_L\boldsymbol{D}_{L-1}(\boldsymbol{\alpha}_{L-1})\Omega(L-1,j)$. The rest of the definitions remain the same.

J Derivation of equation (15)

In this section, we show how to derive Equation J.

$$\begin{aligned}
&\min_{\mathbf{z}_L\in\mathcal{Z}(\mathbf{x},\epsilon)} \quad \max\{\mathbf{c}_{yk}^T\mathbf{z}_L,\mathbf{c}_{a_1k}^T\mathbf{z}_L,\ldots,\mathbf{c}_{a_Mk}^T\mathbf{z}_L\} \\
&= \min_{\mathbf{z}_L\in\mathcal{Z}(\mathbf{x},\epsilon)} \quad \max_{\boldsymbol{\eta}\in\mathcal{P}}\sum_{i=0}^{M}\eta_i c_{a_ik}^T\mathbf{z}_L \\
&\geq \max_{\boldsymbol{\eta}\in\mathcal{P}} \quad \min_{\mathbf{z}_L\in\mathcal{Z}(\mathbf{x},\epsilon)}\sum_{i=0}^{M}\eta_i c_{a_ik}^T\mathbf{z}_L \\
&\geq \max_{\boldsymbol{\eta}\in\mathcal{P}} \quad \max_{\boldsymbol{\alpha},\boldsymbol{\beta}\geq 0}\ \eta_i c_{a_ik}^T\mathbf{z}_L \\
&= \max_{\boldsymbol{\alpha},\boldsymbol{\beta}\geq 0,\boldsymbol{\eta}\in\mathcal{P}}\Big(\sum_{i=0}^{M}\eta_i g_i(\mathbf{x}_0,\boldsymbol{\alpha},\boldsymbol{\beta}) \triangleq G(\mathbf{x}_0,\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\eta})\Big)
\end{aligned}$$

K A simple example on the benefits and pitfalls of having multiple abstain classes

In this example, we provide a simple toy example illustrating:

1. How adding multiple abstain classes can improve the detection of adversarial examples.
2. How detection with multiple abstain classes may suffer from a “model degeneracy" phenomenon.

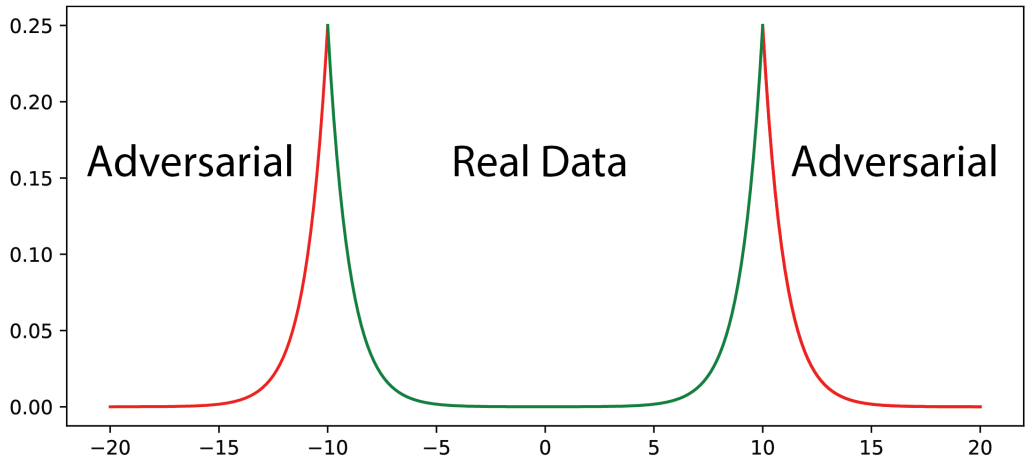

Figure 6: Distribution of adversarial and real data described in the example. While one linear classifier cannot separate the adversarial (red section) and real (green section) data points, two detection classes are capable of detecting adversarial examples.

Example: Consider a simple one dimensional data distributed where the read data is coming from the Laplacian distribution with probability density function $P_r(X = x) = \frac{1}{2}\exp(-|x|)$. Assume that the adversary samples are distributed according to the probability density function $P_a(X = x) = \frac{1}{4}(\exp(-|x - 10|) + \exp(-|x + 10|)$. Assume that $\frac{1}{3}$ data is real, and $\frac{2}{3}$ is coming from adversary. The adversary and the real data is illustrated in Fig 6.

Consider a binary neural network classifier with no hidden layer for detecting adversaries. More specifically, the neural network has two weight vectors w^r and w^a, and the bias values b^r and b^a. The network classifies a sample $\mathbf{x}$ as "real" if $w^r x + b^r > w^a x + b^a$; otherwise, it classifies the sample as out-of-distribution/abstain. The misclassification rate of this classifier is given by:

$$\begin{aligned} P(\text{error}) &= \frac{1}{3}P_{x\sim P_r}(w^a x + b^a > w^r x + b^r) + \frac{2}{3}P_{x\sim P_a}(w^a x + b^a < w^r x + b^r) \\ &= \frac{1}{3}P_{x\sim P_r}(x > \frac{b^r - b^a}{w^a - w^r}) + \frac{2}{3}P_{x\sim P_a}(x < \frac{b^r - b^a}{w^a - w^r}), \end{aligned}$$

where due to symmetry and scaling invariant, without loss of generality we assumed that $w^a - w^r > 0$. Let $t = \frac{b^r - b^a}{w^a - w^r}$. Therefore,

$$P(\text{error}) = \frac{1}{3}\int_t^{+\infty} \frac{1}{2}\exp(-|x|)dx + \frac{2}{3}\int_{-\infty}^{t} \frac{1}{4}(\exp(-|x - 10|) + \exp(-|x + 10|)dx \tag{37}$$

Thus, to find the optimal classifier, we require to determine the optimal t minimizing the above equation. One can numerically verify that the optimal t is given by $t^* = 5$ leading to the minimum misclassification rate of ≈ 0.34. This value is the optimal misclassification rate that can be achieved by our single abstain class neural network.

Now consider a neural network with two abstain classes. Assume that the weights and biases corresponding to the abstain classes are $w_1^a, w_2^a, b_1^a, b_2^a$, and the weight and bias for the real class is given by w_r and b_r. A sample x is classified as a real example if and only if both of the following conditions hold:

$$w^r x + b_r > w_1^a x + b_1^a \tag{38}$$
$$w^r x + b_r > w_2^a x + b_2^a, \tag{39}$$

otherwise, it is classified as an adversarial (out of distribution) sample. The misclassification rate of such classifier is given by:

$$P(\text{error}) = \frac{1}{3}P_{x\sim P_c}(\text{Conditions (38) hold}) + \frac{2}{3}P_{x\sim P_a}(\text{Conditions (38) do not hold}) \tag{40}$$

Claim 1: The point $w_1^a = -1, w_2^a = 1, b_1^a = b_2^a = 0, b^r = 5, w^r = 0$ is a global minimum of (40) with the optimum misclassification rate less than 0.1.

Proof: Define $t_1 = -\frac{b_1^a - b_r}{w_1^a - w^r}, t_2 = -\frac{b_2^a - b_r}{w_2^a - w^r}$. Considering all possible sign cases, it is not hard to see that at the optimal point, $w_1^a - w^r$ and $w_2^a - w^r$ have different signs. Without loss of generality, assume that $w_1^a - w^r < 0$ and $w_2^a - w^r > 0$. Then:

$$P(\text{error}) = \frac{1}{3} P_{x \sim P_c}(x \leq t_1 \vee x \geq t_2) + \frac{2}{3} P_{x \sim P_a}(x \geq t_1 \wedge x \leq t_2) \tag{41}$$

It is not hard to see that the optimal solution is given by $t_1^* = -5$, $t_2^* = 5$. Plugging these values in above equation, we can check that the optimal loss is less than 0.1. ■

Claim 1 shows that by adding an abstain class, the misclassification rate of the classifier goes down from 0.34 to below 0.1. This simple example illustrates the benefit of having multiple abstain classes. Next, we show that by having multiple abstain classes, we are prone to the "model degeneracy" phenomenon.

Claim 2: Let $\bar{w}_1^a = \bar{w}_2^a = 1, \bar{b}_1^a = \bar{b}_2^a = 0, \bar{w}^r = 0, \bar{b}^r = 5$. Then, there exists a point $(\tilde{w}, \tilde{b}) = (\tilde{w}_1^a, \tilde{w}_2^a, \tilde{b}_1^a, \tilde{b}_2^a, \tilde{w}^r, \tilde{b}^r)$ such that $(\tilde{w}, \tilde{b})$ is a local minimum of the loss function in (40) and $\|(\tilde{w}, \tilde{b}) - (\bar{w}, \bar{b})\|_2 \leq 0.1$.

Proof: Let $t_1 = -\frac{b_1^a - b_r}{w_1^a - w^r}, t_2 = -\frac{b_2^a - b_r}{w_2^a - w^r}$. Notice that in a neighborhood of point $(\bar{w}, \bar{b})$, we have $w_1^a - w^r > 0$ and $w_2^a - w^r > 0$. Thus, after the loss function in (40) can be written as:

$$\begin{aligned}
\ell(t_1, t_2) &= \frac{1}{3} P_{x \sim P_c}(x \leq t_1 \vee x \geq t_2) + \frac{2}{3} P_{x \sim P_a}(x \geq t_1 \wedge x \leq t_2) \\
&= \frac{1}{3} P_{x \sim P_r}(x \geq \min(t_1, t_2)) + \frac{2}{3} P_{x \sim P_r}(x \leq \min(t_1, t_2)) \\
&= \frac{1}{3} P_{x \sim P_r}(x \geq z) + \frac{2}{3} P_{x \sim P_r}(x \leq z),
\end{aligned}$$

where $z = \min_{t_1, t_2}$. It suffices to show that the above function has a local minimum close to the point $\bar{z} = 5$ (see [Nouiehed and Razaviyayn, 2021]). Simplifying $\ell(t_1, t_2)$ as a function of z, we have:

$$\ell(t_1, t_2) = h(z) = \frac{1}{6} \exp(-z) + \frac{1}{3} - \frac{1}{6} \exp(-z - 10) + \frac{1}{6} \exp(z - 10)$$

By plotting $h(z)$, we can observe that it has a local minimum close to $\bar{z} = 5$. ■

This claim shows that by optimizing the loss, we may converge to the local optimum $(\tilde{w}, \tilde{b})$ where both abstain classes become essentially the same and we do not utilize the two abstain classes fully.

L Structure of Neural Networks in Section 3

In Section 3 we introduced a toy example in Motivation subsection to show how loser can IBP bounds become when we go from a 2-layer network to a 3-layer network. The structure of the 2-layer neural networks is as follows:

$$\mathbf{z}_2(\mathbf{x}) = \mathbf{V}_2 \text{ReLU}(\mathbf{W}_2 \mathbf{x}),$$

where $\mathbf{x}$ is the 2-dimensional input, $\mathbf{W}_2 = \begin{pmatrix} 1 & -0.5 \\ -0.8 & 1.2 \end{pmatrix}$, and $\mathbf{V}_2 = \begin{pmatrix} -0.2 & -0.8 \\ -1.9 & 1.7 \end{pmatrix}$. The structure of the 3-layer network can be described as follows:

$$\mathbf{z}_3(\mathbf{x}) = \mathbf{U}_3 \text{ReLU}\Big(\mathbf{V}_2 \text{ReLU}(\mathbf{W}_2 \mathbf{x})\Big),$$

where $\mathbf{U}_3 = \begin{pmatrix} -1.1 & -0.9 \\ -1.6 & 1.3 \end{pmatrix}$.

M Societal Impacts

Since current trained neural networks are highly vulnerable to adversarial examples and out-of-distribution samples, it is debatable whether to use such models in mission-critical applications

such as self-driving cars. To address neural networks' safety and reliability concerns, it is crucial to devise mechanisms guaranteeing the robustness of the trained models in uncertain and adversarial environments. The current work proposes a rigorous methodology for training and verifying neural networks against adversarial attacks. From a broader perspective, verifiable guarantees for the performance of artificial intelligence (AI) models reduce the ethical and safety concerns existing around AI systems.