ROBUST AND ACCURATE – COMPOSITIONAL ARCHITECTURES FOR RANDOMIZED SMOOTHING

Miklós Z. Horváth, Mark Niklas Müller, Marc Fischer, Martin Vechev
Department of Computer Science
ETH Zurich, Switzerland
mihorvat@ethz.ch, {mark.mueller,marc.fischer,martin.vechev}@inf.ethz.ch

ABSTRACT

Randomized Smoothing (RS) is considered the state-of-the-art approach to obtain certifiably robust models for challenging tasks. However, current RS approaches drastically decrease standard accuracy on unperturbed data, severely limiting their real-world utility. To address this limitation, we propose a compositional architecture, ACES, which certifiably decides on a per-sample basis whether to use a smoothed model yielding predictions with guarantees or a more accurate standard model without guarantees. This, in contrast to prior approaches, enables both high standard accuracies *and* significant provable robustness. On challenging tasks such as ImageNet, we obtain, e.g., 80.0% natural accuracy and 28.2% certifiable accuracy against ℓ_2 perturbations with $r = 1.0$. We release our code and models at https://github.com/eth-sri/aces.

1 INTRODUCTION

Since the discovery of imperceptible input perturbations that can fool machine learning models, called adversarial examples (Biggio et al., 2013; Szegedy et al., 2014), certifying model robustness has been identified as an essential task to enable their application in safety-critical domains.

Various works have discussed the fundamental trade-off between robustness and accuracy in the empirical setting (Raghunathan et al., 2019; Tsipras et al., 2019; Zhang et al., 2019). However, in the setting of deterministically certified robustness, this Pareto frontier has only recently been explored (Müller et al., 2021). There, due to the poor scaling of deterministic methods to large networks, performance on more challenging tasks is severely limited. In the probabilistic certification setting, recent works aim to jointly increase robustness and accuracy by choosing smoothing parameters per sample (Alfarra et al., 2020), however often at the cost of statistical soundness (Súkeník et al., 2021).

In this work, we build on ideas from Müller et al. (2021) to construct compositional architectures for probabilistic certification and propose corresponding statistically sound and efficient inference and certification procedures based on randomized smoothing (Cohen et al., 2019). More concretely, we propose to use a smoothed selection-mechanism that adaptively chooses on a per-sample basis between a robustified smoothed classifier and a non-robust but highly accurate classifier. We show that the synergy of RS with the proposed compositional architecture allows us to obtain significant robustness at almost no cost in terms of natural accuracy even on challenging datasets such as ImageNet while fully exposing this robustness-accuracy trade-off, even after training.

Main Contributions Our key contributions are:

- We are first to extend compositional architectures to the probabilistic certification setting, combining an arbitrary deep model with a smoothed classifier and selection-mechanism.
- We investigate two selection-mechanisms for choosing, at inference time and on a per-sample basis, between a robust and an accurate classifier and derive corresponding statistically sound prediction and certification algorithms.
- We conduct an extensive empirical investigation of our compositional architectures on ImageNet and CIFAR10 and find that they achieve significantly more attractive trade-offs between robustness and accuracy than any current method. On ImageNet, we, e.g., achieve 15.8% more natural accuracy at the same ACR or 0.14 more ACR at the same natural accuracy.

2 BACKGROUND & RELATED WORK

In this section, we review related work and relevant background.

Adversarial Robustness & Threat Model Let $\boldsymbol{f}\colon \mathbb{R}^d \mapsto \mathbb{R}^m$ be a classifier computing an m-dimensional logit vector, assigning a numerical score to each of the m classes, given a d-dimensional input. Additionally, let $F(\boldsymbol{x}) := \arg\max_i f(\boldsymbol{x})_i$ with $F : \mathbb{R}^d \mapsto [1, \dots, m]$ be the function that outputs the class with the largest score. On a given input $\boldsymbol{x}$ with label y, we say F is (accurately) adversarially robust if it classifies all inputs in a p-norm ball $B^p_\delta(\boldsymbol{x})$ of radius δ around the sample $\boldsymbol{x}$ correctly: $F(\boldsymbol{x}) = F(\boldsymbol{x}') = y, \forall \boldsymbol{x}' \in B^p_\delta(\boldsymbol{x})$. We distinguish between empirical and certified robustness. Empirical robustness is computed by trying to find a counterexample $\boldsymbol{x}' \in B^p_\delta(\boldsymbol{x})$ such that $F(\boldsymbol{x}') \neq F(\boldsymbol{x})$; it constitutes an upper bound to the true robust accuracy. Certified robustness, in contrast, constitutes a sound lower bound. We further distinguish probabilistic and deterministic certification: Deterministic methods compute the reachable set for given input specifications (Katz et al., 2017; Gehr et al., 2018; Raghunathan et al., 2018; Zhang et al., 2018; Singh et al., 2019) to then reason about the output. While providing state-of-the-art guarantees for ℓ_∞ specifications, these methods are computationally expensive and typically limited to small networks. Probabilistic methods (Li et al., 2019; Lécuyer et al., 2019; Cohen et al., 2019) construct a robustified classifier and obtain probabilistic robustness guarantees by introducing noise into the classification process, allowing the certification of much larger models. In this work, we focus on probabilistic certification and an ℓ_2-norm based threat model. Extensions to other threat models are orthogonal to our approach.

Randomized Smoothing Randomized Smoothing (RS) (Cohen et al., 2019) is one of the most popular probabilistic certification methods. The key idea is to generate many randomly perturbed instances of the same sample and to then conduct majority voting over the predictions on these perturbed samples. More concretely, Randomized Smoothing constructs the smoothed classifier $\bar{F}\colon \mathbb{R}^d \mapsto [1, \dots, m]$ by conducting majority voting over a random noise term $\epsilon \sim \mathcal{N}(0, \sigma^2_\epsilon \boldsymbol{I})$:

$$\bar{F}(\boldsymbol{x}) := \arg\max_c \mathbb{E}_{\epsilon \sim \mathcal{N}(0, \sigma^2_\epsilon \boldsymbol{I})}(F(\boldsymbol{x} + \epsilon) = c). \tag{1}$$

For this smoothed classifier $\bar{F}$, we obtain the following robustness guarantee:

Theorem 2.1. (Cohen et al. (2019)). *Let $c_A \in [1, \dots, m]$, $\epsilon \sim \mathcal{N}(0, \sigma^2_\epsilon \boldsymbol{I})$, and $\underline{p_A}, \overline{p_B} \in [0, 1]$. If*

$$\mathcal{P}_\epsilon(F(\boldsymbol{x} + \epsilon) = c_A) \geq \underline{p_A} \geq \overline{p_B} \geq \max_{c \neq c_A} \mathcal{P}_\epsilon(F(\boldsymbol{x} + \epsilon) = c), \tag{2}$$

then $\bar{F}(\boldsymbol{x} + \delta) = c_A$ for all δ satisfying $\|\delta\|_2 < R$ with $R := \frac{\sigma_\epsilon}{2}(\Phi^{-1}(\underline{p_A}) - \Phi^{-1}(\overline{p_B}))$.

Where Φ^{-1} is the inverse Gaussian CDF. The expectation and probabilities in Eqs. (1) and (2), respectively, are computationally intractable. Hence, Cohen et al. (2019) propose to bound them using Monte Carlo sampling and the Clopper-Pearson lemma (Clopper and Pearson, 1934). We denote obtaining a class c_A and radius R fulfilling Theorem 2.1 as *certification* and just obtaining the class as *prediction*. In practice, both are computed with confidence $1 - \alpha$. When this fails, we abstain from making a classification, denoted as $\oslash$. Performance is typically measured in certified accuracy at radius r ($R \geq r$) and average certified radius over samples (ACR). We focus on their trade-off with natural accuracy (NAC) and provide detailed algorithms and descriptions in App. A.

Trade-Off For both empirical and certified methods, it has been shown that there is a trade-off between model accuracy and robustness (Zhang et al., 2019; Xie et al., 2020; Raghunathan et al., 2019; Tsipras et al., 2019). In the case of RS, the parameter σ_ϵ provides a natural way to trade-off certificate strength and natural accuracy (Cohen et al., 2019; Mohapatra et al., 2021).

Compositional Architectures For Deterministic Certification (ACE) To enable efficient robustness-accuracy trade-offs for deterministic certification, Müller et al. (2021) introduced a compositional architecture. The main idea of their ACE architecture is to use a selection model to certifiably predict certification-difficulty, and depending on this, either classify using a model with high certified accuracy, $F_{\text{Certify}} : \mathbb{R}^d \mapsto [1, \dots, m]$, or a model with high natural accuracy, $F_{\text{Core}} : \mathbb{R}^d \mapsto [1, \dots, m]$. Overall, the ACE architecture $F_{\text{ACE}} : \mathbb{R}^d \mapsto [1, \dots, m]$ is defined as

$$F_{\text{ACE}}(\boldsymbol{x}) = F_{\text{Select}}(\boldsymbol{x}) \cdot F_{\text{Certify}}(\boldsymbol{x}) + (1 - F_{\text{Select}}(\boldsymbol{x})) \cdot F_{\text{Core}}(\boldsymbol{x}). \tag{3}$$

Müller et al. (2021) propose two instantiations for the selection-mechanism, $F_{\text{Select}} : \mathbb{R}^d \mapsto \{0, 1\}$: a learned binary classifier and a mechanism selecting F_{Certify} if and only if the entropy of its output is below a certain threshold. In order to obtain a certificate, both F_{Certify} and F_{Select} must be certified.

3 Robustness vs. Accuracy Trade-Off via Randomized Smoothing

Here, we introduce ACES which instantiates ACE (Eq. (3)) with Randomized Smoothing by replacing F_{Select} and F_{Certify} with their smoothed counterparts $\bar{F}_{\text{Select}}$ and $\bar{F}_{\text{Certify}}$, respectively:

$$F_{\text{ACES}}(\boldsymbol{x}) = \bar{F}_{\text{Select}}(\boldsymbol{x}) \cdot \bar{F}_{\text{Certify}}(\boldsymbol{x}) + (1 - \bar{F}_{\text{Select}}(\boldsymbol{x})) \cdot F_{\text{Core}}(\boldsymbol{x}). \tag{4}$$

Note that, due to the high cost of certification and inference of smoothed models, instantiating F_{Core} with significantly larger models than F_{Certify} and F_{Select} comes at a negligible computational cost.

Prediction & Certification Just like other smoothed models (Eq. (1)), ACES (Eq. (4)) can usually not be evaluated exactly in practice but has to be approximated via sampling and confidence bounds. We thus propose CERTIFY (shown in Algorithm 1) to soundly compute the output $F_{\text{ACES}}(\boldsymbol{x})$ and its robustness radius R. Here, SAMPLEWNOISE$(f, \boldsymbol{x}, n, \sigma_\epsilon)$ evaluates n samples of $f(\boldsymbol{x} + \epsilon)$ for $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon \boldsymbol{I})$, and LOWERCONFBND$(m, n, c)$ computes a lower bound to the success probability p for obtaining m successes in n Bernoulli trials with confidence c. Conceptually, we apply the CERTIFY procedure introduced in Cohen et al. (2019) twice, once for $\bar{F}_{\text{Select}}$ and once for $\bar{F}_{\text{Certify}}$. If $\bar{F}_{\text{Select}}$ certifiably selects the certification model, we evaluate $\bar{F}_{\text{Certify}}$ and return its prediction $\hat{c}_A$ along with the minimum certified robustness radius of $\bar{F}_{\text{Select}}$ and $\bar{F}_{\text{Certify}}$. If $\bar{F}_{\text{Select}}$ certifiably selects the core model, we directly return its classification $F_{\text{Core}}(\boldsymbol{x})$ and no certificate ($R = 0$). If $\bar{F}_{\text{Select}}$ does not certifiably select either model, we either return the class that the core and certification model agree on or abstain ($\oslash$). A robustness radius R obtained this way holds with confidence $1 - \alpha$ (Theorem B.1 in App. B). Note that individual tests need to be conducted with $1 - \frac{\alpha}{2}$ to account for multiple testing (Bonferroni, 1936). Please see App. B for a further discussion and PREDICT, an algorithm computing $F_{\text{ACES}}(\boldsymbol{x})$ but not R at a lower computational cost.

Algorithm 1 Certification for ACES

function CERTIFY$(\sigma_\epsilon, \boldsymbol{x}, n_0, n, \alpha)$
- $\texttt{counts}_S^0 \leftarrow$ SAMPLEWNOISE$(F_{\text{Select}}, \boldsymbol{x}, n_0, \sigma_\epsilon)$
- $\texttt{counts}_C^0 \leftarrow$ SAMPLEWNOISE$(F_{\text{Certify}}, \boldsymbol{x}, n_0, \sigma_\epsilon)$
- $\hat{s} \leftarrow 0$ **if** $\texttt{counts}_S^0[0] > \texttt{counts}_S^0[1]$ **else** 1
- $\hat{c}_A \leftarrow$ top indices in $\texttt{counts}_C^0$
- $\texttt{counts}_S \leftarrow$ SAMPLEWNOISE$(F_{\text{Select}}, \boldsymbol{x}, n, \sigma_\epsilon)$
- $\texttt{counts}_C \leftarrow$ SAMPLEWNOISE$(F_{\text{Certify}}, \boldsymbol{x}, n, \sigma_\epsilon)$
- $\underline{p_S} \leftarrow$ LOWERCONFBND$(\texttt{counts}_S[\hat{s}], n, 1 - \frac{\alpha}{2})$
- $\underline{p_A} \leftarrow$ LOWERCONFBND$(\texttt{counts}_C[\hat{c}_A], n, 1 - \frac{\alpha}{2})$
- $\underline{p} \leftarrow \min(\underline{p_A}, \underline{p_S})$
- **if** $\hat{s} = 1 \wedge \underline{p} > \frac{1}{2}$ **return** $\hat{c}_A$ and $R := \sigma_\epsilon \Phi^{-1}(\underline{p})$
- **else if** $\hat{s} = 0 \wedge \underline{p_S} \geq \frac{1}{2}$ **return** $F_{\text{Core}}(\boldsymbol{x})$ and $R := 0$
- **else if** $\hat{c}_A = F_{\text{Core}}(\boldsymbol{x}) \wedge \underline{p_A} \geq \frac{1}{2}$ **return** $\hat{c}_A$ and $R := 0$
- **else return** $\oslash$ and $R := 0$

Selection Model We can apply RS to any binary classifier F_{Select} to obtain a smoothed selection model $\bar{F}_{\text{Select}}$. Like Müller et al. (2021), we consider two selection-mechanisms: i) a separate selection-network framing selection as binary classification and ii) a mechanism based on the entropy of the certification-network's logits $\boldsymbol{f}_{\text{Certify}}(\boldsymbol{x})$ defined as $F_{\text{Select}}(\boldsymbol{x}, \theta) := \mathbb{1}_{\mathcal{H}(\text{softmax}(\boldsymbol{f}_{\text{Certify}}(\boldsymbol{x}))) \leq \theta}$ where $\theta \in \mathbb{R}$ denotes the selection threshold. While a separate selection-network performs much better in the deterministic setting (Müller et al., 2021), we find that in our setting the entropy-based mechanism is even more effective (see App. D.3.2). Thus, we focus our evaluation on an entropy-based selection-mechanism. Using such a selection-mechanism allows us to evaluate ACES for a large range of θ, thus computing the full Pareto frontier (shown in Fig. 1), without reevaluating $\bar{F}_{\text{Certify}}$ and F_{Core}. This makes the evaluation of ACES highly computationally efficient. We can even evaluate all component models separately and compute ACES certificates for arbitrary combinations retrospectively, allowing quick evaluations of new component models.

4 Experimental Evaluation

In this section, we evaluate ACES on the ImageNet and CIFAR10 datasets and demonstrate that it yields much higher average certified radii (ACR) and certified accuracies at a wide range of natural accuracies (NAC) than current state-of-the-art methods. Please see App. C for a detailed description of the experimental setup and App. D for significantly extended results, including different training methods and noise levels σ, showing that the effects discussed here are consistent across a wide range of settings.

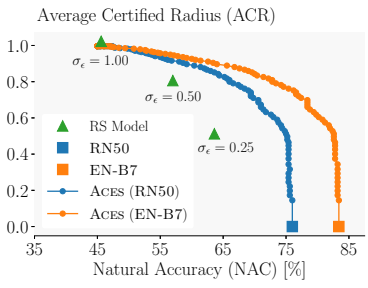

Figure 1: ACR over NAC on ImageNet.

Table 1: Comparison of natural accuracy (NAC), average certified radius (ACR), and certified accuracy and selection rate at various radii on ImageNet with $\sigma_\epsilon = 0.5$. We use a CONSISTENCY trained `ResNet50` as certification-network and an `EfficientNet-B7` as core-network.

θ	NAC	ACR	Certified Accuracy at Radius r							Certified Selection Rate at Radius r						
			0.00	0.25	0.50	0.75	1.00	1.25	1.50	0.00	0.25	0.50	0.75	1.00	1.25	1.50
0.0	83.4	0.000	83.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.1	80.0	0.530	80.0	33.6	32.6	30.2	28.2	25.6	23.0	45.0	40.2	37.2	34.0	31.8	28.2	25.0
0.2	75.4	0.682	75.0	43.6	41.2	38.2	35.8	33.4	30.0	63.8	58.6	55.6	50.6	47.8	45.2	40.8
0.3	68.8	0.744	68.2	48.4	44.4	41.6	39.2	35.6	32.8	78.0	74.2	70.2	66.2	62.8	59.0	55.0
0.6	57.2	0.799	55.4	51.6	48.8	45.0	42.0	39.0	34.6	99.8	99.4	99.0	98.2	97.4	96.6	94.6
1.0	57.2	0.800	55.4	51.6	48.8	45.2	42.2	39.0	34.6	100.0	100.0	100.0	100.0	100.0	100.0	100.0

ACES on ImageNet Fig. 1 compares the average certified radius (ACR) over natural accuracy (NAC) obtained on ImageNet by individual `ResNet50` (green triangles) with those obtained by ACES (dots). We use `ResNet50` with $\sigma_\epsilon = 1.0$ as certification-networks and either another `ResNet50` (blue) or an `EfficientNet-B7` (orange) as the core-network (squares) for ACES. There, the horizontal gap between the individual RS models (triangles) and ACES (orange line) corresponds to the increase in natural accuracy at the same robustness, e.g., 15.8% for $\sigma_\epsilon = 0.5$. We further observe that ACES already dominates the ACR of the individual models, especially at high natural accuracies, when using the small `ResNet50` as core-network and even more so with the stronger `EfficientNet-B7`.

Table 1 shows how the certified accuracy and selection rate (ratio of samples sent to the certification-network) change with the selection threshold θ. Increasing θ from 0.0 to 0.1 only reduces natural accuracy by 3.4% while increasing ACR from 0.0 to 0.530 and certified accuracy at $r = 1.0$ from 0.0% to 28.2%. Similarly, reducing θ from 1.0 to 0.3 loses very little ACR (0.056) and certified accuracy (3.0% at $r = 1.0$) but yields a significant gain in natural accuracy (11.6%).

ACES on CIFAR10 Fig. 2 compares ACES (solid & dashed lines) against a baseline of varying the inference noise levels σ_ϵ (dotted lines) with respect to the robustness accuracy trade-offs obtained on CIFAR10. Using only `ResNet110`, ACES models (solid lines) dominate all individual models across training noise levels $\sigma_t \in \{0.25, 0.5, 1.0\}$ (orange, blue, red). Individual models only reach comparable performance when evaluated at their training noise level. However, covering the full Pareto frontier this way would require training a very large number of networks to match a single ACES model. Using a more precise `LaNet` as core-network for ACES (red dashed line) significantly widens this gap.

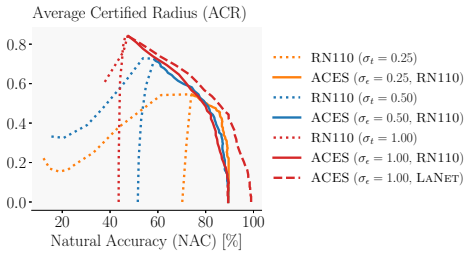

Figure 2: Comparison of ACR over natural accuracy of ACES with different noises σ_ϵ and selection thresholds θ (solid & dashed lines), and individual `ResNet110` evaluated with $\sigma_e \in [0.0, 1.5]$ and trained at $\sigma_t \in \{0.25, 0.5, 1.0\}$.

Selection-Mechanism In Fig. 3, we visualize the distribution of samples that can (blue) and can not (orange) be certified correctly (at $r = 3.0$) over the certification-network's median entropy (over perturbations). Samples to the left of a chosen threshold are assigned to the certification-network and the rest to the core-network. While separation is not perfect, we observe that there is a quick decline in the portion of certifiable samples as entropy increases, indicating that the selection-mechanism works well.

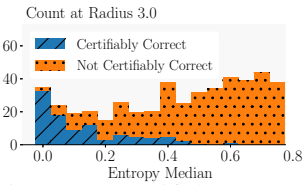

Figure 3: Certifiable correctness over median entropy.

5 CONCLUSION

We extend compositional architectures to probabilistic robustness certification, achieving, for the first time, both high certifiable *and* natural accuracies on the challenging ImageNet dataset. The key component of our ACES architecture is a certified, entropy-based selection-mechanism, choosing, on a per-sample basis, whether to use a smoothed model yielding guarantees or a more accurate standard model for inference. Our experiments show that ACES yields trade-offs between robustness and accuracy that are beyond the reach of current state-of-the-art approaches while being fully orthogonal to other improvements of Randomized Smoothing.

REFERENCES

Motasem Alfarra, Adel Bibi, Philip H. S. Torr, and Bernard Ghanem. Data dependent randomized smoothing. *ArXiv preprint*, abs/2012.04351, 2020.

Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Srndic, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. 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.

Carlo Bonferroni. Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze*, 8, 1936.

C. J. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4), 1934. ISSN 00063444.

Jeremy M. Cohen, Elan Rosenfeld, and J. Zico Kolter. Certified adversarial robustness via randomized smoothing. In *Proc. of ICML*, volume 97, 2019.

Timon Gehr, Matthew Mirman, Dana Drachsler-Cohen, Petar Tsankov, Swarat Chaudhuri, and Martin T. Vechev. AI2: safety and robustness certification of neural networks with abstract interpretation. In *2018 IEEE Symposium on Security and Privacy, SP 2018, Proceedings, 21-23 May 2018, San Francisco, California, USA*, 2018. doi: 10.1109/SP.2018.00058.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, 2016. doi: 10.1109/CVPR.2016.90.

Jongheon Jeong and Jinwoo Shin. Consistency regularization for certified robustness of smoothed classifiers. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Guy Katz, Clark W. Barrett, David L. Dill, Kyle Julian, and Mykel J. Kochenderfer. Reluplex: An efficient SMT solver for verifying deep neural networks. In *Computer Aided Verification - 29th International Conference, CAV 2017, Heidelberg, Germany, July 24-28, 2017, Proceedings, Part I*, volume 10426, 2017. doi: 10.1007/978-3-319-63387-9_5.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial machine learning at scale. In *Proc. of ICLR*, 2017.

Mathias Lécuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. Certified robustness to adversarial examples with differential privacy. In *2019 IEEE Symposium on Security and Privacy, SP 2019, San Francisco, CA, USA, May 19-23, 2019*, 2019. doi: 10.1109/SP.2019.00044.

Bai Li, Changyou Chen, Wenlin Wang, and Lawrence Carin. Certified adversarial robustness with additive noise. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, 2019.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *Proc. of ICLR*, 2018.

Jeet Mohapatra, Ching-Yun Ko, Lily Weng, Pin-Yu Chen, Sijia Liu, and Luca Daniel. Hidden cost of randomized smoothing. In *The 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021, April 13-15, 2021, Virtual Event*, volume 130, 2021.

Mark Niklas Müller, Mislav Balunovic, and Martin T. Vechev. Certify or predict: Boosting certified robustness with compositional architectures. In *Proc. of ICLR*, 2021.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, 2019.

Aditi Raghunathan, Jacob Steinhardt, and Percy Liang. Semidefinite relaxations for certifying robustness to adversarial examples. 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.

Aditi Raghunathan, Sang Michael Xie, Fanny Yang, John C. Duchi, and Percy Liang. Adversarial training can hurt generalization. *ArXiv preprint*, abs/1906.06032, 2019.

Jérôme Rony, Luiz G. Hafemann, Luiz S. Oliveira, Ismail Ben Ayed, Robert Sabourin, and Eric Granger. Decoupling direction and norm for efficient gradient-based L2 adversarial attacks and defenses. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, 2019. doi: 10.1109/CVPR.2019.00445.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115 (3), 2015. doi: 10.1007/s11263-015-0816-y.

Hadi Salman, Jerry Li, Ilya P. Razenshteyn, Pengchuan Zhang, Huan Zhang, Sébastien Bubeck, and Greg Yang. Provably robust deep learning via adversarially trained smoothed classifiers. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, 2019.

Gagandeep Singh, Timon Gehr, Markus Püschel, and Martin T. Vechev. An abstract domain for certifying neural networks. *Proc. ACM Program. Lang.*, 3(POPL), 2019. doi: 10.1145/3290354.

Peter Súkeník, Aleksei Kuvshinov, and Stephan Günnemann. Intriguing properties of input-dependent randomized smoothing. *ArXiv preprint*, abs/2110.05365, 2021.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *Proc. of ICLR*, 2014.

Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proc. of ICML*, volume 97, 2019.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In *Proc. of ICLR*, 2019.

Linnan Wang, Saining Xie, Teng Li, Rodrigo Fonseca, and Yuandong Tian. Sample-efficient neural architecture search by learning actions for monte carlo tree search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. doi: 10.1109/TPAMI.2021.3071343.

Cihang Xie, Mingxing Tan, Boqing Gong, Jiang Wang, Alan L. Yuille, and Quoc V. Le. Adversarial examples improve image recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, 2020. doi: 10.1109/CVPR42600.2020.00090.

Runtian Zhai, Chen Dan, Di He, Huan Zhang, Boqing Gong, Pradeep Ravikumar, Cho-Jui Hsieh, and Liwei Wang. MACER: attack-free and scalable robust training via maximizing certified radius. In *Proc. of ICLR*, 2020.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P. Xing, Laurent El Ghaoui, and Michael I. Jordan. Theoretically principled trade-off between robustness and accuracy. In *Proc. of ICML*, volume 97, 2019.

Huan Zhang, Tsui-Wei Weng, Pin-Yu Chen, Cho-Jui Hsieh, and Luca Daniel. Efficient neural network robustness certification with general activation functions. 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.

A RANDOMIZED SMOOTHING

Algorithm 2 Certification for Randomized Smoothing

function CERTIFY($\sigma_\epsilon, \boldsymbol{x}, n_0, n, \alpha$)
  $\texttt{counts}^0 \leftarrow$ SAMPLEWNOISE($F, \boldsymbol{x}, n_0, \sigma_\epsilon$)
  $\hat{c}_A \leftarrow$ top indices in $\texttt{counts}^0$
  $\texttt{counts} \leftarrow$ SAMPLEWNOISE($F, \boldsymbol{x}, n, \sigma_\epsilon$)
  $\underline{p_A} \leftarrow$ LOWERCONFBND($\texttt{counts}[\hat{c}_A], n, 1-\alpha$)
  if $\underline{p} > \frac{1}{2}$ **return** $\hat{c}_A$ and $R := \sigma_\epsilon \Phi^{-1}(\underline{p})$
  else return $\oslash$ and $R := 0$

Algorithm 3 Prediction for Randomized Smothing

function PREDICT($\sigma_\epsilon, \boldsymbol{x}, n, \alpha$)
  $\texttt{counts} \leftarrow$ SAMPLEWNOISE($F, \boldsymbol{x}, n, \sigma_\epsilon$)
  $\hat{c}_A, \hat{c}_B \leftarrow$ top two indices in $\texttt{counts}$
  $n_A, n_B \leftarrow \texttt{counts}[\hat{c}_A], \texttt{counts}[\hat{c}_B]$
  $\rho_A \leftarrow$ BINOMPVALUE($n_A, n_A + n_B, 0.5$)
  if $\rho_A \leq \alpha$ **return** $\hat{c}_A$
  else return $\oslash$

In this section, we briefly explain the practical certification and inference algorithms CERTIFY and PREDICT, respectively, for a smoothed classifier

$$\bar{F}(\boldsymbol{x}) := \arg\max_c \mathbb{E}_{\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2 \boldsymbol{I})}(F(\boldsymbol{x} + \epsilon) = c)$$

as introduced by Cohen et al. (2019). We first define some components of Algorithms 2 and 3 below before we discuss them in more detail:

SAMPLEWNOISE$(F, x, n, \sigma_\epsilon)$ first samples n inputs $x_1, \ldots, x_n$ as $x_i = x + \epsilon_i$ for $\epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon)$. Then it counts how often F predicts which class for these $x_1, \ldots, x_n$ and returns the corresponding m dimensional array of counts.

LOWERCONFBND$(k, n, 1-\alpha)$ returns a lower bound on the unknown probability p with confidence at least $1-\alpha$ such that $k \sim \mathcal{B}(n, p)$ for the binomial distribution with parameters n and p.

BINOMPVALUE(n_A, n, p) returns the probability of at least n_A success in n Bernoulli trials with success probability p.

Certification We first recall the robustness guarantee for a smoothed classifier (Theorem 2.1):

Theorem 2.1. (Cohen et al. (2019)). *Let $c_A \in [1, \ldots, m]$, $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2 \boldsymbol{I})$, and $\underline{p_A}, \overline{p_B} \in [0, 1]$. If*

$$\mathcal{P}_\epsilon(F(\boldsymbol{x} + \epsilon) = c_A) \geq \underline{p_A} \geq \overline{p_B} \geq \max_{c \neq c_A} \mathcal{P}_\epsilon(F(\boldsymbol{x} + \epsilon) = c), \tag{2}$$

then $\bar{F}(\boldsymbol{x} + \delta) = c_A$ for all δ satisfying $\|\delta\|_2 < R$ with $R := \frac{\sigma_\epsilon}{2}(\Phi^{-1}(\underline{p_A}) - \Phi^{-1}(\overline{p_B}))$.

Unfortunately, computing the exact probabilities $\mathcal{P}_\epsilon(F(\boldsymbol{x} + \epsilon) = c)$ is generally intractable. Thus, to allow practical application, Cohen et al. (2019) propose CERTIFY (Algorithm 2) utilizing Monte Carlo sampling and confidence bounds: First, we draw n_0 samples to determine the majority class $\hat{c}_A$. Then, we draw another n samples to compute a lower bound $\underline{p_A}$ to the success probability, i.e., the probability of the underlying model to predict $\hat{c}_A$ for a perturbed sample, with confidence $1-\alpha$ via the Clopper-Pearson lemma (Clopper and Pearson, 1934). If $\underline{p_A} > 0.5$, we set $\overline{p_B} = 1 - \underline{p_A}$ and obtain radius $R = \sigma_\epsilon \Phi^{-1}(\underline{p_A})$ via Theorem 2.1 with confidence $1-\alpha$, else we abstain (return $\oslash$). See Cohen et al. (2019) for a proof.

Prediction Computing a confidence bound to the success probability with CERTIFY is computationally expensive as the number of samples n is typically large. If we are only interested in computing the class predicted by the smoothed model, we can use the computationally much cheaper PREDICTS (Algorithm 3) proposed by Cohen et al. (2019). Instead of sampling in two separate rounds, we

only draw n samples once and compute the two most frequently predicted classes $\hat{c}_A$ and $\hat{c}_B$ with frequencies n_A and n_B, respectively. Subsequently, we test if the probability of obtaining n_A success in $n_A + n_B$ fair Bernoulli trials is smaller than α, and if so, have with confidence $1 - \alpha$ that the true prediction of the smoothed model is in fact $\hat{c}_A$. See Cohen et al. (2019) for a proof.

Training for Randomized Smoothing To obtain high certified radii via CERTIFY, the base model F has to be trained specifically to cope with the added noise terms ϵ. To achieve this, several training methods have been introduced, which we quickly outline below.

Cohen et al. (2019) propose to use data augmentation with Gaussian noise during training. We refer to this as GAUSSIAN. Salman et al. (2019) suggest SMOOTHADV, combining adversarial training (Madry et al., 2018; Kurakin et al., 2017; Rony et al., 2019) with data augmentation ideas from GAUSSIAN. While effective in improving accuracy, this training procedure comes with a very high computational cost. Zhai et al. (2020) propose MACER as a computationally cheaper alternative with a similar performance by adding a surrogate of the certification radius to the loss and thus more directly optimizing for large radii. Jeong and Shin (2020) build on this approach by replacing this term with a more easily optimizable one and proposing what we refer to as CONSISTENCY.

B PREDICTION & CERTIFICATION FOR ACES

Algorithm 1 Certification for ACES

function CERTIFY($\sigma_\epsilon, \boldsymbol{x}, n_0, n, \alpha$)
  $\texttt{counts}_S^0 \leftarrow$ SAMPLEWNOISE($F_{\text{Select}}, \boldsymbol{x}, n_0, \sigma_\epsilon$)
  $\texttt{counts}_C^0 \leftarrow$ SAMPLEWNOISE($F_{\text{Certify}}, \boldsymbol{x}, n_0, \sigma_\epsilon$)
  $\hat{s} \leftarrow 0$ **if** $\texttt{counts}_S^0[0] > \texttt{counts}_S^0[1]$ **else** 1
  $\hat{c}_A \leftarrow$ top indices in $\texttt{counts}_C^0$
  $\texttt{counts}_S \leftarrow$ SAMPLEWNOISE($F_{\text{Select}}, \boldsymbol{x}, n, \sigma_\epsilon$)
  $\texttt{counts}_C \leftarrow$ SAMPLEWNOISE($F_{\text{Certify}}, \boldsymbol{x}, n, \sigma_\epsilon$)
  $\underline{p_S} \leftarrow$ LOWERCONFBND($\texttt{counts}_S[\hat{s}], n, 1 - \frac{\alpha}{2}$)
  $\underline{p_A} \leftarrow$ LOWERCONFBND($\texttt{counts}_C[\hat{c}_A], n, 1 - \frac{\alpha}{2}$)
  $\underline{p} \leftarrow \min(\underline{p_A}, \underline{p_S})$
  if $\hat{s} = 1 \wedge \underline{p} > \frac{1}{2}$ **return** $\hat{c}_A$ and $R := \sigma_\epsilon \Phi^{-1}(\underline{p})$
  else if $\hat{s} = 0 \wedge \underline{p_S} \geq \frac{1}{2}$ **return** $F_{\text{Core}}(\boldsymbol{x})$ and $R := 0$
  else if $\hat{c}_A = F_{\text{Core}}(\boldsymbol{x}) \wedge \underline{p_A} \geq \frac{1}{2}$ **return** $\hat{c}_A$ and $R := 0$
  else return $\oslash$ and $R := 0$

In this section, we recall the certification approach (Algorithm 1) and introduce the prediction approach (Algorithm 4, below) in detail for ACES as discussed in §3.

Certification For an arbitrary but fixed $\boldsymbol{x}$ we let $c := F_{\text{ACES}}(\boldsymbol{x})$ denote the true output of ACES (Eq. (4)) under exact evaluation of the expectations over perturbations (Eq. (1)) and let

$$R := \begin{cases} \min(R_{\text{Select}}, R_{\text{Certify}}) & \text{if } \bar{F}_{\text{Select}}(\boldsymbol{x}) = 1 \\ 0 & \text{otherwise} \end{cases},$$

where $R_{\text{Select}}, R_{\text{Certify}}$ denote the robustness radius according to Theorem 2.1 for $\bar{F}_{\text{Select}}(\boldsymbol{x})$ and $\bar{F}_{\text{Certify}}(\boldsymbol{x})$, respectively. We now obtain the following guarantees for the outputs of our certification algorithm CERTIFY:

Theorem B.1. *Let $\hat{c}, \hat{R}$ denote the class and robustness radius returned by* CERTIFY *(Algorithm 1) for input $\boldsymbol{x}$. Then, this output $\hat{c}$, computed via sampling, is the true output $F_{\text{ACES}}(\boldsymbol{x} + \delta) =: c = \hat{c} \quad \forall \delta$ with $\|\delta\|_2 \leq \hat{R}$ with confidence at least $1 - \alpha$, if $\hat{c} \neq \oslash$.*

Proof. First, we note that, as CERTIFY (Algorithm 2) in Cohen et al. (2019), our CERTIFY determines $\underline{p_A}$ and $\underline{p_S}$ with probability $1 - \frac{\alpha}{2}$. Thus allowing us to upper bound $\overline{p_B} := 1 - \underline{p_A}$ and giving us $\hat{R}_{\text{Certify}}$ via Theorem 2.1 and similarly $\hat{R}_{\text{Select}}$.

Thus, if $\bar{F}_{\text{Select}}(\boldsymbol{x})$ returns 1 (selecting the certification network) with confidence $1 - \frac{\alpha}{2}$ and $\bar{F}_{\text{Certify}}(\boldsymbol{x})$ returns class c with confidence $1 - \frac{\alpha}{2}$, then we have via union bound with confidence $1 - \alpha$ that $F_{\text{ACES}}(\boldsymbol{x})$ returns $\hat{c} = c$. Further, the probabilities $\underline{p_A}$ and $\underline{p_S}$ induce the robustness radii $\hat{R}_{\text{Select}}$ and $\hat{R}_{\text{Certify}}$, respectively, via Theorem 2.1. Thus we obtain the robustness radius $\hat{R} = \min(\hat{R}_{\text{Select}}, \hat{R}_{\text{Certify}})$ as their minimum.

Should $\bar{F}_{\text{Select}}(\boldsymbol{x}) = 0$ (selecting the core network), with probability $1 - \frac{\alpha}{2}$ we return the deterministically computed $F_{\text{Core}} = \hat{c} = c$, trivially with confidence $1 - \frac{\alpha}{2} \geq 1 - \alpha$. As we only only claim robustness with $\hat{R} = 0$ in this case, the robustness statement is trivially fulfilled.

In case we can not compute the decision of $\bar{F}_{\text{Select}}(\boldsymbol{x})$ with sufficient confidence, but $\bar{F}_{\text{Certify}}(\boldsymbol{x})$ and $F_{\text{Core}}(\boldsymbol{x})$ agree with high confidence, we return the consensus class. We again have trivially from the deterministic F_{Core} and the prediction of $\bar{F}_{\text{Certify}}$ with confidence $1 - \frac{\alpha}{2}$ an overall confidence of $1 - \frac{\alpha}{2} \geq 1 - \alpha$ that indeed $\hat{c} = c$. Finally, in this case we again only claim $\hat{R} = 0$ which is trivially fulfilled. □

Algorithm 4 Prediction for ACES

function PREDICT($\sigma_\epsilon, \boldsymbol{x}, n, \alpha$)
  $\texttt{counts}_S \leftarrow \text{SAMPLEWNOISE}(F_{\text{Select}}, \boldsymbol{x}, n, \sigma_\epsilon$
  $\texttt{counts}_C \leftarrow \text{SAMPLEWNOISE}(F_{\text{Certify}}, \boldsymbol{x}, n, \sigma_\epsilon)$
  $n_0, n_1 \leftarrow \texttt{counts}_S[0], \texttt{counts}_S[1]$
  $\hat{c}_A, \hat{c}_B \leftarrow$ top two indices in $\texttt{counts}_C$
  $n_A, n_B \leftarrow \texttt{counts}_C[\hat{c}_A], \texttt{counts}_C[\hat{c}_B]$
  $\rho_A \leftarrow \text{BINOMPVALUE}(n_A, n_A + n_B, 0.5)$
  if $n_1 > n_0 \wedge \text{BinomPValue}(n_1, n, 0.5) \leq \frac{\alpha}{2} \wedge \rho_A \leq \frac{\alpha}{2}$ **return** $\hat{c}_A$
  else if $n_0 > n_1 \vee \text{BinomPValue}(n_0, n, 0.5) \leq \frac{\alpha}{2}$ **return** $F_{\text{Core}}(\boldsymbol{x})$
  else if $\hat{c}_A = F_{\text{Core}}(\boldsymbol{x}) \wedge \rho_A \leq \frac{\alpha}{2}$ **return** $\hat{c}_A$
  else return $\oslash$

Prediction Let us again consider the setting where for an arbitrary but fixed $\boldsymbol{x}$ we $c := F_{\text{ACES}}(\boldsymbol{x})$ denotes the true output of ACES (Eq. (4)) under exact evaluation of the expectations over perturbations (Eq. (1)). However, now we are only interested in the predicted class $\hat{c}$ and not the robustness radius. We thus introduce PREDICT (Algorithm 4), which is computationally much cheaper than CERTIFY and for which we obtain the following guarantee:

Theorem B.2. *Let $\hat{c}$ be the class returned by* PREDICT *(Algorithm 4) for input $\boldsymbol{x}$. Then, this output computed via sampling is the true output $F_{\text{ACES}}(\boldsymbol{x}) =: c = \hat{c}$ with confidence at least $1 - \alpha$, if $\hat{c} \neq \oslash$ does not abstain.*

Proof. This proof follows analogously to that for CERTIFY (Theorem B.1) from Cohen et al. (2019). □

C EXPERIMENTAL SETUP DETAILS

In this section, we discuss experimental details. We evaluated ACES on the ImageNet (Russakovsky et al., 2015) and the CIFAR10 (Krizhevsky et al., 2009) datasets. For ImageNet, we combine `ResNet50` (He et al., 2016) selection- and certification-networks with `EfficientNet-B7` core-networks (Tan and Le, 2019). For CIFAR10, we use `ResNet110` (He et al., 2016) selection- and certification-networks, and `LaNet` (Wang et al., 2021) core-networks. We implement training and inference in PyTorch (Paszke et al., 2019) and conduct all of our experiments on single GeForce RTX 2080 Ti.

As core-networks, we use pre-trained `EfficientNet-B7`[1] and `LaNet` (Wang et al., 2021) for ImageNet and CIFAR10, respectively. As certification-networks, we use pre-trained `ResNet50` and `ResNet110` from Cohen et al. (2019) (GAUSSIAN), Salman et al. (2019) (SMOOTHADV), and Zhai et al. (2020) (MACER). Additionally, we train smoothed models with CONSISTENCY (Jeong and Shin, 2020)

[1] https://github.com/lukemelas/EfficientNet-PyTorch/tree/master/examples/imagenet

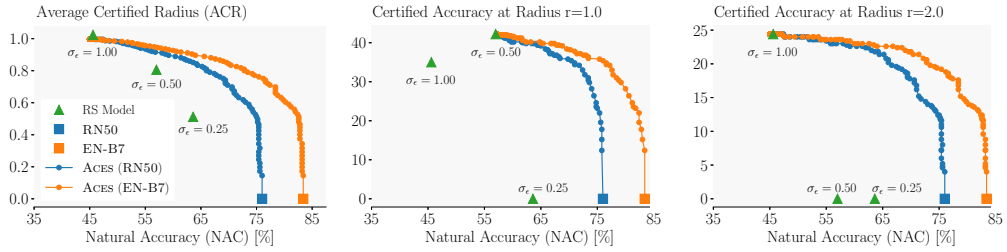

Figure 4: Comparison of ACES (blue and orange dots) and individual smoothed models (green triangles) on ImageNet with CONSISTENCY trained models with respect to average certified radius (left), certified accuracy at $r = 1.0$ (middle), and certified accuracy at $r = 2.0$ (right) over natural accuracy. We use `ResNet50` for individual networks and as certification-networks for all ACES models. We consider ACES models with `ResNet50` and `EfficientNet-B7` core-networks.

using the parameters reported to yield the largest ACR, except on ImageNet with $\sigma_\epsilon = 0.25$ where we use $\eta = 0.5$ and $\lambda = 5$ (there, no parameters were reported).

We follow previous work (Cohen et al., 2019; Salman et al., 2019) and evaluate every 20^{th} image of the CIFAR10 test set and every 100^{th} of the ImageNet test set (Cohen et al., 2019; Jeong and Shin, 2020), yielding 500 test samples for each. For both, we use $n_0 = 100$ and $n = 100'000$ for certification, and $n = 10'000$ for prediction (to report natural accuracy). To obtain an overall confidence of $\alpha = 0.001$ via Bonferroni correction (Bonferroni, 1936), we use $\alpha' = 0.0005$ to certify the selection and the certification model. To compute the entropy, we use the logarithm with basis m (number of classes), such that the resulting entropies are always in $[0, 1]$. Certifying and predicting an ACES model on the 500 test samples we consider takes approximately 23.8 hours on ImageNet, and 10.8 hours on CIFAR10 overall, using one RTX 2080 Ti. This includes computations for a wide range (> 100) values for the selection threshold θ.

D ADDITIONAL EXPERIMENTS

In this section, we provide a significantly extended evaluation focusing on the following aspects:

In App. D.1 and D.2, we evaluate ACES for different training methods and a range of noise levels σ on ImageNet and CIFAR10, respectively.

In App. D.3, we provide an in-depth analysis of the selection-mechanism, considering different measures of selection performance and both entropy-based selection and a separate selection-network.

In App. D.4, we discuss the robustness-accuracy trade-offs obtained by varying the noise level σ_ϵ used at inference.

D.1 ADDITIONAL RESULTS ON IMAGENET

In this section, we evaluate ACES on ImageNet for a wide range of training methods (GAUSSIAN, SMOOTHADV, and CONSISTENCY) and noise levels $\sigma \in \{0.25, 0.50, 1.00\}$. In particular, we provide detailed results on the certified accuracies obtained by ACES in Table 2 and the corresponding certified selection rates in Table 3 for $\sigma_t = \sigma_\epsilon = 0.25$. Similarly, Tables 4 and 5 and Tables 6 and 7 contain results for $\sigma_\epsilon = 0.5$ and $\sigma_\epsilon = 1.0$, respectively.

In Fig. 4, we visualize the trade-off between natural and certified accuracy at fixed radii for ACES (blue and orange dots) and individual smoothed models (green triangles). We observe that ACES achieves significant certified accuracies at natural accuracies not achievable at all by conventional smoothed models.

For example, the highest natural accuracy (63.6%) obtained by one of the CONSISTENCY smoothed models requires $\sigma_\epsilon = 0.25$, leading to a certified accuracy of 0.0% at l_2 radius 2.0. ACES, in contrast, can use a certification-network with $\sigma_\epsilon = 1.0$ to, e.g., obtain a similar natural accuracy of 66.8% and a much higher certified accuracy of 22.6%.

Table 2: Natural accuracy (NAC), average certified radius (ACR) and certified accuracy at different radii on ImageNet with $\sigma_t = \sigma_\epsilon = 0.25$ for a range of threshold parameters θ and an ACES model with entropy selection, a `ResNet50` certification-network and an `EfficientNet-B7` core-network.

Training	θ	NAC	ACR	Certified Accuracy at Radius r								
				0.0	0.25	0.5	0.75	1.0	1.25	1.5	1.75	2.0
GAUSSIAN	0.00	83.4	0.000	83.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	0.10	82.2	0.273	82.2	35.4	27.4	21.4	0.0	0.0	0.0	0.0	0.0
	0.20	80.0	0.382	80.0	47.6	40.2	30.4	0.0	0.0	0.0	0.0	0.0
	0.30	78.6	0.431	78.6	54.6	45.6	35.2	0.0	0.0	0.0	0.0	0.0
	0.40	75.2	0.454	74.6	56.6	47.4	36.6	0.0	0.0	0.0	0.0	0.0
	0.50	72.8	0.464	71.4	58.0	48.4	37.4	0.0	0.0	0.0	0.0	0.0
	0.60	70.4	0.467	68.6	58.4	48.8	37.8	0.0	0.0	0.0	0.0	0.0
	0.70	69.0	0.468	67.0	58.4	49.0	37.8	0.0	0.0	0.0	0.0	0.0
	0.80	68.8	0.468	66.8	58.4	49.0	37.8	0.0	0.0	0.0	0.0	0.0
	0.90	68.8	0.468	66.8	58.4	49.0	37.8	0.0	0.0	0.0	0.0	0.0
	1.00	68.8	0.468	66.8	58.4	49.0	37.8	0.0	0.0	0.0	0.0	0.0
SMOOTHADV	0.00	83.4	0.000	83.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	0.10	82.8	0.269	82.8	31.6	28.6	24.8	0.0	0.0	0.0	0.0	0.0
	0.20	79.6	0.382	79.6	44.0	40.8	36.0	0.0	0.0	0.0	0.0	0.0
	0.30	76.6	0.435	76.6	49.4	45.4	42.0	0.0	0.0	0.0	0.0	0.0
	0.40	72.8	0.469	72.6	53.4	48.6	45.8	0.0	0.0	0.0	0.0	0.0
	0.50	70.4	0.489	70.2	55.6	50.6	47.2	0.0	0.0	0.0	0.0	0.0
	0.60	66.4	0.503	66.0	57.2	52.8	48.2	0.0	0.0	0.0	0.0	0.0
	0.70	65.0	0.508	64.6	57.6	53.4	49.0	0.0	0.0	0.0	0.0	0.0
	0.80	64.4	0.511	64.0	58.0	53.4	49.2	0.0	0.0	0.0	0.0	0.0
	0.90	64.4	0.511	64.0	58.0	53.4	49.2	0.0	0.0	0.0	0.0	0.0
	1.00	64.4	0.511	64.0	58.0	53.4	49.2	0.0	0.0	0.0	0.0	0.0
CONSISTENCY	0.00	83.4	0.000	83.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	0.10	80.4	0.390	80.4	45.0	41.4	36.4	0.0	0.0	0.0	0.0	0.0
	0.20	76.2	0.466	76.2	53.0	49.0	44.8	0.0	0.0	0.0	0.0	0.0
	0.30	71.2	0.492	71.2	57.0	52.2	47.0	0.0	0.0	0.0	0.0	0.0
	0.40	67.8	0.505	67.2	58.6	53.6	48.0	0.0	0.0	0.0	0.0	0.0
	0.50	63.6	0.508	63.2	58.8	53.8	48.4	0.0	0.0	0.0	0.0	0.0
	0.60	63.6	0.509	63.0	58.8	54.0	48.4	0.0	0.0	0.0	0.0	0.0
	0.70	63.6	0.509	63.2	58.8	54.0	48.6	0.0	0.0	0.0	0.0	0.0
	0.80	63.8	0.509	63.2	58.8	54.0	48.6	0.0	0.0	0.0	0.0	0.0
	0.90	63.8	0.509	63.2	58.8	54.0	48.6	0.0	0.0	0.0	0.0	0.0
	1.00	63.8	0.509	63.2	58.8	54.0	48.6	0.0	0.0	0.0	0.0	0.0

Table 3: Natural accuracy (NAC), average certified radius (ACR) and certified selection rate (portion of samples selected for the certification-network) at different radii on ImageNet with $\sigma_t = \sigma_\epsilon = 0.25$ for a range of threshold parameters θ and an ACES model with entropy selection, a `ResNet50` certification-network and an `EfficientNet-B7` core-network.

Training	θ	NAC	ACR	Certified Selection Rate at Radius r								
				0.0	0.25	0.5	0.75	1.0	1.25	1.5	1.75	2.0
GAUSSIAN	0.00	83.4	0.000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	0.10	82.2	0.273	47.0	37.8	29.0	22.2	0.0	0.0	0.0	0.0	0.0
	0.20	80.0	0.382	66.0	57.0	48.8	39.8	0.0	0.0	0.0	0.0	0.0
	0.30	78.6	0.431	76.6	70.4	61.4	53.0	0.0	0.0	0.0	0.0	0.0
	0.40	75.2	0.454	86.2	80.0	72.4	64.6	0.0	0.0	0.0	0.0	0.0
	0.50	72.8	0.464	92.4	87.4	81.8	75.4	0.0	0.0	0.0	0.0	0.0
	0.60	70.4	0.467	97.4	95.4	91.0	85.2	0.0	0.0	0.0	0.0	0.0
	0.70	69.0	0.468	99.8	99.2	97.2	95.0	0.0	0.0	0.0	0.0	0.0
	0.80	68.8	0.468	100.0	100.0	99.8	99.6	0.0	0.0	0.0	0.0	0.0
	0.90	68.8	0.468	100.0	100.0	100.0	100.0	0.0	0.0	0.0	0.0	0.0
	1.00	68.8	0.468	100.0	100.0	100.0	100.0	0.0	0.0	0.0	0.0	0.0
SMOOTHADV	0.00	83.4	0.000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	0.10	82.8	0.269	37.8	34.2	30.8	26.2	0.0	0.0	0.0	0.0	0.0
	0.20	79.6	0.382	56.6	53.0	48.2	43.2	0.0	0.0	0.0	0.0	0.0
	0.30	76.6	0.435	70.2	66.0	61.6	56.6	0.0	0.0	0.0	0.0	0.0
	0.40	72.8	0.469	80.6	76.4	71.8	68.6	0.0	0.0	0.0	0.0	0.0
	0.50	70.4	0.489	88.8	84.6	82.0	79.2	0.0	0.0	0.0	0.0	0.0
	0.60	66.4	0.503	95.6	93.6	91.2	88.2	0.0	0.0	0.0	0.0	0.0
	0.70	65.0	0.508	98.4	98.0	97.6	96.0	0.0	0.0	0.0	0.0	0.0
	0.80	64.4	0.511	100.0	99.8	99.8	99.4	0.0	0.0	0.0	0.0	0.0
	0.90	64.4	0.511	100.0	100.0	100.0	100.0	0.0	0.0	0.0	0.0	0.0
	1.00	64.4	0.511	100.0	100.0	100.0	100.0	0.0	0.0	0.0	0.0	0.0
CONSISTENCY	0.00	83.4	0.000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	0.10	80.4	0.390	55.4	51.0	46.6	40.0	0.0	0.0	0.0	0.0	0.0
	0.20	76.2	0.466	72.4	67.8	61.0	57.0	0.0	0.0	0.0	0.0	0.0
	0.30	71.2	0.492	86.0	80.4	75.2	71.0	0.0	0.0	0.0	0.0	0.0
	0.40	67.8	0.505	93.0	89.8	87.2	82.2	0.0	0.0	0.0	0.0	0.0
	0.50	63.6	0.508	98.4	96.4	94.2	91.8	0.0	0.0	0.0	0.0	0.0
	0.60	63.6	0.509	99.8	99.2	98.8	98.2	0.0	0.0	0.0	0.0	0.0
	0.70	63.6	0.509	100.0	99.8	99.8	99.8	0.0	0.0	0.0	0.0	0.0
	0.80	63.8	0.509	100.0	100.0	100.0	100.0	0.0	0.0	0.0	0.0	0.0
	0.90	63.8	0.509	100.0	100.0	100.0	100.0	0.0	0.0	0.0	0.0	0.0
	1.00	63.8	0.509	100.0	100.0	100.0	100.0	0.0	0.0	0.0	0.0	0.0

Table 4: Natural accuracy (NAC), average certified radius (ACR) and certified accuracy at different radii on ImageNet with $\sigma_t = \sigma_\epsilon = 0.50$ for a range of threshold parameters θ and an ACES model with entropy selection, a `ResNet50` certification-network and an `EfficientNet-B7` core-network.

Training	θ	NAC	ACR	Certified Accuracy at Radius r								
				0.0	0.25	0.5	0.75	1.0	1.25	1.5	1.75	2.0
GAUSSIAN	0.00	83.4	0.000	83.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	0.10	82.4	0.380	82.4	29.2	25.6	23.2	20.0	16.4	13.4	9.8	0.0
	0.20	78.6	0.536	78.4	38.8	33.4	30.6	28.6	24.4	21.2	16.4	0.0
	0.30	74.2	0.619	73.4	44.0	40.2	35.2	31.4	29.0	24.6	19.6	0.0
	0.40	70.8	0.665	69.4	47.8	43.0	38.2	33.2	30.6	26.2	20.2	0.0
	0.50	65.8	0.693	64.4	50.2	44.4	40.6	35.4	31.8	27.0	20.8	0.0
	0.60	62.4	0.712	60.6	51.0	45.6	42.0	36.8	33.0	27.8	21.2	0.0
	0.70	59.8	0.716	57.4	51.4	45.6	42.4	37.2	33.0	28.2	21.4	0.0
	0.80	60.0	0.717	57.4	51.6	45.8	42.4	37.2	33.0	28.2	21.4	0.0
	0.90	59.8	0.717	57.2	51.6	45.8	42.4	37.2	33.0	28.2	21.4	0.0
	1.00	59.8	0.717	57.2	51.6	45.8	42.4	37.2	33.0	28.2	21.4	0.0
SMOOTHADV	0.00	83.4	0.000	83.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	0.10	83.2	0.308	83.2	20.2	18.2	17.2	16.4	14.8	13.0	11.4	0.0
	0.20	81.0	0.486	81.2	31.6	29.6	26.8	24.4	23.6	21.2	19.6	0.0
	0.30	76.8	0.592	77.0	37.6	35.2	33.2	31.2	28.8	26.8	24.0	0.0
	0.40	73.2	0.661	73.4	42.2	39.6	36.6	34.2	31.8	29.8	27.8	0.0
	0.50	68.2	0.716	68.4	46.2	43.0	39.4	36.8	34.0	32.0	30.2	0.0
	0.60	63.4	0.765	63.2	49.8	46.0	42.2	39.6	36.2	34.2	31.0	0.0
	0.70	57.8	0.791	57.4	51.4	47.4	43.4	41.0	37.8	35.4	32.2	0.0
	0.80	55.6	0.806	55.0	52.4	48.6	44.2	41.8	38.6	35.6	32.8	0.0
	0.90	55.6	0.809	55.0	52.6	48.8	44.4	42.2	38.8	35.6	32.8	0.0
	1.00	55.6	0.809	55.0	52.6	48.8	44.4	42.2	38.8	35.6	32.8	0.0
CONSISTENCY	0.00	83.4	0.000	83.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	0.10	80.0	0.530	80.0	33.6	32.6	30.2	28.2	25.6	23.0	19.6	0.0
	0.20	75.4	0.682	75.0	43.6	41.2	38.2	35.8	33.4	30.0	27.0	0.0
	0.30	68.8	0.744	68.2	48.4	44.4	41.6	39.2	35.6	32.8	29.2	0.0
	0.40	62.4	0.777	61.6	50.2	47.6	43.8	40.2	37.4	33.4	30.8	0.0
	0.50	59.2	0.795	57.6	51.4	48.2	45.0	41.8	38.6	34.4	30.8	0.0
	0.60	57.2	0.799	55.4	51.6	48.8	45.0	42.0	39.0	34.6	31.0	0.0
	0.70	57.2	0.800	55.4	51.6	48.8	45.2	42.2	39.0	34.6	31.0	0.0
	0.80	57.2	0.800	55.4	51.6	48.8	45.2	42.2	39.0	34.6	31.0	0.0
	0.90	57.2	0.800	55.4	51.6	48.8	45.2	42.2	39.0	34.6	31.0	0.0
	1.00	57.2	0.800	55.4	51.6	48.8	45.2	42.2	39.0	34.6	31.0	0.0

Table 5: Natural accuracy (NAC), average certified radius (ACR) and certified selection rate (portion of samples selected for the certification-network) at different radii on ImageNet with $\sigma_t = \sigma_\epsilon = 0.50$ for a range of threshold parameters θ and an ACES model with entropy selection, a `ResNet50` certification-network and an `EfficientNet-B7` core-network.

Training	θ	NAC	ACR	Certified Selection Rate at Radius r								
				0.0	0.25	0.5	0.75	1.0	1.25	1.5	1.75	2.0
GAUSSIAN	0.00	83.4	0.000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	0.10	82.4	0.380	35.0	31.6	27.6	24.2	20.6	16.8	13.8	10.2	0.0
	0.20	78.6	0.536	53.2	47.8	41.6	37.2	35.6	30.8	27.2	23.4	0.0
	0.30	74.2	0.619	68.2	61.8	57.2	49.6	45.2	41.2	37.6	32.2	0.0
	0.40	70.8	0.665	78.2	73.0	69.4	62.8	58.2	53.2	47.0	40.6	0.0
	0.50	65.8	0.693	88.4	83.8	79.8	74.4	71.0	65.0	59.2	51.8	0.0
	0.60	62.4	0.712	94.6	91.6	89.6	86.2	82.0	78.8	74.2	65.4	0.0
	0.70	59.8	0.716	99.2	97.6	95.8	94.4	92.2	90.2	88.0	82.0	0.0
	0.80	60.0	0.717	99.8	99.8	99.6	99.6	99.6	99.4	97.8	96.6	0.0
	0.90	59.8	0.717	100.0	100.0	100.0	100.0	100.0	99.8	99.8	99.8	0.0
	1.00	59.8	0.717	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	0.0
SMOOTHADV	0.00	83.4	0.000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	0.10	83.2	0.308	24.6	21.8	19.8	18.6	17.6	15.8	14.0	12.4	0.0
	0.20	81.0	0.486	40.6	38.0	35.4	32.4	30.2	29.0	26.6	25.2	0.0
	0.30	76.8	0.592	52.4	50.4	48.0	45.4	43.8	41.2	37.4	34.6	0.0
	0.40	73.2	0.661	62.8	59.8	57.8	56.2	53.4	51.6	49.6	46.8	0.0
	0.50	68.2	0.716	73.4	71.4	69.2	66.8	64.2	61.2	59.2	56.8	0.0
	0.60	63.4	0.765	86.4	83.6	81.6	78.2	76.6	74.8	71.8	69.0	0.0
	0.70	57.8	0.791	95.4	94.0	92.2	90.4	89.6	87.4	85.2	83.8	0.0
	0.80	55.6	0.806	99.6	99.0	98.8	98.8	98.4	98.0	97.0	95.6	0.0
	0.90	55.6	0.809	100.0	100.0	100.0	100.0	99.8	99.8	99.8	99.8	0.0
	1.00	55.6	0.809	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	0.0
CONSISTENCY	0.00	83.4	0.000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	0.10	80.0	0.530	45.0	40.2	37.2	34.0	31.8	28.2	25.0	21.0	0.0
	0.20	75.4	0.682	63.8	58.6	55.6	50.6	47.8	45.2	40.8	36.4	0.0
	0.30	68.8	0.744	78.0	74.2	70.2	66.2	62.8	59.0	55.0	50.6	0.0
	0.40	62.4	0.777	90.2	85.4	82.8	80.2	76.4	72.6	67.0	63.6	0.0
	0.50	59.2	0.795	96.8	95.2	93.2	90.8	88.0	84.4	81.8	78.4	0.0
	0.60	57.2	0.799	99.8	99.4	99.0	98.2	97.4	96.6	94.6	91.0	0.0
	0.70	57.2	0.800	100.0	100.0	100.0	100.0	100.0	100.0	99.8	99.0	0.0
	0.80	57.2	0.800	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	0.0
	0.90	57.2	0.800	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	0.0
	1.00	57.2	0.800	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	0.0

Table 6: Natural accuracy (NAC), average certified radius (ACR) and certified accuracy at different radii on ImageNet with $\sigma_t = \sigma_\epsilon = 1.00$ for a range of threshold parameters θ and an ACES model with entropy selection, a `ResNet50` certification-network and an `EfficientNet-B7` core-network.

Training	θ	NAC	ACR	Certified Accuracy at Radius r								
				0.0	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0
GAUSSIAN	0.00	83.4	0.000	83.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	0.10	83.0	0.322	83.0	15.0	11.8	9.4	7.6	5.2	4.2	3.2	0.0
	0.20	80.6	0.513	80.6	22.8	18.2	15.2	11.8	9.4	7.4	4.6	0.0
	0.30	75.8	0.650	75.2	28.6	23.6	19.4	14.4	11.2	9.6	6.2	0.0
	0.40	71.0	0.741	70.4	32.4	27.6	22.2	16.8	12.6	10.4	7.4	0.0
	0.50	64.2	0.801	62.4	34.8	29.6	24.4	18.6	13.8	11.6	8.2	0.0
	0.60	56.4	0.846	54.6	37.2	31.6	25.4	19.0	14.6	12.0	8.6	0.0
	0.70	50.0	0.860	46.8	37.8	32.6	25.4	19.2	14.6	12.0	8.8	0.0
	0.80	47.6	0.862	43.8	37.8	32.6	25.8	19.4	14.6	12.0	8.8	0.0
	0.90	47.4	0.862	43.6	37.8	32.6	25.8	19.4	14.6	12.0	8.8	0.0
	1.00	47.4	0.862	43.6	37.8	32.6	25.8	19.4	14.6	12.0	8.8	0.0
SMOOTHADV	0.00	83.4	0.000	83.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	0.10	83.4	0.254	83.4	9.2	8.4	7.2	6.0	5.8	4.6	4.6	0.0
	0.20	82.2	0.407	82.2	14.2	12.2	11.4	10.0	9.2	8.8	7.6	0.0
	0.30	79.8	0.541	79.6	18.4	17.0	16.0	14.2	12.0	10.6	9.8	0.0
	0.40	76.8	0.653	76.6	22.2	20.6	19.0	16.8	15.2	13.4	11.6	0.0
	0.50	70.8	0.755	70.6	26.4	23.8	21.8	19.2	17.0	15.4	13.2	0.0
	0.60	64.2	0.854	63.6	30.0	27.4	24.8	22.4	18.6	16.8	14.6	0.0
	0.70	53.2	0.933	52.6	32.2	30.2	27.2	23.8	20.4	19.2	16.0	0.0
	0.80	44.0	0.985	43.4	34.6	31.2	28.6	25.2	21.8	19.8	16.6	0.0
	0.90	39.8	0.999	39.2	35.2	32.0	29.2	25.6	22.0	19.8	16.6	0.0
	1.00	39.8	0.999	39.2	35.2	32.0	29.2	25.6	22.0	19.8	16.6	0.0
CONSISTENCY	0.00	83.4	0.000	83.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	0.10	82.8	0.375	82.8	14.0	12.6	11.0	9.6	8.6	6.8	5.6	0.0
	0.20	81.8	0.559	81.8	22.2	19.2	15.4	13.2	11.8	10.8	8.0	0.0
	0.30	78.4	0.698	77.8	27.6	24.0	20.2	17.6	14.6	12.4	10.2	0.0
	0.40	72.8	0.800	72.4	31.0	28.2	23.8	20.2	18.0	14.0	10.6	0.0
	0.50	66.8	0.881	66.2	34.0	30.4	25.8	22.6	20.4	14.4	11.6	0.0
	0.60	57.8	0.941	56.8	36.8	32.6	27.6	23.6	21.6	15.8	11.6	0.0
	0.70	51.8	0.979	50.6	39.0	34.0	28.4	24.0	22.0	16.6	12.0	0.0
	0.80	46.0	0.996	44.6	39.4	35.0	29.4	24.4	22.0	16.6	12.0	0.0
	0.90	45.0	0.997	43.2	39.6	35.0	29.4	24.4	22.0	16.6	12.0	0.0
	1.00	45.0	0.997	43.2	39.6	35.0	29.4	24.4	22.0	16.6	12.0	0.0

Table 7: Natural accuracy (NAC), average certified radius (ACR) and certified selection rate (portion of samples selected for the certification-network) at different radii on ImageNet with $\sigma_t = \sigma_\epsilon = 1.00$ for a range of threshold parameters θ and an ACES model with entropy selection, a `ResNet50` certification-network and an `EfficientNet-B7` core-network.

Training	θ	NAC	ACR	Certified Selection Rate at Radius r								
				0.0	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0
GAUSSIAN	0.00	83.4	0.000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	0.10	83.0	0.322	19.8	16.4	12.8	10.2	8.2	5.6	4.4	3.4	0.0
	0.20	80.6	0.513	34.2	28.4	22.6	19.2	16.0	13.4	10.0	7.0	0.0
	0.30	75.8	0.650	48.0	41.0	34.4	29.0	23.8	19.4	15.4	11.2	0.0
	0.40	71.0	0.741	60.2	53.0	47.0	40.8	34.6	28.6	23.6	16.8	0.0
	0.50	64.2	0.801	73.8	65.4	57.8	52.8	47.8	40.4	33.8	24.6	0.0
	0.60	56.4	0.846	85.6	79.6	73.4	66.0	59.8	53.2	47.8	37.2	0.0
	0.70	50.0	0.860	95.8	93.0	89.8	84.6	78.0	70.8	64.6	55.8	0.0
	0.80	47.6	0.862	99.8	99.6	99.4	97.4	95.8	93.6	89.6	82.0	0.0
	0.90	47.4	0.862	100.0	100.0	100.0	99.8	99.8	99.8	99.8	99.6	0.0
	1.00	47.4	0.862	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	0.0
SMOOTHADV	0.00	83.4	0.000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	0.10	83.4	0.254	11.8	10.4	9.4	8.2	6.8	6.2	5.0	5.0	0.0
	0.20	82.2	0.407	20.0	18.0	16.0	14.4	13.2	12.2	11.2	9.8	0.0
	0.30	79.8	0.541	28.2	25.2	23.4	21.8	20.0	16.6	14.2	13.0	0.0
	0.40	76.8	0.653	36.4	33.2	31.4	28.8	27.0	25.2	22.6	19.4	0.0
	0.50	70.8	0.755	49.0	45.0	42.0	38.8	34.6	32.6	29.4	26.4	0.0
	0.60	64.2	0.854	62.4	58.2	54.2	51.2	48.4	44.8	40.2	36.4	0.0
	0.70	53.2	0.933	77.6	75.2	72.2	67.4	64.0	61.0	57.2	52.4	0.0
	0.80	44.0	0.985	94.0	92.4	91.2	89.0	86.4	83.0	79.0	75.0	0.0
	0.90	39.8	0.999	100.0	99.6	99.2	98.8	98.8	98.8	98.2	97.4	0.0
	1.00	39.8	0.999	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	0.0
CONSISTENCY	0.00	83.4	0.000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	0.10	82.8	0.375	19.4	16.2	14.8	13.0	11.0	9.8	7.6	6.4	0.0
	0.20	81.8	0.559	32.2	27.0	24.4	21.4	18.8	16.4	14.6	11.6	0.0
	0.30	78.4	0.698	41.4	39.0	34.8	31.6	27.2	23.4	20.2	16.8	0.0
	0.40	72.8	0.800	51.8	47.8	43.6	40.6	36.4	33.0	28.8	24.0	0.0
	0.50	66.8	0.881	63.6	57.4	53.0	50.0	46.8	42.4	37.6	32.2	0.0
	0.60	57.8	0.941	79.2	74.2	70.2	64.2	58.6	53.4	48.8	45.6	0.0
	0.70	51.8	0.979	90.4	87.4	84.0	80.2	75.4	71.2	67.4	60.4	0.0
	0.80	46.0	0.996	97.6	96.8	96.2	95.2	93.6	90.6	88.2	83.8	0.0
	0.90	45.0	0.997	100.0	100.0	100.0	100.0	99.8	99.8	99.4	98.8	0.0
	1.00	45.0	0.997	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	0.0

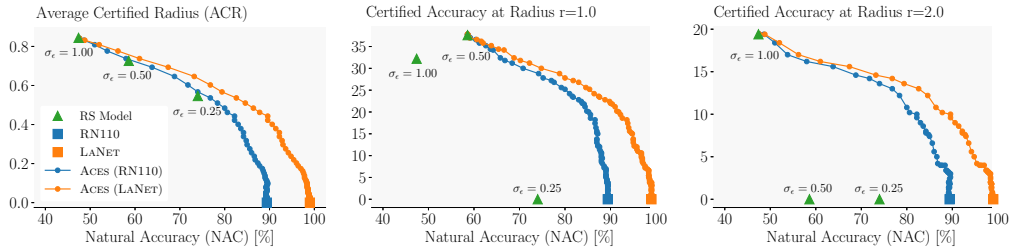

Figure 5: Comparison of ACES (blue and orange dots) and individual smoothed models (green triangles) on CIFAR10 with SMOOTHADV trained models with respect to average certified radius (left), certified accuracy at $r = 1.0$ (middle), and certified accuracy at $r = 2.0$ (right) over natural accuracy. We use `ResNet110` for individual networks and as certification-networks for all ACES models. We consider ACES models with `ResNet110` and `LaNet` core-networks.

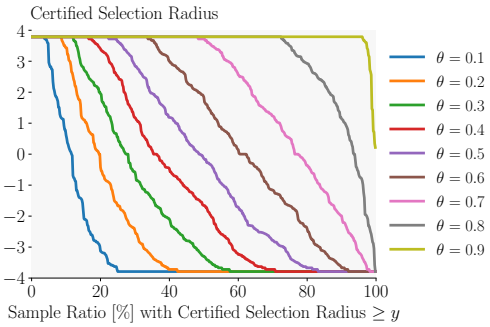

Figure 6: Certified radii of an entropy-based selection mechanism for a range of θ over percentile on ImageNet for a SMOOTHADV trained `ResNet50` model. Positive radii correspond to the selection of the certification-network and negative radii to that of the core-network. A zero radius corresponds to abstentions $\oslash$.

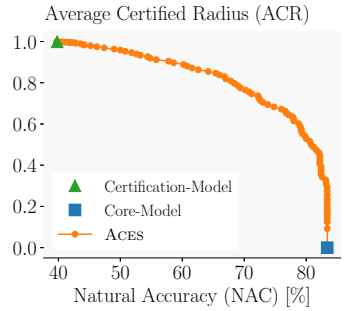

Figure 7: ACR over natural accuracy for an ACES model based on SMOOTHADV trained `ResNet50` certification-network, a corresponding entropy-based selection-network and an `EfficientNet-B7` core-network on ImageNet.

D.2 ADDITIONAL RESULTS ON CIFAR10

In this section, we evaluate ACES on CIFAR10 for a wide range of training methods (GAUSSIAN, SMOOTHADV, MACER, and CONSISTENCY) and noise levels $\sigma \in \{0.25, 0.50, 1.00\}$. In particular, we provide detailed results on the certified accuracies obtained by ACES in Table 8 and the corresponding certified selection rates in Table 9 for $\sigma_t = \sigma_\epsilon = 0.25$. Similarly, Tables 10 and 11 and Tables 12 and 13 contain results for $\sigma_\epsilon = 0.5$ and $\sigma_\epsilon = 1.0$, respectively.

In Fig. 5, we visualize the trade-off between natural and certified accuracy at fixed radii for ACES (blue and orange dots) and individual smoothed models (green triangles). We observe that ACES achieves significant certified accuracies at natural accuracies not achievable at all by conventional smoothed models.

D.3 SELECTION-MECHANISM ABLATION

In this section, we investigate the entropy-based selection-mechanism, introduced in §3, in more detail and compare it to one based on a separate selection-network.

D.3.1 SELECTION CERTIFICATION

In Fig. 6, we visualize the certified radii of the prediction of an entropy-based selection-mechanism based on an SMOOTHADV trained `ResNet50` with $\sigma = 1.00$ for ImageNet. A positive radius corresponds to a certified selection of the certification-network with that radius, and a negative radius corresponds to a certified selection of the core-network. A radius of 0 corresponds to the

Table 8: Natural accuracy (NAC), average certified radius (ACR) and certified accuracy at different radii on CIFAR10 with $\sigma_t = \sigma_\epsilon = 0.25$ for a range of threshold parameters θ and an ACES model with entropy selection, a `ResNet110` certification-network and an `LaNet` core-network.

Training	θ	NAC	ACR	Certified Accuracy at Radius r								
				0.0	0.25	0.5	0.75	1.0	1.25	1.5	1.75	2.00
GAUSSIAN	0.00	99.0	0.000	99.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	0.10	98.0	0.189	98.0	28.8	18.4	9.6	0.0	0.0	0.0	0.0	0.0
	0.20	96.4	0.247	96.6	35.8	24.8	14.0	0.0	0.0	0.0	0.0	0.0
	0.30	94.6	0.303	94.8	41.6	29.6	19.6	0.0	0.0	0.0	0.0	0.0
	0.40	90.4	0.358	90.6	49.8	35.8	22.6	0.0	0.0	0.0	0.0	0.0
	0.50	85.4	0.397	85.2	56.0	40.4	24.2	0.0	0.0	0.0	0.0	0.0
	0.60	81.6	0.416	79.8	59.4	42.2	25.4	0.0	0.0	0.0	0.0	0.0
	0.70	78.2	0.421	76.0	60.0	42.8	25.6	0.0	0.0	0.0	0.0	0.0
	0.80	77.8	0.422	75.4	60.0	42.8	25.6	0.0	0.0	0.0	0.0	0.0
	0.90	77.8	0.422	75.4	60.0	42.8	25.6	0.0	0.0	0.0	0.0	0.0
	1.00	77.8	0.422	75.4	60.0	42.8	25.6	0.0	0.0	0.0	0.0	0.0
SMOOTHADV	0.00	99.0	0.000	99.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	0.10	98.8	0.161	98.8	19.6	17.4	13.4	0.0	0.0	0.0	0.0	0.0
	0.20	98.4	0.222	98.4	27.6	22.6	18.4	0.0	0.0	0.0	0.0	0.0
	0.30	97.4	0.288	97.4	35.4	29.4	24.8	0.0	0.0	0.0	0.0	0.0
	0.40	94.8	0.352	94.6	43.0	37.2	29.6	0.0	0.0	0.0	0.0	0.0
	0.50	92.4	0.414	92.2	50.2	43.4	36.2	0.0	0.0	0.0	0.0	0.0
	0.60	88.0	0.470	88.0	55.2	50.2	41.6	0.0	0.0	0.0	0.0	0.0
	0.70	80.8	0.515	80.2	62.4	53.6	45.2	0.0	0.0	0.0	0.0	0.0
	0.80	76.2	0.538	75.4	65.8	55.8	46.8	0.0	0.0	0.0	0.0	0.0
	0.90	74.2	0.544	73.4	66.8	57.2	47.0	0.0	0.0	0.0	0.0	0.0
	1.00	74.2	0.544	73.4	66.8	57.2	47.0	0.0	0.0	0.0	0.0	0.0
MACER	0.00	99.0	0.000	99.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	0.10	95.8	0.328	96.0	43.6	33.2	23.6	0.0	0.0	0.0	0.0	0.0
	0.20	92.8	0.389	92.6	51.0	39.4	29.2	0.0	0.0	0.0	0.0	0.0
	0.30	90.2	0.438	90.0	56.4	43.8	33.4	0.0	0.0	0.0	0.0	0.0
	0.40	87.0	0.481	86.4	62.8	48.6	37.0	0.0	0.0	0.0	0.0	0.0
	0.50	82.2	0.504	81.4	67.6	51.4	38.0	0.0	0.0	0.0	0.0	0.0
	0.60	80.0	0.513	78.8	68.4	52.0	38.8	0.0	0.0	0.0	0.0	0.0
	0.70	79.0	0.516	77.6	69.0	52.2	39.0	0.0	0.0	0.0	0.0	0.0
	0.80	78.8	0.516	77.4	69.0	52.4	39.0	0.0	0.0	0.0	0.0	0.0
	0.90	78.8	0.516	77.4	69.0	52.4	39.0	0.0	0.0	0.0	0.0	0.0
	1.00	78.8	0.516	77.4	69.0	52.4	39.0	0.0	0.0	0.0	0.0	0.0
CONSISTENCY	0.00	99.0	0.000	99.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	0.10	97.6	0.254	97.6	31.8	27.0	20.6	0.0	0.0	0.0	0.0	0.0
	0.20	96.0	0.323	95.8	39.2	33.0	28.0	0.0	0.0	0.0	0.0	0.0
	0.30	93.8	0.383	93.8	46.4	39.8	33.0	0.0	0.0	0.0	0.0	0.0
	0.40	90.2	0.437	90.4	53.4	46.4	37.4	0.0	0.0	0.0	0.0	0.0
	0.50	85.0	0.485	85.0	59.0	50.0	41.4	0.0	0.0	0.0	0.0	0.0
	0.60	80.6	0.517	80.4	64.2	55.0	44.0	0.0	0.0	0.0	0.0	0.0
	0.70	78.0	0.530	77.4	65.8	56.0	44.4	0.0	0.0	0.0	0.0	0.0
	0.80	76.4	0.535	75.8	66.4	57.0	44.6	0.0	0.0	0.0	0.0	0.0
	0.90	76.0	0.535	75.2	66.4	57.0	44.6	0.0	0.0	0.0	0.0	0.0
	1.00	76.0	0.535	75.2	66.4	57.0	44.6	0.0	0.0	0.0	0.0	0.0

Table 9: Natural accuracy (NAC), average certified radius (ACR) and certified selection rate (portion of samples selected for the certification-network) at different radii on CIFAR10 with $\sigma_t = \sigma_\epsilon = 0.25$ for a range of threshold parameters θ and an ACES model with entropy selection, a `ResNet110` certification-network and an `LaNet` core-network.

Training	θ	NAC	ACR	Certified Selection Rate at Radius r								
				0.0	0.25	0.5	0.75	1.0	1.25	1.5	1.75	2.00
GAUSSIAN	0.00	99.0	0.000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	0.10	98.0	0.189	40.6	29.2	18.4	9.6	0.0	0.0	0.0	0.0	0.0
	0.20	96.4	0.247	51.0	36.4	25.2	14.0	0.0	0.0	0.0	0.0	0.0
	0.30	94.6	0.303	61.2	44.2	30.0	20.6	0.0	0.0	0.0	0.0	0.0
	0.40	90.4	0.358	76.0	56.8	40.6	28.6	0.0	0.0	0.0	0.0	0.0
	0.50	85.4	0.397	87.0	70.0	52.8	36.6	0.0	0.0	0.0	0.0	0.0
	0.60	81.6	0.416	94.4	82.2	67.0	48.4	0.0	0.0	0.0	0.0	0.0
	0.70	78.2	0.421	99.4	93.0	83.0	66.0	0.0	0.0	0.0	0.0	0.0
	0.80	77.8	0.422	100.0	98.6	94.6	86.2	0.0	0.0	0.0	0.0	0.0
	0.90	77.8	0.422	100.0	100.0	99.6	97.4	0.0	0.0	0.0	0.0	0.0
	1.00	77.8	0.422	100.0	100.0	100.0	100.0	0.0	0.0	0.0	0.0	0.0
SMOOTHADV	0.00	99.0	0.000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	0.10	98.8	0.161	23.4	19.8	17.6	13.6	0.0	0.0	0.0	0.0	0.0
	0.20	98.4	0.222	33.6	27.8	22.8	18.6	0.0	0.0	0.0	0.0	0.0
	0.30	97.4	0.288	43.4	36.8	30.0	25.0	0.0	0.0	0.0	0.0	0.0
	0.40	94.8	0.352	53.8	46.6	40.2	32.6	0.0	0.0	0.0	0.0	0.0
	0.50	92.4	0.414	63.4	56.8	50.0	42.8	0.0	0.0	0.0	0.0	0.0
	0.60	88.0	0.470	75.8	67.2	63.0	55.2	0.0	0.0	0.0	0.0	0.0
	0.70	80.8	0.515	90.2	84.2	78.2	72.0	0.0	0.0	0.0	0.0	0.0
	0.80	76.2	0.538	97.6	94.6	91.2	86.8	0.0	0.0	0.0	0.0	0.0
	0.90	74.2	0.544	100.0	99.8	98.2	96.2	0.0	0.0	0.0	0.0	0.0
	1.00	74.2	0.544	100.0	100.0	100.0	100.0	0.0	0.0	0.0	0.0	0.0
MACER	0.00	99.0	0.000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	0.10	95.8	0.328	59.2	44.2	33.4	23.6	0.0	0.0	0.0	0.0	0.0
	0.20	92.8	0.389	72.8	52.8	40.0	29.4	0.0	0.0	0.0	0.0	0.0
	0.30	90.2	0.438	81.6	61.4	45.6	33.8	0.0	0.0	0.0	0.0	0.0
	0.40	87.0	0.481	89.2	72.6	57.6	43.2	0.0	0.0	0.0	0.0	0.0
	0.50	82.2	0.504	94.8	84.0	67.6	53.4	0.0	0.0	0.0	0.0	0.0
	0.60	80.0	0.513	98.4	92.0	80.0	64.8	0.0	0.0	0.0	0.0	0.0
	0.70	79.0	0.516	99.8	97.2	91.0	78.0	0.0	0.0	0.0	0.0	0.0
	0.80	78.8	0.516	100.0	99.4	98.0	92.6	0.0	0.0	0.0	0.0	0.0
	0.90	78.8	0.516	100.0	100.0	99.8	99.0	0.0	0.0	0.0	0.0	0.0
	1.00	78.8	0.516	100.0	100.0	100.0	100.0	0.0	0.0	0.0	0.0	0.0
CONSISTENCY	0.00	99.0	0.000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	0.10	97.6	0.254	38.4	32.6	27.4	20.8	0.0	0.0	0.0	0.0	0.0
	0.20	96.0	0.323	50.8	40.6	33.8	28.4	0.0	0.0	0.0	0.0	0.0
	0.30	93.8	0.383	60.0	49.8	41.4	33.8	0.0	0.0	0.0	0.0	0.0
	0.40	90.2	0.437	71.4	62.0	53.0	44.0	0.0	0.0	0.0	0.0	0.0
	0.50	85.0	0.485	84.2	73.8	62.4	54.2	0.0	0.0	0.0	0.0	0.0
	0.60	80.6	0.517	91.8	85.4	77.4	68.0	0.0	0.0	0.0	0.0	0.0
	0.70	78.0	0.530	97.2	93.8	89.6	83.0	0.0	0.0	0.0	0.0	0.0
	0.80	76.4	0.535	99.4	99.2	97.0	93.6	0.0	0.0	0.0	0.0	0.0
	0.90	76.0	0.535	100.0	99.8	99.2	98.6	0.0	0.0	0.0	0.0	0.0
	1.00	76.0	0.535	100.0	100.0	100.0	100.0	0.0	0.0	0.0	0.0	0.0

Table 10: Natural accuracy (NAC), average certified radius (ACR) and certified accuracy at different radii on CIFAR10 with $\sigma_t = \sigma_\epsilon = 0.50$ for a range of threshold parameters θ and an ACES model with entropy selection, a `ResNet110` certification-network and an `LaNet` core-network.

Training	θ	NAC	ACR	Certified Accuracy at Radius r								
				0.0	0.25	0.5	0.75	1.0	1.25	1.5	1.75	2.00
GAUSSIAN	0.00	99.0	0.000	99.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	0.10	98.6	0.118	98.6	13.0	9.2	6.6	4.6	3.0	1.4	0.8	0.0
	0.20	98.2	0.187	98.2	20.2	15.2	10.4	7.6	4.8	3.0	1.2	0.0
	0.30	96.8	0.264	96.6	27.6	21.0	15.8	11.0	7.2	4.4	2.0	0.0
	0.40	94.6	0.347	94.0	34.6	27.2	21.0	15.8	10.8	6.4	2.8	0.0
	0.50	88.8	0.422	88.6	41.2	32.4	24.8	18.8	13.8	8.0	4.0	0.0
	0.60	82.2	0.484	80.0	47.2	37.8	28.8	21.4	15.0	9.2	4.8	0.0
	0.70	75.2	0.525	70.8	52.8	40.8	31.2	23.0	15.2	9.2	4.8	0.0
	0.80	70.8	0.532	65.8	54.2	41.4	31.8	23.2	15.2	9.2	4.8	0.0
	0.90	70.4	0.533	65.0	54.4	41.4	32.0	23.2	15.2	9.2	4.8	0.0
	1.00	70.4	0.533	65.0	54.4	41.4	32.0	23.2	15.2	9.2	4.8	0.0
SMOOTHADV	0.00	99.0	0.000	99.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	0.10	99.0	0.034	99.0	3.0	2.6	2.2	1.8	1.2	1.0	0.4	0.0
	0.20	99.0	0.060	99.0	4.4	4.2	3.2	3.2	2.8	2.0	1.8	0.0
	0.30	98.6	0.093	98.6	8.0	6.0	5.2	4.4	4.0	3.2	2.6	0.0
	0.40	98.4	0.135	98.4	10.2	9.6	8.2	6.6	5.4	4.4	3.6	0.0
	0.50	96.8	0.189	96.8	14.6	12.2	10.8	9.8	8.0	6.4	4.4	0.0
	0.60	95.0	0.277	95.0	21.8	18.8	15.8	13.6	11.6	9.6	7.2	0.0
	0.70	91.8	0.373	91.8	26.4	23.8	22.0	20.0	16.8	14.2	10.8	0.0
	0.80	85.4	0.499	85.2	36.0	31.4	28.8	25.2	22.8	19.4	16.2	0.0
	0.90	66.6	0.655	65.8	46.0	41.8	38.0	34.2	29.2	25.6	21.4	0.0
	1.00	58.6	0.721	57.0	51.0	45.8	42.2	37.8	32.2	27.8	22.2	0.0
MACER	0.00	99.0	0.000	99.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	0.10	97.4	0.233	97.0	22.4	18.2	13.6	9.2	6.8	5.0	3.4	0.0
	0.20	93.8	0.342	93.8	31.6	26.2	20.6	16.4	11.0	8.0	4.8	0.0
	0.30	90.8	0.437	89.4	39.6	32.2	26.4	20.8	15.8	11.0	6.6	0.0
	0.40	85.0	0.524	84.0	45.2	38.0	32.4	25.2	20.2	14.6	9.0	0.0
	0.50	76.2	0.590	72.8	50.4	42.4	35.6	29.0	23.2	17.0	10.4	0.0
	0.60	68.4	0.635	65.2	53.2	46.4	38.4	30.8	25.2	18.6	11.8	0.0
	0.70	65.4	0.658	62.0	54.0	48.2	40.0	32.8	26.4	19.4	11.8	0.0
	0.80	65.0	0.665	61.6	54.4	48.2	40.4	33.2	26.8	20.2	12.6	0.0
	0.90	65.0	0.665	61.6	54.4	48.2	40.4	33.2	26.8	20.2	12.6	0.0
	1.00	65.0	0.665	61.6	54.4	48.2	40.4	33.2	26.8	20.2	12.6	0.0
CONSISTENCY	0.00	99.0	0.000	99.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	0.10	98.2	0.157	98.2	13.8	10.8	9.6	7.0	5.8	4.4	3.0	0.0
	0.20	97.4	0.250	97.4	21.2	18.2	14.8	12.2	10.0	7.4	5.0	0.0
	0.30	96.4	0.342	96.4	27.6	23.4	21.0	17.0	13.4	11.0	7.8	0.0
	0.40	94.4	0.442	94.4	34.2	30.8	26.2	22.6	18.4	14.4	11.2	0.0
	0.50	89.4	0.534	89.0	41.8	38.2	30.8	26.4	22.2	18.4	12.8	0.0
	0.60	80.8	0.614	80.8	49.4	42.4	37.2	29.8	24.2	20.6	15.4	0.0
	0.70	73.4	0.667	72.8	54.6	46.6	40.6	32.2	26.8	21.2	16.4	0.0
	0.80	67.8	0.696	66.8	57.0	48.8	42.2	33.6	28.0	21.6	16.6	0.0
	0.90	65.8	0.701	64.8	58.0	49.0	42.4	33.6	28.0	21.6	16.6	0.0
	1.00	65.8	0.702	64.6	58.0	49.2	42.4	33.6	28.0	21.6	16.6	0.0

Table 11: Natural accuracy (NAC), average certified radius (ACR) and certified selection rate (portion of samples selected for the certification-network) at different radii on CIFAR10 with $\sigma_t = \sigma_\epsilon = 0.50$ for a range of threshold parameters θ and an ACES model with entropy selection, a `ResNet110` certification-network and an `LaNet` core-network.

Training	θ	NAC	ACR	Certified Selection Rate at Radius r								
				0.0	0.25	0.5	0.75	1.0	1.25	1.5	1.75	2.00
GAUSSIAN	0.00	99.0	0.000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	0.10	98.6	0.118	18.0	13.0	9.2	6.6	4.6	3.0	1.4	0.8	0.0
	0.20	98.2	0.187	27.8	20.6	15.4	10.4	7.6	4.8	3.0	1.2	0.0
	0.30	96.8	0.264	37.2	29.0	22.0	16.2	11.0	7.2	4.4	2.0	0.0
	0.40	94.6	0.347	47.8	37.8	30.4	25.2	19.2	13.6	8.6	4.2	0.0
	0.50	88.8	0.422	62.0	50.2	39.2	31.6	26.0	20.4	13.4	8.2	0.0
	0.60	82.2	0.484	80.6	66.8	57.6	46.4	35.0	27.4	22.4	13.4	0.0
	0.70	75.2	0.525	93.6	85.4	76.6	67.4	55.2	44.0	30.6	22.8	0.0
	0.80	70.8	0.532	99.2	97.0	92.6	85.4	79.2	70.8	55.6	40.4	0.0
	0.90	70.4	0.533	100.0	100.0	99.6	98.2	96.0	91.4	84.8	75.0	0.0
	1.00	70.4	0.533	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	0.0
SMOOTHADV	0.00	99.0	0.000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	0.10	99.0	0.034	3.2	3.0	2.6	2.2	1.8	1.2	1.0	0.4	0.0
	0.20	99.0	0.060	6.0	4.4	4.2	3.2	3.2	2.8	2.0	1.8	0.0
	0.30	98.6	0.093	9.4	8.0	6.0	5.2	4.4	4.0	3.2	2.6	0.0
	0.40	98.4	0.135	12.4	11.0	10.6	9.0	6.8	5.6	4.4	3.8	0.0
	0.50	96.8	0.189	20.2	16.8	13.6	12.0	10.8	8.8	7.2	5.4	0.0
	0.60	95.0	0.277	28.2	26.2	22.8	18.8	15.8	13.6	11.4	9.4	0.0
	0.70	91.8	0.373	38.0	32.8	31.4	28.2	25.6	22.0	18.6	14.8	0.0
	0.80	85.4	0.499	54.2	51.2	47.8	42.8	38.6	34.6	29.6	25.6	0.0
	0.90	66.6	0.655	85.4	82.2	78.2	73.4	68.8	61.8	57.4	52.6	0.0
	1.00	58.6	0.721	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	0.0
MACER	0.00	99.0	0.000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	0.10	97.4	0.233	31.6	23.4	19.0	14.0	9.4	6.8	5.0	3.4	0.0
	0.20	93.8	0.342	45.0	33.8	27.4	21.6	16.8	11.2	8.0	4.8	0.0
	0.30	90.8	0.437	57.8	46.4	35.2	28.6	22.4	16.8	11.4	7.2	0.0
	0.40	85.0	0.524	68.6	57.4	46.8	38.0	30.4	25.0	18.4	13.0	0.0
	0.50	76.2	0.590	85.8	71.0	60.4	49.4	39.8	32.6	26.6	17.4	0.0
	0.60	68.4	0.635	96.0	86.4	76.8	64.8	53.4	44.4	34.6	26.4	0.0
	0.70	65.4	0.658	99.4	97.0	92.2	83.2	75.2	62.8	50.2	36.8	0.0
	0.80	65.0	0.665	100.0	99.8	98.4	96.2	90.6	84.2	75.4	62.2	0.0
	0.90	65.0	0.665	100.0	100.0	100.0	100.0	99.2	97.6	93.8	85.4	0.0
	1.00	65.0	0.665	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	0.0
CONSISTENCY	0.00	99.0	0.000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	0.10	98.2	0.157	17.8	14.2	11.0	9.8	7.0	5.8	4.4	3.0	0.0
	0.20	97.4	0.250	26.6	22.2	19.2	15.6	12.6	10.2	7.4	5.0	0.0
	0.30	96.4	0.342	34.2	29.6	25.0	22.2	18.0	14.2	11.4	7.8	0.0
	0.40	94.4	0.442	44.6	38.0	34.0	29.6	26.6	22.2	17.6	14.0	0.0
	0.50	89.4	0.534	57.2	49.0	44.8	37.4	33.2	30.2	25.6	18.4	0.0
	0.60	80.8	0.614	72.8	66.6	58.6	52.0	45.4	39.2	35.4	29.0	0.0
	0.70	73.4	0.667	87.2	82.2	76.2	70.6	63.0	55.4	49.2	40.8	0.0
	0.80	67.8	0.696	96.8	93.6	90.6	88.0	83.2	76.0	71.2	62.6	0.0
	0.90	65.8	0.701	99.8	99.0	98.4	96.8	95.4	93.2	90.4	85.8	0.0
	1.00	65.8	0.702	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	0.0

Table 12: Natural accuracy (NAC), average certified radius (ACR) and certified accuracy at different radii on CIFAR10 with $\sigma_t = \sigma_\epsilon = 1.00$ for a range of threshold parameters θ and an ACES model with entropy selection, a `ResNet110` certification-network and an `LaNet` core-network.

Training	θ	NAC	ACR	Certified Accuracy at Radius r								
				0.0	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0
GAUSSIAN	0.00	99.0	0.000	99.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	0.10	98.8	0.020	98.8	1.4	0.8	0.2	0.0	0.0	0.0	0.0	0.0
	0.20	98.8	0.054	98.8	4.2	2.4	1.0	0.2	0.0	0.0	0.0	0.0
	0.30	98.4	0.093	98.2	6.2	4.4	2.4	1.0	0.2	0.0	0.0	0.0
	0.40	97.0	0.155	97.4	10.8	6.6	4.0	2.4	0.6	0.2	0.0	0.0
	0.50	95.4	0.239	95.4	14.8	10.0	6.6	4.0	1.8	0.4	0.0	0.0
	0.60	89.8	0.352	89.4	22.2	13.8	9.0	6.2	2.8	0.6	0.0	0.0
	0.70	78.6	0.458	75.8	27.6	19.0	12.2	8.0	4.0	1.8	0.0	0.0
	0.80	64.8	0.528	56.8	32.6	20.8	14.0	9.8	4.6	2.0	0.0	0.0
	0.90	56.4	0.537	46.8	33.6	21.2	14.0	10.0	4.6	2.0	0.0	0.0
	1.00	56.4	0.537	46.8	33.6	21.2	14.0	10.0	4.6	2.0	0.0	0.0
SMOOTHADV	0.00	99.0	0.000	99.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	0.10	99.0	0.000	99.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	0.20	99.0	0.004	99.0	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	0.30	99.0	0.016	99.0	1.4	0.6	0.0	0.0	0.0	0.0	0.0	0.0
	0.40	99.0	0.030	99.0	1.8	1.4	1.2	0.4	0.0	0.0	0.0	0.0
	0.50	98.6	0.048	98.6	2.6	2.4	1.8	1.2	0.2	0.0	0.0	0.0
	0.60	98.4	0.081	98.4	4.2	3.6	2.6	2.0	0.8	0.2	0.0	0.0
	0.70	97.6	0.150	97.4	7.0	5.6	4.4	3.4	2.4	1.2	0.4	0.0
	0.80	92.8	0.318	92.8	15.6	13.2	9.2	7.6	4.8	3.0	1.4	0.0
	0.90	82.6	0.536	82.6	24.2	20.2	16.4	13.0	9.0	6.2	4.0	0.0
	1.00	47.6	0.841	45.4	38.0	32.0	25.0	19.4	14.8	11.2	7.0	0.0
MACER	0.00	99.0	0.000	99.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	0.10	99.0	0.051	99.0	3.2	2.0	1.2	0.8	0.2	0.0	0.0	0.0
	0.20	98.8	0.123	98.8	8.0	4.6	3.2	1.8	1.0	0.6	0.0	0.0
	0.30	96.4	0.218	96.4	14.4	8.6	5.6	3.6	2.2	1.0	0.6	0.0
	0.40	92.0	0.308	92.0	17.2	13.8	8.8	5.6	3.2	1.6	0.8	0.0
	0.50	80.8	0.414	79.6	22.6	17.4	12.0	7.8	4.4	2.6	1.2	0.0
	0.60	67.0	0.522	65.4	28.4	21.6	15.2	10.8	6.0	4.0	1.8	0.0
	0.70	53.8	0.618	53.2	32.4	25.0	18.0	12.4	8.4	5.6	2.6	0.0
	0.80	45.6	0.715	44.4	36.4	29.4	22.0	14.6	10.8	6.4	3.4	0.0
	0.90	44.0	0.784	42.8	37.4	31.0	24.8	18.0	12.8	8.4	4.4	0.0
	1.00	44.0	0.796	42.8	37.4	31.0	25.0	18.4	13.8	9.0	4.8	0.0
CONSISTENCY	0.00	99.0	0.000	99.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	0.10	99.0	0.001	99.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	0.20	98.8	0.017	98.8	1.4	0.6	0.2	0.2	0.0	0.0	0.0	0.0
	0.30	98.6	0.047	98.6	3.2	2.4	1.2	0.4	0.2	0.2	0.0	0.0
	0.40	98.0	0.098	98.0	5.6	4.4	2.8	1.8	0.8	0.2	0.2	0.0
	0.50	96.6	0.184	96.6	10.2	8.0	5.4	4.2	1.6	0.6	0.6	0.0
	0.60	93.6	0.303	93.6	16.2	11.8	9.2	7.0	4.6	2.4	0.8	0.0
	0.70	88.8	0.443	89.0	21.6	17.8	12.8	10.0	7.6	5.0	2.6	0.0
	0.80	74.0	0.600	73.0	26.6	23.2	18.4	14.2	11.4	7.0	3.8	0.0
	0.90	52.8	0.734	51.2	33.6	27.4	20.6	16.8	13.4	9.8	4.6	0.0
	1.00	46.4	0.763	43.8	35.0	28.2	21.4	17.2	14.0	9.8	5.0	0.0

Table 13: Natural accuracy (NAC), average certified radius (ACR) and certified selection rate (portion of samples selected for the certification-network) at different radii on CIFAR10 with $\sigma_t = \sigma_\epsilon = 1.00$ for a range of threshold parameters θ and an ACES model with entropy selection, a `ResNet110` certification-network and an `LaNet` core-network.

Training	θ	NAC	ACR	Certified Selection Rate at Radius r								
				0.0	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0
GAUSSIAN	0.00	99.0	0.000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	0.10	98.8	0.020	4.2	1.4	0.8	0.2	0.0	0.0	0.0	0.0	0.0
	0.20	98.8	0.054	6.4	4.4	2.6	1.0	0.2	0.0	0.0	0.0	0.0
	0.30	98.4	0.093	11.4	6.4	4.6	2.6	1.0	0.2	0.0	0.0	0.0
	0.40	97.0	0.155	17.6	11.8	7.0	4.2	2.8	0.6	0.2	0.0	0.0
	0.50	95.4	0.239	26.6	17.8	11.8	7.4	4.6	2.2	0.4	0.2	0.0
	0.60	89.8	0.352	40.6	30.2	19.8	13.4	8.2	4.4	1.8	0.4	0.0
	0.70	78.6	0.458	62.4	47.8	35.6	25.0	15.2	10.2	5.4	1.4	0.0
	0.80	64.8	0.528	88.6	75.6	61.2	47.2	33.0	22.2	12.6	6.0	0.0
	0.90	56.4	0.537	100.0	97.0	93.2	85.4	70.6	54.8	38.2	22.2	0.0
	1.00	56.4	0.537	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	0.0
SMOOTHADV	0.00	99.0	0.000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	0.10	99.0	0.000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	0.20	99.0	0.004	1.0	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	0.30	99.0	0.016	1.8	1.4	0.6	0.0	0.0	0.0	0.0	0.0	0.0
	0.40	99.0	0.030	2.4	1.8	1.4	1.2	0.4	0.2	0.0	0.0	0.0
	0.50	98.6	0.048	4.0	2.8	2.4	1.8	1.2	0.4	0.2	0.2	0.0
	0.60	98.4	0.081	6.0	4.8	4.2	2.6	2.0	1.0	0.4	0.2	0.0
	0.70	97.6	0.150	12.6	8.2	6.4	5.0	4.0	3.2	1.6	0.6	0.0
	0.80	92.8	0.318	25.2	21.4	17.4	12.8	10.6	7.6	5.0	3.0	0.0
	0.90	82.6	0.536	46.6	39.4	33.2	29.8	25.8	20.8	16.0	11.8	0.0
	1.00	47.6	0.841	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	0.0
MACER	0.00	99.0	0.000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	0.10	99.0	0.051	5.6	3.2	2.0	1.2	0.8	0.2	0.0	0.0	0.0
	0.20	98.8	0.123	12.6	8.0	4.6	3.2	1.8	1.0	0.6	0.0	0.0
	0.30	96.4	0.218	20.8	15.4	9.4	5.6	3.6	2.2	1.2	0.6	0.0
	0.40	92.0	0.308	31.6	21.0	16.2	10.6	6.2	4.2	2.0	1.0	0.0
	0.50	80.8	0.414	50.6	31.8	22.8	15.2	10.2	5.6	3.6	1.4	0.0
	0.60	67.0	0.522	70.8	49.6	34.4	23.4	15.6	9.4	6.2	2.4	0.0
	0.70	53.8	0.618	87.0	70.8	53.0	37.4	24.8	16.2	10.2	5.4	0.0
	0.80	45.6	0.715	98.2	92.2	81.0	64.2	46.2	32.2	18.6	10.4	0.0
	0.90	44.0	0.784	100.0	99.6	98.6	93.0	83.2	64.6	44.4	29.4	0.0
	1.00	44.0	0.796	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	0.0
CONSISTENCY	0.00	99.0	0.000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	0.10	99.0	0.001	0.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	0.20	98.8	0.017	2.2	1.6	0.6	0.2	0.2	0.0	0.0	0.0	0.0
	0.30	98.6	0.047	4.4	3.4	2.6	1.4	0.6	0.2	0.2	0.0	0.0
	0.40	98.0	0.098	9.8	6.4	5.0	3.0	2.0	1.2	0.2	0.2	0.0
	0.50	96.6	0.184	15.8	12.2	9.0	6.2	5.0	2.2	1.0	0.6	0.0
	0.60	93.6	0.303	23.8	21.6	15.2	12.0	8.4	6.2	3.6	1.4	0.0
	0.70	88.8	0.443	35.2	30.4	27.4	21.2	15.2	12.2	8.6	5.0	0.0
	0.80	74.0	0.600	59.4	49.2	42.6	36.4	30.4	24.0	17.2	11.0	0.0
	0.90	52.8	0.734	91.4	83.0	75.8	66.4	57.6	50.2	41.8	29.6	0.0
	1.00	46.4	0.763	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	0.0

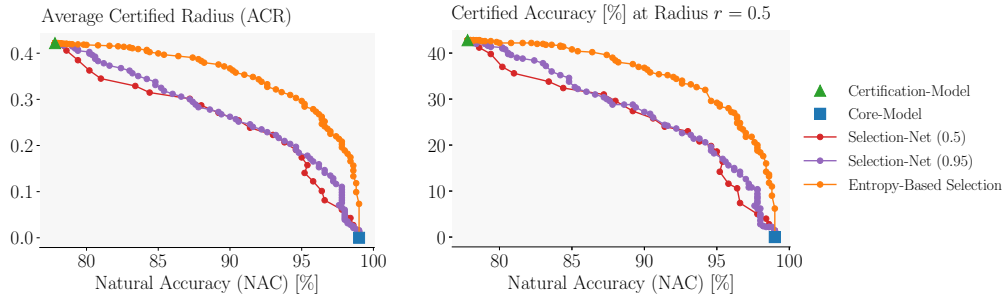

Figure 8: Comparing ACES with the entropy-based selection mechanism (orange) and selection networks (with $\eta \in \{0.5, 0.95\}$, red and purple) on CIFAR10 with respect to ACR (left) and certified accuracy at radius $r = 0.5$ (right) compared to natural accuracy. The certification network (`ResNet110` trained with GAUSSIAN, $\sigma = 0.25$) and core-network (`LaNet`) are fixed.

selection-mechanism abstaining. We generally observe that the selection-mechanism only abstains on very few samples. Further, for most samples and especially at high or low values of θ, (almost) all perturbations lead to the same selection decision and hence the mathematically maximal certified radius (for a given confidence and sample count). This is crucial, as the certified radius obtained for ACES is the minimum of those obtained for the certification-network and selection-mechanism.

D.3.2 TRAINING A SELECTION MODEL

Instead of using an entropy-based selection-mechanism as discussed in §3, we experimented with following Müller et al. (2021) in training a separate binary classifier on this selection task. To generate the labels, we first sample n perturbed instances of every training input and compute the corresponding prediction by the certification-network and determine the count of correct prediction n_y. We then threshold the accuracy of an individual sample over perturbations n_y/n with hyperparameter η to obtain the label $\mathbb{I}_{n_y/n>=\eta}$. We use these labels to then train a binary classifier of the same architecture and using the same training method as for the certification-network.

We instantiate this approach with $n = 1000$, $\eta \in \{0.5, 0.95\}$, and GAUSSIAN training and compare the obtained ACES models with ones using entropy-based selection in Table 14, visualized in Fig. 8. We observe that the entropy-based selection performs significantly better across all natural accuracies than this selection-network based approach. Additionally, the entropy-based mechanism does not need any additional training as it is based on the certification-network. Therefore, we focus all other analysis on entropy-based selection-mechanisms.

D.4 VARYING INFERENCE NOISE MAGNITUDE

Randomized smoothing is based on perturbing the inputs passed to an underlying model with random noise terms ϵ. Varying the magnitude of this noise is a natural way to trade-off robustness and accuracy, considered here as a baseline.

We first vary the evaluation noise level σ_ϵ and training noise level σ_t separately for SMOOTHADV trained `ResNet110` on CIFAR10 and observe that the best ACR is achieved when evaluating a model at (or close to) the noise magnitude it was trained with (see Tables 15 and 16). In Fig. 2, we illustrate a direct comparison of the thus obtained certified accuracies (dotted lines) with those of ACES models for `ResNet110` (solid lines) and `EfficientNet-B7` (dashed lines) core-networks. We generally observe that a) models trained with σ_t performs best with evaluation noise $\sigma_e \approx \sigma_t$ in all settings, except where σ_t is too small to mathematically allow for certification, and b) that reducing the inference noise magnitude often does not improve natural accuracy in sharp contrast to ACES models where much higher natural accuracies can be reached.

Based on this insight and due to the higher computational cost, we vary training and evaluation noise level σ jointly for ImageNet using CONSISTENCY training and show results in Table 17. Again, we observe that ACES models (orange and blue dots) outperform the thus obtained individual smoothed models (green triangles), reaching natural accuracies far beyond what individual smoothed models can, as is illustrated in Fig. 4. Only when purely optimizing for certified accuracy by setting $\theta = 1.0$ is

Table 14: Comparions of natural accuracy (NAC), average certified radius (ACR) and various certified radii via ACES on CIFAR10 with $\sigma_\epsilon = 0.25$. We consider selection networks trained with $\eta \in \{0.5, 0.95\}$ for various threshold parameters θ. All selection and certification networks have a `ResNet110` and all core models have a `LaNet` architecture, and the certification-network was trained with GAUSSIAN.

η	θ	NAC	ACR	Certified Accuracy at Radius r								
				0.0	0.25	0.5	0.75	1.0	1.25	1.5	1.75	2.00
0.50	0.00	99.0	0.000	99.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	0.10	95.2	0.141	95.2	21.4	14.2	8.0	0.0	0.0	0.0	0.0	0.0
	0.20	87.2	0.301	86.0	42.2	31.0	18.0	0.0	0.0	0.0	0.0	0.0
	0.30	77.8	0.422	75.4	60.0	42.8	25.6	0.0	0.0	0.0	0.0	0.0
	0.40	77.8	0.422	75.4	60.0	42.8	25.6	0.0	0.0	0.0	0.0	0.0
	0.50	77.8	0.422	75.4	60.0	42.8	25.6	0.0	0.0	0.0	0.0	0.0
	0.60	77.8	0.422	75.4	60.0	42.8	25.6	0.0	0.0	0.0	0.0	0.0
	0.70	77.8	0.422	75.4	60.0	42.8	25.6	0.0	0.0	0.0	0.0	0.0
	0.80	77.8	0.422	75.4	60.0	42.8	25.6	0.0	0.0	0.0	0.0	0.0
	0.90	77.8	0.422	75.4	60.0	42.8	25.6	0.0	0.0	0.0	0.0	0.0
	1.00	77.8	0.422	75.4	60.0	42.8	25.6	0.0	0.0	0.0	0.0	0.0
0.95	0.00	99.0	0.000	99.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	0.10	98.6	0.021	98.6	3.0	2.2	1.2	0.0	0.0	0.0	0.0	0.0
	0.20	98.0	0.061	98.0	9.0	6.2	2.4	0.0	0.0	0.0	0.0	0.0
	0.30	97.4	0.117	97.2	17.2	10.8	5.8	0.0	0.0	0.0	0.0	0.0
	0.40	95.4	0.178	95.2	26.4	17.0	9.4	0.0	0.0	0.0	0.0	0.0
	0.50	91.6	0.240	91.6	35.4	24.2	14.2	0.0	0.0	0.0	0.0	0.0
	0.60	87.4	0.295	86.8	41.4	30.0	18.0	0.0	0.0	0.0	0.0	0.0
	0.70	83.2	0.356	81.4	49.8	37.2	21.8	0.0	0.0	0.0	0.0	0.0
	0.80	80.2	0.403	78.0	56.4	41.2	25.2	0.0	0.0	0.0	0.0	0.0
	0.90	77.8	0.421	75.4	59.8	42.6	25.6	0.0	0.0	0.0	0.0	0.0
	1.00	77.8	0.422	75.4	60.0	42.8	25.6	0.0	0.0	0.0	0.0	0.0

Table 15: Varying the evaluation noise magnitude σ_ϵ for a `ResNet110` trained using SMOOTHADV and $\sigma_t \in \{0.25, 0.5\}$ or natural training $\sigma_t = 0.0$ on CIFAR10.

Training σ_t	Evaluation σ_ϵ	NAC	ACR	Radius r								
				0.0	0.25	0.5	0.75	1.0	1.25	1.5	1.75	2.0
0.00	0.125	**21.2**	0.038	**18.0**	7.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	0.25	13.0	0.027	13.0	5.4	0.4	0.0	0.0	0.0	0.0	0.0	0.0
	0.375	10.0	0.060	10.0	**9.6**	7.8	1.8	0.0	0.0	0.0	0.0	0.0
	0.5	11.0	0.030	11.4	6.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
	0.625	7.4	0.025	7.0	4.8	1.6	0.0	0.0	0.0	0.0	0.0	0.0
	0.75	8.4	0.047	8.6	7.0	5.2	1.8	0.6	0.0	0.0	0.0	0.0
	0.875	9.2	0.084	9.2	9.0	7.8	6.4	4.4	1.4	0.2	0.0	0.0
	1.0	9.2	0.165	9.2	9.2	9.2	9.2	9.0	8.6	7.0	6.2	2.8
	1.25	9.6	0.207	9.6	**9.6**	**9.6**	9.4	9.0	8.8	8.0	6.8	5.2
	1.5	9.6	**0.210**	9.6	**9.6**	**9.6**	**9.6**	**9.6**	**9.2**	**8.8**	**8.0**	**6.4**
0.25	0.125	72.0	0.301	71.8	63.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	0.25	**74.0**	**0.546**	**73.8**	**66.8**	**57.2**	**46.8**	0.0	0.0	0.0	0.0	0.0
	0.375	61.4	0.542	59.6	53.0	43.8	36.8	**30.2**	**22.2**	0.0	0.0	0.0
	0.5	43.6	0.379	41.6	35.2	29.4	24.4	17.8	12.0	7.8	3.0	0.0
	0.625	33.0	0.250	31.2	24.2	21.0	15.0	9.8	6.6	4.4	2.4	1.8
	0.75	25.4	0.191	24.0	19.6	15.0	11.6	8.4	4.2	2.4	1.6	0.6
	0.875	22.4	0.164	19.8	17.2	13.0	9.8	6.6	4.8	2.0	1.2	1.0
	1.0	18.8	0.154	17.4	15.6	11.8	8.8	6.0	4.6	2.4	1.0	1.0
	1.25	14.8	0.169	13.6	12.0	11.4	8.2	7.6	6.2	5.0	4.4	2.4
	1.5	11.6	0.233	11.2	10.4	10.2	10.2	9.8	9.6	**8.2**	**7.4**	**6.2**
0.50	0.125	52.0	0.224	52.0	46.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	0.25	53.0	0.416	52.2	47.6	43.6	39.2	0.0	0.0	0.0	0.0	0.0
	0.375	55.0	0.589	54.6	48.8	45.6	40.0	36.8	32.0	0.0	0.0	0.0
	0.5	**58.6**	0.726	**57.2**	**50.6**	**45.8**	**42.4**	**37.6**	**32.2**	**27.8**	**22.6**	0.0
	0.625	53.8	**0.729**	52.4	49.2	44.4	39.2	32.8	28.2	24.8	20.6	**16.4**
	0.75	45.6	0.599	43.4	38.0	35.6	31.4	27.4	22.6	18.6	15.0	10.8
	0.875	36.8	0.473	33.8	31.4	26.8	24.8	20.8	16.2	13.6	10.8	9.2
	1.0	30.4	0.390	28.8	25.2	21.4	18.4	15.8	12.6	10.6	9.4	8.2
	1.25	20.6	0.325	18.6	16.8	15.6	12.8	12.0	10.6	9.4	8.0	7.2
	1.5	14.0	0.334	12.8	12.2	11.8	11.8	11.2	10.6	9.8	9.2	8.6

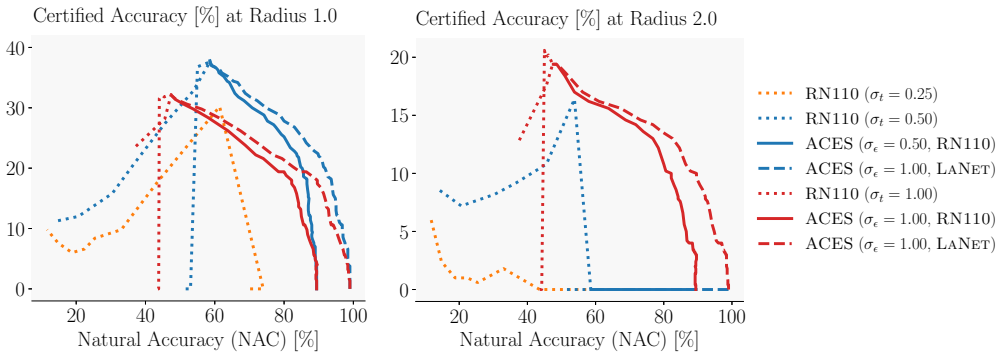

Figure 9: Certified accuracy at fixed radii over natural accuracy for ACES (solid and dashed lines) and individual smoothed models (dotted lines) using SMOOTHADV training. We compare stand alone `ResNet110` (dotted lines) trained with $\sigma_t \in \{0.25, 0.5, 1.0\}$ (orange, blue and red) and evaluated on a wide range of $\sigma_\epsilon \in [0.0, 1.5]$ with corresponding ACES models evaluated at training noise level and based on `ResNet110` for both certification- and core-networks (solid lines) or a `LaNet` core-network (dashed lines).

ACES outperformed by individual models, as the needed Bonferroni correction increases the required confidence leading to a slight drop in ACR from $0.512, 0.806$, and 1.023 to $0.509, 0.800$, and 0.997 for $\sigma_\epsilon = 0.25, 0.5$, and 1.00, respectively.

Table 16: Varying the evaluation noise magnitude σ_ϵ for a SMOOTHADV trained `ResNet110` with $\sigma_t = 1.0$ on CIFAR10.

Training σ_t	Evaluation σ_ϵ	NAC	ACR	Radius r								
				0.0	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0
1.00	0.125	43.6	0.193	43.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	0.25	43.8	0.359	43.6	37.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	0.375	44.0	0.501	43.8	37.8	31.6	0.0	0.0	0.0	0.0	0.0	0.0
	0.5	44.2	0.621	43.4	38.4	31.8	26.2	0.0	0.0	0.0	0.0	0.0
	0.625	45.0	0.716	44.0	39.6	32.0	**27.0**	**20.6**	0.0	0.0	0.0	0.0
	0.75	45.4	0.787	44.8	**39.8**	31.8	27.0	20.0	15.2	0.0	0.0	0.0
	0.875	46.2	0.832	**45.2**	38.6	32.0	26.6	20.2	**15.6**	11.0	0.0	0.0
	1.0	**47.4**	**0.844**	**45.2**	38.0	**32.2**	25.0	19.4	14.8	**11.4**	**7.4**	0.0
	1.25	45.6	0.762	42.2	33.8	28.2	22.2	17.6	13.0	9.4	4.8	**2.6**
	1.5	37.2	0.597	33.2	29.0	23.6	18.0	12.6	9.2	5.6	3.4	1.4

Table 17: Varying training and inference noise magnitude σ for individual CONSISTENCY trained `ResNet50` on ImageNet.

σ_ϵ	NAC	ACR	Radius r								
			0.0	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0
0.25	63.6	0.512	63.0	54.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.50	57.0	0.806	55.4	48.8	42.2	34.8	0.0	0.0	0.0	0.0	0.0
1.00	45.6	1.023	43.2	39.6	35.0	29.4	24.4	22.0	16.6	13.4	0.0