Supplementary Material for "Practical Evaluation of Adversarial Robustness via Adaptive Auto Attack"

A. Introduction

Due to the page limitation of the paper, we further illustrate our method in this supplementary material, which contains the following sections: 1). Detailed quantitative results of the diversified direction $\boldsymbol{w_d}$; 2). The results of the proposed A^3 attack across various defense strategies, datasets, network architectures and metrics.

B. Detailed quantitative results of the diversified direction w_d

In section 3.2 of the main paper, to illustrate that random sampling is sub-optimal, we use ODI [14] to attack 11 defense models, and only give the mean values of $\boldsymbol{w_d}$ at the $\hat{y}$-th (the misclassification label), and the y-th (the ground truth).

In order to observe the detailed quantitative results of the diversified direction $\boldsymbol{w_d}$. In this section, we use ODI [14] to attack 12 defense models, including AWP [17], Proxy [12], Fast [16], Feature Scatter [18], Geometry [21], HYDRA [13], Hypersphere [8], Interpolation [19], Regular [5], MART [15], MMA [1] and Pre-training [4]. The experiment settings are the same as the section 3.2 of the main paper. The CIFAR-10 dataset is used in this experiment, there are a total of 10 categories, with 9 error categories and one ground truth.

Among adversarial examples against different models, we summarize detailed statistic results of the direction of diversification $\boldsymbol{w_d}$ in Fig. 1 and Fig. 2. For each model, there are 9 rows, representing 9 error categories, where "1st" is the error category with the largest output logits, "9th" is the error category with the ninth largest output logits, and so on. There are 10 columns, representing 10 classes (9 error categories and 1 ground truth.), from "1st" to "9th" representing the 9 error categories and "GT" representing the ground truth. For the error categories, we arrange the error categories in descending order according to the output logits of each error category, where the output logits refer to the output logits of the clean example corresponding to the adversarial example we counted. The "i" row and "j" column represent the mean values of $\boldsymbol{w_d}$ on the "j" class when the adversarial example is misclassified as the error category with the "i" largest output logits. For all rows, we initialize their values to 0. We add up the $\boldsymbol{w_d}$ of all adversarial examples that are misclassified as the same row and average them. If none of the adversarial examples are misclassified as a error category, then the values of the corresponding row are 0.

From Fig. 1 and Fig. 2, we have the same observations as section 3.2 of the main paper: 1). The diversified direction $\boldsymbol{w_d}$ disobeys uniform distribution in all cases. 2). The diversified direction for each model has a model-specific bias in the positive/negative direction, specifically, as follows:

(a). The output logits of the error category increases, while the output logits of the ground truth decreases. For most models (e.g., AWP [17], Proxy [12], Fast [16], Geometry [21], HYDRA [13], Hypersphere [8], MART [15], Pre-training [4]), when an adversarial example is misclassified as an error category, the $\boldsymbol{w_d}$ for the error category is mostly positive, *i.e.*, the output logits of the error category increases, while the $\boldsymbol{w_d}$ for the ground truth is mostly negative, *i.e.*, the output logits of the ground truth decreases. This is intuitive because when the output logits of the error category of adversarial examples are greater than the output logits of the ground truth, then the examples are successfully attacked.

(b). The output logits of the error category increases, and the output logits of the ground truth also increases. However, there are some models whose $\boldsymbol{w_d}$ is counter-intuitive, such as Feature Scatter [18], Interpolation [19] and MMA [1]. When adversarial examples are misclassified as an error category, the $\boldsymbol{w_d}$ for the error category is positive, *i.e.*, the output logits of the error category increases, and the $\boldsymbol{w_d}$ for the ground truth is also positive, *i.e.*, the output logits of the ground truth also increases. Although this model has good adversarial robustness against weaker adversarial attack (*i.e.*, PGD [6]), it is poor in adversarial robustness against stronger attacks (*i.e.*, AA and A^3). A potential reason is that these defense models use gradient masks [7], and PGD chooses a bad starting point, which hinders the performance.

(c). The output logits of the error category decreases, and the output logits of the ground truth also increases. The most counter-intuitive is Regular [5], when an adver-

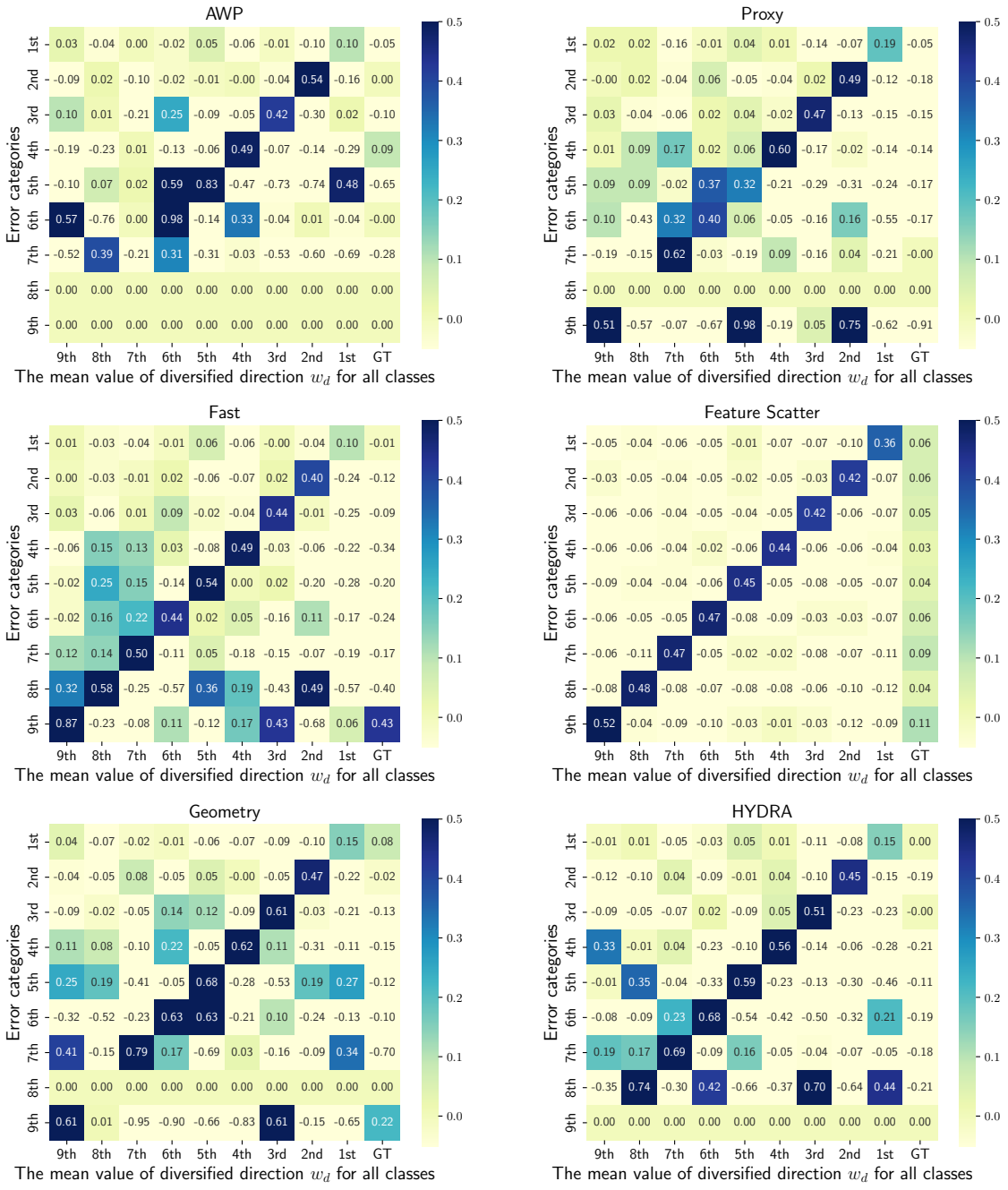

Figure 1. Quantitative statistical results of the diversified direction $\boldsymbol{w_d}$ of adversarial examples on multiple defense models (*i.e.*, AWP [17], Proxy [12], Fast [16], Feature Scatter [18], Geometry [21] and HYDRA [13].). The diversified direction of each model has a model-specific bias in the positive/negative direction. In other words, random sampling is suboptimal.

sarial example is misclassified as an error category, the $\boldsymbol{w_d}$ for the error category is negative, *i.e.*, the output logits of the error category decreases, and the $\boldsymbol{w_d}$ for the ground truth is positive, *i.e.*, the output logits of the ground truth increases. This model also uses gradient masks, which leads to extremely poor adversarial robustness of this model against stronger attacks.

Since the diversified initialization directions of models have some bias, and are not uniformly distributed, generating model-specific initial directions is very important and helps to obtain better performance.

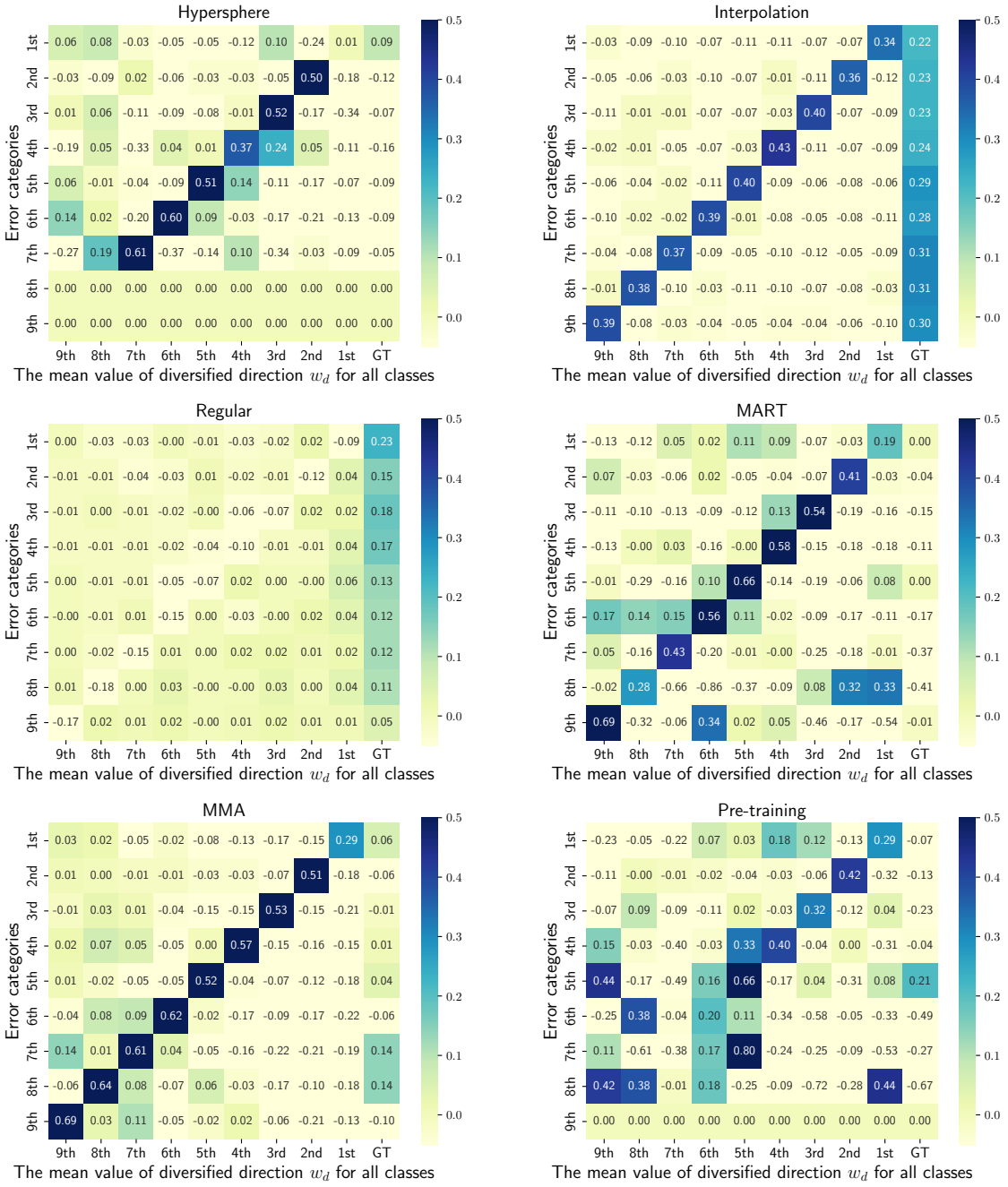

Figure 2. Quantitative statistical results of the diversified direction $\boldsymbol{w_d}$ of adversarial examples on multiple defense models (*i.e.*, Hypersphere [8], Interpolation [19], Regular [5], MART [15], MMA [1], Pre-training [4].). The diversified direction of each model has a model-specific bias in the positive/negative direction. In other words, random sampling is suboptimal.

C. Results of A^3 across various datasets, network architectures and metrics.

In this section, we show the results of the proposed A^3 attack across various defense strategies, datasets, network architectures and metrics. The setup is the same as section 4.1 of the main paper.

Results. As can be seen in Tab. 1. we show the effectiveness of the proposed A^3 across more datasets (e.g., MNIST, CIFAR10, and ImageNet), network architectures (e.g., VGG16, DenseNet161, ShuffleNet, etc.) and metrics (e.g., L_∞ and L_2). The experimental results show that A^3

Defense Method	Dataset	Metrics	Model	Clean	AA			A^3		
	number of test samples			acc	acc	→	←	**acc**	→	←
Undefended	ImageNet(5000)	$L_\infty(\epsilon = 4/255)$	ResNet50	76.74	0.0	0.40	0.39	**0.0**	**0.02(20.0×)**	**0.005(78.0×)**
DARI [11]	ImageNet(5000)	$L_\infty(\epsilon = 4/255)$	WideResNet-50-2	68.46	38.14	15.15	3.82	**38.12 ↓ 0.02**	**2.67(5.67×)**	**1.31(2.90×)**
DARI [11]	ImageNet(5000)	$L_\infty(\epsilon = 4/255)$	ResNet50	64.10	34.66	13.78	3.49	**34.64 ↓ 0.02**	**2.47(5.58×)**	**1.22(2.86×)**
DARI [11]	ImageNet(5000)	$L_\infty(\epsilon = 4/255)$	ResNet18	52.90	25.30	10.10	2.58	**25.16 ↓ 0.14**	**1.96(5.15×)**	**0.96(2.69×)**
DARI [11]	ImageNet(5000)	$L_2(\epsilon = 3.0)$	DenseNet161	66.14	36.52	14.51	3.67	**36.50 ↓ 0.02**	**2.59(5.60×)**	**1.28(2.87×)**
DARI [11]	ImageNet(5000)	$L_2(\epsilon = 3.0)$	VGG16-BN	56.24	29.62	11.79	2.99	**29.62 ↓ 0.00**	**2.20(5.36×)**	**1.08(2.77×)**
DARI [11]	ImageNet(5000)	$L_2(\epsilon = 3.0)$	ShuffleNet	43.16	17.64	7.08	1.85	**17.56 ↓ 0.08**	**1.58(4.48×)**	**0.78(2.37×)**
DARI [11]	ImageNet(5000)	$L_2(\epsilon = 3.0)$	MobileNet-V2	49.62	24.78	9.89	2.52	**24.74 ↓ 0.04**	**1.94(5.10×)**	**0.95(2.65×)**
Fixing Data [9]	CIFAR10(10000)	$L_2(\epsilon = 0.5)$	WideResNet-28-10	91.79	78.80	62.00	15.20	**78.79 ↓ 0.01**	**5.35(11.59×)**	**2.63(5.78×)**
Robustness [2]	CIFAR10(10000)	$L_2(\epsilon = 0.5)$	ResNet50	90.83	69.23	54.56	13.45	**69.21 ↓ 0.02**	**4.72(11.56×)**	**2.32(5.80×)**
Proxy [12]	CIFAR10(10000)	$L_2(\epsilon = 0.5)$	WideResNet-34-10	90.31	76.11	59.89	14.69	**76.10 ↓ 0.01**	**5.18(11.56×)**	**2.55(5.76×)**
Overfitting [10]	CIFAR10(10000)	$L_2(\epsilon = 0.5)$	ResNet18	88.67	67.68	53.34	13.15	**67.64 ↓ 0.04**	**4.61(11.57×)**	**2.27(5.79×)**
ULAT [3]	MNIST(10000)	$L_\infty(\epsilon = 0.3)$	WideResNet-28-10	99.26	96.34	76.05	18.44	**96.31 ↓ 0.03**	**6.53(11.64×)**	**3.22(5.71×)**
TRADES [20]	MNIST(10000)	$L_\infty(\epsilon = 0.3)$	SmallCNN	99.48	92.76	73.12	17.88	**92.71 ↓ 0.05**	**6.33(11.55×)**	**3.12(5.73×)**

Table 1. The results of the proposed A^3 attack across various defense strategies, datasets, network architectures and metrics. The "acc" column shows the robust accuracies of different models. The "→" column shows the iteration number of forward propagation (million), while the "←" column shows the iteration number of backward propagation (million). The "**acc**" column of A^3 shows the difference between the robust accuracies of AA and A^3, the "←" and "→" columns of A^3 show the speedup factors of A^3 relative to AA.

is better than AA on various datasets, model architectures and metrics.

References

[1] Gavin Weiguang Ding, Yash Sharma, Kry Yik Chau Lui, and Ruitong Huang. MMA training: Direct input space margin maximization through adversarial training. In *ICLR*, 2020. 1, 3

[2] Logan Engstrom, Andrew Ilyas, Hadi Salman, Shibani Santurkar, and Dimitris Tsipras. Robustness (python library), 2019. 4

[3] Sven Gowal, Chongli Qin, Jonathan Uesato, Timothy A. Mann, and Pushmeet Kohli. Uncovering the limits of adversarial training against norm-bounded adversarial examples. *CoRR*, abs/2010.03593, 2020. 4

[4] Dan Hendrycks, Kimin Lee, and Mantas Mazeika. Using pre-training can improve model robustness and uncertainty. In *ICML*, 2019. 1, 3

[5] Charles Jin and Martin Rinard. Manifold regularization for adversarial robustness. *CoRR*, abs/2003.04286, 2020. 1, 3

[6] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018. 1

[7] Linh Nguyen, Sky Wang, and Arunesh Sinha. A learning and masking approach to secure learning. In *GameSec*, 2018. 1

[8] Tianyu Pang, Xiao Yang, Yinpeng Dong, Taufik Xu, Jun Zhu, and Hang Su. Boosting adversarial training with hypersphere embedding. In *NeurIPS*, 2020. 1, 3

[9] Sylvestre-Alvise Rebuffi, Sven Gowal, Dan A. Calian, Florian Stimberg, Olivia Wiles, and Timothy A. Mann. Fixing data augmentation to improve adversarial robustness. *CoRR*, abs/2103.01946, 2021. 4

[10] Leslie Rice, Eric Wong, and J. Zico Kolter. Overfitting in adversarially robust deep learning. In *ICML*, 2020. 4

[11] Hadi Salman, Andrew Ilyas, Logan Engstrom, Ashish Kapoor, and Aleksander Madry. Do adversarially robust imagenet models transfer better? In *NeurIPS*, 2020. 4

[12] Vikash Sehwag, Saeed Mahloujifar, Tinashe Handina, Sihui Dai, Chong Xiang, Mung Chiang, and Prateek Mittal. Improving adversarial robustness using proxy distributions. *CoRR*, abs/2104.09425, 2021. 1, 2, 4

[13] Vikash Sehwag, Shiqi Wang, Prateek Mittal, and Suman Jana. HYDRA: pruning adversarially robust neural networks. In *NeurIPS*, 2020. 1, 2

[14] Yusuke Tashiro, Yang Song, and Stefano Ermon. Diversity can be transferred: Output diversification for white- and black-box attacks. *arXiv preprint arXiv:2003.06878*, 2020. 1

[15] Yisen Wang, Difan Zou, Jinfeng Yi, James Bailey, Xingjun Ma, and Quanquan Gu. Improving adversarial robustness requires revisiting misclassified examples. In *ICLR*, 2020. 1, 3

[16] Eric Wong, Leslie Rice, and J. Zico Kolter. Fast is better than free: Revisiting adversarial training. In *ICLR*, 2020. 1, 2

[17] Dongxian Wu, Shu-Tao Xia, and Yisen Wang. Adversarial weight perturbation helps robust generalization. In *NeurIPS*, 2020. 1, 2

[18] Haichao Zhang and Jianyu Wang. Defense against adversarial attacks using feature scattering-based adversarial training. In *NeurIPS*, 2019. 1, 2

[19] Haichao Zhang and Wei Xu. Adversarial interpolation training: A simple approach for improving model robustness. 2019. 1, 3

[20] 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 *ICML*, pages 7472–7482, 2019. 4

[21] Jingfeng Zhang, Jianing Zhu, Gang Niu, Bo Han, Masashi Sugiyama, and Mohan S. Kankanhalli. Geometry-aware instance-reweighted adversarial training. In *ICLR*, 2021. 1, 2