APPENDIX: EMPIRICAL STUDY OF THE DECISION REGION AND ROBUSTNESS IN DEEP NEURAL NETWORKS

Anonymous authors
Paper under double-blind review

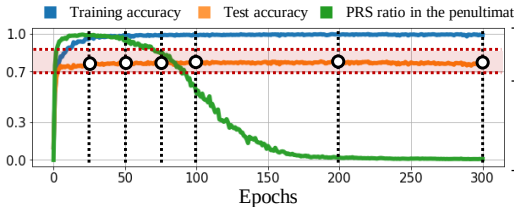
A EXPERIMENT SETUP

In this section, we describe the details to build the model candidates used in Figure 4, 8, 10, and 12 in the main paper. We first select three different structures; (1) convolutional neural network with 6 convolutional blocks (CNN-6), (2) VGG-16, and (3) ResNet-18 and three datasets; (1) MNIST, (2) F-MNIST, and (3) CIFAR-10. In particular, we use bilinear upsampled MNIST and F-MNIST dataset ($28 \times 28 \rightarrow 32 \times 32$). The detailed structure of CNN-6 is described in Table 1. For VGG-16 and ResNet-18, we maintain the original structure[1] for the feature extraction and replace the classifier as same as CNN-6.

Table 1: The architecture of CNN-6.

	Type	Channels/size	Activation
Convolutional Layer	Conv2d [3×3]	32	ReLU
	Conv2d [3×3], MaxPool(2)	64	
	Conv2d [3×3]	128	
	Conv2d [3×3], MaxPool(2)	128	
	Conv2d [3×3], MaxPool(2)	64	
	Conv2d [3×3]	64	
Classifier	Linear	512	ReLU
		128	
		10	

We train[2] basic models with fixed five random seeds (123, 375, 574, 907 and 981) and four batch sizes (64, 128, 512 and 2048). Finally we obtain 20 basic models for each dataset and structure.



Model	Seed	BS	MNIST 25	50	75	100	200	300
CNN-6	375	64	O	O	O	O	O	O
		128	O	O	O	O	O	O
		512	O	O	O	O	O	O
		2048	O	O	O	O	O	O

Part of Table

Figure 14: An illustrative example for the selected candidates and the corresponding part of Table 2. The example corresponds to the CNN-6 with random seed 375, batch size (BS) 128 and MNIST.

For the extensive analysis on the correlation between PRS ratio and properties of network, we extract candidates from each basic model with the grid of epochs ($[25, 50, 75, 100, 200, 300]$). Then we apply the test accuracy ($acc.$) threshold (MNIST: $98\% \leq acc.$, F-MNIST: $90\% \leq acc. \leq 93\%$, and CIFAR-10: $72\% \leq acc. \leq 78\%$) to guarantee the sufficient performance. Figure 14 presents the illustrative example for candidates selection. Table 2 presents used candidates for the experiments. We also provide the statistics of test accuracy for the candidates over each selected epoch in Appendix B.

[1] We use the officially provided network from the Pytorch: https://pytorch.org/vision/stable/models.html

[2] Cross-entropy loss and Adam optimizer with learning rate 10^{-3} is used.

Table 2: Used candidates of models for all experiments in the main paper. Pink box indicates the models which does not satisfy the described condition. The orange/red box indicates network A/B denoted in the main paper respectively. The blue box is described example in Appendix B. BS indicates the batch size.

			MNIST						F-MNIST						CIFAR10					
Model	Seed	BS	25	50	75	100	200	300	25	50	75	100	200	300	25	50	75	100	200	300
CNN-6	123	64	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	x
		128	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	x	o	o
		512	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o
		2048	o	o	o	o	o	o	o	x	o	o	o	o	x	x	x	x	o	o
	375	64	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	x
		128	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o
		512	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o
		2048	o	o	o	o	o	o	o	o	o	o	o	o	x	x	x	o	o	o
	574	64	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	x
		128	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	x	o
		512	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o
		2048	o	o	o	o	o	o	o	o	o	o	o	o	x	x	o	o	o	o
	907	64	o	o	o	o	o	o	o	o	o	o	o	x	o	o	o	o	o	x
		128	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	x	o	o
		512	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o
		2048	o	o	o	o	o	o	o	o	o	o	o	o	x	x	x	o	o	o
	981	64	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	x
		128	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o
		512	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o
		2048	o	o	o	o	o	o	o	o	o	o	o	o	x	x	o	o	o	o
ResNet-18	123	64	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o
		128	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o
		512	o	o	o	o	o	o	x	o	o	o	o	o	o	o	o	o	o	o
		2048	o	o	o	o	o	o	x	o	o	o	o	o	x	x	x	x	x	o
	375	64	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o
		128	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o
		512	o	o	o	o	o	o	x	o	o	o	o	o	x	o	o	o	o	o
		2048	o	o	o	o	o	o	x	o	o	o	o	o	x	x	x	x	o	x
	574	64	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o
		128	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o
		512	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o
		2048	o	o	o	o	o	o	x	o	o	o	o	o	x	x	x	x	x	o
	907	64	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o
		128	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o
		512	o	o	o	o	o	o	x	o	o	o	o	o	o	o	o	o	o	o
		2048	o	o	o	o	o	o	x	x	o	o	o	o	x	x	x	x	o	o
	981	64	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o
		128	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o
		512	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o
		2048	o	o	o	o	o	o	x	o	o	o	o	o	x	x	x	x	x	o
VGG-16	123	64	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	x
		128	o	o	o	o	o	o	o	o	o	o	o	o	o	x	x	x	x	x
		512	o	o	o	o	o	o	o	o	o	o	x	x	o	o	o	x	x	x
		2048	o	o	o	o	o	o	o	o	o	o	o	o	x	x	o	o	o	o
	375	64	o	x	o	o	o	o	o	o	o	o	o	x	o	o	o	x	x	x
		128	o	o	o	o	o	o	o	o	o	o	o	x	x	o	o	o	o	x
		512	o	o	o	o	o	o	o	o	o	o	o	o	x	o	o	o	o	x
		2048	o	o	o	o	o	o	o	o	o	o	o	o	x	x	x	x	x	x
	574	64	o	o	o	o	o	o	o	o	o	o	o	o	x	o	o	o	o	x
		128	o	o	o	o	o	x	o	o	o	o	x	o	o	o	o	o	o	x
		512	o	o	o	o	o	o	o	o	o	o	o	o	x	x	x	x	o	o
		2048	o	o	o	o	o	o	o	o	o	o	o	o	x	x	x	x	x	o
	907	64	o	o	o	o	o	o	o	o	o	x	o	o	o	o	o	o	x	x
		128	o	o	o	o	o	o	o	o	o	o	o	x	x	x	o	x	x	x
		512	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	x	x	x
		2048	x	x	o	o	o	o	o	o	o	o	o	o	x	o	o	o	o	o
	981	64	o	o	o	x	o	x	o	o	o	o	x	o	x	o	x	x	x	x
		128	o	o	o	o	o	o	o	o	o	o	o	o	o	x	x	x	x	x
		512	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	x	x	x
		2048	o	o	o	o	o	o	o	o	o	o	o	o	x	o	o	o	o	o

B MODEL PERFORMANCE

In this section, we provide the statistics of test accuracy for the selected candidates in Table 2. We average the test accuracy for random seeds. For example, the mean and standard deviation in the 25th epoch and batch size (BS) 64 of VGG-16 on MNIST are calculated from blue boxes in Table 2. Dash line indicates that there are no models satisfy described conditions in random seeds.

Table 3: Test accuracy (%) for candidates over random seeds.

		Epoch	25	50	75	100	200	300
		BS	Test Accuracy (%)					
MNIST	CNN-6	64	99.21±0.13	99.13±0.12	99.25±0.10	99.34±0.05	99.23±0.22	99.21±0.15
		128	99.18±0.09	99.21±0.08	99.37±0.08	99.36±0.05	99.43±0.05	99.44±0.05
		512	99.05±0.18	99.21±0.09	99.20±0.22	99.40±0.05	99.42±0.06	99.42±0.03
		2048	98.88±0.16	99.06±0.11	99.04±0.12	99.05±0.11	99.29±0.08	99.30±0.08
	VGG-16	64	99.11±0.31	99.06±0.43	99.21±0.39	99.05±0.55	99.28±0.18	99.31±0.20
		128	99.20±0.34	99.22±0.19	99.22±0.30	99.35±0.15	99.25±0.31	99.26±0.37
		512	99.20±0.12	99.34±0.17	99.36±0.13	99.46±0.09	99.47±0.06	99.46±0.07
		2048	99.10±0.24	99.27±0.06	99.30±0.08	99.21±0.39	99.36±0.12	99.35±0.12
	ResNet-18	64	99.03±0.12	99.24±0.05	99.28±0.05	99.30±0.08	99.30±0.08	99.30±0.10
		128	99.03±0.22	99.15±0.08	99.32±0.05	99.34±0.06	99.34±0.06	99.35±0.06
		512	98.89±0.19	99.13±0.13	99.27±0.19	99.42±0.03	99.44±0.03	99.44±0.03
		2048	98.76±0.34	99.20±0.17	99.34±0.06	99.35±0.06	99.35±0.06	99.35±0.06
F-MNIST	CNN-6	64	91.72±0.37	91.81±0.23	91.83±0.20	91.97±0.17	91.80±0.31	90.82±0.35
		128	91.52±0.19	91.77±0.28	91.95±0.19	92.04±0.26	92.16±0.14	91.73±0.82
		512	91.05±0.24	91.04±0.39	91.52±0.55	91.71±0.27	91.92±0.28	92.49±0.20
		2048	90.43±0.29	90.98±0.23	90.87±0.39	91.06±0.33	91.30±0.24	91.58±0.28
	VGG-16	64	91.92±0.22	92.30±0.65	92.67±0.17	92.11±0.27	92.63±0.23	92.41±0.48
		128	92.14±0.40	92.54±0.44	92.57±0.16	92.71±0.16	92.81±0.17	92.91±0.05
		512	91.74±0.48	92.08±0.25	92.24±0.25	92.62±0.32	92.82±0.05	92.80±0.06
		2048	91.09±0.45	92.03±0.33	92.30±0.25	92.39±0.33	92.50±0.07	92.57±0.29
	ResNet-18	64	90.60±0.24	90.96±0.28	91.02±0.12	91.05±0.15	91.04±0.17	91.01±0.44
		128	90.47±0.22	90.81±0.31	91.00±0.21	91.08±0.23	91.13±0.31	91.23±0.29
		512	90.27±0.25	90.53±0.21	90.57±0.19	90.69±0.23	90.94±0.39	91.19±0.25
		2048	-	90.17±0.23	90.56±0.26	90.41±0.13	90.61±0.33	90.49±0.23
CIFAR-10	CNN-6	64	75.76±1.32	76.39±1.06	77.28±0.45	76.88±0.45	75.76±0.51	-
		128	76.54±0.80	76.88±0.55	76.32±0.41	77.18±0.22	76.42±0.63	76.49±0.87
		512	73.17± 0.97	74.69± 1.17	74.55± 1.49	75.18±0.86	75.35±0.90	75.79±1.11
		2048	-	-	72.79±0.32	73.73±0.56	74.52±0.93	74.59±0.95
	VGG-16	64	74.38±0.96	76.10±1.06	76.73±0.54	77.37±0.61	77.86±0.11	-
		128	75.59±1.90	74.53±0.52	76.44±1.47	76.18±1.05	77.00±1.20	-
		512	75.82±1.01	76.58±1.24	76.98±0.93	76.23±0.00	75.16±1.72	74.87±0.00
		2048	-	74.37±0.71	73.74±1.58	74.94±1.01	75.79±0.77	76.26±1.34
	ResNet-18	64	75.71±0.39	75.98±0.41	76.41±0.75	76.44±0.15	76.83±0.25	76.30±0.52
		128	74.79±0.53	75.75±0.56	76.35± 0.66	76.18±0.52	76.39±0.69	76.32±0.19
		512	72.63±0.45	73.44±0.73	74.20±0.50	74.73±0.61	75.16±0.14	75.38±0.93
		2048	-	-	-	-	72.08±0.00	72.53±0.62

C FAILED ATTACK EXAMPLES

In this section, we provide more failed attack examples described in Section 4.2 of the main paper. We note that the examples are only attacked successfully on Network A.

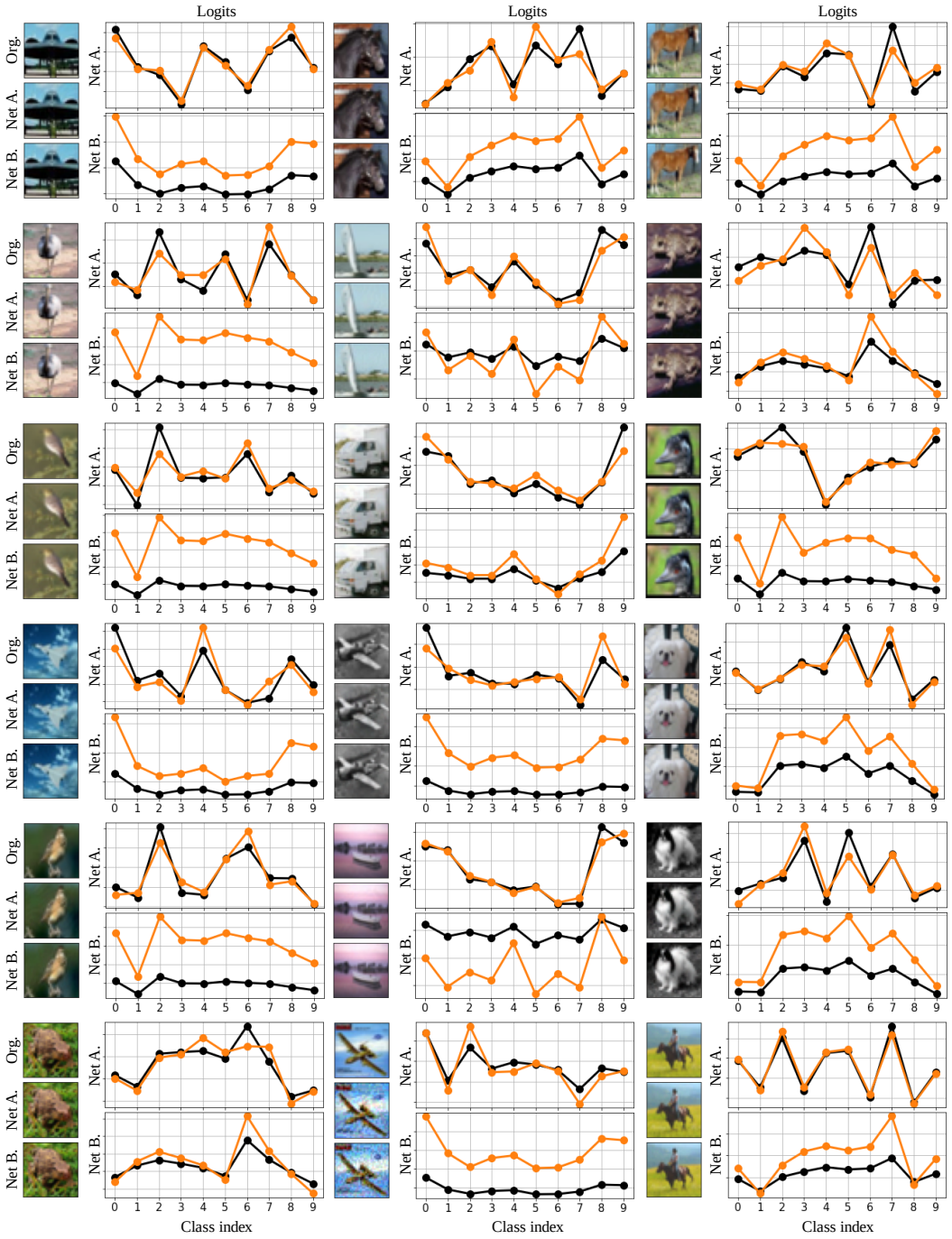


Figure 15: Randomly selected examples for successful attack on Network A but failed attack on Network B. The black line is the original predicted logits and the orange line is the changed logits after attack.

D FEATURE REPRESENTATION ANALYSIS

To quantify the denseness of feature representation which explains how much the samples from the same class are grouped in the feature space on the penultimate layer, we measure the ratio of same labels between K nearest neighbor samples for the given input. Let the training dataset $\mathbf{X}$ and data pair $(x_i, y_i) \in \mathbf{X}$. The denseness for the given data pair (x, y) is defined as,

$$D(\mathbf{X}, K) = \frac{1}{|\mathbf{X}|} \sum_{x,y \in \mathbf{X}} \frac{1}{K} \sum_{i \in \text{NN}_K(x, \mathbf{X})} \mathbf{I}(y_i = y),$$

where $\text{NN}_K(x, \mathbf{X})$ returns indices of K nearest neighbors for feature representation of the given input $f_{L-1:1}(x)$ from training dataset $\mathbf{X}$ and $\mathbf{I}(\cdot)$ is the indicator function. Table 4 presents the average denseness for various random seeds mentioned in Appendix A with various K. The experiment is performed on CNN-6 with a low/high PRS ratio[3]. We identify that low PRS ratio cases (a1, b1, c1) have higher denseness compared to the high PRS ratio cases in almost K. For MNIST dataset, it is difficult to discriminate the difference of denseness along the magnitude of PRS ratio.

Table 4: Average of denseness in feature space for models with a low PRS ratio and a high PRS ratio. We trained CNN-6 model on three datasets, (a) MNIST, (b) F-MNIST and (c) CIFAR-10. Odd rows (a1, b1, c1) indicate the model with a low PRS ratio and even rows (a2, b2, c2) indicate the model with a high PRS ratio.

		K				
	PRS Ratio	10	50	100	200	300
(a1)	0.035±0.029	**0.999±0.001**	**0.998±0.001**	0.997±0.001	0.996±0.002	0.995±0.002
(a2)	0.669±0.088	0.999±0.001	**0.998±0.001**	**0.998±0.001**	**0.997±0.002**	**0.996±0.002**
(b1)	0.022±0.007	**0.981±0.006**	**0.972±0.008**	**0.966±0.009**	**0.957±0.010**	**0.949±0.011**
(b2)	0.946±0.018	0.969±0.003	0.955±0.003	0.948±0.004	0.938±0.004	0.932±0.004
(c1)	0.006±0.000	**0.945±0.006**	**0.920±0.007**	**0.903±0.007**	**0.877±0.007**	**0.857±0.007**
(c2)	1.000±0.000	0.939±0.007	0.904±0.008	0.883±0.009	0.855±0.010	0.836±0.011

[3]We use trained model with batch size of 128 and 2048 to control the PRS ratio

E SIGNIFICANCE TESTS ON THE RELATIONSHIPS

In this section, we provide the quantitative results for the calculated regression lines (in Figure 4, 8, 10, and 12 of the main paper) which represents the relationship between the PRS ratio and each property. We provide (1) the slope of the regression line, and (2) P-value of the fitted line to verify the significance. Table 5 presents the result of slope in each case. The inverse correlation is observed in most investigated properties. The relationship between PRS ratio and sparsity on ResNet-18 shows independence in all datasets. We conjecture that skip connection or batch normalization can cause this phenomenon. Table 6 presents the result of P-value[4] from the significance test of the slope of the regression line. We can identify that most slopes are statistically significant (P-value < 0.05).

Table 5: Slopes of the each regression line between the PRS ratio and properties. FGSM indicates the robust score based on FGSM defined in the main paper. PGD indicates the robust accuracy based on PGD attack with fixed ϵ (MNIST=0.3, F-MNIST=0.1, and CIFAR10=0.0313). IR/CS indicates inclusion ratio and cosine similarity respectively.

		FGSM/PRS	PGD/PRS	Sparsity/PRS	IR/PRS	CS/PRS
CNN-6	MNIST	-0.46	-0.30	-0.49	-0.72	-0.76
	F-MNIST	-0.36	-0.45	-0.23	-0.79	-0.58
	CIFAR-10	-0.11	-0.23	-0.15	-0.80	-0.65
VGG-16	MNIST	-7.72	-14.32	-1.99	-0.53	-14.70
	F-MNIST	-1.89	-2.52	-2.24	-0.48	-3.38
	CIFAR-10	-0.52	-0.86	-0.90	-0.79	0.28
ResNet-18	MNIST	-1.39	-0.80	0.05	-0.53	-2.88
	F-MNIST	-1.08	-0.90	0.08	-0.65	-2.24
	CIFAR-10	-0.18	-0.11	-0.03	-0.71	-1.29

Table 6: P-value from the significance tests for each regression line.

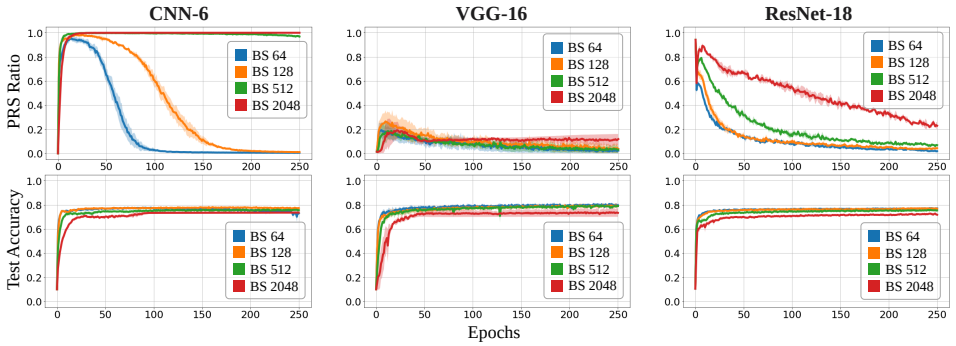
		FGSM/PRS	PGD/PRS	Sparsity/PRS	IR/PRS	CS/PRS
CNN-6	MNIST	2.53E-11	9.43E-05	3.87E-48	1.09E-126	5.17E-29
	F-MNIST	1.61E-21	1.28E-44	9.84E-13	1.78E-133	2.43E-46
	CIFAR-10	2.83E-12	2.36E-22	1.92E-11	5.45E-136	4.35E-54
VGG-16	MNIST	1.90E-03	7.01E-05	2.91E-06	4.66E-43	2.70E-08
	F-MNIST	7.93E-12	4.38E-18	3.79E-43	2.22E-24	4.11E-19
	CIFAR-10	3.60E-05	2.43E-07	9.23E-12	2.07E-09	2.92E-01
ResNet-18	MNIST	1.73E-02	1.38E-02	4.16E-01	1.58E-98	1.36E-11
	F-MNIST	4.08E-05	3.25E-05	1.96E-02	6.38E-38	1.66E-16
	CIFAR-10	1.50E-18	5.23E-13	2.87E-03	2.87E-35	4.20E-14

[4]Computed by *statsmodel* package

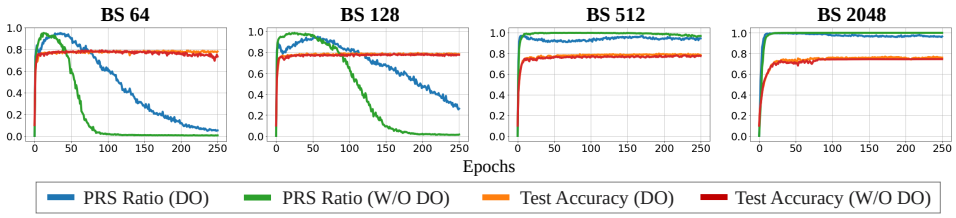
F THE FACTORS WHICH AFFECT THE PRS RATIO

We perform an ablation study on the factors which affects the PRS ratio. First, we identify the relationship between the PRS ratio and batch size (BS) which is one of the factors we used to make the candidates in section A. Figure 16 (a) presents the PRS ratio and test accuracy for training epochs in different BS (64, 128, 512, 2048) and different networks (CNN-6, VGG-16, ResNet-18) on CIFAR-10. In Figure 16 (a), we observe that BS is proportional to the PRS ratio (i.e., The large BS causes the high PRS ratio). Previous work (Yao et al., 2018) provides that training with large batch size can degrade the robustness of the model against the adversarial attack, which is aligned with our observation.

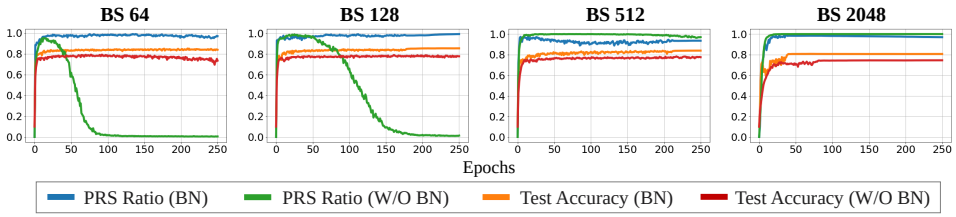
To further investigate other factors, we select two training techniques used in general: (1) Drop out (DO), and (2) Batch normalization (BN). In order to minimize the other influences such as the structural characteristics (e.g., Skip Connection of ResNet), we adopt CNN-6 as the base model in ablation study. Figure 16 (b) presents the PRS ratio and test accuracy for training epochs according to the existence of DO ($p = 0.2$) over various BS on CIFAR-10. We find that DO tends to delay the decrease of PRS ratio in the cases of BS 64 and 128. However, if the network has high PRS ratio (BS 512 and 2048), DO does not affect to change of the PRS ratio. Figure 16 (c) represents the result of the training with the existence of BN for various BS on CIFAR-10. We find that the models with BN have the high PRS ratio in all cases. When we consider the relationship between the PRS ratio and robustness, BN can be considered as the factor which causes adversarial vulnerability. Previous work (Benz et al., 2021) describes the negative effect of BN on the robustness.



(a) Influence of batch size for the PRS ratio in the training with CIFAR-10.



(b) Influence of drop out for the PRS Ratio in the training with CIFAR-10.



(c) Influence of batch normalization for the PRS Ratio in the training with CIFAR-10.

Figure 16: The ablation study for the factors which affect to the PRS ratio.

G PRS RATIO AND ROBUSTNESS FOR ANOTHER ATTACKS

We provide the correlation between PRS ratio and robustness for the FGSM attack (Goodfellow et al., 2014). The experiment is performed for all selected candidates described in Table 2. For the experiments, we take the magnitude of ϵ as follow: MNIST = 0.3, F-MNIST = 0.1, and CIFAR10 = 0.0313 on L_∞ norm. Figure 17 presents the scatter plots between the PRS ratio and robust accuracy with various architectures and datasets. We identify that inverse correlation also holds for FGSM attack.

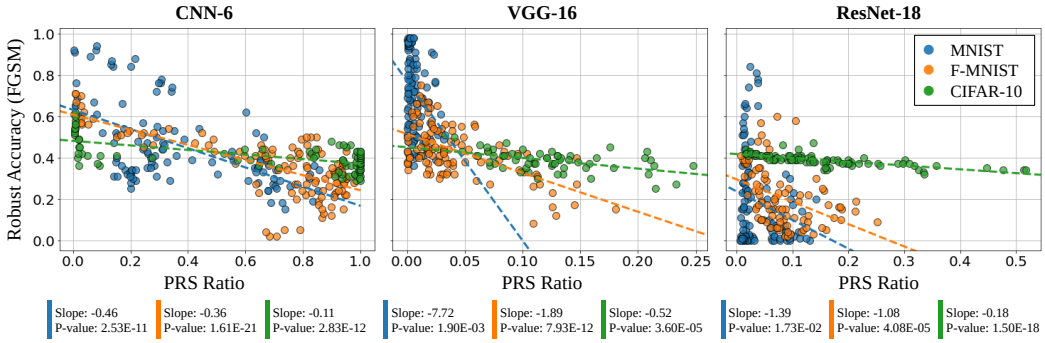


Figure 17: Relationship between PRS ratio and robust accuracy attacked by FGSM method in various models and datasets. The colored dots indicate the independent candidates described in Appendix A. The colored dashed lines indicate the regression slope for each dataset.

We additionally measure the robust accuracy under Auto Attack(AA) (Croce & Hein, 2020) on two networks with different PRS ratio. The used models are selected at the 200th epoch with random seed 907, except for VGG-16 on CIFAR-10 (the 75th epoch). Table 7 presents the robust accuracy for various attacks. We identify that the network with low PRS ratio is more robust than higher one in almost attack methods.

Table 7: Robust accuracy for various attack methods (Higher is better). The magnitude of ϵ is as follow: MNIST = 0.3, F-MNIST = 0.1, and CIFAR10 = 0.0313 on L_∞ norm.

Model	Dataset	Batch Size	PRS Ratio	Clean	FGSM	PGD	AA
CNN6_4	MNIST	128	0.201	99.46	**83.54**	**39.79**	0.00
		2048	0.679	99.39	37.11	11.81	0.00
	fMNIST	128	0.195	92.24	**52.63**	**34.05**	0.00
		2048	0.977	91.05	28.39	2.67	0.00
	CIFAR10	128	0.022	77.60	**44.71**	**41.17**	**23.57**
		2048	1.000	75.66	40.10	25.92	7.48
VGG-16	MNIST	128	0.001	99.18	54.22	36.36	0.00
		2048	0.024	99.35	**55.59**	**36.68**	0.00
	fMNIST	128	0.012	92.66	**45.09**	**30.60**	**14.28**
		2048	0.060	92.44	41.31	17.14	7.99
	CIFAR10	128	0.012	77.90	**44.40**	**33.68**	**21.08**
		2048	0.106	75.05	39.92	30.03	14.22
ResNet-18	MNIST	128	0.013	99.34	**51.15**	**24.88**	0.00
		2048	0.078	99.37	0.40	0.00	0.00
	fMNIST	128	0.027	90.75	**31.56**	**19.94**	**0.60**
		2048	0.078	90.29	26.55	12.98	0.00
	CIFAR10	128	0.073	75.40	**40.27**	**29.87**	**28.52**
		2048	0.334	72.08	32.61	24.51	23.45

H ADVERSARIAL TRAINING AND PRS RATIO

To verify the relationship between adversarial training (AT) and the PRS ratio, we perform the comparison between standard training (ST) and AT.

We first train the ResNet-18 with CIFAR-10 dataset using standard training as the pre-trained model (0th - 150th epoch). We then train two networks from this pre-trained model using (1) standard training, and (2) AT based on the PGD on L_∞ with $\epsilon = 0.0313$. After training, we compare the PRS ratio and the robust accuracy against PGD attack. Figure 18 presents the PRS ratio and the robust accuracy in the training. We observe that the PRS ratio drops and the robust accuracy increases at the same epoch (160th epoch), while those of ST almost maintain.

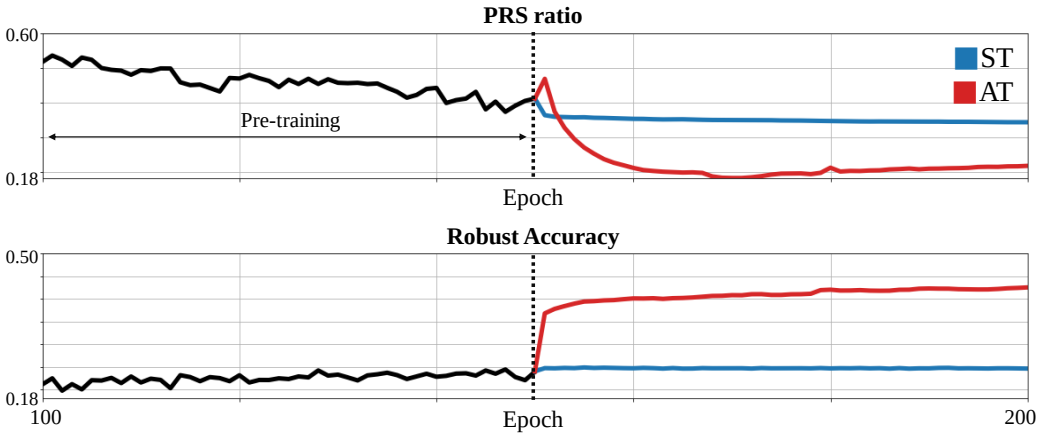


Figure 18: PRS ratio and robust accuracy for two training schemes. Black line indicates the pre-trained network. Red/Blue line indicates standard training and adversarial training, respectively.

I ROBUSTNESS FOR INCLUSION/EXCLUSION TEST SAMPLES

In this section, we provide additional robustness evaluation for different networks and datasets which is mentioned in section 5.

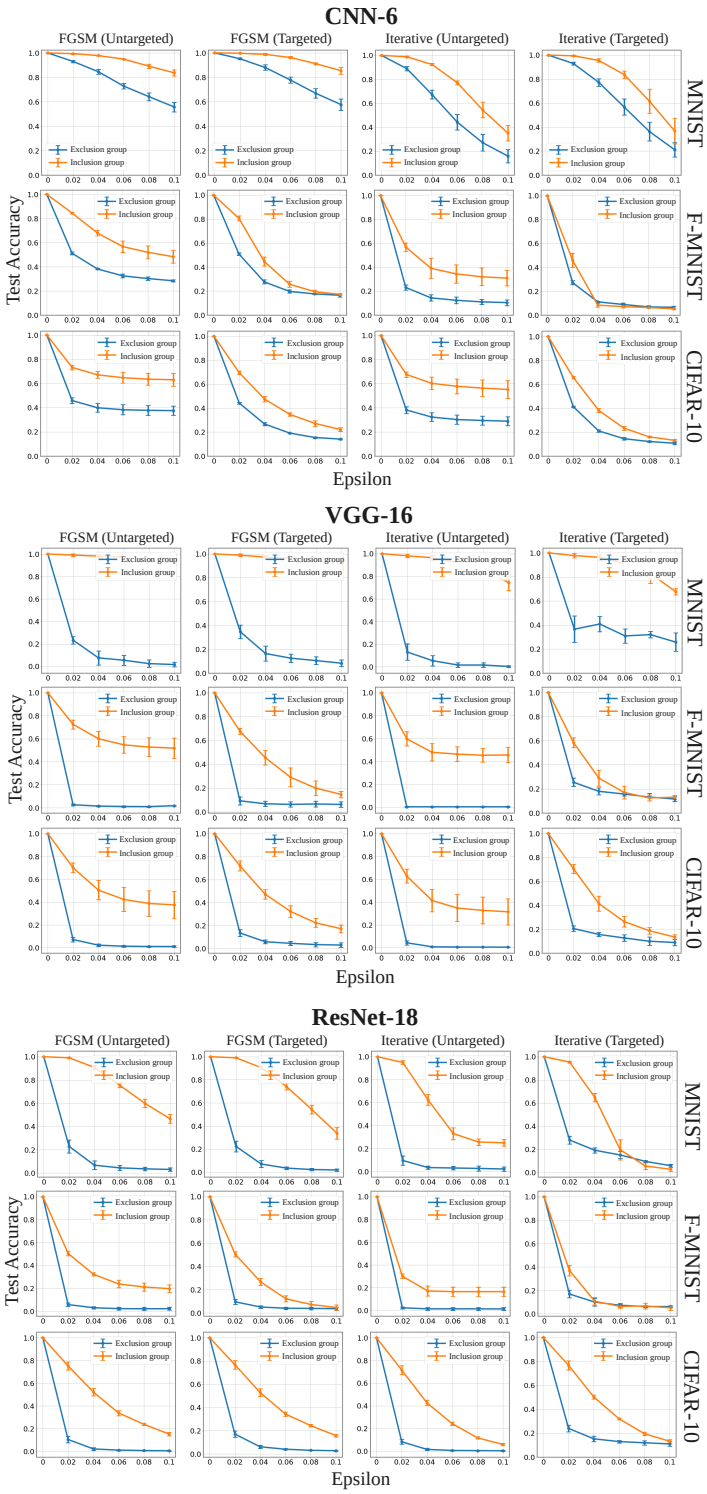
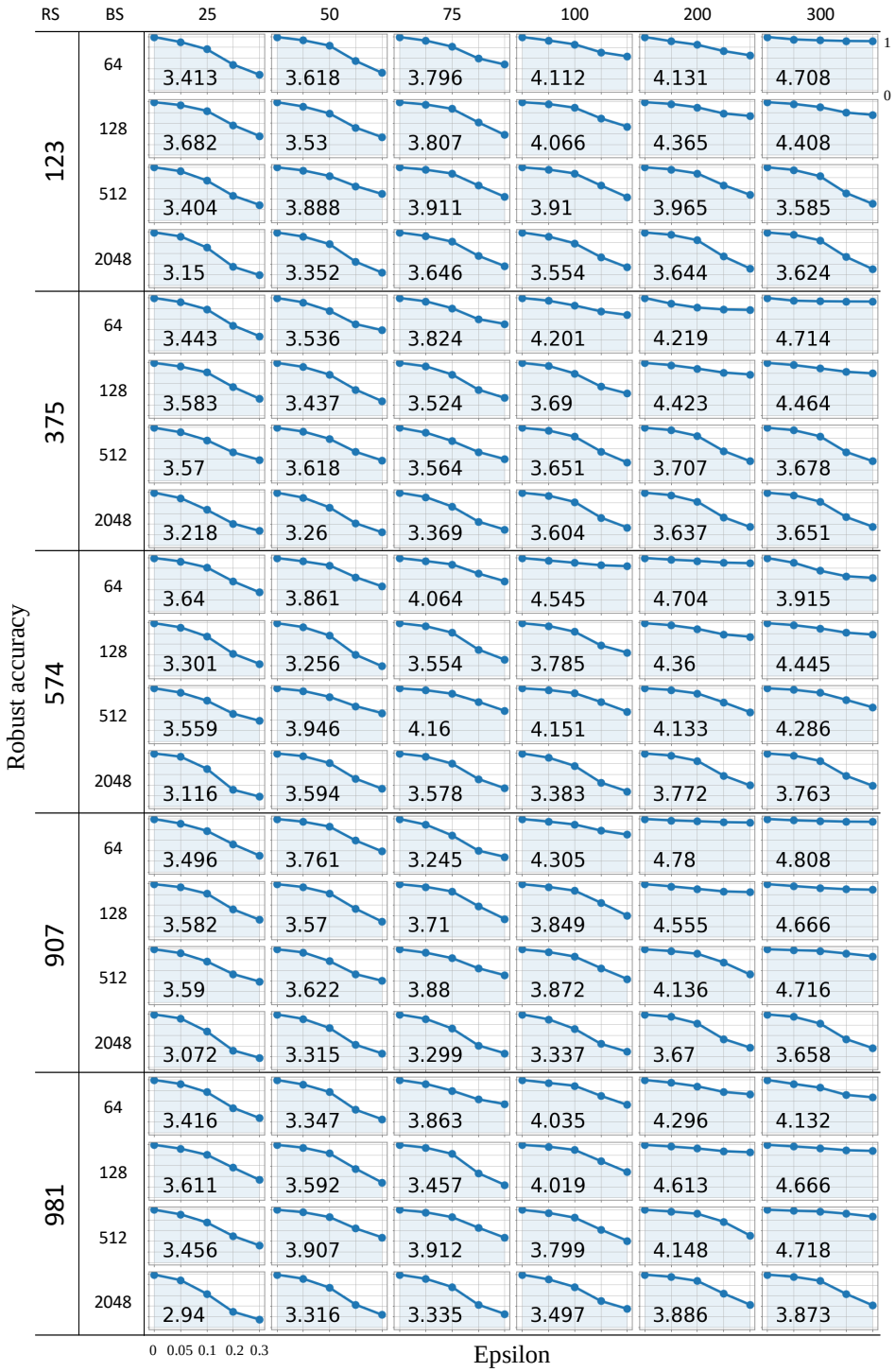


Figure 19: Test accuracy under adversarial attacks on L_∞ norm for exclusion/inclusion group on CNN-6, VGG-16, and ResNet-18. The blue/orange line shows the exclusion/inclusion group.

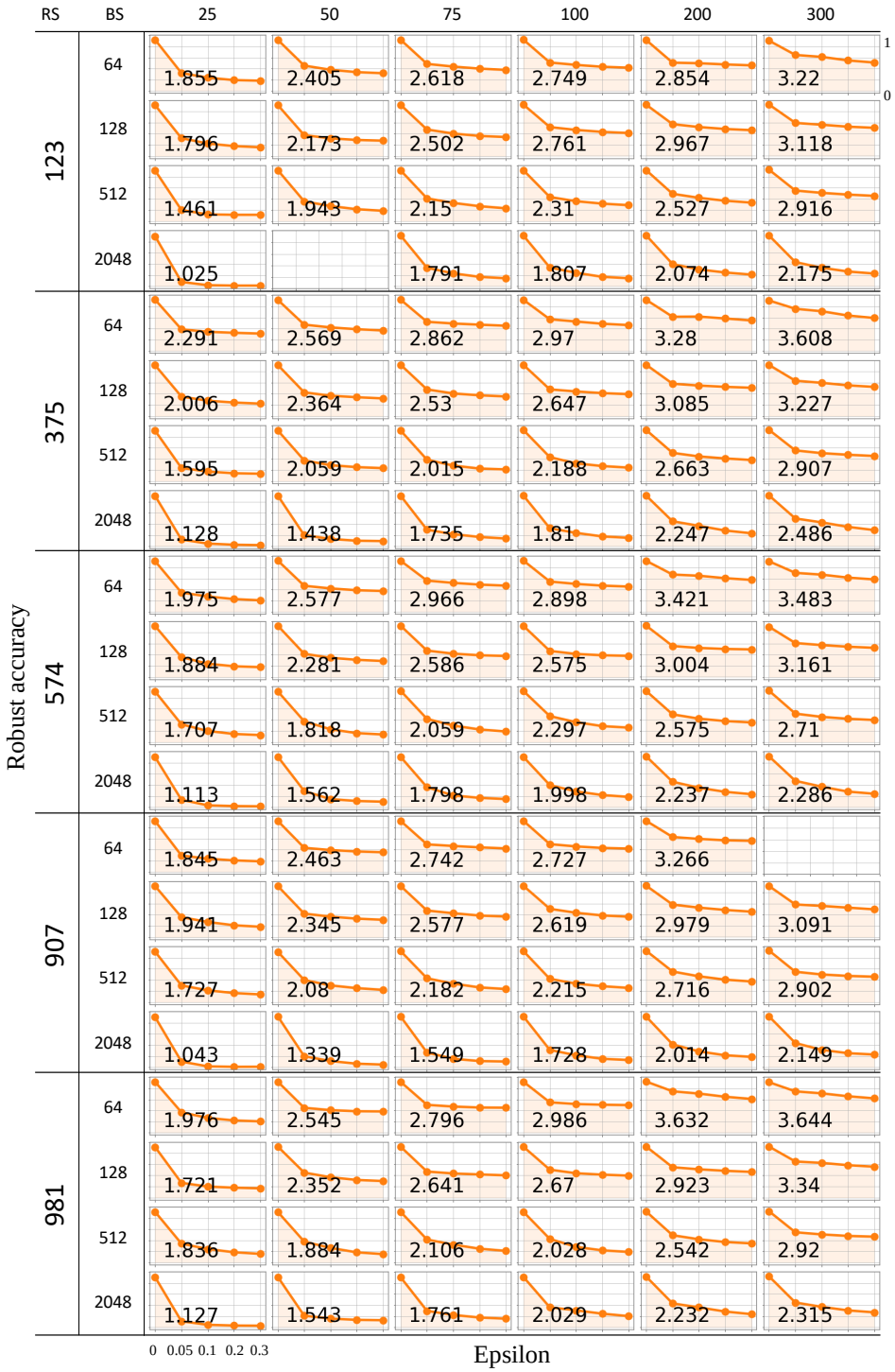
J ROBUST ACCURACY FOR FGSM IN CNN-6 WITH MNIST

We provide the area under the curve for all robust accuracy on various ϵ (left-bottom constant). The row indicates the epoch and RS/BS denotes random seed and batch size, respectively.



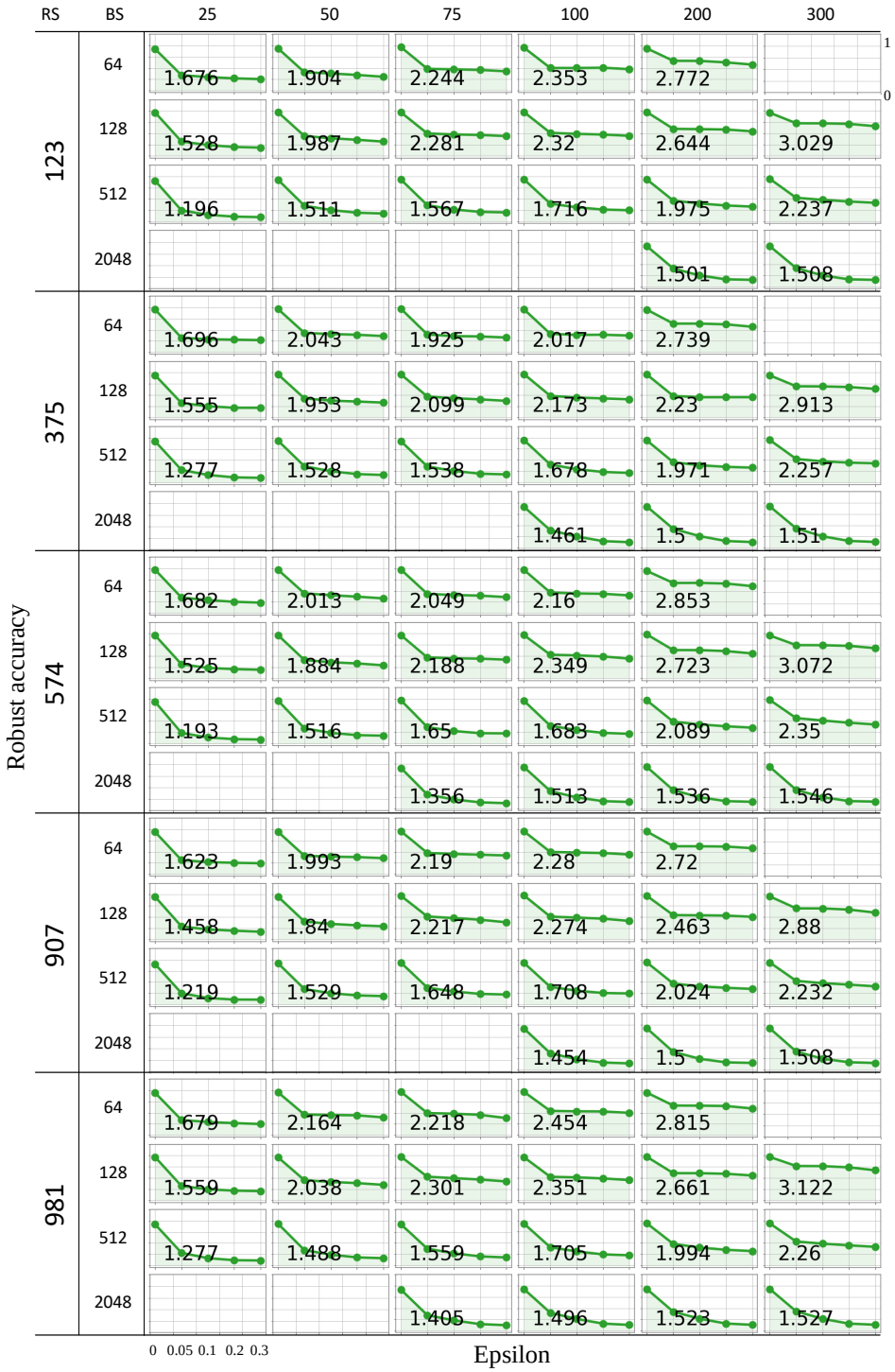
K ROBUST ACCURACY OF FGSM IN CNN-6 WITH F-MNIST

We provide the area under the curve for all robust accuracy on various ϵ (left-bottom constant). The row indicates the epoch and RS/BS denotes random seed and batch size, respectively.



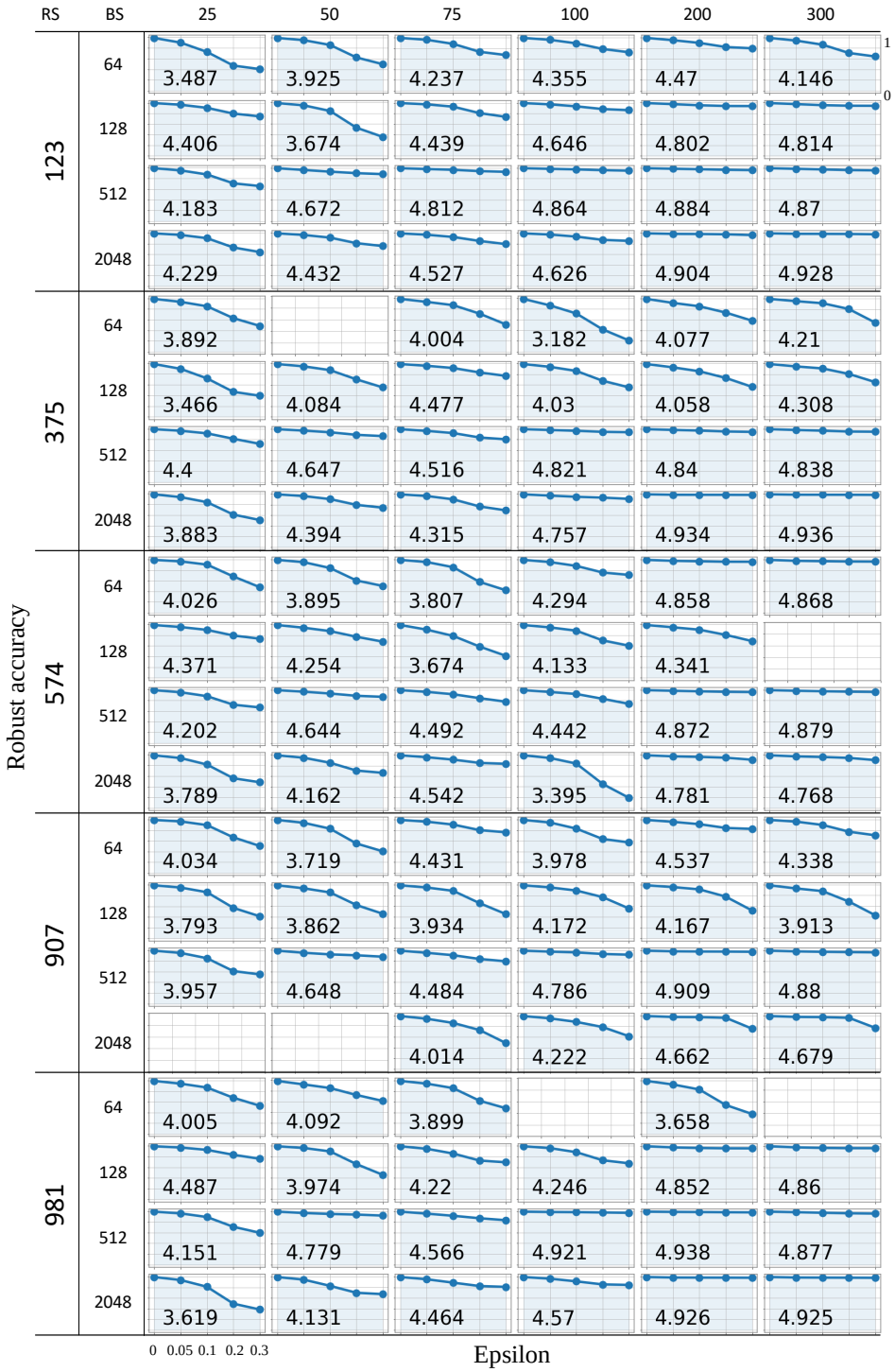
L ROBUST ACCURACY OF FGSM IN CNN-6 WITH CIFAR-10

We provide the area under the curve for all robust accuracy on various ϵ (left-bottom constant). The row indicates the epoch and RS/BS denotes random seed and batch size, respectively.



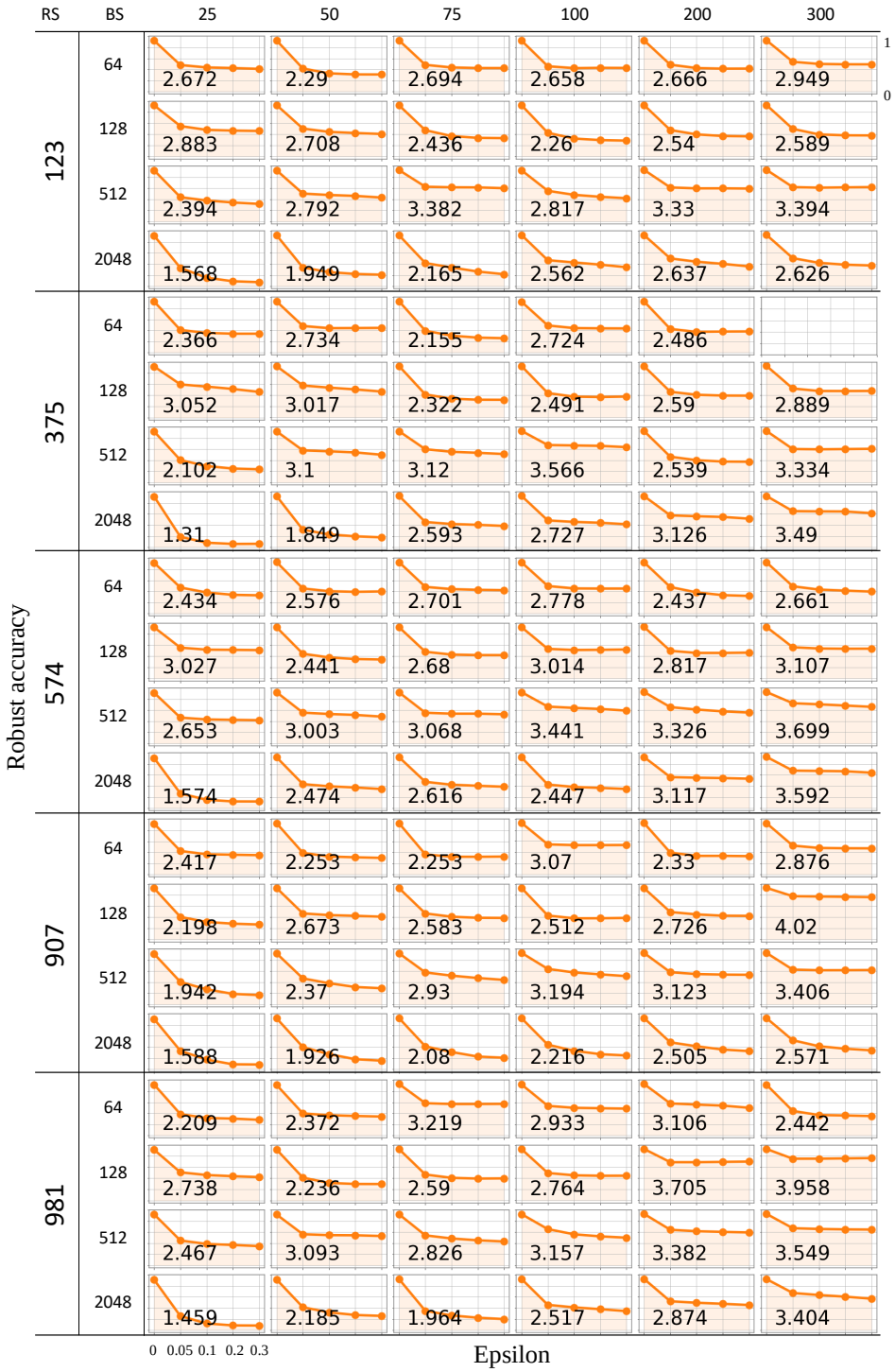
M ROBUST ACCURACY FOR FGSM IN VGG-16 WITH MNIST

We provide the area under the curve for all robust accuracy on various ϵ (left-bottom constant). The row indicates the epoch and RS/BS denotes random seed and batch size, respectively.



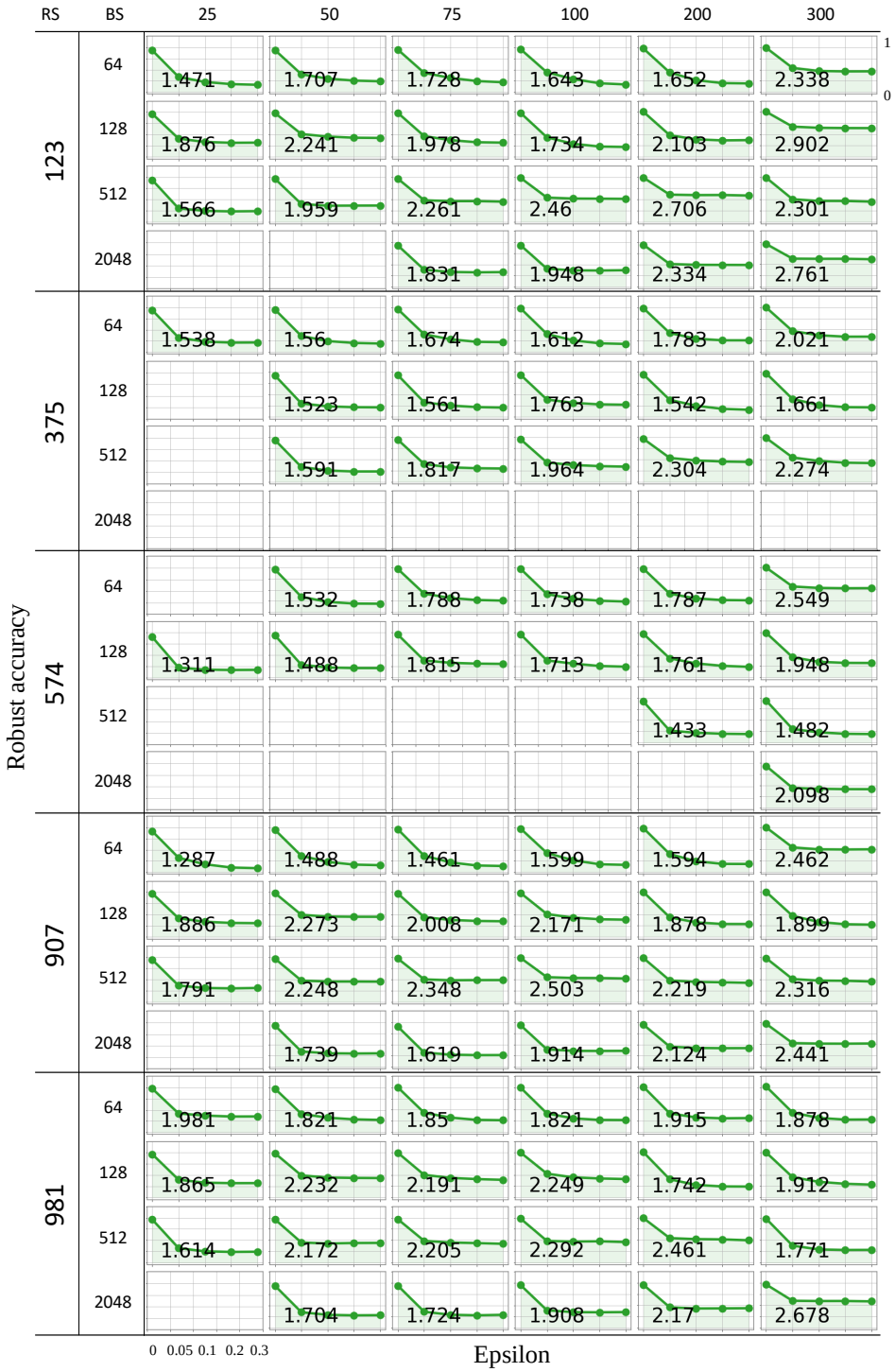
N ROBUST ACCURACY OF FGSM IN VGG-16 WITH F-MNIST

We provide the area under the curve for all robust accuracy on various ϵ (left-bottom constant). The row indicates the epoch and RS/BS denotes random seed and batch size, respectively.



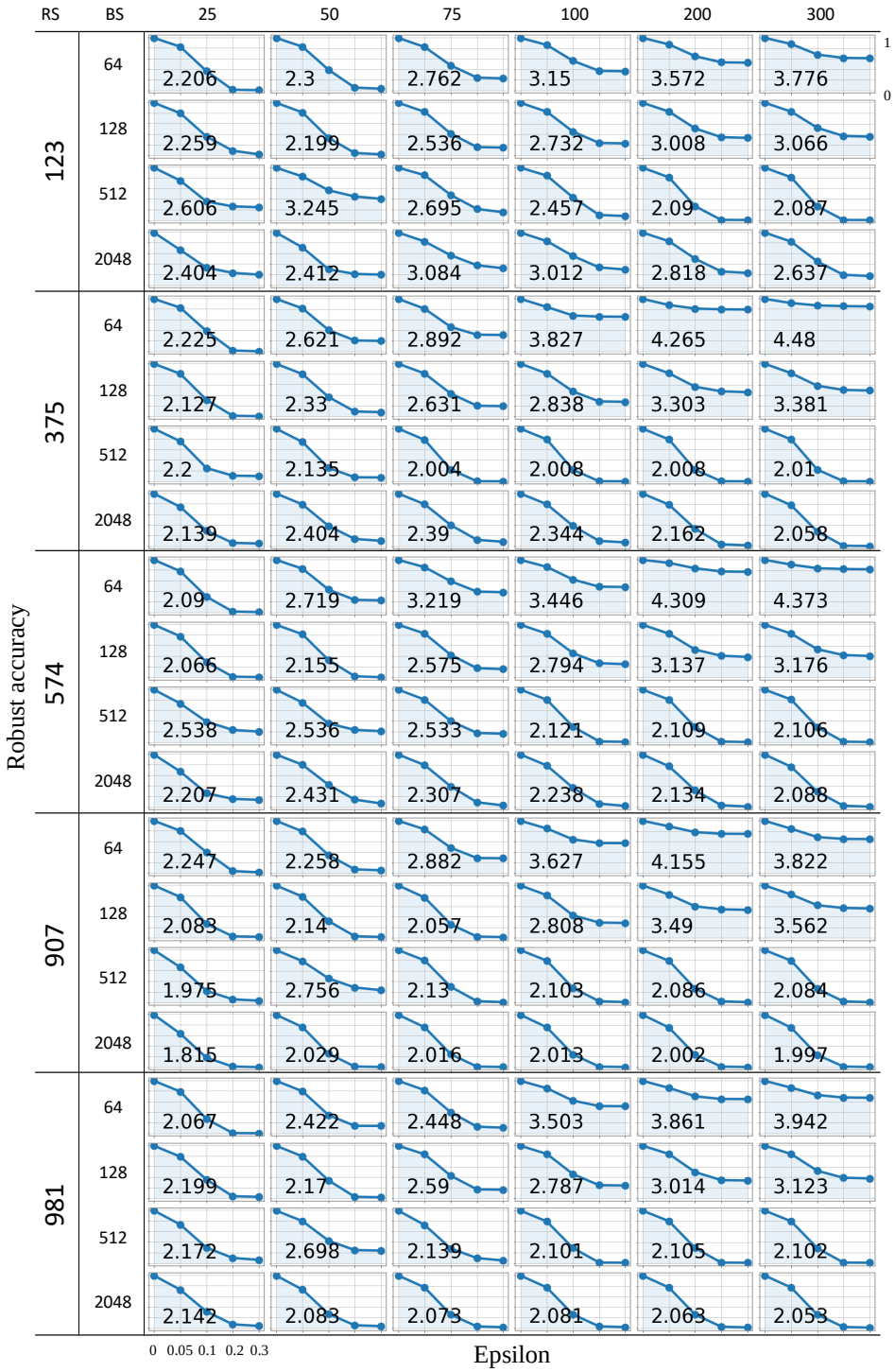
O ROBUST ACCURACY OF FGSM IN VGG-16 WITH CIFAR-10

We provide the area under the curve for all robust accuracy on various ϵ (left-bottom constant). The row indicates the epoch and RS/BS denotes random seed and batch size, respectively.



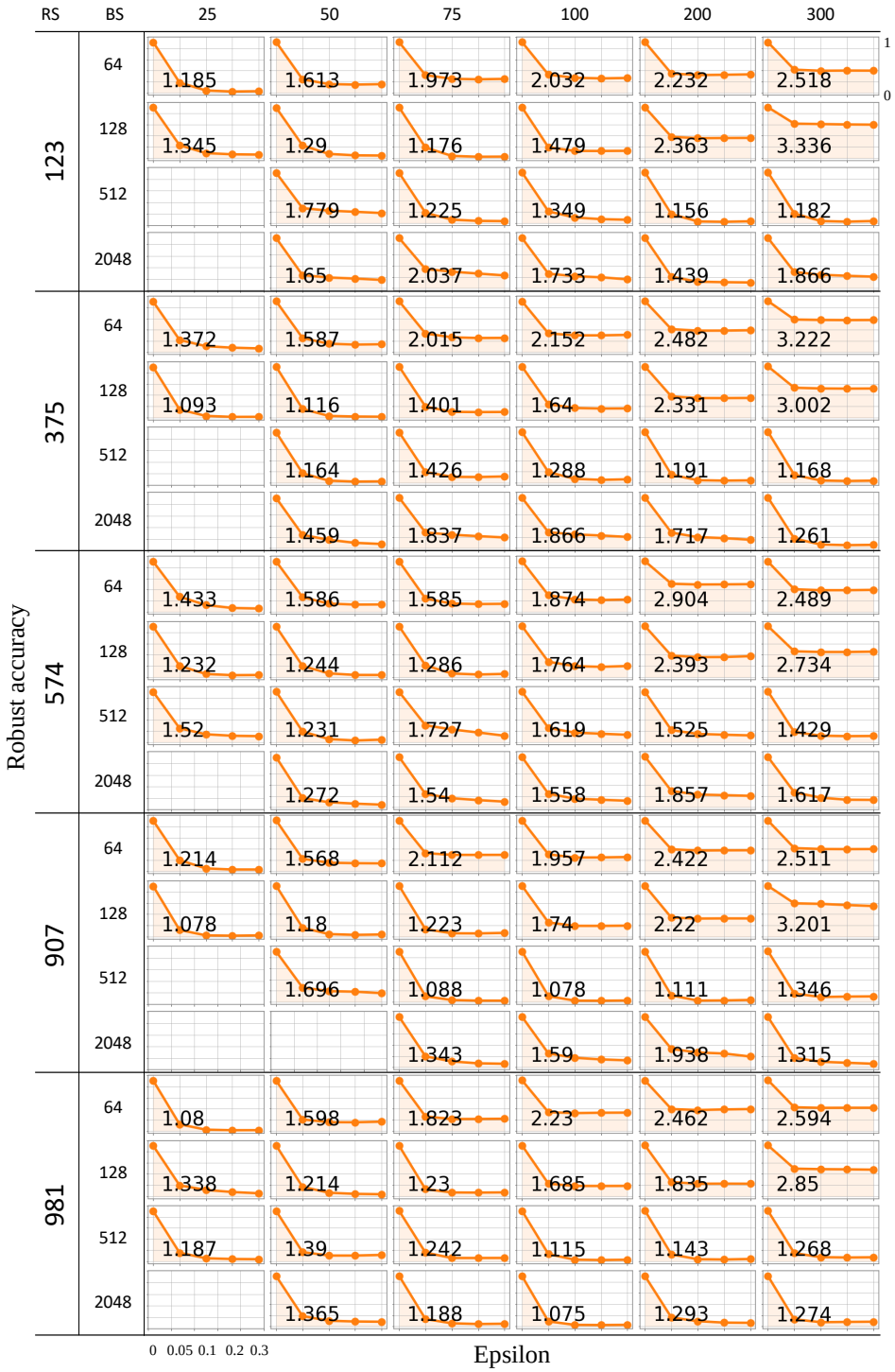
P ROBUST ACCURACY FOR FGSM IN RESNET-18 WITH MNIST

We provide the area under the curve for all robust accuracy on various ϵ (left-bottom constant). The row indicates the epoch and RS/BS denotes random seed and batch size, respectively.



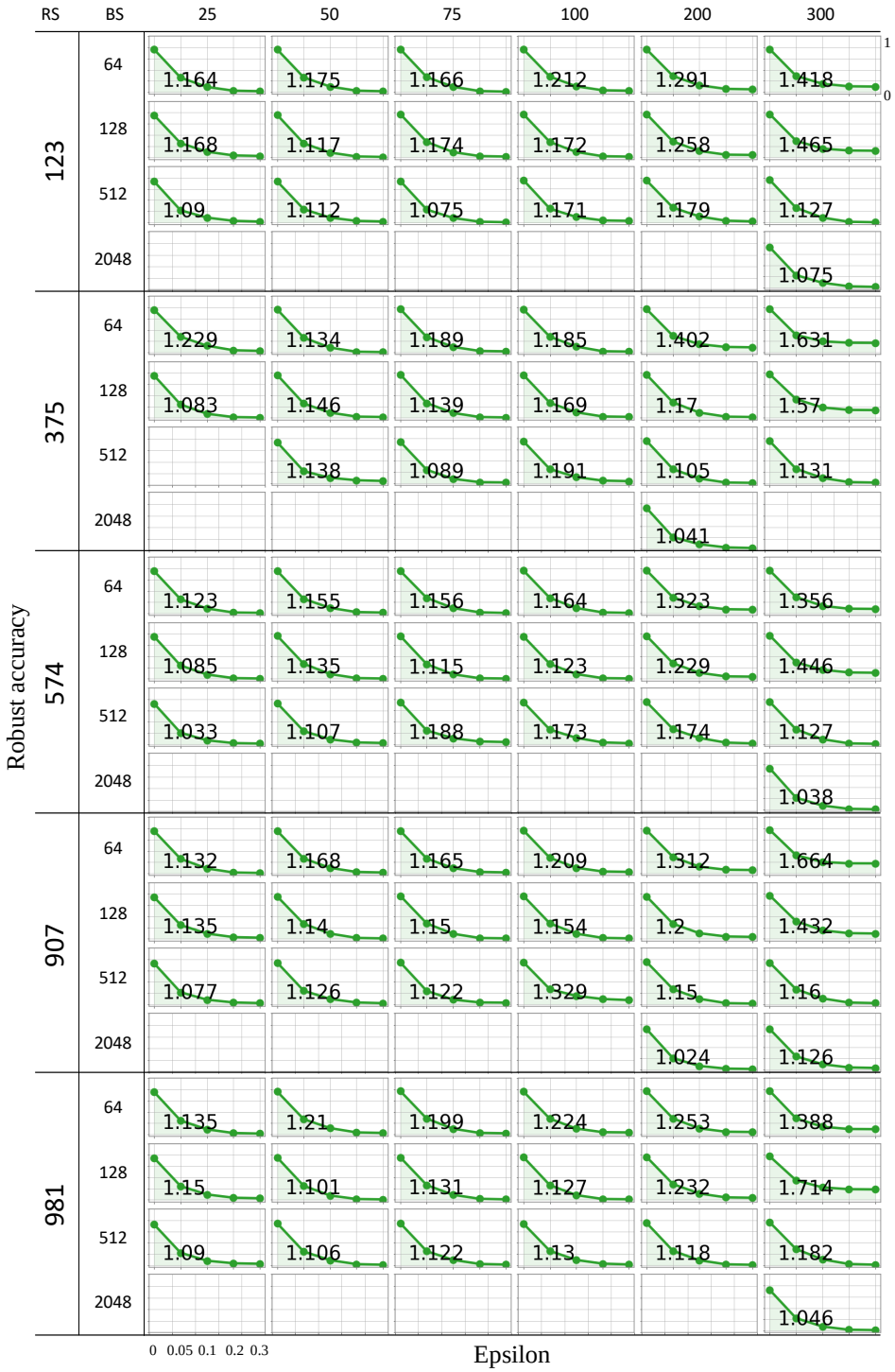
Q ROBUST ACCURACY OF FGSM IN RESNET-18 WITH F-MNIST

We provide the area under the curve for all robust accuracy on various ϵ (left-bottom constant). The row indicates the epoch and RS/BS denotes random seed and batch size, respectively.



R ROBUST ACCURACY OF FGSM IN RESNET-18 WITH CIFAR-10

We provide the area under the curve for all robust accuracy on various ϵ (left-bottom constant). The row indicates the epoch and RS/BS denotes random seed and batch size, respectively.



REFERENCES

Philipp Benz, Chaoning Zhang, and In So Kweon. Batch normalization increases adversarial vulnerability and decreases adversarial transferability: A non-robust feature perspective. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7818–7827, 2021.

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, pp. 2206–2216. PMLR, 2020.

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

Zhewei Yao, Amir Gholami, Qi Lei, Kurt Keutzer, and Michael W Mahoney. Hessian-based analysis of large batch training and robustness to adversaries. *arXiv preprint arXiv:1802.08241*, 2018.