Appendix

A. Performance of Post-pruning

We performed post-pruning for adversarially trained ResNet18 models (with variable sizes w) for CIFAR10 (the same setting as Table 4 in the paper). We found that without retraining, almost all cases show accuracies of 10%/10% (No surprise if we look at Figure 1 in the paper). Then we performed post-pruning with retraining and show the results in following table. Consistent to our key result, pruning from a robust larger model gives better results than training a small model from scratch. In particular, when the difference between the original size and the pruned model's size becomes large, our proposed framework outperforms the post-pruning-with-retraining. For example, the 16-to-1 case is 64.17/37.99 in Table 4 and is only 60.26/36.18 in the following table. Furthermore, we perform additional experiments to verify the importance of ADMM. For LeNet under FashionMNIST, post-pruning and concurrent pruning without using ADMM (proximal gradient descent is used instead) give failure cases when prune rate is large, while ADMM achieves good results under the same training time.

Table A1: Post-pruning (with retraining) for ResNet18 on CIFAR10. Compared to ADMM method, post-pruning without retraining makes almost all models drop to 10.00/10.00.

w	1	2	4	8
2	64.39/38.05	-	-	-
4	62.49/36.77	73.47/43.09	-	-
8	60.40/37.05	72.52/43.34	78.64/45.19	-
16	60.26/36.18	69.47/42.14	78.59/46.17	80.79/46.4

B. Initialization Analysis

The table below contains study of how initialization affects training a small robust model.

Table B1: **Natural test accuracy/adversarial test accuracy** (in %) on **MNIST** (by LeNet with size of $w = 1$) with seven different initialization methods and three optimizers: Adam, SGD, and CosAnneal.

initialization	Adam	SGD	CosAnneal
uniform	**78.86/70.47**	11.35/11.35	11.35/11.35
normal	11.35/11.35	11.35/11.35	11.35/11.35
xavier_uniform[13]	11.35/11.35	11.35/11.35	11.35/11.35
xavier_normal[13]	11.35/11.35	11.35/11.35	11.35/11.35
kaiming_uniform[18]	11.35/11.35	11.35/11.35	11.35/11.35
kaiming_normal[18]	19.68/19.02	11.35/11.35	11.35/11.35
orthogonal	11.35/11.35	11.35/11.35	11.35/11.35

C. Performance against C&W attack

The test accuracy of our proposed framework against C&W ℓ_∞ attack.

Table C1: C&W ℓ_∞ adversarial test accuracy (in %) by the proposed framework on MNIST by LeNet.

w	baseline	1	2	4	8
2	11.35	11.35	-	-	-
4	91.42	89.63	91.75	-	-
8	93.57	92.33	93.83	94.46	-
16	94.78	89.26	91.34	95.08	95.62