We thank all reviewers for their thoughtful and constructive feedback. We are encouraged that the reviewers find our idea of making pruning aware of robust training objectives intuitive (**R1**), our experiments well-informed (**R1**) and extensive (**R2**, **R5**), and success of our method, across different robust training objectives, datasets, architectures, and pruning ratios, an accomplishment (**R2**, **R3**, **R4**). We are pleased that **R1** finds our observation on the existence of robust sub-networks timely and that **R2** finds our extensive supplementary material impressive.

Two major concerns were 1) additional insights on why proposed initialization is highly effective (**R2**, **R3**, **R4**) and 2) additional comparison with Adv-ADMM baseline (**R3**, **R5**). We address both, along with other comments, below and will incorporate all feedback in the updated version of the paper.

[**R2**, **R3**, **R4**] **Why is the proposed initialization effective? Why do random initializations not work so well?** We show in Fig. 1 (left) that with proposed initialization SGD converges faster and to a better pruned network, in comparison to widely used random initializations. This is because with our initialization SGD enjoys much higher magnitude gradients throughout the optimization in the pruning step (Fig. 1 right). We will add in-depth analysis of it in the main paper. [**R2**] **What other initializations were tried?** We now compare with two more initializations, based on Dirac delta function and orthogonal matrices, along with four other widely used initializations.

[**R3**, **R5**] **More comparison with Adv-ADMM:** Following reviewers' suggestion, we now provide a comparison along six more recent architectures (Table 1). Our method achieves better accuracy and robustness, simultaneously, across all of them. Furthermore, when Adv-ADMM fails to even converge for MobileNet, a highly compact network, we achieve non-trivial performance. We already provide comparison at different pruning ratios in Table 2 of the main paper. [**R3**] **Verified robustness?** We use existing techniques IBP, Mixtrain, randomized smoothing to compute verified robustness.

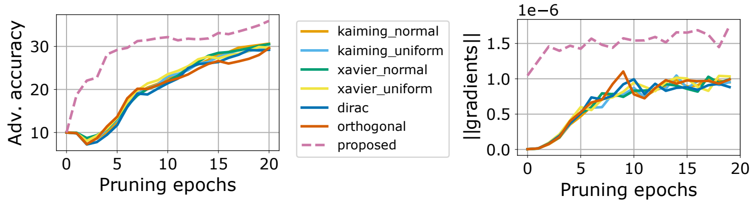

Figure 1: We compare proposed initialization with six other widely used initializations. With proposed initialization in the pruning step, SGD converges faster and to a better architecture (left), since it enjoys higher magnitude gradients throughout (right). (CIFAR10, 99% pruning)

Network	Adv-ADMM	Ours	Δ
ResNet-18	58.7/36.1	69.0/41.6	+10.3/+5.5
ResNet-34	68.8/41.5	71.8/44.4	+3.0/+2.9
ResNet-50	69.1/42.2	73.9/45.3	+4.8/+3.1
WRN-28-2	48.3/30.9	54.2/34.1	+5.9/+3.2
GoogleNet	53.4/33.8	66.7/40.1	+13.3/+6.3
MobileNet-v2	10.0/10.0	39.7/26.4	+29.7/+16.4

Table 1: Comparing test accuracy/robustness (*era*) with Adv-ADMM (CIFAR10 dataset, 99% pruning). Our approach outperforms Adv-ADMM across all network architectures.

[**R5**] **Comparison with other pruning strategies:** Following reviewer's suggestion, we now provide comparison with techniques from each of three pruning paradigms, i.e., pruning before training (SNIP), pruning with training (STR), and pruning followed by re-training (Adv-LWM, Adv-ADMM – these were already covered in the paper). We find that it is not always the case that an existing pruning strategy will also be successful with robust training. For example, SNIP performs equivalent to random pruning, i.e., scratch in Table-1 of main paper, when tested with adversarial training (*era* for SNIP=27.2%, scratch=24.6% at 99% pruning ratio). Similarly, STR with adversarial training achieves only 33.2% *era*, where our approach achieves 41.6% *era* at 99% pruning ratio for the ResNet-18 network and CIFAR10 dataset.

[**R5**] **Better architectures like Wide-ResNet, MobileNet:** We already use Wide-ResNet-28-4 in major experiments (Table-3 in the paper). Following reviewer's suggestion, we demonstrate the success of our method with recent architectures in Table 1. We achieve SOTA robustness in the context of highly compressed networks across multiple architectures. [**R5**] **Authors use L_{benign} in most of their pruning experiments:** We think that this comment is a misunderstanding as we use L_{adv} or L_{ver} in most experiments (L146). We thank the reviewer for providing additional suggestions on related work and clarity, which we will incorporate.

[**R5**] **As long as models are compressed properly, resulting models are robust . . . compare with RIGL, STR:** We find that it is critical for the pruning step to be aware of the robustness objective. However, techniques like RIGL/STR, don't account for the robustness objective while pruning. As suggested by the reviewer, we experimented with their pre-trained checkpoints and observed 0-0.5% robust accuracy. This further validates the importance of our approach.

[**R4**] **More experiments on ImageNet:** Following reviewer's suggestion, we also demonstrate success with provable robustness on ImageNet. Our approach achieves 47% provable robustness, while pre-trained nets has 49% and Adv-LWM achieves 44%, at $||\epsilon||_2 = 0.5$, 90% pruning, using randomized smoothing. Note that this is the *first* work to demonstrate the success of pruning with robust training on the scale of ImageNet. Earlier works [14, 44] only perform experiments with CIFAR10/SVHN. We hope that the reviewer sees this as a *strength*, not a weakness of our paper.

[**R1**] **Combination of two existing ideas?** We clarify that this combination is in itself novel, and that our proposed initialization is a key driver for success. This insight allows us to achieve SOTA accuracy and robustness, for compressed networks, across different pruning ratios, architectures, and datasets (including provable robustness at ImageNet scale).