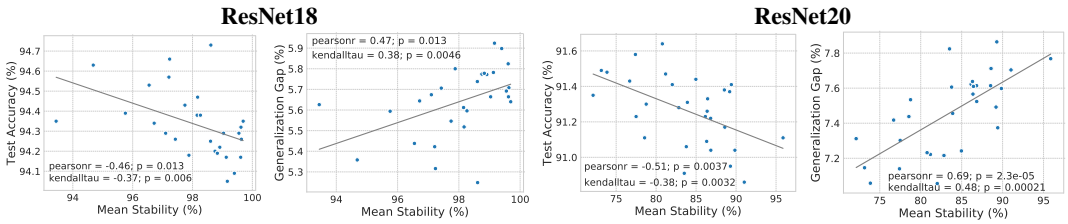
We thank the reviewers for their detailed, valuable reviews. We are glad that the reviewers saw that our work was: "novel and insightful" with "a very intriguing and intuitive explanation for the mechanics of pruning" and "well designed" experiments (R1); possessing an "interesting" core observation about pruning stability and generalization and a "[comprehensive] set of research questions", including a "worthwhile study" of "pruning the largest magnitude filters" and "characterizations of flatness" (R2); and addressing "an interesting phenomenon" with a "proposed metric" and approach that "could lead towards more insight in the training dynamics of sparse neural networks" (R3). We address reviewer comments as much as possible below but will incorporate all feedback in the final version.

R1,R2,R3 **Concern about generality of results due to experimental hyperparameters and overparameterized networks:** We agree with the reviewers' concern and will ensure the final version includes the following data that show the generalization-stability tradeoff is not merely an artifact of the particular setting we studied. Specifically, the following graphs go "beyond the two CIFAR-10 networks" (R3) to include the significantly less-parameterized "ResNet20" (R3). **To show that our results hold without tuned hyperparameters for training and pruning**, we trained ResNet18 and ResNet20 with the exact training configuration described in the CIFAR section of He et al. (2015)—**SGD, data augmentation, weight decay**, etc.; i.e., all of the factors requested by R1 except for dropout. These studies use a "more robust pruning regime" (R2) by pruning **a constant fraction of all layers in all blocks** with the common filter ℓ_1 norm (Li et al., 2017) "for calculating filter importance" (R2). We emphasize that, even when using this "simplest, possible prune setup" and "standard training pipelines" (R1), the generalization-stability tradeoff is clearly present, confirming that our observations do indeed apply to "more general settings" (R1)—and based on the ResNet20 results, seem to apply to less "modern" regimes (R1), though modern regimes inspired our central question.



R2,R3 **Generalization gap vs. test accuracy; train accuracy not reported:** Correct, we neglected to state that all models had 100% training accuracy. With constant training accuracy, higher generalization (test accuracy) implies a smaller generalization gap. Thus, lower stability improves generalization *and* reduces the generalization gap (overfitting)! We will update our manuscript to clearly discuss training accuracies and plot the generalization gaps.

R3 **"Pearson correlation and slope do not give an accurate characterization":** Correct, the graphed relationships are not always linear. The manuscript will add a Kendall rank coefficient (Kendall, 1938; Jiang, 2019), τ, for each Pearson coefficient. Of the 20 statistically significant Pearson tests, 19 of the Kendall tests are statistically significant (one p-value went from 0.04 to 0.07) and all tests had the same sign, further supporting the hypothesized tradeoff.

R2,R3 **Methodology in "main body" and its "clarity":** We will move methodological details to the "main body" (R2, R3), improve their "clarity" (R3), and accent that our method lacks "deal-breakers" (R1) as shown above.

R1, R3 **Hyperparameter choices ("These networks reach much lower accuracy than expected... L1/L2 regularization is disabled"):** Section 2 justified our exposition's focus on less-regularized models, which is not unprecedented: the phenomena explored in the main text of a double descent paper referenced by R1 (Nakkiran et al., 2019) were produced without weight decay—in fact, our CIFAR-100 experiment in Table F.1 explicitly mimicked their approach, including use of "data augmentation" and networks with "residual connections" (R1). As for our "per-layer" (R1) and other pruning settings, these are not "deal-breakers" (again, see above), though our final draft will explain their origin was our initial, unpublished work (excluded to preserve anonymity) that pruned only the last dense layer of a small network. It led to our exploring pruning of the last convolutional layers of VGG11/ResNet18. The present study added pruning of other layers until we saw breakdown points—illustrated in the Section 3.2 "limit studies" appreciated by R2.

R2,R3 **"[DSD] is worth a comparison" and "the claim... is hard to extract":** We thank the reviewer for pointing out this reference and we will include the following discussion in the related work. DSD (Han et al., 2017) shows that pruning based generalization improvements are retained and improved after restoring pruned connections. Relative to DSD, we show that the parameters can reenter at zero *or* their original values (Figure D.2) while achieving the full benefit of pruning and temporary Gaussian noise can replace pruning events to achieve a benefit.

R3 **"[15] is not found to improve generalization":** [15] (LeCun et al., 1990) says OBD improved test error in the last sentence of page 603 and improved "recognition accuracy" in the third sentence of the conclusion.

R1 **Most experiments use VGG; VGG lacks batch norm**: We emphasize that many of our experiments were performed with ResNet18: Figures 2 and B.2, and Tables C.1 and F.1. We will clarify in the text that our VGG11 model **includes batch normalization** and was used in part because it allowed us to create replicates with limited resources.