

1 We thank reviewers sincerely for their detailed feedback. It has enabled us to improve and clarify several sections.

2 **1. The results are not tested on ImageNet / ResNet [R1, R2, R3].** We agree it would be great to scale these
3 experiments to deeper models (ResNet) and larger datasets (ImageNet) and see how well they generalize. As found by
4 Frankle et al in "Stabilizing..." (2019), not all Lottery Ticket (LT) results generalized to ImageNet-sized models. We
5 suspect that some of our results may similarly not generalize (such as mask-1 actions) but some will (such as Supermasks
6 and mask-0 actions because our hypothesis on masking as training is independent of the original initialization of the
7 weights). Due to time and computational constraints we leave this for future work, but have added a note in our paper
8 discussing the lack of generalization results. [See Update a below]

9 **2. Significance of results are unclear; mask criteria curves unclear [R1].** Some of the results were indeed poorly
10 described; we have updated several descriptions and the plot captions and think the clarity has been improved
11 significantly [Updates b, e, f]. Summary: the significance of "magnitude increase" results are two-fold: it shows the LT
12 phenomenon is not exclusive to the large final criterion, and the results are consistent with our hypothesis of masking as
13 training. The mask-1 "significance of the sign" results show, in contrast to the LT paper, that the basin of attraction
14 for a lottery ticket network is actually quite large: anywhere in the correct weight quadrant optimizes well; optimizers
15 encounter difficulty crossing the zero barrier between signs. Finally, "masking as training" proposes an explanation for
16 many of our results (and is predictive of performance of both mask criteria and alternate mask-0 treatments). We have
17 also made the plots easier to interpret by simplifying the confidence bands and highlighting the main takeaways in the
18 caption [Updates d].

19 **3. Missing results on trained performance of additional Supermasks [R2].** This is a great suggestion. We have
20 run these experiments and included them in SI [Update g]. We found that large_nal_same_sign actually slightly
21 underperforms large_nal when trained in the iterative pruning LT paradigm. This is likely because we break 0 ties
22 randomly when a mask criterion does not contain enough 1's, and selecting only weights with the same sign entails the
23 need for much random selection.

24 **4. Questionable use of statistical significance [R2].** Iterative pruning does indeed cause correlation across different
25 pruning percentages. However, we never report p-values aggregated in this way: all statistical tests are performed at a
26 single pruning percentage, aggregating only across five independent runs. We've clarified in the text [Update h].

27 **5. Mask-0 experiments and Fig 5 poorly motivated [R2].** We've removed the unnecessary straw man argument and
28 state the real motivation more simply: to test the hypothesis that the value of frozen weights affects training [Update c].

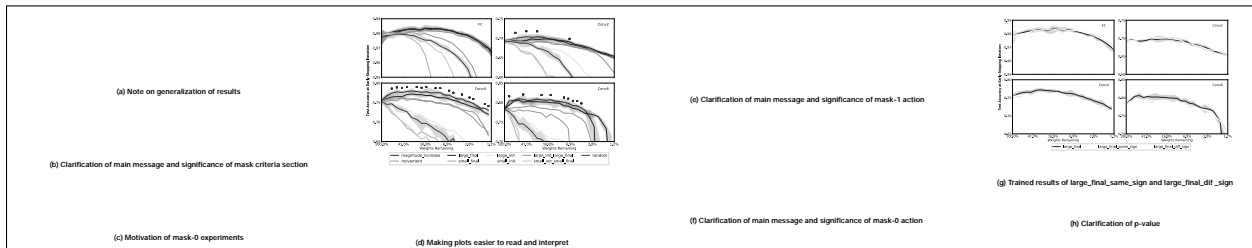
29 **6. Further analysis of masks themselves + principled understanding of lottery tickets [R2, R3].** We agree that
30 it would be a valuable research direction to better understand and provide a theoretical foundation for what makes
31 good masks, as well as their connection to better initializations. We believe our results represent a significant though
32 not consummate step toward better understanding. Relating to the **question on the performance of signed_reinit**
33 **[R3]**, our results hint that using a bimodal distribution to initialize weights may be more beneficial than a Gaussian
34 distribution. This may explain why signed_reinit underperforms signed_resuffle and signed_constant,
35 both of which are initialized using a bimodal distribution.

36 **7. Clarifications on Supermask training [R2, R3].** The Bernoulli is sampled each forward pass, so during training
37 each mini-batch entails a fresh sample of the mask. At test time, we report the average accuracy over 10 random samples
38 of the mask (aggregating accuracy of the single-sample models, not ensembling predictions). As **[R3]** mentions this is
39 akin to learning a per-weight dropout probability, but we should be careful not to interpret this as learned regularization
40 but as the entire learning process itself. **[R3 point 2.3]** We respond to two possible interpretations of your suggestion
41 here: (1) Supermasks are indeed trainable from initial weights that are const magnitude, random sign (Fig 7, right). (2)
42 If signs are randomized *after* training a Supermask, performance degrades to chance (experiment run but not shown).

43 **8. Initialize the learned Supermasks based on a well-performing heuristic mask criteria [R3].** This is a great idea
44 and would likely work, though for now, because it works and is simple, we have trained only from scratch.

45 **9. Mask using other points on optimization trajectory instead of endpoints [R3].** Great idea. We have not yet
46 explored computing masks using (intermediate, final) weights as you suggest instead of (initial, final) weights. We did
47 try masking using (initial, min-val-err) weights, but performance was not as good.

48



49