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

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

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

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

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

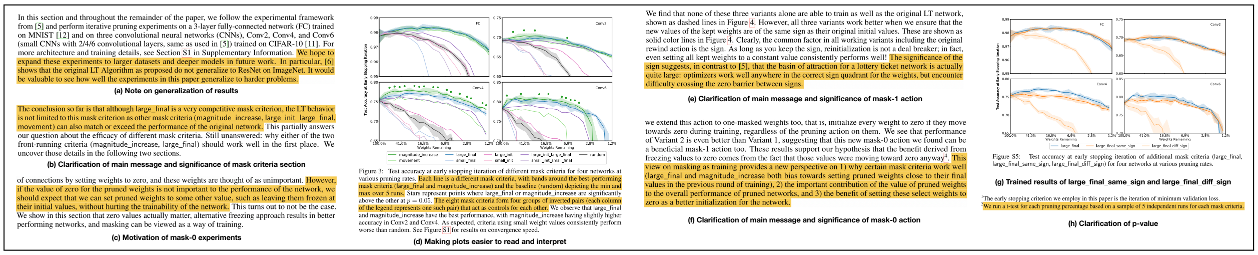
5. Mask-0 experiments and Fig 5 poorly motivated [R2]. We've removed the unnecessary straw man argument and state the real motivation more simply: to test the hypothesis that the value of frozen weights affects training *[Update c]*.

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

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

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

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



In this section and throughout the remainder of the paper, we follow the experimental framework from [5] and perform iterative pruning experiments on a 3-layer fully-connected network (FC) trained on MNIST [12] and on three convolutional neural networks (CNNs), Conv2, Conv4, and Conv6 (small CNNs with 2/4/6 convolutional layers, same as used in [5]) trained on CIFAR-10 [11]. For more architecture and training details, see Section S1 in Supplementary Information. We hope to expand these experiments to larger datasets and deeper models in future work. In particular, [6] shows that the original LT Algorithm as proposed do not generalize to ResNet on ImageNet. It would be valuable to see how well the experiments in this paper generalize to harder problems.

(a) Note on generalization of results

The conclusion so far is that although large_final is a very competitive mask criterion, the LT behavior is not limited to this mask criterion as other mask criteria (magnitude_increase, large_init_large_final, movement) can also match or exceed the performance of the original network. This partially answers our question about the efficacy of different mask criteria. Still unanswered: why either of the two front-running criteria (magnitude_increase, large_final) should work well in the first place. We uncover those details in the following two sections.

(b) Clarification of main message and significance of mask criteria section

of connections by setting weights to zero, and these weights are thought of as unimportant. However, if the value of zero for the pruned weights is not important to the performance of the network, we should expect that we can set pruned weights to some other value, such as leaving them frozen at their initial values, without hurting the trainability of the network. This turns out to not be the case. We show in this section that zero values actually matter, alternative freezing approach results in better performing networks, and masking can be viewed as a way of training.

(c) Motivation of mask-0 experiments

Figure 3: Test accuracy at early stopping iteration of different mask criteria for four networks at various pruning rates. Each line is a different mask criteria, with bands around the best-performing mask criteria (large_final and magnitude_increase) and the baseline (random) depicting the min and max over 5 runs. Stars represent points where large_final or magnitude_increase are significantly above the other at $p = 0.05$. The eight mask criteria form four groups of inverted pairs (each column of the legend represents one such pair) that act as controls for each other. We observe that large_final and magnitude_increase have the best performance, with magnitude_increase having slightly higher accuracy in Conv2 and Conv4. As expected, criteria using small weight values consistently perform worse than random. See Figure S1 for results on convergence speed.

(d) Making plots easier to read and interpret

We find that none of these three variants alone are able to train as well as the original LT network, shown as dashed lines in Figure 4. However, all three variants work better when we ensure that the new values of the kept weights are of the same sign as their original initial values. These are shown as solid color lines in Figure 4. Clearly, the common factor in all working variants including the original rewind action is the sign. As long as you keep the sign, reinitialization is not a deal breaker; in fact, even setting all kept weights to a constant value consistently performs well! The significance of the sign suggests, in contrast to [5], that the basin of attraction for a lottery ticket network is actually quite large: optimizers work well anywhere in the correct sign quadrant for the weights, but encounter difficulty crossing the zero barrier between signs.

(e) Clarification of main message and significance of mask-1 action

we extend this action to one-masked weights too, that is, initialize every weight to zero if they move towards zero during training, regardless of the pruning action on them. We see that performance of Variant 2 is even better than Variant 1, suggesting that this new mask-0 action we found can be a beneficial mask-1 action too. These results support our hypothesis that the benefit derived from freezing values to zero comes from the fact that those values were moving toward zero anyway[4]. This view on masking as training provides a new perspective on 1) why certain mask criteria work well (large_final and magnitude_increase both bias towards setting pruned weights close to their final values in the previous round of training), 2) the important contribution of the value of pruned weights to the overall performance of pruned networks, and 3) the benefit of setting these select weights to zero as a better initialization for the network.

(f) Clarification of main message and significance of mask-0 action

Figure S5: Test accuracy at early stopping iteration of additional mask criteria (large_final, large_final_same_sign, large_final_diff_sign) for four networks at various pruning rates.

(g) Trained results of large_final_same_sign and large_final_diff_sign

[1] The early stopping criterion we employ in this paper is the iteration of minimum validation loss.

[2] We run a t-test for each pruning percentage based on a sample of 5 independent runs for each mask criteria.

(h) Clarification of p-value