

---

# Deconstructing Lottery Tickets: Zeros, Signs, and the Supermask

---

**Hattie Zhou**

Uber

hattie@uber.com

**Janice Lan**

Uber AI

janlan@uber.com

**Rosanne Liu**

Uber AI

rosanne@uber.com

**Jason Yosinski**

Uber AI

yosinski@uber.com

## Abstract

The recent “Lottery Ticket Hypothesis” paper by Frankle & Carbin showed that a simple approach to creating sparse networks (keeping the large weights) results in models that are trainable from scratch, but only when starting from the same initial weights. The performance of these networks often exceeds the performance of the non-sparse base model, but for reasons that were not well understood. In this paper we study the three critical components of the Lottery Ticket (LT) algorithm, showing that each may be varied significantly without impacting the overall results. Ablating these factors leads to new insights for why LT networks perform as well as they do. We show why setting weights to zero is important, how signs are all you need to make the reinitialized network train, and why masking behaves like training. Finally, we discover the existence of Supermasks, masks that can be applied to an untrained, randomly initialized network to produce a model with performance far better than chance (86% on MNIST, 41% on CIFAR-10).

## 1 Introduction

Many neural networks are over-parameterized [3, 4], enabling compression of each layer [4, 21, 8] or of the entire network [14]. Some compression approaches enable more efficient computation by pruning parameters, by factorizing matrices, or via other tricks [8, 10, 13, 16–18, 20–23]. Unfortunately, although sparse networks created via pruning often work well, training sparse networks directly often fails, with the resulting networks underperforming their dense counterparts [16, 8].

A recent work by Frankle & Carbin [5] was thus surprising to many researchers when it presented a simple algorithm for finding sparse subnetworks within larger networks that *are* trainable from scratch. Their approach to finding these sparse, performant networks is as follows: after training a network, set all weights smaller than some threshold to zero, pruning them (similarly to other pruning approaches [9, 8, 15]), rewind the rest of the weights to their initial configuration, and then retrain the network from this starting configuration but with the zero weights frozen (not trained). Using this approach, they obtained two intriguing results.

First, they showed that the pruned networks performed well. Aggressively pruned networks (with 95 percent to 99.5 percent of weights pruned) showed no drop in performance compared to the much larger, unpruned network. Moreover, networks only moderately pruned (with 50 percent to 90 percent of weights pruned) often outperformed their unpruned counterparts. Second, they showed that these pruned networks train well only if they are rewound to their initial state, including the specific initial weights that were used. Reinitializing the same network topology with new weights causes it to train poorly. As pointed out in [5], it appears that the specific combination of pruning mask and weights underlying the mask form a more efficient subnetwork found within the larger network, or, as named by the original study, a lucky winning “Lottery Ticket,” or LT.

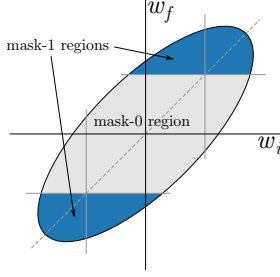


Figure 1: Different mask criteria can be thought of as segmenting the 2D ( $w_i$  = initial weight value,  $w_f$  = final weight value) space into regions corresponding to mask values of 1 vs 0. The ellipse represents in cartoon form the area occupied by the positively correlated initial and final weights from a given layer. The mask criterion shown, identified by two horizontal lines that separate the whole region into mask-1 (blue) areas and mask-0 (grey) areas, corresponds to the large\_final criterion used in [5]: weights with large final magnitude are kept and weights with final values near zero are pruned.

While Frankle & Carbin [5] clearly demonstrated LT networks to be effective, it raises many intriguing questions about the underlying mechanics of these subnetworks. What about LT networks causes them to show better performance? Why are the mask and the initial set of weights so tightly coupled, such that re-initializing the network makes it less trainable? Why does simply selecting large weights constitute an effective criterion for choosing a mask? We attempt to answer these questions by exploiting the essential steps in the lottery ticket algorithm, described below:

0. Initialize a mask  $m$  to all ones. Randomly initialize the parameters  $w$  of a network  $f(x; w \odot m)$
1. Train the parameters  $w$  of the network  $f(x; w \odot m)$  to completion. Denote the initial weights before training  $w_i$  and the final weights after training  $w_f$ .
2. *Mask Criterion.* Use the mask criterion  $M(w_i, w_f)$  to produce a masking score for each currently unmasked weight. Rank the weights in each layer by their scores, set the mask value for the top  $p\%$  to 1, the bottom  $(100 - p)\%$  to 0, breaking ties randomly. Here  $p$  may vary by layer, and we follow the ratios chosen in [5], summarized in Table S1. In [5] the mask selected weights with large final value corresponding to  $M(w_i, w_f) = |w_f|$ .
3. *Mask-1 Action.* Take some action with the weights with mask value 1. In [5] these weights were reset to their initial values and marked for training in the next round.
4. *Mask-0 Action.* Take some action with the weights with mask value 0. In [5] these weights were pruned: set to 0 and frozen during any subsequent training.
5. Repeat from 1 if performing iterative pruning.

In this paper we perform ablation studies along the above three dimensions of variability, considering alternate mask criteria (Section 2), alternate mask-1 actions (Section 3), and alternate mask-0 actions (Section 4). These studies in aggregate reveal new insights for why lottery ticket networks work as they do. Along the way we discover the existence of Supermasks—masks that produce above-chance performance when applied to untrained networks (Section 5). We make our code available at <https://github.com/uber-research/deconstructing-lottery-tickets>.

## 2 Mask criteria

We begin our investigation with a study of different *Mask Criteria*, or functions that decide which weights to keep vs. prune. In this paper, we define the mask for each individual weight as a function of the weight’s values both at initialization and after training:  $M(w_i, w_f)$ . We can visualize this function as a set of decision boundaries in a 2D space as shown in Figure 1. In [5], the mask criterion simply keeps weights with large final value; we refer to this as the large\_final mask,  $M(w_i, w_f) = |w_f|$ .

We experiment with mask criteria based on final weights (large\_final and small\_final), initial weights (large\_init and small\_init), a combination of the two (large\_init\_large\_final and small\_init\_small\_final), and how much weights move (magnitude\_increase and movement). We also include random as a control case, which chooses masks randomly. These nine masks are depicted along with their associated equations in Figure 2. Note that the main difference between magnitude\_increase and movement is that those weights that change sign are more likely to be kept in the movement criterion than the magnitude\_increase criterion.

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

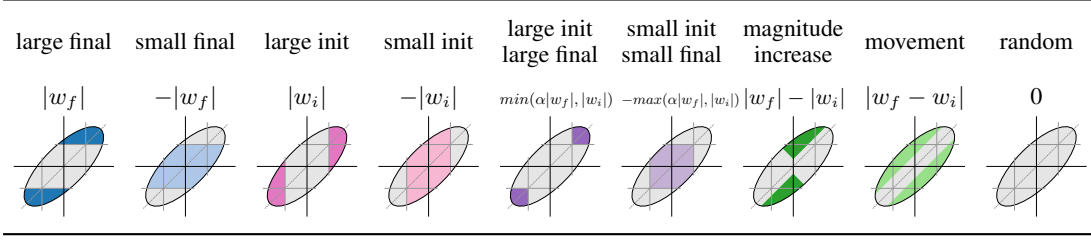


Figure 2: Mask criteria studied in this section, starting with `large_final` that was used in [5]. Names we use to refer to the various methods are given along with the formula that projects each  $(w_i, w_f)$  pair to a score. Weights with the largest scores (colored regions) are kept, and weights with the smallest scores (gray regions) are pruned. The  $x$  axis in each small figure is  $w_i$  and the  $y$  axis is  $w_f$ . In two methods,  $\alpha$  is adjusted as needed to align percentiles between  $w_i$  and  $w_f$ . When masks are created, ties are broken randomly, so a score of 0 for every weight results in random masks.

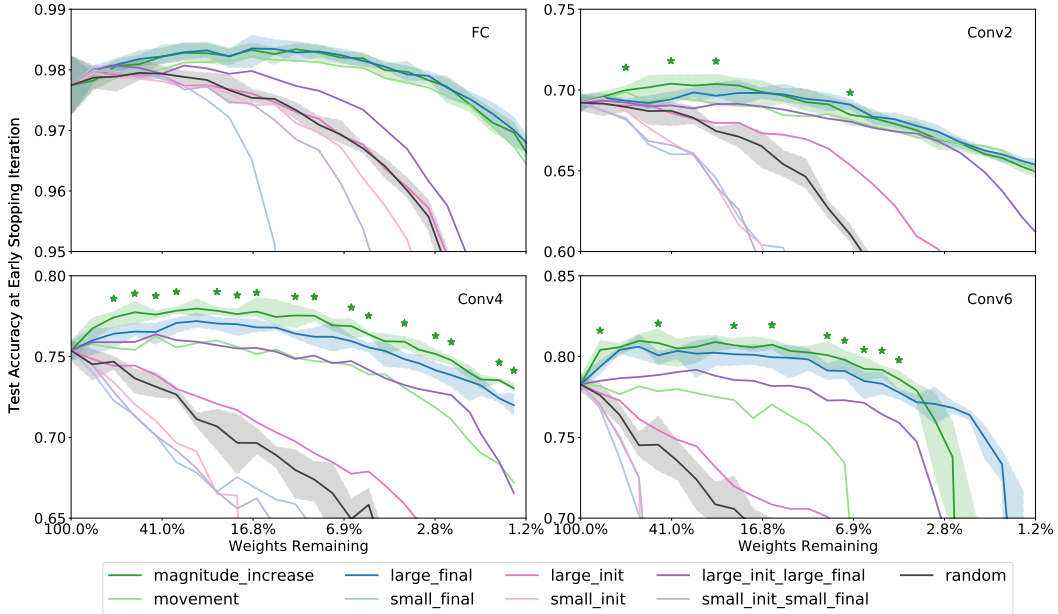


Figure 3: Test accuracy at early stopping iteration of different mask criteria for four networks at various pruning rates. Each line is a different mask criteria, with bands around the best-performing mask criteria (`large_final` and `magnitude_increase`) and the baseline (`random`) depicting the min 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. See Figure S1 for results on convergence speed.

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.

Results of all criteria are shown in Figure 3 for the four networks (FC, Conv2, Conv4, Conv6). The accuracy shown is the test accuracy at an early stopping iteration<sup>1</sup> of training. For all figures in this

<sup>1</sup>The early stopping criterion we employ in this paper is the iteration of minimum validation loss.

paper, the line depicts the mean over five runs, and the band (if shown) depicts the min and max obtained over five runs. In some cases the band is omitted for visual clarity.

Note that the first six criteria as depicted in Figure 2 form three opposing pairs; in each case, we observe when one member of the pair performs better than the random baseline, the opposing member performs worse than it. Moreover, the `magnitude_increase` criterion turns out to work just as well as the `large_final` criterion, and in some cases significantly better<sup>2</sup>.

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.

### 3 Mask-1 actions: the sign-ificance of initial weights

Now that we have explored various ways of choosing which weights to keep and prune, we will consider how we should initialize the kept weights. In particular, we want to explore an interesting observation in [5] which showed that the pruned, skeletal LT networks train well when you rewind to its original initialization, but degrades in performance when you randomly reinitialize the network.

Why does reinitialization cause LT networks to train poorly? Which components of the original initialization are important? To investigate, we keep all other treatments the same as [5] and perform a number of variants in the treatment of 1-masked, trainable weights, in terms of how to reinitialize them before the subnetwork training:

- “Reinit” experiments: reinitialize kept weights based on the original init distribution.
- “Reshuffle” experiments: reinitialize while respecting the original distribution of remaining weights in that layer by reshuffling the kept weights’ initial values.
- “Constant” experiments: reinitialize by setting 1-masked weight values to a positive or negative constant; thus every weight on a layer becomes one of three values:  $-\alpha$ , 0, or  $\alpha$ , with  $\alpha$  being the standard deviation of each layer’s original initialization.

All of the reinitialization experiments are based on the same original networks and use the `large_final` mask criterion with iterative pruning. We include the original LT network (“rewind, large final”) and the randomly pruned network (“random”) as baselines for comparison.

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 an LT 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.

### 4 Mask-0 actions: masking is training

What should we do with weights that are pruned? This question may seem trivial, as deleting them (equivalently: setting them to zero) is the standard practice. The term “pruning” implies the dropping 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.

<sup>2</sup>We run a t-test for each pruning percentage based on a sample of 5 independent runs for each mask criteria.

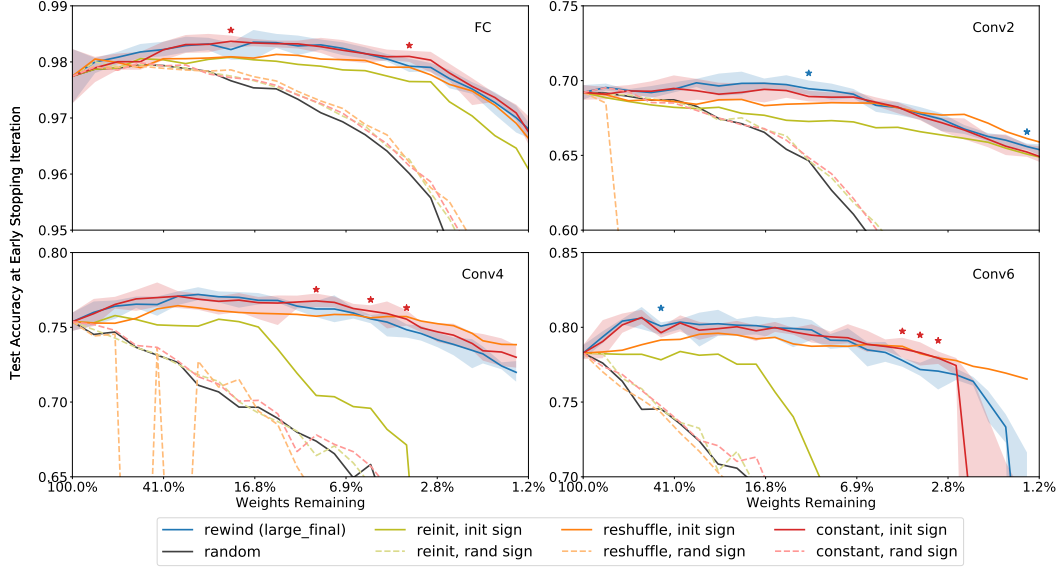


Figure 4: The effects of various 1-actions for four networks at various pruning rates. All reinitialization experiments use the large\_final mask criterion with iterative pruning. Dotted lines represent the three described methods, and solid lines are those three except with each weight having the same sign as its original initialization. Shaded bands around notable runs depict the min and max over 5 runs. Stars represent points where "rewind (large\_final)" or "constant, init sign" is significantly above the other at a  $p < 0.05$  level, showing no difference in performance between the two. The original large\_final and random are included as baselines. See Figure S4 for results on convergence speed.

Typical network pruning procedures [9, 8, 15] perform two actions on pruned weights: set them to zero, and freeze them in subsequent training (equivalent to removing those connections from the network). It is unclear which of these two components leads to the increased performance in LT networks. To separate the two factors, we run a simple experiment: we reproduce the LT iterative pruning experiments in which network weights are masked out in alternating train/mask/rewind cycles, but try an additional treatment: freeze masked weights at their initial values instead of at zero. If zero isn't special, both should perform similarly.

Figure 5 shows the results for this experiment. We find that networks perform significantly better when weights are frozen specifically at zero than at random initial values. For these networks masked via the LT large\_final criterion<sup>3</sup>, zero would seem to be a particularly good value to set pruned weights to. At high levels of pruning, freezing at the initial values may perform better, which makes sense since having a large number of zeros means having lots of dead connections.

So why does zero work better than initial values? One hypothesis is that the mask criterion we use *tends to mask to zero those weights that were headed toward zero anyway*.

To test out this hypothesis, we propose another mask-0 action halfway between freezing at zero and freezing at initialization: for any zero-masked weight, freeze it to zero if it moves toward zero over the course of training, and freeze it at its random initial value if it moves away from zero. We show two variants of this experiment in Figure 5. In the first variant, we apply it directly as stated to zero-masked weights (to be pruned). We see that by doing so we achieve comparable performance to the original LT networks at low pruning rates and better at high pruning rates. In the second variant, 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<sup>4</sup>. This

<sup>3</sup>Figure S3 illustrates why the large\_final criterion biases weights that were moving toward zero during training toward zero in the mask, effectively pushing them further in the direction they were headed.

<sup>4</sup>Additional control variants of this experiment can be seen in Supplementary Information Section S3.

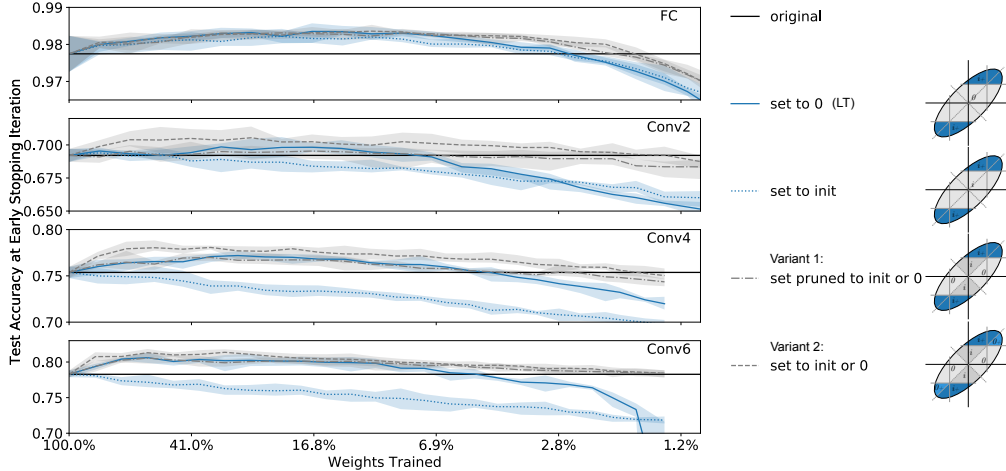


Figure 5: Performance of network pruning using different treatments of pruned weights (mask-0 actions). Horizontal black lines represent the performance of training the original, full network, averaged over five runs. Solid blue lines represent the original LT algorithm, which freezes pruned weights at zero. Dotted blue lines freeze pruned weights at their initial values. Grey lines show the new proposed 0-action—set to zero if they decreased in magnitude by the end of training, otherwise set to their initialization values. Two variants are shown: 1) new treatment applied to only pruned weights (dashdotted grey lines); 2) new treatment applied to all weights (dashed grey lines).

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.

## 5 Supermasks

The hypothesis above suggests that for certain mask criteria, like large\_final, that masking is training: *the masking operation tends to move weights in the direction they would have moved during training*. If so, just how powerful is this training operation? To answer this, we can start from the beginning—not training the network at all, but simply applying a mask to the randomly initialized network.

It turns out that with a well-chosen mask, an untrained network can already attain a test accuracy far better than chance. This might come as a surprise, because if you use a randomly initialized and untrained network to, say, classify images of handwritten digits from the MNIST dataset, you would expect accuracy to be no better than chance (about 10%). But now imagine you multiply the network weights by a mask containing only zeros and ones. In this instance, weights are either unchanged or deleted entirely, but the resulting network now achieves nearly 40 percent accuracy at the task! This is strange, but it is exactly what we observe with masks created using the large\_final criterion.

In randomly-initialized networks with large\_final masks, it is not implausible to have better-than-chance performance since the masks are derived from the training process. The large improvement in performance is still surprising, however, since the only transmission of information from the training back to the initial network is via a zero-one mask based on a simple criterion. We call masks that can produce better-than-chance accuracy without training of the underlying weights “Supermasks”.

We now turn our attention to finding better Supermasks. First, we simply gather all masks instantiated in the process of creating the networks shown in Figure 2, apply them to the original, randomly initialized networks, and evaluate the accuracy without training the network. Next, compelled by the demonstration in Section 3 of the importance of signs and in Section 4 of keeping large weights, we define a new large\_final\_same\_sign mask criterion that selects for weights with large final magnitudes that also maintained the same sign by the end of training. This criterion, as well as the control case of large\_final\_diff\_sign, is depicted in Figure 6. Performances of Supermasks produced by all 10 criteria are included in Figure 7, compared with two baselines: networks untrained

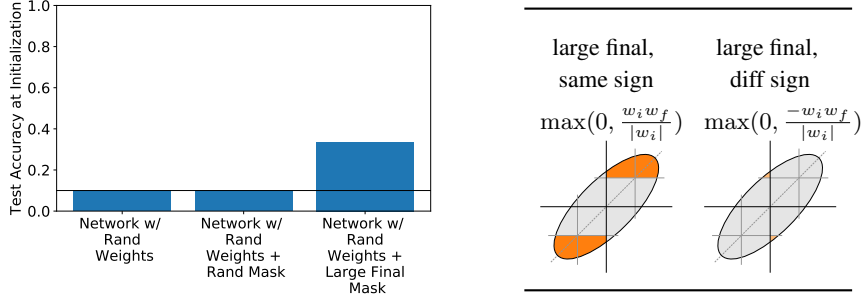


Figure 6: **(left)** Untrained networks perform at chance (10% accuracy) on MNIST, if they are randomly initialized, or randomly initialized and randomly masked. However, applying the large\_final mask improves the network accuracy beyond the chance level. **(right)** The large\_final\_same\_sign mask criterion (left) that tends to produce the best Supermasks. In contrast to the large\_final mask in Figure 1, this criterion masks out the quadrants where the sign of  $w_i$  and  $w_f$  differ. We include large\_final\_diff\_sign (right) as a control.

and unmasked (untrained\_baseline) and networks fully trained (trained\_baseline). For simplicity, we evaluate Supermasks based on one-shot pruning rather than iterative pruning.

We see that large\_final\_same\_sign significantly outperforms the other mask criteria in terms of accuracy at initialization. We can create networks that obtain a remarkable 80% test accuracy on MNIST and 24% on CIFAR-10 without training using this simple mask criterion. Another curious observation is that if we apply the mask to a signed constant (as described in Section 3) rather than the actual initial weights, we can produce even higher test accuracy of up to 86% on MNIST and 41% on CIFAR-10! Detailed results across network architectures, pruning percentages, and these two treatments, are shown in Figure 7.

We find it fascinating that these Supermasks exist and can be found via such simple criteria. As an aside, they also present a method for network compression, since we only need to save a binary mask and a single random seed to reconstruct the full weights of the network.

## 5.1 Optimizing the Supermask

We have shown that Supermasks derived using simple heuristics greatly enhance the performance of the underlying network immediately, with no training involved. In this section we are interested in how far we can push the performance of Supermasks by *training the mask*, instead of training network weights. Similar works in this domain include training networks with binary weights [1, 2], or training masks to adapt a base network to multiple tasks [19]. Our work differs in that the base network is randomly initialized, never trained, and masks are optimized for the original task.

We do so by creating a trainable mask variable for each layer while freezing all original parameters for that layer at their random initialization values. For an original weight tensor  $w$  and a mask tensor  $m$  of the same shape, we have as the effective weight  $w' = w \odot g(m)$ , where  $w_i$  denotes the initial values weights are frozen at,  $\odot$  is element-wise multiplication and  $g$  is a point-wise function that transform a matrix of continuous values into binary values.

We train the masks with  $g(m) = \text{Bern}(S(m))$ , where  $\text{Bern}(p)$  is the bernoulli sampler with probability  $p$ , and  $S(m)$  is the sigmoid function. The bernoulli sampling adds some stochasticity that helps with training, mitigates the bias of all things starting at the same value, and uses in effect the expected value of  $S(m)$ , which is especially useful when they are close to 0.5.

By training the  $m$  matrix with SGD, we obtained up to 95.3% test accuracy on MNIST and 65.4% on CIFAR-10. Results are shown in Figure 7, along with all the heuristic-based, unlearned Supermasks. Note that there is no straightforward way to control for the pruning percentage. Instead, we initialize  $m$  with larger or smaller magnitudes, which nudges the network toward pruning more or less. This allows us to produce masks with the amounts of pruning (percentages of zeros) ranging from 7% to 89%. Further details about the training can be seen in Section S6.



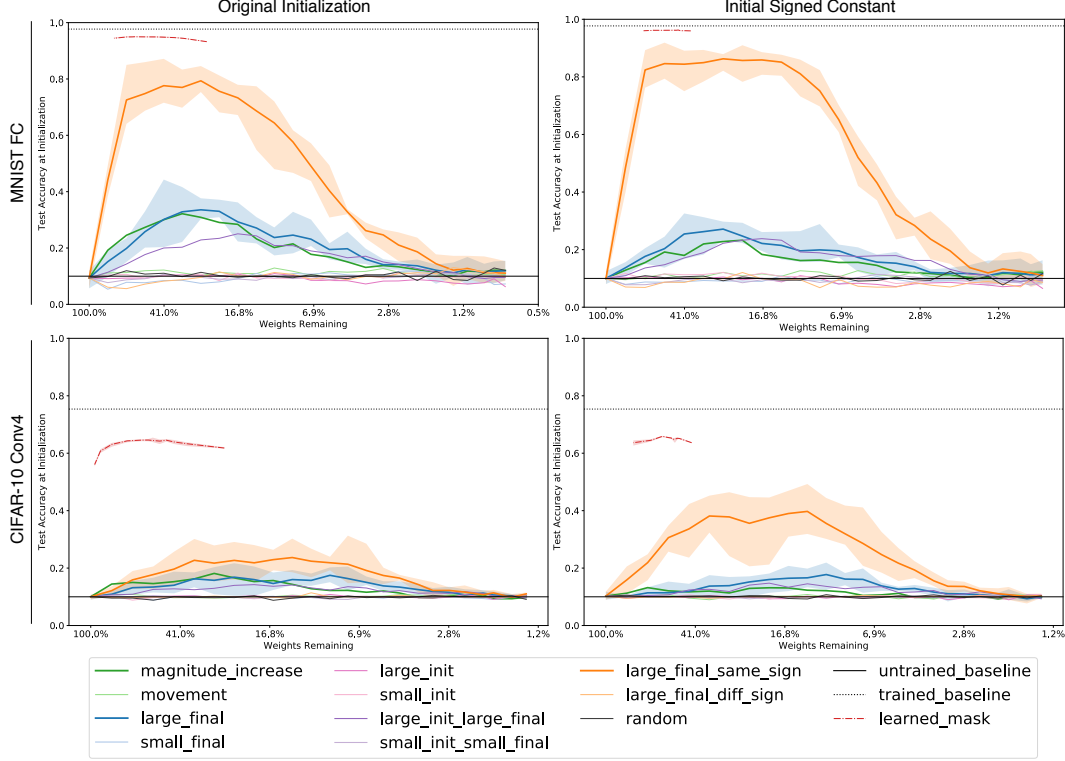


Figure 7: Comparison of Supermask performances in terms of test accuracy on MNIST and CIFAR-10 classification tasks. Subfigures are across two network structures (top: FC on MNIST, bottom: Conv4 on CIFAR-10), as well as 1-action treatments (left: weights are at their original initialization, right: weights are converted to signed constants). No training is performed in any network. Within heuristic based Supermasks (excluding learned\_mask), the large\_final\_same\_sign mask creates the highest performing Supermask by a wide margin. Note that aside from the five independent runs performed to generate uncertainty bands for this plot, every point on this plot is from the same underlying network, just with different masks. See Figure S6 for performance on all four networks.

## 5.2 Dynamic Weight Rescaling

One beneficial trick in Supermask training is to dynamically rescale the values of weights based on the sparsity of the network in the current training iteration. For each training iteration and for each layer, we multiply the underlying weights by the ratio of the total number of weights in the layer over the number of ones in the corresponding mask. Dynamic rescaling leads to significant improvements in the performance of the masked networks, which is illustrated in Table 1.

Table 1 summarizes the best test accuracy obtained through different treatments. The result shows striking improvement of learned Supermasks over heuristic based ones. Learned Supermasks result in performance close to training the full network, which suggests that a network upon initialization already contains powerful subnetworks that work well without training.

## 6 Conclusion

In this paper, we have studied how three components to LT-style network pruning—mask criterion, treatment of kept weights during retraining (mask-1 action), and treatment of pruned weights during retraining (mask-0 action)—come together to produce sparse and performant subnetworks. We proposed the hypothesis that networks work well when pruned weights are set close to their final values. Building on this hypothesis, we introduced alternative freezing schemes and other mask criteria that meet or exceed current approaches by respecting this basic rule. We also showed that the only element of the original initialization that is crucial to the performance of LT networks is the sign,



Table 1: Test accuracy of the best Supermasks with various initialization treatments. Values shown are the max over any prune percentage and averaged over four or more runs. The first two columns show untrained networks with heuristic-based masks, where “init” stands for the initial, untrained weights, and “S.C.” is the signed constant approach, which replaces each random initial weight with its sign as described in Section 3. The next two columns show results for untrained weights overlaid with learned masks; and the two after add the Dynamic Weight Rescaling (DWR) approach. The final column shows the performance of networks with weights trained directly using gradient descent. Bold numbers show the performance of the best Supermask variation.

Network	mask		learned mask		DWR learned mask	DWR learned mask	trained weights
	$\odot$	$\odot$	$\odot$	$\odot$	$\odot$	$\odot$	
	init	S.C.	init	S.C.	init	S.C.	
MNIST FC	79.3	86.3	95.3	96.4	97.8	<b>98.0</b>	97.7
CIFAR Conv2	22.3	37.4	64.4	<b>66.3</b>	65.0	66.0	69.2
CIFAR Conv4	23.	39.7	65.4	66.2	71.7	<b>72.5</b>	75.4
CIFAR Conv6	24.0	41.0	65.3	65.4	76.3	<b>76.5</b>	78.3

not the relative magnitude of the weights. Finally, we demonstrated that the masking procedure can be thought of as a training operation, and consequently we uncovered the existence of Supermasks, which can produce partially working networks without training.

### Acknowledgments

The authors would like to acknowledge Jonathan Frankle, Joel Lehman, Zoubin Ghahramani, Sam Greydanus, and members of the Deep Collective research group at Uber AI for combinations of helpful discussion, ideas, feedback on experiments, and comments on early drafts of this work.

## References

- [1] Matthieu Courbariaux, Yoshua Bengio, and Jean-Pierre David. Binaryconnect: Training deep neural networks with binary weights during propagations. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems* 28, pages 3123–3131. Curran Associates, Inc., 2015.
- [2] Matthieu Courbariaux, Itay Hubara, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. Binarized neural networks: Training deep neural networks with weights and activations constrained to+ 1 or-1. *arXiv preprint arXiv:1602.02830*, 2016.
- [3] Yann Dauphin and Yoshua Bengio. Big neural networks waste capacity. *CoRR*, abs/1301.3583, 2013.
- [4] Misha Denil, Babak Shakibi, Laurent Dinh, Nando de Freitas, et al. Predicting parameters in deep learning. In *Advances in Neural Information Processing Systems*, pages 2148–2156, 2013.
- [5] Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations (ICLR)*, volume abs/1803.03635, 2019.
- [6] Jonathan Frankle, Gintare Karolina Dziugaite, Daniel M. Roy, and Michael Carbin. The lottery ticket hypothesis at scale. *CoRR*, abs/1903.01611, 2019.
- [7] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *International Conference on Artificial Intelligence and Statistics*, pages 249–256, 2010.
- [8] Song Han, Huizi Mao, and William J. Dally. Deep compression: Compressing deep neural network with pruning, trained quantization and huffman coding. *CoRR*, abs/1510.00149, 2015.
- [9] Song Han, Jeff Pool, John Tran, and William Dally. Learning both weights and connections for efficient neural network. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems* 28, pages 1135–1143. Curran Associates, Inc., 2015.
- [10] Babak Hassibi and David G. Stork. Second order derivatives for network pruning: Optimal brain surgeon. In S. J. Hanson, J. D. Cowan, and C. L. Giles, editors, *Advances in Neural Information Processing Systems* 5, pages 164–171. Morgan-Kaufmann, 1993.
- [11] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.
- [12] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- [13] Yann LeCun, John S. Denker, and Sara A. Solla. Optimal brain damage. In D. S. Touretzky, editor, *Advances in Neural Information Processing Systems* 2, pages 598–605. Morgan-Kaufmann, 1990.
- [14] Chunyuan Li, Heerad Farkhoor, Rosanne Liu, and Jason Yosinski. Measuring the Intrinsic Dimension of Objective Landscapes. In *International Conference on Learning Representations*, April 2018.
- [15] Hao Li, Asim Kadav, Igor Durdanovic, Hanan Samet, and Hans Peter Graf. Pruning filters for efficient convnets. *arXiv preprint arXiv:1608.08710*, 2016.
- [16] Hao Li, Asim Kadav, Igor Durdanovic, Hanan Samet, and Hans Peter Graf. Pruning filters for efficient convnets. In *International Conference on Learning Representations (ICLR)*, volume abs/1608.08710, 2017.
- [17] C. Louizos, K. Ullrich, and M. Welling. Bayesian Compression for Deep Learning. *ArXiv e-prints*, May 2017.

- [18] Jian-Hao Luo, Jianxin Wu, and Weiyao Lin. Thinet: A filter level pruning method for deep neural network compression. In *Proceedings of the IEEE international conference on computer vision (ICCV)*, pages 5058–5066, 2017.
- [19] Arun Mallya, Dillon Davis, and Svetlana Lazebnik. Piggyback: Adapting a single network to multiple tasks by learning to mask weights. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 67–82, 2018.
- [20] Pavlo Molchanov, Stephen Tyree, Tero Karras, Timo Aila, and Jan Kautz. Pruning convolutional neural networks for resource efficient inference. In *International Conference on Learning Representations (ICLR)*, 2017.
- [21] Wei Wen, Chunpeng Wu, Yandan Wang, Yiran Chen, and Hai Li. Learning structured sparsity in deep neural networks. *CoRR*, abs/1608.03665, 2016.
- [22] Tien-Ju Yang, Yu-Hsin Chen, and Vivienne Sze. Designing energy-efficient convolutional neural networks using energy-aware pruning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5687–5695, 2017.
- [23] Zichao Yang, Marcin Moczulski, Misha Denil, Nando de Freitas, Alex Smola, Le Song, and Ziyu Wang. Deep fried convnets. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1476–1483, 2015.

# Supplementary Information for: Deconstructing Lottery Tickets: Zeros, Signs, and the Supermask

## S1 Architectures and training hyperparameters

Table S1 contains the architectures used in this study, together with relevant training hyperparameters, based off of experiments in [5].

**Additional details** To compare with results in [5], we used the same training hyperparameters and did not do any additional tuning. Training/test splits were given in both MNIST and CIFAR-10; validation was split randomly from the training set with (55000, 5000) train and val for MNIST and (45000, 5000) for CIFAR.

Our experiments required more computation than regular training procedures, as networks were trained up to 24 times with iterative pruning. We used single GPUs for each experiment (NVIDIA GeForce GTX 1080 Ti) and parallelized by running multiple experiments on multiple GPUs.

Table S1: The architectures used in this paper. Table reproduced and modified from [5]. Conv networks use 3x3 convolutional layers with max pooling followed by fully connected layers. FC layer sizes are from [12]. Initializations are Glorot Normal [7] and activations are ReLu.

<i>Network</i>	MNIST FC	CIFAR-10 Conv2	CIFAR-10 Conv4	CIFAR-10 Conv6
<i>Convolutional Layers</i>		64, 64, pool	64, 64, pool 128, 128, pool	64, 64, pool 128, 128, pool 256, 256, pool
<i>FC Layers</i> None	300, 100, 10	256, 256, 10	256, 256, 10	256, 256, 10
<i>All/Conv Weights</i>	266K	4.3M / 38K	2.4M / 260K	2.3M / 1.1M
<i>Iterations/Batch</i>	50K / 60	20K / 60	25K / 60	30K / 60
<i>Optimizer</i>	Adam 1.2e-3	Adam 2e-4	Adam 3e-4	Adam 3e-4
<i>Pruning Rates</i>	fc20%	conv10% fc20%	conv10% fc20%	conv15% fc20%

## S2 Further mask criteria details

Figure S1 shows the convergence speed and performance of all mask criteria for FC on MNIST and Conv2, 4, 6 on CIFAR-10.

## S3 Further mask-0 action details

In this section, we discuss additional control variants of the mask-0 action experiments shown in Figure 5. In particular, we want to know if the improvement in performance from setting only weights that move towards zero to zero is the result of this particular selection of zero weights, rather than other quirks in the treatment. We run two control cases. In the first control experiment, we randomly freeze a subset of pruned weights to zero, with the number of zero-ed weights equal to the number of

zero-ed weights in "set pruned to init or 0". In the second control experiment, we set weights that *moved away from zero* to zero, reversing the proposed treatment. As shown in Figure S2, both control experiments perform significantly worse than the proposed treatment ("set pruned to init or 0"), with the reversed treatment performing worse than both randomly freezing weights to zero and freezing all pruned weights at initial value. These results are consistent with our hypothesis.

#### S4 Further mask-1 action details

Figure S4 shows the convergence speed and performance of various reinitialization methods for FC on MNIST and Conv2, 4, 6 on CIFAR-10.

#### S5 Further heuristic Supermask details

We show the performance of trained LT networks with mask criteria `large_final_same_sign` and `large_final_diff_sign` in Figure S5. We see that despite having better test accuracy at iteration 0, `large_final_same_sign` slightly underperforms `large_final` when the underlying network is fully trained. This may be due to the fact that the `large_final_same_sign` mask criterion randomly select weights that changed signs to keep when there are not enough same sign weights.

#### S6 Further training details for learning Supermasks

We train the networks with mask  $m$  for each layer (and all regular kernels and biases frozen) with SGD, 0.9 momentum. The {FC, Conv2, Conv4, Conv6} networks respectively had {100, 100, 50, 20} for learning rates and trained for {2000, 2000, 1000, 800} iterations. These hyperparameters may seem absurd, but a network of masks is quite different and cannot train well with typical learning rates. Conv4 and Conv6 showed significant overfitting, thus we used early stopping as we are unable to use standard regularizing techniques. For evaluation, we also use Bernoulli sampling, but average the accuracies over 10 independent samples.

For adjusting the amount pruned, we initialized  $m$  in every layer to be the same constant, which ranged from -5 to 5. In the future it may be worth trying different initializations of  $m$  for each layer for more granular control over per-layer pruning rates. A different method to try would be to add an L1 loss to influence layers to go toward certain values, which may alleviate the cold start problems of some networks not learning anything due to mask values starting too low (effectively having the entire network start at zero).

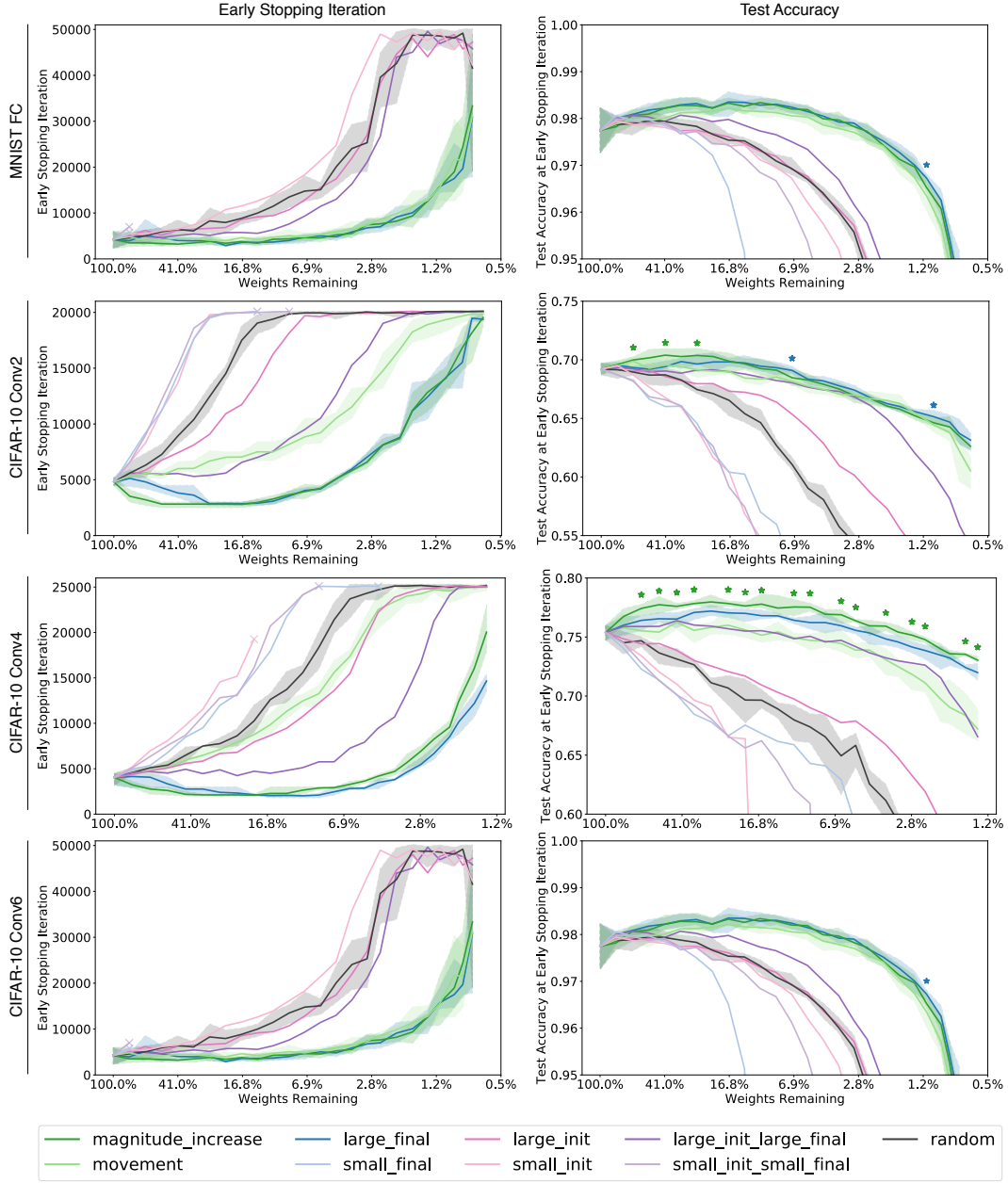


Figure S1: Performance of different mask criteria for four networks at various pruning rates. We show early stopping iteration on the left and test accuracy on the right. Each line is a different mask criteria, with bands around magnitude\_increase, large\_final, movement, and random depicting the min and max over 5 runs. Stars represent points where large\_final or magnitude\_increase are significantly above the other at a  $p < 0.05$  level. large\_final and magnitude\_increase show the best convergence speed and accuracy, with magnitude\_increase having slightly higher accuracy in Conv2 and Conv4. As expected, criteria using small weight values consistently perform worse than random.

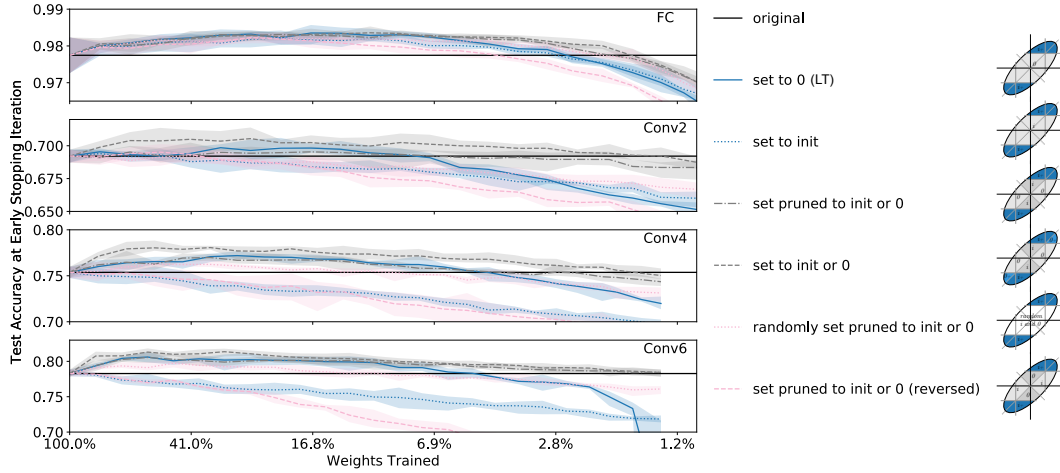


Figure S2: Performance of network pruning using different treatments of pruned weights (mask-0 actions). Horizontal black lines represent the performance of training the original, full network, averaged over five runs. Solid blue lines represent the original LT algorithm, which freezes pruned weights at zero. Dotted blue lines freeze pruned weights at their initial values. Grey lines show the new proposed 0-action—set to zero if they decreased in magnitude by the end of training, otherwise set to their initialization values. Two variants are shown: 1) new treatment applied to only pruned weights (dashdotted grey lines); 2) new treatment applied to all weights (dashed grey lines). Two additional control experiments are shown: 1) randomly freeze a number of pruned weights to zero, with the number equal to the number of zero-ed weights in the new proposed 0-action ("set pruned to init or 0"); 2) reverse the proposed 0-action by setting to zero weights that *increased* in magnitude by the end of training.

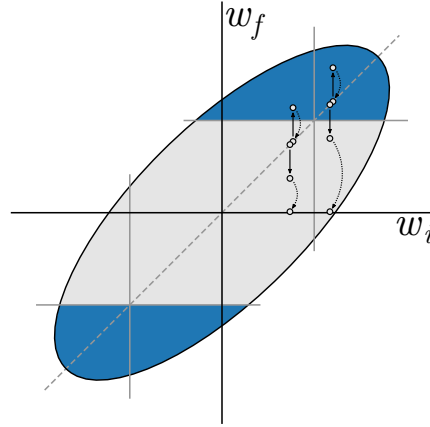


Figure S3: The motion selectivity of the large\_final mask criterion. Although large\_final only selects for large final weights, not directly for weight motion, it biases towards pruning weights that *decreased* in magnitude, since those weights are more likely to have small final magnitudes. As shown by the two paired sets of example weights, those two weights that increased during training (two solid up arrows) are reset to their initial value (dotted arrows) after pruning, resulting in no net motion. In contrast, those two weights that decrease in value during training (two solid down arrows) are set to zero (dotted lines) during pruning, moving them preferentially in the direction they moved during training. This explains why Supermasks may be created by such a simple criterion as large\_final and motivates the more specifically designed, higher performing large\_final\_same\_sign criterion (depicted in Figure 6).



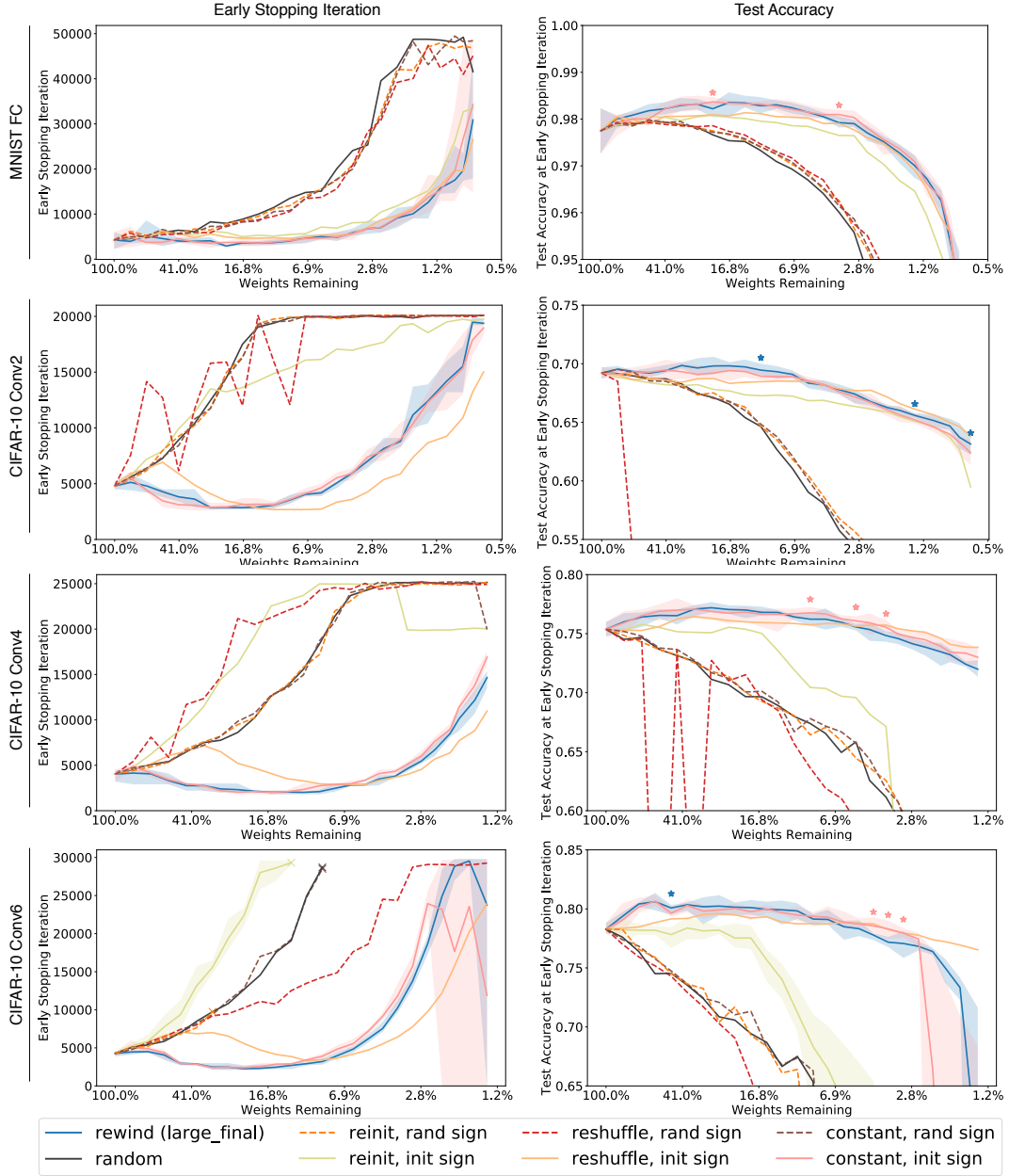


Figure S4: The effects of various 1-actions for the four networks and various pruning rates. Dotted lines represent the three described methods, and solid lines are those except with each weight having the same sign as its original initialization. Shaded bands around notable runs depict the min and max over 5 runs. Stars represent points where "rewind (large\_final)" or "constant, init sign" is significantly above the other at a  $p < 0.05$  level, showing no difference in performance between the two. We also include the original rewinding method and random pruning as baselines. "Reshuffle, init sign" and "constant, init sign" perform similarly to the "rewind" baseline.

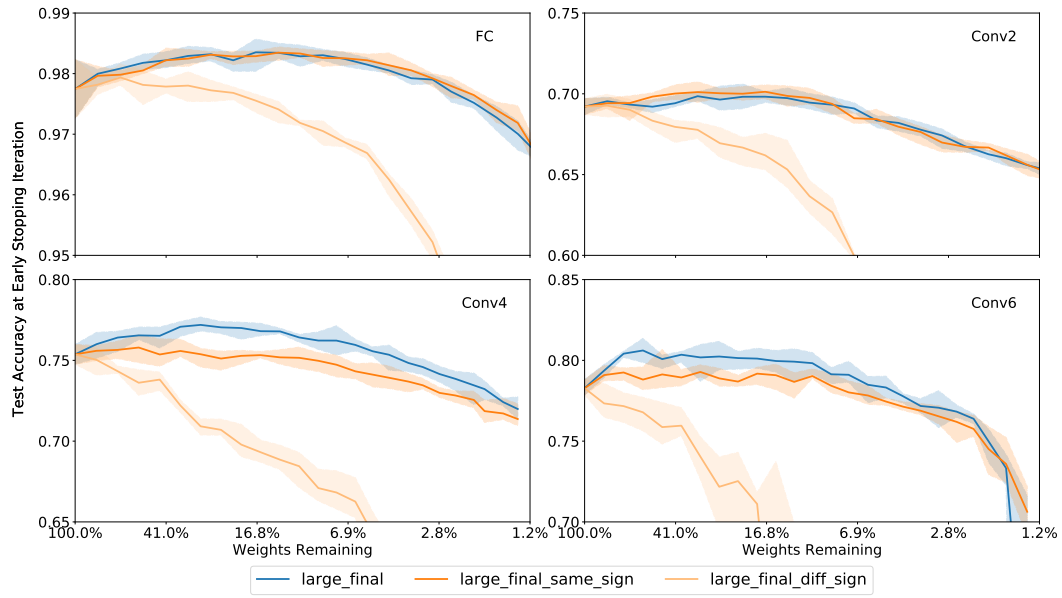


Figure S5: Test accuracy at early stopping iteration of additional mask criteria (large\_final, large\_final\_same\_sign, large\_final\_diff\_sign) for four networks at various pruning rates.

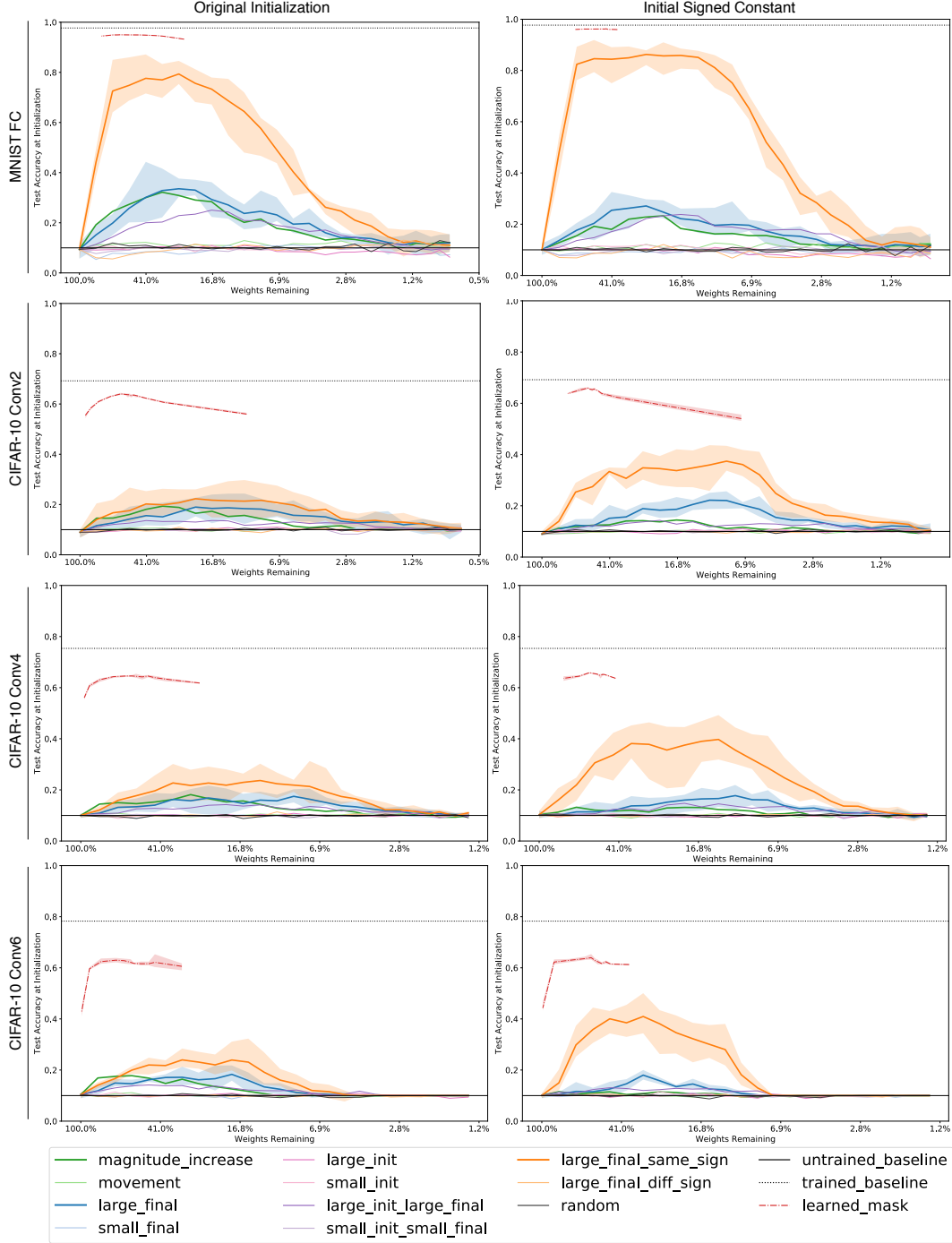


Figure S6: Comparison of Supermask performances in terms of test accuracy on MNIST and CIFAR-10 classification tasks. Subfigures are across various network structures (from top row to bottom: FC on MNIST, Conv2 on CIFAR-10, Conv4 on CIFAR-10, Conv6 on CIFAR-10), as well as 1-action treatments (left: weights are frozen at their original initialization, right: weights are frozen at a signed constant). No training is performed in any network. Weights are frozen at either initialization or constant and various masks are applied. Within heuristic based Supermasks (excluding learned\_mask), the large\_final\_same\_sign mask creates the highest performing Supermask by a wide margin. Note that aside from the five independent runs performed to generate uncertainty bands for this plot, every data point on the plot is the same underlying network, just with different masks.