The Fragility of Polarity: A Perturbative Analysis of the sign Hypothesis in Sparse Networks

Anonymous Author(s)
Affiliation
Address
email

Abstract

The "sign Hypothesis" posits that preserving the initial signs of weights, rather than their exact magnitudes, is critical for successfully training sparse subnetworks found via the Lottery Ticket Hypothesis. While well-supported, the robustness of this sign structure across different task complexities remains unexplored. This paper introduces a framework of systematic, perturbative reinitialization strategies to probe the breaking points of sign-based training. We apply controlled perturbations—including random sign-flipping (Epsilon), targeted randomization (Delta), and threshold shifting (Phi)—to various convolutional architectures across MNIST, CIFAR-10, and CIFAR-100. Our findings reveal a clear "complexity gradient": on MNIST, the sign structure is extremely robust, with networks maintaining >97% accuracy even with 90% of signs flipped. On the more complex CIFAR-10, the classic lottery ticket rewinding strategy ('mask1') becomes fragile, while strategies incorporating learned polarity ('gradual', 'mask0') prove far more resilient. On CIFAR-100, this gap widens, with the choice of reinitialization strategy becoming a critical performance differentiator. A consistent finding across all datasets is a remarkable resilience to sign randomization of the lowest-magnitude weights, suggesting that magnitude pruning is effective precisely because it preserves the weights whose polarity is most critical for guiding optimization.

1 Introduction

The Lottery Ticket Hypothesis (LTH) [1] has reshaped our understanding of network pruning, demonstrating that sparse subnetworks capable of training to full accuracy exist at initialization. The standard algorithm for finding these "winning tickets," Iterative Magnitude Pruning (IMP), relies on rewinding surviving weights to their initial values. This critical step has led to a deeper question: what property of the initial weights is so essential?

A compelling answer was provided by Zhou et al. (2019) [2], who demonstrated that the signs of the initial weights are more important than their specific magnitudes. This "sign Hypothesis" has been reinforced by subsequent works showing that effective pruning algorithms excel at identifying and maintaining this crucial sign information [3, 4].

While the primacy of signs is established, how this principle generalizes across tasks of varying complexity is not well understood. This paper investigates the robustness of the initial sign configuration by asking: How does the fragility of a network's polarity change as the learning task becomes more difficult? Are all signs equally important across different datasets? To answer these questions, we introduce a suite of controlled, perturbative reinitialization strategies. By systematically damaging the polarity of networks trained on MNIST, CIFAR-10, and CIFAR-100, we map the landscape of sign fragility as a function of task complexity.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

Our contributions are: (1) We formalize a framework of perturbative analysis for the sign Hypothesis. (2) We identify a "complexity gradient": the initial sign structure is extremely robust on simple tasks like MNIST but becomes increasingly fragile on more complex datasets like CIFAR-10 and CIFAR-100. (3) We show that while the classic LTH rewinding strategy ('mask1') is effective, its fragility on complex tasks is overcome by strategies that incorporate learned polarity ('gradual', 'mask0'). (4) We consistently find that networks are resilient to sign randomization of low-magnitude weights, providing a deeper justification for magnitude-based pruning.

2 Methodology: Probing Polarity with Perturbations

Our methodology extends the standard IMP workflow. At each pruning iteration, after identifying weights to be pruned based on magnitude, we reinitialize the surviving weights. This reinitialization is governed by two orthogonal components: a *reference strategy*, which determines the source of information for the new weights, and a *perturbation strategy*, which applies controlled noise to the weight signs. All strategies ultimately decouple sign from magnitude by setting the final magnitude of all surviving weights to a constant value (the standard deviation of the initial layer-wise distribution), thereby isolating the effect of the sign configuration.

2.1 Reinitialization Reference Strategies

We investigate four strategies for determining the reference polarity of the surviving weights.

- **Mask1 (Initial Weights):** The reference signs are taken from the network's original weights at initialization ($t = 0$). This is the classic LTH-style rewinding of polarity.
- **Mask0 (Trained Weights):** The reference signs are taken from the weights at the end of the most recent training iteration.
- **Gradual (Mixed):** A stochastic mix where 70% of weights reference the trained signs and 30% reference the initial signs.
- **Standard (Full):** References the trained signs but does not enforce the pruning mask, effectively "reviving" pruned weights.

2.2 Sign Perturbation Methods

After establishing a reference sign for each weight, we apply a perturbation.

Epsilon (ϵ) Perturbation To measure sensitivity to random noise, this strategy randomly flips the signs of a fraction ϵ of the surviving weights.

Delta (δ) Percentile Perturbation To test the hypothesis that the signs of low-magnitude weights are less critical, this strategy targets a fixed percentile, δ, of the surviving weights with the smallest absolute magnitudes and randomizes their signs.

Phi (ϕ) Perturbation This strategy tests a non-standard polarity definition by creating a "dead zone" around zero. Signs are assigned as $+1$ if $w > \phi$ and -1 if $w < \phi$.

3 Experimental Results and Analysis

We conducted experiments across MNIST, CIFAR-10, and CIFAR-100 using various architectures. All models were pruned over 10 iterations with a pruning base of 0.8, reaching a final sparsity of approximately 89%.

3.1 Extreme Robustness on a Simple Task: MNIST

On the MNIST dataset with a fully-connected architecture, the network's polarity configuration is exceptionally robust. Figure 1 shows the final test accuracy under Epsilon perturbation. Remarkably, performance remains almost entirely unaffected, with all strategies maintaining >97% accuracy even

when 90% of the signs are randomly flipped. The 'standard' reinitialization strategy consistently achieves the highest performance, though the differences between strategies are minimal. This suggests that for simple tasks with clear feature separation, the initial sign structure contains a high degree of redundancy, and the optimizer can easily converge to a correct solution even with a heavily damaged polarity map.

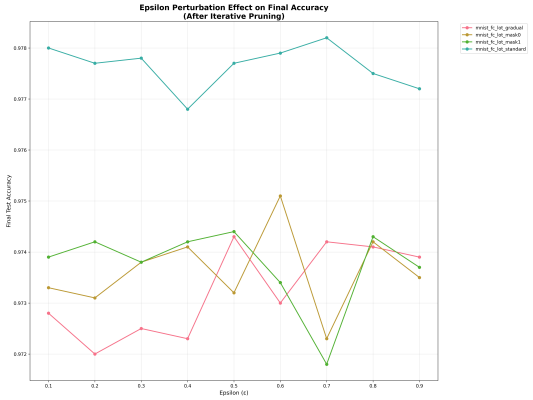

Figure 1: Effect of Epsilon (ϵ) perturbation on a fully-connected network trained on MNIST. Performance is extremely stable across all strategies, with final accuracy remaining above 97% even at high levels of sign perturbation.

Table 1: Best final test accuracy (%) on MNIST across all Epsilon perturbation experiments.

Architecture	Reinitialization Strategy	Best Final Accuracy (%)
Fully-Connected	gradual	97.4
	mask0	97.5
	mask1	97.4
	standard	**97.8**

3.2 Increasing Fragility on a More Complex Task: CIFAR-10

On the more challenging CIFAR-10 dataset, a clear differentiation in robustness emerges. The classic LTH rewinding strategy ('mask1') becomes notably fragile. As shown in the VGG results, the 'mask1' strategy (green line) suffers a catastrophic performance drop under Epsilon perturbation. In contrast, strategies that incorporate learned polarity information ('mask0', 'standard', 'gradual') are far more resilient.

However, when subjected to the more targeted Delta Percentile perturbation, all strategies prove to be highly robust. As seen in Table 2, which summarizes the best performance across architectures, the 'gradual' and 'standard' strategies consistently perform well, often achieving over 70% accuracy. This demonstrates that while the overall sign structure becomes more critical on CIFAR-10, the signs of the lowest-magnitude weights remain less important.

Table 2: Best final test accuracy (%) on CIFAR-10 across all delta percentile perturbation experiments.

Architecture	gradual	mask0	mask1	standard
Conv2	65.6	65.1	64.5	65.3
Conv4	**69.0**	66.9	66.8	66.3
Conv6	**72.2**	69.6	67.2	68.1
VGG	79.6	79.1	72.0	**79.6**

3.3 Polarity as a Critical Performance Differentiator: CIFAR-100

On CIFAR-100, the most complex task, the choice of reinitialization strategy becomes paramount. The results for Delta Percentile perturbation on a VGG network (Table 3) show a wide performance gap between strategies. Here, the 'gradual' and 'mask1' strategies emerge as the top performers, achieving nearly 59% and 56% accuracy, respectively. This is significantly higher than the 'mask0' and 'standard' strategies. This shift in top-performing strategies from CIFAR-10 to CIFAR-100 suggests that on very complex tasks, a strong connection to the initial weight configuration ('mask1') or a careful blend of initial and learned information ('gradual') is necessary to prevent the optimizer from deviating into unproductive regions of the loss landscape.

Table 3: Best final test accuracy (%) on CIFAR-100 with a VGG architecture across all Delta Percentile perturbation experiments.

Architecture	Reinitialization Strategy	Best Final Accuracy (%)
VGG	gradual	**58.9**
	mask1	56.0
	mask0	43.5
	standard	40.8

4 Discussion and Conclusion

Our cross-dataset perturbative analysis reveals a clear "complexity gradient" for the role of weight polarity in sparse networks. On simple tasks like MNIST, the sign structure is incredibly robust, and nearly any sign-preserving reinitialization strategy suffices. As task complexity increases to CIFAR-10, the initial polarity becomes more fragile; the classic LTH rewinding strategy ('mask1') falters under random noise, while strategies that incorporate learned polarity demonstrate superior robustness. On the highly complex CIFAR-100, this differentiation becomes even more stark, with strategies that maintain a strong link to the initial configuration ('gradual', 'mask1') proving most effective.

A unifying principle across all datasets is the network's resilience to targeted, low-magnitude sign perturbation. This provides a powerful, refined justification for the success of Iterative Magnitude Pruning: IMP works because it implicitly preserves the weights whose signs are most critical for defining a favorable optimization landscape. The signs of smaller weights, which are pruned first, are less important and can tolerate significant noise.

As a follow-up, in the future, we wish to make progress on the following question: What happens if instead of IMP we use more structural pruning for instance, spectral pruning? Does the polarity perturbation follow the same pattern?

References

[1] Frankle, J. & Carbin, M. (2019). The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks. In *International Conference on Learning Representations (ICLR)*.

[2] Zhou, H., Lan, J., Liu, R., & Yosinski, J. (2019). Deconstructing lottery tickets: Zeros, signs, and the supermask. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[3] Gadhikar, A., & Burkholz, R. (2024). Masks, Signs, and Learning Rate Rewinding. In *International Conference on Learning Representations (ICLR)*.

[4] Oh, J., Baik, S., & Lee, K. M. (2025). Find a winning sign: Sign is all we need to win the Lottery. In *International Conference on Learning Representations (ICLR)*.

A Appendix

This appendix documents the hyperparameters used in the experiments across different datasets and model architectures.

MNIST

Parameter	Value
Architecture	Fully Connected: $784 \to 300 \to 100 \to 10$
Number of iterations	10
Prune base	0.8
Epochs per iteration	10

Table 4: Hyperparameters for MNIST experiments

CIFAR-10

Parameter	Value
Architectures	conv2, conv4, conv6, vanilla_vgg16
Number of iterations	10
Prune base	0.8
Epochs per iteration	10

Table 5: Hyperparameters for CIFAR-10 experiments

CIFAR-100

Parameter	Value
Architecture	baseline_vgg16
Number of iterations	10
Prune base	0.8
Epochs per iteration	50

Table 6: Hyperparameters for CIFAR-100 experiments

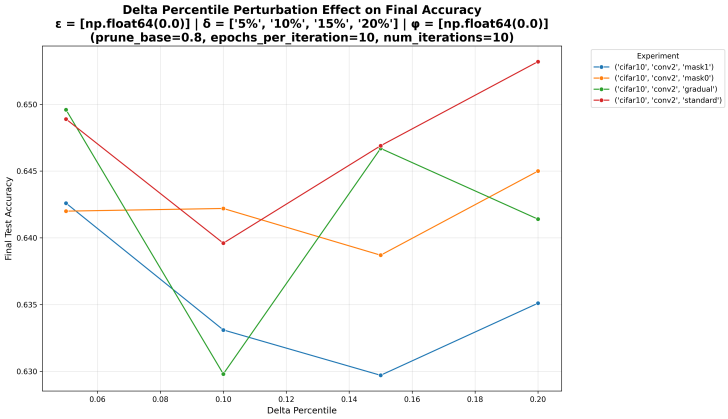

Figure 2: cifar10 conv2

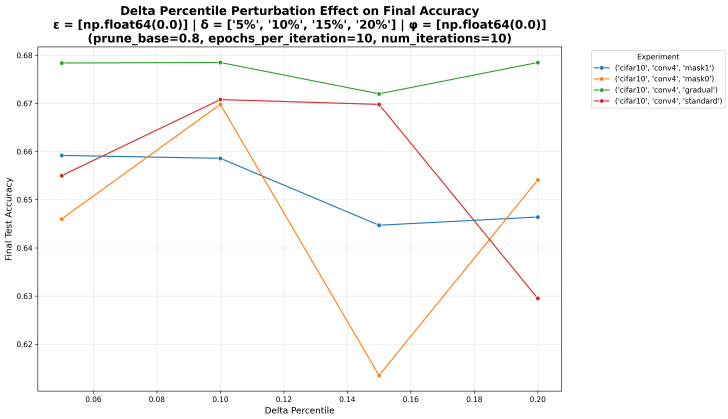

Figure 3: cifar10 conv4

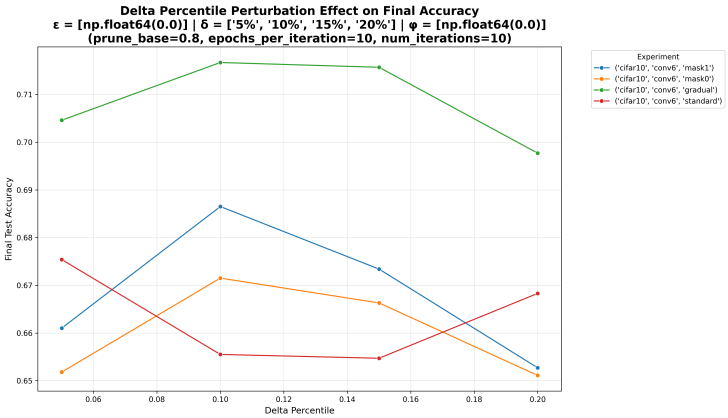

Figure 4: cifar10 conv6

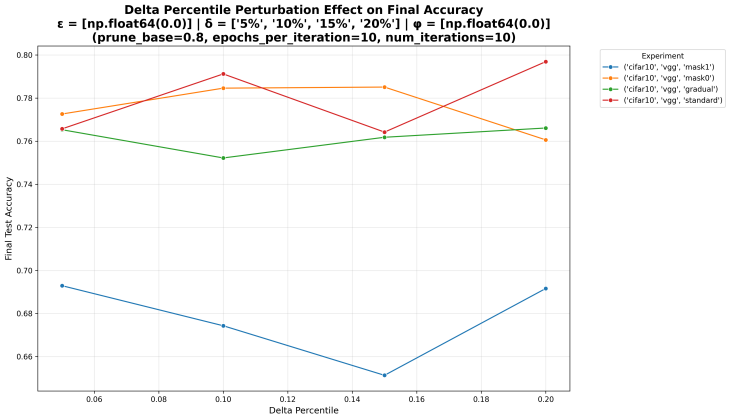

Figure 5: cifar10 vgg16

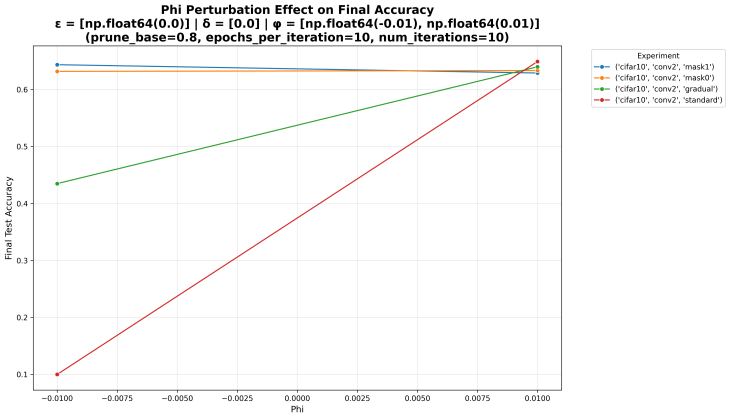

Figure 6: cifar10 conv2

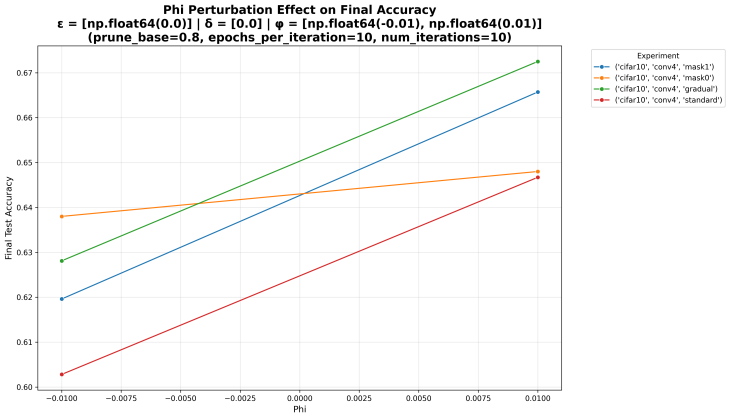

Figure 7: cifar10 conv4

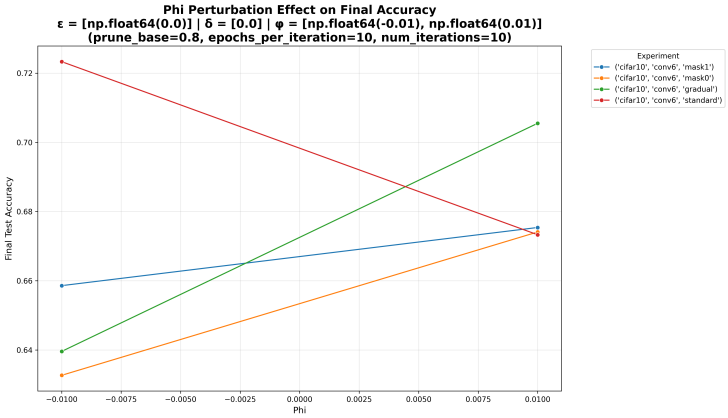

Figure 8: cifar10 conv6

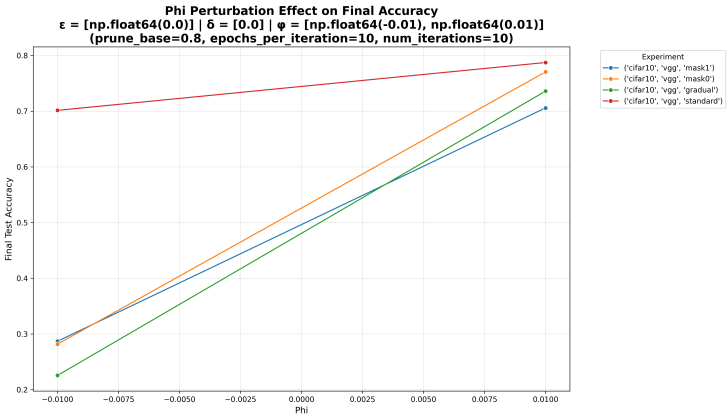

Figure 9: cifar10 vgg16