Extended Abstract Track

Beyond the Zero-Crossing: Finding the Optimal Polarity Threshold for the Sign Hypothesis

Abstract

A central tenet of the Lottery Ticket Hypothesis is that the initial signs of weights are a critical ingredient for success. This "Sign Hypothesis" has always assumed that a weight's polarity is determined by its relationship to zero. We challenge this long-held assumption by introducing a perturbative analysis centered on a "phi-center" (ϕ_{center}) parameter, which establishes a learnable, non-zero threshold for polarity. Our comprehensive experiments show that network performance consistently peaks at a non-zero ϕ_{center}, with its optimal value varying with task complexity. This reveals a "complexity gradient": the polarity definition is irrelevant for simple tasks (MNIST) but becomes paramount for complex ones (CIFAR-10/100). These results offer a more nuanced explanation for the success of magnitude pruning: it effectively removes weights whose initial signs are noisy and misleading because they are evaluated against a suboptimal, zero-centered threshold. To the best of our knowledge, this is the first work to decouple weight polarity from the zero-crossing for lottery ticket pruning. This gives a new direction for understanding initialization and sparsity.

Keywords: Lottery Ticket Hypothesis, Network Pruning, Sign Hypothesis, Sparse Neural Networks, Polarity

1. Introduction

The Lottery Ticket Hypothesis (LTH) (Frankle and Carbin, 2019) has reshaped our understanding of network pruning, demonstrating that sparse subnetworks capable of training to full accuracy exist at initialization. A key insight into this phenomenon is the "Sign Hypothesis," which posits that the signs of the initial weights are more critical than their specific magnitudes for successful subnetwork training (Zhou et al., 2019). This principle has been reinforced by subsequent works showing that effective pruning algorithms excel at identifying and maintaining this crucial sign information (Gadhikar and Burkholz, 2024; Oh et al., 2025).

However, the existing literature implicitly assumes a strict definition of polarity: that the sign is determined by the zero-crossing. None of the previous works have studied whether this exact polarity is the determining factor, or if a small threshold around zero might define a more meaningful boundary. This paper investigates this assumption by asking: Is the boundary for determining a weight's functional sign truly at zero? To answer this, we introduce a novel perturbative strategy based on a "phi-center" (ϕ_{center}) parameter. This parameter establishes a symmetric "dead zone" around zero, allowing us to test whether the network's performance is more sensitive to a weight's sign or its position relative to a small, non-zero threshold.

Our contributions are: (1) We are the first to systematically challenge the strict zero-crossing definition of polarity in the Sign Hypothesis. (2) We introduce the ϕ_{center} framework to probe the robustness of this definition. (3) We discover that the optimal ϕ_{center}

is consistently non-zero across multiple datasets and architectures, suggesting that a dead zone around zero improves performance. (4) We map a "complexity gradient," showing how the importance of polarity strategies and the optimal ϕ_{center} evolve with task difficulty. (5) We provide a refined justification for magnitude pruning: it is effective because it removes low-magnitude weights whose initial signs are noisy and less reliable for guiding optimization.

2. Methodology: Probing Polarity with Perturbations

Our methodology extends the standard Iterative Magnitude Pruning (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 a *reference strategy*, which determines the source of information for the new weights. All strategies decouple sign from magnitude by setting the final magnitude of all surviving weights to a constant value, 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 (wrt ϕ_{center}) are taken from the network's original weights at initialization ($t = 0$). The pruned weights are set to 0. This is the classic LTH-style rewinding of polarity.

- **Mask0 (Trained Weights):** The reference signs (wrt ϕ_{center}) are taken from the weights at the end of the most recent training iteration. The pruned weights are set to 0.

- **Gradual (Mixed):** A stochastic mix where 70% of weights reference the trained signs and 30% reference the initial signs (wrt ϕ_{center}). The pruned weights are set to 0.

- **Standard (Full):** References the trained signs (wrt ϕ_{center}). The pruned weights are set to their trained values at that instant.

2.2. Phi-Center (ϕ_{center}) Perturbation

To test the standard polarity definition, we introduce the ϕ_{center} perturbation. Instead of defining polarity by the sign of a reference weight w, we assign signs based on a threshold: the new sign is $+1$ if $w > \phi_{center}$ and -1 if $w < \phi_{center}$. The polarity is thus determined by the threshold and is relative to ϕ_{center}. This allows us to analyze the network's sensitivity to the signs of weights with very small magnitudes. For each dataset and architecture, the ϕ_{center} values are chosen from the set $\{\pm 5e-05, \pm 5e-04, \pm 1e-04, \pm 1e-03, \pm 1e-01, 0\}$. Evidently 0 is the traditional definition of polarity which exists in the literature.

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, the network's polarity configuration proves to be exceptionally robust, confirming that the sign structure is less critical for simpler tasks. As shown in Table 1, all reinitialization strategies achieve very high and nearly identical performance, with accuracies ranging from 97.5% to 97.8%. The optimizer can easily converge to a high-performing solution, making the precise definition of polarity less critical than on more complex datasets.

Table 1: Best final test accuracy (%) and corresponding optimal ϕ_{center} on MNIST (the interested reader can find the detailed graphs in the appendix B).

Architecture	**gradual**	**mask0**	**mask1**	**standard**
fc_lot	97.75 (0.001)	97.8 (0.0)	97.5 (-5e-05)	97.8 (0.0005)

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

On the more challenging CIFAR-10 dataset, a clear differentiation in strategy performance emerges. Strategies that incorporate learned polarity ('mask0', 'gradual') consistently outperform the classic LTH rewinding ('mask1'). The 'mask0' strategy achieves the highest accuracy, reaching an impressive 84.6% on the VGG architecture.

Performance is highly sensitive to the choice of ϕ_{center}. For all strategies, accuracy peaks at a small, non-zero ϕ_{center} and declines as the threshold moves further from zero. This strongly suggests that a small dead zone is beneficial, while a large one destroys critical polarity information. Table 2 confirms that the optimal ϕ_{center} values are consistently small but non-zero.

Table 2: Best final test accuracy (%) and corresponding optimal ϕ_{center} on CIFAR-10 (detailed graphs are provided in the appendix B).

Architecture	**gradual**	**mask0**	**mask1**	**standard**
Conv2	63.2 (0.0005)	63.4 (-0.0005)	61.6 (0.0005)	62.7 (-0.00005)
Conv4	65.0 (-0.001)	65.2 (0.0001)	62.7 (0.0001)	62.9 (0.001)
Conv6	71.9 (0.0005)	71.2 (0.00005)	67.9 (0.001)	67.4 (0.001)
VGG	81.3 (-0.0001)	**84.6 (0.00005)**	71.0 (-0.001)	80.2 (0.001)

3.3. Polarity as a Critical Performance Differentiator: CIFAR-100

On CIFAR-100, the most complex task, the choice of reinitialization strategy remains critical. The results for the VGG network (Table 3) show a much tighter performance race between strategies, with all achieving around 64% accuracy. The 'gradual' strategy emerges as the top performer, albeit by a narrow margin. This suggests that on very complex tasks, a careful blend of initial and learned polarity is most effective. Crucially, the optimal ϕ_{center} values are again consistently non-zero, reinforcing the finding that the polarity of weights with very small magnitudes is best determined by a learned threshold rather than their sign at initialization.

Table 3: Best final test accuracy (%) and corresponding ϕ_{center} on CIFAR-100 with a VGG architecture (detailed graphs are provided in the appendix B).

Architecture	**Reinitialization Strategy**	**Best Final Accuracy (%) (ϕ_{center})**
VGG	gradual	**64.4** (0.00005)
	mask0	64.2 (0.0005)
	mask1	64.2 (-0.00005)
	standard	64.0 (0.001)

4. Discussion and Conclusion

Our cross-dataset perturbative analysis reveals that the Sign Hypothesis, while foundational, is incomplete. The assumption of a strict zero-crossing for determining polarity does not hold up under scrutiny. Our experiments with the ϕ_{center} parameter consistently show that performance is optimized with a non-zero threshold, creating a "dead zone" where the signs of the lowest-magnitude weights are effectively ignored or corrected. Previous works established the importance of signs, but our work is the first to systematically perturb the *definition* of polarity itself, revealing that the zero-crossing is not a hard boundary but a noisy region.

This finding provides a more refined justification for the success of Iterative Magnitude Pruning. IMP is effective not just because it preserves the critical signs of high-magnitude weights, but because it *removes* low-magnitude weights whose initial signs are ambiguous and less reliable for guiding optimization. Our results demonstrate this directly: by imposing a small threshold via ϕ_{center}, we improve performance by correcting the polarity assignment for these near-zero weights.

As task complexity increases, strategies that incorporate learned polarity ('mask0', 'gradual') prove more robust, indicating that as the loss landscape becomes more complex, relying on initial signs alone is insufficient. As a follow-up, we wish to explore if these polarity perturbation patterns hold when using more structural pruning methods, such as spectral pruning.

References

Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations (ICLR)*, 2019.

Advait Gadhikar and Rebekka Burkholz. Masks, signs, and learning rate rewinding. In *International Conference on Learning Representations (ICLR)*, 2024.

Junghun Oh, Sungyong Baik, and Kyoung Mu Lee. Find a winning sign: Sign is all we need to win the lottery. In *International Conference on Learning Representations (ICLR)*, 2025.

Hattie Zhou, Janice Lan, Rosanne Liu, and Jason Yosinski. Deconstructing lottery tickets: Zeros, signs, and the supermask. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Appendix A. Hyperparameters

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

A.1. MNIST

Table 4: Hyperparameters for MNIST experiments

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

A.2. CIFAR-10

Table 5: Hyperparameters for CIFAR-10 experiments

Parameter	**Value**
Architectures	conv2, conv4, conv6, vgg16
Number of iterations	10
Prune base	0.8
Epochs per iteration	80

A.3. CIFAR-100

Table 6: Hyperparameters for CIFAR-100 experiments

Parameter	Value
Architecture	vgg16_pretrained
Number of iterations	10
Prune base	0.8
Epochs per iteration	50

Appendix B. Experimental Plots

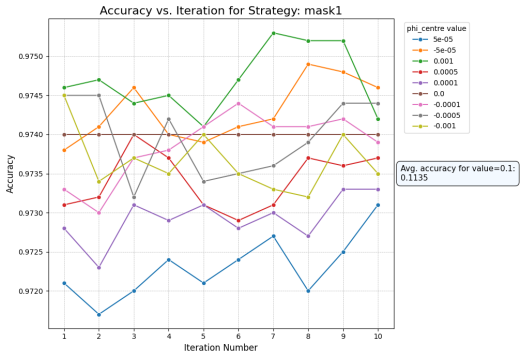

Figure 1: mnist fc_lot mask1

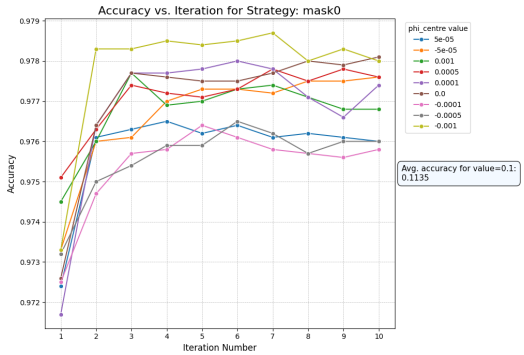

Figure 2: mnist fc_lot mask0

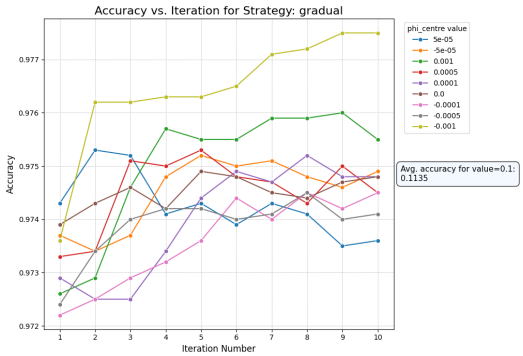

Figure 3: mnist fc_lot gradual

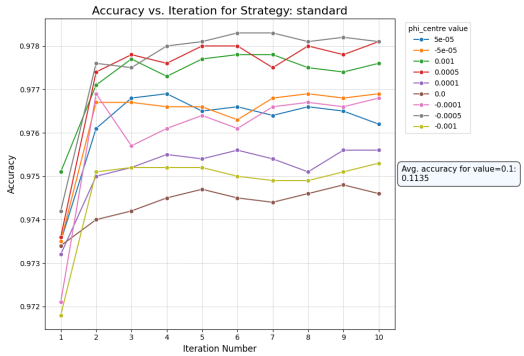

Figure 4: mnist fc_lot standard

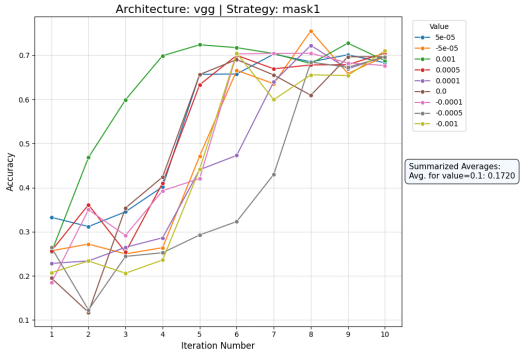

Figure 5: cifar10 vgg mask1

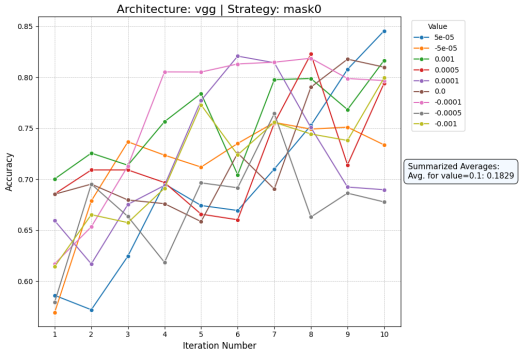

Figure 6: cifar10 vgg mask0

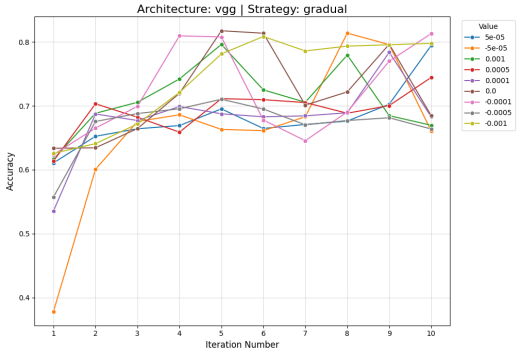

Figure 7: cifar10 vgg gradual

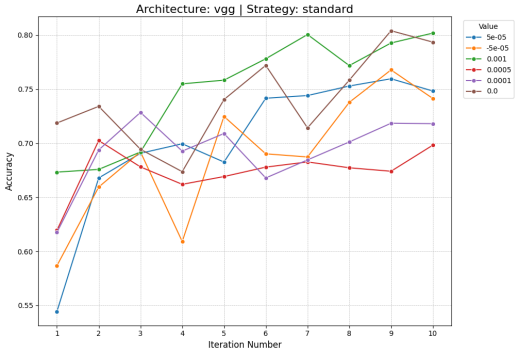

Figure 8: cifar10 vgg standard

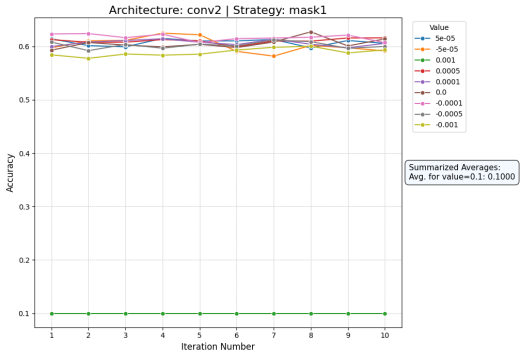

Figure 9: cifar10 conv2 mask1

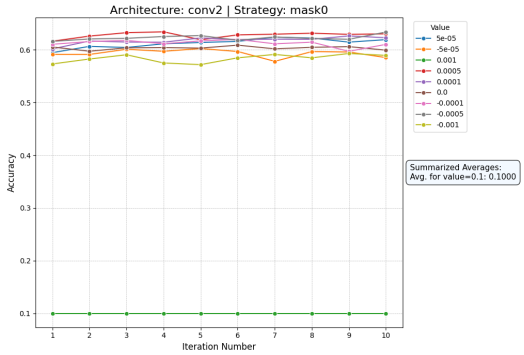

Figure 10: cifar10 conv2 mask0

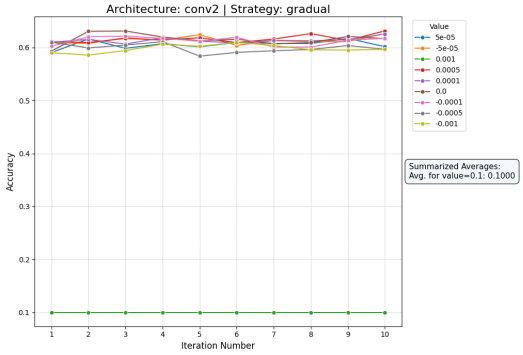

Figure 11: cifar10 conv2 gradual

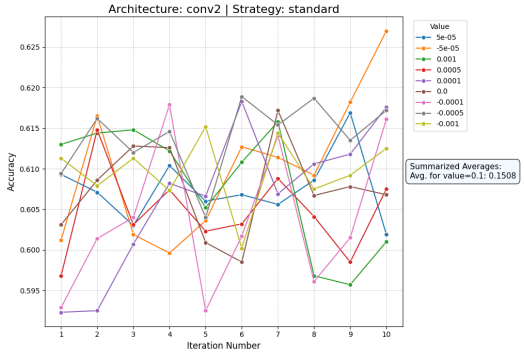

Figure 12: cifar10 conv2 standard

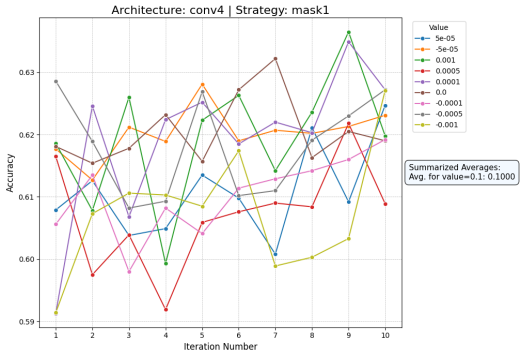

Figure 13: cifar10 conv4 mask1

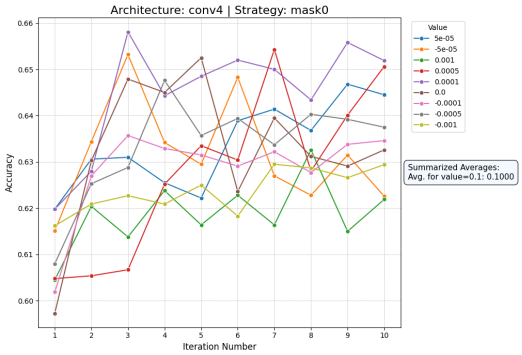

Figure 14: cifar10 conv4 mask0

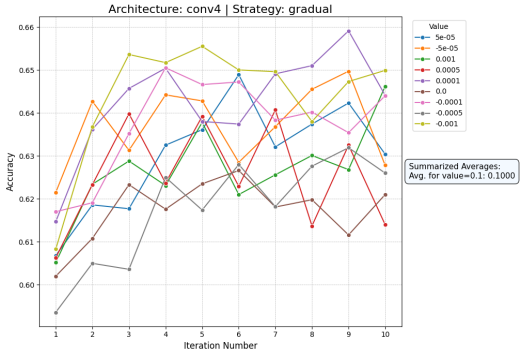

Figure 15: cifar10 conv4 gradual

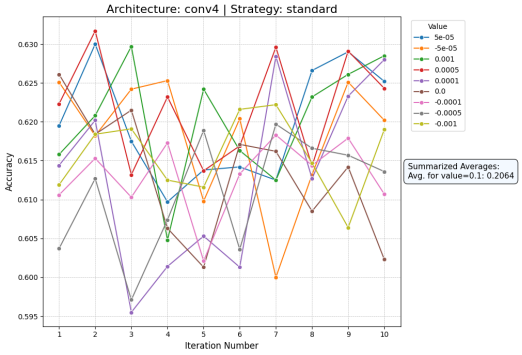

Figure 16: cifar10 conv4 standard

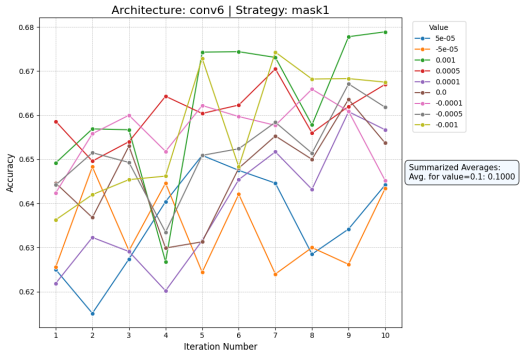

Figure 17: cifar10 conv6 mask1

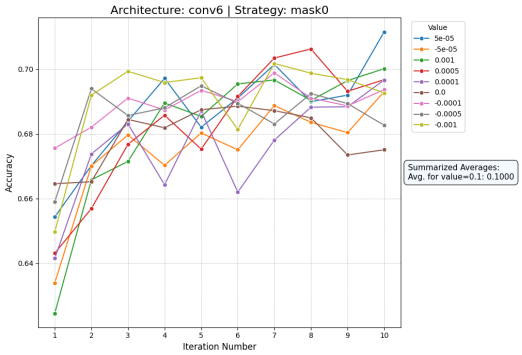

Figure 18: cifar10 conv6 mask0

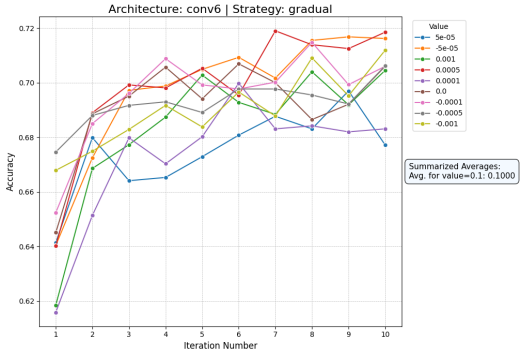

Figure 19: cifar10 conv6 gradual

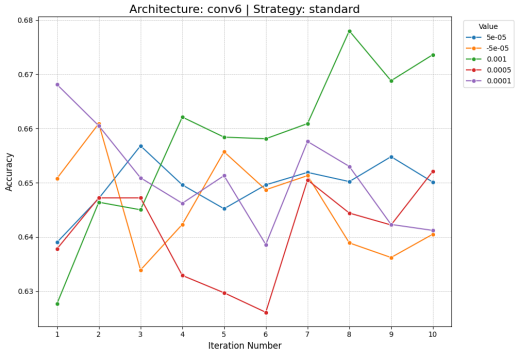

Figure 20: cifar10 conv6 standard

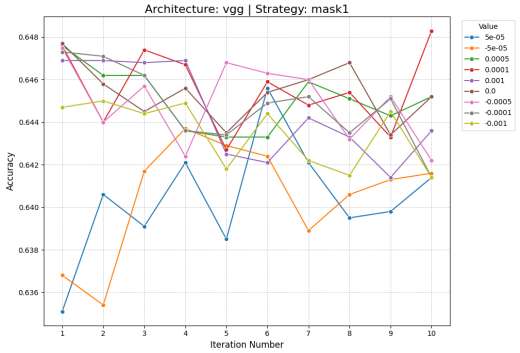

Figure 21: cifar100 vgg mask1

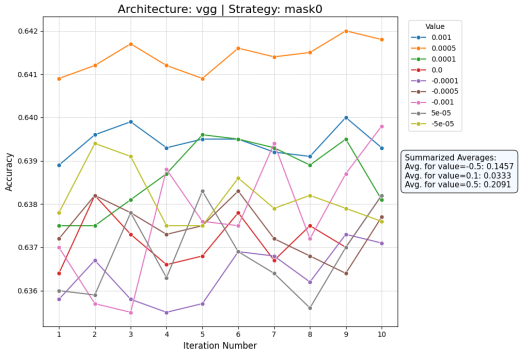

Figure 22: cifar100 vgg mask0

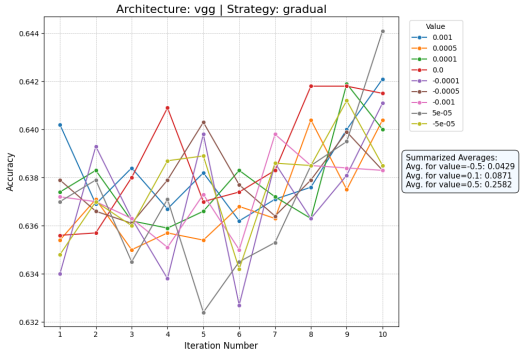

Figure 23: cifar100 vgg gradual

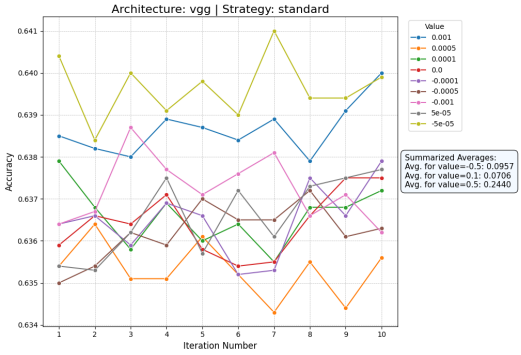

Figure 24: cifar100 vgg standard