Sparse Training from Random Initialization: Aligning Lottery Ticket Masks using Weight Symmetry

UNIVERSITY OF CALGARY · CMI · VECTOR INSTITUTE / INSTITUT VECTEUR · UNIVERSITY OF TORONTO

Mohammed Adnan*,1,3, Rohan Jain*,1, Ekansh Sharma[2,3], Rahul G. Krishnan[2,3], Yani Ioannou[1]

University of Calgary[1] University of Toronto[2] Vector Institute[3]

* Equally contributed 1st authors.

Visualization of Aligning Masks using Weight Symmetry

Figure 1.a)

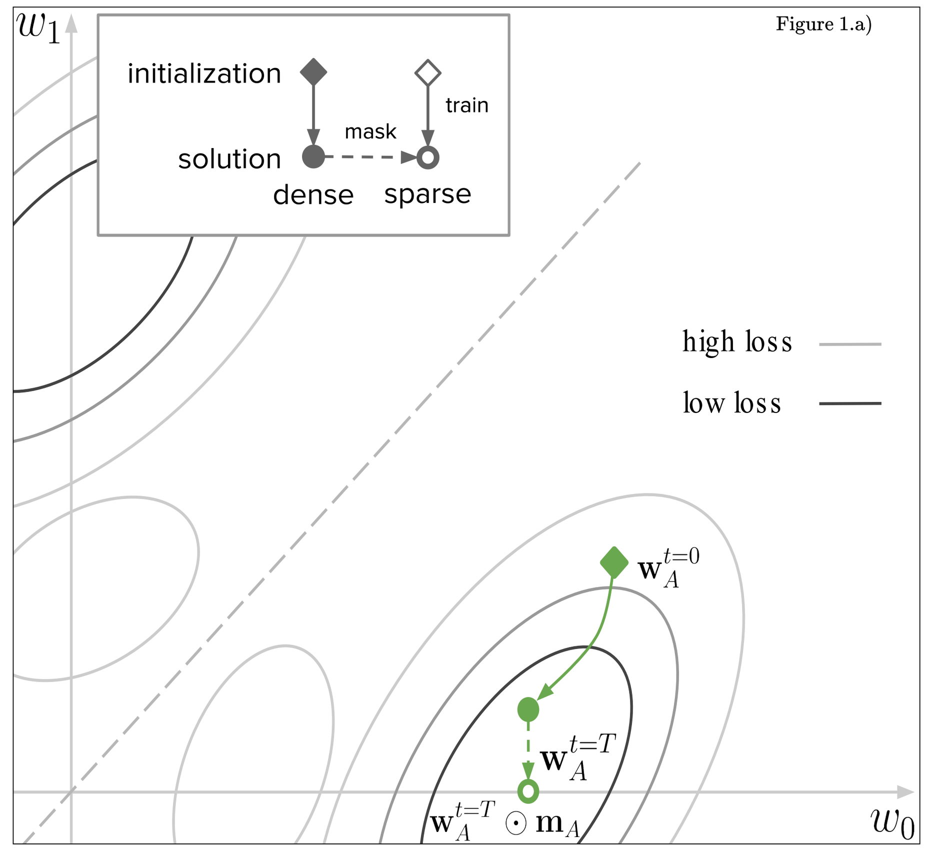
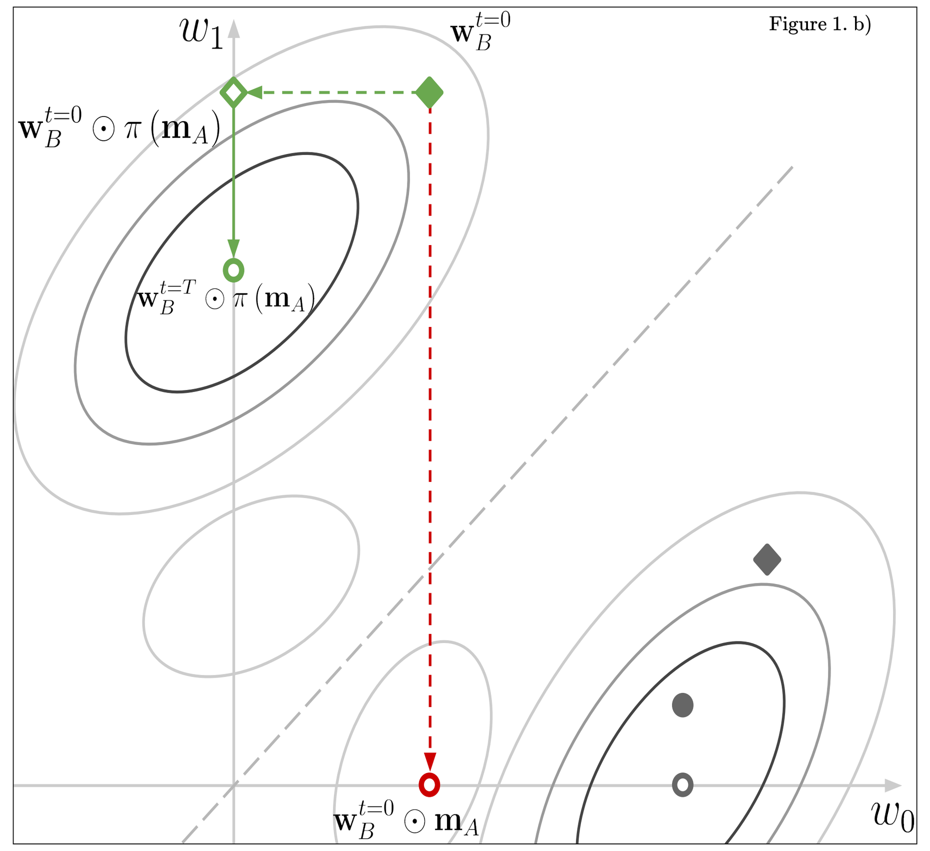


Figure 1. b)

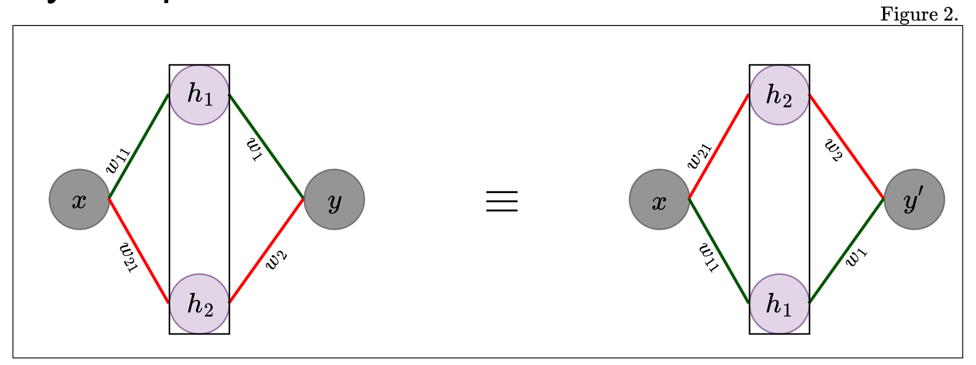


In Figure 1.a), a dense random init., $\mathbf{w}_A^{t=0}$, converges to a dense solution, $\mathbf{w}_A^{t=T}$, which is then pruned with IMP resulting in the mask, $\mathbf{m}_A$.
In Figure 1.b), permuting the mask, $\pi(\mathbf{m}_A)$, to match the (symmetric) basin in which the new initialization, $\mathbf{w}_B^{t=0}$, enables sparse training.

Background

1. **Lottery Ticket Hypothesis (LTH):** identifies sparse sub-networks (i.e. binary masks) that, when trained independently, *can* match dense model performance. [1]
2. NNs are **permutation invariant**: swapping (i.e. permuting) neurons in a layer does not change the underlying function they compute.

Figure 2.



In Figure 2., we show permutation symmetry in a single hidden layer NN, where the outputs, y and y' remain equivalent for the same input.

3. Git Re-Basin claimed that NN loss landscapes nearly contain a **single** solution basin *modulo permutations*. [2]

Motivation

- Motivated by the goal of training a sparse model from a new random initialization.
  - → Frankle et al. demonstrated that training with a highly sparse is mask possible, proposing the LTH [1].

The **key** limitations of LTH: a dense model must be first trained to get a mask, which can only be used with its original random initialization.

- → Lottery tickets do **not** generalize well to new random initializations.
- → LTH consistently converges to very similar solutions to the original pruned model, effectively relearning the same solution [4].

We seek to answer:

- → **How can we train a LTH mask from a different random init. while maintaining good generalization?**
- → **Do models trained with different random initialization but the same LTH mask learn different solutions?**

[1] Frankle et al. The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks, 2019.
[2] Ainsworth et al. Git re-basin: Merging models modulo permutation symmetries, 2023.
[3] Paul et al. Unmasking the lottery ticket hypothesis: What's encoded in a winning ticket's mask?, 2023.
[4] Evci et al. Gradient flow in sparse neural networks and how lottery tickets win, 2022.

NSERC CRSNG · ALBERTA INNOVATES · ICML International Conference On Machine Learning

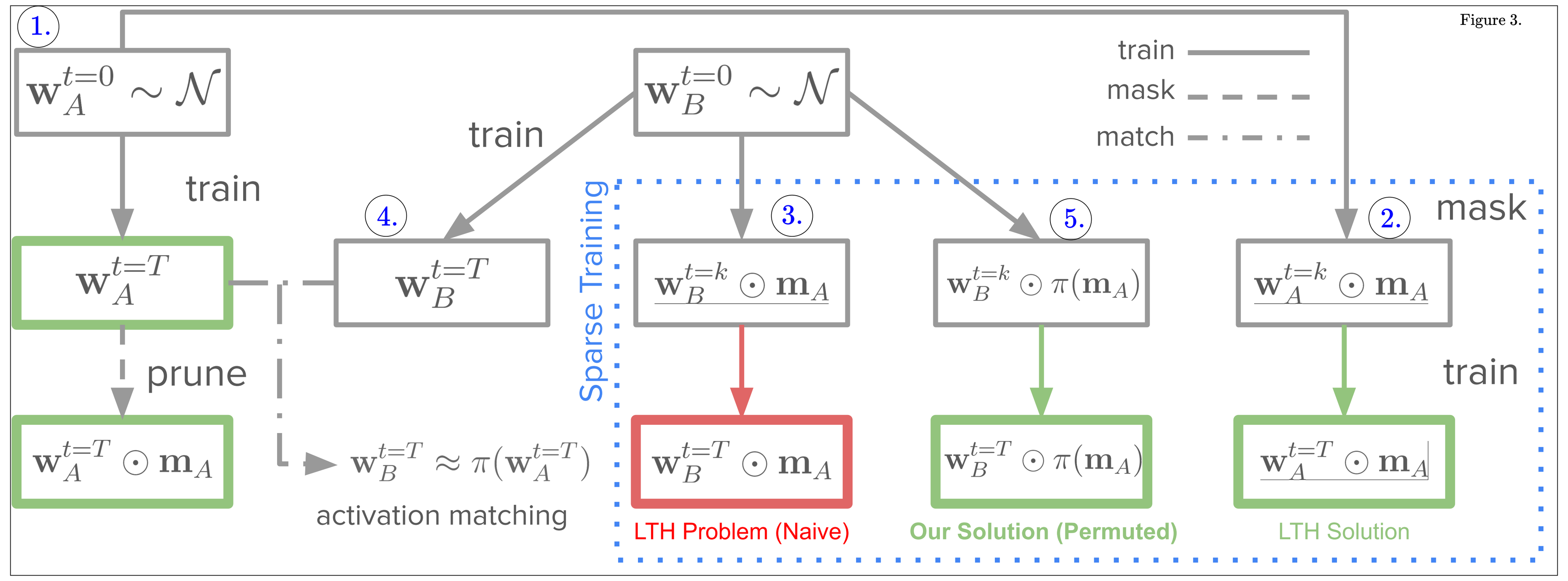
Our Primary Findings

"We found that LTH masks fail to generalize to new random initializations due to ***loss basin misalignment****. To reuse an LTH mask with a different random initialization, we leverage permutation symmetries, to* ***permute the mask to align with the new random initialization optimization basin****."*

- We find that a sparse model (with the permuted mask) with new random initialization can *nearly* match generalization performance of the LTH solution.
- We empirically demonstrate this on CIFAR-10/100 and ImageNet with VGG11 and ResNet models of varying widths.

Methodology

Figure 3.



In Figure 3. we demonstrate the overall framework of our training procedure.

Results: ResNet50/ImageNet + VGG11/CIFAR-10

- **ResNet50**: permuted solution beats the naive solution across all sparsity levels, validating our hypothesis on large datasets.
- **VGG11**: increasing the rewind point, the permuted solution closely matches the accuracy of LTH, while naive solution significantly plateaus.

Figure 4.a)

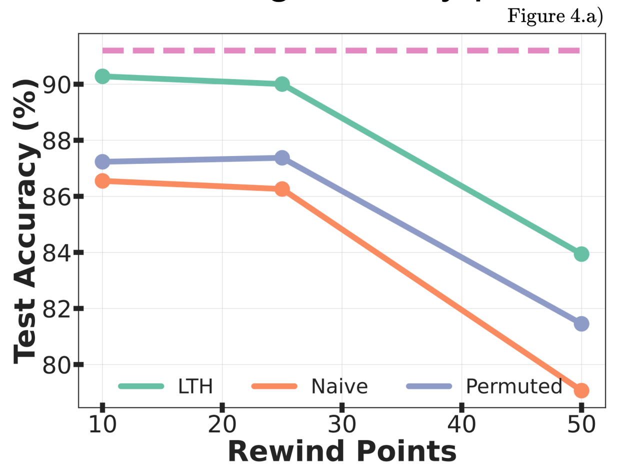
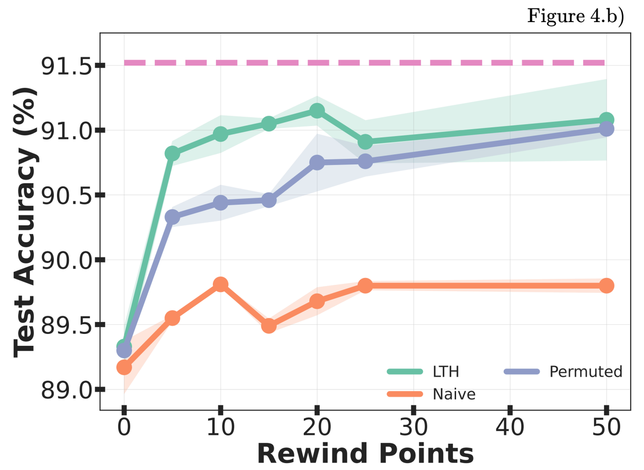


Figure 4.b)



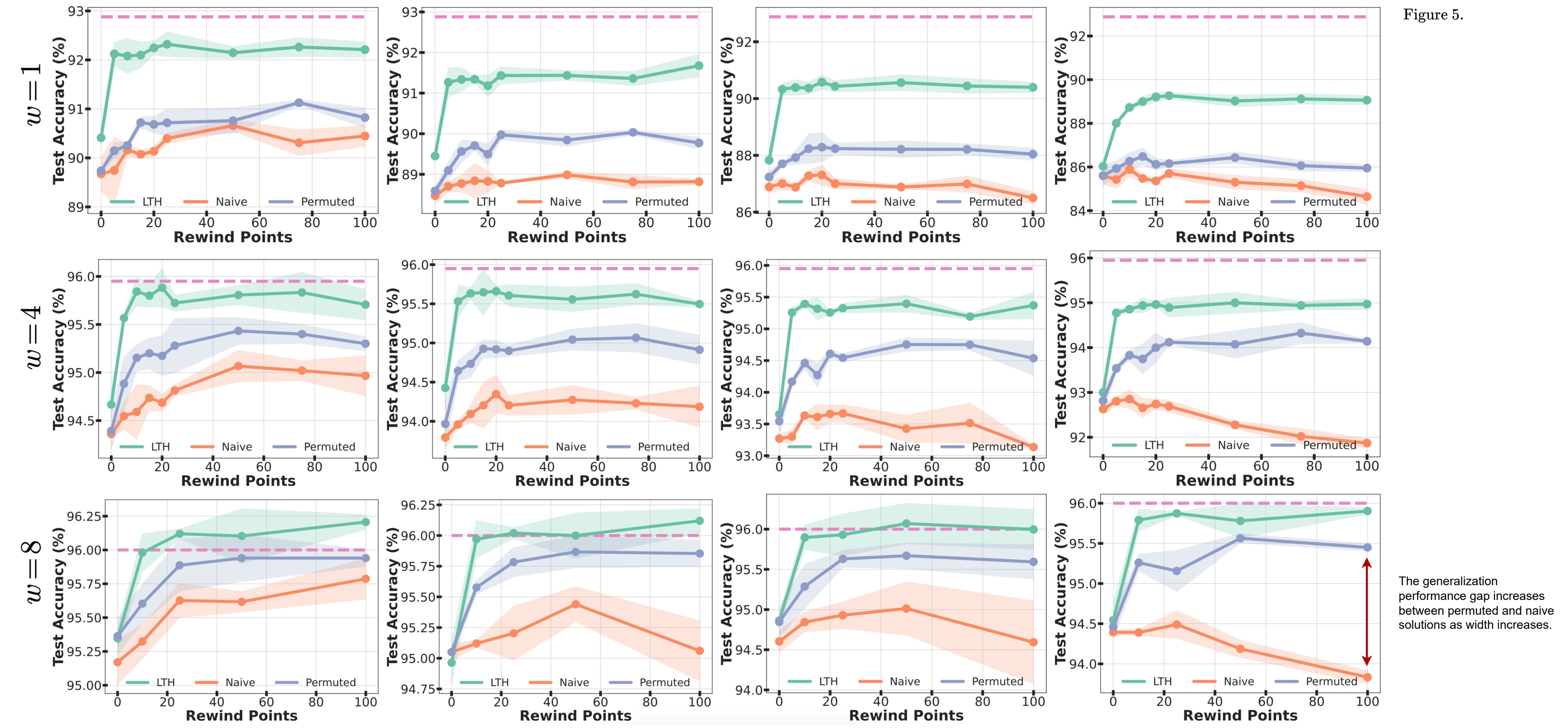
In Figure 4.a) and 4.b) we show the permuted solutions behaviour over ResNet50/ImageNet & VGG11/CIFAR-10 at 90% sparsity, with width = 1, respectively.

Results: ResNet20/CIFAR-10 Observations

- Permuted solution outperforms the naive solution. Sparsity increasing: training becomes harder, widening the gap between permuted and naive solutions.
- Both the LTH & permuted solution do not perform well at a truly random init. (k = 0) but **improves** on increasing the rewind point until plateauing.
- Width increasing: the gap between training from random init. with the permuted mask & the LTH/ dense baseline decreases, unlike training with the naive mask.

Results: ResNet20/CIFAR-10 Plots

Figure 5.



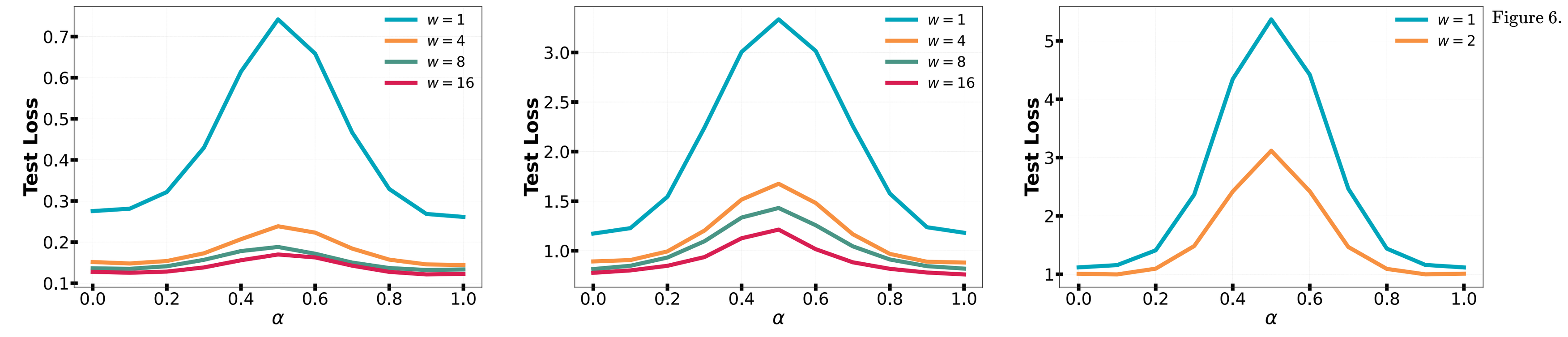
(a) sparsity = 0.80 (b) sparsity = 0.90 (c) sparsity = 0.95 (d) sparsity = 0.97

In Figure 5. Test accuracy of sparse networks solutions vs. increasing rewind points for different sparsity levels and widths, w.

Effect of Model Width

- Larger width exhibits better linear mode connectivity (LMC). As the width of the model increases, the permutation matching algorithm gets more accurate, thereby reducing the loss barrier.

Figure 6.



(a) ResNet20$\times\{w\}$/CIFAR-10 (b) ResNet20$\times\{w\}$/CIFAR-100 (c) ResNet50$\times\{w\}$/ImageNet

Figure 6. Plots showing the linear interpolation between $\pi(\mathbf{w}_A^{t=T})$ and $\mathbf{w}_B^{t=T}$ for various widths.

Ensemble Diversity & Loss Landscape Analysis

- Although the mean test acc. of LTH is higher, ensemble of permuted models achieves better test acc. due to better functional diversity of permuted models.
- We also show, *modulo permutations*, reusing the permuted mask leads to convergence in the same mode as the LTH solution.
- We show for a fixed init., the dense solution and corresponding LTH solution reside within the same loss basin when **variance collapse** is considered. This conclusion presents a new perspective compared to observations made by Paul et al. [3].

Figure 7. a)

Mask	Test Accuracy (%)	Ensemble Acc. (%)	Disagree-ment	KL	JS
ResNet20$\times\{4\}$/CIFAR-100					
none (dense)	78.37 ± 0.059	-	-	-	-
IMP	74.46 ± 0.321	79.27	0.259	1.005	0.372
LTH	$\mathbf{75.35\pm0.204}$	75.99	0.117	0.134	0.038
permuted	72.48 ± 0.356	**77.85**	0.278	0.918	0.327
naive	71.05 ± 0.366	76.15	**0.290**	**0.970**	**0.348**

Figure 7. b)

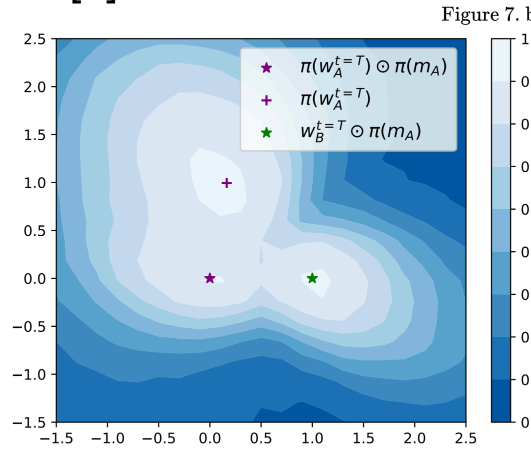
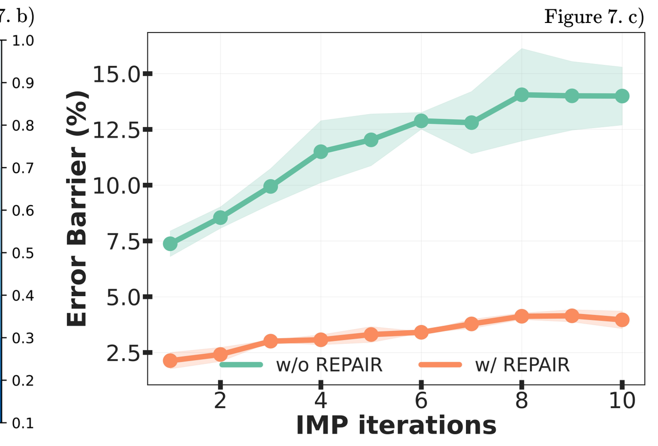


Figure 7. c)



In Figure 7.a): Various measures of function space similarity between the models. In Figure 7. b): 0-1 loss landscape of ResNet20{4}/CIFAR-100.
In Figure 7.c): Error barrier between dense and LTH solutions after accounting for variance collapse.