OMANN : Orthogonal Matrices Alignment for Neural Networks

September 9, 2025

Giovanni Adelfio

Abstract

This project investigates geometry-aware alignments between networks sharing the same architecture, extending Git Re-Basin Framework from permutations to layer-wise orthogonal transformations (Ainsworth et al., 2023). We adopt Procrustes analysis to estimate orthogonal maps that align either weights (PWM) or activations (PAM) (Schonemann, 1966). (Schonemann, 1966). On MLPs, we measure the Residual Alignment Error (RAE) arising from non-commutativity with ReLU and study its effects on merging and cycle consistency. We then explore different post-alignment strategies.

1. Introduction

Building on Git Re-Basin, (Ainsworth et al., 2023),—which aligns independently trained networks via hidden-unit permutations to enable model merging and linear mode connectivity—we generalize the alignment to layer-wise orthogonal transformations that align either weights (PWM) or activations (PAM) across models. We estimate these transforms via the orthogonal Procrustes problem (Schonemann, 1966):

$$\min_{Q\in O(d)} \|XQ-Y\|_F^2, \qquad O(d) := \{\, Q \in \mathbb{R}^{d\times d} \mid Q^\top Q = I \,\}.$$

2. Related Work

Prior work uses permutations to enable model merging and linear mode connectivity, (Ainsworth et al., 2023), but permutations form a discrete subset of the full orthogonal symmetries. Orthogonal Procrustes has recently been applied to model merging to reduce interference by orthogonalizing SVD-based subspaces (Gargiulo et al., 2025). In this work we instead estimate layer-wise orthogonal alignment maps between weight matrices, and then perform LERP, aiming to reduce interference and approximate a shared loss basin. Model composition has also been explored via mode connectivity, weight-space interpolation—both linear (LERP) and spherical (SLERP)—and weight averaging ("model soups") (Wortsman et al., 2022).

Email: Giovanni Adelfio <adelfio.2151753@studenti.uniroma1.it>.

Machine Learning 2025, Sapienza University of Rome, 2nd semester a.y. 2024/2025.

3. Method

Problem setup. We consider two independently trained MLPs, with identical architectures and ReLU nonlinearity, (4 hidden layers, MNIST). For hidden layer $l = 1, \dots, L-1$, we estimate orthogonal, layer-wise alignment maps $Q_l \in O(d_l)$ to align model A to model B. The aligned parameters of A are obtained via the following change-of-basis:

$$\begin{aligned} &Q_0 = I, \quad Q_L = I, \quad \tilde{W}_1 = Q_1^\top W_1^{(A)}, \quad \tilde{b}_1 = Q_1^\top b_1^{(A)}, \\ &\tilde{W}_l = Q_l^\top W_l^{(A)} Q_{l-1}, \quad \tilde{b}_l = Q_l^\top b_l^{(A)} \qquad (1 < l < L), \\ &\tilde{W}_L = W_L^{(A)} Q_{L-1}, \quad \tilde{b}_L = b_L^{(A)}. \end{aligned}$$

PWM — Procrustes Weight Matching. For each hidden layer l, we align the units by solving the following optimization problem:

$$Q_l^\star = \arg\min_{Q\in \mathrm{O}(d_l)} \left\| Q^\top W_l^{(A)} - W_l^{(B)} \right\|_F^2$$

Which can be easily reconducted to the Procrustes problem in Equation 1, by setting $X = W_l^{(A)\top}$ and $Y = W_l^{(B)\top}$. Therefore the optimal solution is given by $Q_l^\star = UV^\top$, where $U\Sigma V^\top = \mathrm{SVD}(X^\top Y) = \mathrm{SVD}(W_l^{(A)} W_l^{(B)\top})$.

It is easy to see that, if the ReLU nonlinearity were absent, the aligned model $\tilde{A}$ would be functionally identical to A, since $Q_i Q_i^\top = I, \forall i$.

PAM — Procrustes Activation Matching. Given a minibatch $\mathcal{B}$, (5000 samples from MNIST train set), we collect either pre-ReLU or post-ReLU activations of layer l for both models:

$$H_l^{(A)} = \phi(Z_l^{(A)}(\mathcal{B})), \quad \textit{and} \quad H_l^{(B)} = \phi(Z_l^{(B)}(\mathcal{B})),$$

where ϕ is either the identity or ReLU function. We then solve the Procrustes problem, as earlier, by substituting

$X = H_l^{(A)\top}$ and $Y = H_l^{(B)\top}$ in Equation 1, and have a closed form solution for $Q_l^\star$.

RAE — Residual Alignment Error. We quantify residual misalignment as the self-induced performance drop when mapping model A into B's basis:

$$\Delta\text{Loss}_{\text{self}}(A \to B) \quad and \quad \Delta\text{Acc}_{\text{self}}(A \to B)$$

(on test set). This proxy reflects the distortion introduced by orthogonal alignment under ReLU non-commutativity.

Cycle consistency. To evaluate cycle consistency we first align model A to B, then align the resulting model back to A. To asses the discrepancy we compare this model with the original A using three metrics: (i) test-set loss, (ii) test-set accuracy, and (iii) the Frobenius Relative Norm Error, (FRNE), between the original weights and those of the twice-aligned model.

Post-alignment strategies. We explore two strategies to mitigate RAE: using PAM as a data-aware refinement, and fine-tuning the aligned model on the training set to recover performance. The latter is esecuted for 2 epochs with Adam optimizer, and learning rate 10^{-4}.

4. Experimental Results

Table 1. RAE — Residual Alignment Error.

#Models	Original	PWM	PAM	PAM, ReLU	PWM+ PAM
Accuracy	97.5%	95%	95%	**86%**	94%
Loss	0.075	0.190	0.180	**0.5711**	0.1559

We observe how PAM on post-ReLU activations incurs the highest RAE, and PWM + PAM does not manage to recover the misalignment incurred by the method.

Cycle consistency. We find that both PWM and PAM exhibit very low cycle inconsistency, for what regards loss(i) and accuracy(ii), while PAM shows a significantly higher weight discrepancy(iii), (FRNE $\sim 10^{-1}$), compared to PWM, ($\sim 10^{-4}$). This is expected since PWM directly aligns the weights as in Equation 3, while PAM focuses on aligning the activations.

5. Discussion and Conclusions

Discussion. Both PAM and PWM methods outperform naive (LERP) and SLERP interpolation, but are hindered by the RAE and underperform permutation weight matching (Ainsworth et al., 2023). However, light fine-tuning of the aligned model recovers performance, enabling seamless, zero-barrier interpolation that outperforms all baselines.

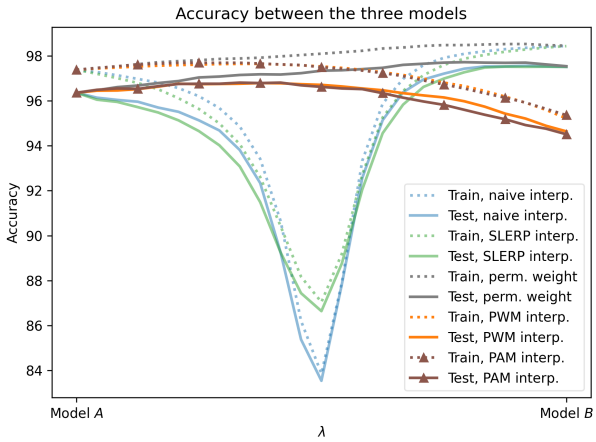

Figure 1. PAM vs PWM vs SLERP vs LERP vs Permutation Weight Matching (Ainsworth et al., 2023).

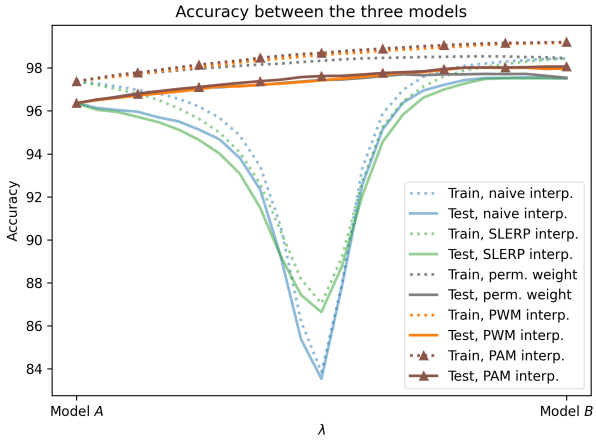

Figure 2. Fine-tuned PWM and PAM vs SLERP vs LERP vs Permutation Weight Matching (Ainsworth et al., 2023)..

Limitations and future work. Our evaluation is limited to MNIST MLPs, broader validation on CNNs/Transformers, and more complex tasks is needed. PAM requires data, and has a relatively high computational cost compared to PWM. A possible future implementation could focus on narrow models, where permutation-based methods are less effective, (Ainsworth et al., 2023).

Conclusions. Relaxing permutation constraints to orthogonal layer-wise alignments is feasible and effective when coupled with light fine-tuning that compensates ReLU-induced residual error. Within MLPs on MNIST, this strategy attains and sometimes surpasses permutation matching, with good cycle consistency and smooth interpolation in weight space.

Source Code. Better data and results visualization in: https://shorturl.at/7X2rl.
Code is available at: https://shorturl.at/RBK6F.

References

Ainsworth, S. K., Hayase, J., and Srinivasa, S. Git re-basin: Merging models modulo permutation symmetries. pp. 1–9, 19,, 2023. URL https://arxiv.org/abs/2209.04836.

Gargiulo, A. A., Crisostomi, D., Bucarelli, M. S., Scardapane, S., Silvestri, F., and Rodolà, E. Task singular vectors: Reducing task interference in model merging. pp. 1–4, 2025. URL https://arxiv.org/abs/2303.17252.

Schonemann, P. H. A generalized solution to the procrustes problem. *Psychometrica*, 31(1):1–10, 1966. URL https://web.stanford.edu/class/cs273/refs/procrustes.pdf.

Wortsman, M., Ilharco, G., Gadre, S. Y., Roelofs, R., Gontijo-Lopes, R., Morcos, A. S., Namkoong, H., Farhadi, A., Carmon, Y., Kornblith, S., and Schmidt, L. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S. (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 23965–23998. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/wortsman22a.html.