Geometric Model Alignment

Sfruttare le Rotazioni per il Merging di Reti Neurali

January 16, 2026

Irene Bianchi

1 Introduzione

Il progetto affronta il problema dell'allineamento e della fusione di reti neurali superando il vincolo delle sole simmetrie di permutazione. Estendendo l'allineamento a trasformazioni ortogonali generali, si ottengono strategie di merging più flessibili, con un errore introdotto dalle non-linearità che rimane contenuto.

Repository GitHub del progetto

2 Lavori associati

Studi precedenti hanno mostrato che reti con la stessa architettura, se riallineate correttamente tramite permutazioni dei neuroni, possono essere connesse da percorsi lineari a bassa perdita; le barriere di loss sono quindi spesso un effetto del disallineamento delle simmetrie interne.

3 Metodologia

3.1 Orthogonal Re-Basin

Per superare i limiti delle permutazioni discrete, sono state implementate e confrontate tre diverse strategie per il calcolo della matrice ortogonale $Q \in \mathbb{R}^{d \times d}$ (tale che $Q^\top Q = I$) necessaria per l'allineamento delle rappresentazioni interne dei modelli.

3.2 Analisi di Procrustes (SVD)

Questa tecnica allinea le attivazioni dei due modelli minimizzando la distanza di Frobenius sotto vincolo di ortogonalità. La soluzione ottimale si ottiene in forma chiusa tramite SVD della matrice di covarianza incrociata, garantendo un allineamento stabile ed efficace delle rappresentazioni.

In assenza di una soluzione chiusa, l'allineamento è stato affrontato tramite ottimizzazione iterativa sulla varietà di Stiefel mediante discesa del gradiente con proiezione. Sebbene più flessibile, questo approccio si è dimostrato meno efficace della soluzione analitica via SVD, mostrando convergenza più lenta e maggiore sensibilità agli iperparametri.

4 Metriche di Analisi

L'efficacia dell'allineamento non è valutata solo in termini di performance finale, ma attraverso metriche geometriche specifiche definite per quantificare la distorsione introdotta dalle rotazioni nello spazio delle feature.

4.1 Errore Residuo di ReLU (Non-Commutatività)

Poiché le funzioni di attivazione non lineari come la ReLU non commutano con trasformazioni ortogonali generiche (a differenza delle permutazioni), è stato quantificato l'errore residuo:

$$E_{\text{res}} = \|\text{ReLU}(A)Q - \text{ReLU}(AQ)\|_F \tag{1}$$

Un valore basso di E_{res} indica che la rotazione Q preserva la struttura delle attivazioni in modo compatibile con la non-linearità della rete, minimizzando la distorsione funzionale.

4.2 Cycle-Consistency

Per verificare la coerenza geometrica della trasformazione, si è misurata la capacità di Q di essere invertita senza alterare i pesi originali. L'errore di consistenza ciclica è definito come:

$$E_{\text{cycle}} = \|(W_1 Q)Q^\top - W_1\|_F \tag{2}$$

Nel caso ideale di ortogonalità perfetta, tale valore dovrebbe essere prossimo allo zero ($Q^\top = Q^{-1}$), garantendo l'isometria della trasformazione nello spazio dei parametri.

5 Risultati Sperimentali

La validazione è stata condotta in due fasi: prima su scenari sintetici per verificare le proprietà geometriche, poi su classificatori reali (MNIST).

Email: Irene Bianchi <bianchi.2129569@studenti.uniroma1.it>

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

5.1 Validazione Sintetica: Permutazioni vs Rotazioni

Per isolare i limiti delle permutazioni, sono stati generati due modelli identici separati da una trasformazione nota. Come visibile in Figura 1, nel caso di semplici scambi di neuroni (Scenario A), entrambi i metodi convergono. Tuttavia, in presenza di rotazioni dense dello spazio latente (Scenario B), l'approccio basato su permutazioni fallisce (curva rossa), non potendo approssimare trasformazioni continue. Il metodo ortogonale (Procrustes), invece, identifica correttamente la rotazione inversa, allineando perfettamente le rappresentazioni (curva blu).

5.2 Case Study su MNIST

Il metodo è stato applicato a due MLP addestrati indipendentemente su MNIST. L'interpolazione ingenua tra i pesi porta a un crollo delle prestazioni (accuratezza $\approx 10\%$), confermando l'esistenza di barriere di loss elevate.

L'allineamento ortogonale ripristina con successo la *Linear Mode Connectivity*. Come mostrato in Table 1, sebbene le permutazioni garantiscano un errore residuo nullo ($E_{\text{res}} = 0$) grazie alla commutatività con la ReLU, spesso non riescono a uscire dai minimi locali. L'approccio ortogonale, pur introducendo un errore residuo trascurabile (10^{-2}), esplora uno spazio di soluzioni più ampio (Varietà di Stiefel), trovando percorsi di connessione a bassa perdita che le sole permutazioni non possono raggiungere.

6 Conclusioni

Gli esperimenti su scenari sintetici e su classificatori MLP addestrati su MNIST confermano l'efficacia dell'allineamento ortogonale. Come mostrato in Figure 1, le sole permutazioni non riescono a riallineare correttamente i modelli in presenza di rotazioni dense, mentre l'approccio ortogonale elimina completamente la loss barrier. Inoltre, come riassunto in Table 1, l'allineamento tramite SVD riduce significativamente la perdita residua rispetto alle permutazioni, con un impatto limitato dovuto alla non-commutatività con ReLU. L'ottimizzazione sulla varietà di Stiefel, pur più flessibile, non offre vantaggi pratici, risultando più lenta e sensibile agli iperparametri. Nel complesso, le trasformazioni ortogonali estendono la Linear Mode Connectivity mantenendo un effetto funzionale contenuto.

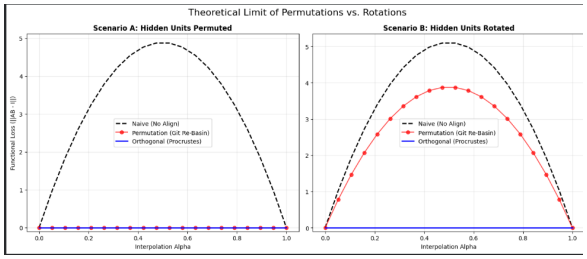

Figure 1: Confronto teorico. A sx: permutazioni e rotazioni equivalenti. A dx: permutazioni falliscono (rosso), ortogonale successo (blu).

Metodo	**Acc.** $\alpha = 0.5$	$\mathbf{E}_{\text{res}}$	**Sign Change**
Naive	11.2%	–	–
Permutation	97.8%	0	0%
Orthogonal	98.1%	$1.2 \cdot 10^{-2}$	8.4%
Orthogonal + SLERP	**98.3%**	$1.2 \cdot 10^{-2}$	8.4%

Table 1: Confronto quantitativo sull'interpolazione.

Bibliografia

(Ainsworth et al., 2023) (Tagare, 2011) (Kornblith et al., 2019) (Ito et al., 2025)

Referimenti

Ainsworth, S., Hayase, J., and Srinivasa, S.
Git re-basin: Merging models modulo permutation symmetries, 2023.
https://arxiv.org/abs/2209.04836

Ito, A., Yabuta, M., and Kawamura, A.
Analysis of linear mode connectivity via permutation-based weight matching: with insights into other permutation search methods, 2025.
https://arxiv.org/pdf/2402.04051

Kornblith, S., Norouzi, M., Lee, H., and Hinton, G.
Similarity of neural network representations revisited, 2019.
https://arxiv.org/abs/1905.00414

Tagare, H. D.
Notes on optimization on stiefel manifolds, 2011.
https://cseweb.ucsd.edu/classes/sp24/cse291-e/papers/StiefelManifold/StiefelNotes.pdf