Appendice Rebasin ortogonale

September 9, 2025

Valeria d'Orsi

1. Introduzione

In questa appendice si riportano alcune tecniche di allineamento, basate su rotazioni e riflessioni,che non si sono rivelate efficaci. Si riportano inoltre degli esperimenti falliti.

2. Failed methods

2.1. Autovettori dei valori singolari delle matrici dei pesi

Nell'articolo *Analisys of linear mode connectivity via permutation-base weight matching* si analizza il Weight Matching (WM) dal punto di vista della funzione computazionale svolta da ciascun layer del modello. In particolare, si osserva che il WM soddisfa la proprietà di Linear Mode Connectivity (LMC) mediante **l'allineamento delle direzioni dei vettori singolari associati ai valori singolari dominanti** nelle matrici dei pesi di ciascun layer. Questo allineamento consente di approssimare efficacemente la funzionalità tra due modelli anche in presenza di una riduzione non significativa della distanza L2 e rende i vettori singolari maggiori coerenti tra i modelli originali e quello ottenuto tramite fusione. Dunque,dato un layer l, con le matrici $\mathbf{W}_l^A$ e $\mathbf{W}_l^B$, corrispondenti ai pesi del layer l nei due modelli da confrontare ,si è ipotizzato che una trasformazione ortogonale appropriata potesse essere ottenuta risolvendo un problema di Procruste tra le matrici dei vettori singolari sinistri, $\mathbf{U}_l^A$ e $\mathbf{U}_l^B$, ottenute tramite decomposizione ai valori singolari (SVD):

$$\mathbf{W}_l^A = \mathbf{U}_l^A \mathbf{\Sigma}_l^A \mathbf{V}_l^{A\top}, \qquad \mathbf{W}_l^B = \mathbf{U}_l^B \mathbf{\Sigma}_l^B \mathbf{V}_l^{B\top}.$$

In tal modo, si è considerato il seguente problema di ottimizzazione per determinare la matrice di rotazione $\mathbf{R}_l \in \mathbb{R}^{d\times d}$:

$$\min_{\mathbf{R}_l \in \mathbb{R}^{d\times d}} \left\| \mathbf{U}_l^A - \mathbf{R}_l \mathbf{U}_l^B \right\|_F \quad \text{t.c.} \quad \mathbf{R}_l^\top \mathbf{R}_l = \mathbf{I}.$$

Email: Valeria d'Orsi <dorsi.2136945@studenti.uniroma1.it>.

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

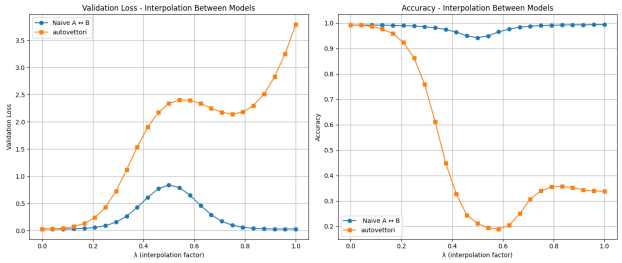

Figure 1. Il grafico mette in evidenza l'inefficacia dell'allineamento degli autovettori sinistri

Tuttavia, si è osservato che, nonostante una risoluzione appropriata del problema, gli autovettori sinistri U'_B, ottenuti dalla successiva decomposizione ai valori singolari del peso ruotato, possono discostarsi da RU_B, compromettendo così l'allineamento funzionale desiderato. (Figure 1)

2.2. Pesi e covarianza

Si è analizzato ulteriormente il problema di Procruste applicato alle matrici dei pesi dei modelli,con la volontà di renderle preventivamente *whitenizzate* per ridurre le differenze dovute a scala e correlazioni interne. Tuttavia ciò ha portato alla realizzazione di un modello ruotato non adeguatamente funzionale.(Figure 2)

Data $W \in \mathbb{R}^{d_{\text{out}}\times d_{\text{in}}}$ matrice dei pesi di un layer; la nuova matrice presa in analisi è $\tilde{W} = C_W^{-\frac{1}{2}} W$, dove la radice quadrata inversa di C_W viene calcolata tramite decomposizione agli autovalori $C_W^{-\frac{1}{2}} = U\Lambda^{-\frac{1}{2}}U^T$. Dunque, viene risolto il problema di Procruste:

$$R^\star = \arg\min_{R\in O(d_{\text{out}})} \|\tilde{W}_1 - R\tilde{W}_2\|_F^2,$$

L'interpolazione ottenuta con questo metodo è risultata globalmente meno efficace di quella ottenuta mediante l'uso della SVD nella risoluzione del problema di Procuste applicato alle matrici dei pesi W.

2.3. Decomposizione Polare

Strettamente affine alla decomposizione a valori singolari è la decomposizione polare.

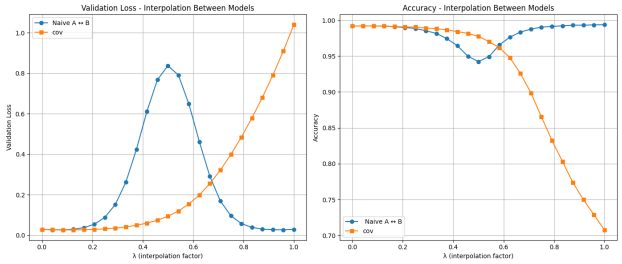

Figure 2. Il grafico evidenzia una ridotta efficacia del modello permutato, che si traduce in un aumento della loss barrier e una riduzione dell'accuracy (metodo pesi e covarianza)

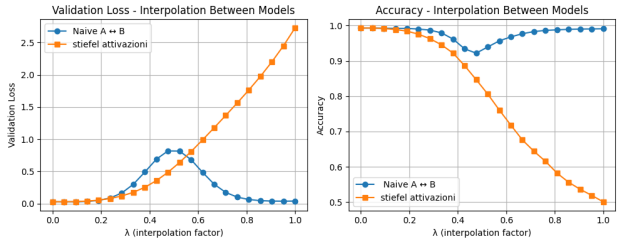

Figure 3. Il grafico validation loss in funzione di lambda con ottimizzazione Stiefel applicata sulle attivazioni

Def: La decomposizione polare di una matrice quadrata A è una fattorizzazione della forma:$A = UP$ dove U è una matrice unitaria che rappresenta una rotazione e P è una matrice hermitiana semidefinita positiva che dilata lo spazio lungo un insieme di assi ortonormali.

Relativamente alla SVD è noto che, data $A = W\Sigma V^*$, con V^* trasposta coniugata di V,

$$P = V\Sigma V^* \qquad U = WV^*$$

Ricordando che la soluzione del problema di Procruste $R^* = \min_{\mathbf{R}\in\mathbb{R}^{d\times d}} \|\mathbf{X} - \mathbf{RY}\|_F \quad \text{t.c.} \quad \mathbf{R}^\top\mathbf{R} = \mathbf{I}$ si ottiene calcolando

$\mathbf{XY}^\top = \mathbf{U}'\mathbf{\Sigma}\mathbf{V}'^\top$ e moltiplicando$\mathbf{U}'\mathbf{V}'^\top$, si comprende che anche U è soluzione del problema.

Dunque,si propone un algoritmo che,data la matrice di correlazione $M = \mathbf{XY}^\top$ calcola prima la matrice simmetrica positiva definita M^TM,ne determina poi la radice quadrata positiva $(M^TM)^{\frac{1}{2}}$ e infine ne calcola l'inversa $(M^TM)^{-\frac{1}{2}}$,ottenendo

$$U = M(M^TM)^{-\frac{1}{2}}.$$

Tuttavia sorgono diverse problematiche: anche se M è reale $\mathbf{M}^\top M$ potrebbe avere valori numerici che, a causa di errori di arrotondamento, fanno sì che la radice quadrata abbia componenti immaginarie piccole ma non nulle. Dunque, pur decidendo di ignorare la parte immaginaria, la validation loss raggiunge durante le interpolazioni valori elevatissimi.

2.4. Ottimizzazioni di Stiefel su attivazioni

Con l'obiettivo di minimizzare la distanza tra le attivazioni di due modelli neurali per favorirne l'allineamento, si sono applicate le tecniche di ottimizzazione sul Stiefel manifold attraverso Cayley transformation sulle attivazioni. Tuttavia l'interpolazione ricavata è risultata molto meno efficiente di quella ottenuta con lo stesso metodo, ma applicato ai pesi.

2.5. Matrici di riflessione

Sebbene la maggior parte degli approcci privilegi le rotazioni è interessante esplorare l'allineamento esclusivamente mediante matrici di riflessione. Date $Z_A, Z_B \in \mathbb{R}^{n\times m}$, si vuole trovare una matrice di riflessione $R = I - 2\mathbf{v}\mathbf{v}^\top, \quad \|\mathbf{v}\| = 1$ che minimizzi

$$\min_{\|\mathbf{v}\|=1} \|Z_A - RZ_B\|_F^2.$$

Si propongono due modi per individuare tale R.

2.5.1. R-AUTOVETTORI

Espandendo la norma della distanza e non considerando i termini costanti rispetto a $\mathbf{v}$ si ottiene

$$\min_{\|\mathbf{v}\|=1} \|Z_A - RZ_B\|_F^2 \quad \Longleftrightarrow \quad \max_{\|\mathbf{v}\|=1} \operatorname{trace}(Z_A^\top RZ_B).$$

Sostituendo $R = I - 2\mathbf{v}\mathbf{v}^\top$ e considerando i termini dipendenti da v si ha

$$\max_{\|\mathbf{v}\|=1} \operatorname{trace}(Z_A^\top RZ_B) \quad \Longleftrightarrow \quad \min_{\|\mathbf{v}\|=1} \mathbf{v}^\top M\mathbf{v},$$

dove$M = \frac{1}{2}\left(Z_BZ_A^\top + Z_AZ_B^\top\right)$

La soluzione $\mathbf{v}$ è dunque l'autovettore di M corrispondente all'autovalore minimo.

2.5.2. OTTIMIZZAZIONE NON LINEARE VINCOLATA SU UNA VARIETÀ SFERICA

L'ottimizzazione viene eseguita tramite discesa del gradiente con l'algoritmo Adam, imponendo a ogni iterazione la normalizzazione del vettore $\mathbf{v} \in \mathbb{R}^n$ affinché $\|\mathbf{v}\| = 1$. La matrice di riflessione $\mathbf{R} = \mathbf{I} - 2\mathbf{v}\mathbf{v}^\top$ viene aggiornata iterativamente per minimizzare la funzione obiettivo:

$$\mathcal{L}(\mathbf{v}) = \|\mathbf{Z}_A - \mathbf{RZ}_B\|_F^2.$$

Al termine dell'ottimizzazione, si ottiene la riflessione che meglio allinea $\mathbf{Z}_B$ a $\mathbf{Z}_A$ in termini di distanza di Frobenius.

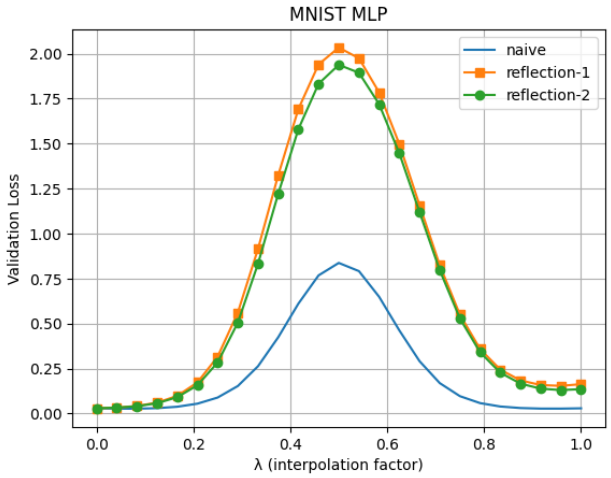

Figure 4. Il grafico rappresenta l'interpolazione lineare tra il modello A e i due modelli Briflesso,che differiscono nella tecnica con cui è stata individuata R. In **Reflection 1**, si è adoperato il metodo descritto nel Paragrafo 2.2.1, in **Reflection 2**, si è seguita la procedura del Paragrafo 2.2.2.

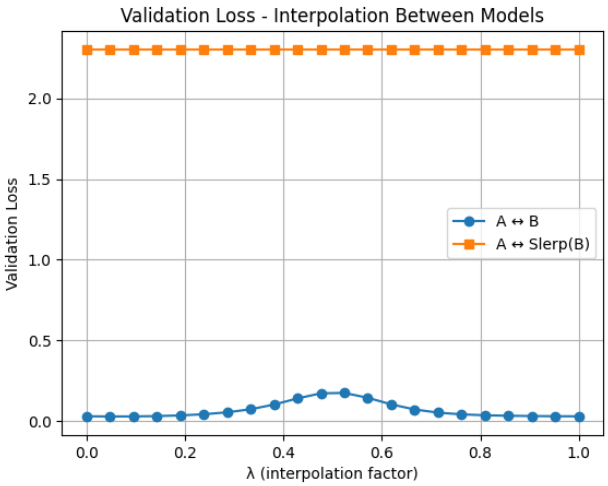

Figure 5. Il grafico rappresenta l'interpolazione sferica tra il modello A e $B_{allineato}$ (tramite SVD) in assenza di interpolazione delle norme

Tuttavia, l'interpolazione lineare tra il modello A e il modello B riflesso risulta meno efficace rispetto a quella effettuata direttamente tra A e B, nonostante una evidente similarità funzionale tra il modello B e la sua versione riflessa. (Figure 4). L'interpolazione sferica risulta invece assai peggiore, raggiungendo una loss barrier pari a 5.4

3. Esperimenti non riusciti

3.1. SLERP

Nello studio della tecnica di interpolazione tramite SLERP (Spherical Linear Interpolation), si è evidenziata l'importanza di interpolare separatamente la norma dei vettori, per poi combinare il risultato con la loro direzione normalizzata. Questo approccio risulta fondamentale poiché SLERP, applicata esclusivamente alle direzioni normalizzate, preserva l'informazione angolare ma ignora la magnitudine dei vettori originali, che contiene dati significativi nell'ambito dei pesi del modello. Interpolare la norma consente dunque di mantenere l'informazione relativa alla scala dei vettori, garantendo che l'interpolazione complessiva rifletta fedelmente sia la direzione sia la grandezza dei parametri, migliorando la coerenza e l'efficacia della tecnica di interpolazione (Figure 5).

3.2. CIFAR-10 con MLP

Come mostrato in *Git-Rebasin*, le tecniche basate su permutazioni risultano efficaci anche su dataset più complessi come CIFAR-10, pur utilizzando un semplice MLP a tre layer nascosti,la cua accuracy di partenza si aggira intorno al 50%. (Figure 6)

Si è quindi analizzato l'impatto delle tecniche di allineamento basate su rotazioni ortogonali nella medesima configurazione. I metodi proposti raggiungono un'accuracy comparabile e nel complesso migliore rispetto all'interpolazione naive, tuttavia il valore della loss barrier aumenta invece di decrescere,ad eccezione di Procruste weight applicato alla matrice di correlazione. Si pensa che ciò possa essere dovuto al fatto che il modello $B_{allineato}$ non sia funzionalmente allineato al modello A (Figure 7).

Loss barrier	
Naive	0.90
svd weight	1.34
Procrustes weight l	1.66
Procrustes weight x	0.47
Procrustes activation	0.97
Stiefel	1.93

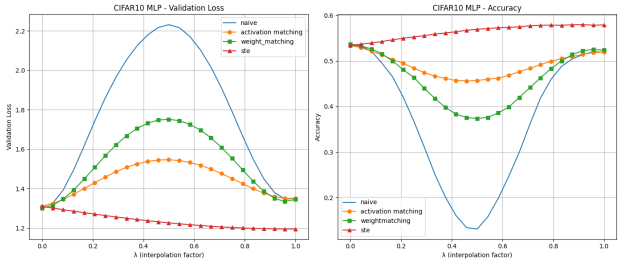

Figure 6. Validation Loss e Accuracy con allineamento basato su permutazione sul dataset CIFAR10

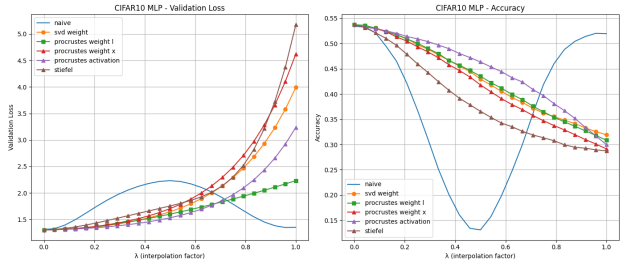

Figure 7. Validation Loss e Accuracy con allineamento basato su rotazioni sul dataset CIFAR10 (nella legenda del grafico i valori le scritte weight x e weight l sono inveritite)