Analisi Empirica della Dinamica di Apprendimento: Double Descent, Neural Collapse e Rappresentazioni Interne via CKA

Report Tecnico sui Modelli 1D

27 febbraio 2026

Sommario

Questo studio esplora la complessa dinamica di ottimizzazione e generalizzazione delle reti neurali profonde attraverso l'analisi di un dataset a una dimensione (MNIST-1D) corrotto con il 20% di rumore sulle etichette. Indaghiamo la presenza del fenomeno del *Deep Double Descent* al variare della larghezza del modello, dimostrando come le metriche di *Neural Collapse* (NC1-NC4) riflettano fedelmente l'errore di test. Inoltre, valutiamo il limite teorico del Neural Collapse al variare della dimensionalità delle feature e utilizziamo il *Centered Kernel Alignment* (CKA) per misurare la similarità delle rappresentazioni tra architetture diverse, epoche di addestramento e inizializzazioni (seed), rivelando che gli strati profondi subiscono derive maggiori e presentano un'alta degenerazione geometrica.

1 Introduzione

La teoria classica del machine learning suggerisce un trade-off bias-varianza a forma di "U", dove l'overfitting peggiora monotonicamente le prestazioni su dati non visti all'aumentare della capacità del modello. Tuttavia, le reti neurali moderne mostrano un comportamento di *Double Descent* [1]: superata la soglia di interpolazione, reti pesantemente sovra-parametrizzate generalizzano in modo eccellente.

Parallelamente, Papyan et al. [2] hanno dimostrato che nell'ultima fase di addestramento emerge un fenomeno geometrico noto come *Neural Collapse* (NC), caratterizzato da una contrazione della varianza intra-classe (NC1) e dalla convergenza delle medie di classe a un *Simplex Equiangular Tight Frame* (ETF) (NC2).

Questo studio replica e analizza questi fenomeni utilizzando varianti di CNN e ResNet1D, arricchendo l'indagine con lo studio delle metriche NC al variare della larghezza e la comparazione della similarità geometrica tramite CKA [3].

2 Metodologia

2.1 Dataset e Rumore

L'esperimento utilizza MNIST-1D, un dataset sintetico a bassa dimensionalità che mantiene le proprietà morfologiche del deep learning. Per indurre il fenomeno del Double Descent, è stato introdotto un **rumore sulle etichette del 20%**. Il rumore forza il modello a esaurire la sua capacità per memorizzare etichette errate in regime critico, esacerbando il picco di errore di generalizzazione.

2.2 Architetture e Ottimizzazione

Sono state confrontate due famiglie di modelli: `ResNet1D` (con connessioni residue e BatchNorm) e `StandardCNN1D`. Gli esperimenti sono stati condotti al variare del parametro di larghezza (*width*) $W \in [2, 64]$. Sono stati confrontati due ottimizzatori: SGD con momento (0.9) e Cosine Annealing, e Adam.

2.3 Metriche di Valutazione

Oltre all'Errore di Test, sono state calcolate le 4 metriche di Neural Collapse:

- **NC1 (Variability):** Rapporto tra covarianza intra-classe e inter-classe.
- **NC2 (Simplex ETF):** Distanza di Frobenius tra la matrice di covarianza delle medie di classe e il Simplex ETF ideale.
- **NC3 (Self-Duality):** Allineamento tra i pesi del classificatore e le medie delle feature di classe.
- **NC4 (NCC Mismatch):** Disaccordo percentuale tra le previsioni della rete e un classificatore *Nearest Class Center*.

Per l'analisi delle rappresentazioni interne è stato utilizzato il **Linear CKA**, che valuta la similarità tra due matrici di attivazioni normalizzando la loro covarianza incrociata rispetto alle varianze individuali.

3 Risultati e Interpretazione

3.1 Exp 1: Double Descent e Neural Collapse (Model-wise)

I grafici in prima riga (Fig. 1) mostrano chiaramente la curva di Double Descent. Attorno ai $10^3 - 10^4$ parametri, l'Errore di Test subisce un picco improvviso. Questo è il **regime di interpolazione critico**: il modello ha esattamente i gradi di libertà necessari per memorizzare il 20% di rumore, risultando in funzioni altamente non lineari e instabili che falliscono sui dati di test. Aumentando ulteriormente i parametri ($> 10^4$), subentra il "blessing of dimensionality": il bias induttivo dell'ottimizzatore (in particolare SGD) predilige soluzioni con norma minore e confini decisionali più lisci, riducendo l'errore.

Il risultato più rilevante è la **perfetta correlazione tra Errore di Test e le metriche NC1, NC2, NC3 e NC4**. Nel punto critico, il Neural Collapse "si rompe": l'elevato sforzo per memorizzare il rumore distrugge la simmetria geometrica dello spazio latente. Superata

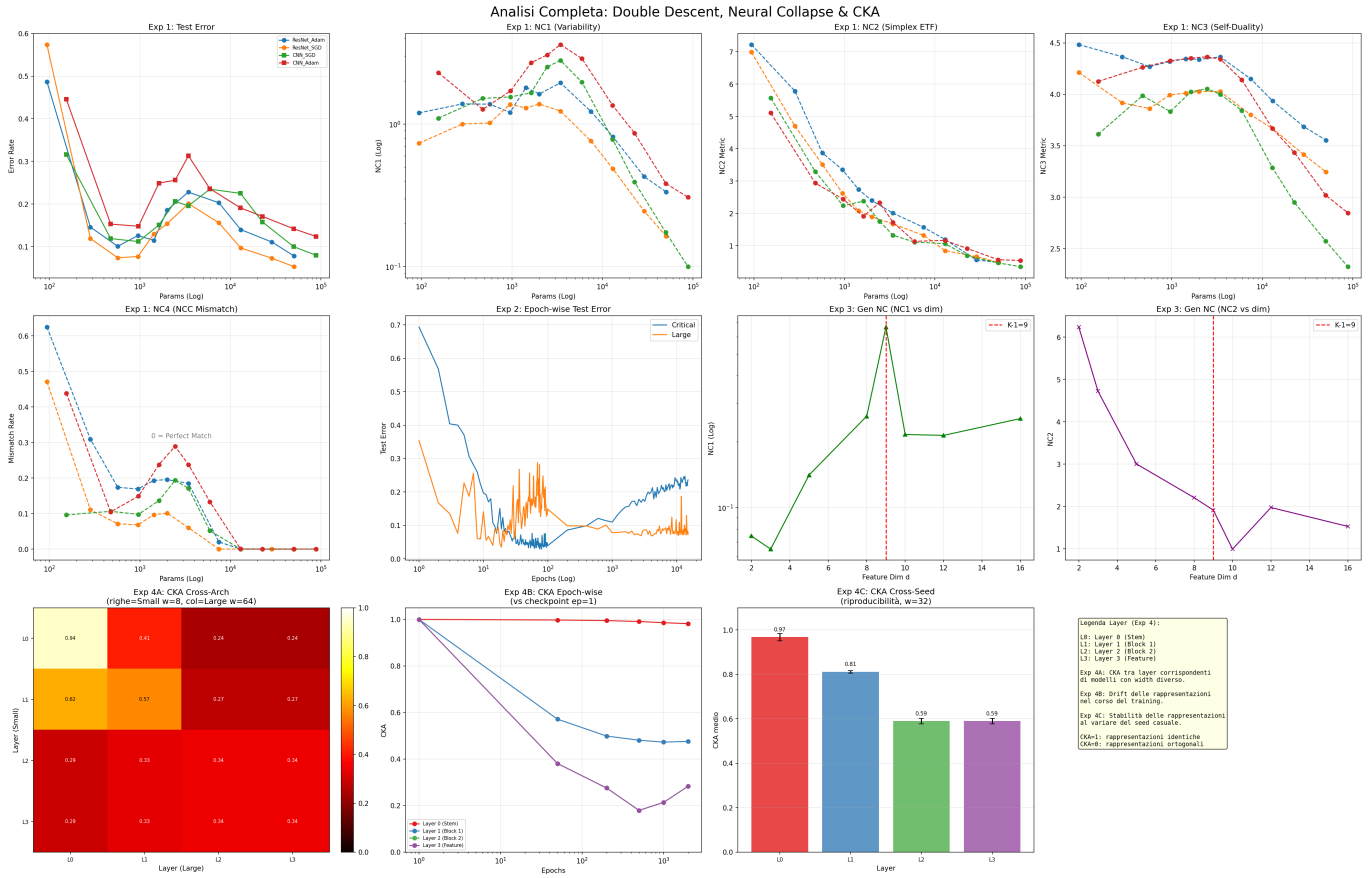

Figura 1: Risultati completi degli esperimenti. La riga 1 mostra l'evoluzione dell'errore e del Neural Collapse (NC1-NC3) al variare dei parametri (Exp 1). La riga 2 illustra NC4 (sinistra), l'evoluzione *epoch-wise* (centro) e il collasso generalizzato rispetto alle dimensioni (destra). La riga 3 mostra le analisi CKA layer-by-layer.

la soglia critica, le rappresentazioni collassano nuovamente verso un Simplex ETF perfetto. Si osserva inolte che **SGD domina nettamente Adam** nei regimi altamente sovra-parametrizzati, favorendo un Neural Collapse più estremo.

3.2 Exp 2: Dinamiche Epoch-wise

Analizzando l'errore lungo le epoche (Fig. 1, riga 2, centro), il modello "Critical" ($W = 12$, posto esattamente sul picco del Double Descent) mostra enormi fluttuazioni. L'addestramento su un dataset rumoroso con capacità limite rende la superficie di perdita frammentata. Al contrario, il modello "Large" ($W = 64$) apprende rapidamente le feature robuste nelle prime epoche e memorizza il rumore solo successivamente, mantenendo un errore di test basso e stabile.

3.3 Exp 3: Limiti Dimensionali del NC

La teoria geometrica impone che, per K classi, un Simplex ETF perfetto possa esistere solo se la dimensione delle feature $d \geq K - 1$. Essendo MNIST-1D composto da 10 classi, ci aspettiamo che il collasso avvenga per $d \geq 9$. I grafici "Gen NC" confermano questa teoria: per $d < 9$, NC2 è elevato poiché lo spazio latente non ha dimensioni sufficienti per ospitare 10 vettori equiangolari simmetrici. Esattamente in corrispondenza della linea tratteggiata rossa ($K - 1 = 9$), la metrica NC2 crolla verticalmente verso lo zero, confermando empiricamente i limiti spaziali previsti dalla matematica del Neural Collapse.

3.4 Exp 4: Analisi delle Rappresentazioni (CKA)

I risultati del CKA offrono una lente d'ingrandimento sul comportamento degli strati intermedi (*layer*):

4A. Cross-Architettura (Small vs Large): L'heatmap rivela che i primi strati (L0, L1) hanno un CKA relativamente alto, indicando che estraggono feature a basso livello simili (es. bordi e frequenze del segnale 1D). Tuttavia, procedendo verso gli strati profondi (L3), il CKA crolla a 0.34. Questa differenza è plausibile: il modello grande possiede i gradi di libertà per collassare le feature verso un ETF, mentre il modello piccolo non riesce a separare le classi in cluster compatti, portando a topologie completamente dissimili prima del classificatore lineare.

4B. Evoluzione Epoch-wise (Drift): Confrontando le attivazioni durante il training contro le attivazioni all'epoca 1, notiamo che lo strato L0 (Stem) rimane quasi invariato ($CKA \approx 1.0$). Questo indica che i filtri di base convogliano rapidamente o subiscono modifiche irrilevanti. Al contrario, L2 e L3 mostrano un forte "drift" (deriva). Le rappresentazioni profonde continuano a riorganizzarsi per decine di migliaia di epoche, spinte dalla necessità di minimizzare le metriche NC (comprimere la varianza intra-classe e separare i centri di classe).

4C. Riproducibilità Cross-Seed: Addestrando lo stesso modello ($W = 32$) con seed casuali differenti, emerge una chiara gerarchia. L0 è quasi deterministico ($CKA = 0.97$). Gli strati profondi (L2, L3) mostrano valori moderati ($CKA \approx 0.59$). Essendo fortemente over-parametrizzata, la rete possiede una **degenerazione geometrica**: vi sono infiniti modi per ruotare uno spazio latente pur formando un Simplex ETF perfetto. Poiché il CKA cattura similarità strutturali, diverse inizializzazioni portano a rotazioni isomorfiche distinte dello spazio finale, risolvendo il task con la stessa accuratezza ma con coordinate latenti dissimili.

4 Conclusioni

Questo studio empirico convalida e intreccia tre dei fenomeni più affascinanti del deep learning moderno. Abbiamo dimostrato che l'incapacità di generalizzare al punto di interpolazione (Double Descent) è intimamente legata alla distruzione della geometria latente (Neural Collapse). L'inserimento del rumore ha evidenziato come le reti massivamente parametrizzate e ottimizzate con SGD riescano a recuperare strutture geometriche ottimali (Simplex ETF) a condizione che la dimensione latente lo permetta ($d \geq K - 1$). Infine, l'analisi CKA ha confermato che tale collasso è un fenomeno dinamico e degenerativo, localizzato quasi esclusivamente negli ultimi strati dell'architettura.

Riferimenti bibliografici

[1] M. Belkin, D. Hsu, S. Ma, S. Mandal. *Reconciling modern machine-learning practice and the classical bias–variance trade-off.* Proceedings of the National Academy of Sciences, 2019.

[2] V. Papyan, X. Y. Han, D. L. Donoho. *Prevalence of neural collapse during the terminal phase of deep learning training.* Proceedings of the National Academy of Sciences, 2020.

[3] S. Kornblith, M. Norouzi, H. Lee, G. Hinton. *Similarity of neural network representations revisited.* International Conference on Machine Learning (ICML), 2019.