NOVA: Non-monotonic Orthogonal Variance Activation per Architetture Neurali Profonde

Preprint — Work in Progress

Autore
Laboratorio di AI e Matematica Applicata

21 febbraio 2026

Abstract

La scelta della funzione di attivazione incide significativamente sulla topologia della loss landscape e sulla capacità espressiva delle reti neurali profonde. In questo articolo preliminare introduciamo **NOVA** (*Non-monotonic Orthogonal Variance Activation*), una funzione ibrida razionale-esponenziale progettata analiticamente. NOVA estende i benefici delle attivazioni gating introducendo un termine di smorzamento razionale che modula la derivata seconda. Dimostriamo analiticamente come NOVA induca una contrazione controllata della varianza all'inizializzazione (richiedendo una scala dei pesi di $\approx 2.8/n_{\text{in}}$) e formalizziamo il fenomeno della *Covariance Shift Collision*, fornendo una spiegazione teorica per il suo degrado prestazionale in architetture basate su Batch Normalization. Attraverso esperimenti esplorativi a run singola su architetture basate su LayerNorm (Mini-ViT, Nano-GPT), osserviamo trend qualitativi incoraggianti rispetto alla baseline GELU. Nello *Scientific Machine Learning*, l'espressività heavy-tailed della derivata seconda di NOVA ha mostrato una riduzione dell'errore residuo sulle Physics-Informed Neural Networks (PINN) di un fattore $\sim 13\times$ su un'equazione di Burgers 1D, suggerendo un forte potenziale per la modellazione di singolarità fisiche.

Contents

1 Formulazione Matematica e Derivazioni

NOVA è definita come la composizione di un meccanismo di gating esponenziale e una regolarizzazione razionale sottrattiva:

$$f(x) = x \cdot \sigma(\beta x) - \frac{x}{1 + (\beta x)^2} \tag{1}$$

dove $\sigma(z) = (1 + e^{-z})^{-1}$ è la funzione logistica standard e β è un parametro continuo che modula la curvatura.

1.1 Analisi del Gradiente e Comportamento Asintotico

La derivata prima $f'(x)$, che governa il flusso del gradiente durante la retropropagazione, è data analiticamente da:

$$f'(x) = \sigma(\beta x) + \beta x \cdot \sigma(\beta x)\big(1 - \sigma(\beta x)\big) - \frac{1 - (\beta x)^2}{\big(1 + (\beta x)^2\big)^2} \tag{2}$$

Per valutare il comportamento in prossimità dell'origine (assumendo $\beta = 1$), calcoliamo $f'(0)$ passo per passo:

$$\begin{aligned} f'(0) &= \sigma(0) + 0 \cdot \sigma(0)\big(1 - \sigma(0)\big) - \frac{1 - 0^2}{\big(1 + 0^2\big)^2} \\ &= \frac{1}{2} + 0 - \frac{1}{1} \\ &= -0.5 \end{aligned}$$

Questo gradiente marcatamente negativo all'origine definisce la *sacca di non-monotonicità* di NOVA. A differenza di GELU o SiLU, dove la concavità negativa è debole, NOVA esercita una forte spinta repulsiva (hard-thresholding implicito) per attivazioni prossime allo zero. Asintoticamente, per $x \to +\infty$, il termine razionale decade a zero ($O(1/x)$) e $\sigma(x) \to 1$, restituendo $f(x) \approx x$. Per $x \to -\infty$, $f(x) \to 0$.

1.2 Conservazione della Varianza (NOVA-Init)

Il mantenimento dell'entropia del segnale nei layer profondi richiede una corretta inizializzazione dei pesi. Nel framework generalizzato di He et al. [1], la varianza ideale è:

$$\mathrm{Var}(W) = \frac{1}{n_{\text{in}} \cdot \mathbb{E}[f(X)^2]}, \quad X \sim \mathcal{N}(0, 1)$$

A differenza di ReLU, per cui $\mathbb{E}[\max(0, X)^2] = 0.5$ (regola di He standard $2/n_{\text{in}}$), NOVA possiede una regione contrattiva. Integrando numericamente l'aspettativa dell'attivazione al quadrato sotto distribuzione normale:

$$\mathbb{E}[f(X)^2] = \int_{-\infty}^{+\infty} \left(x\sigma(x) - \frac{x}{1 + x^2} \right)^2 \frac{e^{-x^2/2}}{\sqrt{2\pi}}\, dx \approx 0.357 \tag{3}$$

Pertanto, per prevenire il *signal decay* nelle prime fasi del training:

$$\mathrm{Var}(W_{\text{NOVA}}) = \frac{1}{n_{\text{in}} \cdot 0.357} \approx \frac{2.801}{n_{\text{in}}} \tag{4}$$

1.3 Derivata Seconda ed Espressività per le PINN

Scomponendo $f(x) = S(x) - R(x)$, dove $S(x) = x\sigma(\beta x)$ è il termine SiLU e $R(x) = x/(1+(\beta x)^2)$ è il termine razionale, deriviamo l'Hessiana 1D di NOVA:

$$f''(x) = \underbrace{2\beta\sigma'(\beta x) + \beta^2 x\,\sigma'(\beta x)\big(1 - 2\sigma(\beta x)\big)}_{S''(x)} - \underbrace{\frac{2\beta^2 x\big(\beta^2 x^2 - 3\big)}{\big(1+\beta^2 x^2\big)^3}}_{R''(x)} \tag{5}$$

La derivata seconda di funzioni esponenziali pure (come GELU) decade rapidamente per $|x| > 2$. In NOVA, il termine $R''(x)$ introduce una coda polinomiale (*heavy-tailed*), che permette all'operatore Laplaciano u_{xx} calcolato dalla rete di catturare variazioni ad alta frequenza (fronti d'urto, discontinuità termiche) che funzioni troppo lisce tendono a sfumare.

2 Contesto Architetturale e Stato dell'Arte

2.1 Confronto con Funzioni Alternative

NOVA si colloca in un ecosistema dove si cerca il bilanciamento tra levigatezza (*smoothness*) e capacità di thresholding. La Tabella 1 riassume le differenze strutturali con le principali funzioni di attivazione.

Table 1: Confronto tassonomico preliminare delle principali funzioni di attivazione.

Funzione	**Formula**	**Monotona**	C^∞	**Coda** f''	**Costo Rel.**
ReLU	$\max(0, x)$	Sì	No	Nulla	$1.0\times$
GELU [2]	$x\Phi(x)$	No	Sì	Esponenziale	$2.5\times$
SiLU/Swish [3]	$x\sigma(x)$	No	Sì	Esponenziale	$2.2\times$
Snake [4]	$x + \sin^2(x)$	No	Sì	Periodica	$4.0\times$
NOVA	Eq. (1)	No	Sì	Polinomiale	$2.8\times$

Rispetto a Snake (ideale per segnali periodici, ma instabile nei Transformer) e SiLU (eccellente per LLM ma debole su operatori differenziali), NOVA fornisce asintoti globali stabili con perturbazioni locali heavy-tailed.

2.2 Teorema della Covariance Shift Collision

Nel nostro studio esplorativo su Convolutional Neural Networks (CNN), abbiamo riscontrato un fallimento critico di NOVA abbinata alla Batch Normalization (BN). Un layer BN standardizza le pre-attivazioni X imponendo $\mu = 0$ e $\sigma^2 = 1$: circa il 68% della massa di probabilità ricade nell'intervallo $x \in [-1, 1]$.

In NOVA, questa regione coincide quasi esattamente con il dominio in cui $f'(x) \leq 0$, raggiungendo il minimo in $f'(0) = -0.5$. Poiché l'entropia differenziale dell'output $h(Y)$ con $Y = f(X)$ è limitata da $\mathbb{E}[\log|f'(X)|]$, forzare la maggior parte del segnale dove la derivata è contrattiva e non-monotona causa un drastico collasso della Mutua Informazione $I(X;Y)$ tra layer adiacenti. Al contrario, la Layer Normalization (usata nei Transformer) non impone una centratura globale, permettendo ai vettori di feature di mantenere una covarianza più naturale e di sfruttare la sacca negativa di NOVA come meccanismo selettivo per il rumore a bassa magnitudine.

2.3 Ablation Study Analitico

Dal punto di vista teorico, il comportamento di NOVA può essere disaccoppiato in due meccanismi:

- **Termine esponenziale** $x\sigma(\beta x)$ **(Gating):** Fornisce il limite asintotico lineare e garantisce la stabilità del gradiente per attivazioni fortemente positive. Senza questo componente, la limitatezza asintotica del termine razionale causerebbe la divergenza della rete.
- **Termine razionale** $x/(1+(\beta x)^2)$ **(Damping):** Modula e approfondisce la sacca non-monotona ed è l'unico responsabile della coda heavy-tailed nella derivata seconda (Eq. (5)), senza la quale le performance su operatori differenziali regredirebbero a quelle di SiLU.

3 Validazione Empirica Preliminare

Tutti gli esperimenti riportati in questa sezione sono **indagini a run singola (singolo seed)**, condotte per valutare trend di convergenza e formare ipotesi preliminari. Studi futuri con replicazione statistica saranno necessari per trarre conclusioni definitive.

3.1 Classificazione Visiva (Mini-ViT su CIFAR-10)

Abbiamo addestrato da zero un Mini-ViT su CIFAR-10 per 10 epoche con Data Augmentation e Layer Normalization. NOVA ha raggiunto una Test Accuracy del 68.31% contro il 67.54% di GELU. Il margine ridotto è compatibile con l'assenza di replicazione statistica; il risultato è riportato come osservazione qualitativa, suggerendo che NOVA non penalizzi la convergenza rispetto a GELU in architetture LayerNorm.

Table 2: Test Accuracy – Mini-ViT su CIFAR-10 (run singola, 10 epoche).

Epoca	NOVA	GELU
1	47.27%	46.97%
5	60.62%	59.18%
10	68.31%	67.54%

3.2 Modellazione Autoregressiva (Nano-GPT)

Abbiamo addestrato un Nano-GPT (≈10M parametri) sul dataset TinyShakespeare per 1000 iterazioni. NOVA ha prodotto una Validation Loss di 1.6949 rispetto a 1.7344 di GELU. L'osservazione suggerisce che il meccanismo di gating potenziato di NOVA possa contribuire a mitigare la saturazione dimensionale nei blocchi Feed-Forward dei Transformer.

Table 3: Validation Loss (Cross-Entropy) – Nano-GPT su TinyShakespeare (run singola).

Step	NOVA	GELU
0	4.3083	4.3501
400	1.9542	1.9747
1000	1.6949	1.7344

3.3 Modelli Generativi Continui (DDPM)

Su una U-Net DDPM per il denoising di immagini, dopo 10 epoche GELU ha ottenuto una MSE Loss di 0.0372 contro 0.0382 di NOVA. Ipotizziamo uno *Smoothness Trade-off*: l'integrazione differenziale (SDE) richiesta dalla Langevin Dynamics nei DDPM favorisce campi vettoriali estremamente lisci, mentre la forte irregolarità della sacca di NOVA introduce microscopiche perturbazioni nel reverse-sampling. Questo definisce un limite applicativo esplicito della funzione.

3.4 Scientific Machine Learning: PINN sull'Equazione di Burgers

Il test qualitativamente più rilevante ha riguardato una PINN addestrata a risolvere l'equazione non-lineare di Burgers 1D, che modella la formazione di onde d'urto fluidodinamiche e richiede il calcolo dell'operatore Laplaciano tramite auto-differenziazione doppia. Dopo 2000 step di ottimizzazione, la Physics Loss (residuo PDE) è stata:

Table 4: Physics Loss (residuo PDE) – PINN sull'Equazione di Burgers 1D (run singola).

Epoca	NOVA	GELU
200	0.07808	0.09757
1000	0.00207	0.00842
2000	0.00027	0.00353

Nonostante l'assenza di replicazione statistica, un abbattimento del residuo di oltre un ordine di grandezza (fattore $\sim 13\times$) suggerisce fortemente che la struttura polinomiale heavy-tailed della derivata seconda di NOVA (Eq. (5)) sia geometricamente superiore alle code esponenziali per approssimare singolarità e variazioni brusche tipiche delle onde d'urto. Questo risultato è coerente con la letteratura sulle attivazioni non-smooth per SciML [5].

4 Efficienza Computazionale (CUDA Fusion)

A causa della sua complessità composita, NOVA in modalità *Eager* su PyTorch richiede l'allocazione ripetuta di tensori temporanei intermedi, generando un Memory Bandwidth Bottleneck. Abbiamo sviluppato un **Fused CUDA Kernel** che sfrutta le istruzioni hardware *Fast Math* (`__expf`, `__fdividef`) direttamente nei registri del multiprocessore.

Un benchmark esplorativo su GPU NVIDIA T4 (tensore FP32 2048×2048, 100 iterazioni) ha prodotto i seguenti tempi per un ciclo Forward+Backward:

Table 5: Latenza Forward+Backward – GPU NVIDIA T4, FP32, tensore 2048×2048.

Funzione	Implementazione	Latency	VRAM Intermedia
NOVA	Python Eager	≈ 4.56 ms	128 MB
GELU	ATen Native Fused	≈ 0.57 ms	0 MB
NOVA	Custom CUDA Fused	≈ 0.92 ms	0 MB

Il kernel fused azzera il memory footprint intermedio e chiude circa il 90% del gap di latenza rispetto a GELU, rendendo NOVA sufficientemente efficiente per il training distribuito. L'overhead residuo (~ 0.35 ms) è attribuibile ai cicli aggiuntivi richiesti dalla divisione razionale rispetto alle sole operazioni esponenziali.

5 Limitazioni e Lavori Futuri

Il presente lavoro costituisce una *Proof of Concept* e presenta le seguenti limitazioni da affrontare in studi successivi:

- **Robustezza Statistica:** Tutti i risultati sperimentali derivano da run singole con singolo seed. La replicazione con seed multipli e test statistici formali è necessaria per validare i trend osservati.
- **Scala:** Gli esperimenti sono stati condotti su architetture ridotte (Mini-ViT, Nano-GPT 10M, PINN 1D). Il comportamento di NOVA a scale LLM ($\geq$7B parametri) e su dataset di pre-training di grandi dimensioni rimane non verificato.
- **Stabilità di β:** Il parametro apprendibile β mostra tendenze alla divergenza in assenza di adeguate strategie di *learning rate warmup*. Un'analisi sistematica della sua dinamica di ottimizzazione è lasciata a lavori futuri.
- **Ambito PDE:** Il vantaggio nelle PINN è stato osservato su equazioni con singolarità (Burgers 1D). Su PDE paraboliche totalmente lisce (es. equazione del calore), i risultati preliminari suggeriscono prestazioni comparabili a GELU, non superiori.

5.1 Euristica d'Uso per Ricercatori

In base alle intuizioni teoriche e ai trend osservati in questa indagine preliminare:

- **Consigliato:** Architetture LayerNorm / RMSNorm (Vision Transformers, LLM Decoder-only); Scientific Machine Learning su PDE con onde d'urto o singolarità.
- **Sconsigliato:** Reti con forte uso di Batch Normalization (es. ResNet); modelli generativi a diffusione continua (DDPM, score-based models).

References

[1] He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1026–1034.

[2] Hendrycks, D., & Gimpel, K. (2016). Gaussian error linear units (GELUs). *arXiv preprint arXiv:1606.08415*.

[3] Ramachandran, P., Zoph, B., & Le, Q. V. (2017). Searching for activation functions. *arXiv preprint arXiv:1710.05941*.

[4] Ziyin, L., Hartwig, T., & Ueda, M. (2020). Neural networks fail to learn periodic functions and how to fix it. *Advances in Neural Information Processing Systems*, 33, 1583–1594.

[5] Raissi, M., Perdikaris, P., & Karniadakis, G. E. (2019). Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378, 686–707.