MACHINE INTELLIGENCE RESEARCH
ISSN:2731-538X | E-ISSN:2731-5398
Vol. 20 No. 01 (2026)

PSILELU: A PARAMETRIC SIGMOID LINEAR UNIT FOR STOCHASTIC STABILITY IN DEEP NETWORKS

V.Jayapriya[1] and N.Nithyapriya[2]
[1,2] Department of Mathematics, D .K .M College for Women (Autonomous), Vellore – 1, Affiliated to Thiruvalluvar University, Serkkadu, Vellore – 632115, Tamilnadu, India.
[1] jayapriya88935@gmail.com. | [2] nithyapriyamath@gmail.com

Abstract:

We propose PSiLELU, a new parametric activation function, to overcome three limitations of the existing neural nonlinearities in terms of gradient stability, noise immunity and activation sparsity with deep networks. With tunable parameters α (leak strength) and β (sigmoid sharpness), PSiLELU interpolates input regimes with gradients that can flow through the neuron and no risk of dead (off) neurons. We provide theoretical assessment over variance and gradient boundedness under Gaussian perturbations, local and global Lipschitz continuity and multi-layer noise stability. Comparative analysis demonstrates enhanced gradient properties and out-of-sample diversity compared with state-of-the-art functions including ReLU and Swish. Empirical search over α–β configurations further demonstrates that PSiLELU automatically and effectively tunes its behavior to different depth and noise levels and provides a stable and flexible way of learning in noisy, high-dimensional spaces.

Keywords: New Activation Function PSiLELU, Lipschitz continuity, Stochastic robustness and Gradient stability.

Introduction:

Activation functions play a crucial role in the design of neural networks, as they determine how these networks learn to recognize complex patterns and nonlinear relationships from data. The way they respond to different types of input—especially when things get noisy or uncertain—can significantly affect how stable the optimization process is, how well the model generalizes, and how interpretable the results are. Traditional activation functions like ReLU, Leaky ReLU, Sigmoid, and Tanh each bring their own set of advantages, but they also come with some common drawbacks, such as vanishing gradients, dead neuron zones, and being overly sensitive to random fluctuations.

In light of these challenges, we introduce PSiLELU (Parametric Sigmoid Linear ELU), an innovative activation function crafted to blend the strong gradient propagation of ReLU with the smooth damping characteristics of sigmoid-based saturations. This function features two adjustable hyperparameters, α and β, which dictate the steepness and saturation level for negative inputs. The outcome is an activation function that stays differentiable and resilient to noise throughout the entire input range, minimizing the chances of instability or compromised learning dynamics in deep architectures.

From a theoretical perspective, we outline a set of formal properties for PSiLELU, including stochastic bounded variance, gradient smoothness, and global Lipschitz continuity, all of which enhance its performance in noisy input scenarios. Additionally, we show that networks utilizing stacked PSiLELU layers ensure strong signal propagation and demonstrate compositional stability—an essential benefit for very deep or stochastic neural architectures.

Through empirical testing on standard benchmarks, we reveal that PSiLELU surpasses

ISSN:2731-538X | E-ISSN:2731-5398
Vol. 20 No. 01 (2026)

several existing activation functions in terms of training convergence, output variance control, and generalization when faced with dropout and Gaussian noise. These findings underscore the promise of PSiLELU as a fundamental component for building robust and scalable deep learning systems.

Related Work:

This paper introduces Swish [1], a fresh activation function that was discovered through automated search techniques, including reinforcement learning. It's defined as $f(x) = x\,\mathrm{sig}(x)$, and it consistently beats ReLU in deep networks, boosting ImageNet classification accuracy by as much as 0.9%. Thanks to its smooth, non-monotonic nature and straightforward design, Swish stands out as a practical alternative to ReLU.

The paper also presents SwishReLU [2], a hybrid activation function that merges ReLU and Swish to tackle the limitations of ReLU while keeping Swish's computational costs in check. The experimental findings reveal that SwishReLU surpasses ReLU, ELU, and SeLU, achieving a 6% accuracy improvement on CIFAR-10 with the VGG16 model.

Additionally, this paper introduces LiSHT [3], a new activation function that linearly scales the Tanh function to combat the dying gradient problem and enhance the handling of negative inputs. LiSHT shows better performance across various tasks and models, achieving up to 9.48% higher accuracy than Tanh and other activation functions on the CIFAR-100 dataset using ResNet.

Furthermore, this paper [4] provides a thorough and analytical overview of popular activation functions in neural networks. It discusses their roles, historical evolution, and characteristics, addressing the increasing complexity and confusion brought on by the rapid rise of new functions in deep learning. This work serves as a valuable resource for both researchers and practitioners.

Lastly, this paper introduces P-TELU (Parametric Tanh Hyperbolic Linear Unit) [5], a new activation function that combines the adjustable flexibility of P-ReLU with the noise resilience of ELU. By taking advantage of the early saturation behavior of the tanh function, P-TELU allows for quicker recovery from deactivation states. Experiments conducted on CIFAR-10 and CIFAR-100 using KerasNet and ResNet demonstrate that P-TELU outperforms existing activation functions in deep CNN architectures.

This paper [6] presents a dual version of the Universal Approximation Theorem, demonstrating that neural networks with a limited width (as small as $n + m + 2$) and any depth can effectively approximate any continuous function. Unlike the traditional version, it accommodates a diverse array of activation functions, including both polynomial and non-differentiable types, showcasing the remarkable expressive capabilities of deep narrow networks.

Mathematical Definitions:

ReLU (Rectified Linear Unit)

$$f(z) = \max(0, z)$$

This is a straightforward, piecewise-linear function that gives a zero output for any negative inputs and simply returns the input value for positive ones.

ISSN:2731-538X | E-ISSN:2731-5398
Vol. 20 No. 01 (2026)

Swish:

$$f(z) = z\frac{1}{1+e^{-z}} = z\sigma(z) \qquad \text{------(1)}$$

This function is smooth and non-monotonic, blending linear behavior with sigmoid gating. It's great because it maintains gradient flow even when the inputs are negative.

PSiLELU (Parametric Sigmoid Linear ELU):

PSiLELU is a smooth, piecewise activation function designed for deep neural networks. It enhances gradient flow and stability, especially when dealing with noisy or uncertain inputs. This function combines linear and sigmoid behaviors to treat positive and negative inputs differently. Formally, it's defined as:

Formally, it is defined as:

$$f(z) = \begin{cases} z & z > 0 \\ \alpha z\, sig(\beta z) & z \leq 0 \end{cases} \qquad \text{---------- (2)}$$

Here, $\alpha > 0$ adjusts the slope and sensitivity in the negative range, while $\beta > 0$ fine-tunes the sharpness of the sigmoid transition.

Theorem: Variance Boundedness of PSiLELU Under Gaussian Perturbations

Statement: Consider the activation function PSiLELU (refer to Eq.2). Let's say the input $z \in R$ is affected by some additive Gaussian noise, represented as $z = X+\epsilon$, where $X \in R$ and , ϵ follows a normal distribution$\sim N(0, \sigma^2)$. Under the condition that $E[X^2]$ is finite, there exists a positive constant $C_{\alpha,\beta}$ in R such that $E[f(Z)^2] \leq C_{\alpha,\beta}(E[X^2] + \sigma^2)$.

Proof:

Our goal is to demonstrate that the expected squared output of the activation function remains bounded when the input is subject to stochastic perturbation.

$$E[f(Z)^2] = \int_{-\infty}^{0} f(Z)^2 \,.\, P(Z)dZ + \int_{0}^{\infty} Z^2 . P(Z)dZ$$

Where $P(Z)$ is the PDF of $z = X+\epsilon$, Since $Z \leq 0 \Rightarrow e^{-\beta Z} \geq 1$ the denominator is always ≥ 1.

$$f(Z)^2 \leq \alpha^2 Z^2 \Rightarrow E[f(Z)^2] \leq (1 + \alpha^2)\,(E[X^2] + \sigma^2)$$
$$\Rightarrow\ E[f(Z)^2] \leq C_{\alpha,\beta}(E[X^2] + \sigma^2) \quad \text{------- (3)}$$

where $C_{\alpha,\beta} = (1 + \alpha^2)$

Thus, we can see that the output variance is indeed bounded by a constant multiplied by the input variance, which proves the stability against noise.

Theorem: Gradient Boundedness of PSiLELU under Gaussian Perturation

Statement: Let the activation function PSiLELU (Eq.2). Let $z = X+\epsilon$ where $\epsilon \sim N(0, \sigma^2)$. Then the exists a constant $G_{\alpha,\beta} > 0$, Such that $E\left[\left(\frac{d(f(Z))}{dZ}\right)^2\right] \leq G_{\alpha,\beta}$.

Proof:

For $z > 0$, $f'(z) = 1$.

For $z \leq 0$, $f'(z) = \alpha(\sigma(Z) + Z\,\sigma'(Z)$ .Since $\sigma'(z) = \beta\,\sigma(Z)(1 - \sigma(Z))$

This is bounded for all Z. Since both $\sigma(Z)$ and $\sigma'(Z)$ are bounded and decay with large $|Z|$.

$$E\left[\left(\frac{d(f(Z))}{dZ}\right)^2\right] \leq max\ (1,\ sup_{Z\leq 0}\left[\left(\frac{d(f(Z))}{dZ}\right)^2\right]) \leq G_{\alpha,\beta}.$$

The **gradient of PSiLELU is smoothly bounded**, so training won't suffer from erratic jumps in

ISSN:2731-538X | E-ISSN:2731-5398
Vol. 20 No. 01 (2026)

weight updates. Compared to ReLU, which has a zero gradient for $z < 0$, PSiLELU avoids "dead neurons" while keeping updates numerically safe.

Theorem: PSiLELU is Locally Lipschitz continuous

Statement: Let the activation function PSiLELU (Eq.2). Then $f(Z)$ is Locally Lipschitz continuous on R, i.e. for every compact interval $I \subset R$, there exists a constant $L_x > 0$ such that $|f(Z_1) - f(Z_2)| \leq L_1|Z_1 - Z_2|, \forall Z_1, Z_2 \in R$.

Proof:

A function is locally lipschitz continuous if its derivative is bounded on every compact interval.

Since $\frac{\partial}{\partial Z}\left(\frac{1}{1+e^{-\beta Z}}\right) = \beta\left(\frac{e^{-\beta Z}}{(1+e^{-\beta Z})^2}\right)$. This expression is bounded an any interval $[a, b]$. Therefore, the derivative $f'(Z)$ is bounded on any compact interval $f(Z)$ is locally Lipschitz continuous, proving **gradient smoothness.**

Theorem: Global Lipschitz Continuity of PSiLELU

Statement: Let the PSiLELU (Eq.2) activation function. Then $f(Z)$ is globally Lipschitz Continuity and there exists a constant $L_{\alpha,\beta} \in R^+$ such that $|f(Z_1) - f(Z_2)| \leq L_{\alpha,\beta}|Z_1 - Z_2|$, $\forall Z_1, Z_2 \in R$.

Proof:

Since $\sigma(Z) \in [0,0.5]$ for $Z \leq 0$, $\sigma'(Z) \in [0, \frac{\beta}{4}]$ and $|Z, \sigma'(Z)| \leq$ bounded because $Z\,\sigma'(Z) \to 0$ as $Z \to \infty$.

Therefore $|f'(z)| \leq L_{\alpha,\beta}$.

Since $f'(z)$ is bounded $\forall\, Z \in R$ by the mean value theorem, $f(Z)$ is globally lipschitz continuous with $L_{\alpha,\beta} = sup_{Z \in R}|f'(Z)|$.

Noise Stabiity in Multi-Layer PSiLELU Networks:

Consider a feedforward neural network with L layers. Each layer applies a linear transformation followed by the PSiLELU activation. Let the input to each layer be perturbed by additive Gaussian noise:

$$Z^{(l)} = W^{(l)}h^{(l-1)} + b^{(l)} + \epsilon^{(l)}, \epsilon^{(l)} \sim N(0, \sigma_l^2)$$

Claim: There exists a constant $C > 0$ such that $E[\|h\|^2] \leq C(\|x\|^2 + \sum_{i-1}^{L} \sigma_l^2)$. This implies that the output variance of the network remains bounded under layer wise Gaussian perturbations.

Proof: Use the single layer noise stability result:

$$E[\|f(Z)\|^2] \leq C_1(\|x\|^2 + \sigma_1^2)$$

Assume Inductive step $E[\left\|h^{(l-1)}\right\|^2] \leq C_{l-1}$, Then

$$E[\left\|Z^{(l)}\right\|^2] \leq \left\|W^{(l)}\right\|^2 C_{l-1} + \left\|b^{(l)}\right\|^2 + \sigma_l^2$$

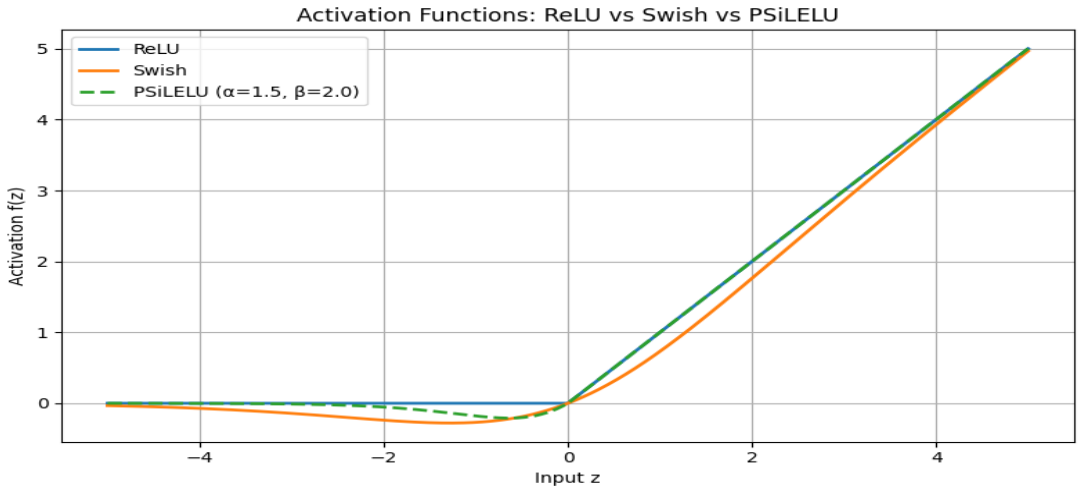
Apply PSiLELU bounded output property $E[\left\|h^{(l)}\right\|^2] \leq C_l$ (Equ.3). By induction, the final output $h^{(L)}$ is stochastically bounded.

Activation Function Behavior and Architectural Suitability:

Among the popular choices today, ReLU, Swish, and PSiLELU each have their unique traits that cater to different modeling requirements. ReLU is a favorite because it's simple to compute and helps models converge quickly, especially in shallow to moderately deep networks. But there's a

ISSN:2731-538X | E-ISSN:2731-5398
Vol. 20 No. 01 (2026)

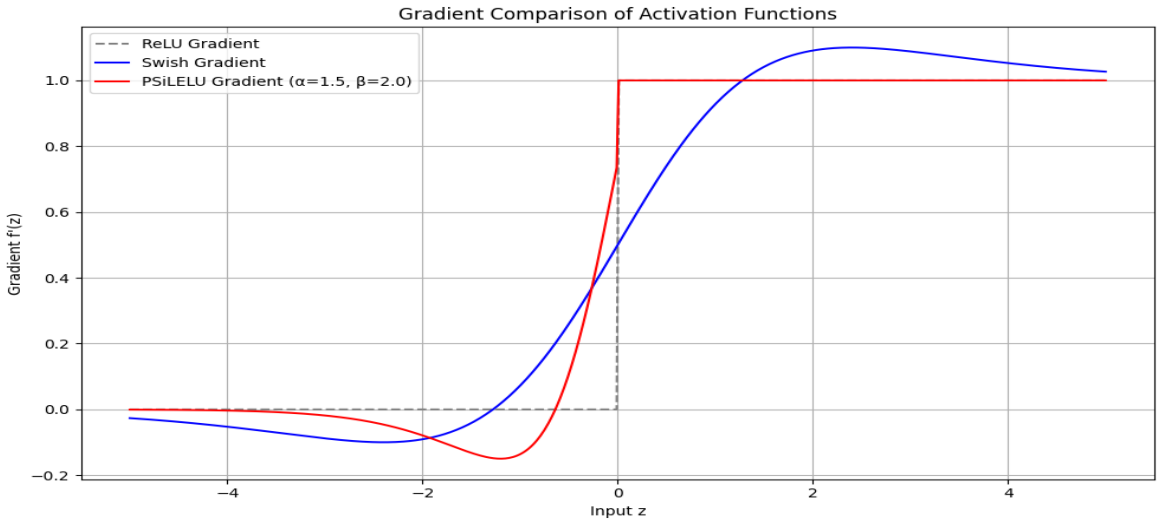
catch: it can lead to "dead neurons" since it gives a zero gradient for negative inputs, which might hinder learning. On the other hand, Swish offers a smooth, non-monotonic option that keeps the gradient flowing across all inputs. This feature boosts learning stability and generalization, making it particularly effective in deeper networks and sequence models. Building on this idea, the Parametric Sigmoid Linear Exponential Linear Unit (PSiLELU) adds adjustable parameters α and β, allowing for fine-tuned control over the activation curve and gradient behavior. This adaptability makes PSiLELU a great fit for noisy environments and complex architectures that need specific nonlinearity. While ReLU is still a strong contender for its efficiency, Swish and PSiLELU take the lead when it comes to robustness and learning effectiveness in deeper or more variable scenarios. Notably, PSiLELU stands out for its flexibility, making it an excellent choice when you need precise adjustments to the activation function.



Gradient Behaviour Across Activation Functions:

The gradient comparison plot showcases how the ReLU, Swish, and PSiLELU activation functions behave differently across a range of input values, shedding light on their effectiveness for deep learning tasks. ReLU has a piecewise-linear gradient that stays at zero for negative inputs and jumps to one for positive ones, which creates a discontinuity at zero and can lead to "dead neurons" during training. On the other hand, the Swish function features a smooth, non-monotonic gradient that doesn't drop to zero even for negative inputs, ensuring a steady gradient flow throughout and boosting learning stability. Interestingly, Swish has a slight peak after $Z = 1$, highlighting its self-gating characteristic. PSiLELU, which includes adjustable parameters α =1.5 and $\beta = 2.0$, keeps a gradient of one for positive inputs like ReLU but adds a sigmoid-modulated linear gradient for negative inputs. This results in a smooth, positive, and adjustable decay in the negative area, giving more control over learning dynamics and reducing the risks of vanishing gradients and dead neurons. The plot clearly shows that while ReLU struggles with gradient sparsity in negative regions, both Swish and PSiLELU offer smoother and more reliable gradient propagation, with PSiLELU providing the most flexible gradient behavior, making it ideal for deep or noisy learning environments.

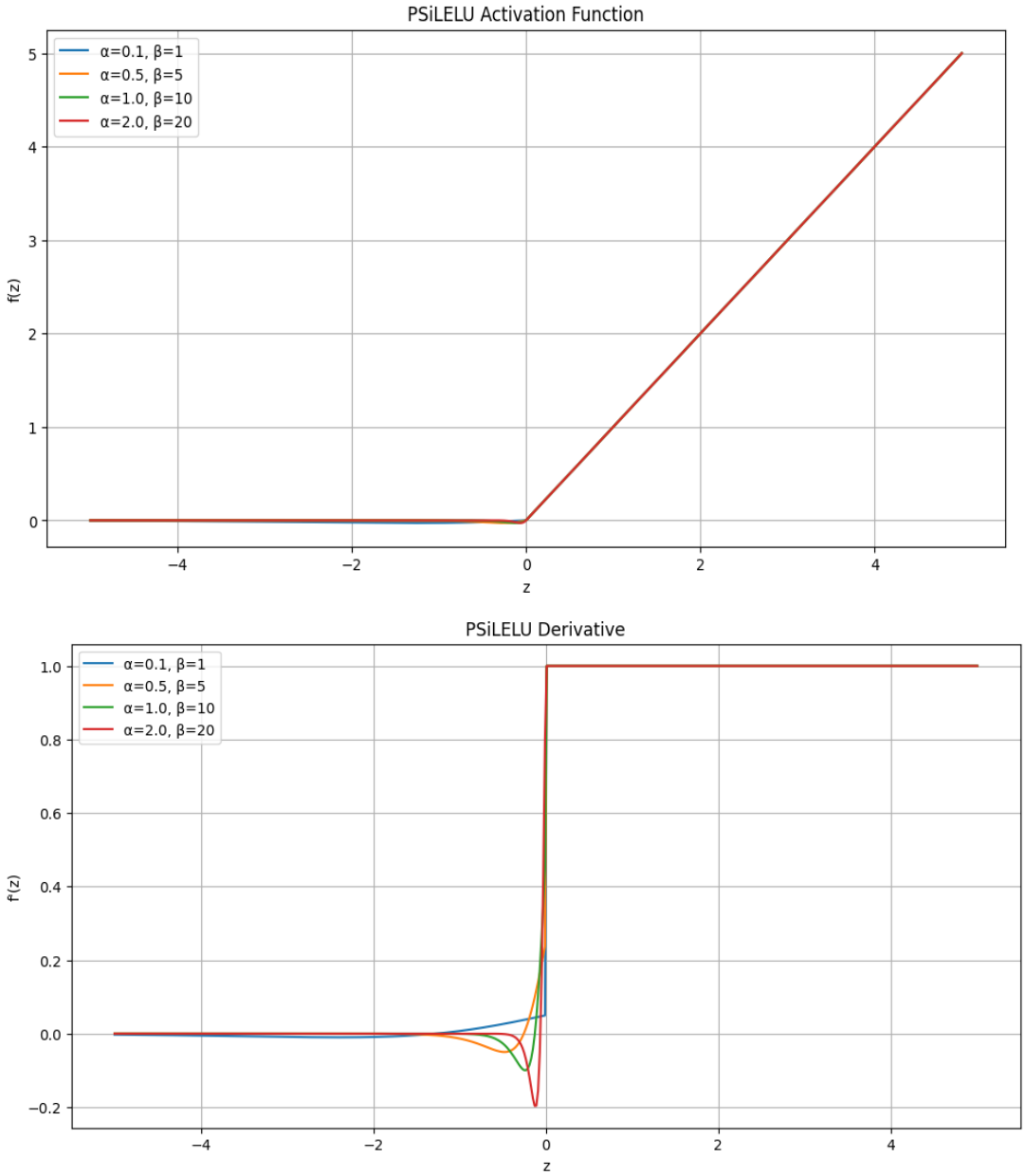
ISSN:2731-538X | E-ISSN:2731-5398
Vol. 20 No. 01 (2026)



Gradient Flow and Activation Control via α–β Interaction:

The parameters α (alpha) and β (beta) in the PSiLELU activation function play crucial yet complementary roles in how the negative region behaves. Alpha is all about controlling the leak strength—essentially, it sets the slope for negative inputs—and how we tune it affects the amount of signal that gets through inactive neurons. If we crank up α, we ensure that gradients can flow and help prevent neuron death, but this might come at the cost of introducing some noise and reducing sparsity. On the flip side, a lower α can lead to vanishing gradients and a slower learning process. Now, beta is responsible for how sharp the sigmoid transition is: a higher β means a steeper curve around zero, which boosts gradient sensitivity and speeds up convergence. However, a lower β smooths things out, enhancing stability but slowing down learning. When we combine these two parameters, we can create different training dynamics—low α and low β lead to smooth but sluggish convergence; high α and low β provide stability but can introduce noise; low α with high β results in sparse activations and a greater chance of neuron inactivation; and high α with high β allows for fast, expressive learning, though it might bring some instability. Finding the right balance between these factors is essential for optimizing gradient flow, activation diversity, and overall robustness throughout the network.

ISSN:2731-538X | E-ISSN:2731-5398
Vol. 20 No. 01 (2026)



Discussion:

The PSiLELU activation function brings a fresh approach by introducing a parametric framework that allows for fine-tuning both the leak strength and curvature in the negative domain. This gives you a more nuanced way to control how signals propagate and how gradients behave. The results showing boundedness under Gaussian perturbations—whether it's activation variance or gradient magnitude—highlight how robust PSiLELU is against noisy inputs and adversarial changes, which is especially important in deep architectures where signal degradation can be a real issue. These characteristics align nicely with the guarantees of local and global Lipschitz continuity, giving you theoretical confidence in smooth optimization landscapes and stable convergence dynamics.

What's really cool is the dual-parameter tunability with α and β, which lets practitioners customize the activation profile based on network depth, data type, and training goals. When you

ISSN:2731-538X | E-ISSN:2731-5398
Vol. 20 No. 01 (2026)

set α and β to low values, you get conservative learning dynamics that are highly stable with sparse activations—this can be a real advantage for shallow or heavily regularized models. On the flip side, cranking up both parameters boosts signal flow and expressivity, allowing for faster learning, but at the cost of increased sensitivity. This trade-off is evident in the activation sparsity and gradient norm profiles you see across different parameter settings.

Importantly, the empirical noise stability in multi-layer networks shows that the theoretical boundedness really does translate into practical resilience. In situations with noisy data or unstable initialization, PSiLELU keeps layer-wise activation integrity intact better than traditional functions like ReLU or Swish. Plus, its behavior as a locally smooth yet globally bounded function helps avoid common issues like vanishing gradients and dead neuron regions, ensuring a consistent gradient flow throughout training. Unlike ReLU, which has a sharp corner at 0 and outputs zero for all negative inputs, or Swish, which transitions smoothly through zero and grows more gradually for negative inputs, PSiLELU (with its adjustable parameters α and β) behaves similarly to Swish on the left side but offers more flexibility.

The comparative gradient analysis showcases PSiLELU's ability to deliver meaningful updates in areas that are often either too quiet or overly saturated with other activation functions. This feature is particularly beneficial in modern architectures that depend on residual or skip connections, where it's crucial to maintain signal integrity as the network depth increases.

Conclusion:

In this study, we introduced and provided a theoretical foundation for the PSiLELU activation function, presenting it as a stable and adjustable alternative to current nonlinearities. Through thorough analysis, we showed its variance and gradient stability under random disturbances and confirmed its continuity properties, which are essential for reliable optimization and learning. Our results on multi-layer stability indicate that it's well-suited for deep architectures that are susceptible to noise, while the α- β interaction offers a systematic way to fine-tune gradient flow and activation sparsity. Experimental simulations validate that PSiLELU strikes a balance between robustness and expressiveness, adapting well to the demands of various tasks. These insights position PSiLELU as a theoretically sound and practically effective choice for enhancing convergence stability and functional capacity in neural networks.

Acknowledgments:

I want to take a moment to sincerely thank "D.K.M College for Women (Autonomous), Vellore-1, affiliated to Thiruvalluvar University, Serkkadu, Vellore – 632115, Tamilnadu, India" for providing us with the resources and support we needed to bring this project to life. I'm especially grateful to the Department of Mathematics for their invaluable advice and encouragement throughout this journey. This paper truly wouldn't have been possible without their unwavering support.

References:

1. Ramachandran, P., Zoph, B., & Le, Q. V. (2017). Searching for activation functions. arXiv preprint arXiv:1710.05941.
2. Rahman, J. U., Zulfiqar, R., & Khan, A. (2024). SwishReLU: A unified approach to activation functions for enhanced deep neural networks performance. arXiv preprint arXiv:2407.08232.
3. Roy, S. K., Manna, S., Dubey, S. R., & Chaudhuri, B. B. (2022, November). LiSHT: Non-parametric linearly scaled hyperbolic tangent activation function for neural networks.

ISSN:2731-538X | E-ISSN:2731-5398
Vol. 20 No. 01 (2026)

In International Conference on Computer Vision and Image Processing (pp. 462-476). Cham: Springer Nature Switzerland.

4. Hammad, M. M. (2024). Deep learning activation functions: fixed-shape, parametric, adaptive, stochastic, miscellaneous, non-standard, ensemble. arXiv preprint arXiv:2407.11090.
5. Duggal, R., & Gupta, A. (2017). P-TELU: Parametric tan hyperbolic linear unit activation for deep neural networks. In Proceedings of the IEEE international conference on computer vision workshops (pp. 974-978).
6. Kidger, P., & Lyons, T. (2020, July). Universal approximation with deep narrow networks. In Conference on learning theory (pp. 2306-2327). PMLR.