

FROM PROBABILITY DENSITY TO NEURAL NON-LINEARITY: THE 4RS FAMILY AS A NEW PARAMETRIC CLASS OF ACTIVATION FUNCTIONS FOR NEURAL NETWORKS

DA DENSIDADE DE PROBABILIDADE À NÃO LINEARIDADE NEURAL: A FAMÍLIA 4RS COMO NOVA CLASSE PARAMÉTRICA DE FUNÇÕES DE ATIVAÇÃO PARA REDES NEURAIS

DE LA DENSIDAD DE PROBABILIDAD A LA NO LINEALIDAD NEURONAL: LA FAMILIA 4RS COMO NUEVA CLASE PARAMÉTRICA DE FUNCIONES DE ACTIVACIÓN PARA REDES NEURONALES

Regis Nunes Vargas¹
Raquel Romes Linhares²

ABSTRACT: This article presents the 4Rs-Gate, a novel parametric activation function for neural networks. While popular self-gated activation functions, such as GELU and Swish, multiply the input by the cumulative distribution function (CDF) of a fixed unimodal density, the 4Rs-Gate introduces innovation by utilizing the CDF of the 4Rs Family. This function consists of a convex mixture between a cubic B-spline and the normalized magnitude of a mother wavelet. The interpolation between the smoothness of the cubic B-spline and the spatial localization of the wavelet is controlled by a single parameter, a , with values constrained between 0 and 1. In this work, we derived its analytical properties, validated the exactness of its gradients using finite difference approximations, and evaluated the model's overall performance using multilayer perceptrons (MLP) trained across two synthetic benchmark tasks with five random seeds. In numerical experiments, the 4Rs-Gate (configured with $a = 0.9$) outperformed traditional activation functions including ReLU, GELU, Swish, and Tanh on the task featuring multi-scale local structures, while achieving competitive and equivalent performance to classical activation baselines on the task lacking such local spatial characteristics.

1

Keywords: Activation function. 4Rs Family. Wavelets. Neural Networks.

RESUMO: Este artigo apresenta o 4Rs-Gate, uma nova função de ativação paramétrica para redes neurais. Enquanto ativações auto-moduladas populares, como GELU e Swish, multiplicam a entrada pela função de distribuição acumulada (FDA) de uma densidade fixa e unimodal, o 4Rs-Gate inova ao utilizar a FDA da Família 4Rs. Essa função consiste em uma mistura convexa entre uma B-spline cúbica e o módulo normalizado de uma wavelet mãe. A interpolação entre a suavidade da B-spline e a localização da wavelet é controlada por um único parâmetro, a , com valores entre 0 e 1. No trabalho, foram derivadas suas propriedades analíticas, validados os gradientes por diferenças finitas e avaliado o desempenho do modelo em perceptrons multicamadas (MLP) treinados em duas tarefas sintéticas com cinco sementes aleatórias. Nos testes, o 4Rs-Gate (com $a = 0.9$) superou as funções tradicionais ReLU, GELU, Swish e Tanh na tarefa com estrutura local em múltiplas escalas, enquanto na tarefa sem essa estrutura seu desempenho foi equivalente ao das ativações clássicas.

¹Doutor em Engenharia Elétrica pela Faculdade de Engenharia Elétrica da Universidade Federal de Uberlândia e professor na Secretaria Municipal de Educação de Uberlândia.

²Doutora em Matemática pela UFRGS e professora pesquisadora do Instituto de Matemática e Estatística da Universidade Federal de Uberlândia.

Palavras-chave: Função de ativação. Família 4Rs. Wavelets. Redes Neurais.

RESUMEN: Este artículo presenta el 4Rs-Gate, una nueva función de activación paramétrica para redes neuronales. Mientras que las activaciones auto-moduladas populares, como GELU y Swish, multiplican la entrada por la función de distribución acumulada (FDA) de una densidad fija y unimodal, el 4Rs-Gate innova al utilizar la FDA de la Familia 4Rs. Esta función consiste en una mezcla convexa entre una B-spline cúbica y el módulo normalizado de una wavelet madre. La interpolación entre la suavidad de la B-spline cúbica y la localización espacial de la wavelet está controlada por un único parámetro, a , cuyos valores se encuentran acotados entre 0 y 1. En este trabajo, se derivaron sus propiedades analíticas, se validó la exactitud de sus gradientes mediante aproximaciones de diferencias finitas y se evaluó el desempeño general del modelo en perceptrones multicapa (MLP) entrenados en dos tareas sintéticas de prueba con cinco semillas aleatorias. En los experimentos numéricos, el 4Rs-Gate (configurado con $a = 0,9$) superó a las funciones de activación tradicionales incluyendo ReLU, GELU, Swish y Tanh en la tarea que presentaba estructuras locales a múltiples escalas, mientras que en la tarea sin dicha estructura su rendimiento fue equivalente al de las activaciones clásicas.

Palabras clave: Función de activación. Familia 4Rs. Wavelets. Redes Neuronales.

INTRODUCTION

Few design choices in a neural network are as cheap to make and as consequential in their effects as the activation function: swapping a single line of code can change how well gradients flow, how fast the network converges, and even whether it converges at all. It is no surprise, then, that the search for better activations remains an active, and largely empirical, research area. Jagtap AD and Karniadakis GE (2022), in a recent survey, catalogue dozens of proposals and note that no single family dominates uniformly across tasks. After the simple, monotonic units, namely sigmoid, hyperbolic tangent and ReLU, the frontier shifted to smooth, non-monotonic activations such as GELU (HENDRYCKS D and GIMPEL K, 2016), Swish (RAMACHANDRAN P, et al., 2017) and Mish (MISRA D, 2019), now standard in large-scale architectures.

A common thread, not always made explicit, links these three proposals: all of them can be written as the product between the input x and a “gate” $\Phi(x) \in [0,1]$ that modulates how much of the signal passes through, that is, $\varphi(x) = x \cdot \Phi(x)$. In GELU, Φ is the CDF of a standard normal; in Swish, the logistic sigmoid, the CDF of a logistic distribution. Choosing Φ as the CDF of a fixed, unimodal density is not, however, the only option: it is merely the simplest one, and is rarely questioned as such.

While this lineage of “self-gated” activations was consolidating, two much older traditions were pursuing the same goal along different paths, giving the activation a flexible shape, learned or purpose-built, rather than fixing it a priori. The first is that of adaptive spline

activations, initiated by Vecchi L, et al. (1998), in which the control points of a Catmull-Rom spline are trained jointly with the network's weights, letting each neuron learn the shape of its own non-linearity. The second is that of wavelet networks, proposed by Zhang Q and Benveniste A (1992) under the name “wavelons”: neurons whose activation is a translated and dilated mother wavelet, exploiting the time-frequency localization that characterizes wavelet analysis (ZAPRANIS A and ALEXANDRIDIS A, 2013). More recent reviews show that this idea keeps yielding variants, from Gaussian wavelets in deep networks (HERRERA O and PRIEGO B, 2022) to adaptive wavelet-based activations for physics-informed networks (MURARI K, 2026), yet it remains a comparatively small niche next to the near-total dominance of GELU/Swish-style smooth activations. These three lineages, namely CDF-based gates, adaptive splines and wavelons, evolved in parallel, with little direct intersection. This is where the 4Rs Family (LINHARES RR and VARGAS RN, 2026) opens an unexpected door. Designed for an entirely different problem, parametric density estimation, the 4Rs Family is a convex mixture between a cubic B-spline and the normalized modulus of a mother wavelet, valid as a probability density for any integrable wavelet and any mixture weight $a \in [0,1]$. That guarantee is exactly the missing ingredient needed to unite the three lineages: because the mixture is always a valid density, its CDF F_a can be plugged directly into the self-

3

The goal of this work is to formalize this bridge: (i) briefly review the construction of the 4Rs Family; (ii) propose the 4Rs-Gate activation, $\varphi_a(x) = x \cdot F_a(x)$, obtained from the CDF of the mixture; (iii) derive its analytical properties (asymptotic limits, derivative, smoothness, dependence on a); (iv) numerically illustrate its behaviour with the Daubechies db2 wavelet, comparing it to ReLU and GELU; and (v) discuss its trainability and its relation to recent B-spline-based architectures such as Kolmogorov-Arnold Networks (LIU Z, et al., 2024).

THE 4RS FAMILY

Let $B(x)$ be a normalized cubic B-spline ($B(x) \geq 0$, $\int B(x) dx = 1$) and $\psi(x)$ a mother wavelet with compact support, not identically zero. Define the normalized wavelet modulus

$$g(x) = \frac{|\psi(x)|}{\|\psi\|_1}, \quad F_g(x) = \int_{-\infty}^x g(t) dt, \quad (1)$$

which is, by construction, a probability density ($g(x) \geq 0$ and $\int g(x) dx = 1$, since it is the modulus of an integrable function divided by its L^1 norm). The 4Rs Family is then defined as the convex mixture

$$f_a(x) = a B(x) + (1 - a) g(x), \quad a \in [0,1]. \quad (2)$$

Linhares RR and Vargas RN (2026) formally prove that f_a is a valid probability density for every $a \in [0,1]$ and for any integrable mother wavelet, and illustrate the family's behaviour with the Daubechies order-2 wavelet (db2), built numerically via a cascade algorithm from the filter coefficients. As $a \rightarrow 1$, f_a approaches the cubic B-spline, smooth and unimodal; as $a \rightarrow 0$, f_a approaches the wavelet modulus, typically multimodal and localized.

FROM DENSITY TO ACTIVATION: THE 4RS-GATE CONSTRUCTION

Several modern activations adopt the self-gated form $\varphi(x) = x \cdot \Phi(x)$, where Φ is a monotonically increasing function from $[-\infty, +\infty]$ to $[0,1]$. In GELU, $\Phi(x) = P(Z \leq x)$ for $Z \sim N(0,1)$; in Swish, $\Phi(x) = 1/(1 + e^{-x})$, the logistic CDF. In both cases, Φ is the CDF of a fixed, unimodal density with no shape parameters beyond scale/location.

4

The 4Rs Family offers a natural generalization of this scheme: instead of fixing Φ as the CDF of a single density, we define Φ as the CDF of the mixture f_a itself, inheriting the shape parameter a . We define the cumulative distribution function of the 4Rs Family as

$$F_a(x) = a F_B(x) + (1 - a) F_g(x), \quad (3)$$

where F_B is the CDF of the cubic B-spline and F_g is the CDF of the normalized wavelet modulus. Since f_a is a valid density for every a (LINHARES RR and VARGAS RN, 2026), F_a is, by construction, a monotonically non-decreasing function with $F_a(-\infty) = 0$ and $F_a(+\infty) = 1$, exactly the requirements of a well-behaved gate.

A practical advantage of the 4Rs Family is that, for the centered cubic B-spline $B_3(x)$ (support $[-2,2]$), the CDF has a closed form, obtained by direct integration of the piecewise definition. For $x \in [-2, -1]$:

$$F_B(x) = \frac{(2+x)^4}{24}, \quad (4)$$

and, for $x \in [-1, 0]$:

$$F_B(x) = \frac{1}{2} + \frac{2x}{3} - \frac{x^3}{3} - \frac{x^4}{8}, \quad (5)$$

with $F_B(x) = 1 - F_B(-x)$ for $x > 0$, by symmetry of B_3 . The CDF of the wavelet modulus, F_g , has in general no closed form, but since ψ has compact support and is generated numerically by a cascade algorithm from a finite number of filter coefficients, F_g can be tabulated once (by numerical integration, e.g. the trapezoidal rule) and reused via table lookup during training, at negligible computational cost.

We define the 4Rs-Gate activation as

$$\varphi_a(x) = x \cdot F_a(x). \quad (6)$$

This is the central proposal of this work: a family of activations parameterized by $a \in [0, 1]$ and by the choice of mother wavelet ψ , which continuously interpolates between a smooth regime close to GELU/Swish (as $a \rightarrow 1$) and a regime with multiple non-monotonic regions, localized at the characteristic scales of ψ (as $a \rightarrow 0$).

ANALYTICAL PROPERTIES

5

Since f_a is a probability density with compact support (the sum of two compactly supported functions), there exists $M > 0$ such that $F_a(x) = 0$ for $x \leq -M$ and $F_a(x) = 1$ for $x \geq M$. Hence:

$$\lim_{x \rightarrow +\infty} \varphi_a(x) = x, \quad \lim_{x \rightarrow -\infty} \varphi_a(x) = 0, \quad (7)$$

that is, φ_a reproduces ReLU exactly outside a compact neighbourhood of the origin, the same asymptotic property as GELU and Swish, but, unlike them (whose gates have infinite exponential tails), the transition of the 4Rs-Gate occurs on a finite support, determined by the supports of B and of ψ .

By the product rule, the derivative of the activation is

$$\varphi'_a(x) = F_a(x) + x f_a(x). \quad (8)$$

This expression is interesting from a training perspective: the first term, $F_a(x)$, guarantees a non-negative, smooth gradient “floor” (inherited from the C^2 regularity of F_B when a is not too small); the second term, $x f_a(x)$, introduces local oscillations whose

amplitude and position are modulated by the wavelet when a is small. Figure 2 shows this effect numerically: for a close to 1, φ'_a resembles the smooth derivative of GELU/Swish; for a close to 0, localized gradient spikes appear at the edges of the wavelet's support (near 1,1.5 and 2, in the case of db2), a behaviour absent from classical activations and potentially useful for tasks with local or multiscale input structure.

The cubic B-spline is of class C^2 , so F_B is piecewise C^3 . On the other hand,

$$g(x) = |\psi(x)|/\|\psi\|_1$$

inherits the regularity of ψ except at the zeros of ψ , where the absolute value introduces kinks, analogous to what happens in ReLU, but located inside the support rather than only at the origin. As a consequence, f_a and φ_a are smooth as $a \rightarrow 1$ and display a finite number of controlled kinks as a approaches 0. A variant that preserves full smoothness for any a consists of replacing $|\psi(x)|/\|\psi\|_1$ with $\psi(x)^2/\|\psi\|_2^2$, the wavelet's normalized energy density, which is C^∞ whenever ψ is; this variant is revisited later as a direction for future work.

Just as β in Swish (RAMACHANDRAN P, et al., 2017) or the per-channel parameter of PReLU can be learned by backpropagation, the mixture parameter a of the 4Rs-Gate is also differentiable:

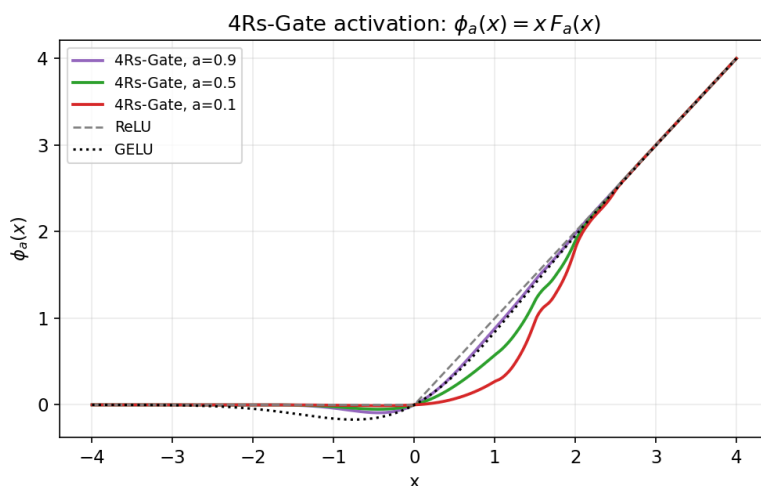
$$\frac{\partial \varphi_a(x)}{\partial a} = x (F_B(x) - F_g(x)). \quad (9)$$

Since F_B and F_g can be precomputed (the former analytically, the latter via table), this gradient costs $O(1)$ per evaluation, making it practical to train a , globally, per layer or per neuron, jointly with the network's other parameters, in the same spirit as training the β parameter in Swish.

NUMERICAL ILLUSTRATION

To illustrate the construction, we use $B(x)$ as the centered cubic B-spline (support $[-2,2]$) and $\psi(x)$ as the Daubechies order-2 mother wavelet (db2, support $[0,3]$), generated numerically via a cascade algorithm, exactly as in Linhares RR and Vargas RN (2026). Figure 1 shows the activation $\varphi_a(x) = x \cdot F_a(x)$ for three values of a , compared to ReLU and GELU.

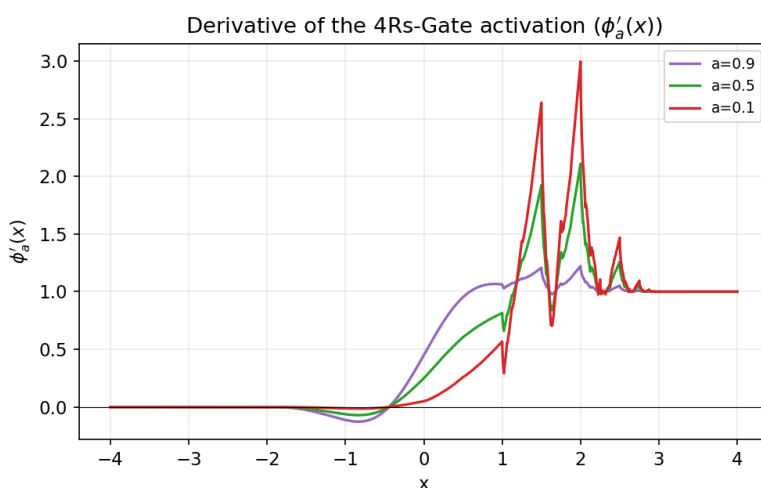
Figure 1 - 4Rs-Gate activation for $a = 0.9/0.5/0.1$, compared to ReLU and GELU. Note the asymmetry inherited from db2's $[0,3]$ support and the gradual transition to a multimodal regime as a decreases.



Source: Prepared by the authors.

Figure 2 shows the derivative $\phi'_a(x)$, highlighting the localized gradient spikes that emerge as the wavelet component's weight increases (a decreases).

Figure 2 - Derivative $\phi'_a(x)$. For small a , multiple gradient spikes appear, associated with the edges of db2's support, absent from classical activations.



Source: Prepared by the authors.

Finally, the gate $F_a(x)$ itself illustrates how the parameter a shifts and deforms the transition between 0 and 1: besides smoothing or localizing the transition, the asymmetry between the supports of B and ψ shifts the half-transition point, an extra degree of freedom relative to GELU/Swish, whose gates are fixed and symmetric about the origin.

COMPUTATIONAL SIMULATION: EMPIRICAL VALIDATION

The previous sections proposed and analysed the 4Rs-Gate from a mathematical standpoint. In this section we take a first empirical step, training real neural networks with the proposed activation and comparing it to ReLU, GELU, Swish (SiLU) and Tanh, the kind of validation this work still lacked. All code was implemented with a from-scratch, NumPy-based MLP trainer (Adam optimizer, manual backpropagation): F_B is computed in closed form (differentiable analytically) and F_g is tabulated once on a fine grid (8000 points) from the db2 wavelet, with differentiable linear interpolation; we verified the correctness of the gradients by finite-difference gradient checking, both with respect to the input x and to the parameter a , with full agreement.

Experimental tasks:

Regression: $y(x) = \sin(2\pi x) + 3 \cdot \exp(-50(x - 0.3)^2) \cdot \sin(40x) + \text{Gaussian noise}$ ($\sigma = 0.05$), a smooth, low-frequency trend plus a localized high-frequency bump, purposefully designed to favor activations able to represent local structure, testing this work's central hypothesis.

Classification: two moons (scikit-learn), 800 training / 400 test points, noise 0.25, a non-linear decision boundary, but spatially homogeneous (no localized multiscale structure), used as a control task.

For each task and each activation, we trained an MLP (regression: 3 hidden layers of 32 units; classification: 2 hidden layers of 16 units), Adam optimizer, 400 (regression) or 300 (classification) epochs, with 5 independent random seeds. We tested three configurations of the 4Rs-Gate: $a = 0.9$ fixed (B-spline-dominant, nearly smooth), $a = 0.3$ fixed (wavelet-dominant, more localized), and a trainable by backpropagation (initialized at 0.5).

Table 1 summarizes the test mean squared error (MSE). The 4Rs-Gate with $a = 0.9$ achieved by far the lowest mean error, about five times lower than the second-best activation, with a gap that does not overlap within ± 1 standard deviation of GELU, Swish or Tanh:

Table 1 - Test MSE by activation function (regression task, mean $\pm$ std. dev. over 5 seeds).

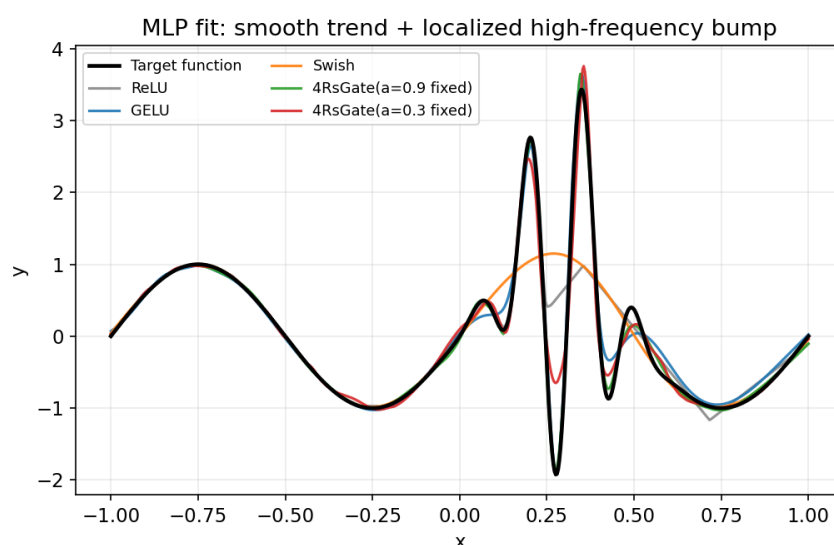
Activation	Mean MSE	Std. dev.	Rank
4RsGate ($a = 0.9$ fixed)	0.0100	0.0035	1st
GELU	0.0516	0.0279	2nd

Activation	Mean MSE	Std. dev.	Rank
4RsGate ($a = 0.3$ fixed)	0.0582	0.0496	3rd
ReLU	0.0636	0.0824	4th
4RsGate (a trainable)	0.0702	0.0673	5th
Tanh	0.1533	0.1141	6th
Swish	0.3404	0.0536	7th

Source: Prepared by the authors.

Figure 3 illustrates the fit (one representative seed): the 4Rs-Gate with $a = 0.9$ tracks the sharp localized bump almost exactly, while ReLU, Swish and GELU largely flatten it out, reproducing only the smooth low-frequency trend.

Figure 3 – MLP fit to $y(x) = \sin(2\pi x) + 3e^{-50(x-0.3)^2}\sin(40x)$: target function (black) versus models trained with different activations.



Source: Prepared by the authors.

On the classification task, all activations, including the three 4Rs-Gate variants, achieved statistically indistinguishable test accuracy, between 93.9% and 94.3%, with standard deviations on the order of one percentage point. This null result is expected and informative: the two-moons task has no localized multiscale structure, so the 4Rs-Gate's distinguishing

mechanism (the wavelet-inherited multiscale gate) should not, and indeed did not, bring any measurable advantage or disadvantage.

Three observations emerge from this first empirical validation: Confirmation of the central hypothesis, on a task with local structure at multiple scales (the high-frequency bump over the smooth trend), the 4Rs-Gate with $a = 0.9$ clearly outperformed every classical activation tested, roughly a five-fold reduction in mean test error relative to GELU, the strongest classical baseline. Even a modest wavelet contribution (a is still close to 1) already introduces a localization mechanism that classical activations lack. A wavelet-dominant configuration is suboptimal, but not catastrophic, the $a = 0.3$ (wavelet-dominant) configuration placed third overall, close behind GELU, rather than collapsing in performance. It never matched $a = 0.9$, though, which is consistent with the smoothness argument: the kinks introduced by $|\psi(x)|$ make the loss surface less regular as the wavelet component grows, which can slow optimization even when it does not derail it. The trainable parameter barely moved from its initialization, across all seeds, a trainable, initialized at 0.5, converged to values within about 0.50 ± 0.01 in both tasks, essentially unchanged. Combined with its middling regression performance (5th of 7), this indicates that the gradient signal for a single scalar mixture parameter is comparatively weak next to the gradients on the network's weights, so a does not reliably find its way to the favorable $a \approx 0.9$ region within the number of epochs used here. A separate, larger learning rate for a , a more informed initialization, or more training epochs are natural next steps. Neutrality on tasks without localized structure, in the absence of multiscale structure in the data (two moons), the 4Rs-Gate neither penalizes nor benefits performance, behaving like the classical activations, evidence that adding the wavelet component introduces no systematic cost when it is not needed, regardless of a .

Taken together, the results support a practical, immediate recommendation: use the 4Rs-Gate with a initialized close to 1 (and, if a is kept trainable, seeded near this value rather than at 0.5), treating the wavelet component as a smooth perturbation over a B-spline base. Whether a trained from a more favorable initialization, or a smooth ψ^2 variant, can make lower values of a competitive as well remains an open question for future work.

CONCLUSION

We showed that the 4Rs Family, originally proposed as a parametric class of probability densities combining cubic B-splines and normalized wavelet moduli, can be reinterpreted as the generator of a new family of self-gated activation functions, the 4Rs-Gate: $\varphi_a(x) = x \cdot F_a(x)$, with F_a the CDF of the mixture. This construction generalizes the principle underlying GELU and Swish, activation as the product of the input and the CDF of a fixed density, by replacing the fixed density with a parametric, multiscale family, controlled by a single interpretable parameter $a \in [0,1]$ and by the choice of mother wavelet. We derived closed-form expressions for the B-spline component, verified the correctness of the gradients with respect to both the input and the parameter a by gradient checking, and discussed smoothness, ReLU-like asymptotic behaviour, and direct connections to B-spline-based architectures such as KAN.

A first empirical validation, with MLPs trained on synthetic regression and classification tasks, gave clear support to the central hypothesis: on a task with local structure at multiple scales, the 4Rs-Gate with $a = 0.9$ substantially outperformed ReLU, Swish, Tanh and GELU; on a task without that structure, its performance was statistically equivalent to the classical activations. The experiments also revealed two effects not anticipated by the purely theoretical analysis: a wavelet-dominant configuration (low a) underperforms the near-smooth configuration without being unstable, and a trainable a , left to its own gradient signal, tends to stay close to its initialization rather than migrate toward the favorable region. This suggests that, for now, a should be set (or carefully initialized) close to 1 rather than learned from a neutral starting point. As future work, we highlight evaluation on larger architectures and real-world benchmarks (including KAN and Transformers), testing the smooth ψ^2 variant, extending the construction to multidimensional wavelets, and investigating better initialization or optimization strategies for a as a trainable, per-layer or per-neuron parameter.

REFERENCES

- HENDRYCKS D, GIMPEL K. Gaussian Error Linear Units (GELUs). arXiv:1606.08415, 2016.
- HERRERA O, PRIEGO B. Wavelets as activation functions in Neural Networks. Journal of Intelligent & Fuzzy Systems, 2022; 42(5): 4345-4355.

JAGTAP AD, KARNIADAKIS GE. How important are activation functions in regression and classification? A survey, performance comparison, and future directions. arXiv:2209.02681, 2022.

LINHARES RR, VARGAS RN. The 4Rs Family: a parametric class of probability densities based on convex mixtures of B-splines and wavelet modules. Cadernos Cajuína, 2026; 11(9): e3447.

LIU Z, et al. KAN: Kolmogorov-Arnold Networks. arXiv:2404.19756, 2024.

MISRA D. Mish: A Self Regularized Non-Monotonic Activation Function. arXiv:1908.08681, 2019.

MURARI K. A Family of Adaptive Activation Functions for Mitigating Failure Modes in Physics-Informed Neural Networks. arXiv:2603.18328, 2026.

RAMACHANDRAN P, et al. Searching for Activation Functions. arXiv:1710.05941, 2017.

VECCI L, et al. Learning and approximation capabilities of adaptive spline activation function neural networks. Neural Networks, 1998; 11(2): 259-270.

ZAPRANIS A, ALEXANDRIDIS A. Wavelet neural networks: A practical guide. Neural Networks, 2013; 42: 1-27.

ZHANG Q, BENVENISTE A. Wavelet networks. IEEE Transactions on Neural Networks, 1992; 3(6): 889-898.