Layered Neural Networks with GELU Activation, a Statistical Mechanics Analysis

Frederieke Richert[1], Michiel Straat[2], Elisa Oostwal[1] and Michael Biehl[1]

1- University of Groningen - Intelligent Systems, Nijenborgh 9, 9747 AG Groningen - The Netherlands
2- Bielefeld University - Center for Cognitive Interaction Technology, Inspiration 1, 33619 Bielefeld - Germany
f.richert@rug.nl

Abstract

Understanding the influence of activation functions on the learning behaviour of neural networks is of great practical interest. The GELU*, being similar to* swish *and* ReLU*, is analysed for soft committee machines in the statistical physics framework of off-line learning. We find phase transitions with respect to the relative training set size, which are always continuous. This result rules out the hypothesis that convexity is necessary for continuous phase transitions. Moreover, we show that even a small contribution of a sigmoidal function like* erf *in combination with* GELU *leads to a discontinuous transition.*

1. Introduction

The GELU activation function [2] is similar to the popular swish [1] and ReLU. Recent work [5] shows that ReLU soft committee machines (SCM) display a continuous phase transition, while SCMs with the sigmoidal erf show a discontinuous transition in the learning curves.

We negate the hypothesis that convexity of the ReLU causes the continuous transition by investigating the nature of the phase transition caused by the non-convex GELU. Furthermore, we construct a hybrid activation function called the ErfGELU.

$$\mathrm{GELU}(x,\gamma) := \frac{x}{2}\left(1+\mathrm{erf}\left[\gamma x/\sqrt{2}\right]\right) \tag{1}$$

$$\mathrm{ErfGELU}(x,\gamma,\delta) := (1-\delta)\,\mathrm{GELU}(x,\gamma) + \delta\,\mathrm{erf}\left[\gamma x/\sqrt{2}\right] \tag{2}$$

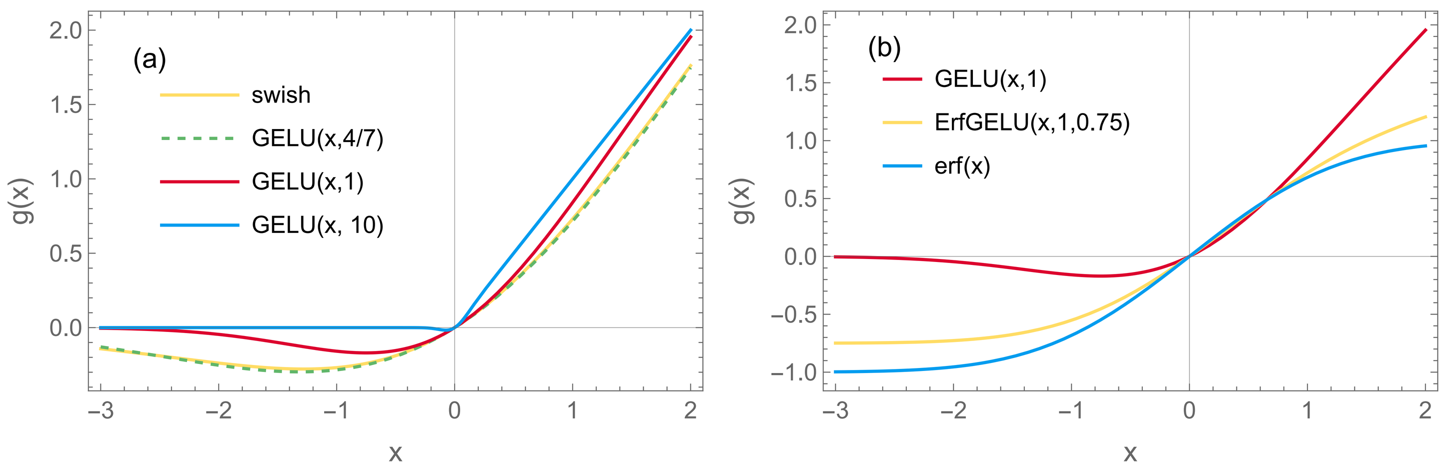
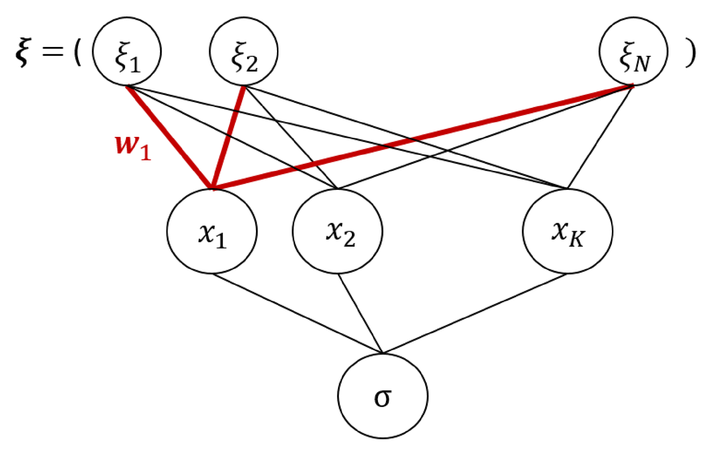


Figure 1: *Part (a) shows the GELU activation function for different γ as compared to the swish. In part (b) the ErfGELU is depicted for different values of δ. For $\delta = 0$ it is the GELU and for $\delta = 1$ the erf is recovered.*

2. Model

The SCMs are analysed in a student-teacher scenario with a trainable student network learning from a matched teacher network representing the task. Given the input vector $\boldsymbol{\xi} \in \mathbb{R}^N$ and the activation function g, the output of the student network σ and the pre-activations $\{x_k\}_{k=1}^K$ are:



$$\sigma(\boldsymbol{\xi}) := \frac{1}{\sqrt{K}} \sum_{k=1}^{K} g(x_k),$$

$$x_k := \frac{\mathbf{w}_k \cdot \boldsymbol{\xi}}{\sqrt{N}}.$$

Accordingly, the output of the teacher network is $\tau(\boldsymbol{\xi}) := \frac{1}{\sqrt{K}} \sum_{m=1}^{K} g(x_m^*)$ with the pre-activation $x_m^* := \mathbf{w}_m^* \cdot \boldsymbol{\xi}/\sqrt{N}$.

In the limit of high input dimension, $N \to \infty$, a suitable off-line training result can be expressed by a Boltzmann-distribution in student weight space. In the high temperature limit $\beta \to 0$, it is dominated by the minima of the free energy, $\beta f = \alpha K \epsilon_g - s$, with $\alpha = \beta P/(KN)$, the entropy s and the generalisation error defined as [3, 4, 5, 6]:

$$\epsilon_g := \left\langle \frac{1}{2K}\left[\sum_{k=1}^{K} g(x_k) - \sum_{m=1}^{K} g(x_m^*)\right]^2 \right\rangle_{\{\boldsymbol{x},\boldsymbol{x}^*\}}. \tag{3}$$

For $N \to \infty$, ϵ_g becomes an average over the pre-activations, which are Gaussian random variables with zero mean and covariances, called order parameters, [3, 4, 5, 6]:

$$Q_{ik} := \langle x_i x_k \rangle = \frac{\boldsymbol{w}_i \cdot \boldsymbol{w}_k}{N}, \qquad R_{in} := \langle x_i x_n^* \rangle = \frac{\boldsymbol{w}_i \cdot \boldsymbol{w}_n^*}{N}. \tag{4}$$

The site-symmetric ansatz [5, 6]:

$$Q_{ik} = \begin{cases} 1, & i = k \\ C, & i \neq k \end{cases} \qquad R_{in} = \begin{cases} R, & i = n \\ S, & i \neq n \end{cases} \tag{5}$$

allows for specialisation of each student vector to one specific teacher vector, where $R > S$, or anti-specialised solutions with $R < S$.

3. Results

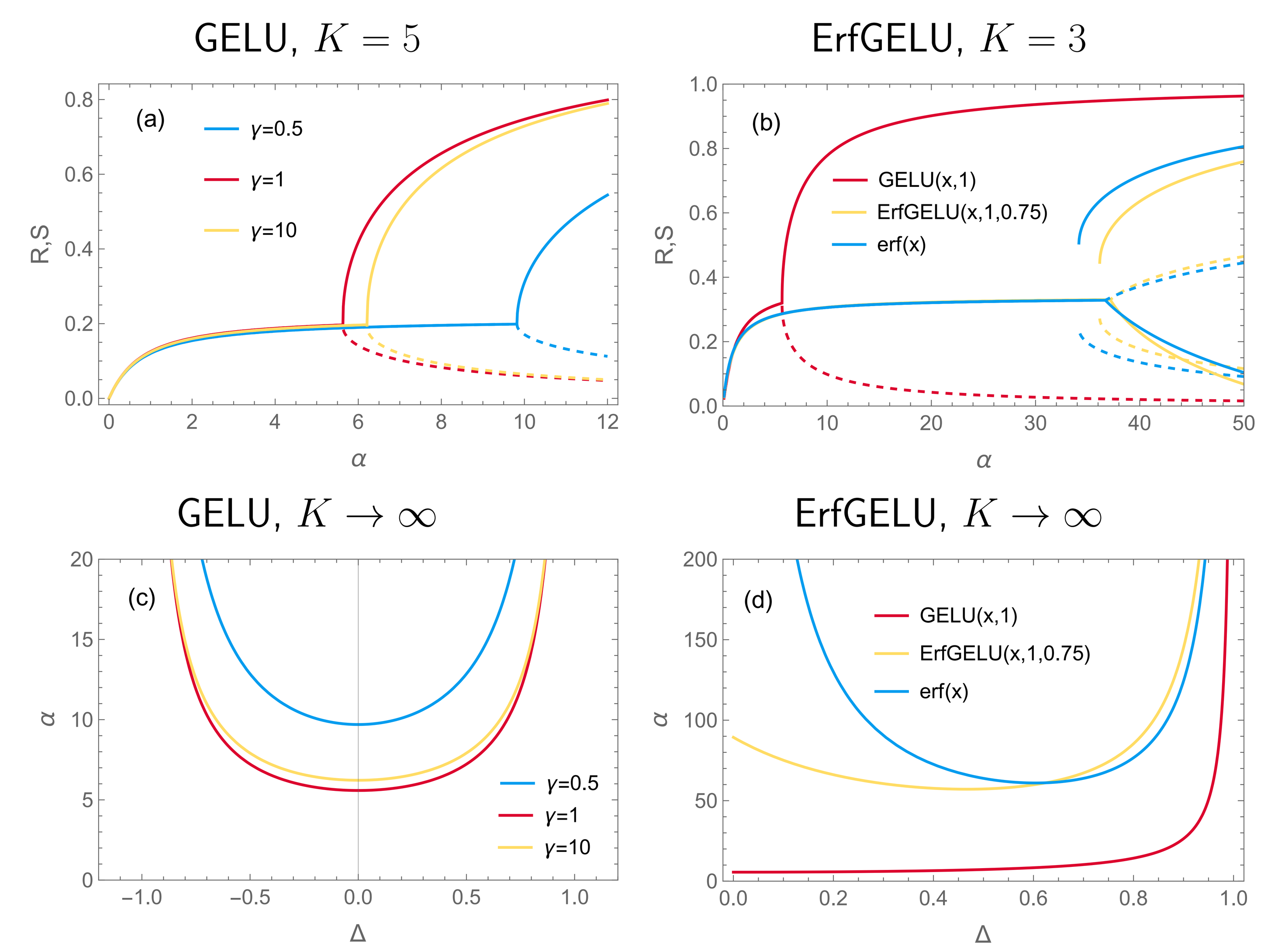


Figure 2: *In (a) and (b) the order parameters R (solid) and S (dashed) reveal the different types of phase transitions. In (a) the network with GELU activation shows a continuous transition for all γ. In contrast, for the ErfGELU, (b), we find a continuous transition only for $\delta = 0$ (GELU). For $\delta > 0$ the transition is discontinuous.*
In (c) and (d) the limit $K \to \infty$ is assumed. Both figures show $\alpha(\Delta)$ with $\Delta = R - S$. For the GELU activation function, (c), α is the solution to $\partial f(\alpha,\gamma,\Delta)/\partial\Delta = 0$. In the ErfGELU case, (d), there is also a dependence on δ. The minimum of $\alpha(\Delta)$ is the smallest possible α minimising the free energy f and the value of Δ at the minimum indicates the type of phase transition: $\Delta_{min} = 0$-continuous transition and $\Delta_{min} > 0$-discontinuous transition.
This shows again that the phase transition for the GELU is continuous (c) and if the activation function also contains a small contribution of the erf, the transition is discontinuous (d).

4. Conclusion

- Using the $\mathbf{GELU}$ activation function leads to a **continuous phase transition** in the SCM.
- **Convexity** of the activation function is **not the distinguishing feature** for a continuous transition.
- A small contribution of the sigmoidal erf to the GELU is sufficient to cause a discontinuous transition.

References

[1] P. Ramachandran, B. Zoph and Q.V. Le, Searching for activation functions, *6th international conference on learning representations*(ICLR2018), 2018.

[2] D. Hendrycks, K. Gimpel, Gaussian Error Linear Units (GELUs), *arXiv e-prints*, 2016.

[3] A.Engel, C. Van den Broeck, *Statistical mechanics of learning*, Cambridge University Press, 2001.

[4] H.S. Seung, H. Sompolinsky and N. Tishby, Statistical mechanics of learning from examples, *Physical Review A*, 45, 8, 1992

[5] E. Oostwal, M. Straat and M. Biehl, Hidden unit specialisation in layered neural networks: ReLU vs. sigmoidal activation, *Physica A*, Vol. 564, 125517, 2021.

[6] M. Ahr, M. Biehl and R. Urbanczik, Statistical physics and practical training of soft-committee machines, *Eur. Phys. J. B*, 10583-588, 1999.

ESANN 2023, 31th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, 4-6 October 2023, Bruges, Belgium