Layered Neural Networks with GELU Activation, a Statistical Mechanics Analysis

Frederike Richert¹, Michiel Straat², Elisa Oostwal¹ and Michael Biehl¹

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

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.

\[ \text{GELU}(x, \gamma) := \frac{1}{2}(1 + \text{erf}(\gamma x / \sqrt{2})) \]  \hspace{1cm} (1)

\[ \text{ErfGELU}(x, \gamma, \delta) := (1 - \delta) \text{GELU}(x, \gamma) + \delta \text{erf}(\gamma x / \sqrt{2}) \] \hspace{1cm} (2)

Figure 1: Part (a) shows the GELU activation function for different \( \gamma \) as compared to the swish. In part (b) the ErfGELU is depicted for different values of \( \delta \). 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 \( \mathbf{x} \in \mathbb{R}^n \) and the activation function \( g \), the output of the student network \( \mathbf{y} \) and the pre-activations \( x_k, i \leq K \) are:

\[ \mathbf{y} = (y_1, \ldots, y_K) \]
\[ \mathbf{x} = (x_1, \ldots, x_K) \]
\[ \sigma(\mathbf{x}) = \left( \frac{1}{\sqrt{K}} \sum_{i=1}^{K} g(x_i) \right) \]
\[ x_i = w_{oi} \cdot \xi + \mathbf{w}_i \cdot \mathbf{x} + n \]

Accordingly, the output of the teacher network is \( \mathbf{\tau}(\mathbf{\xi}) = \frac{1}{\sqrt{K}} \sum_{i=1}^{K} g(x_i) \) with the pre-activation \( x_i = \frac{1}{\sqrt{K}} \sum_{i=1}^{K} g(x_i) \) with the pre-activation. \( \mathbf{x} = \mathbf{w}_i \cdot \mathbf{x} + n \), where \( n \) is the noise.

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 \( T \to 0 \), it is dominated by the minima of the free energy, \( \beta f = \alpha K \epsilon_s - s \), with \( \alpha = \beta P/(KN) \), the entropy \( s \) and the generalisation error defined as [3, 4, 5, 6]:

\[ \epsilon_s = \frac{1}{\sqrt{K}} \sum_{i=1}^{K} g(x_i) - \frac{1}{\sqrt{K}} \sum_{i=1}^{K} g(x_i)/|x_i| \] \hspace{1cm} (3)

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

\[ Q_{ik} := \langle x_i x_k \rangle = \frac{w_{oi} \cdot w_{ok} + R_{ik}}{N}, \quad R_{ik} := \langle x_i x_k \rangle = \frac{w_{oi} \cdot w_{ok}}{N} \] \hspace{1cm} (4)

The site-symmetric ansatz [5, 6]:

\[ Q_{ik} = \begin{cases} 1, & i = k \in C, \ i \neq k \in C \break R_{ij}, & i \neq j \end{cases} \]

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](image)

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 \( \gamma \). 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 \( \epsilon_s(\Delta) \) with \( R = S \). For the GELU activation function, \( \epsilon_s \) is the solution to \( \beta f(\epsilon_s, \Delta) = 0 \). In the ErfGELU case, (d), there is also a dependence on \( \delta \). The minimum of \( \epsilon_s(\Delta) \) is the smallest possible \( \alpha \) minimising the free energy \( f \) and the value of \( \Delta \) at the minimum indicates the type of phase transition: \( \Delta_{\text{conv}} = 0 \)-continuous transition and \( \Delta_{\text{dis}} > 0 \)-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 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