Introduction

The GELU [1] activation function is similar to the popular swish [2] and ReLU. Recent work shows [3] that ReLU soft committee machines (SCM) display a continuous phase transition, while SCMs with the sigmofid erf show a discontinuous transition in the learning curves.

Our result: GELU-SCM \(\rightarrow\) continuous transition

This rules out the hypothesis that the convexity of the ReLU is causing the continuous transition, since the GELU is non-convex and shows the same transition.

\[
GELU(x,y) = \frac{x}{2} \left( 1 + \text{erf} \left( \frac{yx}{\sqrt{2}} \right) \right)
\]

Model

The GELU SCM is analysed in a student-teacher scenario with a trainable student network learning from a matched teacher network representing the task. The output of the student \(\sigma\) and the teacher \(\tau\) with activation function \(g\) are [3]:

\[
\sigma(x) = \frac{1}{\sqrt{K}} \sum_{k=1}^{K} g(x_k), \quad \tau(x) = \frac{1}{\sqrt{K}} \sum_{m=1}^{K} g(x^*_m)
\]

With the dependence on the P-many i.i.d. random input vectors \(\xi\) (with zero-mean, unit-variance components) via the pre-activations [1]:

\[
x_k = w_k \cdot \xi / \sqrt{N}, \quad x^*_m = w^*_m \cdot \xi / \sqrt{N}
\]

The weight \(w_k \in \mathbb{R}^N\) is student weight vector of \(k\)th hidden unit

The weight \(w^*_m \in \mathbb{R}^N\) is teacher weight vector of \(m\)th hidden unit

In the limit of high input dimension, \(N \rightarrow \infty\), a suitable off-line training result can be expressed by a Boltzmann-distribution in student weight space. In the high temperature limit \(\beta = 0\), it is dominated by the minima of the free energy, \(\beta f = aKg - s\), with \(a = \beta P/(KN)\) and \(s\) the activation function independent entropy [3,4,5,6].

\[
g_{\gamma} = \left( \frac{1}{2K} \sum_{k=1}^{K} g(x_k) - \sum_{m=1}^{K} g(x^*_m) \right)^2
\]

For \(N \rightarrow \infty\), the generalisation error \(g_{\gamma}\) becomes an average over the pre-activations, which are Gaussian random variables with zero mean and covariances (\(=\) order parameters) [3,4,5,6]:

\[
R_{\gamma} = \langle x_1x^*_1 \rangle = w_1 \cdot w^*_1 / N \quad Q_{\gamma} = \langle x_1x_2 \rangle = w_1 \cdot w_2 / N
\]

Site-symmetric ansatz:

\[
R_{\gamma} = \delta_{\gamma} R + (1 - \delta_{\gamma}) S \quad Q_{\gamma} = \delta_{\gamma} + (1 - \delta_{\gamma}) C
\]

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

Results

\[\text{Influence of hidden layer size}\]

\[\text{Influence of slope parameter}\]

\[\text{Critical point for different slopes}\]

\[\text{Influence of slope parameter}\]

Conclusion

- GELU in SCM causes a continuous phase transition, independent of the size of the hidden layer \(K\) and the slope parameter \(\gamma\), consistent with similarity of GELU and ReLU.
- Convexity of activation function \(\neq\) cause of continuous phase transition.

References