ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

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

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 success of artificial neural networks in recent years is partly attributed to the use of specific activation functions, such as the prominent ReLU, see e.g. [1]. Here, we study the use of the *Gaussian Error Linear Unit* (GELU) [1, 2] activation:

$$\mathrm{GELU}(x,\gamma) := \tfrac{x}{2}\Big(1+\mathrm{erf}\Big[\gamma x/\sqrt{2}\Big]\Big). \tag{1}$$

The GELU is of practical importance as it displays the same properties as the often used swish activation [1]. In fact, for appropriate choices of its *slope parameter*, see [1], swish and GELU are virtually indistinguishable. Furthermore, for large γ the GELU is a smooth approximation of the popular ReLU activation function, with the approximation becoming exact for $\gamma \to \infty$, see Fig. 1(a).
The statistical mechanics analysis of learning is a well-established approach to describe the typical behaviour of learning systems [3, 4]. Here, we employ this framework to investigate two-layered soft committee machines (SCM) in model scenarios of supervised learning. In a recent analysis [5], it was shown that an SCM with ReLU activation functions displays second-order, continuous phase transitions in the generalisation error as a function of the relative data set size. In contrast, the same architecture equipped with sigmoidal erf activations shows first-order, discontinuous phase transitions in the learning curves [5, 6]. This is a result of practical relevance, because in the presence of a first-order transition, suboptimal states coexist and compete with well performing configurations of the network. This can hinder the success of training significantly as the system may get stuck in these unfavourable states.

*This work is part of the RAFFLES project, funded by the Dutch Research Council (NWO) in the *Open Competitie ENW-M* programme (project number OCENW.M20.287).
M.S. gratefully acknowledges funding by the MKW NRW for the project SAIL, NW21-059A.

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

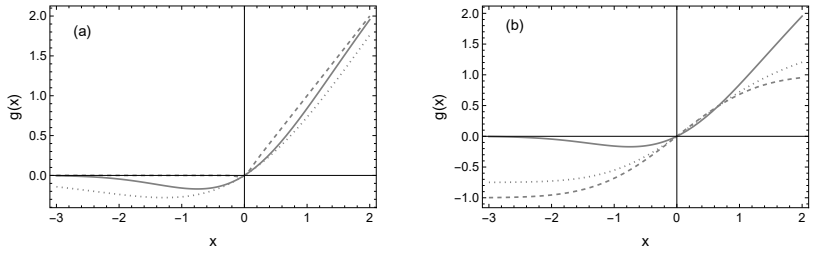

Fig. 1: (a) The GELU for $\gamma = 1$ (solid) and $\gamma = 10$ (dashed), and the swish activation function (dotted) with slope parameter set to 1. (b) The GELU$(x, 1)$ (solid), ErfGELU$(x, 1, 0.75)$ (dotted), cf. Eq. (4), and erf$[x/\sqrt{2}]$ (dashed).

In the following, we investigate the learning behaviour of SCM networks with GELU-activation along the same lines, and show that they exhibit continuous transitions like in the limiting case of the ReLU. Our results indicate that the nature of the phase transition does not depend on the convexity of the activation function, since both the sigmoidal erf and the GELU are non-convex, but exhibit different types of transitions. In addition, we consider activations that can be expressed as a superposition of GELU and erf. We find that even a small contribution of the sigmoidal function leads to a discontinuous transition.

2 Model Setup and Analysis

We briefly summarise the modelling framework and sketch the statistical physics based analysis. For a more detailed presentation we refer to [5]. The learning behaviour of an SCM with GELU activation function is analysed in a student-teacher setting with a fixed teacher network representing the task to be learned by a student network of perfectly matching complexity. Only the first layer weights of the student network are adaptable in the training process. The output of the student network $\sigma(\boldsymbol{\xi})$ and the teacher output $\tau(\boldsymbol{\xi})$ are defined as

$$\sigma(\boldsymbol{\xi}) := \frac{1}{\sqrt{K}} \sum_{k=1}^{K} g\left(\frac{\mathbf{w}_k \cdot \boldsymbol{\xi}}{\sqrt{N}}\right), \qquad \tau(\boldsymbol{\xi}) := \frac{1}{\sqrt{K}} \sum_{m=1}^{K} g\left(\frac{\mathbf{w}_m^* \cdot \boldsymbol{\xi}}{\sqrt{N}}\right), \tag{2}$$

with $g : \mathbb{R} \to \mathbb{R}$ being the GELU activation function of a given slope γ. The weights of both networks are collected in weight vectors, where $\mathbf{w}_k \in \mathbb{R}^N$ contains the weights from the input layer to the k-th hidden unit in the student and $\mathbf{w}_m^* \in \mathbb{R}^N$ is the respective weight vector for the teacher's m-th hidden unit. We consider systems with normalised student weight vectors and orthonormal teacher vectors: $\mathbf{w}_m^* \cdot \mathbf{w}_n^* = \delta_{mn}$.

In the equilibrium statistical mechanics approach to off-line learning the assumption is that the student network is provided with a set of training data of size P, $\{\boldsymbol{\xi}^\mu, \tau(\boldsymbol{\xi}^\mu)\}_{\mu=1}^{P}$. Training is guided by the minimisation of a cost function $E = \sum_{\mu=1}^{P} [\sigma(\boldsymbol{\xi}^\mu) - \tau(\boldsymbol{\xi}^\mu)]^2 / 2$. In the simplest setting, components of the $\boldsymbol{\xi}^\mu$ are assumed to be i.i.d random numbers with zero mean and unit variance [3, 4]. An appropriate training process will eventually lead to an equilibrium state in the weight space of the network which can be described by a Gibbs-Boltzmann

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

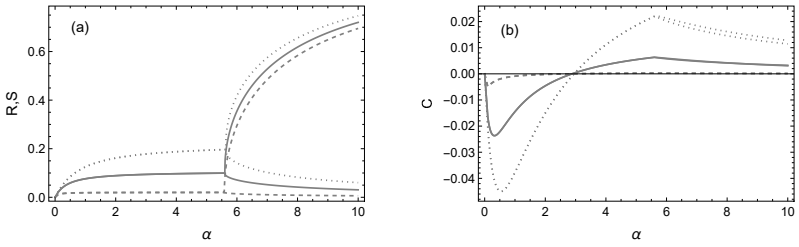

Fig. 2: Learning curves for the GELU with $\gamma = 1$ and different values of the hidden layer size, $K = 5$ (dotted), $K = 10$ (solid) and $K = 50$ (dashed). (a) The order parameters R and S separate after the transition point of the continuous transition with R being the upper curve and S the lower curve for each respective case. (b) The order parameter C also differs in its scaling with K before and after the phase transition at $\alpha_c \approx 5.58$.

density of states proportional to $\exp(-\beta E)$ [3, 4], where β plays the formal role of an inverse temperature. Equilibrium configurations of large networks in the limit $N \to \infty$ and averaged over randomised data sets of a given size, display typical properties of learning systems in terms of their *learning curves*. Here we resort to the simplifying limit of training at high temperature [3, 4, 5]. For $\beta \to 0$, the Gibbs-Boltzmann density is dominated by the minima of the *free energy* $\beta f = \alpha K \epsilon_g - s$, which can be evaluated for a given relative data set size $\alpha = \beta P/(KN)$. The *entropy* term s can be worked out independently of model details such as the activation functions, see e.g. [5]. They do however determine the functional form of the generalisation error ϵ_g, which is defined as the mean squared deviation of student and teacher output, averaged over the data distribution:

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

The respective network output σ, τ only depends on the inputs $\boldsymbol{\xi}$ via the *pre-activations*: $x_k := \mathbf{w}_k \cdot \boldsymbol{\xi}/\sqrt{N}$ and $x_m^* := \mathbf{w}_m^* \cdot \boldsymbol{\xi}/\sqrt{N}$. In the limit $N \to \infty$, x_k and x_n^* become zero mean Gaussian variables according to the central limit theorem [5]. Hence, ϵ_g is obtained as a multi-dimensional Gaussian integral and depends only on the covariances of the pre-activations. These are called order parameters in statistical mechanics jargon and are defined as $Q_{ik} := \langle x_i x_k \rangle = \boldsymbol{w}_i \cdot \boldsymbol{w}_k/N$ and $R_{in} := \langle x_i x_n^* \rangle = \boldsymbol{w}_i \cdot \boldsymbol{w}_n^*/N$. In addition, we have $\langle x_m^* x_n^* \rangle = \boldsymbol{w}_m^* \cdot \boldsymbol{w}_n^*/N = \delta_{mn}$. The order parameters are also referred to as *overlaps* between the weight vectors due to their expression in terms of scalar products.

We consider a simplifying site-symmetric ansatz [5, 6] with $Q_{ik} = \delta_{ik} + (1 - \delta_{ik})C$ and $R_{in} = \delta_{in}R + (1 - \delta_{in})S$. This restriction allows for unspecialised configurations with $R = S$, where all student hidden units perform essentially the same task. The ansatz also admits the case in which the normalised student weight vectors specialise with respect to exactly one of the teacher weight vectors ($R > S$) and perfect agreement with $\epsilon_g = 0$ can be achieved.

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

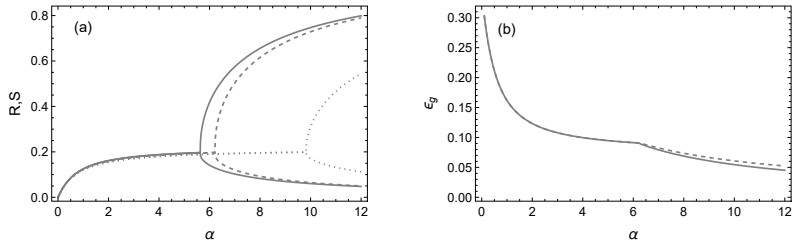

Fig. 3: Learning curves showing the continuous specialisation transition for the GELU for $K = 5$. (a) The order parameters R and S for different values of the slope parameter $\gamma = 0.5$ (dotted), $\gamma = 1$ (solid) and $\gamma = 10$ (dashed). (b) The generalisation error for $\gamma = 10$ after the transition is lower in the specialised case with $R > S$ (solid) than in the anti-specialised case $R < S$ (dashed).

3 Results

Learning for different values of K and γ

In Fig. 2 the learning curves for $\gamma = 1$ and different values of K show that the order parameters R and S are equal for small α. At the critical training set size α_c, the R-values increase while the S-values decrease. In the figure only the specialised solution with $R > S$ is depicted, but a local minimum of f with $R < S$ also exists. The specialised solution can be interpreted as each student weight vector approaching a specific weight vector of the teacher, while the overlap with the remaining teacher vectors becomes small.

For different values of K it can be seen that the plateau value of the order parameters before the transition point decreases as $1/K$, while the relation after the transition is $\Delta = 1 - KS$ with $\Delta = R - S$ [6]. Fig. 2 also shows the order parameter C, which is the overlap between different student weight vectors, for various K. Again, one can observe a strong dependence of the values of C on K, while the point of phase transition remains approximately the same. The value of C scales as $1/(K-1)$ for small α, i.e. in the unspecialised regime and it scales as $1/K^2$ in the specialised state [6].

The dependence of the learning curves on the slope parameter γ for fixed K can be seen in Fig. 3. For small γ the transition happens at large values of α, but for $\gamma \to \infty$ the transition point is $\alpha_c = 2\pi$, the point of the phase transition for the ReLU activation function [5].

Learning for $K \to \infty$

It is instructive to consider the limit where the hidden layer becomes infinitely large, $K \to \infty$ (but with $K \ll N$), since in this limit, the nature of the phase transition becomes particularly clear. Moreover, it allows extending the analysis to learning at finite temperatures as demonstrated in [6]. For $K \to \infty$ and for specialised solutions discussed in the previous section, the ansatz $R = \Delta$, $S = (1 - \Delta)/K$ and $C = 0$ can be justified [6]. The only remaining order parameter is then Δ, so minimising the free energy reduces to finding solutions of

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

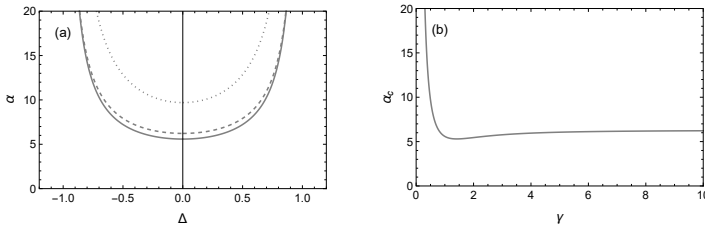

Fig. 4: $K \to \infty$-analysis of the GELU. (a) $\alpha(\gamma, \Delta)$ for $\gamma = 0.5$ (dotted), $\gamma = 1$ (solid) and $\gamma = 10$ (dashed). (b) The value of α where the continuous phase transition happens, α_c, as a function of the slope parameter γ.

$\partial f(\alpha, \gamma, \Delta)/\partial \Delta = 0$. This is not easily solvable for Δ, but straightforward to solve for α, resulting in functions $\alpha(\gamma, \Delta)$, which can be seen for some values of γ in Fig. 4(a). The minimal value of α for which there exists a solution to $\partial f(\alpha, \gamma, \Delta)/\partial \Delta = 0$, i.e. the minimal value of $\alpha(\gamma, \Delta)$, is α_c, at which the phase transition occurs. In case of the GELU activation function α_c is found at $\Delta = 0$ for all $\gamma > 0$. This means that the separation between R and S at the phase transition is zero, i.e. it is a continuous transition for all $\gamma > 0$. The dependence of the transition point α_c on γ is depicted in Fig. 4(b). For very small γ, α_c becomes very large due to the fact that for $\gamma \to 0$ the GELU becomes linear and does not invoke a phase transition. There is an optimal value with the smallest α_c at $\gamma = \sqrt{2}$ and for $\gamma \to \infty$ we recover the value $\alpha_c = 2\pi$ for the ReLU activation function [5].

ErfGELU

In search of a possible explanation for the different learning behaviour exhibited in the off-line learning of SCMs for the different activation functions we propose a convex combination of the GELU and the erf as a hybrid function:

$$\mathrm{ErfGELU}(x, \gamma, \delta) := (1 - \delta)\ \tfrac{x}{2}\left(1 + \mathrm{erf}\left[\gamma x/\sqrt{2}\right]\right) + \delta\ \mathrm{erf}\left[\gamma x/\sqrt{2}\right]\ . \qquad (4)$$

For $\delta = 0$ it is the GELU and for $\delta = 1$ the erf is recovered, while values of $\delta \in (0, 1)$ express an activation function which is a mixture of the two, cf. Fig. 1(b). In Fig. 5(a) the learning curves for different δ at a hidden layer size $K = 3$ show that for $\delta = 0$ we indeed recover the GELU learning curve with a continuous transition. However, for larger values of δ the transition is discontinuous, because a jump in the order parameters is observed, which also depends on δ. The same δ-dependence can be shown for general K including $K \to \infty$. Performing the same analysis as described in the previous section for the ErfGELU leads to the graphs $\alpha(\Delta, \gamma, \delta)$ that can be seen in Fig. 5(b). Again, we see that the minimum of α with respect to Δ depends on δ. For $\delta = 0$ it is at $\Delta = 0$, then shifts to positive values of Δ for larger δ and for $\delta = 1$ we recover the erf value. The derivative of α with respect to Δ can be computed analytically and its values in $\Delta = 0$ are given by $\left.\frac{\partial \alpha(\Delta, \gamma, \delta)}{\partial \Delta}\right|_{\Delta=0} = -\frac{4\pi\gamma^2\left(\gamma^2+1\right)^3\delta^2}{(\gamma^2+2)^4(\delta-1)^4}$.

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

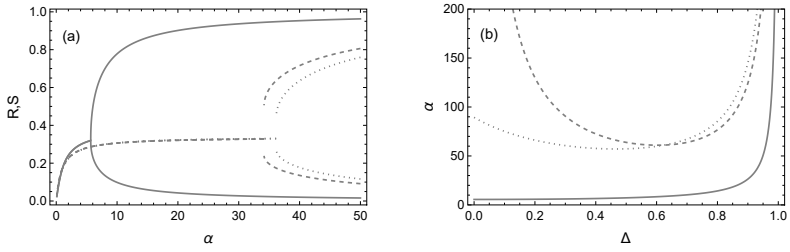

Fig. 5: Specialisation transition for ErfGELU with $\gamma = 1$ and $\delta = 0$, i.e. GELU (solid), $\delta = 0.75$ (dotted) and $\delta = 1$, i.e. erf (dashed). (a) In the learning curves for $K = 3$ after the transition point the upper branch is the R order parameter and the lower branch is the S order parameter. (b) In the $K \to \infty$ case the function $\alpha(\Delta, \gamma, \delta)$ shows minima at different Δ for different δ.

Assuming non-zero γ, this can only be zero for $\delta = 0$, i.e. the only continuous transition occurs for $\delta = 0$, for which the activation function is the GELU.

4 Conclusion

We have shown that the use of the GELU activation function leads to a continuous phase transition in the SCM learning scenario, independent of the size of the hidden layer and the slope parameter γ. This is consistent with the pronounced similarity of GELU and ReLU. One hypothesis that our results rule out is that the convexity of the activation function is necessary for a continuous phase transition, which appears plausible when considering the ReLU in comparison with classical sigmoidal activations. As we have shown, the non-convex GELU also causes a continuous transition. In addition, we studied a superposition of GELU and erf and find that a small contribution of the sigmoidal erf is sufficient to cause a discontinuous transition. Future work will address scenarios in which student and teacher architecture are mismatched with different hidden layer sizes and/or activation functions in student and teacher. This will require the extension of the analysis to training at low temperatures in the annealed approximation or the replica method [3, 4, 6].

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.