university of groningen

Off-line Learning Analysis for Soft Committee Machines with GELU Activation

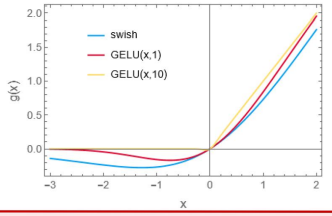
Frederieke Richert, Michiel Straat, Elisa Oostwal, Michael Biehl

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 sigmoidal erf show a discontinuous transition in the learning curves.

Our result: GELU-SCM → 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.

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



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 σ and the teacher τ with activation function g are [3]:

$$\sigma(\xi) \coloneqq \frac{1}{\sqrt{K}}\sum_{k=1}^{K} g(x_k) \qquad \tau(\xi) \coloneqq \frac{1}{\sqrt{K}}\sum_{m=1}^{K} g(x_m^*)$$

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

$$x_k \coloneqq w_k \cdot \xi/\sqrt{N} \qquad x_m^* \coloneqq w_m^* \cdot \xi/\sqrt{N}$$

$w_k \in \mathbb{R}^N$ - student weight vector of *k*-th hidden unit
$w_m^* \in \mathbb{R}^N$ - teacher weight vector of *m*-th hidden unit

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 $\alpha = \beta P/(KN)$ and s the activation function independent entropy [3,4,5,6].

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

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

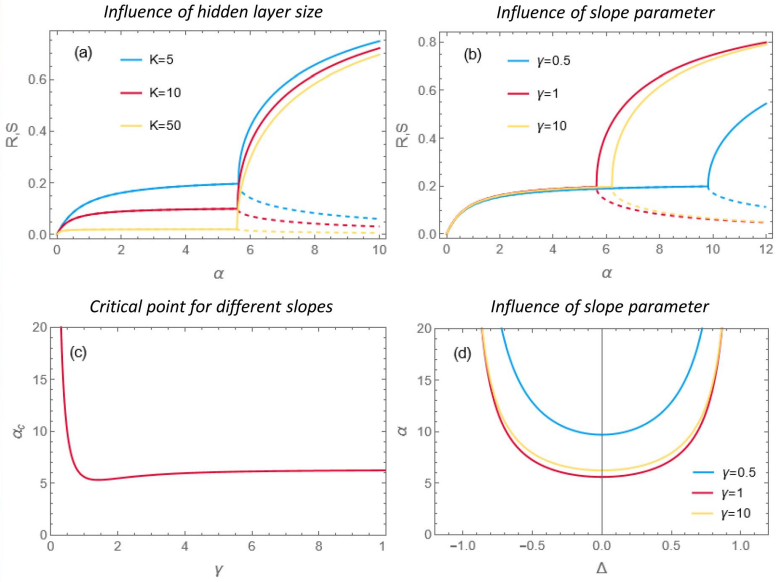
$$R_{in} \coloneqq \langle x_i x_n^* \rangle = w_i \cdot w_n^*/N \qquad Q_{ik} \coloneqq \langle x_i x_k \rangle = w_i \cdot w_k/N$$

Site-symmetric ansatz:

$$R_{in} = \delta_{in} R + (1-\delta_{in}) S \qquad Q_{ik} = \delta_{ik} + (1-\delta_{ik}) C$$

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

Results



R,S-learning curves with continuous transitions for

a) $\gamma = 1$: approximately same transition point α_c; but R,S-value scaling with K

b) K=5 : α_c shifting with γ

For $K \to \infty$

c) Transition points $\alpha_c(\gamma)$: minimum at $\gamma = \sqrt{2}$

d) $\alpha(\Delta)$ with specialisation $\Delta := R - S$: The Δ-value of the minimum of $\alpha(\Delta)$ indicates the type of phase transition. Here, $\Delta_{\min} = 0$ for all $\gamma > 0$, which specifies a continuous transition.

For $K \to \infty$, the scaling ansatz $R = \Delta$, $S = (1-\Delta)/K$ and $C = 0$ for the specialised solution is used [4]. The minimum of the free energy is then given by $\alpha(\Delta)$ for which $f'_{\alpha(\Delta)}(\Delta) = 0$.

Conclusion

- GELU in SCM causes a ***continuous phase transition***, independent of the size of the hidden layer K and the slope parameter γ. (consistent with similarity of GELU and ReLU)
- convexity of activation function ≠ cause of continuous phase transition

References

[1] D. Hendrycks, K. Gimpel, Gaussian Error Linear Units (GELUs), arXiv e-prints, 2016.
[2] P. Ramachandran, B. Zoph and Q.V. Le, Searching for Activation Functions, 6th international conference on learning representations(ICLR2018), 2018.
[3] E. Oostwal, M. Straat and M. Biehl, Hidden unit specialisation in layered neural networks: ReLU vs. sigmoidal activation, Physica A, Vol. 564, 125517, 2021.
[4] M. Ahr, M. Biehl and R. Urbanczik, Statistical physics and practical training of soft committee machines, Eur. Phys. J. B, 10583-588, 1999.
[5] H.S. Seung, H. Sompolinsky and N. Tishby, Statistical mechanics of learning from examples, Physical Review A, 45, 8, 1992
[6] A.Engel, C. Van den Broeck, Statistical Mechanics of Learning, Cambridge University Press, 2001.