ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

On-line Learning Dynamics in Layered Neural Networks with Arbitrary Activation Functions*

Otavio Citton[1,2], Frederieke Richert[1] and Michael Biehl[1] †

1 - Intelligent Systems - Bernoulli Institute, University of Groningen
Nijenborgh 9, 9747 AG Groningen - The Netherlands

2 - Groningen Cognitive Systems and Materials Center (CogniGron),
University of Groningen, Groningen, The Netherlands

Abstract. We revisit and extend the statistical physics based analysis of layered neural networks trained by online gradient descent. We focus on the influence of the hidden unit activation functions on the typical learning behavior in model scenarios. Expanding activation functions in terms of Hermite polynomials enables us to extend the formalism to the analysis of soft committee machines with arbitrary activation in student-teacher scenarios. This approach requires much lower computational effort than naive numerical integration, which is practically infeasible. Moreover, it now becomes possible to treat mismatched scenarios in which the student activation function differs from the one used in the target rule definition. This makes it possible to study realistic models of machine learning.

1 Introduction

The choice of activation function is an important element of specifying a neural network architecture. Hence, knowing the influence this function has on the learning behavior of the network is of practical relevance. We aim at gaining insights into the impact of activation functions in layered neural networks by using methods from the statistical mechanics theory of learning. For on-line learning, i.e. stochastic gradient descent, where an update step is made after the presentation of only one example at a time, the works of Biehl and Schwarze [1] and Saad and Solla [2] derived ordinary differential equations (ODE) describing the learning dynamics of soft committee machines with sigmoidal (erf) activation function, see also [3] for recent extensions. The on-line dynamics for the popular ReLU activation were analysed in [4] along the same lines.

We extend these previous works significantly by presenting a method for studying the learning behavior of soft committee machines with arbitrary activation functions. This is achieved by expanding the activation function in terms of Hermite polynomials. All relevant quantities can be expressed in this formalism and we show that practically feasible truncations of the resulting series expansion achieve sufficient precision.

In the following, we introduce the theoretical framework, describe the differential equations for on-line gradient descent, and outline their expression in terms of Hermite polynomials. In Sec. 3 we show results comparing the previous

*The code used for the analysis can be found in our GitHub repository.

†This work is funded by NWO M1 grant OCENW.M20.287 and CogniGron.

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

analytical treatment for erf, ReLU and GELU activations with the corresponding Hermite series expansion. Furthermore, we present learning curves obtained with the new formalism for settings which cannot be analysed exactly. In particular, we treat cases of mismatched activation in student and target-defining teacher networks. To the best of our knowledge, this is the first statistical physics analysis of such settings. Previous studies were limited to possible mismatch in the size of the hidden layer, but with the same activations in student and teacher.

2 Methods

The network architecture studied here is the so-called Soft Committee Machine (SCM) [2]. It is defined as a two–layer neural network where only the input–to–hidden weights are adjusted during training. All hidden–to–output weights are considered to be constant and equal to *one*. We denote by $\mathbf{w}_i \in \mathbb{R}^d$ the weight vector connecting the input to the i-th hidden neuron and $\mathbf{w} = \{\mathbf{w}_i\}_{i=1}^K$ the set of all learnable parameters. The output of the network for a given input $\boldsymbol{\xi} \in \mathbb{R}^d$ is

$$\sigma(\boldsymbol{\xi}, \mathbf{w}) = \textstyle\sum_{i=1}^K g(x_i), \qquad x_i = \mathbf{w}_i \cdot \boldsymbol{\xi},$$

where g is a nonlinear activation function.

The on-line learning training framework encompasses the presentation of a novel, independent individual example of the form $(\boldsymbol{\xi}^\mu, \tau^\mu)$ at each time step, where $\tau^\mu = \tau(\boldsymbol{\xi}^\mu) \in \mathbb{R}$ is the label of the input $\boldsymbol{\xi}^\mu$. We consider so-called *student-teacher scenarios*, where we parameterize the rule that generates the target labels by a set of M weight vectors $\mathbf{w}^* = \{\mathbf{w}_n^*\}_{n=1}^M$, $\mathbf{w}_n^* \in \mathbb{R}^d$ that can be interpreted as a teacher network with output $\tau(\boldsymbol{\xi}) = \sum_{n=1}^M g(y_n)$ and $y_n = \mathbf{w}_n^* \cdot \boldsymbol{\xi}$.

Training and evaluation of the student performance are based on an error measure that corresponds to the quadratic deviation of the student output from the target. The generalization error

$$\epsilon_g(\mathbf{w}) = \big\langle \epsilon(\boldsymbol{\xi}, \mathbf{w}) \big\rangle_{\boldsymbol{\xi}} \quad \text{with error measure} \quad \epsilon(\boldsymbol{\xi}, \mathbf{w}) = \tfrac{1}{2}\big[\sigma(\boldsymbol{\xi}, \mathbf{w}) - \tau\big]^2$$

is defined as the expected error over the input distribution. Note that the generalization error only depends on the input vector through $x_i = \mathbf{w}_i \cdot \boldsymbol{\xi}$ and $y_n = \mathbf{w}_n^* \cdot \boldsymbol{\xi}$, and, for examples with i.i.d. components with zero mean, the Central Limit Theorem (CLT) implies that, in the limit $d \to \infty$, all quantities $\{x_i, y_n\}$ will be normally distributed with covariance matrix $\mathbf{C}$,

$$\mathbf{C} = \begin{pmatrix} \mathbf{Q} & \mathbf{R} \\ \mathbf{R}^\top & \mathbf{T} \end{pmatrix}, \text{ with } Q_{ik} = \mathbf{w}_i \cdot \mathbf{w}_k,\ R_{in} = \mathbf{w}_i \cdot \mathbf{w}_n^* \text{ and } T_{nm} = \mathbf{w}_n^* \cdot \mathbf{w}_m^*.$$

The R_{in} and Q_{ik} play the role of *order parameters* in the sense that they describe macroscopic properties of the student network, while the T_{nm} are fixed model parameters which specify the teacher network configuration.

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

2.1 Stochastic Gradient Descent and Differential Equation

We consider a stochastic gradient descent rule as the learning algorithm for the SCM, where the student weights are updated at time step μ as:

$$\mathbf{w}_i^{\mu+1} = \mathbf{w}_i^{\mu} - \tfrac{\eta}{d}\nabla_{\mathbf{w}_i}\, \epsilon(\boldsymbol{\xi}^{\mu}, \mathbf{w}^{\mu}) = \mathbf{w}_i^{\mu} + \tfrac{\eta}{d}\delta_i^{\mu}\boldsymbol{\xi}^{\mu},$$

where $\delta_i^{\mu} = g'(x_i^{\mu})\Big(\sum_n g(y_n^{\mu}) - \sum_k g(x_k^{\mu})\Big)$ and η is the learning rate. We assume that at each time step, a *novel* data example is presented to the learning system.

Taking the dot product of the above with $\mathbf{w}_n^*$ and $\mathbf{w}_k^{\mu+1}$, yields, respectively

$$\tfrac{R_{in}^{\mu+1} - R_{in}^{\mu}}{1/d} = \eta\delta_i^{\mu} y_n^{\mu}, \qquad \tfrac{Q_{ik}^{\mu+1} - Q_{ik}^{\mu}}{1/d} = \eta\big(\delta_i^{\mu} x_k^{\mu} + \delta_k^{\mu} x_i^{\mu}\big) + \eta^2 \delta_i^{\mu}\delta_k^{\mu}$$

and, by defining the normalized example number $\bar{\alpha} = \mu/d$ and taking the limit $d \to \infty$ the l.h.s. become derivatives of R_{in} and Q_{ik} with respect to $\bar{\alpha}$. Using the CLT, we conclude that the r.h.s. are equal to their averages, which corresponds to the *self-averaging* property of the order parameters:

$$\tfrac{\mathrm{d}R_{in}}{\mathrm{d}\bar{\alpha}} = \eta\big\langle \delta_i^{\mu} y_n^{\mu}\big\rangle, \qquad \tfrac{\mathrm{d}Q_{ik}}{\mathrm{d}\bar{\alpha}} = \eta\big\langle \delta_i^{\mu} x_k^{\mu} + \delta_k^{\mu} x_i^{\mu}\big\rangle + \mathcal{O}(\eta^2). \tag{1}$$

Here we neglect terms of order η^2 and derive results valid for the regime of small learning rates. This also allows us to rescale the example number with the learning rate as $\alpha = \bar{\alpha}\eta$. According to [2], the ODE can then be written as

$$\tfrac{\mathrm{d}R_{in}}{\mathrm{d}\alpha} = \sum_{m=1}^{M} I_3(\mathbf{C}_{i,K+n,K+m}) - \sum_{j=1}^{K} I_3(\mathbf{C}_{i,K+n,j})$$

$$\tfrac{\mathrm{d}Q_{ik}}{\mathrm{d}\alpha} = \sum_{m=1}^{M}\Big[I_3(\mathbf{C}_{i,k,K+m}) + I_3(\mathbf{C}_{k,i,K+m})\Big] - \sum_{j=1}^{K}\Big[I_3(\mathbf{C}_{i,k,j}) + I_3(\mathbf{C}_{k,i,j})\Big],$$

where I_3 are averages defined as $I_3(\mathbf{A}) = \int g'(z_1)\, z_2\, g(z_3)\, P(\mathbf{z}|\mathbf{A})\, \mathrm{d}z_1\mathrm{d}z_2\mathrm{d}z_3$ with $\mathbf{z} = \begin{pmatrix} z_1 & z_2 & z_3\end{pmatrix}^{\top}$ and $P(\mathbf{z}|\mathbf{A})$ a three-dimensional Gaussian distribution with a general covariance matrix $\mathbf{A}$ and zero mean. $\mathbf{C}_{a,b,c}$ represents a 3×3 correlation matrix obtained from $\mathbf{C}$ by selecting the rows and columns corresponding to the elements a, b and c.

2.2 Hermite Polynomial Representation

For a 3×3 correlation matrix $\boldsymbol{\Sigma}$ with elements $\Sigma_{ij} = \delta_{i,j} + \rho_{ij}(1 - \delta_{i,j})$, the Kibble-Slepian formula [5, 6] (a generalization of Mehler's kernel [7] for higher dimensions) allows us to represent the Gaussian distribution as a product of an uncorrelated Gaussian and a series:

$$P(\mathbf{z}|\boldsymbol{\Sigma}) = P(\mathbf{z}|\mathbf{I}) \sum_{a=0}^{\infty}\sum_{b=0}^{\infty}\sum_{c=0}^{\infty} \tfrac{\rho_{12}^a}{a!}\tfrac{\rho_{13}^b}{b!}\tfrac{\rho_{23}^c}{c!} H_{a+b}(z_1) H_{a+c}(z_2) H_{b+c}(z_3), \tag{2}$$

where $\mathbf{I}$ is the identity matrix and H_n is the n-th (probabilist's) Hermite polynomial. Substituting the above in the expression for I_3 yields

$$I_3(\boldsymbol{\Sigma}) = \sum_{a=0}^{\infty}\sum_{b=0}^{\infty}\sum_{c=0}^{\infty} \tfrac{\rho_{12}^a}{a!}\tfrac{\rho_{13}^b}{b!}\tfrac{\rho_{23}^c}{c!} \left\langle H_{a+b}, g'\right\rangle \left\langle H_{a+c}, H_1\right\rangle \left\langle H_{b+c}, g\right\rangle,$$

where $\langle\cdot,\cdot\rangle$ denotes the inner product $\langle f, g\rangle = \frac{1}{\sqrt{2\pi}} \int f(z) g(z) e^{-\frac{1}{2}z^2}\mathrm{d}z$.

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

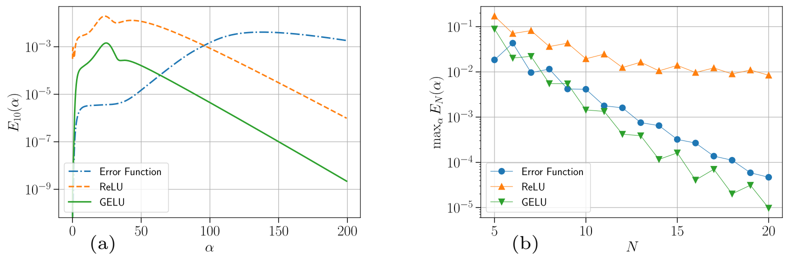

Figure 1: (a) The error $E_N(\alpha) = \|\mathbf{C}(\alpha) - \mathbf{C}_N(\alpha)\|_F \, / \, \|\mathbf{C}(\alpha)\|_F$ is shown for $N = 10$, where $\mathbf{C}_N(\alpha)$ is the covariance matrix using the Hermite approximation with N terms and $\mathbf{C}(\alpha)$ is the analytical covariance matrix, at example number α. (b) Compares the maximum of $E_N(\alpha)$ over α for different values of N.

Note that we can always obtain a correlation matrix with unitary diagonal from our covariance matrix $\mathbf{C}$, by taking $\rho_{ij} = C_{ij}/\sqrt{C_{ii}C_{jj}}$ and scaling the arguments of the functions $g'(z) \to g'(z\sqrt{C_{11}})$, $H_1(z) \to H_1(z\sqrt{C_{22}})$ and $g(z) \to g(z\sqrt{C_{33}})$. Furthermore, using the orthogonality of the polynomials with respect to the above defined inner product, $I_3(\mathbf{\Sigma})$ can be simplified to a single series

$$I_3(\mathbf{\Sigma}) = \sum_{n=0}^{\infty} \frac{\rho_{13}^n}{n!} \big(\rho_{12}\langle H_{n+1}, g'\rangle\langle H_n, g\rangle + \rho_{23}\langle H_n, g'\rangle\langle H_{n+1}, g\rangle\big), \qquad (3)$$

which, for non-pathological[1] activation functions g, can be approximated by truncating the series at sufficiently high order. The result obtained after integrating the ODE using the series with N terms will be denoted by $\mathbf{C}_N(\alpha)$. To calculate the generalization error we use a similar representation for the integrals, details will be published elsewhere.

3 Results and Discussion

We first provide evidence for the validity and usefulness of the series approximation by comparison with analytical solutions available for specific settings [2, 4]. Complementing the results of [8], we also derived the analytical form of I_3 for the GELU activation.

We use the Frobenius norm to quantify the error $E_N(\alpha)$ between the observed covariance matrices (see fig. 1). All results presented here correspond to settings with $K = M = 2$, a *graded teacher* [2] with $T_{nm} = n\,\delta_{n,m}$ and initial condition $Q_{ik}(0) = k10^{-1}\delta_{i,k}$ and $R_{in}(0) = 10^{-3}\delta_{i,n}$.

Furthermore, by using the Hermite polynomials method, we can also derive learning curves for mismatched cases, where the student and teacher network have a different activation function respectively.

Fig. 1 shows that a small number of terms N in the Hermite series expansion suffices to achieve small error between the approximation and the analytical

[1] Here, by "non-pathological" we mean functions $g \in L^2(\mathbb{R})$ whose inner product with Hermite polynomials of order $n > N$, $\langle H_n, g\rangle$, is small when compared to $\sqrt{n!}$. For popular activation functions in machine learning this is usually the case.

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

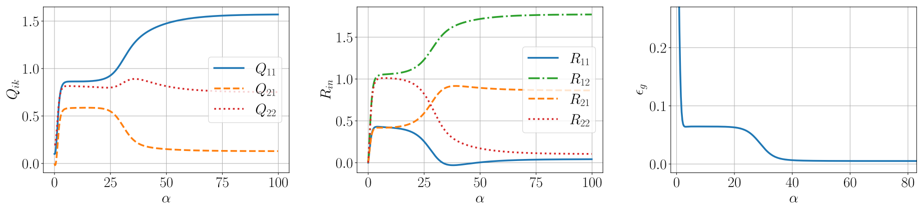

Figure 2: The learning curves (left to right: $Q_{ik}(\alpha), R_{in}(\alpha), \epsilon_g(\alpha)$) for the mismatched case with the swish function as teacher activation and GELU as the student activation function, obtained from the Hermite approximation with $N = 10$ terms.

expression. In fig. 1(a) the graph for the erf is stretched compared to the ones for ReLU and GELU, because the transition happens for larger α. We also notice that in fig. 1(b) the error is systematically larger for the ReLU, which we attribute to the sharp edge in the ReLU, requiring high-order polynomial terms to be closely approximated. Moreover, the GELU and ReLU curves decrease their error when N changes from an odd to an even number, whereas erf behaves vice versa. This can be explained by the definite parity of each Hermite polynomial, which produces larger or smaller contributions for every new term in the expansion depending on the "parity" of the activation function.

Fig. 2 shows the learning curves for a student network with GELU activation function learning from a graded teacher network with swish activation function. Since the GELU and swish are very similar functions the student learns relatively well from the teacher. However, we notice in Fig. 2 that, due to the mismatched activation, the student overlaps do not adjust perfectly to the ones of the teacher, i.e. $Q_{22} < T_{11}$ and $Q_{11} < T_{22}$, and the off-diagonal term Q_{12} converges to a small, but non-zero value.

In Fig. 3 we present the behavior of student networks with various activation functions learning from a teacher with ReLU activation function. We note that the ones with activation similar to the teacher learn the rule with small error, and the students with erf activation and Softplus plateau at a higher value of the generalization error.

4 Conclusion

In this work we introduce a novel way to represent differential equations for order parameters in on-line learning settings in terms of orthogonal polynomials. This new representation allows us to efficiently integrate the dynamics and obtain learning curves for SCMs with arbitrary activation functions. Most importantly, this includes cases of mismatch between student and teacher networks, which constitutes a significant novelty in the field.

One of the main advantages of the method introduced here is its computational efficiency when compared with standard numerical integration methods.

As can be seen from Fig. 1, very few terms in the series suffice to achieve very

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

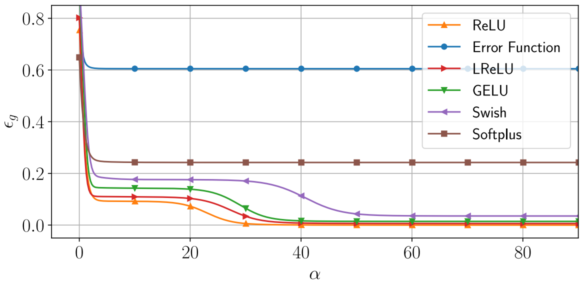

Figure 3: The generalization errors for students with different activation functions learning from a ReLU teacher exhibiting different behaviours. The results were obtained for the Hermite approximation using $N = 10$ terms.

small relative error. Another advantageous feature is that the method allows to interpret the results in terms of properties of the activation function.

In addition, our approach can be extended to include the terms of order η^2 that were ignored in the differential equations (1). These terms include 4-dimensional integrals similar to I_3, which can also be represented as a power series using an extension of Eq. (2) to four dimensions. However, in this case, the simplifications that lead us to Eq. (3) do not apply and the calculation of the six nested sums is computationally expensive.

In parallel, we are working on applying the Hermite polynomial representation for off-line learning, i.e. equilibrium analysis of batch learning processes.

References

[1] M. Biehl and H. Schwarze. Learning by on-line gradient descent. *Journal of Physics A: Mathematical and General*, 28(3):643, Feb 1995.

[2] D. Saad and S. A. Solla. On-line learning in soft committee machines. *Phys. Rev. E*, 52:4225–4243, Oct 1995.

[3] S. Goldt, M. Advani, A. M. Saxe, F. Krzakala, and L. Zdeborová. Dynamics of stochastic gradient descent for two-layer neural networks in the teacher-student setup. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[4] M. Straat and M. Biehl. On-line learning dynamics of ReLU neural networks using statistical physics techniques. In M. Verleysen, editor, *Proc. European Symposium on Artificial Neural Networks (ESANN)*, pages 517–522, 2019.

[5] W. F. Kibble. An extension of a theorem of Mehler's on Hermite polynomials. *Mathematical Proceedings of the Cambridge Philosophical Society*, 41(1):12–15, 1945.

[6] D. Slepian. On the Symmetrized Kronecker Power of a Matrix and Extensions of Mehler's Formula for Hermite Polynomials. *SIAM J. on Mathematical Analysis*, 3(4):606–616, 1972.

[7] F. G. Mehler. Ueber die Entwicklung einer Function von beliebig vielen Variablen nach Laplaceschen Functionen höherer Ordnung. *Journal für die reine und angewandte Mathematik*, 66:161–176, 1866.

[8] F. Richert, M. Straat, E. Oostwal, and M. Biehl. Layered Neural Networks with GELU Activation, a Statistical Mechanics Analysis. In M. Verleysen, editor, *Proc. European Symposium on Artificial Neural Networks (ESANN)*, pages 435–440, 2023.