ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence
and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0.
Available from http://www.i6doc.com/en/.

On-line learning dynamics of ReLU neural networks using statistical physics techniques

Michiel Straat[1] and Michael Biehl[1] *

1- Bernoulli Institute for Mathematics, Computer Science
and Artificial Intelligence, University of Groningen
Nijenborgh 9, 9747AG Groningen, The Netherlands

Abstract. We introduce exact macroscopic on-line learning dynamics of two-layer neural networks with ReLU units in the form of a system of differential equations, using techniques borrowed from statistical physics. For the first experiments, numerical solutions reveal similar behavior compared to sigmoidal activation researched in earlier work. In these experiments the theoretical results show good correspondence with simulations. In overrealizable and unrealizable learning scenarios, the learning behavior of ReLU networks shows distinctive characteristics compared to sigmoidal networks.

1 Introduction

Statistical physics techniques have been used successfully in the theoretical analysis of various machine learning models, including neural networks [1–3] and prototype-based models [3, 4]. In the context of neural networks, several learning scenarios have been studied, e.g., on-line gradient descent learning [1, 5–7], learning in non-stationary environments [3] and batch learning [8]. Macroscopic quantities, the so-called order parameters of the system, aggregate and summarize the usually large number of individual parameters of the machine learning model. In model situations, Central Limit Theorems (CLT) in combination with the consideration of the thermodynamic limit facilitate an exact description of the macroscopic dynamics in the form of a system of ordinary differential equations (ODE). It provides a useful tool to study the behavior of learning theoretically, in order to gain a deeper understanding of the learning process, which could potentially be used to improve algorithms used in practical scenarios. In the context of deep learning, Rectified Linear Unit (ReLU) activation has become popular mainly due to improved empirical performance compared to sigmoidal activation, e.g., see [9]. Here we formulate and study exact macroscopic gradient descent learning dynamics for the Soft Committee Machine (SCM), with the aim of increasing theoretical understanding of the behavior of ReLU activation in neural networks.

*We acknowledge financial support through the Northern Netherlands Region of Smart Factories (RoSF) consortium, see http://www.rosf.nl.

2 Macroscopic ReLU learning dynamics of the SCM

We consider regression where for an input $\boldsymbol{\xi} \in \mathbb{R}^N$ a teacher SCM with M hidden units computes the target output $\tau(\boldsymbol{\xi}) \in \mathbb{R}$ and a student SCM with K hidden units computes the hypothesis $\sigma(\boldsymbol{\xi}) \in \mathbb{R}$:

$$\tau(\boldsymbol{\xi}) = \sum_{n=1}^{M} g(y_n) \quad y_n = \boldsymbol{B}_n \cdot \boldsymbol{\xi}, \quad \sigma(\boldsymbol{\xi}) = \sum_{i=1}^{K} g(x_i) \quad x_i = \boldsymbol{J}_i \cdot \boldsymbol{\xi}\,. \tag{1}$$

Above, $\boldsymbol{B}_n \in \mathbb{R}^N$ and $y_n \in \mathbb{R}$ denote teacher weight vectors and pre-activations, respectively. In case of the student, those are denoted by $\boldsymbol{J}_i \in \mathbb{R}^N$ and $x_i \in \mathbb{R}$. We consider for the activation function $g(x)$: $\mathrm{ReLU}(x) = x\theta(x)$, where $\theta(x)$ is the unit step function. The student weights $\boldsymbol{J}$ are adaptable and we assume that the teacher weights $\boldsymbol{B}$ stay constant, i.e., the target rule remains fixed. In the on-line learning scenario at step μ, a new independent example $\boldsymbol{\xi}^\mu$ is presented from a stream. The direct error for $\boldsymbol{\xi}^\mu$ and the generalization error are defined as:

$$\epsilon(\boldsymbol{J}, \boldsymbol{\xi}^\mu) = \frac{1}{2}(\sigma^\mu - \tau^\mu)^2, \quad \epsilon_g(\boldsymbol{J}) = \langle \epsilon(\boldsymbol{J}, \boldsymbol{\xi}) \rangle_{\boldsymbol{\xi}}\,, \tag{2}$$

where $\langle \cdot \rangle_{\boldsymbol{\xi}}$ denotes averaging over the input distribution. One estimates the input distribution in practice, but here we consider i.i.d. Gaussian random components $\xi_i \sim \mathcal{N}(0,1)$.

For each presentation $\boldsymbol{\xi}^\mu$, the adaptation of the student weight vector $\boldsymbol{J}_i$ is guided by gradient descent on $\epsilon(\boldsymbol{J}^\mu, \boldsymbol{\xi}^\mu)$ with respect to $\boldsymbol{J}_i$, resulting in the update rule:

$$\boldsymbol{J}_i^{\mu+1} = \boldsymbol{J}_i^\mu + \frac{\eta}{N}\delta_i^\mu \boldsymbol{\xi}^\mu, \quad \delta_i^\mu = (\tau^\mu - \sigma^\mu) g'(x_i^\mu) \tag{3}$$

where η is the so-called learning rate which is scaled with the input dimension N. Note that from Equation (3), $g(x)$ should be differentiable. $\mathrm{ReLU}'(0)$ is undefined, but in practice one chooses a value for this rare case.

The choice of i.i.d. components ξ_i makes the CLT apply for large input dimension N. Hence, for large N, the pre-activations x_i and y_n become zero-mean Gaussian variables with properties:

$$\langle x_i x_j \rangle = \boldsymbol{J}_i \cdot \boldsymbol{J}_j = Q_{ij}, \;\; \langle x_i y_n \rangle = \boldsymbol{J}_i \cdot \boldsymbol{B}_n = R_{in}, \;\; \langle y_n y_m \rangle = \boldsymbol{B}_n \cdot \boldsymbol{B}_m = T_{nm}\,. \tag{4}$$

The variables R_{in}, Q_{ik} and T_{nm} are macroscopic variables of the system, so-called *order parameters*. Here we fix the rule properties to $T_{nm} = \delta_{nm}$. Combining the above equations with gradient update Equation (3) yields stochastic update equations for the order parameters directly. In the thermodynamic limit $N \to \infty$, the normalized time variable $\alpha = \mu/N$ can be considered continuous and the order parameters self-average as proved in [10]. Hence, averaging leads to a system of ODEs, e.g., shown in [2], describing exact macroscopic dynamics in the thermodynamic limit. For $g(x) = \mathrm{ReLU}(x)$, the system is:

$$\begin{aligned}
\frac{dR_{in}}{d\alpha} &= \eta \left[\sum_{m=1}^{M} \langle \theta(x_i) y_n y_m \theta(y_m) \rangle - \sum_{j=1}^{K} \langle \theta(x_i) y_n x_j \theta(x_j) \rangle \right] , \\
\frac{dQ_{ik}}{d\alpha} &= \eta \left[\sum_{m=1}^{M} \langle \theta(x_i) x_k y_m \theta(y_m) \rangle - \sum_{j=1}^{K} \langle \theta(x_i) x_k x_j \theta(x_j) \rangle \right] \\
&\quad + \eta \langle x_k \delta_i \rangle + \eta^2 \langle \delta_i \delta_k \rangle ,
\end{aligned} \tag{5}$$

where the term $\eta\langle x_k \delta_i \rangle$ in the second equation is the same as the first term for i and k interchanged. The averages of the form $\langle \theta(u) v w \theta(w) \rangle$ are taken with respect to the 3D joint Gaussian distribution $P(\boldsymbol{x}, \boldsymbol{\Sigma})$, for variable vector $\boldsymbol{x} = (u, v, w)^T$ and covariance matrix $\boldsymbol{\Sigma} = \langle \boldsymbol{x}\boldsymbol{x}^T \rangle$, which is populated with relevant variances and covariances from Equations (4). Integration yields the closed form expression:

$$\langle \theta(u) v w \theta(w) \rangle_{\boldsymbol{\xi}} = \frac{\sigma_{12} \sqrt{\sigma_{11}\sigma_{33} - \sigma_{13}^2}}{2\pi\sigma_{11}} + \frac{\sigma_{23} \sin^{-1}\left(\frac{\sigma_{13}}{\sqrt{\sigma_{11}\sigma_{33}}} \right)}{2\pi} + \frac{\sigma_{23}}{4} , \tag{6}$$

where σ_{ij} denotes the corresponding element of matrix $\boldsymbol{\Sigma}$. For general K and M, the term $\eta^2\langle \delta_i \delta_k \rangle$ consists of averages of the form $\langle w z \theta(u)\theta(v)\theta(w)\theta(z) \rangle_{\boldsymbol{\xi}}$. For now, we only include the η^2 term for $K = M = 1$. For general K and M, we study the dynamics for $\eta \to 0$, neglecting the η^2 term. Combining Equation (6) and (5) gives the closed form macroscopics of the ReLU SCM.

3 Experiments

In this section, we show and discuss for different settings the macroscopic online ReLU dynamics as obtained from the theoretical ODEs from Equation (5). Theoretical results are compared with simulations for sufficiently large N.

We first consider perceptron learning: $M = K = 1$. Initial conditions $(R_0, Q_0) = (0, 0.25)$ correspond to a random initialization of the student weights $\boldsymbol{J}$. For a learning rate $\eta = 0.1$, a numerical solution to the ODE system is shown in Figure 1. One observes an increase in both R and Q, indicating increasing similarity of the student to the rule and increasing weight magnitude. The state $(R, Q) = (1, 1)$ is the perfect solution that corresponds to equality of student and teacher, i.e., $\boldsymbol{J} = \boldsymbol{B}$. In fact, $(R, Q) = (1, 1)$ is a fixed point of the system for all meaningful η. Defining $(r, q) = (R - 1, Q - 1)$, a linearization of the dynamics is $(r', q')^T = \boldsymbol{A}(\eta)(r, q)^T$, where $\boldsymbol{A}(\eta)$ is the Jacobian in the fixed point. The eigenvalues of $\boldsymbol{A}(\eta)$ are given by $\boldsymbol{\lambda} = \{-\eta/2, 1/2\eta^2 - \eta\}$ with corresponding eigenvectors $\boldsymbol{u}_1 = (1/2, 1)^T$ and $\boldsymbol{u}_2 = (0, 1)^T$. As $\lambda_2 \geq 0$ for $\eta \geq 2$, it follows that for $\eta < 2$ the fixed point is asymptotically stable and we define this critical learning rate as $\eta_c = 2$. Figure 1 (right) shows the evolution of ϵ_g for several η. Convergence is slow for $\eta << \eta_c$ but also for $\eta \approx \eta_c$.

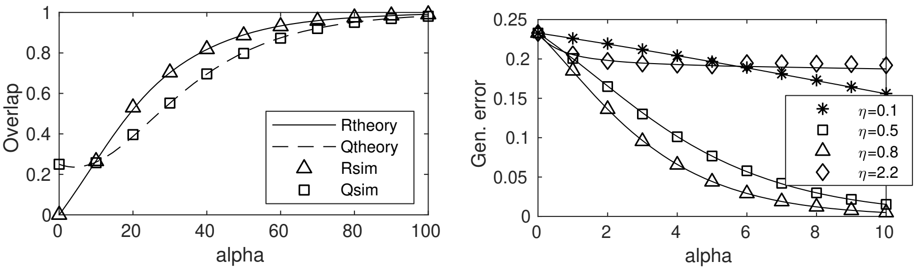

Fig. 1: *Left*: Evolution of order parameters. R and Q with $\eta = 0.1$, $R(0) = 0$ and $Q(0) = 0.25$. *Right*: Evolution of ϵ_g for different η. Note the scale of α. Lines and symbols show theoretical and simulation results, respectively. $N = 1000$ is used in the simulations.

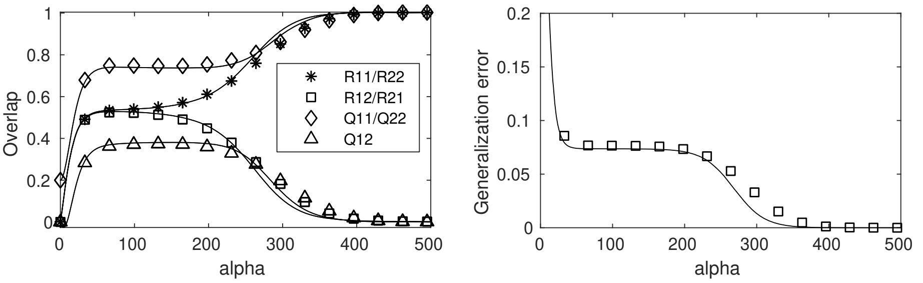

Fig. 2: *Left*: Evolution of order parameters for the case $K = M = 2$ and $\eta = 0.1$. *Right*: Evolution of the generalization error. Symbols show simulation results for $N = 10^4$.

Figure 2 shows dynamics for the ReLU network with $K = M = 2$. Initial conditions are $R_{in} = 10^{-3}\delta_{in}$ and $Q_{11} = 0.2$, $Q_{12} = 0$, $Q_{22} = 0.3$. The learning process is characterized by a suboptimal plateau in which $R_{in} \approx 0.52$ for all i, n, i.e., there is no specialization of students towards specific teachers. The symmetric plateaus are a property of learning in soft committee machines[1, 2] and they arise due to a repulsive fixed point of the system. An expression for the length of the plateau can be found in [11]. From the linearization of the ReLU dynamics, there is one positive eigenvalue that guides the escape: $\lambda_5 = 0.24$ with corresponding eigenvector $\mathbf{u}_5$=(0.5,-0.5,-0.5,0.5,0,0,0)$^{\mathrm{T}}$: It causes the observed specialization of each student towards one teacher. The onset of specialization is associated with a decrease in generalization error, see Figure 2 (right).

	$Q_{11}(\infty)$	$Q_{12}(\infty)$	$Q_{13}(\infty)$	$Q_{22}(\infty)$	$Q_{23}(\infty)$	$Q_{33}(\infty)$
ReLU	1.00	0.00	0.00	0.24	0.25	0.27
Erf	1.00	0.00	0.00	0.00	0.00	1.00

In Figure 3 and the table above, results for $K = 3$ and $M = 2$ are given

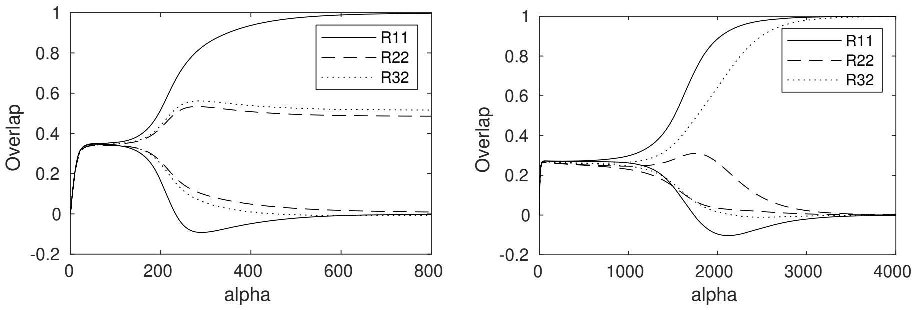

Fig. 3: Evolution of student-teacher overlap parameters for the case $K = 3$ and $M = 2$. *Left*: ReLU activation. *Right*: Erf activation. A pair of the same type of curves shows the correlation of one student unit to each of the two teacher units. The legends point to the upper curve of the pair.

for ReLU activation (left) and sigmoidal Erf activation (right). For the latter, closed form equations can be found in [1, 2]. Non-zero initial conditions are $R_{11} = 10^{-3}, Q_{11} = 0.2, Q_{22} = 0.3, Q_{33} = 0.25$. In both cases, $\boldsymbol{J}_1$ specializes to $\boldsymbol{B}_1$. In the ReLU case, $\boldsymbol{J}_2$ and $\boldsymbol{J}_3$ achieve a similar overlap with $\boldsymbol{B}_2$. From $Q_{22} \approx Q_{33} \approx Q_{23} \approx 0.25$ and $R_{22} \approx R_{32} \approx 0.5$, it follows that $\boldsymbol{J}_2 = \boldsymbol{J}_3 \parallel \boldsymbol{B}_2$ i.e., $\boldsymbol{J}_2 \approx a\boldsymbol{B}_2$ and $\boldsymbol{J}_3 \approx b\boldsymbol{B}_2$ for $a = b = 0.5$ and therefore $\boldsymbol{J}_2 + \boldsymbol{J}_3 = \boldsymbol{B}_2$. Hence, two units of the ReLU student learn both the same teacher unit apart from a scaling and there are in fact infinitely many solutions possible for different a and b, $a + b = 1$. The observed behavior is a consequence of the piece-wise linear property of the ReLU. Such combinations are not possible for the non-linear Erf: In this case, R_{22} decreases to zero due to $Q_{22}(\alpha \to \infty) = 0$, equivalent to $\boldsymbol{J}_2 = \boldsymbol{0}$, effectively removing the unit. In both cases, $\epsilon_g(\alpha \to \infty) = 0$ is achieved, since the rule is learned perfectly.

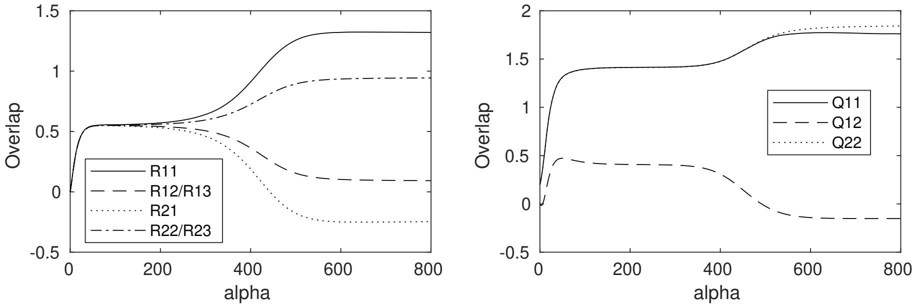

Fig. 4: Overlaps for the ReLU network with $K = 2$ and $M = 3$. *Left*: Evolution of student-teacher overlaps. *Right*: Evolution of student-student overlaps.

In Figure 4, results of the ReLU network for $K = 2$ and $M = 3$ are shown. Initial conditions are $R_{11} = 10^{-3}$, $R_{in} = 0$ for $i, j \neq 1$, $Q_{ii} = 0.2$ and $Q_{i \neq j} = 0$.

$\boldsymbol{J}_1$ mainly specializes to $\boldsymbol{B}_1$. As $R_{22} = R_{23} = 0.94$, it is mainly the case that $\boldsymbol{J}_2 \approx a\boldsymbol{B}_2 + b\boldsymbol{B}_3$ for $a \approx b$. Since the student does not realize the rule, $\epsilon_g(\alpha \to \infty) > 0$.

4 Discussion

We have formulated macroscopic learning dynamics of two-layer neural networks for ReLU activation. Simulation results for the perceptron and the network with two hidden units show good correspondence. For the perceptron, the optimal solution corresponds to a fixed point of the equations which becomes unstable at a critical learning rate. Sub-optimal plateaus appear in the networks that correspond to fixed points, of which the repulsion causes eventually specialization. For the overrealizable case, ReLU units are combined to deal with the extra complexity. The η^2 term that we omitted here should be included in future research to get exact equations for general η. This would also make possible the study of learning rate adaptation schemes within the framework.

References

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

[2] D. Saad and S. A. Solla. On-line learning in soft committee machines. *Phys. Rev. E*, 52:4225–4243, 10 1995.

[3] M. Straat, F. Abadi, C. Hammer, and M. Biehl. Statistical mechanics of on-line learning under concept drift. *Entropy*, 20(10):775, 10 2018.

[4] M. Biehl, A. Ghosh, and B. Hammer. Dynamics and generalization ability of LVQ algorithms. *Journal of Machine Learning Research*, 8:323–360, 2007.

[5] D. Saad and S. A. Solla. Exact solution for on-line learning in multilayer neural networks. *Phys. Rev. Lett.*, 74:4337–4340, May 1995.

[6] R. Vicente and N. Caticha. Functional optimization of online algorithms in multilayer neural networks. *Journal of Physics A: Mathematical and General*, 30(17):L599, 1997.

[7] M. Inoue, H. Park, and M. Okada. On-line learning theory of soft committee machines with correlated hidden units-steepest gradient descent and natural gradient descent. *Journal of the Physical Society of Japan*, 72(4):805–810, 2003.

[8] M. Biehl, E. Schlösser, and M. Ahr. Phase transitions in soft-committee machines. *EPL (Europhysics Letters)*, 44(2):261, 1998.

[9] X. Glorot, A. Bordes, and Y. Bengio. Deep sparse rectifier neural networks. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 315–323. PMLR, 11–13 Apr 2011.

[10] G. Reents and R. Urbanczik. Self-averaging and on-line learning. *Phys. Rev. Lett.*, 80:5445–5448, Jun 1998.

[11] M. Biehl, P. Riegler, and C. Wöhler. Transient dynamics of on-line learning in two-layered neural networks. *Journal of Physics A: Mathematical and General*, 29(16):4769, 1996.