Towards a statistical physics analysis of multilayer ReLU neural networks

Michiel Straat, Michael Biehl

September 12th, 2019

university of groningen
Content

1. On-line learning in a student-teacher scenario
2. Differential equations in the thermodynamic limit
3. Adaptive second layer weights
4. Future work
Towards a statistical physics analysis of multilayer ReLU neural networks

Learning from a teacher network

At timestep $\mu$, the input $\xi^\mu \in \mathbb{R}^N$ is presented.

$$\tau^\mu = g(y^\mu)$$

$y^\mu = B \cdot \xi^\mu$ and $x^\mu = J \cdot \xi^\mu$ are pre-activations and $g(\cdot)$ the activation function.

**Figure:** Teacher with weights $B \in \mathbb{R}^N$

**Figure:** Student with weights $J \in \mathbb{R}^N$
**Macroscopics: Order parameters**

*Order parameters* aggregate the microscopics into a few descriptive parameters.

Overlap \( R = J \cdot B \)

Student magnitude \( Q = J \cdot J \)

Teacher magnitude \( T = B \cdot B = 1 \)

\[
R = \sqrt{Q} \sqrt{T} \cos \phi
\]
On-line learning from a teacher network

Here we assume i.i.d. \( \xi_i \sim \mathcal{N}(0, 1) \) such that \( \langle \xi_i \xi_j \rangle = 0, \quad i \neq j \)
On-line gradient descent

1. Error for the $\mu$th example: $\epsilon^\mu = \frac{1}{2} (\tau^\mu - \sigma^\mu)^2$

2. Update weights $J$ to reduce $\epsilon^\mu$: $J^{\mu+1} = J^\mu + \Delta J$, where $\Delta J = -\frac{\eta}{N} \nabla J \epsilon^\mu$

Weight update

$$J^{\mu+1} = J^\mu + \frac{\eta}{N} \delta^\mu \xi^\mu, \quad \delta^\mu = (\tau^\mu - \sigma^\mu) g'(x^\mu)$$

By a simple substitution of $J^{\mu+1}$, one obtains the recurrences in $R = J \cdot B$ and $Q = J \cdot J$:

$$R^{\mu+1} = R^\mu + \frac{\eta}{N} \delta^\mu y^\mu$$

$$Q^{\mu+1} = Q^\mu + 2 \frac{\eta}{N} \delta^\mu x^\mu + \frac{\eta^2}{N} (\delta^\mu)^2$$
Learning behavior on the level of order parameters

\( \xi \in \mathbb{R}^3 \) i.i.d. \( \xi_i \sim \mathcal{N}(0, 1) \) and \( R(0) = 0, Q(0) = 0.2 \)

\[ \begin{bmatrix} R=J \cdot B \\ Q=J \cdot J \end{bmatrix} \]

\( \text{Time } \alpha = \mu / N \)

\( \rightarrow: B \in \mathbb{R}^3 \)
Learning a rule in higher dimensions

\[ \xi \in \mathbb{R}^N \text{ i.i.d } \xi_i \sim \mathcal{N}(0, 1) \text{ and } R(0) = 0, Q(0) = 0.2 \]

Figure: Learning in \( \mathbb{R}^{60} \)

Time \( \alpha = \mu/N \)
Learning in the thermodynamic limit $N \to \infty$

Order parameters are *self-averaging* $\to$ Deterministic equations in the thermodynamic limit $N \to \infty$ with continuous time $\alpha = \mu/N$. 

Differential equations

\[
dR/d\alpha = \eta \langle \delta y \rangle \xi dQ/d\alpha = 2 \eta \langle \delta x \rangle \xi + \eta^2 \langle \delta^2 \rangle \xi
\]

Pre-activations $x = \sum_{N i=1} J_i \xi_i$ and $y = \sum_{N i=1} B_i \xi_i$ are Gaussians for large $N$ (CLT). Joint density $P(x, y)$ with:

\[
\langle x \rangle = \langle y \rangle = 0 \text{ and } C = (QR R^T)
\]

Averages $\langle \cdot \rangle_{\xi}$ taken over $P(x, y)$ for $g(x) = x \Theta(x)$. 

Michiel Straat, Michael Biehl | September 12th, 2019
Learning in the thermodynamic limit $N \to \infty$

Order parameters are self-averaging $\to$ Deterministic equations in the thermodynamic limit $N \to \infty$ with continuous time $\alpha = \mu/N$.

Differential equations $N \to \infty$

\[
\begin{align*}
\frac{dR}{d\alpha} &= \eta \langle \delta y \rangle \xi \\
\frac{dQ}{d\alpha} &= 2\eta \langle \delta x \rangle \xi + \eta^2 \langle \delta^2 \rangle \xi
\end{align*}
\]
Order parameters are *self-averaging* → Deterministic equations in the thermodynamic limit $N \to \infty$ with continuous time $\alpha = \mu / N$.

### Differential equations $N \to \infty$

$$\frac{dR}{d\alpha} = \eta \langle \delta y \rangle \xi$$
$$\frac{dQ}{d\alpha} = 2\eta \langle \delta x \rangle \xi + \eta^2 \langle \delta^2 \rangle \xi$$

Pre-activations $x = \sum_{i=1}^{N} J_i \xi_i$ and $y = \sum_{i=1}^{N} B_i \xi_i$ are Gaussians for large $N$ (CLT). Joint density $P(x, y)$ with:

$$\langle x \rangle = \langle y \rangle = 0$$
$$C = \begin{pmatrix} Q & R \\ R & T \end{pmatrix}.$$
Towards a statistical physics analysis of multilayer ReLU neural networks

Learning in the thermodynamic limit $N \to \infty$

Order parameters are *self-averaging* $\to$ Deterministic equations in the thermodynamic limit $N \to \infty$ with continuous time $\alpha = \mu/N$.

### Differential equations $N \to \infty$

\[
\frac{dR}{d\alpha} = \eta \langle \delta y \rangle \xi \\
\frac{dQ}{d\alpha} = 2\eta \langle \delta x \rangle \xi + \eta^2 \langle \delta^2 \rangle \xi
\]

Pre-activations $x = \sum_{i=1}^{N} J_i \xi_i$ and $y = \sum_{i=1}^{N} B_i \xi_i$ are Gaussians for large $N$ (CLT). Joint density $P(x, y)$ with:

\[
\langle x \rangle = \langle y \rangle = 0 \text{ and } C = \begin{pmatrix} Q & R \\ R & T \end{pmatrix}.
\]

Averages $\langle \cdot \rangle_\xi$ taken over $P(x, y)$ for $g(x) = x \Theta(x)$. 
Towards a statistical physics analysis of multilayer ReLU neural networks

Soft committee machine

Two-layer network with adaptive first-layer weights.

\[ g(x_1^\mu) \]

\[ g(x_K^\mu) \]

\[ J_{11} \]

\[ J_{21} \]

\[ J_{1K} \]

\[ J_{2K} \]

\[ J_{N1} \]

\[ J_{NK} \]

\[ \xi_1 \]

\[ \xi_2 \]

\[ \xi_N \]

\[ \mu \]

\[ \xi_1^\mu \]

\[ \xi_2^\mu \]

\[ \xi_N^\mu \]

\[ \sigma^\mu \]
Order parameters of the SCM

$M = 3$ teacher hidden units and $K = 3$ student hidden units.

Teacher hidden layer

$T_{nm} = B_n \cdot B_m = \delta_{nm}$

Student hidden layer

$R_{in} = J_i \cdot B_n$

$Q_{ik} = J_i \cdot J_k$
Towards a statistical physics analysis of multilayer ReLU neural networks

**SCM: Solving the ODE system**

\( M = 2 \) teacher units and \( K = 2 \) student.

Initial state: \( R(0) = \begin{pmatrix} 10^{-3} & 0 \\ 0 & 10^{-3} \end{pmatrix}, \quad Q(0) = \begin{pmatrix} 0.2 & 0 \\ 0 & 0.2 \end{pmatrix} \)

---

**Figure:** \( K = M = 2 \) and \( \eta = 0.1 \). Symbols show simulation results for \( N = 10^4 \).
Extension of model with hidden-to-output weights

Introduce adaptive second layer weights:

\[
\sigma^{\mu} = \sum_{i=1}^{K} g(J_i \cdot \xi^\mu) w_i
\]

\[
\tau^{\mu} = \sum_{n=1}^{M} g(B_n \cdot \xi^\mu) v_n
\]

Where we update \( w_i \) with gradient descent: \( w_i^{\mu+1} = w_i^{\mu} - \eta \frac{\partial \epsilon}{\partial w_i} \). From previous research:

- Second layer weights are self-averaging.
- Put second layer weights on a faster timescale.
Two-layer architecture with adaptive second layer weights

\begin{align*}
\xi_1^{\mu} & \rightarrow & \text{Input layer} \\
\xi_2^{\mu} & \rightarrow & \text{Hidden layer} \\
\xi_N^{\mu} & \rightarrow & \text{Output layer}
\end{align*}

\[ g(x_1^{\mu}) + J_{11} + J_{21} + J_{1K} + J_{2K} + J_{NK} \]

\[ w_1 + w_2 \]

\[ \sigma^{\mu} \]
Results

\( K = M = 2 \), teacher second layer weights \( v_1 = 1.2, v_2 = 3 \) and non-adaptive student weights \( w_1, w_2 \). Learning rate \( \eta = 0.1 \) and initial specialization \( R_{11} = R_{22} = 10^{-3} \).

\[ Q_{11} \text{ and } Q_{22} \text{ compensate for the rule’s second layer weights.} \]
$K = M = 2$, teacher second layer weights $v_1 = 1.2$, $v_2 = 3$ and non-adaptive student weights $w_1, w_2$. Learning rate $\eta = 0.1$ and initial specialization $R_{11} = R_{22} = 10^{-3}$.

$\epsilon_g(\alpha \to \infty) = 0 \rightarrow \text{This rule is indeed learnable.}$
Towards a statistical physics analysis of multilayer ReLU neural networks

Results

For a teacher with general weights $v_1, v_2 \in \mathbb{R}$, we need to consider adaptive $w_1, w_2$. Results for simulations with $K = M = 2$ and $N = 1000$:

$v_1 = -1.2, v_2 = -3$
Towards a statistical physics analysis of multilayer ReLU neural networks

Results

For a teacher with general weights $v_1, v_2 \in \mathbb{R}$, we need to consider adaptive $w_1, w_2$. Results for simulations with $K = M = 2$ and $N = 1000$:

Multiple ways of realizing the rule: In this case, lower $|w_1|$ and $|w_2|$ gets compensated by higher $Q_{11}$ and $Q_{22}$, which realizes $\epsilon_g(\alpha \rightarrow \infty) = 0$. 
Future work

- Add second layer updates to the differential equations (straightforward)
- Introduce biases:

\[
\sigma^\mu = \sum_{i=1}^{K} g(J_i \cdot \xi^\mu + \theta_i) w_i \\
\tau^\mu = \sum_{n=1}^{M} g(B_n \cdot \xi^\mu + \phi_n) v_n
\]

The model is now a *universal approximator*, also for ReLU activation.

- Analyses of the ODE system \(N \rightarrow \infty\) for the above scenario, including
  - Optimal learning rates and learning rate adaptation.
  - Different activation function in student and teacher.
  - Complex students (large \(K\)) learning a simple rule (Small \(M\)).

Regularization techniques.
Tree architectures

- Each hidden unit gets a part $\tilde{\xi}_i$ of the input.
- Local potentials $x_i$ are mutually independent in this case.
- Matrices $R, Q$ and $T$ are diagonal.