Towards a statistical physics analysis of multilayer ReLU neural networks

Michiel Straat, Michael Biehl

September 12th, 2019

university of groningen

Content

Learning from a teacher network

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

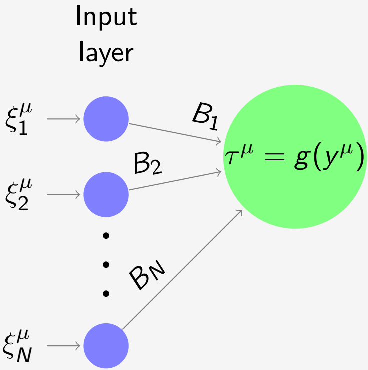


Figure: Teacher with weights $\boldsymbol{B} \in \mathbb{R}^N$

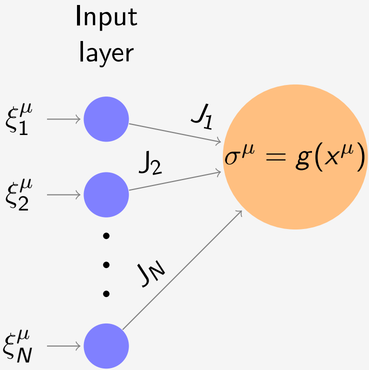


Figure: Student with weights $\boldsymbol{J} \in \mathbb{R}^N$

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

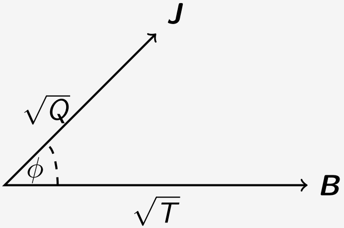
Macroscopics: Order parameters

Order parameters aggregate the microscopics into a few descriptive parameters.

Overlap $R = \boldsymbol{J} \cdot \boldsymbol{B}$

Student magnitude $Q = \boldsymbol{J} \cdot \boldsymbol{J}$

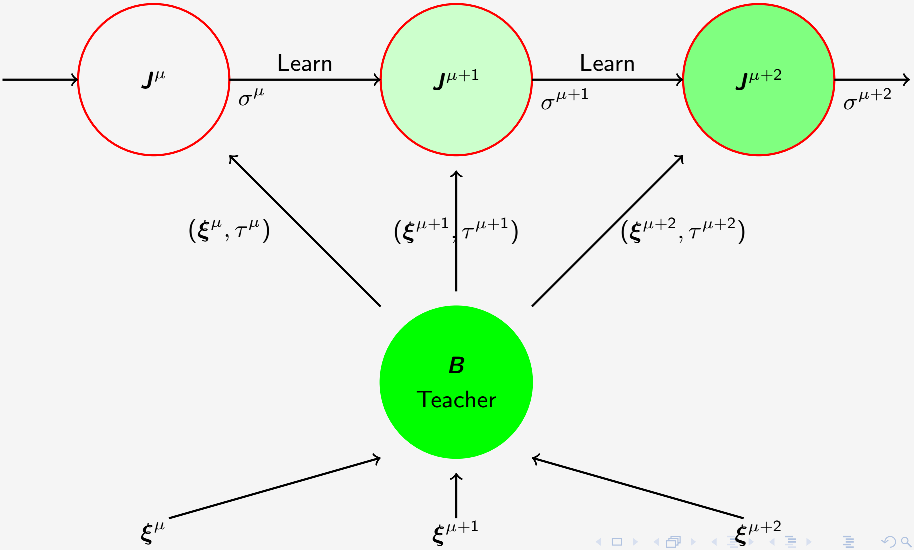
Teacher magnitude $T = \boldsymbol{B} \cdot \boldsymbol{B} = 1$



$$R = \sqrt{Q}\sqrt{T}\cos\phi$$

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 μth example: $\epsilon^{\mu} = \frac{1}{2}(\tau^{\mu} - \sigma^{\mu})^2$
2. Update weights $\boldsymbol{J}$ to reduce ϵ^{μ}: $\boldsymbol{J}^{\mu+1} = \boldsymbol{J}^{\mu} + \Delta\boldsymbol{J}$, where $\Delta\boldsymbol{J} = -\frac{\eta}{N}\nabla_{\boldsymbol{J}}\epsilon^{\mu}$

Weight update

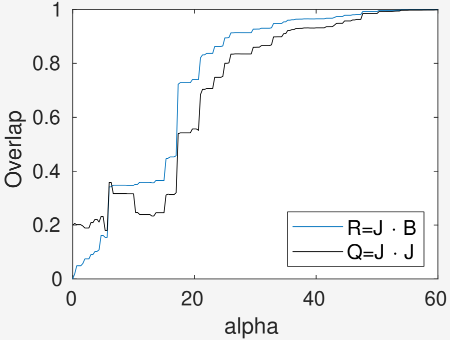
$$\boldsymbol{J}^{\mu+1} = \boldsymbol{J}^{\mu} + \frac{\eta}{N}\delta^{\mu}\boldsymbol{\xi}^{\mu}, \quad \delta^{\mu} = (\tau^{\mu} - \sigma^{\mu})g'(x^{\mu})$$

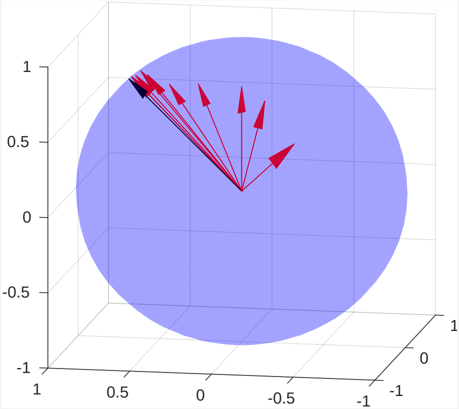
By a simple substitution of $J^{\mu+1}$, one obtains the recurrences in $R = \boldsymbol{J} \cdot \boldsymbol{B}$ and $Q = \boldsymbol{J} \cdot \boldsymbol{J}$:

$$R^{\mu+1} = R^{\mu} + \frac{\eta}{N}\delta^{\mu}y^{\mu}$$
$$Q^{\mu+1} = Q^{\mu} + 2\frac{\eta}{N}\delta^{\mu}x^{\mu} + \frac{\eta^2}{N}(\delta^{\mu})^2$$

Learning behavior on the level of order parameters

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





$\rightarrow$: $\boldsymbol{B} \in \mathbb{R}^3$

Time $\alpha = \mu/N$

Learning a rule in higher dimensions

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

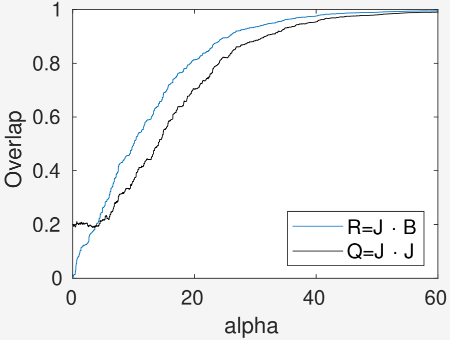


Figure: Learning in $\mathbb{R}^{60}$

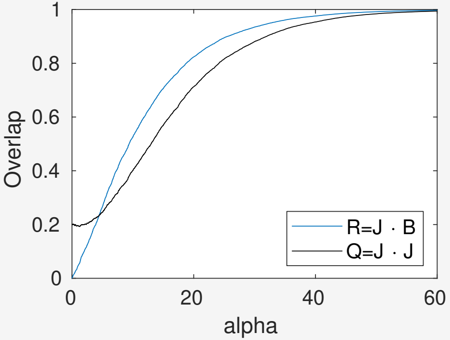


Figure: Learning in $\mathbb{R}^{1000}$

Time $\alpha = \mu/N$

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

Order parameters are *self-averaging* $\rightarrow$ Deterministic equations in the thermodynamic limit $N \to \infty$ with continuous 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 $N \to \infty$

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

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 $N \to \infty$

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

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:

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

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 $N \to \infty$

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

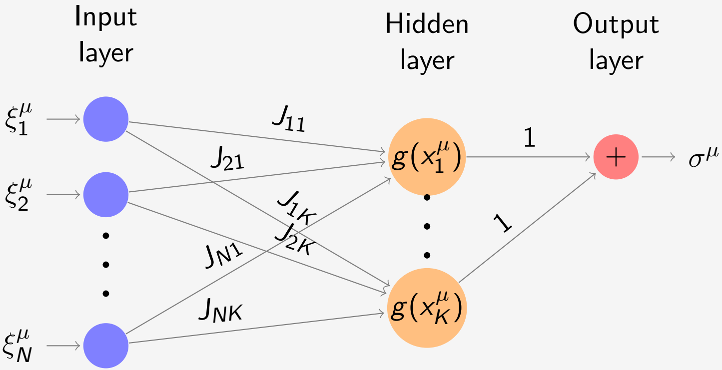
Pre-activations $x = \sum_{i=1}^{N} J_i \xi_i$ and $y = \sum_{i=1}^{N} B_i \xi_i$ are Gaussians for large N (CLT). Joint density $P(x, y)$ with:

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

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

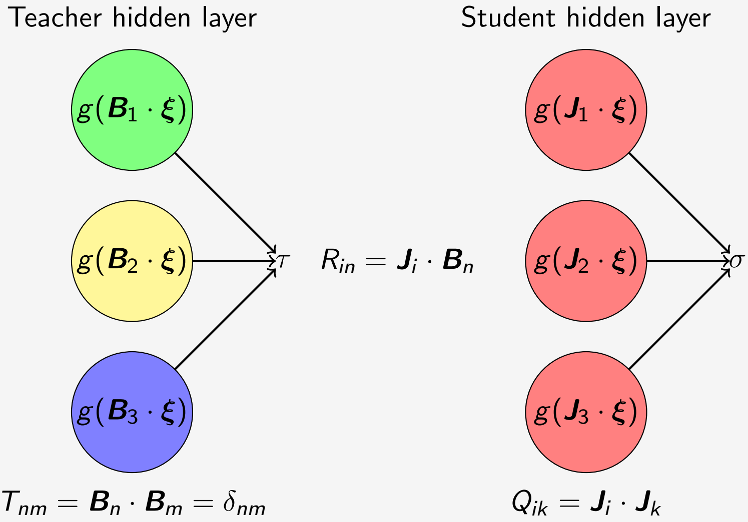
Soft committee machine

Two-layer network with adaptive first-layer weights.



Order parameters of the SCM

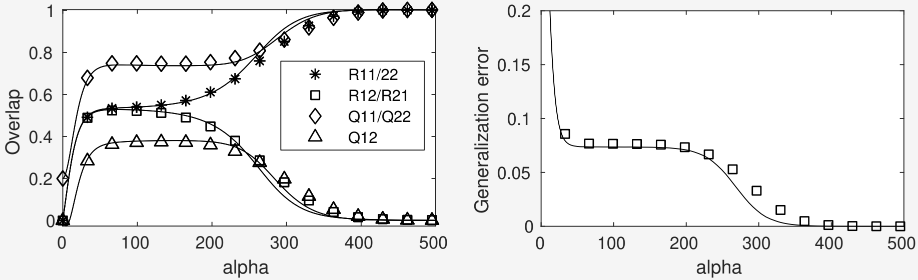
$M = 3$ teacher hidden units and $K = 3$ student hidden units.



SCM: Solving the ODE system

$M = 2$ teacher units and $K = 2$ student.

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



(a) Order parameters (b) ϵ_g

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:

Student output

$$\sigma^\mu = \sum_{i=1}^{K} g(\boldsymbol{J}_i \cdot \boldsymbol{\xi}^\mu) w_i$$

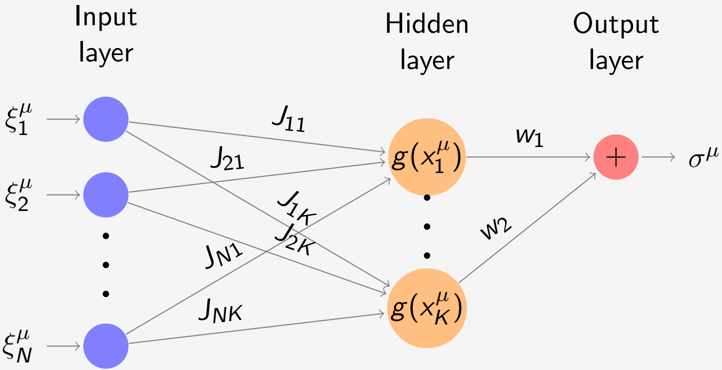
Teacher output

$$\tau^\mu = \sum_{n=1}^{M} g(\boldsymbol{B}_n \cdot \boldsymbol{\xi}^\mu) v_n$$

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:

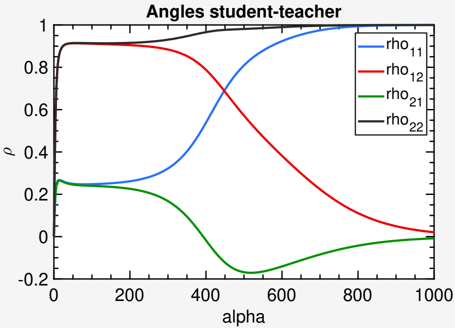
- Second layer weights are self-averaging.
- Put second layer weights on a faster timescale.

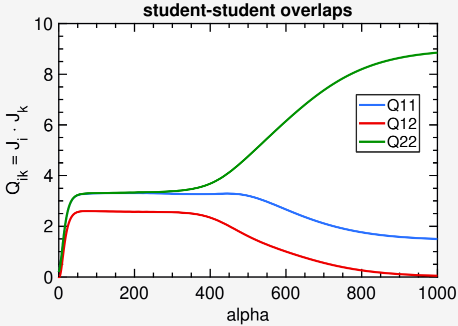
Two-layer architecture with adaptive second layer weights



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 $R_{11} = R_{22} = 10^{-3}$.

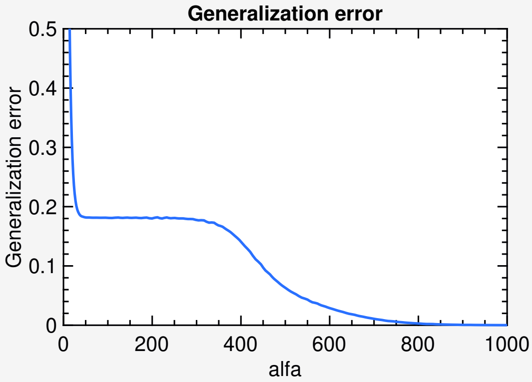




Q_{11} and Q_{22} compensate for the rule's second layer weights.

Results

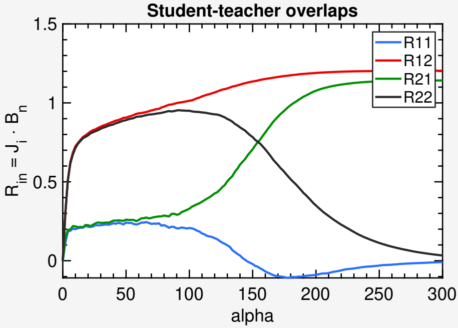
$K = M = 2$, teacher second layer weights $v_1 = 1.2, v_2 = 3$ and non-adaptive student weights w_1, w_2. Learning rate $\eta = 0.1$ and initial specialization $R_{11} = R_{22} = 10^{-3}$.

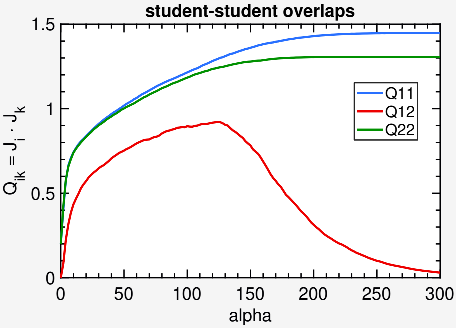


$\epsilon_g(\alpha \to \infty) = 0 \to$ This rule is indeed learnable.

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 $N = 1000$:

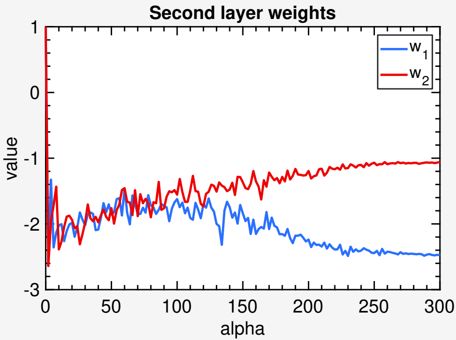


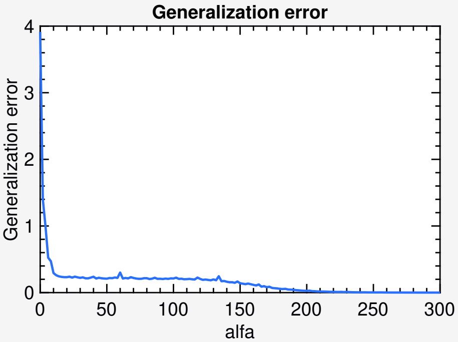


$v1 = -1.2, v2 = -3$

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 $\epsilon_g(\alpha \to \infty) = 0$.

Future work

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

Student output

$$\sigma^\mu = \sum_{i=1}^{K} g(\boldsymbol{J}_i \cdot \boldsymbol{\xi}^\mu + \theta_i) w_i$$

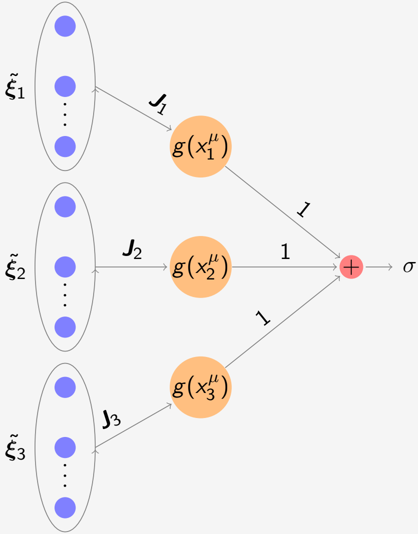
Teacher output

$$\tau^\mu = \sum_{n=1}^{M} g(\boldsymbol{B}_n \cdot \boldsymbol{\xi}^\mu + \phi_n) v_n$$

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

- Analyses of the ODE system ($N \to \infty$) 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.