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

M. Straat, M. Biehl

University of Groningen

April 26, 2019

Content

1. Learning from a teacher network
2. Description in terms of order parameters
3. Evolution of order parameters in the thermodynamic limit
4. Behavior of the ReLU perceptron and Soft Committee Machine

Learning from a teacher network

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

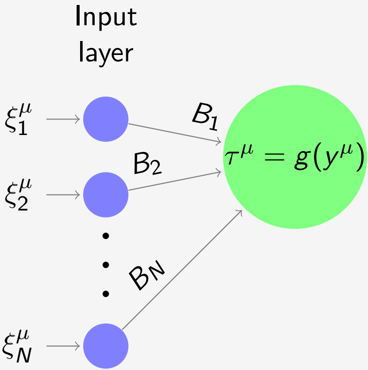


Figure: Teacher with weights $\boldsymbol{B} \in \mathbb{R}^N$

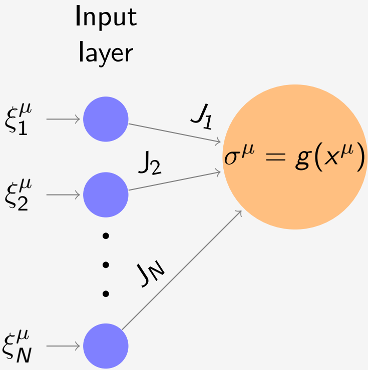
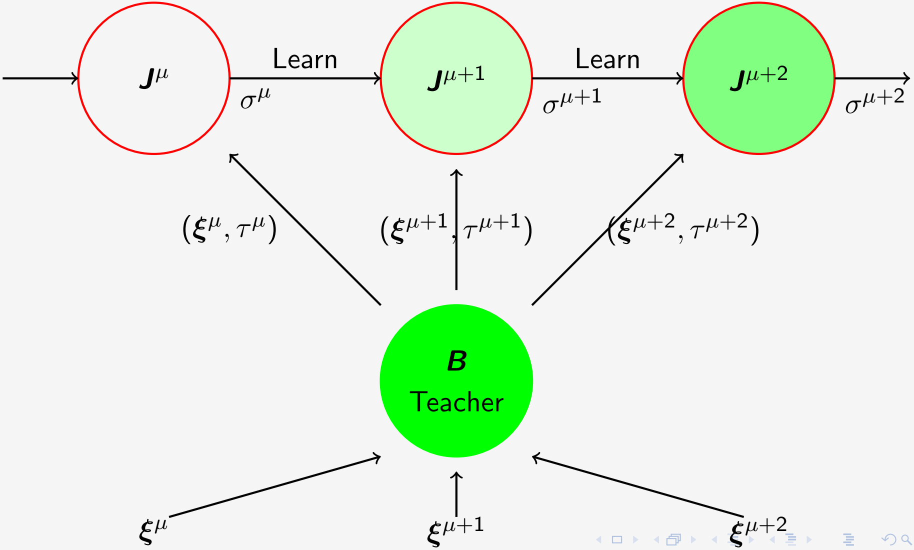


Figure: Student with weights $\boldsymbol{J} \in \mathbb{R}^N$

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

On-line learning from a teacher network



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

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

Generalization error: $\epsilon_g(\boldsymbol{J}) = \langle\epsilon\rangle_{\boldsymbol{\xi}}$

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

The weights $\boldsymbol{J}$ and $\boldsymbol{B}$ are the microscopics of the system.

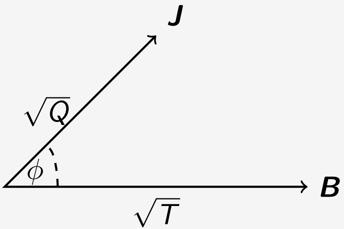
Macroscopics: Order parameters

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



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

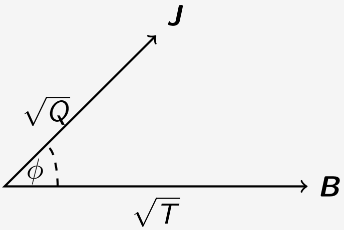
Macroscopics: Order parameters

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



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

$R^{\mu+1}$ and $Q^{\mu+1}$ follow from substituting $\boldsymbol{J}^{\mu+1}$:

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

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

Erf activation

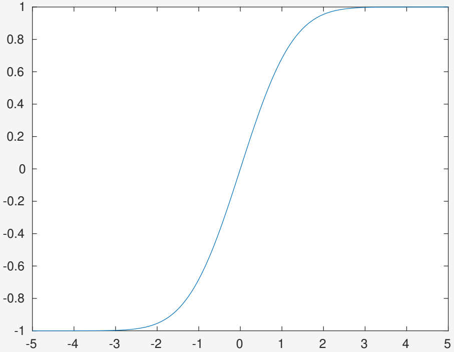


Figure: $g(x) = \mathrm{erf}(x/\sqrt{2})$

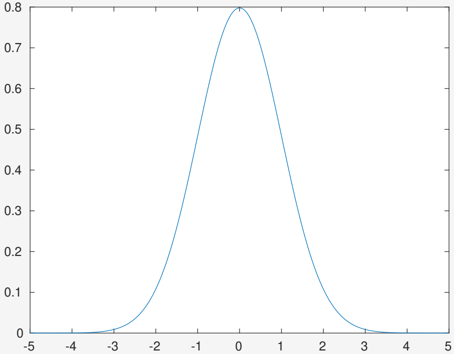
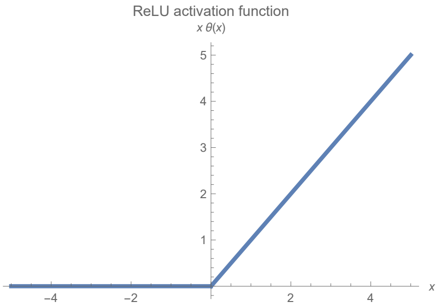


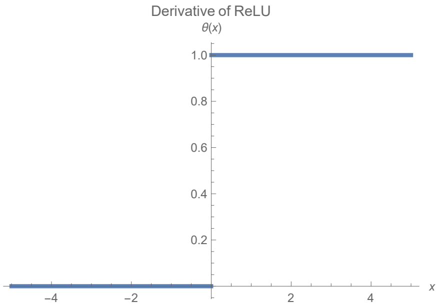
Figure: $g'(x) = \sqrt{2/\pi} e^{-x^2/2}$

ReLU activation

$g(x) = x\Theta(x)$

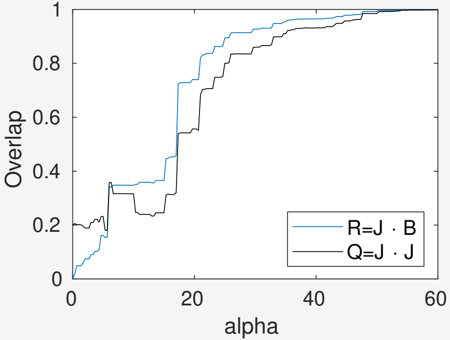


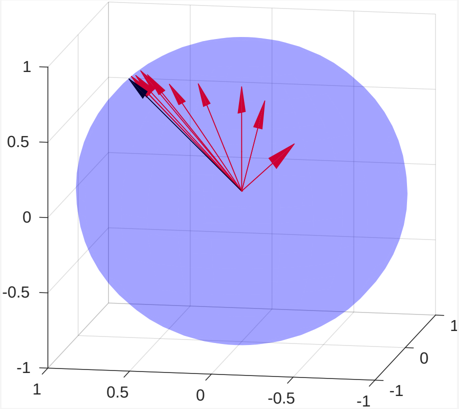
$g'(x) = \Theta(x)$



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 $R(0) = 0, Q(0) = 0.2$

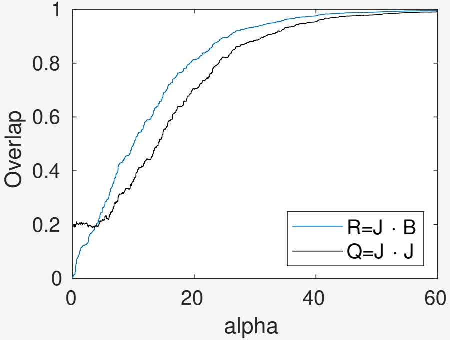


Figure: Learning in $\mathbb{R}^{60}$

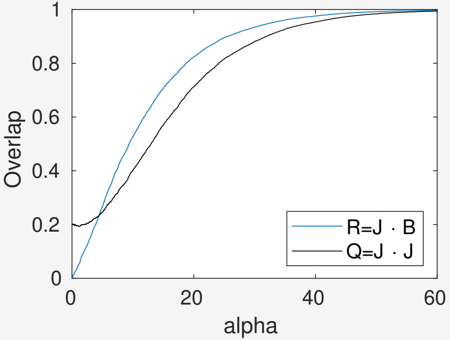


Figure: Learning in $\mathbb{R}^{1000}$

Time $\alpha = \mu/N$

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

Order parameters are *self-averaging* $\to$ 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 $g(x) = x\Theta(x)$.

Solving the ODE system

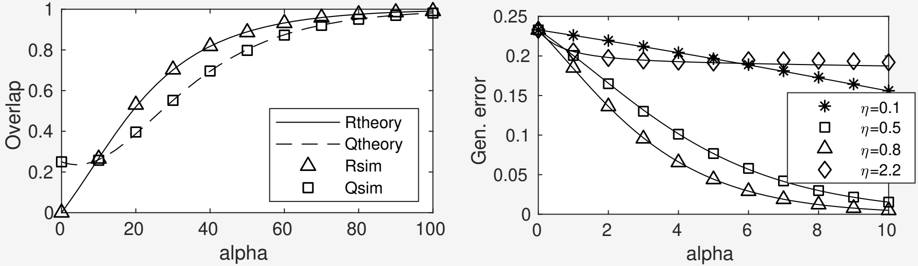


Figure: *Left*: Evolution of R and Q with $\eta = 0.1$, $R(0) = 0$ and $Q(0) = 0.25$. *Right*: Evolution of ϵ_g for different η. Lines and symbols show theoretical and simulation ($N = 1000$) results, respectively.

Soft committee machine

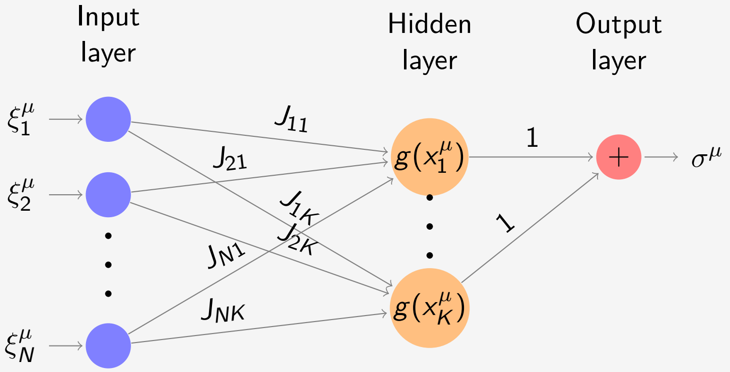


Figure: Soft committee machine with K hidden units.

Weight matrix $\boldsymbol{J} \in \mathbb{R}^{N\times K}$

Student output

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

Teacher output

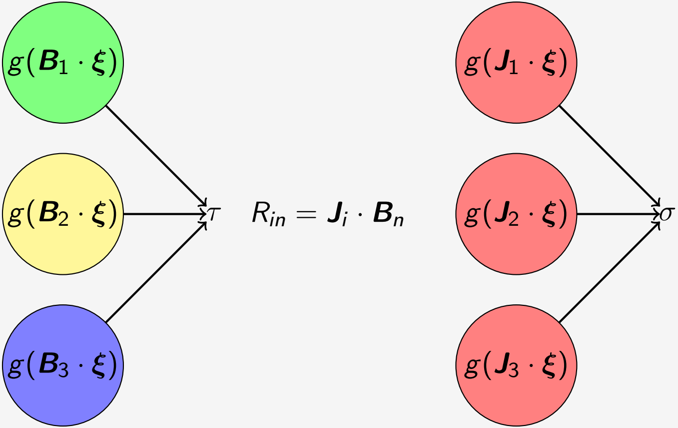
$$\tau^\mu = \sum_{n=1}^{M} g(\boldsymbol{B}_n \cdot \boldsymbol{\xi}^\mu)$$

Order parameters of the SCM

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

Teacher hidden layer

Student hidden layer



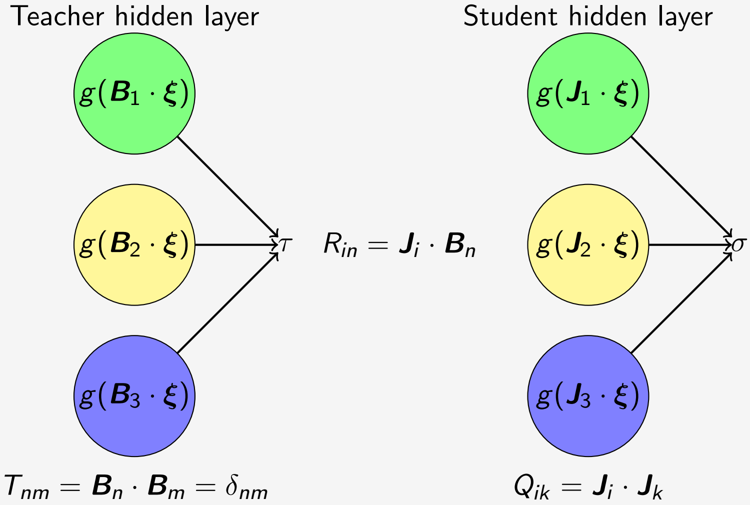
$R_{in} = \boldsymbol{J}_i \cdot \boldsymbol{B}_n$

$T_{nm} = \boldsymbol{B}_n \cdot \boldsymbol{B}_m = \delta_{nm}$

$Q_{ik} = \boldsymbol{J}_i \cdot \boldsymbol{J}_k$

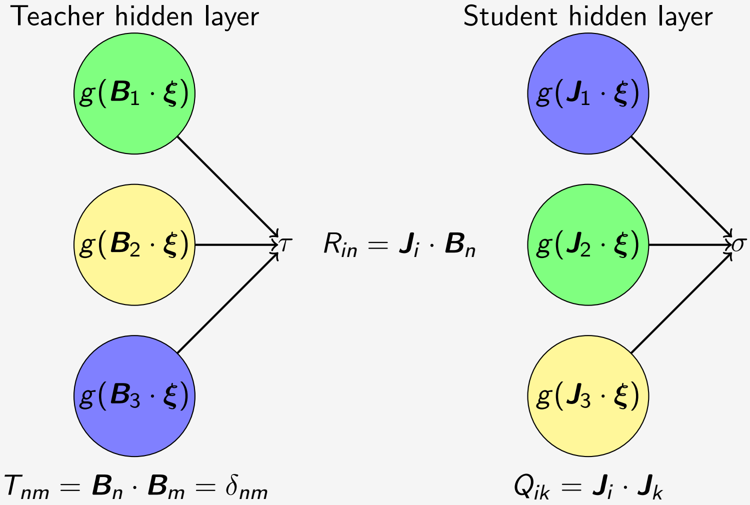
Order parameters of the SCM

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



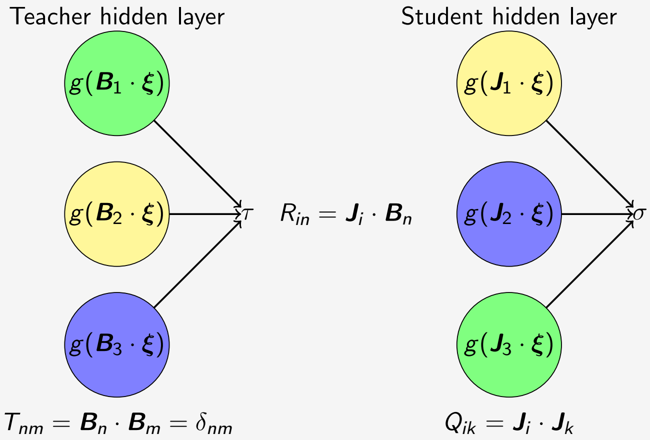
Order parameters of the SCM

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



Order parameters of the SCM

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

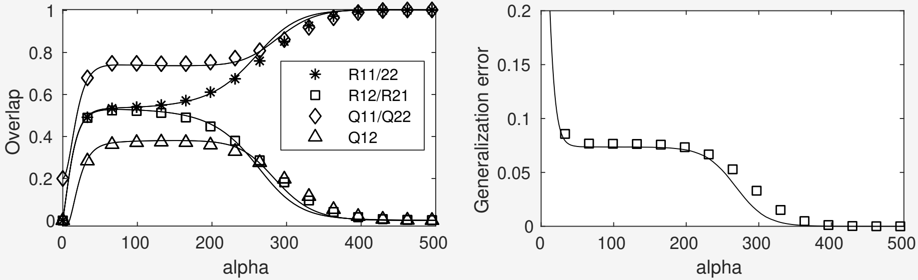


$M!$ possible permutations and therefore realizations of the rule.

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 $N = 10^4$.

SCM: Solving the ODE system

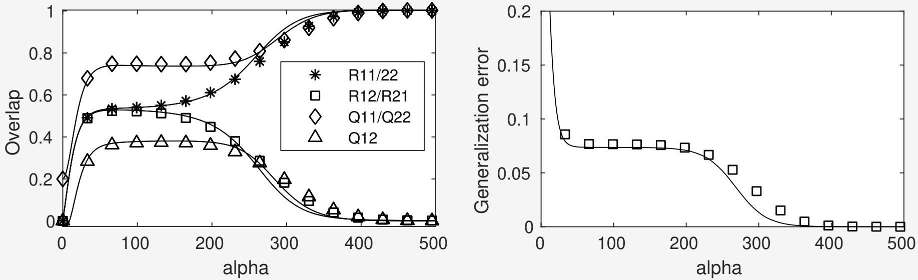
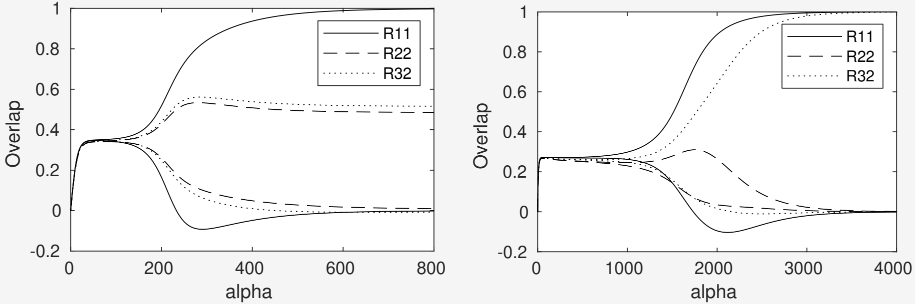


Figure: $K = M = 2$ and $\eta = 0.1$. Symbols show simulation results for $N = 10^4$.

Plateau: $R_{in} = R$, $Q_{ii} = Q$ and $Q_{ik} = C$ (fixed point of ODE).
Eigenvalue $\lambda_5 = 0.24$ with eigenvector $\boldsymbol{u}_5 = (0.5, -0.5, -0.5, 0.5, 0, 0, 0)^T$ guides the escape from symmetry to the start of specialization: $\boldsymbol{J}_1 \to \boldsymbol{B}_1$ and $\boldsymbol{J}_2 \to \boldsymbol{B}_2$

Overrealizable scenarios $K > M$

$M = 2$ teacher units and $K = 3$ student units, $\quad R_{11}(0) = 10^{-3}$



(a) $g(x) = x\Theta(x)$ (b) $g(x) = \mathrm{erf}(x/\sqrt{2})$

ReLU: $\max(\boldsymbol{B}_2 \cdot \boldsymbol{\xi}, 0) = \max(a\boldsymbol{B}_2 \cdot \boldsymbol{\xi}, 0) + \max(b\boldsymbol{B}_2 \cdot \boldsymbol{\xi}, 0)$ for $a + b = 1$

Not possible for non-linear Erf function: $Q_{22} \to 0$

Future work

- Regularization techniques (e.g. Dropout)
- Compare behavior of activation functions
- Concept drift
- Learning rate adaptation
- Adaptive second layer weights
- Extension to more layers

Thank you

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