On-line learning in neural networks with ReLU activation

Michiel Straat

September 19, 2018

Overview

Statistical Mechanics

- Aims to deduce macroscopic properties from microscopic dynamic properties in systems consisting of e.g. $N \approx 10^{23}$ particles.
- Due to Central Limit Theorems (CLT), fluctuations in the macroscopics become negligible $\rightarrow \sigma$ decreases as $O(1/\sqrt{N})$.

Example system: Ideal paramagnet

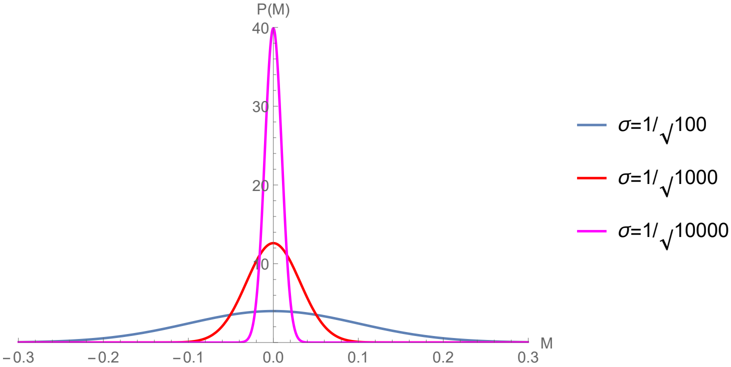
$\uparrow\uparrow\downarrow\uparrow\downarrow\uparrow\cdots\downarrow$
Consider N spins, each spin i has a value S_i:

$$S_i = \begin{Bmatrix} 1, & \text{if } \uparrow \\ -1, & \text{if } \downarrow \end{Bmatrix}.$$

Magnetization:

$$M = \frac{1}{N}\sum_{i=1}^{N} S_i \in [-1, 1]$$

Assume components are i.i.d with $P(S_i = 1) = P(S_i = -1) = \frac{1}{2}$,
$\langle S_i \rangle = 0$ and $\sigma = 1$.
CLT: For large N, approximately $M \sim \mathcal{N}(0, 1/\sqrt{N})$
$\Rightarrow M$ is a deterministic value for $N \to \infty$ (Thermodynamic limit)


P(M)
40
30
20
10
M
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
σ=1/√100
σ=1/√1000
σ=1/√10000

Statistical Physics of online Learning

Online-learning: Uncorrelated examples $\{\boldsymbol{\xi}^{\mu}, \tau^{\mu}\}$ arrive one at the time.

- Previously, online learning in Erf neural networks was characterized using methods of Statistical Mechanics.
- Dynamics of order parameters were formulated, first as difference equations, and in the thermodynamic limit as differential equations.
- Here, the same method is used to characterize online learning in ReLU neural networks.

Student-teacher framework

The target output $\tau(\boldsymbol{\xi})$ is defined by the teacher network. Student tries to learn the rule. $g(\cdot)$ is *activation function*.

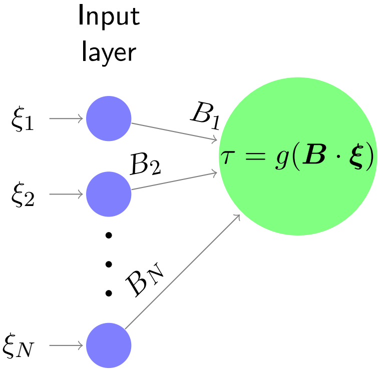


Figure: Teacher with weights $\boldsymbol{B} \in \mathbb{R}^N$

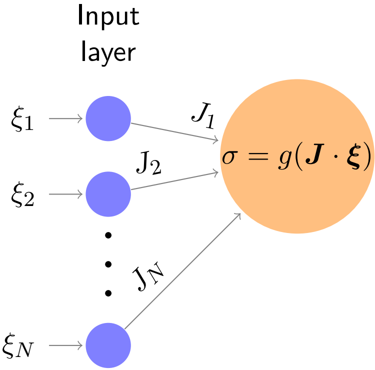


Figure: Student with weights $\boldsymbol{J} \in \mathbb{R}^N$

Generalization error

Teacher
Input activation: $y^\mu = \boldsymbol{B} \cdot \boldsymbol{\xi}^\mu$
Output: $\tau^\mu = g(y^\mu)$

Student
Input activation: $x^\mu = \boldsymbol{J} \cdot \boldsymbol{\xi}^\mu$
Output: $\sigma^\mu = g(x^\mu)$

Error on particular example $\boldsymbol{\xi}^\mu$

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

Generalization error

$$\epsilon_g(\boldsymbol{J}) = \langle \epsilon(\boldsymbol{J}, \boldsymbol{\xi}) \rangle_{\boldsymbol{\xi}}$$

where $\langle \ldots \rangle$ denotes the average over the input distribution.

Assume uncorrelated random components $\xi_i \in \mathcal{N}(0, 1)$.

Gradient descent update rule

Upon presentation of an example $\boldsymbol{\xi}^\mu$, weight vector $\boldsymbol{J}^\mu$ is adapted:

$$\boldsymbol{J}^{\mu+1} = \boldsymbol{J}^\mu - \tfrac{\eta}{N}\nabla_{\boldsymbol{J}}\epsilon(\boldsymbol{J}^\mu, \boldsymbol{\xi}^\mu) =$$
$$\boldsymbol{J}^\mu + \tfrac{\eta}{N}\underbrace{[g(y^\mu) - g(x^\mu)]g'(x^\mu)}_{\delta^\mu}\boldsymbol{\xi}^\mu = \boldsymbol{J}^\mu + \tfrac{\eta}{N}\delta^\mu\boldsymbol{\xi}^\mu$$

- $\frac{\eta}{N}$ is the learning rate scaled by the network size N.
- Actual form of gradient dependent on choice of $g(\cdot)$

Order parameters for large dimension N

$x = \boldsymbol{J} \cdot \boldsymbol{\xi}$, $y = \boldsymbol{B} \cdot \boldsymbol{\xi}$

In the limit $N \to \infty$, the inputs x and y become correlated Gaussian variables according to the Central Limit Theorem, with:

$$\begin{aligned}
&\langle y \rangle = \langle x \rangle = 0 \\
&\langle x^2 \rangle = \sum_{i=1}^{N} \sum_{j=1}^{N} J_i J_j \langle \xi_i \xi_j \rangle = \sum_{i=1}^{N} J_i^2 = ||\boldsymbol{J}||^2 = Q \\
&\langle y^2 \rangle = \sum_{n=1}^{N} \sum_{m=1}^{N} B_n B_m \langle \xi_i \xi_j \rangle = \sum_{n=1}^{N} B_n^2 = ||\boldsymbol{B}||^2 = T = 1 \\
&\langle xy \rangle = \sum_{i=1}^{N} \sum_{n=1}^{N} J_i B_n \langle \xi_i \xi_n \rangle = \sum_{j=1}^{N} J_j B_j = \boldsymbol{J} \cdot \boldsymbol{B} = R
\end{aligned}$$

R and Q are the *order parameters* of the system.

Updates of the order parameters

$$R^{\mu+1} = \boldsymbol{J}^{\mu+1} \cdot \boldsymbol{B} = \underbrace{(\boldsymbol{J}^{\mu} + \frac{\eta}{N}\delta^{\mu}\boldsymbol{\xi}^{\mu})}_{\boldsymbol{J}^{\mu+1}} \cdot \boldsymbol{B}$$

Which leads to the recurrence:

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

Updates of order parameters upon presentation of example $\boldsymbol{\xi}^{\mu}$

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

In the limit $N \to \infty$:

- The scaled time variable $\alpha = \mu/N$ becomes continuous.
- The order parameters become self-averaging.

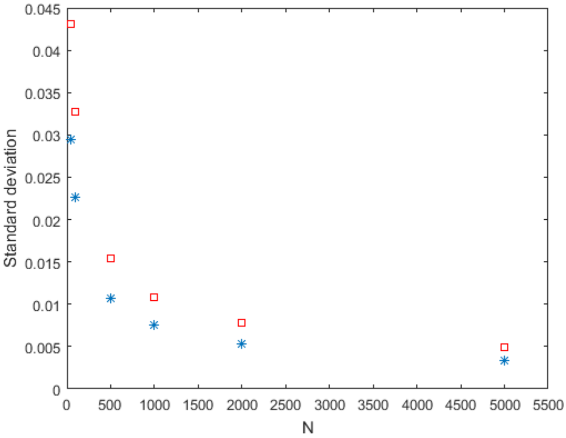


Figure: For fixed $\alpha = 20$, the standard deviation of the order parameters R and Q out of 100 runs for increasing system size N.

$N \to \infty$ (Thermodynamic limit)

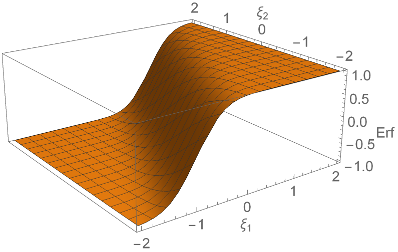
This results in a system of deterministic differential equations for the evolution of the order parameters:

$$\frac{dR}{d\alpha} = \eta\langle\delta y\rangle$$

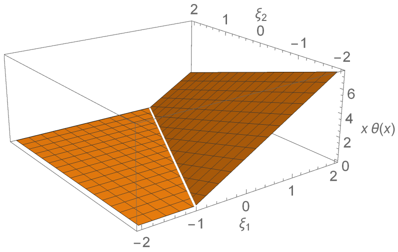
$$\frac{dQ}{d\alpha} = 2\eta\langle\delta x\rangle + \eta^2\langle\delta^2\rangle$$

with $\delta = [g(y) - g(x)]g'(x)$

Choice of activation function



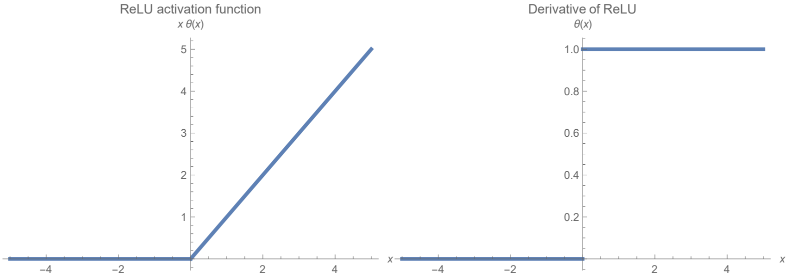
(a) Erf activation



(b) ReLU activation

Figure: Examples of perceptrons with different activation for the same weight vector: $J_1 = 2.5$ and $J_2 = -1.2$.

ReLU



(a) $g(x) = x\theta(x)$ (b) $g'(x) = \theta(x)$

Figure: The ReLU activation function and its derivative.

ReLU Perceptron learning dynamics

$$\begin{aligned}\frac{dR}{d\alpha} = \eta\langle\delta y\rangle &= \eta(\langle g'(x)g(y)y\rangle - \langle g'(x)g(x)y\rangle) \\ &= \eta(\langle y^2\theta(x)\theta(y)\rangle - \langle xy\theta(x)\rangle)\end{aligned}$$

$$\begin{aligned}\frac{dQ}{d\alpha} = 2\eta\langle\delta x\rangle + \eta^2\langle\delta^2\rangle &= 2\eta(\langle g'(x)g(y)x\rangle - \langle g'(x)g(x)x\rangle) + \eta^2\langle\delta^2\rangle \\ &= 2\eta(\langle xy\theta(x)\theta(y)\rangle - \langle x^2\theta(x)\rangle) + \eta^2\langle\delta^2\rangle\end{aligned}$$

The 2D integrals are taken over the joint Gaussian $P(x,y)$ with covariance matrix:
$\Sigma = \begin{pmatrix}\langle x^2\rangle & \langle xy\rangle \\ \langle xy\rangle & \langle y^2\rangle\end{pmatrix} = \begin{pmatrix}Q & R \\ R & 1\end{pmatrix}$

ReLU Perceptron learning dynamics

All averages can be expressed analytically in terms of the order parameters. The following system is obtained:

$$\frac{\partial R}{\partial \alpha} = \eta \left(\frac{T}{4} - \frac{R}{2} + \frac{T \sin^{-1}\left(\frac{R}{\sqrt{T\,Q}}\right)}{2\,\pi} + \frac{R\sqrt{T\,Q - R^2}}{2\,\pi\,Q} \right)$$

$$\frac{\partial Q}{\partial \alpha} = \eta \left(\frac{R}{2} - Q + \frac{\sqrt{T\,Q - R^2}}{\pi} + \frac{\sin^{-1}\left(\frac{R}{\sqrt{T\,Q}}\right) R}{\pi} \right) +$$
$$\eta^2 \left(\frac{T}{4} + \left(\frac{R}{Q} - 2 \right) \frac{\sqrt{QT - R^2}}{2\pi} + (T - 2R) \frac{\sin^{-1}\left(\frac{R}{\sqrt{TQ}}\right)}{2\pi} - \frac{R}{2} + \frac{Q}{2} \right)$$

Integrating the above ODE's numerically yields the evolution of $R(\alpha)$ and $Q(\alpha)$.

Generalization error

$$\epsilon_g(\boldsymbol{J}) = \langle \epsilon(\boldsymbol{J}, \boldsymbol{\xi}) \rangle_{\boldsymbol{\xi}} = \tfrac{1}{2}[\langle g(y)^2 \rangle - 2\langle g(y)g(x) \rangle + \langle g(x)^2 \rangle]$$

For ReLU activation, this yields:

$$\epsilon_g(\boldsymbol{J}) = \tfrac{1}{2}[\langle y^2\theta(y) \rangle - 2\langle xy\theta(x)\theta(y) \rangle + \langle x^2\theta(x) \rangle]$$

Performing the averages yields an analytic expression in terms of order parameters R and Q:

$$\epsilon_g(\alpha) = \tfrac{1}{4} - \left(\frac{\sqrt{Q-R^2}}{2\pi} + \frac{R\sin^{-1}\left(\frac{R}{\sqrt{Q}}\right)}{2\pi} + \frac{R}{4}\right) + \frac{Q}{4}$$

Solving the ODE's for $R(\alpha)$ and $Q(\alpha)$ yields evolution of $\epsilon_g(\alpha)$.

ReLU perceptron: Results order parameters

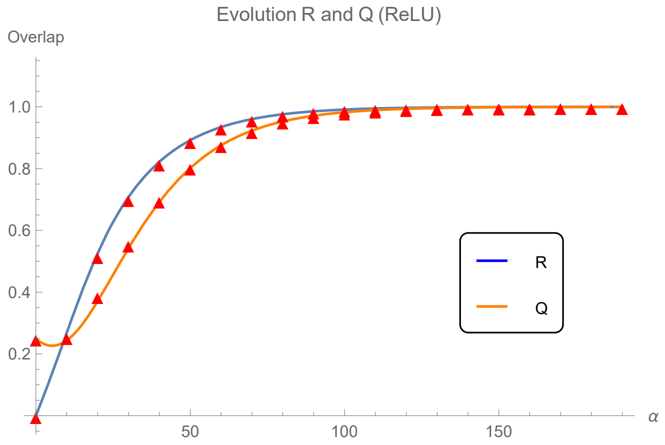
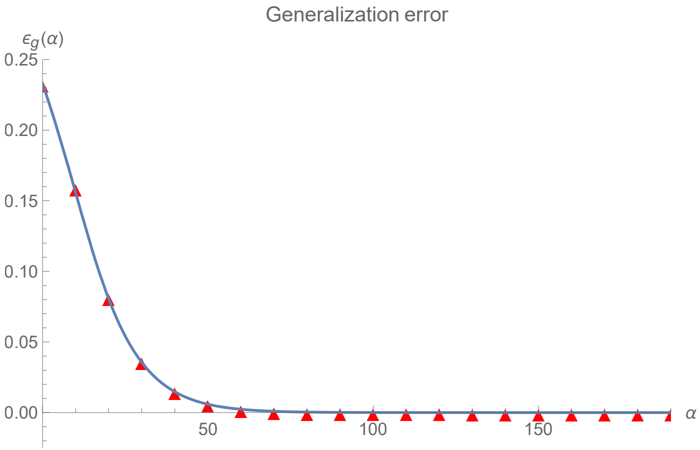


Figure: *solid lines*: Theoretical results with $R(0) = 0$, $Q(0) = 0.25$ and $\eta = 0.1$. *Red triangles*: Simulation with $N = 1000$.

Generalization error result



Stability perfect solution $R = Q = 1$

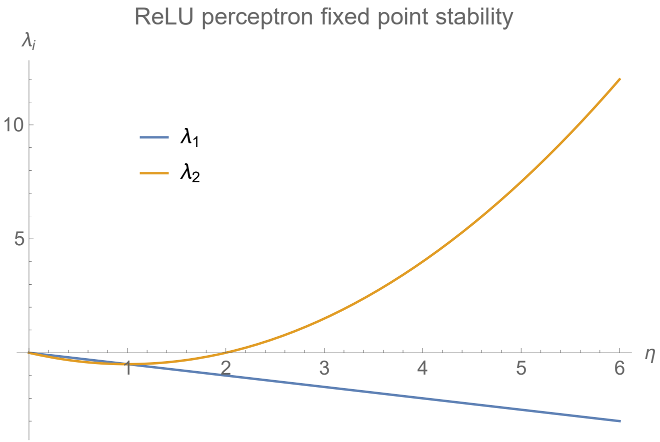
At $R = Q = 1$, $\frac{dR}{d\alpha} = 0$ and $\frac{dQ}{d\alpha} = 0 \rightarrow$ fixed point.

We consider the linear system
$\dot{\boldsymbol{z}} = \boldsymbol{F}\boldsymbol{z} = \begin{pmatrix} -\frac{\eta}{2} & 0 \\ -(\eta-1)\eta & \frac{1}{2}(\eta-2)\eta \end{pmatrix} \begin{pmatrix} R-1 \\ Q-1 \end{pmatrix}$ around the fixed point.

Eigenvalues $\lambda_1(\eta) = -\frac{\eta}{2}$ and $\lambda_2(\eta) = \frac{1}{2}(\eta-2)\eta$ determine stability of the fp.

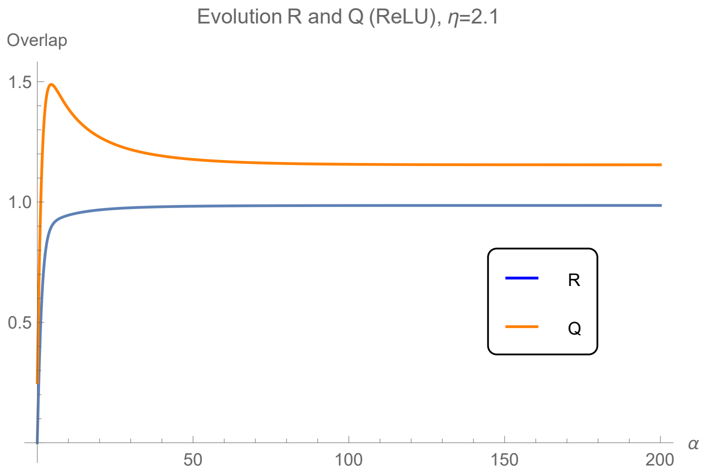
Fixed point stability vs. learning rate η

$$\lambda_1(\eta) = -\frac{\eta}{2},\ \lambda_2(\eta) = \frac{1}{2}(\eta - 2)\eta$$

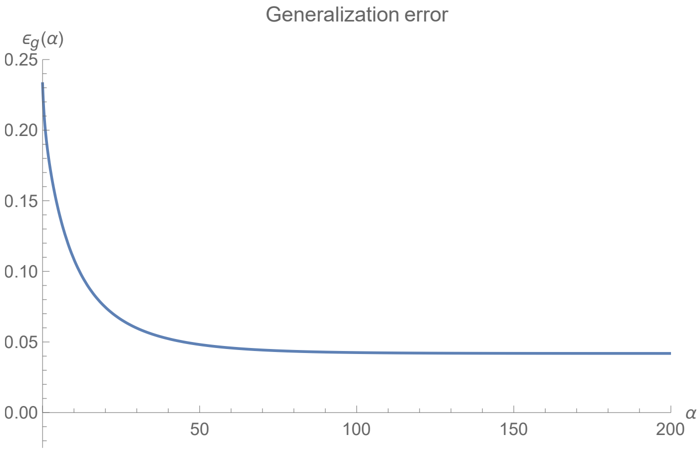


$\eta_c = 2$, eig. vectors: $\boldsymbol{u}_1 = (1/2, 1)^T, \boldsymbol{u}_2 = (0, 1)^T$

$R(\alpha)$ and $Q(\alpha)$ for $\eta = 2.1$



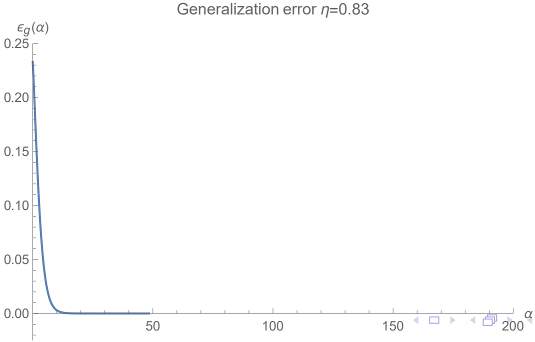
Generalization error for $\eta = 2.1$



Optimal learning rate η_{opt}

An optimal learning rate would have the characteristics:

- Stable at the perfect solution $(R, Q) = (1, 1)$, therefore $\eta_{\text{opt}} < \eta_c$
- Reach the perfect solution the fastest
- $\eta_{\text{opt}} \approx 0.83$



Soft committee machine

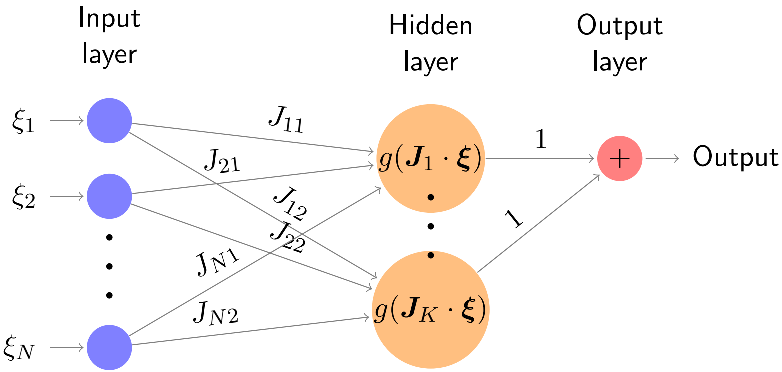


Figure: Soft committee machine with K hidden units.

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 SCM

The given SCM has $K * N$ adaptable weights.

Student inputs	**Teacher inputs**
$x_i = \boldsymbol{J}_i \cdot \boldsymbol{\xi}, \quad i \in [1, 2, ..., K]$	$y_n = \boldsymbol{B}_n \cdot \boldsymbol{\xi}, \quad n \in [1, 2, ..., M]$

$P(x_i, y_n)$ is the $K + M$-dimensional Gaussian with covariance matrix $\boldsymbol{\Sigma} = \begin{pmatrix} Q_{ik} & R_{in} \\ R_{in}^T & T_{nm} \end{pmatrix} \in \mathbb{R}^{(K+M)\times(K+M)}$.

There are $\underbrace{K * M}_{R_{in}} + \underbrace{K(K+1)/2}_{Q_{ik}}$ order parameters and ODE's describing their evolution.

ODE's order parameters SCM

Let δ_i be $g'(x_i)(\tau^\mu - \sigma^\mu)$

$$
\begin{aligned}
\frac{\partial R_{in}}{\partial \alpha} &= \eta \langle \delta_i y_n \rangle \\
&= \eta \left\langle g'(x_i) \left[\sum_{m=1}^{M} g(y_m) - \sum_{j=1}^{K} g(x_j) \right] y_n \right\rangle \\
&= \eta \left[\sum_{m=1}^{M} \langle g'(x_i) y_n g(y_m) \rangle - \sum_{j=1}^{K} \langle g'(x_i) y_n g(x_j) \rangle \right] \\
&= \eta \left[\sum_{m=1}^{M} \langle \theta(x_i) y_n y_m \theta(y_m) \rangle - \sum_{j=1}^{K} \langle \theta(x_i) y_n x_j \theta(x_j) \rangle \right]
\end{aligned}
$$

I_3 integrals ReLU

It turns out the integrals $\langle\theta(u)vw\theta(w)\rangle$ can be expressed analytically:

$\langle\theta(u)vw\theta(w)\rangle = \frac{\sigma_{12}\sqrt{\sigma_{11}\sigma_{33}-\sigma_{13}^2}}{2\pi\sigma_{11}} + \frac{\sigma_{23}\sin^{-1}\left(\frac{\sigma_{13}}{\sqrt{\sigma_{11}\sigma_{33}}}\right)}{2\pi} + \frac{\sigma_{23}}{4}$, and hence:

$$\frac{\partial R_{in}}{\partial\alpha} = \eta\left[\sum_{m=1}^{M}\left(\frac{R_{in}\sqrt{Q_{ii}T_{mm}-R_{im}^2}}{2\pi Q_{ii}} + \frac{T_{nm}\sin^{-1}\left(\frac{R_{im}}{\sqrt{Q_{ii}T_{mm}}}\right)}{2\pi} + \frac{T_{nm}}{4}\right) - \sum_{j=1}^{K}\left(\frac{R_{in}\sqrt{Q_{ii}Q_{jj}-Q_{ij}^2}}{2\pi Q_{ii}} + \frac{R_{jn}\sin^{-1}\left(\frac{Q_{ij}}{\sqrt{Q_{ii}Q_{jj}}}\right)}{2\pi} + \frac{R_{jn}}{4}\right)\right]$$

Student-student overlaps in limit $\eta \to 0$

$\frac{dQ_{ik}}{d\alpha} = \eta(\langle x_i\delta_k\rangle + \langle x_k\delta_i\rangle)\cancel{+\eta^2\langle\delta_i\delta_k\rangle}$

The η^2 term consists of four-dimensional averages I_4, which are omitted initially. Hence, the dynamics are valid for $\eta \to 0$.

$$\frac{\partial Q_{ik}}{\partial\alpha} \approx \eta\Bigg[\sum_{m=1}^{M}\left(\frac{Q_{ik}\sqrt{Q_{ii}T_{mm}-R_{im}^2}}{2\pi Q_{ii}} + \frac{R_{km}\sin^{-1}\left(\frac{R_{im}}{\sqrt{Q_{ii}T_{mm}}}\right)}{2\pi} + \frac{R_{km}}{4}\right) - \sum_{j=1}^{K}\left(\frac{Q_{ik}\sqrt{Q_{ii}Q_{jj}-Q_{ij}^2}}{2\pi Q_{ii}} + \frac{Q_{jk}\sin^{-1}\left(\frac{Q_{ij}}{\sqrt{Q_{ii}Q_{jj}}}\right)}{2\pi} + \frac{Q_{jk}}{4}\right)\Bigg] + \eta\Bigg[\sum_{m=1}^{M}\left(\frac{Q_{ik}\sqrt{Q_{kk}T_{mm}-R_{km}^2}}{2\pi Q_{kk}} + \frac{R_{im}\sin^{-1}\left(\frac{R_{km}}{\sqrt{Q_{kk}T_{mm}}}\right)}{2\pi} + \frac{R_{im}}{4}\right) - \sum_{j=1}^{K}\left(\frac{Q_{ik}\sqrt{Q_{kk}Q_{jj}-Q_{jk}^2}}{2\pi Q_{kk}} + \frac{Q_{ij}\sin^{-1}\left(\frac{Q_{jk}}{\sqrt{Q_{kk}Q_{jj}}}\right)}{2\pi} + \frac{Q_{ij}}{4}\right)\Bigg]$$

Generalization error ReLU SCM

$$\epsilon_g = \frac{1}{2}\left[\sum_{i=1}^{K}\sum_{j=1}^{K}\langle x_i x_j \theta(x_i)\theta(x_j)\rangle - 2\sum_{i=1}^{K}\sum_{m=1}^{M}\langle x_i y_m \theta(x_i)\theta(y_m)\rangle + \sum_{m=1}^{M}\sum_{n=1}^{M}\langle y_m y_n \theta(y_m)\theta(y_n)\rangle\right]$$

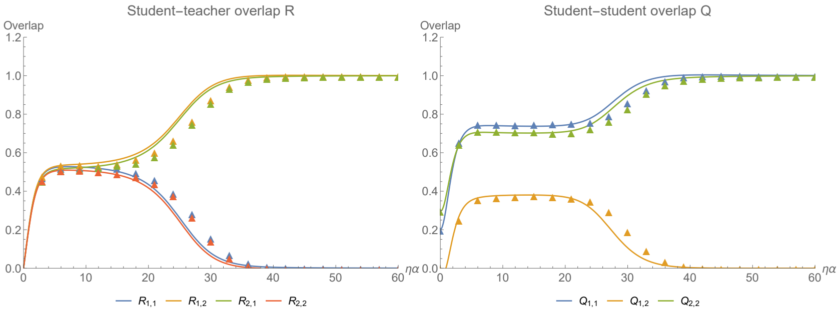
$$\langle uv\theta(u)\theta(v)\rangle = \frac{\sigma_{12}}{4} + \frac{\sqrt{\sigma_{11}\sigma_{22}-\sigma_{12}^2}}{2\pi} + \frac{\sigma_{12}\sin^{-1}\left(\frac{\sigma_{12}}{\sqrt{\sigma_{11}\sigma_{22}}}\right)}{2\pi}$$

Experiment ReLU SCM $M = K = 2$

Teacher SCM with $M = 2$ hidden units and $\boldsymbol{T} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$. Rule is learned by student SCM with $K = 2$ hidden units.
Initial conditions:

$$\boldsymbol{R}(0) = \begin{pmatrix} 0 & 1.2822 * 10^{-3} \\ 1.2822 * 10^{-3} & 0 \end{pmatrix}$$

$$\boldsymbol{Q}(0) = \begin{pmatrix} 0.2 & 0 \\ 0 & 0.3 \end{pmatrix}$$


Student–teacher overlap R
Overlap
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0
10
20
30
40
50
60
ηα
R1,1
R1,2
R2,1
R2,2
Student–student overlap Q
Overlap
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0
10
20
30
40
50
60
ηα
Q1,1
Q1,2
Q2,2

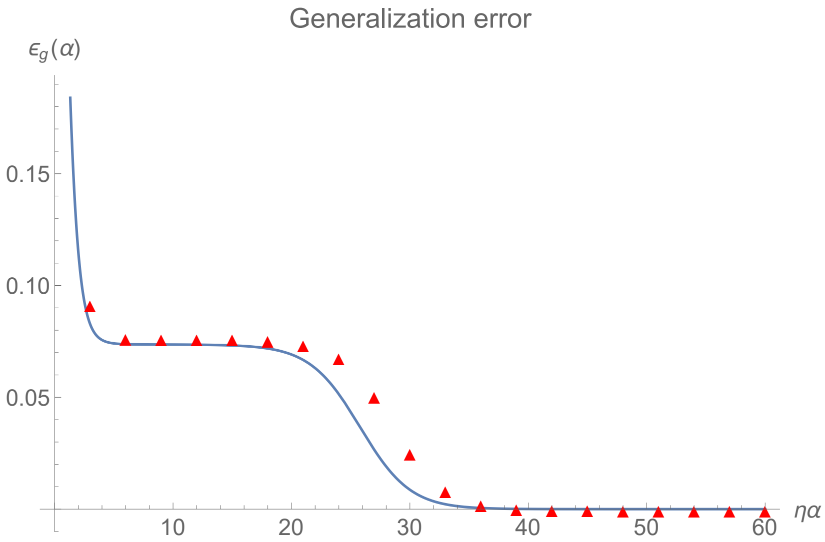
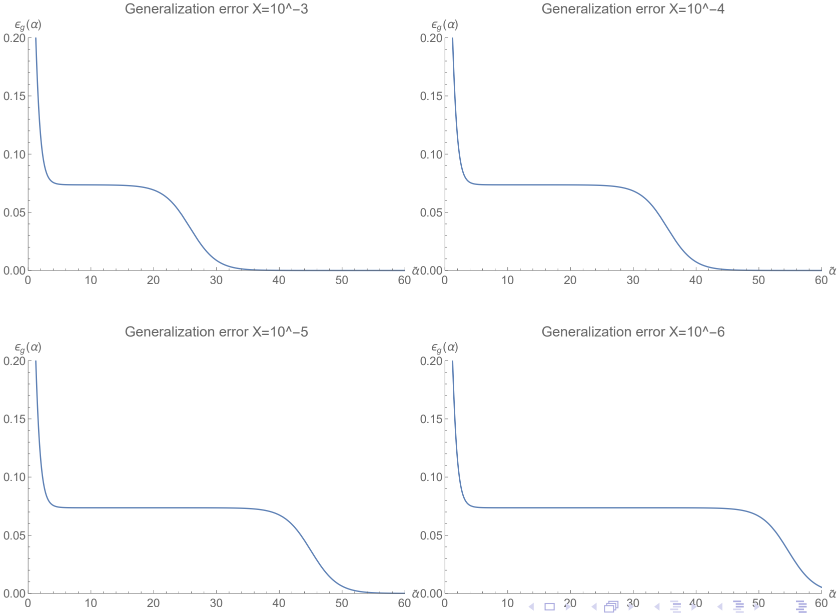


Figure: $\epsilon(\alpha)$ of the ReLU SCM, $K = M = 2$.

Plateau length increases logarithmically with the deviation from symmetry X.



Symmetric plateau

Fixed point associated with plateau:

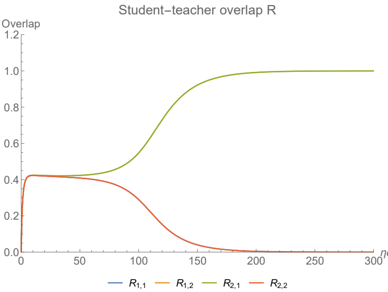
$$\begin{pmatrix} R_{11} \\ R_{12} \\ R_{21} \\ R_{22} \\ Q_{11} \\ Q_{12} \\ Q_{22} \end{pmatrix}_{\text{fix}} \approx \begin{pmatrix} 0.5246 \\ 0.5246 \\ 0.5246 \\ 0.5246 \\ 0.7178 \\ 0.3830 \\ 0.7178 \end{pmatrix}$$

$$\boldsymbol{\lambda} = \{-1.3583, -0.9568, -0.6443, -0.4399, 0.2392, -0.2308, -0.0049\},$$

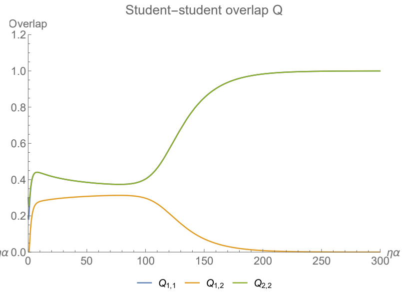
and the fifth eigenvector $\boldsymbol{u}_5$ corresponding to the eigenvalue λ_5 is:

$$\boldsymbol{u}_5 = (0.5, -0.5, -0.5, 0.5, 0, 0, 0)^T$$

Erf SCM $K = M = 2$



(a) $R_{in}(\alpha)$



(b) $Q_{ik}(\alpha)$

Fixed point associated with plateau:

$$\boldsymbol{x}_{\mathsf{fix}} = \begin{pmatrix} R_{11} \\ R_{12} \\ R_{21} \\ R_{22} \\ Q_{11} \\ Q_{12} \\ Q_{22} \end{pmatrix}_{\mathsf{fix}} = \begin{pmatrix} 0.4082 \\ 0.4082 \\ 0.4082 \\ 0.4082 \\ 0.3333 \\ 0.3333 \\ 0.3333 \end{pmatrix} .$$

$$\boldsymbol{\lambda} = \{-1.4682, -0.6922, -0.6108, -0.4086, 0.0682, -0.0192, 0.0103\} .$$

Students are identical in the fixed point. Dominant direction again $\boldsymbol{u}_5 = (0.5, -0.5, -0.5, 0.5, 0, 0, 0)^T$.
$\boldsymbol{u}_7 = (-0.28, -0.28, 0.28, 0.28, -0.58, 0, 0.58)^T$.

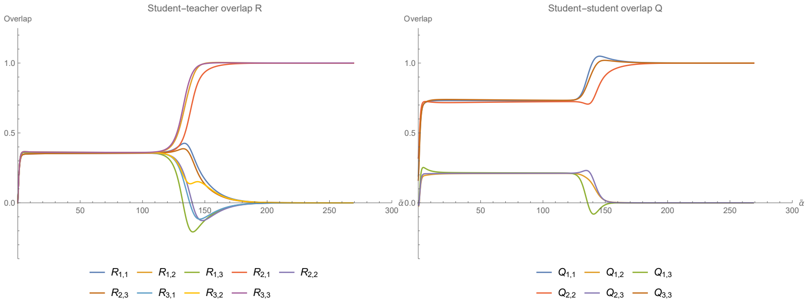
K=M=3

$$\boldsymbol{T} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

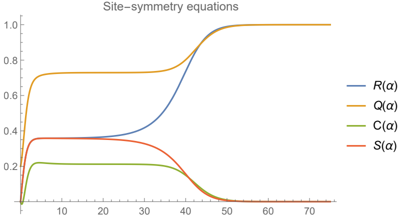
$$R_{in}(0) = U[0, 10^{-12}]$$

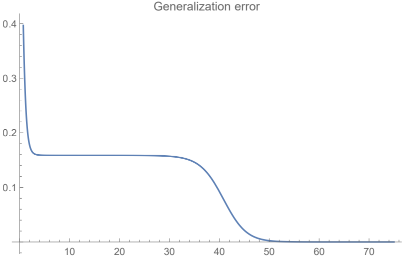
$$Q_{ii}(0) = U[0.1, 0.5] \quad Q_{ij}(0) = U[0, 10^{-12}]$$

ReLU SCM $K = M = 3$


Student–teacher overlap R
Overlap
1.0
0.5
0.0
50
100
150
200
250
300
α̃
R1,1 R1,2 R1,3 R2,1 R2,2
R2,3 R3,1 R3,2 R3,3
Student–student overlap Q
Overlap
1.0
0.5
0.0
50
100
150
200
250
300
α̃
Q1,1 Q1,2 Q1,3
Q2,2 Q2,3 Q3,3

Site symmetry equations



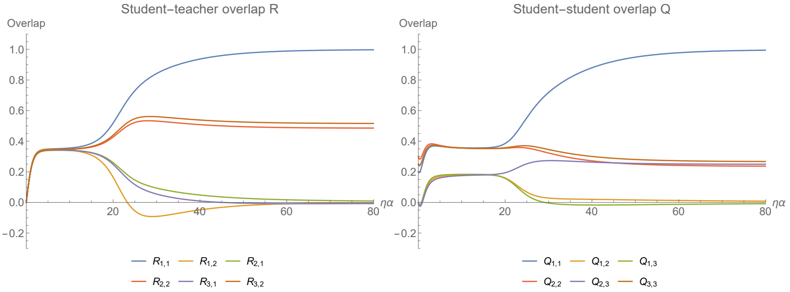


Different learning scenarios

So far, only *realizable* scenarios were studied, i.e. $K = M$.

- $K > M$ (*overrealizable*): more complexity available than needed to represent the rule.
- $K < M$ (*unrealizable*): Rule cannot be represented by the student.

$K = 3, M = 2$, ReLU SCM



$T = \delta_{nm}$, $R_{11} = 10^{-3}$, $Q_{11} = 0.2$, $Q_{22} = 0.3$, $Q_{33} = 0.25$ Two of the student hidden units specialize to one teacher hidden unit.

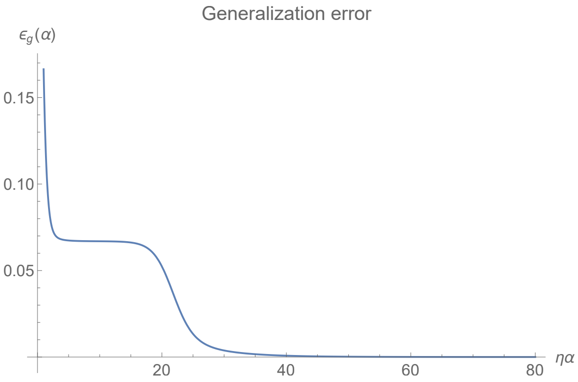


Figure: Generalization error for the overrealizable scenario $(K = 3, M = 2)$

$K = 3, M = 2$, Erf SCM

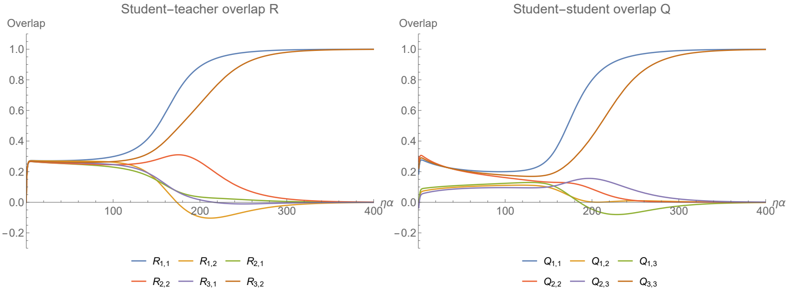


Figure: Two-layer Erf online gradient descent learning in the overrealizable scenario using a student with $K = 3$ and and isotropic teacher with $M = 2$.

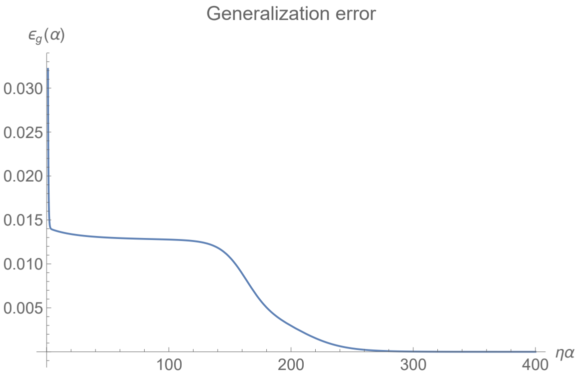


Figure: Generalization error for the overrealizable scenario with a Erf network $(K = 3, M = 2)$

$K = 2, M = 3$, ReLU SCM

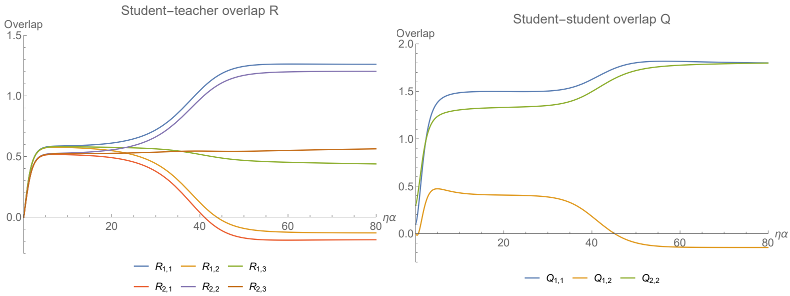


Figure: Online gradient descent learning for an unrealizable case when the rule is a teacher network with $M = 3$ ReLU hidden units and the student is a network with $K = 2$ ReLU hidden units.

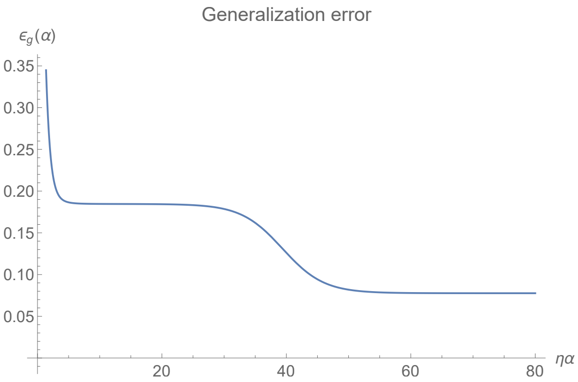


Figure: Generalization error for the overrealizable scenario $(K = 2, M = 3)$.

$$\epsilon_g(\alpha \to \infty) > 0$$

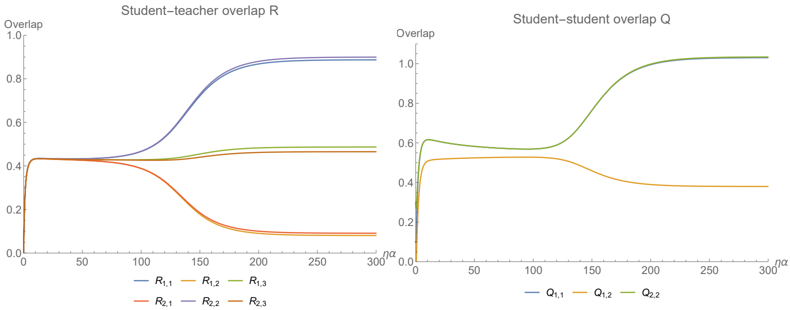


Figure: Online gradient descent learning for an unrealizable case when the rule is an Erf teacher network with $M = 3$ hidden units and the student is an Erf network with $K = 2$ hidden units.

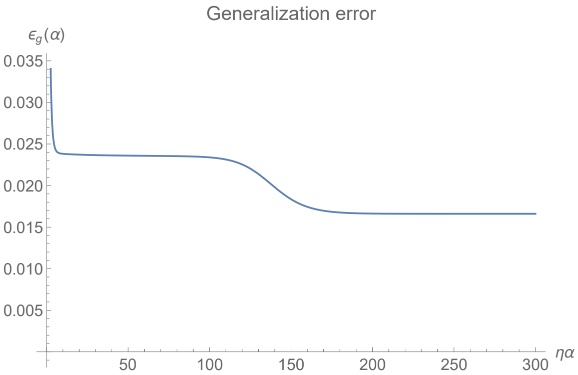


Figure: Generalization error for the unrealizable case in which an Erf student with $K = 2$ learns an Erf teacher with $M = 3$.

Future research

- Include η^2 term.
- Learning dynamics of additional schemes or adaptations, learning rate adaptation.
- Other types of architectures.
- Time-dependent rule.