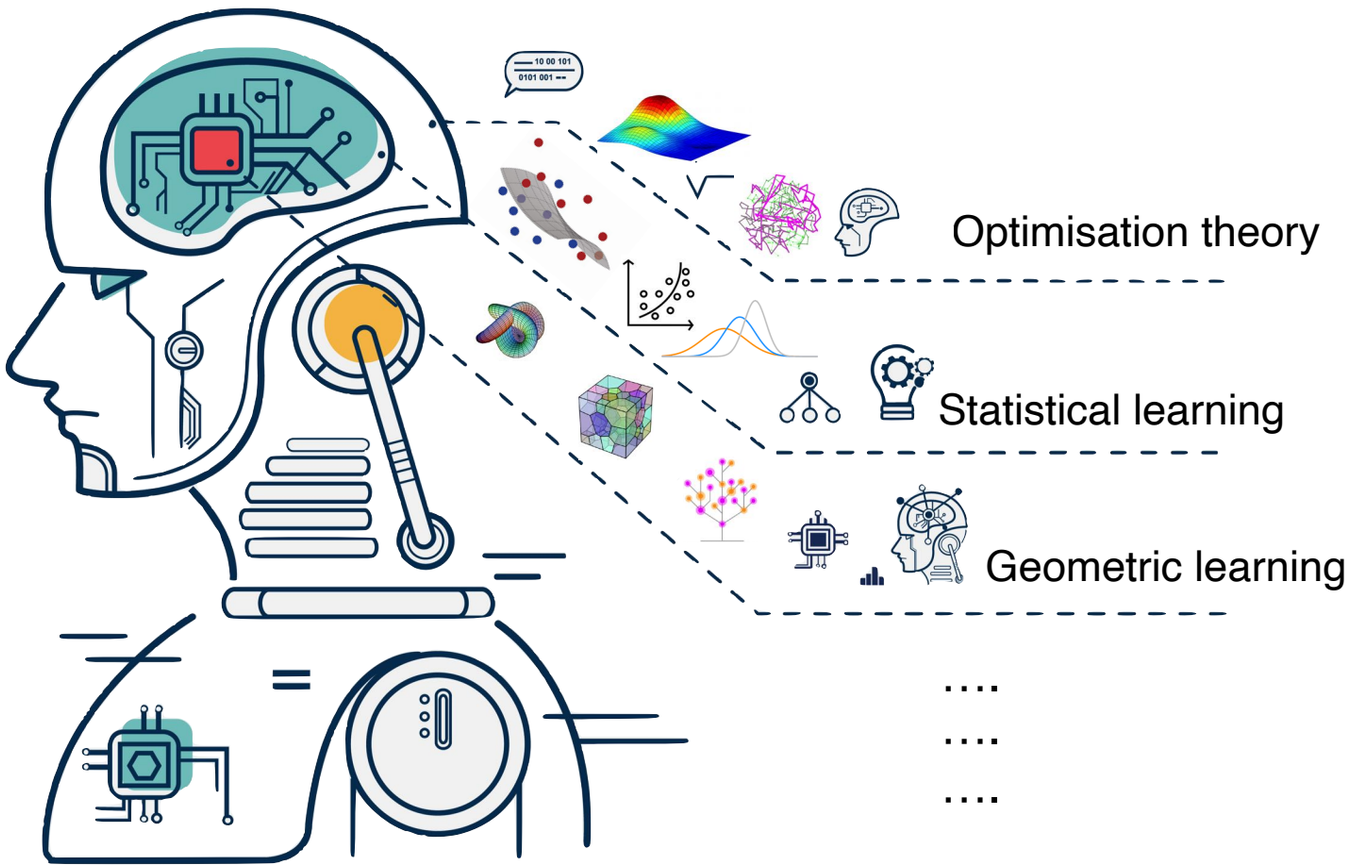
Machine learning:
from mammal's brain to statistical mechanics

Elena Agliari

Dipartimento di Matematica, Sapienza Università di Roma
Istituto Nazionale di Alta Matematica, Roma

Spring School "Bruno Touschek"
Laboratori Nazionali Frascati
16 May 2024

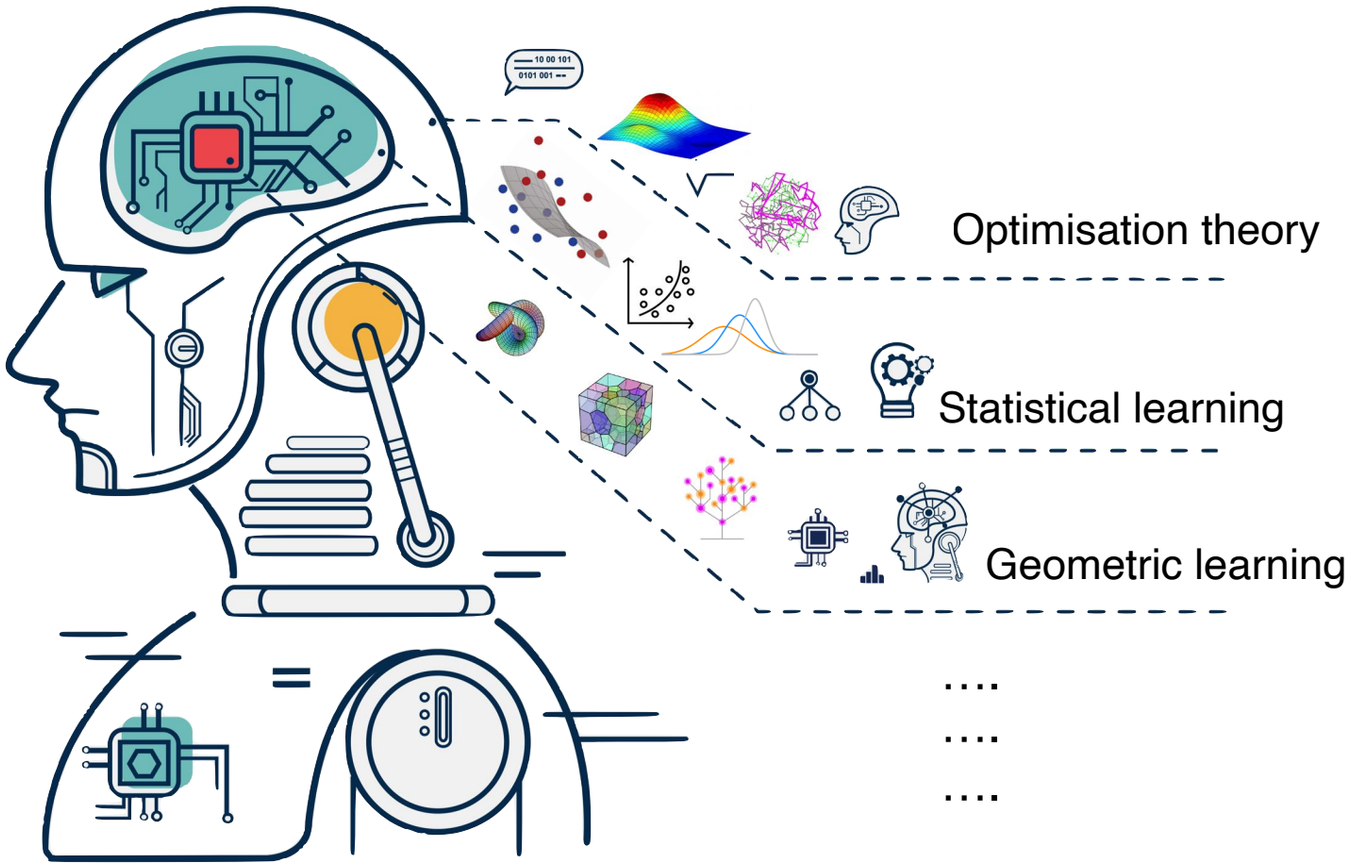
Hard sciences for machine learning



Mathematical control
- design optimal architecture
- estimate hyperparameters
- check dataset size
- ....

→ Sustainability & Interpretability

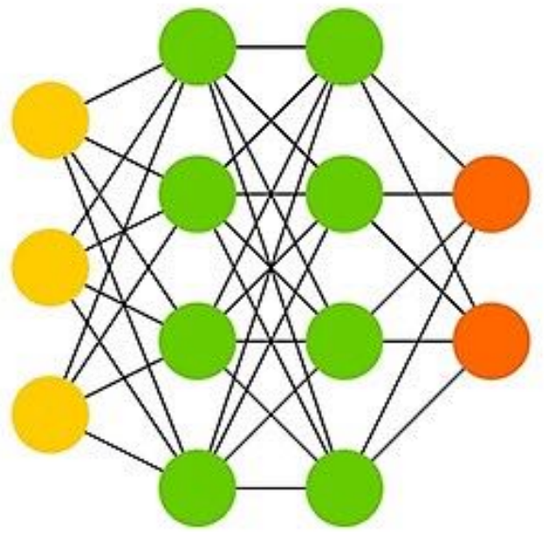
Hard sciences for machine learning



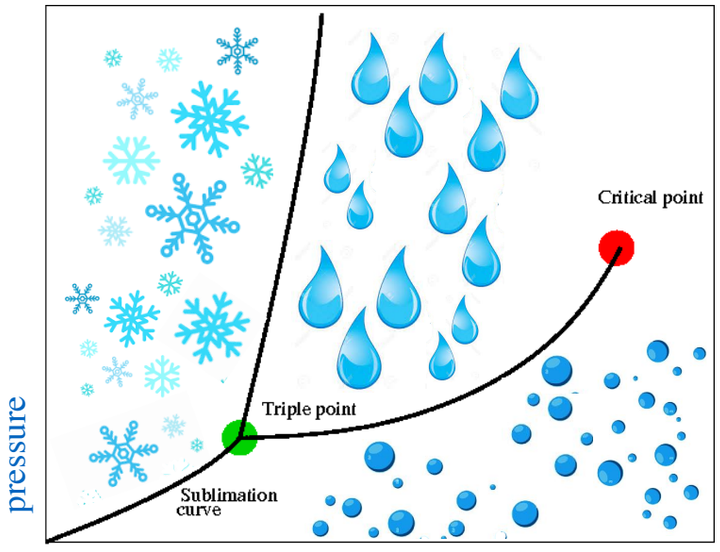
Mathematical control
- design optimal architecture
- estimate hyperparameters
- check dataset size
- ….

→ Sustainability & Interpretability

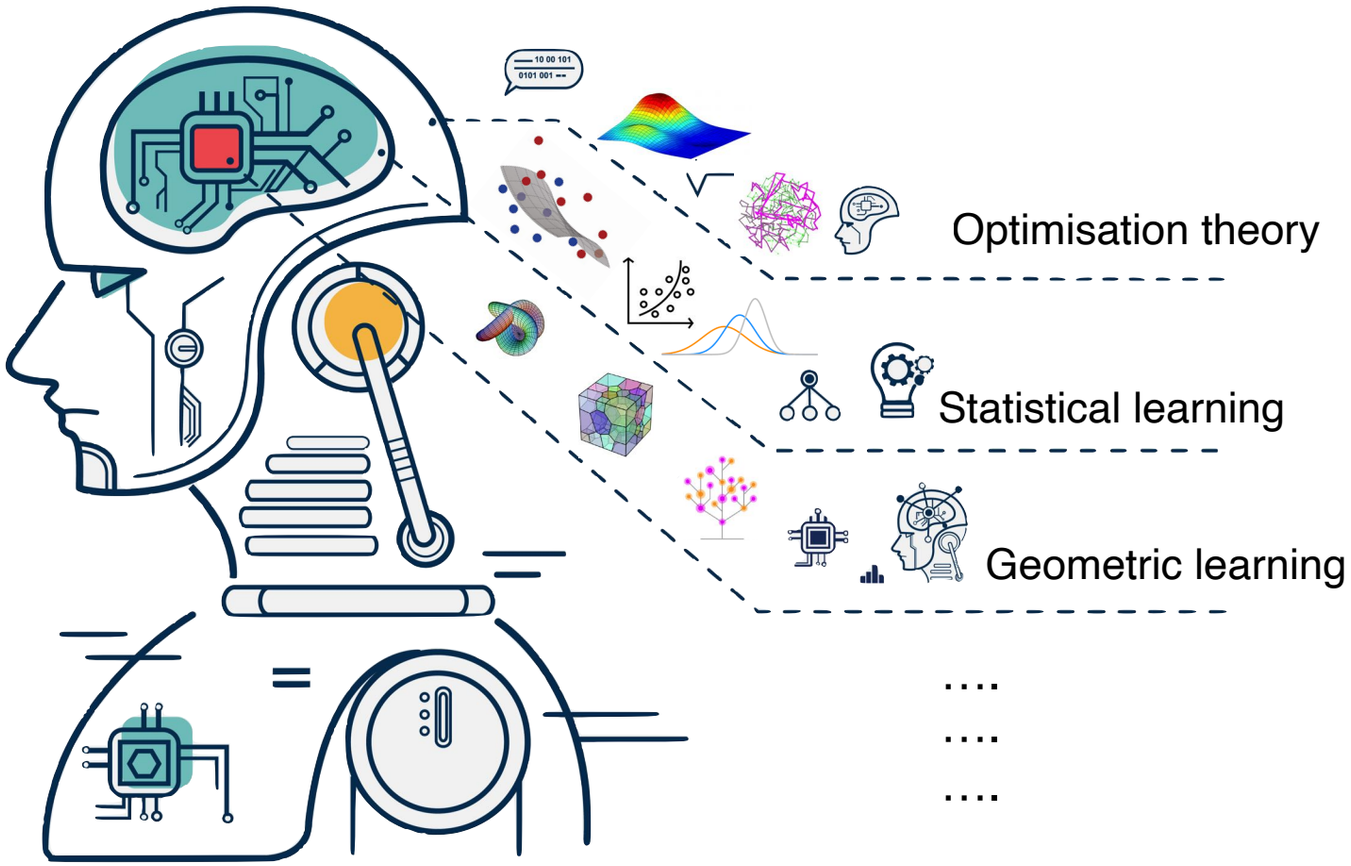
Statistical-mechanics perspective



Information-processing capability as emerging collective behavior



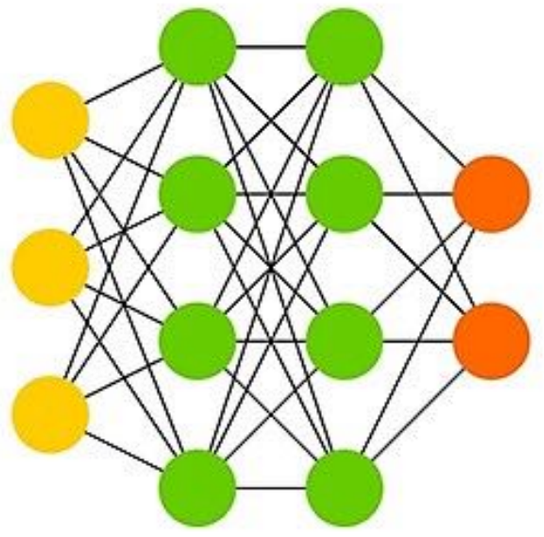
Hard sciences for machine learning



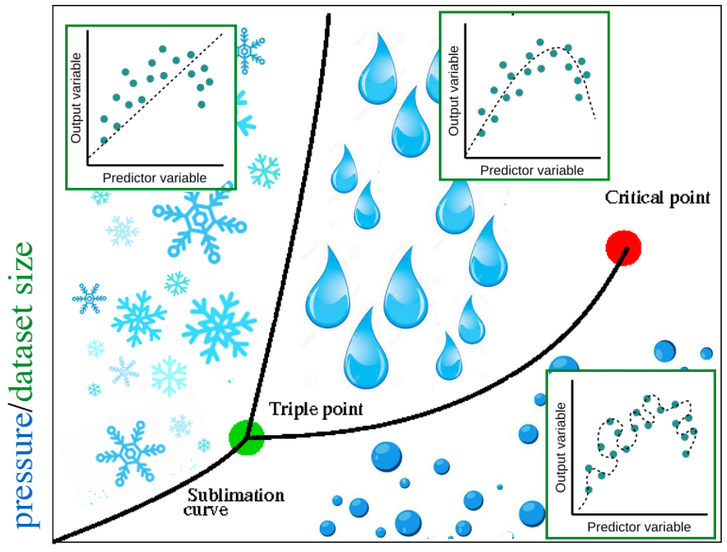
Mathematical control
- design optimal architecture
- estimate hyperparameters
- check dataset size
- ….

→ Sustainability & Interpretability

Statistical-mechanics perspective



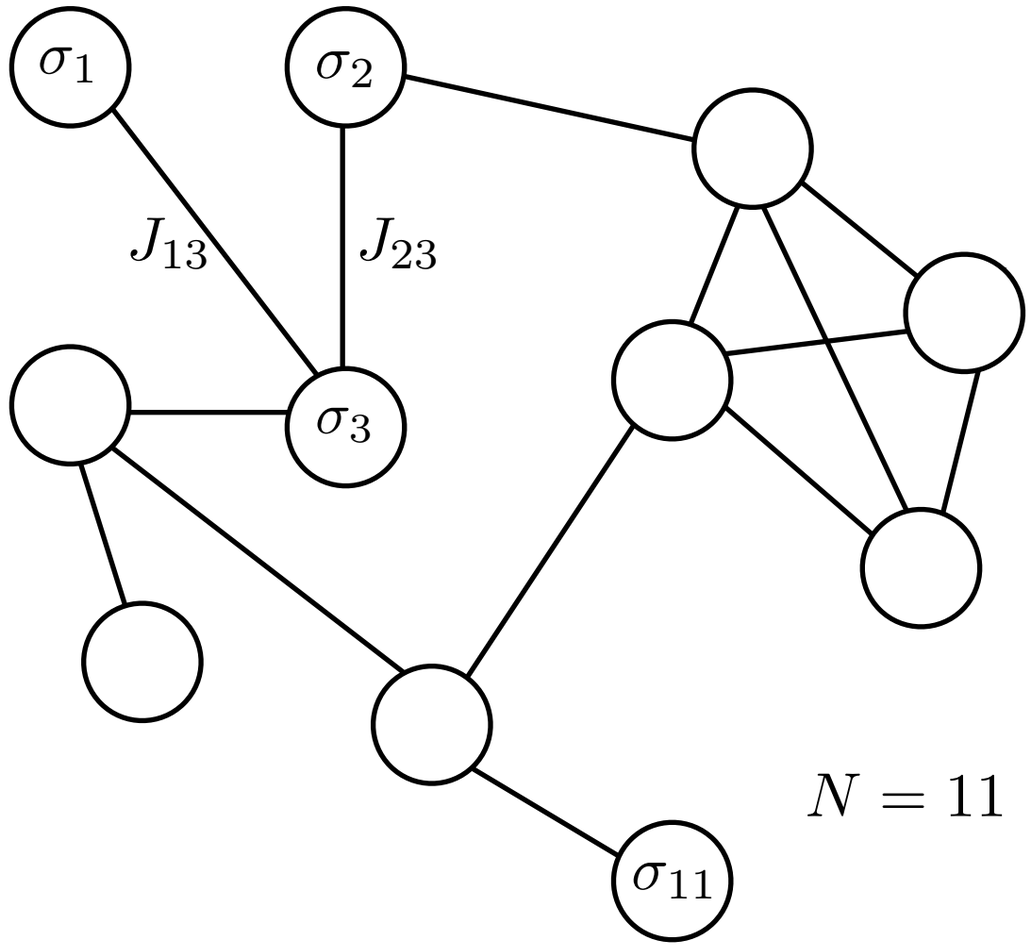
Information-processing capability as emerging collective behaviour



Neural networks: a biology-inspired introduction

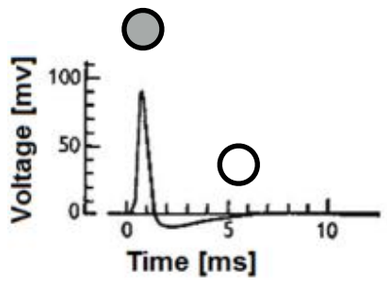
Nodes as neurons with activity $\boldsymbol{\sigma} = \{\sigma_i\}_{i=1,\dots,N}$

Links as synapses with efficacy $\boldsymbol{J} = \{J_{ij}\}_{i,j=1,\dots,N}$



Firing/quiescent neurons

$$\boldsymbol{\sigma} \in \{-1, +1\}^N$$

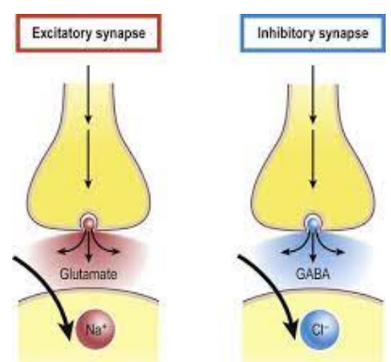


$\{\sigma_i\}$ fast degrees of freedom

Excitatory/Inhibitory synapses

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

$\{J_{ij}\}$ slow degrees of freedom



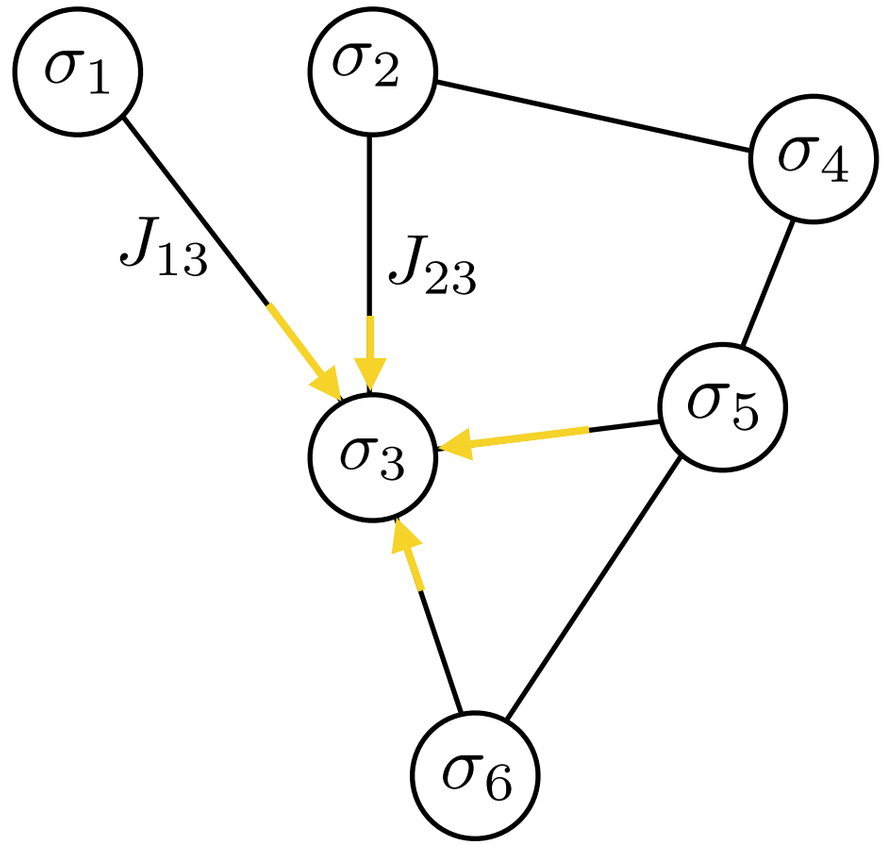
Neural dynamics

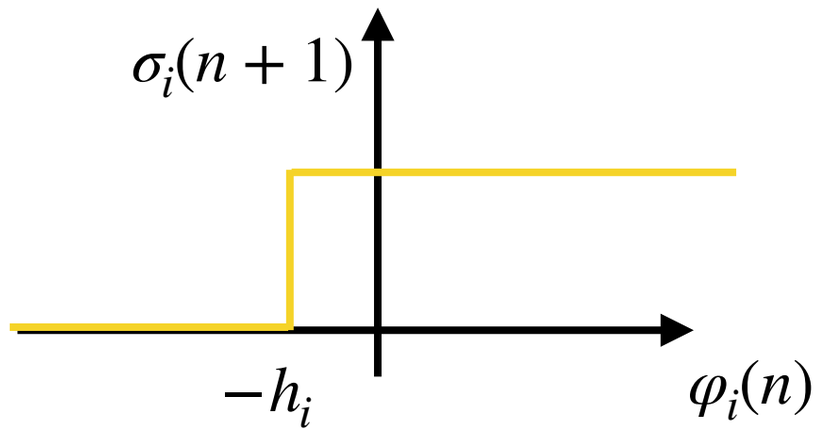
$$\sigma_i(n+1) = \mathrm{sgn}(\varphi_i(n) + h_i)$$

Field acting on *i* at time step *n*

Firing threshold

$$\varphi_i(n) = \sum_{\substack{k=1 \\ k \neq i}}^{N} J_{ik}\sigma_k(n)$$



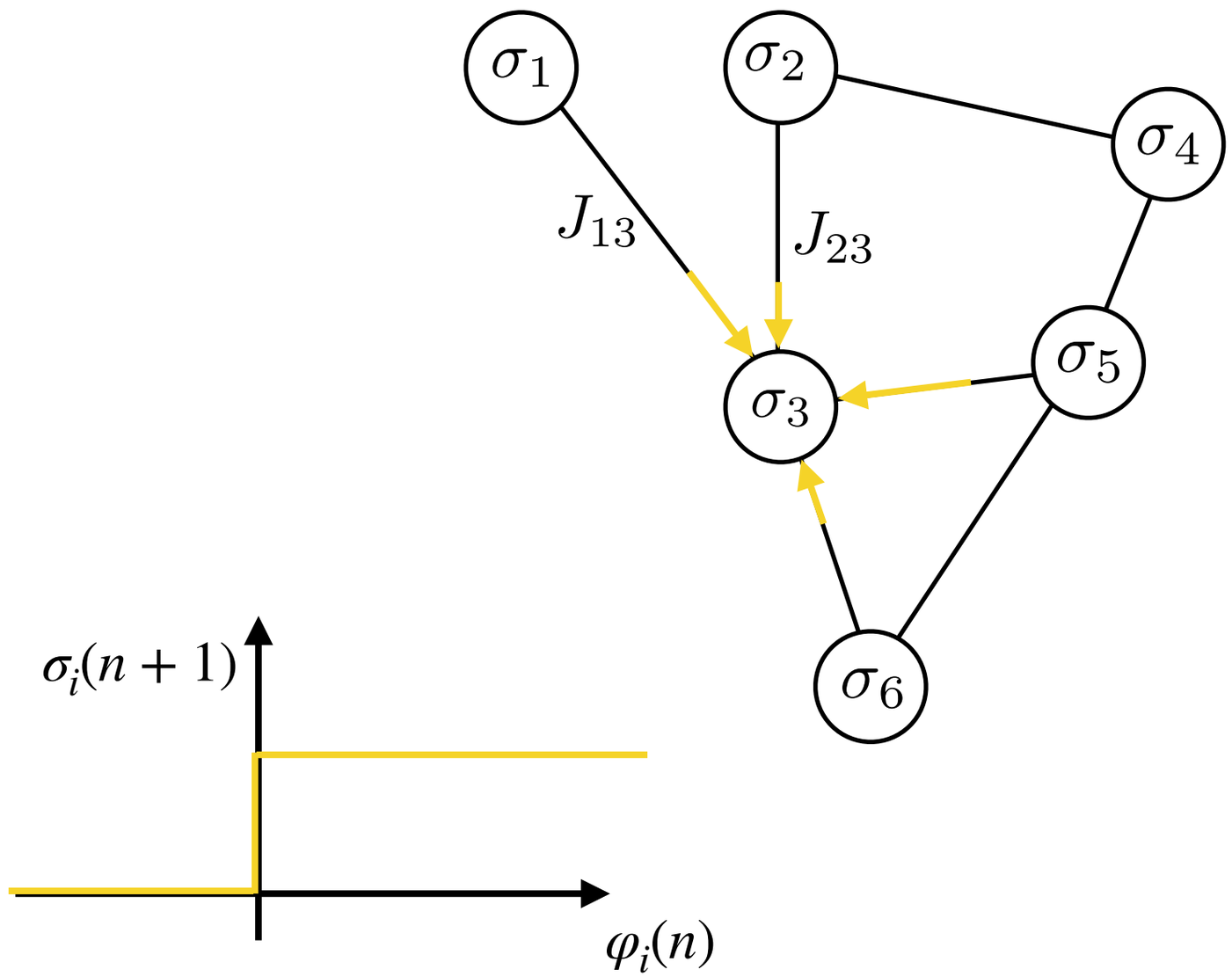


Neural dynamics

$$\sigma_i(n+1) = \mathrm{sgn}(\varphi_i(n))$$

Field acting on *i*
at time step *n*

$$\varphi_i(n) = \sum_{\substack{k=1 \\ k \neq i}}^{N} J_{ik}\sigma_k(n)$$



Neural dynamics

$$\sigma_i(n+1) = \mathrm{sgn}(\varphi_i(n) + \zeta_i T)$$

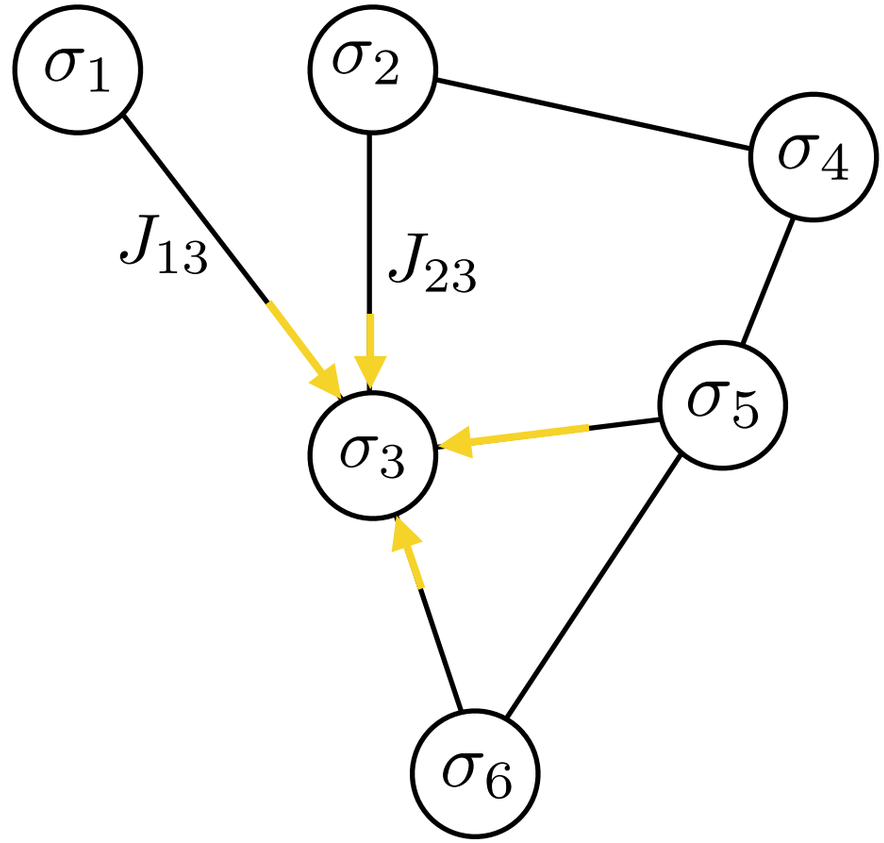
Field acting on *i* at time step *n*

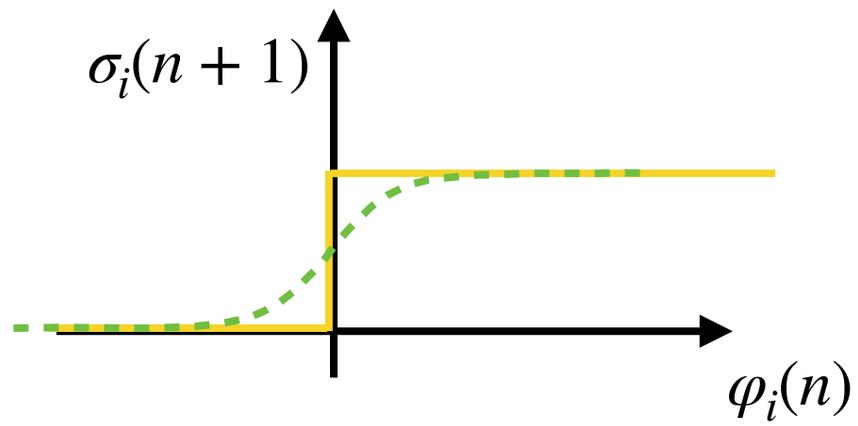
Stochasticity

ζ_i i.i.d. r.v.'s

$T \geq 0$

$$\varphi_i(n) = \sum_{\substack{k=1 \\ k \neq i}}^{N} J_{ik}\sigma_k(n)$$





Neural dynamics

$$\sigma_i(n+1) = \mathrm{sgn}(\varphi_i(n) + \zeta_i T)$$

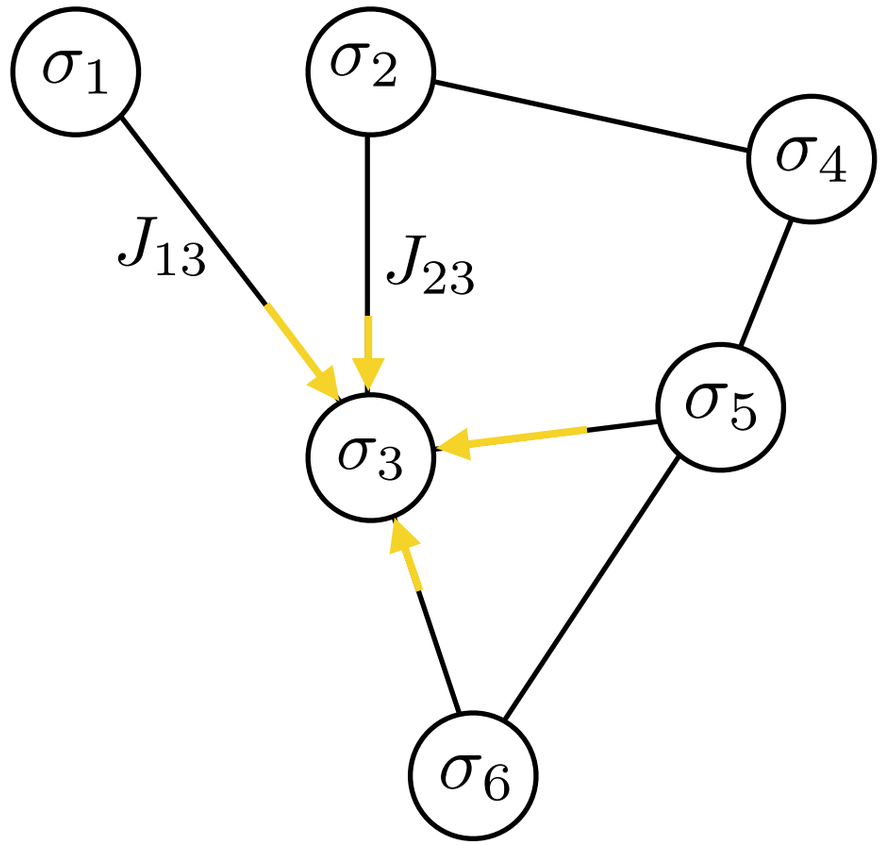
Field acting on *i* at time step *n*

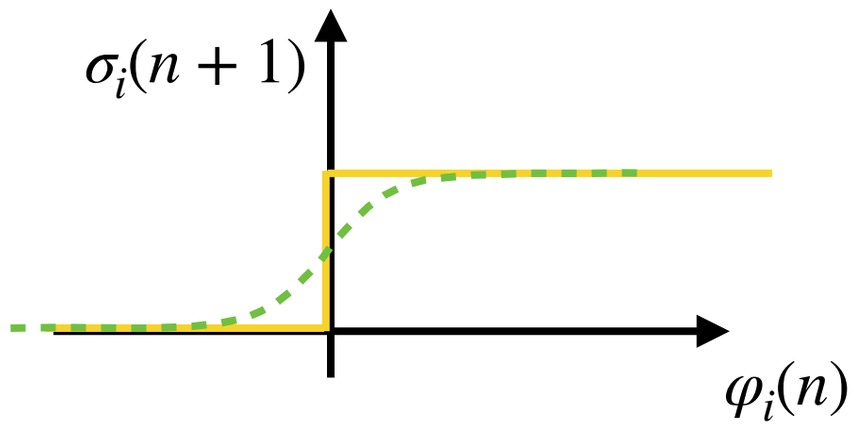
Stochasticity

ζ_i i.i.d. r.v.'s

$T \geq 0$

$$\varphi_i(n) = \sum_{\substack{k=1 \\ k \neq i}}^{N} J_{ik}\sigma_k(n)$$





If $T = 0$, as long as $\boldsymbol{J}$ symmetric,

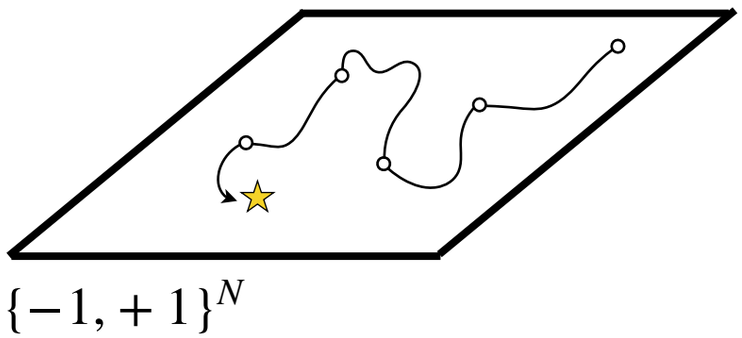
$\mathscr{H}_{N,\boldsymbol{J}}(\boldsymbol{\sigma}) = -\frac{1}{2}\sum_{i,j}\sigma_i J_{ij}\sigma_j$ is a Lyapunov function

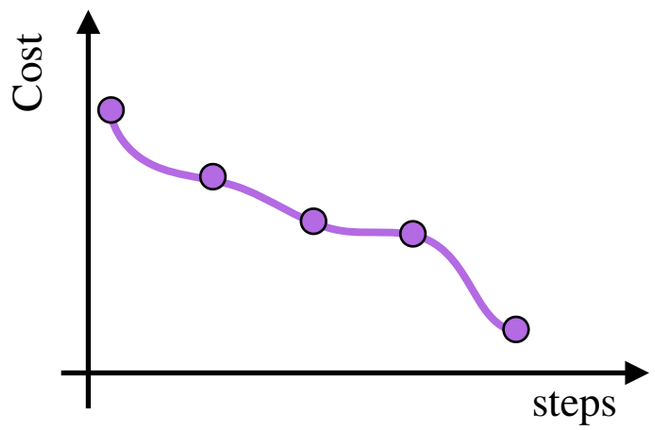
If $T > 0$, as long as $\boldsymbol{J}$ symmetric and $\mathrm{Cum}(\zeta) = [1 + \tanh(\zeta)]/2$,

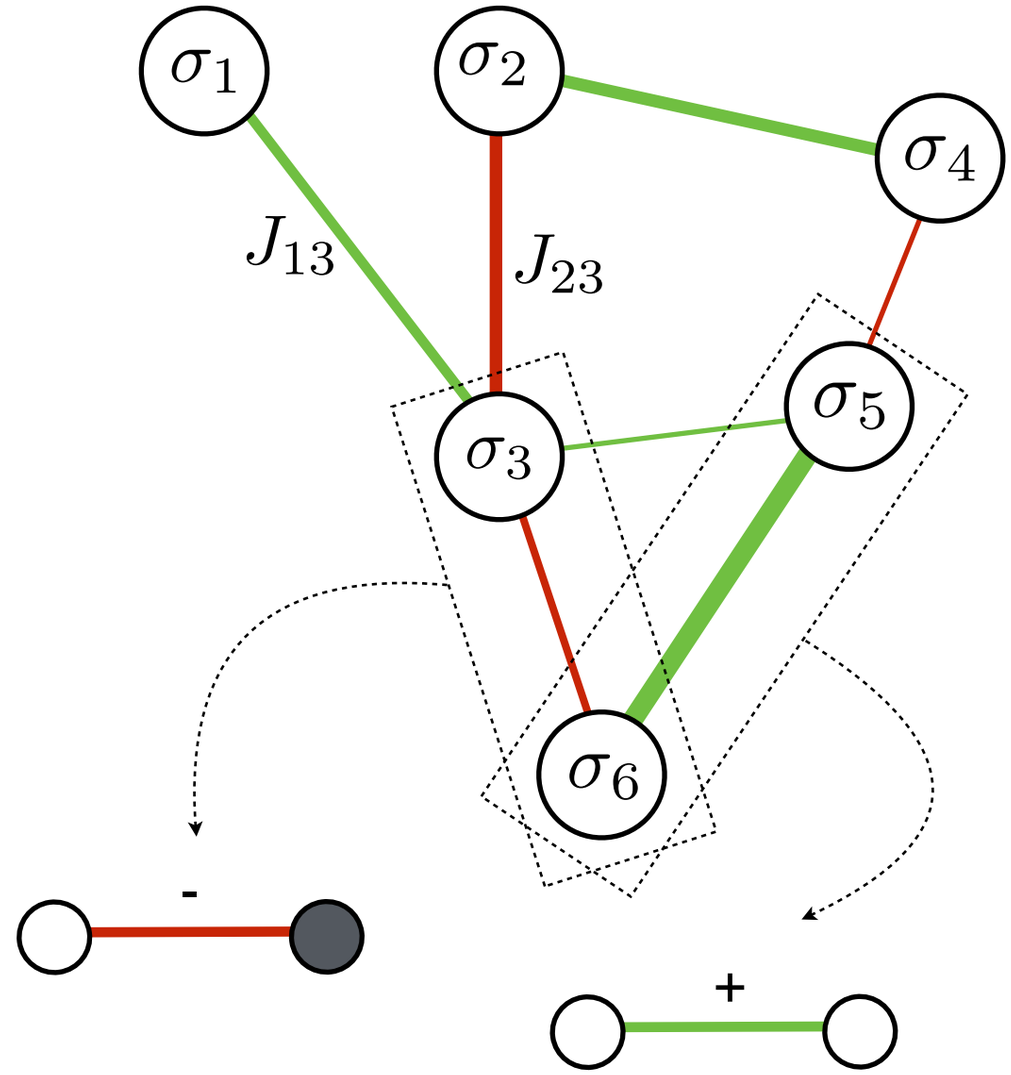
$\mathscr{P}_{N,\beta,\boldsymbol{J}}(\boldsymbol{\sigma}) \propto e^{-\beta\mathscr{H}_{N,\boldsymbol{J}}(\boldsymbol{\sigma})}$, $\beta := T^{-1}$

If $T = 0$, as long as $\boldsymbol{J}$ symmetric,

$\mathscr{H}_{N,\boldsymbol{J}}(\boldsymbol{\sigma}) = -\frac{1}{2}\sum_{i,j}\sigma_i J_{ij}\sigma_j$ is a Lyapunov function



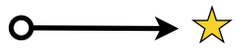




Fixed point dynamics

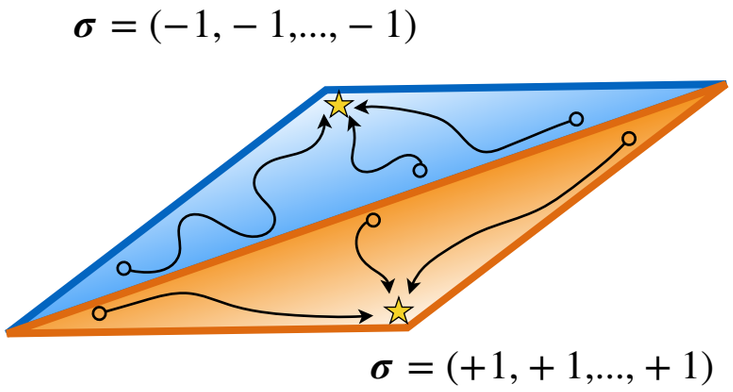
$\boldsymbol{\sigma}^* = \operatorname{argmin} \mathscr{H}_{N,\boldsymbol{J}}(\boldsymbol{\sigma})$

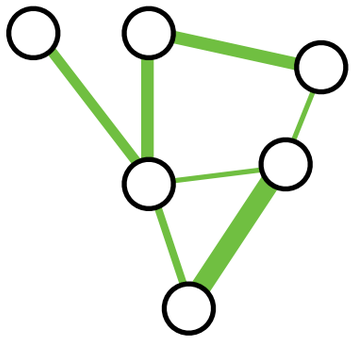
$\boldsymbol{\sigma}_0 \to \boldsymbol{\sigma}^*(\boldsymbol{\sigma}_0, \boldsymbol{J})$

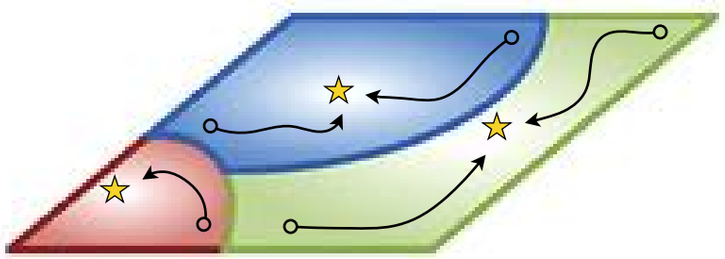


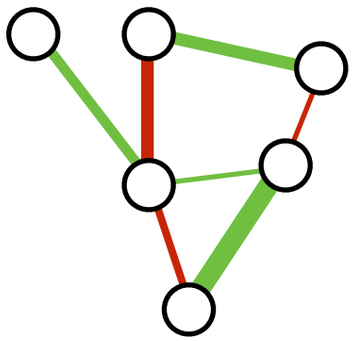
If $T = 0$, as long as $\boldsymbol{J}$ symmetric,

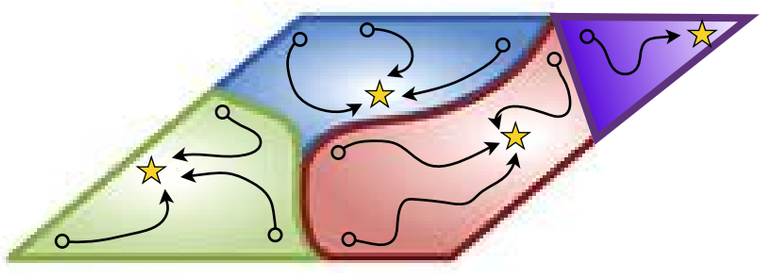
$\mathscr{H}_{N,\boldsymbol{J}}(\boldsymbol{\sigma}) = -\frac{1}{2}\sum_{i,j}\sigma_i J_{ij}\sigma_j$ is a Lyapunov function

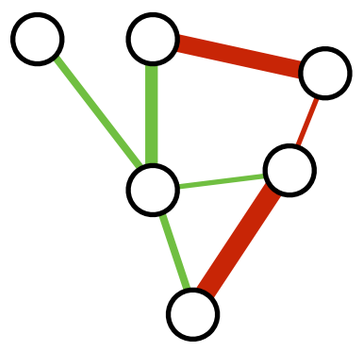


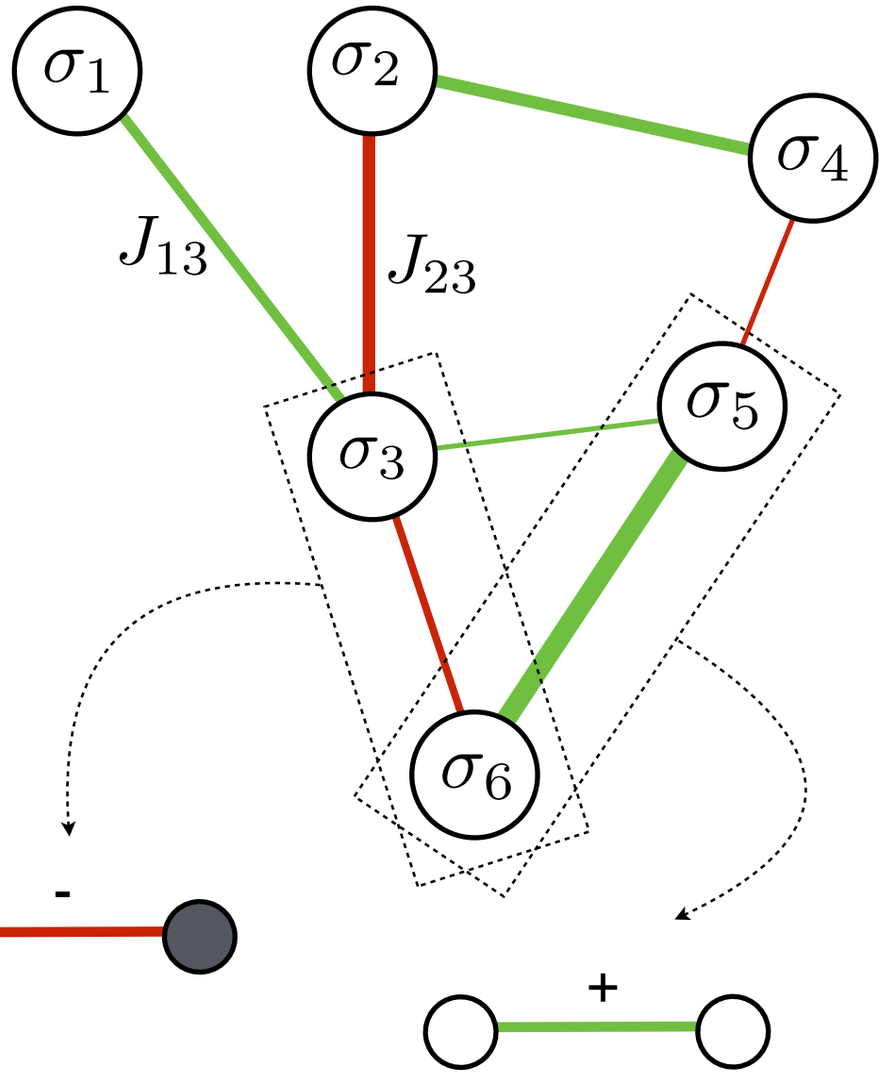












Fixed point dynamics

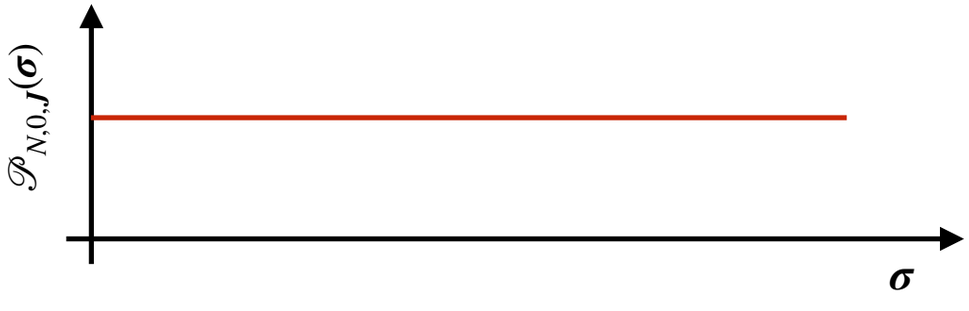
$\boldsymbol{\sigma}^* = \operatorname{argmin} \mathscr{H}_{N,\boldsymbol{J}}(\boldsymbol{\sigma})$

$\boldsymbol{\sigma}_0 \to \boldsymbol{\sigma}^*(\boldsymbol{\sigma}_0, \boldsymbol{J})$

$$\sigma_i(n+1) = \mathrm{sgn}(\varphi_i(n) + \zeta_i T)$$

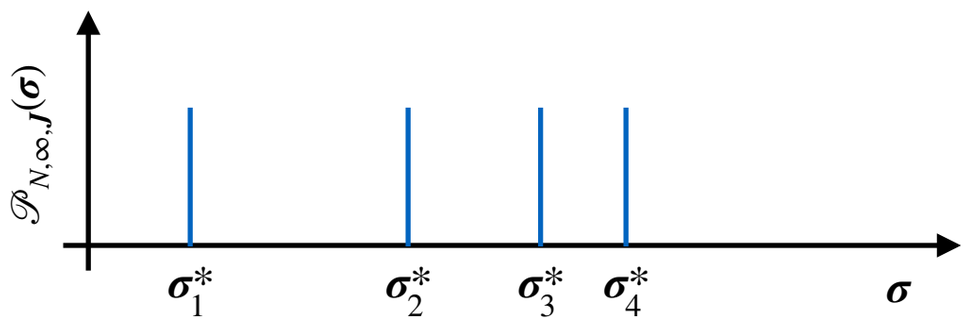
If $T > 0$, as long as $\boldsymbol{J}$ symmetric and $\mathrm{Cum}(\zeta) = [1 + \tanh(\zeta)]/2$,

$$\mathscr{P}_{N,\beta,\boldsymbol{J}}(\boldsymbol{\sigma}) \propto e^{-\beta \mathscr{H}_{N,\boldsymbol{J}}(\boldsymbol{\sigma})},\ \beta := T^{-1}$$



- Noisy limit $T \to \infty$

Fully random dynamics $\sigma_i(n+1) = \pm 1$



- Noiseless limit $T \to 0$

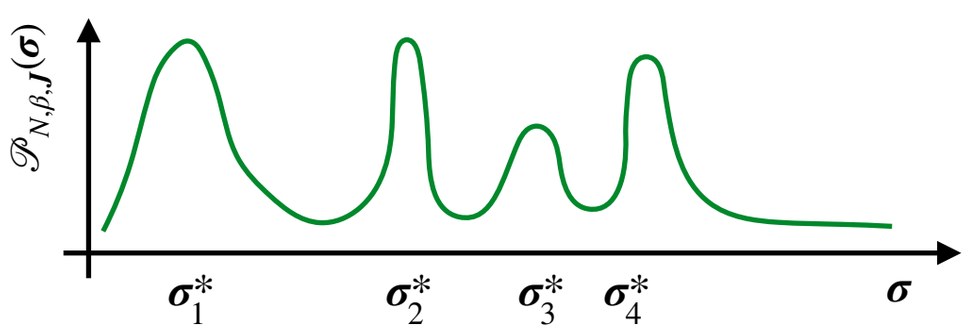
Deterministic dynamics $\sigma_i(n+1) = \mathrm{sgn}(\varphi_i(n))$

$\mathscr{P}_{N,\beta,\boldsymbol{J}}(\boldsymbol{\sigma})$ delta-peaked at $\boldsymbol{\sigma}^*$

Fixed points dynamics

$$\boldsymbol{\sigma}^* = \mathrm{argmin}\, \mathscr{H}_{N,\boldsymbol{J}}(\boldsymbol{\sigma})$$

$$\boldsymbol{\sigma}_0 \to \boldsymbol{\sigma}^*(\boldsymbol{\sigma}_0, \boldsymbol{J})$$



- Intermediate regime $T \in \mathbb{R}^+$

$\mathscr{P}_{N,T,\boldsymbol{J}}(\boldsymbol{\sigma})$ peaked at $\boldsymbol{\sigma}^*$

Starting at $\boldsymbol{\sigma}_0$, likely to end up in neighbourhood of $\sigma^*(\boldsymbol{\sigma}_0, \boldsymbol{J})$

The statistical mechanics setting

Mean-field spin-like model: $\boldsymbol{\sigma} \in \{-1, +1\}^N$, $\boldsymbol{J}$ symm

Hamiltonian $\mathscr{H}_{N,\boldsymbol{J}}(\boldsymbol{\sigma}) = -\frac{1}{2}\sum_{i,j=1}^{N} \sigma_i J_{ij} \sigma_j$

Boltzmann-Gibbs $\mathscr{P}_{N,\beta,\boldsymbol{J}}(\boldsymbol{\sigma}) = \frac{e^{-\beta \mathscr{H}_{N,\boldsymbol{J}}(\boldsymbol{\sigma})}}{\mathscr{Z}_{N,\beta,\boldsymbol{J}}}$, with $\mathscr{Z}_{N,\beta,\boldsymbol{J}} = \sum_{\{\boldsymbol{\sigma}\}} \exp\left[-\beta \mathscr{H}_{N,\boldsymbol{J}}(\boldsymbol{\sigma})\right]$, $\beta := T^{-1}$

The statistical mechanics setting

Mean-field spin-like model: $\boldsymbol{\sigma} \in \{-1, +1\}^N$, $\boldsymbol{J}$ symm

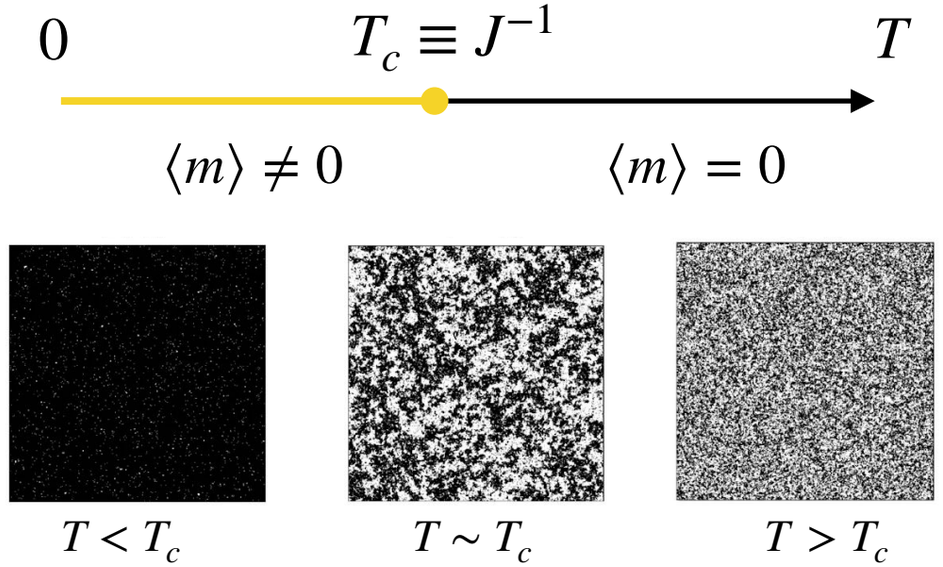
Hamiltonian $\mathscr{H}_{N,\boldsymbol{J}}(\boldsymbol{\sigma}) = -\frac{1}{2}\sum_{i,j=1}^{N} \sigma_i J_{ij} \sigma_j$

Boltzmann-Gibbs $\mathscr{P}_{N,\beta,\boldsymbol{J}}(\boldsymbol{\sigma}) = \frac{e^{-\beta \mathscr{H}_{N,\boldsymbol{J}}(\boldsymbol{\sigma})}}{\mathscr{Z}_{N,\beta,\boldsymbol{J}}}$, with $\mathscr{Z}_{N,\beta,\boldsymbol{J}} = \sum_{\{\boldsymbol{\sigma}\}} \exp\left[-\beta \mathscr{H}_{N,\boldsymbol{J}}(\boldsymbol{\sigma})\right]$, $\beta := T^{-1}$

Curie-Weiss model: $J_{ij} = J > 0, \forall_{i,j}$

(ferromagnet)

$$m_N(\boldsymbol{\sigma}) := \frac{1}{N}\sum_{i=1}^{N} \sigma_i$$



The statistical mechanics setting

Mean-field spin-like model: $\boldsymbol{\sigma} \in \{-1, +1\}^N$, $\boldsymbol{J}$ symm

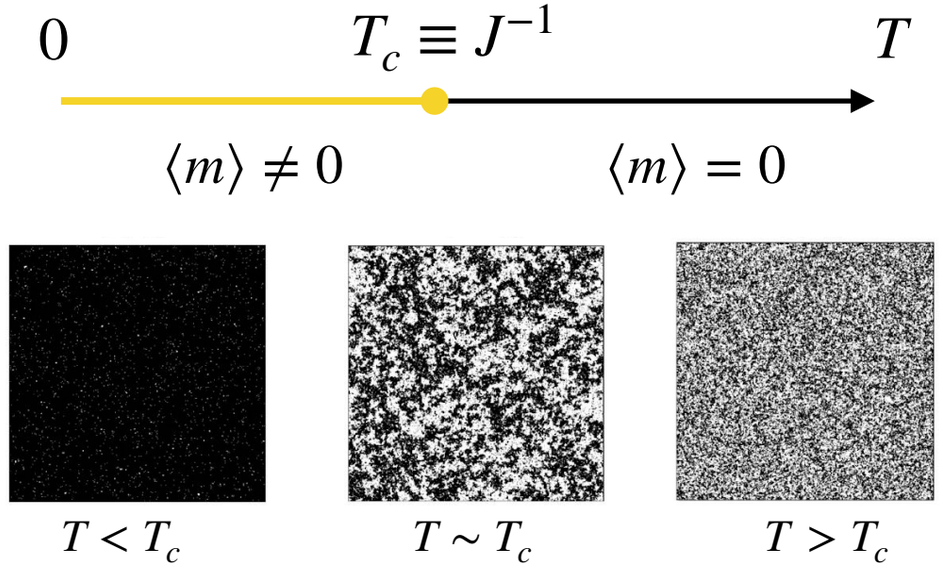
Hamiltonian $\mathscr{H}_{N,\boldsymbol{J}}(\boldsymbol{\sigma}) = -\frac{1}{2}\sum_{i,j=1}^{N} \sigma_i J_{ij} \sigma_j$

Boltzmann-Gibbs $\mathscr{P}_{N,\beta,\boldsymbol{J}}(\boldsymbol{\sigma}) = \frac{e^{-\beta \mathscr{H}_{N,\boldsymbol{J}}(\boldsymbol{\sigma})}}{\mathscr{Z}_{N,\beta,\boldsymbol{J}}}$, with $\mathscr{Z}_{N,\beta,\boldsymbol{J}} = \sum_{\{\boldsymbol{\sigma}\}} \exp\left[-\beta \mathscr{H}_{N,\boldsymbol{J}}(\boldsymbol{\sigma})\right]$, $\beta := T^{-1}$

Curie-Weiss model: $J_{ij} = J > 0, \forall_{i,j}$

(ferromagnet)

$$m_N(\boldsymbol{\sigma}) := \frac{1}{N}\sum_{i=1}^{N} \sigma_i$$



Find explicit expression for *free-energy* $\mathscr{F} := -T \log \mathscr{Z}$

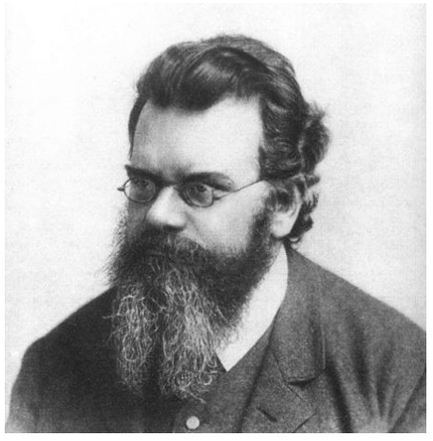
$\rightarrow$ n-th moment of macroscopic observables related to free-energy n-th order derivatives w.r.t. conjugate parameters

Neural networks for retrieval

Mimics *retrieval* capabilities

Pattern recognition
Pattern reconstruction
Denoising
Content addressable memory

CAPTCHA

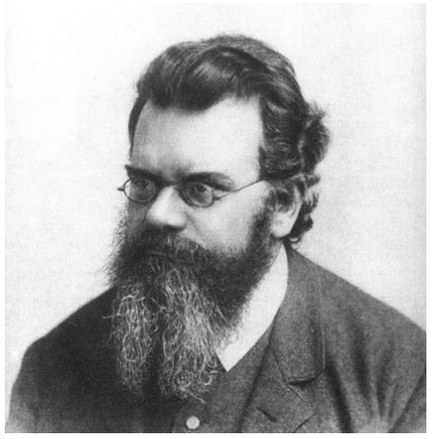


Neural networks for retrieval

Mimics *retrieval* capabilities

Pattern recognition
Pattern reconstruction
Denoising
Content addressable memory

CAPTCHA



Network asked to reconstruct previously learnt vectors, now inputted noisy or incomplete

Set of definite K patterns of length N binary vectors (archetypes)

$$\boldsymbol{\xi}^1, \ldots, \boldsymbol{\xi}^K$$
$$\boldsymbol{\xi}^\mu = (+1, -1, ..., +1) \in \{-1, +1\}^N$$

Patterns

Noisy

input : $\boldsymbol{\sigma}_0$

output : $\boldsymbol{\sigma}^*(\boldsymbol{\sigma}_0)$

Reconstruction

Network asked to reconstruct previously learnt vectors, now inputted noisy or incomplete

Set of definite K patterns of length N binary vectors (archetypes)

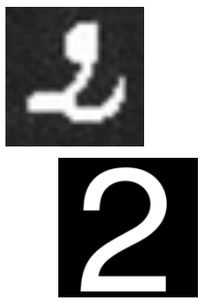
Patterns

$$\boldsymbol{\xi}^1, \ldots, \boldsymbol{\xi}^K$$
$$\boldsymbol{\xi}^\mu = (+1, -1, \ldots, +1) \in \{-1, +1\}^N$$

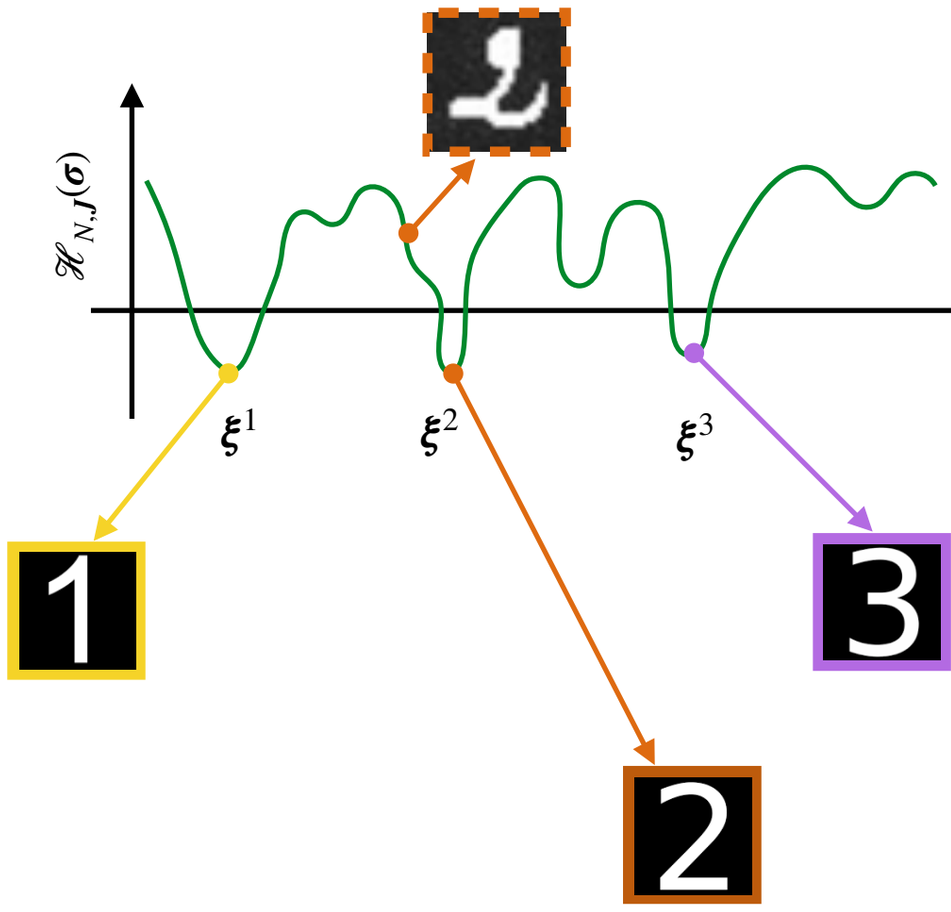
input : $\boldsymbol{\sigma}_0$

output : $\boldsymbol{\sigma}^*(\boldsymbol{\sigma}_0)$

Noisy



Reconstruction

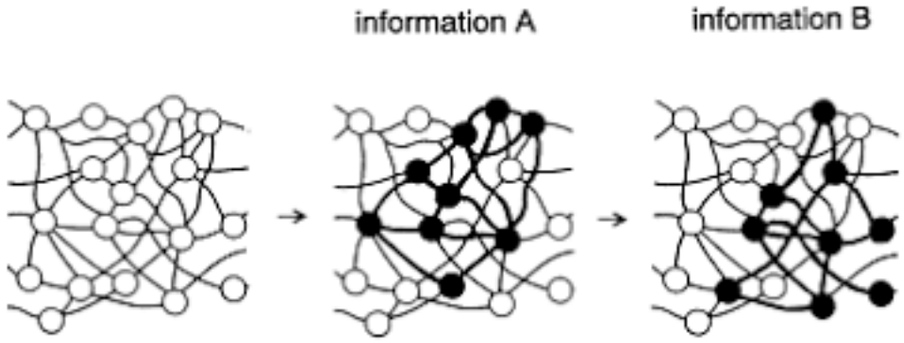


$$\boldsymbol{\sigma}(n=0) = \boldsymbol{\sigma}_0 \rightarrow \boldsymbol{\sigma}^*(\boldsymbol{\sigma}_0) = \lim_{n\to\infty} \boldsymbol{\sigma}(n)\Big|_{\boldsymbol{\sigma}_0}$$

Choose $\boldsymbol{J} = \boldsymbol{J}(\{\xi_i^\mu\}_{i=1,\ldots,N}^{\mu=1,\ldots,K})$ s.t. $\boldsymbol{\xi}^\mu = \underset{\boldsymbol{\sigma}\in\{-1,+1\}^N}{\operatorname{argmin}} \mathscr{H}_{N,\boldsymbol{J}}(\boldsymbol{\sigma})$ for $\mu = 1,\ldots,K$

The Hopfield model (Hopfield model ’82, Pastur&Figotin ’78)

Hebb’s rule $J_{ij} = \frac{1}{N} \sum_{\mu=1}^{K} \xi_i^\mu \xi_j^\mu$

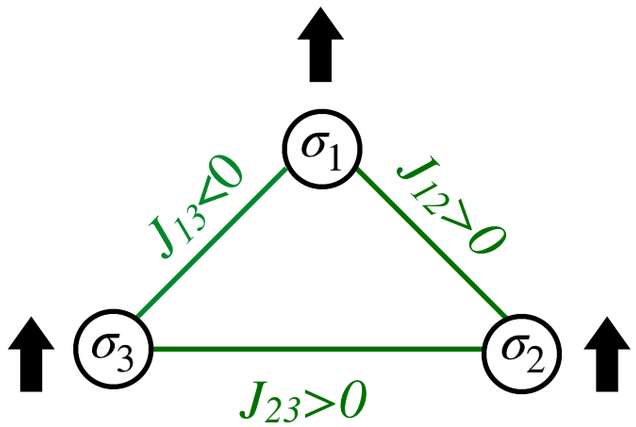


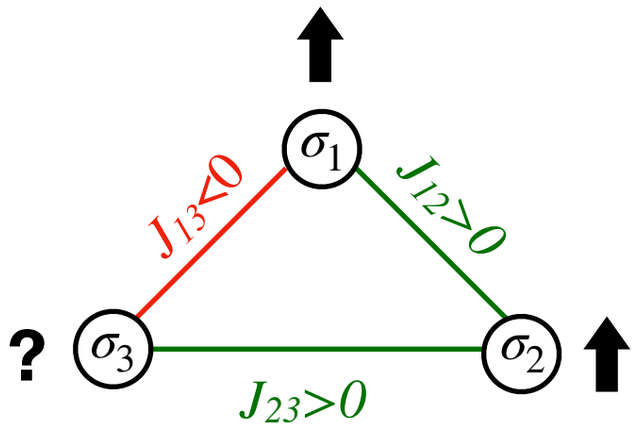
“Neurons out of sync fail to link”

- Biological inspiration
- Low complexity
- Sub-optimal
- Excitatory vs Inhibitory

$$\mathscr{H}_{N,\boldsymbol{J}}(\boldsymbol{\sigma}) = -\sum_{i>j} \sigma_i J_{ij} \sigma_j$$

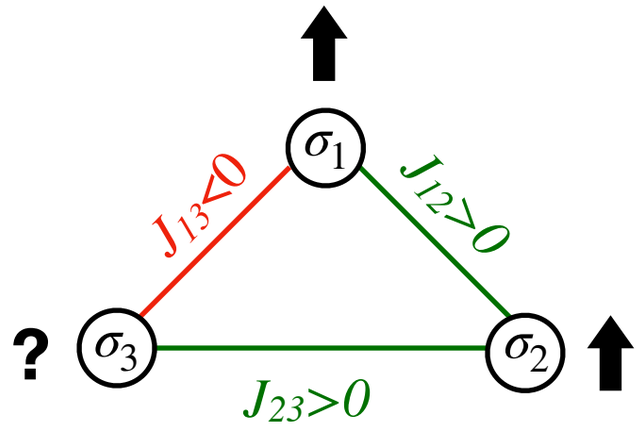
$$J_{ij} = \frac{1}{N} \sum_{\mu=1}^{K} \xi_i^\mu \xi_j^\mu$$





$$\mathscr{H}_{N,\boldsymbol{J}}(\boldsymbol{\sigma}) = -\sum_{i>j} \sigma_i J_{ij} \sigma_j$$

$$J_{ij} = \frac{1}{N} \sum_{\mu=1}^{K} \xi_i^\mu \xi_j^\mu$$



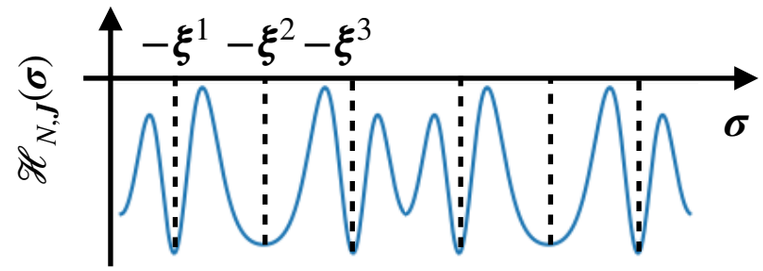
Frustration makes the energy landscape rough
with #local-minima exponentially increasing with N

trivial $K = 0$

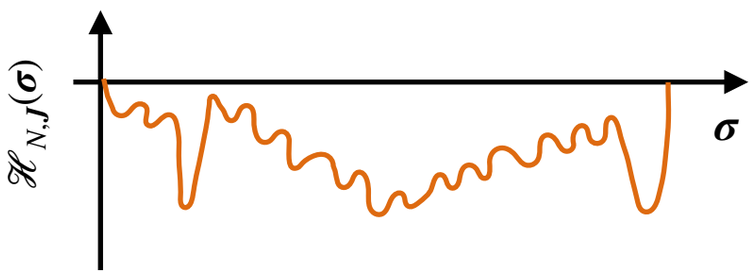
NP-complete combinatorial optimization problem

Load $\alpha := \lim_{N\to\infty} \frac{K}{N},\ \alpha > 0$

simple $K \sim \mathcal{O}(1)$

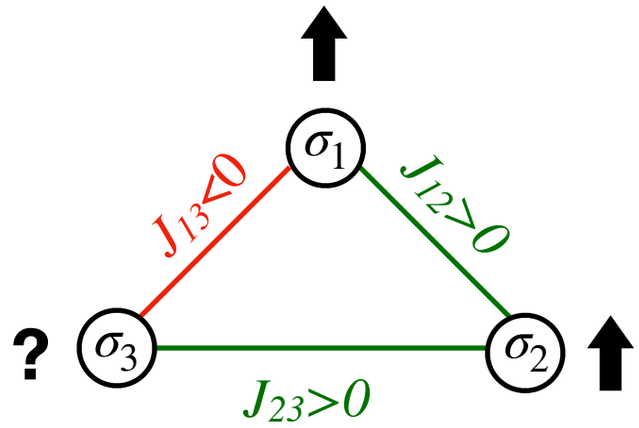


complex $K \sim \mathcal{O}(N)$



$$\mathscr{H}_{N,\boldsymbol{J}}(\boldsymbol{\sigma}) = -\sum_{i>j} \sigma_i J_{ij} \sigma_j$$

$$J_{ij} = \frac{1}{N} \sum_{\mu=1}^{K} \xi_i^\mu \xi_j^\mu$$

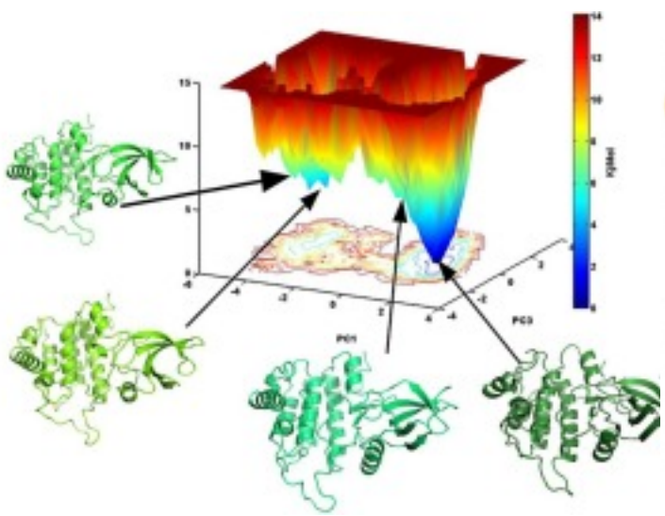


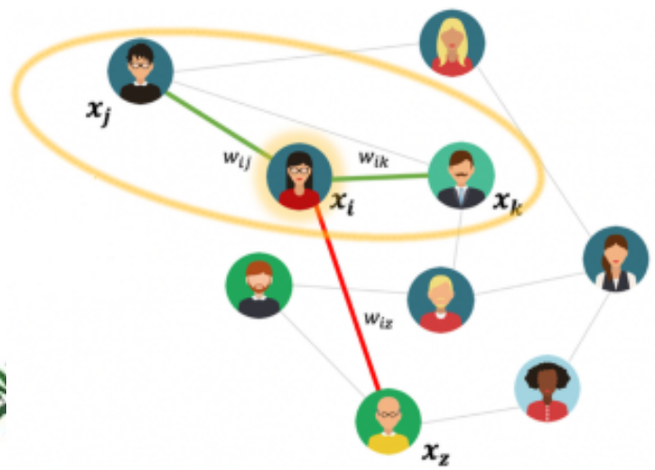
Frustration makes the energy landscape rough
with #local-minima exponentially increasing with N

trivial $K = 0$

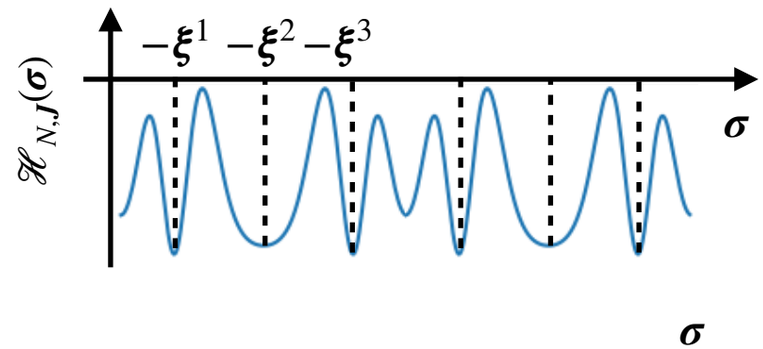
NP-complete combinatorial optimization problem

Load $\alpha := \lim_{N\to\infty} \frac{K}{N},\ \alpha > 0$

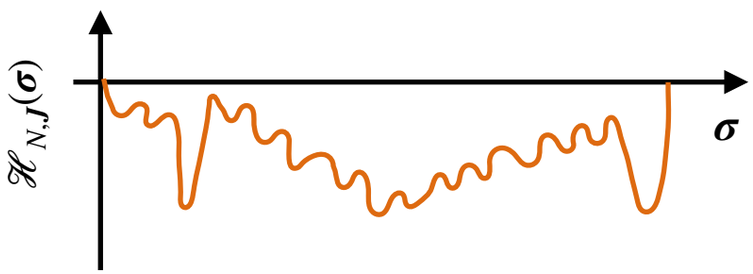




simple $K \sim \mathcal{O}(1)$

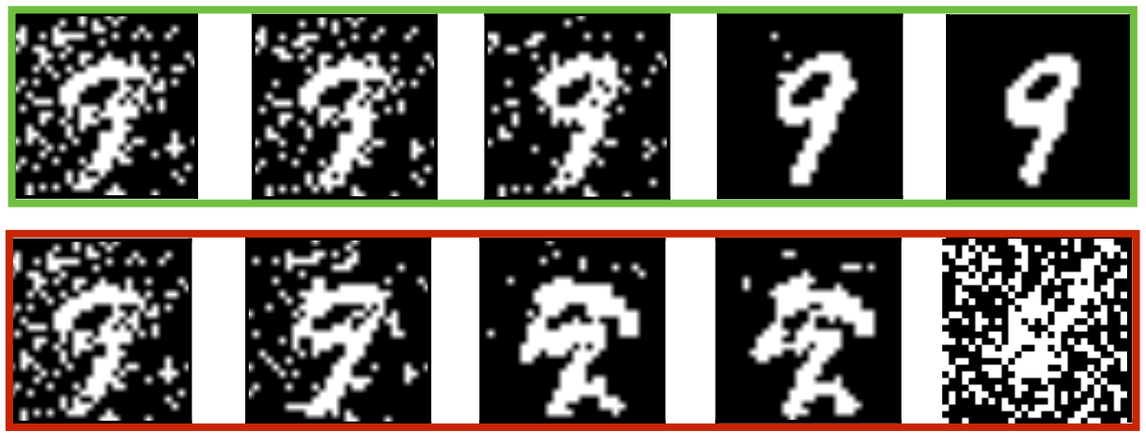


complex $K \sim \mathcal{O}(N)$



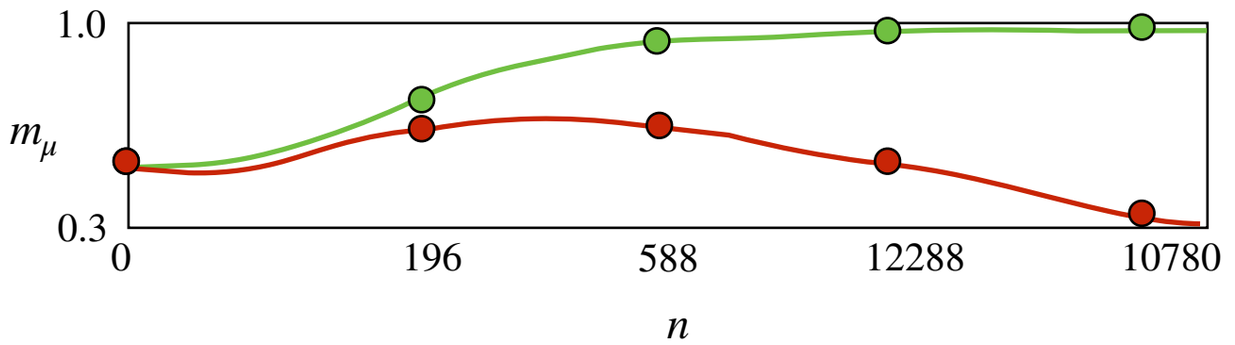
The model *works* (as a pattern reconstructor) if starting "near" a pattern, say $\boldsymbol{\sigma} \approx \boldsymbol{\xi}^\mu$, the dynamics eventually converges to that pattern. Here "near" means that the Hamming distance between the initial point and the pattern is smaller than ϵN, for some $\epsilon > 0$.

e.g., $\boldsymbol{\xi}^\mu =$ 



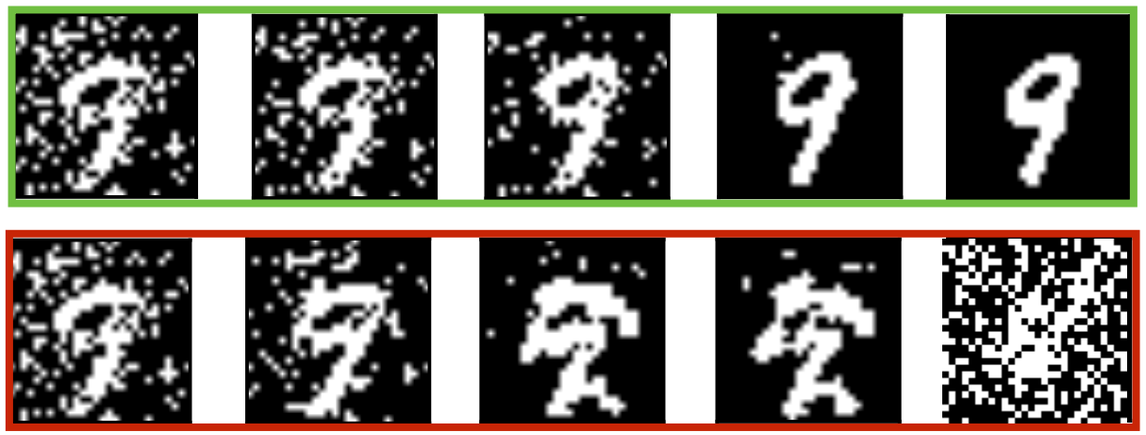
Mattis magnetization $m_\mu := \dfrac{1}{N} \sum_{i=1}^{N} \xi_i^\mu \sigma_i,$

assess retrieval of the μ-th pattern



The model *works* (as a pattern reconstructor) if starting "near" a pattern, say $\boldsymbol{\sigma} \approx \boldsymbol{\xi}^\mu$, the dynamics eventually converges to that pattern. Here "near" means that the Hamming distance between the initial point and the pattern is smaller than ϵN, for some $\epsilon > 0$.

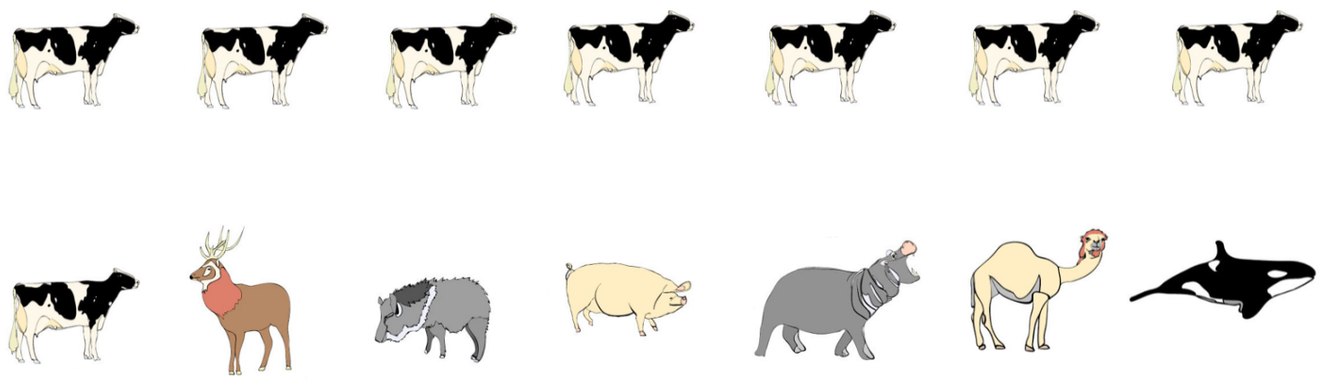
e.g., $\boldsymbol{\xi}^\mu =$ 



Mattis magnetization $m_\mu := \frac{1}{N} \sum_{i=1}^{N} \xi_i^\mu \sigma_i,$

assess retrieval of the μ-th pattern

Overlap $q_{ab} := \frac{1}{N} \sum_{i=1}^{N} \sigma_i^{(a)} \sigma_i^{(b)}$



intrinsic disorder and frustration yield glassy behaviours, captured by the correlation between configurations of two *replicas*

AGS theory

[Amit, Gutfreund, Sompolinsky - Phys. Rev. A '87]

$$\mathscr{H}_{N,\boldsymbol{\xi}}(\boldsymbol{\sigma}) = -\frac{1}{2N}\sum_{\substack{i,j=1\\ i\neq j}}^{N}\sum_{\mu=1}^{K}\xi_i^\mu\xi_j^\mu\sigma_i\sigma_j$$

Mattis magnetization $m_\mu := \frac{1}{N}\sum_{i=1}^{N}\xi_i^\mu\sigma_i,$

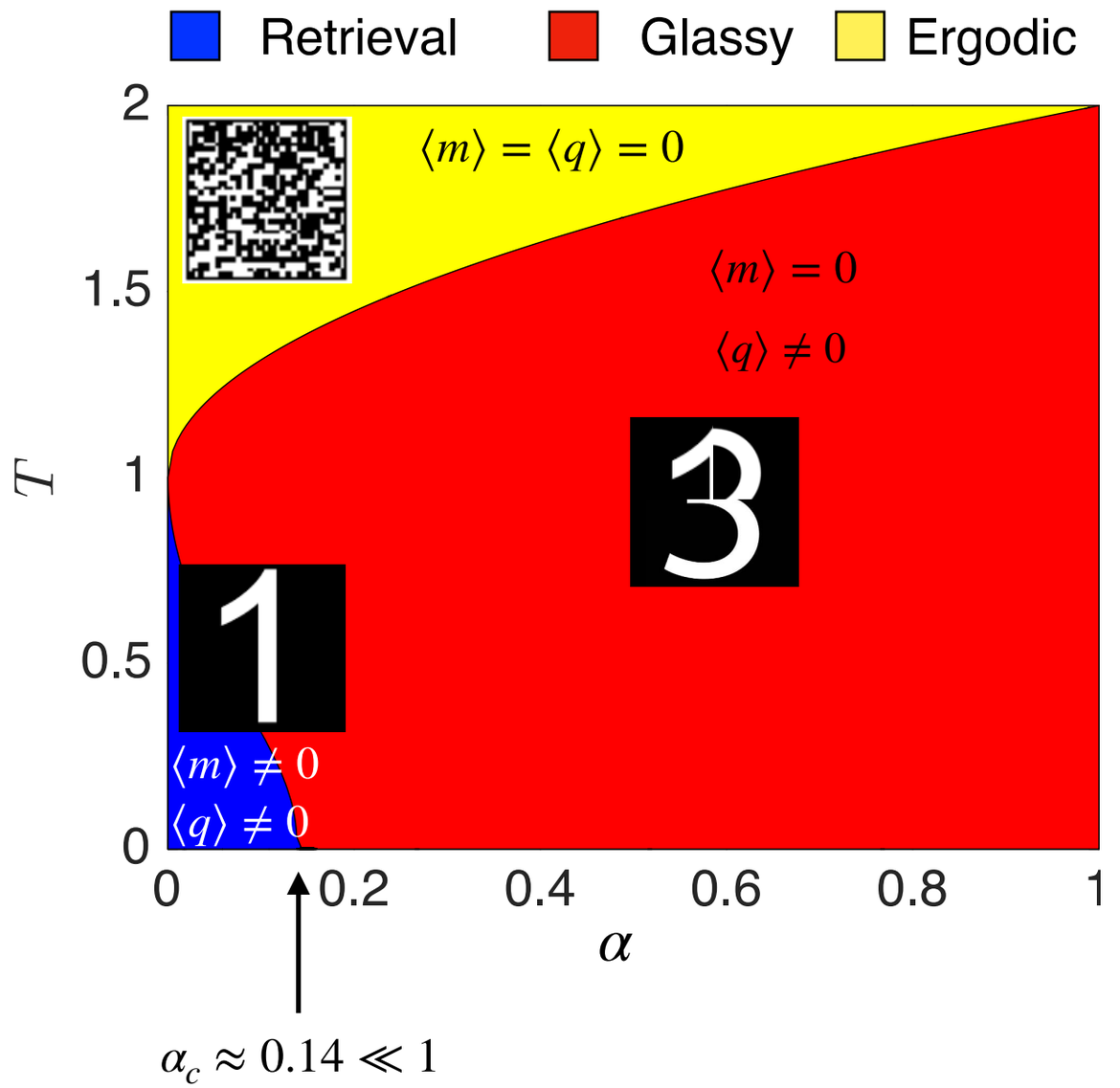
assess retrieval of the μ-th pattern

Overlap $q_{ab} := \frac{1}{N}\sum_{i=1}^{N}\sigma_i^{(a)}\sigma_i^{(b)}$

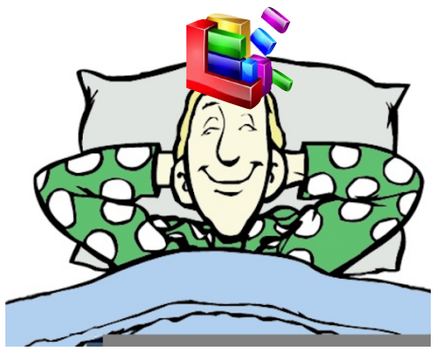
describes correlation between configurations of two *replicas*, namely two different systems endowed with the same couplings $\boldsymbol{J}$

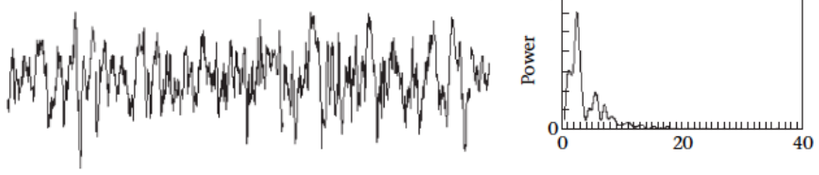
Load $\alpha := \lim_{N\to\infty}\frac{K}{N}$

Noise $T := \beta^{-1}$

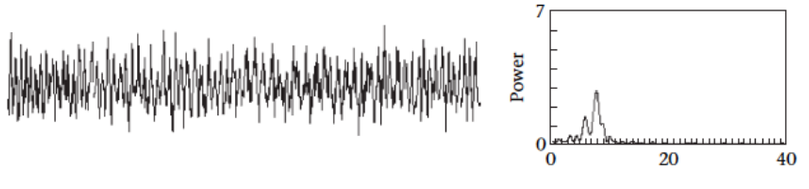


Revise Hebb's rule by iterative protocols





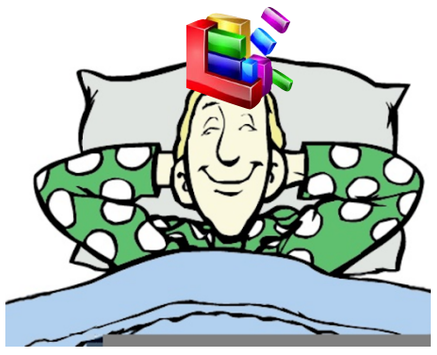
consolidation (SW)

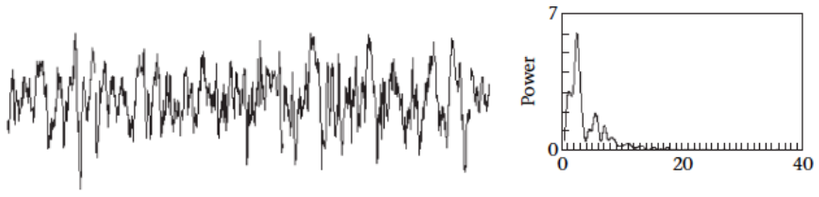


remotion (REM)

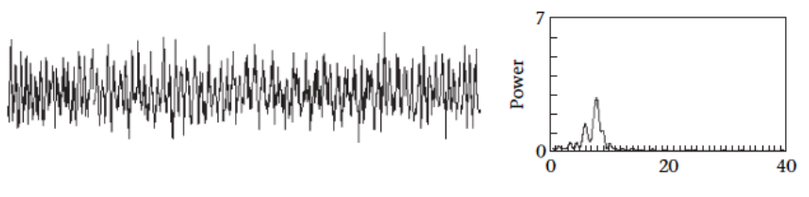
[Crick, Mitchinson - Nature '83; Stickgold - Nature '05; Diekelmann, Born - Nature Rev. Neurosc. '10]

Revise Hebb's rule by iterative protocols





consolidation (SW)



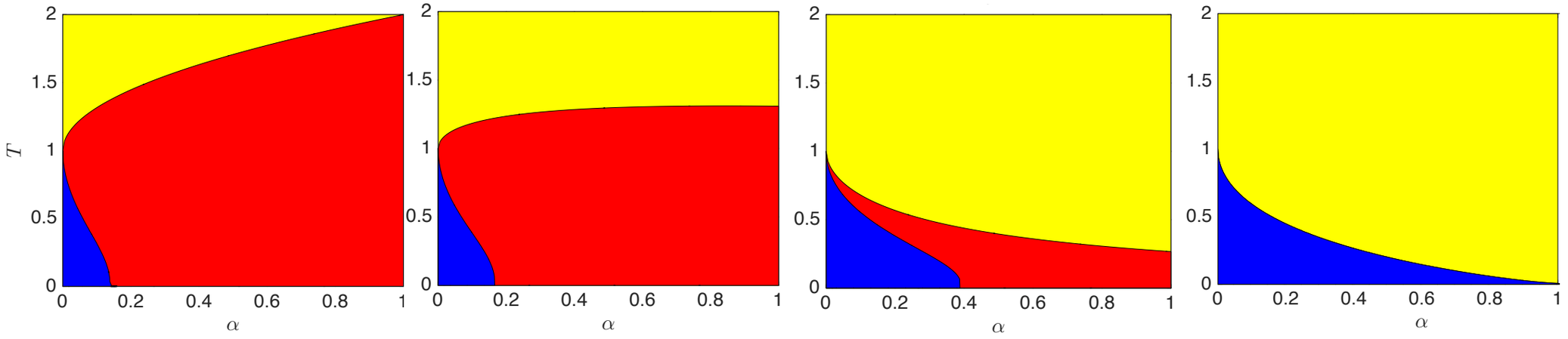
remotion (REM)

[Crick, Mitchinson - Nature '83; Stickgold - Nature '05; Diekelmann, Born - Nature Rev. Neurosc. '10]

Retrieval

Spin-glass

Ergodic



$t = 0$ $t = 0.1$ $t = 1$ $t = 1000$

Allocate more information with the same resources

$$J_{ij}^{(n)} \leftarrow J_{ij}^{(n-1)} + \frac{1}{1+n}[J^{(n-1)} - (J^{(n-1)})^2]$$

$$J_{ij}(t) = \frac{1}{N}\sum_{\mu,\nu} \xi_i^\mu \left(\frac{1+t}{1+t\boldsymbol{C}}\right)_{\mu\nu} \xi_j^\nu$$

$$\underbrace{(1+t)}_{\text{consolidation (SW)}} \cdot \underbrace{(\boldsymbol{I}+t\boldsymbol{C})^{-1}}_{\text{remotion (REM)}}$$

t "sleeping time"

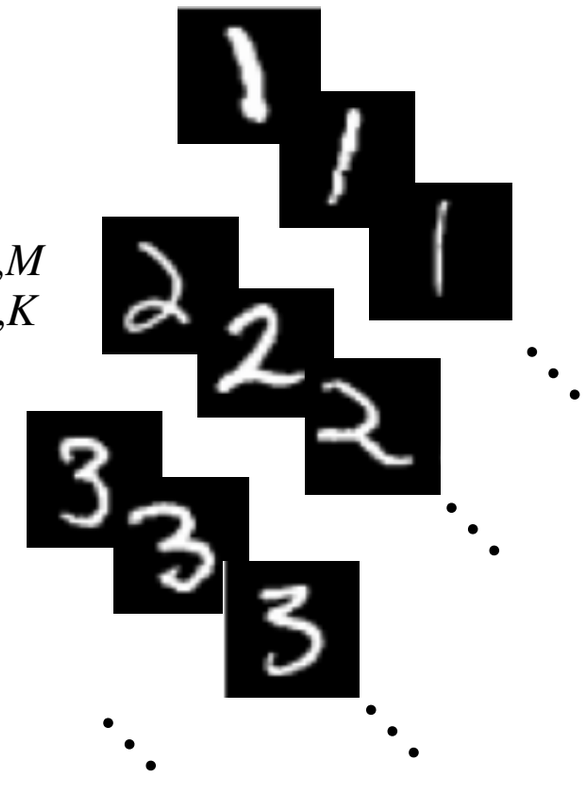
where $C_{\mu\nu} := \frac{1}{N}\sum_{i=1}^{N} \xi_i^\mu \xi_j^\mu$ pattern correlation matrix

Patterns' revision

Replace *archetypes* by a sample of *examples*

K "archetypes" $\{\boldsymbol{\xi}^{\mu}\}_{\mu=1,...,K}$

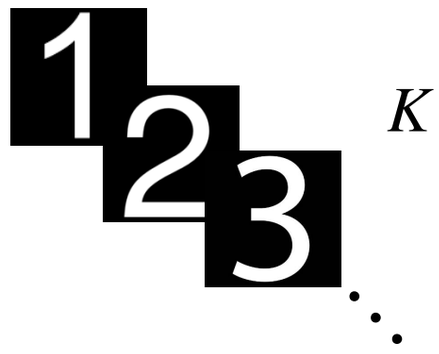
$K \times M$ "blurred" examples $\{\boldsymbol{\eta}^{\mu,a}\}_{\mu=1,...,K}^{a=1,...,M}$



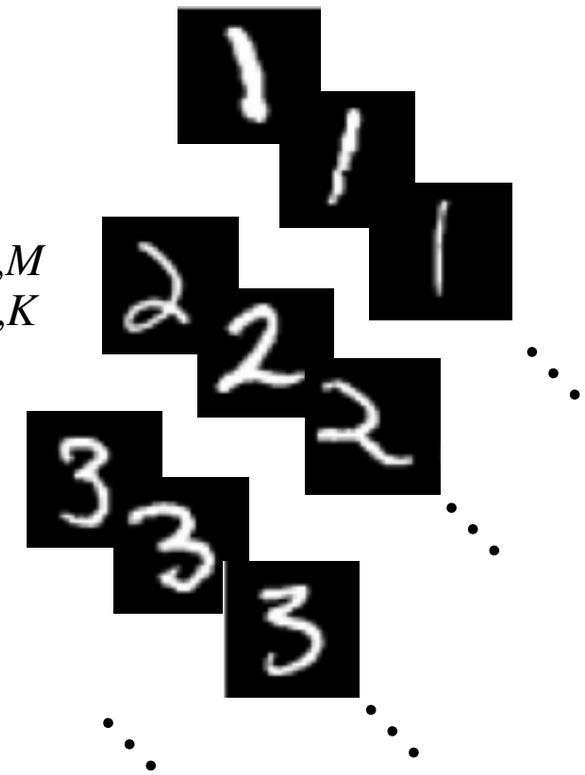
Patterns' revision

Replace *archetypes* by a sample of *examples*

K "archetypes" $\{\boldsymbol{\xi}^{\mu}\}_{\mu=1,...,K}$

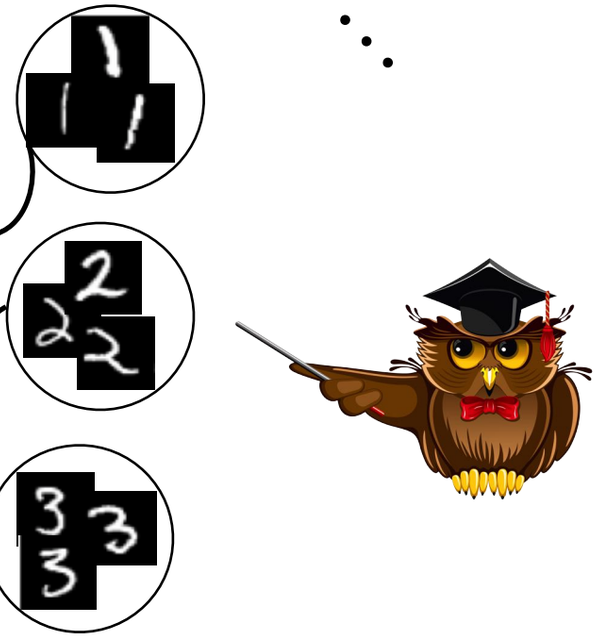


$K \times M$ "blurred" examples $\{\boldsymbol{\eta}^{\mu,a}\}_{\substack{a=1,...,M \\ \mu=1,...,K}}$

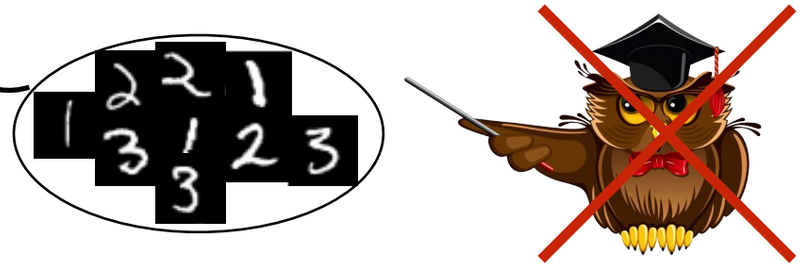


$$J_{ij}(\boldsymbol{\xi}) \propto \sum_{\mu=1}^{K} \xi_i^{\mu} \xi_i^{\mu}$$

$$J_{ij}^{(sup)} \propto \sum_{\mu=1}^{K} \left(\sum_{a=1}^{M} \eta_i^{\mu,a} \right) \left(\sum_{b=1}^{M} \eta_i^{\mu,b} \right)$$

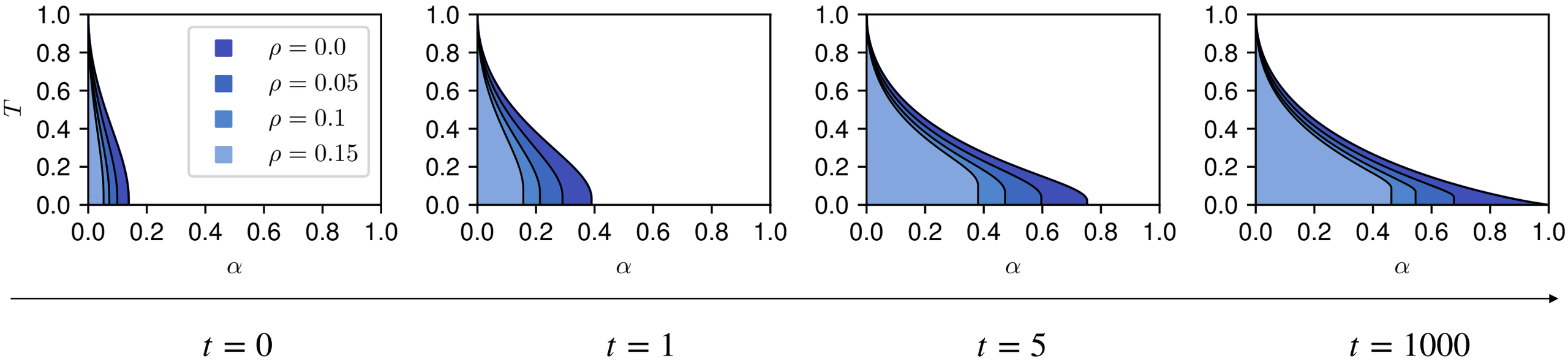


$$J_{ij}^{(unsup)} \propto \sum_{\mu=1}^{K} \sum_{a=1}^{M} \eta_i^{\mu,a} \eta_j^{\mu,a}$$



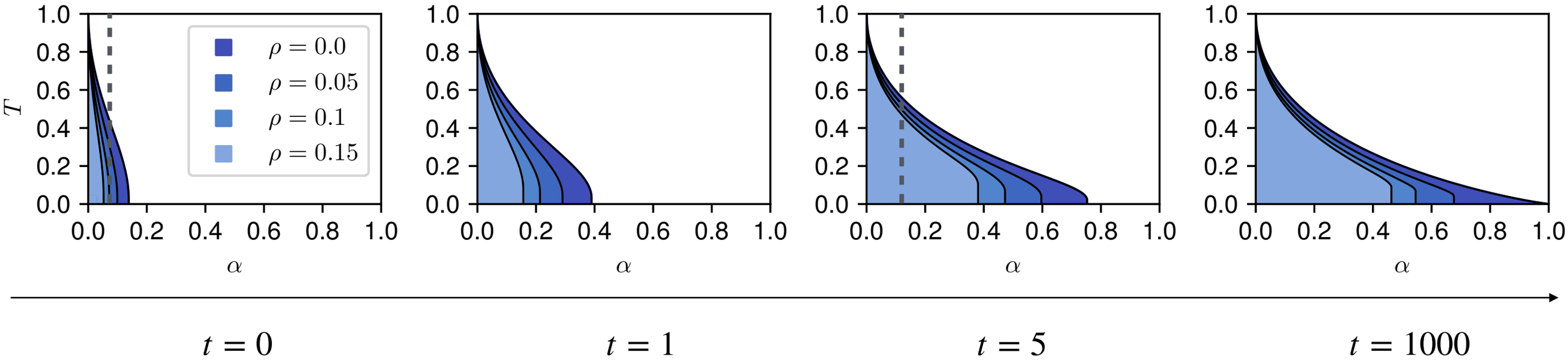
T, α, t

$$\rho \propto \frac{1}{\text{“quality”}} \quad \frac{1}{\text{“quantity”}}$$

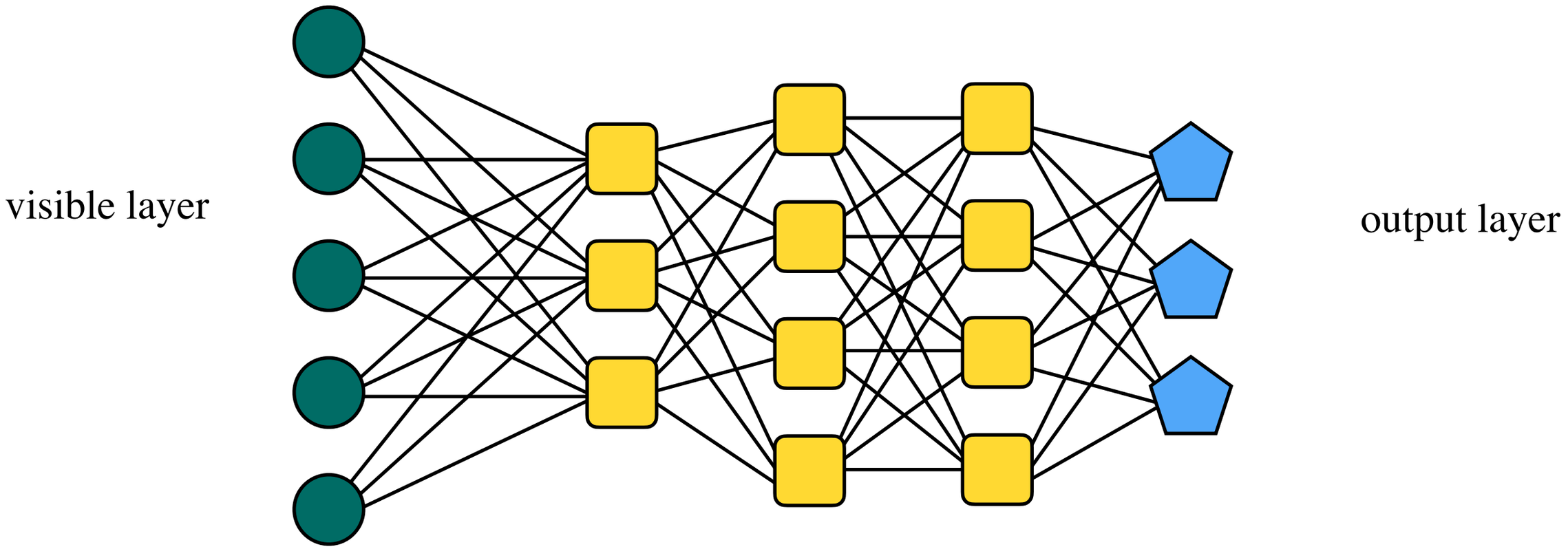


T, α, t

$$\rho \propto \frac{1}{\text{“quality”}} \quad \frac{1}{\text{“quantity”}}$$

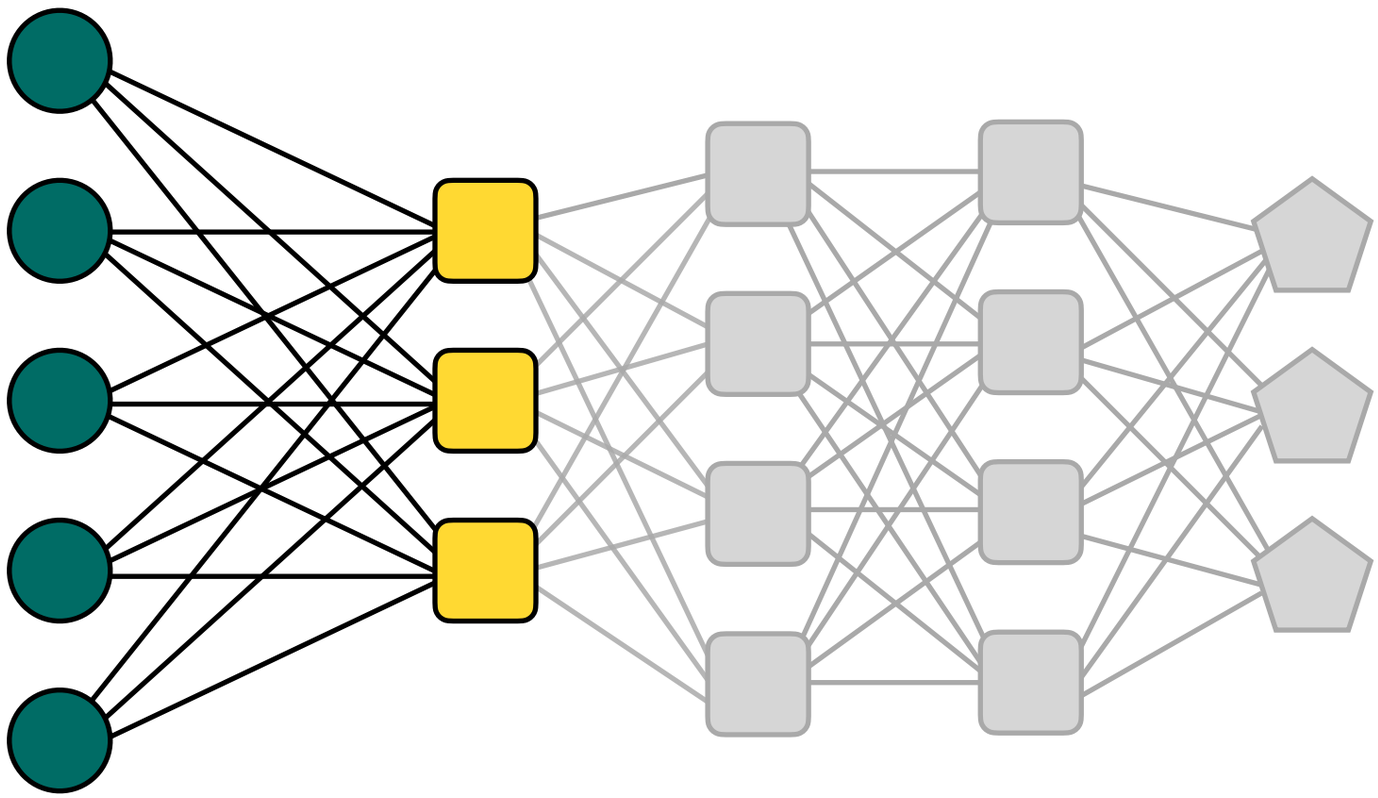


Restricted Boltzmann machines



Restricted Boltzmann machines

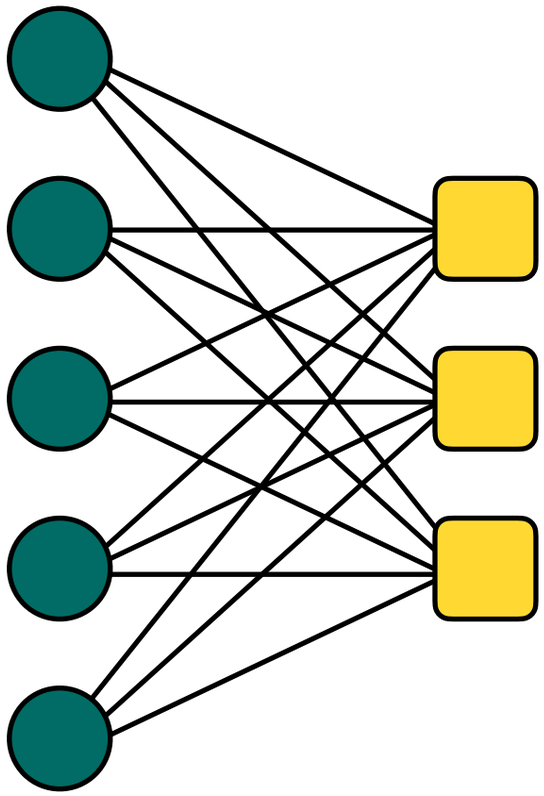
visible layer



Restricted Boltzmann machines

$\boldsymbol{\sigma} \in \{-1, +1\}^N$

visible layer



$\boldsymbol{z} \in \mathbb{R}^K$, Gaussian prior

hidden layer

Two layers composed of $N + K$ neurons

$\boldsymbol{W} \in \mathbb{R}^{N \times K}$ interaction matrix (symmetric, zero eye)

Boltzmann-Gibbs measure $\mathscr{P}_{\boldsymbol{W}}(\boldsymbol{\sigma}, \boldsymbol{z}) \propto e^{\beta \sum_{i,\mu} \sigma_i W_{i,\mu} z_\mu} \; e^{-\frac{\beta z_\mu^2}{2}}$

Stat-mech: bipartite spin-glass $\mathscr{H}^{(\mathrm{RBM})}_{N,K,\boldsymbol{W}}(\boldsymbol{\sigma}, \boldsymbol{z}) = -\sum_{i,\mu}^{N,K} W_{i,\mu} \sigma_i z_\mu$

Task: classification

$$\mathcal{S} = \{(\boldsymbol{x}^{(1)}, y^{(1)}), (\boldsymbol{x}^{(2)}, y^{(2)}), \ldots, (\boldsymbol{x}^{(M)}, y^{(M)})\}$$

$(\boldsymbol{x}^{(i)}, y^{(i)}) \sim_{iid} \mathcal{Q}(\boldsymbol{x}, y)$ unknown, target distr.

$\hat{y} \sim \mathcal{Q}(y \,|\, \boldsymbol{x})$

$(\boldsymbol{x}^{(1)},1)$

$(\boldsymbol{x}^{(2)},3)$

$(\boldsymbol{x}^{(3)},2)$

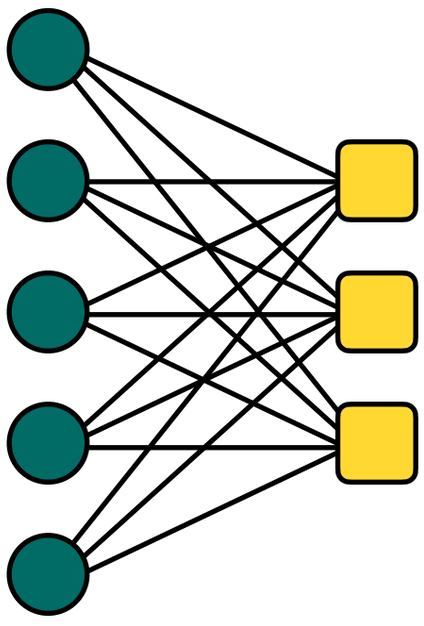
$(\boldsymbol{x}^{(4)},2)$

$(\boldsymbol{x}^{(5)},1)$

$\vdots$

$\boldsymbol{\sigma} \in \{-1, +1\}^N, \;\; N = \#\text{pixels}$

encodes for $\boldsymbol{x}$



$\boldsymbol{z} \in \mathbb{R}^K, \; K = \#\text{classes}$

encodes for y

Ideally: Look for a configuration $\hat{\boldsymbol{W}}$ of the matrix $\boldsymbol{W}$ s.t. $\mathcal{P}_{\hat{\boldsymbol{W}}}(\boldsymbol{\sigma}, \boldsymbol{z}) = \mathcal{Q}(\boldsymbol{\sigma}, \boldsymbol{z})$, namely s.t.

$$\hat{\boldsymbol{W}} = \min_{\boldsymbol{W} \in \mathbb{R}^{N \times K}} D_{\mathrm{KL}}(\mathcal{Q} \| \mathcal{P})$$

Task: classification

$\mathcal{S} = \{(\boldsymbol{x}^{(1)}, y^{(1)}), (\boldsymbol{x}^{(2)}, y^{(2)}), \ldots, (\boldsymbol{x}^{(M)}, y^{(M)})\}$

$(\boldsymbol{x}^{(i)}, y^{(i)}) \sim_{iid} \mathcal{Q}(\boldsymbol{x}, y)$ unknown, target distr.

$\hat{y} \sim \mathcal{Q}(y \,|\, \boldsymbol{x})$

$(\boldsymbol{x}^{(1)},1)$

$(\boldsymbol{x}^{(2)},3)$

$(\boldsymbol{x}^{(3)},2)$

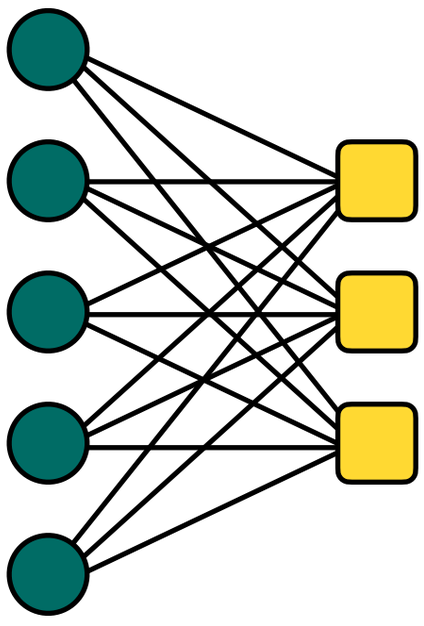
$(\boldsymbol{x}^{(4)},2)$

$(\boldsymbol{x}^{(5)},1)$

$\vdots$

$\boldsymbol{\sigma} \in \{-1, +1\}^N, \;\; N = \#\text{pixels}$

encodes for $\boldsymbol{x}$



$\boldsymbol{z} \in \mathbb{R}^K, \; K = \#\text{classes}$

encodes for y

Ideally: Look for a configuration $\hat{\boldsymbol{W}}$ of the matrix $\boldsymbol{W}$ s.t. $\mathcal{P}_{\hat{\boldsymbol{W}}}(\boldsymbol{\sigma}, \boldsymbol{z}) = \mathcal{Q}(\boldsymbol{\sigma}, \boldsymbol{z})$, namely s.t.

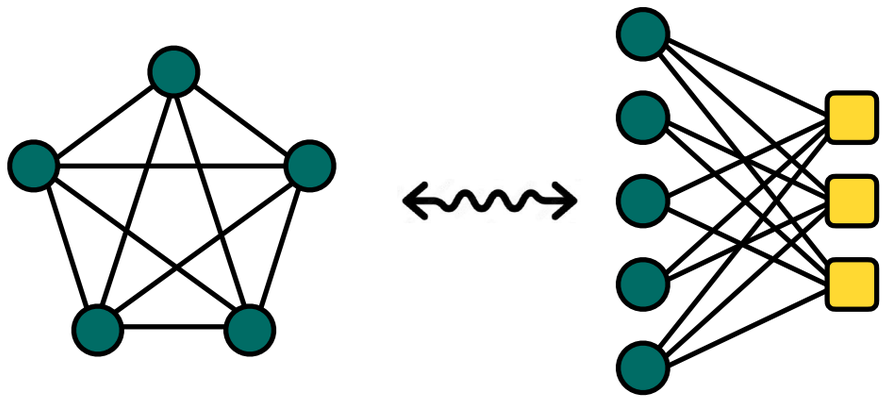
$$\hat{\boldsymbol{W}} = \min_{\boldsymbol{W} \in \mathbb{R}^{N \times K}} D_{\mathrm{KL}}(\mathcal{Q} \| \mathcal{P})$$

In practice: Look for a configuration $\hat{\boldsymbol{W}}$ of the matrix $\boldsymbol{W}$ s.t. $\mathcal{P}_{\hat{\boldsymbol{W}}}(\boldsymbol{\sigma}, \boldsymbol{z}) \approx \tilde{\mathcal{Q}}(\boldsymbol{\sigma}, \boldsymbol{z})$ by applying a gradient descent $W_{i,\mu}^{n+1} = W_{i,\mu}^{n} - \epsilon \dfrac{dD_{\mathrm{KL}}(\tilde{\mathcal{Q}} \| \mathcal{P})}{dW_{i,\mu}}$

Hopfield network / Restricted Boltzmann machines equivalence

$$\boldsymbol{\sigma} \in \{-1, +1\}^N$$

$$\boldsymbol{\xi} \in \{-1, +1\}^{N \times K}$$



$$\boldsymbol{\sigma} \in \{-1, +1\}^N$$

$$\boldsymbol{z} \sim \mathcal{N}(0, \boldsymbol{I}\beta^{-1})$$

$$\boldsymbol{W} = \boldsymbol{\xi}$$

Hubbard-Stratonovich transformation

$$\mathscr{Z}^{(\mathrm{HN})}_{\beta,N,K}(\boldsymbol{\xi}) = \sum_{\boldsymbol{\sigma}} e^{\frac{\beta}{2N}\sum_{i,j,\mu}\xi_i^\mu \xi_j^\mu \sigma_i \sigma_j} = \sum_{\boldsymbol{\sigma}} \int \prod_{\mu} dz_\mu e^{-\frac{\beta z_\mu^2}{2}} e^{\frac{\beta}{\sqrt{N}}\sum_{i,\mu}\sigma_i \xi_i^\mu z_\mu} = \mathscr{Z}^{(\mathrm{RBM})}_{\beta,N,K}(\boldsymbol{\xi})$$

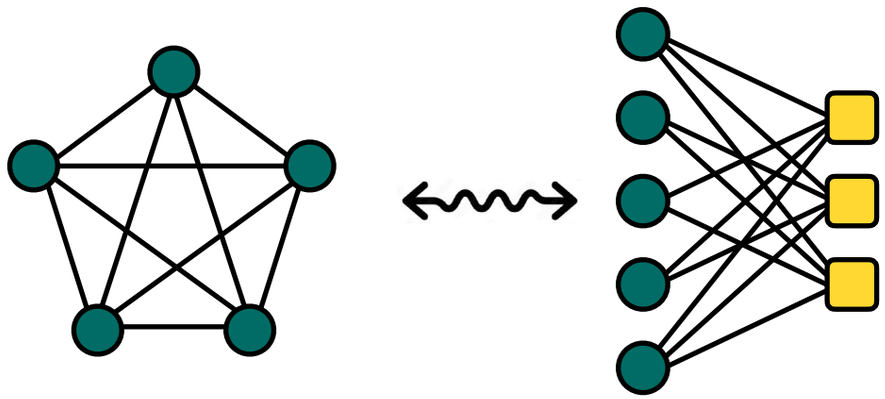
Gaussian integration

$$\mathscr{F} := -T \log \mathscr{Z} \ldots$$

Hopfield network / Restricted Boltzmann machines equivalence

$$\boldsymbol{\sigma} \in \{-1, +1\}^N$$

$$\boldsymbol{\xi} \in \{-1, +1\}^{N \times K}$$



$$\boldsymbol{\sigma} \in \{-1, +1\}^N$$

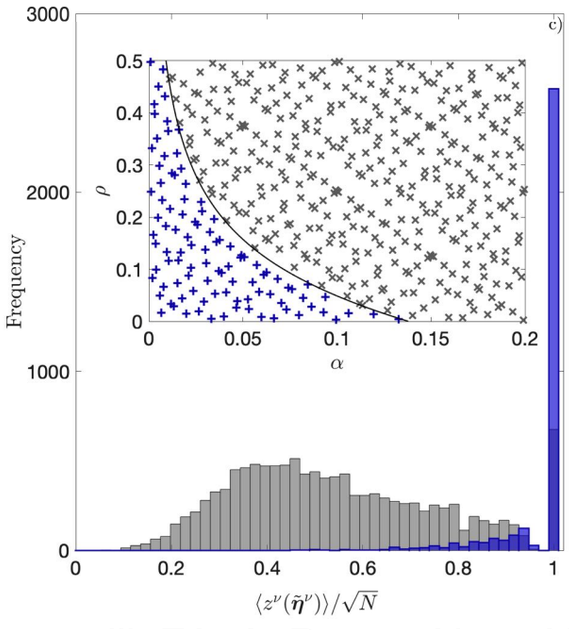
$$\boldsymbol{z} \sim \mathcal{N}(0, \boldsymbol{I}\beta^{-1})$$

$$\boldsymbol{W} = \boldsymbol{\xi}$$

Hubbard-Stratonovich transformation

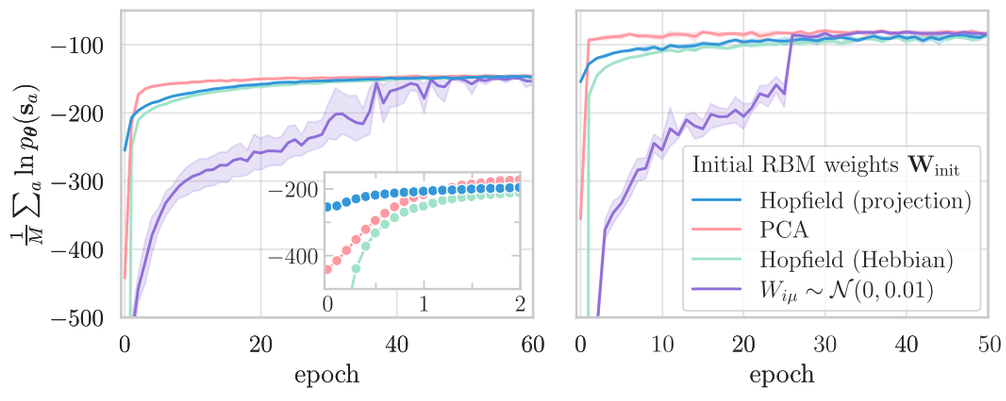
$$\mathcal{Z}^{(\mathrm{HN})}_{\beta,N,K}(\boldsymbol{\xi}) = \sum_{\boldsymbol{\sigma}} e^{\frac{\beta}{2N}\sum_{i,j,\mu}\xi_i^\mu \xi_j^\mu \sigma_i \sigma_j} = \sum_{\boldsymbol{\sigma}} \int \prod_{\mu} dz_\mu e^{-\frac{\beta z_\mu^2}{2}} e^{\frac{\beta}{\sqrt{N}}\sum_{i,\mu}\sigma_i \xi_i^\mu z_\mu} = \mathcal{Z}^{(\mathrm{RBM})}_{\beta,N,K}(\boldsymbol{\xi})$$

Gaussian integration

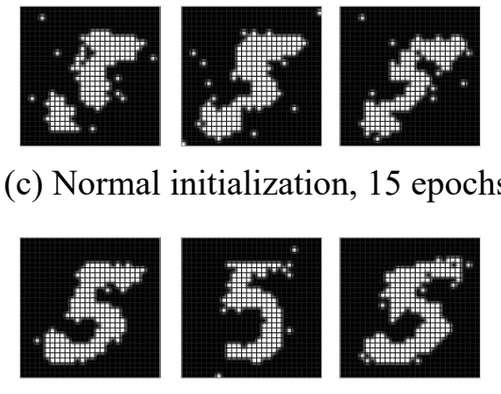


[F.E. Leonelli, EA, A. Barra - Neur. Net. ’21]

[EA, F. Alemanno, A. Barra, G. De Marzo - Neur. Net. ’22]



(a) 10 hidden units (b) 50 hidden units



(c) Normal initialization, 15 epochs

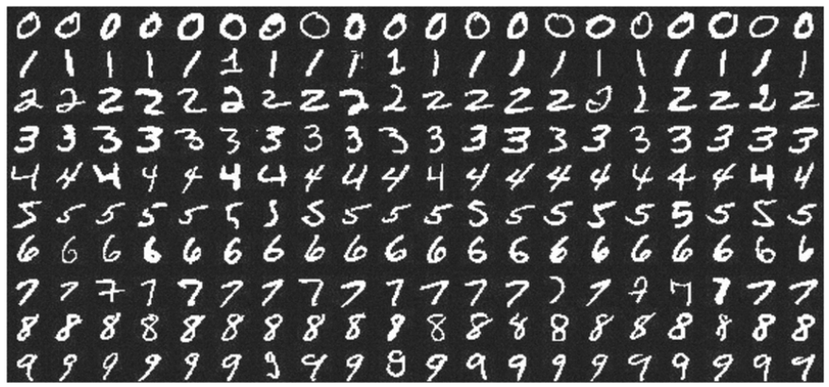
(d) Hopfield initialization, 15 epochs

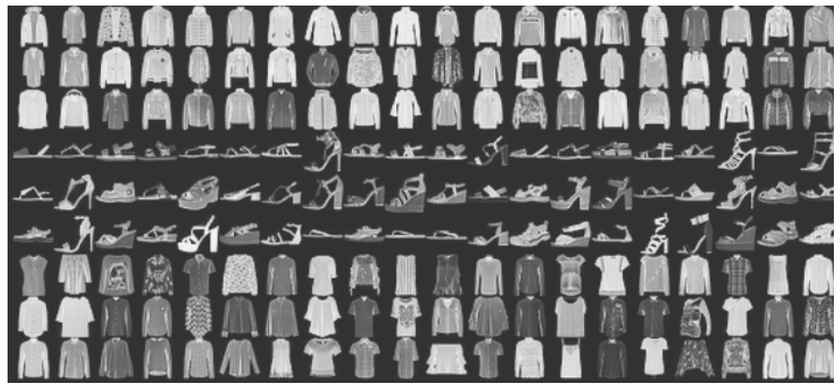
[Smart, Zilman - ICLR ’21]

Dataset and architecture complexities interplay

MNIST and Fashion-MNIST Datasets

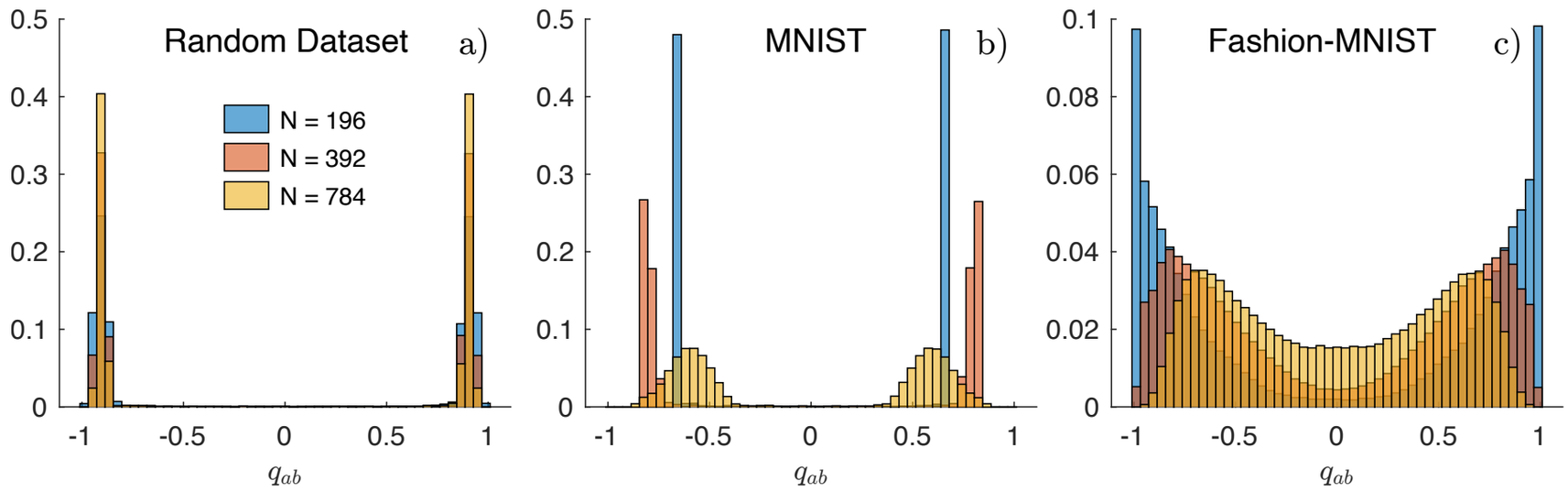
$K = 10$ classes
$M = 6000$ examples each





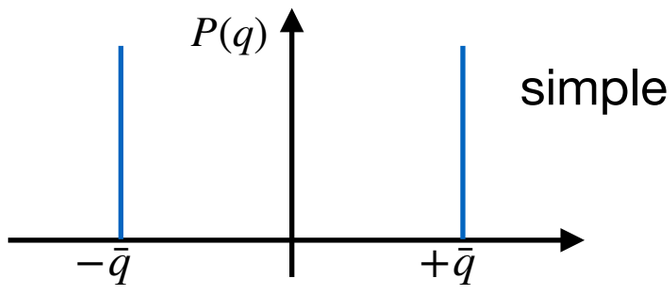
example overlaps

$$\tilde{q}_{ab} = \frac{1}{N}\sum_{i=1}^{N} \eta_i^{\mu,a}\eta_i^{\mu,b}$$

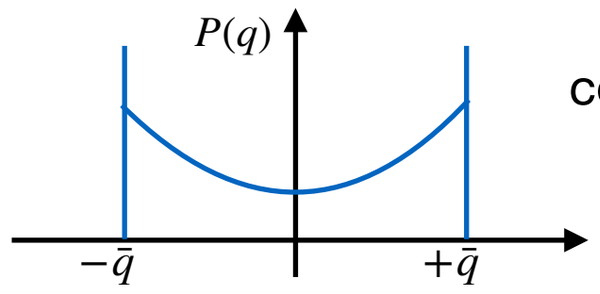


replica overlaps

$$q_{ab} = \frac{1}{N}\sum_{i=1}^{N} \sigma_i^{(a)}\sigma_i^{(b)}$$



simple



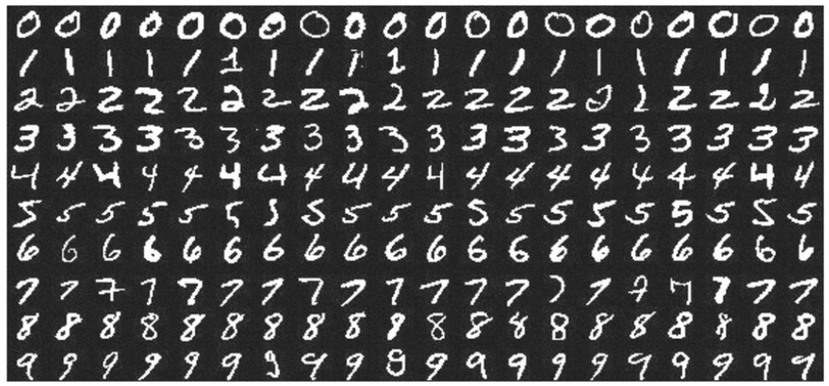
complex

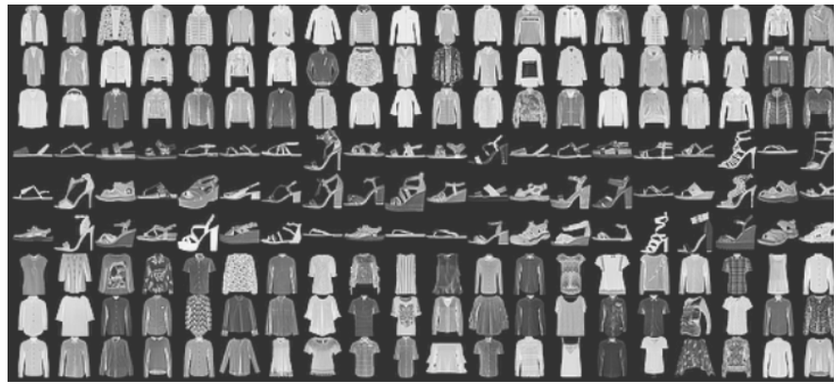
[F. Alemanno, M. Aquaro, I. Kanter, A. Barra, EA - Europhys. Lett. ’23]

Dataset and architecture complexities interplay

MNIST and Fashion-MNIST Datasets

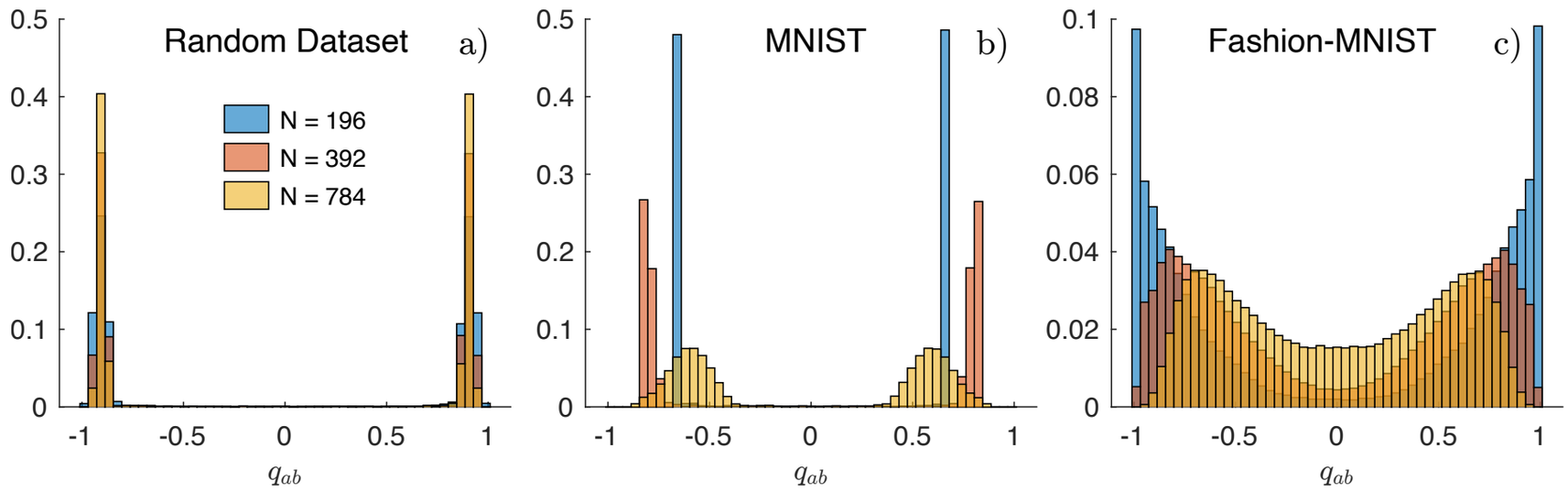
$K = 10$ classes
$M = 6000$ examples each





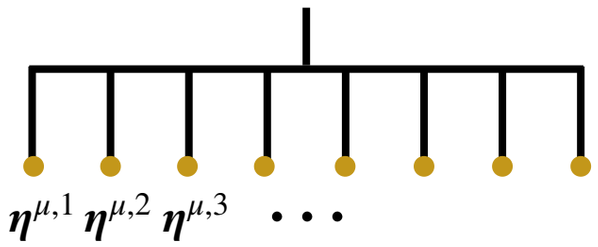
example overlaps

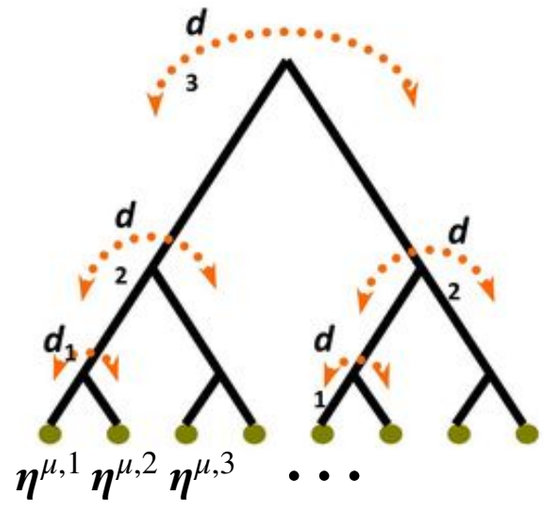
$$\tilde{q}_{ab} = \frac{1}{N}\sum_{i=1}^{N} \eta_i^{\mu,a}\eta_i^{\mu,b}$$



replica overlaps

$$q_{ab} = \frac{1}{N}\sum_{i=1}^{N} \sigma_i^{(a)}\sigma_i^{(b)}$$

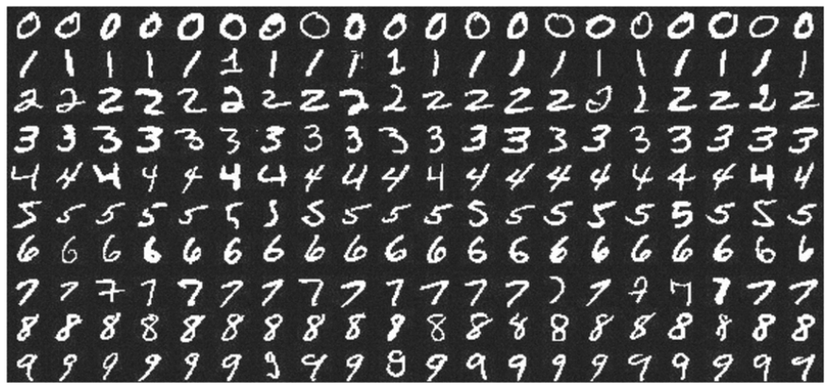


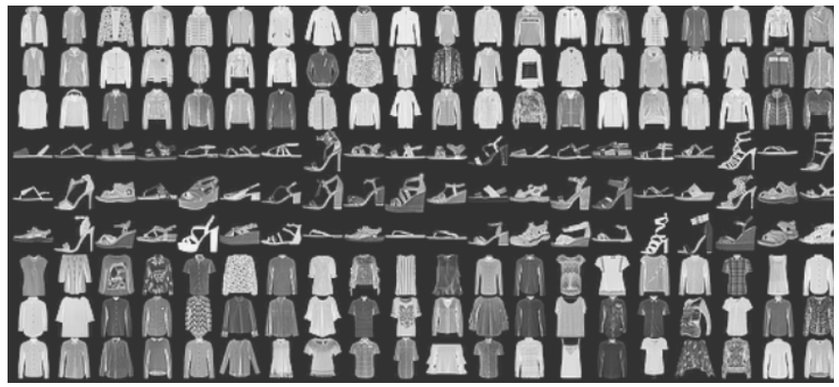


Dataset and architecture complexities interplay

MNIST and Fashion-MNIST Datasets

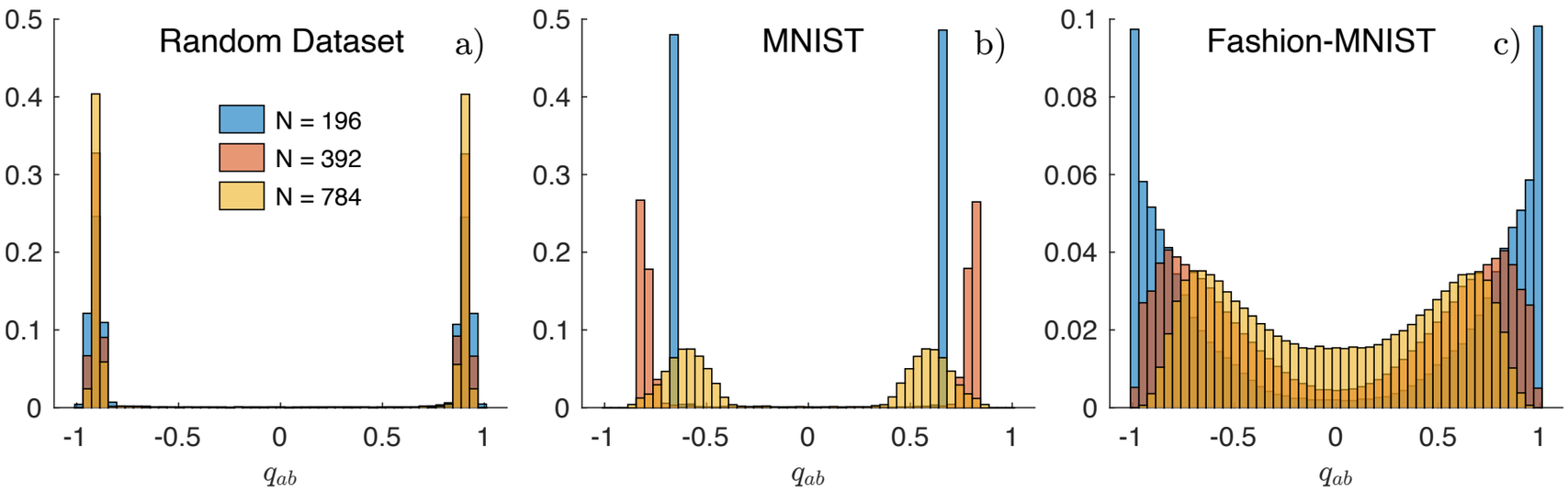
$K = 10$ classes
$M = 6000$ examples each





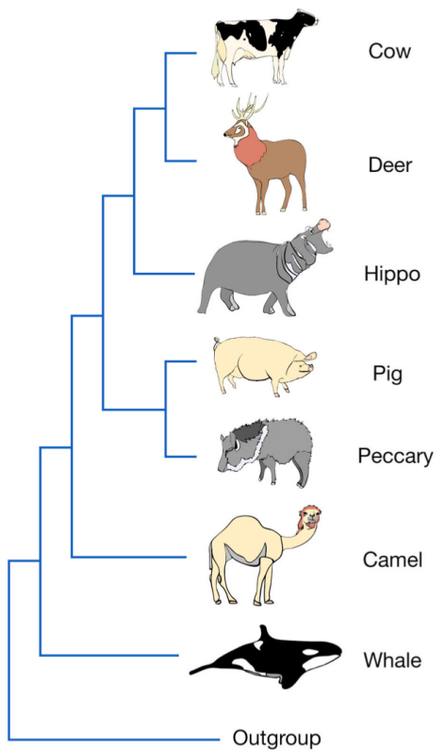
example overlaps

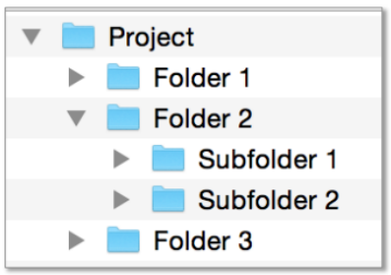
$$\tilde{q}_{ab} = \frac{1}{N}\sum_{i=1}^{N} \eta_i^{\mu,a}\eta_i^{\mu,b}$$

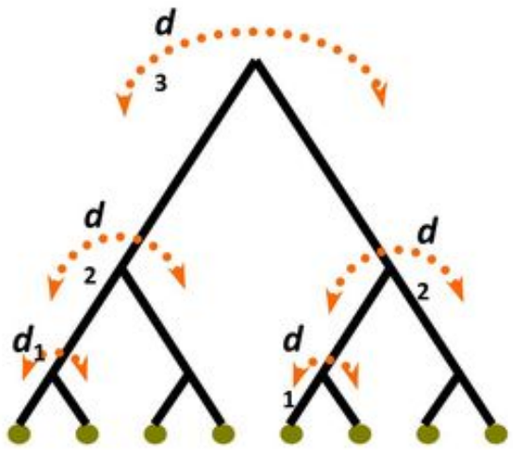


replica overlaps

$$q_{ab} = \frac{1}{N}\sum_{i=1}^{N} \sigma_i^{(a)}\sigma_i^{(b)}$$



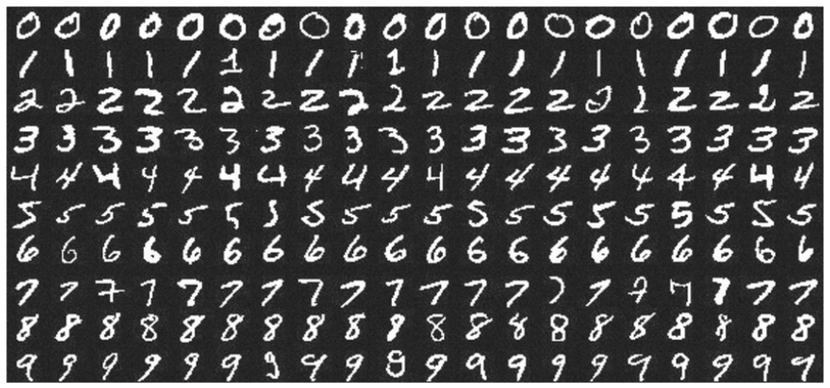


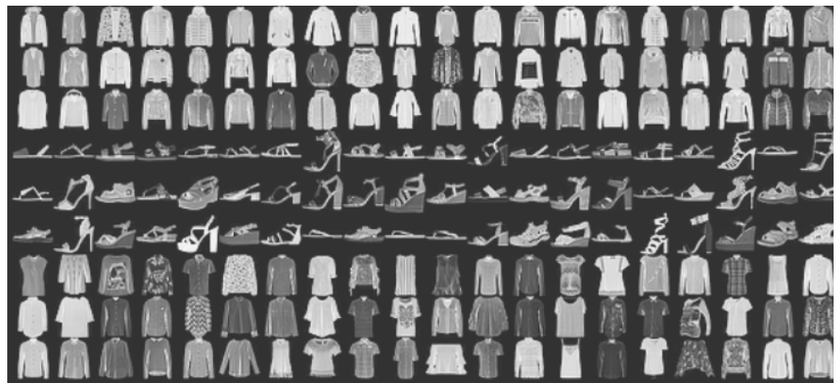


MNIST and Fashion-MNIST Datasets

$K = 10$ classes

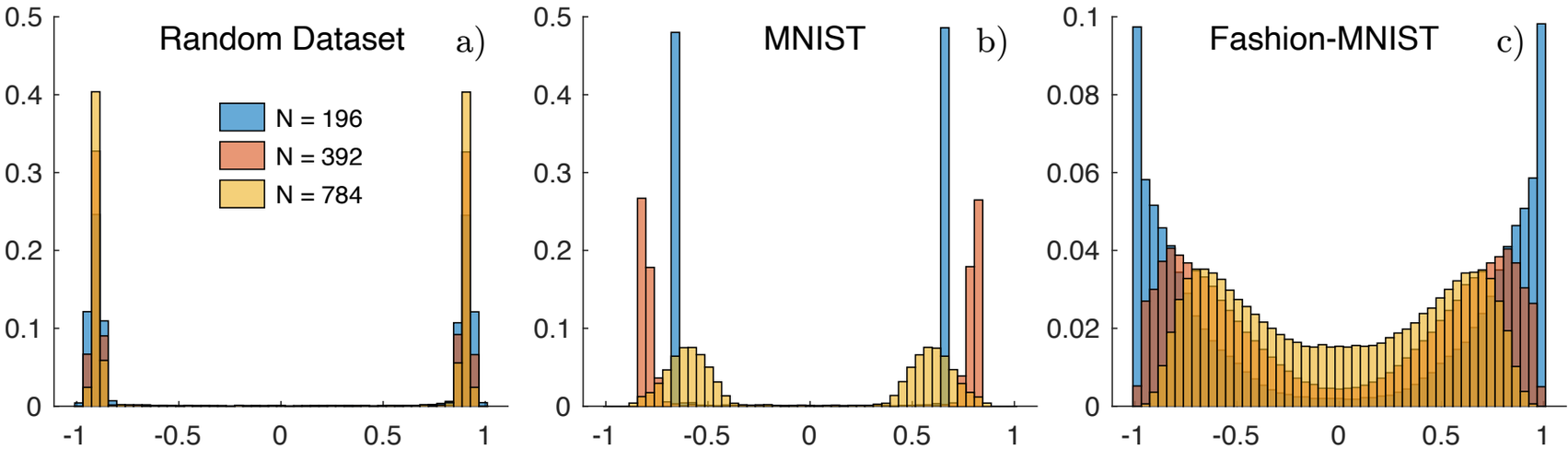
$M = 6000$ examples each

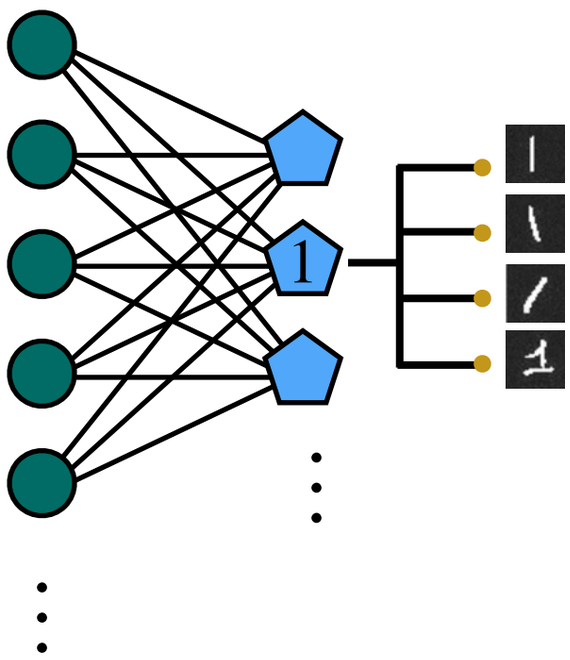




example overlaps

$$\tilde{q}_{ab} = \sum_{i=1}^{N} \xi_i^{(a)} \xi_i^{(b)}$$

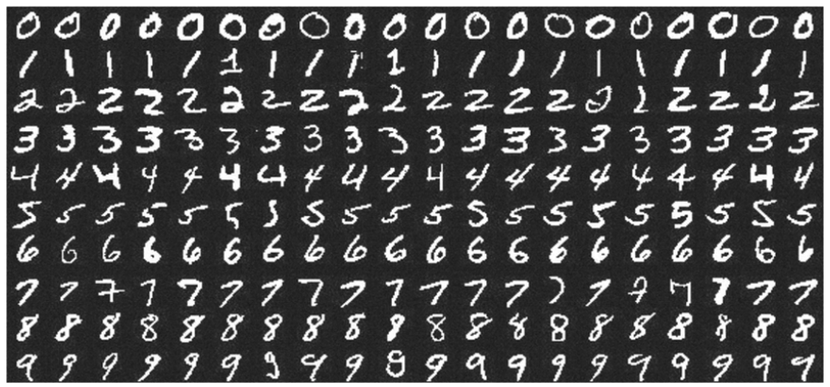


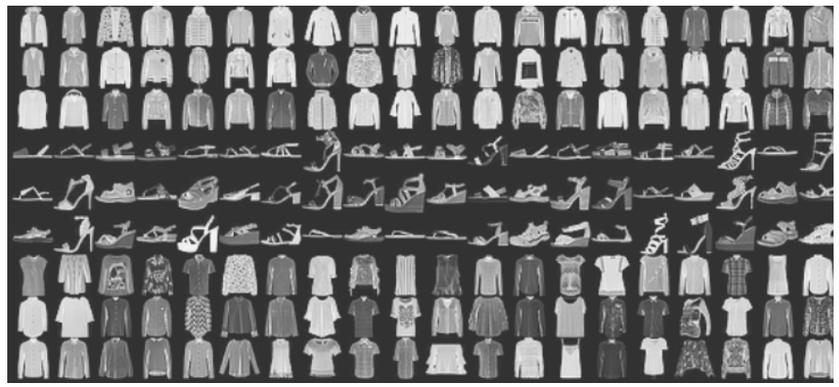


MNIST and Fashion-MNIST Datasets

$K = 10$ classes

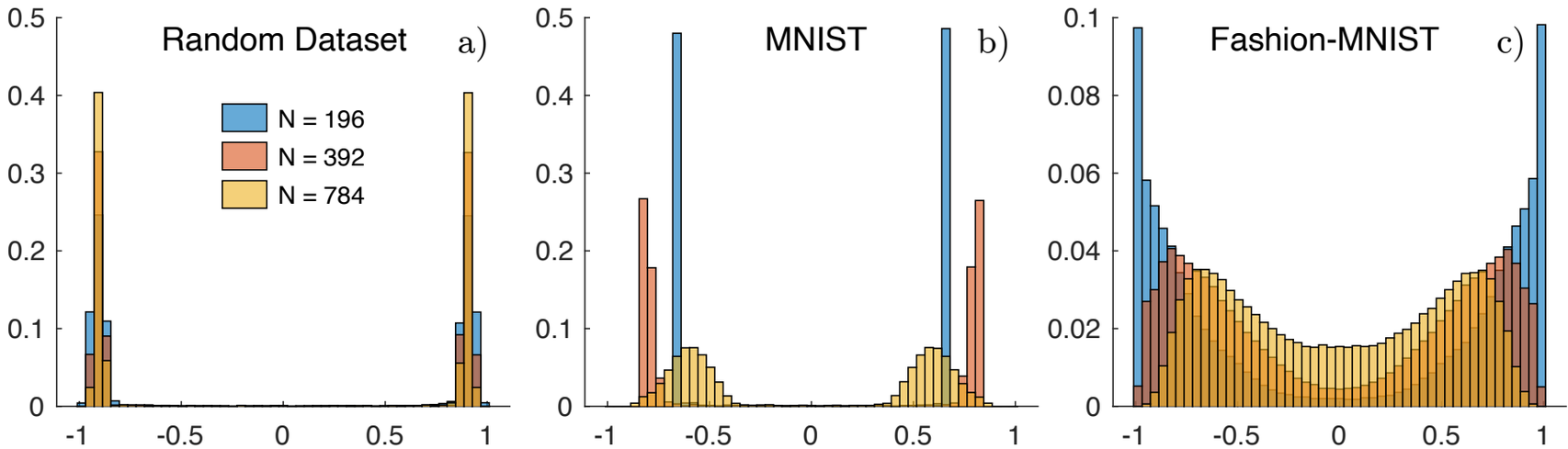
$M = 6000$ examples each

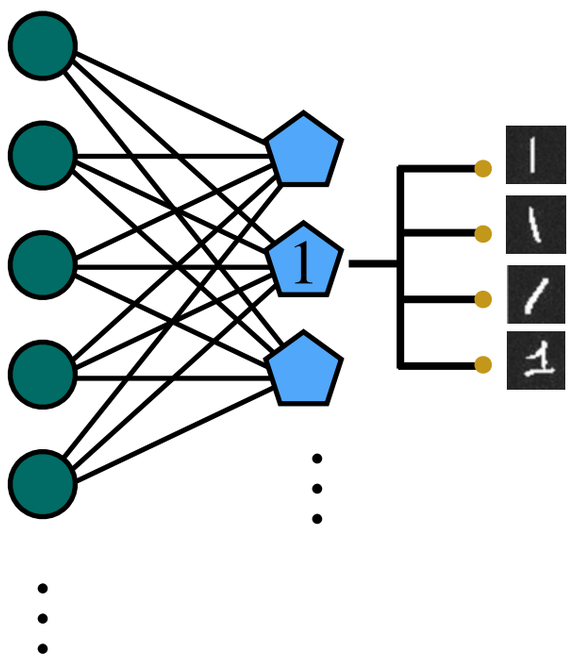


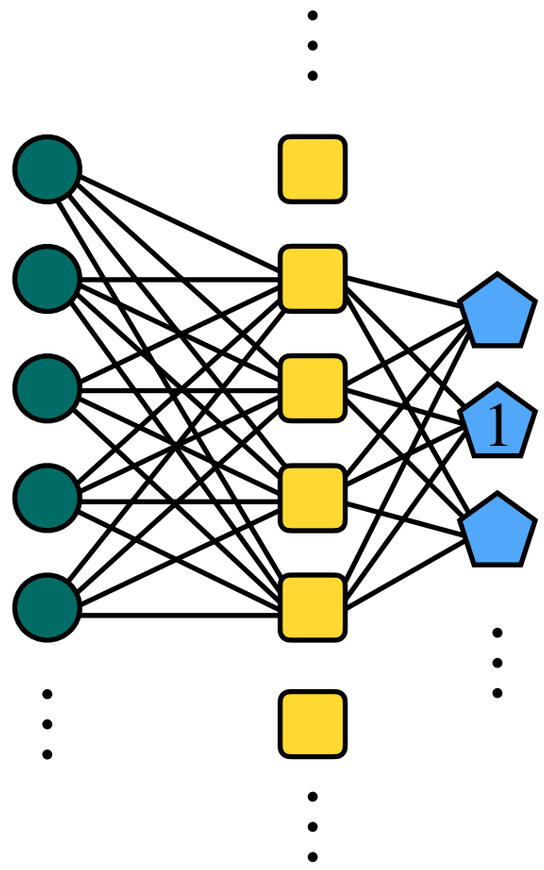


example overlaps

$$\tilde{q}_{ab} = \sum_{i=1}^{N} \xi_i^{(a)} \xi_i^{(b)}$$



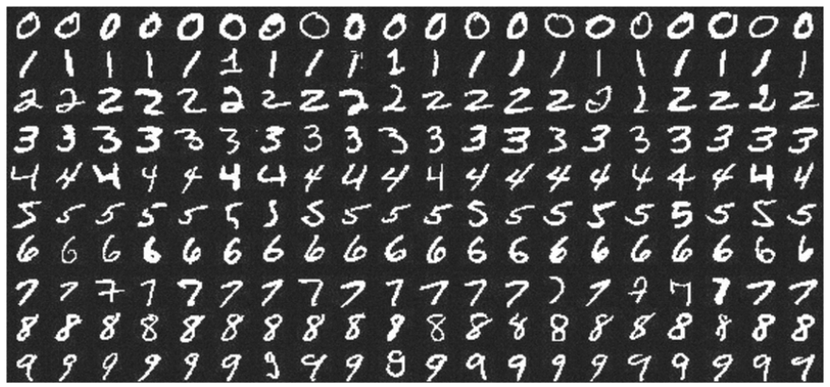


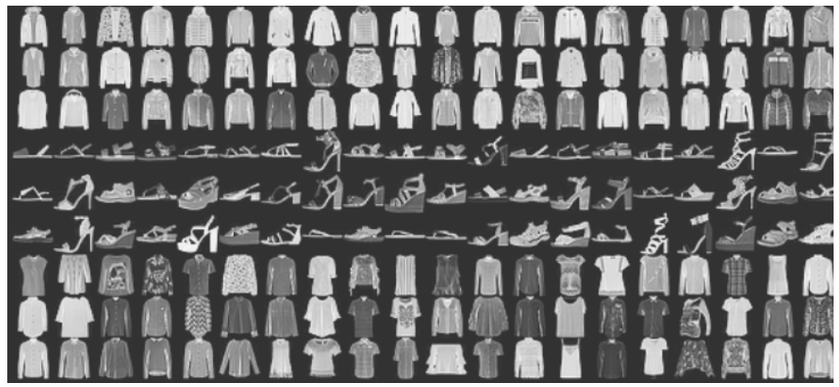


MNIST and Fashion-MNIST Datasets

$K = 10$ classes

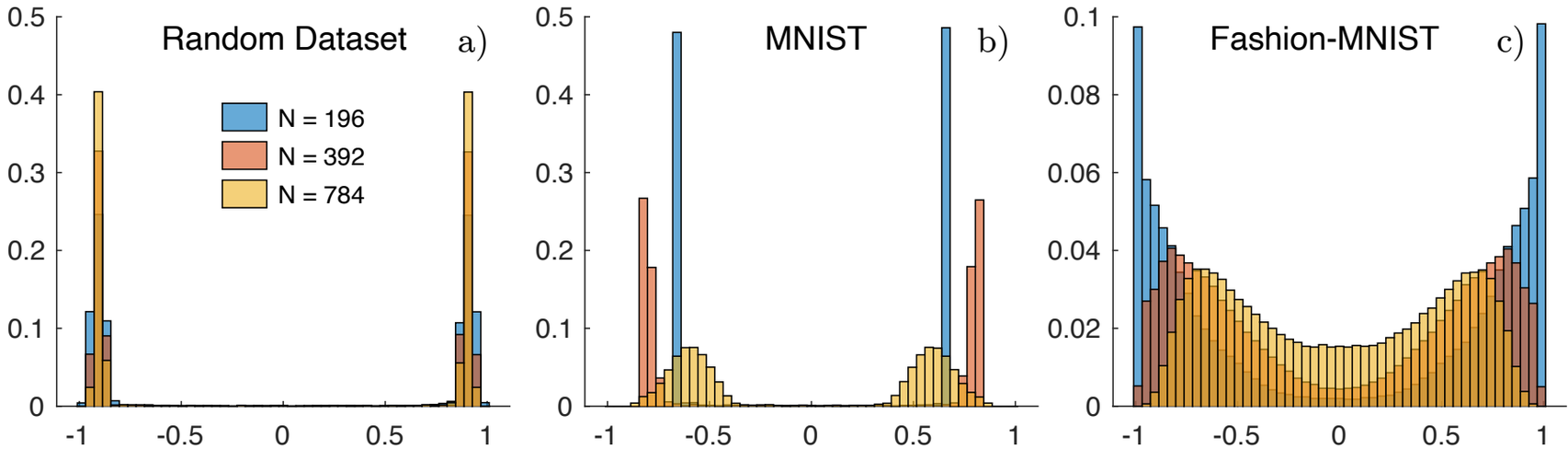
$M = 6000$ examples each

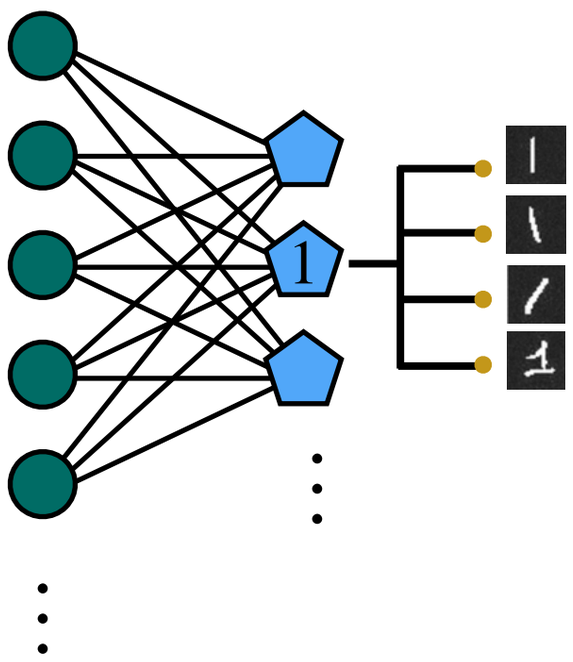


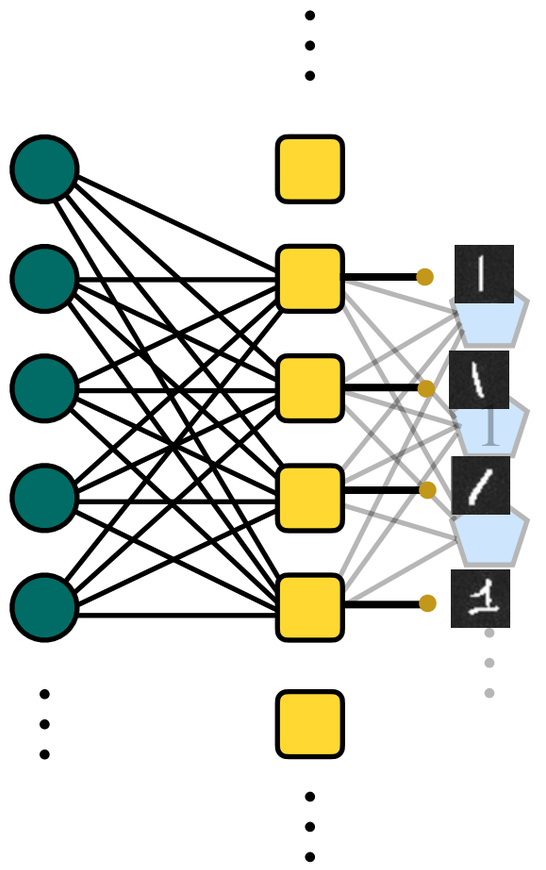


example overlaps

$$\tilde{q}_{ab} = \sum_{i=1}^{N} \xi_i^{(a)} \xi_i^{(b)}$$



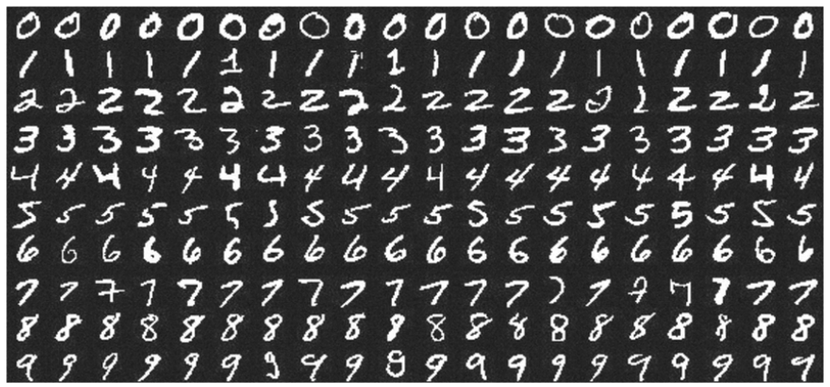


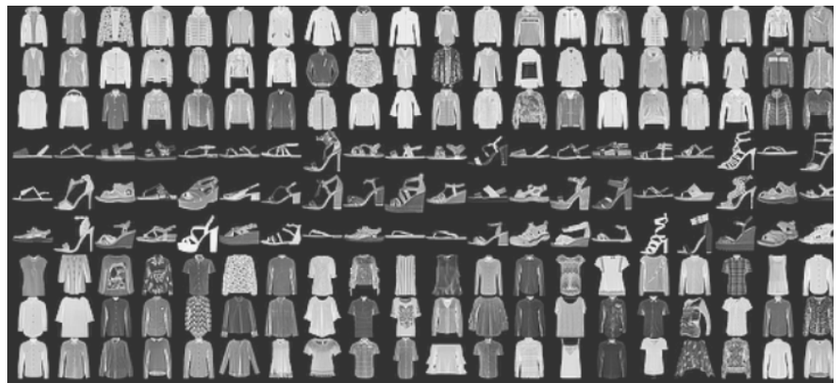


MNIST and Fashion-MNIST Datasets

$K = 10$ classes

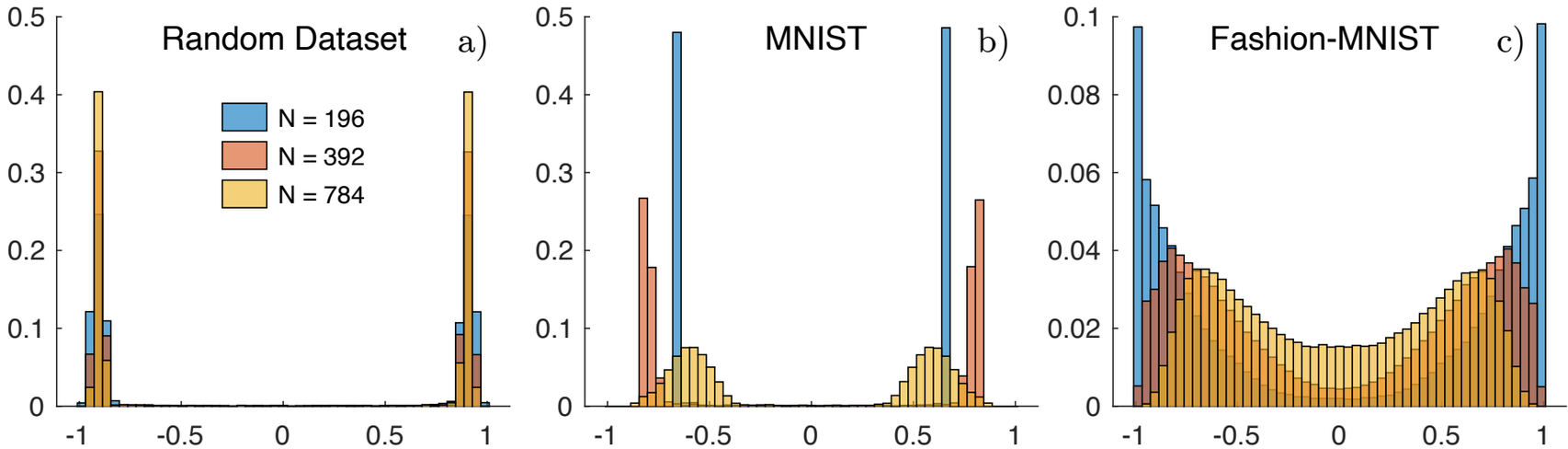
$M = 6000$ examples each

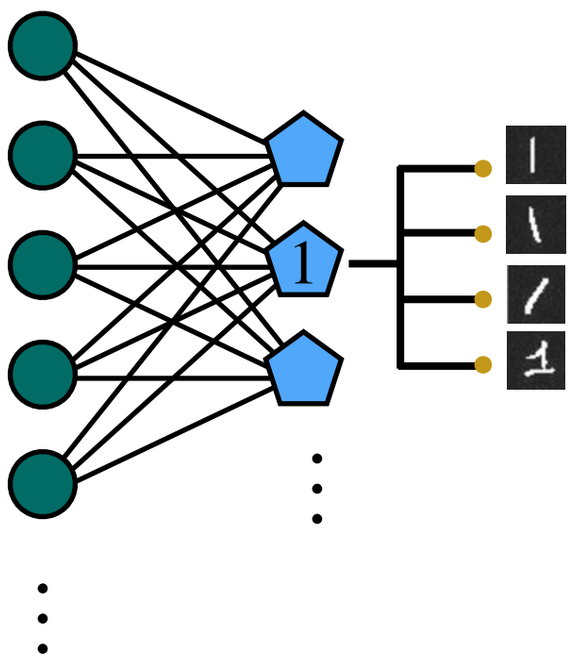


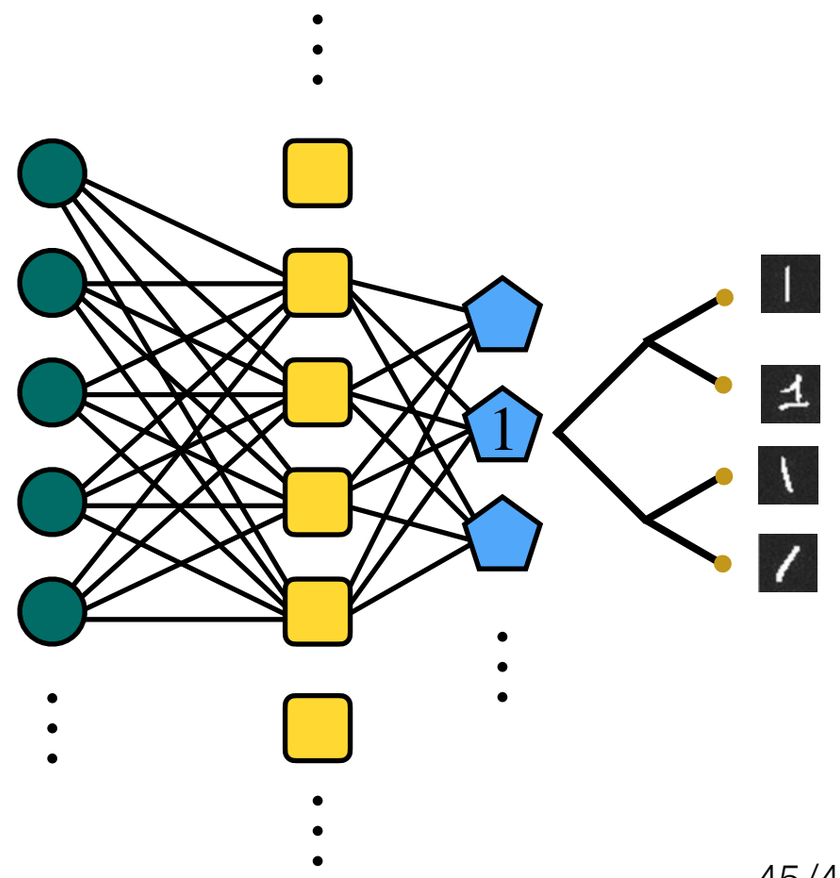


example overlaps

$$\tilde{q}_{ab} = \sum_{i=1}^{N} \xi_i^{(a)} \xi_i^{(b)}$$



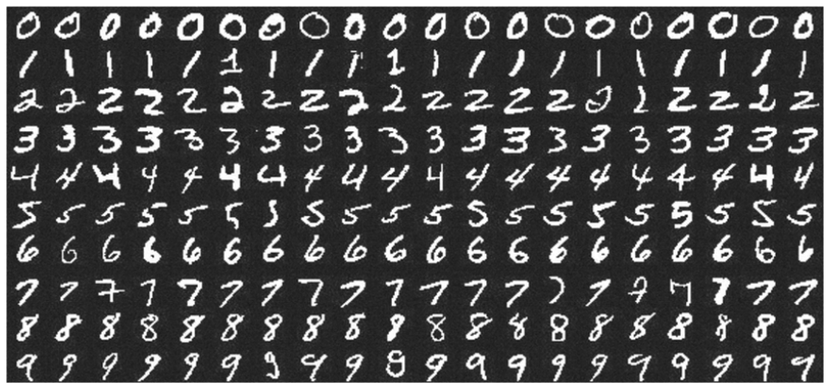


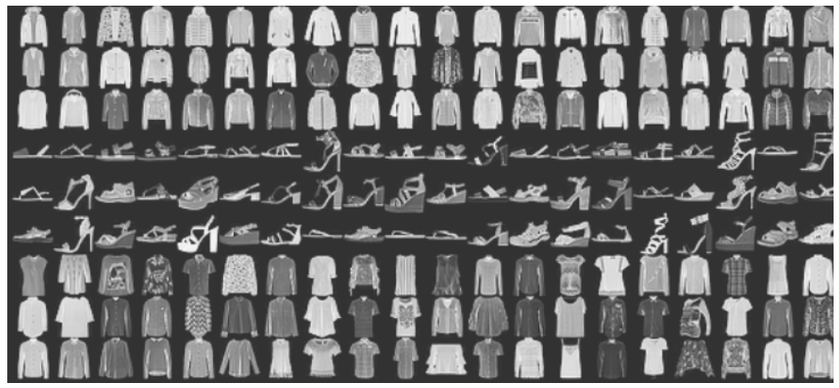


MNIST and Fashion-MNIST Datasets

$K = 10$ classes

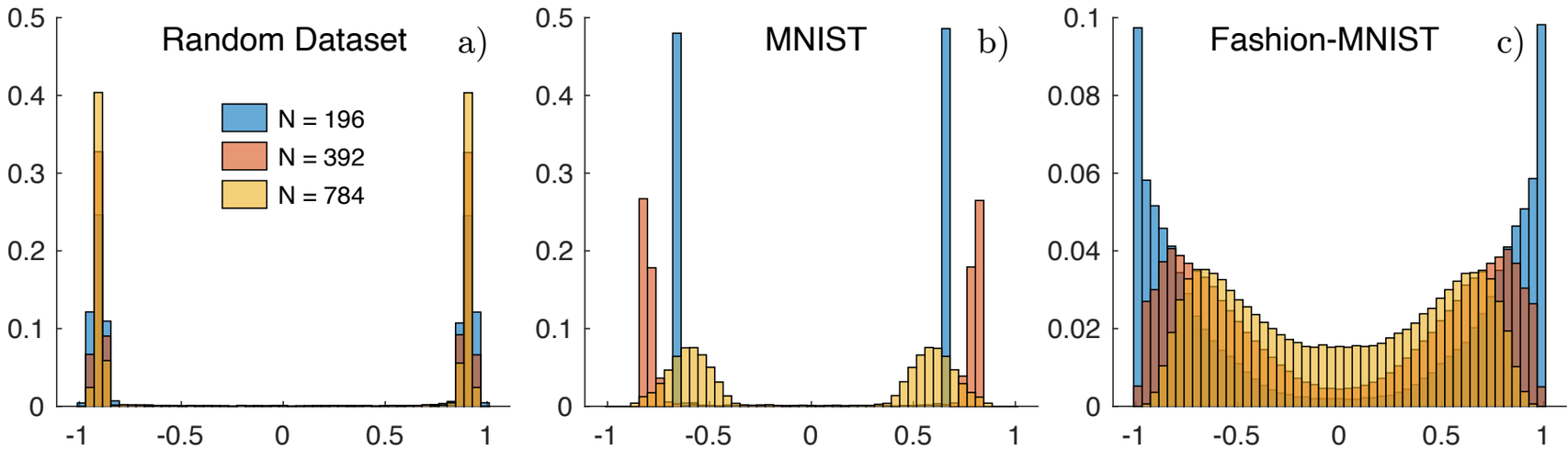
$M = 6000$ examples each

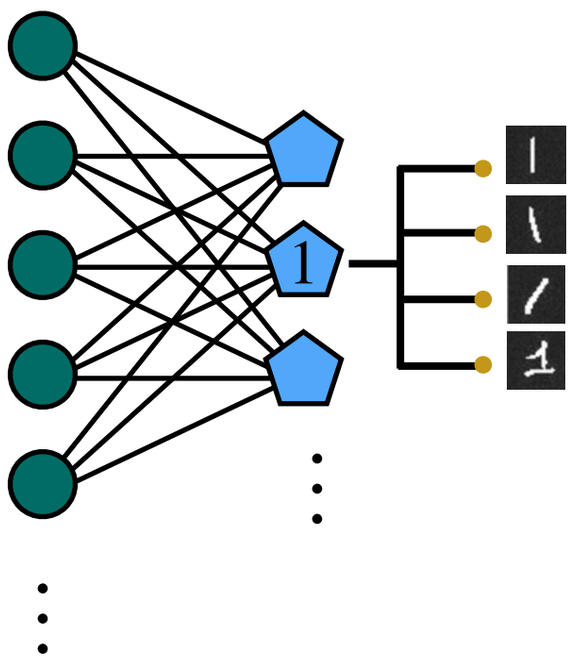


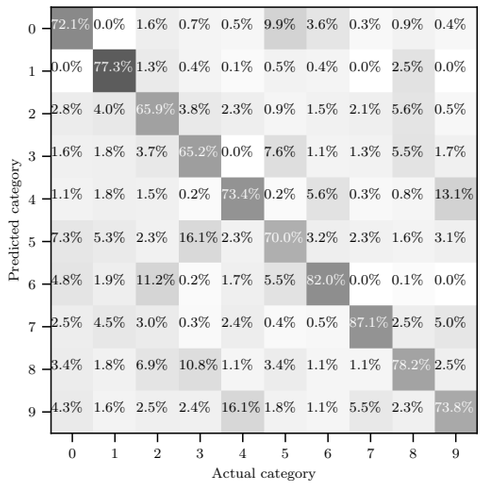


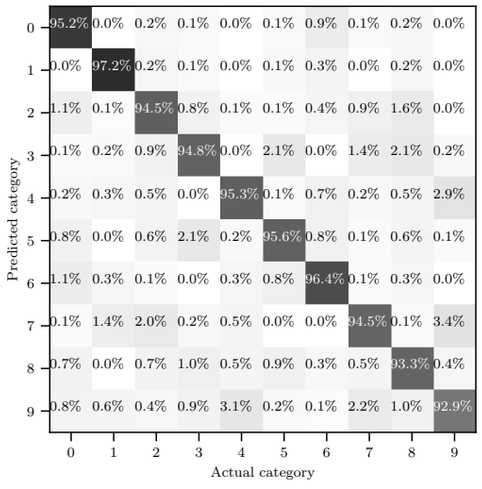
example overlaps

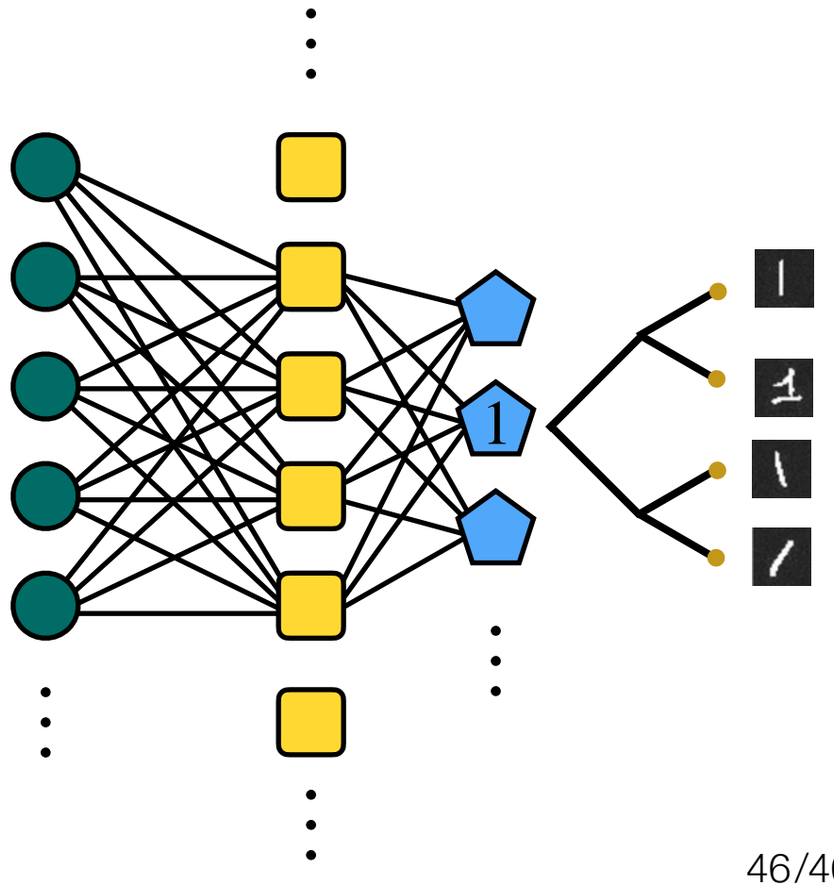
$$\tilde{q}_{ab} = \sum_{i=1}^{N} \xi_i^{(a)} \xi_i^{(b)}$$











THANKS