The 5th East Asian Join SeminEars On Statistical Physics@Beijing

Replica theory on a deep neural network

Cybermedia Center & Department of Physics,
Osaka Univ.

Hajime Yoshino

H. Yoshino,
arXiv1910.09918

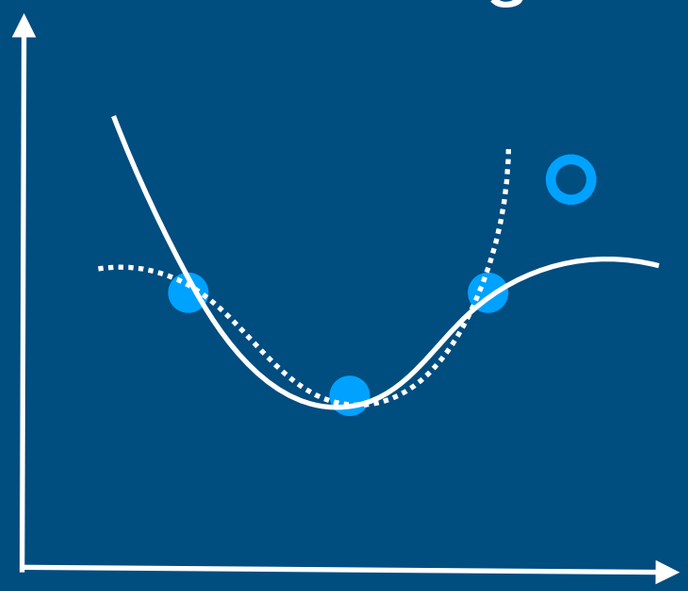
科研費
KAKENHI

Deep networks are typically over-parametrized.

e.g. **data size** $10^6 \ll$ **# of parameters** 10^8

G. Carleo, et. al, arXiv:1903.10563v1

over fitting



Common sense

1. poor generalization

2. many local minimum
learning is difficult
(glassy dynamics)

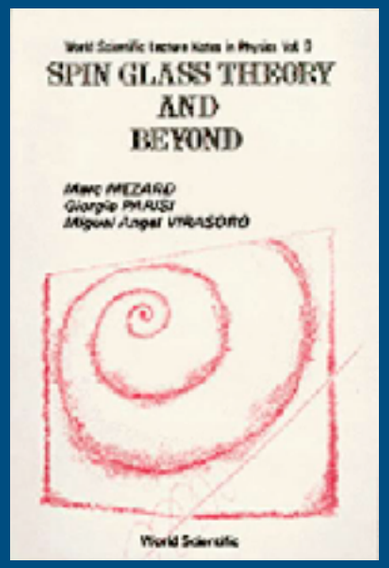
Empirical observations
on deep networks

1. generalization is not bad

???

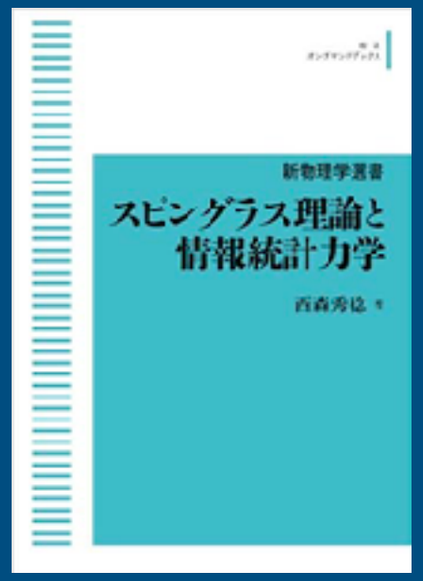
2. learning is not too difficult

???

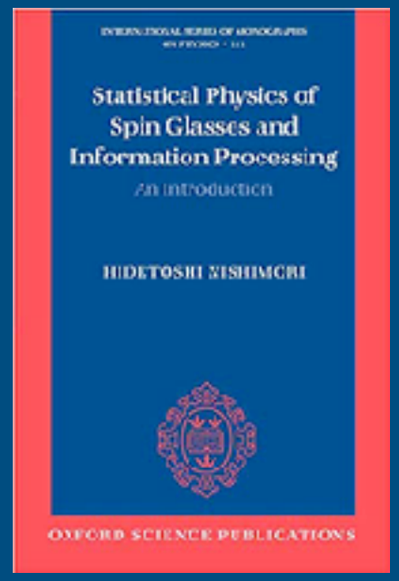


1987

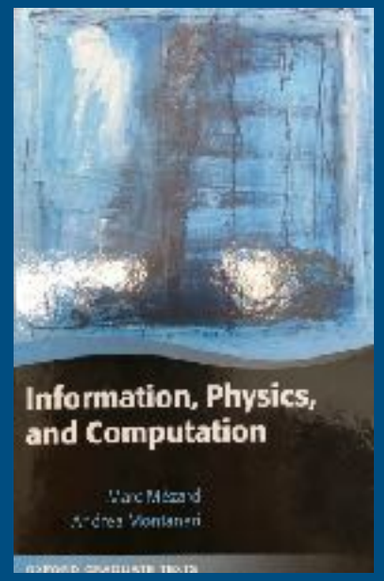
1991



1999



2001



2008

From spin glass to structural glass

p-spin to RFOT **Kirkpatrick-Thirumalai Wolyness (1989)**

more on replicas Franz-Parisi (1995), Monasson (1995), Mezard-Parisi (1999)

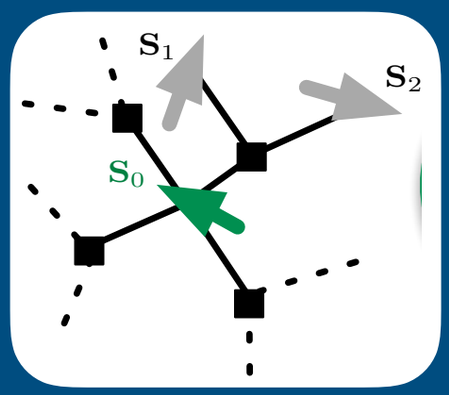
$d \to \infty$ Parisi-Zamponi (2010), Charbonneau-Kurchan-Parisi-Urbani-Zamponi (2014)

shear Yoshino-Mezard (2010), Yoshino-Zamoponi (2014), Rainone-Urbani-Yoshinno-Zamponi (2015)
Jin-Yoshino (2017), Jin-Urbani-Zamponi-Yoshino (2018)

... and back to p-spins... but *with* $M \to \infty$ spin components and without quenched disorder

Using explicit RSB : Parisi-Virasoro (1989)

H. Yoshino, SciPost Phys. 4 (6), 040 (2018)



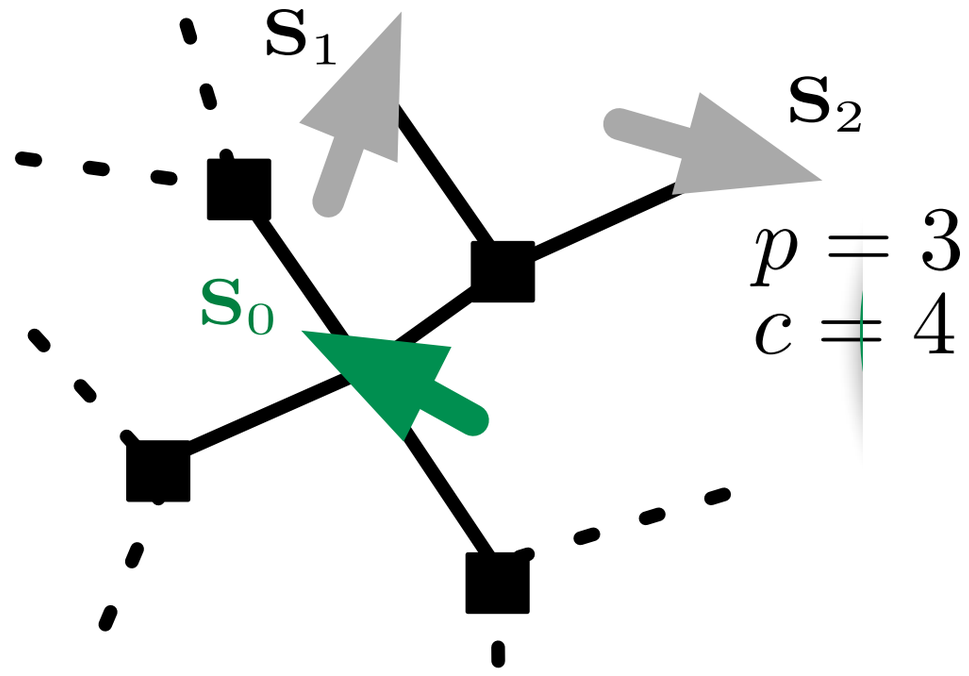
"Disorder-free" vector p-spin models on tree

H. Yoshino, SciPost Phys. 4 (6), 040 (2018)

$$\mathbf{S}_i = (S_i^1, S_i^2, \ldots, S_i^M)$$

$i = 1, 2, \ldots, N$ $\qquad$ $|\mathbf{S_i}^2| = M$

continuous or Ising $\quad S_i^\mu = \pm 1$



Hamiltonian $\quad H = -\sum_{\blacksquare} V(r_{\blacksquare})\,.$

Locally tree like lattice with connectivity c

$$c = \alpha M$$

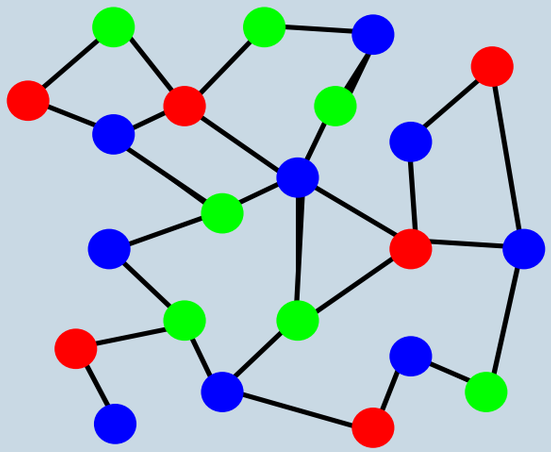
"gap" $\quad r_{\blacksquare} = \delta - \frac{1}{\sqrt{M}} \sum_{\mu=1}^{M} S^\mu_{1(\blacksquare)} S^\mu_{2(\blacksquare)} \cdots S^\mu_{p(\blacksquare)}$

# of factor nodes $\quad N_{\blacksquare} = Nc/p = NM\alpha/p < N^p/p!$

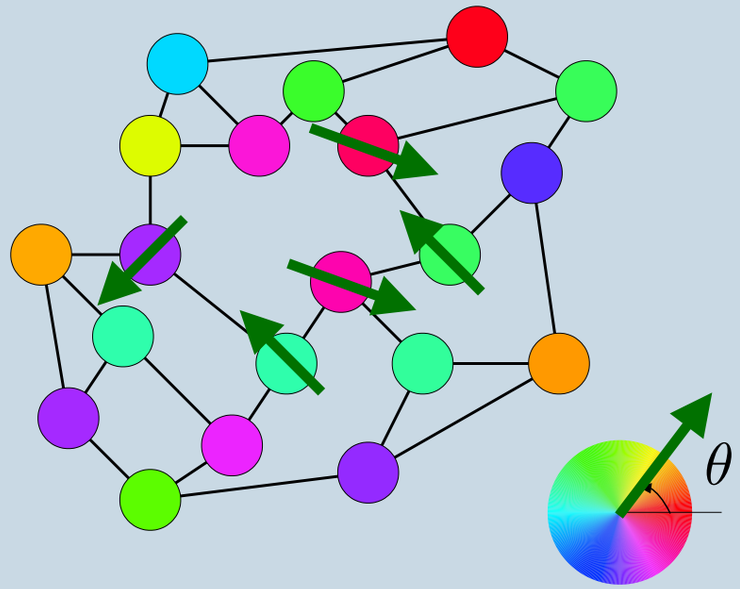
<u>"Inter-mediate sparseness ": high connectivity but not "global coupling"</u>

"Vectorial" constraint satisfaction problems

standard discrete coloring



"continuous" version



antiferromagnetic Potts model

$$H = \sum_{i,j} \delta_{q_i,q_j}$$

"repulsive" vectorial spin model

$$H = \sum_{i,j} V\left(\delta - \frac{\mathbf{S}_i \cdot \mathbf{S}_j}{\sqrt{M}}\right)$$

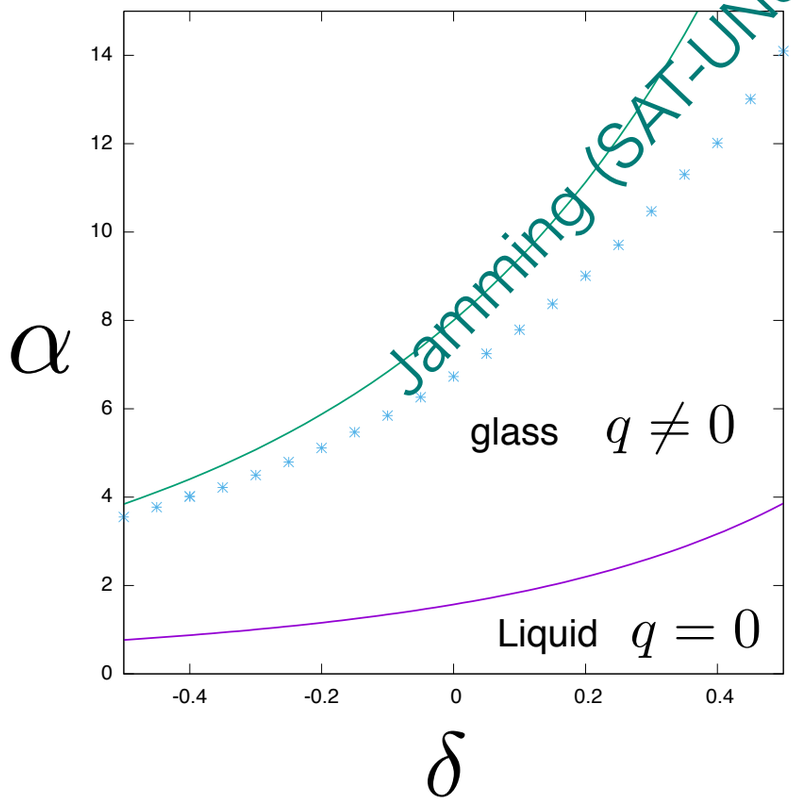
$$V(r) = \lim_{\epsilon \to \infty} \epsilon r^2 \theta(-r)$$

$$M \to \infty$$

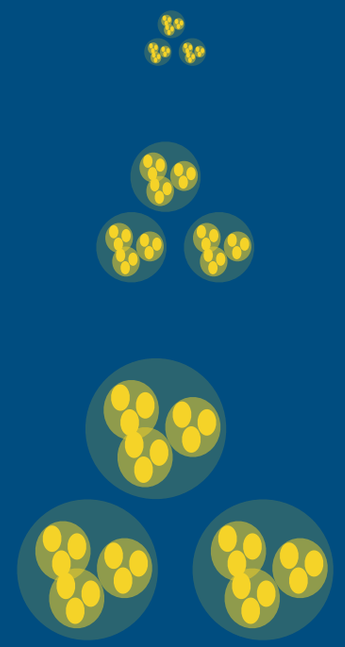
H. Yoshino, SciPost Phys. 4 (6), 040 (2018)

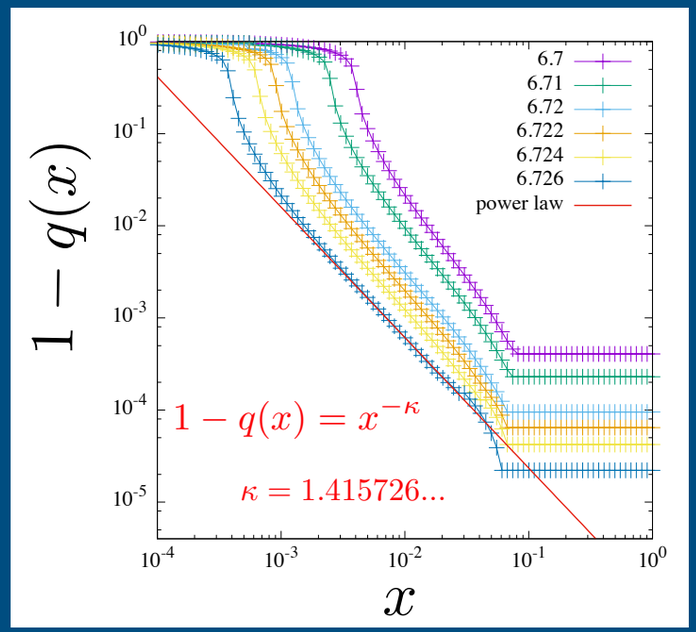
connectivity

$$\alpha = c/M$$



continuous RSB: hierarchical clustering of solutions

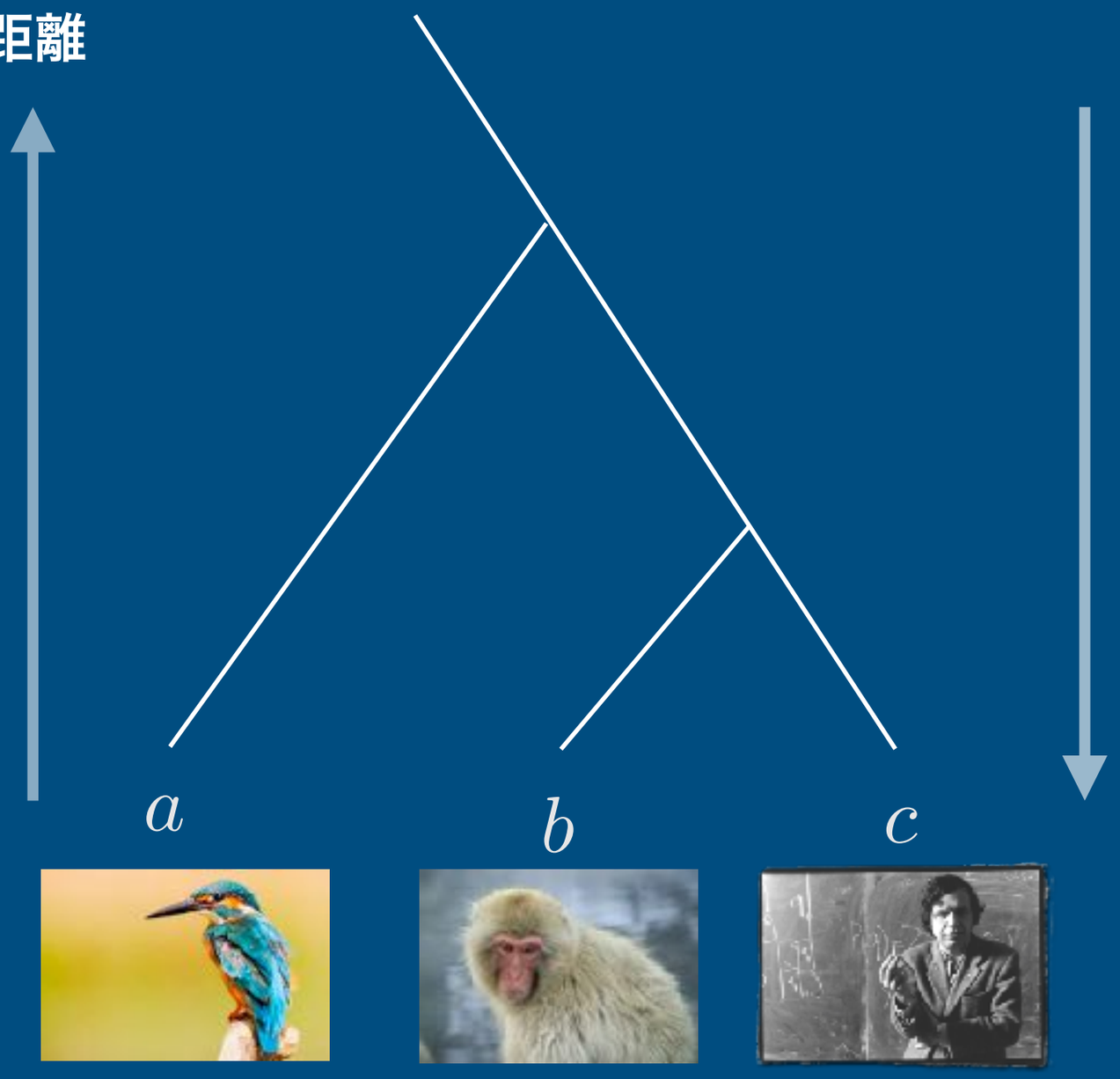




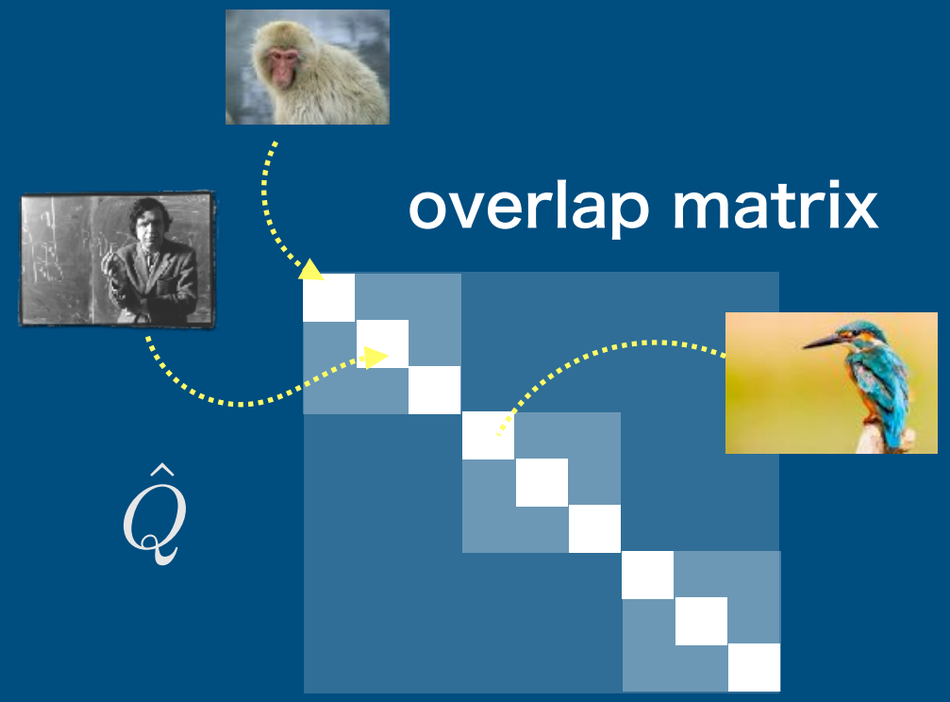
Same jamming criticality as hard spheres and perceptron (p=1)

Franz-Parisi (2016), Franz-Parisi-Sevlev-Urbani-Zamponi (2017)

Replica symmetry braking and ultra-metricity



$$Q(a,b) = \min(Q(a,c), Q(b,c))$$

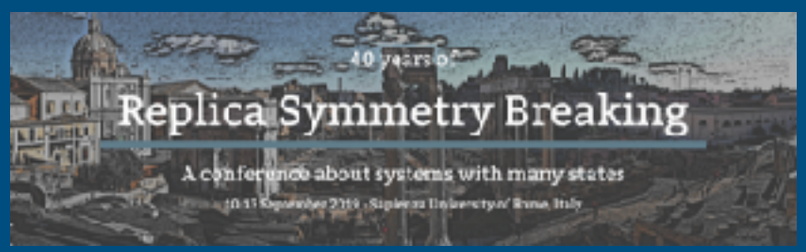


G. Parisi (1979)

first found in the
SK model for spin glass

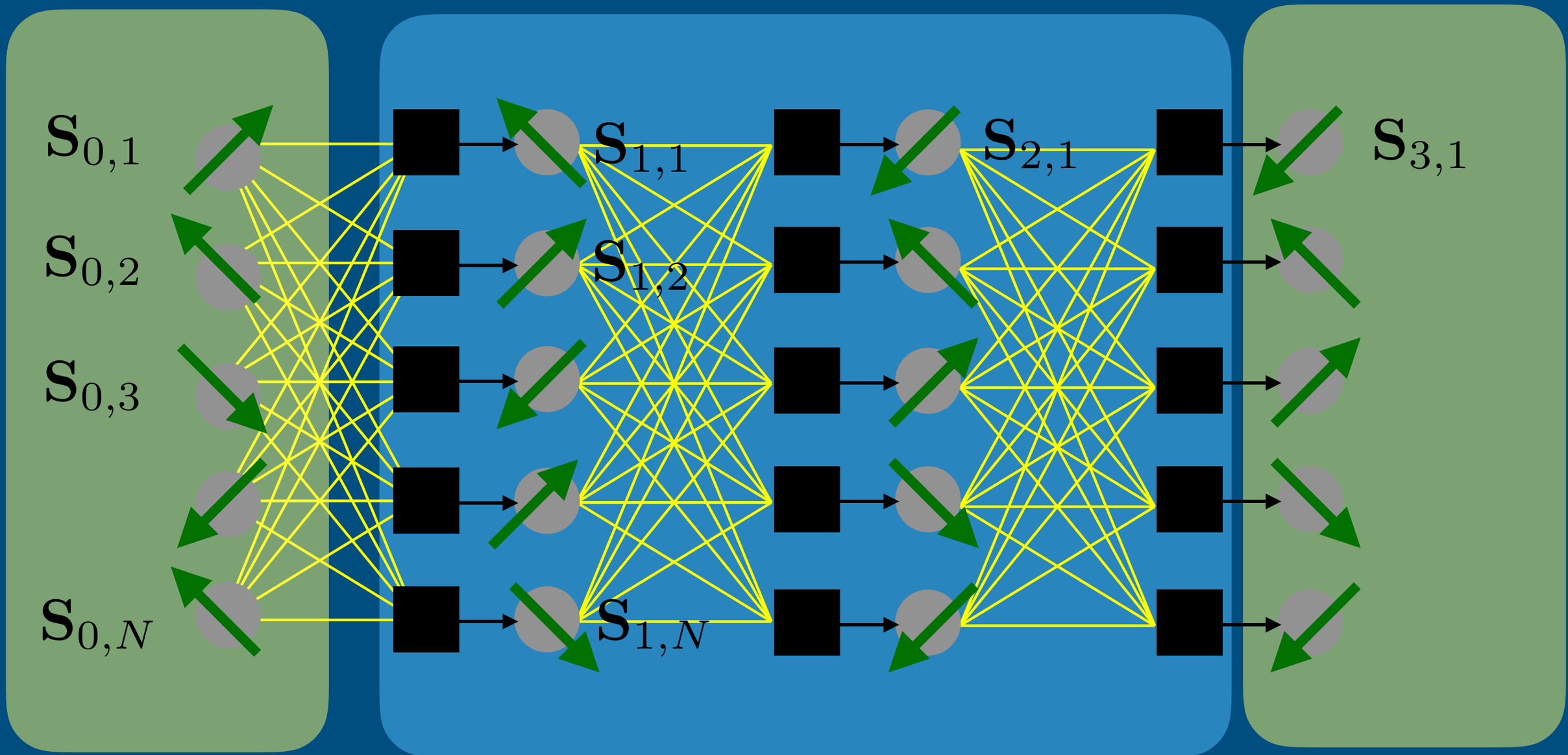
Rigorous proof: M. Talagrand, (2003)

similarity Q



Multi-layer Neural Network

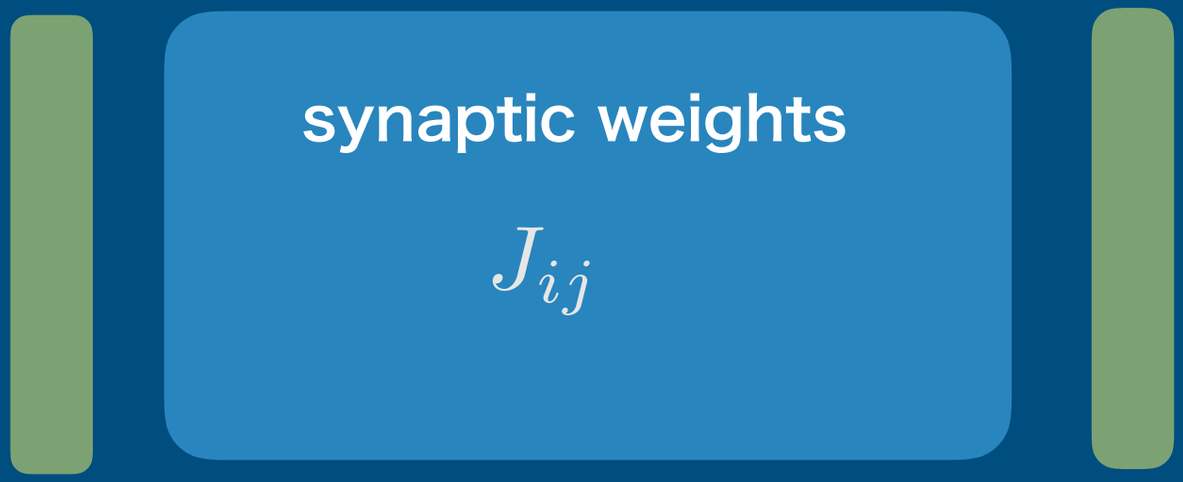
Design weights to satisfy boundary conditions



$$\mathbf{S}_{i,l} = (S^1_{i,l}, S^2_{i,l}, \ldots, S^M_{i,l}) \qquad S^\mu_{i,l} = \pm 1$$

Random inputs/random outpus

a constraint satisfaction problem (CSP)



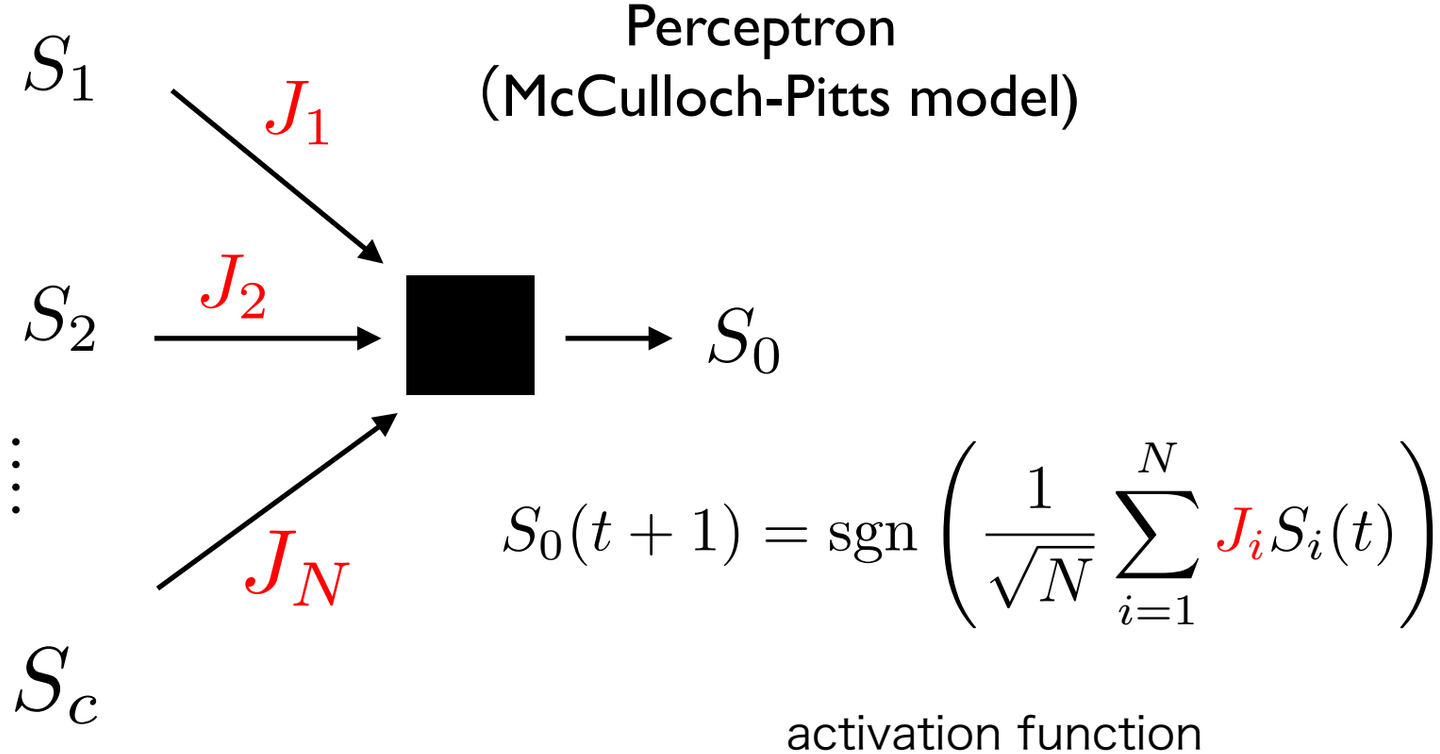
random input

machine

random output

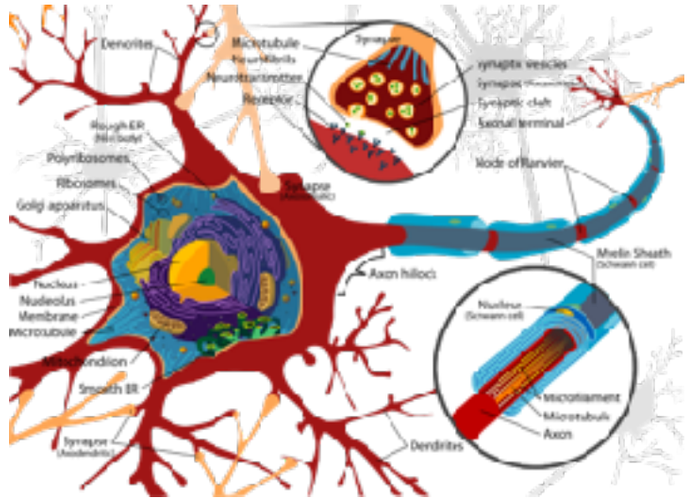
Q: How may different ways the machine can be designed to satisfy the imposed random inputs/outputs ?

Perceptron
(McCulloch-Pitts model)



$$S_0(t+1) = \mathrm{sgn}\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N} J_i S_i(t)\right)$$

activation function



Elisabeth Gardner
(1957-1988)

Statistical mechanics on the “ensemble of fixed points”

fixed point patterns

$$S_i(t) = S_i^\mu \qquad \mu = 1, 2, \ldots, M$$

$$\alpha = \frac{M}{N} \qquad M \to \infty \text{ with fixed } \alpha$$

Gardner volume

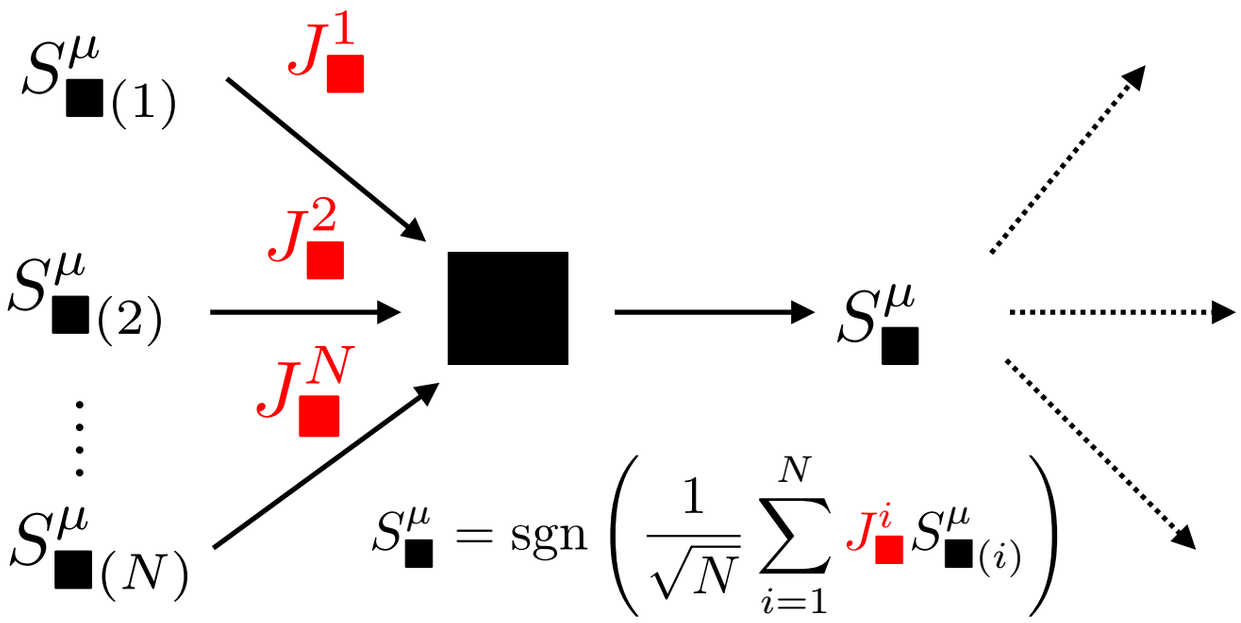
$$V = \int \prod_{j=1}^{N} \frac{dJ_j}{\sqrt{2\pi}} e^{-\frac{J_j{}^2}{2}} \prod_{\mu=1}^{M} e^{-\beta V(r^\mu)}$$

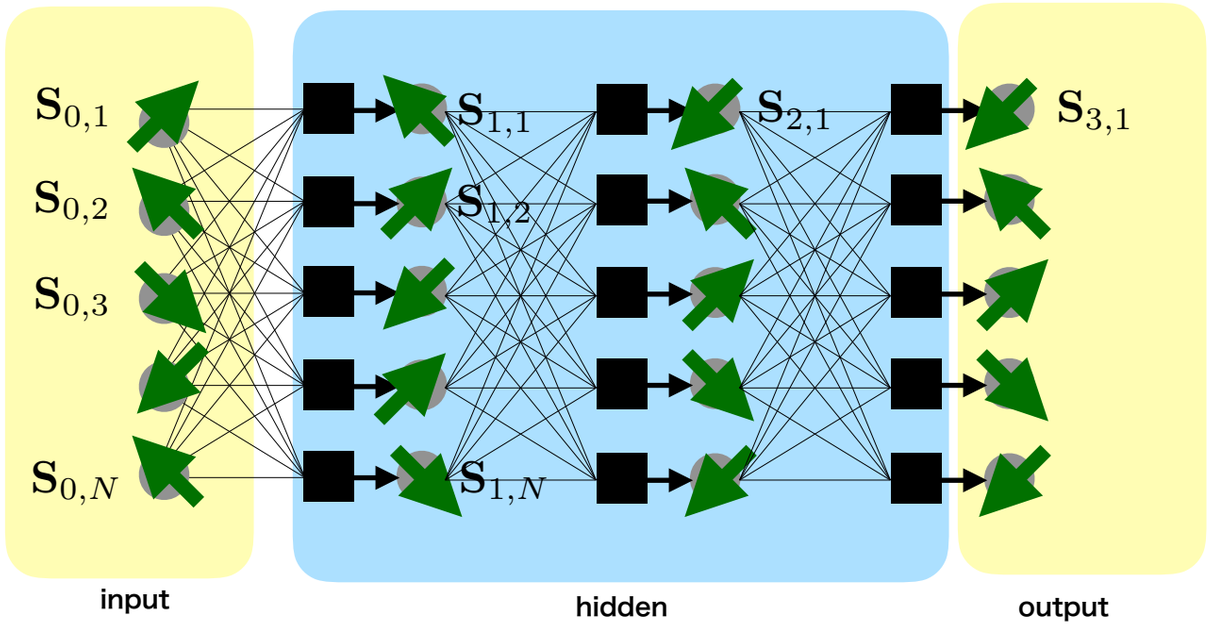
“Hardcore” constraint

$$e^{-\beta V(h)} = \theta(h)$$

“Gap”

$$r^\mu = S_0^\mu \sum_{i=1}^{N} \frac{1}{\sqrt{N}} J_i S_i^\mu$$





$$S^\mu_{L,i}(t+1) = \mathrm{sgn}\left(\frac{1}{\sqrt{N}}\sum_{j=1}^{N} J_{L,i,j}\,\mathrm{sgn}\left(\frac{1}{\sqrt{N}}\sum_{k=1}^{N} J_{L-1,j,k}\cdots\mathrm{sgn}\left(\frac{1}{\sqrt{N}}\sum_{m=1}^{N} J_{1,l,m} S^\mu_{0,m}(t)\right)\right)\right)$$

Usual strategy of learning

(1) define "loss function"

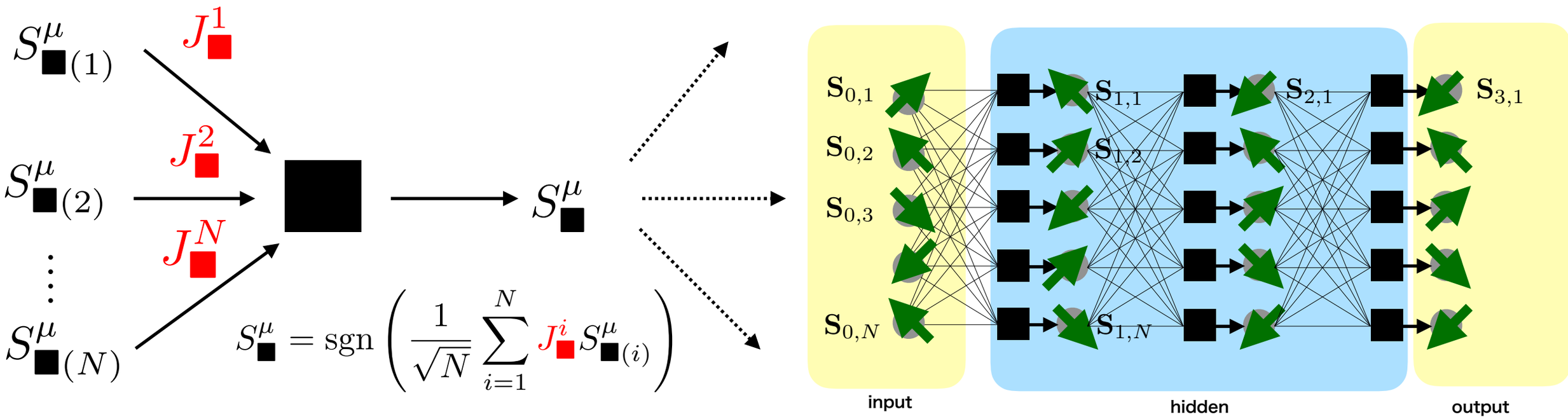
(2) try to minimize the loss function via back-propagation

desired output

e.g. $E = \sum_{i=1}^{N}\sum_{\mu=1}^{M}\left(S^\mu_{L,i} - (S_*)^\mu_{L,i}\right)^2$

e.g. SDG (stochastic gradient descent)

Too much long-ranged, highly convoluted, non-linear interaction! …hard to analyze



Gardner volume generalized for a multi-layer network

trace over hidden variables

$$V(\mathbf{S}(0),\mathbf{S}(L)) = e^{NM\mathcal{S}(\mathbf{S}(0),\mathbf{S}(l))} = \left(\prod_{l=1}^{L-1}\prod_{i=1}^{N}\sum_{S^\mu_{l,i}=\pm 1}\right)\left(\int\prod_{\blacksquare}\prod_{j=1}^{N}\frac{dJ^j_{\blacksquare}}{\sqrt{2\pi}}e^{-\frac{(J^j_{\blacksquare})^2}{2}}\right)e^{-\beta H}$$

Hamiltonian with "short-ranged" interactions

$$H = \sum_{\mu=1}^{M}\sum_{\blacksquare} V(r^\mu_{\blacksquare})$$

"Gap"

$$r^\mu_{\blacksquare} = \sum_{i=1}^{N}\frac{1}{\sqrt{N}} J^i_{\blacksquare} S^\mu_{\blacksquare(i)} S^\mu_{\blacksquare}$$

"Hardcore" constraint

$$e^{-\beta V(h)} = \theta(h)$$

Gaussian approx. or modified model

$$r^\mu_{\blacksquare} = \sum_{i=1}^{N}\frac{1}{\sqrt{N}} J^i_{\blacksquare}\left(\frac{1}{\sqrt{M}}\sum_{\nu=1}^{M}\xi^{\mu\nu} S^\nu_{\blacksquare(i)} S^\nu_{\blacksquare}\right)$$

$\xi^{\mu\nu}$: Gaussian with zero mean and variance 1

(c.f.) internal representation (2-layer): R. Monasson and R. Zecchina (1995)

Replicated Gardner volume

replicas: machines learning in parallel

$$V^n(\mathbf{S}_0,\mathbf{S}_L) = \prod_{a=1}^{n}\left(\prod_{\blacksquare}\mathrm{Tr}_{\mathbf{J}^a_{\blacksquare}}\right)\left(\prod_{\blacksquare\backslash\mathrm{output}}\mathrm{Tr}_{\mathbf{S}^a_{\blacksquare}}\right)\prod_{\mu,\blacksquare,a}e^{-\beta V(r^{\mu}_{\blacksquare,a})}$$

$$r^{\mu}_{\blacksquare,a} = S^{\mu}_{\blacksquare,a}\sum_{i=1}^{N}\frac{1}{\sqrt{N}}J^{i}_{\blacksquare,a}S^{\mu}_{\blacksquare(i),a}$$

Glass order parameters

$$q_{ab,\blacksquare} = \frac{1}{M}\sum_{\mu=1}^{M}(S^{\mu}_{\blacksquare})^a(S^{\mu}_{\blacksquare})^b \qquad Q_{ab,\blacksquare} = \frac{1}{N}\sum_{i=1}^{N}J^{a}_{\blacksquare(i)}J^{b}_{\blacksquare(i)}$$

Replicated free-energy

$$\frac{-\beta\overline{F(\mathbf{S}_0,\mathbf{S}_L)}^{\mathrm{visible}}}{NM} = \frac{\partial_n\overline{V^n(\mathbf{S}_0,\mathbf{S}_L)}^{\mathrm{visible}}\Big|_{n=0}}{NM} = S_n[\{\hat{Q}(l),\hat{q}(l)\}]$$

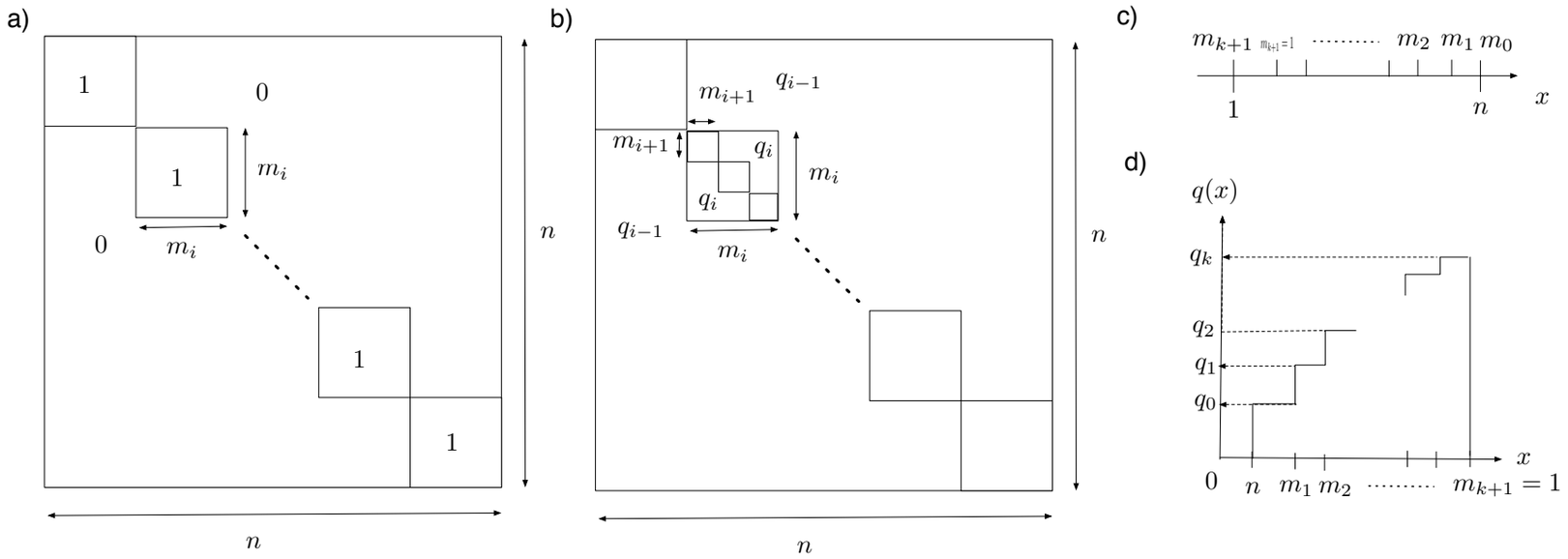
$$S_n[\{\hat{q}(l)\},\{\hat{Q}(l)\}] = \alpha^{-1}\sum_{l=1}^{L}S^{\mathrm{bond}}_{\mathrm{ent}}[\hat{Q}(l)] + \sum_{l=1}^{L-1}S^{\mathrm{spin}}_{\mathrm{ent}}[\hat{q}(l)]$$

$$- \sum_{l=1}^{L}e^{\frac{1}{2}\sum_{ab}q_{ab}(l-1)Q_{ab}(l)q_{ab}(l)\partial_{h_a(l)}\partial_{h_b(l)}}\prod_{a=1}^{n}e^{-\beta V(h_a(l))}\Big|_{h_a(l)=0}$$

$$\alpha = \frac{M}{N}$$

quenched random input/output $\quad q_{ab}(0) = q_{ab}(L) = 1$

■ Parisi's RSB ansatz



overlap distribution function

$$P(q) = \frac{dx(q)}{dq}$$

$$Q_{ab}(l) = \sum_{i=0}^{k+1} Q_i(l)(I_{ab}^{m_i} - I_{ab}^{m_{i+1}}) \qquad l = 1, 2, \ldots, L$$

$$q_{ab}(l) = \sum_{i=0}^{k+1} q_i(l)(I_{ab}^{m_i} - I_{ab}^{m_{i+1}}) \qquad l = 1, 2, \ldots, L-1$$

$$\varepsilon_{ab}(l) = \sum_{i=0}^{k} \varepsilon_i(l)(I_{ab}^{m_i} - I_{ab}^{m_{i+1}}) \qquad l = 1, 2, \ldots, L-1$$

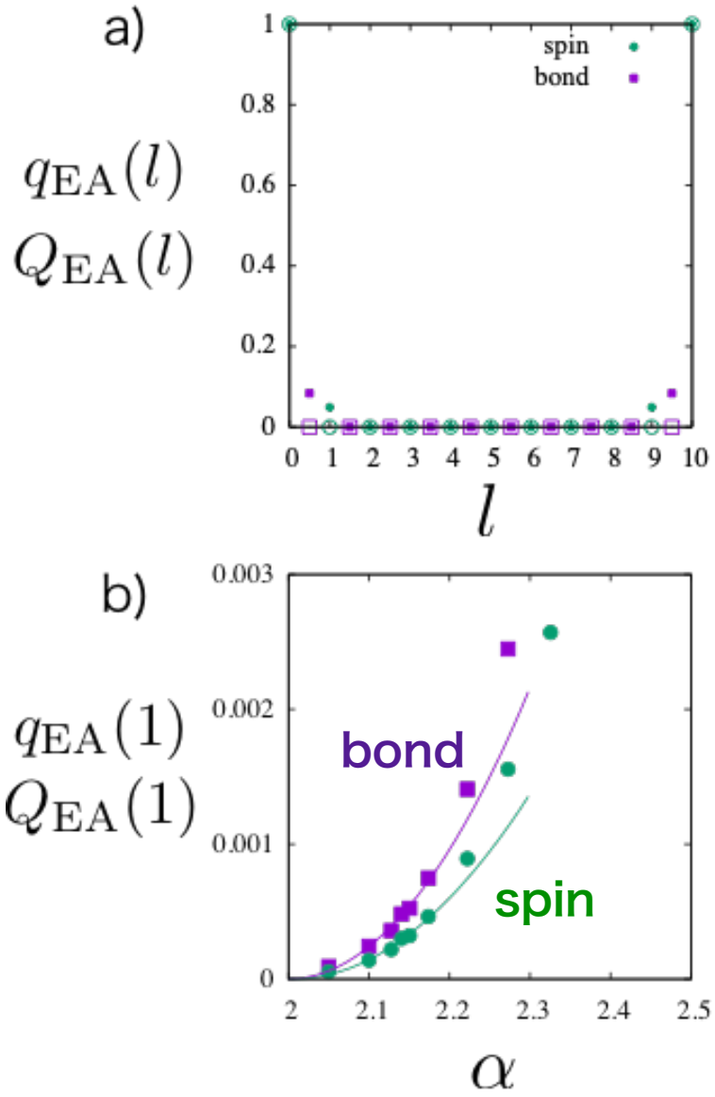
entropic part (bond)

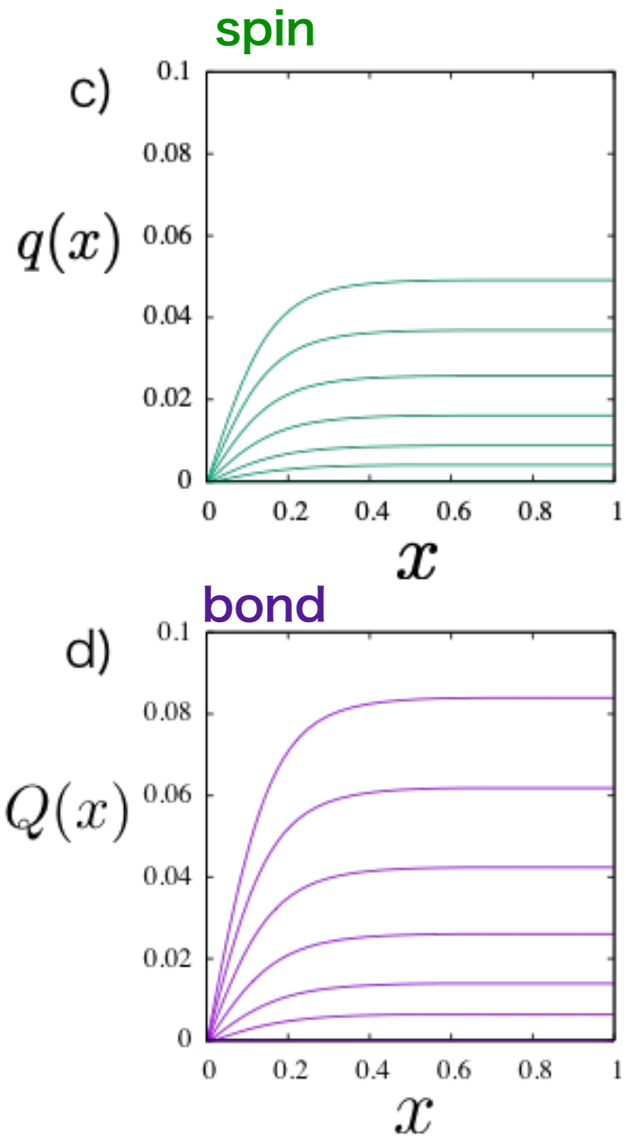
$$S_{\rm ent}[\hat{Q}(l)] = \frac{1}{2}\ln \det \hat{Q}(l)$$

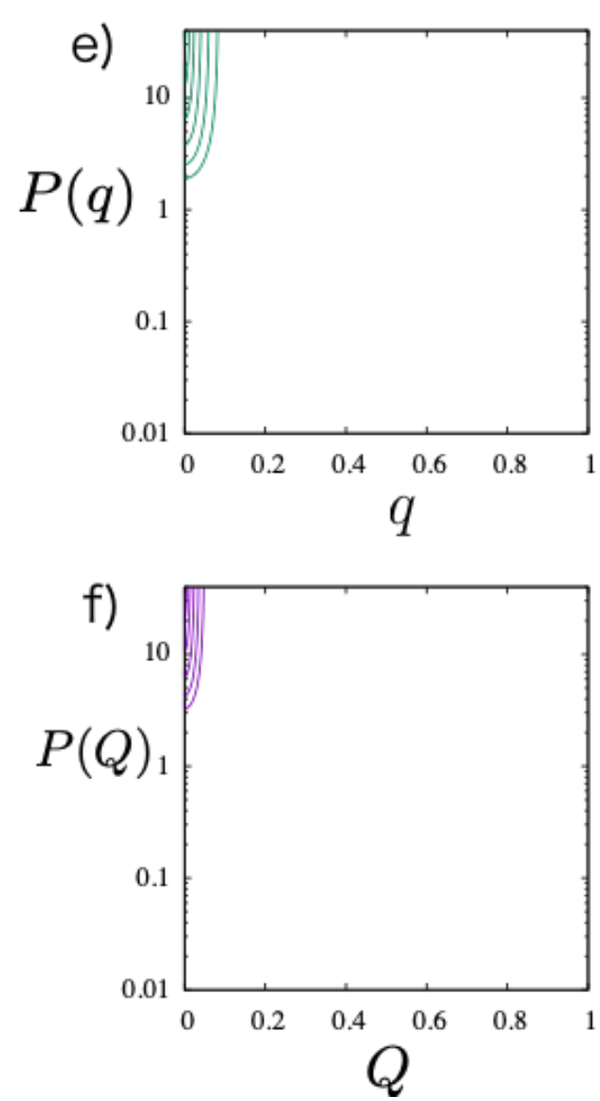
entropic part (spin)

$$S_{\rm ent}[\hat{q}(l)] = \sum_{a<b} \ln e^{-\sum_{a<b}\varepsilon^*_{ab}(l)\frac{\partial^2}{\partial h_a \partial h_b}} \prod_a 2\cosh(h_a)\Big|_{\{h_a=0\}} \qquad q_{ab} = -\frac{\delta}{\delta \varepsilon_{ab}} \ln e^{-\sum_{a<b}\varepsilon_{ab}\frac{\partial^2}{\partial h_a \partial h_b}} \prod_a 2\cosh(h_a)\Big|_{\{h_a=0\}}\Big|_{\varepsilon_{ab}=\varepsilon^*_{ab}[\hat{q}]}$$

■1st Glass transition







$\alpha_{\rm g} \simeq 2.03$

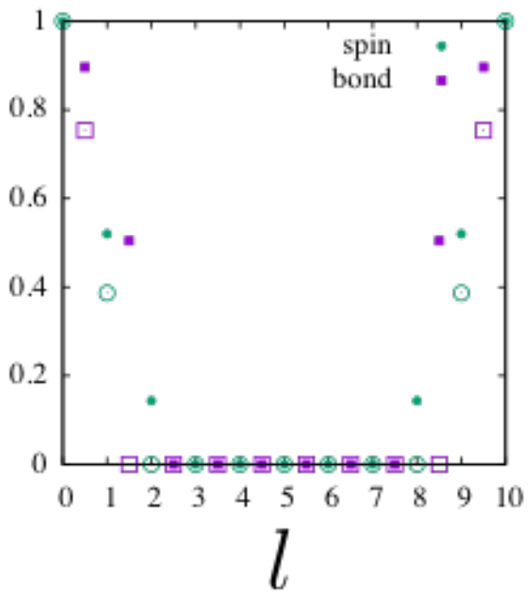
continuous transition to full RSB glass phase
at 1 st & (L-1) th layer
other layers remain in the liquid phase

■2nd Glass transition

a)

$q_{\rm EA}(l)$

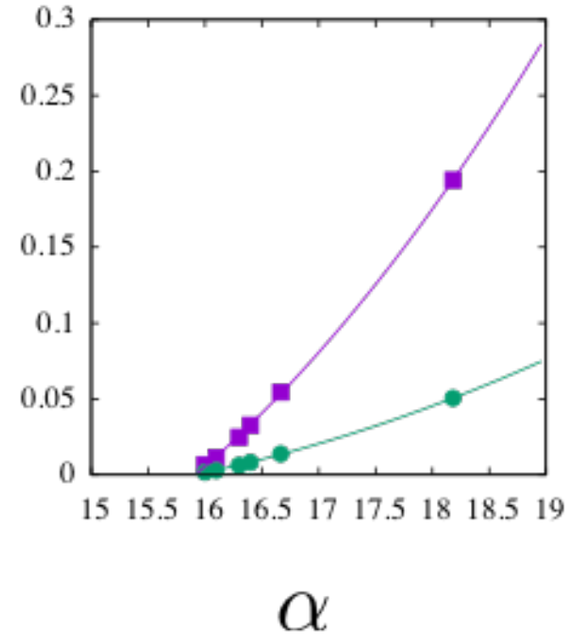
$Q_{\rm EA}(l)$



b)

$q_{\rm EA}(2)$

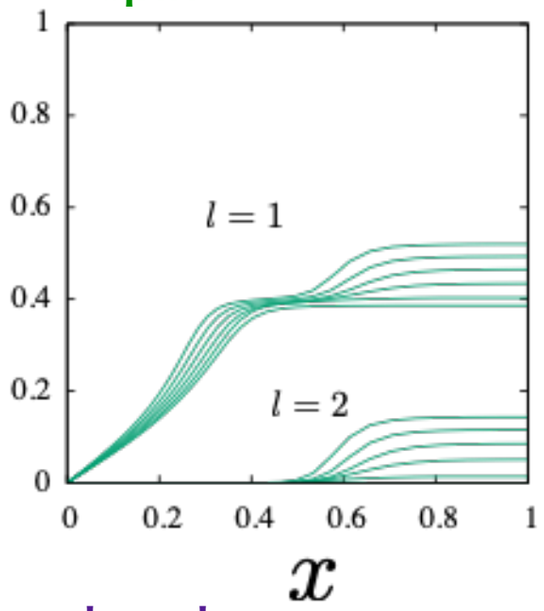
$Q_{\rm EA}(2)$



spin

c)

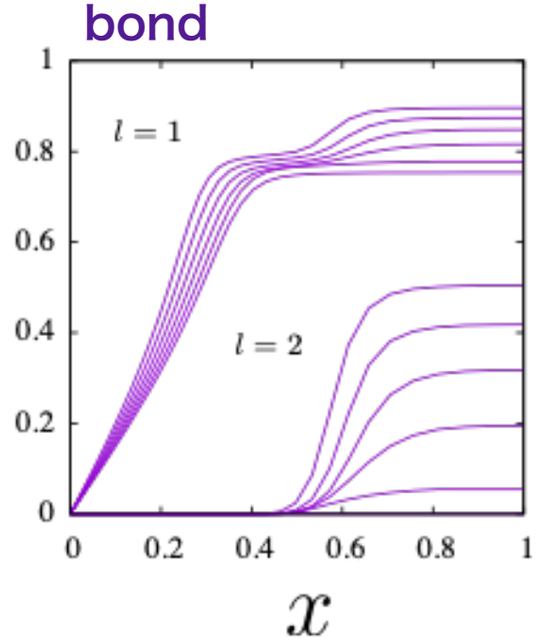
$q(x)$



bond

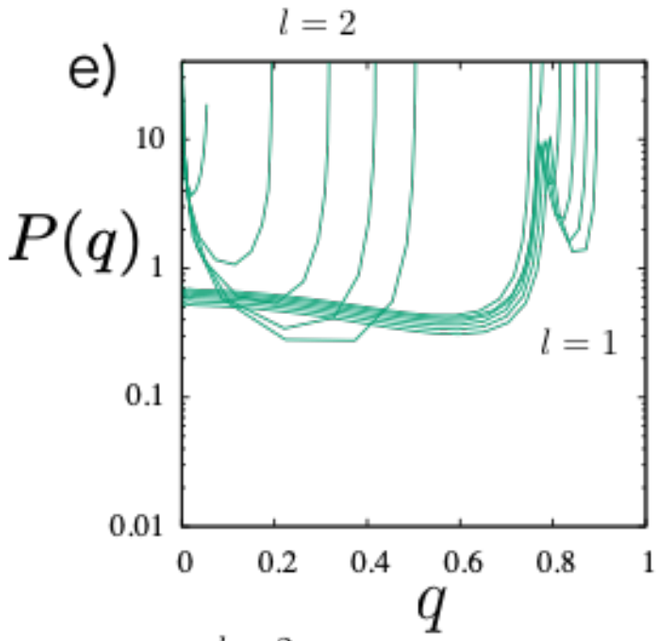
d)

$Q(x)$



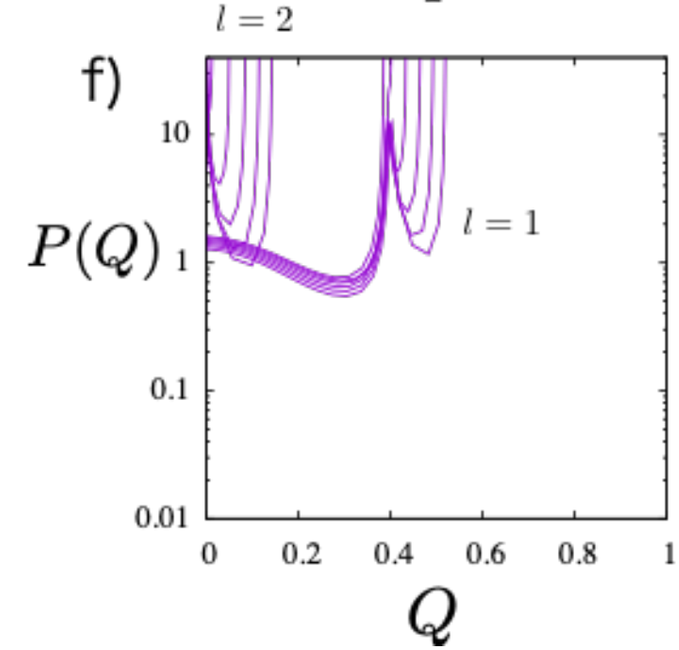
e)

$P(q)$



f)

$P(Q)$

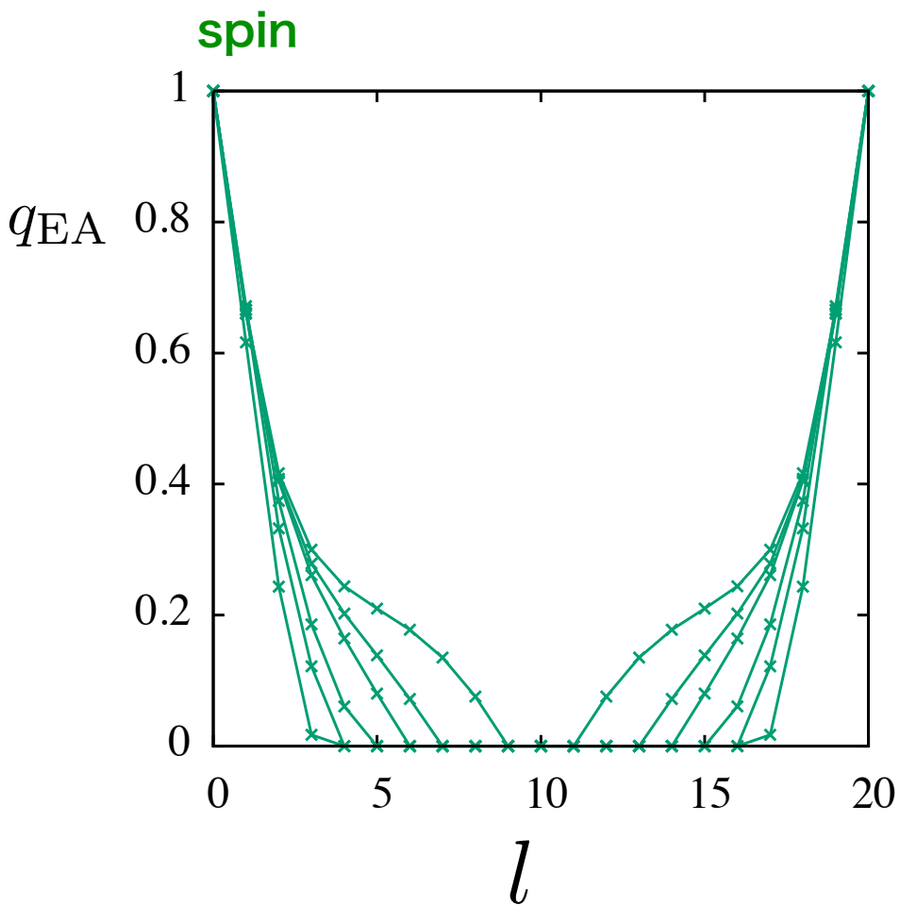


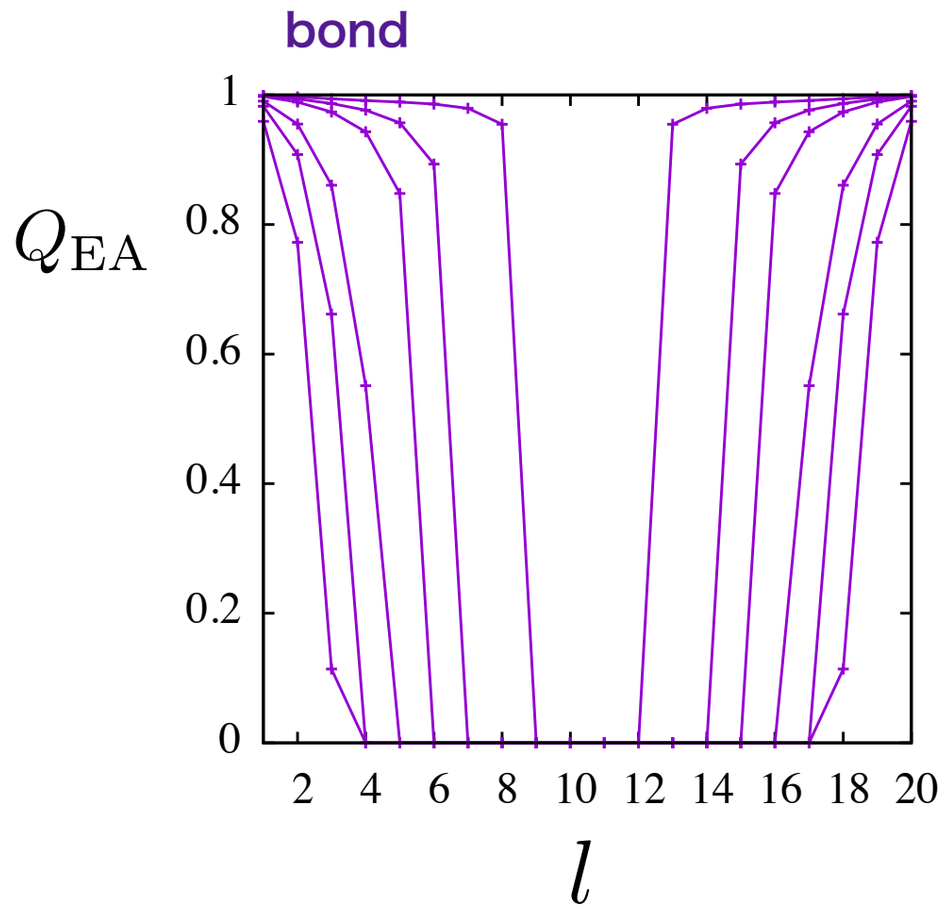
$\alpha_{\rm g}(2) \simeq 15.38$ continuous transition to full RSB glass phase

at 2nd & (L-2) th layer

which also induce 2nd glass transitions at 1st and L-th layer

■Growth of glass phase under larger constrains





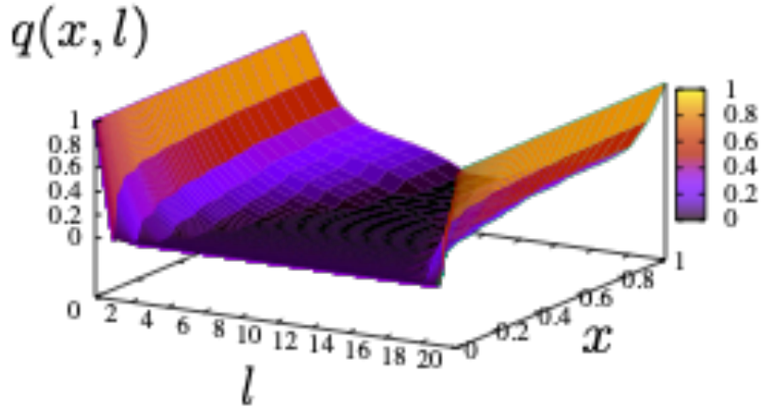
$1/\alpha = 0.02, 0.01, 0.005, 0.001, 0.0005, 0.00025$

"penetration depth" $\xi(\alpha) \propto \ln \alpha$

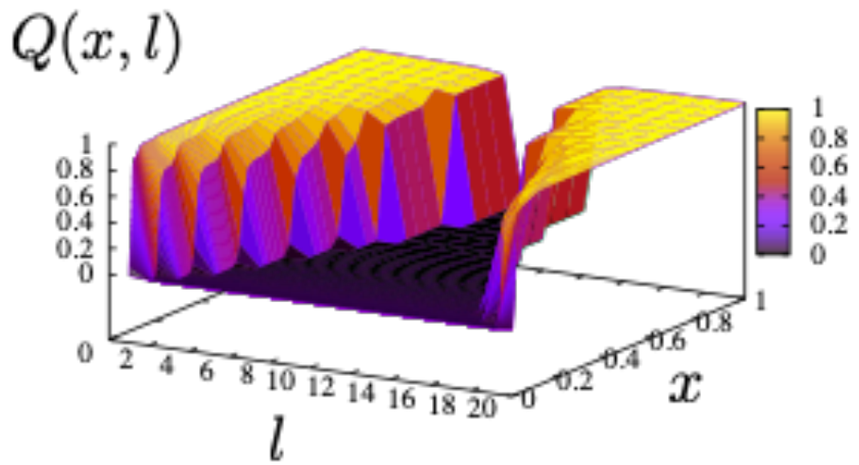
■ More “terraces” under larger constraints

$$\alpha = 4000$$

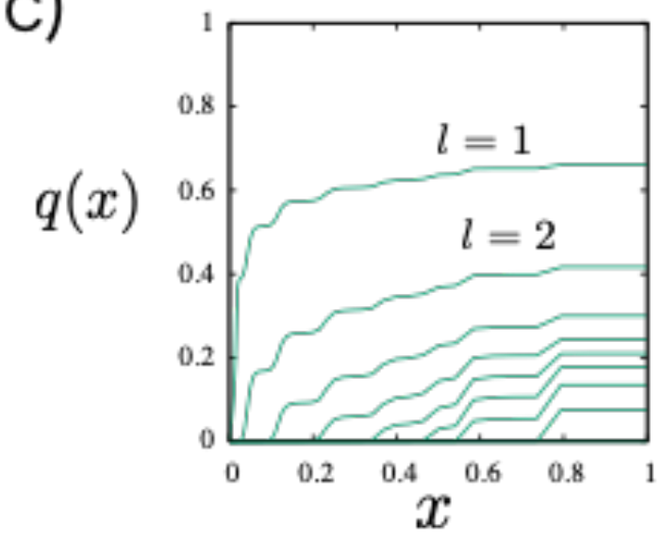
a)



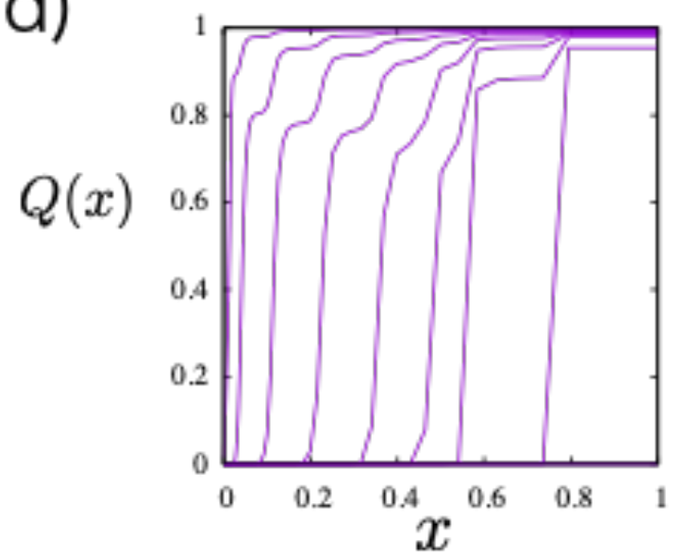
b)



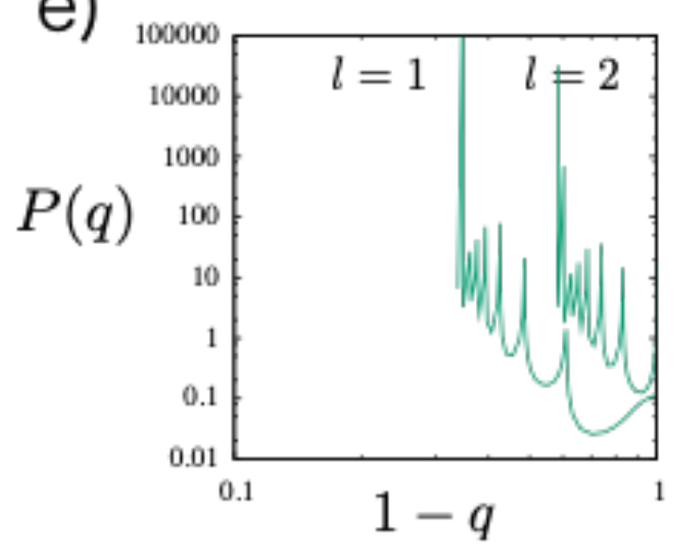
c)



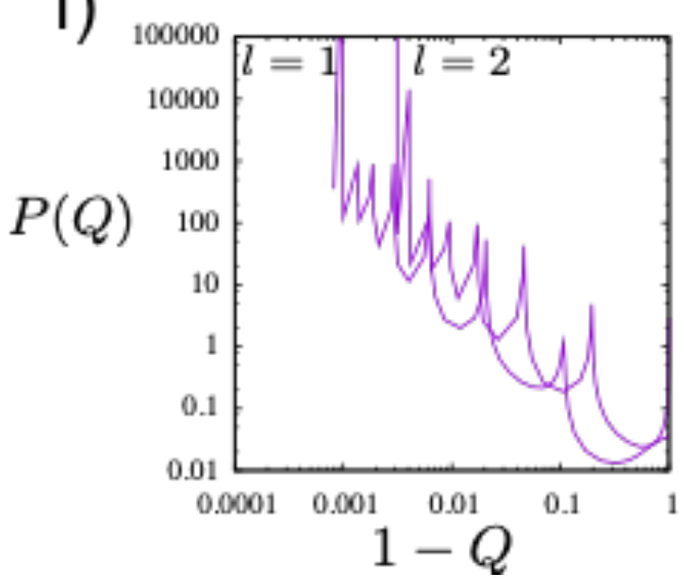
d)



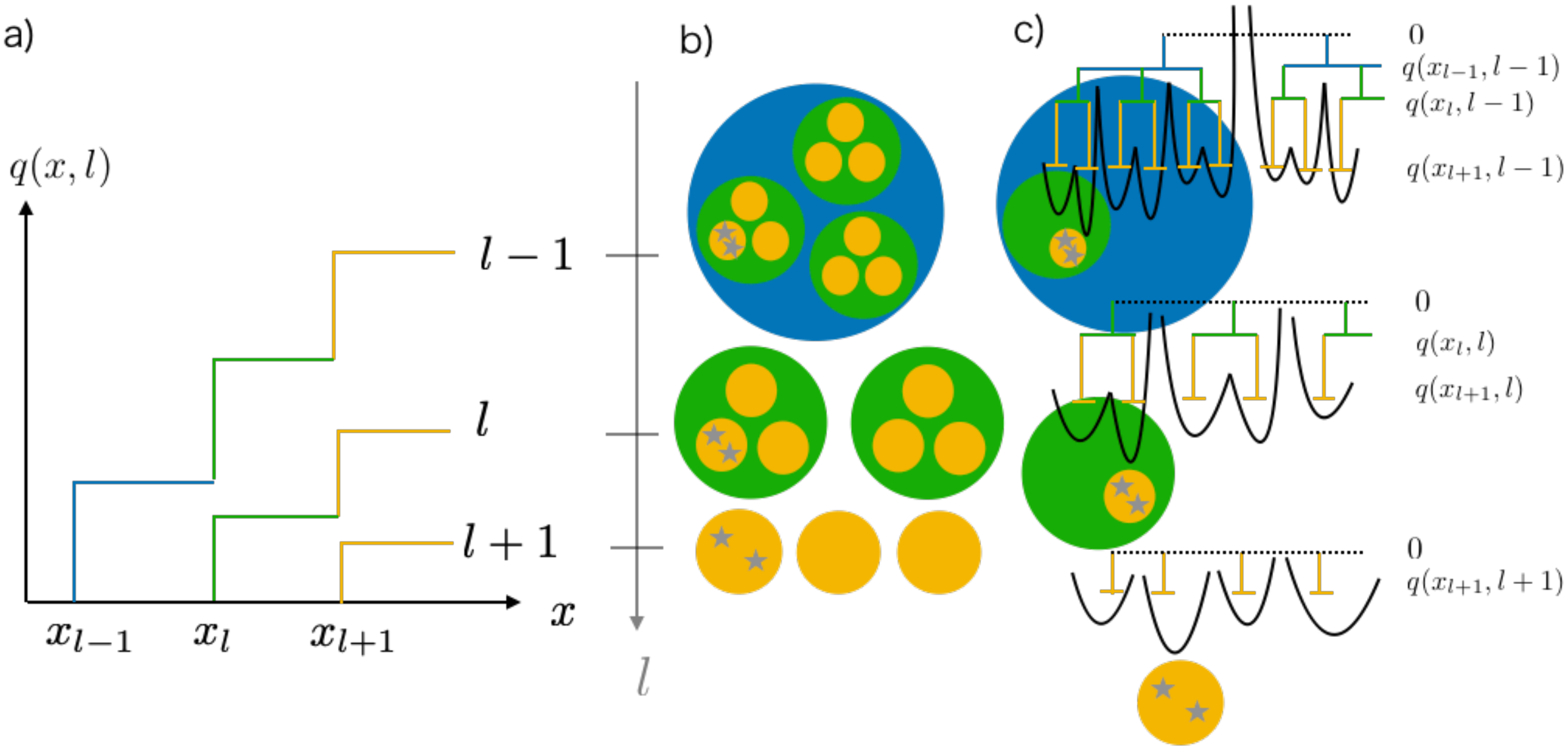
e)



f)



■ Summary: depth dependent free-energy landscape



Teacher student setting

a statistical inference problem (SIP)

1) Training

random test data

Randomly quenched

$J_{ij}^{\mathrm{teacher}}$

$J_{ij}^{\mathrm{student}}$

Out put of teacher

Student is forced to reproduce teacher's output

2) Test

random test data

Randomly quenched

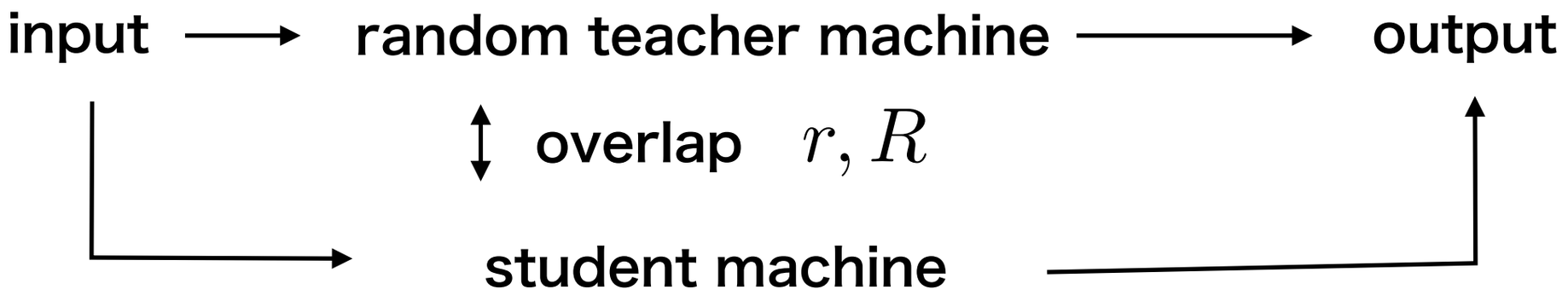
$J_{ij}^{\mathrm{teacher}}$

Now quenched

$J_{ij}^{\mathrm{student}}$

Compare

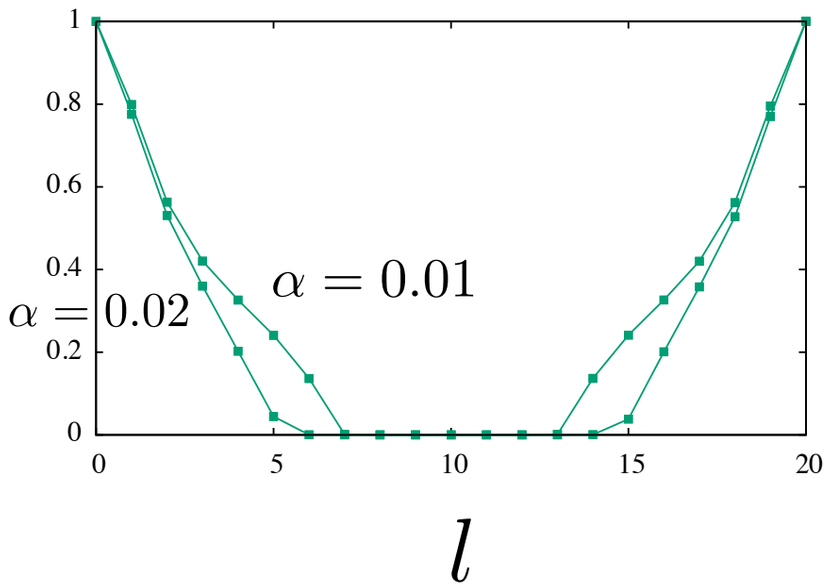
teacher+student machine (inference)



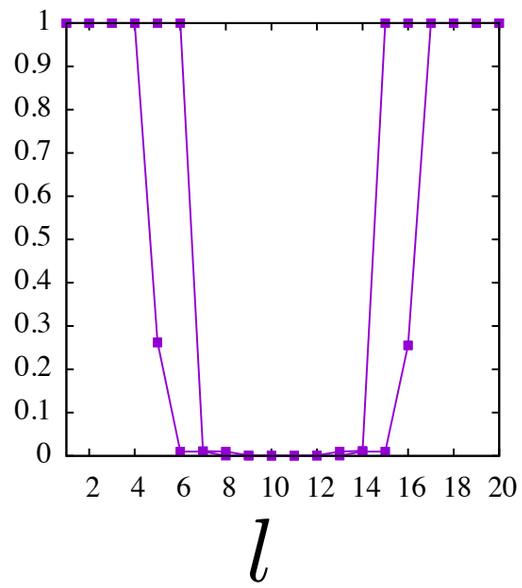
Review: Zdeborova-Krzakala (2017)

1+s replica　　RS solution　　Bays optimal, Nishimori condition

r, q



R, Q

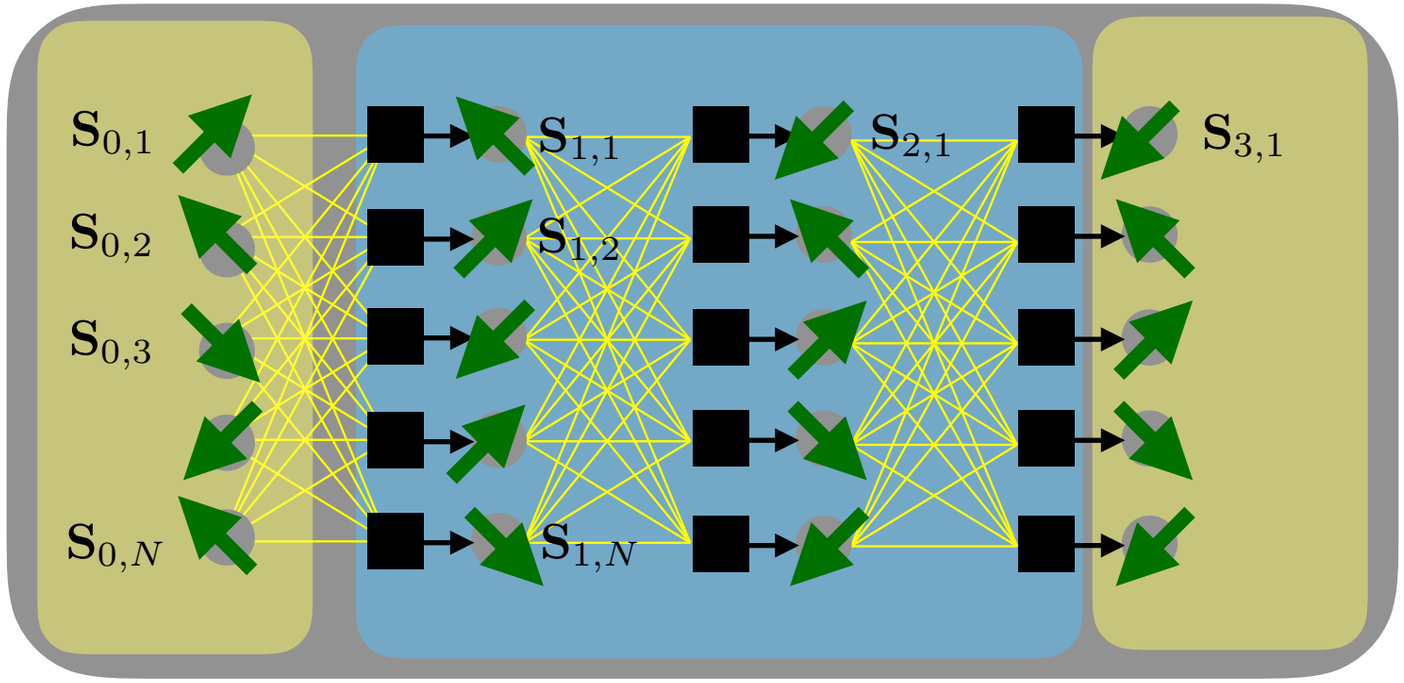


Symmetry breaking field: remanent bias in the liquid phase

$$O(\ln(N)/N)$$

Simulations of learning

■ random inputs/outputs



Monte Carlo simulation

Hamiltonian

$$H = -\sum_{\blacksquare}\sum_{\mu=1}^{M} V\left(r_{\blacksquare}^{\mu}\right)$$

soft-core potential

$$e^{-\beta V(h)} = e^{-\beta\epsilon h^2}\theta(h) \qquad \beta = \frac{1}{k_{\mathrm{B}}T}$$

Gap

$$r_{\blacksquare}^{\mu} = \sum_{i=1}^{N} \frac{1}{\sqrt{N}} J_{\blacksquare}^{i} S_{\blacksquare(i)}^{\mu} S_{\blacksquare}^{\mu}$$

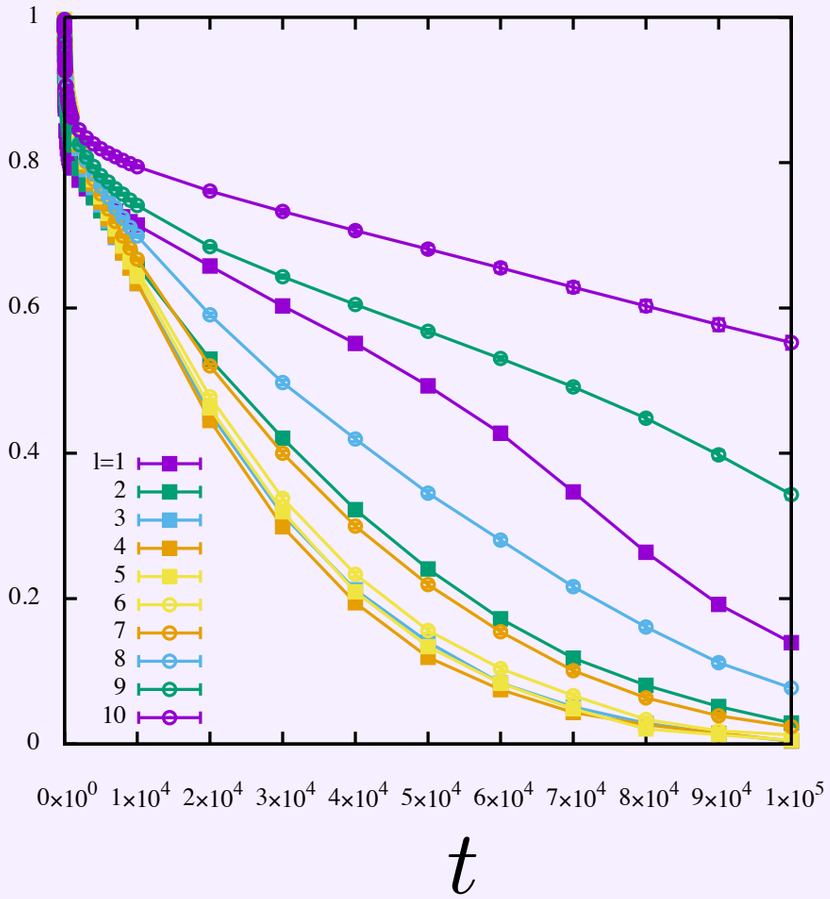
Dynamical variables $J_{\blacksquare}^{i}, S_{\blacksquare}^{\mu}$ $(i = 1, 2, \ldots, N)(\mu = 1, 2, \ldots, M)$ $(\blacksquare = 1, \ldots, LN)$

with random boundaries $\mathbf{S}_{0,i}, \mathbf{S}_{L,i}$ $(i = 1, 2, \ldots, N)$

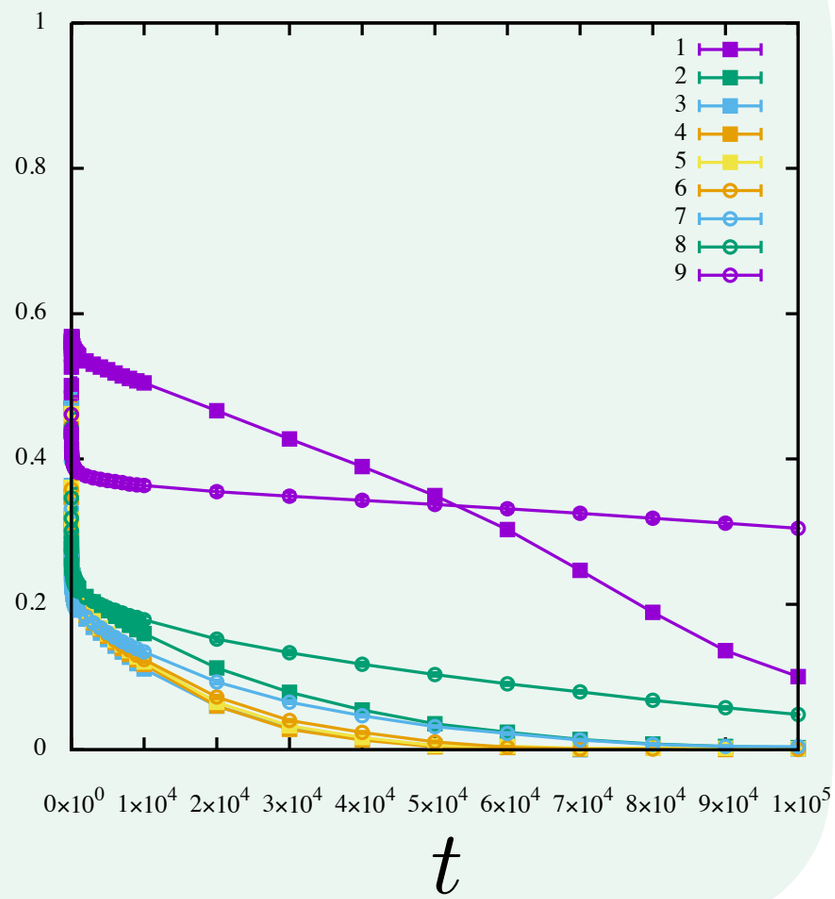
Relaxation of autocorrelation functions

$$C_{\mathrm{bond}}(t,\blacksquare) = \frac{1}{N}\sum_{i=1}^{N}\langle J^i_{\blacksquare}(0)J^i_{\blacksquare}(t)\rangle \qquad C_{\mathrm{spin}}(t,\blacksquare) = \frac{1}{M}\sum_{\mu=1}^{M}\langle S^\mu_{\blacksquare}(0)S^\mu_{\blacksquare}(t)\rangle$$

$C_{\mathrm{bond}}(t,l)$



$C_{\mathrm{spin}}(t,l)$



$N = 20, M = 200(\alpha = 10)$ $L = 10$ $T = 0.015$ 240 samples

■ teacher-student setting

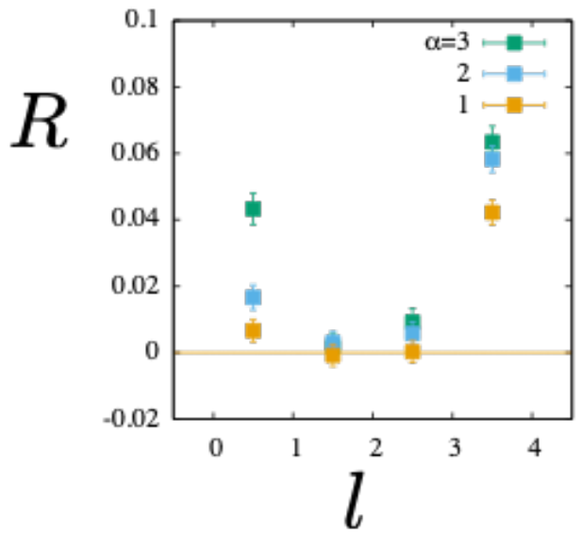
binary perceptron

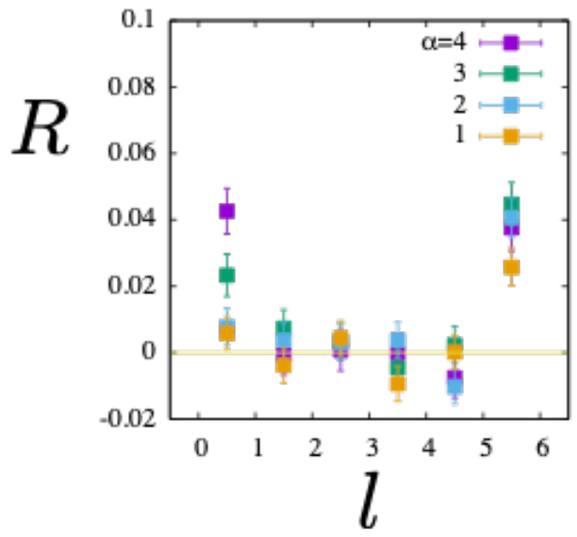
training

test

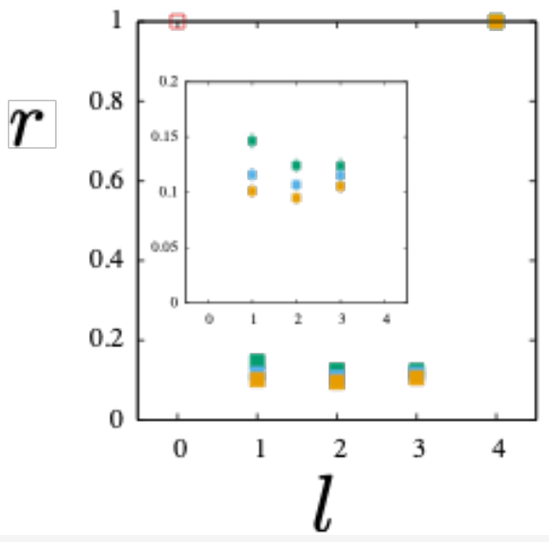
Overlaps between
teacher and student

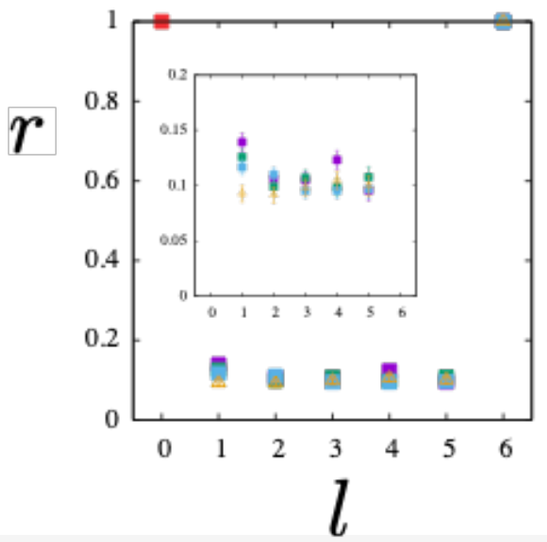
$a)$ bond



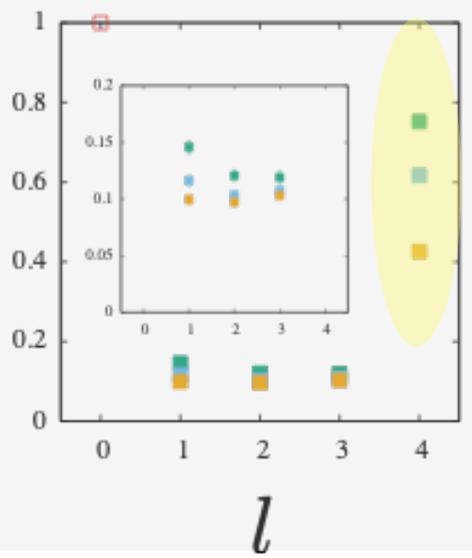


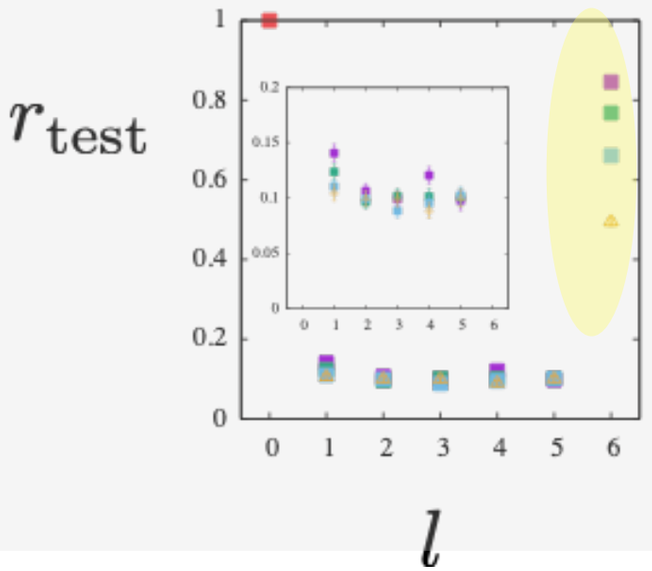
$b)$ spin





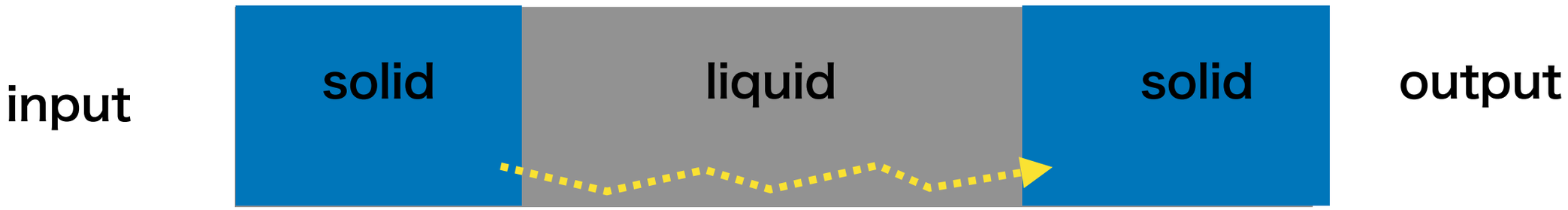
$c)$ spin





Summary

The solution space of over-parametrized DNN
exhibit “solid-liquid-solid” structure



1. Learning : “Liquid phase” helps equilibration

2. Generalization: “crystal phase” enables generalization.
Remanent bias field in the liquid phase plays the role of
symmetry breaking field by which “hidden” crystal
is selected out of may glass solutions.

3. Space evolution of the hierarchal free-energy landscape:
DNN naturally has the power of renormalization: classification,
feature detection

Outlook

Numerical simulations to test theoretical predctions

Various statistical inference problems

Complex systems with heterogeneity

- gene regulatory network,⋯
  functionality vs robustness in biology
- allostericity

ultra-stable glass, rheology

Theories in $d = 1 + \infty$

Like “Landau to Ginzburg-Landau” but more microscopic