IPISeminar Oct 28, 2021

Design space of a deep neural network - its spatial evolution and robustness

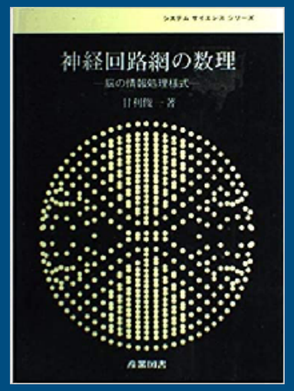
Hajime Yoshino

Cybermedia Center,Osaka University

科研費 KAKENHI

Hajime Yoshino, SciPostPhys. Core 2, 005 (2020).

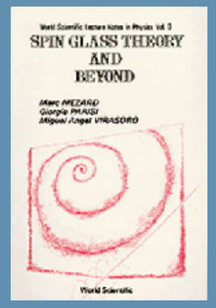
「最近の研究から - 深層ニューラルネットワークの解剖ー統計力学によるアプローチ」日本物理学会誌76巻9号(2021年9月号)



1978

"Deep Learning"

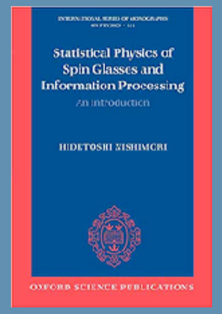
statistical mechanics of disordered systems



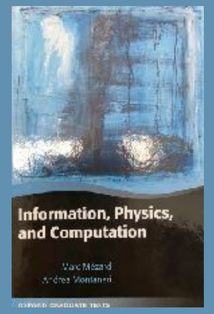
1987

1991

1999

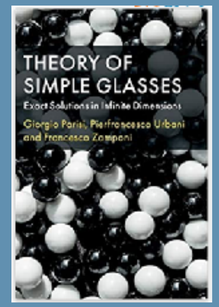


2001



2008

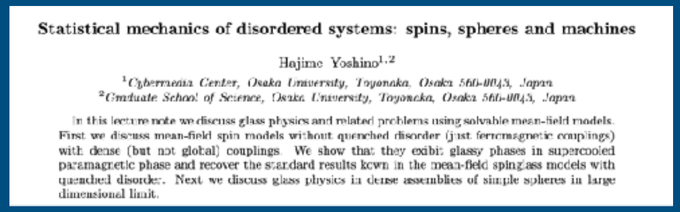
without quenched disorder



2020

spin glass theory and FAR beyond

in progress



Statistical mechanics of disordered systems: spins, spheres and machines

Hajime Yoshino[1,2]
[1]*Cybermedia Center, Osaka University, Toyonaka, Osaka 560-0043, Japan*
[2]*Graduate School of Science, Osaka University, Toyonaka, Osaka 560-0043, Japan*

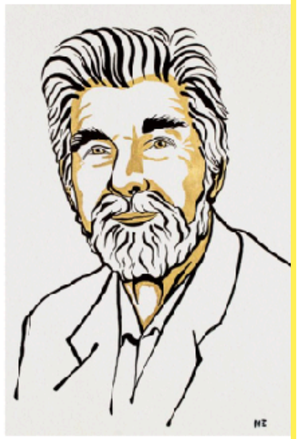
In this lecture note we discuss glass physics and related problems using solvable mean-field models. First we discuss mean-field spin models without quenched disorder (just ferromagnetic couplings) with dense (but not global) couplings. We show that they exibit glassy phases in supercooled paramagnetic phase and recover the standard results kown in the mean-field spinglass models with quenched disorder. Next we discuss glass physics in dense assemblies of simple spheres in large dimensional limit.

The Nobel Prize in Physics 2021

Ill. Niklas Elmehed © Nobel Prize Outreach

Syukuro Manabe

Prize share: 1/4



Ill. Niklas Elmehed © Nobel Prize Outreach

Klaus Hasselmann

Prize share: 1/4

Ill. Niklas Elmehed © Nobel Prize Outreach

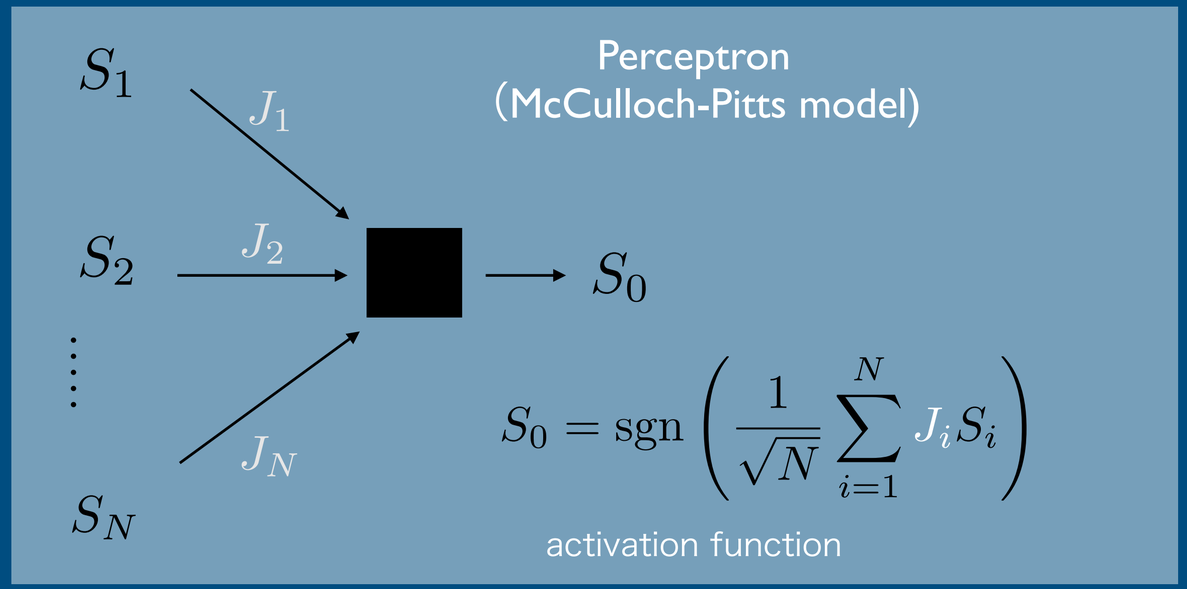
Giorgio Parisi

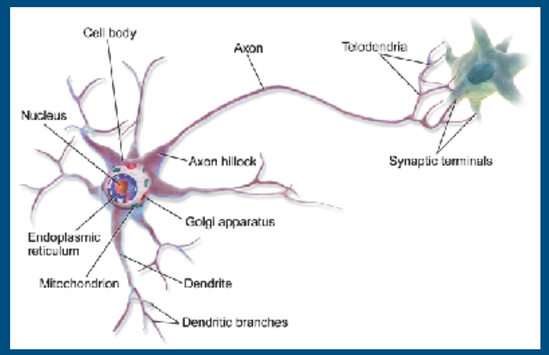
Prize share: 1/2

https://www.nobelprize.org/uploads/2021/10/sciback_fy_en_21.pdf

"for groundbreaking contributions to our understanding of complex physical systems"

Perceptron





Example

# of possible machines

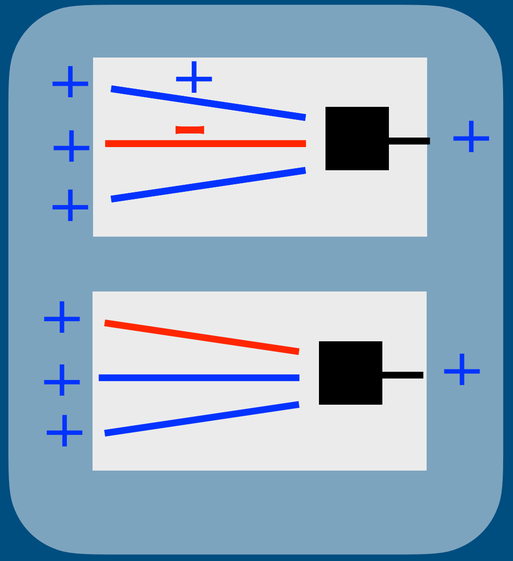
$2^3 = 8$

machine1

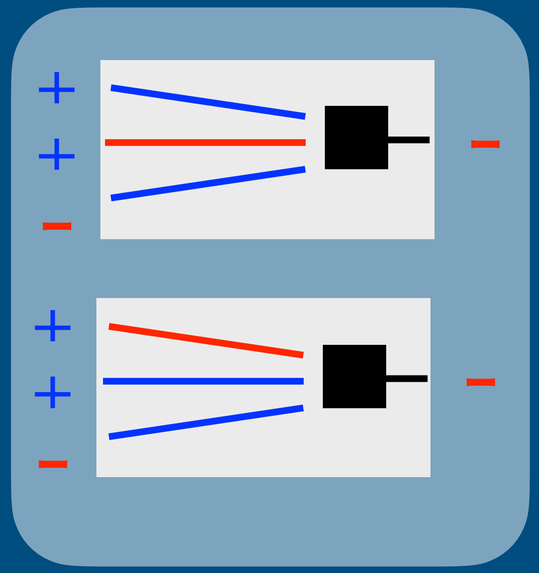
machine2

⋮

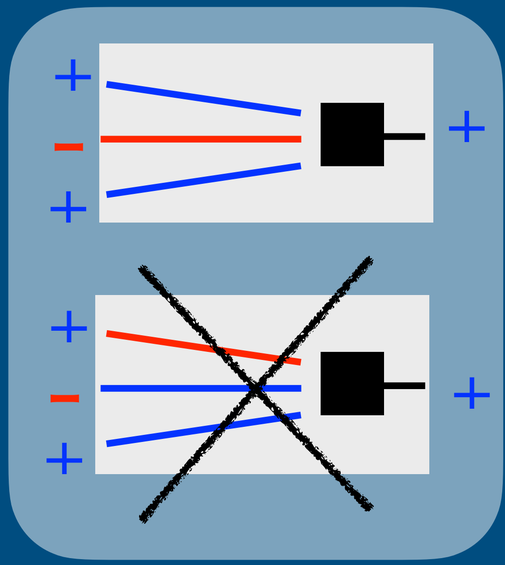
DATA1



DATA2

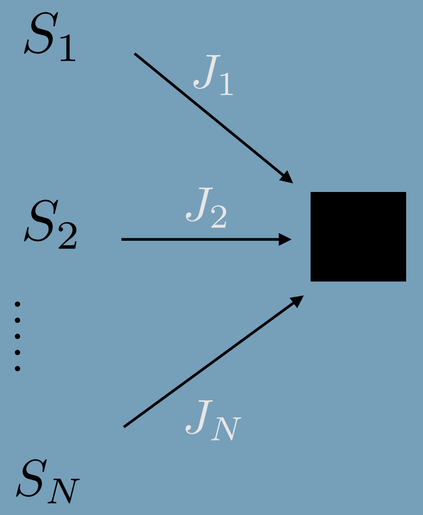


DATA3



……

Perceptron
(McCulloch-Pitts model)



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

activation function

Elisabeth Gardner
(1957-1988)

Statistical mechanics of J_i which meet random constraints

$$S_i^\mu = \pm 1 \qquad \text{pattern} \quad \mu = 1, 2, \ldots, M$$

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

$M \to \infty$ with fixed α

Gardner volume

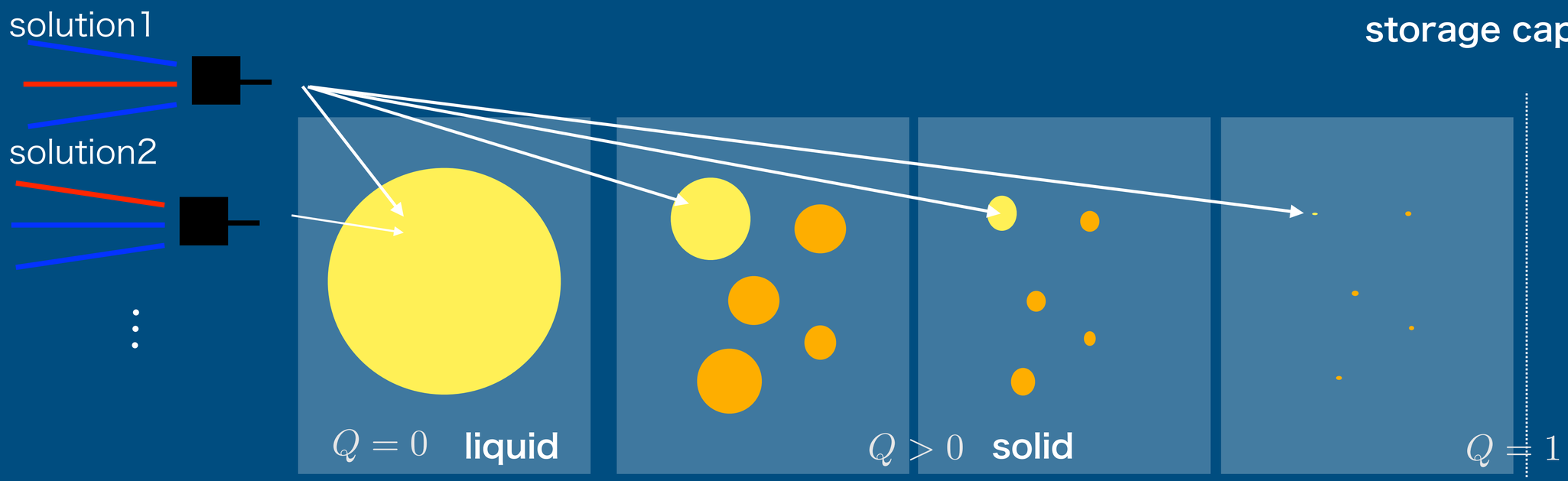
$$V = \int \prod_{j=1}^{N} \frac{dJ_j}{\sqrt{2\pi}} e^{-\frac{{J_j}^2}{2}} \prod_{\mu=1}^{M} e^{-\beta v(r^\mu)}$$

"Hardcore" constraint

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

"Gap" $\quad r^\mu = S_0^\mu \sum_{i=1}^{N} \frac{1}{\sqrt{N}} J_i S_i^\mu - \kappa$

Gardner's volume: design space of a perceptron



Franz-Parisi (2016)

$\kappa < 0$

clustering
glass

jamming
(SAT/UNSAT tradition)

Replica Symmetry Breaking (RSB)

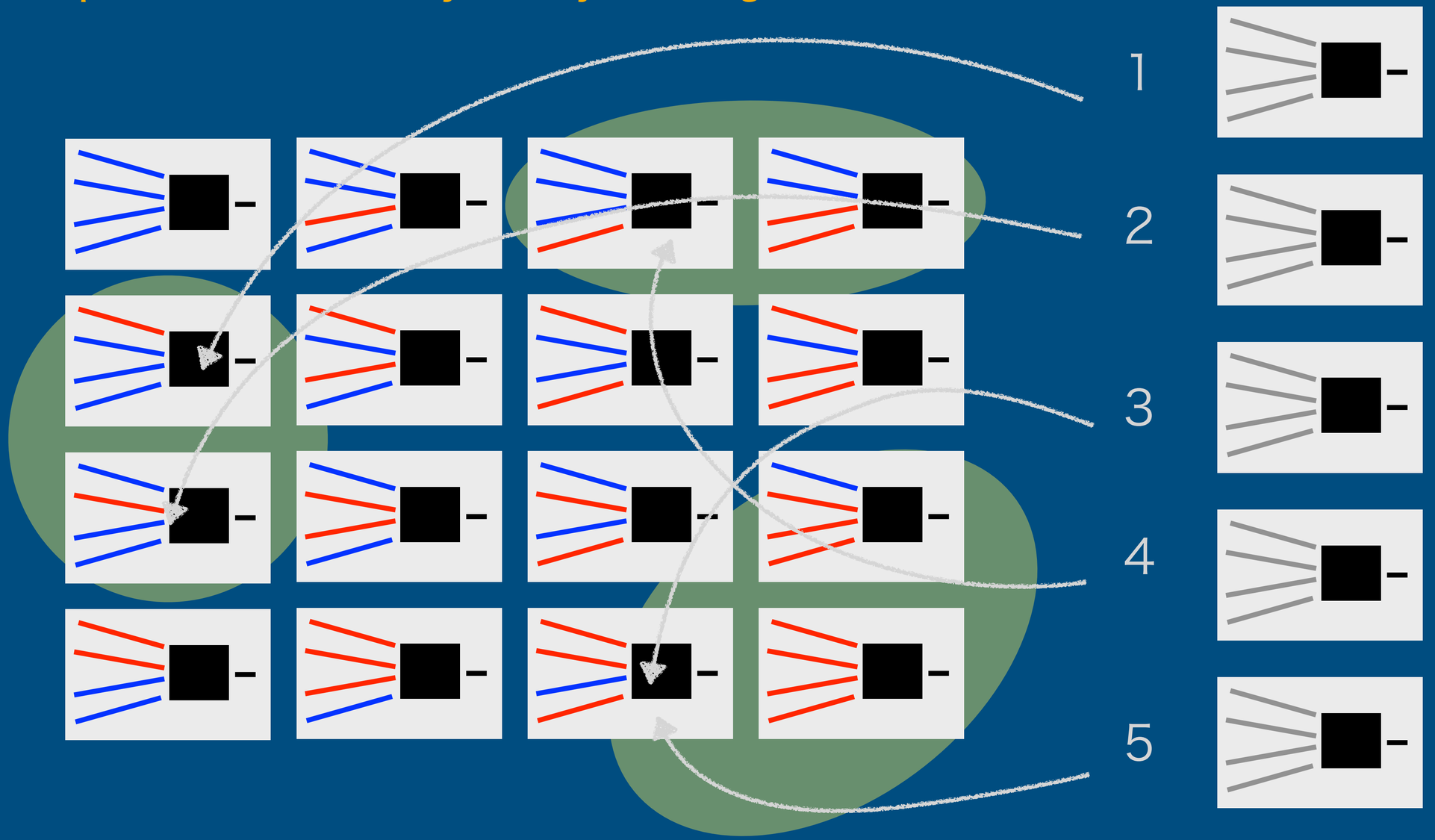
$$\alpha = \frac{M}{N}$$

overlap between machine a and b
learning with the same data

$$Q_{ab} = \frac{1}{N}\sum_{i=1}^{N} J_i^a J_i^b$$

Replica Symmetry Breaking =
Replica “Permulation” Symmetry Breaking

replica=machines learning in parallel
with the same training data

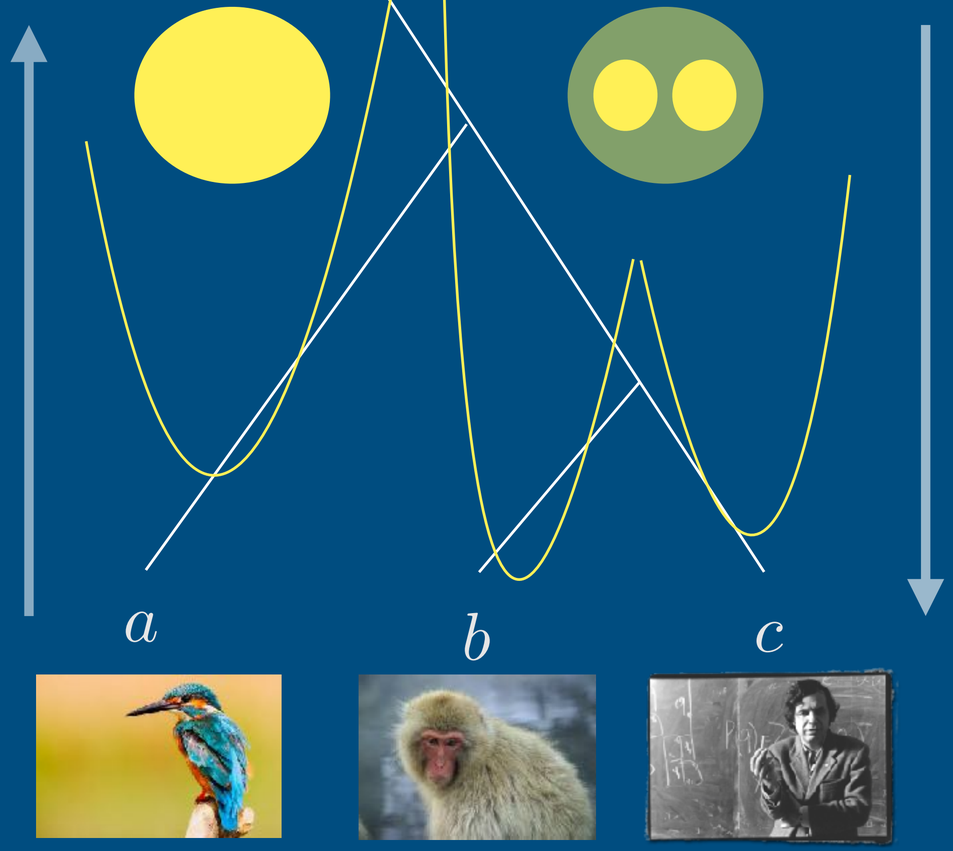


Replica (permutation) symmetry breaking and ultrametricity

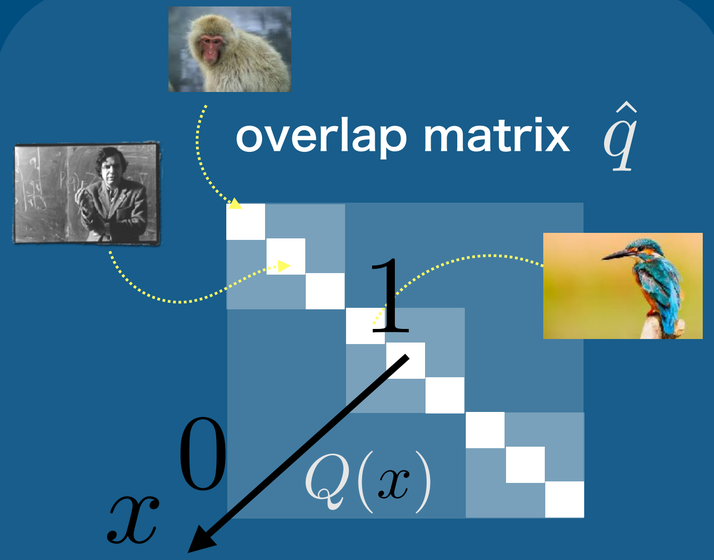
complex
free-energy landscape

ultrametricity

distance



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



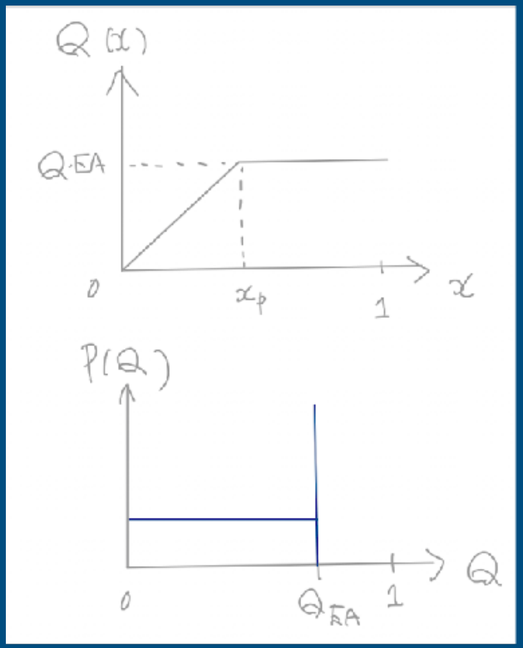
Replica symmetry breaking
G. Parisi (1979)
first found in the SK model
(mean-field spin glass)

Similarity Q

$$Q_{ab} = \frac{1}{N}\sum_{i=1}^{N} J_i^a J_i^b$$

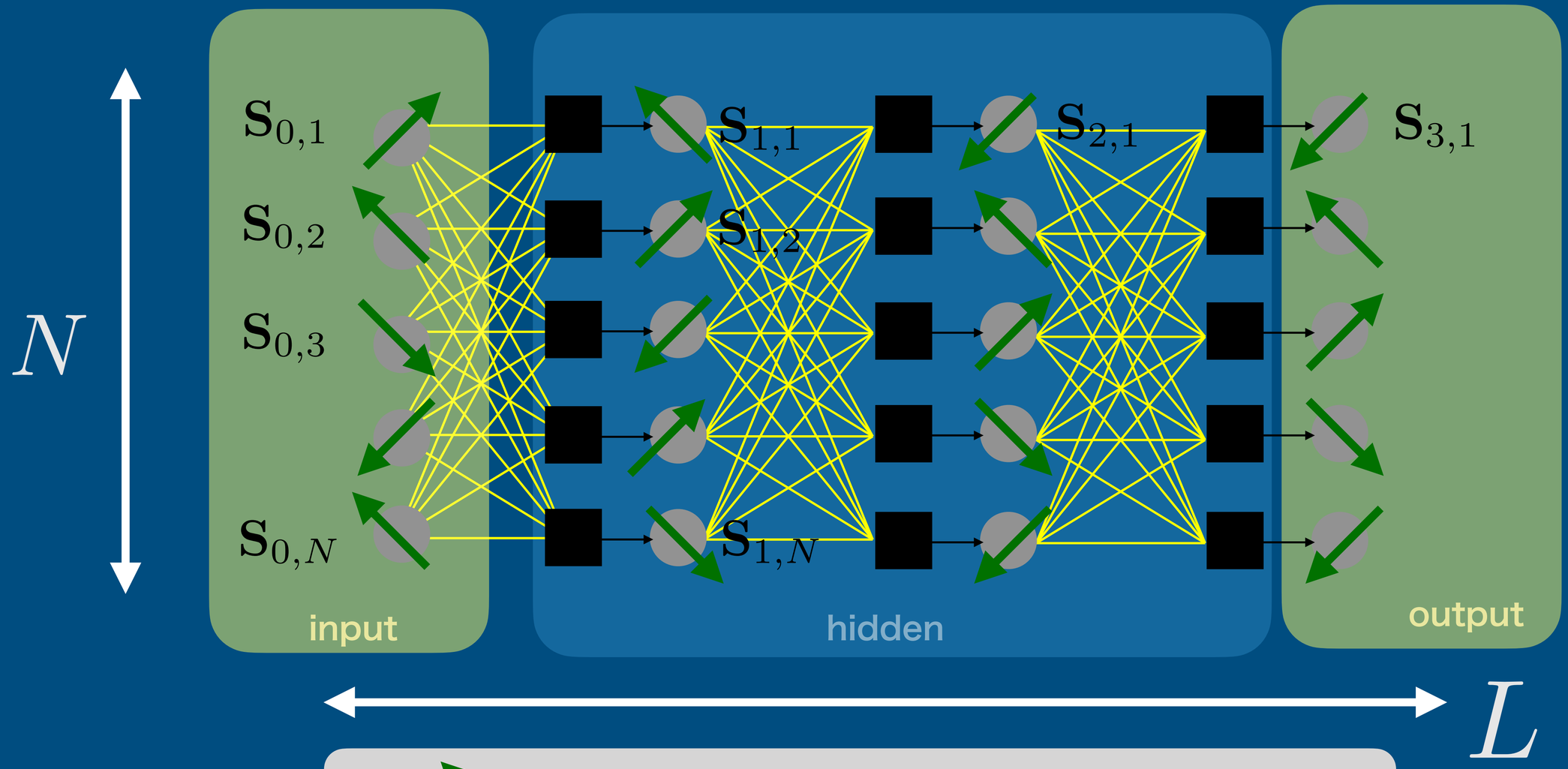
Overlap distribution

$$P(Q) = \frac{dx}{dQ}$$



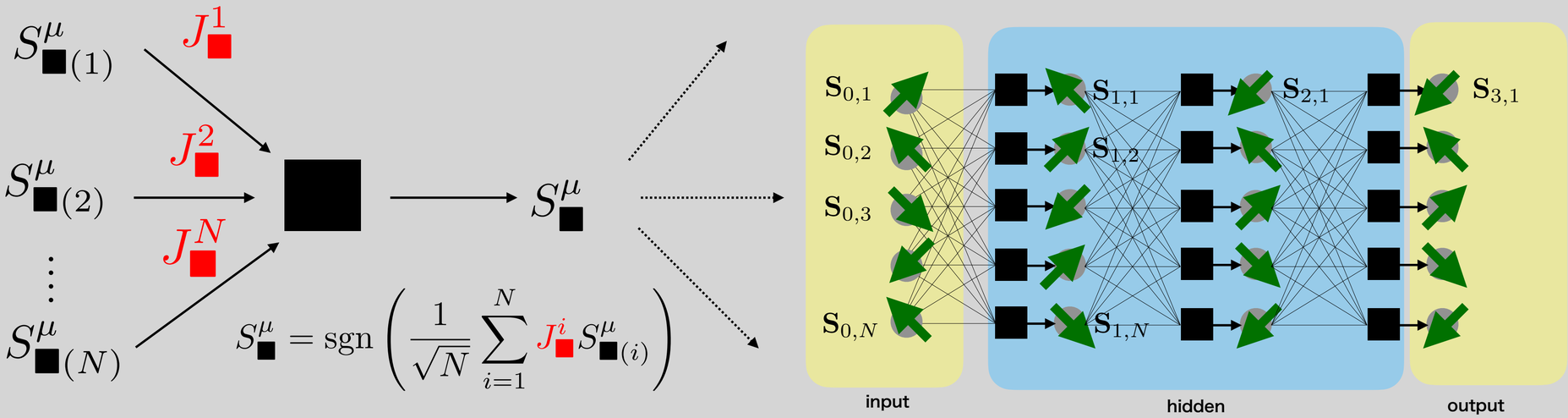
Multi-layer Neural Network

Design weights to satisfy boundary conditions



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

$$\alpha = M/N$$



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

Usual strategy of learning

(1) define “loss function"

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

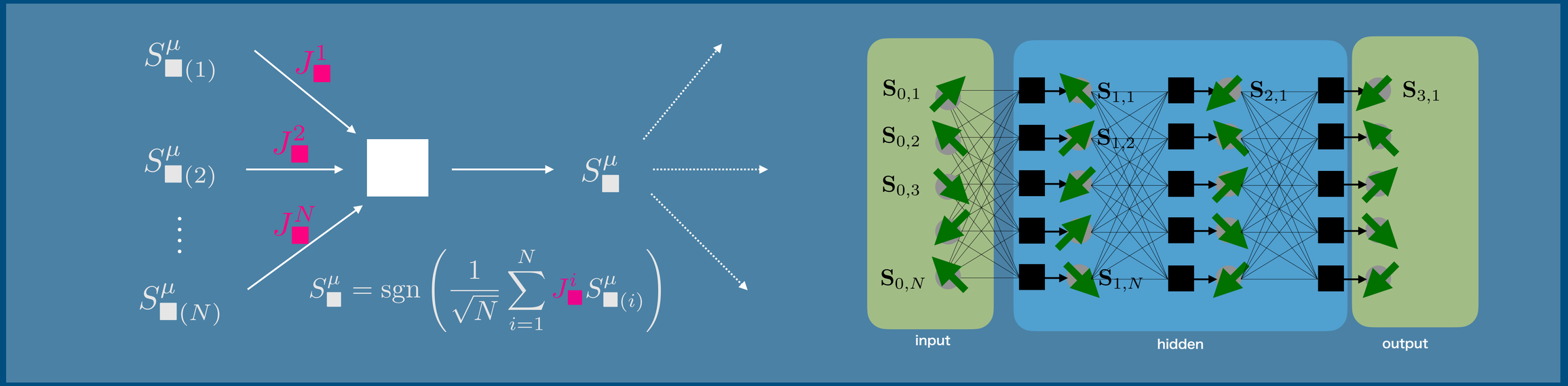
desired output

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 in deep perceptron network



Gardner volume generalized for a multi-layer network

(c.f..) Single percepron: E. Gardner (1987)

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

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

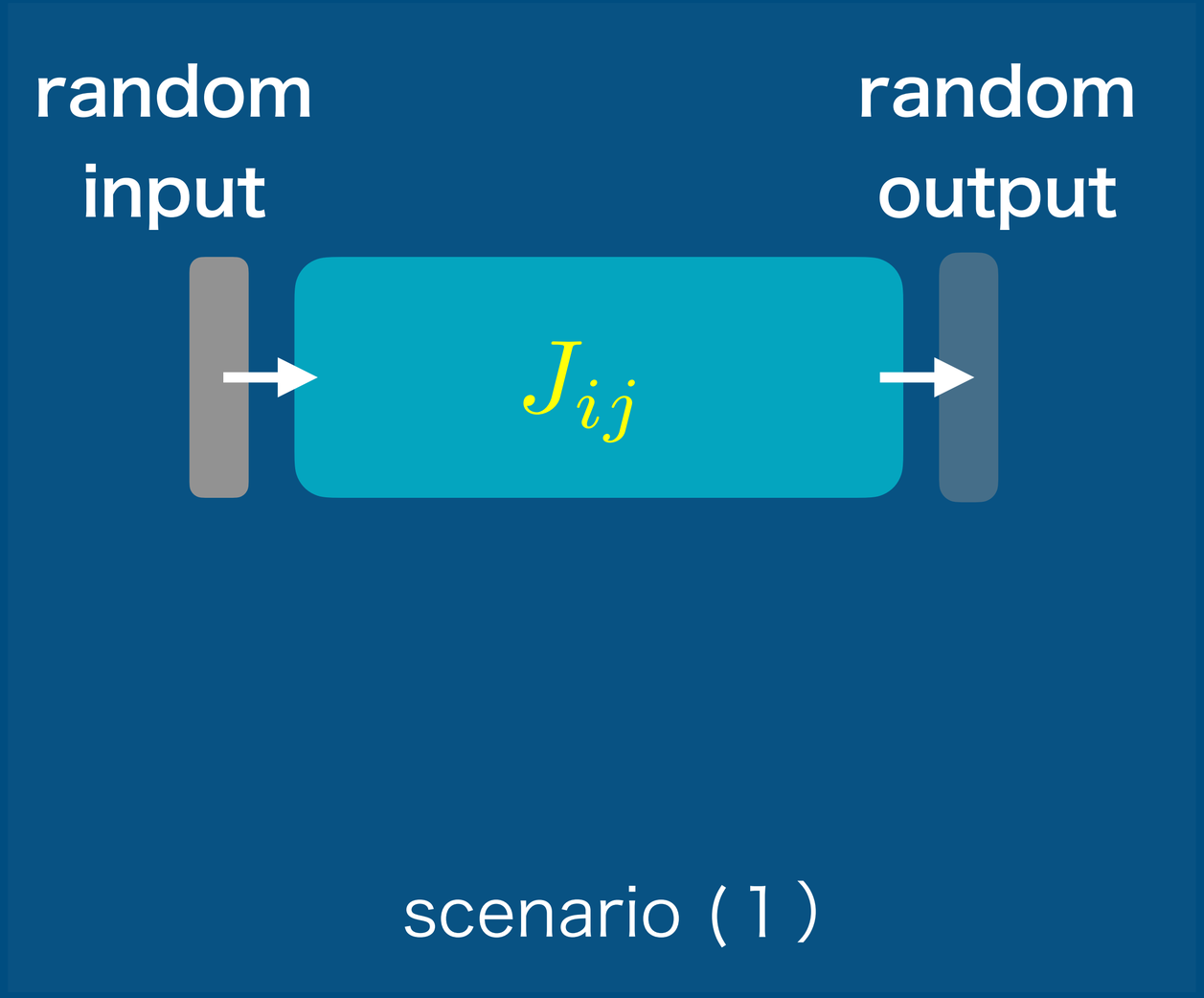
"gap"

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

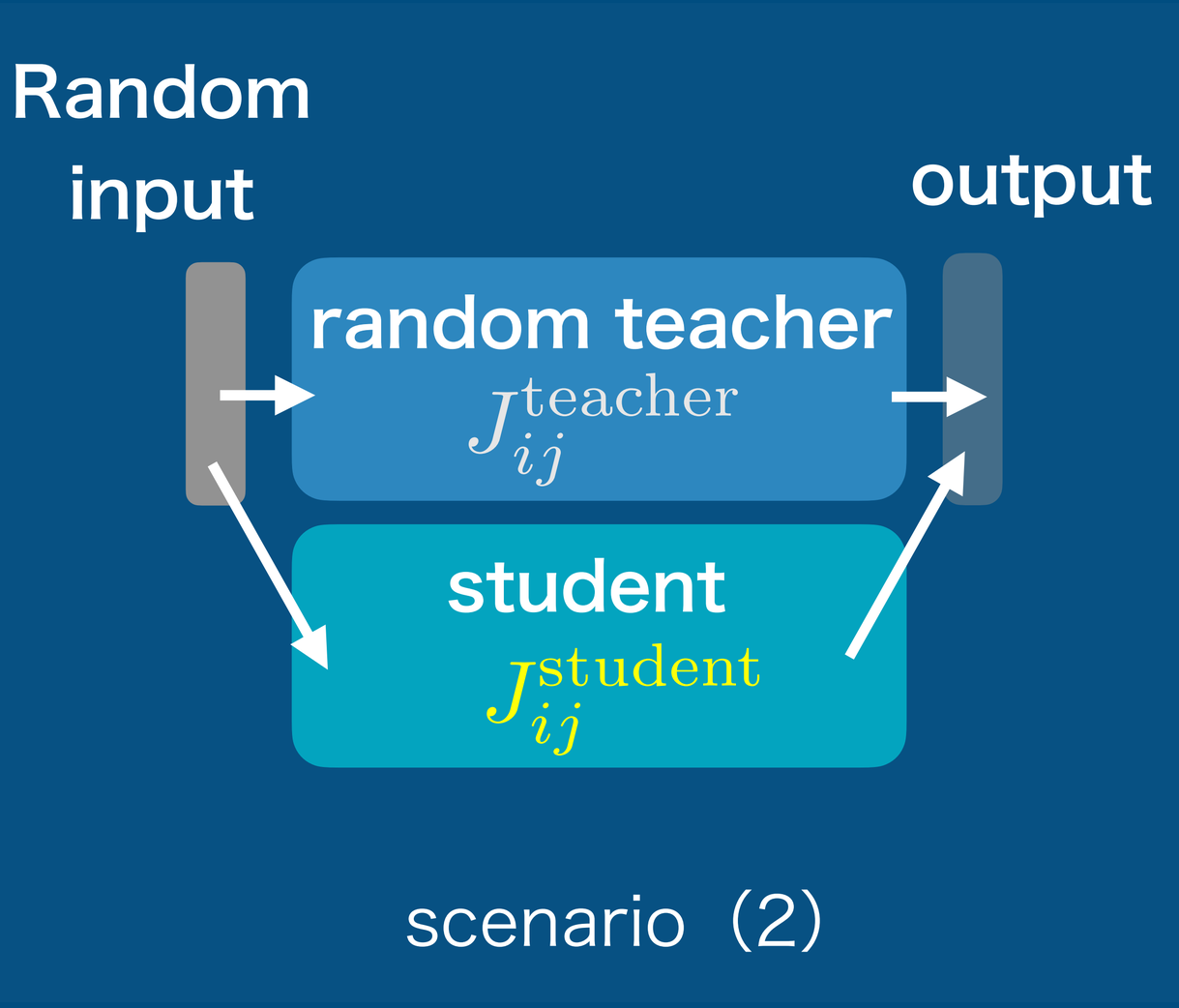
"Hardcore" constraint

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

We study two scenarios of machine learning
with artificially generated training data



scenario (1)

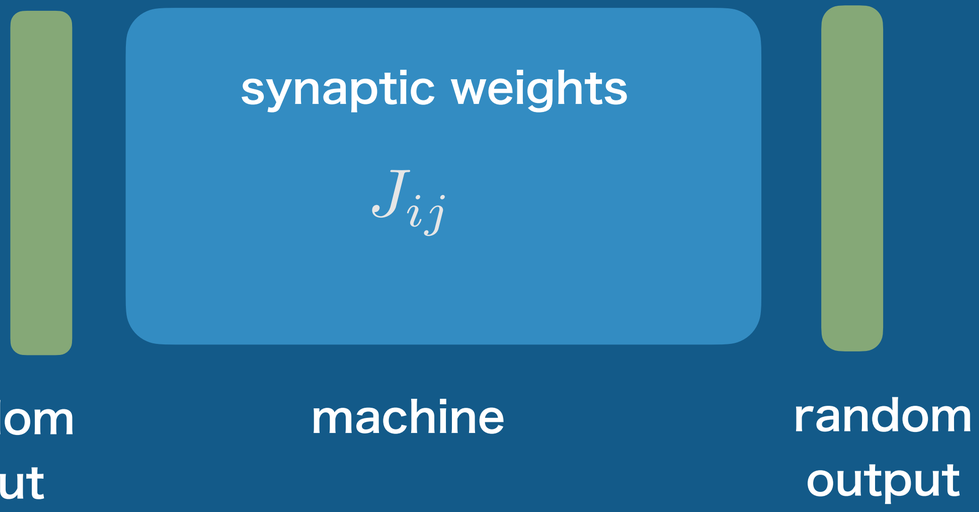


scenario (2)

Scenario (1) Random inputs/random outputs

- a constraint satisfaction problem

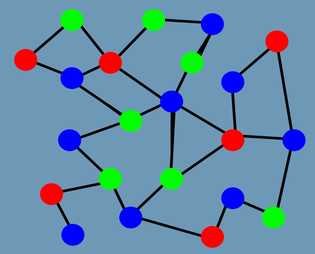
- glass/jamming physics



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

"Random Constraint Satisfaction Problem"
(ランダム制約充足問題)

Example:
Graph coloring



Antiferromagnetic Potts model

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

Connection to glass physics

Clustering transition = glass transition
SAT/UNSAT transition = jamming transition

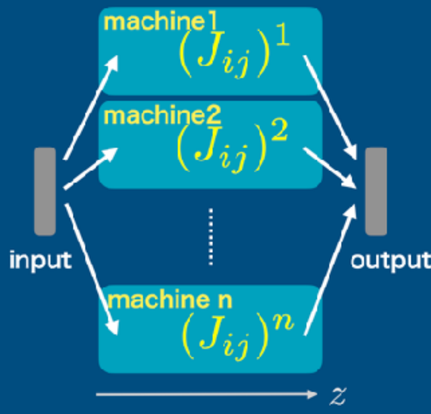
Replicated Gardner volume

replicas: machines learning in parallel with the same data

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

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

replicated machines $a = 1, 2, \ldots, n$



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

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

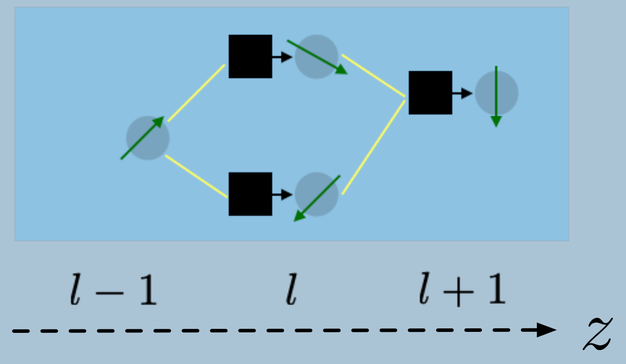
$$S_n[\{\hat{q}(l)\},\{\hat{Q}(l)\}] = \alpha^{-1}\sum_{l=1}^{L}S^{\mathrm{bond}}_{\mathrm{ent}}[\hat{Q}(l)] + \sum_{l=1}^{L-1}S^{\mathrm{spin}}_{\mathrm{ent}}[\hat{q}(l)]$$

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

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

Hajime Yoshino, SciPostPhys. Core 2, 005 (2020).

Loop correction

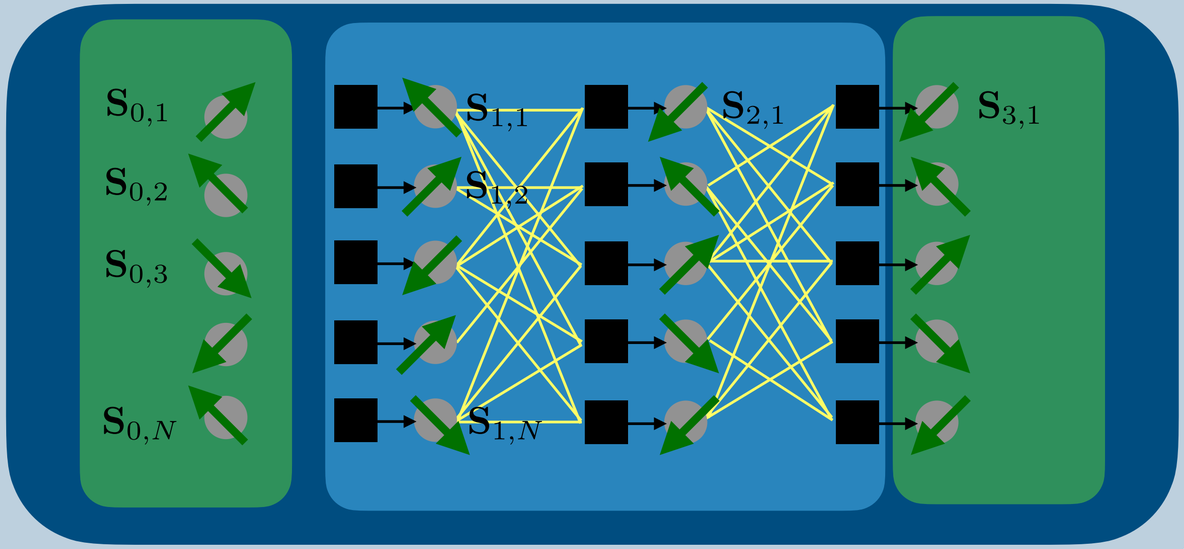


Not global but **dense** coupling - loop corrections become negligible

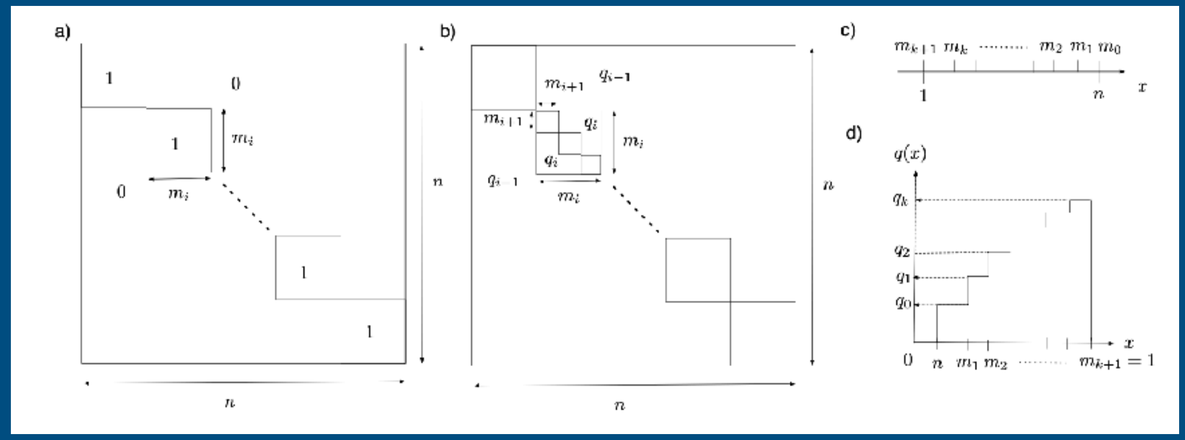
"connectivity" c

$$N \gg c \gg 1$$

$$\alpha = M/c$$

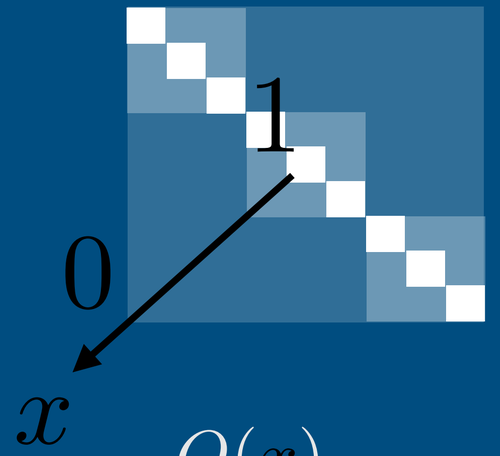


Parisi's RSB ansatz



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

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



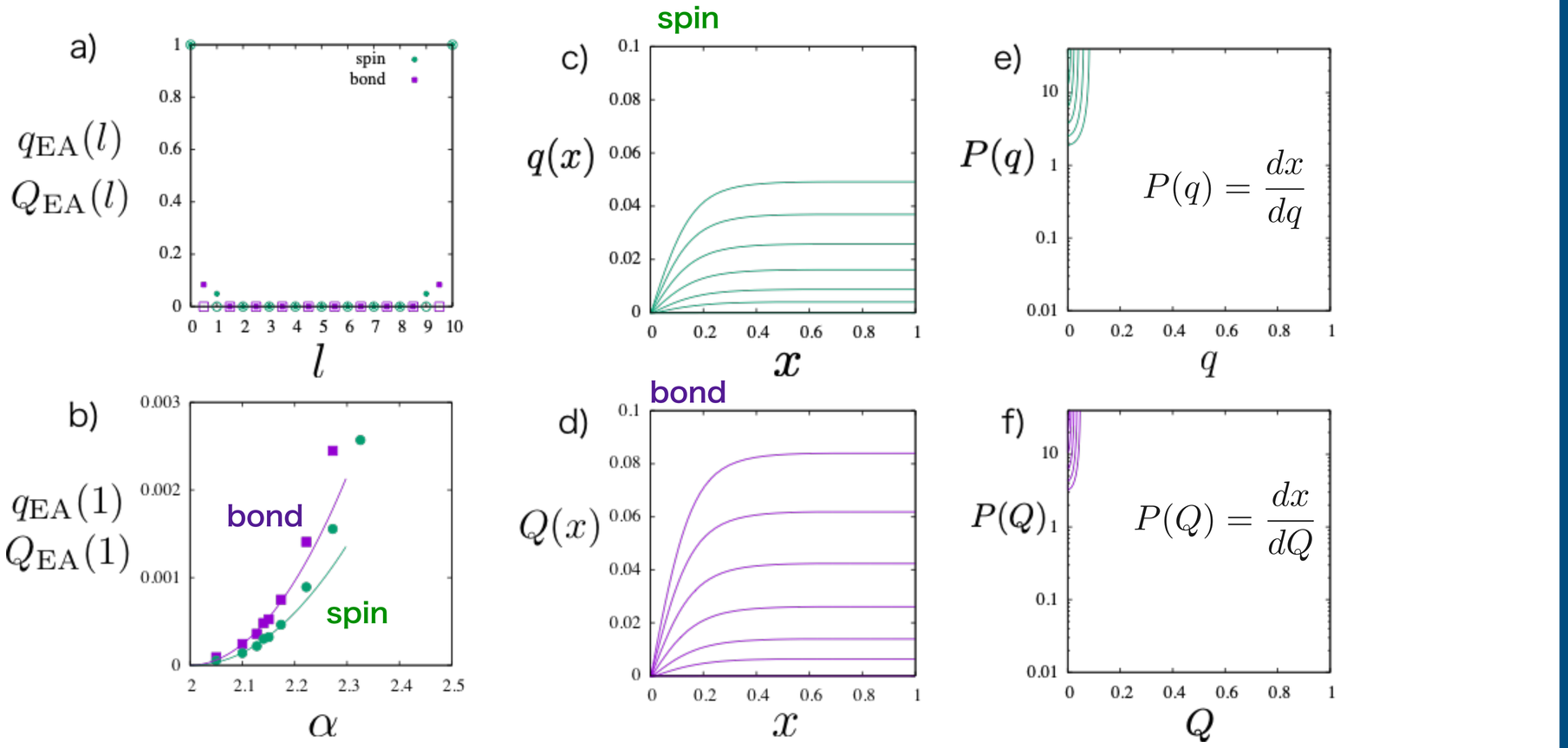
$q(x)$

Input/output boundaries

replicated machines are subjected to the same training data

$$q_{ab}(0) = q_{ab}(L) = 1$$

■1st Glass transition



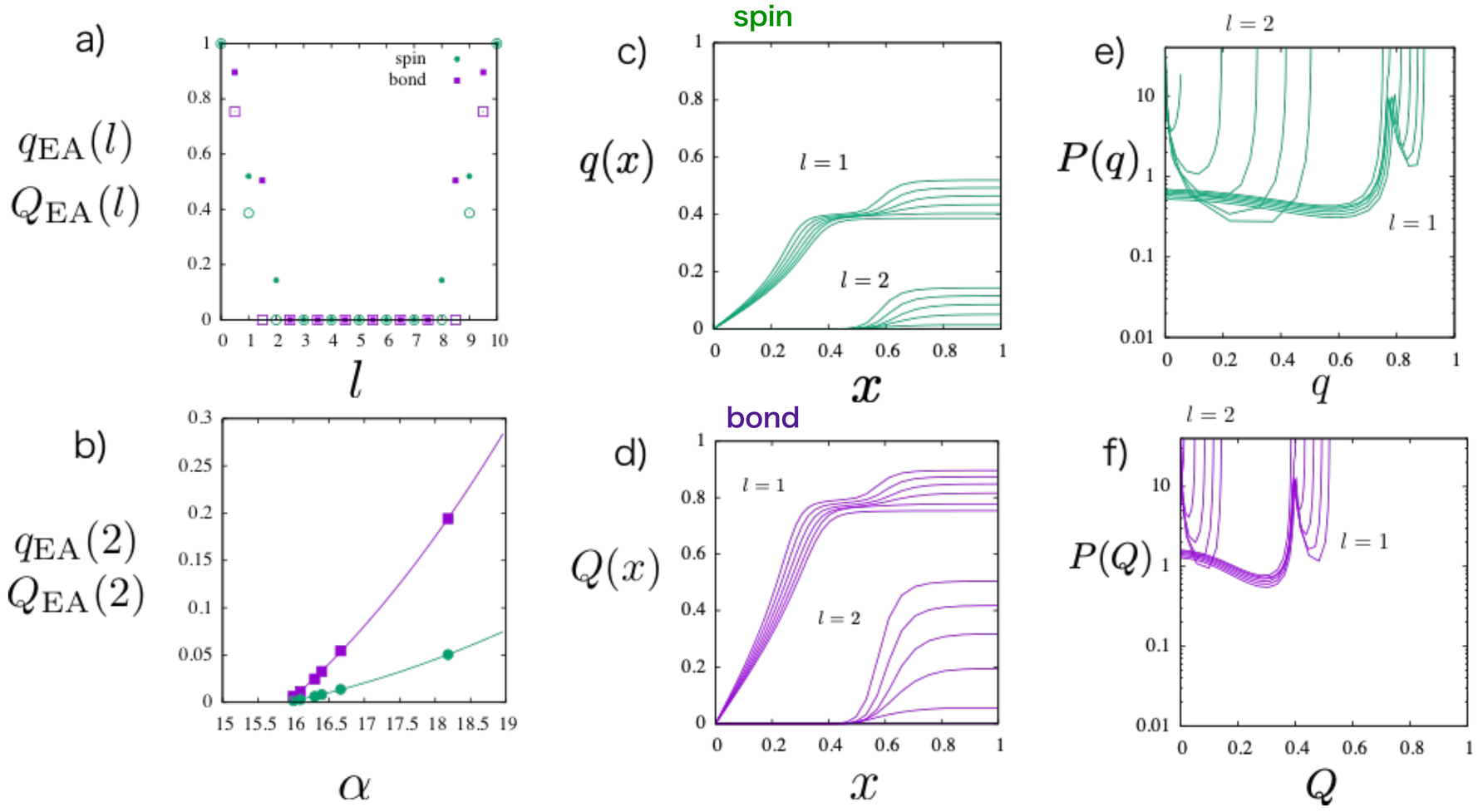
$\alpha_{\rm g} \simeq 2.03$

continuous transition to full RSB glass phase

at 1 st & (L-1) th layer

other layers remain in the liquid phase

■2nd Glass transition



$\alpha_{\rm g}(2) \simeq 15.38$

continuous transition to full RSB glass phase

at 2nd & (L-2) th layer

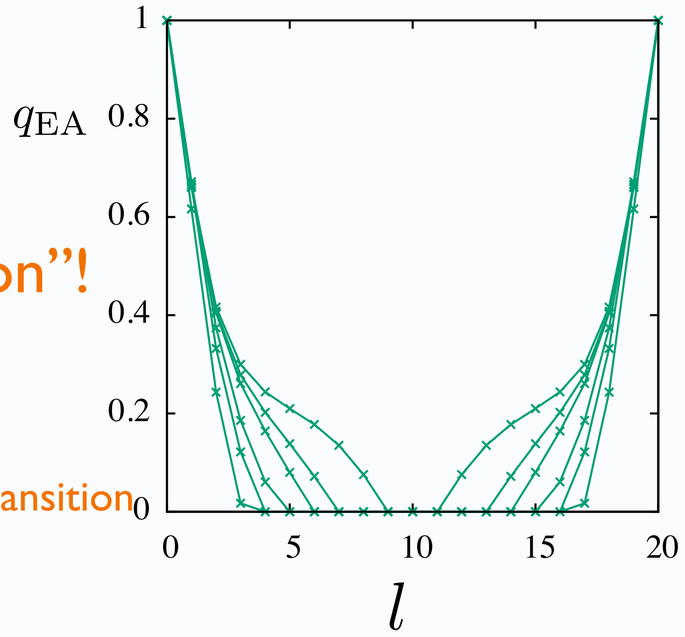
which also induce 2nd glass transitions at 1st and L-th layer

Growth of glass phase with increasing training data

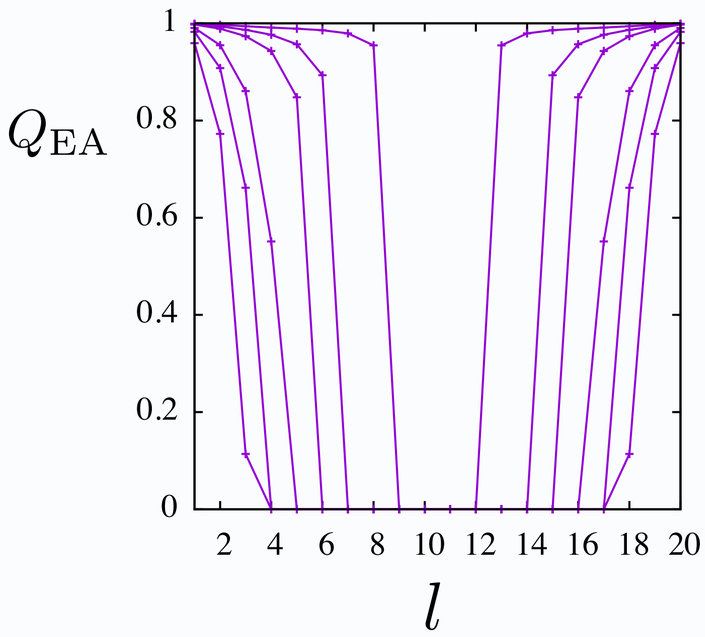
"Wetting transition"!

layer-by-layer 2nd order transition

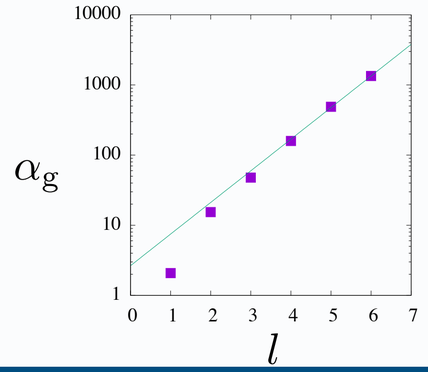
spin



bond



$$\alpha = 50, 100, 200, 1000, 2000, 4000$$



$$\alpha_{\rm g}(l) \sim 2.7(3) e^{1.03(2) l}$$

"penetration depth"

$$\xi(\alpha) \propto \ln \alpha$$

this suggests storage capacity also grows fast with the depth

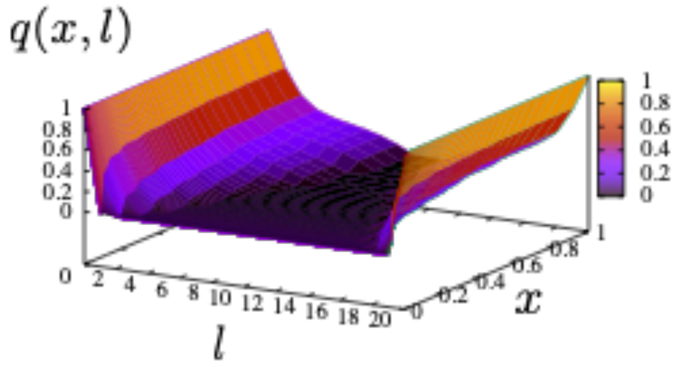
$$\alpha_{\rm J}(l) \propto e^{{\rm const} l}$$

Space-dependent replica-symmetry breaking

$$\alpha = 4000$$

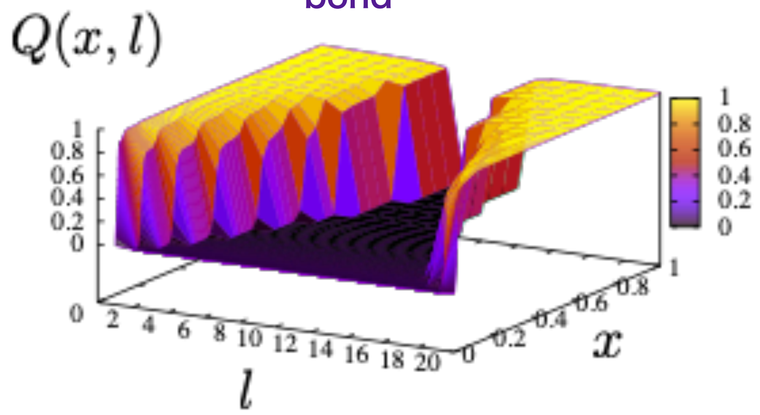
a)

spin

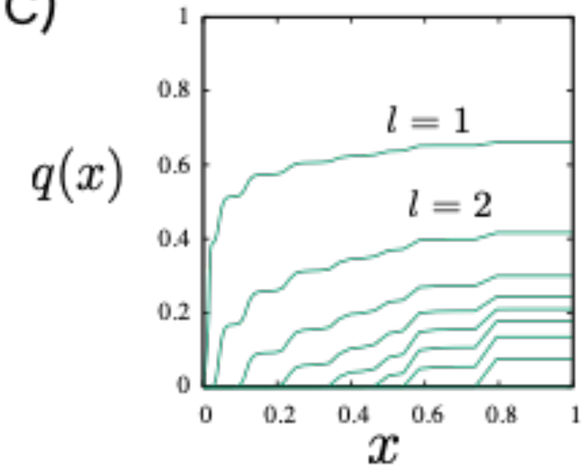


b)

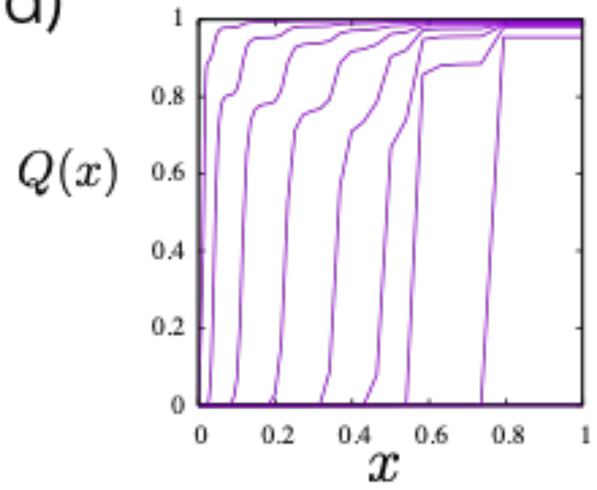
bond



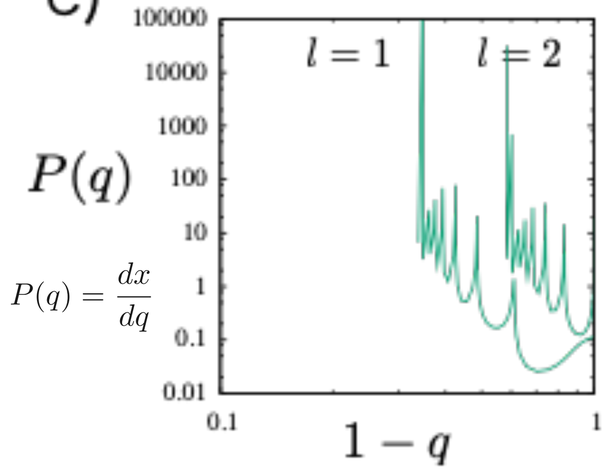
c)



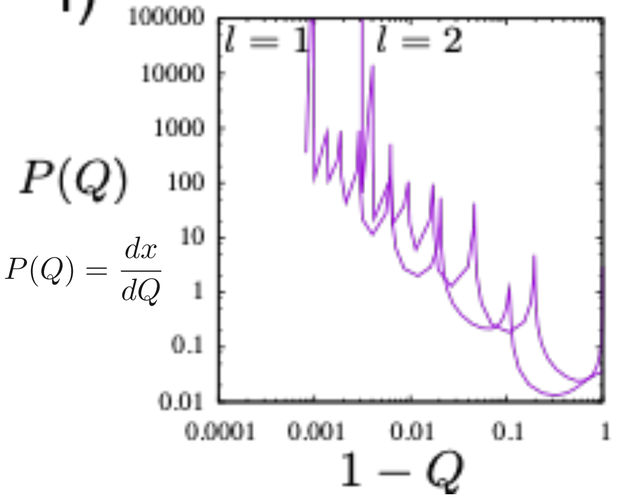
d)

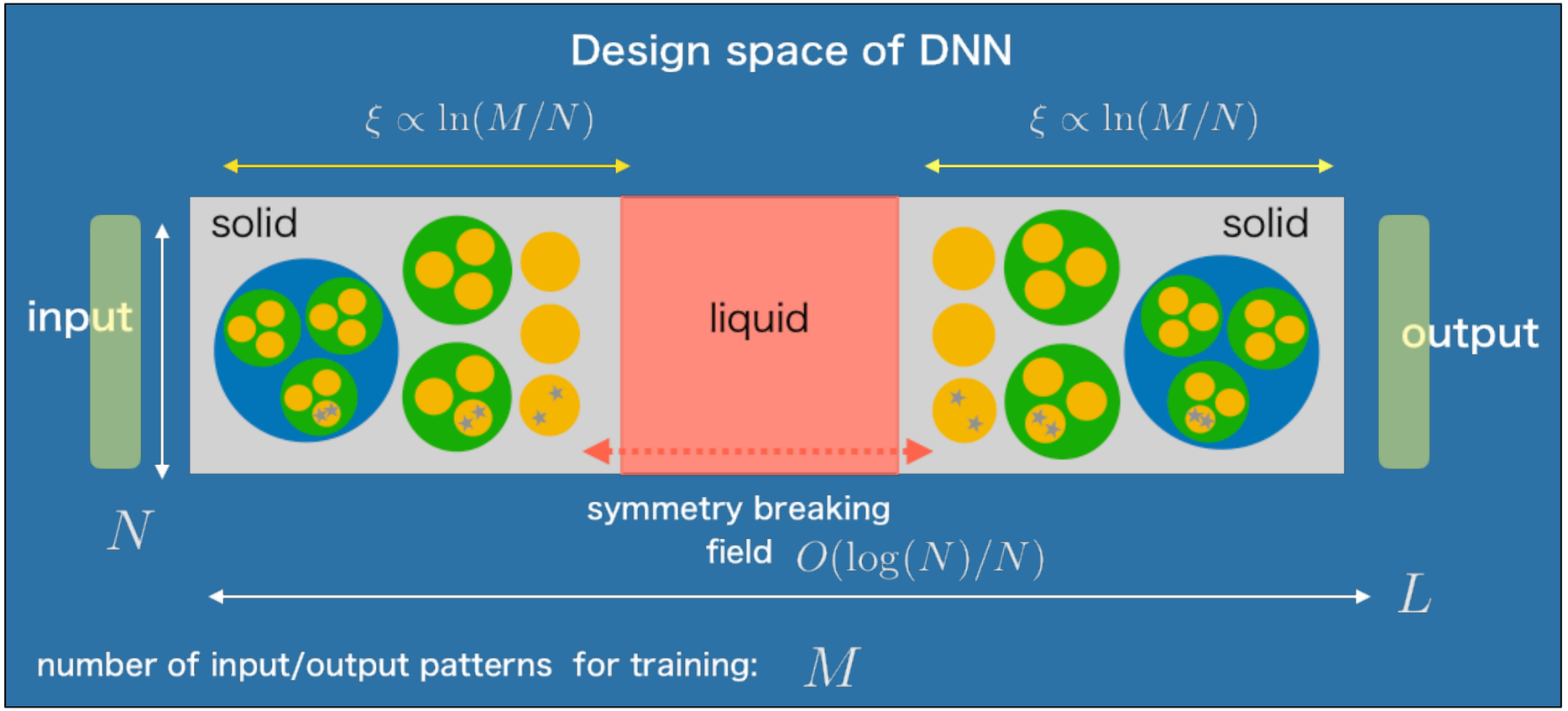


e)



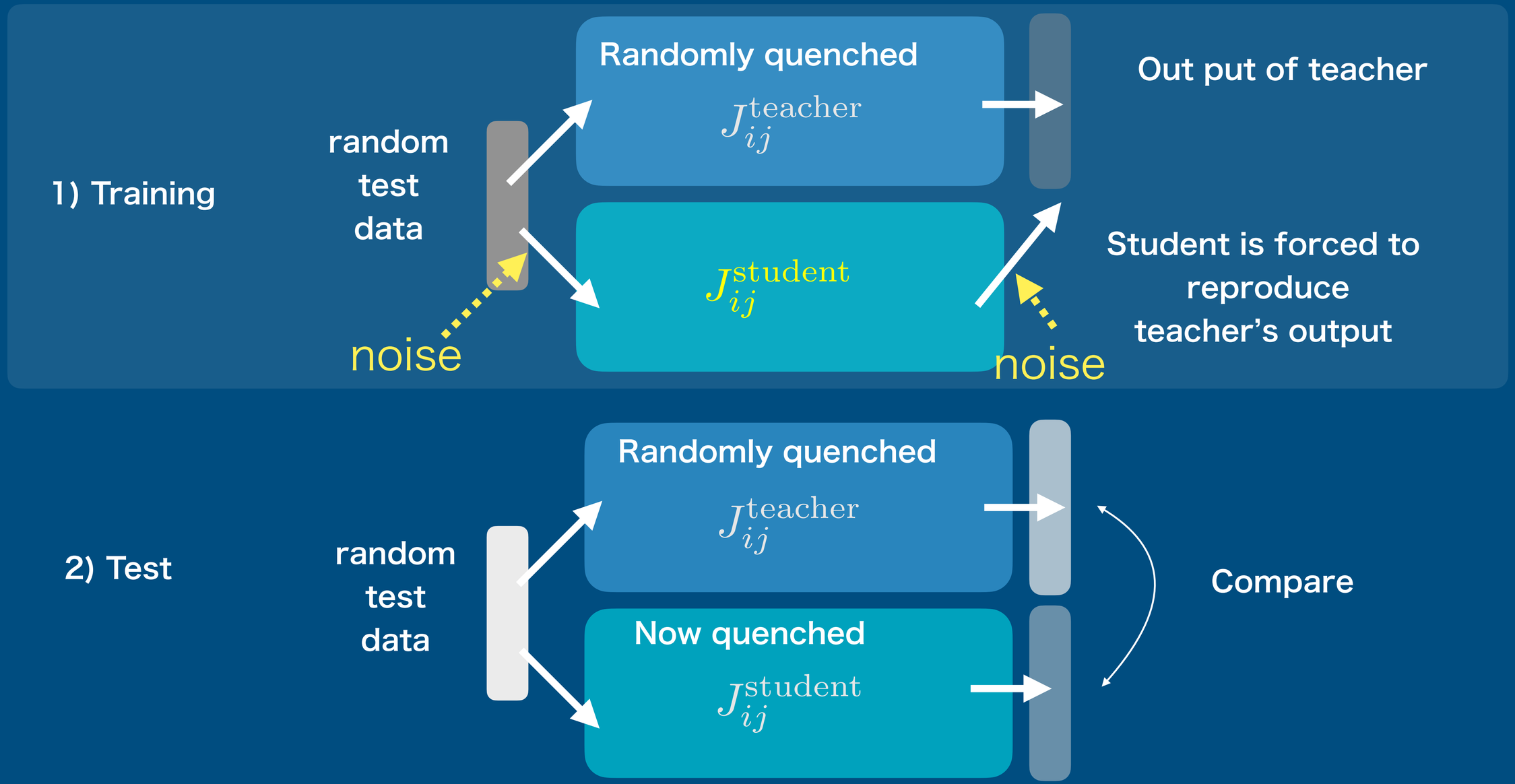
f)





each DNN is NOT a glass

Scenario (2) Teacher student setting - a statistical inference problem



Replicated Gardner volume

$$V^{1+n}(\mathbf{S}_0, \mathbf{S}_L) = \prod_{a=0}^{n} \left(\prod_{\blacksquare} \mathrm{Tr}_{\mathbf{J}^a_{\blacksquare}} \right) \left(\prod_{\blacksquare \setminus \mathrm{output}} \mathrm{Tr}_{\mathbf{S}^a_{\blacksquare}} \right) \prod_{\mu, \blacksquare, a} e^{-\beta v(r^{\mu}_{\blacksquare, a})} \qquad r^{\mu}_{\blacksquare, a} = S^{\mu}_{\blacksquare, a} \sum_{i=1}^{N} \frac{1}{\sqrt{N}} J^{i}_{\blacksquare, a} S^{\mu}_{\blacksquare(i), a}$$

teacher-machine $a = 0$ student-machines $a = 1, 2, \ldots, n$

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 (Franz-Parisi potential)

$$\frac{-\beta \overline{F(\mathbf{S}_0, \mathbf{S}_L)}^{\mathrm{visible}}}{NM} = \frac{\partial_n \overline{V^{1+n}(\mathbf{S}_0, \mathbf{S}_L)}^{\mathrm{visible}} \Big|_{n=0}}{NM} = \partial_n \, S_{1+n}[\{\hat{Q}(l), \hat{q}(l)\}] \Big|_{n=0}$$

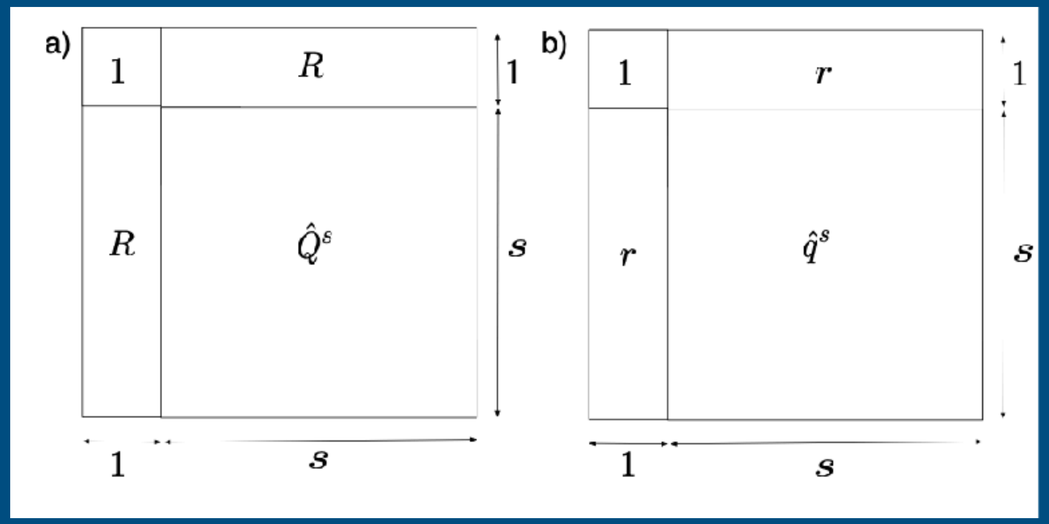
$$S_{1+n}[\{\hat{q}(l)\}, \{\hat{Q}(l)\}] = \alpha^{-1} \sum_{l=1}^{L} S^{\mathrm{bond}}_{\mathrm{ent}}[\hat{Q}(l)] + \sum_{l=1}^{L-1} S^{\mathrm{spin}}_{\mathrm{ent}}[\hat{q}(l)]$$

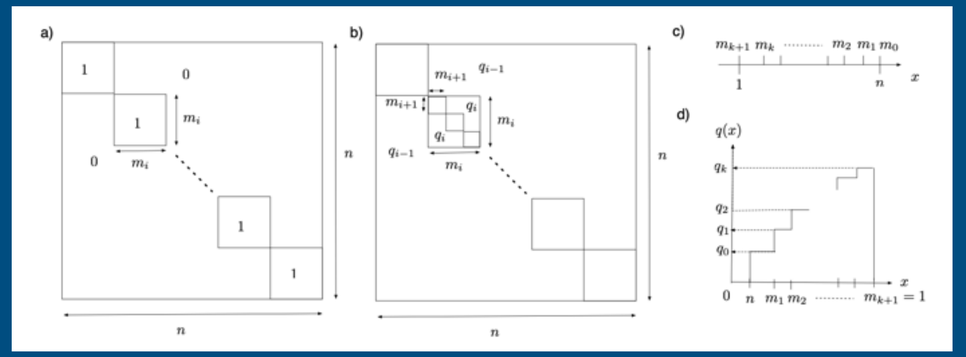
$$- \sum_{l=1}^{L} e^{\frac{1}{2} \sum_{ab} q_{ab}(l-1) Q_{ab}(l) q_{ab}(l) \partial_{h_a(l)} \partial_{h_b(l)}} \prod_{a=0}^{n} e^{-\beta v(h_a(l))} \Big|_{h_a(l)=0}$$

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

Hajime Yoshino, SciPostPhys. Core 2, 005 (2020).

Parisi's RSB ansatz





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

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

Input/output boundaries

overlap among students at the boundaries

$$q_{ab}(0) = q_{ab}(L) = 1$$

overlap between the students and the teacher

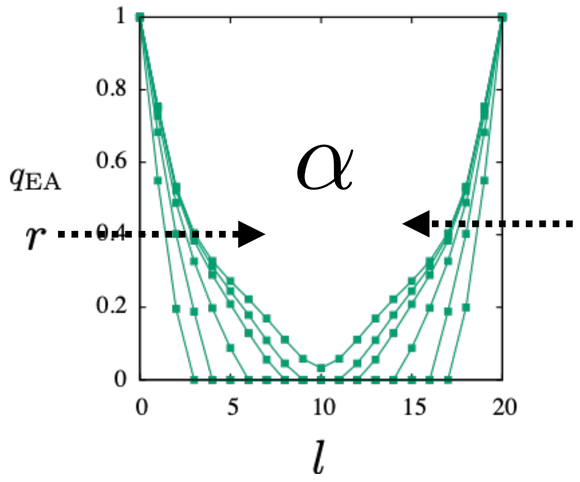
$$r(0) = r(1) = r$$

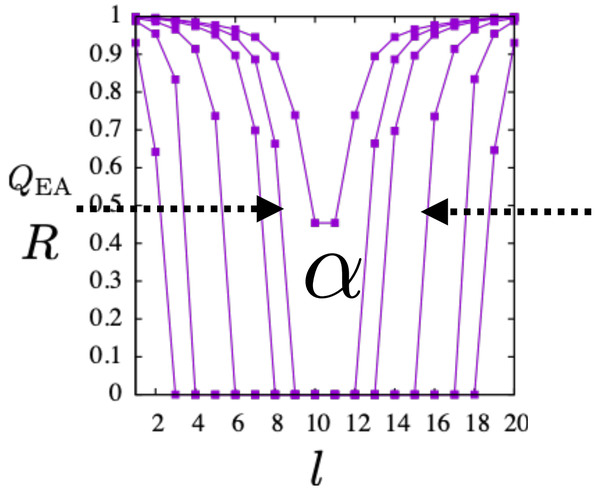
Bayes-optimal limit

Bayes-optimal case (no noise)

Hajime Yoshino, SciPostPhys. Core 2, 005 (2020).

$\alpha = 25, 100, 250, 500, 625, 714.$





"Wetting transition"

layer-by-layer 2nd order transition

Bayes optimal, Nishimori condition $q = r, Q = R$

Replica symmetry (RS) holds

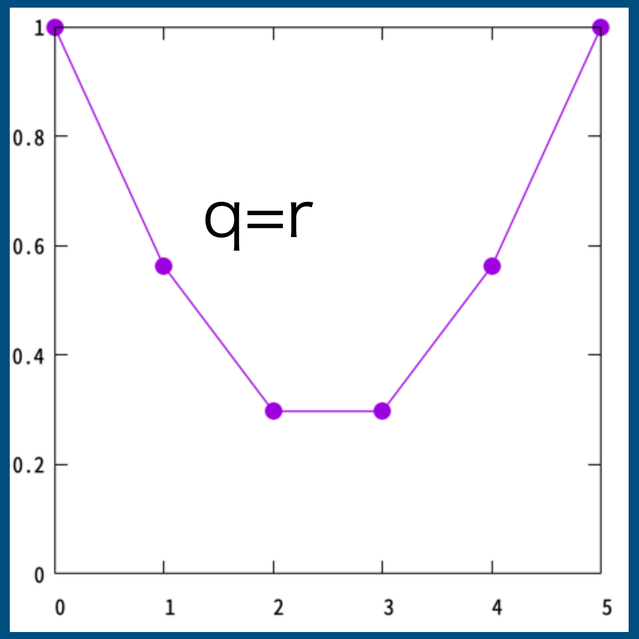
With noise

Spatial profile of the EA order parameter

noise

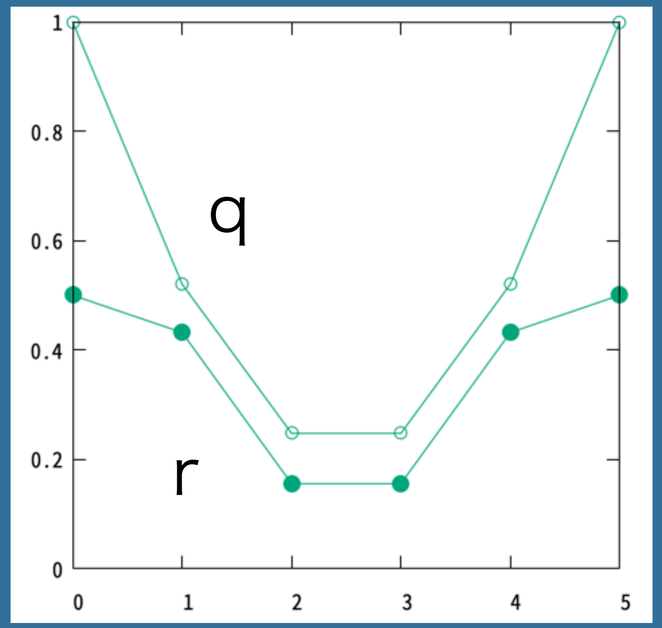
$q_{\mathrm{EA}} = q(1)$

$r = 1$



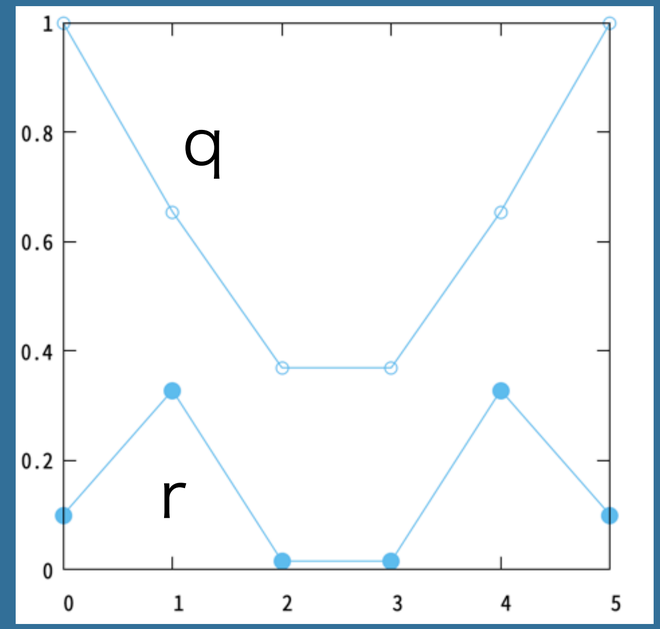
l

$r = 0.5$



l

$r = 0.1$



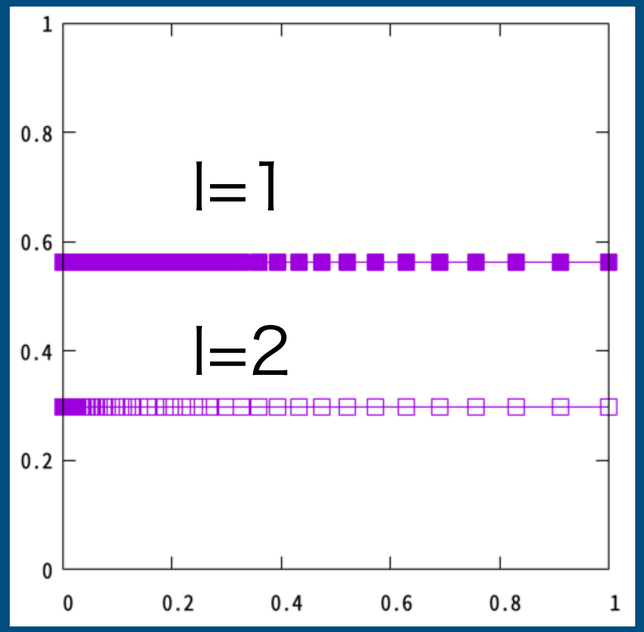
l

$\alpha = 25$

Hierarchical structure of the solutions

$q(x)$

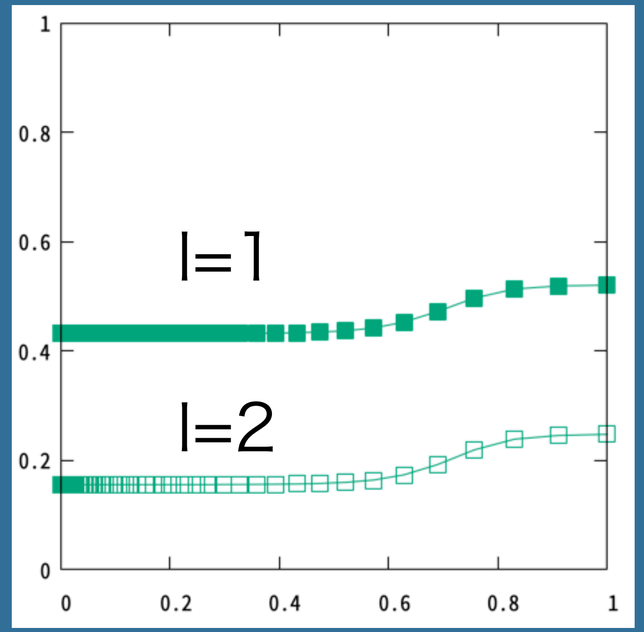
$r = 1$



x

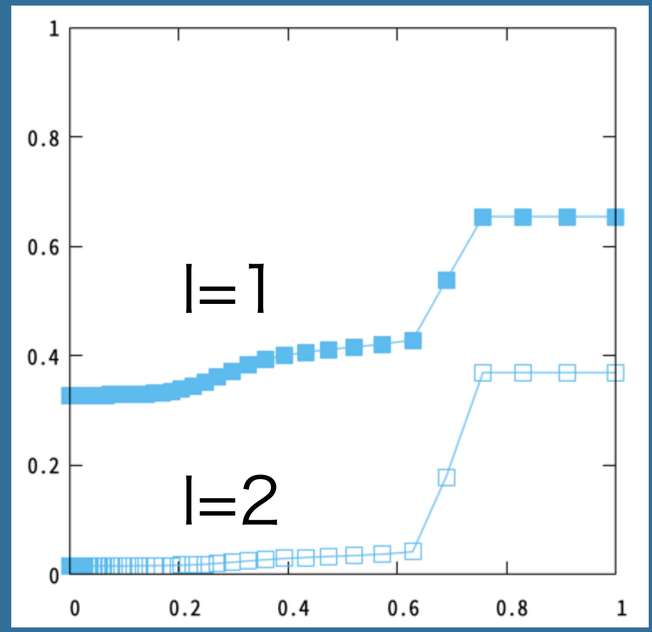
Replica symmetric

$r = 0.5$



x

$r = 0.1$

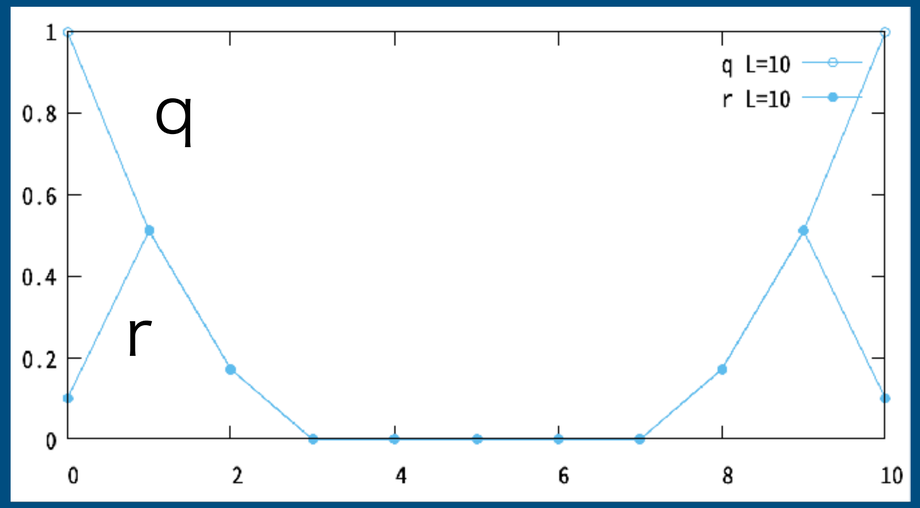


x

Replica symmetry broken

$r = 0.1$

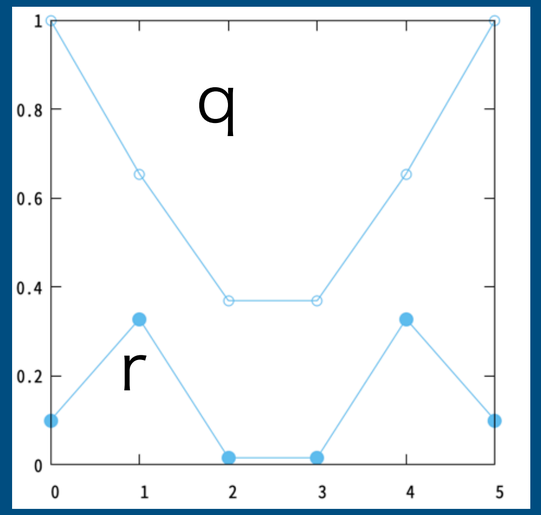
$q_{\mathrm{EA}} = q(1)$



l

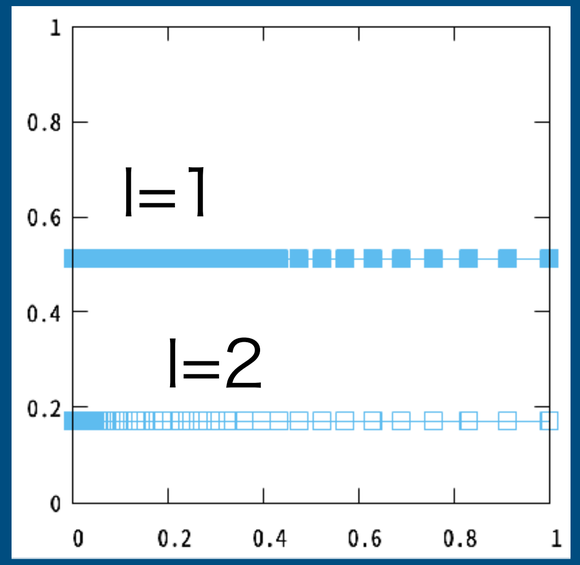
$r = 0.1$

$q_{\mathrm{EA}} = q(1)$



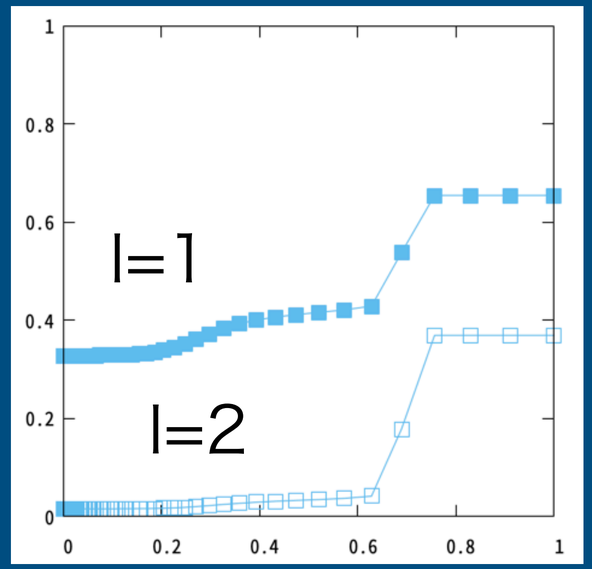
l

$q(x)$



x

$q(x)$



x

Replica symmetric!

- Simulation of learning in a teacher-student setting

t

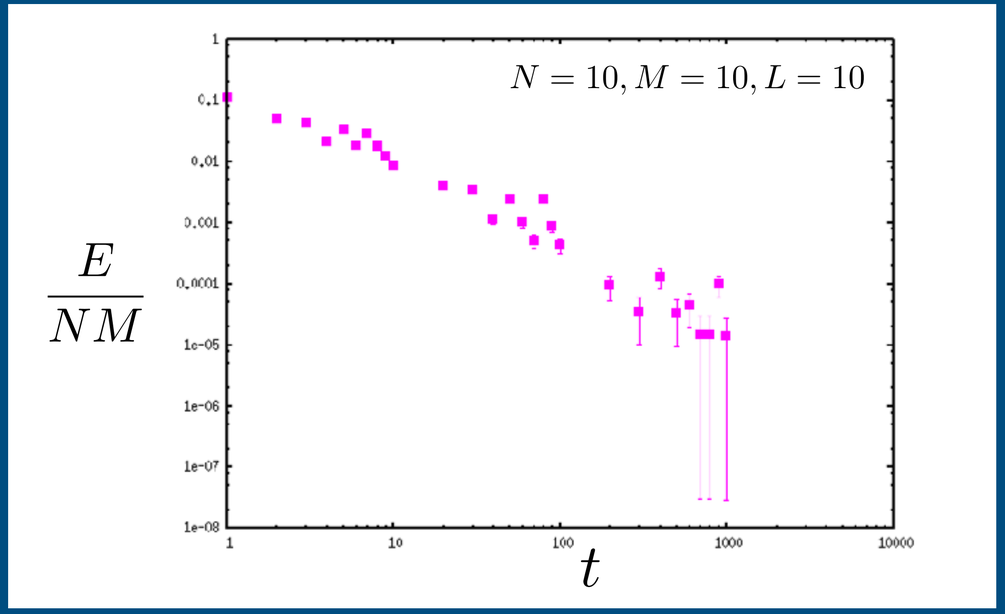
T=0 greedy Monte Carlo

random teacher + student "a" and "b"

loss function $E = \sum_{i=1}^{N} \sum_{\mu=1}^{M} \left(S_{L,i}^{\mu} - (S_*)_{L,i}^{\mu} \right)^2$

1. **"Unlearning"** : start from teacher's configuration

2. **"Learning"** : start from random configuration

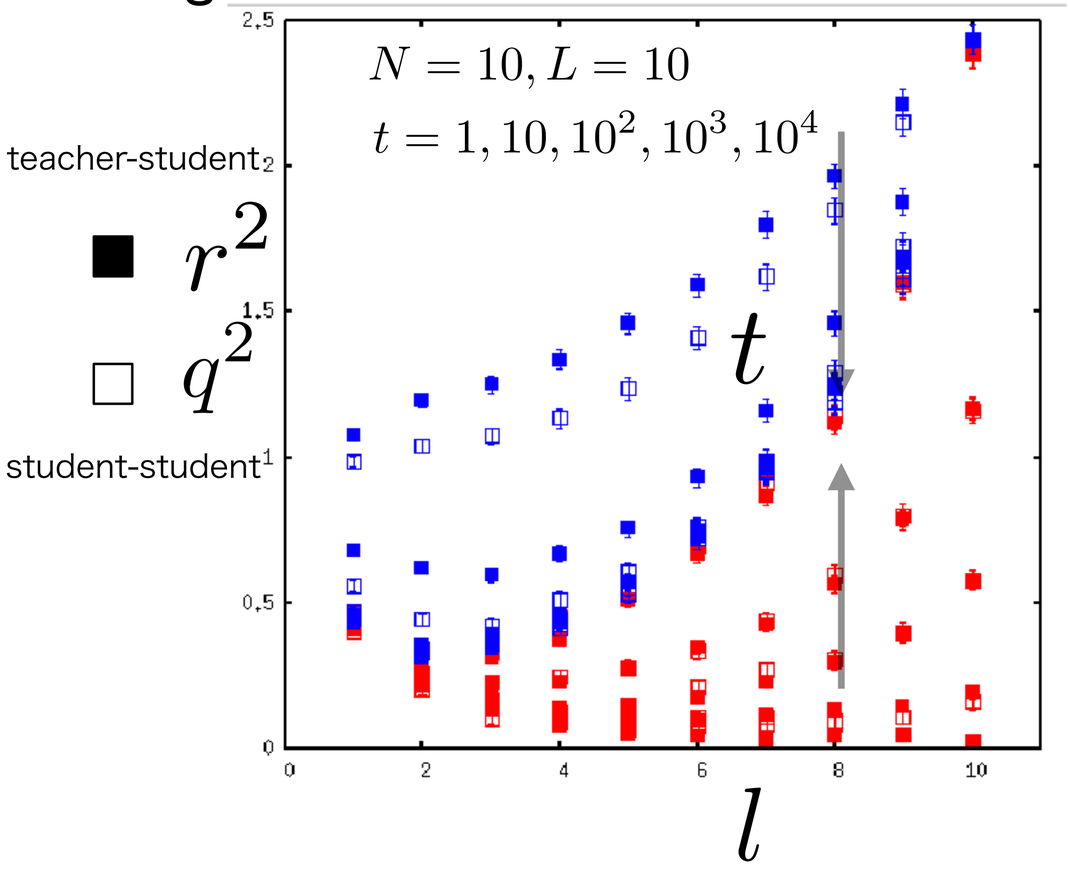


Permutation-invariant overlap

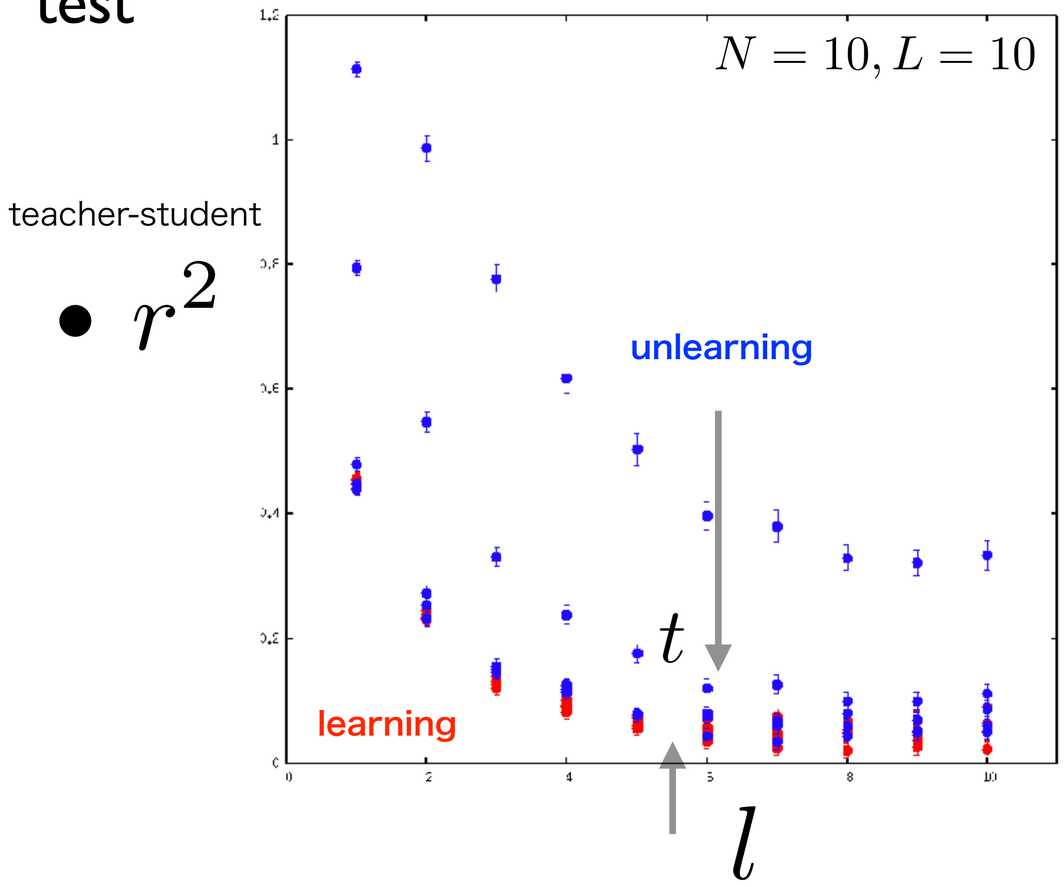
$$q^2 \equiv \frac{1}{N}\sum_{i,j=1}^{N} q_{ij}^2 - \frac{1}{\alpha} \qquad q_{ij} = \frac{1}{M}\sum_{\mu=1}^{M} (S_i^\mu)^a (S_j^\mu)^b$$

$$\alpha = 1$$

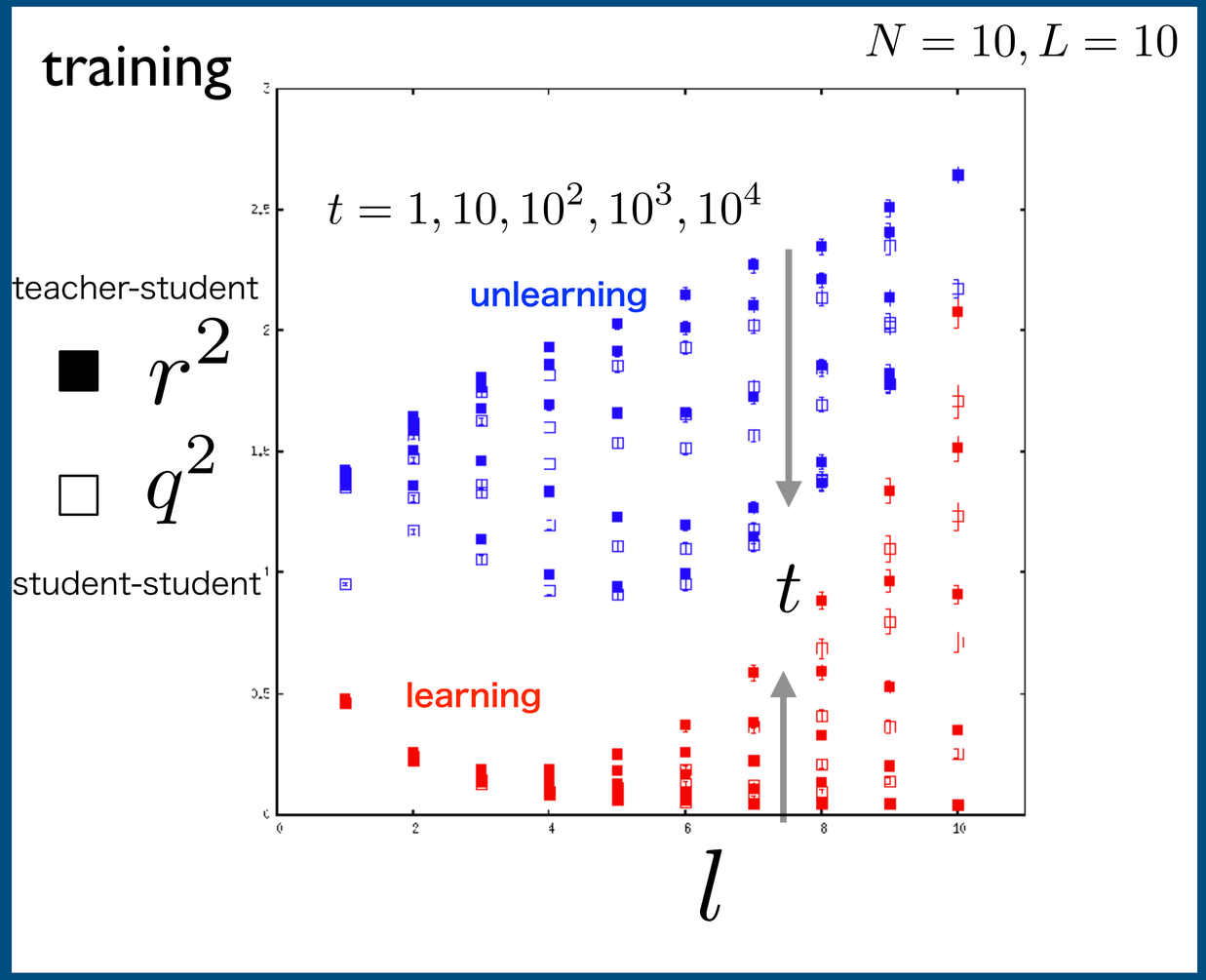
training

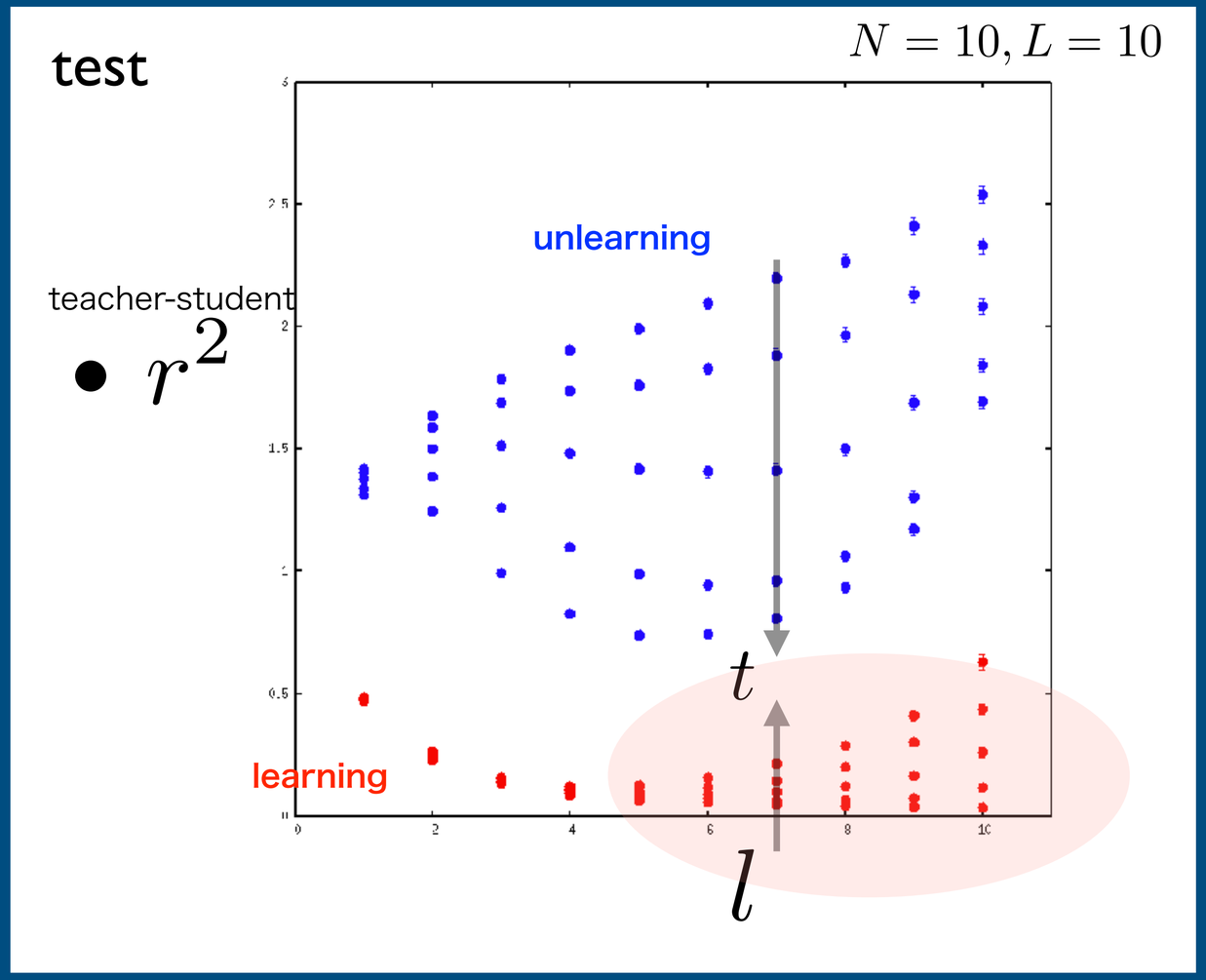


test



$\alpha = 32$


training
$N = 10, L = 10$
$t = 1, 10, 10^2, 10^3, 10^4$
teacher-student
r^2
q^2
student-student
unlearning
learning
t
l


test
$N = 10, L = 10$
teacher-student
r^2
unlearning
learning
t
l

Generalization ability

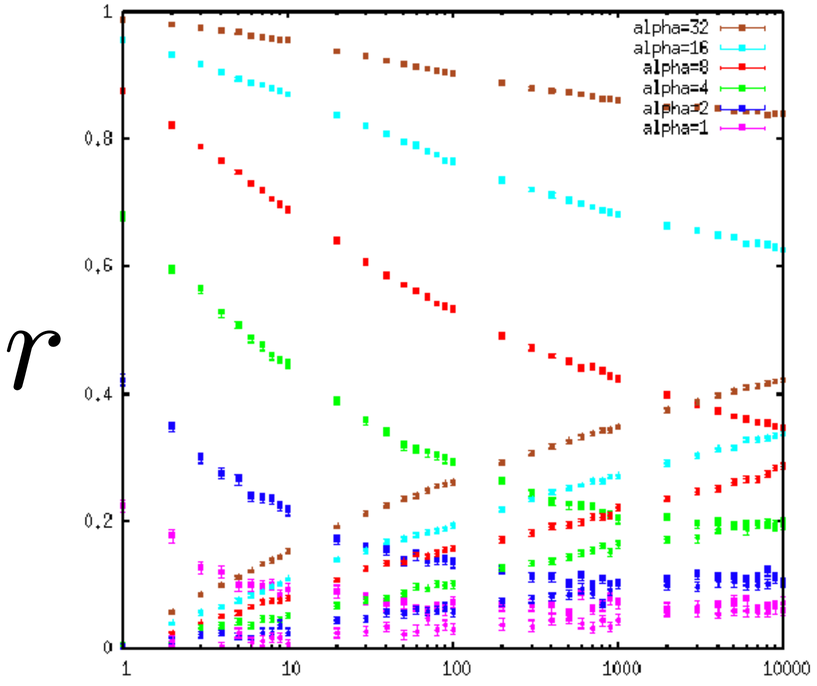
teacher-student overlap in the test

$$r_{\text{test}} = \frac{1}{NM} \sum_{i=1}^{N} \sum_{\mu=1}^{M} (S_i^{\mu})_{\text{test}}^{\text{teacher}} (S_i^{\mu})_{\text{test}}^{\text{student}}$$

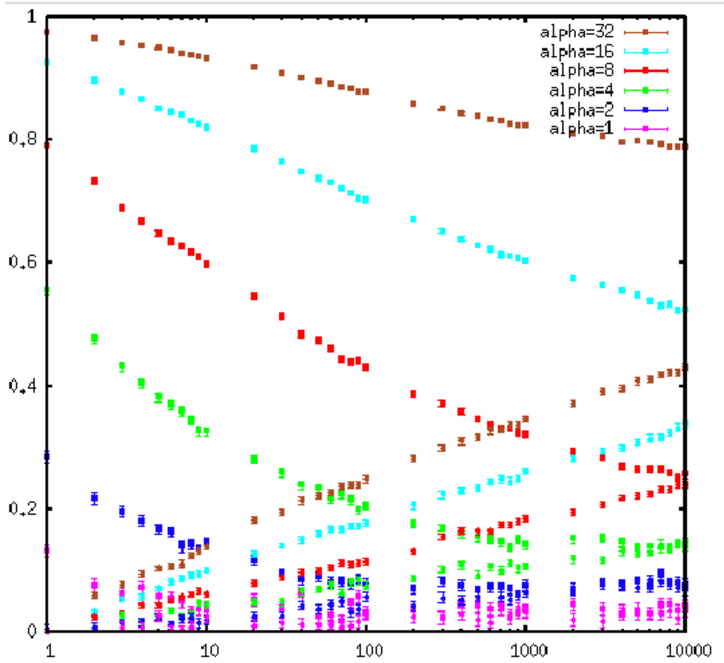
unlearning

learning

$N = 10, L = 5$

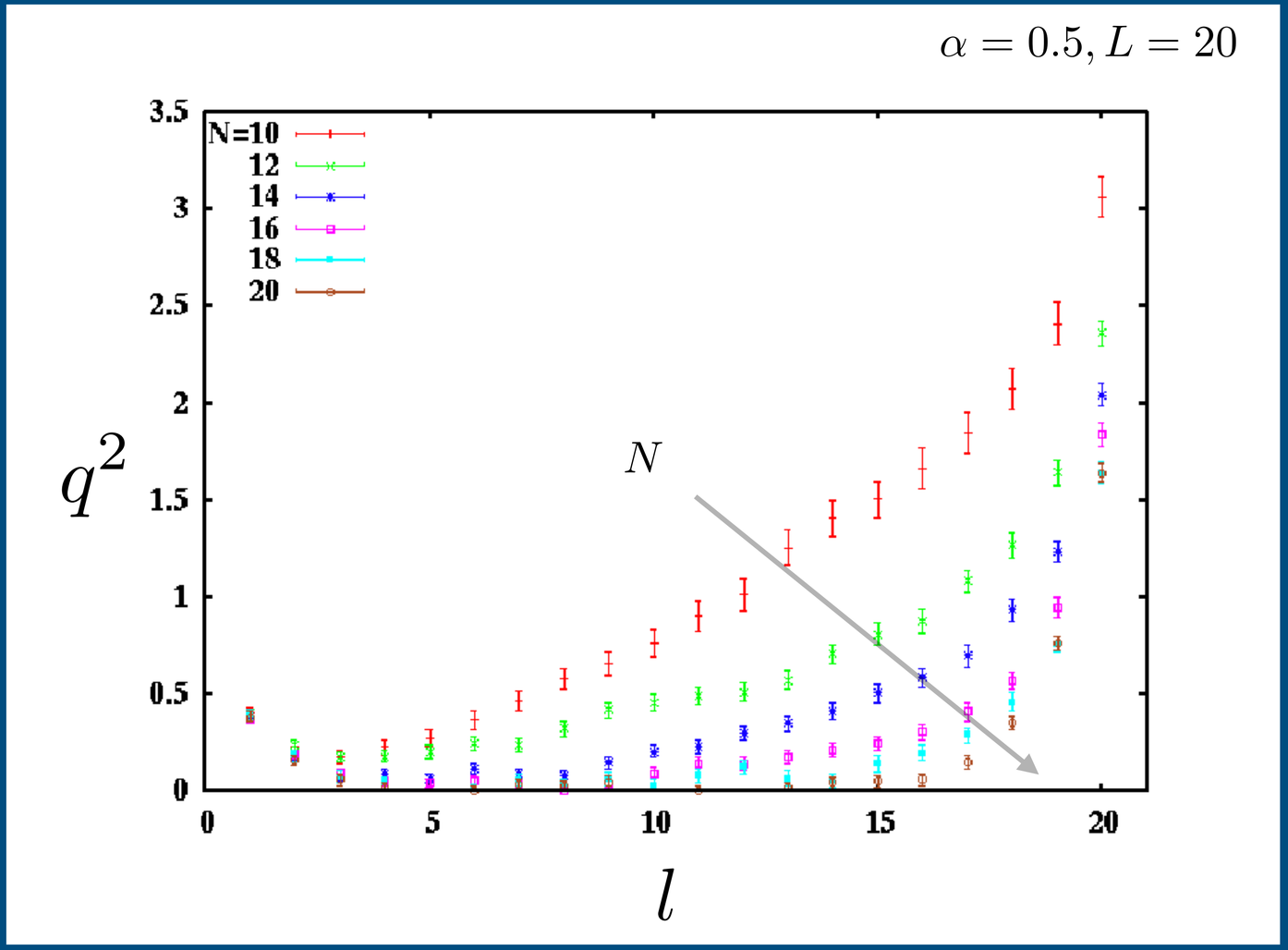


$N = 10, L = 10$



deeper systems generalize as well…

Increasing N with fixed $\alpha = M/N$



more symmetric with larger N

Summary

Construction of replica theory for a deep perceptron network

- random input/output (random constraint satisfaction problem)
- teacher-student scenario (statistical inference) with noise

"Wetting transition" in the design space with/without RSB

Numerical simulations of the teacher-student scenario

Outlook

Goldt, S., Mézard, M., Krzakala, F., & Zdeborová, L. (2020). PRX, 10(4), 041044.

Finite width N effect, hidden manifold model

c.f. MNIST $N = 784$ but $D_{\text{eff}} \simeq 14$

mismatch of architecture

other activation functions: sigmoid, ReLU,….

Simulations with "real data", various algorithms, architectures…