Disordered Systems in Physics, Information Theory and Computer Science

Marc Mézard

Ecole normale supérieure
PSL University

KITP Santa Barbara, January 2019

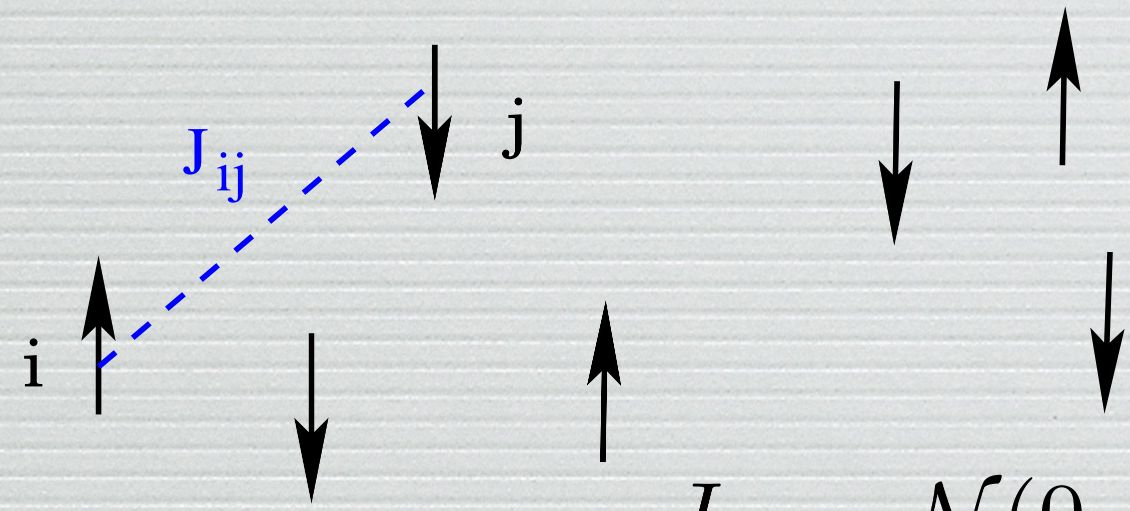
Table of Contents

Chapter One

Ensembles

Spin glasses in the 80's: « ensemble »

CuMn



$$J_{ij} \sim \mathcal{N}(0, 1/N)$$

$$s_i \in \{\pm 1\}$$

$$E_J(s) = -\sum_{ij} J_{ij} s_i s_j$$

$$P_J(s) = \frac{1}{Z_J} e^{-\beta E_J(s)}$$

Strongly disordered system:

Spin glass sample described by the whole set of J_{ij}

$O(N^2)$ parameters (if long range)

$$J_{ij} \sim \mathcal{N}\left(\frac{J_0}{N}, \frac{1}{N}\right)$$

$O(N)$ parameters (if short range)

$J_{ij} = \pm 1$
on Erdös-Renyi graph

Ensemble:

drawn from a probability distribution. eg iid

Thermodynamic limit and self-averaging

E.g. SK model

$$s_i \in \{\pm 1\} \quad J_{ij} \sim \mathcal{N}(0, 1/N)$$

$$E_J(s) = -\sum_{ij} J_{ij} s_i s_j$$

$$Z_J = \sum_{s_1,\dots,s_N} e^{-\beta E_J(s)}$$

$$E_J(s) = O(N)$$

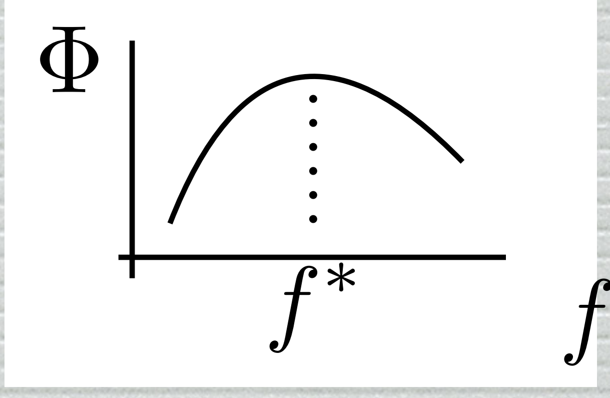
$$Z_J = e^{-\beta N f_J}$$

« Self averaging »

Probability of finding a sample with $f_J = f$: $\quad e^{N\Phi(f)}$

Almost all samples have $f_J = f^*$

therefore they have the same thermodynamics, phase diagram, etc.

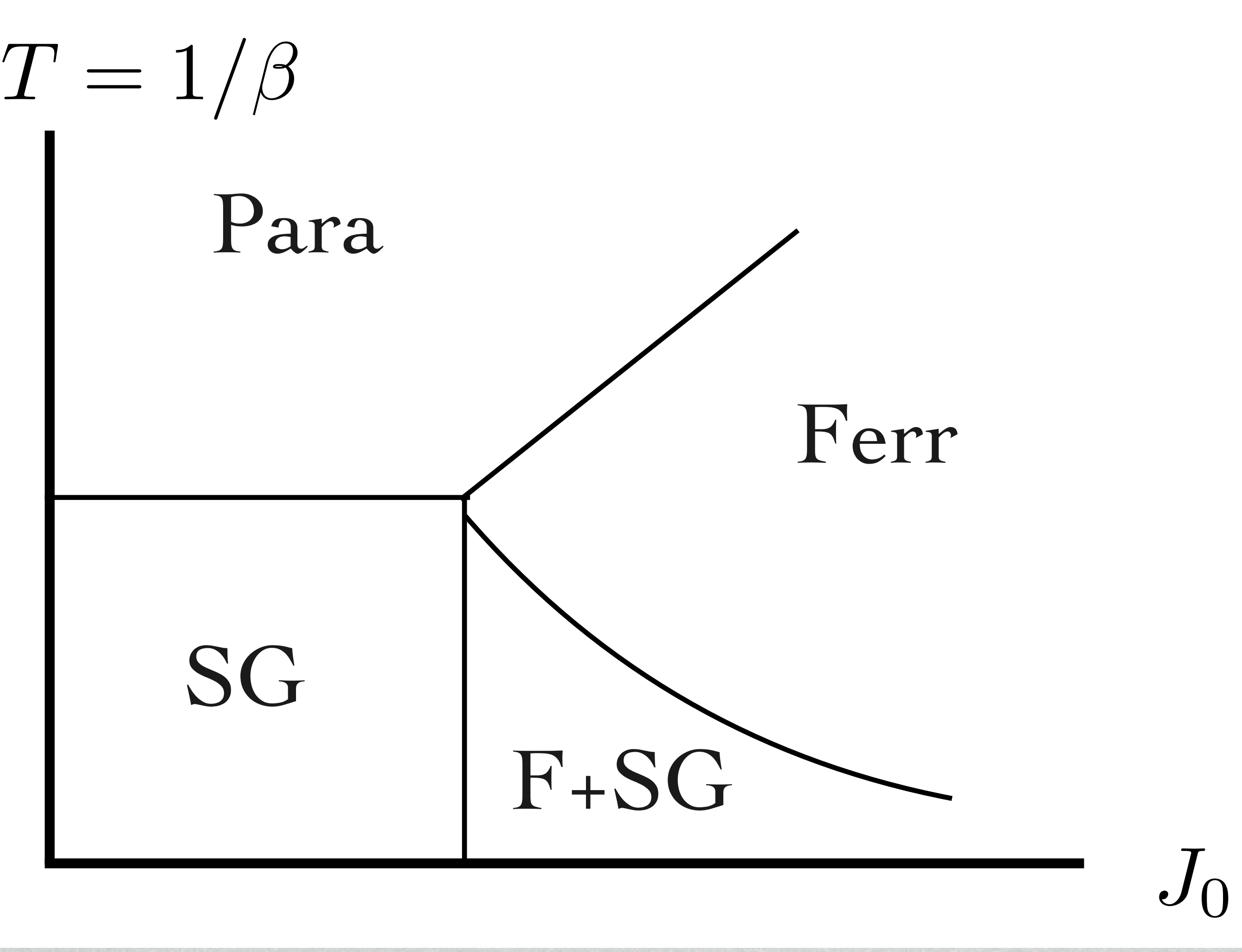


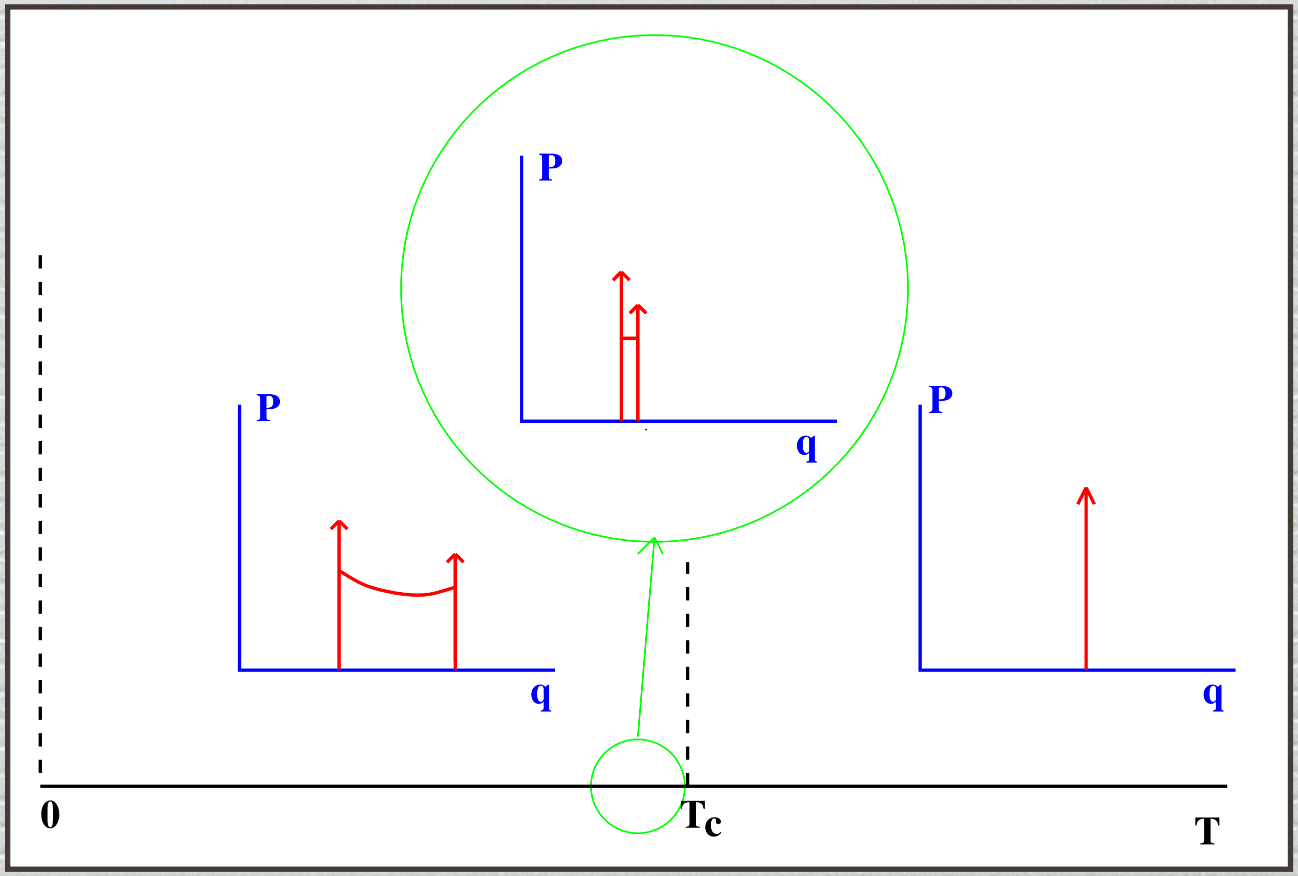
Phase diagram

eg SK model $\overline{J} = \frac{J_0}{N}$ $\overline{J^2} = \frac{1}{N}$

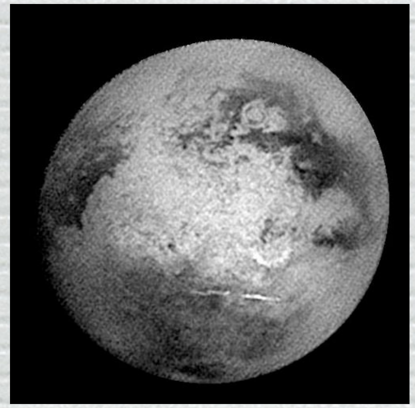
Ferro: $\frac{1}{N}\sum_{i=1}^{N}\langle s_i \rangle > 0$

SG: Prob(two random configs have overlap q)





Ensembles and phase transitions in information transmission: Shannon



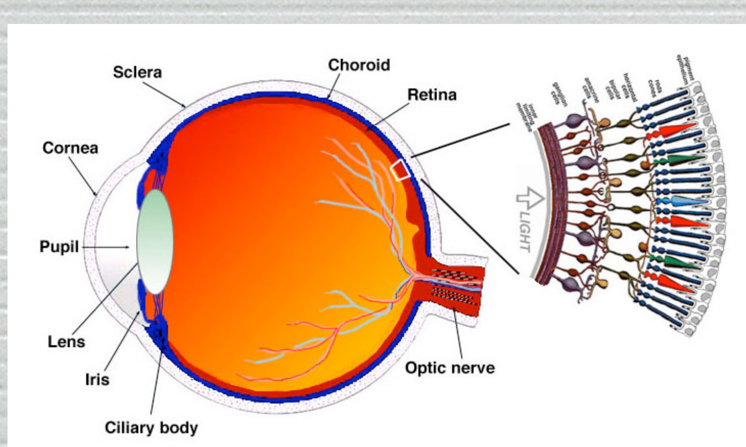
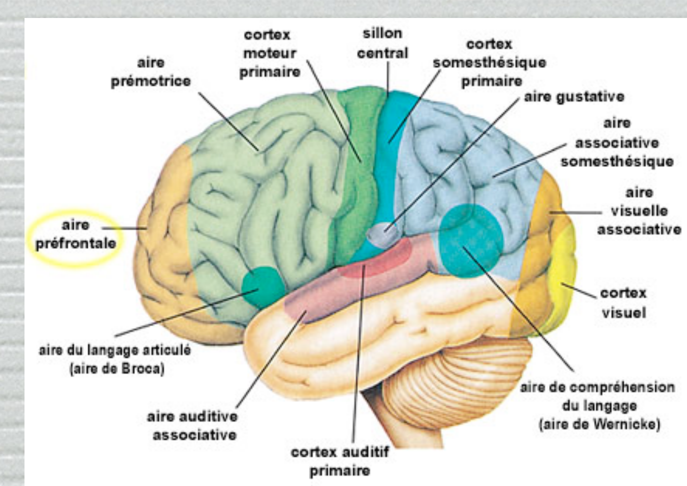
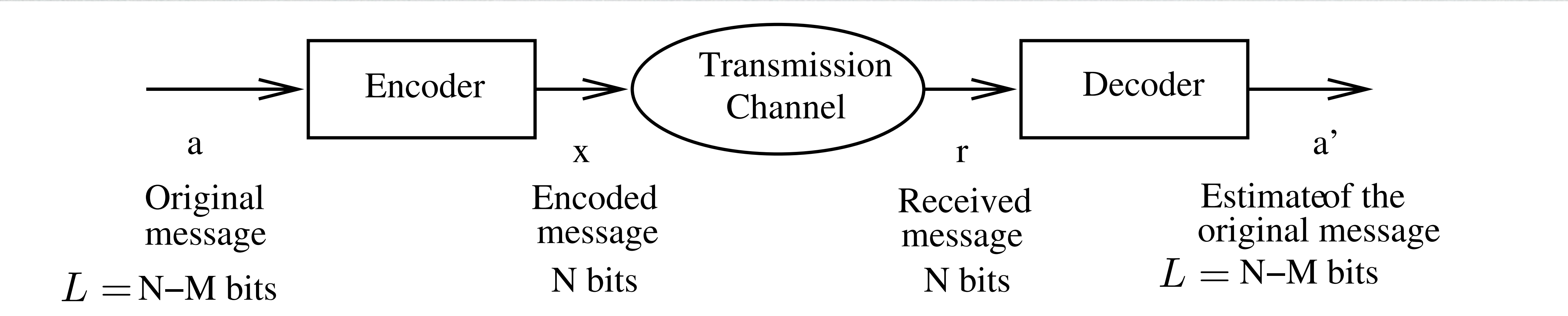


Fig. 1.1. A drawing of a section through the human eye with a schematic enlargement of the retina

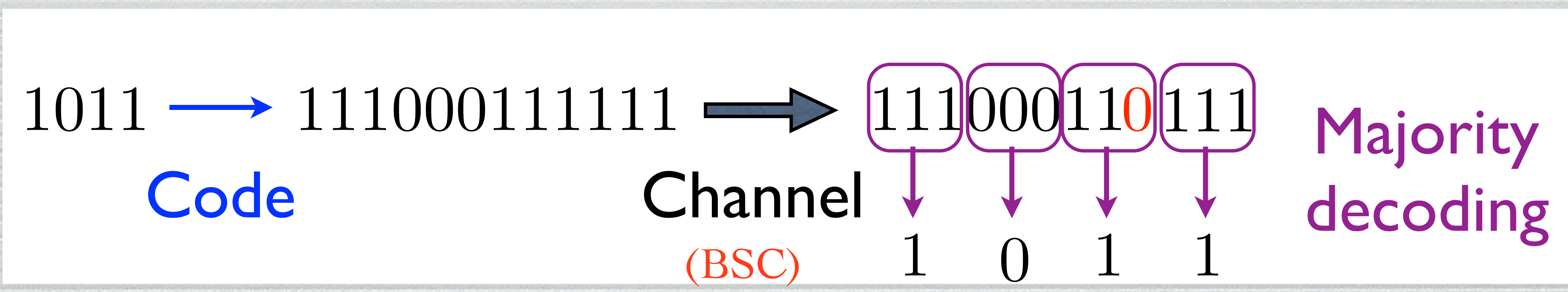


Principle of error correction : redundancy



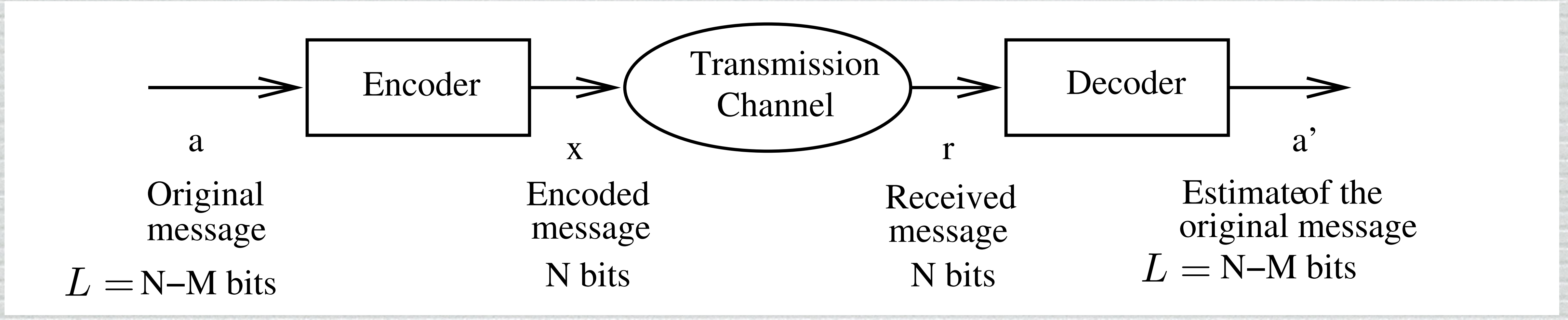
Encoding = add redundancy. Rate L/N

e.g. repetition $0 \rightarrow 000 \quad 1 \rightarrow 111 \quad \text{rate} = 1/3$



error probability $p^3 + 3p^2(1-p) \sim 3p^2$

Principle of error correction : redundancy



Encoding = add redundancy. Rate L/N

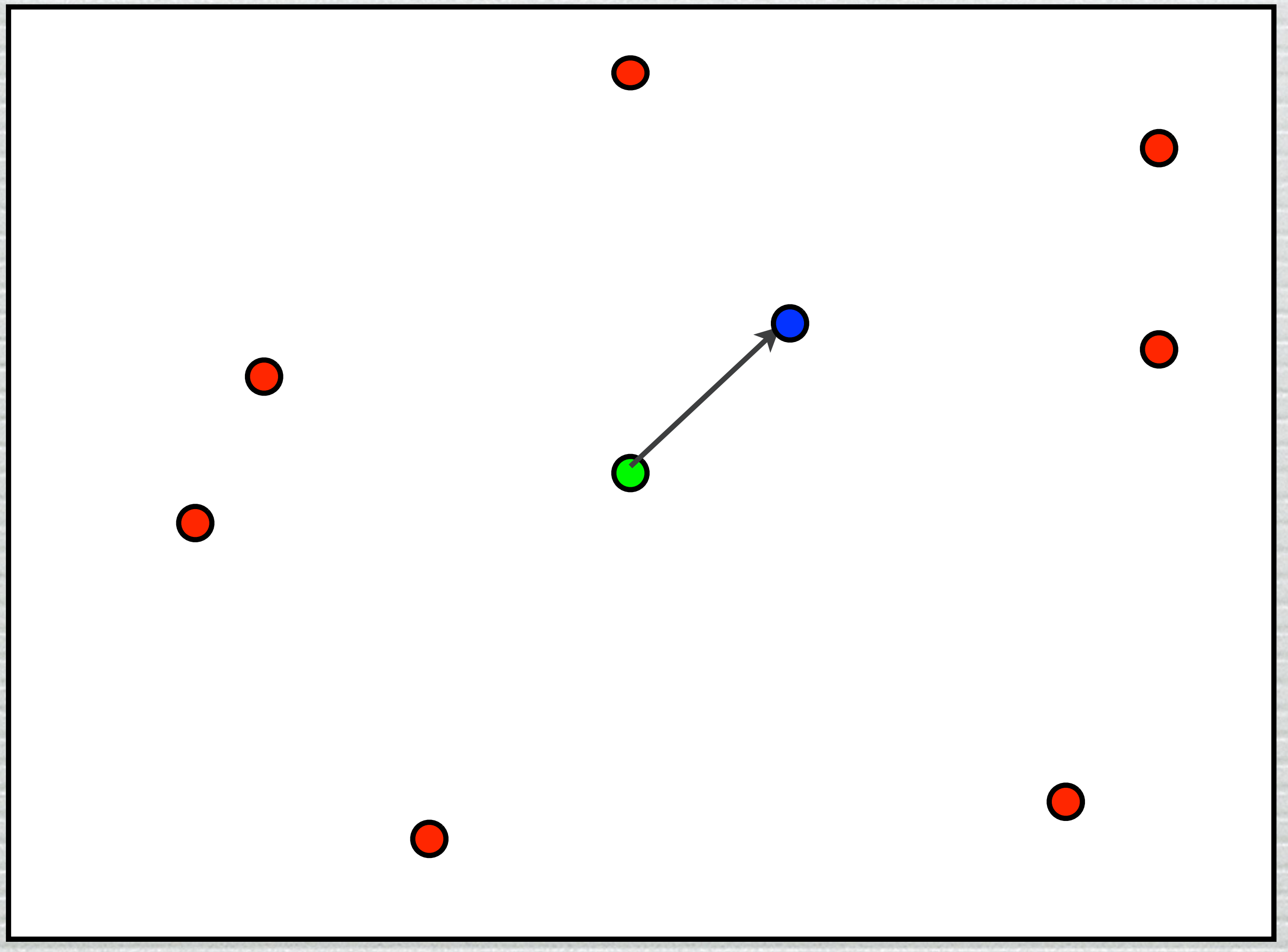
Shannon's theorem: for a given noise level p, one can build a coder/decoder which transmits with **zero error**, iff $r < r_c(p)$

Two ingredients:

- « Thermodynamic limit » $N, L \to \infty$
- Ensemble of Random Codes (~Random Energy Model of spin glasses)

Shannon code ensemble

Unit hypercube
in N dimensions



codewords (random)

sent codeword

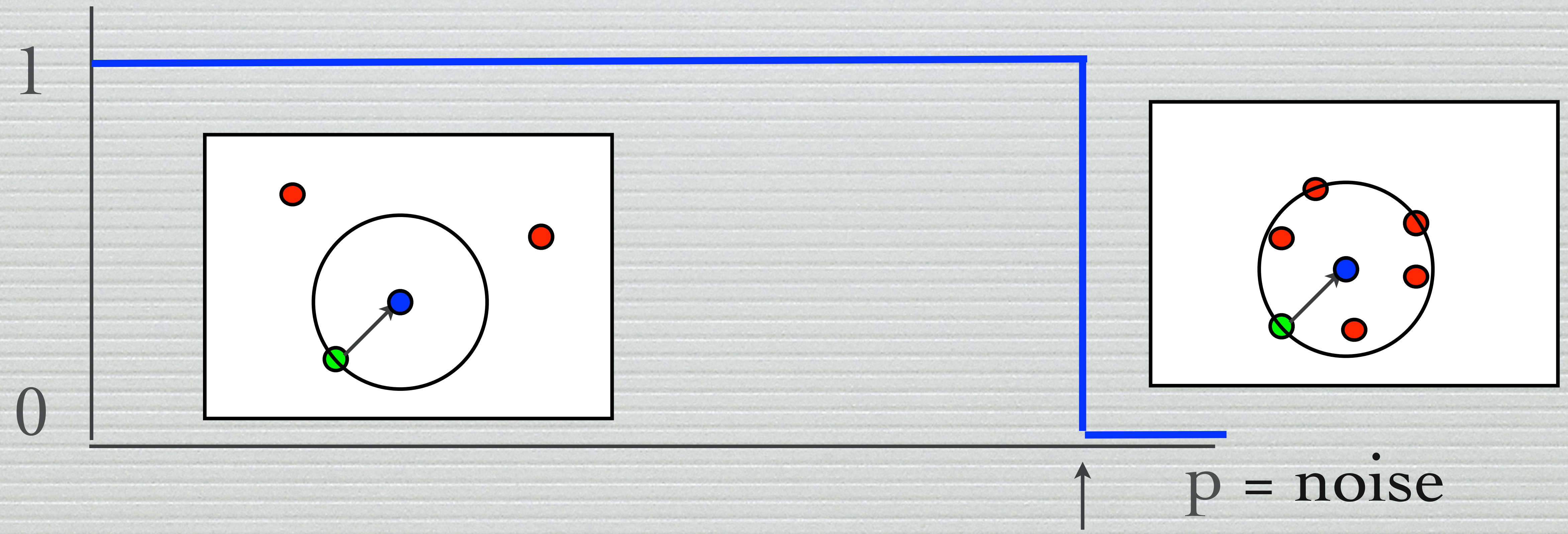
received word

2^{RN} iid random points, uniform distribution

Phase transitions in decoding

Decoding = find closest codeword

Probability of perfect decoding:



Shannon « bound »
geometric phase transition

Ensembles and phase transitions in computer science: Random Satisfiability

N Binary variables $x_i \in \{0,1\}$

M Constraints = clauses, e.g.: $x_1 \vee \overline{x}_2 \vee x_3$

Is there a configuration of the $\{x_i\}$ which satisfies all the constraints?

The grandfather of NP-complete problems. CNF

k-SAT (clauses of length $k \geq 3$) is also NP-complete

Typically hard instances: random k-SAT: Generate each clause with three randomly chosen variables in $\{x_i, \overline{x}_i\}$

Ensemble

Phase transition in the random k-SAT ensemble

Random k-SAT: N variables, M clauses. k variables in each clause, randomly chosen, randomly negated:

Large N limit: $\alpha = M/N$

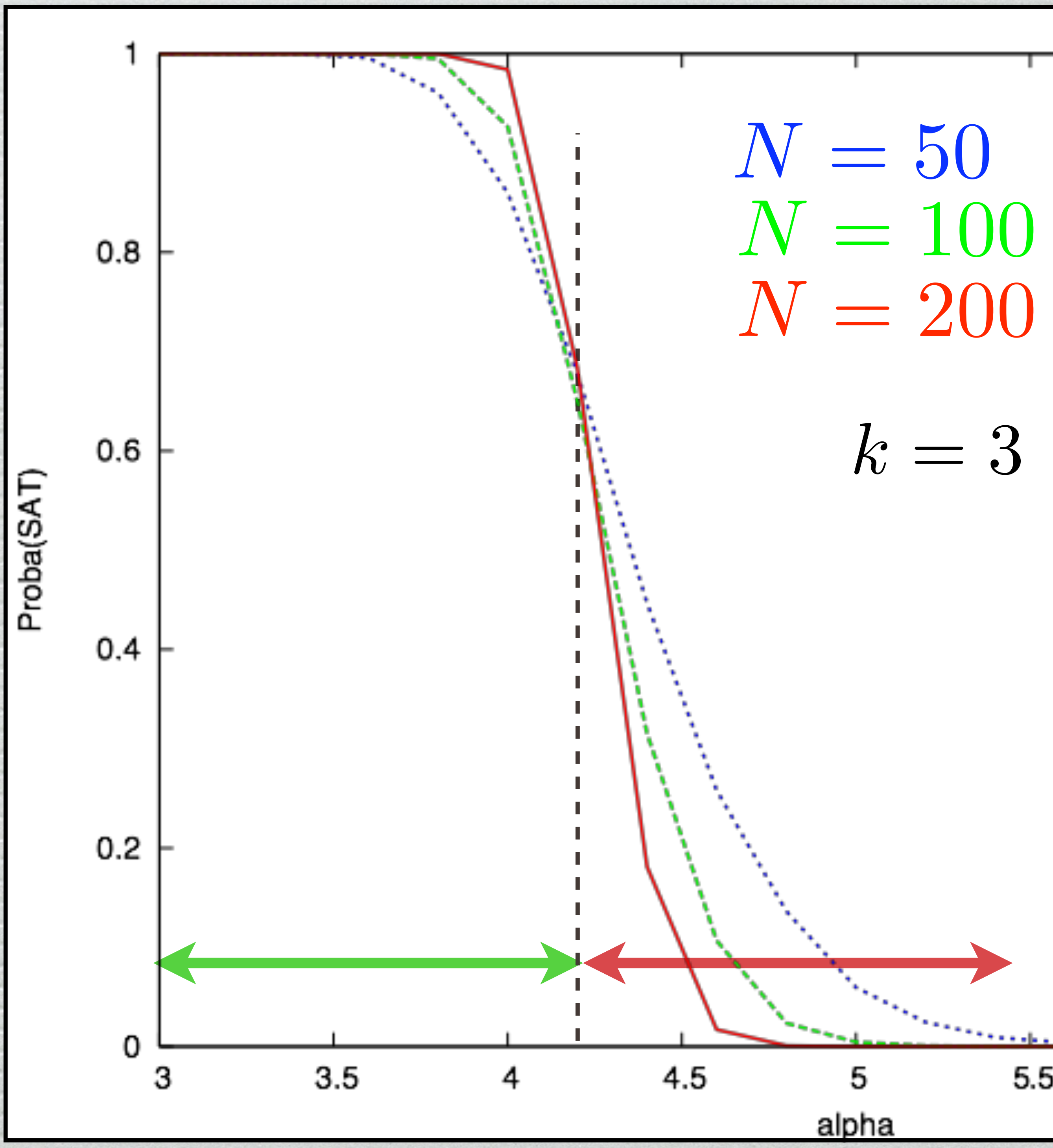
=density of constraints

Phase transition

SAT for $\alpha < \alpha_s$

UNSAT for $\alpha > \alpha_s$

Proven for k large enough by Ding-Sly-Sun (2015), making rigorous the stat phys approach from MM Parisi Zecchina (2002)



Chapter Two

Landscapes

Statistical physics of satisfiability

- many binary variables $x = (x_1, \cdots x_N),\ N \gg 1$
- Cost function $E(x)$ = Number of violated constraints = sum of three-body terms
- Find configuration of lowest cost

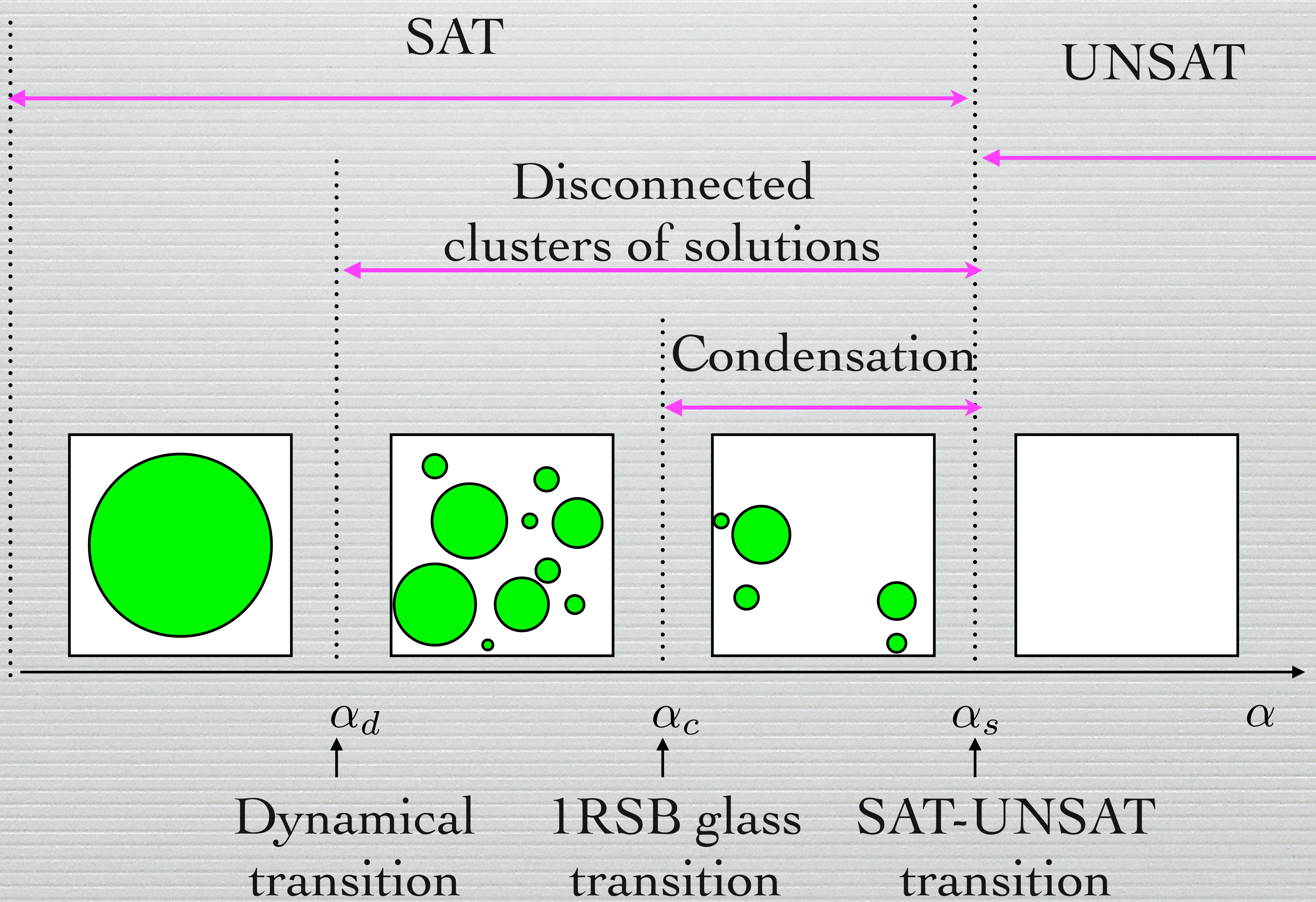
Uniform measure over all SAT assignments

$$P(x) = C\delta_{E(x),0}$$

Kirkpatrick, Selman; Monasson, Zecchina; Biroli, Monasson, Weigt; Mézard, Zecchina; Mézard, Parisi, Zecchina; Krzakala, Montanari, Ricci-Tersenghi, Semerjian, Zdeborova; Coja-Oghlan Panagiotou, Ding Sly Sun…

Results

Random k-Satisfiability: clustering



Clustered SAT phase: a glass phase

$e^{N\Sigma^*}$ clusters. Cluster μ has $\sim e^{Ns_\mu}$ solutions

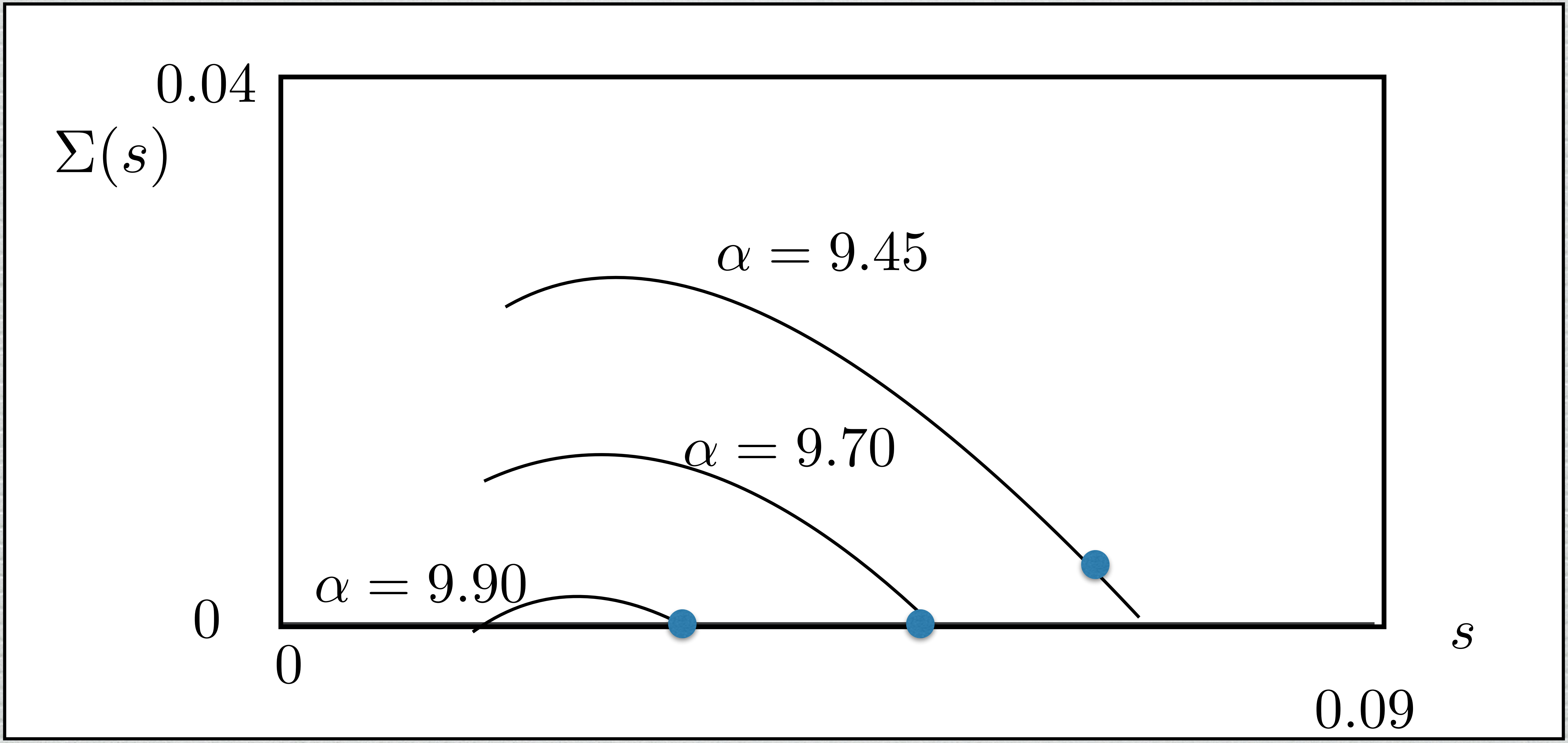
$\sim e^{N\Sigma(s)}$ clusters with $s_\mu = s$

Total number of solutions:

$$e^{N\Sigma^*} = \sum_\mu e^{Ns_\mu} = \int ds \; e^{N[\Sigma(s)+s]}$$

$$\Sigma^* = \max_s (\Sigma(s) + s)$$

4-SAT: Montanari, Ricci-Tersenghi, Semerjian



$\alpha_d = 9.38$ Clusters appear

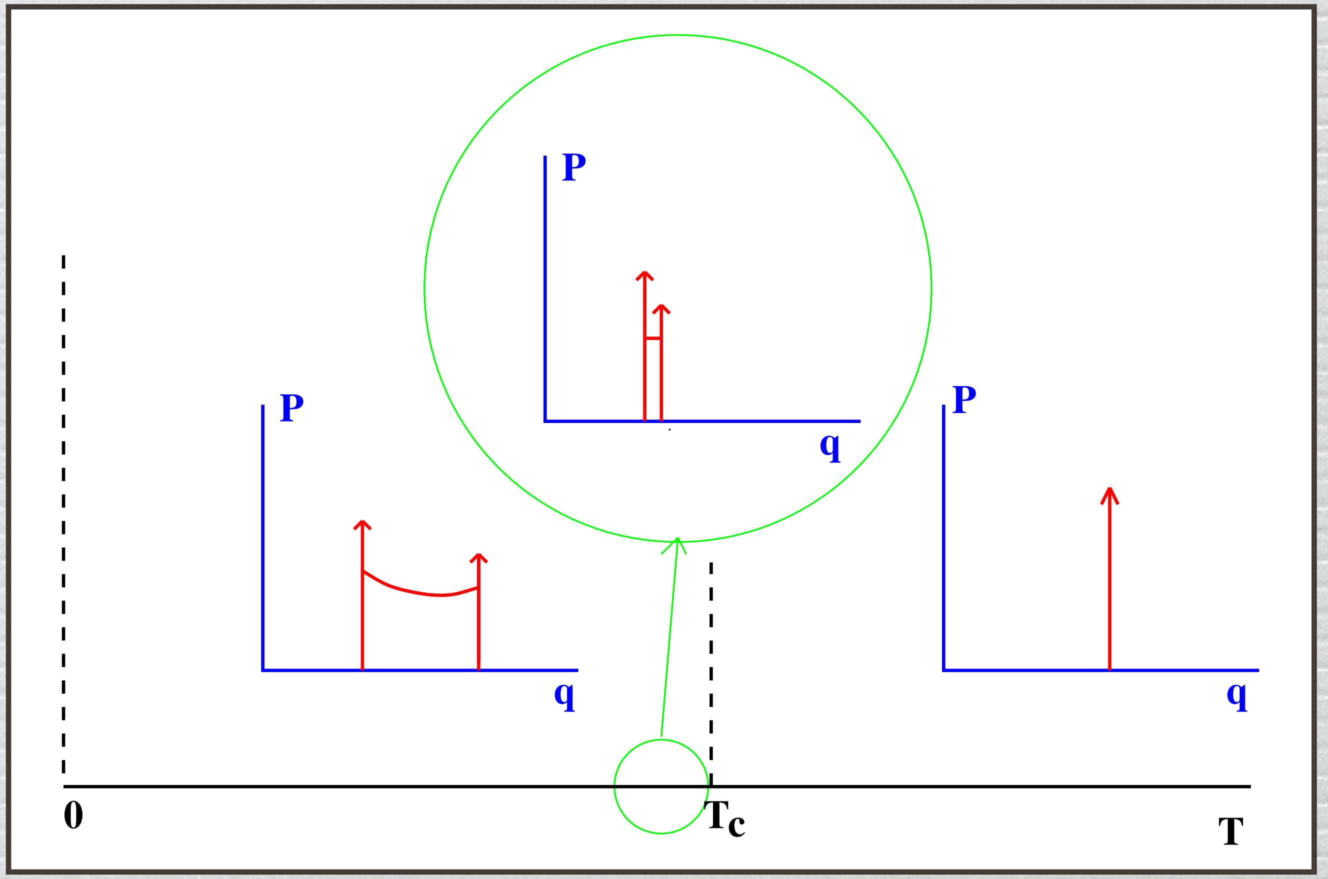
● : $\Sigma^* = \max_s(\Sigma(s) + s)$

$\alpha_c = 9.55$ Condensation on small number of clusters

$\alpha_s = 9.93$ SAT-UNSAT

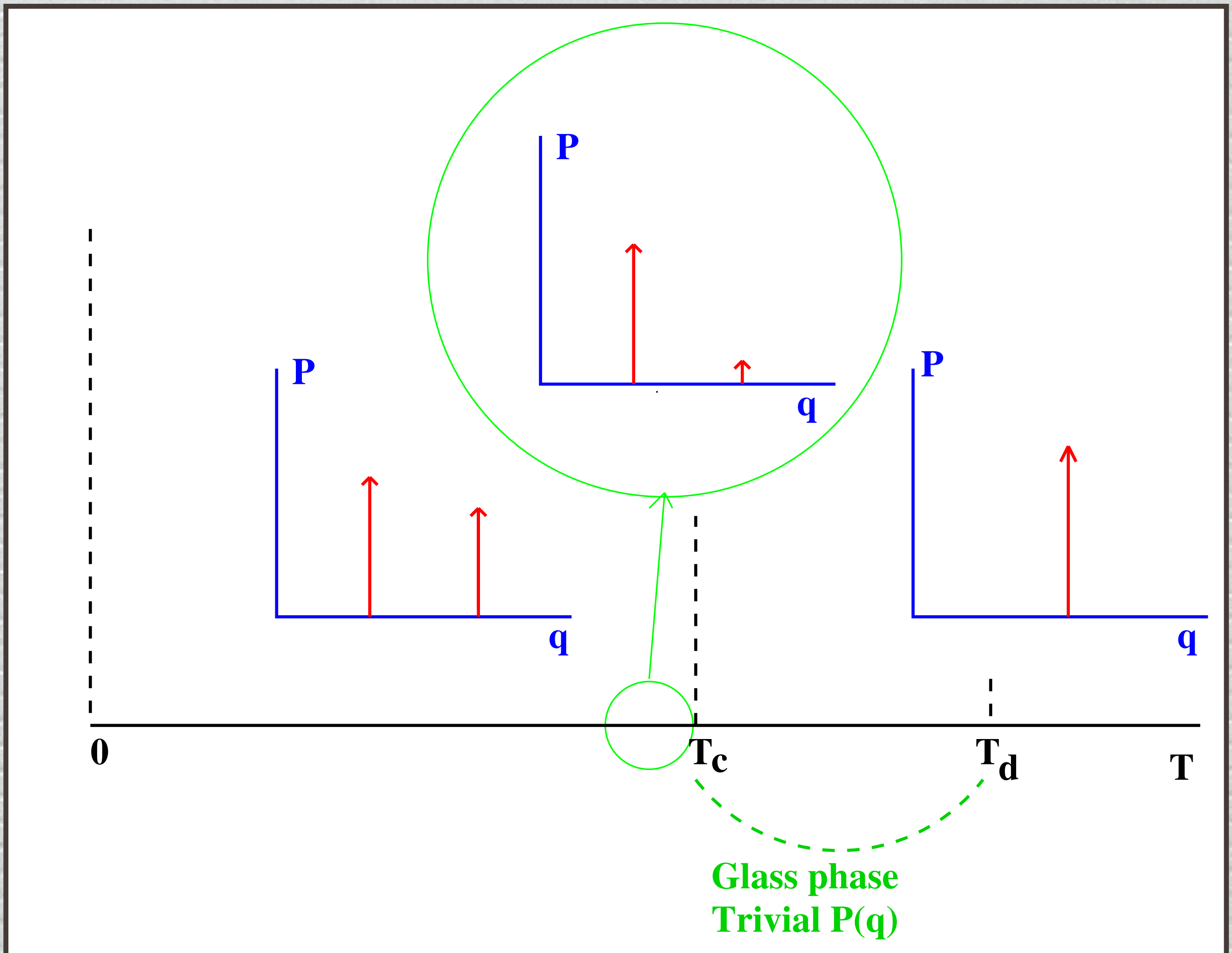
Two families of glasses

Probability (2 random configurations have overlap q)



Continuous transition

« Full replica symmetry breaking »

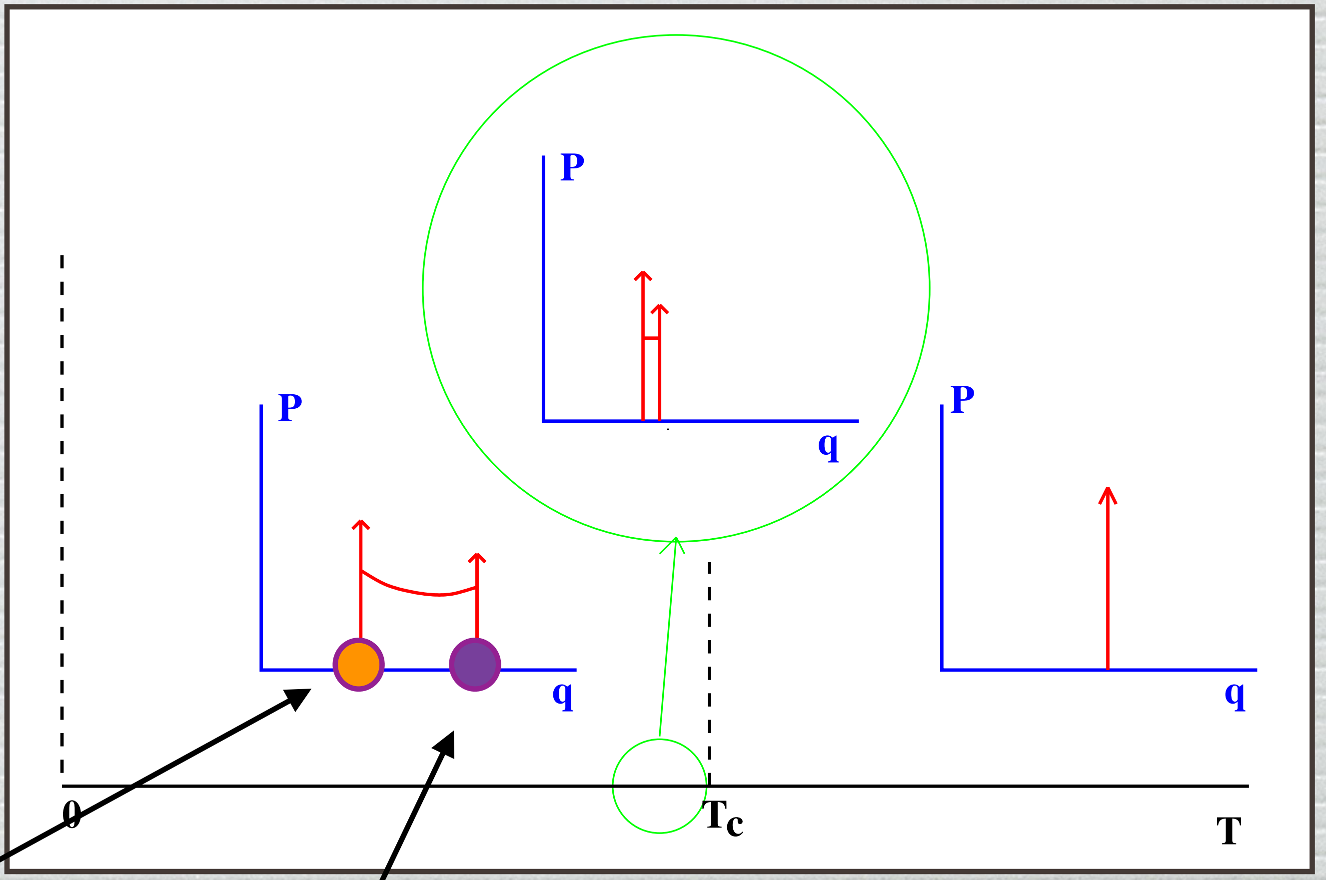


Discontinuous transition

« One step replica symmetry breaking »

Continuous transition

« Full replica symmetry breaking »

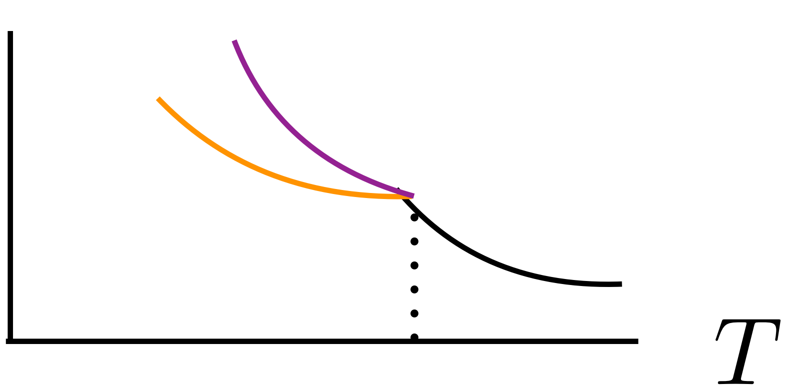


- Two replicas with small repulsion $\epsilon < 0$

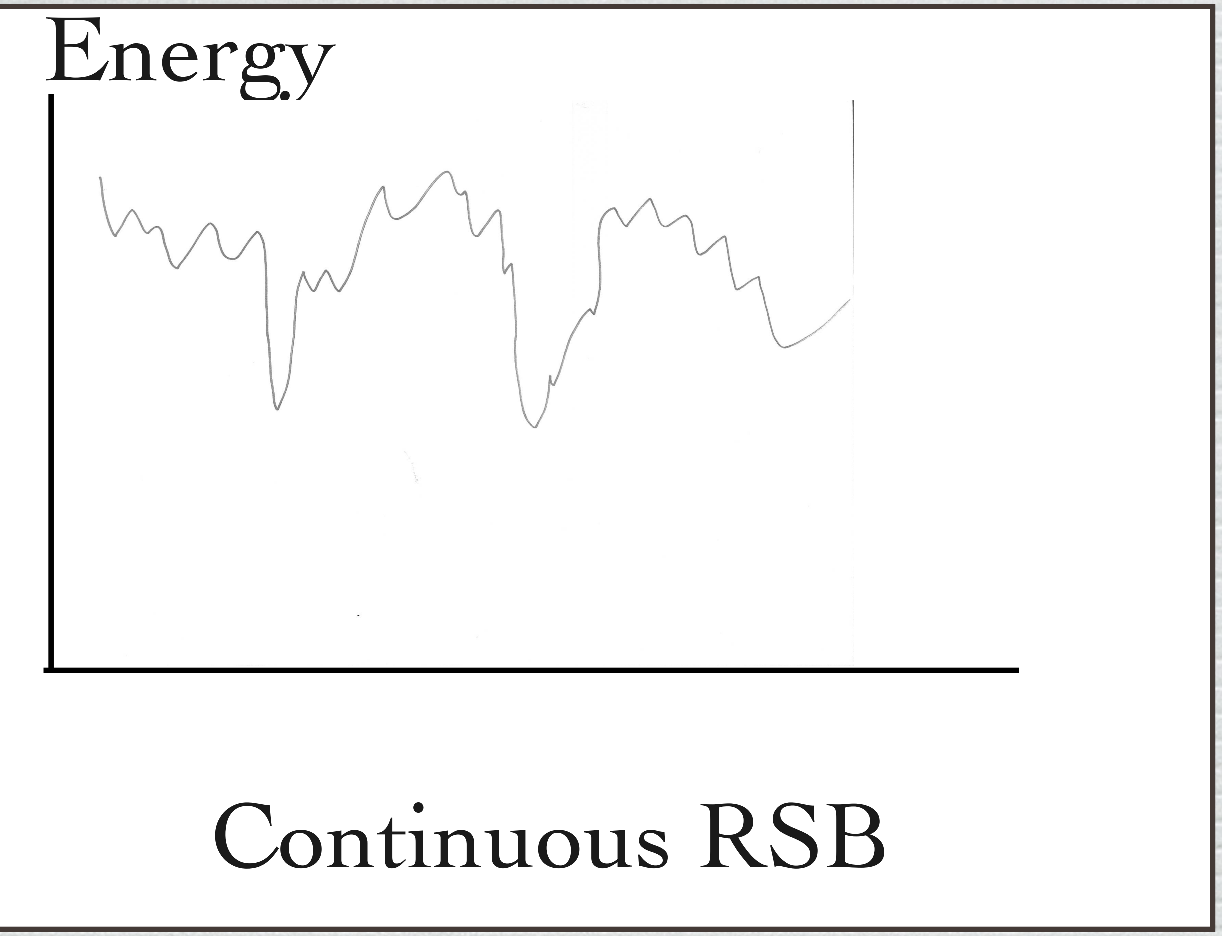
- Two replicas with small attraction $\epsilon > 0$

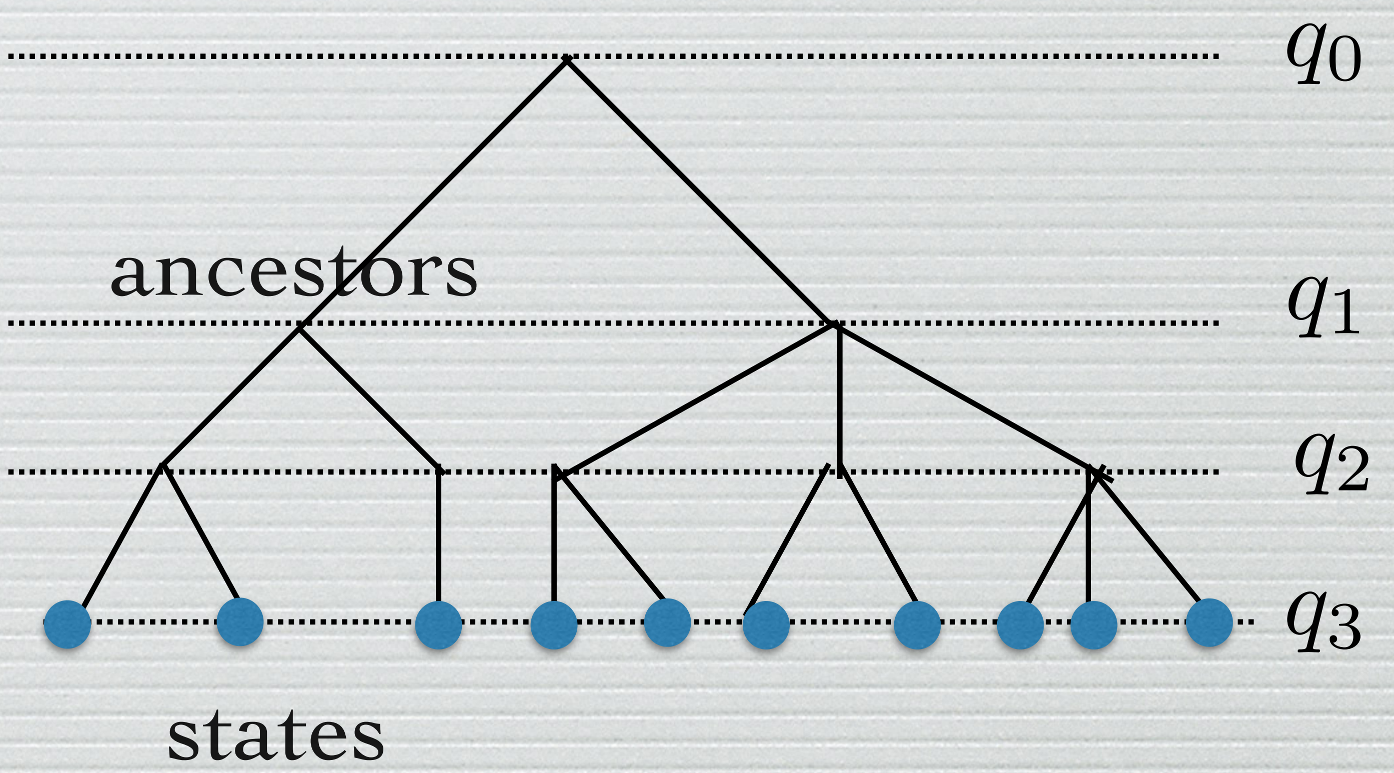
$$P_J(s, s') = \frac{1}{Z} e^{\beta \sum_{i,j} J_{ij}[s_i s_j + s'_i s'_j] + \beta H \sum_i [s_i + s'_i] + \epsilon \sum_i s_i s'_i}$$

$$q = \frac{1}{N} \sum_i \langle s_i s'_i \rangle$$



Spin glass landscape (misleading drawing, but…)





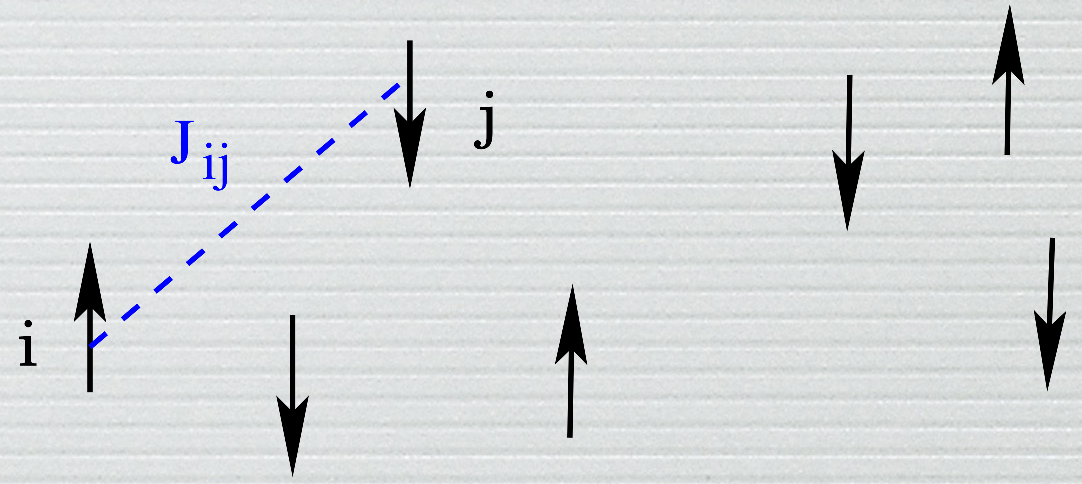
1- Glass « phase » : Many pure states, unrelated by symmetry, organized in a hierarchical « ultrametric » structure

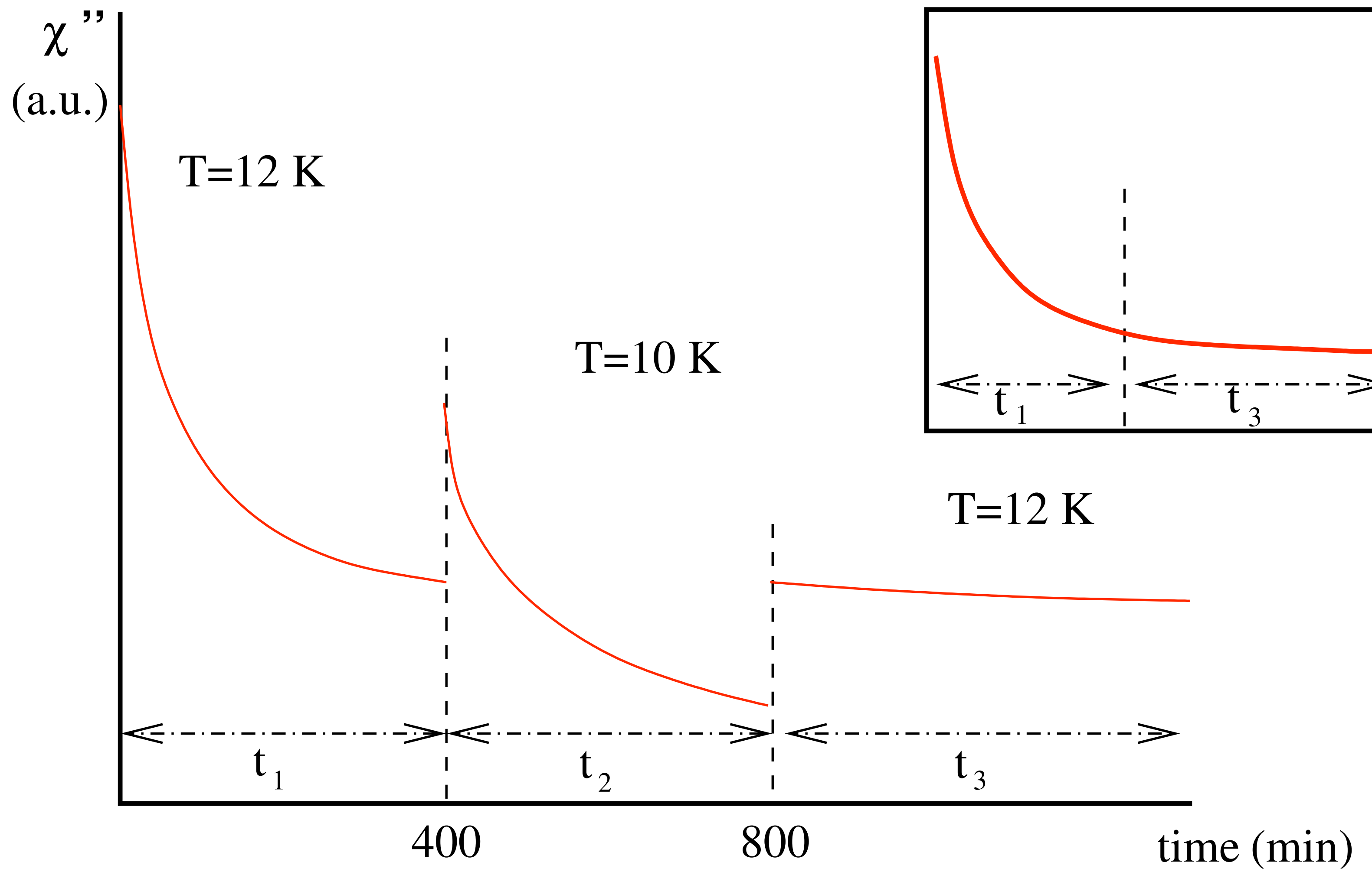
2- Exploit the hierarchical structure for algorithm (Montanari 2019)

Two main techniques, replicas and cavity/TAP

Spin Glasses

Linear response to a small magnetic field:



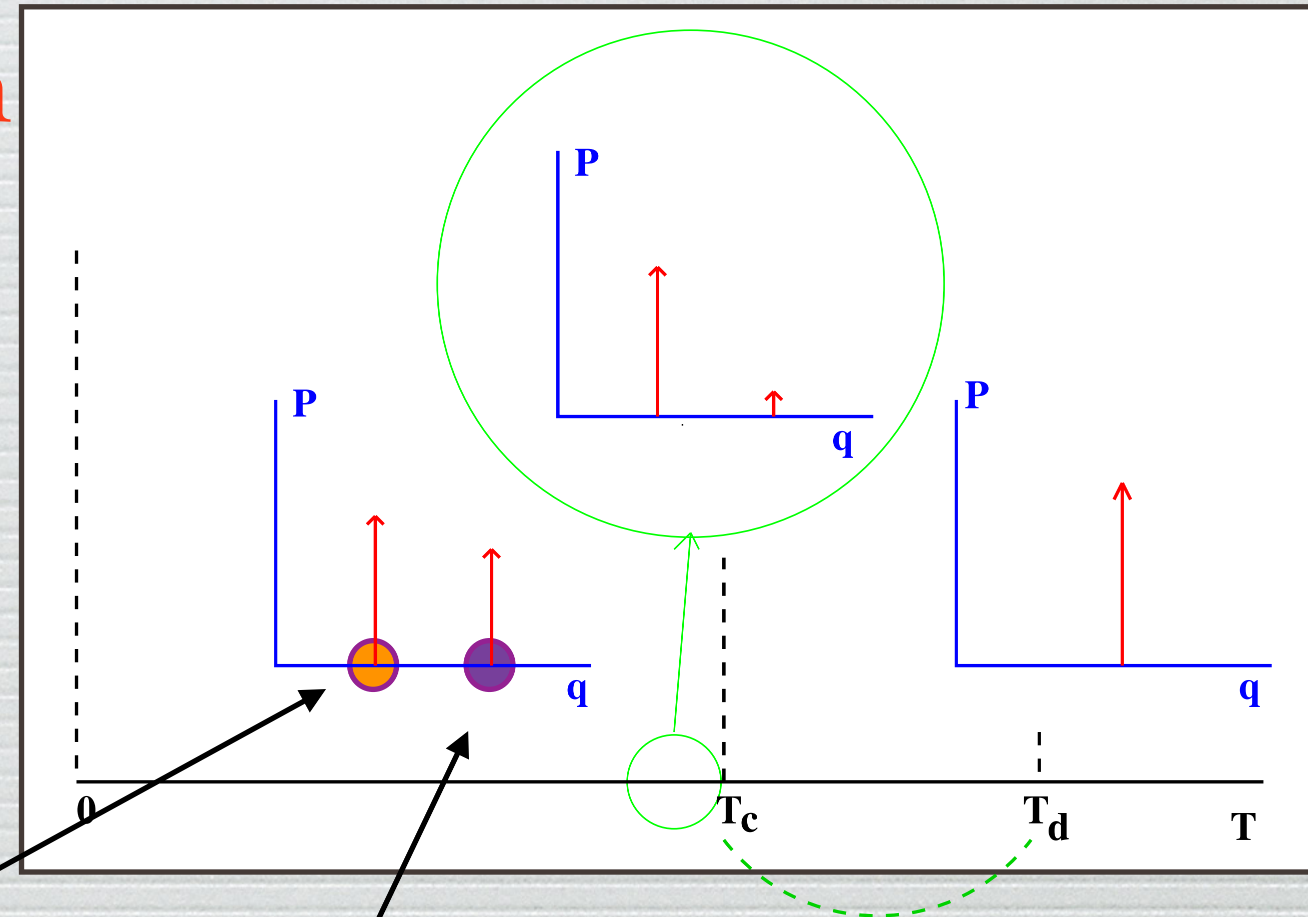


New dynamics
Memory

E. Vincent et al, SPEC

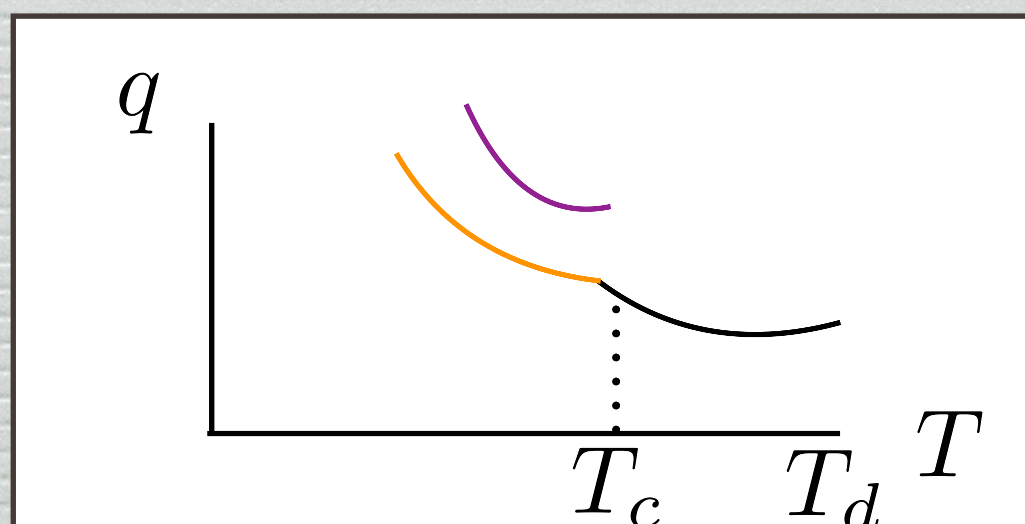
Discontinuous transition

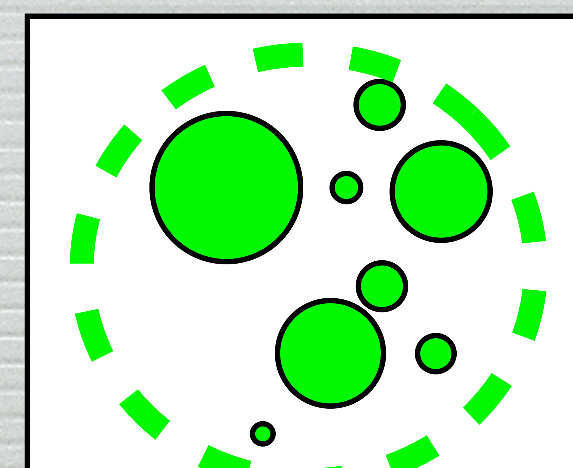
« Discrete (1 step, 2steps…) replica symmetry breaking »



- Two replicas with small repulsion $\epsilon < 0$

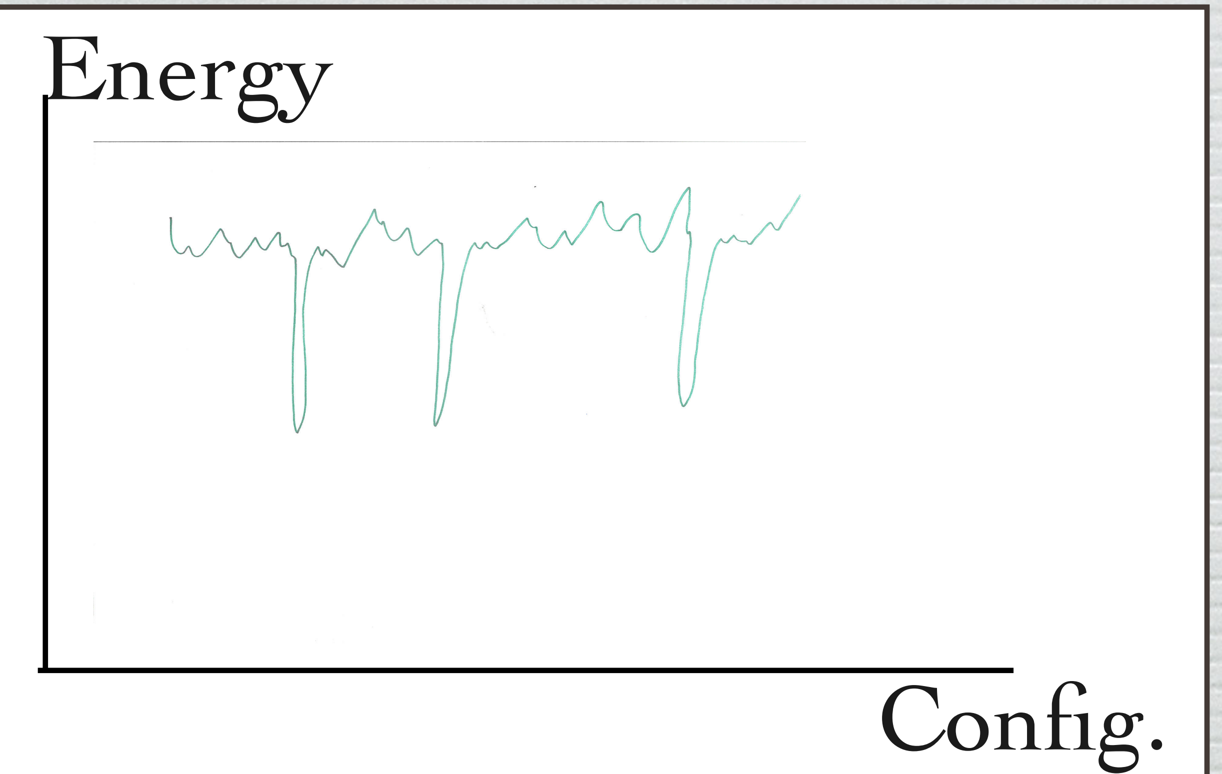
- Two replicas with small attraction $\epsilon > 0$



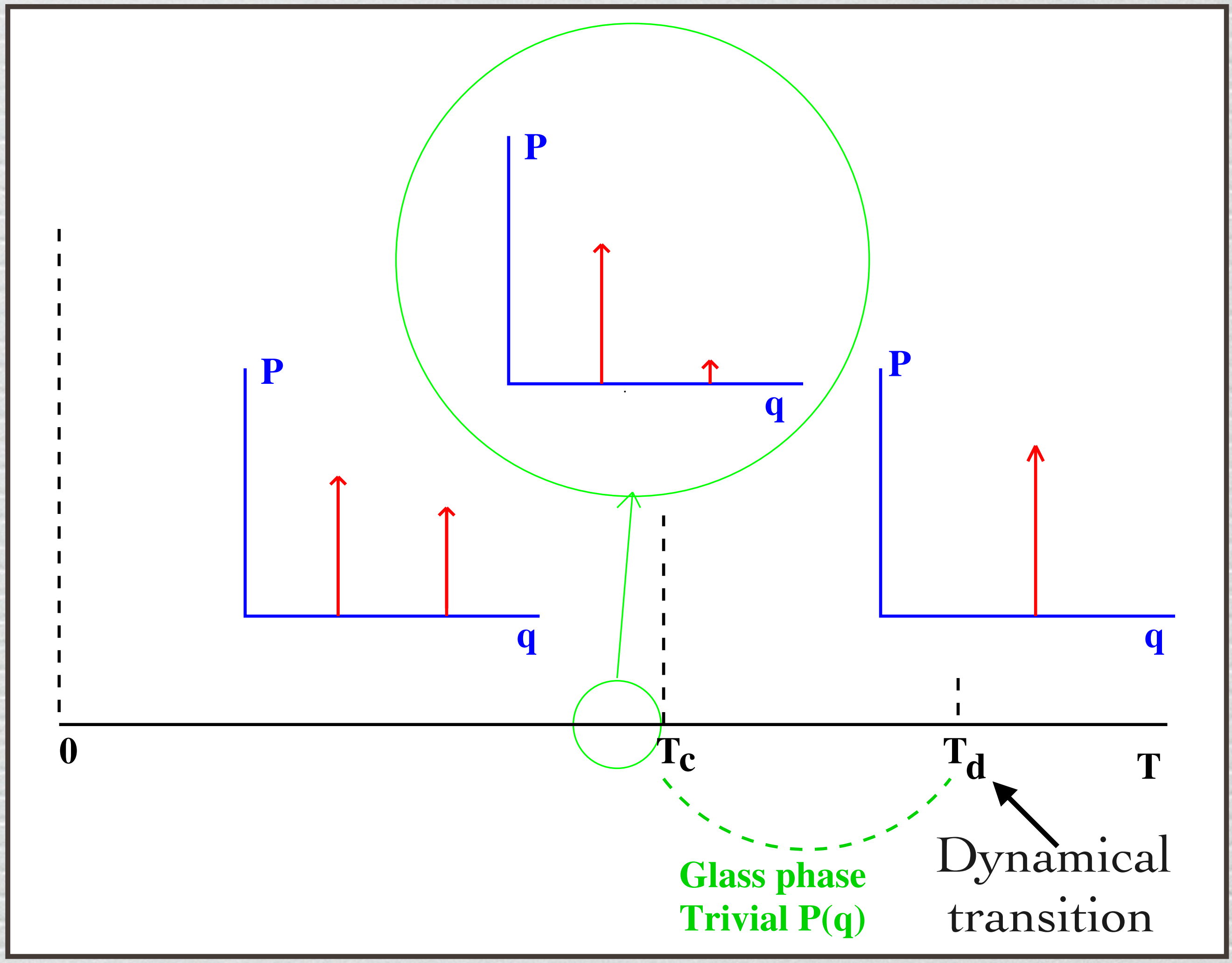


Discontinuous transition

Golf-course landscape : harder to find ground state



Discontinuous RSB



q

If the measure condenses on a small number of clusters: non-trivial P(q)

Otherwise: need to study the measure with two coupled configurations at a fixed distance

Chapter Three

Replicas

Replicas, version 1: analytic continuation

E.g. spin glasses

$$s_i \in \{\pm 1\} \quad J_{ij} \sim \mathcal{N}(0, 1/N)$$

$$E_J(s) = -\sum_{ij} J_{ij} s_i s_j$$

$$f_J = -\frac{1}{\beta N} \log Z_J$$

is self-averaging

$$Z_J = \sum_s e^{-\beta E_J(s)}$$

Compute $\mathcal{E}(f_J)$ average over J

$$E_J(s) = O(N)$$

$$\mathcal{E}(\log Z_J) = \lim_{n \to 0} \mathcal{E}([Z_J^n - 1]/n)$$

$$Z_J = e^{-\beta N f_J}$$

$$Z_J^n = \sum_{s^1, \ldots, s^n} e^{-\beta[E_J(s^1) + \cdots + E_J(s^n)]} \;:$$

n uncoupled replicas, same disorder

$\mathcal{E}(Z_J^n)$: n coupled replicas, no disorder

Replicas, version 1: analytic continuation

$\mathcal{E}(Z_J^n)$: n coupled replicas, no disorder, $\mathcal{S}_n$ symmetry

Analytic continuation $n \to 0$

- Often not unique (Carlson)
- Phase transitions in the $N \to \infty$ thermodynamic limit (spontaneous breaking of $\mathcal{S}_n$ symmetry)

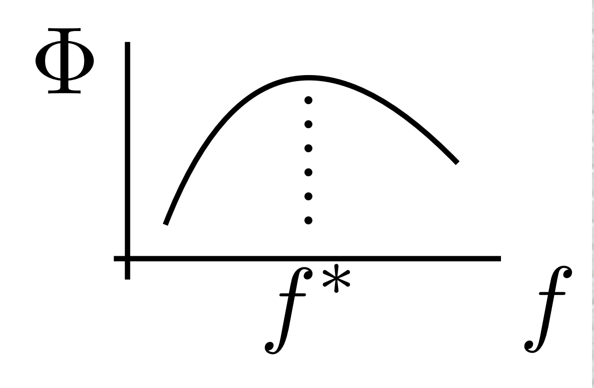
Interchange the $n \to 0$ and $N \to \infty$ limits

« *The Pandora box is open* » (G. Parisi)

Replicas, version 2: large deviations

Free energy of sample J : $\quad f_J = -\dfrac{1}{\beta N} \log Z_J$

Probability of finding a sample with $f_J = f$: $\quad e^{N\Phi(f)}$



Almost all samples have $f_J = f^*$

Reconstruct the large deviation function $\Phi(f)$ and find f^*

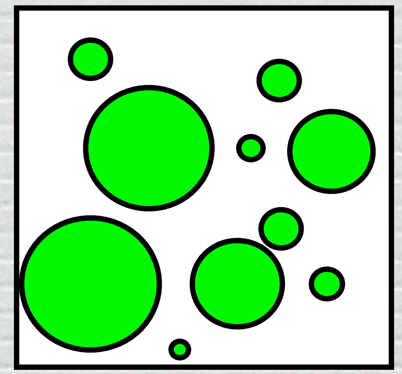
$$\mathcal{E}(Z_J^n) = \int df \; e^{N[-n\beta f + \Phi(f)]}$$

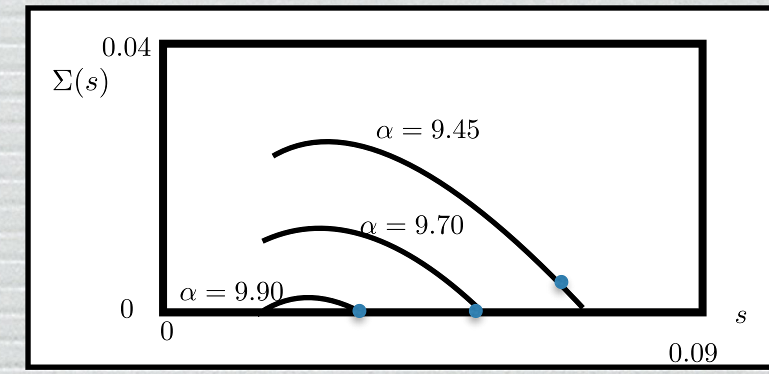
studied in the thermodynamic limit with the Laplace method

Replicas, version 3: metastable states

Glassy phases, even without disorder (eg structural grasses): proliferation of metastable states

eg K-satisfiability





$$Z_J \simeq \sum_\alpha Z_J^\alpha \qquad Z_J^\alpha = e^{-\beta N f_J^\alpha}$$

Complexity $\Sigma_J(f)$: $\sim e^{N\Sigma_J(f)}$ metastable states with $f_J^\alpha = f$

Introduce m replicas (or « clones ») constrained to be in the same states

$$Z_J^{[m]} = \sum_\alpha (Z_J^\alpha)^m = \int df \; e^{N[\Sigma_J(f) - m\beta f]}$$

Can then average over J , with $n \to 0$ replicas 1-step RSB

Replicas « philosophy »

Many pure states or metastable states, sample dependent.
Only the sample knows them.
Compare several « replicas » : configurations generated from the equilibrium measure; measure the distance between them, also in presence of couplings between them; count them (entropy, complexity).

Chapter Four

Algorithms

Analysis of one given sample: mean field

Historical development of mean field equations :

- In homogeneous ferromagnets:
  - Weiss (infinite range, 1907)
  - Bethe Peierls (finite connectivity, 1935)
- In glassy systems:
  - Thouless Anderson Palmer 1977 (infinite range)
  - M. Parisi Virasoro 1986 (infinite range)
  - M. Parisi 2001 (finite connectivity)
- As an algorithm:
  - Gallager 1963
  - Pearl 1986
  - Kabashima Saad 1998
  - M. Parisi Zecchina 2002
  - ...

Mean-Field 111 years ago

Paul Langevin (1905): $$\boxed{M = M_0\, L\left(\frac{B}{T}\right)}\ ; \qquad L(x) = \coth x - 1/x$$

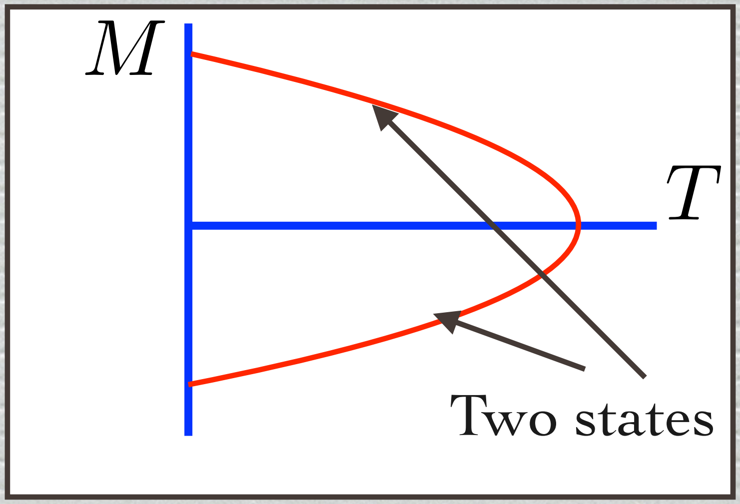
One spin in a magnetic field B

Pierre Weiss (1907): $$\boxed{B = B_{ext} + \alpha M}$$

One spin in a magnet: external field+ field from neighbors

Spontaneous magnetization in zero external field:

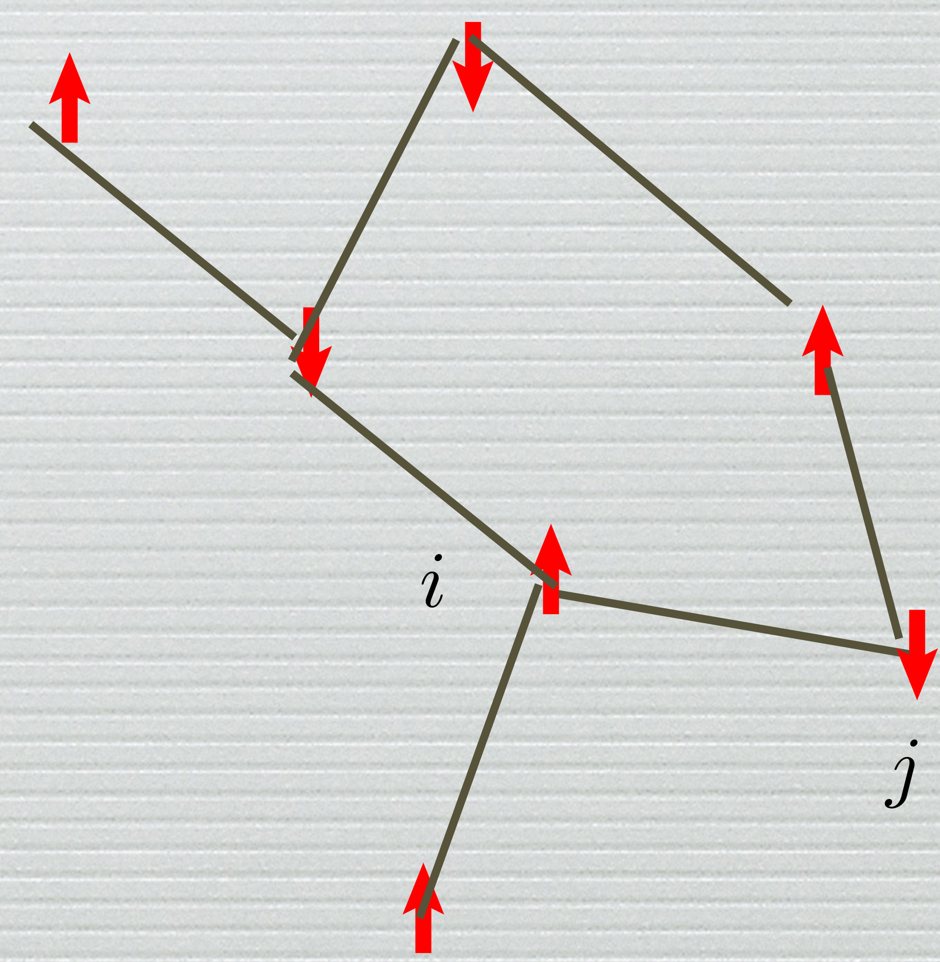
$$\boxed{M = M_0 L\left(\frac{\alpha M}{T}\right)}$$



Simple Mean-Field : Ising model

$$P(S) = \frac{1}{Z}\, e^{-E(S)/T} \qquad E(S) = -\sum_{ij} J_{ij} s_i s_j$$

$$\langle s_i \rangle \simeq \tanh(\beta \sum_j J_{ij} \langle s_j \rangle)$$



N coupled equations for
the local magnetizations $m_i = \langle s_i \rangle$

If homogeneous: $M \simeq \tanh(\beta z J M)$

Generally useless in disordered systems.
Neglects fluctuations. Correct formula:

$$\langle s_i \rangle = \langle \tanh(\beta \sum_j J_{ij} s_j) \rangle$$

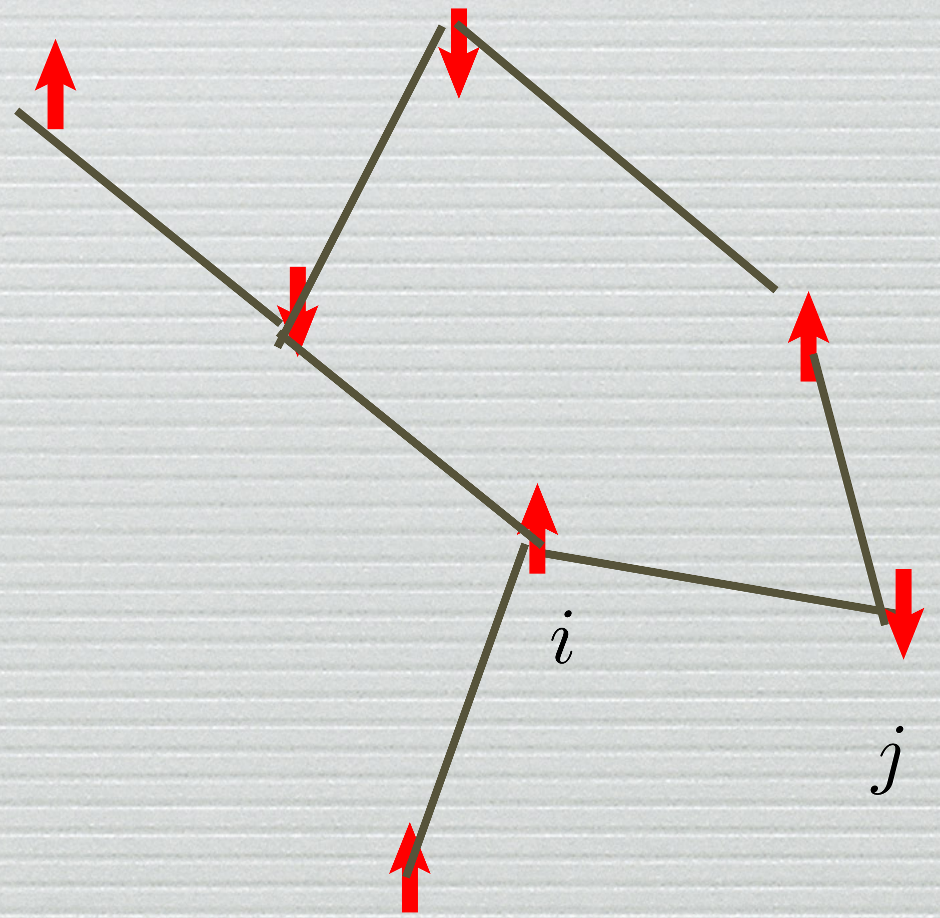
Does not close on $\langle s_i \rangle$

Mean-Field 83 years ago

Hans Bethe (1935)

Rudolf Peierls (1936)

Exact solution for central spin and its neighbors, themselves independent



Mean-Field 83 years ago

Hans Bethe (1935)

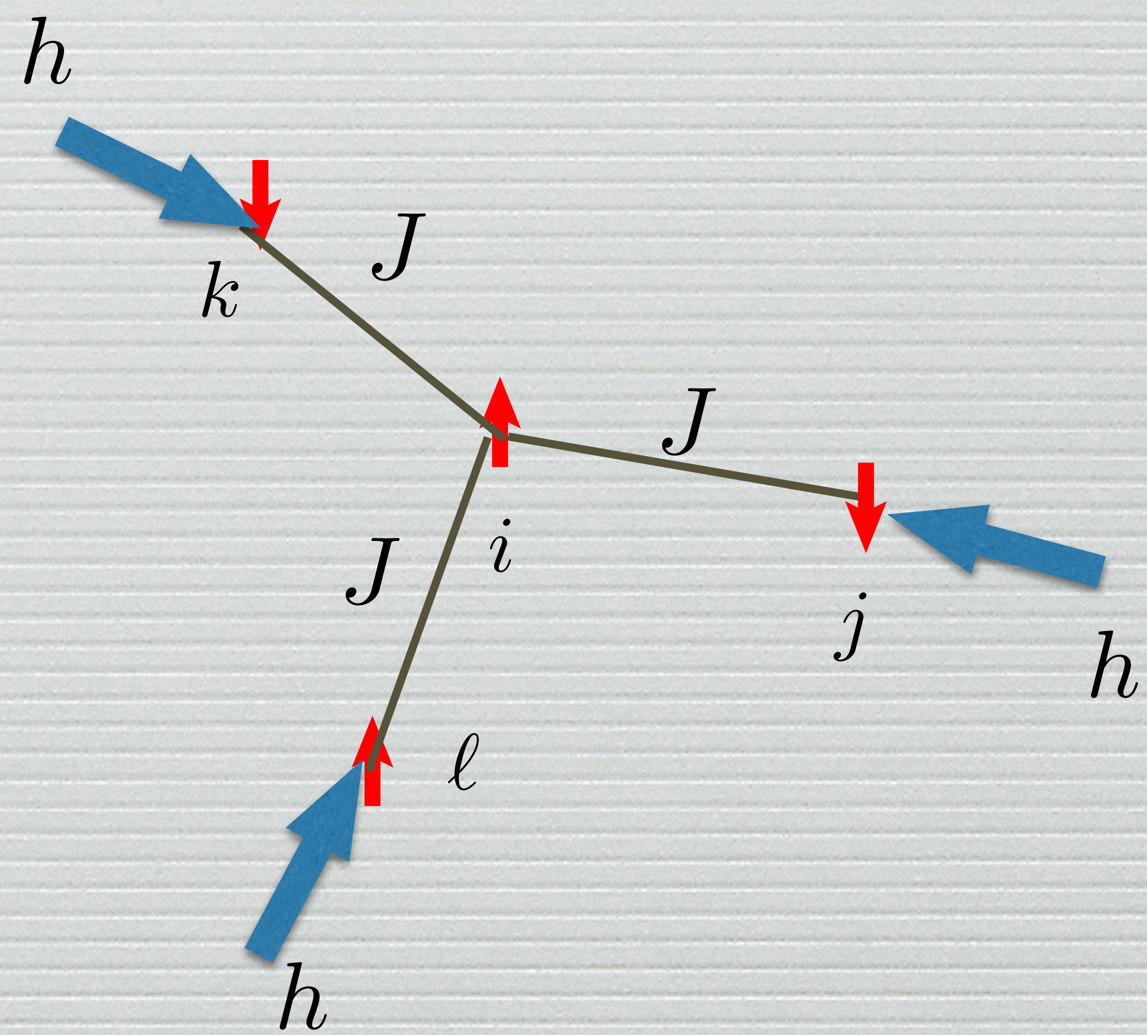
Rudolf Peierls (1936)

Exact solution for central spin and its neighbors, themselves independent

$$P(s_i, s_j, s_k, s_\ell) = \frac{1}{z}\, e^{\beta J s_i [s_j + s_k + s_l]}$$

$$e^{\beta h (s_j + s_k + s_\ell)}$$

$$\boxed{\begin{aligned} h &= \frac{z-1}{\beta} \operatorname{atanh}[\tanh(\beta J) \tanh(\beta h)] \\ M &= \tanh\left(z \operatorname{atanh}[\tanh(\beta J) \tanh(\beta h)]\right) \end{aligned}}$$



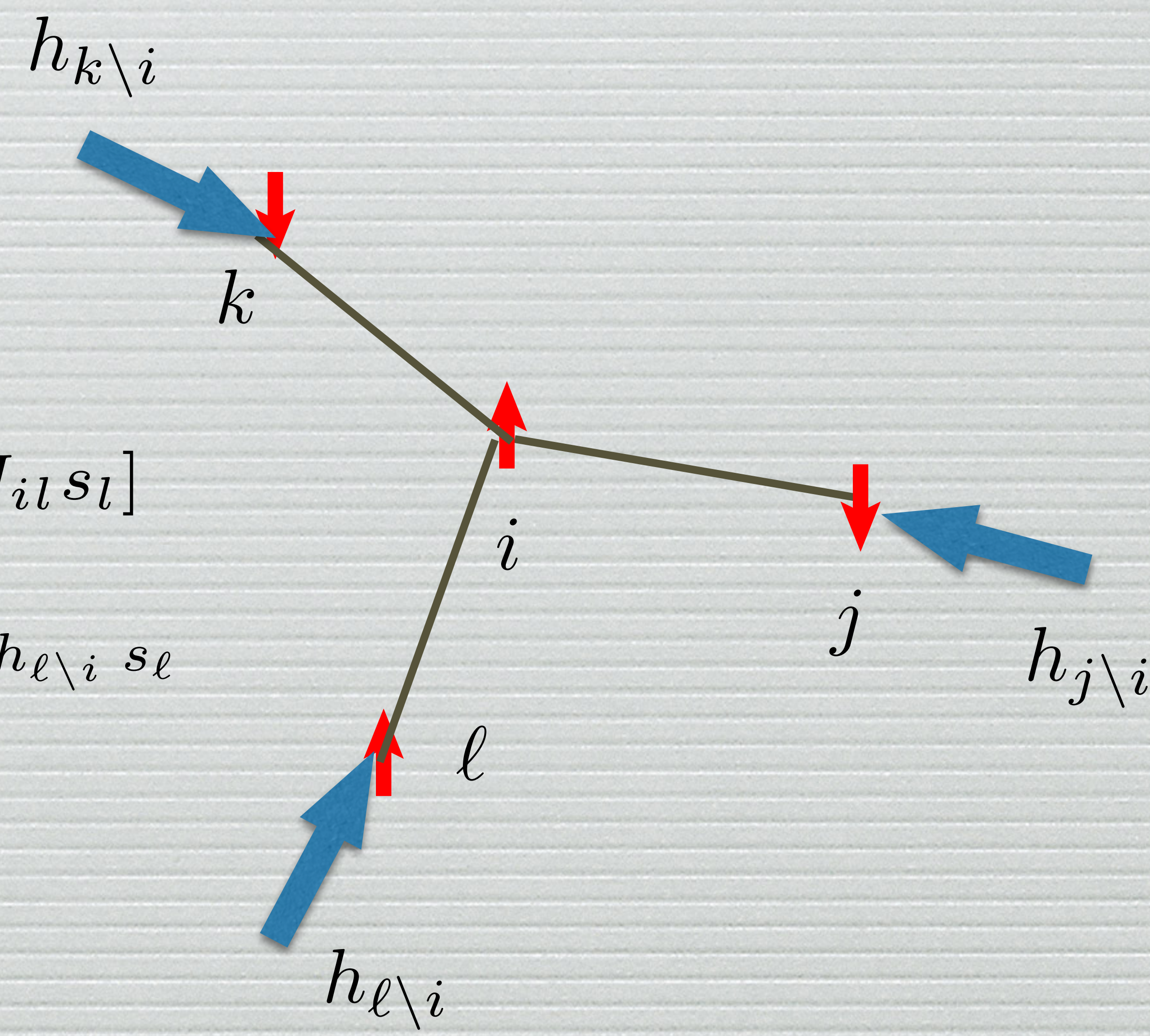
Bethe-Peierls adapted to disordered case

Exact solution for central spin and its neighbors, themselves independent

$$P(s_i, s_j, s_k, s_\ell) = \frac{1}{z}\, e^{\beta J s_i [s_j + s_k + s_l]}$$

$$P(s_i, s_j, s_k, s_\ell) = \frac{1}{z}\, e^{\beta s_i [J_{ij} s_j + J_{ik} s_k + J_{il} s_l]}$$

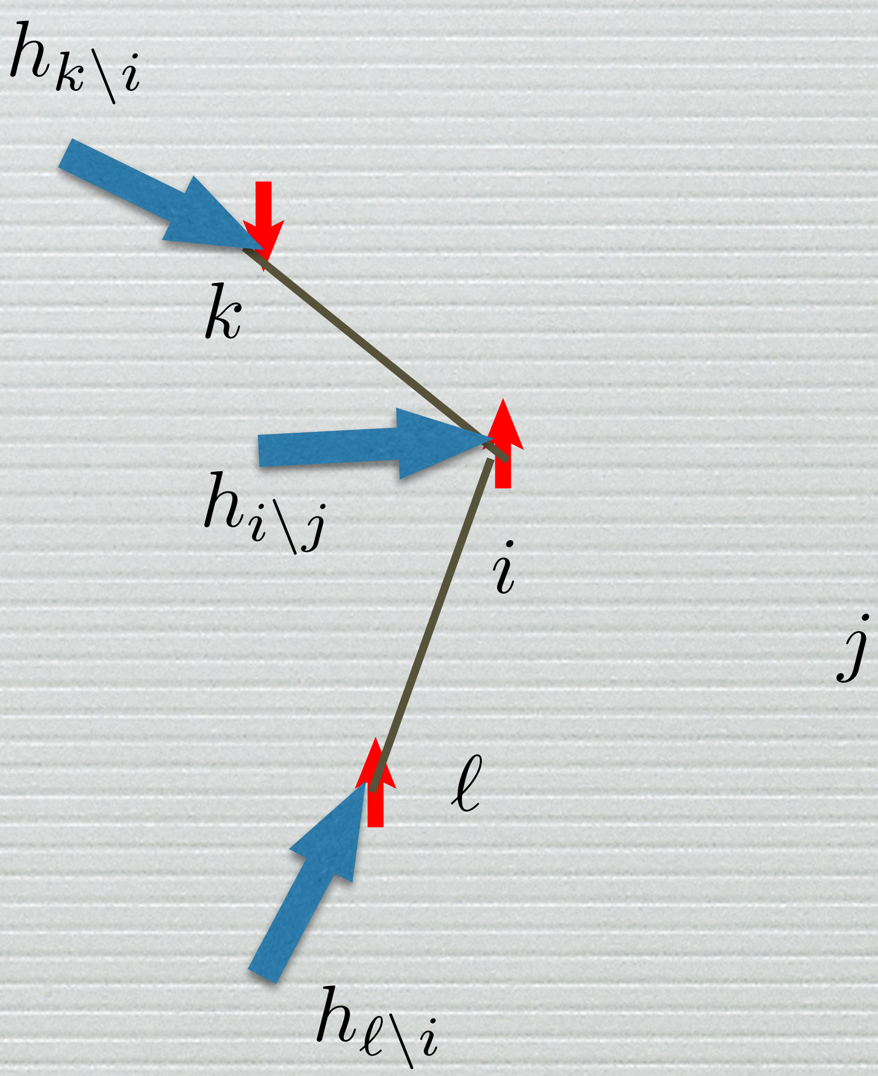
$$e^{\beta h_{j\backslash i}\, s_j}\; e^{\beta h_{k\backslash i}\, s_k}\; e^{\beta h_{\ell\backslash i}\, s_\ell}$$



Bethe-Peierls adapted to disordered case

$h_{i\setminus j} =$ Effective field on i due all of its neighbors in absence of j

$$h_{i\setminus j} = \frac{1}{\beta}\text{atanh}[\tanh(\beta J_{ki})\tanh(\beta h_{k\setminus i})] + \frac{1}{\beta}\text{atanh}[\tanh(\beta J_{\ell i})\tanh(\beta h_{\ell\setminus i})]$$



Bethe-Peierls
Belief Propagation algorithm

$h_{i\setminus j} =$ Effective field on i due all of
its neighbors in absence of j

$$h_{i\setminus j} = \frac{1}{\beta}\mathrm{atanh}[\tanh(\beta J_{ki})\tanh(\beta h_{k\setminus i})]$$
$$+\frac{1}{\beta}\mathrm{atanh}[\tanh(\beta J_{\ell i})\tanh(\beta h_{\ell\setminus i})]$$

N_{edge} coupled equations for the cavity fields

« **BP**» **algorithm**: **iterate** these equations

$$h^{t+1}_{i\setminus j} = f(h^t_{k\setminus i}, h^t_{\ell\setminus i})$$

Generalizable to any constraint satisfaction problem:

$$P(S) = \frac{1}{Z}\prod_a \psi_a(S_{\partial a})$$

A remark: the cavity method

$h_{i\setminus j} =$ Effective field on i due all of its neighbors in absence of j

BP: $\boxed{h_{i\setminus j}^{t+1} = f(h_{k\setminus i}^t, h_{\ell\setminus i}^t)}$

Cavity: statistical analysis of the fixed point. All the messages in the rhs are iid from $P(h)$. The BP equation then leads to a self consistent functional equation for $P(h)$. Sometimes solved by moments (large connectivity), or by population dynamics. Replicas

Cavity seeks a fixed point distribution of $P^{t+1}(h) = F[P^t(h)]$

State evolution does not focus only on fixed-point. It follows the mapping at each iteration generated by the BP iteration. Analytic control of algorithm.

Validity of Mean-field

1) When is simple mean-field exact?

$$\langle s_i \rangle \simeq \tanh(\beta \sum_j J_{ij} \langle s_j \rangle)$$

Ferromagnet with long-range interactions: $J_{ij} = J/N$ (Curie-Weiss)

Fluctuations of $\sum_j J_{ij} s_j$ can be neglected

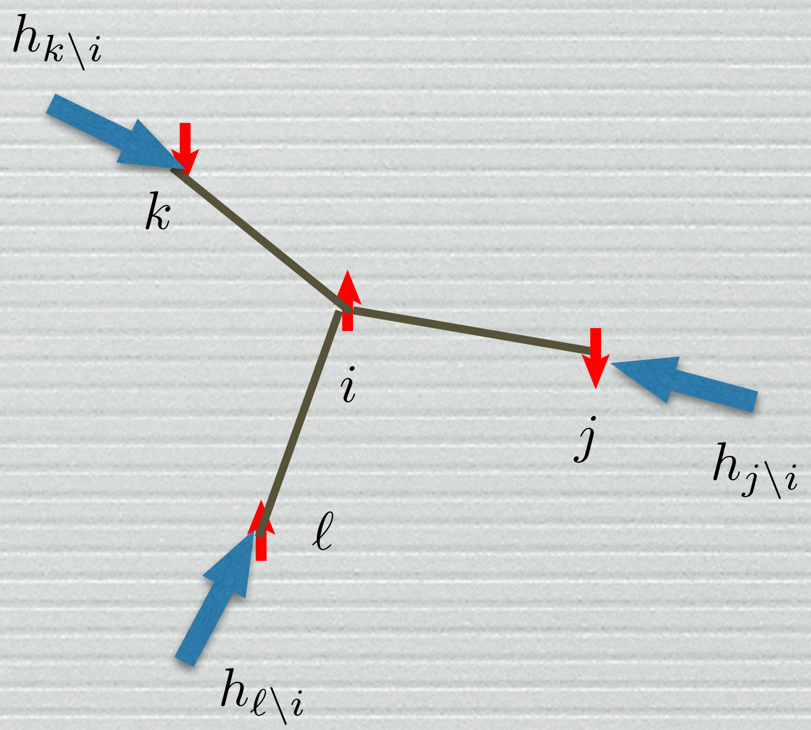
Validity of Mean-field

2) When is BP exact?

$$h_{i\backslash j}^{t+1} = f(h_{k\backslash i}^{t}, h_{\ell\backslash i}^{t})$$

Fluctuations are handled correctly, but beware of correlations

- Exact in one dimension (transfer matrix)
- Exact on a tree (uncorrelated b.c)
- Exact on locally tree-like graphs (Erdös Renyi etc.) **if** correlations decay fast enough (single pure state)
- Exact in infinite range problems (SK) **if** correlations decay fast enough (single pure state)



Validity of Mean-field

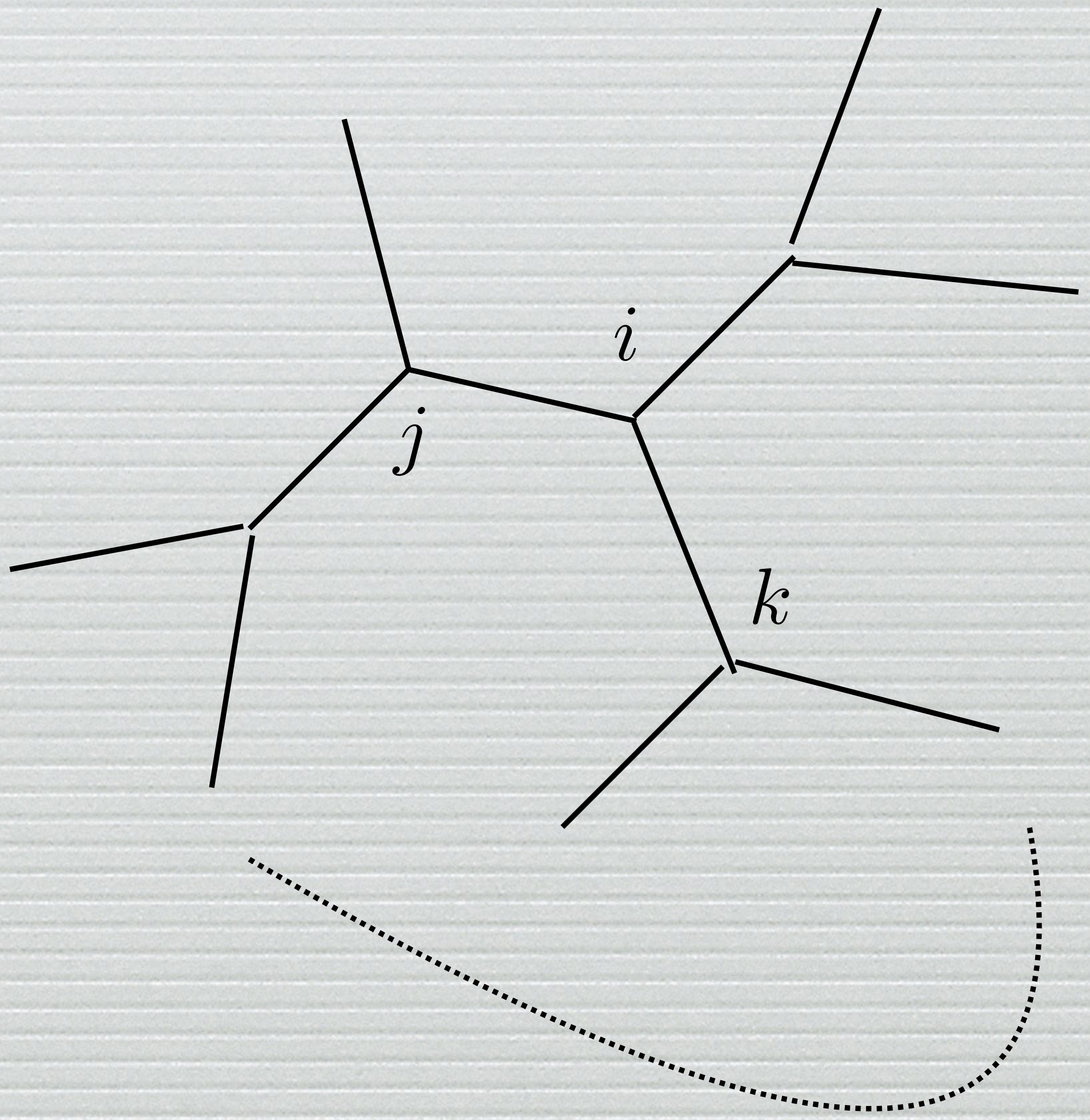
2) When is BP exact?

$$h^{t+1}_{i\backslash j} = f(h^t_{k\backslash i}, h^t_{\ell\backslash i})$$

Typically, j and k are far apart
in absence of i

If correlations decay fast enough
BP is exact asymptotically

Away from phase transitions
Within one pure state



Loop length $O(\log N)$

Three important developments

1) The special case of infinite-range models (TAP 1976, cavity method 1987)

2) What happens if the elementary variables (spins) are real instead of discrete ?

3) What happens in a glass phase, when there are many pure states, and therefore many solutions ?

1)The special case of infinite range models

SK model $J_{ij} = O\left(\frac{1}{\sqrt{N}}\right)$

Correlations can be neglected (in the glass phase : within one pure state)

$$h_{i\setminus j} = \frac{1}{\beta}\sum_{k(\neq i)} \text{atanh}[\tanh(\beta J_{ki})\tanh(\beta h_{k\setminus i})] \simeq \sum_{k(\neq i)} J_{ki}\tanh(\beta h_{k\setminus i})$$

$$H_i = \frac{1}{\beta}\sum_{k} \text{atanh}[\tanh(\beta J_{ki})\tanh(\beta h_{k\setminus i})] \simeq \sum_{k} J_{ki}\tanh(\beta h_{k\setminus i})$$

$$h_{i\setminus j} \simeq H_i - O\left(\frac{1}{\sqrt{N}}\right)$$

Corrections can be handled to first order in perturbation theory, and all the equations close on the N variables H_i $\rightarrow$ TAP equations (AMP)

$$\overset{t+1}{H_i} = \sum_k J_{ki}\tanh(\beta \overset{t}{H_k}) - \beta\tanh(\beta \overset{t-1}{H_i})\sum_k J_{ki}^2[1-\tanh^2(\beta \overset{t-1}{H_k})]$$

Time iteration (Bolthausen): AMP algorithm in information theory

Three important developments

1) The special case of infinite-range models (cavity method 1987)

2) What happens if the elementary variables (spins) are real instead of discrete ?

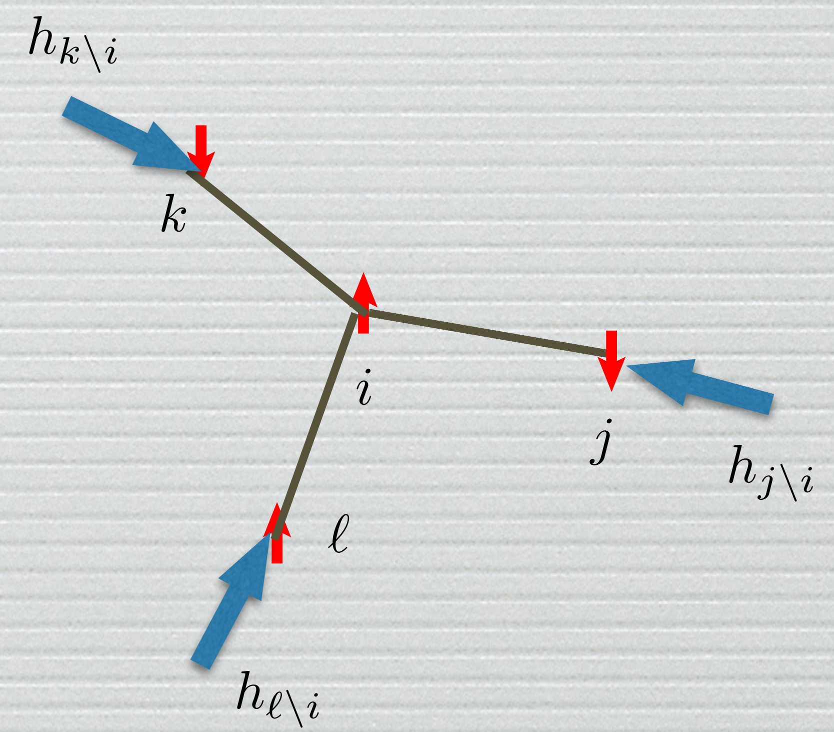
3) What happens in a glass phase, when there are many pure states, and therefore many solutions ?

Real variables

$$h_{i\backslash j}^{t+1} = f(h_{k\backslash i}^t, h_{\ell\backslash i}^t)$$

becomes

$$p_{i\backslash j}(x_i) = F[p_{k\backslash i}(x_k), p_{\ell\backslash i}(x_\ell)]$$



BP messages are cavity probability densities of the local variables.

Simple case : large connectivity $p_{i\backslash j}(x_i)$ approximately Gaussian

Generalized Approximate Message Passing (GAMP).

MM1989: cavity. Rangan 2010 : algorithm,…

Three important developments

1) The special case of infinite-range models (cavity method 1987)

2) What happens if the elementary variables (spins) are real instead of discrete ?

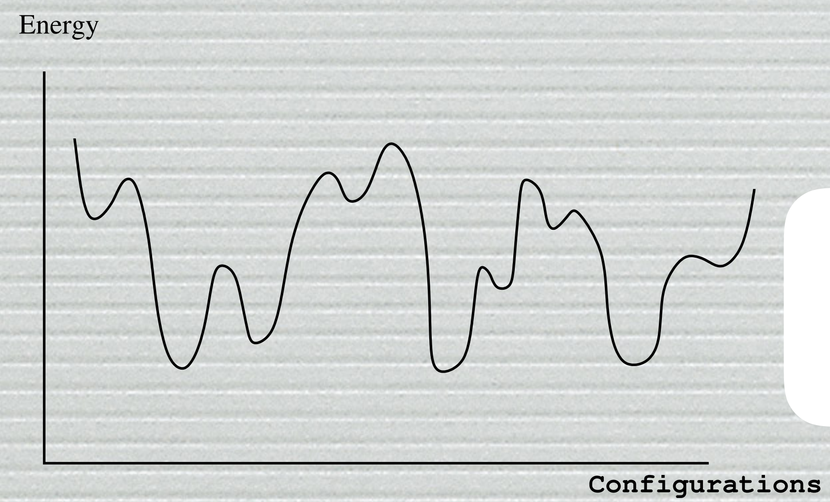
3) What happens in a glass phase, when there are many pure states, and therefore many solutions ?

3) What happens in a glass phase, when there are many pure states, and therefore many solutions ?

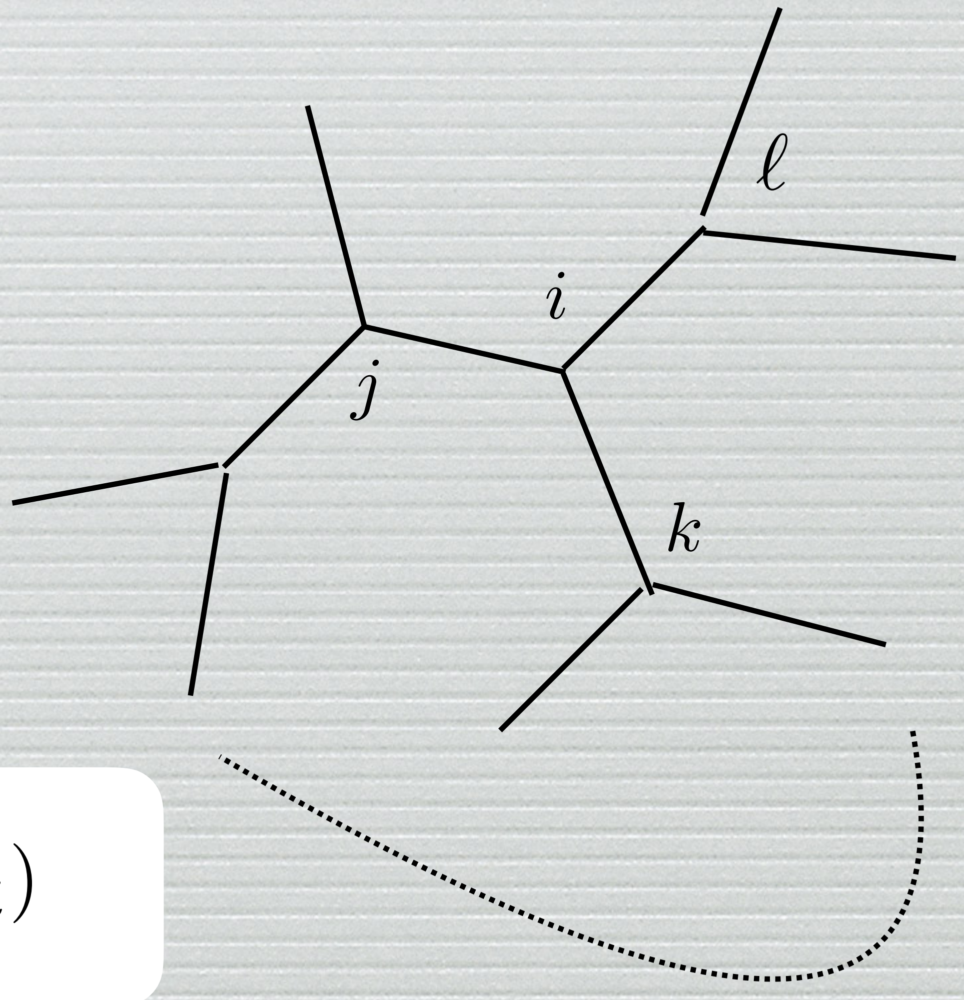
BP equations

$$h_{i\setminus j} = f(h_{k\setminus i}, h_{\ell\setminus i})$$

Correct if, in absence of the i-j interaction, the correlations between k and ℓ can be neglected.



$$h^{\alpha}_{i\setminus j} = f(h^{\alpha}_{k\setminus i}, h^{\alpha}_{\ell\setminus i})$$



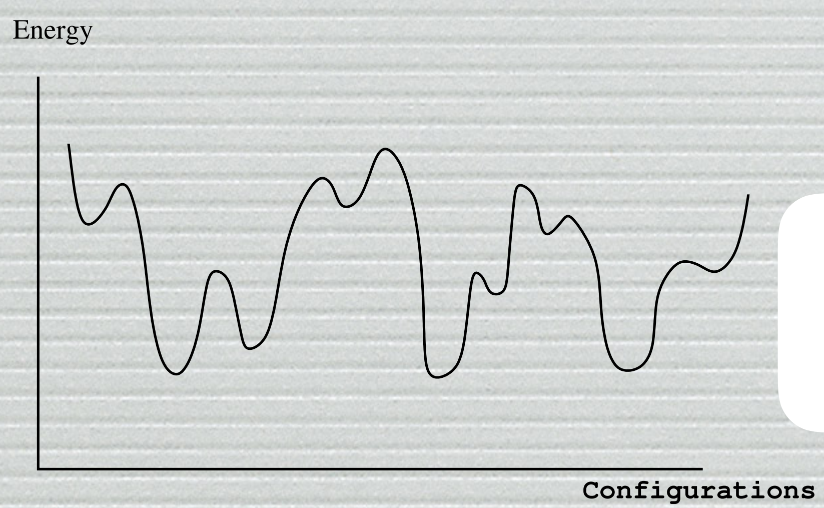
Loop length $O(\log N)$

Glassy phase: many states, many solutions of BP

3) What happens in a glass phase, when there are many pure states, and therefore many solutions ?

BP equations $$h_{i\setminus j} = f(h_{k\setminus i}, h_{\ell\setminus i})$$

Correct if, in absence of the i-j interaction, the correlations between k and ℓ can be neglected.



$$h^{\alpha}_{i\setminus j} = f(h^{\alpha}_{k\setminus i}, h^{\alpha}_{\ell\setminus i})$$

Glassy phase: many states, many solutions of BP

Statistics of $h^{\alpha}_{i\setminus j}$ over the many states α

$P_{i\setminus j}(h)$

related to $P_{k\setminus i}(h)$

$P_{\ell\setminus i}(h)$

Survey propagation

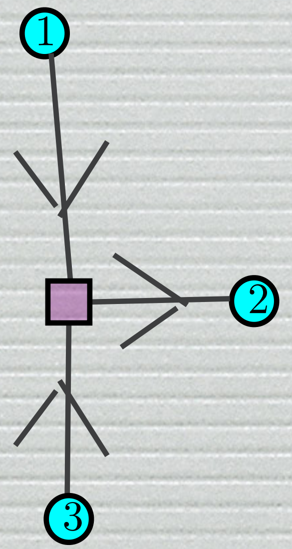
MM Parisi Zecchina

2002

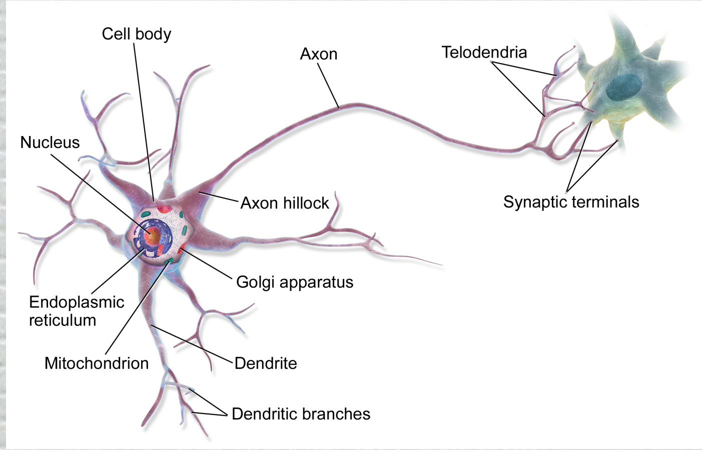
Power of message passing algorithms

Approximate solution of very hard, and very large constraint satisfaction problems, ...FAST! (typically linear time)

- BP: Best decoders for LDPC error correcting codes
- SP: Best solver of random satisfiability problems
- BP: Best algorithm for learning patterns in neural networks (e.g. binary perceptron)
- Data clustering, graph coloring, Steiner trees, etc…
- Fully connected networks : TAP (=AMP). Compressed sensing, linear estimation, near ground-state of SK model (with overlap annealing)



Local, simple update equations:
Each message is updated using information from incoming messages on the same node.
Distributed, solves hard global pb



Chapter Five

Inference

Inference

Infer a hidden rule, or hidden variables, from data.

Restricted sense : find parameters of a probability distribution

Bayesian inference

Unknown parameters	x	Prior	$P(x)$
Measurements	y	Likelihood	$P(y\|x)$

Posterior $$P(\boxed{x}|y) = \frac{P(y|\boxed{x})P(\boxed{x})}{P(y)}$$

Bayesian inference with many unknown and many measurements

Unknown parameters $x = (x_1, \ldots, x_N)$ Large M, N

Measurements $y = (y_1, \ldots, y_M)$ $\alpha = M/N$

Often (but not necessarily):

Independent measurements $$P(y|x) = \prod_\mu P_\mu(y_\mu|x)$$

Factorized prior $$P^0(x) = \prod_i P_i^0(x_i)$$

Posterior $$P(x) = \frac{1}{Z(y)} \left(\prod_i P_i^0(x_i) \right) \exp\left[- \sum_\mu E_\mu(x, y_\mu) \right]$$

$$E_\mu(x, y_\mu) = -\log P_\mu(y_\mu|x)$$

→ **Algorithms** → **Prediction on the quality of inference**

Bayesian inference with many unknown and many measurements

$$P(x) = \frac{1}{Z(y)} \left(\prod_i P_i^0(x_i) \right) \exp\left[- \sum_\mu E_\mu(x, y_\mu) \right]$$

$$E_\mu(x, y_\mu) = -\log P_\mu(y_\mu | x)$$

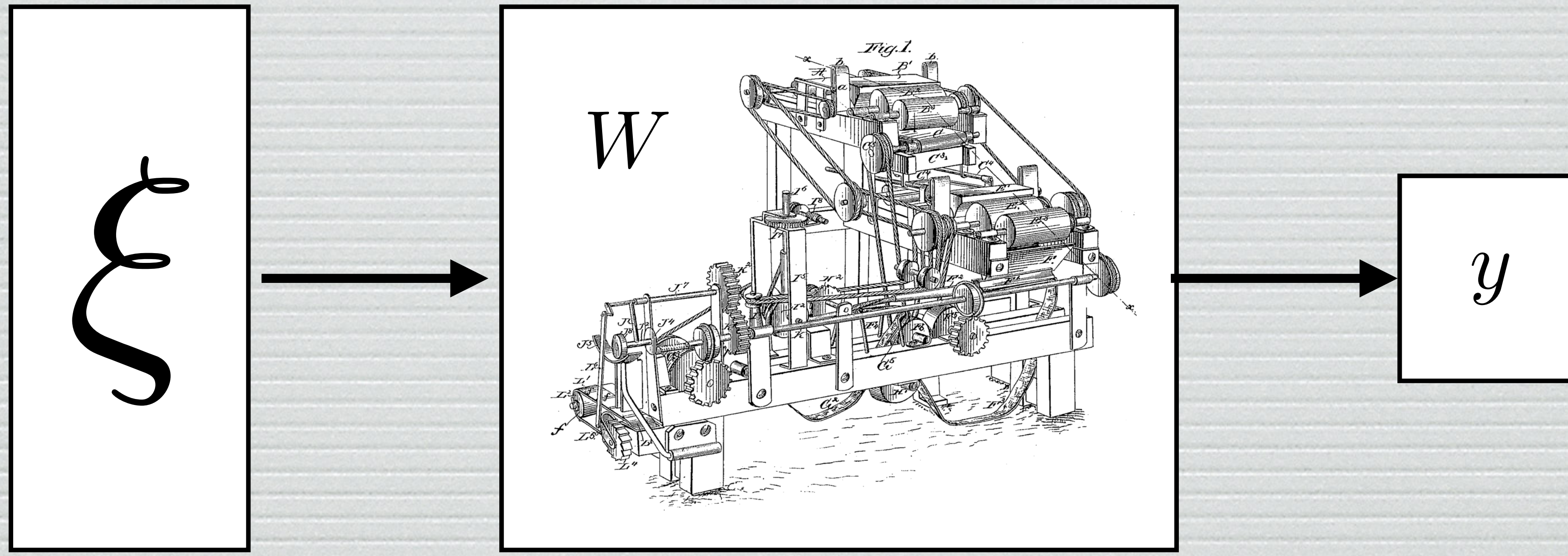
Statistical mechanics. Disordered system

✦Discrete or continuous variables x_i

✦Interactions through $e^{-E_\mu(x,y_\mu)}$ can be

- short-range
- long (or infinite) range

Machine learning

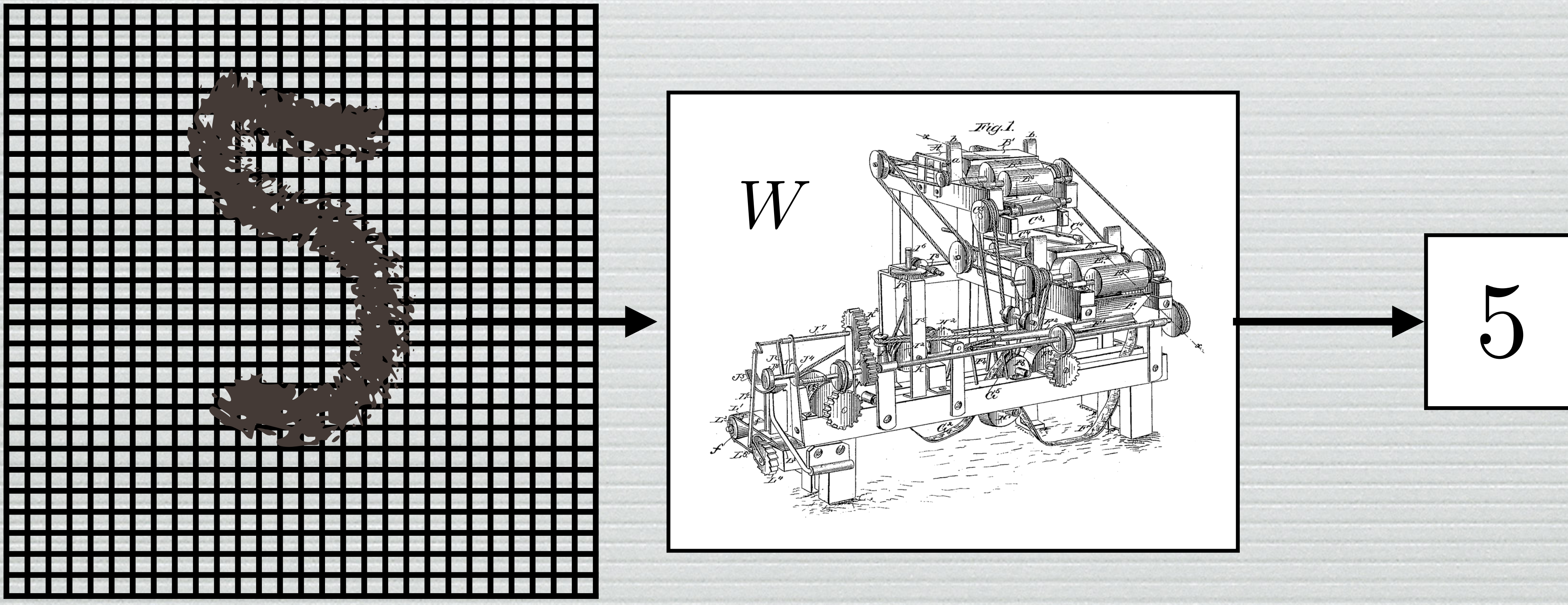


Input ξ

Machine, parameters W

Output y

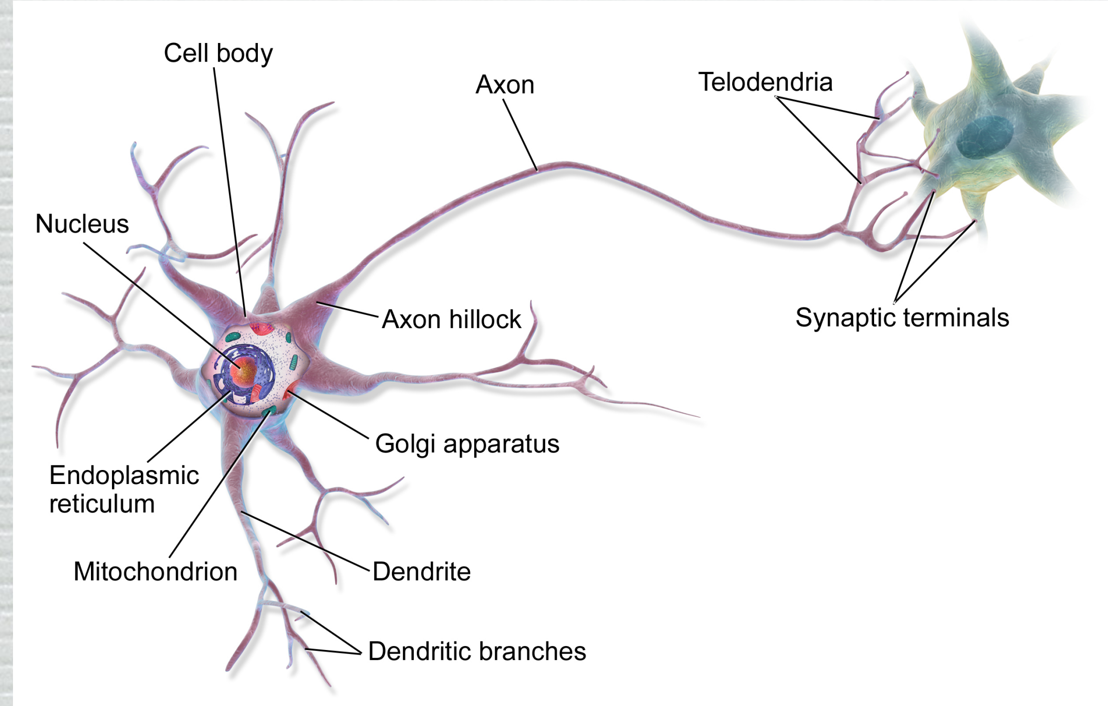
Machine learning

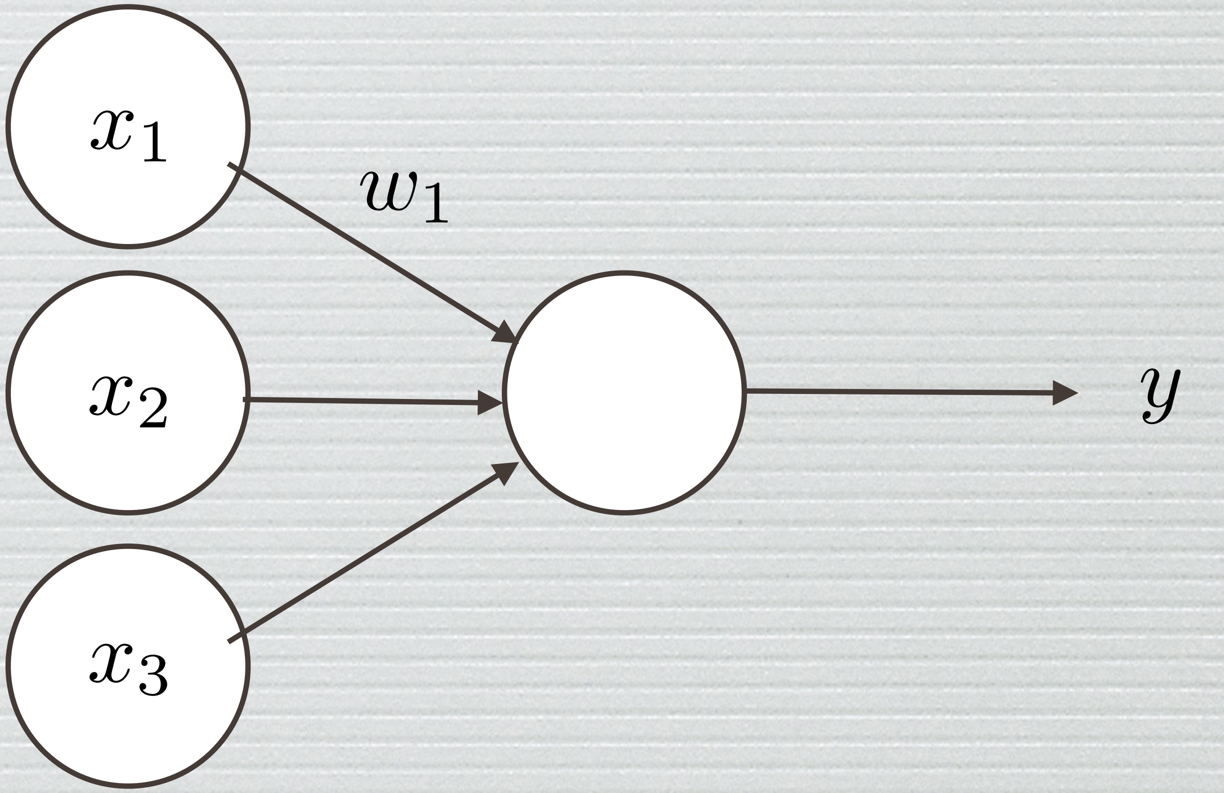


Handwritten digit, 28^2 pixels

Machine, parameters W

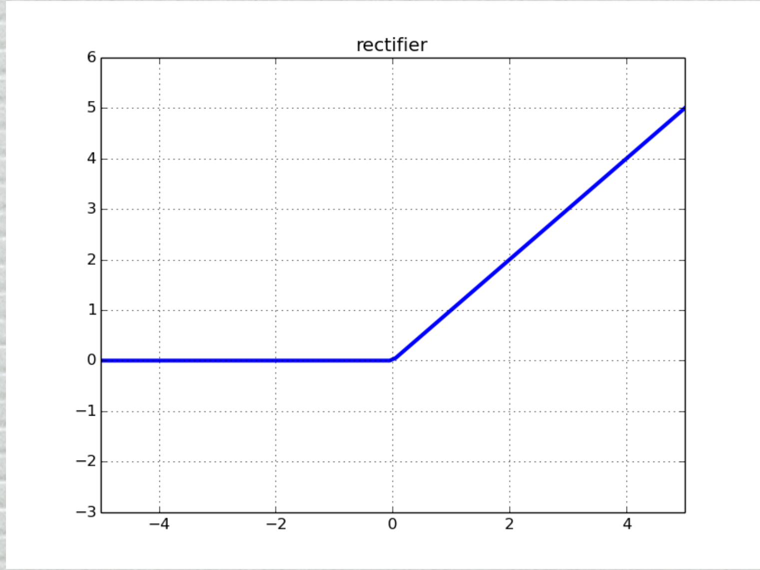
Output the number

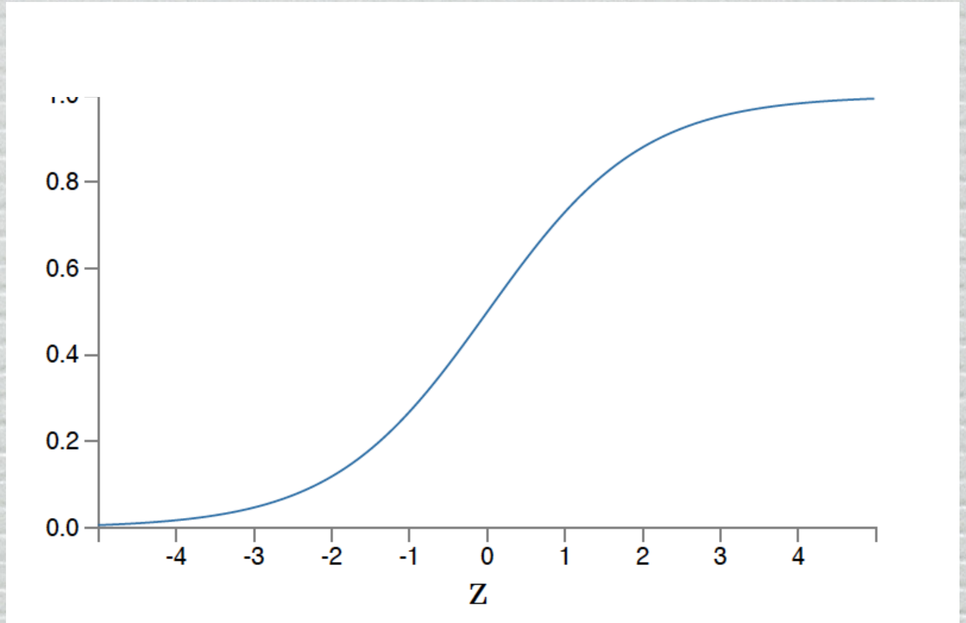




$$\boxed{y = f(w_0 + w_1x_1 + w_2x_2 + w_3x_3)}$$

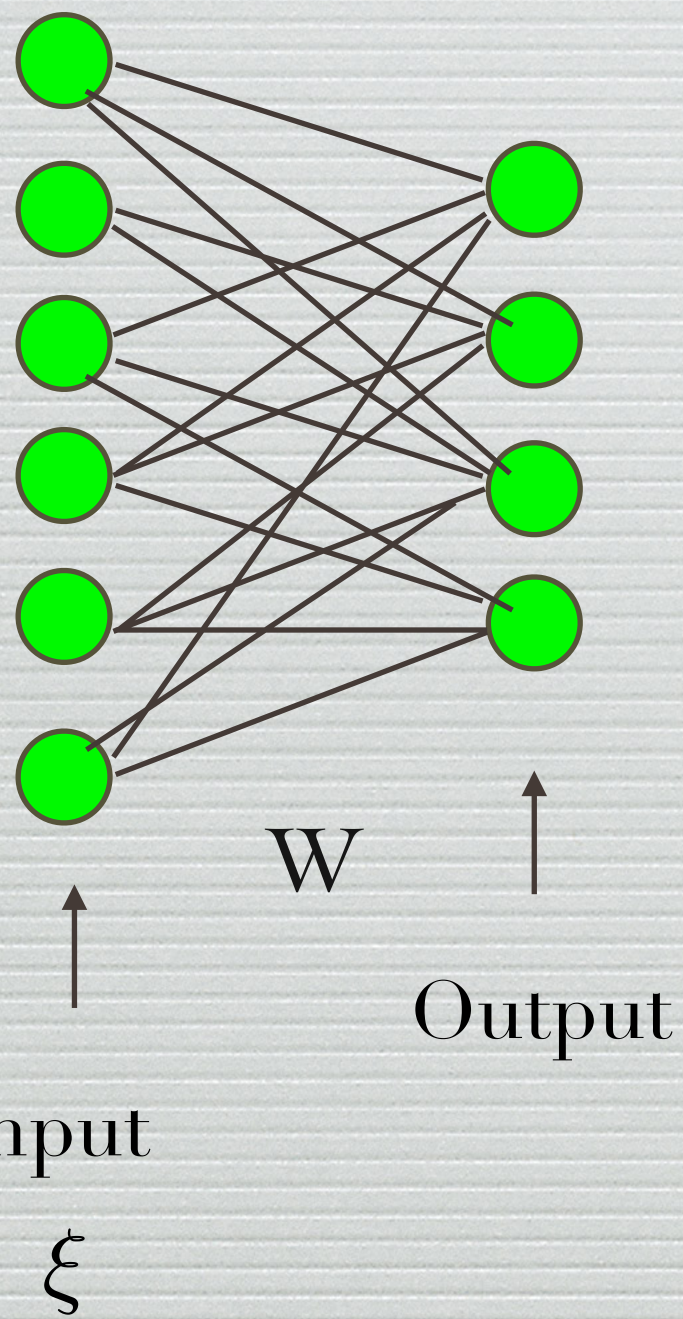
Formal neural network

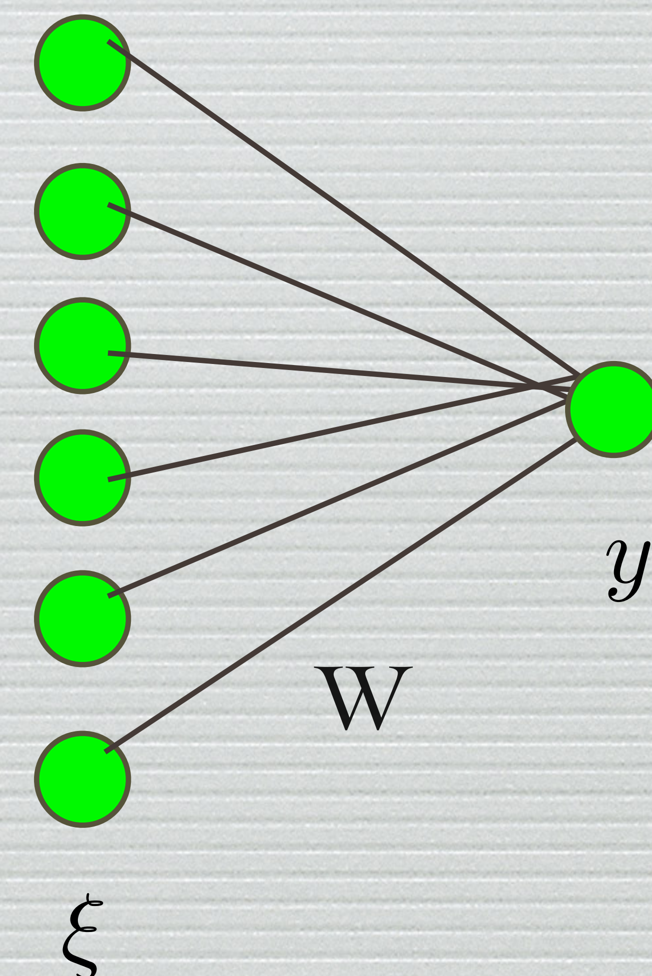




Simple perceptron

Decouples into independent single output machines

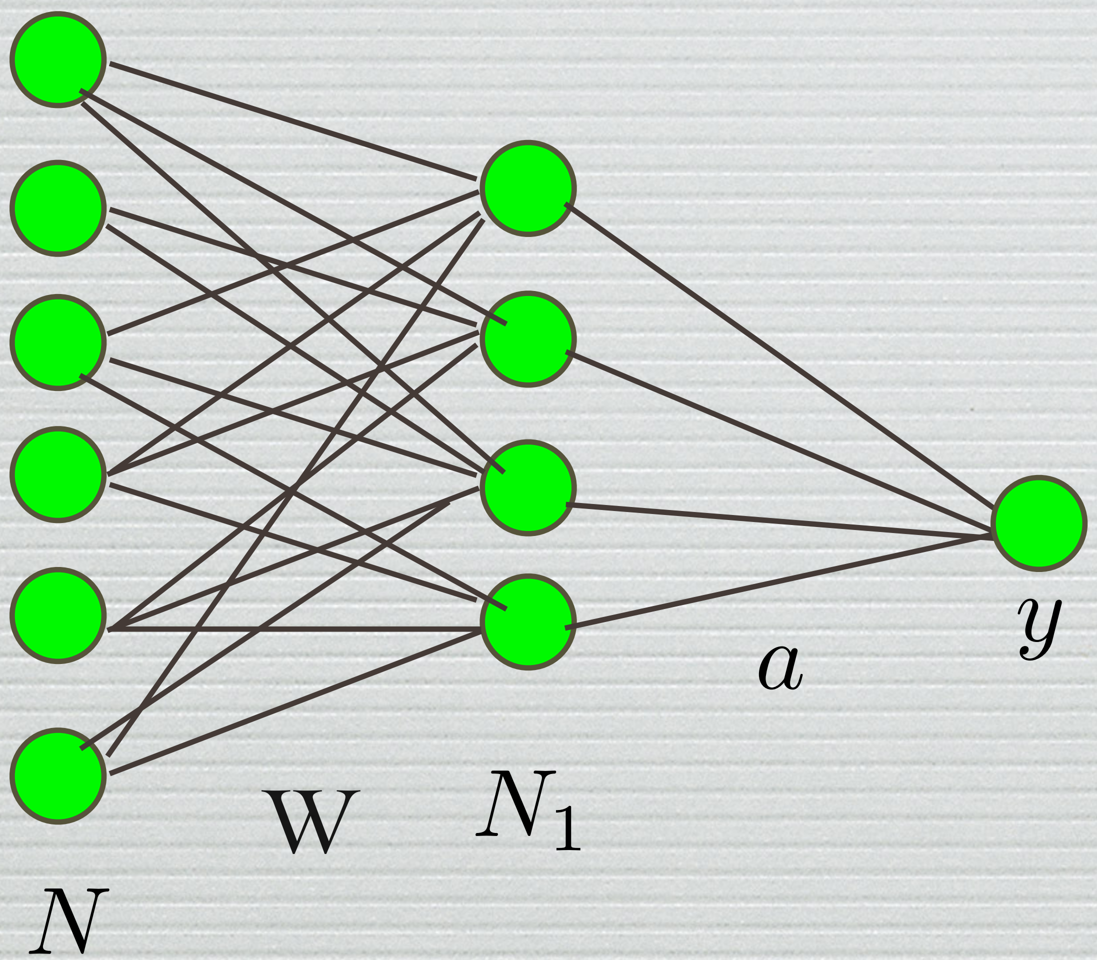




$$y = f(W.\xi)$$

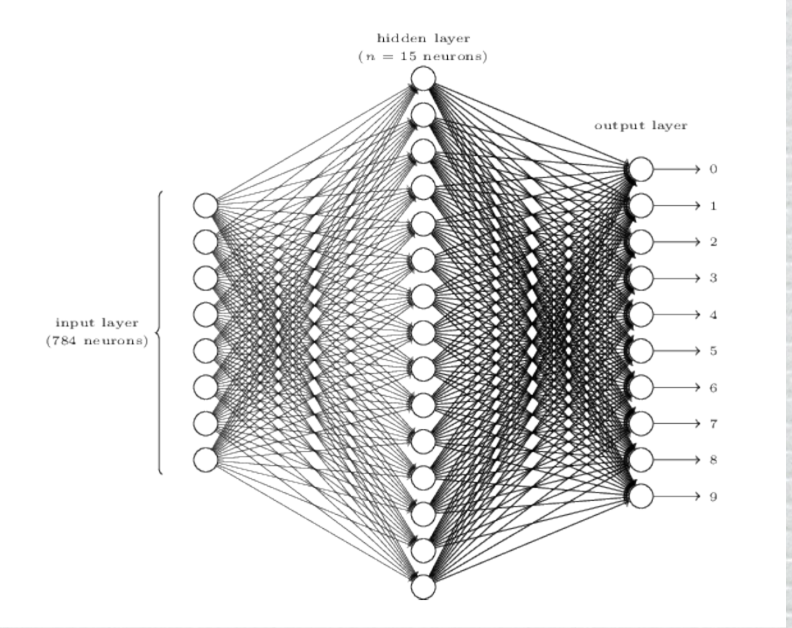
Limited to linearly separable rules

Example of a machine: two-layers feedforward neural network



Support Vector Machines: $$y = \sum_{i=1}^{N_1} a_i f_i(W_i.\xi)$$

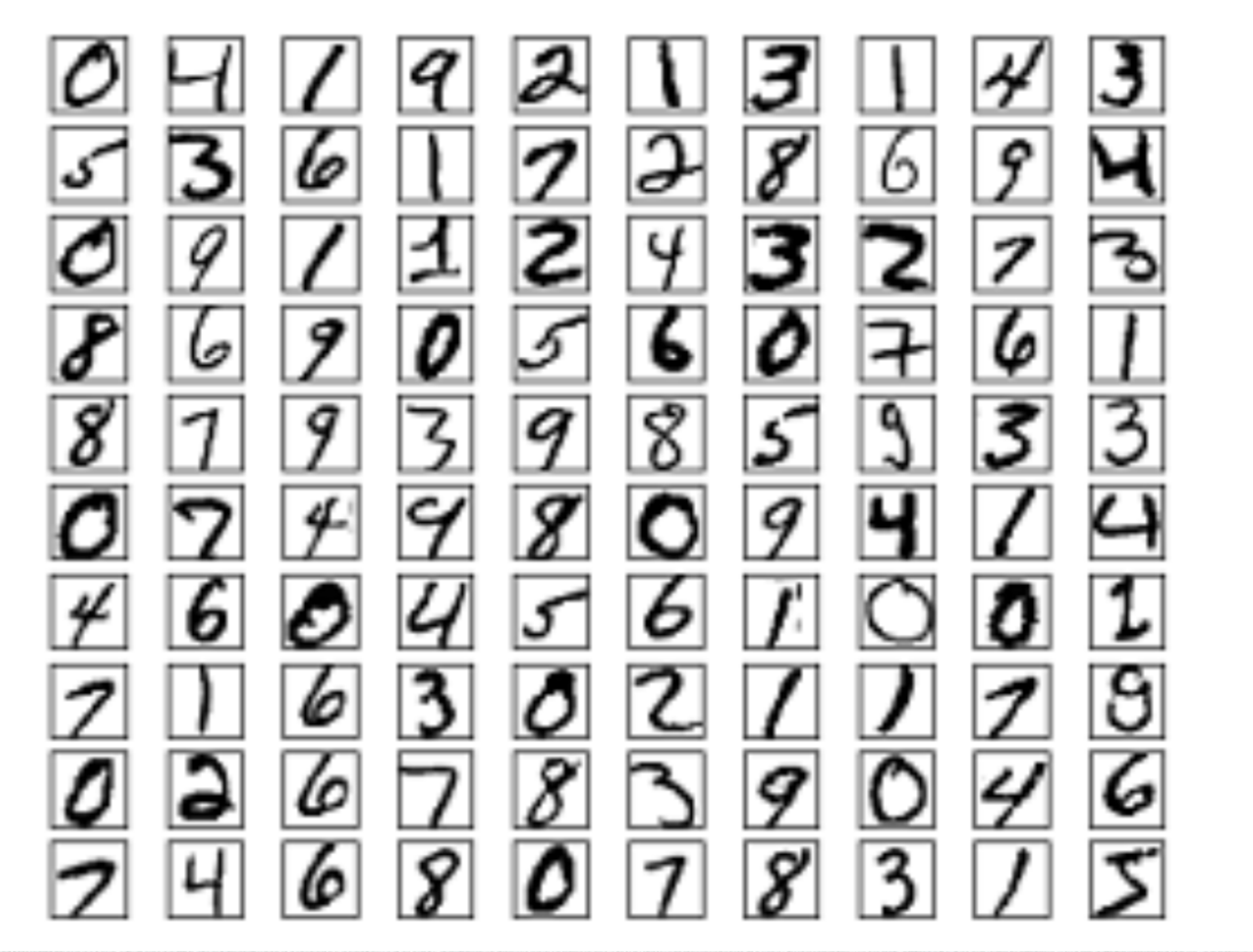
Example of a machine: two-layers feedforward neural network for digits recognition



Neurons: 784 15 10

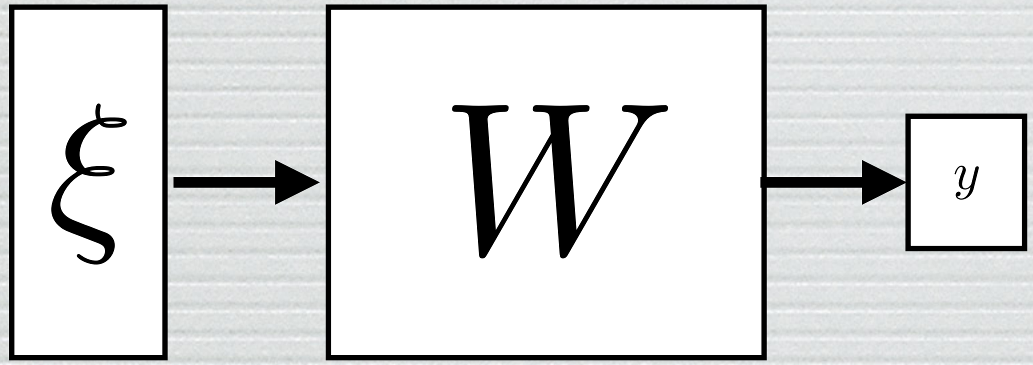
W = all synaptic weights and thresholds: 11925 parameters

Fixed through the study of many examples



MNIST database : 70,000 images of digits, segmented, 28×28 pixels each, greyscale

Machine learning: training



$$y = f(W, \xi)$$

Database = M examples of input-output (ξ_μ, y_μ)

Training = find a set of parameters W such that the machines perform well on the training set

Minimize a training error, e.g. $E_t = \sum_\mu \left[y_\mu - f(W, \xi_\mu) \right]^2$

NB: output could be noisy: $P(y) \propto e^{-E_t/(2\Delta^2)}$

Machine learning: training and generalization

$$\boxed{\xi} \longrightarrow \boxed{W} \longrightarrow \boxed{y}$$

$$y = f(W, \xi)$$

Database = M examples of input-outpu…)

Bayesian learning:

$$P(W|\{\xi_\mu, y_\mu\}) = \frac{1}{Z} P^0 \ldots \mathrm{exp}\left(-\beta \sum_\mu [f(W, \xi_\mu) - y_\mu)]^2\right)$$

Unkno… …ata Prior

Other big « prior »: architecture!

Generalization: having found the best (a « typical ») set of parameters W^*, compute the performance of the machine on some **new data**

$$E_g = \sum_\nu [y_\nu - f(W^*, \xi_\nu)]^2$$

Machine learning: training and generalization

Learning: $$P(W|\{\xi_\mu, y_\mu\}) = \frac{1}{Z} P^0(W) \; \exp\left(-\beta \sum_\mu [f(W,\xi_\mu) - y_\mu)]^2\right)$$

Generalization: $$E_g = \sum_\nu \, [y_\nu - f(W^*, \xi_\nu)]^2$$

Two main issues:
- Algorithmic
- Theoretical

Algorithm: optimization in a large dimensional space, with a disordered « energy function », a priori « glassy ».
Landscape issues!

Theory: Large size OK. But needs a **model of data**. Ideally a generative model, or a smart description of the type of data.
Also very useful for algorithm design and analysis. Ensemble.

Model of data: ensemble

Learning:

$$P(W|\{\xi_\mu, y_\mu\}) = \frac{1}{Z} P^0(W) \ \exp\left(-\beta \sum_\mu [f(W, \xi_\mu) - y_\mu)]^2 \right)$$

Algorithmic studies typically uses one (or several) databases for $\{\xi_\mu, y_\mu\}$: data = quenched disorder

Theoretical analysis usually relies on a generative model of data (« **model of the world** »)

Examples from the 80's: iid patterns

Challenge: Find good generative models of the world

Generative model of data : teacher-student

An important case for theoretical studies of machine learning: **teacher-student.**

Data generated by a teacher. The teacher has his own set of parameters $W = T$

Given an input ξ_μ, the output is $y_\mu = f(T, \xi_\mu)$

If the student knows the architecture of the teacher, and uses the same, he needs to find his own parameters by minimizing the training error:

$$E_t = \sum_\mu \left[f(W, \xi_\mu) - f(T, \xi_\mu) \right]^2$$

Generative model: generate ξ_μ from some input data distribution, generate T from some distribution $P^T(T)$

Generative model of data : teacher-student

Teacher: generates parameters w^* from teacher prior $P^T(w)$

generates data y from teacher prior $P^T(y|w^*)$

Smart student : **knows the teacher's architecture and the generative distribution**. $P^S(W) = P^T(W)$

Bayes optimal: student's prior = teacher's prior

Student seeks a special « planted » configuration w^*

with zero training error : a « **crystal** »

Chapter Six

Correlations

The problem of correlations in the ensemble (the world)

Mean field equations (BP, TAP, AMP) with correlated disorder ?

$$h_{i\backslash j}^{t+1} = f(h_{k\backslash i}^{t}, h_{\ell\backslash i}^{t})$$

$$H_i = \sum_k J_{ki} \tanh(\beta H_k) - \beta \tanh(\beta H_i) \sum_k J_{ki}^2 [1 - \tanh^2(\beta H_k)]$$

Correct only if local quenched disordered variables J_{ki} are **independent**

Beyond independent variables: rotationally invariant disorder

$$J = O^T \, D \, O$$

when O is chosen uniformly in O(N) and D has a limiting distribution of eigenvalues: Parisi Potters 1995, Shinzato Kabashima 2008, Rangan Schniter Fletcher 2016,…

« Usual » TAP equations

$$H_i = \sum_k J_{ki} \tanh(\beta H_k) - \beta \tanh(\beta H_i) \sum_k J_{ki}^2 [1 - \tanh^2(\beta H_k)]$$

must be modified to

$$H_i = \sum_k J_{ki} \tanh(\beta H_k) - \beta \tanh(\beta H_i) G'(1-q)$$

$$q = (1/N) \sum_i \tanh^2(\beta H_i)$$

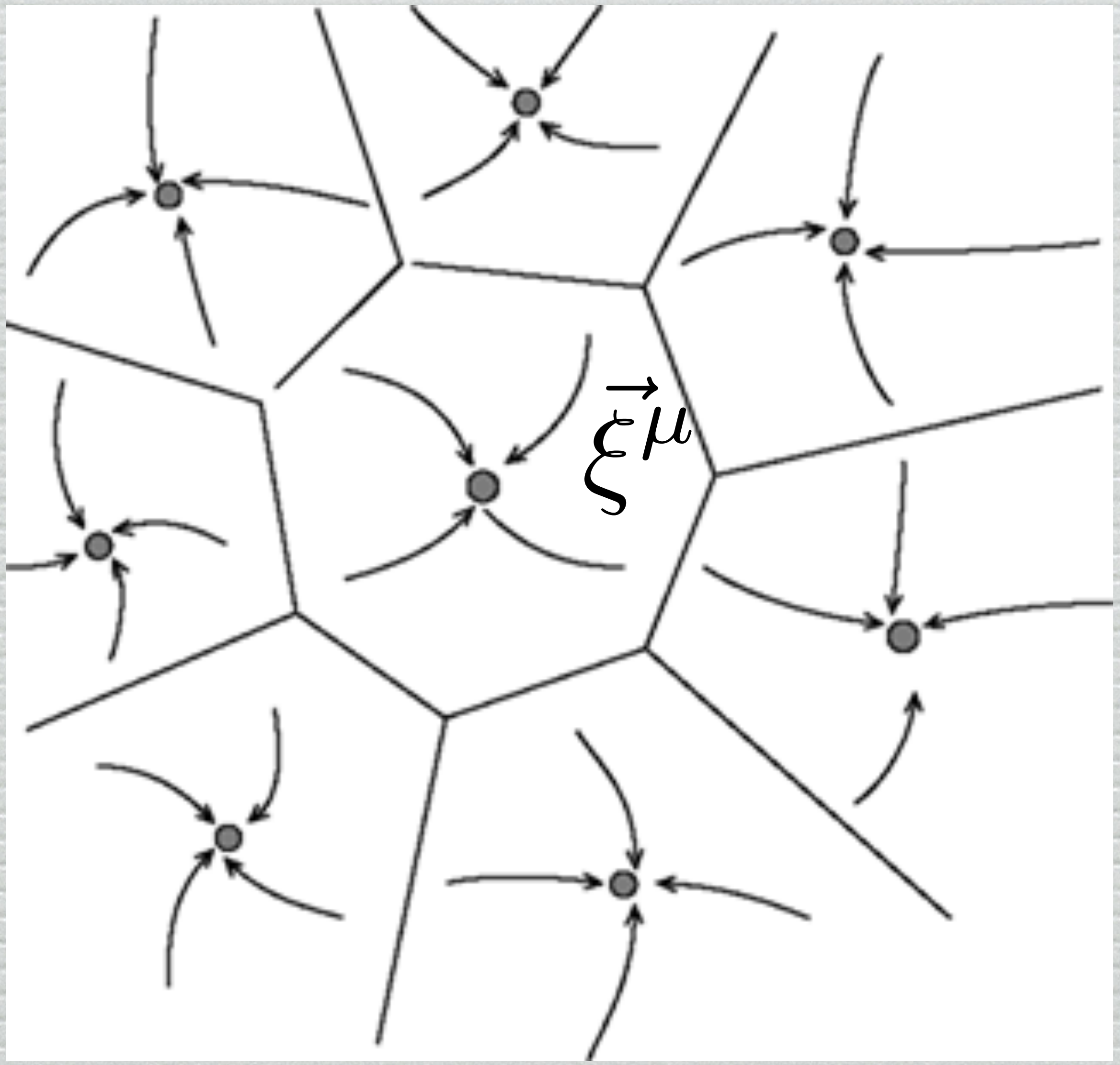
$$G(z) = \mathrm{extr}_\mu \left[\mu z - \int d\lambda D(\lambda) \log(\mu - \lambda) \right] - \log z - 1$$

A special example: Hopfield model

Neurons = N binary spins: $\vec{s} = (s_1, \ldots, s_N)$

$s_i \in \{\pm 1\}$

Patterns to be memorized: $\vec{\xi}^{\mu}$ $\mu = 1, \ldots, P$



Hopfield model

Neurons = N binary spins: $s_i \in \{\pm 1\}$

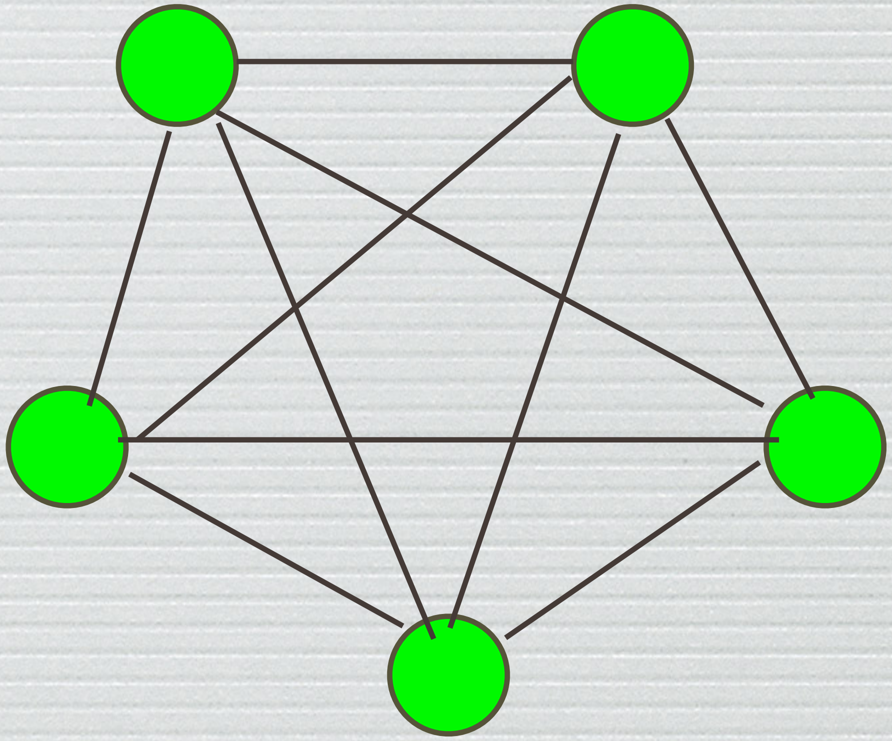
Patterns to be memorized

$$\xi_i^\mu = \pm 1,\ i \in \{1, \dots n\},\ \mu \in \{1, \dots p\}\ ,$$

$$E = -\frac{1}{2} \sum_{i,j} J_{ij} s_i s_j \qquad J_{ij} = \frac{1}{N} \sum_\mu \xi_i^\mu \xi_j^\mu$$

$$P_J(s) = \frac{1}{Z} e^{(\beta/2) \sum_{i,j} J_{ij} s_i s_j} \qquad Z = \sum_s e^{(\beta/2) \sum_{i,j} J_{ij} s_i s_j}$$

Hopfield model



$$E = -\frac{1}{2} \sum_{i,j} J_{ij} s_i s_j$$

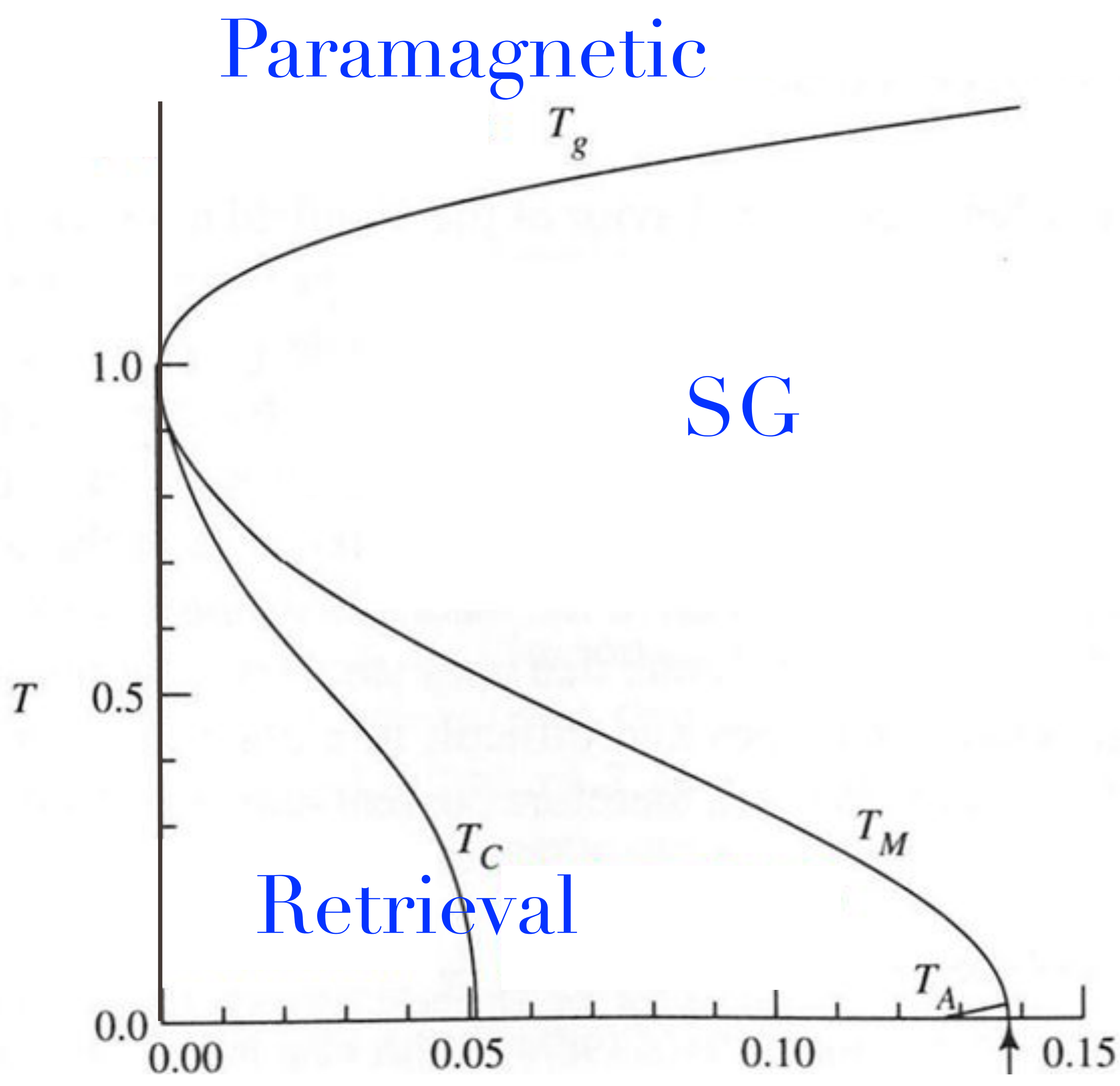
$$J_{ij} = \frac{1}{N} \sum_{\mu} \xi_i^{\mu} \xi_j^{\mu}$$

$$P_J(s) = \frac{1}{Z} e^{(\beta/2) \sum_{i,j} J_{ij} s_i s_j}$$

$$Z = \sum_{s} e^{(\beta/2) \sum_{i,j} J_{ij} s_i s_j}$$

Hopfield model

Phase diagram (Amit Gutfreund Sompolinsky 1985)



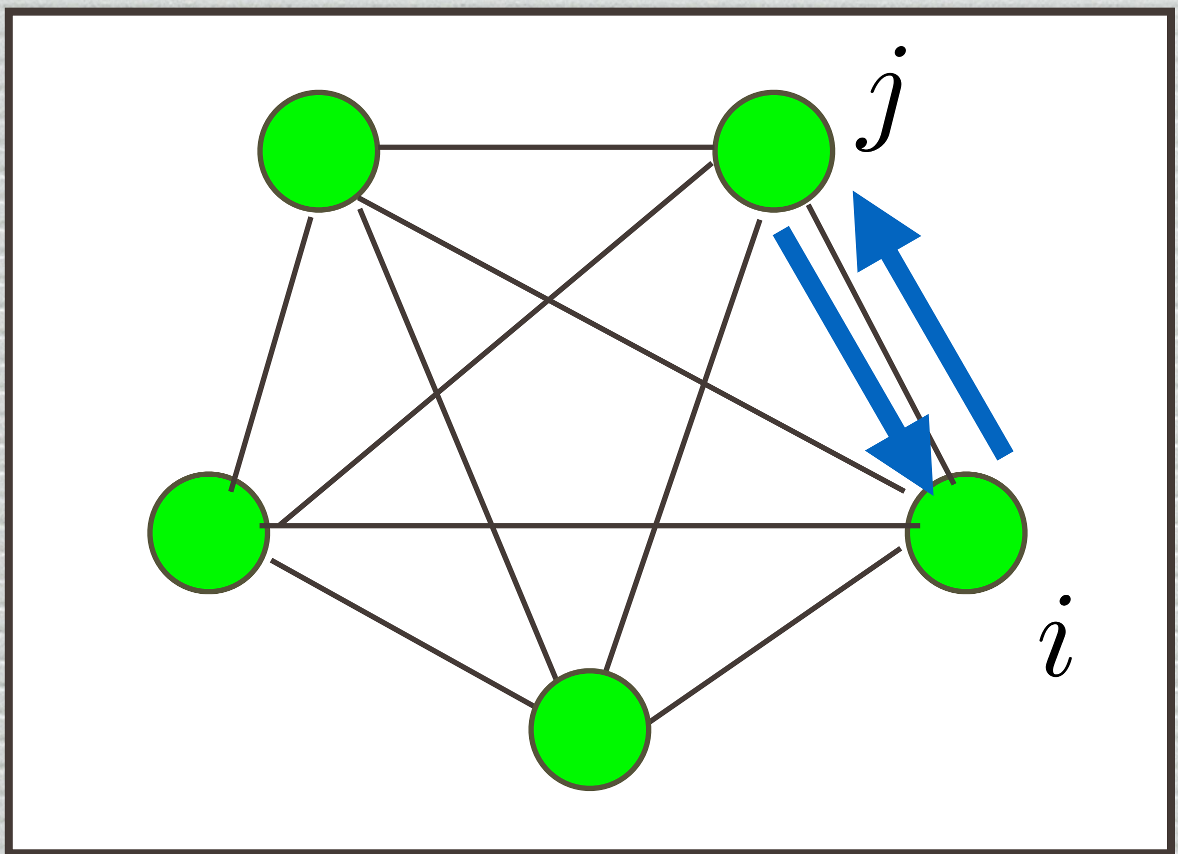
Mean field equations for solving the Hopfield model (find local magnetizations)

First attempt : TAP equations

$$H_i = \sum_k J_{ki} \tanh(\beta H_k) - \beta \tanh(\beta H_i) \sum_k J_{ki}^2 [1 - \tanh^2(\beta H_k)]$$

Disordered and infinite range

WRONG



TAP is valid only if indirect interaction from i to j through other sites can be neglected

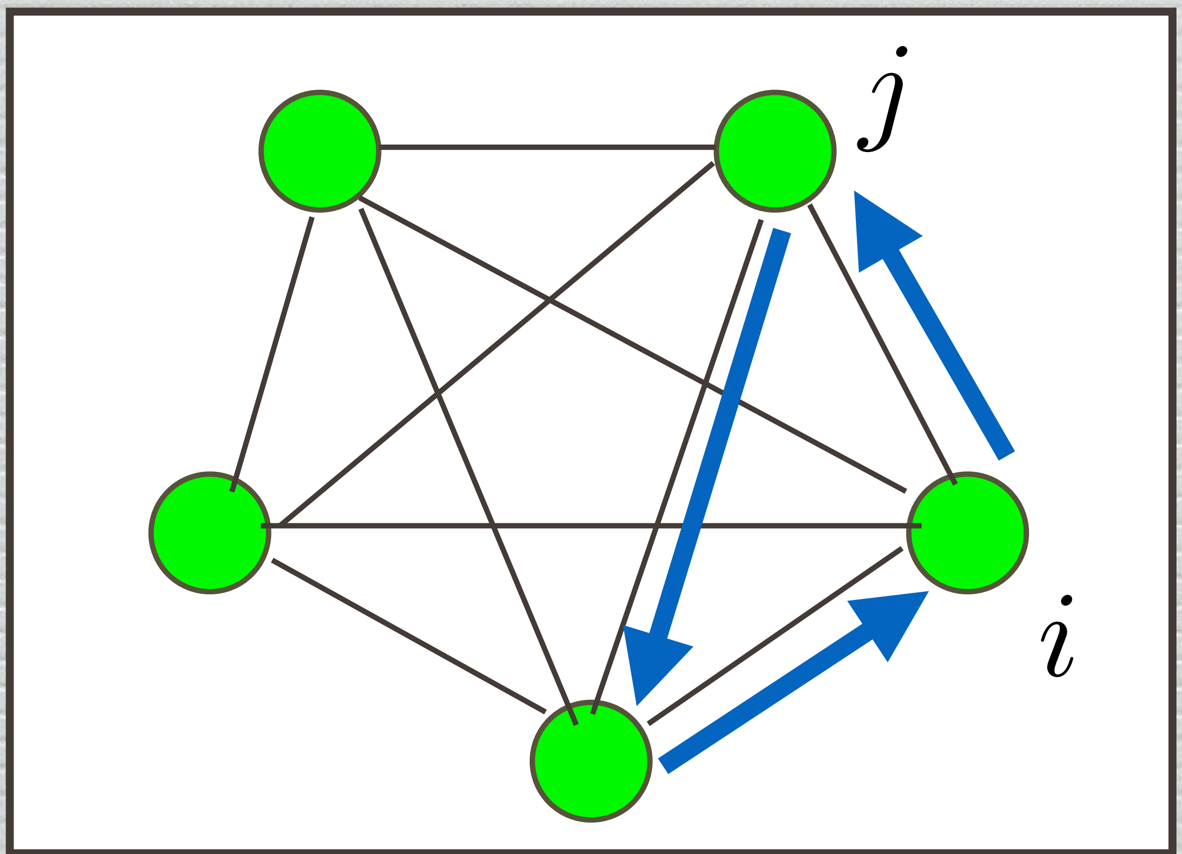
TAP in the Hopfield model: more subtle!

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

$$\overline{J_{ij} J_{jk} J_{ki}} \neq 0$$

Indirect interactions matter

« Naive » TAP does not apply



The Hopfield model as a Restricted Boltzmann Machine

$$Z = \sum_s e^{(\beta/2)\sum_{i,j} J_{ij} s_i s_j} \qquad J_{ij} = \frac{1}{N}\sum_\mu \xi_i^\mu \xi_j^\mu$$

$$Z = \sum_s \exp\left(\frac{\beta}{2N}\sum_\mu \left[\sum_i \xi_i^\mu s_i\right]^2\right)$$

Hubbard Stratonovitch (Gaussian transform) :

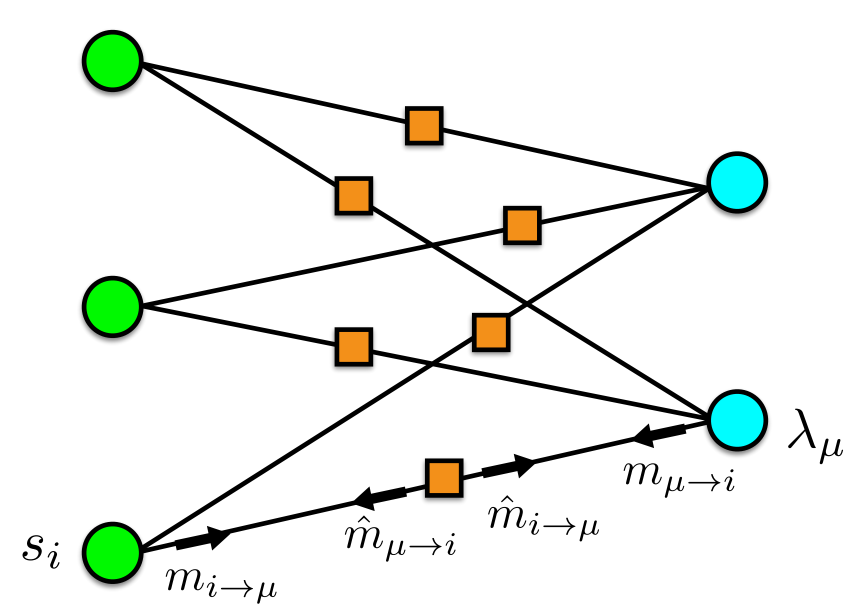
$$Z = \sum_s \int \prod_\mu \frac{d\lambda_\mu}{\sqrt{2\pi\beta}} \; \exp\left[-\frac{\beta}{2}\sum_\mu \lambda_\mu^2 + \beta \sum_{\mu,i} \frac{\xi_i^\mu}{\sqrt{N}} s_i \lambda_\mu\right]$$

$$Z = \sum_s \int \prod_\mu \frac{d\lambda_\mu}{\sqrt{2\pi\beta}} \; \exp\left[-\frac{\beta}{2} \sum_\mu \lambda_\mu^2 + \beta \sum_{\mu,i} \frac{\xi_i^\mu}{\sqrt{N}} s_i \lambda_\mu \right]$$

Spin-variable　　Pattern-variable　　Coupling

Hopfield model is a restricted Boltzmann machine, with a specific set of couplings $\frac{\xi_i^\mu}{\sqrt{N}}$ that store P patterns.

iid couplings



$$Z = \sum_s \int \prod_\mu \frac{d\lambda_\mu}{\sqrt{2\pi\beta}} \ \exp\left[-\frac{\beta}{2}\sum_\mu \lambda_\mu^2 + \beta \sum_{\mu,i} \frac{\xi_i^\mu}{\sqrt{N}} s_i \lambda_\mu\right]$$

Spin-variable Pattern-variable Coupling

$$\langle \lambda_\mu \rangle = \frac{1}{\sqrt{N}} \sum_i \xi_i^\mu \langle s_i \rangle$$

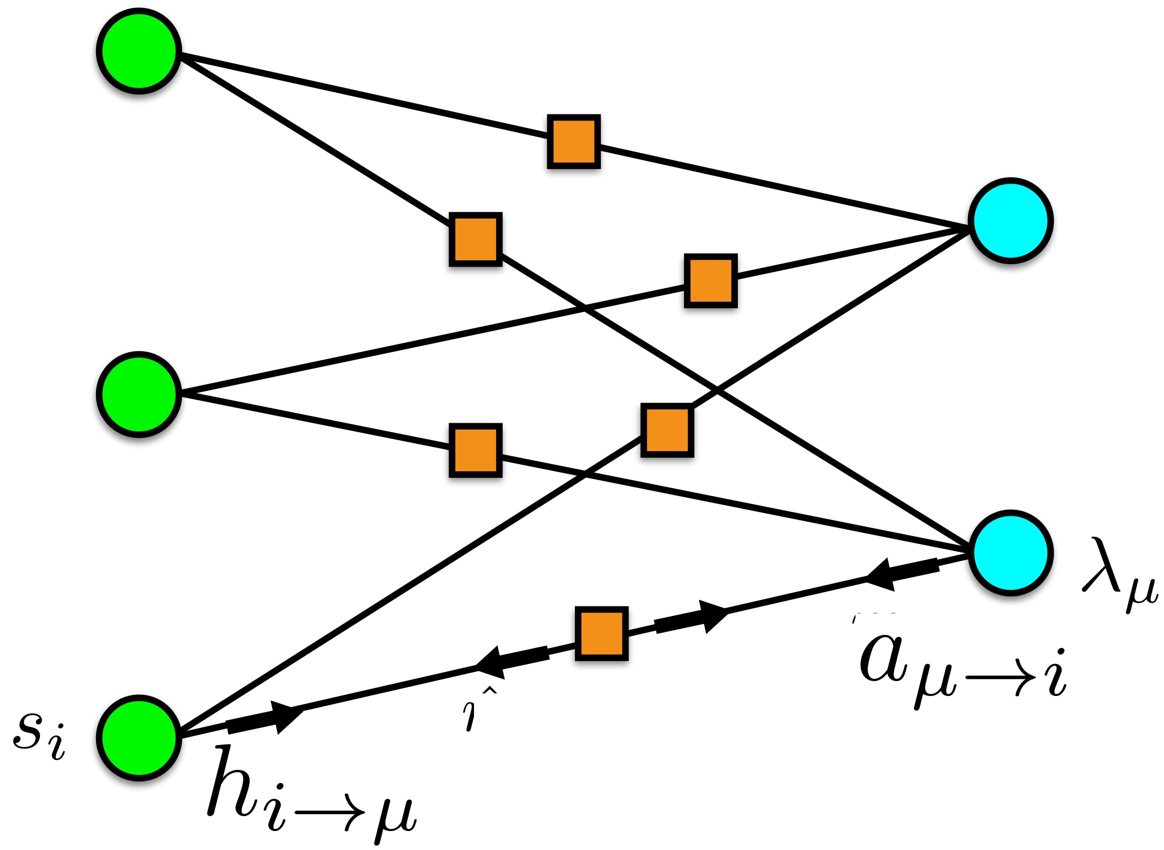
Pattern-variable describes the projection on the pattern

$\Theta(1)$ if uncorrelated

$\Theta(\sqrt{N})$ if spins are polarized towards the pattern

$$h_{i\to\mu} = \sum_{\nu(\neq\mu)} \frac{\xi_i^\nu}{\sqrt{N}} a_{\nu\to i}$$

$$a_{\mu\to i} = \frac{1}{\sqrt{N}} \frac{\sum_{j(\neq i)} \xi_j^\mu \tanh(\beta h_{j\to\mu})}{1-(\beta/N)\sum_{j(\neq i)}[1-\tanh^2(\beta h_{j\to\mu})]}$$



$$m_{i\to\mu}(s_i) \propto \exp(h_{i\to\mu} s_i)$$

$$m_{\mu\to i}(\lambda_\mu)$$

Parameterized in terms of its mean $a_{\mu\to i}$ and variance

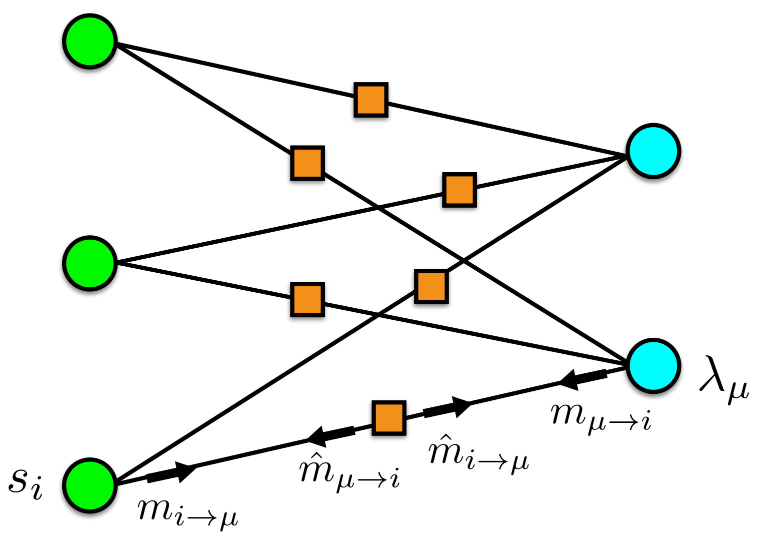
« relaxed BP »

Next step : from relaxed BP to AMP equations

$$h_{i\to\mu} = \sum_{\nu(\neq\mu)} \frac{\xi_i^\nu}{\sqrt{N}} a_{\nu\to i} \quad \simeq \sum_{\nu} \frac{\xi_i^\nu}{\sqrt{N}} a_{\nu\to i} = H_i$$

$$a_{\mu\to i} \simeq A_\mu$$

Work out the correction terms (« cavity »)



AMP equations in the paramagnetic or SG phase

$$H_i \simeq \sum_\nu \frac{\xi_i^\nu}{\sqrt{N}} A_\nu - \frac{\alpha}{1-\beta(1-q)} \tanh(\beta H_i)$$

$$A_\mu = \frac{1}{\sqrt{N}} \sum_j \xi_j^\mu \tanh(\beta H_j)$$

$$q = \frac{1}{N} \sum_i \tanh^2(\beta H_i)$$

First written in MPV 1987, claimed wrong in Nakanishi-Takayama 1997, Shamir Sompolinsky 2000, actually correct. Can be used as an iterative algorithm (with correct time indices)

Towards multilayered networks: structured patterns

Modified Hopfield model: Combinatorial patterns

$$\vec{\xi}^{\mu} = (\xi_1^{\mu}, \cdots, \xi_N^{\mu})$$

$\vec{\xi}^{\mu}$ built from superposition of elementary features $\vec{u}^{r}$

$$\boxed{\vec{\xi}^{\mu} = \frac{1}{\sqrt{\gamma N}} \sum_r v_r^{\mu} \, \vec{u}^{r}}$$

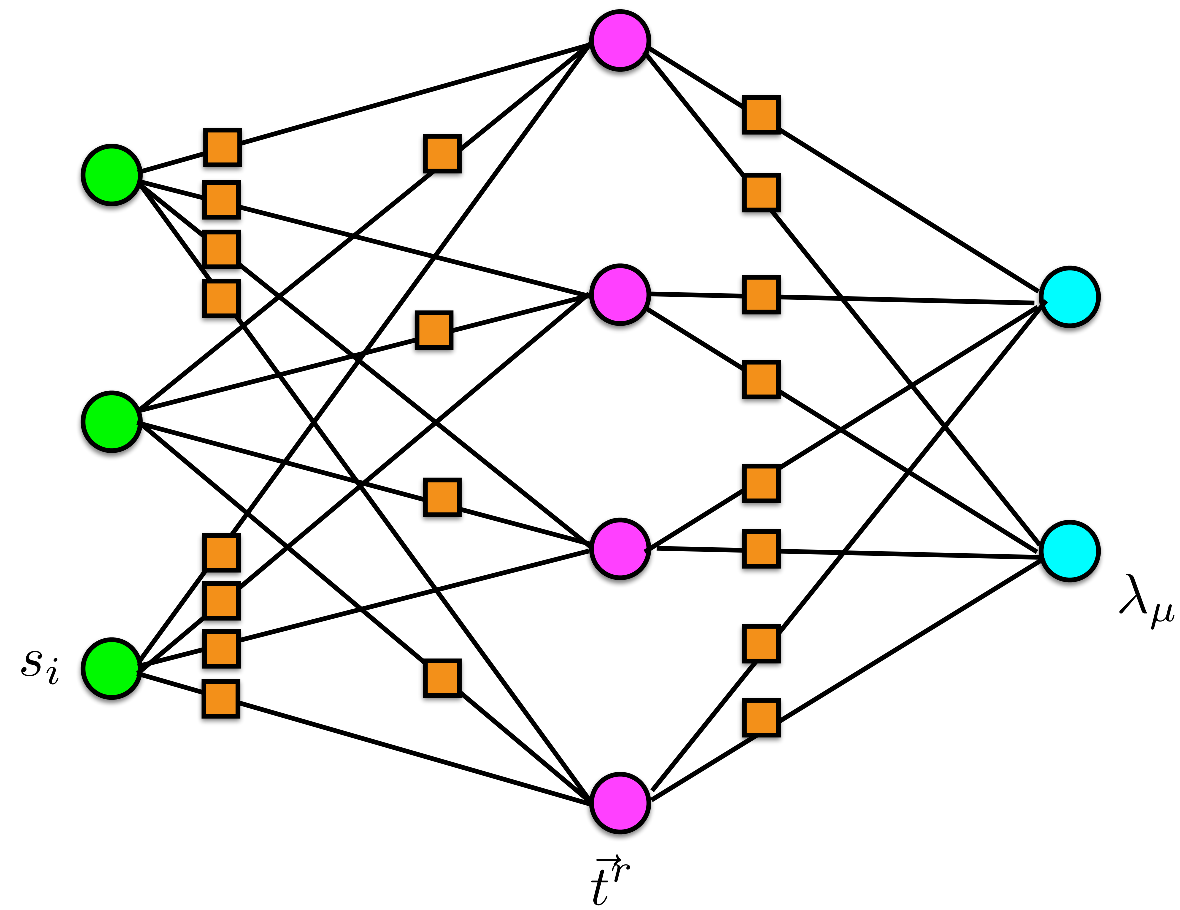
, binary $v_r^{\mu} \in \{\pm 1\}$

TAP equations in the Hopfield model with structured patterns

Modified Hopfield model: Combinatorial patterns

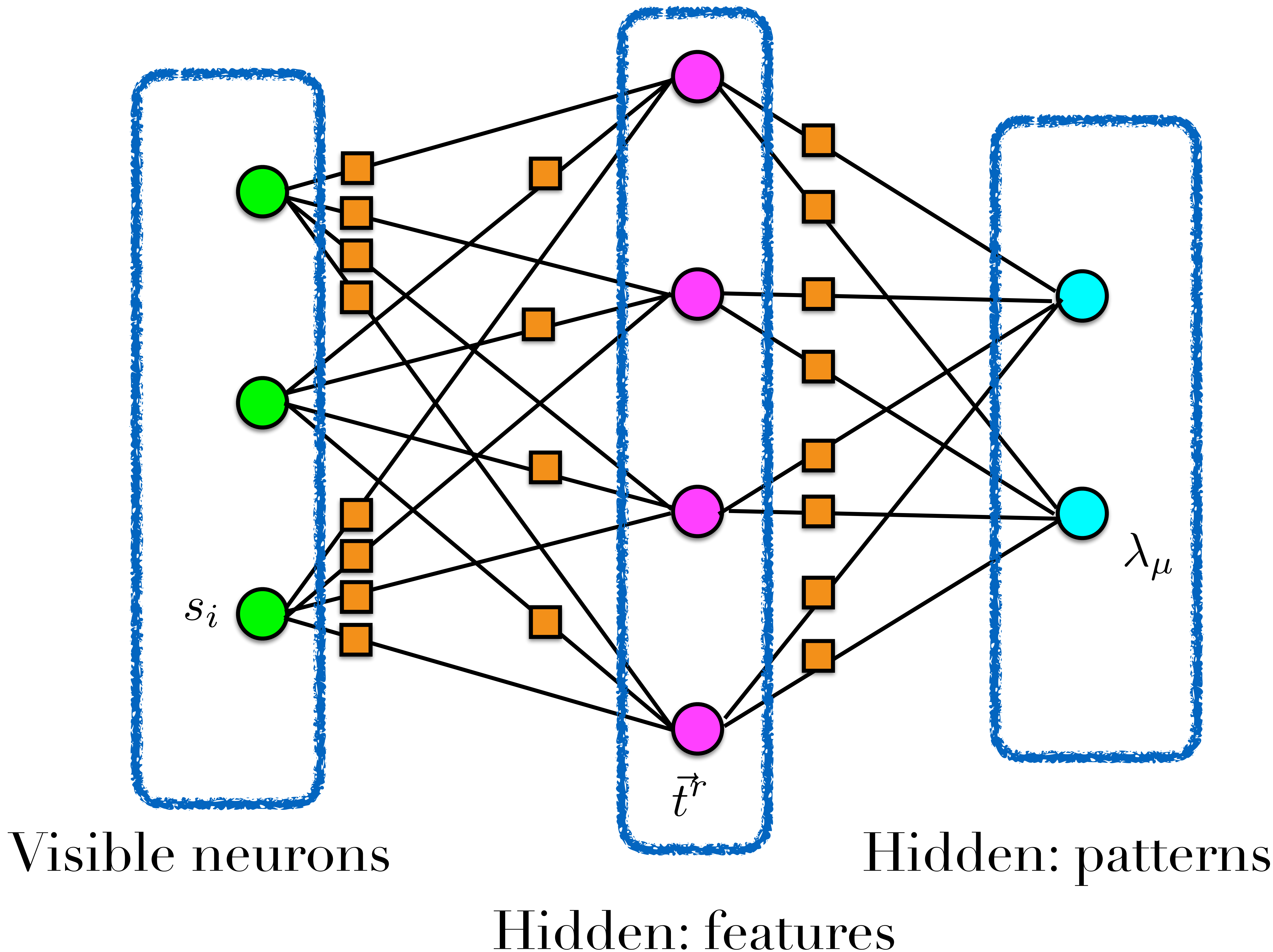
$$\boxed{Z = \sum_s \int \prod_\mu \frac{d\lambda_\mu e^{-\beta\lambda_\mu^2/2}}{\sqrt{2\pi\beta}} \ \exp\left[\frac{\beta}{\sqrt{\gamma}} \sum_{r=1}^{\gamma N} \left(\frac{1}{\sqrt{N}} \sum_i u_i^r s_i\right) \left(\frac{1}{\sqrt{N}} \sum_\mu v_\mu^r \lambda_\mu\right)\right]}$$

Disentangle the last term by another Hubbard Stratonovitch representation



$$Z = \sum_s \int \prod_\mu d\lambda_\mu \int \prod d\vec{t}^r \exp\left[-\frac{\beta}{2}\sum_\mu \lambda_\mu^2 + \beta \sum_{r=1}^{\gamma N}\left(+\frac{1}{\sqrt{\gamma}}U^r V^r - \hat{U}^r U^r - \hat{V}^r V^r\right)\right]$$

$$\exp\left[\frac{\beta}{\sqrt{N}}\sum_{r=1}^{\gamma N}\sum_{i=1}^{N}\hat{U}^r u_i^r s_i + \frac{\beta}{\sqrt{N}}\sum_{r=1}^{\gamma N}\sum_{\mu=1}^{\alpha N}\hat{V}^r v_\mu^r \lambda_\mu + \frac{\beta}{\sqrt{\gamma}}\sum_{r=1}^{\gamma N} U^r V^r\right]$$

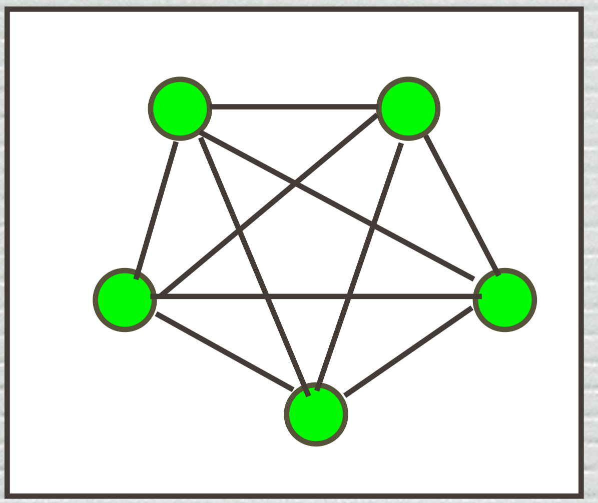


s_i
$\vec{t}^r$
λ_μ
Visible neurons
Hidden: patterns
Hidden: features

TAP equations in the Hopfield model with structured patterns

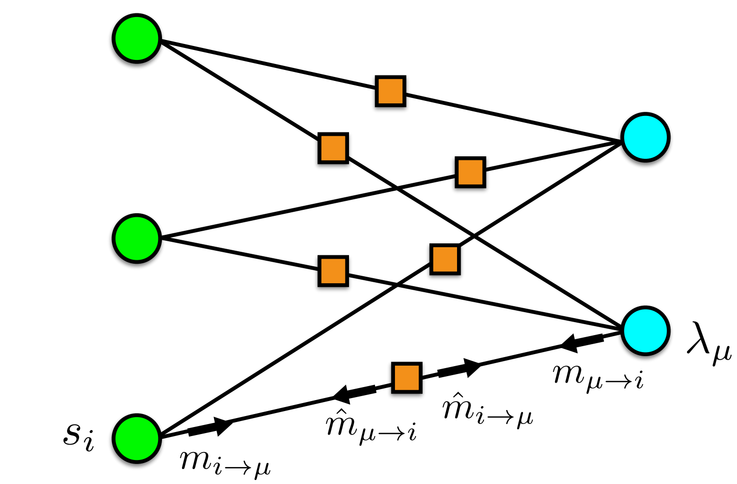
Write the cavity/BP equations. Simplify them to TAP-AMP form, involving: H_i , p_r, A_μ

TAP equations in the Hopfield model with structured patterns



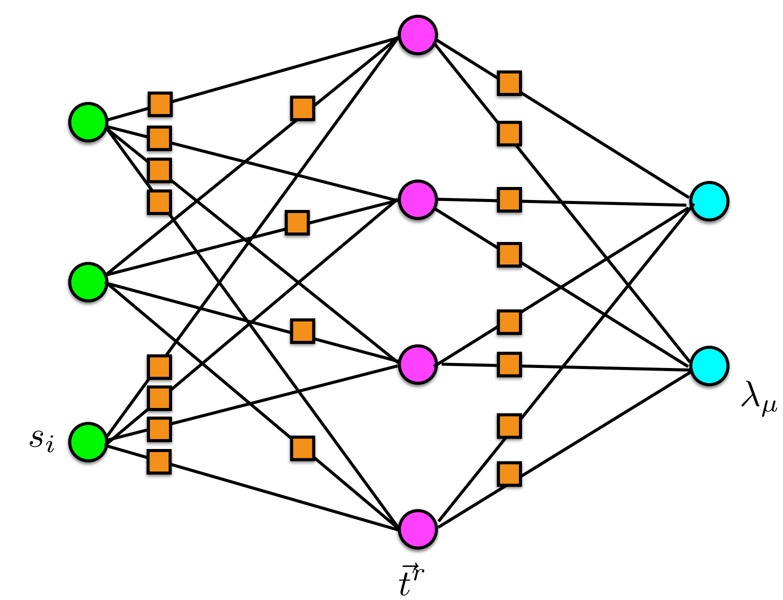
Hopfield model

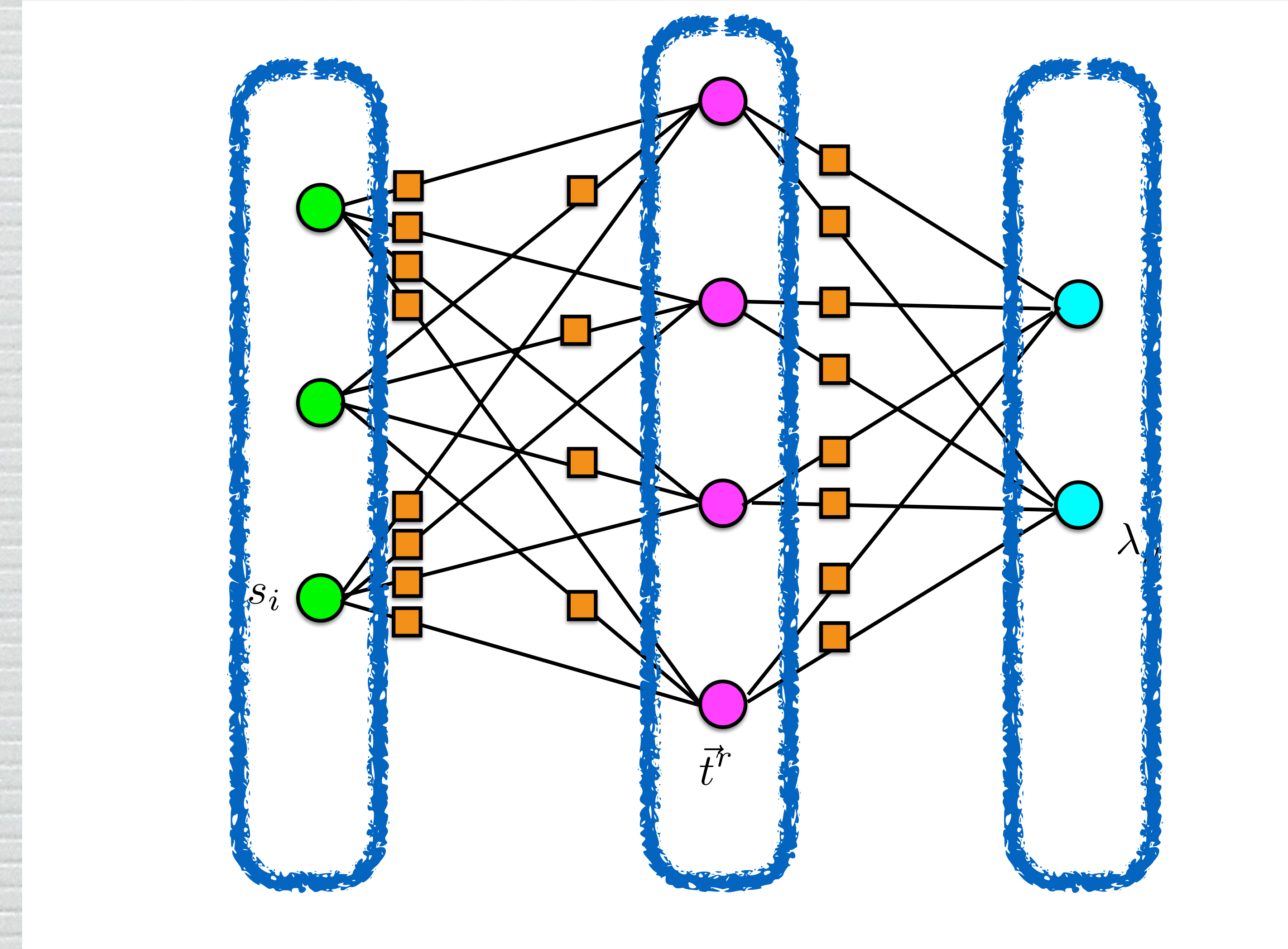
Restricted Boltzmann machine



with combinatorial patterns

Two hidden layers





Hypothesis about the success of deep networks: successive disentanglement of combinatorial correlations?

Visible input Subfeatures Features Patterns

Combinatorial correlations = new type of correlations. Present in images, in semantics, etc.

Take-home messages

- The spin glass cornucopia !

Spin glasses: Totally useless (few grams) of boring material…

Intellectual interest. Tens of thousands of papers over the last 30 years. Some of the most fascinating developments in statistical physics: Glasses, Neural networks, Optimization, Information theory, Evolution, Economy and finance,…

Powerful new concepts. Hidden order known only by the system itself $\rightarrow$ replicas.

Take-home messages

- Inference with many variables = stat phys problem of disordered system. Search of a special configuration (« crystal »)
- Theory needs an ensemble; in machine learning it means a model of data, of the world
- Mean-field approaches provide very powerful algorithms. Used in codes, in linear reconstruction, compressed sensing, tomography, community detection etc. But often tailored on a specific type of data. Limited by a dynamical phase transition

The End