Statistical physics of inference

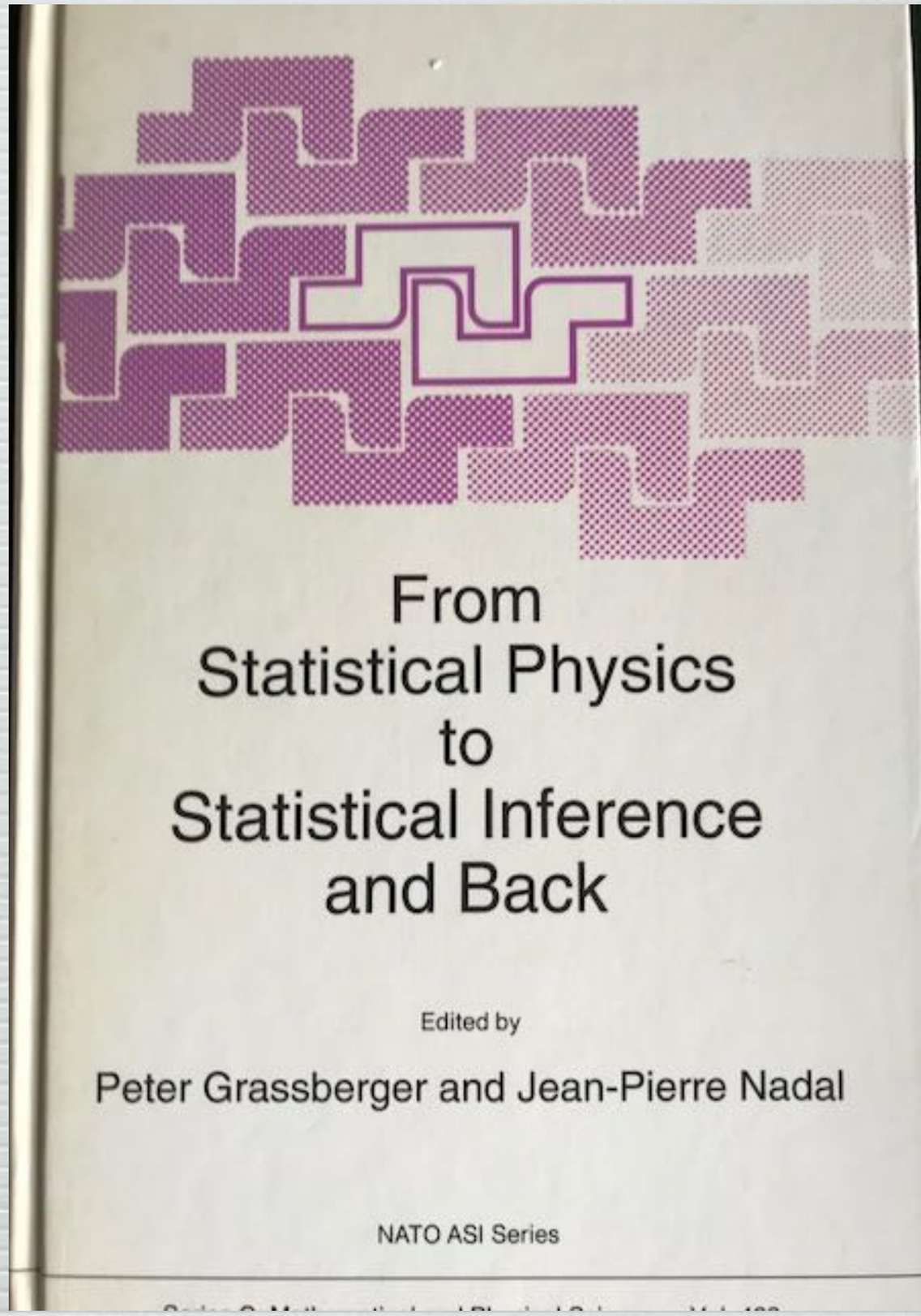
Marc Mézard

École normale supérieure
PSL University

Cargèse, August 21, 2018

Cargèse, September 1992 :

Visionary conference organized by Peter Grassberger and Jean-Pierre Nadal



Contents

Cargèse, September 1992 :

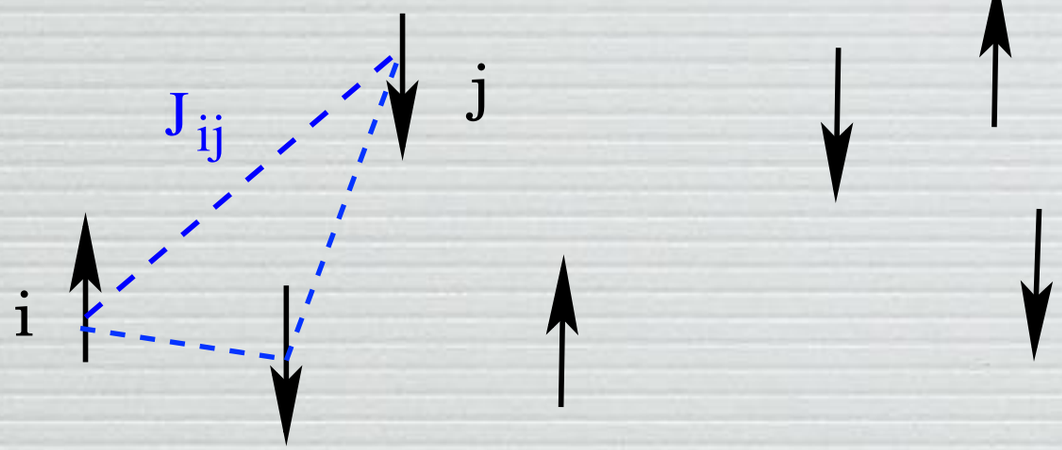
What was the situation then ?

What has changed since then?

What are the most important questions?

Spin glasses in the 80's

CuMn



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

$$E = -\sum_{ij} J_{ij} s_i s_j$$

$$P(s_1, \ldots, s_N) = \frac{1}{Z} e^{-E/T}$$

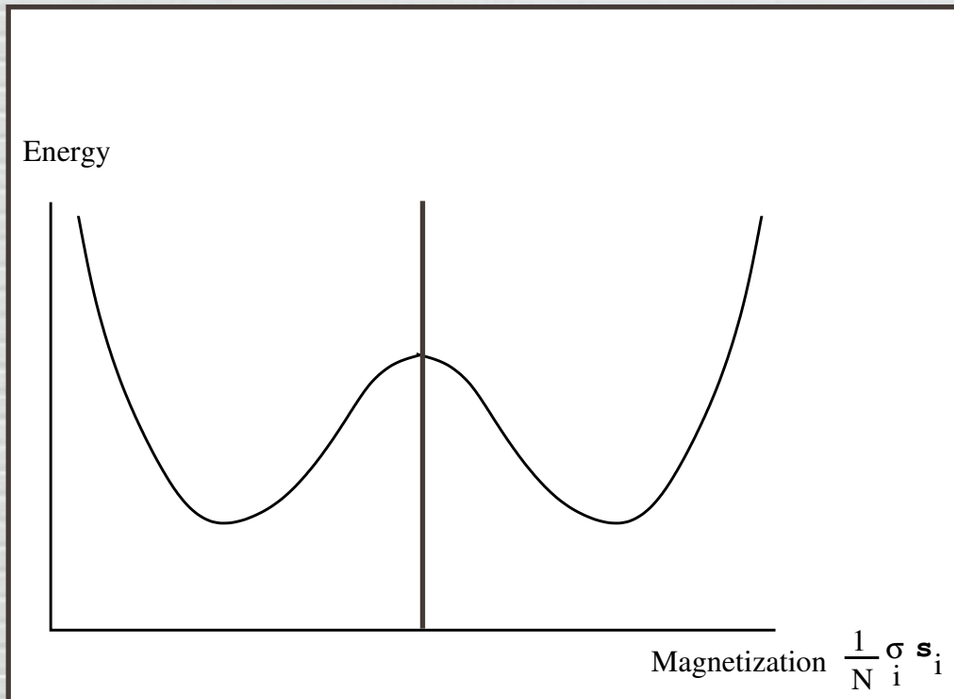
Spin glass: J_{ij} sign depends on ij ⟶ **frustration**

Frustrated triplet: $J_{ij} J_{jk} J_{ki} < 0$

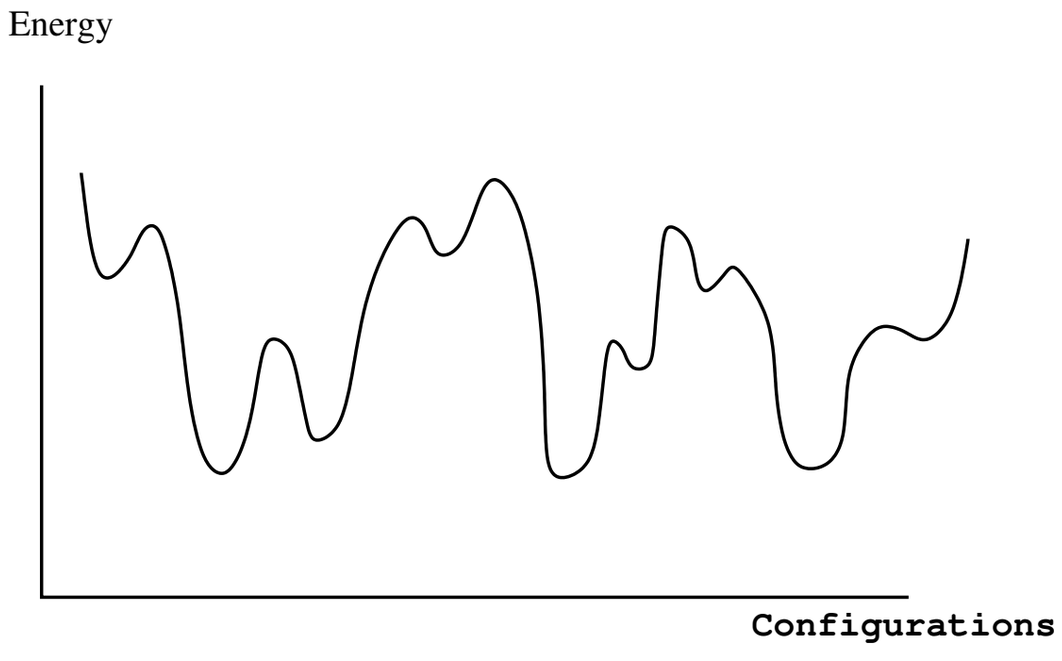
Disorder and frustration as the main building blocks

1975 **EA** ; 1976 **SK** ;1979 **P**; 1982 **P** ; 1984 **MPSTV**; 1985 **MPV** ; 1981 **D** ; 1984 **GM** ...

Mean-field lessons



Ising



Spin glass

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

2- Many metastable states, unrelated by symmetry

3- « True » ground state : fragile to perturbation!

Two main techniques, replicas and cavity

Spin glass as a cornucopia

A- Developments in Physics, not addressed today

1. 3D spin glasses
2. Quantum mechanical disordered systems (many body localization, quantum annealing, etc.)
3. Interfaces and polymers in random media - pinning
4. Heteropolymers and the folding of biopolymers (RNA, proteins)

Spin glass as a cornucopia

B- Developments in the Physics of glasses

1. Glassy dynamics. Rough landscapes, different dynamical behaviors for the two main categories of spin glasses. Aging, modified fluctuation-dissipation relation on long time scales,… CK 1993, B 1992, FP 1995
2. Structural glasses. Relevance of « 1step RSB » for glasses. Spin glasses without disorder. KTW 1989, BM 1994, MPR 1994, PP 1995, M 1995, MP 1999, PZ 2010, …
3. (Connections to evolution theory)

Spin glass as a cornucopia

C- Developments in artificial neural networks - related to SG

1. Initial works. H 1982, AGS 1985
2. Perceptron capacity. C 1965, G 1987, GD 1989, KM 1989, M 1989, …
3. Learning and generalization. G 1990, ST 1990, SST 1992, …
4. Support Vector Machines CV 1995, S
5. (Relation to neurobiology : see Brunel)

Spin glass as a cornucopia

D- Optimization and computer science

1. Simulated annealing KGV1983
2. Random Travelling salesman, matching, assignment
   MV 1984, MP 1985, A 2001…
3. Constraint satisfaction problems : K-SAT, coloring etc.
   MZ 2002, MPZ 2002, KZ 2007, KMRSZ 2007, C 2014, DSS 2015,…

Spin glass as a cornucopia

E- Information theory and codes

1. Error correcting codes S1989, KS 1998, M 2001, FZ 1999, KRU 2010
2. Multi-user detection (CDMA). T 2002, GSV 2005, MT 2006
3. Compressed sensing. CT 2005, D 2006, DMM 2009, RFG 2009, BM 2010, R2012, KMSSZ2012, RSF 2016…

Spin glass as a cornucopia

F- Internal developments since the 80's

1. Finite connectivity spin glasses MP 2000
2. Rigorous results on mean field spin glasses GG1998, T 2003, G 2003, P 2011

Spin glass as a cornucopia

5 main sections (Glasses, Neural networks, Optimization, Information theory, Internal developments) including 13 topics which have seen considerable progress since 1992…

Most relevant :

- Landscape and glassy dynamics
- Finite connectivity systems: cavity method, algorithms
- Codes, Compressed sensing, Information theory

Realizing the importance of algorithmic complexity

Mean field equations can be iterated as algorithms

The replica method can tell about its regimes of convergence

Emergence of a common language, a common scientific field

Statistical inference

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

Restricted sense : find parameters of a probability distribution

Urn with 10.000 balls.

Draw 100, find 70 white balls and 30 black balls.

Best guess for the composition of the urn? How reliable? Probability that it has 6000 white- 4000 black?

If only black and white balls , with fraction x of white, probability to pick-up 70 white balls is $\binom{100}{70} x^{70}(1-x)^{30}$

Log likelihood of x : $L(x) = 70 \log x + 30 \log(1-x)$

Maximum at $x^* = .7$ Probability of .6 : $e^{L(.6)-L(.7)}$

Bayesian inference

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

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

Statistical inference

Challenge = rules with many hidden parameters.

$$x = (x_1, \ldots, x_N) \qquad N \gg 1$$

Many measurements $y = (y_1, \ldots, y_M)$ $M \gg 1$

Measure of the amount of data $\alpha = M/N$

- **Algorithms**
- **Prediction on the quality of inference**, on the performance of the algorithms, on the type of situations where they can be applied

A first simple example : Planted SK model

Bayesian inference with many unknown and many measurements

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

Measurements $y = (y_1, \ldots, y_M)$

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

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

Bayesian inference with many unknown and many measurements

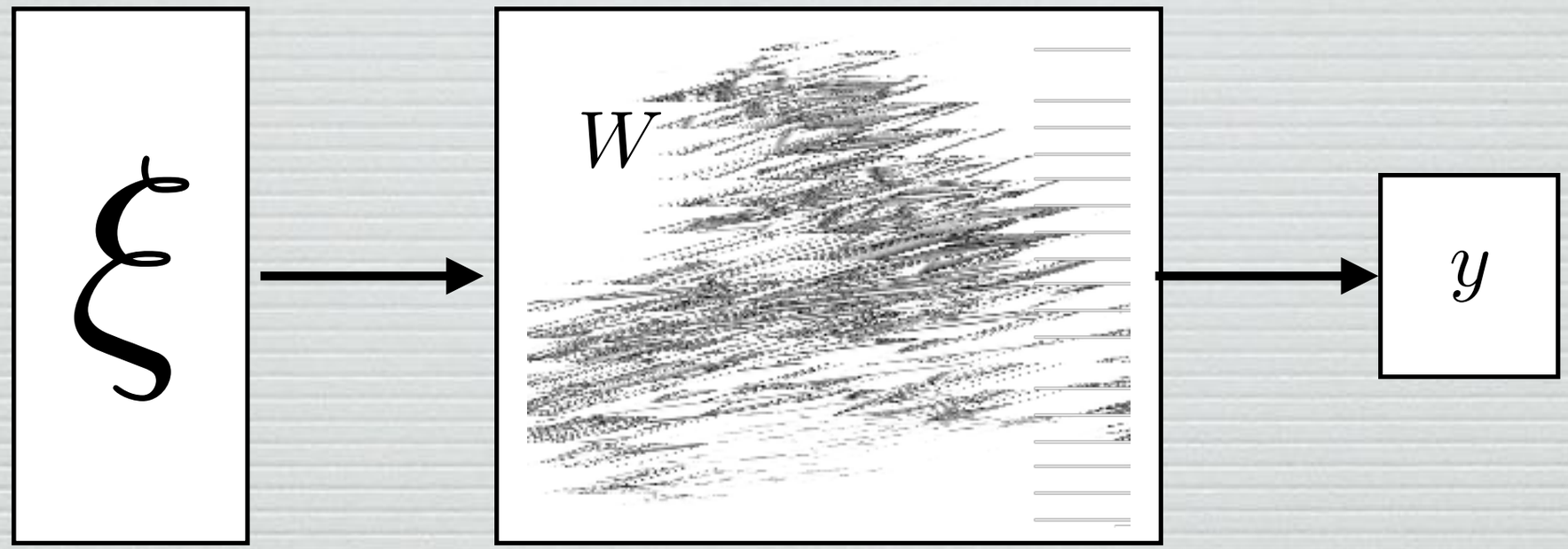
$$P(x) = \frac{1}{Z(y)} \left(\prod_i P_i^0(x_i) \right) \exp \left[- \sum_\mu E_\mu(x, y_\mu) \right]$$

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

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

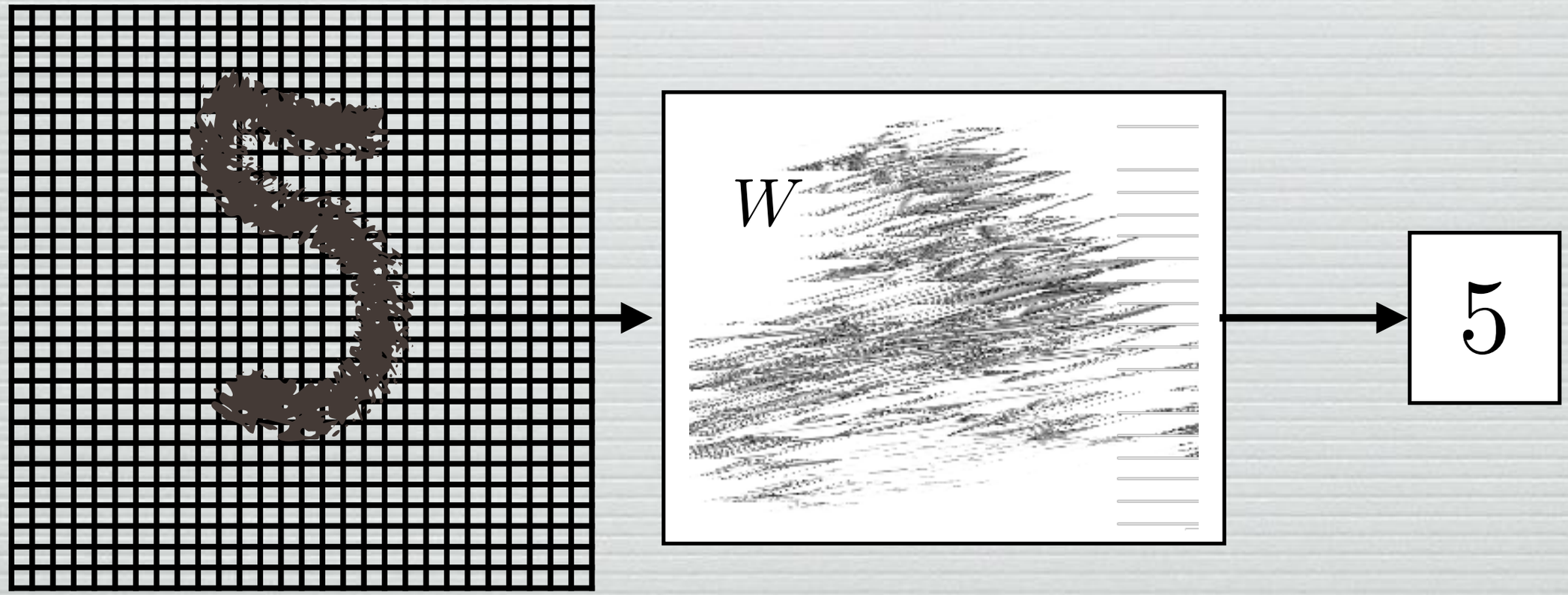


Input ξ

Machine, parameters W

Output y

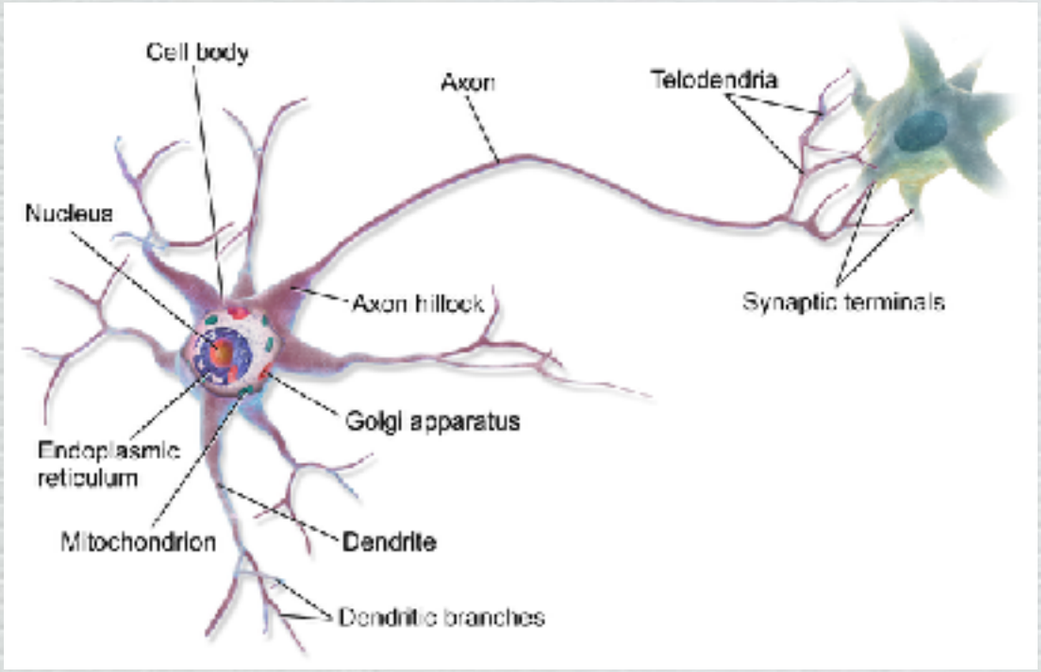
Machine learning

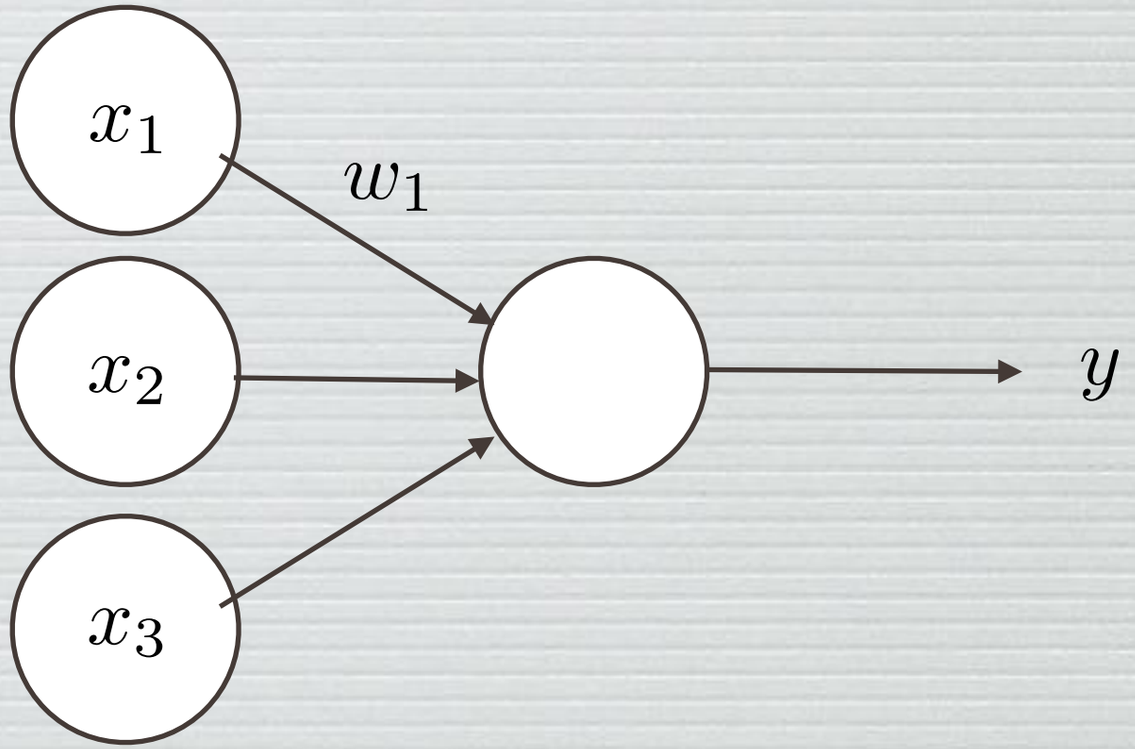


Handwritten digit, 28^2 pixels

Machine, parameters W

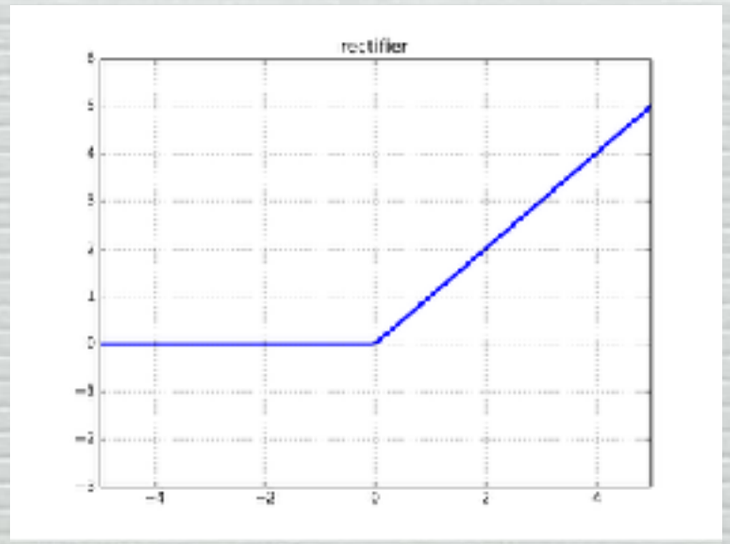
Output the number

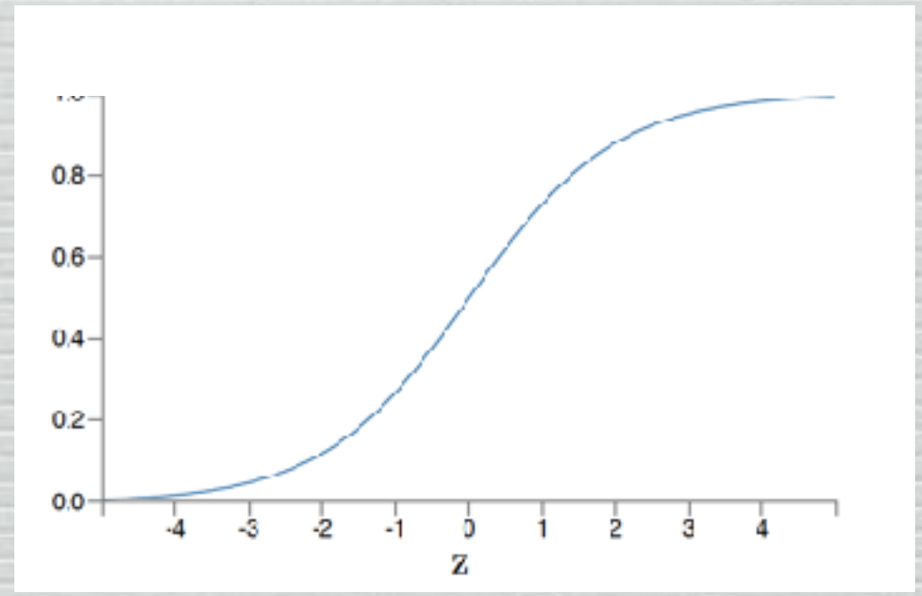




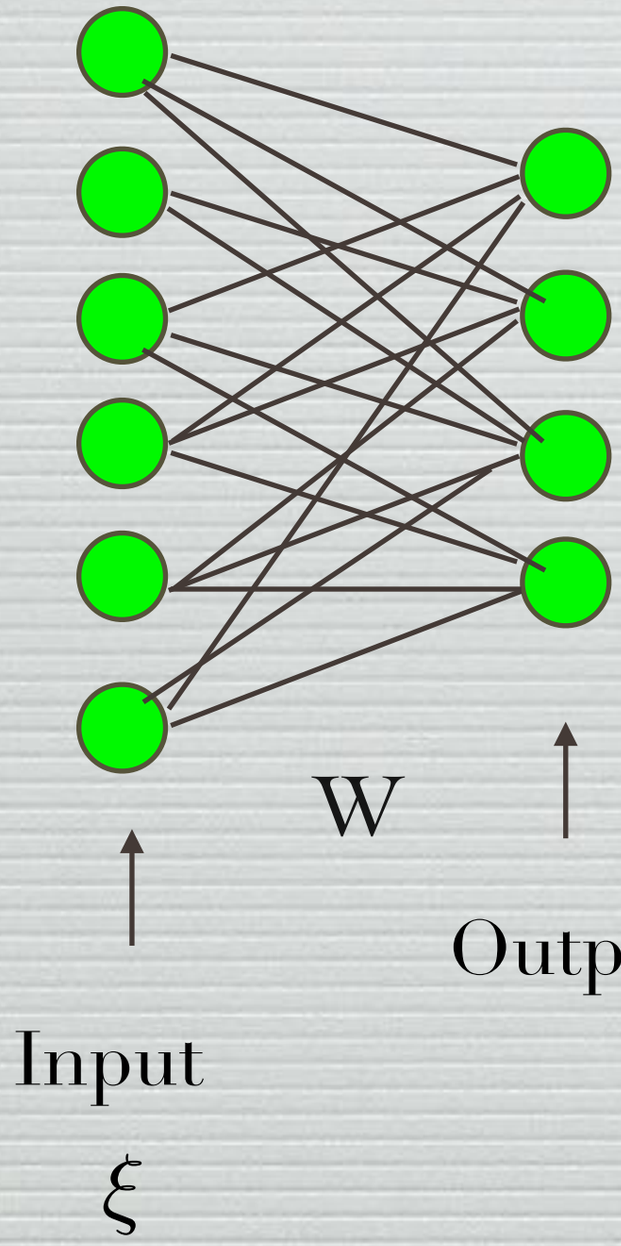
$$y = f(w_0 + w_1 x_1 + w_2 x_2 + w_3 x_3)$$

Formal neural network

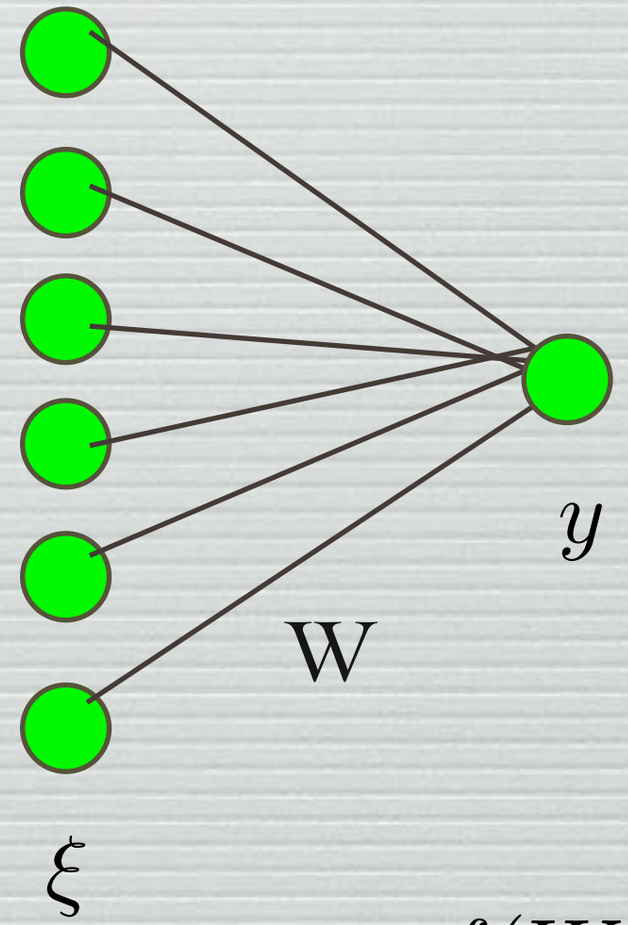




Simple perceptron



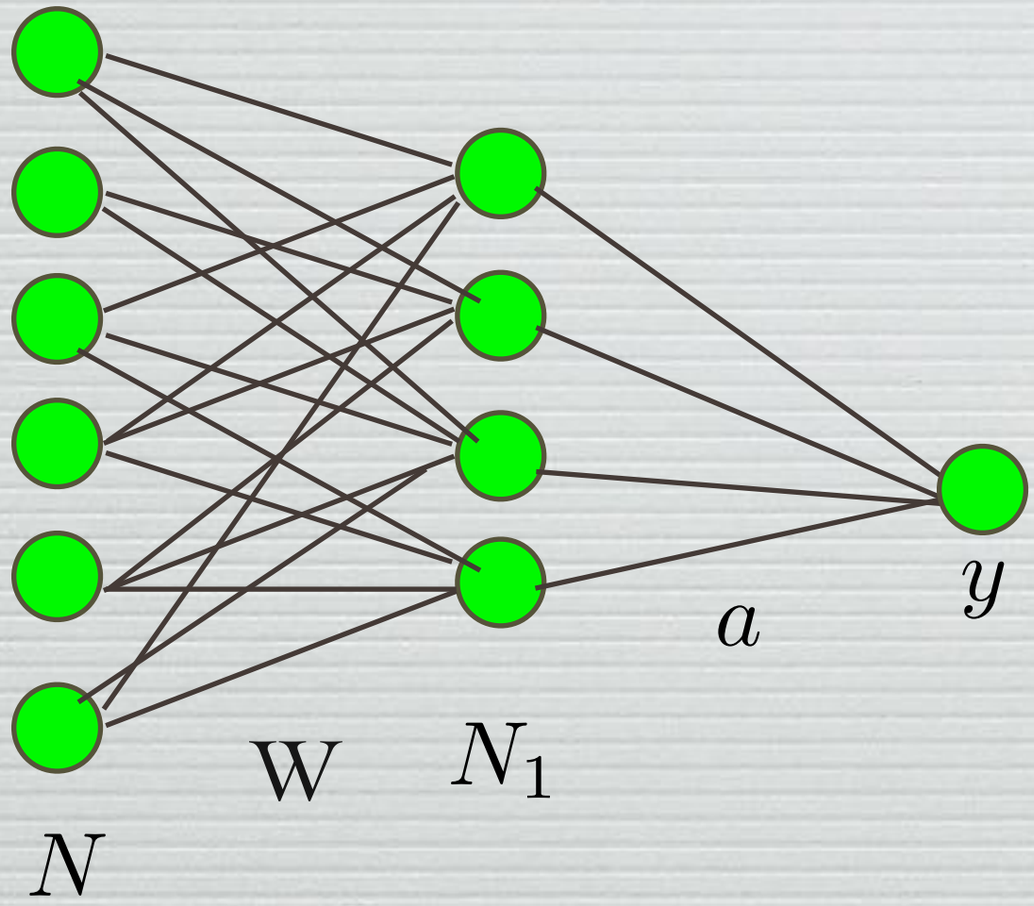
Decouples into independent single output machines



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

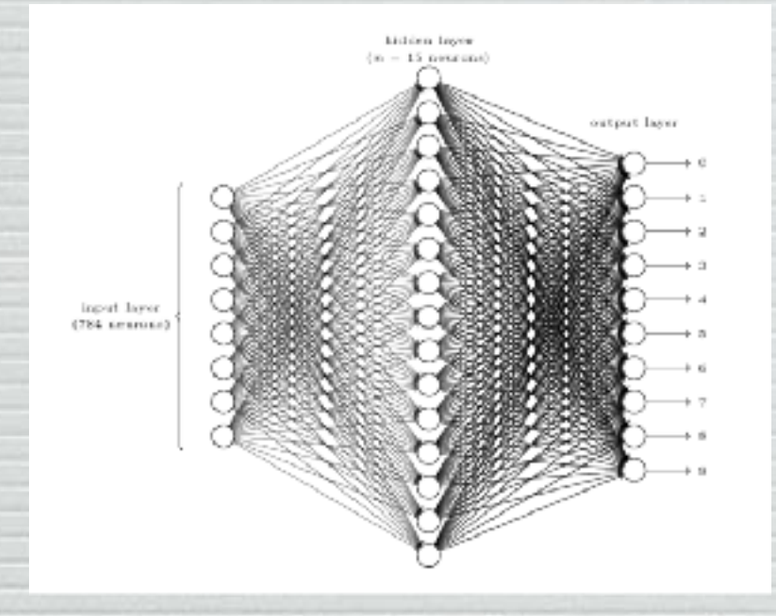
Limited to linearly separable rules

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



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. $E_t = \sum_\mu [y_\mu - f(W, \xi_\mu)]^2$

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

Machine learning: training and generalization

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

Database = M examples of input-outpu)

Bayesian learning:

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

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

Energy landscapes and glasses

In 1992 : Spin glasses, metastable states, $P(q)$, ultrametricity.
But connection to glassy dynamics is at a qualitative level. not so explicit.
Since then : aging, memory, much progress on the use of replicas in structural glasses.

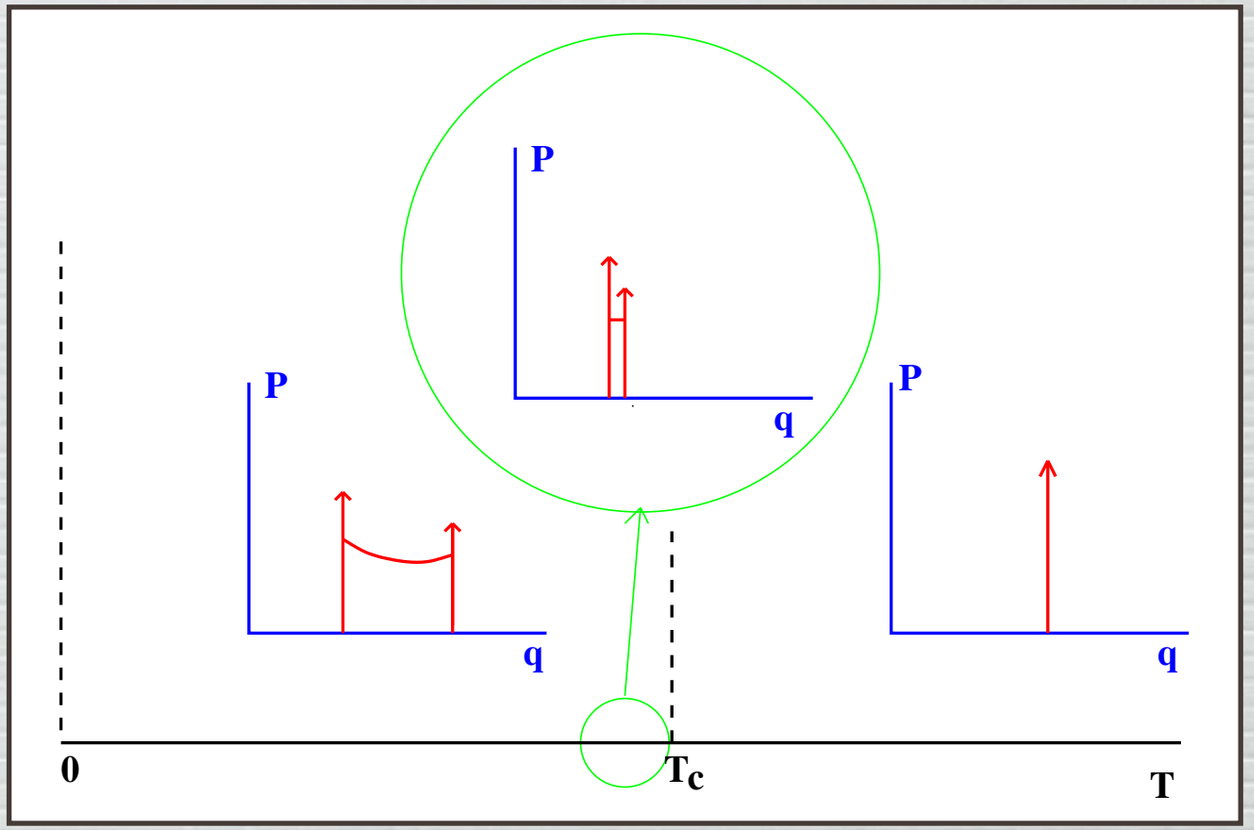
Understanding of the different dynamical properties for each of the two broad classes of glasses.

Two families of glasses

Probability (2 random configurations have overlap q)

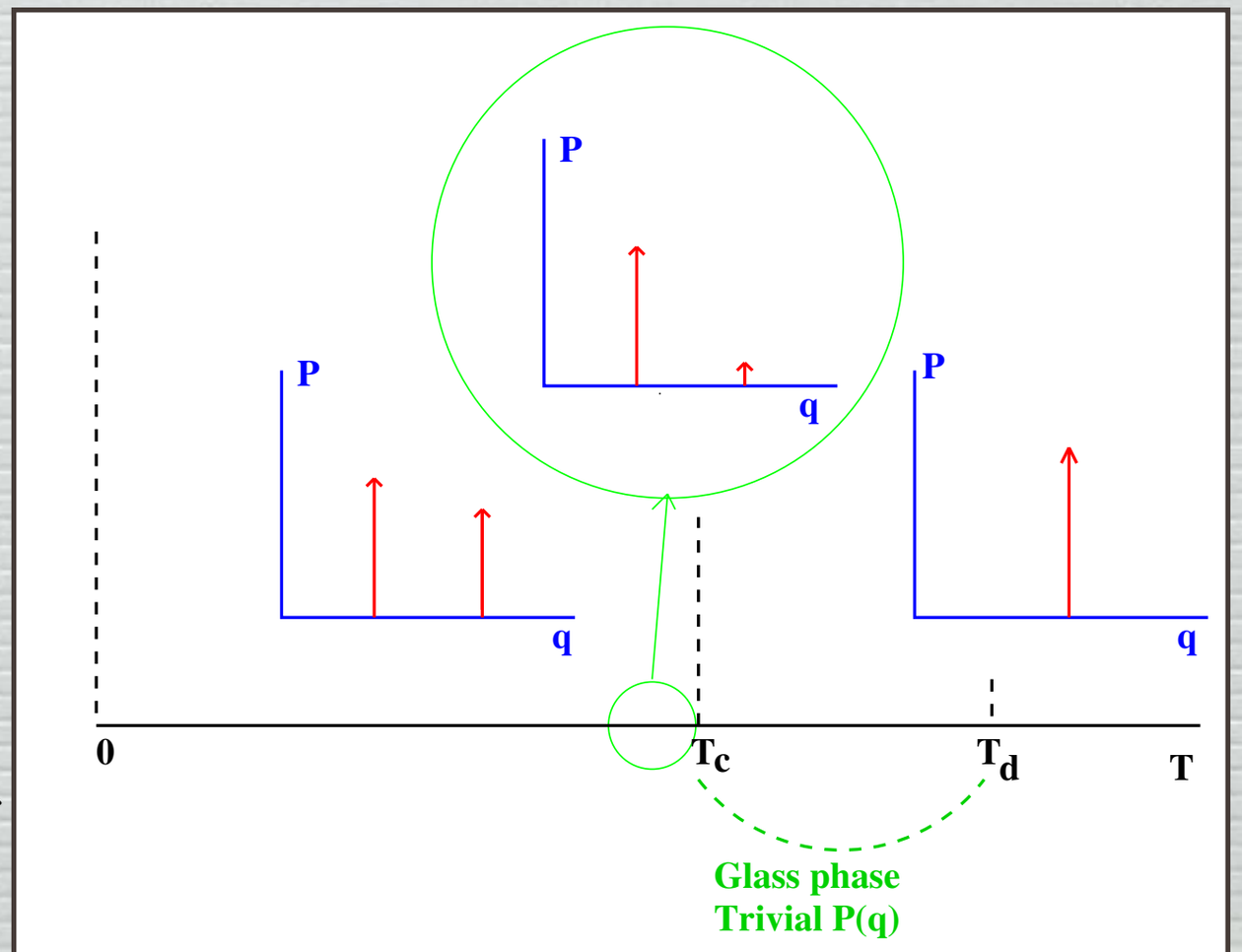
Continuous transition

« Full replica symmetry breaking »



Discontinuous transition

« One step replica symmetry breaking »



Energy landscapes and glasses

Learning:

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

Typical of a glassy system. Experience in glasses tells us that learning should be slow, that the landscape should be full of metastable states, of traps, with a very slow dynamics (when explored e.g. by Monte-Carlo dynamics).

But: deep networks seem to have an « easy learning ». Why ?

Q 1: Many of the speakers in this 2018 Cargèse school have their own interpretation for this puzzling behavior. Which one is correct?

Model of data: ensemble

Learning:

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

Algorithmic studies typically uses one (or several) databases for $\{\xi_\mu, y_\mu\}$

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

Examples from the 80's: iid patterns

Q 2: Find good generative models of the world

Ensemble of data, thermodynamics

Learning:

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

$$W = (W_1, \ldots, W_N)$$

Data generated from a probability distribution $\prod_{\mu=1}^{P} P_{data}(\xi_\mu, y_\mu)$

$Z(\beta, \{\xi_\mu, y_\mu\})$ contains all the information: $E_t = -\partial \log Z/\partial\beta$

Thermodynamic limit: $N \to \infty$ $P \to \infty$ $\alpha = P/N$

$Z \simeq e^{-\beta N \phi}$ ϕ : free energy density. Depends on data

Large deviations: $P_{data}(\phi) \propto e^{NG(\phi)}$

ϕ concentrates around $\phi^* = \text{argmax } G(\phi)$

Replicas

Learning:

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

$$Z \simeq e^{-\beta N \phi} \qquad P_{data}(\phi) \propto e^{N G(\phi)}$$

$$\phi \simeq_{N\to\infty} \phi^* = \text{argmax}\ G(\phi)$$

Replicas: compute $\quad \mathbb{E}_{data}\, Z^n \;=\; \int d\phi\, e^{N[G(\phi) - \beta n \phi]}$

in the limit $\quad n \to 0$

$$Z^n = \int \prod_{a=1}^{n} dW^a\, P^0(W^a)\ \exp\left(-\beta \sum_{\mu,a} \left[f(W^a,\xi_\mu) - y_\mu)\right]^2\right)$$

$$\text{often} \quad = \quad e^{N\ \text{extr}\ \Psi(Q^{ab})} \qquad\qquad Q^{ab} = \langle W^a . W^b \rangle$$

Replica symmetry iff extremum at $\quad Q^{ab} = q + (\tilde{q} - q)\delta_{ab}$

Generative model of data : teacher-student

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

Data generated by a teacher. The teacher has her 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

Smart student : **knows the teacher's architecture and the generative distribution** P^T .

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

$$P^S(W) = P^T(W)$$

Bayes-optimal inference

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

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

Student: prior on w : $P^S(w)$

guess on data-generating process $P^S(y|w)$

seeks w from Bayes $P(w|y) = P^S(y|w)P^S(w)/P^S(y)$

Bayes-optimal inference: if the student knows the generative model for the parameters and the data, she should use them:

$$P^S(w) = P^T(w) \qquad P^S(y|w) = P^T(y|w)$$

Bayes-optimal inference uses:

$$P(w|y) = P^T(y|w)P^T(w)/P^T(y)$$

Seeks a special « planted » configuration w^* : a « **crystal** »

Why Bayes-optimal inference is much simpler

Data generated from $P^T(w^*)P^T(y|w^*)$

Inference seeks w using: $P(w|y) = P^T(y|w)P^T(w)/P^T(y)$

Then: generate $w_1,\ w_2,\ w_3$ from $P(w|y)$

For any function $f(w, w')$ of two weight vectors w, w' :

$$\mathbb{E}_{data} f(w_1, w_2) = \mathbb{E}_{data} f(w^*, w_3)$$

« Nishimori relations » (1980), gradually recognized as a general relation of Bayes-optimal inference (Iba 1998, …). **No glass phase** (but dynamical glass transition is possible)

In replica framework : consequence of $\mathcal{S}_{n+1}$ symmetry

First example of hard « crystal hunting »

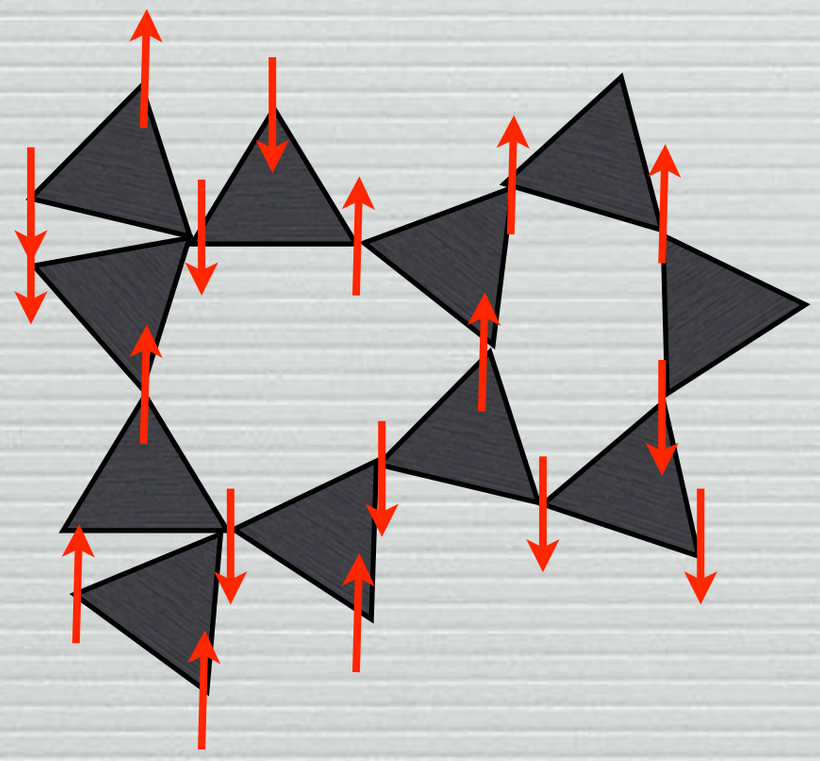
3-spin interaction Mattis model

$$E = -\sum_{ijk} J_{ijk}\sigma_i\sigma_j\sigma_k \qquad J_{ijk} = \tau_i\tau_j\tau_k \qquad \sigma,\ \tau\ \in \{\pm 1\}$$

$$s_i = \sigma_i\tau_i\ \in \{\pm 1\}$$

$s_i = 1$

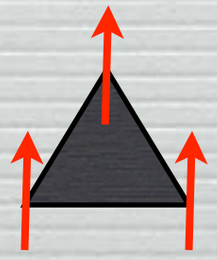
$s_i = -1$



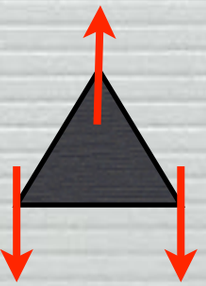
$$E = -\sum_{ijk} s_i s_j s_k$$

Randomly chosen triplets

Lowest E:



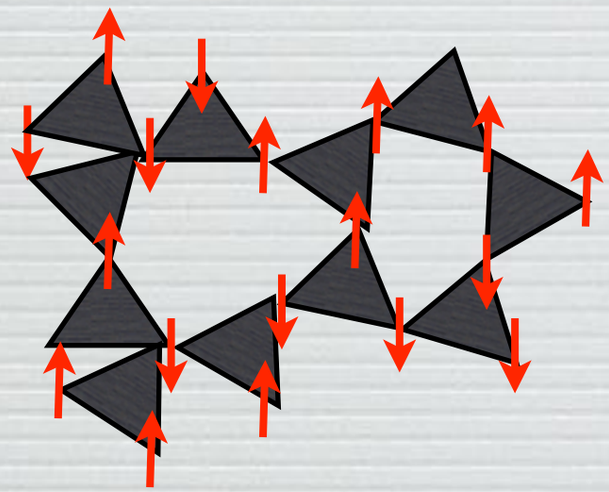
or



or..

Franz, M, Parisi, Ricci-Tersenghi, Weigt, Zecchina 01

Trapped in a glass phase



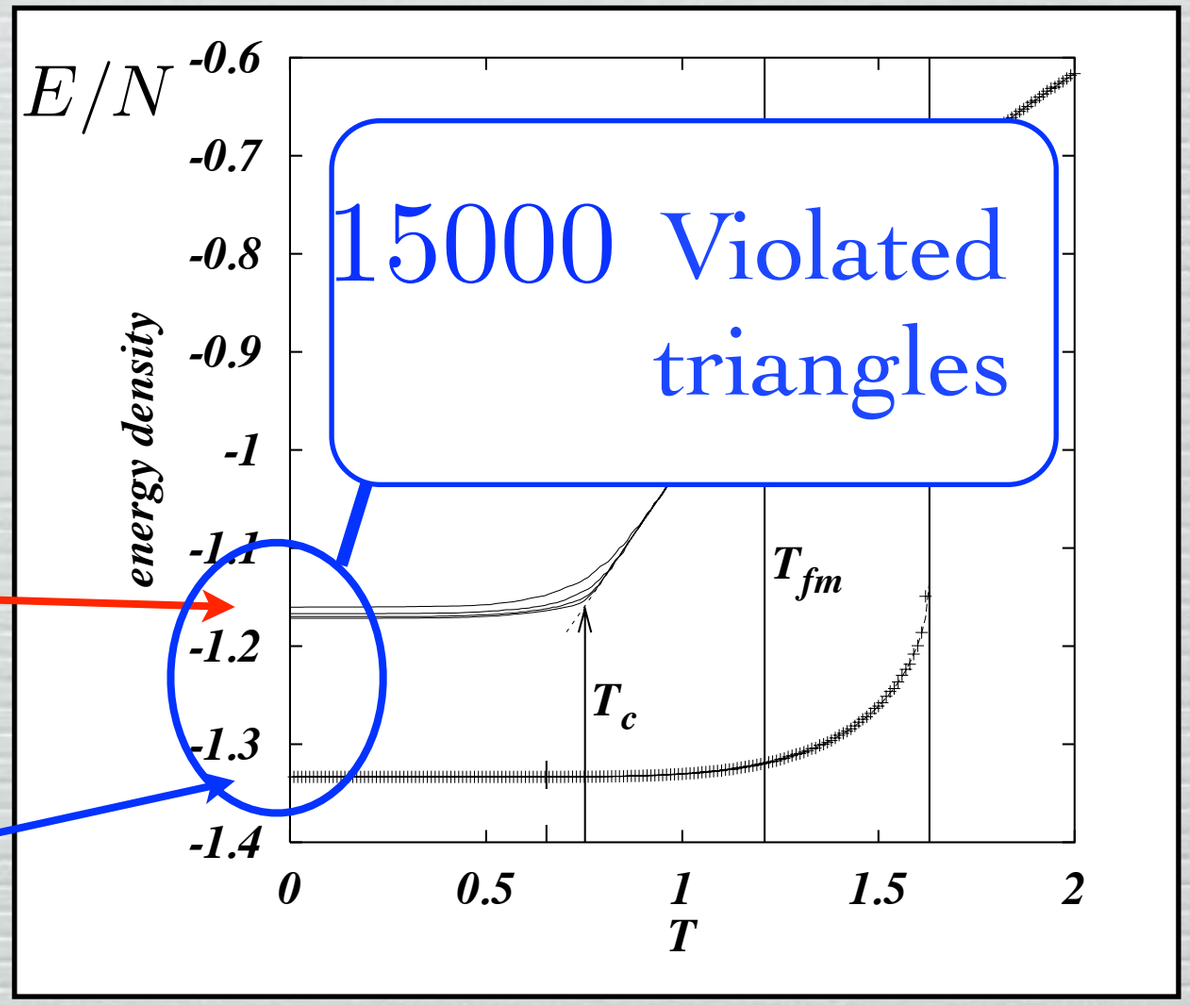
$$E = -\sum_{ijk} s_i s_j s_k$$

$$P(s_1, \ldots, s_N) = \frac{1}{Z} e^{-E/T}$$

10^5 spins, 4 triangles per spin

Metastable states found by simulated annealing 10^4 to 10^7 steps

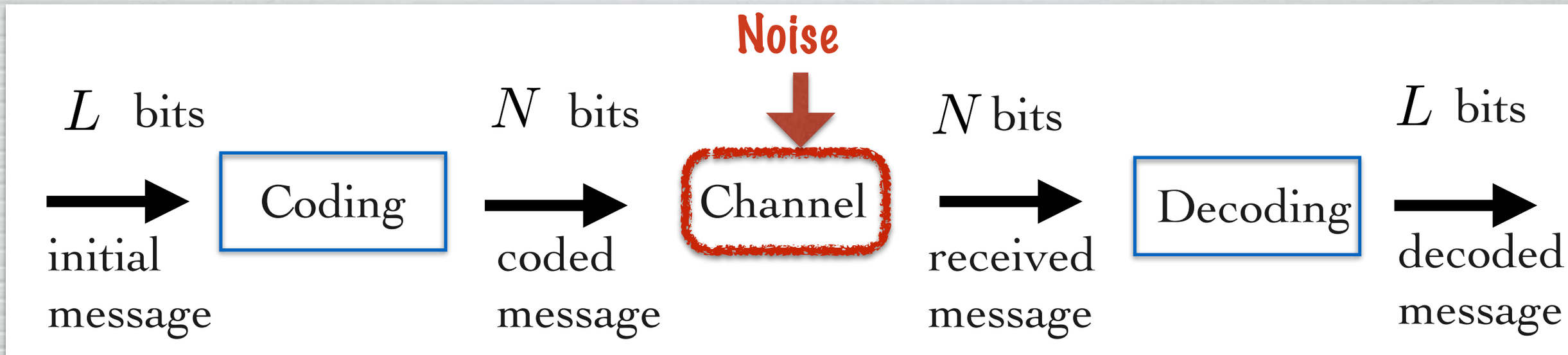
Optimal state, all $s_i = 1$

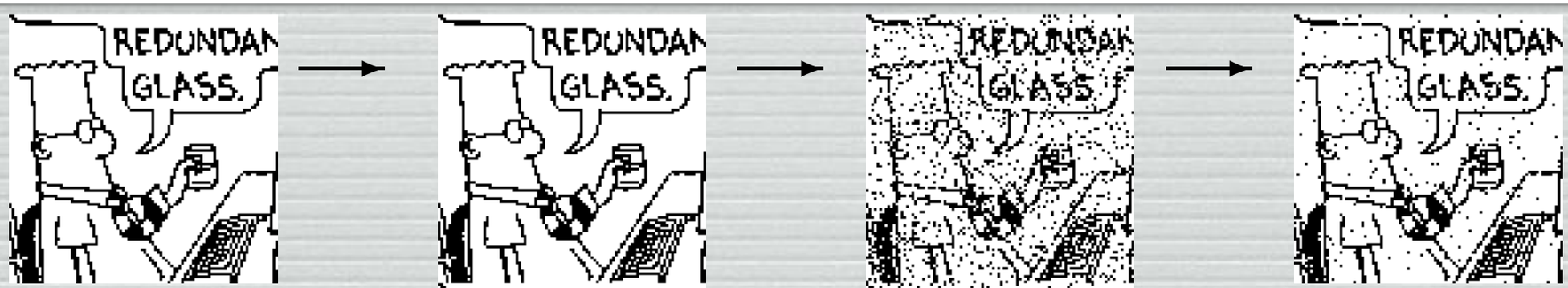


Random first order phase transition at T_c : traps

2nd example : error correction

Coding = add redundancy. $N > L$. Rate = L/N



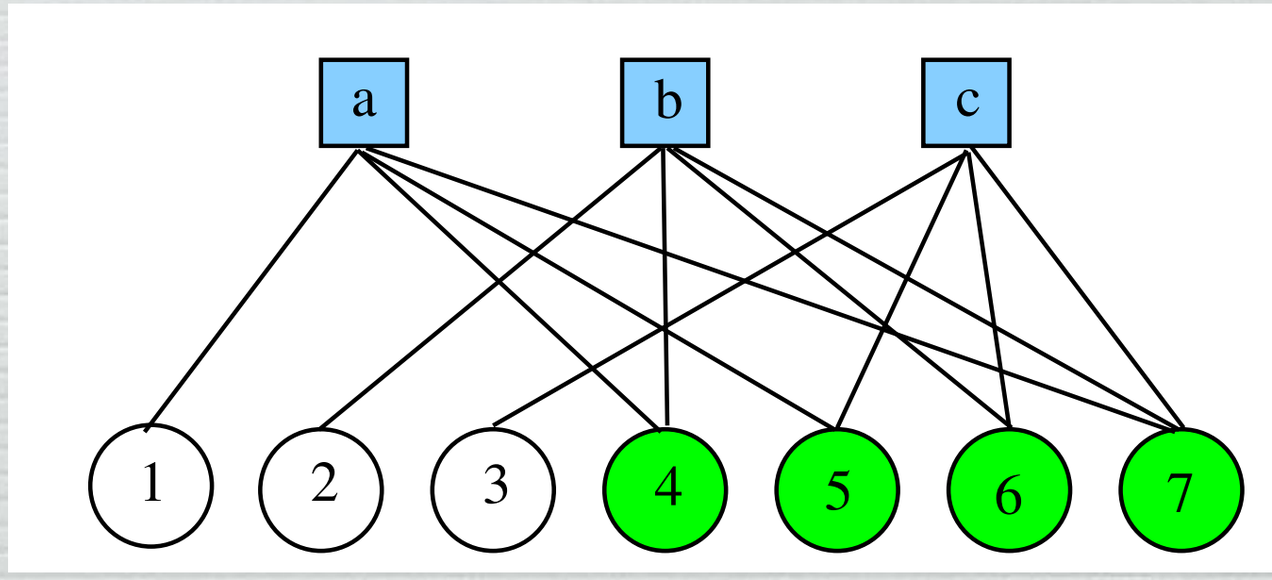


Example : repetition code

Efficient codes : parity checks (LDPC codes)

Add redundancy, with structure allowing to decode

$x_i \in \{0, 1\}$



$$
\begin{aligned}
a: \quad & x_1 + x_4 + x_5 + x_7 = 0 \pmod 2 \\
b: \quad & x_2 + x_4 + x_6 + x_7 = 0 \pmod 2 \\
c: \quad & x_3 + x_5 + x_6 + x_7 = 0 \pmod 2
\end{aligned}
$$

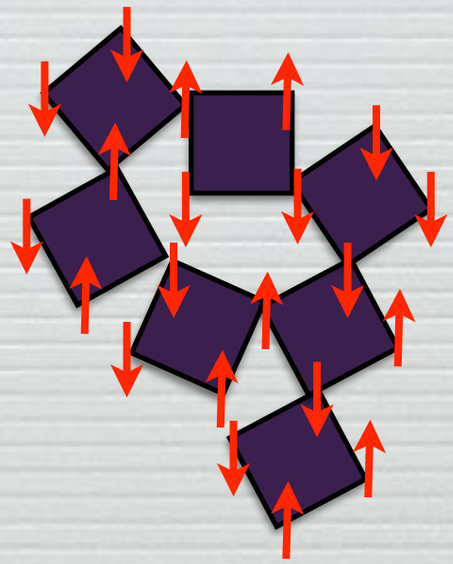
2^4 codewords

among 2^7 words

Efficient codes : parity checks

Add redundancy, with structure allowing to decode

$x_i \in \{0,1\}$



$s_i = (-1)^{x_i}$

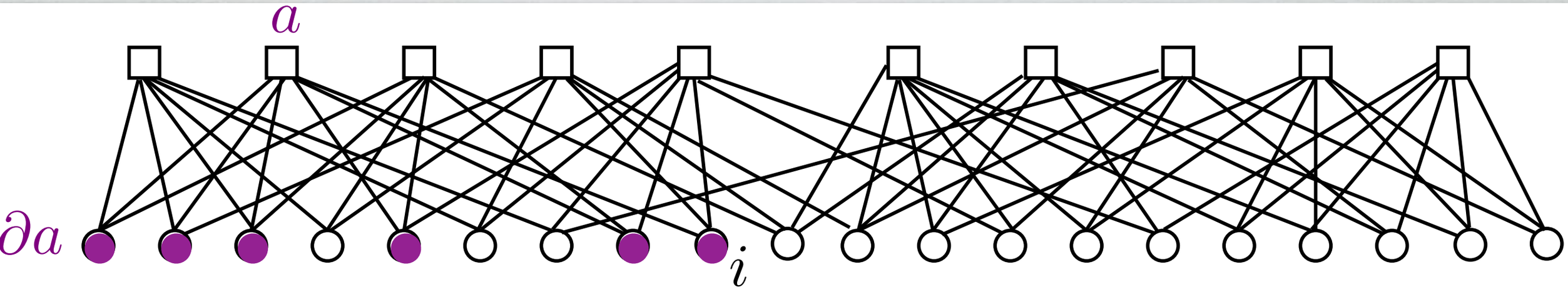
$s_i \in \{\pm 1\}$

$$a: \quad x_1 + x_4 + x_5 + x_7 = 0 \pmod 2 \qquad s_1 s_4 s_5 s_7 = 1$$

$$b: \quad x_2 + x_4 + x_6 + x_7 = 0 \pmod 2 \qquad s_2 s_4 s_6 s_7 = 1$$

$$c: \quad x_3 + x_5 + x_6 + x_7 = 0 \pmod 2 \qquad s_3 s_5 s_6 s_7 = 1$$

Error decoding: « crystal hunting » inference problem

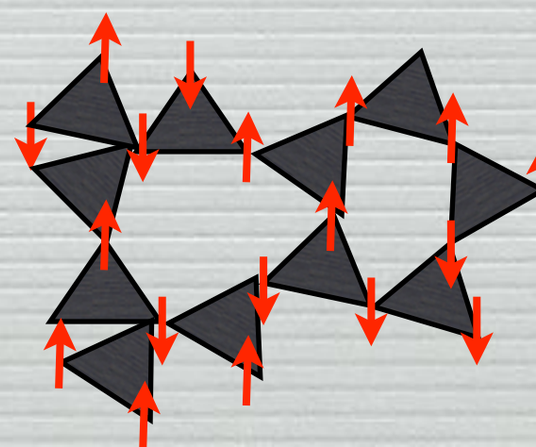


$$P(x_1,\ldots,x_N|\underbrace{y_1,\ldots,y_N}_{\text{received}}) = \frac{1}{Z}\prod_i \psi_i(x_i|y_i) \prod_a \mathbb{I}\left(\sum_{i\in\partial a} x_i = 0 \ (mod\ 2)\right)$$

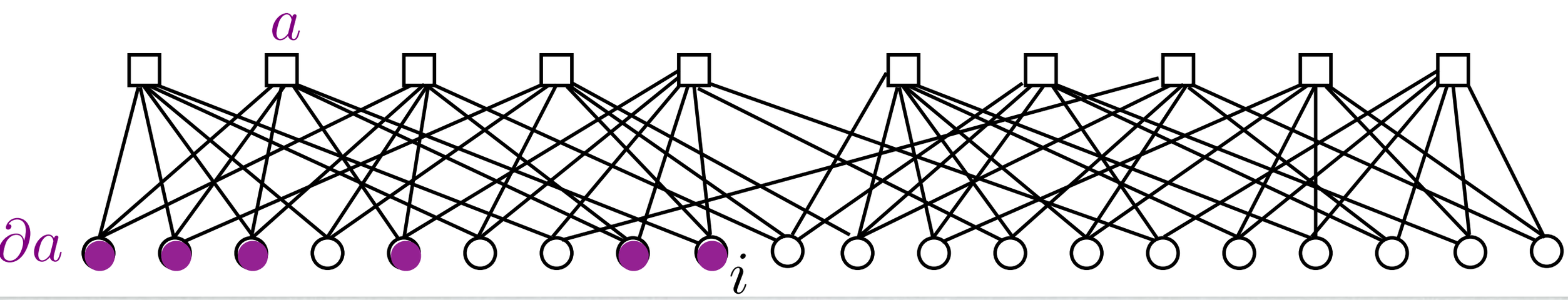
A priori knowledge of the channel

Parity check constraints

Spin glass problem with multispin interactions, discontinuous glass transition (1 step RSB)



Error decoding: inference problem



$$P(x_1,\ldots,x_N|y_1,\ldots,y_N) = \frac{1}{Z}\prod_i \psi_i(x_i|y_i) \prod_a \mathbb{I}\left(\sum_{i\in\partial a} x_i = 0 \ (mod\ 2)\right)$$

One possible decoding algorithm: use belief-propagation mean-field equations relating the local fields $h_{i\setminus j}$ (see later)

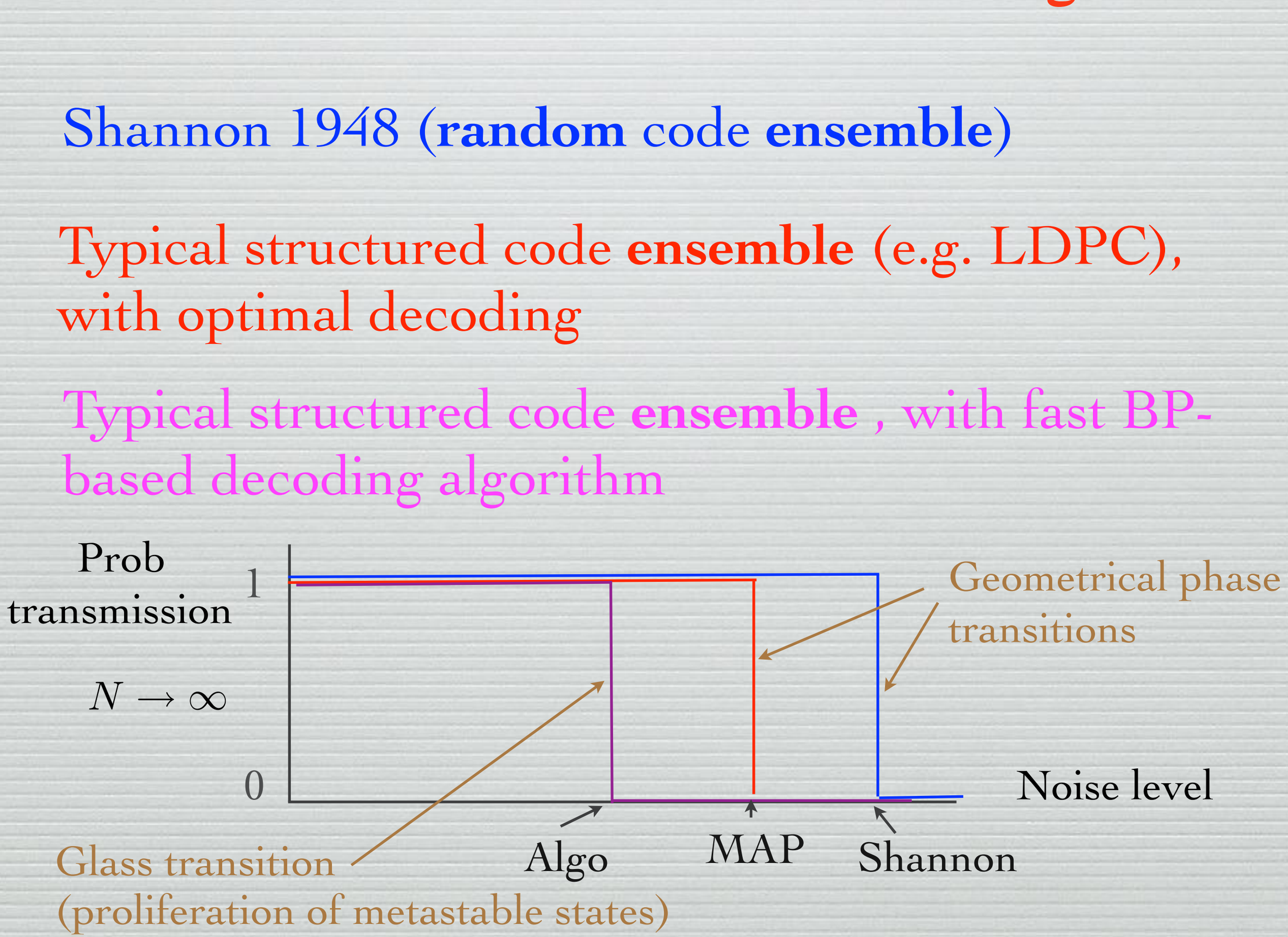
Solve them iteratively (Gallager)

Phase Transitions in Error correcting codes

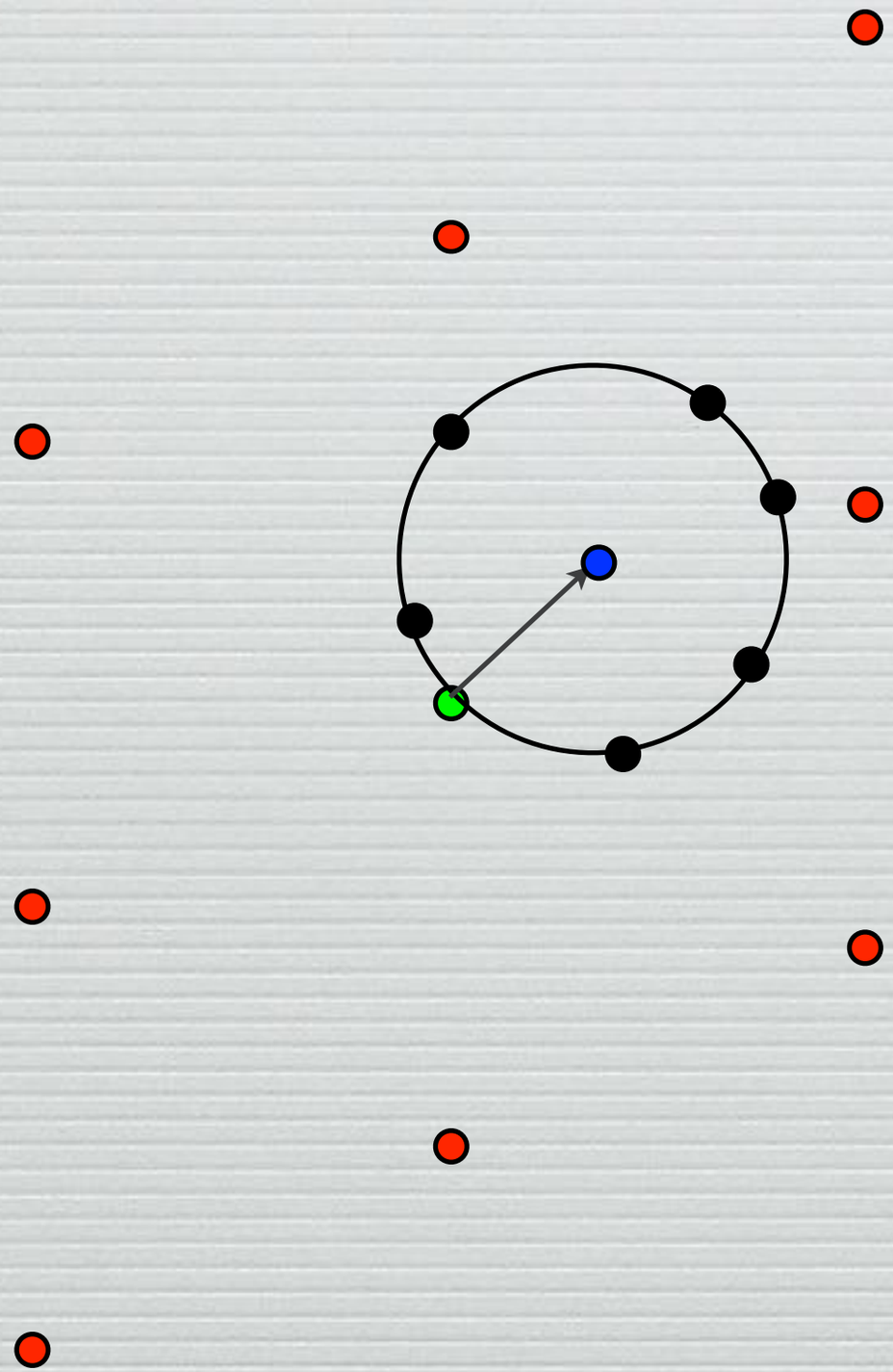
Shannon 1948 (**random** code **ensemble**)

Typical structured code **ensemble** (e.g. LDPC), with optimal decoding

Typical structured code **ensemble** , with fast BP-based decoding algorithm



Error correction: decoding



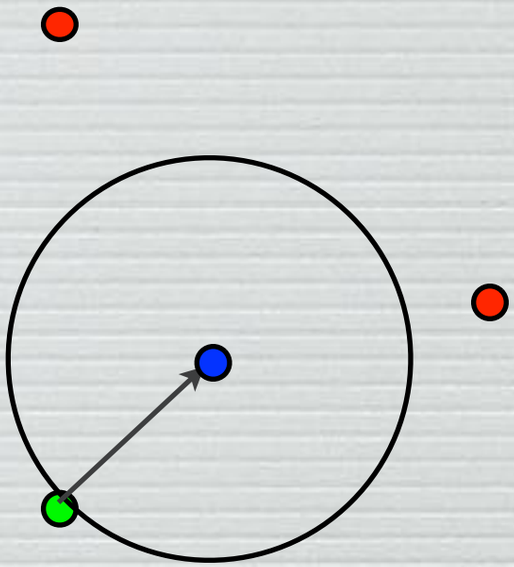
codewords

sent codeword

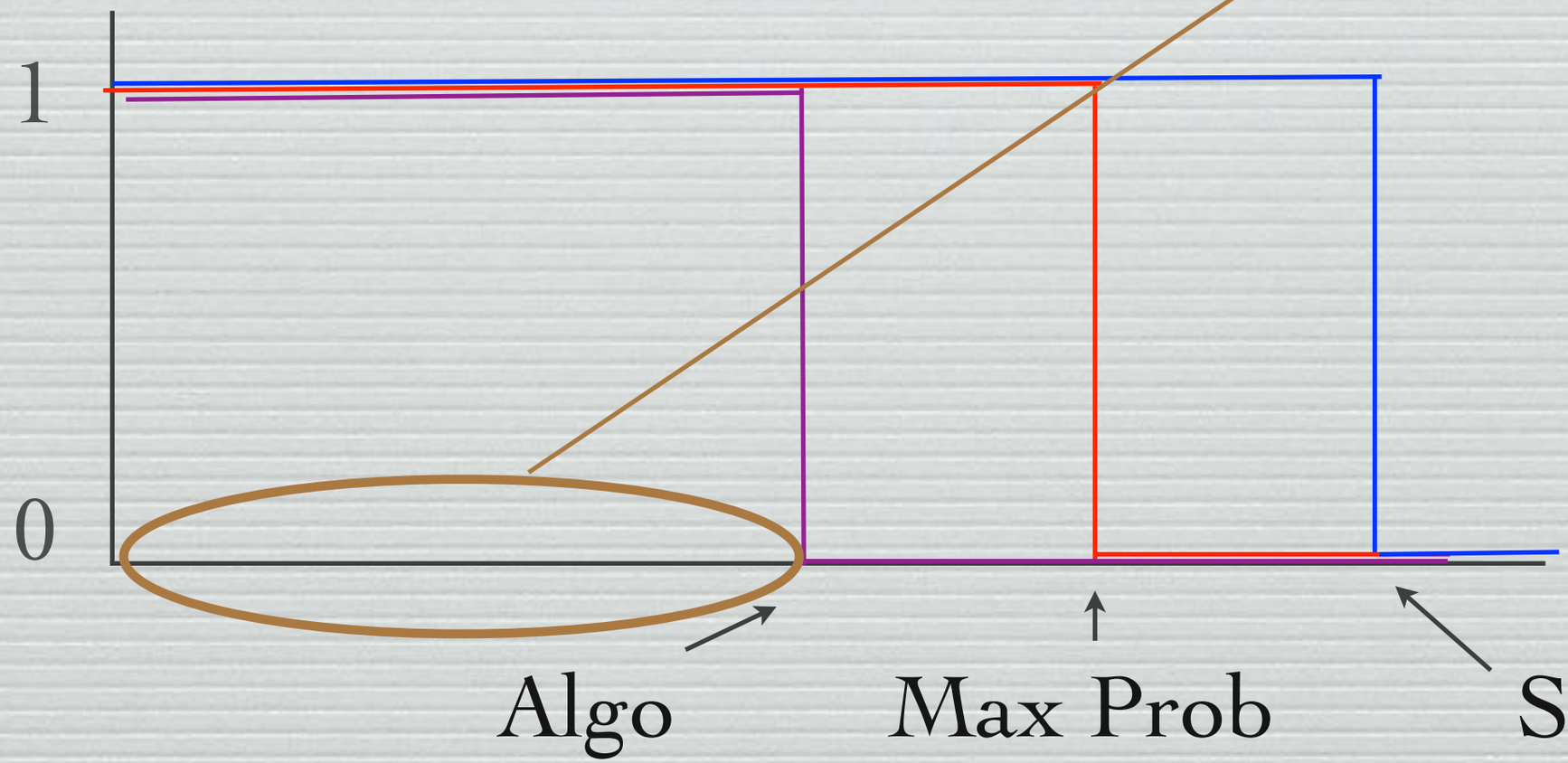
received word

metastable states

Phase transitions in decoding

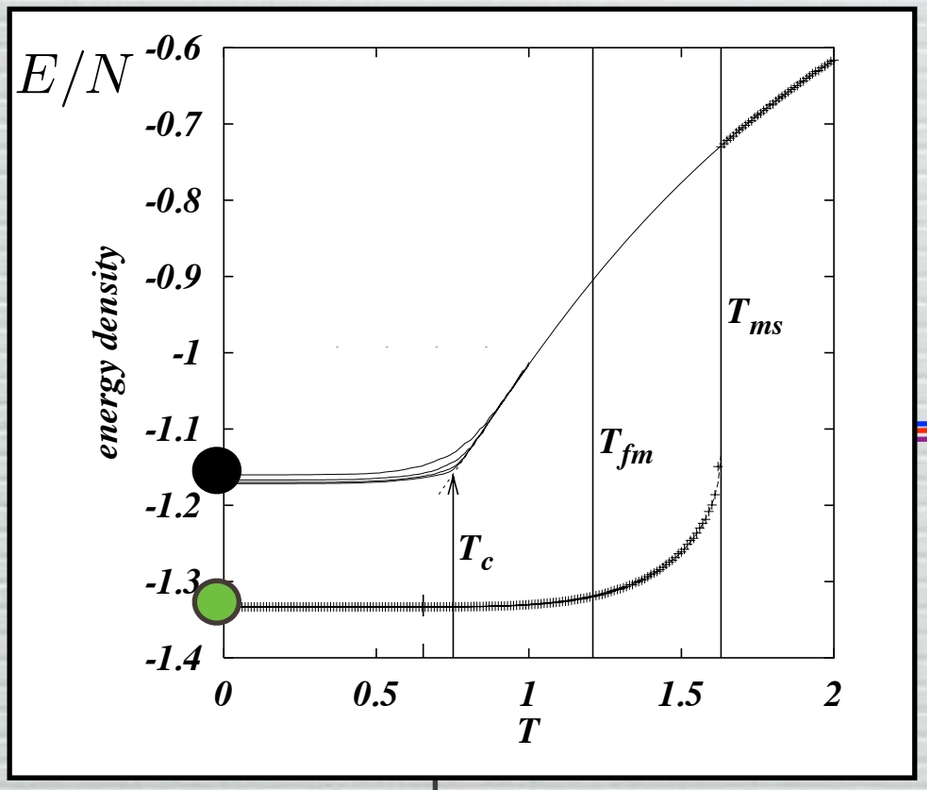


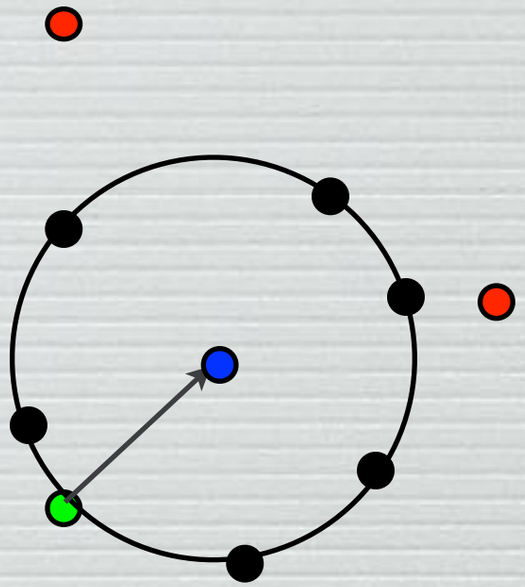
Probability of perfect decoding:



Phase transitions in decoding

Metastables states=traps





0

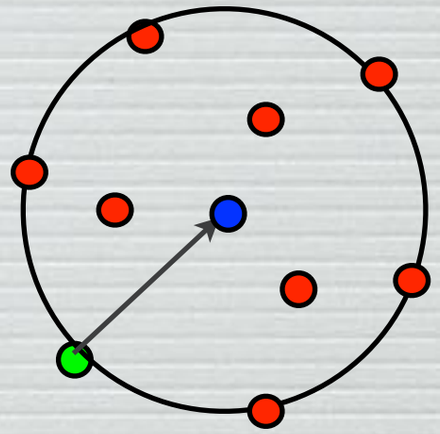
p = noise

Algo

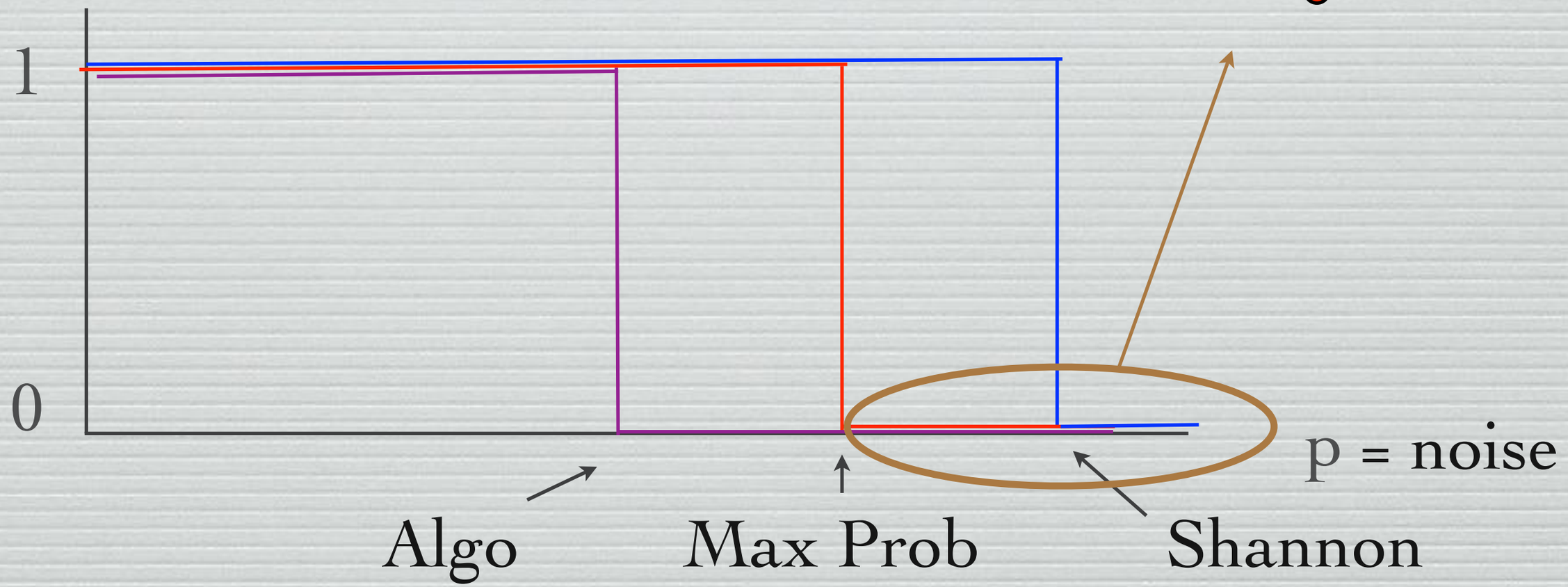
Max Prob

Shannon

Phase transitions in decoding



Probability of perfect decoding:

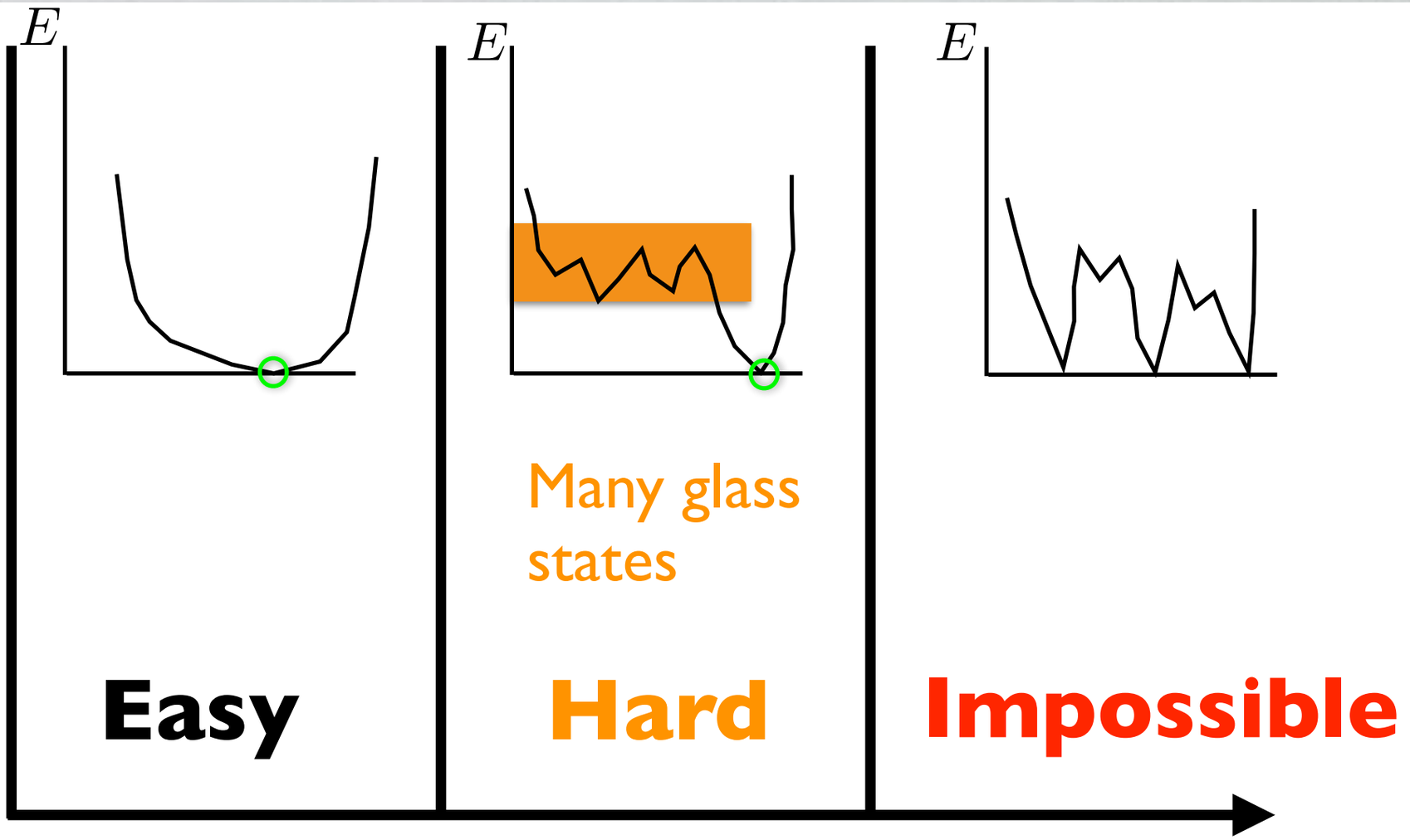


Planted models in Bayes-optimal setting

1-step transition

- q-colouring
- >2 communities
- error correction

Continuous transition: go directly from easy to impossible (eg planted SK)



Q 3: Easy to hard is a dynamical phase transition. Same in BP, simulated annealing. How universal?

NB: Crystal nucleation time can be changed by changing the dynamics (Berthier Biroli, Bouchaud, Tarjus 2018)

Inference with many unknowns : « crystal hunting » with mean-field based algorithms

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

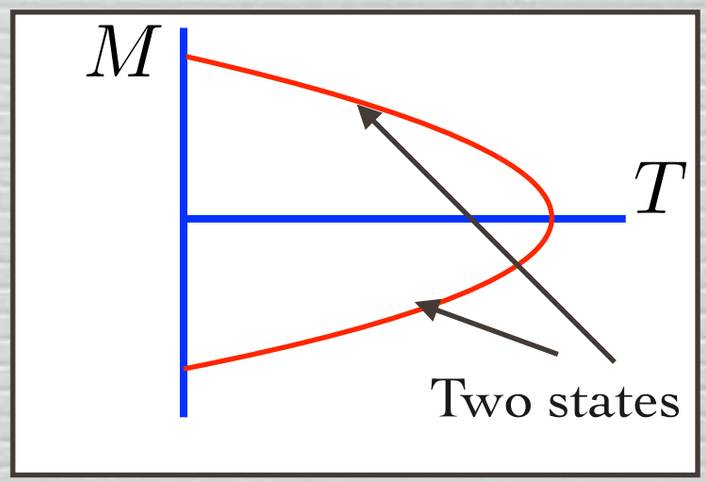
Paul Langevin (1905): $M = M_0\, L\left(\frac{B}{T}\right) \quad ; \quad L(x) = \coth x - 1/x$

Pierre Weiss (1907):

$$B = B_{ext} + \alpha M$$

Spontaneous magnetization in zero external field:

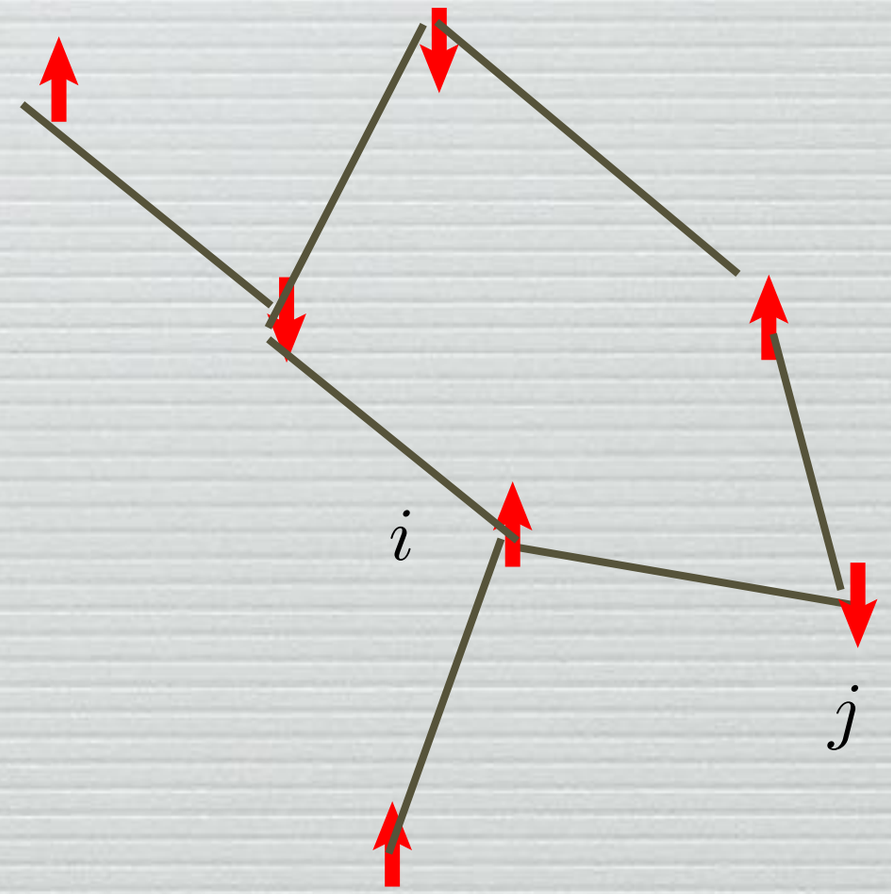
$$M = M_0 L\left(\frac{\alpha M}{T}\right)$$



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:

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

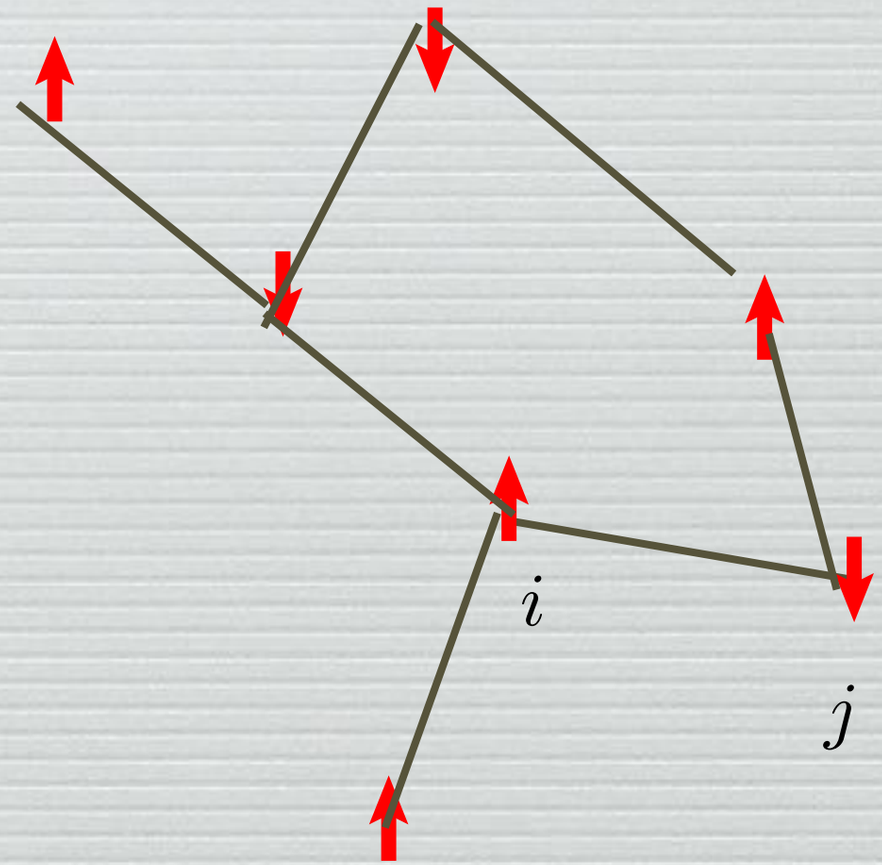
Does not close on $\langle s_i \rangle$

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)

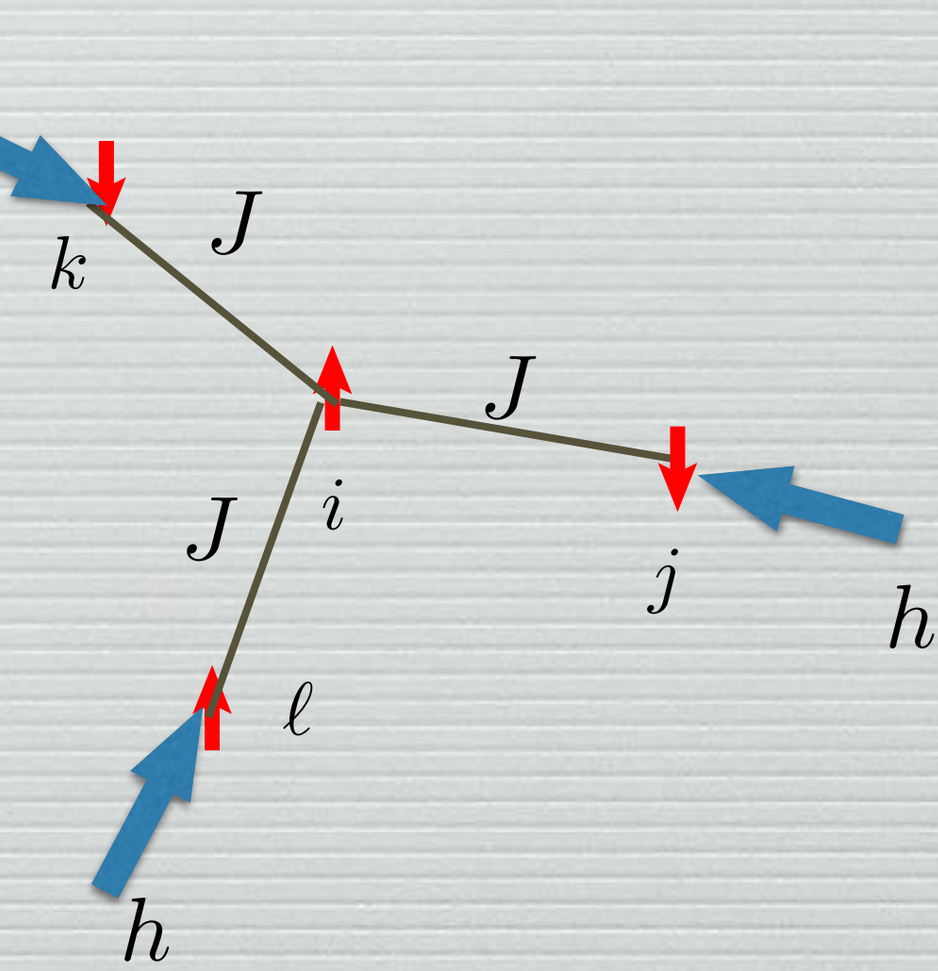
Exact solution for central spin and its neighbors, themselves independent

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

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

$$h = \frac{z-1}{\beta} \mathrm{atanh}[\tanh(\beta J)\tanh(\beta h)]$$

$$M = \tanh\left(z\ \mathrm{atanh}[\tanh(\beta J)\tanh(\beta h)]\right)$$



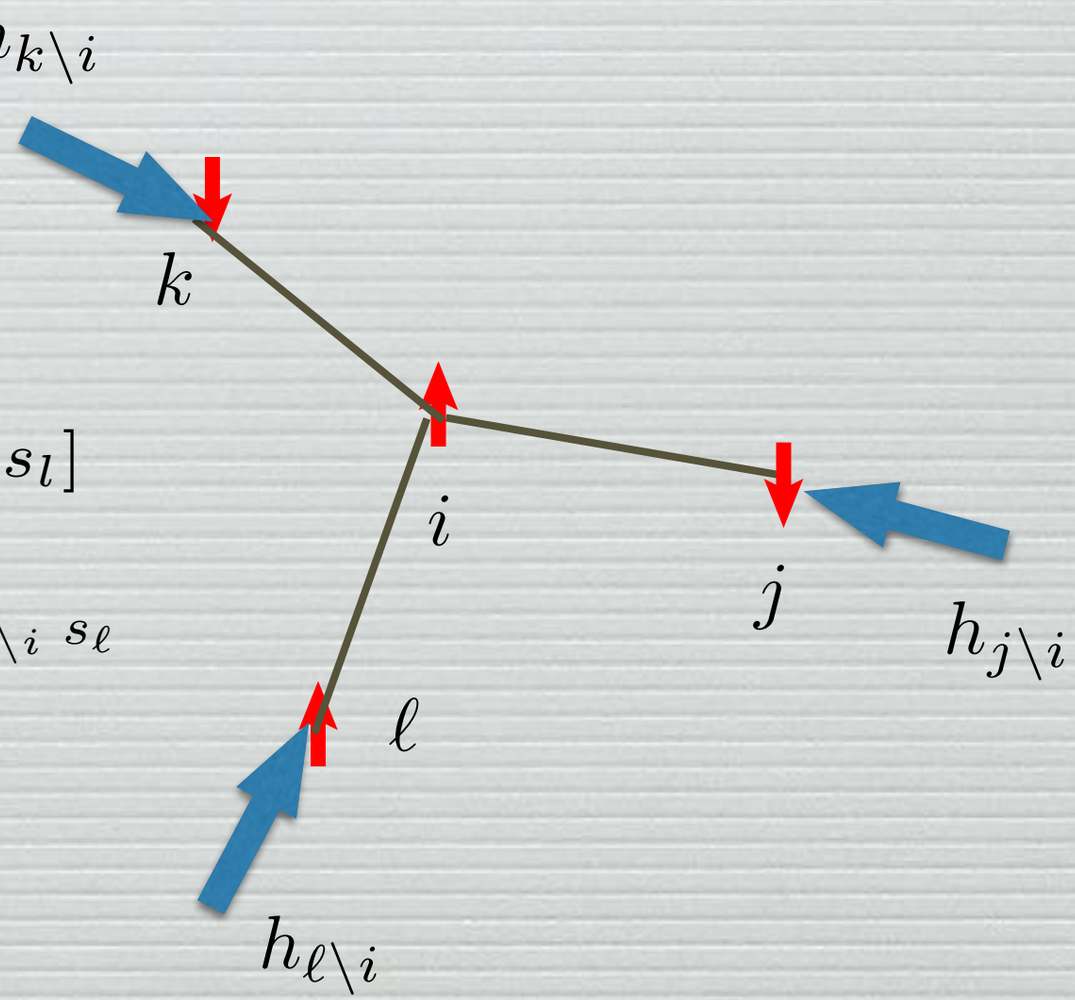
Bethe-Peierls adapted to disordered case

Exact solution for central spin and its neighbors, themselves independent

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

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

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

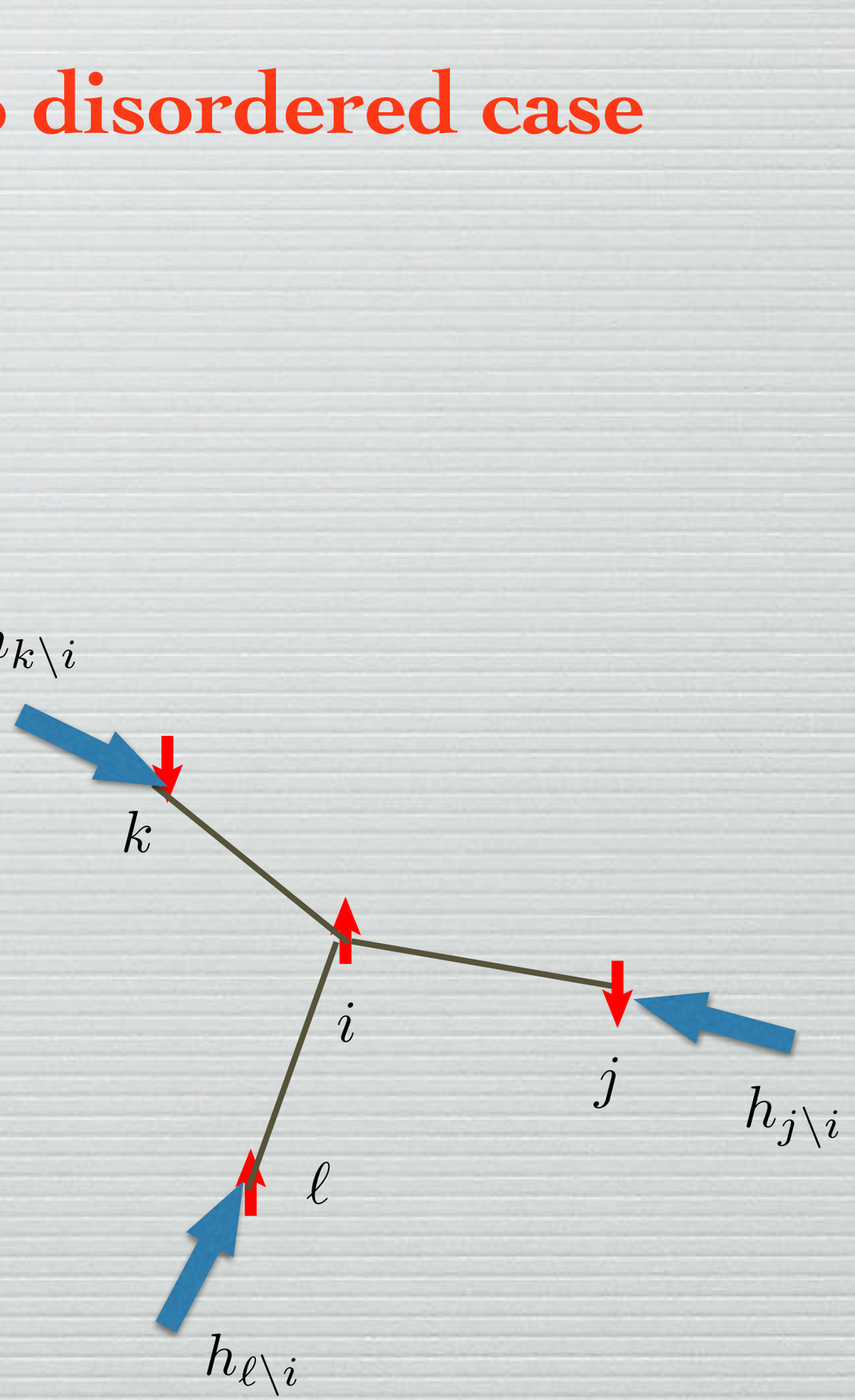


Bethe-Peierls adapted to disordered case

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

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

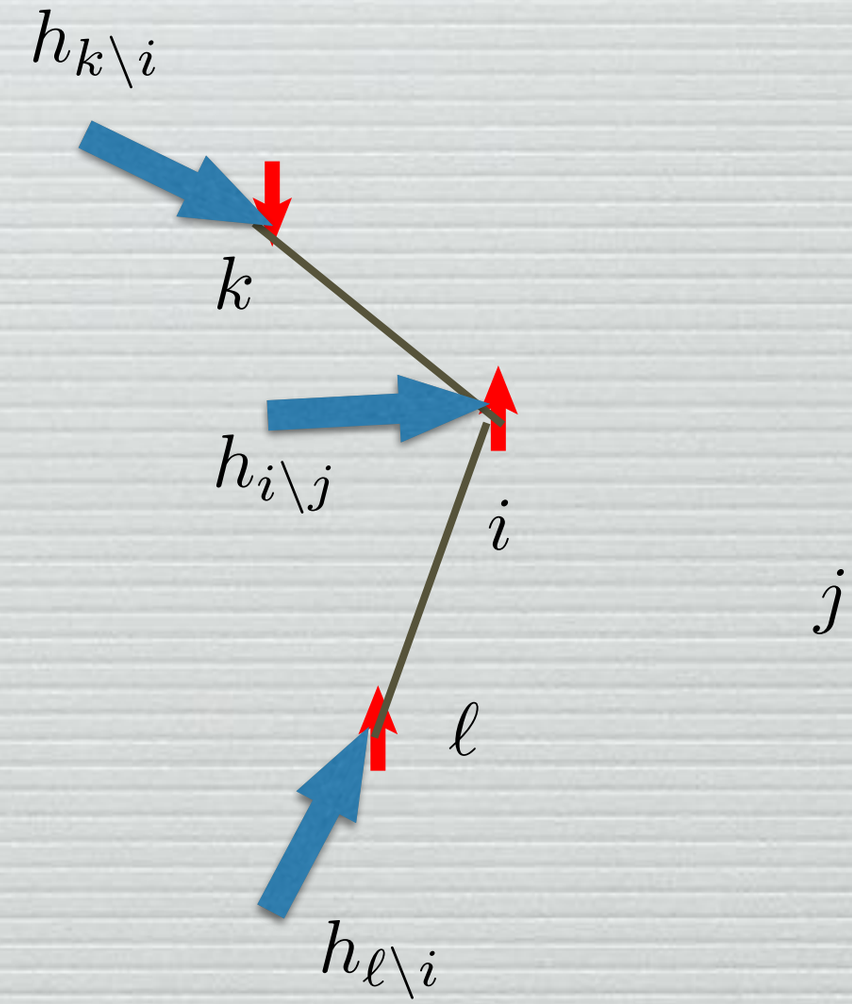
$$+\frac{1}{\beta}\mathrm{atanh}[\tanh(\beta J_{\ell i})\tanh(\beta h_{\ell\setminus i})]$$



Bethe-Peierls adapted to disordered case

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

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



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

« **Belief propagation** » algorithm: **iterate** these equations

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

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

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

State evolution follows the mapping $P^{t+1}(h) = F[P^{t}(h)]$ generated by the BP iteration, and seeks a fixed point distribution

Mean-field algorithms for inference

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

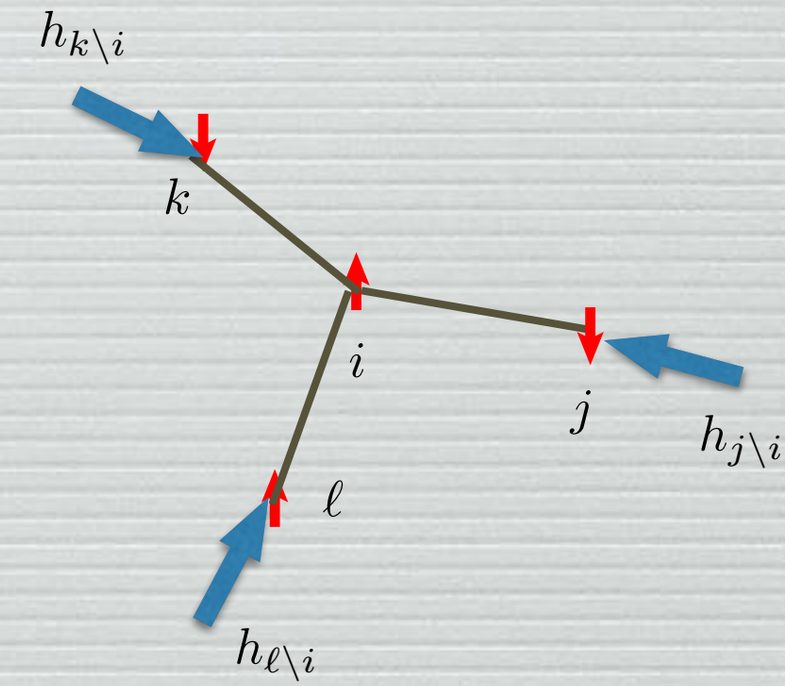
Mean-field algorithms for inference

2) When is BP exact?

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

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)



Mean-field algorithms for inference

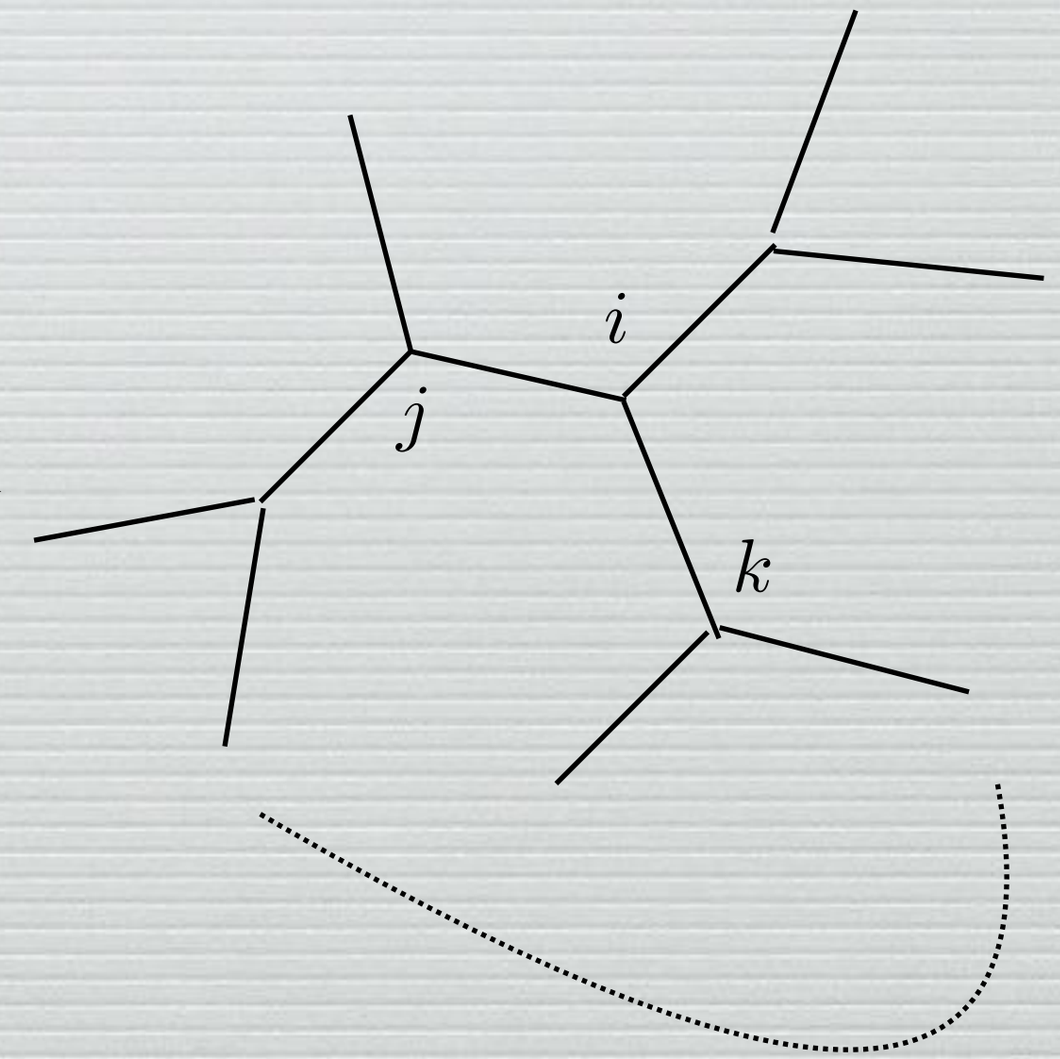
2) When is BP exact?

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

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)

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

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

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

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

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

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 ?

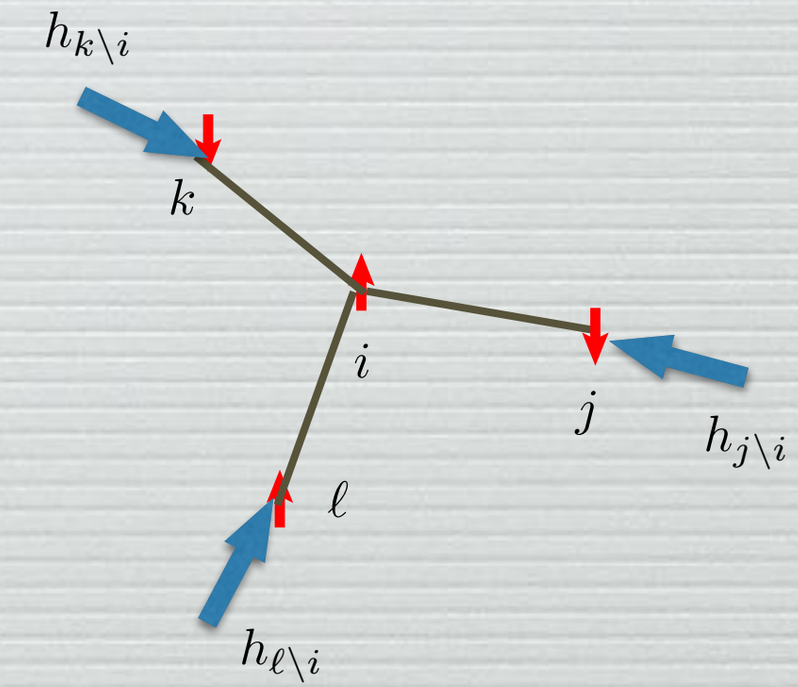
3) What happens in a glass phase, when there are many pure states, and therefore many solutions ?

Real variables

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

becomes

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



BP messages are cavity probability densities of the local variables.
Simple case : large connectivity $p_{i\setminus j}(x_i)$ approximately Gaussian
Generalized Approximate Message Passing (GAMP - MM1989, Rangan 2010,…)

Three important developments

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

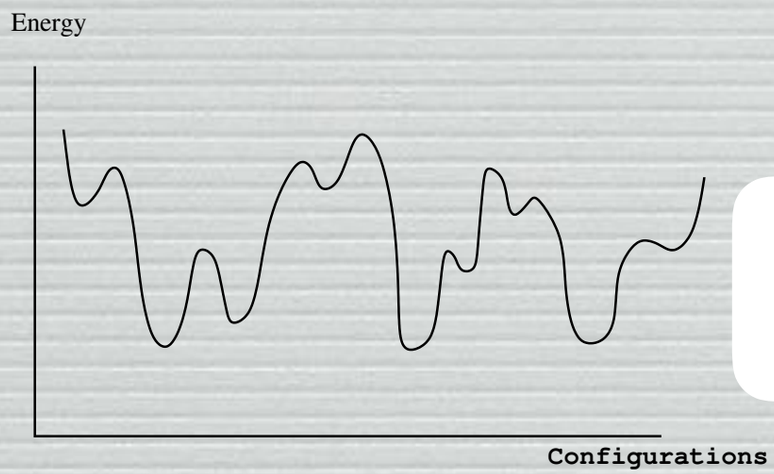
2) What happens if the elementary variables (spins) are real instead of discrete ?

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 ?

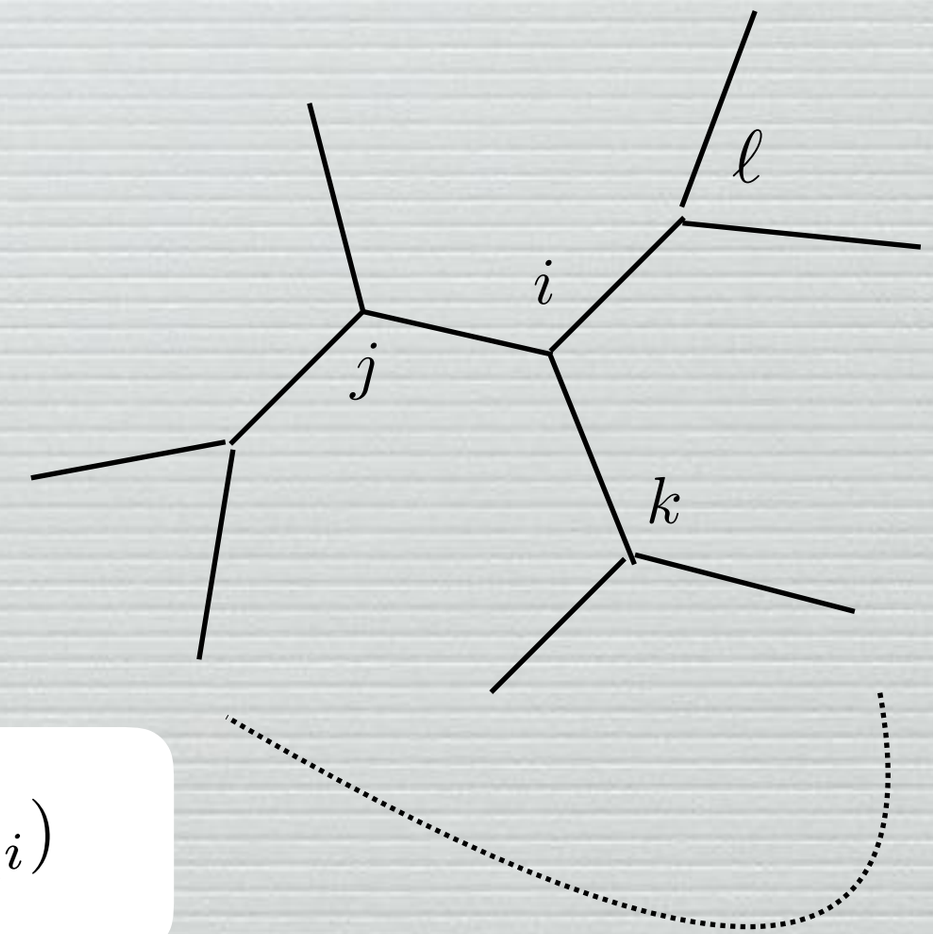
BP equations $$h_{i\setminus j} = f(h_{k\setminus i}, h_{\ell\setminus i})$$

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



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

Glassy phase: many states, many solutions of BP

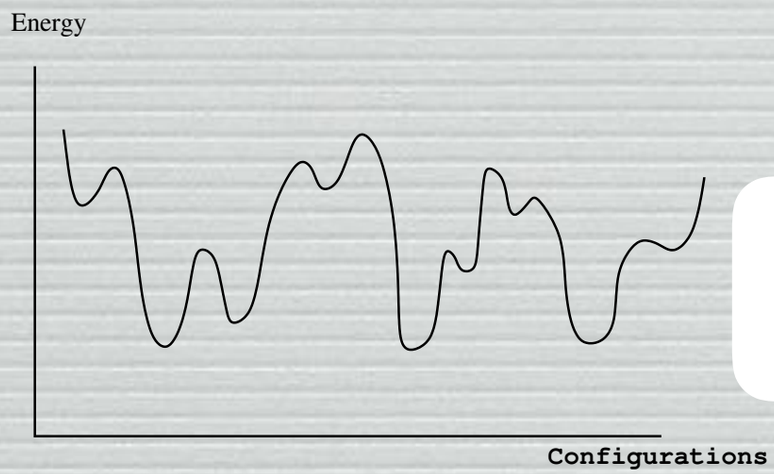


Loop length $O(\log N)$

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

An example of fully connected model: Generalized Linear Regression

The problem of correlations

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

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

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

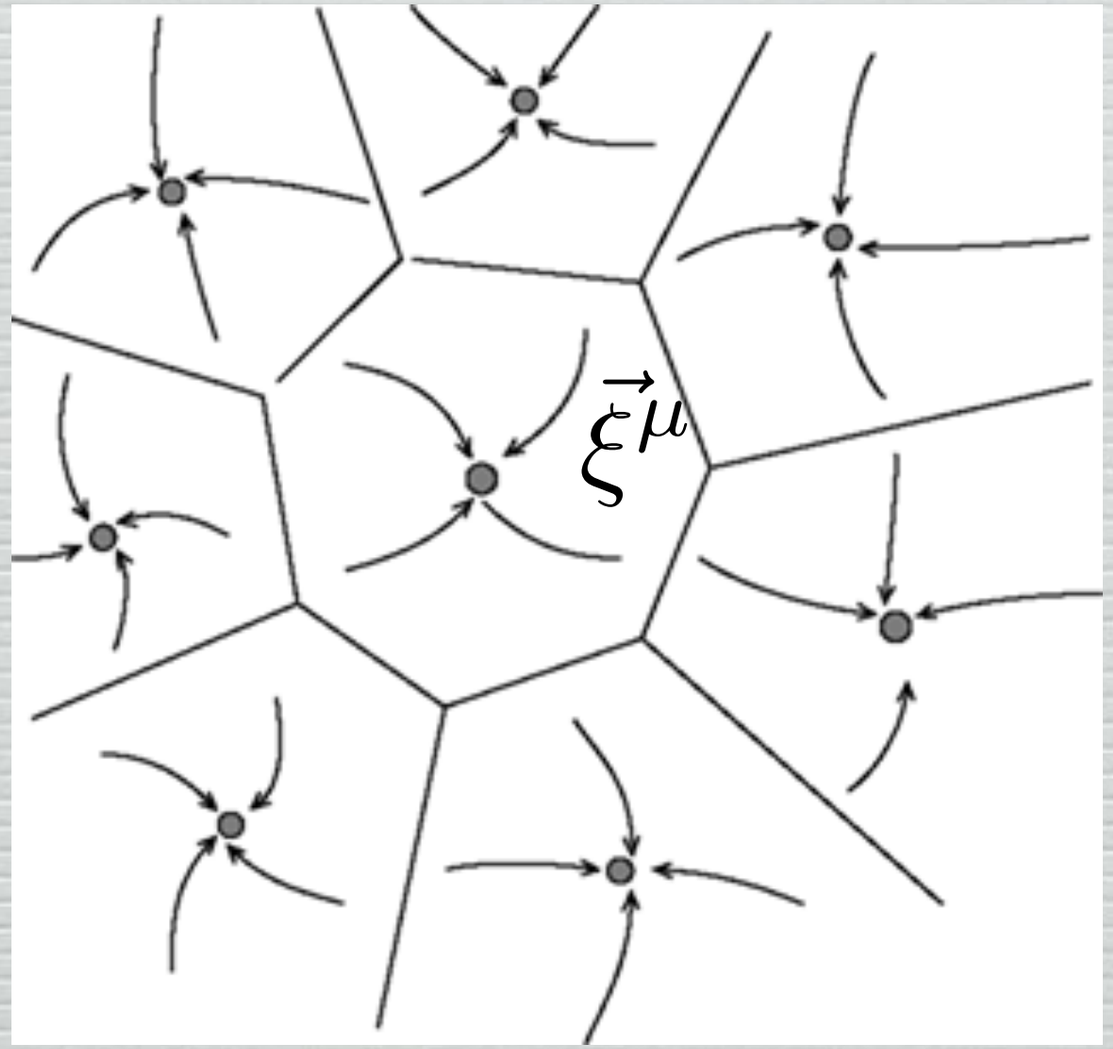
$$G(z) = \mathrm{extr}_\mu \left[\mu z - \int d\lambda D(\lambda) \log(\mu - \lambda) \right] - \log z - 1$$

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

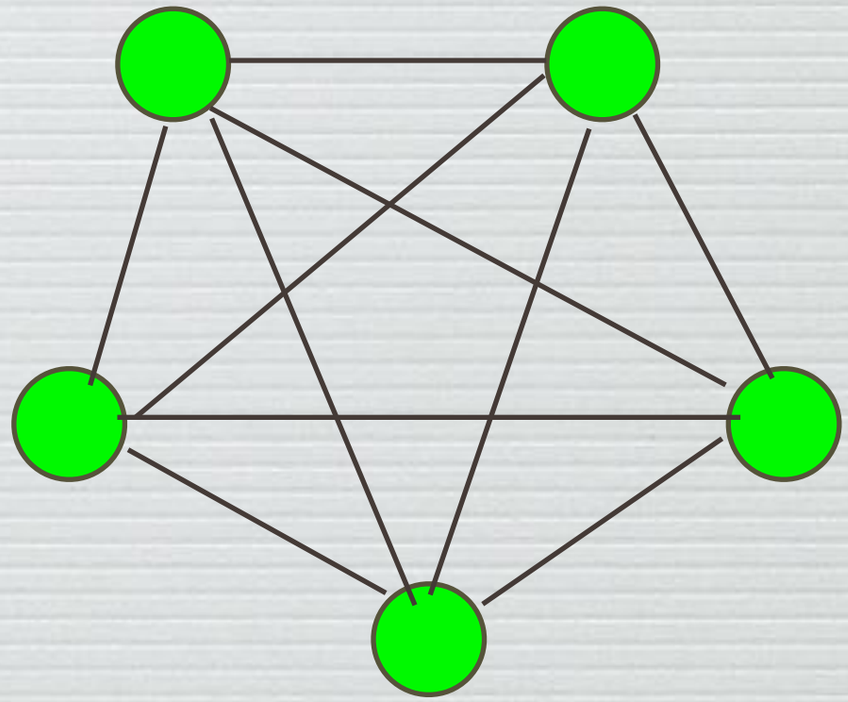
Patterns to be memorized

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

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

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

Hopfield model

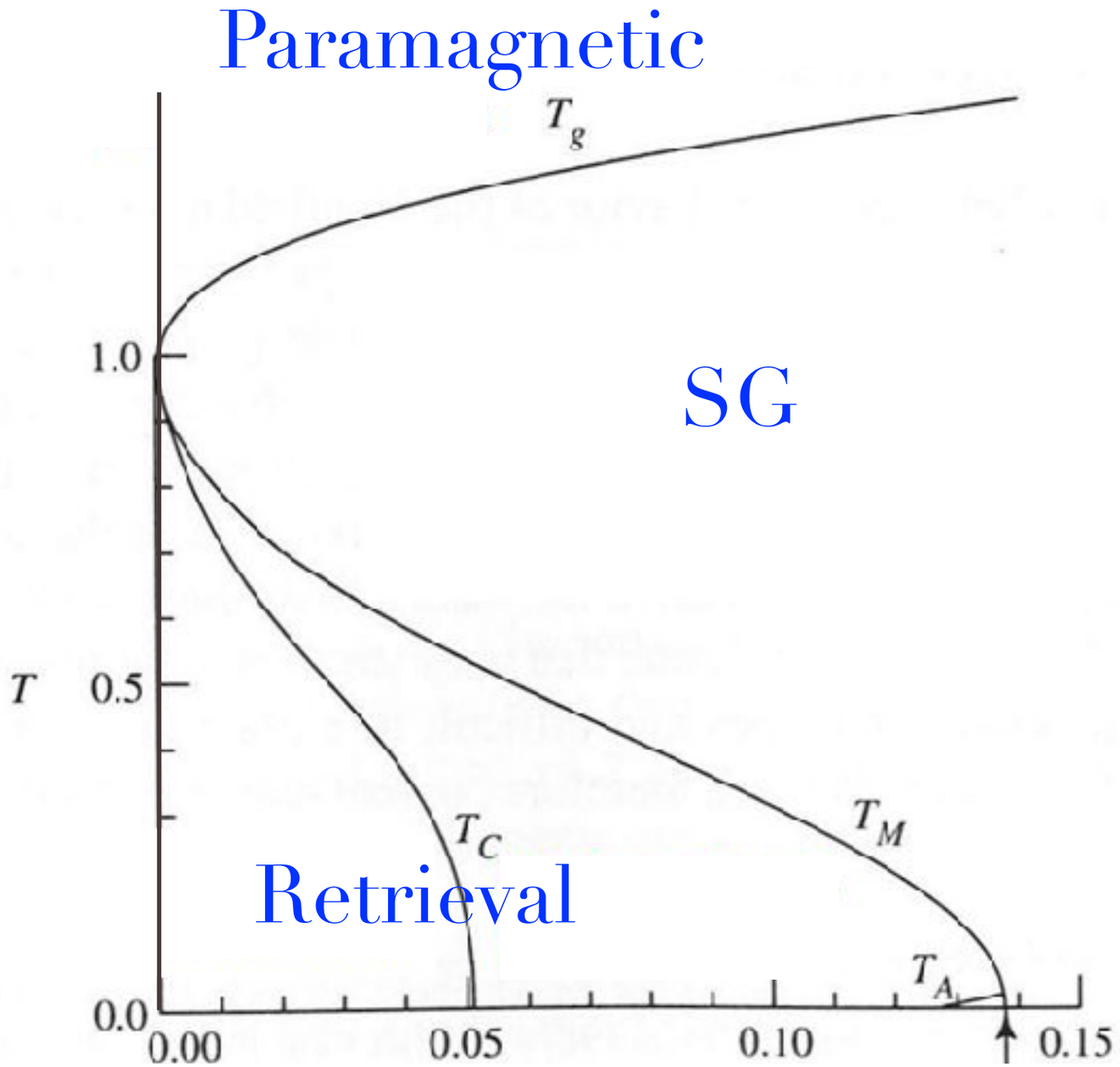


$$E = -\frac{1}{2}\sum_{i,j} J_{ij} s_i s_j \qquad J_{ij} = \frac{1}{N}\sum_{\mu} \xi_i^\mu \xi_j^\mu$$

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

Hopfield model

Phase diagram (Amit Gutfreund Sompolinsky 1985)



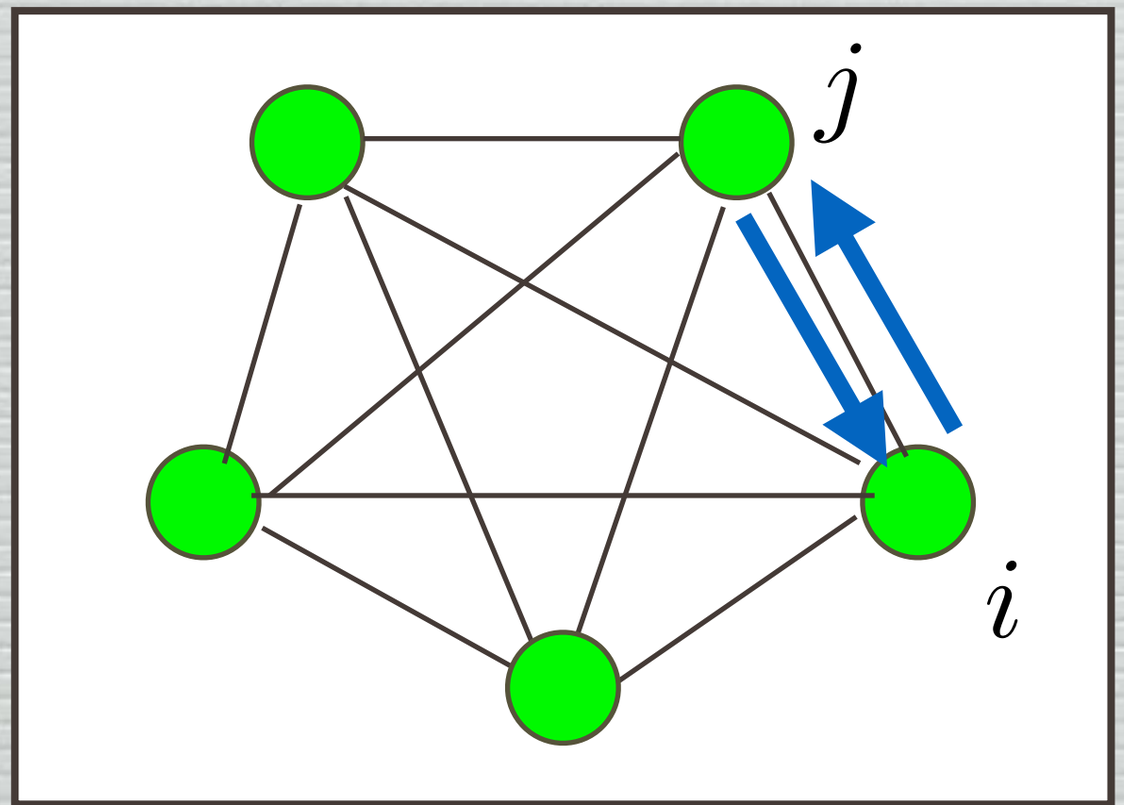
Mean field equations for solving the Hopfield model (find local magnetizations)

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

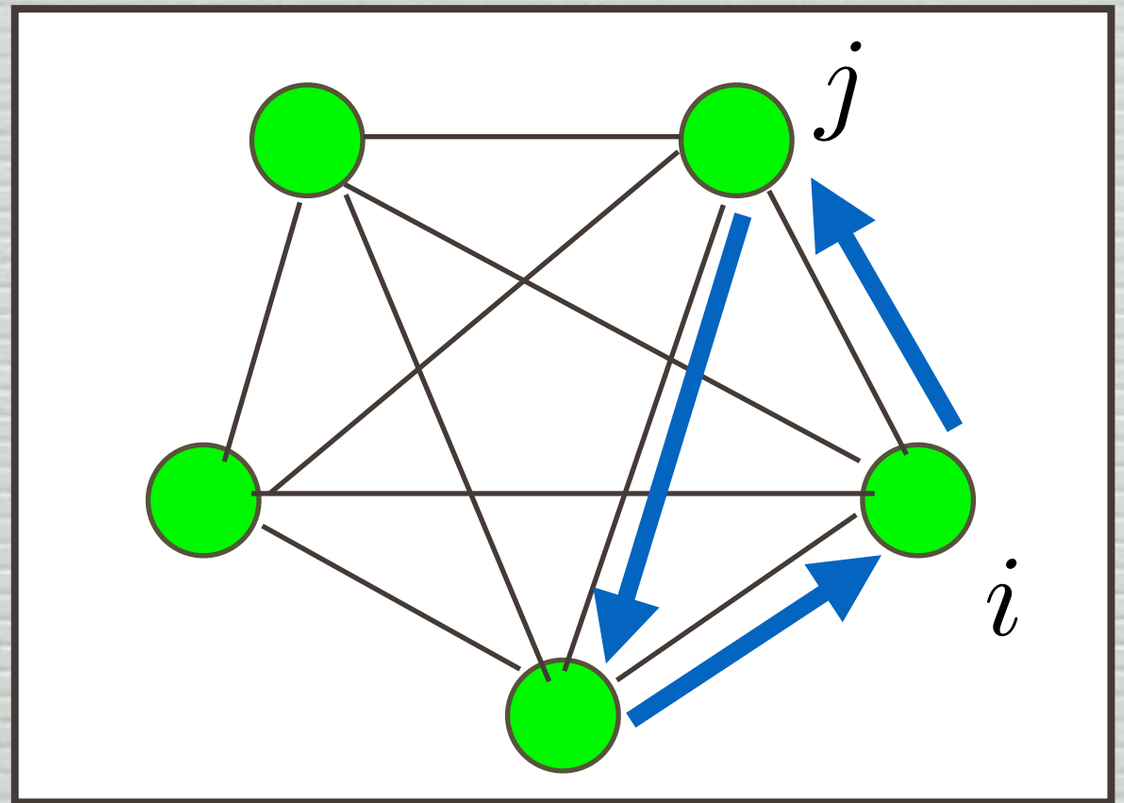


TAP is valid only if indirect interaction from i to j through other sites can be neglected

TAP in the Hopfield model: more subtle!

$$J_{ij} = \frac{1}{N} \sum_{\mu} \xi_i^\mu \xi_j^\mu \qquad \overline{J_{ij} J_{jk} J_{ki}} \neq 0$$

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) :

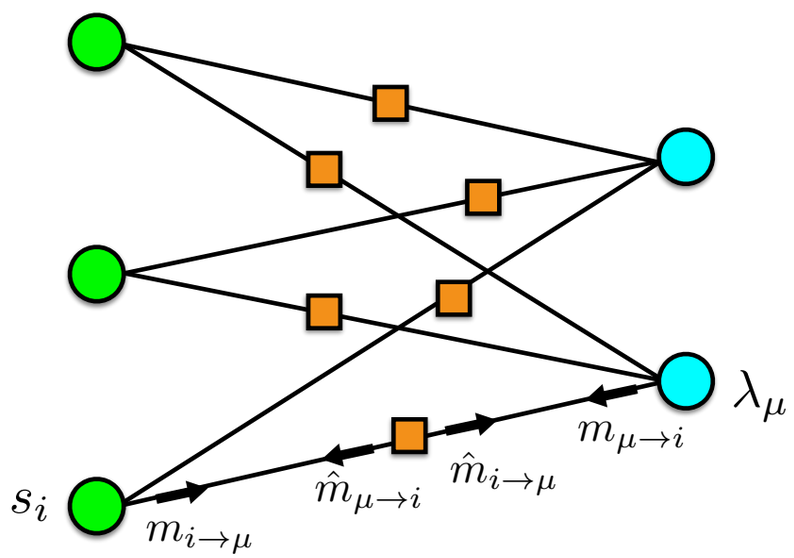
$$Z = \sum_s \int \prod_\mu \frac{d\lambda_\mu}{\sqrt{2\pi\beta}} \exp\left[-\frac{\beta}{2}\sum_\mu \lambda_\mu^2 + \beta \sum_{\mu,i} \frac{\xi_i^\mu}{\sqrt{N}} s_i \lambda_\mu\right]$$

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

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

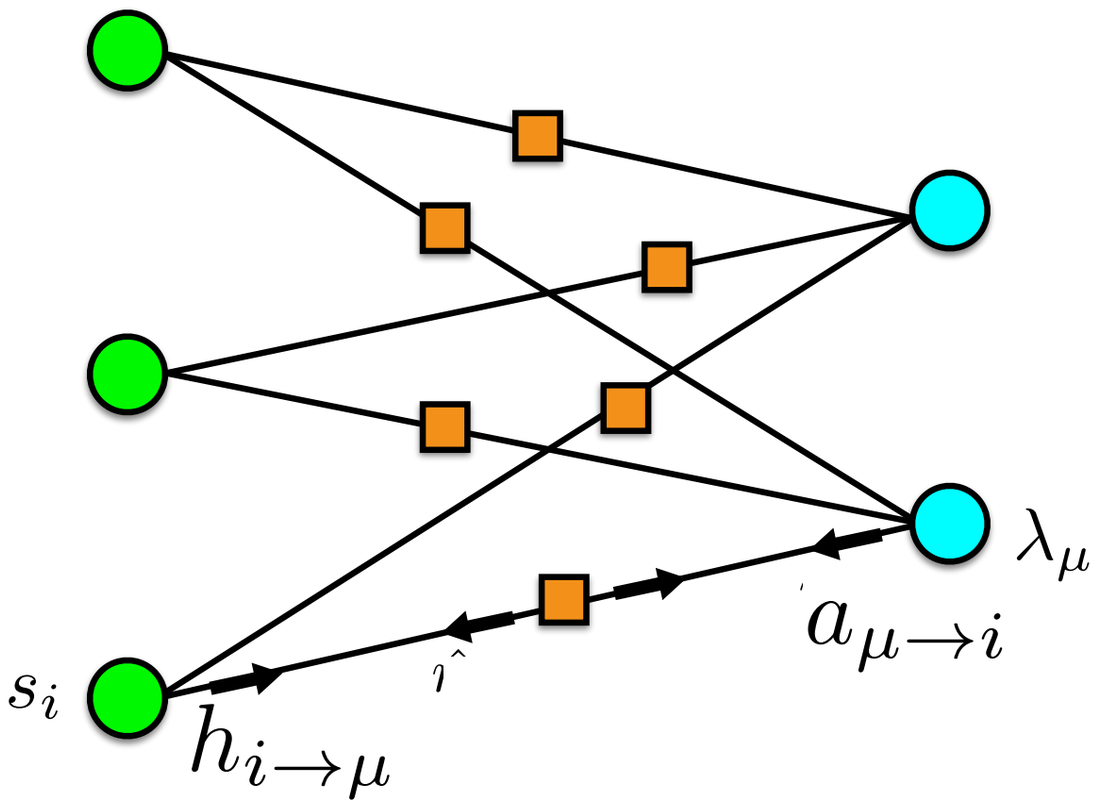
$\Theta(1)$ if uncorrelated

$\Theta(\sqrt{N})$ 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

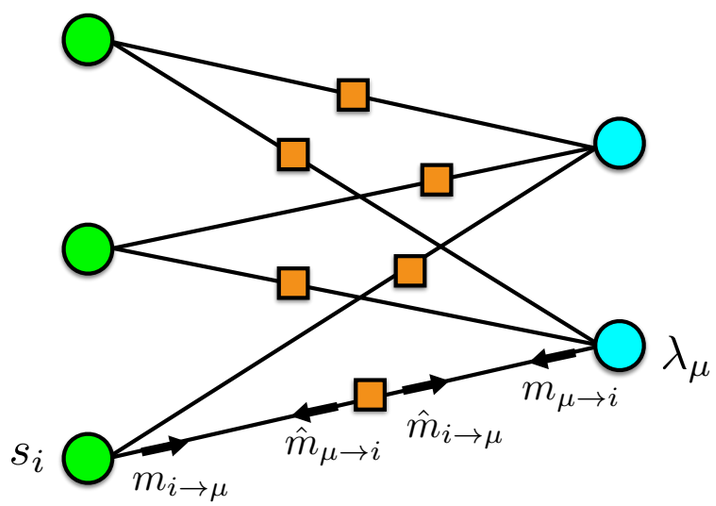
« relaxed BP »

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 $\vec{u}^{r}$

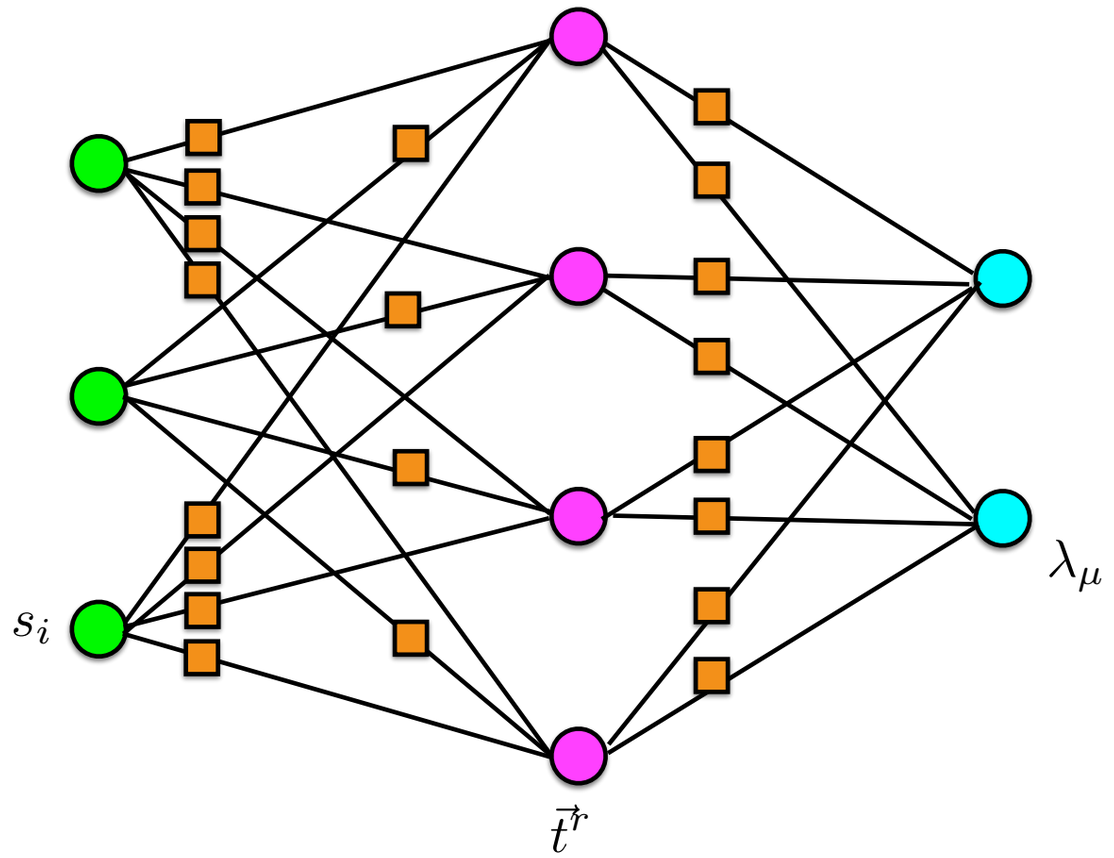
$$\boxed{\vec{\xi}^{\mu} = \frac{1}{\sqrt{\gamma N}} \sum_r v_r^{\mu}\, \vec{u}^{r}}$$, binary $v_r^{\mu} \in \{\pm 1\}$

TAP equations in the Hopfield model with structured patterns

Modified Hopfield model: Combinatorial patterns

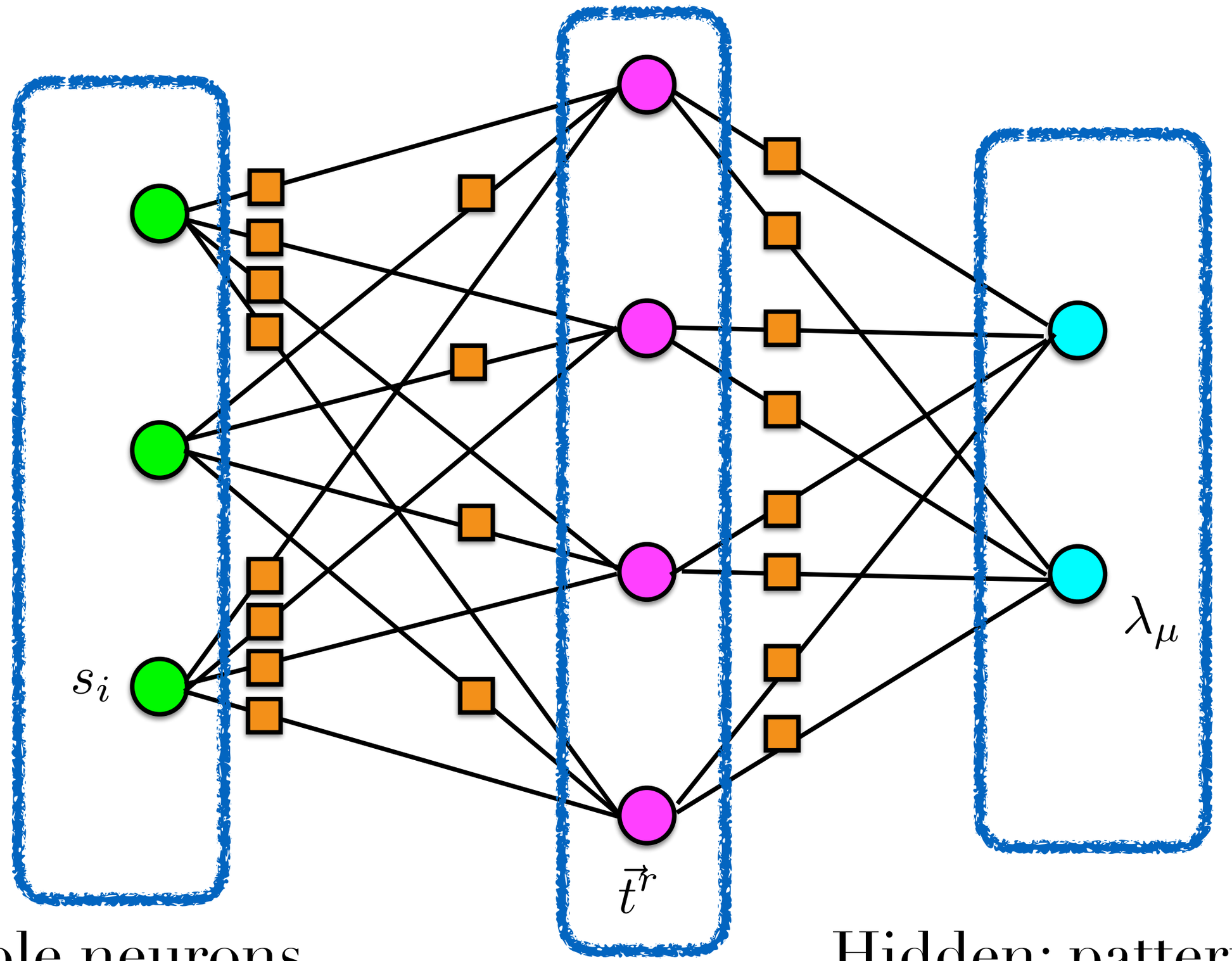
$$Z = \sum_s \int \prod_\mu \frac{d\lambda_\mu e^{-\beta\lambda_\mu^2/2}}{\sqrt{2\pi\beta}} \exp\left[\frac{\beta}{\sqrt{\gamma}} \sum_{r=1}^{\gamma N} \left(\frac{1}{\sqrt{N}} \sum_i u_i^r s_i\right)\left(\frac{1}{\sqrt{N}} \sum_\mu v_\mu^r \lambda_\mu\right)\right]$$

Disentangle the last term by another Hubbard Stratonovitch representation



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

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



Visible neurons

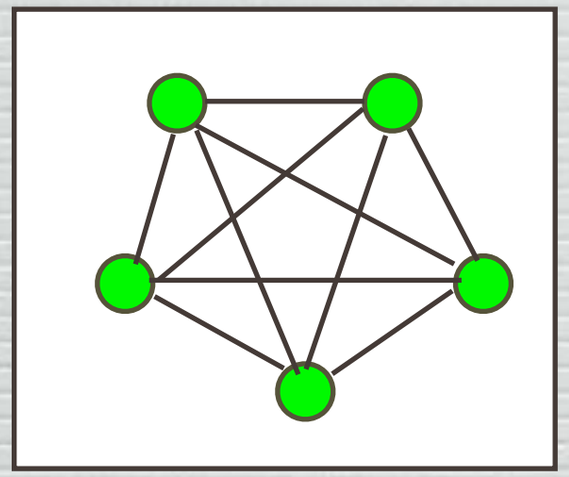
Hidden: features

Hidden: patterns

TAP equations in the Hopfield model with structured patterns

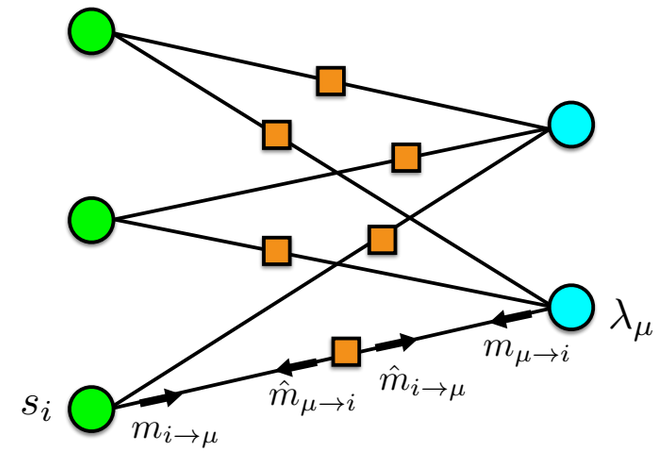
Write the cavity/BP equations. Simplify them to TAP-AMP form, involving: H_i , p_r, A_μ

TAP equations in the Hopfield model with structured patterns



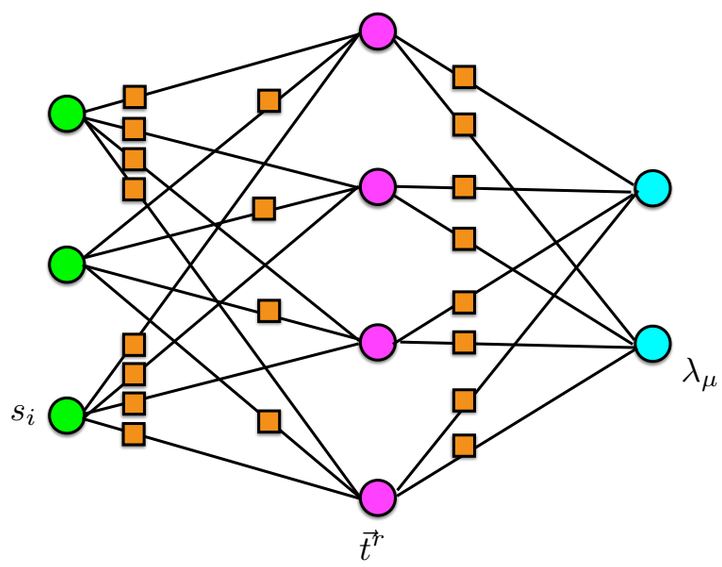
Hopfield model

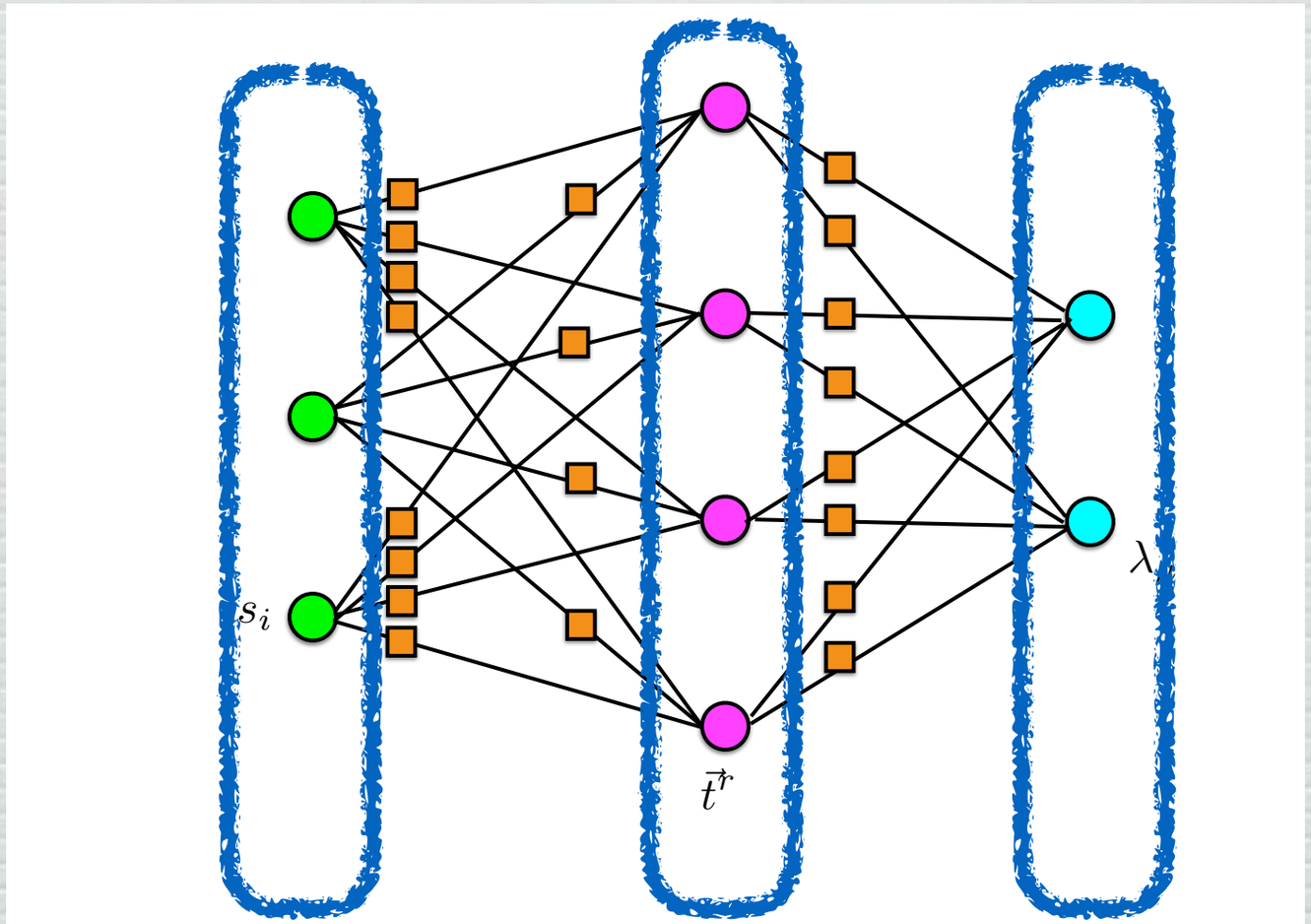
Restricted Boltzmann machine



with combinatorial patterns

Two hidden layers



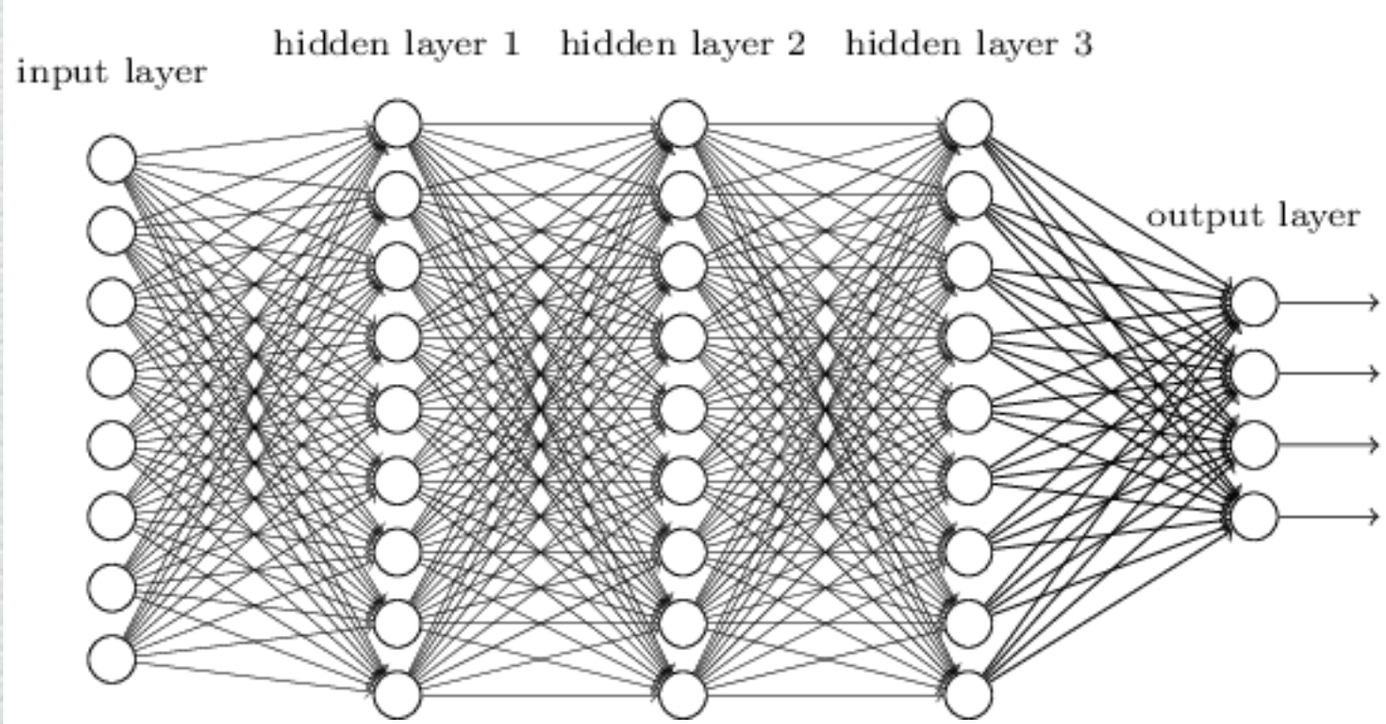


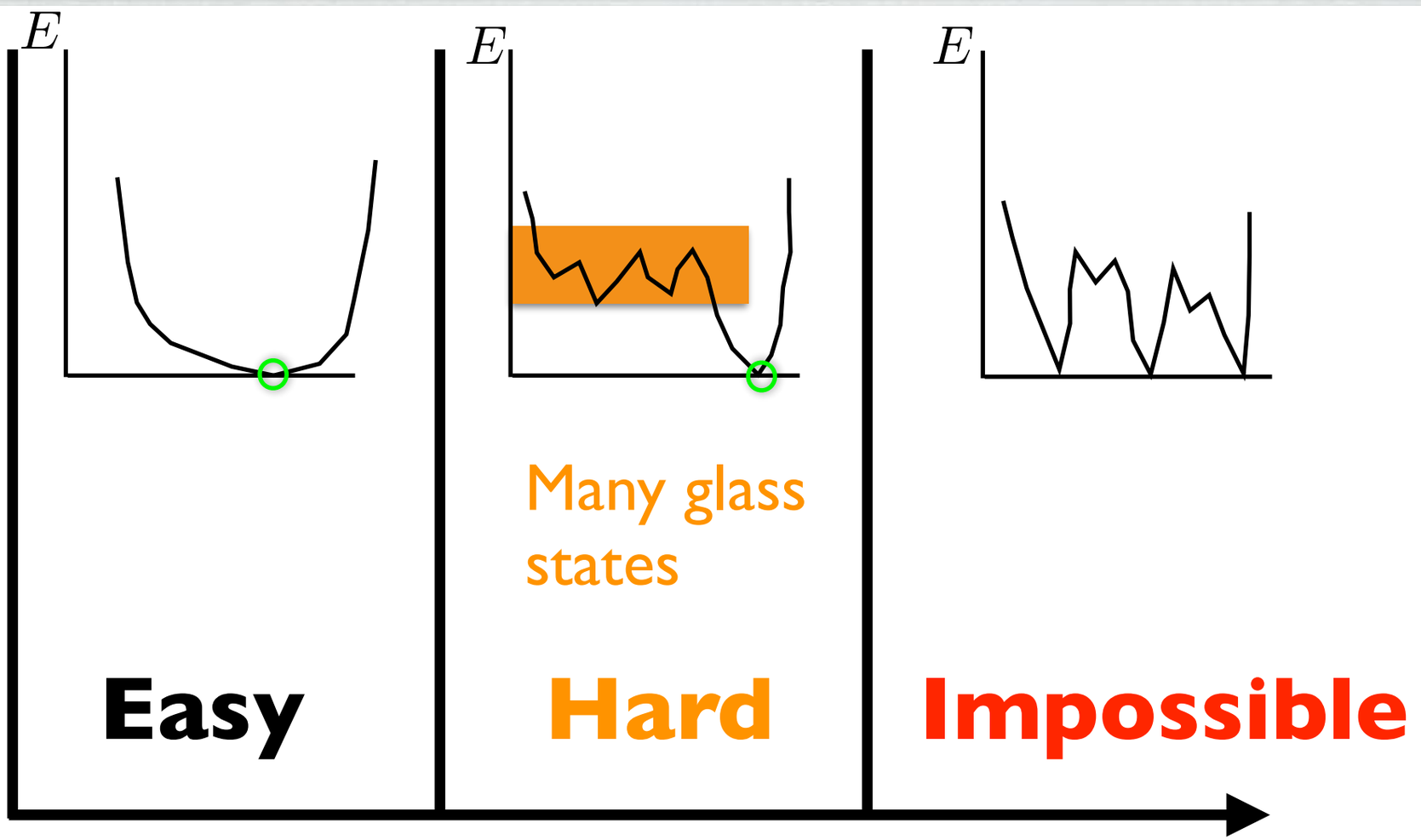
Hypothesis about the success of deep networks: successive disentanglement of combinatorial correlations?

Visible input ➡ Subfeatures ➡ Features ➡ Patterns

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

Machine learning
Deep networks





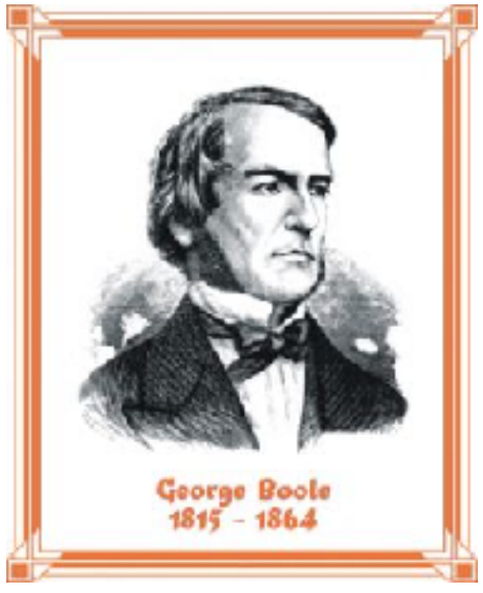
Increasing depth seems to enlarge the « easy » phase. Linked to the structure of data ?

Take-home messages

- Inference with many variables = stat phys problem of disordered system. Search of a special configuration (« crystal »)
- Theory needs 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

Take-home questions

- Q 1: Many of the speakers in this 2018 Cargèse school have their own interpretation for the surprisingly easy learning in deep networks. Which one is correct?
- Q 2: Find good generative models of the world
- Q3: How universal is the dynamical transition which limits practically the inference with mean field algorithms?



THE LAWS OF THOUGHT,

ON WHICH ARE FOUNDED

THE MATHEMATICAL THEORIES OF LOGIC AND PROBABILITIES.

BY

GEORGE BOOLE, LL. D.
PROFESSOR OF MATHEMATICS IN QUEEN'S COLLEGE, CORK.

The general laws of Nature are not, for the most part, immediate objects of perception. They are either inductive inferences from a large body of facts, the common truth in which they express, or, in their origin at least, physical hypotheses of a causal nature serving to explain phænomena with undeviating precision, and to enable us to predict new combinations of them. They are in all cases, and in the strictest sense of the term, probable conclusions, approaching, indeed, ever and ever nearer to certainty, as they receive more and more of the confirmation of experience.

« On the other hand, the knowledge of the laws of the mind does not require as its basis any extensive collection of observations. The general truth is seen in the particular instance, and it is not confirmed by the repetition of instances. We may illustrate this position by an obvious example…

« De omni et nullo » : the maxim of all and none (Aristotelian syllogistic)

- Dogs are mammals.
- Mammals have livers.
- Therefore dogs have livers.

Thanks!