Learning by Local Entropy Maximization:
the effective landscape of neural networks learning algorithms

Riccardo Zecchina

Bocconi University, Milan
Institute for Data Science

Carlo Baldassi (Bocconi, Milan)

Federica Gerace
Carlo Lucibello
Luca Saglietti

Politecnico di Torino
Human Genetics Foundation

Alessandro Ingrosso
Columbia University

Christian Borgs
Jennifer Chayes
Microsoft Research New England

Leon Bottou
Yann Lecun
FB research

Pratik Chaudhari
Stefano Soatto
UCLA

Bert Kappen
Radboud University Nijmegen

Plan of the talk

- Geometrical structure of minima in non-convex random optimization and learning problems
  - Clustering and symmetry breaking
  - The **Local Entropy Measure** reveals the existence of subdominant *high local density regions in weight space.*
  - Accessibility and Local Bayesian predictions
- Algorithms from a local entropy measure
  - The Robust Ensemble: an "out-of-equilibrium" measure
  - Real replicas algorithms: MCMC,SGD and Belief Propagation
  - Connections with DNNs
  - …

What makes a constraint satisfaction problem or a learning problem extracted from a ***natural* distribution** hard to solve?

Basic example: Random K-SAT

- Let $C_K(N)$ be the set of all $2^K \binom{N}{K}$ possible K-clauses on $x_1, x_2, \ldots, x_N$
- Select uniformly, independently and with replacements $M = \alpha N$ clauses from $C_K(N)$ to generate a K-cnf formula $F_N(K, \alpha)$

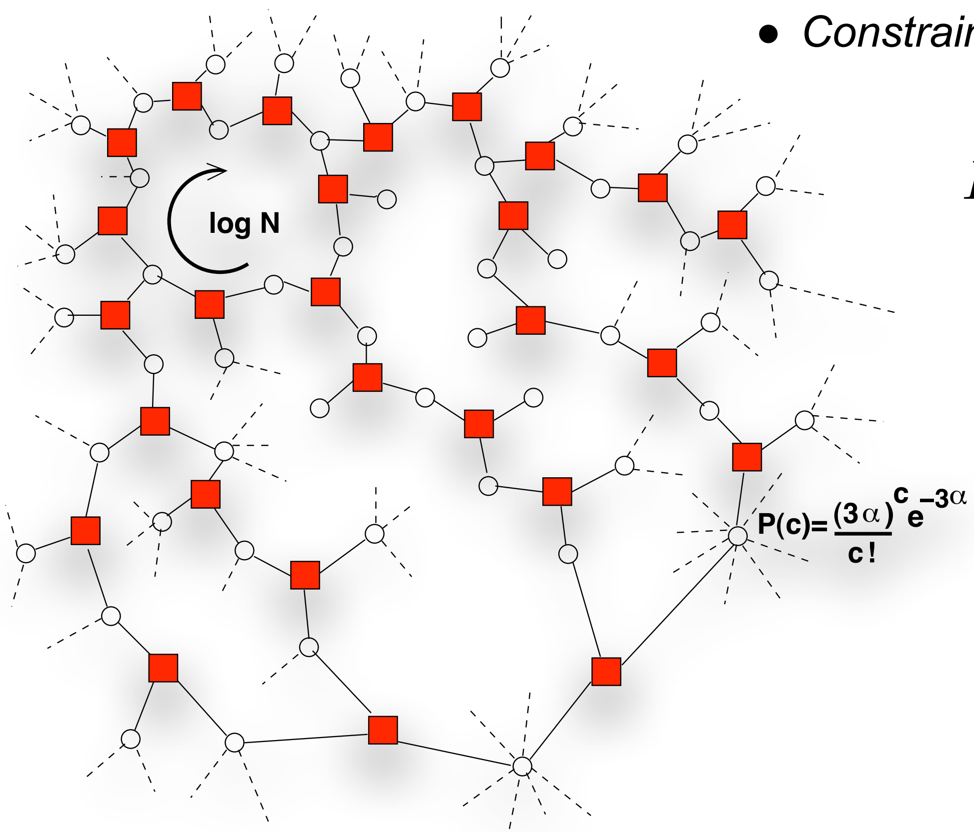
$$F = (x_1 \vee x_{27} \vee \bar{x}_3) \wedge (\bar{x}_{11} \vee x_3 \vee x_2) \wedge \ldots \wedge (x_9 \vee \bar{x}_8 \vee \bar{x}_{30})$$

Question: does $F_N(K, \alpha)$ have a truth assignment?

Factor Graphs for CSPs

- *N discrete variables* $\{x_i\}$ *, e.g., Boolean, spins, colors*
- *Constraints* E_a , $a = 1, ..., M$ *involving vars* $\{x_{i(a)}\}$

$$E_a = \begin{cases} 0 & \text{if } \{x_{i(a)}\} \text{ satisfy constraint} \\ 1 & \text{otherwise} \end{cases}$$

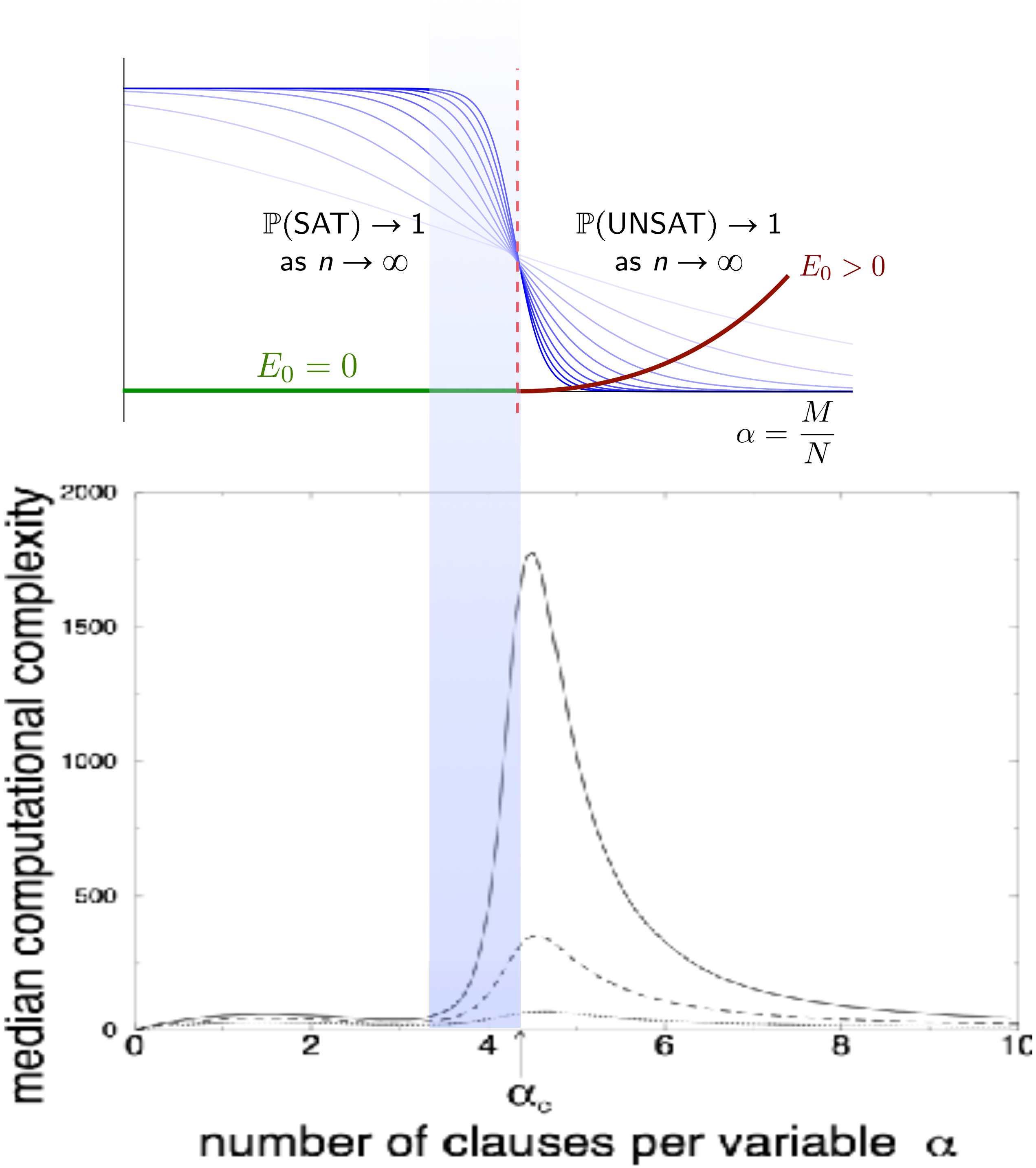


$$E_a = (x_{i_1} \vee \bar{x}_{i_2} \vee x_{i_3})$$

$$M = \alpha N$$

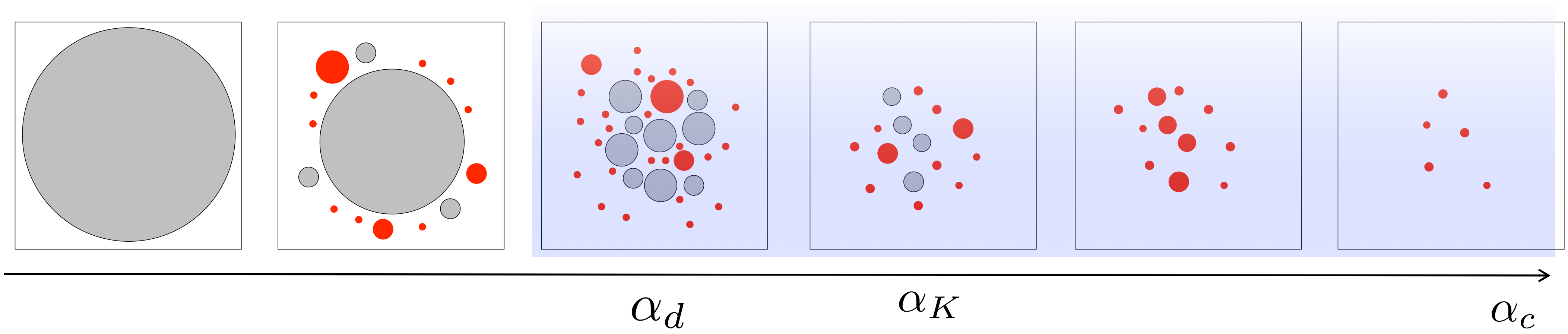
Cost/Energy function:

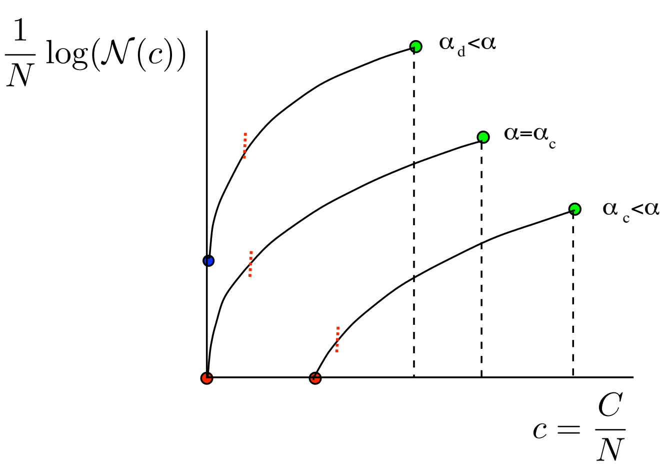
$$E = \sum_{a=1}^{\alpha N} E_a[\{x_{\{i(a)\}}]$$

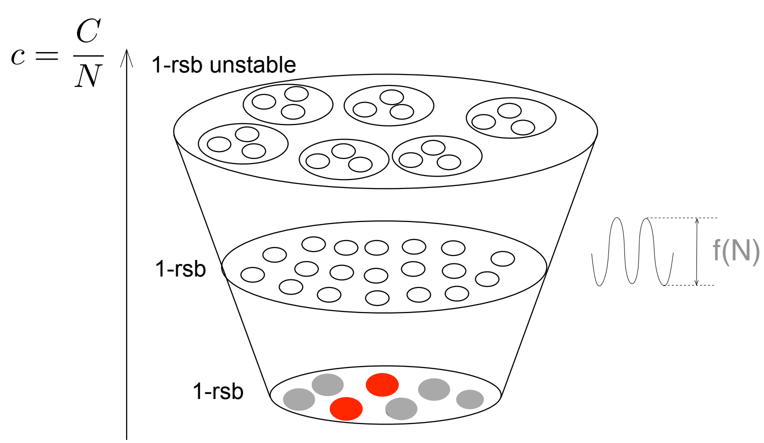


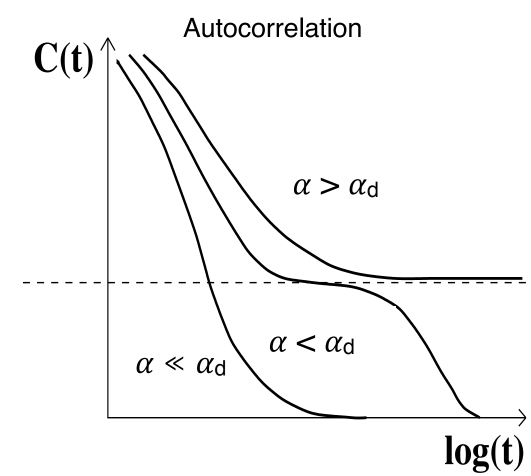
$\mathbb{P}$(SAT) → 1
as $n \to \infty$
$\mathbb{P}$(UNSAT) → 1
as $n \to \infty$
$E_0 > 0$
$E_0 = 0$
$\alpha = \frac{M}{N}$
median computational complexity
2000
1500
1000
500
0
0
2
4
6
8
10
α_c
number of clauses per variable α

Geometry of solutions in random Constraint Satisfaction Problems: Gibbs measure decomposition









Finding isolated solutions is hard. In the last 15 years many physicists, mathematicians and CS have contributed to various aspects of these results … the scenario is by now rigorously established

Gibbs measure decomposition

$$P(\mathbf{w}) = \frac{1}{Z} \prod_{a=1}^{M} \Psi_a(\{w_{(ij)\in a}\})$$

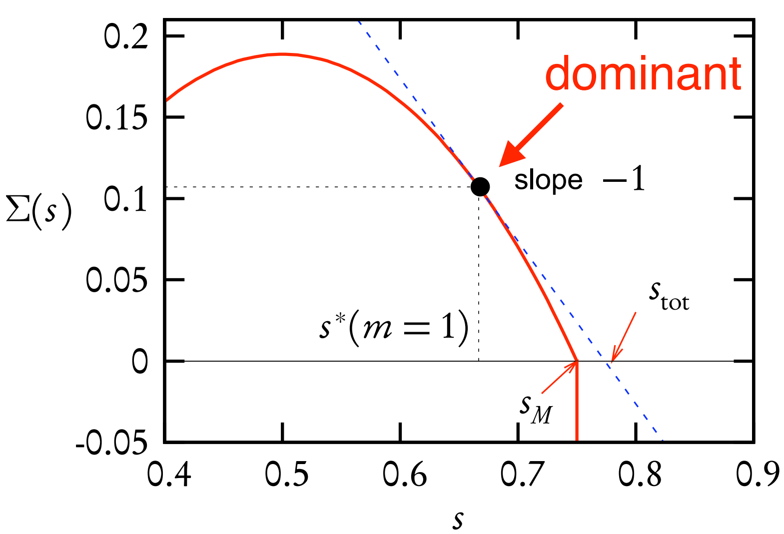
$$P_\ell = \sum_{\{\mathbf{w}\in A_\ell\}} P(\mathbf{w})$$

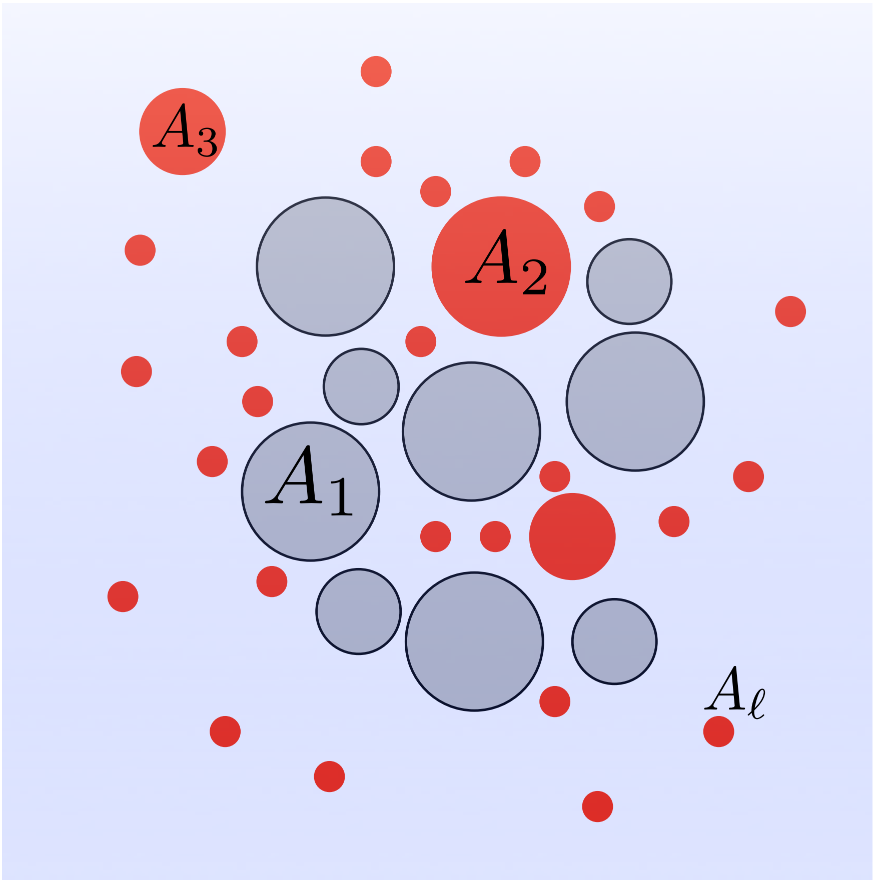
RS: $P_1 = 1$

$P_1 > P_2 > P_3 > \ldots$

1RSB-d: $\mathcal{N} = e^{\Sigma N}$

1RSB-s: $\mathcal{N} = \text{sub-exp}$

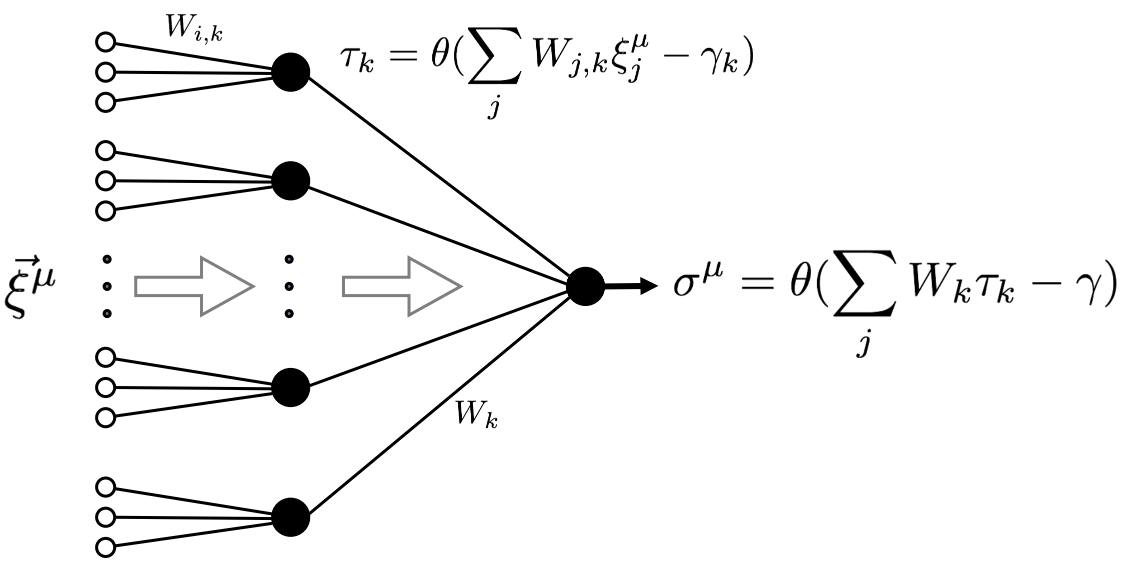




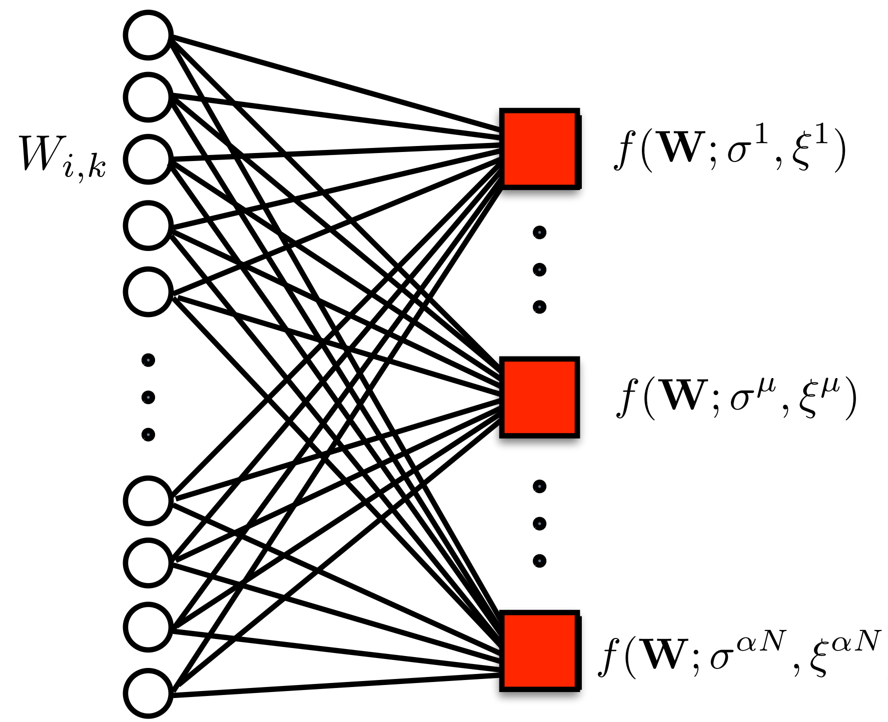
Learning as a CSP problem

Constraints: one for each pattern

$$f(\mathbf{W};\sigma^\mu,\xi^\mu) = \delta\left(\sigma^\mu;\sigma(\mathbf{W},\xi^\mu)\right)$$

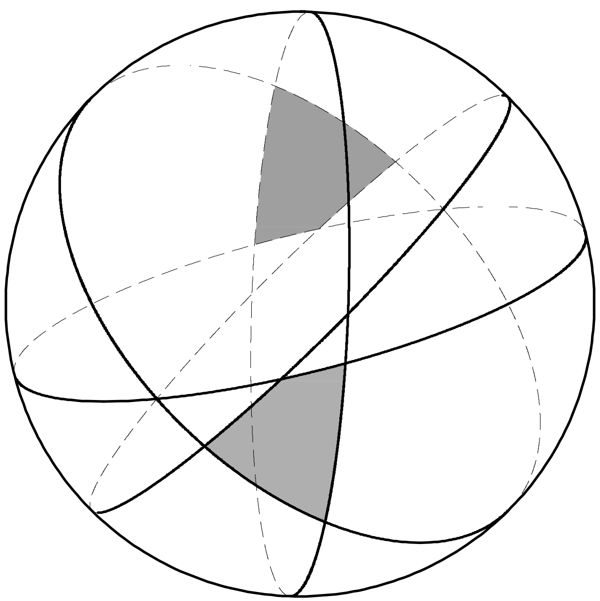


Fully connected factor graph



"Old" (90s) statistical physics results: a relatively similar scenario

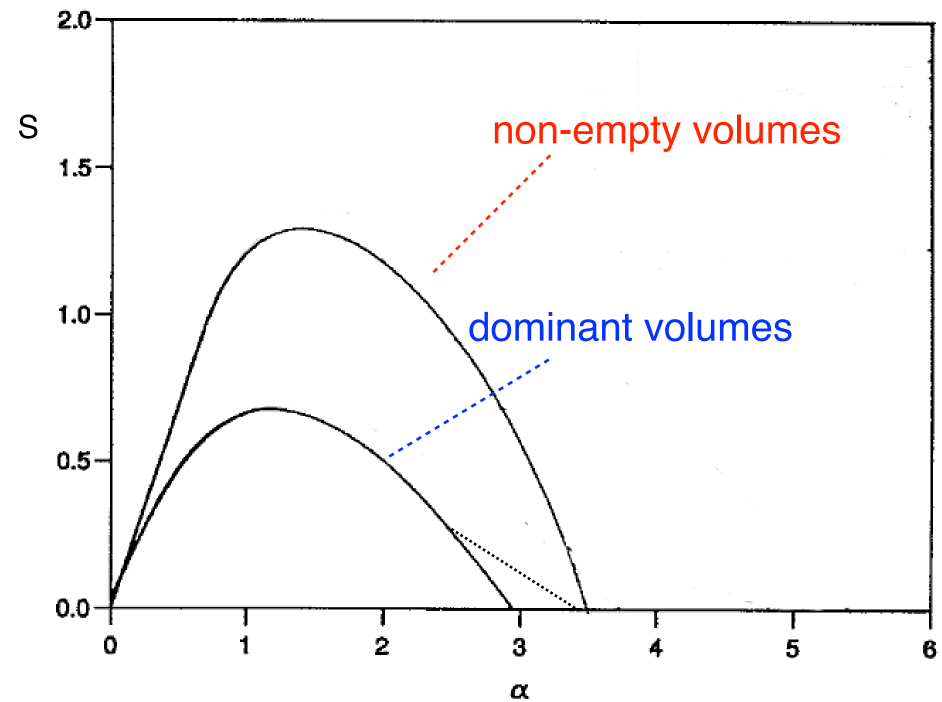
Geometry of space of solution and internal representations in MLP learning random patterns with continuous weights (zero errors landscape)



Fractional volume of weights storing the patterns

$$V = \frac{\int d\mathbf{w}\ \delta(\mathbf{w}^2 - 1) \prod_\mu \delta\left(\sigma^\mu ; \sigma(\mathbf{w}, \xi^\mu)\right)}{\int d\mathbf{w}\ \delta(\mathbf{w}^2 - 1)}$$

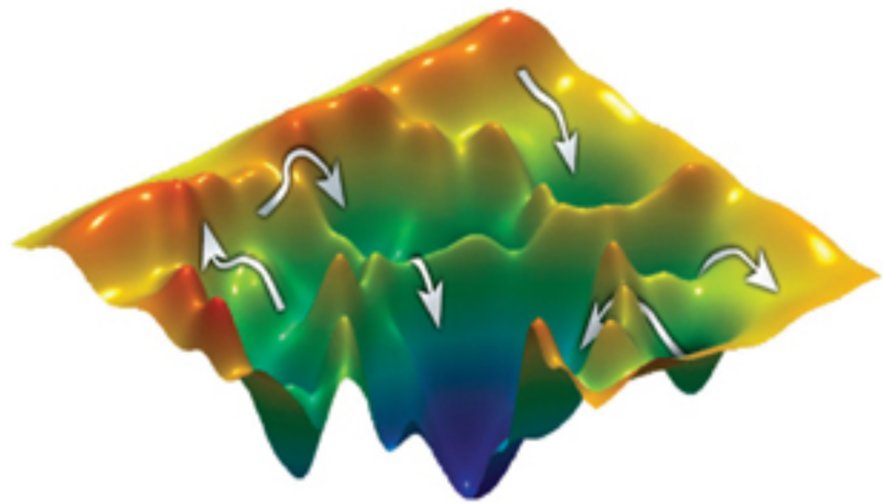
one hidden layer committee NN



Monasson, O'Kane 94, Monasson, Zecchina, 95

How does learning take place in large scale DNNs?

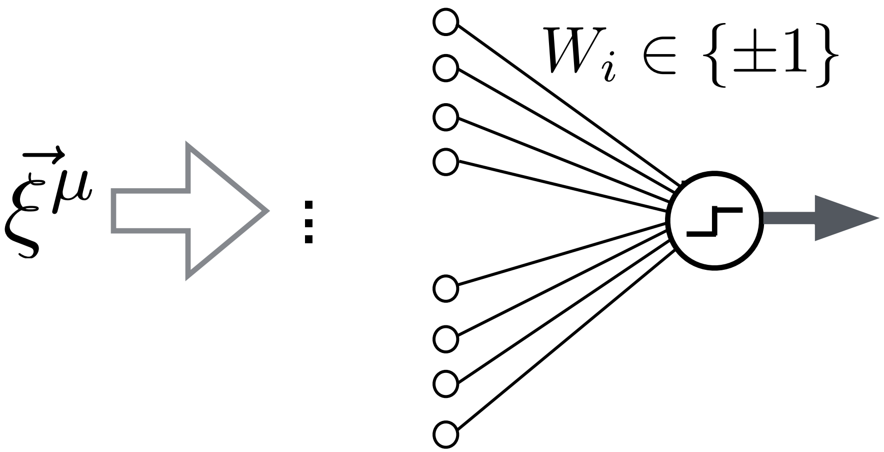
Learning algorithms: Variants of **gradient back-propagation**



However, successful algorithms never “simply” minimize the loss.

Why?

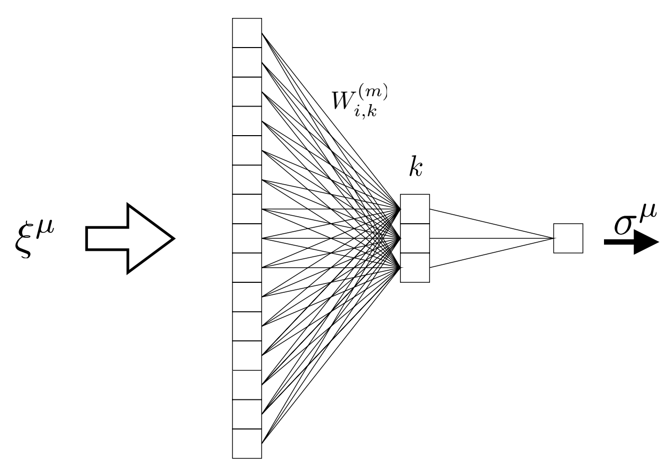
The simplest non-convex neural device: perceptron with discrete weights



$$W_i \in \{\pm 1\} \quad , \quad i = 1, ..., N$$

$$\sigma = \mathrm{sign}\left(\sum_{i=1}^{N} W_i \xi_i^{\mu}\right)$$

Analytical results generalise to arbitrary number of levels and multiple layers



Non-convex minimum *"energy"* problem

Given a set of i.i.d. random examples (p=1/2):

$$\{(\xi_i^\mu = \pm 1, \sigma^\mu = \pm 1)\} \qquad i = 1, \ldots, N \qquad \mu = 1, \ldots, \alpha N$$

Find **W** such that $\sigma^\mu = \sigma(\mathbf{W}, \xi^\mu) \quad \forall \mu$

$$\Rightarrow \boxed{\alpha N \quad \text{constraints on} \quad \{W_i\}}$$

Cost-energy function

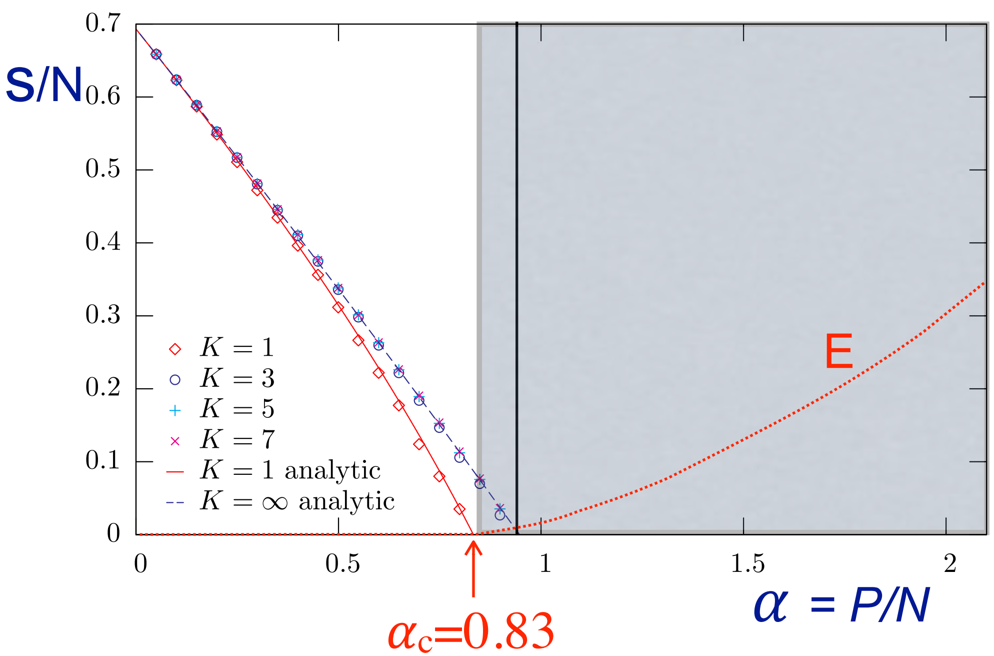
$$E(\mathbf{W}) = \sum_\mu \Theta\left(-\sigma^\mu \mathrm{sgn}(\mathbf{W} \cdot \xi^\mu)\right)$$ = # number of errors

$$\Theta(x) = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases}$$

Phase diagram (~1990)

- At E=0, minima are very narrow and isolated

S= log (# optimal W assignments)



some classical papers:

E. Gardner, E. Gardner B. Derrida, +

W. Krauth, M. Mézard, *J. de Physique* **50**, 3057-3066 (1989) ;E. Barkai, D. Hansel, H. Sompolinsky, *Phys. Rev. A* **45**, 4146-4160 (1992) ; M. Mezard, J. Phys. A 22, 2181 (1989); H.S. Seung, H. Sompolinsky, N. Tishby,Phys. Rev. A 45, 6056 (1992); E. Barkai, I. Kanter, Europhys. Lett14, 107 (1991); R. Penney and D. Sherrington, J. Phys. A 26, 6173(1993)
M. Tsodyks , *Mod.Phys. Lett. B* 4, 713 (1990); D.J. Amit, S. Fusi *NETWORK* 3, 443 (1992); D.J. Amit, S. Fusi, *Neural Computation* **6**, 957-982 (1994);
H. Horner, *Z. Phys. B* **86**, 291-308 (1992)
...

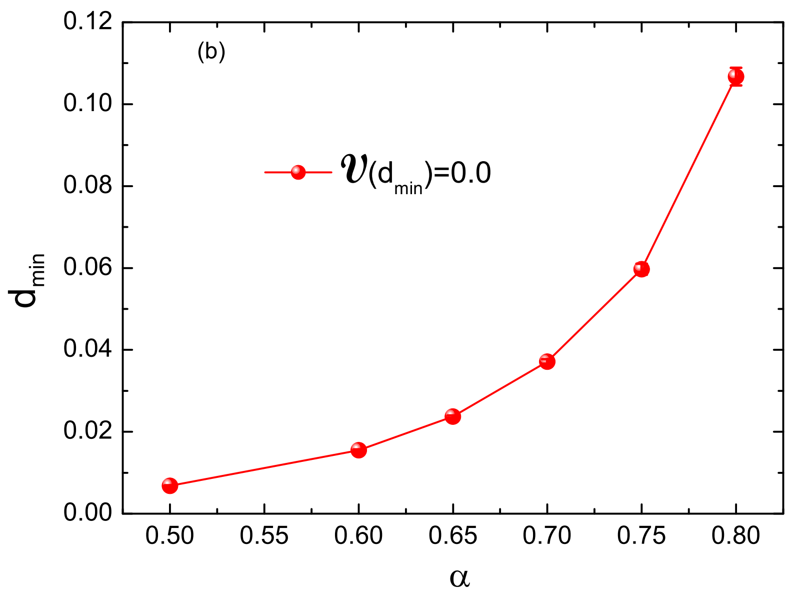
For decades, heuristic local search algorithms were believed to fail in finding solutions for any extensive number of patterns.

Geometry of the space of solutions in the binary perceptron:

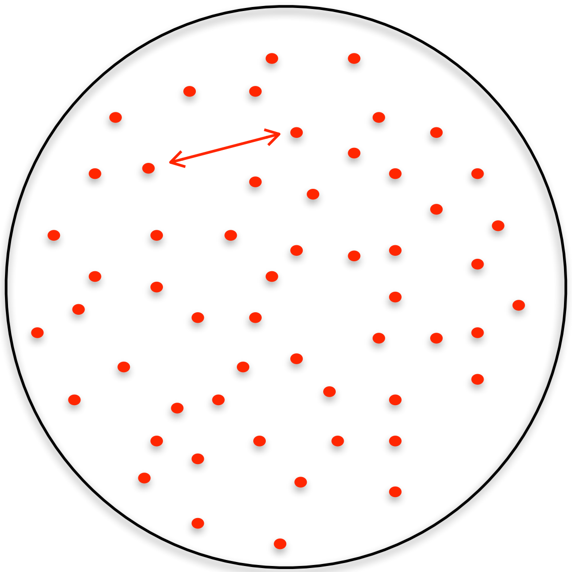
Franz-Parisi potential: entropy at distance **d,** sampling from **typical** solution **J**

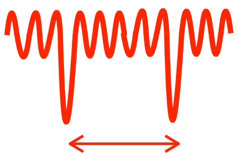
$$F(x) = \left\langle \frac{1}{Z(T')} \sum_{\mathbf{J}} \Theta\left(\frac{1}{\sqrt{N}} \sum_{i=1}^{N} J_i \xi_i^{\mu}\right) \ln \sum_{\mathbf{w}} \Theta\left(\frac{1}{\sqrt{N}} \sum_{i=1}^{N} w_i \xi_i^{\mu}\right) e^{x\mathbf{J}\cdot\mathbf{w}} \right\rangle_{\boldsymbol{\xi}}$$

$d_{\min}(\alpha)$



$d_{\min}(\alpha) \sim O(N)$



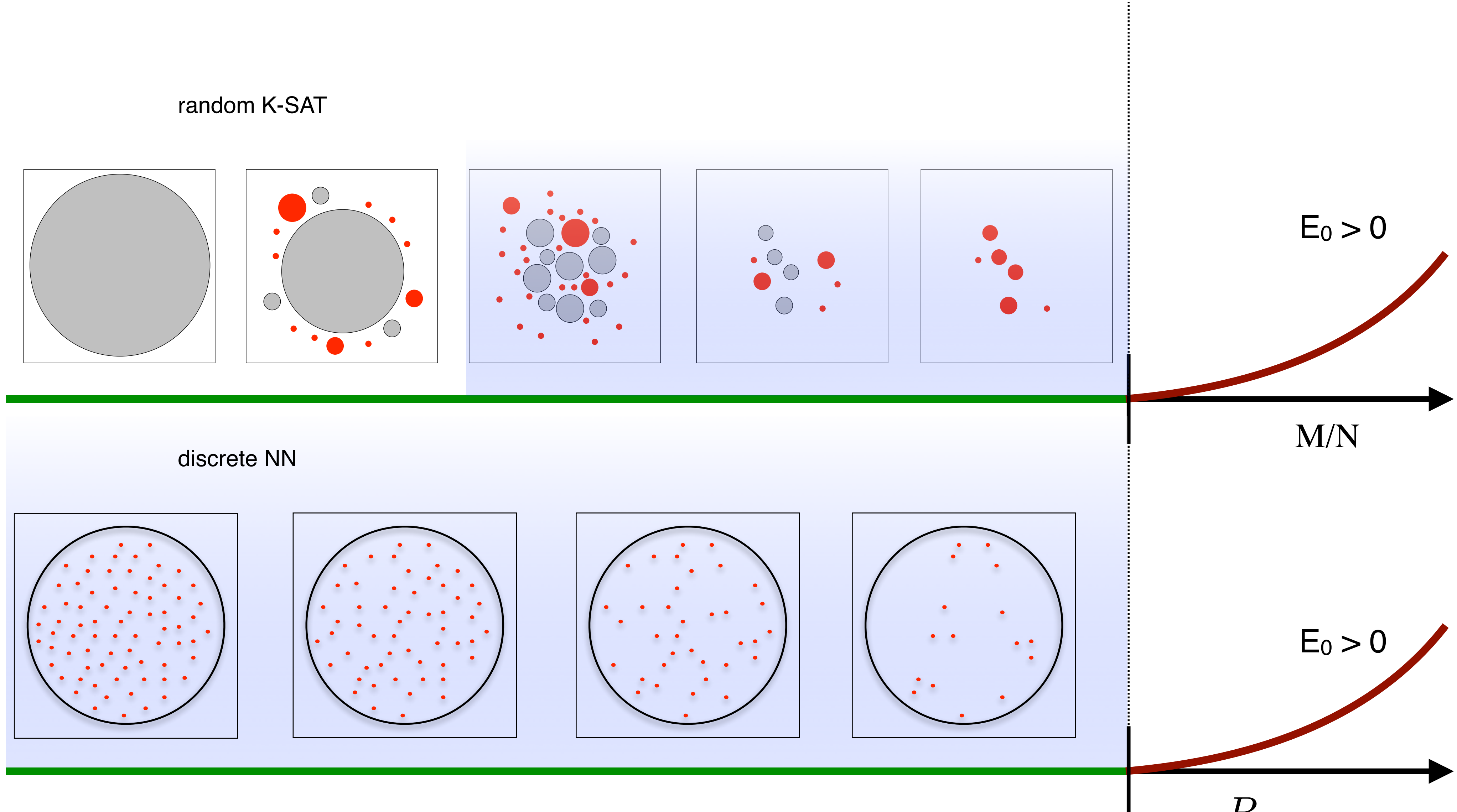


$d_{\min}(\alpha)$

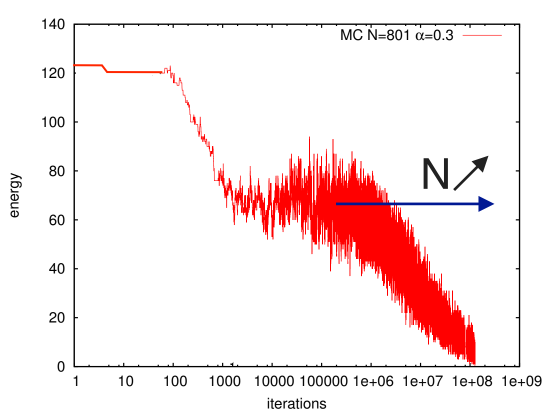
H. Huang, Y. Kabashima (2014) (q_1=1 known since the 80's)

$$\alpha_c = \frac{P_{max}}{N} \simeq 0.83$$

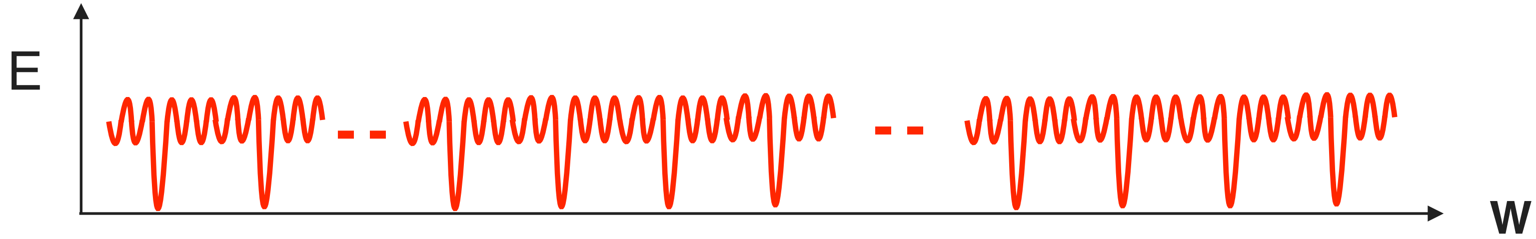
W Krauth, M. Mezard, (1989)



$$\alpha_c = \frac{P_{max}}{N} \simeq 0.83$$



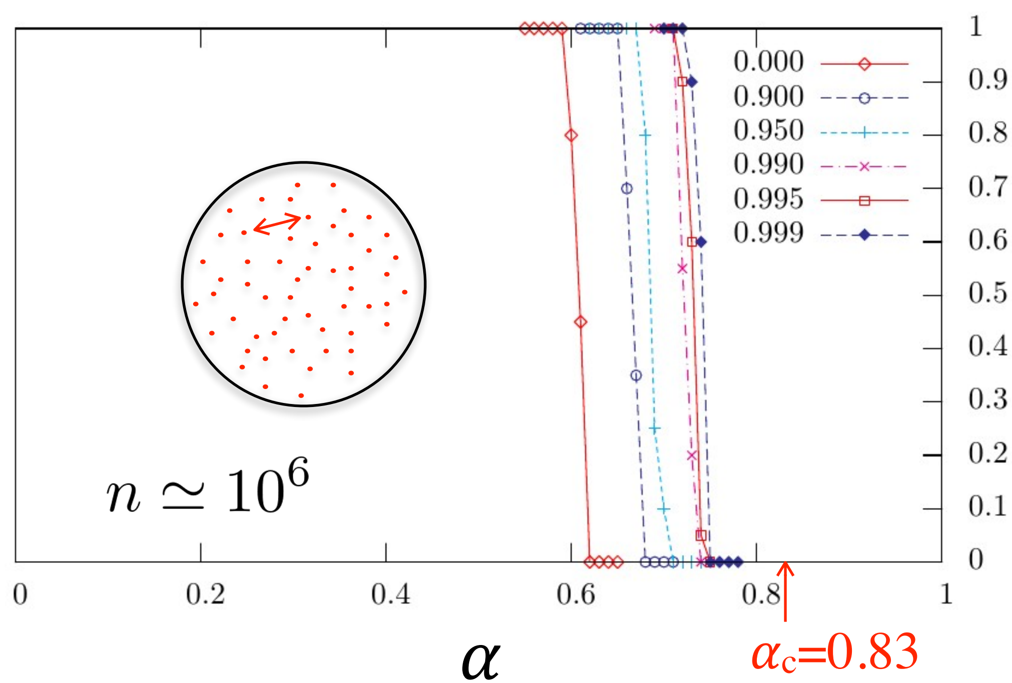
learning impossible for random patterns ?



Golf course for any $\boldsymbol{\alpha}$?   Efficient learning impossible ?

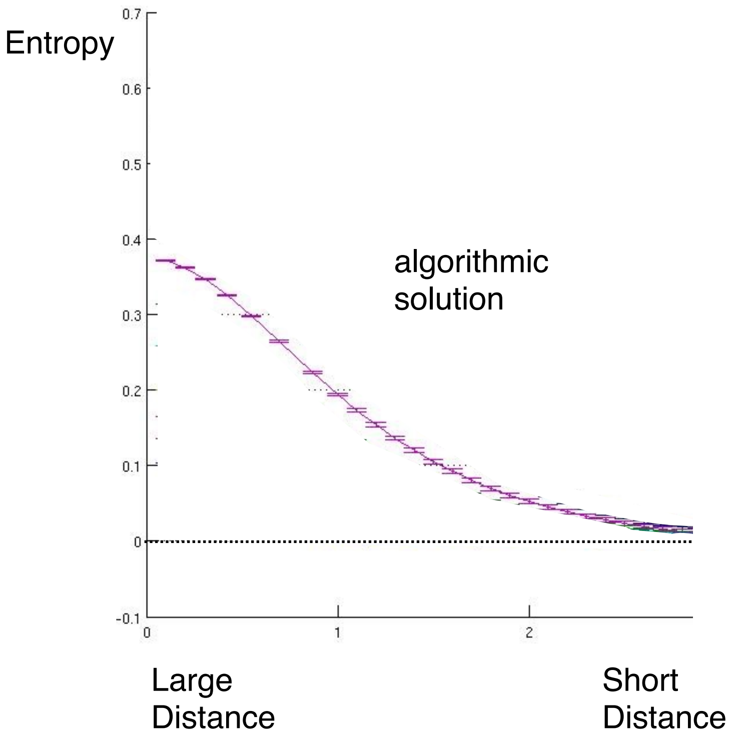
Message-passing algorithms (2006, A. Braunstein RZ) and its simplifications work well

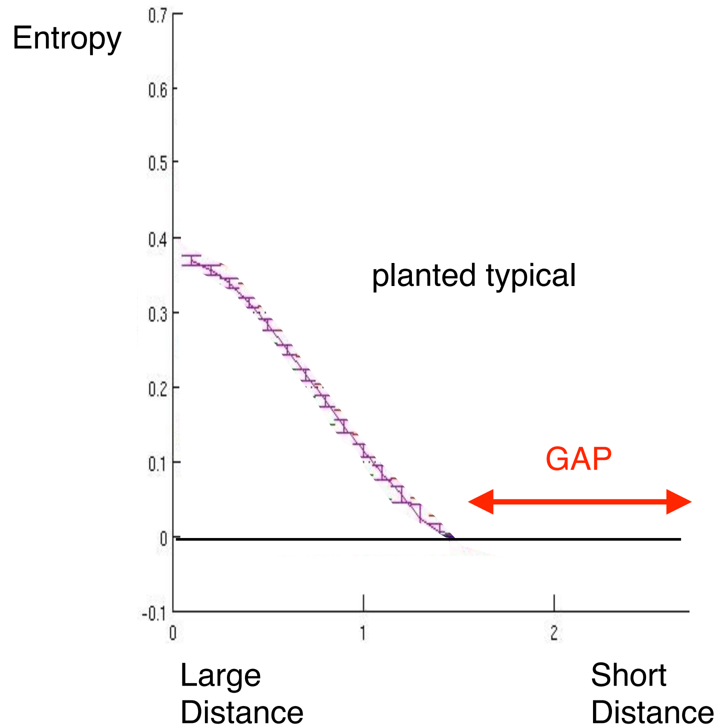
success probability



something unclear …

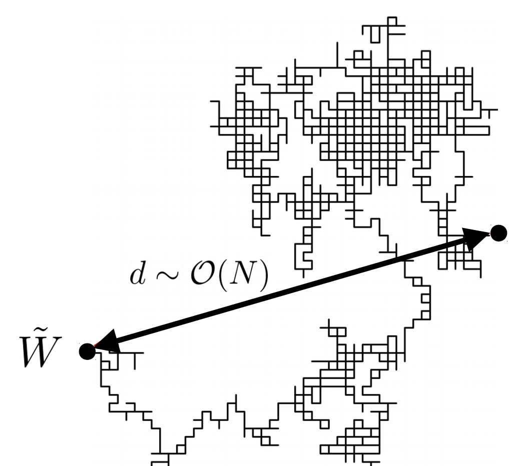
Weight enumerator functions computed with BP, relative to a solution found by an algorithm and to a typical solution (planted)





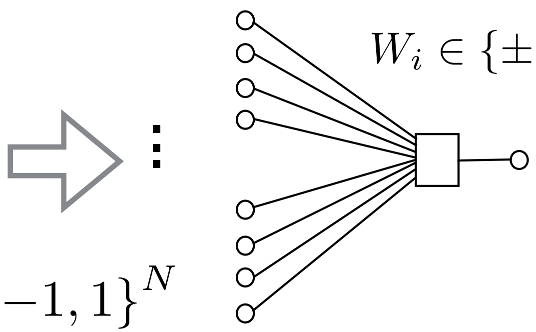
N=1001
samples=50

Evidence for Random Walks analysis:



Learning in rare regions ? Large deviations analysis

$\vec{\xi}^{\mu}$



$W_i \in \{\pm 1\} \quad , \quad i = 1, ..., N$

$\tau(W, \xi) \ = \ \mathrm{sign}(W \cdot \xi)$

$\xi \in \{-1, 1\}^N$

Characteristic function:

$\mathbb{X}_{\xi}(W) \ = \ \prod_{\mu=1}^{\alpha N} \Theta\left(\sigma^{\mu} \tau\left(W, \xi^{\mu}\right)\right)$ = 1 iff all patterns are correctly classified

Number of solutions within Hamming distance d from a given weight vector $\tilde{W}$:

$$\boxed{\mathcal{N}\left(\tilde{W}, d\right) = \sum_{\{W\}} \mathbb{X}_{\xi}(W)\, \delta\left(W \cdot \tilde{W}, N(1-2d)\right)}$$

$$A \cdot B = \sum_{j=1}^{N} A_j B_j$$

Local entropy measure

number of solutions within a distance d $\mathcal{N}\left(\tilde{W}, d\right) = \sum_{\{W\}} \mathbb{X}_\xi(W)\, \delta\left(W \cdot \tilde{W}, N(1-2d)\right)$

"energy" = local entropy

$$\boxed{\mathcal{E}_d(\tilde{W}) \doteq -\log \mathcal{N}(\tilde{W}, d)}$$

Local Entropy Measure

maximally dense for $y \to \infty$

$$\boxed{\mathcal{P}(\tilde{W}) \propto e^{-y\mathcal{E}_d(\tilde{W})}}$$

normalisation

$$Z(d) = \sum_{\{\tilde{W}\}} X_\xi(\tilde{W}) e^{-y\mathcal{E}_d(\tilde{W})}$$

Subdominant Dense Clusters Allow for Simple Learning and High Computational Performance in Neural Networks with Discrete Synapses, C. Baldassi, A. Ingrosso, C. Lucibello, L. Saglietti, and R. Zecchina Phys. Rev. Lett. **115**, 128101 (2015)

By the replica/cavity method we can compute the expectation of the local entropy in the large N limit

$$\mathscr{S}_I(d,y) = -\left\langle \mathscr{E}\left(\tilde{W}\right)\right\rangle_{\xi,\tilde{W}} = \frac{1}{N}\left\langle \log\mathcal{N}\left(\tilde{W},d\right)\right\rangle_{\xi,\tilde{W}}$$

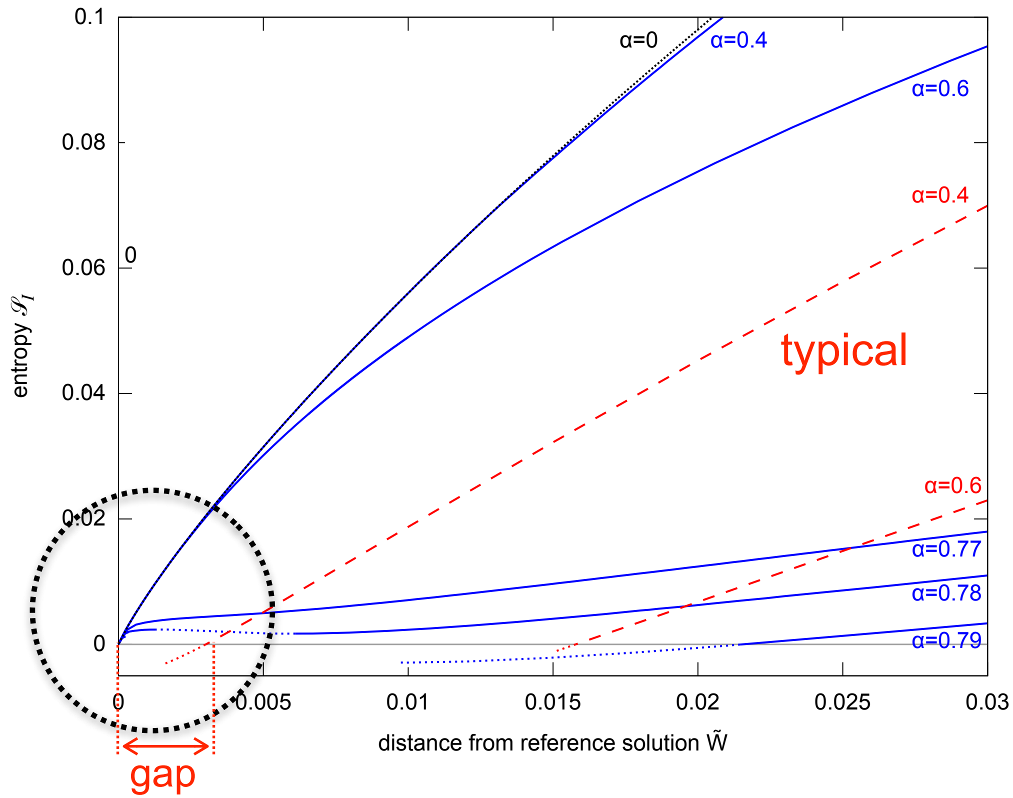
$$\mathscr{S}_I(d,y) = \partial_y\left(y\mathscr{F}(d,y)\right)$$

internal entropy

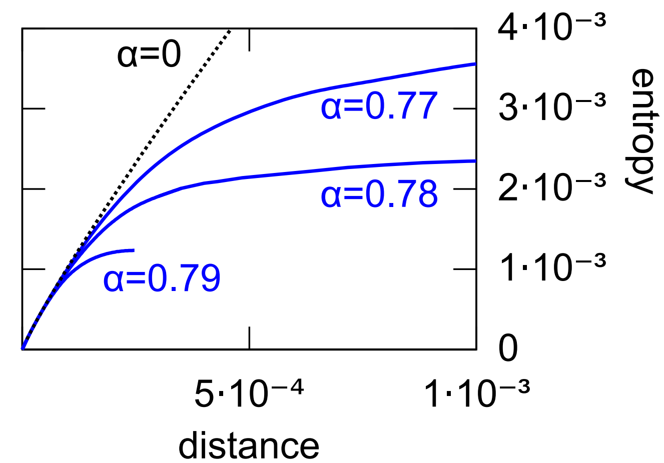
$$\mathscr{S}_E(d,y) = -y\left(\mathscr{F}(d,y) + \mathscr{S}_I(d,y)\right)$$

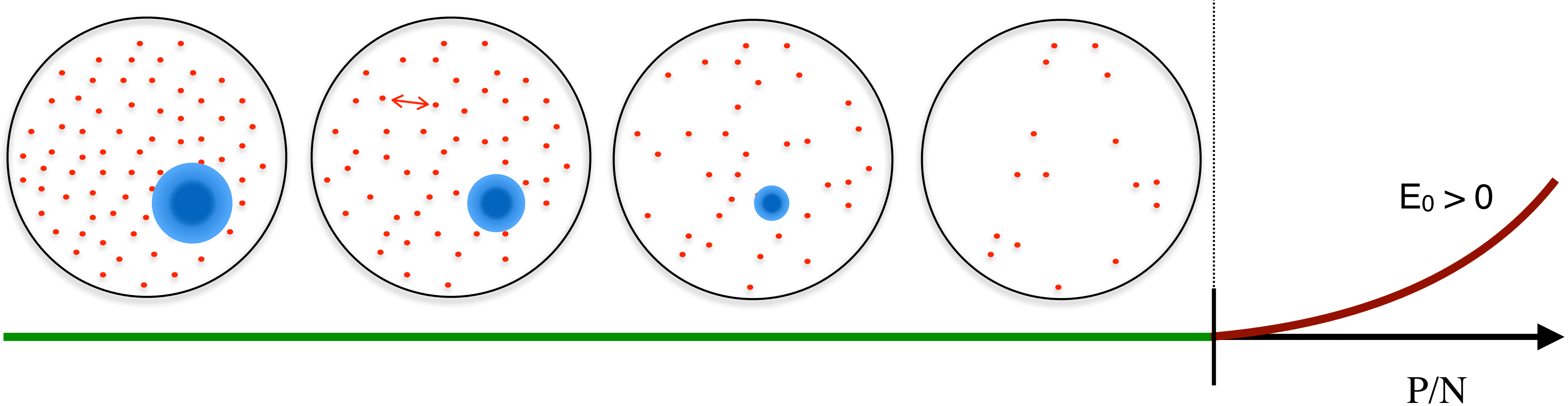
external entropy

Large deviation analysis



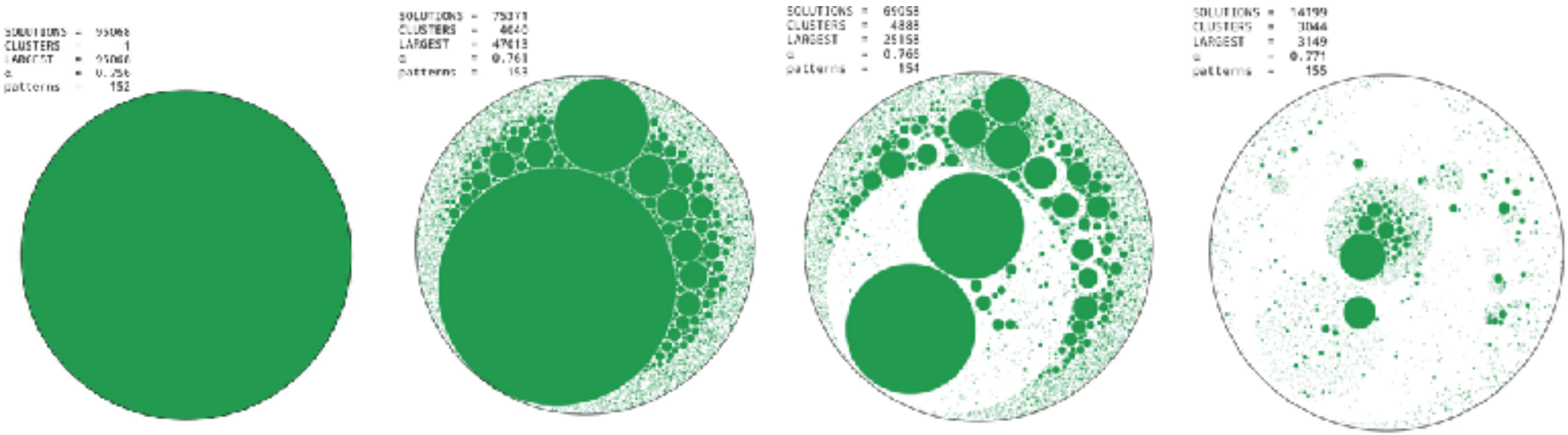
ultra-dense cluster





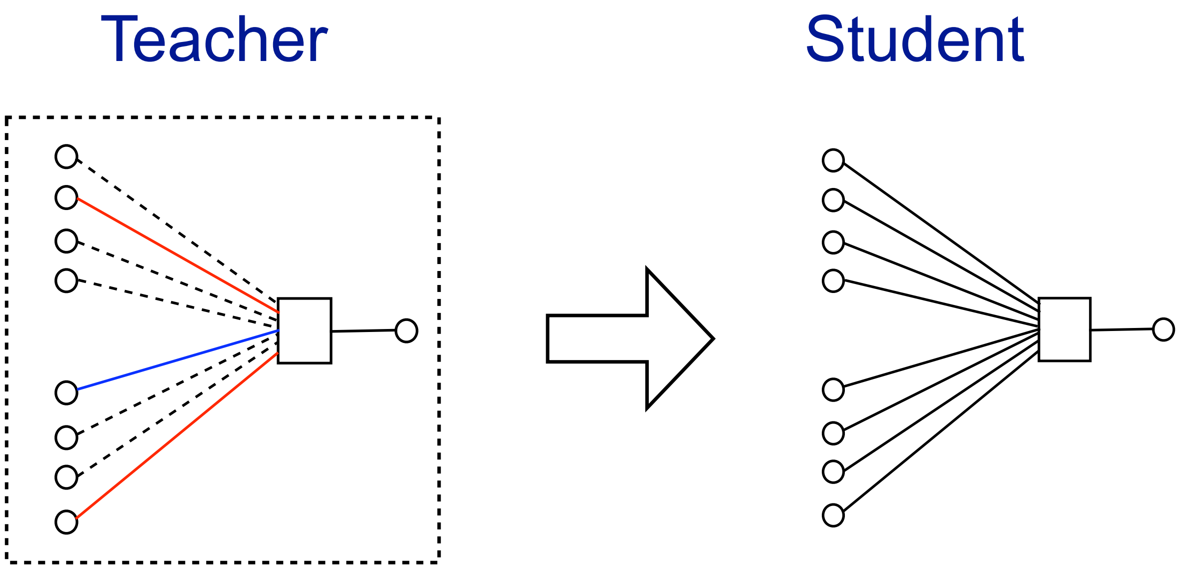
in the 90s we were not aware of this fundamental structural property

close to capacity the dense clusters breaks up



the shape is not spherical …

Making predictions

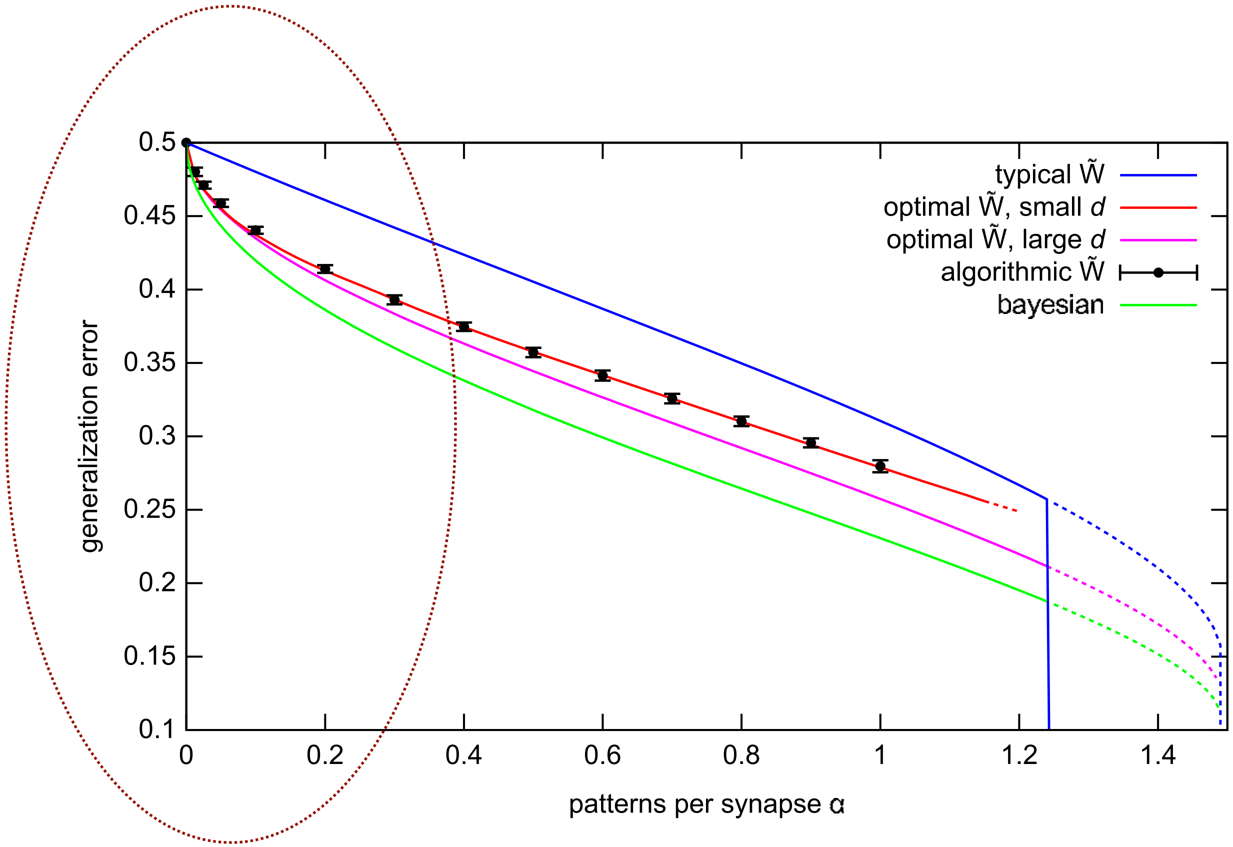


Probability to give the same answer of the teacher on a new input

Optimal Bayesian prediction:

$$P(\sigma|\xi^{new}, \{\xi^{\mu}, \sigma_{\mu}\}) = \int dW\, P(\sigma|W, \xi^{new},)P(W, \{\xi^{\mu}, \sigma_{\mu}\})$$

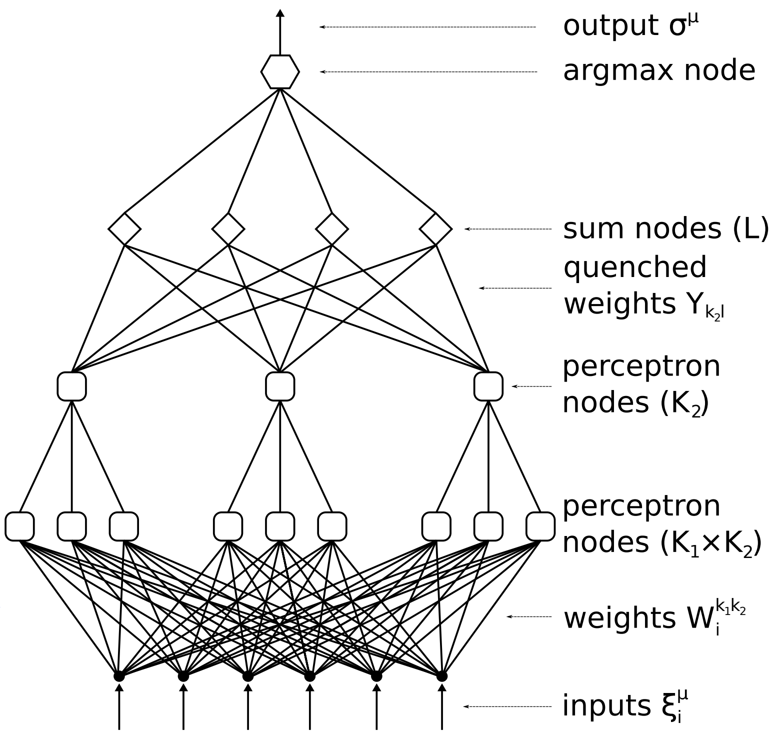
Dense states have the propensity to generalise



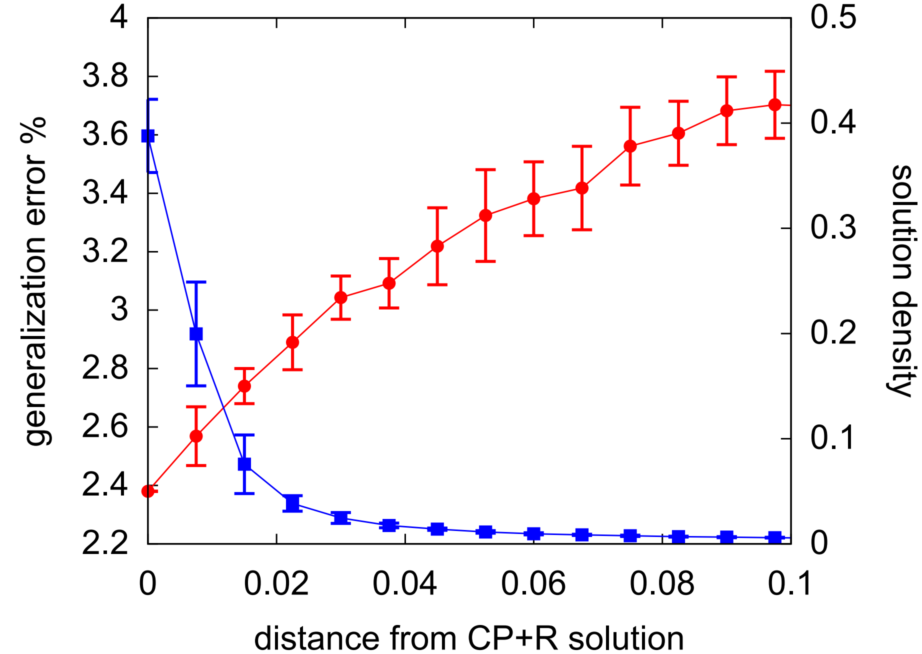
- contribution to the Bayesian integral from the dense cluster
- the Teacher is an isolated weight vector

Subdominant Dense Clusters Allow for Simple Learning and High Computational Performance in Neural Networks with Discrete Synapses
C. Baldassi, A. Ingrosso, C. Lucibello, L. Saglietti, and R. Zecchina Phys. Rev. Lett. **115**, 128101 (2015)

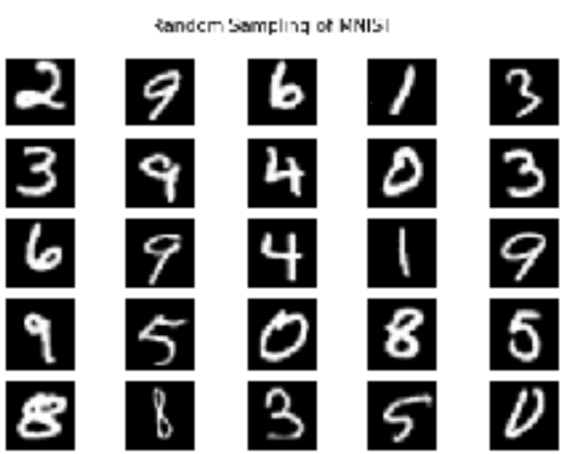
Smallest architecture with zero errors on the training set



learning layer



(70000 patterns)



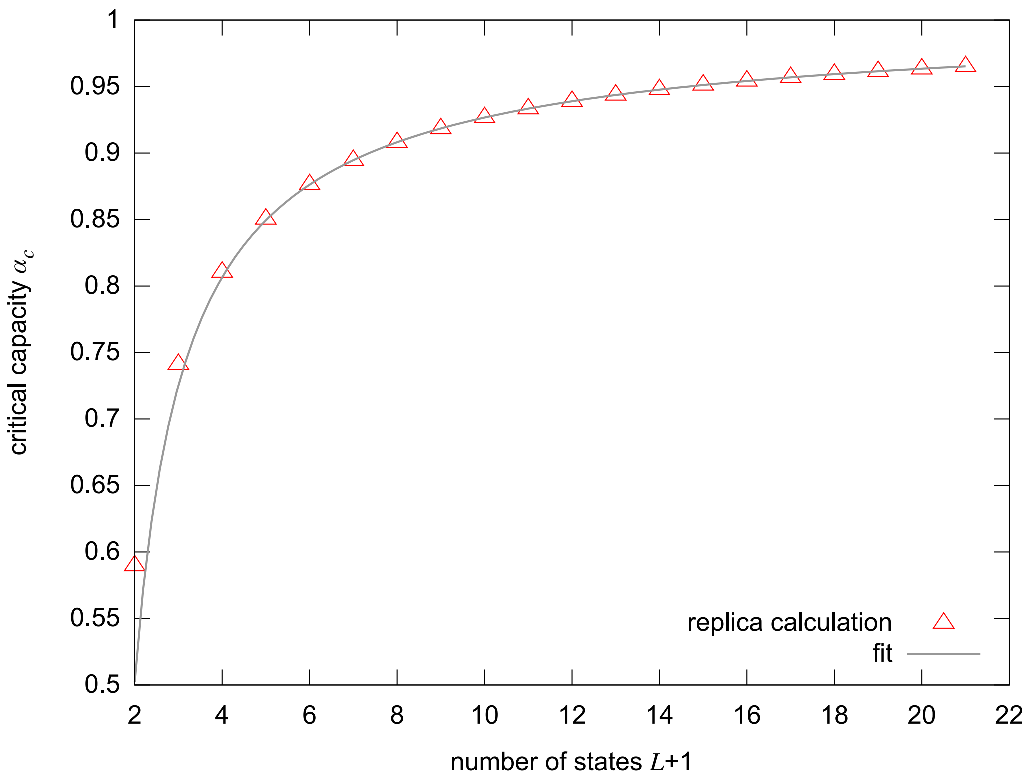
Prediction error ~ 1.2 % with binary weights

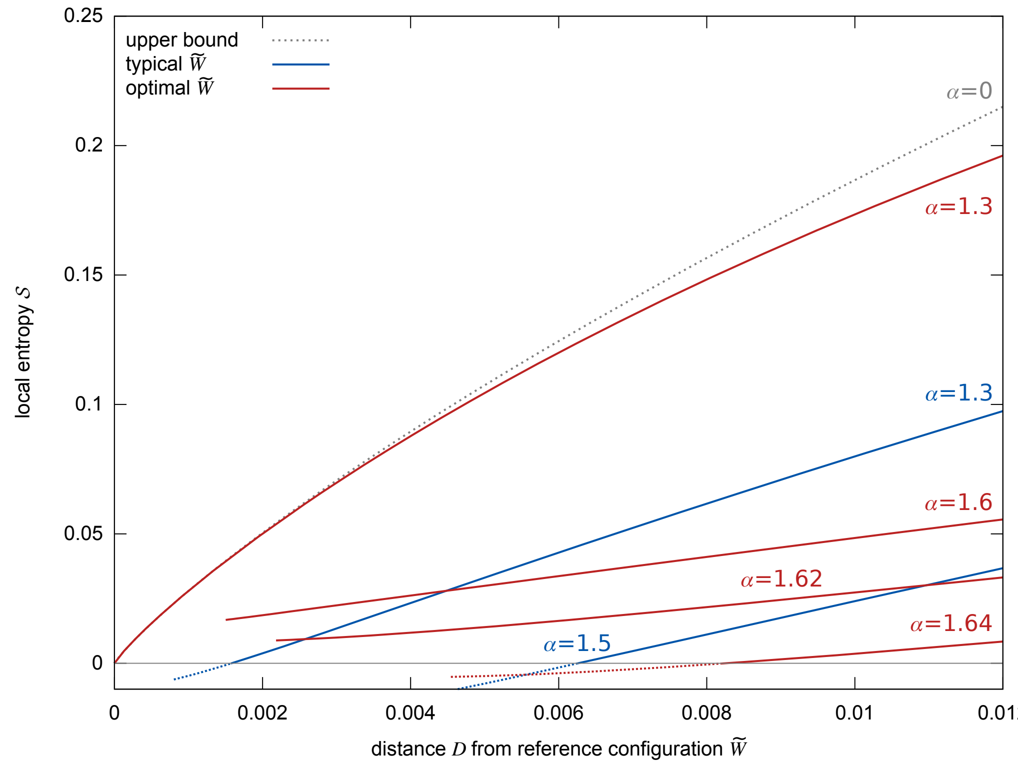
no overfitting with size

Generalisation to multiple state variables

Few levels almost saturate performance

L+1=5





C. Baldassi, Federica Gerace, A. Ingrosso, C. Lucibello, L. Saglietti, R. Zecchina, Phys. Rev. E **93**, 052313 (2016)

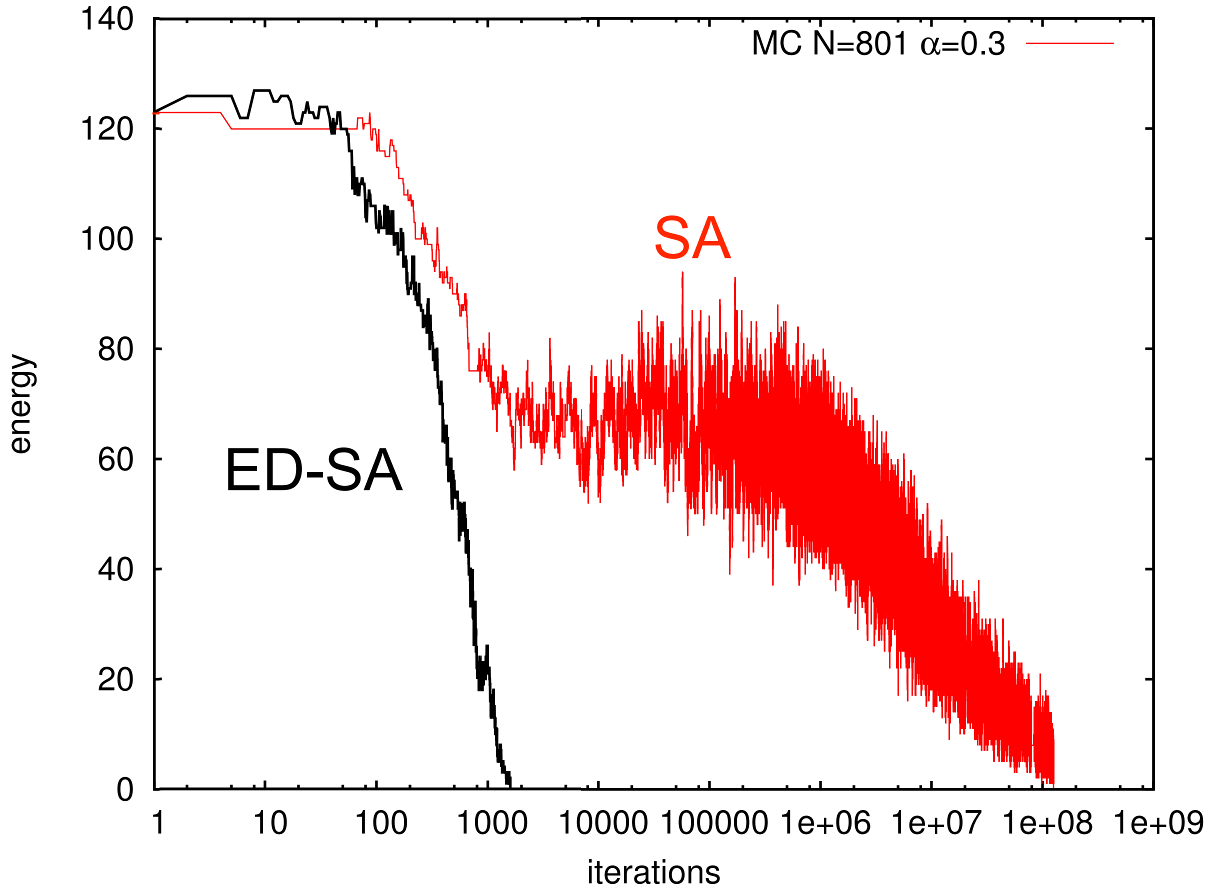
Principled algorithm:
Local Entropy driven Simulated Annealing

Objective Function:
search for configurations which maximize the local entropy
(minimize the "energy")

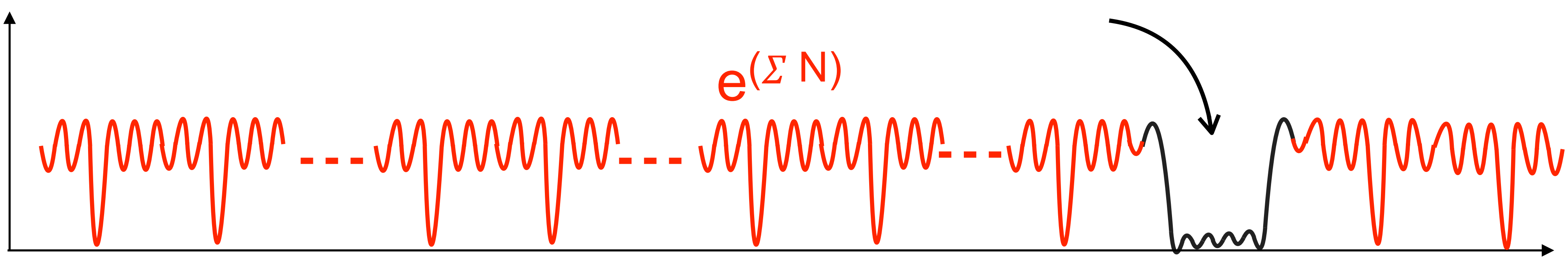
$$\mathscr{E}\left(\tilde{W}\right) = -\log \mathcal{N}\left(\tilde{W}, d\right)$$

1. SA moves

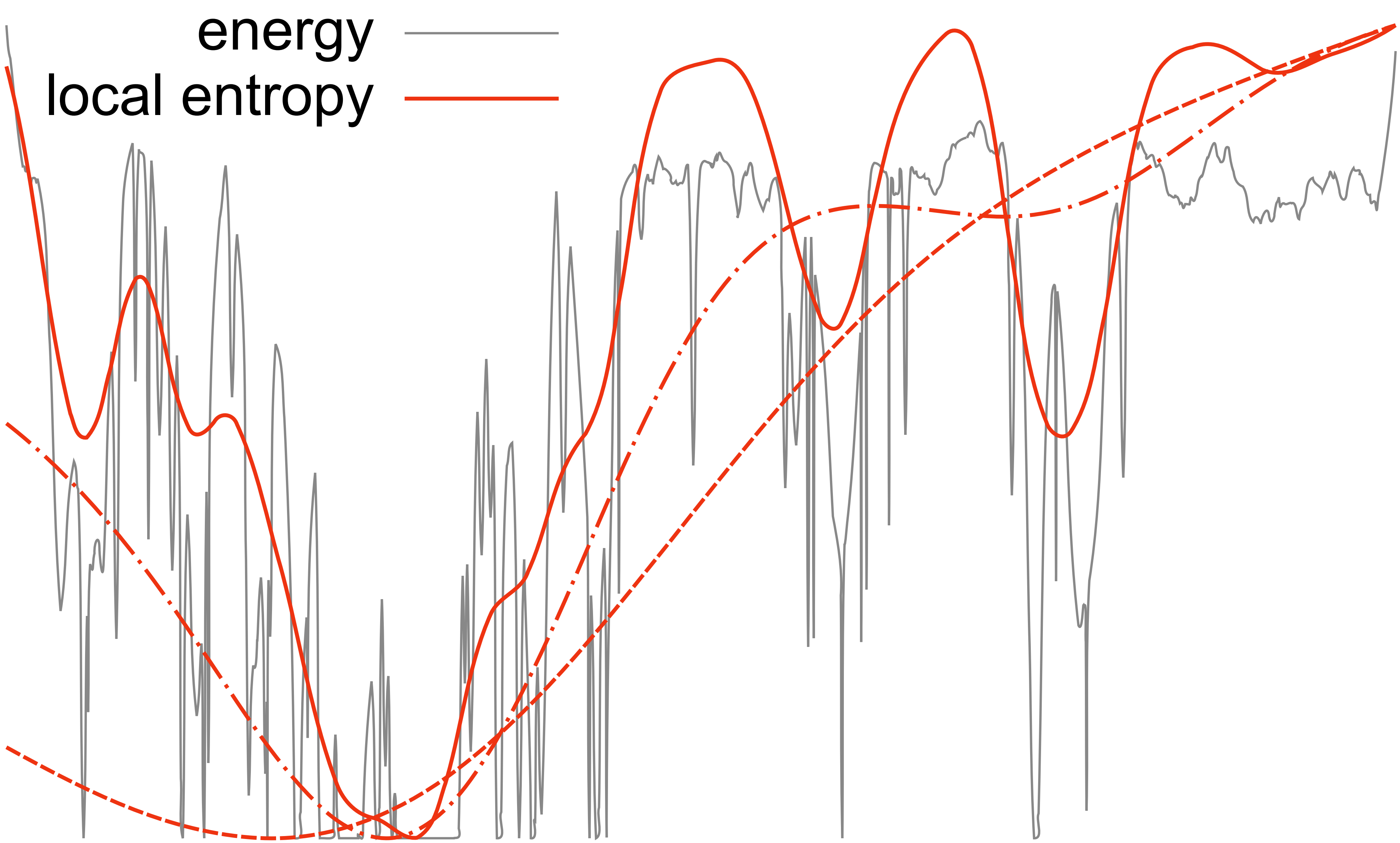
2. Belief Propagation method to estimate the local entropy



$N = 801, \ \alpha = 0.3$



C. Baldassi, A. Ingrosso, C. Lucibello, L. Saglietti, R. Zecchina, J. Stat. Mech. 2016 (2) 023301



Local should not be interpreted as *infinitesimal*: the local entropy is the log of the number of optimal configuration within a hyper-sphere of radius O(N) or fractional volume O(1).

What we have learned from non-convex 1-2 layer NN learning random patterns?

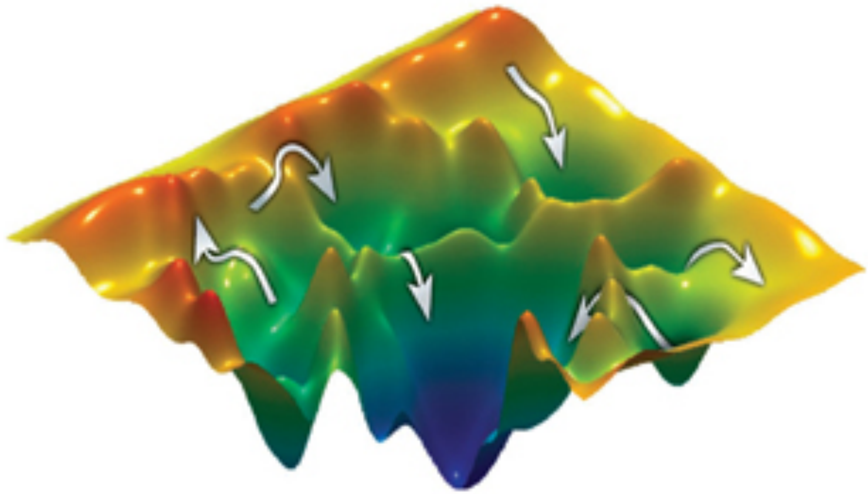
- ✓ The loss function presents an exponential proliferation of metastable states which trap SA or full batch Langevin dynamics

- ✓ HOWEVER, there exist rare dense regions (small but still of extensive size) which are accessible to simple non-detailed-balance stochastic algorithms. These regions have good generalisation capabilities.

- ✓ Accessibility and generalization are not in conflict

- ✓ The Local Entropy Measure amplifies the weight of these regions (from exponentially small to dominant!)

- ✓ shape of dense regions depends on the data, difficult to study analytically even for random patterns back to question:



Successful algorithms never “simply” minimize the loss.

Why?

Because the stationary measure of the stochastic learning process should not be the equilibrium Gibbs measure of the loss function. Many (simple!) out-of-equilibrium processes are attracted by the rare dense states (wide minima).

From Local entropy measure to the Robust Ensemble

$$\mathcal{P}(\tilde{W}) \propto e^{-y\mathcal{E}_d(\tilde{W})} \qquad \mathcal{E}_d(\tilde{W}) \doteq -\log \mathcal{N}(\tilde{W}, d)$$

We may write $\quad \mathcal{P}(\tilde{W}) \propto \lim_{\beta\to\infty} (\sum_{\{W\}} (e^{-\beta E(W)+\gamma W\cdot\tilde{W}})^y$

where γ is a Lagrange multiplier controlling the distance

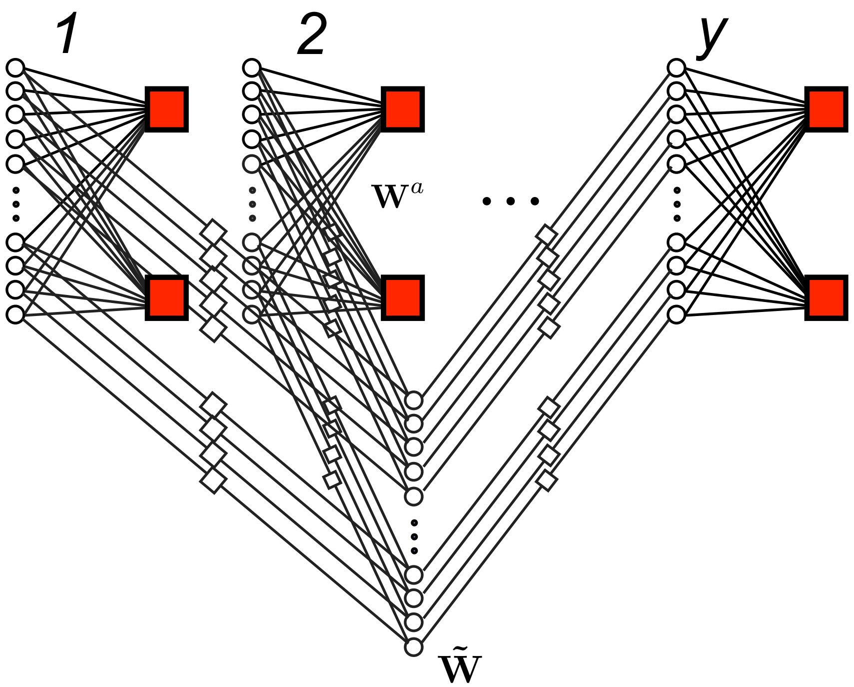
y integer ⇨ multiple real replicas

$$\mathcal{P}(\tilde{W}) \propto \lim_{\beta\to\infty} (\sum_{\{W\}} e^{-\beta E(W)+\gamma W\cdot\tilde{W}})^y = \lim_{\beta\to\infty} \prod_{a=1}^{y} \sum_{\{W^a\}} e^{-\beta E(W^a)+\gamma W^a\cdot\tilde{W}} =$$

$$= \lim_{\beta\to\infty} \sum_{\{W^1,W^2,\ldots,W^y\}} e^{-\beta \sum_{a=1}^{y} E(W^a)+\gamma \sum_{a=1}^{y} W^a\cdot\tilde{W}}$$

$$= \lim_{\beta\to\infty} \sum_{\{W^1,W^2,\ldots,W^y\}} e^{-\beta \sum_{a=1}^{y} E(W^a)+\gamma \sum_{a=1}^{y} \sum_{j=1}^{N} W_j^a \tilde{W}_j}$$

Robust Ensemble



$$\mathcal{P}_{RE}(\tilde{W}, \{W^a\}) \propto e^{-\beta \sum_{a=1}^{y} E(W^a) + \gamma \sum_{a=1}^{y} \sum_{j=1}^{N} W_j^a \tilde{W}_j}$$

Marginalizing the center

$$\hat{\mathcal{P}}_{RE}(\{W^a\}) \propto e^{-\beta(\sum_{a=1}^{y} E(W^a) - \frac{1}{\beta} \sum_j \log(2\cosh(\gamma \sum_{a=1}^{y} W_j^a)))}$$

Expectations of observables: $E[f(\tilde{W})] = \sum_{\tilde{W}} \sum_{\{W^a\}} f(\tilde{W})\, \mathcal{P}_{RE}(\tilde{W}, \{W^a\})$

Replicated MC

$$E(\mathbf{W}) = \sum_{a=1}^{y} E(W^a) - \frac{1}{\beta} \sum_j \log(2\cosh(\gamma \sum_{a=1}^{y} W_j^a))$$

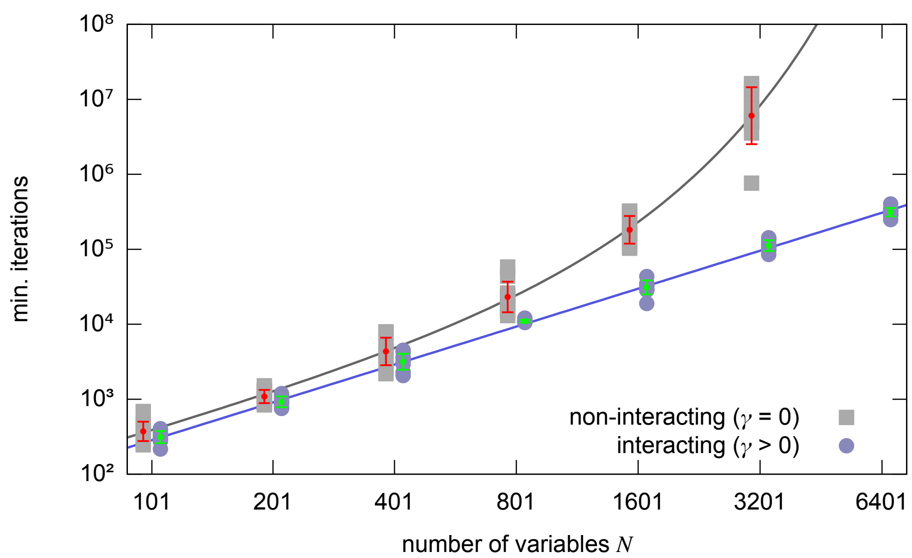
1) $\Delta E = E(\mathbf{W}') - E(\mathbf{W})$ can be computed efficiently when $\mathbf{W}'$ and $\mathbf{W}$ differ in one weight

2) efficient MC sampling for rejection rate reduction (non trivial)

3) most probable of the centroid value: $\tilde{W}_j = \text{sign} \sum_{a=1}^{y} W_j^a$ (typically $E\left(\tilde{W}\right) \leq \langle E\left(W^a\right)\rangle_a$)

$\alpha = 0.3$

$y = 3$



Notice: landscape of local minima could be different from the MC using BP

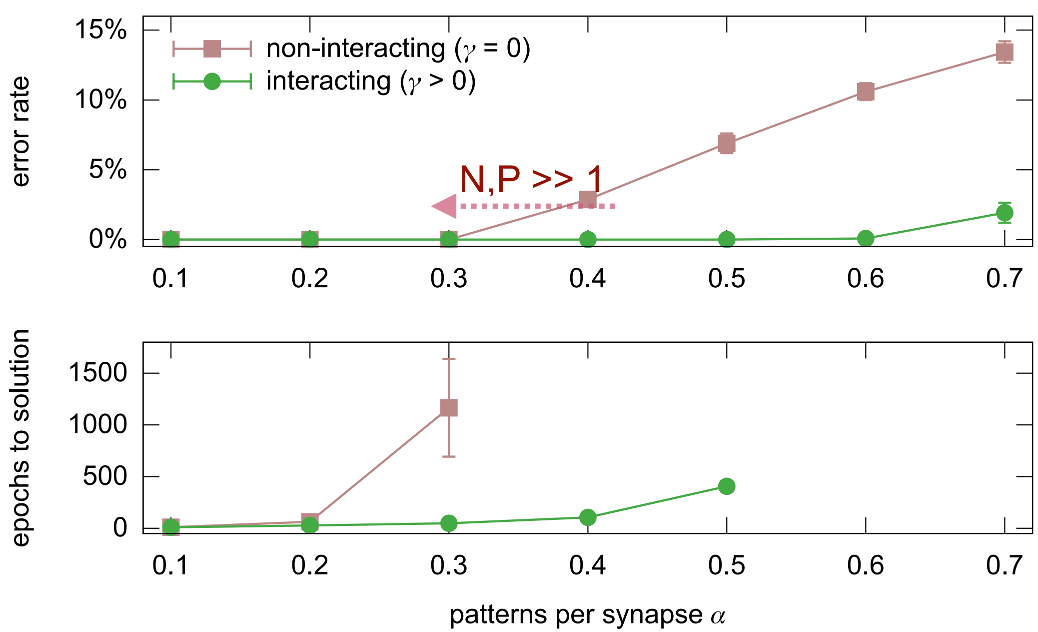
Replicated Stochastic Gradient Descent

$$H\left(\{W^a\}\right) = \sum_{a=1}^{y} E\left(W^a\right) + \frac{1}{\beta}\sum_{j=1}^{N}\log\left(e^{-\frac{\gamma}{2}\sum_{a=1}^{y}\left(W_j^a-1\right)^2} + e^{-\frac{\gamma}{2}\sum_{a=1}^{y}\left(W_j^a+1\right)^2}\right)$$

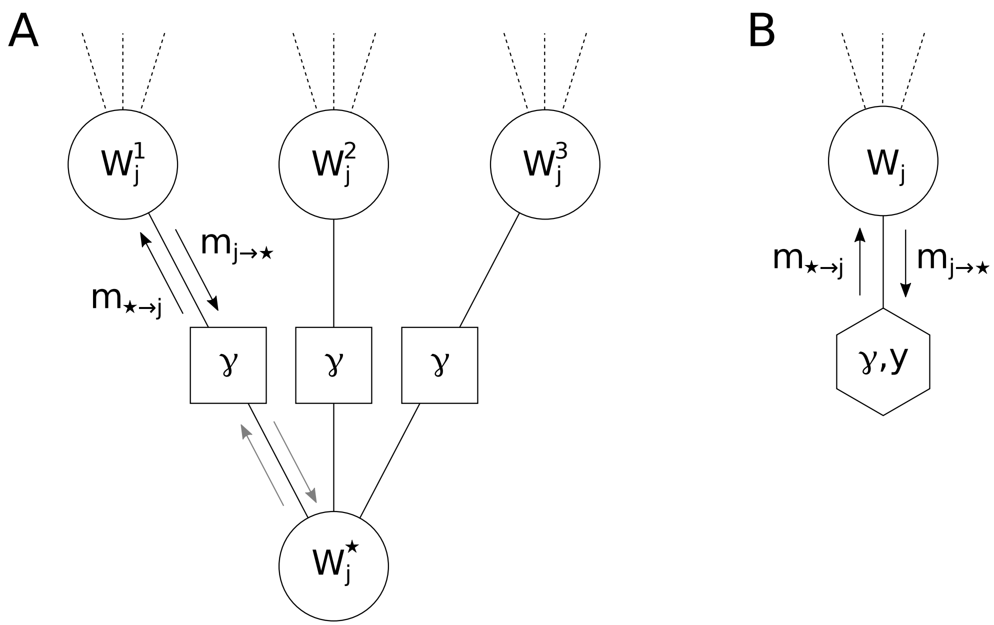
$$\frac{\partial H}{\partial W_i^a}\left(\{W^b\}\right) = \frac{\partial E}{\partial W_i}\left(W\right)\Big|_{W=W^a} + \frac{\gamma}{\beta}\left(\tanh\left(\gamma\sum_{b=1}^{y}W_i^b\right) - W_i^a\right)$$

$$\left(\mathcal{W}_i^a\right)^{t+1} = \left(\mathcal{W}_i^a\right)^{t} - \eta\frac{1}{|m\left(t\right)|}\sum_{\mu\in m(t)}\frac{\partial E^\mu}{\partial W_i}\left(W\right)\Big|_{W=(W^a)^t} + \eta'\left(\tanh\left(\gamma\sum_{b=1}^{y}\left(W_i^b\right)^t\right) - \left(W_i^a\right)^t\right) \qquad \eta' = \frac{\gamma}{\beta\eta}$$

N = 1605 weights   K = 5



Replicated Belief Propagation:
focusing BP (fBP) ~ BP with reinforcement



$$\{W_j^a\}_{a=1}^y$$

$$P_j\left(\{W_j^a\}_{a=1}^y\right) = P_j\left(\sum_{a=1}^y W_j^a\right)$$

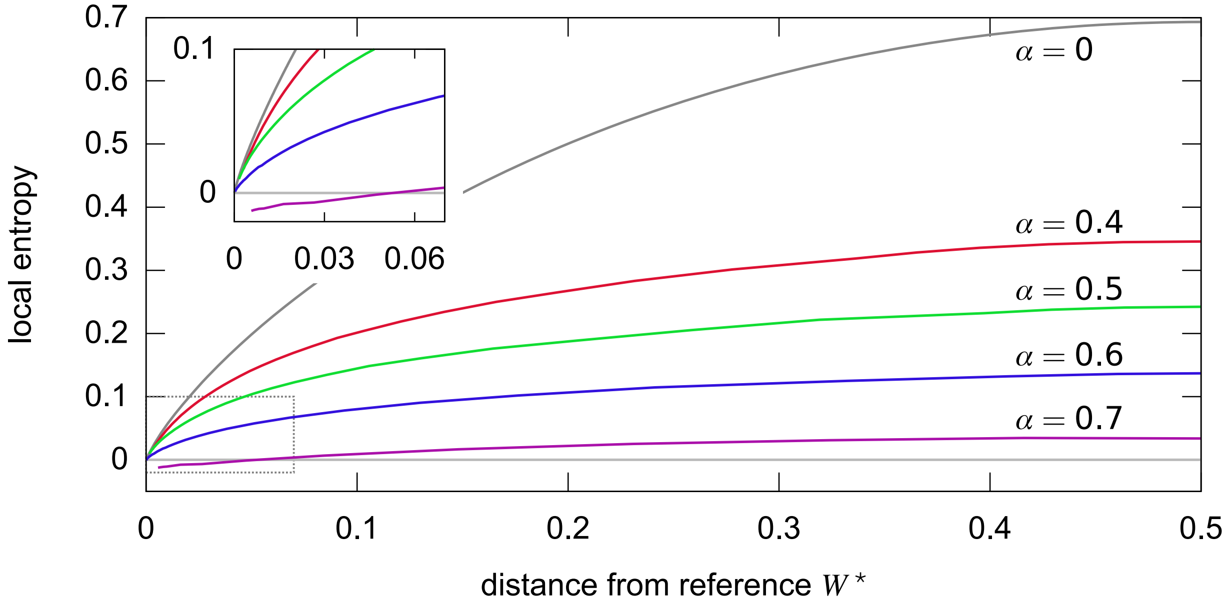
assume that each replica of the system behaves in exactly the same way, and therefore that the same messages are exchanged along the edges of the graph regardless of the replica index. …single system, which is identical to the original one except that each variable now also exchanges messages with y − 1 identical copies of itself through an auxiliary variable (which we can just trace away at this point).

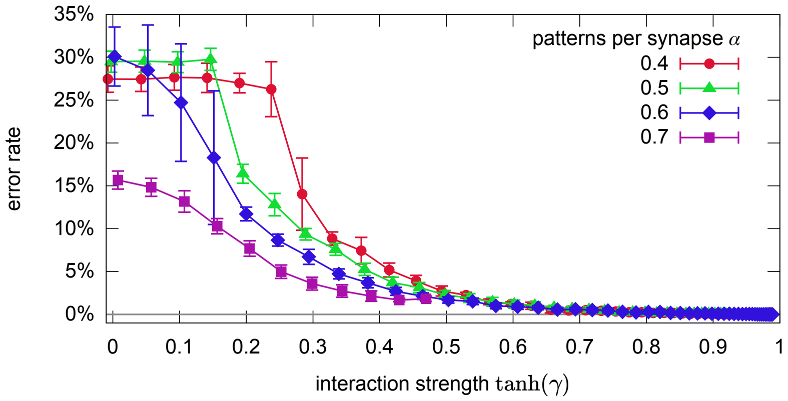
extra message at time t:

$$m_{\star\to j}^{t+1} = \tanh\left((y-1)\tanh^{-1}\left(m_{j\to\star}^t \tanh\gamma\right)\right)\tanh\gamma$$ *focusing BP*

fBP becomes a solver looking for high density regions of solution. Interesting convergence properties (to be further studied).

Replicated BP is also an analytical tool: phase diagram on NN with one hidden layer





committee machine with N = 1605, K = 5, y = 7, increasing γ from 0 to 2.5, averages on 10 samples. Top: local entropy versus distance to the reference W* for various α (error bars not shown for clarity). The topmost grey curve (α = 0) is an upper bound, representing the case where all configurations within some distance are solutions.

Unreasonable effectiveness of learning neural networks: From accessible states and robust ensembles to basic algorithmic schemes, C. Baldassi, C. Borgs, J.T. Chayes, A. Ingrosso, C. Lucibello, L. Saglietti and Riccardo Zecchina, PNAS 113, E7655-E7662 (2016)

The case of Deep Networks

Modified sampling measure: $$P(x') \propto e^{y\Phi(x')}$$

Local free entropy: $$\boxed{\Phi(x') = \log \int_x e^{-\beta f(x)} e^{-\lambda \|x-x'\|^2} dx}$$

where: x are the continuous weights, and $f(x)$ is the loss/energy function

Langevin dynamics:

Algorithm 1: Entropy-SGD algorithm

Input : current weights x, Langevin iterations L
Hyper-parameters : scope γ, learning rate η, SGLD step size η'

// SGLD iterations;
$x', \mu \leftarrow x$;
for $\ell \leq L$ **do**
  $\Xi^\ell \leftarrow$ sample mini-batch;
  $dx' \leftarrow \frac{1}{m} \sum_{i=1}^{m} \nabla_{x'} f\left(x';\, \xi_{\ell_i}\right) - \gamma\, (x - x')$;
  $x' \leftarrow x' - \eta'\, dx' + \sqrt{\eta'}\, \varepsilon\, \mathrm{N}(0, \mathrm{I})$;
  $\mu \leftarrow (1-\alpha)\mu + \alpha\, x'$;

// Update weights;
$x \leftarrow x - \eta\, \gamma\, (x - \mu)$

ENTROPY-SGD: BIASING GRADIENT DESCENT INTO WIDE VALLEYS
Pratik Chaudhari A. Choromanska, S. Soatto, Y. LeCun, C.Baldassi, C. Borgs, J.Chayes, L.Sagun, R. Zecchina, ICLR 2017

Local Entropy & Robust Ensemble ⟺ Elastic Averaging SGD with momentum

$\beta \to \infty$ sampling from

$$\mathcal{P}_{RE}(\tilde{x}) \propto \int dx^1 \ldots dx^y e^{-\beta\phi(\tilde{x},\{x^a\})}$$

$$\phi(\tilde{x},\{x^a\}) = \sum_{a=1}^{y} E(x^a) + \frac{\lambda}{\beta}\sum_{a=1}^{y}\|x^a - \tilde{x}\|^2$$

$$\min_{x^1,\ldots,x^p,\tilde{x}} \sum_{i=1}^{p}\Big(\mathbb{E}[f(x^i,\xi^i)] + \frac{\rho}{2}\|x^i - \tilde{x}\|^2\Big)$$

Work in progress with L. Bottou, L. Sagun, J. Chayes, C. Borgs, C. Baldassi, …

T. Chen, E.B. Fox and C. Guestrin Stochastic gradient hamiltonian monte carlo. 1683–1691, 2014. In ICML, 2014

Deep learning with Elastic Averaging SGD

Sixin Zhang
Courant Institute, NYU
zsx@cims.nyu.edu

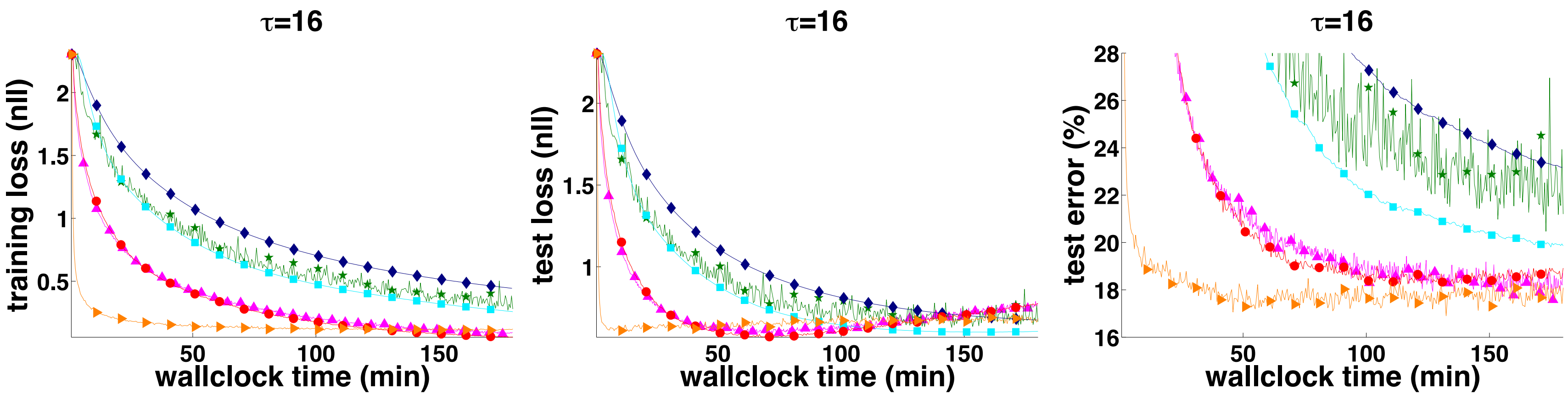
Anna Choromanska
Courant Institute, NYU
achoroma@cims.nyu.edu

Yann LeCun
Center for Data Science, NYU & Facebook AI Research
yann@cims.nyu.edu

Loss function:

$$\min_{x^1,\ldots,x^p,\tilde{x}} \sum_{i=1}^{p} \mathbb{E}[f(x^i,\xi^i)] + \frac{\rho}{2}\|x^i - \tilde{x}\|^2,$$

with *Momentum SGD*.



CIFAR dataset with the 7-layer convolutional neural network.

The idea of Flat Minima is not new:

Hochreiter, Sepp and Schmidhuber, Ju¨rgen. Flat minima. *Neural Computation*, 9(1):1–42, 1997

and many others …

ON LARGE-BATCH TRAINING FOR DEEP LEARNING: GENERALIZATION GAP AND SHARP MINIMA

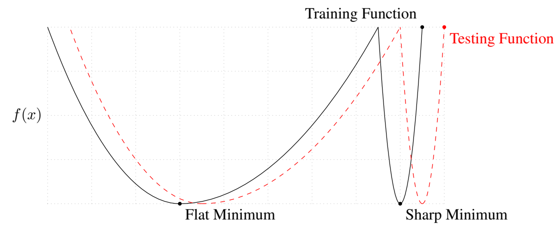
Nitish Shirish Keskar*
Northwestern University
Evanston, IL 60208
keskar.nitish@u.northwestern.edu

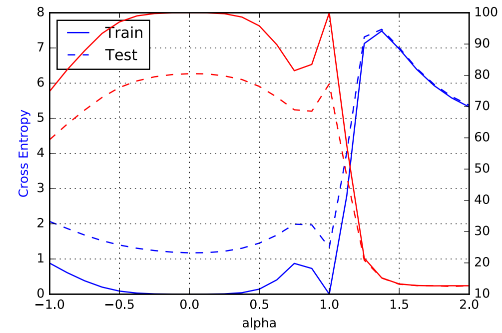
Dheevatsa Mudigere
Intel Corporation
Bangalore, India
dheevatsa.mudigere@intel.com

Jorge Nocedal
Northwestern University
Evanston, IL 60208
j-nocedal@northwestern.edu

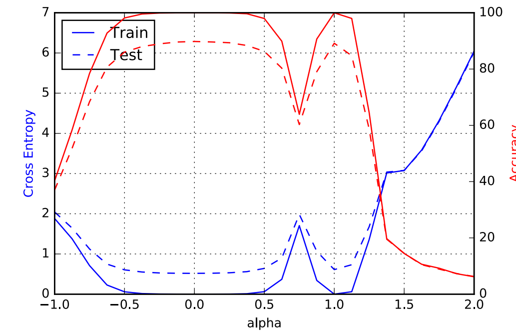
Mikhail Smelyanskiy
Intel Corporation
Santa Clara, CA 95054
mikhail.smelyanskiy@intel.com

Ping Tak Peter Tang
Intel Corporation
Santa Clara, CA 95054
peter.tang@intel.com





(c) C_1



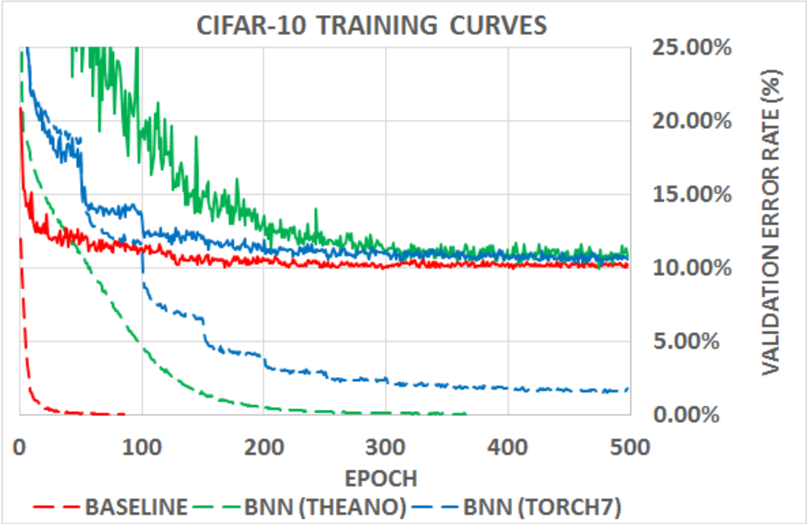
(d) C_2

C_1	(Shallow) Convolutional	Section B.3	CIFAR-10 (Krizhevsky & Hinton, 2009)
C_2	(Deep) Convolutional	Section B.4	CIFAR-10

Binarized Neural Networks: Training Neural Networks with Weights and Activations Constrained to $+1$ or -1

Matthieu Courbariaux*[1] MATTHIEU.COURBARIAUX@GMAIL.COM
Itay Hubara*[2] ITAYHUBARA@GMAIL.COM
Daniel Soudry[3] DANIEL.SOUDRY@GMAIL.COM
Ran El-Yaniv[2] RANI@CS.TECHNION.AC.IL
Yoshua Bengio[1,4] YOSHUA.UMONTREAL@GMAIL.COM

[1]Université de Montréal
[2]Technion - Israel Institute of Technology
[3]Columbia University
[4]CIFAR Senior Fellow
*Indicates equal contribution. Ordering determined by coin flip.



"Although BNNs are slower to train, they are nearly as accurate as 32-bit float DNNs. "

Conclusion and what next

Theoretical framework:

out-of-equilibrium statistical physics and large deviations studies are a key framework for understanding learning phenomena

Next algorithmic developments:

- Accessible dense states in DNN, connections with regularization techniques (dropout), temporal version of local entropy
- An opportunity for acceleration?
- Simple forms of stochastic learning process
- Learning with low precision weights: can we design new neural hardware?
- Generalization to unsupervised learning
- …