Exploring the Role of Liquid States in Neural Networks: From Feedforward Networks to Attractor Models

Riccardo Zecchina

Department of Computing Sciences
Bocconi University, Milan

ellis
European Laboratory for Learning and Intelligent Systems

Plan of the talk

1. Accessible liquid states in non-convex feedforward models and negative margin classifiers

2. Geometry of perfect learning: differences between discrete and continuous weights

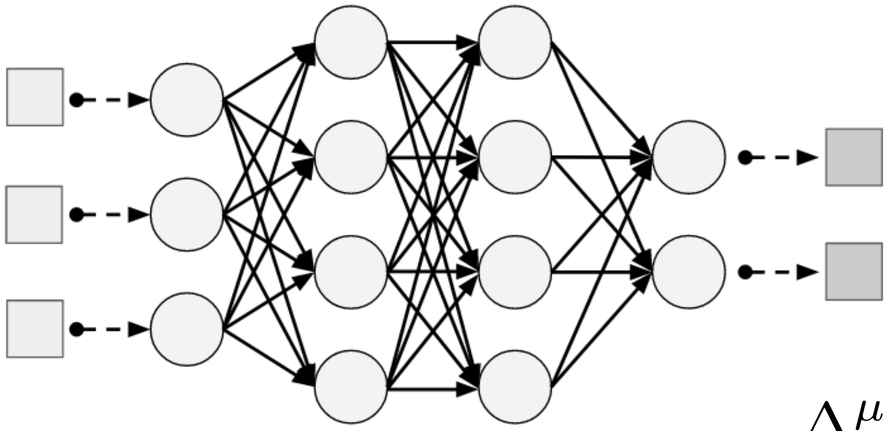
3. Chaotic and stable liquid attractors in recurrent asymmetric neural networks

Basic problems

- Overparametrization & generalization (overfitting under control)
- Non-convexity and first-order algorithms
- Internal representations if feedforward and attractor NN

Let's consider feedforward architecture first

$x^\mu \rightarrow$



$\hat{y}^\mu \leftrightarrow y^\mu$

Training set:

$$\{(x^\mu, y^\mu)\}_{\mu=1,\ldots,M}$$

$$\Delta^\mu \doteq (W^K \cdot \sigma_K(x^\mu))\, y^\mu$$

pre-activations at layer K

$$\hat{y}^\mu = \operatorname{argmax}\left(W^K \sigma\left(W^{K-1} \sigma\left(\ldots \sigma\left(W^2 \sigma\left(W^1 x^\mu\right)\right)\right)\right)\right)$$

output (label)

weights (matrices)

input (vector)

- Energy = “0-1 loss”: number of errors on the training set (not differentiable)

$$\mathcal{L}_{NE} = \sum_{\mu} (1 - \delta(\hat{y}^{\mu}, y^{\mu}))$$

- Surrogate differentiable losses

Cross-entropy: “log-likelihood”, softmax

$$\mathcal{L}_{CE} = -\sum_{\mu} (\hat{\Delta}^{\mu}_{y^{\mu}} - \log \sum_{k} \exp \gamma \hat{\Delta}^{\mu}_{k})$$

$$\Delta^{\mu} \triangleq (W^{K} \cdot \sigma_{K}(x^{\mu}))\, y^{\mu}$$

Mean Square Error, …

The simplest non convex models:
binary perceptron and negative stability spherical perceptrons

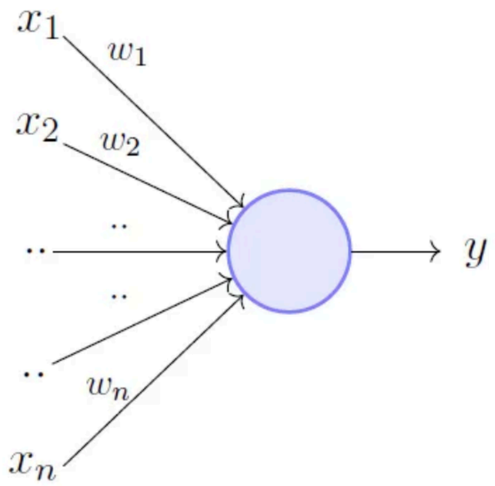
Random Training set: $\{(\bar{x}^\mu, y^\mu)\} \quad \mu = 1, ..., P = \alpha N$

$$x_i^\mu = \pm 1 \quad with \quad p = \frac{1}{2} \quad , \quad i = 1, ..., N$$

$$y^\mu = \pm 1 \quad with \quad p = \frac{1}{2}$$

Control parameter: $$\alpha = \frac{\#\,\text{patterns}}{\#\,\text{weigts}}$$

Learning: find W_i such that



$$y^\mu = \text{Sign}(\sum_{i=1}^{N} W_i x_i^\mu) = \text{Sign}(W \cdot x^\mu) \qquad \forall \mu$$

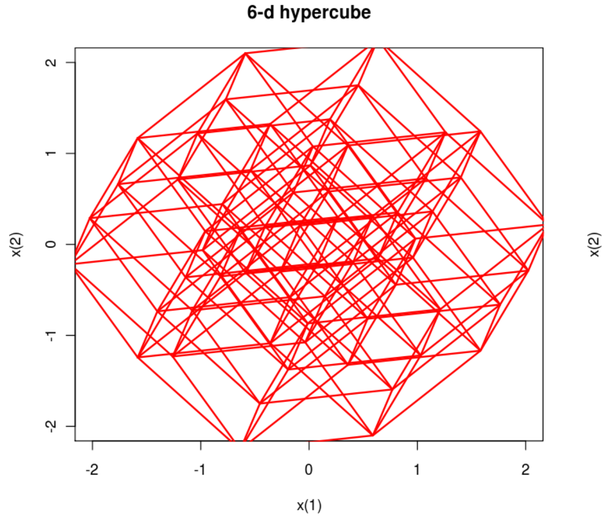
$$1 = \text{Sign}(\Delta^\mu) \quad \text{or} \quad \Delta^\mu > 0$$

with $$\Delta^\mu \equiv y^\mu \sum_{i=1}^{N} W_i x_i^\mu = y^\mu\, W \cdot x^\mu$$ (stability, margin)

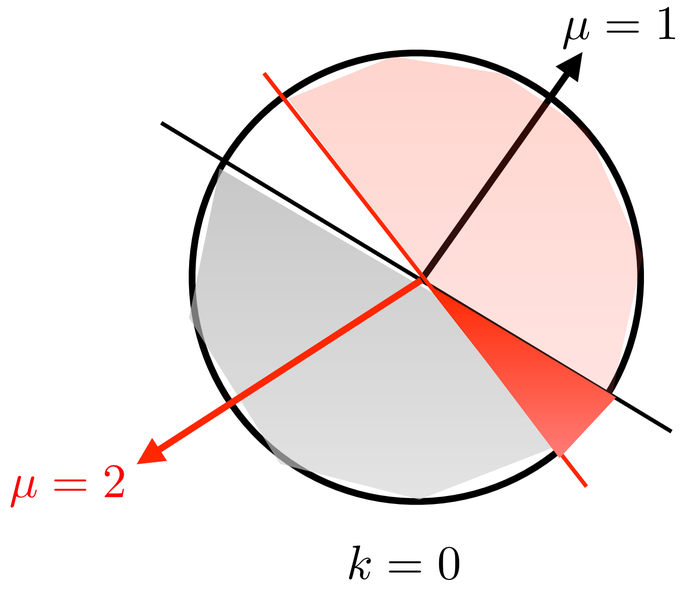
minimal non convex models

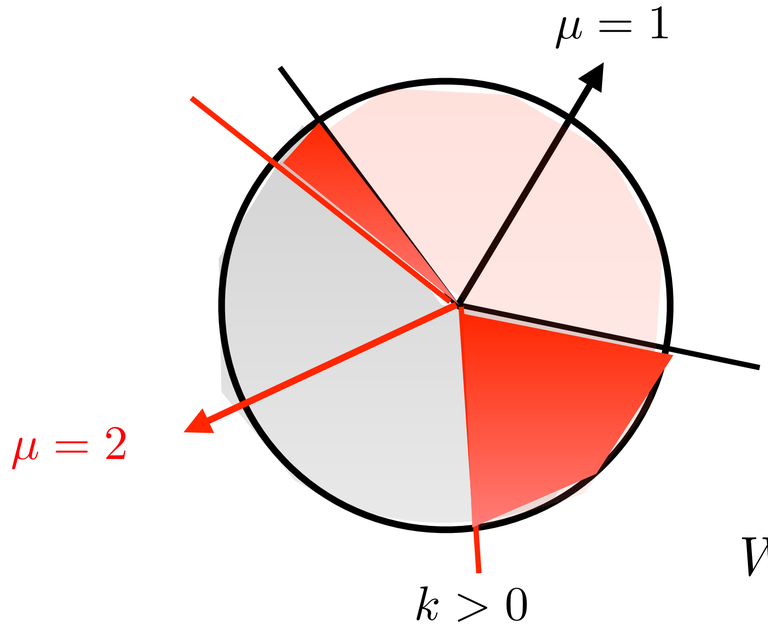
Binary perception: $W_i \in \{-1, +1\}$

With (negative) margin: $W_i \in \mathbb{R}$ & $\sum_{i=1}^{N} W_i^2 = N$

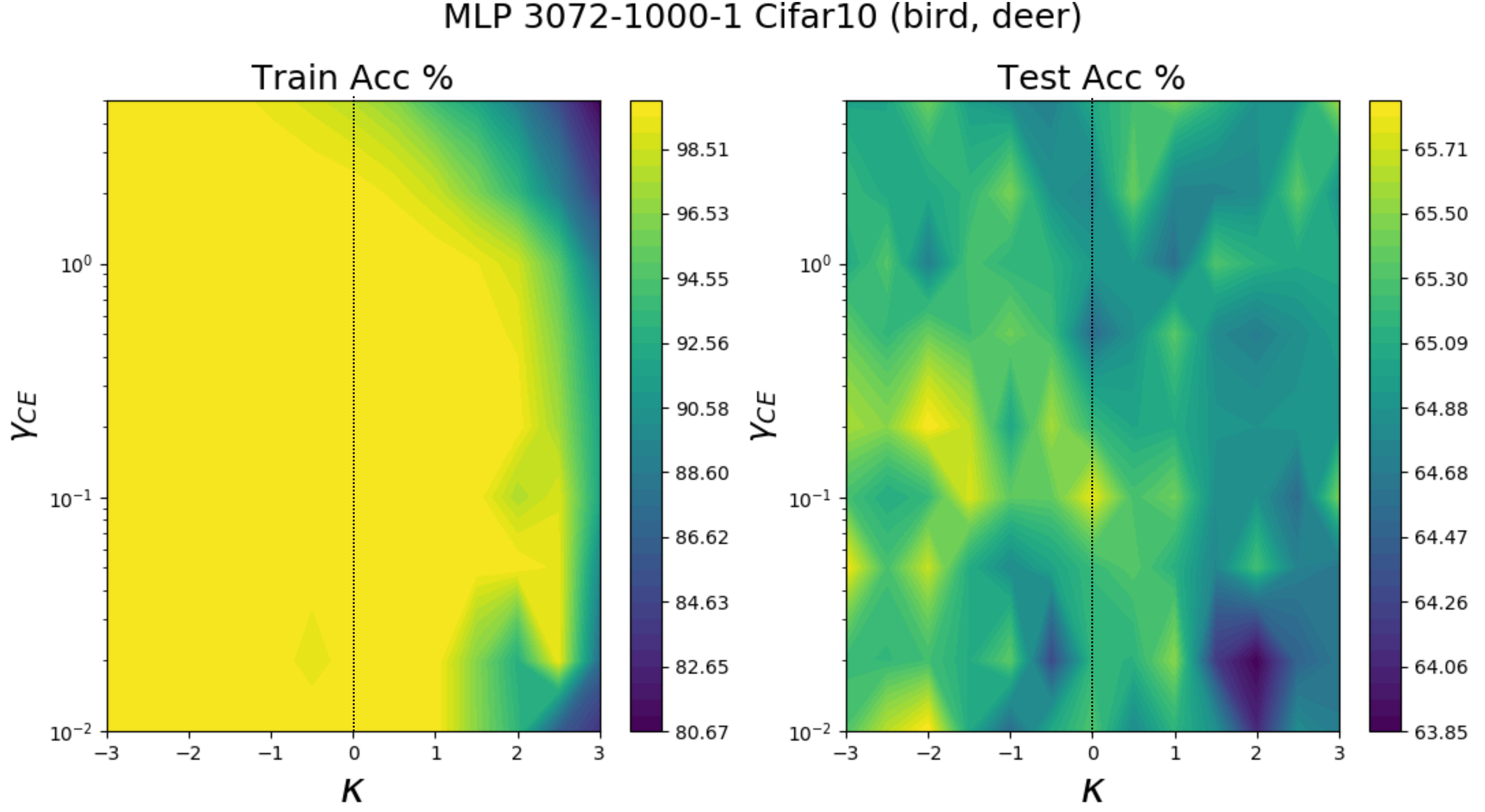


$$1 = \text{Sign}(\Delta^\mu - k) \quad \text{or} \quad \Delta^\mu > k$$



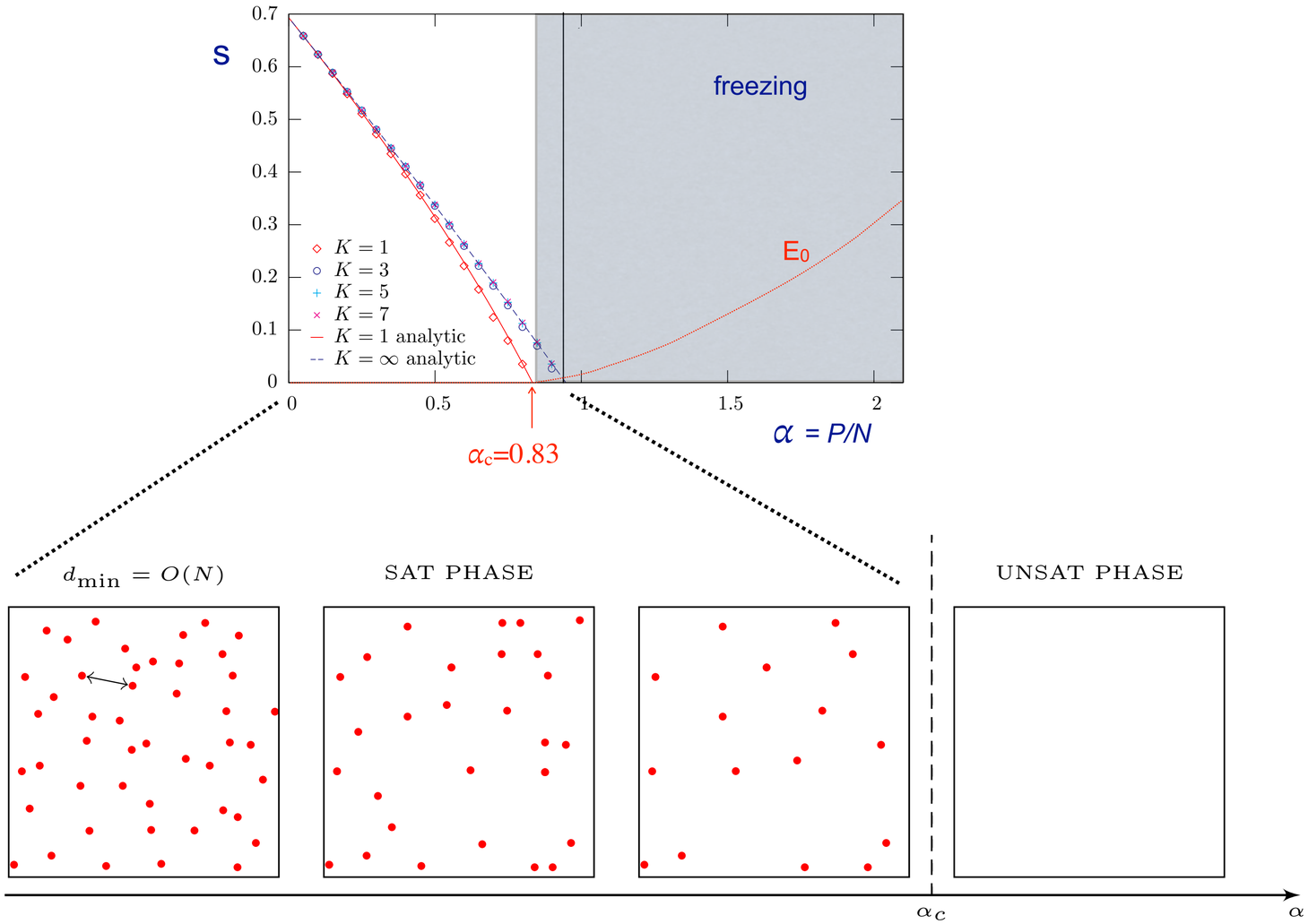


$W \cdot x > -|k|$


MLP 3072-1000-1 Cifar10 (bird, deer)
Train Acc %
Test Acc %
γ_{CE}
K
10^{0}
10^{-1}
10^{-2}
−3
−2
−1
0
1
2
3
98.51
96.53
94.55
92.56
90.58
88.60
86.62
84.63
82.65
80.67
65.71
65.50
65.30
65.09
64.88
64.68
64.47
64.26
64.06
63.85

Start with the Binary perceptron

- The space of solution splits into separated states of vanishing entropy (Krauth, Mézard (1989));
- $\forall \alpha > 0$ typical solutions are isolated (Huang, Kabashima (2014));
- Rigorous: Abbe, Li, Sly (2021), Perkins, Xu (2021), Nakajima, Sun (2022), Aubin, Perkins, Zdeborova (2019)



If this would be the case, locally stable algorithms would fail and learning would be hard:
The so called *Overlap Gap Property* would hold

PERSPECTIVE

The overlap gap property: A topological barrier to optimizing over random structures

David Gamarnik Authors Info & Affiliations

Edited by David L. Donoho, Stanford University, Stanford, CA, and approved August 19, 2021 (received for review May 5, 2021)

October 1, 2021 | 118 (41) e2108492118 | https://doi.org/10.1073/pnas.2108492118

If this would be the case, locally stable algorithms would fail and learning would be hard:
The so called *Overlap Gap Property* would hold

PERSPECTIVE

The overlap gap property: A topological barrier to optimizing over random structures

David Gamarnik Authors Info & Affiliations

Edited by David L. Donoho, Stanford University, Stanford, CA, and approved August 19, 2021 (received for review May 5, 2021)

October 1, 2021 | 118 (41) e2108492118 | https://doi.org/10.1073/pnas.2108492118

This contradicts empirical evidence!

If this would be the case, locally stable algorithms would fail and learning would be hard:
The so called *Overlap Gap Property* would hold

PERSPECTIVE

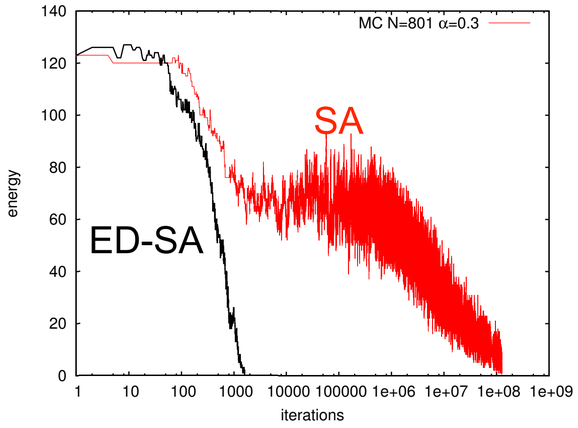
The overlap gap property: A topological barrier to optimizing over random structures

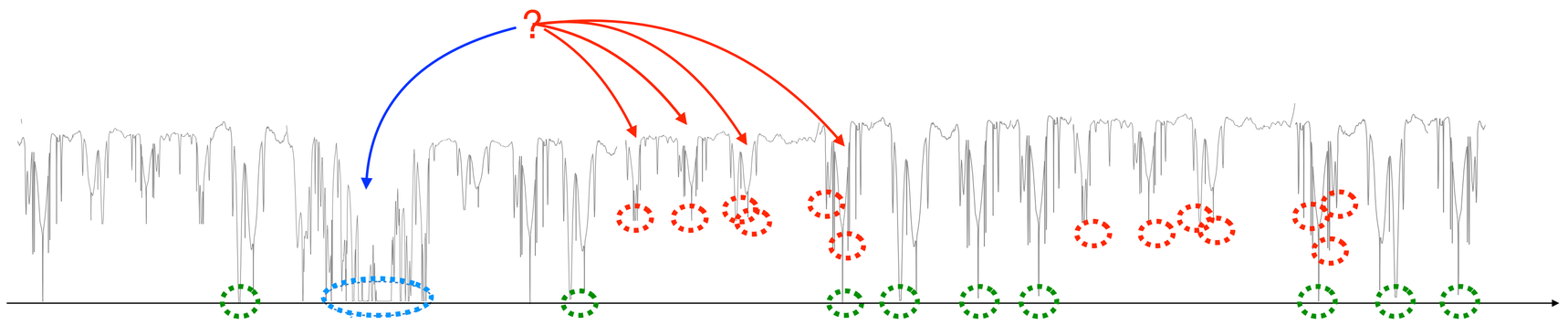
David Gamarnik Authors Info & Affiliations

Edited by David L. Donoho, Stanford University, Stanford, CA, and approved August 19, 2021 (received for review May 5, 2021)

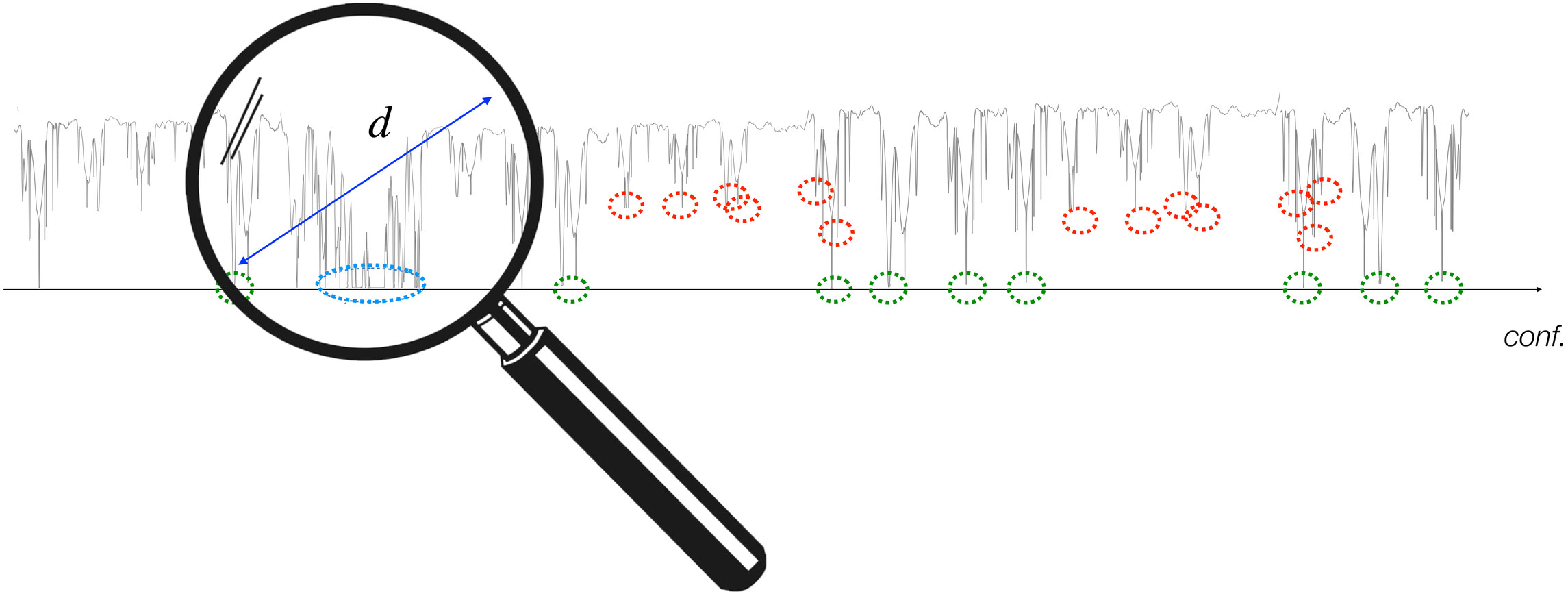
October 1, 2021 | 118 (41) e2108492118 | https://doi.org/10.1073/pnas.2108492118

This contradicts empirical evidence!





conf.

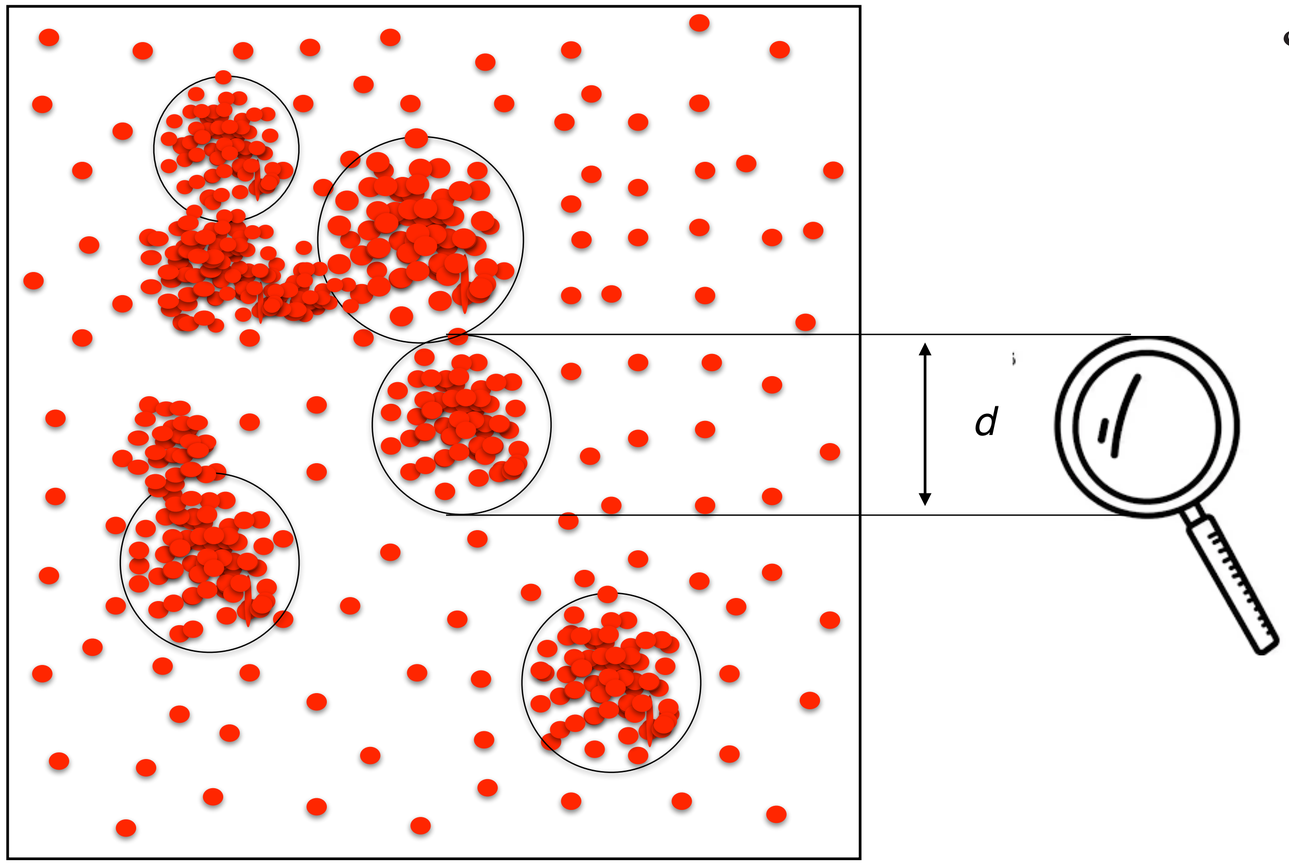

d
conf.

1) Regions observed with a diameter *d*

2) Internal entropy: $\mathcal{S}_I(d,y) = -\langle \mathcal{E}(\tilde{W}) \rangle_{\xi,\tilde{W}} = \frac{1}{N} \langle \log \mathcal{N}(\tilde{W}, d) \rangle_{\xi,\tilde{W}}$ $\mathcal{S}_I(d,y) = \partial_y (y\mathcal{F}(d,y))$

3) External entropy (number of regions of diameter *d* with a given internal entropy):

$\mathcal{S}_E(d,y) = -y[\mathcal{F}(d,y) + \mathcal{S}_I(d,y)]$



Local entropy measure

Bias the statistical measure towards dense (wide, flat) regions

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

solution at distance d

$$\mathbb{X}_{\xi}\left(W\right) = \lim_{\beta \to \infty} e^{-\beta \mathcal{L}_{NE}(W)}$$

indicator function

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

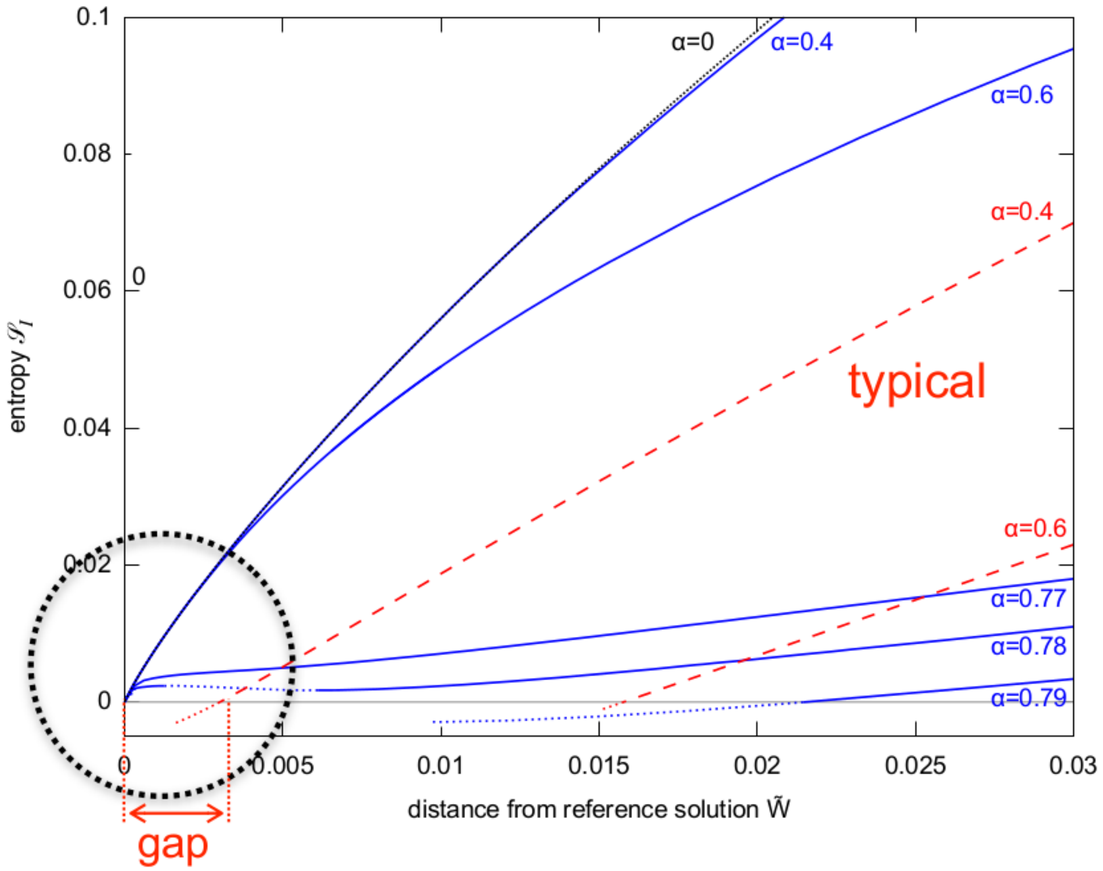
"**local entropy**" (the log of the number of solutions in hypersphere of radius *d*)

$$\Phi(\tilde{W}, \beta, \gamma) = \ln \sum_{\{W\}} e^{-\beta \mathcal{L}_{NE}(W) - \gamma d(\tilde{W}, W)}$$

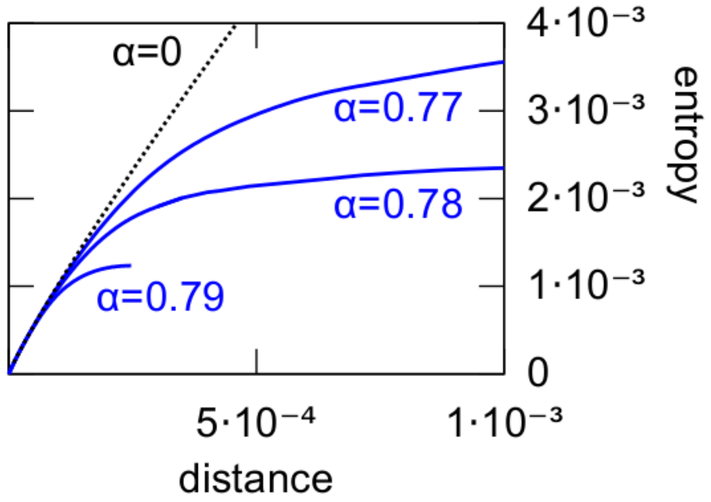
"**local free entropy**"

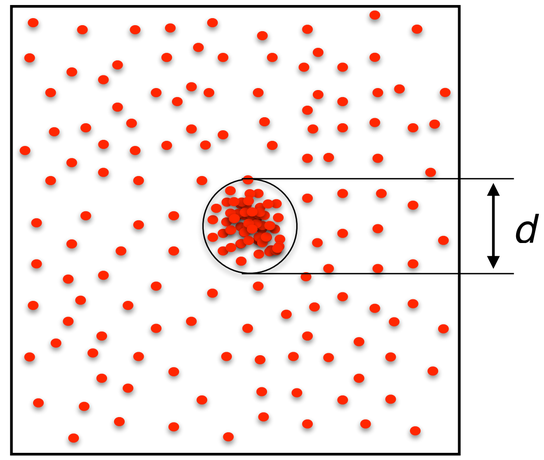
C. Baldassi, A. Ingrosso, C. Lucibello, L. Saglietti, and R. Zecchina Phys. Rev. Lett. 115, 128101 (2015)

Check the existence of subdominant dense regions of solutions in binary networks



ultra-dense cluster





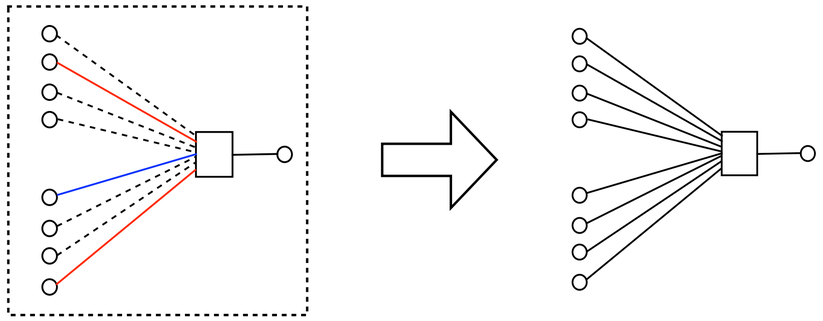
Geometrical phase discontinuous transition: $\alpha_U \simeq 0.75$

C. Baldassi, A. Ingrosso, C. Lucibello, L. Saglietti, and R. Zecchina Phys. Rev. Lett. 115, 128101 (2015)

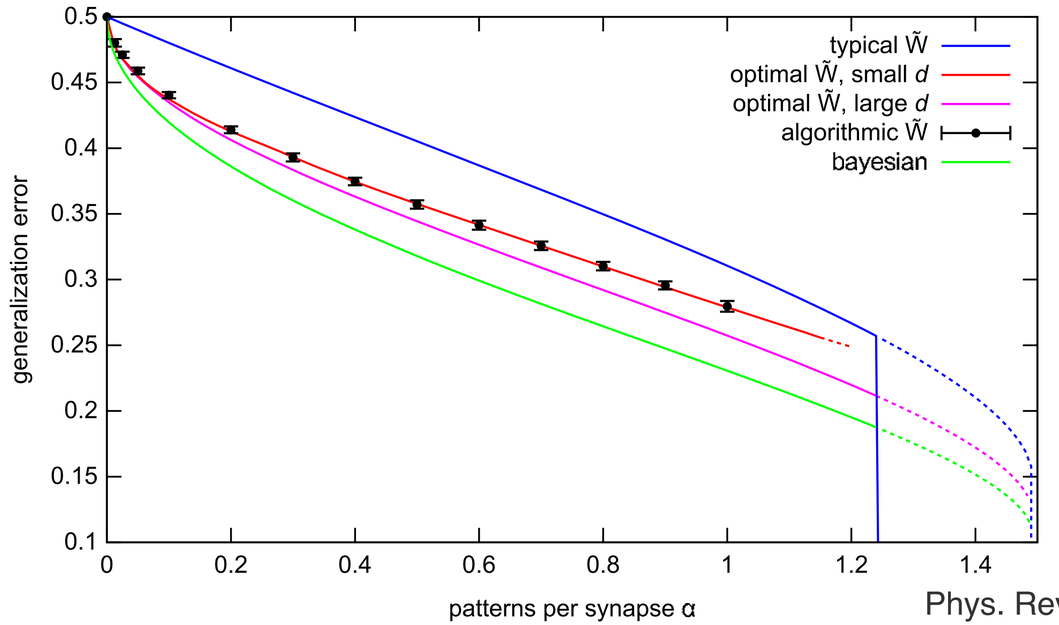
Generalization

Teacher

Student



The teacher is isolated



Phys. Rev. Lett. **115**, 128101 (2015)

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

Algorithmic follow up: real replicas

Local free entropy: $\phi(W,\gamma,\beta) = \log \sum_{W'} e^{-\beta \mathcal{L}_{NE}(W') - \frac{\gamma}{2} d(W,W')}$ $\qquad (\mathcal{L}_{NE} \equiv \mathcal{L}_{0-1})$

Large-deviation partition function: $Z(y,\gamma,\beta',\beta) = \sum_{W} e^{-\beta' \mathcal{L}_{NE}(W) + y\,\phi(W,\gamma,\beta)}$

Assume y integer, and transform the partition function

$$Z(y,\gamma,\beta',\beta) = \sum_{W,\{W_a\}} e^{-\beta' \mathcal{L}_{NE}(W) - \beta \sum_{a=1}^{y} \mathcal{L}_{NE}(W_a) - \frac{\gamma}{2} \sum_{a=1}^{y} d(W,W_a)}$$

interaction

center

y (real) replicas

Algorithmic follow up: real replicas

Local free entropy: $\phi(W,\gamma,\beta) = \log \sum_{W'} e^{-\beta \mathcal{L}_{NE}(W') - \frac{\gamma}{2} d(W,W')}$ $\qquad (\mathcal{L}_{NE} \equiv \mathcal{L}_{0-1})$

Large-deviation partition function: $Z(y,\gamma,\beta',\beta) = \sum_{W} e^{-\beta' \mathcal{L}_{NE}(W) + y\,\phi(W,\gamma,\beta)}$

Assume y integer, and transform the partition function

$$Z(y,\gamma,\beta',\beta) = \sum_{W,\{W_a\}} e^{-\beta' \mathcal{L}_{NE}(W) - \beta \sum_{a=1}^{y} \mathcal{L}_{NE}(W_a) - \frac{\gamma}{2} \sum_{a=1}^{y} d(W,W_a)}$$

interaction

center

y (real) replicas

Local entropy algorithms

- Local Entropy driven Simulated Annealing
- Replicated Message-Passing (aka “focusing Belief Propagation”)
- Replicated Stochastic Gradient Descent (SGD)
- Entropy-SGD: Langevin dynamics to estimate local entropy
- Replicated Greedy Algorithms
- Sharpness Aware Minimization
- Stochastic weights + gradient on the probabilities
- Quantum Annealing delocalization mechanism for finding NN ground states
- …

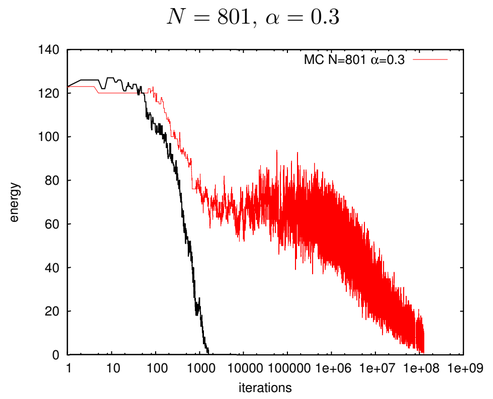
Local Entropy driven Simulated Annealing

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. BP method to estimate the local entropy,
   or use a replicated model



Connection with Monasson's 1-rsb formalism (1995)

1) We are interested in the out-of-equilibrium regime *m >1*

2) q_1 *=1-2 D* is the diameter of the hypersphere. We do not need to maximise over q_1

3) If we find a value of q_1 that maximises the free entropy then we have an out-of-equilibrium state (*m >1*) as in Monasson

$$Z_{1RSB}\left(\beta', y, D\right) = \sum_{\{W^a\}} e^{-\beta' \sum_{a=1}^{y} E(W^a)} \prod_{a>b} \delta\left(d\left(W^a, W^b\right) - ND\right)$$

$$Z_{1RSB}\left(\beta', y\right) = \max_{D} Z_{1RSB}\left(\beta', y, D\right)$$

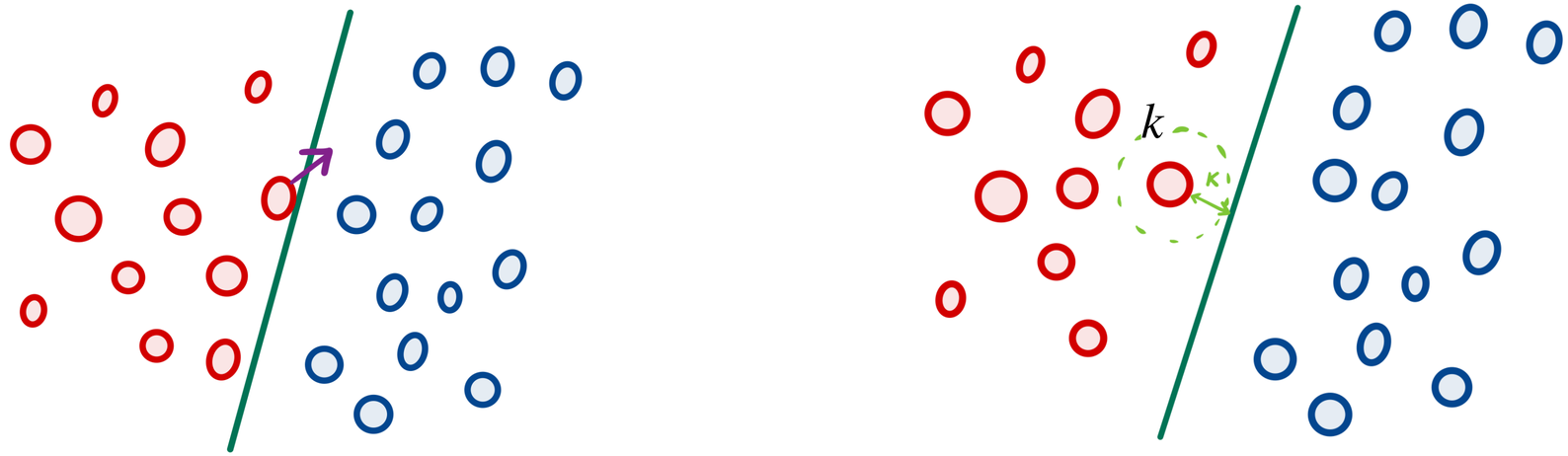
R. Monasson, Physical review letters, 75(15):2847, 1995.

C. Baldassi, F. Pittorino. R. Zecchina, PNAS 117, 161-179 (2020)

- How are these dense regions composed?
- Why they generalise well?
- Do they contain high margin solutions?

Solutions with margin *k*:

$$\mathbb{X}_{\xi,F}(W;\kappa) = \prod_{\mu=1}^{P} \Theta\left(y^{\mu}\sigma^{\mu}_{\text{out}} - \kappa\right)$$

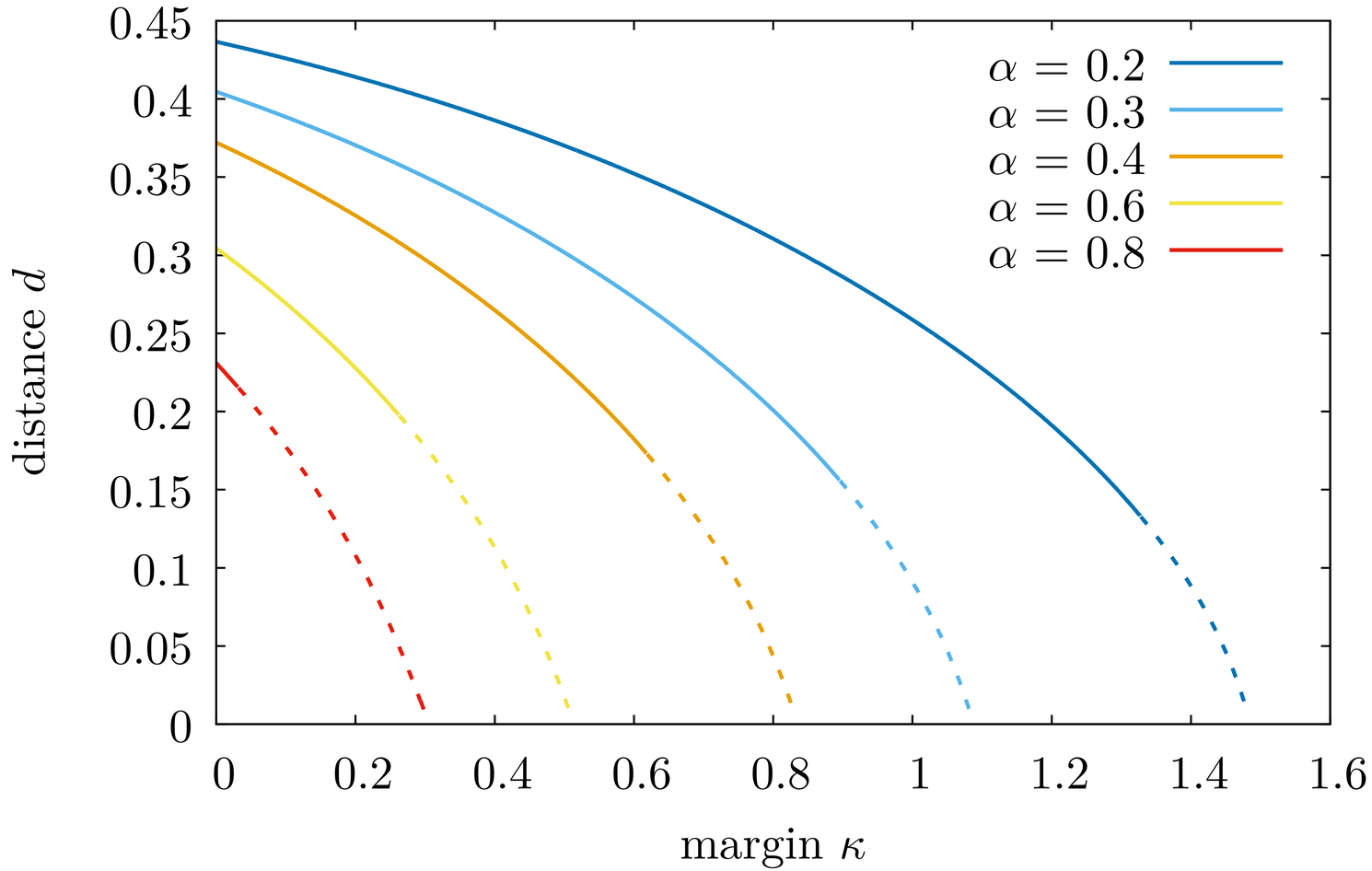


$$E = \sum_{\mu=1}^{P} \Theta\left(-y^{\mu}\sigma^{\mu}_{\text{out}} + \kappa\right)$$

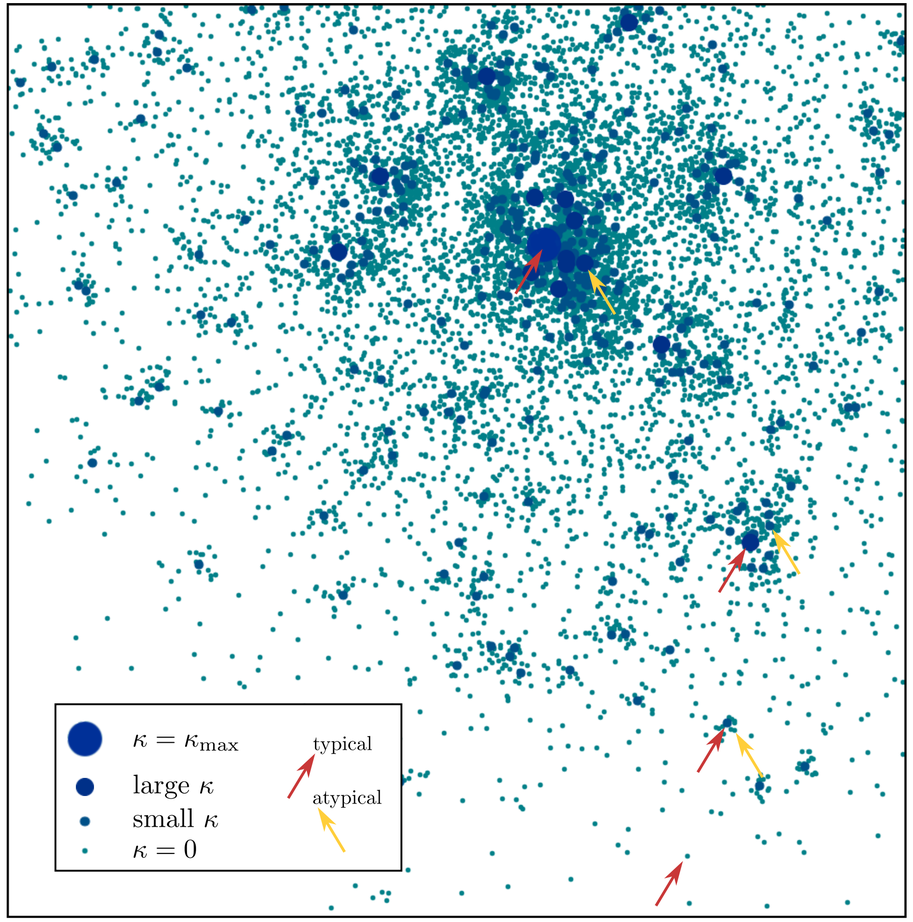
Typical high margin solutions are less but tend to be much closer to each other

The lines change from solid to dashed when the entropy of solutions becomes negative, i.e. when $\kappa = \kappa_{max}$

distance d=(1-q_1)/2
between typical solutions
for a given margin



A wide flat minima arises by the coalescence of (atypical) high margin minima

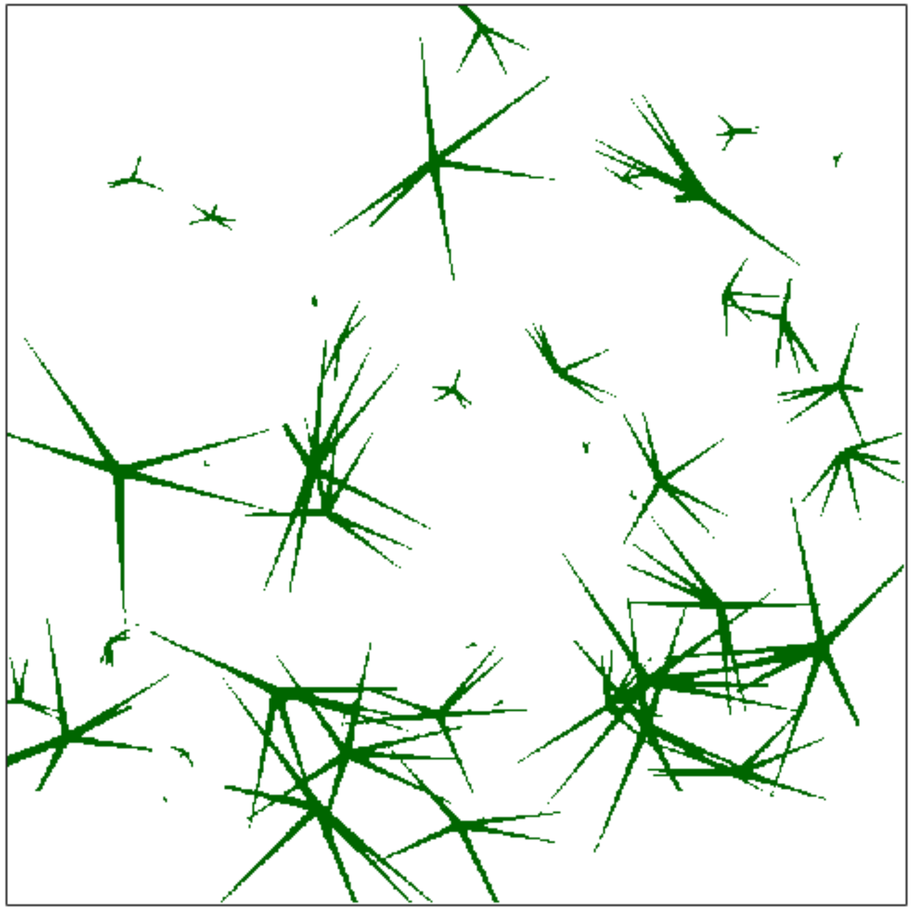


Unveiling the structure of wide flat minima in neural networks

Carlo Baldassi,[1] Clarissa Lauditi,[2] Enrico M. Malatesta,[1] Gabriele Perugini,[1] and Riccardo Zecchina[1]

[1] *Artificial Intelligence Lab, Bocconi University, 20136 Milano, Italy*
[2] *Department of Applied Science and Technology, Politecnico di Torino, 10129 Torino, Italy*

arXiv:2107.01163



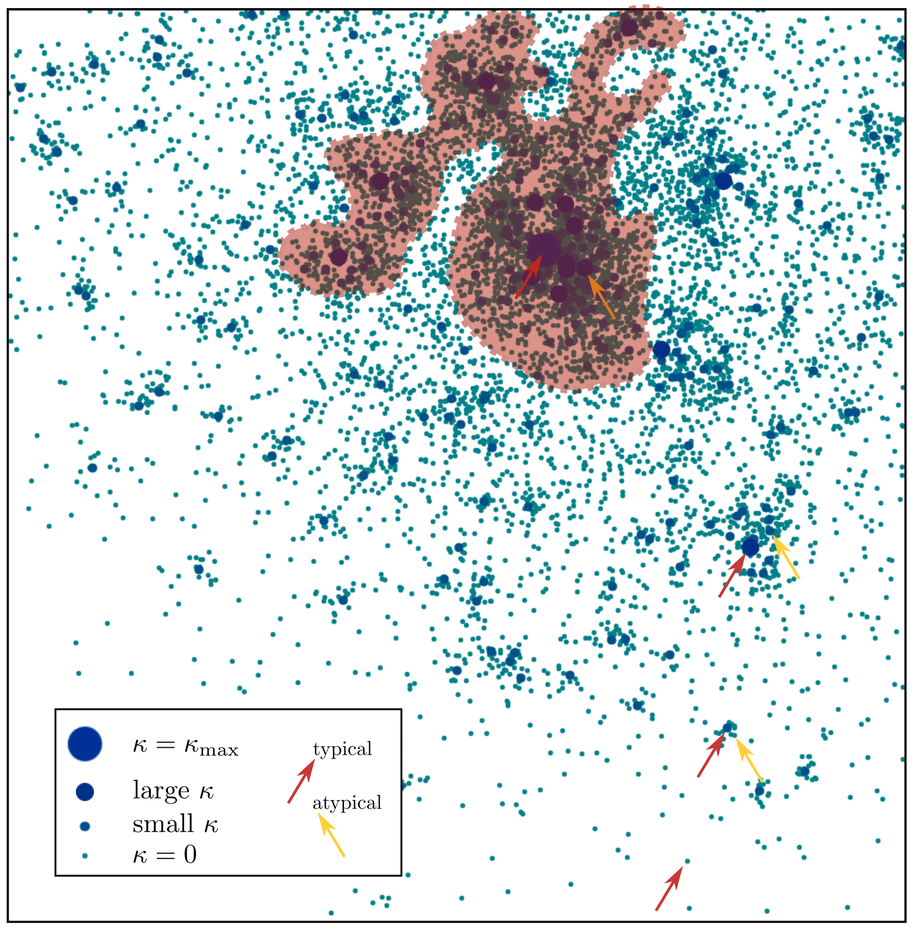
Binary perceptron: efficient algorithms can find solutions in a rare well-connected cluster

Emmanuel Abbe * Shuangping Li † Allan Sly ‡

arXiv:2111.03084

A wide flat minima arises by the coalescence of (atypical) high margin minima

liquid regions

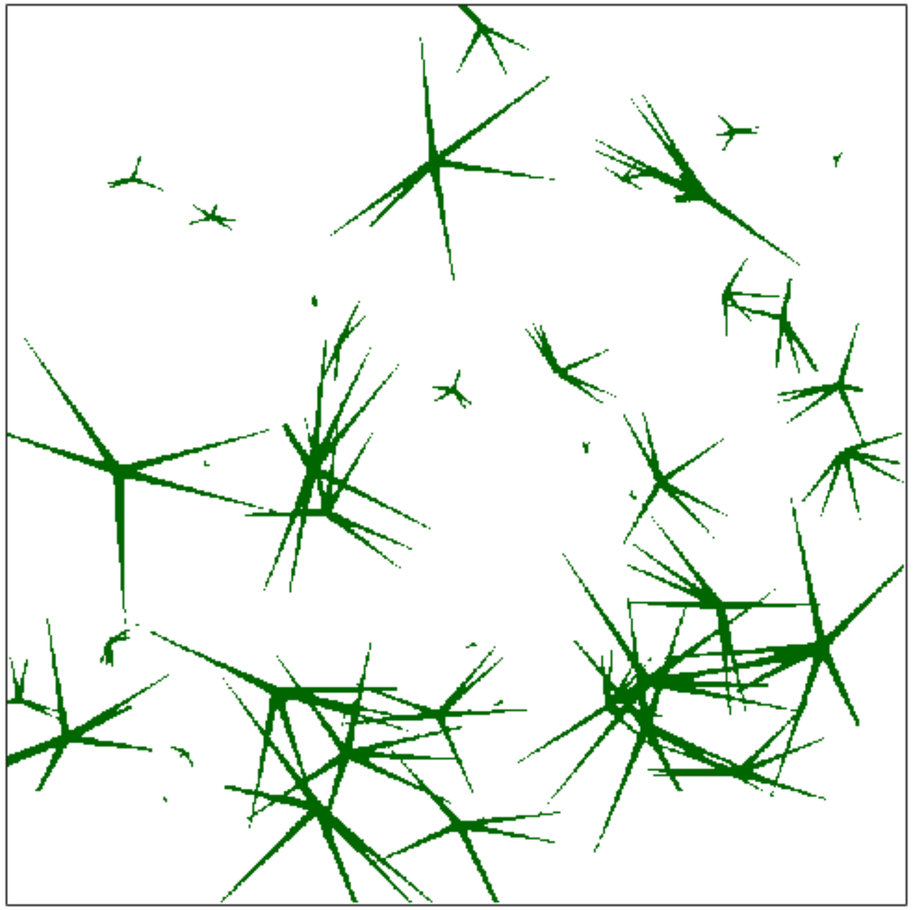


Unveiling the structure of wide flat minima in neural networks

Carlo Baldassi,[1] Clarissa Lauditi,[2] Enrico M. Malatesta,[1] Gabriele Perugini,[1] and Riccardo Zecchina[1]

[1] *Artificial Intelligence Lab, Bocconi University, 20136 Milano, Italy*
[2] *Department of Applied Science and Technology, Politecnico di Torino, 10129 Torino, Italy*

arXiv:2107.01163



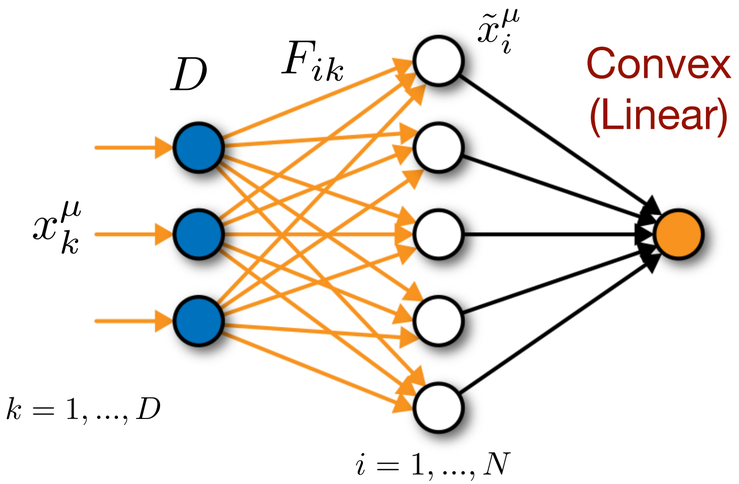
Binary perceptron: efficient algorithms can find solutions in a rare well-connected cluster

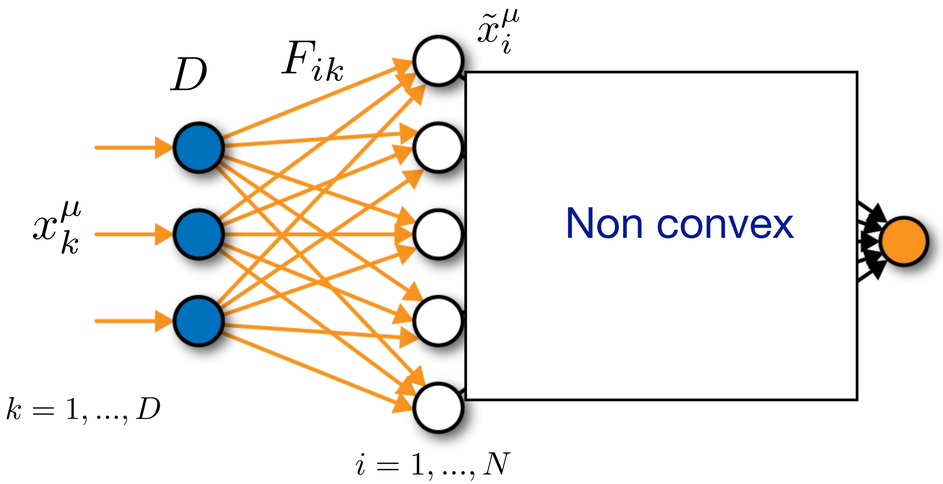
Emmanuel Abbe * Shuangping Li † Allan Sly ‡

arXiv:2111.03084

Overparametrization & non convexity

The non-convex random features model

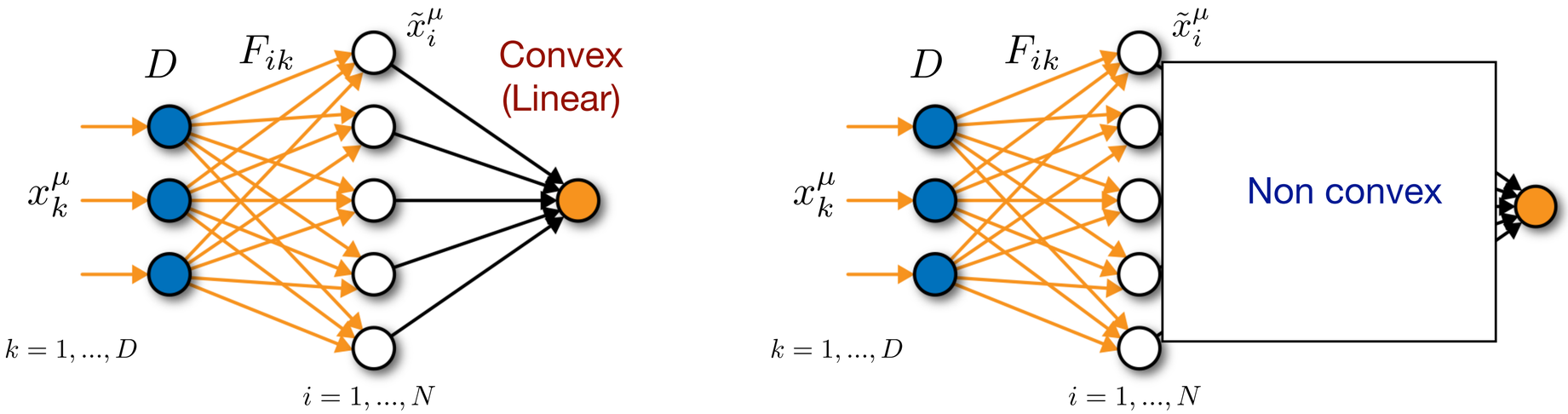




- Gaussian Equivalence Theorem: *in the large N, D, P limit (with $\alpha \equiv P/N$, $\alpha_T \equiv P/D$ fixed)*

$$\tilde{x}_i^\mu = \sigma\left(\frac{1}{\sqrt{D}}\sum_{k=1}^{D} F_{ki}x_k^\mu\right) \to \mu_0 + \frac{\mu_1}{\sqrt{D}}\sum_{k=1}^{D} F_{ki}x_k^\mu + \mu_\star z_i^\mu\,, \quad z_i^\mu \sim \mathcal{N}(0,1)$$

Montanari, Mei (2019); Goldt, Mézard, Krzakala, Zdeborová (2020)

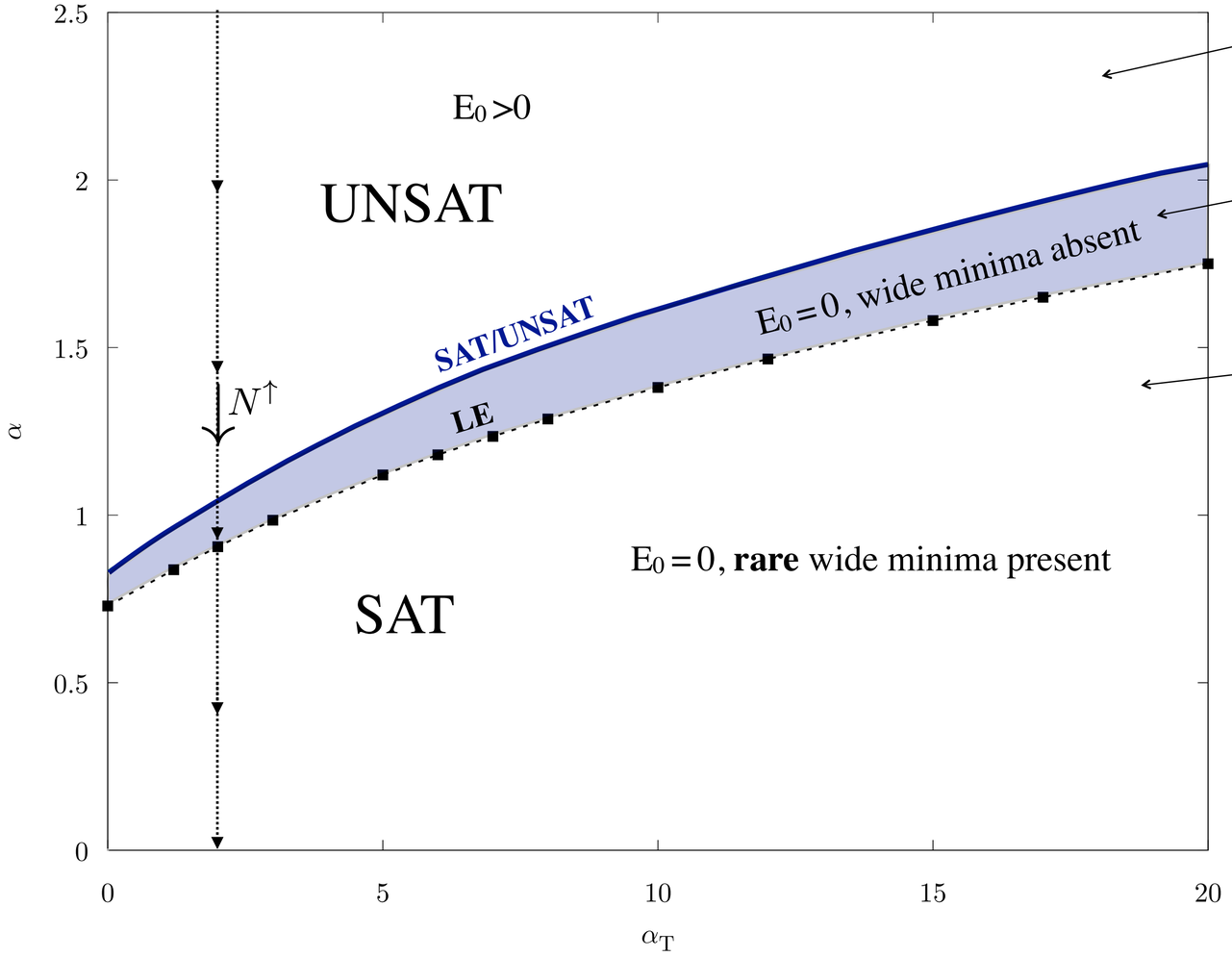


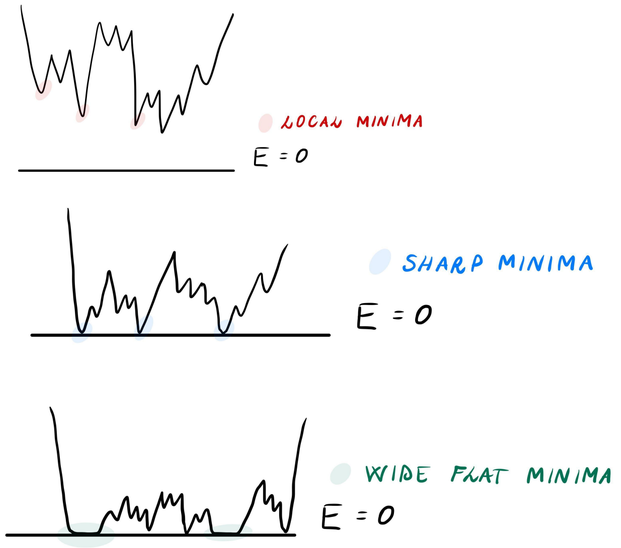
- Statistical Mechanics analysis similar to the underparametrized case: geometrical structure of minimizers, large deviation study of dense regions.

(Baldassi, Malatesta, Lauditi, Pacelli, Perugini, RZ , PRE 2021)

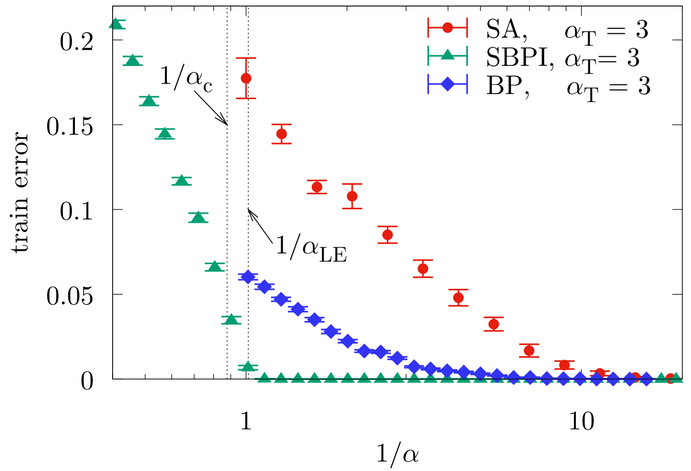
Interpolation Threshold α_c vs Algorithmic Threshold α_{LE} in **discrete** NN

P= # data points $\qquad \alpha \equiv \frac{P}{N}\ ,\ \alpha_T \equiv \frac{P}{D}\ ,\ \alpha_D \equiv \frac{N}{D}\ ,$





Algorithmic experiment

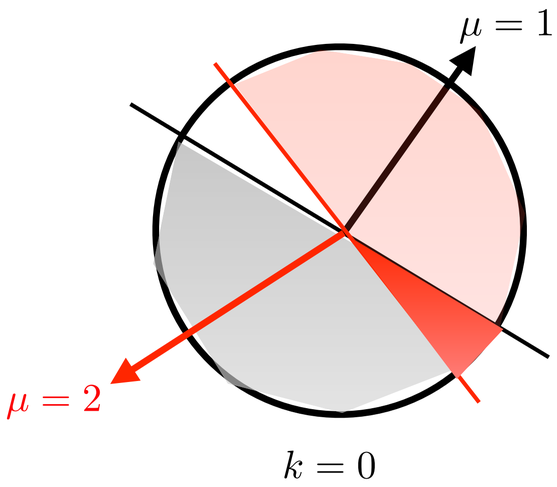


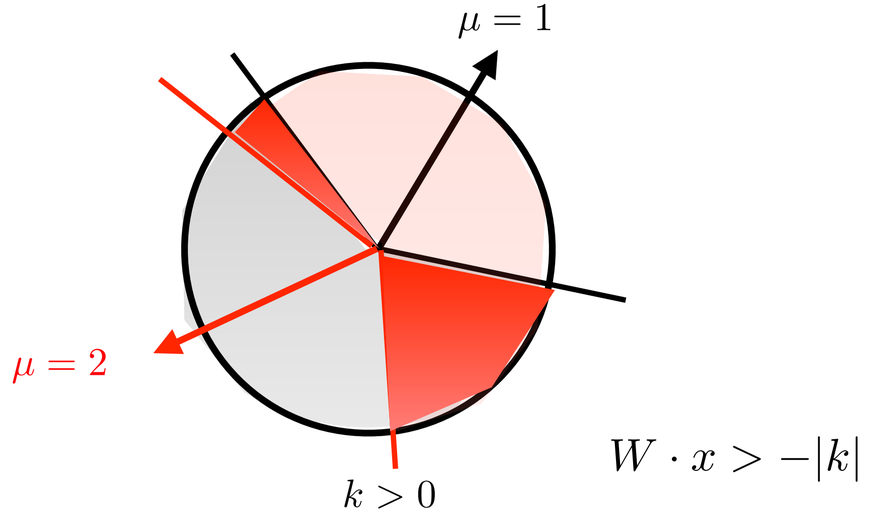
Continuous networks have differences

Let's consider the simplest non convex continuous network: the negative margin spherical perceptron.

$$W_i \in \mathbb{R} \qquad \& \qquad \sum_{i=1}^{N} W_i^2 = N$$

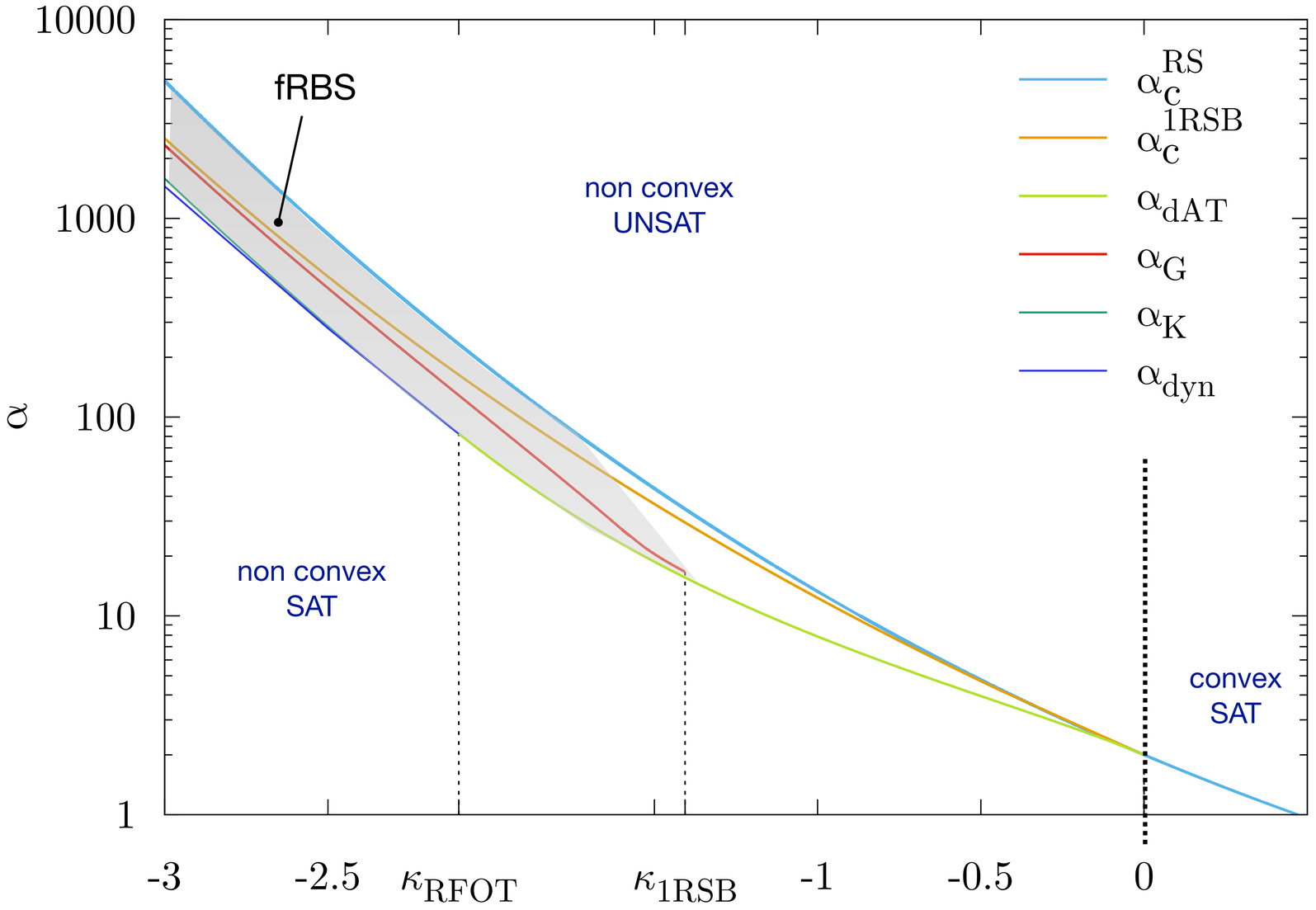
$$1 = \mathrm{Sign}(\Delta^\mu - k) \;\; \text{or} \;\; \Delta^\mu > k$$





$$W \cdot x > -|k|$$

T=0 phase diagram

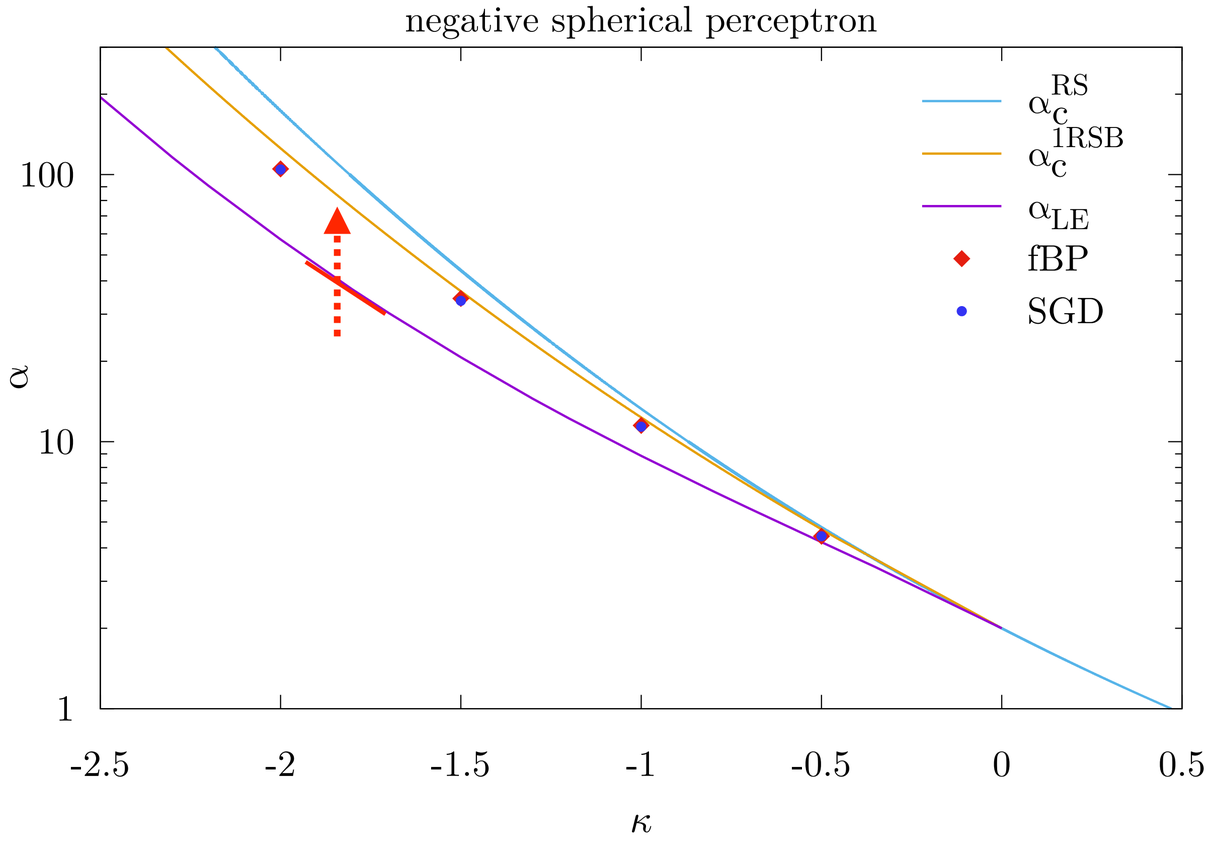


Simplified model for Jamming:

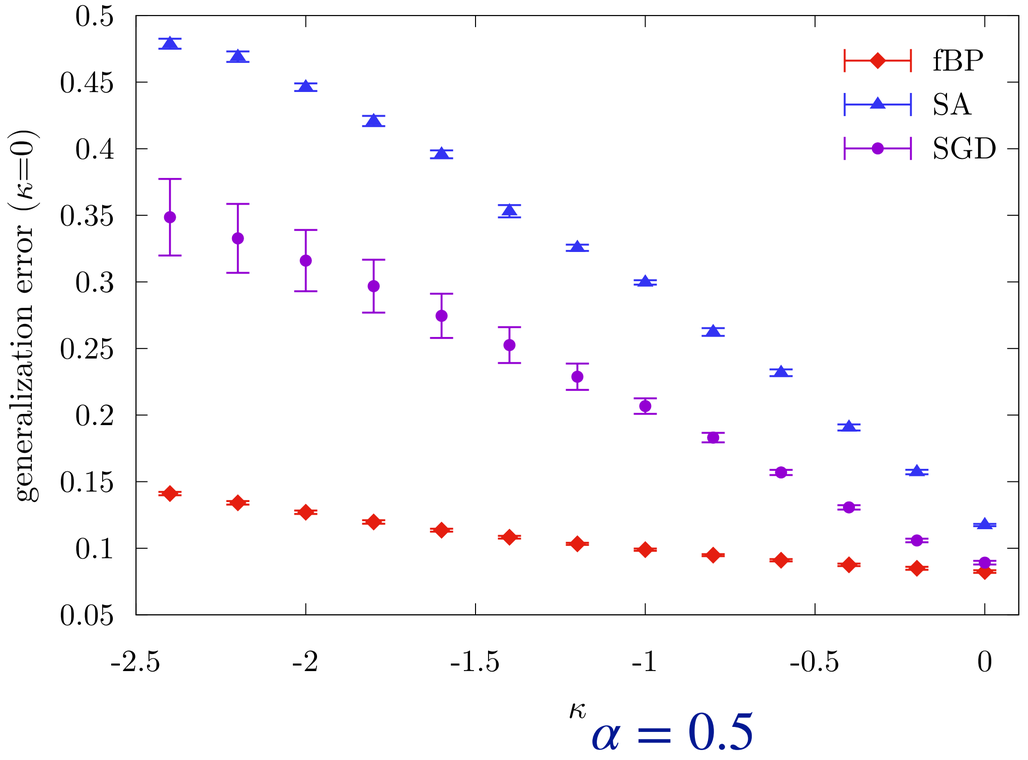
Franz, Parisi, Urbani, Zamponi, …,
and many others

LE transition: Appearance of max entropy minima

capacity



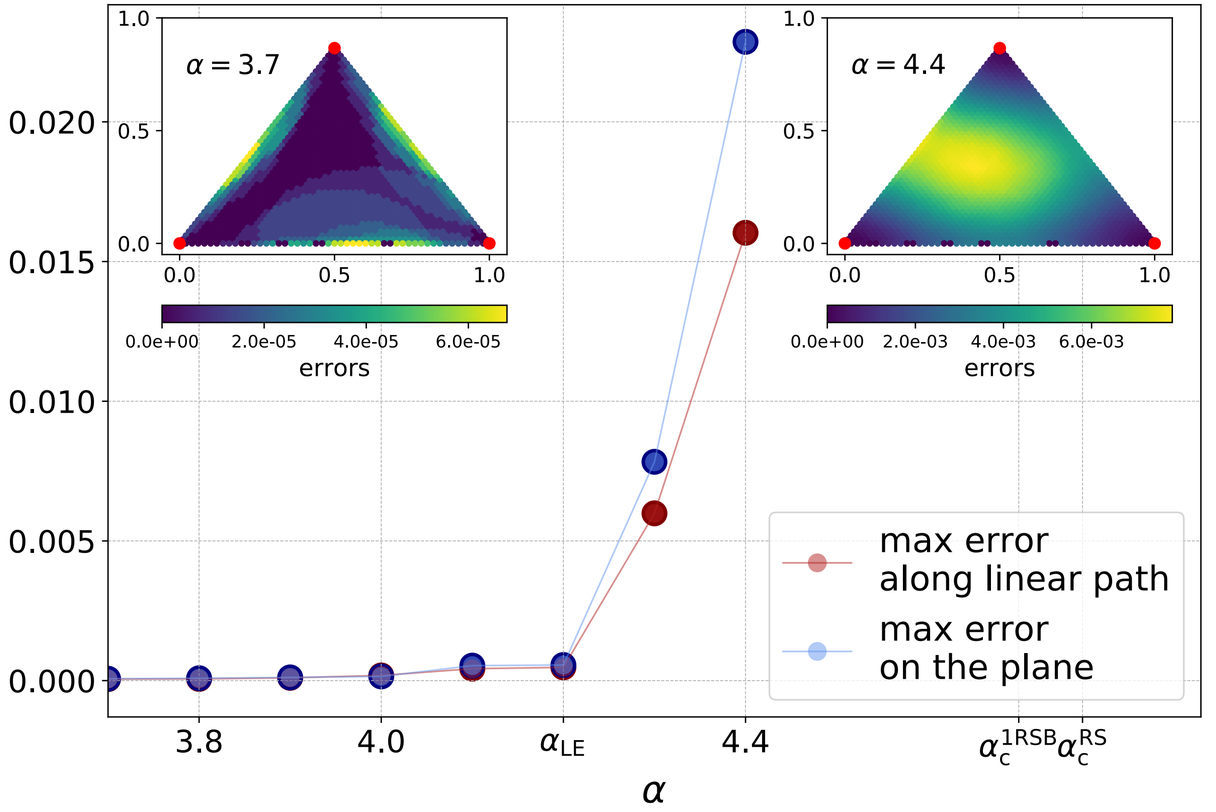
generalization : learning with k<0, test with k=0

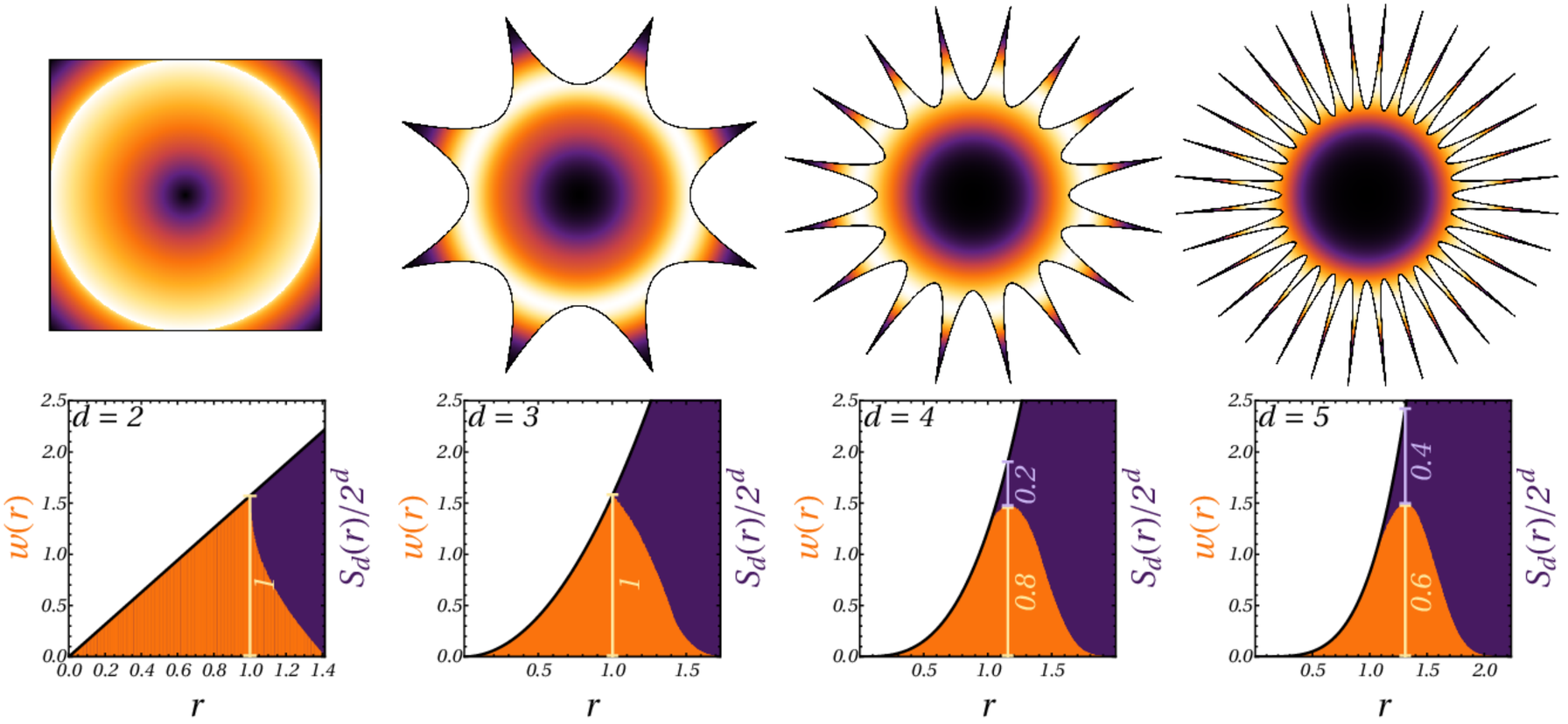


- Not directly related to computational hardness as the model is fRSB.
- Still related to geometry and generalization.

Expected train error of the linear interpolation of *y* replicas

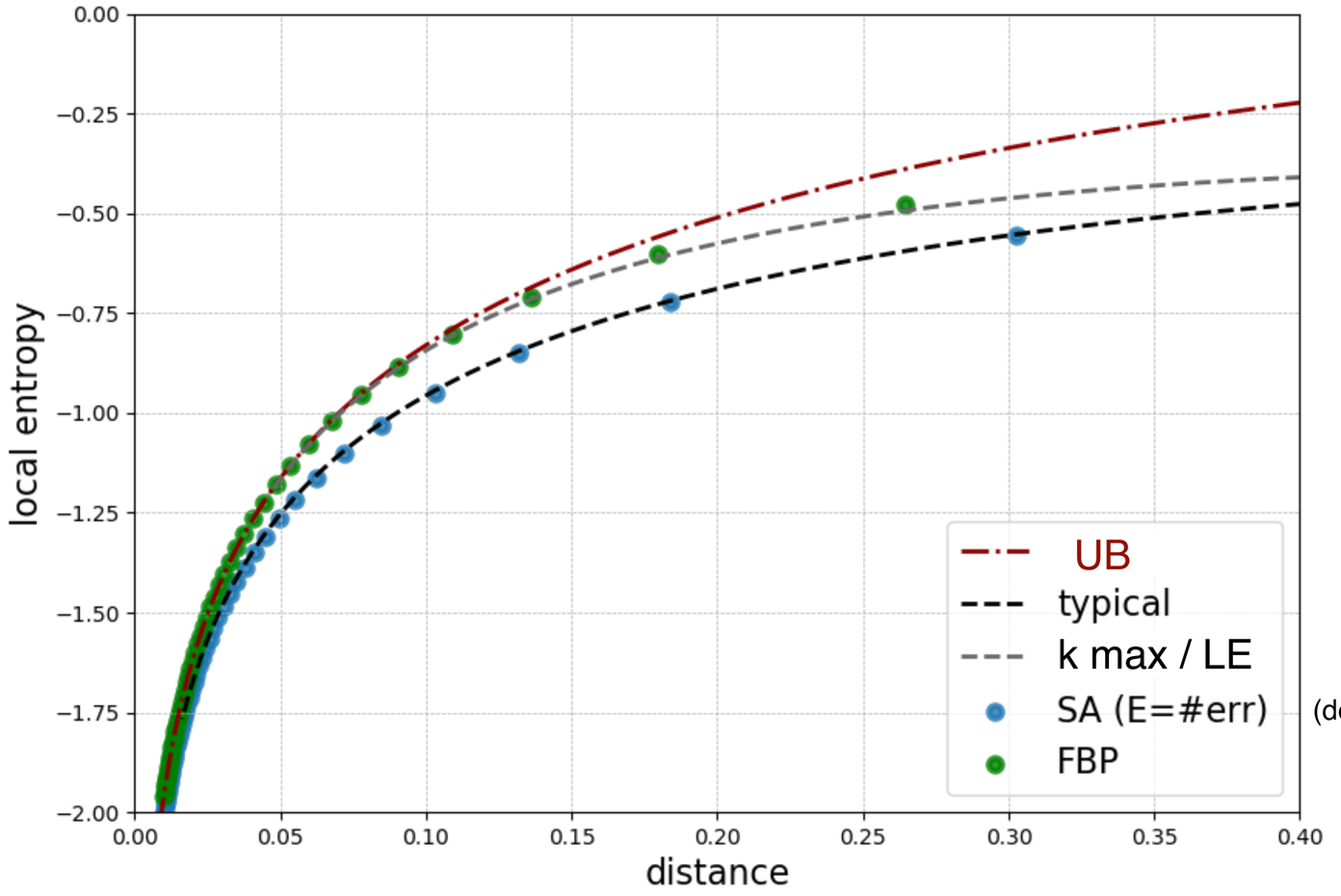
$$\boldsymbol{w}^{\gamma} \equiv \frac{\sum_{a=1}^{y} \gamma_a \boldsymbol{w}_a}{\left\| \sum_{a=1}^{y} \gamma_a \boldsymbol{w}_a \right\|} \qquad \text{with} \qquad \sum_a \gamma_a = 1$$





work by B. Annesi, C. Lauditi, C. Lucibello, E. M. Malatesta, G. Perugini, F. Pittorino, L. Saglietti, 2023

Atypical solutions are surrounded by an exponentially higher number of solutions wrt typical.



α = 2, k=-0.5

Use BP to compute entropy around solutions

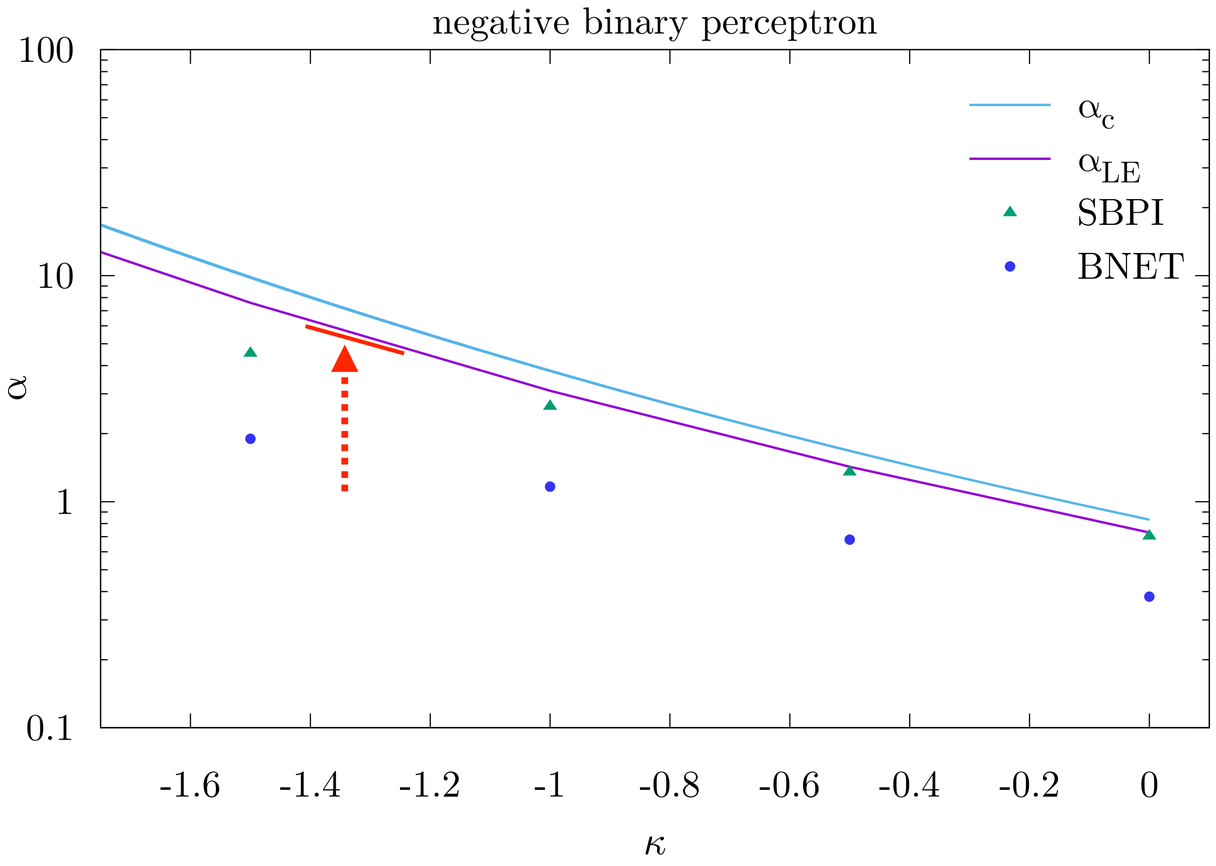
In neural network we expect a generalisation cross over at α_{LE}

Correct regime for overparametrized NN models

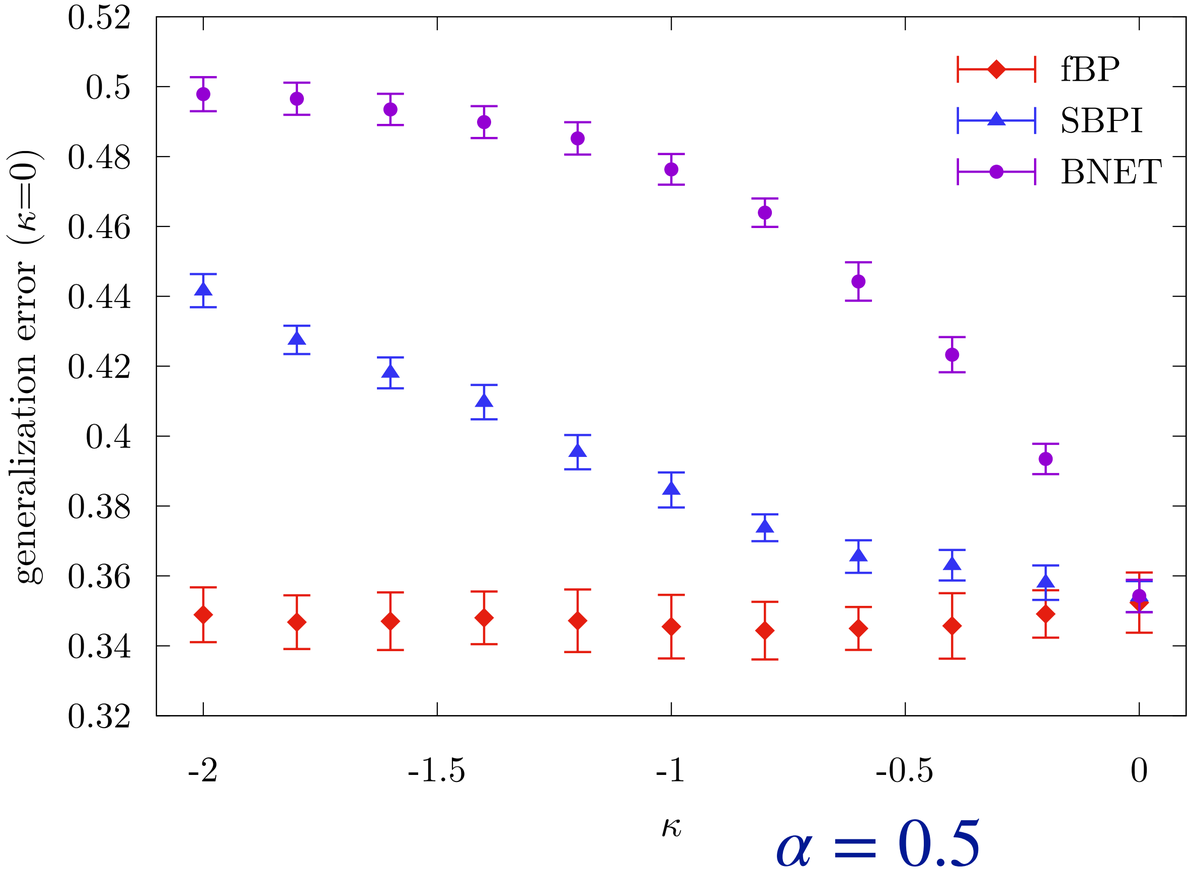
(detailed balance algorithm)

Comparison: LE transition in the binary negative *k* perceptron

capacity



generalization : learning with k<0, test with k=0

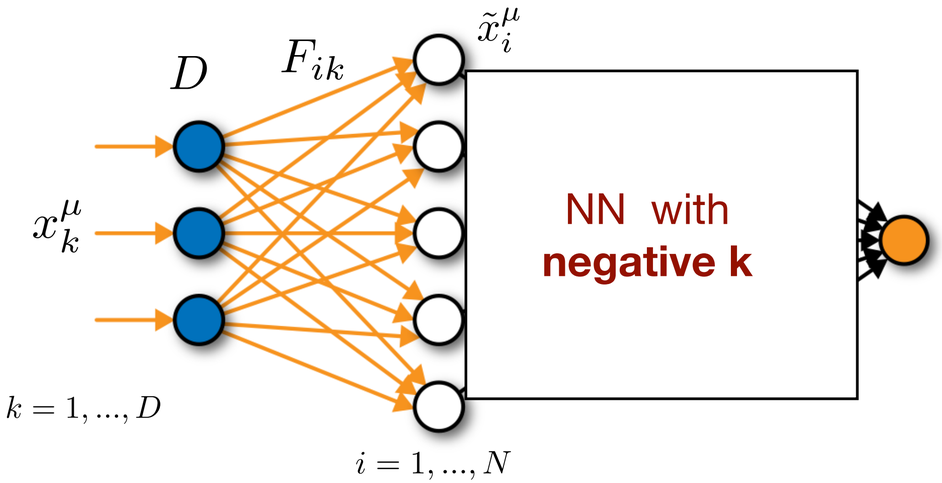


$\alpha = 0.5$

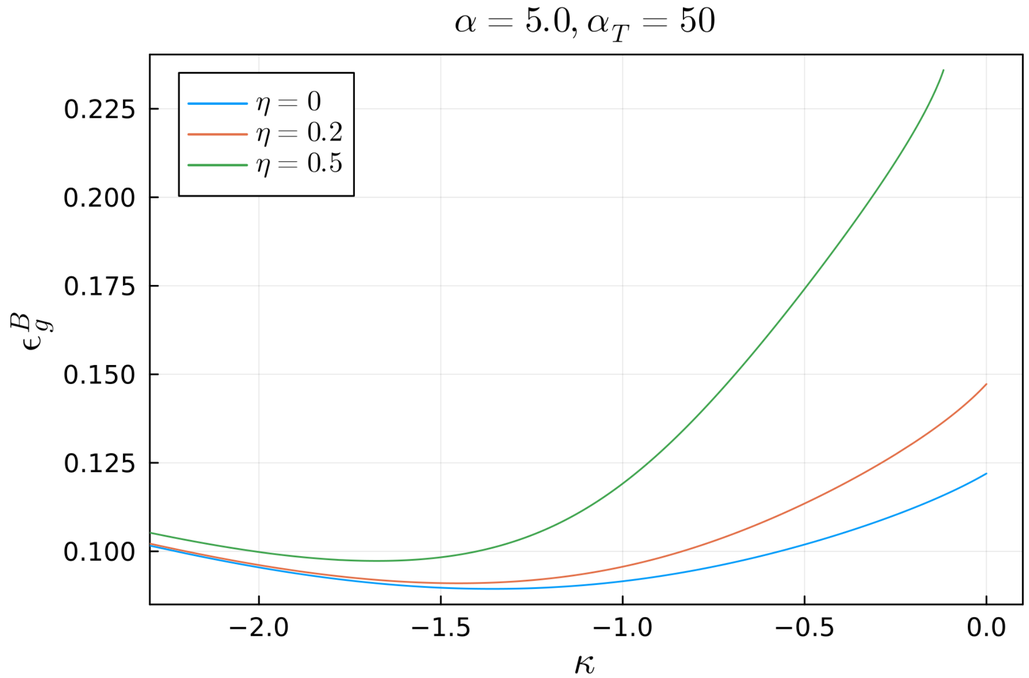
Large networks generalise optimally at negative stabilities in liquid flat regions

- In continuous models, overparameterization and negative k both increase (exponentially) the volume of minimisers (aka solutions)
- We need to find the minimisers which have good generalisations
- Barycentre of liquid regions (locally Bayesian)
- Specific large margin solutions are hard to find at zero training error

Large networks generalise optimally at negative stabilities if in high entropy solutions

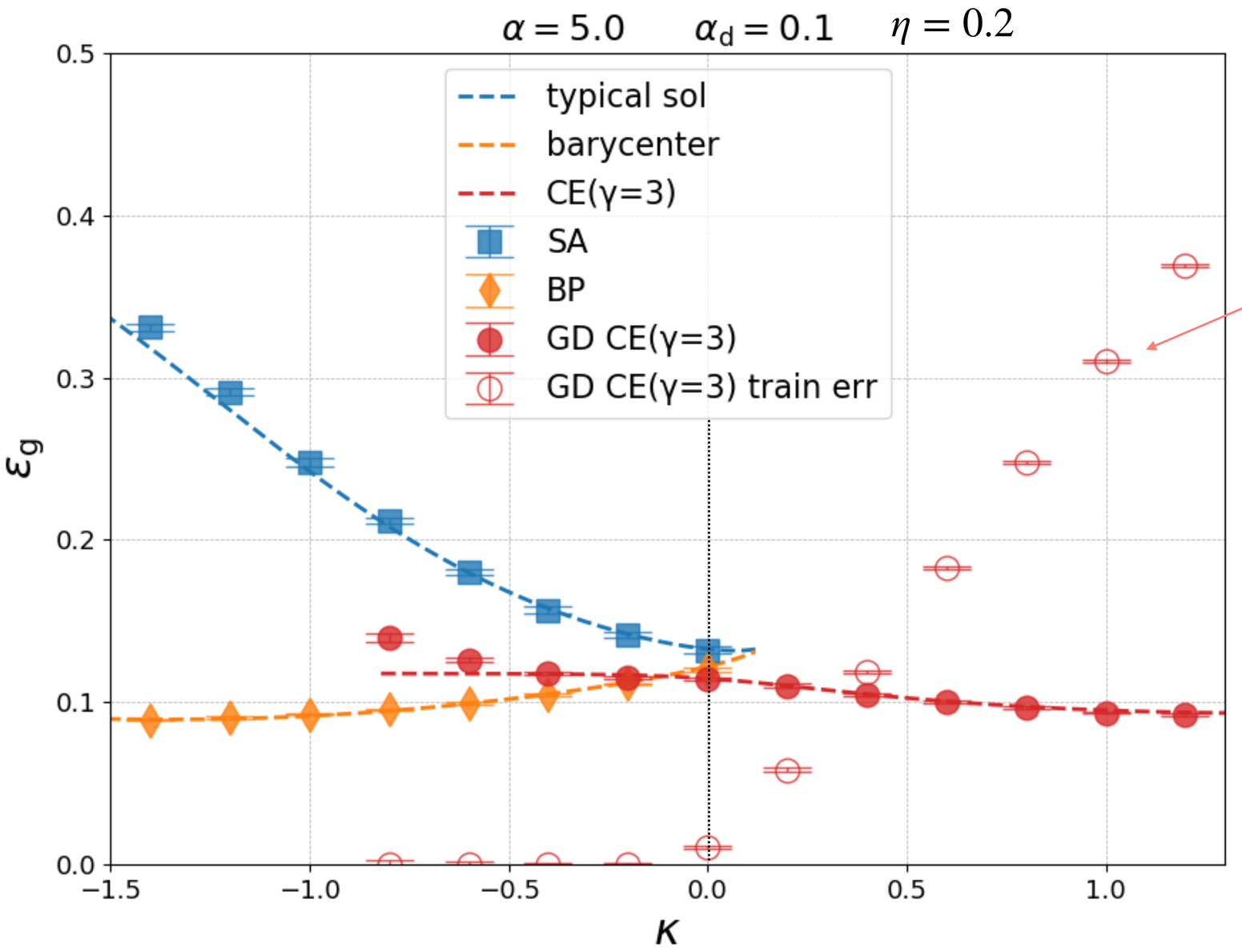


Overparametrized continuous models:

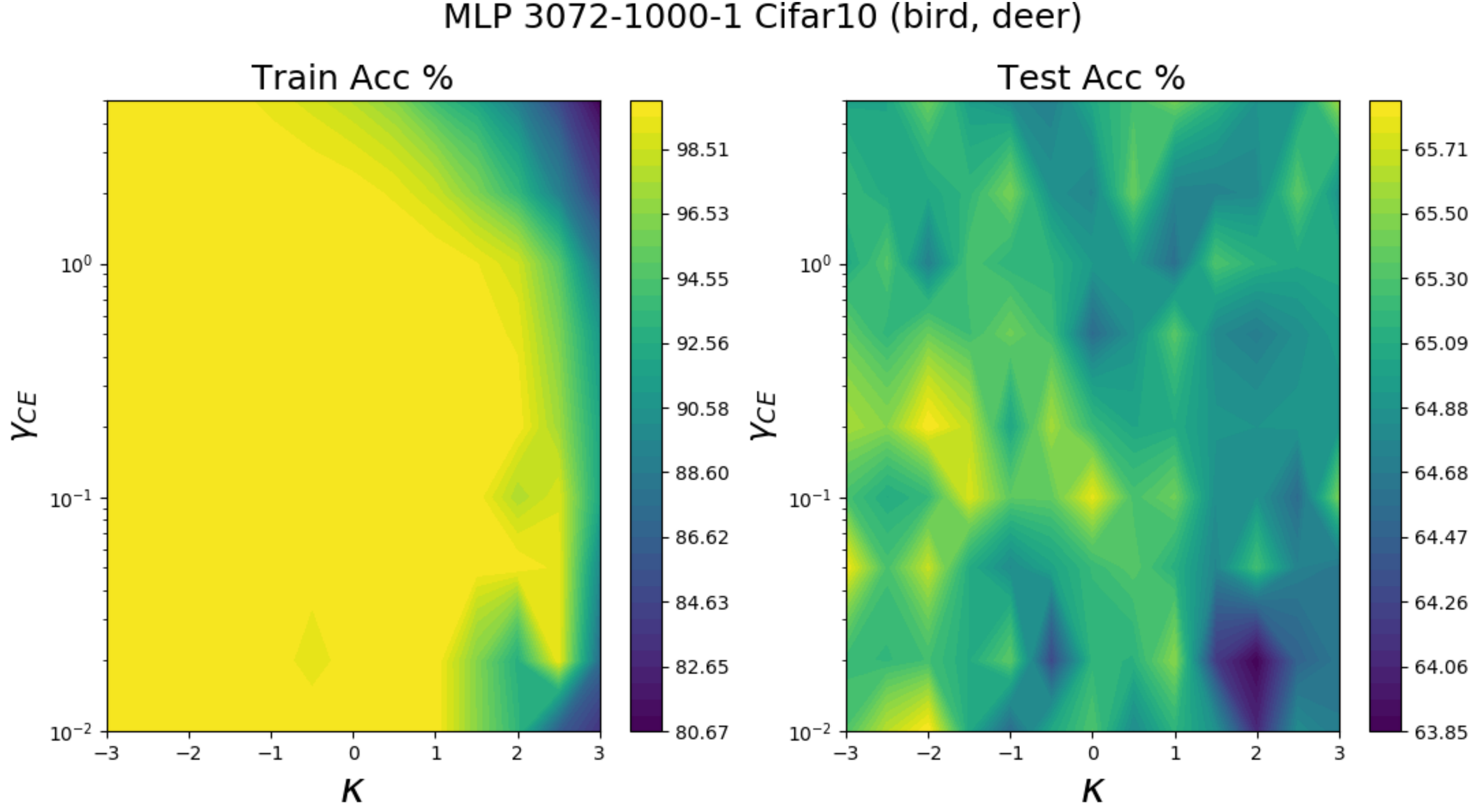


Teacher: input-noisy perceptron

Baldassi, Lauditi, Malatesta, Perugini, Saglietti, Zecchina, in preparation (2023)



ϵ_{train} , GD CE ($\gamma = 3$)


MLP 3072-1000-1 Cifar10 (bird, deer)
Train Acc %
Test Acc %
γ_{CE}
K
10^{0}
10^{-1}
10^{-2}
−3
−2
−1
0
1
2
3
98.51
96.53
94.55
92.56
90.58
88.60
86.62
84.63
82.65
80.67
65.71
65.50
65.30
65.09
64.88
64.68
64.47
64.26
64.06
63.85

On going wok:

- SGD with negative k in the hidden layers
- Smaller architectures with good generalization
- Continual & few-shot learning

Baldassi, Mezard, Zecchina, in preparation (2023)

Baldassi, Lauditi, Malatesta, Mezard, Perugini, Saglietti, Zecchina, in preparation (2023)

Liquid fixed points in recurrent asymmetric neural network

Baldassi, Mezard, Zecchina, in preparation (2023)

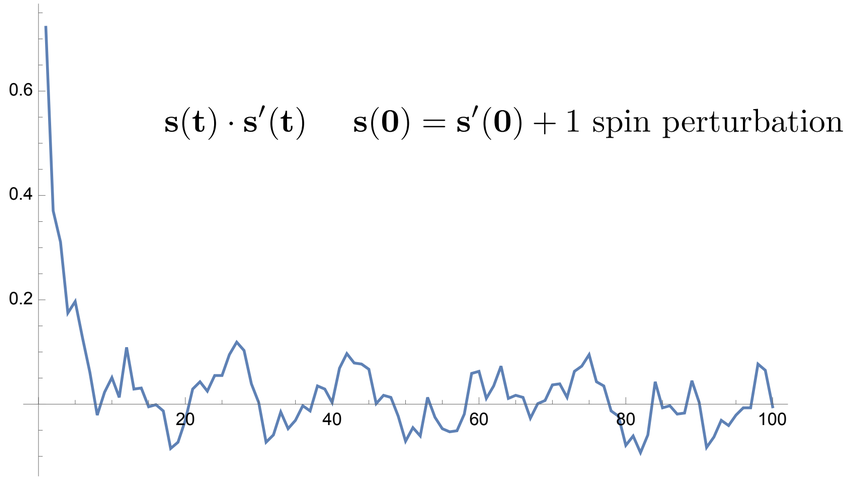
Baldassi, Lauditi, Malatesta, Mezard, Perugini, Saglietti, Zecchina, in preparation (2023)

Some basic results in asymmetric NN

Chaos in Random Neural Networks

H. Sompolinsky, A. Crisanti & H. J. Sommers, PRL 1988

$$\dot{h}_i = -h_i + \sum_{j=1}^{N} J_{ij} S_j = -h_i + \sum_{j=1}^{N} J_{ij} \phi(h_j)$$



Dynamics of Random Neural Networks with Bistable Units

M. Stern, H. Sompolinsky and L. Abbott, PRE 2014

$$\frac{dx_i}{dt} = -x_i + s\tanh(x_i) + g\sum_{j(\neq i)} J_{ij}\tanh(x_j)$$

$$m_i = \tanh\left[sm_i + g\sum_{j(\neq i)} J_{ij} m_j\right]$$

$$m_i = \tanh(x_i)$$

Interactions between the neurons within a cluster are represented in these models by self-coupling *s*

In the limit $g \to \infty$, $s \to \infty$ with fixed $J_D = s/g$ we get back the binary model.

Can an asymmetric random attractor neural network display exponentially many stable attractors (internal representations)?

We expect liquid fixed points to exist

Perceptron storing αN patterns

$$\sigma^\mu = \mathrm{Sign}(\sum_{i=1}^{N} \xi_i^\mu W_i) \quad \forall \mu \quad (\mu = 1, ..., \alpha N)$$

$$W_i = \pm 1$$

$$P(\xi_i^\mu, \xi_j^\nu) = P(\xi_i^\mu)\, P(\xi_j^\nu)$$

$$P(\xi_i^\mu = \pm 1) = \frac{1}{2} \quad P(\sigma^\mu = \pm 1) = \frac{1}{2}$$

Asymmetric recurrent network of neurons

$$S_i = \mathrm{Sign}(\sum_{j=1}^{N} J_{ij} S_j) \quad \forall i \quad (i = 1, ..., N)$$

$$S_i = \pm 1$$

$$P(J_{ij}, J_{i'j'}) = P(J_{ij})\, P(J_{i'j'})$$

$$P(J_{ij} = \pm 1) = \frac{1}{2}$$

In the recurrent NN we have more constraints. To find similar phase space we need to relax them.

One possibility (among others): add a diagonal feedback term $J_{ii} = J_D > 0$

Random binary couplings

$$P(J_{ij}, J_{ji}) = P(J_{ij})P(J_{ji}) \qquad P(J_{ij} = \pm 1) = \frac{1}{2}$$

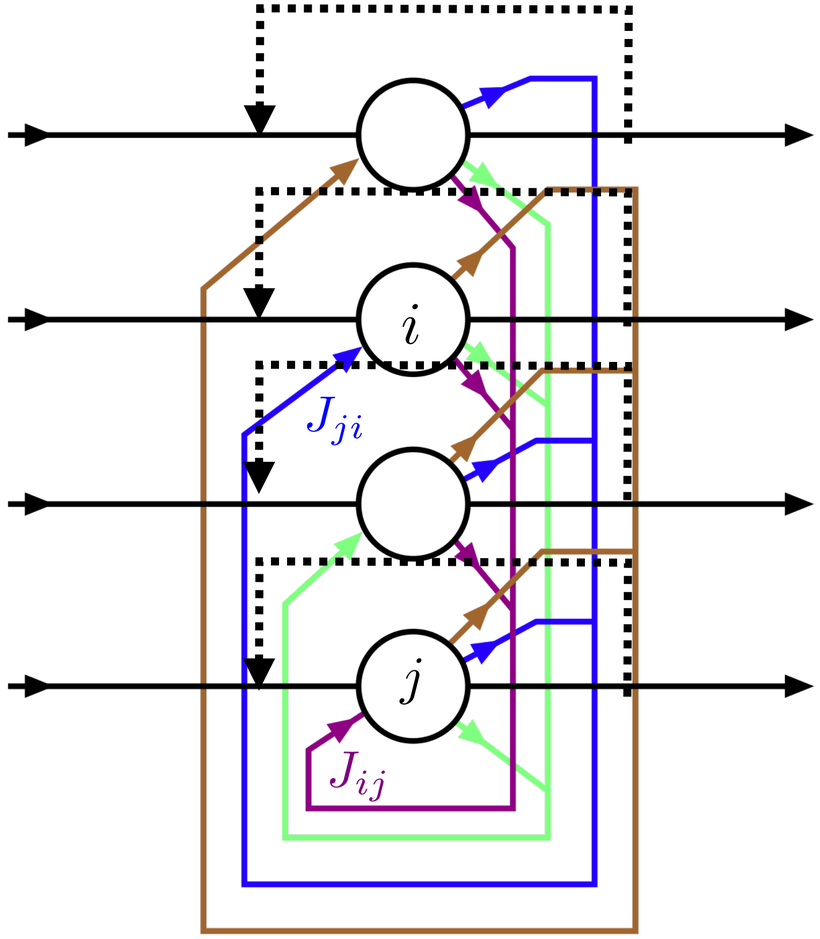
Model: fixed point conditions of the **update dynamics**

$$s_i^{t+1} = \text{Sign}\left(J_D s_i^t + \sum_{i \neq j} J_{ij} s_j^t \right)$$

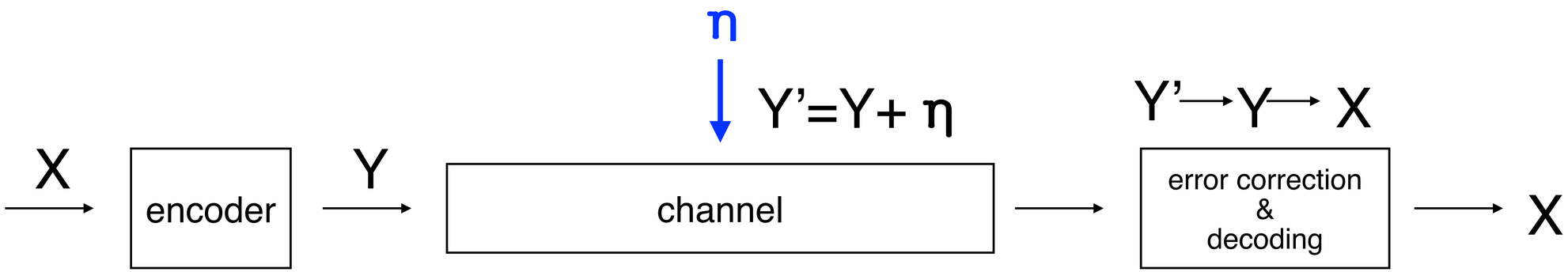
Factors: count violated fixed point constraints

$$E_J(s) = \sum_{i=1}^{N} \left[1 - s_i \text{Sign}\left(J_D s_i + \sum_{j(\neq i)} J_{ij} s_j \right) \right]$$

$J_D \geq 0$ local feedback

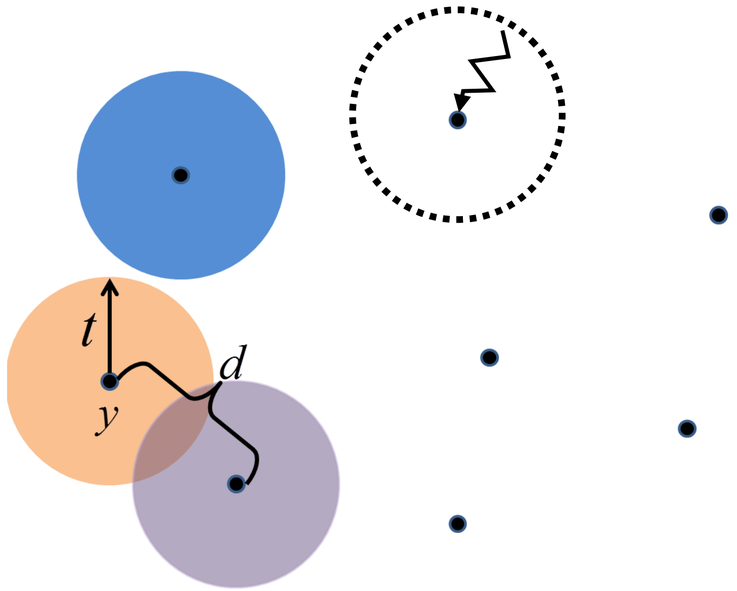


Analogies with error correcting codes



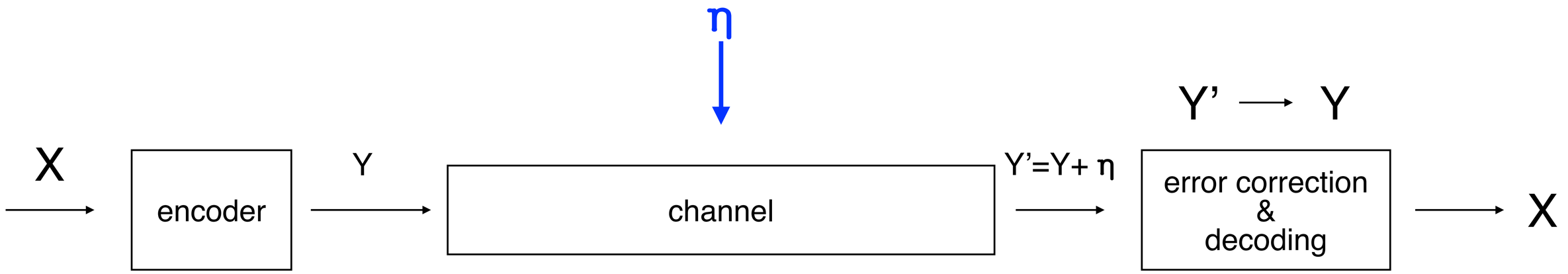
Exponential number of correlated attractive configurations (w.r.t. the decoding dynamics)

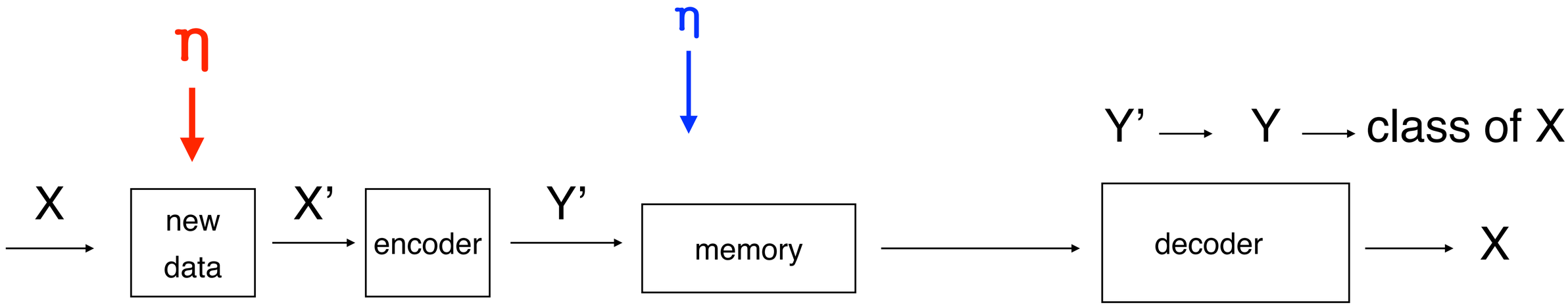
They belong to some subspace, e.g. linear subspace in the case of linear codes (LDPC)



related work by A. Karbasi, A. H. Salavati, A. Shokrollahi, and L. R. Varshney, …

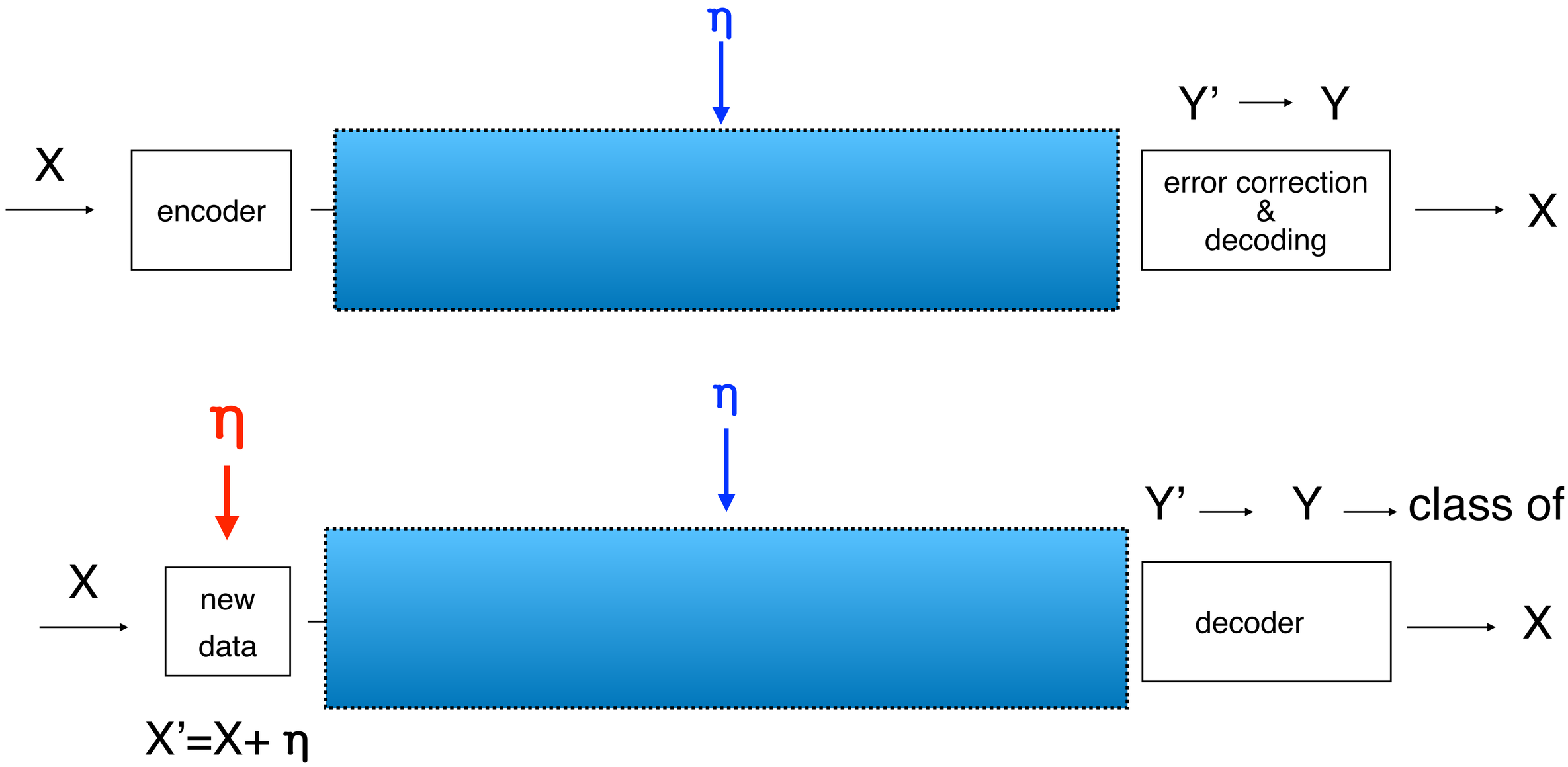
In NN noise will also appear in the input before the encoding





X'=X+ η

In NN noise will also appear in the input before the encoding



Can asymmetric NN perform like sub-optimal ECC?

Can any input mapped to an internal representation (codeword) ?

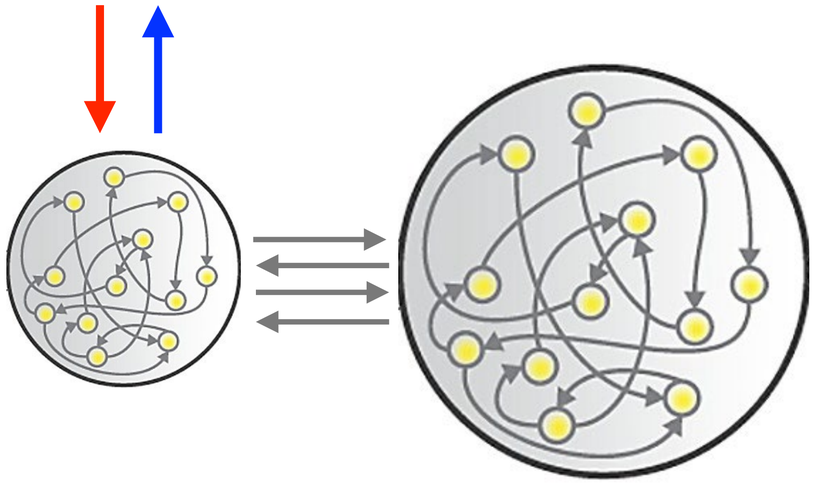
Can they correct an extensive number of errors?

Can asymmetric NN perform like sub-optimal ECC? Yes

Can any input mapped to an internal representation (codeword) ? Yes

Can they correct an extensive number of errors? Yes

Non-linear liquid code



$$\mathbf{s}_A \equiv X$$

$$\mathbf{s}_A \cup \mathbf{s}_B \equiv Y$$

Non-linear encoding:

$$s_i = \text{Sign}\left(J_D s_i + \sum_{j \neq i \in A} J_{ij} s_j + \sum_{j \neq i \in B} J_{ij} s_j \right) \quad \& \quad P(s_i) \propto e^{h x_i} \qquad i \in A$$

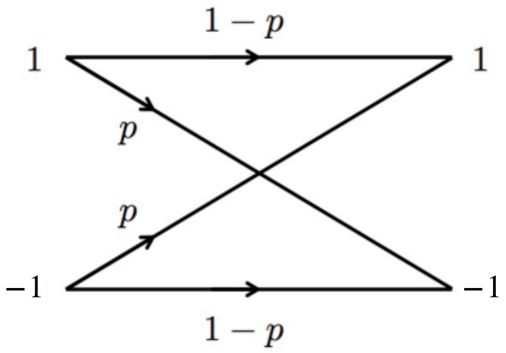
Coding rate:

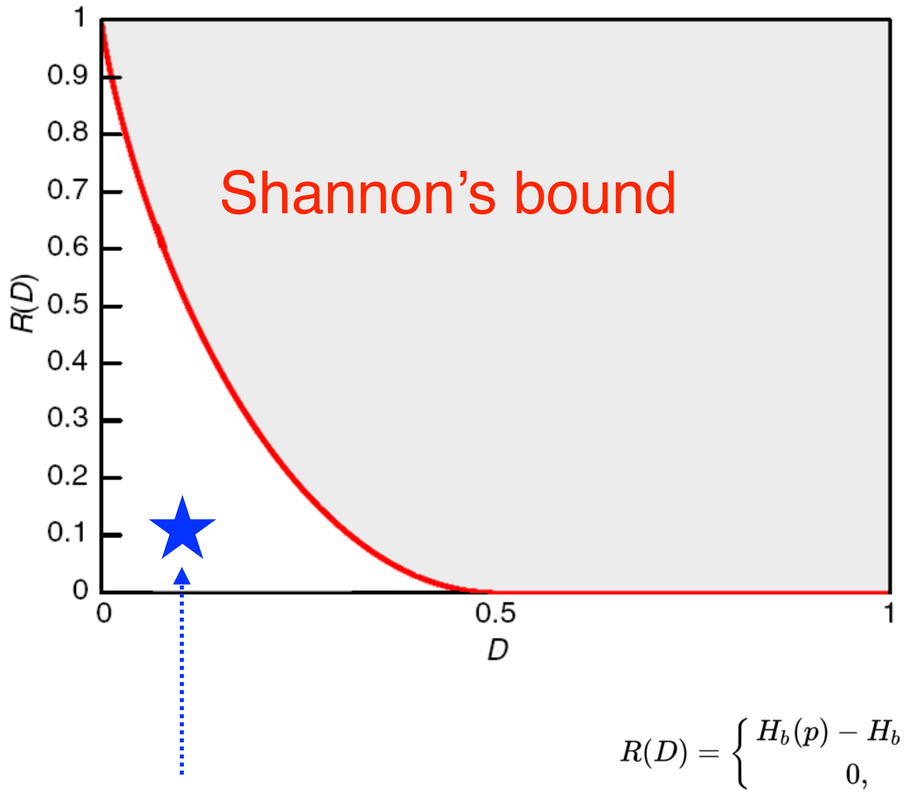
$\mathbf{s}_A$	$\mathbf{s}_B$

$$R \equiv \frac{N_A}{N_A + N_B}$$

Shannon bound and positioning of simple random ANNs

BSC, spin flips





$$R(D) = \begin{cases} H_b(p) - H_b(D), & 0 \le D \le \min(p, 1-p) \\ 0, & D > \min(p, 1-p) \end{cases}$$

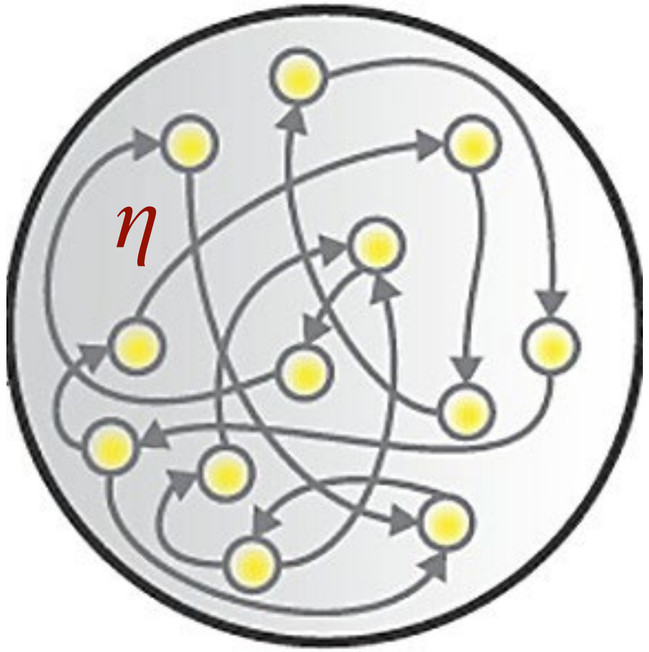
numerical result

(could be done analytically)

In NN we just need addressable and stable internal representations

- Given an external pattern ξ (field)
- Can the network converges to a stable internal representation η ?

ξ ⟶



converge

$$s_i^{t+1} = \mathrm{Sign}(J_D s_i^t + \sum_{j \neq i} J_{ij} s_j^t) \quad , \quad P(s_i) \propto e^{h_i \xi_i} \quad \longrightarrow \quad \eta_i = \mathrm{Sign}(J_D \eta_i^t + \sum_{j \neq i} J_{ij} \eta_j) \quad , \quad \eta = \eta(\xi)$$

$$\{\xi^\mu\} \rightarrow \{\eta^\mu\}$$

Number of configurations which satisfy the fixed point condition, $\beta \to \infty$

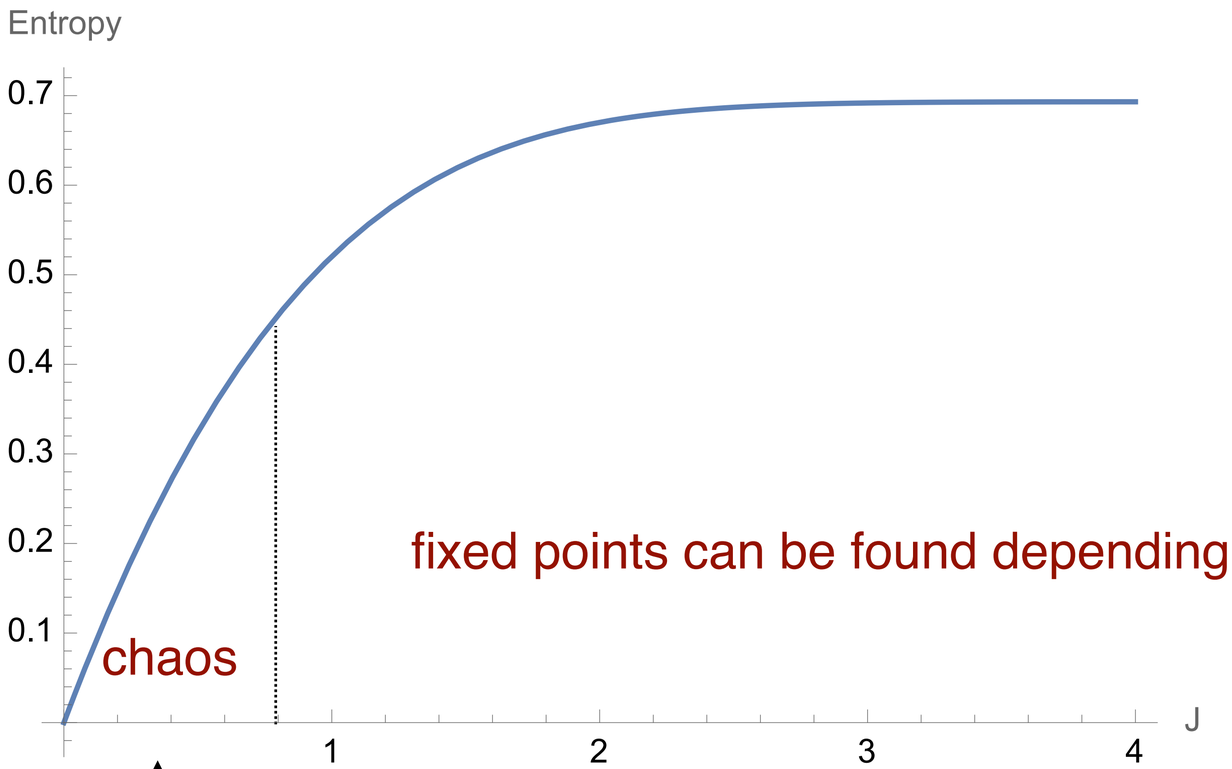
$$Z_J = \sum_s e^{-\beta E_J(s)} = e^{-\beta N} \sum_s e^{\beta \sum_i \mathrm{Sign}\left(J_D + s_i \sum_{j(\neq i)} J_{ij} s_j\right)}$$

First moment gives the **exact** typical number of fixed points:

$$\Phi_1 = -\frac{1}{\beta N} \log \mathbb{E}_J Z_J = 1 - \frac{1}{\beta}\left[\log 2 + \log\left(e^{\beta} H(-J_D) + e^{-\beta} H(J_D)\right)\right]$$

where $$H(x) = \int_x^\infty \frac{dz}{\sqrt{2\pi}} e^{-z^2/2} = \frac{1}{2}\mathrm{Erfc}\left(\frac{x}{\sqrt{2}}\right)$$

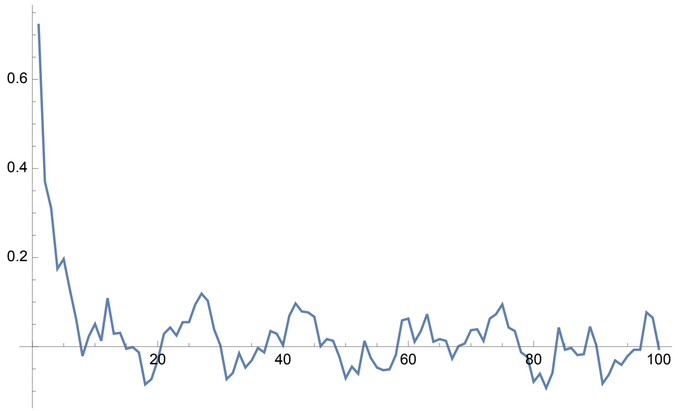
$$S_1 = \frac{1}{N} \log \bar{Z}_J|_{\beta\to\infty} = \log 2 + \log\left[1 - H(J_D)\right]$$



for the simple dynamics

$$s_i^{t+1} = \text{Sign}\left(J_D s_i^t + \sum_{i\neq j} J_{ij} s_j^t\right)$$

$\mathbf{s(t)} \cdot \mathbf{s'(t)}$ $\quad \mathbf{s(0)} = \mathbf{s'(0)} + 1$ spin perturbation



RS, 1-RSB and Local Entropy results

- The RS and 1-RSB solution give back the Annealed (first moment results) for the dominating fixed points (q_0 =0, things are relatively simple)

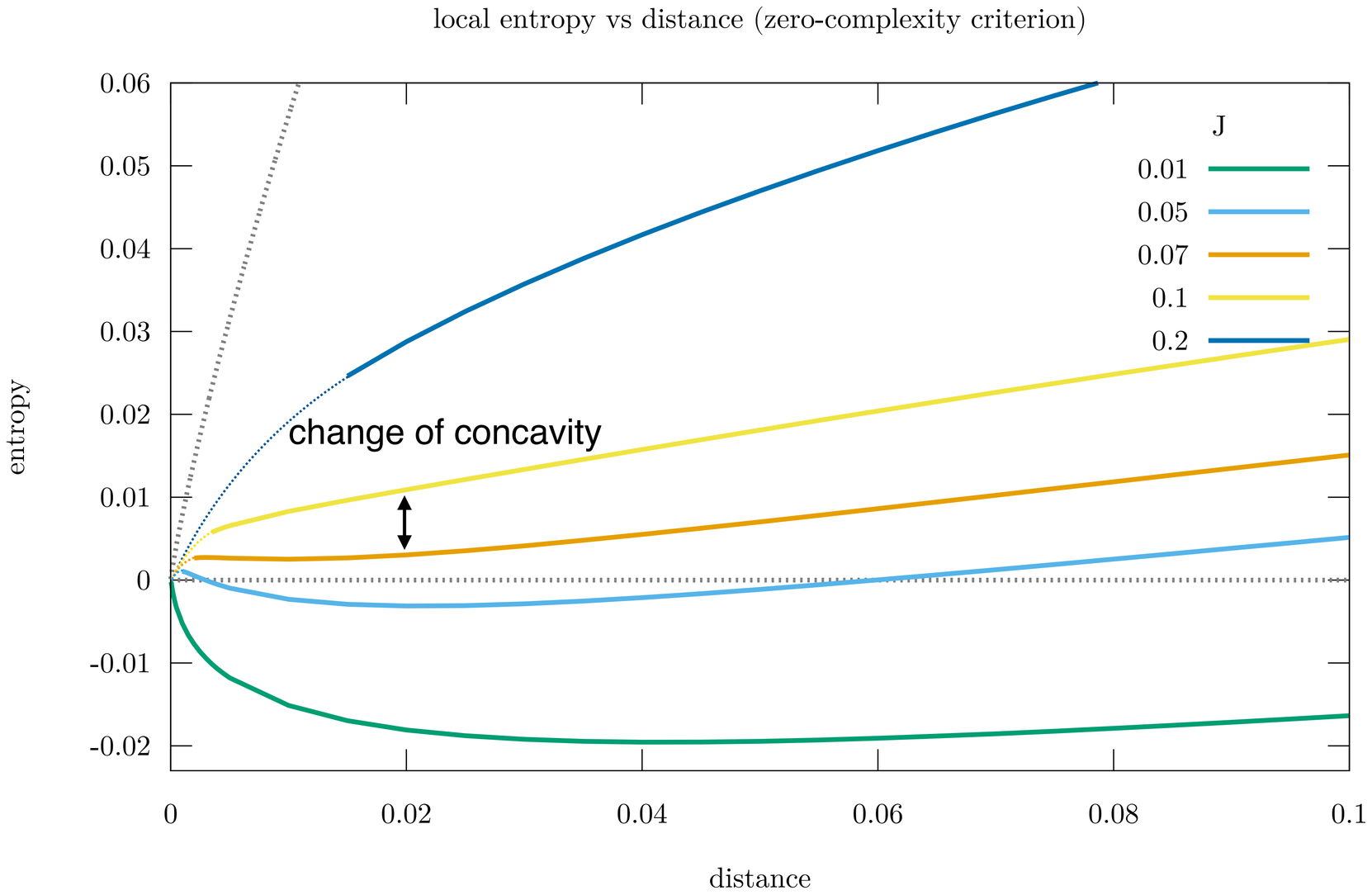
- We can use the 1-RSB formalism to analyse existence of liquid fixed points

$$\Phi = -\beta - \frac{1}{2}\left(r_1\left(1-q_1\right)\right) - \frac{m}{2} q_1 r_1 + \frac{1}{m}\log\left\{\int Dz\, Dt\, \left[e^{z\sqrt{r_1}}\left(e^{\beta}H'_{--} + e^{-\beta}H'_{++}\right) + e^{-z\sqrt{r_1}}\left(e^{\beta}H'_{-+} + e^{-\beta}H'_{+-}\right)\right]^m\right\}$$

$$Q^{ab} = \frac{1}{N}\sum_k s_k^a s_k^b \quad \text{for} \quad a < b$$

$$R^{ab} = \frac{1}{N}\sum_k \hat{h}_k^a \hat{h}_k^b s_k^a s_k^b \quad \text{for} \quad a \leq b$$

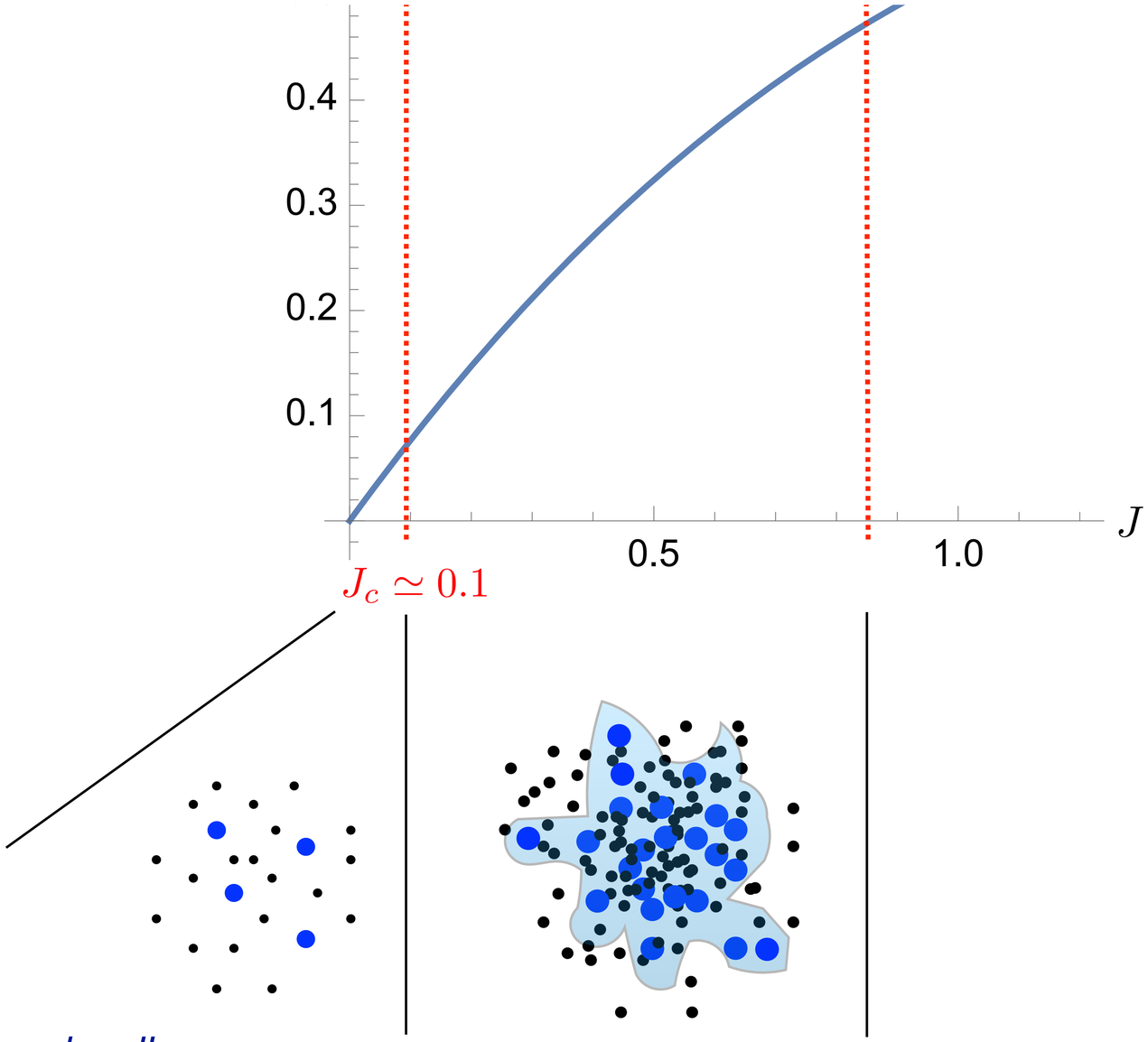
$$H'_{ss'} \equiv H\left(\frac{sJ + s't\sqrt{q_1}}{\sqrt{1-q_1}}\right) \qquad (J_D \equiv J)$$



change of concavity:

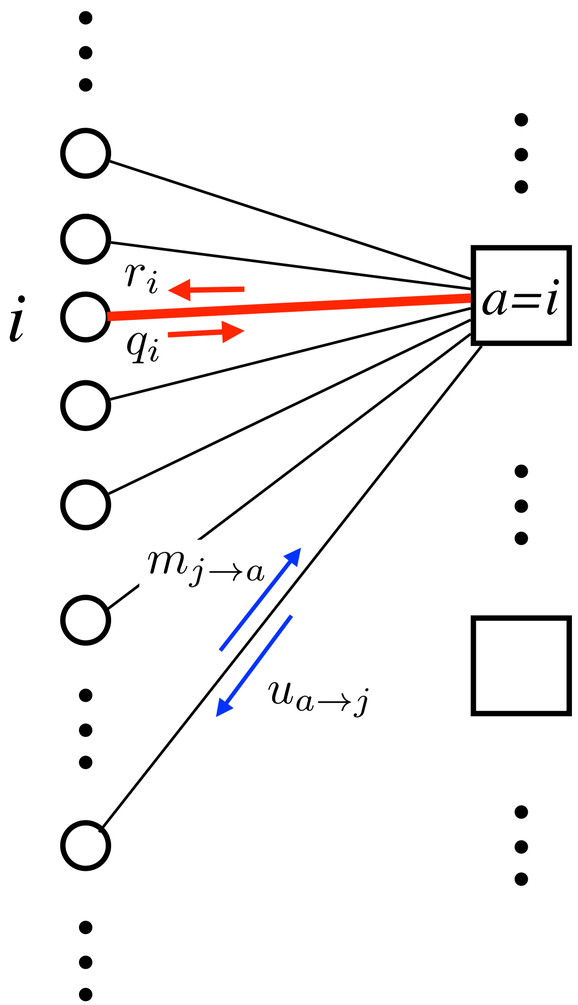
geometrically impossible, the entropy has to increase because of the interchangeability of the centroid and solutions at distance d

$0 \leq J \leq J_c \simeq 0.1$	isolated solutions co-exist with an exponential number of small cluster $d = \frac{1-q}{2} \qquad q \in [0.998, 1]$
$J > J_c$	liquid fixed points: isolated solutions co-exist with an exponential number of extended dense regions which connect the small clusters
$J > J_a(N)$	more fixed points become accessible (strong finite size effects)
$J = \infty$	all configurations are trivially fixed points



Isolated and well separated small clusters, Overlap-Gap-Property (Gamarnik et al.)

Non chaotic local dynamics: fBP dynamics



factor graph:

i, j neural state s_i

a, b; factor (dynamical constraint)

$a, b, i, j \in \{1, ..., N\}$

$u_{a\to j},\ m_{j\to a},\ r_i,\ q_i \quad \in [-1, 1]$

given two magnetizations h, k , define

$$h \otimes k \hat{=} \frac{h+k}{1+hk} = \tanh\left(\operatorname{atanh}(h) + \operatorname{atanh}(k)\right)$$

commutative and associative $\rightarrow$ $\bigotimes_i u_i = \tanh\left(\sum_i \operatorname{atanh}(u_i)\right)$

inverse operation $\rightarrow$ $h \oslash k = h \otimes (-k) = \frac{h-k}{1-hk} = \tanh\left(\operatorname{atanh}(h) - \operatorname{atanh}(k)\right)$

Non-cavity: $\mu_i = \sum_{a\neq i} w_i^a m_{i\to a},\ \sigma_i = \sqrt{\sum_{a\neq i}(1 - m_{i\to a}^2)}$

Cavity: $\mu_{a\to i} = \mu_i - w_i^a m_{i\to a},\ \sigma_{a\to i} = \sqrt{\sigma_i^2 - (1 - m_{i\to a}^2)}$

Simplified fBP equations

1) Approximate with non-cavity fields for each variable, h_i.
2) We update them iteratively; at each iteration we pick a perceptron j and have a special update for h_j (which has the role of the output) vs the others.

$$i \neq j \qquad h_i^{t+1} = h_i^t + \frac{\left(\frac{1+m_j}{2}\right)\left(H^{++} - H^{+-}\right) + \left(\frac{1-m_j}{2}\right)\left(H^{--} - H^{-+}\right)}{\left(\frac{1+m_j}{2}\right)\left(H^{++} + H^{+-}\right) + \left(\frac{1-m_j}{2}\right)\left(H^{--} + H^{-+}\right)}$$

$$H^{s_1 s_2} = H\left(-\frac{s_1\left(\mu - w_j^i m_i\right) + J\sqrt{N} + s_2 w_j^i}{\sigma}\right)$$

$$i = j \qquad h_j^{t+1} = h_j^t + \operatorname{atanh}\left(\frac{r^+ - r^-}{r^+ + r^-}\right) \qquad r^{\pm} = H\left(-\frac{\pm\mu + J}{\sigma}\right)$$

$$\mu = \textstyle\sum_{i \neq j} w_j^i m_i,\ \sigma = \sqrt{\sum_{i \neq j}\left(1 - m_i^2\right)}$$

$$m_i = \operatorname{hardtanh}\left(h_i\right)$$

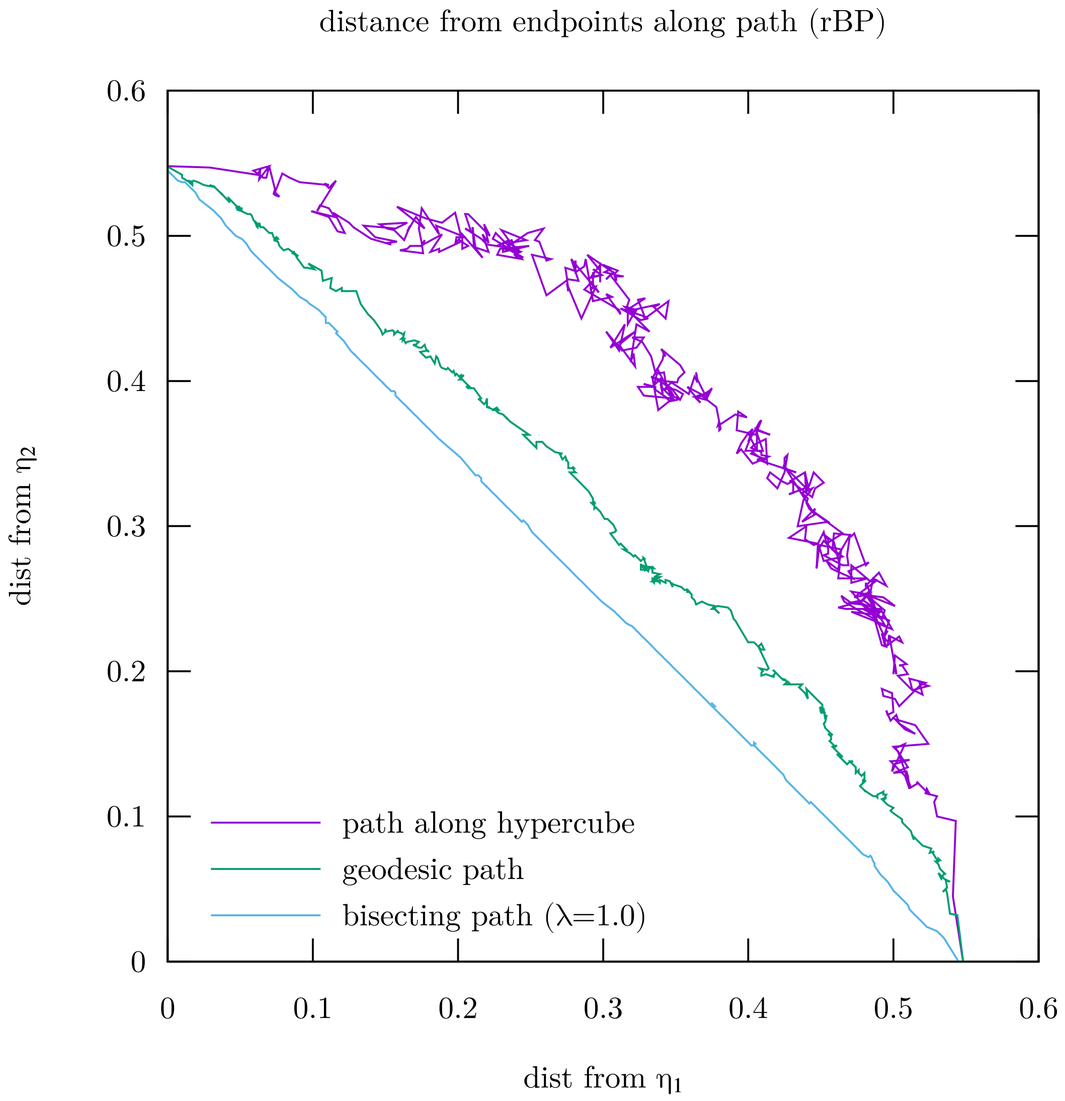
Other dynamical schemes that are effective and simple:

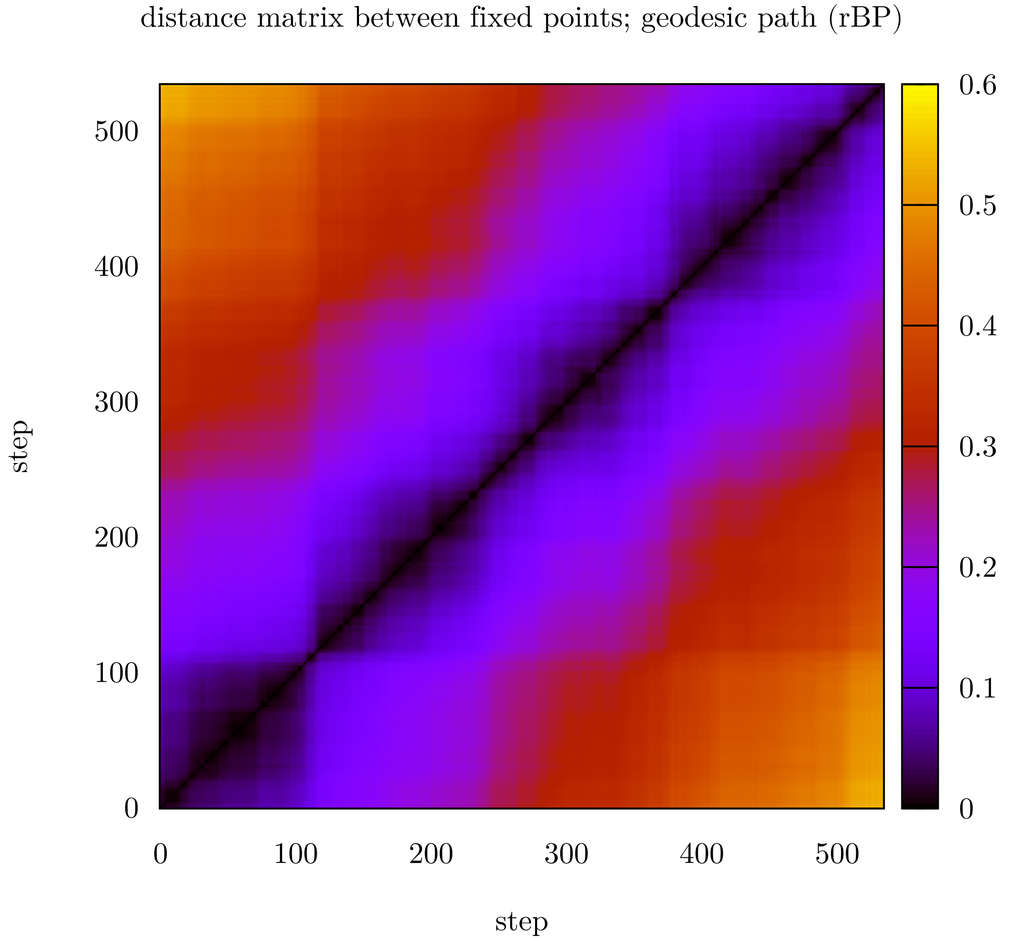
- Gradient Descent on the log-likelihood (1)
- Replicated simple dynamics (2)
- Entropic greedy dynamics ~ entropy-SGD, with m=1 and T=0, (3)
- …

(1) Role of Synaptic Stochasticity in Training Low-Precision Neural Networks, Baldassi, Gerace, Kappen, Lucibello, Saglietti, Tartaglione, Zecchina, Phys. Rev. Lett. **120**, 268103 (2018)

(2) Shaping the learning landscape in neural networks around wide flat minima, Baldassi, Pittorino, ZecchinaPNAS, 117 (2020)

(3) entropy-SGD:biasing gradient descent into wide valleys, Chaudhari, Choromanska, Soatto, LeCun, Baldassi, Borgs, Chayes, Sagun, Zecchina 2017 ICLR, 2019 JSTAT


distance from endpoints along path (rBP)
dist from η_2
dist from η_1
0
0.1
0.2
0.3
0.4
0.5
0.6
path along hypercube
geodesic path
bisecting path (λ=1.0)


distance matrix between fixed points; geodesic path (rBP)
step
step
0
100
200
300
400
500
0
0.1
0.2
0.3
0.4
0.5
0.6

Preliminary numerical results on the dynamics:

1) Finds liquid-stable fixed points
2) These are stable (also for the naive dynamics)
3) One can add external field to find fixed points which are close in Hamming distance the external input.
4) Extensive errors are corrected (given a fixed point, add an external field in its direction with a fraction of fields flipped and minimum J)
5) Fixed points further stabilized by Habbian learning $\eta - \eta$ and $\xi - \eta$
6) Internal representations are non trivially correlated, e.g. they possess a higher capacity if learned as attractors, $\alpha_c \simeq 6$ instead of $\alpha_c = 2$

Conclusion

- Regularization framework for non convex NNs
- Use ANN convergent dynamics as learning modules