Synaptic [classical and quantum] fluctuations as a recipe for robust and efficient neural network training

Carlo Baldassi
Bocconi University, Milano, Italy

R. Zecchina — Bocconi; IIGM Italian Institute for Genomic Medicine

C. Borgs
J. Chayes — Microsoft Research

H.J. Kappen — Radboud University Nijmegen

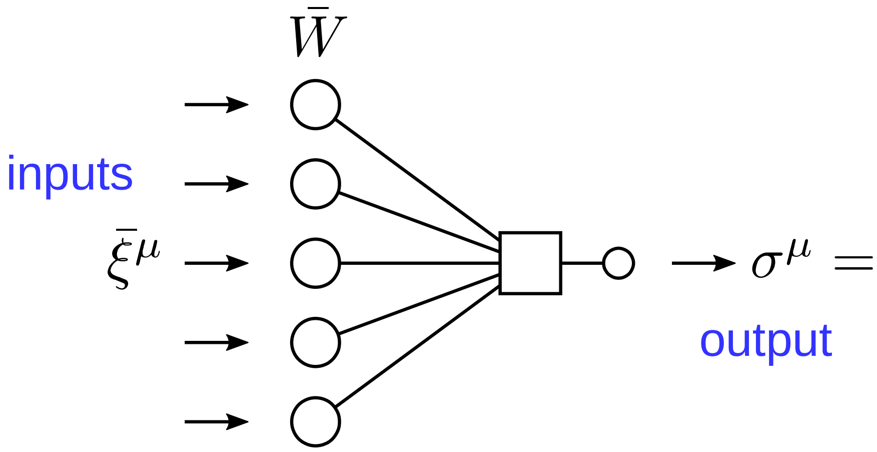
F. Gerace
A. Ingrosso
C. Lucibello
L. Saglietti
E. Tartaglione — Politecnico di Torino

P. Chaudhari
S. Soatto — The University of California, UCLA

Outline (1/4)

Simple (but surprisingly rich) neuronal models → Stat. Phys. Analysis
(replica method)

More complex neural networks, realistic data → numerical experiments



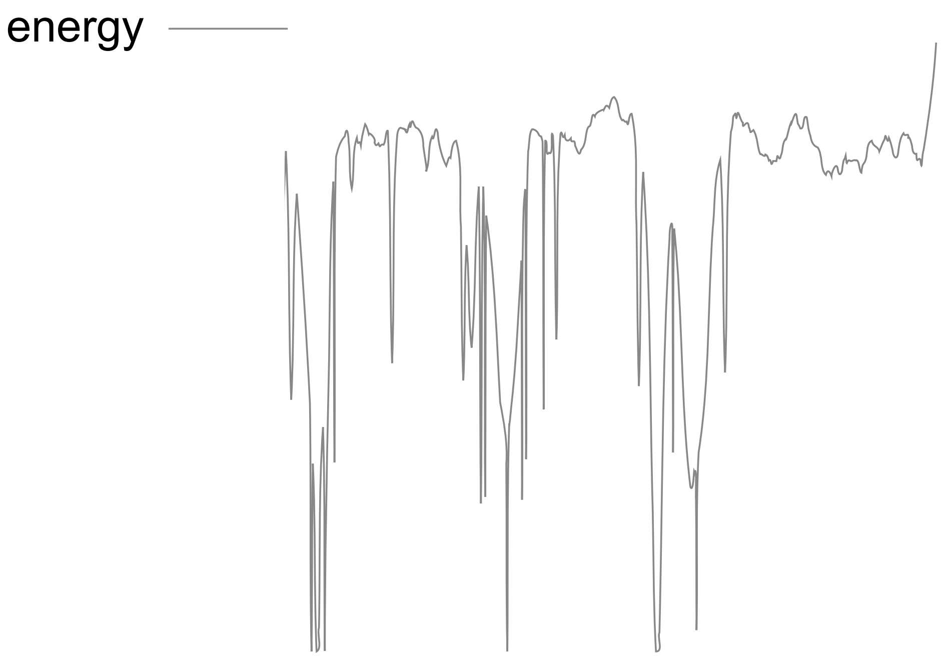
$$\sigma^{\mu} = \mathrm{sign}\left(\sum_{i=1}^{N} \xi_i^{\mu} W_i\right)$$

Outline (2/4)

Learning as an optimization problem

Highly non-convex landscape, many sharp local minima, isolated global minima

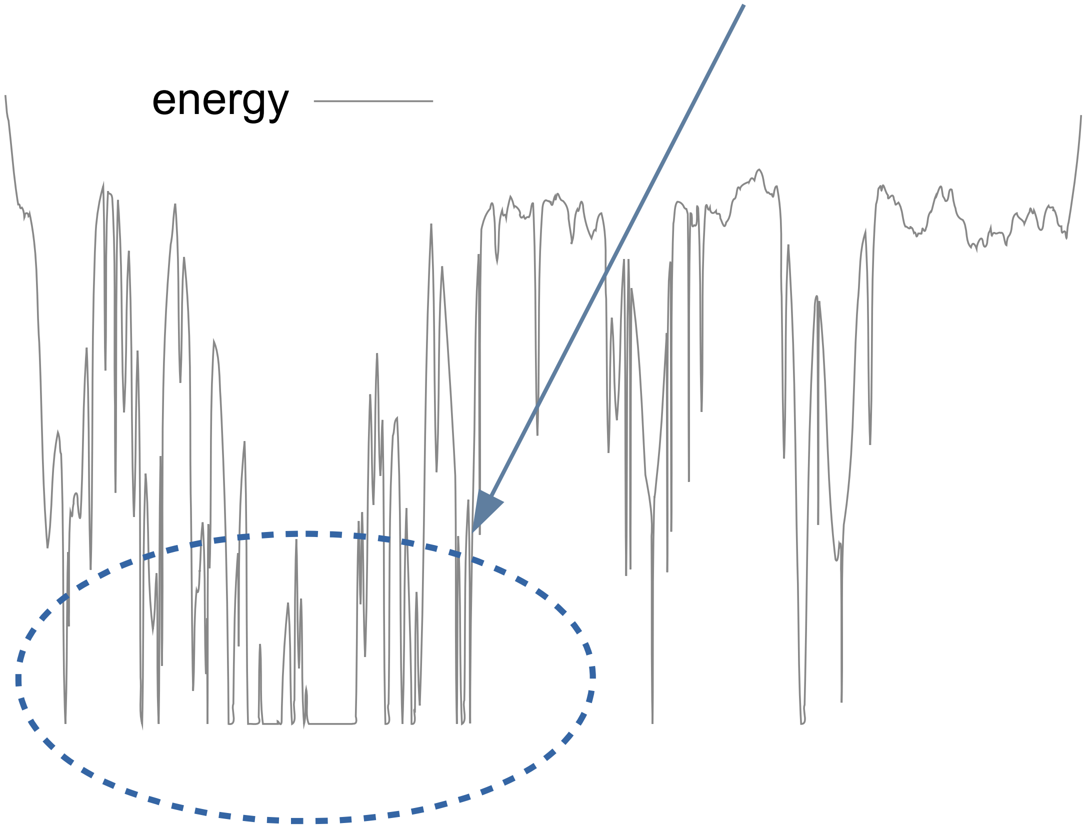
Should be hopeless, yet efficient heuristics exist



Outline (3/4)

Large deviation analysis → hidden dense region (high **local entropy**)

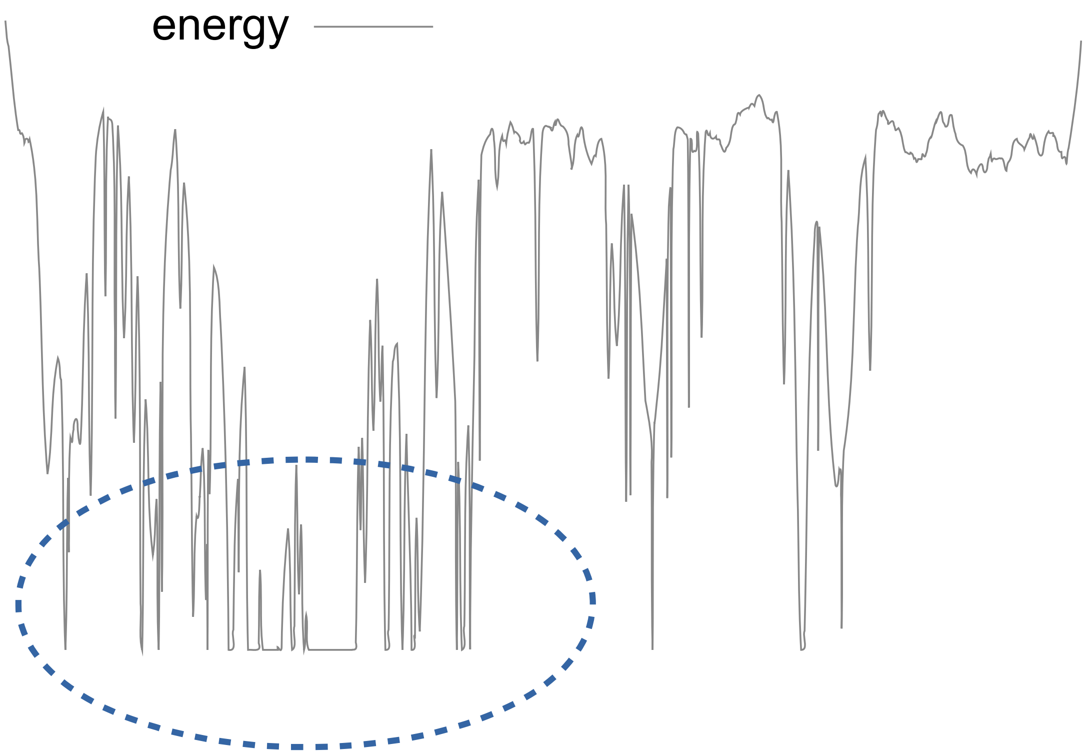
Hidden (eq. analysis ignores it)
Dense → robust → generalizes well
Accessible to simple algorithms



Outline (4/4)

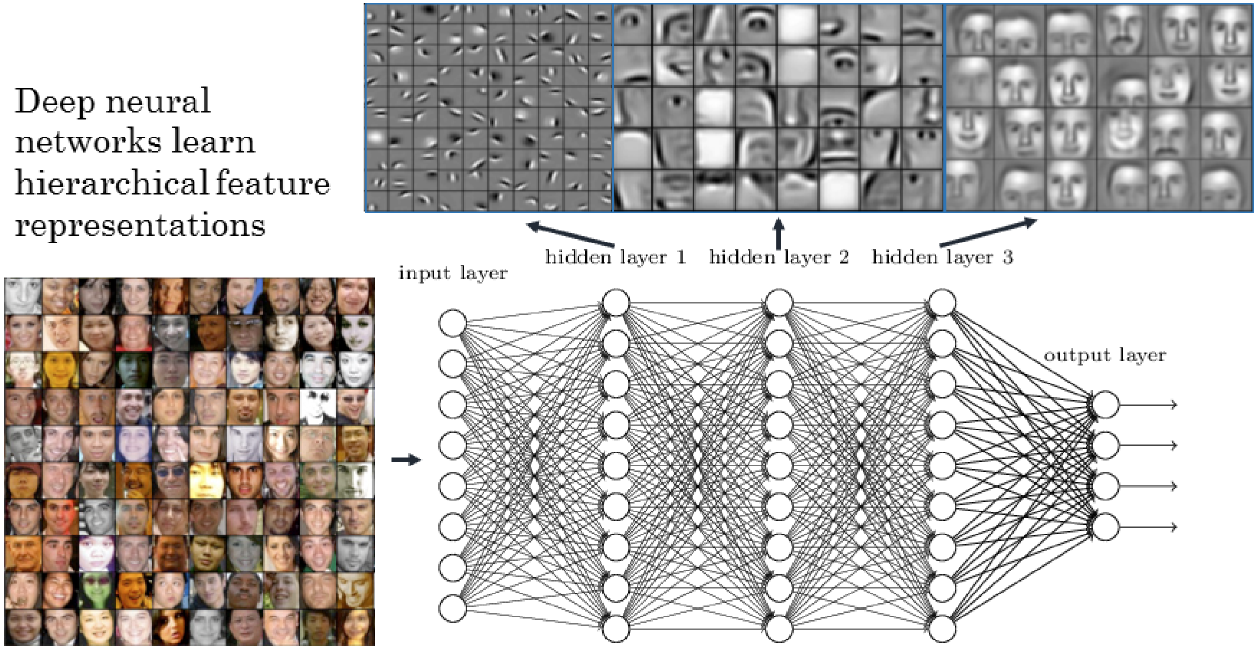
High-density states can be targeted by explicitly designed algorithms

But some processes are attracted to them too. In particular:
1) **Quantum annealing**
2) **Simple gradient on stochastic synapses**



Motivation, in brief

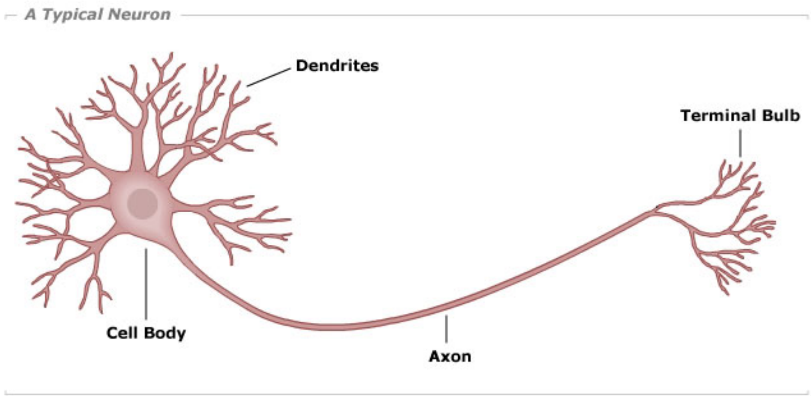
- Deep learning models: impressive results (super-human in some cases), very versatile
- Would benefit from: improved theoretical understanding, reduced resources usage, dedicated hardware
- Biological neurons: low-precision, quite noisy

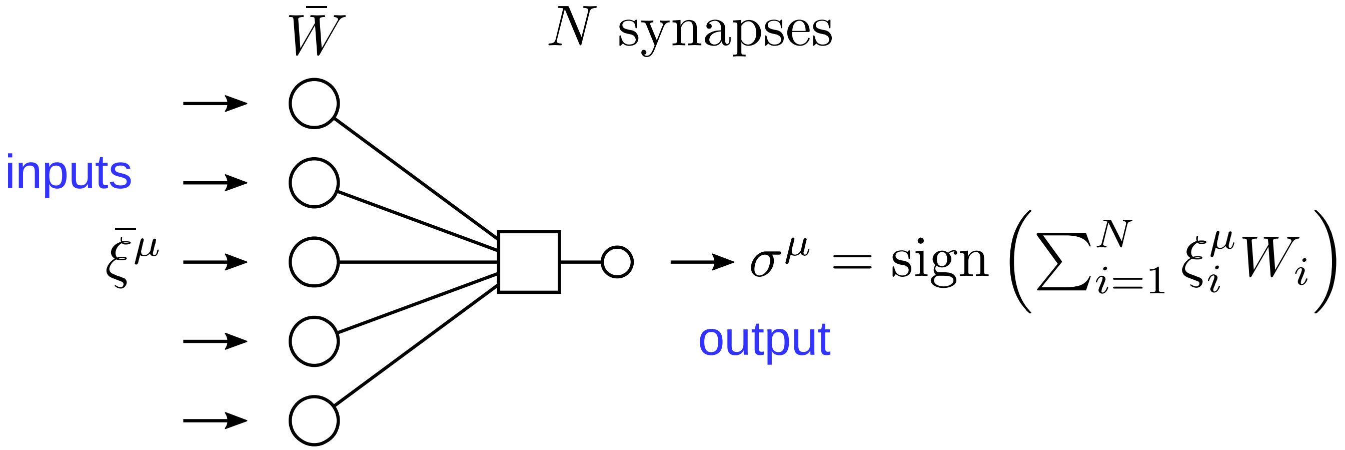


Deep neural networks learn hierarchical feature representations

Simplified neuronal models

- Simplest model neuron: perceptron (one layer, discretized time — no dynamics, no Dale's principle...)

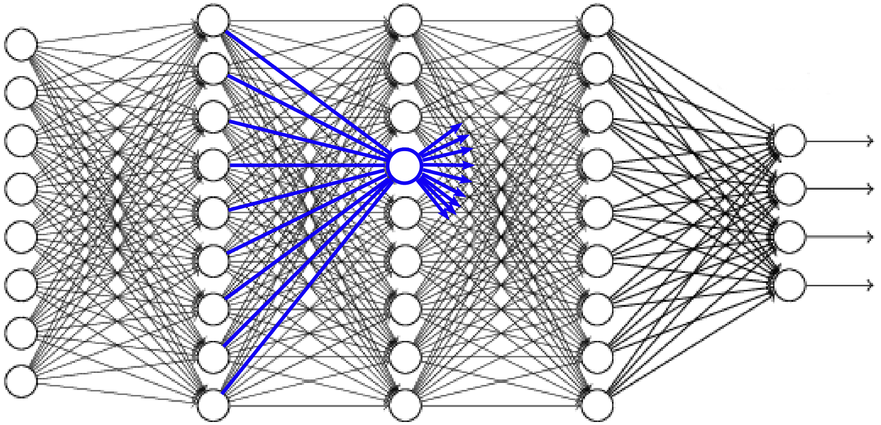


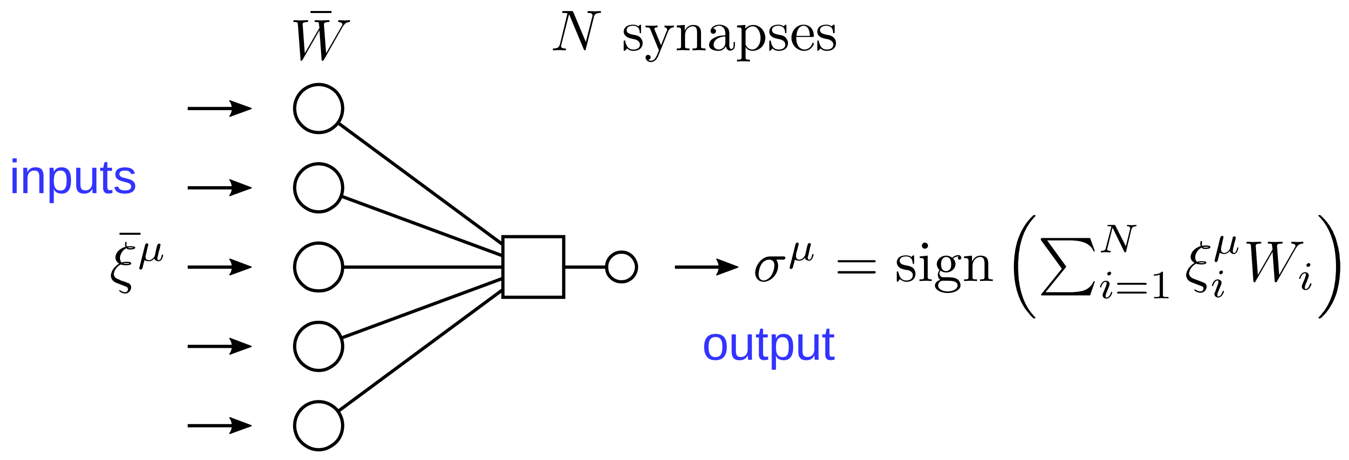


$$\sigma^{\mu} = \mathrm{sign}\left(\sum_{i=1}^{N} \xi_i^{\mu} W_i\right)$$

Simplified neuronal models

- Simplest model neuron: perceptron (building block of DNNs)

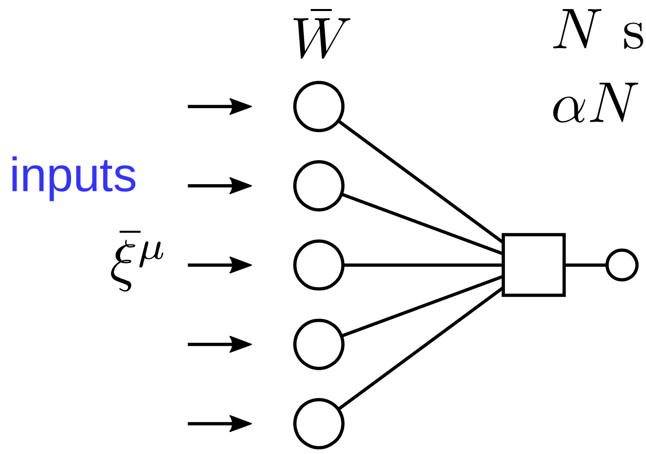




$$\sigma^{\mu} = \text{sign}\left(\sum_{i=1}^{N} \xi_i^{\mu} W_i\right)$$

Simplified neuronal models

- Simplest model neuron: perceptron (building block of DNNs)
- **Learning as an optimization problem**: minimize misclassifications



$\bar{W}$

inputs

$\bar{\xi}^\mu$

N synapses

αN input/output patterns (constraints)

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

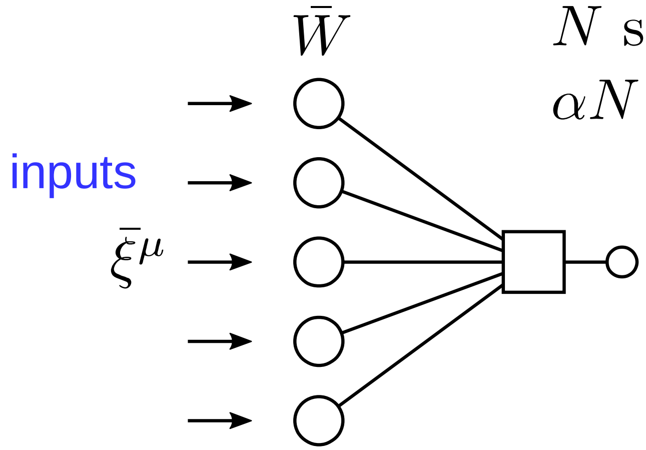
output

desired output

$\boxed{\text{find } W}$

Simplified neuronal models

- Simplest model neuron: perceptron (building block of DNNs)
- **Learning as an optimization problem**: minimize misclassifications
- Random i.i.d. patterns, large N → $\boxed{\text{stat. phys. tools}}$ → phase transitions (e.g. SAT/UNSAT)



N synapses

αN input/output patterns (constraints)

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

output

desired output

$\boxed{\text{find } W}$

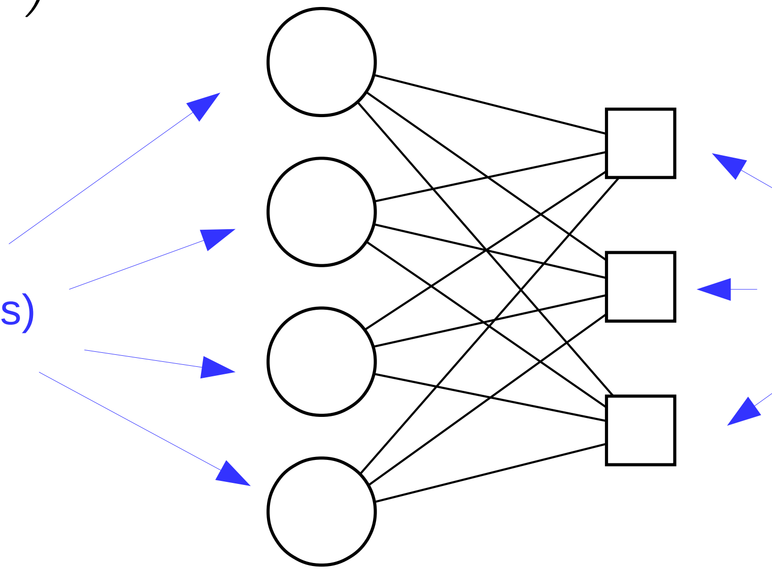
Equilibrium statistical physics of neural networks

- **Errors ~ energy** → Gibbs measure → Stat. Phys. Tools
- SAT/UNSAT phase transition (there is a “critical capacity” α_c)

$E_{\text{min}}=0$ $E_{\text{min}}>0$

$$E\left(W\right) = \sum_{\mu=1}^{\alpha N} \Theta\left(-\sigma_\mu^d \sum_{i=1}^{N} \xi_i^\mu W_i\right)$$

Neural net factor graph
(interactions ~ patterns)

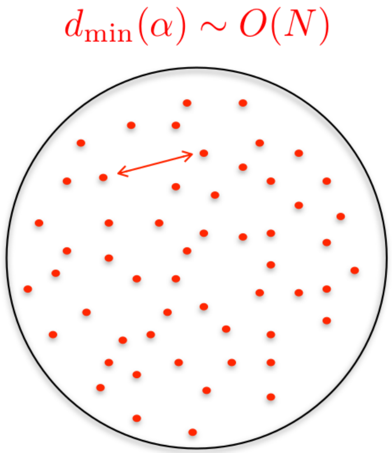


Variables
(synaptic weights)

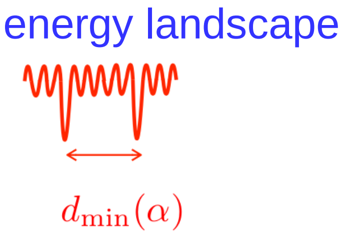
Patterns
(inputs+des. outputs)

Binary (±1) synapses, equilibrium results

- Good capacity: $\alpha_c = 0.83$
- But nasty optimization landscape: **typical** solutions are **isolated**, and immersed in **exponentially many local minima**
- Yet, simple efficient heuristics exists up to to $\alpha_{max} \sim 0.75$. They find **non-isolated** solutions, i.e. **atypical**
- Need to move away from equilibrium analysis



space of configurations



W. Krauth, M. Mézard, J. Phys. France, 1989
H. Huang, Y. Kabashima, Phys. Rev. E, 2014
A. Braunstein, R. Zecchina, PRL, 2006
C. Baldassi, A. Braunstein, et al, PNAS, 2007
C. Baldassi, J. Stat. Phys., 2009
C. Baldassi, A. Braunstein, J. Stat. Mech., 2015

Large deviation analysis: local entropy

- Idea: skew the statistical analysis towards non-isolated regions
- Given a configuration, we want to weigh it by how many solutions surround it — **local entropy**

number of solutions within a distance d

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

where $\mathbb{X}_{\xi}\left(W\right) = \prod_{\mu=1}^{\alpha N}\Theta\left(\sigma^{\mu}\tau\left(W,\xi^{\mu}\right)\right)$

"energy" = local entropy : $\mathscr{E}\left(\tilde{W}\right) = -\log\mathcal{N}\left(\tilde{W},d\right)$

Gibbs distribution with a different energy

free entropy $$Z(d) = \sum_{\{\tilde{W}\}} e^{-y\mathcal{E}(\tilde{W},d)}$$

$$\mathscr{F}\left(d,y\right) = -\frac{1}{Ny}\log\left(\sum_{\{\tilde{W}\}}\mathcal{N}\left(\tilde{W},d\right)^{y}\right)$$

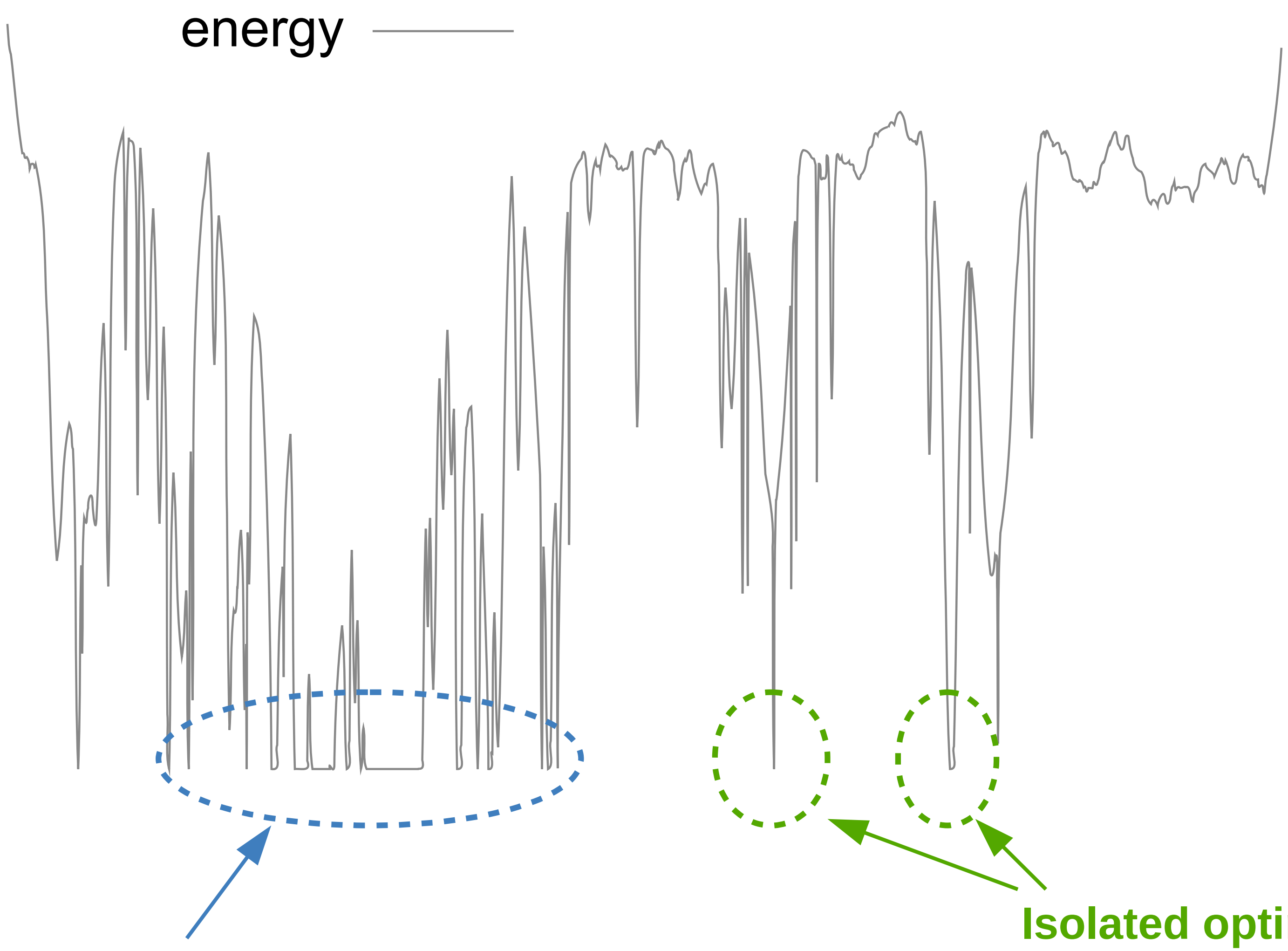
maximally dense cluster $y \to \infty$

C. Baldassi, A. Ingrosso, C. Lucibello, L. Saglietti, R. Zecchina, PRL, 2015

Large deviation analysis: **Results Summary**

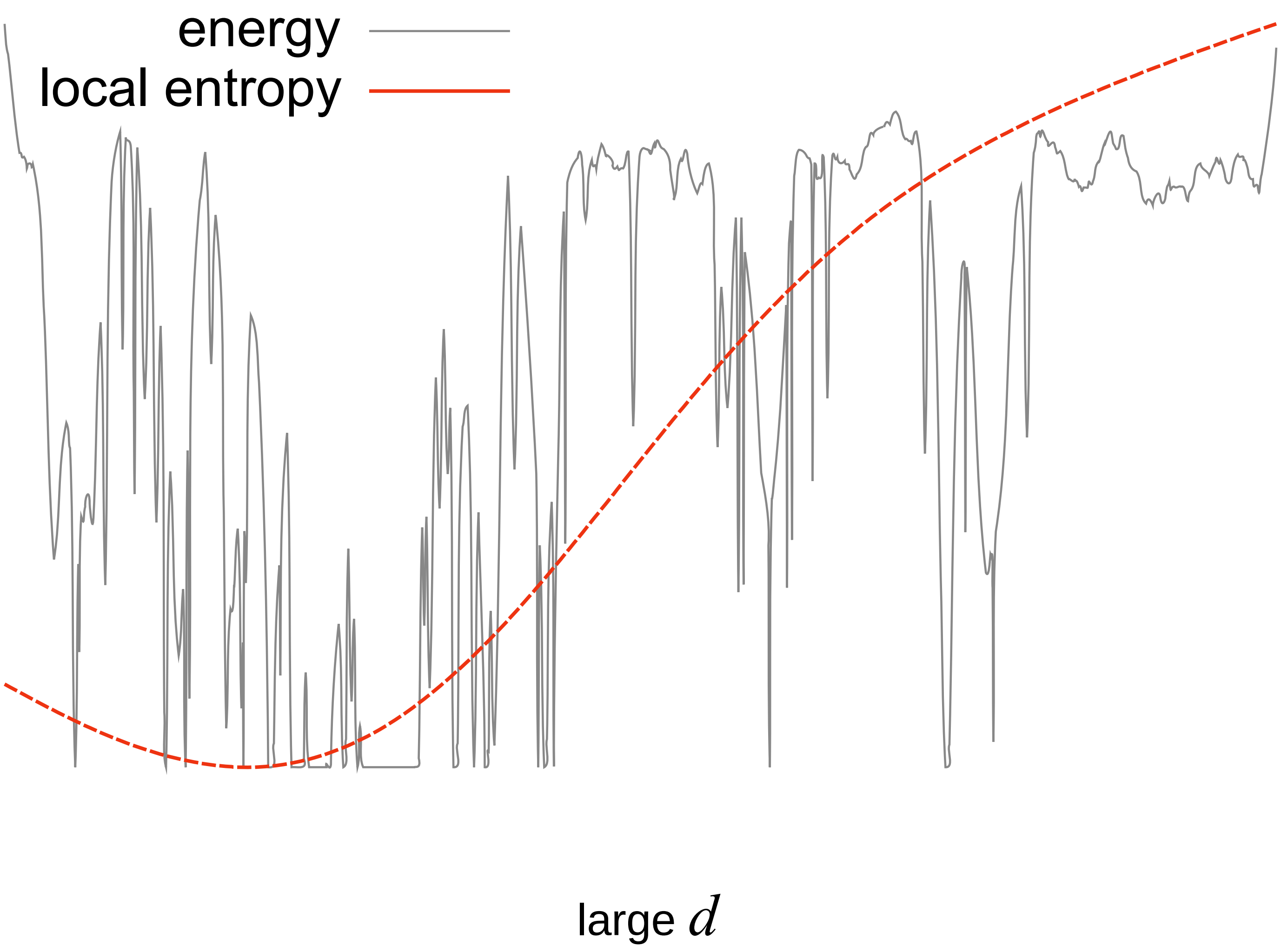
- **Ultra-dense regions exist** at least up to some $\alpha \sim 0.77$ (good agreement with heuristic solvers)
- **Better generalization properties** (quasi-bayesian) [note: dense regions are not planted, they are structural]
- Same phenomenology **also in with arbitrary number of synaptic states**
  - capacity saturates fast with number of states → **high-precision is not needed** (with the right algorithms)
- Same phenomenology **also in deeper networks with more realistic datasets**

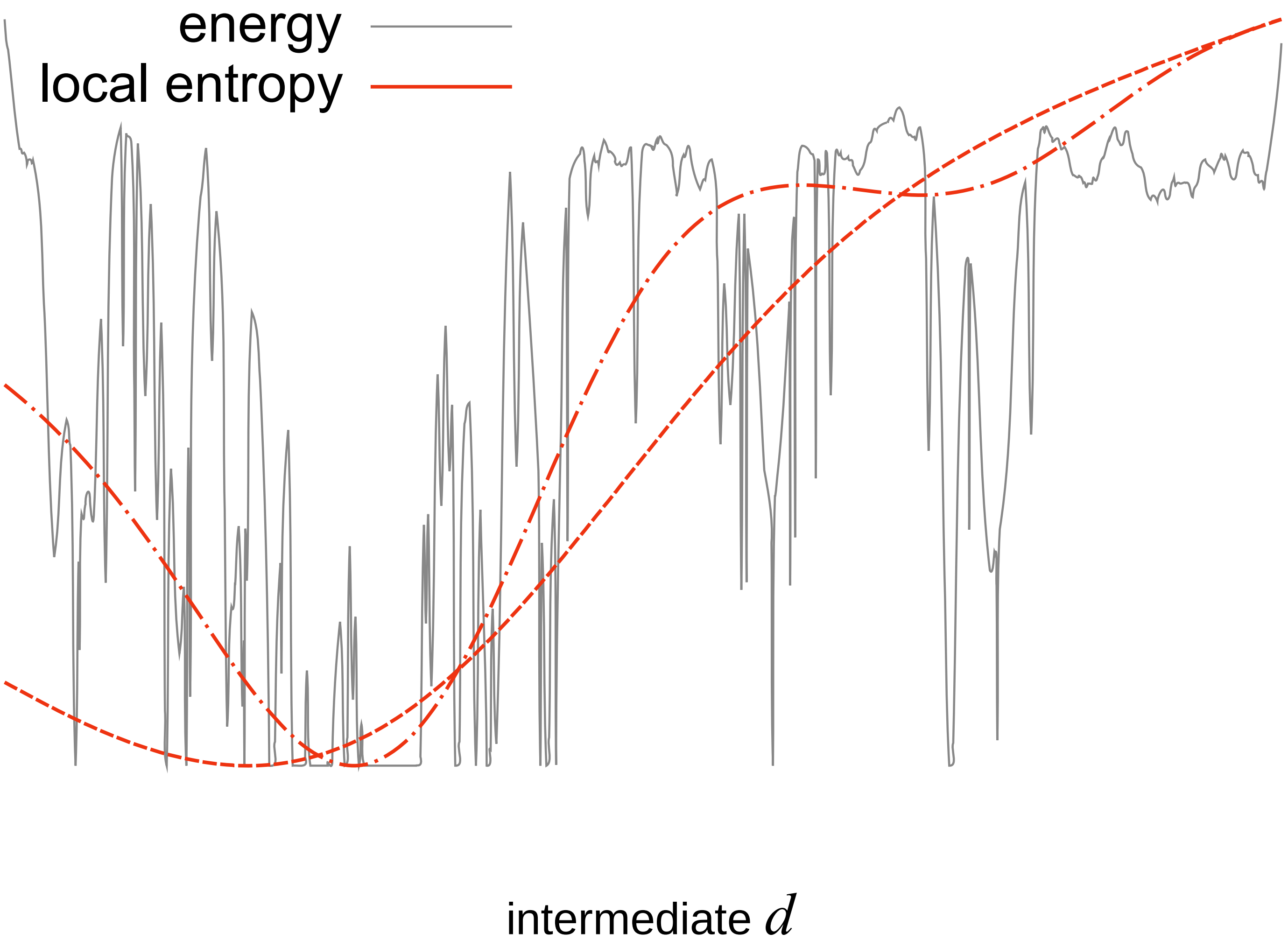
C. Baldassi, A. Ingrosso, C. Lucibello, L. Saglietti, R. Zecchina, PRL, 2015
C. Baldassi, F. Gerace, C. Lucibello, L. Saglietti, R. Zecchina, Phys. Rev. E, 2016

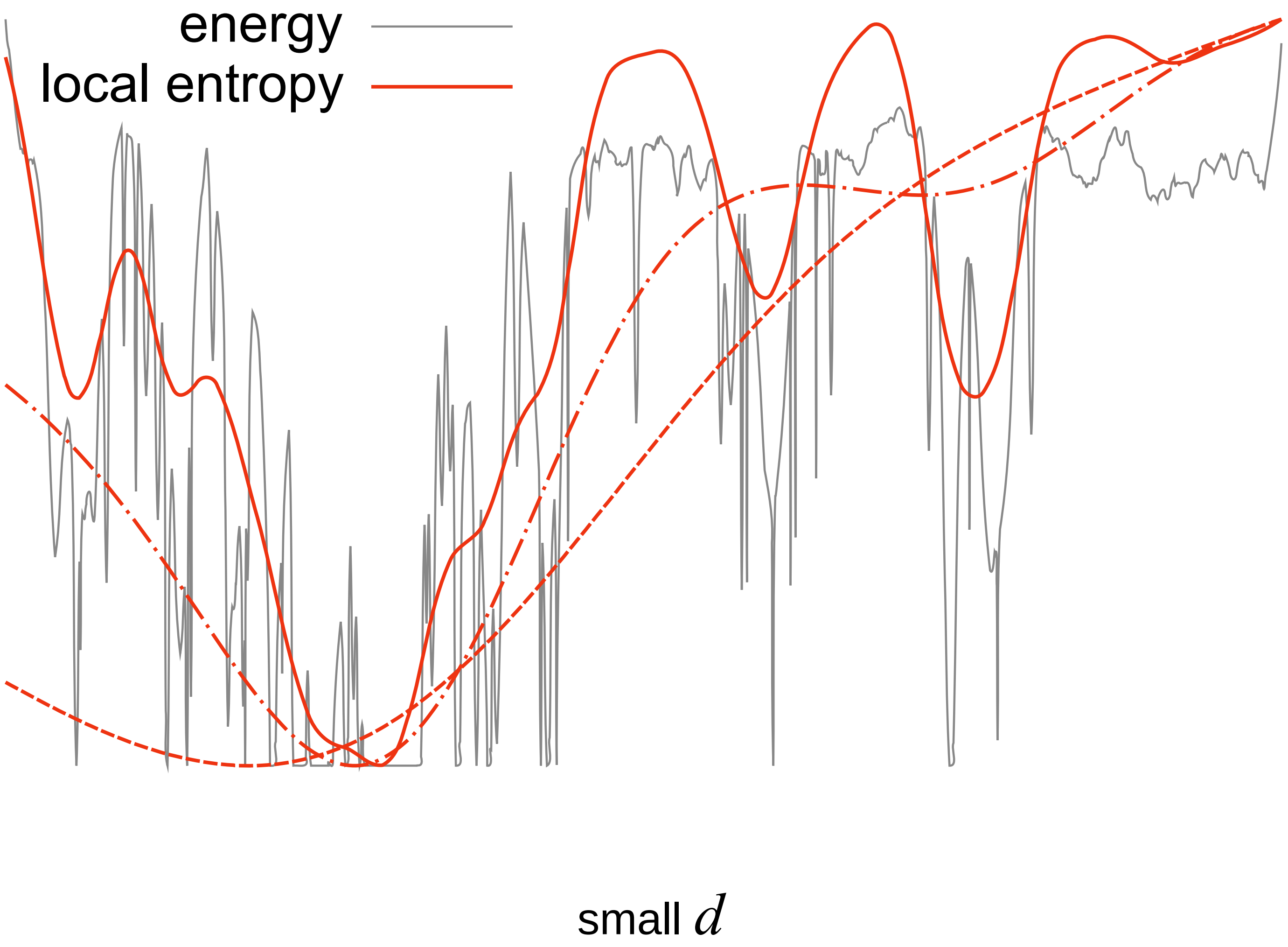


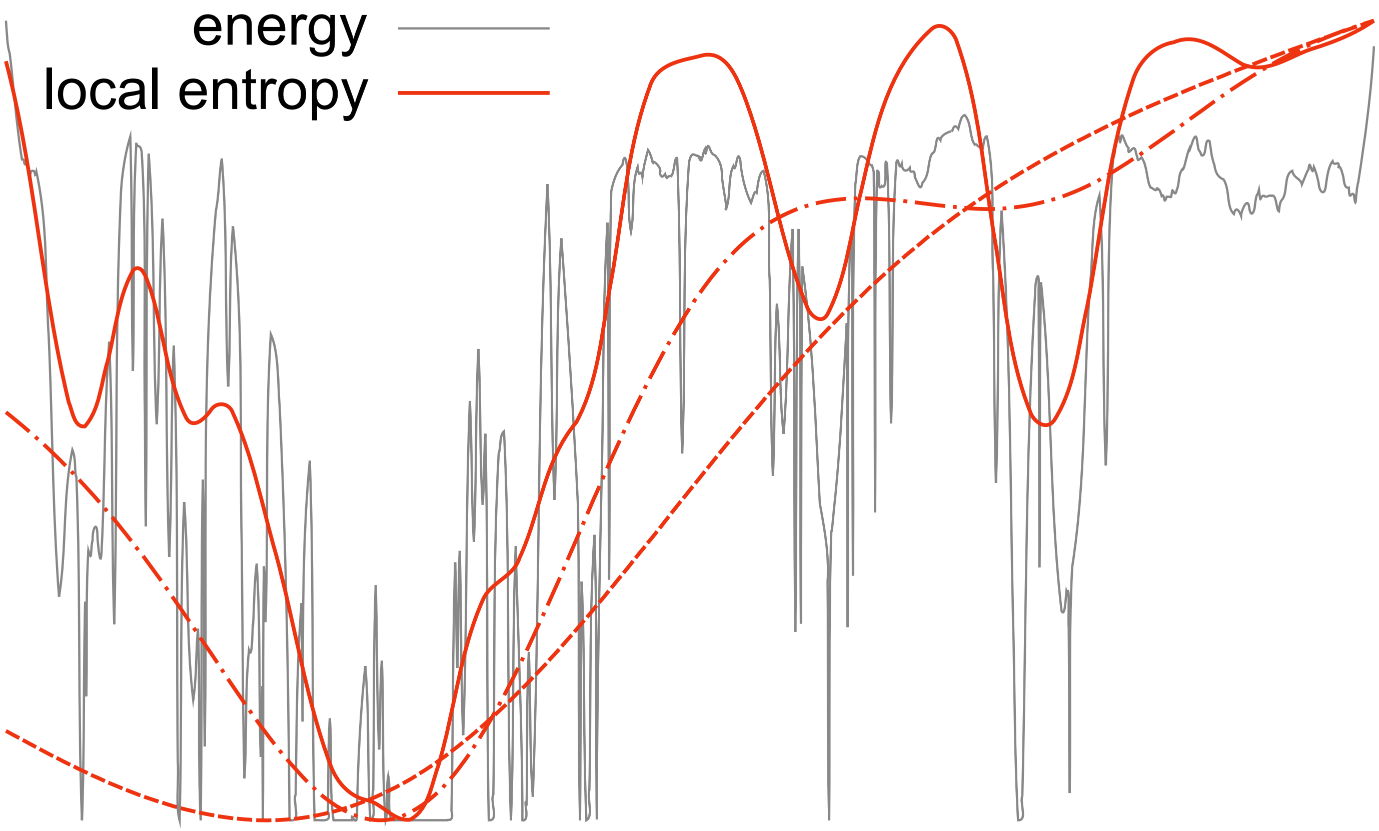
dense region

Isolated optima
(much more numerous than
depicted here)


energy
local entropy
large d


energy
local entropy
intermediate d


energy
local entropy
small d



large to small *d* = "focusing" or "scoping" process

Exploring loc.entropy by interacting replicas

- Assume that the parameter *y* is integer and transform the partition function:

$$P\left(\sigma;\beta,y,\gamma\right) = Z(\beta,y,\gamma)^{-1}\, e^{y\,\Phi(\sigma,\beta,\gamma)}$$

Large deviation measure

Local free entropy

$$\Phi\left(\sigma,\beta,\gamma\right) = \log\sum_{\{\sigma'\}} e^{-\beta E\left(\sigma'\right)-\gamma\, d\left(\sigma,\sigma'\right)}$$

$$\begin{aligned} Z\left(\beta,y,\gamma\right) &= \sum_{\{\sigma^\star\}} e^{y\Phi\left(\sigma^\star,\beta,\gamma\right)} \\ &= \sum_{\{\sigma^\star\}}\sum_{\{\sigma^a\}} e^{-\beta\sum_{a=1}^{y} E\left(\sigma^a\right)-\gamma\sum_{a=1}^{y} d\left(\sigma^\star,\sigma^a\right)} \end{aligned}$$

interaction

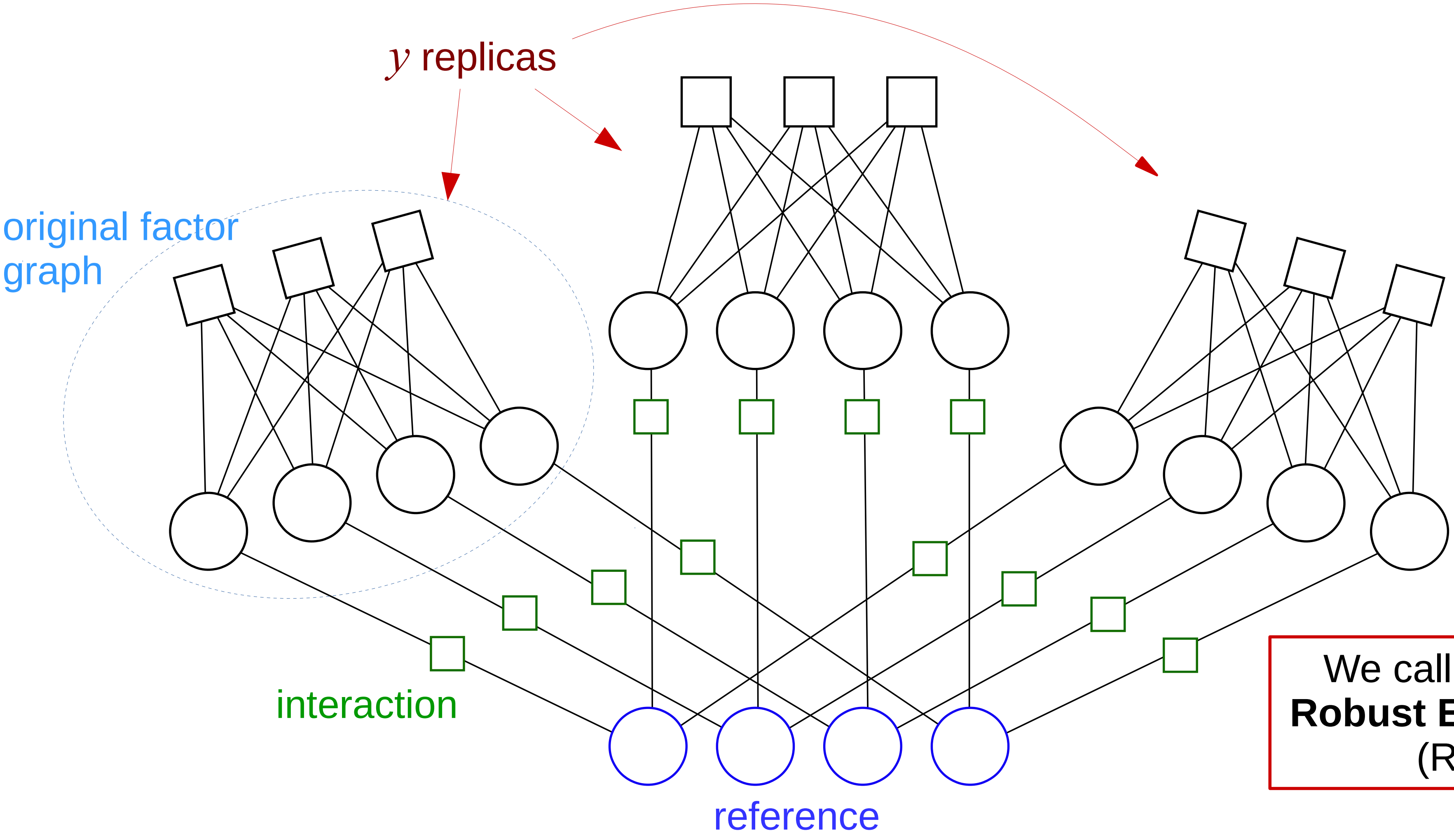
reference
(“center”)

y replicas

C. Baldassi, C. Borgs, J. Chayes, A. Ingrosso, C. Lucibello, L. Saglietti, R. Zecchina, PNAS, 2016

Exploring loc.entropy by interacting replicas

- Assume that the parameter y is integer and transform the partition function: **simple recipe to extend existing algorithms**

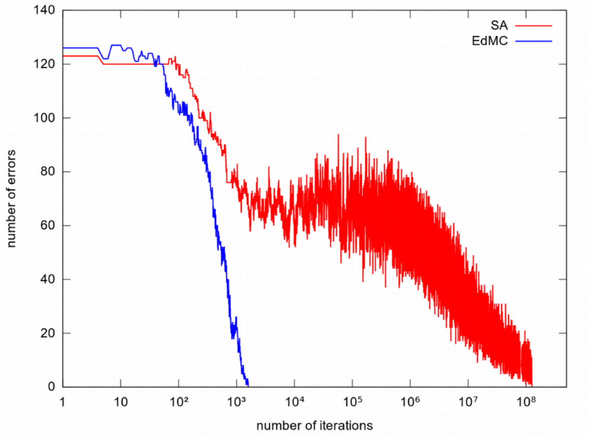


We call this the **Robust Ensemble** (RE)

C. Baldassi, C. Borgs, J. Chayes, A. Ingrosso, C. Lucibello, L. Saglietti, R. Zecchina, PNAS, 2016

Replicated variants of existing algorithms

- Replicated Simulated Annealing → doesn't get stuck → exponential speed-up w.r.t. non-replicated version
- Replicated Belief Propagation → efficient solver + semi-analytical tool for studying dense states + link with reinforcement heuristic
- Replicated Stochastic Gradient Descent (SGD) → dramatic improvement in capacity and speed w.r.t. non-replicated SGD
- (All of these were tested on 2-layer networks: same scenario as the 1-layer perceptron)



C. Baldassi, C. Borgs, J. Chayes, A. Ingrosso, C. Lucibello, L. Saglietti, R. Zecchina, PNAS, 2016

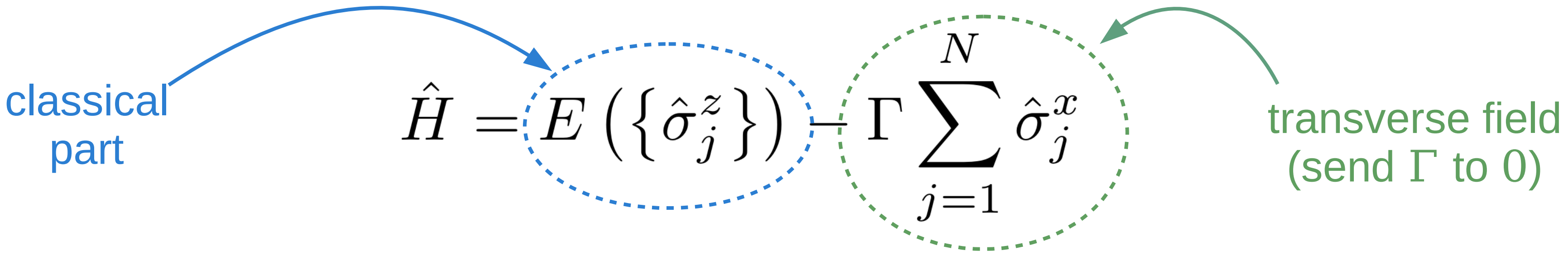
Deep neural networks

- Replicated SGD very deeply related to **EASGD** (Zhang et al. 2015) (11-layers, convolutional architecture, real-life images [ImageNet])
- **Entropy-SGD (2016)**: use Langevin dynamics to estimate local entropy, with excellent results (convolutional networks and recurrent networks, realistic inputs)
- **Parallel Local Entropy (Parle) (2017)**: state-of-the-art results on an array of tasks, but **~3x faster**, distributed learning
- Theoretical and numerical evidence from other research groups: minima are not all the same, wide minima (dense states) generalize better

S. Zhang, A. Choromanska, Y. LeCun, NIPS 2015
P. Chaudhari, A. Choromanska, S. Soatto, Y. LeCun, C. Baldassi, C. Borgs, J. Chayes, L. Sagun, R. Zecchina, ICLR 2016
P. Chaudhari, C. Baldassi, R. Zecchina, S. Soatto, A. Talwalkar, A. Oberman (arXiv 1707.00424), SysML 2018

Quantum annealing

- Quantum annealing strategy: use quantum fluctuations (rather than thermal fluctuations) to overcome energetic barriers
  - Classical energy function + quantum perturbation, slowly send the perturbation to zero



$$\hat{H} = E\left(\{\hat{\sigma}_j^z\}\right) - \Gamma \sum_{j=1}^{N} \hat{\sigma}_j^x$$

classical part

transverse field (send Γ to 0)

- So far: unclear if "true" QA really helps, compared to standard annealing, in any relevant concrete scenario

Suzuki-Trotter transformation

- Partition function transformation → "effective" replicated classical Hamiltonian (with infinite replicas, $y \rightarrow \infty$)

$$H_{\mathrm{eff}}\left(\left\{\sigma_j^a\right\}_{j,a}\right) = \frac{1}{y}\sum_{a=1}^{y} E\left(\left\{\sigma_j^a\right\}_j\right) - \frac{\gamma}{\beta}\sum_{a=1}^{y}\sum_{j=1}^{N}\sigma_j^a\sigma_j^{a+1} - \frac{NK}{\beta}$$

replicated classical part

interaction

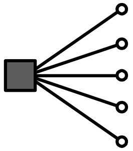
$\Gamma \rightarrow 0 \iff \gamma \rightarrow \infty$

$$\gamma = \frac{1}{2}\log\coth\left(\frac{\beta\Gamma}{y}\right)$$

- Can be simulated with MCMC (finite y) → Quantum Simulated Annealing (QSA)

Quantum annealing vs Robust ensemble

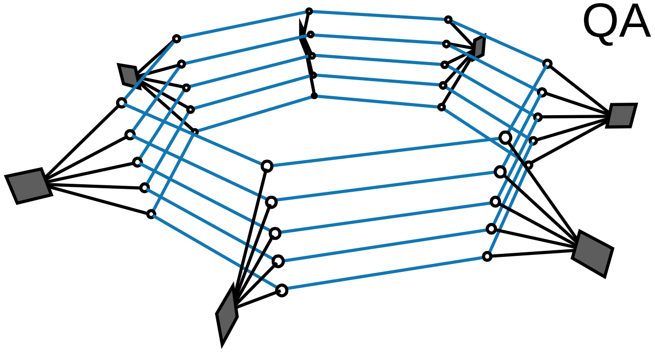
- Effective Hamiltonian after Suzuki-Trotter transformation: very similar to the robust ensemble description...



original factor graph

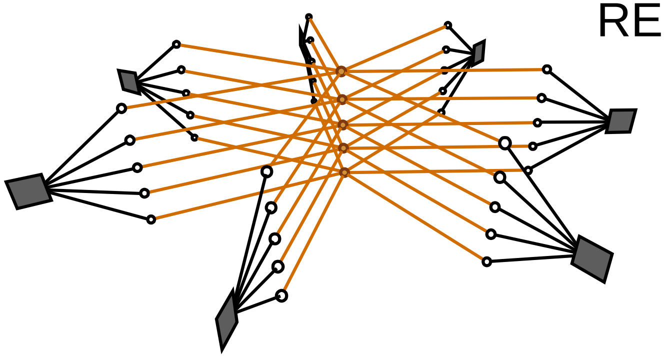
$$H_{\mathrm{eff}}\left(\left\{\sigma_j^a\right\}_{j,a}\right) = \frac{1}{y}\sum_{a=1}^{y} E\left(\left\{\sigma_j^a\right\}_j\right) - \frac{\gamma}{\beta}\sum_{a=1}^{y}\sum_{j=1}^{N}\sigma_j^a\sigma_j^{a+1} - \frac{NK}{\beta}$$

QA



$$H_{\mathrm{eff}}^{\mathrm{RE}}\left(\sigma^\star, \left\{\sigma_j^a\right\}_{j,a}\right) = \sum_{a=1}^{y} E\left(\left\{\sigma_j^a\right\}_j\right) - \frac{\lambda}{\beta}\sum_{a=1}^{y}\sum_{j=1}^{N}\sigma_j^a\sigma_j^\star$$

RE



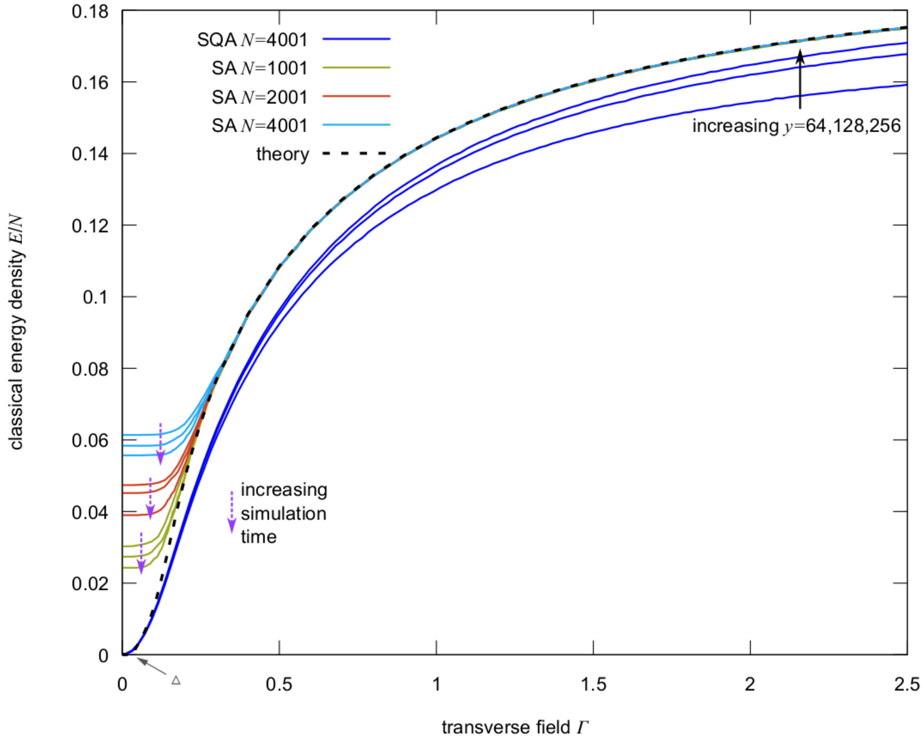
C. Baldassi, R. Zecchina, PNAS 2018 (in press, arXiv:1706.08470)

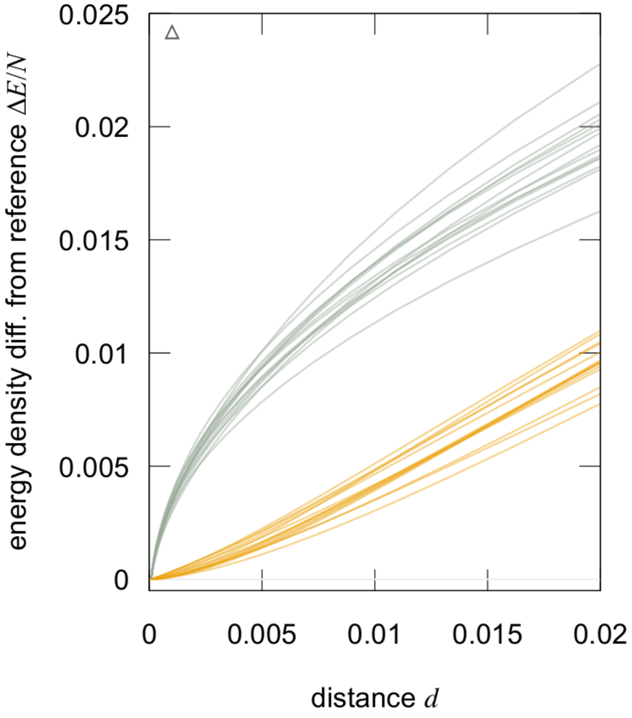
QSA on binary neural networks study

- Analytical calculations + numerical experiments + comparison with true QA in small instances

- Ends up in the dense states (exponential speed-up w.r.t. thermal annealing – a physical device would work in $\sim O(1)$...) (DWave-like)

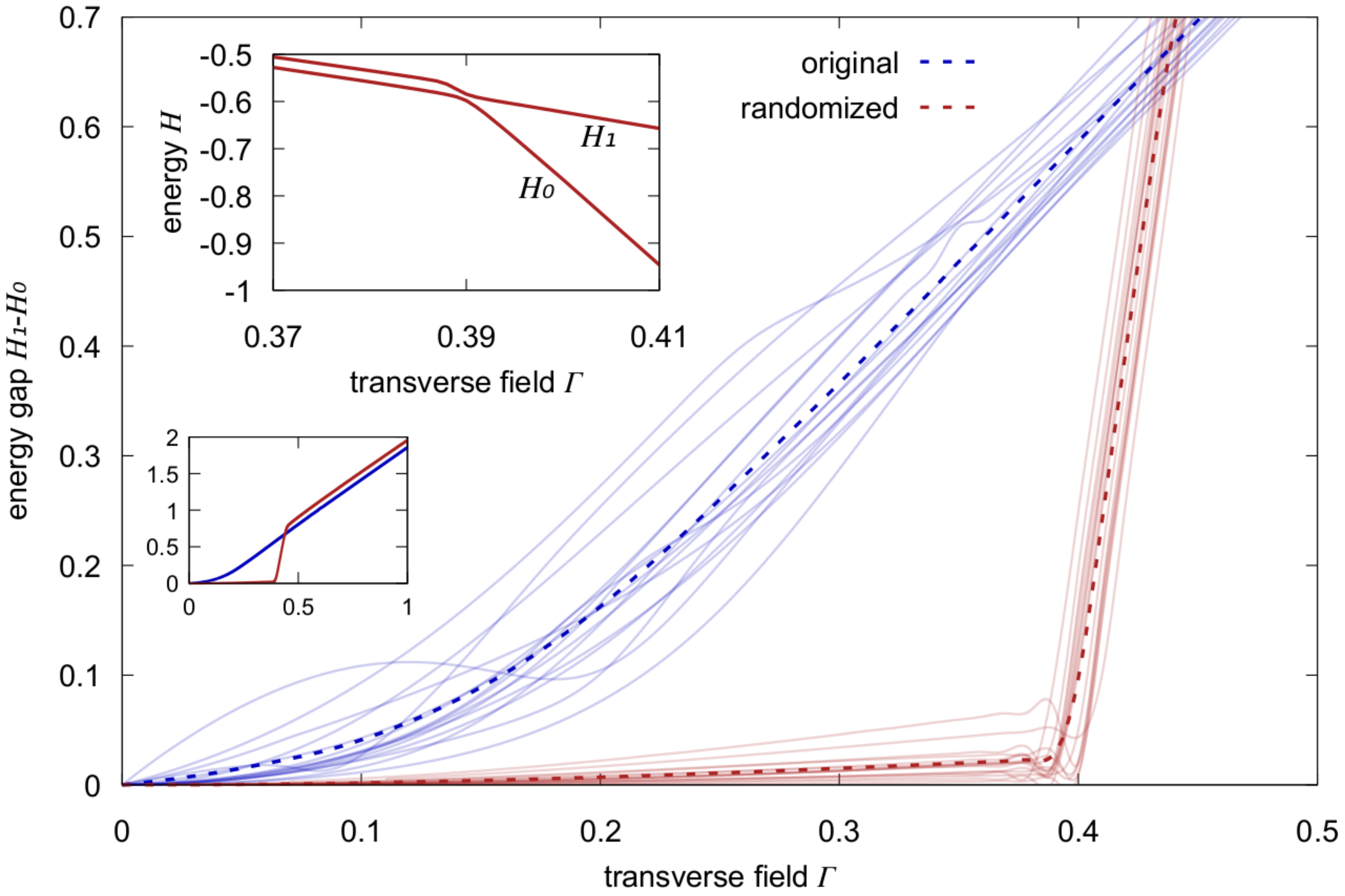
- **QA lowers kinetic energy by delocalizing → favors dense regions**





C. Baldassi, R. Zecchina, PNAS 2018 (in press, arXiv:1706.08470)

Geometric structure is essential



C. Baldassi, R. Zecchina, PNAS 2018 (in press, arXiv:1706.08470)

Stochastic synapses (overview)

- Consider a network with binary **stochastic** synapses, each controlled by a single parameter (a magnetization)
- At each presentation of a pattern, the actual values of the synapses are extracted at random
- Maximum likelihood approach (gradient descent, simulated annealing...)
- **Natural way to enforce robustness**: the network must try to get in the middle of a dense region where almost all solutions are good.
- …and indeed it does → **ends up in high local-entropy states**

  [This is no accident, the intuition is corroborated by a formal similarity with the RE setting]

C. Baldassi, F. Gerace, H.J. Kappen, C. Lucibello, L. Saglietti, E. Tartaglione, R. Zecchina, arXiv:1710.09825, 2017

Stochastic synapses

Standard: $$\arg\max_{W} \ \mathcal{L}(W) := \sum_{\mu \in D} \log P(y^{\mu}|x^{\mu}, W)$$

Bayes: $$P(y|x, D) \propto \int dW \, P(y|x, W) \prod_{\mu \in D} P(y^{\mu}|x^{\mu}, W) P(W)$$

Our approach: $$\arg\max_{\theta} \ \mathcal{L}(\theta) := \sum_{\mu \in D} \log \left[\int dW \ P(y^{\mu}|x^{\mu}, W) Q_{\theta}(W) \right]$$

This is the log-likelihood of a stochastic network where for each pattern (x,y) we sample the synapses W.

More complicated then Standard, less then Bayes. We get some of the goodies of the Bayesian approach (encoding structural priors, implicit robustness) while in the end we solve the Standard problem.

It can be understood as a relaxation of the standard problem.
Near the end of the training we have $Q_{\theta}(W) \to \delta(W - W^{*})$

Now we can use SGD to train binary networks!

C. Baldassi, F. Gerace, H.J. Kappen, C. Lucibello, L. Saglietti, E. Tartaglione, R. Zecchina, arXiv:1710.09825, 2017

Stochastic binary perceptron

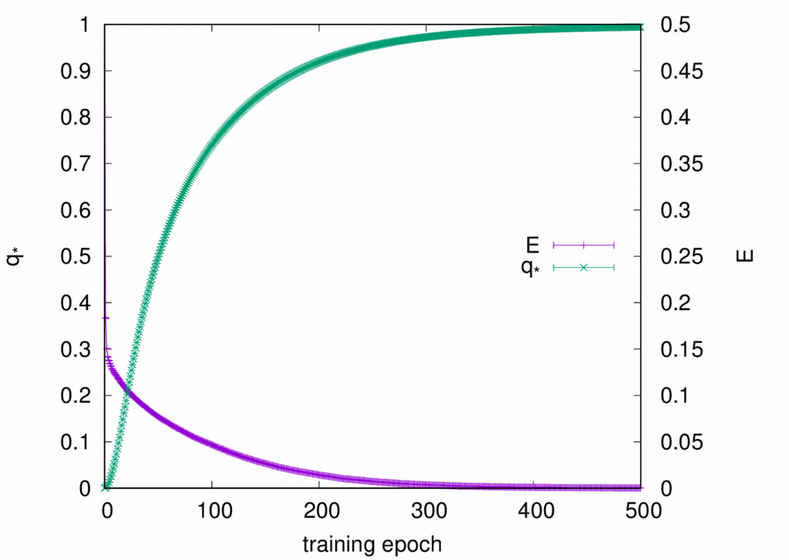
$$Q_m(W) = \prod_{i=1}^{N} \left[\frac{1}{2}(1+m_i)\delta_{W_i,+1} + \frac{1}{2}(1-m_i)\delta_{W_i,-1} \right]$$

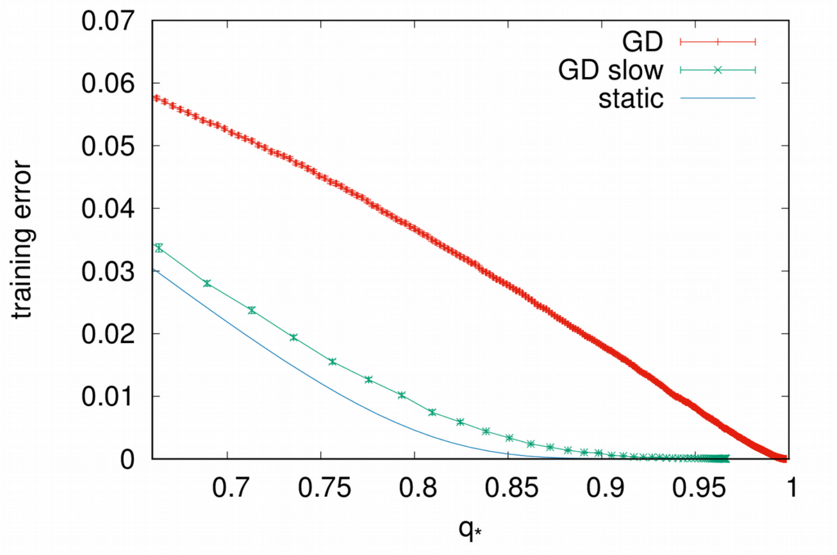
For large N the log-likelihood becomes

$$\mathcal{L}(m) = \sum_{(x,y)\in\mathcal{D}} \log \frac{1}{2} \operatorname{erfc}\left(-\frac{y \sum_i m_i x_i}{\sqrt{2\sum_i (1-m_i^2) x_i^2}} \right)$$

Smoothed-out landscape: we can do Gradient Descent on $\mathcal{L}(m)$ [or any other simple algorithm, actually]

Eventually, we can get to a polarized configuration which is in the middle of a dense region (verified analytically and numerically).





C. Baldassi, F. Gerace, H.J. Kappen, C. Lucibello, L. Saglietti, E. Tartaglione, R. Zecchina, arXiv:1710.09825, 2017

Stochastic binary multi-layer networks

- The procedure can be applied to more complex architectures (e.g. 3 or more hidden layers).
- However, in order to be efficient, we must introduce an approximation (uncorrelated neurons) which can be quite crude.
  - In cases where we can keep the correlations under control, the results are indeed very good (work in progress…)
- Or we could estimate the loglikelihood by sampling (also WIP...)

C. Baldassi, C.Borgs, J. Cheyes, F. Gerace, C. Lucibello, L. Saglietti, E. Tartaglione, R. Zecchina, in preparation, 2018

Conclusions, future directions

- A wide family of stochastic processes is attracted to out-of-equilibrium states with peculiar (and useful) geometric properties in the space of configurations. Can we extend and generalize these results?
- Learning with low precision synapses can be made extremely simple, and the performances are very good: can we improve existing networks/design new neural hardware?
- Accessible states in other problems – by exploring the robust ensemble we can find dense regions as if they were typical
- Applications in inference?
- We want to address unsupervised learning (automatic feature extraction) as well [we have preliminary results, on hold due to time constraints...]

Thanks

R. Zecchina
C. Baldassi

Bocconi

IIGM
Italian Institute for Genomic Medicine

F. Gerace
A. Ingrosso
C. Lucibello
L. Saglietti
E. Tartaglione

POLITECNICO DI TORINO
1859 1906

C. Borgs
J. Chayes

Microsoft® Research

H.J. Kappen

IN·DEI·NOMINE·FELICITER

Radboud University Nijmegen

P. Chaudhari
S. Soatto

THE · UNIVERSITY · OF · CALIFORNIA ·
LET THERE BE LIGHT
UCLA