Using Physical Insights for Machine Learning
IPAM-2019

On the existence of wide flat minima in neural network landscapes: analytic and algorithm approaches

Carlo Baldassi

Artificial Intelligence Lab
Bocconi

Follow-up of
R. Zecchina's talk

C. Baldassi, F. Pittorino, R. Zecchina, arXiv:1905.07833, 2019

Outline

- **Discrete weights**: rare dense clusters of solutions (High-Local-Entropy regions, HLE) exist (shallow networks, random patterns)
- Continuous case: Wide Flat Minima (WFM)
- Replica theory: recovering the Local Entropy large-deviation analysis from the 1-RSB formalism
- Message Passing: BP to estimate local volumes, fBP to find WFM
- Practical algorithms for finding WFM
  - Using Cross-Entropy helps (control the norm)
  - Replication (e.g. from LAL to eLAL)
- Non-random patterns: WFM generalize better; correlation between volumes and hessians
- Randomized real-world dataset: not that much of a difference...

Outline

- Discrete weights: rare dense clusters of solutions (High-Local-Entropy regions, HLE) exist (shallow networks, random patterns)
- **Continuous case**: Wide Flat Minima (WFM)
- Replica theory: recovering the Local Entropy large-deviation analysis from the 1-RSB formalism
- Message Passing: BP to estimate local volumes, fBP to find WFM
- Practical algorithms for finding WFM
  - Using Cross-Entropy helps (control the norm)
  - Replication (e.g. from LAL to eLAL)
- Non-random patterns: WFM generalize better; correlation between volumes and hessians
- Randomized real-world dataset: not that much of a difference...

Outline

- Discrete weights: rare dense clusters of solutions (High-Local-Entropy regions, HLE) exist (shallow networks, random patterns)
- Continuous case: Wide Flat Minima (WFM)
- **Replica theory**: recovering the Local Entropy large-deviation analysis from the 1-RSB formalism
- Message Passing: BP to estimate local volumes, fBP to find WFM
- Practical algorithms for finding WFM
  - Using Cross-Entropy helps (control the norm)
  - Replication (e.g. from LAL to eLAL)
- Non-random patterns: WFM generalize better; correlation between volumes and hessians
- Randomized real-world dataset: not that much of a difference...

Outline

- Discrete weights: rare dense clusters of solutions (High-Local-Entropy regions, HLE) exist (shallow networks, random patterns)
- Continuous case: Wide Flat Minima (WFM)
- Replica theory: recovering the Local Entropy large-deviation analysis from the 1-RSB formalism
- **Message Passing**: BP to estimate local volumes, fBP to find WFM
- Practical algorithms for finding WFM
  - Using Cross-Entropy helps (control the norm)
  - Replication (e.g. from LAL to eLAL)
- Non-random patterns: WFM generalize better; correlation between volumes and hessians
- Randomized real-world dataset: not that much of a difference...

Outline

- Discrete weights: rare dense clusters of solutions (High-Local-Entropy regions, HLE) exist (shallow networks, random patterns)
- Continuous case: Wide Flat Minima (WFM)
- Replica theory: recovering the Local Entropy large-deviation analysis from the 1-RSB formalism
- Message Passing: BP to estimate local volumes, fBP to find WFM
- **Practical algorithms** for finding WFM
  - Using Cross-Entropy helps (control the norm)
  - Replication (e.g. from LAL to eLAL)
- Non-random patterns: WFM generalize better; correlation between volumes and hessians
- Randomized real-world dataset: not that much of a difference...

Outline

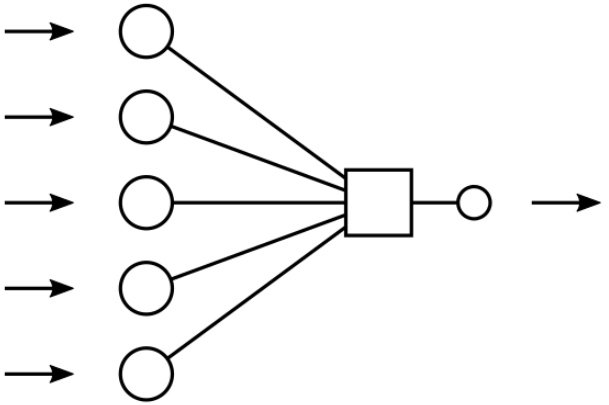
- Discrete weights: rare dense clusters of solutions (High-Local-Entropy regions, HLE) exist (shallow networks, random patterns)
- Continuous case: Wide Flat Minima (WFM)
- Replica theory: recovering the Local Entropy large-deviation analysis from the 1-RSB formalism
- Message Passing: BP to estimate local volumes, fBP to find WFM
- Practical algorithms for finding WFM
  - Using Cross-Entropy helps (control the norm)
  - Replication (e.g. from LAL to eLAL)
- Non-random patterns: **WFM generalize better**; correlation between volumes and hessians
- Randomized real-world dataset: not that much of a difference...

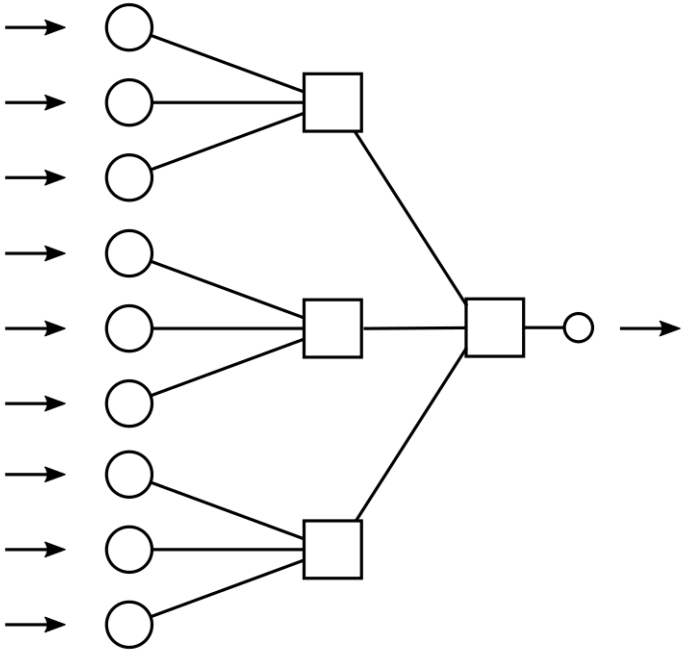
Outline

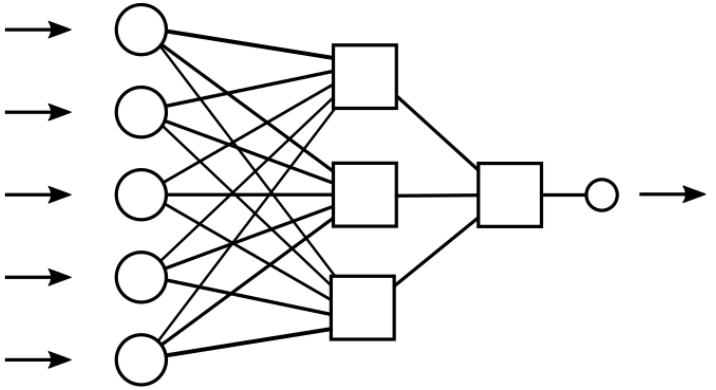
- Discrete weights: rare dense clusters of solutions (High-Local-Entropy regions, HLE) exist (shallow networks, random patterns)
- Continuous case: Wide Flat Minima (WFM)
- Replica theory: recovering the Local Entropy large-deviation analysis from the 1-RSB formalism
- Message Passing: BP to estimate local volumes, fBP to find WFM
- Practical algorithms for finding WFM
  - Using Cross-Entropy helps (control the norm)
  - Replication (e.g. from LAL to eLAL)
- Non-random patterns: WFM generalize better; correlation between volumes and hessians
- **Randomized real-world dataset**: not that much of a difference...

The network models that we studied

- Random input patterns
- The weights W can be continuous or discrete
- No need for differentiable losses

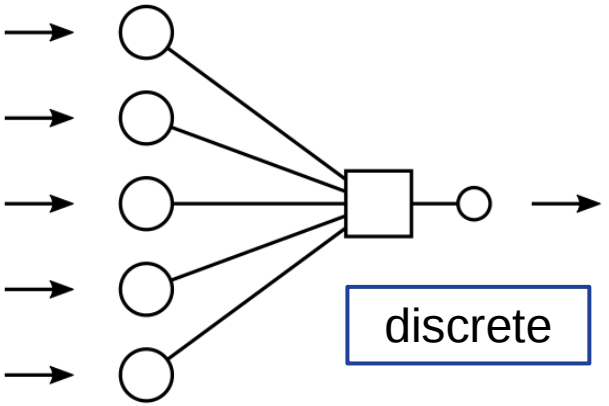


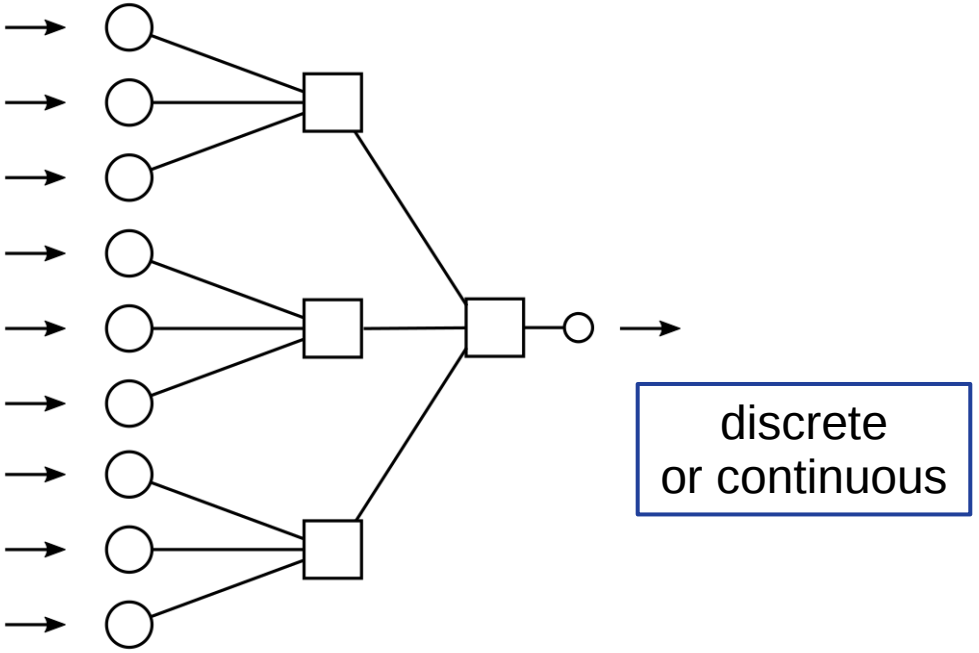


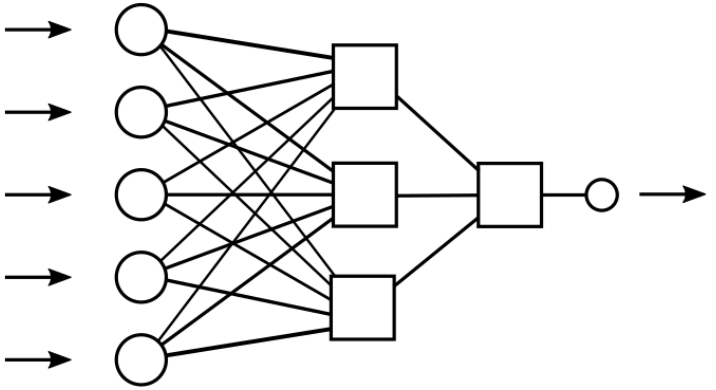


The network models that we studied

- Random input patterns
- The weights W can be continuous or discrete
- No need for differentiable losses

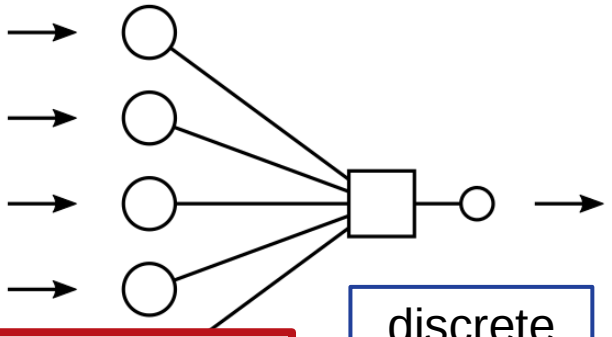






The network models that we studied

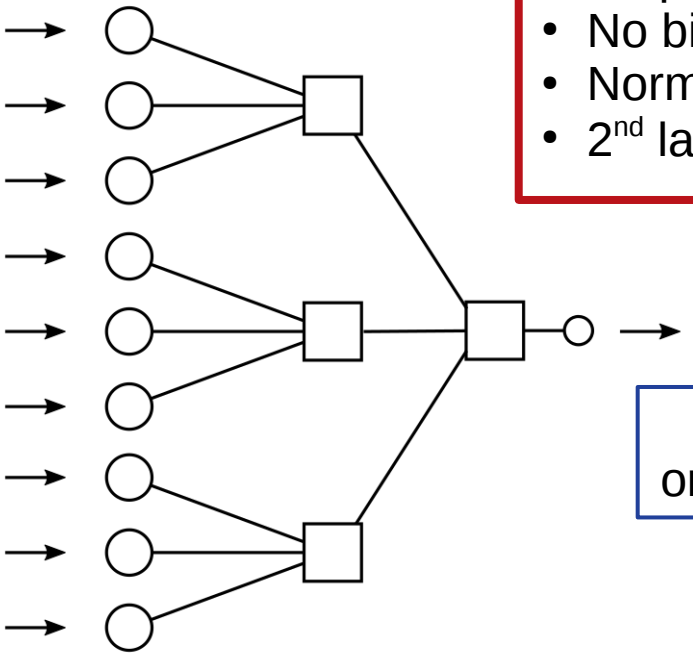
- Random input patterns
- The weights W can be continuous or discrete
- No need for differentiable losses



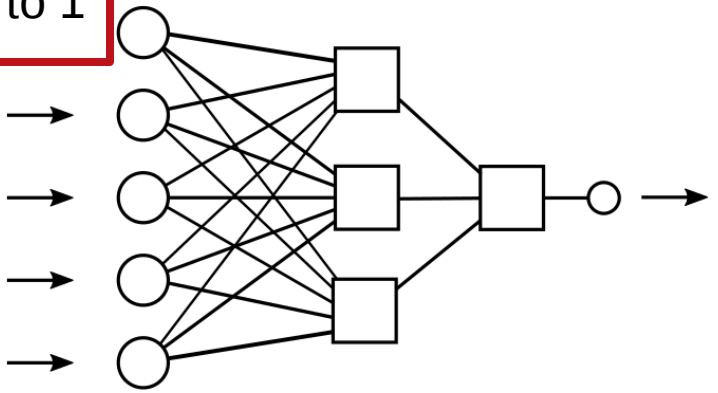
discrete

Throughout the talk

- Step transfer functions
- No bias
- Normalized weights
- 2nd layer weights fixed to 1



discrete
or continuous



Local entropy measure

- Main idea: perform a large-deviation analysis, **bias the statistical measure towards dense (wide, flat) regions**

C. Baldassi, A. Ingrosso, C. Lucibello, L. Saglietti, R. Zecchina, PRL, 2015
C. Baldassi, F. Gerace, C. Lucibello, L. Saglietti, R. Zecchina, PRE, 2016
C. Baldassi, C. Borgs, J. Chayes, A. Ingrosso, C. Lucibello, L. Saglietti, R. Zecchina, PNAS, 2016
C. Baldassi, E. Malatesta, R. Zecchina, PRL, 2019
C. Baldassi, F. Pittorino, R. Zecchina, arXiv:1905.07833, 2019

Local entropy measure

- Main idea: perform a large-deviation analysis, **bias the statistical measure towards dense (wide, flat) regions**
- In practice: define a "**local (free) entropy**" (basically the free energy of a region of the configuration space)

$$\Phi\left(\tilde{W};\beta,\gamma\right)=\log\sum\nolimits_{\{W\}}\exp\left(-\beta\mathcal{L}_{\mathrm{NE}}\left(W\right)-\gamma d\left(W,\tilde{W}\right)\right)$$

C. Baldassi, A. Ingrosso, C. Lucibello, L. Saglietti, R. Zecchina, PRL, 2015
C. Baldassi, F. Gerace, C. Lucibello, L. Saglietti, R. Zecchina, PRE, 2016
C. Baldassi, C. Borgs, J. Chayes, A. Ingrosso, C. Lucibello, L. Saglietti, R. Zecchina, PNAS, 2016
C. Baldassi, E. Malatesta, R. Zecchina, PRL, 2019
C. Baldassi, F. Pittorino, R. Zecchina, arXiv:1905.07833, 2019

Local entropy measure

- Main idea: perform a large-deviation analysis, **bias the statistical measure towards dense (wide, flat) regions**
- In practice: define a "**local (free) entropy**" (basically the free energy of a region of the configuration space)

$$\Phi\left(\tilde{W};\beta,\gamma\right)=\log\sum\nolimits_{\{W\}}\exp\left(-\beta\mathcal{L}_{\mathrm{NE}}\left(W\right)-\gamma d\left(W,\tilde{W}\right)\right)$$

"count all configurations below a certain loss and within a certain distance from a reference"

C. Baldassi, A. Ingrosso, C. Lucibello, L. Saglietti, R. Zecchina, PRL, 2015
C. Baldassi, F. Gerace, C. Lucibello, L. Saglietti, R. Zecchina, PRE, 2016
C. Baldassi, C. Borgs, J. Chayes, A. Ingrosso, C. Lucibello, L. Saglietti, R. Zecchina, PNAS, 2016
C. Baldassi, E. Malatesta, R. Zecchina, PRL, 2019
C. Baldassi, F. Pittorino, R. Zecchina, arXiv:1905.07833, 2019

Local entropy measure

- Main idea: perform a large-deviation analysis, **bias the statistical measure towards dense (wide, flat) regions**
- In practice: define a "**local (free) entropy**" (basically the free energy of a region of the configuration space)

$$\Phi\left(\tilde{W};\beta,\gamma\right)=\log\sum\nolimits_{\{W\}}\exp\left(-\beta\mathcal{L}_{\mathrm{NE}}\left(W\right)-\gamma d\left(W,\tilde{W}\right)\right)$$

"count all configurations below a certain loss and within a certain distance from a reference"

- Instead of minimizing the loss (the energy), **minimize the negative local entropy**

$$\lim_{y\to\infty}\frac{1}{Z}e^{y\Phi\left(\tilde{W}\right)}$$

C. Baldassi, A. Ingrosso, C. Lucibello, L. Saglietti, R. Zecchina, PRL, 2015
C. Baldassi, F. Gerace, C. Lucibello, L. Saglietti, R. Zecchina, PRE, 2016
C. Baldassi, C. Borgs, J. Chayes, A. Ingrosso, C. Lucibello, L. Saglietti, R. Zecchina, PNAS, 2016
C. Baldassi, E. Malatesta, R. Zecchina, PRL, 2019
C. Baldassi, F. Pittorino, R. Zecchina, arXiv:1905.07833, 2019

Using real replicas to study the local entropy

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

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

Local free entropy

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

interaction

$$Z\left(y,\beta,\gamma\right)=\sum_{\{\tilde{W}\}}e^{y\ \Phi\left(\tilde{W};\beta,\gamma\right)}$$

$$=\sum_{\{\tilde{W}\}}\sum_{\{W^a\}}e^{-\beta\sum_{a=1}^{y}E(W^a)-\gamma\sum_{a=1}^{y}d\left(W^a,\tilde{W}\right)}$$

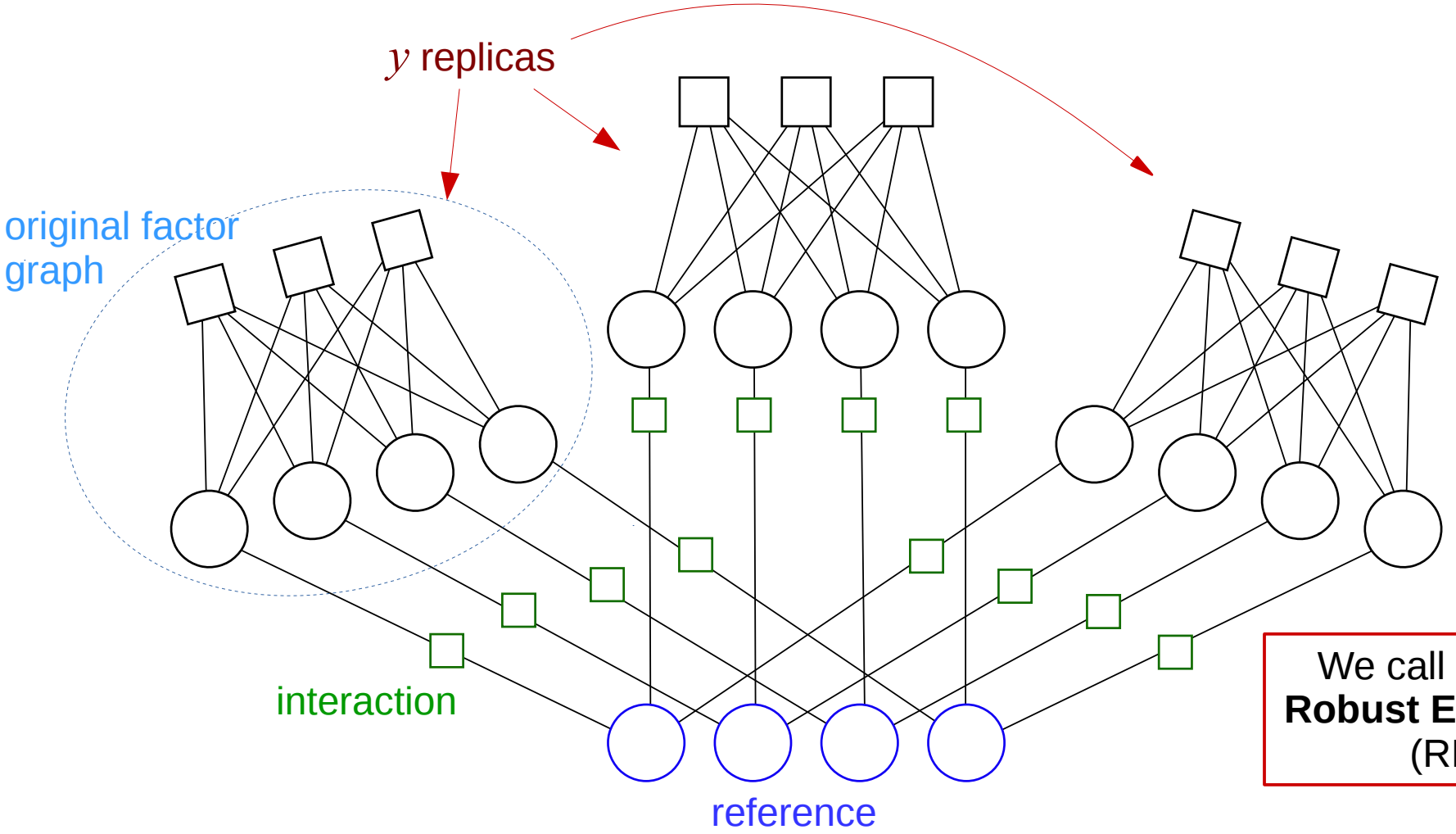
reference
(“center”)

y replicas

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

Using real replicas to study the local entropy

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

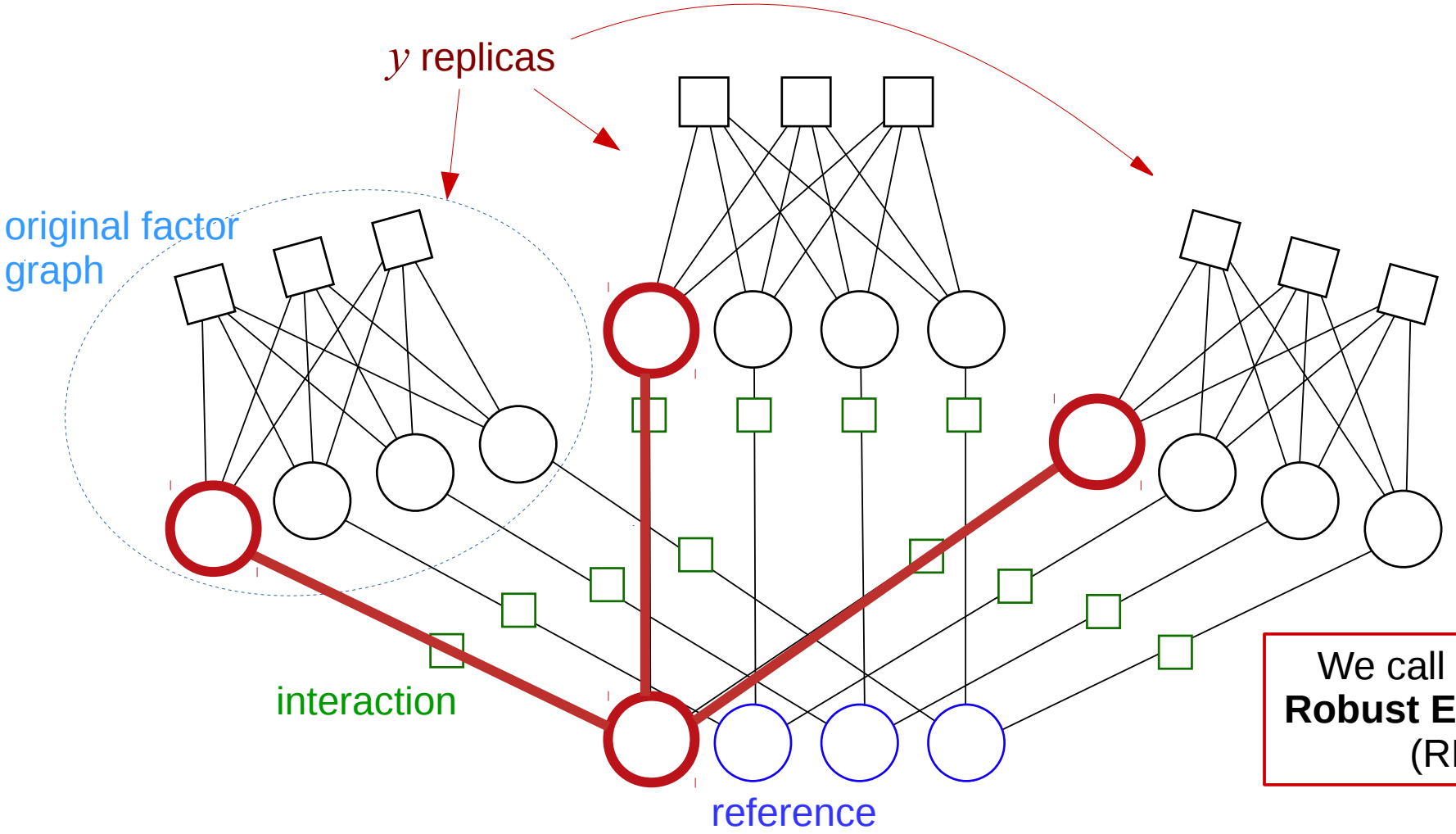


We call this the **Robust Ensemble** (RE)

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

Using real replicas to study the local entropy

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

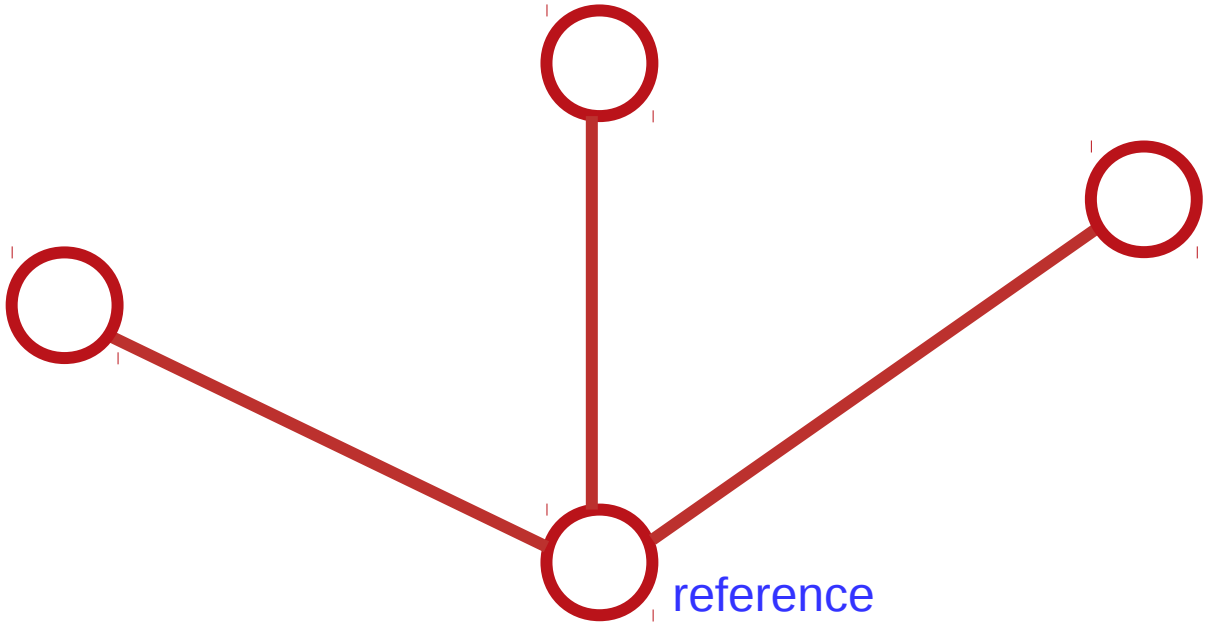


We call this the **Robust Ensemble** (RE)

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

Using real replicas to study the local entropy

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



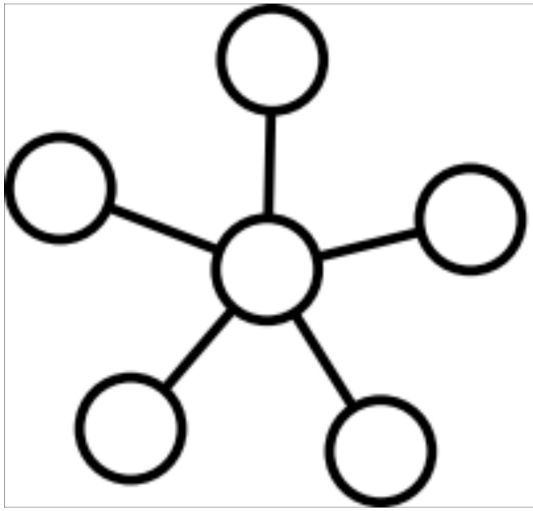
We call this the **Robust Ensemble** (RE)

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

Using real replicas to study the local entropy

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

RE



QMC

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

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

From the 1-RSB formalism to the large-deviations measure

- Start from the ergodicity-breaking scheme of Monasson, PRL (1995)

stuck in a metastable state. We can nevertheless choose a possible direction, given by another field $\sigma(x)$, and compute the free energy of our system when it is weakly pinned by this external quenched field

$$F_\phi[\sigma, g, \beta] = -\frac{1}{\beta} \log \int d\phi(x) \exp\left\{-\beta H[\phi] - \frac{g}{2} \int dx[\sigma(x) - \phi(x)]^2\right\}, \quad (2)$$

where $g > 0$ denotes the strength of the coupling. This free energy (2) will be small when the external perturbing field $\sigma(x)$ lies in a direction corresponding to the bottom of a well of the unperturbed free energy (1). Therefore we should be able to obtain useful information about the free-energy landscape by scanning the entire space of the configurations $\sigma(x)$ to locate all the states in which the system can freeze after spontaneous ergodicity breaking ($g \to 0$) [11]. According to this intuitive idea, we now consider the field $\sigma(x)$ as a thermalized variable with the "Hamiltonian" $F_\phi[\sigma, g, \beta]$. The free energy of the field σ at inverse temperature βm where m is a positive free parameter therefore reads

$$F_\sigma(m, \beta) = \lim_{g \to 0^+} \left(-\frac{1}{\beta m} \log \int d\sigma(x) \exp\{-\beta m F_\phi[\sigma, g, \beta]\} \right). \quad (3)$$

When the ratio m between the two temperatures is an integer, one can easily integrate $\sigma(x)$ in Eq. (3) after having introduced m copies $\phi^\rho(x)$ ($\rho = 1, \ldots, m$) of the original field to obtain the relation

$$F_\sigma(m, \beta) = \lim_{g \to 0^+} \left(-\frac{1}{\beta m} \log \int \prod_{\rho=1}^{m} d\phi^\rho(x) \exp\left\{-\beta \sum_\rho H[\phi^\rho] - \frac{g}{2m} \sum_{\rho<\lambda} \int dx[\phi^\rho(x) - \phi^\lambda(x)]^2\right\} \right) \quad (4)$$

up to an irrelevant additive constant. The meaning of relation (4) is clear. We have to study the statics of m interacting systems, with an attractive coupling between themselves. An alternative formulation of the intuitive

the field σ by introducing n replicas and averaging over the quenched disorder, we obtain from Eq. (4) the same free-energy functional $\mathcal{F}_\phi$ where the number of replicas $\phi^{a\rho}(x)$ now equals $n \times m$ and with an additional term

R. Monasson, PRL, 1995

From the 1-RSB formalism to the large-deviations measure

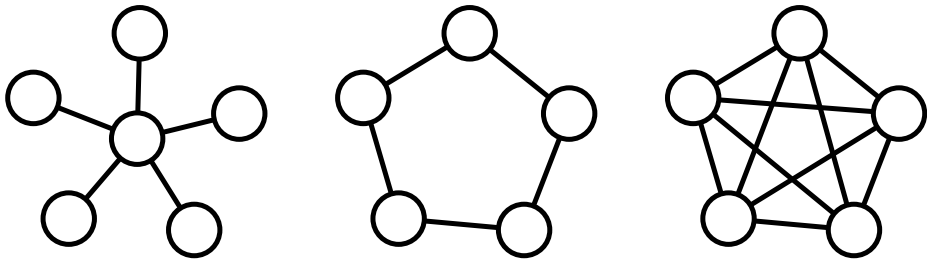
- Start from the ergodicity-breaking scheme of Monasson, PRL (1995)
  - Our replicas however are **real, not virtual**
  - **Don't send the field g to zero**; instead, fix it
  - **Send the number of real replicas m to infinity** (or a large number, e.g. 5 or 10…)
- Also: the 1-RSB cavity method equations (Mézard and Montanari, Oxford Univ. Press 2009, ch. 19) can be used by slightly adapting them
  - Binary weights → propagate only two quantities per message, reflecting the distribution of the magnetizations
  - Doesn't work well as a solver though (symmetry-breaking effects). At least at zero temperature.

R. Monasson, PRL, 1995
M. Mézard, A. Montanari, Oxford Univ. Press, 2009
C. Baldassi, F. Pittorino, R. Zecchina, arXiv:1905.07833, 2019

From the 1-RSB formalism to the large-deviations measure

- In practice (general case):
  - Compute the 1-RSB action (or look it up in the literature...)
  - Express the overlap q_1 in terms of a distance between replicas D (if necessary, change variables)
  - Solve the usual saddle point equations, except for the one for q_1
  - Deduce the scalings for $y \to \infty$ and take the limit [note this surely neglects RSB effects...]

RE QMC 1-RSB-like



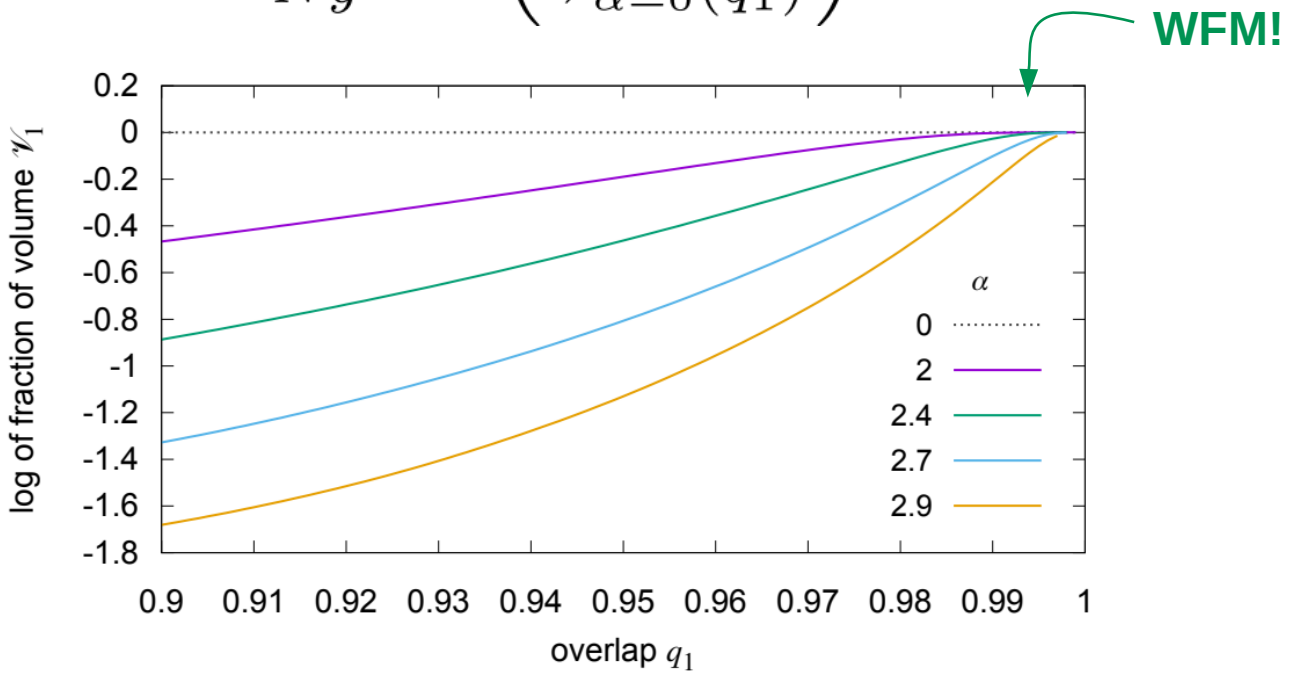
C. Baldassi, F. Pittorino, R. Zecchina, arXiv:1905.07833, 2019

WFM in continuous committee machines

- Study the normalized local volume as a function of the distance
  - Start from the 1-RSB solution, fix q_1, take $y \to \infty$, at $\beta=0$ (we're neglecting further RSB effects here). Compute

$$\frac{1}{Ny} \log \left(\frac{\mathcal{V}_\alpha(q_1)}{\mathcal{V}_{\alpha=0}(q_1)} \right)$$

$K=3$

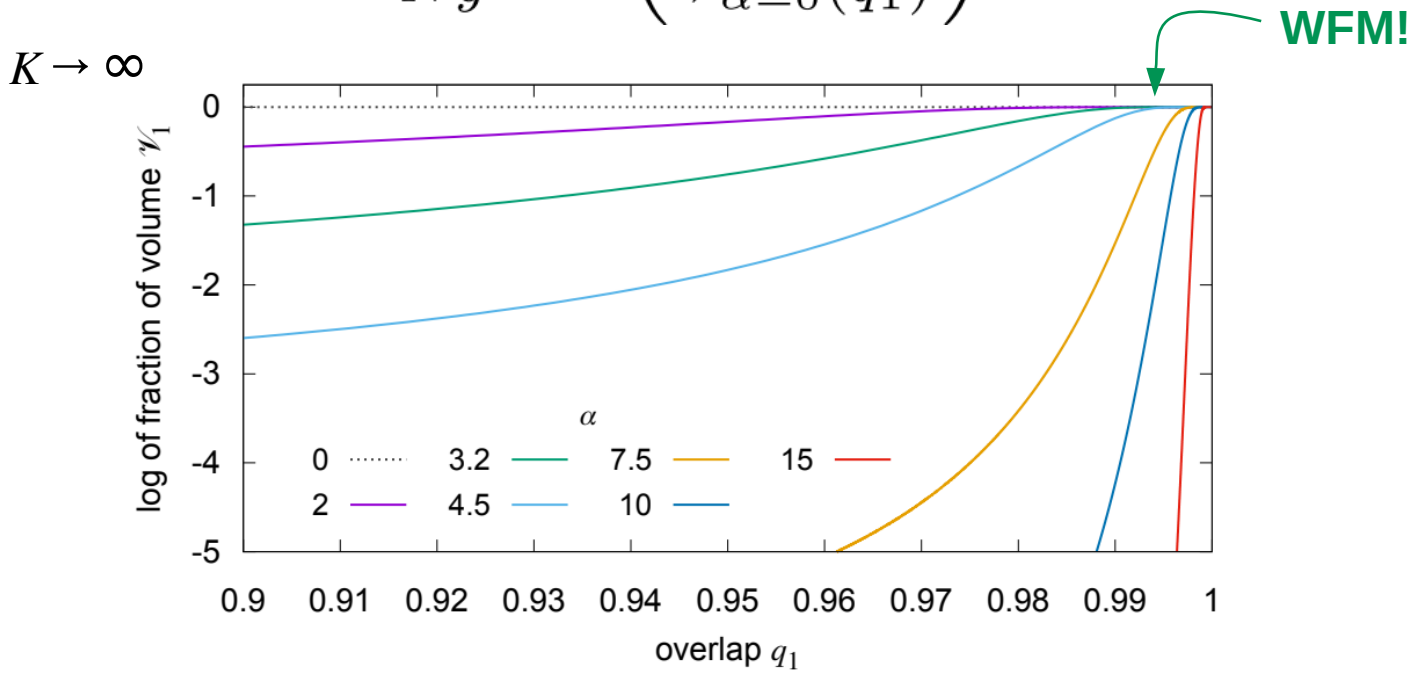


E. Barkai, D. Hansel, H. Sompolinsky, PRA, 1992
C. Baldassi, F. Pittorino, R. Zecchina, arXiv:1905.07833, 2019

WFM in continuous committee machines

- Study the normalized local volume as a function of the distance
  - Start from the 1-RSB solution, fix q_1, take $y \to \infty$, at $\beta=0$ (we're neglecting further RSB effects here). Compute

$$\frac{1}{Ny} \log \left(\frac{\mathcal{V}_\alpha(q_1)}{\mathcal{V}_{\alpha=0}(q_1)} \right)$$



E. Barkai, D. Hansel, H. Sompolinsky, PRA, 1992
C. Baldassi, F. Pittorino, R. Zecchina, arXiv:1905.07833, 2019

Parity machines do not have WFM

- Replace the final output $\sigma_{\rm out} = {\rm sgn}\left(\sum_{\ell=1}^{K} \tau_\ell\right)$ with $\sigma_{\rm out} = \prod_{\ell=1}^{K} \tau_\ell$
  - These devices are intimately related to error-correcting codes. One does **not** want WFM when correcting codes!
- Repeat the previous computation: the curves are flat at $-\alpha \log 2$
- **Claim: The existence of WFM is a structural property of a model**
- Open question: **How to characterize this better? Are NNs special?**

C. Baldassi, F. Pittorino, R. Zecchina, arXiv:1905.07833, 2019

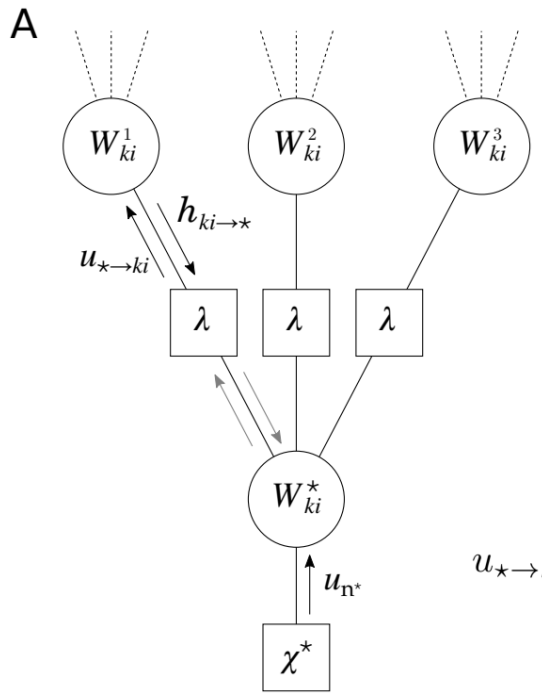
BP on the continuous committee machine

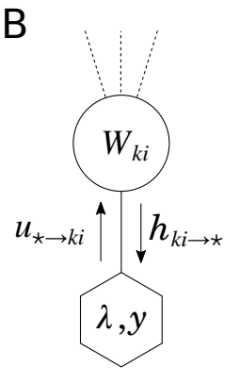
- **Messages are distributions**
- The tricky part is that the "factor→variable" messages are not Gaussian and not normalizable
- Yet, we **only need to propagate two quantities per message** (encoded as m/v and $1/\mathrm{v}$, such that message composition = addition) to close the equations in the large N limit
- **Control the norms** with a self-adjusting external field (one quantity, the same for all nodes)
- Adding an **extra field: explore a region around a given configuration** (weight-enumerator function)

C. Baldassi, F. Pittorino, R. Zecchina, arXiv:1905.07833, 2019

Focusing-BP (fBP): find WFM

- Write **BP on the replicated (RE) factor graph**; **assume symmetry**
- Effect: a "pseudo-factor-node" → self-reinforcement term





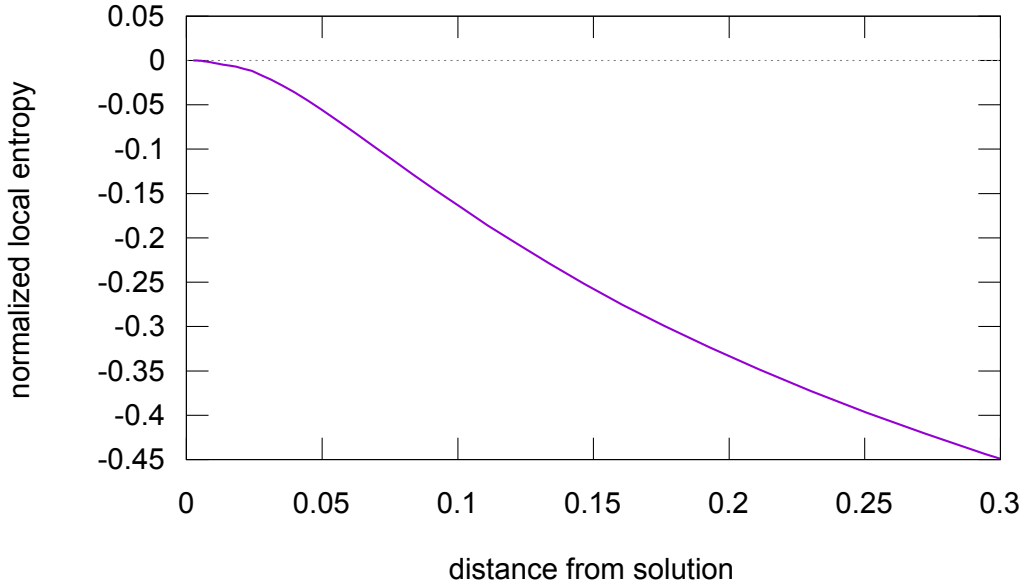
$$u_{\star\to ki} = (((h_{ki\to\star} * \lambda) \uparrow (y-1)) \otimes u_{n^\star}) * \lambda$$

- (Joining variables, taking the magnetization → alternative derivation of a pseudo-1-RSB scheme)

C. Baldassi, C. Borgs, J. Chayes, A. Ingrosso, C. Lucibello, L. Saglietti, R. Zecchina, PNAS, 2016
C. Baldassi, F. Pittorino, R. Zecchina, arXiv:1905.07833, 2019

Focusing-BP (fBP): find WFM - results

- Highest capacity of all the algorithms tested (well within the 1-RSB phase), widest minima
- Only works for uncorrelated random patterns though (for now). Also, tree-like architectures are much easier (breaking the permutation symmetry is possible but fiddly).



C. Baldassi, F. Pittorino, R. Zecchina, arXiv:1905.07833, 2019

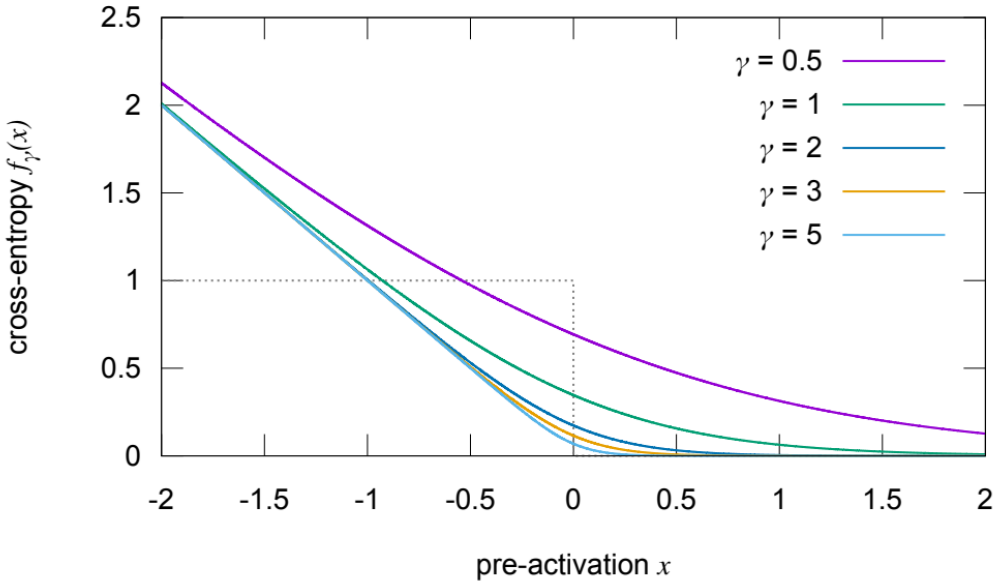
The role of the Cross Entropy loss

- Replica study of the CE loss landscape on the binary perceptron case

$$f_\gamma(x) = -\frac{x}{2} + \frac{1}{2\gamma}\log\left(2\cosh\left(\gamma x\right)\right)$$

- **Usually $\gamma{=}1$ → implicitly set by the norm. We control it explicitly**

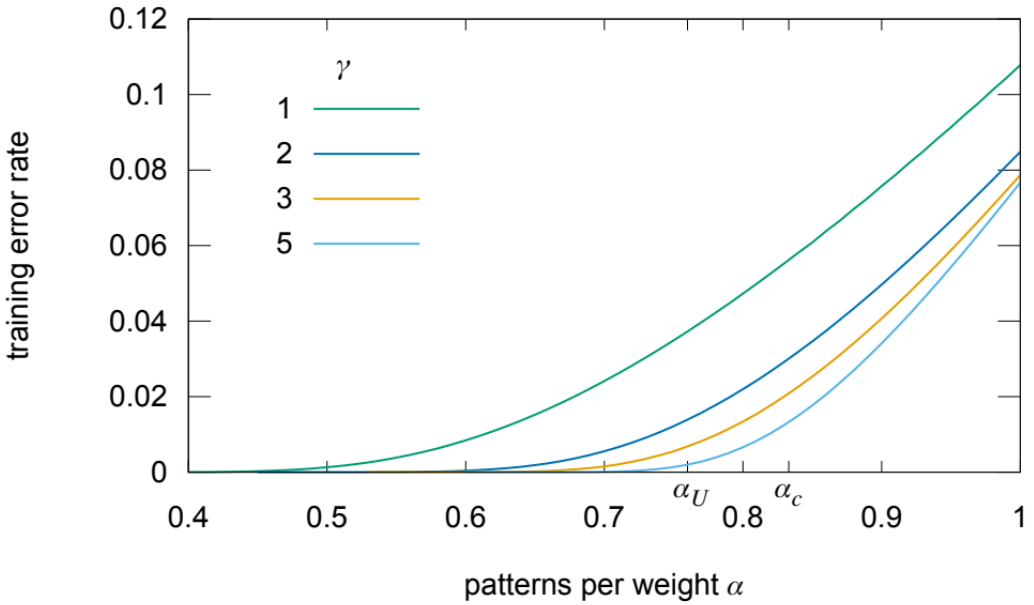
- Small γ → robustness effect
- Large γ → tends to $\mathrm{ReLU}(\text{-}x)$

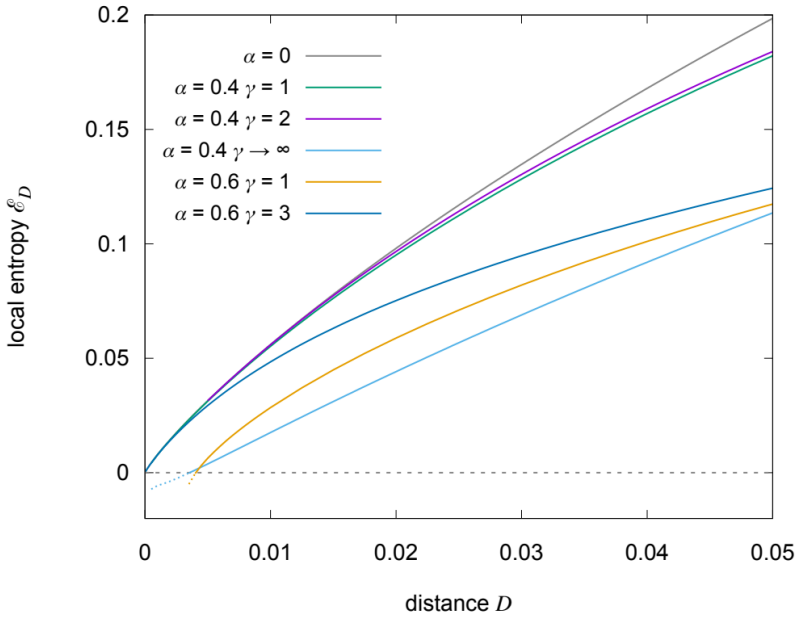


C. Baldassi, F. Pittorino, R. Zecchina, arXiv:1905.07833, 2019

The role of the Cross Entropy loss

- Replica analysis: **The minima of the CE loss end up in the middle of HLE regions as long as the norm is within a certain range**
- Checked numerically with Monte Carlo simulations
- For deeper networks, the effect is limited to the last layer (unless ReLUs are used, without batch-norm...)

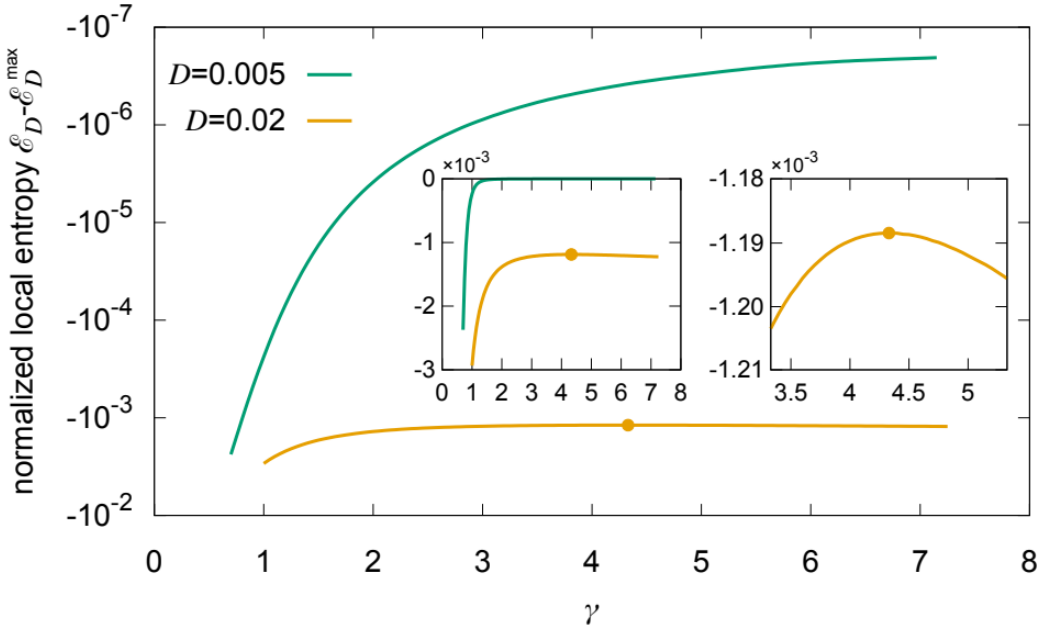




C. Baldassi, F. Pittorino, R. Zecchina, arXiv:1905.07833, 2019

The role of the Cross Entropy loss

- **There's an optimal gamma**. The "good" range is rather wide though.
- Physical significance possibly tied to the detailed geometry of the HLE region (not very clear so far… fractal?)
- Relevance for algorithms possibly mild, but also unclear



C. Baldassi, F. Pittorino, R. Zecchina, arXiv:1905.07833, 2019

Robust greedy algorithms: From LAL to entropic-LAL (eLAL)

- Least Action Learning (**LAL**): extension of the perceptron algorithm to the committee machine
  - apply perceptron rule to the most easily fixable hidden unit
  - very **greedy**
  - very **fast**, high capacity
  - **not gradient-based**, doesn't require differentiability
  - ends up in **narrow minima** (estimated with BP)
- Add an **entropic component (eLAL)**
  - replicated system (RE-like)
  - interaction between replicas
  - **still fast, much wider minima**
- Unfortunately **unclear how to extend it to deeper networks**

G. Mitchison, R. Durbin, Biol. Cybernetics, 1989
C. Baldassi, F. Pittorino, R. Zecchina, arXiv:1905.07833, 2019

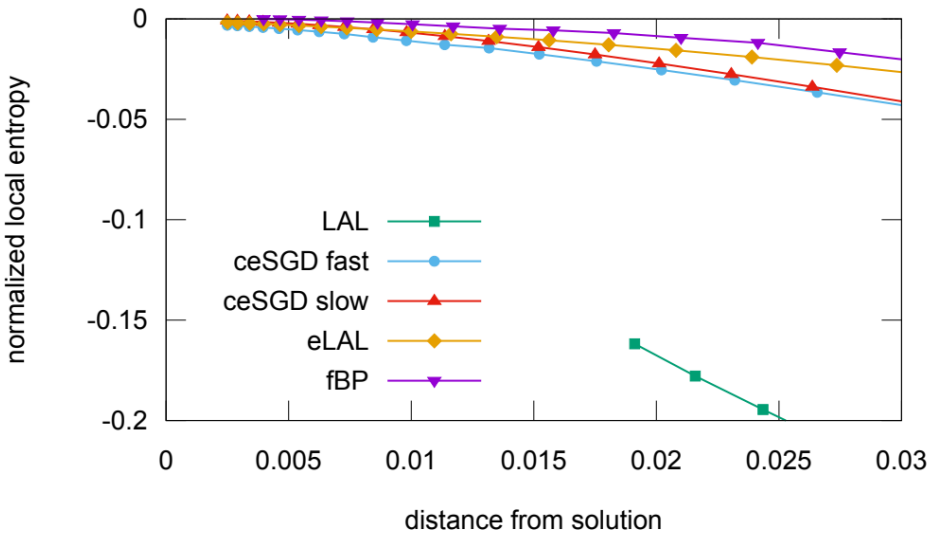
Numerical experiments: random patterns

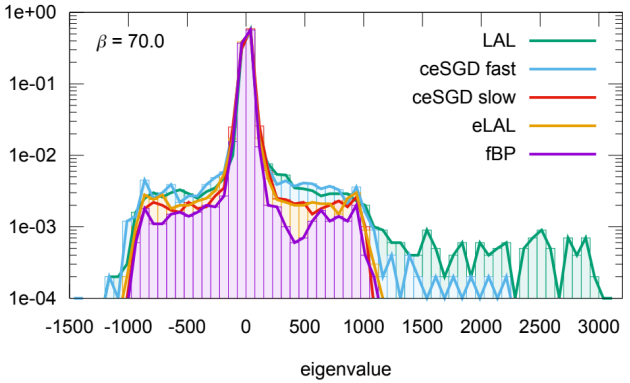
- Set up: tree committee machine with $K=9$, $\alpha=1$ as close as possible to the theory
- 5 algorithms
  - **fBP**
  - **LAL**
  - **eLAL**
  - **ceSGD-fast** [use $\tanh(\beta x)$ activations, polarize β and γ gradually]
  - **ceSGD-slow** [same, grow the norms (β,γ) more slowly]
- Comparison of the local volumes and the spectrum of the Hessians → good correlation

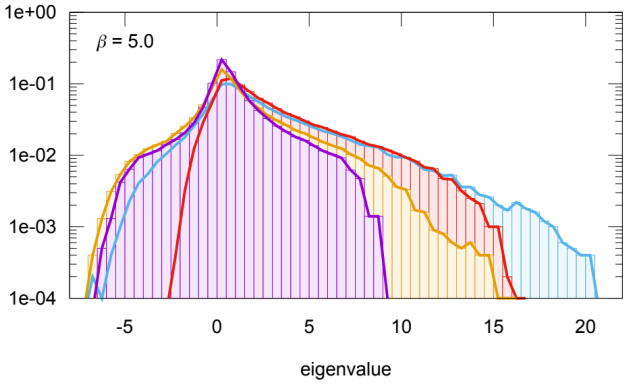
C. Baldassi, F. Pittorino, R. Zecchina, arXiv:1905.07833, 2019

Numerical experiments: random patterns

- Set up: tree committee machine with $K=9$, $\alpha=1$ as close as possible to the theory
- Comparison of the local volumes and the spectrum of the Hessians → good correlation



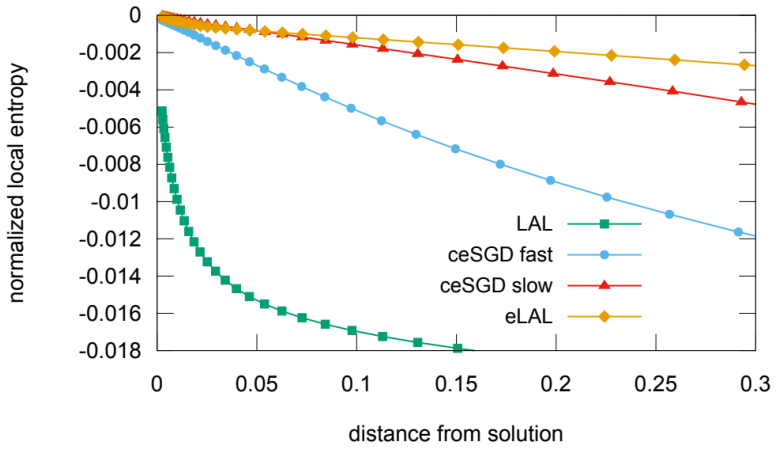


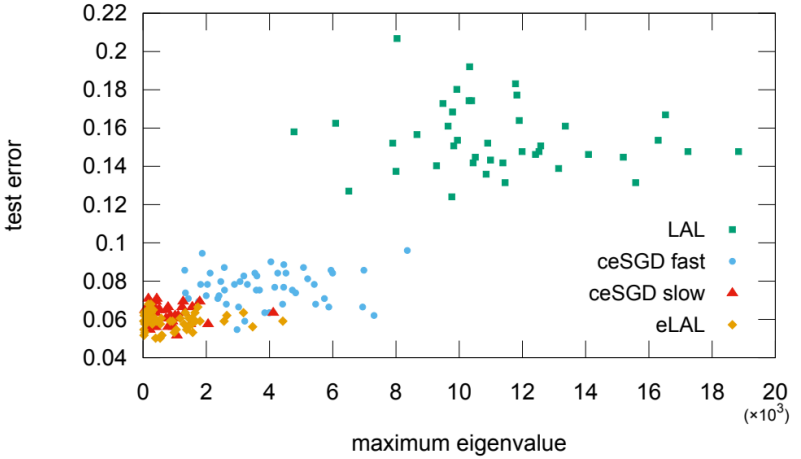


C. Baldassi, F. Pittorino, R. Zecchina, arXiv:1905.07833, 2019

Numerical experiments: reduced binarized Fashion-MNIST

- Set up: fully connected committee machine with $K=9$
- Two (hard but not impossible) classes (*dress* vs *coat*), binarized patterns (250 per class)
  - Patterns are binary and balanced, but biased and correlated
- 4 algorithms (no fBP, correlated patterns)
- Local volumes and Hessians correlate with generalization scores

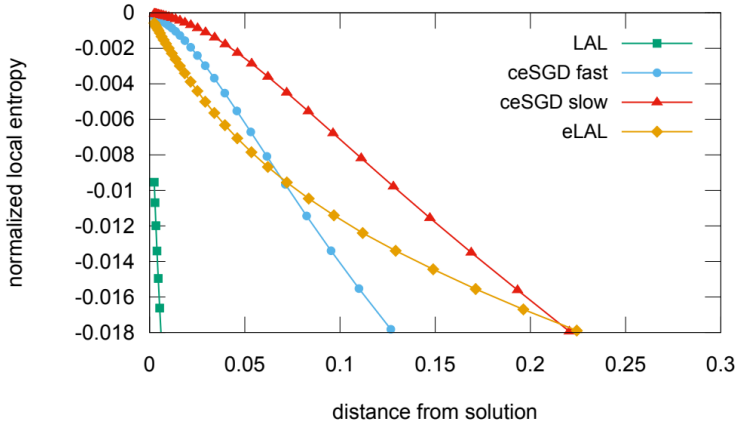


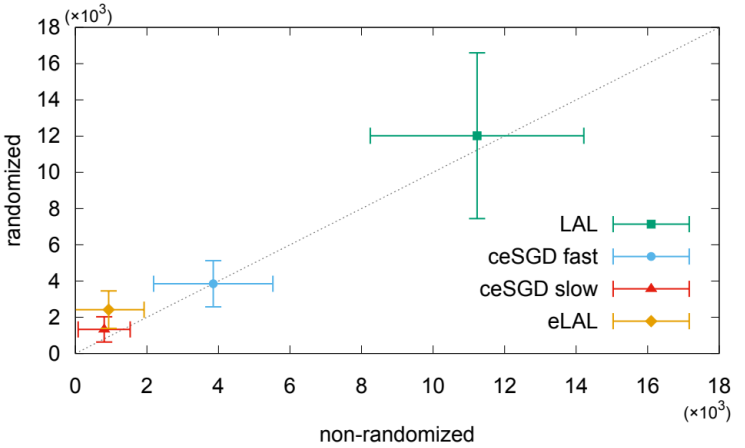


C. Baldassi, F. Pittorino, R. Zecchina, arXiv:1905.07833, 2019

Numerical experiments: randomized Fashion-MNIST

- Shuffle the data across patterns at each pixel location
  - keep biases
  - destroy correlations
- Slightly harder problem
- Results are overall similar to the previous ones
  - eLAL affected more than SGD (perhaps tuning?); still way better than LAL

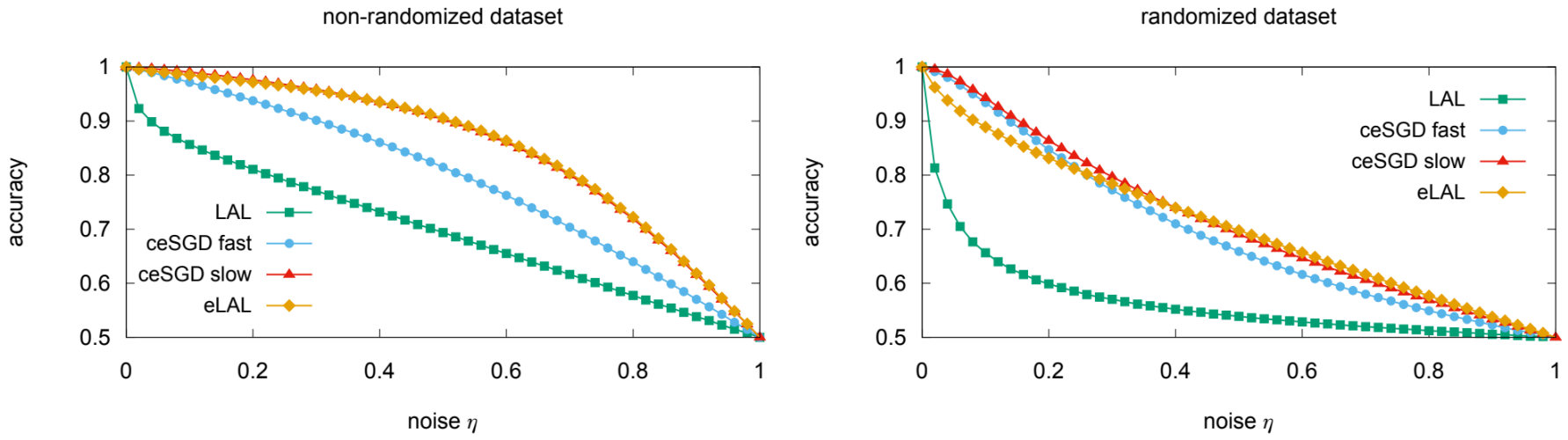




C. Baldassi, F. Pittorino, R. Zecchina, arXiv:1905.07833, 2019

Numerical experiments: robustness to input perturbation

- Tested on Fashion-MNIST and randomized-Fashion-MNIST
- Add a colored (pixel-dependent) noise, measure the training error degradation [attempt to measure a "generalization precursor"]
- Consistent with the other measures (volume, generalization etc.)



C. Baldassi, F. Pittorino, R. Zecchina, arXiv:1905.07833, 2019

Conclusions, open problems, future directions

- Wide Flat Minima seem to be a structural property of NNs
- They seem to generalize better (other results on deeper networks from other groups)
- Some algorithmic techniques help finding WFM
  - Some are well known and widely used (CE loss, ReLU, …)
  - Some aren't (replication [Elastic-SGD], Entropy-SGD, eLAL, ...)
- Role of pattern structure still to be investigated more in detail
- Going deeper: WIP
- Also TODO: Detailed description of the geometrical structure of the HLE/WFM
- Which distance function to use? (unclear esp. for multi-layer networks)