Matrix Factorization with Neural Networks of Associative Memory

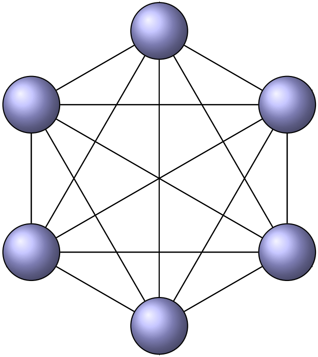
Francesco Camilli
Joint work with M. Mézard

The Abdus Salam International Center for Theoretical Physics

YHD, Trieste, May 30th, 2023

Hopfield network[1]

Hopfield Network is a model for associative memory. The task of the network is the storage of P patterns, $\boldsymbol{\xi}^\mu$ that are recalled if we provide an input sufficiently close to one of them.



The memory effect is given by the interaction between neurons:

$$J_{ij} = \frac{1}{N}\sum_{\mu=1}^{P} \xi_i^\mu \xi_j^\mu \,, \quad \xi_i^\mu = \pm 1$$

that tend to make the patterns ground states for the energy:

$$E(\boldsymbol{J}, \boldsymbol{\sigma}) = -\frac{1}{2}\sum_{i,j=1}^{N} J_{ij}\sigma_i\sigma_j \,, \quad \sigma_i = \pm 1$$

[1] Hopfield, J., PNAS. 79 (8) 2554-2558 (1982)

Retrieval

There are several alternatives (not limited to the following):

- the classical "neural dynamics": $\sigma_i^{t+1} = \mathrm{sign}(h_i^t)\,,\ h_i^t = \sum_{j\neq i} J_{ij}\sigma_j^t$;
- Simulated Annealing on the energy $E(\boldsymbol{J}, \boldsymbol{\sigma})$
- Message passing algorithms (like AMP) on the Boltzmann-Gibbs measure $\langle\cdot\rangle \propto \exp\big(-\beta E(\boldsymbol{J}, \boldsymbol{\sigma})\big)$ at very high β

They all aim at sampling the BG measure at low temperature!

They all work provided a good initialization is provided!

Retrieval

There are several alternatives (not limited to the following):

- the classical "neural dynamics": $\sigma_i^{t+1} = \mathrm{sign}(h_i^t)\,,\ h_i^t = \sum_{j\neq i} J_{ij}\sigma_j^t$;
- Simulated Annealing on the energy $E(\boldsymbol{J}, \boldsymbol{\sigma})$
- Message passing algorithms (like AMP) on the Boltzmann-Gibbs measure $\langle\cdot\rangle \propto \exp\big(-\beta E(\boldsymbol{J}, \boldsymbol{\sigma})\big)$ at very high β

They all aim at sampling the BG measure at low temperature!

They all work provided a good initialization is provided!

Retrieval

There are several alternatives (not limited to the following):

- the classical "neural dynamics": $\sigma_i^{t+1} = \mathrm{sign}(h_i^t)\,,\ h_i^t = \sum_{j\neq i} J_{ij}\sigma_j^t$;
- Simulated Annealing on the energy $E(\boldsymbol{J},\boldsymbol{\sigma})$
- Message passing algorithms (like AMP) on the Boltzmann-Gibbs measure $\langle\cdot\rangle \propto \exp\big(-\beta E(\boldsymbol{J},\boldsymbol{\sigma})\big)$ at very high β

They all aim at sampling the BG measure at low temperature!

They all work provided a good initialization is provided!

Pattern interference

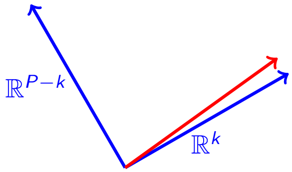
The patterns are typically *i.i.d.* drawn from a centered P_ξ. Which means:

$$\frac{\boldsymbol{\xi}^\mu \cdot \boldsymbol{\xi}^\nu}{N} = \delta_{\mu\nu} + O(N^{-1/2})\,.$$

Pattern interference[2] occurse when $\alpha = P/N > 0$.

Intuitive explanation

$\mathrm{x} \sim \langle \cdot \rangle$ can have a $O(N)$ projection only onto a finite number of patterns. The remaining ones are $O(\sqrt{N})$.



[2] D.J. Amit, H. Gutfreund and H. Sompolinsky, in Phys. Rev. Lett. 55, 1530 (1985)

High rank matrix factorization

$$\boldsymbol{Y} = \boldsymbol{\xi} \cdot \boldsymbol{\xi}^{\intercal} + \sqrt{\Delta}\boldsymbol{Z}$$

The task is to recover an $N \times P$ matrix $\boldsymbol{\xi} = (\xi_i^\mu)_{i \leq N}^{\mu \leq P}$ with $P/N = \alpha \geq 0$ from

$$Y_{ij} = \sum_{\mu=1}^{P} \frac{\xi_i^\mu \xi_j^\mu}{\sqrt{N}} + \sqrt{\Delta} Z_{ij}\,, \quad \boldsymbol{Z} \text{ Wigner}\,.$$

Bayes optimal approach: in the hypothesis of $\xi_i^\mu \overset{\text{iid}}{\sim} P_\xi$

$$\Phi = \frac{1}{NP} \mathbb{E} \log \int \prod_{i=1}^{N} \prod_{\mu=P}^{N} dP_\xi(X_i^\mu) \exp\left[\frac{1}{2\Delta\sqrt{N}} \mathrm{Tr}\, \boldsymbol{Y}\boldsymbol{X}\boldsymbol{X}^T - \frac{\mathrm{Tr}(\boldsymbol{X}\boldsymbol{X}^T)^2}{4\Delta N}\right]$$

High rank matrix factorization

$$\boldsymbol{Y} = \boldsymbol{\xi} \cdot \boldsymbol{\xi}^{\intercal} + \sqrt{\Delta}\boldsymbol{Z}$$

The task is to recover an $N \times P$ matrix $\boldsymbol{\xi} = (\xi_i^\mu)_{i \leq N}^{\mu \leq P}$ with $P/N = \alpha \geq 0$ from

$$Y_{ij} = \sum_{\mu=1}^{P} \frac{\xi_i^\mu \xi_j^\mu}{\sqrt{N}} + \sqrt{\Delta} Z_{ij}\,, \quad \boldsymbol{Z} \text{ Wigner}\,.$$

Bayes optimal approach: in the hypothesis of $\xi_i^\mu \overset{\text{iid}}{\sim} P_\xi$

$$\Phi = \frac{1}{NP}\mathbb{E}\log\int\prod_{i=1}^{N}\prod_{\mu=P}^{N} dP_\xi(X_i^\mu)\exp\left[\frac{1}{2\Delta\sqrt{N}}\mathrm{Tr}\,\boldsymbol{Y}\boldsymbol{X}\boldsymbol{X}^T - \frac{\mathrm{Tr}(\boldsymbol{X}\boldsymbol{X}^T)^2}{4\Delta N}\right]$$

Why is it interesting?

In its asymmetric version $\boldsymbol{Y} = \boldsymbol{AB} + \sqrt{\Delta}\boldsymbol{Z}$ matrix factorization is employed for:

- image and video restoration;
- image and video inpainting;

where usually one imposes $\boldsymbol{B}$ is sparse and the columns of $\boldsymbol{A}$ form an overcomplete basis.

We also have

- Recommendation systems
- A high rank version of spiked models!

When P is finite, the model reduces to the low rank matrix estimation, which is a well studied problem in the Bayes-optimal setting[3].

[3] Among other refs: T. Lesieur, F. Krzakala and L. Zdeborová, doi:10.1109/ALLERTON.2015.7447070.

Why is it interesting?

In its asymmetric version $\boldsymbol{Y} = \boldsymbol{AB} + \sqrt{\Delta}\boldsymbol{Z}$ matrix factorization is employed for:

- image and video restoration;
- image and video inpainting;

where usually one imposes $\boldsymbol{B}$ is sparse and the columns of $\boldsymbol{A}$ form an overcomplete basis.

We also have

- Recommendation systems
- A high rank version of spiked models!

When P is finite, the model reduces to the low rank matrix estimation, which is a well studied problem in the Bayes-optimal setting[3].

[3] Among other refs: T. Lesieur, F. Krzakala and L. Zdeborová, doi:10.1109/ALLERTON.2015.7447070.

A related problem: matrix denoising and the RIE[4]

Task: reconstruct a symmetric matrix $\boldsymbol{S}$ given the observations:

$$\boldsymbol{Y} = \boldsymbol{S} + \sqrt{\Delta}\boldsymbol{Z}\,, \quad \in \mathbb{R}^{N\times N}\,,$$

with $O(1)$ eigenvalues.

Cleaning procedure

$$\hat{\boldsymbol{\lambda}}_S = \boldsymbol{\lambda}_Y - 2\Delta\mathcal{H}[\rho_Y](\boldsymbol{\lambda}_Y)\,, \quad \hat{\boldsymbol{S}} = \boldsymbol{O}\hat{\boldsymbol{\lambda}}_S\boldsymbol{O}^T$$

$\rho_Y =$ spectral density of $\boldsymbol{Y}$, $\mathcal{H} =$ Hilbert transform.

In our case $\boldsymbol{S}$ is $\frac{\boldsymbol{\xi}\boldsymbol{\xi}^\intercal}{N}$ and we will measure its performance via the matrix MSE (mMSE):

$$\text{mMSE} := \frac{1}{2N}\left\|\hat{\boldsymbol{S}} - \frac{\boldsymbol{\xi}\boldsymbol{\xi}^\intercal}{N}\right\|^2$$

[4] J. Bun, R. Allez, J.-P. Bouchaud, and M. Potters, IEEE Transactions on Information Theory, 62(12):7475–7490, 2016.

Decimation: a sub-optimal yet feasible approach

We look for one column $\boldsymbol{\xi}^\mu$ of $\boldsymbol{\xi}$ at a time.

Decimation scheme:

1. Assume we are able to produce an estimate of $\boldsymbol{\xi}^P$, denoted $\boldsymbol{\eta}^P$ sampling from a certain measure;
2. Subtract the corresponding rank 1 contribution from $\boldsymbol{Y}$:
$$\boldsymbol{Y}_1 = \boldsymbol{Y} - \frac{\boldsymbol{\eta}^P \boldsymbol{\eta}^{P\,T}}{\sqrt{N}}$$
3. Replace $\boldsymbol{Y}_0 \equiv \boldsymbol{Y}$ with $\boldsymbol{Y}_1$ and produce another estimate for $\boldsymbol{\xi}^{P-1}$;
4. Repeat untill the P-th step when $\boldsymbol{Y}_P = \boldsymbol{Y}_1 - \frac{\boldsymbol{\eta}^1 \boldsymbol{\eta}^{1\,T}}{\sqrt{N}} = \boldsymbol{Y} - \sum_{\mu=1}^{P} \frac{\boldsymbol{\eta}^\mu \boldsymbol{\eta}^{\mu\,T}}{\sqrt{N}}$.

Decimation: a sub-optimal yet feasible approach

We look for one column $\boldsymbol{\xi}^\mu$ of $\boldsymbol{\xi}$ at a time.

Decimation scheme:

1. Assume we are able to produce an estimate of $\boldsymbol{\xi}^P$, denoted $\boldsymbol{\eta}^P$ sampling from a certain measure;
2. Subtract the corresponding rank 1 contribution from $\boldsymbol{Y}$:
$$\boldsymbol{Y}_1 = \boldsymbol{Y} - \frac{\boldsymbol{\eta}^P \boldsymbol{\eta}^{P\,T}}{\sqrt{N}}$$
3. Replace $\boldsymbol{Y}_0 \equiv \boldsymbol{Y}$ with $\boldsymbol{Y}_1$ and produce another estimate for $\boldsymbol{\xi}^{P-1}$;
4. Repeat untill the P-th step when $\boldsymbol{Y}_P = \boldsymbol{Y}_1 - \frac{\boldsymbol{\eta}^1 \boldsymbol{\eta}^{1\,T}}{\sqrt{N}} = \boldsymbol{Y} - \sum_{\mu=1}^P \frac{\boldsymbol{\eta}^\mu \boldsymbol{\eta}^{\mu\,T}}{\sqrt{N}}$.

R-th decimation step

Let $R = 0, 1, 2, \dots, P-1$ be the number of patterns already estimated. Define

$$\boldsymbol{Y}_R = \boldsymbol{Y} - \sum_{\mu=P-R+1}^{P} \frac{\boldsymbol{\eta}^\mu \boldsymbol{\eta}^{\mu\intercal}}{\sqrt{N}} .$$

We assume the $R+1$-th estimate is sampled from $\frac{1}{\mathcal{Z}_R} e^{-\beta E(\boldsymbol{x}|\boldsymbol{Y}_R)} dP_\xi(\boldsymbol{x})$ with

$$-E(\boldsymbol{x}|\boldsymbol{Y}_R) = \frac{1}{2\sqrt{N}} \mathrm{Tr}\boldsymbol{Y}_R \boldsymbol{x}\boldsymbol{x}^\intercal - \frac{\|\boldsymbol{x}\|^4}{4N} =$$

$$= \frac{\sqrt{\Delta}}{2\sqrt{N}} \sum_{i,j=1}^{N} Z_{ij} x_i x_j + \frac{1}{2N} \sum_{\mu=1}^{P} \left(\sum_{i=1}^{N} \xi_i^\mu x_i \right)^2 - \frac{1}{2N} \sum_{\mu=P-R+1}^{P} \left(\sum_{i=1}^{N} \eta_i^\mu x_i \right)^2 - \frac{\|\boldsymbol{x}\|^4}{4N} .$$

- The second set of terms is the energy of the Hopfield model!
- The third set of terms repels from already found patterns.

R-th decimation step

Let $R = 0, 1, 2, \dots, P-1$ be the number of patterns already estimated. Define

$$\boldsymbol{Y}_R = \boldsymbol{Y} - \sum_{\mu=P-R+1}^{P} \frac{\boldsymbol{\eta}^\mu \boldsymbol{\eta}^{\mu\intercal}}{\sqrt{N}}\,.$$

We assume the $R+1$-th estimate is sampled from $\frac{1}{\mathcal{Z}_R} e^{-\beta E(\boldsymbol{x}|\boldsymbol{Y}_R)} dP_\xi(\boldsymbol{x})$ with

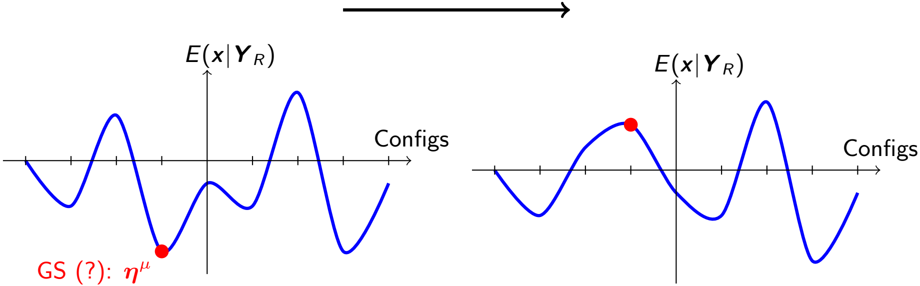
$$-E(\boldsymbol{x}|\boldsymbol{Y}_R) = \frac{1}{2\sqrt{N}} \mathrm{Tr}\, \boldsymbol{Y}_R \boldsymbol{x}\boldsymbol{x}^\intercal - \frac{\|\boldsymbol{x}\|^4}{4N} =$$

$$= \frac{\sqrt{\Delta}}{2\sqrt{N}} \sum_{i,j=1}^{N} Z_{ij} x_i x_j + \frac{1}{2N} \sum_{\mu=1}^{P} \left(\sum_{i=1}^{N} \xi_i^\mu x_i \right)^2 - \frac{1}{2N} \sum_{\mu=P-R+1}^{P} \left(\sum_{i=1}^{N} \eta_i^\mu x_i \right)^2 - \frac{\|\boldsymbol{x}\|^4}{4N}\,.$$

- The second set of terms is the energy of the Hopfield model!
- The third set of terms repels from already found patterns.

Acting on the energy landscape

$$\boldsymbol{Y} \mapsto \boldsymbol{Y} - \frac{\boldsymbol{\eta}^{\mu}\boldsymbol{\eta}^{\mu\intercal}}{\sqrt{N}}$$



Noise sources

It is important to notice that decimation is affected by three noise sources:

a) The initial Gaussian noise $\boldsymbol{Z}$;

b) Patter interference, tuned by $\alpha = P/N$, rank over dimensionality, due to the similarity with Hopfield;

c) Decimation itself! Indeed, the $\boldsymbol{\eta}$'s are only estimates of the patterns.

The ultimate goal of decimation is **to decrease the effective rank of the hidden matrix**, so to tune down noise source b).

Does decimation corrupt itself?

Noise sources

It is important to notice that decimation is affected by three noise sources:

a) The initial Gaussian noise $\boldsymbol{Z}$;

b) Patter interference, tuned by $\alpha = P/N$, rank over dimensionality, due to the similarity with Hopfield;

c) Decimation itself! Indeed, the $\boldsymbol{\eta}$'s are only estimates of the patterns.

The ultimate goal of decimation is **to decrease the effective rank of the hidden matrix**, so to tune down noise source b).

Does decimation corrupt itself?

Decimation free entropies

Each decimation step, say $R+1$, has its own free entropy:

$$\Phi_{R+1} := \frac{1}{N}\mathbb{E}\log\int dP_\xi(\boldsymbol{x})\exp\Big(-\beta E(\boldsymbol{x}|\boldsymbol{Y}_R)\Big)$$

whose limit is computed with the *replica method*. The replica symmetric ansatz yields

RS free entropy

$$\Phi_{R+1} = \mathrm{Extr}\Big\{\Phi_0(\alpha, m^{[0,t]}, \beta; q, r, u) - \beta\frac{m^2}{2} + \\ + \mathbb{E}_{Z,\xi}\log\int dP_\xi(x)\exp\Big(\big(Z\sqrt{r}+\beta m\xi\big)x - \frac{u+r}{2}x^2\Big)\Big\}$$

- $m^{[0,t]}$, $t = R/P$ collection of the previous retrieval accuracies $m^\mu = \frac{\boldsymbol{\xi}^\mu\cdot\boldsymbol{\eta}^\mu}{N}$;
- m: at stationarity is the $R+1$-th retrieval accuracy;
- r: multiplies a std Gaussian, tunes amplitude of the noise.

Decimation free entropies

Each decimation step, say $R+1$, has its own free entropy:

$$\Phi_{R+1} := \frac{1}{N}\mathbb{E}\log\int dP_\xi(\boldsymbol{x})\exp\Big(-\beta E(\boldsymbol{x}|\boldsymbol{Y}_R)\Big)$$

whose limit is computed with the *replica method*. The replica symmetric ansatz yields

RS free entropy

$$\Phi_{R+1} = \mathrm{Extr}\Big\{\Phi_0(\alpha, m^{[0,t]}, \beta; q, r, u) - \beta\frac{m^2}{2} + \\ + \mathbb{E}_{Z,\xi}\log\int dP_\xi(x)\exp\Big(\big(Z\sqrt{r}+\beta m\xi\big)x - \frac{u+r}{2}x^2\Big)\Big\}$$

- $m^{[0,t]}$, $t = R/P$ collection of the previous retrieval accuracies $m^\mu = \frac{\boldsymbol{\xi}^\mu\cdot\boldsymbol{\eta}^\mu}{N}$;
- m: at stationarity is the $R+1$-th retrieval accuracy;
- r: multiplies a std Gaussian, tunes amplitude of the noise.

(Some) saddle point equations

Recall

$$\langle \cdot \rangle_{\xi_i^\mu, Z} = \frac{\int dP_\xi(x) e^{(Z\sqrt{r}+\beta m^\mu \xi_i^\mu)x - \frac{r+u}{2}x^2} (\cdot)}{\int dP_\xi(x) e^{(Z\sqrt{r}+\beta m^\mu \xi_i^\mu)x - \frac{r+u}{2}x^2}}$$

Saddle point equations

(r) $r = r_a + r_b + r_c$:

$$r_a = \beta^2 \Delta q \,, \quad r_b = \frac{(1-t)\alpha\beta^2 q}{(1-\beta(1-q))^2} \,,$$

$$r_c = 2\alpha t \beta^2 q \int_0^t d\tau \, f(\tau)[1 + 2\beta^2(1-q)^2 f(\tau)] \,.$$

(m) $m = \mathbb{E}_{\xi,Z} \xi \langle X \rangle_{\xi,Z}$

(q) $q = \mathbb{E}_{\xi,Z} \langle X \rangle^2_{\xi,Z}$.

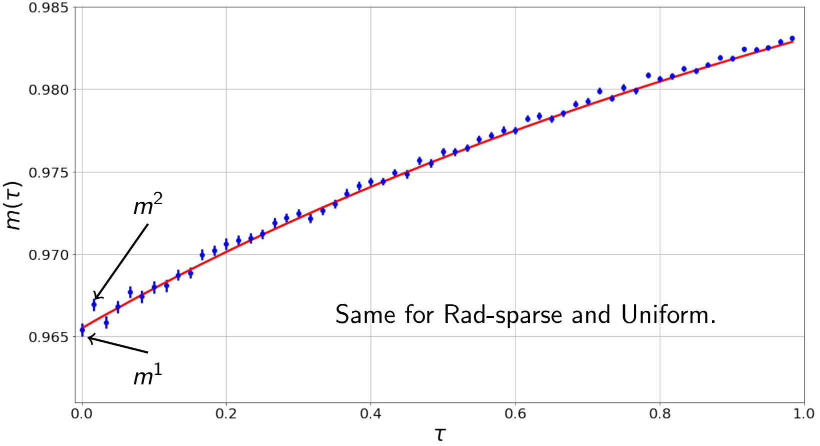
Main questions

Q1 Does decimation corrupt itself (too much)? Would it be able to retrieve all the patterns?

Q2 Is there an efficient way to implement it?

Q1: Decimation is a valid procedure!

Here $\alpha = 0.03$, $\Delta = 1/\beta = 0.15$, $N = 2000$, $P_\xi =$Rad. Blue dots are AMP initialized close to patterns.



Q2: The ground state "oracle" for Ising spins

Made up of three ingredients:

- Simulated annealing on $E(\boldsymbol{\sigma}|\boldsymbol{Y}_R)$: we set the i-th spin up with probability

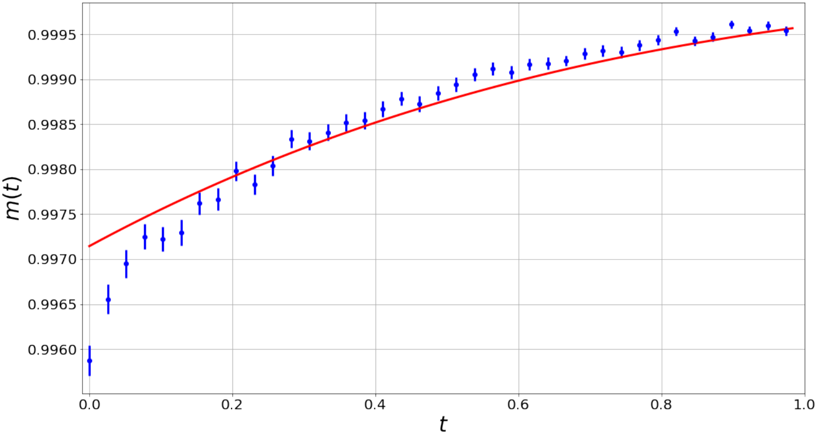
$$\frac{1}{1+\exp\big(-2\beta_k h_{i,k}\big)}\,,\quad \beta_k = 1 + Ck\,, h_{i,k} = (\boldsymbol{Y}_R\boldsymbol{\sigma}_k)_i$$

  $k = 1, 2, \dots$.

- Decimation $\boldsymbol{Y}_R \mapsto \boldsymbol{Y}_R - \frac{\boldsymbol{\eta}^\mu \boldsymbol{\eta}^{\mu\intercal}}{\sqrt{N}}$
- Restarting criterion: due to the ragged landscape, the SA often gets stuck in metastable states. Hence, we compute the energy of the found configuration and test it against the expected energy of a GS, accessible through the theory.

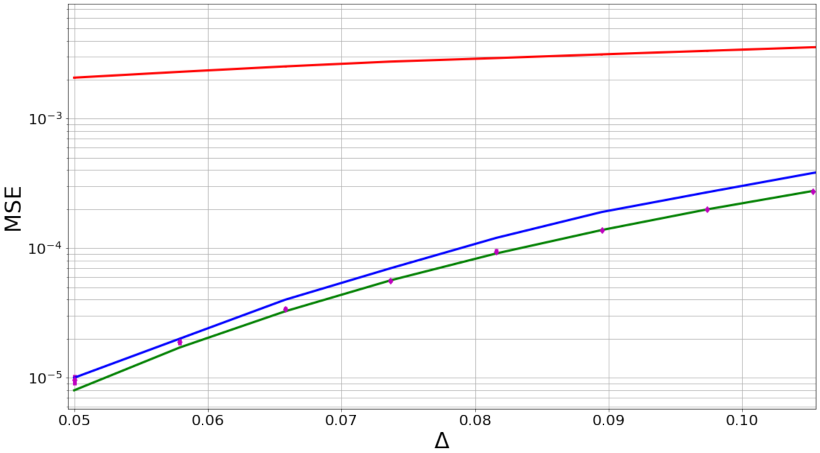
GS oracle

Here $N = 1300$, with P_ξ =Rad, $\alpha = 0.03$, $\Delta = 0.08$, $\beta \to \infty$. **No informative initialization needed!**

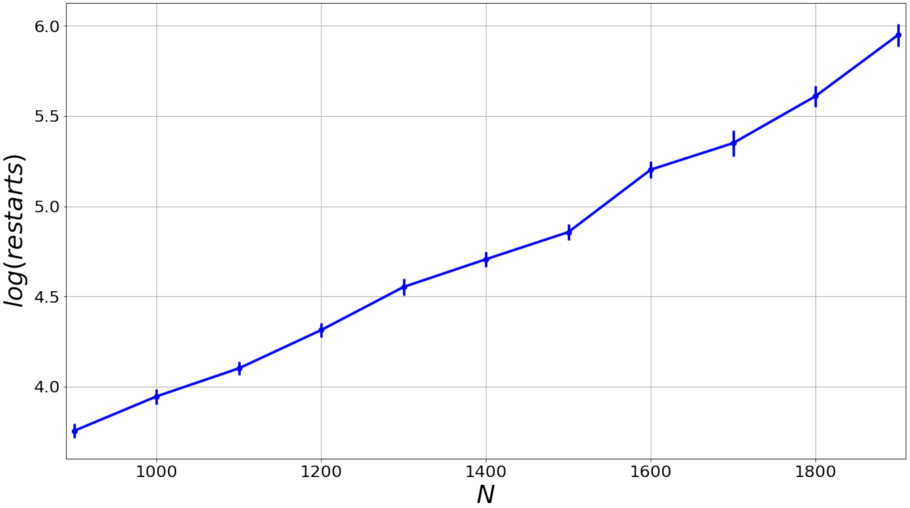


Decimation performance

Red: RIE, blue: decimation at $\beta = 1/\Delta$, green: decimation at $\beta \to \infty$.



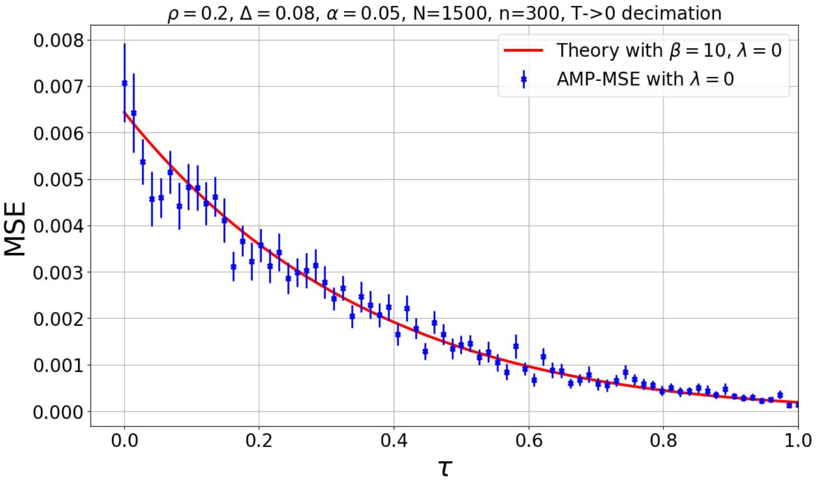
… but not efficient!



Rademacher sparse prior

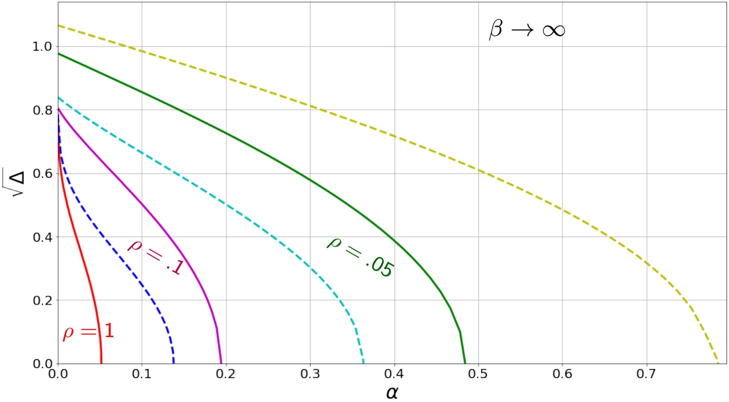
The above results hold for any well behaved prior, including

$$P_\xi = \frac{\rho}{2}(\delta_{-1} + \delta_1) + (1-\rho)\delta_0 \,, \quad \rho \in (0,1] \,.$$



Sparsity helps in theory...

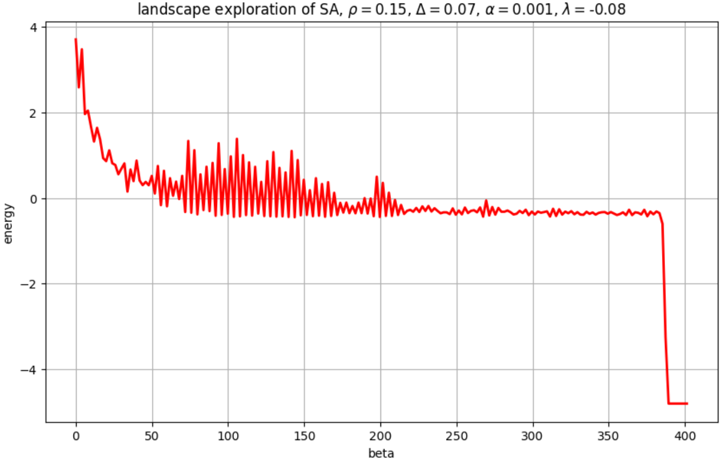
We now know that decimation gets through as long as the initial step has a good retrieval accuracy. Then it only improves. Sparsity helps to widen the regions of the phase space where retrieval is possible.



but we have a problem...

... but it creates a golf course landscape

If we monitor the SA routine with strong sparsity, e.g. $\rho = 0.15$, and few patterns to find, we see that the energy stays constant for a large number of iterations before SA is lucky enough to find a pit!



Thanks!

YHD, Trieste, May 30th, 2023

The Abdus Salam International Centre for Theoretical Physics

United Nations Educational, Scientific and Cultural Organization

IAEA International Atomic Energy Agency

Some References

1. Hopfield, J., *Neural networks and physical systems with emergent collective computational abilities*, PNAS. 79 (8) 2554-2558 (1982)
2. D.J. Amit, H. Gutfreund and H. Sompolinsky, *Storing Infinite Numbers of Patterns in a Spin-Glass Model of Neural Networks*, Physical Review Letters 55, 1530 (1985)
3. Bun, J., Allez, R., Bouchaud, J. P., and Potters, M., *Rotational invariant estimator for general noisy matrices*, IEEE Transactions on Information Theory, 62(12), 7475-7490 (2016)
4. T. Lesieur, F. Krzakala and L. Zdeborová, *MMSE of probabilistic low-rank matrix estimation: Universality with respect to the output channel*, 2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton), pp. 680-687, doi:10.1109/ALLERTON.2015.7447070.
5. Camilli, F. and Mézard, M., *Matrix factorization with neural networks*, Physical Review E (2023, to appear), doi:10.48550/arXiv.2212.02105

and many more... write me if you are interested! `fcamilli@ictp.it`