On convergence of Approximate Message Passing

Francesco Caltagirone[1], Florent Krzakala[2] and Lenka Zdeborova[1]

[1] Institut de Physique Théorique, CEA Saclay

[2] LPS, Ecole Normale Supérieure, Paris

Compressed Sensing

$$\mathbf{y} = F\mathbf{x} + \xi$$

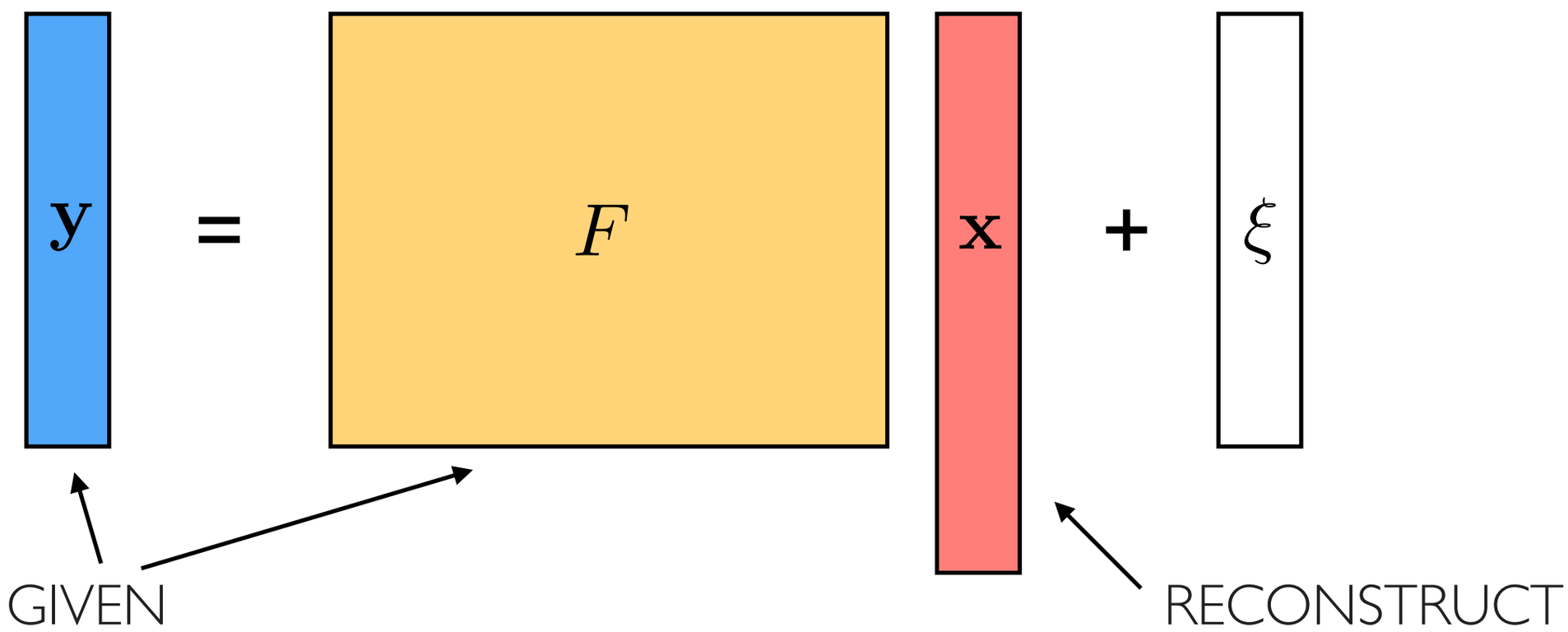
the signal is an N components vector
only K<N components are non-zero

the measurement is an M<N components vector

$$\rho = \frac{K}{N} \qquad \alpha = \frac{M}{N}$$

MxN random matrix with i.i.d. elements

white noise with variance $\langle \xi^2 \rangle = \Delta$



Standard techniques

Minimization of the l_0 norm under linear constraint

$$\min_{\mathbf{x}} ||\mathbf{x}||_0 \quad \text{with} \quad F\mathbf{x} = \mathbf{y}$$

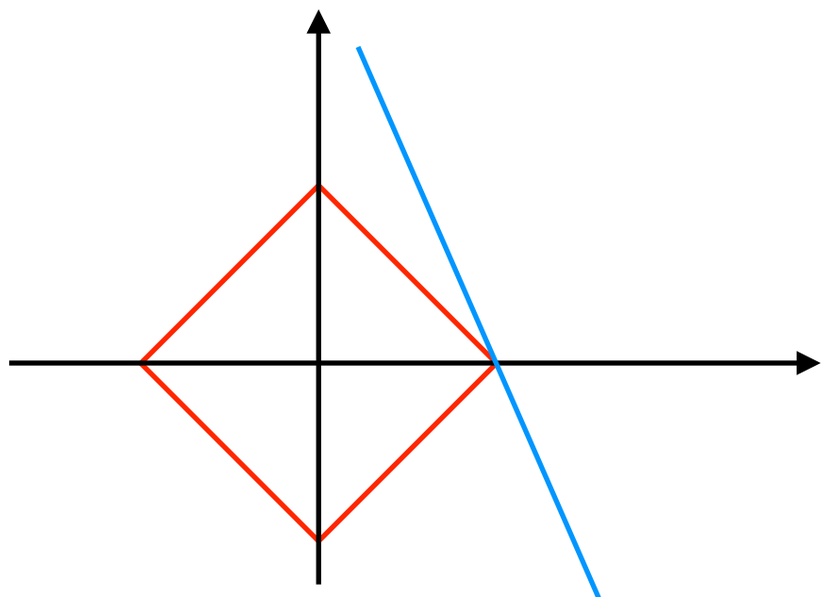
$||\mathbf{x}||_0$ = number of non-zero elements

non-convex norm, exponentially hard to find

Candès, Tao, Donoho ⟶ Minimization of the l_1 norm

$$||\mathbf{x}||_1 = \sum_{i=1}^{N} |x_i|$$ convex norm, easy to minimize

The l_1 norm well approximates the l_0 norm



The Donoho-Tanner line

$$\mathbf{y} = F \mathbf{x} + \xi$$

$\alpha = 1$

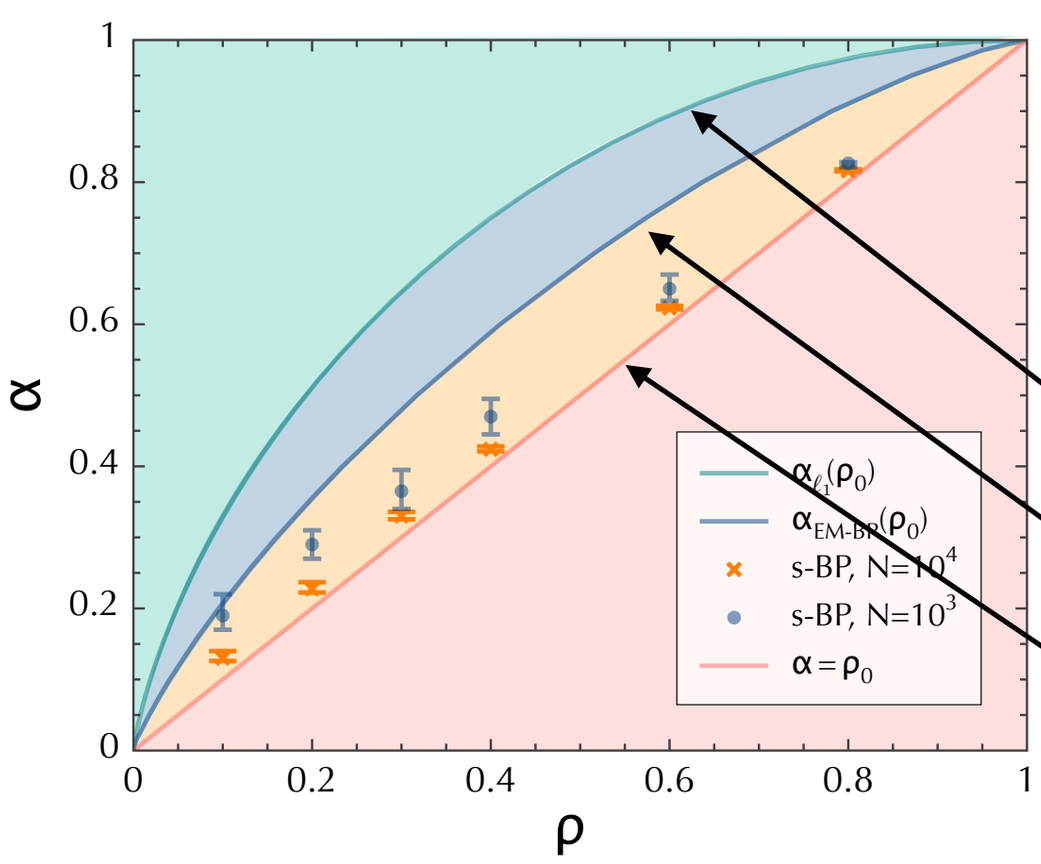
Square matrix, we can invert it

$\alpha < 1$

Rectangular matrix, under-determined

$K = \rho N \qquad M = \alpha N$

$\alpha = \rho$ Information-theoretical limit



Let us consider the noiseless case with a measurement matrix with i.i.d. elements distributed according to a gaussian with zero mean and a variance of order 1/N.

Donoho-Tanner, L1 minimization

AMP Bayesian

Information-Theoretical limit

Setting and motivation

Bayesian setting

GOAL → Reconstruct the signal, given the measurement vector, the measurement matrix and a prior knowledge of the (sparse) distribution of signal elements

Approximate Message Passing

Donoho, Maleki, Montanari (2009)

→ Powerful algorithm.
Convergence issues.

Setting and motivation

$$\mathbf{y} = F\mathbf{x} + \xi \quad \vdots \quad F_{\mu i} = \frac{\gamma}{N} + \frac{1}{\sqrt{N}}\mathcal{N}(0,1) \quad \vdots \quad P(x) = (1-\rho)\delta(x) + \rho\mathcal{N}(0,1)$$

Simplest case in which Approximate Massage Passing (AMP) has convergence problems.

If the mean is sufficiently large then AMP displays violent divergencies.

This kind of divergencies are observed in many other cases and are the main obstacle to a wide use of AMP.

In this simple case there are workarounds that ensure convergence, like a "mean-removal" procedure.
BUT it is interesting because want to understand the origin of the non-convergence that, we argue, is of the same nature in more complicated settings.

Bayesian Inference with Belief Propagation

Bayes formula ⟶ $$P(\mathbf{x}|\mathbf{F},\mathbf{y}) = \frac{P(\mathbf{x}|\mathbf{F})P(\mathbf{y}|\mathbf{F},\mathbf{x})}{P(\mathbf{y}|\mathbf{F})}$$

Conditional probability of the measurement vector ⟶ $$P(\mathbf{y}|\mathbf{F},\mathbf{x}) = \prod_{\mu=1}^{M} \frac{1}{\sqrt{2\pi\Delta_\mu}} e^{-\frac{1}{2\Delta_\mu}(y_\mu - \sum_{i=1}^{N} F_{\mu i} x_i)^2}$$

$$P(\mathbf{x}|\mathbf{F},\mathbf{y}) = \frac{1}{Z(\mathbf{y},\mathbf{F})} \prod_{i=1}^{N} [(1-\rho)\delta(x_i) + \rho\phi(x_i)] \prod_{\mu=1}^{M} \frac{1}{\sqrt{2\pi\Delta_\mu}} e^{-\frac{1}{2\Delta_\mu}(y_\mu - \sum_{i=1}^{N} F_{\mu i} x_i)^2}$$

$$x_i^\star = \int \mathrm{d}x_i \, x_i \, \nu_i(x_i)$$

$$\nu_i(x_i) \equiv \int_{\{x_j\}_{j\neq i}} P(\mathbf{x}|\mathbf{F},\mathbf{y})$$

MMSE estimator

$$E = \sum_{i=1}^{N} (x_i - s_i)^2 / N$$

Takes an exponential time, unfeasible

Bayes optimal setting

If we know exactly the prior distribution on the signal elements
and on the noise we are in the so-called BAYES OPTIMAL setting

In the following we will consider that this is the case.
When it is not the case, the prior can be efficiently learned
adding a step to the algorithm that I will present.
(I will not talk about this)

Belief Propagation (Cavity method)

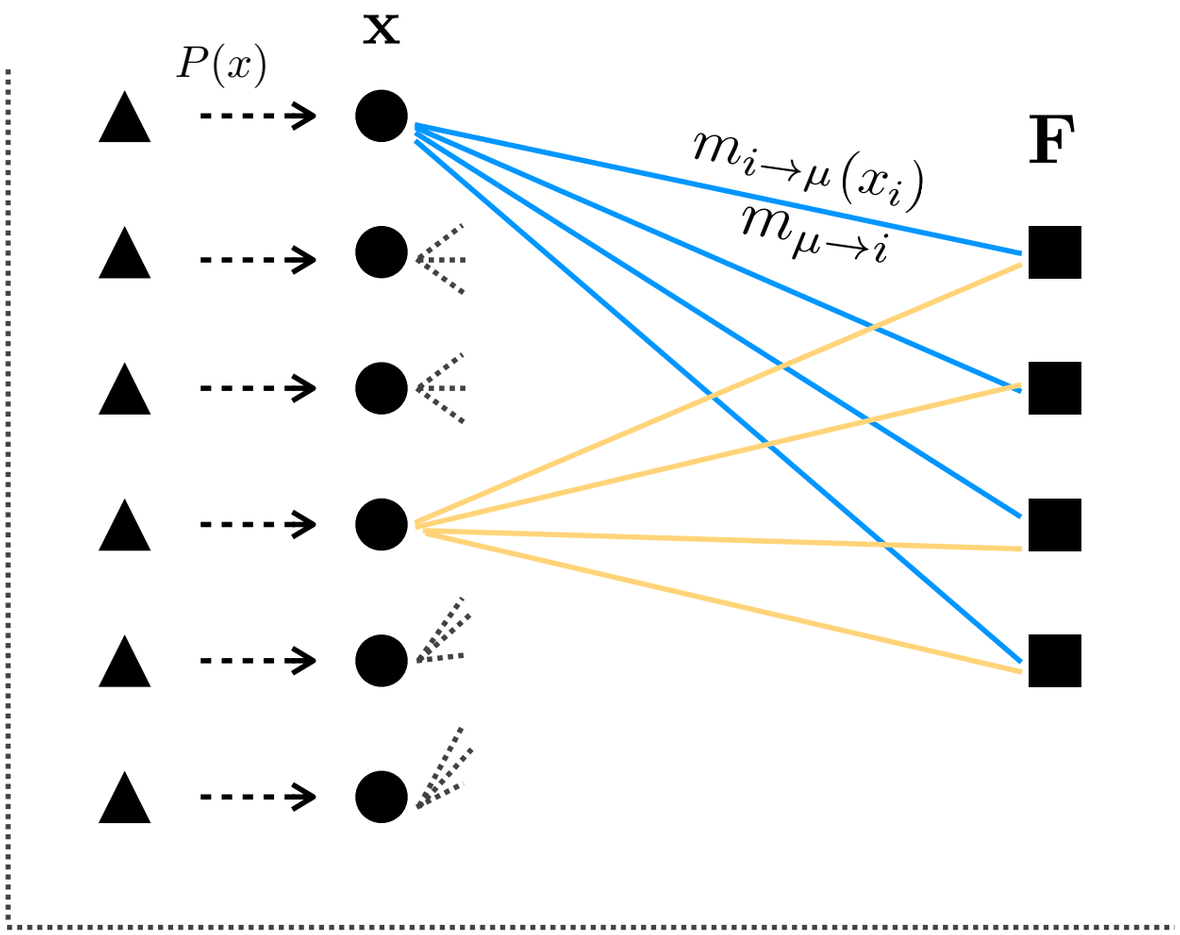
Two kinds of nodes: factors (matrix lines) and variables (signal elements)

We can introduce a third kind of nodes: the prior distribution on the signal elements, local field.

Belief propagation works for:
locally tree-like graphs or densely and weakly connected graphs.

Messages represent an approximation to the marginal distribution of a variable.

Messages are updated according to a sequential or parallel schedule until convergence (fixed point).



Belief Propagation, r-BP and AMP

BP

$\mathcal{O}(N^2)$

continuous messages

↓ *projection*

r-BP

$\mathcal{O}(N^2)$

numbers

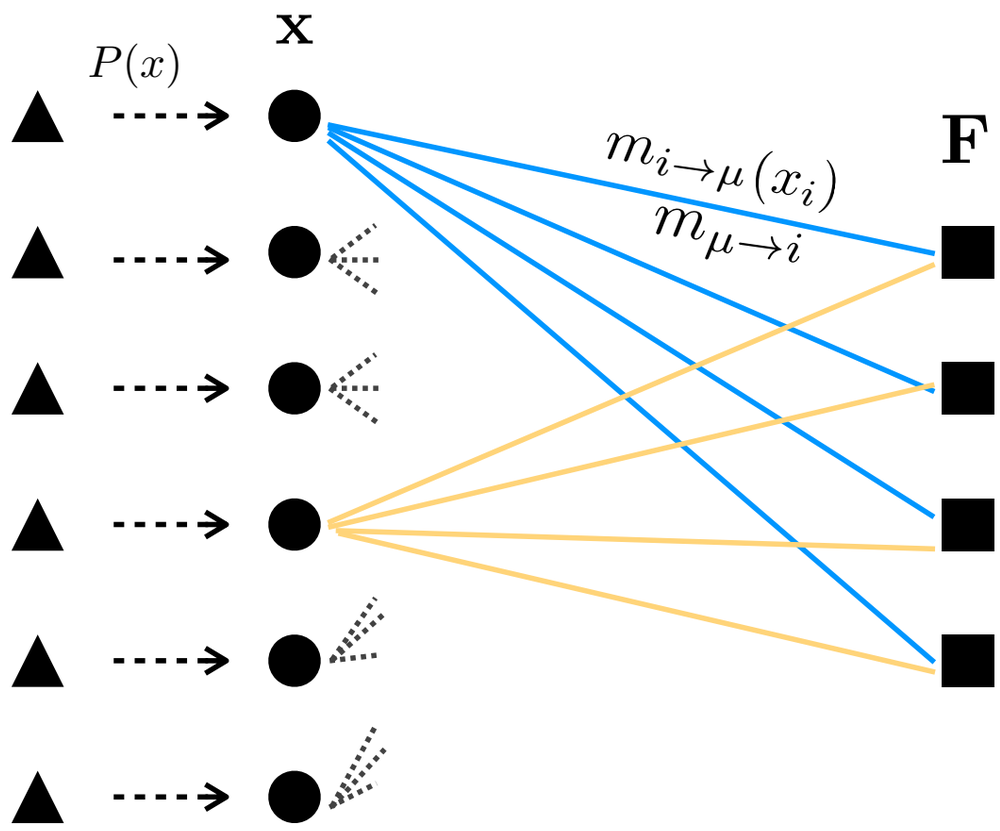
↓ *dense matrix, TAP*

AMP

$\mathcal{O}(N^2)$

operations

Donoho, Maleki, Montanari (2009)
Krzakala et al. (2012)



For the last step one assumes parallel update

In this case, fast matrix multiplication algorithms can be applied, reducing the complexity to

$$N\log(N)$$

AMP Algorithm

$$V_\mu^{t+1} = \sum_i F_{\mu i}^2 \, v_i^t \,, \tag{1}$$

$$\omega_\mu^{t+1} = \sum_i F_{\mu i} \, a_i^t - \frac{(y_\mu - \omega_\mu^t)}{\Delta + V_\mu^t} \sum_i F_{\mu i}^2 \, v_i^t \,, \tag{2}$$

$$(\Sigma_i^{t+1})^2 = \left[\sum_\mu \frac{F_{\mu i}^2}{\Delta + V_\mu^{t+1}} \right]^{-1} , \tag{3}$$

$$R_i^{t+1} = a_i^t + \frac{\sum_\mu F_{\mu i} \frac{(y_\mu - \omega_\mu^{t+1})}{\Delta + V_\mu^{t+1}}}{\sum_\mu \frac{F_{\mu i}^2}{\Delta + V_\mu^{t+1}}} \,, \tag{4}$$

$$a_i^{t+1} = f_1 \left((\Sigma_i^{t+1})^2, R_i^{t+1} \right) , \tag{5}$$

$$v_i^{t+1} = f_2 \left((\Sigma_i^{t+1})^2, R_i^{t+1} \right) . \tag{6}$$

The performance of the algorithm can be evaluated through

$$E^t = \frac{1}{N} \sum_{i=1}^{N} (s_i - a_i^t)^2$$

$$V^t = \frac{1}{N} \sum_{i=1}^{N} v_i$$

$f_k(\Sigma^2, R)$ → k-th connected cumulants w.r.t. the measure

$$\mathcal{Q}(x) = \frac{1}{Z(\Sigma^2, R)} P(x) \frac{e^{-\frac{(x-R)^2}{2\Sigma^2}}}{\sqrt{2\pi\Sigma^2}}$$

a_i and v_i are the AMP estimators for the mean and variance of the i-th signal component.

AMP Algorithm

$$V_\mu^{t+1} = \sum_i F_{\mu i}^2 \, v_i^t \,, \tag{1}$$

$$\omega_\mu^{t+1} = \sum_i F_{\mu i} \, a_i^t - \frac{(y_\mu - \omega_\mu^t)}{\Delta + V_\mu^t} \sum_i F_{\mu i}^2 \, v_i^t \,, \tag{2}$$

$$(\Sigma_i^{t+1})^2 = \left[\sum_\mu \frac{F_{\mu i}^2}{\Delta + V_\mu^{t+1}} \right]^{-1} , \tag{3}$$

$$R_i^{t+1} = a_i^t + \frac{\sum_\mu F_{\mu i} \frac{(y_\mu - \omega_\mu^{t+1})}{\Delta + V_\mu^{t+1}}}{\sum_\mu \frac{F_{\mu i}^2}{\Delta + V_\mu^{t+1}}} \,, \tag{4}$$

$$a_i^{t+1} = f_1 \left((\Sigma_i^{t+1})^2 , R_i^{t+1} \right) , \tag{5}$$

$$v_i^{t+1} = f_2 \left((\Sigma_i^{t+1})^2 , R_i^{t+1} \right) . \tag{6}$$

The performance of the algorithm can be evaluated through

$$E^t = \frac{1}{N} \sum_{i=1}^{N} (s_i - a_i^t)^2$$

$$V^t = \frac{1}{N} \sum_{i=1}^{N} v_i$$

$$F_{\mu i} = \frac{\gamma}{N} + \frac{1}{\sqrt{N}} \mathcal{N}(0, 1)$$

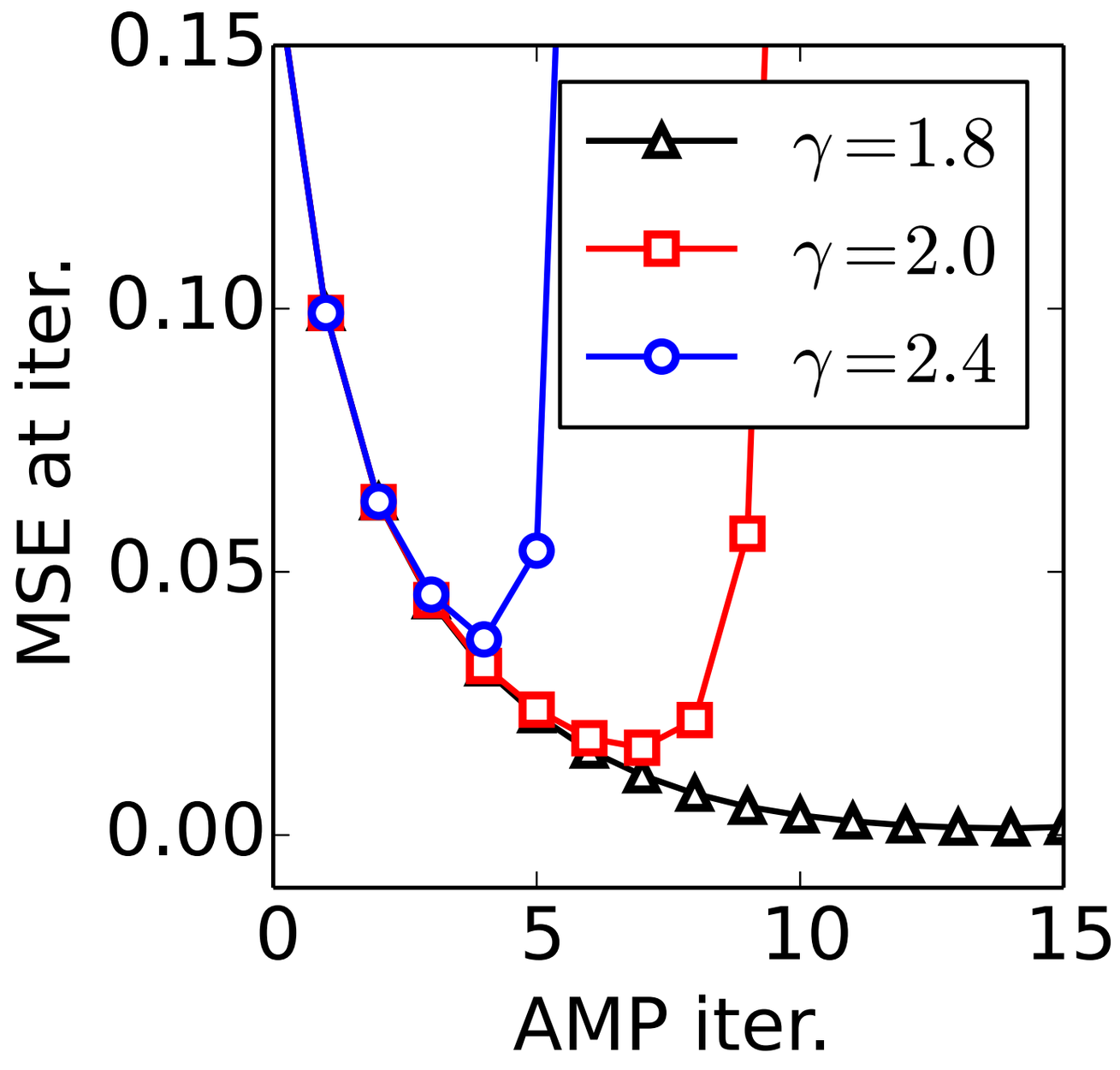
The AMP algorithm does NOT depend explicitly on the value of the mean of the matrix.

Convergence

Bayes optimal case.

Given a certain (sufficiently high) measurement ratio.

Very small or zero noise.



State Evolution (infinite N)

Bayati, Montanari (rigorous in the zero-mean case) '11
Krzakala et al. (replicas in the zero-mean case) '12
Caltagirone, Krzakala, Zdeborova (replicas in the non-zero-mean case) '14

State evolution is the asymptotic analysis of the average performance of the inference algorithm when the size of the signal goes to infinity.

It gives a good indication of what happens in a practical situation if the size of the signal is sufficiently large.

It can be obtained rigorously in simple cases and non rigorously with the replica method in more involved cases.

State Evolution (infinite N)

Bayati, Montanari (rigorous in the zero-mean case) '11
Krzakala et al. (replicas in the zero-mean case) '12
Caltagirone, Krzakala, Zdeborova (replicas in the non-zero-mean case) '14

$$E^t = \frac{1}{N}\sum_{i=1}^{N}(s_i - a_i^t)^2 \qquad V^t = \frac{1}{N}\sum_{i=1}^{N} v_i \qquad D^t = \frac{1}{N}\sum_{j}(s_j - a_j^t)$$

$$V^{t+1} = \int \mathrm{d}s\, P(s) \int \mathcal{D}z \times f_2\left(\frac{\Delta + V^t}{\alpha}, s + z\mathcal{A}(E^t, D^t) + \gamma^2 D^t\right)$$

$$E^{t+1} = \int \mathrm{d}s\, P(s) \int \mathcal{D}z \times \left[s - f_1\left(\frac{\Delta + V^t}{\alpha}, s + z\mathcal{A}(E^t, D^t) + \gamma^2 D^t\right)\right]^2$$

$$D^{t+1} = \int \mathrm{d}s\, P(s) \int \mathcal{D}z \times \left[s - f_1\left(\frac{\Delta + V^t}{\alpha}, s + z\mathcal{A}(E^t, D^t) + \gamma^2 D^t\right)\right]$$

with $\qquad \mathcal{A}(E^t, D^t) = \sqrt{\frac{E^t + \Delta + \gamma^2 (D^t)^2}{\alpha}}$

If the mean is zero the density evolution that does not depend on D

The Nishimori Condition

Bayes optimal setting → $E^t = V^t \qquad D^t = 0$

↓

$E^{t+1} = V^{t+1} \qquad D^{t+1} = 0$

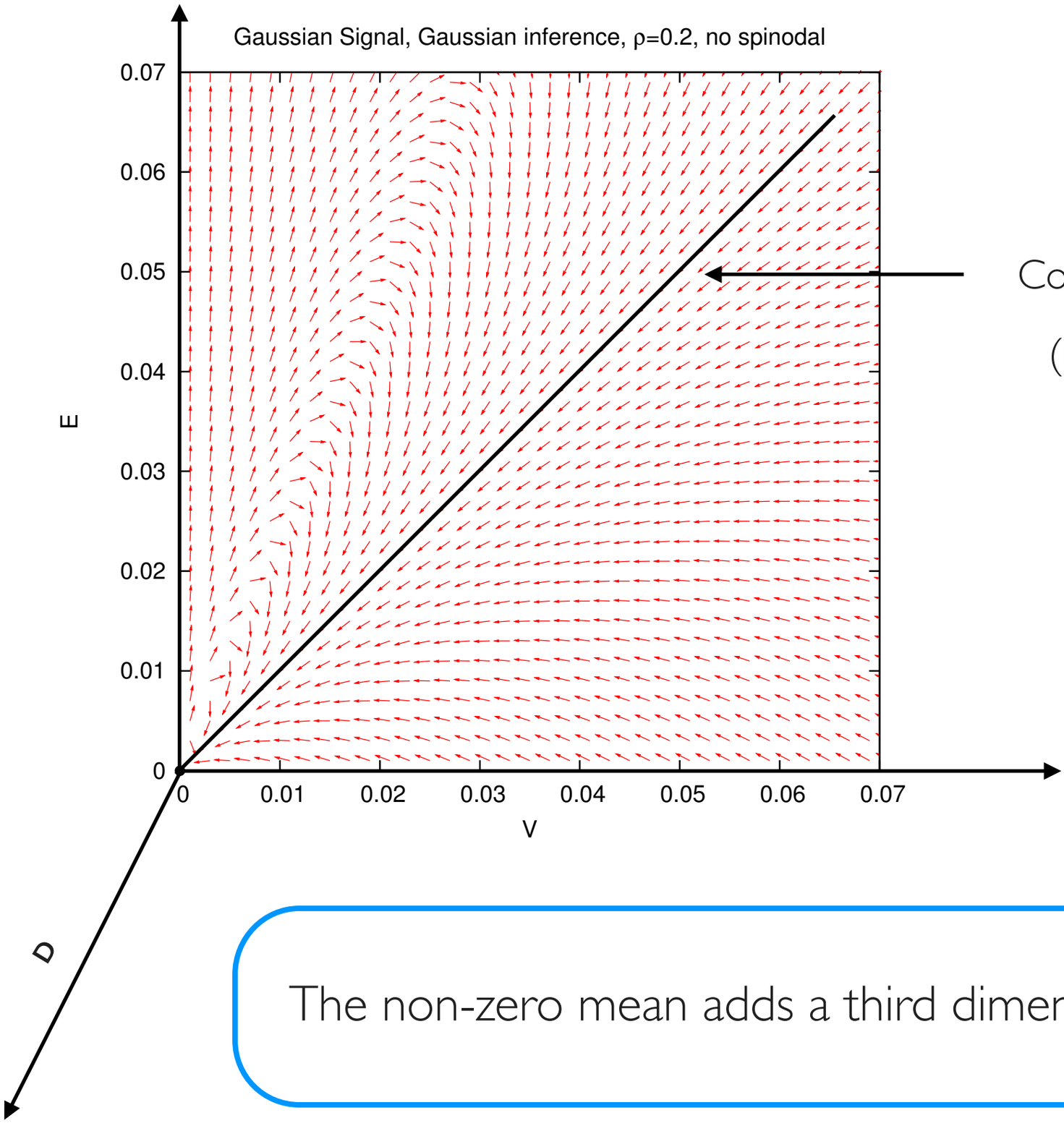
Therefore, analytically, if the evolution starts (exactly) on the Nishimori Line it stays on it until convergence.

BUT

What is the effect of small perturbations with respect to the NL?

- Very small fluctuations due to numerical precision in the DE
- Fluctuations due to finite size in the AMP algorithm

Zero-mean case

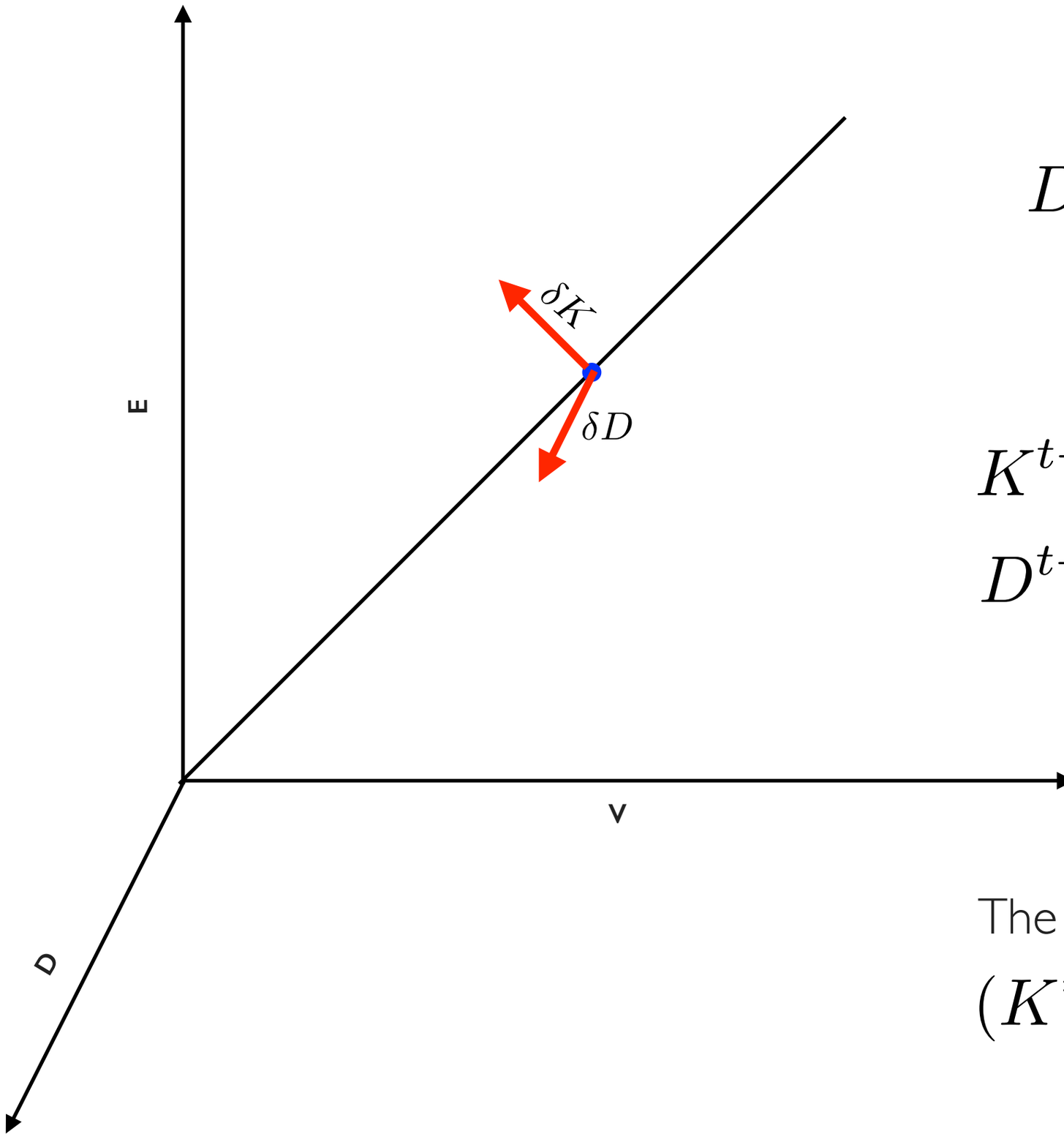


Convergence on the NL

(Bayati, Montanari)

The non-zero mean adds a third dimension to the phase space!

Stability Analysis (I)



$$D \qquad K = V - E$$

$$K^{t+1} = f_K(V^t, K^t, D^t)$$

$$D^{t+1} = f_D(V^t, K^t, D^t)$$

The NL is a "fixed line":

$$(K^* = 0, D^* = 0)$$

Stability Analysis (II)

We linearize the equations with

$$\delta K^t = K^t - K^*$$
$$\delta D^t = D^t - D^*$$

$$\begin{pmatrix} \delta K^{t+1} \\ \delta D^{t+1} \end{pmatrix} = \mathcal{M} \cdot \begin{pmatrix} \delta K^t \\ \delta D^t \end{pmatrix}$$

$$\mathcal{M} = \begin{pmatrix} \partial_K f_K(V^t, 0, 0) & \partial_D f_K(V^t, 0, 0) \\ \partial_K f_D(V^t, 0, 0) & \partial_D f_D(V^t, 0, 0) \end{pmatrix}$$

Stability Analysis (II)

We linearize the equations with

$$\delta K^t = K^t - K^*$$
$$\delta D^t = D^t - D^*$$

$$\begin{pmatrix} \delta K^{t+1} \\ \delta D^{t+1} \end{pmatrix} = \mathcal{M} \cdot \begin{pmatrix} \delta K^t \\ \delta D^t \end{pmatrix}$$

$$\mathcal{M} = \begin{pmatrix} \partial_K f_K(V^t, 0, 0) & 0 \\ 0 & \partial_D f_D(V^t, 0, 0) \end{pmatrix}$$

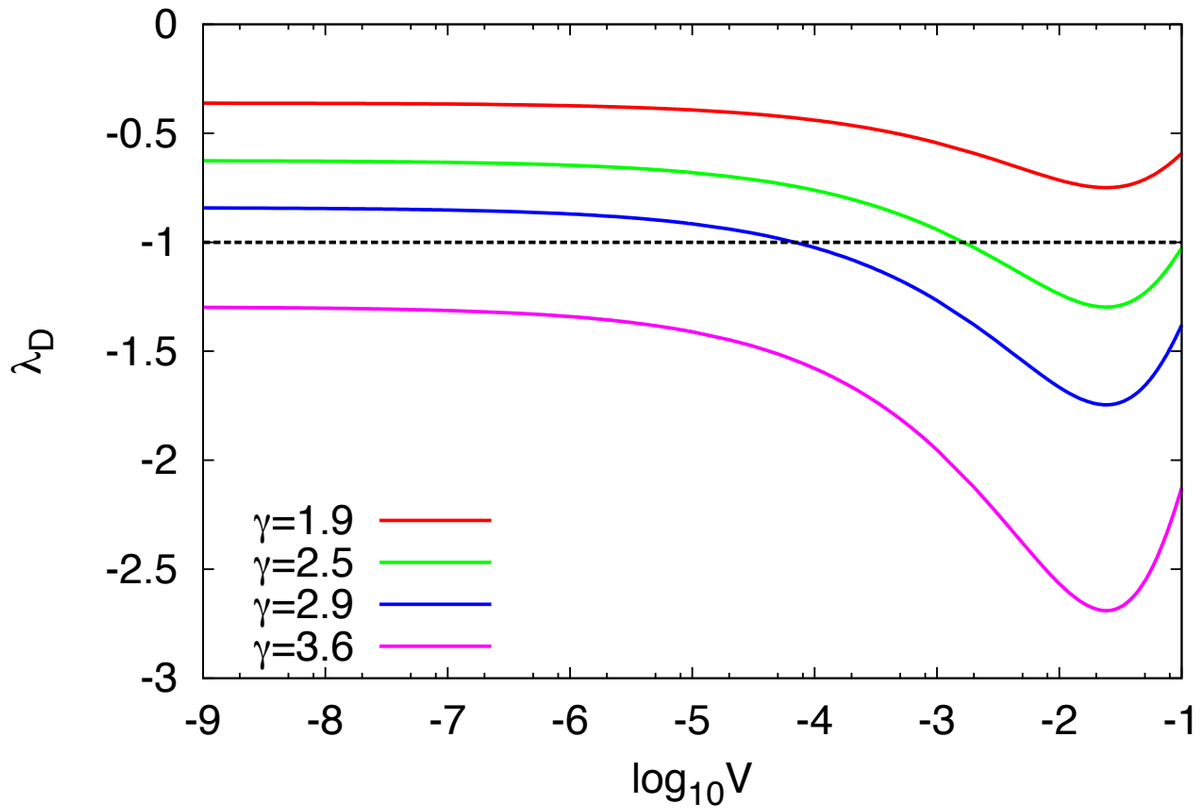
When the signal is Gauss-Bernoulli with zero mean, the off-diagonal terms vanish.

Stability Analysis (II)

$$\partial_D f_D(V^t) = -\frac{\alpha\gamma^2}{\Delta + V^t} \int ds\, P(s) \int \mathcal{D}z\, f_2\left(A^2, s + zA\right) = -\frac{\alpha\gamma^2 V^{t+1}}{\Delta + V^t}\,, \qquad \lambda_D$$

$$\partial_K f_K(V^t) = -\frac{1}{2}\frac{1}{\Delta + V^t} \int ds\, P(s) \int \mathcal{D}z \left\{ f_4\left(A^2, s + zA\right) + 2(f_2\left(A^2, s + zA\right))^2 \right.$$
$$\left. +2\left[f_1\left(A^2, s + zA\right) - s\right] f_3\left(A^2, s + zA\right) \right\}\,, \qquad \lambda_K$$

$\rho = 0.1 \qquad \alpha = 0.3 \qquad \Delta = 10^{-10}$



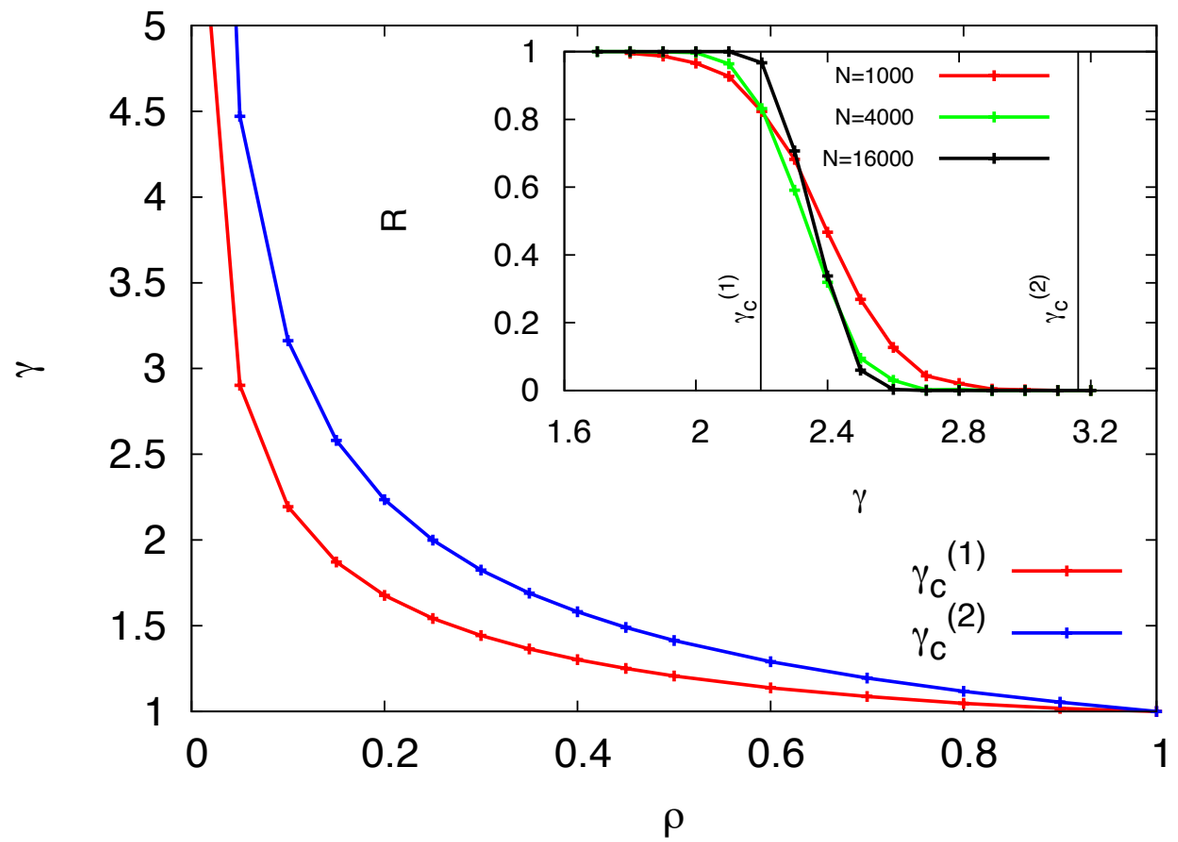
- $\gamma < \gamma_c^{(1)}$ the eigenvalue is always less than 1 in modulus.
- $\gamma_c^{(1)} < \gamma < \gamma_c^{(2)}$ the eigenvalue becomes larger than 1 in a limited region.
- $\gamma > \gamma_c^{(2)}$ the eigenvalue is larger than 1 in modulus down to the fixed point.

Density Evolution and AMP

For zero measurement noise both the critical values do NOT depend on the undersampling rate α .

For weak noise only the second critical value has a very weak dependence on both Δ and α .

[Inset] Convergence Rate of the AMP algorithm for different signal sizes.



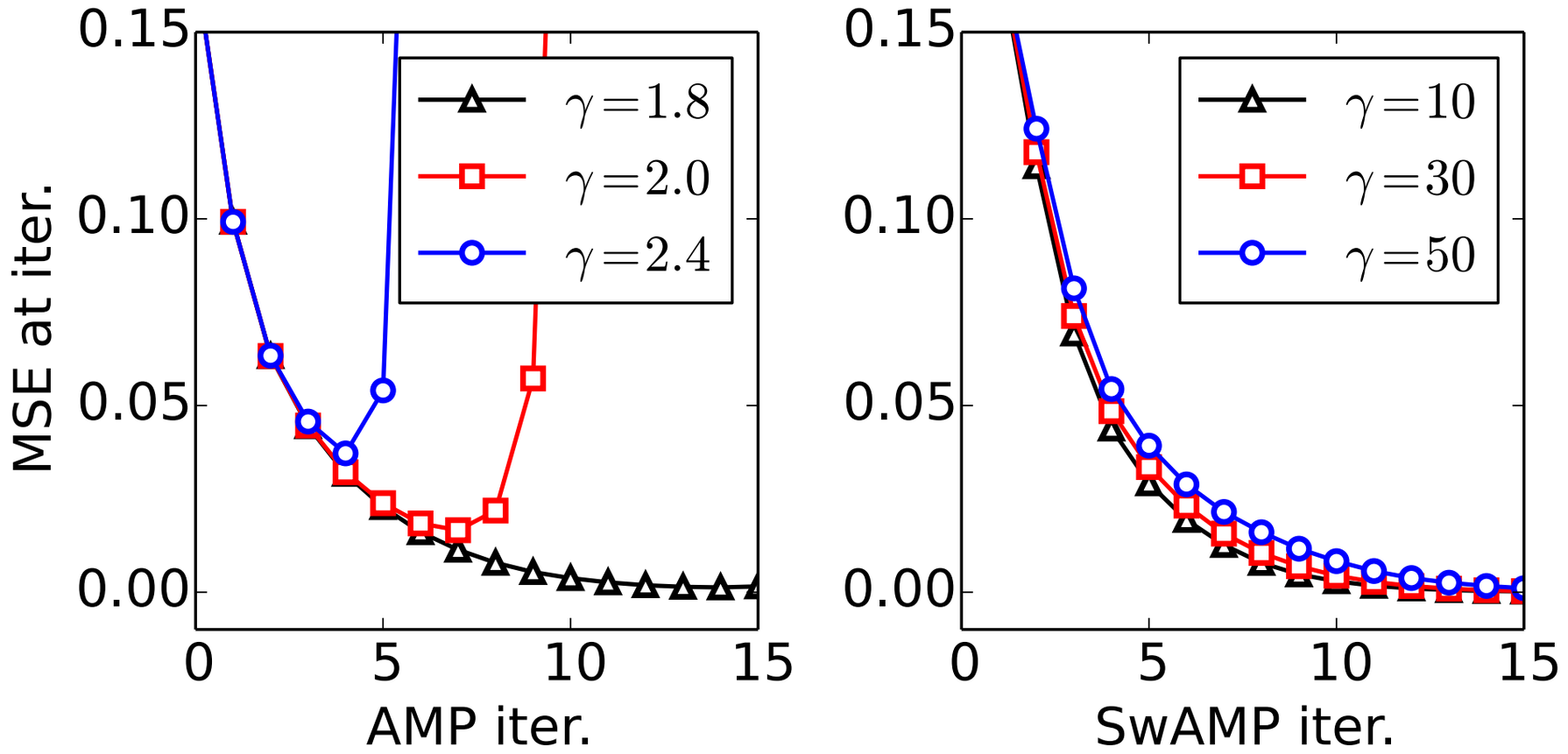
$$N \to \infty$$

The transition becomes sharper and sharper

It is expected to move towards the second critical value and behave similarly to the density evolution.

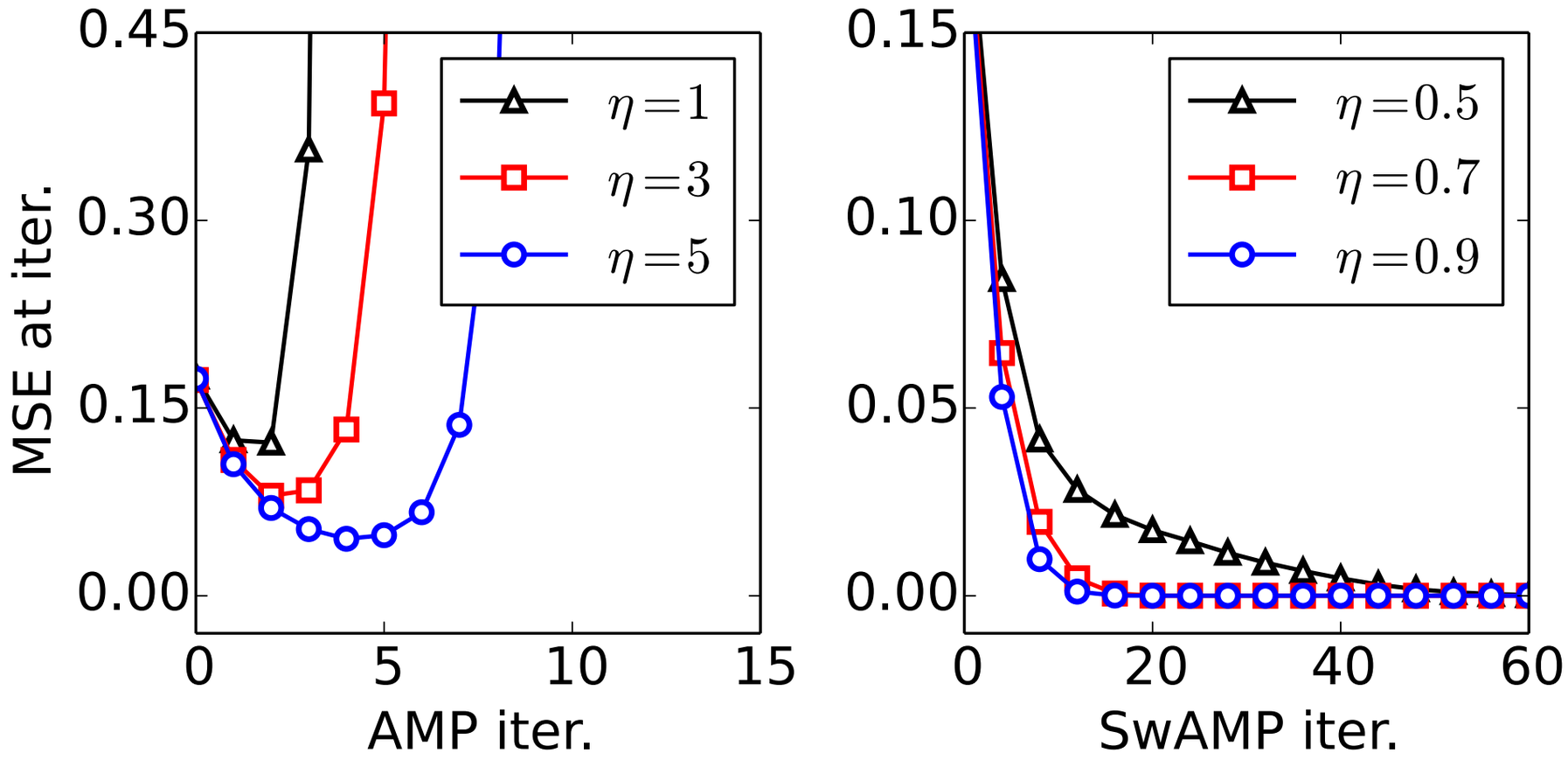
SwAMP algorithm, a possible solution

Manoel, Krzakala, Tramel, Zdeborova (2014)



With random sequential update convergence problems disappear.

SwAMP algorithm, a possible solution



LOW RANK!

Very effective solution that works well in many interesting cases!

It is not a universal solution.

Looses the property of involving only matrix multiplications.

Caltagirone, Manoel, Krzakala, Tramel, Zdeborova. arXiv:1401.6384

Rangan, Schniter, Fletcher arXiv:1406.4311

Kabashima

Conclusions and Perspectives

- We found that the origin of the convergence problems is an instability of the Nishimori Line
- We provided a possible solution with the SwAMP algorithm.

- Relate this kind of instability in the density evolution to the shape of the replica potential.
- Perform the same kind of analysis for the case of dictionary learning.

THANK YOU!