BEYOND MEAN-FIELD APPROXIMATION

AURÉLIEN DECELLE

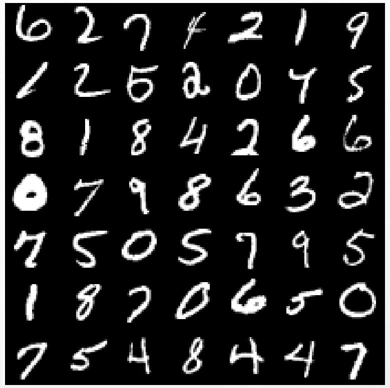
LABORATOIRE DE RECHERCHE EN INFORMATIQUE

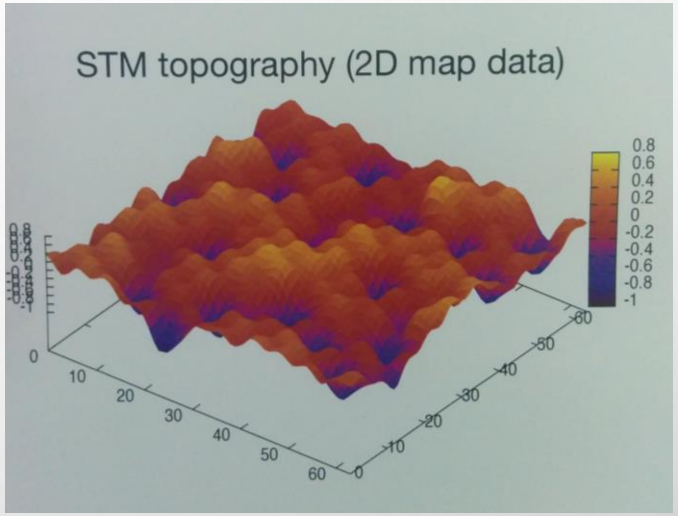
UNIVERSITÉ PARIS SUD

MOTIVATIONS

Why inverse problems ?

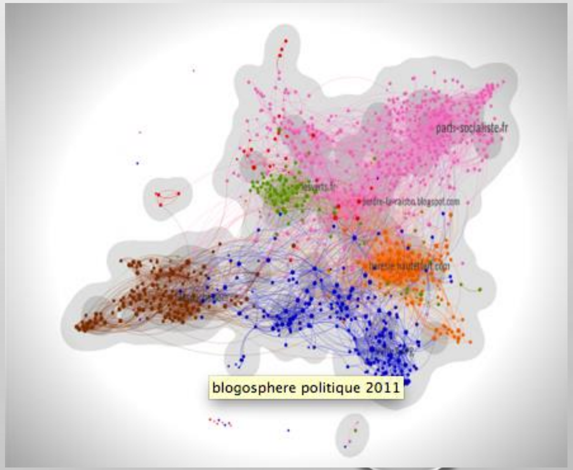
- ❑ In Machine Learning → online recognition tasks





- ❑ In Physics → understanding a physical system from observations

- ❑ In social science → getting insight of latent properties



HOW HARD ?

Direct problems are already hard : understanding equilibrium properties can be (very) challenging (e.g. spin glasses)

Inverse problems can be harder : ideally maximizing the likelihood would involve to compute the partition function many times

In particular, serious problems can appear because if

- Overfitting
- Non-convex functions
- Slow convergence in the direct problem

HOW HARD ?

Depending on the system, different optimization scheme can be adopted

Mean-field

Pseudo-likelihood

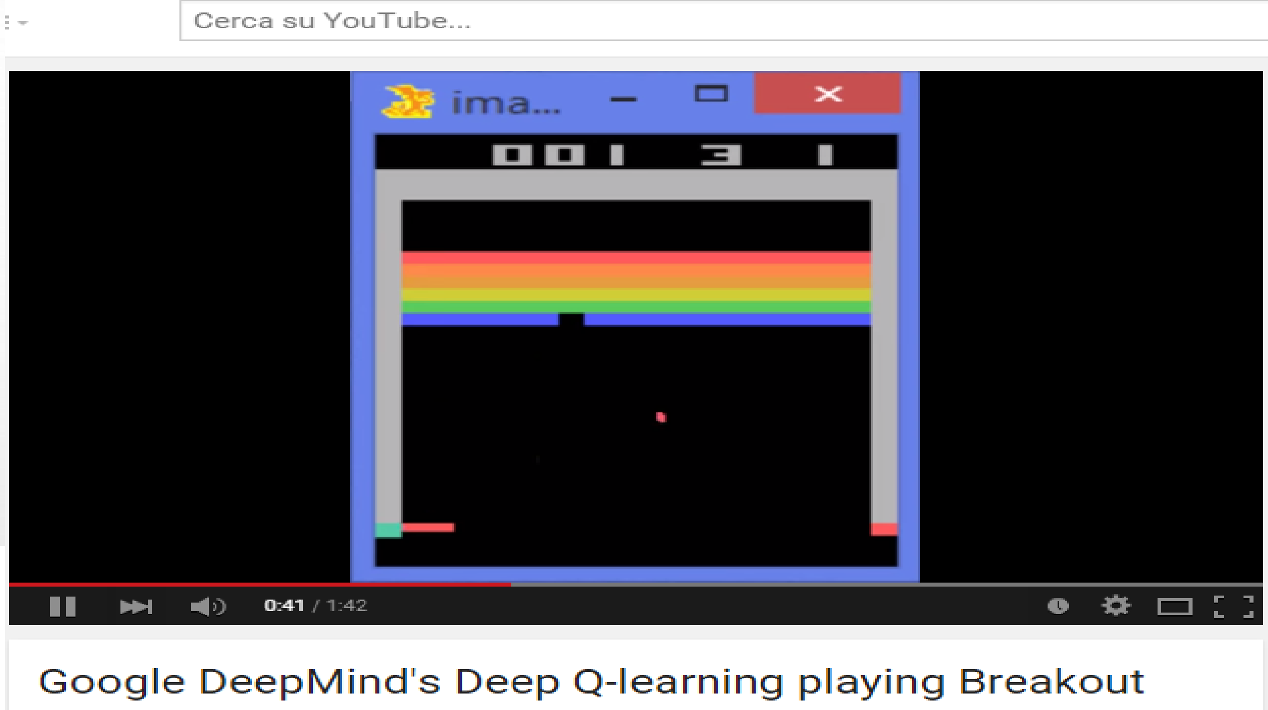
Contrastic Divergence

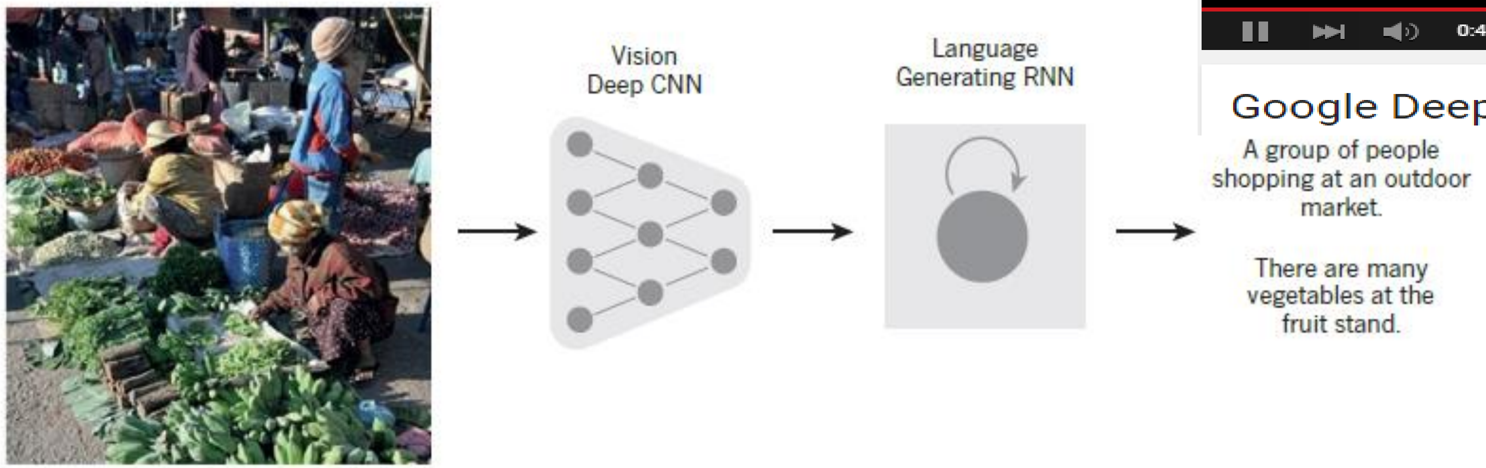
Cluster expansion

Others

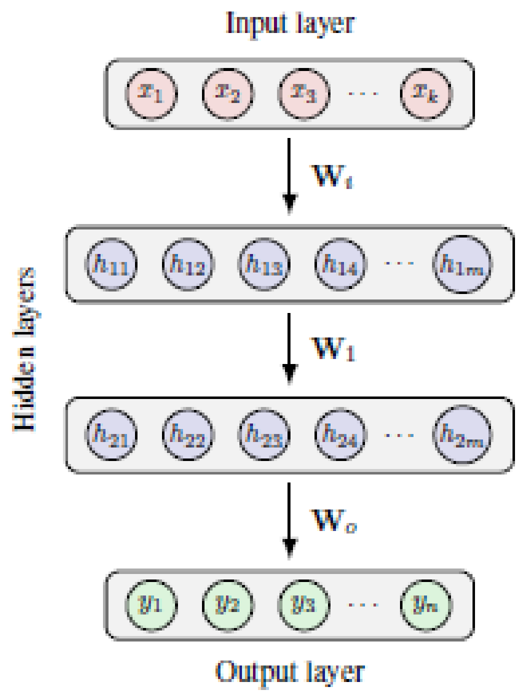
DEEP LEARNING

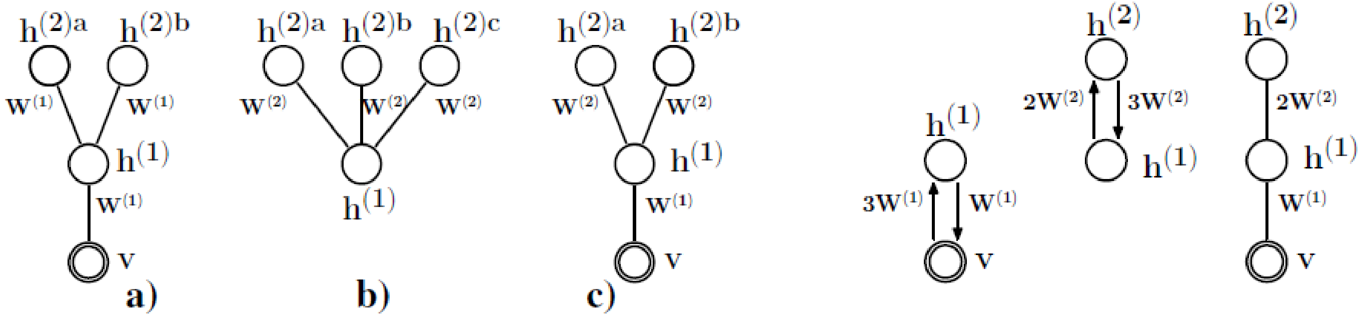
A **stop** sign is on a road with a mountain in the background

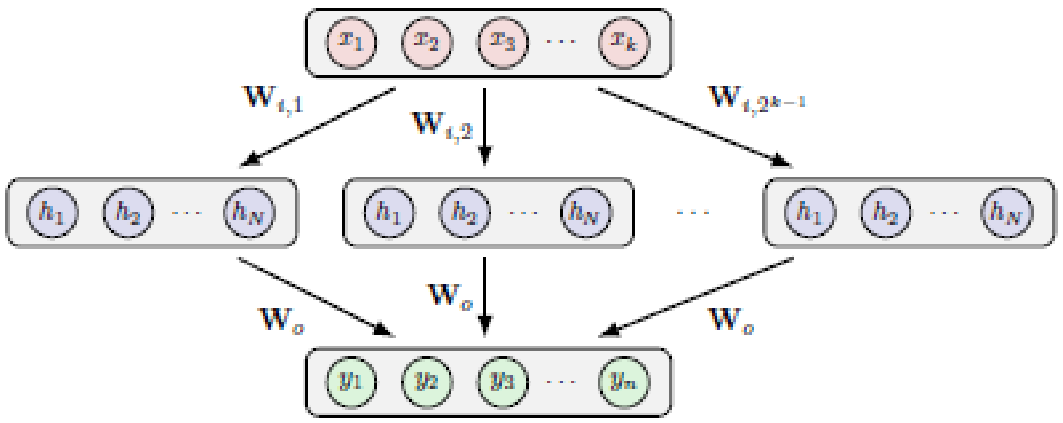




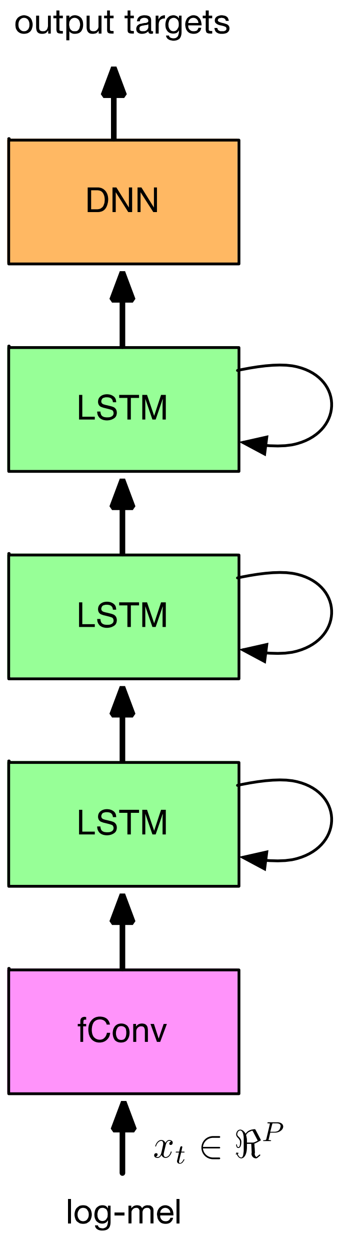
ICML STUFFS







# DNN Layers	WER
0	18.0 (LSTM)
1	17.8
2	**17.6**
3	17.6



WHY IT IS NEEDED TO GO BEYOND MF

MF is mapping the distribution of the data onto a particular form of probability distribution

$$\min_{\vartheta} KL(p_{data} || p_{target}(\vartheta))$$

nMF $\quad p_{nMF}(\vartheta) = \prod_i p_i(s_i)$

Bethe approx $\quad p_{BA}(\vartheta) = \prod_{ij} \frac{p_{ij}(s_i,s_j)}{p_i(s_i)p_j(s_j)} \prod_i p_i(s_i)$

WHY IT IS NEEDED TO GO BEYOND MF

What about when the system can not be describe by this particular form of distribution ?

- Long-range correlations
- Very specific topology
- Presence of hidden nodes

$\oplus$ how to put prior information ?

OTHER METHODS ?

Pseudo-Likelihood

- Trade off between complexity and the level of approximation
- Consistent for infinite sampling
- Can deal with priors

But overfit

Max likelihood

- Same as the two last points of above

But overfit and can be very slow

OTHER METHODS ?

Adaptive cluster exansion

- Avoid overfitting
- Consistently develop cluster of larger sizes

But it is hard to write it …

Contrastic divergence

- Very fast
- A trade off can be found between speed and exactness

Overfit, and can be bad if very slow convergence !

Minimum Probabilistic Flow

- Fast to converge
- Consistent

But probably does not work well for small sampling.

PSEUDO-LIKELIHOOD METHOD

- Principle
- Comparison with MF
- Regularization
- Decimation
- Generalisation and extension

SETTINGS

We consider the following problem :
A system of discrete variables $s_i = 1, \dots, q$ (ok let's say $s_i = \pm 1$ in the following)
- Interacting by pairs and having biases.

$$\boldsymbol{\mathcal{H}} = \sum_{<i,j>} J_{ij} s_i s_j + \sum_i h_i s_i \qquad \mathrm{p}(\vec{s}) = \frac{e^{-\beta \boldsymbol{\mathcal{H}}(\vec{s})}}{Z}$$

Then, a set of configuration is collected : $\{\vec{s}^{(a)}\}_{a=1,..,M}$
Using them, it is possible to compute the likelihood

Reconstruction error $\varepsilon^2 = \frac{\sum (J_{ij} - J_{ij}^*)^2}{\sum J_{ij}^2}$

SETTINGS

The likelihood function

Proba of observing the configurations $= \prod_a \frac{e^{-\beta \boldsymbol{\mathcal{H}}(\vec{s}^{(a)})}}{Z}$

Define the log-likelihood $\mathcal{L} = \sum_a (-\beta \boldsymbol{\mathcal{H}}(\vec{s}^{(a)}) - \log(Z))$

$$\frac{\partial \mathcal{L}}{\partial J_{ij}} \propto < s_i s_j >_{data} - < s_i s_j >_{model}$$

Problem of maximization …
How to compute average values efficiently ?

PSEUDO–LIKELIHOOD

Goal : find a function that can be maximize and would infer correctly the Js

$$p(\vec{s}) = p(s_i|\vec{s}_{j\setminus \mathrm{i}}) \sum_{s_i} p(\vec{s}) = \boxed{p(s_i|\vec{s}_{j\setminus \mathrm{i}})} p(\vec{s}_{j\setminus \mathrm{i}})$$

$$p(s_i|\vec{s}_{j\setminus \mathrm{i}}) = \frac{e^{-\beta s_i (\sum_j J_{ij} s_j + h_i)}}{2\cosh(\beta\, (\sum_j J_{ij} s_j + h_i)\,)}$$ can be minimized !

Ekeberg et al. : Protein foldings
??? : training RBM

PSEUDO–LIKELIHOOD

Can we have theoretical insight ? Yes, for gibbs infinite sampling, the maximum is correct !

Consider : $\mathcal{PL}_i = \sum_a \log(p(s_i|\vec{s}_{j\setminus\mathrm{i}}))$ we replace the distribution over the data by Boltzmann

$$\mathcal{PL}_i = \sum_{\mathcal{C}} \frac{e^{-\beta \mathcal{H}_G\,(\vec{s}^{\mathcal{C}})}}{Z_G} \log(p\left(s_i^{\mathcal{C}} \middle| \vec{s}_{j\setminus\mathrm{i}}^{\mathcal{C}}\right))$$

The maximum is reached when the couplings from $\mathcal{H}_G$ and $\mathcal{H}$ of are equals

PSEUDO–LIKELIHOOD

When no hidden variables are present, the PL is convex !
Therefore only one maxima exists !

The PL can be minimized without too much trouble using for instance

- Newton method
- Gradient descent

And the complexity goes as $O(N^2M)$

Let's understand how this works and how it compares to MF

RECALL OF THE SETTING

A set of M equilibrium configurations $\{\vec{s}^{(k)}\}, k = 1,..,M$
On one side we use the MF equations

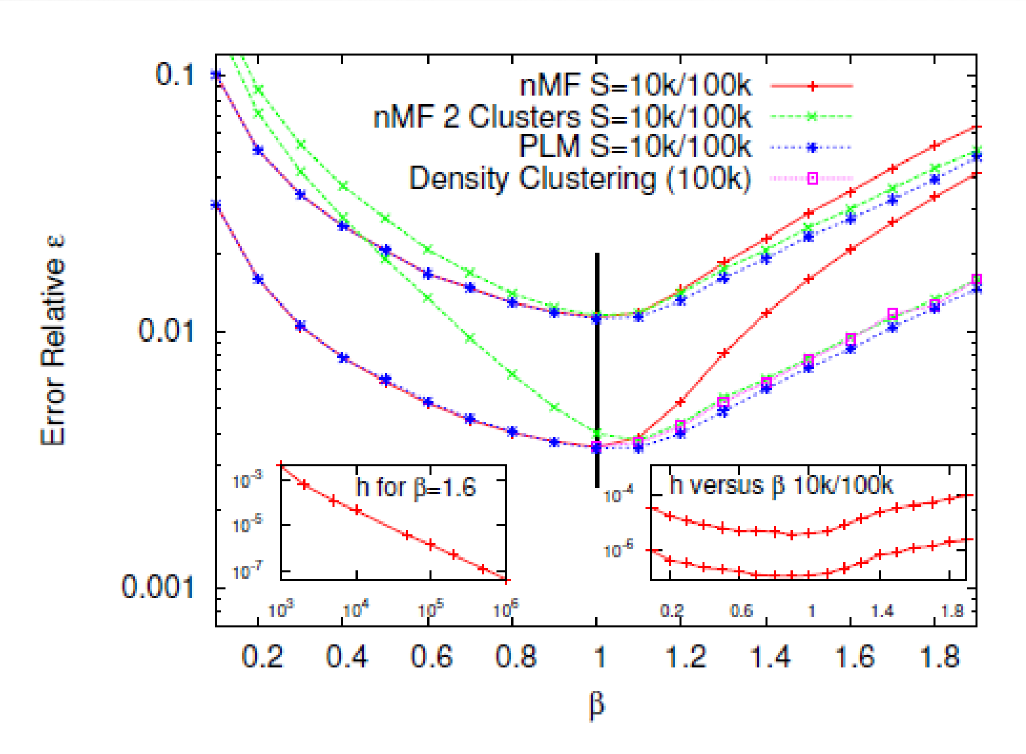
$$m_i = \tanh(\sum_j J_{ij} m_j + h_j) \qquad J_{ij} = -c_{ij}^{-1}$$

On the other side we maximize the Pseudo-Likelihood distributions

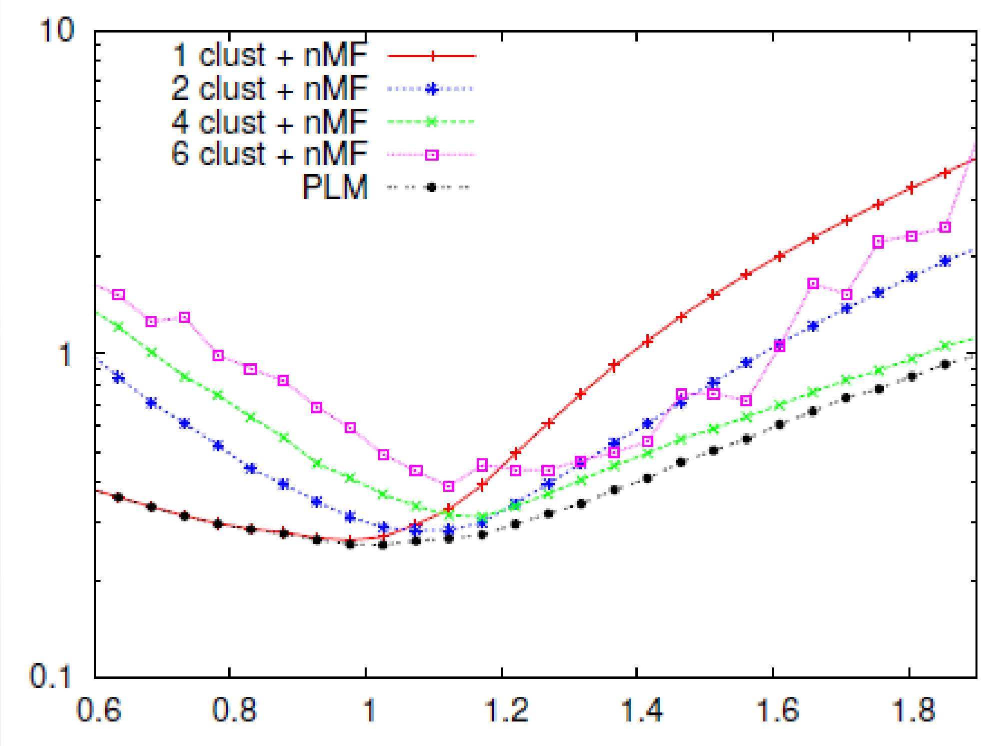
$$\mathcal{PL}_i = \sum_k \log(1 + e^{-2\beta s_i^{(k)} \sum J_{ij} s_j^{(k)}}) \ \forall i$$

MEAN-FIELD AND PLM

Curie-Weiss $J_{ij} = -1/N$ with N=100 spins

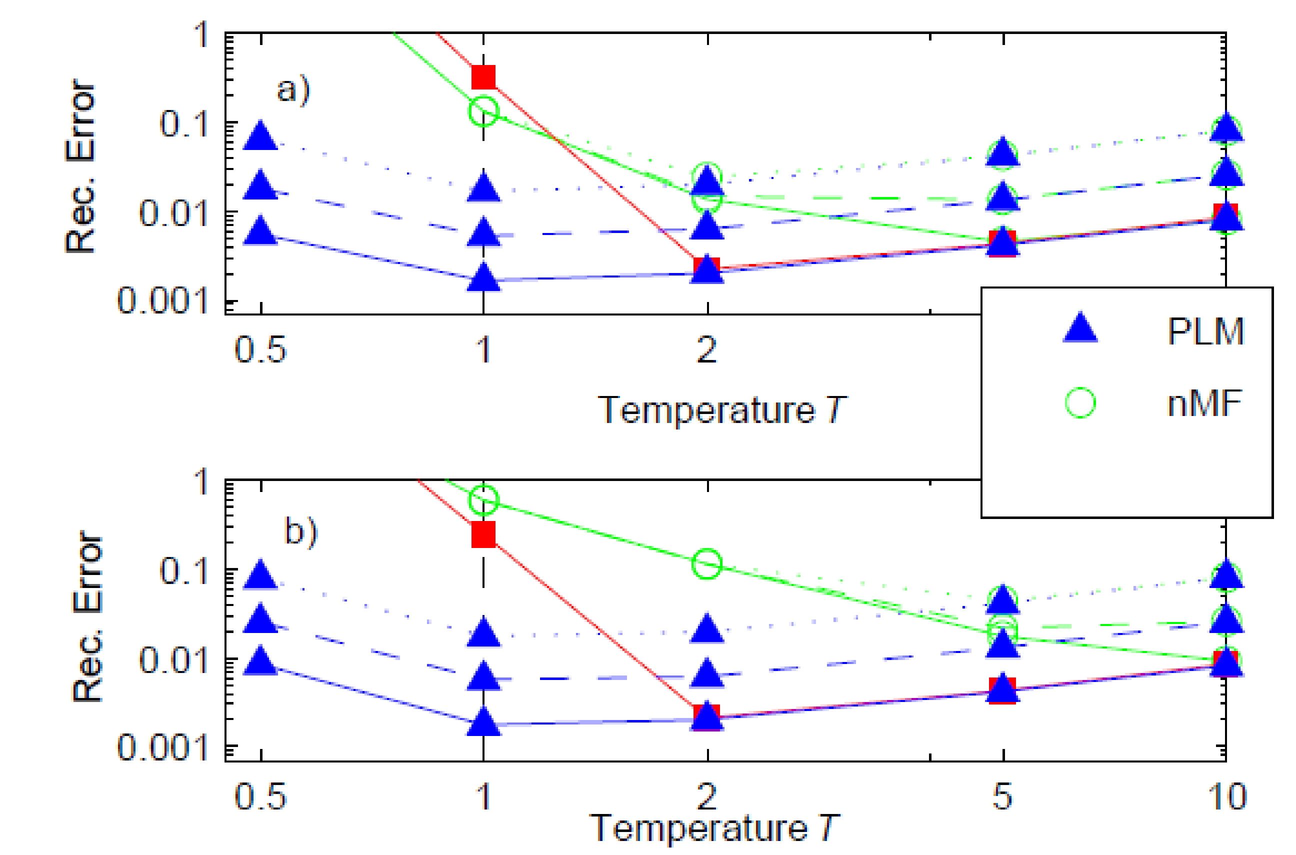


Hopfield $J_{ij} = \sum \xi_i^a \xi_j^a$ with N=100 spins and two states, M=100k

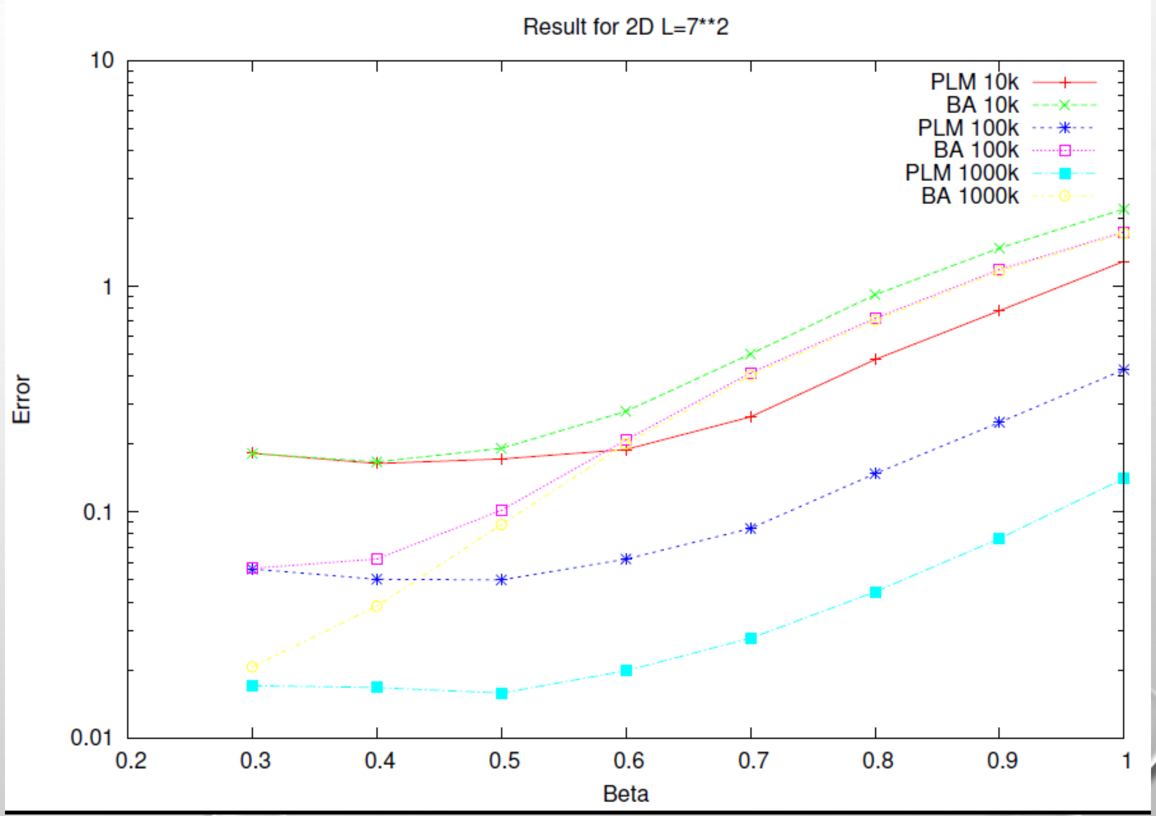


MEAN-FIELD AND PLM

SK model, N=64, with M=10^6, 10^7, 10^8

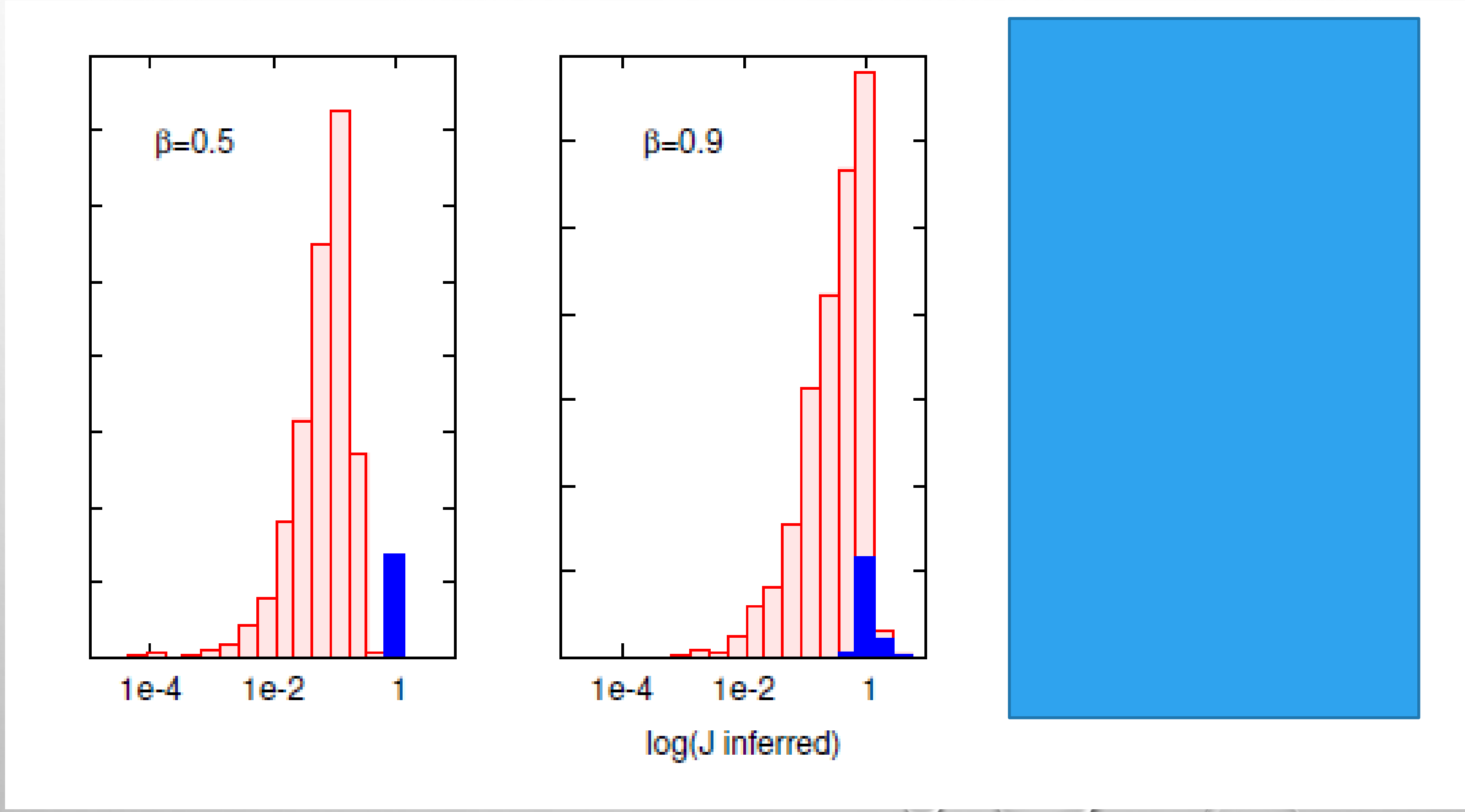


2D model, $J_{ij} = -1$, N=49, with M=10^4, 10^5, 10^6



E. Aurell and M. Ekeberg 2012

WHAT ABOUT THE STRUCTURE ?



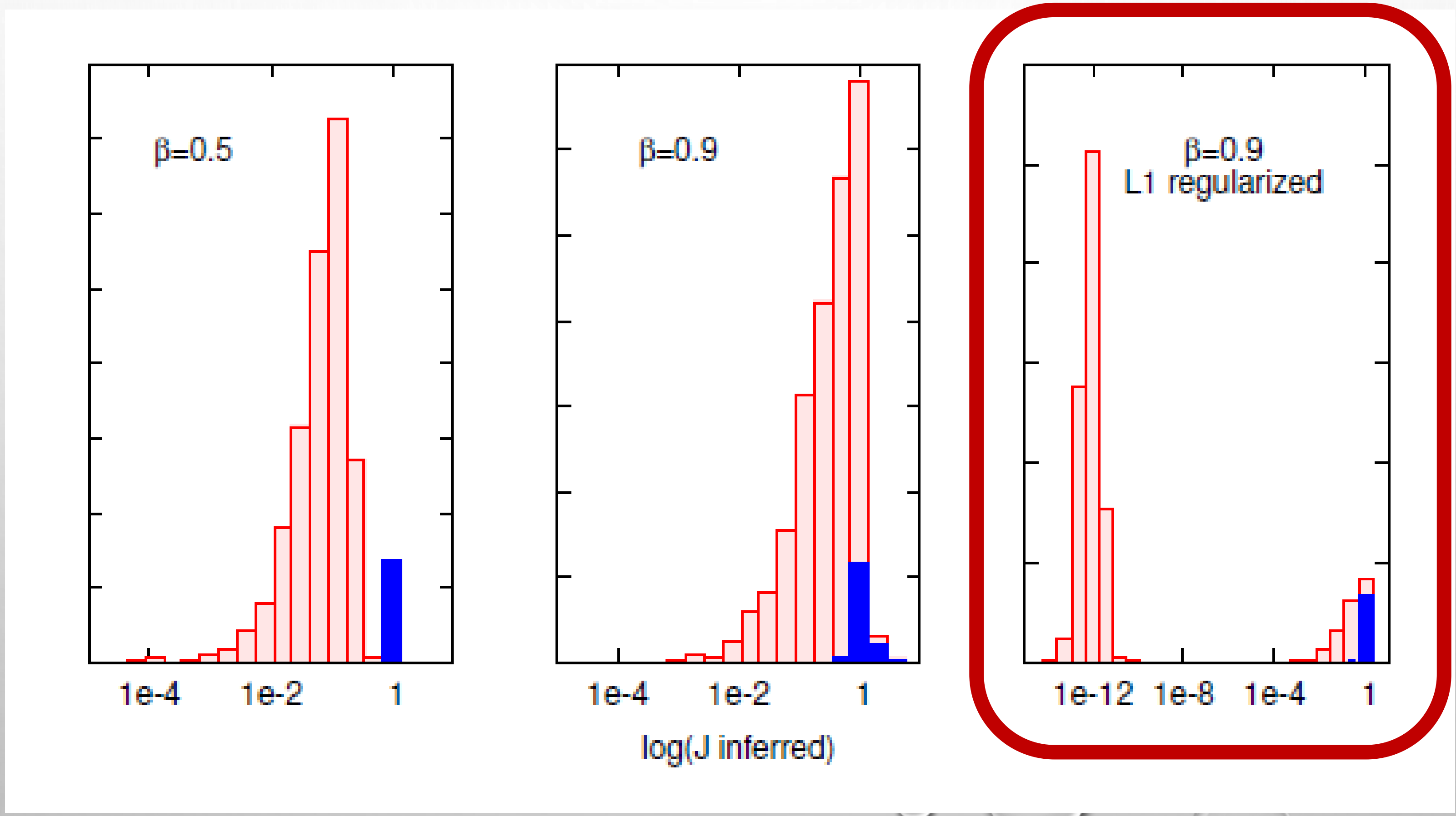
WHAT ABOUT THE STRUCTURE ?

How does the L1-norm is included in PLM ?

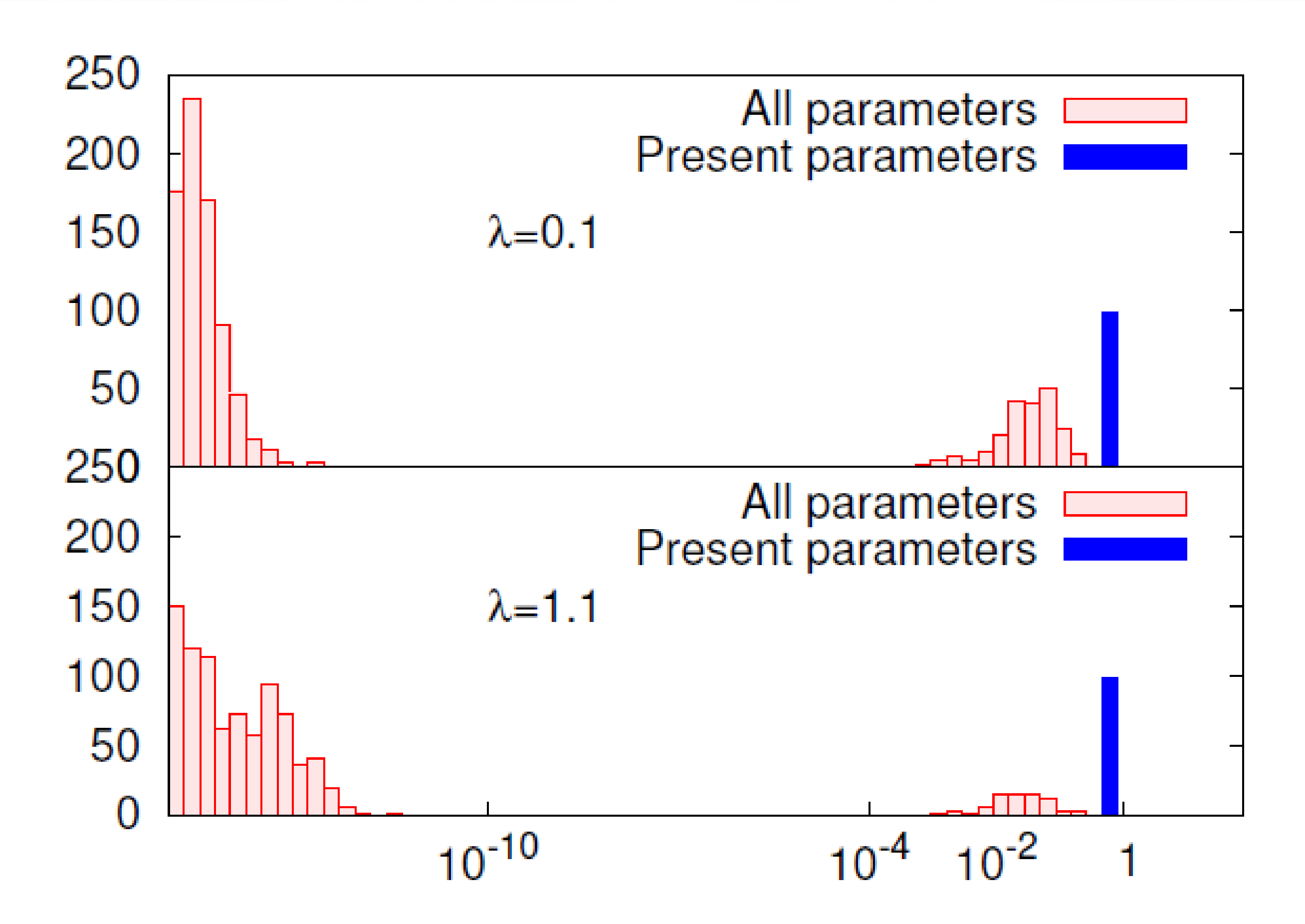
$$\mathcal{PL}_i = \sum_k \log\left(1 + e^{-2\beta s_i^{(k)} \sum J_{ij} s_j^{(k)}}\right) - \lambda \sum_j |J_{ij}| \ \forall i$$

Leads to sparse solution … how to fix λ ?

WHAT ABOUT THE STRUCTURE ?



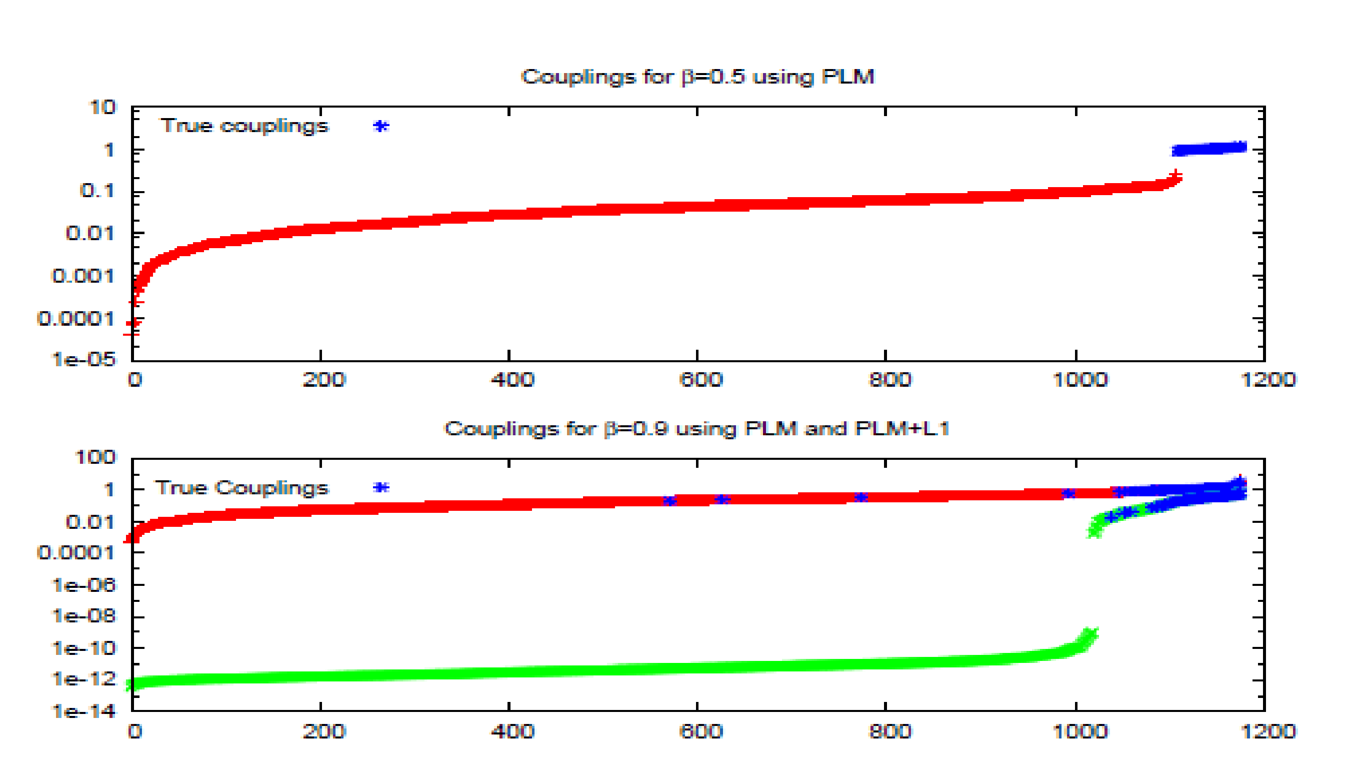
WHAT ABOUT THE STRUCTURE ?



VERY SIMPLE IDEA : DECIMATION

Progressively decimating parameters with a small absolute values
Not NEW :

- In optimization problem using BP (Montanari et al.)
- Brain damage (Lecun)



DECIMATION ALGORITHM

Given a set of equilibrium configurations and all unfixed paramaters

1. Maximize the Pseudo-Likelihood function over all non-fixed variables
2. Decimate the $\rho(t)$ smallest variables (in magnitude) and fixed them
3. If (criterion is reached)
    1. exit
4. Else
    1. $t \leftarrow t + 1$
    2. goto 1.

Join work with F. Ricci-Tersenghi

DECIMATION ALGORITHM

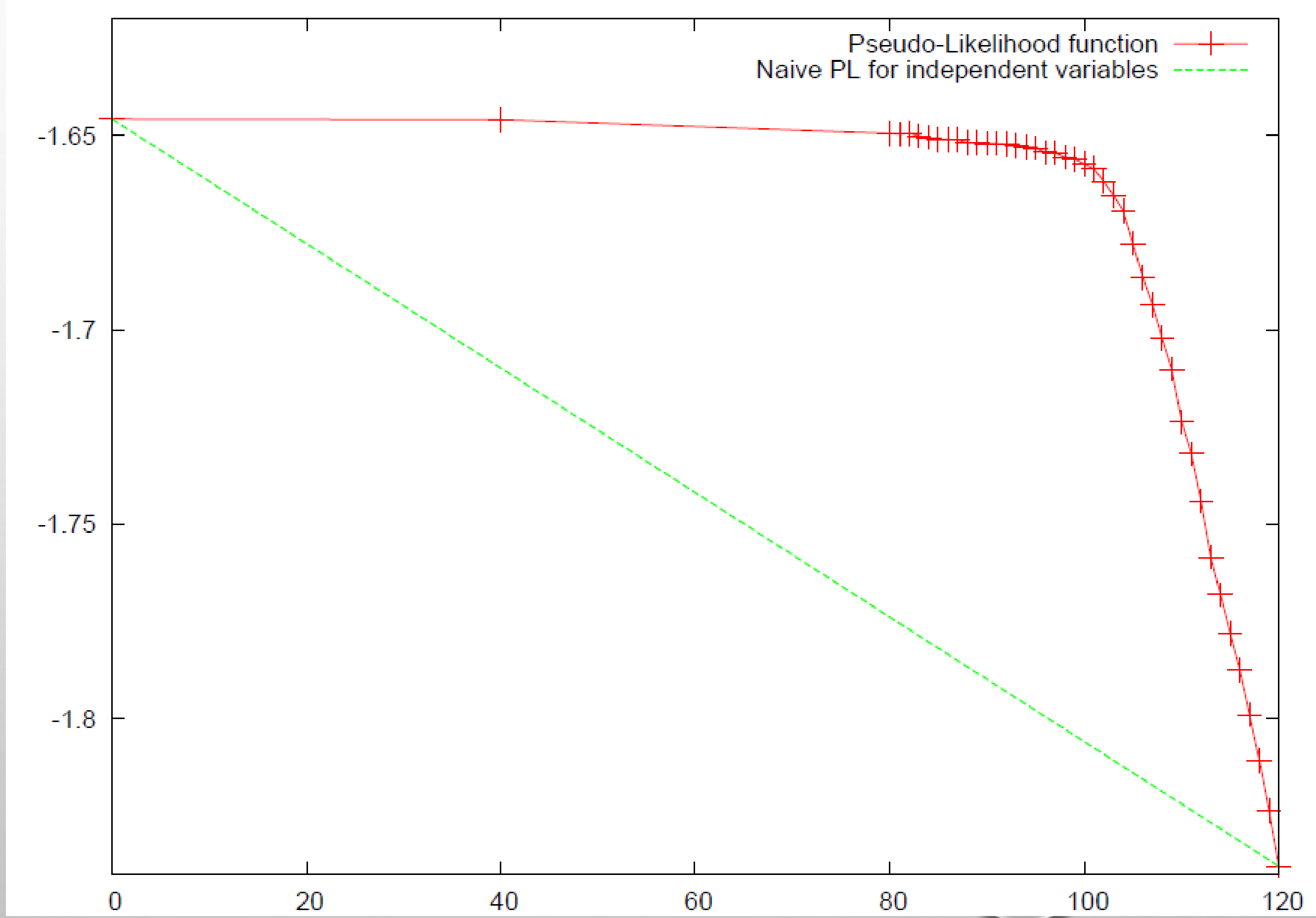
Given a set of equilibrium configurations and all unfixed paramaters

1. Maximize the Pseudo-Likelihood function over all non-fixed variables
2. Decimate the $\rho(t)$ smallest variables (in magnitude) and fixed them
3. If (criterion is reached)
    1. exit
4. Else
    1. $t \leftarrow t + 1$
    2. goto 1.

????

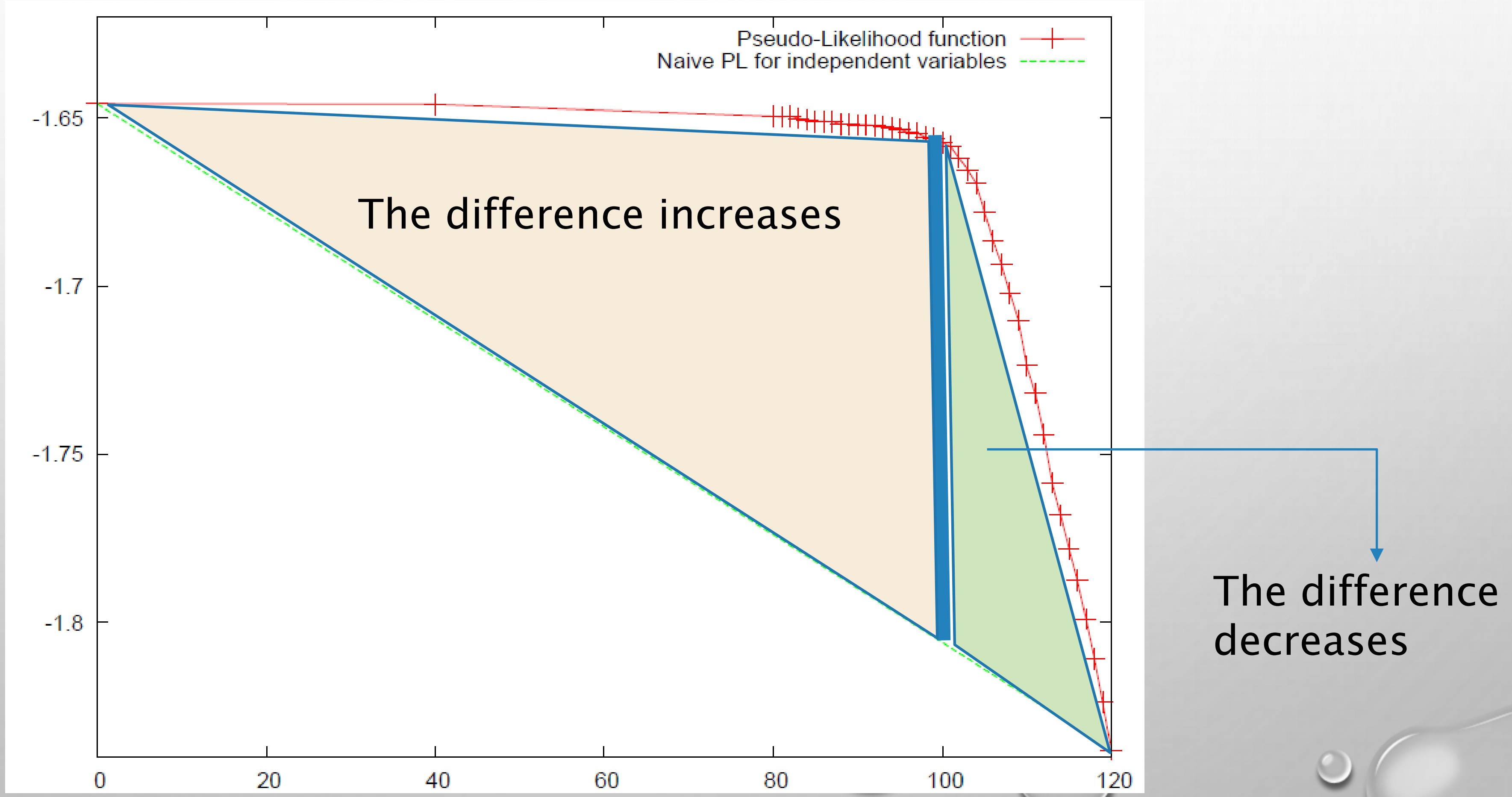
CAN YOU GUESS THE CRITERION ?

Random graph with 16 nodes



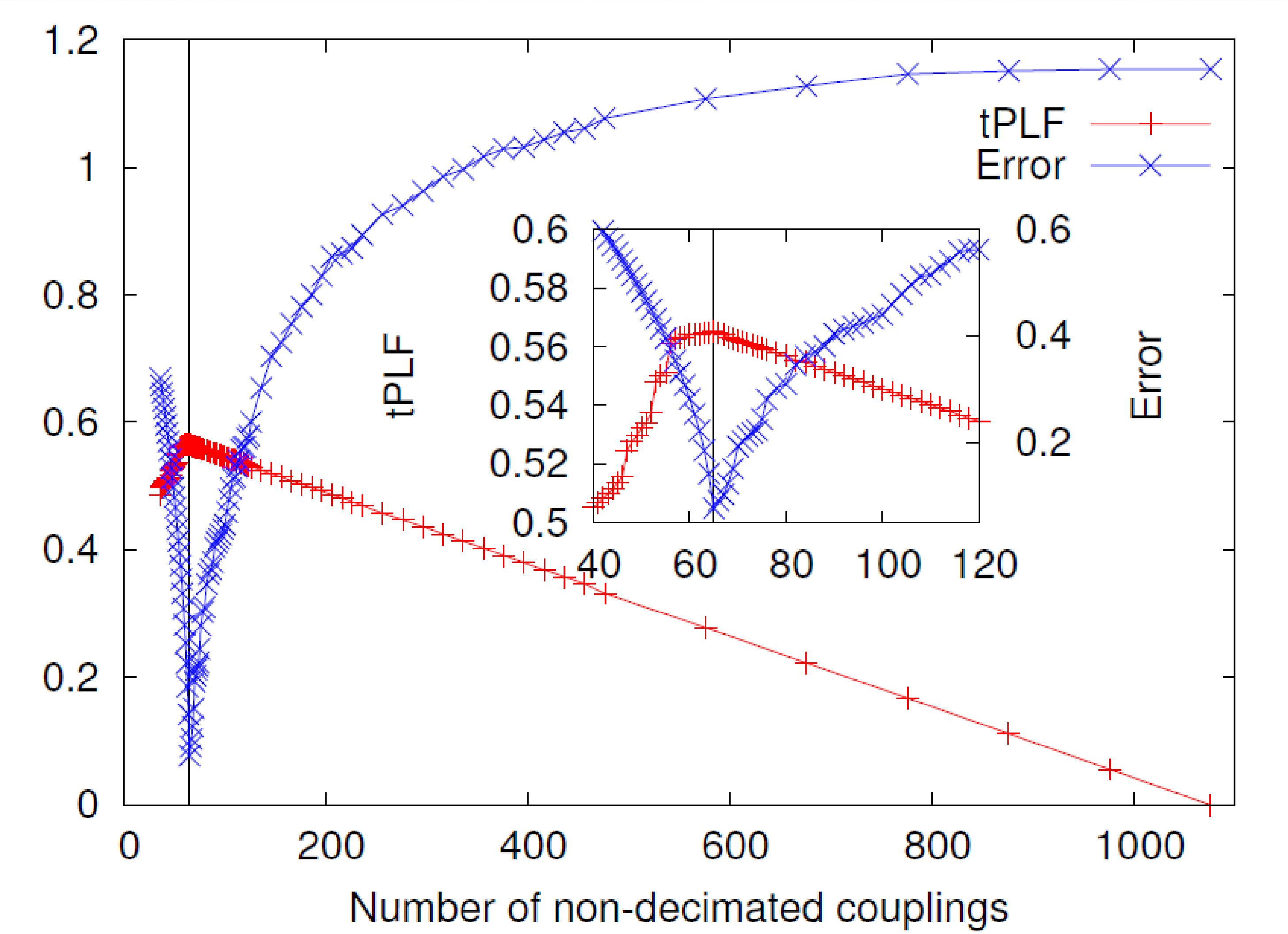
CAN YOU GUESS THE CRITERION ?

Random graph
with 16 nodes

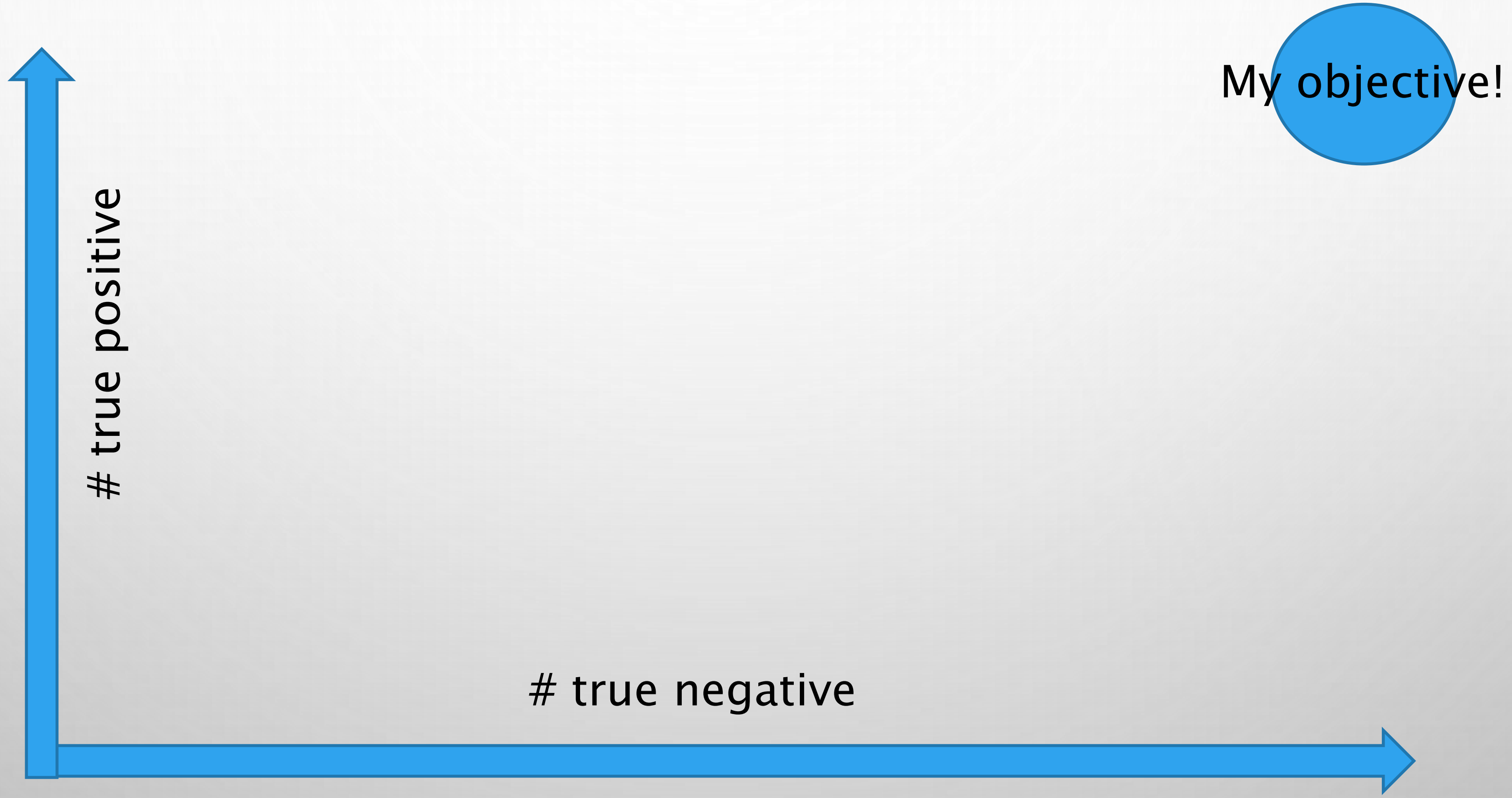


HOW DOES IT LOOK!

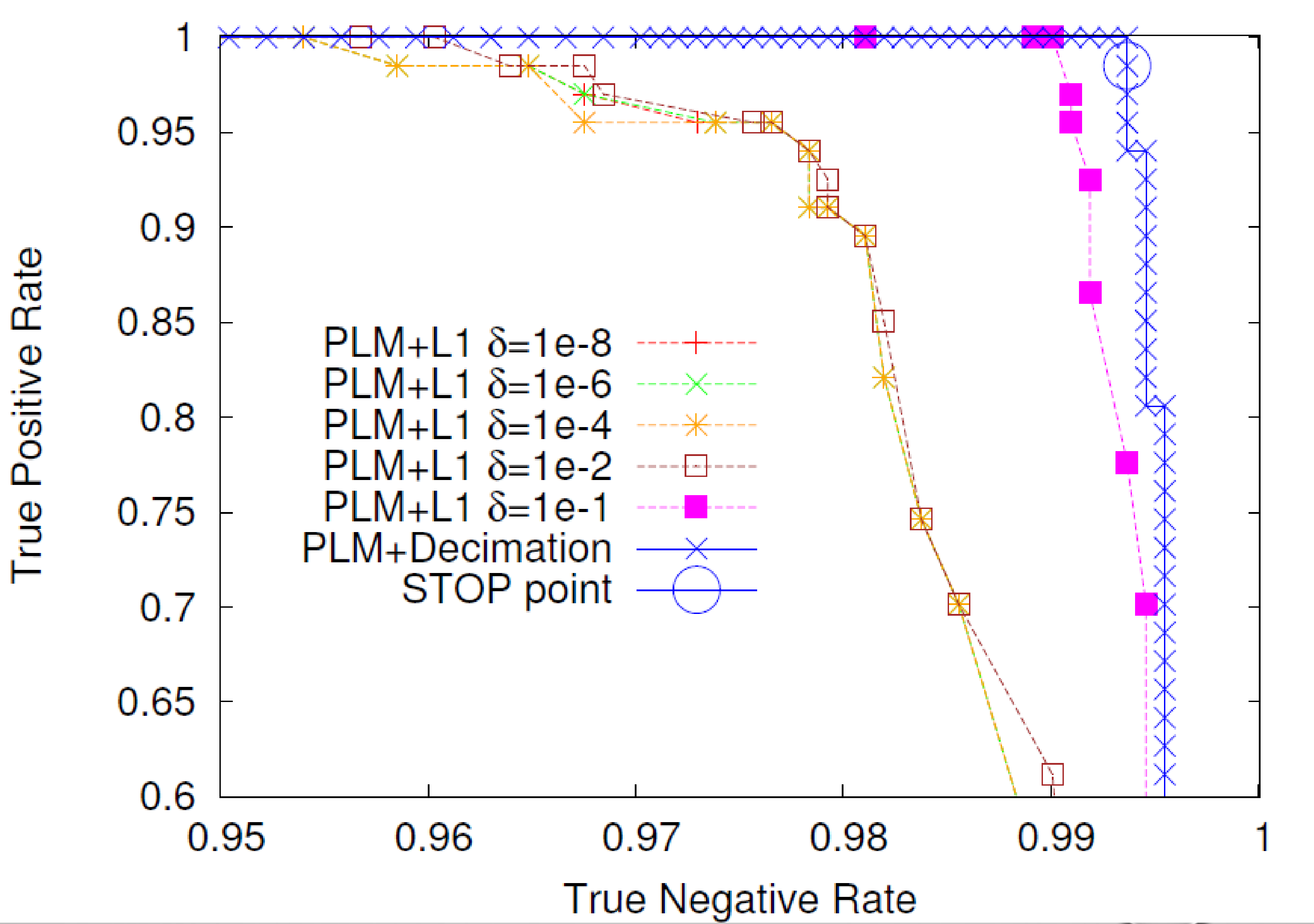
2D ferro model
M=4500
$\boldsymbol{\beta}$=0.8



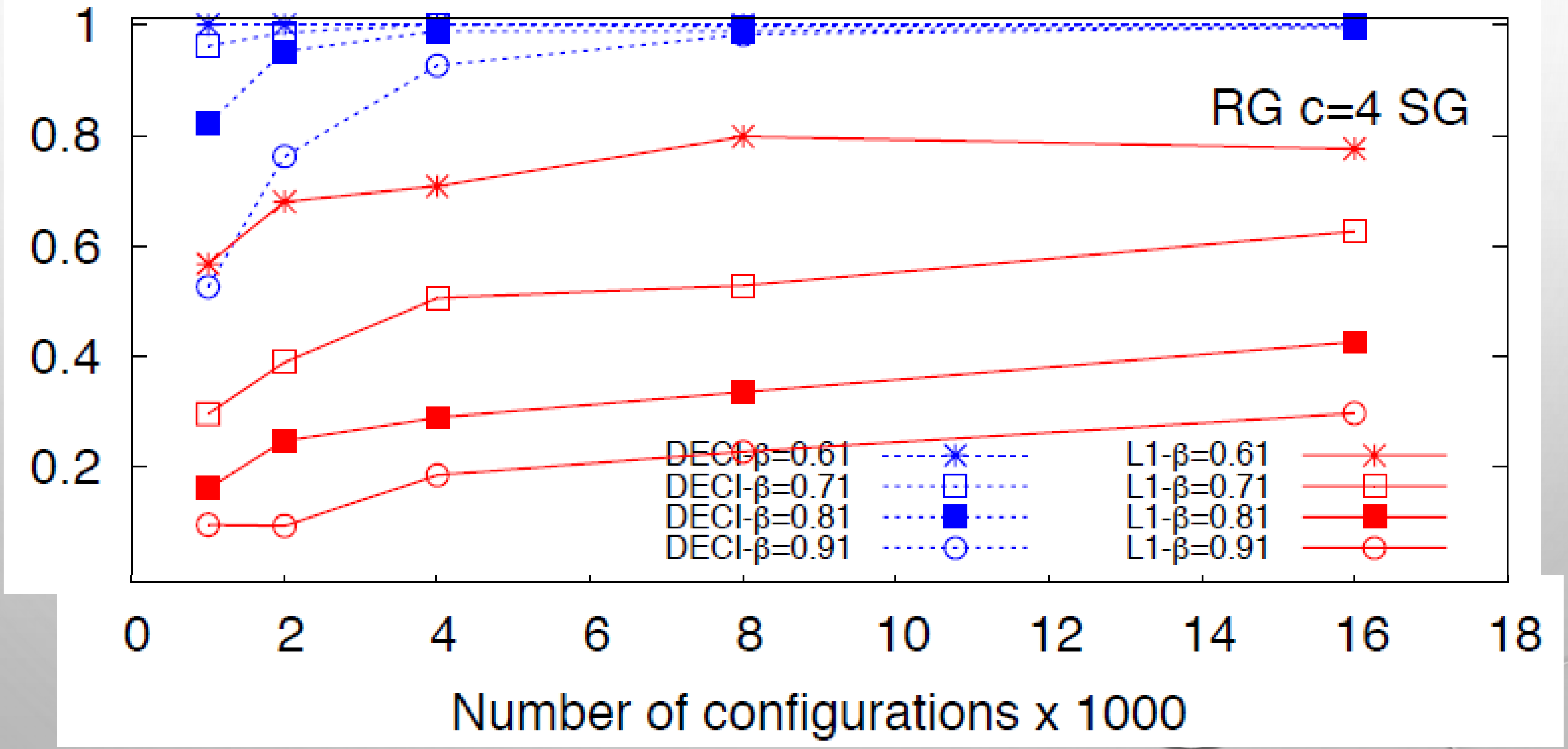
COMPARISON WITH L1 : ROC



COMPARISON WITH L1 : ROC



SOME MORE COMPARISONS (IF TIME)



TO BE CONTINUED …

Can be adapted for the max-likelihood of the parallel dynamics (A.D and P. Zhang)

$$\mathrm{p}(\vec{s}(t+1)|\vec{s}(t)) = \prod_i \frac{e^{-\beta s_i(t+1)(\sum_j J_{ij} s_j(t)+h_i)}}{2\cosh(\beta(\sum_j J_{ij} s_j(t) + h_i)\,)}$$

Has been applied to « detection of cheating by decimation algorithm »
Shogo Yamanaka, Masayuki Ohzeki, A.D.

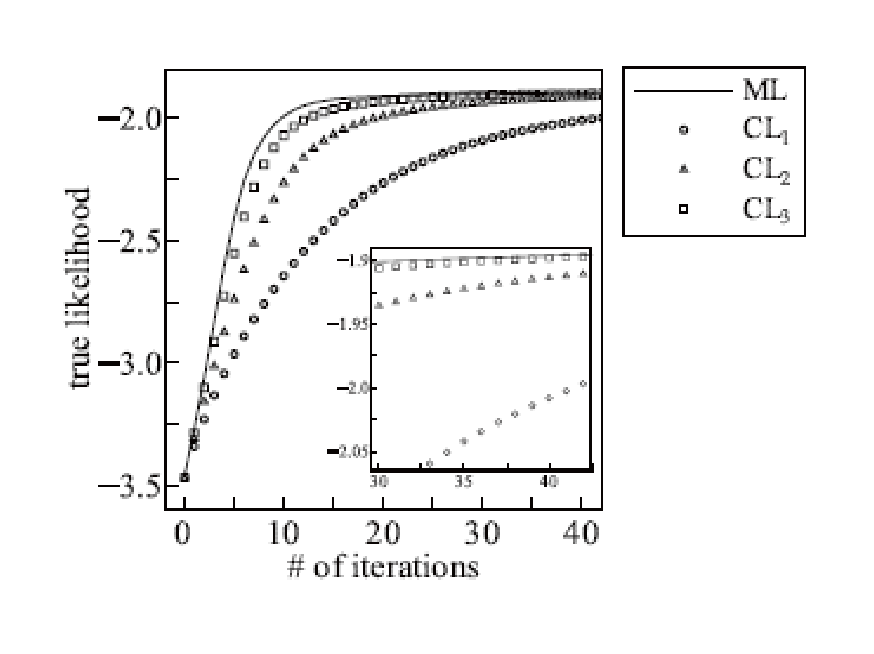
EXTENSION ?

The PLM relies on the evaluation of the one-point marginal, why not use two-points or more ?
"Composite Likelihood Estimation for Restricted Boltzmann machines" by Yasuda et al.

Define $\mathcal{PL}^k = \frac{1}{\#k\text{-}tuples} \sum_{k-uple\ c} \sum_{data} p(\vec{s}_c^{\,(data)} | \vec{s}_{\bar{c}}^{\,(data)})$
They show that

$$\mathcal{PL}^1 \leq \mathcal{PL}^2 \leq \cdots \leq \mathcal{PL}^k \leq \cdots \leq \boxed{\mathcal{PL}^N}$$

True Likelihood !



EXTENSION : THREE-BODY INTERACTIONS

The maximum likelihood can be seen as a maximum entropy problem where we would like to fit the 2-point correlations and local bias !

$$\mathcal{H} = \sum_{i<j} J_{ij} s_i s_j + \sum_i h_i s_i$$

There are already a lot of parameters $O(N^2)$
What if the system « could » have n-body interactions ?

$$\mathcal{H} = \sum_{i<j} J_{ij} s_i s_j + \sum_i h_i s_i + \sum_{i<j<k} J_{ij} s_i s_j s_k + \cdots$$

EXTENSION : THREE-BODY INTERACTIONS

We need to find an **indicator** that there could be new interactions

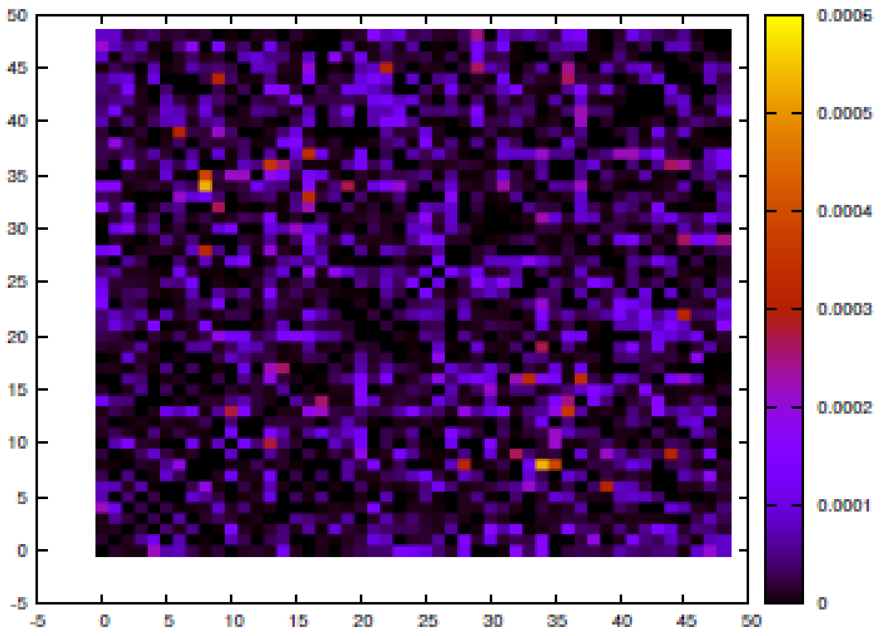
Let's consider the following experience

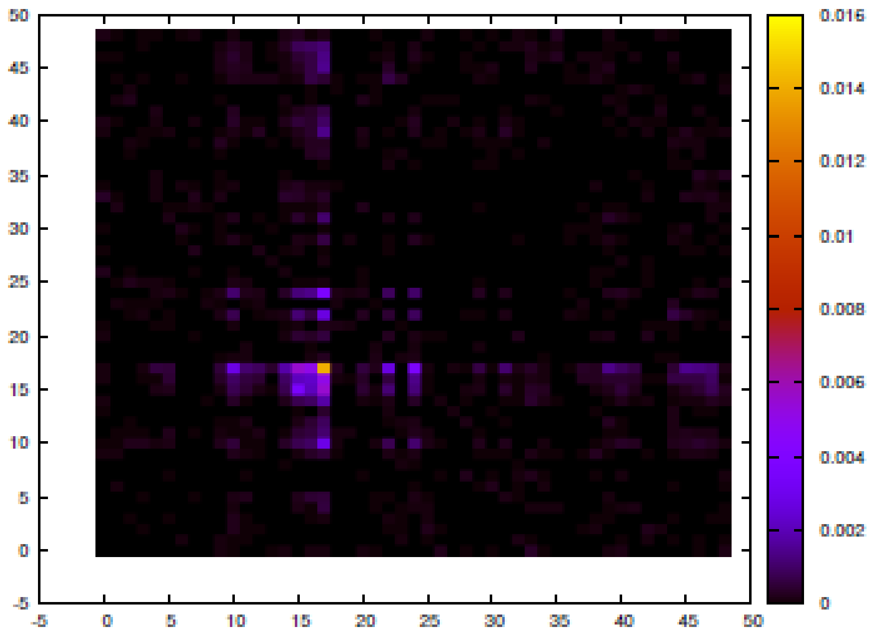
- Take a system S1, 2D ferro without field
- Take a system S2, 2D ferro without field but with some 3B interactions
- Make the inference on the two models with a pairwise model and a model with 3B interactions included

EXTENSION : THREE–BODY INTERACTIONS

Error on the correlation matrix

LEFT : S1 (whatever model I use for inferences)
RIGHT : S2 when doing inference with the wrong model

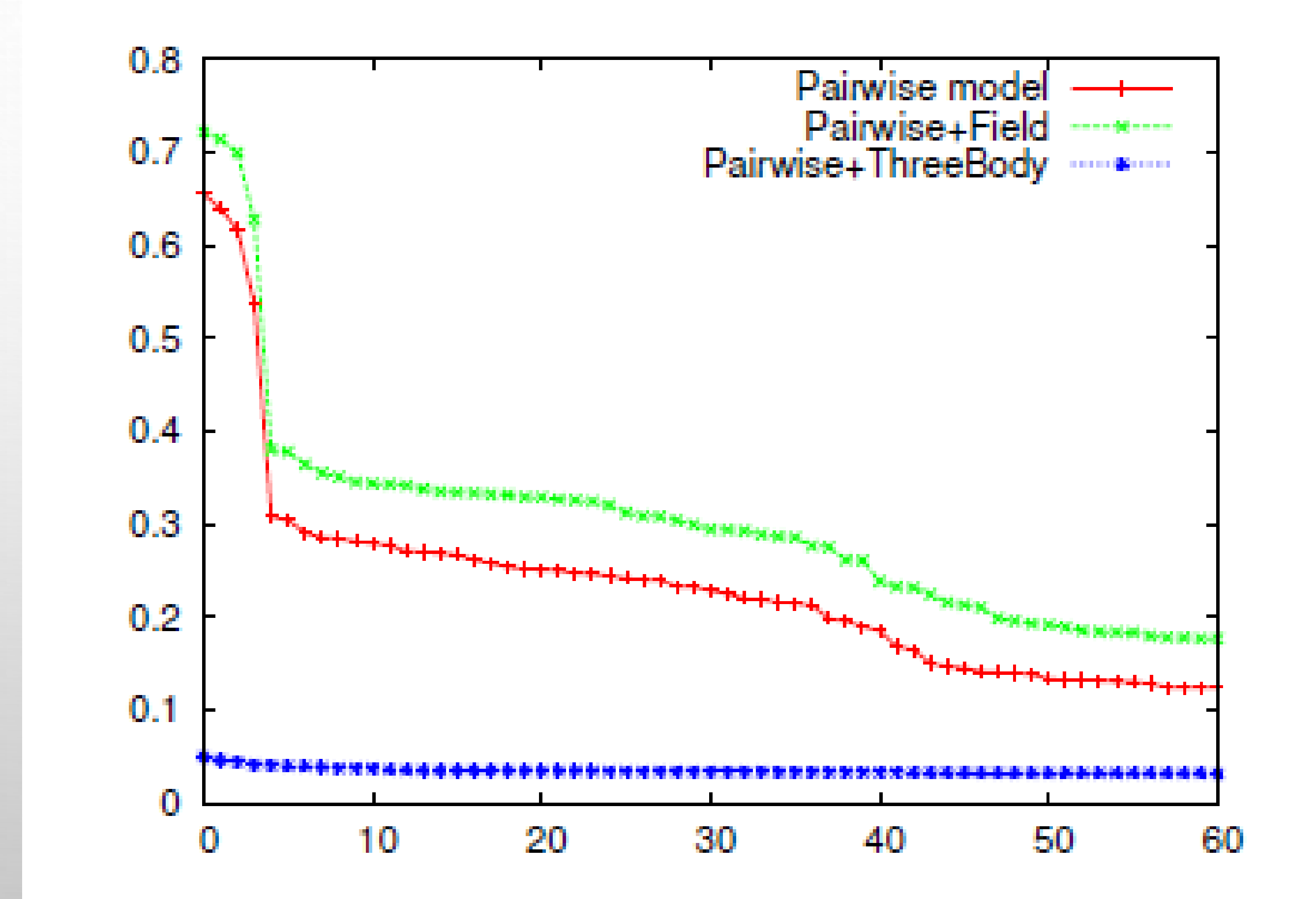




EXTENSION : THREE–BODY INTERACTIONS

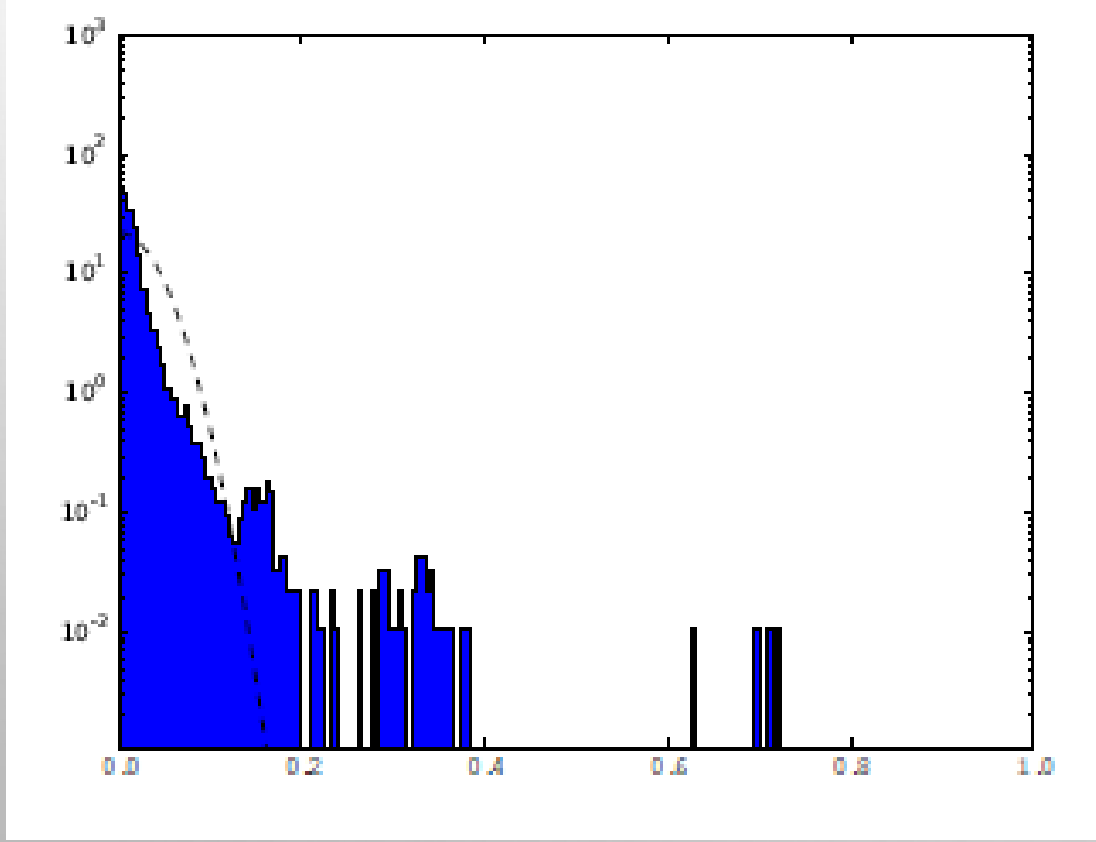
Take the error on the 3points correlation functions, plot them by decreasing order!

Can you guess how many three–body interactions there are ?

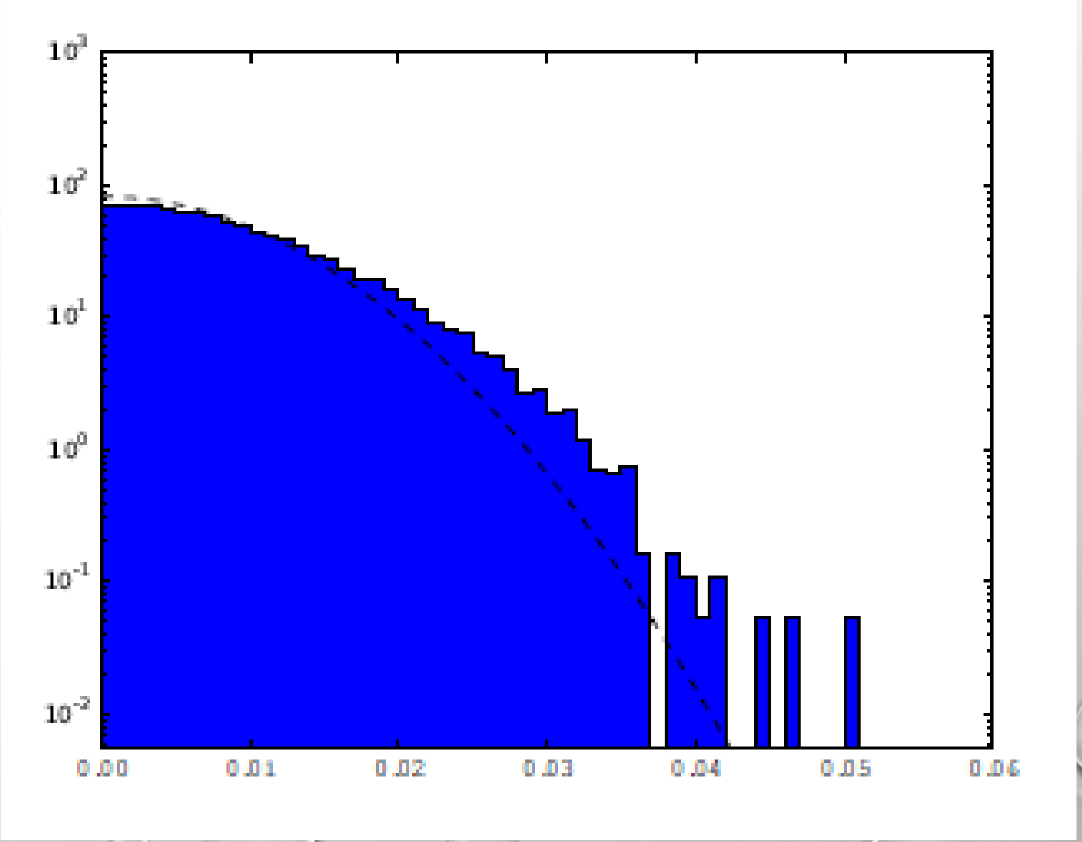


EXTENSION : THREE-BODY INTERACTIONS

– Wrong model –
Histogram of the error on the 3p-corr



– Correct model –
Histogram of the error on the 3p-corr



SUMMARY – CONCLUSION

- Beyond MF method : perform much better on non-trivial topology
  (or strong coupling regime)

- Recovering exact or approximate structure (by Decimation)
  (without the need of fixing parameters)

- Detection many-body interactions inside high order correlations
  « Generalizing » max-ent

As seen : PLM can be extend to become better and better at the cost of complexity!