TAP equations for invariant coupling matrices

Manfred Opper (TUB),
joint work with
B Cakmak (Aalborg), L Bachschmid–Romano (TUB), O Winther (DTU)

February 27, 2017

Outline

- Cavity method for dense Ising networks
- TAP equations for random coupling matrices
- Dynamics of solving TAP equations
- Learning curves for Inverse Ising model

Simplest model: Ising

$$\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_N) \in \{\pm 1\}^N$$

$$P(\boldsymbol{\sigma}) = \frac{1}{Z} \exp\left[\sum_{i<j}^{N} J_{ij}\sigma_i\sigma_j + \sum_{i=1}^{N} H_i\sigma_i\right]$$

Simplest model: Ising

$\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_N) \in \{\pm 1\}^N$

$$P(\boldsymbol{\sigma}) = \frac{1}{Z} \exp\left[\sum_{i<j}^{N} J_{ij}\sigma_i\sigma_j + \sum_{i=1}^{N} H_i\sigma_i\right]$$

Try to compute 'exact mean field equations' marginals for $m_i \doteq < S_i >$ when $\mathbf{J}$ large random matrix $\rightarrow$ TAP equations (Thouless, Anderson and Palmer 78).

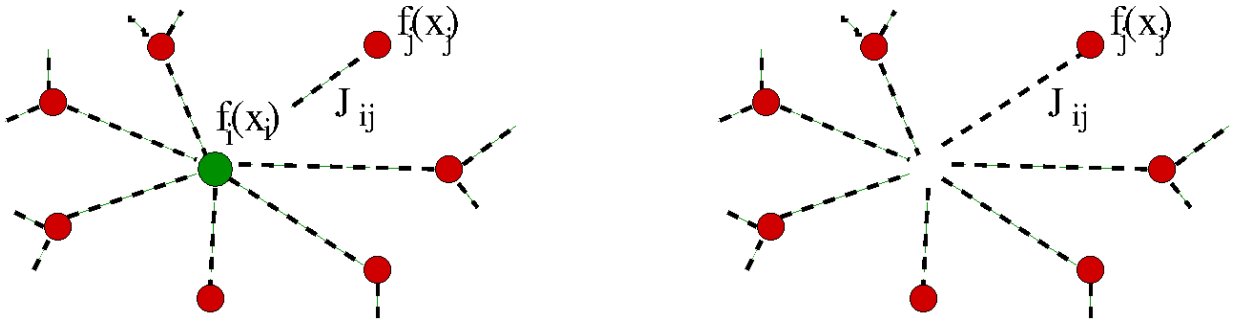
Cavity method

- The marginal at node i can be derived from the joint distribution

$$p_i(\sigma, h) \propto \; e^{\sigma(h+H_i)} \; p_{\setminus i}(h)$$

with the 'cavity field' distribution

$$p_{\setminus i}(h) = \sum_{\boldsymbol{\sigma}\setminus i} \delta\left(h - \sum_{j\neq i} J_{ij}\sigma_j\right) \; p_{\setminus i}(\boldsymbol{\sigma}_{\setminus i})$$



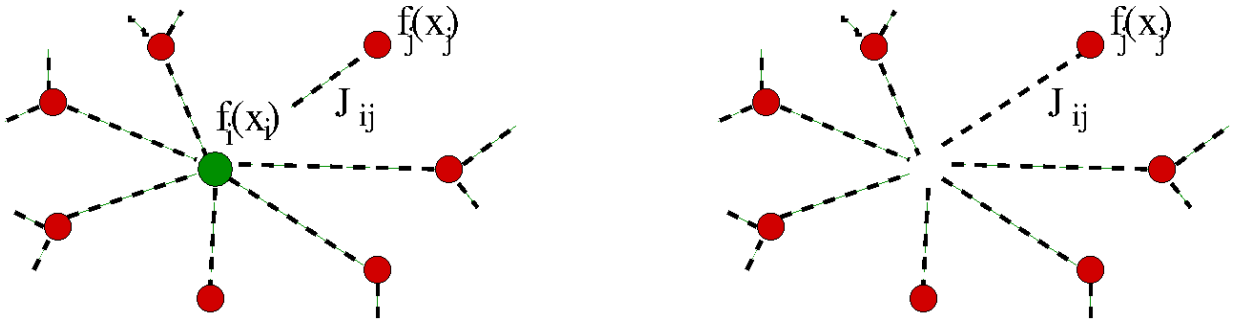
Cavity method

- The marginal at node i can be derived from the joint distribution

$$p_i(\sigma, h) \propto \; e^{\sigma(h+H_i)} \; p_{\setminus i}(h)$$

with the 'cavity field' distribution

$$p_{\setminus i}(h) = \sum_{\boldsymbol{\sigma}\setminus i} \delta\left(h - \sum_{j\neq i} J_{ij}\sigma_j\right) \; p_{\setminus i}(\boldsymbol{\sigma}_{\setminus i})$$



- Approximate $p_{\setminus i}(h)$ by Gaussian (central limit theorem) $p_{\setminus i}(h) = \mathcal{N}(a_i, V_i)$.

TAP equations:

- Magnetisation $m_i = < \sigma_i >$ given by

$$m_i = \tanh\left(H_i + \sum_j J_{ij} m_j - V_i m_i \right)$$

- Susceptibility matrix

$$\boldsymbol{\chi} \doteq \frac{\partial \mathbf{m}}{\partial \mathbf{H}} = (\boldsymbol{\Lambda} - \mathbf{J})^{-1}$$

  where $\lambda_i = V_i + \frac{1}{\chi_{ii}}$.

- V_i determined by consistency $\chi_{ii} = 1 - m_i^2$

Selfaveraging cavity variance

- Random matrix assumption $\mathbf{J} \sim \mathbf{O}^\top \mathbf{J}\mathbf{O}$ (Parisi and Potters 95)

Selfaveraging cavity variance

- Random matrix assumption $\mathbf{J} \sim \mathbf{O}^\top \mathbf{J}\mathbf{O}$ (Parisi and Potters 95)
- If $V \equiv \mathrm{R}_{\mathbf{J}}(\bar{\chi})$ with $\bar{\chi} = \frac{1}{N}\sum_i \chi_{ii}$ one can show

$$\left[(\mathbf{\Lambda} - \mathbf{J})^{-1}\right]_{ii} \to \chi_{ii}$$

as $N \to \infty$.

Selfaveraging cavity variance

- Random matrix assumption $\mathbf{J} \sim \mathbf{O}^\top \mathbf{J}\mathbf{O}$ (Parisi and Potters 95)
- If $V \equiv \mathrm{R}_{\mathbf{J}}(\bar{\chi})$ with $\bar{\chi} = \frac{1}{N}\sum_i \chi_{ii}$ one can show

$$\left[(\mathbf{\Lambda} - \mathbf{J})^{-1}\right]_{ii} \to \chi_{ii}$$

  as $N \to \infty$.
- The R–transform is defined as

$$\mathrm{R}_{\mathbf{J}}(s) \triangleq \mathrm{S}_{\mathbf{J}}^{-1}(s) - 1/s$$

  where

$$\mathrm{S}_{\mathbf{J}}(z) \triangleq \frac{1}{N}\mathrm{Tr}\,(z\mathbf{I} - \mathbf{J})^{-1}$$

Algorithms for solving TAP equations ?

- Analysis of dynamics of random systems usually hard !

Algorithms for solving TAP equations ?

- Analysis of dynamics of random systems usually hard !
- Exact dynamics converging to TAP For SK model by Bolthausen (2014).
- Other results by Kabashima (2003) in the context of the CDMA and by Donoho, Maleki, Montanari (2009) for compressed sensing.

Generating functional approach

- Candidate algorithm could be of the form

$$\mathbf{m}(t) = \tanh\left(\{\mathbf{h}(\tau), \mathbf{m}(\tau)\}_{\tau=0}^{t-1}\right)$$
$$\mathbf{h}(t) = \mathbf{H} + \mathbf{J}\mathbf{m}(t)$$

Generating functional approach

- Candidate algorithm could be of the form

$$\mathbf{m}(t) = \tanh\left(\{\mathbf{h}(\tau), \mathbf{m}(\tau)\}_{\tau=0}^{t-1}\right)$$
$$\mathbf{h}(t) = \mathbf{H} + \mathbf{J}\mathbf{m}(t)$$

- Try average case analysis: use generating functional $\langle Z(\{\mathbf{I}(t)\})\rangle_{\mathbf{J}}$ where

$$Z(\{\mathbf{I}(t)\}) = \int \prod_{t=0}^{T-1} \left\{ \mathrm{d}\mathbf{m}(t)\mathrm{d}\mathbf{h}(t)\, \delta(\mathbf{m}(t) - \tanh\left(\{\mathbf{h}(\tau), \mathbf{m}(\tau)\}_{\tau=0}^{t-1}\right)) \right.$$
$$\left. \delta(\mathbf{h}(t) - \mathbf{H} - \mathbf{J}\mathbf{m}(t)) e^{i\mathbf{h}(t)^\top \mathbf{I}(t)} \right\}.$$

Effective stochastic dynamics with memory

$$\mathbf{m}(t) = \tanh\left(\{\mathbf{h}(\tau), \mathbf{m}(\tau)\}_{\tau=0}^{t-1}\right)$$

$$\mathbf{h}(t) = \mathbf{H} + \sum_{s=0}^{t-1} \hat{\mathcal{G}}(t,s)\mathbf{m}(s) + \phi(t) \ .$$

with $\phi(t)$ discrete time Gaussian process and order parameters

Effective stochastic dynamics with memory

$$\mathbf{m}(t) = \tanh\left(\{\mathbf{h}(\tau), \mathbf{m}(\tau)\}_{\tau=0}^{t-1}\right)$$

$$\mathbf{h}(t) = \mathbf{H} + \sum_{s=0}^{t-1} \hat{\mathcal{G}}(t,s)\mathbf{m}(s) + \boldsymbol{\phi}(t) \ .$$

with $\boldsymbol{\phi}(t)$ discrete time Gaussian process and order parameters

$$\hat{\mathcal{G}} = \mathrm{R}_{\mathbf{J}}(\mathcal{G}) \qquad \mathcal{G}(t,\tau) = \left\langle \frac{\partial m_i(t)}{\partial \phi_i(\tau)} \right\rangle_{\phi}$$

and something more complicated for C_ϕ.

Effective stochastic dynamics with memory

$$\mathbf{m}(t) = \tanh\left(\{\mathbf{h}(\tau), \mathbf{m}(\tau)\}_{\tau=0}^{t-1}\right)$$

$$\mathbf{h}(t) = \mathbf{H} + \sum_{s=0}^{t-1} \hat{\mathcal{G}}(t,s)\mathbf{m}(s) + \boldsymbol{\phi}(t) \ .$$

with $\boldsymbol{\phi}(t)$ discrete time Gaussian process and order parameters

$$\hat{\mathcal{G}} = \mathrm{R}_{\mathbf{J}}(\mathcal{G}) \qquad \mathcal{G}(t,\tau) = \left\langle \frac{\partial m_i(t)}{\partial \phi_i(\tau)} \right\rangle_{\phi}$$

and something more complicated for C_ϕ.
Memory could be bad for convergence !

Single step memory algorithm

- **Idea:** Try to **erase the memory** terms

$$\hat{\mathcal{G}}(t,\tau) = 0 \qquad \forall \tau \neq t-1$$

Single step memory algorithm

- **Idea:** Try to **erase the memory** terms

$$\hat{\mathcal{G}}(t,\tau) = 0 \qquad \forall \tau \neq t-1$$

- and subtract: $\mathbf{J}\mathbf{m}(\tau) - \hat{\mathcal{G}}(\tau, \tau-1)\mathbf{m}(\tau-1)$.

Single step memory algorithm

- **Idea:** Try to **erase the memory** terms

$$\hat{\mathcal{G}}(t,\tau) = 0 \qquad \forall \tau \neq t-1$$

- and subtract: $\mathbf{Jm}(\tau) - \hat{\mathcal{G}}(\tau,\tau-1)\mathbf{m}(\tau-1)$. Consistency with definition of $G(t,\tau)$ and TAP asymptotics suggests

$$\mathbf{m}(t+1) = \tanh(\psi(t))$$

$$\psi(t) = Q(t)\sum_{\tau=0}^{t} a_{t+1-\tau}\frac{\mathbf{H} + \mathbf{Jm}(\tau) - \hat{\mathcal{G}}(\tau,\tau-1)\mathbf{m}(\tau-1)}{Q(\tau-1)(1-q(\tau))}$$

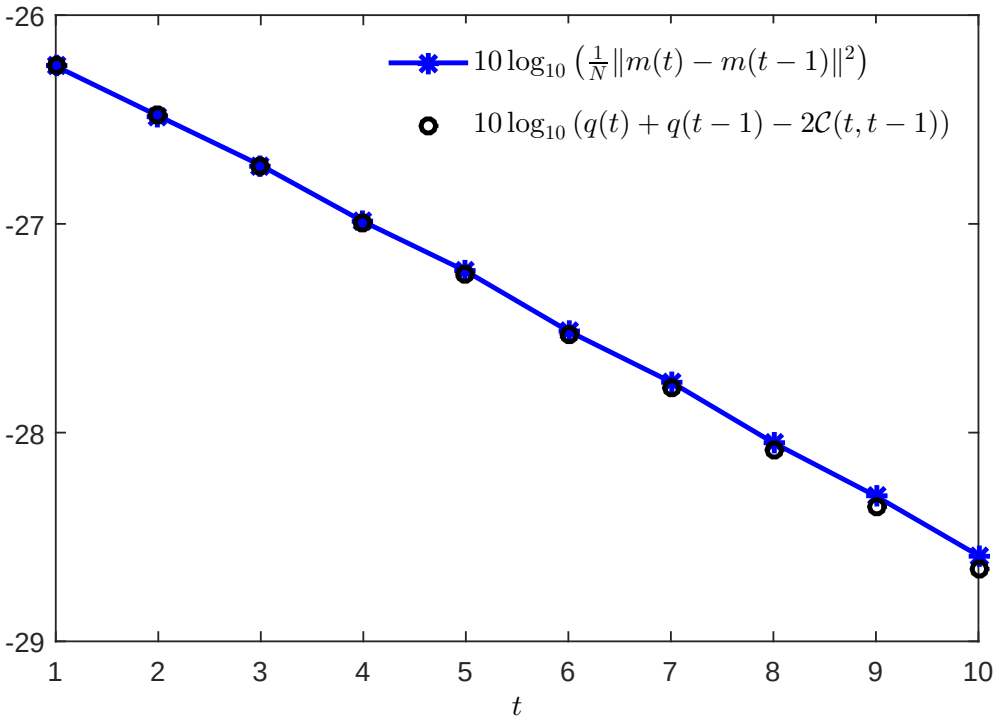
where we define $\hat{\mathcal{G}}(t,t-1) = \frac{1-q(t)}{1-q(t-1)}\mathrm{R}_{\mathbf{J}}(1-q(t-1))$
$Q(t) = \prod_{\tau=0}^{t}\mathrm{R}_{\mathbf{J}}(1-q(\tau))$ and the coefficients a_k via
$\mathrm{R}^{-1}(x) = \sum_{n=1}^{\infty} a_n x^n$.

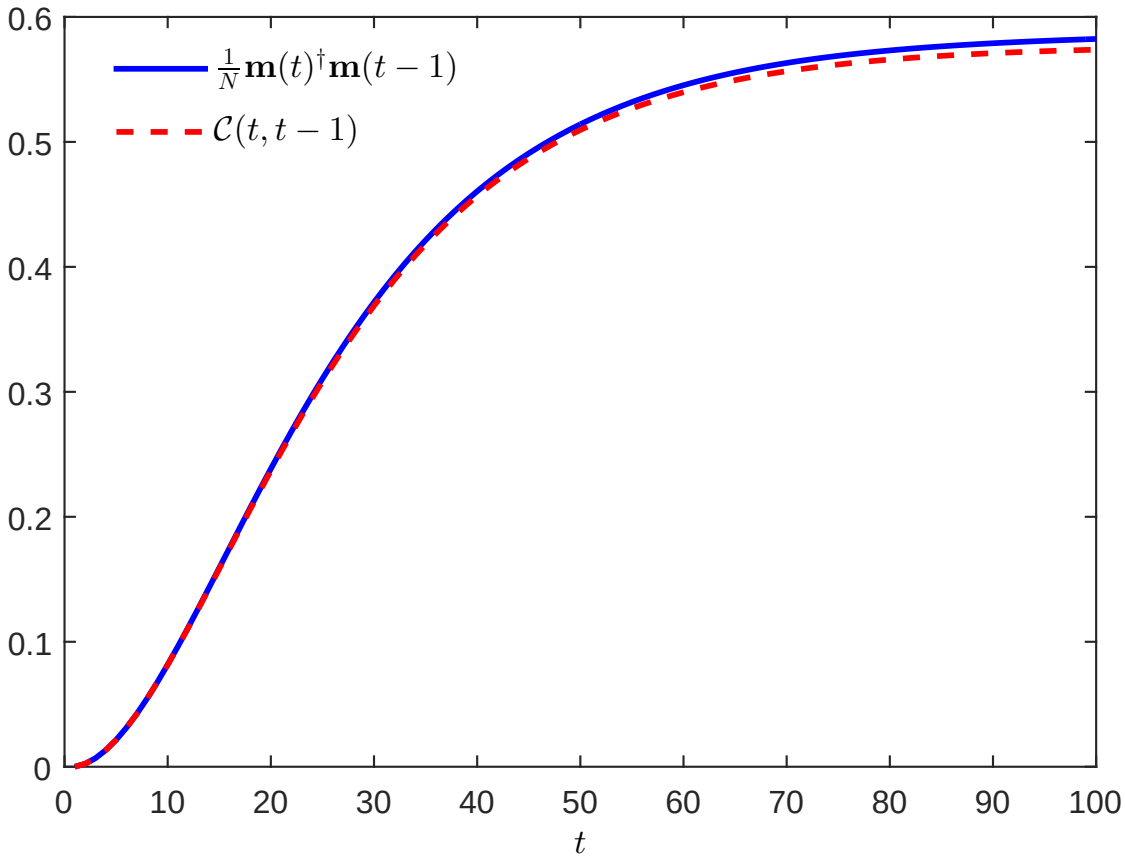
Examples

- For $\mathbf{J}$ Gaussian i.i.d. (Sherrington Kirkpatrick model) agrees with Bolthausen's (2014) result
- For $-\mathbf{J} \sim$ Wishart coincides with AMP algorithm introduced by Kabashima (2003) in the context of the CDMA and by Donoho, Maleki, Montanari (2009) for compressed sensing.

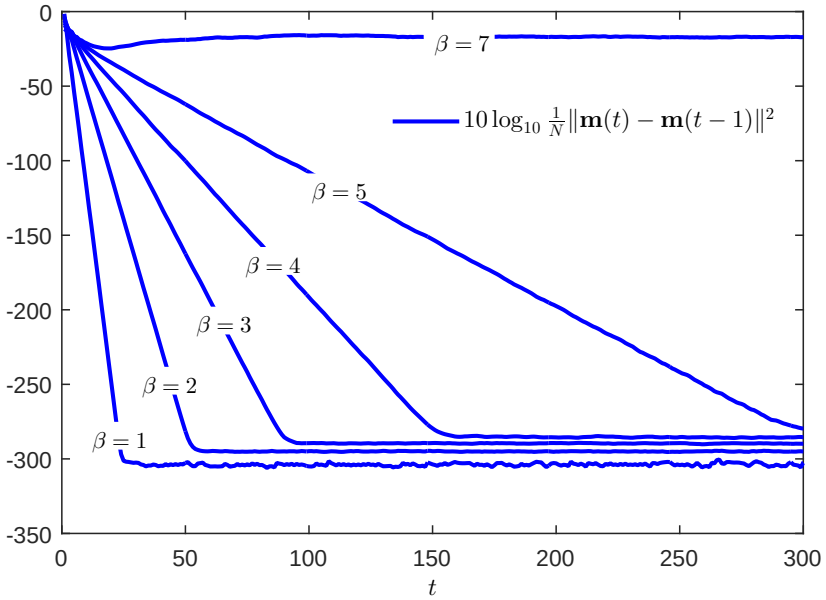
Random orthogonal ensemble: Analytical results vs Simulation

$\mathrm{R}^{-1}(x) = x/(\beta^2 - x^2)$ and $a_n = \frac{1}{\beta^{n+1}}$ for n odd and $a_n = 0$ else.
$N = 2^{14}$, $\beta = 20$ and $h_i = 1$, single realisation of $\mathbf{J}$.

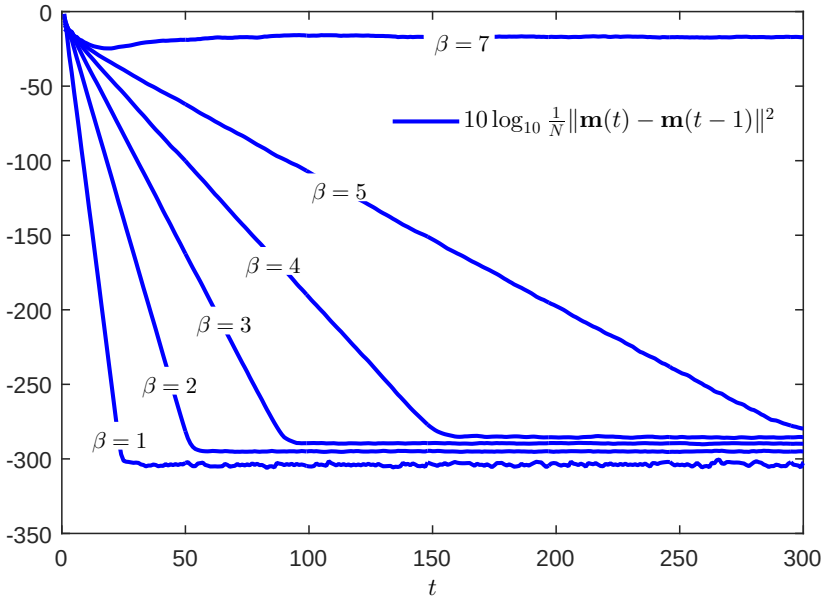



0.6
0.5
0.4
0.3
0.2
0.1
0
0
10
20
30
40
50
60
70
80
90
100
t
$\frac{1}{N}\mathbf{m}(t)^\dagger\mathbf{m}(t-1)$
$\mathcal{C}(t,t-1)$

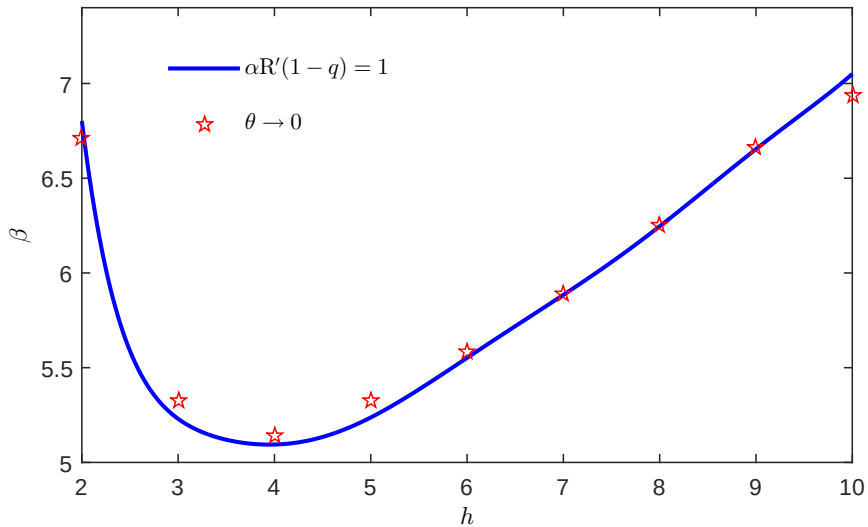
Convergence for $N = 2^{14}$, $H_i = 2$ (Simulations)



Convergence for $N = 2^{14}$, $H_i = 2$ (Simulations)

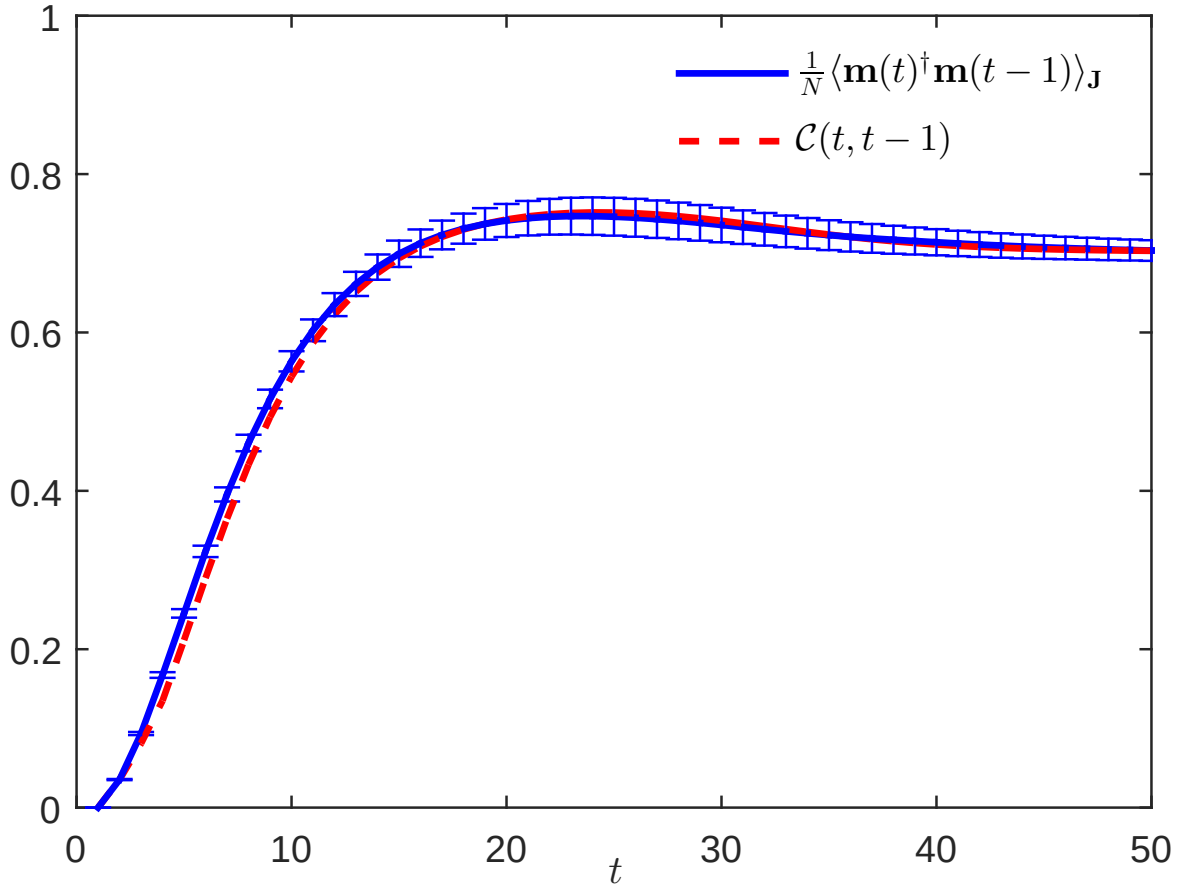


Stability of fixed point (Almeida–Thouless line)



Analytical results vs Simulation

Random orthogonal ensemble, region of instability: $N = 2^{12}$, $\beta = 10$ and $H_i = 2$, 5×10^3 Realisations of $\mathbf{J}$



Inverse Ising problem

- Learn couplings from Ising data $\boldsymbol{\sigma}^1, \ldots, \boldsymbol{\sigma}^M$ generated from

$$P(\boldsymbol{\sigma}|\mathbf{J}) = Z_{\text{Ising}}^{-1} \exp\left[\beta \sum_{i<j} J_{ij}\sigma_i\sigma_j\right]$$

Inverse Ising problem

- Learn couplings from Ising data $\boldsymbol{\sigma}^1, \ldots, \boldsymbol{\sigma}^M$ generated from

$$P(\boldsymbol{\sigma}|\mathbf{J}) = Z_{\text{Ising}}^{-1} \exp\left[\beta \sum_{i<j} J_{ij}\sigma_i\sigma_j\right]$$

- Local approach: $\mathbf{W} = (W_1, \ldots, W_N) \doteq \sqrt{N}(J_{01}, \ldots, J_{0N-1})$ for **spin 0** using cost function

$$E(\mathbf{W}) = \sum_{k=1}^{M} \mathcal{E}(\mathbf{W}; \boldsymbol{\sigma}^k)$$

where $\mathcal{E}(\mathbf{W}; \boldsymbol{\sigma}) = \Phi(\sigma_0, h)$ and $h \doteq \frac{1}{\sqrt{N}} \sum_{j\neq 0} W_j\sigma_j$

Inverse Ising problem

- Learn couplings from Ising data $\boldsymbol{\sigma}^1, \ldots, \boldsymbol{\sigma}^M$ generated from

$$P(\boldsymbol{\sigma}|\mathbf{J}) = Z_{\text{Ising}}^{-1} \exp\left[\beta \sum_{i<j} J_{ij}\sigma_i\sigma_j\right]$$

- Local approach: $\mathbf{W} = (W_1, \ldots, W_N) \doteq \sqrt{N}(J_{01}, \ldots, J_{0N-1})$ for **spin** 0 using cost function

$$E(\mathbf{W}) = \sum_{k=1}^{M} \mathcal{E}(\mathbf{W}; \boldsymbol{\sigma}^k)$$

where $\mathcal{E}(\mathbf{W}; \boldsymbol{\sigma}) = \Phi(\sigma_0, h)$ and $h \doteq \frac{1}{\sqrt{N}} \sum_{j\neq 0} W_j\sigma_j$

- Example: **pseudo–likelihood**

$$\mathcal{E}(\mathbf{W}; \boldsymbol{\sigma}) = -\ln P(\sigma_0|\boldsymbol{\sigma}_{\setminus 0}, \mathbf{W})$$

Average case learning performance

- Define partition function of student couplings

$$Z = \int d\mathbf{W} \exp[-\nu E(\mathbf{W})],$$

- Order parameters from $\nu \to \infty$ limit of quenched free energy

$$F = -N^{-1}\nu^{-1}\overline{\ln Z} = -\lim_{n\to 0} N^{-1}\nu^{-1}\frac{\partial}{\partial n}\ln \overline{Z^n},$$

Average case learning performance

- Define partition function of student couplings

$$Z = \int d\mathbf{W} \exp[-\nu E(\mathbf{W})],$$

- Order parameters from $\nu \to \infty$ limit of quenched free energy

$$F = -N^{-1}\nu^{-1}\overline{\ln Z} = -\lim_{n\to 0} N^{-1}\nu^{-1}\frac{\partial}{\partial n}\ln \overline{Z^n},$$

- Quenched averages over the fields performed using **cavity method**:

$$P(\sigma_0, h_*, h_1, \ldots h_n) = \frac{1}{Z_0} e^{\beta\sigma_0 h_*} p_{\text{cav}}(h_*, h_1, \ldots h_n)$$

where $p_{\text{cav}}(h_*, h_1, \ldots h_n)$ is a **multivariate Gaussian** with covariance

$$\langle h_a h_b \rangle = \frac{1}{N}\sum_{i,j\neq 0} W_j^a C_{ij}^{\setminus 0} W_j^b \qquad C_{ij}^{\setminus 0} = \langle \sigma_i \sigma_j \rangle_{\setminus 0}$$

- Reconstruction error

$$\varepsilon = N^{-1}\overline{(\mathbf{W}^* - \mathbf{W})^2} = (q - \frac{R^2}{V})T + Y(1 - \frac{R}{V})^2.$$

depends on the teacher statistics

$$\begin{aligned} Y &\doteq N^{-1}(\mathbf{W}^*)^2 \\ V &\doteq \frac{1}{N}\sum_{i,j\neq 0} W_j^* C_{ij}^{\setminus 0} W_j^* \\ T &\doteq \frac{1}{N}\mathrm{Tr}\mathbf{C}^{-1} = \beta^2 V + 1 \end{aligned}$$

- q and R are derived from the free energy ($\alpha = M/N$):

- Reconstruction error

$$\varepsilon = N^{-1}\overline{(\mathbf{W}^* - \mathbf{W})^2} = (q - \frac{R^2}{V})T + Y(1 - \frac{R}{V})^2.$$

depends on the teacher statistics

$$\begin{aligned} Y &\doteq N^{-1}(\mathbf{W}^*)^2 \\ V &\doteq \frac{1}{N}\sum_{i,j\neq 0} W_j^* C_{ij}^{\backslash 0} W_j^* \\ T &\doteq \frac{1}{N}\mathrm{Tr}\mathbf{C}^{-1} = \beta^2 V + 1 \end{aligned}$$

- q and R are derived from the free energy ($\alpha = M/N$):

$$F = -\operatorname*{extr}_{q,R,x} \max_z \left\{ \frac{1}{2}\frac{q - R^2/V}{x} + \frac{1}{2}\alpha\, e^{-\frac{\beta^2 R^2}{2q}} \sum_{\sigma_0} \int \mathcal{D}u\, e^{-\frac{\beta\sigma_0 R}{\sqrt{q}}u} \left[-\frac{z^2}{2} - \Phi(\sigma_0, \sqrt{x}z + \sqrt{q}u) \right] \right\} .$$

Maximum Likelihood (mean field)

- Definition

$$J_{ij}^{ML-MF} = -\frac{1}{\beta}(\hat{\mathbf{C}}^{-1})_{ij}$$

Maximum Likelihood (mean field)

- Definition

$$J_{ij}^{ML-MF} = -\frac{1}{\beta}(\hat{\mathbf{C}}^{-1})_{ij}$$

- Using matrix inversion lemma

$$W_i^{ML-MF} = \sqrt{N} J_{0i}^{ML-MF} = \frac{\sqrt{N}}{\beta}\phi_0 \sum_{j\neq 0}(\hat{C}^{-1})_{ji}\hat{C}_{0j}$$

where $\phi_0 = \frac{1}{1-\sum_{i,j\neq 0}\hat{C}_{0i}(\hat{C}^{-1})_{ij}\hat{C}_{0j}}$ is selfaveraging.

Maximum Likelihood (mean field)

- Definition

$$J_{ij}^{ML-MF} = -\frac{1}{\beta}(\hat{\mathbf{C}}^{-1})_{ij}$$

- Using matrix inversion lemma

$$W_i^{ML-MF} = \sqrt{N} J_{0i}^{ML-MF} = \frac{\sqrt{N}}{\beta}\phi_0 \sum_{j\neq 0}(\hat{C}^{-1})_{ji}\hat{C}_{0j}$$

where $\phi_0 = \frac{1}{1-\sum_{i,j\neq 0}\hat{C}_{0i}(\hat{C}^{-1})_{ij}\hat{C}_{0j}}$ is selfaveraging.

- W_i^{ML-MF} minimises local cost function

$$E(\mathbf{W}) = \frac{\alpha\beta}{2\phi_0}\sum_{i,j\neq 0} W_i \hat{C}_{ij} W_j - \alpha\sqrt{N}\sum_{j\neq 0}\hat{C}_{0j}W_j,$$

Maximum Likelihood (mean field)

- Definition

$$J_{ij}^{ML-MF} = -\frac{1}{\beta}(\hat{\mathbf{C}}^{-1})_{ij}$$

- Using matrix inversion lemma

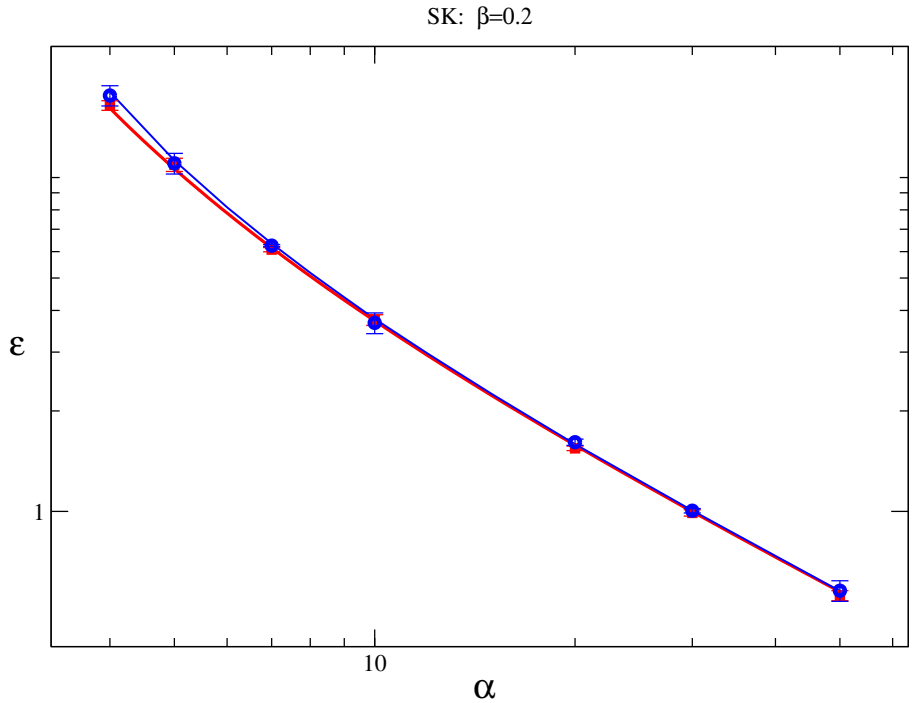
$$W_i^{ML-MF} = \sqrt{N} J_{0i}^{ML-MF} = \frac{\sqrt{N}}{\beta}\phi_0 \sum_{j\neq 0}(\hat{C}^{-1})_{ji}\hat{C}_{0j}$$

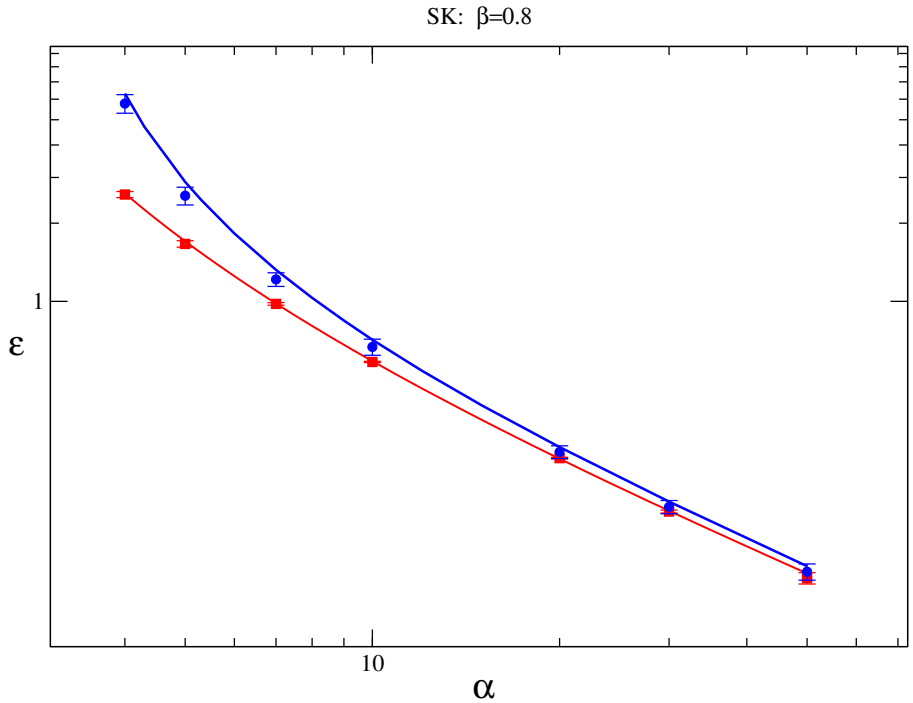
where $\phi_0 = \frac{1}{1-\sum_{i,j\neq 0}\hat{C}_{0i}(\hat{C}^{-1})_{ij}\hat{C}_{0j}}$ is selfaveraging.

- W_i^{ML-MF} minimises local cost function

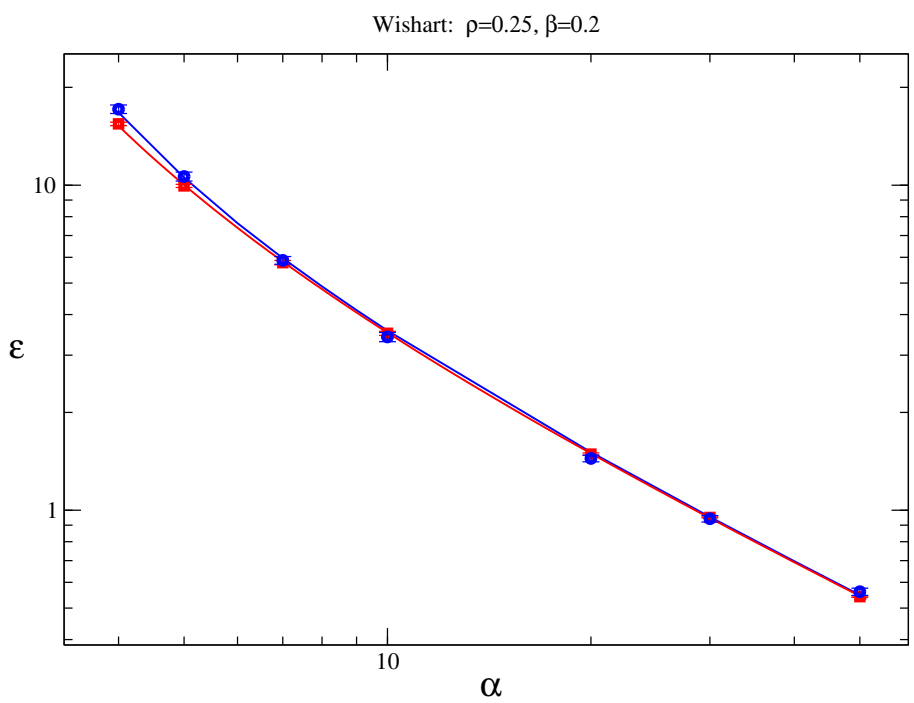
$$E(\mathbf{W}) = \frac{\alpha\beta}{2\phi_0}\sum_{i,j\neq 0} W_i \hat{C}_{ij} W_j - \alpha\sqrt{N}\sum_{j\neq 0}\hat{C}_{0j}W_j,$$

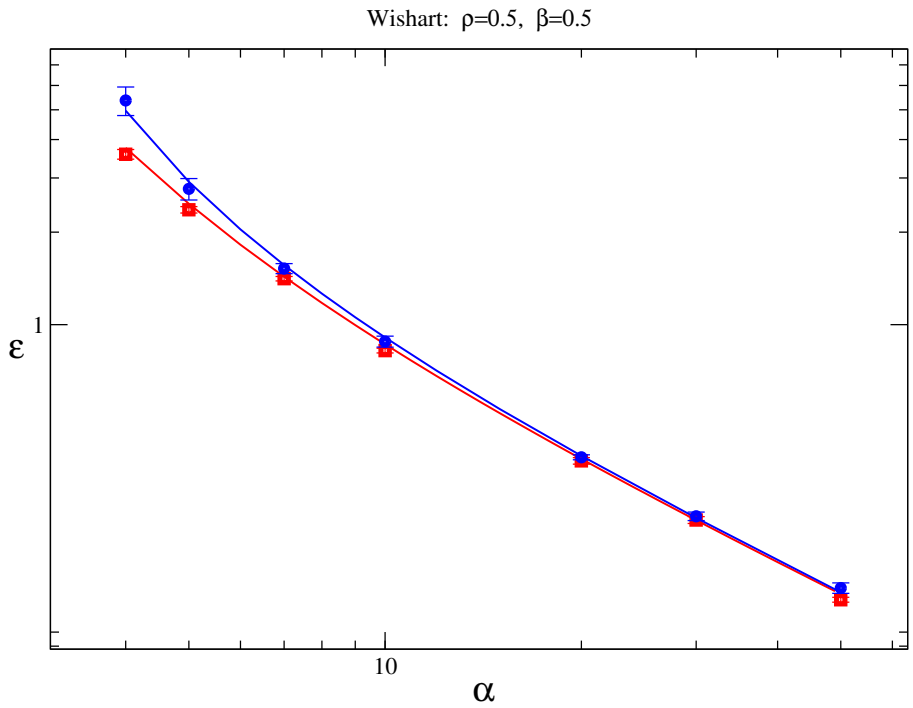
$$E(\mathbf{W}^{ML-MF}) = -\frac{\alpha N\phi_0}{2\beta}\sum_{i,j\neq 0}\hat{C}_{0i}(\hat{C}^{-1})_{ij}\hat{C}_{0j} = \frac{\alpha N(1-\phi_0)}{2\beta}.$$





SK-model: $J_{ij}^* \sim \mathcal{N}(0, 1/N)$





Hopfield–Wishart $J_{ij}^{*} = \frac{1}{N} \sum_{\mu=1}^{\rho N} \xi_i^{\mu} \xi_j^{\mu}$

Asymptotics of error

$$\varepsilon \simeq \frac{c}{\alpha} \qquad \alpha \to \infty$$

where

$$c_{\text{PLM}} = \frac{1}{\beta^2} \frac{1}{\left\langle 1 - \tanh^2(\beta h_*) \right\rangle_{h_*}} \frac{1}{N} \text{Tr} \mathbf{C}^{-1}$$

$$c_{\text{MF–ML}} = \frac{1 + \beta^2 V}{\beta^2} \frac{1}{N} \text{Tr} \mathbf{C}^{-1}$$

Outlook

- Try to estimate R transforms for real data.
- Optimise estimators for Inverse Ising problem.