Random matrix theory and the dynamics of Expectation Propagation

Manfred Opper and Burak Çakmak
AI Group TU Berlin

May 2, 2019

Outline

- Motivation
  1. Probabilistic Inference
  2. Cavity method
  3. TAP equations for Ising model
  4. EP algorithm (recurrent network dynamics)
- Analysis of algorithm dynamics
  1. Dynamical functional approach
  2. Explicit solution
  3. Comparison with simulations
- Understanding the results and robustness

Probabilistic Inference

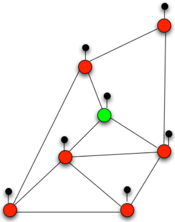
Posterior distribution of hidden variables $\mathbf{x}$ given observed data $\mathbf{y}$

$$p(\mathbf{x}|\mathbf{y}) = \frac{p(\mathbf{x},\mathbf{y})}{p(\mathbf{y})}$$

- Marginal probability of the data $p(\mathbf{y}) = \int d\mathbf{x}\; p(\mathbf{x},\mathbf{y})$ requires high dimensional integrals (or sums).
- Similar problems for the computation of marginals $p_i(x_i|\mathbf{y}) = \int d\mathbf{x}_{\setminus i}\; p(\mathbf{x}|\mathbf{y})$,

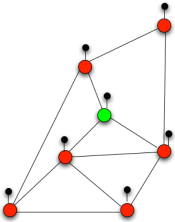
Gaussian latent variable models

$$p(\mathbf{x}|\mathbf{y}) = \frac{1}{Z} e^{-\frac{1}{2}\sum_{ij} x_i K_{ij} x_j} \prod_{k=1}^{N} f_k(x_k)$$



Gaussian latent variable models

$$p(\mathbf{x}|\mathbf{y}) = \frac{1}{Z} e^{-\frac{1}{2}\sum_{ij} x_i K_{ij} x_j} \prod_{k=1}^{N} f_k(x_k)$$



Examples:

- Gaussian process classification: $f_k(x_k) = \text{'sigmoid'}\,(y_k x_k)$ with $y_k = \pm 1$.
- Compressed sensing: $\mathbf{y} = \mathbf{H}\mathbf{x} + \boldsymbol{\epsilon}$ with $K \times N$ matrix $\mathbf{A}$.
  Sparsity prior $p_0(\mathbf{x}) = \prod_{k=1}^{N} \left((1-\rho)\delta(x_k) + \frac{\rho}{\sqrt{2\pi\sigma^2}} e^{-\frac{x_k^2}{2\sigma^2}} \right)$
- ...

. . . Ising model

$$P(\mathbf{x}) \propto \exp\left[\sum_{i<j} x_i J_{ij} x_j + \sum_i h_i x_i\right]$$

with discrete 'spin variables' $x_i = \pm 1$.
Write as Gaussian latent variable model:

$$p(\mathbf{x}) = e^{\sum_{k<l} x_k J_{kl} x_l} \prod_k f_k(x_k)$$

by taking

$$f_k(x) = e^{h_k x}\left\{\delta(x-1) + \delta(x+1)\right\} \ .$$

Cavity approach of statistical physics

(Mezard, Parisi, Virasoro 1987)

$$p(\mathbf{x}) \propto \exp\left[\sum_{i<j} x_i J_{ij} x_j\right] \prod_k f_k(x_k)$$

Suppose we are interested in node i

$$p(x_1, \ldots, x_{i-1}, \underline{x_i}, x_{i+1}, \ldots, x_N) \propto f_i(x_i) \exp[x_i \underbrace{\sum_{j \in \mathcal{N}(i)} J_{ij} x_j}_{h_i}]\, p_{\setminus i}(\mathbf{x} \setminus i)$$

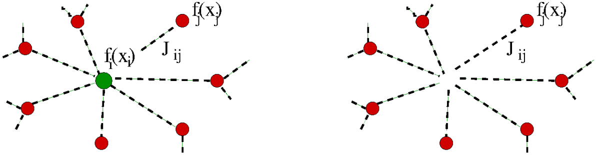
with $p_{\setminus i}(\mathbf{x} \setminus i)$ obtained by deleting node i.

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

$$p_i(x, h) \propto f_i(x)\; e^{x(h+h_i)}\; p_{\setminus i}(h)$$

with the 'cavity field' distribution

$$p_{\setminus i}(h) = \int \delta\left(h - \sum_{j \neq i} J_{ij} x_j\right)\; p_{\setminus i}(\mathbf{x}_{\setminus i}) d\mathbf{x}_{\setminus i}$$



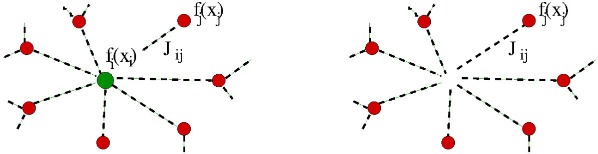
Dense graphs & weak dependencies: approximate inference

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

$$p_i(x,h) \propto f_i(x)\; e^{x(h+h_i)}\; p_{\setminus i}(h)$$

with the 'cavity field' distribution

$$p_{\setminus i}(h) = \int \delta\left(h - \sum_{j\neq i} J_{ij} x_j\right)\; p_{\setminus i}(\mathbf{x}_{\setminus i}) d\mathbf{x}_{\setminus i}$$



- Approximate $p_{\setminus i}(h)$ by Gaussian $p_{\setminus i}(h) \approx \mathcal{N}(a_i, V_i)$. Then

$$p_i(x) = \frac{1}{Z_i}\; f_i(x)\; \exp\left[a_i x + \frac{V_i}{2} x^2\right]$$

TAP Equations

- Within the Gaussian cavity approximation

$$a_i = \sum_j J_{ij} m_j - V_i m_i$$

with $m_j = E[x_j]$.

TAP Equations

- Within the Gaussian cavity approximation

$$a_i = \sum_j J_{ij} m_j - V_i m_i$$

  with $m_j = E[x_j]$.
- Neglecting dependencies

$$V_i = \sum_{jk} J_{ij} J_{ik} \mathrm{VAR}_{\setminus i}(x_i, x_j) \approx \sum_j J_{ij}^2 \, \mathrm{VAR}(x_j)$$

- TAP equations, *D J Thouless, P W Anderson & R J Palmer*, 1977)

$$m_i = \tanh\left(\sum_j J_{ij} m_j - m_i \sum_j J_{ij}^2 (1 - m_j^2) + h_i\right)$$

TAP Equations

- Within the Gaussian cavity approximation

$$a_i = \sum_j J_{ij} m_j - V_i m_i$$

with $m_j = E[x_j]$.

- Neglecting dependencies

$$V_i = \sum_{jk} J_{ij} J_{ik} \mathrm{VAR}_{\setminus i}(x_i, x_j) \approx \sum_j J_{ij}^2 \, \mathrm{VAR}(x_j)$$

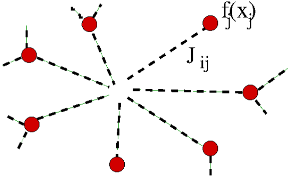
- TAP equations, *D J Thouless, P W Anderson & R J Palmer*, 1977)

$$m_i = \tanh\left(\sum_j J_{ij} m_j - m_i \sum_j J_{ij}^2 (1 - m_j^2) + h_i\right)$$

believed to be correct (in high temperature phase) for *Sherrington–Kirkpatrick model*, i.e. random couplings $J_{ij} \sim \mathcal{N}(0, \frac{c}{N})$!

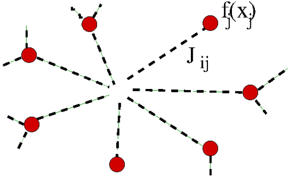
(adaptive) TAP equations:

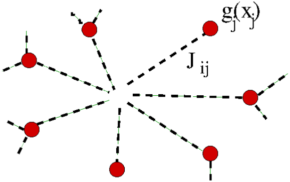
(MO and O Winther, 2000)



(adaptive) TAP equations:

(MO and O Winther, 2000)

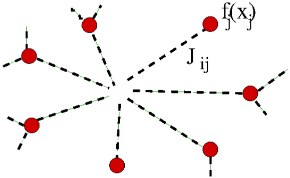


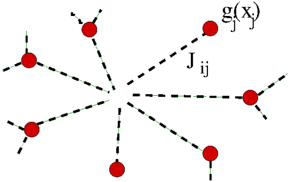


- Assume cavity field variances V_i depend only on moments $E[x_j]$ and $E[x_j^2]$ of surrounding variables (G Parisi, M Potters, 1995) .

(adaptive) TAP equations:

(MO and O Winther, 2000)





- Assume cavity field variances V_i depend only on moments $E[x_j]$ and $E[x_j^2]$ of surrounding variables (G Parisi, M Potters, 1995) .
- Work with an auxiliary Gaussian model where $f_i(x) = e^{-\frac{1}{2}\Lambda_i x^2 + \gamma_i x}$
- Hence, we have for all i we must have matching of 2nd moments

$$\mathrm{VAR}[x_i] = \left[(\mathbf{\Lambda} - \mathbf{J})^{-1}\right]_{ii} = \frac{1}{\Lambda_i - V_i}$$

- Leads to set of nonlinear self–consistent equations.

Expectation Propagation

- Efficient approximate inference algorithm (if convergent) introduced by Tom Minka (2001), applicable to discrete and continuous variables (and hybrid). Solves TAP fixed point equations
- Often excellent results for Gaussian latent variable models
- EP applications:
  `http://research.microsoft.com/en-us/um/people/minka/papers/ep/roadmap.html`
- Disadvantages: Variance updates costly ! Convergence properties unclear.

The questions

- Can we simplify EP
- and understand their 'typical' properties
- for large systems under random matrix assumptions for $\mathbf{J}$?

Some related results for AMP algorithms

- Analysis of message passing algorithm for TAP equations for SK–model (Bolthausen, 2014)
- Approximate message passing algorithm (Donoho, Maleki, Montanari, 2009) for compressed sensing.
- Analysis by statistical mechanics, phase diagrams, achieving of thresholds (Krzakala, Mézard, Sausset, Sun, Zdeborová, 2012)
- Rigorous analysis for matrices with random i.i.d. matrix elements (Bayati, Montanari, 2011, Bayati, Lelarge, Montanari 2015).
- VAMP algorithms by Rangan, Schniter, Fletcher (2016)

Costly variance conditions for 2nd moments

- Given χ_i for $i = 1, \ldots, N$: Find diagonal matrix $\mathbf{\Lambda} = \mathrm{diag}\,(\lambda_1, \ldots, \lambda_N)$ such that

$$[(\mathbf{\Lambda} - \mathbf{J})^{-1}]_{ii} = \chi_i$$

Costly variance conditions for 2nd moments

- Given χ_i for $i = 1, \ldots, N$: Find diagonal matrix $\mathbf{\Lambda} = \mathrm{diag}\,(\lambda_1, \ldots, \lambda_N)$ such that

$$[(\mathbf{\Lambda} - \mathbf{J})^{-1}]_{ii} = \chi_i$$

- Can we get an approximation to this computation if $\mathbf{J}$ is **'random'** ?
- Consider matrices of the form $\mathbf{J} \doteq \mathbf{O}^\top \mathbf{D} \mathbf{O}$ where $\mathbf{D}$ is a (deterministic) diagonal matrix and $\mathbf{O}$ random orthogonal (rotation).

Solution under freeness assumption for $N \to \infty$

- Define $\mathbf{\Lambda} \doteq \mathrm{diag}(\frac{1}{\chi_i}) - \mathrm{R}_{-\mathbf{J}}\left(-\frac{1}{N}\sum_i \chi_i\right)\mathbf{I}$
- Assume $\mathbf{\Lambda}$ and $\mathbf{K}$ asympt. free (B Çakmak and MO, ISIT 2018):

$$\lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N}\left(\left[(\mathbf{\Lambda}-\mathbf{J})^{-1}\right]_{ii} - \chi_i\right)^2 = 0$$

Random matrix TAP equations

For constant external fields $h_i \equiv h$ the approximate fixed point equations for $\mathbf{m} \equiv E[\mathbf{x}]$ are given by (G Paris & M Potters, 1995)

$$\begin{aligned} \mathbf{m} &= \mathrm{Th}(\boldsymbol{\gamma}) \\ \boldsymbol{\gamma} &= \mathbf{J}\mathbf{m} - \mathrm{R}_{\mathbf{J}}(\chi)\mathbf{m} \\ \chi &= \mathbb{E}_u[\mathrm{Th}'(\sqrt{(1-\chi)\mathrm{R}'_{\mathbf{J}}(\chi)}\,u)]. \end{aligned}$$

where we define

$$\mathrm{Th}(x) \doteq \tanh(h + x).$$

and $u \sim \mathcal{N}(0, 1)$.

(modified) EP algorithm for Ising model

- Initialise $\gamma(0) = \sqrt{(1-\chi)\mathrm{R}'}\mathbf{u}$ where $\mathbf{u} \sim \mathcal{N}(0, \mathbf{I})$
- Iterate

(modified) EP algorithm for Ising model

- Initialise $\gamma(0) = \sqrt{(1-\chi)\mathrm{R}'}\mathbf{u}$ where $\mathbf{u} \sim \mathcal{N}(0, \mathbf{I})$
- Iterate (similar to a **recurrent NN**)

$$\tilde{\gamma}(t) = \frac{1}{\chi}\mathrm{Th}(\gamma(t-1)) - \gamma(t-1)$$
$$\gamma(t) = \mathbf{A}\tilde{\gamma}(t)$$

for $t = 1, 2, 3, \ldots$ with the *time-independent* matrix

$$\mathbf{A} \doteq \frac{1}{\chi}(\lambda\mathbf{I} - \mathbf{J})^{-1} - \mathbf{I}.$$

- λ and χ are solutions of the (pre–computed) scalar equations

$$\mathrm{R}_{\mathbf{J}}(\chi) = \lambda - \frac{1}{\chi}$$
$$\chi = \mathbb{E}_u[\mathrm{Th}'(\sqrt{(1-\chi)\mathrm{R}'_{\mathbf{J}}(\chi)}u)]$$

Analysis: Generating functional approach

- Consider discrete time dynamics of the form

$$\tilde{\gamma}(t) = f(\gamma(t-1))$$
$$\gamma(t) = \mathbf{A}\tilde{\gamma}(t)$$

Analysis: Generating functional approach

- Consider discrete time dynamics of the form

$$\tilde{\gamma}(t) = f(\gamma(t-1))$$
$$\gamma(t) = \mathbf{A}\tilde{\gamma}(t)$$

- Marginal dynamics of $\gamma_i(t)$ derived from generating functional $E_{\mathbf{A}}\left[Z\{\mathbf{l}(t)\}\right]$
(Martin, Siggia, Rose, 1973, Sompolinsky & Zippelius, 1981)

$$Z\{\mathbf{l}(t)\} \doteq \int \prod_{t=1}^{T} \mathrm{d}\tilde{\gamma}(t)\mathrm{d}\gamma(t)\, \delta\left(\tilde{\gamma}(t) - f(\gamma(t-1))\right) \times \\ \times\, \delta\left(\gamma(t) - \mathbf{A}\tilde{\gamma}(t)\right) e^{\mathrm{i}\sum_i \gamma_i(t) l_i(t)}$$

- Replace Dirac $\delta(\cdot)$ by Fourier representation
- Perform expectation over disorder

$$E_{\mathbf{A}}\left[e^{i\left\{\sum_t \hat{\gamma}(t)^\top \mathbf{A}\tilde{\gamma}(t)\right\}}\right]$$

- For rotational invariant **A**, the degrees of freedom of resulting non–random model are decoupled by order parameters which become self–averaging for $N \to \infty$.
- Order parameters introduce couplings in time.
- Exact for $N \to \infty$, number of steps T finite !

$N \to \infty$: Effective stochastic dynamics with memory

- The resulting effective stochastic process of single variables is given by

$$\tilde{\gamma}(t) = f(\gamma(t-1))$$
$$\gamma(t) = \sum_{s<t} \hat{\mathcal{G}}(t,s)\tilde{\gamma}(s) + \phi(t)$$

- $\hat{\mathcal{G}}$ is a $T \times T$ matrix defined by the matrix function

$$\hat{\mathcal{G}} \doteq \mathrm{R}_{\mathbf{A}}(\mathcal{G})$$

- $\mathcal{G}$ is a $T \times T$ **susceptibility** matrix

$$\mathcal{G}(t,s) \doteq \mathbb{E}\left[\frac{\partial \tilde{\gamma}(t)}{\partial \phi(s)}\right].$$

The zero-mean Gaussian process $\{\phi(t)\}$ has a covariance matrix given by

$$\mathcal{C}_\phi = \sum_{n=1}^{\infty} c_{\mathbf{A},n} \sum_{k=0}^{n-2} \mathcal{G}^k \tilde{\mathcal{C}} (\mathcal{G}^\top)^{n-2-k}$$

where

$$\tilde{\mathcal{C}}(t,s) \doteq \mathbb{E}[\tilde{\gamma}(t)\tilde{\gamma}(s)].$$

and the $c_{\mathbf{A},n}$ are **free cumulants** defined by the R–transform $\mathrm{R}_{\mathbf{A}}$.

But things actually simplify a bit ...

For the specific choice of $f(x) \doteq \frac{1}{\chi}\mathrm{Th}(x) - x$ and $\mathbf{A} \doteq \frac{1}{\chi}(\lambda\mathbf{I} - \mathbf{J})^{-1} - \mathbf{I}$, we get

But things actually simplify a bit ...

For the specific choice of $f(x) \doteq \frac{1}{\chi}\mathrm{Th}(x) - x$ and $\mathbf{A} \doteq \frac{1}{\chi}(\lambda\mathbf{I} - \mathbf{J})^{-1} - \mathbf{I}$, we get

$$\begin{aligned}\tilde{\gamma}(t) &= f(\gamma(t-1)) \\ \gamma(t) &= \phi(t)\end{aligned}$$

where the $\phi(t)$ are Gaussian random variables with a covariance computed recursively

$$\begin{aligned}\mathcal{C}(t,s) &= \frac{g(\mathcal{C}(t-1,s-1))}{1/\mathrm{R}' - \chi^2} \\ \mathcal{C}(t,t) &= (1-\chi)\mathrm{R}'\end{aligned}$$

and we have defined

$$g(x) \doteq \mathbb{E}[\mathrm{Th}(\gamma_1)\mathrm{Th}(\gamma_2)] - \chi^2 x$$

Convergence and Asymptotics $T \to \infty$

- To analyse convergence, study

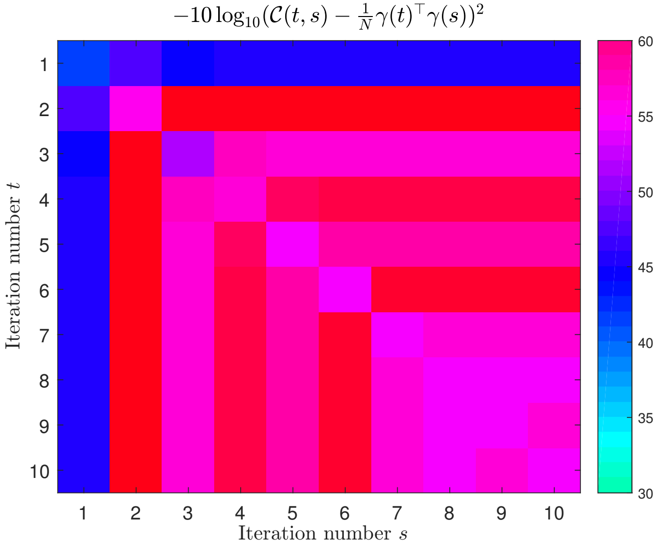
$$\Delta(t,s) \doteq \lim_{N\to\infty} \frac{1}{N}\mathbb{E}[\|\gamma(t)-\gamma(s)\|^2]$$

- Result: If $\frac{1-\eta\mathrm{R}'}{1-\chi^2\mathrm{R}'} < 1$ (AT line):
  Convergence (from random initial conditions) with rate

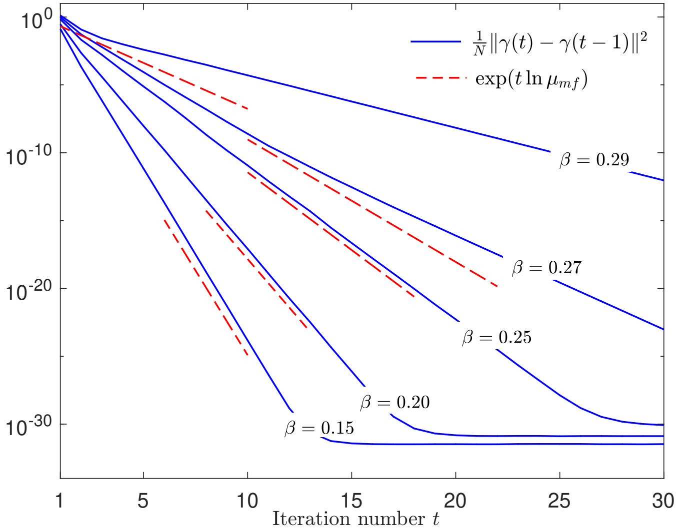
$$\lim_{t\to\infty} \frac{1}{t}\ln\Delta(t,\infty) = \ln\left(1 - \frac{1-\eta\mathrm{R}'}{1-\chi^2\mathrm{R}'}\right)$$

  where $\eta \doteq \mathbb{E}_u[(\mathrm{Th}'(\sqrt{(1-\chi)\mathrm{R}'(\chi)}u))^2]$.

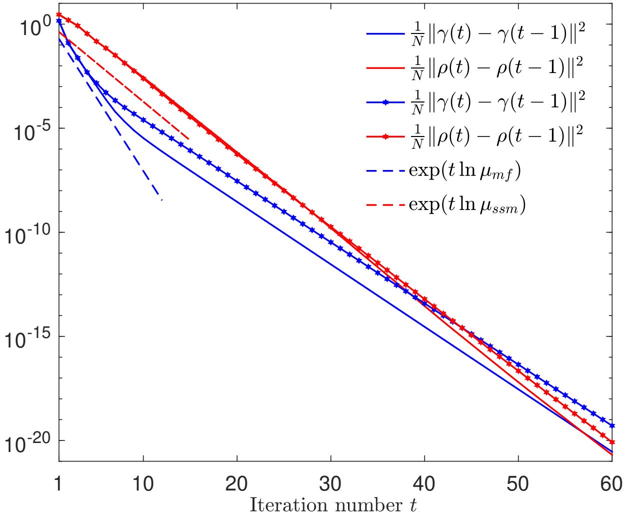
Comparison with simulations



$\mathbf{J} = \beta \mathbf{X}_2^\top \mathbf{X}_1^\top \mathbf{X}_1 \mathbf{X}_2$ ('2 layer–Hopfield'), $N = 10^4$.



$N = 10^4$, critical $\beta = 0.35$ ('2 layer–Hopfield')



2 realisations and comparison with previously defined algorithm on: 'Single layer Hopfield' **J**.

Understanding the results

- Linearisation:

$$\frac{\partial \gamma_i(t)}{\partial \gamma_j(t-1)} = (\mathbf{A}\mathbf{D}(t-1))_{ij}$$

where $\phi(\mathbf{A}) = \phi(\mathbf{D}(t-1)) = 0$ and $\phi(\ldots) \doteq \lim \frac{1}{N}\mathrm{Tr}(\ldots)$

Understanding the results

- Linearisation:

$$\frac{\partial \gamma_i(t)}{\partial \gamma_j(t-1)} = (\mathbf{AD}(t-1))_{ij}$$

where $\phi(\mathbf{A}) = \phi(\mathbf{D}(t-1)) = 0$ and $\phi(\ldots) \doteq \lim \frac{1}{N}\mathrm{Tr}(\ldots)$

- Leads to vanishing of **susceptibility** by freeness (self averaging)

$$\frac{\partial \gamma_i(t)}{\partial \gamma_i(t')} = \left(\prod_{\tau=t'}^{t-1} \mathbf{AD}(\tau) \right)_{ii} \to 0$$

Understanding the results

- Linearisation:
$$\frac{\partial \gamma_i(t)}{\partial \gamma_j(t-1)} = (\mathbf{AD}(t-1))_{ij}$$
where $\phi(\mathbf{A}) = \phi(\mathbf{D}(t-1)) = 0$ and $\phi(\ldots) \doteq \lim \frac{1}{N}\mathrm{Tr}(\ldots)$
- Leads to vanishing of **susceptibility** by freeness (self averaging)
$$\frac{\partial \gamma_i(t)}{\partial \gamma_i(t')} = \left(\prod_{\tau=t'}^{t-1} \mathbf{AD}(\tau)\right)_{ii} \to 0$$
- Small random perturbation of fixed–point (use freeness):
$$\frac{1}{N}\|\delta\boldsymbol{\gamma}(T)\|^2 \simeq C\phi(\mathbf{A}^2)^T\phi(\mathbf{D}^2)^T$$
coincides with asymptotics calculated from dynamical functional method.

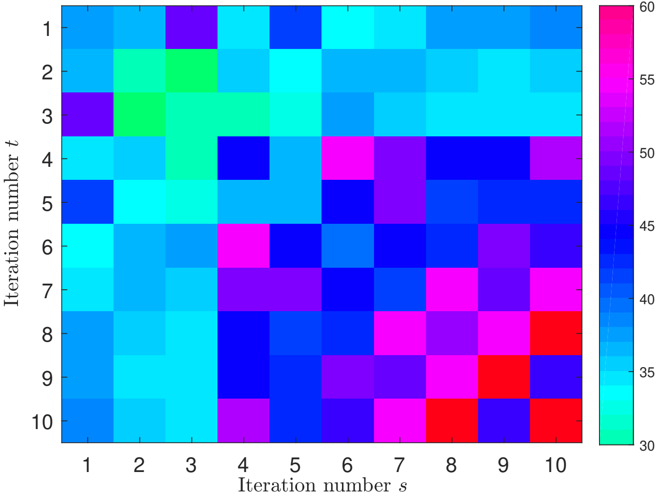
Robustness: Hadamard ensemble

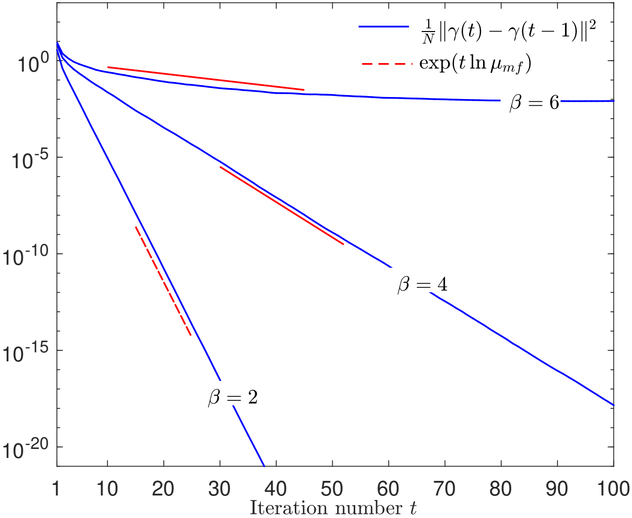
- Define non–rotational invariant ensemble

$$\mathbf{J} = \beta \tilde{\mathbf{O}}^\top \mathbf{D}_\rho \tilde{\mathbf{O}} \quad \text{with} \quad \tilde{\mathbf{O}} \doteq \frac{1}{\sqrt{N}} \mathbf{H}_N \mathbf{Z}.$$

- $\mathbf{Z}$ and random diagonal with $z_i = \pm 1$ and $\mathbf{D}_\rho$ random diagonal $d_i = \pm 1$ with $|\{d_i = 1\}| = \rho N$.
- $\mathbf{H}_N$ is the $N \times N$ **Hadamard** matrix. $\tilde{\mathbf{O}}$ is an orthogonal matrix with $\tilde{O}_{ij} = \pm \frac{1}{\sqrt{N}}$.

$$-10\log_{10}(\mathcal{C}(t,s) - \tfrac{1}{N}\gamma(t)^\top\gamma(s))^2$$





$\beta_c = 6.8,\ N = 2^{13}$.

Outlook

- Generalise to other EP problems
- Model of real data ?
- Learning of matrices

Publications

- **Details of dynamical functional method:**
  A theory of solving TAP equations for Ising models with general invariant random matrices, M. Opper, B. Çakmak and O. Winther, Journal of Physics A: Mathematical and Theoretical 49, 114002 (2016).
- **Simplifying EP using free probability:**
  Expectation Propagation for Approximate Inference: Free Probability Framework, B. Çakmak and M. Opper, ISIT (2018).
- **Explicit solution to dynamics:**
  Memory-free dynamics for the TAP equations of Ising models with arbitrary rotation invariant ensembles of random coupling matrices
  Authors: B. Çakmak and M. Opper, ISIT (2019).
  arXiv:1901.08583v1

Appendix: Stieltjes–transform and R–transform

- **Stieltjes–transform**

$$\mathrm{G}_{\mathbf{A}}(z) \doteq \phi\left(\mathbf{A} - z\mathbf{I}\right)^{-1}$$

with $\phi(\mathbf{A}) \doteq \lim \frac{1}{N}\mathrm{Tr}\mathbf{A}$.

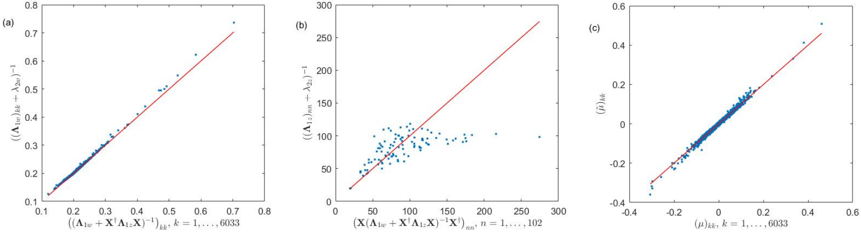
- and its functional inverse

$$z_{\mathbf{A}}(s) \triangleq \mathrm{G}_{\mathbf{A}}^{-1}(s)$$

- The R–transform is defined as

$$\mathrm{R}_{\mathbf{A}}(s) \doteq z_{\mathbf{A}}(-s) - 1/s \tag{6}$$

Real data



$K = 6033$, $N = 102$.