Expectation Consistent Approximate Inference

Ole Winther

Informatics and Mathematical Modelling
Technical University of Denmark
DK-2800 Lyngby, Denmark

owi@imm.dtu.dk

In collaboration with **Manfred Opper**

ISIS
School of Electronics and Computer Science
University of Southampton
SO17 1BJ, United Kingdom

mo@ecs.soton.ac.uk

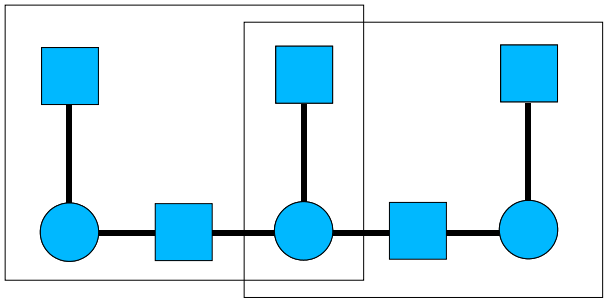
Motivation

- Contemporary machine learning uses complex flexible probabilistic models.
- Bayesian inference is typically intractable.
- Approximate polynomial complexity methods needed.
- VB, Bethe, EP and EC: Use tractable factorization of original model.
- EC: Expectation Consistency between 2 distributions, e.g. discrete and Gaussian

Exact Inference in Tree Graphs

Bethe – tree factorization, e.g.

$$p(\mathbf{x}) = \frac{1}{Z} f_{12} f_{13} f_1 f_2 f_3$$



Write $p(\mathbf{x})$ in terms of marginals $q_i(x_i)$ and $q_{ij}(x_i, x_j)$

$$p(\mathbf{x}) = q(\mathbf{x}) = \frac{q_{12}(x_1, x_2) q_{23}(x_2, x3)}{q_2(x_2)}$$

$$Z = \frac{Z_{12} Z_{23}}{Z_1}$$

Message-parsing: Effective inference for $p(\mathbf{x})$ discrete or Gaussian.

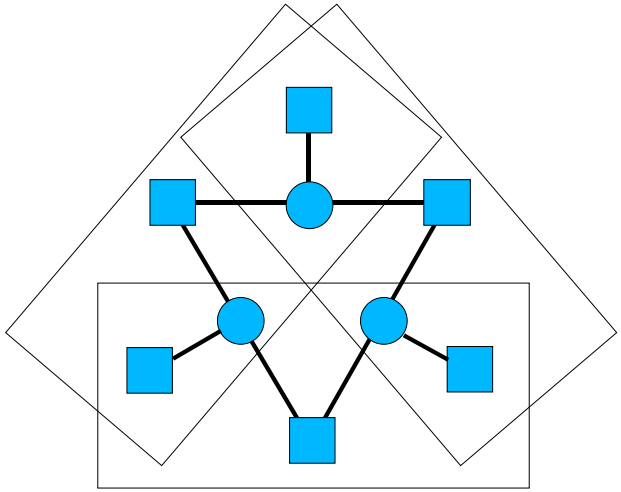
Bethe Approximation

Bethe approximation – treat $p(\mathbf{x})$, e.g.

$$p(\mathbf{x}) = \frac{1}{Z} f_{12} f_{23} f_{13} f_1 f_2 f_3$$

as if it was a tree-graph

$$q(\mathbf{x}) = \frac{q_{12}(x_1, x_2) q_{23}(x_2, x3) q_{13}(x_1, x_3)}{q_1(x_1) q_2(x_2) q_3(x_3)} .$$



Works extremely well in "sparse systems" - e.g. low density decoding.

Disadvantage over-counting – $q(\mathbf{x})$ not a density.

Variational Bayes (VB)

Minimize KL-divergence in restricted tractable family $q(\mathbf{x}) = \prod_i q_i(x_i)$:

$$q_i(x_i) = \underset{q_i(x_i)}{\operatorname{argmin}} \, \mathrm{KL}\left[q(\mathbf{x})||p(\mathbf{x})\right] \propto \exp\left\langle \ln p(\mathbf{x}) \right\rangle_{q \setminus q_i(x_i)}$$

Example Gaussian:

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \mathbf{m}, \mathbf{C}) \qquad \rightarrow \qquad q(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \mathbf{m}^q, \mathbf{C}^q)$$

$$\mathbf{m}^q = \mathbf{m} \qquad \text{and} \qquad C^q_{ij} = \delta_{ij} \frac{1}{\left[\mathbf{C}^{-1}\right]_{ii}}$$

In general (factorized) VB reliable on mean, but under-estimates width of distribution (see e.g. MacKay, 2003, Opper & Winther 2004).

Important for parameter-estimation (see e.g. Minka & Lafferty).

Motivating EC and Overview

We are looking for a tractable approximation that

- can handle "dense graphs" (better than Bethe+).
- estimate correlations (better than VB).

Free energy

Why it works – central limit theorem.

Algorithmics and connection to EP

Simulations, conclusions and outlook

Expectation Consistent (EC) free energy

Calculate partition function

$$Z = \int d\mathbf{x}\, f(\mathbf{x}) = \int d\mathbf{x}\, f_q(\mathbf{x}) f_r(\mathbf{x})$$

Problem: Z intractable – integral not analytical and/or summation exponential in number of variables N.

Introduce **tractable distribution** $q(\mathbf{x})$

$$q(\mathbf{x}) = \frac{1}{Z_q(\boldsymbol{\lambda}_q)} f_q(\mathbf{x}) \exp(\boldsymbol{\lambda}_q^T \mathbf{g}(\mathbf{x}))$$

Z_q can be calculated in polynomial time.

$$\begin{aligned} Z &= Z_q \frac{Z}{Z_q} = Z_q \frac{\int d\mathbf{x} f_r(\mathbf{x}) f_q(\mathbf{x}) \exp\left((\boldsymbol{\lambda}_q - \boldsymbol{\lambda}_q)^T \mathbf{g}(\mathbf{x})\right)}{\int d\mathbf{x} f_q(\mathbf{x}) \exp \boldsymbol{\lambda}_q^T \mathbf{g}(\mathbf{x})} \\ &= Z_q \left\langle f_r(\mathbf{x}) \exp\left(-\boldsymbol{\lambda}_q^T \mathbf{g}(\mathbf{x})\right)\right\rangle_q \end{aligned}$$

Free energy

Free energy exact:

$$-\ln Z = -\ln Z_q - \ln \left\langle f_r(\mathbf{x}) \exp\left(-\boldsymbol{\lambda}_q^T \mathbf{g}(\mathbf{x})\right)\right\rangle_q$$

Variational approximation use Jensen: $\ln \langle f(\mathbf{x})\rangle \geq \langle \ln f(\mathbf{x})\rangle$

$$-\ln Z \leq -\ln Z_q - \langle \ln f_r(\mathbf{x})\rangle_q + \boldsymbol{\lambda}_q^T \langle \mathbf{g}(\mathbf{x})\rangle_q$$

Find $\boldsymbol{\lambda}_q$ by minimizing the upper bound.

Better to average over $f_r(\mathbf{x}) \exp\left(-\boldsymbol{\lambda}_q^T \mathbf{g}(\mathbf{x})\right)$ **approximately.**

Retain **more averaging** in that way.

Expectation consistent approximation

Define $\mathbf{g}(\mathbf{x})$ such that both

$$
\begin{aligned}
q(\mathbf{x}) &= \frac{1}{Z_q(\boldsymbol{\lambda}_q)} f_q(\mathbf{x}) \exp(\boldsymbol{\lambda}_q^T \mathbf{g}(\mathbf{x})) \\
r(\mathbf{x}) &= \frac{1}{Z_r(\boldsymbol{\lambda}_r)} f_r(\mathbf{x}) \exp(\boldsymbol{\lambda}_r^T \mathbf{g}(\mathbf{x}))
\end{aligned}
$$

are tractable.

Excludes some models tractable in the variational approach (without further approximations).

Example I – the Ising model

Binary variables – spins – $x_i = \pm 1$ with pairwise interactions

$$
\begin{aligned}
f_q(\mathbf{x}) &= \prod_i \Psi_i(x_i) \\
\psi_i(x_i) &= [\delta(x_i + 1) + \delta(x_i - 1)] e^{\theta_i x_i} \\
f_r(x) &= \exp\left(\sum_{i>j} x_i J_{ij} x_j\right) = \exp\left(\frac{1}{2}\mathbf{x}^T \mathbf{J}\mathbf{x}\right)
\end{aligned}
$$

E.g. set $g(\mathbf{x})$ to first and second order

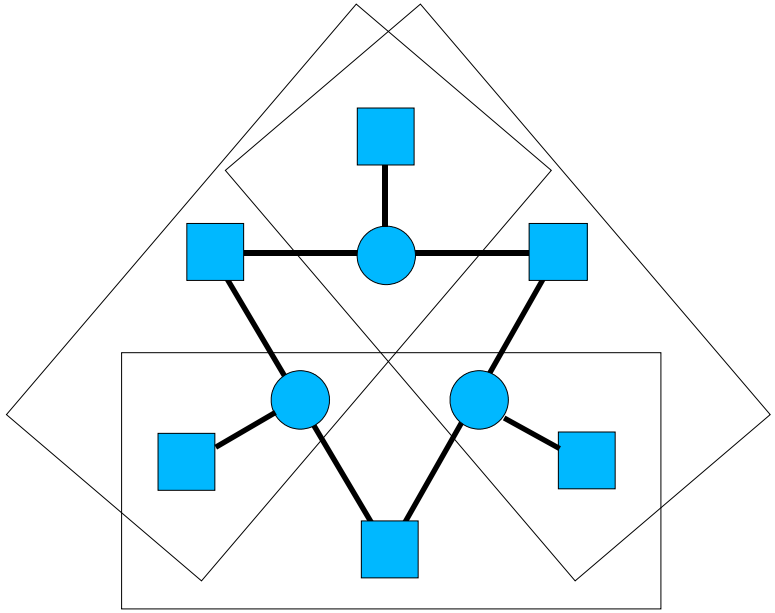
$$
\mathbf{g}(\mathbf{x}) = \left(x_1, -\frac{x_1^2}{2}, x_2, -\frac{x_2^2}{2}, \ldots, x_N, -\frac{x_N^2}{2}\right)
$$

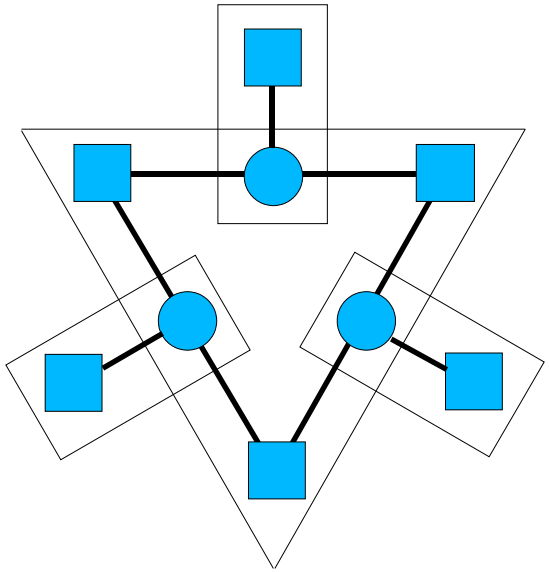
$q(\mathbf{x})$ – a factorized binary distribution

$r(\mathbf{x})$ – multivariate Gaussian.

Interpretation of $\mathbf{g}(\mathbf{x})$ will be clear shortly.

Bethe and EC factorization





$$Z^{\text{Bethe}} = \frac{Z_{12} Z_{23} Z_{13}}{Z_1 Z_2 Z_3} \ .$$

Z^{EC} will be similar in spirit:

$$Z^{\text{EC}} = \frac{Z_q Z_r}{Z_{\text{s(eparator)}}} \ .$$

Example II – Gaussian processes

Supervised learning: Inputs $\mathbf{x}_1, \ldots, \mathbf{x}_N$ and targets $t_1, \ldots, t_N$.

Gaussian process prior over functions $\mathbf{y} = (y(\mathbf{x}_1), \ldots, y(\mathbf{x}_N))$:

$$p(\mathbf{y}) = \frac{1}{\sqrt{(2\pi)^N \det \mathbf{C}}} \exp\left(-\frac{1}{2}\mathbf{y}^T \mathbf{C}^{-1} \mathbf{y}\right)$$

Likelihood, observation model: $p(t|y(\mathbf{x}))$, e.g. noise-free classification

$$p(t|y(\mathbf{x})) = \Theta(ty(\mathbf{x}))$$

$$Z = \int d\mathbf{y} \prod_i p(t_i|y(\mathbf{x}_i)) p(\mathbf{y})$$

Same structure as ex. I – factorized and multivariate Gaussian (Opper&Winther,2000; Minka 2001).

Expectation Consistent (Helmholtz) Free Energy

Exchange average wrt $q(\mathbf{x})$ with one over simpler distribution $s(\mathbf{x})$.

$$s(\mathbf{x}) = \frac{1}{Z_s(\boldsymbol{\lambda}_s)} \exp\left(\boldsymbol{\lambda}_s^T \mathbf{g}(\mathbf{x})\right)$$

Approximation:

$$\left\langle f_r(\mathbf{x}) \exp\left(-\boldsymbol{\lambda}_q^T \mathbf{g}(\mathbf{x})\right)\right\rangle_q \approx \left\langle f_r(\mathbf{x}) \exp\left(-\boldsymbol{\lambda}_q^T \mathbf{g}(\mathbf{x})\right)\right\rangle_s$$

Parameters $\boldsymbol{\lambda}_q, \boldsymbol{\lambda}_s$ to be optimized in suitable way:

$$\begin{aligned} -\ln Z &\approx -\ln Z_q - \ln \left\langle f_r(\mathbf{x}) \exp\left(-\boldsymbol{\lambda}_q^T \mathbf{g}(\mathbf{x})\right)\right\rangle_s \\ &= -\ln \int d\mathbf{x} f_q(\mathbf{x}) \exp\left(\boldsymbol{\lambda}_q^T \mathbf{g}(\mathbf{x})\right) \\ &\quad -\ln \int d\mathbf{x} f_r(\mathbf{x}) \exp\left((\boldsymbol{\lambda}_s - \boldsymbol{\lambda}_q)^T \mathbf{g}(\mathbf{x})\right) \\ &\quad +\ln \int d\mathbf{x} \exp\left(\boldsymbol{\lambda}_s^T \mathbf{g}(\mathbf{x})\right) \end{aligned}$$

Determining the Parameters

Expectation consistency:

$$\frac{\partial \ln Z^{\mathrm{EC}}}{\partial \boldsymbol{\lambda}_q} = 0 \quad : \quad \langle \mathbf{g}(\mathbf{x}) \rangle_q = \langle \mathbf{g}(\mathbf{x}) \rangle_r$$

$$\frac{\partial \ln Z^{\mathrm{EC}}}{\partial \boldsymbol{\lambda}_s} = 0 \quad : \quad \langle \mathbf{g}(\mathbf{x}) \rangle_r = \langle \mathbf{g}(\mathbf{x}) \rangle_s$$

where

$$q(\mathbf{x}) = \frac{1}{Z_q(\boldsymbol{\lambda}_q)} f_q(\mathbf{x}) \exp(\boldsymbol{\lambda}_q^T \mathbf{g}(\mathbf{x}))$$

$$r(\mathbf{x}) = \frac{1}{Z_r(\boldsymbol{\lambda}_r)} f_r(\mathbf{x}) \exp(\boldsymbol{\lambda}_r^T \mathbf{g}(\mathbf{x})) \qquad \text{with} \qquad \boldsymbol{\lambda}_r = \boldsymbol{\lambda}_s - \boldsymbol{\lambda}_q$$

$$s(\mathbf{x}) = \frac{1}{Z_r(\boldsymbol{\lambda}_s)} \exp(\boldsymbol{\lambda}_s^T \mathbf{g}(\mathbf{x}))$$

$$Z \approx \frac{Z_r Z_q}{Z_s}$$

Approximation symmetric in $q(\mathbf{x})$ and $r(\mathbf{x})$. $s(\mathbf{x})$ is the "separator".

Why it Works

Neither q or r are good approximations to p.

But **marginal distributions** and **moments** can be precise!

$\mathbf{g}(\mathbf{x}) = \left(x_1, -\frac{x_1^2}{2}, \ldots, x_N, -\frac{x_N^2}{2}\right)$ and $\boldsymbol{\lambda} = (\gamma_1, \Lambda_1, \ldots, \gamma_N, \Lambda_N)$:

$$q(\mathbf{x}) = \prod_i q_i(x_i) \qquad q_i(x_i) \propto \Psi_i(x_i) \exp\left(\gamma_{q,i} x_i - \Lambda_{q,i} x_i^2\right) \ .$$

The central limit theorem saves us: the details of the distribution of the marginalized variables not important, only first and second moments. Cavity method (Onsager 1936, Mezard, Parisi & Virasoro 1987).

Exact under some conditions: "dense models", many variables, no dominating interactions and not too strong interactions.

Other complications such as non-ergodicity (RSB).

Non-trivial estimates in EC

- Marginal distributions $q(x_i)$ (factorized moments)

$$q(\mathbf{x}) \propto \prod_i \Psi_i(x_i) \exp(\boldsymbol{\gamma}_q^T \mathbf{x} - \mathbf{x}^T \boldsymbol{\Lambda}_q \mathbf{x}/2)$$

$$q(x_i) \propto \Psi_i(x_i) \exp(\gamma_{q,i} x_i - x_i^2 \Lambda_{q,i}/2) \ .$$

- Correlations $r(\mathbf{x})$ global Gaussian approximation

$$r(\mathbf{x}) \propto \exp(\boldsymbol{\gamma}_r^T \mathbf{x} - \mathbf{x}^T (\boldsymbol{\Lambda}_r - \mathbf{J}) \mathbf{x}/2)$$

  Covariance $C(x_i, x_j) = \langle x_i x_j \rangle_{r(\mathbf{x})} - \langle x_i \rangle_{r(\mathbf{x})} \langle x_j \rangle_{r(\mathbf{x})} = \left[(\boldsymbol{\Lambda}_r - \mathbf{J})^{-1} \right]_{ij}$.

- The free energy $-\ln Z^{\mathsf{EC}} \approx -\ln Z$.
  Z is the *marginal likelihood* (or *evidence*) of the model.

- Supervised learning, Predictive distribution and leave-one-out (Opper & Winther, 2000).

Non-Convex Optimization

Partition function $Z(\boldsymbol{\lambda}) = \int d\mathbf{x} f(\mathbf{x}) \exp\left(\boldsymbol{\lambda}^T \mathbf{g}(\mathbf{x})\right)$ is convex in $\boldsymbol{\lambda}$:

$$\mathbf{H} = \frac{\partial^2 \ln Z}{\partial \boldsymbol{\lambda}^T \boldsymbol{\lambda}} = \left\langle \mathbf{g}(\mathbf{x}) \mathbf{g}(\mathbf{x})^T \right\rangle - \left\langle \mathbf{g}(\mathbf{x}) \right\rangle \left\langle \mathbf{g}(\mathbf{x}) \right\rangle^T \quad .$$

EC non-convex optimization – like Bethe and variational.

$$\begin{aligned} -\ln Z^{\mathsf{EC}}(\boldsymbol{\lambda}_q, \boldsymbol{\lambda}_s) &= -\ln Z_q(\boldsymbol{\lambda}_q) - \ln Z_r(\boldsymbol{\lambda}_s - \boldsymbol{\lambda}_q) + \ln Z_s(\boldsymbol{\lambda}_s) \\ &= -\ln \int d\mathbf{x} f_q(\mathbf{x}) \exp\left(\boldsymbol{\lambda}_q^T \mathbf{g}(\mathbf{x})\right) \\ &\quad -\ln \int d\mathbf{x} f_r(\mathbf{x}) \exp\left((\boldsymbol{\lambda}_s - \boldsymbol{\lambda}_q)^T \mathbf{g}(\mathbf{x})\right) \\ &\quad +\ln \int d\mathbf{x} \exp\left(\boldsymbol{\lambda}_s^T \mathbf{g}(\mathbf{x})\right) \end{aligned}$$

Optimize with **single loop (no warranty)** or **double loop (slow)**.

Single Loop – Objective

Expectation consistency

$$\langle \mathbf{g}(\mathbf{x}) \rangle_q = \langle \mathbf{g}(\mathbf{x}) \rangle_r = \langle \mathbf{g}(\mathbf{x}) \rangle_s$$

with

$$\begin{aligned} q(\mathbf{x}) &= \frac{1}{Z_q(\boldsymbol{\lambda}_q)} f_q(\mathbf{x}) \exp(\boldsymbol{\lambda}_q^T \mathbf{g}(\mathbf{x})) \\ r(\mathbf{x}) &= \frac{1}{Z_r(\boldsymbol{\lambda}_r)} f_r(\mathbf{x}) \exp(\boldsymbol{\lambda}_r^T \mathbf{g}(\mathbf{x})) \quad \text{with} \quad \boldsymbol{\lambda}_r = \boldsymbol{\lambda}_s - \boldsymbol{\lambda}_q \\ s(\mathbf{x}) &= \frac{1}{Z_r(\boldsymbol{\lambda}_s)} \exp(\boldsymbol{\lambda}_s^T \mathbf{g}(\mathbf{x})) \end{aligned}$$

Sending messages $r \rightarrow q \rightarrow r \rightarrow \ldots$ and make s consistent.

Single Loop – Propagation Algorithms

1. Send messages from r to q

   - Calculate separator $s(\mathbf{x})$.
     Solve for $\boldsymbol{\lambda}_s$: $\langle \mathbf{g}(\mathbf{x}) \rangle_s = \boldsymbol{\mu}_r(t) \equiv \langle \mathbf{g}(\mathbf{x}) \rangle_{r(\mathbf{x};t)}$

   - Update $q(\mathbf{x})$: $\boldsymbol{\lambda}_q(t+1) := \boldsymbol{\lambda}_s - \boldsymbol{\lambda}_r(t)$

2. Send messages from q to r

   - Calculate separator $s(\mathbf{x})$.
     Solve for $\boldsymbol{\lambda}_s$: $\langle \mathbf{g}(\mathbf{x}) \rangle_s = \boldsymbol{\mu}_q(t+1) \equiv \langle \mathbf{g}(\mathbf{x}) \rangle_{q(\mathbf{x};t+1)}$

   - Update $r(\mathbf{x})$: $\boldsymbol{\lambda}_r(t+1) := \boldsymbol{\lambda}_s - \boldsymbol{\lambda}_q(t+1)$

Expectation Propagation (EP): sequential factor-by-factor update.

Single Loop Details

$q(\mathbf{x})$ non-Gaussian, factorized or on a spanning tree and $r(\mathbf{x})$ multi-variate Gaussian. Complexity $\mathcal{O}(N^3)$.

Factorized moments $\mathbf{g}(\mathbf{x}) = \left(x_1, -\frac{x_1^2}{2}, x_2, -\frac{x_2^2}{2}, \ldots, x_N, -\frac{x_N^2}{2}\right)$:

Gaussian $s(\mathbf{x}) = \prod_i s_i(x_i)$ and $s_i(x_i) \propto \exp\left(\gamma_{s,i} x_i - \Lambda_{s,i} x_i^2/2\right)$.

Moment matching to mean and variance of q and r:

$$\gamma_{s,i} := m_i/v_i \qquad \text{and} \qquad \Lambda_{s,i} := 1/v_i \ .$$

All second moments on a spanning tree:

$q(\mathbf{x})$ moments can be inferred by (exact) **message parsing**.

$s(\mathbf{x})$ multi-variate Gaussian on a spanning tree, solve using **tree-decomposition** of Z.

Double Loop – EC (Gibbs) free energy

Gibbs free energy definition (Lagrangian dual of $\ln Z(\boldsymbol{\lambda})$):

$$G(\boldsymbol{\mu}) = \max_{\boldsymbol{\lambda}} \left\{ -\ln Z(\boldsymbol{\lambda}) + \boldsymbol{\lambda}^T \boldsymbol{\mu} \right\}$$

convex in generalized moments $\boldsymbol{\mu} = \langle \mathbf{g}(\mathbf{x}) \rangle = \frac{\partial \ln Z(\boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}}$.

EC Gibbs free energy (non-convex in $\boldsymbol{\mu}$)

$$\begin{aligned} G^{\mathsf{EC}}(\boldsymbol{\mu}) &= G_q(\boldsymbol{\mu}) + G_r(\boldsymbol{\mu}) - G_s(\boldsymbol{\mu}) \\ &= \max_{\boldsymbol{\lambda}_q, \boldsymbol{\lambda}_r} \min_{\boldsymbol{\lambda}_s} \{ -\ln Z_q(\boldsymbol{\lambda}_q) - \ln Z_r(\boldsymbol{\lambda}_r) + \ln Z_s(\boldsymbol{\lambda}_s) \\ &\qquad + \boldsymbol{\mu}^T (\boldsymbol{\lambda}_q + \boldsymbol{\lambda}_r - \boldsymbol{\lambda}_s) \} \\ -\ln Z^{\mathsf{EC}} &= \min_{\boldsymbol{\mu}} G^{\mathsf{EC}}(\boldsymbol{\mu}) \end{aligned}$$

Helmholtz from $\min_{\boldsymbol{\mu}} G^{\mathsf{EC}}(\boldsymbol{\mu})$: $\boldsymbol{\lambda}_q + \boldsymbol{\lambda}_r - \boldsymbol{\lambda}_s = 0$ to eliminate $\boldsymbol{\lambda}_r$.

Double loop details

Outer loop: bound the concave term $-G_s(\boldsymbol{\mu})$ by

$$-G_s(\boldsymbol{\mu}) \geq -G_s(\boldsymbol{\mu}^*) - \frac{\partial G_s(\boldsymbol{\mu}^*)}{\partial \boldsymbol{\mu}^T}(\boldsymbol{\mu} - \boldsymbol{\mu}^*) = -(\boldsymbol{\lambda}_s^*)^T(\boldsymbol{\mu} - \boldsymbol{\mu}^*)$$

where $\boldsymbol{\mu}^*$ is current estimate and $\boldsymbol{\lambda}_s^* = \boldsymbol{\lambda}_s(\boldsymbol{\mu}^*)$.

Eliminate $\boldsymbol{\lambda}_r$ from $\min_{\boldsymbol{\mu}} G^{\text{EC,ubound}}(\boldsymbol{\mu})$: $\boldsymbol{\lambda}_q + \boldsymbol{\lambda}_r - \boldsymbol{\lambda}_s^* = 0$

Inner loop: Solve concave problem in $\boldsymbol{\lambda}_q$:

$$\max_{\boldsymbol{\lambda}_q} \{-\ln Z_q(\boldsymbol{\lambda}_q) - \ln Z_r(\boldsymbol{\lambda}_s^* - \boldsymbol{\lambda}_q)\} : \langle \mathbf{g}(\mathbf{x}) \rangle_q = \langle \mathbf{g}(\mathbf{x}) \rangle_r$$

After convergence update $\boldsymbol{\mu}^* = \langle \mathbf{g}(\mathbf{x}) \rangle_q$

Simulations – Ising Models

N binary variables with pairwise interactions J_{ij}:

$$\begin{aligned} p(\mathbf{x}) &= \frac{1}{Z}\prod_i \Psi_i(x_i)\exp\left(\frac{1}{2}\mathbf{x}^T\mathbf{J}\mathbf{x}\right) \\ \Psi_i(x_i) &= [\delta(x_i+1)+\delta(x_i-1)]e^{\theta_i x_i} \end{aligned}$$

Look at the approximation for the

- One-variable marginals $p(x_i)=\frac{1+x_i m_i}{2}$, mean $m_i=\langle x_i\rangle$.
- Two-variable marginals $p(x_i,x_j)=\frac{x_i x_j C_{ij}}{4}+p(x_i)p(x_j)$, covariance $C_{ij}=\langle x_i x_j\rangle-\langle x_i\rangle\langle x_j\rangle$.
- Free energy $G=-\ln Z$.

Methods Compared

- Exact.
- Factorized expectation consistent.
- Spanning tree structured expectation consistent.
- Bethe (and Kikuchi) approximation.
- Log-determinant relaxation (Wainwright & Jordan, 2002).

Scenario I: Kappen and Albers

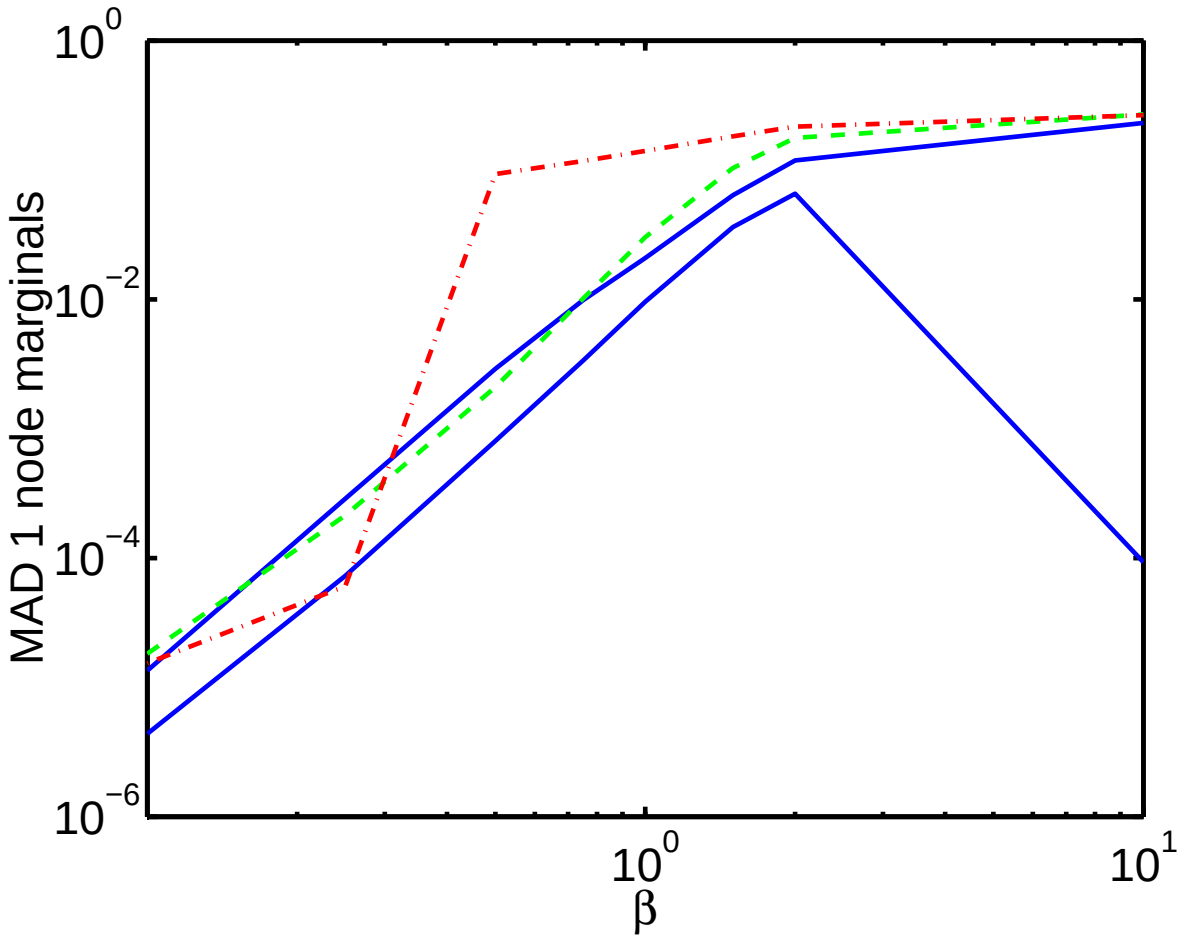
$N = 10$, $J_{ij} = \beta w_{ij}$, $w_{ij} \sim \mathcal{N}(0,1)$ and $\beta \in [0.1; 10]$.

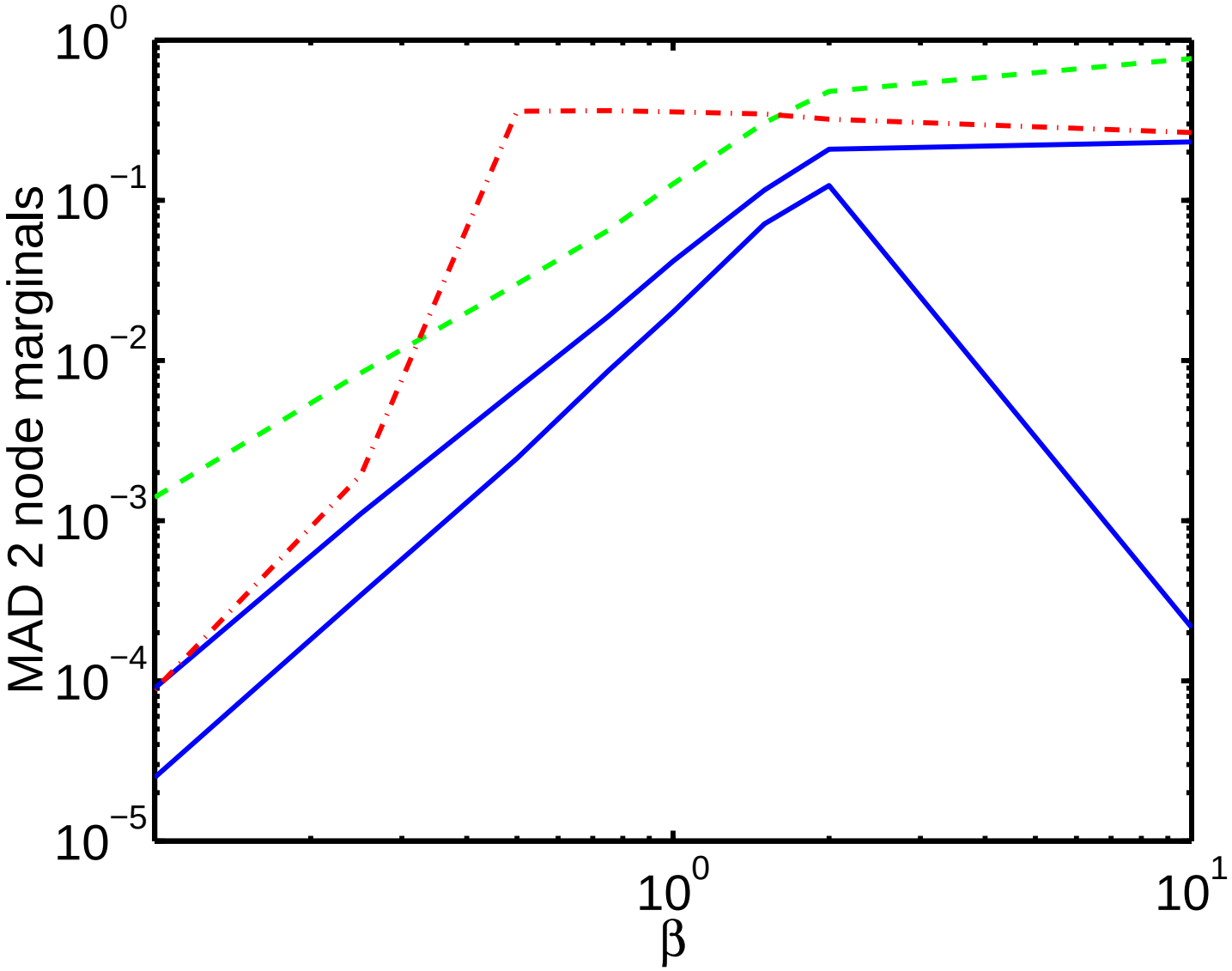
Error-measures:

$$\begin{aligned}
\text{MAD}_1 &= \max_i |p(x_i = 1) - p(x_i = 1|\text{Method})| \\
\text{MAD}_2 &= \max_{i,j} \max_{x_i=\pm 1, x_j=\pm 1} \left|p(x_i, x_j) - p(x_i, x_j|\text{Method})\right| \\
\text{AD Free energy} &= \left|G - G^{\text{Method}}\right|
\end{aligned}$$

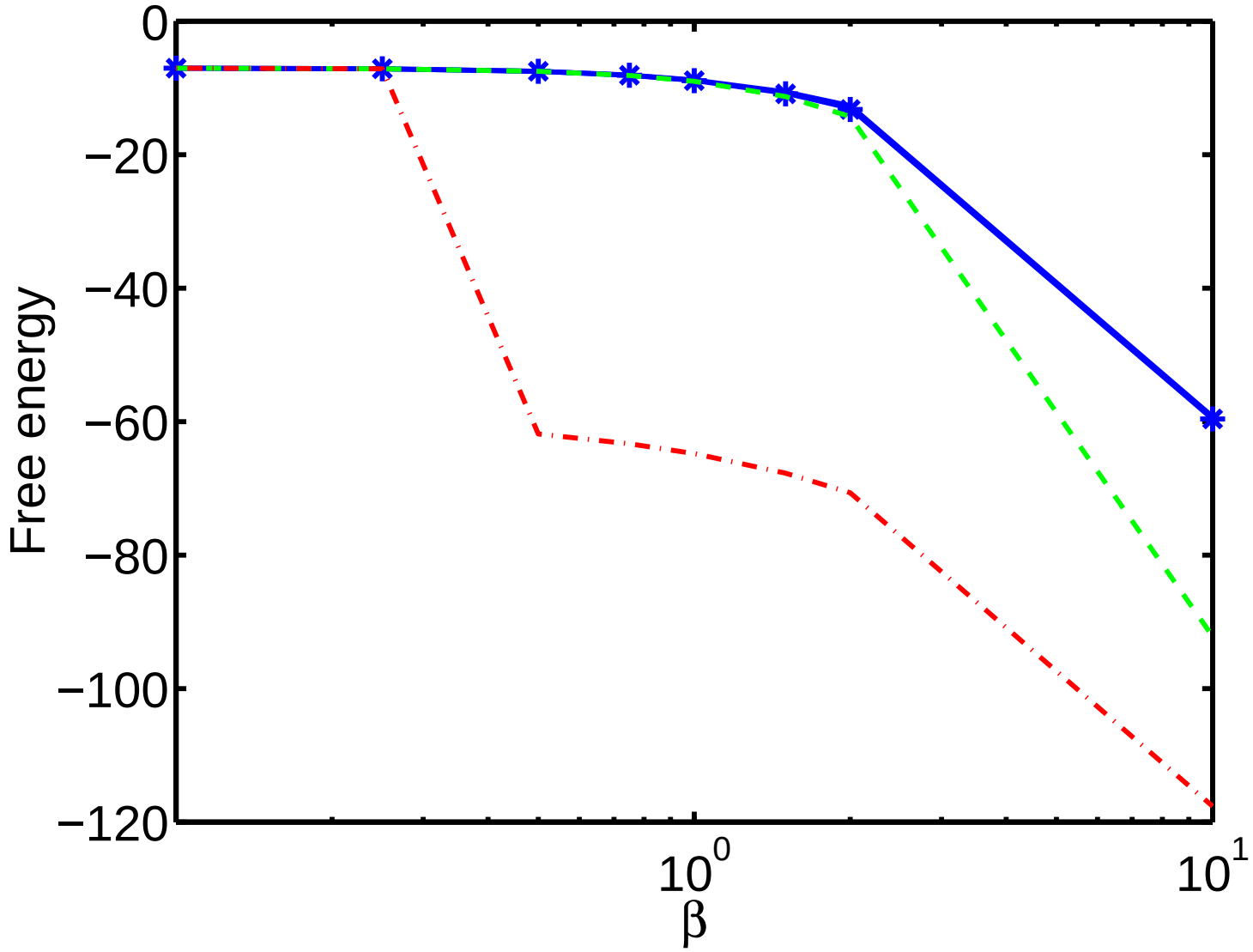
In EC, the non-trivial correlation estimates: $C_{ij} = \left[(\mathbf{\Lambda}_r - \mathbf{J})^{-1}\right]_{ij}$ is used for the two-variables marginals.

Maximal absolute deviation (MAD) for one-variable marginals. Blue upper full line: EC factorized, blue lower full line EC tree, green dashed line: Bethe and red dash-dotted line: Kikuchi.

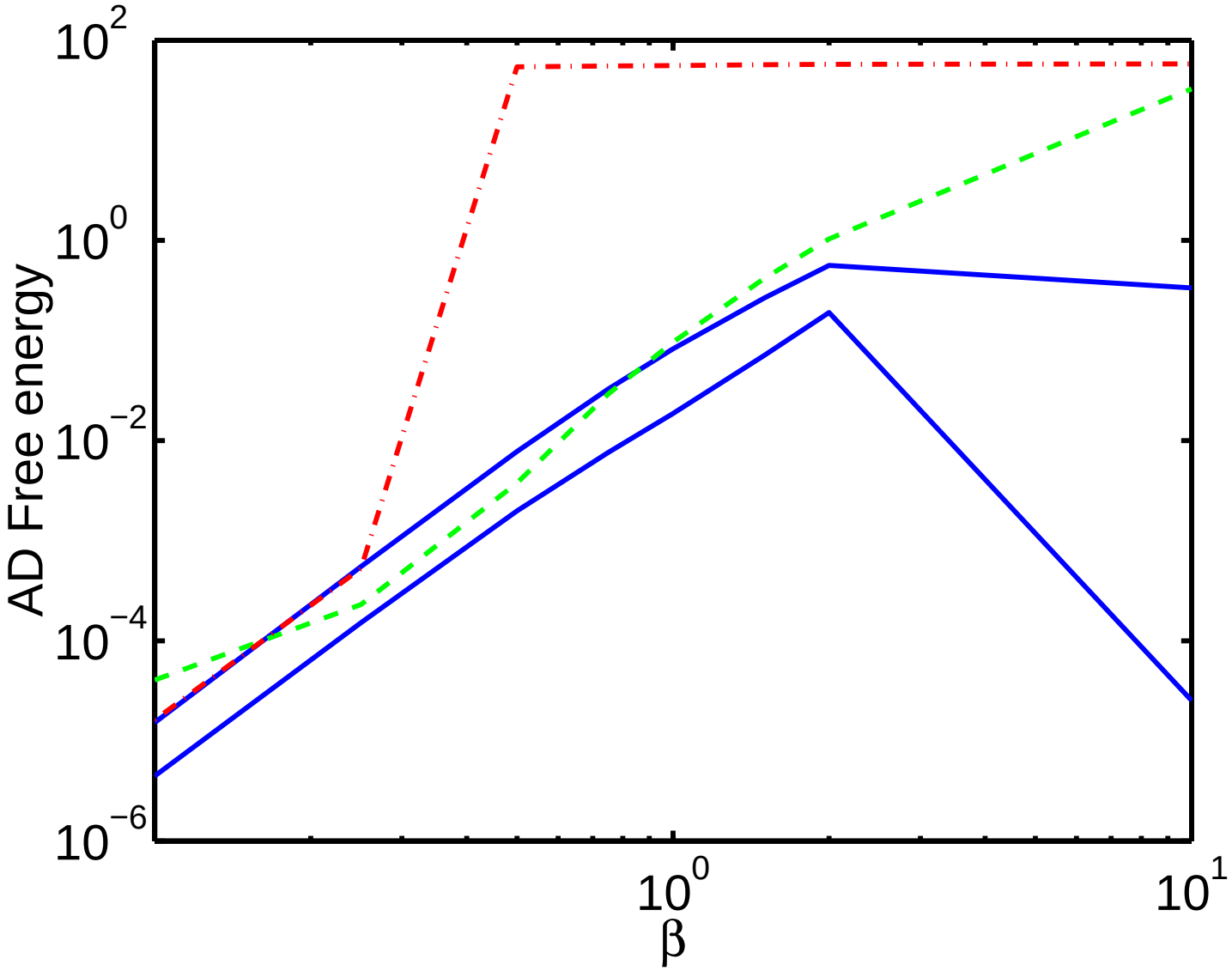




ec016



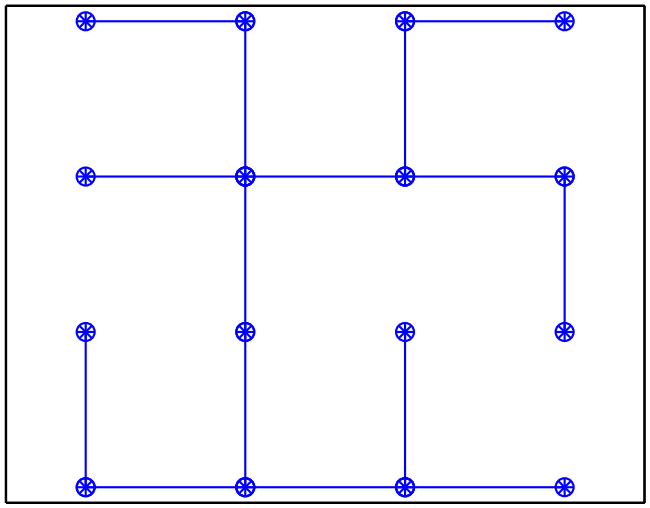
ec016



ec016

Scenario II: Wainwright and Jordan

$N = 16$
Fully connected or
4-by-4 nearest neighbor grid.



Coupling strength:

- repulsive (anti-ferromagnetic) $J_{ij} \sim \mathcal{U}[-2d_{\text{coup}}, 0]$,
- mixed $J_{ij} \sim \mathcal{U}[-d_{\text{coup}}, +d_{\text{coup}}]$ and
- attractive (ferromagnetic) $J_{ij} \sim \mathcal{U}[0, +2d_{\text{coup}}]$ with $d_{\text{coup}} > 0$.

θ_i from *uniform* distribution: $\theta_i \sim \mathcal{U}[-d_{\text{obs}}, d_{\text{obs}}]$ with $d_{\text{obs}} = 0.25$.

Problem type			Method					
			SP	LD	EC factorized		EC tree	
Graph	Coupling	d_{coup}	Mean	Mean	Mean $\pm$ std	Max	Mean $\pm$ std	Max
Full	Repulsive	0.25	0.037	0.020	0.003 $\pm$ 0.002	0.00	0.0017 $\pm$ 0.0011	0.007
	Repulsive	0.50	0.071	0.018	0.031 $\pm$ 0.045	0.20	0.0143 $\pm$ 0.0141	0.102
	Mixed	0.25	0.004	0.020	0.002 $\pm$ 0.002	0.00	0.0013 $\pm$ 0.0008	0.005
	Mixed	0.50	0.055	0.021	0.022 $\pm$ 0.030	0.17	0.0151 $\pm$ 0.0204	0.163
	Attractive	0.06	0.024	0.027	0.004 $\pm$ 0.002	0.01	0.0025 $\pm$ 0.0014	0.007
	Attractive	0.12	0.435	0.033	0.117 $\pm$ 0.090	0.30	0.0211 $\pm$ 0.0307	0.159
Grid	Repulsive	1.0	0.294	0.047	0.153 $\pm$ 0.123	0.58	0.0031 $\pm$ 0.0021	0.013
	Repulsive	2.0	0.342	0.041	0.198 $\pm$ 0.135	0.49	0.0021 $\pm$ 0.0010	0.009
	Mixed	1.0	0.014	0.016	0.011 $\pm$ 0.010	0.08	0.0018 $\pm$ 0.0011	0.006
	Mixed	2.0	0.095	0.038	0.082 $\pm$ 0.081	0.32	0.0068 $\pm$ 0.0053	0.028
	Attractive	1.0	0.440	0.047	0.125 $\pm$ 0.104	0.36	0.0028 $\pm$ 0.0018	0.013
	Attractive	2.0	0.520	0.042	0.177 $\pm$ 0.125	0.41	0.0024 $\pm$ 0.0022	0.016

Error measure (averaged over 100 trials)

$$\text{MeanAD} = \sum_i |p(\mathbf{x}_i = 1) - p(\mathbf{x}_i = 1|\text{Method})|/N \ .$$

SP = Sum Product = Bethe

LD = log determinant relaxation.

Further Approximations – Iterate EC

Belief networks

$$f(\mathbf{x}) = \prod_i \psi_i(x_i) \prod_k \phi_k \left(\sum_j w_{kj} x_j \right)$$

Set $f_q(\mathbf{x}) = \prod_i \psi_i(x_i)$ and $f_r(\mathbf{x}) = \prod_k \phi_k \left(\sum_j w_{kj} x_j \right)$.

$r(\mathbf{x})$ not tractable – change of variables $u_k = \sum_j w_{kj} x_j$,

$$r(\mathbf{u}) = \prod_k \phi_k (u_k) \exp \left(\frac{1}{2} \mathbf{u}^T \mathbf{J} \mathbf{u} + \mathbf{h}^T \mathbf{u} \right)$$

Split into new factors: $\widehat{f}_q(\mathbf{u})$ and $\widehat{f}_r(\mathbf{u})$: tractable $\widehat{q}(\mathbf{u})$ and $\widehat{r}(\mathbf{u})$.

Iterate EC – Mixture Models

Example Bayes mixture of Gaussians:

$$p(\mathbf{Y}, \{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}) = \prod_i \left(\sum_k \pi_k p(\mathbf{y}_i | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right) p(\{\pi_k\}) p(\{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}) \ ,$$

$\mathbf{x} = \{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}$.

$$f(\mathbf{x}) = \prod_i^{N_{\text{ex}}} f_i(\mathbf{x}) p_0(\mathbf{x})$$

Iterate approximation N_{ex} times to get tractable $q_i(\mathbf{x})$:

$$\begin{aligned} q_i(\mathbf{x}) &= \frac{1}{Z_i(\boldsymbol{\lambda})} f_i(\mathbf{x}) \exp(\boldsymbol{\lambda}^T \mathbf{g}(\mathbf{x})) p_0(\mathbf{x}) \\ s(\mathbf{x}) &= \frac{1}{Z_0(\boldsymbol{\lambda})} \exp(\boldsymbol{\lambda}^T \mathbf{g}(\mathbf{x})) p_0(\mathbf{x}) \end{aligned}$$

Free Energy Mixture Models

$$
\begin{aligned}
-\ln Z^{\mathrm{EC}} &= -\ln Z_0(\sum_i \boldsymbol{\lambda}_{q,i}) - \sum_i \ln Z_i(\boldsymbol{\lambda}_{s,i} - \boldsymbol{\lambda}_{q,i}) + \sum_i \ln Z_0(\boldsymbol{\lambda}_{s,i}) \\
Z_0(\boldsymbol{\lambda}) &= \int d\mathbf{x} \exp(\boldsymbol{\lambda}^T \mathbf{g}(\mathbf{x})) p_0(\mathbf{x}) \\
Z_i(\boldsymbol{\lambda}) &= \int d\mathbf{x} f_i(\mathbf{x}) \exp(\boldsymbol{\lambda}^T \mathbf{g}(\mathbf{x})) p_0(\mathbf{x})
\end{aligned}
$$

Expectation consistency: $\sum_{i'} \boldsymbol{\lambda}_{q,i'} = \boldsymbol{\lambda}_{s,i} = \boldsymbol{\lambda}_s$

$$
-\ln Z^{\mathrm{EC}} = -\sum_i \ln Z_i(\sum_{i' \neq i} \boldsymbol{\lambda}_{q,i'}) + (N_{\mathrm{ex}} - 1) \ln Z_0(\sum_i \boldsymbol{\lambda}_{q,i})
$$

Similar to Aspect model (Minka & Lafferty).

Beyond EC – Higher Order Models

(Not so) **low density parity check decoding**

$$p(\mathbf{x}) \propto \prod_m^M \exp\left(J_m \prod_{i_m} x_{i_m}\right) \prod_i^N \exp(h_i x_i)$$

Bayesian treatment of **linear models**

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \boldsymbol{\epsilon} \ .$$

Distributions over $\mathbf{A}$, $\mathbf{x}$ and $\boldsymbol{\epsilon}$.

Mean field theory:
identify statistics that can be approximated with Gaussians.

Summary and Conclusions

Expectation consistent global approximations $q(\mathbf{x})$ and $r(\mathbf{x})$.

Approximation works because of CLT. We are averaging more (than in variational Bayes) and not over-counting (as opposed to loopy BP).

Non-trivial estimates of correlations.

Closely related to Minka's EP and Opper & Winther's adaptive TAP.

Not possible to use for all models (where variational and/or Bethe apply). Further approximations needed.

Expectation Consistent Free Energie for Approximate Inference

NIPS poster

Manfred Opper and Ole Winther

ISIS
School of Electronics and Computer Science
University of Southampton
SO17 1BJ, United Kingdom
mo@ecs.soton.ac.uk

Informatics and Mathematical Modelling
Technical University of Denmark
DK-2800 Lyngby, Denmark
owi@imm.dtu.dk

Abstract

We propose a novel framework for deriving approximations for intractable probabilistic models. This framework is based on a free energy (negative log marginal likelihood) and can be seen as a generalization of adaptive TAP [1-3] and expectation propagation (EP) [4,5] The free energy is constructed from two approximating distributions which encode different aspects of the intractable model such a single node constraints and couplings and are by construction consistent on a chosen set of moments. We test the framework on a difficult benchmark problem with binary variables on fully connected graphs and 2D grid graphs. We find good performance using sets of moments which either specify factorized nodes or a spanning tree on the nodes (structured approximation). Surprisingly, the Bethe approximation gives very inferior results even on grids.

Approximate inference

Compute expectations over distribution

$$\boxed{p(\mathbf{x}) = \frac{1}{Z} f(\mathbf{x})}$$

with random variables $\mathbf{x} = (x_1, x_2, \ldots, x_N)$ and *partition function* $Z = \int d\mathbf{x} f(\mathbf{x})$.

Intractability arises either because the necessary sums are over a too large number of variables or because multivariate integrals cannot be evaluated exactly.

Many application areas: Loopy belief propagation, mixture models, factor models, independent component analysis, Gaussian processes, bootstrap methods for kernel machines, etc.

Tractability from simpler forms

In a typical scenario, $f(\mathbf{x})$ is expressed as a product of two functions

$$f(\mathbf{x}) = f_1(\mathbf{x}) f_2(\mathbf{x}) \tag{1}$$

with $f_{1,2}(\mathbf{x}) \geq 0$, where f_1 is "simple" enough to allow for tractable computations. Approximate inference (e.g. variational) make substitution

$$f_2(\mathbf{x}) \rightarrow \exp\left(\boldsymbol{\lambda}^T \mathbf{g}(\mathbf{x})\right) \equiv \exp\left(\sum_{j=1}^{K} \lambda_j g_j(\mathbf{x})\right)$$

such that computations becomes tractable. But how to choose $\boldsymbol{\lambda}$?

In the *expectation consistent* framework: $\boldsymbol{\lambda}$ is chosen such that two different *global approximations* $q(\mathbf{x})$ and $r(\mathbf{x})$ agree on a chosen *set of moments* of the distributions: $\langle \mathbf{g}(\mathbf{x}) \rangle_{q(\mathbf{x})} = \langle \mathbf{g}(\mathbf{x}) \rangle_{r(\mathbf{x})}$.

It is convenient to use the Gibbs free energy to formalize this.

We will discuss relation to other approaches at the end!

Gibbs free energy – Two-stage optimization

Introduce *trial distribution* $q(\mathbf{x})$.

Step 1: fix a set of *generalized moments* $\langle \mathbf{g}(\mathbf{x}) \rangle_q$ Definition of Gibbs Free Energy $G(\boldsymbol{\mu})$:

$$\boxed{G(\boldsymbol{\mu}) = \min_q \{KL(q,p) \mid \langle \mathbf{g}(\mathbf{x}) \rangle_q = \boldsymbol{\mu}\} - \ln Z} \tag{2}$$

with KL-divergence

$$KL(q,p) = \int d\mathbf{x}\; q(\mathbf{x}) \ln \frac{q(\mathbf{x})}{p(\mathbf{x})} \; . \tag{3}$$

Step 2: Optimize wrt. moments

$$\min_{\boldsymbol{\mu}} G(\boldsymbol{\mu}) = -\ln Z \quad \text{and} \quad \langle \mathbf{g} \rangle = \operatorname{argmin}_{\boldsymbol{\mu}} G(\boldsymbol{\mu}) \; . \tag{4}$$

Gibbs free energy – properties

Explicit form of the optimized trial distribution, $Z(\boldsymbol{\lambda}) = \int d\mathbf{x}\, f(\mathbf{x}) \exp\left(\boldsymbol{\lambda}^T \mathbf{g}(\mathbf{x}\right.$

$$\boxed{q(\mathbf{x}) = \frac{f(\mathbf{x})}{Z(\boldsymbol{\lambda})} \exp\left(\boldsymbol{\lambda}^T \mathbf{g}(\mathbf{x})\right) \ ,} \tag{5}$$

The set of *Lagrange parameters* $\boldsymbol{\lambda} = \boldsymbol{\lambda}(\boldsymbol{\mu})$ (often called messages in belief propagation) is chosen such that the conditions $\langle \mathbf{g}(\mathbf{x}) \rangle_q = \boldsymbol{\mu}$ are fulfilled, i.e. $\boldsymbol{\lambda}$ satisfies

$$\frac{\partial \ln Z(\boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}} = \boldsymbol{\mu} \ . \tag{6}$$

Inserting the optimized trial distribution in $G(\boldsymbol{\mu})$:

$$\boxed{G(\boldsymbol{\mu}) = -\ln Z(\boldsymbol{\lambda}(\boldsymbol{\mu})) + \boldsymbol{\lambda}^T(\boldsymbol{\mu})\boldsymbol{\mu} = \max_{\boldsymbol{\lambda}} \left\{ -\ln Z(\boldsymbol{\lambda}) + \boldsymbol{\lambda}^T \boldsymbol{\mu} \right\} \ ,} \tag{7}$$

i.e. G is the *Legendre transform* or *dual* of $-\ln Z(\boldsymbol{\lambda})$ and is convex.

Examples – factorized, tree and Gaussian

Completely factorized, i.e. $p(\mathbf{x}) = \prod_i \psi_i(x_i)$. For simplicity we will consider biased binary variables: $\Psi_i(x_i) = [\delta(x_i + 1) + \delta(x_i - 1)]e^{\theta_i x_i}$ and fix the first moments $\mathbf{m} = \langle \mathbf{x} \rangle$. Denoting the conjugate Lagrange parameters by $\boldsymbol{\gamma}$:

$$G(\mathbf{m}) = \sum_i G_i(m_i) \quad \text{with} \quad G_i(m_i) = \max_{\gamma_i} \{-\ln Z_i(\gamma_i) + m_i \gamma_i\} \quad (8)$$

and $Z_i(\gamma_i) = \int dx_i \; \Psi_i(\xi) e^{\gamma_i x_i} = 2\cosh(\gamma_i + \theta_i)$.

Tree-connected graph. For the case where either the couplings and the moments together define a tree-connected graph, we can write the free energy in term of single- and two-node free energies. Considering again completely factorized binary variables, all non-trivial moments on the graph $(ij) \in \mathcal{G}$ are the means $\mathbf{m}$ and correlations of linked nodes $M_{ij} = \langle x_i x_j \rangle$:

$$G(\mathbf{m}, \{M_{ij}\}_{(ij)\in\mathcal{G}}) = \sum_{(ij)\in\mathcal{G}} G_{ij}(m_i, m_j, M_{ij}) + \sum_i (1 - n_i) G_i(m_i) \; , \quad (9)$$

where $G_{ij}(m_i, m_j, M_{ij})$ is the two-node free energy defined in a similar fashion as the one-node free energy, n_i the number of links to node i and $G_i(m_i)$ is the one-node free energy.

Gaussian distribution. We set $\boldsymbol{\mu} = (\mathbf{m}, \mathbf{M})$ with all first moments $\mathbf{m}$ and an arbitrary subset of second moments $\mathbf{M}$ for a Gaussian model $\Psi_i(x_i) \propto \exp[a_i x_i - \frac{b_i}{2} x_i^2]$ and $p(\mathbf{x}) \propto \prod_i \Psi_i(x_i) \exp(\mathbf{x}^T \mathbf{J} \mathbf{x}/2)$. We introduce conjugate variables $\boldsymbol{\gamma}$ and $-\boldsymbol{\Lambda}/2$. $\boldsymbol{\gamma}$ can be eliminated analytically, whereas we get a log-determinant maximization problem for $\boldsymbol{\Lambda}$ [6]:

$$\begin{aligned} G(\mathbf{m}, \mathbf{M}) &= -\frac{1}{2}\mathbf{m}^T \mathbf{J} \mathbf{m} - \mathbf{m}^T \mathbf{a} + \frac{1}{2} \sum_i M_{ii} b_i \\ &\quad + \max_{\boldsymbol{\Lambda}} \left\{ \frac{1}{2} \ln \det(\boldsymbol{\Lambda} - \mathbf{J}) - \frac{1}{2} \operatorname{Tr} \boldsymbol{\Lambda} (\mathbf{M} - \mathbf{m}\mathbf{m}^T) \right\} . \end{aligned} \tag{10}$$

Exact interpolation for Gibbs free energy

Introduce smooth interpolant on $f_2(\mathbf{x}) \to f_2(\mathbf{x}, t)$,

$$f_2(\mathbf{x}, t = 0) = 1 \quad \text{and} \quad f_2(\mathbf{x}, t = 1) = f_2(\mathbf{x}) \ , \qquad 0 \leq t \leq 1$$

$$q(\mathbf{x}|t) = \frac{1}{Z_q(\boldsymbol{\lambda}, t)} f_1(\mathbf{x}) f_2(\mathbf{x}, t) \exp\left(\boldsymbol{\lambda}^T \mathbf{g}(\mathbf{x})\right) \tag{11}$$

$$G_q(\boldsymbol{\mu}, t) = \max_{\boldsymbol{\lambda}} \left\{ - \ln Z_q(\boldsymbol{\lambda}, t) + \boldsymbol{\lambda}^T \boldsymbol{\mu} \right\} \ . \tag{12}$$

Interpolation between exact $G(\boldsymbol{\mu}) = G_q(\boldsymbol{\mu}, t = 1)$ and 'free model'

$$G_q(\boldsymbol{\mu}, 1) - G_q(\boldsymbol{\mu}, 0) = \int_0^1 dt \, \frac{dG_q(\boldsymbol{\mu}, t)}{dt} = - \int_0^1 dt \left\langle \frac{d \ln f_2(\mathbf{x}, t)}{dt} \right\rangle_{q(\mathbf{x}|t)} \ .$$

because $\frac{\partial \ln Z(\boldsymbol{\lambda}, t)}{\partial t} = \left\langle \frac{d \ln f_2(\mathbf{x}, t)}{dt} \right\rangle_{q(\mathbf{x}|t)}$ and saddlepoint condition:

$$\frac{dG(\boldsymbol{\mu}, t)}{dt} = - \frac{\partial \ln Z(\boldsymbol{\lambda}, t)}{\partial t} + \left(\boldsymbol{\mu} - \frac{\partial \ln Z(\boldsymbol{\lambda}, t)}{\partial \boldsymbol{\lambda}} \right) \frac{d\boldsymbol{\lambda}^T}{dt} = - \frac{\partial \ln Z(\boldsymbol{\lambda}, t)}{\partial t} \ .$$

Expectation consistent (EC) approximation

Introduce second tractable familiy (the first is the 'free' distribution $q(\mathbf{x}, t = 0)$)

$$r(\mathbf{x}|t) = \frac{1}{Z_r(\boldsymbol{\lambda}, t)} f_2(\mathbf{x}, t) \exp\left(\boldsymbol{\lambda}^T \mathbf{g}(\mathbf{x})\right) \ , \tag{13}$$

Note the $f_1(\mathbf{x})$-factor does not appear. Again parameters $\boldsymbol{\lambda}$ will be chosen to guarantee *consistency for the expectations* of $\mathbf{g}$, i.e. $\langle \mathbf{g}(\mathbf{x}) \rangle_{r(\mathbf{x}|t)} = \boldsymbol{\mu}$

Using $r(\mathbf{x}|t)$ instead of $q(\mathbf{x}|t)$ gives us the central approximation

$$G_q(\boldsymbol{\mu}, 1) - G_q(\boldsymbol{\mu}, 0) \approx -\int_0^1 dt \left\langle \frac{d \ln f_2(\mathbf{x}, t)}{dt} \right\rangle_{r(\mathbf{x}|t)} = G_r(\boldsymbol{\mu}, 1) - G_r(\boldsymbol{\mu}, 0) \ .$$

The last equality holds because q and r contain the same (exponential) family.

$$\boxed{G_q(\boldsymbol{\mu}, 1) \approx G_q(\boldsymbol{\mu}, 0) + G_r(\boldsymbol{\mu}, 1) - G_r(\boldsymbol{\mu}, 0) \equiv G^{\mathsf{EC}}(\boldsymbol{\mu}) \ .}$$

Simplified notation: $G_q \equiv G_q(\boldsymbol{\mu}, 0)$, $G_r \equiv G_r(\boldsymbol{\mu}, 1)$ and $G_s \equiv G_r(\boldsymbol{\mu}, 0)$.

Variational approximation

The main advantage of the EC approximation is that it takes into account interaction of the variables in the $f_2(\mathbf{x})$ part retained in $r(\mathbf{x})$ distribution. The EC approximation can also be justified by central limit theory (cavity) arguments. See below for a discussion of when to expect the different types of approximations to work well.

If the interpolant is $f_2(\mathbf{x},t) = [f_2(\mathbf{x})]^t$, we can recover the *variational approximation* by replacing the average over $q(\mathbf{x}|t)$ with an average over the free model, $q(\mathbf{x}|0)$:

$$\begin{aligned} G(\boldsymbol{\mu}) &\approx G(\boldsymbol{\mu},0) - \int_0^1 dt \left\langle \frac{d \ln f_2(\mathbf{x},t)}{dt} \right\rangle_{q(\mathbf{x}|0)} \\ &= G(\boldsymbol{\mu},0) - \langle \ln f_2(\mathbf{x}) \rangle_{q(\mathbf{x}|0)} = G^{\text{var}}(\boldsymbol{\mu}) \ . \end{aligned}$$

In the variational approx. we are neglecting even more of the interaction part!

Pairwise potentials

$$p(\mathbf{x}) = \frac{1}{Z} \prod_{\alpha} \Psi_{\alpha}(\mathbf{x}_{\alpha}) \ \exp\left(\sum_{i<j} x_i J_{ij} x_j \right) \ , \tag{14}$$

where the $\mathbf{x}_{\alpha}$ denote *tractable non-Gaussian* potentials defined on disjoint subsets of variables $\mathbf{x}_{\alpha}$, (e.g. factorized or a spanning tree).

Fix $m_i = \langle x_i \rangle$ and $M_{ij} = \langle x_i x_j \rangle$ and take as our second tractable family $r(\mathbf{x})$, the *Gaussian part of* $p(\mathbf{x})$, i.e. $f_2(\mathbf{x}) = \exp\left(\sum_{i<j} x_i J_{ij} x_j\right)$, then G_r and G_s will be free energies of a Gaussian with $\mathbf{J}$ and $\mathbf{J} = \mathbf{0}$, respectively:

$$\begin{aligned} G^{\mathsf{EC}}(\mathbf{m}, \mathbf{M}) \ = \ & G_q(\mathbf{m}, \mathbf{M}, 0) - \frac{1}{2} \mathbf{m}^T \mathbf{J} \mathbf{m} \\ & + \max_{\mathbf{\Lambda}} \left\{ \frac{1}{2} \ln \det(\mathbf{\Lambda} - \mathbf{J}) - \frac{1}{2} \operatorname{Tr} \mathbf{\Lambda} (\mathbf{M} - \mathbf{m}\mathbf{m}^T) \right\} \\ & - \max_{\mathbf{\Lambda}} \left\{ \frac{1}{2} \ln \det \mathbf{\Lambda} - \frac{1}{2} \operatorname{Tr} \mathbf{\Lambda} (\mathbf{M} - \mathbf{m}\mathbf{m}^T) \right\} \ , \end{aligned} \tag{15}$$

where the free energy $G_q(\mathbf{m}, \mathbf{M}, 0)$ will depend explicitly upon the potentials $\Psi_\alpha(\mathbf{x}_\alpha)$.

What can we get non-trivial estimates for in EC?

The two complementary approximations $q(\mathbf{x})$ and $r(\mathbf{x})$ (here exemplified for pairwise interactions) give:

- Marginal distributions

  $$q(\mathbf{x}) \propto \prod_i \Psi_i(x_i) \exp(\boldsymbol{\gamma}_q^T \mathbf{x} - \mathbf{x}^T \boldsymbol{\Lambda}_q \mathbf{x}/2)$$

  is tractable and includes the *non-trivial constraints* on the variables. For e.g. factorized moments, the marginals are:

  $$q(x_i) \propto \Psi_i(x_i) \exp(\gamma_{q,i} x_i - x_i^2 \Lambda_{q,i}/2) \ .$$

- Correlations

  $$r(\mathbf{x}) \propto \exp(\boldsymbol{\gamma}_r^T \mathbf{x} - \mathbf{x}^T (\boldsymbol{\Lambda}_r - \mathbf{J}) \mathbf{x}/2)$$

  is a global Gaussian approximation with non-trivial covariance $C(x_i, x_j) = \langle x_i x_j \rangle_{r(\mathbf{x})} - \langle x_i \rangle_{r(\mathbf{x})} \langle x_j \rangle_{r(\mathbf{x})} = \left[(\boldsymbol{\Lambda}_r - \mathbf{J})^{-1} \right]_{ij}$.

- The free energy $G^{\mathrm{EC}} \approx -\ln Z$. Useful in Bayesian statistics since Z is the *marginal likelihood* (or *evidence*) of the model.

EC free energy is upper bounded by variational free energy

Algorithmics

Solving the optimization problem $\min_{\boldsymbol{\mu}} G^{\mathrm{EC}}(\boldsymbol{\mu})$ is non-trivial: it may be non-convex:

$$G^{\mathrm{EC}}(\boldsymbol{\mu}) = G_q(\boldsymbol{\mu}) + G_r(\boldsymbol{\mu}) - G_s(\boldsymbol{\mu})$$

because it is a non-convex combination of (convex) free energies. We can use

- Guaranteed convergent – double loop, variational bounding [7].
- Gradient methods directly on $G(\boldsymbol{\mu})$ (or unconstrained transformation of $\boldsymbol{\mu}$, e.g. for $x_i = \pm 1$ use $\gamma_i = \tanh^{-1}(m_i)$ instead of mean value m_i.
- Expectation propagation [4,5,8].

Guaranteed convergent – Variational bounding, double loop

The basic idea is to minimize a decreasing sequence of *convex upper bounds* to G_{EC} [7,8,9]. Linearize concave term $-G_s(\boldsymbol{\mu})$ at the present iteration $\boldsymbol{\mu}^*$, $G_s(\boldsymbol{\mu}) \geq G_s^{\mathsf{lbound}}(\boldsymbol{\mu}) = -C_* + \boldsymbol{\mu}^T \boldsymbol{\lambda}_s^*$, $C_* \equiv \ln Z_q(\boldsymbol{\lambda}_s^*)$ and $\boldsymbol{\lambda}_s^* = \boldsymbol{\lambda}_s(\boldsymbol{\mu}^*)$.

$$
\begin{aligned}
G_{\mathsf{EC}}(\boldsymbol{\mu}) &\leq G_q(\boldsymbol{\mu}) + G_r(\boldsymbol{\mu}) - \boldsymbol{\mu}^T \boldsymbol{\lambda}_s^* + C_* \\
&= \min_{\boldsymbol{\mu}} \max_{\boldsymbol{\lambda}_q, \boldsymbol{\lambda}_r} \left\{ -\ln Z_q(\boldsymbol{\lambda}_q) - \ln Z_r(\boldsymbol{\lambda}_r) + \boldsymbol{\mu}^T (\boldsymbol{\lambda}_q + \boldsymbol{\lambda}_r - \boldsymbol{\lambda}_s^*) + C_* \right\} \\
&= \max_{\boldsymbol{\lambda}_q, \boldsymbol{\lambda}_r} \{ -\ln Z_q(\boldsymbol{\lambda}_q) - \ln Z_r(\boldsymbol{\lambda}_r) | \boldsymbol{\lambda}_q + \boldsymbol{\lambda}_r = \boldsymbol{\lambda}_s^* \} + C_* \\
&= \max_{\boldsymbol{\lambda}_r} \{ -\ln Z_q(\boldsymbol{\lambda}_s^* - \boldsymbol{\lambda}_r) - \ln Z_r(\boldsymbol{\lambda}_r) + C_* \} \quad . \qquad (16)
\end{aligned}
$$

Double loop recipe

1. Outer loop: For fixed old value $\boldsymbol{\mu}^*$, bound the concave term $-G_s(\boldsymbol{\mu})$ by $-G_s^{\text{lbound}}(\boldsymbol{\mu})$ go get the convex upper bound to $G_{\text{EC}}(\boldsymbol{\mu})$.

2. Inner loop: Solve the concave maximization problem

$$\max_{\boldsymbol{\lambda}_r} \mathcal{L} \quad \text{with} \quad \mathcal{L} = -\ln Z_q(\boldsymbol{\lambda}_s^* - \boldsymbol{\lambda}_r) - \ln Z_r(\boldsymbol{\lambda}_r) \ . \tag{17}$$

   Inserting the solution into $\boldsymbol{\mu}(\boldsymbol{\lambda}_r) = \langle \mathbf{g}(\mathbf{x}) \rangle_r$ gives new value $\boldsymbol{\mu}^* = \boldsymbol{\mu}$.

Currently, we either solve the non-linear inner-loop optimization by a sequential approach that are computationally efficient when G_r

is the free energy of a multivariate Gaussian or by interior point methods [6,10,11].

Unfortunately, this approach can be very slow especially for hard problems.

Expectation propagation (EP) [4,5,8]

EP can be interpreted as a greedy algorithm [8] for minimizing the EC free energy: Cycling over the factors

$$\tilde{\Psi}_i(x_i) = \exp(\boldsymbol{\lambda}_{r,i}^T \mathbf{g}_i(x_i)) \quad \text{and} \quad r(\mathbf{x}) = \prod_i \tilde{\Psi}_i(x_i) f_2(\mathbf{x})$$

for simplicity assuming that they contain only one variable.

1. Deletion of factor from $r(\mathbf{x})$**:** $r^{\backslash i}(\mathbf{x}) \propto r(\mathbf{x})/\tilde{\Psi}_i(x_i)$ or

$$r_i^{\backslash i}(x_i) \propto s_i(x_i)/\tilde{\Psi}_i(x_i) \propto \exp[(\boldsymbol{\lambda}_{s,i} - \boldsymbol{\lambda}_{r,i})^T \mathbf{g}_i(x_i)] \ ,$$

where $s_i(x_i) \propto \exp[\boldsymbol{\lambda}_{s,i}^T \mathbf{g}_i(x_i)]$ is the marginal distribution of $r(\mathbf{x})$.

2. Incorporate evidence $\Psi_i(x_i)$**:** $q_i(x_i) \propto \Psi_i(x_i) r_i^{\backslash i}(x_i)$ **and update sufficient statistics** $\mathbf{m}_i(x_i) = \langle \mathbf{g}_i(x_i) \rangle_{q_i}$**:**

3. Update factor $\tilde{\Psi}_i(x_i) \propto s_i(x_i)/r^{\setminus i}(\mathbf{x})$**:** First set the marginal distribution $s_i(x_i)$ to *match the moments*: $\mathbf{m}_i(x_i) = \langle \mathbf{g}_i(x_i) \rangle_{q_i}$. Finally, recalculate all the sufficient statistics from the new $r(\mathbf{x})$: $\mathbf{m} = \langle \mathbf{g}(\mathbf{x}) \rangle_r$.

We can identify all steps as sequentially updating the distributions and moments: $\boldsymbol{\lambda}_s$, $\boldsymbol{\lambda}_q$, $\boldsymbol{\mu}$, $\boldsymbol{\lambda}_s$, $\boldsymbol{\lambda}_r$, $\boldsymbol{\mu}$,... using the saddlepoint conditions on $\boldsymbol{\mu}$, $\boldsymbol{\lambda}_q$, $\boldsymbol{\lambda}_r$ and $\boldsymbol{\lambda}_s$.

This greedy approach is fast when it converges, unfortunately, it often fails especially for hard (non-convex?) problems.

Simulations – Ising models

N binary variables with pairwise interactionse J_{ij}:

$$\begin{aligned} p(\mathbf{x}) &= \frac{1}{Z}\prod_i \Psi_i(x_i)\exp\left(\frac{1}{2}\mathbf{x}^T\mathbf{J}\mathbf{x}\right) \\ \Psi_i(x_i) &= [\delta(x_i+1)+\delta(x_i-1)]e^{\theta_i x_i} \end{aligned}$$

Look at the approximation for the

- One-variable marginals $p(x_i) = \frac{1+x_i m_i}{2}$, mean $m_i = \langle x_i\rangle$.
- Two-variable marginals $p(x_i,x_j) = \frac{x_i x_j C_{ij}}{4} + p(x_i)p(x_j)$, covariance $C_{ij} = \langle x_i x_j\rangle - \langle x_i\rangle\langle x_j\rangle$.
- Free energy $G = -\ln Z$.

EC in practice – choosing $\mathbf{g}(\mathbf{x})$

Factorized restricted: consistency on $\langle x_i \rangle$, $i = 1, \ldots, N$ and $\sum_i \langle x_i^2 \rangle$

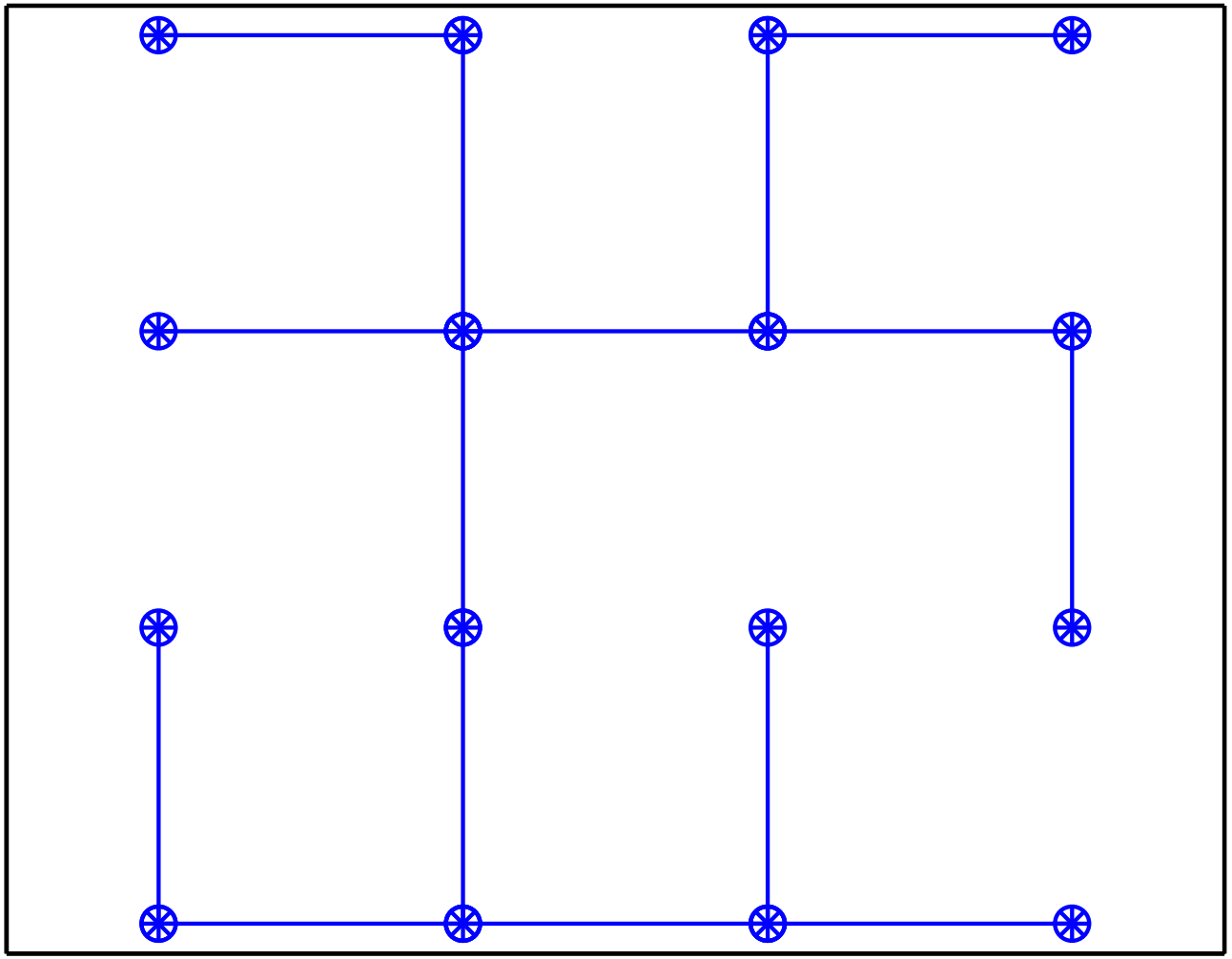
$$\begin{aligned} \mathbf{g}(\mathbf{x}) &= \left(x_1, \ldots, x_N, -\frac{1}{2} \sum_i \frac{x_i^2}{2} \right) \\ \boldsymbol{\lambda} &= (\gamma_1, \ldots, \gamma_N, \Lambda) \end{aligned}$$

Factorized: consistency on $\langle x_i \rangle$ and $\langle x_i^2 \rangle = 1$, $i = 1, \ldots, N$

$$\begin{aligned} \mathbf{g}(\mathbf{x}) &= \left(x_1, -\frac{x_1^2}{2}, \ldots, x_N, -\frac{x_N^2}{2} \right) \\ \boldsymbol{\lambda} &= (\gamma_1, \Lambda_1, \ldots, \gamma_N, \Lambda_N) \end{aligned}$$

Structured – spanning tree: as above and $M_{ij} = \langle x_i x_j \rangle$, $(ij) \in \mathcal{G}$

$$\begin{aligned} q(\mathbf{x}) &= \prod_{(ij) \in \mathcal{G}} \frac{q_{ij}(x_i, x_j)}{q_i(x_i) q_j(x_j)} \prod_i q_i(x_i) \\ G(\mathbf{m}, \{M_{ij}\}_{(ij) \in \mathcal{G}}) &= \sum_{(ij) \in \mathcal{G}} G_{ij}(m_i, m_j, M_{ij}) + \sum_i (1 - n_i) G_i(m_i) \end{aligned}$$



ec015

Methods compared

- Exact.
- Factorized expectation consistent.
- Spanning tree structured expectation consistent.
- Bethe (and Kikuchi) approximation.
- Log-determinant relaxation (Wainwright & Jordan, 2002).

Scenario I: Kappen and Albers

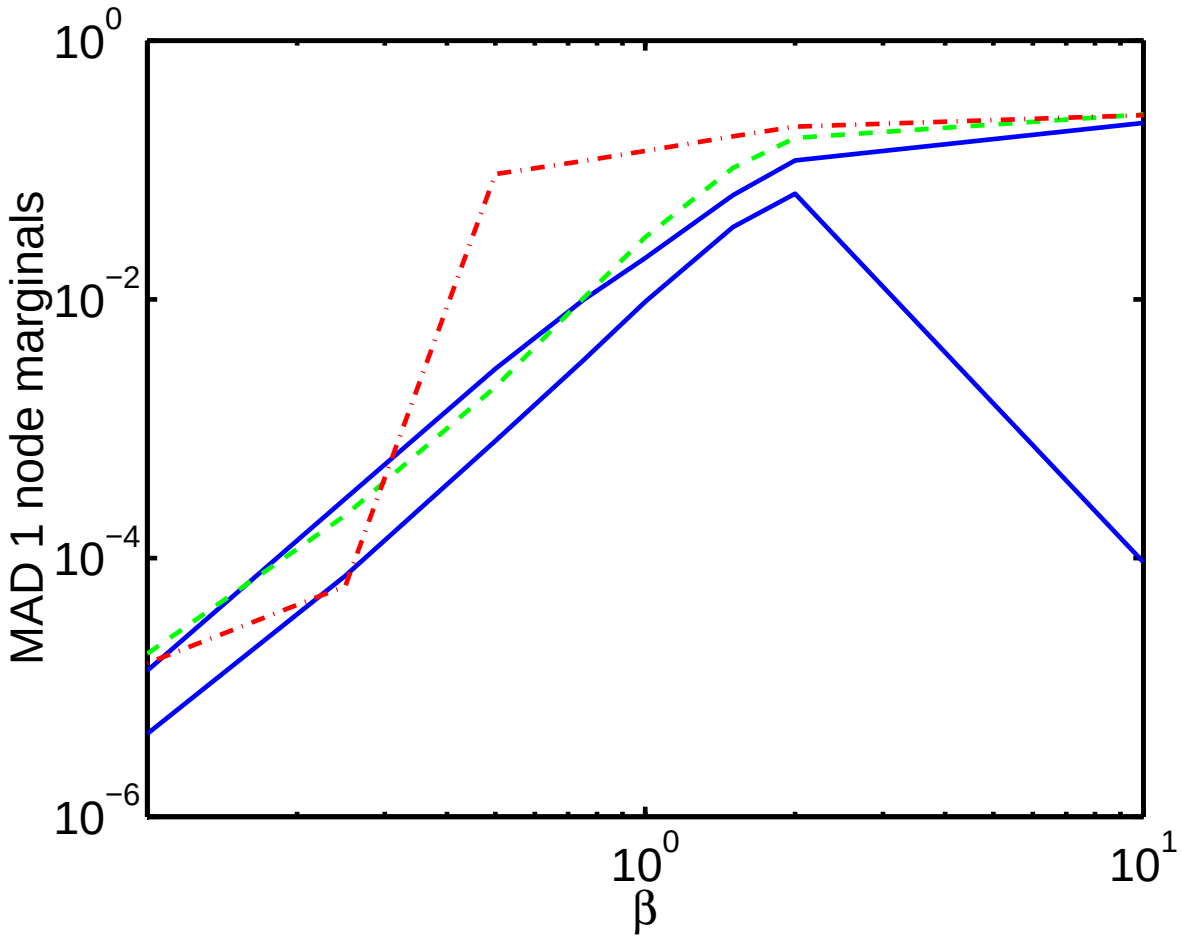
$N = 10$, $J_{ij} = \beta w_{ij}$, $w_{ij} \sim \mathcal{N}(0,1)$ and $\beta \in [0.1; 10]$.

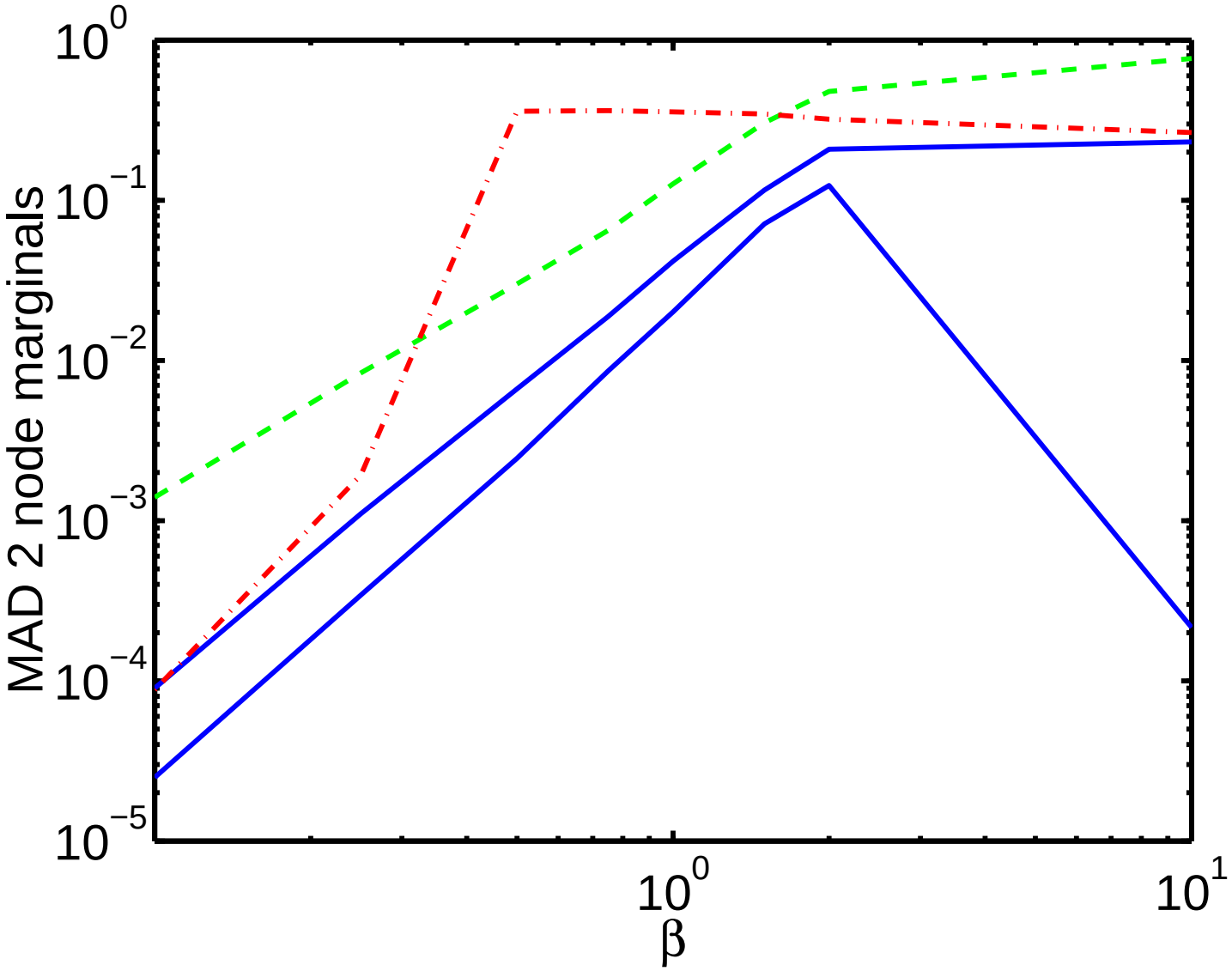
Error-measures:

$$\begin{aligned}
\text{MAD}_1 &= \max_i |p(x_i = 1) - p(x_i = 1|\text{Method})| \\
\text{MAD}_2 &= \max_{i,j} \max_{x_i=\pm 1, x_j=\pm 1} \left|p(x_i, x_j) - p(x_i, x_j|\text{Method})\right| \\
\text{AD Free energy} &= \left|G - G^{\text{Method}}\right|
\end{aligned}$$

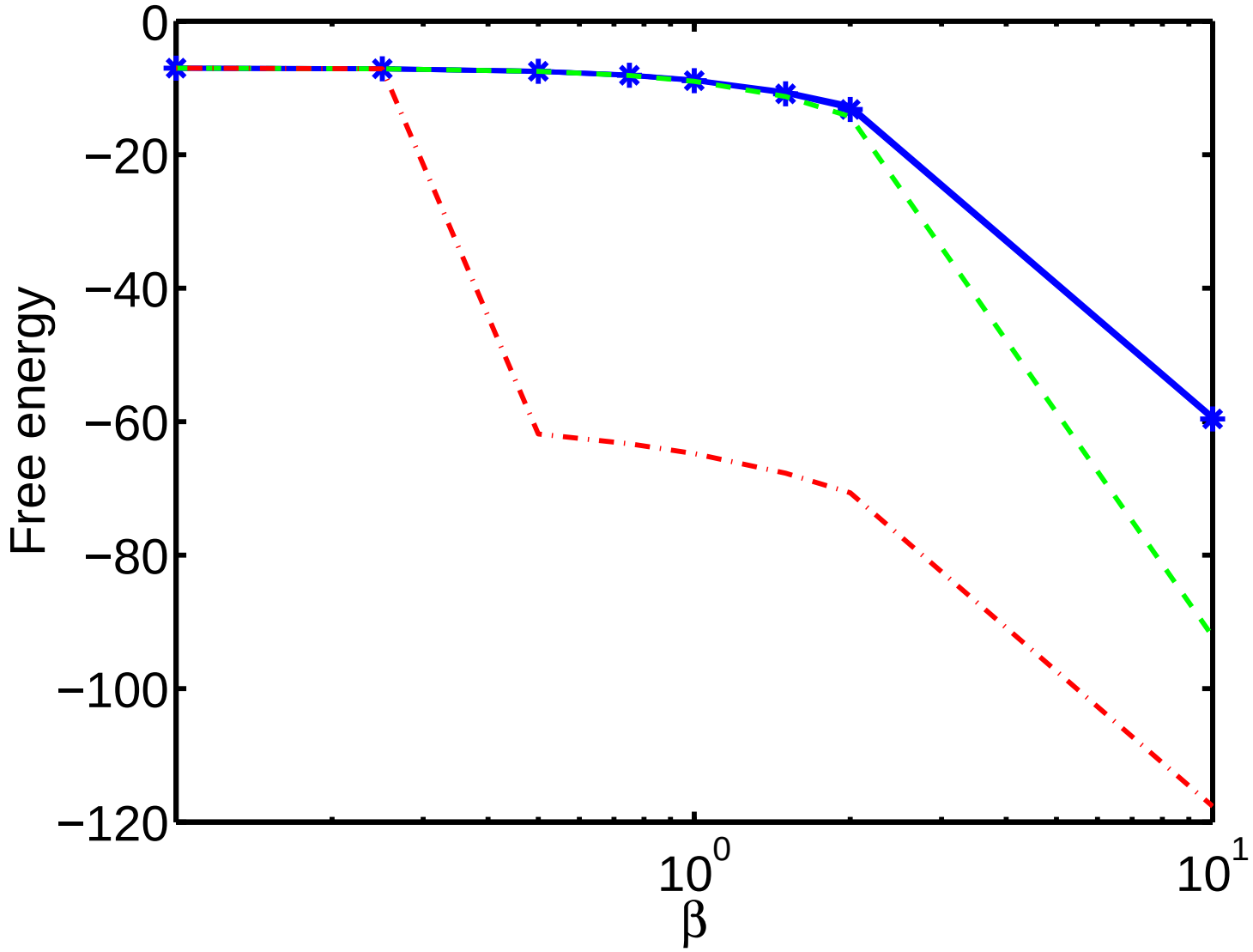
In EC, the non-trivial correlation estimates: $C_{ij} = \left[(\mathbf{\Lambda}_r - \mathbf{J})^{-1}\right]_{ij}$ is used for the two-variables marginals.

Maximal absolute deviation (MAD) for one-variable marginals. Blue upper full line: EC factorized, blue lower full line EC tree, green dashed line: Bethe and red dash-dotted line: Kikuchi.

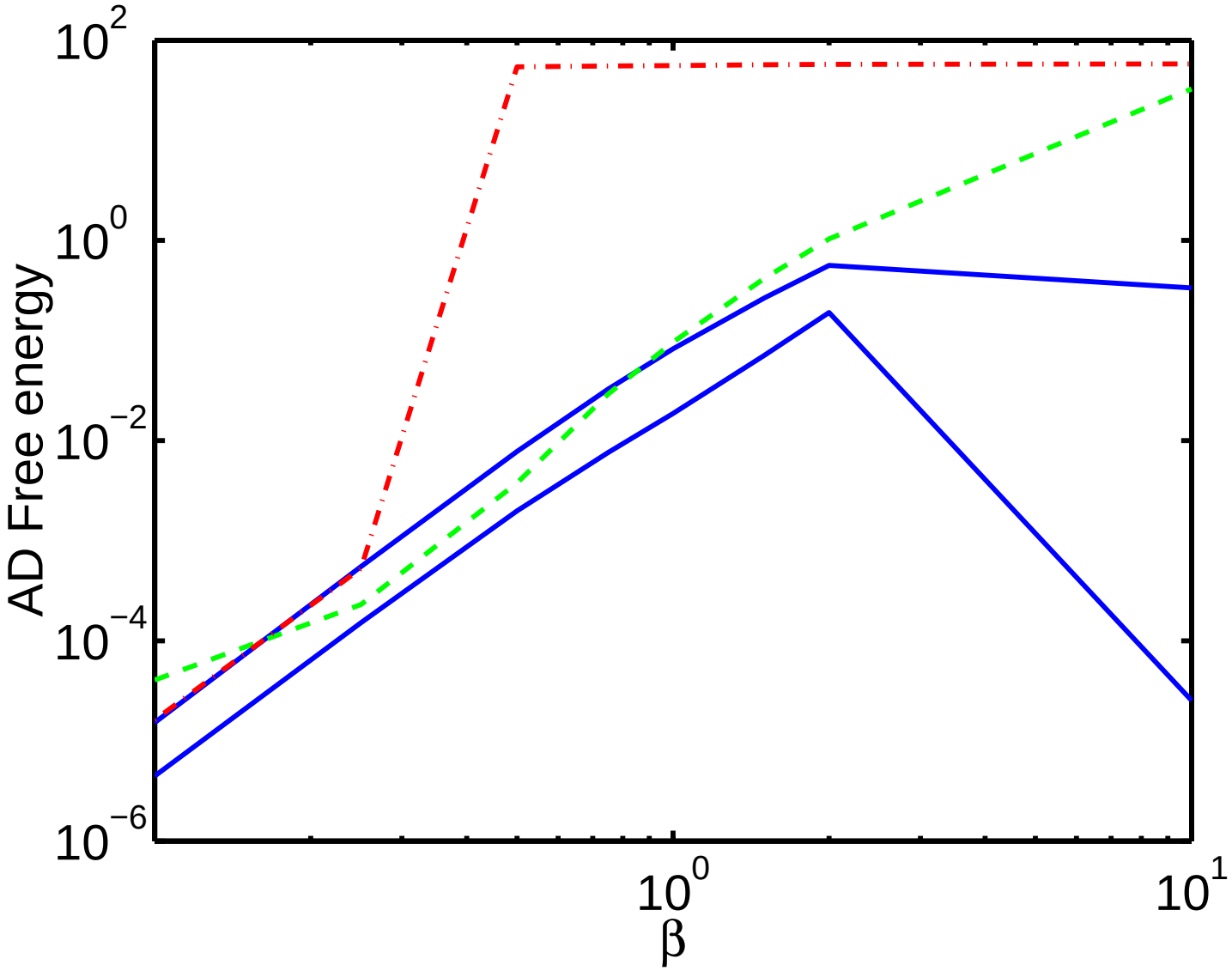




ec016



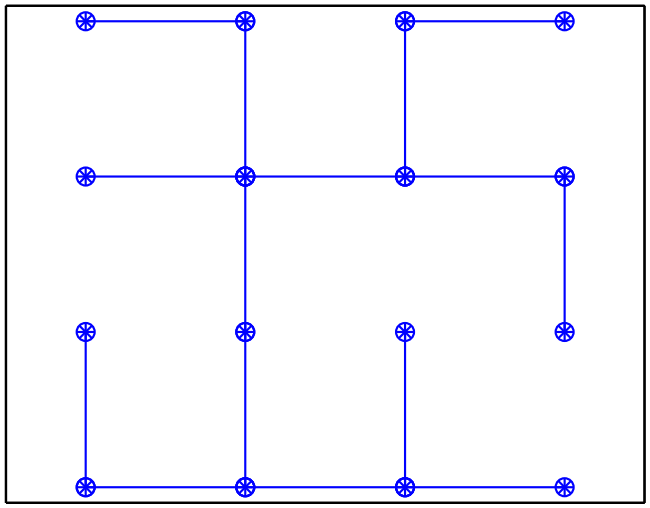
ec016



ec016

Scenario II: Wainwright and Jordan

$N = 16$
Fully connected or
4-by-4 nearest neighbor grid.



Coupling strength:

- repulsive (anti-ferromagnetic) $J_{ij} \sim \mathcal{U}[-2d_{\text{coup}}, 0]$,
- mixed $J_{ij} \sim \mathcal{U}[-d_{\text{coup}}, +d_{\text{coup}}]$ and
- attractive (ferromagnetic) $J_{ij} \sim \mathcal{U}[0, +2d_{\text{coup}}]$ with $d_{\text{coup}} > 0$.

θ_i from *uniform* distribution: $\theta_i \sim \mathcal{U}[-d_{\text{obs}}, d_{\text{obs}}]$ with $d_{\text{obs}} = 0.25$.

Problem type			Method					
			SP	LD	EC factorized		EC tree	
Graph	Coupling	d_{coup}	Mean	Mean	Mean $\pm$ std	Max	Mean $\pm$ std	Max
Full	Repulsive	0.25	0.037	0.020	0.003 $\pm$ 0.002	0.00	0.0017 $\pm$ 0.0011	0.007
	Repulsive	0.50	0.071	0.018	0.031 $\pm$ 0.045	0.20	0.0143 $\pm$ 0.0141	0.102
	Mixed	0.25	0.004	0.020	0.002 $\pm$ 0.002	0.00	0.0013 $\pm$ 0.0008	0.005
	Mixed	0.50	0.055	0.021	0.022 $\pm$ 0.030	0.17	0.0151 $\pm$ 0.0204	0.163
	Attractive	0.06	0.024	0.027	0.004 $\pm$ 0.002	0.01	0.0025 $\pm$ 0.0014	0.007
	Attractive	0.12	0.435	0.033	0.117 $\pm$ 0.090	0.30	0.0211 $\pm$ 0.0307	0.159
Grid	Repulsive	1.0	0.294	0.047	0.153 $\pm$ 0.123	0.58	0.0031 $\pm$ 0.0021	0.013
	Repulsive	2.0	0.342	0.041	0.198 $\pm$ 0.135	0.49	0.0021 $\pm$ 0.0010	0.009
	Mixed	1.0	0.014	0.016	0.011 $\pm$ 0.010	0.08	0.0018 $\pm$ 0.0011	0.006
	Mixed	2.0	0.095	0.038	0.082 $\pm$ 0.081	0.32	0.0068 $\pm$ 0.0053	0.028
	Attractive	1.0	0.440	0.047	0.125 $\pm$ 0.104	0.36	0.0028 $\pm$ 0.0018	0.013
	Attractive	2.0	0.520	0.042	0.177 $\pm$ 0.125	0.41	0.0024 $\pm$ 0.0022	0.016

Error measure (averaged over 100 trials)

$$\text{MeanAD} = \sum_i |p(\mathbf{x}_i = 1) - p(\mathbf{x}_i = 1|\text{Method})|/N \ .$$

SP = Sum Product = Bethe

LD = log determinant relaxation.

There is no universal best approximation.

The need for more than a framework!

Understanding the 'physics' of the problem is neecssary:

- Sparse: use Bethe approximation and extensions (loopy belief propagation).
- Dense: Use central limit theorem (or cavity) arguments and extensions (replica symmetry breaking).
- In between: structured extensions of the above.

Conclusion

EC provides a framework for approximate inference.

Relation to other approaches: variational (Bayes), adaptive TAP, expectation propagation, Bethe+, EP, log-determinant relaxation.

Outlook

EC for RSB.

More efficient algorithms needed – variational bounding can be very slow when the problem is hard. E.g. EP is fast, but doesn't converge on hard cases.

Perhaps relax the requirement for complete consistency of complementary approximations.

References

[1] M. Opper and O. Winther. Gaussian processes for classification: Mean field algorithms. *Neural Computation*, 12:2655–2684, 2000.

[2] M. Opper and O. Winther. Adaptive and self-averaging Thouless-Anderson-Palmer mean field theory for probabilistic modeling. *Phys. Rev. E*, 64:056131, 2001

[3] M. Opper and O. Winther. Tractable approximations for probabilistic models: The adaptive Thouless-Anderson-Palmer mean field approach. *Phys. Rev. Lett.*, 86:3695, 2001

[4] T. P. Minka. Expectation propagation for approximate Bayesian inference. In *UAI 2001*, pages 362–369, 2001

[5] T. Minka and Y. Qi. Tree-structured approximations by expectation propagation. In S. Thrun, L. Saul, and B. Schölkopf, editors, *NIPS 16*. MIT Press, Cambridge, MA, 2004.

[6] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.

[7] A. L. Yuille and A. Rangarajan. The concave-convex procedure. *Neural Comput.*, 15(4): 915–936, 2003.

[8] T. Heskes and O. Zoeter, Expectation propagation for approximate inference in dynamic Bayesian networks. In A. Darwiche and N. Friedman, editors, Proceedings UAI-2002" , 216–233, 2002.

[9] T. Heskes, K. Albers, and H. Kappen. Approximate inference and constrained optimization. In *UAI-03*, pages 313–320, San Francisco, CA, 2003. Morgan Kaufmann Publishers.

[10] M. J. Wainwright and M. I. Jordan, "Semidefinite methods for approximate inference on graphs with cycles," Tech. Rep. UCB/CSD-03-1226, UC Berkeley CS Division, 2003.

[11] M. Wainwright and M. I. Jordan, "Semidefinite relaxations for approximate inference on graphs with cycles," in *NIPS 16*. MIT Press, Cambridge, MA, 2004.

[12] J. S. Yedidia, W. T. Freeman, and Y. Weiss. Generalized belief propagation. In T. K. Leen, T. G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems 13*, pages 689–695, 2001.

ecbib