Expectation Propagation

Manfred Opper

Lecture 1

- Inference and variational approximations (mean field & Gaussian)
- The 'other KL' and assumed density filtering
- Expectation propagation as an algorithm
- TAP equations
- Free energy from TAP

Lecture 2

- EP free energy
- Correcting EP: Cluster expansion
- Correcting EP: Cumulant expansion
- Applications

Probabilistic Inference: the problem

For a joint distribution $p(\mathbf{x}, \mathbf{y})$ of hidden variables $\mathbf{x}$ (or parameter θ in a Bayesian setting) and observed data $\mathbf{y}$ the posterior is given by

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

- The computation of the marginal probability of the data $p(\mathbf{y}) = \int d\mathbf{x}\, p(\mathbf{x}, \mathbf{y})$ (evidence) requires high dimensional sums or integrals and is often intractable.

- For the same reasons we often can't compute marginals $p_i(x_i|\mathbf{y})$, or expectations using these densities.

The Variational Approximation

- Approximate $p(\mathbf{x}|\mathbf{y})$ by $q(\mathbf{x}) \in \mathcal{F}$ where $\mathcal{F}$ tractable family of distributions such that the Kullback-Leibler divergence

$$KL(q,p) = \int d\mathbf{x} q(\mathbf{x}) \ \ln \frac{q(\mathbf{x})}{p(\mathbf{x}|\mathbf{y})} \geq 0$$

  is minimized.

- From $p(\mathbf{x}|\mathbf{y}) = \frac{p(\mathbf{x},\mathbf{y})}{p(\mathbf{y})}$, we get an **upper bound** for any q

$$-\ln p(\mathbf{y}) \leq F(q) \doteq \int d\mathbf{x} \, q(\mathbf{x}) \ln q(\mathbf{x}) - E_q[\ln p(\mathbf{x},\mathbf{y})]$$

  with the **variational free energy** $F(q)$

Example 1: Mean field approximation

- Factorizing probability distribution

$$q(\mathbf{x}) = \prod_{i=1}^{M} q_i(x_i)$$

- Optimal selfconsistent solution: $q_i^*(x) = \frac{1}{Z_i} \exp\left\{E_{\setminus i}[\ln p(\mathbf{x}, \mathbf{y})]\right\}$ with $E_{\setminus i}[\ldots]$ the average over all variables except x_i.

- Applicable to discrete and continuous random variables.

- Neglects dependencies but linear response corrections possible.

Example 2: Gaussian approximation

- Gaussian densities $q(\mathbf{x}) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

- Variational free energy $F(q) = -\frac{N}{2}\log 2\pi - \frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{N}{2} - E_q[\log p(\mathbf{x}, \mathbf{y}|\boldsymbol{\theta})]$

- Selfconsistency equations

$$
\begin{aligned}
0 &= E_q\left[\frac{\partial \log p(\mathbf{x}, \mathbf{y}|\boldsymbol{\theta})}{\partial x_i}\right] \\
(\boldsymbol{\Sigma}^{-1})_{ij} &= -E_q\left[\frac{\partial^2 \log p(\mathbf{x}, \mathbf{y}|\boldsymbol{\theta})}{\partial x_i \partial x_j}\right]
\end{aligned}
$$

- Applicable to continuous variables only (no constraints allowed).

Other popular (in machine learning) approximations

- Loopy belief propagation and its extensions:

  Exact on trees, numerically nontrivial when applied to continuous random variables.

- Expectation Propagation:

  Applicable to discrete and constrained continuous random variables, allows for dependencies.

Motivation: Minimising the other KL

- The reverse KL divergence is

$$KL(p,q) = \int d\mathbf{x}\, p(\mathbf{x}|\mathbf{y}) \ln \frac{p(\mathbf{x}|\mathbf{y})}{q(\mathbf{x})} = \text{const} - \int d\mathbf{x}\, p(\mathbf{x}|\mathbf{y}) \ln q(\mathbf{x})$$

- If $q(\mathbf{x}) = \prod_i q_i(x_i)$, we have to minimize

$$-\sum_i \int dx\, p_i(x|\mathbf{y}) \ln q_i(x)$$

which is minimized by the true marginal $q_i = p_i$.

- On the other hand for exponential families

$$q_\theta(\mathbf{x}) \propto b(\mathbf{x})\ \exp[\boldsymbol{\theta}^\top \boldsymbol{\phi}(\mathbf{x}) + g(\boldsymbol{\theta})]\ .$$

the optimal θ must be chosen such that the general moments match $E_q[\boldsymbol{\phi}(\mathbf{x})] = E_p[\boldsymbol{\phi}(\mathbf{x})]$. **In general: Intractable !**

Examples of exponential families

- Multivariate Gaussian densities $\phi(\mathbf{x}) = (\mathbf{x}, -\frac{1}{2}\mathbf{x}\mathbf{x}^\top)$

  and $\theta = (\gamma, \lambda)$ yields $q_\theta(\mathbf{x}) \propto \exp[-\frac{1}{2}\mathbf{x}^\top \lambda \mathbf{x} + \gamma^\top \mathbf{x}]$.

- Multinomial model: Let $x \in \{1, \ldots, K\}$. Set $\phi(x) = (\phi_1(x), \ldots, \phi_K(x))$ with $\phi_j(x) = 1$ if $x = j$ and $= 0$ else. Hence with $\theta = (\theta(1), \ldots, \theta(K))$ we have $e^{\theta^\top \phi(x)} = e^{\theta(x)}$

Assumed Density Filtering

- Assume data arrive sequentially: $D_{t+1} = y_1, y_2, \ldots, y_{t+1}$

- Exact update of posterior

$$p(\mathbf{x}|D_{t+1}) = \frac{p(y_{t+1}|\mathbf{x})p(\mathbf{x}|D_t)}{\int d\mathbf{x} p(y_{t+1}|\mathbf{x})p(\mathbf{x}|D_t)}.$$

- Replace $p(\mathbf{x}|D_{t+1})$ by parametric approximation $q_\theta(\mathbf{x})$ using the following steps:

  - Update:

$$q_\theta(t)(\mathbf{x}|y_{t+1}) = \frac{p(y_{t+1}|\mathbf{x})q_{\theta(t)}(\mathbf{x})}{\int d\mathbf{x} p(y_{t+1}|\mathbf{x})q_{\theta(t)}(\mathbf{x})}.$$

  - Project: Minimize

$$KL\left(q_\theta(t)(\cdot|y_{t+1})||q_\theta(\cdot)\right)$$

Example: Bayesian classifier

- Classification model: $y = \text{sign}[h_{\mathbf{w}}(s)] = \pm 1$ where $h_{\mathbf{w}}(s) = \sum_j w_j \psi_j(s)$.

- Probit likelihood:

$$p(y|\mathbf{w}, s) = \frac{1}{2} + \int_0^{y h_{\mathbf{w}}(s)} g(t) dt$$

  with $g(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2}$.

- Gaussian prior distribution over weights $p_0(\mathbf{w}) \propto e^{-\frac{1}{2}\sum_j w_j^2}$

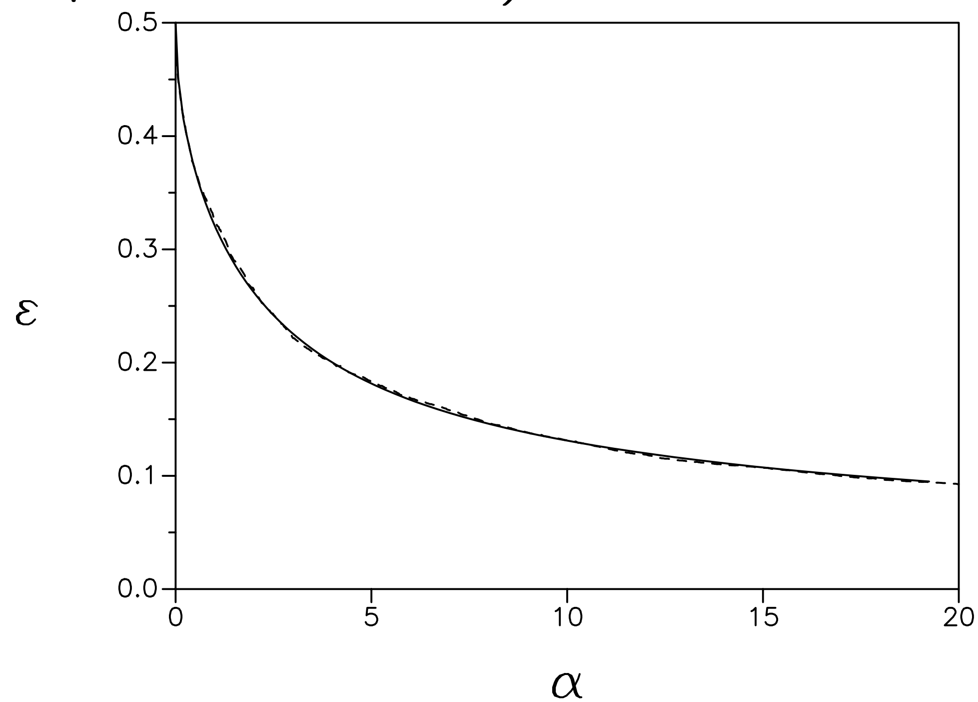
- Posterior distribution $p(\mathbf{w}|D_n) = \frac{1}{Z} p_0(\mathbf{w}) \prod_{i=1}^n p(y_i|\mathbf{w}, s_i)$

- Parametric approximation $q_\theta(\mathbf{x}) \sim \mathcal{N}(\widehat{\mathbf{w}}, \mathbf{C})$

- Moments of $q_{\theta(t)}(\mathbf{w}|y_{t+1}) \propto p(y_{t+1}|\mathbf{w} s_{t+1}) q_{\theta(t)}(\mathbf{w})$ easily computable: $p(y_{t+1}|\mathbf{w}, s_{t+1})$ depends only on single Gaussian $\sum_j w_j \psi_j(s_{t+1})$!

Toy application:

Learning curve for toy $d = 50$ model (probit likelihood, spherical Gaussian inputs, realizable random target, $\alpha \doteq \frac{\#\text{data}}{d}$). Dashed line: Bayes optimal (batch – replica calculation).



For finite t: Result depends on order of presentation of data terms.

Gaussian latent variable models

- Set $x_i \doteq h_{\mathbf{w}}(s_i)$

- write the posterior as

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

Assumed Density Filtering

- Assume target density written as a product of terms

$$p(\mathbf{x}) = \frac{1}{Z} f_0(\mathbf{x}) \prod_{i=1}^{N} f_i(\mathbf{x})$$

- Update: $\widehat{q}(\mathbf{x}) \propto f_{n+1}(\mathbf{x}) q(\mathbf{x})$

- Project: Minimize $KL(\widehat{q}|q)$ wrt $q \in$ exponential family $\rightarrow q^{\text{new}}(\mathbf{x})$

- For exponential families $q(\mathbf{x}) \propto \exp[\lambda^\top \phi(\mathbf{x})]$
$\rightarrow$ matching of moments $\langle \phi(\mathbf{x}) \rangle_q = \langle \phi(\mathbf{x}) \rangle_{\widehat{q}}$.

Expectation - Propagation (Tom Minka)

$$p(\mathbf{x}) = \frac{1}{Z} f_0(\mathbf{x}) \prod_{i=1}^{N} f_i(\mathbf{x})$$

with $f_0 \in$ exponential family. Initialize $g_i(\mathbf{x})_i = 1$ and **repeat until convergence**

- Choose $i > 0$, <u>remove</u> terms g_i i.e. construct $q_{\setminus i}(\mathbf{x}) \propto q(\mathbf{x})/g_i(\mathbf{x})$
- <u>Update:</u> $q_i(\mathbf{x}) = f_i(\mathbf{x}) q_{\setminus i}(\mathbf{x})$
- <u>Project:</u> Minimize $KL(q_i || q)$ for $q \in$ exponential family $\rightarrow q^{\text{new}}(\mathbf{x})$
- <u>Refine terms:</u> $g_i^{\text{new}}(\mathbf{x}) \propto \frac{q^{\text{new}}(\mathbf{x})}{q_{\setminus i}(\mathbf{x})} \propto \frac{q^{\text{new}}(\mathbf{x}) g_i(\mathbf{x})}{q(\mathbf{x})}$

At convergence

- Approximation by $q(\mathbf{x}) \propto f_0(\mathbf{x}) \prod_i g_i(\mathbf{x})$ with **tractable** g_i's.

- q and the **tilted distributions**

$$q_i(\mathbf{x}) \propto f_i(\mathbf{x}) q_{\setminus i}(\mathbf{x}) = q(\mathbf{x}) \frac{f_i(\mathbf{x})}{g_i(\mathbf{x})}$$

  have a set of equal moments

$$\langle \phi(\mathbf{x}) \rangle_q = \langle \phi(\mathbf{x}) \rangle_{q_i}$$

  for $i = 1, \ldots, n$.

EP Comments

- Fast Algorithm (if convergent), applicable to discrete and continuous variables.

- Excellent results for Gaussian latent variable models

- Depends on factorization and exponential family chosen for the g_i.

- Match Ising variables and multivariate Gaussians ($\mathrm{KL} = \infty$) ?

Examples: Discrete variables on graph

- Discrete variables $x_i \in \{1, \dots, K\}$

$$p(\mathbf{x}) \propto \prod_k e^{\theta_k(x_k)} \prod_{(ij)} e^{\theta_{ij}(x_i, x_j)}$$

- Tractable approximation (factorizing):

$$q(\mathbf{x}) \propto \prod_k e^{\theta_k(x_k)} \prod_{(ij)} e^{\lambda_{i\to j}(x_j) + \lambda_{j \to i}(x_i)}$$

- Tilted distribution (edge (uv) removed).

$$q_{uv}(\mathbf{x}) \propto q(\mathbf{x}) e^{\theta_{uv}(x_u, x_v) - \lambda_{u \to v}(x_v) - \lambda_{v \to u}(x_u)}$$

- Moment matching

$$
\begin{aligned}
q^{\mathsf{new}}(x_u) &= \sum_{\mathbf{x} \backslash x_u} q_{uv}(\mathbf{x}) \\
q^{\mathsf{new}}(x_v) &= \sum_{\mathbf{x} \backslash x_v} q_{uv}(\mathbf{x})
\end{aligned}
$$

Gaussian latent variable model

- The model

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

- Exponential family terms $g_i(x) = e^{\gamma_i x_i - \frac{1}{2}\Lambda_i x_i^2}$

- Approximation

$$q(\mathbf{x}) \propto \exp[-\frac{1}{2}\mathbf{x}^\top \mathbf{K}\mathbf{x} - \frac{1}{2}\sum_{i=1}^{N} \Lambda_i x_i^2 + \gamma^\top \mathbf{x}]$$

.

- EP updates

  Iterate until convergence:

  1. Choose a site i
  2. Remove γ_i, Λ_i, Integrate out all variables in q except $x_i \to$ marginal $q_i(x_i)$,
  3. Compute $\langle x_i \rangle$ and $\langle x_i^2 \rangle$ from $q_i(x_i)$
  4. Moment matching: recompute marginal $q(x_i)$
  5. Recompute γ_i and Λ_i

TAP equations

- Sherrington–Kirkpatrick model for N Ising spins $S_i = \pm 1$ with random couplings $J_{ij} \sim \mathcal{N}(0, 1/N)$

$$P(\mathbf{S}) \propto \exp\left[\sum_{i<j} S_i J_{ij} S_j + \sum_i S_i \theta_i\right]$$

- Mean field equations (TAP equations, after *Thouless, Anderson & Palmer*)

$$\langle S_i \rangle \approx \tanh\left(\sum_j J_{ij}\langle S_j\rangle - \langle S_i\rangle \sum_j J_{ij}^2 (1 - \langle S_j\rangle^2) + \theta_i\right)$$

Perturbative (Plefka) approach

- Gibbs free energy.

$$G(\mathbf{m}) = \min_q \{KL(q||p) \mid \langle \mathbf{S} \rangle_q = \mathbf{m}\} - \ln Z$$

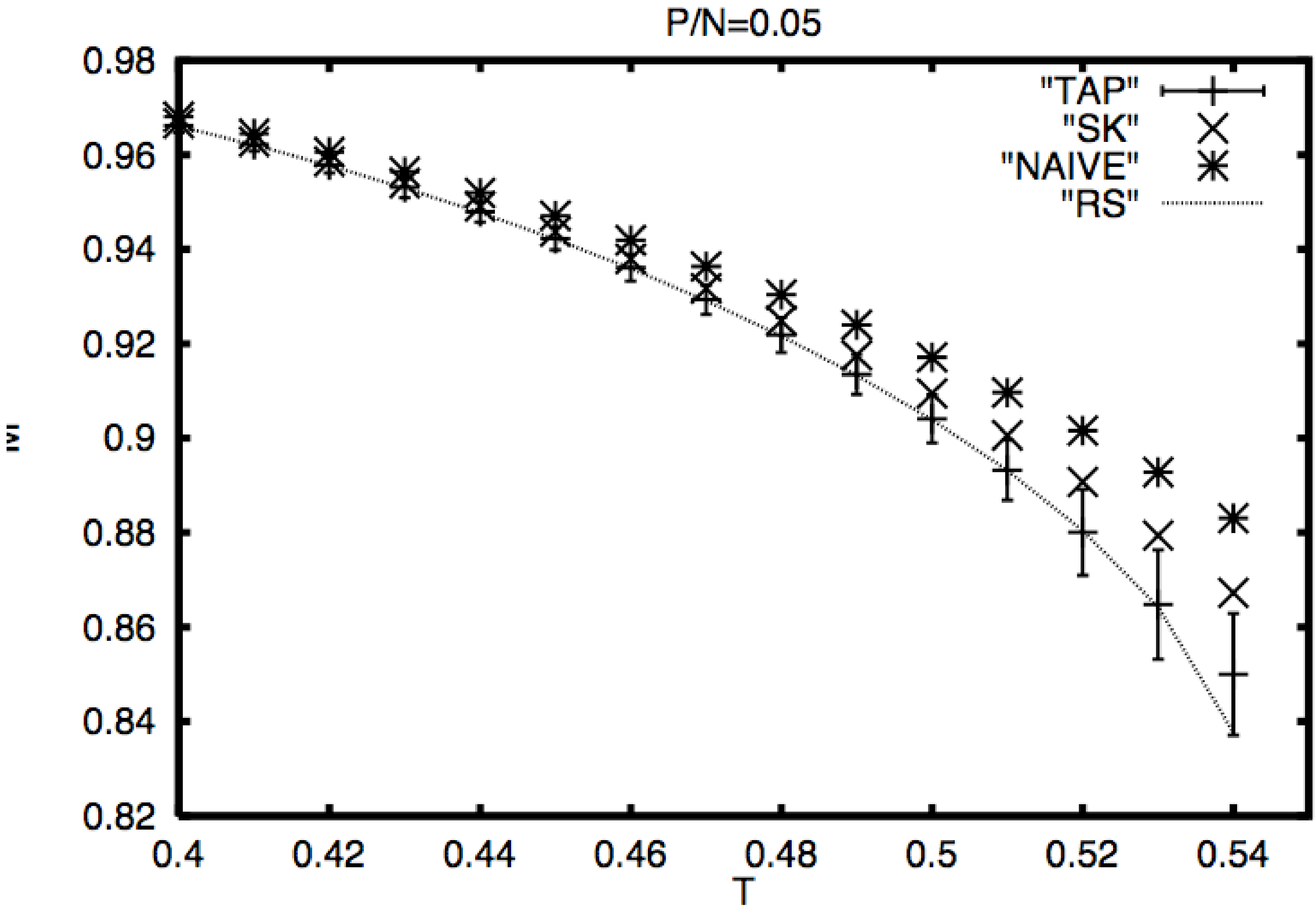
- Define one parameter family of models

$$P_t(\mathbf{S}) \propto \exp\left[t \sum_{i<j} S_i J_{ij} S_j + \sum_i S_i \theta_i\right]$$

  Perturbative approach (Plefka): Expand $G_t(\mathbf{m})$ to $\mathcal{O}(t^2)$ yields TAP equations.

- Information geometric interpretation and related derivations (Tanaka, Bhattacharyya & Keerthi, Amari & Ikeda & Shimokawa, Kappen & Wiegerinck):)

- Unfortunately Not exact for other random matrix ensembles! Proper correction to naive MF depends on statistics of $\mathbf{J}$!



(Hopfield Model: Kabashima & Saad):
$J_{ij} = \sum_{\mu=1}^{\alpha N} \xi_i^\mu \xi_j^\mu$ with i.i.d. ξ_i^μ of variance $\frac{\beta}{N}$

Consider slightly more general class of models

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

allows for latent Gaussian models but also discrete variables (spins) by taking

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

.

Cavity approach

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

Hence

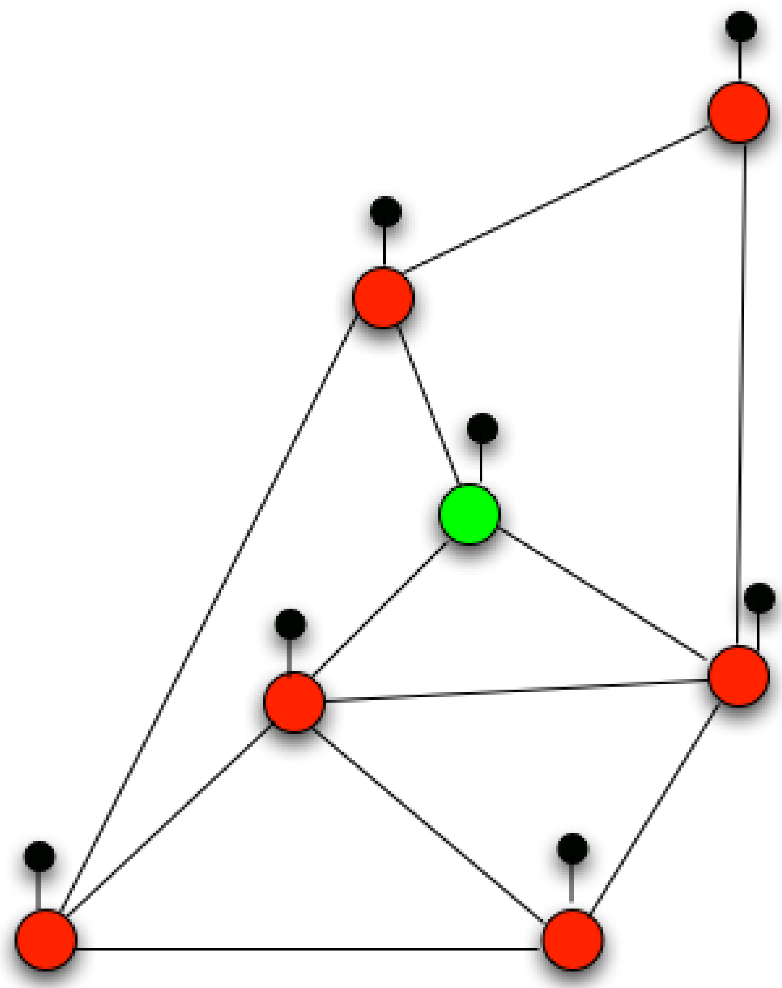
$$p_i(x_i, \mathbf{x}_{\mathcal{N}(i)}) \propto f_i(x_i)\,\mathsf{exp}[x_i h_i(\mathbf{x}_{\mathcal{N}(i)})]\; p_{\backslash i}(\mathbf{x}_{\mathcal{N}(i)})$$

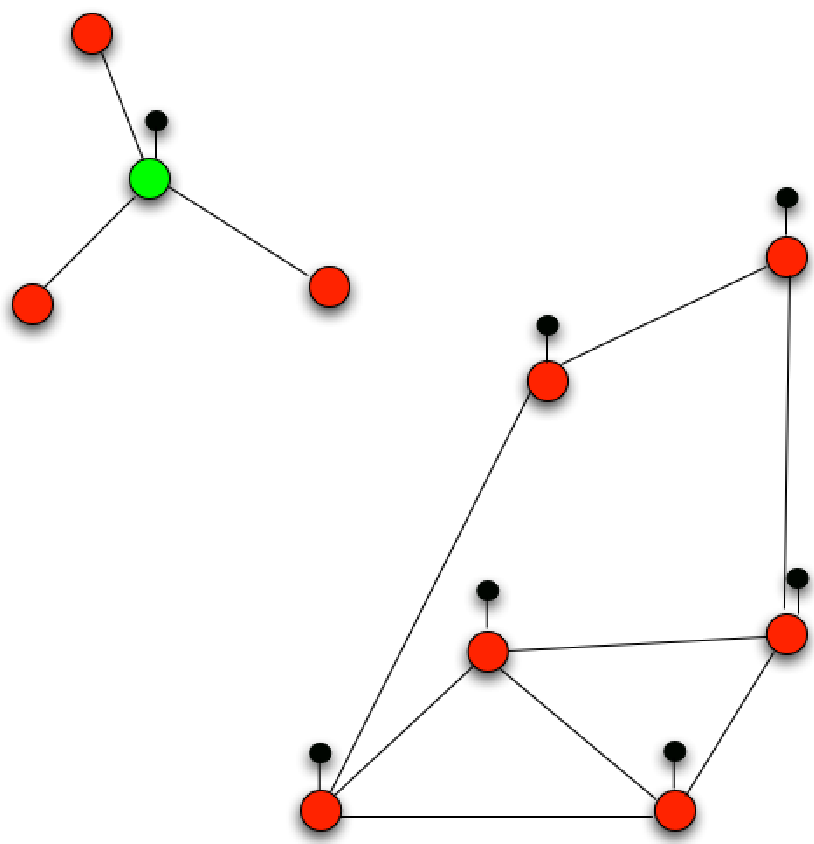
We can write

$$p_i(x, h) \propto f_i(x)\; e^{xh}\; p_{\backslash i}(h)$$

when we introduce the 'cavity field' distribution

$$p_{\backslash i}(h) = \sum_{\mathbf{x}_{\mathcal{N}(i)}} \delta\left(h - \sum_{j \in \mathcal{N}(i)} J_{ij} x_j\right)\; p_{\backslash i}(\mathbf{x}_{\mathcal{N}(i)})$$





Weak dependencies:

- Approximate $p_{\backslash i}(h)$ by Gaussian (central limit theorem) $p_{\backslash i}(h) \propto \exp[-\frac{(h-a_i)^2}{2V_i}]$.

$$p_i(x) \approx \int p_i(x,h)dh \propto f_i(x) \int e^{xh}\, p_{\backslash i}(h)dh$$
$$= \frac{1}{Z_i}\, f_i(x)\ \exp\left[a_i x + \frac{V_i}{2}x^2\right]$$

Derive set of nonlinear equations for $2N$ unknowns γ_i, V_i!

- We use

$$\langle h_i \rangle = \int dx \int p_i(x,h)\ h\ dh = \frac{1}{Z_i}\int dx\ f_i(x)\ \int dh\ h\ e^{xh}\ p_{\backslash i}(h) \approx a_i + V_i \langle x_i \rangle$$

TAP Equations

- Hence, using $\langle h_i \rangle = \sum_j J_{ij} \langle x_j \rangle$ we get

$$a_i = \sum_j J_{ij} \langle x_j \rangle - V_i \langle x_i \rangle$$

- Naive computation

$$\begin{aligned} V_i &= \sum_{jk} J_{ij} J_{ik} \left(\langle x_j x_k \rangle_{\backslash i} - \langle x_j \rangle_{\backslash i} \rangle \langle x_k \rangle_{\backslash i} \right) \\ &\approx \sum_j J_{ij}^2 \left(\langle x_j^2 \rangle - \langle x_j \rangle^2 \right) \end{aligned}$$

leads us back to the SK Onsager term

(adaptive) TAP equations:

- Replace surrounding nodes by auxiliary model with $f_i(x) \to g_i(x) = e^{-\frac{1}{2}\Lambda_i x^2 + \gamma_i x}$ with γ_i, Λ_i chosen s.t. moments $\langle x_i \rangle$ and $\langle x_i^2 \rangle$. Assume we get the same cavity fields (generalizes an idea of Parisi & Potters).

- Let

$$Z_i = \int dx\, f_i(x)\ \exp\left[a_i x + \frac{V_i}{2} x^2\right] \qquad \tilde{Z}_i = \int dx\, g_i(x)\ \exp\left[a_i x + \frac{V_i}{2} x^2\right]$$

- Hence, we have

$$\begin{aligned} \langle x_i \rangle &= \frac{d}{da_i} \ln Z_i = \frac{d}{da_i} \ln \tilde{Z}_i = \frac{\gamma_i + a_i}{\Lambda_i - V_i} \\ \langle x_i^2 \rangle - \langle x_i \rangle^2 &= \frac{d^2}{da_i^2} \ln Z_i = \frac{d^2}{da_i^2} \ln \tilde{Z}_i = \frac{1}{\Lambda_i - V_i} \end{aligned}$$

On the other hand, by direct computation

$$\begin{aligned} \langle x_i \rangle &= ((\mathbf{\Lambda} - \mathbf{J})^{-1} \boldsymbol{\gamma})_i \\ \langle x_i^2 \rangle - \langle x_i \rangle^2 &= \left[(\mathbf{\Lambda} - \mathbf{J})^{-1}\right]_{ii} \end{aligned}$$

- Eliminiating $a_i, \Lambda_i, \gamma_i, V_i$ (numerically) we get closed set of equations for moments $\langle x_i \rangle$, $\langle x_i^2 \rangle$ for $i = 1, \ldots, N$.

- This corresponds to the fixed points of EP, when applied to latent Gaussian model family and projections to multivariate Gaussians of the form

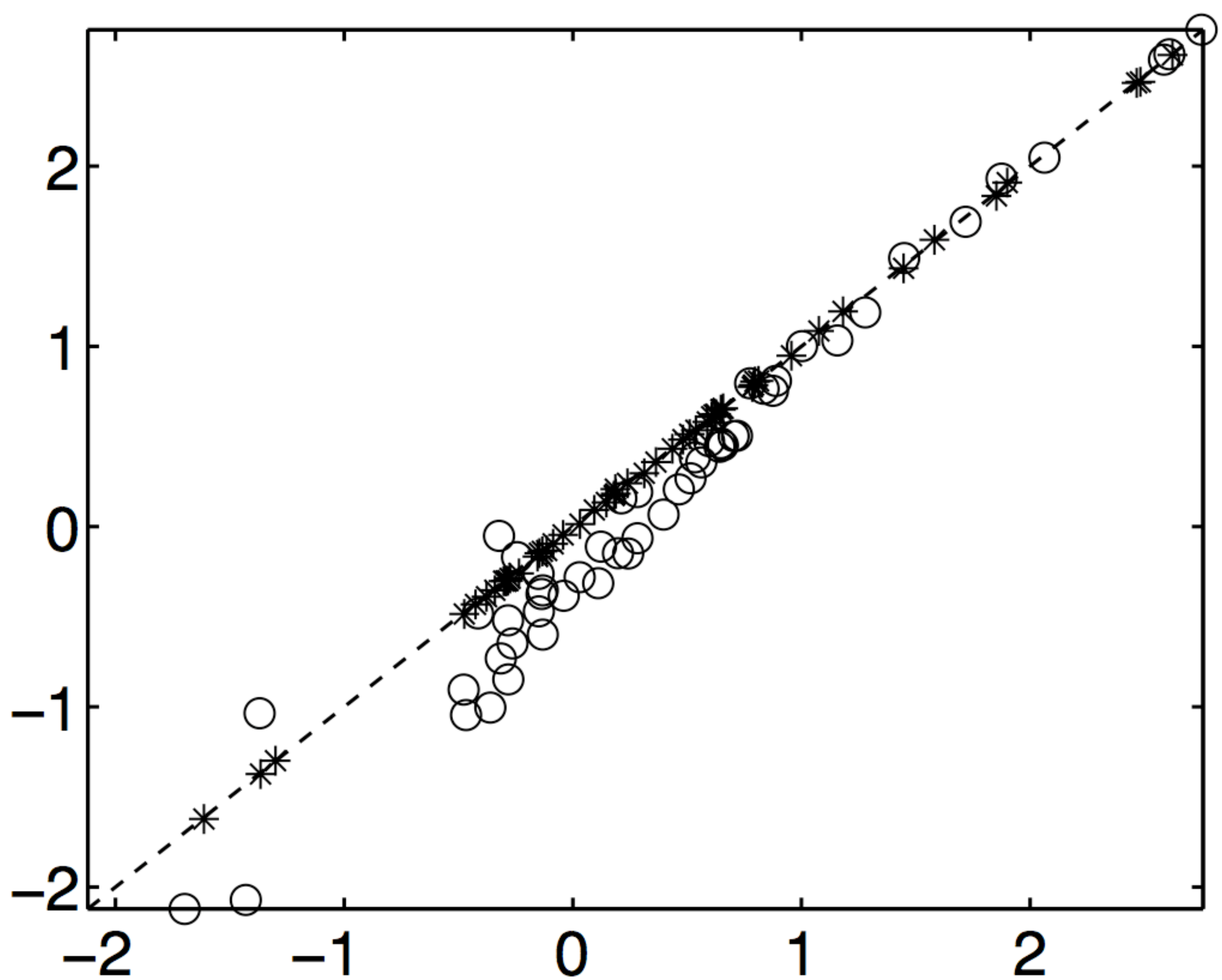
$$q(\mathbf{x}) \propto \exp[\frac{1}{2}\mathbf{x}^\top \mathbf{J}\mathbf{x} - \frac{1}{2}\sum_{i=1}^{N} \Lambda_i x_i^2 + \gamma^\top \mathbf{x}]$$

- Moment matching makes sense, even when KL projection $= \infty$!

Consistency of cavity field: Bayes classifier

Remove variable x_i from system and compute average cavity field $\langle h_i \rangle_{\setminus i} \doteq \sum_j J_{ij} \langle x_j \rangle_{\setminus i}$

- "exactly": by solving the TAP equations on $N-1$ variable system $\to \langle h_i \rangle_{\setminus i}^{(\mathrm{N}-1)}$.
- Using the generalized TAP approximation for $p_{\setminus i}$.
- **Next page:** $y_i \langle h_i \rangle_{\setminus i}^{(\mathrm{N}-1)}$ as function of $y_i \langle h_i \rangle_{\setminus i}$



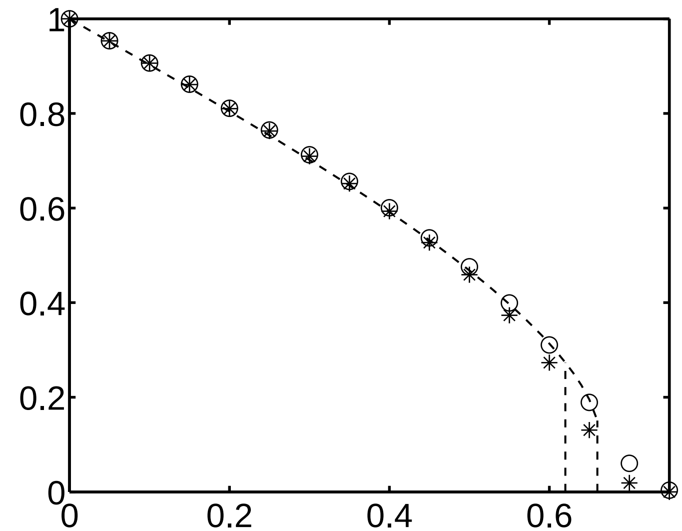
2
1
0
−1
−2
−2
−1
0
1
2

Toy Model: Linear regression with binary variables

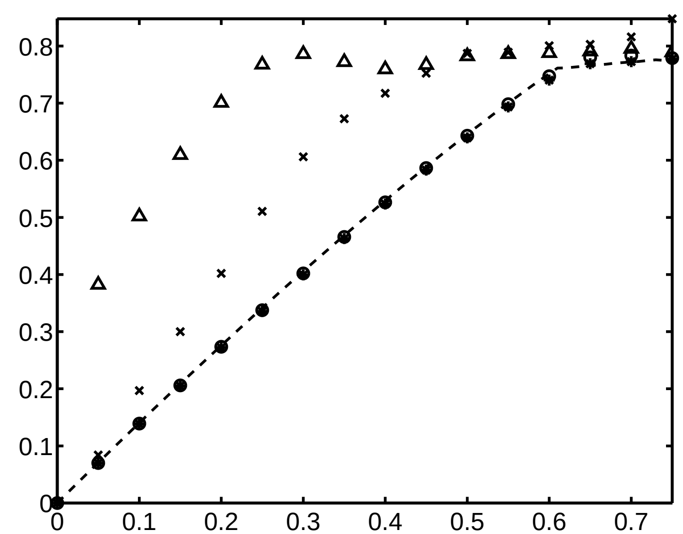
$y(k) = \sum_{i=1}^{N} x_i \frac{s_i(k)}{\sqrt{N}} + \sigma_0 \xi(k)$ with $x_i = \pm 1$. $s_i(k)$ i.i.d. Gaussian of unit variance and $\sigma_0^2 = 0.2$.

Testerror vs $\frac{\#\text{data}}{\#\text{variables}}$

($N = 60$ & asymptotic analytical result)



Minimal Free Energy



Thermodynamic limit

- Assume $f_i(x) = f(x)$, and the statistics of $\mathbf{J}$ defined by generating function
$$\frac{1}{N} \ln \left[e^{\frac{1}{2}\text{Trace}\,(\mathbf{AJ})} \right]_{\mathbf{J}} \simeq \text{Trace}\, G(\mathbf{A}/N)$$

- Assume $V_i = V$ self–averaging.

- Disorder average: $\langle \ln \det(\mathbf{\Lambda} - \mathbf{J}) \rangle_J = \sum_i \ln(\lambda_i - \hat{r}) + N r \hat{r} - 2NG(r)$

- Order parameter equations give
$$\begin{aligned} r &= \frac{1}{N} \sum_i \frac{1}{\lambda_i - \hat{r}} = \frac{1}{N} \sum_i \left(\langle x_i^2 \rangle - \langle x_i \rangle^2 \right) \equiv \chi \to V = \hat{r} \\ \hat{r} &= 2G'(r) \end{aligned}$$

- This yields $V = G'(\chi)$ and agrees with known results of (Parisi & Potters)

Free energy from cavity approach

- Introduce variable interaction strength

$$p_t(\mathbf{x}) = \frac{1}{Z} \exp\left[t \sum_{(ij)} x_i J_{ij} x_j \right] \prod_k f_k(x_k)$$

- Gibbs Free energy

$$G_t(\mathbf{m}, \mathbf{M}) = \min_q \left\{ KL(q||p) \; | \langle S_i \rangle_q = m_i; \langle S_i^2 \rangle_q = M_i, \forall i \right\} - \ln Z_t$$

- Differentiating gives

$$\frac{\partial G_t(\mathbf{m}, \mathbf{M})}{\partial t} = -\frac{1}{2} \sum_{i,j} m_i J_{ij} m_j - \frac{1}{2} \mathrm{Tr}(\mathbf{C}_t \mathbf{J})$$

- Inserting Gaussian approximation $\mathbf{C}_t \approx (\mathbf{\Lambda} - t\mathbf{J})^{-1}$ and integrating, we obtain

$$G \equiv G_1 = G_0 - \frac{1}{2}\sum_{ij} m_i J_{ij} m_j$$

$$-\frac{1}{2}\ln\det(\mathbf{\Lambda} - \mathbf{J}) - \frac{1}{2}\sum_i V_i(M_i - m_i^2) + \frac{1}{2}\sum_i \ln(M_i - m_i^2)$$

- This can be written as

$$G = G^{\text{Gauss}} + G_0 - G_0^{\text{Gauss}}$$