(a) First, write down the log-joint probability for a single observation-source pair $\log(p(\mathbf{s}, \mathbf{x}))$. Rearrange the terms to form a sum of log-factors on $\mathbf{s}$ (assuming $\mathbf{x}$ is observed), each defined either on a single source variable, or on a pair:

$$\log(p(\mathbf{s}, \mathbf{x})) = \sum_i \log f_i(s_i) + \sum_{ij} \log g_{ij}(s_i, s_j).$$

Relate your result to the Boltzmann Machine. [Remember that, since the sources s are binary, $s_i^2 = s_i$.]

(b) Next, derive a message passing scheme to find iterative approximations $\tilde{f}_i$ and $\tilde{g}_{ij}$ to each factor. Start your derivation from the KL divergence $\mathbf{KL}[p\|q]$ and identify clearly each time you make an approximate step. You don't need to make all of the EP approximations: which one(s) is(are) missing?

Give the final message-passing scheme in terms of updates to the natural parameters of the site approximations. There will be two different types of update: for the $\tilde{f}_i$ and the $\tilde{g}_{ij}$ respectively.

(c) Rewrite your message passing approximation to use factored approximate messages. Explain how this leads to a loopy BP algorithm.

(d) Describe a Bayesian method for selecting K, the number of hidden binary variables using EP. Does your method pose any computational difficulties and if so how would you tackle them?

Implement the EP/loopy-BP algorithm that you derived in the previous question, and compare your results to those of the variational mean-field algorithm.

$$* \log p(x|s) = \log (2\pi\sigma^2)^{-D/2} \exp\left\{ \left(\underset{\sim}{x} - \sum_{k=1}^{K} s_k \underset{\sim}{\mu_k}\right)^T \left(\underset{\sim}{x} - \sum_{k=1}^{K} s_k \underset{\sim}{\mu_k}\right) / 2\sigma^2 \right\}$$

$$= -\frac{D}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}\left\{ \underset{\sim}{x}^T\underset{\sim}{x} + \sum_{m,k=1}^{K} \underset{\sim}{\mu_k}^T\underset{\sim}{\mu_m} \cdot s_k \cdot s_m - 2\sum_{k=1}^{K} \underset{\sim}{\mu_k}^T \underset{\sim}{x}\, s_k \right\}.$$

$$= -\frac{D}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}\left\{ \underset{\sim}{x}^T\underset{\sim}{x} + \sum_{m\neq k}^{K}\sum_{k=1}^{K} s_k s_m \underset{\sim}{\mu_k}^T\underset{\sim}{\mu_m} + \sum_{k=1}^{K} s_k \underset{\sim}{\mu_k}^T\underset{\sim}{\mu_k} - 2\sum_{k=1}^{K} \underset{\sim}{\mu_k}^T \underset{\sim}{x}^{(n)} s_k \right\}$$

$$* \log p(s|\pi) = \log \prod_{k=1}^{K} \pi_k^{s_k} (1-\pi_k)^{(1-s_k)}$$

$$= \sum_{k=1}^{K} s_k \log \pi_k + (1-s_k)\log(1-\pi_k)$$

$$\log p(s, x) = \log p(x|s) + \log p(s|\pi)$$

$$= -\frac{1}{2\sigma^2}\left\{ \sum_{m\neq k}^{K}\sum_{k=1}^{K} \underset{\sim}{\mu_k}^T\underset{\sim}{\mu_m} s_k s_m \right\} + \sum_{k=1}^{K}\left\{ -\frac{1}{2\sigma^2}\left(\underset{\sim}{\mu_k}^T\underset{\sim}{\mu_k} - 2\underset{\sim}{\mu_k}^T\underset{\sim}{x}\right) + \log\pi_k - \log(1-\pi_k) \right\} s_k.$$

$$+ \text{constant}.$$

$$= \left\{ \sum_{m\neq k}^{K}\sum_{k=1}^{K} \left[-\frac{1}{2\sigma^2}\underset{\sim}{\mu_k}^T\underset{\sim}{\mu_m} \right] s_k s_m \right\} + \left\{ \sum_{k=1}^{K} \left[\frac{1}{\sigma^2}\underset{\sim}{\mu_k}^T\underset{\sim}{x} - \frac{1}{2\sigma^2}\underset{\sim}{\mu_k}^T\underset{\sim}{\mu_k} + \log\pi_k - \log(1-\pi_k) \right] s_k \right\}$$

① ②

$$+ \text{constant}.$$

$$= \left\{ \sum_{i\neq j} w_{ij} s_i s_j \right\} + \left\{ \sum_i \theta_i s_i \right\} \quad \text{Boltzmann Machine}.$$

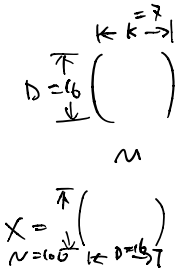
where $w_{ij} = -\frac{1}{2\sigma^2}\underset{\sim}{\mu_i}^T\underset{\sim}{\mu_j}$, $\theta_i = \frac{1}{\sigma^2}\underset{\sim}{\mu_i}^T\underset{\sim}{x} - \frac{1}{2\sigma^2}\underset{\sim}{\mu_i}^T\underset{\sim}{\mu_i} + \log\pi_i - \log(1-\pi_i)$

$$\Rightarrow p(x, s) \propto \exp\left(\sum_{i\neq j} w_{ij} s_i s_j \right) \cdot \exp\left(\sum_i \theta_i s_i \right) = \prod_{j\neq i} g_{ij}(s_i, s_j) \prod_i f_i(s_i).$$

① ②

$$= \prod_{i\neq j} \exp(w_{ij} s_i s_j) \cdot \prod_i \exp(\theta_i s_i).$$

$f_i(s_i)$.



(b) Note that the ② part in (a) can be regarded as a linear combination of S_k (sufficient statistics) and natural parameters. Therefore it is in exponential family. So we don't need to approximate f_i

But the ① part in (a) is not in exponential family.

We use EP / Loopy BP to approximate ① part.

Let $\tilde{g}_{ij}(s_i, s_j) = \exp(\beta_{ji} s_i + \beta_{ij} s_j)$ be bernoulli distribution to approximate

$g_{ij}(s_i, s_j) = \exp(w_{ij} s_i s_j)$ to ensure tractability.

Note $p(s) = \prod_i f_i(s_i) \cdot \prod_{i,j} g_{ij}(s_i, s_j)$

Rewrite $q(s) = \prod_i f_i(s_i) \prod_{i,j} \tilde{g}_{ij}(s_i, s_j)$.

The cavity distribution:

$$q_{\neg g_{ij}}(s) = \prod_k f_k(s_k) \cdot \prod_{(m,n)\neq(i,j)} \tilde{g}_{mn}(s_m, s_n)$$

$$\Rightarrow \tilde{g}_{ij}(s_i, s_j) = \arg\min_{\tilde{g}} KL\left\{ g_{ij}(s_i, s_j)\, q_{\neg g_{ij}} \,\middle\|\, \tilde{g}_{ij}(s_i, s_j) \cdot q_{\neg g_{ij}} \right\}$$

Left part of KL:

$$g_{ij}(s_i, s_j) q_{\neg g_{ij}} = \exp\left\{ w_{ij} s_i s_j + \underbrace{\sum_k \theta_k s_k}_{\text{constant expand from } s_i, s_j} + \sum_{k\neq i,j} \left(\beta_{ki} s_i + \beta_{kj} s_j\right) \right\}$$

$$\propto \exp\left\{ w_{ij} s_i s_j + \theta_i s_i + \theta_j s_j + \sum_{k\neq i,j} \left(\beta_{ki} s_i + \beta_{kj} s_j\right) \right\}$$

Right part of KL:

$$\tilde{q}_{ij}(s_i,s_j)\,q_{\neg ij} = \exp\left\{\beta_{ji}s_i + \beta_{ij}s_j + \sum_k \theta_k s_k + \sum_{k\neq i,j}\left(\beta_{ki}s_i + \beta_{kj}s_j\right)\right\}$$

$$\propto \exp\left\{\beta_{ji}s_i + \beta_{ij}s_j + \theta_i s_i + \theta_j s_j + \sum_{k\neq i,j}\left(\beta_{ki}s_i + \beta_{kj}s_j\right)\right\}.$$

$$= \exp\left\{\left(\beta_{ji}s_i + \theta_i s_i + \sum_{k\neq i,j}\beta_{ki}s_i\right) + \left(\beta_{ij}s_j + \theta_j s_j + \sum_{k\neq i,j}\beta_{kj}s_j\right)\right\}$$

$$= \exp\left\{\left(\beta_{ji} + \theta_i + \sum_{k\neq i,j}\beta_{ki}\right)s_i + \left(\beta_{ij} + \theta_j + \sum_{k\neq i,j}\beta_{kj}\right)s_j\right\}$$

Then do moment matching:

* In order to match s_i, we integrate out (sum out) s_j in the left equation: note $s_j = 0$ or $s_j = 1$ ($\sum_{s_j}$ means sum over $s_j=1$ and $s_j=0$)

$$\sum_{s_j} g_{ij}(s_i,s_j)\,q_{\neg ij} = \exp\left\{w_{ij} + \theta_i s_i + \theta_j + \sum_{k\neq i,j}\left(\beta_{ki}s_i + \beta_{kj}\right)\right\} + \exp\left\{\theta_i s_i + \sum_{k\neq i,j}\beta_{ki}s_i\right\}$$

$$= \exp\left\{\left(w_{ij} + \theta_i + \sum_{k\neq i,j}\beta_{ki}\right)s_i + \left(\theta_j + \sum_{k\neq i,j}\beta_{kj}\right)\right\} + \exp\left\{\left(\theta_i + \sum_{k\neq i,j}\beta_{ki}\right)s_i\right\}$$

$$\sum_{s_i}\sum_{s_j} g_{ij}(s_i,s_j)\,q_{\neg ij} = \exp\left\{\left(w_{ij} + \theta_i + \sum_{k\neq i,j}\beta_{ki}\right) + \left(\theta_j + \sum_{k\neq i,j}\beta_{kj}\right)\right\} + \exp\left\{\left(\theta_i + \sum_{k\neq i,j}\beta_{ki}\right)\right\}$$

$$+ \exp\left\{\left(\theta_j + \sum_{k\neq i,j}\beta_{kj}\right)\right\} + 1\,.$$

$$\Rightarrow \mathbb{E}_p(s_i) = \frac{\sum_{s_i=1}\sum_{s_j} g_{ij}(s_i,s_j)\,q_{\neg ij}}{\sum_{s_i}\sum_{s_j} g_{ij}(s_i,s_j)\,q_{\neg ij}}$$

$$= \frac{\exp\left\{\left(w_{ij} + \theta_i + \sum_{k\neq i,j}\beta_{ki}\right)\overset{=1}{\cancel{s_i}} + \left(\theta_j + \sum_{k\neq i,j}\beta_{kj}\right)\right\} + \exp\left\{\left(\theta_i + \sum_{k\neq i,j}\beta_{ki}\right)\overset{=1}{\cancel{s_i}}\right\}}{\exp\left\{\left(w_{ij} + \theta_i + \sum_{k\neq i,j}\beta_{ki}\right) + \left(\theta_j + \sum_{k\neq i,j}\beta_{kj}\right)\right\} + \exp\left\{\left(\theta_i + \sum_{k\neq i,j}\beta_{ki}\right)\right\} + \exp\left\{\left(\theta_j + \sum_{k\neq i,j}\beta_{kj}\right)\right\} + 1}$$

Note $\mathbb{E}_q(S_i) = \mathbb{E}_p(S_i)$, where under the distribution q,

$\tilde{q}_{ij}(s_i, s_j) q_{\neg ij}$ is Bernoulli distribution by our assumption.

and $\tilde{q}_{ij}(s_i, s_j) q_{\neg ij} = \exp\left\{ \left(\beta_{ji} + \theta_i + \sum_{k \neq i,j} \beta_{ki}\right) s_i + \left(\beta_{ij} + \theta_j + \sum_{k \neq i,j} \beta_{kj}\right) s_j \right\}$

$$= \exp\left\{ \left(\beta_{ji} + \theta_i + \sum_{k \neq i,j} \beta_{ki}\right) s_i \right\} \cdot \exp\left\{ \left(\beta_{ij} + \theta_i + \sum_{k \neq i,j} \beta_{ki}\right) s_i \right\}.$$

Match the part with s_i → Nature Parameter

Recall for Bernulli Distribution: $p(s|\pi) = \pi^s (1-\pi)^{(1-s)}$

$$= \exp\left\{ \log\left(\frac{\pi}{1-\pi}\right) s + \log(1-\pi) \right\}$$

Natural Parameter.

Also note $\mathbb{E}_p(s) = \pi$

Therefore the update for natural parameters under

$$\beta^*_{ji} + \theta_i + \sum_{k \neq i,j} \beta_{ki} = \log \frac{\mathbb{E}_q(s_i)}{1 - \mathbb{E}_q(s_i)}.$$

$$\Rightarrow \beta^*_{ji} = \log \frac{\mathbb{E}_q(s_i)}{1 - \mathbb{E}_q(s_i)} - \theta_i - \sum_{k \neq i,j} \beta_{ki}$$

Denote $\theta_i + \sum_{k \neq i,j} \beta_{ki} = \nu_i$

$\theta_j + \sum_{k \neq i,j} \beta_{kj} = \nu_j$

$= \log(\exp(-\nu_i))$

$$= \log\left(\frac{\frac{\exp(w_{ij} + \nu_i + \nu_j) + \exp(\nu_i)}{1 + \exp(w_{ij} + \nu_i + \nu_j) + \exp(\nu_i) + \exp(\nu_j)}}{1 - \frac{\exp(w_{ij} + \nu_i + \nu_j) + \exp(\nu_i)}{1 + \exp(w_{ij} + \nu_i + \nu_j) + \exp(\nu_i) + \exp(\nu_j)}} \right) + \log(\exp(-\nu_i))$$

$$= \log\left(\frac{\exp(w_{ij} + \nu_i + \nu_j) + \exp(\nu_i)}{1 + \exp(\nu_j)} \right) + \log \exp(-\nu_i)$$

$$= \log\left(\frac{\exp(w_{ij} + \nu_j) + 1}{1 + \exp(\nu_j)} \right)$$

Conclude:

$$\beta^*_{ji} = \log\left(\frac{\exp(w_{ij} + \nu_j) + 1}{\exp(\nu_j) + 1}\right), \quad \beta^*_{ij} = \log\left(\frac{\exp(w_{ij} + \nu_i) + 1}{\exp(\nu_i) + 1}\right).$$

where: $\nu_i = \theta_i + \sum_{k \neq i,j} \beta_{ki}$

$$\nu_j = \theta_j + \sum_{k \neq i,j} \beta_{kj}$$

(c) Write pairwise factor $g_{ij}(s_i, s_j)$ into message passing:

$$g_{ij}(s_i, s_j) \approx \tilde{g}_{ij}(s_i, s_j) = M_{i\to j}(s_j) M_{j\to i}(s_i)$$

Rewrite $q(s)$ into beliefs

$$q(s) = \prod_i f_i(s_i) \prod_{i,j} \tilde{g}_{ij}(s_i, s_j) = \prod_i \left(f_i(s_i) \prod_{i,j} M_{i\to j}(s_j) M_{j\to i}(s_i) \right) = \prod_i b_i(s_i)$$

Cavity distribution can be rewrite into:

$$q_{\neg ij}(s_i, s_j) = \int \prod_{s_k} f_k(s_k) \prod_{(m,n)\neq(i,j)} \tilde{g}_{mn}(s_m, s_n)$$

$$= f_i(s_i) f_j(s_j) \prod_{\alpha \in ne(i)\setminus j} M_{\alpha\to i}(s_i) \prod_{\beta \in ne(j)\setminus i} M_{\beta\to j}(s_j)$$

$$\Rightarrow \left\{ M^{new}_{i\to j}, M^{new}_{j\to i} \right\} = \underset{M_{i\to j}, M_{j\to i}}{\operatorname{argmin}} KL \left\{ g_{ij}(s_i, s_j) q_{\neg ij}(s_i, s_j) \,\middle\|\, M_{i\to j}(s_j) M_{j\to i}(s_i) q_{\neg ij}(s_i, s_j) \right\}$$

where $\tilde{g}_{ij}(s_i, s_j) = \exp(\beta_{ji} s_i + \beta_{ij} s_j) = \exp(\beta_{ij} s_j) \cdot \exp(\beta_{ji} s_i)$

$$= M_{i\to j}(s_j) \cdot M_{j\to i}(s_i)$$

This is the same as:

$$\tilde{g}_{ij}(s_i, s_j) = \underset{\tilde{g}}{\operatorname{argmin}} KL \left\{ g_{ij}(s_i, s_j) q_{\neg ij}(s_i, s_j) \,\middle\|\, \tilde{g}_{ij}(s_i, s_j) q_{\neg ij}(s_i, s_j) \right\}$$

We then get updated messages by:

$\sum_{s_i} f_i(s_i) g_{ij}(s_i, s_j) \prod_{\alpha \in ne(i)\setminus j} M_{\alpha\to i}(s_i)$, project into exponential family, get $M^{new}_{i\to j}$

Similar for $M^{new}_{j\to i}$

(d) By the Automatic Relevance determination (ARD) described in lecture. Apply it to loopy BP

Then we do a "Hyper" M-step to optimise hyperparameter K

$$K \leftarrow \arg\max \langle \log P(\mu|\pi) \rangle_{q\mu}.$$

In this step, the dimension of S is selected, effectively learning the number of K.

Solution for computational difficulty:
Introduce additionally latent variables

3. Implement EP.

As we derived in 2(b), we have:

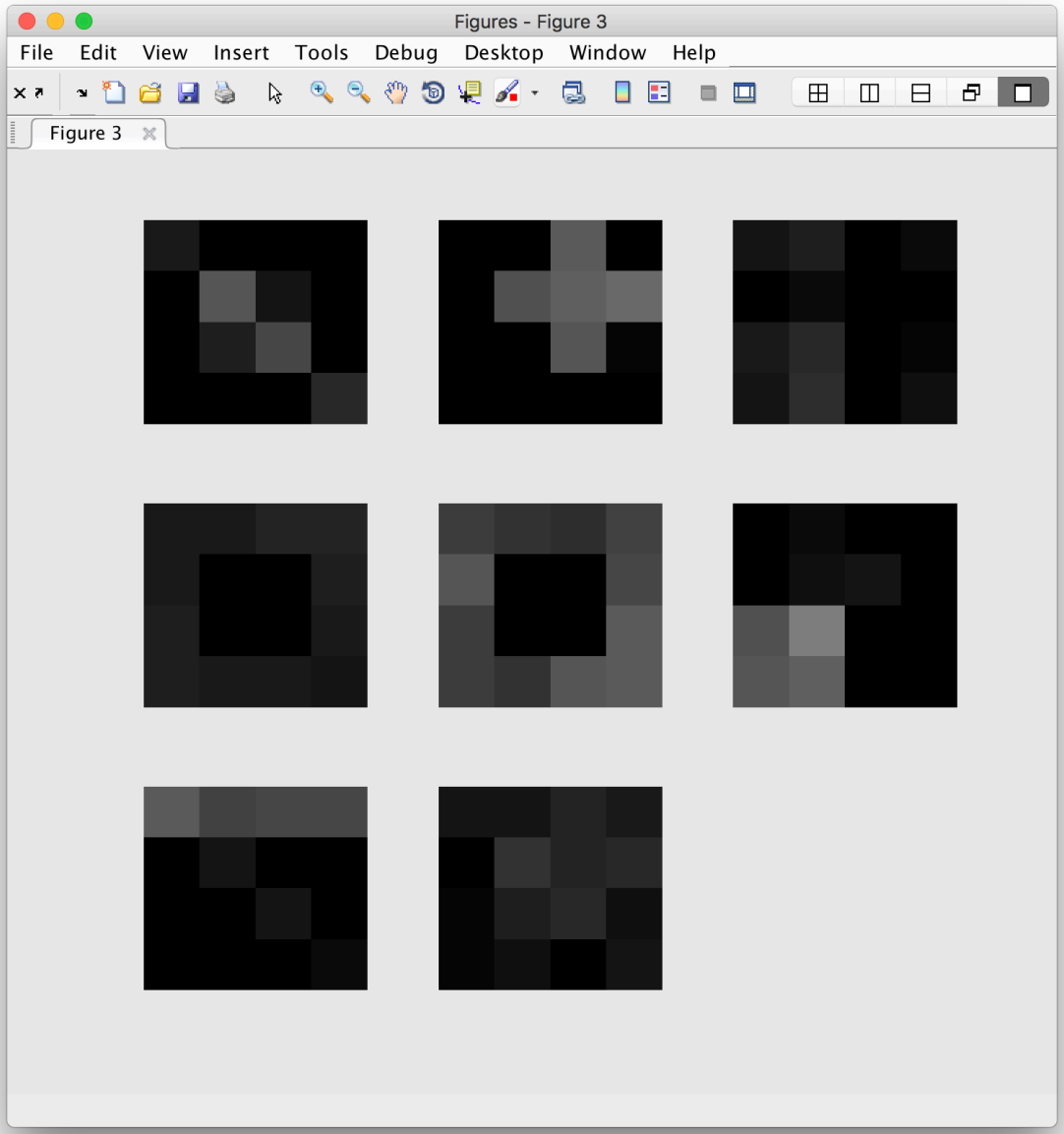
$$\beta^*_{ji} + \theta_i + \sum_{k \neq i,j} \beta_{ki} = \log \frac{\mathbb{E}_q(s_i)}{1 - \mathbb{E}_q(s_i)}.$$

Therefore $\mathbb{E}_q(s_i) = \text{sigmoid}\left(\beta^*_{ji} + \theta_i + \sum_{k \neq i,j} \beta_{ki}\right)$

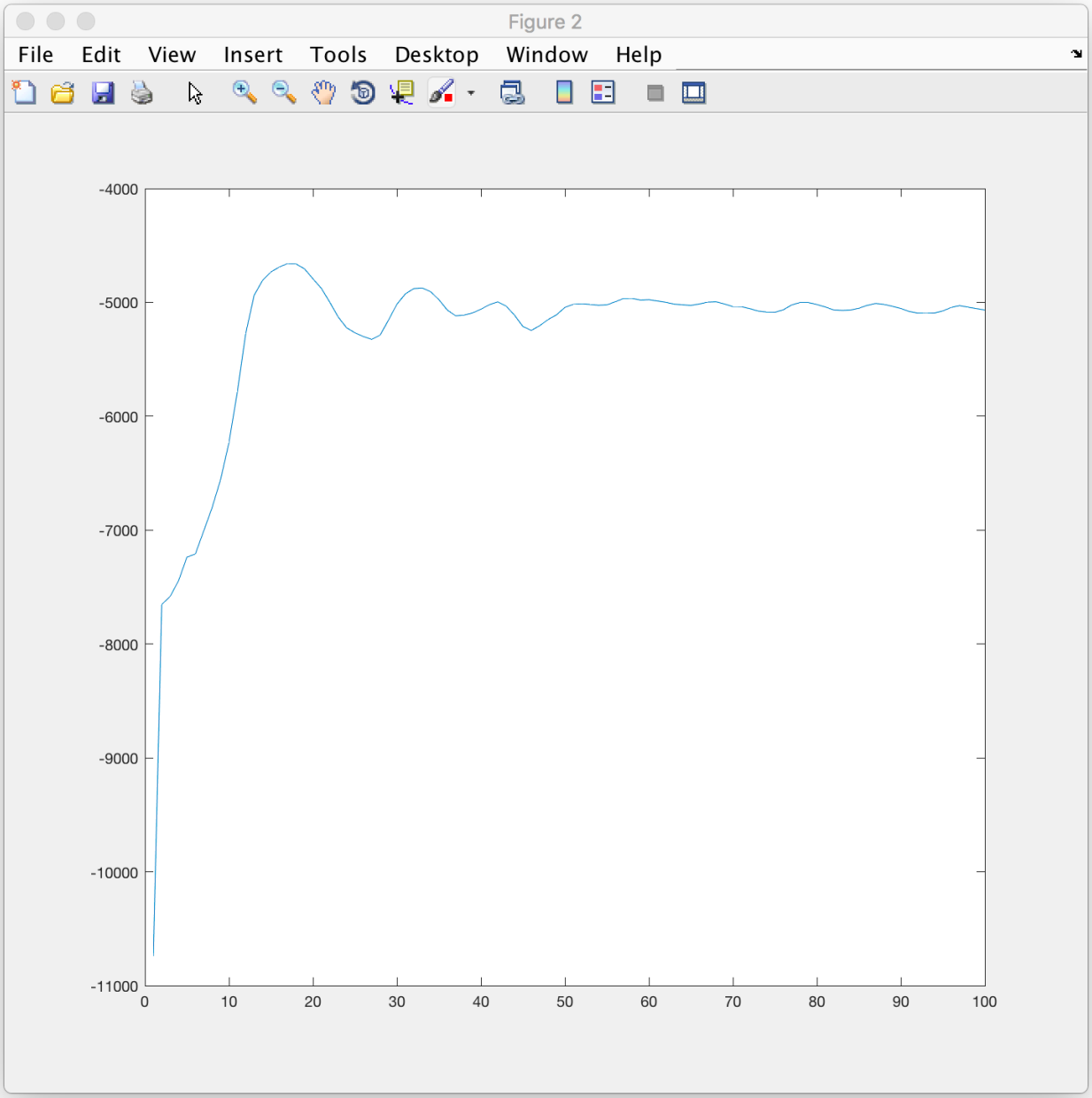
$$= \text{sigmoid}\left(\underbrace{\beta^*_{ji} + \lambda_i}_{XX(i)}\right).$$

RESULTS:

1. Features Learnt from EP:

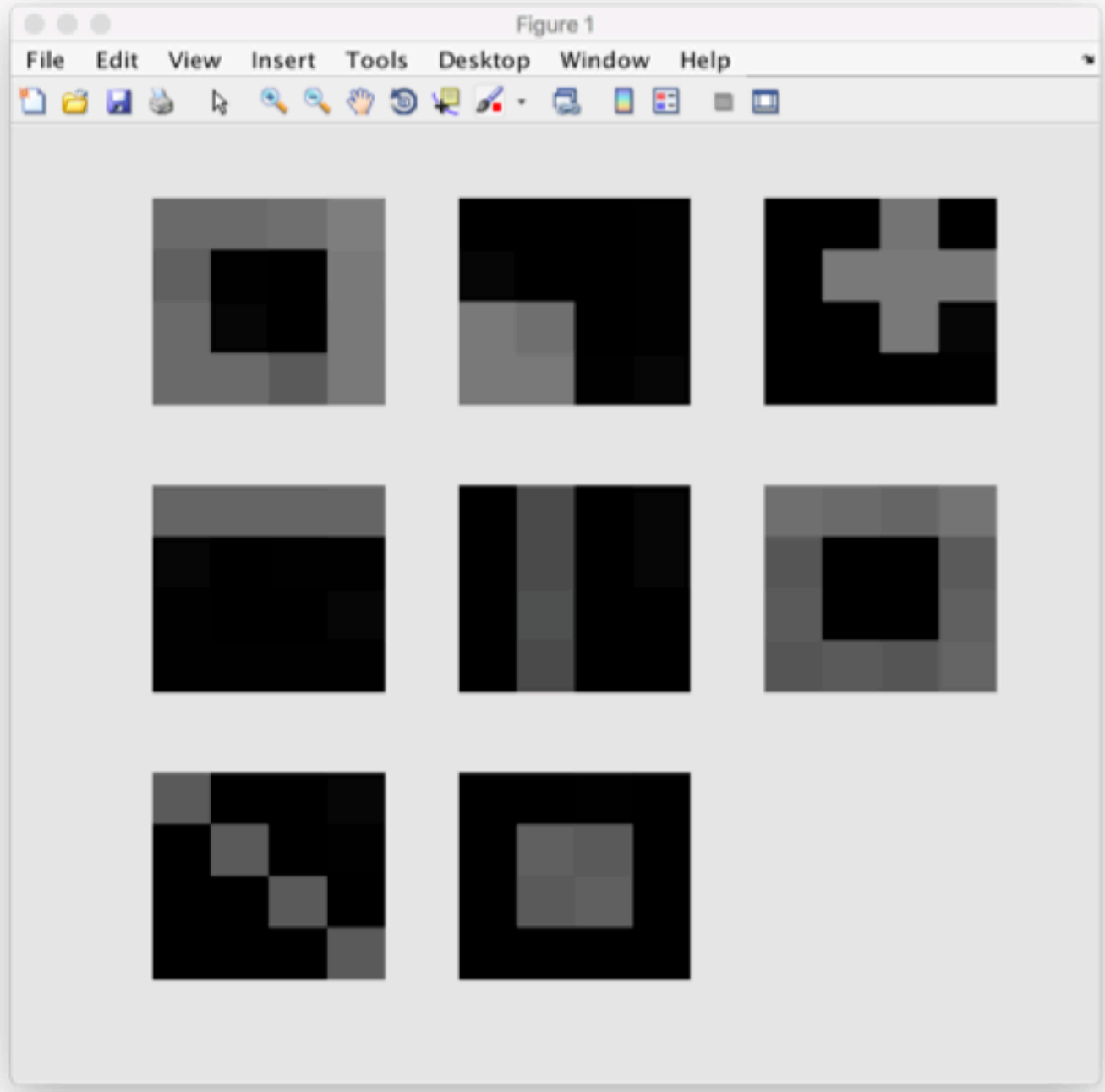


2. Corresponding Free energy with 100 iterations



The 'free energy' for EP does not always increase.

3. Compare the results with Mean Field:



Comparison with features learnt from EP:

It seems that my mean field feature is better than the EP feaures. This might be biased as I ran 100 times mean field in assignment 5 and chose the one with the best free energy. But on average, the features learnt by mean field is not very clear and very similar to the features we learnt by EP here.

In theory, EP should have better results. But here we might have some comvergence issues. I have tried power EP update as described in the last slide in the EP lecture, which gives similar results.