So far we have considered two variational approaches to the general Boltzmann distribution where pair-wise connections might be defined between all nodes in the graph. However, we are interested in the adaptation of the extended mean field approximation to the restricted Boltzmann machine.

0.1. Adaptation of EMF to RBM

Adaptation of the extended mean field approximation derived in the first chapter to the case of the RBM is rather straightforward. Let's divide set of spins into visible and hidden variables along with corresponding biases (a and b for visible and hidden units respectively). We will denote by $\mathbf{m^v} = \{m_i\}_{i=1}^N$ and $\mathbf{m^h} = \{m_i\}_{j=1}^M$ corresponding sets of magnetizations where N and M are the sizes of the visible and hidden layers accordingly. The energy in the BM models is set to 1 thus we set β to 1 as well. This leads to the following free energy expansion (up to the third term) in the new setting:

$$\begin{aligned} F^{EMF}(\mathbf{m^v}, \mathbf{m^h}) \simeq\ & H(\mathbf{m^v}, \mathbf{m^h}) \\ & - \sum_i a_i m_i^v - \sum_j b_j m_j^h \\ & - \sum_{i,j} \left(m_i^v w_{ij} m_j^h + \frac{w_{ij}^2}{2}(m_i^v - (m_i^v)^2)(m_j^h - (m_j^h)^2) \right) \\ & - \sum_{i,j} \left(\frac{2w_{ij}^3}{3}(m_i^v - (m_i^v)^2)(\frac{1}{2} - m_i^v)(m_j^h - (m_j^h)^2)(\frac{1}{2} - m_j^h) \right), \end{aligned} \tag{1}$$

where $H(\cdot)$ denotes the entropy of magnetizations. In the case of the RBM, the third term consists only of the sum of pair connection because the coupled triplets are excluded by the bipartite structure of the RBM [**?**]. To recover the true free energy we set the external fields to $\mathbf{0}$ which by conjugacy yields the self-consistency constraints $\frac{\mathrm{d}G}{\mathrm{d}\mathbf{m}} = \mathbf{0}$. This stationary condition might be interpreted as a requirement that in the equilibrium where magnetizations perfectly describes the average configuration of spins under the Boltzmann measure, the variational free energy reaches its minimum. This leads to the following constraint on the i-th visible magnetization:

$$\frac{\partial F^{EMF}}{\partial m_i^v} = 1 + \ln m_i - 1 - \ln(1 - m_i^v) - R = 0 \tag{2}$$

where

$$R = a_i + \sum_j w_{ij} m_j^h - \sum_j w_{ij}^2 \left(m_i^v - \frac{1}{2} \right) (m_j^h - (m_j^h)^2) + \sum_j \frac{w_{ij}^3}{3} \left(m_i^v - (3m_i^v)^2 + 2(m_i^v)^3 \right) (m_j^h - (m_j^h)^2)(\frac{1}{2} - m_j^h).$$

This can be regrouped as:

$$\ln \left(\frac{m_i^v}{1 - m_i^v} \right) = R$$

which leads to the following

$$m_i^v = \frac{\exp(R)}{1 + \exp(R)} = \mathrm{sigm}(R) \tag{3}$$

where $\mathrm{sigm}(x) = (1 + e^{-x})^{-1}$. Similar condition can be obtained for $\{m_j^h\}_{j=1}^M$. These consistency relations can be defined for an arbitrary order of the approximation. Thus, the hidden and visible magnetizations are the solutions of a set of non-linear equations that can be recognized as the extended mean field equations for a spin system. We can pose a question how to efficiently define a schedule of updates of magnetizations which will eventually satisfy self-consistency constraints. This will allow us to compute extended mean field approximation for the partition function **??**.

0.2. Schedule of updates

The choice of the update procedure is of crucial importance for the convergence of the magnetizations. It was observed in the case of mean field updates for Boltzmann machines that updates have to be run sequentially [**?**]. Similarly, in the case of the extended mean field approximation, it was proposed that an iterative, asynchronous algorithm may serve as update rules [**?**] following positive theoretical results proved in the context of random spin glass model. However, there are many heuristically reasonable ways to perform such sequential updates as well as it is interesting how different procedures might affect the convergence. Thus, I will analyse three different update rules for magnetizations on a toy model and on the real life data set example. The updates here are considered only up to the second order.

0.2.1. Asynchronously

The structure of the RBM suggests that the updates might be performed layer-wise. At each iteration, the whole hidden layer is updated with visible magnetizations fixed at the values from the previous step. This can be written using the time index t in the following way:

$$\begin{aligned}
\mathbf{m}^h[t+1] &= \mathrm{sigm}\left[\mathbf{b} + W\mathbf{m}^v[t] - \left(\mathbf{m}^h[t] - \frac{1}{2}\right)^T \odot W^2\left(\mathbf{m}^v[t] - (\mathbf{m}^v[t])^2\right)\right], \\
\mathbf{m}^v[t+1] &= \mathrm{sigm}\left[\mathbf{a} + W^T\mathbf{m}^h[t+1] - \left(\mathbf{m}^v[t] - \frac{1}{2}\right) \odot (W^2)^T\left(\mathbf{m}^h[t+1] - (\mathbf{m}^h[t+1])^2\right)\right],
\end{aligned} \tag{4}$$

where $\odot$ denotes Hadamard product.

0.2.2. Sequentially

Previous procedure takes advantage of the bipartite structure of the model. However, we might consider updates not in the vectorize way but rather in sequential manner:

$$\begin{aligned}
m_i^h[t+1] &= \mathrm{sigm}\left[b_i + \sum_j \left(w_{ij}m_j^v[t] - w_{ij}^2(m_i^h[t] - \frac{1}{2})\left(m_j^v[t] - (m_j^v[t])^2\right)\right)\right], \\
m_{j=i+1}^v[t+1] &= \mathrm{sigm}\Bigg[a_i + \sum_{l\neq i} \left(w_{ij}m_l^h[t] - w_{ij}^2(m_l^h[t] - \frac{1}{2})(m_j^v[t] - (m_j^v[t])^2)\right) \\
&\qquad + \left(w_{ij}m_i^h[t+1] - w_{ij}^2(m_i^h[t+1] - \frac{1}{2})(m_j^v[t] - (m_j^v[t])^2)\right)\Bigg]
\end{aligned} \tag{5}$$

where $i \in \{1, ..., M\}$, $j \in \{1, ..., N\}$. This implies imbalance in numbers of updates performed between hidden and visible layers if $N \neq M$.

0.2.3. Parallelly

Finally, one could consider parallel updates where both visible and hidden magnetizations are updated at the same time. This might be summarized as follows:

$$\begin{aligned}
\mathbf{m}^h[t+1] &= \mathrm{sigm}\left[\mathbf{b} + W\mathbf{m}^v[t] - \left(\mathbf{m}^h[t] - \frac{1}{2}\right)^T \odot W^2\left(\mathbf{m}^v[t] - (\mathbf{m}^v[t])^2\right)\right], \\
\mathbf{m}^v[t+1] &= \mathrm{sigm}\left[\mathbf{a} + W^T\mathbf{m}^h[t] - \left(\mathbf{m}^v[t] - \frac{1}{2}\right) \odot (W^2)^T\left(\mathbf{m}^h[t] - (\mathbf{m}^h[t])^2\right)\right].
\end{aligned} \tag{6}$$

This schedule of updates poses a risk that the model might not learn the proper transfer of information from one layer to another which is in contrast with the structure of the RBM.

Figure 1 presents graphically all proposed procedures.

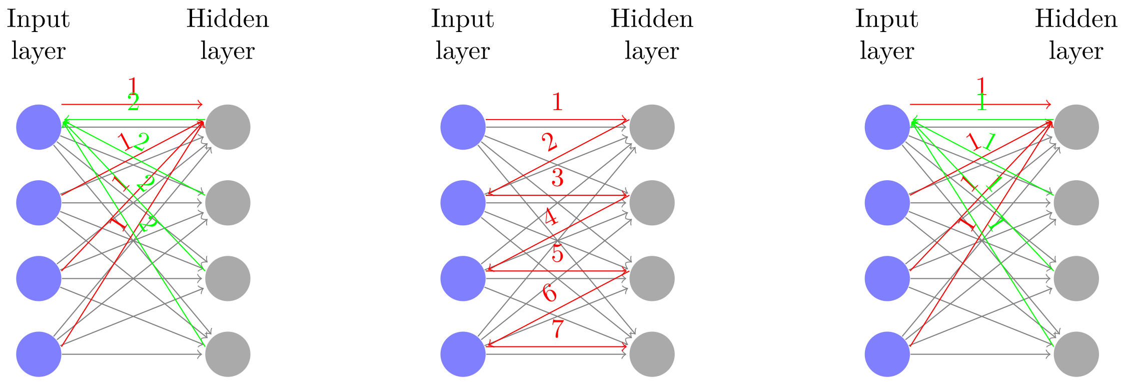

Figure 1: Graphical visualisations of three different schedule of updates considered for the RBM toy model. Numbers above the coloured arrows denotes the order of updates.

In the case of fixed point algorithms, it is a common practice to use damped updates [**?**] where as a new value for the given magnetization we take weighted average of the its value from the previous step and after performing an update. The weight hyper-parameter λ is usually in the range $[0, 1]$. Damping operation helps in avoiding unnecessary artefacts and oscillations. In all experiments conducted in this and the following chapters, updates will be damped with λ set to 0.5 following other authors [**?**], [**?**].

0.3. Toy models

As it was mentioned in the previous chapter unlike naive mean field approach, the TAP approximation doesn't provide us with an upper or lower bound for the variational free energy. In order to adapt the EMF approximation to the RBM model we set β to 1 which means that the temperature is also 1 while the approximation was derived for an infinite temperature. Thus, the radius of convergence for the Taylor expansion might be not big enough to obtain reliable estimate of magnetizations. That is why, two toy models (a grid model and a small RBM) were created in order to perform an exact inference which will allow us to assess the quality of the EMF approximation before turning to real data set which requires much bigger and powerful modelling structures. The analysis will be made assuming that the parameters of the models are known a priori.

0.3.1. Grid toy model

A small grid toy model was considered of size 4×4 with periodic boundary conditions in order to avoid edge effects – Figure 2 shows this model from the graphical model's perspective. The nature of this model implies that the sequential updates of magnetizations seems as the most natural way to obtain a statistics of the system in the equilibrium and that is why only updates of the form 5 will be considered here. In this case each magnetization m_i is updated one at a time using equation 3.

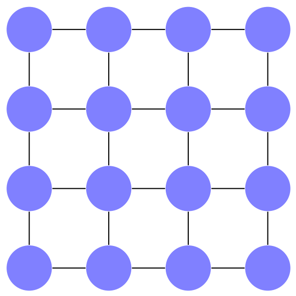

Figure 2: 4×4 grid toy model used for an exact inference.

Initially, the external field was set to 0 and I considered the case when all couplings have the same value ranging from -1 to 1. As it was expected, the naive mean field approach is an upper bound for the variational free energy. However, even in the case of this small model the TAP approximation for different values of couplings is either upper or lower bound. We can see that the approximation is closest to the ground truth when the couplings are close to zero. This is consistent with the fact that the approximation was performed around point where the temperature T is infinite which means that spins are independent – small values of couplings imitate this state.

Another computational inference problem that can be evaluate thanks to the TAP method is computing a mode of the marginal density for a given spin – in this case we can estimate average value of the spin under the Boltzmann distribution. The right plot in the Figure 3 shows the mean squared error (MSE) between the real and estimated magnetizations for all spins. In this case, the TAP approach provides much better estimates than the naive method – we can see that adding a second term to the approximation allows to properly model the connections in the system between the spins.

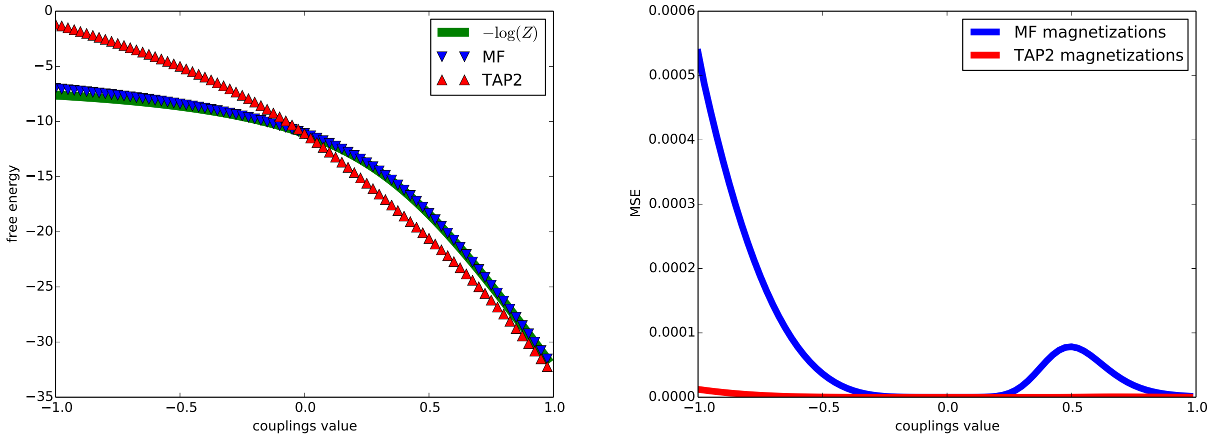

Figure 3: Comparison of two variational approaches – free energy estimates (left)with the true free energy (green line) and MSE between real and estimated magnetisations (right) as a function of the couplings strength ranging from -1 to 1.

In the next experiment, all couplings were initialised to random values around "mean" strength which varies from 0 to 1 and then randomly assigned with positive or negative sign. The results are similar to the one observed previously (Figure 4). The naive approach gives consistently better approximation for the $-\ln Z$ while the TAP method performs better in the case of estimating an average value of spin.

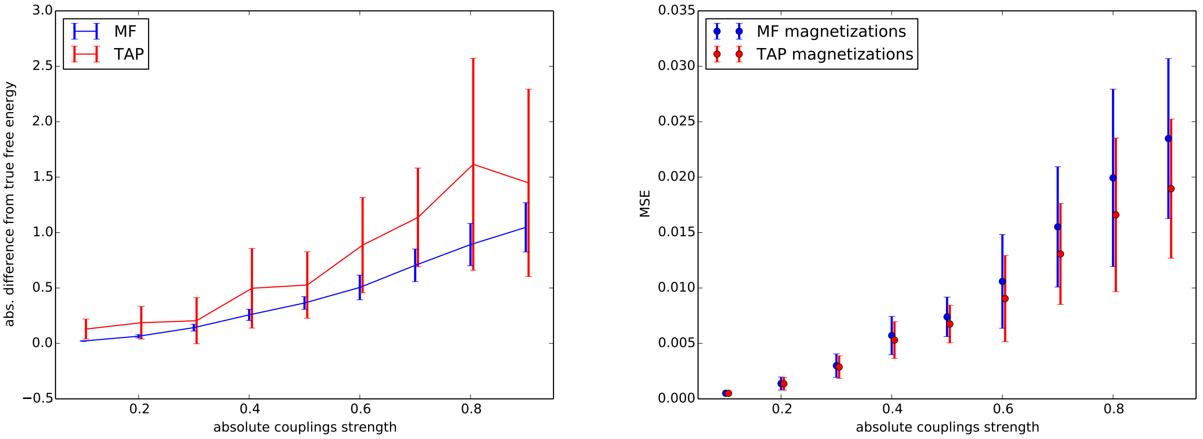

Figure 4: Absolute difference between true free energy and the one computed by naive and extended mean field approach (left) an MSE between real and estimated magnetisations (right) as a function of the absolute value of couplings strength.

TODO: external fields.

0.3.2. RBM toy model

Due to the different structure of connections between states, the RBM toy model is much tougher to approximate. This will lead to substantially different results in the performance on toy model and it is another suggestion to use the extended mean field approach on the real data set.

Unlike in the previous case, there is no strong heuristics how the updates of self-consistency relations should be performed. The literature suggests that in the case of the naive approach it is necessary to run self-consistency equations sequentially [**?**]. To assess the impact on to final estimates, all three different schedules of updates will be considered here. Following the analysis from the previous section, initially all couplings were set to the same value ranging from -1 to 1 (Figure 5).

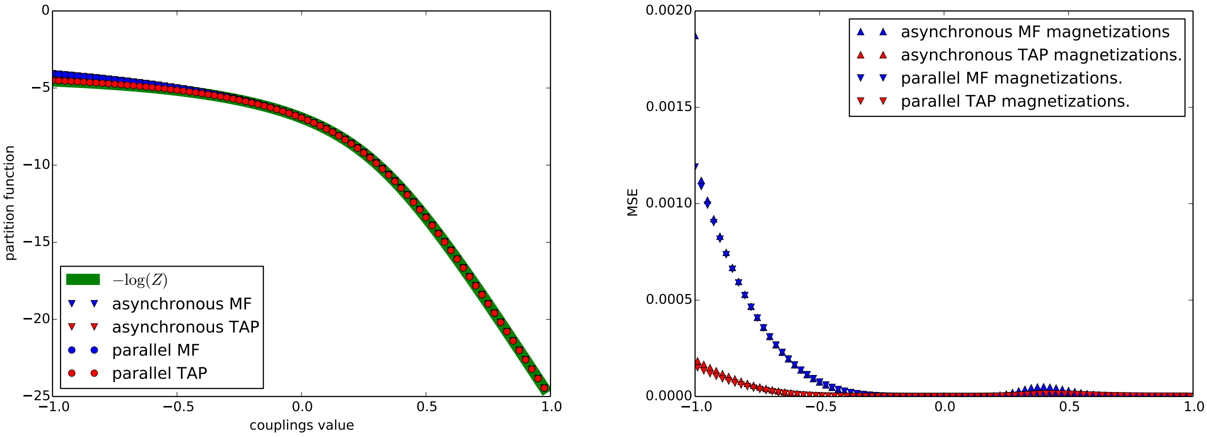

Figure 5: Comparison of two variational approaches – free energy estimates (left)with the true free energy (green line) and MSE between real and estimated magnetisations (right) as a function of the couplings strength ranging from -1 to 1.

The estimation of the free energy is almost exact in the case of the TAP method while the naive mean field method again provides a slightly biased upper bound. As it was the case on the grid model, the magnetizations estimated using extended approximation are very precise while MF magnetizations shows discrepancies from true values when connections become stronger in the model. No significant differences were observed between different schedules of updates.

Unlike the case of the grid model, when couplings were random with randomly assigned negative or positive sign, TAP approximation yields consistently much better estimates which most of the time are exact at the same time having much smaller variance (Figure 6). Again, differences between schedules of updates were negligible and thus the results for sequential updates weren't included as they were almost identical to the ones obtained with asynchronous ones.

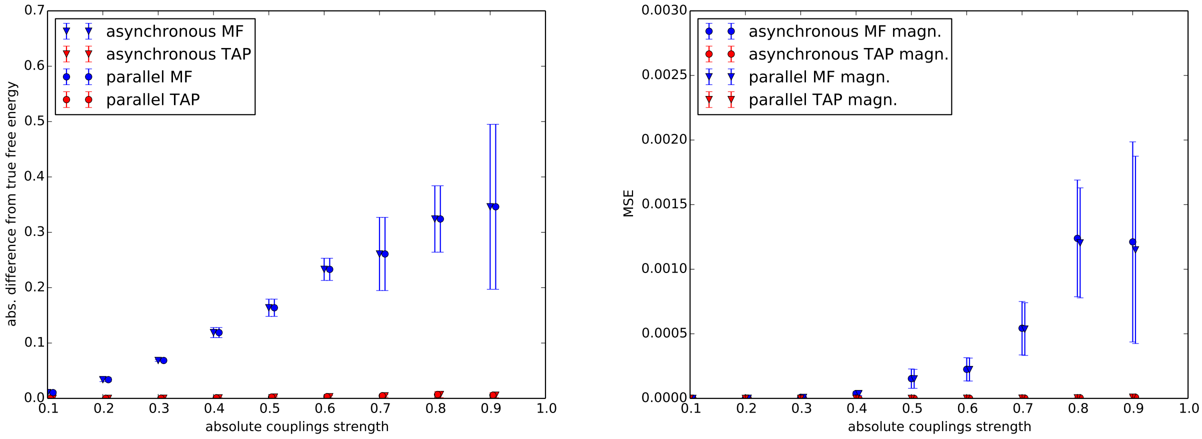

Figure 6: Absolute difference between true free energy and the one computed by naive and extended mean field approach (left) an MSE between real and estimated magnetisations (right) as a function of the absolute value of couplings strength.

The randomness associated with choosing the sign of connections might have averaged the overall statistics of the model, which in turn might affect the effectiveness of different schedules of updates. Thus, to assess how robust the analysed extended method is along with different schedules, the couplings were chosen again randomly around given mean value but this time the sign of the weight was chosen sequentially. Figure 7 presents the results:

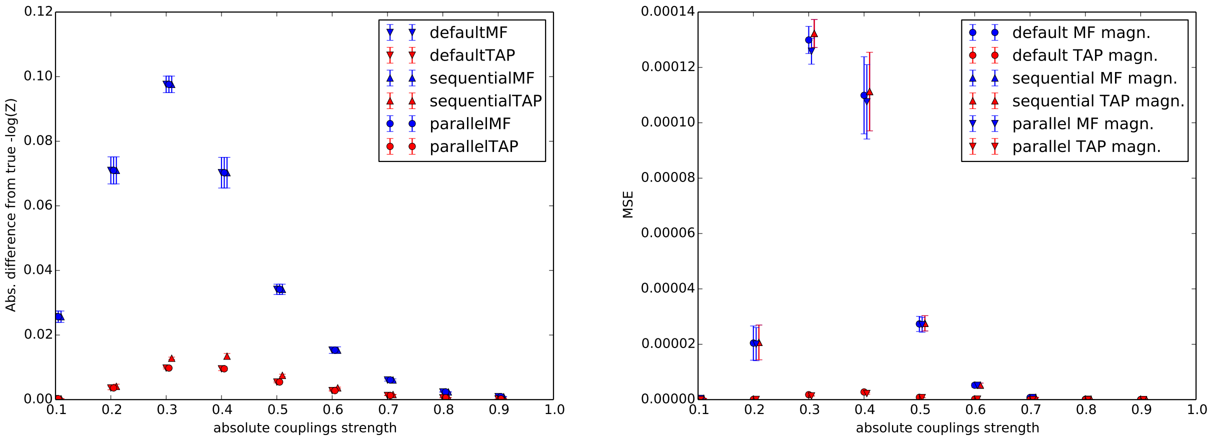

Figure 7: Absolute difference between true free energy and the one computed by naive and extended mean field approach (left) an MSE between real and estimated magnetisations (right) as a function of the absolute value of couplings strength with sequential changes of signs.

Talk about seqiential

1. radius of convergence - works fine 2. udates - no sequential 3.