University of Cambridge

Thesis

Paweł Budzianowski, pfb30, Clare Hall College

Contents

1. Introduction

2. Extended Mean Field Approximation

2.1. Graphical models as Markov Random Fields

One of the basic concepts in statistical modelling are graphical models which greatly help in analysing multivariate phenomenon. Visualization by graphs help with efficient development and understanding analysed models while complex computations can be performed Consider a graph $G = (V, E)$ which consists of a finite set of vertices V and a collection of edges $E \subset V \times V$. Each edge $e_i \in E$ joins two vertices and in general may have a direction. The vertex $v \in V$ may be seen as a random variable X_v defined on some space $\mathcal{X}_v$ that may be either continuous or discrete. We will use the notion of *clique* which is a subset of V in which all nodes are pairwise connected. One of the most useful family of models are Markov Random Fields which are type undirected random fields which satisfies global Markov property, specifically:

Definition 1 *[**koller2009probabilistic**] An undirected graphical model G is a Markov Random Field (MRF) if for any node X_v in the graph the following conditional property holds:*

$$P(X_i|X_{G\setminus i}) = P(X_i|X_{N(i)})$$

where $X_{G\setminus i}$ denotes all the nodes except X_i, and $X_{N(i)}$ denotes the set of all vertices connected to X_i.

Thus, the MRF has a desired property that any two nodes are conditionally independent given some evidence nodes that separates them. This property is closely related with the notion of factorization of the joint probability distribution:

Definition 2 *A probability distribution $P(athbfX)$ on an undirected graphical model G factorizes over G if there exists a set of non-negative functions (potentials) $\{\psi_C\}_{C\in\mathcal{C}}$ functions on cliques that cover all the nodes and edges of G and we can write:*

$$P(X_1, X_2, ..., X_n) = \frac{1}{Z}\prod_{C\in\mathcal{C}} \psi_C(x_C)$$

where $\mathcal{C}$ is a set of all cliques in G and Z is a normalization constant $Z = \sum_{\mathbf{x}} \prod_{C\in\mathcal{C}} \psi_C(X_C)$ which is often called a partition function.

Theorem 1 (Hammersley-Clifford, 1972?) *Strictly positive distribution $P(\mathbf{X})$ is MRF w.r.t an undirected graph G if and only if it factorizes over G.*

This theorem ensures us that there exits a general factorization form of the distribution of MRFs. It follows from the strict positivity of P that we can write:

$$p(x_1, x_2, ..., x_n) = \frac{1}{Z}\prod_{C\in\mathcal{C}} \psi_C(x_C) = \frac{1}{Z} e^{\sum_{C\in\mathcal{C}} \ln \psi_C(x_C)} = \frac{1}{Z} e^{-E(x)}$$

where $E(x)$ is called an energy function. This general form of distribution is usually defined as *Gibbs distribution*. Hence, the probability distribution of every MRF can be expressed as in **??**. This relationship allow us to take advantage of Moreover, this form of distribution is a natural candidate to approximate and model phenomenon which can be seen as graphical models. In next sections we will analyse one of examples of Gibbs which will is powerful enough to approximate any probability distribution (TODO - find concept of apprximation by BM)

2.2. Boltzmann distribution

In this thesis, a primary undirected graphical model (or MRF) which will be analysed in the general form has the joint distribution:

$$p(x_1, x_2, ..., x_n) = \frac{1}{Z}\exp(-E(x_1, x_2, ..., x_n)$$

where E is the *energy* of the system of the form:

$$E(\mathbf{X}) = -\sum_{(ij)} w_{ij} x_i x_j - \sum_i \theta_i x_i$$

The pair-wise potential function have here the form:

$$\psi_{i,j} = \exp(x_i w_{ij} x_j)$$

while the magnetic fields defined as:

$$\psi_i = \exp(a_i x_i)$$

The name comes from the Boltzmann distribution which extensively used in physics to compute the energy of the system of particles. This model proves to be very useful for many applications such as the error-correcting code, computer vision, medical diagnosis or statistical mechanics [**yedidia2001idiosyncratic**]. This model may represents the statistical dependencies between different variables through the weight link w_{ij} as well as the evidence for the specific variable. However, computing the partition function requires to sum over a number of states that grows exponentially with the number of variables and is untractable even for a small number of variables. That is why, we have to resort to some tractable approximation which one of them will be analysed in next section.

2.3. Statistical Perspective

Consider energy based model (initially without latent variables) of the form:

$$P(\mathbf{s}) = \frac{e^{-E(\mathbf{s})}}{\sum_{\mathbf{s}} e^{-E(\mathbf{s})}} = \frac{1}{Z} e^{-E(\mathbf{s})} \tag{1}$$

where energy is defined as:

$$E \equiv E(\mathbf{s}) = -\sum_{(ij)} s_i w_{ij} s_j - \sum_i a_i s_i$$

If we restrict each node to have two states, $s_i \in \{0, 1\}$, we obtain well-known the Ising model [**yedidia2001idiosyncratic**]. Restricting the w_{ij} to be positive we obtain the ferromagnetic Ising model. Finally, assuming that the w_{ij} are chosen from a random distribution, we obtain the Ising spin glass model. Instead of imposing some restrictions on model structure we will try to find a approximate distribution Q (which poses useful characteristics) that minimizes relative entropy often called Kullback-Leibler divergence:

$$KL(Q||P) = \mathbb{E}_Q \left(\ln \frac{Q}{P} \right) = \sum_{\mathbf{s}} Q(\mathbf{s}) \ln \frac{Q(\mathbf{s})}{P(\mathbf{s})} \tag{2}$$

The KL-divergence is non-symmetric measure of the difference between two distributions which is always non-negative. Substituting P from 1 into previous equation gives us:

$$KL(Q||P) = \ln Z + E[Q] - H[Q]$$

where H stands for entropy of the distribution Q, $\ln Z$ is the *free energy* and $E[Q] = \sum_{\mathbf{s}} Q(\mathbf{s}) E(\mathbf{s})$ is called the *variational energy* [**opper2001advanced**]. The partition function Z doesn't depend on Q and we need to only focus on minimizing the variational free energy:

$$F[Q] = E[Q] - H[Q]. \tag{3}$$

2.4. Mean Field Approximation

2.5. Extended Mean Field Approximation

Mean field approach (TODO: murphy here why MF):

$$Q(\mathbf{s}) = \prod_i q_i(x_i) \tag{4}$$

MF theory is exact for infinite-ranged Ising model. In the case of spin glass model Thouless, Anderson and Palmer (TAP) showed the corresponding Gibbs free energy with additional "Onsager term".

$$\ln Z = \ln \sum_{\mathbf{s}} \exp(-E(\mathbf{s})) = \ln \sum_{\mathbf{s}} q(\mathbf{s}) \frac{\exp(-E(\mathbf{s}))}{q(\mathbf{s})} \geqslant \sum_{\mathbf{s}} q(\mathbf{s}) \ln \frac{\exp(-E(\mathbf{s}))}{q(\mathbf{s})} = -\mathbb{E}(E(\mathbf{s})) + H(Q) = -F \tag{5}$$

The bound F on log Z is called *the mean field free energy.*

We will minimize 3 where we instead of assuming Q to be a product distribution we require that our distribution has to satisfy:

$$\mathbb{E}_Q(\mathbf{S}) = \mathbf{m} \tag{6}$$

where $\mathbb{E}$ refers to the average configuration under the Boltzmann measure). Thus, the Gibbs' free energy is defined as

$$G(\mathbf{m}) = \min_Q \{E(Q) - H(Q) \mid \mathbb{E}(\mathbf{S}) = \mathbf{m}\} \tag{7}$$

The constrained optimization problem can be transformed into unconstrained using Lagrange multipliers, i.e.:

$$E(Q) - H(Q) - \sum_i \lambda_i (S_i - m_i) \tag{8}$$

which leads to the new form of the free energy with auxiliary fields:

$$F(\boldsymbol{\lambda}) = \sum_{\mathbf{s}} \exp(-E(\mathbf{s}) - \sum_i \lambda_i s_i) \tag{9}$$

Using this new free energy in the equation 7 yields:

$$G(\mathbf{m}, \boldsymbol{\lambda}) = \sum_i \lambda_i m_i - \ln \sum_{\mathbf{s}} \exp(-E(\mathbf{s}) + \sum_i \lambda_i s_i) \tag{10}$$

Through the Legendre transform we can conjugate both unknown parameters and the condition 6 on the λ_i can be introduced.

$$G(\mathbf{m}) = \max_{\boldsymbol{\lambda}} \{\sum_i \lambda_i m_i - \ln \sum_{\mathbf{s}} \exp(-E(\mathbf{s}) + \sum_i \lambda_i s_i)\} \tag{11}$$

where now the the maximizing auxiliary field $\boldsymbol{\lambda}^*(\mathbf{m})$ is the inverse function of $\mathbf{m}(\boldsymbol{\lambda}) = \frac{\mathrm{d}F}{\mathrm{d}\boldsymbol{\lambda}}$. As we operate on the concave function, the Legendre transform is its own inverse and and we can restore the free energy by setting the auxiliary field to zero which yields:

$$F = \min_{\mathbf{m}} G(\mathbf{m}). \tag{12}$$

Now lets assume inifinite temperature. - expand there Here Georges: Variational free energy with a set of external auxiliary fields (Lagrange multipliers):

$$-\beta G = \ln \sum_{\mathbf{s}} \exp\left(\beta \sum_{(ij)} w_{ij} s_i s_j + \beta \sum_i a_i s_i + \sum_i \lambda_i(\beta)(s_i - m_i)\right)$$

We are interested in expanding $-\beta G(\beta, \mathbf{m})$ around $\beta = 0$ where the spins are entirely controlled by their auxiliary fields:

$$-\beta G = -(\beta G)_{\beta=0} - \left(\frac{\partial(\beta G)}{\partial \beta}\right)_{\beta=0} \beta - \left(\frac{\partial^2(\beta G)}{\partial \beta^2}\right)_{\beta=0} \frac{\beta^2}{2} - ...$$

At $\beta = 0$ the spins are uncorrelated:

$$m_i = \mathbb{E}_{\beta=0}(s_i) = \frac{\exp(\lambda_i(0))}{\exp(\lambda_i(0)) + 1} = \text{sigmoid}(\lambda_i(0)) \tag{13}$$

and

$$\begin{aligned}
-(\beta G)_{\beta=0} &= \ln \sum_{\mathbf{s}} \exp\left(\sum_i \lambda_i(0)(s_i - m_i)\right) \\
&= \ln \left\{\sum_{s_1} \exp\left(\lambda_1(0)(s_1 - m_1)\right) ... \sum_{s_n} \exp\left(\lambda_n(0)(s_n - m_n)\right)\right\} \\
&= \ln \left\{(\exp(\lambda_i(0)) + 1) \exp(-\lambda_1(0) m_1) ... (\exp(\lambda_i(0)) + 1) \exp(-\lambda_1(0) m_n)\right\} \\
&= \sum_i \left(\ln\left[1 + \left(\frac{m_i}{1 - m_i}\right)\right] - \lambda_i(0) m_i\right)
\end{aligned} \tag{14}$$

Using 13, we obtain:

$$\lambda_i(0) = \text{logit}(m_i) = \ln\left(\frac{m_i}{1-m_i}\right)$$

and thus:

$$\begin{aligned}-(\beta G)_{\beta=0} &= \sum_i \left\{\ln\left(\frac{1}{1-m_i}\right) - m_i \ln\left(\frac{m_i}{1-m_i}\right)\right\} \\ &= -\sum_i \left[m_i \ln(m_i) + (1-m_i)\ln(1-m_i)\right]\end{aligned}$$

Yedida and Georges showed how to create the Taylor expansion beyond $O(\beta^2)$ (TODO referring to MF and TAP variant) (derivation in Appendix):

$$\begin{aligned}-\beta G = &-\sum_i \left[m_i \ln(m_i) + (1-m_i)\ln(1-m_i)\right] \\ &+ \beta \sum_{(ij)} w_{ij} m_i m_j + \beta \sum_i a_i m_i \\ &+ \frac{\beta^2}{2} \sum_{(ij)} w_{ij}^2 (m_i - m_i^2)(m_j - m_j^2) \\ &+ \frac{2\beta^3}{3} \sum_{(ij)} w_{ij}^3 (m_i - m_i^2)(\frac{1}{2} - m_i)(m_j - m_j^2)(\frac{1}{2} - m_j)\end{aligned}$$

TODO:Add self consistency relations:

The variational free energy $\mathcal{F}$ is highly multimodal in real world examples in which case all stationary points either maximum, minimum or saddle satisfies the self-consistency relations **??**.

2.6. Marginal polytope - bounds for the partition function

Although it is very straightforward to obtain naive mean field approach from the TAP expansions, unlike the former this method doesn't bound in any way the variational free energy. Moreover, the approximation was based on the Taylor expansion which poses a threat that the radius of convergence of the expansion will be too small to obtain robust results for the different values of β [**yedidia2001idiosyncratic**]. There are a few examples in statistical physics where this method works very reliably in a wide variety of temperatures [**plefka1982convergence**] however in general there aren't any theoretical foundations for the robustness of the TAP expansion.

2.7. Boltzmann Machine

A Boltzmann machine is a network of symmetrically coupled stochastic binary units with both visible-to-visible and hidden-to-hidden lateral connections with energy function defined as: [**ackley1985learning**]

2.7.1. Approximator for any distribution

An example of such structure presents Figure **??**.

2.8. RBMs

Representational power - bengio -any distibution can be modelled by RBM. TODO.

$$E(\mathbf{v}, \mathbf{h}) = -\sum_i a_i v_i - \sum_j b_j h_j - \sum_{i,j} v_i w_{ij} h_j$$

$$P(\mathbf{v}, \mathbf{h}) = \frac{1}{Z} e^{-E(\mathbf{v}, \mathbf{h})}$$

$$Z = \sum_{\mathbf{v},\mathbf{h}} e^{-E(\mathbf{v},\mathbf{h})}$$

$$F = -\ln Z$$

$$F^c(\mathbf{v}) = \ln\left(\sum_{\mathbf{h}} e^{-E(\mathbf{v},\mathbf{h})}\right) = \ln\left(\sum_{h_1} e^{-E(\mathbf{v},h_1)} \dots \sum_{h_n} e^{-E(\mathbf{v},h_n)}\right)$$

$$\mathcal{L} = \ln P(\mathbf{v}) = F^c(\mathbf{v}) - F$$

where F is the *free energy* of the RBM.

2.8.1. Exploiting the RBM structure

The restrictions imposed on the structure allows for efficient computation of the clamped free energy because the hidden variables are independent given the state of the visible variables and vice versa and we can write:

$$\begin{aligned} p(\mathbf{h}|\mathbf{v}) &= \prod_{i=1}^{m} p(h_i|\mathbf{v}) \\ p(\mathbf{v}|\mathbf{h}) &= \prod_{i=1}^{n} p(v_i|\mathbf{h}) \end{aligned} \tag{15}$$

Clamped free energy can be written in the form:

$$\begin{aligned} \mathcal{F}^c(\mathbf{v}) =& \sum_{\mathbf{h}} e^{-E(\mathbf{v},\mathbf{h})} = e^{\mathbf{b}'\mathbf{v}} \sum_{h_1} \dots \sum_{h_m} e^{-E(\mathbf{v},\mathbf{h})} \\ =& e^{\mathbf{b}'\mathbf{v}} \sum_{h_1} e^{h_1(c_1 + W_{1\bullet}\mathbf{v})} \dots \sum_{h_m} e^{h_m(c_m + W_{m\bullet}\mathbf{v})} \\ =& e^{\mathbf{b}'\mathbf{v}} \prod_{j=1}^{m} \left(1 + e^{c_i + W_{i\bullet}\mathbf{v}}\right) \end{aligned} \tag{16}$$

Appendix

Following [**georges1991expand**] lets define energy as:

$$E = -\sum_{ij} w_{ij} s_i s_j - \sum_i a_i s_i \tag{17}$$

and introduce the following operator:

$$U \equiv E - \mathbb{E}(E) - \sum_i \frac{\partial \lambda_i(\beta)}{\partial \beta}(s_i - m_i) \tag{18}$$

which poses useful property:

$$\mathbb{E}(U) = 0$$

For any other operator O we then have:

$$\frac{\partial \mathbb{E}(O)}{\partial \beta} = \mathbb{E}\left(\frac{\partial O}{\partial \beta}\right) - \mathbb{E}(OU) \tag{19}$$

Now, the first derivative from the Taylor expansion is:

$$\begin{aligned} \frac{\partial(\beta G)}{\partial \beta} =& \frac{\sum_{\mathbf{s}} \exp\left(\beta \sum_{(ij)} w_{ij} s_i s_j + \beta \sum_i a_i s_i + \sum_i \lambda_i(\beta)(s_i - m_i)\right)\left(\sum_{(ij)} w_{ij} s_i s_j + \sum_i a_i s_i + \sum_i \frac{\partial \lambda_i(\beta)}{\partial \beta}(s_i - m_i)\right)}{\sum_{\mathbf{s}} \exp\left(\beta \sum_{(ij)} w_{ij} s_i s_j + \sum_i a_i s_i + \sum_i \lambda_i(\beta)(s_i - m_i)\right)} \\ =& \sum_{(ij)} w_{ij} \mathbb{E}(s_i s_j) + \sum_i a_i \mathbb{E}(s_i) + \frac{\partial \lambda_i(\beta)}{\partial \beta} \sum_i \mathbb{E}(s_i - m_i) \end{aligned}$$

In the case of $\beta = 0$ we have:

$$\left.\frac{\partial(\beta G)}{\partial \beta}\right|_{\beta=0} = \sum_{(ij)} w_{ij} m_i m_j + \sum_i a_i m_i$$

Using 19 we obtain: A.1.4:

$$\frac{\partial m_i}{\partial \beta} = 0 = \frac{\partial \mathbb{E}(s_i)}{\partial \beta} = \mathbb{E}\left(\frac{\partial s_i}{\partial \beta}\right) - \mathbb{E}(s_i U) = -\mathbb{E}(s_i U) = -\mathbb{E}(U(s_i - m_i)) \tag{20}$$

Thus A.1.5:

$$\begin{aligned} \frac{\partial U}{\partial \beta} = & \frac{\partial H}{\partial \beta} - \frac{\partial \mathbb{E}(H)}{\partial \beta} - \sum_i \frac{\partial^2 \lambda_i(\beta)}{\partial \beta^2}(s_i - m_i) \\ = & \mathbb{E}(U^2) - \sum_i \frac{\partial^2 \lambda_i(\beta)}{\partial \beta^2}(s_i - m_i) \end{aligned} \tag{21}$$

The second derivative has the form A.1.6:

$$\begin{aligned} \frac{\partial^2 U}{\partial \beta^2} = & 2\mathbb{E}\left(\frac{\partial U}{\partial \beta} U\right) - \mathbb{E}(U^3) - \sum_i \frac{\partial^3 \lambda_i(\beta)}{\partial \beta^3}(s_i - m_i) \\ = & -\mathbb{E}(U^3) - \sum_i \frac{\partial^3 \lambda_i(\beta)}{\partial \beta^3}(s_i - m_i) \end{aligned} \tag{22}$$

Coming back to our expansion of free energy using formulas derived above we now can calculate A.1.7:

$$\frac{\partial(\beta G)}{\partial \beta} = \mathbb{E}(E) - \sum_i \frac{\partial \lambda_i(\beta)}{\partial \beta} \mathbb{E}(s_i - m_i) = \mathbb{E}(E) \tag{23}$$

and A.1.8:

$$\frac{\partial^2(\beta G)}{\partial \beta^2} = \mathbb{E}\left(\frac{\partial E}{\partial \beta}\right) - \mathbb{E}(EU) = -\mathbb{E}(U^2)$$

$$\frac{\partial^3(\beta G)}{\partial \beta^3} = -2\mathbb{E}\left(U \frac{\partial U}{\partial \beta}\right) + \mathbb{E}(U^3) = \mathbb{E}(U^3) \tag{24}$$

$$\frac{\partial^4(\beta G)}{\partial \beta^4} = 3\mathbb{E}\left(U^2 \frac{\partial U}{\partial \beta}\right) - \mathbb{E}(U^4) = 3\left(\mathbb{E}(U^2)\right)^2 - 3\sum_i \frac{\partial^2 \lambda_i(\beta)}{\partial \beta^2} \mathbb{E}(U^2(s_i - m_i)) - \mathbb{E}(U^4)$$

TS was considered around point $\beta = 0$. Using derivations from above we obtain again a 'naive' term:

$$\left.\frac{\partial(\beta G)}{\partial \beta}\right|_{\beta=0} = \mathbb{E}_{\beta=0}(E) = -\sum_{(ij)} w_{ij} m_i m_j - \sum_i a_i m_i = -\frac{1}{2}\sum_i \sum_j w_{ij} m_i m_j - \sum_i a_i m_i \tag{25}$$

Consider now

$$\left.\frac{\partial(\beta G)}{\partial m_i \partial \beta}\right|_{\beta=0} = -\sum_{j \neq i} w_{ij} m_j - a_i \tag{26}$$

On the other hand:

$$\left.\frac{\partial(\beta G)}{\partial m_i \partial \beta}\right|_{\beta=0} = \left.\frac{\partial(\beta G)}{\partial \beta \partial m_i}\right|_{\beta=0} = \frac{\partial}{\partial \beta}\mathbb{E}(\lambda_i(\beta)) = \left.\frac{\partial \lambda_i(\beta)}{\partial \beta}\right|_{\beta=0} \tag{27}$$

Substituting 27 into 18 gives us A.12:

$$\begin{aligned} U_{\beta=0} = & -\sum_{(ij)} w_{ij} s_i s_j - \sum_i a_i s_i + \frac{1}{2}\sum_i \sum_j w_{ij} m_i m_j + \sum_i a_i m_i + \sum_i \left(\sum_{j \neq i} w_{ij} m_j + a_i\right)(s_i - m_i) \\ = & -\sum_{(ij)} w_{ij} s_i s_j - \frac{1}{2}\sum_{(ij)} w_{ij} m_i m_j + \sum_i \sum_{j \neq i} w_{ij} s_i m_j \\ = & -\sum_{(ij)} w_{ij}(s_i - m_i)(s_j - m_j) = -\sum_l w_l y_l \end{aligned} \tag{28}$$

where y_l stands for the 'link' operator $w_l = w_{ij}$ and $y_l = (s_i - m_i)(s_j - m_j)$ which poses useful properties:

$$\begin{aligned}
\mathbb{E}(y_l)_{\beta=0} &= \mathbb{E}(s_i s_j) - m_j \mathbb{E}(s_i) - m_i \mathbb{E}(s_j) + m_i m_j = 0 \\
\mathbb{E}(y_l(s_i - m_i))_{\beta=0} &= m_j - m_j - m_i^2 m_j + m_i^2 m_j \\
&\quad - m_i^2 m_j + m_i^2 m_j + m_i^2 m_j - m_i^2 m_j \\
&= 0
\end{aligned} \tag{29}$$

Finally, if $k \neq l$ then:

$$\mathbb{E}(y_k y_l) = \mathbb{E}(y_k)\mathbb{E}(y_l) = 0$$

while for $k = l$ we have:

$$\begin{aligned}
\mathbb{E}((s_i - m_1)^2 (s_j - m_j)^2) &= m_i m_j - 2m_i m_j^2 + m_i m_j^2 - 2m_i^2 m_j + 4m_1^2 m_j^2 \\
&\quad - 2m_i^2 m_j^2 + m_i^2 m_j - 2m_i^2 m_j^2 + m_i^2 m_j^2 \\
&= (m_i - m_i^2)(m_j - m_j^2)
\end{aligned} \tag{30}$$

Using properties from 30 in equations 24 we can derive **??** A1.13:

$$\begin{aligned}
\left.\frac{\partial^2(\beta G)}{\partial \beta^2}\right|_{\beta=0} &= -\mathbb{E}(U^2)_{\beta=0} \\
&= -\sum_{l_i l_2} w_{l_i} w_{l_2} \mathbb{E}_{\beta=0}(y_{l_1} y_{l_2}) \\
&= -\sum_{(i,j)} w_{ij}^2 (m_i - m_i^2)(m_j - m_j^2)
\end{aligned}$$

which yields the TAP-Onsager term.

To obtain the next term for the Taylor expansion we need to compute $\mathbb{E}(y_{l_1} y_{l_2} y_{l_3})$ term and by definition the structure of the RBM model doesn't admit triangles in its corresponding factor graphs. Thus, we need to consider on the case when $l_1 = l_2 = l_3$:

$$\begin{aligned}
\mathbb{E}((s_i - m_1)^3 (s_j - m_j)^3) &= m_i m_j - 3m_i m_j^2 + 2m_i m_j^2 + 2m_i m_j^3 - 3m_i^2 m_j + 2m_i^3 m_j \\
&\quad + 9m_i^2 m_j^2 - 6m_i^3 m_j^2 - 6m_i^2 m_j^3 + 4m_i^3 m_j^3 \\
&= 4(m_i - m_i^2)(\frac{1}{2} - m_i)(m_j - m_j^2)(\frac{1}{2} - m_j)
\end{aligned} \tag{31}$$

$$\left.\frac{\partial^3(\beta G)}{\partial \beta^3}\right|_{\beta=0} = \frac{2\beta^3}{3} \sum_{(ij)} w_{ij}^3 (m_i - m_i^2)(\frac{1}{2} - m_i)(m_j - m_j^2)(\frac{1}{2} - m_j)$$

3. Evaluation on the toy models

So far we have considered general graphical model where pair-wise connections might be defined between all the nodes. However, we are interested in the adaptation of the EMF to the restricted Boltzmann machine.

3.1. Adaptation of TAP to RBM

To obtain the division between visible and hidden layers lets define visible and hidden magnetizations denoted by $\mathbf{m^v}$ and $\mathbf{m^h}$ respectively. The energy in the BM model 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:

$$\begin{aligned} G(\mathbf{m^v}, \mathbf{m^h}) \simeq & H(\mathbf{m^v}, \mathbf{m^h}) \\ & - \sum_i a_i m_i^v - \sum_j b_j m_j^h \\ & - \sum_{i,j} \left(m_i^v w_{ij} m_j^h + \frac{w_{ij}^2}{2} (m_i^v - (m_i^v)^2)(m_j^h - (m_j^h)^2) \right) \\ & - \sum_{i,j} \left(\frac{2w_{ij}^3}{3} (m_i^v - (m_i^v)^2)(\frac{1}{2} - m_i^v)(m_j^h - (m_j^h)^2)(\frac{1}{2} - m_j^h) \right) \end{aligned} \tag{32}$$

In the case of the RBM, the third term consists of the sum of pair connection because the coupled triplets are excluded by the bipartite structure of the RBM [**gabrie2015training**]. 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:

$$\frac{\partial G}{\partial m_i^v} = \frac{m_i^v}{m_i^v} + \ln m_i - \frac{1 - m_i^v}{1 - m_i^v} - \ln(1 - m_i^v) - m = 0 \tag{33}$$

This can be regrouped as:

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

The

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

Similar condition can be obtained for $\mathbf{m^h}$. 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. This creates a question how to efficiently define a schedules of update of magnetizations that will eventually satisfy these constraints which will allow us to compute extended mean field approximation for the partition function **??**.

3.2. 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 [**welling2002new**]. Similarly, in the case of the extended mean field approximation, it was proposed that an iterative, asynchronous algorithm may serve as update rules [**gabrie2015training**] following positive theoretical results proved in the context of random spin glass. However, there are many heuristically reasonable ways perform such sequential updates as well as it is interesting how different procedures might affect the convergence. Thus, I will analyse three different updates rule for magnetizations on a toy model and on the real life data set example. The updates here are considered only up to the third order.

3.2.1. Asynchronous

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

$$\begin{aligned} \mathbf{m^h}[t+1] &= \mathrm{sigm}\left[\mathbf{b} + W\mathbf{m^v}[t] - \left(\mathbf{m^h}[t] - \frac{1}{2} \right)^T \odot W^2 \left(\mathbf{m^v}[t] - (\mathbf{m^v}[t])^2 \right) \right] \\ \mathbf{m^v}[t+1] &= \mathrm{sigm}\left[\mathbf{a} + W^T\mathbf{m^h}[t+1] - \left(\mathbf{m^v}[t] - \frac{1}{2} \right) \odot (W^2)^T \left(\mathbf{m^h}[t+1] - (\mathbf{m^h}[t+1])^2 \right) \right], \end{aligned} \tag{34}$$

where $\odot$ denotes Hadamard product.

3.2.2. Sequential

Previous procedure takes advantage of the bipartite structure of the model. However, we might consider updates not in the vectorize way but rather by sequential updates. In general number of hidden units differ from the visible ones and we can sequentially update either hidden or visible magnetizations. Here the procedure is sequential for the hidden layer:

$$\begin{aligned} m_i^h &= \text{sigm}\left[b_i + \sum_j \left(w_{ij} m_j^v - w_{ij}^2 (m_i^h - \frac{1}{2})\left(m_j^v - (m_j^v)^2\right)\right)\right] \\ \mathbf{m^v} &= \text{sigm}\left[\mathbf{a} + W^T \hat{\mathbf{m}}^{\mathbf{h}} - \left(\mathbf{m^v} - \frac{1}{2}\right) \odot (W^2)^T \left(\hat{\mathbf{m}}^{\mathbf{h}} - (\hat{\mathbf{m}}^{\mathbf{h}})^2\right)\right] \end{aligned} \tag{35}$$

where $i \in \{1, ..., \#\text{ of hidden nodes}\}$ and $\hat{\mathbf{m}}^{\mathbf{h}}$ is a magnetization vector with i-th value beign updated. This implies imbalance in numbers of updates performed between hidden and visible layers.

3.2.3. Parallel

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

$$\begin{aligned} \mathbf{m^h}[t+1] &= \text{sigm}\left[\mathbf{b} + W\mathbf{m^v}[t] - \left(\mathbf{m^h}[t] - \frac{1}{2}\right)^T \odot W^2 \left(\mathbf{m^v}[t] - (\mathbf{m^v}[t])^2\right)\right] \\ \mathbf{m^v}[t+1] &= \text{sigm}\left[\mathbf{a} + W^T\mathbf{m^h}[t] - \left(\mathbf{m^v}[t] - \frac{1}{2}\right) \odot (W^2)^T \left(\mathbf{m^h}[t] - (\mathbf{m^h}[t])^2\right)\right], \end{aligned} \tag{36}$$

This schedule of updates pose a risk that the model might not learn the proper transfer of information from one layer to another as the RBM implies and is designed for.

Figure 3.2.3 presents graphically all proposed procedures. In the case of a fixed point algorithms, it is a common

Figure 1: rbm model - same couplings

practice to use damped updates [**murphy2012machine**] of the form:

$$m_i^t = (1 - \lambda) m_i^{t-1} + \lambda(\text{update}),$$

for $0 < \lambda < 1$. This helps in avoiding unnecessary artefacts and oscillations. In all experiments conducted in this and the following chapters, the updates will be damped with λ set to 0.5.

3.3. 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 our case, to specialize to the RBM model we set β to 1 which means that the temperature is also 1 while the approximation was derived with infinite temperature. This suggests Those two facts

3.3.1. Grid toy model

The analysis will be made assuming that the parameters of the model are known. A small grid toy model was considered of size 4×4 with periodic boundary conditions in order to avoid edge effects – Figure 3.3.1 (left) shows this model from graphical models perspective. The nature of the models implies that the asynchronous updates of magnetizations seems as the most natural way to obtain a statistics of the system in the equilibrium. In this case each magnetization m_i is updated one at a time using equation **??**.

Figure 2: Grid toy model used for an exact inference.

Initially, the external field was set to 0 and considered the case where 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 3.3.2 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 model the connections in the system between the spins.

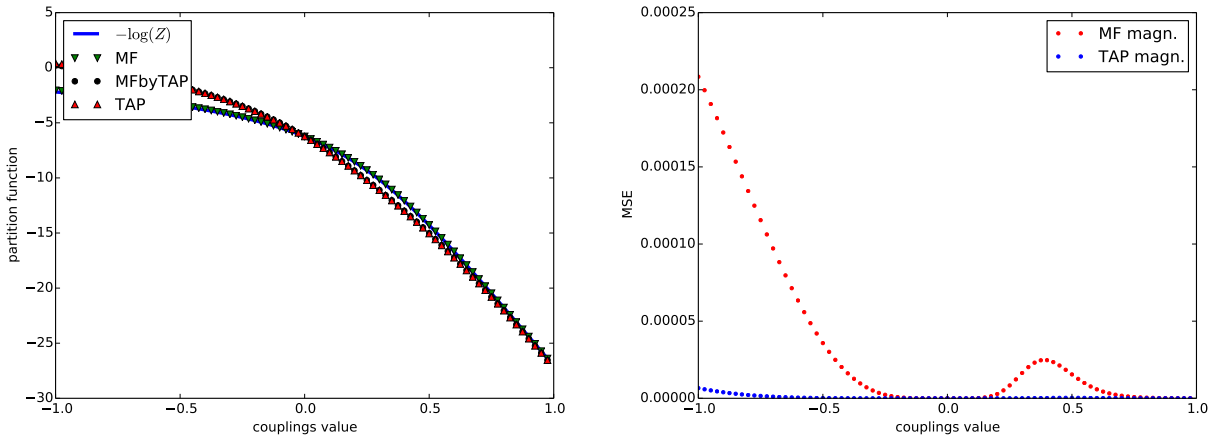

Figure 3: grid model - same couplings

In the next experiment, all couplings were initialised to random values around "mean" strength which varies from 0 to 1 and randomly assigned with $-$ sign. The results are similar to the one observed previously - Figure 3.3.1. The naive approach gives consistently better approximation for the $-\ln Z$ while the TAP method performs better in the case of estimating average value of spin.

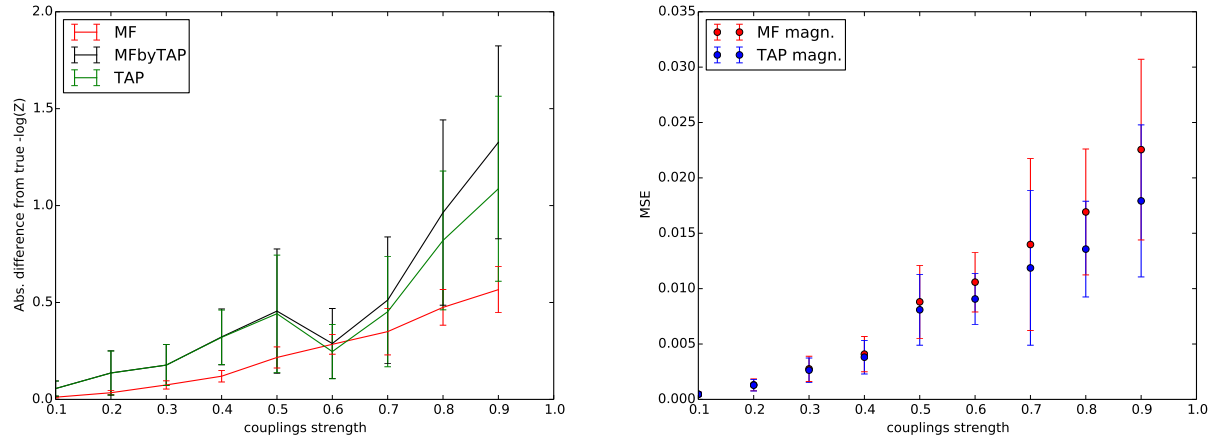

Figure 4: grid model - different couplings

TODO: add third way of noise? -1 +1 TODO: external fields.

3.3.2. RBM toy model

As the – As we will observe, due to the different structure of connections between states, Unlike in the previous case, there is no strong heuristics how the updates of self-consistency relations should be performed. The

literature suggests that the updates in the case of the naive approach it is necessary to run self-constiency equations sequentialy [**welling2002new**]. That is why, I considered three ways in which we can update the magnetizations. (sequential), (parallel) . The last way of performing updates follows from the application of [**bolthausen2014iterative**] (default)

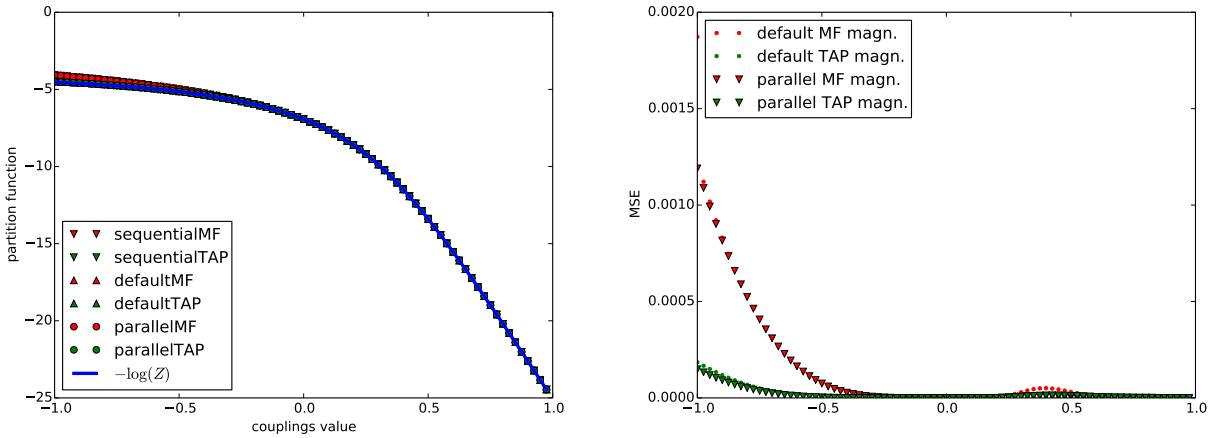

Figure 5: rbm model - same couplings

The results presented in Figure 3.3.2 are similar to the ones obtained with.

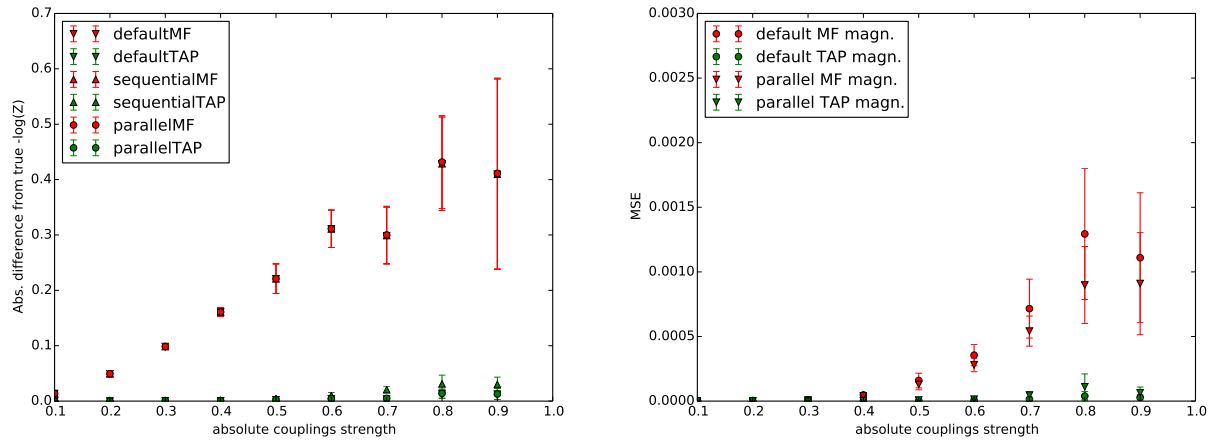

Figure 6: rbm model - different couplings

The results - CLT - why? TODO - might be caused

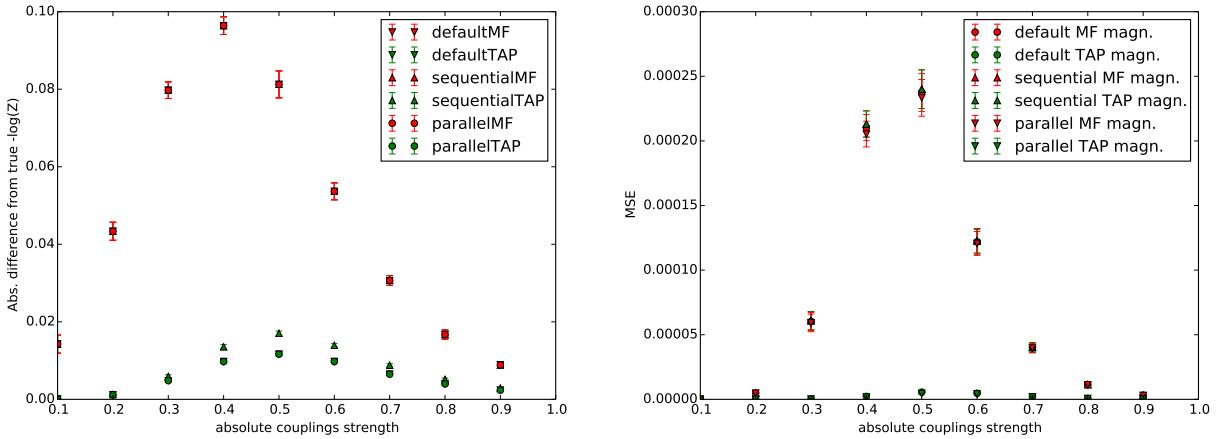

Figure 7: rbm model - different signs

4. Learning of Boltzmann Machines

4.1. Usupervised learning

The results of experiments on toy models suggest that the initial unsatisfactory results with naive mean field approaches [**tieleman2008training**] might be greatly improved if add additional terms responsible for better estimation of (connections) between the spins.

[**dayan1999unsupervised**]

Our general goal is to maximize the probability of $\mathcal{D}$ under the MRF distributions – thus we are looking for the vector of parameters θ that maximize the likelihood given the training data, i.e.

$$\max_{\theta} \ln \mathcal{L}(\theta|\mathcal{D}) = \max_{\theta} \ln \prod_{i=1}^{N} p(\mathbf{v}_i|\theta) = \max_{\theta} \sum_{i=1}^{N} \ln p(\mathbf{v}_i|\theta) \tag{37}$$

In most of the cases, it is not possible to find the analytical solution for the maximum likelihood parameters and we need to resort to some approximation methods.

4.2. Unsupervised Pre-training of Neural Networks

add Erham here

4.3. Training of Boltzmann Machines

As it

$$\theta^{t+1} = \theta^t + \eta \frac{\partial}{\partial \theta^t} \ln \mathcal{L}(\theta|\mathcal{D}) \tag{38}$$

This relies on the fact that the gradient w.r.t. parameters θ informs us how fast function increases in the current point θ^t. By taking appropriately small learning rate, these iterative updates might converge to the maximum of the function. However, there is no guarantees that this procedure will lead to obtaining global maximum.

TODO - stochastic gradient descent - theoretical results - writeabout it and read theoeretical results.

Learning Restricted Boltzmann machines relies on gradient ascent of the log-likelihood. The gradient of the log-likelihood from given a training example $\mathbf{v}$ takes the form:

$$\begin{aligned}\frac{\partial \log \mathcal{L}(\theta|\mathbf{v})}{\partial \theta} &= \frac{\partial \mathcal{F}^c}{\partial \theta} - \frac{\partial \mathcal{F}}{\partial \theta} \\ &= -\frac{\sum_{\mathbf{h}} e^{-E(\mathbf{v},\mathbf{h})} \frac{\partial E(\mathbf{v},\mathbf{h})}{\partial \theta}}{\sum_{\mathbf{h}} e^{-E(\mathbf{v},\mathbf{h})}} + \frac{\sum_{\mathbf{v},\mathbf{h}} e^{-E(\mathbf{v},\mathbf{h})} \frac{\partial E(\mathbf{v},\mathbf{h})}{\partial \theta}}{\sum_{\mathbf{v},\mathbf{h}} e^{-E(\mathbf{v},\mathbf{h})}} \\ &= -\sum_{\mathbf{h}} p(\mathbf{h}|\mathbf{v}) \frac{\partial E(\mathbf{v},\mathbf{h})}{\partial \theta} + \sum_{\mathbf{v},\mathbf{h}} p(\mathbf{v},\mathbf{h}) \frac{\partial E(\mathbf{v},\mathbf{h})}{\partial \theta} \\ &= -\mathbb{E}_{p(\mathbf{h}|\mathbf{v})}\left(\frac{\partial E(\mathbf{v},\mathbf{h})}{\partial \theta}\right) + \mathbb{E}_{p(\mathbf{v},\mathbf{h})}\left(\frac{\partial E(\mathbf{v},\mathbf{h})}{\partial \theta}\right)\end{aligned} \tag{39}$$

As we can see the gradient is the difference of two expectations – the expected value of the gradient of the energy function under the model distribution and under the conditional distribution of the hidden variables given the observed variables $\mathbf{v}$. Thanks to the restriction imposed on the structure of the BM, the first term can be computed analytically 51. However, as it was mentioned previously, direct calculations of the second term leads to the complexity that is exponential in the number of variables in the model.

4.4. MCMC Sampling

The

4.5. Contrastive Divergence

4.5.1. Persistent CD

4.6. Learning in the TAP case

4.6.1. Gradients

Eq. 11. Gradients:

$$w_{ij}^{t+1} = w_{ij}^t + \eta \triangle w_{ij}^{t+1}$$

$$\triangle w_{ij}^{t+1} \propto \frac{\partial \mathcal{L}}{\partial w_{ij}} \simeq -\frac{\partial F}{\partial w_{ij}} - \frac{\partial F^{EMF}}{\partial w_{ij}}$$

$$\begin{aligned}\frac{\partial F^{EMF}}{\partial w_{ij}} =& - m_i^v m_j^h - w_{ij}^t (m_i^v - (m_i^v)^2)(m_j^h - (m_j^h)^2) \\ & - 2 w_{ij}^2 (m_i^v - (m_i^v)^2)(\frac{1}{2} - m_i^v)(m_j^h - (m_j^h)^2)(\frac{1}{2} - m_j^h)\end{aligned}$$

$$\frac{\partial \mathcal{L}}{\partial a_i} = \frac{\partial F^{EMF}}{\partial a_i} = -m_i^v$$

$$\frac{\partial \mathcal{L}}{\partial b_j} = \frac{\partial F^{EMF}}{\partial b_j} = -m_j^h$$

An example of such structure presents Figure **??**.

4.7. Approximating the likelihood

TODO - describe problems and procedure.

4.7.1. Pseudo approximation

The problems mentioned above makes training such structure very difficult because we cannot observe directly progress along learning. Thus, we need to resort to some approximations. One of the most popular approaches is due to Besag [**besag1972nearest**]. Consider the following approximation:

$$P(\mathbf{s}) = \prod_i p(s_i|s_1, ..., s_{i-1}) \approx \prod_i p(s_i|s_1, ..., s_{i-1}, s_{i+1}, ..., s_n) = \prod_i p(s_i|s_{-i}) \tag{40}$$

where the first equation comes from the chain rule. Here we assume that particular marginals given all other dimensions are independent of each other.

Theorem 2 *TODO: Assume that* $\mathbf{x}$ *is generated I.ID. by a distribution* $p(\mathbf{x}; \theta)$.

If the analysed phenomena has many dimensions this approximation is still computationally expensive. Thus, another step is to choose only one marginal as a proxy, i.e.

$$\log PL(\mathbf{s}) = N\mathbb{E}\left(\log P(s_i|\mathbf{s}_{-i})\right), \tag{41}$$

where $i \sim U(1, N)$.

In the case of the analysed model we obtain the following form using Monte Carlo approximation (TODO: monte carlo approximation):

$$\log PL(\mathbf{s}) \approx N \log\left(\frac{\exp\{-F(\mathbf{s})\}}{\exp\{-F(\hat{\mathbf{s}})\} + \exp\{-F(\mathbf{s})\}}\right), \tag{42}$$

where $\hat{\mathbf{s}}$ represents the vector $\mathbf{s}$ with i-th variable of flipped, i.e. $1 - s_i$.

4.8. Real scale model

4.8.1. MNIST data set

4.9. Comparison

In order to test the efficiency of the EMF learning algorithm I used three approximations of **??** – at first-order (MF), second-order (TAP2) and third order (TAP3).

All trained models used the same set-up of free parameters. The purpose of this experiment is to compare different RBM trainings thus following [**gabrie2015training**] I didn't use the adaptive learning rate which was set to 0.005, learning was performed using mini-batch learning with 100 training points per batch. The couplings matrix was randomly initialised using normal distribution with zero mean and variance set to 0.01. This allows to compare the procedures in the their "raw" forms.

However, the EMF approximation was performed around the infinite temperature were the spins are independent. This means that the couplings should have small values – this can be enforced using regularization which at the same times allows for a better regularization. From probabilistic perspective it can be seen as adding a weighted prior over the parameters (maximum a posteriori). This leads to a new criterion which will be maximized of the form:

$$E(\theta, \mathcal{D}) = \ln \mathcal{L}(\theta|\mathcal{D}) - \lambda R(\theta). \tag{43}$$

In all experiments I used Laplacian prior $R(\theta) = \|\theta\|_1$ ($L1$ regularization) with the weight λ set to 0.01.

Figure **??** presents COmment Figure **??** shows comment

Figure 8: Pseudo LL.

Figure 9: EMF LL.

5. Chapter 4 - Applications

5.1. Updates

In the chapter 3 different schedules of updates were analysed on the toy model where the parameters of the model were known a priori and no substantial discrepancies were observed in terms of the quality of approximation. However, in the case of the `MNIST` data set estimated magnetizations allow us to perform learning of unknown parameters. Thus, in this case we combine uncertainty related to both magnetizations and parameters – this may lead to substantial differences in performance. Figure

Figure 10: updatesTAP2

Figure 11: updatesMF

Figure 12: updatesTAP3

5.2. AIS vs TAP

Make several runs of AIS to get reasonable approximation!

5.2.1. Annealed Importance Sampling (AIS)

The most widely used technique is based on a very simple identity. Assume we have two distributions $p_A = \frac{1}{Z_A} p_A^*(\mathbf{x})$, $p_B = \frac{1}{Z_B} p_B^*(\mathbf{x})$ where $p^*()$ denotes unnormalized distribution and Z_A, Z_B are partition functions. Assuming that a proposal p_A distribution p_A supports tractable sampling and tractable evaluation of both the unnormalized distribution $p_A^*(\mathbf{x})$ and the partition function Z_A we can use the following relation:

$$\begin{aligned} Z_B &= \int p_B^*(\mathbf{x}) \mathrm{d}\mathbf{x} \\ &= \int \frac{p_A(\mathbf{x})}{p_A(\mathbf{x})} p_B^*(\mathbf{x}) \mathrm{d}\mathbf{x} \\ &= Z_A \int \frac{p_B^*(\mathbf{x})}{p_A^*(\mathbf{x})} p_A(\mathbf{x}) \mathrm{d}\mathbf{x} \end{aligned} \tag{44}$$

Sampling from the tractable distribution, we can derive Monte Carlo estimator of the ratio between partition functions:

$$\frac{Z_B}{Z_A} \approx \frac{1}{N} \sum_{i=1}^{N} \frac{p_B^*(\mathbf{x}^{(i)})}{p_A^*(\mathbf{x}^{(i)})} = \hat{r}_{SIS} \tag{45}$$

where $\mathbf{x}^{(i)}$ comes from p_A. Assuming that distribution p_A is close to p_B, the estimator from 41 called simple importance sampling proves to work well [**minka2005divergence**]. However, in high-dimensional spaces where p_B is usually multimodal as it is considered in this thesis, the variance of the estimator from 41 might be very high.

The idea presented above might be improved by following the classic approach from probabilistic optimization i.e. simulated annealing. The idea is to introduce intermediate distributions that will allow to bridge the gap between two considered distributions p_A and p_B [**jarzynski1997nonequilibrium**], [**neal2001annealed**].

Consider a sequence of distributions $p_0, p_1, ..., p_M$ where $p_0 = p_A$ and $p_M = p_B$. If the intermediate distributions p_m and p_{m+1} are close enough, a simple estimator from 41 can be used to estimate each ratio $\frac{Z_{m+1}}{Z_m}$. Using the the following identity:

$$\frac{Z_M}{Z_0} = \frac{Z_1}{Z_0} \frac{Z_2}{Z_1} \cdots \frac{Z_M}{Z_{M-1}} \tag{46}$$

those intermediate ratios are then combined to obtain the estimate of $\frac{Z_B}{Z_A}$. There is no need to compute the normalizing constants of any intermediate distributions. The intermediate distributions are chosen to suit a given problem domain. However in most cases, we are able to draw exact samples only from the first tractable distribution p_A. In order to sample from intermediate distribution we have be able to draw a sample $\mathbf{x}^{'}$ given $\mathbf{x}$ using Markov chain transition operator $T_m(\mathbf{x}^{'}|\mathbf{x})$ that leaves $p_m(\mathbf{x})$ invariant, i.e.:

$$\int T_m(\mathbf{x}^{'}|\mathbf{x})p_m(\mathbf{x})\mathrm{d}\mathbf{x} = p_m(\mathbf{x}^{'}) \tag{47}$$

These transition operators represent the probability density of transitioning from state $\mathbf{x}$ to $\mathbf{x}^{'}$ [**salakhutdinov2008learning**]. Having obtained the sequence of samples from the intermediate distributions we can obtain the improved estimator of the ratio between partition functions following the procedure:

Algorithm 1 Annealed Importance Sampling.

Set p_A and p_B with appropriate parameters
for $i \in \{1, ..., N\}$ **do**
  sample $\mathbf{x}_1$ from $p_0 = p_A$
  sample $\mathbf{x}_2$ via $T_1(\mathbf{x}_2|\mathbf{x}_1)$
  ...
  sample $\mathbf{x}_M$ via $T_M(\mathbf{x}_M|\mathbf{x}_{M-1})$
  $r_{AIS}^{(i)} = \frac{p_1^*(\mathbf{x}_1)}{p_0^*(\mathbf{x}_1)} \frac{p_2^*(\mathbf{x}_2)}{p_1^*(\mathbf{x}_2)} ... \frac{p_M^*(\mathbf{x}_M)}{p_{M-1}^*(\mathbf{x}_M)}$
end for
$\hat{r}_{AIS} = \frac{1}{N} \sum_{i=1}^{N} \hat{r}_{AIS}^{(i)}$

It was proven that the variance of $\hat{r}_{AIS}$ will be proportional to $1/MN$ assuming we used sufficiently large numbers of intermediate distributions M [**neal2001annealed**]. Moreover, the estimate of Z_M/Z_0 will be unbiased if each ratio is estimated using $N = 1$ and a sample $\mathbf{x}^m$ is obtained using Markov chain starting at previous sample. This follows from the observation that the AIS procedure is an simple importance sampling defined on an extended state space $\mathcal{X} = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_M)$.

The procedure described above can be adapted to the RBM case – assume that we have estimated parameters θ_B of the model that we want to evaluate. Following [**salakhutdinov2008quantitative**] as a tractable starting distribution p_A we can use "clamped" restricted Boltzmann machine where there is no hidden layer. The sequence of intermediate distribution is then defined as:

$$p_m(\mathbf{v}) = \frac{1}{Z_m} p_m^*(\mathbf{v}) = \frac{1}{Z_m} \sum_{\mathbf{h}} \exp(-E_m(\mathbf{v}, \mathbf{h})) \tag{48}$$

where $m = 0, ..., M$, $\mathbf{h} = \mathbf{h_B}$, and the energy function has the form:

$$E_m(\mathbf{v}, \mathbf{h}) = (1 - \beta_m)E(\mathbf{v}; \theta_A) + \beta_m E(\mathbf{v}, \mathbf{h}; \theta_B) \tag{49}$$

where $\beta_m \in [0, 1]$ with $\beta_m = 0$ yielding p_A and $\beta_m = 1$ giving p_B. Annealing slowly the "temperature" from infinity to zero we gradually moves from the state space of proposal distribution to the space defined by the untractable distribution. Following the approach from **??** we can obtain transition operators for hidden and visible variables:

$$\begin{aligned} p(h^A|\mathbf{v}) &= \sigma((1 - \beta)NONE \\ p(h^B) &= \end{aligned} \tag{50}$$

Algorithm 2 Appendix – Implementation for BM and RBM.

$\theta, \phi \leftarrow$.
while not converged in θ, ϕ **do**
  Pick subset of size $\mathbf{x}_{1:M}$ from the full dataset uniformly at random.
  Compute $g \leftarrow \nabla_{\theta,\phi} \tilde{\mathcal{L}}^B \theta, \phi; \mathbf{x}_{1:M}, \epsilon_{1:M}$
end while

Figure 13: AIS for TAP2 and TAP3

5.2.2. Comparison

Two models were estimated with 5 of magnetizations to compare the quality of the approximation of the variational free energy. Figure **??** presents the estimates of 51 for the `TAP2` and `TAP3` models along with AIS estimates for different numbers of runs.

TAP2, TAP3

5.3. Deep RBM

renormalization group erham In the previous chapter it was argued that the unsupervised pre-training . science Deep learning methods

Following the approach from The breakthrough to effective training strategies for deep architectures came in 2006 with the CD algorithm for training Deep Belief networks (DBN) [**hinton2006reducing**]. DBNs are generative graphical models with many hidden layers of hidden causal variables which joint distribution has the following form:

$$p(\mathbf{x}, \mathbf{h}^1, \mathbf{h}^2, ..., \mathbf{h}^l) = p(\mathbf{x}|\mathbf{h}^1)P(\mathbf{h}^1|\mathbf{h}^2)...P(\mathbf{h}^{l-2}|\mathbf{h}^{l-1})P(\mathbf{h}^{l-1}|\mathbf{h}^l). \tag{51}$$

It was shown that adding an extra layers always improve a lower bound **??** on the training data if the number of feature detectors per layer is sufficiently large and the weights are initialized correctly. It was empirically proven that Figure **??** depicts the exemplary deep belief network. DBNs can be formed using a greedy layer-wise unsupervised training of stacked RBMs – algorithm **??** presents how the process folows: http://www.yann-ollivier.org/rech/publs/deeptrain.pdf - show the picture This simple and intuitive algorithm proved to be an

Algorithm 3 Learning Deep Belief Nets.

Train the first layer as an RBM, learning $P(\mathbf{x} = \mathbf{h^0}, \mathbf{h^1}$
for $l \in \{2, ..., L\}$ **do**
  Pass the mean activities $\mathbf{x}^l = P(\mathbf{h^l}|\mathbf{h^{l-1}}$ which become a representation of the input at the layer l.
  Train the l-th layer treating it as an RBM with $\mathbf{x}^l$ as an input.
end for

effective way of pretraining deep structures which laid the foundations of the resurgence of deep neural networks. Originally, the building blocks are trained following constrastive divergence procedure. However, the positive results obtained using extended mean-field approximation suggests that we may follow this procedure

The guarantee that we improve the bound is no longer valid if the size of subsequent hidden layers is not large enough however it was empiracally proven that such approach still can learn an effective generative model. After pretraining multiple layers of feature detectors, the model can be “unfolded” to form an autoencoder structures where the decoder network uses transposed weigths of the encoder network. At this stage, such network might be considered as feed forward deep neural architecture and might be used as a starting point for supervised fine-tuning with respect to any training criterion that depends on the learnt representation **??**.

The Figure **??** presents the reconstructions produced by the 25-dimensional deep autoencoder

Figure 14: DRBM TAP2 TAP3 PCD

Figure **??** shows the mean squared error on training and validation data sets for different methods for pre-training the RBMs.

Figure 15: MSE TAP2 TAP3 PCD on training and validation data set

5.4. Semi RBM

5.4.1. Exploiting the SRBM structure

Clamped free energy can be written in the form:

$$\begin{aligned}\mathcal{F}^c(\mathbf{v}) =& \sum_{\mathbf{h}} e^{-E(\mathbf{v},\mathbf{h})} = e^{\mathbf{b}'\mathbf{v}} \sum_{h_1} \dots \sum_{h_m} e^{-E(\mathbf{v},\mathbf{h})} \\ =& e^{\mathbf{b}'\mathbf{v}} \sum_{h_1} e^{h_1(c_1+W_{1\bullet}\mathbf{v})} \dots \sum_{h_m} e^{h_m(c_m+W_{m\bullet}\mathbf{v})} \\ =& e^{\mathbf{b}'\mathbf{v}} \prod_{j=1}^{m} \left(1+e^{c_i+W_{i\bullet}\mathbf{v}}\right)\end{aligned} \tag{52}$$

5.5. Boltzmann Machine

$$E(\mathbf{v},\mathbf{h}) = -\sum_i a_i v_i - \sum_j b_j h_j - \sum_{i,j} v_i w_{ij} h_j = -\frac{1}{2}\mathbf{v}^T\mathbf{V}\mathbf{v} - \frac{1}{2}\mathbf{h}^T\mathbf{J}\mathbf{h} - \mathbf{h}^T\mathbf{W}\mathbf{v}$$

This approximation this time will be used two times, first for $P(\mathbf{h}|\mathbf{v})$ and the second time for the variational free energy. We have:

$$KL\left(Q(\mathbf{h}|\mathbf{v})\|P(\mathbf{h}|\mathbf{v})\right)$$

5.6. GBRBM

6. Conclusions

7. Appendix