Training Restricted Boltzmann Machine Using High-Temperature Expansion

Paweł Budzianowski

Department of Engineering
University of Cambridge

This dissertation is submitted for the degree of
Master of Philosophy

Clare Hall College

August 2016

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others. This dissertation contains fewer than $15,000$ words including appendices, bibliography, footnotes, tables and equations and has fewer than 20 figures.

Paweł Budzianowski
August 2016

Abstract

This is where you write your abstract ...

Table of contents

List of figures **vii**

1 Introduction **1**

2 Extended mean field approximation **3**

2.1 Graphical models as Markov random fields 3
2.2 Boltzmann distribution 4
  2.2.1 Statistical physics perspective 5
  2.2.2 (Naive) Mean field approximation 6
  2.2.3 Extended mean field approximation (EMF) 7
  2.2.4 EMF approximation of the free energy 9
  2.2.5 Boltzmann machine 10

3 Evaluation on toy models **13**

3.1 Adaptation of EMF to RBM 13
3.2 Schedule of updates 14
  3.2.1 Asynchronously 15
  3.2.2 Sequentially 15
  3.2.3 Parallelly 15
3.3 Toy models 16
  3.3.1 Grid toy model 16
  3.3.2 RBM toy model 18

4 Learning of Boltzmann machines **21**

4.1 Unsupervised learning 21
  4.1.1 Training of Boltzmann Machines 21
  4.1.2 Monte Carlo methods 23
  4.1.3 Contrastive Divergence 24
  4.1.4 Learning using extended mean field approximation 25
  4.1.5 Approximating the log-likelihood 26
  4.1.6 Real scale model – MNIST data set 26
  4.1.7 Comparison of both approaches 27
  4.1.8 Generated samples from the models 28

5 Various experiments **31**

5.1 Comparison of schedules of updates 31

5.1.1 Evaluation of EMF approximation . . . 31
5.2 Annealed Importance Sampling . . . 33
5.2.1 Comparison . . . 35
5.3 Deep structures . . . 36
5.3.1 Unsupervised pre-training of deep neural networks . . . 36
5.3.2 Deep belief nets . . . 37
5.3.3 Reconstruction analysis . . . 38

References **43**

Appendix A EMF derivation **45**

List of figures

2.1 Boltzmann Machines . . . 11

3.1 1 . . . 16
3.2 1 . . . 17
3.3 Grid toy model – the couplings strength . . . 17
3.4 1 . . . 18
3.5 1 . . . 19
3.6 1 . . . 19
3.7 1 . . . 20
3.8 1 . . . 20

4.1 1 . . . 24
4.2 1 . . . 27
4.3 1 . . . 28
4.4 1 . . . 29

5.1 Estimates of log-likelihood with different updates . . . 32
5.2 1 . . . 35
5.3 1 . . . 36
5.4 Deep belief net . . . 37
5.5 Reconstruction of digits with basic models . . . 38
5.6 Reconstructions of digits with deep structures . . . 39
5.7 1 . . . 40

Chapter 1

Introduction

The resurgence of deep neural architectures over couple last years allowed us to train very powerful statistical model which yields substantial improvements in many areas of machine learning like speech or vision recognition, image processing or machine translation. One of the main reasons for the resurgence of deep architectures is przypisany to the effective unsupervised pre-training of those structures with restricted Boltzmann machines (RBMs) – a special case of general energy-based model where there aren't connections between nodes from the same layer.

It has been shown empirically that unsupervised pre-training of deep structures .. [4]

A new algorithm - deterministic - previous didn't work.

Chapter 1 We will start by embedding BMs in to the framework of probabilistic graphical models known as Markov random fields which enables us to obtain many theoretical results and well-developed algorithms

Chapter 2

Chapter 3

Chapter 2

Extended mean field approximation

2.1. Graphical models as Markov random fields

One of the basic concepts in the theory of statistical modelling are graphical models which greatly help in analysing multivariate phenomena. Visualizations by graphs help in efficient development and understanding of analysed models while complex computations can be performed exploiting the graph properties. 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. Moreover, an important concept related with every graph structure is the notion of clique which is a subset of V in which all nodes are pairwise connected. One of the most useful class of graphical models is a Markov random field (MRF) which is a type undirected random field that satisfies global Markov property, specifically:

Definition 1 *An undirected graphical model G is a Markov random field if for any node X_v in the graph the following conditional property holds:*

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

where $X_{G\backslash i}$ 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 separate them. This property is closely related with the notion of factorization of the joint probability distribution:

Definition 2 *A probability distribution $P(\mathbf{X})$, $\mathbf{X} = (X_1, ..., X_n)$, defined on an undirected graphical model G factorizes over G if there exists a set of non-negative functions (potentials)* *on cliques $\{\psi_C\}_{C\in\mathcal{C}}$ 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.

The following theorem shows a direct connection between those two family of probability distributions that will be heavily exploited in the following sections:

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

Theorem 1 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:

$$p(x_1, x_2, ..., x_n) = \frac{1}{Z}\prod_{C\in\mathcal{C}}\psi_C(x_C) = \frac{1}{Z}e^{\sum_{C\in\mathcal{C}}\ln\psi_C(x_C)} = \frac{1}{Z}e^{-E(x)} \tag{2.1}$$

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 positive MRF can be expressed as in 2.1. This relationship allows us to take advantage of both approaches to statistical modelling as we can perform inference exploiting a graph structure as well as algebraic properties of the Gibbs family. Moreover, this form of distribution is a natural candidate to approximate and model phenomena which can be also seen as graphical models. In next sections we will analyse one particular class of Gibbs distribution which is powerful enough to approximate any probability distribution.

2.2. Boltzmann distribution

In this thesis, an undirected graphical model (which can be also seen a MRF) that will be extensively analysed is the Boltzmann distribution which in the most general form has the following joint distribution:

$$p(x_1, x_2, ..., x_n) = \frac{1}{Z}\exp\left(-\frac{1}{T}E(x_1, x_2, ..., x_n)\right) \tag{2.2}$$

where T is the temperature of the system and E is the *energy* of the system defined as:

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

and $Z = \sum_{\mathbf{x}} \exp(-\frac{1}{T}E(x_1, x_2, ..., x_n))$ is the normalization constant often called the partition function. The pair-wise potential function has here the form:

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

while the magnetic field is defined as:

$$\psi_i = \exp(\theta_i x_i).$$

Wide range of distributions having the form of 2.2 is extensively used in physics to compute the energy of the system of particles. This model proves to be very useful in many other applications such as the error-correcting code, computer vision, medical diagnosis or statistical mechanics [27]. This model may represent 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 summation over a number of states that grows exponentially with the number of variables and is intractable even for a small number of variables. That is why, we have to resort to some tractable approximations which two of them will be considered in next sections.

2.2.1. Statistical physics perspective

Following the notation from the statistical physics, consider a graphical model over a set of random variables $\mathbf{s}$ taking the "spin" values $\{0, 1\}$. In the context of statistical physics, these values might represent the orientations of magnets in a field, or the existence of particles in a gas. Lets consider the Boltzmann distribution for such system:

$$P(\mathbf{s}) = \frac{e^{-\frac{1}{T}E(\mathbf{s})}}{\sum_{\mathbf{s}} e^{-\frac{1}{T}E(\mathbf{s})}} = \frac{1}{Z} e^{-\frac{1}{T}E(\mathbf{s})} \tag{2.3}$$

where energy is defined as:

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

This yields the well-known Ising model which plays a primarily role in the analysis of phase transitions in many physical systems. 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 [27].

As it was mentioned previously, the number of configurations in the system scales exponentially with the number of variables which forces us to resort to some kind of approximations. Instead of imposing some restrictions on the model structure, we will try to find an approximate distribution Q that poses useful characteristics and minimizes the relative entropy often called the Kullback-Leibler divergence:

$$KL(Q||P) = \mathbb{E}_Q\left(\ln\frac{Q}{P}\right) = \sum_{\mathbf{s}} Q(\mathbf{s})\ln\frac{Q(\mathbf{s})}{P(\mathbf{s})}. \tag{2.4}$$

The KL-divergence is a non-symmetric measure of the difference between two distributions which is always non-negative. Substituting P from 2.3 into the previous equation yields:

$$KL(Q||P) = \ln Z + \frac{1}{T}\mathbb{E}[Q] - H[Q]$$

where H stands for entropy of the distribution Q, $\ln Z$ is the *free energy* and $\mathbb{E}[Q] = \sum_{\mathbf{s}} Q(\mathbf{s})E(\mathbf{s})$ is called the *variational energy* where $\mathbb{E}$ refers to the average configuration under the Boltzmann measure [19]. The partition function Z doesn't depend on Q and we need to only focus on minimizing the variational free energy:

$$F[Q] := \mathbb{E}[Q] - TH[Q]. \tag{2.5}$$

At equilibrium i.e. when the approximate distribution would equal the desired one the KL-divergence is 0 and the variational free energy is equal to the Helmholtz free energy defined by $\mathcal{F} := -T\ln Z$.

2.2.2. (Naive) Mean field approximation

The most widely used approximation to the family of models defined in 2.3 is the mean field approximation which is obtained by taking as an approximator the family of distribution that factorizes as following:

$$Q(\mathbf{s}) = \prod_i q_i(s_i) \tag{2.6}$$

which results in neglecting the dependency between the random variables. The variational free energy in this case takes the form:

$$F^{MF} = -\sum_{(ij)}\sum_{s_i,s_j} w_{ij}q_i(x_i)q_j(x_j) - \sum_i\sum_{s_i}\theta_i q_i(x_i) + T\sum_i\sum_{s_i} q_i(s_i)\ln q_i(s_i) \tag{2.7}$$

and the energy for a single spin is:

$$E(s_i) = -\theta_i s_i - s_i\sum_j w_{ij}m_j \tag{2.8}$$

where neighbour spins are replaced by certain effective mean fields which are defined as:

$$m_i = \mathbb{E}_{q_i}(s_i), \quad i \in \{1, ..., N\}. \tag{2.9}$$

In terms of magnetizations, 2.7 becomes:

$$F^{MF} = -\sum_{(ij)} w_{ij} m_i m_j - \sum_i \theta_i m_i + T \sum_i \left[m_i \ln m_i + (1 - m_i) \ln(1 - m_i) \right]. \tag{2.10}$$

Minimizing 2.10 with respect to magnetizations yields the so-called mean field stationary conditions:

$$m_i = \mathrm{sigm}\left(\frac{1}{T} \sum_j w_{ij} m_j + \frac{1}{T} \theta_i \right), \quad i \in \{1, ..., N\} \tag{2.11}$$

where N is the number of spins in the model. These equations are usually run sequentially. As the free energy is convex [25], these updates can be seen as coordinate descent in $\mathbf{m}$ that guarantees to obtain some stable solution. However, there might exist many solutions to 2.11 as well as some of them might not be even local minima. Nonetheless, the MF approach is exact for the infinite-ranged Ising model where each the node is connected to every other node and all couplings w_{ij} are positive and equal[11].

Additionally, the variational mean field approximation yields an upper bound on the exact free energy as the following holds:

$$\begin{aligned} \ln Z = \ln \sum_{\mathbf{s}} \exp(-\frac{1}{T} E(\mathbf{s})) &= \ln \sum_{\mathbf{s}} Q(\mathbf{s}) \frac{\exp(-\frac{1}{T} E(\mathbf{s}))}{Q(\mathbf{s})} \\ \geqslant \sum_{\mathbf{s}} Q(\mathbf{s}) \ln \frac{\exp(-\frac{1}{T} E(\mathbf{s}))}{Q(\mathbf{s})} &= -\frac{1}{T} \mathbb{E}_Q(E(\mathbf{s})) + H(Q) \end{aligned} \tag{2.12}$$

where the middle inequality follows from the concavity of the log function and application of Jensen's inequality. We arrive at the bound by reversing the inequality:

$$\mathcal{F} = -T \ln Z \leqslant \mathbb{E}[Q] - T H[Q] = F[Q]. \tag{2.13}$$

2.2.3. Extended mean field approximation (EMF)

At the expense of loosing the rigorous upper bound on the Helmholtz free energy, we might consider a different approximation for the magnetization dependent variational free energy [7]. We will minimize 2.5 where instead of assuming Q to be a product distribution we require that magnetizations has appropriate values, i.e.:

$$\mathbb{E}_Q(\mathbf{s}) = \mathbf{m}. \tag{2.14}$$

where $\mathbf{m}$ is fixed. Thus, the variational free energy is now defined as:

$$\beta F(\mathbf{m}) = \min_Q \{ E(Q) - H(Q) \mid \mathbb{E}(\mathbf{S}) = \mathbf{m} \} \tag{2.15}$$

where β was introduced as a reciprocal of temperature – this will allow us to a perform useful expansion w.r.t β later on. The constrained optimization problem can be transformed into unconstrained using Lagrange multipliers, i.e.:

$$E(Q)-H(Q)-\sum_i \lambda_i(\mathbb{E}(s_i)-m_i). \tag{2.16}$$

Thus, the minimizing distribution has the form:

$$Q_{\mathbf{m}}(\mathbf{s})=\frac{1}{Z}e^{-E(\mathbf{s})+\lambda_i s_i} \tag{2.17}$$

with partition function $Z=\sum_{\mathbf{s}} e^{-E(\mathbf{s})+\sum_i \lambda_i s_i}$. Using this distribution back into 2.15 along with making auxiliary fields λ temperature-dependant and suppressing (for the moment) the λ and $\{m_i\}$ dependence of F we arrive at the objective function:

$$-\beta F=\ln\sum_{\mathbf{s}}\exp\left(\beta\sum_{(ij)}w_{ij}s_is_j+\beta\sum_i\theta_i s_i+\sum_i\lambda_i(\beta)(s_i-m_i)\right) \tag{2.18}$$

Lets now expand $-\beta F$ around $\beta=0$:

$$-\beta F=-(\beta F)_{\beta=0}-\left(\frac{\partial(\beta F)}{\partial\beta}\right)_{\beta=0}\beta-\left(\frac{\partial^2(\beta F)}{\partial\beta^2}\right)_{\beta=0}\frac{\beta^2}{2}-\dots \tag{2.19}$$

In this case, the spins are entirely controlled by their auxiliary fields. Although it not a desired assumption, it will allow us to obtain useful form of the expansion. Magnetizations are fixed equal to $\mathbb{E}_Q(\mathbf{s})$, particularly for $\beta=0$ which gives an important conjugate relation between magnetizations and auxiliary fields:

$$m_i=\mathbb{E}_{\beta=0}(s_i)=\frac{\exp(\lambda_i(0))}{\exp(\lambda_i(0))+1}=\mathrm{sigm}(\lambda_i(0)) \tag{2.20}$$

We can now choose which variables use in derivations and this is a purely dependent on mathematical convenience. As the equation 2.20 is easy to invert, we will work on the magnetizations. The first term from the 2.19 takes now the form:

$$\begin{aligned}
-(\beta F)_{\beta=0}=&\ln\sum_{\mathbf{s}}\exp\left(\sum_i\lambda_i(0)(s_i-m_i)\right)\\
=&\ln\left\{\sum_{s_1}\exp\left(\lambda_1(0)(s_1-m_1)\right)\dots\sum_{s_n}\exp\left(\lambda_n(0)(s_n-m_n)\right)\right\}\\
=&\ln\left\{(\exp(\lambda_i(0))+1)\exp(-\lambda_1(0)m_1)\dots(\exp(\lambda_i(0))+1)\exp(-\lambda_1(0)m_n)\right\}\\
=&\sum_i\left\{\ln\left(\frac{1}{1-m_i}\right)-m_i\ln\left(\frac{m_i}{1-m_i}\right)\right\}\\
=&-\sum_i\left[m_i\ln(m_i)+(1-m_i)\ln(1-m_i)\right]
\end{aligned} \tag{2.21}$$

where using 2.20, we replace auxiliary variables by:

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

As we can see, this is exactly the mean field entropy from the equation 2.10. Next, the first derivative is:

$$-\left.\frac{\partial(\beta F)}{\partial\beta}\right|_{\beta=0} = \sum_{(ij)} w_{ij}\mathbb{E}_{\beta=0}(s_i s_j) + \sum_i \theta_i \mathbb{E}_{\beta=0}(s_i) - \sum_i \left.\frac{\partial\lambda_i(\beta)}{\partial\beta}\right|_{\beta=0} \mathbb{E}(s_i - m_i) \tag{2.22}$$

and as it was observed earlier, at $\beta = 0$ the spins are independent and the expectation in the first term factorizes. Thus, we have:

$$-\left.\frac{\partial(\beta F)}{\partial\beta}\right|_{\beta=0} = \sum_{(ij)} w_{ij} m_i m_j + \sum_i \theta_i m_i. \tag{2.23}$$

Vomparing 2.23 and 2.21 with 2.10 we can see that we have already recovered the simple mean field approximation. Yedida and Georges [7] showed how to continue this expansion to the arbitrarily high order (derivation in Appendix). However, in next chapters the expansion only up to the third order will be used:

TODO ADDD ONSAGER TAP NAMES

$$\begin{aligned}
-\beta F^{EMF} = & -\sum_i [m_i \ln(m_i) + (1-m_i)\ln(1-m_i)] \\
& + \beta \sum_{(ij)} w_{ij} m_i m_j + \beta \sum_i \theta_i m_i \\
& + \frac{\beta^2}{2} \sum_{(ij)} w_{ij}^2 (m_i - m_i^2)(m_j - m_j^2) \\
& + \frac{2\beta^3}{3} \sum_{(ij)} w_{ij}^3 (m_i - m_i^2)(\frac{1}{2} - m_i)(m_j - m_j^2)(\frac{1}{2} - m_j) \\
& + \beta^3 \sum_{(ijk)} w_{ij} w_{jk} w_{ki} (m_i - m_i^2)(m_j - m_j^2)(m_k - m_k^2) + \dots
\end{aligned}$$

where (ijk) stands for coupled triplets of nodes. Contrary to the mean field approximation, the extended approach takes into account all distinct pairs and triplets of spins. This will lead to significant improvements over naive mean field approach in learning graphical models from Boltzmann family.

2.2.4. EMF approximation of the free energy

Although it is very straightforward to obtain naive mean field approximation from the extended approach, unlike the former, in general case this method doesn't bound in any way the free energy $-\ln Z$. This follows from the fact that we don't enforce any constraint regarding marginal or joint probabilities. 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 [27]. There are a few examples in statistical physics where this method works very reliably in a wide variety of temperatures [20] however in general there aren't any theoretical foundations for the robustness of this expansion. In the next chapter this approach will be tested on various toy models to assess the quality of the approximation.

2.2.5. Boltzmann machine

A particular example from the family of distributions defined in 2.2 is a Boltzmann machine [1] which has a two-layer architecture with N visible units $\mathbf{v} = (v_1, ..., v_N)$ and M hidden units $\mathbf{b} = (h_1, ..., h_M)$ that can take values 0 or 1. The energy function has the form:

$$E(\mathbf{v},\mathbf{h}) = -\sum_i a_i v_i - \sum_j b_j h_j - \sum_{i,j} v_i W_{ij} h_j - \sum_{i<j} v_i V_{ij} v_j - \sum_{i<j} h_i J_{ij} h_j,$$

where W_{ij}, V_{ij}, H_{ij} are real valued couplings between visible and hidden, visible and visible and hidden and hidden units respectively for $i \in \{1, ..., n\}$, $j \in \{1, ..., M\}$. An example of such structure presents Figure 2.1 (left). The connections between units from the same layer makes this model hard to operate with – for example even with given visible units, we are not able to compute the marginal probability $p(\mathbf{v})$ as this requires summation that scales exponentially with number of hidden units.

Restricted Boltzmann machine

A restricted Boltzmann machine (RBM) is a special case of Boltzmann machine which overcomes difficulties associated with Boltzmann machines at the same time preserving the approximating power. The energy function takes the simplified form:

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

The graph of an RBM has connections between visible and hidden units but not between any variables from the same layer (Figure 2.1, right). This results in independence between variables from the same layer given the state of the other layer. The RBM can be interpreted as a stochastic neural network, where units and connections correspond to neurons and synaptics respectively [5].

Approximator of any distribution

The power of RBM comes from the fact that with data-dependent number of hidden units they become non-parametric and possess universal approximation properties [12]. It can be shown that with additional hidden units there exist weight values for these new units that

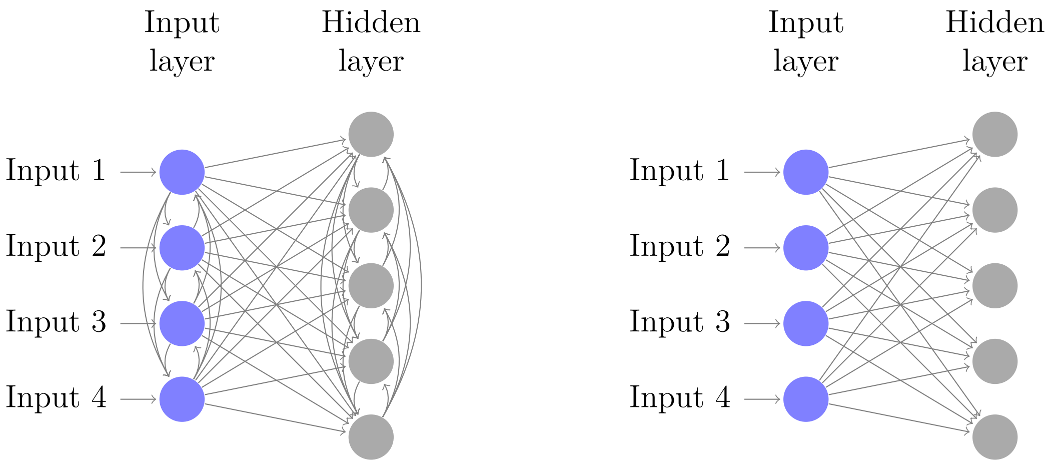

Fig. 2.1 Exemplary graphs of Boltzmann machine (left) and restricted Boltzmann Machine (right) with 4 visible and 5 hidden units.

guarantee improvement in increasing the log-likelihood of observed data. Taking this process to extreme, we can obtain a model with an unlimited expressive power:

Theorem 2 (LeRoux-Bengio, 2010) *Any distribution over* $\{0,1\}^n$ *can be approximated arbitrarily well (in the sense of the KL divergence) with an RBM with* $k+1$ *hidden units where* k *is is the number of input vectors whose probability is not* 0*.*

This theorem shows that an RBM is the natural candidate for modelling an arbitrary distribution where we are interested in learning powerful generative model. In the next chapters, analysed models will not have more hidden units than visible ones thus we lose the guarantee of learning an unbiased approximate distribution. Nonetheless, the experiments show that even then the models that are learnt provide effective generative approximator of an unknown distribution.

Exploiting the RBM structure

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

$$\begin{aligned} p(\mathbf{h}|\mathbf{v}) &= \prod_{i=1}^{M} p(h_i|\mathbf{v}), \\ p(\mathbf{v}|\mathbf{h}) &= \prod_{i=1}^{N} p(v_i|\mathbf{h}). \end{aligned} \tag{2.24}$$

The conditional probability of a single variable being one is also explicitly available:

$$\begin{aligned} p(h_i = 1|\mathbf{v}) = p(h_i = 1|\mathbf{h}_{-i}, \mathbf{v}) = \frac{p(h_i = 1, \mathbf{h}_{-i}, \mathbf{v})}{p(\mathbf{h}_{-i}, \mathbf{v})} \\ = \frac{\exp(-E(h_i = 1, \mathbf{h}_{-i}, \mathbf{v}))}{\exp(-E(h_i = 1, \mathbf{h}_{-i}, \mathbf{v})) + \exp(-E(h_i = 0, \mathbf{h}_{-i}, \mathbf{v}))} \\ = \frac{1}{1 + \exp(\sum_{n=1}^{N} W_{i,n} v_n + a_n)} \\ = \mathrm{sigm}(\sum_{n=1}^{N} W_{i,n} v_n + b_i)) \end{aligned} \tag{2.25}$$

and following the same steps we can show that:

$$p(v_j = 1|\mathbf{h}) = \mathrm{sigm}(\sum_{m=1}^{M} W^T_{i,m} h_m + a_j)). \tag{2.26}$$

The independence between the variables in one layer makes sampling from conditional distributions 2.25 and 2.26 easy to perform. This will be crucial for effective learning of this model when we don't know a priori the parameters. Moreover the nominator from the $p(\mathbf{v})$ factorizes over hidden variables and we can write:

$$\begin{aligned} \sum_{\mathbf{h}} e^{-E(\mathbf{v},\mathbf{h})} = e^{\mathbf{b}'\mathbf{v}} \sum_{h_1} \dots \sum_{h_m} e^{-E(\mathbf{v},\mathbf{h})} \\ = e^{\mathbf{b}'\mathbf{v}} \sum_{h_1} e^{h_1(c_1 + W_{1\bullet}\mathbf{v})} \dots \sum_{h_m} e^{h_m(c_m + W_{m\bullet}\mathbf{v})} \\ = e^{\mathbf{b}'\mathbf{v}} \prod_{j=1}^{m} \left(1 + e^{c_i + W_{i\bullet}\mathbf{v}}\right) \end{aligned} \tag{2.27}$$

where $W_{i\bullet}$ denotes the i-th row of the matrix W. These properties will be heavily exploited later on when we will be interested in computing the probability of observed data points.

Chapter 3

Evaluation on toy models

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.

3.1. 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 $\mathbf{m}^{\mathbf{v}} = \{m_i\}_{i=1}^{N}$ and $\mathbf{m}^{\mathbf{h}} = \{m_i\}_{j=1}^{M}$ 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:

$$\begin{aligned}
F^{EMF}(\mathbf{m}^{\mathbf{v}},\mathbf{m}^{\mathbf{h}}) \simeq\ & H(\mathbf{m}^{\mathbf{v}},\mathbf{m}^{\mathbf{h}}) \\
& -\sum_i a_i m_i^v - \sum_j b_j m_j^h \\
& -\sum_{i,j}\left(m_i^v w_{ij} m_j^h + \frac{w_{ij}^2}{2}(m_i^v - (m_i^v)^2)(m_j^h - (m_j^h)^2)\right) \\
& -\sum_{i,j}\left(\frac{2w_{ij}^3}{3}(m_i^v - (m_i^v)^2)(\frac{1}{2} - m_i^v)(m_j^h - (m_j^h)^2)(\frac{1}{2} - m_j^h)\right),
\end{aligned} \tag{3.1}$$

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 [6]. 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:

$$\frac{\partial F^{EMF}}{\partial m_i^v} = 1 + \ln m_i - 1 - \ln(1 - m_i^v) - R = 0, \tag{3.2}$$

where

$$\begin{aligned} R = & a_i + \sum_j w_{ij} m_j^h - \sum_j w_{ij}^2 \left(m_i^v - \frac{1}{2} \right) \left(m_j^h - (m_j^h)^2 \right) \\ & + \sum_j \frac{w_{ij}^3}{3} \left(m_i^v - (3m_i^v)^2 + 2(m_i^v)^3 \right) (m_j^h - (m_j^h)^2)(\frac{1}{2} - m_j^h). \end{aligned} \tag{3.3}$$

This can be regrouped as:

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

which leads to the following relation:

$$m_i^v = \frac{\exp(R)}{1 + \exp(R)} = \text{sigm}(R), \tag{3.4}$$

where $\text{sigm}(x) = (1 + e^{-x})^{-1}$, and 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 2.24.

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 they *must* be run sequentially in order to yield proper estimates [26]. Similarly, in the case of the extended mean field approximation, it was proposed that an iterative, asynchronous algorithm may serve as an update rule [6] 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 toy models and on the real life data set example. The updates here are considered only up to the second order.

3.2.1. 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 to the values from the previous step. This can be written using the time index t in the following way:

$$\begin{aligned}\mathbf{m}^h[t+1] &= \operatorname{sigm}\left[\mathbf{b} + W\mathbf{m}^v[t] - \left(\mathbf{m}^h[t] - \frac{1}{2}\right)^T \odot W^2 \left(\mathbf{m}^v[t] - (\mathbf{m}^v[t])^2\right)\right], \\ \mathbf{m}^v[t+1] &= \operatorname{sigm}\left[\mathbf{a} + W^T\mathbf{m}^h[t+1] - \left(\mathbf{m}^v[t] - \frac{1}{2}\right) \odot (W^2)^T \left(\mathbf{m}^h[t+1] - (\mathbf{m}^h[t+1])^2\right)\right],\end{aligned} \tag{3.5}$$

where $\odot$ denotes Hadamard product.

3.2.2. 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:

$$\begin{aligned}m_i^h[t+1] &= \operatorname{sigm}\left[b_i + \sum_j \left(w_{ij}m_j^v[t] - w_{ij}^2(m_i^h[t] - \frac{1}{2})\left(m_j^v[t] - (m_j^v[t])^2\right)\right)\right], \\ m_{j=i+1}^v[t+1] &= \operatorname{sigm}\left[a_i + \sum_{l\neq i} \left(w_{ij}m_l^h[t] - w_{ij}^2(m_l^h[t] - \frac{1}{2})(m_j^v[t] - (m_j^v[t])^2)\right)\right. \\ &\qquad \left. + \left(w_{ij}m_i^h[t+1] - w_{ij}^2(m_i^h[t+1] - \frac{1}{2})(m_j^v[t] - (m_j^v[t])^2)\right)\right]\end{aligned} \tag{3.6}$$

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

3.2.3. 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:

$$\begin{aligned}\mathbf{m}^h[t+1] &= \operatorname{sigm}\left[\mathbf{b} + W\mathbf{m}^v[t] - \left(\mathbf{m}^h[t] - \frac{1}{2}\right)^T \odot W^2 \left(\mathbf{m}^v[t] - (\mathbf{m}^v[t])^2\right)\right], \\ \mathbf{m}^v[t+1] &= \operatorname{sigm}\left[\mathbf{a} + W^T\mathbf{m}^h[t] - \left(\mathbf{m}^v[t] - \frac{1}{2}\right) \odot (W^2)^T \left(\mathbf{m}^h[t] - (\mathbf{m}^h[t])^2\right)\right].\end{aligned} \tag{3.7}$$

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 3.1 presents graphically all proposed procedures.

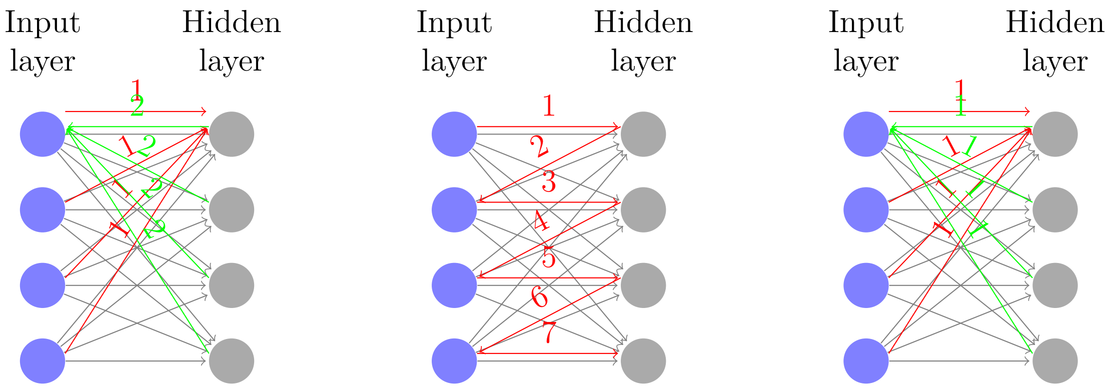

Fig. 3.1 Graphical visualisations of three different schedule of updates considered for the RBM toy model. Numbers above the coloured arrows denotes the order of updates.

In the case of fixed point algorithms, it is a common practice to use damped updates [16] where as a new value for the given magnetization we take weighted average of its value from the previous step and after performing an update. The weight hyper-parameter λ is usually in the range of $[0, 1]$. 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 [6], [26].

3.3. Toy models

As it was mentioned in the previous chapter, unlike the naive mean field approach, the extended approximation doesn't yield niether an upper nor lower bound for the variational free energy. Moreover, 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*. The toy grid model will have 16 nodes while the toy RBM model will have 20 nodes (10 hidden and 10 visible).

3.3.1. 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 3.2 shows this model from the graphical model's perspective. The nature of this model implies that the sequential updates of magnetizations seem as the most natural way to obtain statistics of the system in the equilibrium and that is why only updates of the form 3.6 will be considered here. In this case each magnetization m_i is updated one at a time using equation 3.4.

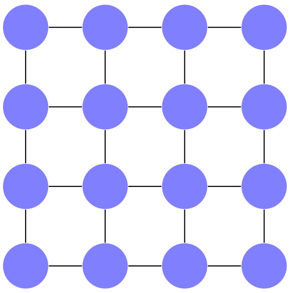

Fig. 3.2 4 × 4 grid toy model used for an exact inference.

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 becomes 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 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 between the spins in the system.

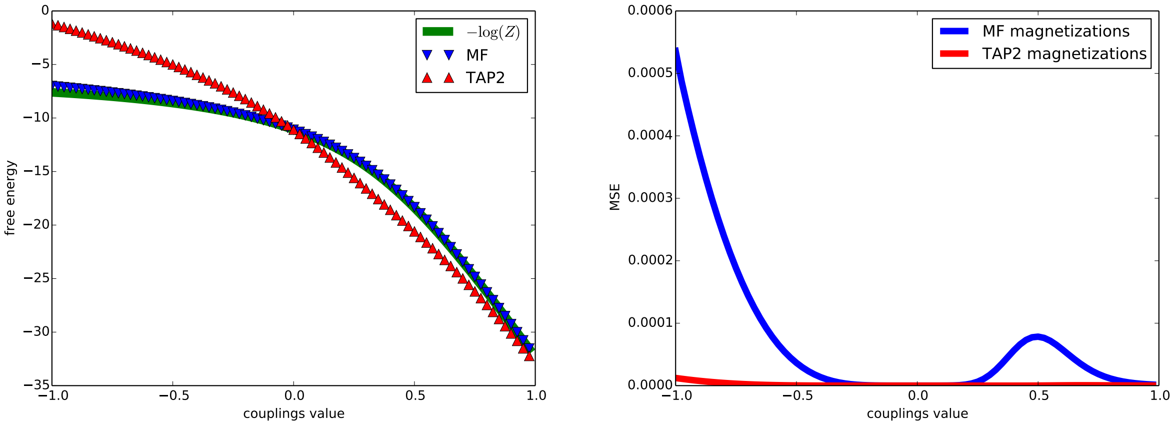

Fig. 3.3 Comparison of two variational approaches – free energy estimates (left) with the true free energy (green line) and MSE between real and estimated magnetisations (right) as a function of the couplings strength ranging from -1 to 1.

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 3.4). 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.

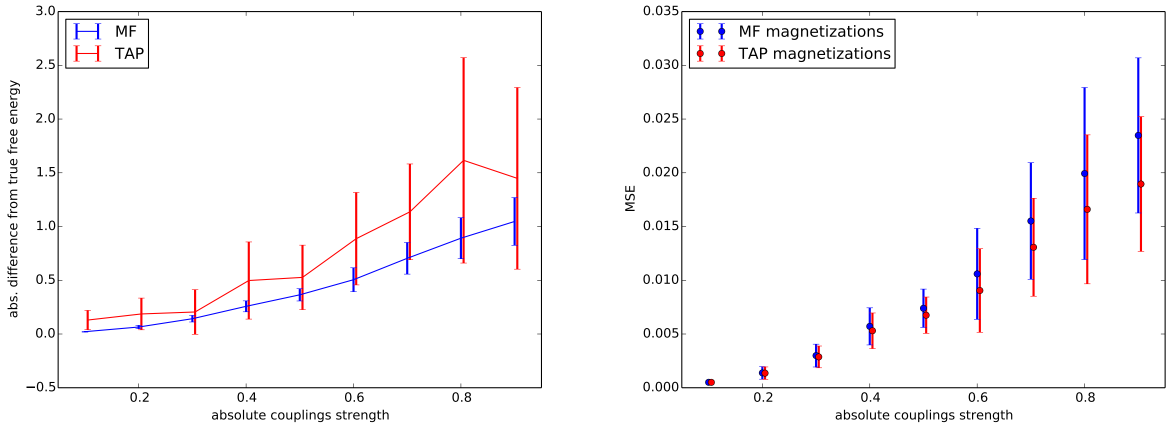

Fig. 3.4 Absolute difference between true free energy and the one computed by the naive and the extended mean field approach (left) and MSE between real and estimated magnetisations (right) as a function of the absolute value of couplings strength.

The addition of external fields didn't change substantially the performance of methods and the results are not included here.

3.3.2. RBM toy model

Due to the different structure of connections between states, the RBM toy model is less trivial to approximate. This should lead to considerably different results in the performance comparing to the toy mode and it will be an another motivation 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 [26]. To assess the impact on the 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).

Unlike the case of the grid model, 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 and thus results for sequential updates weren't included here.

When couplings were random with randomly assigned negative or positive signs, TAP approximation again yields consistently much better estimates that most of the time are

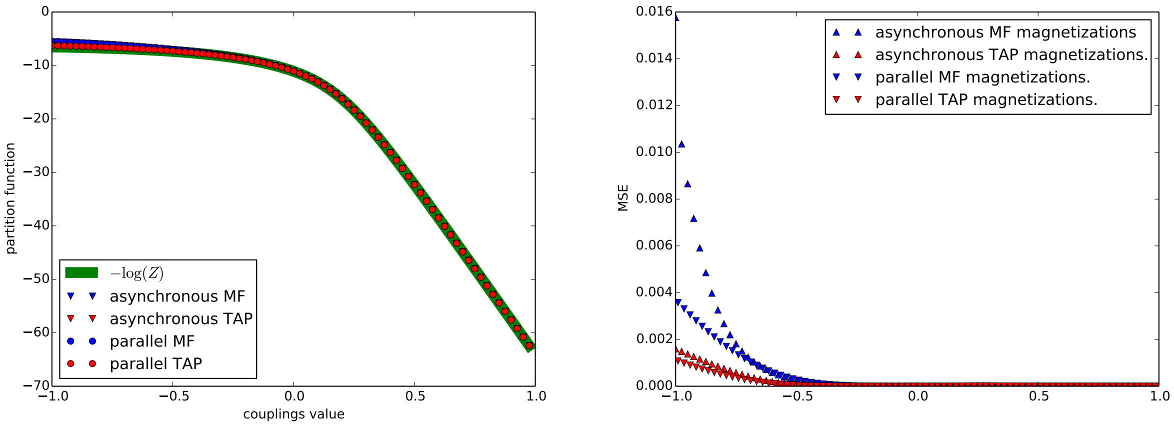

Fig. 3.5 Comparison of two variational approaches – free energy estimates (left)with the true free energy (green line) and MSE between real and estimated magnetisations (right) as a function of the couplings strength ranging from −1 to 1.

exact at the same time having much smaller variance (Figure 3.6). 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 iterations.

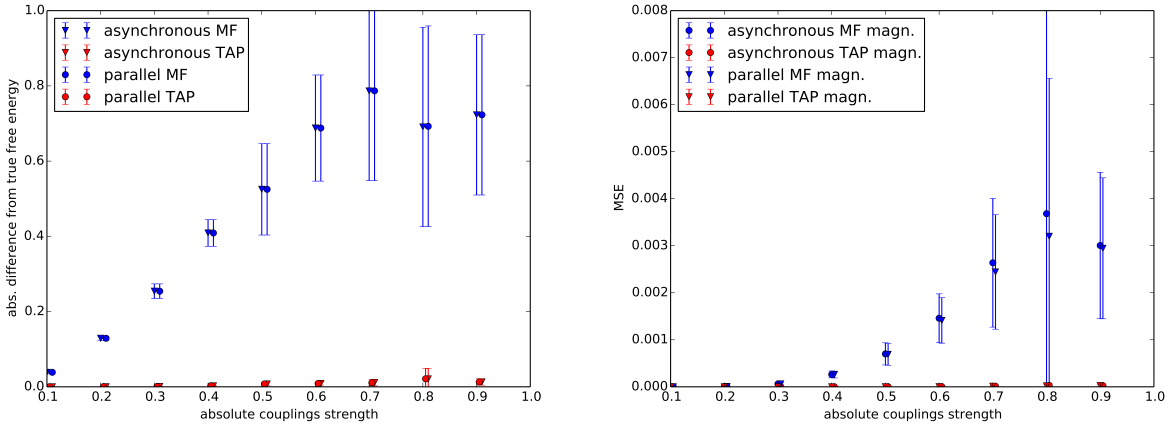

Fig. 3.6 Absolute difference between the true free energy and the one computed by the naive and the extended mean field approach (left) and MSE between real and estimated magnetisations (right) as a function of the absolute value of couplings strength.

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 with different schedules, the couplings were chosen again randomly around given mean values but this time the sign of the weight was chosen sequentially (Figure 3.7).

This time, TAP estimates are not exact for couplings with an absolute strength around values 0.2 or 0.3. Moreover, the sequential updates yield very poor estimates for the magnetizations – the MSE is almost the same as with the naive approach. The addition of external fields

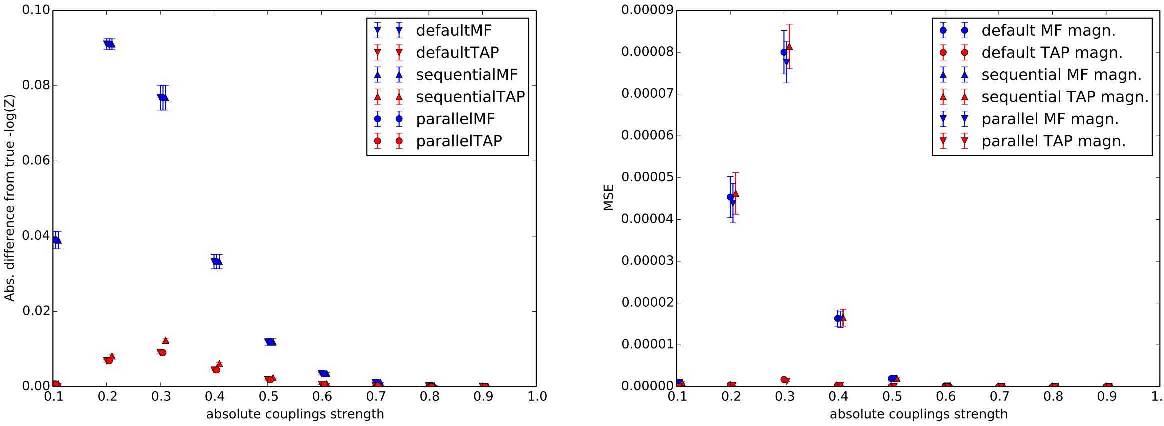

Fig. 3.7 Absolute difference between the true free energy and the one computed by the naive and the extended mean field approach (left) and MSE between real and estimated magnetisations (right) as a function of the absolute value of couplings strength with sequentially opposite signs.

this time extensify the differences in performance between updates (Figure 3.8). The MSE between true and estimated magnetizations is higher for the TAP model with sequential updates than for the corresponding naive model.

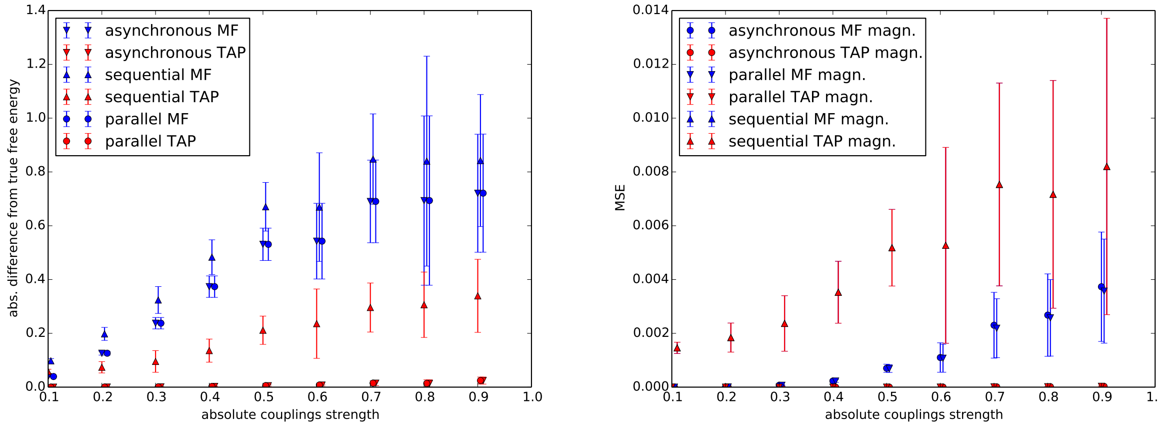

Fig. 3.8 Absolute difference between true free energy and the one computed by naive and extended mean field approach (left) an MSE between real and estimated magnetisations (right) as a function of the absolute value of couplings strength with sequential changes of signs.

Taking into consideration all results we can infer that the extended approximation yields robust results even though the expansion is performed at $\beta = 1$ and the couplings between the spins might have strong connections. Moreover, the real benefits of modelling the relationship between the spins can be observed when the MRF has the RBM structure. In this case the estimates of the free energy and the magnetizations were almost exact for the EMF-based model. However, contrary to the heuristics parallel updates perform just as well as the asynchronous ones while the sequential updates might drastically worsen the quality of estimates for the EMF-based models.

Chapter 4

Learning of Boltzmann machines

4.1. Unsupervised learning

So far it was assumed that the couplings in analysed structures (along with bias terms) were known a priori. However, in general when we analyse some phenomena we don't know this values and we are interested in learning an unknown distribution Q based on some observed data $\mathcal{D}$. The theoretical results suggests that the RBM structure is a natural candidate for approximating underlying distribution from which the data were generated. Thus, the unsupervised learning in this case consists of learning the parameters θ of the approximate distribution P. Therefore, our general goal is to maximize the probability of $\mathcal{D}$ under the MRF distribution i.e. we are looking for the vector of parameters θ that maximize the likelihood given the training data:

$$\max_{\theta} \ln \mathcal{L}(\theta|\mathcal{D}) = \max_{\theta} \ln \prod_{i=1}^{N} p(\mathbf{v}_i|\theta) = \max_{\theta} \sum_{i=1}^{N} \ln p(\mathbf{v}_i|\theta) \tag{4.1}$$

where N is the size of $\mathcal{D}$.

The experiments on toy models suggest that the initial unsatisfactory results with naive mean field approaches [24] might be greatly improved if we include additional terms responsible for connections between the spins.

4.1.1. Training of Boltzmann Machines

With large graphical models, it is not possible to find an analytical solution to the maximum likelihood estimation of parameters and we need to resort to some approximation methods. That is also the case of the RBM and learning the parameters of this structure relies on the gradient ascent of the log-likelihood. At time t during training, the update of the vector containing all parameters of the RBM θ has the form:

$$\theta^t = \theta^{t-1} + \eta \frac{\partial}{\partial \theta^{t-1}} \ln \mathcal{L}(\theta|\mathcal{D}). \tag{4.2}$$

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-1}. By taking appropriately small learning rate, these iterative updates converge to stationary points. With large data set it is common to use a stochastic gradient ascent method [21] where we sample a minibatch of datapoints and take a noisy gradient estimate which results in the update rule:

$$\theta^{t+1} = \theta^t + \eta \frac{1}{M} \frac{\partial}{\partial \theta^t} \sum_{m=1}^{M} \ln \mathcal{L}(\theta|\mathbf{x}^{(m)}), \tag{4.3}$$

where M is the size of the minibatch. It can be shown that updates via 4.3 guarantee to converge to a local optimum under weak conditions [Bottou].

For a given data point $\mathbf{v}$ the log-likelihood can be seen as the difference between two energies:

$$\mathcal{L} = \ln P(\mathbf{v}) = -\ln(\sum_{\mathbf{h}} e^{-E(\mathbf{v},\mathbf{h})}) - \ln Z = F^c(\mathbf{v}) + F \tag{4.4}$$

where F is the *free energy* of the RBM and F^c denotes the clamped free energy as we operate on the fixed visible units $\mathbf{v}$. The gradient of the log-likelihood w.r.t θ given a training example $\mathbf{v}$ takes the form:

$$\begin{aligned}
\frac{\partial \log \mathcal{L}(\theta|\mathbf{v})}{\partial \theta} &= \frac{\partial F^c}{\partial \theta} - \frac{\partial F}{\partial \theta} \\
&= -\frac{\sum_{\mathbf{h}} e^{-E(\mathbf{v},\mathbf{h})} \frac{\partial E(\mathbf{v},\mathbf{h})}{\partial \theta}}{\sum_{\mathbf{h}} e^{-E(\mathbf{v},\mathbf{h})}} + \frac{\sum_{\mathbf{v},\mathbf{h}} e^{-E(\mathbf{v},\mathbf{h})} \frac{\partial E(\mathbf{v},\mathbf{h})}{\partial \theta}}{\sum_{\mathbf{v},\mathbf{h}} e^{-E(\mathbf{v},\mathbf{h})}} \\
&= -\sum_{\mathbf{h}} p(\mathbf{h}|\mathbf{v}) \frac{\partial E(\mathbf{v},\mathbf{h})}{\partial \theta} + \sum_{\mathbf{v},\mathbf{h}} p(\mathbf{v},\mathbf{h}) \frac{\partial E(\mathbf{v},\mathbf{h})}{\partial \theta} \\
&= -\mathbb{E}_{p(\mathbf{h}|\mathbf{v})} \left(\frac{\partial E(\mathbf{v},\mathbf{h})}{\partial \theta} \right) + \mathbb{E}_{p(\mathbf{v},\mathbf{h})} \left(\frac{\partial E(\mathbf{v},\mathbf{h})}{\partial \theta} \right)
\end{aligned} \tag{4.5}$$

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 Boltzmann machine, the clamped free energy can be computed explicitly. 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.

4.1.2. Monte Carlo methods

The second expectation from the gradient in 4.5 is intractable to compute explicitly in the case of large models and we have to resort to some kind of approximations. Monte Carlo methods rely on stochastic generations of random variables w.r.t. the desired expectation needs to be computed. Denote by:

$$\theta = \mathbb{E}_p(f(X)) = \int f(\mathbf{x})p(\mathbf{x})\mathrm{d}\mathbf{x}$$

the quantity of interest where $X \sim p(\cdot)$. The Monte Carlo estimate has the form:

$$\hat{\theta} = \frac{1}{N}\sum_{i=1}^{N} f(\mathbf{x}^i)$$

where $\mathbf{x}^i$, $i \in \{1,...,N\}$ are random samples from X and N is the number of samples. This simple procedure provides unbiased and consistent estimate of θ as $n \to \infty$.

Markov chain Monte Carlo

Monte Carlo method relies on the fact that we are able to generate independent random samples from the distribution of interest. In the case of the RBM, we are not able to generate random samples $\{\mathbf{v},\mathbf{h}\}$ from the complex joint posterior to approximate the expectation of interest. However, we can use Monte Carlo Markov chain (MCMC) framework to generate approximate samples from the joint distribution $p(\mathbf{v},\mathbf{h})$.

A discrete stochastic process $X = \{X_t, t \in \mathbb{N}\}$ which takes values in discrete set S is a Markov chain if the Markov property holds, i.e.

$$p_{ij}^t = P(X_t = j | X_{t-1} = i, ..., X_0 = i_0) = P(X_t = j | X_{t-1} = i)$$

for every $t \in \mathbb{N}$ and $i, j, i_0 \in S$. In the case of the discrete process, we usually operate on the transition matrix defined as $\mathbf{P} = (p_{ij})_{i,j \in S}$. The fundamental concept of the theory of the MCMC is stationarity or a stationary distribution π for which it holds $\pi = \mathbf{P}\pi$. MCMC methods focus on constructing an appropriate Markov chain that converges to the desired distribution.

Gibbs sampling

A particular class of MCMC algorithms is the Gibbs sampling algorithm which enables us to produce samples from the joint probability distribution using full conditional distributions. This method is also often called "block-at-a-time" as the transition probabilities are related with subblocks of the vector $\mathbf{x}$. Let $\mathbf{x}$ be divided into two blocks of variables $\mathbf{x}_1$ and $\mathbf{x}_2$. The

Gibbs sampler subsequently generates samples from $\mathbf{x}_1^i = p(\mathbf{x}_1|\mathbf{x}_2)$ and $\mathbf{x}_2^i = p(\mathbf{x}_2|\mathbf{x}_1)$ which forms samples from the joint $(\mathbf{x}_1^i, \mathbf{x}_2^i)$ assuming we reached a convergence of the chain.

In the case of the RBM, the structure of the model suggests that we can divide the variables from the joint into two blocks – visible and hidden units. No connections between variables from the same layer enables us efficiently sample from conditionals $p(\mathbf{v}|\mathbf{h})$ and $p(\mathbf{h}|\mathbf{v})$ using **??**.

4.1.3. Contrastive Divergence

The main challenge related with MCMC methods is the computational burden related with ensuring that the Markov chain has been run sufficiently long to ensure convergence to a stationary distribution. However, it was proven empirically that the chain might be run only a few steps in order to train an effective model [8] which is called contrastive divergence (CD) learning.

There are two steps which differ CD from the naive MCMC sampling to approximating the second expectation from the gradient 4.5. Firstly, instead of running the Markov chain until it obtains a stationary distribution, the chain is initialized using training data point $\mathbf{v}^0$ from the training data set. Secondly, the Gibbs chain is run only for k steps (CD-k) where k is usually smaller than 20. Figure 4.1 presents the procedure for the CD-1:

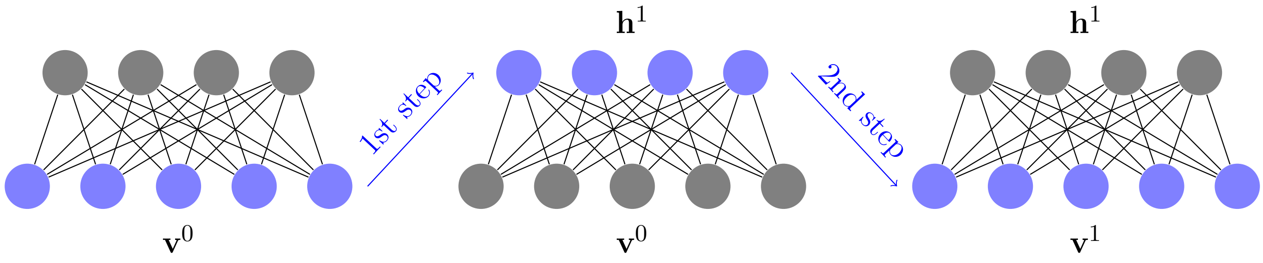

Fig. 4.1 The first step of the Gibbs sampler for the RBM for a particular data point $\mathbf{v}^0 \in \mathcal{D}$.

The approximation to the gradient by the single data point $\mathbf{v}^0$ in the case of CD-k takes the form:

$$-\sum_{\mathbf{h}} p(\mathbf{h}|\mathbf{v}^0)\frac{\partial E(\mathbf{v}^0,\mathbf{h})}{\partial\theta} + \sum_{\mathbf{h}} p(\mathbf{h}|\mathbf{v}^k)\frac{\partial E(\mathbf{v}^k,\mathbf{h})}{\partial\theta} \tag{4.6}$$

It should be noted here that as we run the Gibbs chain only a few (k) steps, the samples $\{\mathbf{v}^k, \mathbf{h}^k\}$ don't come from the stationary distribution and the approximation 4.6 is biased as it doesn't maximize the likelihood of the data but the difference of two KL-divergences [8], [5]:

$$KL(Q|P) - KL(P_k|P)$$

where Q is the empirical distribution and P_k is the distribution after k step of the Gibbs chain and this explains the name of the algorithm.

Persistent contrastive divergence

It was observed that the contrastive divergence procedure still requires many steps to be run in order to learn a good generative model. The rate of learning might be significantly improved when we don't reinitialize the Markov chains with a new training batch in order to obtain a sample $\{\mathbf{v}_i^k\}_{i=1}^N$ where N is the size of the batch but rather keep "persistent" chains (PCD) [24]. Thus, the starting state for the Gibbs chain is equal to the last step from the previous update. The assumption made here is that between parameter updates, the model changes only slightly in terms of parameters' values[17]. Thus, the initialization from the last state of the Gibbs chain taken from the previous model should be closer to the model distribution. The empirical results suggest to keep one persistent chain per one training data point in a batch.

4.1.4. Learning using extended mean field approximation

The stochastic procedure described in the previous section can be exchanged with the fully deterministic approach as the log-likelihood in the case of the EMF approximation has the form:

$$\mathcal{L} = \ln P(\mathbf{v}) = F^c(\mathbf{v}) - F^{EMF}. \tag{4.7}$$

As the first term from 4.4 can be computed explicitly, it is independent from the approach taken during training and we only have to derive the updates using the EMF approximation of the free energy.

Let's now fix visible and hidden magnetizations $\{\mathbf{m}^v, \mathbf{m}^h\}$. The gradient of the log-likelihood w.r.t a coupling parameter W_{ij} up to the third-order term is:

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

while the updates for the bias terms are just negative of the fixed-point magnetizations:

$$\begin{aligned}\frac{\partial F^{EMF}}{\partial a_i} &= -m_i^v, \\ \frac{\partial F^{EMF}}{\partial b_j} &= -m_j^h.\end{aligned} \tag{4.8}$$

Thus, the training procedure using a deterministic approach goes as follows: given a data point $\mathbf{v}$ we obtain expected values of the hidden units $\mathbf{h} = \mathrm{sigm}(W\mathbf{v} + \mathbf{b})$ which are starting points for magnetizations, i.e. $\mathbf{m}_0^v = \mathbf{v}$ and $\mathbf{m}_0^h = \mathbf{h}$. Then, we perform an iterative algorithm (which can have the form as presented in the previous chapter) until convergence to obtain magnetizations $\{\mathbf{m}^v, \mathbf{m}^h\}$ that satisfy self-consistency relations. Those magnetizations can then be used to obtain gradient w.r.t the parameters of the model and to compute the approximation of the free energy.

4.1.5. Approximating the log-likelihood

The problems related with intractability of the partition function makes training such structure very difficult as we cannot observe directly progress of learning. Thus, we need to resort to some approximations. One of the most popular approaches to measure progress in training RBMs is due to Besag [2] – consider the following approximation of n-dimensional distribution

$$P(\mathbf{x};\theta)=\prod_i p(x_i|x_1,...,x_{i-1};\theta)\approx\prod_i p(x_i|x_1,...,x_{i-1},x_{i+1},...,x_n;\theta)=\prod_i p(x_i|x_{-i};\theta):=PL(\mathbf{x};\theta) \tag{4.9}$$

where the first equation comes from the chain rule and x_{-i} denotes the set of all variables except variable x_i. We assume here that marginals given all other are independent of each other. The likelihood has then the form:

$$\ln PL(\mathbf{x};\theta)=\sum_i \ln P(x_i|x_{i-1};\theta). \tag{4.10}$$

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.

$$\ln PL(\mathbf{x};\theta)=n\ln P(x_i|\mathbf{x}_{-i};\theta), \tag{4.11}$$

where i is randomly chosen from $\{1,2,...,n\}$. It can be shown that this pseudo-likelihood is maximized by the true parameters of the model. In the case of the RBM, this estimator takes especially efficient form:

$$\ln PL(\mathbf{x};\theta)\approx n\log\left(\frac{\exp\{-F^c(\mathbf{x})\}}{\exp\{-F^c(\hat{\mathbf{x}})\}+\exp\{-F^c(\mathbf{x})\}}\right)=n\ln\left(\mathrm{sigm}(F^c(\hat{\mathbf{x}})-F^c(\mathbf{x})\right) \tag{4.12}$$

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

4.1.6. Real scale model – MNIST data set

The data set that will be used for the comparison and the evaluation of EMF and CD training algorithms is the MNIST set [13] which is a well-known benchmark image classification dataset that consists of 60000 training and 10000 testing images of digit numbers. They are represented on 28-by-28 grey-scale grid of pixels. Thus, the first visible layers in all analysed models consists of 784 visible units. Following [6], [22] all images were rescaled to $\{0,1\}$ and binarized by setting all non-zero pixels to 1 in all experiments. The data set was divided into 600 mini-batches which results in 100 training points per batch.

4.1.7. Comparison of both approaches

In order to test the efficiency of the EMF learning algorithm, I used three expansions of **??** – up to the first-order (MF), second-order (TAP2) and third order (TAP3) term. Moreover, I varied the number of iterations of self-consistency relations (3 and 10) using asynchronous updates of the form 3.5 to mimic the idea from the contrastive divergence approach. As a benchmark, two models were trained following the stochastic training (CD1, CD10).

Furthermore, all models described above were trained using persistent approach (PMF, PTAP2, PTAP3, PCD). In the case of the EMF approximation, the magnetizations of a batch from the previous update are the starting points in the next update [6]. Similarly to PCD, this idea is based on the fact that between updates the model changes only slightly and it should improve the convergence to the new fixed point magnetizations.

All models were trained 10 times using the same set-up of free parameters with 500 units. The purpose of this experiment is to compare different RBM trainings thus following [6] I didn't use the adaptive learning rate which was set to 0.005, learning was performed using mini-batch updates 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. Thus, in general couplings should have small values – this can be enforced using regularization which at the same times allows for a better generalization. From probabilistic perspective this can be seen as adding a weighted prior over the parameters (maximum a posteriori training). The criterion that will be maximized has now the form:

$$E(\theta,\mathcal{D}) = \ln \mathcal{L}(\theta|\mathcal{D}) - \lambda R(\theta) \tag{4.13}$$

where $R(\cdot)$ is the regularizer and $\lambda \in \mathbb{R}_+$ is a hyper-parameter which controls the effective power of the regularization. In all experiments Laplacian prior $R(\theta) = \|\theta\|_1$ ($L1$ regularization) was used with λ set to 0.01.

Figure 4.2 presents the pseudo log-likelihood 4.12 (left) and EMF log-likelihood 4.7 for the non-persistent training procedure. Firstly, by the visual inspection both approximation yield very similar results for each analysed model. However, the EMF estimates are much less noisy at a lower computational cost.

Fig. 4.2 Per-sample pseudo log-likelihood (left) and EMF log-likelihood (right) on the validation set of the MNIST data set divided by number of all units in the model (1284) across first training 50 epochs for RBMs models trained stochastically and deterministically. Error bars shows the standard deviations of 10 trained models using a particular version of training.

Secondly, results for the MF-10[1] confirms the findings from the literature – the naive mean field approach is not able to learn an effective model. Moreover, the results for the CD, TAP2 and TAP3 are very similar. There are not significant differences between models with 3 or 10 iterations of self-consistency relations which shows that the deterministic approach is not computationally expensive.

As it was expected, the best results in terms of the EMF log-likelihood are achieved by EMF methods. However, the results for the CD models suggest that the EMF log-likelihood may be used as a reliable indicator of progress during training as those models weren't constructed to optimize over this objective [6].

Figure 4.3 presents the results for persistent versions of models analysed above. There are not significant differences comparing to However, as it was expected the samples from the models trained using persistent chains are of much higher quality.

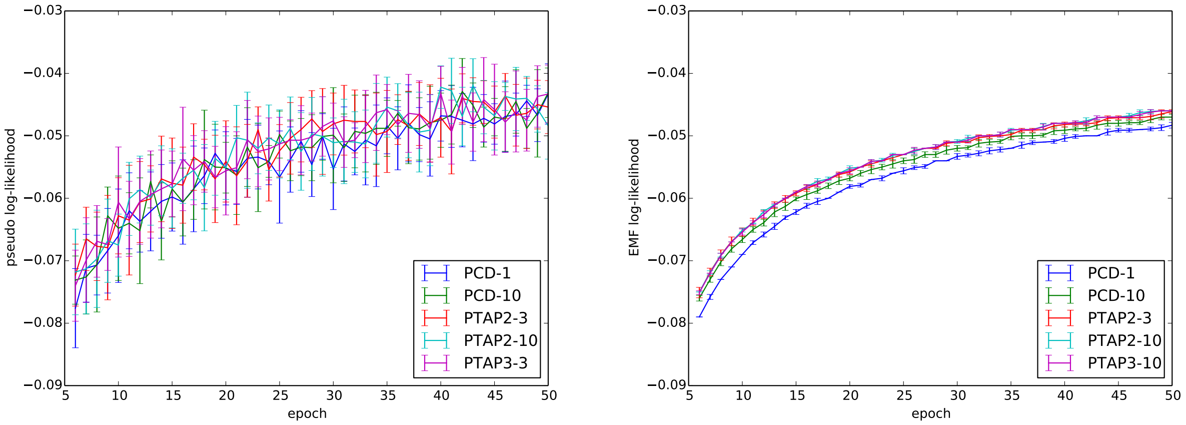

Fig. 4.3 Per-sample pseudo log-likelihood (left) and EMF log-likelihood (right) for the same models trained using persistent Gibbs chains. Results of the naive method weren't included.

Finally, in persistent and non-persistent versions of models the addition of the third order term from the EMF expansion 2.24 doesn't provide improvement over the TAP model. This might be partially explained by the fact that estimated weights are in general smaller than 1 (in absolute value) which are then used at the order of 3 in self-consistency equations and hence don't affect significantly estimations.

4.1.8. Generated samples from the models

The restricted Boltzmann machine can also be seen as a generative model that can creates new samples that will resemble the observed data provided that it approximates well the unknown distribution that generated the original, observed samples. Three models were trained with 500 hidden units using persistent chains – with the contrastive divergence algorithm, and two models using naive and extended mean field approach respectively. Figure

---

[1] The results for MF-3 weren't included as it was very similar to the MF-10

4.4 shows samples generated from those models where between samples 1000 chain steps were performed. The samples produced by the naive mean field methos provides the samples of

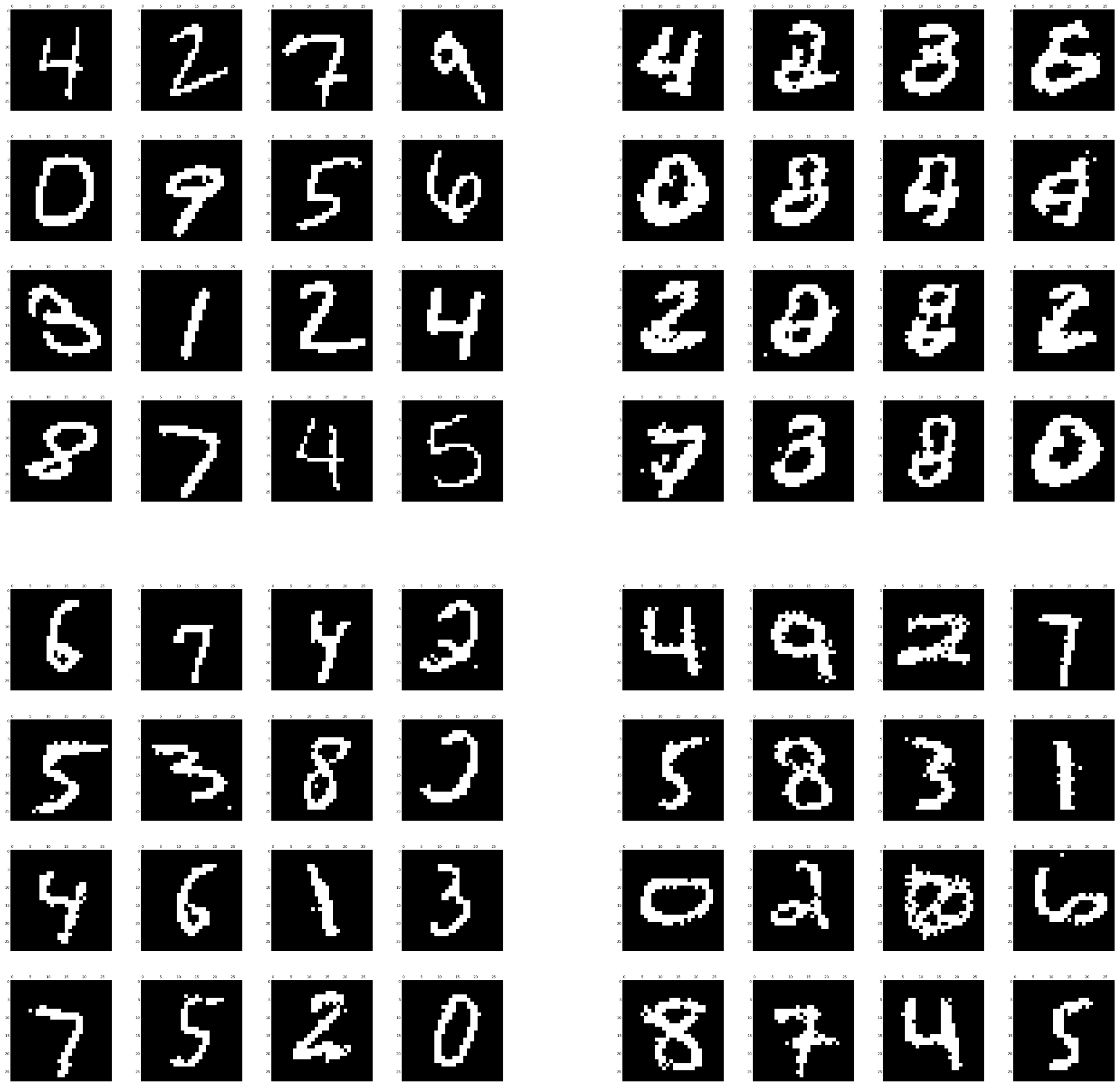

Fig. 4.4 Comparison of samples generated by 500 hidden unit RBM trained using naive mean field approach (top right), extended mean field approximation up to the second-term order (bottom right) and with contrastive divergence (bottom left) to the original digits from the MNIST data (top left). All models were trained using persistent chains.

the poorest quality that are often non-identifiable as one of 10 digits. As we can see, the addition of the second-order term greatly improves also the generetive qualitative of the learned model. It might be argued that samples produced by the PCD and EMF models are of similar quality.

Chapter 5

Various experiments

5.1. Comparison of schedules of updates

In the chapter 2 different schedules of updates were analysed on toy models where the parameters of the model were known *a priori* and no substantial discrepancies were observed in terms of the quality of approximation between asynchronous and parallel schedules. Taking into consideration the performance of the sequential updates on both toy models, this schedule wasn't considered in the evaluation on the real data set.

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 with both magnetizations and parameters – this may lead to substantial differences in performance. Figure 5.1 shows the normalized estimates of log-likelihood of three models trained using the extended mean field approximation up to the first-order term (MF), up to the second-order term (TAP2) and up to the third-order term (TAP3), each trained with two nchedules of updates.

Surprisingly, only with the naive mean field approximation we can observe that the parallel schedule provides slightly better results in terms of the approximated log-likelihood. In general, there are no significant differences between two considered schedules even though the parallel form of updates doesn't reflect the way both layers of the RBM 'communicate' with each other. Moreover, it seems that small number of fixed point iterations doesn't deteriorate the performance. This suggests that asynchronous updates with only 3 iterative updates of magnetizations yields consistently competitive results at the same time being the most computationally inexpensive form of schedule.

5.1.1. Evaluation of EMF approximation

The similiarity of the estimates of the log-likelihood using the pseudo log-likelihood method and with the extended mean field approach suggests that the high-temperature expansion

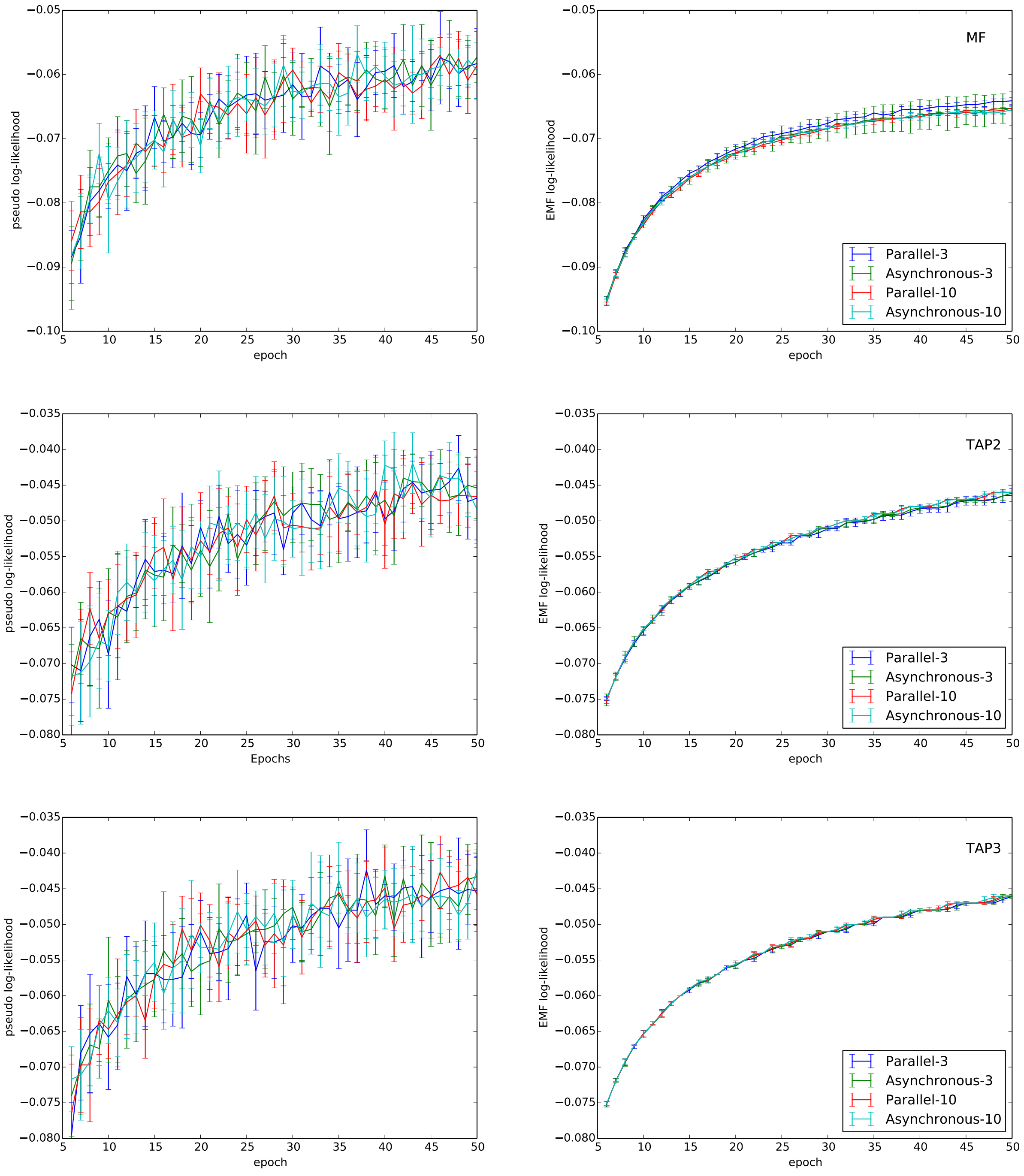

Fig. 5.1 Per-sample pseudo log-likelihood (left) and EMF log-likelihood (right) on the validation set of the MNIST data divided by number of all units in the model (1284) across first 50 training epochs for RBMs models trained with different schedules and number of updates. Error bars show standard deviation of 10 trained models using a particular version of training.

method might be a computationally inexpensive metric of the partition function and the log-likelihood in general for RBMs. However, the pseudo log-likelihood is a very noisy approximation thus as a benchmark for comparison of the effectiveness of the EMF-based estimates I will use the annealed importance sampling (AIS) [18] – very popular method for computing the partition function. Naturally, this comparison assumes that the AIS estimate are close to the true value which in general might no be the case.

5.2. Annealed Importance Sampling

The AIS method is based on a very simple identity – assume we have two distributions $p_A = \frac{1}{Z_A}p^*_A(\mathbf{x})$ and $p_B = \frac{1}{Z_B}p^*_B(\mathbf{x})$, where $p^*(\dot{)}$ denotes an unnormalized distribution and Z_A, Z_B are partition functions of distribution p_A and B respectively. If the proposal 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:

$$\begin{aligned} Z_B &= \int p^*_B(\mathbf{x})\mathrm{d}\mathbf{x} \\ &= \int \frac{p_A(\mathbf{x})}{p_A(\mathbf{x})} p^*_B(\mathbf{x})\mathrm{d}\mathbf{x} \\ &= Z_A \int \frac{p^*_B(\mathbf{x})}{p^*_A(\mathbf{x})} p_A(\mathbf{x})\mathrm{d}\mathbf{x}. \end{aligned} \tag{5.1}$$

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

$$\frac{Z_B}{Z_A} \approx \frac{1}{N}\sum_{i=1}^{N} \frac{p^*_B(\mathbf{x}^{(i)})}{p^*_A(\mathbf{x}^{(i)})} = \hat{r}_{SIS} \tag{5.2}$$

where $\mathbf{x}^{(i)}$ comes from p_A. Assuming that distribution p_A is close to p_B, the estimator from 5.2 called simple importance sampling proves to work well [Minka et al.]. However, in high-dimensional spaces where p_B is usually multimodal as it is considered in this thesis the variance of the estimator from 5.2 might be very high.

The idea presented above might be improved by following the classic approach from the 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 [10], [18].

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 5.2 can be used to estimate each ratio $\frac{Z_{m+1}}{Z_m}$. Using the identity:

$$\frac{Z_M}{Z_0} = \frac{Z_1}{Z_0}\frac{Z_2}{Z_1}...\frac{Z_M}{Z_{M-1}} \tag{5.3}$$

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.:

$$\int T_m(\mathbf{x}^{'}|\mathbf{x})p_m(\mathbf{x})\mathrm{d}\mathbf{x} = p_m(\mathbf{x}^{'}) \tag{5.4}$$

These transition operators represent the probability density of transitioning from state $\mathbf{x}$ to $\mathbf{x}^{'}$ [22]. 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 $T_M(\mathbf{x}_M|\mathbf{x}_{M-1})$
    $r_{AIS}^{(i)} = \frac{p_1^*(\mathbf{x}_1)}{p_0^*(\mathbf{x}_1)}\frac{p_2^*(\mathbf{x}_2)}{p_1^*(\mathbf{x}_2)}\cdots\frac{p_M^*(\mathbf{x}_M)}{p_{M-1}^*(\mathbf{x}_M)}$
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 [18]. 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 [23] 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:

$$p_m(\mathbf{v}) = \frac{1}{Z_m}p_m^*(\mathbf{v}) = \frac{1}{Z_m}\sum_{\mathbf{h}}\exp(-E_m(\mathbf{v},\mathbf{h})) \tag{5.5}$$

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

$$E_m(\mathbf{v},\mathbf{h}) = (1-\beta_m)E(\mathbf{v};\theta_A) + \beta_m E(\mathbf{v},\mathbf{h};\theta_B) \tag{5.6}$$

where $\beta_m \in [0,1]$ with $\beta_m = 0$ yielding p_A and $\beta_m = 1$ giving p_B. By 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.

5.2.1. Comparison

Two models were estimated based on the extended mean field approximation – up to the second-order term (TAP2), and up to the third-order term (TAP3), to compare the quality of the approximation of the variational free energy. Each model was re-estimated 10 times using persistent chains with 10 iterations of self-consistency relations with asynchronous updates. Taking into consideration the inherent variability of the AIS method, 10 runs of AIS were performed to obtain an average estimate. A sequence of βs is required to set the "tempo" of annealing – following [22] 1000 β_k were spaced uniformly from 0 to 0.5, $4,000$ β_k were spaced uniformly from 0.5 to 0.9, and $5,000$ β_k were spaced uniformly from 0.9 to 1.0, with a total of 10,000 intermediate distributions.

Firstly, both methods were compared on a small toy model where we can compute the true partition function explicitly – an RBM with 15 hidden units. Figure 5.2 presents the true value of the parition function along with its estimates from the TAP2 and TAP3 models and using the AIS method.

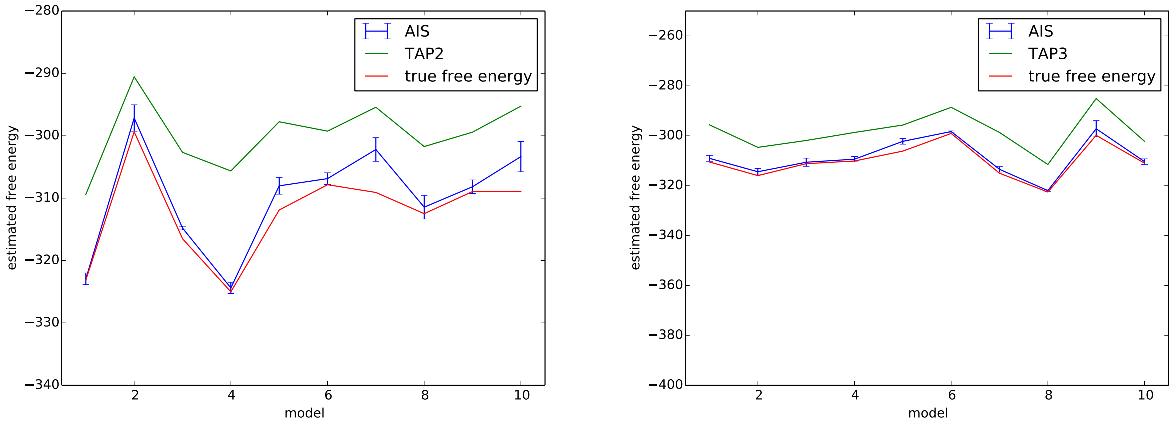

Fig. 5.2 Free energy estimates using two forms of extended mean field approximations and AIS estimates for 10 trained models.

The results indicate that the AIS works rather well as an approximator of the true free energy. However, the EMF approximation tends to be biased and the addition of the third-order term improves the estimates only slightly. In both cases the approximation consistently overestimates the true value of $\mathcal{F}$ on average by about 5%.

Having justified the usage of the AIS estimates as a benchmark, we can compare both methods on the RBM with 500 hidden units. As a sanity check, the AIS was re-estimated in this case 100 times as the results from **??** show that the variance is then lower than 1%. Figure 5.2 presents the estimates of **??** for the TAP2 and TAP3 models along with AIS estimates.

Firstly, as it was expected the learned models using two high-energy expansion yield very similar estimates of the free energy. Secondly, in both cases they give consistently biased upper bound approximation for the $\mathcal{F}$, assuming that the AIS method gives an accurate estimation of **??**. As in the case of the toy model, the difference is about 5% in value on

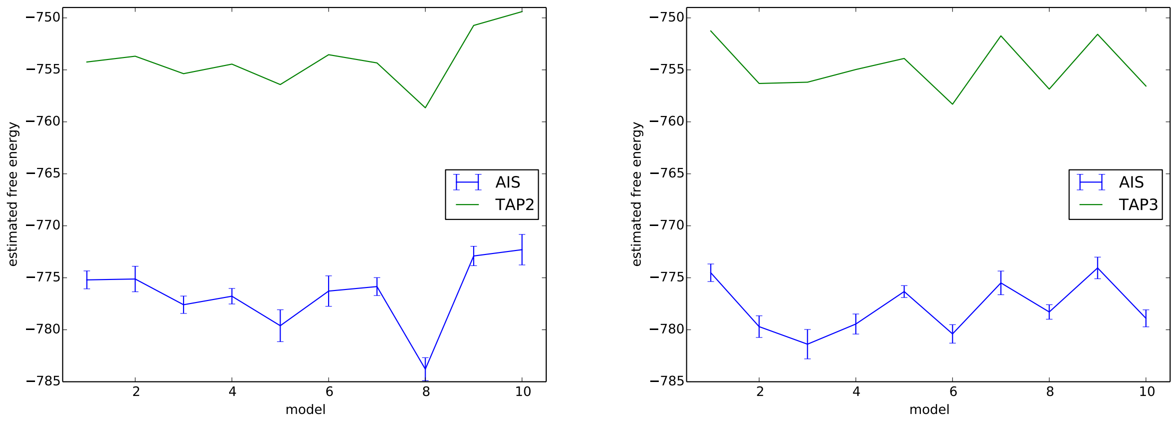

Fig. 5.3 Free energy estimates using two forms of extended mean field approximations and AIS estimates for 10 trained models.

average. The mean squared error between the average AIS estimate and the TAP2 method was 505.747 while for the the approximation including third-order the MSE was 534.193. This suggests that even though the extended mean field approximation enable us to learn good generative models, the approximation of the free energy is very biased. However, the computational cost of estimation is about 10^{-5} smaller comparing to the AIS.

5.3. Deep structures

5.3.1. Unsupervised pre-training of deep neural networks

Recent major advances in tasks like speech recognition, vision recognition or machine translation were obtained thanks to effective training of deep neural structures. Deep learning methods allow to learn high-level features based on the composition of lower level characteristics. However, the objective function in general is highly non-convex and multimodal over high-dimensional parameter space. This naturally leads to many challanges in training such deep structures.

The breakthrough that lead to effective training strategies for deep architectures came in 2006 with the CD algorithm for training deep belief networks (DBN) [9]. The deep structure might be formed by stacking single RBMs pre-trained using contrastive divergence approach. It was empirically shown that unsupervised pre-training guides the learning towards basins of attraction of minima that support better generalization from the training data set **??**. Thus, such pre-trained networks much better generalize and are more robust comparing to networks trained only in supervised fashion.

In the previous chapter it was shown that unsupervised learning using the extended mean-field approach can be used to learn a good quality generative model of the observed phenomena.

Those results motivate learning deep structures using layer-wise greedily pre-training where instead of using CD for training RBMs we might consider fully deterministic approach.

5.3.2. Deep belief nets

DBNs are generative graphical models with many hidden layers of hidden causal variables which joint distribution has the following form:

$$p(\mathbf{x},\mathbf{h}^1,\mathbf{h}^2,...,\mathbf{h}^l) = p(\mathbf{x}|\mathbf{h}^1)P(\mathbf{h}^1|\mathbf{h}^2)...P(\mathbf{h}^{l-2}|\mathbf{h}^{l-1})P(\mathbf{h}^{l-1}|\mathbf{h}^l). \tag{5.7}$$

It was shown that adding an extra layers always improve a lower bound on the log-likelihood on the training data if the number of feature detectors per layer is sufficiently large and the weights are initialized correctly. 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. Figure 5.4 depicts the exemplary deep belief network.

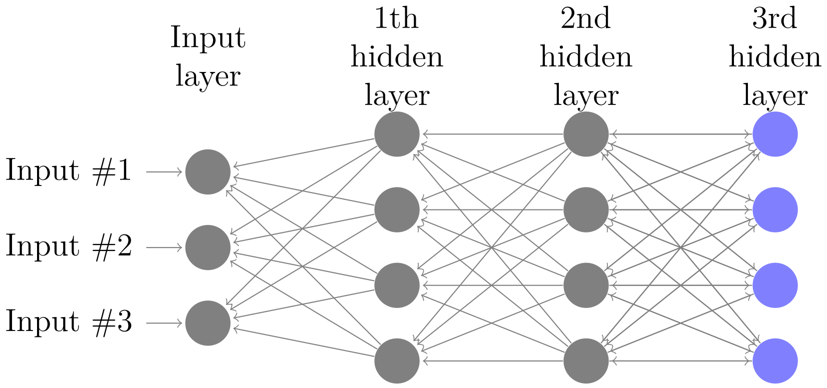

Fig. 5.4 An exemplary deep belief net with 3 hidden layers where the last two layers form a single RBM.

Similarly to the case of the single RBM, DBN may be seen as a general approximator of any probability distribution on $\{0,1\}^n$ and we can prove [15] :

Theorem 3 (Guido-Ay, 2010) *Let* $n = \frac{2^b}{2} + b,\ b \in \mathbb{N},\ b \geqslant 1$. *A DBN containing* $\frac{2^n}{2(n-b)}$ *hidden layers of size* n *is a universal approximator of distributions on* $\{0,1\}^n$.

DBNs can be formed using a greedy layer-wise unsupervised training of stacked RBMs – Algorithm 2 presents how the process folows. This simple and intuitive algorithm proved to be an effective way of pre-training deep structures that laid the foundations for the resurgence of deep neural networks. Originally, the building blocks are trained following constrastive divergence procedure. However, the positive results obtained using the extended mean-field approximation suggests that we may follow Algorithm 2 with fully deterministic approach.

Algorithm 2 Learning procedure for 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
```

5.3.3. Reconstruction analysis

After pre-training multiple layers of feature detectors, the model can be "unfolded" to form an autoencoder structure 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 **??**.

To motivate the analysis of deep structures formed of stacked RBMs, I trained different three models that compress the information in the data to 25 dimensional manifold – using principal component analysis with first 25 dimension with the highest variance, Check, a single RBMs with 25 hidden units and a deep belief net that consists of three hiddent layers of sizes 500, 250 and 25. Both models were trained using an extended mean field approach up to the second-order term with persistent chains.

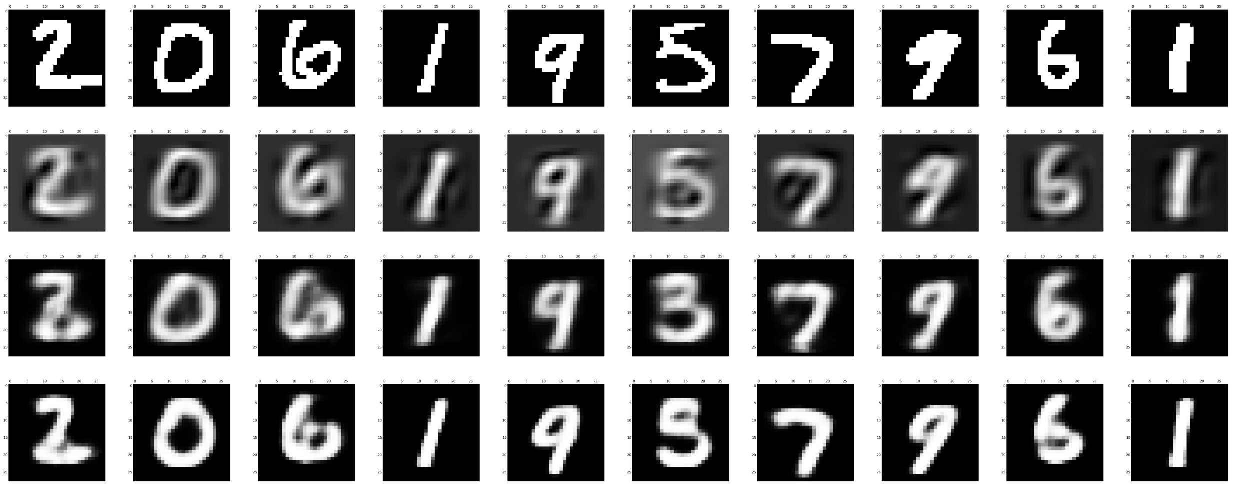

Fig. 5.5 Reconstructions of MNIST digits (top row) generated by PCA-25 (second row), single RBM with 25 hiddent units (third row) and deep belief net (bottom row).

As we can see, the linear transformation to the small dimensional space creates very noisy encoding system. The reconstructions show that the numbers lie very closely on the space from which we decode. The reconstructions provided by the single RBM substanitally improves the quality of the reconstructions and it can be infer that the learned manifold better differentiate between the digits. However, although the background is properly delimited from the numbers, digits are still very blurry. It might be argued that the deep structure creates the most sharply-outlined numbers. From this we might infer that deep belief net is able to creates more coarser and coarser features that in the process of the encoding are able to generate clear-cut digits as the learned meanifold in the smallest layer is properly partitioned.

Similarly to the case of the shallow structures, we might compare the effectiveness of the EMF approach to the CD learning in the case of the greedily layer-wise training of DBN. Figure 5.6 presents the reconstructions of randomly chosen samples from the validation data set produced by deep autoencoders trained with four different methods of pre-training DBNs. The autoencoder consists of three hidden layers of sizes 500, 250 and 25 accordingly. Extended mean field approximation was considered up to the first-order (MF), second-order (TAP2) and third-order (TAP3) terms. One model was also trained using CD procedure. At each layer 50 updates through the entire data sets were performed using 10 iterations of asynchronous updates. In each case magnetization or Gibbs chains were persistent.

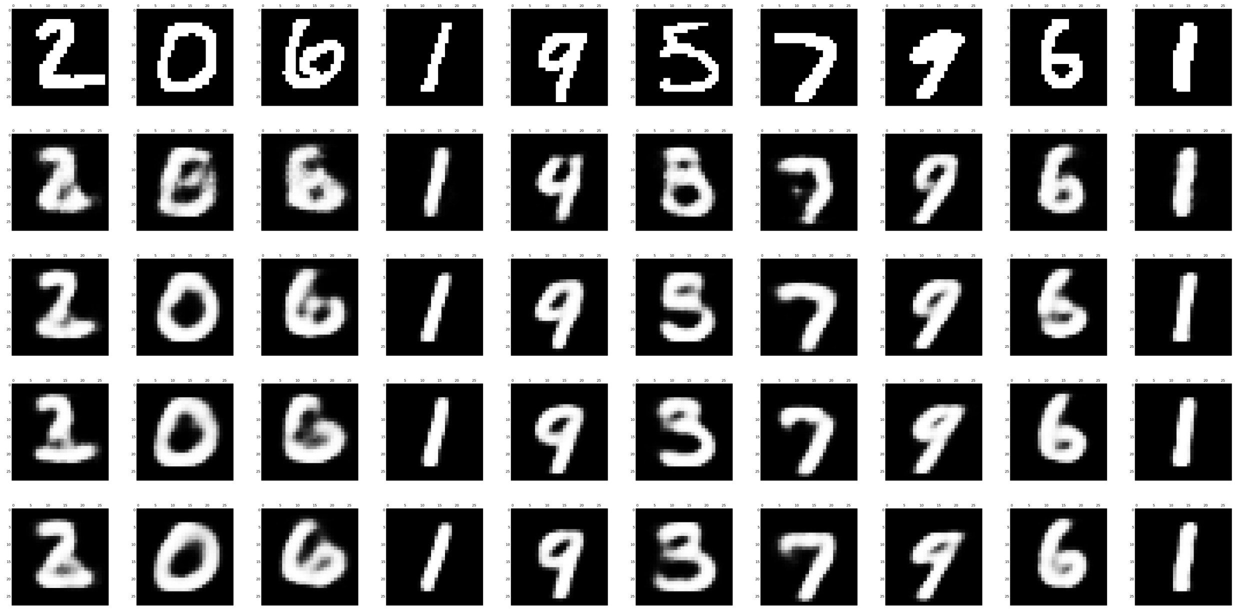

Fig. 5.6 Reconstructions of MNIST digits (top row) generated by four different deep belief nets trained using naive mean field approach (second row), EMF up to the second-order term (third row), EMF up to the third-order term (fourth row) and with PCD (bottom row).

By the visual inspection, it might be argued that the reconstruction created by TAP2, TAP3 and PCD are of similar quality and they are more identifiable than those produced by the MF. It can be observed how the autoencoder learnt by EMF or PCD recovers a a smoothed version of the original digit – an "average" representative of a given number. Surprisingly, the addition of the third-order term leads somewhat to deterioration of the quality in reconstructions which can be observed especially in the case of the first (2) and sixth (5) number. The average squared errors on training and validation data sets (Figure 5.7) confirms the visual assessment of reconstructions. The mean field approximation obtains the highest score while TAP2 and TAP3's scores are slightly higher than with training DBN using PCD approach.

Those results confirms the observations from the previous chapter and shows that additional higher-order approximations substantially improves the quality of learned magnetizations which in turns helps learning a better generative model.

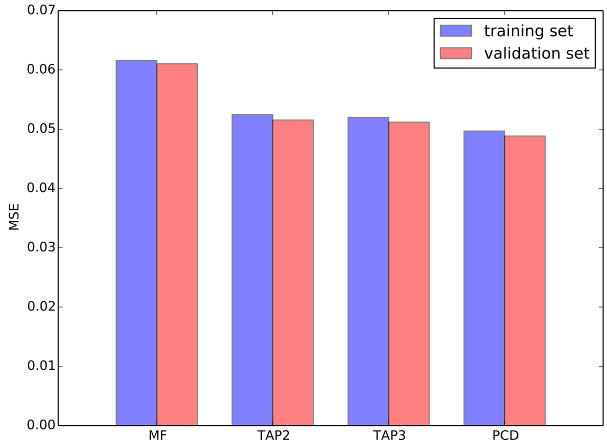

Fig. 5.7 MSE of reconstructions for four different models on training and validation sets.

Conclusions

The analysed paper provides a clear introduction to a new methodology for performing variational inference. We managed to obtain the same results as the authors on both data sets which was partially possible thanks to the straightforward explanation of the proposed method and the detailed description of the experiments conducted. We found the model structure very robust to changes of parameters of the network. Moreover, our experiments show that the performance of VAEB is superior to that of the vanilla AE architecture and it is resistant to superfluous latent variables thanks to automatic regularisation via the KL-term. Our implementation of variational inference on both the parameters θ and the latent variables $\mathbf{z}$ performed disappointingly poorly which might be partially explained by overly-restrictive prior and we plan to further investigate this problem. The code needed to replicate our results is available at https://github.com/budzianowski/VAEB. We found this project incredibly interesting and we hope to undertake further work on the extensions proposed in the previous section.

References

[1] Ackley, D. H., Hinton, G. E., and Sejnowski, T. J. (1985). A learning algorithm for boltzmann machines. *Cognitive science*, 9(1):147–169.

[2] Besag, J. E. (1972). Nearest-neighbour systems and the auto-logistic model for binary data. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 75–83.

[Bottou] Bottou, L. Online learning and stochastic approximations. *On-line learning in neural networks*, 17(9):142.

[4] Erhan, D., Bengio, Y., Courville, A., Manzagol, P.-A., Vincent, P., and Bengio, S. (2010). Why does unsupervised pre-training help deep learning? *Journal of Machine Learning Research*, 11(Feb):625–660.

[5] Fischer, A. and Igel, C. (2012). An introduction to restricted boltzmann machines. In *Iberoamerican Congress on Pattern Recognition*, pages 14–36. Springer.

[6] Gabrié, M., Tramel, E. W., and Krzakala, F. (2015). Training restricted boltzmann machine via the thouless-anderson-palmer free energy. In *Advances in Neural Information Processing Systems*, pages 640–648.

[7] Georges, A. and Yedidia, J. S. (1991). How to expand around mean-field theory using high-temperature expansions. *Journal of Physics A: Mathematical and General*, 24(9):2173.

[8] Hinton, G. E. (2002). Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800.

[9] Hinton, G. E. and Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507.

[10] Jarzynski, C. (1997). Nonequilibrium equality for free energy differences. *Physical Review Letters*, 78(14):2690.

[11] Kirkpatrick, S. and Sherrington, D. (1978). Infinite-ranged models of spin-glasses. *Physical Review B*, 17(11):4384.

[12] Le Roux, N. and Bengio, Y. (2008). Representational power of restricted boltzmann machines and deep belief networks. *Neural computation*, 20(6):1631–1649.

[13] LeCun, Y. and Cortes, C. (1998). The mnist database of handwritten digits. Technical report.

[Minka et al.] Minka, T. et al. Divergence measures and message passing. Technical report.

[15] Montufar, G. and Ay, N. (2010). Refinements of universal approximation results for deep belief networks and restricted boltzmann machines. *arXiv preprint arXiv:1005.1593*.

[16] Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT Press.

[17] Neal, R. M. (1992). Connectionist learning of belief networks. *Artificial intelligence*, 56(1):71–113.

[18] Neal, R. M. (2001). Annealed importance sampling. *Statistics and Computing*, 11(2):125–139.

[19] Opper, M. and Saad, D. (2001). *Advanced mean field methods: Theory and practice*. MIT press.

[20] Plefka, T. (1982). Convergence condition of the tap equation for the infinite-ranged ising spin glass model. *Journal of Physics A: Mathematical and general*, 15(6):1971.

[21] Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407.

[22] Salakhutdinov, R. (2008). Learning and evaluating boltzmann machines. Technical report.

[23] Salakhutdinov, R. and Murray, I. (2008). On the quantitative analysis of deep belief networks. In *Proceedings of the 25th international conference on Machine learning*, pages 872–879. ACM.

[24] Tieleman, T. (2008). Training restricted boltzmann machines using approximations to the likelihood gradient. In *Proceedings of the 25th international conference on Machine learning*, pages 1064–1071. ACM.

[25] Wainwright, M. J. and Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1-2):1–305.

[26] Welling, M. and Hinton, G. E. (2002). A new learning algorithm for mean field boltzmann machines. In *International Conference on Artificial Neural Networks*, pages 351–357. Springer.

[27] Yedidia, J. (2001). An idiosyncratic journey beyond mean field theory. *Advanced mean field methods: Theory and practice*, pages 21–36.

s

Appendix A

EMF derivation

Following [7] lets define energy as:

$$E = -\sum_{ij} w_{ij} s_i s_j - \sum_i \theta_i s_i \tag{A.1}$$

and introduce the following operator:

$$U \equiv E - \mathbb{E}(E) - \sum_i \frac{\partial \lambda_i(\beta)}{\partial \beta}(s_i - m_i) \tag{A.2}$$

which poses useful property – $\mathbb{E}(U) = 0$. For any other operator O we then have:

$$\frac{\partial \mathbb{E}(O)}{\partial \beta} = \mathbb{E}\left(\frac{\partial O}{\partial \beta}\right) - \mathbb{E}(OU). \tag{A.3}$$

Now, the first derivative from the Taylor expansion is:

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

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

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

Using A.3 we obtain:

$$\frac{\partial m_i}{\partial \beta} = 0 = \frac{\partial \mathbb{E}(s_i)}{\partial \beta} = \mathbb{E}\left(\frac{\partial s_i}{\partial \beta}\right) - \mathbb{E}(s_i U) = -\mathbb{E}(s_i U) = -\mathbb{E}(U(s_i - m_i)). \tag{A.4}$$

The first derivative of the operator U has the form:

$$\begin{aligned}\frac{\partial U}{\partial\beta} &= \frac{\partial E}{\partial\beta}-\frac{\partial\mathbb{E}(E)}{\partial\beta}-\sum_i\frac{\partial^2\lambda_i(\beta)}{\partial\beta^2}(s_i-m_i)\\ &= \mathbb{E}(U^2)-\sum_i\frac{\partial^2\lambda_i(\beta)}{\partial\beta^2}(s_i-m_i)\end{aligned}\tag{A.5}$$

and the second derivative is:

$$\begin{aligned}\frac{\partial^2 U}{\partial\beta^2} &= 2\mathbb{E}\left(\frac{\partial U}{\partial\beta}U\right)-\mathbb{E}(U^3)-\sum_i\frac{\partial^3\lambda_i(\beta)}{\partial\beta^3}(s_i-m_i)\\ &= -\mathbb{E}(U^3)-\sum_i\frac{\partial^3\lambda_i(\beta)}{\partial\beta^3}(s_i-m_i)\end{aligned}\tag{A.6}$$

The expansion of free energy using formulas derived above can now be reformulated in terms of the operator U:

$$\frac{\partial(\beta F)}{\partial\beta}=\mathbb{E}(E)-\sum_i\frac{\partial\lambda_i(\beta)}{\partial\beta}\mathbb{E}(s_i-m_i)=\mathbb{E}(E)\tag{A.7}$$

and the higher orders:

$$\begin{aligned}\frac{\partial^2(\beta F)}{\partial\beta^2} &= \mathbb{E}\left(\frac{\partial E}{\partial\beta}\right)-\mathbb{E}(EU)=-\mathbb{E}(U^2),\\ \frac{\partial^3(\beta F)}{\partial\beta^3} &= -2\mathbb{E}\left(U\frac{\partial U}{\partial\beta}\right)+\mathbb{E}(U^3)=\mathbb{E}(U^3).\end{aligned}\tag{A.8}$$

Taylor expansion was considered around point $\beta=0$. Using derivations from A.7 we obtain again a 'naive' term:

$$\left.\frac{\partial(\beta F)}{\partial\beta}\right|_{\beta=0}=\mathbb{E}_{\beta=0}(E)=-\sum_{(ij)}w_{ij}m_im_j-\sum_i\theta_im_i=-\frac{1}{2}\sum_i\sum_j w_{ij}m_im_j-\sum_i\theta_im_i.\tag{A.9}$$

Consider now:

$$\left.\frac{\partial(\beta F)}{\partial m_i\partial\beta}\right|_{\beta=0}=-\sum_{j\neq i}w_{ij}m_j-\theta_i.\tag{A.10}$$

On the other hand:

$$\left.\frac{\partial(\beta F)}{\partial m_i\partial\beta}\right|_{\beta=0}=\left.\frac{\partial(\beta F)}{\partial\beta\partial m_i}\right|_{\beta=0}=\frac{\partial}{\partial\beta}\mathbb{E}(\lambda_i(\beta))=\left.\frac{\partial\lambda_i(\beta)}{\partial\beta}\right|_{\beta=0}\tag{A.11}$$

Substituting A.11 into A.2 gives us:

$$\begin{aligned}
U_{\beta=0} &= -\sum_{(ij)} w_{ij} s_i s_j - \sum_i \theta_i s_i + \frac{1}{2}\sum_i \sum_j w_{ij} m_i m_j + \sum_i \theta_i m_i + \sum_i \left(\sum_{j \neq i} w_{ij} m_j + \theta_i \right) (s_i - m_i) \\
&= -\sum_{(ij)} w_{ij} s_i s_j - \frac{1}{2}\sum_{(ij)} w_{ij} m_i m_j + \sum_i \sum_{j \neq i} w_{ij} s_i m_j \\
&= -\sum_{(ij)} w_{ij} (s_i - m_i)(s_j - m_j) = -\sum_l w_l y_l
\end{aligned} \tag{A.12}$$

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

$$\begin{aligned}
\mathbb{E}(y_l)_{\beta=0} &= \mathbb{E}(s_i s_j) - m_j \mathbb{E}(s_i) - m_i \mathbb{E}(s_j) + m_i m_j = 0 \\
\mathbb{E}(y_l (s_i - m_i))_{\beta=0} &= m_j - m_j - m_i^2 m_j + m_i^2 m_j \\
&\quad - m_i^2 m_j + m_i^2 m_j + m_i^2 m_j - m_i^2 m_j \\
&= 0.
\end{aligned} \tag{A.13}$$

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:

$$\begin{aligned}
\mathbb{E}((s_i - m_1)^2 (s_j - m_j)^2) &= m_i m_j - 2m_i m_j^2 + m_i m_j^2 - 2m_i^2 m_j + 4m_1^2 m_j^2 \\
&\quad - 2m_i^2 m_j^2 + m_i^2 m_j - 2m_i^2 m_j^2 + m_i^2 m_j^2 \\
&= (m_i - m_i^2)(m_j - m_j^2).
\end{aligned} \tag{A.14}$$

Using properties from y_l in equations A.8 we can derive:

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

which yields the TAP-Onsager term. To obtain the next term in 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 only the case when $l_1 = l_2 = l_3$:

$$\begin{aligned}
\mathbb{E}((s_i - m_1)^3 (s_j - m_j)^3) &= m_i m_j - 3m_i m_j^2 + 2m_i m_j^2 + 2m_i m_j^3 - 3m_i^2 m_j + 2m_i^3 m_j \\
&\quad + 9m_i^2 m_j^2 - 6m_i^3 m_j^2 - 6m_i^2 m_j^3 + 4m_i^3 m_j^3 \\
&= 4(m_i - m_i^2)(\frac{1}{2} - m_i)(m_j - m_j^2)(\frac{1}{2} - m_j)
\end{aligned} \tag{A.15}$$

and the third-order term in the case of the RBM structure:

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