University of Cambridge

Thesis

Paweł Budzianowski, pfb30, Clare Hall College

Contents

1. Introduction

2. Extended mean field approximation

2.1. Graphical models as Markov random fields

One of the basic concepts in 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:*

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

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

Thus, the MRF has a desired property that any two nodes are conditionally independent given some evidence nodes that 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:

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

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 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:

$$p(x_1, x_2, ..., x_n) = \frac{1}{Z}\exp\left(-\frac{1}{T}E(x_1, x_2, ..., x_n)\right) \tag{2}$$

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

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

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 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 [18]. 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.

2.3. Statistical 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:

$$P(\mathbf{s}) = \frac{e^{-\frac{1}{T}E(\mathbf{s})}}{\sum_{\mathbf{s}} e^{-\frac{1}{T}E(\mathbf{s})}} = \frac{1}{Z} e^{-\frac{1}{T}E(\mathbf{s})} \tag{3}$$

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 [18].

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:

$$KL(Q||P) = \mathbb{E}_Q \left(\ln \frac{Q}{P} \right) = \sum_{\mathbf{s}} Q(\mathbf{s}) \ln \frac{Q(\mathbf{s})}{P(\mathbf{s})}. \tag{4}$$

The KL-divergence is a non-symmetric measure of the difference between two distributions which is always non-negative. Substituting P from 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* [12]. The partition function Z doesn't depend on Q and we need to only focus on minimizing the variational free energy:

$$F[Q] \equiv \mathbb{E}[Q] - TH[Q]. \tag{5}$$

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 $F \equiv -T \ln Z$.

2.4. Mean field approximation

The most widely used approximation to the family of models defined in 3 is the mean field approximation which is obtained by taking as an approximator the family of distribution that factorizes, i.e.

$$Q(\mathbf{s}) = \prod_i q_i(s_i) \tag{6}$$

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

$$F^{MF} = -\sum_{(ij)} \sum_{s_i,s_j} w_{ij} q_i(x_i) q_j(x_j) - \sum_i \sum_{s_i} \theta_i q_i(x_i) + T \sum_i \sum_{s_i} q_i(s_i) \ln q_i(s_i) \tag{7}$$

and the energy for a single spin is:

$$E(s_i) = -\theta_i s_i - \sum_j w_{ij} s_i m_j \tag{8}$$

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

$$m_i = \mathbb{E}_{q_i}(s_i), \quad i \in \{1, ..., N\} \tag{9}$$

where $\mathbb{E}$ refers to the average configuration under the Boltzmann measure. In terms of magnetizations, 7 becomes:

$$F^{MF} = -\sum_{(ij)} w_{ij} m_i m_j - \sum_i \theta_i m_i + T \sum_i \left[m_i \ln m_i + (1 - m_i) \ln(1 - m_i) \right]. \tag{10}$$

Minimizing 10 with respect to magnetizations yields the so-called mean field stationary conditions:

$$m_i = \text{sigm}\left(\frac{1}{T} \sum_j w_{ij} m_j + \frac{1}{T} \theta_i \right), \quad i \in \{1, ..., N\} \tag{11}$$

where N is the number of spins in the model. These equations are usually run sequentially. As the free energy is convex [17], 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 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[8].

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

$$\begin{aligned} \ln Z &= \ln \sum_{\mathbf{s}} \exp(-\frac{1}{T} E(\mathbf{s})) = \ln \sum_{\mathbf{s}} Q(\mathbf{s}) \frac{\exp(-\frac{1}{T} E(\mathbf{s}))}{Q(\mathbf{s})} \\ &\geqslant \sum_{\mathbf{s}} Q(\mathbf{s}) \ln \frac{\exp(-\frac{1}{T} E(\mathbf{s}))}{Q(\mathbf{s})} = -\frac{1}{T} \mathbb{E}_Q(E(\mathbf{s})) + H(Q) \end{aligned} \tag{12}$$

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:

$$F = -T \ln Z \leqslant \mathbb{E}[Q] - TH[Q] = F[Q]. \tag{13}$$

2.5. Extended mean field approximation

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 [6]. We will minimize 5 where instead of assuming Q to be a product distribution we require that magnetizations has appropriate values, i.e.:

$$\mathbb{E}_Q(\mathbf{s}) = \mathbf{m}. \tag{14}$$

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

$$\beta F(\mathbf{m}) = \min_Q \{ E(Q) - H(Q) \mid \mathbb{E}(\mathbf{S}) = \mathbf{m} \} \tag{15}$$

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.:

$$E(Q) - H(Q) - \sum_i \lambda_i(\mathbb{E}(s_i) - m_i). \tag{16}$$

Thus, the minimizing distribution has the form:

$$Q_{\mathbf{m}}(\mathbf{s}) = \frac{1}{Z} e^{-E(\mathbf{s}) + \lambda_i s_i} \tag{17}$$

with partition function $Z = \sum_{\mathbf{s}} e^{-E(\mathbf{s}) + \sum_i \lambda_i s_i}$. Using this distribution back into 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:

$$-\beta F = \ln \sum_{\mathbf{s}} \exp\left(\beta \sum_{(ij)} w_{ij} s_i s_j + \beta \sum_i \theta_i s_i + \sum_i \lambda_i(\beta)(s_i - m_i)\right) \tag{18}$$

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

$$-\beta F = -(\beta F)_{\beta=0} - \left(\frac{\partial(\beta F)}{\partial \beta}\right)_{\beta=0} \beta - \left(\frac{\partial^2(\beta F)}{\partial \beta^2}\right)_{\beta=0} \frac{\beta^2}{2} - ... \tag{19}$$

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:

$$m_i = \mathbb{E}_{\beta=0}(s_i) = \frac{\exp(\lambda_i(0))}{\exp(\lambda_i(0)) + 1} = \text{sigm}(\lambda_i(0)) \tag{20}$$

We can now choose which variables use in derivations and this is a purely dependent on mathematical convenience. As the equation 20 is easy to invert, we will work on the magnetizations. The first term from the 19 takes now the form:

$$\begin{aligned}
-(\beta F)_{\beta=0} &= \ln \sum_{\mathbf{s}} \exp\left(\sum_i \lambda_i(0)(s_i - m_i)\right) \\
&= \ln \left\{ \sum_{s_1} \exp\left(\lambda_1(0)(s_1 - m_1)\right) ... \sum_{s_n} \exp\left(\lambda_n(0)(s_n - m_n)\right) \right\} \\
&= \ln \left\{ (\exp(\lambda_i(0)) + 1) \exp(-\lambda_1(0) m_1) ... (\exp(\lambda_i(0)) + 1) \exp(-\lambda_1(0) m_n) \right\} \\
&= \sum_i \left\{ \ln\left(\frac{1}{1 - m_i}\right) - m_i \ln\left(\frac{m_i}{1 - m_i}\right) \right\} \\
&= -\sum_i \left[m_i \ln(m_i) + (1 - m_i) \ln(1 - m_i) \right]
\end{aligned} \tag{21}$$

where using 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 10. Next, the first derivative is:

$$-\left.\frac{\partial(\beta F)}{\partial \beta}\right|_{\beta=0} = \sum_{(ij)} w_{ij} \mathbb{E}_{\beta=0}(s_i s_j) + \sum_i \theta_i \mathbb{E}_{\beta=0}(s_i) - \sum_i \left.\frac{\partial \lambda_i(\beta)}{\partial \beta}\right|_{\beta=0} \mathbb{E}(s_i - m_i) \tag{22}$$

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

$$-\left.\frac{\partial(\beta F)}{\partial \beta}\right|_{\beta=0} = \sum_{(ij)} w_{ij} m_i m_j + \sum_i \theta_i m_i. \tag{23}$$

Vomparing 23 and 21 with 10 we can see that we have already recovered the simple mean field approximation. Yedida and Georges [6] 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:

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

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.

2.6. 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 [18]. There are a few examples in statistical physics where this method works very reliably in a wide variety of temperatures [13] 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.

2.7. Boltzmann machine

A particular example from the family of distributions defined in 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 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.

2.7.1. 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 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 [4].

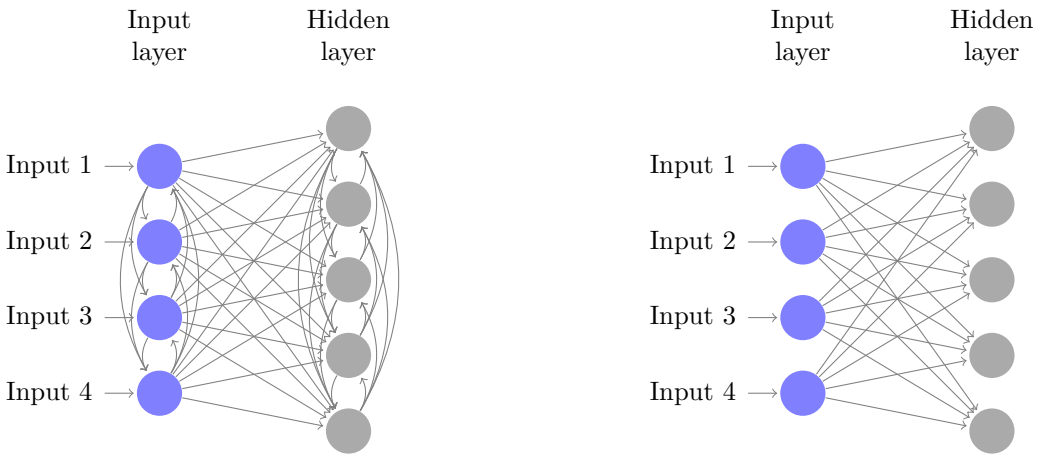

Figure 1: Exemplary graphs of Boltzmann machine (left) and restricted Boltzmann Machine (right) with 4 visible and 5 hidden units.

2.7.2. 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 [9]. It can be shown that with additional hidden units there exist weight values for these new units that 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.

2.7.3. 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:

$$\begin{aligned} p(\mathbf{h}|\mathbf{v}) &= \prod_{i=1}^{M} p(h_i|\mathbf{v}), \\ p(\mathbf{v}|\mathbf{h}) &= \prod_{i=1}^{N} p(v_i|\mathbf{h}). \end{aligned} \tag{24}$$

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

$$\begin{aligned} p(h_i = 1|\mathbf{v}) = p(h_i = 1|\mathbf{h}_{-i}, \mathbf{v}) &= \frac{p(h_i = 1, \mathbf{h}_{-i}, \mathbf{v})}{p(\mathbf{h}_{-i}, \mathbf{v})} \\ &= \frac{\exp(-E(h_i = 1, \mathbf{h}_{-i}, \mathbf{v}))}{\exp(-E(h_i = 1, \mathbf{h}_{-i}, \mathbf{v})) + \exp(-E(h_i = 0, \mathbf{h}_{-i}, \mathbf{v}))} \\ &= \frac{1}{1 + \exp(\sum_{n=1}^{N} W_{i,n} v_n + a_n)} \\ &= \text{sigm}(\sum_{n=1}^{N} W_{i,n} v_n + b_i)) \end{aligned} \tag{25}$$

and following the same steps we can show that:

$$p(v_j = 1|\mathbf{h}) = \text{sigm}(\sum_{m=1}^{M} W_{i,m}^T h_m + a_j)). \tag{26}$$

The independence between the variables in one layer makes sampling from conditional distributions 25 and 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:

$$\begin{aligned}\sum_{\mathbf{h}} e^{-E(\mathbf{v},\mathbf{h})} &= e^{\mathbf{b}'\mathbf{v}} \sum_{h_1} ... \sum_{h_m} e^{-E(\mathbf{v},\mathbf{h})} \\ &= e^{\mathbf{b}'\mathbf{v}} \sum_{h_1} e^{h_1(c_1+W_{1\bullet}\mathbf{v})} ... \sum_{h_m} e^{h_m(c_m+W_{m\bullet}\mathbf{v})} \\ &= e^{\mathbf{b}'\mathbf{v}} \prod_{j=1}^{m} \left(1 + e^{c_i+W_{i\bullet}\mathbf{v}}\right)\end{aligned} \tag{27}$$

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.

3. Evaluation on the toy models

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:

$$\max_{\theta} \ln \mathcal{L}(\theta|\mathcal{D}) = \max_{\theta} \ln \prod_{i=1}^{N} p(\mathbf{v}_i|\theta) = \max_{\theta} \sum_{i=1}^{N} \ln p(\mathbf{v}_i|\theta) \tag{28}$$

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

The experiments on toy models suggest that the initial unsatisfactory results with naive mean field approaches [16] might be greatly improved if we include additional terms responsible for connections between the spins.

4.2. 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:

$$\theta^t = \theta^{t-1} + \eta \frac{\partial}{\partial \theta^{t-1}} \ln \mathcal{L}(\theta|\mathcal{D}). \tag{29}$$

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 [14] where we sample a minibatch of datapoints and take a noisy gradient estimate which results in the update rule:

$$\theta^{t+1} = \theta^t + \eta \frac{1}{M} \frac{\partial}{\partial \theta^t} \sum_{m=1}^{M} \ln \mathcal{L}(\theta|\mathbf{x}^{(m)}), \tag{30}$$

where M is the size of the minibatch. It can be shown that updates via 30 guarantee to converge to a local optimum under weak conditions [3].

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

$$\mathcal{L} = \ln P(\mathbf{v}) = -\ln(\sum_{\mathbf{h}} e^{-E(\mathbf{v},\mathbf{h})}) - \ln Z = F^c(\mathbf{v}) + F \tag{31}$$

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:

$$\begin{aligned}
\frac{\partial \log \mathcal{L}(\theta|\mathbf{v})}{\partial \theta} &= \frac{\partial F^c}{\partial \theta} - \frac{\partial F}{\partial \theta} \\
&= -\frac{\sum_{\mathbf{h}} e^{-E(\mathbf{v},\mathbf{h})} \frac{\partial E(\mathbf{v},\mathbf{h})}{\partial \theta}}{\sum_{\mathbf{h}} e^{-E(\mathbf{v},\mathbf{h})}} + \frac{\sum_{\mathbf{v},\mathbf{h}} e^{-E(\mathbf{v},\mathbf{h})} \frac{\partial E(\mathbf{v},\mathbf{h})}{\partial \theta}}{\sum_{\mathbf{v},\mathbf{h}} e^{-E(\mathbf{v},\mathbf{h})}} \\
&= -\sum_{\mathbf{h}} p(\mathbf{h}|\mathbf{v}) \frac{\partial E(\mathbf{v},\mathbf{h})}{\partial \theta} + \sum_{\mathbf{v},\mathbf{h}} p(\mathbf{v},\mathbf{h}) \frac{\partial E(\mathbf{v},\mathbf{h})}{\partial \theta} \\
&= -\mathbb{E}_{p(\mathbf{h}|\mathbf{v})} \left(\frac{\partial E(\mathbf{v},\mathbf{h})}{\partial \theta} \right) + \mathbb{E}_{p(\mathbf{v},\mathbf{h})} \left(\frac{\partial E(\mathbf{v},\mathbf{h})}{\partial \theta} \right)
\end{aligned} \tag{32}$$

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.

4.3. Monte Carlo methods

The second expectation from the gradient in 32 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 $n \to \infty$.

4.3.1. 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.

4.3.2. 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 **??**.

4.4. 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 [7] 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 32. 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 2 presents the procedure for the CD-1:

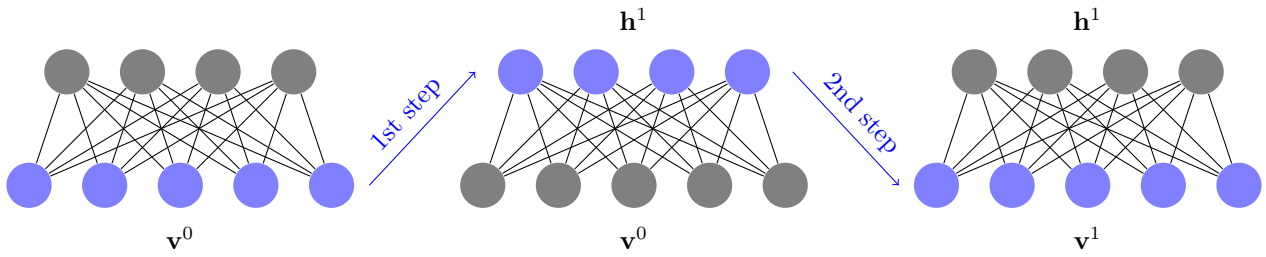

Figure 2: The first step of the Gibbs sampler for the RBM for a particular data point $\mathbf{v}^0 \in \mathcal{D}$.

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

$$-\sum_{\mathbf{h}} p(\mathbf{h}|\mathbf{v}^0)\frac{\partial E(\mathbf{v}^0,\mathbf{h})}{\partial\theta} + \sum_{\mathbf{h}} p(\mathbf{h}|\mathbf{v}^k)\frac{\partial E(\mathbf{v}^k,\mathbf{h})}{\partial\theta} \tag{33}$$

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 33 is biased as it doesn't maximize the likelihood of the data but the difference of two KL-divergences [7], [4]:

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

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.

4.4.1. 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) [16]. 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[11]. 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.

4.5. 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:

$$\mathcal{L} = \ln P(\mathbf{v}) = F^c(\mathbf{v}) - F^{EMF}. \tag{34}$$

As the first term from 31 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:

$$\begin{aligned}\frac{\partial F^{EMF}}{\partial a_i} &= -m_i^v, \\ \frac{\partial F^{EMF}}{\partial b_j} &= -m_j^h.\end{aligned} \tag{35}$$

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 $\mathbf{h} = \text{sigm}(W\mathbf{v}+\mathbf{b})$ 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.

4.6. 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

$$P(\mathbf{x};\theta) = \prod_i p(x_i|x_1,...,x_{i-1};\theta) \approx \prod_i p(x_i|x_1,...,x_{i-1},x_{i+1},...,x_n;\theta) = \prod_i p(x_i|x_{-i};\theta) \coloneqq PL(\mathbf{x};\theta) \tag{36}$$

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:

$$\ln PL(\mathbf{x};\theta) = \sum_i \ln P(x_i|x_{i-1};\theta). \tag{37}$$

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.

$$\ln PL(\mathbf{x};\theta) = n \ln P(x_i|\mathbf{x}_{-i};\theta), \tag{38}$$

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:

$$\ln PL(\mathbf{x};\theta) \approx n \log\left(\frac{\exp\{-F^c(\mathbf{x})\}}{\exp\{-F^c(\hat{\mathbf{x}})\} + \exp\{-F^c(\mathbf{x})\}}\right) = n \ln\left(\text{sigm}(F^c(\hat{\mathbf{x}}) - F^c(\mathbf{x})\right) \tag{39}$$

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

4.7. 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 [10] 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 [5], [15] 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.

4.8. 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 **??** 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 [5]. 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 [5] 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:

$$E(\theta, \mathcal{D}) = \ln \mathcal{L}(\theta|\mathcal{D}) - \lambda R(\theta) \tag{40}$$

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 3 presents the pseudo log-likelihood 39 (left) and EMF log-likelihood 34 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.

Figure 3: 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 [5].

Figure 4 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.

Figure 4: 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 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.

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

5. Chapter 4 - Applications

6. Conclusions

7. Appendix

References

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

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

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

[4] Asja Fischer and Christian Igel. An introduction to restricted boltzmann machines. In *Iberoamerican Congress on Pattern Recognition*, pages 14–36. Springer, 2012.

[5] Marylou Gabrié, Eric W Tramel, and Florent Krzakala. Training restricted boltzmann machine via the thouless-anderson-palmer free energy. In *Advances in Neural Information Processing Systems*, pages 640–648, 2015.

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

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

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

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

[10] Yann LeCun and Corinna Cortes. The mnist database of handwritten digits. Technical report, 1998.

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

[12] Manfred Opper and David Saad. *Advanced mean field methods: Theory and practice*. MIT press, 2001.

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

[14] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[15] Ruslan Salakhutdinov. Learning and evaluating boltzmann machines. Technical report, 2008.

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

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

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