0.1. 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.

0.2. 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:

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

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 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 [?]. 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.

0.3. 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 [?].

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* where $\mathbb{E}$ refers to the average configuration under the Boltzmann measure [?]. The partition function Z doesn't depend on Q and we need to only focus on minimizing the variational free energy:

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

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 $\mathcal{F} := -T \ln Z$.

0.4. (Naive) 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 as following:

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

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:

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

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

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

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:

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

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

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:

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

0.5. 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 [**?**]. 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:

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

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:

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

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 [**?**] 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

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

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.

0.6. 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 [**?**]. There are a few examples in statistical physics where this method works very reliably in a wide variety of temperatures [**?**] 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.

0.7. Boltzmann machine

A particular example from the family of distributions defined in 2 is a Boltzmann machine [**?**] 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.

0.7.1. 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 [**?**].

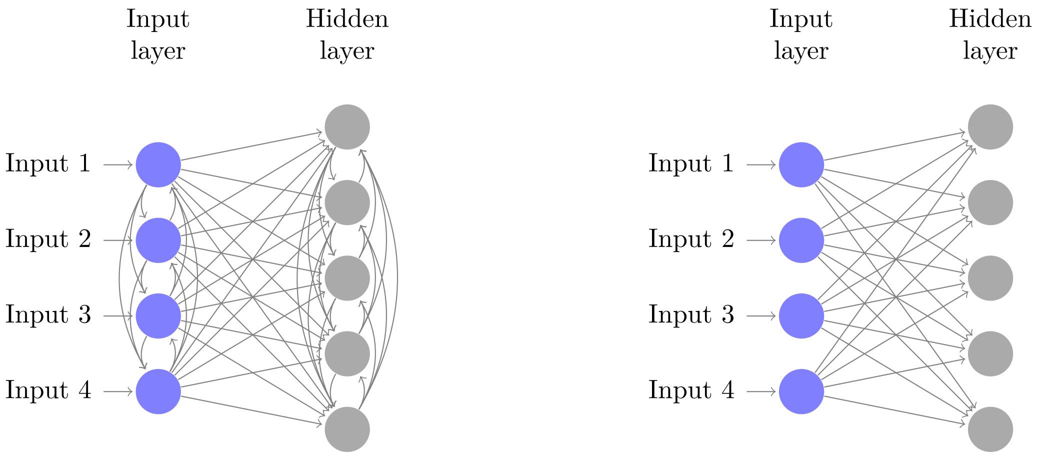

Figure 1: Exemplary graphs of Boltzmann machine (left) and restricted Boltzmann Machine (right) with 4 visible and 5 hidden units.

0.7.2. 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 [**?**]. 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.

0.7.3. 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:

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

and following the same steps we can show that:

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

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.