# PGM - Variational methods for sigmoid belief networks

Caio Corro

## Contents



## 1 Introduction

Computing the probability of an observation requires to marginalize out latent (i.e. non observed) variables (if any). For, many models, this marginalization is too expensive to be computed exactly. There exists two standard approximation methods: stochastic approximations and variational approximations. These notes present variational approximation techniques for two layers sigmoid belief networks that were introduced in [Saul et al., 1996] and [Jaakkola and Jordan, 1996].

The main idea is as follows. Assume a joint probability distribution $p(\mathcal{X}, \mathcal{Y})$ where $\mathcal{X}$ and $\mathcal{Y}$ are the observed and latent random variables, respectively. We wish to estimate the evidence (log-probability) of an observation $\boldsymbol{x}$:

$$\log(\mathcal{X} = \boldsymbol{x}) = \log \mathbb{E}_{p(\mathcal{Y})}[p(\mathcal{X} = \boldsymbol{x}|\mathcal{Y})]$$

This quantity can be useful if we want to estimate the probability that a given point comes from the distribution or to learn the parameters of the distribution by maximizing the log-likelihood of a dataset. In many cases, the expectation (i.e. the marginalization) leads to an intractable evidence. Therefore, we may want to approximate it using parameterized functions:

$$\begin{aligned} & \log(\mathcal{X} = \boldsymbol{x}) \geq f(\boldsymbol{x}, \phi) \\ \text{or} \quad & \log(\mathcal{X} = \boldsymbol{x}) \leq h(\boldsymbol{x}, \epsilon), \end{aligned}$$

where we assume that computing $f(\boldsymbol{x}, \phi)$ and $h(\boldsymbol{x}, \epsilon)$ is easy. The parameters $\phi$ and $\epsilon$ are variational parameters that control the quality of the bounds, i.e. we assume that these bounds are true for any $\phi$ and $\epsilon$. That is, we replaced an intractable expectation with an easy to compute parameterized function. Note that, unfortunately, finding the best possible variational parameters can be a difficult problem.

## 2 Two layers sigmoid belief networks (SBN)

Let $\mathcal{X}$ be the observed random variable, taking values in $[0, 1]^d$, and $\mathcal{Y}$ the latent random variable, taking values in $[0, 1]^k$. A two layers sigmoid belief networks (SBN) defines a joint probability distribution

on $\mathcal{X}$ and $\mathcal{Y}$ as follows:

$$\begin{aligned} p_\theta(\boldsymbol{x}, \boldsymbol{y}) &= p_\theta(\boldsymbol{y}) p_\theta(\boldsymbol{x}|\boldsymbol{y}) \\ &= \prod_{i=1}^{k} p_\theta(y_i) \prod_{i=1}^{d} p_\theta(x_i|\boldsymbol{y}) \\ &= \prod_{i=1}^{k} \frac{\exp(y_i a_i)}{1+\exp(a_i)} \prod_{i=1}^{d} \frac{\exp(x_i(\boldsymbol{B}_i \boldsymbol{y} + c_i)}{1+\exp(\boldsymbol{B}_i \boldsymbol{y} + c_i)} \end{aligned}$$

where $\theta = \{\boldsymbol{a}, \boldsymbol{B}, \boldsymbol{c}\}$ are the parameters of the model with $\boldsymbol{a} \in \mathbb{R}^k$, $\boldsymbol{B} \in \mathbb{R}^{d \times k}$ and $\boldsymbol{c} \in \mathbb{R}^d$. In this model, the evidence is intractable as it requires to sum of $2^k$ possible assignments for the latent variable $\mathcal{Y}$.

Generating data from a trained model is easy via ancestral sampling: first sample from the latent variable distribution (independent Bernoullis) and then sample from the conditional distribution of observed variables (independent Bernoullis again). The generative story can be described as follows:

1. $\boldsymbol{y} \sim p_\theta(\mathcal{Y})$
2. $\boldsymbol{x} \sim p_\theta(\mathcal{X}|\mathcal{Y} = \boldsymbol{y})$

## 2.1 Lower bound on the evidence

In this section, we derive a lower bound on the evidence based on the introduction of a proposal distribution and Jensen's inequality. The resulting lower bound is not equivalent to the "standard" ELBO, as one of the term of the ELBO is intractable for SBNs. Contrary to Gaussian Mixture Models, the posterior distribution cannot be computed exactly (as it requires summing over $2^k$ terms in the denominator of the Bayes rule). Therefore, we cannot just introduce a proposal distribution that is made equal to the true posterior as in the Expectation step of EM for GMM. Instead, we introduce a proposal distribution that factorizes across latent variable, i.e. each latent variable is independent wrt to other latent variable. In other words, the proposal distribution is defined as follows:

$$q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}) = \prod_{i=1}^{k} (\phi_i^{(\boldsymbol{x})})^{z_i} (1 - \phi^{(\boldsymbol{x})})_i^{1-z_i}$$

where $\boldsymbol{\phi}^{(\boldsymbol{x})} \in [0,1]^k$ are the parameters of the proposal distribution associated with observation $\boldsymbol{x}$. We can observe that they are independent Bernoullis. This technique of simplifying the distribution of the proposal is called Mean Field Theory in statistical physics.

The bound we present here is a simplified version of the one proposed by [Saul et al., 1996]. We first derive the evidence lower bound via Jensen's inequality:

$$\begin{aligned} \log p_\theta(\boldsymbol{x}) &= \log \sum_{\boldsymbol{y} \in Y} p_\theta(\boldsymbol{y}) p_\theta(\boldsymbol{x}|\boldsymbol{y}) \\ &\geq \underbrace{\mathbb{E}_{q_\phi(\mathcal{Y}|\boldsymbol{x})}[\log p_\theta(\mathcal{Y})]}_{\textbf{(a)}} + \underbrace{\mathbb{E}_{q_\phi(\mathcal{Y}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\mathcal{Y})]}_{\textbf{(b)}} + \underbrace{H^{\mathrm{S}}[q_\phi(\mathcal{Y}|\boldsymbol{x})]}_{\textbf{(c)}} \end{aligned}$$

where $H^{\mathrm{S}}$ is the Shannon entropy. Although this equation doesn't seem simpler (we still need to sum over $2^k$ values in the tree expectations), we show in the following that the mean field assumption on $q$ allows to compute these expectations efficiently (except **(b)** which requires to derive another bound on this term).

Term **(a)** can be rewritten as:

$$\begin{aligned}\mathbb{E}_{q_\phi(\mathcal{Y}|\boldsymbol{x})}[\log p_\theta(\mathcal{Y})] &= \mathbb{E}_{q_\phi(\mathcal{Y}|\boldsymbol{x})}\left[\log \prod_{i=1}^{k} \frac{\exp(\mathcal{Y}_i a_i)}{1+\exp(a_i)}\right] \\ &= \mathbb{E}_{q_\phi(\mathcal{Y}|\boldsymbol{x})}\left[\sum_{i=1}^{k} \big(\mathcal{Y}_i a_i - \log(1+\exp(a_i))\big)\right]\end{aligned}$$

Note that by linearity of the expectation, we can separate the two part of the substraction. Moreover, the term $\log(1+\exp(a_i))$ does not depends on $Y$, therefore:

$$\begin{aligned}&= \sum_{i=1}^{k} \mathbb{E}_{q_\phi(\mathcal{Y}|\boldsymbol{x})}\left[\mathcal{Y}_i a_i\right] - \sum_{i=1}^{k} \log(1+\exp(a_i)) \\ &= \sum_{i=1}^{k} a_i \phi_i^{(\boldsymbol{x})} - \sum_{i=1}^{k} \log(1+\exp(a_i)) \\ &= \langle \boldsymbol{a}, \boldsymbol{\phi}^{(\boldsymbol{x})} \rangle - \sum_{i=1}^{k} \log(1+\exp(a_i))\end{aligned}$$

which is easy to compute, we don't need to sum over $2^k$ terms in this form.

Term **(b)** can be rewritten as:

$$\begin{aligned}\mathbb{E}_{q_\phi(\mathcal{Y}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\mathcal{Y})] &= \mathbb{E}_{q_\phi(\mathcal{Y}|\boldsymbol{x})}\left[\log \prod_{i=1}^{d} \frac{\exp(x_i(\boldsymbol{B}_i\mathcal{Y}+c_i))}{1+\exp(\boldsymbol{B}_i\mathcal{Y}+c_i)}\right] \\ &= \mathbb{E}_{q_\phi(\mathcal{Y}|\boldsymbol{x})}\left[\sum_{i=1}^{d} \big(x_i(\boldsymbol{B}_i\mathcal{Y}+c_i) - \log(1+\exp(\boldsymbol{B}_i\mathcal{Y}+c_i))\big)\right] \\ &= \sum_{i=1}^{d} \mathbb{E}_{q_\phi(\mathcal{Y}|\boldsymbol{x})}\left[\big(x_i(\boldsymbol{B}_i\mathcal{Y}+c_i)\big)\right] - \sum_{i=1}^{d} \mathbb{E}_{q_\phi(\mathcal{Y}|\boldsymbol{x})}\left[\log(1+\exp(\boldsymbol{B}_i\mathcal{Y}+c_i))\right]\end{aligned}$$

As for term **(a)**, the first argument of the substraction can be simplified. However, the second argument cannot be simplified directly and requires to sum over $2^k$ values. We derive a lower bound on this term using Jensen's inequality again (therefore, we actually compute a lower bound to the standard ELBO):

$$\begin{aligned}&\geq \sum_{i=1}^{d} x_i \mathbb{E}_{q_\phi(\mathcal{Y}|\boldsymbol{x})}\left[\sum_{j=1}^{k} \boldsymbol{B}_{i,k}\mathcal{Y}_k\right] + \sum_{i=1}^{d} x_i c_i \\ &\quad - \sum_{i=1}^{d} \log\left(\mathbb{E}_{q_\phi(\mathcal{Y}|\boldsymbol{x})}\left[1+\exp(c_i)\prod_{j=1}^{k}\exp(B_{i,j}\mathcal{Y}_j)\right]\right) \\ &= \langle \boldsymbol{x}, \boldsymbol{B}\boldsymbol{\phi}^{(\boldsymbol{x})} \rangle + \langle \boldsymbol{x}, \boldsymbol{c} \rangle - \sum_{i=1}^{d} \log\left(1+\exp(c_i)\prod_{j=1}^{k}\left(1-\phi_j^{(\boldsymbol{x})}+\phi_j^{(\boldsymbol{x})}\exp(B_{i,j})\right)\right),\end{aligned}$$

which is easy to compute.

Lastly, the entropy term **(c)** can be rewritten as:

$$\begin{aligned} H^{\mathrm{S}}[q_\phi(\mathcal{Y}|\boldsymbol{x})] &= \mathbb{E}_{q_\phi(\mathcal{Y}|\boldsymbol{x})}[\log q_\phi(\mathcal{Y}|\boldsymbol{x})] \\ &= \mathbb{E}_{q_\phi(\mathcal{Y}|\boldsymbol{x})}\left[\log\left(\prod_{i=1}^{k} q_\phi(\mathcal{Y}_i|\boldsymbol{x})\right)\right] \\ &= \sum_{i=1}^{k} \mathbb{E}_{q_\phi(\mathcal{Y}|\boldsymbol{x})}\left[\log q_\phi(\mathcal{Y}_i|\boldsymbol{x})\right] \end{aligned}$$

Note that the value inside the expectation depends only on a single random variable element $\mathcal{Y}_i$, therefore it simplifies to:

$$\begin{aligned} &= \sum_{i=1}^{k} \mathbb{E}_{q_\phi(\mathcal{Y}_i|\boldsymbol{x})}\left[\log q_\phi(\mathcal{Y}_i|\boldsymbol{x})\right] \\ &= \sum_{i=1}^{k} \left(\phi_i^{(\boldsymbol{x})} \log \phi_i^{(\boldsymbol{x})} + (1-\phi_i^{(\boldsymbol{x})}) \log(1-\phi_i^{(\boldsymbol{x})})\right) \end{aligned}$$

which is tractable.

# 3 Upper bound on the evidence

We now show how to derive an upper bound on the evidence. Although it is not a good idea to rely on this upper bound for training a SBN (i.e. at training time we want to maximize a lower bound, hopping that maximizing a lower bound will also push up the actual evidence of the training data), it can be useful for example to estimate the quality of the lower bound derived in the previous section by computing the gap between the two bounds.

The method described in this section is a simplification of the approach proposed in [Jaakkola and Jordan, 1996]. The probability of an observation is defined as:

$$p_\theta(\boldsymbol{x}) = \sum_{\boldsymbol{y} \in Y} p_\theta(\boldsymbol{y}) p_\theta(\boldsymbol{x}|\boldsymbol{y})$$

As $1 - \sigma(u) = \sigma(-u)$, we can write:

$$= \sum_{\boldsymbol{y} \in Y} p_\theta(\boldsymbol{y}) \prod_{i=1}^{d} \sigma((2x_i - 1)(\boldsymbol{B}_i \boldsymbol{y} + c_i))$$

We can replace the sigmoid by its variational formulation, i.e. $\sigma(u) = \inf_{\epsilon \in [0,1]} \exp(\epsilon u - H^{\mathrm{FD}}[\epsilon])$ where $H^{\mathrm{FD}}[\epsilon] = -\epsilon \log \epsilon - (1-\epsilon)\log(1-\epsilon)$ if the Fermi-Dirac entropy. Note that we need one variational parameter $\epsilon_i$ per observed variable dimension. We obtain:

$$= \sum_{\boldsymbol{y} \in Y} p_\theta(\boldsymbol{y}) \prod_{i=1}^{d} \inf_{\epsilon_i \in [0,1]} \exp(\epsilon_i (2x_i - 1)(\boldsymbol{B}_i \boldsymbol{y} + c_i) - H^{\mathrm{FD}}[\epsilon_1])$$

The objectives in infinimizations are strictly positive, therefore we can take out the infinimizations from the product:

$$= \sum_{\boldsymbol{y} \in Y} p_\theta(\boldsymbol{y}) \inf_{\substack{\epsilon_i \in [0,1], \\ \forall i \in \{1 \dots d\}}} \prod_{i=1}^{d} \exp(\epsilon_i (2x_i - 1)(\boldsymbol{B}_i \boldsymbol{y} + c_i) - H^{\text{FD}}[\epsilon_1])$$

$$= \sum_{\boldsymbol{y} \in Y} p_\theta(\boldsymbol{y}) \inf_{\substack{\epsilon_i \in [0,1], \\ \forall i \in \{1 \dots d\}}} \exp(-\sum_{i=1}^{d} H^{\text{FD}}[\epsilon_1]) \exp(\sum_{i=1}^{d} \epsilon_i (2x_i - 1)(\boldsymbol{B}_i \boldsymbol{y} + c_i))$$

$$= \sum_{\boldsymbol{y} \in Y} p_\theta(\boldsymbol{y}) \inf_{\substack{\epsilon_i \in [0,1], \\ \forall i \in \{1 \dots d\}}} \exp(-\sum_{i=1}^{d} H^{\text{FD}}[\epsilon_1]) \exp(\sum_{i=1}^{d} \epsilon_i c_i (2x_i - 1)) \prod_{j=1}^{k} \exp(\sum_{i=1}^{d} \epsilon_i \boldsymbol{B}_{i,j} (2x_i - 1))^{y_j}$$

Note that the outside sum is an expectation:

$$= \mathbb{E}_{p_\theta(\mathcal{Y})} \left[ \inf_{\substack{\epsilon_i \in [0,1], \\ \forall i \in \{1 \dots d\}}} \exp(-\sum_{i=1}^{d} H^{\text{FD}}[\epsilon_1]) \exp(\sum_{i=1}^{d} \epsilon_i c_i (2x_i - 1)) \prod_{j=1}^{k} \exp(\sum_{i=1}^{d} \epsilon_i \boldsymbol{B}_{i,j} (2x_i - 1))^{y_j} \right]$$

We obtain an upper bound by moving the expectation inside the infinimizations. To understand why this gives us an upper bound, instead of finding the optimal variational parameters for each latent variable assignments $\boldsymbol{y}$, now the assignment is shared across all possible $\boldsymbol{y} \in Y$. As it is an infinimization, which is then "less expressive", we obtain an upperound.

$$\leq \inf_{\substack{\epsilon_i \in [0,1], \\ \forall i \in \{1 \dots d\}}} \exp(-\sum_{i=1}^{d} H^{\text{FD}}[\epsilon_1]) \exp(\sum_{i=1}^{d} \epsilon_i c_i (2x_i - 1)) \prod_{j=1}^{k} \mathbb{E}_{p_\theta(\mathcal{Y}_i | \boldsymbol{x})} \left[ \exp(\sum_{i=1}^{d} \epsilon_i \boldsymbol{B}_{i,j} (2x_i - 1))^{y_j} \right]$$

$$= \inf_{\substack{\epsilon_i \in [0,1], \\ \forall i \in \{1 \dots d\}}} \exp(-\sum_{i=1}^{d} H^{\text{FD}}[\epsilon_1]) \exp(\sum_{i=1}^{d} \epsilon_i c_i (2x_i - 1)) \prod_{j=1}^{k} \left( 1 - \sigma(a_i) + \sigma(a_i) \exp(\sum_{i=1}^{d} \epsilon_i \boldsymbol{B}_{i,j} (2x_i - 1)) \right)$$

Computing the objective in the infinimizations in the last expression are all tractable. Therefore, we have the following upper bound on the evidence:

$$\log p_\theta(\boldsymbol{x}) \leq \left( -\sum_{i=1}^{d} H^{\text{FD}}[\epsilon_1] + \sum_{i=1}^{d} \epsilon_i c_i (2x_i - 1) + \sum_{j=1}^{k} \left( 1 - \sigma(a_i) + \sigma(a_i) \exp(\sum_{i=1}^{d} \epsilon_i \boldsymbol{B}_{i,j} (2x_i - 1)) \right) \right)$$

where $\boldsymbol{\epsilon} \in [0,1]^d$ are the variational parameters of the bound. Although this bound is easy to compute, this does not mean that computing the parameters $\boldsymbol{\epsilon}$ that leads to the tightness bound possible is easy. Methods to compute the variational parameters are described in [Jaakkola and Jordan, 1996].

# 4 Expectation-Maximization algorithm

In this section, we show how the variational lower bound can be used to learn the parameters $\theta$ of a sigmoid belief network. The Expectation-Maximization algorithm is a learning algorithm the relies on block-coordinate ascent to optimize the evidence lower bound (ELBO). Given a dataset $\mathcal{D} = \{\boldsymbol{x}^{(i)}\}_{i=1}^{n}$, the log-likelihood of the dataset and its ELBO are defined as follows:

$$\max_{\theta \in \Theta} \sum_{\boldsymbol{x} \in \mathcal{D}} \log p_\theta(\boldsymbol{x}) \geq \max_{\theta \in \Theta, \boldsymbol{\phi} \in \Phi} \sum_{\boldsymbol{x} \in \mathcal{D}} \textsc{Elbo}(\boldsymbol{x}, \theta, \boldsymbol{\phi})$$

The two step of the EM algorithms are:

1. Expectation step: maximize over variational parameters $\boldsymbol{\phi}$;

2. Maximization step: maximize over model parameters $\theta$.

The iterative optimization process simply interleaves between E and M steps. In the case of GMMs, both steps enjoy simple closed-form expression. This is not the case of SBNs. We show below how the E step for SBN can be approximated efficiently. For the M step, one can simply rely on a single step of gradient ascent.

To simplify notation, for each training point $\boldsymbol{x} \in \mathcal{D}$ we introduce vector $\boldsymbol{w}^{(\boldsymbol{x})} \in \mathbb{R}^d$ and matrix $\boldsymbol{U}^{(\boldsymbol{x})} \in \mathbb{R}^{d,k}$ defined as follows:

$$w_j^{(\boldsymbol{x})} = \exp(c_j) \prod_{l=1}^{k} (1 - \phi_l^{(\boldsymbol{x})} + \phi_l^{(\boldsymbol{x})} \exp(B_{j,l}))$$

$$U_{j,l}^{(\boldsymbol{x})} = \frac{\partial}{\partial \phi_l^{(\boldsymbol{x})}} \log(1 + w_j^{(\boldsymbol{x})}) = \frac{w_j^{(\boldsymbol{x})}}{1 + w_j^{(\boldsymbol{x})}} \times \frac{\exp(B_{j,l}) - 1}{1 - \phi_l^{(\boldsymbol{x})} + \phi_l^{(\boldsymbol{x})} \exp(B_{j,l})}$$

The objective of the E step is defined as follows:

$$\max_{\boldsymbol{\phi} \in \Phi} \sum_{\boldsymbol{x} \in \mathcal{D}} \textsc{Elbo}(\boldsymbol{x}, \theta, \boldsymbol{\phi}) = \max_{\substack{\boldsymbol{\phi}^{(\boldsymbol{x})} \in [0,1]^k, \\ i \in \{1 \dots n\}}} \sum_{\boldsymbol{x} \in \mathcal{D}} \left( \begin{array}{l} \langle \boldsymbol{a}, \boldsymbol{\phi}^{(\boldsymbol{x})} - \sum_{j=1}^{k} \log(1 + \exp(a_j)) \\ + \langle \boldsymbol{x}, \boldsymbol{B} \boldsymbol{\phi}^{(\boldsymbol{x})} \rangle + \langle \boldsymbol{x}, \boldsymbol{c} \rangle - \sum_{j=1}^{d} \log \left( 1 + w_j^{(\boldsymbol{x})} \right) \\ + \sum_{j=1}^{k} (\phi_j^{(\boldsymbol{x})} \log \phi_j^{(\boldsymbol{x})} + (1 - \phi_j^{(\boldsymbol{x})}) \log(1 - \phi^{(\boldsymbol{x})})_j) \end{array} \right)$$

Ignoring the constraints on $\boldsymbol{\phi}$, by first order optimality conditions we have:

$$\log \frac{\phi_m^{(\boldsymbol{x})}}{1 - \phi_m^{(\boldsymbol{x})}} = a_m + \langle \boldsymbol{x}^{(i)}, \boldsymbol{B}_{-,m} \rangle - \sum_{j=1}^{d} U_{j,m}^{(\boldsymbol{x})}$$

$$\phi_m^{(\boldsymbol{x})} = \sigma \left( a_m + \langle \boldsymbol{x}^{(i)}, \boldsymbol{B}_{-,m} \rangle - \sum_{j=1}^{d} U_{j,m}^{(\boldsymbol{x})} \right)$$

where $\boldsymbol{B}_{-,m}$ denotes column $m$ of matrix $\boldsymbol{B}$. Note that in the last equation, each $\phi_m^{(\boldsymbol{x})}$ depends on the values in $\boldsymbol{\phi}^{(\boldsymbol{x})}$ via the matrix $\boldsymbol{U}^{(\boldsymbol{x})}$. However, each vector $\phi^{(\boldsymbol{x})}$ is independent of the proposal parameters of other datapoints. We can solve the equations defined by first order optimality conditions via standard iterative methods, i.e. we initialize $\boldsymbol{\phi}^{(\boldsymbol{x})}$ randomly and update its value with the right-hand side of the last equation for a prefixed number of iterations. The resulting solution will always satisfies constraints on $\boldsymbol{\phi}^{(\boldsymbol{x})}$.

# References


[Jaakkola and Jordan, 1996] Jaakkola, T. S. and Jordan, M. I. (1996). Computing upper and lower bounds on likelihoods in intractable networks. In *Proceedings of the Twelfth Conference on Uncertainty in Artificial Intelligence (UAI1996).*

[Saul et al., 1996] Saul, L. K., Jaakkola, T., and Jordan, M. I. (1996). Mean field theory for sigmoid belief networks. *Journal of artificial intelligence research*, 4:61–76.