Notes on Mean Field Theory for Sigmoid Belief Networks

Caio Corro

September 18, 2021

This are my personal notes on the paper "Mean Field Theory for Sigmoid Belief Networks" (Saul, Jaakkola and Jordan).

1 Definition

Let $Z \in \{0,1\}^m$ and $X \in \{0,1\}^n$ be the sets of latent and observed random variables, respectively. We assume that the data is generated as follows:

$$\begin{aligned} &(1) \quad \boldsymbol{z} \sim p(Z;\mu) && \text{where } \mu = \sigma(\boldsymbol{a}) \\ &(2) \quad \boldsymbol{x} \sim p(X|Z=\boldsymbol{z};\phi) && \text{where } \phi = \sigma(\boldsymbol{B}\boldsymbol{z}+\boldsymbol{c}) \end{aligned}$$

where σ is the element-wise sigmoid function and $\boldsymbol{a} \in \mathbb{R}^m$, $\boldsymbol{B} \in \mathbb{R}^{n \times m}$ and $\boldsymbol{c} \in \mathbb{R}^n$ are the parameters of the model. The distributions $p(Z)$ and $p(X|Z)$ are distributions over vector of binary values where we assume independence between coordinates, i.e. $p(Z) = \prod_i p(Z_i)$ and $p(X|Z) = \prod_i p(X_i|Z)$. In order to to highlight the dependency between parameters and distribution, we will write $p(Z;\boldsymbol{a})$ and $p(X|Z;\boldsymbol{B},\boldsymbol{c})$. We will omit the parameters when clear from context.

During training, the goal is to maximize the log-likehood of a dataset of observed values $\boldsymbol{x}$ by tuning the parameters of the model. In theory, this can be done via gradient descent by marginalizing over latent variables, i.e. computing $\log p(\boldsymbol{x}) = \log \sum_{\boldsymbol{z}} p(\boldsymbol{x})p(\boldsymbol{x}|\boldsymbol{z})$. However, as we use a neural parameterization of the conditional distribution, it cannot be computed efficiently with a closed form solution. It is therefore required to sum over the 2^m possible latent variable assignments, which is prohibitive in practice. A second solution would be to rely on the Expectation-Maximization algorithm, but similar issues appear when computing the posterior distribution in the expectation step.

Using mean field theory, we show how we can efficiently derive a lower bound on the true likelihood which can be optimized via gradient ascent.

2 Mean Field Theory

We assume a parameterized proposal distribution where each coordinate is independent:

$$q(\boldsymbol{z}|\boldsymbol{x};\bar{\boldsymbol{\mu}}) = \prod_i q(z_i|\boldsymbol{x};\bar{\mu}_i)$$

where $\bar{\boldsymbol{\mu}} \in [0,1]^m$ are the parameters of the proposal distribution, i.e. each $q(z_i|\boldsymbol{x};a_i)$ is a Bernoulli parameterized by $\bar{\mu}_i$. We omit to explicitly note the dependency between $\bar{\boldsymbol{\mu}}$ and $\boldsymbol{x}$ for simplicity, but one proposal distribution per datapoint is computed in practice (i.e. the proposal is not amortized on the dataset). In this section, we highlight properties of statistical independency that will be useful in the following.

First, we show that the entropy of the joint distribution can be computed efficiently. We have:

$$H[q(Z_i|\boldsymbol{x})] = -\sum_{z_i} p(z_i|\boldsymbol{x}) = -p(0|\boldsymbol{x}) - p(1|\boldsymbol{x}) = -\bar{\mu}_i \log \bar{\mu}_i - (1-\bar{\mu}_i)\log(1-\bar{\mu}_i)$$

As each coordinate is independent, we can show that the entropy of the joint distribution decomposition as the sum of entropies:

$$H[q(Z|\boldsymbol{x})] = \sum_i H[q(Z_i|\boldsymbol{x})] = \sum_i -\bar{\mu}_i \log \bar{\mu}_i - (1-\bar{\mu}_i)\log(1-\bar{\mu}_i)$$

Note that the partial derivative wrt any given $\bar{\mu}_j$ as the form:

$$\begin{aligned}
\frac{\partial}{\partial \bar{\mu}_j} H[q(Z|\boldsymbol{x})] &= \frac{\partial}{\partial \bar{\mu}_j} \sum_i -\bar{\mu}_i \log \bar{\mu}_i - (1-\bar{\mu}_i) \log(1-\bar{\mu}_i) \\
&= \frac{\partial}{\partial \bar{\mu}_j} - \bar{\mu}_j \log \bar{\mu}_j - (1-\bar{\mu}_j) \log(1-\bar{\mu}_j) \\
&= -\log \bar{\mu}_j - 1 + \log(1-\bar{\mu}_j) + 1 \\
&= -\log \bar{\mu}_j + \log(1-\bar{\mu}_j) \\
&= -\log \frac{\bar{\mu}_j}{1-\bar{\mu}_j}
\end{aligned}$$

3 Efficient approximation of the log-likelihood

In this section we show that under the mean field assumption we can derive tractable lower bound of the true likelihood that can be optimized via gradient ascent. Let $q(\boldsymbol{z}|\boldsymbol{x};\bar{\boldsymbol{\mu}}) = \prod_i q(z_i|\boldsymbol{x};\bar{\mu}_i)$ be the proposal mean field distribution. The usual evidence lower bound is derived as follows:

$$\begin{aligned}
\log p(\boldsymbol{x}) &= \log \sum_{\boldsymbol{z}} p(\boldsymbol{z}) p(\boldsymbol{x}|\boldsymbol{z}) \\
&= \log \sum_{\boldsymbol{z}} \frac{q(\boldsymbol{z}|\boldsymbol{x})}{q(\boldsymbol{z}|\boldsymbol{x})} p(\boldsymbol{z}) p(\boldsymbol{x}|\boldsymbol{z})
\end{aligned}$$

By Jensen's inequality:

$$\begin{aligned}
&\geq \sum_{\boldsymbol{z}} q(\boldsymbol{z}|\boldsymbol{x}) \log \frac{p(\boldsymbol{z}) p(\boldsymbol{x}|\boldsymbol{z})}{q(\boldsymbol{z}|\boldsymbol{x})} \\
&= \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})} \left[\log p(\boldsymbol{z})\right] + \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})} \left[p(\boldsymbol{x}|\boldsymbol{z})\right] + H[q]
\end{aligned}$$

Note that the last term is constant wrt to the network parameters, so its gradient is not required. We are left with the two expectations. We show how to compute them efficiently, meaning that gradient wrt the parameters can be deduced via the backpropagation algorithm.

First, note that we have:

$$\log p(\boldsymbol{z};\boldsymbol{a}) = \log \prod_i p(z_i;\boldsymbol{a}) = \sum_i \log p(z_i;a_i) \qquad \text{where} \qquad \log p(z_i;a_i) = \begin{cases} \sigma(a_i) & \text{if } z_i = 1 \\ 1-\sigma(a_i) & \text{otherwise} \end{cases}$$

We can rewrite the probabilities as follow:

$$\begin{aligned}
\sigma(a_i) &= \frac{\exp(a_i)}{1+\exp(a_i)} = \frac{\exp(a_i z_i)}{1+\exp(a_i)} \quad \text{when } z_i = 1 \\
1-\sigma(a_i) &= \frac{1+\exp(a_i)}{1+\exp(a_i)} - \frac{\exp(a_i)}{1+\exp(a_i)} = \frac{1}{1+\exp(a_i)} = \frac{\exp(a_i z_i)}{1+\exp(a_i)} \quad \text{when } z_i = 0
\end{aligned}$$

We can therefore write:

$$\log p(\boldsymbol{z};\boldsymbol{a}) = \sum_i \log \frac{\exp(a_i z_i)}{1+\exp(a_i)} = \sum_i \left(a_i z_i - \log\left(1+\exp(a_i)\right)\right)$$

Taking the expectation under the proposal distribution, we have:

$$\begin{aligned}
\mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})} \left[\log p(\boldsymbol{z})\right] &= \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})} \left[\sum_i \left(a_i z_i - \log\left(1+\exp(a_i)\right)\right)\right] \\
&= \sum_i \left(\mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})} \left[a_i z_i\right] - \log\left(1+\exp(a_i)\right)\right) \\
&= \sum_i \left(\mathbb{E}_{q(z_i|\boldsymbol{x})} \left[(a_i z_i)\right] - \log\left(1+\exp(a_i)\right)\right) \\
&= \sum_i \left(a_i \bar{\mu}_i - \log\left(1+\exp(a_i)\right)\right)
\end{aligned}$$

Note that this term is easy to compute.

For the second term in the ELBO, note that we have:

$$\begin{aligned}
\log p(\boldsymbol{x}|\boldsymbol{z};\boldsymbol{B},\boldsymbol{c}) &= \sum_i \log p(x_i|\boldsymbol{z};\boldsymbol{B},\boldsymbol{c}) \\
&= \sum_i \log \frac{\exp(x_i(c_i + \sum_j B_{i,j} z_j))}{1 + \exp(c_i + \sum_j B_{i,j} z_j)} \\
&= \sum_i \left(c_i x_i + \sum_j B_{i,j} z_j x_i - \log(1 + \exp(c_i + \sum_j B_{i,j} z_j)) \right)
\end{aligned}$$

Taking the expectation wrt the proposal distribution, we have:

$$\begin{aligned}
\mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}\left[\log p(\boldsymbol{x}|\boldsymbol{z})\right] &= \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}\left[\sum_i \left(c_i x_i + \sum_j B_{i,j} z_j x_i - \log(1 + \exp(c_i + \sum_j B_{i,j} z_j)) \right)\right] \\
&= \sum_i \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}\left[c_i x_i + \sum_j B_{i,j} z_j x_i - \log(1 + \exp(c_i + \sum_j B_{i,j} z_j))\right] \\
&= \sum_i c_i x_i + \sum_j B_{i,j} \bar{\mu}_j x_i - \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}\left[\log(1 + \exp(c_i + \sum_j B_{i,j} z_j))\right]
\end{aligned}$$

Unfortunately the last term is intractable as we cannot break down the expectation over the joint distribution in simpler computation. However, we can derive a lower bound using Jensen's inequality:

$$\begin{aligned}
&\geq \sum_i c_i x_i + \sum_j B_{i,j} \bar{\mu}_j x_i - \log\left(\mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}\left[1 + \exp(c_i)\prod_j \exp(B_{i,j} z_j)\right]\right) \\
&= \sum_i c_i x_i + \sum_j B_{i,j} \bar{\mu}_j x_i - \log\left(1 + \exp(c_i)\prod_j (\bar{\mu}_j \exp(B_{i,j}) + 1 - \bar{\mu}_j)\right)
\end{aligned}$$

which can be efficiently computed.

As such, we showed that if we are given the proposal distribution parameters $\bar{\boldsymbol{\mu}}$ we can efficiently learn the parameters $\boldsymbol{a}, \boldsymbol{B}$ and $\boldsymbol{c}$ of the SBN via gradient ascent. In the next section, we show how we can compute the optimal proposal mean field distribution.

4 Proposal distribution estimation

To compute the parameters $\bar{\boldsymbol{\mu}}$, i.e. the expectation step, we can maximize the ELBO wrt to $\bar{\boldsymbol{\mu}}$:

$$\begin{aligned}
\arg\max_{\bar{\boldsymbol{\mu}}} \quad & \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}[\log p(\boldsymbol{z}) p(\boldsymbol{x}|\boldsymbol{z})] + H[q] \\
\text{s.t.} \quad & \bar{\mu}_i \geq 0 \qquad \forall i \\
& \bar{\mu}_i \leq 1 \qquad \forall i
\end{aligned}$$

We will ignore the inequality constraints for simplicity. A more formal derivation should include them, but their contribution to the stationarity conditions will vanish as the associated dual variables are fixed to zero by complementary slackness and the fact the each $\bar{\mu}_i$ will have a sigmoidal shape.

Hence, if we focus on the unconstrained variant of the problem that is concave and differentiable, we can derive optimal parameters $\bar{\boldsymbol{\mu}}$ it by solving the following system of equations:

$$\nabla_{\bar{\boldsymbol{\mu}}}\left(\mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}[\log p(\boldsymbol{z}) p(\boldsymbol{x}|\boldsymbol{z})] + H[q]\right) = \boldsymbol{0}$$

However, as we saw in the previous section, the expectation step is intractable. We replace this objective by

the lower bound from previous section and obtain the following partial derivative constraint:

$$\begin{aligned}
\log \frac{\bar{\mu}_k}{1-\bar{\mu}_k} &= \frac{\partial}{\partial \bar{\mu}_k}\left(\sum_i c_i x_i + \sum_j B_{i,j}\bar{\mu}_j x_i - \log\left(1+\exp(c_i)\prod_j(\bar{\mu}_j\exp(B_{i,j})+1-\bar{\mu}_j)\right)\right) \\
&= B_{i,k}x_i - \frac{1}{1+\exp(c_i)\prod_j(\bar{\mu}_j\exp(B_{i,j})+1-\bar{\mu}_j)}\frac{\partial}{\partial\bar{\mu}_k}\left(1+\exp(c_i)\prod_j(\bar{\mu}_j\exp(B_{i,j})+1-\bar{\mu}_j)\right) \\
&= B_{i,k}x_i - \frac{(\exp(B_{i,k})-1)\exp(c_i)\prod_{j\neq k}(\bar{\mu}_j\exp(B_{i,j})+1-\bar{\mu}_j)}{1+\exp(c_i)\prod_j(\bar{\mu}_j\exp(B_{i,j})+1-\bar{\mu}_j)} \\
\bar{\mu}_k &= \sigma\left(B_{i,k}x_i - \frac{(\exp(B_{i,k})-1)\exp(c_i)\prod_{j\neq k}(\bar{\mu}_j\exp(B_{i,j})+1-\bar{\mu}_j)}{1+\exp(c_i)\prod_j(\bar{\mu}_j\exp(B_{i,j})+1-\bar{\mu}_j)}\right)
\end{aligned}$$

Note that this does not result in a closed form expression for $\bar{\boldsymbol{\mu}}$. However, we can compute iteratively a solution to this set of equations as follows:

$$\bar{\mu}_k^{(t+1)} = \sigma\left(B_{i,k}x_i - \frac{(\exp(B_{i,k})-1)\exp(c_i)\prod_{j\neq k}(\bar{\mu}_j^{(t)}\exp(B_{i,j})+1-\bar{\mu}_j^{(t)})}{1+\exp(c_i)\prod_j(\bar{\mu}_j^{(t)}\exp(B_{i,j})+1-\bar{\mu}_j^{(t)})}\right)$$