

## 0.1. Unsupervised learning

The results of experiments on toy models suggest that the initial unsatisfactory results with naive mean field approaches [tieleman2008training] might be greatly improved if add additional terms responsible for better estimation of (connections) between the spins.

[dayan1999unsupervised]

Our general goal is to maximize the probability of  $\mathcal{D}$  under the MRF distributions – thus we are looking for the vector of parameters  $\theta$  that maximize the likelihood given the training data, i.e.

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

In most of the cases, it is not possible to find the analytical solution for the maximum likelihood parameters and we need to resort to some approximation methods.

## 0.2. Unsupervised Pre-training of Neural Networks

add Erham here

## 0.3. Training of Boltzmann Machines

As it

$$\theta^{t+1} = \theta^t + \eta \frac{\partial}{\partial \theta^t} \ln \mathcal{L}(\theta|\mathcal{D}) \quad (2)$$

This relies on the fact that the gradient w.r.t. parameters  $\theta$  informs us how fast function increases in the current point  $\theta^t$ . By taking appropriately small learning rate, these iterative updates might converge to the maximum of the function. However, there is no guarantees that this procedure will lead to obtaining global maximum.

TODO - stochastic gradient descent - theoretical results - writeabout it and read theoretical results.

Learning Restricted Boltzmann machines relies on gradient ascent of the log-likelihood. The gradient of the log-likelihood from given a training example  $\mathbf{v}$  takes the form:

$$\begin{aligned} \frac{\partial \log \mathcal{L}(\theta|\mathbf{v})}{\partial \theta} &= \frac{\partial \mathcal{F}^c}{\partial \theta} - \frac{\partial \mathcal{F}}{\partial \theta} \\ &= - \frac{\sum_{\mathbf{h}} e^{-E(\mathbf{v}, \mathbf{h})} \frac{\partial E(\mathbf{v}, \mathbf{h})}{\partial \theta}}{\sum_{\mathbf{h}} e^{-E(\mathbf{v}, \mathbf{h})}} + \frac{\sum_{\mathbf{v}, \mathbf{h}} e^{-E(\mathbf{v}, \mathbf{h})} \frac{\partial E(\mathbf{v}, \mathbf{h})}{\partial \theta}}{\sum_{\mathbf{v}, \mathbf{h}} e^{-E(\mathbf{v}, \mathbf{h})}} \\ &= - \sum_{\mathbf{h}} p(\mathbf{h}|\mathbf{v}) \frac{\partial E(\mathbf{v}, \mathbf{h})}{\partial \theta} + \sum_{\mathbf{v}, \mathbf{h}} p(\mathbf{v}, \mathbf{h}) \frac{\partial E(\mathbf{v}, \mathbf{h})}{\partial \theta} \\ &= -\mathbb{E}_{p(\mathbf{h}|\mathbf{v})} \left( \frac{\partial E(\mathbf{v}, \mathbf{h})}{\partial \theta} \right) + \mathbb{E}_{p(\mathbf{v}, \mathbf{h})} \left( \frac{\partial E(\mathbf{v}, \mathbf{h})}{\partial \theta} \right) \end{aligned} \quad (3)$$

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 BM, the first term can be computed analytically ???. 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.

## 0.4. MCMC Sampling

The

## 0.5. Contrastive Divergence

### 0.5.1. Persistent CD

## 0.6. Learning in the TAP case

### 0.6.1. Gradients

Eq. 11. Gradients:

$$\begin{aligned}w_{ij}^{t+1} &= w_{ij}^t + \eta \Delta w_{ij}^{t+1} \\ \Delta w_{ij}^{t+1} &\propto \frac{\partial \mathcal{L}}{\partial w_{ij}} \simeq -\frac{\partial F}{\partial w_{ij}} - \frac{\partial F^{EMF}}{\partial w_{ij}} \\ \frac{\partial F^{EMF}}{\partial w_{ij}} &= -m_i^v m_j^h - w_{ij}^t (m_i^v - (m_i^v)^2)(m_j^h - (m_j^h)^2) \\ &\quad - 2w_{ij}^2 (m_i^v - (m_i^v)^2) \left(\frac{1}{2} - m_i^v\right) (m_j^h - (m_j^h)^2) \left(\frac{1}{2} - m_j^h\right) \\ \frac{\partial \mathcal{L}}{\partial a_i} &= \frac{\partial F^{EMF}}{\partial a_i} = -m_i^v \\ \frac{\partial \mathcal{L}}{\partial b_j} &= \frac{\partial F^{EMF}}{\partial b_j} = -m_j^h\end{aligned}$$

An example of such structure presents Figure ??.

## 0.7. Approximating the likelihood

TODO - describe problems and procedure.

Figure 1: updatesMF

Figure 2: updatesTAP2

### 0.7.1. Pseudo approximation

The problems mentioned above makes training such structure very difficult because we cannot observe directly progress along learning. Thus, we need to resort to some approximations. One of the most popular approaches is due to Besag [besag1972nearest]. Consider the following approximation:

$$P(\mathbf{s}) = \prod_i p(s_i | s_1, \dots, s_{i-1}) \approx \prod_i p(s_i | s_1, \dots, s_{i-1}, s_{i+1}, \dots, s_n) = \prod_i p(s_i | s_{-i}) \quad (4)$$

where the first equation comes from the chain rule. Here we assume that particular marginals given all other dimensions are independent of each other.

**Theorem 1** *TODO: Assume that  $\mathbf{x}$  is generated I.ID. by a distribution  $p(\mathbf{x}; \theta)$ .*

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.

$$\log PL(\mathbf{s}) = N\mathbb{E}(\log P(s_i|\mathbf{s}_{-i})), \quad (5)$$

where  $i \sim U(1, N)$ .

In the case of the analysed model we obtain the following form using Monte Carlo approximation (TODO: monte carlo approximation):

$$\log PL(\mathbf{s}) \approx N \log \left( \frac{\exp\{-F(\mathbf{s})\}}{\exp\{-F(\hat{\mathbf{s}})\} + \exp\{-F(\mathbf{s})\}} \right), \quad (6)$$

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

## 0.8. Real scale model

### 0.8.1. MNIST data set

## 0.9. Comparison

In order to test the efficiency of the EMF learning algorithm I used three approximations of ?? – at first-order (MF), second-order (TAP2) and third order (TAP3).

All trained models used the same set-up of free parameters. The purpose of this experiment is to compare different RBM trainings thus following [gabrie2015training] I didn't use the adaptive learning rate which was set to 0.005, learning was performed using mini-batch learning 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 where the spins are independent. This means that the couplings should have small values – this can be enforced using regularization which at the same times allows for a better regularization. From probabilistic perspective it can be seen as adding a weighted prior over the parameters (maximum a posteriori). This leads to a new criterion which will be maximized of the form:

$$E(\theta, \mathcal{D}) = \ln \mathcal{L}(\theta|\mathcal{D}) - \lambda R(\theta). \quad (7)$$

In all experiments I used Laplacian prior  $R(\theta) = \|\theta\|_1$  (L1 regularization) with the weight  $\lambda$  set to 0.01.

Figure 0.9 presents COMment Figure ?? shows comment

Figure 3: Pseudo LL.

Figure 4: EMF LL.