0.1. Usupervised 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 θ that maximize the likelihood given the training data, i.e.

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

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

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

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. 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 theoeretical 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:

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

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:

$$w_{ij}^{t+1} = w_{ij}^{t} + \eta \triangle w_{ij}^{t+1}$$

$$\triangle w_{ij}^{t+1} \propto \frac{\partial \mathcal{L}}{\partial w_{ij}} \simeq -\frac{\partial F}{\partial w_{ij}} - \frac{\partial F^{EMF}}{\partial w_{ij}}$$

$$\begin{aligned}\frac{\partial F^{EMF}}{\partial w_{ij}} =& - m_i^v m_j^h - w_{ij}^t (m_i^v - (m_i^v)^2)(m_j^h - (m_j^h)^2) \\ & - 2 w_{ij}^2 (m_i^v - (m_i^v)^2)(\frac{1}{2} - m_i^v)(m_j^h - (m_j^h)^2)(\frac{1}{2} - m_j^h)\end{aligned}$$

$$\frac{\partial \mathcal{L}}{\partial a_i} = \frac{\partial F^{EMF}}{\partial a_i} = -m_i^v$$

$$\frac{\partial \mathcal{L}}{\partial b_j} = \frac{\partial F^{EMF}}{\partial b_j} = -m_j^h$$

An example of such structure presents Figure **??**.

0.7. Approximating the likelihood

TODO - describe problems and procedure.

0.7.1. Annealed Importance Sampling (AIS)

The most widely used technique is based on a very simple identity. Assume we have two distributions $p_A = \frac{1}{Z_A} p_A^*(\mathbf{x})$, $p_B = \frac{1}{Z_B} p_B^*(\mathbf{x})$ where $p^*()$ denotes unnormalized distribution and Z_A, Z_B are partition functions. Assuming that a proposal p_A distribution p_A supports tractable sampling and tractable evaluation of both the unnormalized distribution $p_A^*(\mathbf{x})$ and the partition function Z_A we can use the following relation:

$$\begin{aligned}Z_B &= \int p_B^*(\mathbf{x}) \mathrm{d}\mathbf{x} \\ &= \int \frac{p_A(\mathbf{x})}{p_A(\mathbf{x})} p_B^*(\mathbf{x}) \mathrm{d}\mathbf{x} \\ &= Z_A \int \frac{p_B^*(\mathbf{x})}{p_A^*(\mathbf{x})} p_A(\mathbf{x}) \mathrm{d}\mathbf{x}\end{aligned} \tag{4}$$

Sampling from the tractable distribution, we can derive Monte Carlo estimator of the ratio between partition functions:

$$\frac{Z_B}{Z_A} \approx \frac{1}{N} \sum_{i=1}^{N} \frac{p_B^*(\mathbf{x}^{(i)})}{p_A^*(\mathbf{x}^{(i)})} = \hat{r}_{SIS} \tag{5}$$

where $\mathbf{x}^{(i)}$ comes from p_A. Assuming that distribution p_A is close to p_B, the estimator from 5 called simple importance sampling proves to work well [**minka2005divergence**]. However, in high-dimensional spaces where p_B is usually multimodal as it is considered in this thesis, the variance of the estimator from 5 might be very high.

The idea presented above might be improved by following the classic approach from probabilistic optimization i.e. simulated annealing. The idea is to introduce intermediate distributions that will allow to bridge the gap between two considered distributions p_A and p_B [**jarzynski1997nonequilibrium**], [**neal2001annealed**].

Consider a sequence of distributions $p_0, p_1, ..., p_M$ where $p_0 = p_A$ and $p_M = p_B$. If the intermediate distributions p_m and p_{m+1} are close enough, a simple estimator from 5 can be used to estimate each ratio $\frac{Z_{m+1}}{Z_m}$. Using the the following identity:

$$\frac{Z_M}{Z_0} = \frac{Z_1}{Z_0}\frac{Z_2}{Z_1}...\frac{Z_M}{Z_{M-1}} \tag{6}$$

those intermediate ratios are then combined to obtain the estimate of $\frac{Z_B}{Z_A}$. There is no need to compute the normalizing constants of any intermediate distributions. The intermediate distributions are chosen to suit a given problem domain. However in most cases, we are able to draw exact samples only from the first tractable distribution p_A. In order to sample from intermediate distribution we have be able to draw a sample $\mathbf{x}^{'}$ given $\mathbf{x}$ using Markov chain transition operator $T_m(\mathbf{x}^{'}|\mathbf{x})$ that leaves $p_m(\mathbf{x})$ invariant, i.e.:

$$\int T_m(\mathbf{x}^{'}|\mathbf{x})p_m(\mathbf{x})\mathrm{d}\mathbf{x} = p_m(\mathbf{x}^{'}) \tag{7}$$

These transition operators represent the probability density of transitioning from state $\mathbf{x}$ to $\mathbf{x}^{'}$ [**salakhutdinov2008learning**]. Having obtained the sequence of samples from the intermediate distributions we can obtain the improved estimator of the ratio between partition functions following the procedure:

Algorithm 1 Annealed Importance Sampling.

Set p_A and p_B with appropriate parameters
for $i \in \{1, ..., N\}$ **do**
  sample $\mathbf{x}_1$ from $p_0 = p_A$
  sample $\mathbf{x}_2$ via $T_1(\mathbf{x}_2|\mathbf{x}_1)$
  ...
  sample $\mathbf{x}_M$ via $T_M(\mathbf{x}_M|\mathbf{x}_{M-1})$
  $r_{AIS}^{(i)} = \frac{p_1^*(\mathbf{x}_1)}{p_0^*(\mathbf{x}_1)}\frac{p_2^*(\mathbf{x}_2)}{p_1^*(\mathbf{x}_2)}...\frac{p_M^*(\mathbf{x}_M)}{p_{M-1}^*(\mathbf{x}_M)}$
end for
$\hat{r}_{AIS} = \frac{1}{N}\sum_{i=1}^{N}\hat{r}_{AIS}^{(i)}$

It was proven that the variance of $\hat{r}_{AIS}$ will be proportional to $1/MN$ assuming we used sufficiently large numbers of intermediate distributions M [**neal2001annealed**]. Moreover, the estimate of Z_M/Z_0 will be unbiased if each ratio is estimated using $N = 1$ and a sample $\mathbf{x}^m$ is obtained using Markov chain starting at previous sample. This follows from the observation that the AIS procedure is an simple importance sampling defined on an extended state space $\mathcal{X} = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_M)$.

The procedure described above can be adapted to the RBM case – assume that we have estimated parameters θ_B of the model that we want to evaluate. Following [**salakhutdinov2008quantitative**] as a tractable starting distribution p_A we can use "clamped" restricted Boltzmann machine where there is no hidden layer. The sequence of intermediate distribution is then defined as:

$$p_m(\mathbf{v}) = \frac{1}{Z_m}p_m^*(\mathbf{v}) = \frac{1}{Z_m}\sum_{\mathbf{h}}\exp(-E_m(\mathbf{v}, \mathbf{h})) \tag{8}$$

where $m = 0, ..., M$, $\mathbf{h} = \mathbf{h_B}$, and the energy function has the form:

$$E_m(\mathbf{v}, \mathbf{h}) = (1 - \beta_m)E(\mathbf{v}; \theta_A) + \beta_m E(\mathbf{v}, \mathbf{h}; \theta_B) \tag{9}$$

where $\beta_m \in [0, 1]$ with $\beta_m = 0$ yielding p_A and $\beta_m = 1$ giving p_B. Annealing slowly the "temperature" from infinity to zero we gradually moves from the state space of proposal distribution to the space defined by the untractable distribution. Following the approach from **??** we can obtain transition operators for hidden and visible variables:

$$\begin{aligned} p(h^A|\mathbf{v}) &= \sigma((1-\beta)NONE \\ p(h^B) &= \end{aligned} \tag{10}$$

Algorithm 2 Appendix – Implementation for BM and RBM.

$\theta, \phi \leftarrow .$
while not converged in θ, ϕ **do**
  Pick subset of size $\mathbf{x}_{1:M}$ from the full dataset uniformly at random.
  Compute $g \leftarrow \nabla_{\theta,\phi} \tilde{\mathcal{L}}^B \theta, \phi; \mathbf{x}_{1:M}, \epsilon_{1:M}$
end while

0.7.2. 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:

$$P(\mathbf{s}) = \prod_i p(s_i|s_1, ..., s_{i-1}) \approx \prod_i p(s_i|s_1, ..., s_{i-1}, s_{i+1}, ..., s_n) = \prod_i p(s_i|s_{-i}) \tag{11}$$

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.

$$\log PL(\mathbf{s}) = N\mathbb{E}\left(\log P(s_i|\mathbf{s}_{-i})\right), \tag{12}$$

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):

$$\log PL(\mathbf{s}) \approx N \log\left(\frac{\exp\{-F(\mathbf{s})\}}{\exp\{-F(\hat{\mathbf{s}})\} + \exp\{-F(\mathbf{s})\}}\right), \tag{13}$$

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 following

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 variance set to 0.001. 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. 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 the new criterion which we will maximize of the form:

$$E(\theta, \mathcal{D}) = \ln \mathcal{L}(\theta|\mathcal{D}) - \lambda R(\theta) \tag{14}$$

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