Deep Learning

Bastien BRIER
bastien.brier@student.ecp.fr

January 5, 2017

Assignment 2

1 Analytic exercise

Given a set of training samples $\mathbf{X} = \{\mathbf{x}^1, ...\mathbf{x}^M\}$, the log-likelihood $\mathbf{X}$ is:

$$\begin{aligned}
S(\mathbf{X}, \mathbf{W}) &= \sum_{m=1}^{M} \log P(\mathbf{x}^m; \mathbf{W}) \\
&= \sum_{m=1}^{M} \log \sum_{\mathbf{h}} \frac{1}{Z} \exp(-E(\mathbf{x}^m, \mathbf{h}; \mathbf{W})) \\
&= \sum_{m=1}^{M} \log \frac{1}{Z} \sum_{\mathbf{h}} \exp(-E(\mathbf{x}^m, \mathbf{h}; \mathbf{W})) \\
&= \sum_{m=1}^{M} \log \sum_{\mathbf{h}} \exp(-E(\mathbf{x}^m, \mathbf{h}; \mathbf{W})) - \sum_{m=1}^{M} \log Z
\end{aligned}$$

with

$$E(\mathbf{x}, \mathbf{h}; \mathbf{W}) = E(\mathbf{y}; \mathbf{W}) = -\frac{1}{2}\mathbf{y}^T\mathbf{W}\mathbf{y} \text{ (with } \mathbf{y} = \begin{bmatrix} \mathbf{x} \\ \mathbf{h} \end{bmatrix})$$

$$Z = \sum_{\mathbf{h}} \sum_{\mathbf{x}} \exp(-E(\mathbf{x}, \mathbf{h}; \mathbf{W})) = \sum_{\mathbf{y}} \exp(-E(\mathbf{y}; \mathbf{W}))$$

As derivation is linear, we can compute the derivative of the first part of the equation and then the second, so:

$$\begin{aligned}
\frac{\partial \log \sum_{\mathbf{h}} \exp(-E(\mathbf{x}^m, \mathbf{h}; \mathbf{W}))}{\partial w_{k,l}} &= \frac{1}{\sum_{\mathbf{h}} \exp(-E(\mathbf{x}^m, \mathbf{h}; \mathbf{W}))} \times \frac{\partial \sum_{\mathbf{h}} \exp(-E(\mathbf{x}^m, \mathbf{h}; \mathbf{W}))}{\partial w_{k,l}} \\
&= \frac{1}{\sum_{\mathbf{h}} \exp(-E(\mathbf{x}^m, \mathbf{h}; \mathbf{W}))} \times \sum_{\mathbf{h}} \frac{\partial \exp(-E(\mathbf{x}^m, \mathbf{h}; \mathbf{W}))}{\partial w_{k,l}} \\
&= \frac{1}{\sum_{\mathbf{h}} \exp(-E(\mathbf{x}^m, \mathbf{h}; \mathbf{W}))} \times \sum_{\mathbf{h}} \exp(-E(\mathbf{x}^m, \mathbf{h}; \mathbf{W})) \times \frac{\partial(-E(\mathbf{x}^m, \mathbf{h}; \mathbf{W}))}{\partial w_{k,l}} \\
&= \frac{\sum_{\mathbf{h}} \mathbf{y}_k \mathbf{y}_l \exp(-E(\mathbf{x}^m, \mathbf{h}; \mathbf{W}))}{\sum_{\mathbf{h}} \exp(-E(\mathbf{x}^m, \mathbf{h}; \mathbf{W}))}
\end{aligned}$$

Moreover, we know that:

$$P(\mathbf{h}|\mathbf{x}^m; \mathbf{W}) = \frac{\exp(-E(\mathbf{x}^m, \mathbf{h}; \mathbf{W}))}{\sum_{\mathbf{h}} \exp(-E(\mathbf{x}^m, \mathbf{h}; \mathbf{W}))}$$

Then:

$$\begin{aligned}
\frac{\partial \log \sum_{\mathbf{h}} \exp(-E(\mathbf{x}^m, \mathbf{h}; \mathbf{W}))}{\partial w_{k,l}} &= \sum_{\mathbf{h}} \mathbf{y}_k \mathbf{y}_l P(\mathbf{h}|\mathbf{x}^m; \mathbf{W}) \\
&= \langle \mathbf{y}_k \mathbf{y}_l \rangle_{P(\mathbf{h}, \mathbf{x}^m; \mathbf{W})}
\end{aligned}$$

We compute the derivation of the second part of the equation:

$$\begin{aligned}
\frac{\partial \log Z}{\partial w_{k,l}} &= \frac{1}{Z} \times \frac{\partial \sum_{\mathbf{h},\mathbf{x}} \exp(-E(\mathbf{x}, \mathbf{h}; \mathbf{W}))}{\partial w_{k,l}} \\
&= \frac{1}{Z} \sum_{\mathbf{h},\mathbf{x}} \frac{\partial \exp(-E(\mathbf{x}, \mathbf{h}; \mathbf{W}))}{\partial w_{k,l}} \\
&= \frac{1}{Z} \sum_{\mathbf{h},\mathbf{x}} \frac{\partial(-E(\mathbf{x}, \mathbf{h}; \mathbf{W}))}{\partial w_{k,l}} \times \exp(-E(\mathbf{x}, \mathbf{h}; \mathbf{W})) \\
&= \frac{1}{Z} \sum_{\mathbf{h},\mathbf{x}} \mathbf{y}_k \mathbf{y}_l \exp(-E(\mathbf{x}, \mathbf{h}; \mathbf{W})) \\
&= \frac{1}{Z} \sum_{\mathbf{y}} \mathbf{y}_k \mathbf{y}_l \exp(-E(\mathbf{y}; \mathbf{W}))
\end{aligned}$$

We know that:

$$P(\mathbf{y}; \mathbf{W}) = \frac{1}{Z} \exp(-E(\mathbf{y}; \mathbf{W}))$$

Therefore:

$$\begin{aligned}\frac{\partial \log Z}{\partial w_{k,l}} &= \sum_{\mathbf{y}} \mathbf{y}_k \mathbf{y}_l P(\mathbf{y}; \mathbf{W}) \\ &= \sum_{\mathbf{h},\mathbf{x}} \mathbf{y}_k \mathbf{y}_l P(\mathbf{h}, \mathbf{x}; \mathbf{W}) \\ &= \langle \mathbf{y}_k \mathbf{y}_l \rangle_{P(\mathbf{h},\mathbf{x};\mathbf{W})}\end{aligned}$$

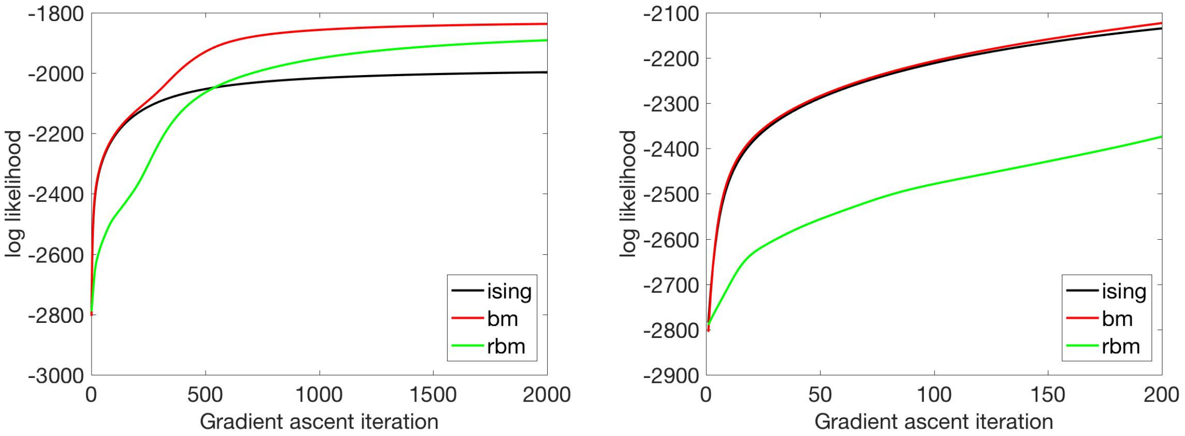
Putting it all together:

$$\begin{aligned}\frac{\partial S(\mathbf{X}, \mathbf{W})}{\partial w_{k,l}} &= \sum_{m=1}^{M} \langle \mathbf{y}_k \mathbf{y}_l \rangle_{P(\mathbf{h},\mathbf{x}^m;\mathbf{W})} - \sum_{m=1}^{M} \langle \mathbf{y}_k \mathbf{y}_l \rangle_{P(\mathbf{h},\mathbf{x};\mathbf{W})} \\ &= \sum_{m=1}^{M} [\langle \mathbf{y}_k \mathbf{y}_l \rangle_{P(\mathbf{h},\mathbf{x}^m;\mathbf{W})} - \langle \mathbf{y}_k \mathbf{y}_l \rangle_{P(\mathbf{h},\mathbf{x};\mathbf{W})}]\end{aligned}$$

2 Exact summations

We use brute-force summations to obtain the exact value of the log-likelihood for three models : Ising Model, Boltzmann Machine and Restricted Boltzmann Machine.
We implemented gradient-ascent and here are the plots obtained.

Figure 1: Brute-force summation, 8 hidden layers

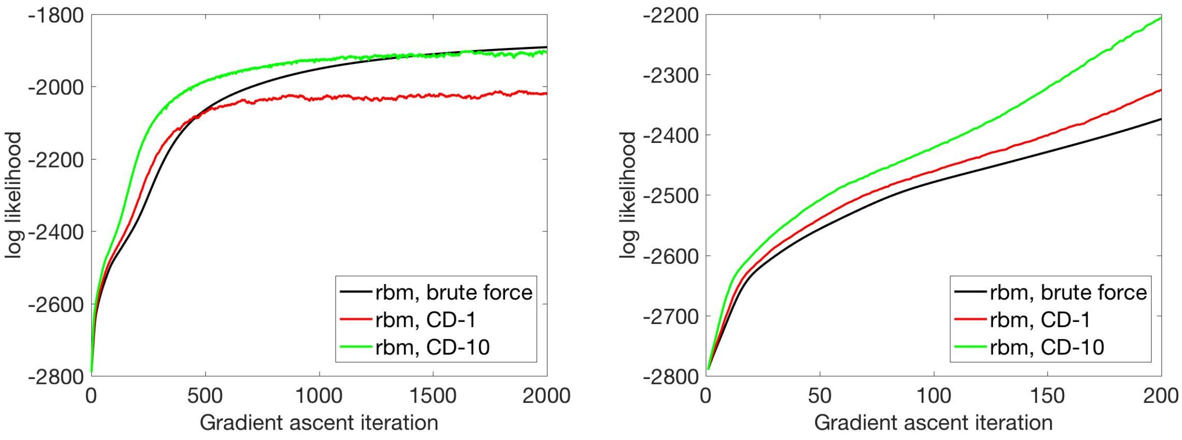


As expected, the best performance is reached by the full Boltzmann machine, as it is fully connected between observable and hidden layers. As the first estimations are based on the observable variables, the Ising model and Boltzmann Machine perform way better than the Restricted BM in the right plot. But then, the hidden layers improve the accuracy of the summation and, from the 500th iteration, the Restricted Boltzmann Machine has better performance than the Ising Model. In the end, the performance of the BM and RBM are better than the Ising Model.

3 Block-Gibbs sampling and Contrastive Divergence

After having implemented Block-Gibbs sampling, we used contrastive divergence with L = 1 and L = 10 to train a Restricted Boltzmann Machine with 8 hidden layers. Here are the plots obtained.

Figure 2: Contrastive Divergence, 8 hidden layers



As expected, after 2000 iterations, the best result is obtained by the brute-force summation. But, the CD approximations have better results in the beginning. The principle of Gibbs sampling is that we condition x (visible) and sample h, and then condition h and sample x, L times. Is is then not surprising that the CD with L=10 has better results than the one with L=1. For 2000 iterations, the result of the CD with L=10 is really close to the brute-force summation.