Restricted Boltzmann Machine

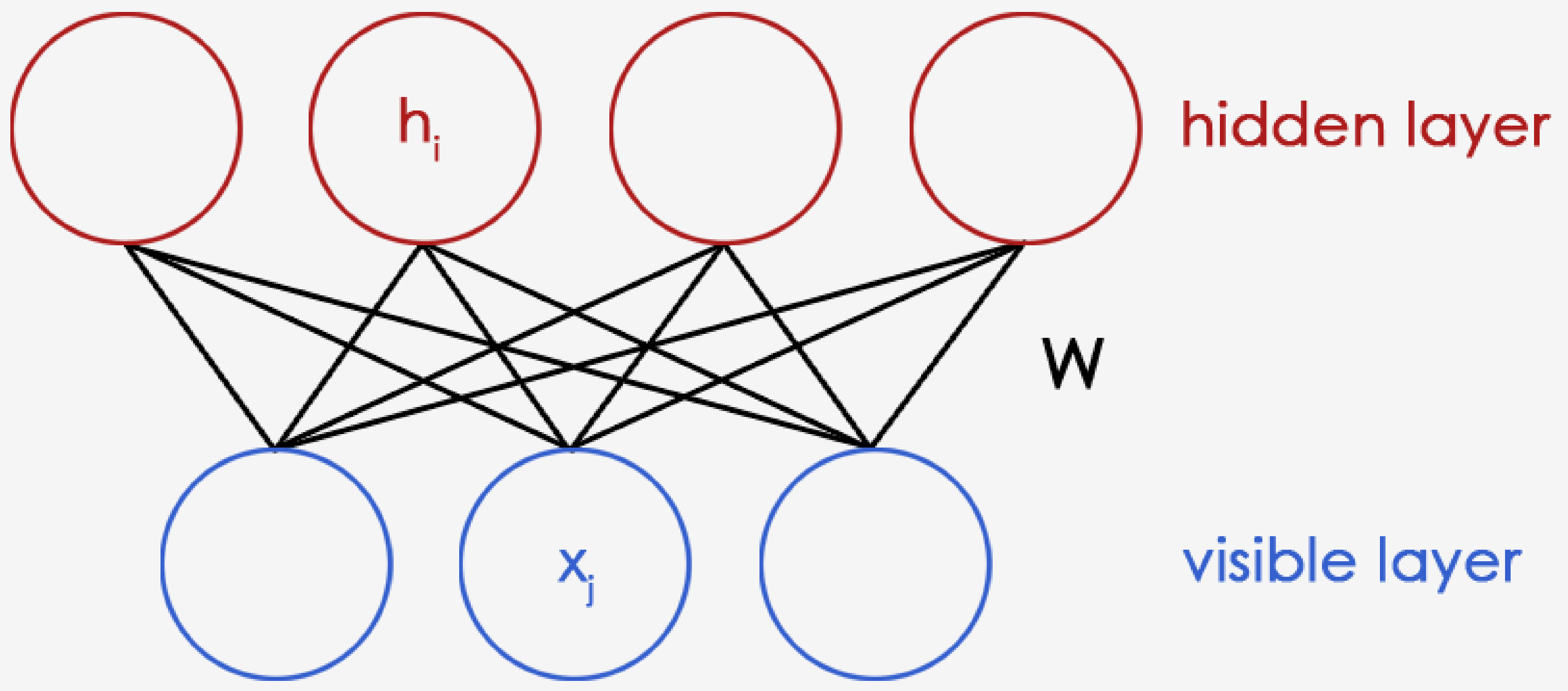
A Probabilistic Graphical Model course's project

Chaïmaa Kadaoui, Othman Sbai, Xi Shen

école normale supérieure paris—saclay

Definition

A Restricted Boltzman Machine is an undirected graph which can be decomposed in two layers.



It is *restricted* because no connection between nodes of the same layer is supposed. The matrix W characterizes the connections between the two layers. For simplification, we suppose that the variables $\mathbf{x}$ and $\mathbf{h}$ are binary.

Energy and probability

We introduce the two bias vectors c and b to define the energy of this graph:

$$E(x,h) = -h^\top W x - c^\top x - b^\top h$$

We can write the joint probability of $\mathbf{x}$ and $\mathbf{h}$ as:

$$p(x,h) = \exp(-E(x,h))/Z$$

Z is the partition parameter which can be computed by calculating all the possibles values of $\mathbf{x}$ and $\mathbf{h}$. In practice, this parameter is intractable.

Inference We can prove that given $\mathbf{x}$, h_j follows a Bernoulli:

$$p(h_j = 1|x) = sigm(b_j + W_{j.}x)$$

$W_{j.}$ is the j-th row of W and $sigm$ defines the sigmoid function.

Contrastive Divergence

To train the RBM with our data $\mathbf{x}^{(t)}$, we would like to minimize the average loss: $\frac{1}{T}\sum_t l(f(\mathbf{x}^{(t)})) = \frac{1}{T}\sum_t -\log p(\mathbf{x}^{(t)})$ We will apply a stochastic gradient descent. We derive each term of the sum with respect to our model parameter θ as follow (where $\mathbb{E}_h$ and $\mathbb{E}_{x,h}$ are expectations of h and (x,h) respectively):

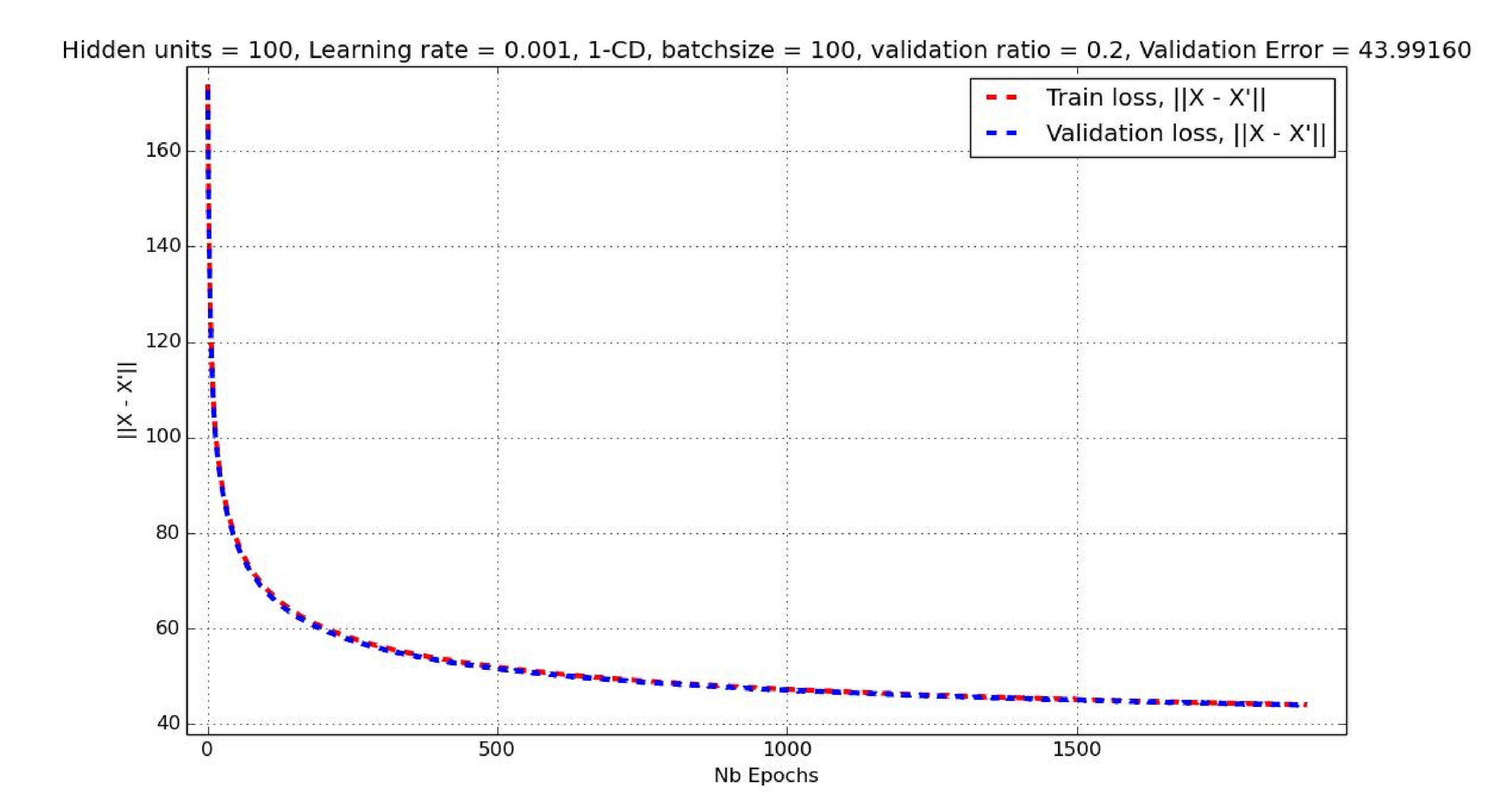
$$\frac{\partial -\log p(\mathbf{x}^{(t)})}{\partial \theta} = \mathbb{E}_h\left[\frac{\partial E(\mathbf{x}^{(t)},h)}{\partial \theta}|\mathbf{x}^{(t)}\right] - \mathbb{E}_{x,h}\left[\frac{\partial E(\mathbf{x},h)}{\partial \theta}\right]$$

The first term is called the *positive phase* and the second term the *negative phase*. Because of the difficulty to compute the seconde term, we will use an algorithm called *Contrastive Divergence*. The key points of the algorithm are:

- We estimate the expectation $\mathbb{E}_{x,h}$ by sampling a single point $\tilde{\mathbf{x}}$.
- To do so, we use Gibbs sampling in chain (we apply it k times).
- We initialize our Gibbs sampling with $\mathbf{x^{(t)}}$.

Results

We implemented the algorithm and applied it to the MNIST dataset.



Interpretation

References

[1] Hinton G. *A practical guide to training restricted Boltzmann machines[J]. Momentum, 2010, 9(1): 926.*

[2] Fischer A. *Training restricted boltzmann machines[J]. KI-Künstliche Intelligenz, 2015, 29(4): 441-444.*

[3] Bengio, Y., Delalleau, O. *Justifying and generalizing contrastive divergence.* Neural Computation 21(6), 1601n621 (2009)