ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

**Unsupervised and reinforcement learning in neural networks**

Week 8

Lecturer: Wulfram Gerstner

Assistants: Carlos Stein Naves de Brito
Christian Pozzorini

# Restricted Boltzmann Machines

## Introduction

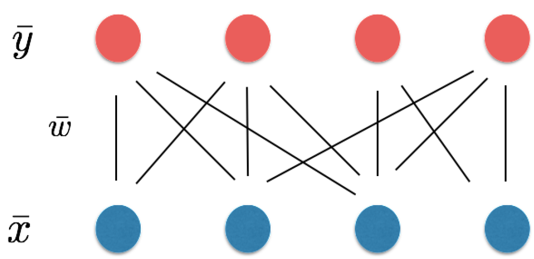
A restricted Boltzmann machine (RBM) is a generative stochastic neural network that can learn a probability distribution over its set of inputs. RBMs are a special case of Boltzmann machines, with the restriction that their neurons form a bipartite graph, with an input layer and a hidden layer, with no lateral connections.

The standard RBM has binary units, $x_i, y_j \in \{0, 1\}$, and a connection matrix of weights $W = (w_{i,j})$ connecting hidden units $y_j$ and visible units $x_i$, as well as bias weights (offsets) $a_i$ for the visible units and $b_j$ for the hidden units.

The weight $w_{i,j}$ modulates the probability of $x_i$ and $y_j$ having the same value, so that the probability of some activation pattern $(\bar{x}, \bar{y})$ is given by

$$P(\bar{x}, \bar{y}) = \frac{1}{Z} e^{\sum_i a_i x_i + \sum_j b_j y_j + \sum_{i,j} w_{i,j} x_i y_j}. \qquad (1)$$

Inspired on thermodynamics, the exponent is defined as the negative energy of the configuration $E(\bar{x}, \bar{y}) = -\sum_i a_i x_i - \sum_j b_j y_j - \sum_{i,j} w_{i,j} x_i y_j$. $Z$ is the normalization factor (partition function) defined as the sum of $e^{-E(\bar{x},\bar{y})}$ over all possible configurations.



# 1 Inference

We want to calculate the posterior probability of $y_k$ given a data sample $\bar{x}^\mu$, $P(y_k = 1|\bar{x}^\mu)$.

Step 1: Show that if the probability distribution of two variables is separable, $P(x_1, x_2) = A^{-1} f(x_1) g(x_2)$, for some normalization constant $A = \sum_{\{x1,x2\}} f(x_1)g(x_2)$ and arbitrary functions $f$ and $g$, then $x_1$ and $x_2$ are independent.

Step 2: Show that the hidden units are independent given the visible units

$$P(\bar{y}|\bar{x}^\mu) = \prod_j P(y_j|\bar{x}^\mu). \tag{2}$$

Step 3: Show that

$$\frac{P(y_k = 1|\bar{x}^\mu)}{P(y_k = 0|\bar{x}^\mu)} = e^{b_k + \sum_i w_{i,k} x_i^\mu}. \tag{3}$$

Step 4: Noticing that $P(y_k = 1|\bar{x}^\mu) + P(y_k = 0|\bar{x}^\mu) = 1$, show that

$$P(y_k = 1|\bar{x}^\mu) = \sigma(b_k + \sum_i w_{i,k} x_i^\mu), \tag{4}$$

where $\sigma(u) = \frac{1}{1+e^{-u}}$ is the sigmoid function.

What is the value of $P(x_m = 1|\bar{y})$?

# 2 Learning

Show that the gradient ascent on the log-likelihood of the data is given by

$$\Delta w_{i,j} \propto \frac{\partial log P(\bar{x}^\mu, \bar{y}^\mu)}{\partial w_{i,j}} = x_i^\mu y_j^\mu - \sum_{\{\tilde{x},\tilde{y}\}} \tilde{x}_i \tilde{y}_j p(\tilde{x}, \tilde{y}), \tag{5}$$

where the sum is over all possible activations vectors $\{\tilde{x}, \tilde{y}\}$.

Conclude that the learning rule is given by $\Delta w_{i,j} \propto \langle x_i y_j \rangle_{data} - \langle x_i y_j \rangle_{model}$, where this second term is an anti-Hebbian rule for samples generated by the model.