Reproduction of the article: Nonequilibrium Thermodynamics of Restricted Boltzmann Machines

Prof. Carlos Leonardo Viviescas Date: October 18, 2018

1 RBM explained

An RBM is a Markov Random Field defined on a bipartite undirected graph formed by two layers of non-interacting variables. The visible nodes v which consist of m units representing the observable data, and n hidden units h to capture the dependences between the observed variables. The state of the system is represented with $s = (v, h)$ where $v = v_i, i = 1, ..., m., h = h_j, j = 1, ..., n.$. The random variables (v, h) take values 0 or 1 and an energy function is defined for a given configuration $\lambda = (a_i, b_j, W_{ij})$ of the nodes.

$$E\left(s,\lambda\right) = -\sum_i \sum_j W_{ij} h_j v_i - \sum_i a_i x_i - \sum_j b_j h_j$$

$$E\left(\mathbf{x},\mathbf{h}\right) = -\mathbf{h}^T\mathbf{W}\mathbf{x} - \mathbf{c}^T\mathbf{x} - \mathbf{b}^T\mathbf{h}$$

The probability of a state s is given by.

$$p\left(s,\lambda\right) = \frac{e^{-E(s,\lambda)}}{Z}$$

Using Bayes theorem the conditional probability can be written as.

$$p\left(\mathbf{h}|\mathbf{v}\right) = \frac{p\left(\mathbf{v},\mathbf{h}\right)}{\sum_{\mathbf{h'}} p\left(\mathbf{v},\mathbf{h'}\right)}$$

expanding the terms

$$\begin{aligned}
p\left(\mathbf{h}|\mathbf{v}\right) &= \frac{p\left(\mathbf{v},\mathbf{h}\right)}{\sum_{\mathbf{h'}} p\left(\mathbf{v},\mathbf{h'}\right)} \\
&= \frac{e^{\left(\mathbf{h}^T\mathbf{W}\mathbf{v}+\mathbf{a}^T\mathbf{v}+\mathbf{b}^T\mathbf{h}\right)}/Z}{\sum_{h'\epsilon\{0,1\}^H} e^{(h^TWv+a^Tv+b^Th)}/Z} \\
&= \frac{e^{\sum_j (h_jW_j.v+b_jh_j)}}{\sum_{h'_1\epsilon\{0,1\}}\cdots\sum_{h'_H\epsilon\{0,1\}} e^{\sum_j (h'_jW_j.v+b_jh'_j)}} \\
&= \frac{\prod_j e^{(h_jW_j.v+b_jh_j)}}{\sum_{h'_1\epsilon\{0,1\}}\cdots\sum_{h'_H\epsilon\{0,1\}} \prod_j e^{(h'_jW_j.v+b_jh'_j)}} \\
&= \frac{\prod_j e^{(h_jW_j..v+b_jh_j)}}{\left(\sum_{h'_1\epsilon\{0,1\}} e^{(h'_1W_1..v+b_1h'_1)}\right)\cdots\left(\sum_{h'_H\epsilon\{0,1\}} e^{(h'_HW_H..v+b_Hh'_H)}\right)} \\
&= \frac{\prod_j e^{(h_jW_j.v+b_jh_j)}}{\prod_j\left(\sum_{h'_j\epsilon\{0,1\}} e^{(h'_jW_j.v+b_jh'_j)}\right)} \\
&= \prod_j \frac{e^{(h_jW_j..v+b_jh_j)}}{1+e^{(W_j..v+b_j)}} \\
&= \prod_j p\left(h_j|v\right)
\end{aligned}$$

2 First law of thermodynamics in RBMs

From above is immediately that for $h_j = 1$ we have

$$\begin{aligned} p\left(h_j = 1 |\, \mathbf{v}\right) &= \frac{exp\left(W_j \cdot \mathbf{v} + b_j\right)}{1 + exp\left(W_j \cdot \mathbf{v} + b_j\right)} \\ &= \frac{exp\left(W_j \cdot \mathbf{v} + b_j\right)}{1 + exp\left(W_j \cdot \mathbf{x} + b_j\right)} \left(\frac{exp\left(-W_j \cdot \mathbf{v} - b_j\right)}{exp\left(-W_j \cdot \mathbf{v} - b_j\right)}\right) \\ &= \frac{1}{exp\left(-W_j \cdot \mathbf{v} - b_j\right) + 1} \\ &= \frac{1}{1 + exp\left[-\left(W_j \cdot \mathbf{v} + b_j\right)\right]} \\ &= \sigma(W_j \cdot \mathbf{v} + b_j) \end{aligned}$$

where $\sigma\left(x\right)$ is known as the *logistic function*

Due to equation (2) in the paper, and because $\beta \neq 1$

$$p\left(h_j = 1 |\, \mathbf{v}\right) = \sigma(\beta W_j \cdot \mathbf{v} + \beta b_j)$$

Since W_j is a row vector, and x is a column vector, in terms of their components and identifying x as v,

$$p\left(h_j = 1 |\, \mathbf{v}\right) = \sigma(\beta \sum_i W_{ij} v_i + \beta b_j)$$

In an analogous way,

$$p\left(\mathbf{v}|\mathbf{h}\right) = \prod_j p\left(v_j|\mathbf{h}\right)$$

and consequently, for $v_i = 1$ we have

$$p\left(v_i = 1|\mathbf{h}\right) = \sigma\left(\beta a_i + \beta \sum_j h_j W_{ij}\right)$$

Realize that in this case, W_i is a column vector, and h_j is a row vector. In vector form, last expression can be rewritten as

$$p\left(v_i = 1|\mathbf{h}\right) = \sigma\left(\beta\left(a_i + \mathbf{h}^T \cdot W_i\right)\right)$$

For $K = 1$ step, detailed balance reads for a constant λ

$$
\begin{aligned}
\frac{p_\lambda^{(1)}\left(s \to s^{'}\right)}{p_\lambda^{(1)}\left(s^{'} \to s\right)} &= \frac{p_\lambda\left(v'|h\right) p_\lambda\left(h'|v'\right)}{p_\lambda\left(h|v'\right) p_\lambda\left(v|h\right)} \\
&= \frac{\frac{\cancel{p_\lambda(h|v')} p_\lambda(v') p_\lambda(h'|v')}{p_\lambda(h)}}{\cancel{p_\lambda(h|v')} p_\lambda\left(v|h\right)} \\
&= \frac{p_\lambda\left(v'\right) p_\lambda\left(h'|v'\right)}{p_\lambda\left(h\right) p_\lambda\left(v|h\right)} \\
&= \frac{\frac{\cancel{p_\lambda(v')} p_\lambda(v'|h') p_\lambda(h')}{\cancel{p_\lambda(v')}}}{p_\lambda\left(h\right) p_\lambda\left(v|h\right)} \\
&= \frac{p_\lambda\left(v'|h'\right) p_\lambda\left(h'\right)}{p_\lambda\left(h\right) p_\lambda\left(v|h\right)} \\
&= \frac{\frac{p_\lambda(v',h') \cancel{p_\lambda(h')}}{\cancel{p_\lambda(h')}}}{\frac{\cancel{p_\lambda(h)} p_\lambda(v,h)}{\cancel{p_\lambda(h)}}} \\
&= \frac{p_\lambda\left(v', h'\right)}{p_\lambda\left(v, h\right)} \\
&= \frac{p_\lambda\left(s'\right)}{p_\lambda\left(s\right)}
\end{aligned}
$$

For K steps the transition probability may be written in terms of one step transition.

$$
p_\lambda^{(K)}(s \to s') = \sum_{s_1,\ldots,s_{K-1}} \prod_{i=0}^{K-1} p_\lambda^{(1)}(s_i \to s_{i+1})
$$

The detailed balance condition for K steps is given by

$$
\begin{aligned}
\frac{p_\lambda^{(K)}(s \to s')}{p_\lambda^{(K)}(s' \to s)} &= \frac{\sum_{s_1,\ldots,s_{K-1}} \prod_{i=0}^{K-1} p_\lambda^{(1)}(s_i \to s_{i+1})}{\sum_{s_1,\ldots,s_{K-1}} \prod_{i=0}^{K-1} p_\lambda^{(1)}(s_{i+1} \to s_i)} \\
&= \frac{\sum_{s_1,\ldots,s_{K-1}} p_\lambda(s_0 \to s_1) p_\lambda(s_1 \to s_2) \ldots.. p_\lambda(s_{K-1} \to s_K)}{\sum_{s_1,\ldots,s_{K-1}} p_\lambda(s_1 \to s_0) p_\lambda(s_2 \to s_1) \ldots.. p_\lambda(s_K \to s_{K-1})} \\
&= \frac{\sum_{s_1,\ldots,s_{K-1}} p_\lambda(s_1) p_\lambda(s_2) \ldots.. p_\lambda(s_K)}{\sum_{s_1,\ldots,s_{K-1}} p_\lambda(s_0) p_\lambda(s_1) \ldots.. p_\lambda(s_{K-1})} \\
&= \frac{\sum_{s_1,\ldots,s_{K-1}} p_\lambda(s_K)}{\sum_{s_1,\ldots,s_{K-1}} p_\lambda(s_0)} \\
&= \frac{p_\lambda(s_K)}{p_\lambda(s_0)}
\end{aligned}
$$

3 Fluctuations theorems and the second law

3.1 Crooks fluctuation theorem and the second law

Here the article focus on the fluctuations theorems, known as the Crooks Fluctuations Theorems, the first part is to obtain the Crooks theorem from the conditions previously establish for the probabilities distribution. First its considered that the initial and final distributions are the equilibrium distribution. And a trajectory γ' which is a collection of states that defined a path (i.e. $\gamma' = (s_k, ..., s_0)$. One can formulate a detailed balance condition for the probability of the trajectory γ:

$$\frac{P(\gamma)}{P(\gamma')} = \frac{P(s_0, ..., s_K)}{P(s_K, ..., s_0)} \tag{1}$$

Where the trajectory is chopped into little pieces as follows :

$$P(s_0, ..., s_K) = p_{\lambda_1}(s_0 \to s_1)p_{\lambda_2}(s_1 \to s_2).....p_{\lambda_{K-1}}(s_{K-1} \to s_K) \tag{2}$$

Where for any $i \in 0, ..., K$

$$p(s_i \to s_{i+1}) = p_{\lambda_i}(s_{i+1}|s_i) \tag{3}$$

For the initial state and the last one, there is not conditional probability but instead the initial and final distribution in equilibrium $p_{eq}(s_0)$ and $p_{eq}(s_K)$.Putting the preceding considerations in 1 it can be written:

$$\frac{P(s_0, ..., s_k)}{P(s_k, ..., s_0)} = \prod_{i=0}^{K-1} \frac{p_{eq}(s_0)p_{\lambda_{i+1}}(s_{i+1}|s_i)}{p_{eq}(s_K)p_{\lambda_{i+1}}(s_i|s_{i+1})}$$

Where is also used the fact that the process is Markovian so the state i only depends on the state $i-1$. Now using the detailed balance condition in which we take $s' = s_{i+1}$ and $s = s_i$ and using the respective configuration of weights for each state. We get :

$$\frac{P(s_0, ..., s_k)}{P(s_k, ..., s_0)} = \prod_{i=0}^{K-1} \frac{p_{\lambda_i}(s_i)}{p_{\lambda_{i+1}}(s_i)} \frac{p_{\lambda_K}(s_K)}{p_{\lambda_0}(s_0)} \frac{p_{eq}(s_0)}{p_{eq}(s_K)} \tag{4}$$

Arranging the terms related with s_0 and s_K we get the equation (15) of the article. from this, the only thing that remains is to plug in the original definition for the probability

$$p_\lambda = \frac{1}{Z(\beta, \lambda)} e^{-\beta E(s,\lambda)}$$

In the last result, which is made next:

$$
\begin{aligned}
\frac{P(s_0,...,s_k)}{P(s_k,...,s_0)} &= \frac{p_{eq}(s_0)p_{\lambda_K}(s_K)}{p_{\lambda_0}(s_0)p_{\lambda_{eq}}(s_K)}\prod_{i=0}^{K-1}\frac{p_{\lambda_i}(s_i)}{p_{\lambda_{i+1}}(s_i)} \\
&= \frac{p_{eq}(s_0)p_{\lambda_K}(s_K)}{p_{\lambda_0}(s_0)p_{\lambda_{eq}}(s_K)}\prod_{i=0}^{K-1}\frac{e^{-\beta E(s_i,\lambda_i)}/Z(\beta,\lambda_i)}{e^{-\beta E(s_i,\lambda_{i+1})}/Z(\beta,\lambda_{i+1})} \\
&= \frac{p_{eq}(s_0)p_{\lambda_K}(s_K)}{p_{\lambda_0}(s_0)p_{\lambda_{eq}}(s_K)}\frac{e^{-\beta E(s_0,\lambda_0)}/Z(\beta,\lambda_0)e^{-\beta E(s_1,\lambda_1})/Z(\beta,\lambda_1).....e^{-\beta E(s_{K-1},\lambda_{K-1})}/Z(\beta,\lambda_{K-1})}{e^{-\beta E(s_0,\lambda_1)}/Z(\beta,\lambda_1)e^{-\beta E(s_1,\lambda_2})/Z(\beta,\lambda_2).....e^{-\beta E(s_{K-1},\lambda_K)}/Z(\beta,\lambda_K)} \\
&= \frac{p_{eq}(s_0)p_{\lambda_K}(s_K)}{p_{\lambda_0}(s_0)p_{\lambda_{eq}}(s_K)}e^{\beta(\sum_{i=0}^{K-1}(E(s_i,\lambda_{i+1})-E(s_i,\lambda_i)))}\frac{Z(\beta,\lambda_K)}{Z(\beta,\lambda_0)} \\
&= \frac{p_{eq}(s_0)p_{\lambda_K}(s_K)}{p_{\lambda_0}(s_0)p_{\lambda_{eq}}(s_K)}e^{\beta(\sum_{i=0}^{K-1}(E(s_i,\lambda_{i+1})-E(s_i,\lambda_i)))}e^{\beta(\beta^{-1}Ln(Z(\beta,\lambda_K)/Z(\beta,\lambda_0)))} \\
&= \frac{p_{eq}(s_0)p_{\lambda_K}(s_K)}{p_{\lambda_0}(s_0)p_{\lambda_{eq}}(s_K)}e^{\beta(\sum_{i=0}^{K-1}(E(s_i,\lambda_{i+1})-E(s_i,\lambda_i))+(\beta^{-1}Ln(Z(\beta,\lambda_K)/Z(\beta,\lambda_0)))} \\
&= \frac{p_{eq}(s_0)p_{\lambda_K}(s_K)}{p_{\lambda_0}(s_0)p_{\lambda_{eq}}(s_K)}e^{\beta(W-\Delta F)}
\end{aligned}
$$

Considering the initial and final distribution to be in equilibrium and summing over all possible trajectories with the same work we obtain.

$$\frac{P_{s_0\to s_K(W)}}{P_{s_K\to s_0(-W)}} = e^{\beta(W-\Delta F)}$$

which is the Crooks fluctuation theorem.

Using the definition of the Shannon Entropy

$$S(\beta,\lambda) = -\sum_s p_\lambda(s)log(p_\lambda(s))$$

we can derivate the change in the entropy between the states s_0 and s_K, first we calculate the entropy in both states :

$$S(\beta,\lambda_k) = -\sum_s \frac{1}{Z(\beta,\lambda)}e^{-\beta E(s,\lambda)}log\left(\frac{1}{Z(\beta,\lambda)}e^{-\beta E(s,\lambda)}\right) \tag{5}$$

$$S(\beta,\lambda_k) = -\left(\frac{1}{Z(\beta,\lambda_k)}log\left(\frac{1}{Z(\beta,\lambda_k)}\right)\sum_s e^{-\beta E(s,\lambda_k)} + \sum_s \beta E(s,\lambda_k)\right) \tag{6}$$

$$S(\beta,\lambda_k) = -\left(\frac{1}{Z(\beta,\lambda_k)}log\left(\frac{1}{Z(\beta,\lambda_k)}\right)Z(\beta,\lambda_k) + \beta\langle E(\lambda_k)\rangle\right) \tag{7}$$

$$S(\beta,\lambda_k) = log(Z(\beta,\lambda_k)) + \beta\langle E(\lambda_k)\rangle \tag{8}$$

Doing exactly the same for λ_0 and taking the difference between the two values of entropy we get :

$$S(\beta,\lambda_k) - S(\beta,\lambda_0) = log(Z(\beta,\lambda_k)) + \beta\langle E(\lambda_k)\rangle - log(Z(\beta,\lambda_0)) - \beta\langle E(\lambda_0)\rangle \tag{9}$$

$$= \beta\langle\Delta E\rangle + log\left(\frac{log(Z(\beta,\lambda_k)}{log(Z(\beta,\lambda_0)}\right) \tag{10}$$

Now using the first of the thermodynamic the previous result can be written as :

$$\Delta S = \beta\langle\Delta Q\rangle + \beta\langle\Delta W\rangle - \beta\Delta F \tag{11}$$

Remembering that the change in the entropy is bigger than the change of the heat for the system ($\Delta S \geq \beta\Delta Q$) at constant temperature it can be establish that :

$$\beta\langle\Delta Q\rangle + \beta\langle\Delta W\rangle - \beta\Delta F \geq \Delta Q \tag{12}$$

$$\beta\langle\Delta W\rangle - \beta\Delta F \geq 0 \tag{13}$$

Which is a common statement for the second law of the thermodynamics

3.2 Heat exchange fluctuation theorem

In a RBM prepared in thermal equilibrium with a temperature T_1 and then placed in contact with a reservoir with temperature T_2 with a configuration Λ there will be a non-equilibrium fluctuation for the heat Q.

The probability $P(\Delta E)$ of finding the energy variation ΔE after K steps in the dynamics is given in terms of the joint probability of states.

$$\begin{aligned}
P^K(\Delta E) &= \sum_{ss'} p_2^K(s \to s')p_1(s)\delta(E' - E - \Delta E) \\
&= \sum_{ss'} p_2^K(s \to s')\frac{e^{-\beta_1 E(e,\lambda)}}{Z(\beta_1,\lambda)}e^{-\beta_1 E(s',\lambda)}e^{\beta_1 E(s',\lambda)}\delta(E' - E - \Delta E) \\
&= e^{\beta_1\Delta E}\sum_{ss'} p_2^K(s \to s')p_1(s')\delta(E' - E - \Delta E) \\
&= e^{\beta_1\Delta E}\sum_{s's} p_2^K(s' \to s)p_1(s)\delta(E' - E + \Delta E)
\end{aligned}$$

Using the detailed balance condition for the term $p_2^K(s' \to s)$ we obtain

$$\begin{aligned}
P^K(\Delta E) &= e^{\beta_1\Delta E}\sum_{s's} p_2^K(s \to s')p_1(s)\frac{p_2(s')}{p_2(s)}\delta(E' - E + \Delta E) \\
&= e^{(\beta_1-\beta_2)\Delta E}\sum_{s's} p_2^K(s \to s')p_1(s)\delta(E' - E + \Delta E) \\
&= e^{(\beta_1-\beta_2)\Delta E}P^K(-\Delta E)
\end{aligned}$$

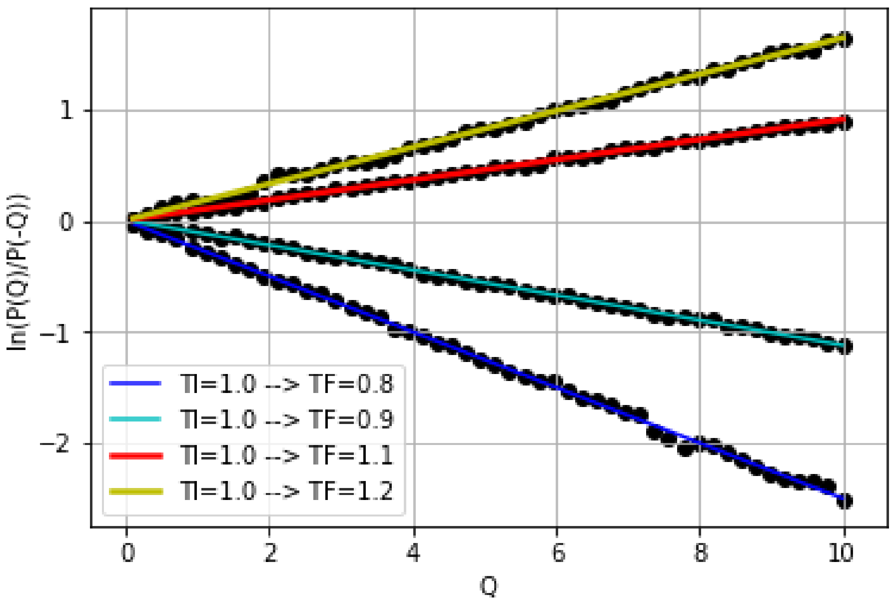
In the absence of work the change of energy is totally due to the heat Q. Finally we get

$$\frac{P^K(Q)}{P^K(-Q)} = e^{(\beta_1-\beta_2)Q}$$

3.3 Numerical verification of XFT identity

In order to verify the XFT identity a RBM with 784 visible units and 500 hidden units was trained with the MNIST dataset and setting $\beta = 1$, then an ensemble of $2 \cdot 10^6$ RBMs with the same λ was put in equilibrium with the temperature $T_1 = 1$, this was done by setting a random initial state for the visible and hidden units and performing 300 Gibbs steps such that a state

s was obtained. The state s which is in equilibrium at temperature $T_1 = 1$ and has energy $E(s) = E$ was taken as the initial state to perform one Gibbs step at temperature T_2 such that the state s' with energy $E' = E(s')$ is obtained, then the energy variation $\Delta E = E' - E$ is stored. The $2 \cdot 10^6$ values of ΔE are used to estimate $P(\Delta E)$ and $P(-\Delta E)$ and finally the $\log(P(\Delta E)/P(-\Delta E))$ is plotted and compare with the predictions from the XFT theorem.



4 Unsupervised learning as a thermodynamic process

4.1 Contrastive Divergence

Likelihood function: Given a sample an a parametric family of distributions that could have generate the sample, the likelihood function is a function that associates to each parameter the probability of observing the given sample.

The likelihood function is not a probability density function. Let us suppose two parameters θ_1 and θ_2 such that $\mathcal{L}(\theta_1|x) > \mathcal{L}(\theta_2|x)$. Then, the sample we actually observed is more likely to have occurred if $\theta = \theta_1$ than if $\theta = \theta_2$, this means θ_1 is more plausible value for θ than θ_2.

Given $\mathcal{L}\left(\lambda = \{a_i, b_j, w_{ij}\} = \theta, D = \{v_i\}_{i=1}^N\right)$, let us calculate the variation of the log likelihood function as follows,

$$
\begin{aligned}
\frac{\partial}{\partial\theta}\left[\mathcal{L}\left(\lambda,D\right)\right] &= \frac{\partial}{\partial\theta}\sum_{i=1}^{N} log\ p\left(v_i\right) \\
&= \frac{\partial}{\partial\theta} log\ p\left(\boldsymbol{v}\right) \\
&= \frac{\partial}{\partial\theta}\sum_{\boldsymbol{h}} log\ p\left(\boldsymbol{v},\boldsymbol{h}\right) \\
&= \frac{\partial}{\partial\theta}\left[log\left(\sum_{\boldsymbol{h}}\frac{e^{-\beta E(\boldsymbol{v},\boldsymbol{h})}}{Z}\right)\right] \\
&= \frac{1}{\sum_{\boldsymbol{h}}\frac{e^{-\beta E(\boldsymbol{v},\boldsymbol{h})}}{Z}}\left[\frac{1}{Z}\sum_{\boldsymbol{h}}\frac{\partial}{\partial}e^{-\beta E(\boldsymbol{v},\boldsymbol{h})}+\left(-\frac{1}{Z^2}\right)\sum_{\boldsymbol{h}}e^{-\beta E(\boldsymbol{v},\boldsymbol{h})}\right] \\
&= \frac{Z}{\sum_{\boldsymbol{h}}e^{-\beta E(\boldsymbol{v},\boldsymbol{h})}}\left[-\frac{\beta}{Z}\sum_{\boldsymbol{h}}e^{-\beta E(\boldsymbol{v},\boldsymbol{h})}\frac{\partial}{\partial}E\left(\boldsymbol{v},\boldsymbol{h}\right)-\frac{1}{Z^2}\sum_{\boldsymbol{h}}e^{-\beta E(\boldsymbol{v},\boldsymbol{h})}\frac{\partial}{\partial\theta}Z\right] \\
&= -\sum_{\boldsymbol{h}}\left[\frac{Z}{\sum_{\boldsymbol{h}}e^{-\beta E(\boldsymbol{v},\boldsymbol{h})}}\frac{\beta}{Z}e^{-\beta E(\boldsymbol{v},\boldsymbol{h})}\frac{\partial}{\partial}E\left(\boldsymbol{v},\boldsymbol{h}\right)\right]-\frac{Z}{\sum_{\boldsymbol{h}}e^{-\beta E(\boldsymbol{v},\boldsymbol{h})}}\frac{1}{Z^2}\sum_{\boldsymbol{h}}e^{-\beta E(\boldsymbol{v},\boldsymbol{h})}\frac{\partial}{\partial\theta}Z \\
&= -\beta\sum_{\boldsymbol{h}}\left[\sum_{\boldsymbol{v}}p\left(\boldsymbol{v}\right)p\left(\boldsymbol{h}|\boldsymbol{v}\right)\frac{\partial}{\partial}E\left(\boldsymbol{v},\boldsymbol{h}\right)\right]-\frac{1}{Z}\frac{\partial}{\partial\theta}Z \\
&= -\beta\sum_{\boldsymbol{v},\boldsymbol{h}}p\left(\boldsymbol{v}\right)p\left(\boldsymbol{h}|\boldsymbol{v}\right)\frac{\partial}{\partial}E\left(\boldsymbol{v},\boldsymbol{h}\right)-\frac{1}{Z}\frac{\partial}{\partial\theta}\left[\sum_{\boldsymbol{v},\boldsymbol{h}}e^{-\beta E(\boldsymbol{v},\boldsymbol{h})}\right] \\
&= -\beta\sum_{\boldsymbol{v},\boldsymbol{h}}p_D\left(\boldsymbol{v}\right)p_\lambda\left(\boldsymbol{h}|\boldsymbol{v}\right)\frac{\partial}{\partial}E\left(\boldsymbol{v},\boldsymbol{h}\right)+\beta\left[\sum_{\boldsymbol{v},\boldsymbol{h}}\frac{e^{-\beta E(\boldsymbol{v},\boldsymbol{h})}}{Z}\frac{\partial}{\partial\theta}E\left(\boldsymbol{v},\boldsymbol{h}\right)\right] \\
&= -\beta < \frac{\partial}{\partial\theta}E\left(\boldsymbol{v},\boldsymbol{h}\right) >_D +\beta\sum_{\boldsymbol{v},\boldsymbol{h}}p_\lambda\left(\boldsymbol{v},\boldsymbol{h}\right)\frac{\partial}{\partial\theta}E\left(\boldsymbol{v},\boldsymbol{h}\right) \\
&= -\beta\left(< \frac{\partial}{\partial\theta}E\left(\boldsymbol{v},\boldsymbol{h}\right) >_D - < \frac{\partial}{\partial\theta}E\left(\boldsymbol{v},\boldsymbol{h}\right) >_\lambda\right)
\end{aligned}
$$

The learning rules are given by

$$
\theta_{\tau+1} = \theta_\tau + \eta\cdot\left[\beta\left(< \frac{\partial}{\partial\theta}E\left(\boldsymbol{v},\boldsymbol{h}\right) >_D - < \frac{\partial}{\partial\theta}E\left(\boldsymbol{v},\boldsymbol{h}\right) >_\lambda\right)\right]
$$

such that

$$
\Rightarrow \Delta\theta = \theta_{\tau+1}-\theta_\tau = -\eta\beta\left(< \frac{\partial}{\partial\theta}E\left(\boldsymbol{v},\boldsymbol{h}\right) >_D - < \frac{\partial}{\partial\theta}E\left(\boldsymbol{v},\boldsymbol{h}\right) >_\lambda\right)
$$

Due to $\theta = \{a_i, b_j, w_{ij}\}$ and that energy is given by

$$
E\left(\boldsymbol{v},\boldsymbol{h}\right) = -\sum_i a_i v_i - \sum_j b_j h_j - \sum_{i,j} v_i h_j w_{i,j}
$$

,

it follows that

$$\begin{aligned}\Delta a_i &= -\eta\beta\left(< \frac{\partial}{\partial a_i}E\left(\boldsymbol{v},\boldsymbol{h}\right) >_D - < \frac{\partial}{\partial a_i}E\left(\boldsymbol{v},\boldsymbol{h}\right) >_{\lambda}\right) \\ &= -\eta\beta\left(< -v_i >_D - < -v_i >_{\lambda}\right) \\ &= \eta\beta\left(< v_i >_D - < v_i >_{\lambda}\right)\end{aligned}$$

,

analogously for b_j and $w_{i,j}$,

$$\begin{aligned}\Delta b_i &= -\eta\beta\left(< \frac{\partial}{\partial b_i}E\left(\boldsymbol{v},\boldsymbol{h}\right) >_D - < \frac{\partial}{\partial b_i}E\left(\boldsymbol{v},\boldsymbol{h}\right) >_{\lambda}\right) \\ &= -\eta\beta\left(< -h_i >_D - < -h_i >_{\lambda}\right) \\ &= \eta\beta\left(< h_i >_D - < h_i >_{\lambda}\right)\end{aligned}$$

,

$$\begin{aligned}\Delta w_{i,j} &= -\eta\beta\left(< \frac{\partial}{\partial w_{i,j}}E\left(\boldsymbol{v},\boldsymbol{h}\right) >_D - < \frac{\partial}{\partial w_{i,j}}E\left(\boldsymbol{v},\boldsymbol{h}\right) >_{\lambda}\right) \\ &= -\eta\beta\left(< -v_ih_j >_D - < -v_ih_j >_{\lambda}\right) \\ &= \eta\beta\left(< v_ih_j >_D - < v_ih_j >_{\lambda}\right)\end{aligned}$$

,

The second part of the change in the parameters $< \frac{\partial}{\partial\theta}E\left(\boldsymbol{v},\boldsymbol{h}\right) >_{\lambda}$ cannot be calculated explicitly since it requires knowing the value of partition function for the RBM, hence, a Markov Chain Monte Carlo (MCMC) method can be used to sample from the equilibrium distribution and use this samples to estimate that average. In practice a MCMC method would require too much computational time for big architectures, so the Gibbs sampling is performed for a finite number of steps k and the average is taken from the samples obtained after this process.

The Gibbs sampling dynamics consist of sampling independently the visible and the hidden variables which can be done since those are conditionally independent. Given a vector v_0 of the visible layer a Gibbs step consist of sampling a vector h_1 from the hidden layer according to the probability $p(h|v_0)$ and then sampling a vector v_1 from the visible layer according to the probability $p(v|h_1)$, the joint state $\{v_1, h_1\}$ is the result of one Gibbs step. After performing n Gibbs steps the state $\{v_n, h_n\}$ is obtained and if $n \to \infty$ then $\{v_\infty, h_\infty\}$ would be distributed according to $p(v_\infty, h_\infty)$.

4.2 Stochastic thermodynamics of CD

Convergence contrastive for n Gibbs samples through the Kullback-Leibler divergence is given by

$$
\begin{aligned}
CD_n &= \sum_s p_D log \frac{p_D}{p_\lambda} - \sum_s p_n log \frac{p_n}{p_\lambda} \\
&= \sum_s p_D \left[log\ P_D - log \left(\frac{e^{-\beta E(s,\lambda)}}{Z} \right) \right] - \sum_s p_n \left[log\ P_n - log \left(\frac{e^{-\beta E(s,\lambda)}}{Z} \right) \right] \\
&= \sum_s p_D \left[log\ P_D - \left(-\beta E(s,\lambda) - log\ Z \right) \right] - \sum_s p_n \left[log\ P_n - \left(-\beta E(s,\lambda) - log\ Z \right) \right] \\
&= -\beta \sum_s p_n E(S,\lambda) + \beta \sum_s p_D E(s,\lambda) - \sum_s p_n log\ p_n + \sum_s p_D log\ p_D + \cancelto{0}{\sum_s log\ Z \left(p_D - p_n \right)}
\end{aligned}
$$

From the averages of the function $f(\mathbf{v}, \mathbf{h})$ and the definition of Shannon entropy,

$$
\begin{aligned}
-\beta \sum_s p_n(s) E(s,\lambda) &\to -\beta < E(s,\lambda) >_n \\
\beta \sum_s p_D(s) E(s,\lambda) &\to \beta < E(s,\lambda) >_D \\
-\sum_s p_n(s)\, log\ p_n(s) &\to S(\beta, n) = S_n \\
\sum_s p_D(s)\, log\ p_D(s) &\to -S(\beta, D) = S_0
\end{aligned}
$$

such that

$$CD_n = -\beta \left(< E(s,\lambda) >_n - < E(s,\lambda) >_D \right) + S_n + S_0$$

Now, when $n \to \infty$,

$$\Delta S = \beta < Q > + \beta < W > - \beta \Delta F$$

and finally,

$$
\begin{aligned}
CD_{n \to \infty} &= \beta < Q > + \beta < W > - \beta \Delta F - \beta < Q > \\
&= \beta < W > - \beta \Delta F
\end{aligned}
$$

5 Application in estimation of the partition function

Annealed Importance Sampling **AIS** is a Monte Carlo algorithm based on sampling from a sequence of distributions which interpolate between a tractable initial distribution and the intractable target distribution. It returns a set of weighted samples, and in the limit of infinitely many intermediate distributions, the variance of the weights approaches zero. The most common use is in estimating partition functions.

The construction of AIS is as follows:

1. Let $p_\lambda(x)$ be our target distribution

2. Let $p_{\lambda_0}(x)$ be our proposal distribution. The only requirement for $p_{\lambda_0}(x)$ is that we can sample independent point from it. It does not matter is $p_{\lambda_0}(x)$ is closed to $p_\lambda(x)$

3. Let $p_{\lambda_k}(x)$ a sequence of intermediate distributions from $p_{\lambda_0}(x)$ to $p_{\lambda}(x)$. The requirement over $p_{\lambda_k}(x)$ is that $p_{\lambda_k}(x) \neq 0$ for $p_{\lambda_{k-1}}(x) \neq 0$.

4. Define a local transition probabilities $T_j\left(x, x^{'}\right)$

5. Then, we need:
   - Sample an independent point from $x_{n-1} \sim p_{\lambda_k}(x)$
   - Sample x_{n-2} from x_{n-1} from Markov Chain Monte Carlo with rate transition T_{n-1}
   - Sample as before until x_0 from x_1 from Markov Chain Monte Carlo with rate transition T_1

In the original AIS formulation for RBMs, it is defined

$$\begin{aligned} p^*_{\lambda}(v) &= \sum_h p_{\lambda}(s = (\boldsymbol{v}, \boldsymbol{h}))\, Z(\beta, \lambda) \\ &= p_{\lambda}(v)\, Z(\beta, \lambda) \end{aligned}$$

From above definition follows

$$\begin{aligned} \sum_v p^*_{\lambda}(v) &= \sum_v \sum_h p_{\lambda}(s = (\boldsymbol{v}, \boldsymbol{h}))\, Z(\beta, \lambda) \\ &= Z(\beta, \lambda) \sum_{v,h} p_{\lambda}(v)\, Z(\beta, \lambda) \\ &= Z(\beta, \lambda), \end{aligned}$$

besides, due to $p_{\lambda}(v)$ definition, we have

$$p_{\lambda_0}(v) = \frac{p^*_{\lambda_0}(v)}{Z(\beta, \lambda_0)}$$

$$\Rightarrow Z(\beta, \lambda_0) = \frac{p^*_{\lambda_0}(v)}{p_{\lambda_0}(v)}$$

which leads us to

$$\begin{aligned} \frac{Z(\beta, \lambda)}{Z(\beta, \lambda_0)} &= \frac{\sum_v p^*_{\lambda}(v)}{\frac{p^*_{\lambda_0}(v)}{p_{\lambda_0}(v)}} \\ &= \sum_v \frac{p^*_{\lambda}(v)}{p^*_{\lambda_0}(v)} p_{\lambda_0}(v) \\ &= < \frac{p^*_{\lambda}(v)}{p^*_{\lambda_0}(v)} >_{p_{\lambda_0}} \end{aligned}$$

On the other hand, by definition $Z(\beta, \lambda) = \sum_s e^{-\beta E(s,\lambda)}$ and $p_{\lambda_0}(v) = \frac{e^{-\beta E(s,\lambda_0)}}{Z(\beta,\lambda_0)}$, thus

$$\begin{aligned} \frac{Z(\beta, \lambda)}{Z(\beta, \lambda_0)} &= \sum_s e^{-\beta E(s,\lambda)} p_{\lambda_0}(s)\, e^{\beta E(s,\lambda_0)} \\ &= \sum_s p_{\lambda_0}(s)\, e^{-\beta[E(s,\lambda) - E(s,\lambda_0)]} \end{aligned}$$

In order to reach Jarzynski equality,

$$\begin{aligned}
\prod_{k=0}^{K-1} \frac{Z(\beta, \lambda_{k+1})}{Z(\beta, \lambda_k)} &= \frac{\cancel{Z(\beta, \lambda_1)}}{Z(\beta, \lambda_0)} \frac{\cancel{Z(\beta, \lambda_2)}}{\cancel{Z(\beta, \lambda_1)}} \times \cdots \times \frac{\cancel{Z(\beta, \lambda_{K-1})}}{\cancel{Z(\beta, \lambda_{K-2})}} \frac{Z(\beta, \lambda_K)}{\cancel{Z(\beta, \lambda_{K-1})}} \\
&= \frac{Z(\beta, \lambda_K)}{Z(\beta, \lambda_0)} \\
&= \sum_s p_{\lambda_0}(s) \, e^{-\beta[E(s,\lambda_K) - E(s,\lambda_0)]} \\
&= \sum_s p_{\lambda_0}(s) \, e^{-\beta W} \\
&= < e^{-\beta W} >_{p_{\lambda_0}},
\end{aligned}$$

but we have that $-log\left[\frac{Z(\beta,\lambda_K)}{Z(\beta,\lambda_0)}\right] = \beta \Delta F$, then, $\frac{Z(\beta,\lambda_K)}{Z(\beta,\lambda_0)} = e^{-\beta \Delta F}$, finally, getting Jarzynski equality:

$$< e^{-\beta W} >= e^{-\beta \Delta F}$$

References

[1] Domingos S. P. Salazar *Nonequilibrium thermodynamics of Restricted Boltzmann Machines*, DOI 10.1103/PhysRevE.96.022131