A Beginner's Tutorial of Restricted Boltzmann Machines

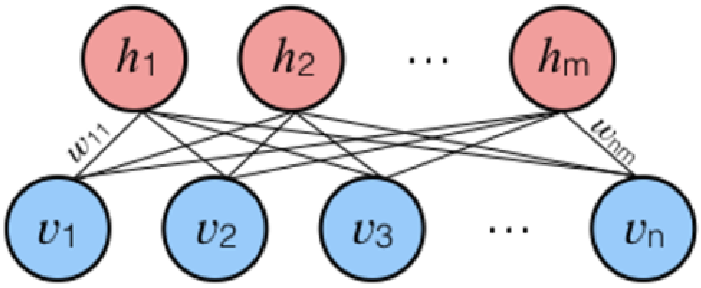
Yiping Cheng
School of Electronic and Information Engineering
Beijing Jiaotong University, Beijing 100044, China
ypcheng@bjtu.edu.cn

Abstract

Restricted Boltzmann machines (RBMs) are the building blocks of some deep learning networks. However, despite their importance, it is our perception that some very important derivations about the RBM are missing in the literature, and a beginner may feel RBM very hard to understand. We provide here these missing derivations. We cover the classic Bernoulli-Bernoulli RBM and the Gaussian-Bernoulli RBM, but leave out the "continuous" RBM as it is believed not as mature as the former two. This tutorial can be used as a companion or complement to the famous RBM paper "Training restricted Boltzmann machines: An introduction" by Fisher and Igel.

1 Introduction

A restricted Boltzmann machine (RBM) is actually a parameterized model of probability distributions. Given some observations, i.e. the training data, the parameters of the model can be learned, which is actually done by using an optimization algorithm to maximize the likelihood function, where the optimization algorithm is usually *stochastic gradient ascent*. An RBM consists of two types of units: the visible units which can be observed, and the hidden units which cannot be observed. Connections only exist between the visible units and the hidden units, but not between different visible units, nor between different hidden units. So the visible units can be viewed as constituting the first layer, and the hidden units as constituting the second layer. Each visible unit corresponds to a component of an observation (e.g., one visible unit for each pixel of an image). For detailed descriptions of RBM see [1, 2]. The architecture of an RBM is depicted as follows.



There are two well-received RBM types: Bernoulli-Bernoulli RBM and Gaussian-Bernoulli RBM, which we will discuss in this paper. The biggest difference between the two is that the visible units are binary in the former but continuous in the latter. The hidden units are binary in both types of RBM. In the literature there are formulations of RBM that the hidden units are also continuous, which are called *continuous RBMs*. We will not cover this type of RBM here as it is believed by the present author that the existing continuous RBM formulations are not as mature as the previous two.

As a researcher with decades of experience but a newcomer to the deep learning field, I feel RBM quite hard to understand, at least much harder than the convolutional neural network. And I now know the reason. It is

doi:10.20944/preprints202003.0337.v1

because all papers and tutorials of RBM have not provided detailed derivations of some very important facts about the RBM. These authors may feel the facts trivial or obvious, but actually they are not, at least as felt by the present author. Therefore, our purpose here is to provide the missing derivations neatly in one paper so that beginners of this field can benefit from it.

This paper should be used as a companion for other RBM tutorials, e.g. [1, 2]. Concepts that are explained very well in the existing tutorials will not be explained here, as we only provide missing derivations.

Notation: Throughout this paper, Bernoulli-Bernoulli RBM will be abbreviated as BB-RBM, and Gaussian-Bernoulli RBM will be abbreviated as GB-RBM. We denote the number of visible units by I and the number of hidden units by J. The vector of visible units is denoted by $\boldsymbol{v}$ and the vector of hidden units by $\boldsymbol{h}$. We define the sigmoid function $\sigma(\cdot)$ by

$$\sigma(x) = \frac{e^x}{1+e^x} \tag{1.1}$$

which we will use quite often later.

2 Derivations for BB-RBM

In this section, we consider BB-RBMs, which are RBMs with binary visible and hidden units, i.e., $\boldsymbol{v} \in \{0,1\}^I$ and $\boldsymbol{h} \in \{0,1\}^J$. It is important to note that RBM is an energy-based model of probability distributions. In this spirit, the BB-RBM with parameters $\boldsymbol{\theta}$ represents a probability distribution

$$p(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\theta}) = \frac{e^{E(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\theta})}}{Z(\boldsymbol{\theta})} \tag{2.1}$$

where

$$Z(\boldsymbol{\theta}) = \sum_{\boldsymbol{u},\boldsymbol{g}} e^{E(\boldsymbol{u},\boldsymbol{g};\boldsymbol{\theta})} \tag{2.2}$$

is the normalizing constant which is called the partition function and $E(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\theta})$ is the (actually negative) energy function, which is defined as

$$E(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\theta}) = \sum_{i=1}^{I}\sum_{j=1}^{J} v_i w_{ij} h_j + \sum_{i=1}^{I} b_i v_i + \sum_{j=1}^{J} c_j h_j \tag{2.3}$$

where $\boldsymbol{\theta} = \boldsymbol{w} \in \mathbb{R}^{I\times J}, \boldsymbol{b} \in \mathbb{R}^I, \boldsymbol{c} \in \mathbb{R}^J$ is the set of model parameters. Please be warned that the E in this paper is actually $-E$ (negative of energy) in other papers. I believe defining E this way helps to reduce some clutter in later derivations.

2.1 Definition of $p(\boldsymbol{v};\boldsymbol{\theta})$

We already have

$$p(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\theta}) = \frac{e^{E(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\theta})}}{Z(\boldsymbol{\theta})} = \frac{e^{E(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\theta})}}{\sum_{\boldsymbol{u},\boldsymbol{g}} e^{E(\boldsymbol{u},\boldsymbol{g};\boldsymbol{\theta})}}.$$

And the probability mass function of the visible variables is thus

$$p(\boldsymbol{v};\boldsymbol{\theta}) = \sum_{\boldsymbol{h}} p(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\theta}) = \frac{\sum_{\boldsymbol{h}} e^{E(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\theta})}}{\sum_{\boldsymbol{u},\boldsymbol{g}} e^{E(\boldsymbol{u},\boldsymbol{g};\boldsymbol{\theta})}}. \tag{2.4}$$

Then

$$\frac{p(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\theta})}{p(\boldsymbol{v};\boldsymbol{\theta})} = \frac{e^{E(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\theta})}}{\sum_{\boldsymbol{g}} e^{E(\boldsymbol{v},\boldsymbol{g};\boldsymbol{\theta})}}. \tag{2.5}$$

2.2 General formula for the gradient of the log-likelihood

Learning the RBM parameters $\boldsymbol{\theta}$ is through the maximization of the log-likelihood over the training samples. The log-likelihood, for a single training sample $\boldsymbol{v}$, is given by

$$L(\boldsymbol{\theta};\boldsymbol{v}) = \log p(\boldsymbol{v};\boldsymbol{\theta}) = \log \sum_{\boldsymbol{h}} p(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\theta}) = \log \sum_{\boldsymbol{h}} e^{E(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\theta})} - \log \sum_{\boldsymbol{u},\boldsymbol{g}} e^{E(\boldsymbol{u},\boldsymbol{g};\boldsymbol{\theta})}. \tag{2.6}$$

doi:10.20944/preprints202003.0337.v1

$$\begin{aligned}
\frac{\partial L(\boldsymbol{\theta};\boldsymbol{v})}{\partial\theta} &= \frac{\partial \log \sum_{\boldsymbol{h}} e^{E(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\theta})}}{\partial\theta} - \frac{\partial \log \sum_{\boldsymbol{u},\boldsymbol{g}} e^{E(\boldsymbol{u},\boldsymbol{g};\boldsymbol{\theta})}}{\partial\theta} \\
&= \frac{\sum_{\boldsymbol{h}} e^{E(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\theta})} \frac{\partial E(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\theta})}{\partial\theta}}{\sum_{\boldsymbol{g}} e^{E(\boldsymbol{v},\boldsymbol{g};\boldsymbol{\theta})}} - \frac{\sum_{\boldsymbol{u},\boldsymbol{g}} e^{E(\boldsymbol{u},\boldsymbol{g};\boldsymbol{\theta})} \frac{\partial E(\boldsymbol{u},\boldsymbol{g};\boldsymbol{\theta})}{\partial\theta}}{Z(\boldsymbol{\theta})} \\
&= \sum_{\boldsymbol{h}} \frac{e^{E(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\theta})}}{\sum_{\boldsymbol{g}} e^{E(\boldsymbol{v},\boldsymbol{g};\boldsymbol{\theta})}} \frac{\partial E(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\theta})}{\partial\theta} - \sum_{\boldsymbol{u},\boldsymbol{g}} \frac{e^{E(\boldsymbol{u},\boldsymbol{g};\boldsymbol{\theta})}}{Z(\boldsymbol{\theta})} \frac{\partial E(\boldsymbol{u},\boldsymbol{g};\boldsymbol{\theta})}{\partial\theta} \\
&= \sum_{\boldsymbol{h}} \frac{p(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\theta})}{p(\boldsymbol{v};\boldsymbol{\theta})} \frac{\partial E(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\theta})}{\partial\theta} - \sum_{\boldsymbol{u},\boldsymbol{g}} p(\boldsymbol{u},\boldsymbol{g};\boldsymbol{\theta}) \frac{\partial E(\boldsymbol{u},\boldsymbol{g};\boldsymbol{\theta})}{\partial\theta} \\
&= \sum_{\boldsymbol{h}} \frac{p(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\theta})}{p(\boldsymbol{v};\boldsymbol{\theta})} \frac{\partial E(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\theta})}{\partial\theta} - \sum_{\boldsymbol{u}} p(\boldsymbol{u};\boldsymbol{\theta}) \sum_{\boldsymbol{g}} \frac{p(\boldsymbol{u},\boldsymbol{g};\boldsymbol{\theta})}{p(\boldsymbol{u};\boldsymbol{\theta})} \frac{\partial E(\boldsymbol{u},\boldsymbol{g};\boldsymbol{\theta})}{\partial\theta}. \qquad (2.7)
\end{aligned}$$

2.3 Derivation of $p(\boldsymbol{h}|\boldsymbol{v};\boldsymbol{\theta}) = \frac{p(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\theta})}{p(\boldsymbol{v};\boldsymbol{\theta})}$

In view of (2.3),

$$\begin{aligned}
e^{E(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\theta})} &= e^{\sum_{i=1}^I b_i v_i} \cdot e^{\sum_{i=1}^I \sum_{j=1}^J v_i w_{ij} h_j + \sum_{j=1}^J c_j h_j} = e^{\sum_{i=1}^I b_i v_i} \cdot e^{\sum_{j=1}^J h_j (c_j + \sum_{i=1}^I v_i w_{ij})} \\
&= e^{\sum_{i=1}^I b_i v_i} \cdot \prod_{j=1}^J e^{h_j (c_j + \sum_{i=1}^I v_i w_{ij})}. \qquad (2.8)
\end{aligned}$$

$$\begin{aligned}
\sum_{\boldsymbol{g}} e^{E(\boldsymbol{v},\boldsymbol{g};\boldsymbol{\theta})} &= e^{\sum_{i=1}^I b_i v_i} \cdot \sum_{g_1=0}^1 \sum_{g_2=0}^1 \cdots \sum_{g_J=0}^1 \prod_{j=1}^J e^{g_j (c_j + \sum_{i=1}^I v_i w_{ij})} \\
= e^{\sum_{i=1}^I b_i v_i} \cdot \prod_{j=1}^J \sum_{g_j=0}^1 e^{g_j (c_j + \sum_{i=1}^I v_i w_{ij})} &= e^{\sum_{i=1}^I b_i v_i} \cdot \prod_{j=1}^J (1 + e^{c_j + \sum_{i=1}^I v_i w_{ij}}). \qquad (2.9)
\end{aligned}$$

So

$$p(\boldsymbol{h}|\boldsymbol{v};\boldsymbol{\theta}) = \frac{p(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\theta})}{p(\boldsymbol{v};\boldsymbol{\theta})} = \frac{e^{E(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\theta})}}{\sum_{\boldsymbol{g}} e^{E(\boldsymbol{v},\boldsymbol{g};\boldsymbol{\theta})}} = \prod_{j=1}^J \frac{e^{h_j (c_j + \sum_{i=1}^I v_i w_{ij})}}{1 + e^{c_j + \sum_{i=1}^I v_i w_{ij}}}. \qquad (2.10)$$

2.4 Derivation of $\sum_{\boldsymbol{h}} \frac{p(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\theta})}{p(\boldsymbol{v};\boldsymbol{\theta})} \frac{\partial E(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\theta})}{\partial\theta}$

$$\begin{aligned}
\sum_{\boldsymbol{h}} \frac{p(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\theta})}{p(\boldsymbol{v};\boldsymbol{\theta})} \frac{\partial E(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\theta})}{\partial w_{ij}} &= \sum_{\boldsymbol{h}} \prod_{k=1}^J \frac{e^{h_k (c_k + \sum_{i=1}^I v_i w_{ik})}}{1 + e^{c_k + \sum_{i=1}^I v_i w_{ik}}} v_i h_j \\
= v_i \sum_{h_1=0}^1 \cdots \sum_{h_{j-1}=0}^1 \sum_{h_{j+1}=0}^1 \cdots \sum_{h_J=0}^1 \sum_{h_j=0}^1 \prod_{k=1,k\neq j}^J & \frac{e^{h_k (c_k + \sum_{i=1}^I v_i w_{ik})}}{1 + e^{c_k + \sum_{i=1}^I v_i w_{ik}}} \cdot \frac{e^{h_j (c_j + \sum_{i=1}^I v_i w_{ij})}}{1 + e^{c_j + \sum_{i=1}^I v_i w_{ij}}} h_j \\
= v_i \sum_{h_1=0}^1 \cdots \sum_{h_{j-1}=0}^1 \sum_{h_{j+1}=0}^1 \cdots \sum_{h_J=0}^1 \prod_{k=1,k\neq j}^J & \frac{e^{h_k (c_k + \sum_{i=1}^I v_i w_{ik})}}{1 + e^{c_k + \sum_{i=1}^I v_i w_{ik}}} \cdot \sum_{h_j=0}^1 \frac{e^{h_j (c_j + \sum_{i=1}^I v_i w_{ij})}}{1 + e^{c_j + \sum_{i=1}^I v_i w_{ij}}} h_j \\
= v_i \frac{e^{c_j + \sum_{i=1}^I v_i w_{ij}}}{1 + e^{c_j + \sum_{i=1}^I v_i w_{ij}}} \sum_{h_1=0}^1 \cdots \sum_{h_{j-1}=0}^1 \sum_{h_{j+1}=0}^1 \cdots \sum_{h_J=0}^1 & \prod_{k=1,k\neq j}^J \frac{e^{h_k (c_k + \sum_{i=1}^I v_i w_{ik})}}{1 + e^{c_k + \sum_{i=1}^I v_i w_{ik}}} \\
&= v_i \frac{e^{c_j + \sum_{i=1}^I v_i w_{ij}}}{1 + e^{c_j + \sum_{i=1}^I v_i w_{ij}}} \prod_{k=1,k\neq j}^J \sum_{h_k=0}^1 \frac{e^{h_k (c_k + \sum_{i=1}^I v_i w_{ik})}}{1 + e^{c_k + \sum_{i=1}^I v_i w_{ik}}}
\end{aligned}$$

$$= v_i \frac{e^{c_j + \sum_{i=1}^{I} v_i w_{ij}}}{1 + e^{c_j + \sum_{i=1}^{I} v_i w_{ij}}}. \tag{2.11}$$

In a similar manner, we have

$$\sum_{\boldsymbol{h}} \frac{p(\boldsymbol{v}, \boldsymbol{h}; \boldsymbol{\theta})}{p(\boldsymbol{v}; \boldsymbol{\theta})} \frac{\partial E(\boldsymbol{v}, \boldsymbol{h}; \boldsymbol{\theta})}{\partial b_i} = v_i, \tag{2.12}$$

$$\sum_{\boldsymbol{h}} \frac{p(\boldsymbol{v}, \boldsymbol{h}; \boldsymbol{\theta})}{p(\boldsymbol{v}; \boldsymbol{\theta})} \frac{\partial E(\boldsymbol{v}, \boldsymbol{h}; \boldsymbol{\theta})}{\partial c_j} = \frac{e^{c_j + \sum_{i=1}^{I} v_i w_{ij}}}{1 + e^{c_j + \sum_{i=1}^{I} v_i w_{ij}}}. \tag{2.13}$$

2.5 Detailed formula for the gradient of the log-likelihood

$$\frac{\partial L(\boldsymbol{\theta}; \boldsymbol{v})}{\partial w_{ij}} = v_i \cdot \sigma(c_j + \sum_{i=1}^{I} v_i w_{ij}) - \sum_{\boldsymbol{u}} p(\boldsymbol{u}; \boldsymbol{\theta}) \cdot u_i \cdot \sigma(c_j + \sum_{i=1}^{I} u_i w_{ij}) \tag{2.14}$$

$$\frac{\partial L(\boldsymbol{\theta}; \boldsymbol{v})}{\partial b_i} = v_i - \sum_{\boldsymbol{u}} p(\boldsymbol{u}; \boldsymbol{\theta}) \cdot u_i \tag{2.15}$$

$$\frac{\partial L(\boldsymbol{\theta}; \boldsymbol{v})}{\partial c_j} = \sigma(c_j + \sum_{i=1}^{I} v_i w_{ij}) - \sum_{\boldsymbol{u}} p(\boldsymbol{u}; \boldsymbol{\theta}) \cdot \sigma(c_j + \sum_{i=1}^{I} u_i w_{ij}). \tag{2.16}$$

2.6 Derivation of $p(\boldsymbol{v}|\boldsymbol{h}; \boldsymbol{\theta}) = \frac{p(\boldsymbol{v}, \boldsymbol{h}; \boldsymbol{\theta})}{p(\boldsymbol{h}; \boldsymbol{\theta})}$

In Gibbs sampling for the calculation of the second term (model average) on the right-hand side of (2.14-16), both $p(\boldsymbol{h}|\boldsymbol{v}; \boldsymbol{\theta})$ and $p(\boldsymbol{v}|\boldsymbol{h}; \boldsymbol{\theta})$ are needed. $p(\boldsymbol{h}|\boldsymbol{v}; \boldsymbol{\theta})$ is derived in Section 2.3, and we now derive $p(\boldsymbol{v}|\boldsymbol{h}; \boldsymbol{\theta})$ here. In view of (2.3),

$$e^{E(\boldsymbol{v}, \boldsymbol{h}; \boldsymbol{\theta})} = e^{\sum_{j=1}^{J} c_j h_j} \cdot e^{\sum_{i=1}^{I} \sum_{j=1}^{J} v_i w_{ij} h_j + \sum_{i=1}^{I} b_i v_i} = e^{\sum_{j=1}^{J} c_j h_j} \cdot e^{\sum_{i=1}^{I} v_i (\sum_{j=1}^{J} w_{ij} h_j + b_i)}$$

$$= e^{\sum_{j=1}^{J} c_j h_j} \cdot \prod_{i=1}^{I} e^{v_i (b_i + \sum_{j=1}^{J} w_{ij} h_j)}. \tag{2.17}$$

$$\sum_{\boldsymbol{u}} e^{E(\boldsymbol{u}, \boldsymbol{h}; \boldsymbol{\theta})} = e^{\sum_{j=1}^{J} c_j h_j} \cdot \sum_{u_1=0}^{1} \sum_{u_2=0}^{1} \cdots \sum_{u_I=0}^{1} \prod_{i=1}^{I} e^{u_i (b_i + \sum_{j=1}^{J} w_{ij} h_j)}$$

$$= e^{\sum_{j=1}^{J} c_j h_j} \cdot \prod_{i=1}^{I} \sum_{u_i=0}^{1} e^{u_i (b_i + \sum_{j=1}^{J} w_{ij} h_j)} = e^{\sum_{j=1}^{J} c_j h_j} \cdot \prod_{i=1}^{I} (1 + e^{b_i + \sum_{j=1}^{J} w_{ij} h_j}). \tag{2.18}$$

So

$$p(\boldsymbol{v}|\boldsymbol{h}; \boldsymbol{\theta}) = \frac{p(\boldsymbol{v}, \boldsymbol{h}; \boldsymbol{\theta})}{p(\boldsymbol{h}; \boldsymbol{\theta})} = \frac{e^{E(\boldsymbol{v}, \boldsymbol{h}; \boldsymbol{\theta})}}{\sum_{\boldsymbol{u}} e^{E(\boldsymbol{u}, \boldsymbol{h}; \boldsymbol{\theta})}} = \prod_{i=1}^{I} \frac{e^{v_i (b_i + \sum_{j=1}^{J} w_{ij} h_j)}}{1 + e^{b_i + \sum_{j=1}^{J} w_{ij} h_j}}. \tag{2.19}$$

3 Derivations for GB-RBM

In this section, we consider GB-RBMs, which are RBMs with continuous visible units and binary hidden units, i.e., $\boldsymbol{v} \in \mathbb{R}^I$ and $\boldsymbol{h} \in \{0, 1\}^J$. The GB-RBM with parameters $\boldsymbol{\theta}$ represents a probability distribution

$$p(\boldsymbol{v}, \boldsymbol{h}; \boldsymbol{\theta}) = \frac{e^{E(\boldsymbol{v}, \boldsymbol{h}; \boldsymbol{\theta})}}{Z(\boldsymbol{\theta})} \tag{3.1}$$

where

$$Z(\boldsymbol{\theta}) = \underset{\boldsymbol{u}\in\mathbb{R}^I}{\int\cdots\int} \sum_{\boldsymbol{g}} e^{E(\boldsymbol{u},\boldsymbol{g};\boldsymbol{\theta})} d\boldsymbol{u} \tag{3.2}$$

is the normalizing constant which is called the partition function and $E(\boldsymbol{v}, \boldsymbol{h}; \boldsymbol{\theta})$ is the negative energy function, which is defined as

$$E(\boldsymbol{v}, \boldsymbol{h}; \boldsymbol{\theta}) = \sum_{i=1}^{I}\sum_{j=1}^{J} a v_i w_{ij} h_j - \sum_{i=1}^{I} \frac{a^2}{2}(v_i - b_i)^2 + \sum_{j=1}^{J} c_j h_j \tag{3.3}$$

where $\boldsymbol{\theta} = \boldsymbol{w} \in \mathbb{R}^{I\times J}, \boldsymbol{b} \in \mathbb{R}^I, \boldsymbol{c} \in \mathbb{R}^J$ is the set of model parameters. In (3.3), a is supposed to be the inverse of the standard deviation of the visible variables (we assume they share a standard deviation here, but we can also assume otherwise). In practice, with proper preprocessing of the data, a is usually treated as a hyper-parameter, i.e. not included in the trained parameters. See [3, 4] for discussions.

3.1 Definition of $p(\boldsymbol{v}; \boldsymbol{\theta})$

We already have

$$p(\boldsymbol{v}, \boldsymbol{h}; \boldsymbol{\theta}) = \frac{e^{E(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\theta})}}{Z(\boldsymbol{\theta})} = \frac{e^{E(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\theta})}}{\int\cdots\int_{\boldsymbol{u}\in\mathbb{R}^I}\sum_{\boldsymbol{g}} e^{E(\boldsymbol{u},\boldsymbol{g};\boldsymbol{\theta})} d\boldsymbol{u}}.$$

And the probability density function of the visible variables is thus

$$p(\boldsymbol{v}; \boldsymbol{\theta}) = \sum_{\boldsymbol{h}} p(\boldsymbol{v}, \boldsymbol{h}; \boldsymbol{\theta}) = \frac{\sum_{\boldsymbol{h}} e^{E(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\theta})}}{\int\cdots\int_{\boldsymbol{u}\in\mathbb{R}^I}\sum_{\boldsymbol{g}} e^{E(\boldsymbol{u},\boldsymbol{g};\boldsymbol{\theta})} d\boldsymbol{u}}. \tag{3.4}$$

Then

$$\frac{p(\boldsymbol{v}, \boldsymbol{h}; \boldsymbol{\theta})}{p(\boldsymbol{v}; \boldsymbol{\theta})} = \frac{e^{E(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\theta})}}{\sum_{\boldsymbol{g}} e^{E(\boldsymbol{v},\boldsymbol{g};\boldsymbol{\theta})}}. \tag{3.5}$$

3.2 General formula for the gradient of the log-likelihood

Learning the RBM parameters $\boldsymbol{\theta}$ is through the maximization of the log-likelihood over the training samples. The log-likelihood, for a single training sample $\boldsymbol{v}$, is given by

$$\begin{aligned} L(\boldsymbol{\theta}; \boldsymbol{v}) &= \log p(\boldsymbol{v}; \boldsymbol{\theta}) = \log \sum_{\boldsymbol{h}} p(\boldsymbol{v}, \boldsymbol{h}; \boldsymbol{\theta}) \\ &= \log \sum_{\boldsymbol{h}} e^{E(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\theta})} - \log \underset{\boldsymbol{u}\in\mathbb{R}^I}{\int\cdots\int} \sum_{\boldsymbol{g}} e^{E(\boldsymbol{u},\boldsymbol{g};\boldsymbol{\theta})} d\boldsymbol{u}. \end{aligned} \tag{3.6}$$

$$\begin{aligned} \frac{\partial L(\boldsymbol{\theta}; \boldsymbol{v})}{\partial \theta} &= \frac{\partial \log \sum_{\boldsymbol{h}} e^{E(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\theta})}}{\partial \theta} - \frac{\partial \log \int\cdots\int_{\boldsymbol{u}\in\mathbb{R}^I}\sum_{\boldsymbol{g}} e^{E(\boldsymbol{u},\boldsymbol{g};\boldsymbol{\theta})} d\boldsymbol{u}}{\partial \theta} \\ &= \frac{\sum_{\boldsymbol{h}} e^{E(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\theta})} \frac{\partial E(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\theta})}{\partial \theta}}{\sum_{\boldsymbol{g}} e^{E(\boldsymbol{v},\boldsymbol{g};\boldsymbol{\theta})}} - \frac{\int\cdots\int_{\boldsymbol{u}\in\mathbb{R}^I}\sum_{\boldsymbol{g}} e^{E(\boldsymbol{u},\boldsymbol{g};\boldsymbol{\theta})} \frac{\partial E(\boldsymbol{u},\boldsymbol{g};\boldsymbol{\theta})}{\partial \theta} d\boldsymbol{u}}{Z(\boldsymbol{\theta})} \\ &= \sum_{\boldsymbol{h}} \frac{e^{E(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\theta})}}{\sum_{\boldsymbol{g}} e^{E(\boldsymbol{v},\boldsymbol{g};\boldsymbol{\theta})}} \frac{\partial E(\boldsymbol{v}, \boldsymbol{h}; \boldsymbol{\theta})}{\partial \theta} - \underset{\boldsymbol{u}\in\mathbb{R}^I}{\int\cdots\int} \sum_{\boldsymbol{g}} \frac{e^{E(\boldsymbol{u},\boldsymbol{g};\boldsymbol{\theta})}}{Z(\boldsymbol{\theta})} \frac{\partial E(\boldsymbol{u}, \boldsymbol{g}; \boldsymbol{\theta})}{\partial \theta} d\boldsymbol{u} \\ &= \sum_{\boldsymbol{h}} \frac{p(\boldsymbol{v}, \boldsymbol{h}; \boldsymbol{\theta})}{p(\boldsymbol{v}; \boldsymbol{\theta})} \frac{\partial E(\boldsymbol{v}, \boldsymbol{h}; \boldsymbol{\theta})}{\partial \theta} - \underset{\boldsymbol{u}\in\mathbb{R}^I}{\int\cdots\int} \sum_{\boldsymbol{g}} p(\boldsymbol{u}, \boldsymbol{g}; \boldsymbol{\theta}) \frac{\partial E(\boldsymbol{u}, \boldsymbol{g}; \boldsymbol{\theta})}{\partial \theta} d\boldsymbol{u} \\ &= \sum_{\boldsymbol{h}} \frac{p(\boldsymbol{v}, \boldsymbol{h}; \boldsymbol{\theta})}{p(\boldsymbol{v}; \boldsymbol{\theta})} \frac{\partial E(\boldsymbol{v}, \boldsymbol{h}; \boldsymbol{\theta})}{\partial \theta} - \underset{\boldsymbol{u}\in\mathbb{R}^I}{\int\cdots\int} p(\boldsymbol{u}; \boldsymbol{\theta}) \cdot \sum_{\boldsymbol{g}} \frac{p(\boldsymbol{u}, \boldsymbol{g}; \boldsymbol{\theta})}{p(\boldsymbol{u}; \boldsymbol{\theta})} \frac{\partial E(\boldsymbol{u}, \boldsymbol{g}; \boldsymbol{\theta})}{\partial \theta} \cdot d\boldsymbol{u}. \end{aligned} \tag{3.7}$$

3.3 Derivation of $p(\boldsymbol{h}|\boldsymbol{v};\boldsymbol{\theta}) = \frac{p(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\theta})}{p(\boldsymbol{v};\boldsymbol{\theta})}$

In view of (3.3),

$$\begin{aligned} e^{E(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\theta})} &= e^{-\sum_{i=1}^{I}\frac{a^2}{2}(v_i-b_i)^2}\cdot e^{\sum_{i=1}^{I}\sum_{j=1}^{J}av_iw_{ij}h_j+\sum_{j=1}^{J}c_jh_j} \\ &= e^{-\sum_{i=1}^{I}\frac{a^2}{2}(v_i-b_i)^2}\cdot e^{\sum_{j=1}^{J}h_j(c_j+\sum_{i=1}^{I}av_iw_{ij})} \\ &= e^{-\sum_{i=1}^{I}\frac{a^2}{2}(v_i-b_i)^2}\cdot\prod_{j=1}^{J}e^{h_j(c_j+\sum_{i=1}^{I}av_iw_{ij})}. \end{aligned} \tag{3.8}$$

$$\begin{aligned} \sum_{\boldsymbol{g}}e^{E(\boldsymbol{v},\boldsymbol{g};\boldsymbol{\theta})} &= e^{-\sum_{i=1}^{I}\frac{a^2}{2}(v_i-b_i)^2}\cdot\sum_{g_1=0}^{1}\sum_{g_2=0}^{1}\cdots\sum_{g_J=0}^{1}\prod_{j=1}^{J}e^{g_j(c_j+\sum_{i=1}^{I}av_iw_{ij})} \\ &= e^{-\sum_{i=1}^{I}\frac{a^2}{2}(v_i-b_i)^2}\cdot\prod_{j=1}^{J}\sum_{g_j=0}^{1}e^{g_j(c_j+\sum_{i=1}^{I}av_iw_{ij})} \\ &= e^{-\sum_{i=1}^{I}\frac{a^2}{2}(v_i-b_i)^2}\cdot\prod_{j=1}^{J}(1+e^{c_j+\sum_{i=1}^{I}av_iw_{ij}}). \end{aligned} \tag{3.9}$$

So

$$p(\boldsymbol{h}|\boldsymbol{v};\boldsymbol{\theta}) = \frac{p(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\theta})}{p(\boldsymbol{v};\boldsymbol{\theta})} = \frac{e^{E(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\theta})}}{\sum_{\boldsymbol{g}}e^{E(\boldsymbol{v},\boldsymbol{g};\boldsymbol{\theta})}} = \prod_{j=1}^{J}\frac{e^{h_j(c_j+\sum_{i=1}^{I}av_iw_{ij})}}{1+e^{c_j+\sum_{i=1}^{I}av_iw_{ij}}}. \tag{3.10}$$

3.4 Derivation of $\sum_{\boldsymbol{h}}\frac{p(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\theta})}{p(\boldsymbol{v};\boldsymbol{\theta})}\frac{\partial E(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\theta})}{\partial\theta}$

$$\begin{aligned} \sum_{\boldsymbol{h}}\frac{p(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\theta})}{p(\boldsymbol{v};\boldsymbol{\theta})}\frac{\partial E(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\theta})}{\partial w_{ij}} &= \sum_{\boldsymbol{h}}\prod_{k=1}^{J}\frac{e^{h_k(c_k+\sum_{i=1}^{I}av_iw_{ik})}}{1+e^{c_k+\sum_{i=1}^{I}av_iw_{ik}}}av_ih_j \\ = av_i\sum_{h_1=0}^{1}\cdots\sum_{h_{j-1}=0}^{1}\sum_{h_{j+1}=0}^{1}\cdots\sum_{h_J=0}^{1}\sum_{h_j=0}^{1}\prod_{k=1,k\neq j}^{J}&\frac{e^{h_k(c_k+\sum_{i=1}^{I}av_iw_{ik})}}{1+e^{c_k+\sum_{i=1}^{I}av_iw_{ik}}}\cdot\frac{e^{h_j(c_j+\sum_{i=1}^{I}av_iw_{ij})}}{1+e^{c_j+\sum_{i=1}^{I}av_iw_{ij}}}h_j \\ = av_i\sum_{h_1=0}^{1}\cdots\sum_{h_{j-1}=0}^{1}\sum_{h_{j+1}=0}^{1}\cdots\sum_{h_J=0}^{1}\prod_{k=1,k\neq j}^{J}&\frac{e^{h_k(c_k+\sum_{i=1}^{I}av_iw_{ik})}}{1+e^{c_k+\sum_{i=1}^{I}aw_{kj}v_i}}\cdot\sum_{h_j=0}^{1}\frac{e^{h_j(c_j+\sum_{i=1}^{I}av_iw_{ij})}}{1+e^{c_j+\sum_{i=1}^{I}av_iw_{ij}}}h_j \\ = av_i\frac{e^{c_j+\sum_{i=1}^{I}av_iw_{ij}}}{1+e^{c_j+\sum_{i=1}^{I}av_iw_{ij}}}\sum_{h_1=0}^{1}\cdots\sum_{h_{j-1}=0}^{1}\sum_{h_{j+1}=0}^{1}\cdots\sum_{h_J=0}^{1}&\prod_{k=1,k\neq j}^{J}\frac{e^{h_k(c_k+\sum_{i=1}^{I}av_iw_{ik})}}{1+e^{c_k+\sum_{i=1}^{I}av_iw_{ik}}} \\ &= av_i\frac{e^{c_j+\sum_{i=1}^{I}av_iw_{ij}}}{1+e^{c_j+\sum_{i=1}^{I}av_iw_{ij}}}\prod_{k=1,k\neq j}^{J}\sum_{h_k=0}^{1}\frac{e^{h_k(c_k+\sum_{i=1}^{I}av_iw_{ik})}}{1+e^{c_k+\sum_{i=1}^{I}av_iw_{ik}}} \\ &= av_i\frac{e^{c_j+\sum_{i=1}^{I}av_iw_{ij}}}{1+e^{c_j+\sum_{i=1}^{I}av_iw_{ij}}}. \end{aligned} \tag{3.11}$$

In a similar manner, we have

$$\sum_{\boldsymbol{h}}\frac{p(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\theta})}{p(\boldsymbol{v};\boldsymbol{\theta})}\frac{\partial E(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\theta})}{\partial b_i} = a^2(v_i-b_i) \tag{3.12}$$

$$\sum_{\boldsymbol{h}}\frac{p(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\theta})}{p(\boldsymbol{v};\boldsymbol{\theta})}\frac{\partial E(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\theta})}{\partial c_j} = \frac{e^{c_j+\sum_{i=1}^{I}av_iw_{ij}}}{1+e^{c_j+\sum_{i=1}^{I}av_iw_{ij}}}. \tag{3.13}$$

doi:10.20944/preprints202003.0337.v1

3.5 Detailed formula for the gradient of the log-likelihood

$$\frac{\partial L(\boldsymbol{\theta};\boldsymbol{v})}{\partial w_{ij}} = av_i \cdot \sigma(c_j + \sum_{i=1}^{I} av_i w_{ij}) - \sum_{\boldsymbol{u}} p(\boldsymbol{u};\boldsymbol{\theta}) \cdot au_i \cdot \sigma(c_j + \sum_{i=1}^{I} au_i w_{ij}) \tag{3.14}$$

$$\frac{\partial L(\boldsymbol{\theta};\boldsymbol{v})}{\partial b_i} = a^2 v_i - \sum_{\boldsymbol{u}} p(\boldsymbol{u};\boldsymbol{\theta}) \cdot a^2 u_i \tag{3.15}$$

$$\frac{\partial L(\boldsymbol{\theta};\boldsymbol{v})}{\partial c_j} = \sigma(c_j + \sum_{i=1}^{I} av_i w_{ij}) - \sum_{\boldsymbol{u}} p(\boldsymbol{u};\boldsymbol{\theta}) \cdot \sigma(c_j + \sum_{i=1}^{I} au_i w_{ij}). \tag{3.16}$$

3.6 Derivation of $p(\boldsymbol{v}|\boldsymbol{h};\boldsymbol{\theta}) = \frac{p(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\theta})}{p(\boldsymbol{h};\boldsymbol{\theta})}$

In Gibbs sampling for the calculation of the second term (model average) on the right-hand side of (3.14-16), both $p(\boldsymbol{h}|\boldsymbol{v};\boldsymbol{\theta})$ and $p(\boldsymbol{v}|\boldsymbol{h};\boldsymbol{\theta})$ are needed. $p(\boldsymbol{h}|\boldsymbol{v};\boldsymbol{\theta})$ is derived in Section 3.3, and we now derive $p(\boldsymbol{v}|\boldsymbol{h};\boldsymbol{\theta})$ here. In view of (3.3),

$$\begin{aligned} e^{E(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\theta})} &= e^{\sum_{j=1}^{J} c_j h_j} \cdot e^{\sum_{i=1}^{I}\sum_{j=1}^{J} av_i w_{ij} h_j - \sum_{i=1}^{I} \frac{a^2}{2}(v_i-b_i)^2} \\ &= e^{\sum_{j=1}^{J} c_j h_j} \cdot e^{\sum_{i=1}^{I}[av_i \sum_{j=1}^{J} w_{ij} h_j - \frac{a^2}{2}(v_i-b_i)^2]} \\ &= e^{\sum_{j=1}^{J} c_j h_j} \cdot \prod_{i=1}^{I} e^{av_i \sum_{j=1}^{J} w_{ij} h_j - \frac{a^2}{2}(v_i-b_i)^2}. \end{aligned} \tag{3.17}$$

$$\begin{aligned} \underset{\boldsymbol{u}\in\mathbb{R}^I}{\int\cdots\int} e^{E(\boldsymbol{u},\boldsymbol{h};\boldsymbol{\theta})} d\boldsymbol{u} &= e^{\sum_{j=1}^{J} c_j h_j} \cdot \underset{\boldsymbol{u}\in\mathbb{R}^I}{\int\cdots\int} \prod_{i=1}^{I} e^{au_i \sum_{j=1}^{J} w_{ij} h_j - \frac{a^2}{2}(u_i-b_i)^2} d\boldsymbol{u} \\ &= e^{\sum_{j=1}^{J} c_j h_j} \cdot \prod_{i=1}^{I} \int_{-\infty}^{\infty} e^{au_i \sum_{j=1}^{J} w_{ij} h_j - \frac{a^2}{2}(u_i-b_i)^2} du_i \\ &= e^{\sum_{j=1}^{J} c_j h_j} \cdot \prod_{i=1}^{I} e^{\frac{(2ab_i + \sum_{j=1}^{J} w_{ij} h_j)(\sum_{j=1}^{J} w_{ij} h_j)}{2}} \int_{-\infty}^{\infty} e^{-\frac{[a(u_i-b_i) - \sum_{j=1}^{J} w_{ij} h_j]^2}{2}} du_i \\ &= e^{\sum_{j=1}^{J} c_j h_j} \cdot \prod_{i=1}^{I} e^{\frac{(2ab_i + \sum_{j=1}^{J} w_{ij} h_j)(\sum_{j=1}^{J} w_{ij} h_j)}{2}} \frac{\sqrt{2\pi}}{a}. \end{aligned} \tag{3.18}$$

So

$$p(\boldsymbol{v}|\boldsymbol{h};\boldsymbol{\theta}) = \frac{p(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\theta})}{p(\boldsymbol{h};\boldsymbol{\theta})} = \frac{e^{E(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\theta})}}{\underset{\boldsymbol{u}\in\mathbb{R}^I}{\int\cdots\int} e^{E(\boldsymbol{u},\boldsymbol{h};\boldsymbol{\theta})} d\boldsymbol{u}} = \prod_{i=1}^{I} \frac{a}{\sqrt{2\pi}} e^{-\frac{[a(v_i-b_i) - \sum_{j=1}^{J} w_{ij} h_j]^2}{2}}. \tag{3.19}$$

References

[1] Y. Bengio. Learning deep architectures for AI. *Foundations and Trends in Machine Learning*, 2(1):1–55, 2009.

[2] A. Fisher and C. Igel. Training restricted Boltzmann machines: An introduction. *Pattern Recognition*, 47(1):25–39, 2014.

[3] G. Hinton. A practical guide to training restricted Boltzmann machines. Technical Report UTML TR 2010–003, Department of Computer Science, University of Toronto, 2010.

[4] K. Cho, A. Ilin, and T. Raiko. Improved learning of Gaussian-Bernoulli restricted Boltzmann machines. In T. Honkela et al., editor, *ICANN 2011, Part I*, number 6791 in LNCS, pages 10–17. Springer, 2011.