

GROUP2505 – Log-likelihood analysis for Restricted Boltzmann Machines

Elias Maria Bonasera, Alberto Casellato, Nicola Garbin, and Francesco Pazzocco
(Dated: March 27, 2025)

[illegible]

1. INTRODUCTION

The main goal of this assignment is to explore how the performance of an RBM changes for different choices of the hyperparameters of the model, using the MNIST digits as the database; in particular using the log-likelihood evaluation we explore the trend of the model during the learning process as the number of epochs increases.

2. RESTRICTED BOLTZMANN MACHINES

Restricted Boltzmann Machines (RBM) are a powerful kind of generative models designed to accomplish training processes relatively fast. In RBMs, a set of D binary visible units i of state $\{v_i\}_{i=1,\dots,D}$ is symmetrically connected to a set of L binary hidden units μ of state $\{h_\mu\}_{\mu=1,\dots,L}$; the continuous weight $w_{i\mu}$ quantifies the strength between unit i and unit μ (see Figure 1). RBMs use an energy function to define the probability distribution over the input data. In the training process the energy of the configuration is minimized by adjusting the parameters θ . One of the most adopted training algorithm is contrastive divergence which allows to approximate the gradient of the likelihood to update the parameters. During this process, a cyclic Gibbs sampling is performed setting the visible units given the hidden ones and vice versa, according to the following probabilities p :

$$p(h_\mu = 1 \mid \mathbf{v}) = \sigma(b_\mu + \sum_i v_i w_{i\mu}) \quad (1)$$

$$p(v_i = 1 \mid \mathbf{h}) = \sigma(a_i + \sum_{\mu} h_{\mu} w_{i\mu}) \quad (2)$$

where $\sigma(x) = 1/(1 + e^{-x})$ is the logistic sigmoid function, a_i is the bias of the i -th visible unit and b_μ is the bias of the μ -th hidden unit; they act shifting the sigmoid function $\sigma(x)$. The absence of links among units of the same type simplifies the training process. Moreover, the number of iterations of Eq. 1 and Eq. 2 can be set to 2 if real data is used to fix $\{v_i\}_{i=1,\dots,D}$ in the first place.

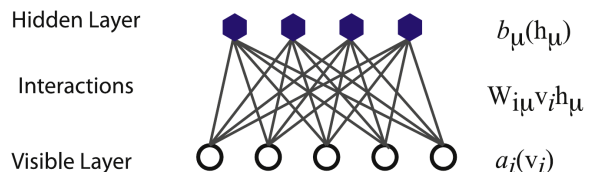


FIG. 1. A Restricted Boltzmann Machine (RBM) is made up of visible units, denoted as v_i , and hidden units, represented as h_μ . These units engage with one another through interactions characterized by the weights $w_{i\mu}$. Notably, there are no direct interactions among the visible units or among the hidden units themselves.

The goodness of a model is evaluated by computing the log-likelihood \mathcal{L} of the data:

$$\mathcal{L} = \frac{1}{M} \sum_{m=1}^M \ell_{\theta}(v^{(m)}) \quad (3)$$

$$\ell_{\theta}(v^{(m)}) = \log \sum_h e^{-E(v,h)} - \log Z \quad (4)$$

where M is the number of data points, Z is the partition function, $E(v, h)$ in the exponential argument $e^{-E(v, h)} = G(h) \prod_i e^{H_i(h)v_i}$ is the energy function, with $G(h) = \prod_\mu e^{b_\mu h_\mu}$ and $H_i(h) = a_i + \sum_\mu w_{i\mu} h_\mu$, and, in Eq. 4, the index h of the summation represents each possible state of the hidden layer.

3. METHODS

3.1. Theoretical implementation

Regarding the computation of the partition function Z (in Eq. 4), its implementation depends on the set of values which each unit can be set on. Since the direct computation of the Z function is unfeasible, due to the summation of all possible state combinations of the system, a possible approach is to split the summation separately over v and h .

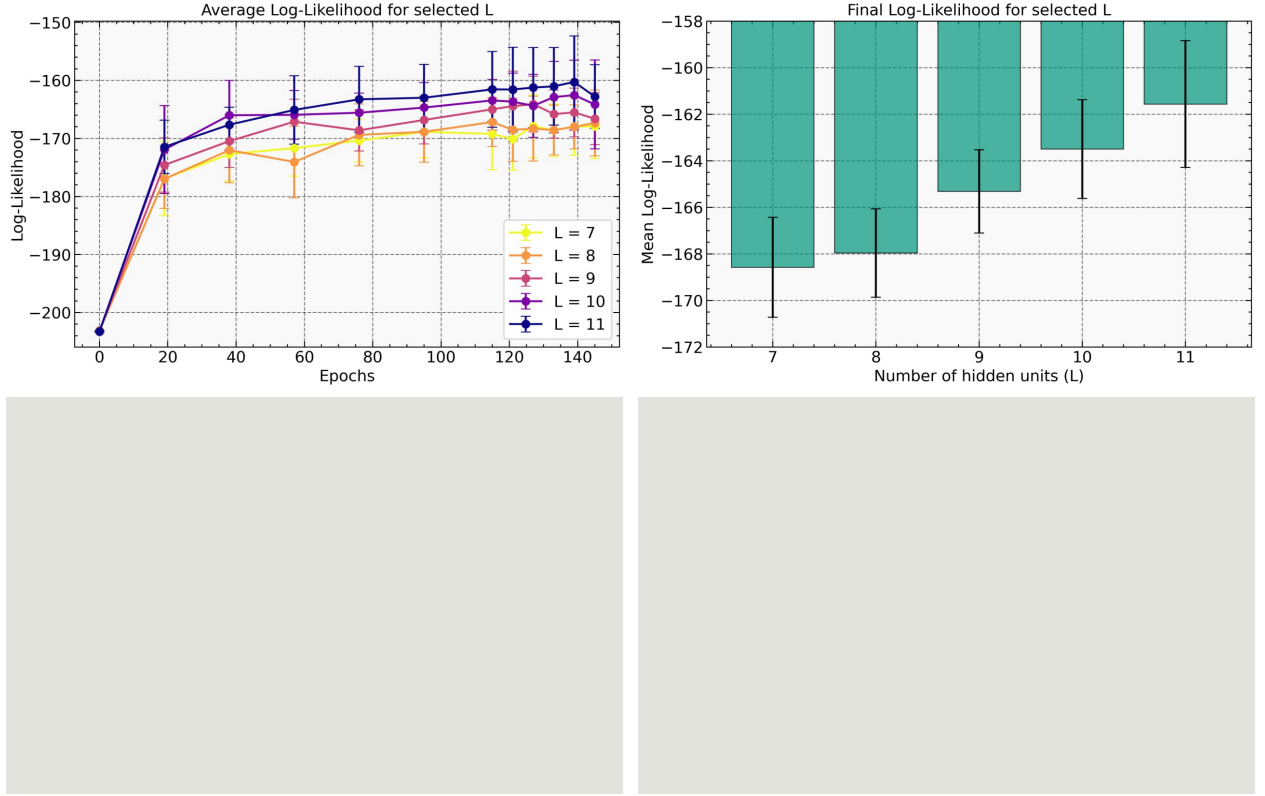


FIG. 2. Description of the panels: (a)..... (b)... etc. This caption should give enough info on the content of figures to make them mostly readable without consulting the main text. However, repetitions with the main text should e avoided if possible. If this format is difficult to frame in the page you want, just break it into multiple single figures.

For the Bernulli case the units can assume the values $\{0, 1\}$ and the partition function form can be manipulated to obtain the following expression:

$$Z = \sum_v \sum_h e^{-E(v,h)} = q^D \sum_h G(h) \prod_{i=1}^D \frac{1 + e^{H_i(h)}}{q} \quad (5)$$

where q is the average of $1 + e^{H_i(h)}$ $\forall i, h$ that we introduce to avoid overflow. Concerning the training process of RBM, \mathbf{v} is set using real data. Weights $w_{i\mu}$ are initialized sampling values from a Gaussian distribution of mean 0 and standard deviation of $2/\sqrt{L+D}$. Biases b_μ of hidden units are initialized to 0. Thus \mathbf{h} is evaluated by Eq. 1 before passed to Eq. 2 in order to compute back \mathbf{v} . In such evaluation it is convenient to set $a_i = \log[p_i/(1-p_i)]$; fixed the i -th visible unit, p_i is defined as the number of times such unit is on over the whole training array set, normalized over the size of the set. Referring to Eq. 2, the choice of shifting the sigmoid $\sigma(x)$ by the function $\log[p_i/(1-p_i)]$ of the average p_i ensures that the units can activate properly, even given initial weights $w_{i\mu}$ close to 0.

The initialization of biases a_i , firstly proposed by Hinton, makes hidden units able to activate and differentiate better their activation based on data, avoiding

the risk of getting stuck on values near to 0.

3.2. Computation

We define two Python3 classes, one for the computation of the log-likelihood and the other for defining the RBM structure and managing all the input and output of the training process. Referring to Eq. 4, in order to address the problem of overflow of $e^{H_i(h)v_i}$, we approximate the summation of powers over h as a sum of the biggest ones. The purpose of this choice is to extract the biggest term out of the summation.

Regarding the expression of the partition function shown in Eq. 5, it changes form for Spin case, characterized by units taking values from $\{-1, 1\}$. Manipulating the second member of the equation one gets:

$$Z = \sum_v \sum_h e^{-E(v,h)} = \sum_h G(h) \prod_{i=1}^D 2 \cosh(H_i(h)) \quad (6)$$

For Eq. 6 the $2 \cosh(H_i(h))$ term is approximated to just $e^{H_i(h)}$ by shifting all exponents away from negative values after adding a positive value and properly compensating that out of the summation. Thus the same

ZZZZZZZZZZZZZZZZZZ ZZZZZ ZZZZZZZZZZZ ZZZZZZZ Z ZZZZZZZZZZZ
ZZZZ ZZZZZZZZZZZ ZZZZZZZZZZZZZZZZZZZ ZZZZZ ZZZZZZZZZZZ ZZZZZZZ
Z ZZZZZZZZZZZ ZZZZZ ZZZZZZZZZZZ. ZZZZZZZZZZZZZZZZZ ZZZZZ
ZZZZZZZZZZ ZZZZZZZ Z ZZZZZZZZZZZ ZZZZZ ZZZZZZZZZZZ
ZZZZZZZZZZZZZZZZZZ ZZZZZ ZZZZZZZZZZZ ZZZZZZZ Z ZZZZZZZZZZZ
ZZZZ ZZZZZZZZZZZ ZZZZZZZZZZZZZZZZZZZ ZZZZZ ZZZZZZZZZZZ ZZZZZZZ
Z ZZZZZZZZZZZ ZZZZZ ZZZZZZZZZZZ ZZZZZZZZZZZZZZZZZ ZZZZZ
ZZZZZZZZZZ ZZZZZZZ Z ZZZZZZZZZZZ ZZZZZ ZZZZZZZZZZZ
ZZZZZZZZZZZZZZZZZZ ZZZZZ ZZZZZZZZZZZ ZZZZZZZ Z ZZZZZZZZZZZ
ZZZZ ZZZZZZZZZZZ.

ZZZZZZZZZZZZZZZZZZ ZZZZZ ZZZZZZZZZZZ ZZZZZZZ Z ZZZZZZZZZZZ
ZZZZ ZZZZZZZZZZZ ZZZZZZZZZZZZZZZZZZZ ZZZZZ ZZZZZZZZZZZ ZZZZZZZ
Z ZZZZZZZZZZZ ZZZZZ ZZZZZZZZZZZ ZZZZZZZZZZZZZZZZZ ZZZZZ
ZZZZZZZZZZ ZZZZZZZ Z ZZZZZZZZZZZ ZZZZZ ZZZZZZZZZZZ
ZZZZZZZZZZZZZZZZZZ ZZZZZ ZZZZZZZZZZZ ZZZZZZZ Z ZZZZZZZZZZZ
ZZZZ ZZZZZZZZZZZ ZZZZZZZZZZZZZZZZZZZ ZZZZZ ZZZZZZZZZZZ ZZZZZZZ Z
ZZZZZZZZZZZZ ZZZZZ ZZZZZZZZZZZ.

ZZZZZZZZZZZZZZZZZZ ZZZZZ ZZZZZZZZZZZ ZZZZZZZ Z ZZZZZZZZZZZ
ZZZZ ZZZZZZZZZZZ ZZZZZZZZZZZZZZZZZZZ ZZZZZ ZZZZZZZZZZZ ZZZZZZZ
Z ZZZZZZZZZZZ ZZZZZ ZZZZZZZZZZZ ZZZZZZZZZZZZZZZZZ ZZZZZ
ZZZZZZZZZZ ZZZZZZZ Z ZZZZZZZZZZZ ZZZZZ ZZZZZZZZZZZ
ZZZZZZZZZZZZZZZZZZ ZZZZZ ZZZZZZZZZZZ ZZZZZZZ Z ZZZZZZZZZZZ
ZZZZ ZZZZZZZZZZZ ZZZZZZZZZZZZZZZZZZZ ZZZZZ ZZZZZZZZZZZ ZZZZZZZ Z
ZZZZZZZZZZZZ ZZZZZ ZZZZZZZZZZZ.

5. CONCLUSIONS

Discuss the key aspects that we can take home from this work.

Check if your text is light, swift, and correct in exposing its passages.

APPENDIX

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \gamma \cdot \eta_t \cdot \text{sign}(\boldsymbol{\theta}) \quad (7)$$

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta_t \nabla_{\boldsymbol{\theta}} E(\boldsymbol{\theta}) \quad (8)$$

$$\begin{cases} \mathbf{g}_t = \nabla_{\boldsymbol{\theta}} E(\boldsymbol{\theta}) \\ \mathbf{s}_t = \beta \mathbf{s}_{t-1} + (1 - \beta) \mathbf{g}_t^2 \\ \boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta_t \frac{\mathbf{g}_t}{\sqrt{\mathbf{s}_t + \epsilon}} \end{cases} \quad (9)$$

-
- [1] Pankaj Mehta, Marin Bukov, Ching-Hao Wang, Alexandre G.R. Day, Clint Richardson, Charles K. Fisher, David J. Schwab, A high-bias, low-variance introduction to Machine Learning for physicists Physics Reports **810**, 1–124 (2019).
[2] A. Einstein, Int. J. There Here **20**, 125–133 (1910).

Assignment score grid

Structure: the exposition follow a logic order	8
Clarity: the text is brief enough, avoids complicated sentences and specifies all concepts and links	8
Depth: the text is not a shallow repetition of notions, there emerges a good understanding	8
Rigor: the analysis of the results is precise, quantitatively, and convincing	8
Innovation: new methods/ideas are introduced; conclusions beyond what introduced in the class	4