

# Lab Exercise 2: Restricted Boltzmann Machine

Caio Corro

## 1 Introduction

In this lab exercise, you will build a Restricted Boltzmann Machine (RBM) with continuous observed variables and binary latent variables. Hence, we have the following sample spaces:

- $\mathbf{x} \in \mathbb{R}^2$  for the observed variables,
- $\mathbf{z} \in \{0, 1\}^n$  for the latent variables.

The joint probability distribution over all random variables is defined as follows:

$$p(\mathbf{x}, \mathbf{z}; \theta) = \frac{\exp(-e(\mathbf{x}, \mathbf{z}; \theta))}{\int_{\mathbf{x}'} \sum_{\mathbf{z}'} \exp(-e(\mathbf{x}', \mathbf{z}'; \theta)) d\mathbf{x}'} = \frac{\exp(-e(\mathbf{x}, \mathbf{z}; \theta))}{Z(\theta)} = Z(\theta)^{-1} \exp(-e(\mathbf{x}, \mathbf{z}; \theta))$$

where  $f : \mathbb{R}^2 \times \{0, 1\}^n \rightarrow \mathbb{R}$  is the energy function parameterized by  $\theta$  and  $Z(\theta)$  is the partition function. Note that the integral and the sum are over all possible values for the observed and latent variables, respectively. As mentioned in the course, in a RBM we have the following conditional independence by construction of the energy function:

- $p(\mathbf{x}|\mathbf{z}; \theta) = p(x_1|\mathbf{z}; \theta)p(x_2|\mathbf{z}; \theta)$
- $p(\mathbf{z}|\mathbf{x}; \theta) = \prod_{i=1}^n p(z_i|\mathbf{x}; \theta)$

Moreover, both  $p(\mathbf{x}|\mathbf{z}; \theta)$  and  $p(\mathbf{z}|\mathbf{x}; \theta)$  are well known probability distributions. In the example of the course, they were both independent Bernoullis. In this lab exercise,  $p(\mathbf{x}|\mathbf{z}; \theta)$  are independent Gaussians and  $p(\mathbf{z}|\mathbf{x}; \theta)$  are independent Bernoullis (see below for a derivation). This fact allows us to rely on efficient Gibbs sampling in order to sample from the joint distribution  $p(\mathbf{x}, \mathbf{z}; \theta)$ : we start from a random assignation of either  $\mathbf{z}$  or  $\mathbf{x}$ , and then iteratively sample from  $p(\mathbf{x}|\mathbf{z}; \theta)$  and  $p(\mathbf{z}|\mathbf{x}; \theta)$ . For example:

1.  $\forall i : \mathbf{z}_i^{(1)} \sim \mathbb{B}(0.5)$  where  $\mathbb{B}(0.5)$  is an unbiased Bernoulli distribution,
2.  $\mathbf{x}^{(2)} \sim p(\mathbf{x}^{(2)} | \mathbf{z}^{(1)}; \theta)$ ,
3.  $\mathbf{z}^{(2)} \sim p(\mathbf{z}^{(2)} | \mathbf{x}^{(2)}; \theta)$ ,
4.  $\mathbf{x}^{(3)} \sim p(\mathbf{x}^{(3)} | \mathbf{z}^{(2)}; \theta)$ ,
5.  $\mathbf{z}^{(3)} \sim p(\mathbf{z}^{(3)} | \mathbf{x}^{(3)}; \theta)$ ,
6. ...

And we keep one or several couple of values  $\langle \mathbf{x}^{(t)}, \mathbf{z}^{(t)} \rangle$  from this Markov chain as samples.

## 2 Energy function

In this lab exercise, the RBM energy function is defined as follows:

$$e(\mathbf{x}, \mathbf{z}; \theta) = \sum_{i=1}^2 \frac{(x_i - b_i)^2}{2\sigma_i^2} - \sum_{i=1}^2 \sum_{j=1}^n \frac{x_i}{\sigma_i^2} W_{ij} z_j - \sum_{j=1}^n c_j z_j$$

where  $\theta = \{\mathbf{b}, \sigma, \mathbf{W}, \mathbf{c}\}$  is the set of parameters that must be learned from data:

- $\mathbf{b} \in \mathbb{R}^2$ ,
- $\sigma \in \mathbb{R}_+^2$  (that is  $\sigma_1 > 0$  and  $\sigma_2 > 0$ ),
- $\mathbf{W} \in \mathbb{R}^{2 \times n}$ ,
- $\mathbf{c} \in \mathbb{R}^n$ .

We can compute analytically the conditional probability of  $x_1$ :

$$\begin{aligned}
p(x_1|x_2, \mathbf{z}; \theta) &= \frac{p(x_1, x_2, \mathbf{z}; \theta)}{p(x_2, \mathbf{z}; \theta)} \\
&= \frac{p(x_1, x_2, \mathbf{z}; \theta)}{\int_{x'_1} p(x'_1, x_2, \mathbf{z}; \theta) dx'_1} \\
&= \frac{Z(\theta)^{-1} \exp\left(-\frac{(x_1-b_1)^2}{2\sigma_1^2} - \frac{(x_2-b_2)^2}{2\sigma_2^2} + \sum_{j=1}^n \frac{x_1}{\sigma_1^2} W_{1j} z_j + \sum_{j=1}^n \frac{x_2}{\sigma_2^2} W_{2j} z_j + \sum_{j=1}^n c_j z_j\right)}{\int_{x'_1} Z(\theta)^{-1} \exp\left(-\frac{(x'_1-b_1)^2}{2\sigma_1^2} - \frac{(x_2-b_2)^2}{2\sigma_2^2} + \sum_{j=1}^n \frac{x'_1}{\sigma_1^2} W_{1j} z_j + \sum_{j=1}^n \frac{x_2}{\sigma_2^2} W_{2j} z_j + \sum_{j=1}^n c_j z_j\right) dx'_1}
\end{aligned}$$

Remember the exponential rule  $\exp(a+b) = \exp(a)\exp(b)$ . In the denominator, we can get out of the integral the partition function term  $Z(\theta)^{-1}$  and all terms in the exponential that do not depend on  $x'_1$  thanks to the exponential rule. Then, these terms will cancel with their counterparts in the nominator, so we end up with:

$$= \frac{\exp\left(-\frac{(x_1-b_1)^2}{2\sigma_1^2} + \sum_{j=1}^n \frac{x_1}{\sigma_1^2} W_{1j} z_j\right)}{\int_{x'_1} \exp\left(-\frac{(x'_1-b_1)^2}{2\sigma_1^2} + \sum_{j=1}^n \frac{x'_1}{\sigma_1^2} W_{1j} z_j\right) dx'_1}$$

which shows the conditional independence of  $x_1$  given  $x_2$  as the latter value does not appear in the equation anymore. Let  $a = \sum_{j=1}^n W_{1j} z_j$ . Then we can rewrite the argument of the exponential in the numerator argument as:

$$\frac{ax_1}{\sigma_1^2} - \frac{(x_1-b_1)^2}{2\sigma_1^2} = \frac{2ax_1 - x_1^2 - b_1^2 + b_1x_1}{2\sigma_1^2} = \frac{2ax_1 - x_1^2 - b_1^2 + b_1x_1 + a^2 - a^2 + 2ab_1 - 2ab_1}{2\sigma_1^2} = \frac{-(x_1-a-b_1)^2 + a^2 + 2ab_1}{2\sigma_1^2}$$

By using this relation and similarly in the denominator, we can rewrite the conditional probability as:

$$= \frac{\exp\left(\frac{-(x_1 - \sum_{j=1}^n z_j W_{1j} - b_1)^2 + (\sum_{j=1}^n W_{1j} z_j)^2 + 2 \sum_{j=1}^n W_{1j} z_j b_1}{2\sigma_1^2}\right)}{\int_{x'_1} \exp\left(\frac{-(x'_1 - \sum_{j=1}^n W_{1j} z_j - b_1)^2 + (\sum_{j=1}^n W_{1j} z_j)^2 + 2 \sum_{j=1}^n W_{1j} z_j b_1}{2\sigma_1^2}\right) dx'_1}$$

Again, we can simplify by removing terms that do not depend on  $x'_1$  or  $x_1$  by using the exponential rule and taking them out of the integral in the denominator:

$$\begin{aligned}
&= \frac{\exp\left(\frac{-(x_1 - \sum_{j=1}^n W_{1j} z_j - b_1)^2}{2\sigma_1^2}\right)}{\int_{x'_1} \exp\left(\frac{-(x'_1 - \sum_{j=1}^n W_{1j} z_j - b_1)^2}{2\sigma_1^2}\right) dx'_1} \\
&= \frac{\exp\left(-\frac{1}{2} \left(\frac{x_1 - \sum_{j=1}^n W_{1j} z_j - b_1}{\sigma_1}\right)^2\right)}{\int_{x'_1} \exp\left(-\frac{1}{2} \left(\frac{x'_1 - \sum_{j=1}^n W_{1j} z_j - b_1}{\sigma_1}\right)^2\right) dx'_1}
\end{aligned}$$

Now, this value looks like a PDF where the denominator act as a normalizer that ensures that the integral of the PDF is equal to 1. You can observe that the numerator is an unnormalized Gaussian, therefore the normalization term is equal to  $\sigma_1 \sqrt{2\pi}$ . This results can also be shown using closed form formula for Gaussian function integration.<sup>1</sup> By rewriting the denominator as  $\sigma_1 \sqrt{2\pi}$  we have:

$$\begin{aligned}
&= \frac{1}{\sigma_1 \sqrt{2\pi}} \exp\left(-\frac{1}{2} \left(\frac{x_1 - (\sum_{j=1}^n W_{1j} z_j + b_1)}{\sigma_1}\right)^2\right) \\
&= \mathcal{N}\left(x_1 \middle| b_1 + \sum_{j=1}^n W_{1j} z_j, \sigma_1^2\right)
\end{aligned}$$

that is  $p(x_1|x_2, \mathbf{z}; \theta) = p(x_1|\mathbf{z}; \theta)$  is a Gaussian with mean  $b_1 + \sum_{j=1}^n W_{1j} z_j$  and standard deviation  $\sigma_1^2$ . We can do a similar computation for the other observed variable:

$$p(x_2|x_1, \mathbf{z}; \theta) = p(x_2|\mathbf{z}; \theta) = \mathcal{N}\left(x_2 \middle| b_2 + \sum_{j=1}^n W_{2j} z_j, \sigma_2^2\right)$$

<sup>1</sup>See [https://en.wikipedia.org/wiki/Gaussian\\_integral](https://en.wikipedia.org/wiki/Gaussian_integral)

Similarly, we can compute the conditional probability of an observation. If we take the specific case  $p(z_1|x, z_2...z_n; \theta)$ :

$$\begin{aligned} p(z_1|x, z_2...z_n; \theta) &= \frac{\exp(-e(x, [z_1, z_2...z_n]; \theta))}{\sum_{z'_1} \exp(-e(x, [z'_1, z_2...z_n]; \theta))} \\ &= \frac{\exp\left(-\sum_{i=1}^2 \frac{(x_i - b_i)^2}{2\sigma_i^2} + \sum_{i=1}^2 \sum_{j=1}^n \frac{x_i}{\sigma_i^2} W_{ij} z_j + \sum_{j=1}^n c_j z_j\right)}{\sum_{z'_1} \exp\left(-\sum_{i=1}^2 \frac{(x_i - b_i)^2}{2\sigma_i^2} + \sum_{i=1}^2 \frac{x_i}{\sigma_i^2} W_{i1} z'_1 + c_j z'_1 + \sum_{i=1}^2 \sum_{j=2}^n \frac{x_i}{\sigma_i^2} W_{ij} z_j + \sum_{j=2}^n c_j z_j\right)} \end{aligned}$$

As previously, there are many terms that cancels:

$$= \frac{\exp\left(\sum_{i=1}^2 \frac{x_i}{\sigma_i^2} W_{i1} z_1 + c_1 z_1\right)}{\sum_{z'_1} \exp\left(\sum_{i=1}^2 \frac{x_i}{\sigma_i^2} W_{i1} z'_1 + c_1 z'_1\right)}$$

The variable  $z'_1$  can either be 0 or 1 in the denominator's sum. In the case it is 0, the exponential simplifies as  $\exp(0) = 1$ . Moreover, we can factorize by  $z_1$  in the numerator's exponential, therefore:

$$= \frac{\exp\left(z_1 \left(\sum_{i=1}^2 \frac{x_i}{\sigma_i^2} W_{i1} + c_1\right)\right)}{1 + \exp\left(\sum_{i=1}^2 \frac{x_i}{\sigma_i^2} W_{i1} + c_1\right)}$$

Hence, we can see that  $z_1$  is conditionally independent of other latent variables  $p(z_1|x, z_2...z_n; \theta) = p(z_1|x; \theta)$ . Moreover, from the formula we deduce that  $p(z_1|x; \theta)$  is a Bernoulli distribution with parameter  $\mu = \text{sigmoid}(\sum_{i=1}^2 \frac{x_i}{\sigma_i^2} W_{i1} + c_1)$  which we write as:

$$= \mathcal{B}\left(z_1 \middle| \text{sigmoid}\left(c_1 + \sum_{i=1}^2 \frac{x_i}{\sigma_i^2} W_{i1}\right)\right)$$

A similar computation can be done for other latent variables. Note that for both condition distributions over observed and latent variables we can use tensor operations instead of computing parameters for each variable independently. Try to visualize how does it works!

### 3 Training

The parameter  $\theta$  must be set so that the probability of the training data is maximized. To do this, we reformulate the probability of observation using the free energy of  $\mathbf{x}$  denoted  $\mathcal{F}(\mathbf{x}; \theta)$  which can be computed in closed form:

$$\begin{aligned} p(\mathbf{x}; \theta) &= \sum_{\mathbf{z}} p(\mathbf{x}, \mathbf{z}; \theta) \\ &= \sum_{\mathbf{z}} \frac{\exp(-e(\mathbf{x}, \mathbf{z}; \theta))}{\int_{\mathbf{x}'} \sum_{\mathbf{z}'} \exp(-e(\mathbf{x}', \mathbf{z}'; \theta)) d\mathbf{x}'} \\ &= \frac{\sum_{\mathbf{z}} \exp\left(-\sum_{i=1}^2 \frac{(x_i - b_i)^2}{2\sigma_i^2} + \sum_{i=1}^2 \sum_{j=1}^n \frac{x_i}{\sigma_i^2} W_{ij} z_j + \sum_{j=1}^n c_j z_j\right)}{\int_{\mathbf{x}'} \sum_{\mathbf{z}'} \exp\left(-\sum_{i=1}^2 \frac{(x'_i - b_i)^2}{2\sigma_i^2} + \sum_{i=1}^2 \sum_{j=1}^n \frac{x'_i}{\sigma_i^2} W_{ij} z'_j + \sum_{j=1}^n c_j z'_j\right) d\mathbf{x}'} \end{aligned}$$

We can use the fact that  $\exp(a + b) = \exp(a) \exp(b)$  to rewrite both exponentials:

$$= \frac{\sum_{\mathbf{z}} \left[ \exp\left(-\sum_{i=1}^2 \frac{(x_i - b_i)^2}{2\sigma_i^2}\right) \prod_{j=1}^n \exp\left(c_j z_j + \sum_{i=1}^2 \frac{x_i}{\sigma_i^2} W_{ij} z_j\right) \right]}{\int_{\mathbf{x}'} \left[ \sum_{\mathbf{z}'} \exp\left(-\sum_{i=1}^2 \frac{(x'_i - b_i)^2}{2\sigma_i^2}\right) \prod_{j=1}^n \left(c_j z'_j + \sum_{i=1}^2 \frac{x'_i}{\sigma_i^2} W_{ij} z'_j\right) \right] d\mathbf{x}'}$$

Note that the first exponential term in the nominator does not depend on  $\mathbf{z}$ , so we can take it out of the sum, and similarly in the denominator:

$$= \frac{\exp\left(-\sum_{i=1}^2 \frac{(x_i - b_i)^2}{2\sigma_i^2}\right) \sum_{\mathbf{z}} \left[ \prod_{j=1}^n \exp\left(c_j z_j + \sum_{i=1}^2 \frac{x_i}{\sigma_i^2} W_{ij} z_j\right) \right]}{\int_{\mathbf{x}'} \exp\left(-\sum_{i=1}^2 \frac{(x'_i - b_i)^2}{2\sigma_i^2}\right) \sum_{\mathbf{z}'} \left[ \prod_{j=1}^n \left(c_j z'_j + \sum_{i=1}^2 \frac{x'_i}{\sigma_i^2} W_{ij} z'_j\right) \right] d\mathbf{x}'}$$

Now we can rewrite using the distributivity of multiplication. The sum will simplify as a sum over a single latent variable, see the course for a more detailed explanation:

$$= \frac{\exp\left(-\sum_{i=1}^2 \frac{(x_i - b_i)^2}{2\sigma_i^2}\right) \prod_{j=1}^n \left[ \sum_{z_j} \exp\left(c_j z_j + \sum_{i=1}^2 \frac{x_i}{\sigma_i^2} W_{ij} z_j\right) \right]}{\int_{\mathbf{x}'} \exp\left(-\sum_{i=1}^2 \frac{(x'_i - b_i)^2}{2\sigma_i^2}\right) \prod_{j=1}^n \left[ \sum_{z'_j} \left(c_j z'_j + \sum_{i=1}^2 \frac{x'_i}{\sigma_i^2} W_{ij} z'_j\right) \right] d\mathbf{x}'}$$

The latent variable  $z_j$  can only take two values (0 and 1) in the sum in the numerator. When  $z_j = 0$ , we have  $\exp(0) = 1$ . By applying the same trick in the nominator and denominator, we have:

$$= \frac{\exp\left(-\sum_{i=1}^2 \frac{(x_i - b_i)^2}{2\sigma_i^2}\right) \prod_{j=1}^n \left[1 + \exp\left(c_j + \sum_{i=1}^2 \frac{x_i}{\sigma_i^2} W_{ij}\right)\right]}{\int_{\mathbf{x}'} \exp\left(-\sum_{i=1}^2 \frac{(x'_i - b_i)^2}{2\sigma_i^2}\right) \prod_{j=1}^n \left[1 + \exp\left(c_j + \sum_{i=1}^2 \frac{x'_i}{\sigma_i^2} W_{ij}\right)\right] d\mathbf{x}'}$$

Now observe that if we can take the exponential-log of the numerator and then use the fact that  $\log(ab) = \log a + \log b$  to change the product by a sum (moreover the log-exp in the first term then cancel - again, check the course for an explanation of this trick) we have:

$$\begin{aligned} &= \frac{\exp\left(-\sum_{i=1}^2 \frac{(x_i - b_i)^2}{2\sigma_i^2} + \sum_{j=1}^n \log\left[1 + \exp\left(c_j + \sum_{i=1}^2 \frac{x_i}{\sigma_i^2} W_{ij}\right)\right]\right)}{\int_{\mathbf{x}'} \exp\left(-\sum_{i=1}^2 \frac{(x'_i - b_i)^2}{2\sigma_i^2} + \sum_{j=1}^n \log\left[1 + \exp\left(c_j + \sum_{i=1}^2 \frac{x'_i}{\sigma_i^2} W_{ij}\right)\right]\right) d\mathbf{x}'} \\ &= \frac{\exp(-\mathcal{F}(\mathbf{x}; \theta))}{\int_{\mathbf{x}'} \exp(-\mathcal{F}(\mathbf{x}'; \theta)) d\mathbf{x}'} \end{aligned}$$

where  $\mathcal{F}(\mathbf{x}; \theta) = \sum_{i=1}^2 \frac{(x_i - b_i)^2}{2\sigma_i^2} - \sum_{j=1}^n \log\left[1 + \exp\left(c_j + \sum_{i=1}^2 \frac{x_i}{\sigma_i^2} W_{ij}\right)\right]$  is the free energy of  $\mathbf{x}$ . Note that  $\mathcal{F}(\mathbf{x}; \theta)$  can be computed efficiently because we transformed the original sum over an exponential number of values  $\sum_{\mathbf{z}}$  (sum over  $2^n$  possible assignation for  $\mathbf{z}$ !) as  $n$  sums. The marginalization in the denominator is still intractable, but during training we will approximate its contribution to the gradient via Monte Carlo estimation.

Let  $D = \{\mathbf{x}^{(1)} \dots \mathbf{x}^{(d)}\}$  be the training set. During training, we seek to maximize the log probability of the dataset:

$$\begin{aligned} \hat{\theta} &= \arg \max_{\theta} \sum_{i=1}^d \log p(\mathbf{x}^{(d)}; \theta) \\ &= \arg \min_{\theta} \sum_{i=1}^d \mathcal{L}(\theta, \mathbf{x}^{(d)}) \end{aligned}$$

where  $\mathcal{L}(\theta, \mathbf{x}) = -\log p(\mathbf{x}; \theta)$  is the loss function. The gradient of the loss function is defined as:

$$\begin{aligned} \nabla \mathcal{L}(\theta, \mathbf{x}) &= \nabla -\log p(\mathbf{x}; \theta) \\ &= \nabla -\log \frac{\exp(-\mathcal{F}(\mathbf{x}; \theta))}{\int_{\mathbf{x}'} \exp(-\mathcal{F}(\mathbf{x}'; \theta)) d\mathbf{x}'} \\ &= \nabla \mathcal{F}(\mathbf{x}; \theta) + \nabla \log \left( \int_{\mathbf{x}'} \exp(-\mathcal{F}(\mathbf{x}'; \theta)) d\mathbf{x}' \right) \\ &= \nabla \mathcal{F}(\mathbf{x}; \theta) + \frac{1}{\int_{\mathbf{x}'} \exp(-\mathcal{F}(\mathbf{x}'; \theta)) d\mathbf{x}'} \nabla \left( \int_{\mathbf{x}'} \exp(-\mathcal{F}(\mathbf{x}'; \theta)) d\mathbf{x}' \right) \end{aligned}$$

Similarly to sums, the gradient of an integral is the integral of the gradient:

$$\begin{aligned} &= \nabla \mathcal{F}(\mathbf{x}; \theta) + \frac{1}{\int_{\mathbf{x}'} \exp(-\mathcal{F}(\mathbf{x}'; \theta)) d\mathbf{x}'} \int_{\mathbf{x}'} \nabla \exp(-\mathcal{F}(\mathbf{x}'; \theta)) d\mathbf{x}' \\ &= \nabla \mathcal{F}(\mathbf{x}; \theta) + \frac{1}{\int_{\mathbf{x}'} \exp(-\mathcal{F}(\mathbf{x}'; \theta)) d\mathbf{x}'} \int_{\mathbf{x}'} \exp(-\mathcal{F}(\mathbf{x}'; \theta)) \nabla -\mathcal{F}(\mathbf{x}'; \theta) d\mathbf{x}' \\ &= \nabla \mathcal{F}(\mathbf{x}; \theta) + \int_{\mathbf{x}'} \frac{\exp(-\mathcal{F}(\mathbf{x}'; \theta))}{\int_{\mathbf{x}''} \exp(-\mathcal{F}(\mathbf{x}''; \theta)) d\mathbf{x}''} \nabla -\mathcal{F}(\mathbf{x}'; \theta) d\mathbf{x}' \\ &= \nabla \mathcal{F}(\mathbf{x}; \theta) + \int_{\mathbf{x}'} p(\mathbf{x}'; \theta) \nabla -\mathcal{F}(\mathbf{x}'; \theta) d\mathbf{x}' \\ &= \nabla \mathcal{F}(\mathbf{x}; \theta) - \mathbb{E}_{\mathbf{x}' \sim p(\mathbf{x}'; \theta)} [\nabla \mathcal{F}(\mathbf{x}'; \theta)] \end{aligned}$$

Computing the expectation exactly is intractable but we can rely on Monte Carlo estimation:

$$\simeq \nabla \mathcal{F}(\mathbf{x}; \theta) - \frac{1}{K} \sum_{k=1}^K \nabla \mathcal{F}(\tilde{\mathbf{x}}^{(k)}; \theta)$$

where  $\tilde{\mathbf{x}}^{(k)} \sim p(\tilde{\mathbf{x}}^{(k)}; \theta)$  are  $K$  samples from the model distribution. To sample from the distribution, we can rely on Gibbs sampling Markov Chain Monte Carlo as it is easy to sample from both conditional distributions. In practice, we will rely on the contrastive divergence objective, i.e. we take a single sample using a single step of a Markov chain initialized with the datapoint  $\mathbf{x}$ .