LCPB 24-25 exercise 2 (Restricted Boltzmann Machines, RBMs)

We want to study the performances of an RBM applied to some digits of the MNIST database. Choose the set of digits to analyze, such as the set (0,1,2) of the lecture. Of course, the more diverse are the digits in the set, the harder is the learning task.

1. Initialization of biases

Explain the initialization of visible biases proposed by Hinton.

2. Log-likelihood

Take an RBM with L=3 hidden units and compute the log-likelihood $\mathscr{L}$ of data during training (see the notes). Does it grow during training? Does it depend on the number of contrastive divergence steps? Then compare the log-likelihood $\mathscr{L}$ at the end of training for L=3, 4, 5, 6, … up to a maximum L of your choice.

3. Hyperparameters

Try RBMs with different hyperparameters. These variations *may* include: gradient descent type and parameters, SPINS=True or False, POTTS=True or False, regularization ($\gamma>0$)…
Do they improve the final log-likelihood $\mathscr{L}$?

Note 1

Again, keep into account that cross validation or any other indicatior of statistical fluctuations is useful to support statements quantitatively.

Note 2

The package **itertools** is useful for generating the set of possible hidden states, see https://docs.python.org/3/library/itertools.html

```
Q=4
import itertools as it
conf = it.product((0,1), repeat=Q)
all_conf=list(conf)
for z in all_conf:
    print(z)
```

4. OPTIONAL: hidden space activation

For a selected RBM, check which are the hidden units activated by each digit of the MNIST. Plot the probability distribution of each hidden state for each class (that is, label corresponding to the digit, e.g., 0,1,2 in the class example) in the MNIST. A hidden state is one out of 2^L possible states (or L states if one-hot encoding was selected). Is there any correlation between the activated hidden units?

5. OPTIONAL: energy barriers

We noted that the contrastive divergence (CD) algorithm does not switch hidden representations very much and continues to give back the same digit over and over. One may wonder which is the height of the energy barrier that separates the typical (x,z) energy of configurations when x is say the digit "1" from those typical for the digit "2".
Try inventing a modified CD method forcing the transition from say a "1" to a "2", and measuring the increase and decrease in energy of intermediate (x,z) configurations during the transition. Finding high intermediate energies should explain the observed quasi-ergodicity of the CD algorithm.