# Comments Lab 4

## 4.1

* There is convergence because the reconstruction loss (which represents an estimation of the gradient of log-likelihood function) decreases until convergence. The noise appears because we are using a stochastic gradient algorithm.
* If we use 500 and 200 the obtained result in terms of reconstruction loss is quite similar. This is maybe because, as we have many training samples that are so redundant [2], 200 hidden units are enough. The main difference is that with 200 hidden units it takes a little bit more time to converge (see graphs)
* If we see the weights to connect one hidden node to all visble nodes (25 random hidden nodes) we can see features learnt by them and how they are better learnt along epochs.