A restricted Boltzmann machine (RBM) was trained using 3 visible and *M* = 1,2,4,8 hidden neurons using the CD-k algorithm using *k* = 100, a learning rate of η = 0.1 and 1000 training steps. To compute the Kullback-Leibler divergence I iterated the dynamics of the RBM after training it using the parameters mentioned and counted the frequency at which the different patterns occur using 1000 number of samples (N\_outer) and updating 1000 (N\_inner) times. This was then done for all the different amount of hidden neurons *M* and the Kullback-Leibler (KL) divergence was then plotted as a function of the number of hidden neurons as can be seen in Figure 1. The KL divergence plotted in Figure 1 is an average of 50 independent trainings to understand what divergence is to be expected at that number of hidden neurons.

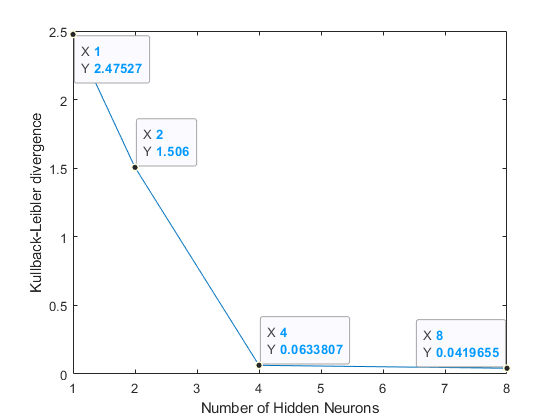


Figure 1. The KL divergence for different number of hidden neurons on an average of 50 trainings.

There is a clear switch from having 2 or 1 hidden neurons to having 4 or 8 which also makes sense since the theoretical limit to the XOR is 3 hidden neurons, and the network does not seem to improve that much compared with 8 hidden neurons. Although not tested an even higher number of hidden neurons might cause overfitting.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Nh | X1 | X2 | X3 | X4 | X5 | X6 | X7 | X8 |
| 1 | 0.1333 | 0.1349 | 0.0989 | 0.2007 | 0.0584 | 0.1131 | 0.1834 | 0.0773 |
| 2 | 0.2224 | 0.1927 | 0.1745 | 0.1822 | 0.0604 | 0.0753 | 0.0632 | 0.0292 |
| 4 | 0.2491 | 0.2674 | 0.2509 | 0.2314 | 0.0002 | 0.0002 | 0.0005 | 0.0003 |
| 8 | 0.2409 | 0.2391 | 0.2555 | 0.2632 | 0.0006 | 0.0002 | 0.0002 | 0.0002 |

Table 1. Percentage of patterns converged to the different patterns (x1-4 is the trained patterns), for *M* different number of hidden neurons.

As can be seen in Table 1 at 8 hidden neurons the patterns that was used as training (x1, x2, x3, x4) with a distribution of ¼=0.25 which they seem to diverge to and the other patterns x5, x6, x7, x8 to 0 which is to be expected. When calculating the average of the KL divergence I ignore the ones who blow up when a pb is 0 although this is very rare and should not affect the divergence too much.