Neural Information Processing: 2016-2017
Assignment 1, Notes to answers

10th April 2017

Linsker's Infomax

In this assignment we analyse ways to maximize information transmission in linear networks. We assume the two-dimensional input $\mathbf{u}$ is a Gaussian distribution signal with correlation matrix $Q = \begin{pmatrix} 1 & q \\ q & 1 \end{pmatrix}$ with $q = 1/2$. The output of the network is given by $\mathbf{v} = W.\mathbf{u} + \mathbf{n}$, where $\mathbf{n}$ is independent Gaussian noise with variance σ^2.

Question 1: We first analyse Linkers's approach (see lecture notes). Give an expression for $H(\mathbf{v})$. What is it in the limit $\sigma \to 0$? (5 points max)

- *Use that for Gaussian distributed variables,* $H(\boldsymbol{v}) = \frac{1}{2} \log \det \langle \boldsymbol{v}\boldsymbol{v}^T \rangle$*, and that* $\langle \boldsymbol{v}\boldsymbol{v}^T \rangle = WQW^T + \sigma^2 I$*, so that in the limit* $H(\boldsymbol{v}) = \log(|detW|) + \frac{1}{2} \log \det Q$.

- *Others took Q to mean the normalized correlation matrix. This gets tricky when the* $\langle \boldsymbol{v} \rangle \neq \mathbf{0}$*, but otherwise things are as above.*

We compare two weight constraints: 1) the weight matrix is constrained such that $w_{k1}^2 + w_{k2}^2 = 1$ for all k, or, 2) the elements are constrained as $|w_{ij}| \leq 1$.

Question 2: Maximize the mutual information under either constraint in the limit of zero noise. What invariances do you encounter? (5 points max)

- *From Question 1 we see that we need to maximize* $|w_{11}w_{22} - w_{12}w_{21}|$*, subject to those constraints.*

- *First constraint: The optimization can for instance be done by using Lagrange multipliers, one finds that the two columns of W have to be orthogonal (and unit length), this is basically a whitening solution; any mixing reduces entropy. Another way is to check that* $W = I$ *maximizes entropy, but because only* $|\det W|$ *matters, any orthogonal transformation will also do.*

- *Second constraint: Now the maximum is attained for* $W = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$ *and any sign permutation that preserves* $|\det W|$*. Note, here directly using Lagrange multipliers won't work for this case, you can use 'slack' variables.*

- *These invariances make sense: it should not matter if we swap neuron1 with neuron2, nor should the sign of linear transformation that they do matter.*

Question 3: To study the case of non-zero noise, maximize the information numerically with respect to the weight matrix. Matlab (but not octave) has the function `fmincon()`. Plot the result as a function of σ. Why is numerical optimization tricky? (5 points max)

- *To do the numerical optimization it is best to take different starting values. Otherwise, you can not see whether there are local solutions. A good way is to take random starting values and slowly vary σ. If the obtained information as a function of σ is a smooth curve it is a case of degeneracy, if it jagged (e.g. jumping between two values) there are local minima.*

- *For the 1st constraint, the orthogonal symmetry in the solutions found in question 2 disappears as soon as $\sigma \neq 0$.*

- *For the 1st constraint, you should find a smooth transition from the solution $W = I$ to $W = 1/\sqrt{2}\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$ (and permutations). While the solution is degenerate (multiple solutions leading to the same information), there are no local minima. Nevertheless, fmincon might get stuck in the second solution when it is initialized there.*

- *For the 2nd constraint, there is a sudden transition from $\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$ to $\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$ at large σ. The transition is at a similar value of σ. Here there are local minima, as can be seen by given different initial conditions for a given σ.*

- *It does not make sense to optimize without any constraint; the weights will diverge.*

- *Note that fmincon can give a warning about local solutions, but that does not mean that there are local solutions unequal to the global one.*

- *The numerical optimization is made harder by the symmetries in the optimal solutions. It makes therefore sense to sort the solution. In the limit of zero noise we found above even a continuous symmetry. An alternative would be to explicitly eliminate two of the weights (e.g. $w_{12} = \pm\sqrt{(1 - w_{11}^2}$, or by using an angle to parametrize the weight vector $(\cos(a), \sin(a))$) so that only an unconstrained optimization is left, which is easier to deal with.*

Question 4: Examine the optimal weight matrix in the case of 3 inputs and 3 outputs, and $Q_{ij} = \delta_{ij} + (1 - \delta_{ij})q$ for either constraint. (5 points max)

- *In general the information is higher for 3x3 than for 2x2 , and again decreases with increasing noise.*

- *Again when exploring numerically, one should check the role of the initial conditions, ideally including random ones.*

- *For the 1st constraint, essentially the same picture emerges as above, but the solutions are much more degenerate. (I think there are still not local minima, but I am not 100% sure).*

- *For the 2nd constraint, one finds again extremal solutions and that the determinant for small σ is 4, and . Again there are extremal solutions ($|w_{ij}| = 1$) with local minima.*

- *Note that one needs $q \geq -1/2$ for Q to be a valid correlation matrix.*