Exercises for week 2
Hebbian rules and ICA

Exercise 1: Optimality criteria for PCA: variance and optimal reconstruction

In lecture 1 we found that PCA is a result of Hebbian learning. We now ask whether PCA (and Hebbian learning rules!) can be derived from optimality criteria.

As usual we have a set of P input patterns $\mathbf{x}^\mu$ with $1 \leq \mu \leq P$. We assume that $E_{data}[\mathbf{x}] = 0$.

A (i) Define an output $y = \sum_k w_k x_k$.

Derive the batch update rule then the single-sample update rule to maximize the variance $E_{data}[y^2]$ by gradient ascent.

(ii) Turn the update equation of gradient ascent into a differential equation. Compare your result to the equation we found in Lecture 1. want component-wise or matrix?

(iii) Assume now that the weight vector is normalized, i.e., we maximize variance for a normalized vector. Express the weight vector in terms of the Eigenvectors of the correlation matrix. Convince yourself that the variance is maximal if the only nonzero component is the projection on the first Eigenvector.

(iv) Go back to point (i) and switch to a presentation in terms of the vector component. Intepret the result as a Hebbian learning rule and identify the presynaptic and postsynaptic terms.

B. The aim of an autoencoder is to compress a set high-dimensional data points into a low-dimensional representation such that a reconstruction of the input is possible at minimal loss. Assume a linear autoencoder consisting of one hidden layer of a single neuron $y = \sum_k w_k x_k$. The weights from the hidden layer to the output are $w_k^{\text{out}} = w_k$.

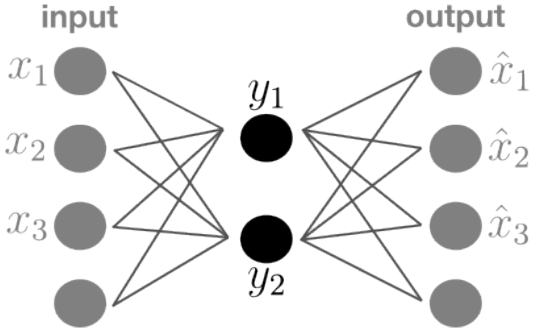

Figure 1: Architecture of an autoencoder with one hidden layer of two hidden neurons.

(i) Minimize the reconstruction error

$\frac{1}{P}\sum_\mu [\mathbf{x}^\mu - \hat{\mathbf{x}}^\mu]^2$

where $\hat{x}_k^\mu = w_k^{\text{out}} y$.

Derive first two separate batch rules, one for the output weights and one for the input weights.

(ii) Then turn to a presentation in terms of the vector component. Intepret the result as a Hebbian learning rule and identify the presynaptic and postsynaptic terms. What is the difference between the two rules? How are they related to the Oja rule?

(iii) Repeat the same calculation but assuming that there are two neurons in the hidden layer. Interpret the resulting online-rule as an interaction between the two hidden neuron. What is this interaction?

(iv) Think about the relevance of these results for the interpretation of the learning rule. Would you agree with the following statement: “Hebbian learning rules are able to find ’good representations’, in the sense that an optimal reconstruction of the stimulus WOULD be possible (even if we do not implement the reconstruction process)'.

Exercise 2: Some important fun facts about independence and correlation

2.1 Prove that if x_1 and x_2 are statistically independent, they are also uncorrelated:

$$p(x_1, x_2) = p_1(x_1)p_2(x_2) \Rightarrow \langle (x_1 - \langle x_1 \rangle)(x_2 - \langle x_2 \rangle) \rangle = 0.$$

2.2 Prove: Given two functions h_1, h_2 and two independent random variables x_1 and x_2, the expectation value of the product of h_1 and h_2 factorizes in the product of the expectation values :

$$p(x_1, x_2) = p_1(x_1)p_2(x_2) \;\Rightarrow\; E\{h_1(x_1)h_2(x_2)\} = E\{h_1(x_1)\}E\{h_2(x_2)\}$$

2.3 For N-dimensional data, the Gaussian distribution has the form:

$$p(\vec{x}) = \frac{1}{\sqrt{(2\pi)^N det(C)}} \exp\left(-\frac{1}{2}(\vec{x} - \vec{\mu})^T C^{-1}(\vec{x} - \vec{\mu})\right) ,$$

where and $\vec{\mu}$ is the mean of the data, C their covariance matrix, and $det(C)$ its determinant.

Suppose that C has elements $C_{ij} \neq 0$ for all i, j indicating that the variables x_i and x_j are correlated. First, convince yourself (without calculation) that after transformation to the coordinate system of Eigenvectors, the new coordinates $\tilde{x}$ are uncorrelated even if the Eigenvalues are not identical $\lambda_n \neq \lambda_m$ for all n, m.

Show with a short calculation (one line) that the variables are not just uncorrelated but also statistically independent. Hence, what you need to show is:

$$\forall i, j : \langle (x_i - \langle x_i \rangle)(x_j - \langle x_j \rangle) \rangle = 0 \;\Rightarrow\; p(\vec{x}) = \prod_i p_i(x_i)\,.$$

2.4 Now assume that the data is whitened (each component has zero mean and unit variance, and the components are pairwise decorrelated) so that $\lambda_n = \lambda_m$ for all n, m. Show that any rotation $\{\vec{y}^\mu = R\vec{x}^\mu\}$ of the data is also whitened.

Hint: R is a rotation matrix, iff $RR^T = E$, E being the identity matrix.

Exercise 3: ICA as Hebbian Learning

Consider an ICA algorithm that aims at maximizing $J(\vec{w}) = \langle F(y) \rangle$, where $y = \vec{w}^T\vec{x}$ and $F(y) = \frac{1}{a}\log\cosh(ay)$. The maximization is done by gradient ascent.

3.1 Show that: $\frac{dF}{dy} = \tanh(ay)$.

3.2 Calculate $\frac{dF}{dw_j}$ for $y = \sum w_k x_k$.

3.3 Show that a gradient ascent on $J(\vec{w}) = \langle F(\vec{w}^T\vec{x}) \rangle$ leads to a Hebbian rule.
(Hint: Make the transition from a batch rule to an online rule).

Exercise 4: A few fun facts on Kurtosis

Students who do not want to do the calculation can make use of the statements in 4.1-4.3 as a table of results.

Variance is defined as $var(x) = E\{x^2\} - E\{x\}^2$, kurtosis as $\kappa(x) = E\{x^4\} - 3(E\{x^2\})^2$.

For each of the following distributions, calculate the variance and prove the given kurtosis:

4.1 the Gaussian distribution, with kurtosis $\kappa = 0$:

$$p(x) = \sqrt{\frac{a}{\pi}} \exp\left(-ax^2\right) .$$

Hint: $x^2 \exp(-ax^2) = -\frac{d}{da} \exp(-ax^2)$.

4.2 the uniform distribution, with kurtosis $\kappa = -\frac{6}{5}$:

$$p(x) = \begin{cases} \frac{1}{2\sqrt{3}} & \text{if } |x| \leq \sqrt{3} \\ 0 & \text{otherwise.} \end{cases}$$

4.3 the exponential distribution (Laplace distribution), with kurtosis $\kappa = 3$:

$$p(x) = \frac{1}{\sqrt{2}} \exp(-\sqrt{2}|x|).$$

4.4 In the above examples, do distributions with 'longer tails' than the Gaussian yield smaller or larger kurtosis? Do you think that this observation about 'tails' and kurtosis can be transformed into a general statement?

Exercise 5: Kurtosis maximization

Remember that the kurtosis is defined as $\kappa(x) = E[x^4] - 3E[x^2]^2$. Suppose that we have two independent variables x_1 and x_2 both of zero mean, but we measure some arbitrary mixture. In class we have argued in a hand-waving fashion that Kurtosis is maximal (or sometimes minimal) if the direction of projection yields one of the independent variables. In this exercise we have a special case, where we can explicitly show this. We mix two variables of known kurtosis, and then apply a projection in an arbitrary direction which gives a variable y. For this mixed variable y we maximize kurtosis. The calculation is a bit lengthy, but for those of you who have doubts why ICA works, it may provide useful insights. Here are the steps of the calculation:

5.1 Show that the kurtosis of $y = x_1 + x_2$ is given by $\kappa(y) = \kappa(x_1) + \kappa(x_2)$.

5.2 Show that the kurtosis of $y = \alpha x$ is given by $\kappa(y) = \alpha^4 \kappa(x)$.

5.3 Use 1 and 2 to show that the kurtosis of $y = \sqrt{a} x_1 + \sqrt{1-a} x_2$, $a \in [0, 1]$, is given by

$$\kappa(y) = a^2 \kappa(x_1) + (1-a)^2 \kappa(x_2) .$$

5.4 Let $\kappa(x_1) = c$ and $\kappa(x_2) = d$ be the kurtosis of x_1 and x_2. Assume that both signals are super-Gaussian and that $0 < c < d$. Show that the kurtosis of the mixture $y = \sqrt{a} x_1 + \sqrt{1-a} x_2$ has maxima for $a = 0$ and $a = 1$, and that $a = 0$ is the global maximum.

5.5 Which value(s) of a maximize the kurtosis if the signals x_1 and x_2 are sub-Gaussian: $c < d < 0$?