ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Unsupervised and reinforcement learning in neural networks
Week 4

Lecturer: Marc-Oliver Gewaltig

Assistants: Dane Corneil
Berat Denizdurduran

Independent Component Analysis

Exercise 1

In class, it was argued that a mixture of statistically independent sources tends to be more Gaussian that the sources themselves. This argument served as the basis for ICA algorithms that rely on non-Gaussianity. In this exercise, we want you to show that the non-Gaussianity argument does not rely on the summation of a large number of statistically independent sources, but that it works already for two sources.
Remember that the kurtosis is defined as $\kappa(x) = E[x^4] - 3E[x^2]^2$. Now let x_1 and x_2 be statistically independent and let both have zero mean.

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

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

1.3 Use 1.1 and 1.2 to show that the kurtosis of $y = \sqrt{a}x_1 + \sqrt{1-a}x_2$, $a \in [0, 1]$, is given by

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

1.4 Let $\kappa(x_1) = c$ and $\kappa(x_2) = d$ be the kurtoses 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.

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

Exercise 2

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.

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

2.2 Calculate $\frac{dF}{dw_j}$ for $y = \sum_k w_k x_k$.

2.3 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 3

In the previous exercise, we discussed a simple ICA algorithm based on gradient ascent. Here, we will go one step further and maximize the non-Gaussianity of the mixture using the Newton method, that yields a faster convergence. The resulting learning algorithm is known as *fastICA*.

3.1 We want to maximize the measure of non-Gaussianity F under the constraint of a normalized weight vector, i. e. $\vec{w}^T\vec{w} = 1$. This corresponds to finding the maximum of the function $J(\vec{w}) = \langle F(\vec{w}^T\vec{x})\rangle$. Derive the Taylor expansion $J^*(\vec{w})$ of $J(\vec{w})$ around $\vec{w_0}$ up to second order in $\vec{w}$.

3.2 A Newton step consists of setting the next value $\vec{w}_{new}$ to the vector that maximizes the second-order approximation J^* around the previous weight vector $\vec{w_0}$. Show that this leads to the *fastICA* update rule:

$$\vec{w}_{new} = \langle g(\vec{w_0}^T\vec{x})\vec{x}\rangle - \langle g'(\vec{w_0}^T\vec{x})\rangle\vec{w_0},$$

with $g := \frac{dJ(y)}{dy}$ and $g' = \frac{dg(y)}{dy}$.

(**Hint:** Make the approximation that $\langle g'(\vec{w_0}^T\vec{x})\vec{x}\vec{x}^T\rangle \approx \langle g'(\vec{w_0}^T\vec{x})\rangle \cdot \langle\vec{x}\vec{x}^T\rangle$ and exploit the fact that the data is pre-whitened, $\langle\vec{x}\vec{x}^T\rangle = E$ with identity matrix E. Finally, remember that the weight vector gets re-normalized to unity in the *fastICA* algorithm after the above update rule is applied.)

Exercise 4

In this exercise we see that PCA extracts the direction of maximum variance: For a zero-mean data set $D = \{x_1, ..., x_N\}$, try to find a direction $\vec{w}$ for which the variance of the projected elements $\vec{y} = \vec{w}^T\vec{x}$ has maximum variance. You should write an optimization problem with constraint $\|w\| = 1$ where $\|.\|$ denotes the vector norm and use Lagrange multiplier method to solve it.

Hint: In mathematical optimization, the method of Lagrange multipliers provides a strategy for finding the local maxima and minima of a function subject to equality constraints. Consider the optimization problem: maximize $f(w)$ subject to $g(w) = c$. We need both f, g to have continuous first partial derivatives. In this method, a new variable λ called a Lagrange multiplier is introduced and the Lagrange function defined by $\mathcal{L}(w, \lambda) = f(w) - \lambda(g(w) - c)$ is studied. If w_0 is a maximizer of $f(w)$ for the original constraint problem, then there exists λ_0 such that (w_0, λ_0) is a stationary point for the Lagrange function $\mathcal{L}$. Stationary points are those points where the partial derivatives of $\mathcal{L}$ are zero.

Exercise 5

Assume that a set of signals y_i^t are statistically independent (and have zero mean) and that consequently, their time-delayed covariance matrix is diagonal: $C = \langle y_i^t(y_j^{t-\tau})^T\rangle = \lambda_i(\tau)\delta_{ij}$. Show that for any matrix R, the time-delayed covariance matrix C^* of the signals $\vec{x}^t = R^T\vec{y}^t$ is symmetric.