UML2012
Exercise set 2
Solutions to be presented in the 27.3.2012 session

**Exercise 1:**
If the function $J$ maps a matrix $W \in \mathbb{R}^{n\times m}$ to $\mathbb{R}$, the gradient is defined as

$$\nabla J(W) = \begin{pmatrix} \frac{\partial J(W)}{\partial W_{11}} & \cdots & \frac{\partial J(W)}{\partial W_{1m}} \\ \vdots & \ddots & \vdots \\ \frac{\partial J(W)}{\partial W_{n1}} & \cdots & \frac{\partial J(W)}{\partial W_{nm}} \end{pmatrix}. \tag{1}$$

Alternatively, it is defined to be the matrix $\nabla J(W)$ such that

$$J(W + \varepsilon \mathbf{e}_i \mathbf{e}_j^T) = J(W) + \varepsilon \mathbf{e}_i^T \nabla J(W) \mathbf{e}_j + o(\varepsilon), \qquad (\varepsilon \to 0) \tag{2}$$

Here, $\mathbf{e}_i$ is a vertical *column* vector which is everywhere zero but in slot $i$ where it is 1. Number of elements in $\mathbf{e}_i$ and $\mathbf{e}_j$ depend on where they are used. Here $\mathbf{e}_i\mathbf{e}_j^T$ is a $(n \times m)$-matrix, where we assume $1 \leq i \leq n$, $1 \leq j \leq m$, and $\mathbf{e}_i \in \mathbb{R}^{n\times 1}$, $\mathbf{e}_j^T \in \mathbb{R}^{1\times m}$.

Sometimes, the second definition is more convenient because you can avoid multiplying out matrices. Use either of the two definitions to find $\nabla J(W)$ in the following cases (here: $\mathbf{u} \in \mathbb{R}^n, \mathbf{v} \in \mathbb{R}^m, A \in \mathbb{R}^{n\times m}, f : \mathbb{R} \to \mathbb{R}$).

$$\begin{aligned} &1. \quad J(W) = \mathbf{u}^T W \mathbf{v} \\ &2. \quad J(W) = \mathbf{u}^T (W + A) \mathbf{v} \\ &3. \quad J(W) = \sum_{n'=1}^{n} f(W_{n'*}\mathbf{v}), \quad \text{where } W_{n'*} \text{ are the horizontal rows} \\ &4. \quad J(W) = \mathbf{u}^T W^{-1} \mathbf{v}, \quad \text{where we assume } n = m. \end{aligned} \tag{3}$$

Hint: Prove $(W + \varepsilon H)^{-1} = W^{-1} - \varepsilon W^{-1} H W^{-1} + O(\varepsilon^2)$ first.

**Exercise 2:**
In this exercise, we calculate the gradient of

$$J(W) = \log|\det(W)| \tag{4}$$

using what we learned in previous exercise sessions.

**2.1** Assume that $\mathbf{u}_1, \ldots, \mathbf{u}_N$ are linearly independent eigenvectors of $W$, and form a matrix $U = (\mathbf{u}_1, \ldots, \mathbf{u}_N)$. Let's then define $V = (U^{-1})^T$, and define vectors

$\mathbf{v}_1, \ldots, \mathbf{v}_N$ so that $V = (\mathbf{v}_1, \ldots, \mathbf{v}_N)$. Show that the eigenvalues $\lambda_n$ can be written by formula

$$\lambda_n = \mathbf{v}_n^T W \mathbf{u}_n \tag{5}$$

**2.2** We can consider $\lambda_n$, $\mathbf{u}_n$ and $\mathbf{v}_n$ to be functions of $W$, so that they can be written as $\lambda_n(W)$, $\mathbf{u}_n(W)$ and $\mathbf{v}_n(W)$. Calculate a formula for $\nabla\lambda_n(W)$, by substituting $\lambda_n(W) = \mathbf{v}_n(W)^T W \mathbf{u}_n(W)$. You can assume that $\mathbf{u}_n(W)$ and $\mathbf{v}_n(W)$ are differentiable functions of $W$.

**2.3** Use your formula for $\nabla\lambda_n(W)$ to obtain a formula for $\nabla J(W)$, by first writing $J(W)$ in terms of the eigenvalues $\lambda_n(W)$.

**2.4** Show that

$$\nabla J(W) = (W^{-1})^T \qquad \left(\text{meaning } \frac{\partial J(W)}{\partial W_{nn'}} = (W^{-1})_{n'n}\right) \tag{6}$$

You should also recall results from the last week's exercise set.

**Exercise 3:**

A Gaussian random vector $\mathbf{x}$ of dimension $m$ has the density

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{m/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})\right], \tag{7}$$

where $\Sigma$ is the covariance matrix and $\boldsymbol{\mu}$ is the mean.

**3.1** Given independently distributed data $\mathbf{x}_1, \ldots, \mathbf{x}_N$ where each sample $\mathbf{x}_k$ is a Gaussian random vector with mean $\boldsymbol{\mu}$ and covariance matrix $\Sigma$, formulate the log-likelihood $\ell(\boldsymbol{\mu}, \Sigma)$.

**3.2** Calculate the gradient of $\ell(\boldsymbol{\mu}, \Sigma)$ with respect to $\boldsymbol{\mu}$ and $\Sigma$. You can use the results from the exercises 1 and 2.

**3.3** Conclude that the ML estimate for $\boldsymbol{\mu}$ is the sample mean

$$\hat{\boldsymbol{\mu}} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}_n, \tag{8}$$

and that the ML estimate for $\Sigma$ is the sample covariance matrix

$$\hat{\Sigma} = \frac{1}{N} \sum_{n=1}^{N} (\mathbf{x}_n - \bar{\mathbf{x}})(\mathbf{x}_n - \bar{\mathbf{x}})^T, \tag{9}$$

where $\bar{\mathbf{x}} = \hat{\boldsymbol{\mu}}$.

**Exercise 4:**

**4.1** Let $\mathbf{w} \in \mathbb{R}^n$ be a vector which reduces the dimension of $\mathbf{x} \in \mathbb{R}^n$ from $n$ to one via $z = \mathbf{w}^T\mathbf{x} = \sum_i w_i x_i$. Assume also that you want to reconstruct $\mathbf{x}$ from $z$ via

$$\hat{\mathbf{x}} = z\mathbf{w}. \tag{10}$$

In this exercise, we will show that taking for $\mathbf{w}$ the first principal component is the optimal way of reducing the dimension if the optimality criterion is the squared reconstruction error $J$

$$J(\mathbf{w}) = \mathrm{E}(||\mathbf{x} - \hat{\mathbf{x}}||^2) = \mathrm{E}\left(\sum_j (x_j - w_j z)^2\right). \tag{11}$$

Here $z$ and $\hat{\mathbf{x}}$ are considered to be functions of $\mathbf{w}$, so that $z = z(\mathbf{w})$ and $\hat{\mathbf{x}} = \hat{\mathbf{x}}(\mathbf{w})$.

Prove that minimizing $J(\mathbf{w})$ with constraint $\|\mathbf{w}\| = 1$ is equivalent to maximizing $\mathbf{w}^T E(\mathbf{x}\mathbf{x}^T)\mathbf{w}$ (with the same constraint).

**4.2** Assume that $\lambda_1 > \cdots > \lambda_n$ are some fixed numbers, and that parameters $m_1, \ldots, m_n$ must satisfy $m_{n'} \geq 0$ and $m_1 + \cdots + m_n = 1$. Prove, that under these constraints, the quantity

$$\lambda_1 m_1 + \cdots + \lambda_n m_n \tag{12}$$

is maximized by choosing $(m_1, m_2, \ldots, m_n) = (1, 0, \ldots, 0)$.

Advice: You want to prove, that if $(m_1, m_2, \ldots, m_n) \neq (1, 0, \ldots, 0)$ (antithesis), then

$$\lambda_1 > \lambda_1 m_1 + \cdots + \lambda_n m_n. \tag{13}$$

This is equivalent to

$$\lambda_1 > \lambda_2 \frac{m_2}{1 - m_1} + \cdots + \lambda_n \frac{m_n}{1 - m_1}. \tag{14}$$

You can use induction step from here.

**4.3** Assume that $A$ is some real symmetric matrix with distinct eigenvalues. Prove that if the vector $\mathbf{w}$ must satisfy $\|\mathbf{w}\| = 1$, the quantity $\mathbf{w}^T A\mathbf{w}$ will be maximized with respect to $\mathbf{w}$ precisely when $\mathbf{w}$ is an eigenvector of $A$ corresponding to the largest eigenvalue.

**4.4** Deduce that minimizing $J(\mathbf{w})$ with constraint $\|\mathbf{w}\| = 1$ is equivalent to finding an eigenvector of $E(\mathbf{x}\mathbf{x}^T)$ with the largest eigenvalue.