Philipp Lintl
12152498
lintl.philipp@gmail.com

Homework Assignment 1
Machine Learning 1, 18/19

2018-11-12

1 Basic Linear Algebra and Derivatives

1.1

Given $\mathbf{A} = \begin{bmatrix} 2 & 3 & 4 \\ 3 & 4 & 5 \\ 4 & 5 & 6 \end{bmatrix}$ and $\mathbf{B} = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{bmatrix}$ and $\mathbf{b} = \begin{bmatrix} 2 & 3 & 4 \end{bmatrix}^T$

(a) Compute **AB**:

$$\mathbf{AB} = \begin{bmatrix} 2 & 3 & 4 \\ 3 & 4 & 5 \\ 4 & 5 & 6 \end{bmatrix} \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 2+4 & 2+3 & 3+4 \\ 3+5 & 3+4 & 4+5 \\ 4+6 & 4+5 & 5+6 \end{bmatrix} = \begin{bmatrix} 6 & 5 & 7 \\ 8 & 7 & 9 \\ 10 & 9 & 11 \end{bmatrix}$$

(b) Are **A** and **B** invertible? How do you test for it? If so, calculate the inverses. Provide details oly for your analysis of **A**.

A matrix **A** is invertible, if the row (respectively column) vectors are independent of each other. That means, that they can not be exposed as a linear combination of the others. A quick and easy way to test for invertibility is enabled by calculating the determinant of the respective matrix. According to $det(X) = 0 \rightarrow$ X is not invertible. As asked in the task description, only A requires detailed analysis. Therefore, the determimant is calculated first:

$$\begin{aligned} det(\mathbf{A}) &= 2 \cdot 4 \cdot 6 + 3 \cdot 5 \cdot 4 + 4 \cdot 3 \cdot 5 - (4 \cdot 4 \cdot 4 + 5 \cdot 5 \cdot 2 + 6 \cdot 3 \cdot 3) \\ &= 48 + 60 + 60 - 64 - 50 - 54 = 0 \end{aligned}$$

The calculation is obtained via a trick which specifically for 3×3 matrices works as follows (Cramers rule would also have worked):

The matric **A** is extended by its first two columns on the right. Now, its elements can be multiplied diagonally. So, now an extended 3×5 matrix with rows denoted by $i = 1, \ldots, 3$, columns by $j = 1, \ldots, 5$ and entries by A_{ij} is subject to calculation.

$$\begin{bmatrix} 2 & 3 & 4 & 2 & 3 \\ 3 & 4 & 5 & 3 & 4 \\ 4 & 5 & 6 & 4 & 5 \end{bmatrix}$$

The first multiplication consists of terms $A_{11} \cdot A_{22} \cdot A_{33} \cdot$ which is summed with $A_{12} \cdot A_{23} \cdot A_{34} \cdot$ and $A_{13} \cdot A_{24} \cdot A_{35} \cdot$. Then three multiplied terms $A_{31} \cdot A_{22} \cdot A_{13} \cdot$, $A_{32} \cdot A_{23} \cdot A_{14} \cdot$ and $A_{33} \cdot A_{24} \cdot A_{15} \cdot$ are subtracted.

As seen, $det(\mathbf{A}) = 0$ yielding that **A** is in fact not invertible.

B on the other hand is invertible, as $det(\mathbf{B}) = 2$. $\mathbf{B}^{-1}$ now is obtained with the Gauss-Jordan elimination. Shortely described, the matrix to be inverted is given on the left and the identity matrix **I** is given on the right. Then, **B** is transformed via rowwise (or columnwise) subtractions among the rows (or columns), such that the Identity matrix **I** is gotten on the left. The respective inverse $\mathbf{B}^{-1}$ is then given on the right, provided all the transformations are conducted in the same fashion. For B this emerges as:

$$\left[\begin{array}{ccc|ccc} 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \end{array}\right] \xrightarrow[III-I]{} \left[\begin{array}{ccc|ccc} 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & -1 & 1 & -1 & 0 & 1 \end{array}\right] \xrightarrow[III+II]{} \left[\begin{array}{ccc|ccc} 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 2 & -1 & 1 & 1 \end{array}\right] \xrightarrow[\frac{III}{2} \; and \; II-III]{}$$

$$\left[\begin{array}{ccc|ccc} 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} \\ 0 & 0 & 1 & -\frac{1}{2} & \frac{1}{2} & \frac{1}{2} \end{array}\right] \xrightarrow[I-II]{} \left[\begin{array}{ccc|ccc} 1 & 0 & 0 & \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} \\ 0 & 1 & 0 & \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} \\ 0 & 0 & 1 & -\frac{1}{2} & \frac{1}{2} & \frac{1}{2} \end{array}\right]$$

Thus,

$$\mathbf{B}^{-1} = \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} & \frac{1}{2} \end{bmatrix} = \frac{1}{2} \cdot \begin{bmatrix} 1 & -1 & 1 \\ 1 & 1 & -1 \\ -1 & 1 & 1 \end{bmatrix}$$

To test the result, $\mathbf{B}\mathbf{B}^{-1} = \mathbf{I}$ must hold.

$$\mathbf{B}\mathbf{B}^{-1} = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{bmatrix} \frac{1}{2} \cdot \begin{bmatrix} 1 & -1 & 1 \\ 1 & 1 & -1 \\ -1 & 1 & 1 \end{bmatrix} = \frac{1}{2} \begin{bmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \mathbf{I}$$

(c) Is **AB** invertible? Why (not)?

No, it is not invertible, as a calculation rule regarding determinants shows:

$$det(\mathbf{AB}) = \underbrace{det(\mathbf{A})}_{=0} \cdot det(\mathbf{B}) = 0$$

So, det(**AB**) is zero which means that **AB** is not invertible.

(d) Compare the solution set for the systems $\mathbf{Bx} = \mathbf{b}$ and $\mathbf{Ax} = \mathbf{b}$. What can we say about the second system due to the invertibility property you determined previously?

For $\mathbf{Bx} = \mathbf{b}$, the solution is obtained once again by Gaussian elimination:

$$\left[\begin{array}{ccc|c} 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 2 \\ 1 & 0 & 1 & 5 \end{array}\right] \underset{III-I}{\longrightarrow} \left[\begin{array}{ccc|c} 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 2 \\ 0 & -1 & 1 & 4 \end{array}\right] \underset{III+II \text{ and } \frac{III}{2}}{\longrightarrow} \left[\begin{array}{ccc|c} 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 2 \\ 0 & 0 & 1 & 3 \end{array}\right]$$

$\rightarrow x_3 = 3$, $x_2 + x_3 = 2 \rightarrow x_2 = -1$ and $x_1 + x_2 = 1 \rightarrow x_1 = 2$, which yields $x = \begin{bmatrix} 2 \\ -1 \\ 3 \end{bmatrix}$. Once again testes by

$$\mathbf{Bx} = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 2 \\ -1 \\ 3 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 5 \end{bmatrix} = \mathbf{b},$$

so **Bx=b** holds.
For the second system, we dont even need to start calculating, as non invertibility of **A** implies, that the column/row vectors are not independent and thus at least one row/column can be represented as 0's or in other words can be spared. That way, the equation system is not solvable anymore, as no or infinitely many solutions exist.

1.2

Find the derivative of the folowing functions with respect to x.

(a) $f_1(x) = \frac{2}{x^2} + x^{-7} + x^3$

$$\frac{\partial}{\partial x} f_1(x) = -\frac{4}{x^3} - 7x^{-8} + 3x^2$$

(b) $f_2(x) = xe^{-\sqrt[5]{x}}$

$$\begin{aligned} \frac{\partial}{\partial x} f_2(x) &= xe^{-\sqrt[5]{x}} \left(-\frac{1}{5} x^{-\frac{4}{5}}\right) + e^{-\sqrt[5]{x}} \\ &= e^{-\sqrt[5]{x}} \left(-\frac{1}{5} x^{\frac{1}{5}} + 1\right) \\ &= e^{-\sqrt[5]{x}} \left(-\frac{1}{5} \sqrt[5]{x} + 1\right) \end{aligned}$$

(c) $f_3(x) = \frac{1}{x} + ln(x^2)$

$$\begin{aligned}\frac{\partial}{\partial x} f_3(x) &= -\frac{1}{x^2} + \frac{1}{x^2} \cdot 2x \\ &= \frac{2x-1}{x^2}\end{aligned}$$

(d) $\sigma(x) = \frac{1}{1+e^{-x}}$

$$\begin{aligned}\frac{\partial}{\partial x}\sigma(x) &= -\left(1+e^{-x}\right)^{-2} \cdot \left(e^{-x} \cdot -1\right) \\ &= \frac{e^{-x}}{\left(1+e^{-x}\right)^2}\end{aligned}$$

(e) $f_4(x) = \max\{0, x\}$

$$\frac{\partial}{\partial x} f_4(x) = \begin{cases} 0, & \text{x <= 0.} \\ 1, & \text{x > 0.} \end{cases}$$

What is the shape of the following gradients:

(g) $\frac{\partial f(x)}{\partial x}$ with $f : \mathbb{R} \to \mathbb{R}, x \in \mathbb{R}$

As $f(x)$ is a one-dimensional, reel function, this gradient then has the shape of a 1×1 vector, which in this case is a scalar made of the derivative of f(x) according to x.

(h) $\frac{df(\mathbf{x})}{d\mathbf{x}}$ with $f : \mathbb{R}^n \to \mathbb{R}, x \in \mathbb{R}^n$

Now, **x** is a n-dimensional vector which is mapped on a scalar by f and can be represented as such $f(x_1, \ldots, x_n)$. As specified, the gradient is defined as a row vector, which affects the shape in the following way:

$$\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = \left[\frac{\partial f(\mathbf{x})}{\partial x_1}, \ldots, \frac{\partial f(\mathbf{x})}{\partial x_n}\right]$$

Thus, the gradient for this function is a n-dimensional row vector consisting of the partial derivatives with respect to $x_1, \ldots, x_n$ in the according order.

(i) $\frac{\partial \mathbf{f}(\mathbf{x})}{\partial \mathbf{x}}$ with $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^m, x \in \mathbb{R}^n$

Now, **x** is a n-dimensional vector which maps on a m-dimensional vector, meaning that there are m functions involved in the mapping. $f(x_1, \ldots, x_n) = (f_1(x_1, \ldots, x_n), \ldots, f_m(x_1, \ldots, x_m))$. The resulting gradient has the following shape:

$$\frac{\partial \mathbf{f}(\mathbf{x})}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial f_1(\mathbf{x})}{\partial x_1} & \cdots & \frac{\partial f_1(\mathbf{x})}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m(\mathbf{x})}{\partial x_1} & \cdots & \frac{\partial f_m(\mathbf{x})}{\partial x_n} \end{bmatrix}$$

Thus, the gradient for this function is a $m \times n$ matrix, which has the partial derivatives according to all x_j with j $\in \{1, \ldots, n\}$ of each f_i with i $\in \{1, \ldots, m\}$ in its rows.

FInd the gradient of the folowing functions. Make their shapes explicit.

(j) $\frac{\partial(\mathbf{x})}{\partial \mathbf{x}}$ with $f(\mathbf{x}) = 2exp\left(x_2 - ln(x_1^{-1}) - sin(x_3 x_1^2)\right), \mathbf{x} \in \mathbb{R}^3$

$$\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial (\mathbf{x})}{\partial x_1} & \frac{\partial f(\mathbf{x})}{\partial x_2} & \frac{\partial f(\mathbf{x})}{\partial x_3} \end{bmatrix} \quad with$$

$$\frac{\partial f(\mathbf{x})}{\partial x_1} = 2exp\left(x_2 - ln(x_1^{-1}) - sin(x_3 x_1^2)\right) \cdot \left(\frac{1}{x_1} - 2x_3 x_1 cos(x_3 x_1^2)\right)$$

$$\frac{\partial f(\mathbf{x})}{\partial x_2} = 2exp\left(x_2 - ln(x_1^{-1}) - sin(x_3 x_1^2)\right) \cdot 1$$

$$\frac{\partial f(\mathbf{x})}{\partial x_3} = 2exp\left(x_2 - ln(x_1^{-1}) - sin(x_3 x_1^2)\right) \cdot (-x_1^2 cos(x_3 x_1^2))$$

(k) $\nabla_y h$ with $h(y) = (g \circ f)(y)$, where $g(\mathbf{x}) = x_1^3 + exp(x_2)$ and $\mathbf{x} := \mathbf{f}(y) = [ysin(y), ycos(y)]^T$. First show your understanding of the application of the chain rule in this example before "plugging in"the actual derivatives.

In other words $x_1 = ysin(y)$, $x_2 = ycos(y)$ which then leads to the formal derivation of the searched gradient:

$$\nabla_y h = \nabla_y (g \circ f) = \frac{\partial g}{\partial \mathbf{x}} \cdot \frac{\partial \mathbf{x}}{\partial y}, \quad \text{whereas}$$
$$\frac{\partial g}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial g}{\partial x_1} & \frac{\partial g}{\partial x_2} \end{bmatrix} \text{ and } \quad \frac{\partial \mathbf{x}}{\partial y} = \begin{bmatrix} \frac{\partial x_1}{\partial y} \\ \\ \frac{\partial x_2}{\partial y} \end{bmatrix}.$$

Then looking at the respective terms:

$$\frac{\partial g}{\partial x_1} = 3x_1^2 \cdot \underbrace{\frac{\partial x_1}{\partial x_1}}_{=1} = 3x_1^2\,; \quad \frac{\partial g}{\partial x_2} = exp(x_2) \cdot \underbrace{\frac{\partial x_2}{\partial x_2}}_{=1} = exp(x_2)$$
$$\frac{\partial x_1}{\partial y} = sin(y) + ycos(y)\,; \quad \frac{\partial x_2}{\partial y} = cos(y) - ysin(y)$$

Plugged into the general forms of above:

$$\frac{\partial g}{\partial \mathbf{x}} = \begin{bmatrix} 3x_1^2 & exp(x_2) \end{bmatrix} \quad \text{and}$$
$$\frac{\partial \mathbf{x}}{\partial y} = \begin{bmatrix} sin(y) + ycos(y) \\ \\ cos(y) - ysin(y) \end{bmatrix}$$

Now, the entire gradient can be assembled as such:

$$\begin{aligned} \nabla_y h = \nabla_y (g \circ f) &= \frac{\partial g}{\partial \mathbf{x}} \cdot \frac{\partial \mathbf{x}}{\partial y} \\ &= \begin{bmatrix} 3x_1^2 & exp(x_2) \end{bmatrix} \cdot \begin{bmatrix} sin(y) + ycos(y) \\ \\ cos(y) - ysin(y) \end{bmatrix} \\ &= 3x_1^2 \cdot (sin(y) + ycos(y)) + exp(x_2) \cdot (cos(y) - ysin(y)) \\ &= 3(ysin(y))^2 \cdot (sin(y) + ycos(y)) + exp(ycos(y)) \cdot (cos(y) - ysin(y)) \end{aligned}$$

(l) We now assume that $\mathbf{x} := \mathbf{f}(y, z) = [ysin(y) + z, ycos(y) + z^2]^T$. Provide $\nabla_{y,z} h$. Hint: To determine the correct shape of $\nabla_{y,z} h$, view the input pair y and z as vecor $[y, z]^T$.

In other words $x_1 = ysin(y) + z$, $x_2 = ycos(y) + z^2$ which then leads to the formal derivation of the searched

gradient:

$$\nabla_{y,z} h = \nabla_{y,z}(g \circ f) = \frac{\partial g}{\partial \mathbf{x}} \cdot \frac{\partial \mathbf{x}}{\partial y \partial z}, \quad \text{whereas}$$

$$\frac{\partial g}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial g}{\partial x_1} & \frac{\partial g}{\partial x_2} \end{bmatrix} \text{ and } \quad \frac{\partial \mathbf{x}}{\partial y \partial z} = \begin{bmatrix} \frac{\partial x_1}{\partial y} & \frac{\partial x_1}{\partial z} \\ \frac{\partial x_2}{\partial y} & \frac{\partial x_2}{\partial z} \end{bmatrix}$$

Then looking at the respective terms from above and the changed ones:

$$\begin{aligned} \frac{\partial g}{\partial x_1} &= 3x_1^2\,; \quad \frac{\partial g}{\partial x_2} = exp(x_2) \\ \frac{\partial x_1}{\partial y} &= sin(y) + ycos(y)\,; \quad \frac{\partial x_1}{\partial z} = 1 \\ \frac{\partial x_2}{\partial y} &= cos(y) - ysin(y)\,; \quad \frac{\partial x_2}{\partial z} = 2z, \end{aligned}$$

which assembled leads to the following final gradient:

$$\begin{aligned} \nabla_{y,z} h = \nabla_{y,z}(g \circ f) &= \frac{\partial g}{\partial \mathbf{x}} \cdot \frac{\partial \mathbf{x}}{\partial y \partial z} \\ &= \begin{bmatrix} \frac{\partial g}{\partial x_1} & \frac{\partial g}{\partial x_2} \end{bmatrix} \cdot \begin{bmatrix} \frac{\partial x_1}{\partial y} & \frac{\partial x_1}{\partial z} \\ \frac{\partial x_2}{\partial y} & \frac{\partial x_2}{\partial z} \end{bmatrix} \\ &= \begin{bmatrix} 3x_1^2 & exp(x_2) \end{bmatrix} \cdot \begin{bmatrix} ysin(y) + ycos(y) & 1 \\ cos(y) - ysin(y) & 2z \end{bmatrix} \\ &= \begin{bmatrix} 3x_1^2 \cdot (sin(y) + ycos(y)) + exp(x_2) \cdot (cos(y) - ysin(y)) \\ 3x_1^2 + exp(x_2) 2z \end{bmatrix}^T \\ &= \begin{bmatrix} 3(ysin(y) + z)^2 \cdot (sin(y) + ycos(y)) + exp(ycos(y) + z^2) \cdot (cos(y) - ysin(y)) \\ 3(ysin(y) + z)^2 + exp(ycos(y) + z^2) 2z \end{bmatrix}^T \end{aligned}$$

1.3

The following questions are good practive in manipulating vectors and matrices and they are very important for solving for posterior distributions. Compute the following gradients, assuming $\mathbf{\Sigma^{-1}}$ is symmetric, positive semi-definite and invertible.

(a) $\nabla_\mu \mathbf{x}^T \mathbf{\Sigma^{-1}} \mu$

Some transformations and suppose $\sigma_{i,j}^2 = \Sigma_{i,j}^{-1}$:

$$\begin{aligned} \mathbf{x}^T \mathbf{\Sigma}^{-1} \mu &= \big(x_1, \ldots, x_n\big) \cdot \mathbf{\Sigma}^{-1} \cdot \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix} \\ &= \big((x_1 \sigma_{1,1}^2 + \cdots + x_n \sigma_{n,1}^2), \ldots, (x_1 \sigma_{n,1}^2 + \cdots + x_n \sigma_{n,n}^2)\big) \cdot \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix} \\ &= \sum_i \sum_j x_i \sigma_{ij}^2 \mu_j, \quad \text{with an individual partial derivative of this looking like the following} \\ \frac{\partial \mathbf{x}^T \mathbf{\Sigma}^{-1} \mu}{\partial \mu_k} &= \sum_i x_i \sigma_{ik}^2 \end{aligned}$$

If those partial derivatives are taken with respect to each dimension, the shape of this gradient then arises as $\mathbb{R}^{1\times n}$ and in matrix form:

$$\nabla_\mu \mathbf{x}^T \mathbf{\Sigma^{-1}} \mu = \frac{\partial \mathbf{x}^T \mathbf{\Sigma^{-1}} \mu}{\partial \mu} = \left(\sum_i x_i \sigma_{i1}^2, \ldots, \sum_i x_i \sigma_{in}^2\right) = \mathbf{x}^T \mathbf{\Sigma}^{-1},$$

(b) $\nabla_\mu \mu^T \mathbf{\Sigma^{-1}} \mu$

Again suppose $\sigma_{ij}^2 = \Sigma_{ij}^{-1}$:,

$$\begin{aligned}
\mu^T \mathbf{\Sigma}^{-1} \mu &= \left(\mu_1, \ldots, \mu_n\right) \cdot \mathbf{\Sigma}^{-1} \cdot \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix} \\
&= \left(\mu_1 \sigma_{11}^2 + \cdots + \mu_n \sigma_{n1}^2 \cdot \mu_1, \ldots, (\mu_1 \sigma_{n,1} + \cdots + \mu_n \sigma_{n,n}) \cdot \mu_n\right) \cdot \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix} \\
&= (\mu_1^2 \sigma_{1,1}^2 + \cdots + \mu_1 \mu_n \sigma_{n,1}^2) + \ldots + (\mu_n \mu_1 \sigma_{n,1}^2 + \cdots + \mu_n^2 \sigma_{n,n}^2)
\end{aligned}$$

As $\mathbf{\Sigma}^{-1}$ is symmetric, $\sigma_{ij}^2 = \sigma_{ji}^2$ holds. With that information, the equation can similarly to before be represented as such:

$$\mu^T \mathbf{\Sigma}^{-1} \mu = \sum_i \sum_j \mu_i \sigma_{ij} \mu_j$$

Again, one partial derivative looks like:

$$\frac{\partial \mu^T \mathbf{\Sigma}^{-1} \mu}{\partial \mu_k} = \sum_i \mu_i \sigma_{ik}^2 + \sum_j \sigma_{kj}^2 \mu_j \overset{symmetric}{=} 2 \sum_j \mu_j \sigma_{jk}^2$$

Then putting all of the partial derivativies together:

$$\nabla_\mu \mu^T \mathbf{\Sigma}^{-1} \mu = \begin{pmatrix} 2\sum_j \mu_j \sigma_{j1}^2 & \ldots & 2\sum_j \mu_j \sigma_{jn}^2 \end{pmatrix} = 2 \cdot \mu^T \mathbf{\Sigma}^{-1}$$

(c) $\nabla_{\mathbf{W}} \mathbf{f}$ where $\mathbf{f} = \mathbf{W}\mathbf{x}$ and assume $\mathbf{W} \in \mathbb{R}^{2\times 3}$ and $\mathbf{x} \in \mathbb{R}^3$. Follow example of 5.11 of the book *mathematics for machine learning* to solve this.

$$\nabla_{\mathbf{W}} \mathbf{f} = \begin{bmatrix} \frac{\partial f_1}{\partial \mathbf{W}} \\ \frac{\partial f_2}{\partial \mathbf{W}} \end{bmatrix} \in \mathbb{R}^{2\times 3\times 2},$$

with following characteristics:

- $\frac{\partial f_i}{\partial W} \in \mathbb{R}^{1\times 3\times 2}$
- $\frac{\partial f_i}{\partial W_{ij}} = x_j$
- $\frac{\partial f_i}{W_{i,:}} = \mathbf{x}^T \in \mathbb{R}^{1\times 3\times 1}$
- $\frac{\partial f_i}{\partial W_{j\neq i,:}} = \mathbf{0}^T \in \mathbb{R}^{1\times 3\times 1}$
- $f_i = \sum_{j=1}^3 A_{ij} x_j$ which for i=1 is for instance: $f_1 = A_{11} \cdot x_1 + A_{12} \cdot x_2 + A_{13} \cdot x_3$
- $\frac{\partial f_i}{\partial A_{iq}} = x_q$

If we now partially look at the terms:

$$\frac{\partial f_1}{\partial \mathbf{W}} = \begin{pmatrix} x_1 & 0 \\ x_2 & 0 \\ x_3 & 0 \end{pmatrix} \quad and \quad \frac{\partial f_2}{\partial \mathbf{W}} = \begin{pmatrix} 0 & x_1 \\ 0 & x_2 \\ 0 & x_3 \end{pmatrix}$$

So in total, the searched gradient looks like:

$$\nabla_{\mathbf{W}}\mathbf{f} = \begin{pmatrix} \frac{\partial f_1}{\partial \mathbf{W}} \\ \frac{\partial f_2}{\partial \mathbf{W}} \end{pmatrix} = \begin{pmatrix} \begin{pmatrix} x_1 & 0 \\ x_2 & 0 \\ x_3 & 0 \end{pmatrix} \\ \begin{pmatrix} 0 & x_1 \\ 0 & x_2 \\ 0 & x_3 \end{pmatrix} \end{pmatrix}$$

and thus is of shape $\mathbb{R}^{2\times 3\times 2}$

(d) $\nabla_{\mathbf{W}}\mathbf{f}$, where $f = (\mu - \mathbf{y}(\mathbf{W},\mathbf{x})^T\mathbf{\Sigma}^{-1}(\mu - \mathbf{y}(\mathbf{W},\mathbf{x}))$ and $\mathbf{y}(\mathbf{W},\mathbf{x}) = \mathbf{W}^T\mathbf{x}$. Assume $\mathbf{W} \in \mathbb{R}^{M\times K}$ Follow these steps:

1. Compute $\frac{\partial f}{\partial W_{i,j}}$- Hint: Use the chainrule by computing $\frac{\partial f}{\partial \mathbf{y}_k}\frac{\partial \mathbf{y}_k}{\partial W_{i,j}}$. Hint: You might find the notation $\mathbf{\Sigma}^{-1}_{:;k}$ helpful to define the kth row in the matrix $\mathbf{\Sigma}^{-1}$.

   With $\frac{\partial f(\mathbf{y})}{\partial \mathbf{W}} = \frac{\partial f}{\partial \mathbf{y}} \cdot \frac{\partial y}{\partial \mathbf{W}}$ in mind, at first f is multiplied:

   $$\begin{aligned} f &= \mu^T\mathbf{\Sigma}^{-1}\mu - \mathbf{y}^T\mathbf{\Sigma}^{-1}\mu - \mu^T\mathbf{\Sigma}^{-1}\mathbf{y} + \mathbf{y}^T\mathbf{\Sigma}^{-1}\mathbf{y} \\ &= \mu^T\mathbf{\Sigma}^{-1}\mu - 2\mu^T\mathbf{\Sigma}^{-1}\mathbf{y} + \mathbf{y}^T\mathbf{\Sigma}^{-1}\mathbf{y} \end{aligned}$$

   Thus the derivative $\frac{\partial f}{\partial \mathbf{y}}$ emerges as:

   $$\frac{\partial f}{\partial \mathbf{y}} = -2\mu^T\mathbf{\Sigma}^{-1} + 2\mathbf{y}^T\mathbf{\Sigma}^{-1} \quad \in \mathbb{R}^{1\times K}$$

   Then also

   $$\frac{\partial \mathbf{y}}{\partial \mathbf{W}} = \begin{bmatrix} \frac{\partial y_1}{\partial \mathbf{W}} \\ \vdots \\ \frac{\partial y_k}{\partial \mathbf{W}} \end{bmatrix} \in \mathbb{R}^{K\times(K\times M)}$$

   Similar to before the following hold:

   - $y_i = \sum_{j=1}^{M} W_{ji}x_j$ and $\frac{\partial y_i}{W_{ji}} = x_j$ $(\star)$
   - $\frac{\partial y_i}{\partial W_{:,i}} = \mathbf{x}^T \in \mathbb{R}^{1\times M}$
   - $\frac{\partial y_i}{\partial W_{:,j\neq i}} = \mathbf{0}^T \in \mathbb{R}^{1\times M}$
   - Putting them together yields: for a single y_i: $\frac{\partial y_i}{\partial \mathbf{W}} = \begin{bmatrix}\mathbf{0} & \dots & \mathbf{x} & \dots & \mathbf{0}\end{bmatrix}^T \in \mathbb{R}^{K\times M}$.

   All of those are used when assembled for $\frac{\partial f}{\partial W_{ij}}$. At first $\frac{\partial f}{\partial y_k}$ is needed:

   $$\frac{\partial f}{\partial y_k} = -2\mu^T\mathbf{\Sigma}^{-1}_{:,k} + 2\mathbf{y}^T\mathbf{\Sigma}^{-1}_{:,k} = 2(\mathbf{y}-\mu)^T\mathbf{\Sigma}^{-1}_{:,k}$$

   So,

   $$\frac{\partial f}{\partial W_{ij}} = \sum_k \frac{\partial f}{\partial y_k}\frac{\partial y_k}{\partial W_{ij}}$$

   Now, the partial results from above come into play:

   $$\sum_k \frac{\partial f}{\partial y_k}\frac{\partial y_k}{\partial W_{ij}} = 2(\mathbf{y}-\mu)^T\mathbf{\Sigma}^{-1}_{:,j}\cdot x_i \quad \in \mathbb{R}^{1\times 1} \quad (\star)$$

2. Combine your insights to answer the original question.
   As a final result, $\nabla_{\mathbf{W}}\mathbf{f}$ emerges as:

   $$\nabla_{\mathbf{W}}\mathbf{f} = \begin{pmatrix} 2(\mathbf{y}-\mu)^T\mathbf{\Sigma}^{-1}_{:,1}\cdot x_1 & \dots & 2(\mathbf{y}-\mu)^T\mathbf{\Sigma}^{-1}_{:,K}\cdot x_1 \\ \vdots & \ddots & \vdots \\ 2(\mathbf{y}-\mu)^T\mathbf{\Sigma}^{-1}_{:,1}\cdot x_M & \dots & 2(\mathbf{y}-\mu)^T\mathbf{\Sigma}^{-1}_{:,K}\cdot x_M \end{pmatrix},$$

   which has the same shape as the **W** matrix $\in \mathbb{R}^{M\times K}$

2 Probability Theory

For these questions you will practice manipulating probabilities and probability density functions using the sum and product rules.

2.1

Suppose you meet Bart. Bart is a quiet, introvert man in hus mid-fourties. He is very shy and withdrawn, invariably helpful but with little interest in people or in the world reality. A meek and tidy soul, he has a need for order and structure and a passion for detail.

(a) What do you think: Is Bart a farmer or is he a librarian?

Well, he could be either. I personally see him more as a librarian, but thats just a matter of stereotype i guess.
(b) Suppose we formulate our belief such that we think that Bart is a librarian, given that he is an introvert, with 0.8 probability. If he were an extrovert, that probability is 0.1. Since the Netherlands is a social country, you assume that the probability of meeting an extrovert is 0.7. Define the random variables and values they can take on, both with symbols and numerically.

- random variable $X : \{L, F\}$ representing whether Bart is a farmer (F) or a Lirbarian (L)
- random variable $Y : \{I, E\}$ indicating that the person you meet is an extrovert (E) and introvert (I).
- $P(X = L|Y = I) = 0.8$ representing the probability that Bart is a Librarian given he is an introvert.
- $P(X = L|Y = E) = 0.1$ representing the probability that Bart is a Librarian given he is an extrovert.
- $P(Y = E) = 0.7$ probability that the person you meet is an extrovert.
- $P(Y = I) = 0.3$ probability that the person you meet is an introvert.

(c) What is the probability that Bart is a librarian?
So $P(X = L$ is sought. Therefore we use the ... rule:

$$\begin{aligned} P(X = L) &= P(X = L|Y = I) \cdot P(Y = I) + P(X = L|Y = E) \cdot P(Y = E) \\ &= 0.8 \cdot 0.3 + 0.1 \cdot 0.7 \\ &= 0.31 \end{aligned}$$

(d) Suppose you looked up the actual statistics and find out that there are 1000 times more farmers than librarians in the Netherlands. Additionally, Bart'd friend is telling you that if Bart were a librarian, the probability of him being an extrovert would be low, and equal to 0.1. Given this information, how should we update our belief in that Bart is a librarian, if he is an introvert? How does does influence the answer to the previous question: What is the probability that Bart is a librarian?

So according to Barts friend: $P(Y = E|X = L) = 0.1$ (thats why $P(Y = I|X = L) = 0.9$) and there are 1000 more farmers than librarians. The second information means the following:

$$P(X = F) = 1000 \cdot P(X = L)$$

Or in other words, that $P(X = L) = \frac{1}{1001}$, as $P(X = F) = \frac{1000}{1001}$. Now according to Bayes rule:

$$\begin{aligned} P(X = L|Y = I) &= \frac{P(Y = I|X = L) \cdot P(X = L)}{P(Y = I)} \\ &= \frac{0.9 \cdot \frac{1}{1001}}{0.3} \\ &= \frac{3}{1001} \end{aligned}$$

2.2

For this question you will compute the expression for the posterior parameter distributions for a simple data problem. Assume we observe N univariate data points $\{x_1, x_2, \ldots, x_N\}$. Further, we assume that they are generated by observing the outcome of a coin flip. Therefore the random variable representing the outcome of the coin flip is described by a Bernoulli distribution. The coin takes on the values 0,1, where the probability of throwing a 1 is equal to $\rho \in [0, 1]$- A Bernoulli random variable c has the following probability density function: $p(x) = \rho^x(1-\rho)^{(1-\rho)}$.

(a) Write down the expression for the likelihood of the ovserved datapoints.

Likelihood of the observed data points:

$$\begin{aligned} p(\mathbf{x}|\rho) &\underbrace{=}_{iid} \prod_{i=1}^{N} p(x_i|\rho) \\ &= \prod_{i=1}^{N} \rho^{x_i}(1-\rho)^{(1-x_i)} \end{aligned}$$

(b) Assume you observed the following coin throws: [1,1]. Which value of ρ *most likely* generated this data-set?

$x_1 = 1; x_2 = 1$; $\mathbf{x} = \{1, 1\}$. It is *most likely* that given x_1 and x_2 ρ was $\rho = 1$. As, when the likelihood for the both samples is written:

$$\begin{aligned} p(\mathbf{x}|\rho) &= p(x_1|\rho) \cdot (x_2|\rho) \\ &= \rho^1(1-\rho)^{(1-1)} \cdot \rho^1(1-\rho)^{(1-1)} \\ &= \rho \cdot \rho \end{aligned}$$

If we now look for the ρ leading to biggest likelihood, then $\rho = 1$ is the case, as then ρ^2 will be the largest: $\hat{\rho} = argmax_{\rho \in [0,1]} \rho^2 \rightarrow \hat{\rho} = 1$

We now assume some prior $p(\rho)$ over our model's parameter ρ.

(c) Write down the general expression for the posterior over the parameter ρ assuming we observe the dataset D. Indicate which terms correspond to *prior, likelihood, evidence* and *posterior*. (Do not bother to fill in the (bernoulli) form of the likelihood).

$$p(\rho|D) = \frac{p(D|\rho) \cdot p(\rho)}{p(D)},$$

whereas

- prior: $p(\rho)$
- posterior: $p(\rho|D)$
- evidence: $p(D|\rho)$

(d) Assume our prior $p(\rho)$ reflects our belief that the coin is fair. Under this prior, how will our posterior belief into the value of ρ change, assuming we observed the throws [1,1]?

Fair coin $\rightarrow$ equal prior: $\rho = 0.5$.
As not noted otherwise, the uniform distribution $\mathcal{U} \in$ [0,1] is chosen for a prior distribution over ρ. With that information, $p(\rho = 0.5) = \frac{1}{1-0} = 1$ holds (density of uniform distribution). As determined in (b), $p(\mathbf{x}|\rho) = \rho^2$. So the denominator in the general form of the posterior then is according to the following:

$$\begin{aligned} p(D = \mathbf{x}) = p(D = \{1, 1\}) &= \int_0^1 p(\rho) \cdot p(D|\rho) d\rho \\ &= \int_0^1 1 \cdot \rho^2 d\rho = \left[\frac{\rho^3}{3}\right]_0^1 = \frac{1}{3} \end{aligned}$$

$$p(\rho|D) = \frac{p(D|\rho) \cdot p(\rho)}{p(D)} \quad \Rightarrow \quad p(\rho|D) = \frac{0.5^2 \cdot 1}{\frac{1}{3}} = \frac{3}{4}$$