# Statistical Machine Learning 2018

Exercises and answers, week 1

7 September 2018

## TUTORIAL

## Exercise 1

Calculate the gradient $\nabla f$ of the following functions $f(\mathbf{x})$. In the left column, $\mathbf{x} = (x_1, x_2, x_3)$. In the right column, $\mathbf{x} = (x_1, \ldots, x_n)$.

a) $f(x_1, x_2, x_3) = a_1 x_1 + a_2 x_2 + a_3 x_3$

b) $f(x_1, x_2, x_3) = x_2$

c) $f(x_1, x_2, x_3) = x_1 x_2 x_3$

d) $f(x_1, x_2, x_3) = x_1^{k_1} x_2^{k_2} x_3^{k_3}$

e) $f(\mathbf{x}) = \sum_{i=1}^n a_i x_i$

f) $f(\mathbf{x}) = x_i$

g) $f(\mathbf{x}) = \prod_{i=1}^n x_i$

h) $f(\mathbf{x}) = \prod_{i=1}^n x_i^{k_i}$

Note: often it suffices to write down the partial derivative $\partial f / \partial x_j$ (Can you tell why?).

ANSWER: a) $(a_1, a_2, a_3)$, in other words $\partial f / \partial x_i = a_i,\ i = 1 \ldots 3$
b) $(0, 1, 0)$, in other words $\partial f / \partial x_j = \delta_{2j},\ j = 1 \ldots 3$ (Kronecker delta, see slides)
c) $(x_2 x_3, x_1 x_3, x_1 x_2)$
d) $(k_1 x_1^{k_1 - 1} x_2^{k_2} x_3^{k_3}, k_2 x_1^{k_1} x_2^{k_2 - 1} x_3^{k_3}, k_3 x_1^{k_1} x_2^{k_2} x_3^{k_3 - 1})$ (where the $k_i x_i^{k_i - 1}$ is understood as 0 if $k_i = 0$)
e) $(a_1, \ldots, a_n)$ in other words $\partial f / \partial x_j = a_j$
f) $(\delta_{i1}, \ldots, \delta_{in})$ in other words $\partial f / \partial x_j = \delta_{ij}$
g) Note that $\prod_{i=1}^n x_i = (\prod_{i=1, i \neq j}^n x_i) x_j$, so $\partial f / \partial x_j = \prod_{i=1, i \neq j}^n x_i$
h) $\partial f / \partial x_j = k_j x^{k_j - 1} \prod_{i=1, i \neq j}^n x_i^{k_i}$
To describe a vector say $\vec{u} = (u_1, u_2, \ldots, u_j, \ldots, u_n)$, it suffices to give the expresssion of an arbitrary component $u_j$. So $u_j$ is some expression that contains $j$'s. All components and so the complete vector can then be reconstructed by filling in the appropriate component number for $j$. E.g. if you look for $u_2$, take the general expression for $u_j$ and substitute all the $j$'s by a 2. Now the gradient $\nabla f$ is also a vector. Its $j$-th component is just $\partial f / \partial x_j$ , which is therefore sufficient to describe the vector $\nabla f$.

In many cases, it is convenient to only write down the abstract $j$ component. However, it should be remembered that the gradient is an object with $n$ components, and that it is sometimes more convenient to write down all the components. I think this could be argued for e.g. the gradient in b), $\nabla f = (0, 1, 0)$.

## Exercise 2

The function

$$f(x, y) = 2x^2 - xy + y^2 - x + y + 5.5 \tag{1}$$

has a unique minimum $(x^*, y^*)$. Calculate this point.

ANSWER: Partial derivatives of $f$ are given by

$$\begin{aligned}\frac{\partial f}{\partial x} &= 4x - y - 1 \\ \frac{\partial f}{\partial y} &= 2y - x + 1\end{aligned}$$

Setting equal to zero yields two equations for $x$ and $y$. Solve the first to get: $y = 4x - 1$. Substituting in the second then gives: $8x - 2 - x + 1 = 7x - 1 = 0 \Rightarrow x^* = 1/7$, and so $y^* = -3/7$.

(As a side remark: it is indeed a *minimum* since the Hessian, the matrix of second order partial derivatives, is positive definite, meaning that $x^T M x > 0$ for all vectors $x$. An equivalent statement is that the eigenvalues $\lambda_i$ of the matrix $M$ are all positive.)

# Exercise 3

Calculate the minimum $x^*$ of the following two functions.

1. $f(x) = \sum_{i=1}^{n}(x - a_i)^2$

   ANSWER:

$$\begin{aligned}\frac{df(x)}{dx} &= 2\sum_{i=1}^{n}(x - a_i) = 0 \\ &\Rightarrow \sum_{i=1}^{n} x = \sum_{i=1}^{n} a_i \\ &\Rightarrow nx = \sum_{i=1}^{n} a_i \\ &\Rightarrow x = \frac{1}{n}\sum_{i=1}^{n} a_i,\end{aligned}$$

   so $x^*$ is the mean of the $a_i$'s.

   There are several things to note:

   (a) The derivative of a sum is a sum of derivatives.

   (b) $x$ has no subindex $i$. Therefore, $\sum_{i=1}^{n} x = \underbrace{x + x + \ldots + x}_{n \text{ times}} = nx$.

   (c) This minimization can be seen as a *least squares problem*: given data $a_i$, find $x$ that gives the best fit such that the sum of the squares of the errors $(x - a_i)$ is minimal. The solution is the data mean.

2. $f(x) = \sum_{i=1}^{n} \alpha_i (x - a_i)^2 \quad$ (with $\alpha_i > 0$)

ANSWER:

$$\begin{aligned}\frac{df(x)}{dx} &= 2\sum_{i=1}^{n}\alpha_i(x-a_i)=0\\ &\Rightarrow \sum_{i=1}^{n}\alpha_i x=\sum_{i=1}^{n}\alpha_i a_i\\ &\Rightarrow x=\frac{\sum_{i=1}^{n}\alpha_i a_i}{\sum_{i=1}^{n}\alpha_i}\end{aligned}$$

This minimization can be seen as a *weighted least square problem*: given data $a_i$, find $x$ that gives the best fit such that the sum of the squares of the errors $(x-a_i)$ weighted by $\alpha_i$ is minimal. The solution is the weighted mean. Here, $x^*$ is the weighted average of $a_i$ with weights $\alpha_i$. The factor in the denominator (noemer) is for normalization (just as $n$ is in the previous case).

# Exercise 4

(see Bishop, appendix C, eq.C.1) A matrix $\mathbf{M}$ has elements $M_{ij}$ (with $i$ the row and $j$ the column index). The transposed matrix $\mathbf{M}^{\mathrm{T}}$ has elements $(\mathbf{M}^{\mathrm{T}})_{ij}=M_{ji}$. By writing out the matrix product using index notation show that

$$(\mathbf{AB})^{\mathrm{T}}=\mathbf{B}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}, \tag{2}$$

where $\mathbf{A}$ is a $M\times N$ matrix and $\mathbf{B}$ is a $N\times P$ matrix.

Hint: $\mathbf{C}=\mathbf{AB}$ corresponds to $C_{ij}=\sum_{k=1}^{N}A_{ik}B_{kj}$

ANSWER: $(\mathbf{A})_{ij}=A_{ij}$, $(\mathbf{A}^{\mathrm{T}})_{ij}=A_{ji}$, $(\mathbf{AB})_{ij}=\sum_k A_{ik}B_{kj}$ so

$$\begin{aligned}((\mathbf{AB})^{\mathrm{T}})_{ij} &= (\mathbf{AB})_{ji}=\sum_k A_{jk}B_{ki}=\sum_k B_{ki}A_{jk}\\ &= \sum_k(\mathbf{B}^{\mathrm{T}})_{ik}(\mathbf{A}^{\mathrm{T}})_{kj}=(\mathbf{B}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}})_{ij}\end{aligned}$$

# Exercise 5

(see Bishop, Exercise 1.1) Consider the $M$-th order polynomial

$$y(x;\mathbf{w})=w_0+w_1x+\ldots+w_Mx^M=\sum_{j=0}^{M}w_jx^j \tag{3}$$

and the error function

$$E(\mathbf{w})=\frac{1}{2}\sum_{n=1}^{N}\{y(x_n;\mathbf{w})-t_n\}^2 \tag{4}$$

with $x_n,t_n$ the input/output pairs from the data set. Define the error per data point as

$$E_n(\mathbf{w})=\frac{1}{2}\{y(x_n;\mathbf{w})-t_n\}^2 \tag{5}$$

(so $E=\sum_{n=1}^{N}E_n$). Note that $x$ = 1-dimensional, and that in this exercise the super-indices $i,j$ represent 'power'.

1. Calculate the gradient of the error per data point $E_n$:

$$\nabla E_n \quad (= (\frac{\partial E_n}{\partial w_0}, \ldots, \frac{\partial E_n}{\partial w_M})^T\ ). \tag{6}$$

ANSWER: Use the chain rule on $E_n(\mathbf{w}) = \frac{1}{2}\{u(\mathbf{w})\}^2$ with $u(\mathbf{w}) = y(x_n; \mathbf{w}) - t_n$. Then for the components of the gradient

$$\begin{aligned}
\frac{\partial E_n}{\partial w_i} &= \frac{\partial E_n}{\partial u} \frac{\partial u}{\partial w_i} \\
&= u(\mathbf{w}) \frac{\partial u(\mathbf{w})}{\partial w_i} \\
&= (y(x_n; \mathbf{w}) - t_n) \frac{\partial y(x_n; \mathbf{w}) - t_n}{\partial w_i} \\
&= \left( \sum_{j=0}^{M} w_j (x_n)^j - t_n \right) \frac{\partial}{\partial w_i} \left[ \sum_{k=0}^{M} w_k x_n^k - t_n \right] \\
&= \left( \sum_{j=0}^{M} w_j x_n^j - t_n \right) x_n^i \\
&= \sum_{j=0}^{M} w_j x_n^{i+j} - t_n x_n^i
\end{aligned}$$

Note that $x_n^i$ means: $x_n$ to-the-power-of $i$. Note that in general $x^a x^b = x^{a+b}$, e.g. $2^3 2^4 = 2^7$ If you got this answer by direct differentiation e.g. by writing out the $y$'s in terms of $w$'s, without the use of an $u$, that is of course also ok.

2. Calculate the gradient of the total error $E$.

ANSWER: The total error $E$ is the sum of the errors per datapoint $E_n$. Since the gradient is a linear function of its operands: $\nabla(f + g) = \nabla f + \nabla g$, the gradient of the total error is the sum of the gradients of the error per datapoint:

$$\nabla E = \sum_{n=1}^{N} \nabla E_n$$

with$\nabla E_n$ as above. So,

$$\frac{\partial E}{\partial w_i} = \sum_{n=1}^{N} \Big( \sum_{j=0}^{M} w_j x_n^{i+j} - t_n x_n^i \Big)$$

3. Show that the partial derivatives can be written as

$$\frac{\partial E}{\partial w_i} = \sum_{j=0}^{M} A_{ij} w_j - T_i \tag{7}$$

with $A_{ij}$ and $T_i$ defined as

$$A_{ij} = \sum_{n=1}^{N} x_n^{i+j} \qquad T_i = \sum_{n=1}^{N} t_n x_n^i. \tag{8}$$

ANSWER: Substituting the result for the components of $\nabla E_n$ into (2) we have

$$\begin{aligned}
\frac{\partial E}{\partial w_i} &= \sum_{n=1}^{N} \Big( \sum_{j=0}^{M} w_j x_n^{i+j} - t_n x_n^i \Big) \\
&= \sum_{n=1}^{N} \sum_{j=0}^{M} w_j x_n^{i+j} - \sum_{n=1}^{N} t_n x_n^i \\
&= \sum_{j=0}^{M} \sum_{n=1}^{N} x_n^{i+j} w_j - \sum_{n=1}^{N} t_n x_n^i \\
&= \sum_{j=0}^{M} A_{ij} w_j - T_i
\end{aligned}$$

4. When $E$ is minimal it holds that $\nabla E = 0$ (i.e., all partial derivatives are zero). Using this, show that in the minimum of $E$ the parameters $\mathbf{w}$ satisfy

$$\sum_{j=0}^{M} A_{ij} w_j = T_i. \tag{9}$$

ANSWER: In the last result, setting all partial derivatives equal to zero implies that when the error is minimal then

$$\sum_{j=0}^{M} A_{ij} w_j - T_i = 0 \;\Rightarrow\; \sum_{j=0}^{M} A_{ij} w_j = T_i.$$

5. Verify that for a single data point $\{x_1, t_1\}$ the optimal solution for a first order polynomial through the origin takes the form

$$w_1 = \frac{1}{A_{11}} T_1 \tag{10}$$

ANSWER: A first order polynomial through the origin implies an equation of the form $y(x; \mathbf{w}) = w_1 x$, i.e. $w_0 = 0$. That means that, with $M = 1$, (9) reduces to

$$\sum_{j=1}^{1} w_j A_{ij} = w_1 A_{i1} = T_i \tag{11}$$

This holds for both equations, i.e. $i = 0$ and $i = 1$, and so choosing the latter and dividing both sides by $A_{11}$ gives the result to show.

6. Show that for an arbitrary data set $\{x_n, t_n\}$ the optimal solution for an M-th order polynomial takes the form

$$\mathbf{w} = \mathbf{A}^{-1}\mathbf{T} \tag{12}$$

ANSWER:

Rewriting summation as matrix multiplication the set of equations (9) becomes

$$\begin{aligned}\sum_{j=0}^{M} w_j A_{ij} &= \sum_{j=0}^{M} A_{ij} w_{j1} \\ &= \big(\mathbf{A}\mathbf{w}\big)_{i1} = T_{i1}\end{aligned}$$

Combining all $i$ components into matrix form this corresponds to

$$\mathbf{A}\mathbf{w} = \mathbf{T}$$

As left-multiplying by $\mathbf{A}^{-1}$ gives $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$, the identity matrix, we get back equation (12):

$$\mathbf{w} = \mathbf{A}^{-1}\mathbf{T}$$

7. One technique that is often used to control the over-fitting phenomenon is *regularization*. Consider adding a penalty term to the squared error loss that takes the form of the sum-of-squares of all coefficients. The error function becomes:

$$\tilde{E}(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n;\mathbf{w}) - t_n\}^2 + \frac{\lambda}{2}||\mathbf{w}||^2, \tag{13}$$

where $||\mathbf{w}||^2 = \mathbf{w}^{\mathrm{T}}\mathbf{w} = \sum_{j=0}^{M} w_j^2$. Write down the set of coupled linear equations for the modified error function, analogous to the case without regularization:

$$\sum_{j=0}^{M} w_j \tilde{A}_{ij} = \tilde{T}_i. \tag{14}$$

Compare $\tilde{A}_{ij}$ and $\tilde{T}_i$ to $A_{ij}$ and $T_i$.

ANSWER:

We have previously determined that:

$$\frac{\partial E}{\partial w_i} = \sum_{j=0}^{M} A_{ij} w_j - T_i.$$

$$\begin{aligned}\tilde{E}(\mathbf{w}) = E(\mathbf{w}) + \frac{\lambda}{2}||\mathbf{w}||^2 \implies \frac{\partial \tilde{E}}{\partial w_i} &= \frac{\partial E}{\partial w_i} + \lambda w_i \\ &= \sum_{j=0}^{M} A_{ij} w_j - T_i + \lambda w_i \\ &= \sum_{j=0}^{M} A_{ij} w_j - T_i + \lambda \sum_{j=0}^{M} \delta_{ij} w_j \\ &= \sum_{j=0}^{M} (A_{ij} + \lambda\delta_{ij}) w_j - T_i \\ &= \sum_{j=0}^{M} w_j \tilde{A}_{ij} - \tilde{T}_i\end{aligned}$$

We conclude that $\tilde{A}_{ij} = A_{ij} + \lambda\delta_{ij}$ and $\tilde{T}_i = T_i$.

## Exercise 6

In this exercise, we will have a closer look at the gradient descent algorithm for function minimization. When the function to be minimized is $E(\mathbf{x})$, the gradient descent iteration is

$$\mathbf{x}_{n+1} = \mathbf{x}_n - \eta \nabla E(\mathbf{x}_n) \tag{15}$$

where $\eta > 0$ is the so-called learning-rate.

1. Consider the function $E(x) = \frac{\lambda}{2}(x-a)^2$ with parameters $\lambda > 0$, and $a$ arbitrary.

   (a) Write down the gradient descent iteration rule. Verify that the minimum of $E$ is $a$ and that $a$ is a fixed point[1] of the gradient descent iteration rule.

   ANSWER:
   $$x_{n+1} = x_n - \eta\lambda(x_n - a) = (1-\eta\lambda)x_n + \eta\lambda a$$
   The minimum of $E$ is $x^* = a$ (for this value $E$ is zero, for any other value, it is larger). Fixed point: fill in $a = (1-\eta\lambda)a + \eta\lambda a = a$.

   (b) Show that the algorithm converges in one step if $\eta = 1/\lambda$.

   ANSWER: With $\eta = 1/\lambda$,
   $$x_{n+1} = x_n - (x_n - a) = a$$
   So $x_1 = a$ for any $x_0$.

   (c) Define $d_n = x_n - a$. Show that if $0 < \eta < 1/\lambda$, subsequent $d_n$'s have the same signs. Also show that if $\eta > 1/\lambda$, subsequent $d_n$'s have opposite signs.

   ANSWER: In terms of $d_n$ the iteration rule is
   $$d_{n+1} = d_n - \eta\lambda d_n = (1-\eta\lambda)d_n$$
   If $0 < \eta < 1/\lambda$ then $(1-\eta\lambda) > 0$ and if $\eta > 1/\lambda$ then $(1-\eta\lambda) < 0$

   (d) The distance to the fixed point is $|d_n|$. Show that $|d_{n+1}| = |(1-\eta\lambda)||d_n|$. Show that this implies that the algorithm converges to the fixed point if $0 < \eta < 2/\lambda$, and that it diverges if $\eta > 2/\lambda$.

   ANSWER: In terms of $d_n$ the iteration rule is
   $$d_{n+1} = d_n - \eta\lambda d_n = (1-\eta\lambda)d_n$$
   so
   $$|d_n| = |(1-\eta\lambda)|^n |d_0|$$
   If $0 < \eta < 2/\lambda$, then $|(1-\eta\lambda)| < 1$ and $|(1-\eta\lambda)|^n \to 0$. If $\eta > 2/\lambda$ then $|(1-\eta\lambda)| > 1$ and $|(1-\eta\lambda)|^n \to \infty$

2. Consider now the function $E(x,y) = \frac{\lambda_1}{2}(x-a_1)^2 + \frac{\lambda_2}{2}(y-a_2)^2$ with parameters $0 < \lambda_1 < \lambda_2$, and $a_i$ arbitrary.

   (a) Write down the gradient descent iteration rule. Verify that the minimum of $E$ is a fixed point.

   ANSWER:
   $$x_{n+1} = (1-\eta\lambda_1)x_n + \eta\lambda_1 a_1 \tag{16}$$
   $$y_{n+1} = (1-\eta\lambda_2)y_n + \eta\lambda_2 a_2 \tag{17}$$
   The minimum of $E$ is $(a_1, a_2)$. Two equations are decoupled. Same as previous.

[1] A fixed point $x^*$ of an iteration $x_{n+1} = F(x_n)$ satisfies $x^* = F(x^*)$.

(b) We want to find the learning rate $\eta$ that leads to the fasted convergence in both $x$ and $y$ direction. This optimal learning rate is the one for which both $|1 - \eta\lambda_1|$ and $|1 - \eta\lambda_2|$ are as small as possible. For the optimal learning rate, the equation $|1-\eta\lambda_1| = |1-\eta\lambda_2|$ must therefore hold. Since $\lambda_1 < \lambda_2$, this can only hold if $\eta\lambda_1 < 1$ and $\eta\lambda_2 > 1$.

- Show that solving the equation leads to $\eta^* = 2/(\lambda_2 + \lambda_1)$ (which is the optimal learning rate). What happens if $\eta$ is smaller than the optimal value? What happens if it is larger?

ANSWER: The solution is to set $|1 - \eta\lambda_1| = |1 - \eta\lambda_2|$, where $\eta\lambda_1 < 1$ and $\eta\lambda_2 > 1$. So $1 - \eta\lambda_1 = \eta\lambda_2 - 1$. So $\eta = 2/(\lambda_2 + \lambda_1)$. When $\eta$ smaller: slows down in the flat direction. $\eta$ larger: more overshoot in the steep direction, causing slowing down.

(c) What is the value of $|1-\eta^*\lambda_i|$ in both directions? What does this say about the applicability of gradient descent to functions with steep hills and flat valleys (i.e., if $\lambda_2 \gg \lambda_1$)?

ANSWER:

$$\left|1 - 2\frac{\lambda_i}{\lambda_2 + \lambda_1}\right| = \left|\frac{\lambda_2 - \lambda_1}{\lambda_2 + \lambda_1}\right|$$

If $\lambda_2 \gg \lambda_1$, then this value is approximately $1-2\lambda_1/\lambda_2$, which is only a little bit smaller than 1, i.e. gradient descent will converge only very slowly.

# BONUS PRACTICE

## Exercise 7

In analyzing problems in which a sigma-summation symbol is involved, it is sometimes helpful to write out the sum. By writing out the sum, I mean e.g.,

$$\sum_{i=1}^{5} x_i = x_1 + x_2 + x_3 + x_4 + x_5$$

or more general

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \ldots + x_n \ .$$

- Show, by explicitly writing out the sums, rearranging terms, and using brackets where needed, that the following four equations hold:

$$\sum_{i=1}^{3}(ax_i) = a\big(\sum_{i=1}^{3} x_i\big) \tag{18}$$

$$\sum_{i=1}^{3}\big(\sum_{j=1}^{2} a_{ij}\big) = \sum_{j=1}^{2}\big(\sum_{i=1}^{3} a_{ij}\big) \tag{19}$$

$$\sum_{i=1}^{3}\big(\sum_{j=1}^{2} x_i y_j\big) = \big(\sum_{i=1}^{3} x_i\big)\big(\sum_{j=1}^{2} y_j\big) \tag{20}$$

$$\sum_{i=1}^{3} a = 3a \tag{21}$$

ANSWER:
Show (18):

$$\begin{aligned}\sum_{i=1}^{3}\big(ax_i\big) &= ax_1 + ax_2 + ax_3 \\ &= a(x_1 + x_2 + x_3) \\ &= a\big(\sum_{i=1}^{3} x_i\big)\end{aligned}$$

Show (19):

$$\begin{aligned}\sum_{i=1}^{3}\big(\sum_{j=1}^{2} a_{ij}\big) &= \sum_{i=1}^{3}(a_{i1} + a_{i2}) \\ &= (a_{11} + a_{12}) + (a_{21} + a_{22}) + (a_{31} + a_{32}) \\ &= (a_{11} + a_{21} + a_{31}) + (a_{12} + a_{22} + a_{32}) \\ &= \sum_{j=1}^{2}(a_{1j} + a_{2j} + a_{3j}) \\ &= \sum_{j=1}^{2}\big(\sum_{i=1}^{3} a_{ij}\big)\end{aligned}$$

Show (20):

$$\sum_{i=1}^{3}\big(\sum_{j=1}^{2} x_i y_j\big) = (x_1y_1 + x_1y_2) + (x_2y_1 + x_2y_2) + (x_3y_1 + x_3y_2)$$
$$= x_1(y_1 + y_2) + x_2(y_1 + y_2) + x_3(y_1 + y_2)$$
$$= (x_1 + x_2 + x_3)(y_1 + y_2)$$
$$= \big(\sum_{i=1}^{3} x_i\big)\big(\sum_{j=1}^{2} y_j\big)$$

Show (21):

$$\sum_{i=1}^{3} a = a + a + a$$
$$= 3a$$

# Exercise 8

Calculate the gradient $\nabla f$ of

$$f(\vec{h}) = \sum_{i=1}^{n} p_i h_i - \ln\Big(\sum_{i=1}^{n} \exp(h_i)\Big) \tag{22}$$

ANSWER:
$\vec{h}$ is a vector of $n$ components $(h_1, \ldots, h_n)$. The function $f(\vec{h})$ is a scalar function of these $n$ components; the $p_i$ are constants. The gradient $\nabla f$ is then the vector of partial derivatives of $f$ w.r.t. each component $h_j$. Since

$$\frac{\partial}{\partial h_j}\left[\sum_{i=1}^{n} p_i h_i\right] = p_j$$

and

$$\frac{\partial}{\partial h_j}\left[\sum_{i=1}^{n} \exp(h_i)\right] = \exp(h_j)$$

application of the chain rule to (22) gives

$$\frac{\partial f}{\partial h_j} = p_j - \frac{\exp(h_j)}{\sum_{i=1}^{n} \exp(h_i)}$$

Side remark: this $f$ is related to a so-called likelihood function (will be treated later in the course).

# Exercise 9

Compute the minimum $x^*$ of

$$f(x) = a\ln(x) + \frac{b}{2x^2} \tag{23}$$

with $a > 0$, $b > 0$ and $x > 0$. Express your answer in terms of $a$ and $b$. (Note: $\ln(x)' = 1/x$).

ANSWER: Calculate gradient (slope) of $f$, set equal to zero and solve for $x^*$

$$\begin{aligned} \frac{a}{x} - bx^{-3} = 0 \quad &\Rightarrow \quad a - bx^{-2} = 0 \\ &\Rightarrow \quad ax^2 - b = 0 \\ &\Rightarrow \quad x = \sqrt{b/a} \end{aligned}$$

Side remark: this $f$ is also related to (another) likelihood function (will also be treated later in the course).

# Exercise 10

(see Bishop, eq.C.8 and C.9) The trace $\mathsf{Tr}\left(\mathbf{A}\right)$ of a square matrix $\mathbf{A}$ is defined as the sum of the elements on the main diagonal:

$$\mathsf{Tr}\left(\mathbf{A}\right) = \sum_{i=1}^{N} A_{ii} \tag{24}$$

1. Prove by writing out in terms of indices that

$$\mathsf{Tr}\left(\mathbf{AB}\right) = \mathsf{Tr}\left(\mathbf{BA}\right) \tag{25}$$

ANSWER: $\mathsf{Tr}\left(\mathbf{A}\right) = \sum_i A_{ii}$, so

$$\mathsf{Tr}\left(\mathbf{AB}\right) = \sum_i \sum_k A_{ik} B_{ki} = \sum_k \sum_i B_{ki} A_{ik} = \mathsf{Tr}\left(\mathbf{BA}\right)$$

2. Show that from this symmetry it follows that the trace is *cyclic*:

$$\mathsf{Tr}\left(\mathbf{ABC}\right) = \mathsf{Tr}\left(\mathbf{CAB}\right) = \mathsf{Tr}\left(\mathbf{BCA}\right) \tag{26}$$

ANSWER: Use

$$\mathbf{ABC} = \mathbf{A}(\mathbf{BC})$$

and take $(\mathbf{BC})$ to be a single matrix in the trace. The symmetry property (25) then implies:

$$\mathsf{Tr}\left(\mathbf{A}(\mathbf{BC})\right) = \mathsf{Tr}\left((\mathbf{BC})\mathbf{A}\right)$$

etc.

# Exercise 11

(see Bishop, eq.C.20) The derivative of a matrix $\mathbf{A}$ with elements $A_{ij}$ depending on $x$ is the matrix $\partial\mathbf{A}/\partial x$ with elements $\partial A_{ij}/\partial x$. Show, by writing out in elements, that

$$\frac{\partial}{\partial x}(\mathbf{AB}) = \frac{\partial \mathbf{A}}{\partial x}\mathbf{B} + \mathbf{A}\frac{\partial \mathbf{B}}{\partial x} \tag{27}$$

ANSWER:

$$\frac{\partial}{\partial x}\left(\sum_k A_{ik} B_{kj}\right) = \sum_k \frac{\partial A_{ik}}{\partial x} B_{kj} + \sum_k A_{ik} \frac{\partial B_{kj}}{\partial x}$$