Exercise 2

Machine Learning I

2A-1.

a) Ordinary Multiplication is not defined for vectors. The dot product is no ordinary multiplication, as it does not satisfy e.g. field axioms.

b) $\boldsymbol{a}\boldsymbol{a}^T = \begin{pmatrix} 1 & 0 & 2 \\ 0 & 0 & 0 \\ 2 & 0 & 4 \end{pmatrix}$

c) $\boldsymbol{a}^T\boldsymbol{a} = 5.$

2A-2.

a)

We want to decompose an arbitrary matrix $\mathbf{A}$ into:

$$A = W_S + W_a$$

where W_S is a symmetric matrix and W_a is skew symmetric.

We basically have the constraints:

$$\begin{aligned} w_{ij} &= w_{ij}^S + w_{ij}^A \\ w_{ji} &= w_{ji}^S + w_{ji}^A \end{aligned}.$$

Setting

$$w_{ij}^S = \frac{1}{2}(w_{ij} + w_{ji})$$

$$w_{ij}^W = \frac{1}{2}(w_{ij} - w_{ji})$$

satisfies the system of equations and maintains the symmetric properties. This decomposition is comparable to the Euler decomposition of the complex sine and cosine functions.

b)

Assumption: The x_i satisfy field axioms (especially with regards to multiplicative commutativity).

Consider:

$$w_{ij}x_ix_j = \left[\frac{1}{2}(w_{ij}+w_{ji}) + \frac{1}{2}(w_{ij}-w_{ji})\right]x_ix_j$$

$$w_{ji}x_jx_i = \left[\frac{1}{2}(w_{ji}+w_{ij}) + \frac{1}{2}(w_{ji}-w_{ij})\right]x_jx_i$$

Addition now lead to pairwise cancellation of w_{ij}^W:

$$w_{ij}x_ix_j + w_{ji}x_jx_i = (w_{ij}+w_{ji})x_jx_i + \underbrace{0}_{w_{ij}^W+w_{ji}^W}.$$

Because the polynomial in

$$\sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij}x_{ij}x_{ji}$$

allows $i = j$, the contribution of the skew symmetric matrix vanishes. This is convenient, as it only collapses monomials $x_{ij} \cdot x_{ji}, x_j \cdot x_i$ that are linearly dependent anyway.

c)

In a symmetric matrix, each row i has i entries.

$$\begin{pmatrix} a_1 & 0 & \dots & 0 \\ a_2 & a_3 & \dots & 0 \\ \dots & \dots & \dots & 0 \\ a_{\frac{d(d-1)}{2}} & a_{\frac{d(d-1)}{2}+1} & \dots & a_{\frac{d(d+1)}{2}} \end{pmatrix}$$

Consequently, the number of entries is equivalent to the sum of natural numbers up until d. But this sum can already be expressed in closed form by the well-known Gauss formula:

$\sum_{i=1}^{d} i = \underbrace{\frac{d(d+1)}{2}}_{Gauss\ Sum}.$

2A-3.

Auxiliary calculation

Inverse of a 2x2 Matrix:

$$\begin{aligned} \Sigma^{-1} &= \frac{1}{\Sigma_{11}\Sigma_{22} - \Sigma_{22}\Sigma_{12}} \begin{pmatrix} \Sigma_{22} & -\Sigma_{12} \\ -\Sigma_{12} & \Sigma_{22} \end{pmatrix} \\ &= \frac{1}{\sigma_a^2\sigma_b^2 - Cov(x_a,x_b)^2} \begin{pmatrix} \sigma_b^2 & Cov(x_a,x_b) \\ Cov(x_a,x_b) & \sigma_a^2 \end{pmatrix} \end{aligned}$$

Additionally, explicit calculation for $(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$ with $D = 2$ gives:

$$\begin{pmatrix} x_a-\mu_a \\ x_b-\mu_b \end{pmatrix}^T \mathbf{\Sigma}^{-1} \begin{pmatrix} x_a-\mu_a \\ x_b-\mu_b \end{pmatrix} = \frac{\left[\sigma_b^2(x_a-\mu_a)^2 - 2Cov(x_a,x_b)(x_a-\mu_a)(x_b-\mu_b) + \sigma_a^2(x_b-\mu_b)^2\right]}{\sigma_a^2\sigma_b^2 - Cov(x_a,x_b)^2}$$

Let $\mathbf{x}$ be two dimensional. Recovery of x_a by marginalizing x_b out:

$$\begin{aligned}
P(x_a) &= \int_{-\infty}^{+\infty} (2\pi)^{-\frac{4}{2}} |\mathbf{\Sigma}|^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})} dx_b \\
&= \int_{-\infty}^{+\infty} (2\pi)^{-\frac{4}{2}} \frac{1}{\sqrt{\sigma_a^2\sigma_b^2 - Cov(x_a,x_b)^2}} e^{-\frac{\left[\sigma_b^2(x_a-\mu_a)^2 - 2Cov(x_a,x_b)(x_a-\mu_a)(x_b-\mu_b) + \sigma_a^2(x_b-\mu_b)^2\right]}{2\left(\sigma_a^2\sigma_b^2 - Cov(x_a,x_b)^2\right)}} dx_b \\
&= \frac{1}{\sqrt{2\pi\sigma_a^2}} \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{\sigma_b^2 - \frac{Cov(x_a,x_b)^2}{\sigma_a^2}}} e^{-\frac{\left[\sigma_b^2(x_a-\mu_a)^2 - 2Cov(x_a,x_b)(x_a-\mu_a)(x_b-\mu_b) + \sigma_a^2(x_b-\mu_b)^2\right]}{2\sigma_a^2\left(\sigma_b^2 - \frac{Cov(x_a,x_b)^2}{\sigma_a^2}\right)}} dx_b
\end{aligned}$$

Let $\sigma_b^2 - \frac{Cov(x_a,x_b)^2}{\sigma_a^2} = \sigma^2$. Then:

$$\begin{aligned}
P(x_a) &= \frac{1}{\sqrt{2\pi\sigma_a^2}} \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\left[\sigma_b^2(x_a-\mu_a)^2 - 2Cov(x_a,x_b)(x_a-\mu_a)(x_b-\mu_b) + \sigma_a^2(x_b-\mu_b)^2\right]}{2\sigma_a^2\sigma^2}} dx_b \\
&= \frac{1}{\sqrt{2\pi\sigma_a^2}} e^{-\frac{1}{2\sigma_a^2\sigma^2}\sigma_b^2(x_a-\mu_a)^2} \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\left[-2Cov(x_a,x_b)(x_a-\mu_a)(x_b-\mu_b) + \sigma_a^2(x_b-\mu_b)^2\right]}{2\sigma_a^2\sigma^2}} dx_b \\
&= \frac{1}{\sqrt{2\pi\sigma_a^2}} e^{-\frac{1}{2\sigma_a^2\sigma^2}\sigma_b^2(x_a-\mu_a)^2} \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\left[-2\frac{Cov(x_a,x_b)}{\sigma_a^2}(x_a-\mu_a)(x_b-\mu_b) + (x_b-\mu_b)^2\right]}{2\sigma^2}} dx_b
\end{aligned}$$

Let us isolate the first exponent and simplify it:

$$
\begin{aligned}
-\frac{1}{2\sigma_a^2\sigma^2}\sigma_b^2(x_a-\mu_a)^2 &= -\frac{\sigma_b^2(x_a-\mu_a)^2}{2\sigma_a^2\left(\sigma_b^2-\frac{Cov(x_a,x_b)^2}{\sigma_a^2}\right)} \\
&= -\frac{\sigma_b^2(x_a-\mu_a)^2}{2\sigma_a^2\sigma_b^2\left(1-\frac{Cov(x_a,x_b)^2}{\sigma_a^2\sigma_b^2}\right)} \\
&= -\frac{\sigma_b^2(x_a-\mu_a)^2-\frac{Cov(x_a,x_b)^2}{\sigma_a^2}(x_a-\mu_a)^2+\frac{Cov(x_a,x_b)^2}{\sigma_a^2}(x_a-\mu_a)^2}{2\sigma_a^2\sigma_b^2\left(1-\frac{Cov(x_a,x_b)^2}{\sigma_a^2\sigma_b^2}\right)} \\
&= -\frac{(x_a-\mu_a)^2\left[1-\frac{Cov(x_a,x_b)^2}{\sigma_a^2\sigma_b^2}\right]+\frac{Cov(x_a,x_b)^2}{\sigma_a^2\sigma_b^2}(x_a-\mu_a)^2}{2\sigma_a^2\left(1-\frac{Cov(x_a,x_b)^2}{\sigma_a^2\sigma_b^2}\right)} \\
&= -\frac{1}{2}\left[(x_a-\mu_a)^2+\frac{Cov(x_a,x_b)^2}{\sigma_a^2\sigma_a^2\sigma^2}(x_a-\mu_a)^2\right]
\end{aligned}
$$

Continuation of previous marginalization:

$$
\begin{aligned}
P(x_a) &= \frac{1}{\sqrt{2\pi\sigma_a^2}}e^{-\frac{1}{2}\left[(x_a-\mu_a)^2+\frac{Cov(x_a,x_b)^2}{\sigma_a^2\sigma_a^2\sigma^2}(x_a-\mu_a)^2\right]}\int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{\left[-2\frac{Cov(x_a,x_b)}{\sigma_a^2}(x_a-\mu_a)(x_b-\mu_b)+(x_b-\mu_b)^2\right]}{2\sigma^2}}dx_b \\
&= \frac{1}{\sqrt{2\pi\sigma_a^2}}e^{-\frac{1}{2}(x_a-\mu_a)^2}\int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{\left[\frac{Cov(x_a,x_b)^2}{\sigma_a^2\sigma_a^2}(x_a-\mu_a)^2-2\frac{Cov(x_a,x_b)}{\sigma_a^2}(x_a-\mu_a)(x_b-\mu_b)+(x_b-\mu_b)^2\right]}{2\sigma^2}}dx_b \\
&= N(x_a;\mu_a,\sigma_a^2)\int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{\left((x_b-\mu_b)-\frac{Cov(x_a,x_b)}{\sigma_a^2}(x_a-\mu_a)\right)^2}{2\sigma^2}}dx_b \\
&= N(x_a;\mu_a,\sigma_a^2)\underbrace{\int_{-\infty}^{+\infty}N(x_b;\mu,\sigma^2)dx_b}_{=1} \\
&= N(x_a;\mu_a,\sigma_a^2)
\end{aligned}
$$

2A-4.

Let the $L(\mu,\sigma)$ be defined as:

$$
L(\mu,\sigma)=\prod_{n=1}^{N}p(x_n;\mu,\sigma^2)=(2\pi\sigma^2)^{-\frac{N}{2}}e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i-\mu)^2}
$$

We try to find values for which the gradient $\nabla L(\mu, \sigma)$ vanishes. Afterwards, we verify if the found solutions are local/global maxima.

$$\boxed{\begin{array}{l} \mathit{Maximize}\ L(\mu, \sigma) \\ \mathit{Subject\ to}\ \nabla L(\mu, \sigma) = \left(\frac{\partial L}{\partial \mu} \frac{\partial L}{\partial \sigma}\right) = 0 \end{array}}$$

Because L is strictly positive on $\mathbb{R} \times \mathbb{R}^{+} \backslash \{0\}$, the position of extreme values is invariant under logarithmic transformations. This helps to facilitate easier differentiation, as products turn into sums:

$$\ln L(\mu, \sigma) = -\frac{N}{2} \ln 2\pi\sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{N} (x_i - \mu)^2$$

Calculation of $\frac{\partial L}{\partial \sigma}$:

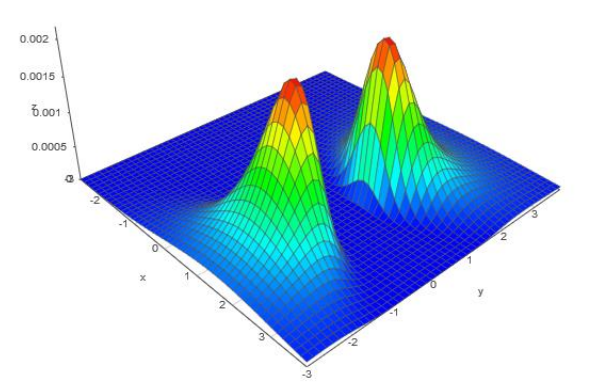
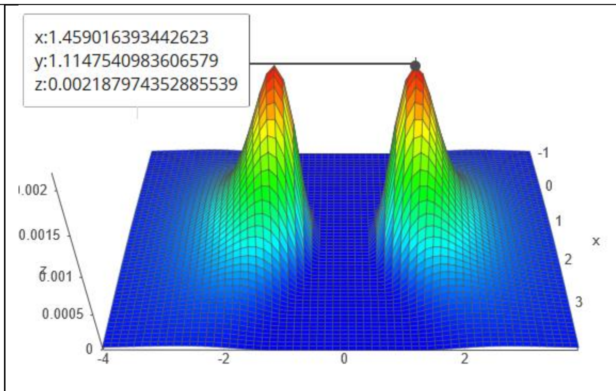
$$\begin{aligned} \frac{\partial \ln L(\sigma)}{\partial \sigma} &= 0 \\ \Leftrightarrow \quad -\frac{N}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^{N} (x_i - \mu)^2 &= 0 \\ \Leftrightarrow \quad \frac{1}{n} \sum_{i=1}^{N} (x_i - \mu)^2 &= \sigma^2 \end{aligned}$$

Calculation of $\frac{\partial L}{\partial \mu}$:

$$\begin{aligned} \frac{\partial \ln L(\sigma)}{\partial \mu} &= 0 \\ \Leftrightarrow \quad \frac{1}{\sigma^2} \sum_{i=1}^{N} (x_i - \mu) &= 0 \\ \Leftrightarrow \quad \sum_{i=1}^{N} x_i &= N\mu \\ \Leftrightarrow \quad \frac{1}{N} \sum_{i=1}^{N} x_i &= \mu \end{aligned}$$

To show that these values indeed maximize L is left as an exercise to the reader (just take the Hessian $H(\mu, \sigma)$ and see if $H\left(\frac{1}{n} \sum_{i=1}^{N} x_i\,, \sqrt{\frac{1}{n} \sum_{i=1}^{N} (x_i - \mu)^2}\right) < 0$).

For given data $\mathbf{x}$, we can visually inspect the validity of the alleged extrema.

	
Visualization of $L(\mu, \sigma)$ for the dataset $\mathbf{x} = (0,1,2,3)$ and $y = \sigma$, $x = \mu$. The sample and variance averages $\bar{\mu} = 1.5, \sigma = \sqrt{1{,}31}$ indeed	On the unconstrained domain $\mathbb{R}^2$ L reaches two symmetric maxima. The restriction $R^+ \times R$ removes negative values of σ.

Conveniently the estimators that maximize L are the sample variance and sample mean. This sample variance is biased. The unbiased estimate would be $\frac{1}{n-1}\sum_{i=1}^{N}(x_i - \mu)^2$, see Bessel's Correction.

Please remember, existence of partial derivatives does not imply that L is differentiable. But since $\frac{\partial L}{\partial \sigma}$ and $\frac{\partial L}{\partial \mu}$ are also continuous, L is differentiable.

2A-5.

Lazy version:

If we assume that Y is already normal, we only need to find the mean vector and covariance matrix:

$$\begin{aligned} E[\mathbf{y}] &= E[\mathbf{Ax} + \mathbf{b}] \\ &= \mathbf{A\mu} + \mathbf{b} \\ Var[\mathbf{y}] &= Var[\mathbf{Ax} + \mathbf{b}] \\ &= \underbrace{\mathbf{A}Var(\boldsymbol{x})\mathbf{A}^T}_{linearity} \ . \\ &= \mathbf{A\Sigma A^T} \end{aligned}$$

Complete version:

Prerequisites

We require the following properties of matrices $\mathbf{A}, \mathbf{B}$:

$$(1) \qquad (\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$$

$$(2) \quad |\mathbf{A}^{-1}| = \left|\left(\mathbf{A}^{1/2}\mathbf{A}^{1/2}\right)^{-1}\right|$$

$$(3) \qquad |\mathbf{A}| = |\mathbf{A}^T|$$

All three properties are satisfied by any invertible complex matrix.

Additionally, we use the change of variable formula:

If

$$y = g(\mathbf{x})$$

then:

$$f_Y(\mathbf{y}) = f_X\left(g^{-1}(\mathbf{y})\right) \cdot \left|\frac{dg^{-1}(y)}{dy}\right| . \mathbf{A}^{-1}(\mathbf{y} - \boldsymbol{b}) = g^{-1}(y)$$

In this case we have: $\mathbf{A}^{-1}(\mathbf{y} - \boldsymbol{b}) = g^{-1}(y)$

This can now be plugged into the pdf f_X of $\mathbf{x}$:

$$\begin{aligned} f_y(\mathbf{y}) \quad &= (2\pi)^{-\frac{D}{2}} \cdot \left|\boldsymbol{\Sigma}^{-\frac{1}{2}}\right| \cdot e^{-\frac{1}{2}\left[\left(\mathbf{A}^{-1}(\mathbf{y}-\boldsymbol{b})-\boldsymbol{\mu}\right)^T \boldsymbol{\Sigma}^{-1}\left(\mathbf{A}^{-1}(\mathbf{y}-\boldsymbol{b})-\boldsymbol{\mu}\right)\right]} \cdot \left|\mathbf{A}^{-1}\right| \\ &= (2\pi)^{-\frac{D}{2}} \cdot |\mathbf{A}^T\boldsymbol{\Sigma}\mathbf{A}|^{-\frac{1}{2}} \cdot e^{-\frac{1}{2}\left[(\mathbf{y}-\boldsymbol{b}-\boldsymbol{\mu})^T {\mathbf{A}^{-1}}^T \boldsymbol{\Sigma}^{-1}\mathbf{A}^{-1}(\mathbf{y}-\boldsymbol{b}-\boldsymbol{\mu})\right]} \\ &= N(\mathbf{y}; \boldsymbol{\mu} - \mathbf{b}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T) \end{aligned}$$