Exercise 2

Machine Learning I

2A-1.

a) Ordinary Multiplication is not defined for vectors. Alternatives exist however, see e.g. the Hadamard product.

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

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

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

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)

Just plug the previously found solution into the polynomial:

$$\begin{aligned}
\sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij}x_i x_j &= \sum_{i=1}^{D}\sum_{j=1}^{D}\left[\underbrace{\frac{1}{2}(w_{ij}+w_{ji})}_{w_{ij}^S} + \underbrace{\frac{1}{2}(w_{ij}-w_{ji})}_{w_{ij}^A}\right]x_i x_j \\
&= \sum_{i=1}^{D}\sum_{j=1}^{D}\left[\frac{1}{2}(w_{ij}+w_{ji})\right]x_i x_j + \sum_{i=1}^{D}\sum_{j=1}^{D}\left[\frac{1}{2}(w_{ij}-w_{ji})\right]x_i x_j \\
&= \sum_{i=1}^{D}\sum_{j=1}^{D}\left[\frac{1}{2}(w_{ij}+w_{ji})\right]x_i x_j + \underbrace{\frac{1}{2}\left(\sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij}x_i x_j - \sum_{i=1}^{D}\sum_{j=1}^{D} w_{ji}x_i x_j\right)}_{=0} \\
&= \sum_{i=1}^{D}\sum_{j=1}^{D}\left[\frac{1}{2}(w_{ij}+w_{ji})\right]x_i x_j
\end{aligned}$$

c)

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

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

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.

Prerequisites

Given Σ with the following block structure:

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

where $\Sigma_{11} \in \mathbb{C}^{k\times d}, \Sigma_{12} \in \mathbb{C}^{k\times n-k}, \Sigma_{21} \in \mathbb{C}^{n-k\times k}, \Sigma_{22} \in \mathbb{C}^{n-k\times n-k}$.

As seen in [1], every matrix that is in the same equivalence class as the matrix below, is a valid inverse if Σ_{11}, Σ_{22} are Hermitian (i.e. symmetric if matrices are real):

$$\Sigma^{-1} = \begin{pmatrix} \Sigma_{11}^{-1} + \Sigma_{11}^{-1}\Sigma_{12}[\Sigma_{22} - \Sigma_{12}^{*}\Sigma_{11}^{-1}\Sigma_{12}]^{-1}\Sigma_{12}^{*}\Sigma_{11}^{-1} & -\Sigma_{11}^{-1}\Sigma_{12}(\Sigma_{22} - \Sigma_{12}^{*}\Sigma_{11}^{-1}\Sigma_{12})^{-1} \\ -(\Sigma_{22} - \Sigma_{12}^{*}\Sigma_{11}^{-1}\Sigma_{12})^{-1}\Sigma_{12}^{*}\Sigma_{11}^{-1} & (\Sigma_{22} - \Sigma_{12}^{*}\Sigma_{11}^{-1}\Sigma_{12})^{-1} \end{pmatrix}$$

Note: $*$ denotes the Hermitian transpose.
Also:

$$\Sigma_{21}^{*} = \Sigma_{12}$$

Additionally, we decompose data $\mathbf{x}$ into:

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix}, \mathbf{x}_a \in \mathbb{R}^k, \mathbf{x}_b \in \mathbb{R}^{n-k}.$$

If you compare sums, you will notice:

$$\mathbf{x}_a^T \Sigma_{21} \mathbf{x}_b = \mathbf{x}_b^T \Sigma_{21}^T \mathbf{x}_a = \mathbf{x}_b^T \Sigma_{12} \mathbf{x}_a$$

For any vector $\mathbf{x}$ with aforementioned dimensions.

Let us first focus on $\left(\mathbf{x}_a^T - \boldsymbol{\mu}_a^T, \mathbf{x}_b^T - \boldsymbol{\mu}_b^T\right)^T \boldsymbol{\Sigma}^{-1} \begin{pmatrix} \mathbf{x}_a - \boldsymbol{\mu}_a \\ \mathbf{x}_b - \boldsymbol{\mu}_b \end{pmatrix}$:

$$\begin{aligned}
(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu}) &= (\mathbf{x}-\boldsymbol{\mu})^T \begin{pmatrix} \Sigma_{11}^{-1} + \Sigma_{11}^{-1}\Sigma_{12}[\Sigma_{22} - \Sigma_{12}^{*}\Sigma_{11}^{-1}\Sigma_{12}]^{-1}\Sigma_{12}^{*}\Sigma_{11}^{-1} & -\Sigma_{11}^{-1}\Sigma_{12}(\Sigma_{22} - \Sigma_{12}^{*}\Sigma_{11}^{-1}\Sigma_{12}) \\ -(\Sigma_{22} - \Sigma_{12}^{*}\Sigma_{11}^{-1}\Sigma_{12})\Sigma_{12}^{*}\Sigma_{11}^{-1} & (\Sigma_{22} - \Sigma_{12}^{*}\Sigma_{11}^{-1}\Sigma_{12})^{-1} \end{pmatrix} (\mathbf{x}-\boldsymbol{\mu}) \\
&= \mathbf{x}_a^T[\Sigma_{11}^{-1} + \Sigma_{11}^{-1}\Sigma_{12}[\Sigma_{22} - \Sigma_{12}^{*}\Sigma_{11}^{-1}\Sigma_{12}]^{-1}\Sigma_{12}^{*}\Sigma_{11}^{-1}]\mathbf{x}_a + 2\mathbf{x}_b^T[-\Sigma_{11}^{-1}\Sigma_{12}(\Sigma_{22} - \Sigma_{12}^{*}\Sigma_{11}^{-1}\Sigma_{12})]\mathbf{x}_a \\
&\quad -2\boldsymbol{\mu}_a[-\Sigma_{11}^{-1}\Sigma_{12}(\Sigma_{22} - \Sigma_{12}^{*}\Sigma_{11}^{-1}\Sigma_{12})]\boldsymbol{\mu}_b + \mathbf{x}_b^T(\Sigma_{22} - \Sigma_{12}^{*}\Sigma_{11}^{-1}\Sigma_{12})^{-1}\mathbf{x}_b \\
&= \mathbf{x}_a^T\Sigma_{11}^{-1}\mathbf{x}_a + \mathbf{x}_a^T[\Sigma_{11}^{-1}\Sigma_{12}[\Sigma_{22} - \Sigma_{12}^{*}\Sigma_{11}^{-1}\Sigma_{12}]^{-1}\Sigma_{12}^{*}\Sigma_{11}^{-1}]\mathbf{x}_a \\
&\quad -2(\mathbf{x}_b - \boldsymbol{\mu}_b)^T[\Sigma_{11}^{-1}\Sigma_{12}(\Sigma_{22} - \Sigma_{12}^{*}\Sigma_{11}^{-1}\Sigma_{12})](\mathbf{x}_b - \boldsymbol{\mu}_b) \\
&\quad +(\mathbf{x}_b - \boldsymbol{\mu}_b)^T(\Sigma_{22} - \Sigma_{12}^{*}\Sigma_{11}^{-1}\Sigma_{12})^{-1}(\mathbf{x}_b - \boldsymbol{\mu}_b) \\
&= \mathbf{x}_a^T\Sigma_{11}^{-1}\mathbf{x}_a \\
&\quad +[(\mathbf{x}_b - \boldsymbol{\mu}_b) - \Sigma_{12}^{*}\Sigma_{11}^{-1}(\mathbf{x}_a - \boldsymbol{\mu}_a)]^T(\Sigma_{22} - \Sigma_{12}^{*}\Sigma_{11}^{-1}\Sigma_{12})^{-1}[(\mathbf{x}_b - \boldsymbol{\mu}_b) - \Sigma_{12}^{*}\Sigma_{11}^{-1}(\mathbf{x}_a - \boldsymbol{\mu}_a)].
\end{aligned}$$

Now that the exponents are separated, they can be factored out of the integral:

$$p(\mathbf{x}_a) \propto e^{-\frac{1}{2}[\mathbf{x}_a^T\Sigma_{11}^{-1}\mathbf{x}_a]} + \int_{\Omega} e^{-\frac{1}{2}[(\mathbf{x}_b-\boldsymbol{\mu}_b)-\Sigma_{12}^{*}\Sigma_{11}^{-1}(\mathbf{x}_a-\boldsymbol{\mu}_a)]^T(\Sigma_{22}-\Sigma_{12}^{*}\Sigma_{11}^{-1}\Sigma_{12})^{-1}[(\mathbf{x}_b-\boldsymbol{\mu}_b)-\Sigma_{12}^{*}\Sigma_{11}^{-1}(\mathbf{x}_a-\boldsymbol{\mu}_a)]} d\Omega.$$

This can now be normalized by the appropriate factors so that the integral is and our resulting distribution is normal.

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.

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

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

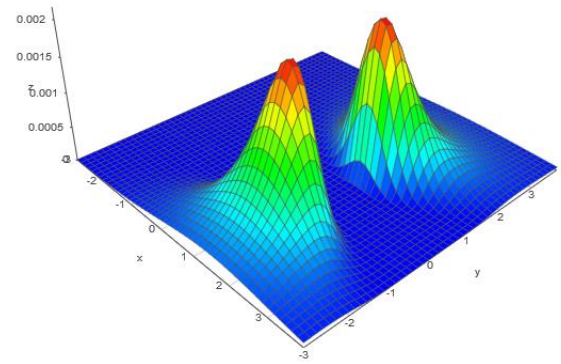
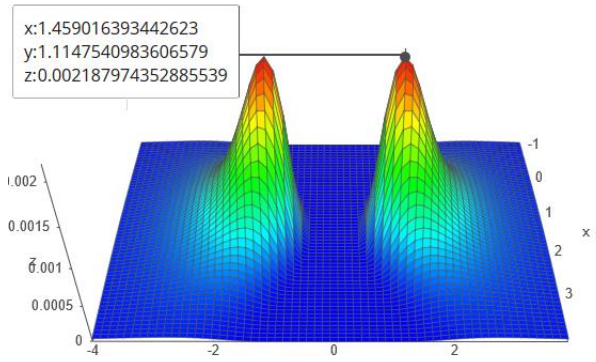
$$\begin{aligned} & \frac{\partial \ln L(\sigma)}{\partial \sigma} && = 0 \\ \Leftrightarrow \quad & -\frac{N}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{N}(x_i - \mu)^2 && = 0 \\ \Leftrightarrow \quad & \frac{1}{n}\sum_{i=1}^{N}(x_i - \mu)^2 && = \sigma^2 \end{aligned}$$

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 represent the maxima.	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 $\mathbf{A}, \mathbf{B}$:

$$(1) \quad (\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$$
$$(2) \quad |\mathbf{A}^{-1}| = \left|\left(\mathbf{A}^{1/2}\mathbf{A}^{1/2}\right)^{-1}\right|$$
$$(3) \quad |\mathbf{A}| = |\mathbf{A}^T|$$

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 $\mathbf{x}$:

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