

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

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

2A-2.

a)

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

$$\mathbf{A} = \mathbf{W}_S + \mathbf{W}_a$$

where \mathbf{W}_S is a symmetric matrix and \mathbf{W}_a is skew symmetric.

We basically have the constraints:

$$w_{ij} = w_{ij}^S + w_{ij}^A$$

$$w_{ji} = w_{ji}^S + w_{ji}^A$$

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 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:

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

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:

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

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

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

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

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

Let us isolate the first exponent and simplify it:

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

Continuation of previous marginalization:

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

2 A -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.

$ \begin{aligned} &\text{Maximize } L(\mu, \sigma) \\ &\text{Subject to } \nabla L(\mu, \sigma) = \left(\frac{\partial L}{\partial \mu} \frac{\partial L}{\partial \sigma} \right) = 0 \end{aligned} $
--

Because L is strictly positive on $\mathbb{R} \times \mathbb{R}^+ \setminus \{0\}$, 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 $\frac{\partial L}{\partial \sigma}$:

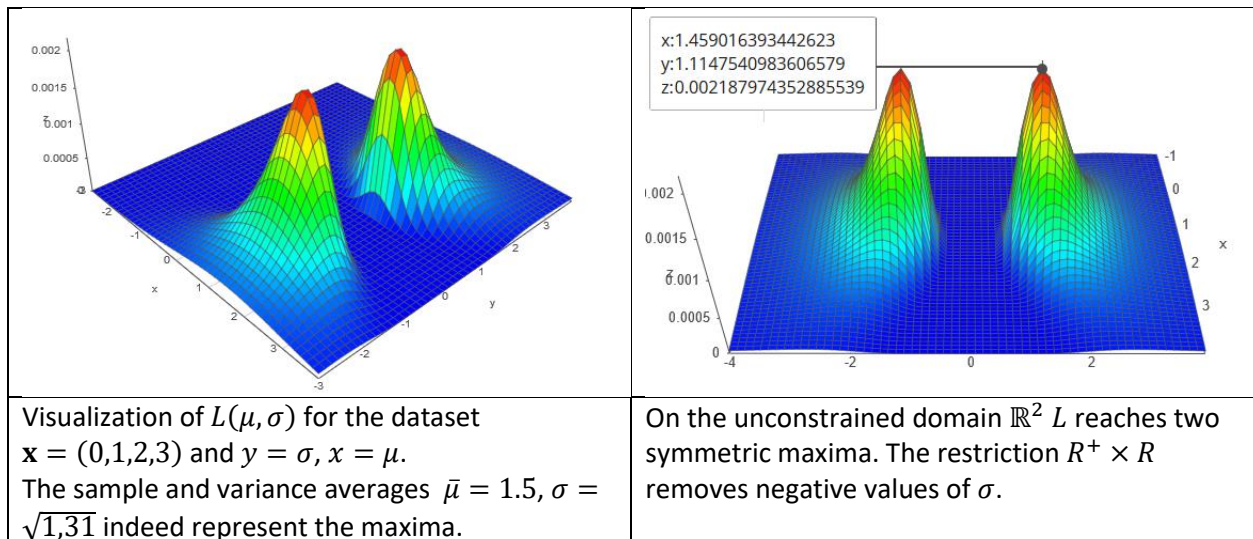
$$\begin{aligned}
& \frac{\partial \ln L(\sigma)}{\partial \sigma} = 0 \\
\Leftrightarrow & -\frac{N}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^N (x_i - \mu)^2 = 0 \\
\Leftrightarrow & \frac{1}{n} \sum_{i=1}^N (x_i - \mu)^2 = \sigma^2
\end{aligned}$$

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

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

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.



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:

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

Complete version:

Prerequisites

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

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

All three properties are satisfied by any invertible complex matrix.

Additionally, we use the change of variable formula:

If

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

then:

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

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

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

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