University of Cyprus
MAI613 - Research Methodologies and Professional Practices in AI
Lecturer: Stelios Timotheou

Exercises on Probability and Distributions

Exercises

Solve Exercises 6.1, 6.2a, 6.2b and 6.2c (compute only the mean not the mode and the median), 6.4, 6.5, 6.6, and 6.12 from the book:

- M.P. Deisenroth, A. A. Faisal, and C. S. Ong, "Mathematics for Machine Learning," Cambridge University Press, 2020. https://mml-book.github.io/book/mml-book.pdf

Note: These exercises are intended for self-assessment purposes. Their solutions will be uploaded in one week from now.

6

Probability and Distributions

Exercises

6.1 Consider the following bivariate distribution $p(x, y)$ of two discrete random variables X and Y.

Y \ X	x_1	x_2	x_3	x_4	x_5
y_1	0.01	0.02	0.03	0.1	0.1
y_2	0.05	0.1	0.05	0.07	0.2
y_3	0.1	0.05	0.03	0.05	0.04

Compute:

a. The marginal distributions $p(x)$ and $p(y)$.
b. The conditional distributions $p(x|Y = y_1)$ and $p(y|X = x_3)$.

The marginal and conditional distributions are given by

$$
\begin{aligned}
p(x) &= [0.16, 0.17, 0.11, 0.22, 0.34]^\top \\
p(y) &= [0.26, 0.47, 0.27]^\top \\
p(x|Y = y_1) &= [0.01, 0.02, 0.03, 0.1, 0.1]^\top \\
p(y|X = x_3) &= [0.03, 0.05, 0.03]^\top .
\end{aligned}
$$

6.2 Consider a mixture of two Gaussian distributions (illustrated in Figure 6.4),

$$
0.4\mathcal{N}\left(\begin{bmatrix}10\\2\end{bmatrix}, \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}\right) + 0.6\mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}8.4 & 2.0\\2.0 & 1.7\end{bmatrix}\right).
$$

a. Compute the marginal distributions for each dimension.
b. Compute the mean, mode and median for each marginal distribution.
c. Compute the mean and mode for the two-dimensional distribution.

Consider the mixture of two Gaussians,

$$
p\left(\begin{bmatrix}x_1\\x_2\end{bmatrix}\right) = 0.4\mathcal{N}\left(\begin{bmatrix}10\\2\end{bmatrix}, \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}\right) + 0.6\mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}8.4 & 2.0\\2.0 & 1.7\end{bmatrix}\right).
$$

a. Compute the marginal distribution for each dimension.

$$p(x_1) = \int 0.4\left(\mathcal{N}\left(\begin{bmatrix}10\\2\end{bmatrix}, \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}\right) + 0.6\mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}8.4 & 2.0\\2.0 & 1.7\end{bmatrix}\right)\right) \mathrm{d}x_2 \tag{6.1}$$

$$= 0.4\int \mathcal{N}\left(\begin{bmatrix}10\\2\end{bmatrix}, \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}\right) \mathrm{d}x_2 + 0.6\int \mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}8.4 & 2.0\\2.0 & 1.7\end{bmatrix}\right) \mathrm{d}x_2 \tag{6.2}$$

$$= 0.4\mathcal{N}\big(10,\, 1\big) + 0.6\mathcal{N}\big(0,\, 8.4\big)\,.$$

From (6.1) to (6.2), we used the result that the integral of a sum is the sum of the integrals.
Similarly, we can get

$$p(x_2) = 0.4\mathcal{N}\big(2,\, 1\big) + 0.6\mathcal{N}\big(0,\, 1.7\big)\,.$$

b. Compute the mean, mode and median for each marginal distribution.
Mean:

$$\mathbb{E}(x_1) = \int x_1 p(x_1)\mathrm{d}x_1$$

$$= \int x_1 \left(0.4\mathcal{N}\big(x_1 \,|\, 10,\, 1\big) + 0.6\mathcal{N}\big(x_1 \,|\, 0,\, 8.4\big)\right) \mathrm{d}x_1 \tag{6.3}$$

$$= 0.4\int x_1\mathcal{N}\big(x_1 \,|\, 10,\, 1\big)\mathrm{d}x_1 + 0.6\int x_1\mathcal{N}\big(x_1 \,|\, 0,\, 8.4\big)\mathrm{d}x_1 \tag{6.4}$$

$$= 0.4 \cdot 10 + 0.6 \cdot 0 = 4\,.$$

From step (6.3) to step (6.4), we use the fact that for $Y \sim \mathcal{N}\big(\mu,\, \sigma\big)$, where $\mathbb{E}[Y] = \mu$.
Similarly,

$$\mathbb{E}(x_2) = 0.4 \cdot 2 + 0.6 \cdot 0 = 0.8\,.$$

Mode: In principle, we would need to solve for

$$\frac{\mathrm{d}p(x_1)}{\mathrm{d}x_1} = 0 \quad \text{and} \quad \frac{\mathrm{d}p(x_2)}{\mathrm{d}x_2} = 0\,.$$

However, we can observe that the modes of each individual distribution are the peaks of the Gaussians, that is the Gaussian means for each dimension. Median:

$$\int_{-\infty}^{a} p(x_1)\mathrm{d}x_1 = \frac{1}{2}\,.$$

c. Compute the mean and mode for the 2 dimensional distribution.
Mean:
From (6.30), we know that

$$\mathbb{E}[\begin{bmatrix}x_1\\x_2\end{bmatrix}] = \begin{bmatrix}\mathbb{E}[x_1]\\\mathbb{E}[x_2]\end{bmatrix} = \begin{bmatrix}4\\0.8\end{bmatrix}\,.$$

Mode:
The two dimensional distribution has two peaks, and hence there are

two modes. In general, we would need to solve an optimization problem to find the maxima. However, in this particular case we can observe that the two modes correspond to the individual Gaussian means, that is

$$\begin{bmatrix}10\\2\end{bmatrix} \text{ and } \begin{bmatrix}0\\0\end{bmatrix}.$$

6.4 There are two bags. The first bag contains four mangos and two apples; the second bag contains four mangos and four apples.
We also have a biased coin, which shows "heads" with probability 0.6 and "tails" with probability 0.4. If the coin shows "heads". we pick a fruit at random from bag 1; otherwise we pick a fruit at random from bag 2.
Your friend flips the coin (you cannot see the result), picks a fruit at random from the corresponding bag, and presents you a mango.
What is the probability that the mango was picked from bag 2?
Hint: Use Bayes' theorem.
We apply Bayes' theorem and compute the posterior $p(b_2 \mid m)$ of picking a mango from bag 2.

$$p(b_2 \mid m) = \frac{p(m \mid b_2)p(b_2)}{p(m)}$$

where

$$p(m) = p(b_1)p(m \mid b_1) + p(b_2)p(m \mid b_2) = \frac{3}{5}\frac{2}{3} + \frac{2}{5}\frac{1}{2} = \frac{2}{5} + \frac{1}{5} = \frac{3}{5} \quad \text{Evidence}$$

$$p(b_2) = \tfrac{2}{5} \quad \text{Prior}$$

$$p(m \mid b_2) = \tfrac{1}{2} \quad \text{Likelihood}$$

Therefore,

$$p(b_2 \mid m) = \frac{p(m \mid b_2)p(b_2)}{p(m)} = \frac{\frac{2}{5}\frac{1}{2}}{\frac{3}{5}} = \frac{1}{3}\,.$$

6.5 Consider the time-series model

$$\begin{aligned} \boldsymbol{x}_{t+1} &= \boldsymbol{A}\boldsymbol{x}_t + \boldsymbol{w}\,, \quad \boldsymbol{w} \sim \mathcal{N}\big(\boldsymbol{0},\, \boldsymbol{Q}\big) \\ \boldsymbol{y}_t &= \boldsymbol{C}\boldsymbol{x}_t + \boldsymbol{v}\,, \quad \boldsymbol{v} \sim \mathcal{N}\big(\boldsymbol{0},\, \boldsymbol{R}\big)\,, \end{aligned}$$

where $\boldsymbol{w}, \boldsymbol{v}$ are i.i.d. Gaussian noise variables. Further, assume that $p(\boldsymbol{x}_0) = \mathcal{N}\big(\boldsymbol{\mu}_0,\, \boldsymbol{\Sigma}_0\big)$.

a. What is the form of $p(\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_T)$? Justify your answer (you do not have to explicitly compute the joint distribution).
b. Assume that $p(\boldsymbol{x}_t \mid \boldsymbol{y}_1, \ldots, \boldsymbol{y}_t) = \mathcal{N}\big(\boldsymbol{\mu}_t,\, \boldsymbol{\Sigma}_t\big)$.
   1. Compute $p(\boldsymbol{x}_{t+1} \mid \boldsymbol{y}_1, \ldots, \boldsymbol{y}_t)$.
   2. Compute $p(\boldsymbol{x}_{t+1}, \boldsymbol{y}_{t+1} \mid \boldsymbol{y}_1, \ldots, \boldsymbol{y}_t)$.
   3. At time $t{+}1$, we observe the value $\boldsymbol{y}_{t+1} = \hat{\boldsymbol{y}}$. Compute the conditional distribution $p(\boldsymbol{x}_{t+1} \mid \boldsymbol{y}_1, \ldots, \boldsymbol{y}_{t+1})$.

1. The joint distribution is Gaussian: $p(\boldsymbol{x}_0)$ is Gaussian, and $\boldsymbol{x}_{t+1}$ is a linear/affine transformations of $\boldsymbol{x}_t$. Since affine transformations leave the Gaussianity of the random variable invariant, the joint distribution must be Gaussian.
2. 1. We use the results from linear transformation of Gaussian random variables (Section 6.5.3). We immediately obtain

$$\begin{aligned} p(\boldsymbol{x}_{t+1} \mid \boldsymbol{y}_{1:t}) &= \mathcal{N}\big(\boldsymbol{x}_{t+1} \mid \boldsymbol{\mu}_{t+1\,|\,t},\, \boldsymbol{\Sigma}_{t+1\,|\,t}\big) \\ \boldsymbol{\mu}_{t+1\,|\,t} &:= \boldsymbol{A}\boldsymbol{\mu}_t \\ \boldsymbol{\Sigma}_{t+1\,|\,t} &:= \boldsymbol{A}\boldsymbol{\Sigma}_t\boldsymbol{A}^\top + \boldsymbol{Q}\,. \end{aligned}$$

   2. The joint distribution $p(\boldsymbol{x}_{t+1}, \boldsymbol{y}_{t+1} \mid \boldsymbol{y}_{1:t})$ is Gaussian (linear transformation of random variables). We compute every component of the Gaussian separately:

$$\begin{aligned} &\mathbb{E}[\boldsymbol{y}_{t+1} \mid \boldsymbol{y}_{1:t}] = \boldsymbol{C}\boldsymbol{\mu}_{t+1\,|\,t} =: \boldsymbol{\mu}^y_{t+1\,|\,t} \\ &\mathbb{V}[\boldsymbol{y}_{t+1} \mid \boldsymbol{y}_{1:t}] = \boldsymbol{C}\boldsymbol{\Sigma}_{t+1\,|\,t}\boldsymbol{C}^\top + \boldsymbol{R} =: \boldsymbol{\Sigma}^y_{t+1\,|\,t} \\ &\operatorname{Cov}[\boldsymbol{x}_{t+1}, \boldsymbol{y}_{t+1} \mid \boldsymbol{y}_{1:t}] = \operatorname{Cov}[\boldsymbol{x}_{t+1}, \boldsymbol{C}\boldsymbol{x}_{t+1} \mid \boldsymbol{y}_{1:t}] = \boldsymbol{\Sigma}_{t+1\,|\,t}\boldsymbol{C}^\top =: \boldsymbol{\Sigma}_{xy} \end{aligned}$$

   Therefore,

$$p(\boldsymbol{x}_{t+1}, \boldsymbol{y}_{t+1} \mid \boldsymbol{y}_{1:t}) = \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_{t+1\,|\,t} \\ \boldsymbol{\mu}^y_{t+1\,|\,t} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{t+1\,|\,t} & \boldsymbol{\Sigma}_{xy} \\ \boldsymbol{\Sigma}_{yx} & \boldsymbol{\Sigma}^y_{t+1\,|\,t} \end{bmatrix}.\right)$$

3. We obtain the desired distribution (again: Gaussian) by applying the rules for Gaussian conditioning of the joint distribution in B):

$$
\begin{aligned}
p(\boldsymbol{x}_{t+1} \mid \boldsymbol{y}_{1:t+1}) &= \mathcal{N}\big(\boldsymbol{\mu}_{t+1\mid t+1},\, \boldsymbol{\Sigma}_{t+1\mid t+1}\big) \\
\boldsymbol{\mu}_{t+1\mid t+1} &= \boldsymbol{\mu}_{t+1\mid t} + \boldsymbol{\Sigma}_{xy}(\boldsymbol{\Sigma}^{y}_{t+1\mid t})^{-1}(\hat{\boldsymbol{y}} - \boldsymbol{\mu}^{y}_{t+1\mid t}) \\
\boldsymbol{\Sigma}_{t+1\mid t+1} &= \boldsymbol{\Sigma}_{t+1\mid t} - \boldsymbol{\Sigma}_{xy}(\boldsymbol{\Sigma}^{y}_{t+1\mid t})^{-1}\boldsymbol{\Sigma}_{yx}
\end{aligned}
$$

6.6 Prove the relationship in (6.44), which relates the standard definition of the variance to the raw-score expression for the variance.

The formula in (6.43) can be converted to the so called raw-score formula for variance

$$
\begin{aligned}
\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2 &= \frac{1}{N}\sum_{i=1}^{N}\left(x_i^2 - 2x_i\mu + \mu^2\right) \\
&= \frac{1}{N}\sum_{i=1}^{N}x_i^2 - \frac{2}{N}\mu\sum_{i=1}^{N}x_i + \mu^2 \\
&= \frac{1}{N}\sum_{i=1}^{N}x_i^2 - \mu^2 \\
&= \frac{1}{N}\sum_{i=1}^{N}x_i^2 - \left(\frac{1}{N}\sum_{i=1}^{N}x_i\right)^2 .
\end{aligned}
$$

6.11 **Iterated Expectations.**
Consider two random variables x, y with joint distribution $p(x, y)$. Show that

$$\mathbb{E}_X[x] = \mathbb{E}_Y\left[\mathbb{E}_X[x \mid y]\right].$$

Here, $\mathbb{E}_X[x \mid y]$ denotes the expected value of x under the conditional distribution $p(x \mid y)$.

$$\begin{aligned}
&\mathbb{E}_X[x] = \mathbb{E}_Y\left[\mathbb{E}_X[x \mid y]\right] \\
&\quad \iff \int x p(x) dx = \iint x p(x \mid y) p(y) dx dy \\
&\quad \iff \int x p(x) dx = \iint x p(x, y) dx dy \\
&\quad \iff \int x p(x) dx = \int x \int p(x, y) dy dx = \int x p(x) dx\,,
\end{aligned}$$

which proves the claim.

6.12 **Manipulation of Gaussian Random Variables.**
Consider a Gaussian random variable $\boldsymbol{x} \sim \mathcal{N}\left(\boldsymbol{x} \mid \boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x\right)$, where $\boldsymbol{x} \in \mathbb{R}^D$. Furthermore, we have

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b} + \boldsymbol{w}\,,$$

where $\boldsymbol{y} \in \mathbb{R}^E$, $\boldsymbol{A} \in \mathbb{R}^{E \times D}$, $\boldsymbol{b} \in \mathbb{R}^E$, and $\boldsymbol{w} \sim \mathcal{N}\left(\boldsymbol{w} \mid \boldsymbol{0}, \boldsymbol{Q}\right)$ is independent Gaussian noise. "Independent" implies that $\boldsymbol{x}$ and $\boldsymbol{w}$ are independent random variables and that $\boldsymbol{Q}$ is diagonal.

a. Write down the likelihood $p(\boldsymbol{y} \mid \boldsymbol{x})$.

$$\begin{aligned}
p(\boldsymbol{y} \mid \boldsymbol{x}) &= \mathcal{N}\left(\boldsymbol{y} \mid \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}, \boldsymbol{Q}\right) \\
&= Z \exp\left(-\tfrac{1}{2}(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x} - \boldsymbol{b})^\top \boldsymbol{Q}^{-1}(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x} - \boldsymbol{b})\right) \\
\text{where} \quad Z &= (2\pi)^{-E/2} \left|\boldsymbol{Q}\right|^{-1/2}
\end{aligned}$$

b. The distribution $p(\boldsymbol{y}) = \int p(\boldsymbol{y} \mid \boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x}$ is Gaussian. Compute the mean $\boldsymbol{\mu}_y$ and the covariance $\boldsymbol{\Sigma}_y$. Derive your result in detail.
$p(\boldsymbol{y})$ is Gaussian distributed (affine transformation of the Gaussian random variable $\boldsymbol{x}$) with mean $\boldsymbol{\mu}_y$ and covariance matrix $\boldsymbol{\Sigma}_y$. We obtain

$$\begin{aligned}
\boldsymbol{\mu}_y &= \mathbb{E}_Y[\boldsymbol{y}] = \mathbb{E}_{X,W}[\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b} + \boldsymbol{w}] = \boldsymbol{A}\mathbb{E}_X[\boldsymbol{x}] + \boldsymbol{b} + \mathbb{E}_W[\boldsymbol{w}] \\
&= \boldsymbol{A}\boldsymbol{\mu}_x + \boldsymbol{b}, \quad \text{and} \\
\boldsymbol{\Sigma}_y &= \mathbb{E}_Y[\boldsymbol{y}\boldsymbol{y}^\top] - \mathbb{E}_Y[\boldsymbol{y}]\mathbb{E}_Y[\boldsymbol{y}]^\top \\
&= \mathbb{E}_{X,W}\left[(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b} + \boldsymbol{w})(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b} + \boldsymbol{w})^\top\right] - \boldsymbol{\mu}_y\boldsymbol{\mu}_y^\top \\
&= \mathbb{E}_{X,W}\left[\boldsymbol{A}\boldsymbol{x}\boldsymbol{x}^\top\boldsymbol{A}^\top + \boldsymbol{A}\boldsymbol{x}\boldsymbol{b}^\top + \boldsymbol{A}\boldsymbol{x}\boldsymbol{w}^\top + \boldsymbol{b}\boldsymbol{x}^\top\boldsymbol{A}^\top + \boldsymbol{b}\boldsymbol{b}^\top + \boldsymbol{b}\boldsymbol{w}^\top\right. \\
&\quad \left. + \boldsymbol{w}(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b})^\top + \boldsymbol{w}\boldsymbol{w}^\top)\right] - \boldsymbol{\mu}_y\boldsymbol{\mu}_y^\top .
\end{aligned}$$

We use the linearity of the expected value, move all constants out of the expected value, and exploit the independence of $\boldsymbol{w}$ and $\boldsymbol{x}$:

$$\begin{aligned}
\boldsymbol{\Sigma}_y = \boldsymbol{A}\mathbb{E}_X[\boldsymbol{x}\boldsymbol{x}^\top]\boldsymbol{A}^\top + \boldsymbol{A}\mathbb{E}_X[\boldsymbol{x}]\boldsymbol{b}^\top + \boldsymbol{b}\mathbb{E}_X[\boldsymbol{x}^\top]\boldsymbol{A}^\top + \boldsymbol{b}\boldsymbol{b}^\top + \mathbb{E}_W[\boldsymbol{w}\boldsymbol{w}^\top] \\
- (\boldsymbol{A}\boldsymbol{\mu}_x + \boldsymbol{b})(\boldsymbol{A}\boldsymbol{\mu}_x + \boldsymbol{b})^\top ,
\end{aligned}$$

where we used our previous result for $\boldsymbol{\mu}_y$. Note that $\mathbb{E}_W[\boldsymbol{w}] = \boldsymbol{0}$. We continue as follows:

$$\begin{aligned}
\boldsymbol{\Sigma}_y &= \boldsymbol{A}\mathbb{E}_X[\boldsymbol{x}\boldsymbol{x}^\top]\boldsymbol{A}^\top + \boldsymbol{A}\boldsymbol{\mu}_x\boldsymbol{b}^\top + \boldsymbol{b}\boldsymbol{\mu}_x^\top\boldsymbol{A}^\top + \boldsymbol{b}\boldsymbol{b}^\top + \boldsymbol{Q} \\
&\quad - \boldsymbol{A}\boldsymbol{\mu}_x\boldsymbol{\mu}_x^\top\boldsymbol{A}^\top - \boldsymbol{A}\boldsymbol{\mu}_x\boldsymbol{b}^\top - \boldsymbol{b}\boldsymbol{\mu}_x^\top\boldsymbol{A}^\top - \boldsymbol{b}\boldsymbol{b}^\top \\
&= \boldsymbol{A}\underbrace{\left(\mathbb{E}_X[\boldsymbol{x}\boldsymbol{x}^\top] - \boldsymbol{\mu}_x\boldsymbol{\mu}_x^\top\right)}_{=\boldsymbol{\Sigma}_x}\boldsymbol{A}^\top + \boldsymbol{Q} \\
&= \boldsymbol{A}\boldsymbol{\Sigma}_x\boldsymbol{A}^\top + \boldsymbol{Q} .
\end{aligned}$$

Alternatively, we could have exploited

$$\begin{aligned}
\mathbb{V}_{\boldsymbol{y}}[\boldsymbol{y}] &= \mathbb{V}_{\boldsymbol{x},\boldsymbol{w}}[\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b} + \boldsymbol{w}] \overset{\text{i.i.d.}}{=} \mathbb{V}_{\boldsymbol{x}}[\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}] + \mathbb{V}_{\boldsymbol{w}}[\boldsymbol{w}] \\
&= \boldsymbol{A}\mathbb{V}_{\boldsymbol{x}}\boldsymbol{A}^\top + \boldsymbol{Q} = \boldsymbol{A}\boldsymbol{\Sigma}_x\boldsymbol{A}^\top + \boldsymbol{Q} .
\end{aligned}$$

c. The random variable $\boldsymbol{y}$ is being transformed according to the measurement mapping

$$\boldsymbol{z} = \boldsymbol{C}\boldsymbol{y} + \boldsymbol{v} ,$$

where $\boldsymbol{z} \in \mathbb{R}^F$, $\boldsymbol{C} \in \mathbb{R}^{F \times E}$, and $\boldsymbol{v} \sim \mathcal{N}(\boldsymbol{v} \mid \boldsymbol{0}, \boldsymbol{R})$ is independent Gaussian (measurement) noise.

- Write down $p(\boldsymbol{z} \mid \boldsymbol{y})$.

$$\begin{aligned}
p(\boldsymbol{z} \mid \boldsymbol{y}) &= \mathcal{N}(\boldsymbol{z} \mid \boldsymbol{C}\boldsymbol{y}, \boldsymbol{R}) \\
&= (2\pi)^{-F/2} \mid \boldsymbol{R} \mid^{-1/2} \exp\left(-\tfrac{1}{2}(\boldsymbol{z} - \boldsymbol{C}\boldsymbol{y})^\top\boldsymbol{R}^{-1}(\boldsymbol{z} - \boldsymbol{C}\boldsymbol{y})\right) .
\end{aligned}$$

- Compute $p(\boldsymbol{z})$, i.e., the mean $\boldsymbol{\mu}_z$ and the covariance $\boldsymbol{\Sigma}_z$. Derive your result in detail.
  Remark: Since $\boldsymbol{y}$ is Gaussian and $\boldsymbol{z}$ is a linear transformation of $\boldsymbol{y}$, $p(\boldsymbol{z})$ is Gaussian, too. Let's compute its moments (similar to the "time update"):

$$\begin{aligned}\boldsymbol{\mu}_z &= \mathbb{E}_Z[\boldsymbol{z}] = \mathbb{E}_{Y,V}[\boldsymbol{C}\boldsymbol{y} + \boldsymbol{v}] = \boldsymbol{C}\mathbb{E}_Y[\boldsymbol{y}] + \underbrace{\mathbb{E}_V[\boldsymbol{v}]}_{=\boldsymbol{0}} \\ &= \boldsymbol{C}\boldsymbol{\mu}_y \,.\end{aligned}$$

  For the covariance matrix, we compute

$$\begin{aligned}\boldsymbol{\Sigma}_z &= \mathbb{E}_Z[\boldsymbol{z}\boldsymbol{z}^\top] - \mathbb{E}_Z[\boldsymbol{z}]\mathbb{E}_Z[\boldsymbol{z}]^\top \\ &= \mathbb{E}_{Y,V}\left[(\boldsymbol{C}\boldsymbol{y} + \boldsymbol{v})(\boldsymbol{C}\boldsymbol{y} + \boldsymbol{v})^\top\right] - \boldsymbol{\mu}_z\boldsymbol{\mu}_z^\top \\ &= \mathbb{E}_{Y,V}\left[\boldsymbol{C}\boldsymbol{y}\boldsymbol{y}^\top\boldsymbol{C}^\top + \boldsymbol{C}\boldsymbol{y}\boldsymbol{v}^\top + \boldsymbol{v}(\boldsymbol{C}\boldsymbol{y})^\top + \boldsymbol{v}\boldsymbol{v}^\top)\right] - \boldsymbol{\mu}_z\boldsymbol{\mu}_z^\top \,.\end{aligned}$$

  We use the linearity of the expected value, move all constants out of the expected value, and exploit the independence of $\boldsymbol{v}$ and $\boldsymbol{y}$:

$$\boldsymbol{\Sigma}_z = \boldsymbol{C}\mathbb{E}_Y[\boldsymbol{y}\boldsymbol{y}^\top]\boldsymbol{C}^\top + \mathbb{E}_V[\boldsymbol{v}\boldsymbol{v}^\top] - \boldsymbol{C}\boldsymbol{\mu}_y\boldsymbol{\mu}_y^\top\boldsymbol{C}^\top \,,$$

  where we used our previous result for $\boldsymbol{\mu}_z$. Note that $\mathbb{E}_V[\boldsymbol{v}] = \boldsymbol{0}$. We continue as follows:

$$\begin{aligned}\boldsymbol{\Sigma}_y &= \boldsymbol{C}\underbrace{\left(\mathbb{E}_Y[\boldsymbol{y}\boldsymbol{y}^\top] - \boldsymbol{\mu}_y\boldsymbol{\mu}_y^\top\right)}_{=\boldsymbol{\Sigma}_x}\boldsymbol{C}^\top + \boldsymbol{R} \\ &= \boldsymbol{C}\boldsymbol{\Sigma}_y\boldsymbol{C}^\top + \boldsymbol{R} \,.\end{aligned}$$

d. Now, a value $\hat{\boldsymbol{y}}$ is measured. Compute the posterior distribution $p(\boldsymbol{x} \mid \hat{\boldsymbol{y}})$.
   Hint for solution: This posterior is also Gaussian, i.e., we need to determine only its mean and covariance matrix. Start by explicitly computing the joint Gaussian $p(\boldsymbol{x}, \boldsymbol{y})$. This also requires us to compute the cross-covariances $\mathrm{Cov}_{\boldsymbol{x},\boldsymbol{y}}[\boldsymbol{x}, \boldsymbol{y}]$ and $\mathrm{Cov}_{\boldsymbol{y},\boldsymbol{x}}[\boldsymbol{y}, \boldsymbol{x}]$. Then apply the rules for Gaussian conditioning.
   We derive the posterior distribution following the second hint since we do not have to worry about normalization constants:
   Assume, we know the joint Gaussian distribution

$$p(\boldsymbol{x}, \boldsymbol{y}) = \mathcal{N}\left(\begin{bmatrix}\boldsymbol{x}\\ \boldsymbol{y}\end{bmatrix} \,\middle|\, \begin{bmatrix}\boldsymbol{\mu}_x\\ \boldsymbol{\mu}_y\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_x & \boldsymbol{\Sigma}_{xy}\\ \boldsymbol{\Sigma}_{yx} & \boldsymbol{\Sigma}_y\end{bmatrix}\right) , \tag{6.11}$$

   where we defined $\boldsymbol{\Sigma}_{xy} := \mathrm{Cov}_{\boldsymbol{x},\boldsymbol{y}}[\boldsymbol{x}, \boldsymbol{y}]$.
   Now, we apply the rules for Gaussian conditioning to obtain

$$\begin{aligned}p(\boldsymbol{x} \mid \hat{\boldsymbol{y}}) &= \mathcal{N}\big(\boldsymbol{x} \mid \boldsymbol{\mu}_{x\mid y}, \boldsymbol{\Sigma}_{x\mid y}\big) \\ \boldsymbol{\mu}_{x\mid y} &= \boldsymbol{\mu}_x + \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_y^{-1}(\hat{\boldsymbol{y}} - \boldsymbol{\mu}_y) \\ \boldsymbol{\Sigma}_{x\mid y} &= \boldsymbol{\Sigma}_x - \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_y^{-1}\boldsymbol{\Sigma}_{yx} \,.\end{aligned}$$

Looking at (6.11), it remains to compute the cross-covariance term $\boldsymbol{\Sigma}_{xy}$ (the marginal distributions $p(\boldsymbol{x})$ and $p(\boldsymbol{y})$ are known and $\boldsymbol{\Sigma}_{yx} = \boldsymbol{\Sigma}_{xy}^\top$):

$$\boldsymbol{\Sigma}_{xy} = \mathrm{Cov}[\boldsymbol{x}, \boldsymbol{y}] = \mathbb{E}_{X,Y}[\boldsymbol{x}\boldsymbol{y}^\top] - \mathbb{E}_X[\boldsymbol{x}]\mathbb{E}_Y[\boldsymbol{y}]^\top = \mathbb{E}_{X,Y}[\boldsymbol{x}\boldsymbol{y}^\top] - \boldsymbol{\mu}_x\boldsymbol{\mu}_y^\top ,$$

where $\boldsymbol{\mu}_x$ and $\boldsymbol{\mu}_y$ are known. Hence, it remains to compute

$$\begin{aligned}\mathbb{E}_{X,Y}[\boldsymbol{x}\boldsymbol{y}^\top] &= \mathbb{E}_X[\boldsymbol{x}(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b} + \boldsymbol{w})^\top] = \mathbb{E}_X[\boldsymbol{x}\boldsymbol{x}^\top]\boldsymbol{A}^\top + \mathbb{E}_X[\boldsymbol{x}]\boldsymbol{b}^\top \\ &= \mathbb{E}_X[\boldsymbol{x}\boldsymbol{x}^\top]\boldsymbol{A}^\top + \boldsymbol{\mu}_x\boldsymbol{b}^\top .\end{aligned}$$

With

$$\boldsymbol{\mu}_x\boldsymbol{\mu}_y^\top = \boldsymbol{\mu}_x\boldsymbol{\mu}_x^\top\boldsymbol{A}^\top + \boldsymbol{\mu}_x\boldsymbol{b}^\top$$

we obtain the desired cross-covariance

$$\begin{aligned}\boldsymbol{\Sigma}_{xy} &= \mathbb{E}_X[\boldsymbol{x}\boldsymbol{x}^\top]\boldsymbol{A}^\top + \boldsymbol{\mu}_x\boldsymbol{b}^\top - \boldsymbol{\mu}_x\boldsymbol{\mu}_x^\top\boldsymbol{A}^\top - \boldsymbol{\mu}_x\boldsymbol{b}^\top \\ &= \big(\mathbb{E}_X[\boldsymbol{x}\boldsymbol{x}^\top] - \boldsymbol{\mu}_x\boldsymbol{\mu}_x^\top\big)\boldsymbol{A}^\top = \boldsymbol{\Sigma}_x\boldsymbol{A}^\top .\end{aligned}$$

And finally

$$\begin{aligned}\boldsymbol{\mu}_{x\,|\,y} &= \boldsymbol{\mu}_x + \boldsymbol{\Sigma}_x\boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{\Sigma}_x\boldsymbol{A}^\top + \boldsymbol{Q})^{-1}(\hat{\boldsymbol{y}} - \boldsymbol{A}\boldsymbol{\mu}_x - \boldsymbol{b}) , \\ \boldsymbol{\Sigma}_{x\,|\,y} &= \boldsymbol{\Sigma}_x - \boldsymbol{\Sigma}_x\boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{\Sigma}_x\boldsymbol{A}^\top + \boldsymbol{Q})^{-1}\boldsymbol{A}\boldsymbol{\Sigma}_x .\end{aligned}$$