Statistical Machine Learning 2018

Assignment and answers 2
Deadline: 28th of October 2018

Instructions:

- You can work **alone or in pairs** (= max 2 people). **Write the full name and S/U-number of all team members on the first page of the report.**
- Write a **self-contained report** with the answers to each question, **including** comments, derivations, explanations, graphs, etc. This means that the elements and/or intermediate steps required to derive the answer have to be in the report. (Answers like 'No' or 'x=27.2' by themselves are not sufficient, even when they are the result of running your code.)
- If an exercise specifically asks for code, put **essential code snippets** in your answer to the question in the report, and explain briefly what the code does. In addition, hand in **complete (working and documented) source code** (MATLAB recommended, other languages are allowed but not "supported").
- In order to avoid extremely verbose or poorly formatted reports, we impose a **maximum page limit** of 20 pages, including plots and code, with the following formatting: fixed **font size** of 11pt on an **A4 paper**; **margins** fixed to 2cm on all sides. All figures should have axis labels and a caption or title that states to which exercise (and part) they belong.
- Upload reports to **Brightspace** as a **single pdf** file: '`SML_A2_<Namestudent(s)>.pdf`' and one zip-file with the executable source/data files (e.g. matlab m-files). For those working in pairs, only one team member should upload the solutions.
- Assignment 2 consists of 3 exercises, weighted as follows: 3 points, 2 points, and 5 points. The **grading** will be based solely on the report pdf file. The source files are considered supplementary material (e.g. to verify that you indeed did the coding).
- For any problems or questions, send us an email, or just ask.
  Email addresses: `tomc@cs.ru.nl` and `b.kappen@science.ru.nl`

Exercise 1 – weight 3

The financial services department of an insurance company receives numerous phone calls each day from people who want to make a claim against their policy. Most claims are genuine, however about 1 out of every 6 are thought to be fraudulent. To tackle this problem the company has installed a trial version of a software voice-analysis system that monitors each conversation and gives a numerical score z between 0 and 1, depending on allegedly suspicious vocal intonations of the customer. Unfortunately, nobody seems to know anymore how to interpret the score in this particular version of the system ...

Tests revealed that the conditional probability density of z, given that a claim was valid ($c = 1$) or false ($c = 0$) are

$$\begin{aligned} p(z|c=0) &= \alpha_0(1-z^2)\,, \\ p(z|c=1) &= \alpha_1 z(z+1)\,. \end{aligned}$$

1. Compute the normalization constants α_0 and α_1. How does the z score relate to the validity of the claim? What values for z would you expect when the claim is valid / false?

   ANSWER: The normalization constants come from $\int p(z|c=0)\mathrm{d}z = \int p(z|c=1)\mathrm{d}z = 1$.

   $$\int p(z|c=0)\mathrm{d}z = \alpha_0 \int_0^1 (1-z^2)\mathrm{d}z = \alpha_0 \left[z - \frac{1}{3}z^3\right]_0^1 = \alpha_0 \left[(1-\frac{1}{3}) - (0-0)\right] = \alpha_0 \frac{2}{3} = 1$$
   $$\int p(z|c=1)\mathrm{d}z = \alpha_1 \int_0^1 (z^2+z)\mathrm{d}z = \alpha_1 \left[\frac{1}{3}z^3 + \frac{1}{2}z^2\right]_0^1 = \alpha_1 \left[(\frac{1}{3}+\frac{1}{2}) - (0+0)\right] = \alpha_1 \frac{5}{6} = 1$$

   We conclude that $\alpha_0 = \frac{3}{2}$ and $\alpha_1 = \frac{6}{5}$.

   If we plot the conditional probability densities of z given c, like in Figure 1, we notice that when the claim is false ($c = 0$), the density of z is decreasing from $z = 0$ to $z = 1$. In contrast, when the claim is valid ($c = 1$), the density of z is increasing from $z = 0$ to $z = 1$. This suggests that the test score is likely to have higher values if the claim is valid and likely to have lower values if the claim is false. We can confirm this by looking at the expected values of z:

   $$\mathbb{E}[Z|c=0] = \int z p(z|c=0)\mathrm{d}z = \frac{3}{2}\int_0^1 (z-z^3)\mathrm{d}z = \frac{3}{2}\left[\frac{1}{2}z^2 - \frac{1}{4}z^4\right]_0^1 = \frac{3}{8}$$
   $$\mathbb{E}[Z|c=1] = \int z p(z|c=1)\mathrm{d}z = \frac{6}{5}\int_0^1 (z^3+z^2)\mathrm{d}z = \frac{6}{5}\left[\frac{1}{4}z^4 + \frac{1}{3}z^3\right]_0^1 = \frac{7}{10}$$

   We see that the expected value of Z given $c = 1$ is higher than the expected value of Z given $c = 0$, in agreement with our intuition.

2. Use the sum and product rule to show that the probability distribution function $p(z)$ can be written as

   $$p(z) = \frac{(3z+1)(z+1)}{4}. \tag{1}$$

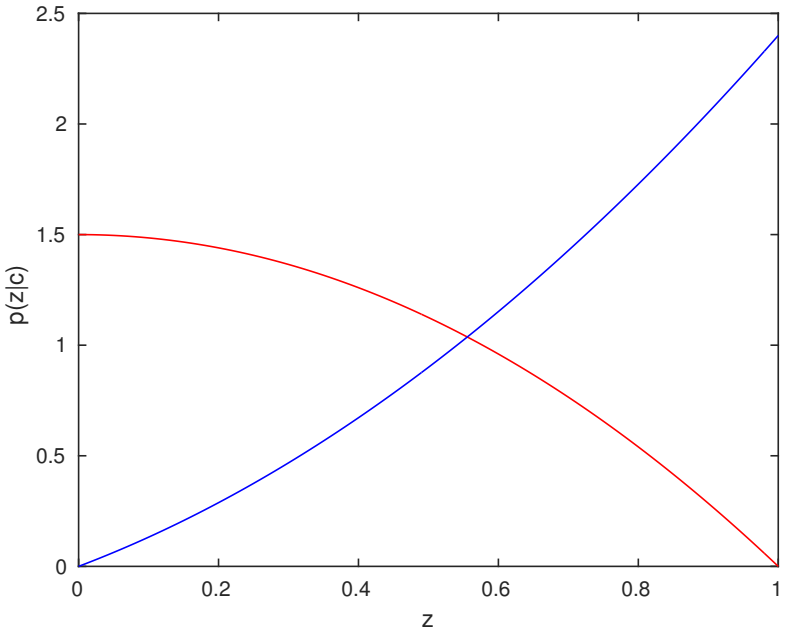

Figure 1: The red line corresponds to the conditional probability density $p(z|c=0)$, while the blue line corresponds to the conditional probability density $p(z|c=1)$.

ANSWER: From the description, we have $p(c=0)=\frac{1}{6}$ and $p(c=1)=\frac{5}{6}$, giving

$$\begin{aligned}
p(z) &= \sum_c p(z|c)p(c) \\
&= p(z|c=0)p(c=0)+p(z|c=1)p(c=1) \\
&= \frac{3}{2}(1-z^2)\cdot\frac{1}{6}+\frac{6}{5}z(z+1)\cdot\frac{5}{6} \\
&= \frac{1}{4}(1-z^2)+z(z+1) \\
&= \frac{1}{4}(1-z)(1+z)+z(z+1) \\
&= (\frac{1-z}{4}+z)(z+1) \\
&= \frac{(3z+1)(z+1)}{4}.
\end{aligned}$$

To check our result, we see if $p(z)$ is properly normalized, by integrating over z:

$$\begin{aligned}
1 &= \int_0^1 \frac{(3z+1)(z+1)}{4}\mathrm{d}z \\
&= \frac{1}{4}\int_0^1 (3z^2+4z+1)\mathrm{d}z \\
&= \frac{1}{4}(z^3+2z^2+z)|_0^1 \\
&= \frac{1}{4}(1+2+1-0-0-0) \\
&= 1. \quad \square
\end{aligned}$$

3. Use Bayes' rule to compute the posterior probability distribution function $p(c|z)$. Plot these distributions in MATLAB as a function of z. How can these posterior probabilities help in making a decision regarding the validity of the claim?

   ANSWER: Applying Bayes' rule for calculating the requested distribution gives:

$$\begin{aligned} p(c=0|z) &= \frac{p(z|c=0)p(c=0)}{p(z)} \\ &= \frac{\frac{3}{2}(1-z^2)\cdot\frac{1}{6}}{\frac{(3z+1)(z+1)}{4}} \\ &= \frac{\frac{1}{4}(1-z)(1+z)}{\frac{1}{4}(3z+1)(z+1)} \\ &= \frac{1-z}{3z+1}, \end{aligned}$$

$$\begin{aligned} p(c=1|z) &= 1-p(c=0|z) \\ &= \frac{4z}{3z+1}. \end{aligned}$$

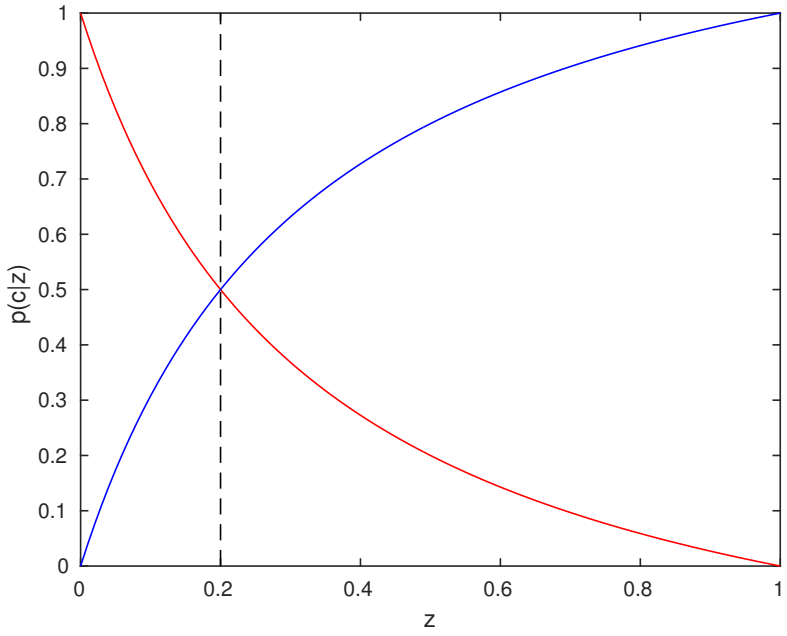

Figure 2: The posterior probabilities $p(c|z)$ are shown as a function of z. The red line corresponds to the values of $p(c=0|z)$, while the blue line corresponds to the values of $p(c=1|z)$. The vertical dashed line represents the decision boundary.

   The posterior probabilities of $c=0$ and $c=1$, which are shown as a function of z in Figure 2, are a measure of confidence in the validity / falsehood of a claim. For any particular value of the score z, $p(c=0|z) > p(c=1|z)$ implies that it is more likely for the claim to be false. Analogously, $p(c=0|z) < p(c=1|z)$ implies that it is more likely for the claim to be valid.

4. Compute the optimal decision boundary (based on our numerical score z) that minimizes the misclassification rate. For which z should we classify $c=0$ (false) and for which z should we classify $c=1$ (valid)? Explain your decision.

ANSWER: In order to minimize the misclassification rate, we should always choose the class that, given a particular value of the score z, has the maximum posterior probability $p(c|z)$ (see Bishop §1.5.1). In our case, we have to choose between two classes, namely $c = 0$ (false) and $c = 1$ (valid). In Figure 2, we can see that $p(c = 0|z) > p(c = 1|z)$ when the score is close to zero, while $p(c = 0|z) < p(c = 1|z)$ when the score is close to one. The decision boundary lies at the intersection of the two lines, where the posterior probabilities are equal.

To get the decision boundary, we find the score z for which $p(c = 0|z) = p(c = 1|z)$:

$$p(c = 0|z) = p(c = 1|z) \iff \frac{1 - z}{3z + 1} = \frac{4z}{3z + 1} \iff 1 - z = 4z \iff z = \frac{1}{5}.$$

We classify $c = 0$ if $z < \frac{1}{5}$ because in that region $p(c = 0|z) > p(c = 1|z)$ (see Figure 2).

We classify $c = 1$ if $z > \frac{1}{5}$ because in that region $p(c = 0|z) < p(c = 1|z)$ (see Figure 2).

If $z = \frac{1}{5}$, then $p(c = 0|z) = p(c = 1|z) = 0.5$, so either decision is just as good.

NB: One can always minimize the misclassification rate directly. First, we set a decision boundary at $z = \theta$, so that the decision regions for this problem become $\mathcal{R}_0 = [0, \theta)$ and $\mathcal{R}_1 = [\theta, 1]$. We can then compute the misclassification rate for these decision regions and then minimize this quantity with respect to θ.

5. Compute the misclassification rate, given the optimal decision boundary determined previously. Interpret the result you have obtained. Is the z score useful in determining the validity of the claim? Compare this with your prior guess from 1.

ANSWER: In the previous part, we determined that $\mathcal{R}_0 = [0, \frac{1}{5})$ (the region where we classify $c = 0$) and $\mathcal{R}_1 = [\frac{1}{5}, 1]$ (the region where we classify $c = 1$).

$$\begin{aligned}
\int_{\mathcal{R}_0} p(z, c = 1)\mathrm{d}z + \int_{\mathcal{R}_1} p(z, c = 0)\mathrm{d}z &= \int_{\mathcal{R}_0} p(z|c = 1)p(c = 1)\mathrm{d}z + \int_{\mathcal{R}_1} p(z|c = 0)p(c = 0)\mathrm{d}z \\
&= \int_0^{\frac{1}{5}} \frac{6}{5}(z^2 + z) \cdot \frac{5}{6}\mathrm{d}z + \int_{\frac{1}{5}}^1 \frac{3}{2}(1 - z^2) \cdot \frac{1}{6}\mathrm{d}z \\
&= \int_0^{\frac{1}{5}} (z^2 + z)\mathrm{d}z + \int_{\frac{1}{5}}^1 \frac{1}{4}(1 - z^2)\mathrm{d}z \\
&= \left(\frac{z^3}{3} + \frac{z^2}{2}\right)\Bigg|_0^{\frac{1}{5}} + \frac{1}{4}\left(z - \frac{z^3}{3}\right)\Bigg|_{\frac{1}{5}}^1 \\
&= \left(\frac{1}{3 \cdot 5^3} + \frac{1}{2 \cdot 5^2} - 0 - 0\right) + \frac{1}{4}\left(1 - \frac{1}{3} - \frac{1}{5} + \frac{1}{3 \cdot 5^3}\right) \\
&= \left(\frac{2}{6 \cdot 5^3} + \frac{15}{6 \cdot 5^3}\right) + \frac{1}{4}\left(\frac{2 \cdot 5^3}{3 \cdot 5^3} - \frac{3 \cdot 5^2}{3 \cdot 5^3} + \frac{1}{3 \cdot 5^3}\right) \\
&= \frac{17}{6 \cdot 5^3} + \frac{1}{4}\left(\frac{7 \cdot 5^2}{3 \cdot 5^3} + \frac{1}{3 \cdot 5^3}\right) \\
&= \frac{17}{6 \cdot 5^3} + \frac{1}{4}\frac{176}{3 \cdot 5^3} \\
&= \frac{105}{6 \cdot 5^3} \\
&= 14\%
\end{aligned}$$

We can compare this result with the misclassification rate obtained for a score that does not provide any information regarding the claim. To see this, we set the conditional probability

distributions of z given c to be uniform: $p(z|c=1) = p(z|c=0) = 1$. The misclassification rate then becomes:

$$\int_{\mathcal{R}_0} p(z, c=1)\mathrm{d}z + \int_{\mathcal{R}_1} p(z, c=0)\mathrm{d}z = \int_{\mathcal{R}_0} \frac{5}{6}\mathrm{d}z + \int_{\mathcal{R}_1} \frac{1}{6}\mathrm{d}z$$

We can see from the above formulation that, in order to minimize the misclassification rate, we should always make the decision that a claim is valid. This means that we set the decision regions to $\mathcal{R}_0 = \varnothing$ and $\mathcal{R}_1 = [0, 1]$. For these decision regions, the computed misclassification rate is 16.67%.

Intuitively, since we know that a claim is (five times) more likely to be valid than false and we have no other information to help us in making the decision, it is logical to guess that any claim is valid. The misclassification rate can be interpreted as the fraction of claims that are false, even though we classify them as valid.

We notice that with the help of our voice-analysis system, we achieve a better misclassification rate of 14%, which means that the z score is indeed useful for our decision making process.

Exercise 2 – weight 2

The government of the United Kingdom has decided to call a referendum regarding the country's European Union membership. The citizens of the UK will be asked the following question at the referendum: "Should the United Kingdom remain a member of the European Union or leave the European Union?". The European Commission (EC) is interested in the potential outcome of this referendum and has contracted a polling agency to study this issue.

Suppose that a person's vote follows a Bernoulli distribution with parameter θ and suppose that the EC's prior distribution for θ, the proportion of British citizens that would be in favor of leaving the EU, is beta distributed with $\alpha = 90$ and $\beta = 110$.

1. Determine the mean and variance of the prior distribution. Plot the prior density function.

   ANSWER: For determining the mean and the variance of the beta distribution see, for example, formulas (2.15) and (2.16) in Bishop:

   $$\mathbb{E}[\theta] = \frac{\alpha}{\alpha+\beta} = \frac{90}{200} = 0.45;$$
   $$\mathrm{var}[\theta] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = \frac{90 \cdot 110}{200^2 \cdot 201} \approx 0.001231.$$

   To plot the beta prior density (see Figure 3), you can use for example the function `betapdf` in Matlab.

2. A random sample of 1000 British citizens is taken, and 60% of the people polled support leaving the European Union. What are the posterior mean and variance for θ? Plot the posterior density function together with the prior density. Explain how the data from the sample changed the prior belief.

   ANSWER:

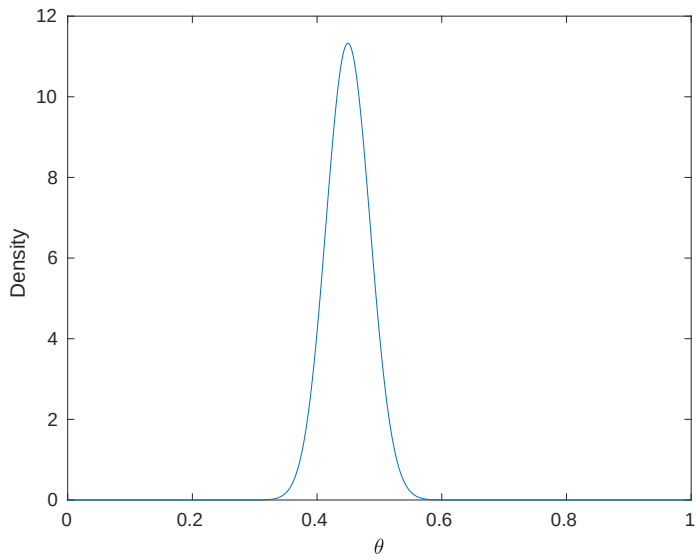

Figure 3: Probability density function of $\text{Beta}(\theta|90, 110)$.

Assuming each vote is independent, we have a sum of independent random Bernoulli trials, which means that the number of votes for leaving the EU, which we denote by S in the sample follows a binomial distribution:

$$p(S|\theta) = \text{Bin}(S|1000, \theta) = \binom{1000}{600} \theta^S (1-\theta)^{1000-S}.$$

The beta distribution is the conjugate prior for the binomial distribution (Bishop, §2.1.1), which means that the posterior of θ given the observed random sample also follows a beta distribution. This result can easily be derived by applying Bayes' rule. To get the parameters of the posterior beta distribution, we add up the number of observations in the sample to the effective number of observations in the prior:

$$p(\theta|S) = \text{Beta}(\theta|600 + 90, 400 + 110) = \text{Beta}(\theta|690, 510).$$

Again, using formulas (2.15) and (2.16) from Bishop, the mean and variance of $p(\theta|S)$ are:

$$\begin{aligned} \mathbb{E}[\theta|S] &= \frac{690}{690 + 510} = 0.575; \\ \text{var}[\theta|S] &= \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = \frac{690 \cdot 510}{1200^2 \cdot 1201} \approx 0.0002. \end{aligned}$$

We can see that the mean has increased significantly compared to the mean of the prior, which suggests that the chance for a 'Leave' win is higher than what was previously thought. At the same time, the variance of the distribution over θ has decreased (equivalently, the precision has increased), which means that we are more confident in the value of θ after having observed more data. To see this change even better, we show the prior and posterior distribution over θ together in Figure 4, where the dashed vertical line corresponds to $\theta = 0.5$.. The posterior distribution has shifted close to $\theta = 0.6$, but not exactly at 0.6 due to the influence of the

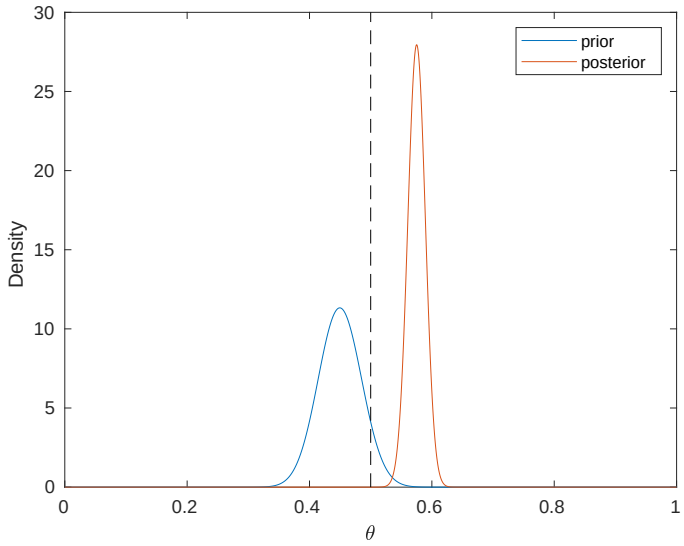

Figure 4: Prior and posterior probability density function of θ.

prior. The shift is significant because the observed sample size is larger than the effective sample size of the prior ($\alpha + \beta = 200$) and, as a result, the observed sample weighs more than the prior information. Since most of the probability mass now corresponds to $\theta > 0.5$, it suggests a higher expected chance of the 'leave' side winning, whereas before it looked like 'remain' had the higher expected chance.

3. Examine the effect of changing the prior hyperparameters (α, β) on the posterior by looking at several other hyperparameter configurations. Which values for α and β correspond to a non-informative prior? What is the interpretation of α and β for the beta prior? What does the choice of α and β in Question 1 tell you about the strength of the prior belief?

ANSWER: The hyperparameters α and β can be interpreted as the effective (pseudo-)number of prior observations corresponding to $vote = 1$ ('leave') and $vote = 0$ ('remain'), respectively. By increasing the sum $\alpha + \beta$, we increase the total (pseudo-)number of prior observations, and therefore the strength in our prior belief. This can be observed in Figure 5 by noticing that for higher values of α and β, the prior (and posterior) distribution has less variance. Furthermore, for higher values for α and β, the posterior shifts more and more towards the prior, indicating that the prior carries a higher weight relative to the data. By changing the ratio $\frac{\alpha}{\beta}$, we shift the mean and the mode of the prior distribution, which indicate the expected value and most likely value, respectively, of θ. This can be seen in Figure 6, where we have kept the same number of effective prior samples ($\alpha + \beta$). We also notice that the posterior distribution does not change much with the prior because the 1000 observations in the sample carry significantly more weight than the 200 effective prior observations.

A non-informative prior is a distribution that gives us as little information as possible about the parameter under study, in our case θ. A typical choice for a non-informative prior is the uniform distribution, since its probability density is the same for every possible parameter value. For $\alpha = 1$ and $\beta = 1$, the Beta prior becomes a uniform distribution (check for yourself), so this is our non-informative choice. The non-informativeness of this prior can be seen in Figure 7. The prior distribution has uniform density across the parameter space, while the posterior distribution has its mode exactly at $\theta = 0.6$, which is equal to the proportion of citizens in the sample that are in favor of leaving the EU (60%). This means

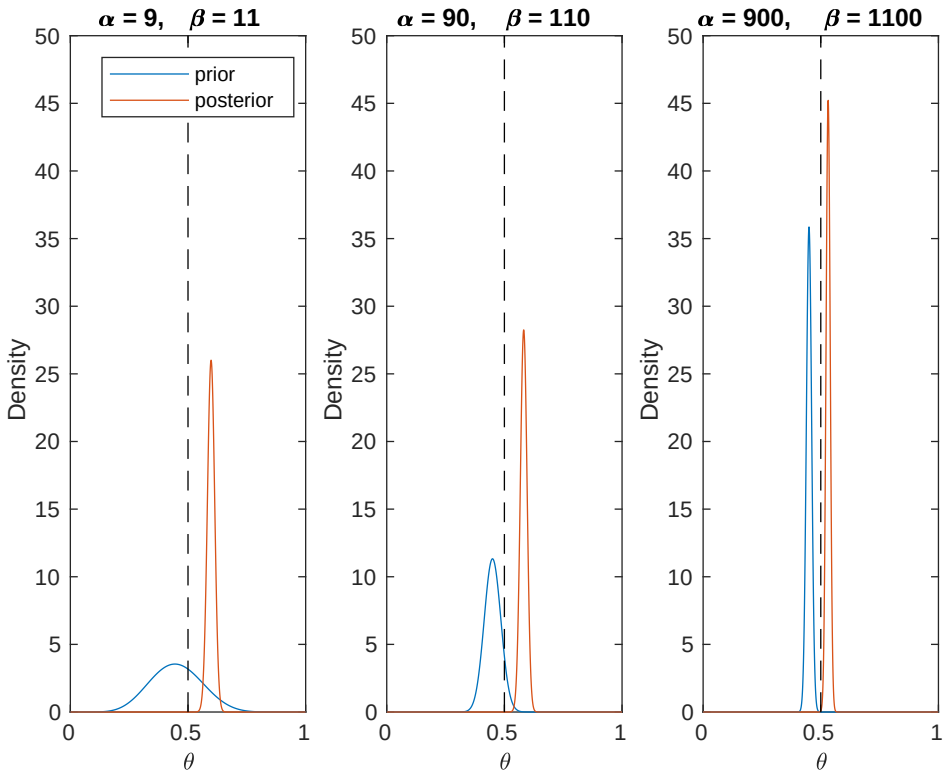

Figure 5: By varying the sum of α and β while keeping their ratio constant, we change the variance (precision) of the prior belief.

that the maximum a-posteriori (MAP) value of θ is completely determined by the data and thus the prior has no effect.

4. Imagine you are now a reporter for the polling agency and you have been sent on field duty to gather more data. Your mission is to go out on the streets and randomly survey people on their thoughts regarding the upcoming referendum. Given all the available information you have acquired, what is the probability that the first person you talk to will vote 'Leave'?

   Hint: Derive the predictive distribution for the next vote using the posterior distribution for θ computed in Question 2. For a reminder on predictive distributions, see subsection 1.2.6 in Bishop, in particular Equation (1.68).

   ANSWER:

   We want to predict the value of a new vote given the data collected and sample across all possible parameter (θ) values. To do this, we derive the predictive distribution by and integrating out θ in the joint posterior distribution over the new observation and θ given the data S from the previous random poll:

$$p(vote = 1|S) = \int_0^1 p(vote = 1, \theta|S)\mathrm{d}\theta = \int_0^1 p(vote = 1|\theta)p(\theta|S)\mathrm{d}\theta = \int_0^1 \theta p(\theta|S)\mathrm{d}\theta = \mathbb{E}[\theta|S].$$

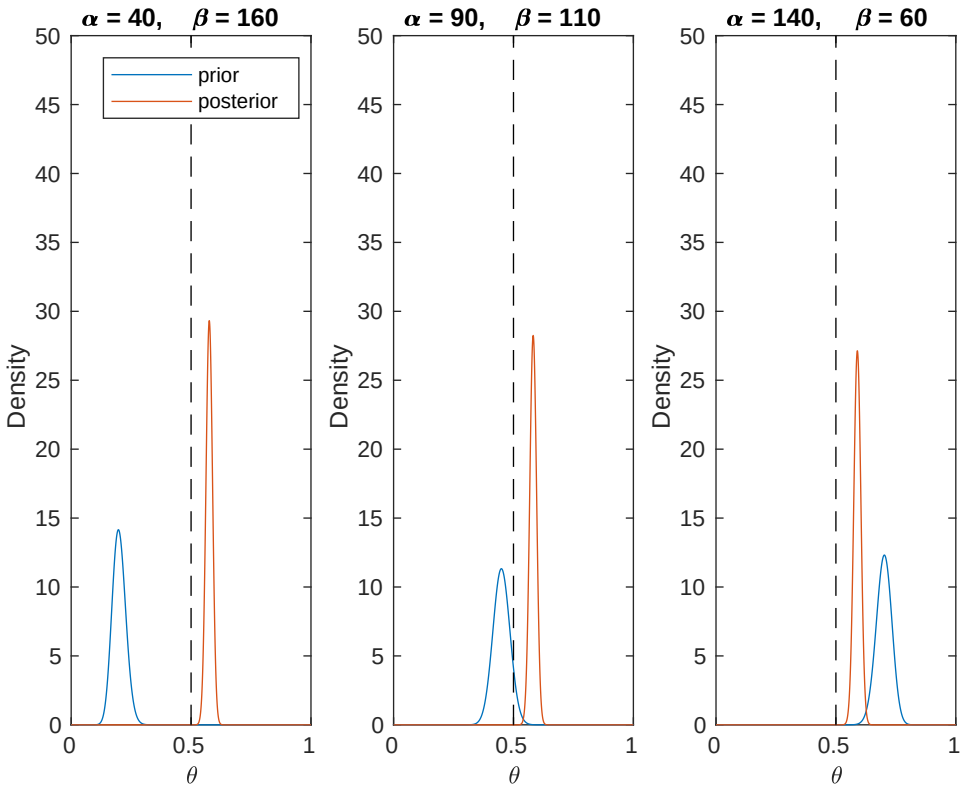

Figure 6: By varying the ratio between α and β while keeping the sum constant, we change the mean (mode) of the prior distribution.

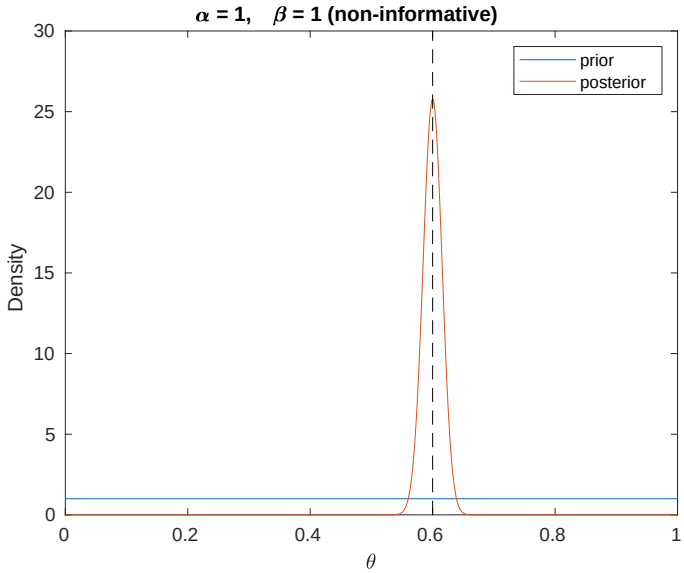

Figure 7: For $\alpha = \beta = 1$, the beta prior becomes a uniform distribution. The mode of the posterior is then determined by the data and is equal to the empirical proportion of leave voters in the sample (60%).

Since we know that the posterior over θ has a Beta(690, 510) distribution, we can immediately get the mean from Bishop (2.15):

$$p(v = 1|S) = \mathbb{E}[\theta|S] = \frac{690}{1200} = 0.575.$$

In conclusion, the probability of the first person surveyed answering 'Leave' is 57.5%.

Exercise 3 – Sequential learning (weight 5)

Part 1 – Obtaining the prior

Consider a four dimensional variable $[x_1, x_2, x_3, x_4]^T$, distributed according to a multivariate Gaussian with mean $\tilde{\boldsymbol{\mu}} = [1, 0, 1, 2]^T$ and covariance matrix $\tilde{\boldsymbol{\Sigma}}$ given as

$$\tilde{\boldsymbol{\Sigma}} = \left(\begin{array}{cc|cc} 0.14 & -0.3 & 0.0 & 0.2 \\ -0.3 & 1.16 & 0.2 & -0.8 \\ \hline 0.0 & 0.2 & 1.0 & 1.0 \\ 0.2 & -0.8 & 1.0 & 2.0 \end{array}\right) \tag{2}$$

We are interested in the conditional distribution over $[x_1, x_2]^T$, given that $x_3 = x_4 = 0$. We know this conditional distribution will also take the form of a Gaussian:

$$p\big([x_1, x_2]^T \,|\, x_3 = x_4 = 0\big) = \mathcal{N}([x_1, x_2]^T|\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p) \tag{3}$$

for which the mean and covariance matrix are most easily expressed in terms of the (partitioned) precision matrix (see Bishop,§2.3.1).

1. Use the partitioned precision matrix $\tilde{\boldsymbol{\Lambda}} = \tilde{\boldsymbol{\Sigma}}^{-1}$ to give an explicit expression for the mean $\boldsymbol{\mu}_p$ and covariance matrix $\boldsymbol{\Sigma}_p$ of this distribution and calculate their values. (This distribution will be taken as the *prior* information for the rest of this exercise, hence the subscript p). You may use the MATLAB command `inv` to calculate matrix inverses.

   ANSWER: The precision matrix $\tilde{\boldsymbol{\Lambda}}$ is given as (via MATLAB or the matrix inversion formula)

   $$\tilde{\boldsymbol{\Lambda}} = \left(\begin{array}{cc|cc} 60 & 50 & -48 & 38 \\ 50 & 50 & -50 & 40 \\ \hline -48 & -50 & 52.4 & -41.4 \\ 38 & 40 & -41.4 & 33.4 \end{array}\right) \tag{4}$$

   Bishop, eq.2.73 and 2.75, describe the covariance matrix and mean of a subset $\mathbf{x}_a$ of a multivariate Gaussian $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$, conditional on the subset $\mathbf{x}_b$ of the remaining components, in terms of the partitioned precision matrix $\boldsymbol{\Lambda}$

   $$\begin{aligned} \boldsymbol{\Sigma}_{a|b} &= \boldsymbol{\Lambda}_{aa}^{-1} \\ \boldsymbol{\mu}_{a|b} &= \boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b) \end{aligned}$$

   For $\mathbf{x}_a = [x_1, x_2]^T$ and $\mathbf{x}_b = [x_3, x_4]^T = [0, 0]^T$, this gives

   $$\begin{aligned} \boldsymbol{\Sigma}_p &= \begin{bmatrix} 0.1 & -0.1 \\ -0.1 & 0.12 \end{bmatrix} \\ \boldsymbol{\mu}_p &= \begin{bmatrix} 1.0 \\ 0.0 \end{bmatrix} - \begin{bmatrix} 0.1 & -0.1 \\ -0.1 & 0.12 \end{bmatrix} \cdot \begin{bmatrix} -48 & 38 \\ -50 & 40 \end{bmatrix} \cdot \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix} - \begin{bmatrix} 1 \\ 2 \end{bmatrix} \right) = \begin{bmatrix} 0.8 \\ 0.8 \end{bmatrix} \end{aligned}$$

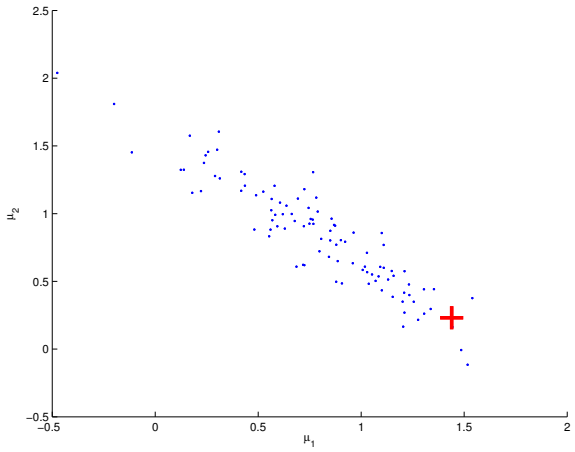

Figure 8: Several data points drawn from the distribution in (3). The single "true" point $\boldsymbol{\mu}_t$ is indicated with a red cross.

2. [MATLAB] - Create a function that can generate random number pairs, distributed according to the distribution in (3). Initialize your random generator and then draw a *single* pair

$$\boldsymbol{\mu}_t = [\mu_{t1}, \mu_{t2}]^T \tag{5}$$

from this distribution. (These will be the 'true' means, hence the subscript t).

Hint: you can use the MATLAB function `mvnrnd` (which resides in the `Statistics` toolbox[1]).

ANSWER: Drawing a single pair $\boldsymbol{\mu}_t = [x_1, x_2]$ from $\mathcal{N}(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p)$ is realized in MATLAB by the multivariate normal distribution function `mvnrnd`, with proper initialization, as

```
randn('state', 12345);
mu_t = mvnrnd(mu_p, Sigma_p, 1);
```

See example code in file `a007.m` and Figure 8.

Note: You can also write your own multivariate Gaussian random function starting from the zero mean, univariate `randn` function. In that case, if $\mathbf{y}$ is a vector of independent, zero mean Gaussian variables with unit variance, then

$$\boldsymbol{\mu}_t = \boldsymbol{\mu}_p + \mathbf{L}_p \mathbf{y}$$

has the desired distribution, where $\mathbf{L}_p$ represents the *Cholesky decomposition* of covariance matrix $\boldsymbol{\Sigma}_p$, defined as

$$\mathbf{L}_p \mathbf{L}_p^T = \boldsymbol{\Sigma}_p$$

which can be calculated in MATLAB by the command `L = chol(Sigma,'lower')`. (This is analogous to the univariate case, where $x = \mu + \sigma y$ is Gaussian with mean μ and variance σ^2. See also Bishop, §11.1.1.).

3. [MATLAB] - Make a plot of the probability density of the distribution (3).

Hint: use the MATLAB function `mvnpdf` (which resides in the `Statistics` toolbox[1]) to calculate the probability density of a multivariate Gaussian random variable. The MATLAB functions `meshgrid` and `surf` may also prove useful.

ANSWER: See example code in file `a007.m` and the resulting plot in Figure 9.

[1] In OCTAVE, you have to install the `Statistics` package in order to use this function.

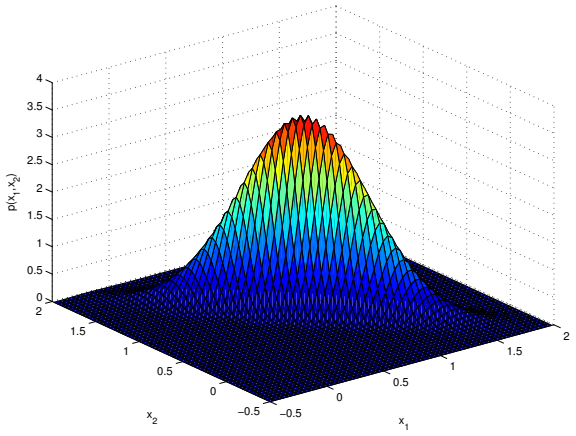

Figure 9: Probability density of the distribution (3).

Part 2 – Generating the data

Here we assume we are dealing with a 2d-Gaussian data generating process

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{6}$$

For the mean $\boldsymbol{\mu}$, we will use the value $\boldsymbol{\mu}_t$ drawn in (5) in order to *generate* the data. Subsequently, we will pretend that we do not know this "true" value $\boldsymbol{\mu}_t$ of $\boldsymbol{\mu}$, and estimate $\boldsymbol{\mu}$ from the data. For the covariance matrix $\boldsymbol{\Sigma}$ we will use the "true" value

$$\boldsymbol{\Sigma}_t = \begin{pmatrix} 2.0 & 0.8 \\ 0.8 & 4.0 \end{pmatrix} \tag{7}$$

to generate the data.

1. [MATLAB] - Generate at least 1000 data pairs $\{x_i, y_i\}$, distributed according to (6) with $\boldsymbol{\mu} = \boldsymbol{\mu}_t$ and $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_t$ and save them to file in a plain-text format.

   Hint: Use MATLAB functions `save` and `load` with the `-ascii` flag.

   ANSWER: Use the $\boldsymbol{\Sigma}_t$ from (7), and the $\boldsymbol{\mu}_t$ generated before, to draw 1000 data points from a multivariate Gaussian, and save them to file as

   ```
   randn('state', 56789);
   Sigma_t = [2.0 , 0.8; 0.8 , 4.0 ];
   N = 1000;
   dat = mvnrnd(mu_t, Sigma_t, N);
   %...
   save(s, 'dat', '-ascii');
   ```

   For further details, see `a007.m`. A scatter plot of the generated data is given Figure 10.

2. From now on, we will assume (pretend) the 'true' mean $\boldsymbol{\mu}_t$ is unknown and estimate $\boldsymbol{\mu}$ from the data. Calculate the maximum likelihood estimate of $\boldsymbol{\mu}_{\mathrm{ML}}$ and $\boldsymbol{\Sigma}_{\mathrm{ML}}$ for the data, and also an unbiased estimate of $\boldsymbol{\Sigma}$ (see Bishop, §2.3.4). Compare with the true values $\boldsymbol{\mu}_t$ and $\boldsymbol{\Sigma}_t$.

   ANSWER: The maximum likelihood estimate of the mean $\boldsymbol{\mu}_{\mathrm{ML}}$ is simply the average of all the datapoints (for each component): eq.2.121. This is an unbiased estimate (eq.2.123).

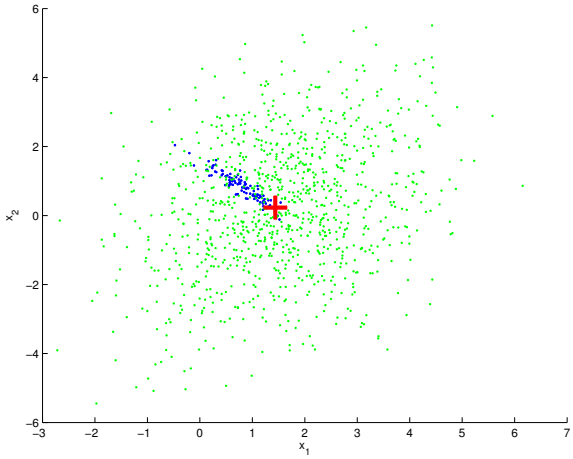

Figure 10: Scatter plot of the noisy observations (green points) of the "true" data (red cross) drawn from the prior distribution (3) (blue points).

The maximum likelihood estimate for the covariance matrix $\mathbf{\Sigma}_{\mathrm{ML}}$ is likewise an average over the contributions $(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})^T$ (each a 2d-vector times its transposed, giving a 2x2 matrix). This estimate would be too low, but from eq.2.125 we see that dividing by $N-1$ instead of N results in the requested unbiased estimator. See `a007.m` for an example implementation of these estimators.

Part 3 – Sequential learning algorithms

We will now estimate the mean $\boldsymbol{\mu}$ from the generated data and the known variance $\mathbf{\Sigma}_t$ *sequentially*, i.e., by considering the data points one-by-one.

1. [MATLAB] - Write a procedure that processes the data points $\{\mathbf{x}_n\}$ in the generated file one-by-one, and after each step computes an updated estimate of $\boldsymbol{\mu}_{\mathrm{ML}}$, the maximum likelihood of the mean (using Bishop, eq.2.126).

   ANSWER: Sequential learning implies that we start from some value $\boldsymbol{\mu}^{(0)}$ and calculate a new estimate each time a new datapoint is observed. After a datapoint is processed it can be discarded (only the updated estimate is needed for the next step).
   For maximum likelihood this results in the simple updating scheme of eq.2.126

$$\begin{aligned} \boldsymbol{\mu}_{\mathrm{ML}}^{(n)} &= \boldsymbol{\mu}_{\mathrm{ML}}^{(n-1)} + \frac{1}{n}(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}}^{(n-1)}) \\ \boldsymbol{\mu}_{\mathrm{ML}}^{(0)} &= (0,0)^T \end{aligned}$$

   in words: the new, updated estimate after observing the n-th datapoint is simply the previous estimate plus 1-over-n times the difference between this estimate and the observed value $\mathbf{x}_n$. Therefore only the last estimate $\boldsymbol{\mu}_{\mathrm{ML}}^{(n)}$ needs to be retained for the next update.

Now we also use the prior information $p(\boldsymbol{\mu}) = \mathcal{N}(\boldsymbol{\mu}|\boldsymbol{\mu}_p, \mathbf{\Sigma}_p)$. From the prior, the generated data and the known variance $\mathbf{\Sigma}_t$, we will estimate the mean $\boldsymbol{\mu}$.

2. Work out the details of sequential Bayesian inference (see eq.2.144) for the mean $\boldsymbol{\mu}$. Apply Bayes' theorem in eq. 2.113-2.117 at each step $n = 1, \ldots, N$ to compute the new posterior mean $\boldsymbol{\mu}^{(n)}$ and covariance $\mathbf{\Sigma}^{(n)}$ after a new point $(\mathbf{x}_n)$ has arrived from the old posterior mean $\boldsymbol{\mu}^{(n-1)}$ and covariance $\mathbf{\Sigma}^{(n-1)}$. Use this updated posterior as the prior in the next step. The first step starts from the original prior (3).
   Note: do not confuse the posterior $\mathbf{\Sigma}^{(n)}$ with the known $\mathbf{\Sigma}_t$ of the data generating process.

For some more hints, see appendix.

ANSWER: In this case, we need to take the available prior information $p(\boldsymbol{\mu}) = \mathcal{N}(\boldsymbol{\mu}|\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p)$ as well as the data into account. For this we can use Bayes' theorem for Gaussians (eq.2.113-7). However, instead of using the given prior and writing a likelihood for the entire dataset, eq. 2.144 suggests we can also view this as a sequential update for each of the points separately, where the prior for point $\mathbf{x}_n$ is just the posterior distribution after observing $n-1$ data points. At each step we have to compute the posterior from Bayes' theorem for Gaussian variables:

$$\begin{aligned} p(\mathbf{x}) &= \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}) \\ p(\mathbf{y}|\mathbf{x}) &= \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{L}^{-1}) \\ &\Rightarrow \\ p(\mathbf{x}|\mathbf{y}) &= \mathcal{N}(\mathbf{x}|\mathbf{m}_n, \mathbf{S}_n) \end{aligned}$$

where $\mathbf{S}_n = (\boldsymbol{\Lambda} + \mathbf{A}^T\mathbf{L}\mathbf{A})^{-1}$ and $\mathbf{m}_n = \mathbf{S}_n\{\mathbf{A}^T\mathbf{L}(\mathbf{y} - \mathbf{b}) + \boldsymbol{\Lambda}\boldsymbol{\mu}\}$.
Matching the parameters (standard form $\leftrightarrow$ exercise)

$$\begin{aligned} \mathbf{x} &\leftrightarrow \boldsymbol{\mu} \\ \boldsymbol{\mu} &\leftrightarrow \boldsymbol{\mu}_p^{(n)} \\ \boldsymbol{\Lambda}^{-1} &\leftrightarrow \boldsymbol{\Sigma}_p^{(n)} \\ \mathbf{y} &\leftrightarrow \mathbf{x}_n \\ \mathbf{A} &\leftrightarrow \mathbf{I}_2 \\ \mathbf{b} &\leftrightarrow (0,0)^T \\ \mathbf{L}^{-1} &\leftrightarrow \boldsymbol{\Sigma}_t \end{aligned}$$

Set the prior equal to the previous posterior: $\boldsymbol{\mu}_p^{(n)} = \mathbf{m}_{n-1}$ and $\boldsymbol{\Sigma}_p^{(n)} = \mathbf{S}_{n-1}$. Then calculate the parameters $\boldsymbol{\Lambda}$, $\mathbf{L}$ and $\mathbf{S}_n$ as needed

$$\begin{aligned} \boldsymbol{\Lambda} &= \mathbf{S}_{n-1}^{-1} \\ \mathbf{L} &= \boldsymbol{\Sigma}_t^{-1} \\ \mathbf{S}_n &= (\mathbf{S}_{n-1}^{-1} + \boldsymbol{\Sigma}_t^{-1})^{-1} \\ \mathbf{m}_n &= \mathbf{S}_n\left\{\boldsymbol{\Sigma}_t^{-1} \cdot \mathbf{x}_n + \mathbf{S}_{n-1}^{-1} \cdot \mathbf{m}_{n-1}\right\} \end{aligned}$$

For $n = 1$, $\mathbf{m}_0$ and $\mathbf{S}_0$ simply correspond to the prior (3), before the first data point $\mathbf{x}_1$ is observed.

3. [MATLAB] - Write a procedure that processes the data points $\{\mathbf{x}_n\}$ in the generated file one-by-one, and after each step computes an updated estimate of $\boldsymbol{\mu}_{\text{MAP}}$ - the maximum of the posterior distribution, using the results of the previous exercise.

ANSWER: The mode of a Gaussian is also its mean, so we have $\boldsymbol{\mu}_{\text{MAP}}^{(n)} = \mathbf{m}_n$, and therefore for the variance $\boldsymbol{\Sigma}_{\text{MAP}}^{(n)} = \mathbf{S}_n$. This results in the following MAP update scheme

$$\begin{aligned} \boldsymbol{\Sigma}_{\text{MAP}}^{(n)} &= \left((\boldsymbol{\Sigma}_{\text{MAP}}^{(n-1)})^{-1} + \boldsymbol{\Sigma}_t^{-1}\right)^{-1} \\ \boldsymbol{\mu}_{\text{MAP}}^{(n)} &= \boldsymbol{\Sigma}_{\text{MAP}}^{(n)} \cdot \left(\boldsymbol{\Sigma}_t^{-1} \cdot \mathbf{x}_n + (\boldsymbol{\Sigma}_{\text{MAP}}^{(n-1)})^{-1} \cdot \boldsymbol{\mu}_{\text{MAP}}^{(n-1)}\right) \\ \boldsymbol{\Sigma}_{\text{MAP}}^{(0)} &= \boldsymbol{\Sigma}_p \\ \boldsymbol{\mu}_{\text{MAP}}^{(0)} &= \boldsymbol{\mu}_p \end{aligned}$$

See the actual implementation in `a007.m`.

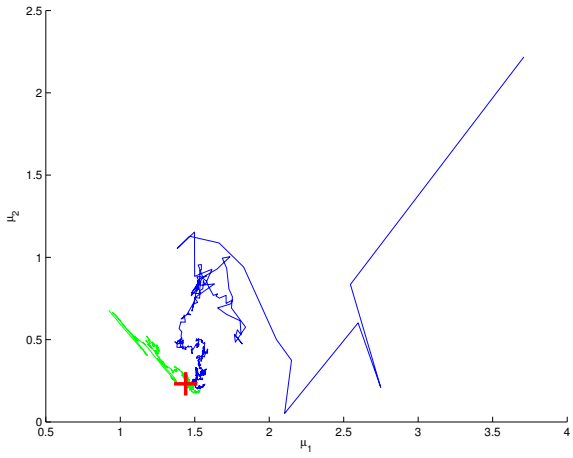

Figure 11: Blue: 2D-plot joining points $(\mu_{\text{ML}}^{(n)}, \mu_{\text{ML}}^{(n+1)})$ for successive n. Green: 2D-plot joining points $(\mu_{\text{MAP}}^{(n)}, \mu_{\text{MAP}}^{(n+1)})$ for successive n.

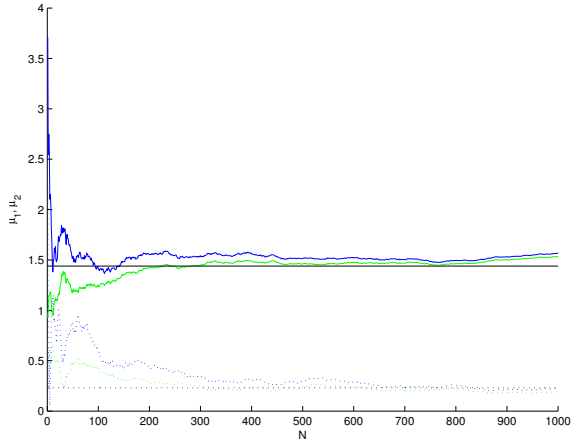

Figure 12: Estimates of μ_1 (solid lines) and μ_2 (dotted lines) as a function of n. Blue: ML, green: MAP, black: true values.

4. [MATLAB] - Plot both estimates (ML and MAP) in a single graph (1d or 2d) as a function of the number of data points observed. Indicate the true values $\{\mu_{t1}, \mu_{t2}\}$ as well. Evaluate your result.

   ANSWER: Make sure you store the values for $\boldsymbol{\mu}_{\text{ML}}$ and $\boldsymbol{\mu}_{\text{MAP}}$ at each intermediate step n and use these to plot against each other. Useful graphs to get an impression of the convergence behaviour are:

   - Lineplots of components of $\boldsymbol{\mu}_{\text{ML}}^{(n)}$ and $\boldsymbol{\mu}_{\text{MAP}}^{(n)}$ vs. n (see Figure 12);
   - 2D-plot joining points $(\boldsymbol{\mu}_{\text{ML}}^{(n)}, \boldsymbol{\mu}_{\text{ML}}^{(n+1)})$ for successive n.
   - combinations of $\boldsymbol{\mu}_{\text{ML}}$ and $\boldsymbol{\mu}_{\text{MAP}}$ components in a single plot (see Figure 11).
   - the final posterior distribution (see Figure 13).

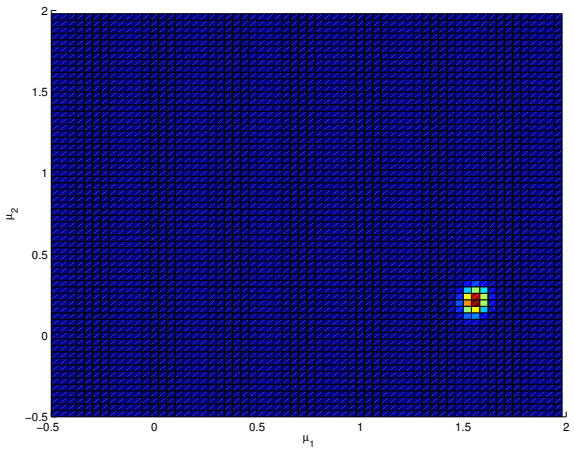

Figure 13: Final posterior distribution $\mathcal{N}(\boldsymbol{\mu}_{\text{MAP}}^{(1000)}, \boldsymbol{\Sigma}_{\text{MAP}}^{(1000)})$.

4 Hints

Below are some hints for **Exercise 3 - Part 3 - Question 2**.

Bayes rule is also valid if earlier acquired information is taken into account. For example, if this is earlier seen data $D_{n-1} = \{x_1, \ldots, x_{n-1}\}$. Bayes rule conditioned on this earlier data is

$$P(\mu|x_n, D_{n-1}) \propto P(\mu|D_{n-1})P(x_n|\mu, D_{n-1}).$$

Since$D_n = \{x_1, \ldots, x_n\}$ this is written more conveniently as

$$P(\mu|D_n) \propto P(\mu|D_{n-1})P(x_n|\mu, D_{n-1}).$$

If, given the model parameters μ, the probability distribution of x_n is independent of earlier data D_{n-1}, we can further reduce this to

$$P(\mu|D_n) \propto P(\mu|D_{n-1})P(x_n|\mu)$$

You should be able to see the relation with (2.144) and see in particular that the factor between square brackets in (2.144) is to be identified with $P(\mu|D_{n-1})$.

Another important insight is that if $P(\mu|D_{n-1})$ and $P(x_n|\mu)$ are of the form (2.113) and (2.114), i.e., if $P(\mu|D_{n-1})$ is a Gaussian distribution over μ with a certain mean and covariance (you are free to give these any name, e.g. $\mu^{(n-1)}$, $\Sigma^{(n-1)}$) and if $P(x_n|\mu)$ is also Gaussian with a mean that is linear μ, then you can use (2.116) and (2.117) to compute the posterior $P(\mu|D_n)$, which therefore is also Gaussian.

So it is your task to show this. To do this you have to figure out the mapping of the variables and parameters in the current exercise, i.e., what is the correspondence between $\mu, x_n, \Sigma_t, \mu^{(n-1)}, \Sigma^{(n-1)}$ etc. with $x, \mu, \Lambda, y, A, b, L$. Don't forget that some quantities can also be zero or and other may be identity matrices.