KTH Electrical Engineering

Pattern Recognition Problems

Solutions Manual

ARNE LEIJON, ET AL.

Stockholm 2012-10-12

KTH Electrical Engineering
Royal Institute of Technology

Teacher's Notes www.kth.se/social/course/EN2202/

Chapter 1

Introduction

1.1.a We know that $E[Z_1] = E[Z_2] = 0$, $\text{var}[Z_1] = \sigma_1^2$, $\text{var}[Z_2] = \sigma_2^2$, and $\text{cov}[Z_1, Z_2] = 0$. Using elementary rules for the expected value, and definitions,

$$\boldsymbol{\mu}_X = E[\boldsymbol{X}] = \boldsymbol{p}_1 E[Z_1] + \boldsymbol{p}_2 E[Z_2] = \boldsymbol{p}_1 0 + \boldsymbol{p}_2 0 = \mathbf{0} \tag{1.1}$$

$$\begin{aligned} C_X = \text{cov}[\boldsymbol{X}] = E\left[(\boldsymbol{X} - \boldsymbol{\mu}_X)(\boldsymbol{X} - \boldsymbol{\mu}_X)^T\right] = \\ = E\left[(\boldsymbol{p}_1 Z_1 + \boldsymbol{p}_2 Z_2)(\boldsymbol{p}_1 Z_1 + \boldsymbol{p}_2 Z_2)^T\right] = \\ = \boldsymbol{p}_1 E\left[Z_1^2\right] \boldsymbol{p}_1^T + \boldsymbol{p}_2 E\left[Z_2^2\right] \boldsymbol{p}_2^T - \boldsymbol{p}_1 E[Z_1 Z_2] \boldsymbol{p}_2^T - \boldsymbol{p}_2 E[Z_1 Z_2] \boldsymbol{p}_1^T = \\ = \sigma_1^2 \boldsymbol{p}_1 \boldsymbol{p}_1^T + \sigma_2^2 \boldsymbol{p}_2 \boldsymbol{p}_2^T = PDP^T \end{aligned} \tag{1.2}$$

with

$$P = (\boldsymbol{p}_1, \boldsymbol{p}_2); \quad D = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix} \tag{1.3}$$

1.1.b Using the previous general results,

$$E[\boldsymbol{X}] = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \tag{1.4}$$

$$\text{cov}[\boldsymbol{X}] = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} 9 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} = \begin{pmatrix} 10 & 8 \\ 8 & 10 \end{pmatrix} \tag{1.5}$$

1.2.a If $A\boldsymbol{e}_k = \lambda_k \boldsymbol{e}_k$, then $Ac\boldsymbol{e}_k = cA\boldsymbol{e}_k = c\lambda_k \boldsymbol{e}_k$. Thus, if $\boldsymbol{e}_k$ is an eigenvector, then $c\boldsymbol{e}_k$ is also an eigenvector, for any c.

1.2.b The characteristic equation is

$$\det \begin{pmatrix} 10 - \lambda & 8 \\ 8 & 10 - \lambda \end{pmatrix} = \lambda^2 - 20\lambda + 100 - 64 = 0 \tag{1.6}$$

with solutions $\lambda_1 = 18$ and $\lambda_2 = 2$. For the corresponding eigenvector we have

$$\begin{pmatrix} 10 & 8 \\ 8 & 10 \end{pmatrix} \begin{pmatrix} e_{11} \\ e_{12} \end{pmatrix} = 18 \begin{pmatrix} e_{11} \\ e_{12} \end{pmatrix} \tag{1.7}$$

$$\begin{cases} -8e_{11} + 8e_{12} = 0 \\ 8e_{11} - 8e_{12} = 0 \end{cases} \tag{1.8}$$

As expected, the two equations are linearly dependent, because we are free to choose the scale factor for the eigenvector. Both equations lead to the result $e_{11} = e_{12}$. To normalize the eigenvector we choose

$$\boldsymbol{e}_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \tag{1.9}$$

The same approach gives the other eigenvector

$$\boldsymbol{e}_2 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ -1 \end{pmatrix} \tag{1.10}$$

1.2.c Using $C\boldsymbol{e} = \lambda \boldsymbol{e}$ we get $C^{-1}C\boldsymbol{e} = C^{-1}\lambda \boldsymbol{e}$, and therefore

$$C^{-1}\boldsymbol{e} = \lambda^{-1}C^{-1}C\boldsymbol{e} = \lambda^{-1}\boldsymbol{e}$$

1.2.d Using the hint we obtain the two scalar equations

$$\overline{\boldsymbol{e}}_k^T C \boldsymbol{e}_k = \overline{\boldsymbol{e}}_k^T \lambda_k \boldsymbol{e}_k$$
$$\overline{\boldsymbol{e}}_k^T \overline{C}^T \boldsymbol{e}_k = \overline{\boldsymbol{e}}_k^T \overline{\lambda}_k \boldsymbol{e}_k$$

As C is real and symmetric, we have $C = \overline{C}^T$, and the two left-hand sides are equal, so the two right-hand sides must also be equal. Thus, $\lambda_k = \overline{\lambda}_k$, so λ_k must be real.

1.2.e We have two equations

$$\boldsymbol{e}_k^T C \boldsymbol{e}_l = \boldsymbol{e}_k^T \lambda_l \boldsymbol{e}_l$$
$$\boldsymbol{e}_l^T C \boldsymbol{e}_k = \boldsymbol{e}_l^T \lambda_k \boldsymbol{e}_k$$

Transposing the second equation, we get

$$\boldsymbol{e}_k^T C^T \boldsymbol{e}_l = \boldsymbol{e}_k^T \lambda_k \boldsymbol{e}_l$$

where the left side is the same as in the first equation, because C is symmetric. The difference between the first and third equations is then

$$0 = (\lambda_k - \lambda_l) \boldsymbol{e}_k^T \boldsymbol{e}_l$$

Thus, if $(\lambda_k - \lambda_l) \neq 0$, then the eigenvectors must be orthogonal, i.e. $\boldsymbol{e}_k^T \boldsymbol{e}_l = 0$.

On the other hand, if two different eigenvectors correspond to equal eigenvalues, $\lambda_k = \lambda_l$, then any linear combination $\alpha \boldsymbol{e}_k + \beta \boldsymbol{e}_l$ is also an eigenvector, so it is always possible to make two eigenvectors to be orthogonal, by proper choices of α and β.

1.2.f Using the hint, we note that $P^T P P^{-1} = P^T I = P^T$, and $P^T P P^{-1} = I P^{-1}$, and therefore $P^T = P^{-1}$. The rest is easy:

$$
\begin{aligned}
P P^T &= P P^{-1} = I \\
C P &= (\lambda_1 \boldsymbol{e}_1, \ldots, \lambda_K \boldsymbol{e}_K) = P \Lambda \\
P^T C P &= P^T P \Lambda = \Lambda \\
C = C P P^T &= P \Lambda P^T = \boldsymbol{e}_1 \lambda_1 \boldsymbol{e}_1^T + \ldots + \boldsymbol{e}_K \lambda_K \boldsymbol{e}_K^T = \sum_{k=1}^{K} \lambda_k \boldsymbol{e}_k \boldsymbol{e}_k^T \\
C^{-1} &= (P \Lambda P^T)^{-1} = (P^T)^{-1} \Lambda^{-1} P^{-1} = P \Lambda^{-1} P^T
\end{aligned}
$$

1.8

$$
\begin{aligned}
F_Y(y) = P\left[Y \leq y\right] = P\left[X^2 \leq y\right] = \\
= P\left[-\sqrt{y} \leq X \leq +\sqrt{y}\right] = \int_{-\sqrt{y}}^{\sqrt{y}} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx
\end{aligned}
$$

$$
\begin{aligned}
f_Y(y) = \frac{d}{dy} F_Y(y) = \\
= \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(\sqrt{y}-\mu)^2}{2\sigma^2}} \frac{1}{2\sqrt{y}} - \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(-\sqrt{y}-\mu)^2}{2\sigma^2}} \frac{-1}{2\sqrt{y}} = \\
= \frac{1}{\sqrt{2\pi}\sigma} \frac{1}{2\sqrt{y}} e^{-\frac{y+\mu^2}{2\sigma^2}} \left(e^{\frac{\mu\sqrt{y}}{\sigma^2}} + e^{-\frac{\mu\sqrt{y}}{\sigma^2}} \right)
\end{aligned}
$$

Chapter 2

Conditional Probability

2.1 The place of the car is the random variable S with possible outcomes 1, 2, or 3. You first selected door 1. Then the host opens either door 2, if $S \neq 2$, or door 3, if $S \neq 3$. We note the resulting probabilities in the following table.

State $S = i$	$S = 1$	$S = 2$	$S = 3$	Sum
Prior $P(S = i) =$	1/3	1/3	1/3	1
$P(X = 2 \mid S = i) =$	1/2	0	1	
$P(X = 3 \mid S = i) =$	1/2	1	0	
$P(X = 2 \cap S = i) =$	1/6	0	1/3	1/2
$P(X = 3 \cap S = i) =$	1/6	1/3	0	1/2
$P(S = i \mid X = 2) =$	1/3	0	2/3	1

Obviously, you can double your chance of winning the car by moving to door 3, if the game show host showed that the car was not behind door 2.

2.2.a For any random variables X and Y, we have

$$E[X + Y] = E[X] + E[Y] \tag{2.1}$$

$$\text{var}[X + Y] = \text{var}[X] + \text{var}[Y] - \text{cov}[X, Y] \tag{2.2}$$

Thus, for the *independent* variables X and Y, we have

$$f_Z(z) = \frac{1}{\sigma_Z \sqrt{2\pi}} e^{-\frac{(z-\mu_Z)^2}{2\sigma_Z^2}} \tag{2.3}$$

where

$$\mu_Z = \mu_X + \mu_Y \tag{2.4}$$

$$\sigma_Z^2 = \sigma_X^2 + \sigma_Y^2 \tag{2.5}$$

2.2.b For given fixed outcome $X = x_1$, the variability of the sum Z includes only the random variability of Y, as $Z = Y + x_1$. Therefore, we can immediately determine

$$E[Z|X = x_1] = E[Y] + x_1 = \mu_Y + x_1 \tag{2.6}$$

$$\text{var}[Z|X = x_1] = \text{var}[Y] = \sigma_Y^2 \tag{2.7}$$

$$f_{Z|X}(z|x_1) = \frac{1}{\sigma_Y\sqrt{2\pi}} e^{-\frac{(z-(\mu_Y+x_1))^2}{2\sigma_Y^2}} \tag{2.8}$$

2.2.c Following the hint, we have the joint density

$$\begin{aligned} f_{Z,X}(z,x) &\propto e^{-\frac{(z-\mu_Y-x)^2}{2\sigma_Y^2} - \frac{(x-\mu_X)^2}{2\sigma_X^2}} = \\ &= e^{-\frac{x^2}{2\sigma_Y^2} - \frac{x^2}{2\sigma_X^2} + \frac{2x(z-\mu_Y)}{2\sigma_Y^2} + \frac{2x\mu_X}{2\sigma_X^2} + \ldots} \end{aligned} \tag{2.9}$$

where the ... denote a constant that is independent of z and x. The exponent is clearly a second-degree polynomial in x. Thus, the conditional density for X, given Z, is Gaussian, with some mean $\mu_{X|Z}$ and variance $\sigma_{X|Z}$ that remain to be determined. Then the conditional density function for a given $Z = z_1$ can be written as

$$f_{X|Z}(x|z_1) \propto e^{-\frac{x^2}{2\sigma_{X|Z}^2} + \frac{2x\mu_{X|Z}}{2\sigma_{X|Z}^2} + \ldots} \tag{2.10}$$

Identifying the coefficients for x^2 and x in Eqs. (**??**) and (**??**), we see that

$$\frac{1}{\sigma_{X|Z}^2} = \frac{1}{\sigma_Y^2} + \frac{1}{\sigma_X^2} \tag{2.11}$$

$$\frac{\mu_{X|Z}}{\sigma_{X|Z}^2} = \frac{z - \mu_Y}{\sigma_Y^2} + \frac{\mu_X}{\sigma_X^2} \tag{2.12}$$

$$\sigma_{X|Z}^2 = \frac{\sigma_X^2\sigma_Y^2}{\sigma_X^2 + \sigma_Y^2} \tag{2.13}$$

$$\mu_{X|Z}(z_1) = \frac{\sigma_X^2(z_1 - \mu_Y) + \sigma_Y^2\mu_X}{\sigma_X^2 + \sigma_Y^2} \tag{2.14}$$

Note that the conditional mean depends on the observed outcome $Z = z_1$, but the conditional variance is independent of z_1.

2.3.a X is $N(0,1)$, if $S = 0$, and $N(1,2^2)$, if $S = 1$. This statement can also be expressed as

$$f_{X|S}(x|i) = \begin{cases} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}, & i = 0 \\ \frac{1}{2\sqrt{2\pi}} e^{-\frac{(x-1)^2}{8}}, & i = 1 \end{cases} \tag{2.15}$$

2.3.b

$$\begin{aligned} f_X(x) &= \sum_{i=0}^{1} f_{X|S}(x|i)P_S(i) = \\ &= \frac{1}{2}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2} + \frac{1}{2}\frac{1}{2\sqrt{2\pi}}e^{-\frac{(x-1)^2}{8}} \end{aligned} \tag{2.16}$$

2.3.c

$$P_{S|X}(i|x) = \frac{f_{X|S}(x|i)P_S(i)}{f_X(x)} = \frac{f_{X|S}(x|i)P_S(i)}{\sum_{k=0}^{1} f_{X|S}(x|k)P_S(k)} \tag{2.17}$$

$$P_{S|X}(0|x) = \frac{e^{-\frac{(x)^2}{2}}\frac{1}{2}}{\frac{1}{2}e^{-\frac{(x)^2}{2}} + \frac{1}{2}\frac{1}{2}e^{-\frac{(x-1)^2}{8}}} = \frac{1}{1+\frac{1}{2}e^{-\frac{(x-1)^2}{8}+\frac{x^2}{2}}} = \tag{2.18}$$

$$= \frac{1}{1+\frac{1}{2}e^{\frac{3x^2+2x-1}{8}}} \tag{2.19}$$

For $x = 0.3$ this expression yields $P_{S|X}(0|0.3) \approx 0.670$.

2.4.a The conditional mean and variance is

$$E[Z|S=i] = iE[X] = i \tag{2.20}$$

$$\text{var}[Z|S=i] = i^2 \,\text{var}[X] = i^2 \tag{2.21}$$

Thus, the conditional density function is

$$f_{Z|S}(z|i) = \frac{1}{i\sqrt{2\pi}}e^{-\frac{(z-i)^2}{2i^2}} \tag{2.22}$$

2.4.b

$$\begin{aligned} f_Z(z) &= \sum_i f_{Z|S}(z|i)P_S(i) = \frac{0.8}{\sqrt{2\pi}}e^{-\frac{(z-1)^2}{2}} + \frac{0.2}{2\sqrt{2\pi}}e^{-\frac{(z-2)^2}{8}} = \\ &= \frac{0.8}{\sqrt{2\pi}}e^{-\frac{(z-1)^2}{2}} + \frac{0.1}{\sqrt{2\pi}}e^{-\frac{(z-2)^2}{8}} \end{aligned} \tag{2.23}$$

2.4.c Using Bayes rule as usual, the conditional probability mass for S is

$$P_{S|Z}(i|z_1) = \frac{f_{Z|S}(z_1|i)P_S(i)}{\sum_{k=1}^{2} f_{Z|S}(z_1|k)P_S(k)} \tag{2.24}$$

$$\begin{aligned} P_{S|Z}(1|z_1) &= \frac{\frac{0.8}{\sqrt{2\pi}}e^{-\frac{(z_1-1)^2}{2}}}{\frac{0.8}{\sqrt{2\pi}}e^{-\frac{(z_1-1)^2}{2}} + \frac{0.1}{\sqrt{2\pi}}e^{-\frac{(z_1-2)^2}{8}}} = \\ &= \frac{8}{8 + e^{-\frac{(z_1-2)^2}{8}+\frac{(z_1-1)^2}{2}}} = \\ &= \frac{8}{8 + e^{(3z_1^2-4z_1)/8}} \end{aligned} \tag{2.25}$$

$$\begin{aligned} P_{S|Z}(2|z_1) &= 1 - P_{S|Z}(1|z_1) = \\ &= \frac{1}{1 + 8e^{(-3z_1^2+4z_1)/8}} \end{aligned} \tag{2.26}$$

2.4.d The conditional probability for X is most easily expressed as a probability *mass* function, as only two discrete values of X are possible, either $X = z_1$, if $S = 1$, or $X = z_1/2$, if $S = 2$. Thus, using the result from the previous sub-problem,

$$P_{X|Z}(z_1|z_1) = P_{S|Z}(1|z_1) = \frac{8}{8 + e^{(3z_1^2-4z_1)/8}} \tag{2.27}$$

$$P_{X|Z}(z_1/2|z_1) = P_{S|Z}(2|z_1) = \frac{1}{1 + 8e^{(-3z_1^2+4z_1)/8}} \tag{2.28}$$

$$P_{X|Z}(x|z_1) = 0, \quad \text{otherwise} \tag{2.29}$$

The probability mass function is now properly normalized to give the correct sum across all possible outcomes, as

$$\sum_{x\in\{z_1,z_1/2\}} P_{X|Z}(x|z_1) = 1 \tag{2.30}$$

This probability *mass* function can also be formally expressed as a probability *density* function, using a Dirac δ, as

$$f_{X|Z}(x|z_1) = \frac{8\delta(x - z_1)}{8 + e^{(3z_1^2-4z_1)/8}} + \frac{\delta(x - z_1/2)}{1 + 8e^{(-3z_1^2+4z_1)/8}} \tag{2.31}$$

Here, the probability density is also properly normalized to give the correct integral across all possible outcomes, as

$$\int_{-\infty}^{\infty} f_{X|Z}(x|z_1)dx = 1 \tag{2.32}$$

Chapter 3

Bayesian Classification

3.1

$$\mathrm{P}_c = \mathrm{P}(\{12345\} \mid computer) = \mathrm{P}(\{24153\} \mid computer)$$

That is, the probability that the sequence {12345} is generated by a computer is equal to the probability that the sequence {24153}is generated by a computer equals to P_c.
Let P_{h_1} be the probability that the sequence {12345} is generated by a human and P_{h_2} be the probability that the sequence {24153} is generated by a human. It is resonable to assume that $\mathrm{P}_{h_1} > \mathrm{P}_{h_2}$.

$$\mathrm{P}_{h_1} = \mathrm{P}(\{12345\} \mid human)$$

$$\mathrm{P}_{h_2} = \mathrm{P}(\{24153\} \mid human)$$

$$S = \begin{cases} 1 & : \quad x(human; computer) \\ 2 & : \quad x(computer; human) \end{cases}$$

$$\mathrm{P}_{S|X}(1 \mid x) \quad = \quad \frac{\mathrm{P}_{X|S}(x \mid 1)\mathrm{P}_S(1)}{\mathrm{P}_X(x)} \tag{3.1}$$

$$\begin{aligned} \mathrm{P}_{X|S}(x \mid 1) &= \mathrm{P}((\{12345\}; \{24153\}) \mid (human; computer)) \text{the two sequences are independent} \\ &= \mathrm{P}(\{12345\} \mid human)\mathrm{P}(\{24153\} \mid computer) \\ &= \mathrm{P}_{h_1}\mathrm{P}_c \end{aligned}$$

similarly

$$\mathrm{P}_{X|S}(x \mid 2) = \mathrm{P}_{h_2}\mathrm{P}_c$$

$$\begin{aligned} \mathrm{P}_X(x) &= \sum_{i=1}^{2} \mathrm{P}_{X|S}(x \mid i)\mathrm{P}_S(i) \\ &= \frac{1}{2}\mathrm{P}_{h_1}\mathrm{P}_c + \frac{1}{2}\mathrm{P}_{h_2}\mathrm{P}_c \end{aligned}$$

substitute in 2

$$\begin{aligned} \mathrm{P}_{S|X}(1 \mid x) &= \frac{\frac{1}{2}\mathrm{P}_{h_1}\mathrm{P}_c}{\frac{1}{2}\mathrm{P}_{h_1}\mathrm{P}_c + \frac{1}{2}\mathrm{P}_{h_2}\mathrm{P}_c} \\ &= \frac{\mathrm{P}_{h_1}}{\mathrm{P}_{h_1} + \mathrm{P}_{h_2}} \end{aligned}$$

similarly

$$\mathrm{P}_{S|X}(2 \mid x) = \frac{\mathrm{P}_{h_2}}{\mathrm{P}_{h_1} + \mathrm{P}_{h_2}}$$

Optimal MAP decision about the source is achieved by choosing the source category corresponding to the maximum aposteriori probability.
Since $\mathrm{P}_{h_1} > \mathrm{P}_{h_2} \Rightarrow \mathrm{P}_{S|X}(1 \mid x) > \mathrm{P}_{S|X}(2 \mid x)$ and hence we decide on (human;computer) corresponding to the source category $S = 1$

3.2.a

$$f_{X|S}(x, i) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu_i)^2/2\sigma^2}$$

To obtain minimum error probability when the a priori probabilities are equal,we use the ML decision rule.
The discriminant function can be defined as

$$\begin{aligned} g_i(x) &= \ln f_{X|S}(x, i) \\ &= \ln \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu_i)^2/2\sigma^2} \\ &= \ln \frac{1}{\sqrt{2\pi}\sigma} - \frac{(x-\mu_i)^2}{2\sigma^2} \end{aligned}$$

The first term of the above equation is constant and can be removed from the discriminant function.

$$g_i(x) = -\frac{x^2 - 2x\mu_i + \mu_i^2}{2\sigma^2}$$

The term x^2 appears identically in all discriminant functions, so we can remove it to get

$$g_i(x) = x\mu_i - \frac{\mu_i^2}{2}$$

Thus the two discriminant functions is written as

$$g_1(x) = x\mu_1 - \frac{\mu_1^2}{2}$$

$$g_2(x) = x\mu_2 - \frac{\mu_2^2}{2}$$

we can combine the two functions to form single discriminant function as

$$\begin{aligned} g(x) &= g_1(x) - g_2(x) \\ &= (\mu_1 - \mu_2)\left[x - \frac{\mu_1 + \mu_2}{2}\right] \end{aligned}$$

The decision rule is then formulated as

$$d(x) = \begin{cases} 1 & \text{if } g(x) > 0 \\ 2 & \text{otherwise} \end{cases}$$

or

$$d(x) = \begin{cases} 1 & \text{if } sgn(x - x_{th}) = sgn(\mu_1 - \mu_2) \\ 2 & \text{otherwise} \end{cases}$$

where $x_{th} = \frac{\mu_1+\mu_2}{2}$, assume $\mu_1 > \mu_2$, the decision rule will be

$$d(x) = \begin{cases} 1 & \text{if } x > x_{th} \\ 2 & \text{otherwise} \end{cases}$$

3.2.b

$$\begin{aligned} \mathrm{P}(error) &= \mathrm{P}(d(x) = 1 \cap S = 2) + \mathrm{P}(d(x) = 2 \cap S = 1) \\ &= \mathrm{P}(d(x) = 1 \mid S = 2)\mathrm{P}_S(2) + \mathrm{P}(d(x) = 2 \mid S = 1)\mathrm{P}_S(1) \\ &= \mathrm{P}(x > x_{th} \mid S = 2) \times \frac{1}{2} + \mathrm{P}(x < x_{th} \mid S = 1) \times \frac{1}{2} \\ &= \frac{1}{2} \times \mathrm{P}(x > 0 \mid S = 2) + \frac{1}{2} \times \mathrm{P}(x < 0 \mid S = 1) \\ &= \frac{1}{2} \times \mathrm{P}(x + 2 > 2 \mid S = 2) + \frac{1}{2} \times \mathrm{P}(x - 2 < -2 \mid S = 1) \\ &= \frac{1}{2} \times (1 - \Phi(2)) + \frac{1}{2} \times \Phi(-2) \\ &= (1 - \Phi(2)) \\ &= 1 - 0.9772 \\ &= 0.023 \end{aligned}$$

3.3.a *Decision Criterion*: We must use the Minimum Risk in this case. *Discriminant functions*: The conditional expected loss function for decisions

$i = 1 \dots N$ is given by

$$\begin{aligned} R(i \mid \boldsymbol{x}) &= \sum_{j=1}^{N} L(d(\boldsymbol{x}) = i \mid S = j) P_{S|\boldsymbol{X}}(j \mid \boldsymbol{x}) \\ &= 0 + \sum_{j(i \neq j)}^{N} c P_{S|\boldsymbol{X}}(j \mid \boldsymbol{x}) \\ &= c(1 - P_{S|\boldsymbol{X}}(i \mid \boldsymbol{x})) \end{aligned}$$

For decision $i = N + 1$ the cost is $R(i \mid \boldsymbol{x}) = r$, thus we have

$$R(i \mid \boldsymbol{x}) = \begin{cases} c(1 - P_{S|\boldsymbol{X}}(i \mid \boldsymbol{x})), & i = 1 \dots N \\ r, & i = N + 1 \end{cases}$$

The minimum-risk decision can be equivalently expressed as

$$d(\boldsymbol{x}) = \operatorname*{argmax}_i g'_i(\boldsymbol{x})$$

with preliminary discriminant functions $g'_i(\boldsymbol{x}) = -R(i \mid \boldsymbol{x})$, i.e.,

$$g'_i(\boldsymbol{x}) = \begin{cases} c P_{S|\boldsymbol{X}}(i \mid \boldsymbol{x}) - c, & i = 1, \dots, N \\ -r, & i = N + 1 \end{cases}$$

Simplify Discriminant functions: The decision remains optimal if we scale all discriminant functions with the same positive factor, or add some constant to all of them. Thus, an equivalent set of discriminant functions are

$$g''_i(\boldsymbol{x}) = \begin{cases} P_{S|\boldsymbol{X}}(i \mid \boldsymbol{x}), & i = 1, \dots, N \\ 1 - \frac{r}{c}, & i = N + 1 \end{cases}$$

The decision rule $d(\boldsymbol{x}) = \operatorname{argmax}_i g''_i(\boldsymbol{x})$ is equivalent to the rule formulated in the problem text.

3.3.b if $r = 0$ this means that the cost of a reject is always less than the cost of an erroneous decision, thus we always choose $N + 1$

3.3.c if $r > c$ never choose $N + 1$

3.3.d *Simplify discriminant functions*, continued: We know from part 3.3.a that the discriminant functions $g''_i(\boldsymbol{x})$ give minimum expected cost. We continue to transform all of them. For this purpose, we apply Bayes' rule

$$\begin{aligned} f_{\boldsymbol{X}}(\boldsymbol{x}) P_{S|\boldsymbol{X}}(i \mid \boldsymbol{x}) &= f_{\boldsymbol{X}|S}(\boldsymbol{x} \mid i) P_S(i) \\ f_{\boldsymbol{X}}(\boldsymbol{x}) &= \sum_{j=1}^{N} f_{\boldsymbol{X}|S}(\boldsymbol{x} \mid j) P_S(j) \end{aligned}$$

As $f_{\boldsymbol{X}}(\boldsymbol{x})$ is positive, we can scale all the preliminary discriminant functions as $g_i(\boldsymbol{x}) = f_{\boldsymbol{X}}(\boldsymbol{x})g_i''(\boldsymbol{x})$, to get

$$g_i(\boldsymbol{x}) = f_{\boldsymbol{X}}(\boldsymbol{x})P_{S|\boldsymbol{X}}(i \mid \boldsymbol{x}) = f_{\boldsymbol{X}|S}(\boldsymbol{x} \mid i)P_S(i), \quad i = 1, \ldots, N$$
$$g_{N+1}(\boldsymbol{x}) = \left(1 - \frac{r}{c}\right) f_{\boldsymbol{X}}(\boldsymbol{x}) = \left(1 - \frac{r}{c}\right) \sum_{j=1}^{N} f_{\boldsymbol{X}|S}(\boldsymbol{x} \mid j)P_S(j)$$

These functions are precisely the ones defined in the problem text.

3.4.a

$$\begin{aligned} \boldsymbol{\mu}_{\boldsymbol{X}} = E[\boldsymbol{X}] &= E[\boldsymbol{a}_i + \boldsymbol{W}] \\ &= E[\boldsymbol{a}_i] + E[\boldsymbol{W}] \\ &= \boldsymbol{a}_i \end{aligned}$$

$$\begin{aligned} C = \operatorname{cov}[\boldsymbol{X}] &= E[(\boldsymbol{X} - \boldsymbol{\mu}_{\boldsymbol{X}})(\boldsymbol{X} - \boldsymbol{\mu}_{\boldsymbol{X}})^T] \\ &= E[(\boldsymbol{a}_i + \boldsymbol{W} - \boldsymbol{a}_i)(\boldsymbol{a}_i + \boldsymbol{W} - \boldsymbol{a}_i)^T] \\ &= E[\boldsymbol{W}\boldsymbol{W}^T] \\ &= E\begin{bmatrix} w_1 w_1 & w_1 w_2 \\ w_2 w_1 & w_2 w_2 \end{bmatrix} \\ &= \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \end{aligned}$$

$$f_{\boldsymbol{X}|S}(\boldsymbol{x} \mid i) = \frac{1}{2\pi\sqrt{\det(C)}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{a}_i)^T C^{-1}(\boldsymbol{x}-\boldsymbol{a}_i)}$$

3.4.c

$$\begin{aligned} g_i(\boldsymbol{x}) &= \ln(f_{\boldsymbol{X}|S}(\boldsymbol{x} \mid i)\mathrm{P}_S(i)) \\ &= \ln(\frac{1}{2\pi\sqrt{\det(C)}}) - \frac{1}{2}(\boldsymbol{x} - \boldsymbol{a}_i)^T C^{-1}(\boldsymbol{x} - \boldsymbol{a}_i) + \ln \mathrm{P}_S(i) \\ &= -\frac{1}{2}[\boldsymbol{x}^T C^{-1}\boldsymbol{x} - \boldsymbol{x}^T C^{-1}\boldsymbol{a}_i - \boldsymbol{a}_i^T C^{-1}\boldsymbol{x} + \boldsymbol{a}_i^T C^{-1}\boldsymbol{a}_i] + \ln \mathrm{P}_S(i) \\ &= -\frac{1}{2}[-2\boldsymbol{a}_i^T C^{-1}\boldsymbol{x} + \boldsymbol{a}_i^T C^{-1}\boldsymbol{a}_i] + \ln \mathrm{P}_S(i) \\ &= \boldsymbol{a}_i^T C^{-1}\boldsymbol{x} - \frac{\boldsymbol{a}_i^T C^{-1}\boldsymbol{a}_i}{2} + \ln \mathrm{P}_S(i) \end{aligned}$$

The two discriminant functions are

$$\begin{aligned} g_0(\boldsymbol{x}) &= \boldsymbol{a}_0^T C^{-1}\boldsymbol{x} - \frac{\boldsymbol{a}_0^T C^{-1}\boldsymbol{a}_0}{2} + \ln \mathrm{P}_S(0) \\ g_1(\boldsymbol{x}) &= \boldsymbol{a}_1^T C^{-1}\boldsymbol{x} - \frac{\boldsymbol{a}_1^T C^{-1}\boldsymbol{a}_1}{2} + \ln \mathrm{P}_S(1) \end{aligned}$$

We can define a single discriminant function as

$$\begin{aligned} g'(\boldsymbol{x}) &= g_0(\boldsymbol{x}) - g_1(\boldsymbol{x}) \\ &= (\boldsymbol{a}_0^T - \boldsymbol{a}_1^T)C^{-1}\boldsymbol{x} + \ln\frac{\mathrm{P}_S(0)}{\mathrm{P}_S(1)} - \frac{\boldsymbol{a}_0^T C^{-1}\boldsymbol{a}_0}{2} + \frac{\boldsymbol{a}_1^T C^{-1}\boldsymbol{a}_1}{2} \\ &= \boldsymbol{q}^T\boldsymbol{x} - y_{th} \end{aligned}$$

The decision function is then given by

$$d(\boldsymbol{x}) = \begin{cases} 0 & \text{if } g(\boldsymbol{x}) > 0 \\ 1 & \text{otherwise} \end{cases}$$

we can define a slightly modified discriminant function as $Y = g(\boldsymbol{X}) = \boldsymbol{q}^T\boldsymbol{X}$. The decision function will then be

$$d(\boldsymbol{x}) = \begin{cases} 0 & \text{if } Y > y_{th} \\ 1 & \text{if } Y < y_{th} \end{cases}$$

Where

$$y_{th} = -\ln\frac{\mathrm{P}_S(0)}{\mathrm{P}_S(1)} + \frac{\boldsymbol{a}_0^T C^{-1}\boldsymbol{a}_0}{2} - \frac{\boldsymbol{a}_1^T C^{-1}\boldsymbol{a}_1}{2}$$

For $\boldsymbol{a}_0 = \begin{pmatrix} 5 \\ 0 \end{pmatrix}$, $\boldsymbol{a}_1 = \begin{pmatrix} 0 \\ 5 \end{pmatrix}$, and $C = \begin{pmatrix} 100 & 50 \\ 50 & 100 \end{pmatrix}$, we have

$$\begin{aligned} \boldsymbol{q}^T &= (5,0) - (0,5)\begin{pmatrix} 100 & 50 \\ 50 & 100 \end{pmatrix}^{-1} \\ &= (5,-5)\frac{1}{7500}\begin{pmatrix} 100 & -50 \\ -50 & 100 \end{pmatrix} \\ &= \frac{1}{150}(5,-5)\begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix} \\ &= \frac{1}{10}(1,-1) \end{aligned}$$

3.5.a *Criterion*: As both source categories are equally probable, we use the ML criterion. The conditional density functions are

$$\begin{aligned} f_{\boldsymbol{X}|S}(\boldsymbol{x} \mid i) &= \frac{1}{\sqrt{2\pi}^K\sqrt{\det(\boldsymbol{C})}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{a}_i)^T\boldsymbol{C}^{-1}(\boldsymbol{x}-\boldsymbol{a}_i)} \\ &= \frac{1}{\sqrt{2\pi\sigma^2}^K} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{a}_i)^T\frac{1}{\sigma^2}\boldsymbol{I}(\boldsymbol{x}-\boldsymbol{a}_i)} \end{aligned}$$

Discriminant functions: We define preliminary discriminant functions

$$\begin{aligned} g_i'(\boldsymbol{x}) &= \ln f_{\boldsymbol{X}|S}(\boldsymbol{x} \mid i) = \\ &= \ln(\frac{1}{\sqrt{2\pi\sigma^2}^K}) - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{a}_i)^T \frac{1}{\sigma^2}\boldsymbol{I}(\boldsymbol{x}-\boldsymbol{a}_i) \\ &= \ln(\frac{1}{\sqrt{2\pi\sigma^2}^K}) - \frac{1}{2\sigma^2}[\boldsymbol{x}^T\boldsymbol{x} - \boldsymbol{x}^T\boldsymbol{a}_i - \boldsymbol{a}_i^T\boldsymbol{x} + \boldsymbol{a}_i^T\boldsymbol{a}_i] \end{aligned}$$

Simplify discriminant functions: Omitting the second-degree terms and constant terms and scale factors that are equal among the discriminant functions, we define a new set of functions

$$\begin{aligned} g_i(\boldsymbol{x}) &= -\frac{1}{2}[-\boldsymbol{x}^T\boldsymbol{a}_i - \boldsymbol{a}_i^T\boldsymbol{x} + \boldsymbol{a}_i^T\boldsymbol{a}_i] \\ &= -\frac{1}{2}[-2\boldsymbol{a}_i^T\boldsymbol{x} + \boldsymbol{a}_i^T\boldsymbol{a}_i]) \\ &= \boldsymbol{a}_i^T\boldsymbol{x} - \frac{\boldsymbol{a}_i^T\boldsymbol{a}_i}{2} \end{aligned}$$

The two discriminant functions are

$$\begin{aligned} g_1(\boldsymbol{x}) &= \boldsymbol{a}_1^T\boldsymbol{x} - \frac{\boldsymbol{a}_1^T\boldsymbol{a}_1}{2} \\ g_2(\boldsymbol{x}) &= \boldsymbol{a}_2^T\boldsymbol{x} - \frac{\boldsymbol{a}_2^T\boldsymbol{a}_2}{2} \end{aligned}$$

We can define a single discriminant function as

$$\begin{aligned} g(\boldsymbol{x}) &= g_1(\boldsymbol{x}) - g_2(\boldsymbol{x}) \\ &= \boldsymbol{a}_1^T\boldsymbol{x} - \frac{\boldsymbol{a}_1^T\boldsymbol{a}_1}{2} - \boldsymbol{a}_2^T\boldsymbol{x} + \frac{\boldsymbol{a}_2^T\boldsymbol{a}_2}{2} \\ &= (\boldsymbol{a}_1 - \boldsymbol{a}_2)^T\left[\boldsymbol{x} - \frac{\boldsymbol{a}_1 + \boldsymbol{a}_2}{2}\right] \end{aligned}$$

Decision rule: The final decision function is then

$$\begin{aligned} d(\boldsymbol{x}) &= \begin{cases} 1, & g(\boldsymbol{x}) > 0 \\ 2, & g(\boldsymbol{x}) < 0 \end{cases} \\ &= \begin{cases} 1, & (\boldsymbol{a}_1 - \boldsymbol{a}_2)^T(\boldsymbol{x} - \boldsymbol{x}_{\text{th}}) > 0 \\ 2, & \text{otherwise} \end{cases} \end{aligned}$$

where $\boldsymbol{x}_{\text{th}} = (\boldsymbol{a}_1 + \boldsymbol{a}_2)/2$.

3.5.b To find the minimum probability of error, define a scalar random decision variable Y as

$$Y = g(\boldsymbol{X}) = (\boldsymbol{a}_1 - \boldsymbol{a}_2)^T(\boldsymbol{X} - \frac{\boldsymbol{a}_1 + \boldsymbol{a}_2}{2})$$

Y is a linear combination of the Gaussian random variable elements X_k of the random vector $\boldsymbol{X}$, and is therefore also Gaussian.

$$\begin{aligned}
\mu_{Y,1} = E\left[Y \mid S=1\right] &= E\left[(\boldsymbol{a}_1-\boldsymbol{a}_2)^T(\boldsymbol{X}-\frac{\boldsymbol{a}_1+\boldsymbol{a}_2}{2})\right] \\
&= (\boldsymbol{a}_1-\boldsymbol{a}_2)^T(E\left[\boldsymbol{X} \mid S=1\right]-\frac{\boldsymbol{a}_1+\boldsymbol{a}_2}{2}) \\
&= (\boldsymbol{a}_1-\boldsymbol{a}_2)^T(\boldsymbol{a}_1-\frac{\boldsymbol{a}_1+\boldsymbol{a}_2}{2}) \\
&= (\boldsymbol{a}_1-\boldsymbol{a}_2)^T(\boldsymbol{a}_1-\boldsymbol{a}_2)/2 \\
&= \frac{\|\boldsymbol{a}_1-\boldsymbol{a}_2\|^2}{2} \\
&= \frac{d^2}{2}
\end{aligned}$$

$$\mu_{Y,2} = E[Y \mid S=2] = -\mu_{Y,1} = -\frac{d^2}{2}$$

The variance of Y, given signal category $S=i$, is

$$\begin{aligned}
\text{var}\left[Y \mid S=i\right] &= E\left[(Y-\mu_{Y,i})^2 \mid S=i\right] = \\
&= E\left[(\boldsymbol{a}_1-\boldsymbol{a}_2)^T(\boldsymbol{X}-\boldsymbol{\mu}_{X,i})(\boldsymbol{X}-\boldsymbol{\mu}_{X,i})^T(\boldsymbol{a}_1-\boldsymbol{a}_2) \mid S=i\right] = \\
&= (\boldsymbol{a}_1-\boldsymbol{a}_2)^T \underbrace{E\left[(\boldsymbol{X}-\boldsymbol{\mu}_{X,i})(\boldsymbol{X}-\boldsymbol{\mu}_{X,i})^T \mid S=i\right]}_{\text{cov}[\boldsymbol{X}]}(\boldsymbol{a}_1-\boldsymbol{a}_2) = \\
&= (\boldsymbol{a}_1-\boldsymbol{a}_2)^T(\sigma^2\boldsymbol{I})(\boldsymbol{a}_1-\boldsymbol{a}_2) = \\
&= \|\boldsymbol{a}_1-\boldsymbol{a}_2\|^2\sigma^2
\end{aligned}$$

We see that the variance is actually the same, regardless of the signal category S. To simplify the calculations we can transform the decision variable Y by dividing with its standard deviation $d\sigma$, to get a new variable $Z = Y/d\sigma$, which has variance equal to 1

$$\begin{aligned}
\mu_{Z,1} &= E\left[Z \mid S=1\right] = E[Y/d\sigma \mid S=1] = \frac{d}{2\sigma} = \frac{d'}{2} \\
\mu_{Z,2} &= -\mu_{Z,1} = -\frac{d}{2\sigma} = -\frac{d'}{2}
\end{aligned}$$

The probability of error is given by

$$\begin{aligned}
P_e &= P\left[Y > 0 \mid S = 2\right] P_S(2) + P\left[Y \leq 0 \mid S = 1\right] P_S(1) = \\
&= P\left[Z > 0 \mid S = 2\right] P_S(2) + P\left[Z \leq 0 \mid S = 1\right] P_S(1) = \\
&= P\left[Z + \frac{d}{2\sigma} > 0 + \frac{d}{2\sigma} \mid S = 2\right] P_S(2) + \\
&+ P\left[Z - \frac{d}{2\sigma} \leq 0 - \frac{d}{2\sigma} \mid S = 1\right] P_S(1) = \\
&= \frac{1}{2}(1 - \Phi(d/2\sigma)) + \frac{1}{2}\Phi(-d/2\sigma) = \\
&= \frac{1}{2}(1 - \Phi(d/2\sigma)) + \frac{1}{2}(1 - \Phi(d/2\sigma)) = 1 - \Phi(d/2\sigma)
\end{aligned}$$

3.6.a For the two possible source categories, (coin#0 or coin#1) we define two discriminant functions as

$$\begin{aligned}
g_0(\underline{x}) = \mathrm{P}_{\underline{X}|S}(\underline{x} \mid 0) &= (1/2)^K \\
g_1(\underline{x}) = \mathrm{P}_{\underline{X}|S}(\underline{x} \mid 1) &= p^k(1-p)^{K-k}
\end{aligned}$$

We can combine the two functions into one single function defined as

$$g(\underline{x}) = g_1(\underline{x}) - g_0(\underline{x})$$

The decision function is defined as

$$d(\underline{x}) = \begin{cases} 1 & \text{if } g(\underline{x}) > 0 \\ 0 & \text{if } g(\underline{x}) \leq 0 \end{cases}$$

$$\begin{aligned}
g_1(\underline{x}) - g_0(\underline{x}) &> 0 \\
p^k(1-p)^{K-k} &> (1/2)^K \\
\ln(p^k(1-p)^{K-k}) &> \ln(1/2)^K \\
k\ln p + (K-k)\ln(1-p) &> K\ln(1/2) \\
k &> \frac{K\ln\frac{1}{2(1-p)}}{\ln\frac{p}{1-p}}
\end{aligned}$$

The decision function is then defined as

$$d(\underline{x}) = \begin{cases} 1 & \text{if } k > k_{\text{th}} \\ 0 & \text{if } k \leq k_{\text{th}} \end{cases}$$

where $k_{\text{th}} = \frac{K\ln\frac{1}{2(1-p)}}{\ln\frac{p}{1-p}}$.

3.6.b For $K = 5$ and $p = 0.6$ we have

$$k_{\text{th}} = \frac{5 \ln(\frac{1}{2(1-0.6)})}{\ln(\frac{0.6}{1-0.6})} \approx 2.7517$$

$$\mathrm{P}_{\tilde{K}|S} \sim \text{Binomial} \begin{pmatrix} K \\ p \end{pmatrix}$$

Thus

$$\mathrm{P}_{\tilde{K}|S}(k \mid 0) \sim \text{Binomial} \begin{pmatrix} K \\ 0.5 \end{pmatrix}$$

$$\mathrm{P}_{\tilde{K}|S}(k \mid 1) \sim \text{Binomial} \begin{pmatrix} K \\ 0.6 \end{pmatrix}$$

The probability of error is given by

$$\begin{aligned}
\mathrm{P}(\mathit{error}) &= \mathrm{P}(\tilde{K} \leq k_{\text{th}} \mid S = 1)\mathrm{P}_S(1) + \mathrm{P}(\tilde{K} > k_{\text{th}} \mid S = 0)\mathrm{P}_s(0) \\
&= \frac{1}{2}\mathrm{P}(\tilde{K} \leq 2 \mid S = 1) + \frac{1}{2}\mathrm{P}(\tilde{K} \geq 3 \mid S = 0) \\
&= \frac{1}{2}\sum_{k=0}^{2} \mathrm{P}_{\tilde{K}|S}(k \mid 1) + \frac{1}{2}\sum_{k=3}^{5} \mathrm{P}_{\tilde{K}|S}(k \mid 0) \\
&= \frac{1}{2}\sum_{k=0}^{2} \begin{pmatrix} 5 \\ k \end{pmatrix} 0.6^k 0.4^{5-k} + \frac{1}{2}\sum_{k=3}^{5} \begin{pmatrix} 5 \\ k \end{pmatrix} (1/2)^5 \\
&\approx 0.4087
\end{aligned}$$

3.7 *Decision Criterion*: As the source categories are not necessarily equally probable, we must use the MAP criterion.

Discriminant functions: We have an observed feature vector $\boldsymbol{x}$, considered as an outcome of the random vector $\boldsymbol{X}$ with independent binary elements X_k, with exactly known probability mass distribution

$$f_{X_k|S}(x_k \mid j) = \begin{cases} p_{kj}, & x_k = 1 \\ 1 - p_{kj}, & x_k = 0 \end{cases}$$

This can be more conveniently written as

$$f_{X_k|S}(x_k \mid j) = p_{kj}^{x_k}(1 - p_{kj})^{1-x_k}$$

As the feature elements are independent, the conditional probability mass distributions for the complete feature vector are

$$f_{\boldsymbol{X}|S}(\boldsymbol{x} \mid j) = \prod_{k=1}^{K} p_{kj}^{x_k}(1 - p_{kj})^{1-x_k}, \quad j \in \{1, \ldots, N_s\}$$

Using the MAP rule, discriminant functions can be defined as

$$\begin{aligned} g_j(\boldsymbol{x}) = \ln f_{\boldsymbol{X}|S}(\boldsymbol{x} \mid j) P_S(j) = \\ = \ln P_S(j) + \sum_{k=1}^{K} x_k \ln p_{kj} + (1 - x_k) \ln(1 - p_{kj}) = \\ = \ln P_S(j) + \sum_{k=1}^{K} x_k \left(\ln p_{kj} - \ln(1 - p_{kj})\right) + \ln(1 - p_{kj}) = \\ = \boldsymbol{w}_j^T \boldsymbol{x} + w_{0j} \end{aligned}$$

where $\boldsymbol{w}_j$ is a vector with elements

$$w_{kj} = \ln p_{kj} - \ln(1 - p_{kj}) = \ln \frac{p_{kj}}{1 - p_{kj}}$$

and

$$w_{0j} = \ln P_S(j) + \sum_{k=1}^{K} \ln(1 - p_{kj})$$

These discriminant functions are already linear, as desired, and do not need further simplifications.

Decision function: With these discriminant functions, the optimal decision function is

$$d(\boldsymbol{x}) = \underset{j \in \{1,\ldots,N_s\}}{\operatorname{argmax}} \; g_j(\boldsymbol{x})$$

3.8.a The mean and covariance of $\boldsymbol{X}$ are

$$\begin{aligned} \boldsymbol{\mu}_{\boldsymbol{X}} = E[\boldsymbol{X}] &= E[\boldsymbol{u}_S + \boldsymbol{W}] \\ &= E[\boldsymbol{u}_S] + E[\boldsymbol{W}] \\ &= \boldsymbol{u}_S \end{aligned}$$

$$\begin{aligned} \operatorname{cov}(\boldsymbol{X}) &= E[(\boldsymbol{X} - \boldsymbol{\mu}_{\boldsymbol{X}})(\boldsymbol{X} - \boldsymbol{\mu}_{\boldsymbol{X}})^T] \\ &= E[(\boldsymbol{u}_S + \boldsymbol{W} - \boldsymbol{u}_S)(\boldsymbol{u}_S + \boldsymbol{W} - \boldsymbol{u}_S)^T] \\ &= E[\boldsymbol{W}\boldsymbol{W}^T] \\ &= C_W \end{aligned}$$

The conditional probability density of the vector $\boldsymbol{X}$ is given by

$$f_{\boldsymbol{X}|S}(\boldsymbol{x} \mid j) = \frac{1}{\sqrt{2\pi}^L \sqrt{\det(C_W)}} e^{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{u}_j)^T C_W^{-1} (\boldsymbol{x} - \boldsymbol{u}_j)}$$

The MAP discriminant functions are defined as

$$\begin{aligned} g_j(\boldsymbol{x}) &= \ln(f_{\boldsymbol{X}|S}(\boldsymbol{x} \mid j)\mathrm{P}_S(j)) \\ &= \ln \frac{1}{\sqrt{2\pi}^L \sqrt{\det(C_W)}} - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{u}_j)^T C_W^{-1}(\boldsymbol{x}-\boldsymbol{u}_j) + \ln \mathrm{P}_S(j) \\ &= -\frac{1}{2}(\boldsymbol{x}^T C_W^{-1}\boldsymbol{x} - \boldsymbol{x}^T C_W^{-1}\boldsymbol{u}_j - \boldsymbol{u}_j^T C_W^{-1}\boldsymbol{x} + \boldsymbol{u}_j^T C_W^{-1}\boldsymbol{u}_j) + \ln \mathrm{P}_S(j) \\ &= \boldsymbol{u}_j^T C_W^{-1}\boldsymbol{x} - \frac{\boldsymbol{u}_j^T C_W^{-1}\boldsymbol{u}_j}{2} + \ln \mathrm{P}_S(j) \end{aligned}$$

where we have used the symmetry of C_W and additive terms that do not depend on j have been ignored. We thus get

$$\begin{aligned} g_0(\boldsymbol{x}) &= \boldsymbol{u}_0^T C_W^{-1}\boldsymbol{x} - \frac{\boldsymbol{u}_0^T C_W^{-1}\boldsymbol{u}_0}{2} + \ln \mathrm{P}_S(0) \\ g_1(\boldsymbol{x}) &= \boldsymbol{u}_1^T C_W^{-1}\boldsymbol{x} - \frac{\boldsymbol{u}_1^T C_W^{-1}\boldsymbol{u}_1}{2} + \ln \mathrm{P}_S(1) \end{aligned}$$

The two discriminant functions can be combined into a single function as

$$\begin{aligned} g(\boldsymbol{x}) &= g_0(\boldsymbol{x}) - g_1(\boldsymbol{x}) \\ &= \boldsymbol{u}_0^T C_W^{-1}\boldsymbol{x} - \boldsymbol{u}_1^T C_W^{-1}\boldsymbol{x} - \frac{\boldsymbol{u}_0^T C_W^{-1}\boldsymbol{u}_0}{2} + \frac{\boldsymbol{u}_1^T C_W^{-1}\boldsymbol{u}_1}{2} + \ln \frac{\mathrm{P}_S(0)}{\mathrm{P}_S(1)} \\ &= (\boldsymbol{u}_0^T - \boldsymbol{u}_1^T) C_W^{-1}\boldsymbol{x} - \frac{\boldsymbol{u}_0^T C_W^{-1}\boldsymbol{u}_0}{2} + \frac{\boldsymbol{u}_1^T C_W^{-1}\boldsymbol{u}_1}{2} + \ln \frac{\mathrm{P}_S(0)}{\mathrm{P}_S(1)} \\ &= \boldsymbol{q}^T\boldsymbol{x} - y_{\mathrm{th}} \end{aligned}$$

where

$$y_{\mathrm{th}} = \frac{1}{2}\left(\boldsymbol{u}_0^T C_W^{-1}\boldsymbol{u}_0 - \boldsymbol{u}_1^T C_W^{-1}\boldsymbol{u}_1\right) + \ln \frac{\mathrm{P}_S(1)}{\mathrm{P}_S(0)}$$

We can define a new discriminant function as $Y = \boldsymbol{q}^T\boldsymbol{X}$ and then the decision function will be

$$d(\boldsymbol{x}) = \begin{cases} 0 & \text{if } Y > y_{\mathrm{th}} \\ 1 & \text{if } Y \leq y_{\mathrm{th}} \end{cases}$$

3.8.b The previous calculations show that

$$\boldsymbol{q} = C_W^{-1}(\boldsymbol{u}_0 - \boldsymbol{u}_1)$$

3.9.a For $S = i \in \{1, 2\}$ we have the Gaussian probability density function

$$f_{\boldsymbol{X}|S}(\boldsymbol{x} \mid i) = \frac{1}{2\pi} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^T(\boldsymbol{x}-\boldsymbol{\mu}_i)}$$

Define two discriminant functions as

$$\begin{aligned} g_i(\boldsymbol{x}) &= \ln f_{\boldsymbol{X}|S}(\boldsymbol{x} \mid i) \\ &= \ln(1/2\pi) - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^T(\boldsymbol{x}-\boldsymbol{\mu}_i) \\ &= -\frac{1}{2}(\boldsymbol{x}^T\boldsymbol{x} - 2\boldsymbol{\mu}_i^T\boldsymbol{x} + \boldsymbol{\mu}_i^T\boldsymbol{\mu}_i) \\ &= \boldsymbol{\mu}_i^T\boldsymbol{x} - \frac{\boldsymbol{\mu}_i^T\boldsymbol{\mu}_i}{2} \\ &= \boldsymbol{\mu}_i^T\boldsymbol{x}, \end{aligned}$$

where additive terms independent of i have been ignored (they will cancel out later). This includes the terms

$$\boldsymbol{\mu}_1^T\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2^T\boldsymbol{\mu}_2 = d'^2$$

We can further simplify the discriminant functions as

$$\begin{aligned} g_1(\boldsymbol{x}) &= \boldsymbol{\mu}_1^T\boldsymbol{x} = (d', 0)\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = d'x_1 \\ g_2(\boldsymbol{x}) &= \boldsymbol{\mu}_2^T\boldsymbol{x} = (0, d')\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = d'x_2 \end{aligned}$$

and combine them into a single function

$$g(\boldsymbol{x}) = g_1(\boldsymbol{x}) - g_2(\boldsymbol{x}) = d'(x_1 - x_2)$$

In the Two-Interval Two-alternative Forced Choice method(2I2AFC), d' is considered to be positive, so we can divide the above discriminant function by the positive number d' to get the new discriminant function

$$g(\boldsymbol{x}) = x_1 - x_2$$

The decision function is then

$$d(\boldsymbol{x}) = \begin{cases} 1 & \text{if } g(\boldsymbol{x}) > 0 \\ 2 & \text{if } g(\boldsymbol{x}) \leq 0 \end{cases} = \begin{cases} 1 & \text{if } x_1 > x_2 \\ 2 & \text{otherwise} \end{cases}$$

3.9.b To find the probability of a correct response, we define a new random variable $Y = g(\boldsymbol{X}) = X_1 - X_2$. Since Y is a linear combination of two Gaussian random variables, it is also Gaussian and defined by its mean and variance. These are easy to calculate as

$$E[Y \mid S = 1] = E[X_1 - X_2] = E[X_1] - E[X_2] = d' - 0 = d'$$

$$E[Y \mid S = 2] = E[X_1 - X_2] = E[X_1] - E[X_2] = 0 - d' = -d'$$

$$\text{var}\,[Y \mid S = 1] = \text{var}\,[X_1 - X_2] = \text{var}\,[X_1] + \text{var}\,[X_2] = 1 + 1 = 2$$

(since X_1 and X_2 are independent) and

$$\text{var}\,[Y \mid S = 1] = \text{var}\,[Y \mid S = 2]$$

The probability of correct response is thus

$$\begin{aligned}
\text{P}(\textit{correct}) &= \text{P}(Y > 0 \mid S = 1)\text{P}_S(1) + \text{P}(Y \leq 0 \mid S = 2)\text{P}_S(2) \\
&= \frac{1}{2}\text{P}\left(\frac{Y - d'}{\sqrt{2}} > -\frac{d'}{\sqrt{2}}\right) + \frac{1}{2}\text{P}\left(\frac{Y + d'}{\sqrt{2}} \leq \frac{d'}{\sqrt{2}}\right) \\
&= \frac{1}{2}(1 - \Phi(-\frac{d'}{\sqrt{2}})) + \frac{1}{2}\Phi(\frac{d'}{\sqrt{2}}) \\
&= \frac{1}{2}\Phi(\frac{d'}{\sqrt{2}}) + \frac{1}{2}\Phi(\frac{d'}{\sqrt{2}}) \\
&= \Phi(\frac{d'}{\sqrt{2}})
\end{aligned}$$

Chapter 4

Practical Classification

Chapter 5

Hidden Markov Model

5.1.a The transition probability matrix A has a left-right structure, thus the word model is a left-right HMM. It is not ergodic since previous states of the Markov chain affect long-term behavior—if you for instance are in state 3 at time t, you can never return to state 1 or 2 at any future time $t+n$, regardless of how large $n > 0$ is.

5.1.b When applying the Forward Algorithm, many terms will be zero due to the left-right structure of the HMM, and the fact that certain observations are impossible for some states, leading to zeroes in B. For brevity, we will only show the computations for the nonzero elements.

For $t = 1$ we perform an initialization step following the equations

$$\alpha_{j,1}^{temp} = q_j b_j(z_1); \quad c_1 = \sum_{k=1}^{N} \alpha_{k,1}^{temp}; \quad \hat{\alpha}_{j,1} = \frac{\alpha_{j,1}^{temp}}{c_1}$$

This gives

$$\alpha_{1,1}^{temp} = q_1 b_1(1) = 1 \cdot 1 = 1; \quad \alpha_{2,1}^{temp} = \alpha_{3,1}^{temp} = 0$$

$$c_1 = \sum_{k=1}^{3} \alpha_{k,1}^{temp} = 1 + 0 + 0 = 1$$

so

$$\hat{\alpha}_{1,1} = \frac{1}{1} = 1; \quad \hat{\alpha}_{2,1} = \hat{\alpha}_{3,1} = 0$$

For $t = 2, 3, \ldots$ we perform forward steps according to the formulas

$$\alpha_{j,t}^{temp} = b_j(z_t) \sum_{i=1}^{N} \hat{\alpha}_{i,t-1} a_{ij}; \quad c_t = \sum_{k=1}^{N} \alpha_{k,t}^{temp}; \quad \hat{\alpha}_{j,t} = \frac{\alpha_{j,t}^{temp}}{c_t}$$

For $t = 2$ we get

$$\begin{aligned} \alpha_{2,2}^{temp} &= b_2(2) \sum_{i=1}^{3} \hat{\alpha}_{i,1} a_{i2} \\ &= 0.5(1 \cdot 0.7 + 0 \cdot 0.5 + 0 \cdot 0) \\ &= 0.35 \end{aligned}$$

while

$$\alpha_{1,2}^{temp} = 0; \quad \alpha_{3,2}^{temp} = 0$$

so

$$c_2 = 0.35; \quad \hat{\alpha}_{2,2} = 1; \quad \hat{\alpha}_{1,2} = \hat{\alpha}_{3,2} = 0$$

For $t = 3$

$$\alpha_{2,3}^{temp} = 0.1(0 \cdot 0.7 + 1 \cdot 0.5 + 0 \cdot 0) = 0.05$$

$$\alpha_{3,3}^{temp} = 0.6(0 \cdot 0 + 1 \cdot 0.5 + 0 \cdot 1) = 0.3$$

$$c_3 = 0.35; \quad \hat{\alpha}_{1,3} = 0; \quad \hat{\alpha}_{2,3} = \frac{1}{7}; \quad \hat{\alpha}_{3,3} = \frac{6}{7}$$

For $t = 4$

$$\alpha_{2,4}^{temp} = 0.1(0 + \frac{1}{7}0.5 + \frac{6}{7}0) = \frac{1}{140}$$

$$\alpha_{3,4}^{temp} = 0.6(0 + \frac{1}{7}0.5 + \frac{6}{7}1) = \frac{78}{140}$$

$$c_4 = \frac{79}{140}; \quad \hat{\alpha}_{1,4} = 0; \quad \hat{\alpha}_{2,4} = \frac{1}{79}; \quad \hat{\alpha}_{3,4} = \frac{78}{79}$$

Finally, for $t = 5$ we get

$$\alpha_{3,5}^{temp} = 0.1(0 + \frac{1}{79}0.5 + \frac{78}{79}1) \approx 0.0994$$

$$c_5 = 0.0994; \quad \hat{\alpha}_{1,5} = \hat{\alpha}_{2,5} = 0; \quad \hat{\alpha}_{3,5} = 1$$

The complete table becomes

t	1	2	3	4	5
$\hat{\alpha}_{1,t}$	1	0	0	0	0
$\hat{\alpha}_{2,t}$	0	1	1/7	1/79	0
$\hat{\alpha}_{3,t}$	0	0	6/7	78/79	1
c_t	1	0.35	0.35	79/140	0.0994

and the requested probability can be calculated as

$$\begin{aligned} \mathrm{P}(\underline{Z} = \underline{z}|\lambda) &= \mathrm{P}(z_1 z_2 z_3 z_4 z_5|\lambda) \\ &= \mathrm{P}(\{1, 2, 4, 4, 1\}|\lambda) \\ &= \mathrm{P}(1|\lambda)\mathrm{P}(2|1, \lambda)\mathrm{P}(4|2, 1, \lambda)\mathrm{P}(4|4, 2, 1, \lambda)\mathrm{P}(1|4, 4, 2, 1, \lambda) \\ &= c_1 c_2 c_3 c_4 c_5 \\ &= 1 \times 0.35 \times 0.35 \times 0.5641 \ldots \times 0.0994 \ldots \\ &\approx 0.0069 \end{aligned}$$

5.1.c This problem can be solved by counting all the possible state sequences. Since we have a three-state HMM, we always start in state 1 at $t = 1$, and can only stay in the current state i or jump to the next state $i + 1 \leq 3$, there can be at the most two transitions in any of the possible sequences.

There is only one sequence with no transitions; the Markov chain then stays in state 1 all the time. There is also only one sequence where the transition to state 2 occurs at $t = 5$; there is no time for additional transitions. There are two sequences where the transition to state 2 occurs at $t = 4$; the Markov chain can either stay in state 2 or move to state 3 at $t = 5$. Similarly, there are three sequences where the transition to state 2 occurs at $t = 3$, and four sequences where the transition to state 2 occurs at $t = 2$. Summing up, there are $1 + 1 + 2 + 3 + 4 = 11$ possible state sequences.

5.1.d An efficient way to determine the most likely hidden state sequence given Markov chain parameters $\lambda = \{q, A, B\}$ and a set of observations $\underline{z} = (1, 2, 4, 4, 1)$ is to use the Viterbi algorithm, which has many similarities to the Forward Algorithm. For brevity we shall only show computations for the nonzero terms that are calculated using this procedure.

For $t = 1$ we initalize the Viterbi partial-sequence vector elements as

$$\chi_{j,1} = q_j b_j(z_1)$$

giving

$$\chi_{1,1} = q_1 b_1(1) = 1 \cdot 1 = 1; \quad \chi_{2,1} = \chi_{3,1} = 0$$

For $t = 2, 3, \ldots$ we apply Viterbi Forward steps

$$\chi_{j,t} = b_j(z_t) \max_i \chi_{i,t-1} a_{ij}; \quad \zeta_{j,t} = \operatorname*{argmax}_i \chi_{i,t-1} a_{ij}$$

For $t = 2$, in particular, we get

$$\chi_{2,2} = 0.5 \max_i \{1 \cdot 0.7,\, 0,\, 0\} = 0.35; \quad \chi_{1,2} = \chi_{3,2} = 0$$

and (since $\chi_{1,1}$ is the only nonzero term from the previous time step)

$$\zeta_{1,2} = \zeta_{2,2} = 1$$

The calculation of $\zeta_{3,2}$ is ambiguous since all terms in the argmax_i are zero. However, this simply means that we can choose any of the legal i-values.

For $t = 3$ we get

$$\chi_{2,3} = 0.1 \max_i \{0,\, 0.35 \cdot 0.5,\, 0\} = 0.0175$$

$$\chi_{3,3} = 0.6 \max_i \{0,\, 0.35 \cdot 0.5,\, 0\} = 0.105$$

and (since $\chi_{2,2}$ is the only nonzero term from the previous time step)

$$\zeta_{2,3} = \zeta_{3,3} = 2$$

For $t = 4$ we get

$$\chi_{2,4} = 0.1 \max_i\{0,\ 0.0175 \cdot 0.5,\ 0.105 \cdot 0\} = 8.75 \cdot 10^{-4}$$

$$\chi_{3,4} = 0.6 \max_i\{0,\ 0.0175 \cdot 0.5,\ 0.105 \cdot 1\} = 0.6 \cdot 0.105 \cdot 1 = 0.063$$

and the corresponding indices for the maxima

$$\zeta_{2,4} = 2; \quad \zeta_{3,4} = 3$$

Finally, at $t = 5$ we get $\chi_{2,5} = 0$ since $b_2(1) = 0$, while

$$\chi_{3,5} = 0.1 \max_i\{0,\ 8.75 \cdot 10^{-4} \cdot 0.5,\ 0.063 \cdot 1\} = 0.0063$$

The backpointer variables are similar to before,

$$\zeta_{2,4} = 2; \quad \zeta_{3,4} = 3$$

Like with the Forward Algorithm, we can collect the results in a table

t	1	2	3	4	5
$\chi_{1,t}$	1	0	0	0	0
$\chi_{2,t}$	0	0.35	0.0175	$8.75 \cdot 10^{-4}$	0
$\chi_{3,t}$	0	0	0.105	0.063	0.0063
$\zeta_{1,t}$	**N/A**	1			
$\zeta_{2,t}$	N/A	**1**	2	2	2
$\zeta_{3,t}$	N/A		**2**	**3**	**3**

Since the $\chi_{3,5}$ is the greatest partial-sequence probability vector element at $t = 5$, we follow the path given by the backpointers using $s_t^\star = \zeta_{s_t^\star+1,t+1}$ starting from $\zeta_{3,5}$ (bolded in the table). We then recover the most likely hidden state sequence, $\underline{S}^\star = \{1, 2, 3, 3, 3\}$.

5.3.a

$$\mathrm{P}(S_{19} = j | x_1 \ldots x_{19}) = \frac{\mathrm{P}(S_{19} = j, X_{19} = x_{19} | x_1 \ldots x_{18})}{\mathrm{P}(X_{19} = x_{19} | x_1 \ldots x_{18})}$$

$$\mathrm{P}(X_{19} = x_{19} | x_1 \ldots x_{18}) = \sum_j \mathrm{P}(S_{19} = j, X_{19} = x_{19} | x_1 \ldots x_{18})$$

$$\begin{aligned}
\mathrm{P}(S_{19} = j, X_{19} = x_{19} | x_1 \ldots x_{18}) &= \mathrm{P}(X_{19} = x_{19} | S_{19} = j, x_1 \ldots x_{18}) \mathrm{P}(S_{19} = j | x_1 \ldots x_{18}) \\
&= \mathrm{P}(X_{19} = x_{19} | S_{19} = j) \mathrm{P}(S_{19} = j | x_1 \ldots x_{18}) \\
&= b_j(x_{19}) \mathrm{P}(S_{19} = j | x_1 \ldots x_{18}) \\
&= b_{j3} \mathrm{P}(S_{19} = j | x_1 \ldots x_{18})
\end{aligned}$$

$$\begin{aligned}\mathrm{P}(S_{19}=j|x_1\dots x_{18}) &= \sum_i \mathrm{P}(S_{19}=j|S_{18}=i,x_1\dots x_{18})\mathrm{P}(S_{18}=i|x_1\dots x_{18})\\ &= \sum_i \mathrm{P}(S_{19}=j|S_{18}=i)\mathrm{P}(S_{18}=i|x_1\dots x_{18})\\ &= \sum_i a_{ij}\hat{\alpha}_{i,18}\end{aligned}$$

Therefore

$$\mathrm{P}(S_{19}=j,X_{19}=x_{19}|x_1\dots x_{18}) = b_{j3}\left[\sum_{i=1}^{2} a_{ij}\hat{\alpha}_{i,18}\right]$$

for $j=1$

$$\begin{aligned}\mathrm{P}(S_{19}=1,X_{19}=x_{19}|x_1\dots x_{18}) &= 0.2\times(0.9\times 0.3+0.2\times 0.7)\\ &= 0.082\end{aligned}$$

for $j=2$

$$\begin{aligned}\mathrm{P}(S_{19}=2,X_{19}=x_{19}|x_1\dots x_{18}) &= 0.3\times(0.1\times 0.3+0.8\times 0.7)\\ &= 0.177\end{aligned}$$

Therefore

$$\mathrm{P}(S_{19}=j|x_1\dots x_{19}) = \begin{cases}\frac{0.082}{0.082+0.177}=0.3166 & \text{for } j=1\\ \frac{0.177}{0.082+0.177}=0.6834 & \text{for } j=2\end{cases}$$

5.3.b We use the previous result and continue as

$$P(S_{19}=j|x_1\dots x_{19}x_{20},\lambda) = \frac{P(S_{19}=j,x_{20}|x_1\dots x_{19},\lambda)}{P(x_{20}|x_1\dots x_{19},\lambda)}$$

$$\begin{aligned}P(S_{19}=j,x_{20}|x_1\dots x_{19},\lambda) &= \sum_k P(S_{19}=j,S_{20}=k,x_{20}|x_1\dots x_{19},\lambda) =\\ &= \sum_k a_{jk}b_{k4}P(S_{19}=j|x_1\dots x_{19},\lambda) =\\ &\approx \begin{cases}(0.9\cdot 0.1+0.1\cdot 0.4)\cdot 0.3166 = 0.041158, & j=1\\ (0.2\cdot 0.1+0.8\cdot 0.4)\cdot 0.6834 = 0.232356, & j=2\end{cases}\end{aligned}$$

$$P(S_{19}=1|x_1\dots x_{19}x_{20},\lambda) \approx \begin{cases}\frac{0.041158}{0.041158+0.232356}\approx 0.1505, & j=1\\ \frac{0.232356}{0.041158+0.232356}\approx 0.8495, & j=2\end{cases}$$

5.4 It is easiest to use vector notation. We denote the state probability vector at time t as $\boldsymbol{p}_t$, with elements $p_{i,t} = P(S_t = i)$. Then, with the definitions of the transition and backward transition matrices, we have

$$\begin{aligned}\boldsymbol{p}_t &= A^T\boldsymbol{p}_{t-1}\\ \boldsymbol{p}_{t-1} &= C\boldsymbol{p}_t\end{aligned}$$

Thus,

$$C = (A^T)^{-1}$$

5.5.a The initial state probability vector is given by

$$P = \begin{pmatrix} 0.2 \\ 0.2 \\ 0.2 \\ 0.4 \end{pmatrix}$$

The transition probability matrix is given by

$$A = \begin{pmatrix} 0.8 & 0.2 & 0 & 0 \\ 0.2 & 0.8 & 0 & 0 \\ 0 & 0 & 0.8 & 0.2 \\ 0 & 0 & 0.1 & 0.9 \end{pmatrix}$$

The given Markov chain is stationary since $P = A^T P$

5.5.b *Not ergodic*, because the Markov chain is *reducible*. It is not possible to go from states 1 or 2 to states 3 or 4, and vice versa.

5.5.c Define the initial state probability vector as:

$$P = \begin{pmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \end{pmatrix}$$

The state probabilities at time t+1 are determined by:

$$\begin{aligned} \mathrm{P}(S_{t+1} = j) &= \sum_{i=1}^{N} \mathrm{P}(S_{t+1} = j | S_t = i) \mathrm{P}(S_t = i) \\ &= \sum_{i=1}^{N} a_{ij} \mathrm{P}(S_t = i) \end{aligned}$$

$$\begin{aligned} \mathrm{P}(S_{t+1} = 1) &= \sum_{i=1}^{4} \mathrm{P}(S_{t+1} = 1 | S_t = i) \mathrm{P}(S_t = i) \\ \Rightarrow p_1 &= \sum_{i=1}^{4} a_{i1} \mathrm{P}(S_t = i) \\ \Rightarrow p_1 &= 0.8 \times p_1 + 0.2 \times p_2 + 0 + 0 \\ \Rightarrow p_1 &= p_2 \end{aligned}$$

$$\begin{aligned} \mathrm{P}(S_{t+1}=3) &= \sum_{i=1}^{4} \mathrm{P}(S_{t+1}=3|S_t=i)\mathrm{P}(S_t=i) \\ \Rightarrow p_3 &= \sum_{i=1}^{4} a_{i3}\mathrm{P}(S_t=i) \\ \Rightarrow p_3 &= 0+0+0.8\times p_3+0.1\times p_4 \\ \Rightarrow p_3 &= \frac{p_4}{2} \end{aligned}$$

we can chose $p_1 = a$ and $p_3 = b$, the initial state probabilities are then given by:

$$P = \begin{pmatrix} a \\ a \\ b \\ 2b \end{pmatrix}$$

with

$$2a + 3b = 1$$
$$a \geq 0 \text{ and } b \geq 0$$

5.6.c The HMM is stationary as there are eigenvectors with positive elements and eigenvalues =1. Eigenvectors with eigenvalues 1 are good candidates as stationary state distribution vectors, but we must normalize the eigenvectors such that their sum is 1. Any linear combination of the two eigenvectors is a stationary distribution, i.e., any vector of the following form:

$$\boldsymbol{p} = (1-b)\begin{pmatrix} 2/3 \\ 0 \\ 1/3 \end{pmatrix} + b\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 2(1-b)/3 \\ b \\ (1-b)/3 \end{pmatrix}$$

5.6.d The Markov chain is *reducible*, thus not ergodic.

5.7.a In order to be in a state for exactly d time instants, given the state has been entered, the Markov chain must jump back to the same state $d-1$ times, and then jump to some other state. The probability of this sequence of events is

$$f_{D_i}(d) = P(D_i = d) = a_{ii}^{d-1}(1-a_{ii}) \tag{5.1}$$

This is a geometric distribution.

5.7.b The expected duration is

$$\begin{aligned} E[D_i] &= \sum_{d=1}^{\infty} d \cdot f_{D_i}(d) = \\ &= (1-a_{ii})\sum_{d=1}^{\infty} d a_{ii}^{d-1} = \frac{1-a_{ii}}{a_{ii}}\sum_{d=1}^{\infty} d a_{ii}^{d} = \text{/Beta Sec. 8.6/} \\ &= \frac{1-a_{ii}}{a_{ii}}\frac{a_{ii}}{(1-a_{ii})^2} = \frac{1}{1-a_{ii}} \end{aligned} \tag{5.2}$$

5.8.a We define the probability-mass vector for the state distribution at time t as $\boldsymbol{p}_t$ with elements $p_{it} = P(D \geq t, S_t = i)$, for $1 \leq i \leq N$. At time $t = 1$, this probability vector is obviously the initial state probability vector, i.e., $\boldsymbol{p}_1 = q$. Let us denote the transposed transition matrix, without the last column, as the square matrix $\boldsymbol{C}$ with elements $c_{ij} = a_{ji}$ for $1 \leq i, j \leq N$. Then the state distribution at time t is, for any $i \in \{1, \ldots, N\}$,

$$\begin{aligned} p_{it} &= P\left[D \geq t \cap S_t = i\right] = \sum_{j=1}^{N} P\left[D \geq t \cap S_t = i \cap S_{t-1} = j\right] = \\ &= \sum_{j=1}^{N} P\left[D \geq t \cap S_t = i | S_{t-1} = j \cap D \geq t-1\right] P\left[S_{t-1} = j \cap D \geq t-1\right] = \\ &= \sum_{j=1}^{N} P\left[S_t \leq N \cap S_t = i | S_{t-1} = j \cap D \geq t-1\right] P\left[S_{t-1} = j \cap D \geq t-1\right] = \\ &= \sum_{j=1}^{N} P\left[S_t = i | S_{t-1} = j\right] P\left[S_{t-1} = j \cap D \geq t-1\right] = \\ &= \sum_{j=1}^{N} a_{ji} p_{jt} \end{aligned} \tag{5.3}$$

i.e.,

$$\boldsymbol{p}_t = \boldsymbol{C}\boldsymbol{p}_{t-1} = \boldsymbol{C}^{t-1} q \tag{5.4}$$

The probabilty that the HMM is still working at time t, i.e. it has not yet reached the END state, is then

$$P(D \geq t) = \sum_{i=1}^{N} p_{it} = \mathbf{1}\boldsymbol{p}_t = \mathbf{1}\boldsymbol{C}^{t-1} q \tag{5.5}$$

where $\mathbf{1}$ is a row vector where all N elements are 1. Thus, the duration probability-mass function is

$$f_D(d) = P\left[D = d\right] = \mathbf{1}\boldsymbol{C}^{d-1} q - \mathbf{1}\boldsymbol{C}^{d} q = \mathbf{1}(\boldsymbol{I} - \boldsymbol{C})\boldsymbol{C}^{d-1} q \tag{5.6}$$

where $\boldsymbol{I}$ is the identity matrix.

5.8.b The expected duration is

$$\begin{aligned} E\left[D\right] = \sum_{d=1}^{\infty} d f_D(d) &= \mathbf{1}\left(\sum_{d=1}^{\infty} d(\boldsymbol{I} - \boldsymbol{C})\boldsymbol{C}^{d-1}\right) q = \\ &= \mathbf{1}\left(\sum_{d=1}^{\infty} d\boldsymbol{C}^{d-1} - d\boldsymbol{C}^{d}\right) q = \\ &= \mathbf{1}\left(\sum_{d=0}^{\infty} (d+1)\boldsymbol{C}^{d} - d\boldsymbol{C}^{d}\right) q = \mathbf{1}\left(\sum_{d=0}^{\infty} \boldsymbol{C}^{d}\right) q \end{aligned} \tag{5.7}$$

By analogy with the similar scalar geometric series, we guess that the infinite matrix power sum is

$$\sum_{d=0}^{\infty} \boldsymbol{C}^d = (\boldsymbol{I} - \boldsymbol{C})^{-1} \tag{5.8}$$

provided that the sum converges. To prove this conjecture, we consider the matrix product

$$(\boldsymbol{I} - \boldsymbol{C}) \sum_{d=0}^{\infty} \boldsymbol{C}^d = \sum_{d=0}^{\infty} \boldsymbol{C}^d - \sum_{d=0}^{\infty} \boldsymbol{C}^{d+1} = \sum_{d=0}^{\infty} \boldsymbol{C}^d - \sum_{d=1}^{\infty} \boldsymbol{C}^d = \boldsymbol{C}^0 = \boldsymbol{I} \tag{5.9}$$

We thus obtain the final result as

$$E\left[D\right] = \mathbf{1} \left(\sum_{d=0}^{\infty} \boldsymbol{C}^d \right) q = \mathbf{1} \left(\boldsymbol{I} - \boldsymbol{C} \right)^{-1} q \tag{5.10}$$

5.9.a As all word sequences are equally probable, we can use the *Maximum-Likelihood* decision rule. The pronunciations of different words in the sequence are conditionally independent, given a particular speaker. The likelihood for any word sequence $\underline{w} = (w_1, \ldots, w_J)$ is therefore

$$\begin{aligned} Q(\underline{w}) = P(\underline{\boldsymbol{x}}_1, \ldots, \underline{\boldsymbol{x}}_J | w_1, \ldots, w_J) = \sum_{m=1}^{M} P(\underline{\boldsymbol{x}}_1, \ldots, \underline{\boldsymbol{x}}_J \cap S = m | w_1, \ldots, w_J) = \\ = \sum_{m=1}^{M} P(\underline{\boldsymbol{x}}_1, \ldots, \underline{\boldsymbol{x}}_J | (w_1, \ldots, w_J) \cap S = m) P(S = m) = \\ = \sum_{m=1}^{M} \frac{1}{M} \prod_{j=1}^{J} e^{L_{j,w_j,m}} = \frac{1}{M} \sum_{m=1}^{M} e^{\sum_{j=1}^{J} L_{j,w_j,m}} \end{aligned} \tag{5.11}$$

This likelihood must be evaluated for each of all the possible word sequences, and the highest value is selected. Thus, we choose the most probable word sequence as

$$\underline{\hat{w}} = \underset{w_1,\ldots,w_J}{\operatorname{argmax}} \sum_{m=1}^{M} e^{\sum_{j=1}^{J} L_{j,w_j,m}} \tag{5.12}$$

It may be possible to find an algorithm, similar to the Viterbi algorithm, to solve the maximization search across word sequences recursively.

5.9.b Here, for minimal error probability at each word position j in the sequence, we select the word with the greatest conditional probability, i.e.

$$\hat{w}_j = \underset{n}{\operatorname{argmax}}\, P(W_j = n | \underline{\boldsymbol{x}}_1, \ldots, \underline{\boldsymbol{x}}_J) \tag{5.13}$$

We calculate the conditional word probability at position j, given the complete observed data sequence, as

$$P(W_j = n|\underline{\boldsymbol{x}}_1, \ldots, \underline{\boldsymbol{x}}_J) = \sum_{m=1}^{M} P(S = m \cap W_j = n|\underline{\boldsymbol{x}}_1, \ldots, \underline{\boldsymbol{x}}_J) =$$
$$= \sum_{m=1}^{M} P(W_j = n|S = m \cap \underline{\boldsymbol{x}}_1, \ldots, \underline{\boldsymbol{x}}_J) P(S = m|\underline{\boldsymbol{x}}_1, \ldots, \underline{\boldsymbol{x}}_J) \quad (5.14)$$

The maximal word probability is clearly also equal to the probability of a correct decision. We facilitate the computation by first calculating, once and for all, the conditional speaker probabilities

$$p_m = P(S = m|\underline{\boldsymbol{x}}_1, \ldots, \underline{\boldsymbol{x}}_J) = \frac{P(\underline{\boldsymbol{x}}_1, \ldots, \underline{\boldsymbol{x}}_J|S = m)}{\sum_k P(\underline{\boldsymbol{x}}_1, \ldots, \underline{\boldsymbol{x}}_J|S = k)} \quad (5.15)$$

where

$$P(\underline{\boldsymbol{x}}_1, \ldots, \underline{\boldsymbol{x}}_J|S = m) =$$
$$= \prod_{j=1}^{J} P(\underline{\boldsymbol{x}}_j|S = m, \underline{\boldsymbol{x}}_1, \ldots, \underline{\boldsymbol{x}}_{j-1}) = \quad (5.16)$$
$$= \prod_{j=1}^{J} \sum_{n=1}^{N} P(\underline{\boldsymbol{x}}_j \cap W_j = n|S = m, \underline{\boldsymbol{x}}_1, \ldots, \underline{\boldsymbol{x}}_{j-1}) = \quad (5.17)$$
$$= \prod_{j=1}^{J} \sum_{n=1}^{N} P(W_j = n|S = m, \underline{\boldsymbol{x}}_1, \ldots, \underline{\boldsymbol{x}}_{j-1}) \cdot$$
$$\cdot P(\underline{\boldsymbol{x}}_j|W_j = n, S = m, \underline{\boldsymbol{x}}_1, \ldots, \underline{\boldsymbol{x}}_{j-1}) = \quad (5.18)$$
$$= \prod_{j=1}^{J} \sum_{n=1}^{N} P(W_j = n|S = m) P(\underline{\boldsymbol{x}}_j|W_j = n, S = m) = \quad (5.19)$$
$$= \prod_{j=1}^{J} \sum_{n=1}^{N} \frac{1}{N} e^{L_{jnm}} = \quad (5.20)$$
$$= \frac{1}{N^J} \prod_{j=1}^{J} \sum_{n=1}^{N} e^{L_{jnm}} \quad (5.21)$$

The same speaker probabilities p_m can then be used for all words in the sequence. The remaining factor is

$$P(W_j = n|S = m \cap \underline{\boldsymbol{x}}_1, \ldots, \underline{\boldsymbol{x}}_J) = \frac{P(\underline{\boldsymbol{x}}_j \cap W_j = n|S = m)}{\sum_l P(\underline{\boldsymbol{x}}_j \cap W_j = l|S = m)} =$$
$$= \frac{P(W_j = n|S = m) P(\underline{\boldsymbol{x}}_j|W_j = n, S = m)}{\sum_l P(W_j = l|S = m) P(\underline{\boldsymbol{x}}_j|W_j = l, S = m)} = \frac{e^{L_{jnm}}}{\sum_l e^{L_{jlm}}} \quad (5.22)$$

Thus, the final decision rule is

$$\hat{w}_j = \underset{n}{\operatorname{argmax}} \sum_{m=1}^{M} \frac{e^{L_{jnm}}}{\sum_l e^{L_{jlm}}} p_m \tag{5.23}$$

where

$$p_m = P(S = m|\underline{\boldsymbol{x}}_1, \ldots, \underline{\boldsymbol{x}}_J) = \frac{\prod_{j=1}^{J} \left(\sum_{n=1}^{N} e^{L_{jnm}} \right)}{\sum_k \prod_{j=1}^{J} \left(\sum_{n=1}^{N} e^{L_{jnk}} \right)} \tag{5.24}$$

Chapter 6

HMM Training

Chapter 7

Expectation Maximization

7.1.a We have

$$y(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{A}\boldsymbol{x} = \sum_i \sum_j x_i a_{ij} x_j \tag{7.1}$$

$$\frac{\partial y}{\partial x_k} = \sum_i x_i a_{ik} + \sum_j a_{kj} x_j = \sum_j (a_{jk} + a_{kj}) x_j \tag{7.2}$$

Including all possible values of k, this can be written as

$$\frac{\partial y}{\partial \boldsymbol{x}} = (\boldsymbol{A}^T + \boldsymbol{A})\boldsymbol{x} \tag{7.3}$$

7.1.b We have

$$y(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{A}\boldsymbol{x} = \sum_i \sum_j x_i a_{ij} x_j \tag{7.4}$$

$$\frac{\partial y}{\partial a_{ij}} = x_i x_j \tag{7.5}$$

Including all possible values of i and j, this can be written as

$$\frac{\partial y}{\partial \boldsymbol{A}} = \boldsymbol{x}\boldsymbol{x}^T \tag{7.6}$$

7.1.c The determinant of any square matrix $\boldsymbol{A}$ with elements a_{ij} can be developed as a sum along any row i or any column j, as

$$y(\boldsymbol{A}) = \det \boldsymbol{A} = \sum_k a_{ik} C_{ik} \tag{7.7}$$

or

$$y(\boldsymbol{A}) = \det \boldsymbol{A} = \sum_k a_{kj} C_{kj} \tag{7.8}$$

where $C_{ij} = (-1)^{i+j} D_{ij}$ is a *cofactor* of $\boldsymbol{A}$, where D_{ij} is the determinant of the sub-matrix obtained by deleting row i and column j of $\boldsymbol{A}$. Thus,

$$\frac{\partial y}{\partial a_{ij}} = C_{ij} \tag{7.9}$$

Including all possible values of i and j, this can be written as a matrix

$$\frac{\partial y}{\partial \boldsymbol{A}} = \boldsymbol{C} = \begin{pmatrix} C_{11} & C_{12} & \cdots & \cdots \\ C_{21} & \cdots & \cdots & \cdots \\ \vdots & \cdots & C_{ij} & \cdots \\ \vdots & \cdots & \cdots & \cdots \end{pmatrix} \tag{7.10}$$

As the inverse matrix can also be expressed in terms of cofactors, as

$$\boldsymbol{A}^{-1} = \frac{1}{\det \boldsymbol{A}} \boldsymbol{C}^T \tag{7.11}$$

the result can be written as

$$\frac{\partial y}{\partial \boldsymbol{A}} = (\det \boldsymbol{A})(\boldsymbol{A}^{-1})^T \tag{7.12}$$

7.1.d Using the same cofactor expansion as in the previous example, the partial derivative with respect to a particular element a_{ij} of $\boldsymbol{A}$ can be written as

$$\frac{\partial \ln \det \boldsymbol{A}}{\partial a_{ij}} = \frac{1}{\det \boldsymbol{A}} \frac{\partial \det \boldsymbol{A}}{\partial a_{ij}} = \frac{1}{\det \boldsymbol{A}} C_{ij} \tag{7.13}$$

Observing that element i, j of the inverse matrix $\boldsymbol{A}^{-1}$ can also be written in terms of cofactors, as

$$\left[\boldsymbol{A}^{-1}\right]_{ij} = \frac{1}{\det \boldsymbol{A}} C_{ji} \tag{7.14}$$

we see that

$$\frac{\partial \ln \det \boldsymbol{A}}{\partial a_{ij}} = \left[\boldsymbol{A}^{-1}\right]_{ji} \tag{7.15}$$

Including all values of i and j, this can be written as a matrix

$$\frac{\partial \ln \det \boldsymbol{A}}{\partial \boldsymbol{A}} = (\boldsymbol{A}^{-1})^T \tag{7.16}$$

7.1.e As $\det \boldsymbol{A}^{-1} = 1/\det \boldsymbol{A}$, we can simply exchange $\boldsymbol{A}$ with $\boldsymbol{A}^{-1}$ in the previous problem, to obtain

$$\frac{\partial \ln \det \boldsymbol{A}}{\partial \boldsymbol{A}^{-1}} = -\frac{\partial \ln \det \boldsymbol{A}^{-1}}{\partial \boldsymbol{A}^{-1}} = -\boldsymbol{A}^T \tag{7.17}$$

7.2.a There are only two possible sequences of coins used, either $\underline{S} = (1212\ldots)$ or $\underline{S} = (2121\ldots)$, with equal probability. In the first sequence the asymmetrical coin was used in all the even-numbered trials, and in the other sequence it was used in all the odd-numbered trials.

In the observed sequence, we see $x_t = 1$ at K_o times out of N_o odd-numbered time positions, and also at K_e times out of N_e even-numbered time positions. Thus,

$$\mathrm{P}(\underline{x}|\underline{S} = (2121\ldots), \theta) = \theta^{K_o}(1-\theta)^{N_o-K_o}0.5^{N_e}$$

$$\mathrm{P}(\underline{x}|\underline{S} = (1212\ldots), \theta) = \theta^{K_e}(1-\theta)^{N_e-K_e}0.5^{N_o}$$

Using Bayes Rule,

$$\mathrm{P}(\underline{S}, \underline{x}|\theta) = \mathrm{P}(\underline{x}|\underline{S}, \theta)\mathrm{P}(\underline{S})$$

As $P(\underline{S}) = 0.5$ for both state sequences,

$$\begin{aligned}\mathrm{P}(\underline{S} = (2121\ldots), \underline{x}|\theta) &= \theta^{K_o}(1-\theta)^{N_o-K_o}0.5^{N_e} \times 0.5 \\ &= \theta^{K_o}(1-\theta)^{N_o-K_o}0.5^{N_e+1}\end{aligned}$$

$$\begin{aligned}\mathrm{P}(\underline{S} = (1212\ldots), \underline{x}|\theta) &= \theta^{K_e}(1-\theta)^{N_e-K_e}0.5^{N_o} \times 0.5 \\ &= \theta^{K_e}(1-\theta)^{N_e-K_e}0.5^{N_o+1}\end{aligned}$$

Let

$$A(\theta) = \theta^{K_o}(1-\theta)^{N_o-K_o}0.5^{N_e+1}$$

and

$$B(\theta) = \theta^{K_e}(1-\theta)^{N_e-K_e}0.5^{N_o+1}$$

Apply the EM algorithm:

$$\begin{aligned}Q(\theta', \theta) &= E_{\underline{S}}\left[\ln \mathrm{P}(\underline{S}, \underline{x}|\theta')|\underline{x}, \theta\right] \\ &= \sum_{\underline{S}} \mathrm{P}(\underline{S}|\underline{x}, \theta) \ln \mathrm{P}(\underline{S}, \underline{x}|\theta') \\ &= \frac{A(\theta)}{A(\theta)+B(\theta)} \ln \theta'^{K_o}(1-\theta')^{N_o-K_o}0.5^{N_e+1} + \\ &\quad + \frac{B(\theta)}{A(\theta)+B(\theta)} \ln \theta'^{K_e}(1-\theta')^{N_e-K_e}0.5^{N_o+1}\end{aligned}$$

At the maximum point, the partial derivative of the function $Q(\theta', \theta)$ with respect to θ' (i.e., $\frac{\partial q(\theta', \theta)}{d\theta'}$) must be zero. We get the solution

$$\begin{aligned}\theta' &= \frac{A(\theta)K_o + B(\theta)K_e}{A(\theta)N_o + B(\theta)N_e} \\ &= \frac{\theta^{K_o}(1-\theta)^{N_o-K_o}0.5^{N_e+1}K_o + \theta^{K_e}(1-\theta)^{N_e-K_e}0.5^{N_o+1}K_e}{\theta^{K_o}(1-\theta)^{N_o-K_o}0.5^{N_e+1}N_o + \theta^{K_e}(1-\theta)^{N_e-K_e}0.5^{N_o+1}N_e}\end{aligned}$$

7.4

$$Q(\lambda', \lambda) = \sum_{i_1} \dots \sum_{i_N} \mathrm{P}[\underline{S} = (i_1, \dots, i_N)|\boldsymbol{x}, \lambda] \ln \mathrm{P}[\underline{S} = (i_1, \dots, i_N), \boldsymbol{x}|\lambda']$$

The activity of each cell is independent of each other:

$$\begin{aligned}
Q(\lambda', \lambda) &= \sum_{i_1} \dots \sum_{i_N} \mathrm{P}[\underline{S} = (i_1, \dots, i_N)|\boldsymbol{x}, \lambda] \ln \prod_j \mathrm{P}[S_j = i_j, x_j|\lambda'] = \\
&= \sum_{i_1} \dots \sum_{i_N} \mathrm{P}[\underline{S} = (i_1, \dots, i_N)|\boldsymbol{x}, \lambda] \sum_j \ln \mathrm{P}[S_j = i_j, x_j|\lambda'] = \\
&= \sum_j \sum_{i_1} \dots \sum_{i_N} \mathrm{P}[\underline{S} = (i_1, \dots, i_N)|\boldsymbol{x}, \lambda] \ln \mathrm{P}[S_j = i_j, x_j|\lambda'] = \\
&= \sum_j \sum_i \sum_{i_1} \dots \sum_{i_N} \mathrm{P}[\underline{S} = (i_1, \dots, i_N), S_j = i|\boldsymbol{x}, \lambda] \ln \mathrm{P}[S_j = i, x_j|\lambda'] = \\
&= \sum_j \sum_i \mathrm{P}[S_j = i|\boldsymbol{x}, \lambda] \ln \mathrm{P}[S_j = i, x_j|\lambda'] \cdot \\
&\qquad \cdot \underbrace{\sum_{i_1} \dots \sum_{i_N} \mathrm{P}[\underline{S} = (i_1, \dots, i_N)|S_j = i, \boldsymbol{x}, \boldsymbol{\lambda}]}_{=1} = \\
&= \sum_j \sum_i \mathrm{P}[S_j = i|x_j, \lambda] \ln \mathrm{P}[S_j = i, x_j|\lambda']
\end{aligned}$$

$$\begin{aligned}
Q(\lambda', \lambda) = \sum_j \mathrm{P}[S_j = 1|x_j, \lambda](-c_1' D + x_j \ln(c_1' D) - x_j!) + \\
+ \sum_j \mathrm{P}[S_j = 2|x_j, \lambda](-c_2' D + x_j \ln(c_2' D) - x_j!)
\end{aligned}$$

Remove terms independent of λ':

$$Q(\lambda', \lambda) = \sum_j \mathrm{P}[S_j = 1|x_j, \lambda](-c_1' D + x_j \ln(c_1' D)) + \mathrm{P}[S_j = 2|x_j, \lambda](-c_2' D + x_j \ln(c_2' D))$$

$$\frac{\partial Q(\lambda', \lambda)}{\partial c_1'} = \sum_j -D\mathrm{P}[S_j = 1|x_j, \lambda] + \frac{\mathrm{P}[S_j = 1|x_j, \lambda] x_j}{c_1'} = 0$$

$$\begin{aligned}
c_1' &= \frac{\sum_j \mathrm{P}[S_j = 1|x_j, \lambda] x_j}{D \sum_j \mathrm{P}[S_j = 1|x_j, \lambda]} \\
c_2' &= \frac{\sum_j \mathrm{P}[S_j = 2|x_j, \lambda] x_j}{D \sum_j \mathrm{P}[S_j = 2|x_j, \lambda]}
\end{aligned}$$

The probability can be expanded with Bayes' rule:

$$\mathrm{P}[S_j = k|x_j, \lambda] = \frac{\mathrm{P}[x_j|S_j = k, \lambda]\mathrm{P}[S_j = k]}{\sum_i \mathrm{P}[x_j|S_j = i, \lambda]\mathrm{P}[S_j = i]}$$

7.5 We regard the internal discrete variable U_t as a HMM state by mapping

$$U_t = +1 \Leftrightarrow S_t = 1$$
$$U_t = -1 \Leftrightarrow S_t = 2$$

Given any value c' for the unknown constant, we can then define conditional probability density functions for the output samples X_t, as

$$b_1(x) = f_{X_t|S_t}(x|1) = \frac{1}{\sqrt{2\pi}\sigma_W} e^{-(x-c')^2/2\sigma_W^2};$$
$$b_2(x) = f_{X_t|S_t}(x|2) = \frac{1}{\sqrt{2\pi}\sigma_W} e^{-(x+c')^2/2\sigma_W^2};$$

For any particular state sequence $\underline{S} = (i_1 \dots i_T)$ we can express

$$\ln P[\underline{S} = (i_1 \dots i_T) \cap (x_1 \dots x_T)|\lambda, c'] =$$
$$= \ln q_{i_1} b_{i_1}(x_1) a_{i_1 i_2} b_{i_2}(x_2) \dots b_{i_T}(x_T) = \sum_{t=1}^{T} \ln b_{i_t}(x_t) + D$$

Here D includes all the terms that do not depend on the parameter c' and can therefore be ignored in the following. Using the already available γ values, we redefine $Q(c', c)$ as

$$\begin{aligned}
Q(c', c) &= \sum_{i_1=1}^{2} \sum_{i_2=1}^{2} \dots \sum_{i_T=1}^{2} P[\underline{S} = (i_1 \dots i_T)|\underline{x}, \lambda, c] \sum_{t=1}^{T} \ln b_{i_t}(x_t) \\
&= \sum_{t=1}^{T} \sum_{i_t=1}^{2} P[S_t = i_t|\underline{x}, \lambda, c] \ln b_{i_t}(x_t) \cdot \\
&\quad \underbrace{\sum_{i_1} \dots \sum_{i_{t-1}} \sum_{i_{t+1}} \dots \sum_{i_T} P[(i_1 \dots i_{t-1} i_{t+1} \dots i_T)|S_t = i_t, \underline{x}, \lambda, c]}_{=1} \\
&= \sum_{t=1}^{T} \sum_{i_t=1}^{2} \gamma_{i_t,t} \ln b_{i_t}(x_t) \\
&= \sum_{t=1}^{T} -\gamma_{1,t}(x_t - c')^2/2\sigma_W^2 - \gamma_{2,t}(x_t + c')^2/2\sigma_W^2 - \ln \sqrt{2\pi}\sigma_W \\
\frac{dQ(c', c)}{dc'} &= \sum_{t=1}^{T} -\gamma_{1,t}(x_t - c')(-1)/\sigma_W^2 - \gamma_{2,t}(x_t + c')/\sigma_W^2 \\
&= \frac{1}{\sigma_W^2} \sum_{t=1}^{T} x_t(\gamma_{1,t} - \gamma_{2,t}) - c' \underbrace{(\gamma_{1,t} + \gamma_{2,t})}_{=1}
\end{aligned}$$

The derivative is zero for

$$c' = \frac{1}{T} \sum_{t=1}^{T} x_t(\gamma_{1,t} - \gamma_{2,t})$$

which is the desired update equation.

7.7

$$
\begin{aligned}
\gamma_{im,t} &= P\left[S_t = i \cap U_t = m | \underline{\boldsymbol{x}}, \lambda\right] = /\text{Bayes' rule}/ \\
&= P\left[U_t = m | S_t = i, \underline{\boldsymbol{x}}, \lambda\right] \underbrace{P\left[S_t = i | \underline{\boldsymbol{x}}, \lambda\right]}_{\gamma_{i,t}} = /\text{cond. indep., given state}/ \\
&= \gamma_{i,t} P\left[U_t = | S_t = i, \boldsymbol{x}_t, \lambda\right] = /\text{Bayes}/ \\
&= \gamma_{i,t} \frac{P\left[U_t = m \cap \boldsymbol{x}_t | S_t = i, \lambda\right]}{\sum_k P\left[U_t = k \cap \boldsymbol{x}_t | S_t = i, \lambda\right]} = /\text{Bayes}/ \\
&= \gamma_{i,t} \frac{P\left[U_t = m | S_t = i, \lambda\right] P\left[\boldsymbol{x}_t | U_t = m \cap S_t = i, \lambda\right]}{\sum_k P\left[U_t = k | S_t = i, \lambda\right] P\left[\boldsymbol{x}_t | U_t = k \cap S_t = i, \lambda\right]} = \\
&= \gamma_{i,t} \frac{w_{im} g(\boldsymbol{x}_t, \mu_{im}, C_{im})}{\sum_k w_{ik} g(\boldsymbol{x}_t, \mu_{ik}, C_{ik})}
\end{aligned}
\tag{7.18}
$$

Chapter 8

Bayesian Learning

8.1.a Assuming each result in observed sequence is independent, the probability mass for the complete sequence is

$$P\left[\underline{X} = \underline{x}\right] = \prod_{t=1}^{T} w^{x_t}(1-w)^{1-x_t} \tag{8.1}$$

The log-likelihood function is then

$$L(w) = \ln P\left[\underline{X} = \underline{x}\right] = \sum_{t=1}^{T} x_t \ln w + (1-x_t)\ln(1-w) \tag{8.2}$$

To find the ML estimate as $\hat{w}_{ML} = \operatorname{argmax}_w L(w)$, we solve the equation

$$0 = \frac{dL(w)}{dw} = \sum_t \frac{x_t}{w} - \frac{1-x_t}{1-w} \tag{8.3}$$

Denoting the count of correct responses as $z = \sum_{t=1}^{T} x_t$, the equation can be written as

$$\frac{z}{w} - \frac{T-z}{1-w} = \frac{z - Tw}{w(1-w)} = 0 \tag{8.4}$$

with solution $w = z/T$. Thus, the ML estimate is

$$\hat{w}_{ML} = \frac{z}{T} = \frac{1}{T}\sum_{t=1}^{T} x_t \tag{8.5}$$

8.1.b The probability mass function for a binary outcome X is

$$f_{X|W}(x \mid w) = w^x(1-w)^{1-x} \tag{8.6}$$

where the only possible outcome values are $x = 0$ or $x = 1$. The Jeffreys prior is derived directly from its definition in a few steps, as

$$\ln f_{X|W}(x \mid w) = x \ln w + (1-x)\ln(1-w) + \text{const.} \tag{8.7}$$

$$\frac{\partial \ln f_{X|W}(x \mid w)}{\partial w} = \frac{x}{w} - \frac{1-x}{1-w} \tag{8.8}$$

$$\left(\frac{\partial \ln f_{X|W}(x \mid w)}{\partial w}\right)^2 = \frac{x^2}{w^2} + \frac{(1-x)^2}{(1-w)^2} - \frac{2x(1-x)}{w(1-w)} \tag{8.9}$$

This expression is valid for both the possible outcomes of X. We now regard the expression as a function of the random variable X instead of the outcome x, as

$$Y = \left(\frac{\partial \ln f_{X|W}(X \mid w)}{\partial w}\right)^2 = \frac{X^2}{w^2} + \frac{(1-X)^2}{(1-w)^2} - \frac{2X(1-X)}{w(1-w)} \tag{8.10}$$

and calculate the expected value of this transformed random variable

$$I(w) = E[Y] = E\left[\left(\frac{\partial \ln f_{X|W}(X \mid w)}{\partial w}\right)^2\right] = \\ = \frac{E[X^2]}{w^2} + \frac{E[(1-X)^2]}{(1-w)^2} - \frac{E[2X(1-X)]}{w(1-w)} \tag{8.11}$$

For the binary random variable X, we know that $E[X] = w$ and $\text{var}[X] = E[X^2] - E[X]^2 = w(1-w)$. Thus, we have $E[X^2] = w^2 + w(1-w)$ and $E(1-X)^2 = (1-w)^2 + w(1-w)$. Substituting these known values, we obtain the Fisher Information

$$I(w) = E\left[\left(\frac{\partial \ln f_{X|W}(X \mid w)}{\partial w}\right)^2\right] = \\ = \frac{w^2 + w(1-w)}{w^2} + \frac{E[(1-w)^2 + w(1-w)]}{(1-w)^2} - \frac{2w - w^2 - w(1-w)}{w(1-w)} = \\ = \frac{1}{w} + \frac{1}{1-w} = \frac{1}{w(1-w)} \tag{8.12}$$

Thus, the Jeffreys prior is

$$f_W(w) \propto \sqrt{I(w)} = \frac{1}{\sqrt{w(1-w)}} = \frac{1}{w^{0.5}(1-w)^{0.5}} \tag{8.13}$$

This is obviously a Beta density with parameters $a = b = 0.5$.

8.2.a The prior parameter density is uniform, i.e. we select hyperparameters $a = b = 1$ in both models. Then the prior parameter density is constant, $f(w \mid 1, 1) = 1$.

As both models H_0, H_1 are equally probable, we can use the ML criterion to choose the better model. We denote the sequence of binary observations $\underline{x}_1, \underline{x}_2$ with $x_{t,i} = 1$ indicating a correct response, and $x_{t,i} = 0$ otherwise.

The observed data can be summarized as $z_1 = \sum_{t=1}^{T} x_{t,1}$ and similarly for z_2.

The probability of the observed binary observation sequences given the models, are

$$\begin{aligned} P\left[\underline{x}_1, \underline{x}_2 \mid H_0\right] &= \int_0^1 P\left[\underline{x}_1 \cap W = w \mid H_0\right] P\left[\underline{x}_2 \cap W = w \mid H_0\right] dw = \\ &= \int_0^1 w^{z_1}(1-w)^{T-z_1} w^{z_2}(1-w)^{T-z_2} dw = \\ &= \frac{\Gamma(z_1+z_2+1)\Gamma(2T-z_1-z_2+1)}{\Gamma(2T+2)} = \\ &= \frac{(z_1+z_2)!(2T-z_1-z_2)!}{(2T+1)!}; \end{aligned} \tag{8.14}$$

$$\begin{aligned} P\left[\underline{x}_1, \underline{x}_2 \mid H_1\right] &= \int_0^1 P\left[\underline{x}_1 \cap W_1 = w_1 \mid H_1\right] dw_1 \int_0^1 P\left[\underline{x}_2 \cap W_2 = w_2 \mid H_1\right] dw_2 = \\ &= \int_0^1 w_1^{z_1}(1-w_1)^{T-z_1} dw_1 \int_0^1 w_2^{z_2}(1-w_2)^{T-z_2} dw_2 = \\ &= \frac{\Gamma(z_1+1)\Gamma(T-z_1+1)}{\Gamma(T+2)} \frac{\Gamma(z_2+1)\Gamma(T-z_2+1)}{\Gamma(T+2)} = \\ &= \frac{z_1!(T-z_1)!}{(T+1)!} \frac{z_2!(T-z_2)!}{(T+1)!}; \end{aligned} \tag{8.15}$$

The conditional probability of the models, given observations, are

$$P\left[H_0 \mid \underline{x}_1, \underline{x}_2\right] = \frac{P\left[\underline{x}_1, \underline{x}_2 \mid H_0\right]}{P\left[\underline{x}_1, \underline{x}_2 \mid H_0\right] + P\left[\underline{x}_1, \underline{x}_2 \mid H_1\right]} \tag{8.16}$$

$$P\left[H_1 \mid \underline{x}_1, \underline{x}_2\right] = 1 - P\left[H_0 \mid \underline{x}_1, \underline{x}_2\right] \tag{8.17}$$

8.2.b We just use the results of the previous sub-problem and insert for the given numerical values to find the answer (see also MatLab `BayesModelComp2`):

$$\begin{aligned} P\left[H_0 \mid z_1 = 4, z_2 = 6\right] &= 0.579 \\ P\left[H_1 \mid z_1 = 4, z_2 = 6\right] &= 0.421 \end{aligned}$$

Note interesting discussion given in the Answers.

8.3.a The prior density function for A is (disregarding unimportant constants)

$$f_A(a) \propto e^{-\frac{(a-\mu_0)^2}{2\sigma_0^2}}$$

The conditional density for the complete observed sequence, given any particular outcome of A is (disregarding constants again)

$$f_{\underline{X}|A}(x_1,\ldots,x_t,\ldots,x_T \mid a) \propto \prod_{t=1}^{T} e^{-\frac{(x_t-ax_{t-1})^2}{2c^2}} = \\ = e^{-\frac{1}{2c^2}(\sum_t x_t^2 - 2ax_tx_{t-1} + a^2x_{t-1}^2)}$$

The probability density of the parameter, given the observations, is then

$$f_{A|\underline{X}}(a \mid \underline{x}) \propto f_{\underline{X},A}(x_1,\ldots,x_t,\ldots,x_T,a) = f_{\underline{X}|A}(x_1,\ldots,x_t,\ldots,x_T \mid a) f_A(a) \propto \\ \propto e^{-\frac{1}{2c^2}\left(a^2\left(c^2/\sigma_0^2+\sum_t x_{t-1}^2\right)-2a\left(\mu_0c^2/\sigma_0^2+\sum_t x_tx_{t-1}\right)+\cdots\right)}$$

where the $\cdots$ in the exponent represents the remaining expression that is independent of a.

As the exponent is a quadratic expression in a, it is clear that the posterior density for A must be Gaussian. We simply denote its mean and variance after T observations by μ_T and σ_T^2, and express the posterior density as

$$f_{A|\underline{X}}(a \mid \underline{x}) \propto e^{-\frac{1}{2\sigma_T^2}\left(a^2-2\mu_Ta+\cdots\right)}$$

and then just identify

$$\frac{1}{\sigma_T^2} = \frac{c^2/\sigma_0^2 + \sum_t x_{t-1}^2}{c^2} \tag{8.18}$$

$$\frac{\mu_T}{\sigma_T^2} = \frac{\mu_0c^2/\sigma_0^2 + \sum_t x_tx_{t-1}}{c^2} \tag{8.19}$$

$$\tag{8.20}$$

which yields, finally,

$$\mu_T = \frac{\mu_0c^2/\sigma_0^2 + \sum_t x_tx_{t-1}}{c^2/\sigma_0^2 + \sum_t x_{t-1}^2} \tag{8.21}$$

$$\sigma_T^2 = \frac{c^2}{c^2/\sigma_0^2 + \sum_t x_{t-1}^2} \tag{8.22}$$

To account for the fact that we had no prior knowledge about A we can just let $\sigma_0 \to \infty$ in these expressions.

8.3.b We start from the previous expressions (8.18) and (8.19) to obtain

$$\frac{1}{\sigma_T^2} = \frac{c^2/\sigma_0^2 + \sum_{t=1}^{T} x_{t-1}^2}{c^2} = \\ = \frac{c^2/\sigma_0^2 + \sum_{t=1}^{T-1} x_{t-1}^2 + x_{T-1}^2}{c^2} = \frac{1}{\sigma_{T-1}^2} + \frac{x_{T-1}^2}{c^2} \tag{8.23}$$

and, similarly,

$$\frac{\mu_T}{\sigma_T^2} = \frac{\mu_0 c^2/\sigma_0^2 + \sum_{t=1}^{T} x_t x_{t-1}}{c^2} =$$

$$= \frac{\mu_0 c^2/\sigma_0^2 + \sum_{t=1}^{T-1} x_t x_{t-1} + x_T x_{T-1}}{c^2} = \frac{\mu_{T-1}}{\sigma_{T-1}^2} + \frac{x_T x_{T-1}}{c^2} \quad (8.24)$$

$$\mu_T = \frac{\sigma_T^2}{\sigma_{T-1}^2}\mu_{T-1} + \frac{\sigma_T^2}{c^2} x_T x_{T-1} \quad (8.25)$$

8.3.c For a future sample we know that

$$X_{T+1} = Ax_T + cW_{T+1} \quad (8.26)$$

where the conditional density of A, given the previous observations $\underline{x} = (x_1, \ldots, x_T)$, is Gaussian with known mean μ_T and variance σ_T^2, and W_{T+1} is still Gaussian with mean 0 and variance 1. Therefore, the predictive distribution for X_{T+1} must also be Gaussian, with conditional mean and variance as

$$E\left[X_{T+1} \mid \underline{x}\right] = \mu_T x_T + 0 \quad (8.27)$$

$$\text{var}\left[X_{T+1} \mid \underline{x}\right] = \sigma_T^2 x_T^2 + c^2 \quad (8.28)$$

8.5.a The prior density of supplied voltage U, using only the company guarantee, is

$$f_U(u) = \frac{1}{\sigma_0\sqrt{2\pi}} e^{-\frac{(u-\mu_0)^2}{2\sigma_0^2}}$$

with mean $\mu_0 = 225$ V, and $\sigma_0 = 5$ V. Given any supplied voltage, any single measured value X is conditionally Gaussian

$$f_{X|U}(x \mid u) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-u)^2}{2\sigma^2}}$$

with mean u and standard deviation σ =3 V, given $U = u$. Thus, the joint density of the first measured voltage x_1 =215 V and the parameter is

$$f_{X,U}(x_1, u) = f_{X|U}(x_1 \mid u) f_U(u) \propto e^{-\frac{(x_1-u)^2}{2\sigma^2}} e^{-\frac{(u-\mu_0)^2}{2\sigma_0^2}} =$$

$$= e^{-\frac{1}{2}\left[u^2\left(\frac{1}{\sigma^2}+\frac{1}{\sigma_0^2}\right)+2u\left(\frac{x_1}{\sigma^2}+\frac{\mu_0}{\sigma_0^2}\right)+\ldots\right]}$$

where the ... include terms that do not depend on u. As the exponent is a second-degree polynomial in u, the posterior density for U is Gaussian

$$f_{U|X_1}(u \mid x_1) \propto e^{-\frac{1}{2}\frac{(u-\mu_1)^2}{\sigma_1^2}}$$

with mean μ_1 and variance σ_1^2, identified as

$$\begin{aligned}\frac{1}{\sigma_1^2} &= \frac{1}{\sigma^2} + \frac{1}{\sigma_0^2} = \frac{\sigma^2 + \sigma_0^2}{\sigma^2 \sigma_0^2} \\ \frac{\mu_1}{\sigma_1^2} &= \frac{x_1}{\sigma^2} + \frac{\mu_0}{\sigma_0^2} \\ \mu_1 &= \sigma_1^2 \left(\frac{x_1}{\sigma^2} + \frac{\mu_0}{\sigma_0^2} \right) = \frac{\sigma_0^2 x_1 + \sigma^2 \mu_0}{\sigma^2 + \sigma_0^2}\end{aligned}$$

A check of physical dimensions verifies that μ_1 has dimension V, and σ_1^2 has dimension V^2, as it should be.

8.5.b Now the joint density of all measurements and the parameter U is

$$f_{\underline{X},U}(\underline{x}, u) = f_U(u) \prod_{n=1}^{N} f_{X_n|U}(x_n \mid u) \propto e^{-\frac{\sum_n (x_n - u)^2}{2\sigma^2}} e^{-\frac{(u-\mu_0)^2}{2\sigma_0^2}} = \\ = e^{-\frac{1}{2}\left[u^2 \left(\frac{N}{\sigma^2} + \frac{1}{\sigma_0^2} \right) + 2u \left(\frac{\sum_n x_n}{\sigma^2} + \frac{\mu_0}{\sigma_0^2} \right) + \ldots \right]}$$

where the ... include terms that do not depend on u. Again, the exponent is a second-degree polynomial in u, and we can identify the posterior density as Gaussian

$$f_{U|\underline{X}}(u \mid \underline{x}) \propto e^{-\frac{1}{2}\frac{(u-\mu_N)^2}{\sigma_N^2}}$$

with mean μ_N and variance σ_N^2:

$$\begin{aligned}\frac{1}{\sigma_N^2} &= \frac{N}{\sigma^2} + \frac{1}{\sigma_0^2} = \frac{\sigma^2 + N\sigma_0^2}{\sigma^2 \sigma_0^2} \\ \frac{\mu_N}{\sigma_N^2} &= \frac{\sum_n x_n}{\sigma^2} + \frac{\mu_0}{\sigma_0^2} \\ \mu_N &= \sigma_N^2 \left(\frac{\sum_n x_n}{\sigma^2} + \frac{\mu_0}{\sigma_0^2} \right) = \frac{\sigma_0^2 \sum_n x_n + \sigma^2 \mu_0}{N\sigma_0^2 + \sigma^2}\end{aligned}$$

With increasing number of measurements the influence of the prior diminishes, and the result converges towards

$$\mu_N \to \frac{1}{N} \sum_{n=1}^{N} x_n$$

which is intuitively reasonable.

8.6.c The log-likelihood function is

$$L(w) = \ln f_{X|W}(x \mid w) = x \ln w - w - \ln x! \tag{8.29}$$

so the Fisher information is

$$I(w) = -E_X \left[\frac{\partial^2 L}{\partial w^2} \right] = \frac{E[X]}{w^2} = \frac{1}{w} \tag{8.30}$$

Thus, the Jeffreys prior is

$$f_W(w) \propto \frac{1}{\sqrt{w}} \tag{8.31}$$

This can be seen as the limit of a gamma density with shape $a_0 = 0.5$ and inverse scale $b_0 \to 0$.

Chapter 9

Approximate Bayesian Learning

9.1.a The conditional probability density of the observed sequence, given the parameter is

$$f_{\underline{X}|\Theta}(\underline{x} \mid \theta) = \prod_{t=1}^{T} f_{X_t|\Theta}(x_t \mid \theta) \tag{9.1}$$

With some normalization constant D, the exact posterior parameter density can be expressed, formally, as

$$f_{\Theta|\underline{X}}(\theta \mid \underline{x}) = \frac{1}{D} f_{\underline{X}|\Theta}(\underline{x} \mid \theta) f_{\Theta}(\theta) = \\ = \frac{1}{D}\frac{1}{c}\prod_{t=1}^{T}\left(0.5\frac{1}{\sqrt{2\pi}}e^{-x_t^2/2} + 0.5\frac{1}{\sqrt{2\pi}}e^{-(x_t-\theta)^2/2}\right) \tag{9.2}$$

The normalization constant is

$$D = \int_{-\infty}^{\infty} \frac{1}{c}\prod_{t=1}^{T}\left(0.5\frac{1}{\sqrt{2\pi}}e^{-x_t^2/2} + 0.5\frac{1}{\sqrt{2\pi}}e^{-(x_t-\theta)^2/2}\right) d\theta \tag{9.3}$$

For any $T \geq 1$, the first term inside the product in the integrand will cause a constant term that does not include θ, and therefore, D will always become infinite. For example, for $T = 1$,

$$D = \int_{-\infty}^{\infty} \frac{1}{c} 0.5\frac{1}{\sqrt{2\pi}}e^{-x_1^2/2}\, d\theta + \int_{-\infty}^{\infty} \frac{1}{c} 0.5\frac{1}{\sqrt{2\pi}}e^{-(x_1-\theta)^2/2}\, d\theta \tag{9.4}$$

where the second integral is nicely finite, but the first integral is infinite, for any constant c. Even for a very large T, the result will include an infinite term, as

$$D = \int_{-\infty}^{\infty} \frac{1}{c}\left(0.5\frac{1}{\sqrt{2\pi}}\right)^T e^{-\sum_t x_t^2/2}\, d\theta + \ldots \tag{9.5}$$

Of course, the normalization difficulty disappears if we choose a proper prior that approaches the uniform non-informative one only asymptotically.

9.1.c The joint log likelihood for the parameter and the observations is

$$L(\theta) = \ln f_{\Theta,\underline{X}}(\theta, \underline{x}) = \ln f_{\Theta}(\theta) + \sum_{t=1}^{T} \ln f_{X_t|\Theta}(x_t \mid \theta) = \\ = -\frac{\theta^2}{2\sigma_0^2} + \sum_{t=1}^{T} \ln\left(e^{-x_t^2/2} + e^{-(x_t-\theta)^2/2}\right) + \text{const.} \quad (9.6)$$

To find the Taylor approximation we need the first and second derivatives,

$$L'(\theta) = -\frac{\theta}{\sigma_0^2} + \sum_{t=1}^{T} \underbrace{\frac{e^{-(x_t-\theta)^2/2}}{e^{-x_t^2/2} + e^{-(x_t-\theta)^2/2}}}_{\gamma_t(\theta)} (x_t - \theta) \quad (9.7)$$

$$L''(\theta) = -\frac{1}{\sigma_0^2} + \sum_{t=1}^{T} -\gamma_t(\theta) + \gamma_t'(\theta)(x_t - \theta) \quad (9.8)$$

where

$$\gamma_t'(\theta) = \frac{e^{-(x_t-\theta)^2/2}(x_t-\theta)}{e^{-x_t^2/2} + e^{-(x_t-\theta)^2/2}} - \frac{e^{-(x_t-\theta)^2/2} e^{-(x_t-\theta)^2/2}(x_t-\theta)}{\left(e^{-x_t^2/2} + e^{-(x_t-\theta)^2/2}\right)^2} = \\ = \left(\gamma_t(\theta) - \gamma_t^2(\theta)\right)(x_t - \theta) \quad (9.9)$$

At the point $\theta = \mu$ where $L(\theta)$ is maximal, we must have $L'(\theta) = 0$, i.e.,

$$\mu = \frac{\sum_t \gamma_t(\mu) x_t}{1/\sigma_0^2 + \sum_t \gamma_t(\mu)} \quad (9.10)$$

We can find this solution by iterating

$$\mu \leftarrow \frac{\sum_t \gamma_t(\mu) x_t}{1/\sigma_0^2 + \sum_t \gamma_t(\mu)} \quad (9.11)$$

Note that, if $\sigma_0 \to \infty$, this is a similar update equation as in the VI approach. The difference is that the γ_t in the VI iteration also accounted for the posterior variance σ^2 for Θ.

The Taylor approximation around μ is

$$L(\theta) \approx \frac{L''(\mu)}{2}(\theta - \mu)^2 + C \quad (9.12)$$

Using this approximation, the posterior density is

$$f_{\Theta|\underline{X}}(\theta \mid \underline{x}) \propto e^{L(\theta)} \approx e^C e^{\frac{L''(\mu)(\theta-\mu)^2}{2}} \quad (9.13)$$

Thus, the approximation is Gaussian, with mean μ and inverse variance

$$\frac{1}{\sigma^2} = -L''(\mu) = \frac{1}{\sigma_0^2} + \sum_{t=1}^{T} \gamma_t(\mu) - \Big(\gamma_t(\mu) - \gamma_t^2(\mu)\Big)\,(x_t - \mu)^2 \tag{9.14}$$

Again, this is similar to the VI solution. The terms involving $(x_t - \mu)^2$ become very small if γ_t are near zero or one, i.e., for all observations for which it is rather clear to which component they "belong". In the worst case, when $\mu \approx 0$, the responsibility factors are $\gamma_t \approx 0.5$ for all t, so then

$$\frac{1}{\sigma^2} \approx \frac{1}{\sigma_0^2} + 0.75 \sum_{t=1}^{T} \gamma_t(\mu) \tag{9.15}$$

i.e., about 75% of the nominal result in the VI approach. On the other hand, in the VI approach all the γ_t factors were slightly smaller than here, because they already included an effect of the posterior uncertainty σ^2.

9.2.a *Observation Model*: If parameters and previous observations were exactly known, each new sample X_t is Gaussian with mean ax_{t-1} and variance c^2, normally, and variance $c^2 + d^2$, if a disturbance occurred i.e., if $Z_t = 1$. This can be formally expressed as

$$\begin{aligned} f_{X_t|X_{t-1},Z_t,A}(x_t \mid x_{t-1}, z_t, a) = \\ = \begin{cases} \frac{1}{\sqrt{2\pi}c} e^{-\frac{(x_t - a x_{t-1})^2}{2c^2}}, & z_t = 0 \\ \frac{1}{\sqrt{2\pi}\sqrt{c^2+d^2}} e^{-\frac{(x_t - a x_{t-1})^2}{2(c^2+d^2)}}, & z_t = 1 \end{cases} \end{aligned} \tag{9.16}$$

The conditional probability density of the complete observed sequence can then be expressed as

$$\begin{aligned} f_{\underline{X}|\underline{Z},A}(\underline{x} \mid \underline{z}, a) = \prod_{t=1}^{T} f_{X_t|X_{t-1},Z_t,A}(x_t \mid x_{t-1}, z_t, a) = \\ = \prod_{t=1}^{T} \left(\frac{1}{\sqrt{2\pi}c} e^{-\frac{(x_t - a x_{t-1})^2}{2c^2}} \right)^{1-z_t} \cdot \left(\frac{1}{\sqrt{2\pi}\sqrt{c^2+d^2}} e^{-\frac{(x_t - a x_{t-1})^2}{2(c^2+d^2)}} \right)^{z_t} \end{aligned} \tag{9.17}$$

Prior Distributions: The probability of a disturbance is

$$\begin{cases} P\left[Z_t = 1 \mid W = w\right] = w \\ P\left[Z_t = 0 \mid W = w\right] = 1 - w \end{cases} \tag{9.18}$$

This can also be expressed as a probability mass function

$$f_{Z_t|W}(z_t \mid w) = w^{z_t}(1-w)^{1-z_t} \tag{9.19}$$

For the hyper-parameter W we assume a prior Beta distribution

$$f_W(w) \propto w^{\alpha_0 - 1}(1-w)^{\beta_0 - 1} \tag{9.20}$$

For the auto-regresssive coefficient A we assume a zero-mean Gaussian prior:

$$f_A(a) = \frac{1}{\sqrt{2\pi}\sigma_0} e^{-\frac{a^2}{2\sigma_0^2}} \tag{9.21}$$

Now it is easy to write the joint log-probability of all observations, hidden variables, and parameters, as

$$\begin{aligned} \ln f_{\underline{X},\underline{Z},W,A}(\underline{x},\underline{z},w,a) &= \ln f_{\underline{X}|\underline{Z},A}(\underline{x} \mid \underline{z},a) f_{Z_t|W}(z_t \mid w) f_W(w) f_A(a) = \\ = \sum_{t=1}^{T} -z_t \left(\ln \sqrt{c^2+d^2} + \frac{(x_t - a x_{t-1})^2}{2(c^2+d^2)} \right) &- (1-z_t) \left(\ln c + \frac{(x_t - a x_{t-1})^2}{2c^2} \right) + \\ + \sum_{t=1}^{T} z_t \ln w + (1-z_t)\ln(1-w) &+ \\ + (\alpha_0 - 1)\ln w + (\beta_0 - 1)\ln(1-w) - \frac{a^2}{2\sigma_0^2} &+ \ldots \end{aligned} \tag{9.22}$$

where remaining terms ... are constant and do not include any of the random variables.

We must now try to find a form for the suggested factorized approximation $\ln q_{A,W}(a,w) q_{\underline{Z}}(\underline{z})$ that is similar to this total log-probability expression. As (9.22) includes no terms involving both a and w, we immediately see that the additional factorization $q_{A,W}(a,w) = q_A(a) q_W(w)$ appears naturally, without any further enforced approximation. Also, as the log-probability expression includes only separate terms for each t, we can also immediately factorize

$$q_{\underline{Z}}(\underline{z}) = \prod_{t=1}^{T} q_{Z_t}(z_t) \tag{9.23}$$

without introducing any further approximation. Also, the expression includes only terms with either z_t or $(1-z_t)$, so the posterior $q_{Z_t}(z_t)$ is, like the prior, still just a binary (Bernoulli) probability mass of the form

$$q_{Z_t}(z_t) = \gamma_t^{z_t}(1-\gamma_t)^{1-z_t} \tag{9.24}$$

with a mean $\gamma_t = E_{q_{Z_t}}[Z_t \mid \underline{X}]$ indicating the posterior probability that a disturbance was present at time t. For the update of the other variables, we use the value of γ_t that was obtained in the previous iterations. The new updated γ_t will be found later.

We can now apply the general VI solution for the factorized approximation: To find $q_A(a)$, we just take the expected value over all the other

variables, using $q_{\underline{Z}}$ and q_W. We need to consider only the terms that involve a to obtain

$$\begin{aligned}\ln q_A(a) &= E_{q_{\underline{Z}}}\left[\ln f_{\underline{X},\underline{Z},W,A}(\underline{x},\underline{z},w,a)\right] + \ldots = \\ &= -\frac{a^2}{2\sigma_0^2} - \sum_t (1-\gamma_t)\frac{(x_t - a x_{t-1})^2}{2c^2} + \gamma_t \frac{(x_t - a x_{t-1})^2}{2(c^2+d^2)} + \ldots \end{aligned} \tag{9.25}$$

As this is a second-degree polynomial in a, we conclude that q_A must be a Gaussian probability density function, i.e.,

$$\ln q_A(a) = -\frac{(a-\mu_T)^2}{2\sigma^2} + \ldots = -\frac{a^2}{2\sigma_T^2} + \frac{a\mu_T}{\sigma_T^2} + \ldots \tag{9.26}$$

which is completely specified by its mean μ_T and variance σ_T^2. To identify these hyper-parameters, we re-write (9.25) as

$$\begin{aligned}\ln q_A(a) = -\frac{a^2}{2\sigma_0^2} - \frac{a^2}{2}\sum_t \underbrace{\left(\frac{1-\gamma_t}{c^2} + \frac{\gamma_t}{c^2+d^2}\right)}_{\eta_t} x_{t-1}^2 + \\ + a\sum_t \underbrace{\left(\frac{1-\gamma_t}{c^2} + \frac{\gamma_t}{c^2+d^2}\right)}_{\eta_t} x_t x_{t-1} + \ldots \end{aligned} \tag{9.27}$$

Thus, using the weight factors η_t, we identify

$$\frac{1}{\sigma_T^2} = \frac{1}{\sigma_0^2} + \sum_{t=1}^{T} \eta_t x_{t-1}^2 \tag{9.28}$$

$$\frac{\mu_T}{\sigma_T^2} = \sum_{t=1}^{T} \eta_t x_t x_{t-1} \tag{9.29}$$

$$\mu_T = \frac{\sum_t \eta_t x_t x_{t-1}}{\frac{1}{\sigma_0^2} + \sum_t \eta_t x_{t-1}^2} \tag{9.30}$$

To update $q_W(w)$, we only use the terms in (9.22) that include w. This gives

$$\begin{aligned}\ln q_W(w) = \sum_t \gamma_t \ln w + (1-\gamma_t)\ln(1-w) + \\ + (\alpha_0 - 1)\ln w + (\beta_0 - 1)\ln(1-w) + \ldots = \\ = (\alpha_T - 1)\ln w + (\beta_T - 1)\ln(1-w) + \ldots \end{aligned} \tag{9.31}$$

Thus, the posterior $q_W(w) \propto w^{\alpha_T - 1}(1-w)^{\beta_T - 1}$ is still a beta density, with updated hyper-parameters

$$\alpha_T = \alpha_0 + \sum_{t=1}^{T} \gamma_t \tag{9.32}$$

$$\beta_T = \beta_0 + \sum_{t=1}^{T} (1-\gamma_t) \tag{9.33}$$

Now, finally, we must find expressions for γ_t to be used in the next round. To find the updated $q_{Z_t}(z_t)$, we only use the terms in (9.22) that include z_t, and take the expected value over the other variables, using the recently updated q_A and q_W:

$$\begin{aligned}\ln q_{Z_t}(z_t) = &-z_t \ln \sqrt{c^2+d^2} - (1-z_t)\ln c \\ &- z_t \frac{E_{q_A}\left[(x_t - Ax_{t-1})^2\right]}{2(c^2+d^2)} - (1-z_t)\frac{E_{q_A}\left[(x_t - Ax_{t-1})^2\right]}{2c^2} + \\ &+ z_t E_{q_W}\left[\ln W\right] + (1-z_t)E_{q_W}\left[\ln(1-W)\right] + \ldots = \\ &= z_t r_{1,t} + (1-z_t)r_{0,1} + \ldots \end{aligned} \tag{9.34}$$

To get final expressions for the factors $r_{0,t}$ and $r_{1,t}$, we can write the $E_{q_A}[\,\cdot\,]$ factor as

$$y_t^2 = E_{q_A}\left[(x_t - Ax_{t-1})^2\right] = x_t^2 - 2\mu_T x_t x_{t-1} + (\mu_T^2 + \sigma_T^2)x_{t-1}^2 \tag{9.35}$$

For the beta distribution q_W, expected values are known as

$$E_{q_W}[W] = \frac{\alpha_T}{\alpha_T + \beta_T} \tag{9.36}$$

$$E_{q_W}[\ln W] = \psi(\alpha_T) - \psi(\alpha_T + \beta_T) \tag{9.37}$$

$$E_{q_W}[\ln(1-W)] = \psi(\beta_T) - \psi(\alpha_T + \beta_T) \tag{9.38}$$

where $\psi(x) = \frac{d\ln\Gamma(x)}{dx}$ is known as the *digamma* function. Using these expressions for $E_{q_W}[\,\cdot\,]$ and $E_{q_A}[\,\cdot\,]$, we obtain

$$r_{1,t} = \psi(\alpha_T) - \psi(\alpha_T + \beta_T) - \ln\sqrt{c^2+d^2} - \frac{y_t^2}{2(c^2+d^2)} \tag{9.39}$$

$$r_{0,t} = \psi(\beta_T) - \psi(\alpha_T + \beta_T) - \ln c - \frac{y_t^2}{2c^2} \tag{9.40}$$

Here, the term $\psi(\alpha_T+\beta_T)$ which is common in both $r_{0,t}$ and $r_{1,t}$ can actually be omitted, as it only contributes a common scale factor that will be divided out anyway in the normalization later. Then, we can finally express the updated probability mass for Z_t as

$$q_{Z_t}(z_t) \propto (e^{r_{1,t}})^{z_t}\,(e^{r_{0,t}})^{1-z_t} \tag{9.41}$$

which can be normalized as

$$q_{Z_t}(z_t) = \gamma_t^{z_t}(1-\gamma_t)^{1-z_t} \tag{9.42}$$

by choosing

$$\gamma_t = \frac{e^{r_{1,t}}}{e^{r_{0,t}} + e^{r_{1,t}}} \tag{9.43}$$

Now, all hyper-parameters have been updated, and can be used in the next iteration round. The only remaining issue is how to initialize the VI learning. In this case, the disturbance causes an increase in the variance of X_t. Therefore, we might crudely assign $\gamma_t = 1$ for all t where x_t^2 is greater than some threshold. The threshold can be chosen as, e.g., the 90-th percentile across all x_t^2, if we initially guess that the disturbance occurs with a small probability.

9.2.b For a future sample we know that

$$X_{T+1} = Ax_T + cU_{T+1} + dZ_{T+1}V_{T+1} \tag{9.44}$$

where the conditional density of A, given the previous observations $\underline{x} = (x_1, \ldots, x_T)$, is Gaussian with known mean μ_T and variance σ_T^2, and U_{T+1} and V_{T+1} are still Gaussian with mean 0 and variance 1. Therefore, the predictive distribution for X_{T+1} must also be Gaussian, with conditional mean and variance as

$$E\left[X_{T+1} \mid \underline{x}\right] = \mu_T x_T + 0 \tag{9.45}$$

$$\text{var}\left[X_{T+1} \mid \underline{x}\right] = \sigma_T^2 x_T^2 + c^2 + d^2 E_{q_{Z_t}}\left[Z_{T+1}\right] \tag{9.46}$$

where $E\left[Z_{T+1}\right] = E_{q_W}\left[W\right] = \alpha_T/\beta_T$, as determined by the final posterior beta distribution for W.