KTH Electrical Engineering

Solution Examples to Exam in Pattern Recognition EN2202

Date: Friday Oct 19, 2012, 08:00 – 13:00

Place: D32, D34.

Allowed: Beta (or corresponding), calculator with empty memory. No notes!

Grades: A: 31p; B: 27p; C: 23p; D: 20p; E: 17; of max 25p + 10p project bonus.

Language: Swedish or English.

Results: Friday, Nov 9.

Review: At KTH-S3/STEX, Osquldas v. 10.

Contact: Arne Leijon, 070 274 6904; Gustav Eje Henter, 070 526 0185

Good Luck!

Please do the **Course Evaluation**! See the course web page.

1 (Directional Classifier) A classifier is designed to use a microphone array to distinguish between two sound sources, using only a series of estimates of the *direction* to the currently active source.

For each short time frame, a feature extractor uses the input signals from all microphones to estimate the direction to the source in the horizontal plane. To avoid the problem that angles θ and $\theta + 2\pi$ are exactly equivalent, the direction angle θ_t estimated from signal frame no. t is transformed into a two-element vector with norm 1, as

$$\boldsymbol{x}_t = \begin{pmatrix} \cos\theta_t \\ \sin\theta_t \end{pmatrix}$$

The general *von Mises-Fisher* probability distribution has been defined for directional data represented by a D-dimensional random vector $\boldsymbol{X}$, normalized to unit length $\|\boldsymbol{X}\|^2 = \boldsymbol{X}^T\boldsymbol{X} = 1$. The von-Mises-Fisher probability density function for such a vector is

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{\lambda^{D/2-1}}{(2\pi)^{D/2} I_{D/2-1}(\lambda)} e^{\lambda \boldsymbol{\mu}^T \boldsymbol{x}}$$

Here the parameter $\boldsymbol{\mu}$ is also normalized as $\|\boldsymbol{\mu}\|^2 = \boldsymbol{\mu}^T\boldsymbol{\mu} = 1$ and indicates the most probable direction for $\boldsymbol{X}$, and $\lambda \geq 0$ is a concentration parameter. For high positive values of λ, the probability density for $\boldsymbol{X}$ is sharply peaked near $\boldsymbol{\mu}$. If $\lambda = 0$, the probability density is constant for all directions $\boldsymbol{X}$. The function $I_\nu(\lambda)$ in the normalization factor is the *modified Bessel function* of the first kind and order ν. This function can be regarded as known, and it is available numerically as `besseli` in Matlab.

For the present application, where $D = 2$, you have used training data to estimate von-Mises-Fisher distribution parameters $(\boldsymbol{\mu}_1, \lambda)$ and $(\boldsymbol{\mu}_2, \lambda)$ for each of the two different sound sources $S = 1$ and $S = 2$. The concentration parameter λ is equal for both sources. The two sources occur with known prior probabilities $P[\,S = 1\,] = p_1$ and $P[\,S = 2\,] = p_2 = 1 - p_1$.

(a) Calculate the posterior probabilities $P[\,S = i \mid \underline{\boldsymbol{x}}\,]$ that each of the two sound sources was active, given a sequence $\underline{\boldsymbol{x}} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T)$ of observed feature vectors. The feature vectors are recorded for non-overlapping signal segments, so that all feature vectors in the sequence can be regarded as conditionally independent and identically distributed, given the known von-Mises-Fisher distribution for the actual sound source. (3p)

Solution:

Decision Criterion: As both source categories may have unequal prior probabilities, the classifier must use a MAP decision. The posterior source probabilities are, by Bayes' rule,

$$P[\,S = i \mid \underline{\boldsymbol{x}}\,] = \frac{f_{\underline{\boldsymbol{X}}|S}(\underline{\boldsymbol{x}} \mid i) p_i}{f_{\underline{\boldsymbol{X}}}(\underline{\boldsymbol{x}})} \tag{1}$$

Here, the denominator is identical for both $i = 1$ and $i = 2$, as

$$f_{\underline{\boldsymbol{X}}}(\underline{\boldsymbol{x}}) = f_{\underline{\boldsymbol{X}}|S}(\underline{\boldsymbol{x}} \mid 1) p_1 + f_{\underline{\boldsymbol{X}}|S}(\underline{\boldsymbol{x}} \mid 2) p_2$$

We can define *Discriminant Functions*, as

$$g_i(\underline{\boldsymbol{x}}) = \ln f_{\underline{\boldsymbol{X}}|S}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T \mid i) p_i \tag{2}$$

and then obtain the posterior probabilities by normalization, as

$$P\,[\,S=1\mid \underline{\boldsymbol{x}}\,] = \frac{e^{g_1(\underline{\boldsymbol{x}})}}{e^{g_1(\underline{\boldsymbol{x}})}+e^{g_2(\underline{\boldsymbol{x}})}} = \frac{e^{g_1(\underline{\boldsymbol{x}})-g_2(\underline{\boldsymbol{x}})}}{e^{g_1(\underline{\boldsymbol{x}})-g_2(\underline{\boldsymbol{x}})}+1}$$
$$P\,[\,S=2\mid \underline{\boldsymbol{x}}\,] = 1-P\,[\,S=1\mid \underline{\boldsymbol{x}}\,]$$

As the vectors $\boldsymbol{x}_t$ in the observed sequence are independent and equally distributed, and $D=2$, we have

$$f_{\underline{\boldsymbol{X}}|S}(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_T\mid i) = \prod_{t=1}^{T} f_{\boldsymbol{X}_t|S}(\boldsymbol{x}_t\mid i) = \prod_{t=1}^{T}\frac{1}{2\pi I_0(\lambda)}e^{\lambda\boldsymbol{\mu}_i^T\boldsymbol{x}_t} \tag{3}$$

Thus,

$$g_i(\underline{\boldsymbol{x}}) = \ln p_i - T\ln 2\pi I_0(\lambda) + \sum_{t=1}^{T}\lambda\boldsymbol{\mu}_i^T\boldsymbol{x}_t \tag{4}$$

The posterior conditional probabilities can also be expressed more simply as

$$P\,[\,S=1\mid \underline{\boldsymbol{x}}\,] = \frac{e^{g(\underline{\boldsymbol{x}})}}{e^{g(\underline{\boldsymbol{x}})}+1}$$
$$P\,[\,S=2\mid \underline{\boldsymbol{x}}\,] = \frac{1}{e^{g(\underline{\boldsymbol{x}})}+1}$$

using the single discriminant function

$$g(\underline{\boldsymbol{x}}) = g_1(\underline{\boldsymbol{x}})-g_2(\underline{\boldsymbol{x}}) = \ln\frac{p_1}{1-p_1}+\sum_{t=1}^{T}\left(\lambda\boldsymbol{\mu}_1^T\boldsymbol{x}_t-\lambda\boldsymbol{\mu}_2^T\boldsymbol{x}_t\right) = \ln\frac{p_1}{1-p_1}+\lambda(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^T\sum_{t=1}^{T}\boldsymbol{x}_t \tag{5}$$

where the common terms in the two separate discriminant functions have cancelled each other.

(b) Show that a classifier can make optimal decisions for minimal error probability, using only a single *linear* discriminant function $g(\underline{\boldsymbol{x}})$, given the observed feature-vector sequence $\underline{\boldsymbol{x}}$. (2p)

Solution: *Decision Rule*: Optimal MAP decisions are obtained by the simple decision rule

$$d(\underline{\boldsymbol{x}}) = \begin{cases} 1, & P\,[\,S=1\mid \underline{\boldsymbol{x}}\,] \geq P\,[\,S=2\mid \underline{\boldsymbol{x}}\,] \\ 2, & \text{otherwise} \end{cases} \tag{6}$$

or equivalently using the single discriminant function directly, as

$$d(\underline{\boldsymbol{x}}) = \begin{cases} 1, & g(\underline{\boldsymbol{x}}) \geq 0 \\ 2, & g(\underline{\boldsymbol{x}}) < 0 \end{cases} \tag{7}$$

We have already shown above, that the single discriminant function can be expressed as

$$g(\underline{\boldsymbol{x}}) = \ln\frac{p_1}{1-p_1} + \lambda(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^T\sum_{t=1}^{T}\boldsymbol{x}_t \tag{8}$$

This single discriminant function is obviously a linear combination of all the observed feature-vector elements.

2 Determine for each of the following statements whether it is *true* or *false*.
No motivation is required, and given motivations will not be considered, but you should be certain about your choice. A correct answer gives +1 point, no answer gives 0 points, but an incorrect answer gives −1 point! A negative sum will be counted as 0. The final result on this problem can be any integer from 0p to the maximum of (5p).

(a) The initial probability vector

$$q = \begin{pmatrix} 0.5 \\ 0.25 \\ 0.25 \end{pmatrix}, \text{ and transition matrix } A = \begin{pmatrix} 0 & 0.5 & 0.5 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix},$$

together define an *ergodic* Markov chain.
Solution: FALSE. This Markov chain is periodic, as state 1 will be re-visited every second transition.

(b) The probability density function of a GMM with M components, having weight factors $w_m > 0$ for all $m \in \{1, \ldots, M\}$, has M local maxima.
Solution: FALSE. The GMM might have fewer than M maxima, e.g., if the mean parameter is identical in different components. The GMM can also have more than M maxima, e.g., if some component densities are overlapping and add to form a peak in some region where none of the components have their main peak.

(c) The joint distribution of a sequence of N consecutive samples from an M-component Gaussian mixture model (GMM) can equivalently be described by an M-state HMM with Gaussian output distributions.
Solution: TRUE. If the initial state probability vector of the HMM, and all the transition probability vectors for all M HMM states, are equal to the vector of mixture weights of the GMM, and the HMM output distributions are equal to the corresponding GMM component distributions, then the HMM is exactly equivalent to the GMM.

(d) An optimal ML classifier is designed to calculate log-probabilities $\ln P[\,\boldsymbol{x} \mid S = n\,]$ for all source alternatives $n \in \{1, \ldots, N_s\}$, for an observed Gaussian feature vector $\boldsymbol{x}$, containing K feature elements. The calculated log-probabilities $\ln P[\,\boldsymbol{x} \mid S = n\,]$ tend to become smaller, if the number of feature elements, K, is increased.
Solution: FALSE. The probability density for $\boldsymbol{X}$ depends entirely on the numerical scale of the features. If all the feature values are typically very small, $|x_k| \ll 1$, the probability density is typically $\gg 1$ for the $\boldsymbol{x}$ values that occur in practice, and increases with the number of elements. Conversely, if feature elements are numerically large, the probability density will be spread out and small everywhere, and decrease as the number of observed features is increased.

(e) A general mixture model for a scalar random variable X can be formally defined as

$$f_X(x) = \sum_{m=1}^{M} w_m g_m(x),$$

where $g_m(x)$ represents the probability distribution of the m-th mixture component. Here, $g_m(x)$ can be any type of distribution, e.g., Gaussian, gamma, beta, or discrete. However, if the

components $g_m(x)$, for all m, are probability-mass functions for different *discrete* distributions, the total mixture model is exactly equivalent to only a single probability mass function $g(x)$ defined as a vector with elements $g(x_k) = p_k$ for each possible outcome x_k.
Solution: TRUE. A weighted sum of M discrete probability mass functions is just equal to a single probability mass function, as

$$P[X = x_k] = \sum_m w_m P[X = x_k \mid m] = \sum_m w_m g_m(x_k) = p_k$$

3 You can observe some elements of the output sequence $\underline{x} = (x_1, \ldots, x_t, \ldots)$ from a discrete hidden Markov source, but you do not know the corresponding internal state sequence $\underline{S} = (S_1, \ldots, S_t, \ldots)$ in the source. The initial state probability vector is

$$q = \begin{pmatrix} 0.2 \\ 0.8 \end{pmatrix}, \text{with elements } P(S_1 = i).$$

The state transition probability matrix is

$$A = \begin{pmatrix} 0.6 & 0.4 \\ 0.1 & 0.9 \end{pmatrix}, \text{with elements } a_{ij} = P(S_{t+1} = j \mid S_t = i).$$

The output probability matrix is

$$B = \begin{pmatrix} 0.1 & 0.4 & 0.5 \\ 0.7 & 0.2 & 0.1 \end{pmatrix}, \text{with elements } b_{ik} = P(X_t = k \mid S_t = i).$$

(a) Calculate $P(X_5 = 1)$. (2p)
Solution: The HMM is stationary with state probabilities $P[S_t = 1] = p_1 = 0.2$ and $P[S_t = 2] = p_2 = 0.8$ for any t. This can be verified by checking that $A^T q = q$. Thus,

$$\begin{aligned} P(X_5 = 1) = P(X_5 = 1 \mid S_5 = 1)P[S_5 = 1] + P(X_5 = 1 \mid S_5 = 2)P[S_5 = 2] = \\ = b_{11}p_1 + b_{21}p_2 = 0.1 \cdot 0.2 + 0.7 \cdot 0.8 = 0.58 \end{aligned} \quad (9)$$

(b) Calculate $P(X_8 = 1 \mid X_6 = 1 \cap S_6 = 2 \cap X_7 = 1 \cap S_7 = 1 \cap S_9 = 1 \cap X_{10} = 3)$. (3p)

Solution: We can disregard some events that do not influence the conditional probability. The event $X_8 = 1$ is conditionally independent of X_6, X_7, X_{10}, S_6 given S_7, S_9. Therefore,

$$\begin{aligned} P\,[\,X_8 = 1 \mid X_6 = 1 \cap S_6 = 2 \cap X_7 = 1 \cap S_7 = 1 \cap S_9 = 1 \cap X_{10} = 3\,] = \\ = P\,[\,X_8 = 1 \mid S_7 = 1 \cap S_9 = 1\,] = \\ = \frac{P\,[\,X_8 = 1 \cap S_9 = 1 \mid S_7 = 1\,]}{P\,[\,S_9 = 1 \mid S_7 = 1\,]} = \\ = \frac{\sum_{i=1}^{2} P\,[\,S_8 = i \cap X_8 = 1 \cap S_9 = 1 \mid S_7 = 1\,]}{\sum_{i=1}^{2} P\,[\,S_8 = i \cap S_9 = 1 \mid S_7 = 1\,]} = \\ = \frac{\sum_{i=1}^{2} a_{1i} b_{i1} a_{i1}}{\sum_{i=1}^{2} a_{1i} a_{i1}} = \frac{0.6 \cdot 0.1 \cdot 0.6 + 0.4 \cdot 0.7 \cdot 0.1}{0.6 \cdot 0.6 + 0.4 \cdot 0.1} = 0.16 \end{aligned} \quad (10)$$

4 (Bayesian Learning) You have observed a sequence of scalar feature values $\underline{x} = (x_1, \ldots, x_T)$, where each element x_t is regarded as an outcome of a random variable X_t with a *uniform* distribution, and all variables in the sequence are statistically independent and identically distributed. The probability density for X_t is constant in the range $0 < X_t \leq W$, and zero outside this range. However, the upper limit W of the range is totally unknown. You will now apply Bayesian learning for W, using the observed sequence $\underline{x}$.

(a) Determine the non-informative (possibly improper) *Jeffreys prior* density for W. (1p)
Hint: The Jeffreys prior density for a scalar parameter W that determines the distribution of a random variable X is

$$f_W(w) = \sqrt{I(w)}, \text{ where } I(w) = E_X\left[\left(\frac{\partial \ln f_{X|W}(X \mid w)}{\partial w}\right)^2 \middle| W = w\right]$$

Solution: As all observations are i.i.d., it is enough to consider a single feature X_t, called just X in the following. The conditional feature distribution, given the parameter, is

$$f_{X|W}(x \mid w) = \begin{cases} 1/w & 0 < x \leq w \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

Thus,

$$\ln f_{X|W}(x \mid w) = \begin{cases} -\ln w & 0 < x \leq w \\ \ln 0 & \text{otherwise} \end{cases} \tag{12}$$

(To be formal, we can treat $\ln 0$ as a constant here, and let this value approach $-\infty$ later, at the end of all calculations.)

$$\left(\frac{\partial \ln f_{X|W}(x \mid w)}{\partial w}\right)^2 = \begin{cases} 1/w^2 & 0 < x \leq w \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

(We can argue that the derivative of $\ln 0$ is actually zero, by considering $\ln 0$ as a constant value that approaches $-\infty$ in the limit, but we take the derivative of the constant before we actually let its value approach infinity.)

$$E_X\left[\left(\frac{\partial \ln f_{X|W}(X \mid w)}{\partial w}\right)^2 \middle| W = w\right] = \int_0^w \frac{1}{w}\frac{1}{w^2}\,dx = \frac{1}{w^2} \tag{14}$$

Thus, the Jeffreys prior is

$$f_W(w) \propto \frac{1}{w} \tag{15}$$

which is an improper density function as it cannot be normalized.

(b) Determine the exact posterior probability density function for the parameter W, expressed in terms of the observed sequence $\underline{x}$ or its elements x_t. (2p)
Hint: Any conjugate prior density may be used, if you could not find the Jeffreys prior.
Solution: The conditional distribution for the complete observed feature sequence, given the parameter, is

$$f_{\underline{X}|W}(x_1, \ldots, x_T \mid w) = \begin{cases} \prod_{t=1}^T 1/w, & x_t \leq w, \text{ for all } t \\ 0, & \text{otherwise} \end{cases} \tag{16}$$

For a general conjugate prior we may choose

$$f_W(w) \propto \begin{cases} w^{-\alpha_0}, & u_0 \leq w \\ 0, & \text{otherwise} \end{cases} \tag{17}$$

The Jeffreys prior would then be obtained with $\alpha_0 = 1$ and $u_0 \to 0$. Thus, the posterior density for the parameter is

$$f_{W|\underline{X}}(w \mid x_1, \ldots, x_T) \propto f_{\underline{X}|W}(x_1, \ldots, x_T \mid w) f_W(w) = \begin{cases} 1/w^{\alpha_T}, & u_T \leq w \\ 0, & \text{otherwise} \end{cases} \tag{18}$$

where $\alpha_T = T + \alpha_0$, and $u_T = \max(u_0, \max_t x_t)$. We must normalize the posterior density as

$$f_{W|\underline{X}}(w \mid x_1, \ldots, x_T) = \frac{1}{C} \frac{1}{w^{\alpha_T}}, \quad u_T \leq w \tag{19}$$

where

$$C = \int_{u_T}^{\infty} w^{-\alpha_T} dw = \frac{u_T^{-\alpha_T+1}}{\alpha_T - 1} \text{ for } \alpha \neq 1 \tag{20}$$

Thus, finally, the posterior density can be written as

$$f_{W|\underline{X}}(w \mid \underline{x}) = \begin{cases} \frac{\alpha_T - 1}{u_T} \left(\frac{u_T}{w}\right)^{\alpha_T}, & u_T \leq w \\ 0, & \text{otherwise} \end{cases} \tag{21}$$

This formulation allows a simple dimensionality test: assuming that x_t and w and u_T all have dimension V, the posterior density has dimension 1/V, as it should.

A similar estimation problem for a uniform *discrete* distribution has historical interest and is discussed at
`http://en.wikipedia.org/wiki/German_tank_problem`

(c) Determine the predictive probability density for any future observation x_{T+1}, given the training sequence $\underline{x}$. (2p)

Solution: For any potential future observation x, the predictive probability density is

$$f_{X_{T+1}|\underline{X}}(x \mid \underline{x}) = \int_{-\infty}^{\infty} f_{X_{T+1}|W}(x \mid w) f_{W|\underline{X}}(w \mid \underline{x})\, dw = \\ = \int_{u_T}^{\infty} f_{X_{T+1}|W}(x \mid w) \frac{\alpha_T - 1}{u_T} \left(\frac{u_T}{w}\right)^{\alpha_T} dw \tag{22}$$

As we have

$$f_{X_{T+1}|W}(x \mid w) = \begin{cases} 1/w & 0 < x \leq w \\ 0 & \text{otherwise} \end{cases} \tag{23}$$

we must separate two cases:

If the new observation is in the range $0 < x \leq u_T$ the integration range does not depend on x, so the probability density is constant:

$$f_{X_{T+1}|\underline{X}}(x \mid \underline{x}) = \int_{u_T}^{\infty} \frac{1}{w} \frac{\alpha_T - 1}{u_T} \left(\frac{u_T}{w}\right)^{\alpha_T} dw = \frac{\alpha_T - 1}{\alpha_T u_T} \tag{24}$$

If the new observation is above the previous upper limit $u_T < x$, the integrand is zero for $u_T < w < x$ and the integral becomes a function of x:

$$f_{X_{T+1}|\underline{X}}(x \mid \underline{x}) = \int_x^\infty \frac{1}{w} \frac{\alpha_T - 1}{u_T} \left(\frac{u_T}{w}\right)^{\alpha_T} dw = \frac{\alpha_T - 1}{\alpha_T u_T} \left(\frac{u_T}{x}\right)^{\alpha_T} \tag{25}$$

Summarizing these results,

$$f_{X_{T+1}|\underline{X}}(x \mid \underline{x}) = \frac{\alpha_T - 1}{\alpha_T u_T} \begin{cases} 0, & x \leq 0 \\ 1, & 0 < x \leq u_T \\ \left(\frac{u_T}{x}\right)^{\alpha_T}, & u_T < x \end{cases} \tag{26}$$

Again, a dimensionality check shows that this result is reasonable. It is also possible to verify that $\int_0^\infty f_{X_{T+1}|\underline{X}}(x \mid \underline{x})\, dx = 1$ as it should.

5 (Expectation Maximization) You have observed a sequence $\underline{\boldsymbol{x}} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T)$ of column vectors $\boldsymbol{x}_t$, where each vector $\boldsymbol{x}_t$ encodes a measured *direction* to a sound source in the horizontal plane as

$$\boldsymbol{x}_t = \begin{pmatrix} \cos\theta_t \\ \sin\theta_t \end{pmatrix}$$

Thus, each vector $\boldsymbol{x}_t$ has $D = 2$ elements and is normalized as $\|\boldsymbol{x}_t\|^2 = \boldsymbol{x}_t^T \boldsymbol{x}_t = 1$. Each observed vector is regarded as an outcome of a random vector $\boldsymbol{X}_t$ with a *von Mises-Fisher* distribution, and all vectors in the sequence are statistically independent and identically distributed. As the dimensionality is $D = 2$, the probability density for X_t is defined as

$$f_{\boldsymbol{X}_t|\boldsymbol{W}}(\boldsymbol{x}_t \mid \boldsymbol{w}, \lambda) = \frac{1}{2\pi I_0(\lambda)} e^{\lambda \boldsymbol{w}^T \boldsymbol{x}_t}$$

Here the parameter $\boldsymbol{w}$ is also normalized as $\|\boldsymbol{w}\|^2 = 1$ and indicates the most probable direction for $\boldsymbol{X}$, and $\lambda \geq 0$ is a concentration parameter. For high positive values of λ, the probability density for $\boldsymbol{X}$ is sharply peaked near $\boldsymbol{w}$. If $\lambda = 0$ the probability density is constant for all directions $\boldsymbol{X}_t$. The function $I_\nu(\lambda)$ in the normalization factor is the *modified Bessel function* of the first kind and order ν. This function can be regarded as known, and it is available numerically as `besseli` in Matlab.

The parameter $\boldsymbol{w}$ is regarded as an outcome of a random vector $\boldsymbol{W}$, whereas the concentration parameter λ is regarded as just a fixed unknown constant. You will now apply Bayesian learning for $\boldsymbol{W}$, combined with an EM procedure to estimate λ, using the observed sequence $\underline{\boldsymbol{x}}$.

(a) Assume that an estimate λ_{old} for the concentration parameter is known, and that the prior density for $\boldsymbol{W}$ also has the von Mises-Fisher form, as

$$f_{\boldsymbol{W}}(\boldsymbol{w}) = \frac{1}{2\pi I_0(\alpha_0)} e^{\alpha_0 \boldsymbol{m}_0^T \boldsymbol{w}}$$

Determine the posterior probability density $f_{\boldsymbol{W}|\underline{\boldsymbol{X}}}(\boldsymbol{w} \mid \underline{\boldsymbol{x}}, \lambda_{old})$, and specify its hyper-parameters. (2p)

Solution: The posterior density is

$$\begin{aligned} f_{\boldsymbol{W}|\underline{\boldsymbol{X}}}(\boldsymbol{w} \mid \underline{\boldsymbol{x}}, \lambda_{old}) &\propto f_{\underline{\boldsymbol{X}}|\boldsymbol{W}}(\underline{\boldsymbol{x}} \mid \boldsymbol{w}, \lambda_{old}) f_{\boldsymbol{W}}(\boldsymbol{w}) = \\ &= f_{\boldsymbol{W}}(\boldsymbol{w}) \prod_{t=1}^{T} f_{\boldsymbol{X}_t|\boldsymbol{W}}(\boldsymbol{x}_t \mid \boldsymbol{w}, \lambda_{old}) \propto e^{\alpha_0 \boldsymbol{m}_0^T \boldsymbol{w}} \prod_{t=1}^{T} e^{\lambda_{old} \boldsymbol{w}^T \boldsymbol{x}_t} = e^{\alpha_0 \boldsymbol{m}_0^T \boldsymbol{w} + \lambda_{old} \boldsymbol{w}^T \sum_t \boldsymbol{x}_t} = \\ &= e^{\alpha_0 \boldsymbol{m}_0^T \boldsymbol{w} + \lambda_{old} \left(\sum_t \boldsymbol{x}_t\right)^T \boldsymbol{w}} = e^{\alpha \boldsymbol{m}^T \boldsymbol{w}} \end{aligned} \quad (27)$$

This is obviously just another von Mises-Fisher density function. Knowing that the central parameter $\boldsymbol{m}$ should have norm $\|\boldsymbol{m}\| = 1$, we identify the posterior hyper-parameters as

$$\alpha = \left\| \alpha_0 \boldsymbol{m}_0 + \lambda_{old} \sum_{t=1}^{T} \boldsymbol{x}_t \right\| \quad (28)$$

$$\boldsymbol{m} = \frac{1}{\alpha} \left(\alpha_0 \boldsymbol{m}_0 + \lambda_{old} \sum_{t=1}^{T} \boldsymbol{x}_t \right) \quad (29)$$

Including the known normalization factor for the von Mises-Fisher density, we can summarize the result as

$$f_{\boldsymbol{W}|\underline{\boldsymbol{X}}}(\boldsymbol{w} \mid \underline{\boldsymbol{x}}, \lambda_{old}) = \frac{1}{2\pi I_0(\alpha)} e^{\alpha \boldsymbol{m}^T \boldsymbol{w}} \tag{30}$$

To apply a uniform prior density, we can further simplify the solution by choosing $\alpha_0 = 0$. Then,

$$\bar{\boldsymbol{x}} = \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{x}_t \tag{31}$$

$$\alpha = \lambda_{old} T \left\| \bar{\boldsymbol{x}} \right\| \tag{32}$$

$$\boldsymbol{m} = \frac{\bar{\boldsymbol{x}}}{\left\| \bar{\boldsymbol{x}} \right\|} \tag{33}$$

(b) Using the estimated posterior probability density for $\boldsymbol{W}$ from the previous step, apply the EM approach to find an improved point estimate λ_{new} for the concentration parameter. (3p)
Hint: With the notation used here, the EM approach should maximize the help function

$$Q(\lambda', \lambda_{old}) = E_{\boldsymbol{W}} \left[\ln f_{\boldsymbol{W}, \underline{\boldsymbol{X}}}(\boldsymbol{W}, \underline{\boldsymbol{x}} \mid \lambda') \mid \underline{\boldsymbol{x}}, \lambda_{old} \right]$$

by varying the free parameter λ'. If the EM approach leads to a non-linear equation involving modified Bessel functions, it is sufficient to just formulate the equation. It is not required to find an explicit solution to the equation. The derivative $I_0'(\lambda) = I_1(\lambda)$ is also a known function.

For a random vector $\boldsymbol{W}$ with dimension $D = 2$ and a von Mises-Fisher density function $f_{\boldsymbol{W}}(\boldsymbol{w}) \propto e^{\alpha \boldsymbol{m}^T \boldsymbol{w}}$, the expected value is

$$E\left[\boldsymbol{W}\right] = \frac{I_1(\alpha)}{I_0(\alpha)} \boldsymbol{m}$$

Solution: The complete log-probability needed for the EM help function is

$$\ln f_{\boldsymbol{W}, \underline{\boldsymbol{X}}}(\boldsymbol{w}, \underline{\boldsymbol{x}} \mid \lambda') = \ln f_{\boldsymbol{W}}(\boldsymbol{w}) + \ln f_{\underline{\boldsymbol{X}}|\boldsymbol{W}}(\underline{\boldsymbol{x}} \mid \boldsymbol{w}, \lambda') =$$

$$= -\ln 2\pi I_0(\alpha_0) + \alpha_0 \boldsymbol{m}_0^T \boldsymbol{w} - T \ln 2\pi I_0(\lambda') + \sum_{t=1}^{T} \lambda' \boldsymbol{w}^T \boldsymbol{x}_t \tag{34}$$

For the EM help function we must calculate the expected value across $\boldsymbol{W}$ for this expression, using the previously estimated posterior density $f_{\boldsymbol{W}|\underline{\boldsymbol{X}}}(\boldsymbol{w} \mid \underline{\boldsymbol{x}}, \lambda_{old})$. Thus,

$$Q(\lambda', \lambda_{old}) = E_{\boldsymbol{W}|\underline{\boldsymbol{x}}, \lambda_{old}} \left[\ln f_{\boldsymbol{W}, \underline{\boldsymbol{X}}}(\boldsymbol{W}, \underline{\boldsymbol{x}} \mid \lambda') \mid \underline{\boldsymbol{x}}, \lambda_{old} \right] =$$

$$= \underbrace{-\ln 2\pi I_0(\alpha_0) + \alpha_0 \boldsymbol{m}_0^T E_{\boldsymbol{W}} \left[\boldsymbol{W} \mid \underline{\boldsymbol{x}}, \lambda_{old} \right]}_{C} - T \ln 2\pi I_0(\lambda') + \lambda' E_{\boldsymbol{W}} \left[\boldsymbol{W} \mid \underline{\boldsymbol{x}}, \lambda_{old} \right]^T \sum_{t=1}^{T} \boldsymbol{x}_t \tag{35}$$

We need to consider in detail only the terms that depend on λ'. Using the given hint for the expected value $E_{\boldsymbol{W}}[\boldsymbol{w} \mid \underline{\boldsymbol{x}}, \lambda_{old}]$, we have

$$Q(\lambda', \lambda_{old}) = C - T \ln I_0(\lambda') + \lambda' \frac{I_1(\alpha)}{I_0(\alpha)} \boldsymbol{m}^T \sum_{t=1}^{T} \boldsymbol{x}_t \tag{36}$$

where α and $\boldsymbol{m}$ were determined using the previous concentration parameter λ_{old}. A stationary point where Q might be maximal must satisfy the equation

$$\frac{\partial Q}{\partial \lambda'} = -T \frac{I_0'(\lambda')}{I_0(\lambda')} + \frac{I_1(\alpha)}{I_0(\alpha)} \boldsymbol{m}^T \sum_{t=1}^{T} \boldsymbol{x}_t = 0 \tag{37}$$

Applying the known derivative given in the hint, the EM update equation can be expressed as

$$\frac{I_1(\lambda')}{I_0(\lambda')} = \frac{I_1(\alpha)}{I_0(\alpha)} \boldsymbol{m}^T \left(\frac{1}{T} \sum_{t=1}^{T} \boldsymbol{x}_t \right) \tag{38}$$

Substituting the value of $\boldsymbol{m}$ obtained with the special case of the uniform prior concentration $\alpha_0 = 0$, the EM update equation can also be written as

$$\frac{I_1(\lambda')}{I_0(\lambda')} = \frac{I_1(\alpha)}{I_0(\alpha)} \left\| \bar{\boldsymbol{x}} \right\|, \text{ where } \bar{\boldsymbol{x}} = \frac{1}{T} \sum_t \boldsymbol{x}_t, \text{ and } \alpha = \lambda_{old} T \left\| \bar{\boldsymbol{x}} \right\| \tag{39}$$