UNIVERSITY COLLEGE LONDON

EXAMINATION FOR INTERNAL STUDENTS

MODULE CODE : **COMP0086**

ASSESSMENT PATTERN : **COMP0086A7PD**

MODULE NAME : **COMP0086 - Probabilistic and Unsupervised Learning**

LEVEL: : **Postgraduate**

DATE : **04-May-2022**

TIME : **10:00**

Controlled Condition Exam: 3 Hours exam

You cannot submit your work after the date and time shown on AssessmentUCL – you must ensure to allow sufficient time to upload and hand in your work

This paper is suitable for candidates who attended classes for this module in the following academic year(s):

Year
2021/22

Additional material	N/A
Special instructions	N/A
Exam paper word count	N/A

TURN OVER

[intentionally blank]

PROBABILISTIC AND UNSUPERVISED LEARNING
FINAL EXAMINATION
COMP0086 2021-22

This examination contains SIX questions, of which you should answer FIVE.

Each question is worth 20 marks. If you have time, you may attempt all 6 questions, but only your 5 best scores will count towards your final mark. Make sure you manage your time carefully, read through ALL questions first before attempting any, and don't spend too much time on any one question. Calculators are allowed, but are unlikely to be useful.

Always provide justification and show any intermediate work for your answers. A correct but unsupported answer may not receive any marks.

In the following, bold symbols represent vectors. When asked to provide algorithms, you may provide equations or use MATLAB/Octave notation for matrix algebra. You may assume the availability of standard matrix functions such as `inv`, `pinv`, `eig` or `svd`. You may **not** assume statistical functions, including `mean` and `cov`. If you choose to use MATLAB notation, you will not lose marks for syntactic errors, as long as your intention is clear and correct.

You might find it useful to consult the following table of standard distributions:

Name	Domain	Symbol	Density or Probability fn
Gaussian (Normal)	$\mathbb{R}$	$x \sim \mathcal{N}\left(\mu, \sigma^2\right)$	$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$
Multivariate Normal	$\mathbb{R}^D$	$\mathbf{x} \sim \mathcal{N}\left(\boldsymbol{\mu}, \Sigma\right)$	$\lvert 2\pi\Sigma\rvert^{-1/2} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})}$
Bernoulli	$\{0, 1\}$	$x \sim \mathsf{Bernoulli}(p)$	$p^x(1-p)^{1-x}$
Binomial	$\mathbb{Z}_{0-N}$	$x \sim \mathsf{Binom}(p)$	$\binom{N}{x} p^x (1-p)^{N-x}$
Multinomial	$[\mathbb{Z}_{0-N}]^D$	$\mathbf{x} \sim \mathsf{Multinom}(\mathbf{p})$	$\frac{N!}{x_1!\, x_2! \ldots x_D!} \prod_{d=1}^{D} p_d^{x_d}$
Poisson	$\mathbb{Z}_{0+}$	$x \sim \mathsf{Poisson}(\mu)$	$\mu^x e^{-\mu}/x!$
Beta	$[0, 1]$	$x \sim \mathsf{Beta}(\alpha, \beta)$	$\frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}$
Exponential	$\mathbb{R}_+$	$x \sim \mathsf{Exp}(\lambda)$	$\lambda e^{-\lambda x}$
Gamma	$\mathbb{R}_+$	$x \sim \mathsf{Gamma}(\alpha, \beta)$	$\frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$
Student-t	$\mathbb{R}$	$x \sim \mathsf{Student\text{-}t}(\alpha)$	$\frac{\Gamma\left(\frac{\alpha+1}{2}\right)}{\sqrt{\pi\alpha}\Gamma\left(\frac{\alpha}{2}\right)} \left(1 + \frac{x^2}{\alpha}\right)^{-\frac{\alpha+1}{2}}$
Inverse-Wishart	$\mathbb{S}_+^p$	$X \sim \mathsf{InvWish}(\Psi, \nu)$	$\frac{\lvert\Psi\rvert^{\frac{\nu}{2}}}{2^{\frac{\nu p}{2}}\Gamma_p\left(\frac{\nu}{2}\right)} \lvert X\rvert^{-\frac{\nu+p+1}{2}} e^{-\frac{1}{2}\mathsf{Tr}\left[\Psi X^{-1}\right]}$
Dirichlet	$[0, 1]^D$	$\mathbf{x} \sim \mathsf{Dirichlet}(\boldsymbol{\alpha})$	$\frac{\Gamma\left(\sum_{d=1}^{D} \alpha_d\right)}{\prod_{d=1}^{D} \Gamma(\alpha_d)} \prod_{d=1}^{D} x_d^{\alpha_d - 1}$

1. **Conjugacy in regression**

Consider linear regression for input-output pairs $(\mathbf{x}_n, y_n)$, $n = 1 \dots N$ with weight vector $\mathbf{w}$ and output noise variance σ^2:

$$y_n \sim \mathcal{N}\left(\mathbf{w}^\mathsf{T}\mathbf{x}_n, \sigma^2\right)$$

(a) What is a conjugate prior? *[2 marks]*

(b) State the form of the conjugate prior for $\mathbf{w}$, assuming that σ^2 is known. You do not need to derive the result. *[2 marks]*

Now consider the case where σ^2 is unknown and is estimated along with $\mathbf{w}$. We might choose an inverse-gamma prior on σ^2:

$$p(\sigma^2) = \mathsf{InvGamma}(\sigma^2; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}(\sigma^2)^{-\alpha-1}e^{-\beta/\sigma^2}$$

(c) Find the joint posterior on σ^2 and $\mathbf{w}$ under independent priors on $\mathbf{w}$ (of the form you identified above) and inverse-gamma on σ^2. Your notation may be simpler if you introduce a matrix $X = [\mathbf{x}_1 \dots \mathbf{x}_N]$ and row vector $Y = [y_1 \dots y_N]$. Simplify your expression (discarding constants of proportionality) to isolate the functional dependence on σ^2 and $\mathbf{w}$ as far as possible. *[6 marks]*

Your result to the previous part should have shown that the independent prior defined above is *not* conjugate. The correct conjugate prior is a joint distribution with the form:

$$p(\sigma^2, \mathbf{w}) = \mathsf{InvGamma}(\sigma^2; \alpha, \beta)\, \mathcal{N}\left(\mathbf{w}; \boldsymbol{\mu}, \sigma^2 A^{-1}\right)$$

You may take $\boldsymbol{\mu} = 0$.

(d) Show that this is indeed a conjugate prior, and derive the updates to the parameters α, β, $\boldsymbol{\mu}$ and A induced by the likelihood. *[8 marks]*

(e) In lecture we considered Bayesian learning of parameters within A (for example, the diagonal entries) by maximising the evidence or marginal likelihood obtained by integrating out $\mathbf{w}$ alone, while maximising over σ^2. Do you think the results for these parameters would have been different if we had integrated out both $\mathbf{w}$ and σ^2? You do not need to derive the result – just argue whether it will be the same or different based on your results here. *[2 marks]*

2. **Optimisation**

Consider a Markov chain $X_1, X_2, \ldots$ with states $1, \ldots, K$ and transition matrix $P \in \mathbb{R}^{K \times K}$. The entry in row i and column j is

$$P_{ij} = \text{ probability that } X_{t+1} = i \text{ given that } X_t = j \ ,$$

so the sum of each column is 1. We observe the first n states $X_1 = x_1, \ldots, X_n = x_n$ the chain visits, and denote by N_{ij} the transition counts

$$N_{ij} = \text{ number of times that } x_{t+1} = i \text{ and } x_t = j \text{ for } t < n \ .$$

(a) Suppose $f_1, \ldots, f_K$ are convex functions and $g_1, \ldots, g_K$ linear functions, all from $\mathbb{R}^d \to \mathbb{R}$. Show that the convex optimisation problem

$$\min \sum_{k \leq K} f_k(x_k) \qquad \text{subject to } g_1(x_1) = 0, \ldots, g_K(x_K) = 0 \tag{A}$$

has the same solution(s) as are obtained by solving the K separate problems

$$\min f_k(x_k) \text{ subject to } g_k(x_k) = 0 \qquad \text{for } k = 1, \ldots, K \ . \tag{B}$$

[3 marks]

Recall that a statistical model is a set of probability distributions. Suppose we assume the model

$$M := \{ \text{ all Markov chains with states } 1, \ldots, K \ \} \ .$$

(That means any initial distribution is permitted.) We want to determine the maximum likelihood estimator $\hat{P}$ of P.

(b) Write out the log-likelihood of P as a function of the transition counts. Define a convex (!) optimisation problem with constraints whose solution is $\hat{P}$.
Hint: The logarithm is a concave function. If $\alpha, \alpha' \geq 0$ and h and h' are convex functions, then $\alpha h + \alpha' h'$ is also a convex function. *[6 marks]*

(c) Consider the jth column $\hat{P}_j$ of $\hat{P}$. Specify a convex optimisation problem that has solution $\hat{P}_j$ and involves only the transition counts for the j column, $N_{1j}, \ldots, N_{Kj}$. Show that

$$\hat{P}_{kj} = \frac{N_{kj}}{\sum_{i \leq K} N_{ij}}$$

for each $k \leq K$ solves this problem. *[8 marks]*

Now suppose we are given a probability distribution π on the set $1, \ldots, K$, and we change our model to

$$M_\pi := \{ \text{ all Markov chains with states } 1, \ldots, K \text{ and invariant distribution } \pi \ \} \ .$$

(d) State maximum likelihood estimation in this model as a convex optimisation problem. Can (a) again be used to split the problem into K separate problems? *[3 marks]*

3. **Derivatives in the EM algorithm**

Consider a general latent variable model with observed data $\mathcal{X}$, latent variables $\mathcal{Z}$ and parameters θ. The log-likelihood function is given by

$$\ell(\theta) = \log \int d\mathcal{Z}\ P(\mathcal{X}, \mathcal{Z}|\theta)$$

(a) How is the free energy defined? Show that it is a lower bound on the log likelihood. *[4 marks]*

(b) Show that the (generalised) M-step gradient evaluated after an exact E-step is (under some regularity conditions) equal to the gradient of log likelihood with respect to the parameters. *[6 marks]*

(c) Consider a mixture of spherical Gaussians:

$$p(z_n) = \mathsf{Discrete}(\boldsymbol{\pi})$$
$$p(\mathbf{x}_n|z_n) = \prod_k \left(\mathcal{N}\left(\boldsymbol{\mu}_k, I\right)\right)^{\delta_{z_n,k}}$$

where $z_n \in 1 \ldots K$ and $\delta_{z_n,k}$ is an indicator taking the value 1 if $z_n = k$ and 0 otherwise.

Compute the exact M-step update for the parameter $\boldsymbol{\mu}_k$, as well as the gradient and show that the exact step changes $\boldsymbol{\mu}_k$ along the direction of the gradient. *[6 marks]*

(d) Is this generally true? That is, is the change in parameters resulting from an exact (i.e. non-generalised) M-step always aligned with the gradient direction? Explain. *[4 marks]*

4. **Markov chains and sampling algorithms**

The first few questions below develop a (very primitive version of) a method known as the Propp-Wilson algorithm. The final question is a separate problem.

Consider a finite set $\mathbf{K} = \{1, \ldots, K\}$. Let f_θ be a function $\mathbf{K} \to \mathbf{K}$ parametrized by some value θ. (Simple examples would be a constant function with value $f_\theta(x) = \theta$, for all x and some $\theta \in \mathbf{K}$, or the "discrete ReLU" function $\max\{\theta, x\}$.) We generate parameter values $\Theta_1, \ldots, \Theta_n$ at random, and define

$$F_n(x) := f_{\Theta_1} \circ \cdots \circ f_{\Theta_n}(x) = f_{\Theta_1}(f_{\Theta_2}(\cdots f_{\Theta_n}(x) \cdots))$$

and reversely

$$G_n(x) := f_{\Theta_n} \circ \cdots \circ f_{\Theta_1}(x) \; .$$

(a) We fix any x, and generate $\Theta_1, \ldots, \Theta_n$ either i.i.d., or from a Markov chain, or from a hidden Markov model. In which of these cases is the sequence $(G_1(x), G_2(x), \ldots, G_n(x))$ a Markov chain? *[3 marks]*

(b) Now assume $\Theta_1, \ldots, \Theta_n$ are i.i.d. For some fixed x, do $F_n(x)$ and $G_n(x)$ have the same distribution? Does the sequence $(F_1(x), F_2(x), \ldots)$ have the same distribution as $(G_1(x), G_2(x), \ldots)$? *[3 marks]*

Suppose we can find functions f_θ and a distribution for the Θ_i such that F_n *always* becomes a constant function eventually, that is, there are some random values N in $\mathbb{N}$ and X in $\mathbf{K}$ such that

$$F_N(x) = X$$

for all initial values $x \in \mathbf{K}$. (We avoid details here, but that is indeed possible!)

(c) Let π be the distribution of X. What is the distribution of $F_{N+1}(x)$? *[3 marks]*

(d) Suppose you have access to the random values $\Theta_1, \Theta_2, \ldots$. Specify an algorithm, as a single while loop, that takes input x and outputs a single draw from π. *[3 marks]*

(e) Which method would be preferable to generate a single draw from π: The algorithm in (d), or a Metropolis-Hastings algorithm with invariant distribution π? *[3 marks]*

(f) We want to sample from a distribution with a very simple "staircase" density p (the thick line in the left figure):

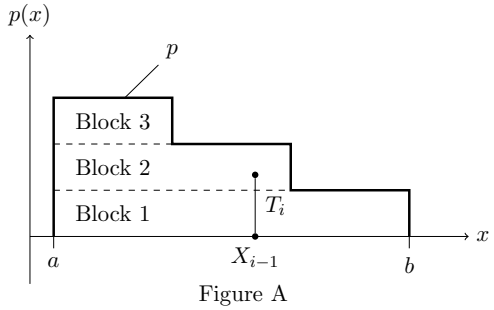

Figure A

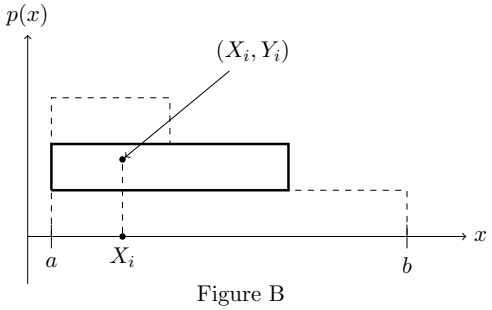

Figure B

We divide the area under p into three blocks, and define a Markov chain that generate samples $x_1, x_2, \ldots$ as follows. Start with any $X_1 \in [a, b]$. For each $i > 1$:

- Draw $T_i \sim \text{Uniform}[0, p(X_{i-1})]$
- Select the block k which contains (X_{i-1}, T_i). (In the example in the figure, this would be block 2.)
- Sample a point (X_i, Y_i) uniformly from the area in block k.
- Discard the vertical coordinate Y_i and keep X_i as the ith sample.

Is this a valid sampler for p? More precisely: If we assume that the previous sample X_{i-1} is distributed according to p, is X_i marginally distributed according to p? *[5 marks]*

5. **The entropy term in the free energy**

In this problem, we consider the role of the entropy term in the free energy, by trying to understand what happens if we neglect it. That is, given data $\{\mathbf{x}_i\}$, we consider a generic mixture model with M components and parameters $\boldsymbol{\pi}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M$:

$$s_i \sim \mathsf{Discrete}(\boldsymbol{\pi}) \qquad s_i \in \{1 \ldots M\};\ \pi_m \geq 0;\ \sum_{m=1}^{M} \pi_m = 1$$
$$\mathbf{x}_i | s_i{=}m \sim P_m(\mathbf{x}_i; \boldsymbol{\theta}_m) \qquad \mathbf{x}_i \in \mathbb{R}^D\,.$$

We introduce posterior responsibilities $r_i^m \geq 0$ with $\sum_{m=1}^{M} r_i^m = 1$ for each observation i, and try to maximize

$$\widetilde{\mathcal{F}}(\{r_i^m\}, \{\boldsymbol{\theta}_m\}, \boldsymbol{\pi}) = \sum_i \sum_m r_i^m \log P(\mathbf{x}_i, s_i{=}m | \boldsymbol{\theta}_m, \boldsymbol{\pi})\,,$$

with respect to $\{r_i^m\}$, $\{\boldsymbol{\theta}_m\}$ and $\boldsymbol{\pi}$.

Constrained optimisation with Lagrange multipliers as used in class is not helpful here: When we try to find the optima with respect to r_i^m, we would obtain derivatives in which the r_i^m do not appear.

(a) We try a geometric approach: Taking the case $M = 2$, sketch the contour lines of $\widetilde{\mathcal{F}}$ as a function of the responsibilities for the first data point, as well as the constraint surface $\sum_m r_1^m = 1$. Argue that the maximum will typically be found when $r_1^m = 1$ for some m. What determines which m? *[5 marks]*

(b) Derive the correct free energy for this model, showing explicitly that it lower bounds the log-likelihood. You should start from the log-likelihood for the model above, and complete the derivation using the specific model notation — **do not** reproduce a generic derivation and substitute the model-specific values at the final step. *[5 marks].*

(c) Show that the incorrect free energy $\widetilde{\mathcal{F}}$ is also a lower bound on the log likelihood. *[3 marks]*

Now consider a continuum of free energies $\mathcal{F}_\tau, \tau \in [0, 1]$ in which the entropy term is premultiplied by τ.

(d) Derive the optimal responsibilities for $\mathcal{F}_\tau$ when $\tau > 0$. Then take the limit $\tau \to 0$ and show that you obtain the same result as you found geometrically. *[5 marks]*

(e) What well-known algorithm will result if you maximise $\widetilde{F}$ for a mixture of Gaussians with fixed identity covariance matrices and unknown means, and fixed mixing proportions $\pi_m = 1/M$. *[2 marks]*

6. **HMM posteriors**.

Consider a hidden Markov model (HMM) for observations $\mathbf{x}_t \in \mathbb{R}^D, t = 1 \dots T$ with:

hidden states	$s_t \in \{1 \dots K\}, t = 1 \dots T,$
initial state distribution	$P(s_1{=}i) = \pi_i,$
transition matrix	$P(s_t{=}j \mid s_{t-1}{=}i) = \Phi_{ij},$
and output distributions	$P(\mathbf{x}_t \mid s_t{=}i) = A_i(\mathbf{x}_t)\,.$

(a) Give the general Bayesian filtering recursion for $P(s_t|\mathbf{x}_1 \dots \mathbf{x}_t)$ in terms of the parameters and $P(s_{t-1}|\mathbf{x}_1 \dots \mathbf{x}_{t-1})$. How does this recursion depend on conditional independence in the HMM? *[4 marks]*

(b) Given forward and backward messages:

$$\alpha_t(i) = P(\mathbf{x}_1, \dots, \mathbf{x}_t, s_t{=}i | \Phi, \{A_j\}, \boldsymbol{\pi})$$
$$\beta_t(i) = P(\mathbf{x}_{t+1}, \dots, \mathbf{x}_T | s_t{=}i, \Phi, \{A_j\})$$

and a prior on the rows of Φ:

$$\Phi_{i,\cdot} \sim \mathsf{Dirichlet}(\boldsymbol{\lambda}) \qquad \boldsymbol{\lambda} \in \mathbb{R}_+^K$$

derive the *maximum a posteriori* (MAP) M-step update for the transition matrix Φ by finding the analytic maximimum of the free energy.

[8 marks]

(c) Now introduce a set of diagonal matrices

$$A_t = \begin{bmatrix} A_1(\mathbf{x}_t) & 0 & 0 & \dots & 0 \\ 0 & A_2(\mathbf{x}_t) & 0 & \dots & 0 \\ 0 & 0 & \ddots & & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ 0 & 0 & 0 & \dots & A_K(\mathbf{x}_t) \end{bmatrix}$$

and derive a closed-form expression for the likelihood as a (long!) matrix-vector product. *[5 marks]*

(d) Based on this expression, write down a formal expression for the posterior distribution on the transition matrix Φ (assuming that all the other parameters are known). Do you think direct maximisation of this posterior is possible? Why (not)? *[3 marks]*