IMPERIAL COLLEGE LONDON

TIMED REMOTE ASSESSMENTS 2021-2022

MEng Honours Degree in Electronic and Information Engineering Part IV
MEng Honours Degree in Mathematics and Computer Science Part IV
MEng Honours Degrees in Computing Part IV
MSc Advanced Computing
MSc Artificial Intelligence
MSc Computing
MSc in Computing (Specialism)
for Internal Students of the Imperial College of Science, Technology and Medicine

This paper is also taken for the relevant assessments for the Associateship of the City and Guilds of London Institute

PAPER COMP70019=COMP97061=COMP97062

PROBABILISTIC INFERENCE

Friday 25 March 2022, 10:00
Writing time: 120 minutes
Upload time: 30 minutes

Answer ALL THREE questions
Open book assessment

This time-limited remote assessment has been designed to be open book. You may use resources which have been identified by the examiner to complete the assessment and are included in the instructions for the examination. You must not use any additional resources when completing this assessment.

The use of the work of another student, past or present, constitutes plagiarism. Giving your work to another student to use constitutes an offence. Collusion is a form of plagiarism and will be treated in a similar manner. This is an individual assessment and thus should be completed solely by you. The College will investigate all instances where an examination or assessment offence is reported or suspected, using plagiarism software, vivas and other tools, and apply appropriate penalties to students. In all examinations we will analyse exam performance against previous performance and against data from previous years and use an evidence-based approach to maintain a fair and robust examination. As with all exams, the best strategy is to read the question carefully and answer as fully as possible, taking account of the time and number of marks available.

Paper contains 3 questions

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

Answer *all 3* questions.

Read the question carefully to ensure you answer the question correctly.

Useful formulae

Probability distributions:

- •Bernoulli $p(x|\mu) = \mu^x(1-\mu)^{1-x}, \quad x \in \{0,1\}$
- •Binomial $p(m|N,\mu) = \binom{N}{m}\mu^m(1-\mu)^{N-m}$
- •Beta $\text{Beta}(\mu|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\mu^{\alpha-1}(1-\mu)^{\beta-1}$
- •Gaussian $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = |2\pi\boldsymbol{\Sigma}|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\intercal\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right), \ \boldsymbol{x} \in \mathbb{R}^D$
- •Gamma $\text{Gamma}(\tau|a,b) = \frac{1}{\Gamma(a)}b^a\tau^{a-1}\exp(-b\tau)$
- •Wishart $\mathcal{W}(\boldsymbol{\Sigma}|\boldsymbol{W},\nu) = B|\boldsymbol{\Sigma}|^{\frac{\nu-D-1}{2}}\exp\left(-\frac{1}{2}\text{tr}(\boldsymbol{W}^{-1}\boldsymbol{\Sigma})\right), \ \boldsymbol{\Sigma} \in \mathbb{R}^{D\times D}$

Other:

- •KL divergence $\text{KL}[p(\mathbf{x})||q(\mathbf{x})] := \int p(\mathbf{x})\log\left(\frac{p(\mathbf{x})}{q(\mathbf{x})}\right)\mathrm{d}\mathbf{x}$
- •Woodbury $(A+UCV)^{-1} = A^{-1} - A^{-1}U\left(C^{-1}+VA^{-1}U\right)^{-1}VA^{-1}$
- •Gaussian conditioning. For a joint Gaussian density

$$p\left(\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix}\right) = \mathcal{N}\left(\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix}; \begin{bmatrix}\mathbf{m_x}\\ \mathbf{m_y}\end{bmatrix}, \begin{bmatrix}\mathbf{\Sigma_{xx}} & \mathbf{\Sigma_{xy}}\\ \mathbf{\Sigma_{yx}} & \mathbf{\Sigma_{yy}}\end{bmatrix}\right), \tag{1}$$

we have the conditional density

$$p(\mathbf{x}\,|\,\mathbf{y}) = \mathcal{N}\Big(\mathbf{x}; \quad \mathbf{m_x} + \mathbf{\Sigma_{xy}}\mathbf{\Sigma_{yy}^{-1}}(\mathbf{y}-\mathbf{m_y}), \quad \mathbf{\Sigma_{xx}} - \mathbf{\Sigma_{xy}}\mathbf{\Sigma_{yy}^{-1}}\mathbf{\Sigma_{yx}}\Big). \tag{2}$$

- •Gaussian CDF $\Phi(x) = \int_{-\infty}^{x}\mathcal{N}(x';0,1)\mathrm{d}x'$. Remember: $\Phi(-x) = 1-\Phi(x)$.

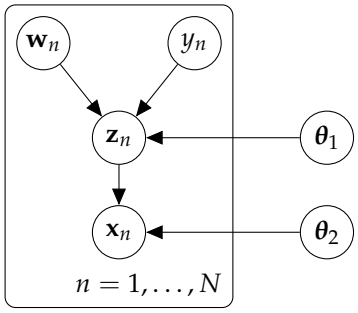
x	-3.0	-2.5	-2.0	-1.5	-1.0	-0.5	0.0
$\Phi(x)$	0.00135	0.00621	0.0227	0.0668	0.159	0.309	0.5

Notation:

- •For a matrix $X \in \mathbb{R}^{N\times D}$ consisting of N vectors in $\mathbb{R}^D$, we use $f(X) \in \mathbb{R}^N$ to denote the function $f : \mathbb{R}^D \to \mathbb{R}$ evaluated at all points in X.
- •Similarly, for a function of two arguments $k : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}$, we use $k(X_1, X_2) \in \mathbb{R}^{N_1\times N_2}$ to denote k evaluated at all pairs of points between $X_1 \in \mathbb{R}^{N_1\times D}$ and $X_2 \in \mathbb{R}^{N_2\times D}$.

1 **Inference & Graphical Models**

Consider the graphical model:



a State the joint density in terms of the conditionals implied by the factorisation expressed in the graphical model.

b Determine whether the following conditional independencies hold. Briefly state through which node paths are open/closed. Half of the marks are assigned for the correct answer, the other half for the reasoning.

i) $\mathbf{x}_1 \perp\!\!\!\perp \mathbf{x}_2$

ii) $\mathbf{x}_1 \perp\!\!\!\perp \mathbf{x}_2 | \boldsymbol{\theta}_1, \boldsymbol{\theta}_2$

iii) $\mathbf{w}_n \perp\!\!\!\perp y_n$

iv) $\mathbf{w}_n \perp\!\!\!\perp y_n | \mathbf{x}_n$

v) $\mathbf{w}_n \perp\!\!\!\perp y_n | \boldsymbol{\theta}_1$

vi) $\mathbf{x}_1 \perp\!\!\!\perp \mathbf{x}_2 | \boldsymbol{\theta}_2$

vii) $\mathbf{x}_1 \perp\!\!\!\perp \mathbf{x}_2 | \boldsymbol{\theta}_2, \mathbf{z}_1$

viii) $\mathbf{x}_n \perp\!\!\!\perp \mathbf{w}_n | \boldsymbol{\theta}_1$

c Write the following distributions in terms of the conditionals defined in the graphical model. We denote the collection $\{\mathbf{x}_n\}_{n=1}^N$ as X, and similar for Z. If the answer to an earlier question is useful, you may use it, rather than going to the conditionals of the graphical models. Factorise as far as possible, and integrate over the smallest dimension possible.

i) $p(\mathbf{z}_n | \boldsymbol{\theta}_1)$

ii) $p(\mathbf{x}_n | \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$

iii) $p(Z)$

iv) $p(X | \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$

v) $p(\mathbf{z}_n | \mathbf{x}_n, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$

The three parts carry, respectively, 10%, 40%, and 50% of the marks.

2 Heteroskedastic noise

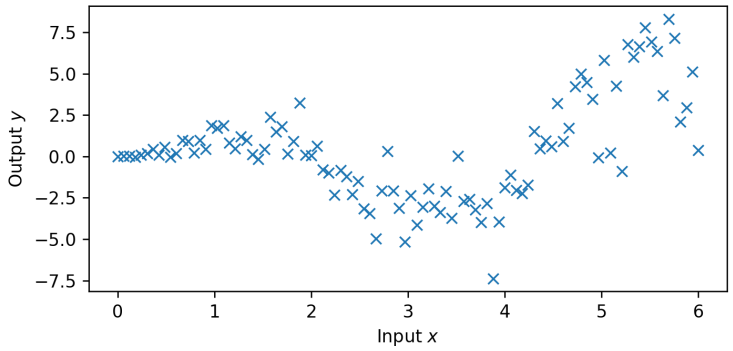


Fig. 1: Regression dataset consisting of input-output pairs (x_n, y_n). Notice how the amount of noise increases to the right.

The standard Gaussian process model consists of a prior on function values and likelihood $p(y_n|f(x_n),\boldsymbol{\theta}) = \mathcal{N}\left(y_n; f(x_n), \sigma^2\right)$. Training data consists of a set of inputs $X = \{x_n\}_{n=1}^N$ and their corresponding outputs $\mathbf{y} = \{y_n\}_{n=1}^N$. We assume a zero mean function and a squared exponential kernel $k_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{x}')$. We collect the hyperparameters of the likelihood and prior into a single vector $\boldsymbol{\theta}$. We do not condition on any regression inputs in our notation.

The predictive distribution for a GP at a new input x^* is:

$$p(y^*|\mathbf{y},\boldsymbol{\theta}) = \mathcal{N}\Big(y^*; \quad k_{\boldsymbol{\theta}}(x^*,X)(k_{\boldsymbol{\theta}}(X,X)+\sigma^2 I_N)^{-1}\mathbf{y}, \\ \sigma^2 + k_{\boldsymbol{\theta}}(x^*,x^*) - k_{\boldsymbol{\theta}}(x^*,X)(k_{\boldsymbol{\theta}}(X,X)+\sigma^2 I_N)^{-1}k_{\boldsymbol{\theta}}(X,x^*)\Big)$$

a For this part, we take the hyperparameters so that $k_{\boldsymbol{\theta}}(x,x) = 1$. In the case where all our datapoints are the same, we have $k_{\boldsymbol{\theta}}(X,X) = \mathbf{1}\mathbf{1}^\intercal$, where $\mathbf{1} \in \mathbb{R}^N$ is filled with ones. We also take x^* to be the same as all training inputs.

i) Using the Woodbury identity, show that in this case the predictive variance is equal to $\sigma^2 + \frac{\sigma^2}{\sigma^2+N}$. When you apply the Woodbury identity, clearly state which matrices you use as A, U, C, and V.

ii) What happens to the predictive variance as $N \to \infty$?

iii) Describe how the predictive distribution $p(y^*|\mathbf{y},\boldsymbol{\theta})$ of such a model would be inadequate for the dataset in Fig. 1, even as the number of data becomes large.

Consider the following regression model:

$$p(\mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\eta}) = \prod_{n=1}^{N} p(y_n|\boldsymbol{\theta}, \boldsymbol{\eta}) p(\boldsymbol{\theta}) p(\boldsymbol{\eta}) \,, \tag{3}$$

$$p(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}; 0, I_M) \,, \tag{4}$$

$$p(\boldsymbol{\eta}) = \mathcal{N}(\boldsymbol{\eta}; 0, I_L) \,, \tag{5}$$

$$p(y_n|\boldsymbol{\theta}, \boldsymbol{\eta}) = \mathcal{N}\left(y_n; \boldsymbol{\phi}(x_n)^\intercal \boldsymbol{\theta}, (\boldsymbol{\psi}(x_n)^\intercal \boldsymbol{\eta})^2\right) \,, \tag{6}$$

where $\boldsymbol{\psi}(x) = [x \quad 1]^\intercal$, and $N = 100$.

b i) Draw the graphical model.

ii) What property of this model makes it more suitable to the dataset in Fig. 1?

iii) Is it possible to find the posterior $p(\boldsymbol{\theta}|\mathbf{y}, \boldsymbol{\eta})$ in closed form? (max 4 sentences)

iv) Is it possible to find the posterior $p(\boldsymbol{\theta}, \boldsymbol{\eta}|\mathbf{y})$ in closed form? (max 4 sentences)

c i) Which quantity would you optimise to find a point estimate for $\boldsymbol{\eta}$? You do not need to calculate an explicit form, but you should comment on whether you can compute it.

ii) Describe why a point estimate for $\boldsymbol{\eta}$ may suffice in this case. (max 3 sentences)

The three parts carry, respectively, 45%, 35%, and 20% of the marks.

3 Variational Inference in Gaussian Processes

a We start with two different joint densities: $q(a,b)$ and $p(a,b)$. Show that

$$\mathrm{KL}[q(a,b)||p(a,b)] = \mathrm{KL}[q(a)||p(a)] \tag{7}$$

if $q(b|a) = p(b|a)$. Use only the definition of the KL divergence and the rules of probability. Make sure you make every step clear (3-5 steps). You may want to briefly state what manipulation you used.

b We have a Gaussian process prior with zero mean and arbitrary kernel $k_{\boldsymbol{\theta}} : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}$. The prior density on function values is $p(f(X), f(Z)|\boldsymbol{\theta})$, where X, Z are sets of inputs of size N and $M \ll N$ respectively, and $\boldsymbol{\theta}$ are kernel hyperparameters. In our notation, we drop conditioning on regression inputs.

We also have a second distribution

$$q(f(X), f(Z)) = \mathcal{N}\left(\begin{bmatrix} f(Z) \\ f(X) \end{bmatrix}; \begin{bmatrix} \mathbf{m} \\ \mathbf{Cm} \end{bmatrix}, \begin{bmatrix} \mathbf{S} & \mathbf{SC}^\intercal \\ \mathbf{CS} & K(X,X) - \mathbf{C}(K(Z,Z) - \mathbf{S})\mathbf{C}^\intercal \end{bmatrix}\right), \tag{8}$$

with $\mathbf{C} = K(X,Z)K(Z,Z)^{-1}$. This distribution depends on parameters which we do not explicitly condition on.

Show that $\mathrm{KL}[q(f(Z), f(X))||p(f(Z), f(X)|\boldsymbol{\theta})] = \mathrm{KL}[q(f(Z))||p(f(Z)|\boldsymbol{\theta})]$.

c We now incorporate our GP prior into a model with **arbitrary** conditionally independent likelihood terms $p(y_n|f(x_n))$. This set-up includes regression ($y_n \in \mathbb{R}$) and classification ($y_n \in \{-1, +1\}$).

Combining the usual derivation with the results from the previous questions leads to a variational inference scheme for Gaussian processes with the following ELBO:

$$\log p(\mathbf{y}|\boldsymbol{\theta}) \geq \mathbb{E}_{q(f(\mathbf{x}_n))}[\log p(y_n|f(\mathbf{x}_n))] - \mathrm{KL}[q(f(Z))||p(f(Z)|\boldsymbol{\theta})]\,,$$

with the property that

$$\log p(\mathbf{y}|\boldsymbol{\theta}) - \mathrm{ELBO} = \mathrm{KL}[q(f(Z), f(X))||p(f(Z), f(X))|\mathbf{y}, \boldsymbol{\theta})]\,.$$

We can only evaluate the likelihood pointwise.

i) Which parameters does the predictive approximate posterior $q(f(\mathbf{x}_n))$ depend on?

ii) Now assume Z is fixed. If we only want to improve the quality of the approximation, which parameters should we optimise? Why? (max 3 sentences)

iii) What type of algorithm would you use to optimise this ELBO? Why? Include both computational and mathematical considerations (max 5 sentences)

iv) State what you need to compute for the optimisation algorithm. State the procedure you would use, and why it is applicable in this case. (max 4 sentences)

The three parts carry, respectively, 25%, 30%, and 45% of the marks.