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

Answer *3 out of 4* questions. You will only receive marks for the first 3 questions attempted, unless clearly marked which 3 should be counted.

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

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

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

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 **Bayesian Optimisation & Decision Theory**

a Answer in $\sim$ 3 sentences each.

i) Describe the exploration-exploitation trade-off.

ii) How do acquisition functions encourage exploration in Bayesian Optimisation? Take Lower Confidence Bound (LCB) as an example.

b Two students live on the same street and need to commute to university. They can either go by **tube** or **bike**. The tube is fast on average and reliable, while the bike can be very fast if all traffic lights are green but is slower on average. The probability distribution for the lengths of the commute in minutes for each transport option can be approximated by Gaussians:

$$p(t_{\text{tube}}) = \mathcal{N}\left(t_{\text{tube}}; 30, 2^2\right), \quad p(t_{\text{bike}}) = \mathcal{N}\left(t_{\text{bike}}; 35, 10^2\right). \tag{3}$$

While Gaussians are unrealistic in the tails, the moments of the true distribution are well-approximated by the Gaussian.

It is 8:35am. Both students minimise expected loss.

i) Steven wants to maximise his expected time in the library, so his loss function is $L(t, a) = t_a$, where the action a can be either "bike" or "tube". What is Steven's preferred commuting method?

ii) Paula has an exam at 9am today. She will not be admitted if she is late, so her loss function is

$$L(t, a) = \begin{cases} 0 & \text{if } t \leq 25 \text{ min} \\ 100 & \text{if } t > 25 \text{ min} \end{cases}. \tag{4}$$

What is Paula's preferred commuting method?

c To increase her chances of making the exam, Paula decides to take a taxi. The taxi driver knows of routes A and B. Route A can either be clear and fast ($Q = 0$) or jammed and slow ($Q = 1$). The time for route B follows a Gaussian.

$$t_{\text{A}} = \begin{cases} 10 \text{ min} & \text{if } Q = 0 \\ 50 \text{ min} & \text{if } Q = 1 \end{cases}, \tag{5}$$

$$P(Q) = \begin{cases} 0.5 & \text{for } Q = 0 \\ 0.5 & \text{for } Q = 1 \end{cases}, \tag{6}$$

$$p(t_{\text{B}}) = \mathcal{N}\left(t_{\text{B}}; 25, 5^2\right). \tag{7}$$

The driver suggests the following strategy:

- Drive the first 2.5 minutes of route A to see whether it is clear (i.e. to observe Q).
- At that point, decide whether to keep going with route A, or to drive back 2.5 minutes and take route B.

i) What is Paula's expected loss for choosing route A and sticking with it, and for choosing route B?

ii) When following the driver's strategy, describe the optimal decisions for when Q is observed to be 0 and 1, and their corresponding expected losses.

iii) Compute the overall expected loss for this strategy. Should this strategy be chosen over simply going for route A or B?

The three parts carry, respectively, 25%, 25%, and 50% of the marks.

2 Non-linear Regression & Gaussian Processes

a You are solving a regression problem with a Gaussian process prior with a squared exponential kernel:

$$k(x, x') = \sigma_f^2 \exp\left(-\frac{(x-x')^2}{2\ell^2}\right) . \tag{8}$$

The information you have about the problem places constraints on your Gaussian process prior on functions. You know that knowledge of $f(x)$ should leave variance of $f(x+3)$ at ≥ 0.95 times the prior variance. However, the variance of $f(x+0.1)$ should be ≤ 0.01 times the prior variance.

i) What hyperparameter does this place bounds on?

ii) Derive the variance of the prior conditional $p(f(z) \,|\, f(x))$.

iii) State the bounds that your prior information places on the hyperparameter.

b The goal of your regression is to predict the amount of recycled plastic that is produced by a process, and deposited in a large vat. Each time a measurement is taken, the vat is emptied. Then the process parameters are adjusted for a new measurement. Due to an error, the vat was not emptied between two parameter settings and only a single measurement of the vat was taken. I.e. a measurement was taken of the sum of two outputs:

$$s = f(\mathbf{x}_1) + f(\mathbf{x}_2) . \tag{9}$$

We do not want to waste this measurement, so we include it in our posterior, which can be written as

$$\begin{aligned} p(f(\mathbf{x}^*) \,|\, f(X), s) = \mathcal{N}\Bigg(f(\mathbf{x}^*);\, &\begin{bmatrix}\mathbf{c}_1^\intercal & c_2\end{bmatrix} \begin{bmatrix}\boldsymbol{K} & \mathbf{k}_1 \\ \mathbf{k}_1^\intercal & k_2\end{bmatrix}^{-1} \begin{bmatrix} f(X) \\ s \end{bmatrix}, \\ &k(\mathbf{x}^*, \mathbf{x}^*) - \begin{bmatrix}\mathbf{c}_1^\intercal & c_2\end{bmatrix} \begin{bmatrix}\boldsymbol{K} & \mathbf{k}_1 \\ \mathbf{k}_1^\intercal & k_2\end{bmatrix}^{-1} \begin{bmatrix} \mathbf{c}_1 \\ c_2 \end{bmatrix} \Bigg) , \end{aligned} \tag{10}$$

where $X \in \mathbb{R}^{N \times D}$ denotes a collection of N input points, $\mathbf{c}_1 \in \mathbb{R}^N$, $\boldsymbol{K} \in \mathbb{R}^{N+1 \times N+1}$, and $c_2 \in \mathbb{R}$. In terms of the *unspecified* prior kernel $k(\mathbf{x}, \mathbf{x}')$, state:

i) $\mathbf{c}_1$, and $\boldsymbol{K}$.

ii) c_2, $\mathbf{k}_1$, and k_2.

You can use the usual notation of evaluating $k(\cdot, \cdot')$ at points arranged in a matrix $X \in \mathbb{R}^{N \times D}$, i.e. $k(X, \mathbf{x}) \in \mathbb{R}^{N \times 1}$ with the ith element being $k(X_{i,:}, \mathbf{x})$, and $k(X, X) \in \mathbb{R}^{N \times N}$ with the i, jth element being $k(X_{i,:}, X_{j,:})$.

c Consider a finite basis function model

$$f(x) = \sum_{i=1}^{B} w_i \exp\left(-(x - c_i)^2\right) = \boldsymbol{\phi}(x)^\intercal \mathbf{w}, \qquad 0 \leq c_i \leq 10, \tag{11}$$

with $p(\mathbf{w}) = \mathcal{N}(\mathbf{w}; 0, \mathbf{I})$. We observe data through the likelihood $p(y_n \,|\, f(\mathbf{x}_n)) = \mathcal{N}(y_n; f(\mathbf{x}_n), \sigma^2)$. Assume $B \leq 10^4$, and $\sigma = 0.1$.

i) Derive an expression for the variance of $f(x)$ under the prior.

ii) Derive an upper bound for the variance of $f(x)$ for $x \geq 20$ in terms of the number of basis functions B.

iii) What is the consequence for predictive variances of $f(x)$ for $x \geq 20$? Why may this lead to poor predictions? (Total no more than 5 sentences.)

iv) Consider performing regression on a function that was zero for $x < 1$ and $x > 9$ (but observed with noise). We observe two points, one at $x = 5$ and one at $x = 20$. Model A is the finite basis function model, while model B has a Gaussian process prior with a Squared Exponential kernel. For $1 \leq x \leq 9$, both models have marginal variances of $f(x)$ that are within 10^{-6} of each other. Which model has a higher marginal likelihood?

The three parts carry, respectively, 30%, 35%, and 35% of the marks.

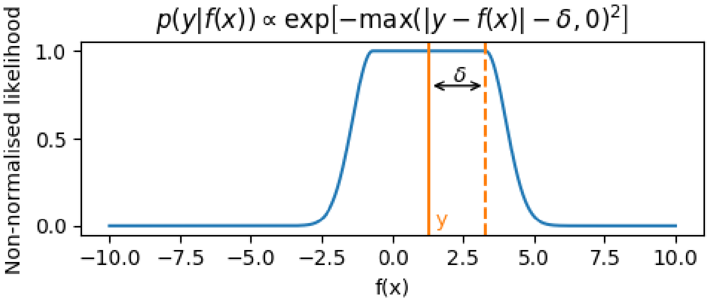

Fig. 1: Likelihood with uniform region and Gaussian tails for some setting of τ and δ.

3 Approximate Inference

Consider a finite basis function model

$$f(\mathbf{x}) = \boldsymbol{\phi}(x)^\intercal \mathbf{w}\,, \qquad\qquad \phi_i = \exp\left(-(\mathbf{x}-\mathbf{c}_i)^2\right), \tag{12}$$

with $p(\mathbf{w}) = \mathcal{N}(\mathbf{w}; 0, \mathbf{I})$.

a The data is observed through a likelihood that is uniform for a region (see fig 1), after which it decays with Gaussian tails:

$$p(y_n \,|\, \mathbf{w}) = Z^{-1} \exp\left[-\frac{\tau}{2}\max(|y_n - f(\mathbf{x}_n)| - \delta, 0)^2\right], \quad 1 \le n \le N. \tag{13}$$

i) Derive a Laplace approximation for this model. You can assume $\mathbf{w}^* = \operatorname{argmax}_{\mathbf{w}} \log p(\mathbf{y} \,|\, \mathbf{w}) + \log p(\mathbf{w})$ is known, and that $\mathbf{w}^*$ lies in the uniform region of the likelihood. State your result as the parameters of the distribution implied by the Laplace approximation.

ii) What is the deficiency in this Laplace approximation, as compared to the true posterior? Pay special attention to the case where the number of observations N grows.

iii) A variational approximation $q(\mathbf{w}) = \mathcal{N}(\mathbf{w}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ with variational parameters $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ is used instead. Would this result in a smaller or larger posterior variance? Explain using the most appropriate form of the variational lower bound. ($\sim$ 6 sentences.)

b Consider now the likelihood to be $p(y_n \,|\, \mathbf{w}) = \mathcal{N}\left(y_n; \boldsymbol{\phi}(\mathbf{x}_n)^\intercal \mathbf{w}, \sigma^2\right)$.

i) What type of distribution would the true posterior be? Give the name of the distribution. And very briefly state why.

ii) Derive a Laplace approximation for the model with the Gaussian likelihood. Derive and state all parameters of the distribution implied by the Laplace approximation.

iii) What is the KL divergence to the true posterior? In this case, would the result of the Laplace approximation also be a good variational approximation?

The two parts carry, respectively, 45% and 55% of the marks.

4 **Variational Autoencoders** When answering this question, choose clear and consistent notation. The most important thing is to emphasise the parameters that quantities depend on. Especially when relevant for gradients.

a i) By applying Jensen's inequality, show that

$$\text{ELBO} = \mathbb{E}_{q(\mathbf{z})}[\log p(\mathbf{x}\,|\,\mathbf{z})] - \text{KL}[q(\mathbf{z})||p(\mathbf{z})] \leq \log p(\mathbf{x}) \tag{14}$$

ii) What is the gap Δ in the lower bound to the marginal likelihood? I.e. derive

$$\Delta = \log p(\mathbf{x}) - \text{ELBO}\,. \tag{15}$$

b i) Write down the integral describing the exact log likelihood for the model that the variational autoencoder performs inference in. Be sure to emphasise (conditional) independencies by factorising where possible. Assume N data points are observed.

ii) What are the factorisation and distributional assumptions made for the variational distribution in the variational autoencoder? Write down the form of $q(\mathbf{z})$, emphasising the (conditional) independencies by factorising where possible. State the name of the density assumed.

iii) If the variational distribution stated above was parameterised directly, how many numbers would need to be optimised?

iv) How does the variational autoencoder avoid scaling the number of variational parameters with the number of observations N? Write down how the variational distribution is parameterised.

c i) For stochastic optimisation to converge to a local minimum, what is the main requirement on the gradient estimator? State the requirement in terms of a gradient estimator $\hat{g}(\boldsymbol{\psi})$ and the objective function $\mathcal{L}(\boldsymbol{\psi})$.

ii) Using the reparameterisation trick, derive estimators for the gradients of the variational and model parameters of the ELBO of the variational autoencoder.

iii) How does the variance of the gradient estimators decrease if more Monte Carlo samples are used?

The three parts carry, respectively, 30%, 30%, and 40% of the marks.

1 **Bayesian Optimisation & Decision Theory**

a Answer in $\sim$ 3 sentences each.

i) Describe the exploration-exploitation trade-off.

ii) How do acquisition functions encourage exploration in Bayesian Optimisation? Take Lower Confidence Bound (LCB) as an example.

i) *When choosing an action to minimise a long-term loss, you need to choose whether to take the action which is currently the best under your beliefs (exploit), or whether to take an action which may teach you a lot, which allows you to make better decisions in the future (explore). This is a trade-off, because often actions that are good in expectation, are ones where you are certain about the outcome, and will therefore not teach you very much.*

ii) *Acquisition functions search trade off evaluating where the GP mean is low, and where the variance (uncertainty) is high. High variance indicates that much information can be found by evaluating there (exploration), whereas low mean tells us that in expectation we will believe to find a good point. LCB does this explicitly by evaluating a weighted sum of the mean and predicted standard deviation.*

Marks: $\overline{\mathbf{5}}$

b Two students live on the same street and need to commute to university. They can either go by **tube** or **bike**. The tube is fast on average and reliable, while the bike can be very fast if all traffic lights are green but is slower on average. The probability distribution for the lengths of the commute in minutes for each transport option can be approximated by Gaussians:

$$p(t_{\text{tube}}) = \mathcal{N}\left(t_{\text{tube}}; 30, 2^2\right), \quad p(t_{\text{bike}}) = \mathcal{N}\left(t_{\text{bike}}; 35, 10^2\right). \tag{1}$$

While Gaussians are unrealistic in the tails, the moments of the true distribution are well-approximated by the Gaussian.

It is 8:35am. Both students minimise expected loss.

i) Steven wants to maximise his expected time in the library, so his loss function is $L(t, a) = t_a$, where the action a can be either "bike" or "tube". What is Steven's preferred commuting method?

ii) Paula has an exam at 9am today. She will not be admitted if she is late, so her loss function is

$$L(t,a) = \begin{cases} 0 & \text{if } t \leq 25 \text{ min} \\ 100 & \text{if } t > 25 \text{ min} \end{cases}. \tag{2}$$

What is Paula's preferred commuting method?

i) $\mathbb{E}_{t_{bike}}[L(t, bike)] = 35$, $\mathbb{E}_{t_{tube}}[L(t, tube)] = 30$. *So by the principle of minimum expected loss, we choose to take the tube.*

ii) *Again, by the principle of minimum expected loss:*

$$\begin{aligned}
\mathbb{E}_{t_{bike}}[L(t, bike)] &= \int_{25}^{\infty} 100 \cdot \mathcal{N}\left(t_{bike}; 30, 10^2\right) \mathrm{d}t_{bike} && (3)\\
&= 100 \cdot \Bigg((1 - \Phi\left(\frac{25-35}{10}\right)\Bigg) && (4)\\
&= 100 \cdot (1 - \Phi(-1)) && (5)\\
\mathbb{E}_{t_{tube}}[L(t, tube)] &= \int_{25}^{\infty} 100 \cdot \mathcal{N}\left(t_{tube}; 30, 2^2\right) \mathrm{d}t_{tube} && (6)\\
&= 100 \cdot \left(1 - \Phi\left(\frac{25-30}{2}\right)\right) && (7)\\
&= 100 \cdot (1 - \Phi(-2.5)) && (8)\\
& && (9)
\end{aligned}$$

We choose the bike.

Marks: **5**

c To increase her chances of making the exam, Paula decides to take a taxi. The taxi driver knows of routes A and B. Route A can either be clear and fast ($Q = 0$) or jammed and slow ($Q = 1$). The time for route B follows a Gaussian.

$$t_{\mathrm{A}} = \begin{cases} 10 \text{ min} & \text{if } Q = 0 \\ 50 \text{ min} & \text{if } Q = 1 \end{cases}, \tag{10}$$

$$P(Q) = \begin{cases} 0.5 & \text{for } Q = 0 \\ 0.5 & \text{for } Q = 1 \end{cases}, \tag{11}$$

$$p(t_{\mathrm{B}}) = \mathcal{N}\left(t_{\mathrm{B}}; 25, 5^2\right). \tag{12}$$

The driver suggests the following strategy:

- Drive the first 2.5 minutes of route A to see whether it is clear (i.e. to observe Q).
- At that point, decide whether to keep going with route A, or to drive back 2.5 minutes and take route B.

i) What is Paula's expected loss for choosing route A and sticking with it, and for choosing route B?

ii) When following the driver's strategy, describe the optimal decisions for when Q is observed to be 0 and 1, and their corresponding expected losses.

iii) Compute the overall expected loss for this strategy. Should this strategy be chosen over simply going for route A or B?

---

i) *Simple expectations again, but results are needed for later.*

$$\mathbb{E}_{t_A}[L(t,A)] = 0.5 \cdot 0 + 0.5 \cdot 100 = 50 \tag{13}$$

$$\begin{aligned}\mathbb{E}_{t_B}[L(t,B)] &= \int_{25}^{\infty} 100\mathcal{N}\left(t_B; 25, 5^2\right) \mathrm{d}t_B \\ &= 1 - \Phi\left(\frac{25-25}{5}\right) = 0.5\end{aligned} \tag{14}$$

ii) *For $Q = 0$:*

$$\mathbb{E}_{t_A \,|\, Q=0}[L(t,A)] = 0 \tag{15}$$

$$\begin{aligned}\mathbb{E}_{t_B \,|\, Q=0}[L(t,B)] &= \int p(t_B) 100 \cdot \mathbb{1}(5 + t_B > 25) \mathrm{d}t_B \\ &= 100 \int_{20}^{\infty} \mathcal{N}\left(t_B; 25, 5^2\right) \mathrm{d}t_B \\ &= 100\left(1 - \Phi\left(\frac{20-25}{5}\right)\right) \\ &= 100(1 - \Phi(-1)) = 84\end{aligned} \tag{16}$$

So the optimal choice is to continue on route A.

For $Q = 1$:

$$\mathbb{E}_{t_A \,|\, Q=1}[L(t,A)] = 100 \tag{17}$$

$$\mathbb{E}_{t_B \,|\, Q=1}[L(t,B)] = 84 \tag{18}$$

iii) *Overall expected loss is* $\mathbb{E}_Q\left[\mathbb{E}_{t|Q}[L(t,d)]\right] = 0.5 \cdot 0 + 0.5 \cdot 84 = 42$, *since in the inner expectatio we take the decision which minimises the loss at that point. This was calculated in the previous part of the question.*

Marks: **10**

The three parts carry, respectively, 25%, 25%, and 50% of the marks.

2 Non-linear Regression & Gaussian Processes

a You are solving a regression problem with a Gaussian process prior with a squared exponential kernel:

$$k(x, x') = \sigma_f^2 \exp\left(-\frac{(x-x')^2}{2\ell^2}\right). \tag{19}$$

The information you have about the problem places constraints on your Gaussian process prior on functions. You know that knowledge of $f(x)$ should leave variance of $f(x+3)$ at ≥ 0.95 times the prior variance. However, the variance of $f(x+0.1)$ should be ≤ 0.01 times the prior variance.

i) What hyperparameter does this place bounds on?

ii) Derive the variance of the prior conditional $p(f(z) \mid f(x))$.

iii) State the bounds that your prior information places on the hyperparameter.

i) *The lengthscale. We are considering* relative *reductions of variance based on input distance. The lengthscale is the only parameter that does this.*

ii) *Can be obtained from the Gaussian conditioning formula:*

$$\begin{aligned}\mathbb{V}[f(z) \mid f(x)] &= k(z,z) - k(z,x)k(x,x)^{-1}k(x,z) \\ &= \sigma_f^2\left(1 - \exp\left(-\frac{(x-z)^2}{\ell^2}\right)\right). \end{aligned} \tag{20}$$

iii) *First find the point of equality:*

$$\frac{\mathbb{V}[f(z) \mid f(x)]}{\mathbb{V}[f(z)]} = c \tag{21}$$

$$1 - \exp\left(-\frac{(x-z)^2}{\ell^2}\right) = c \tag{22}$$

$$\ell = d \cdot \log^{-\frac{1}{2}} \frac{1}{1-c} \tag{23}$$

Condition 1: $c = 0.95\,, x - z = 3$. *Need a short lengthscale to ensure variance ratio is* $\geq c$, *so* $\ell \leq 1.733$.

Condition 2: $c = 0.01\,, x - z = 0.1$. *Need a long lengthscale to ensure variance ratio is* $\leq c$, *so* $\ell \geq 0.9975$.

Marks: $\overline{\mathbf{6}}$

b The goal of your regression is to predict the amount of recycled plastic that is produced by a process, and deposited in a large vat. Each time a measurement is taken, the vat is emptied. Then the process parameters are adjusted for a new measurement. Due to an error, the vat was not emptied between two parameter settings and only a single measurement of the vat was taken. I.e. a measurement was taken of the sum of two outputs:

$$s = f(\mathbf{x}_1) + f(\mathbf{x}_2)\,. \tag{24}$$

We do not want to waste this measurement, so we include it in our posterior, which can be written as

$$\begin{aligned} p(f(\mathbf{x}^*)\,|\,f(X), s) = \mathcal{N}\Bigg(& f(\mathbf{x}^*); \begin{bmatrix}\mathbf{c}_1^\intercal & c_2\end{bmatrix} \begin{bmatrix}\boldsymbol{K} & \mathbf{k}_1 \\ \mathbf{k}_1^\intercal & k_2\end{bmatrix}^{-1} \begin{bmatrix} f(X) \\ s\end{bmatrix}, \\ & k(\mathbf{x}^*, \mathbf{x}^*) - \begin{bmatrix}\mathbf{c}_1^\intercal & c_2\end{bmatrix} \begin{bmatrix}\boldsymbol{K} & \mathbf{k}_1 \\ \mathbf{k}_1^\intercal & k_2\end{bmatrix}^{-1} \begin{bmatrix}\mathbf{c}_1 \\ c_2\end{bmatrix}\Bigg), \end{aligned} \tag{25}$$

where $X \in \mathbb{R}^{N\times D}$ denotes a collection of N input points, $\mathbf{c}_1 \in \mathbb{R}^N$, $\boldsymbol{K} \in \mathbb{R}^{N+1\times N+1}$, and $c_2 \in \mathbb{R}$. In terms of the *unspecified* prior kernel $k(\mathbf{x}, \mathbf{x}')$, state:

i) $\mathbf{c}_1$, and $\boldsymbol{K}$.

ii) c_2, $\mathbf{k}_1$, and k_2.

You can use the usual notation of evaluating $k(\cdot, \cdot')$ at points arranged in a matrix $X \in \mathbb{R}^{N\times D}$, i.e. $k(X, \mathbf{x}) \in \mathbb{R}^{N\times 1}$ with the ith element being $k(X_{i,:}, \mathbf{x})$, and $k(X, X) \in \mathbb{R}^{N\times N}$ with the i,jth element being $k(X_{i,:}, X_{j,:})$.

This question regards Gaussian conditioning. First write down the full joint distribution:

$$p(f(\mathbf{x}^*), f(X), s) = \mathcal{N}\left(\begin{bmatrix} f(\mathbf{x}^*) \\ f(X) \\ s\end{bmatrix}; \begin{bmatrix}0\\0\\0\end{bmatrix}, \begin{bmatrix} k(\mathbf{x}^*, \mathbf{x}^*) & k(\mathbf{x}^*, X) & c_2 \\ k(\mathbf{x}^*, X) & k(X, X) & \mathbf{k}_1 \\ c_2 & \mathbf{k}_1^T & k_2\end{bmatrix}\right) \tag{26}$$

$$\begin{aligned} \mathbf{k}_1 &= \mathbb{C}[f(X), s] = \mathbb{C}[f(X), f(\mathbf{x}_1) + f(\mathbf{x}_2)] \\ &= \mathbb{E}_f[f(X)(f(\mathbf{x}_1) + f(\mathbf{x}_2))] = k(X, \mathbf{x}_1) + k(X, \mathbf{x}_2) && (27) \\ c_2 &= \mathbb{C}[f(\mathbf{x}^*), s] = \mathbb{C}[f(\mathbf{x}^*), f(\mathbf{x}_1) + f(\mathbf{x}_2)] \\ &= k(\mathbf{x}^*, \mathbf{x}_1) + k(\mathbf{x}^*, \mathbf{x}_2) && (28) \\ k_2 &= \mathbb{V}[s] = k(\mathbf{x}_1, \mathbf{x}_1) + 2k(\mathbf{x}_1, \mathbf{x}_2) + k(\mathbf{x}_2, \mathbf{x}_2) && (29) \\ \boldsymbol{K} &= k(X, X) && (30) \end{aligned}$$

Marks: **7**

c Consider a finite basis function model

$$f(x) = \sum_{i=1}^{B} w_i \exp\left(-(x - c_i)^2\right) = \boldsymbol{\phi}(x)^{\mathsf{T}}\mathbf{w}, \qquad 0 \le c_i \le 10, \tag{31}$$

with $p(\mathbf{w}) = \mathcal{N}(\mathbf{w}; 0, \mathbf{I})$. We observe data through the likelihood $p(y_n \,|\, f(\mathbf{x}_n)) = \mathcal{N}\left(y_n; f(\mathbf{x}_n), \sigma^2\right)$. Assume $B \le 10^4$, and $\sigma = 0.1$.

i) Derive an expression for the variance of $f(x)$ under the prior.

ii) Derive an upper bound for the variance of $f(x)$ for $x \ge 20$ in terms of the number of basis functions B.

iii) What is the consequence for predictive variances of $f(x)$ for $x \ge 20$? Why may this lead to poor predictions? (Total no more than 5 sentences.)

iv) Consider performing regression on a function that was zero for $x < 1$ and $x > 9$ (but observed with noise). We observe two points, one at $x = 5$ and one at $x = 20$. Model A is the finite basis function model, while model B has a Gaussian process prior with a Squared Exponential kernel. For $1 \le x \le 9$, both models have marginal variances of $f(x)$ that are within 10^{-6} of each other. Which model has a higher marginal likelihood?

i) *From definition of variance*

$$\begin{aligned}\mathbb{V}[f(x)] &= \mathbb{E}_{\mathbf{w}}[\boldsymbol{\phi}(x)^{\mathsf{T}}\mathbf{w}\mathbf{w}^{\mathsf{T}}\boldsymbol{\phi}(x)] = \boldsymbol{\phi}(x)^{\mathsf{T}}\boldsymbol{\phi}(x) \\ &= \sum_{i=1}^{B}\sum_{j=1}^{B} \exp\left(-(x - c_i)^2\right)\exp\left(-(x - c_j)^2\right)\end{aligned} \tag{32}$$

ii) *Upper bound comes from all basis functions having $c_i = 10$, and measuring the variance at $x = 20$.*

$$\begin{aligned}\mathbb{V}[f(x)] &\le B^2 \exp(-(20 - 10)^2) \qquad \textit{for } x \ge 20 \\ &= B^2 \cdot 3.72 \cdot 10^{-44}\end{aligned} \tag{33}$$

iii) *Variances decrease in the posterior. If we predict in the region $x \ge 20$, we will be very confident that the prediction is very close to zero, due to the small posterior predictive variance. There are many functions for which this is not the case.*

iv) *Model A has the higher marginal likelihood. The marginal likelihood can be decomposed as $p(y_1)p(y_2 \,|\, y_1)$. If we take y_1 to be the output of point $x = 5$, the first term in the marginal likelihood will be roughly the same for both models. Model A however, will have a much larger likelihood for the second point, as it predicts zero with a small variance.*

Marks: **$\bar{7}$**

The three parts carry, respectively, 30%, 35%, and 35% of the marks.

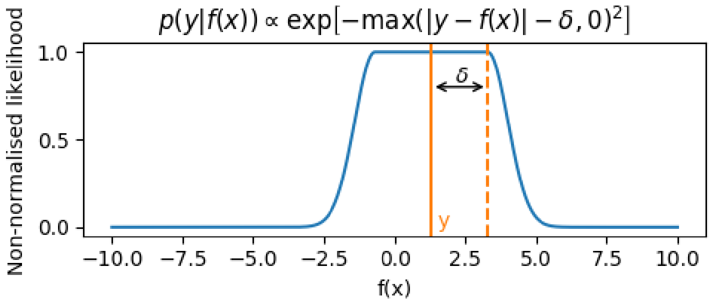

Fig. 1: Likelihood with uniform region and Gaussian tails for some setting of τ and δ.

3 Approximate Inference

Consider a finite basis function model

$$f(\mathbf{x}) = \boldsymbol{\phi}(x)^\intercal \mathbf{w}\,, \qquad \phi_i = \exp\left(-(\mathbf{x}-\mathbf{c}_i)^2\right)\,, \tag{34}$$

with $p(\mathbf{w}) = \mathcal{N}(\mathbf{w}; 0, \mathbf{I})$.

a The data is observed through a likelihood that is uniform for a region (see fig 1), after which it decays with Gaussian tails:

$$p(y_n \,|\, \mathbf{w}) = Z^{-1} \exp\left[-\frac{\tau}{2}\max(|y_n - f(\mathbf{x}_n)| - \delta, 0)^2\right], \quad 1 \leq n \leq N\,. \tag{35}$$

i) Derive a Laplace approximation for this model. You can assume $\mathbf{w}^* = \operatorname{argmax}_{\mathbf{w}} \log p(\mathbf{y} \,|\, \mathbf{w}) + \log p(\mathbf{w})$ is known, and that $\mathbf{w}^*$ lies in the uniform region of the likelihood. State your result as the parameters of the distribution implied by the Laplace approximation.

ii) What is the deficiency in this Laplace approximation, as compared to the true posterior? Pay special attention to the case where the number of observations N grows.

iii) A variational approximation $q(\mathbf{w}) = \mathcal{N}(\mathbf{w}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ with variational parameters $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ is used instead. Would this result in a smaller or larger posterior variance? Explain using the most appropriate form of the variational lower bound. ($\sim$ 6 sentences.)

i) *Laplace approximation:*

$$\log p(\mathbf{w}\,|\,\mathbf{y}) \approx \log q(\mathbf{w}) = -\log Z + \left[\frac{\partial}{\partial \mathbf{w}} \log p(\mathbf{w}\,|\,\mathbf{y})\right]_{\mathbf{w}=\mathbf{w}^*} (\mathbf{w}-\mathbf{w}^*) +$$

$$(\mathbf{w}-\mathbf{w}^*)^T \left[\frac{\partial^2}{\partial \mathbf{w}^2} \log p(\mathbf{w}\,|\,\mathbf{y})\right]_{\mathbf{w}=\mathbf{w}^*} (\mathbf{w}-\mathbf{w}^*) \tag{36}$$

$$\left[\frac{\partial}{\partial \mathbf{w}} \log p(\mathbf{w}\,|\,\mathbf{y})\right]_{\mathbf{w}=\mathbf{w}^*} = 0 \tag{37}$$

$$(\mathbf{w}-\mathbf{w}^*)^T \left[\frac{\partial^2}{\partial \mathbf{w}^2} \log p(\mathbf{w}\,|\,\mathbf{y})\right]_{\mathbf{w}=\mathbf{w}^*} (\mathbf{w}-\mathbf{w}^*) = -\boldsymbol{I} \tag{38}$$

$$\implies q(\mathbf{w}) = \mathcal{N}(\mathbf{w}; \mathbf{w}^*, \boldsymbol{I}) \tag{39}$$

ii) *The approximate posterior variance is always the same as the prior variance! We never become more certain.*

iii) *The variational bound can be stated as*

$$\mathcal{L} = const - \mathrm{KL}[q(\mathbf{w})||p(\mathbf{w}\,|\,\mathbf{y})]\,. \tag{40}$$

The KL divergence heavily penalises placing mass in $q(\mathbf{w})$ where there is none in $p(\mathbf{w}\,|\,\mathbf{y})$. The Laplace approximation has the same variance as the prior, but with a different mean, so it definitely places mass where $p(\mathbf{w}\,|\,\mathbf{y})$ does not. Variational inference would avoid this by reducing the variance.

Marks: **9**

b Consider now the likelihood to be $p(y_n\,|\,\mathbf{w}) = \mathcal{N}\left(y_n; \boldsymbol{\phi}(\mathbf{x}_n)^\intercal \mathbf{w}, \sigma^2\right)$.

i) What type of distribution would the true posterior be? Give the name of the distribution. And very briefly state why.

ii) Derive a Laplace approximation for the model with the Gaussian likelihood. Derive and state all parameters of the distribution implied by the Laplace approximation.

iii) What is the KL divergence to the true posterior? In this case, would the result of the Laplace approximation also be a good variational approximation?

i) *A Gaussian distribution.*

ii) *Equate first derivative to zero, then find the quadratic terms.*

$$
\begin{aligned}
\frac{\partial}{\partial \mathbf{w}} \log p(\mathbf{w} \mid \mathbf{y}) = \frac{\partial}{\partial \mathbf{w}} \Big[-\frac{1}{2\sigma^2} \sum_{n=1}^{N} (y_n - 2\boldsymbol{\phi}(x_n)^{\intercal} \mathbf{w} y_n + \\
\mathbf{w}^{\intercal} \boldsymbol{\phi}(x_n) \boldsymbol{\phi}(x_n)^{\intercal} \mathbf{w}) - \frac{1}{2} \mathbf{w}^{\intercal} \mathbf{w} \Big] = 0
\end{aligned}
\tag{41}
$$

$$
\implies \mathbf{w}^* = \left[\sum_{n=1}^{N} \boldsymbol{\phi}(\mathbf{x}_n) \boldsymbol{\phi}(\mathbf{x}_n)^{\intercal} + \sigma^2 \boldsymbol{I} \right]^{-1} \sum_{n=1}^{N} \boldsymbol{\phi}(\mathbf{x}_n) y_n \tag{42}
$$

$$
-\boldsymbol{\Sigma}^{-1} = -\frac{1}{2\sigma^2} \sum_{n=1}^{N} 2\boldsymbol{\phi}(\mathbf{x}_n) \boldsymbol{\phi}(\mathbf{x}_n)^{\intercal} - \boldsymbol{I} \tag{43}
$$

$$
\boldsymbol{\Sigma} = \left[\frac{1}{\sigma^2} \sum_{n=1}^{N} \boldsymbol{\phi}(x_n) \boldsymbol{\phi}(x_n)^{\intercal} + \boldsymbol{I} \right]^{-1} \tag{44}
$$

$$
q(\mathbf{w}) = \mathcal{N}(\mathbf{w}; \mathbf{w}^*, \boldsymbol{\Sigma}) \tag{45}
$$

iii) *The KL divergence to the true posterior is zero, meaning the Laplace approximation is exact! Since the KL is measured by the variational approximation, the result of the Laplace approximation would also be excellent by variational standards.*

Marks: **11**

The two parts carry, respectively, 45% and 55% of the marks.

4 **Variational Autoencoders** When answering this question, choose clear and consistent notation. The most important thing is to emphasise the parameters that quantities depend on. Especially when relevant for gradients.

a i) By applying Jensen's inequality, show that

$$\text{ELBO} = \mathbb{E}_{q(\mathbf{z})}[\log p(\mathbf{x} \,|\, \mathbf{z})] - \text{KL}[q(\mathbf{z})||p(\mathbf{z})] \leq \log p(\mathbf{x}) \tag{46}$$

ii) What is the gap Δ in the lower bound to the marginal likelihood? I.e. derive

$$\Delta = \log p(\mathbf{x}) - \text{ELBO}\,. \tag{47}$$

i) *Standard.*

$$\begin{aligned}
\log p(\mathbf{x}) &= \log \int p(\mathbf{x} \,|\, \mathbf{z}) p(\mathbf{z}) \mathrm{d}\mathbf{z} && (48)\\
&= \log \int p(\mathbf{x} \,|\, \mathbf{z}) p(\mathbf{z}) \frac{q(\mathbf{z})}{q(\mathbf{z})} \mathrm{d}\mathbf{z} && (49)\\
&= \log \mathbb{E}_{q(\mathbf{z})}\left[\frac{p(\mathbf{x} \,|\, \mathbf{z}) p(\mathbf{z})}{q(\mathbf{z})}\right] && (50)\\
&\geq \mathbb{E}_{q(\mathbf{z})}\left[\log \frac{p(\mathbf{x} \,|\, \mathbf{z}) p(\mathbf{z})}{q(\mathbf{z})}\right] && (51)\\
&= \mathbb{E}_{q(\mathbf{z})}[\log p(\mathbf{x} \,|\, \mathbf{z})] - \text{KL}[q(\mathbf{z})||p(\mathbf{z})] && (52)
\end{aligned}$$

ii) *Perform subtraction:*

$$\begin{aligned}
\Delta &= \log p(\mathbf{x}) - \mathbb{E}_{q(\mathbf{x})}\left[\log \frac{p(\mathbf{x}, \mathbf{z})}{q(\mathbf{z})}\right] && (53)\\
&= -\mathbb{E}_{q(\mathbf{x})}\left[\log \frac{p(\mathbf{x}, \mathbf{z})}{q(\mathbf{z}) p(\mathbf{x})}\right] && (54)\\
&= -\mathbb{E}_{q(\mathbf{x})}\left[\log \frac{p(\mathbf{z} \,|\, \mathbf{x})}{q(\mathbf{z})}\right] && (55)\\
&= \text{KL}[q(\mathbf{z})||p(\mathbf{z} \,|\, \mathbf{x})] && (56)
\end{aligned}$$

Marks: **6**

b i) Write down the integral describing the exact log likelihood for the model that the variational autoencoder performs inference in. Be sure to emphasise (conditional) independencies by factorising where possible. Assume N data points are observed.

ii) What are the factorisation and distributional assumptions made for the variational distribution in the variational autoencoder? Write down the form of $q(\mathbf{z})$, emphasising the (conditional) independencies by factorising where possible. State the name of the density assumed.

iii) If the variational distribution stated above was parameterised directly, how many numbers would need to be optimised?

iv) How does the variational autoencoder avoid scaling the number of variational parameters with the number of observations N? Write down how the variational distribution is parameterised.

i) *Result can be written down by inspection.*

$$\begin{aligned} p(\mathbf{x}) &= \int p(\mathbf{x} \,|\, \mathbf{z}) p(\mathbf{z}) \mathrm{d}\mathbf{z} \tag{57} \\ &= \int \prod_{n=1}^{N} p(\mathbf{x}_n \,|\, \mathbf{z}_n) p(\mathbf{z}_n) \mathrm{d}\mathbf{z}_n \tag{58} \\ &= \prod_{n=1}^{N} \int p(\mathbf{x}_n \,|\, \mathbf{z}_n) p(\mathbf{z}_n) \mathrm{d}\mathbf{z}_n \tag{59} \\ & \tag{60} \end{aligned}$$

ii) $q(\mathbf{z}) = \prod_{n=1}^{N} \mathcal{N}(\mathbf{z}_n; \boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)$.

iii) $ND + ND(D+1)/2$

iv) *A single neural network is used to parameterise a mapping from an observation* $\mathbf{x}_n$ *to the mean* $\boldsymbol{\mu}_n$ *and covariance* $\boldsymbol{\Sigma}_n$ *of the variational distribution. This neural network is shared between all data points.*

Marks: $\overline{\mathbf{6}}$

c i) For stochastic optimisation to converge to a local minimum, what is the main requirement on the gradient estimator? State the requirement in terms of a gradient estimator $\hat{g}(\boldsymbol{\psi})$ and the objective function $\mathcal{L}(\boldsymbol{\psi})$.

ii) Using the reparameterisation trick, derive estimators for the gradients of the variational and model parameters of the ELBO of the variational autoencoder.

iii) How does the variance of the gradient estimators decrease if more Monte Carlo samples are used?

i) *The gradient estimator has to be* unbiased.

$$\mathbb{E}[\hat{g}(\boldsymbol{\psi})] = \nabla_{\boldsymbol{\psi}} \mathcal{L}(\boldsymbol{\psi}) \tag{61}$$

ii) *Start by expressing the reparameterised ELBO, then take derivatives.*

$$\begin{aligned}
\mathcal{L} &= \sum_{n=1}^{N} \mathbb{E}_{q(\mathbf{z}_n)}[\log p(\mathbf{x}_n \mid \mathbf{z}_n)] - \underbrace{\mathrm{KL}[q(\mathbf{z}_n)||p(\mathbf{z}_n)]}_{\textit{closed form}} && (62)\\
&= \sum_{n=1}^{N} \mathbb{E}_{q(\boldsymbol{\epsilon})}\left[\log p(\mathbf{x}_n \mid \mathbf{z}_n = h_{\boldsymbol{\phi}}(\mathbf{x}, \boldsymbol{\epsilon}))\right] - \mathrm{KL}[q(\mathbf{z}_n)||p(\mathbf{z}_n)] && (63)\\
\hat{g}_{\boldsymbol{\phi}} &= \nabla_{\boldsymbol{\phi}} \log p(\mathbf{x}_n \mid \mathbf{z}_n = h_{\boldsymbol{\phi}}(\mathbf{x}, \boldsymbol{\epsilon})) - \nabla_{\boldsymbol{\phi}} \mathrm{KL}[q(\mathbf{z}_n)||p(\mathbf{z}_n)] && (64)\\
&= \nabla_{\mathbf{z}} \log \mathcal{N}\left(\mathbf{x}_n; f(\mathbf{z}), \sigma^2 \boldsymbol{I}\right)|_{\mathbf{z}=h_{\boldsymbol{\phi}}(\mathbf{x},\boldsymbol{\epsilon})} \nabla_{\boldsymbol{\phi}} h(\mathbf{x}_n, \boldsymbol{\epsilon}) \\
&\qquad - \nabla_{\boldsymbol{\phi}} \mathrm{KL}[q(\mathbf{z}_n)||p(\mathbf{z}_n)] && (65)\\
\hat{g}_{\boldsymbol{\psi}} &= \nabla_{\boldsymbol{\psi}} \log \mathcal{N}\left(\mathbf{x}_n; f_{\boldsymbol{\psi}}(h(\mathbf{x}_n, \boldsymbol{\epsilon})), \sigma^2 \boldsymbol{I}\right) && (66)
\end{aligned}$$

Marks: $\overline{\mathbf{8}}$

The three parts carry, respectively, 30%, 30%, and 40% of the marks.