70019 Probabilistic Inference: Study Guide

Mark van der Wilk*
m.vdwilk@imperial.ac.uk

January 30, 2023

1 Overview

This document will provide:

- references to some prerequisites and useful background material,
- clarification on notation conventions used in the course,
- some useful identities (some of these assumed knowledge in the exam),
- exercises and their answers.

This document will be continuously updated during the course with more exercises, pointers and identities. Please e-mail me about any errors or comments.

2 Background material

You will be expected to have a *firm* understanding of Mathematics for Machine Learning. In the explanations, I will be manipulating probabilities and expectations freely, as discussed in Mathematics for Machine Learning. If steps are difficult, I encourage you to raise this on the course EdStem page, or during a Q&A session.

- Basic probability: sample spaces, disjoint events (summation of probabilities), independent events (multiplication of probabilities). See Walpole and Myers [2012], Ch2 (Imperial Library, or search Google for reading options).
- Probability densities. See Deisenroth et al. [2020] §6.2.
- Sum, product & Bayes' rules. See Deisenroth et al. [2020] §6.3.
- Unconstrained continuous optimisation. See Deisenroth et al. [2020] §7.1.
- Linear algebra and matrix decompositions. See Deisenroth et al. [2020] ch 4 (and Chs 2 and 3 for basics).
- A familiarity with linear basis-function regression. See Deisenroth et al. [2020] ch 9.

3 Notation of probabilities

In this course we will use the notation for probabilities that is common in machine learning. The main advantage is that this notation is shorter, although it does leave certain things implicit.

- We generally denote outcomes of random variables without referring explicitly to the random variable itself. For example, when we refer to an outcome $\mathbf{x}$, we implicitly know there is a random variable that can take this value. We usually denote this as the capital, for example here X.

*Many thanks to teaching assistants Anish Dhir, Seth Nabarro, and Filippo Valdettaro for their improvements to the document.

- If $\mathbf{x} \in \mathbb{N}^D \implies p(\mathbf{x})$ is a probability mass function, i.e. $p(\mathbf{x}) = P(X = \mathbf{x})$, i.e. the probability of the random variable X takes value $\mathbf{x}$.
- If $\mathbf{x} \in \mathbb{R}^D \implies p(\mathbf{x})$ is a density. So in this case, $P(a < \mathbf{X} < b) = \int_a^b p(\mathbf{x})\mathrm{d}\mathbf{x}$.
- If I want to be explicit about the random variable that we are evaluating the density/mass of, I will write e.g. $p_{X,Y}(\mathbf{x}, \mathbf{y}) = p_{X|Y}(\mathbf{x}|\mathbf{y})p_Y(\mathbf{y})$.
- Expectations are denoted as $\mathbb{E}_{p(\mathbf{x})}[f(\mathbf{x})] = \int p(\mathbf{x})f(\mathbf{x})\mathrm{d}\mathbf{x}$, or with respect to a different density $\mathbb{E}_{q(\mathbf{x}))}[h(\mathbf{x})] = \int q(\mathbf{x})h(\mathbf{x})\mathrm{d}\mathbf{x}$.
- Often, densities and pmfs can be discussed in exactly the same way, if we think of the density of a discrete RV as a sum of delta functions. I.e. $p(\mathbf{x}) = \sum_o \delta(\mathbf{x} - \mathbf{x}_o)p_o$, where $\{\mathbf{x}_o\}$ are all the possible outcomes that X can take, and p_o are their corresponding probabilities. This allows us to write an expectation as an integral, regardless of whether the RV is continuous or discrete, because for discrete RVs we get:

$$\mathbb{E}_{p(\mathbf{x})}[f(\mathbf{x})] = \int p(\mathbf{x})f(\mathbf{x})\mathrm{d}\mathbf{x} = \int \sum_o \delta(\mathbf{x} - \mathbf{x}_o)p_o f(\mathbf{x})\mathrm{d}\mathbf{x} = \sum_o f(\mathbf{x}_o)p_o \,. \tag{1}$$

- There are some reasons to cringe at this notation, but it is the norm, and it's actually very convenient in many ways.

4 Mathematical identities

4.1 Assumed knowledge

These identities are useful and very common. They are also easy to derive from first principles, and so will be assumed knowledge in the course. I would recommend to be fluent in deriving these.

4.1.1 Properties of covariances

For general random vectors $\mathbf{x}, \mathbf{y}$ for which the means and variances exist, we have:

- $\mathbb{E}_{\mathbf{x},\mathbf{y}}[\mathbf{x} + \mathbf{y}] = \mathbb{E}_{\mathbf{x}}[\mathbf{x}] + \mathbb{E}_{\mathbf{y}}[\mathbf{y}]$
- $\mathbb{V}_{\mathbf{x}}[\mathbf{x}] = \mathbb{E}_{\mathbf{x}}[\mathbf{x}\mathbf{x}^\intercal] - \mathbb{E}_{\mathbf{x}}[\mathbf{x}]\mathbb{E}_{\mathbf{x}}[\mathbf{x}]^\intercal$
- $\mathbb{V}_{\mathbf{x},\mathbf{y}}[\mathbf{x} + \mathbf{y}] = \mathbb{V}_{\mathbf{x}}[\mathbf{x}] + \mathbb{V}_{\mathbf{y}}[\mathbf{y}]$, if $\mathbf{x}$ and $\mathbf{y}$ are independent
- $\mathbb{V}_{\mathbf{x},\mathbf{y}}[\mathbf{x} - \mathbf{y}] = \mathbb{V}_{\mathbf{x}}[\mathbf{x}] + \mathbb{V}_{\mathbf{y}}[\mathbf{y}]$, if $\mathbf{x}$ and $\mathbf{y}$ are independent
- $\mathbb{V}_{\mathbf{x}}[c\mathbf{x}] = c^2\mathbb{V}_{\mathbf{x}}[\mathbf{x}]$
- $\mathbb{V}_{\mathbf{x},\mathbf{y}}[\mathbf{x} + \mathbf{y}] = \mathbb{V}_{\mathbf{x}}[\mathbf{x}] + \mathbb{V}_{\mathbf{y}}[\mathbf{y}] + \mathrm{Cov}_{\mathbf{x},\mathbf{y}}[\mathbf{x}, \mathbf{y}] + \mathrm{Cov}_{\mathbf{x},\mathbf{y}}[\mathbf{y}, \mathbf{x}]$
- $\mathrm{Cov}_{\mathbf{x},\mathbf{y}}[\mathbf{x}, \mathbf{y}] = 0$ if $\mathbf{x}, \mathbf{y}$ are independent (but not necessarily the other way round!)
- Subscripts of the covariance matrix of vector-valued random variables determine the ordering of the axes of the matrix. So for $\mathbf{x} \in \mathbb{R}^D$ and $\mathbf{y} \in \mathbb{R}^E$, we have $\mathbf{\Sigma}_{\mathbf{x}\mathbf{y}} \in \mathbb{R}^{D\times E}$ with

$$\begin{aligned}\mathbf{\Sigma}_{\mathbf{x}\mathbf{y}} &= \mathrm{Cov}[\mathbf{x}, \mathbf{y}] = \mathbb{E}_{p(\mathbf{x},\mathbf{y})}[(\mathbf{x} - \mathbf{m}_{\mathbf{x}})(\mathbf{y} - \mathbf{m}_{\mathbf{y}})^\intercal] \\ &= \mathbb{E}[\mathbf{x}\mathbf{y}^\intercal] - \mathbf{m}_{\mathbf{x}}\mathbf{m}_{\mathbf{y}}^\intercal \,, \qquad (2) \\ \implies [\mathbf{\Sigma}_{\mathbf{x}\mathbf{y}}]_{ij} &= \mathrm{Cov}[x_i, y_j] \,. \qquad (3)\end{aligned}$$

- Covariance matrices are symmetric by definition.
- Covariance matrices are always positive semidefinite (PSD), i.e. $\mathbf{a}^\intercal\mathbf{\Sigma}\mathbf{a} \geq 0, \forall \mathbf{a}$. This comes from the fact that for a random variable $\mathbf{x}$ with covariance $\mathbf{\Sigma}$, we can define a scalar random variable $\mathbf{a}^\intercal\mathbf{x}$ for a constant $\mathbf{a}$. Its variance must be $\mathbf{a}^\intercal\mathbf{\Sigma}\mathbf{a}$, and variances are always positive.

4.1.2 Properties of Gaussians

- The family of Gaussian distributions is **closed under linear transformations**. I.e. transforming the outcome of a Gaussian random vector $\mathbf{x}$ by a matrix $\boldsymbol{A}$ ($\boldsymbol{A}\mathbf{x}$) will also be Gaussian distributed (see above for its variance).

  This is the **single most important** property of Gaussians that leads to many of its other properties.

- Gaussians are closed under **marginalisation** (take $\boldsymbol{A}$ to be a row vector with a element being 1), i.e. for a Gaussian $p(\mathbf{x}, \mathbf{y})$ we have

$$p(\mathbf{x}) = \int p(\mathbf{x}, \mathbf{y}) \mathrm{d}\mathbf{y} = \int \mathcal{N}\left(\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix}; \begin{bmatrix} \mathbf{m_x} \\ \mathbf{m_y} \end{bmatrix}, \begin{bmatrix} \mathbf{\Sigma_{xx}} & \mathbf{\Sigma_{xy}} \\ \mathbf{\Sigma_{yx}} & \mathbf{\Sigma_{yy}} \end{bmatrix} \right) \mathrm{d}\mathbf{y} = \mathcal{N}(\mathbf{x}; \mathbf{m_x}, \mathbf{\Sigma_{xx}}) \,. \tag{4}$$

4.2 Identities that will be provided

These identities are useful and will be provided in an exam.

- Gaussian probability density function (pdf) with input $\mathbf{x} \in \mathbb{R}^D$, which in my notes I designate by $\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \mathbf{\Sigma})$ is

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \mathbf{\Sigma}) = (2\pi)^{-\frac{D}{2}} |\mathbf{\Sigma}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^\intercal \mathbf{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right) . \tag{5}$$

- For a joint Gaussian density

$$p\left(\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} \right) = \mathcal{N}\left(\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix}; \begin{bmatrix} \mathbf{m_x} \\ \mathbf{m_y} \end{bmatrix}, \begin{bmatrix} \mathbf{\Sigma_{xx}} & \mathbf{\Sigma_{xy}} \\ \mathbf{\Sigma_{yx}} & \mathbf{\Sigma_{yy}} \end{bmatrix} \right) , \tag{6}$$

  we have the conditional density

$$p(\mathbf{x} \,|\, \mathbf{y}) = \mathcal{N}\big(\mathbf{x}; \quad \mathbf{m_x} + \mathbf{\Sigma_{xy}} \mathbf{\Sigma_{yy}}^{-1} (\mathbf{y} - \mathbf{m_y}), \quad \mathbf{\Sigma_{xx}} - \mathbf{\Sigma_{xy}} \mathbf{\Sigma_{yy}}^{-1} \mathbf{\Sigma_{yx}}\big) \,. \tag{7}$$

5 Exercises

I have marked questions that use skills that are *particularly* important for the exams by an asterisk (*). Other questions are there to deepen your understanding.

5.1 Inference

In the lectures, I recommended:

- Bent coin [MacKay, 2003, §3.2]
- Legal evidence [MacKay, 2003, §3.4]
- Burglar alarm [MacKay, 2003, ch 21]

Answers are provided in the book.

5.2 Probability of Everything

Question 1 (Regression model)**.** Consider a standard regression model, like we saw in Mathematics for ML [Deisenroth et al., 2020]. Following the procedure from "Building Probabilistic Models" (slide 6) factorise the joint distribution in terms of easily defined conditional distributions. Justify the factorisations briefly.

Question 2 (Regression posterior)**.** Express the weight posterior $p(\boldsymbol{\theta}|X, \mathbf{y})$ in terms of distributions defined in the previous question. Describe the rule you use in each step of your derivation. Does the posterior depend on the choice of $p(X)$?

Question 3 (Regression prediction*)**.** You want to make a prediction for some new input $\mathbf{x}_*$.

- Which new random variables should you introduce to your PoE?

- Which conditional distributions should they have? Justify your choice.
- Which distribution should you find when making a prediction? Derive this in the same way as the previous question.

I could have phrased this question in a less vague way by giving you the specific random variables to add, and telling you to find the distribution of interest. This would turn the problem into a plain mathematics problem. This, however, is how you will encounter problems in practice. Sometimes I have to spend some time thinking about which distribution is actually the answer to the question I'm asking. This requires understanding the problem and its relation to reality. The mathematics alone is not enough.

Question 4 (Hierarchical models*)**.** Consider a joint (i.e. PoE) that factorises into tractable densities as

$$p(\mathbf{y}, \mathbf{f}, \boldsymbol{\theta}) = \prod_{n=1}^{N} p(y_n|f_n)p(\mathbf{f}|\boldsymbol{\theta})p(\boldsymbol{\theta})\,. \tag{8}$$

In terms of tractable densities, find

- the posterior $p(\mathbf{f}|\mathbf{y}, \boldsymbol{\theta})$ (does this depend on $p(\boldsymbol{\theta})$?),
- $p(\mathbf{y}|\boldsymbol{\theta})$,
- $p(\boldsymbol{\theta}|\mathbf{y})$ in terms of $p(\mathbf{y}|\boldsymbol{\theta})$.

Make sure to include the derivation.

Also answer the following:

- Is $p(\mathbf{y}|\boldsymbol{\theta}) = \prod_n p(y_n|\boldsymbol{\theta})$?
- Is $p(y_n|f_n) = p(y_n|f_n, \boldsymbol{\theta})$?
- Is $p(y_n|\mathbf{f}) = p(y_n|f_n)$?
- Does $p(\mathbf{y}|\boldsymbol{\theta})$ depend on $p(\mathbf{f}|\boldsymbol{\theta})$?

Question 5 (Gaussian Processes*)**.** Later in the course we will look at Gaussian Processes. Here, we will already consider the structure of its joint, and how to manipulate it. Consider a joint that factorises as

$$p(\mathbf{y}, \mathbf{y}^*, \mathbf{f}, \mathbf{f}^*|X, X^*, \boldsymbol{\theta}) = \left[\prod_{n=1}^{N} p(y_n|f_n, \mathbf{x}_n)\right]\left[\prod_{t=1}^{N^*} p(y_t^*|f_t^*, \mathbf{x}_t^*)\right] p(\mathbf{f}, \mathbf{f}^*|X, X^*)\,, \tag{9}$$

with the additional knowledge that

$$p(\mathbf{f}|X, X^*) = p(\mathbf{f}|X)\,, \qquad \text{and} \qquad p(\mathbf{f}^*|X, X^*) = p(\mathbf{f}^*|X^*)\,, \tag{10}$$

and the convention that $X = \{\mathbf{x}_n\}_{n=1}^N$, $\mathbf{x}_n \in \mathbb{R}^D$, and $X \in \mathbb{R}^{N\times D}$.

- Find $p(\mathbf{f}|\mathbf{y}, X)$ in terms of tractable densities.
- Find $p(\mathbf{y}^*, \mathbf{f}^*, \mathbf{f}|\mathbf{y}, X, X^*)$ in terms of tractable densities.
- Given that you know how to find $p(\mathbf{f}|\mathbf{y}, X)$ in terms of tractable densities, find $p(\mathbf{y}^*, \mathbf{f}^*, \mathbf{f}|\mathbf{y}, X, X^*)$ in terms of tractable densities and $p(\mathbf{f}|\mathbf{y}, X)$.
- Is $p(\mathbf{f}|\mathbf{y}, X) = p(\mathbf{f}|\mathbf{y}, X, X^*)$?
- Based on the joint density given, can you find $p(\mathbf{f})$ or $p(\mathbf{y})$? What additional information do you need?

5.3 Graphical Models

Question 6 (Conditional independence and Information)**.** Consider random variables X, Y, Z with density $p(x, y, z)$. Now assume that $X \perp\!\!\!\perp Y \mid Z$. Show that

$$p(x|y, z) = p(x|z)\,. \tag{11}$$

This can be interpreted in terms of information. Probability distributions quantify our belief over possible outcomes of an observation. If this belief does not change if we additionally condition on an outcome $Y = y$, then it does not contain any additional information. Information theory actually quantifies this using concepts like *mutual information*.

Question 7 (Hierarchical models)**.** Draw a graphical model for the joint density in question 4.

Question 8 (Reflection)**.** Consider the situation of the lighting example from lecture 1. Draw the graphical model that corresponds to the factorisation assumption made in the joint. Imagine having two objects with each different colours, but illuminated by the same light source.

Question 9 (Filtering)**.** Consider the joint distribution (probability of everything) over discrete variables $\mathbf{x}_{:t} = (x_1, x_2, \ldots, x_t), x_i \in (1, 2, \ldots, D_x)$ and $\mathbf{y}_{:t} = (y_1, y_2, \ldots, y_t), y_i \in \{1, 2, \ldots, D_y\}$

$$p(\mathbf{x}_{:t}, \mathbf{y}_{:t}) = p(x_1)p(y_1|x_1) \prod_{i=2}^{t} p(x_i|x_{i-1})p(y_i|x_i). \tag{12}$$

You may assume that all distributions in the above equation are known.

- Draw the graphical model for this PoE.
- Suppose we observe $\mathbf{y}_{:t}$ and want to compute the posterior on the final state of the latent trajectory $p(x_t|\mathbf{y}_{:t})$. This is known as filtering. Using Bayes' rule, we can write this posterior as

$$p(x_t|\mathbf{y}_{:t}) = \sum_{x_1=1}^{D_x} \sum_{x_2=1}^{D_x} \cdots \sum_{x_{t-1}=1}^{D_x} \frac{p(\mathbf{x}_{:t}, \mathbf{y}_{:t})}{p(\mathbf{y}_{:t})}. \tag{13}$$

  Note that we do not need to calculate $p(\mathbf{y}_{:t})$ explicitly, as we can compute $p(x_t, \mathbf{y}_{:t})$ and then normalise for a categorical x_t. Given each x_i can take one of D_x possible values, how many terms would we need to sum over if we were to directly apply the marginalisation above?

- Assume instead that we have access to $p(x_{t-1}|\mathbf{y}_{:t-1})$, write down $p(x_t|y_{:t})$ as a function of this distribution and others distributions present in PoE (12). (Hint: start by writing out $p(x_t, y_t, x_{t-1}|\mathbf{y}_{:t-1})$). How many terms do we need to sum over to go from $p(x_{t-1}|\mathbf{y}_{:t-1})$ to $p(x_t|\mathbf{y}_{:t})$?
- If we assume an initial prior distribution $p(x_1)$, we can compute $p(x_1|y_1)$ and recursively compute the posterior at each step $p(x_i|\mathbf{y}_{:i})$ in light of a new observation y_i. This process is known as sequential filtering. How many terms must we sum over to compute $p(x_t|\mathbf{y}_{:t})$ with this approach? Compare with the naïve inference procedure in (13).

5.4 Conjugate Priors

Question 10 (Finding a Conjugate Prior)**.** Consider a coin flipping experiment, where we observe N_1 heads and N_0 tails from a Bernoulli random variable, with an unknown probability of success θ, giving the likelihood for the data $D = \{N_1, N_0\}$:

$$p(D|\theta) = \theta^{N_1} \cdot (1 - \theta)^{N_0} = \ell(\theta)\,. \tag{14}$$

Derive a conjugate prior for this likelihood.

Hints:

- Remember the form of the posterior: $p(\theta|D) \propto \ell(\theta)p(\theta)$.
- What form should $p(\theta)$ take, and what should its parameters be, so that after multiplying by $\ell(\theta)$, you get a posterior distribution with the same form, but different parameters.

- You don't need to find the normalising constant, but the answer will provide it. This requires knowing some integrals. This is a good skill to have, but not the focus of the course.

Question 11 (General Conjugate Prior for Exponential Family)**.** For a conjugate family likelihood

$$\ell(\eta) = p(x|\eta) = h(x)\exp(\eta^\intercal t(x) - A(\eta)) \tag{15}$$

- show that the conjugate prior is given by

$$p(\eta|\tau, n_0) = H(\tau, n_0)\exp(\tau^\intercal \eta - n_0 A(\eta))\,, \tag{16}$$

- and put the conjugate prior into the natural parameterisation.

Question 12 (Natural Form)**.** Put the following exponential family distributions into their natural form:

- Gamma distribution: $p(x;\alpha,\beta) = \frac{x^{\alpha-1}\beta^\alpha \exp(-\beta x)}{\Gamma(\alpha)}$.
- Beta distribution: $p(x;\alpha,\beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}$.
- Univariate Gaussian distribution: $p(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right)$.

Question 13 (Exponential Family Conjugate Priors)**.** Using exponential family relations, derive the conjugate priors for:

1. The Bernoulli likelihood.
2. A Gaussian likelihood with unknown precision $\nu = 1/\sigma^2$, but known mean μ.

Question 14 (Multivariate Gaussian Conjugacy)**.** Consider a multivariate Gaussian likelihood for $p(\mathbf{y}|\mathbf{f}) = \mathcal{N}\big(\mathbf{y};\mathbf{f},\sigma^2\mathrm{I}_N\big)$. Show that the prior $p(\mathbf{f}) = \mathcal{N}(\mathbf{f};0,\mathrm{K})$ is conjugate by placing both distributions into their natural exponential family form.

5.5 From Linear Models to GPs

Question 15 (Complete certainty)**.** Show that for a Bayesian Linear Regression model with M basis functions, the variance of the posterior becomes zero when you observe M data points with zero observation noise.

5.6 Gaussian processes

Question 16 (Distributions over functions)**.** A Gaussian process is defined as a "collection of a possibly infinite number of random variables, any finite number of which is Gaussian distributed". When using a Gaussian process as a distribution over functions, how do we use the random variables in this collection?

Question 17 (Gaussian Process shapes)**.** Consider the Gaussian process notation

$$p(\mathbf{y},\mathbf{y}^*) = \mathcal{N}\left(\begin{bmatrix}\mathbf{y}\\ \mathbf{y}^*\end{bmatrix}; \mathbf{0}, \begin{bmatrix}k(X,X) & k(X,X^*)\\ k(X^*,X) & k(X^*,X^*)\end{bmatrix}\right),. \tag{17}$$

Given that $\mathbf{y}\in\mathbb{R}^N$ and $\mathbf{y}^*\in\mathbb{R}^{N^*}$, what are the shapes of $k(X,X)$, $k(X,X^*)$, $k(X^*,X)$, $k(X^*,X^*)$

Question 18 (Determining a GP*)**.** The properties of a Gaussian process are fully determined by its mean and covariance function. T/F?

Question 19 (GP prior density*)**.** Write down the probability density of the function values at input locations $X = \{\mathbf{x}_1,\ldots,\mathbf{x}_5\}$ of a GP with mean function $\mu(\mathbf{x})$ and covariance function $k(\mathbf{x},\mathbf{x}')$.

Question 20 (GP posterior*)**.** To do regression over an unknown function $f(\cdot)$, we use a GP prior $f(\cdot)\sim\mathcal{GP}(0,k(\mathbf{x},\mathbf{x}'))$. We observe N noisy observations $\mathbf{y} = \{y_1,\ldots,y_N\}$ which are i.i.d. distributed according to the likelihood $p(y_n|f(\mathbf{x}_n),\mathbf{x}_n) = \mathcal{N}\big(y_n;f(\mathbf{x}_n),\sigma^2\big)$. What is the posterior over the function values at an arbitrary set of input locations X^*.

Question 21 (Marginalisation)**.** Why do we not need to consider any points other than the training and testing points when doing regression over an entire function?

Question 22 (Predicting observations*)**.** How does the predictive distribution of a new *observation* at an input point $\mathbf{x}^*$ differ from the posterior over a function value $f(\mathbf{x}^*)$ (calculated in Q4)?

Question 23 (Weights to covariances*)**.** Say we parameterise multivariate linear functions as $f(\mathbf{x}) = \boldsymbol{\theta}_1^\intercal \mathbf{x} + \theta_2$, and we define a prior over them by defining the priors over parameters as

$$p(\boldsymbol{\theta}_1) = \mathcal{N}(\boldsymbol{\theta}_1; 0, v_1 I)\,, \qquad p(\theta_2) = \mathcal{N}(\theta_2; 0, v_2)\,. \tag{18}$$

This actually defines a Gaussian process over $f(\cdot)$. Can you find its mean and covariance functions? Can you apply the same procedure to get the covariance function over functions parameterised as quadratics or higher-order polynomials?

Question 24 (Sums of GPs*)**.** We want to learn $f(\cdot)$, which can be decomposed as a sum of two functions:

$$f(\mathbf{x}) = f_1(\mathbf{x}) + f_2(\mathbf{x})\,. \tag{19}$$

If we know that $f_1 \perp\!\!\!\perp f_2$, and that the covariance functions for $f_1(\cdot)$ and $f_2(\cdot)$ are $k_1(\cdot,\cdot')$ and $k_2(\cdot,\cdot')$ respectively, what will the prior over $f(\cdot)$ be? Specify the type of distribution and its parameters. If we make observations at $f(X)$ what is the posterior for $f_1(\cdot)$?

5.7 Model selection

Question 25 (Hyperparameter Conjugacy)**.** Consider GP prior with a Squared Exponential kernel and uniform priors on its hyperparameters. Can you find the posterior on the hyperparameters $p(\boldsymbol{\theta}|\mathbf{y})$ in closed form?

More questions for this part of the lectures is contained in the computer-based lab session. See EdStem for the link to the online Colab notebook.

5.8 Gaussian process computations and limitations

Question 26 (Sparse approximation*)**.** A white noise covariance has the form $k(x, x') = \sigma_f^2 \mathbb{1}(x - x')$, where $\mathbb{1}(\cdot)$ takes the value 1 if its input is 0. If input data is drawn from $\mathbf{x} \sim \mathcal{N}(0, 1)$, does this covariance have a high-quality low-rank approximation to the covariance matrix? What about if the input data is drawn from a discrete distribution with 5 possible values?

Question 27 (Limitations of stationarity*)**.** Imagine learning the function

$$f(x) = \sin(\log(1 + \exp^x)) \tag{20}$$

from data in the range $-10 < x < 10$. Explain how the predictions of a stationary covariance like the Squared Exponential would inadequately extrapolate, and how this holds back the value of the marginal likelihood.

5.9 Decision theory

Question 28 (Gold prospecting*)**.** This question was discussed in lectures, but is discussed more in-depth here. This example is also discussed in MacKay [2003, ch. 36], but the answer I include will have a bit more detail.

We are gold miners looking to buy one of N plots of land to mine on. Based on our knowledge of the areas, we have prior distributions on the total profit x_n that will come from each mine:

$$p(x_n) = \mathcal{N}\big(x_n; \mu_n, \sigma_n^2\big)\,. \tag{21}$$

We can choose to do an exploratory mining operation on *one* plot of land to gain some more information about the profit we could obtain. If we do this, we obtain a noisy observation of the true profit we would obtain from the mine through the likelihood

$$p(d_n \,|\, x_n) = \mathcal{N}\big(d_n; x_n, \sigma^2\big)\,. \tag{22}$$

However, we incur a cost of c_n for doing an exploratory dig at location n. We aim to maximise our expected utility based on the information we have. Our utility is simply the total profit. Answer the following questions:

- If we do not do an exploratory dig, what action would we take?
- If we do consider doing an exploratory dig, when would we do one?
- If we do decide to do an exploratory dig, what is the change in our expected utility?

Question 29 (Timing*)**.** In question 28, we saw that the possibility of acquiring new information always resulted in a higher expected utility. Would we still improve the expected utility if

- we had to mine in the location that we prospect?
- we had to choose the site to prospect at the same time as choosing the site to mine?

Derive the expected utility in these cases.

Question 30 (Probability of improvement*)**.** In question 28 we saw that prospecting would still be worth it for large differences in prior means if the uncertainty was large enough. For prospecting sites that have equivalent expected utility but increasing difference in prior mean returns (and therefore increasing prior uncertainties), what happens to the *probability of improvement*? I.e. the probability that we actually would make a switch based on d_p.

5.10 Bayesian Optimisation

Question 31 (Single observation*)**.** You are performing Bayesian optimisation on a black-box function, and you have made N observations, summarised in the dataset $\mathcal{D} = \{\mathbf{x}_n, y_n\}_{n=1}^n$. You get to perform one more evaluation for free. By using decision theory, derive the optimal place to evaluate next, based on your current belief. Which acquisition function have you derived?

5.11 Laplace Approximation

Question 32 (Gaussian process classification*)**.** You have a Gaussian process prior $p(f)$, and a Bernoulli likelihood for a binary label:

$$p(\mathbf{y} \mid f(X), X) = \prod_{n=1}^{N} p(y_n \mid f(\mathbf{x}_n), \mathbf{x}_n) = p_n^{y_n}(1-p_n)^{1-y_n} \tag{23}$$

$$p_n = \sigma(f(\mathbf{x}_n)) = \frac{1}{1+e^{-f(\mathbf{x}_n)}}\,. \tag{24}$$

- What is the condition for $f(X)$ that must hold at the maximum of the posterior distribution?
- Using a suitable notation, what is the covariance of the Laplace approximation? You can assume that $\mathbf{f}^* = \operatorname{argmax}_{\mathbf{f}} p(f(X) \mid \mathbf{y})$.
- Clearly state the Laplace approximation you end up with.
- (Jumping ahead) Derive a Monte Carlo estimate for the predictive distribution.

Question 33 (Gaussian process regression*)**.** You have a Gaussian process prior $p(f)$, and a *full-rank* Gaussian likelihood $p(\mathbf{y} \mid f(X)) = \mathcal{N}(\mathbf{y}; f(X), \boldsymbol{\Sigma})$. Using the Laplace approximation, derive an approximate posterior. How good is this posterior?

5.12 Sampling and Monte Carlo

Question 34 (Rejection sampling*)**.** Rejection sampling is an algorithm to generate samples from an unnormalised probability distribution $\tilde{p}(\mathbf{x})$. We know a constant k such that

$$kq(\mathbf{x}) \geq \tilde{p}(\mathbf{x})\,, \qquad \forall \mathbf{x}\,. \tag{25}$$

Rejection sampling works by drawing pairs of samples

$$\mathbf{x}^{[s]} \sim q(\mathbf{x})\,, \qquad u^{[s]} \mid \mathbf{x}^{[s]} \sim \text{Uniform}\Big(0, k \cdot q(\mathbf{x}^{[s]})\Big)\,. \tag{26}$$

We return the samples that have $u^{[s]} < \tilde{p}(\mathbf{x})$.

- What event should we condition on to find the distribution of samples generated by rejection sampling? I.e. write down Bayes' rule with the condition that holds for samples we return.
- Prove that the distribution returned by rejection sampling is from the distribution of interest $p(\mathbf{x}) = \frac{1}{Z}\tilde{p}(\mathbf{x})$, where $Z = \int \tilde{p}(\mathbf{x})\mathrm{d}\mathbf{x}$.

Question 35 (Importance sampling*)**.** In the lectures we saw a proof for the unbiasedness of an importance sampling estimator of the integral $I = \int f(\mathbf{x})p(\mathbf{x})\mathrm{d}\mathbf{x}$. We noted that the estimate needed exact evaluations of $p(\mathbf{x})$ (i.e. in its normalised form).

- What is the mean of the importance sampling estimator if we evaluate the unnormalised $\tilde{p}(\mathbf{x})$ instead of the normalised $p(\mathbf{x})$? I.e. if we used the estimator

$$\hat{J} = \frac{1}{S}\sum_{s=1}^{S} f(\mathbf{x}^{[s]})\frac{\tilde{p}(\mathbf{x}^{[s]})}{q(\mathbf{x}^{[s]})}\,, \qquad \mathbf{x}^{[s]} \sim q(\mathbf{x})\,. \tag{27}$$

- How does $\mathbb{E}_{q(\mathbf{x})}\left[\hat{J}\right]$ differ from the quantity we are trying to estimate?
- Construct a Monte Carlo estimate for the factor by which the unnormalised importance sampling estimate differs from
- Can you construct an estimator from $\hat{J}$ that computes the right value as $S \to \infty$?

Question 36 (Variance of importance sampling*)**.** Say, we want to estimate the marginal likelihood

$$I = \int p(\mathbf{y}\,|\,\mathbf{x})p(\mathbf{x})\mathrm{d}\mathbf{x}\,. \tag{28}$$

- Construct an importance sampling estimate for the marginal likelihood.
- What is the proposal distribution that gives the minimum variance estimate of the marginal likelihood?

Question 37 (Variance of importance sampling II*)**.** Ex 29.13 in MacKay.

Question 38 (Metropolis-Hastings)**.** Prove that the Metropolis-Hastings method gives a Markov Chain that converges to a target distribution $p(\mathbf{x})$.

Question 39 (Independent Markov Chains)**.** Consider the process of generating samples from two independent Markov chains. By considering the joint density between their states, show that the final samples from both Markov chains are also independent.

Question 40 (GP hyperparameters for non-conjugate likelihoods)**.** Consider a Gaussian process model with an arbitrary (i.e. possibly non-conjugate) likelihood $p(y_n|f(\mathbf{x}_n))$.

1. State the integral that computes the predictive distribution $p(y^*|\mathbf{y})$, where y^* is the observation for the input x^*. Note that we integrated out the hyperparameters $\boldsymbol{\theta}$ here.
2. State the Monte Carlo approximation to this integral.
3. State the procedure for computing samples for the Monte Carlo approximation.

Question 41 (GP hyperparameters for regression)**.** Consider GP regression. Describe the whole process for computing the predictive distribution $p(y^*|\mathbf{y})$ using Metropolis-Hastings if the Markov Chain:

1. only runs on the hyperparameters $\boldsymbol{\theta}$,
2. runs on the hyperparameters and function values,
3. runs on the hyperparameters and function values and the proposal distribution for the function values is $p(f(X)|\mathbf{y}, \boldsymbol{\theta})$.

Pay special attention to the acceptance probabilities.

5.13 Variational Inference

Question 42 (Independence and KL divergence)**.** Show that for $q(\mathbf{x}) = \prod_n q(x_n)$ and $p(\mathbf{x}) = \prod_n p(x_n)$ the KL divergence simplifies as

$$\mathrm{KL}[q(\mathbf{x})||p(\mathbf{x})] = \sum_n \mathrm{KL}[q(x_n)||p(x_n)]\,. \tag{29}$$

Question 43 (Reparameterisation gradient*)**.** You have a Gaussian process prior $p(f)$, and a Bernoulli likelihood for a binary label:

$$p(\mathbf{y} \mid f(X), X) = \prod_{n=1}^{N} p(y_n \mid f(\mathbf{x}_n), \mathbf{x}_n) = p_n^{y_n}(1-p_n)^{1-y_n} \tag{30}$$

$$p_n = \sigma(f(\mathbf{x}_n)) = \frac{1}{1+e^{-f(\mathbf{x}_n)}}\,. \tag{31}$$

- Derive the variational lower bound using a Gaussian variational distribution.
- Derive the reparameterisation gradient for the expected log-likelihood term.

Question 44 (Bayesian neural networks)**.** Derive a variational lower bound and its reparameterisation gradients for a neural network, using a mean-field Gaussian distribution.

Question 45 (Variational Autoencoders)**.** Derive the gradient estimators for Variational Autoencoders for the model parameters and recognition model parameters.

Question 46 (Bayesian Variational Autoencoders)**.** Consider a Variational Autoencoder, but now we place a prior on the neural network weights $p(\mathbf{w})$.

- Derive the ELBO for this case, where we also want to approximate the posterior on the weights. Assume independence between $\mathbf{w}$ and each $\mathbf{z}_n$ in the variational approximation.
- Which gradient estimator can you use for this model? No need to derive the estimator.

Question 47 (Imputation with a VAE)**.** You are given an image where certain pixels have been lost. How would you use a VAE to find what the full picture looked like?

6 Answers

6.1 Question 1 – Regression model

We have the following variables: data outputs $\mathbf{y} = \{y_n\}_{n=1}^N$, data inputs $X = \{\mathbf{x}_n\}_{n=1}^N$, parameters $\boldsymbol{\theta}$. Our key assumption is that we want to find some function $f_{\boldsymbol{\theta}}(\cdot)$ from noisy observations $\mathbf{y}$. Our probability of everything becomes:

$$\begin{aligned}
p(\mathbf{y}, X, \boldsymbol{\theta}) &= p(\mathbf{y}|X, \boldsymbol{\theta})p(X, \boldsymbol{\theta}) && \text{Always possible.} && (32)\\
&= \prod_n p(y_n|\mathbf{x}_n, \boldsymbol{\theta})p(X, \boldsymbol{\theta}) && \text{Noise for each observation is iid.} && (33)\\
&= \prod_n p(y_n|\mathbf{x}_n, \boldsymbol{\theta})p(X)p(\boldsymbol{\theta}) && \text{Function is not influenced by where you evaluate it.} && (34)
\end{aligned}$$

For the most common regression model, we choose the tractable densities for the conditionals:

$$p(y_n|\mathbf{x}_n, \boldsymbol{\theta}) = \mathcal{N}\left(y_n; \boldsymbol{\phi}(\mathbf{x}_n)^\intercal \boldsymbol{\theta}, \sigma^2\right), \tag{35}$$

$$p(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}; 0, v\mathrm{I})\,. \tag{36}$$

6.2 Question 2 – Regression posterior

Given our regression model, we have a known and tractable expression for $p(\mathbf{y}|X, \boldsymbol{\theta})$ and $p(\boldsymbol{\theta})$. Therefore, we can rewrite the relevant joint distribution as

$$\begin{aligned} p(\mathbf{y}, X, \boldsymbol{\theta}) &= p(\mathbf{y}|X, \boldsymbol{\theta})p(X, \boldsymbol{\theta}) && (37)\\ &= \prod_n p(y_n|\mathbf{x}_n, \boldsymbol{\theta})p(X, \boldsymbol{\theta}) && \text{Noise for each observation is iid.} \quad (38)\\ &= \prod_n p(y_n|\mathbf{x}_n, \boldsymbol{\theta})p(X)p(\boldsymbol{\theta}) && \boldsymbol{\theta} \text{ and } X \text{ are independent} \quad (39) \end{aligned}$$

On the other hand, we are interested in the posterior $p(\boldsymbol{\theta}|\mathbf{y}, X)$, which we rewrite in terms of the joint as:

$$p(\boldsymbol{\theta}|\mathbf{y}, X) = \frac{p(\mathbf{y}, X, \boldsymbol{\theta})}{p(\mathbf{y}, X)} \tag{40}$$

Substituting in the previous expression for the joint then gives

$$p(\boldsymbol{\theta}|\mathbf{y}, X) = \frac{\prod_n p(y_n|\mathbf{x}_n, \boldsymbol{\theta})p(X)p(\boldsymbol{\theta})}{p(\mathbf{y}, X)} \tag{41}$$

What is left is to rewrite $p(\mathbf{y}, X)$ in terms of known quantities. Once again, we start from the joint distribution and this time marginalise over $\boldsymbol{\theta}$. Using our previous factorisation of the joint results in

$$\begin{aligned} p(\mathbf{y}, X) &= \int p(\mathbf{y}, X, \boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} && (42)\\ &= \int \prod_n p(y_n|\mathbf{x}_n, \boldsymbol{\theta})p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}p(X) && (43) \end{aligned}$$

where $p(X)$ is independent of θ so can be taken outside the integral. Combining with the previous expression for $p(\boldsymbol{\theta}|\mathbf{y}, X)$ results in

$$p(\boldsymbol{\theta}|\mathbf{y}, X) = \frac{\prod_n p(y_n|\mathbf{x}_n, \boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{y}|X)}, \qquad p(\mathbf{y}|X) = \int \prod_n p(y_n|\mathbf{x}_n, \boldsymbol{\theta})p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}\,. \tag{44}$$

We have now rewritten the posterior in terms of $p(y_n|\mathbf{x}_n, \boldsymbol{\theta})$ (the likelihood for each data point from the regression model) and $p(\boldsymbol{\theta})$ (the prior over parameters). The $p(X)$ factors canceled out so this posterior does not depend on $p(X)$, as is to be expected since it is a posterior conditioned on X.

6.3 Question 3 – Regression prediction*

6.3.1 Which new random variables should you introduce to your PoE?

To make a prediction, we will use the observed data points $X, \mathbf{y}$, with notation identical to the previous questions, but we we will also need to consider some additional variables for the predicted data point. To this end we introduce the new variables $\mathbf{x}_*, \mathbf{y}_*$ to denote the variables representing the prediction. Just like in the previous questions we will be employing a regression model, parametrised by $\boldsymbol{\theta}$ with Gaussian noise likelihood.

6.3.2 Which conditional distributions should they have? Justify your choice

We will assume that the same model that generates the observed data will also generate the next data point which we wish to predict the distribution of. Our model posits that for any given location $\mathbf{x}$ and parametrisation $\boldsymbol{\theta}$ the probability of y will be given by $p(y|\boldsymbol{\theta}, \mathbf{x}) = \mathcal{N}\left(y; \boldsymbol{\phi}(\mathbf{x})^\intercal\boldsymbol{\theta}, \sigma^2\right)$ for some constant σ^2. We assume this to be true for each individual observed pair x_n, y_n and the condition that the model generates the next point consistently with how the observed data was generated will imply that $p(\mathbf{y}_*|\boldsymbol{\theta}, \mathbf{x}_*) = \mathcal{N}\left(\mathbf{y}_*; \boldsymbol{\phi}(\mathbf{x}_*)^\intercal\boldsymbol{\theta}, \sigma^2\right)$ also.

6.3.3 Which distribution should you find when making a prediction? Derive this in the same way as the previous question

We wish to make a prediction of what new value $\mathbf{y}_*$ we will get at the new given location $\mathbf{x}_*$ given that we have already observed data $X, \mathbf{y}$. The quantity of interest is therefore the distribution of this new $\mathbf{y}_*$ under the conditions $p(\mathbf{y}_*|X, \mathbf{y}, \mathbf{x}_*)$.

Therefore, in line with previous questions, we can factorise the joint distribution as

$$p(\mathbf{y}, \mathbf{y}_*, X, \mathbf{x}_*, \boldsymbol{\theta}) \overset{AT}{=} p(\mathbf{y}, \mathbf{y}_*, \boldsymbol{\theta}|X, \mathbf{x}_*)p(X, \mathbf{x}_*) \tag{45}$$

$$\overset{AT}{=} p(\mathbf{y}, \mathbf{y}_*|\boldsymbol{\theta}, X, \mathbf{x}_*)p(\boldsymbol{\theta}|X, \mathbf{x}_*)p(X, \mathbf{x}_*) \tag{46}$$

$$\overset{MA}{=} p(\mathbf{y}, \mathbf{y}_*|\boldsymbol{\theta}, X, \mathbf{x}_*)p(X, \mathbf{x}_*)p(\boldsymbol{\theta}) \tag{47}$$

$$\overset{MA}{=} \prod_n p(y_n|\mathbf{x}_n, \boldsymbol{\theta})p(\mathbf{y}_*|\mathbf{x}_*, \boldsymbol{\theta})p(X, \mathbf{x}_*)p(\boldsymbol{\theta}) \tag{48}$$

where 'AT' denotes a step that is always true and 'MA' one which is valid only due to the assumptions behind the particular model we are considering. In particular we used the independence between the points at which the function is evaluated, $(X, \mathbf{x}_*)$, and $\boldsymbol{\theta}$ for the first 'MA' step and the independence of all y values for the second - both of which are taken as true from our model assumptions.

As in the previous questions, to write this in a tractable form we use our expression for the joint and marginalise over the relevant variables:

$$p(\mathbf{y}_*|X, \mathbf{y}, \mathbf{x}_*) = \frac{p(X, \mathbf{y}, \mathbf{x}_*, \mathbf{y}_*)}{p(X, \mathbf{y}, \mathbf{x}_*)} \tag{49}$$

$$= \frac{\int p(X, \mathbf{y}, \mathbf{x}_*, \mathbf{y}_*, \theta)\mathrm{d}\boldsymbol{\theta}}{\int p(X, \mathbf{y}, \mathbf{x}_*, \mathbf{y}_*, \theta)\mathrm{d}\boldsymbol{\theta}\mathrm{d}\mathbf{y}_*} \tag{50}$$

$$= \frac{\left[\int \prod_n p(y_n|\mathbf{x}_n, \boldsymbol{\theta})p(\mathbf{y}_*|\mathbf{x}_*, \boldsymbol{\theta})p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}\right]p(X, \mathbf{x}_*)}{\left[\int \prod_n p(y_n|\mathbf{x}_n, \boldsymbol{\theta})p(\mathbf{y}_*|\mathbf{x}_*, \boldsymbol{\theta})p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}\mathrm{d}\mathbf{y}_*\right]p(X, \mathbf{x}_*)} \tag{51}$$

We can simplify the denominator performing the $\mathbf{y}_*$ integral by noting that no other term other than $p(\mathbf{y}_*|\mathbf{x}_*, \boldsymbol{\theta})$ depends on $\mathbf{y}_*$ and that $\int p(\mathbf{y}_*|\mathbf{x}_*, \boldsymbol{\theta})\mathrm{d}\mathbf{y}_* = 1$, and after cancelling factors of $p(X, \mathbf{x}_*)$ we get the answer

$$p(\mathbf{y}_*|X, \mathbf{y}, \mathbf{x}_*) = \frac{\int \prod_n p(y_n|\mathbf{x}_n, \boldsymbol{\theta})p(\mathbf{y}_*|\mathbf{x}_*, \boldsymbol{\theta})p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}}{\int \prod_n p(y_n|\mathbf{x}_n, \boldsymbol{\theta})p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}} \tag{52}$$

Finally, note that this result could be obtained a number of different ways depending on what the starting point for the 'probability of everything' is taken to be or how equivalent versions of Bayes' theorem are applied - regardless of how you achieved an answer you are still encouraged to think about which steps are always true and which are instead consequences of the specific model we are considering. If you'd like to check whether your reasoning is correct you are welcome to discuss this with a TA during the sessions after the live lecture.

6.4 Question 4 – Hierarchical models*

We are told that the probability of everything factorises as

$$p(\mathbf{y}, \mathbf{f}, \boldsymbol{\theta}) = \prod_{n=1}^{N} p(y_n|f_n)p(\mathbf{f}|\boldsymbol{\theta})p(\boldsymbol{\theta}). \tag{53}$$

Note that this factorisation is an assumption specific to the model, not a relation which holds in general.

6.4.1 Functional posterior

To derive the posterior over function values, $p(\mathbf{f}|\mathbf{y}, \boldsymbol{\theta})$ we can first apply Bayes' rule and substitute in our probability of everything for the numerator. If we assume tractable densities, the denominator can be

computed by marginalising out $\mathbf{f}$ from this PoE.

$$p(\mathbf{f}|\mathbf{y},\boldsymbol{\theta}) \overset{AT}{=} \frac{p(\mathbf{y},\mathbf{f},\boldsymbol{\theta})}{p(\mathbf{y},\boldsymbol{\theta})} = \frac{\prod_{n=1}^{N} p(y_n|f_n)p(\mathbf{f}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{y},\boldsymbol{\theta})} \tag{54}$$

$$= \frac{\prod_{n=1}^{N} p(y_n|f_n)p(\mathbf{f}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(\mathbf{y},\mathbf{f}',\boldsymbol{\theta})\mathrm{d}\mathbf{f}'} \tag{55}$$

$$= \frac{\prod_{n=1}^{N} p(y_n|f_n)p(\mathbf{f}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int \prod_{n=1}^{N} p(y_n|f'_n)p(\mathbf{f}'|\boldsymbol{\theta})p(\boldsymbol{\theta})\mathrm{d}\mathbf{f}'} \tag{56}$$

$$= \frac{\cancel{p(\boldsymbol{\theta})}\prod_{n=1}^{N} p(y_n|f_n)p(\mathbf{f}|\boldsymbol{\theta})}{\cancel{p(\boldsymbol{\theta})}p(\mathbf{y}|\boldsymbol{\theta})} \tag{57}$$

$$\text{where } p(\mathbf{y}|\boldsymbol{\theta}) = \int \prod_{n=1}^{N} p(y_n|f_n)p(\mathbf{f}|\boldsymbol{\theta})\mathrm{d}\mathbf{f}. \tag{58}$$

Note that $\mathbf{f}'$ and $\mathbf{f}$ are the same variable, but we use dash to highlight that the integration is separate from other occurrences of $\mathbf{f}$ in the expression. We see that the prior over $\boldsymbol{\theta}$ can be taken outside the integral in the denominator and cancels with the same term in the numerator. Thus the posterior over function values does **not** depend on the prior on parameters $p(\boldsymbol{\theta})$.

We can also see this in a different way, by noting that

$$p(\mathbf{f}|\mathbf{y},\boldsymbol{\theta}) \overset{AT}{=} \frac{p(\mathbf{f},\mathbf{y}|\boldsymbol{\theta})}{p(\mathbf{y}|\boldsymbol{\theta})} \overset{MA}{=} \frac{\prod_{n=1}^{N} p(y_n|f_n)p(\mathbf{f}|\boldsymbol{\theta})}{p(\mathbf{y}|\boldsymbol{\theta})}, \tag{59}$$

and then finding $p(\mathbf{y}|\boldsymbol{\theta})$ as above.

6.4.2 Data given parameters

We could compute $p(\mathbf{y}|\boldsymbol{\theta})$ by starting with the probability of everything, marginalising out $\mathbf{f}$ and then conditioning the resulting joint $p(\mathbf{y},\boldsymbol{\theta})$ by dividing it by $p(\boldsymbol{\theta})$. However, we have already done this work in answering the question above (see (58)).

6.4.3 Parameter posterior

First applying Bayes' rule, our distribution of interest can be written as

$$p(\boldsymbol{\theta}|\mathbf{y}) \overset{AT}{=} \frac{p(\boldsymbol{\theta},\mathbf{y})}{p(\mathbf{y})} \tag{60}$$

$$\overset{AT}{=} \frac{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{y})}. \tag{61}$$

We have already computed $p(\mathbf{y}|\boldsymbol{\theta})$ in (58) and $p(\boldsymbol{\theta})$ is known, so we are left with $p(\mathbf{y})$ which we can calculate by marginalising $\boldsymbol{\theta}$ and $\mathbf{f}$ from the PoE

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\boldsymbol{\theta})\int \prod_{n=1}^{N} p(y_n|f_n)p(\mathbf{f}|\boldsymbol{\theta})\mathrm{d}\mathbf{f}}{\int\int \prod_{n=1}^{N} p(y_n|f'_n)p(\mathbf{f}'|\boldsymbol{\theta}')p(\boldsymbol{\theta}')\mathrm{d}\boldsymbol{\theta}'\mathrm{d}\mathbf{f}'}. \tag{62}$$

6.4.4 Is $p(\mathbf{y}|\boldsymbol{\theta}) = \prod_n p(y_n|\boldsymbol{\theta})$?

We can see from the given factorisation of the joint distribution that $p(\mathbf{y}|\mathbf{f}) = \prod_n p(y_n|f_n)$, so we may be inclined to believe that $p(\mathbf{y}|\boldsymbol{\theta})$ factorises in a similar way. However, looking at the form of $p(\mathbf{y}|\boldsymbol{\theta})$ (58), this is not the case, as we need to integrate out $\mathbf{f}$ which ties the elements of $\mathbf{y}$ in $p(\mathbf{y}|\boldsymbol{\theta})$. More specifically, for $p(\mathbf{y}|\boldsymbol{\theta}) = \prod_n p(y_n|\boldsymbol{\theta})$ to hold, we would need $p(\mathbf{f}|\boldsymbol{\theta}) = \prod_n p(f_n|\boldsymbol{\theta})$, which would allow the integral over $\mathbf{f}$ in (58) to be split into the product of integrals over each f_n. Each of these integrals would form a factor $p(y_n|\boldsymbol{\theta})$. However, $p(\mathbf{f}|\boldsymbol{\theta}) \neq \prod_n p(f_n|\boldsymbol{\theta})$ in general, and the model assumptions do not suggest $p(\mathbf{f}|\boldsymbol{\theta})$ factorises in this case, so we conclude that $p(\mathbf{y}|\boldsymbol{\theta}) \neq \prod_n p(y_n|\boldsymbol{\theta})$.

6.4.5 Is $p(y_n|f_n) = p(y_n|f_n, \boldsymbol{\theta})$?

Any distribution over variables $\mathbf{y}$, $\mathbf{f}$ and $\boldsymbol{\theta}$ can be conditioned and marginalised as

$$p(y_n|f_n, \boldsymbol{\theta}) \overset{AT}{=} \frac{p(y_n, f_n|\boldsymbol{\theta})}{p(f_n|\boldsymbol{\theta})} \tag{63}$$

$$\overset{AT}{=} \frac{\int p(y_n, \mathbf{f}|\boldsymbol{\theta}) \mathrm{d}\mathbf{f}_{\neq n}}{p(f_n|\boldsymbol{\theta})}. \tag{64}$$

Now using the model specification to expand the integrand in the numerator

$$p(y_n|f_n, \boldsymbol{\theta}) \overset{MA}{=} \frac{\int p(y_n|f_n) p(\mathbf{f}|\boldsymbol{\theta}) \mathrm{d}\mathbf{f}_{\neq n}}{p(f_n|\boldsymbol{\theta})} \overset{AT}{=} \frac{p(y_n|f_n) \int p(\mathbf{f}|\boldsymbol{\theta}) \mathrm{d}\mathbf{f}_{\neq n}}{p(f_n|\boldsymbol{\theta})} \tag{65}$$

$$\overset{AT}{=} \frac{p(y_n|f_n) \cancel{p(f_n|\boldsymbol{\theta})}}{\cancel{p(f_n|\boldsymbol{\theta})}}. \tag{66}$$

We thus conclude that $p(y_n|f_n) = p(y_n|f_n, \boldsymbol{\theta})$ holds.

6.4.6 Is $p(y_n|\mathbf{f}) = p(y_n|f_n)$?

For any distribution we can write

$$p(y_n|\mathbf{f}) \overset{AT}{=} \int p(\mathbf{y}|\mathbf{f}) \mathrm{d}\mathbf{y}_{\neq n} \tag{67}$$

$$\overset{AT}{=} \frac{1}{p(\mathbf{f})} \int p(\mathbf{y}, \mathbf{f}) \mathrm{d}\mathbf{y}_{\neq n} \tag{68}$$

$$\overset{AT}{=} \frac{1}{p(\mathbf{f})} \int \int p(\mathbf{y}, \mathbf{f}, \boldsymbol{\theta}) \mathrm{d}\mathbf{y}_{\neq n} \mathrm{d}\boldsymbol{\theta}. \tag{69}$$

Now substituting in the PoE for the model

$$p(y_n|\mathbf{f}) \overset{MA}{=} \frac{1}{p(\mathbf{f})} \int \int \prod_{i=1}^{N} p(y_i|f_i) p(\mathbf{f}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) \mathrm{d}\mathbf{y}_{\neq n} \mathrm{d}\boldsymbol{\theta} \tag{70}$$

$$\overset{MA}{=} \frac{p(y_n|f_n)}{p(\mathbf{f})} \int \left(\prod_{i \neq n} \cancel{\int p(y_i|f_i) dy_i} \right)^{1} p(\mathbf{f}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} \overset{AT}{=} \frac{p(y_n|f_n)}{p(\mathbf{f})} \int p(\mathbf{f}, \boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} \tag{71}$$

$$\overset{AT}{=} \frac{p(y_n|f_n)}{\cancel{p(\mathbf{f})}} \cancel{p(\mathbf{f})} \tag{72}$$

and so $p(y_n|\mathbf{f}) = p(y_n|f_n)$ holds in this case.

6.4.7 Does $p(\mathbf{y}|\boldsymbol{\theta})$ depend on $p(\mathbf{f}|\boldsymbol{\theta})$?

We might see $p(\mathbf{y}|\boldsymbol{\theta})$ and assume that it does not depend on $p(\mathbf{f}|\boldsymbol{\theta})$ because we do not see any explicit conditioning on $\mathbf{f}$. However for the factorisation given, we compute $p(\mathbf{y}|\boldsymbol{\theta})$ as in (58) which requires integrating over $p(\mathbf{f}|\boldsymbol{\theta})$. We thus conclude that $p(\mathbf{y}|\boldsymbol{\theta})$ does indeed depend on $p(\mathbf{f}|\boldsymbol{\theta})$.

6.5 Question 5 – Gaussian Processes*

6.5.1 Find $p(\mathbf{f}|\mathbf{y}, X)$

Using Bayes rule, we know that

$$p(\mathbf{f}|\mathbf{y}, X) \overset{\text{AT}}{=} \frac{p(\mathbf{y}|\mathbf{f}, X) p(\mathbf{f}|X)}{p(\mathbf{y}|X)}. \tag{73}$$

$$\tag{74}$$

We can now write the numerator in terms of tractable densities:

$$p(\mathbf{y}|\mathbf{f}, X) = \prod_{n=1}^{N} p(y_n|f_n, \mathbf{x}_n), \qquad \text{Due to iid noise assumption.} \tag{75}$$

$$p(\mathbf{f}|X) = \int p(\mathbf{f}, \mathbf{f}^*|X, X^*)d\mathbf{f}^*. \qquad \text{Due to marginalisation and } p(\mathbf{f}|X, X^*) = p(\mathbf{f}|X). \tag{76}$$

$$\tag{77}$$

Now, looking at the denominator

$$p(\mathbf{y}|X) = \int p(\mathbf{y}, \mathbf{f}|X)d\mathbf{f}, \qquad \text{Due to marginalisation.} \tag{78}$$

$$p(\mathbf{y}, \mathbf{f}|X) = p(\mathbf{y}|\mathbf{f}, X)p(\mathbf{f}|X), \qquad \text{Always true due to chain rule.} \tag{79}$$

$$\tag{80}$$

where we have already found $p(\mathbf{f}|X)$.

6.5.2 Find $p(\mathbf{y}^*, \mathbf{f}^*, \mathbf{f}|\mathbf{y}, X, X^*)$

For $p(\mathbf{y}^*, \mathbf{f}^*, \mathbf{f}|\mathbf{y}, X, X^*)$, we will apply the chain rule for probabilities

$$p(\mathbf{y}^*, \mathbf{f}^*, \mathbf{f}|\mathbf{y}, X, X^*) = p(\mathbf{y}^*|\mathbf{f}^*, \mathbf{f}, \mathbf{y}, X, X^*)p(\mathbf{f}, \mathbf{f}^*|\mathbf{y}, X, X^*). \qquad \text{Always true.} \tag{81}$$

We can now write the numerator in terms of tractable densities:

$$p(\mathbf{y}^*|\mathbf{f}^*, \mathbf{f}, \mathbf{y}, X, X^*) = p(\mathbf{y}^*|\mathbf{f}^*, X^*), \qquad \text{Model assumptions.} \tag{82}$$

$$p(\mathbf{f}, \mathbf{f}^*|\mathbf{y}, X, X^*) = \frac{p(\mathbf{y}|\mathbf{f}, \mathbf{f}^*, X, X^*)p(\mathbf{f}, \mathbf{f}^*|X, X^*)}{p(\mathbf{y}|X, X^*)} \qquad \text{Always true.} \tag{83}$$

$$= \frac{p(\mathbf{y}|\mathbf{f}, X)p(\mathbf{f}, \mathbf{f}^*|X, X^*)}{p(\mathbf{y}|X, X^*)}. \qquad \text{Model assumptions.} \tag{84}$$

The simplification of terms due to model assumptions are due to the factorisation of the probability of everything. The denominator can be expanded similarly to the first part of the question,

$$p(\mathbf{y}|X, X^*) = \int p(\mathbf{y}|\mathbf{f}, X, X^*)p(\mathbf{f}|X, X^*)d\mathbf{f} \tag{85}$$

$$= \int p(\mathbf{y}|\mathbf{f}, X)p(\mathbf{f}|X)d\mathbf{f} \qquad \text{Model assumptions.} \tag{86}$$

$$= p(\mathbf{y}|X). \tag{87}$$

6.5.3 Find $p(\mathbf{y}^*, \mathbf{f}^*, \mathbf{f}|\mathbf{y}, X, X^*)$ in terms of tractable densities and $p(\mathbf{f}|\mathbf{y}, X)$

We have already found $p(\mathbf{y}^*, \mathbf{f}^*, \mathbf{f}|\mathbf{y}, X, X^*)$. However looking at the term

$$p(\mathbf{f}, \mathbf{f}^*|\mathbf{y}, X, X^*) = \frac{p(\mathbf{y}|\mathbf{f}, X)p(\mathbf{f}, \mathbf{f}^*|X, X^*)}{p(\mathbf{y}|X)}, \tag{88}$$

we can write the last term in the numerator as

$$p(\mathbf{f}, \mathbf{f}^*|X, X^*) = p(\mathbf{f}^*|\mathbf{f}, X, X^*)p(\mathbf{f}|X), \qquad \text{Chain rule.} \tag{89}$$

$$p(\mathbf{f}^*|\mathbf{f}, X, X^*) = \frac{p(\mathbf{f}^*, \mathbf{f}|X, X^*)}{p(\mathbf{f}|X)}. \qquad \text{Bayes rule.} \tag{90}$$

We can now see that

$$p(\mathbf{f}, \mathbf{f}^*|\mathbf{y}, X, X^*) = p(\mathbf{f}|\mathbf{y}, X)p(\mathbf{f}^*|\mathbf{f}, X, X^*). \tag{91}$$

The advantage here is that if we have already computed $p(\mathbf{f}|\mathbf{y}, X)$, we much more easily compute $p(\mathbf{f}, \mathbf{f}^*|\mathbf{y}, X, X^*)$.

6.5.4 Is $p(\mathbf{f}|\mathbf{y}, X) = p(\mathbf{f}|\mathbf{y}, X, X^*)$?

We have already found $p(\mathbf{f}|\mathbf{y}, X)$, and we can write

$$p(\mathbf{f}|\mathbf{y}, X, X^*) = \frac{p(\mathbf{y}|\mathbf{f}, X, X^*)p(\mathbf{f}|X, X^*)}{p(\mathbf{y}|X, X^*)}. \tag{92}$$

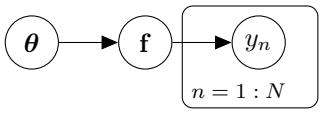
$$\tag{93}$$

We know that $p(\mathbf{f}|X, X^*) = p(\mathbf{f}|X)$ and we already found that $p(\mathbf{y}|X, X^*) = p(\mathbf{y}|X)$. $p(\mathbf{y}|\mathbf{f}, X, X^*) = p(\mathbf{y}|\mathbf{f}, X)$ is obvious from model assumptions.

6.5.5 Can you find $p(\mathbf{f})$ or $p(\mathbf{y})$?

Based on the joint, we cannot find $p(\mathbf{f})$ or $p(\mathbf{y})$ as we require $p(X, X^*)$ to evaluate these densities.

6.6 Question 7 – Hierarchical models



As an alternative, you could perform an always true product rule decomposition on $\mathbf{f}$ as $p(\mathbf{f}) = p(f_1)p(f_2|f_1)p(f_3|f_2, f_1)\ldots$. This would result in a more complicated graphical model, which would also represent the dependence of y_n on *only* f_n.

6.7 Question 8 – Reflection

In the case of one object of colour C under illumination I reflecting light L the meaningful factorisation of the joint distribution into tractable distributions was $p(C, L, I) = p(L|C, I)p(C, I) = p(L|C, I)P(C)p(I)$. Intuitively, given a colour and illumination, we could determine what the reflected light looks like ($p(L|C, I)$ term) and we assume that the illumination and 'intrinsic' colour of the object are independent variables ($p(C, I) = p(C)p(I)$).

In the case with two objects, each will have its own intrinsic colour, which we denote as C_1 and C_2, and will reflect light differently with variables L_1 and L_2. Since we have on further information about the relationship of the items, it is safe to assume their intrinsic colours will be independent, so $p(C_1, C_2) = p(C_1)p(C_2)$. Similarly, there is no relationship between colour of the objects and illumination so $p(C_1, C_2, I) = p(C_1)p(C_2)p(I)$. Moreover, it is safe to assume that L_1 will depend on C_1 and I only and C_2 will depend only on C_2 and I and so $p(L_1, L_2|C_1, C_2, I) = p(L_1|L_2, C_1, C_2, I)p(L_2|C_1, C_2, I) = p(L_1|C_1, I)p(L_2|C_2, I)$. Hence, the joint will factorise as

$$p(C_1, C_2, L_1, L_2, I) = p(L_1, L_2|C_1, C_2, I)p(C_1, C_2, I) \tag{94}$$

$$= p(L_1, |C_1, I)p(L_2|C_2, I)p(C_1)p(C_2)p(I) \tag{95}$$

which is represented in the graphical model in Figure 1.

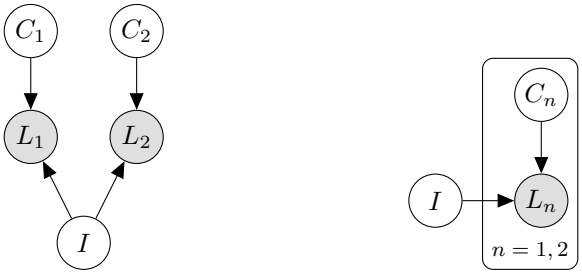


Figure 1: Graphical model representation of the reflection model. The graphical model on the right a compact but equivalent version of the one on the left.

6.8 Question 9 – Filtering

6.8.1 Graphical Model Illustration

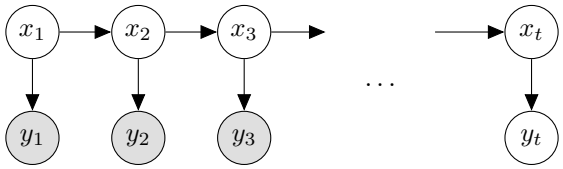


Figure 2: Directed graphical model for the joint distribution given.

6.8.2 Naïve inference

Each x_i may take one of D_x values, and for brute force marginalisation as per $p(x_t|\mathbf{y}_{:t}) = \sum_{x_1,\ldots,x_t} p(\mathbf{x}_{:t}, \mathbf{y}_{:t})/p(\mathbf{y}_{:t})$, we would need to sum over *all possible variable combinations*. Therefore, we would need sum over $(D_x)^t$ terms. Strictly speaking, if we neglect calculating $p(\mathbf{y}_{:t})$ explicitly, and instead reach the posterior by computing $p(x_t, \mathbf{y}_{:t})$ and then normalising over x_t as $p(x_t|\mathbf{y}_{:t}) = p(x_t, \mathbf{y}_{:t})/\sum_{x_t} p(x_t, \mathbf{y}_{:t})$, we would need to compute an additional sum over x_t. Total terms to sum over would therefore be $(D_x)^t + D_x$.

6.8.3 Recursive inference

If we know $p(x_{t-1}|\mathbf{y}_{:t-1})$, we can write the posterior at step t as

$$p(x_t|\mathbf{y}_{:t}) \overset{\text{AT}}{=} \sum_{x_{t-1}=1}^{D_x} \frac{p(x_t, y_t, x_{t-1}|\mathbf{y}_{:t-1})}{p(y_t|\mathbf{y}_{:t-1})} \tag{96}$$

$$\overset{\text{MA}}{=} \sum_{x_{t-1}=1}^{D_x} \frac{p(x_t|x_{t-1})p(y_t|x_t)p(x_{t-1}|\mathbf{y}_{:t-1})}{p(y_t|\mathbf{y}_{:t-1})}. \tag{97}$$

So to get from $p(x_{t-1}|\mathbf{y}_{:t-1})$ to the posterior at t, we need to sum over a single variable x_{t-1}, which can take one of D_x values. Therefore, there are D_x terms in the sum. As above, we would have to sum over an additional D_x to normalise.

Also, notice that:

$$p(x_{t+1}|y_{1:t}) \overset{\text{AT}}{=} \sum_{x_t} p(x_{t+1}, x_t|y_{1:t}) \tag{98}$$

$$\overset{\text{AT}}{=} \sum_{x_t} p(x_{t+1}|x_t, y_{1:t})p(x_t|y_{1:t}) \tag{99}$$

$$\overset{\text{MA}}{=} \sum_{x_t} p(x_{t+1}|x_t)p(x_t|y_{1:t})\,, \tag{100}$$

which is in the numerator of what was derived above.

6.8.4 Comparison

We are asked to evaluate how many terms we must sum over when we carry out sequential filtering to get $p(x_t|\mathbf{y}_{:t})$. In this case we need to sum over x_{i-1} to calculate the update at each step for $1 < i < t$, i.e. $t-1$ times. We also need to normalise over x_t to get the final posterior. We thus calculate t sums, each with D_x terms, summing over a total of D_x^t terms.

How does this compare with the brute force approach in 6.8.2? Our answers show that using sequential filtering reduces the complexity from exponential to linear in t. This is a massive saving, particularly for large t.

6.9 Question 10 – Finding a Conjugate Prior

We are asked to derive a conjugate prior $p(\theta)$ for a binomial likelihood $p(D|\theta) = \theta^{N_1} \cdot (1-\theta)^{N_0}$. The posterior for a conjugate prior has the same form as the prior, but with different parameters, so we know that

$$p(\theta|D) = \frac{1}{Z} p(\theta) p(D|\theta) \tag{101}$$

$$= \frac{1}{Z} p(\theta) \theta^{N_1} \cdot (1-\theta)^{N_0} \tag{102}$$

shares the form of $p(\theta)$. Z is a normalising constant which, as the question reminds us, we need not evaluate.

We can see that choosing a prior with similar form to the likelihood, we can simply collect the exponents of θ and $1-\theta$ to get the corresponding parameters of the posterior. i.e.

$$p(\theta) = p(\theta; a, b) = \frac{1}{Z_\theta} \theta^a \cdot (1-\theta)^b \tag{103}$$

$$p(\theta|D) = \frac{1}{Z'} \theta^{a+N_1} \cdot (1-\theta)^{b+N_0} = p(\theta; a+N_1, b+N_0) \tag{104}$$

where $Z' = Z \cdot Z_\theta$ and a, b are chosen prior parameters (or hyperparameters). Our conjugate prior is in fact a *beta* distribution, which is more commonly written as

$$p(\theta; \alpha, \beta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{\mathrm{B}(\alpha, \beta)}. \tag{105}$$

$\mathrm{B}(.,.)$ is known as the beta function. By comparing with the Bernoulli likelihood, we can see that $\alpha - 1$ and $\beta - 1$ act as pseudo-counts.

6.10 Question 11 – General Conjugate Prior for Exponential Family

6.10.1 Show that the conjugate prior is given by $p(\eta|\tau, n_0) = H(\tau, n_0) \exp(\tau^\intercal \eta - n_0 A(\eta))$:

A prior is conjugate if the posterior is of the same family as the prior. The given likeliood is

$$\ell(\eta) = p(x|\eta) = h(x) \exp(\eta^\intercal t(x) - A(\eta)), \tag{106}$$

and the prior is

$$p(\eta|\tau, n_0) = H(\tau, n_0) \exp(\tau^\intercal \eta - n_0 A(\eta)). \tag{107}$$

We can find the posterior using

$$p(\eta|x, \tau, n_0) \propto p(\eta|\tau, n_0) p(x|\eta). \tag{108}$$

Computing this quantity is straightforward

$$p(\eta|\tau, n_0) p(x|\eta) = h(x) H(\tau, n_0) \exp(\eta^\intercal t(x) - A(\eta) + \tau^\intercal \eta - n_0 A(\eta)) \tag{109}$$

$$= H'(x, \tau, n_0) \exp((t(x) + \tau)^\intercal \eta - (1+n_0) A(\eta)), \tag{110}$$

where $H'(x, \tau, n_0) = h(x) H(\tau, n_0)$. As this is the same form as the prior, the prior is conjugate.

6.10.2 Put the conjugate prior into the natural parameterisation.

The natural parametrisation of the prior is as follows:

$$p(\eta|\tau, n_0) = \exp\left(\begin{bmatrix} \tau \\ n_0 \end{bmatrix}^\intercal \begin{bmatrix} \eta \\ -A(\eta) \end{bmatrix} - (-\log H(\tau, n_0)) \right). \tag{111}$$

To write this in the natural parametrisation, we need to notice that τ, n_0 are parameters of the prior and can be written with an inner product on functions of η. The left over function of the parameters of the prior $H(\tau, n_0)$ can be taken inside the exponential, hence it can be written as the the log partition function.

6.11 Question 12 – Natural Form

The natural form for an exponential family distribution is

$$p(x|\eta) = h(x)\exp(\eta^\intercal t(x) - A(\eta)). \tag{112}$$

6.11.1 Gamma Distribution

The pdf for a gamma distribution is

$$p(x;\alpha,\beta) = \frac{x^{\alpha-1}\beta^\alpha \exp(-\beta x)}{\Gamma(\alpha)} \tag{113}$$

$$= \exp((\alpha-1)\log x - \beta x - (\log\Gamma(\alpha) - \alpha\log\beta)) \tag{114}$$

Comparing to (112) we see that

$$h(x) = 1 \tag{115}$$

$$\eta(\alpha,\beta) = \begin{bmatrix}\alpha-1\\-\beta\end{bmatrix} = \begin{bmatrix}\eta_1\\\eta_2\end{bmatrix} \tag{116}$$

$$t(x) = \begin{bmatrix}\log x\\ x\end{bmatrix} \tag{117}$$

$$A(\eta) = \log\Gamma(\eta_1+1) - (\eta_1+1)\log(-\eta_2). \tag{118}$$

Equivalently, we could have chosen $\eta_1 = \alpha$, $A(\eta) = \log\Gamma(\eta_1) - \eta_1\log(-\eta_2)$ and $h(x) = 1/x$.

6.11.2 Beta Distribution

Rearranging the pdf for a beta distribution, we have

$$p(x;\alpha,\beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)} \tag{119}$$

$$= \exp((\alpha-1)\log x + (\beta-1)\log(1-x) - \log B(\alpha,\beta)) \tag{120}$$

$$= \exp(-\log x - \log(1-x))\exp(\alpha\log x + \beta\log(1-x) - \log B(\alpha,\beta)) \tag{121}$$

$$= \frac{1}{x(1-x)}\exp(\alpha\log x + \beta\log(1-x) - \log B(\alpha,\beta)). \tag{122}$$

Comparing to (112) we see that

$$h(x) = \frac{1}{x(1-x)} \tag{123}$$

$$\eta(\alpha,\beta) = \begin{bmatrix}\alpha\\\beta\end{bmatrix} = \begin{bmatrix}\eta_1\\\eta_2\end{bmatrix} \tag{124}$$

$$t(x) = \begin{bmatrix}\log x\\ \log(1-x)\end{bmatrix} \tag{125}$$

$$A(\eta) = \log B(\eta_1,\eta_2). \tag{126}$$

6.11.3 Univariate Gaussian Distribution

The pdf for a univariate Gaussian distribution is

$$p(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right). \tag{127}$$

Putting this into natural form, we have

$$p(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}}\exp\left(-\log\sigma - \frac{1}{2\sigma^2}\left(x^2 - 2x\mu + \mu^2\right)\right) \tag{128}$$

$$= \frac{1}{\sqrt{2\pi}}\exp\left(-\left(\frac{\mu^2}{2\sigma^2} + \log\sigma\right) - \frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x\right). \tag{129}$$

Comparing to (112) we see that

$$h(x) = h = \frac{1}{\sqrt{2\pi}} \tag{130}$$

$$\eta(\mu, \sigma) = \begin{bmatrix} \frac{\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2} \end{bmatrix} = \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} \tag{131}$$

$$t(x) = \begin{bmatrix} x \\ x^2 \end{bmatrix} \tag{132}$$

$$A(\eta) = -\left(\frac{\eta_1^2}{4\eta_2} + \frac{1}{2} \log(-2\eta_2) \right). \tag{133}$$

6.12 Question 13 – Exponential Family Conjugate Priors

6.12.1 Bernoulli Likelihood

The likelihood of the probability of success θ for a Bernoulli-distributed random variable is given by

$$p(x|\theta) = \theta^x (1-\theta)^{(1-x)}. \tag{134}$$

Putting this into the exponential family form given in the question

$$p(x|\theta) = \exp \log \left(\theta^x (1-\theta)^{(1-x)} \right) \tag{135}$$

$$= \exp \left(x \log \frac{\theta}{1-\theta} + \log(1-\theta) \right) \tag{136}$$

we see that

$$h(x) = 1 \tag{137}$$

$$\eta = \log \frac{\theta}{1-\theta} \tag{138}$$

$$t(x) = x \tag{139}$$

$$A(\eta) = -\log \left(1 - \frac{e^\eta}{1+e^\eta} \right) = \log(1+e^\eta). \tag{140}$$

Starting from the general exponential family conjugate prior we have shown in the previous question, we can find the Bernoulli conjugate prior

$$p(\eta|\tau, n_0) = H(\tau, n_0) \exp(\tau^\intercal \eta - n_0 A(\eta)) \tag{141}$$

$$p(\theta|\tau, n_0) = H(\tau, n_0) \exp \left(\tau^\intercal \log \frac{\theta}{1-\theta} + n_0 \log(1-\theta) \right) \tag{142}$$

$$= H(\tau, n_0) \theta^\tau (1-\theta)^{n_0 - \tau}. \tag{143}$$

As discussed in Section 6.9, this is a beta distribution, with parameters $\alpha = \tau + 1$ and $\beta = n_0 - \tau + 1$. Its normalising constant is defined by the beta function: $H(\tau, n_0) = 1/B(\alpha, \beta)$.

6.12.2 Gaussian likelihood with unknown precision but known mean

The Gaussian likelihood, for a precision $\nu = 1/\sigma^2$ is

$$p(x|\mu, \nu) = \sqrt{\frac{\nu}{2\pi}} \exp \left(-\frac{\nu}{2} (x-\mu)^2 \right) \tag{144}$$

$$= \sqrt{\frac{\nu}{2\pi}} \exp \left(-\frac{\nu}{2} (x^2 - 2x\mu + \mu^2) \right). \tag{145}$$

The parameters of the exponential family form are

$$h(x) = (2\pi)^{-\frac{1}{2}} \tag{146}$$

$$\eta = \nu \tag{147}$$

$$t(x) = -\frac{1}{2} (x^2 - 2x\mu) \tag{148}$$

$$A(\eta) = \frac{1}{2} (\nu\mu^2 - \log \nu). \tag{149}$$

Note that we have used the fact that μ is known and therefore does not appear in our natural parameters. Starting from the general exponential family conjugate prior, we can write

$$p(\eta|\tau, n_0) = H(\tau, n_0)\exp(\tau^\intercal \eta - n_0 A(\eta)) \tag{150}$$

$$p(\nu|\mu, \tau, n_0) = H(\tau, n_0)\exp\left(\tau\nu - n_0\frac{1}{2}\left(\nu\mu^2 - \log\nu\right)\right) \tag{151}$$

$$p(\nu|\mu, \tau, n_0) = H(\tau, n_0)\exp\left(\nu\left(\tau - \frac{n_0\mu^2}{2}\right)\right)\nu^{\frac{n_0}{2}}. \tag{152}$$

This is in fact an *inverse gamma* distribution on σ^2. Inverse gamma distributions are usually written in the form

$$p(x|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\left(\frac{1}{x}\right)^{\alpha+1}\exp(-\beta/x) \tag{153}$$

where $\Gamma(\alpha)$ is known as the gamma function.

6.13 Question 14 – Multivariate Gaussian Conjugacy

This question addresses multivariate Gaussian distributions: we will be setting all distributions in their natural form so we first tackle how to put in exponential family form the general multivariate Gaussian with mean $\boldsymbol{\mu}$ and covariance matrix Σ. For an N dimensional distribution with likelihood model $p(x; \boldsymbol{\mu}, \Sigma)$, we start by rewriting

$$\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi|\Sigma|)^{\frac{N}{2}}}\exp -\left(\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\intercal\Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu})\right) \tag{154}$$

$$= \frac{1}{(2\pi)^{\frac{N}{2}}}\exp\left(-\frac{1}{2}\mathbf{x}^\intercal\Sigma^{-1}\mathbf{x} + \mathbf{x}^\intercal\Sigma^{-1}\boldsymbol{\mu} - \frac{1}{2}\boldsymbol{\mu}^\intercal\Sigma^{-1}\boldsymbol{\mu} - \frac{N}{2}\log|\Sigma|\right). \tag{155}$$

To write this in the required exponential form, we have to rearrange the term in the exponential so that somehow we have a 'sufficient statistic' $t(x)$ term that is a function of the data dotted with vector η that is some function of the parameters $\boldsymbol{\mu}, \Sigma$. This is straightforward for the term $\mathbf{x}^\intercal\Sigma^{-1}\boldsymbol{\mu}$, and for the term quadratic in $\mathbf{x}$ we can use the following trick. We have

$$\mathbf{x}^\intercal\Sigma^{-1}\mathbf{x} = \sum_{i,j} x_i\Sigma^{-1}_{i,j}x_j \tag{156}$$

$$= \sum_{i,j} x_i\Sigma^{-1}_{i,j}x_j \tag{157}$$

$$= \sum_{i,j} \Sigma^{-1}_{i,j}x_i x_j \tag{158}$$

$$= \overrightarrow{(\mathbf{x}\mathbf{x}^\intercal)}\overrightarrow{(\Sigma^{-1})}, \tag{159}$$

where $\mathbf{x}\mathbf{x}^\intercal$ is a matrix with element i, j equal to $x_i x_j$ and the notation $\overrightarrow{(\cdot)}$ denotes that we are flattening the relevant matrix into a vector (preserving the ordering so that all matrices are flattened in the same way). We have now written the quadratic term in the required form of a dot product (between two N^2 dimensional vectors) where one is a function of $\mathbf{x}$ and the other is in terms of the distribution parameters $\boldsymbol{\mu}, \Sigma$.

Substituting into our previous expression gives

$$\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{\frac{N}{2}}}\exp\left(-\frac{1}{2}\overrightarrow{\mathbf{x}\mathbf{x}^\intercal}\overrightarrow{\Sigma^{-1}} + \mathbf{x}^\intercal\Sigma^{-1}\boldsymbol{\mu} - \frac{1}{2}\boldsymbol{\mu}^\intercal\Sigma^{-1}\boldsymbol{\mu} - \frac{N}{2}\log|\Sigma|\right) \tag{160}$$

$$= \frac{1}{(2\pi)^{\frac{N}{2}}}\exp\left(\begin{bmatrix}\mathbf{x} \\ \overrightarrow{\mathbf{x}\mathbf{x}^\intercal}\end{bmatrix}^\intercal\begin{bmatrix}\Sigma^{-1}\mu \\ -\overrightarrow{\Sigma^{-1}}/2\end{bmatrix} - \left(\frac{1}{2}\boldsymbol{\mu}^\intercal\Sigma^{-1}\boldsymbol{\mu} + \frac{N}{2}\log|\Sigma|\right)\right) \tag{161}$$

and comparing with the standard form of the exponential family $p(\mathbf{x}|\eta) = h(x)\exp(\eta^\intercal t(x) - A(\eta))$, we can now make the immediate identifications $h(x) = (2\pi)^{-\frac{N}{2}}$, $t(x) = \begin{bmatrix}\mathbf{x} & \overrightarrow{\mathbf{x}\mathbf{x}^\intercal}\end{bmatrix}^\intercal$ and $\eta = \begin{bmatrix}\Sigma^{-1}\boldsymbol{\mu} & -\overrightarrow{\Sigma^{-1}}/2\end{bmatrix}^\intercal$.

Finally, we identify $A(\eta) = \frac{1}{2}\boldsymbol{\mu}^\intercal\Sigma^{-1}\boldsymbol{\mu} + \frac{N}{2}\log|\Sigma|$, which we could in principle rewrite as a function of the elements of η to determine $A(\eta)$ explicitly. For this question this implicit form of $A(\eta)$ will suffice.

Applying our general formula to the particular multivariate Gaussians of interest to the question we recover for the likelihood

$$p(\mathbf{y}|\mathbf{f}) = \mathcal{N}\big(\mathbf{y}; \mathbf{f}, \sigma^2\mathrm{I}_N\big) = \frac{1}{(2\pi)^{\frac{N}{2}}}\exp\left(\begin{bmatrix}\mathbf{y}\\ \overrightarrow{\mathbf{y}\mathbf{y}^\intercal}\end{bmatrix}^\intercal\begin{bmatrix}\mathbf{f}/\sigma^2\\ -\mathrm{I}_N/(2\sigma^2)\end{bmatrix} - \frac{1}{2\sigma^2}\mathbf{f}^\intercal\mathbf{f} - \frac{N}{2}\log\sigma^2\right) \tag{162}$$

with $h_l(y) = (2\pi)^{-\frac{N}{2}}$, $\eta_l = \begin{bmatrix}\mathbf{f}/\sigma^2 & -\mathrm{I}_N/(2\sigma^2)\end{bmatrix}^\intercal$, $t_l(y) = \begin{bmatrix}\mathbf{y} & \overrightarrow{\mathbf{y}\mathbf{y}^\intercal}\end{bmatrix}^\intercal$ and implicit $A_l(\eta_l) = \frac{1}{2\sigma^2}\mathbf{f}^\intercal\mathbf{f} - \frac{N}{2}\log\sigma^2$.

Note that in this case the likelihood above could also be simplified as

$$p(\mathbf{y}|\mathbf{f}) = \frac{1}{(2\pi)^{\frac{N}{2}}}\exp\left(-\frac{1}{2\sigma^2}\mathbf{y}^\intercal\mathbf{y} + \frac{1}{\sigma^2}\mathbf{y}^\intercal\mathbf{f} - \frac{1}{2\sigma^2}\mathbf{f}^\intercal\mathbf{f} - \frac{N}{2}\log\sigma^2\right) \tag{163}$$

$$= \frac{1}{(2\pi)^{\frac{N}{2}}}\exp\left(\begin{bmatrix}\mathbf{y}\\ \mathbf{y}^\intercal\mathbf{y}\end{bmatrix}^\intercal\begin{bmatrix}\mathbf{f}/\sigma^2\\ -1/(2\sigma^2)\end{bmatrix} - \frac{1}{2\sigma^2}\mathbf{f}^\intercal\mathbf{f} - \frac{N}{2}\log\sigma^2\right) \tag{164}$$

which would result in an alternative but equivalent parametrisation in exponential family form with lower dimensional η and $t(y)$ vectors.

Moving on to the prior, we recover

$$\mathcal{N}(\mathbf{f}; 0, \mathrm{K}) = \frac{1}{(2\pi)^{\frac{N}{2}}}\exp\left(\begin{bmatrix}\mathbf{f}\\ \overrightarrow{\mathbf{f}\mathbf{f}^\intercal}\end{bmatrix}^\intercal\begin{bmatrix}\vec{0}\\ -\frac{1}{2}\overrightarrow{\mathrm{K}^{-1}}\end{bmatrix} - \frac{N}{2}\log|\mathrm{K}|\right) \tag{165}$$

with corresponding parametrisation $h_p(f) = (2\pi)^{-\frac{N}{2}}$, $\eta_p = \begin{bmatrix}\vec{0} & -\frac{1}{2}\overrightarrow{\mathrm{K}^{-1}}\end{bmatrix}$, $t_p(f) = \begin{bmatrix}\mathbf{f} & \overrightarrow{\mathbf{f}\mathbf{f}^\intercal}\end{bmatrix}^\intercal$ and $A_p(\eta_p) = \frac{N}{2}\log|\mathrm{K}|$.

Now, to ascertain that the two distributions are conjugate, we look at the form of the posterior. By multiplying the two, we see it will be proportional to

$$p(f|y) \propto \exp\left(\begin{bmatrix}\mathbf{y}\\ \overrightarrow{\mathbf{y}\mathbf{y}^\intercal}\end{bmatrix}^\intercal\begin{bmatrix}\mathbf{f}/\sigma^2\\ -\mathrm{I}_N/(2\sigma^2)\end{bmatrix} - \frac{1}{2\sigma^2}\mathbf{f}^\intercal\mathbf{f} - \frac{N}{2}\log\sigma^2 + \begin{bmatrix}\mathbf{f}\\ \overrightarrow{\mathbf{f}\mathbf{f}^\intercal}\end{bmatrix}^\intercal\begin{bmatrix}\vec{0}\\ -\frac{1}{2}\overrightarrow{\mathrm{K}^{-1}}\end{bmatrix} - \frac{N}{2}\log|\mathrm{K}|\right) \tag{166}$$

$$\propto \exp\left(\underbrace{\begin{bmatrix}\mathbf{f}\\ \overrightarrow{\mathbf{f}\mathbf{f}^\intercal}\end{bmatrix}^\intercal}_{t(f)}\underbrace{\begin{bmatrix}\mathbf{y}/\sigma^2\\ -\frac{1}{2}\left(\overrightarrow{\mathrm{K}^{-1}} + \overrightarrow{\mathrm{I}_N/\sigma^2}\right)\end{bmatrix}}_{\eta} - \underbrace{\left(\mathbf{y}^\intercal\mathrm{K}^{-1}\mathbf{y} + \frac{N}{2}(\log\sigma^2 + \log|\mathrm{K}|)\right)}_{A(\eta)}\right) \tag{167}$$

and we have the posterior in exponential family form (so the two distributions are conjugate), with the normalisation constant h that could be found by normalising the expression with respect to $\mathbf{f}$.

6.14 Question 15 – Complete certainty

We assume the following model

$$p(\boldsymbol{\theta}) = \mathcal{N}(\mathbf{0}, \mathbf{I}_M), \tag{168}$$

$$p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\Phi}(\mathbf{X})\boldsymbol{\theta}, \sigma^2\mathbf{I}_N), \tag{169}$$

where $\boldsymbol{\Phi}(\mathbf{X})$ is a $N \times M$ matrix and M is the number of basis functions. The posterior on $\boldsymbol{\theta}$ will be Gaussian as

$$p(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}) \propto p(\mathbf{y}|\boldsymbol{\theta}, \mathbf{X})p(\boldsymbol{\theta}), \tag{170}$$

and both the likelihood and prior are Gaussian, and the likelihood is a linear function of the parameter. Looking at just the variance (you should be able to derive this on your own), we get

$$\Sigma_{\boldsymbol{\theta}|\mathbf{y},\mathbf{X}} = \mathbf{I}_M - \boldsymbol{\Phi}(\mathbf{X})^\intercal\big[\boldsymbol{\Phi}(\mathbf{X})\boldsymbol{\Phi}(\mathbf{X})^\intercal + \sigma^2\mathbf{I}_N\big]^{-1}\boldsymbol{\Phi}(\mathbf{X}). \tag{171}$$

If we have M distinct datapoints, $\boldsymbol{\Phi}(\mathbf{X})$ is an $M \times M$ matrix and is full rank and hence invertible. We take $\sigma^2 \to 0$ as that is the likelihood noise. Then,

$$\Sigma_{\boldsymbol{\theta}|\mathbf{y},\mathbf{X}} = \mathbf{I}_M - \boldsymbol{\Phi}(\mathbf{X})^\intercal(\boldsymbol{\Phi}(\mathbf{X})^\intercal)^{-1}\boldsymbol{\Phi}(\mathbf{X})^{-1}\boldsymbol{\Phi}(\mathbf{X}). \tag{172}$$

$$= \mathbf{0} \tag{173}$$

6.15 Question 16 – Distributions over functions

For an alternative explanation, see Rasmussen and Williams [2006, p. 13].

A scalar function $f : \mathbb{R}^D \to \mathbb{R}$ can be seen as a collection of scalar output variables indexed by some input $\mathbf{x} \in \mathbb{R}^D$. In this case where inputs lie in the vector space $\mathbb{R}^D$, the number of variables in the collection is infinite[1], since there are an infinite number of inputs.[2] We can consider finite subsets of the output variables that correspond to finite subsets of inputs. I.e. for a set of N inputs

$$X = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}, \tag{174}$$

which we can also denote as a matrix $X \in \mathbb{R}^{N \times D}$, we have the corresponding output set

$$f(X) = \{f(\mathbf{x}_1), \dots, f(\mathbf{x}_N)\}, \tag{175}$$

which we alternatively denote as a vector $f(X) \in \mathbb{R}^N$. If the entire function $f(\cdot)$ is a random variable with some distribution, this will also mean that the function outputs $f(X)$ will be a random variable with some distribution.

Going back to the abstract definition of a Gaussian process, we see that its definition as a collection of random variables matches what we found must hold for a function-valued random variable. The collection of random variables that form the Gaussian process can be identified with the random variables of the function outputs at specified inputs. The next part of the definition specifies that a finite subset of the random variables in the Gaussian process must be Gaussian distributed, i.e. for some arbitrary set of inputs X, the corresponding set of outputs $f(X)$ is Gaussian distributed

$$f(\cdot) \sim \mathcal{GP} \implies p(f(X)) = \mathcal{N}\Big(f(X); \boldsymbol{\mu}_{f(X)}, \boldsymbol{\Sigma}_{f(X)}\Big), \tag{176}$$

with some mean $\boldsymbol{\mu}_{f(X)} \in \mathbb{R}^N$ and covariance matrix $\boldsymbol{\Sigma}_{f(X)} \in \mathbb{R}^{N \times N}$. The value of X determines *which* random variables in the collection of the GP we are considering, and therefore must also determine the value of $\boldsymbol{\mu}_{f(X)}$ and $\boldsymbol{\Sigma}_{f(X)}$.

So, long story short: When we use a Gaussian process as a distribution over functions, the random variables in the collection that makes up the GP are used as the *outputs of the function at some input locations*.

6.16 Question 18 – Determining a GP*

True.

6.17 Question 19 – GP prior density*

The mean function $\mu : \mathbb{R}^D \to \mathbb{R}$ and covariance function[3] $k : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}$ determine the mean vector and covariance matrix of the distribution over function values for some input. For example, for the inputs X, we have

$$p(f(X)) = \mathcal{N}\Big(f(X); \boldsymbol{\mu}_{f(X)}, \boldsymbol{\Sigma}_{f(X)}\Big), \tag{177}$$

$$\text{with } \Big[\boldsymbol{\mu}_{f(X)}\Big]_n = \mu(\mathbf{x}_n), \quad \text{and } \big[\boldsymbol{\Sigma}_{f(X)}\big]_{nn'} = k(\mathbf{x}_n, \mathbf{x}_{n'}), \tag{178}$$

where $[\dots]_n$ denotes indexing a vector or matrix. We will often denote the covariance matrix of a common set of inputs (e.g. training inputs X) as $\mathbf{K}$. Alternatively, we can denote the matrix resulting from evaluating the covariance function at two sets of inputs $X \in \mathbb{R}^{N \times D}, X^* \in \mathbb{R}^{N^* \times D}$ as

$$K(X, X^*) \in \mathbb{R}^{N \times N^*}. \tag{179}$$

[1] Or more specifically, *uncountably* infinite, but for the purpose of this course this is a detail of minor importance.

[2] In the course, we will consider only inputs in $\mathbb{R}^D$. However you can have functions on many different types of inputs, leading to different sizes of the set of output variables. For example, you could have a function from a finite set: $\{0, 1, 2, 3, 4\} \to \mathbb{R}$.

[3] The covariance function is often referred to as the *kernel*.

6.18 Question 20 – GP posterior*

In inference problems, the goal is to go from a joint probability distribution over all relevant variables (which comes from the model specification) to a conditional distribution given observations (the posterior). The posterior is obtained by manipulating the joint using the rules of probability. Often Bayes' rule is enough, but sometimes other applications of the product or sum rules are necessary [MacKay, 2003, p. 24]. I will go through this derivation multiple ways, and in quite a lot of detail. I will pay special attention to the manipulations of the probability distributions on the way, as it is important to be familiar with how to do this. If you struggle with this, I would recommend looking over chapters 1–3 of MacKay [2003], and perhaps doing a few exercises.

6.18.1 Understanding and manipulating the joint

The joint distribution is usually specified through a set of conditional probability distributions. For Gaussian process regression, the problem specification gives:

1. The *prior* distribution over the function values at any inputs, through the definition of the prior over functions being a GP. For a set of inputs X, we can refer to the prior density over corresponding function values as $p(f(X))$.
2. The conditional distribution of any observation y_n at an input location $\mathbf{x}_n$, conditioned on full knowledge of the underlying function value. This distribution is known as the *likelihood*, and is written as $p(y_n \mid f(\mathbf{x}_n), \mathbf{x}_n)$ in this case, and has the density $\mathcal{N}\big(y_n; f(\mathbf{x}_n), \sigma^2\big)$.

We need to consider the function values at both the training points $f(X)$ and testing points $f(X^*)$, in addition to the outputs $\mathbf{y}$. This gives us the joint distribution $p(f(X), f(X^*), \mathbf{y} \mid X)$. We can factorise the joint along the conditional independence relationships that the model was specified with:

$$p(f(X), f(X^*), \mathbf{y} \mid X) = p(f(X), f(X^*))p(\mathbf{y} \mid f(X), X) \tag{180}$$

$$= p(f(X), f(X^*)) \prod_{n=1}^{N} p(y_n \mid f(\mathbf{x}_n), \mathbf{x}_n)\,. \tag{181}$$

We can also go one step further and factorise the prior using eqs. (6) and (7):

$$p(f(X), f(X^*)) = \mathcal{N}\left(\begin{bmatrix} f(X) \\ f(X^*) \end{bmatrix}; 0, \begin{bmatrix} \mathbf{K_{ff}} & \mathbf{K_{ff^*}} \\ \mathbf{K_{ff^*}^\intercal} & \mathbf{K_{f^*f^*}} \end{bmatrix}\right) \tag{182}$$

$$\implies p(f(X^*) \mid f(X)) = \mathcal{N}\big(f(X^*); \mathbf{K_{ff^*}^\intercal K_{ff}^{-1}} f(X), \quad \mathbf{K_{f^*f^*}} - \mathbf{K_{ff^*}^\intercal K_{ff}^{-1} K_{ff^*}}\big)\,, \tag{183}$$

where $\mathbf{K_{ff}} = k(X, X)$, $\mathbf{K_{ff^*}} = k(X, X^*)$, and $\mathbf{K_{f^*f^*}} = k(X^*, X^*)$. This gives an overall factorised joint of

$$p(f(X), f(X^*), \mathbf{y} \mid X) = p(f(X))p(f(X^*) \mid p(f(X))) \prod_{n=1}^{N} p(y_n \mid f(\mathbf{x}_n), \mathbf{x}_n) \tag{184}$$

This conditional independence relationship can be represented as a graphical model, which often helps in reasoning about the model (fig. 3)

Figure 3: Graphical model representations of Gaussian process regression, with conditional independence of $\mathbf{y}$ represented explicitly on the right.

We are tasked with finding the posterior over function values at an arbitrary set of input locations X^*, given that we observe a training set consisting of N inputs and corresponding outputs $\{\mathbf{x}_n, y_n\}_{n=1}^N$. We arrange the inputs into a matrix $X \in \mathbb{R}^{N \times D}$, and the outputs into a vector $\mathbf{y} \in \mathbb{R}^N$.

6.18.2 Method 1: Starting with the appropriate joint

One representation of Bayes' rule that we can apply is the joint distribution divided by the conditional. In our case this becomes

$$p(f(X^*) \,|\, \mathbf{y}, X) = \frac{p(\mathbf{y}, f(X^*) \,|\, X)}{p(\mathbf{y} \,|\, X)} \,. \tag{185}$$

However, the model specification did not give us the density of $p(\mathbf{y}, f(X^*) \,|\, X)$, so we will have to find it ourselves. We can find it by marginalising (or *integrating out*) the unused variable $f(X)$ from the joint using the sum rule:

$$p(\mathbf{y}, f(X^*)) = \int p(\mathbf{y}, f(X), f(X^*) \,|\, X) \mathrm{d}f(X) \,. \tag{186}$$

We know the joint is a Gaussian, as the conditionals are all Gaussians which only interact linearly. This makes the marginal Gaussian too. We could perform the integration by hand, but given the knowledge that $p(\mathbf{y}, f(X^*) \,|\, X)$ is Gaussian, we could also just compute its means and covariances. The means are all zero. The main useful insight into finding the covariances, is that the Gaussian likelihood can be rewritten as

$$\mathbf{y} = f(X) + \boldsymbol{\epsilon} \,, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}\big(0, \sigma^2 \mathbf{I}_N\big) \,. \tag{187}$$

This makes the covariances

$$\begin{aligned}
\mathrm{Cov}[\mathbf{y}, f(X^*)] &= \mathbb{E}_{f,\boldsymbol{\epsilon}}[(f(X) + \boldsymbol{\epsilon}) f(X^*)^\intercal] = \mathbb{E}_{f,\boldsymbol{\epsilon}}[f(X) f(X^*)^\intercal] + \cancelto{0}{\mathbb{E}_{f,\boldsymbol{\epsilon}}[\boldsymbol{\epsilon} f(X^*)^\intercal]} && (188)\\
&= k(X, X^*) = \mathbf{K}_{\mathbf{f}\mathbf{f}^*} \,, && (189)\\
\mathrm{Cov}[\mathbf{y}, \mathbf{y}] &= \mathbb{E}_{f,\boldsymbol{\epsilon}}\big[(f(X) + \boldsymbol{\epsilon})(f(X) + \boldsymbol{\epsilon})^\intercal\big] = k(X, X) + \sigma^2 \mathbf{I} = \mathbf{K}_{\mathbf{f}\mathbf{f}} + \sigma^2 \mathbf{I} \,, && (190)\\
\mathrm{Cov}[f(X^*), f(X^*)] &= k(X^*, X^*) = \mathbf{K}_{\mathbf{f}^*\mathbf{f}^*} \,. && (191)
\end{aligned}$$

The posterior can now be found by simple Gaussian conditioning, using eqs. (6) and (7):

$$\begin{aligned}
p(\mathbf{y}, f(X^*) \,|\, X) &= \mathcal{N}\left(\begin{bmatrix} \mathbf{y} \\ f(X^*) \end{bmatrix}; 0, \begin{bmatrix} \mathbf{K}_{\mathbf{f}\mathbf{f}} + \sigma^2 \mathbf{I}_N & \mathbf{K}_{\mathbf{f}\mathbf{f}^*} \\ \mathbf{K}_{\mathbf{f}\mathbf{f}^*}^\intercal & \mathbf{K}_{\mathbf{f}^*\mathbf{f}^*} \end{bmatrix}\right) && (192)\\
\implies p(f(X^*) \,|\, \mathbf{y}, X) &= \mathcal{N}\big(f(X^*); \mathbf{K}_{\mathbf{f}\mathbf{f}^*}^\intercal (\mathbf{K}_{\mathbf{f}\mathbf{f}} + \sigma^2 \mathbf{I})^{-1} \mathbf{y}, \quad \mathbf{K}_{\mathbf{f}^*\mathbf{f}^*} - \mathbf{K}_{\mathbf{f}\mathbf{f}^*}^\intercal (\mathbf{K}_{\mathbf{f}\mathbf{f}} + \sigma^2 \mathbf{I})^{-1} \mathbf{K}_{\mathbf{f}\mathbf{f}^*}\big) \,. && (193)
\end{aligned}$$

This completes the exercise.

6.18.3 Method 2: Finding a larger posterior then marginalising

Alternatively, we may want to apply Bayes' rule directly. However, our joint also contains $f(X)$, which we weren't interested in representing in our posterior. We can also solve the problem by applying Bayes' rule to find the posterior $p(f(X), f(X^*) \,|\, \mathbf{y}, X)$, and then marginalising out $f(X)$. We start by applying Bayes' rule to the usual prior-likelihood factorisation of the joint:

$$p(f(X), f(X^*) \,|\, \mathbf{y}, X) = \frac{p(\mathbf{y} \,|\, f(X), X) p(f(X), f(X^*))}{p(\mathbf{y} \,|\, X)} \,. \tag{194}$$

The first thing we can notice is that using the factorisation of the prior from eq. (183), we can actually write the posterior over both training and testing points in terms of the posterior over only the training points:

$$\begin{aligned}
p(f(X), f(X^*) \,|\, \mathbf{y}, X) &= \frac{p(\mathbf{y} \,|\, f(X), X) p(f(X))}{p(\mathbf{y} \,|\, X)} p(f(X^*) \,|\, f(X)) && (195)\\
&= p(f(X) \,|\, \mathbf{y}) p(f(X^*) \,|\, f(X)) && (196)
\end{aligned}$$

We then marginalise out $f(X)$ to get the posterior we are interested in:

$$p(f(X^*) \,|\, \mathbf{y}) = \int p(f(X) \,|\, \mathbf{y}) p(f(X^*) \,|\, f(X)) \mathrm{d}f(X) \,. \tag{197}$$

Performing these computations requires exactly the same procedure of forming the joint and performing Gaussian conditioning (eqs. (6) and (7)) as in the previous way. Once the joint $p(f(X), f(X^*) \,|\, \mathbf{y})$ is obtained, marginalisation can be performed simply by dropping the rows and columns related to $f(X)$ from the covariance. Overall the process is a bit more laborious than the other method.[4]

[4] However, it does give us the nice insight that the only part of the posterior that *really* changes due to the data is $p(f(X) \,|\, \mathbf{y})$, and that any change in the posterior on $f(X^*)$ only occurs *through* the effect on $f(X)$.

6.19 Question 21 – Marginalisation

The Gaussian process prior defines a joint distribution over all finite subsets of the infinite set of function values. This raises the question of *which* finite subset we need to represent in order to do inference. We know that we at least need to represent $f(X)$ for the training data X, since these function values are involved in the likelihood. If we want to make predictions over some new subset of input points X^*, these need to be involved in inference as well. The question remains, do we need to represent any other points?

We only need to represent variables that influence the inference over variables we are interested in. To consider whether we need to represent any other function values, we can include them in our posterior analysis, integrate them out, and see whether they influence our beliefs in what we are interested in. We can do this by repeating the derivation in section 6.18.3, but with extra function values $f(\tilde{X})$ included.

$$p(f(X), f(X^*), f(\tilde{X}) \mid \mathbf{y}, X) = \frac{p(\mathbf{y} \mid f(X), X)p(f(X))}{p(\mathbf{y} \mid X)} p(f(X^*) \mid f(X))p(f(\tilde{X}) \mid f(X), f(X^*)) \tag{198}$$

$$= p(f(X) \mid \mathbf{y})p(f(X^*) \mid f(X))p(f(\tilde{X}) \mid f(X), f(X^*)) \tag{199}$$

By integrating out $f(\tilde{X})$ we see that the posterior given by this procedure is the same as the posterior which we obtained without considering the extra points $f(\tilde{X})$. Because this holds for *any* set of other points $\tilde{X}$, this shows that including any other set of points doesn't change the outcome of our inference procedure, and therefore we don't have to include them in the first place.

If you look at the derivation, you actually see that the fact that the prior over function values defines consistent distributions over function values. The distributions are consistent in the sense that if we consider the prior joint $p(f(X), f(X^*))$ and integrate out $f(X^*)$, then we get the same distribution as if we had considered the prior $p(f(X))$ directly. This is a direct consequence of a Gaussian process being a distribution over functions.

It is, of course, possible to define an infinite collection of probability densities that do not have this property. For example, defining the precision of a Gaussian using a covariance function. See Rasmussen and Williams [2006, §2.2] for another explanation.

6.20 Question 22 – Predicting observations*

An observation y_n is modelled by the function value $f(\mathbf{x}_n)$ with some noise added. This question is asking for a posterior distribution over a different random variable than question 20 (i.e. y_n, rather than $f(\mathbf{x}_n)$). Once you have the posterior of $f(\mathbf{x}_n)$, you can use the usual manipulations of Gaussian distributions to find out that you just need to add $\sigma^2\mathbf{I}_{N^*}$ to the covariance, where N^* is the number of test points. You could also apply Bayes' rule to the joint distribution

$$p(\mathbf{y}^* \mid \mathbf{y}, X, X^*) = \frac{p(\mathbf{y}, \mathbf{y}^* \mid X, X^*)}{p(\mathbf{y} \mid X)} \tag{200}$$

$$= \mathcal{N}\left(\begin{bmatrix} \mathbf{y} \\ \mathbf{y}^* \end{bmatrix}; 0, \begin{bmatrix} \mathbf{K}_{\mathbf{ff}} + \sigma^2\mathbf{I}_N & \mathbf{K}_{\mathbf{ff}^*} \\ \mathbf{K}_{\mathbf{ff}^*}^\intercal & \mathbf{K}_{\mathbf{f}^*\mathbf{f}^*} + \sigma^2\mathbf{I}_{N^*} \end{bmatrix} \right), \tag{201}$$

as in section 6.18.2.

6.21 Question 23 – Weights to covariances*

In the question, a function is defined through a specific setting of the parameters. By making the parameters a random variable, we also make the function a random variable. By noting that we place Gaussian distributions on the parameters, and the relationship between the function value and parameters is linear, we see that the distribution over function values is also Gaussian. We can find the covariance function by simply finding the covariance between the function at two points. We first note that the mean function is zero by taking the mean.

$$k(\mathbf{x}, \mathbf{x}') = \mathbb{E}_{\boldsymbol{\theta}_1, \theta_2}[f(\mathbf{x})f(\mathbf{x}')] = \mathbb{E}_{\boldsymbol{\theta}_1, \theta_2}\left[(\mathbf{x}^\intercal\boldsymbol{\theta}_1 + \theta_2)(\mathbf{x}'^\intercal\boldsymbol{\theta}_1 + \theta_2)\right] \tag{202}$$

$$= \mathbf{x}^\intercal\mathbf{x}'v_1 + v_2 \tag{203}$$

6.22 Question 24 – Sums of GPs*

The prior over any set of points $f(X)$ is the sum of two independent random variables. Therefore, variances add.

$$f(X) = f_1(X) + f_2(X)\,, \qquad f_1(X) \sim \mathcal{N}(0, \boldsymbol{k}_1(X,X))\,, \qquad f_2(X) \sim \mathcal{N}(0, k_2(X,X))\,, \tag{204}$$

$$p(f(X)) = \mathcal{N}(0, k_1(X,X) + k_2(X,X)) \tag{205}$$

Now, we want the posterior over $f_1(\cdot)$. We follow the same procedure as in question 20, by first finding the joint over all variables of interest, and then using the Gaussian conditioning formula.

$$\begin{aligned}\mathrm{Cov}[f(X), f_1(X^*)] &= \mathbb{E}[(f_1(X) + f_2(X))f_1(X^*)^\intercal] = \mathbb{E}[f_1(X)f_1(X^*)^\intercal] + \cancelto{0}{\mathbb{E}[f_1(X)f_2(X^*)^\intercal]} \\ &= k_1(X, X^*) \end{aligned} \tag{206}$$

$$p\left(\begin{bmatrix} f(X) \\ f_1(X^*)\end{bmatrix}\right) = \mathcal{N}\left(\begin{bmatrix} f(X) \\ f_1(X^*)\end{bmatrix}; \mathbf{0}, \begin{bmatrix} k_1(X,X) + k_2(X,X) & k_1(X,X^*) \\ k_1(X^*,X) & k_1(X^*,X^*)\end{bmatrix}\right) \tag{207}$$

$$\begin{aligned}\implies p(f_1(X^*) \,|\, f(X)) = \mathcal{N}\big(&f_1(X); k_1(X^*,X)[k_1(X,X) + k_2(X,X)]^{-1} f(X), \\ &k_1(X^*,X^*) - k_1(X^*,X)[k_1(X,X) + k_2(X,X)]^{-1} k(X,X^*)\big)\end{aligned} \tag{208}$$

6.23 Question 25 – Hyperparameter Conjugacy

6.23.1 Can you find the posterior on the hyperparameters $p(\boldsymbol{\theta}|\mathbf{y})$ in closed form?

No, because the kernel (covariance) depends on l through some complicated non-linear function which we can't integrate over. Hence, we can't define a valid density in closed form that has the same functional form as a covariance matrix of a normal distribution with the non linearities imposed by the kernel.

6.23.2 When can we find a posterior in closed form?

To find the posterior of the hyperparameters we can write, using Bayes rule, (leaving out dependence on $\mathbf{X}$ for this question)

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{y})} \tag{209}$$

where $p(\mathbf{y}|\boldsymbol{\theta}) = \int p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\boldsymbol{\theta})d\boldsymbol{\theta}$ and $p(\mathbf{y}) = \int p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}$. We know that $p(\mathbf{y}|\boldsymbol{\theta})$ is the marginal likelihood, and if we can find the posterior in closed form, we can compute the normaliser above $p(\mathbf{y})$. So, we ignore the normaliser for now. The given GP prior is

$$p(\mathbf{f}|\boldsymbol{\theta}) = \mathcal{N}(\mathbf{0}, \mathbf{K}_{SE}), \tag{210}$$

$$K_{SE}(x, x') = \lambda^2 \exp\left(-\frac{(x-x')^2}{2l^2}\right). \tag{211}$$

This will give us a marginal likelihood of

$$p(\mathbf{y}|\boldsymbol{\theta}) = (2\pi)^{-n/2}|\mathbf{K}_{SE} + \sigma_n^2\mathbf{I}|^{-1/2} \exp\left(-\frac{1}{2}\mathbf{y}^\intercal(\mathbf{K}_{SE} + \sigma_n^2\mathbf{I})^{-1}\mathbf{y}\right) \tag{212}$$

where we have assumed $p(\mathbf{y}|\mathbf{f}) = \mathcal{N}(\mathbf{f}, \sigma_n^2\mathbf{I})$, and $\mathbf{y} \in \mathbb{R}^n$. We can see that this is a Gaussian with covariance $\mathbf{K}_{SE} + \sigma_n^2$. We know from conjugacy of Gaussian distributions that the conjugate prior for the mean of a multivariate Gaussian is a multivariate Gaussian. The conjugate prior for a covariance matrix $\boldsymbol{\Sigma}$ is an inverse Wishart distribution (multivariate form of the inverse Gamma)

$$p(\boldsymbol{\Sigma}|\boldsymbol{\Psi}, \nu) = \frac{|\boldsymbol{\Psi}|^{\nu/2}}{2^{\nu d/2}\Gamma_d(\frac{\nu}{2})}|\boldsymbol{\Sigma}|^{-(\nu+d+1)/2} \exp\left(-\frac{1}{2}\mathrm{Tr}(\Psi\boldsymbol{\Sigma}^{-1})\right), \tag{213}$$

where $\boldsymbol{\Psi} \in \mathbb{R}^{d\times d}$ is a scale matrix and Γ_d is a multivariate Gamma function. It is easy to see why this is conjugate to the covariance function, it has a term that matches the inverse square root of the determinant $|\boldsymbol{\Sigma}|^{-(\nu+d+1)/2}$, and a term that matches the inverse in the exponential $\exp\big(-\frac{1}{2}\mathrm{Tr}(\Psi\boldsymbol{\Sigma}^{-1})\big)$. Thus, if we put an inverse Wishart prior on the covariance of the marginal likelihood, $\boldsymbol{\Sigma} = \mathbf{K}_{SE} + \sigma_n^2\mathbf{I}$, then we could

derive the posterior in closed form. However, if we wish to put a prior on the hyperparameters of $\mathbf{K}_{SE}$, λ and l, the conjugate prior needs to take into account the form of the kernel $\mathbf{K}_{SE}$. For λ, if we have $\sigma_n^2 = 0$, we can see that it essentially is a variance term, and the conjugate prior would be the inverse Gamma. For l, with $\sigma_n^2 = 0$, we would need a distribution that has a term with a determinant of the exponential of the inverse random variable, as well as a term that had an exponential of an exponential of the random variable. As there is no tractable distribution (that we know of) with these properties, we can't find the posterior for l in closed form.

6.24 Question 26 – Sparse approximation*

6.24.1 Gaussian input density

With probability 1, no two inputs will ever be equal, so the covariance matrix will be equal to the identity:

$$k(X, X) = \boldsymbol{I} \,. \tag{214}$$

In this case, all eigenvalues equal 1, so there is no eigenvalue decay. Low-rank approximations are accurate when some eigenvalues are much smaller than others. Given that this is not the case here, we cannot get an accurate approximation.

6.24.2 Discrete distribution

There will be only 5 input points that can be drawn. So if $N \geq 5$, then $k(X, X)$ will have repeated rows and columns. The maximum rank of $k(X, X)$ is 5, with 5 eigenvalues of 1, and the rest being zero. We therefore can find an exact rank 5 approximation.

6.25 Question 27 – Limitations of stationarity*

6.26 Question 28 – Gold prospecting*

6.26.1 No exploratory dig

If we do not do an exploratory dig, our utility function is simply the true profit of the mine that we pick:

$$U(\mathbf{x}, n) = x_n \,. \tag{215}$$

To find the expected utility, we take the expectation of all unknown quantities under the distributions that represent our belief about them. In this case, this is simply the prior:

$$\mathbb{E}_{p(\mathbf{x})}[U(\mathbf{x}, n)] = \mathbb{E}_{p(\mathbf{x}_n)}[x_n] = \mu_n \,. \tag{216}$$

The action we take (i.e. which site we choose to mine, indicated by m) maximises the expected utility, and is therefore

$$m = \underset{n}{\operatorname{argmax}}\, \mu_n \,, \tag{217}$$

i.e. we pick the site with the largest prior mean.

6.26.2 Exploratory dig

Here, we will apply the 3 steps of decision theory from lectures to the exploratory dig scenario. We will see that applying the principles can require a bit more thought in this complicated situation, but that the principles remain the same.

If we do choose to do an exploratory dig, we have a more complicated utility function, because we have two actions to take. We need to choose the site p of where to prospect, and m of where to mine. The outcome utility becomes

$$U = -c_p + x_m \,. \tag{218}$$

In the way the utility is written now, we have two actions that we could optimise over. However, this does not take into account that 1) we take the action to decide where to prospect first, and 2) the choice of where to mine can depend on where we prospected and what data we observed. We also know that

after prospecting at site p and gaining the data d_p, we will choose where to mine using the same utility maximisation principle as earlier. So, we can write:

$$m(d_p) = \underset{n}{\operatorname{argmax}}\, \mathbb{E}_{p(x_n \mid d_p)}[x_n] = \underset{n}{\operatorname{argmax}}\, \mu'_n(d_p)\,. \tag{219}$$

The location that we mine at m and the mean of our posterior belief of its value μ'_n are now functions of the data we observe after prospecting d_p. We write these quantities as explicit functions of d_p to highlight this dependence. We substitute this into the utility to obtain

$$U(\mathbf{x}, p, d_p) = -c_p + x_{m(d_p)}\,. \tag{220}$$

We can now take expectations over all random variables:

$$\begin{aligned}\mathbb{E}_{p(\mathbf{x},d_p)}[U(\mathbf{x}, p, d_p)] &= -c_p + \mathbb{E}_{p(d_p)}\left[\mathbb{E}_{p(\mathbf{x} \mid d_p)}\left[x_{m(d_p)}\right]\right] &(221)\\ &= -c_p + \mathbb{E}_{p(d_p)}\left[\max_n \mu'_n(d_p)\right]. &(222)\end{aligned}$$

We can simplify the expectation over the max by noticing that for $n \neq p$, $\mu'_n = \mu_n$, i.e. the posterior mean of the profit only changes for the site that we prospect at p. We can therefore write

$$\mathbb{E}_{p(\mathbf{x},d_p)}[U(\mathbf{x}, p, d_p)] = -c_p + \mathbb{E}_{p(d_p)}\left[\max\left\{\mu'_p(d_p), \mu_1\right\}\right], \tag{223}$$

where

$$\mu_1 = \max_{n \neq p} \mu_n\,, \tag{224}$$

i.e. the highest prior mean of all sites except the one we prospect at. Since μ'_p is now the only quantity which depends on d_p, we can alternatively write the expectation over the distribution of μ'_p. We find this distribution by

$$\begin{aligned}p(x_p \mid d_p) = \frac{p(d_p \mid x_p)p(x_p)}{p(d_p)} &= \mathcal{N}\left(x_p; \underbrace{\frac{d_p\sigma_p^2 + \mu_p\sigma^2}{\sigma^2 + \sigma_p^2}}_{\mu'_p}, \left(\frac{1}{\sigma^2} + \frac{1}{\sigma_p^2}\right)^{-1}\right), &(225)\\ p(d_p) &= \mathcal{N}\left(d_p; \mu_p, \sigma^2 + \sigma_p^2\right), &(226)\\ \implies p(\mu'_p) &= \mathcal{N}\left(\mu'_p; \mu_p, \sigma_p^2\frac{\sigma_p^2}{\sigma^2 + \sigma_p^2}\right), &(227)\end{aligned}$$

which can all be derived by applying Gaussian identities (you should know how to do this). We can now write the expectation as

$$\begin{aligned}\mathbb{E}_{p(\mathbf{x},d_p)}[U(\mathbf{x}, p, d_p)] &= -c_p + \mathbb{E}_{p(\mu'_p)}\left[\max\left\{\mu'_p, \mu_1\right\}\right] &(228)\\ &= -c_p + \left(\int_{-\infty}^{\mu_1} p(\mu'_p)\mathrm{d}\mu'_p\right)\mu_1 + \int_{\mu_1}^{\infty} p(\mu'_p)\mu'_p\mathrm{d}\mu'_p &(229)\\ &= -c_p + \underbrace{\left(P(\mu'_p < \mu_1)\right)\mu_1}_{\text{term 1}} + \underbrace{\int_{\mu_1}^{\infty} p(\mu'_p)\mu'_p\mathrm{d}\mu'_p}_{\text{term 2}}\,. &(230)\end{aligned}$$

We have now calculated the expected utility given the action of prospecting at site p, and it is useful to take a time to interpret the terms. **Term 1** gives us the contribution of what happens when the data tells us that the prospected site is worse than the best alternative. In this case, we choose the best alternative, and gain an expected return of μ_1. **Term 2** gives us the contribution if the data tells that the prospected site is best. We choose site p, and gain a return that is described by our belief over how good the site is $p(\mu'_p)$.

We now have computed the utility *given that we make the choice to prospect at site p.* To properly solve the problem, we need to choose whether or not we should prospect, and if we choose to, which site p to prospect at. To decide this, we calculate the quantity Δ_p, which is the difference in expected utility

between prospecting at site p, and not prospecting at all. If this quantity is negative for all sites, we do not prospect, otherwise we choose p for which Δ_p is maximum.

$$\Delta_p = -c_p + \int_{-\infty}^{\infty} p(\mu'_p)\big[\max\{\mu'_p, \mu_1\} - \max\{\mu_p, \mu_1\}\big]\mathrm{d}\mu'_p \tag{231}$$

$$= -c_p + \begin{cases} \displaystyle\int_{\mu_1}^{\infty} p(\mu'_p)\big(\mu'_p - \mu_1\big)\mathrm{d}\mu'_p & \text{if } \mu_1 \geq \mu_p \\ \displaystyle\int_{-\infty}^{\mu_1} p(\mu'_p)\big(\mu_1 - \mu'_p\big)\mathrm{d}\mu'_p & \text{if } \mu_1 < \mu_p \end{cases} \tag{232}$$

It is also useful to understand both cases in this last equation:

Case 1) computes the expected improvement in return for prospecting at site p, if we previously *would not* have chosen it. This happens if $\mu_1 \geq \mu_p$.

Case 2) computes the expected improvement in return for prospecting at site p, if we otherwise *would* have chosen it. This happens if $\mu_1 < \mu_p$.

In both cases, the terms compute *the value of being able to switch*. This needs to outweigh the cost c_p in order for Δ_p to be great than zero, and prospecting to be worth it. Figure 4 plots Δ_p as a function of the difference in prior mean of the candidate prospecting site and the best alternative $\mu_n - \mu_1$, and the uncertainty of the return for the candidate site σ_n. Note how uncertainty increases the value of exploration, and that larger uncertainties are needed to make exploration worth it for larger deviations in the mean.

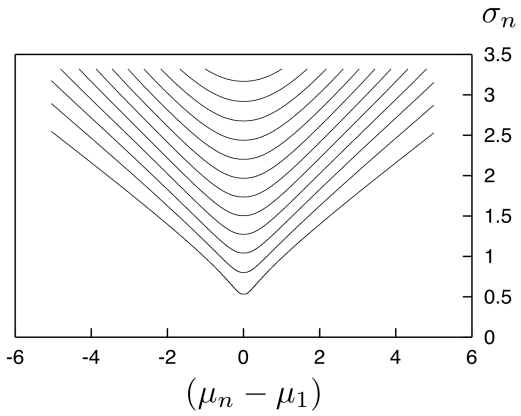


Figure 4: Contour plot for Δ_n, with values increasing towards the top. Reproduced from MacKay [2003, ch. 36].

6.27 Question 29 – Timing*

No, in both cases the mining location is fixed *before* observing the data from the prospecting. This means that the utility is not a function of the data. For both cases, the expected utility is

$$\mathbb{E}_{p(\mathbf{x})}[U(\mathbf{x}, p, m)] = \mathbb{E}_{\mathbf{x}}[-c_p + x_m] = -c_p + \mu_m \tag{233}$$

where the answer to the first part of the question has $m = p$.

This shows that the difficulty (and the benefit!) comes from the fact that future decisions are a function of data that will be observed in the future. Because we have beliefs over what this future data will be, we can still take the expectations.

6.28 Question 30 – Probability of improvement*

First, let's compute the probability that we would actually make a switch based on the data obtained from prospecting. The site we prospect at p is the only one that has a posterior mean μ'_p which is affected

by the data from the prospecting d_p. We denote the highest prior mean of all other sites as μ_1 (eq. (224)). In this explanation, we only consider the case that $\mu_1 > \mu_p$, as the opposite case is similar.

Remember eq. (219), which states that we choose the site to mine based on the highest posterior mean return given the data. For the case that $\mu_1 > \mu_p$, we will only switch our decision to site p if $\mu'_p > \mu_1$. We calculate the probability of this happening using the same $p(\mu'_p)$ as before:

$$\begin{aligned} P(\mu'_p > \mu_1) &= \int_{\mu_1}^{\infty} p(\mu'_p)\mathrm{d}\mu'_p = \int_{\mu_1}^{\infty} \mathcal{N}\left(\mu'_p; \mu_p, \sigma_p^2 \frac{\sigma_p^2}{\sigma^2 + \sigma_p^2}\right)\mathrm{d}\mu'_p \\ &= 1 - \int_{-\infty}^{\mu_1} \mathcal{N}\left(\mu'_p; \mu_p, \sigma_p^2 \frac{\sigma_p^2}{\sigma^2 + \sigma_p^2}\right)\mathrm{d}\mu'_p \\ &= 1 - \Phi\left(\frac{\mu_1 - \mu_p}{\sigma_p^2}\sqrt{\sigma^2 + \sigma_p^2}\right) = 1 - \Phi\left(\frac{\mu_1 - \mu_p}{s}\right), \end{aligned} \tag{234}$$

where

$$s^2 = \sigma_p^2 \frac{\sigma_p^2}{\sigma^2 + \sigma_p^2}, \qquad \text{and} \qquad \Phi(x) = \int \mathcal{N}(x';,0,1)\mathrm{d}x' . \tag{235}$$

We can compare this to the expected improvement Δ_p for the case that $\mu_1 > \mu_p$ (from eq. (232)):

$$\Delta_p = -c_p + \int_{\mu_1}^{\infty} p(\mu'_p)(\mu'_p - \mu_1)\mathrm{d}\mu'_p \tag{236}$$

$$= -c_p + \int_{\mu_1}^{\infty} \mathcal{N}(\mu'_p; \mu_p, s^2)(\mu'_p - \mu_1)\mathrm{d}\mu'_p . \tag{237}$$

Using the equation for the mean of a truncated Gaussian [Wikipedia, 2020], we can find the solution to the integral (keep in mind that in our integral we do not normalise the truncated Gaussian):

$$\Delta_p = -c_p + \left(1 - \Phi\left(\frac{\mu_1 - \mu_p}{s}\right)\right)(\mu_p - \mu_1) + \phi\left(\frac{\mu_1 - \mu_p}{s}\right)s , \tag{238}$$

where we write the standard Gaussian density as $\phi(\cdot)$.

We have now computed the probability of improvement (eq. (234)), as well as the explicit form for the expected improvment (eq. (238)). We want to

- see how the probability of improvement changes,
- for constant expected improvement,
- as $\mu_p - \mu_1$ increases (and therefore σ_p increases as well).

Since it is difficult to see how σ_p needs to change to keep the expected improvement contant, we instead notice that keeping $(\mu_1 - \mu_p)/s$ constant keeps the probability of improvement constant. This implies

$$\frac{\mu_1 - \mu_p}{s} = \frac{(\mu_1 - \mu_p)\sqrt{\sigma^2 + \sigma_p^2}}{\sigma_p^2} = c = \text{const} \qquad \Longrightarrow \qquad \sigma_p^2 = \frac{\sigma^2(\mu_1 - \mu_p)^2}{c^2 - (\mu_1 - \mu_p)^2} . \tag{239}$$

By looking at eq. (238), we see that for a constant $(\mu_1 - \mu_p)/s$, with increasing $\mu_1 - \mu_p$, the expected improvement monotonically increases. This implies that if we instead keep the expected improvement fixed while increasing $\mu_1 - \mu_p$, then the ratio $(\mu_1 - \mu_p)/s$ would have to decrease.

So to summarise, a large difference in means $\mu_1 - \mu_p$ can still lead to a given expected improvement if σ_p^2 is large enough. However, for larger gaps in means, the probability that we will see an improvement becomes smaller.

6.29 Question 31 – Single observation*

We are given a Gaussian process posterior for N data points in $(X, \mathbf{y})$

$$p(f \,|\, \mathbf{y}, X) = \mathcal{GP}\big(\underbrace{k(\mathbf{x}, X)k(X, X)^{-1}\mathbf{y}}_{\mu'(\mathbf{x})}, \quad \underbrace{k(\mathbf{x}, \mathbf{x}') - k(\mathbf{x}, X)k(X, X)^{-1}k(X, \mathbf{x}')}_{k'(\mathbf{x},\mathbf{x}')}\big) . \tag{240}$$

Our utility function is

$$U(\mathbf{x}, f) = \max\{f(\mathbf{x}), f(\mathbf{x}^*)\} \tag{241}$$

where $\mathbf{x}^*$ is the input location with the best function value that we have seen so far. To find the next place to evaluate, we simply apply next two steps of decision theory: finding the expected utility, and optimising it. We first find the expected utility:

$$\mathbb{E}_{p(f \mid \mathbf{y}, X)}[U(\mathbf{x}, f)] = \mathbb{E}_{p(f \mid \mathbf{y}, X)}[\max\{f(\mathbf{x}), f(\mathbf{x}^*)\}]\,. \tag{242}$$

We know that in a noiseless Gaussian process posterior, the predictive variance is zero for observed points. Since $f(\mathbf{x}^*)$ is an observed point (contained in the data $\mathcal{D} = (X, \mathbf{y})$), it will not be uncertain. Therefore, we only need to take the expectation over the function at the candidate point $\mathbf{x}$:

$$\begin{aligned}
&\mathbb{E}_{p(f \mid \mathbf{y}, X)}[U(\mathbf{x}, f)] = \mathbb{E}_{p(f(\mathbf{x}) \mid \mathbf{y}, X)}[\max\{f(\mathbf{x}), f(\mathbf{x}^*)\}] \\
&= \int_{-\infty}^{\infty} \mathcal{N}(f(\mathbf{x}); \mu'(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')) \max\{f(\mathbf{x}), f(\mathbf{x}^*)\} \mathrm{d}f(\mathbf{x}) \\
&= \int_{-\infty}^{f(\mathbf{x}^*)} \mathcal{N}(f(\mathbf{x}); \mu'(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')) f(\mathbf{x}^*) \mathrm{d}f(\mathbf{x}) + \int_{f(\mathbf{x}^*)}^{\infty} \mathcal{N}(f(\mathbf{x}); \mu'(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')) f(\mathbf{x}) \mathrm{d}f(\mathbf{x}) \\
&= f(\mathbf{x}^*)\Phi\left(\frac{f(\mathbf{x}^*) - \mu'(\mathbf{x})}{\sqrt{k'(\mathbf{x}, \mathbf{x})}}\right) + \left(1 - \Phi\left(\frac{f(\mathbf{x}^*) - \mu'(\mathbf{x})}{\sqrt{k'(\mathbf{x}, \mathbf{x})}}\right)\right)\mu'(\mathbf{x}) + \phi\left(\frac{f(\mathbf{x}^*) - \mu'(\mathbf{x})}{\sqrt{k'(\mathbf{x}, \mathbf{x})}}\right)\sqrt{k'(\mathbf{x}, \mathbf{x})} \\
&= \mu'(\mathbf{x}) + \underbrace{(f(\mathbf{x}^*) - \mu'(\mathbf{x}))\Phi\left(\frac{f(\mathbf{x}^*) - \mu'(\mathbf{x})}{\sqrt{k'(\mathbf{x}, \mathbf{x})}}\right) + \phi\left(\frac{f(\mathbf{x}^*) - \mu'(\mathbf{x})}{\sqrt{k'(\mathbf{x}, \mathbf{x})}}\right)\sqrt{k'(\mathbf{x}, \mathbf{x})}}_{\text{Expected improvement}}\,.
\end{aligned} \tag{243}$$

Then we pick the point that has the largest value of the expression above. This can be done through gridding or numerical optimisation.

Notice how we want to pick the point with the highest mean *plus* the expected improvement over the earlier evaluation $f(\mathbf{x})$. Simply picking the point with the highest $\mu'(\mathbf{x})$ would be what we would do if we did not have any existing evaluations. However, since we can always fall back on our previous evaluation with a reward of $f(\mathbf{x})$, decision theory tells us to explore more aggressively.

6.30 Question 32 – Gaussian process classification*

This discussion follows Rasmussen and Williams [2006, §3.4.1], although they use a slightly different likelihood. For completeness, I restate the Laplace approximation for an unnormalised probability distribution $\tilde{p}(\mathbf{f})$

$$q(\mathbf{f}) = \mathcal{N}\big(\mathbf{f}; \mathbf{f}^*, -\nabla_{\mathbf{f}}^2 \log \tilde{p}(\mathbf{f})|_{\mathbf{f}^*}^{-1}\big)\,, \qquad \mathbf{f}^* = \operatorname*{argmax}_{\mathbf{f}} \tilde{p}(\mathbf{f})\,. \tag{244}$$

This comes from a second order Taylor series expansion of $\tilde{p}(\mathbf{f})$

$$p(\mathbf{f}) \approx q(\mathbf{f}) = -\log Z + \underbrace{\nabla_{\mathbf{f}} \log \tilde{p}(\mathbf{f})|_{\mathbf{f}^*}}_{=0}(\mathbf{f} - \mathbf{f}^*) + \frac{1}{2}(\mathbf{f} - \mathbf{f}^*)^\intercal \nabla_{\mathbf{f}}^2 \log \tilde{p}(\mathbf{f})|_{\mathbf{f}^*}(\mathbf{f} - \mathbf{f}^*)\,, \tag{245}$$

which implies a Gaussian approximation.

6.30.1 Condition at the optimum

$$\begin{aligned}
\nabla_{\mathbf{f}}\left[\log p(\mathbf{f}) + \sum_{n=1}^{N} \log p(y_n \mid f(\mathbf{x}_n), \mathbf{x}_n)\right] &= \nabla_{\mathbf{f}}\left[-\frac{1}{2}\log 2\pi|\boldsymbol{K}| - \frac{1}{2}\mathbf{f}^\intercal \boldsymbol{K}^{-1}\mathbf{f} + \sum_{n=1}^{N} \log p(y_n \mid f(\mathbf{x}_n), \mathbf{x}_n)\right] \\
&= -\boldsymbol{K}^{-1}\mathbf{f} + \sum_{n=1}^{N} \nabla_{\mathbf{f}} \log p(y_n \mid f(\mathbf{x}_n), \mathbf{x}_n) \qquad (246) \\
\implies \mathbf{f} &= \boldsymbol{K}\sum_{n=1}^{N} \nabla_{\mathbf{f}} \log p(y_n \mid f(\mathbf{x}_n), \mathbf{x}_n) \qquad (247)
\end{aligned}$$

If you struggle to take first derivatives w.r.t. vectors, remember that for $f(\mathbf{x}) \in \mathbb{R}, \mathbf{x} \in \mathbb{R}^D$ we have

$$\nabla_{\mathbf{x}} f(\mathbf{x}) \in \mathbb{R}^D \quad \text{with } [\nabla_{\mathbf{x}} f(\mathbf{x})]_d = \frac{\partial f(\mathbf{x})}{\partial x_d}\,. \tag{248}$$

You can obtain the derivative of any vector-input function by writing it as a sum, and then taking the partial derivative, i.e. for $f(\mathbf{x}) = \mathbf{x}^\intercal \boldsymbol{A}\mathbf{x}$ with symmetric $\boldsymbol{A}$

$$\begin{aligned}[\nabla_{\mathbf{x}} f(\mathbf{x})]_d &= \frac{\partial}{\partial x_d} \sum_{i=1}^{D}\sum_{j=1}^{D} x_i A_{ij} x_j = \sum_{ij}\left[\frac{\partial x_i}{\partial x_d} A_{ij} x_j + x_i A_{ij} \frac{\partial x_j}{\partial x_d}\right] = 2\sum_j A_{dj} x_j \\ &= [2\boldsymbol{A}\mathbf{x}]_d \end{aligned} \tag{249}$$

This uses that $\frac{\partial x_i}{\partial x_d}$ is only 1 when $i = d$.

6.30.2 Covariance of Laplace approximation

From the definition of the Laplace approximation (you should know this) the covariance is equal to the negative inverse Hessian of the objective function at the optimum. Let's find an expression for the Hessian.

$$\begin{aligned}\nabla^2_{\mathbf{f}}\left[\log p(\mathbf{f}) + \sum_{n=1}^{N} \log p(y_n \,|\, f(\mathbf{x}_n), \mathbf{x}_n)\right] &= \nabla^2_{\mathbf{f}}\left[-\frac{1}{2}\log 2\pi|\boldsymbol{K}| - \frac{1}{2}\mathbf{f}^\intercal \boldsymbol{K}^{-1}\mathbf{f} + \sum_{n=1}^{N} \log p(y_n \,|\, f(\mathbf{x}_n), \mathbf{x}_n)\right] \\ &= \nabla_{\mathbf{f}}\left[-\boldsymbol{K}^{-1}\mathbf{f} + \sum_{n=1}^{N} \nabla_{\mathbf{f}} \log p(y_n \,|\, f(\mathbf{x}_n), \mathbf{x}_n)\right] \\ &= -\boldsymbol{K}^{-1} + \sum_{n=1}^{N} \nabla^2_{\mathbf{f}} \log p(y_n \,|\, f(\mathbf{x}_n), \mathbf{x}_n) \\ &= -\boldsymbol{K}^{-1} - \boldsymbol{W}\end{aligned}$$

Here $\boldsymbol{W}$ is diagonal with $W_{nn} = \nabla^2_{f(\mathbf{x}_n)} \log p(y_n \,|\, f(\mathbf{x}_n), \mathbf{x}_n)$, as each likelihood term only depends on a single element in $\mathbf{f}$, i.e. $f(\mathbf{x}_n)$. To be completely explicit, one should actually compute the terms W_{nn}, but this is standard, as it is simply differentiation w.r.t. one variable.

6.30.3 Clearly state the approximation

So, overall we get the approximation

$$q(\mathbf{f}) = \mathcal{N}\big(\mathbf{f}; \mathbf{f}^*, (\boldsymbol{K}^{-1} + \boldsymbol{W})^{-1}\big)\,. \tag{250}$$

6.30.4 Monte Carlo approximation

Remember that we denoted $\mathbf{f} = f(X)$.

$$p^* = p(y^* \,|\, \mathbf{x}^*, \mathbf{y}, X) = \int p(y^* \,|\, f(\mathbf{x}^*), \mathbf{x}^*) p(f(\mathbf{x}^*), f(X) \,|\, \mathbf{x}^*, \mathbf{y}, X) \mathrm{d}f(\mathbf{x}^*) \mathrm{d}f(X) \tag{251}$$

$$= \int p(y^* \,|\, f(\mathbf{x}^*), \mathbf{x}^*) p(f(\mathbf{x}^*) \,|\, f(X), X, \mathbf{x}^*) p(f(X) \,|\, \mathbf{y}, X) \mathrm{d}f(X) \mathrm{d}f(\mathbf{x}^*) \tag{252}$$

$$\approx \int p(y^* \,|\, f(\mathbf{x}^*), \mathbf{x}^*) \left[\int p(f(\mathbf{x}^*) \,|\, f(X), X, \mathbf{x}^*) q(f(X)) \mathrm{d}f(X)\right] \mathrm{d}f(\mathbf{x}^*) \tag{253}$$

The inner integral can be performed using the standard Gaussian manipulations, to get

$$q(f(\mathbf{x}^*)) = \int p(f(\mathbf{x}^*) \,|\, f(X), X, \mathbf{x}^*) q(f(X)) \mathrm{d}f(X) \tag{254}$$

$$= \mathcal{N}\Big(f(\mathbf{x}^*); k(\mathbf{x}^*, X)\boldsymbol{K}^{-1}\mathbf{f}^*, \tag{255}$$

$$k(\mathbf{x}^*, \mathbf{x}^*) - k(\mathbf{x}^*, X)\boldsymbol{K}^{-1}\big[\boldsymbol{K} - (\boldsymbol{K}^{-1} + \boldsymbol{W})^{-1}\big]\boldsymbol{K}^{-1}k(X, \mathbf{x}^*)\Big)\,. \tag{256}$$

You should be able to derive this. The trick is to note that

$$f(\mathbf{x}^*) = k(\mathbf{x}^*, X)\boldsymbol{K}^{-1}\mathbf{f} + \epsilon\,, \tag{257}$$

$$\mathbf{f} \sim q(\mathbf{f})\,, \qquad \epsilon \sim \mathcal{N}\big(0, k(\mathbf{x}^*, \mathbf{x}^*) - k(\mathbf{x}^*, X)\boldsymbol{K}^{-1}k(X, \mathbf{x}^*)\big)\,. \tag{258}$$

The Monte Carlo estimate becomes

$$\hat{p}^* = \frac{1}{S}\sum_{s=1}^{S} p(y^* \,|\, f(\mathbf{x}^*) = f^{(s)}, \mathbf{x}^*)\,, \qquad f^{(s)} \overset{iid}{\sim} q(f(\mathbf{x}^*))\,. \tag{259}$$

6.31 Question 33 – Gaussian process regression*

Following the same procedure as in the previous question, we find the mean and covariance for the Laplace approximation.

$$\log p(\mathbf{f} \,|\, \mathbf{y}, X) = \text{const} + -\frac{1}{2}\mathbf{f}^\intercal\boldsymbol{K}^{-1}\mathbf{f} - \frac{1}{2}(\mathbf{f}-\mathbf{y})^\intercal\boldsymbol{\Sigma}^{-1}(\mathbf{f}-\mathbf{y}) \tag{260}$$

$$= \text{const} + -\frac{1}{2}\mathbf{f}^\intercal\boldsymbol{K}^{-1}\mathbf{f} - \frac{1}{2}\mathbf{f}^\intercal\boldsymbol{\Sigma}^{-1}\mathbf{f} + \mathbf{f}^\intercal\boldsymbol{\Sigma}^{-1}\mathbf{y} - \frac{1}{2}\mathbf{y}^\intercal\boldsymbol{\Sigma}^{-1}\mathbf{y} \tag{261}$$

$$\nabla_{\mathbf{f}} \log p(\mathbf{f} \,|\, \mathbf{y}, X) = -\boldsymbol{K}^{-1}\mathbf{f} - \boldsymbol{\Sigma}^{-1}\mathbf{f} + \boldsymbol{\Sigma}^{-1}\mathbf{y} = 0 \tag{262}$$

$$\implies \mathbf{f}^* = (\boldsymbol{K}^{-1} + \boldsymbol{\Sigma}^{-1})^{-1}\boldsymbol{\Sigma}^{-1}\mathbf{y} \tag{263}$$

$$= \boldsymbol{K}(\boldsymbol{K} + \boldsymbol{\Sigma})^{-1}\mathbf{y} \tag{264}$$

$$\nabla^2_{\mathbf{f}} \log p(\mathbf{f} \,|\, \mathbf{y}, X) = -\boldsymbol{K}^{-1} - \boldsymbol{\Sigma}^{-1} = -(\boldsymbol{K}^{-1} + \boldsymbol{\Sigma}^{-1}) \tag{265}$$

$$\implies q(\mathbf{f}) = \mathcal{N}\big(\mathbf{f}; \boldsymbol{K}(\boldsymbol{K} + \boldsymbol{\Sigma})^{-1}\mathbf{y}, (\boldsymbol{K}^{-1} + \boldsymbol{\Sigma}^{-1})^{-1}\big) \tag{266}$$

We can apply Woodbury to get this into a more familiar form...

$$q(\mathbf{f}) = \mathcal{N}\big(\mathbf{f}; \boldsymbol{K}(\boldsymbol{K} + \boldsymbol{\Sigma})^{-1}\mathbf{y}, \boldsymbol{K} - \boldsymbol{K}(\boldsymbol{K} + \boldsymbol{\Sigma})^{-1}\boldsymbol{K}\big) \tag{267}$$

This looks familiar.

6.32 Question 34 – Rejection sampling*

Lecture.

6.33 Question 35 – Importance sampling*

Lecture.

6.34 Question 36 – Variance of importance sampling*

Lecture.

6.35 Question 40 – GP hyperparameters for non-conjugate likelihoods

6.35.1 State the integral for $p(y^*|\mathbf{y})$ for a GP

We can write this as

$$p(y^*|\mathbf{y}) = \int\int\int p(y^*|f^*)p(f^*|\mathbf{f}, \boldsymbol{\theta})p(\mathbf{f}, \boldsymbol{\theta}|\mathbf{y})\mathrm{d}f^*\mathrm{d}\boldsymbol{\theta}\mathrm{d}\mathbf{f}. \tag{268}$$

6.35.2 State the Monte Carlo approximation to this integral

We can write the above integral in terms of expectations

$$p(y^*|\mathbf{y}) = \mathbb{E}_{p(\mathbf{f},\boldsymbol{\theta}|\mathbf{y})}\left[\mathbb{E}_{p(f^*|\mathbf{f},\boldsymbol{\theta})}[p(y^*|f^*)]\right]. \tag{269}$$

The inner expectation over $p(f^*|\mathbf{f},\boldsymbol{\theta})$ is intractable in general (for a non-conjugate likelihood $p(y^*|f^*)$). Similarly the posterior $p(\mathbf{f}|\mathbf{y})$ cannot be found in closed form for a general likelihood, and the hyperparameter posterior $p(\boldsymbol{\theta}|\mathbf{y})$ is usually intractable even if the likelihood is Gaussian (see Question 25 – Hyperparameter Conjugacy). Thus the joint posterior $p(\mathbf{f},\boldsymbol{\theta}|\mathbf{y})$ is intractable, as is its expectation.

Given these intractabilities, we resort to the following Monte Carlo approximation

$$p(y^*|\mathbf{y}) = \frac{1}{K}\sum_{k}^{K} p(y^*|f^{*(k)}), \quad f^{*(k)} \sim p(f^*|\mathbf{f}^{(k)},\boldsymbol{\theta}^{(k)}), \quad \boldsymbol{\theta}^{(k)},\mathbf{f}^{(k)} \sim p(\mathbf{f},\boldsymbol{\theta}|\mathbf{y}). \tag{270}$$

6.35.3 State the procedure for computing samples for the Monte Carlo approximation

We need to draw samples from $p(\mathbf{f},\boldsymbol{\theta}|\mathbf{y})$ and $p(f^*|\mathbf{f},\boldsymbol{\theta})$. While $p(\mathbf{f},\boldsymbol{\theta}|\mathbf{y})$ is intractable, we can sample using the Metropolis algorithm (or another MCMC algorithm). We would target the tractable joint density

$$p(\mathbf{f},\boldsymbol{\theta},\mathbf{y}) = p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \tag{271}$$

where $p(\boldsymbol{\theta})$ is prior on hyperparameters. This is justified as $p(\mathbf{f},\boldsymbol{\theta},\mathbf{y})$ is proportional to the joint posterior $p(\mathbf{f},\boldsymbol{\theta}|\mathbf{y})$ (w.r.t $\boldsymbol{\theta}$ and $\mathbf{f}$). Note that $\mathbf{y}$ is the observed data and so is fixed in our target density.

For our sampler, we could use a symmetric proposal $\hat{T}(\mathbf{s}'|\mathbf{s}^{(t)}) = \mathcal{N}(\mathbf{s}^{(t)},\boldsymbol{\Sigma})$ where state $\mathbf{s} = [\mathbf{f}^\intercal,\boldsymbol{\theta}^\intercal,\mathbf{y}^\intercal]^\intercal$. $\boldsymbol{\Sigma}$ could diagonal, but we could improve sampling efficiency by choosing the submatrix $\boldsymbol{\Sigma}_\mathbf{f}$ to reflect the covariance of the GP on $\mathbf{f}$. As $\hat{T}(\mathbf{s}'|\mathbf{s}^{(t)}) = \hat{T}(\mathbf{s}^{(t)}|\mathbf{s}')$, the proposal densities cancel out in the acceptance probability which reduces to $p(\mathbf{s}')/p(\mathbf{s}^{(t)})$.

After sampling $\mathbf{f}^{(t)},\boldsymbol{\theta}^{(t)}$, we can compute $p(f^*|\mathbf{f}^{(t)},\boldsymbol{\theta}^{(t)})$ analytically (Gaussian conditioning), and then sample from the resulting multivariate Gaussian.

6.36 Question 39 – Independent Markov Chains

Consider two independent Markov chains generated by a transition operator T such that $p(\mathbf{x}^{(t+1)}|\mathbf{x}^{(1)},\cdots,\mathbf{x}^{(t)}) = T(\mathbf{x}^{(t+1)}|\mathbf{x}^{(t)})$, and denote the samples from chain $i \in \{1,2\}$ at time t as $x_i^{(t)}$. The question asks us to show independence between $x_1^{(\tau)}$ and $x_2^{(\tau)}$ where τ refers to the timestep of the final states mentioned in the question.

Consider the joint $p(\mathbf{x}_1^{(\tau)},\mathbf{x}_2^{(\tau)})$. The way the samples from this distribution are drawn is by applying the transition operator T independently to the previous samples $\mathbf{x}_1^{(\tau-1)}$ and $\mathbf{x}_2^{(\tau-1)}$. Therefore, we have

$$\begin{aligned}
p(\mathbf{x}_1^{(\tau)},\mathbf{x}_2^{(\tau)}) &= p(\mathbf{x}_1^{(\tau)},\mathbf{x}_2^{(\tau)}|\mathbf{x}_1^{(\tau-1)},\mathbf{x}_2^{(\tau-1)})p(\mathbf{x}_1^{(\tau-1)},\mathbf{x}_2^{(\tau-1)}) && (272)\\
&= T(\mathbf{x}_1^{(\tau)}|\mathbf{x}_1^{(\tau-1)})T(\mathbf{x}_2^{(\tau)}|\mathbf{x}_2^{(\tau-1)})p(\mathbf{x}_1^{(\tau-1)},\mathbf{x}_2^{(\tau-1)}) && (273)\\
&\cdots && (274)\\
&= \prod_{t=1}^{\tau} T(\mathbf{x}_1^{(t)}|\mathbf{x}_1^{(t-1)})T(\mathbf{x}_2^{(t)}|\mathbf{x}_2^{(t-1)})p(\mathbf{x}_1^{(0)},\mathbf{x}_2^{(0)}) && (275)
\end{aligned}$$

after propagating the effect of the transition operator for each of the τ steps. Since the Markov chains are initialised independently, we can take the initial states as chosen independently at random from some identical distribution $p(\mathbf{x}^{(0)})$. The joint starting distribution will therefore factorise as $p(\mathbf{x}_1^{(0)},\mathbf{x}_2^{(0)}) = p(\mathbf{x}_1^{(0)})p(\mathbf{x}_2^{(0)})$. Hence we can write

$$\begin{aligned}
p(\mathbf{x}_1^{(\tau)},\mathbf{x}_2^{(\tau)}) &= \prod_{t=1}^{\tau} T(\mathbf{x}_1^{(t)}|\mathbf{x}_1^{(t-1)})T(\mathbf{x}_2^{(t)}|\mathbf{x}_2^{(t-1)})p(\mathbf{x}_1^{(0)})p(\mathbf{x}_2^{(0)}) && (276)\\
&= \left(\prod_{t=1}^{\tau} T(\mathbf{x}_1^{(t)}|\mathbf{x}_1^{(t-1)})p(\mathbf{x}_1^{(0)})\right)\left(\prod_{t=1}^{\tau} T(\mathbf{x}_2^{(t)}|\mathbf{x}_2^{(t-1)})p(\mathbf{x}_2^{(0)})\right) && (277)\\
&= p(\mathbf{x}_1^{(\tau)})p(\mathbf{x}_2^{(\tau)}), && (278)
\end{aligned}$$

and the joint factors into two terms $p(\mathbf{x}_i^{(\tau)}) = \prod_{t=1}^{\tau} T(\mathbf{x}_i^{(t)}|\mathbf{x}_i^{(t-1)})p(\mathbf{x}_i^{(0)})$, and we conclude that the samples from the Markov chains at time τ are independent.

6.37 Question 43 – Reparameterisation gradient*

You should be able to derive a variational lower bound, which for this model becomes:

$$\mathcal{L} = \sum_{n=1}^{N} \mathbb{E}_{q(f(X))}[\log p(\mathbf{y} \mid f(\mathbf{x}_n), \mathbf{x}_n)] - \mathrm{KL}[q(f(X))||p(f(X))] \tag{279}$$

For the reparameterisation gradient, we first rewrite the expected log-likelihood term as

$$\begin{aligned}\mathbb{E}_{q(f(X))}[\log p(\mathbf{y} \mid f(\mathbf{x}_n), \mathbf{x}_n)] &= \mathbb{E}_{q(f(\mathbf{x}_n))}[y_n \log[\sigma(f(\mathbf{x}_n))] + (1-y_n)\log[1-\sigma(f(\mathbf{x}_n))]] &(280)\\ &= \mathbb{E}_{q(\epsilon)}[y_n \log[\sigma(\mu_n + \sigma_n\epsilon)] + (1-y_n)\log[1-\sigma(\mu_n+\sigma_n\epsilon)])]\,, &(281)\end{aligned}$$

where μ_n and σ_n are the mean and stddev of $q(f(\mathbf{x}_n))$. Some crucial observations:

- We only need to sample 1D Gaussians, since each term in the expected log-likelihood depends only on the marginal of $q(f(X))$, $q(f(\mathbf{x}_n))$.
- We still get a tighter bound by parameterising $q(f(X))$ as a full-rank Gaussian, i.e. $q(f(X)) = \mathcal{N}(f(X), \boldsymbol{\mu}, \boldsymbol{\Sigma})$, since the KL divergence depends on the correlations.

Now we develop an expression for the reparameterisation gradient estimator of one term in the sum. The final estimator will be a sum over a minibatch of points. We start with a general expression for the reparameterisation gradient. Note that since the KL divergence can be computed in closed-form, I will not include it in the derivation, as it just requires differentiation of a computable quantity.

$$\begin{aligned}\nabla_\mu \mathbb{E}_{q(\epsilon)}[\log p(y_n \mid f(\mathbf{x}) = \mu + \sigma\epsilon, \mathbf{x})] &= \mathbb{E}_{q(\epsilon)}\left[\nabla_{f(\mathbf{x})} \log p(y_n \mid f(\mathbf{x}), \mathbf{x})|_{f(\mathbf{x})=\mu+\sigma\epsilon} \cdot \nabla_\mu(\mu+\sigma\epsilon)\right] &(282)\\ &= \mathbb{E}_{q(\epsilon)}\left[\nabla_{f(\mathbf{x})} \log p(y_n \mid f(\mathbf{x}), \mathbf{x})|_{f(\mathbf{x})=\mu+\sigma\epsilon}\right] &(283)\\ \nabla_\sigma \mathbb{E}_{q(\epsilon)}[\log p(y_n \mid f(\mathbf{x}) = \mu + \sigma\epsilon, \mathbf{x})] &= \mathbb{E}_{q(\epsilon)}\left[\nabla_{f(\mathbf{x})} \log p(y_n \mid f(\mathbf{x}), \mathbf{x})|_{f(\mathbf{x})=\mu+\sigma\epsilon} \cdot \epsilon\right] &(284)\end{aligned}$$

The Monte Carlo estimator can be found simply by sampling from $q(\epsilon)$.

This is roughly the level of detail you would need to do on the exam. Write out the chain rule until you get to the reparameterisation of the variational distribution. Intermediate terms like

$$\nabla_{f(\mathbf{x})} \log p(y_n \mid f(\mathbf{x}), \mathbf{x})|_{f(\mathbf{x})=\mu+\sigma\epsilon} \tag{285}$$

do not have to be differentiated unless asked explicitly.

6.38 Question 44 – Bayesian neural networks

A neural network is a function that is parameterised through weights $\mathbf{w}$. So we write the neural network as $f(\mathbf{x}; \mathbf{w})$. We assume a general likelihood which depends only on the output for a single input, i.e. $p(y_n \mid f(\mathbf{x}_n; \mathbf{w}))$, and a Gaussian prior on the weights $p(\mathbf{w}) = \mathcal{N}\left(\mathbf{w}; 0, \lambda^{-1}\right)$. The variational lower bound becomes

$$\mathcal{L} = \sum_{n=1}^{N} \mathbb{E}_{q(\mathbf{w})}[\log p(y_n \mid f(\mathbf{x}; \mathbf{w}))] - \mathrm{KL}[q(\mathbf{w})||p(\mathbf{w})]\,. \tag{286}$$

We assume a mean-field variational distribution $q(\mathbf{w}) = \prod_i \mathcal{N}\left(w_i; \mu_i, \sigma_i^2\right)$.

In the previous question, each likelihood term depended only on a single variational distribution. In this question, we need to sample *all* the weights. We write our reparameterisation trick as

$$\mathbf{w} = \boldsymbol{\mu} + \boldsymbol{\sigma} \circ \boldsymbol{\epsilon} \tag{287}$$

where $\circ$ denotes elementwise multiplication. Now we can write the gradients as

$$\begin{aligned}\nabla_{\boldsymbol{\mu}} \mathbb{E}_{q(\boldsymbol{\epsilon})}[\log p(y_n \mid f(\mathbf{x}; \mathbf{w}))] &= \mathbb{E}_{q(\boldsymbol{\epsilon})}\left[\nabla_{f(\mathbf{x};\mathbf{w})} \log p(y_n \mid f(\mathbf{x}; \mathbf{w}))|_{\mathbf{w}=\boldsymbol{\mu}+\boldsymbol{\sigma}\boldsymbol{\epsilon}} \nabla_{\mathbf{w}} f(\mathbf{x}; \mathbf{w})|_{\mathbf{w}=\boldsymbol{\mu}+\boldsymbol{\sigma}\boldsymbol{\epsilon}}\right] &(288)\\ \nabla_{\boldsymbol{\sigma}} \mathbb{E}_{q(\boldsymbol{\epsilon})}[\log p(y_n \mid f(\mathbf{x}; \mathbf{w}))] &= \mathbb{E}_{q(\boldsymbol{\epsilon})}\left[\nabla_{f(\mathbf{x};\mathbf{w})} \log p(y_n \mid f(\mathbf{x}; \mathbf{w}))|_{\mathbf{w}=\boldsymbol{\mu}+\boldsymbol{\sigma}\boldsymbol{\epsilon}} \nabla_{\mathbf{w}} f(\mathbf{x}; \mathbf{w})|_{\mathbf{w}=\boldsymbol{\mu}+\boldsymbol{\sigma}\boldsymbol{\epsilon}} \circ \boldsymbol{\epsilon}\right] &(289)\end{aligned}$$

and find Monte Carlo estimates in the usual way. Note that we again apply the chain rule until we differentiate through the reparameterisation.

6.39 Question 45 – Variational Autoencoders

For a single data point we have the ELBO

$$\mathcal{L}_n = \mathbb{E}_{q(\mathbf{z}_n)}[\log p(\mathbf{x}_n \mid \mathbf{z}_n; \theta)] - \mathrm{KL}[q(\mathbf{z}_n)||p(\mathbf{x}_n)] \,. \tag{290}$$

The approximate posterior is parameterised through a recognition network with parameters ϕ, so we denote

$$q(\mathbf{x}_n; \phi) = \mathcal{N}(\mathbf{x}_n; \boldsymbol{\mu}(\mathbf{x}_n, \phi), \boldsymbol{\Sigma}(\mathbf{x}_n, \phi)) \,. \tag{291}$$

Writing using the reparamterisation trick

$$\mathcal{L}_n = \mathbb{E}_{q(\boldsymbol{\epsilon})}\Big[\log p(\mathbf{x}_n; \mathbf{z}_n = \boldsymbol{\mu}(\mathbf{x}_n, \phi) + \boldsymbol{\Sigma}(\mathbf{x}_n, \phi)^{\frac{1}{2}}\boldsymbol{\epsilon})\Big] - \mathrm{KL} \,. \tag{292}$$

We can now take gradients. We start with the gradient of the model parameters θ

$$\nabla_\theta \mathcal{L}_n = \mathbb{E}_{q(\boldsymbol{\epsilon})}\big[\nabla_{\bar{\mathbf{x}}_n} \log \mathcal{N}\big(\mathbf{x}_n; \bar{\mathbf{x}}_n, \sigma^2\big)|_{\bar{\mathbf{x}}_n = f_\theta(\mathbf{z}_n)} \nabla_\theta f_\theta(\mathbf{z}_n = \boldsymbol{\mu} \dots)\big] \,. \tag{293}$$

Similar for the parameters of the inference network ϕ.

References

Marc Peter Deisenroth, A. Aldo Faisal, and Cheng Soon Ong (2020). *Mathematics for machine learning*. Cambridge University Press.

David J. C. MacKay (2003). *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press.

Carl Edward Rasmussen and Christopher K.I. Williams (2006). *Gaussian processes for machine learning*. MIT press, Cambridge, MA, USA.

Ronald E. Walpole and Raymond H. Myers (2012). *Probability & statistics for engineers & scientists*. Pearson Education Limited.

Wikipedia (2020). "Truncated normal distribution". In: *Wikipedia*.