The Australian National University
School of Computing
Convenor: Lexing Xie
Other Contributors: Alexander Soen, Chamin Hewa Koneputugodage

Semester 1, 2022
Assignment 1
Weight: 20%

COMP4670/8600: Statistical Machine Learning

Release Date. 23rd Feburary 2022.
Due Date. 28th March 2022 at 1200 AEDT.
Maximum credit. 100 Marks for COMP4670 and 120 Marks for COMP8600 (see Appendix A).

Section 1: Bayesian Estimate vs Probabilistic Hunches (5 Total Points)

Background: Daniel Kahneman and Amos Tversky are the first to systematically describe the various human judgement biases in situations about uncertain outcomes, or involving probabilistic reasoning. In this question, we pitch human intuitions (including our own) against statistical inference, on several examples presented in their original 1972 paper[1].

Gender balance in high school classes. There are two programs in a high school. Boys are a majority (65%) in program A, and a minority (45%) in program B. There is an equal number of classes in each of the two programs; with each class consisting of an equal number of students. You enter a class at random, and observe that 55% of the students are boys.

Question 1.1: Your hunch (0 Points)

In 30 seconds or less, what is your guess – does the class belong to program A or to program B?

Question 1.2: Computing the probabilities (4 Points)

Can you compute probabilities of the class being in program A and program B given the observation, and from there reason about which program is more likely?

Hint: You may need to write out expressions involving the number of students in this class, say n, which is fixed, but not provided in this question. Do the probability of the observed class being in program A depend on class size n?
You may want to consider the ratio of likelihoods.

Question 1.3: Well-reasoned hunch (1 Points)

Without computing the posterior probabilities for program A or B, could you make a reasoned argument, based on properties of the binomial distribution, that program B is more likely?

The variance of a binomial will be larger for p = .45 than for p = .65. Therefore, there will be more variability in the number of boys in Program B's classes, and it is more likely that a class in Program B will have as many as 55% boys.

[1] Kahneman, D. and Tversky, A., 1972. Subjective probability: A judgment of representativeness. Cognitive psychology, 3(3), pp.430-454.

Section 2: Exponential Families (60 Total Points)

In the following question, we will explore an important family of distributions in machine learning: the *exponential family* (not to be confused with an exponential distribution). Although this section and its questions are littered with mathematical and technical details, one should not miss the elegance and simplicity of exponential family distributions. Through the following questions, we will establish how this family of distribution generalises distributions you have seen before and explore how they allow for different forms of geometric ideas.

First, let us define the exponential family distributions we will be considering in this question.

Definition 1 (Exponential Family[2])**.** Given a function $\boldsymbol{u} : \mathbb{R} \to \mathbb{R}^m$, we denote an exponential family distribution as $\textsc{Exp}(\boldsymbol{u}, \boldsymbol{\eta})$, where $\boldsymbol{\eta} \in \mathcal{P} \subset \mathbb{R}^m$ designates the m-dimensional parameters of the distribution within an exponential family[3]. The corresponding densities of the distributions are given by

$$q(x \mid \boldsymbol{\eta}) = \exp(\boldsymbol{\eta}^\top \boldsymbol{u}(x) - \psi(\boldsymbol{\eta})), \tag{2.1}$$

where

$$\psi(\boldsymbol{\eta}) = \log \int \exp(\boldsymbol{\eta}^\top \boldsymbol{u}(x)) \, \mathrm{d}x. \tag{2.2}$$

The function $\boldsymbol{u}$ is called the sufficient statistics of the exponential family and the function ψ is called the log-partition function of the exponential family.

Remark. Notice that Definition 1 represent a broad set of special cases of the definition given in the Bishop textbook Eq. (2.194), with $h(\boldsymbol{x}) = 1$ and $g(\boldsymbol{\eta}) = \exp(-\psi(\boldsymbol{\eta}))$. For notational convenience in the tasks specified below, we use this definition.

We will also take the **assumption** that any random variable defined with respect to any exponential family distribution $x \sim \textsc{Exp}(\boldsymbol{u}, \boldsymbol{\eta})$ is a continuous random variable in this section (for simplicity). (For technical reasons, we also **assume** that $\boldsymbol{u}(\cdot)$ is continuous for each of its dimensions and so are its partial derivatives.)

Question 2.1: Verifying valid distribution (5 Points)

Show that $q(x \mid \boldsymbol{\eta})$ given by Eqs. (2.1) and (2.2) is a valid probability density function.

To start our exploration of general exponential families, we will consider its application to Bayesian estimation. In particular, we look at the example of considering a simple 1-dimensional Gaussian distribution.

Question 2.2: A Bayesian example (Part 1) (13 Points)

Suppose that we have a likelihood and prior distribution given by,

$$\begin{aligned} x \mid \mu &\sim \mathcal{N}(\mu, \sigma^2) \\ \mu &\sim \textsc{Exp}(\boldsymbol{u}, \boldsymbol{\eta}), \end{aligned}$$

where $\mathcal{N}$ is a 1-dimensional Gaussian distribution with mean μ and standard deviation σ. Show that $\mu \mid x \sim \textsc{Exp}(\hat{\boldsymbol{u}}, \hat{\boldsymbol{\eta}})$, for some $\hat{\boldsymbol{u}}$ and $\hat{\boldsymbol{\eta}}$.

Make sure to clearly specify the function $\hat{\boldsymbol{u}}$ and parameters $\hat{\boldsymbol{\eta}}$.

Hint: The integral for $q(\mu)$ does not need to be simplified / solved.

To move from this general case, we make the addition assumption that μ is also Gaussian by considering sufficient statistics given by $\boldsymbol{u}(x) = (x, x^2)$.

[2]This is actually a simplified version of an exponential family. A more general version can defined, *i.e.*, multi-dimensional and non-unit base measure.

[3]The exponential family for a fixed $\boldsymbol{u}$ is the set of distributions given by $\mathcal{M}_{\boldsymbol{u}} = \{\textsc{Exp}(\boldsymbol{u}, \boldsymbol{\eta}) \mid \boldsymbol{\eta} \in \mathcal{P}\}$. Two exponential family distributions share the same exponential family if and only if their $\boldsymbol{u}$'s are the same.

Question 2.3: A Bayesian example (Part 2) (7 Points)

Suppose that we have the same likelihood as per Question 2.2 but with a Gaussian prior distribution

$$x \mid \mu \sim \mathcal{N}(\mu, \sigma^2)$$
$$\mu \sim \mathcal{N}(\mu_0, \sigma_0^2).$$

First, find the parameters of an exponential family such that $\text{Exp}(\boldsymbol{u}, \boldsymbol{\eta}) = \mathcal{N}(\mu_0, \sigma_0^2)$ (in distribution) with $\boldsymbol{u} = (x, x^2)$.

Then use the result of Question 2.2 to find the parameters of $\mu \mid x$.

Another aspect which makes exponential families convenient to work with in Bayesian inference is the evaluation of *maximum likelihood estimation* (MLE). For multiple data points $\{x_i\}_{i=1}^N$ and a general probability distribution $q(x \mid \vartheta)$ with parameter ϑ, the MLE corresponds to the maximization of:

$$\ell(\vartheta) = \frac{1}{N} \sum_{i=1}^{N} \log q(x_i \mid \vartheta). \tag{2.3}$$

Let us consider the estimation of a parameter $\boldsymbol{\eta}$ of $\text{Exp}(\boldsymbol{u}, \boldsymbol{\eta})$ given a single data point x. Given that this is an exponential family distribution, the log-likelihood is presented in a relatively simple form:

$$\log q(x \mid \boldsymbol{\eta}) = \boldsymbol{\eta}^\top \boldsymbol{u}(x) - \psi(\boldsymbol{\eta}). \tag{2.4}$$

In practice, we often maximise the log-likelihood by minimise the negative of Eq. (2.4), *i.e.*, minimise the negative log-likelihood. The maximisation of this quantity for exponential families has many interesting properties which can be explored (for the interested reader see the convex conjugate).

Another important object in statistics is the *KL-divergence*, which is defined by:

$$D_{\text{KL}}[p(x) : q(x)] = \int p(x) \log \left(\frac{p(x)}{q(x)} \right) \, \mathrm{d}x. \tag{2.5}$$

We will show an equivalence in MLE and the minimisation of the KL-divergences.

Question 2.4 [COMP8600]: MLE is equivalent to KL minimisation (10 Points)

Let $\{x_i\}_{i=1}^N$ denote a set of data points we will use to fit an exponential family distribution $\text{Exp}(\boldsymbol{u}, \boldsymbol{\eta})$. With this, let the corresponding empirical distribution be defined as

$$\bar{q}(x) = \frac{1}{N} \sum_{i=1}^{N} \delta(x - x_i), \tag{2.6}$$

where $\delta(\cdot)$ is the Dirac delta function.

With $q(x \mid \boldsymbol{\eta})$ as the p.d.f. of $\text{Exp}(\boldsymbol{u}, \boldsymbol{\eta})$, show that the minimisation of $D_{\text{KL}}[\bar{q}(x) : q(x \mid \boldsymbol{\eta})]$ is equivalent to the MLE of $\boldsymbol{\eta}$ using all data points x_i.

Then, with the additional assumption that $\boldsymbol{u}$ is a linear map, show that the minimisation of $D_{\text{KL}}[\bar{q}(x) : q(x \mid \boldsymbol{\eta})]$ is equivalent to MLE using a single data point consisting of the mean of all data points x_i.

Hint: For the first part, use the property that for any continuous function f we have

$$\int \delta(x - a) \cdot f(x) \, \mathrm{d}x = f(a).$$

Remark. Notice that the first part of Question 2.4 actually holds for any type of distribution and is not limited to just considering exponential family distributions.

As we have seen from the preceding questions, the parameters $\boldsymbol{\eta}$ are key in the analysis of exponential families. These sets of parameters are called the *natural parameters* of the exponential family. Despite their usefulness, sometimes it is more convenience to use a different set of parameters when analysing exponential family distributions. The aforementioned set of parameters we will consider are the distribution's *expectation parameters* $\boldsymbol{\lambda}$, which corresponds to the expectation of the sufficient statistics for a specific parametrisation $\boldsymbol{\eta}$.

In particular, there is a very nice connection between the expectation parameters $\boldsymbol{\lambda}$ and the log-partition function $\psi(\cdot)$:

$$\boldsymbol{\lambda} \overset{\text{def}}{=} \mathop{\mathbb{E}}_{x \sim \textsc{Exp}(\boldsymbol{u}, \boldsymbol{\eta})} [\boldsymbol{u}(x)] = \nabla \psi(\boldsymbol{\eta}). \tag{2.7}$$

Although one could think of $\boldsymbol{\lambda}$ as a function of $\boldsymbol{\eta}$, we make this correspondence implicit. If there are multiple natural and expectation parameters we are interested in, we will subscript them to remove ambiguity, *i.e.*, $\boldsymbol{\eta}_i$ corresponds to $\boldsymbol{\lambda}_i$ or more concretely $\boldsymbol{\lambda}_i = \nabla \psi(\boldsymbol{\eta}_i)$.

Eq. (2.7) is key to the next few properties of exponential families we will explore.

One reason why the expectation parameters are of interest is that the KL-divergence of these exponential families can be characterised by these parameters. For two distributions $\textsc{Exp}(\boldsymbol{u}, \boldsymbol{\eta}_1)$ and $\textsc{Exp}(\boldsymbol{u}, \boldsymbol{\eta}_2)$ within the same exponential family, we shorthand the KL-divergence of densities by $D_{\mathrm{KL}}[\boldsymbol{\eta}_1 : \boldsymbol{\eta}_2] \overset{\text{def}}{=} D_{\mathrm{KL}}[q(x \mid \boldsymbol{\eta}_1) : q(x \mid \boldsymbol{\eta}_2)]$.

Question 2.5: KL-Divergence for exponential families (15 Points)

Show that the KL-divergence of distributions $\textsc{Exp}(\boldsymbol{u}, \boldsymbol{\eta}_1)$ and $\textsc{Exp}(\boldsymbol{u}, \boldsymbol{\eta}_2)$ within the same exponential family is given by:

$$D_{\mathrm{KL}}[\boldsymbol{\eta}_1 : \boldsymbol{\eta}_2] = \psi(\boldsymbol{\eta}_2) - \psi(\boldsymbol{\eta}_1) - \boldsymbol{\lambda}_1^\top (\boldsymbol{\eta}_2 - \boldsymbol{\eta}_1).$$

Hint: Use Eq. (2.7).

Divergences are an important statistical tool to measure the 'closeness' of distributions. However, they typically do not exhibit properties which are typical for more common distance function (metric functions). For example, the common Euclidean distance is symmetrical and satisfies the triangle inequality. The KL-divergence **does not** have these properties. However despite this, within an exponential family a generalisation of the Pythagorean theorem exists, where a different notion of "right angle" is used and the squared Euclidean distance is replaced by KL-divergence:

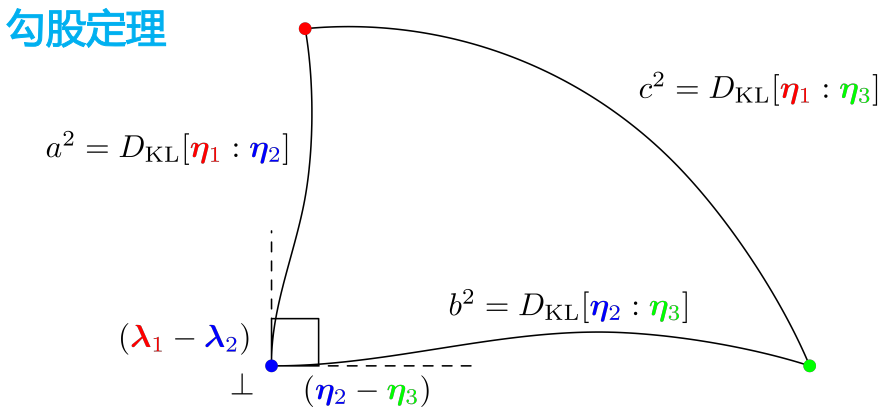

Figure 1: Question 2.6. A Pythagorean Theorem for KL-divergence.

Question 2.6: Pythagorean Theorem for exponential families (10 Points)

Let $\textsc{Exp}(\boldsymbol{u}, \boldsymbol{\eta}_1)$, $\textsc{Exp}(\boldsymbol{u}, \boldsymbol{\eta}_2)$, and $\textsc{Exp}(\boldsymbol{u}, \boldsymbol{\eta}_3)$ be distributions within the same exponential family. Furthermore, let

$$a^2 = D_{\mathrm{KL}}[\boldsymbol{\eta}_1 : \boldsymbol{\eta}_2]; \quad b^2 = D_{\mathrm{KL}}[\boldsymbol{\eta}_2 : \boldsymbol{\eta}_3]; \quad c^2 = D_{\mathrm{KL}}[\boldsymbol{\eta}_1 : \boldsymbol{\eta}_3].$$

Prove that Fig. 1 holds. That is, show that $a^2 + b^2 = c^2$ iff the difference in natural parameters $(\boldsymbol{\eta}_2 - \boldsymbol{\eta}_3)$ is perpendicular to the difference in expectation parameters $(\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2)$.

Section 3: Utilising Expectation Maximisation (55 Total Points)

In this set of questions we are going to explore the Expectation Maximisation (EM) algorithm in greater detail. We will first consider EM for Exponential family Mixture Models (EMMs), which should mirror many of the aspects which Gaussian Mixture Models (GMMs) have. We then adapt EM to Bayesian Linear Regression (BLR) and BLR with non-gaussian priors. A light refresher of the EM algorithm with notation used in this section can be found in Appendix B.

Understanding EM with EMM. Let us consider a mixture of K exponential family distributions $\text{Exp}(\boldsymbol{u}, \boldsymbol{\eta}_k)$ (all of which come from the same exponential family, *i.e.*, they all share the same sufficient statistics function $\boldsymbol{u}(\cdot)$) with m-dimensional parameters $\{\boldsymbol{\eta}_k \in \mathbb{R}^m\}_{k=1}^K$; and N data points sampled from the mixture $\{x_n\}_{n=1}^N$, $x_n \in \mathbb{R}$, where the sampling is done with the ratio $\pi_1, \dots, \pi_K$ with $0 < \pi_k < 1$ and $\sum_{k=1}^K \pi_k = 1$. Note then that our data is $X = [x_n]_{n=1}^N \in \mathbb{R}^{1\times N}$, and our parameters are $\boldsymbol{H} = [\boldsymbol{\eta}_k]_{k=1}^K$ and $\pi = [\pi]_{k=1}^K$. We denote both parameters as $\vartheta = \{\boldsymbol{H}, \pi\}$ to follow the preceding notation in Appendix B. In summary, to sample a single sample from the mixture we have:

$$z \sim \text{Categorical}(\pi)$$
$$x \sim \text{Exp}(\boldsymbol{u}, \boldsymbol{\eta}_z),$$

where $\text{Categorical}(\pi)$ is the categorical distribution, sampling values in 1 to K (1-of-K coding) with probabilities determined by π.

As suggested by the above equations, the use of a (set of) latent variables is useful. Hence, we define Z as the set of latent variables, similar to their definition in Bishop for GMMs. As such, we have similar equation summarising the main components of our EMM (*c.f.* Bishop Section 9.2):

$$p(x_n \mid \vartheta) \stackrel{\text{def}}{=} \sum_{k=1}^K \pi_k q(x \mid \boldsymbol{\eta}_k) \tag{3.1}$$
$$\pi_k \stackrel{\text{def}}{=} p(z_{nk} = 1) \tag{3.2}$$
$$p(x_n \mid z_n, \vartheta) \stackrel{\text{def}}{=} \prod_{k=1}^K \left(q(x_n \mid \boldsymbol{\eta}_k)\right)^{z_{nk}} \tag{3.3}$$
$$p(z_n) \stackrel{\text{def}}{=} \prod_{k=1}^K \pi_k^{z_{nk}}. \tag{3.4}$$

For notations sake, we will denote the densities for exponential family distributions by $q(x \mid \boldsymbol{\eta})$, following the convention set in Definition 1.

Question 3.1: Deriving the EMM expectation step (8 Points)

Derive the expectation we will maximise in EMM EM. In particular, show that

$$\sum_Z p(Z \mid X, \vartheta^{\text{old}}) \log p(X, Z \mid \vartheta) = \sum_{n=1}^N \sum_{k=1}^K \gamma^{\text{old}}(z_{nk}) \left(\log \pi_k + \log q(x_n \mid \boldsymbol{\eta}_k)\right) \tag{3.5}$$

where

$$\gamma^{\text{old}}(z_{nk}) = \frac{\pi_k^{\text{old}} q(x_n \mid \boldsymbol{\eta}_k^{\text{old}})}{\sum_j \pi_j^{\text{old}} q(x_n \mid \boldsymbol{\eta}_j^{\text{old}})}. \tag{3.6}$$

Hint: You will first need to derive the complete likelihood $p(X, Z \mid \vartheta^{\text{old}})$ and may want to use the identity

$$\sum_Z \{p(Z \mid X, \vartheta^{\text{old}}) z_{nk}\} = \sum_{z_n} \{p(z_n \mid X, \vartheta^{\text{old}}) z_{nk}\} = p(z_{nk} = 1 \mid x_n, \vartheta^{\text{old}}) \tag{3.7}$$

which holds because Z is independent in n and because z_n is a 1-of-K coding. Note that the summation over Z is in fact the summation over the possible latent variable matrices $Z = [z_{nk}]_{n,k=1}^{N,K}$.

Remark. Notice that these equation are similar to those presented in Bishop, namely Eqs. (9.40) and (9.39). Further note that Eq. (9.39) is incorrect (see errata on page 19).

It follows that in the E-step, we need to calculate $\left[\gamma^{\text{old}}(z_{nk})\right]_{n=1,k=1}^{N,K}$.

Now that we have the objective to maximise (the expectation to maximise), as given by Eq. (3.5), the EM updates can be derived. However, unlike in regular GMMs, the update that we initially derive do not correspond to the original *natural parameters* of the exponential distributions $\boldsymbol{H}$. Instead we actually get an update to the *expectation parameters* of the exponential distributions $\boldsymbol{\Lambda} \stackrel{\text{def}}{=} [\boldsymbol{\lambda}_k]_{k=1}^K$. That is, we get the update $\boldsymbol{H}^{\text{old}} \to \boldsymbol{\Lambda}$.

Despite this, there is in-fact a method to "re-project" the updated expectation parameters back to natural parameters. We will use without proof, that we have the transformation

$$\boldsymbol{\eta} = \nabla\varphi(\boldsymbol{\lambda}), \tag{3.8}$$

where φ is a convex function (you will not need to know the details[4]). Note the similarities of the transformation to that in Eq. (2.7).

Question 3.2: Deriving the EMM maximisation step (7 Points)

Derive the maximisation updates for EMM EM for mixing parameters π and expectation parameters $\boldsymbol{\Lambda}$. That is, assuming that Eq. (3.5) is concave in π and $\boldsymbol{H}$, show that

$$\pi_k = \frac{N_k}{N} \tag{3.9}$$

and

$$\boldsymbol{\lambda}_k = \frac{1}{N_k}\sum_{n=1}^{N}(\boldsymbol{u}(x_n)\cdot\gamma^{\text{old}}(z_{nk})), \tag{3.10}$$

where $N_k = \sum_{n=1}^{N}\gamma^{\text{old}}(z_{nk})$.

Thus show that we have the update equation for the natural parameters $\boldsymbol{H}$ given by

$$\boldsymbol{\eta}_k = \nabla\varphi\left(\frac{1}{N_k}\sum_{n=1}^{N}(\boldsymbol{u}(x_n)\cdot\gamma^{\text{old}}(z_{nk}))\right). \tag{3.11}$$

Hint: Use Lagrange multipliers for Eq. (3.9) to enforce the constraint $\sum_k \pi_k = 1$ and use Eq. (2.7) for Eq. (3.10).

Remark. The update in Eq. (3.11) can be thought of as a series of optimization and projection steps, as earlier eluded to. We do an optimization step to give the next $\boldsymbol{\lambda}$ and project the expectation parameters back into the natural parameter space.

Now that we have our update equation, its time for the algorithm's implementation. In particular, you are tasked to implement the EMM algorithm in its generality (as per prior equations and questions). The algorithm will then be tested in the simple case where the exponential families of the mixtures are simplified to Gaussian distributions (and thus simplifying EMMs to GMMs). In particular, we will use sufficient statistics function $\boldsymbol{u}(x) = (x, x^2)$, as presented in Question 2.3.

Question 3.3: EMM implementation (15 Points)

Implement `weighted_probs()`, `e_step_EMM()`, `m_step_EMM()` in `emm_question.py`.

You can test your implemented algorithm by utilising `implementation_viewer.ipynb`, a Jupyter notebook which includes as visualisation of 1-dimension GMMs.

Hint: Note that you can do this coding question without completing the correspond theory questions: the equations for the E- and M-steps are given in Question 3.1 and 3.2.

[4] It turns out that φ is the convex conjugate of the log-partition function ψ.

Bayesian Linear Regression. We now revisit Bayesian Linear Regression (BLR) for a more complex use case of EM. In particular, given a dataset $X = (\phi(X'), t) = \{(\phi(x'_n), t_n)\}_{n=1}^N \in \mathbb{R}^{N\times(M+1)}$ of training data, our model assumes that $t_n = y(\phi(x'_n), w) + \varepsilon$ where ϕ is our feature map, ε is small Gaussian noise, and w has a Gaussian prior. We will define $\Phi = \phi(X') \in \mathbb{R}^{N\times M}$ and $\phi_n = \phi(x'_n) \in \mathbb{R}^M$ to simplify notation.

This clearly is not a mixture model, but it fits the criteria for EM: we have incomplete data from which we cannot describe the full likelihood, but have other variables which would make it possible to. It is not possible to specify the probability of the data X and the parameters w by themselves, we need the parameters of the two gaussian distributions mentioned, which will be β^{-1} and α^{-1} respectively. Specifically,

$$p(t_n \mid \phi_n, w, \beta) = \mathcal{N}(t_n \mid y(\phi_n, w), \beta^{-1}) \tag{3.12}$$

$$p(w \mid \alpha) = \mathcal{N}(w \mid 0, \alpha^{-1} I). \tag{3.13}$$

It turns out it is easier to consider w as the latent variables which we will take the expectation over and $\{\alpha, \beta\}$ as the parameters of the model. Thus the complete data likelihood is

$$p(X, Z \mid \vartheta) = p(X, w \mid \alpha, \beta) = \prod_{n=1}^N p((t_n, \phi_n), w \mid \alpha, \beta) = \prod_{n=1}^N p(t_n \mid \phi_n, w, \beta) p(w \mid \alpha). \tag{3.14}$$

Question 3.4 [COMP8600]: Deriving the BLR expectation step (5 Points)

Derive the expectation we will maximise in BLR EM. In particular, show that

$$\mathop{\mathbb{E}}_{Z\mid X,\vartheta^{\text{old}}}[\log p(t, w \mid \alpha, \beta)] = \frac{N}{2}\log\left(\frac{\beta}{2\pi}\right) - \frac{\beta}{2}\sum_{n=1}^N \mathbb{E}\left[(t_n - w^\top \phi_n)^2\right] + \frac{M}{2}\log\left(\frac{\alpha}{2\pi}\right) - \frac{\alpha}{2}\mathbb{E}[w^\top w]. \tag{3.15}$$

With the objective for the expectation step of BLR EM, the update parameters for BLR can now be derived.

Question 3.5 [COMP8600]: Deriving the BLR maximisation step (Part 1) (2 Points)

Derive the maximisation updates for BLR EM That is, assuming that Eq. (3.15) is concave in α and β, maximise the Eq. (3.15) w.r.t. α to get

$$\alpha = \frac{M}{\mathbb{E}[w^\top w]}; \tag{3.16}$$

and maximise Eq. (3.15) w.r.t. β to get

$$\beta = \frac{N}{\sum_{n=1}^N \mathbb{E}\left[(t_n - w^\top \phi_n)^2\right]}. \tag{3.17}$$

Remark. Note that Eqs. (3.16) and (3.17) are the update rules for the BLR M-step.

Despite having update equations Eqs. (3.16) and (3.17) as presented in Question 3.5, these updates cannot be calculated directly as is. First we need to compute the denominators of each of the equations. Thankfully, there are some simple equation which can be derived for each of these quantities.

Question 3.6 [COMP8600]: Deriving the BLR maximisation step (Part 2) (3 Points)

Show that the following equations hold:

$$\mathbb{E}[w^\top w] = m_N^\top m_N + \mathrm{tr}(S_N) \tag{3.18}$$

$$\sum_{n=1}^{N} \mathbb{E}\left[(t_n - w^\top \phi_n)^2\right] = \|t - \Phi m_N\|_2^2 + \mathrm{tr}(\Phi^\top \Phi S_N), \tag{3.19}$$

where $m_N = \beta S_N \Phi^t t$ and $S_N^{-1} = \alpha I + \beta \Phi^\top \Phi$.

Hint: you may use that $w \sim \mathcal{N}(m_N, S_N)$ as per (3.49) in Bishop.

You may also use the following identities: given vector of random variables $v \in \mathbb{R}^n$ with mean $\mu \in \mathbb{R}^n$ and covariance $\Sigma \in \mathbb{R}^{n \times n}$, and constant vector $a \in \mathbb{R}^n$ and constant matrix $A \in \mathbb{R}^{n \times n}$

$$\mathbb{E}[v^\top a] = \mu^\top a \tag{3.20}$$

$$\mathbb{E}[v^\top A v] = \mathrm{tr}\left(\mu^\top A \mu\right) + \mathrm{tr}\left(A\Sigma\right). \tag{3.21}$$

Question 3.7: Implementing BLR within the EM framework (15 Points)

Implement `single_EM_iter_blr()` in `blr_question.py`.

You can test your implemented algorithm by utilising `implementation_viewer.ipynb`, a Jupyter notebook which includes as visualisation of the BLR EM algorithm.

Hint: Note that you can do these coding questions without having done the above theory questions: the equations for the E- and M-steps are given in the question text of Question 3.4 to 3.6.

A Mark Allocation

The following section details the mark allocation for both COMP4670 and COMP8600 students.

COMP8600 students are assigned all questions listed in the assignment (both gray and blue background). The total number of marks available is 120 marks.

COMP4670 students are given the choice of either completing only the non-COMP8600 questions (gray background) or all questions in the assignment (both gray and blue background). In either case, the student will receive the best total mark with respect to each selection of questions. Thus, an attempt in additional question will **not** lower your grade in any circumstance.

In addition to these base marks, students will receive 2 extra marks (capped at 100 or 120 marks) if their theory side of their assignment is submitted in LaTeX. We warn, that this may take a significant amount of time to complete depending on the student's familiarity; but encourage the use and application of typesetting for this assignment. A short post on piazza will be made regarding tips for learning LaTeX.

A template file on overleaf is also provided for students at: `http://quicklink.anu.edu.au/a4me,`.

In summary, the marks for students can be calculated as the following:

$$\text{COMP8600} = \frac{\min\{120, \text{LaTeX} + \text{All\_Questions}\}}{120}$$
$$\text{COMP4670} = \max\left\{\text{COMP8600}, \frac{\min\{100, \text{LaTeX} + \text{COMP4670\_Questions}\}}{100}\right\},$$

where $\text{LaTeX} = 2$; and a list of question to complete can be found below.

For COMP4670 students, the following consists of the questions to complete:

- Question 1.1: Your hunch — 0 Marks
- Question 1.2: Computing the probabilities — 4 Marks
- Question 1.3: Well-reasoned hunch — 1 Marks

- Question 2.1: Verifying valid distribution — 5 Marks
- Question 2.2: A Bayesian example (Part 1) — 13 Marks
- Question 2.3: A Bayesian example (Part 2) — 7 Marks
- Question 2.5: KL-Divergence for exponential families — 15 Marks
- Question 2.6: Pythagorean Theorem for exponential families — 10 Marks

---

- Question 3.1: Deriving the EMM expectation step — 8 Marks
- Question 3.2: Deriving the EMM maximisation step — 7 Marks
- Question 3.3: EMM implementation step — 15 Marks
- Question 3.7: BLR implementation step — 15 Marks

For COMP8600 students, the following additional questions are set to be complete:

- Question 2.4: MLE is equivalent to KL minimisation — 10 Marks

---

- Question 3.4: Deriving the BLR expectation step — 5 Marks
- Question 3.5: Deriving the BLR maximisation step (Part 1) — 2 Marks
- Question 3.6: Deriving the BLR maximisation step (Part 2) — 3 Marks

B The EM Algorithm.

The goal of EM is to find parameters ϑ that maximise a log likelihood $\log \ell(\vartheta) = \log p(X \mid \vartheta)$, X being known data, where $p(X \mid \vartheta)$ is unknown but adding unknown latent variables Z to give $p(X, Z \mid \vartheta)$ is known.

We can consider $\{X, Z\}$ the complete data set with complete data likelihood $p(X, Z \mid \vartheta)$, as given X and Z and ϑ we can calculate $p(X, Z \mid \vartheta)$, and $\{X\}$ as our incomplete known data set with incomplete data likelihood $p(X \mid \vartheta)$, as just given X and ϑ we cannot compute $p(X \mid \vartheta)$. Thus to maximise the incomplete data likelihood we rewrite it in terms of our known quantity by marginalising over the latent variables to get

$$\log \ell(\vartheta) = \log p(X \mid \vartheta) = \log \left(\sum_Z p(X, Z \mid \vartheta) \right). \tag{B.1}$$

It turns out this maximisation is usually hard to do in closed form, but would be easy if we knew what the values Z were (i.e. we had the complete data set), which leads to us using an iterative scheme that can find local maximums of this likelihood (the EM algorithm).

In the EM algorithm, we instead maximise the expectation of the complete log likelihood w.r.t. latent variables. Note that the probability of the latent variables depends on the parameters, which we are trying to maximise, so we formulate this expectation using previous values for the parameters:

$$\max_{\vartheta} \mathop{\mathbb{E}}_{Z \mid X, \vartheta^{\text{old}}} \log p(X, Z \mid \vartheta) = \sum_Z p(Z \mid X, \vartheta^{\text{old}}) \log p(X, Z \mid \vartheta). \tag{B.2}$$

This leads to an iterative scheme where each iteration is usually done in two steps:

- **Expectation Step / E-step**: calculate $p(Z \mid X, \vartheta^{\text{old}})$, allowing us to formulate the expectation;
- **Maximisation Step / M Step**: maximise the expectation to find a new estimate for the parameters $\vartheta^{\text{new}} = \max_\vartheta \sum_Z p(Z \mid X, \vartheta^{\text{old}}) \log p(X, Z \mid \vartheta)$.