Exam April 2023

(Q1) d-separation

(A) True or False? Justify your answer, e.g., give an open path, if any, or specify the relevant blocking variables. Give justifications (1.5p per question)

1. a and e are d-separated by c in Fig.1
2. a and c are d-separated by $\emptyset$ in Fig.1
3. a and e are d-separated by b in Fig.1

(B) Draw a DAG that is different from the DAG in Fig. 1, but Markov equivalent to it. Justfy your answer. (1.5p)

(Q2) Bayes rule

(A) In this question you must model a problem with 4 binary variables: G ('Gray'), V('Vancouver'), R ('Rain') and S ('sad'). Consider a Bayesian network for these variables with structure and conditional distributions as shown in Figure 2. Write down the expression for $p(S = 1|S = 1)$ in terms of $\alpha, \beta, \gamma, \delta$ (4p)

(B) The goal in modeling is often to calculate a predictive distribution p(x*|x), where x denotes some observed data and x* correspondingly unobserved future data. Let θ represent the parameters of the model and assume a model, i.e., a Bayesian network, as follows: $x^* \leftarrow \theta \rightarrow x$. Derive the formula for $p(x^*|x)$ using the maximum likelihood (ML) approach, and compare it with the Bayesian approach. How is maximum likelihood a special case of the Bayesian approach? (2p)

+ Code + Markdown

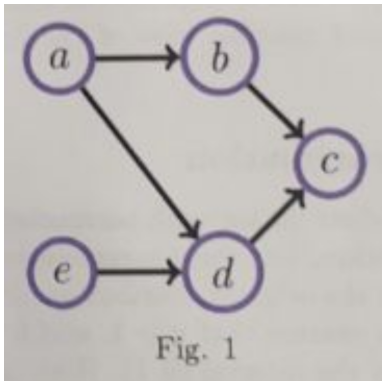
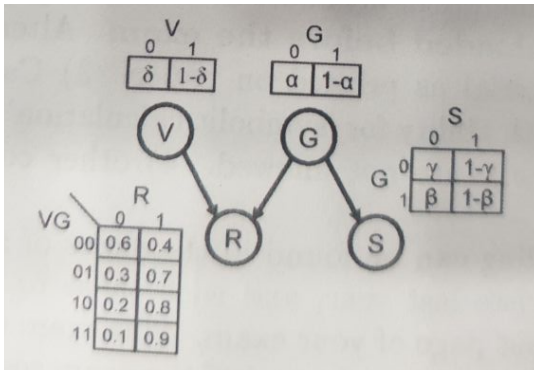


Fig. 1

(3) Laplace approximation

Approximate the Beta distribution with parameters a and b, Beta(x|a,b) using the Laplace approximation, i.e., the approximating distribution is the Gaussian centered at the mode of the original distribution. Parameters a and b are known constants, and you can assume that a > 1, and b > 1, such that the Beta distribution has a mode in the interval (0,1).

Hint: use $E(x) = -\log Beta(x|a, b)$ as the starting point (6p)



(Q3) Marginal likelihood

(A) Suppose the data $\mathbf{x} = \{x_n\}_{n=1}^N$ are distributed as $x_n \overset{\text{i.i.d}}{\sim} Exp(\lambda)$ and assume a prior $\lambda \sim Gamma(\alpha, \beta)$. Derive the marginal likelihood $p(\mathbf{x}|\alpha, \beta)$.Hint: The Gamma prior is conjugate to the Exponential likelihood. (4p)

(B) Give two example of possible uses of the calculated marginal likelihood $p(\mathbf{x}|\alpha, \beta)$. (2p)

(4) Laplace approximation

Approximate the Beta distribution with parameters a and b, Beta(x|a,b) using the Laplace approximation, i.e., the approximating distribution is the Gaussian centered at the mode of the original distribution. Parameters a and b are known constants, and you can assume that a > 1, and b > 1, such that the Beta distribution has a mode in the interval (0,1).

Hint: use $E(x) = -\log Beta(x|a, b)$ as the starting point (6p)

(Q5) EM algorithm

Consider a simple factor analysis model:

$$v_n \sim N_2(\mathbf{a}h_n; \lambda^{-1}I), n = 1, ..., N$$

$$h_n \sim N(0, 1), n = 1, ..., N$$

where $v_n \in R^2$ and $h_n \in R$ for all $n = 1, ..., N$. Parameters of the model are the loading matrix (a vector in this case), $\mathbf{a} \in R^2$, and precision (inverse variance) $\lambda^2 \in R$.

(A) Derive and simplify the complete data log-likelihood. (2p)

(B) It can be shown that the posterior distribution $p(h_n|v_n; a_0; \lambda_0)$, where $w_0; \lambda_0$ are current estimates of the parameters, is a Gaussian $\mathcal{N}(h_n|\mu_n; \lambda_z^{-1})$ with certain μ_n and λ_z. Derive formulas for μ_n and σ_z. (2p)

(C) Derive the Q function needed in the E step of the EM algorithm, and express it using μ_n and λ_z (2p)

Hint 1: You can solve C even if you did not solve B because the solution to C can be given using μ_n and λ_z.

Hint 2: Completing the square.

Hint 3: $Var(X) = E(X^2) - E(X)^2$

Exam April 2021

Q1) Bayes' rule

Consider three binary variables x_1, x_2, and x_3. Their joint distribution factorizes as $p(x_1, x_2, x_3) = p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_2)$, where $p(x_1 = 0) = 0.4$, $p(x_2 = 0 \mid x_1 = 0) = 0.7$, $p(x_2 = 0 \mid x_1 = 1) = 0.5$, $p(x_3 = 0 \mid x_2 = 0) = 0.9$, and $p(x_3 = 0 \mid x_2 = 1) = 0.2$.

A) Draw the DAG corresponding to the model. (1p)

p(x1=0)
0.4

x1	p(x2=0)
0	0.7
1	0.5

x2	p(x3=0)
0	0.9
1	0.2

B) Compute $p(x_2 = 1 \mid x_3 = 1)$. (5p)

$$
\begin{aligned}
&p(x_2 = 1) = \\
&p(x_1 = 0)p(x_2 = 1|x_1 = 0) + \\
&p(x_1 = 1)p(x_2 = 1|x_1 = 1) \\
&= 0.4 \times 0.3 + 0.6 \times 0.5 = 0.42
\end{aligned}
$$

$$
\begin{aligned}
&p(x_3 = 1) = \\
&p(x_1 = 0)p(x_2 = 0|x_1 = 0)p(x_3 = 1|x_2 = 0) + \\
&p(x_1 = 0)p(x_2 = 1|x_1 = 0)p(x_3 = 1|x_2 = 1) + \\
&p(x_1 = 1)p(x_2 = 0|x_1 = 1)p(x_3 = 1|x_2 = 0) + \\
&p(x_1 = 1)p(x_2 = 1|x_1 = 1)p(x_3 = 1|x_2 = 1) = \\
&0.4 \times 0.7 \times 0.1 + \\
&0.4 \times 0.3 \times 0.8 + \\
&0.6 \times 0.5 \times 0.1 + \\
&0.6 \times 0.5 \times 0.8 = \\
&0.394
\end{aligned}
$$

$$
p(x_2 = 1|x_3 = 1) = \frac{p(x_3 = 1|x_2 = 1)p(x_2 = 1)}{p(x_3 = 1)} = 0.8 \times 0.42/0.394 = \frac{168}{197} = 0.8527
$$

d-separation

Definition 20 (d-connection, d-separation)**.** If G is a directed graph in which $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Z}$ are disjoint sets of vertices, then $\mathcal{X}$ and $\mathcal{Y}$ are d-connected by $\mathcal{Z}$ in G if and only if there exists an undirected path U between some vertex in $\mathcal{X}$ and some vertex in $\mathcal{Y}$ such that for every collider C on U, either C or a descendent of C is in $\mathcal{Z}$, and no non-collider on U is in $\mathcal{Z}$.

$\mathcal{X}$ and $\mathcal{Y}$ are d-separated by $\mathcal{Z}$ in G if and only if they are not d-connected by $\mathcal{Z}$ in G.

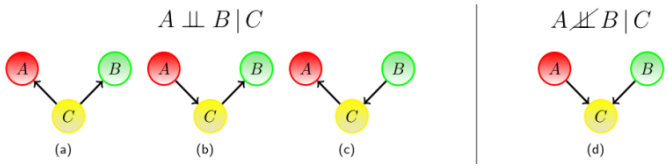
One may also phrase this as follows. For every variable $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, check every path U between x and y. A path U is said to be *blocked* if there is a node w on U such that either

1. w is a collider and neither w nor any of its descendants is in $\mathcal{Z}$.
2. w is not a collider on U and w is in $\mathcal{Z}$.

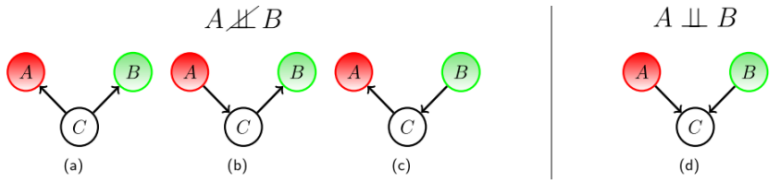
If all such paths are blocked then $\mathcal{X}$ and $\mathcal{Y}$ are d-separated by $\mathcal{Z}$.

If the variable sets $\mathcal{X}$ and $\mathcal{Y}$ are d-separated by $\mathcal{Z}$, they are independent conditional on $\mathcal{Z}$ in all probability distributions such a graph can represent.

- Possible BNs with three nodes and two links



- In (a), (b), and (c), A and B are **conditionally independent** given C.
  - $p(a, b|c) = p(a|c)p(b|c)$
- In (d), A and B are not conditionally independent given C
  - $p(a, b|c) \propto p(a)p(b)p(c|a, b)$



- In (a), (b), and (c), A and B are marginally dependent
- In (d) the variables A and B are **marginally independent**

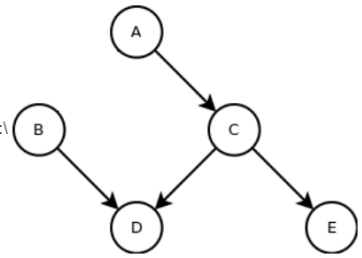
$$p(a,b) = \sum_c p(a,b,c) = \sum_c p(a)p(b)p(c|a,b) = p(a)p(b)$$

conditional independence is not commutative in general, even given a set of condition variables.

Conditional independence is a property of probability distributions that specifies that two variables are independent given a third variable, or set of variables, that serves as a conditioning context. Specifically, if X and Y are conditionally independent given Z, denoted as X ⊥ Y | Z, then knowing the value of Z provides no additional information about the relationship between X and Y beyond what is already implied by their marginal distributions.

However, conditional independence does not necessarily commute, meaning that X ⊥ Y | Z does not imply Y ⊥ X | Z. This is because the set of condition variables, Z, may affect the relationships between X and Y in different ways, depending on the specific distribution being considered. In other words, the direction of conditioning matters in determining whether two variables are conditionally independent.

From the description given by the expert, we get the following DAG:\



The conditional independencies corresponding to the expert's description are:

$E \perp\!\!\!\perp A \mid C$
$E \perp\!\!\!\perp B \mid C$
$E \perp\!\!\!\perp D \mid C$
$D \perp\!\!\!\perp E \mid \{B, C\}$ (already implicit in the above statement)
$D \perp\!\!\!\perp A \mid \{B, C\}$
$C \perp\!\!\!\perp B \mid A$

Note that C is a direct cause of D and E, which is why $C \not\perp\!\!\!\perp D \mid A$ and $C \not\perp\!\!\!\perp E \mid A$. Also note that the expert did not say anything about a possible causal dependence between A and B. However, the above conditional independence statements would remain the same even if we added an edge between A and B (in either direction).

The joint distribution factorized according to the DAG is:

http://bayes.cs.ucla.edu/BOOK-2K/d-sep.html

two adjacent nodes in a DAG (directed acyclic graph) cannot be d-separated (separated by observed variables). If two nodes in a DAG are adjacent (i.e., connected by a single edge), they have parent-child relationship (in whatever order), and therefore, there will always axist the path U that d-connects them via this edge, no matter what set of conditions Z. Therefore, we can remove adjacent nodes from consideration.

Rule 1: *x* and *y* are *d*-connected if there is an unblocked path between them (one that doesnt have any colliders)

Rule 2: *x* and *y* are *d*-connected, conditioned on a set *Z* of nodes, if there is a collider-free path between *x* and *y* that traverses no member of *Z.* If no such path exists, we say that *x* and *y* are *d*-separated by *Z,* We also say then that every path between *x* and *y* is "blocked" by *Z.*

Rule 3: *If a collider is a member of the conditioning set Z, or has a descendant in Z, then it no longer blocks any path that traces this collider.*

We have 15 distinct pairs, removing pairs with adjacent nodes A-B, A-D, B-E, B-C, C-E, C-F, D-F => We have 8 pairs left to consider. The colliders are E and F

Strategy: list out all possible paths for each pair, then for each path, cross out the paths that have the colliders. These are unconditional blocks. With the unblocked paths left, pick out at least 1 node from each path and append to the set Z, because if we dont do so, there will be an unblocked path not going through any elements of Z. By simplicity, we can also union all the blocked paths into set Z. Note: a node can be a collider on a graph as a whole, but with respect to a certain path, it is not necessarily a collider.

Q2) Bayesian networks

List all pairs of variables that are d-separated in the DAG in Fig. 1; for each pair of d-separated variables, give one set that d-separates those variables. (4p)

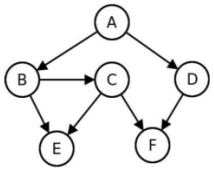


Figure 1

Pair A - C:

+ A-B-C not blocked
+ A-B-E-C blocked
+ A-D-F-C blocked

=> {B} d-separates A and C because B is not a collider on U (path A-B-C) and B is in the set

A - E
- + A-B-E not blocked
- + A-B-C-E not blocked
- + Others blocked

=> Pick one node from each unblocked path. Since B is on both paths,
=> {B} d-separates A and E. {B, C} also blocks the path

A - F
- + A-D-F not blocked
- + A-B-C-F not blocked
- + A-B-E-C-F blocked

=> {D-B}, {D, C} or {D,B,C} d-separates A and F

B - D
- + B-A-D not blocked
- + B-C-F-D blocked
- + B-E-C-F-D blocked

=> {A} d-separates B and D

B - F
- + B-C-F not blocked
- + B-E-C-F blocked
- + B-A-D-F not blocked

=> {C, A}, {C, D} or {C, A, D} d-separates B and F

C - D
- + C-F-D blocked
- + C-B-A-D not blocked
- + C-E-B-A-D blocked

=> {B}, {A} or {B, A} d-separates C and D

D - E
- + D-A-B-E not blocked
- + D-F-C-E blocked
- + D-F-C-B-E blocked

=> {B}, {A} or {B, A} d-separates D and E

E - F
- + E-C-F not blocked
- + E-B-A-D-F not blocked
- + E-C-B-A-D-F not blocked

=> {C, B}, {C, A}, {C,D} and their combinations d-separates E and F

Q3) Variational inference

Assume that N observations x_n, $n = 1, \ldots, N$ have been generated from the following mixture model:

$$p(x_n|\tau, \lambda_1, \lambda_2) = \tau N(x_n|0, \lambda_1^{-1}) + (1-\tau)N(0, \lambda_2^{-1}),$$

where λ_1 and λ_2 are the unknown precisions (inverse variances) of the two components, and τ is the mixing coefficient. Assume the following prior distributions:

$$\tau \sim Beta(\alpha_0, \alpha_0), \quad \lambda_1 \sim Gamma(a_0, b_0), \quad \lambda_2 \sim Gamma(c_0, d_0).$$

A) Define the model using latent variables $\mathbf{z} = \{z_i\}_{i=1}^N$. (1p)

We formulate the model using latent variables $z_n = (z_1, ..., z_N)$ which explicitly specify the component responsible for generating observation x_n. In detail,

$$z_n = (z_{n1}, z_{n2})^T = \begin{cases} (1,0)^T, & (x_n \text{ is from } \mathcal{N}(x_n|0, \lambda_1^{-1})) \\ (0,1)^T, & (x_n \text{ is from } \mathcal{N}(x_n|0, \lambda_2^{-1})) \end{cases}$$

and place a prior on the latent variables

$p(\mathbf{z}|\tau) = \prod_{n=1}^N \tau^{z_{n1}}(1-\tau)^{z_{n2}}$

The likelihood in the latent variable model is given by

$p(\mathbf{x}|\mathbf{z}, \lambda_1, \lambda_2) = \prod_{n=1}^N \mathcal{N}(x_n|0, \lambda_1^{-1})^{z_{n1}} \mathcal{N}(x_n|0, \lambda_1^{-1})^{z_{n2}}$

The joint distribution of all observed ($\mathbf{x}$) and unobserved variables ($\mathbf{z}, \tau, \lambda_1, \lambda_2$) factorizes as follows

$p(\mathbf{x}, \mathbf{z}, \tau, \lambda_1, \lambda_2) = p(\tau)p(\lambda_1)p(\lambda_2)p(\mathbf{z}|\tau)p(\mathbf{x}|\mathbf{z}, \lambda_1, \lambda_2)$

and the log of the joint distribution can correspondingly be written as

$\log p(\mathbf{x}, \mathbf{z}, \tau, \lambda_1, \lambda_2) = \log p(\tau) + \log p(\lambda_1) + \log p(\lambda_2) + \log p(\mathbf{z}|\tau) + \log p(\mathbf{x}|\mathbf{z}, \lambda_1, \lambda_2)$

We approximate the posterior distribution $p(\mathbf{z}, \tau, \lambda_1, \lambda_2|\mathbf{x})$ using the factorized variational distribution $q(\mathbf{z})q(\tau)q(\lambda_1)q(\lambda_2)$

B) Derive the variational update for λ_2. You can assume the mean-field approximation:

$$q(\mathbf{z},\tau,\lambda_1,\lambda_2) = q(\lambda_1)q(\lambda_2)q(\tau)\prod_n q(z_n)$$

and assume the other factors are given by

$$q(\tau) = Beta(\tau|\alpha_n,\beta_n), \quad q(z_{n1}) = Bernoulli(z_{n1}|r_{n1}), \quad q(\lambda_1) = Gamma(\lambda_1|a_n,b_n).$$

(5p)

Simple example

- Model: assume that we have observations $\mathbf{x} = (x_1,\ldots,x_N)$ s.t.

$$p(x_n|\theta,\tau) = (1-\tau)N(x_n|0,1) + \tau N(x_n|\theta,1)$$

Prior:

$$\tau \sim Beta(\alpha_0,\alpha_0) \qquad \theta \sim N(0,\beta_0^{-1})$$

Formulation using latent variables $\mathbf{z} = (z_1,\ldots,z_n)$:

$$p(\mathbf{z}|\tau) = \prod_{n=1}^N \tau^{z_{n2}}(1-\tau)^{z_{n1}}$$
$$p(\mathbf{x}|\mathbf{z},\theta) = \prod_{n=1}^N N(x_n|0,1)^{z_{n1}} N(x_n|\theta,1)^{z_{n2}}$$

- *simple_vb_example.pdf*, and the next exercise.

Important points

- Variational Bayes aims to find a tractable approximation $q(\mathbf{z})$ for the posterior distribution $p(\mathbf{z}|\mathbf{x})$.
- $q(\mathbf{z})$ is found by maximizing the ELBO $\mathcal{L}(q)$ or, equivalently, by minimizing $KL(q||p)$.
- Mean-field VB: if $q(\mathbf{z}) = \prod_{i=1}^M q_i(\mathbf{z}_i)$, factor $q_j(\mathbf{z}_j)$ can be updated using

$$\log q_j^*(\mathbf{z}_j) = E_{q(\mathbf{z}_{\setminus j})}[\log p(\mathbf{x},\mathbf{z})] + \text{const.}$$

- Variational approximation for a fully Bayesian model with prior distributions avoids some of the problems related to the ML estimation of the GMM (overfitting, singularities).

Update of factor $q(\lambda_2)$

$\log q^*(\lambda_2) = E_{q(\mathbf{z})q(\tau)q(\lambda_1)}[\log p(\mathbf{x}, \mathbf{z}, \tau, \lambda_1, \lambda_2)]$

=> $\log q^*(\lambda_2) = E_{q(\mathbf{z})q(\tau)q(\lambda_1)}[\log p(\tau) + \log p(\lambda_1) + \log p(\lambda_2) + \log p(\mathbf{z}|\tau) + \log p(\mathbf{x}|\mathbf{z}, \lambda_1, \lambda_2)]$

To derive the variational update for λ_2, we need to find the optimal $q(\lambda_2)$ that minimizes the KL divergence between the true posterior $p(z, \tau, \lambda_1, \lambda_2 \mid x)$ and the approximating distribution $q(z)q(\tau)q(\lambda_1)q(\lambda_2)$. This can be done by applying the coordinate ascent variational inference (CAVI) algorithm.

The CAVI update for $q(\lambda_2)$ is given by taking the expectation of the log joint distribution with respect to all other factors and then exponentiating the result. We need to keep only the terms dependent on λ_2 (having λ_2 in the term). The rest terms are constant with respect to this factor can be added to the constant "C"

=> $\log q^*(\lambda_2) = E_{q(\mathbf{z})q(\tau)q(\lambda_1)}[\log p(\lambda_2)] + E_{q(\mathbf{z})q(\tau)q(\lambda_1)}[\log p(\mathbf{x}|\mathbf{z}, \lambda_1, \lambda_2)] + C$

=> $\log q^*(\lambda_2) = \log p(\lambda_2) + E_{q(\mathbf{z})q(\tau)q(\lambda_1)}[\log p(\mathbf{x}|\mathbf{z}, \lambda_1, \lambda_2)] + C$ (E.q 1)

Additionally, we have

$E_{q(\mathbf{z})q(\tau)q(\lambda_1)}[\log p(\mathbf{x}|\mathbf{z}, \lambda_1, \lambda_2)] = \prod_{n=1}^{N} \mathcal{N}(x_n|0, \lambda_1^{-1})^{z_{n1}} \mathcal{N}(x_n|0, \lambda_2^{-1})^{z_{n2}}$

=> $E_{q(\mathbf{z})q(\tau)q(\lambda_1)}[\log p(\mathbf{x}|\mathbf{z}, \lambda_1, \lambda_2)] = E_{q(\mathbf{z})q(\tau)q(\lambda_1)}[\sum_{n=1}^{N} z_{n1} \log \mathcal{N}(x_n|0, \lambda_1^{-1}) + z_{n2} \log \mathcal{N}(x_n|0, \lambda_2^{-1})]$. We can drop the term that is independent of λ_2, which can be treated as a constant

=> $E_{q(\mathbf{z})q(\tau)q(\lambda_1)}[\log p(\mathbf{x}|\mathbf{z}, \lambda_1, \lambda_2)] = E_{q(\mathbf{z})q(\tau)q(\lambda_1)}[\sum_{n=1}^{N} z_{n2} \log \mathcal{N}(x_n|0, \lambda_2^{-1})] + C$

By definition, $E_{q(z_n)}[\sum_{n=1}^{N} z_{nk}] = \sum_{n=1}^{N} r_{nk}$ is the expected responsibility of component k for observation x_n according to Bernoulli distribution

=> $E_{q(\mathbf{z})q(\tau)q(\lambda_1)}[\log p(\mathbf{x}|\mathbf{z}, \lambda_1, \lambda_2)] = \sum_{n=1}^{N} r_{n2} \log \mathcal{N}(x_n|0, \lambda_2^{-1}) + C$ (E.q 3)

Plugging (2)(3) into equation (1), we have:

$\log q^*(\lambda_2) = (c_0 - 1) \log \lambda_2 - d_0 \lambda_2 + \sum_{n=1}^{N} r_{n2} \log \mathcal{N}(x_n|0, \lambda_2^{-1}) + C$

=> $\log q^*(\lambda_2) = (c_0 - 1)\log\lambda_2 - d_0\lambda_2 + \sum_{n=1}^{N} r_{n2}\log[(2\pi\lambda_2^{-1})^{-1/2}\exp(\frac{1}{2}(x_n - 0)^2(\lambda_2^{-1})^{-1})] + C$

=> $\log q^*(\lambda_2) = (c_0 - 1)\log\lambda_2 - d_0\lambda_2 + \sum_{n=1}^{N} r_{n2}\log[(2\pi)^{-1/2}\lambda_2^{1/2}\exp(\frac{1}{2}x_n^2\lambda_2)] + C$

=> $\log q^*(\lambda_2) = (c_0 - 1)\log\lambda_2 - d_0\lambda_2 + \sum_{n=1}^{N} r_{n2}\left[-\frac{1}{2}\log(2\pi) + \frac{1}{2}\log\lambda_2 + \frac{1}{2}x_n^2\lambda_2\right] + C$

=> $\log q^*(\lambda_2) = (c_0 - 1)\log\lambda_2 - d_0\lambda_2 - \frac{N}{2}r_{n2}\log(2\pi) + \frac{N}{2}r_{n2}\log\lambda_2 + \frac{1}{2}\sum_{n=1}^{N} r_{n2}x_n^2\lambda_2 + C$

=> $\log q^*(\lambda_2) = (c_0 - 1)\log\lambda_2 - d_0\lambda_2 - \frac{1}{2}\sum_{n=1}^{N} r_{n2}\log(2\pi) + \frac{1}{2}\sum_{n=1}^{N} r_{n2}\log\lambda_2 + \frac{1}{2}\sum_{n=1}^{N} r_{n2}x_n^2\lambda_2 + C$

Removing constant term $\sum_{n=1}^{N} r_{n2}\log(2\pi)$:

=> $\log q^*(\lambda_2) = (c_0 - 1)\log\lambda_2 - d_0\lambda_2 + \frac{1}{2}\sum_{n=1}^{N} r_{n2}\log\lambda_2 + \frac{1}{2}\sum_{n=1}^{N} r_{n2}x_n^2\lambda_2 + C$

=> $\log q^*(\lambda_2) = (c_0 + \frac{1}{2}\sum_{n=1}^{N} r_{n2} - 1)\log\lambda_2 - (d_0 + \frac{1}{2}\sum_{n=1}^{N} r_{n2}x_n^2)\lambda_2 + C$

=> $q^*(\lambda_2) \propto \exp(c_0 + \frac{1}{2}\sum_{n=1}^{N} r_{n2} - 1)\log\lambda_2 - (d_0 + \frac{1}{2}\sum_{n=1}^{N} r_{n2}x_n^2)\lambda_2$, which resembles the Gamma distribution, as the prior of λ_2 is Gamma distribution, which is conjugate to the posterior:

$q^*(\lambda_2) = \text{Gamma}(c_N, d_N)$

where

$c_N = c_0 + \frac{1}{2}\sum_{n=1}^{N} r_{n2}$

$d_N = d_0 + \frac{1}{2}\sum_{n=1}^{N} r_{n2}x_n^2$

Q4) EM algorithm

Consider N observations x_n, $n = 1, \ldots, N$, from a two-component mixture of binomial distributions

$$p(x_n \mid \theta, q_1, q_2) = \theta Bin(x_n \mid q_1) + (1 - \theta)Bin(x_n \mid q_2).$$

A) Represent the model using latent variables and derive the E step of the expectation maximization. In the end, simplify the Q-function, $Q(\theta, q_1, q_2|\theta^0, q_1^0, q_2^0)$, where θ^0, q_1^0, q_2^0 are the current values of the parameters. (4p)

Idea of the EM algorithm (1/2)

- Let X denote the observed data, and θ model parameters. The goal in maximum likelihood is to find $\widehat{\theta}$:

$$\widehat{\theta} = \arg\max_{\theta} \{\log p(X|\theta)\}$$

- If model contains latent variables Z, the log-likelihood is given by

$$\log p(X|\theta) = \log \left\{ \sum_{Z} p(X, Z|\theta) \right\},$$

  which may be difficult to maximize analytically
- Possible solutions: 1) numerical optimization, 2) the EM algorithm (expectation-maximization)

Idea of the EM algorithm (2/2)

- X: **observed** data, Z: **unobserved** latent variables
- $\{X, Z\}$: **complete** data, X: **incomplete** data
- In EM algorithm, we assume that the complete data log-likelihood:

$$\log p(X, Z|\theta)$$

  is easy to maximize.
- Problem: Z is not observed
- Solution: maximize

$$\begin{aligned} Q(\theta, \theta_0) &\equiv E_{Z|X,\theta_0} [\log p(X, Z|\theta)] \\ &= \sum_{Z} p(Z|X, \theta_0) \log p(X, Z|\theta) \end{aligned}$$

  where $p(Z|X, \theta_0)$ is the posterior distribution of the latent variables computed using the current parameter estimate θ_0

EM algorithm in detail

Goal: maximize $\log p(X|\theta)$ w.r.t. θ

1. Initialize θ_0
2. **E-step** Evaluate $p(Z|X, \theta_0)$, and then compute

$$Q(\theta, \theta_0) = E_{Z|X,\theta_0}\left[\log p(X, Z|\theta)\right] = \sum_Z p(Z|X, \theta_0) \log p(X, Z|\theta)$$

3. **M-step** Evaluate θ^{new} using

$$\theta^{new} = \arg\max_\theta Q(\theta, \theta_0).$$

   Set $\theta_0 \leftarrow \theta^{new}$
4. Repeat **E** and **M** steps until convergence

We formulate the model using latent variables $z_n = (z_1, ..., z_N)$ which explicitly specify the component responsible for generating observation x_n. In detail,

$$z_n = (z_{n1}, z_{n2})^T = \begin{cases} (1,0)^T, & (x_n \text{ is from } Bin(x_n|q_1)) \\ (0,1)^T, & (x_n \text{ is from } Bin(x_n|q_2)) \end{cases}$$

and place a prior on the latent variables

$p(\mathbf{z}|\theta) = \prod_{n=1}^N \theta^{z_{n1}}(1-\theta)^{z_{n2}}$

The likelihood in the latent variable model is given by

$p(\mathbf{x}|\mathbf{z}, q_1, q_2) = \prod_{n=1}^N Bin(x_n|q_1)^{z_{n1}} Bin(x_n|q_2)^{z_{n2}}$

The joint distribution of all observed ($\mathbf{x}$) and unobserved variables ($\mathbf{z}, \theta, q_1, q_2$) factorizes as follows

$p(\mathbf{x}, \mathbf{z}, \theta, q_1, q_2) = p(\theta)p(q_1)p(q_2)p(\mathbf{z}|\theta)p(\mathbf{x}|\mathbf{z}, q_1, q_2)$

and the log of the joint distribution can correspondingly be written as

$\log p(\mathbf{x}, \mathbf{z}, \theta, q_1, q_2) = \log p(\theta) + \log p(q_1) + \log p(q_2) + \log p(\mathbf{z}|\theta) + \log p(\mathbf{x}|\mathbf{z}, q_1, q_2)$

The complete data log-likelihood is the joint log likelihood of both the observed variable, $\mathbf{x}$ and the latent variable, $\mathbf{z}$. In the EM-algorithm we will maximize the expectation of the log-likelihood of the complete data $(\mathbf{x}, \mathbf{z})$. The log-likelihood is:

$\log p(\mathbf{x}, \mathbf{z}|\theta, q_1, q_2) = \log \left\{ \prod_{n=1}^{N} p(x_n, z_n|\theta, q_1, q_2) \right\} = \sum_{n=1}^{N} \log p(x_n, z_n|\theta, q_1, q_2)$

=> $\log p(\mathbf{x}, \mathbf{z}|\theta, q_1, q_2) = \sum_{n=1}^{N} \log \left(p(x_n|z_n, q_1, q_2) p(z_n|\theta) \right)$

=> $\log p(\mathbf{x}, \mathbf{z}|\theta, q_1, q_2) = \sum_{n=1}^{N} \log \left(\theta^{z_{n1}} Bin(x_n|q_1)^{z_{n1}} (1-\theta)^{z_{n2}} Bin(x_n|q_2)^{z_{n2}} \right)$

=> $\log p(\mathbf{x}, \mathbf{z}|\theta, q_1, q_2) = \sum_{n=1}^{N} z_{n1} \log \left(\theta Bin(x_n|q_1) \right) + z_{n2} \log \left((1-\theta) Bin(x_n|q_2) \right)$

=> $\log p(\mathbf{x}, \mathbf{z}|\theta, q_1, q_2) = \sum_{n=1}^{N} \left(z_{n1} \log \theta + z_{n1} \log Bin(x_n|q_1) + z_{n2} \log(1-\theta) + z_{n2} \log Bin(x_n|q_2) \right)$ (E.q 1)

E-step 1^0

Compute the posterior distribution of the latent variables, given the current estimate θ^0 of θ, q_1^0 of q_1 and q_2^0 of q_2:

$p(z_{n1} = 1|x_n, \theta^0, q_1^0, q_2^0) \propto \log p(z_{n1} = 1, x_n|\theta, q_1, q_2) = p(z_{n1} = 1) p(x_n|z_n, \theta^0, q_1^0, q_2^0) = \theta^0 Bin(x_n|q_1^0)$ (E.q 2)

$p(z_{n2} = 1|x_n, \theta^0, q_1^0, q_2^0) \propto \log p(z_{n1} = 1, x_n|\theta, q_1, q_2) = p(z_{n2} = 1) p(x_n|z_n, \theta^0, q_1^0, q_2^0) = (1-\theta_0) Bin(x_n|q_2^0)$ (E.q 3)

By normalizing these two equations E.q 2 and E.q 3, we get:

$\gamma(z_{n1}) = p(z_{n1} = 1|x_n, \theta_0, q_1^0, q_2^0) = \dfrac{\theta^0 Bin(x_n|q_1^0)}{\theta^0 Bin(x_n|q_1^0) + (1-\theta_0) Bin(x_n|q_2^0)}$ (E.q 4)

$\gamma(z_{n2}) = p(z_{n2} = 1|x_n, \theta_0, q_1^0, q_2^0) = \dfrac{(1-\theta^0) Bin(x_n|q_2^0)}{\theta^0 Bin(x_n|q_1^0) + (1-\theta_0) Bin(x_n|q_2^0)}$ (E.q 5)

E-step 2^0

Evaluate the expectation of the complete data log-likelihood over the posterior distribution of the latent variables in E.q 4 and E.q 5

=> $\log p(\mathbf{x}, \mathbf{z}|\theta, q_1, q_2) = \sum_{n=1}^{N} \left(z_{n1} \log \theta + z_{n1} \log Bin(x_n|q_1) + z_{n2} \log(1-\theta) + z_{n2} \log Bin(x_n|q_2) \right)$

$\mathcal{Q}(\theta, q_1, q_2|\theta^0, q_1^0, q_2^0) = E_{z|x, \theta^0, q_1^0, q_2^0}[\log p(\mathbf{x}, \mathbf{z}|\theta, q_1, q_2)]$

=> $\mathcal{Q}(\theta, q_1, q_2|\theta^0, q_1^0, q_2^0) = \sum_{n=1}^{N} \sum_{k=1}^{2} [p(\mathbf{z_{nk}}|\mathbf{x_n}, \theta^0, q_1^0, q_2^0) \log p(\mathbf{x_n}, \mathbf{z_{nk}}|\theta, q_1, q_2)]$

=> $\mathcal{Q}(\theta, q_1, q_2|\theta^0, q_1^0, q_2^0) =$
$\sum_{n=1}^{N} [p(\mathbf{z_{n1}}|\mathbf{x_n}, \theta^0, q_1^0, q_2^0) \log p(\mathbf{x_n}, \mathbf{z_{n1}}|\theta, q_1, q_2) +$
$p(\mathbf{z_{n2}}|\mathbf{x_n}, \theta^0, q_1^0, q_2^0) \log p(\mathbf{x_n}, \mathbf{z_{n2}}|\theta, q_1, q_2)]$

=> $\mathcal{Q}(\theta, q_1, q_2|\theta^0, q_1^0, q_2^0) = \sum_{n=1}^{N} \gamma(z_{n1}) \log(\theta Bin(x_n|q_1)) + \gamma(z_{n2}) \log((1-\theta) Bin(x_n|q_2))$

=> $\mathcal{Q}(\theta, q_1, q_2|\theta^0, q_1^0, q_2^0) = \sum_{n=1}^{N} \gamma(z_{n1}) \log(\theta) + \gamma(z_{n1}) \log Bin(x_n|q_1) + \gamma(z_{n2}) \log(1-\theta) +$
$\gamma(z_{n2}) \log Bin(x_n|q_2)$ (answer)

B) Derive the M-step for the θ parameter. (2p)
The binomial distribution has a probability mass function of the form

$$f(k|m,q) = p(x_n = k) = \binom{m}{k} q^k (1-q)^{m-k},$$

where $0 \leq k \leq m$ is an integer. You can treat m as a known constant.

M-step

Maximize $\mathcal{Q}(\theta, q_1, q_2 | \theta^0, q_1^0, q_2^0)$ with respect to θ, q_1 and q_2.

Maximizing for θ:

$$\frac{d}{d\theta}\mathcal{Q}(\theta;\theta_0) = \frac{d}{d\theta}\sum_{n=1}^{N} \gamma(z_{n1})\log(\theta) + \gamma(z_{n1})\log Bin(x_n|q_1) + \gamma(z_{n2})\log(1-\theta) + \gamma(z_{n2})\log Bin(x_n|q_2)$$

$$\frac{d}{d\theta}\mathcal{Q}(\theta;\theta_0) = \sum_{n=1}^{N}\frac{\gamma(z_{n1})}{\theta} + 0 - \frac{\gamma(z_{n2})}{1-\theta} + 0 = \sum_{n=1}^{N}\frac{\gamma(z_{n1})}{\theta} - \frac{\gamma(z_{n2})}{1-\theta} = \frac{N_1}{\theta} - \frac{N_2}{1-\theta}$$

where we have defined $N_2 = \sum_{n=1}^{N}\gamma(z_{n2})$; which can be interpreted as the effective number of observations assigned to the component 2. Similarly, we can also define $N_1 = \sum_{n=1}^{N}\gamma(z_{n1})$ for the first component

Setting $\frac{d}{d\theta}\mathcal{Q}(\theta;\theta_0) = 0$, we get the result for θ

$$\frac{N_1}{\theta} - \frac{N_2}{1-\theta} = 0 => \theta = \frac{N_1}{N_1+N_2} \text{ (answer)}$$

11.2 Expectation Maximisation

The EM algorithm is a convenient and general purpose iterative approach to maximising the likelihood under missing data/hidden variables[187]. It is generally straightforward to implement and can achieve large jumps in parameter space, particularly in the initial iterations.

Algorithm 11.1 Expectation Maximisation. Compute Maximum Likelihood value for data with hidden variables. Input: a distribution $p(x|\theta)$ and dataset $\mathcal{V}$. Returns ML candidate θ.

1: $t = 0$ ▷ Iteration counter
2: Choose an initial setting for the parameters θ^0. ▷ Initialisation
3: **while** θ not converged (or likelihood not converged) **do**
4: $t \leftarrow t + 1$
5: **for** $n = 1$ to N **do** ▷ Run over all datapoints
6: $q_t^n(h^n|v^n) = p(h^n|v^n, \theta^{t-1})$ ▷ E step
7: **end for**
8: $\theta^t = \arg\max_\theta \sum_{n=1}^{N} \langle \log p(h^n, v^n|\theta) \rangle_{q_t^n(h^n|v^n)}$ ▷ M step
9: **end while**
10: **return** θ^t ▷ The max likelihood parameter estimate.

Q5) Stochastic variational inference

Explain in your own words, using examples and formulas when needed, the following concepts.
A) The difference between variational parameters, model parameters, and prior parameters. (3p)
B) Reparametrization trick. (3p)

A) The difference between variational parameters, model parameters, and prior parameters. (3p)

Variational parameters, model parameters, and prior parameters are all crucial components in probabilistic modeling. However, they serve different purposes and play distinct roles.

Model parameters refer to the parameters that define the statistical model. These are the parameters that we are interested in estimating or inferring, such as the mean or variance of a Gaussian distribution. These parameters are often denoted by Greek letters, such as μ and σ^2, and are typically fixed values that determine the behavior of the model. The values of model parameters can be learned from data using techniques such as maximum likelihood estimation or Bayesian inference. For example, in linear regression, the model parameters are the slope and intercept of the regression line.

Prior parameters, also known as hyperparameters, refer to the parameters of the prior distribution that we place over the model parameters. These parameters represent our prior beliefs about the distribution of the model parameters. Prior parameters are often denoted by Greek letters with a subscript, such as α and β, and are typically set by the modeler before any data is observed. The values of prior parameters can be chosen based on prior knowledge or by using techniques such as cross-validation. For example, in a Gaussian mixture model, the prior parameters might be the mean and variance of the prior distributions for the component means and variances.

Variational parameters are introduced in variational inference, a powerful technique for approximating complex posterior distributions. In variational inference, we posit a family of tractable distributions over the latent variables and introduce variational parameters to optimize the fit between the approximating distribution and the true posterior. These parameters are often denoted by Greek letters with a subscript, such as ϕ and θ, and are typically learned from data using optimization techniques such as gradient descent. For example, in a latent Dirichlet allocation model, the variational parameters might be the probabilities of each topic in each document.

In summary, model parameters, prior parameters, and variational parameters are all important concepts in probabilistic modeling. Model parameters define the statistical model, prior parameters reflect prior beliefs about the model parameters, and variational parameters are introduced to approximate complex posterior distributions in variational inference. Understanding the roles and relationships between these parameters is essential for successful probabilistic modeling.

B) Reparameterization trick. (3p)

The reparametrization trick is a powerful technique used in probabilistic modeling to enable efficient and stable gradient-based optimization. The idea behind the reparametrization trick is to re-express a random variable in terms of a different set of parameters that are deterministic and differentiable, allowing us to compute gradients with respect to those parameters. By doing so, we can take advantage of modern optimization methods such as stochastic gradient descent to efficiently train complex probabilistic models.

The reparametrization trick is often used in the context of variational inference and deep generative models such as variational autoencoders. In these models, we typically want to compute gradients of the expected log-likelihood with respect to the model parameters. However, this expectation is typically intractable to compute analytically, and therefore we need to use techniques such as Monte Carlo integration to estimate it. The reparametrization trick allows us to generate samples from the distribution of interest that are differentiable with respect to the parameters, enabling us to use gradient-based optimization to optimize the model. For example, in a variational autoencoder, we can use the reparametrization trick to generate samples from the approximate posterior distribution, which can then be used to compute the gradients needed to optimize the model parameters.

Overall, the reparametrization trick is a valuable technique that enables efficient and stable optimization of complex probabilistic models. By re-expressing random variables in terms of deterministic and differentiable parameters, we can compute gradients with respect to those parameters and use modern optimization methods to train models that would otherwise be intractable to optimize. The reparametrization trick has become a key tool in deep probabilistic modeling and is used in a wide range of applications, from image and speech recognition to drug discovery and personalized medicine.

Distribution reference

$$N(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \quad \text{(Gaussian)}$$

$$\text{Gamma}(x|a,b) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx}, \quad a>0, b>0, x>0$$

$$\text{Bernoulli}(k|p) = \begin{cases} p, \text{if } k=1 \\ 1-p, \text{ if } k=0 \end{cases}, \quad k \in \{0,1\},\ 0 \le p \le 1.$$

$$\text{Beta}(x|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)+\Gamma(b)} x^{a-1}(1-x)^{b-1}, \quad x \in [0,1], a>0, b>0, \Gamma \text{ is the Gamma function.}$$

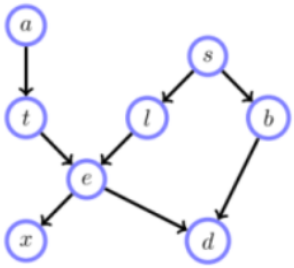
Exam June 2021

Q1) Bayesian networks

A) Are the following conditional independence statements true or false for variables in Fig. 1 (on last page)? Justify your answer by specifying paths between the variables and the blocking variables (if any). (correct answer and justification: 1p per question).

1. tuberculosis $\perp\!\!\!\perp$ smoking | shortness of breath
2. lung cancer $\perp\!\!\!\perp$ bronchitis | smoking
3. visit to Asia $\perp\!\!\!\perp$ smoking | lung cancer, shortness of breath

Fig. 1

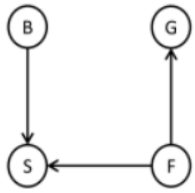


(Barber, Fig. 3.15)

x = Positive X-ray
d = Dyspnea (Shortness of breath)
e = Either Tuberculosis or Lung Cancer
t = Tuberculosis
l = Lung Cancer
b = Bronchitis
a = Visited Asia
s = Smoker

B) Compute the probability $p(F = 0|S = 0)$, where the structure of the model and the required conditional probabilities are specified in Fig. 2. All variables are assumed to have two possible states, 0 and 1.(3p)

Fig. 2



P(B=1)
0.95

P(F=1)
0.8

F	P(G=1|F)
0	0.01
1	0.95

B	F	P(S=1|F,B)
0	0	0.001
1	0	0.01
0	1	0.02
1	1	0.98

Q2) EM algorithm

Consider N i.i.d. observations x_n, $n = 1, \ldots, N$, from a two-component mixture model of exponential distributions

$$p(x_n|\theta, \lambda_1, \lambda_2) = \theta \,\mathrm{Exp}(x_n|\lambda_1) + (1 - \theta)\,\mathrm{Exp}(x_n|\lambda_2)$$

with parameters $(\theta, \lambda_1, \lambda_2)$.
A) Represent the model using latent variables and derive the Q-function of the EM algorithm. (4.5p)
B) Derive the M step update for the λ_1 parameter. (1.5p)

(A) Represent the model using latent variables and derive the Q-function of the EM algorithm. (4.5p)

We formulate the model using latent variables $z_n = (z_1, ..., z_N)$ which explicitly specify the component responsible for generating observation x_n. In detail,

$$z_n = (z_{n1}, z_{n2})^T = \begin{cases} (1,0)^T, & (x_n \text{ is from } Exp(x_n|\lambda_1)) \\ (0,1)^T, & (x_n \text{ is from } Exp(x_n|\lambda_2)) \end{cases}$$

and place a prior on the latent variables

$p(\mathbf{z}|\theta) = \prod_{n=1}^{N} \theta^{z_{n1}}(1-\theta)^{z_{n2}}$

The likelihood in the latent variable model is given by

$p(\mathbf{x}|\mathbf{z}, \lambda_1, \lambda_2) = \prod_{n=1}^{N} Exp(x_n|\lambda_1)^{z_{n1}} Exp(x_n|\lambda_2)^{z_{n2}}$

The joint distribution of all observed ($\mathbf{x}$) and unobserved variables ($\mathbf{z}, \theta, \lambda_1, \lambda_2$) factorizes as follows

$p(\mathbf{x}, \mathbf{z}, \theta, \lambda_1, \lambda_2) = p(\theta)p(\lambda_1)p(\lambda_2)p(\mathbf{z}|\theta)p(\mathbf{x}|\mathbf{z}, \lambda_1, \lambda_2)$

and the log of the joint distribution can correspondingly be written as

$\log p(\mathbf{x}, \mathbf{z}, \theta, \lambda_1, \lambda_2) = \log p(\theta) + \log p(\lambda_1) + \log p(\lambda_2) + \log p(\mathbf{z}|\theta) + \log p(\mathbf{x}|\mathbf{z}, \lambda_1, \lambda_2)$

We approximate the posterior distribution $p(\mathbf{z}, \theta, \lambda_1, \lambda_2|\mathbf{x})$ using the factorized variational distribution $\lambda(\mathbf{z})\lambda(\theta)\lambda(\lambda_1)\lambda(\lambda_2)$

Derive the Q-function of the EM algorithm.

The complete data log-likelihood is the joint log likelihood of both the observed variable, $\mathbf{x}$ and the latent variable, $\mathbf{z}$. In the EM-algorithm we will maximize the expectation of the log-likelihood of the complete data $(\mathbf{x}, \mathbf{z})$. The log-likelihood is:

$\log p(\mathbf{x}, \mathbf{z}|\theta, \lambda_1, \lambda_2) = \log \left\{ \prod_{n=1}^{N} p(x_n, z_n|\theta, \lambda_1, \lambda_2) \right\} = \sum_{n=1}^{N} \log p(x_n, z_n|\theta, \lambda_1, \lambda_2)$

=> $\log p(\mathbf{x}, \mathbf{z}|\theta, \lambda_1, \lambda_2) = \sum_{n=1}^{N} \log \left(p(x_n|z_n, \lambda_1, \lambda_2)p(z_n|\theta) \right)$

=> $\log p(\mathbf{x}, \mathbf{z}|\theta, \lambda_1, \lambda_2) = \sum_{n=1}^{N} \log \left(\theta^{z_{n1}} Exp(x_n|\lambda_1)^{z_{n1}} (1-\theta)^{z_{n2}} Exp(x_n|\lambda_2)^{z_{n2}} \right)$ (from the prior and likelihood above)

=> $\log p(\mathbf{x}, \mathbf{z}|\theta, \lambda_1, \lambda_2) = \sum_{n=1}^{N} z_{n1} \log \left(\theta Exp(x_n|\lambda_1) \right) + z_{n2} \log \left((1-\theta) Exp(x_n|\lambda_2) \right)$

=> $\log p(\mathbf{x}, \mathbf{z}|\theta, \lambda_1, \lambda_2) = \sum_{n=1}^{N} \left(z_{n1} \log \theta + z_{n1} \log Exp(x_n|\lambda_1) + z_{n2} \log(1-\theta) + z_{n2} \log Exp(x_n|\lambda_2) \right)$ (E.q 1)

E-step 1^0

Compute the posterior distribution of the latent variables, given the current estimate θ^0 of θ, λ_1^0 of λ_1 and λ_2^0 of λ_2:

$p(z_{n1} = 1|x_n, \theta^0, \lambda_1^0, \lambda_2^0) \propto \log p(z_{n1} = 1, x_n|\theta, \lambda_1, \lambda_2) = p(z_{n1} = 1)p(x_n|z_n, \theta^0, \lambda_1^0, \lambda_2^0) = \theta^0 Exp(x_n|\lambda_1^0)$ (E.q 2)

$p(z_{n2} = 1|x_n, \theta^0, \lambda_1^0, \lambda_2^0) \propto \log p(z_{n1} = 1, x_n|\theta, \lambda_1, \lambda_2) = p(z_{n2} = 1)p(x_n|z_n, \theta^0, \lambda_1^0, \lambda_2^0) = (1-\theta_0) Exp(x_n|\lambda_2^0)$ (E.q 3)

By normalizing these two equations E.q 2 and E.q 3, we get:

$$\gamma(z_{n1}) = p(z_{n1} = 1|x_n, \theta_0, \lambda_1^0, \lambda_2^0) = \frac{\theta^0 Exp(x_n|\lambda_1^0)}{\theta^0 Exp(x_n|\lambda_1^0) + (1-\theta_0) Exp(x_n|\lambda_2^0)}$$ (E.q 4)

$$\gamma(z_{n2}) = p(z_{n2} = 1|x_n, \theta_0, \lambda_1^0, \lambda_2^0) = \frac{(1-\theta^0) Exp(x_n|\lambda_2^0)}{\theta^0 Exp(x_n|\lambda_1^0) + (1-\theta_0) Exp(x_n|\lambda_2^0)}$$ (E.q 5)

E-step 2^0

The Q-function of the EM algorithm is derived by taking the expectation of the complete-data log-likelihood with respect to the posterior distribution of the latent variables. In this case, the complete-data log-likelihood is given by $\log p(\mathbf{x}, \mathbf{z}, \theta, \lambda_1, \lambda_2)$, and the posterior distribution of the latent variables is $p(\mathbf{z} \mid \mathbf{x}, \theta, \lambda_1, \lambda 2)$. Evaluate the expectation of the complete data log-likelihood over the posterior distribution of the latent variables in E.q 4 and E.q 5

$\mathcal{Q}(\theta, \lambda_1, \lambda_2 | \theta^0, \lambda_1^0, \lambda_2^0) = E_{z|x,\theta^0,\lambda_1^0,\lambda_2^0}[\log p(\mathbf{x}, \mathbf{z} | \theta, \lambda_1, \lambda_2)]$

$=> \mathcal{Q}(\theta, \lambda_1, \lambda_2 | \theta^0, \lambda_1^0, \lambda_2^0) = \sum_{n=1}^{N} \sum_{k=1}^{2} [p(\mathbf{z_{nk}} | \mathbf{x_n}, \theta^0, \lambda_1^0, \lambda_2^0) \log p(\mathbf{x_n}, \mathbf{z_{nk}} | \theta, \lambda_1, \lambda_2)]$

$=> \mathcal{Q}(\theta, \lambda_1, \lambda_2 | \theta^0, \lambda_1^0, \lambda_2^0) =$
$\sum_{n=1}^{N} [p(\mathbf{z_{n1}} | \mathbf{x_n}, \theta^0, \lambda_1^0, \lambda_2^0) \log p(\mathbf{x_n}, \mathbf{z_{n1}} | \theta, \lambda_1, \lambda_2) +$
$p(\mathbf{z_{n2}} | \mathbf{x_n}, \theta^0, \lambda_1^0, \lambda_2^0) \log p(\mathbf{x_n}, \mathbf{z_{n2}} | \theta, \lambda_1, \lambda_2)]$

$=> \mathcal{Q}(\theta, \lambda_1, \lambda_2 | \theta^0, \lambda_1^0, \lambda_2^0) = \sum_{n=1}^{N} \gamma(z_{n1}) \log(\theta Exp(x_n | \lambda_1)) + \gamma(z_{n2}) \log((1-\theta) Exp(x_n | \lambda_2))$

$=> \mathcal{Q}(\theta, \lambda_1, \lambda_2 | \theta^0, \lambda_1^0, \lambda_2^0) = \sum_{n=1}^{N} \gamma(z_{n1}) \log(\theta) + \gamma(z_{n1}) \log Exp(x_n | \lambda_1) + \gamma(z_{n2}) \log(1-\theta) +$
$\gamma(z_{n2}) \log Exp(x_n | \lambda_2)$ (answer)

(B) Derive the M step update for the λ_1 parameter (1.5p)

Maximizing for θ_1:

$\frac{d}{d\theta} \mathcal{Q}(\theta; \theta_0) = \frac{d}{d\theta} \sum_{n=1}^{N} \gamma(z_{n1}) \log(\theta) + \gamma(z_{n1}) \log Exp(x_n | \lambda_1) + \gamma(z_{n2}) \log(1-\theta) + \gamma(z_{n2}) \log Exp(x_n | \lambda_2)$

$\frac{d}{d\theta} \mathcal{Q}(\theta; \theta_0) = \sum_{n=1}^{N} \frac{\gamma(z_{n1})}{\theta} + 0 - \frac{\gamma(z_{n2})}{1-\theta} + 0 = \sum_{n=1}^{N} \frac{\gamma(z_{n1})}{\theta} - \frac{\gamma(z_{n2})}{1-\theta} = \frac{N_1}{\theta} - \frac{N_2}{1-\theta}$

where we have defined $N_2 = \sum_{n=1}^{N} \gamma(z_{n2})$; which can be interpreted as the effective number of observations assigned to the component 2. Similarly, we can also define $N_1 = \sum_{n=1}^{N} \gamma(z_{n1})$ for the first component

Setting $\frac{d}{d\theta} \mathcal{Q}(\theta; \theta_0) = 0$, we get the result for θ

$$\frac{N_1}{\theta} - \frac{N_2}{1-\theta} = 0 => \theta = \frac{N_1}{N_1 + N_2} \text{ (answer)}$$

Maximizing for λ_1:

$$\frac{d}{d\lambda_1}\mathcal{Q}(\lambda_1;\lambda_1^0) = \frac{d}{d\lambda_1}\sum_{n=1}^{N}\gamma(z_{n1})\log(\theta)+\gamma(z_{n1})\log Exp(x_n|\lambda_1)+\gamma(z_{n2})\log(1-\theta)+\gamma(z_{n2})\log Exp(x_n|\lambda_2)$$

First we need to calculate the derivative of the exponential distribution with respect to λ_1:

$$\frac{d}{d\lambda_1}Exp(x_n|\lambda_1) = \frac{d}{d\lambda_1}(\lambda_1\exp(-\lambda_1 x_n)) = \exp(-\lambda_1 x_n)(1-\lambda_1 x_n)$$

$$=> \frac{d}{d\lambda_1}\log Exp(x_n|\lambda_1) = \frac{1}{Exp(x_n|\lambda_1)}\exp(-\lambda_1 x_n)(1-\lambda_1 x_n) = \frac{\exp(-\lambda_1 x_n)(1-\lambda_1 x_n)}{\lambda_1\exp(-\lambda_1 x_n)} = \frac{1}{\lambda_1}-x_n$$

Plugging in the main derivative equation:

$$\frac{d}{d\lambda_1}\mathcal{Q}(\lambda_1;\lambda_1^0) = \sum_{n=1}^{N}0+0+\gamma(z_{n1})(\frac{1}{\lambda_1}-x_n)+0 = \sum_{n=1}^{N}\gamma(z_{n1})(\frac{1}{\lambda_1}-x_n)$$

Setting $\frac{d}{d\lambda_1}\mathcal{Q}(\lambda_1;\lambda_1^0)=0$, we get the result for λ_1

$$\sum_{n=1}^{N}\gamma(z_{n1})(\frac{1}{\lambda_1}-x_n)=0 => \sum_{n=1}^{N}\left[\frac{\gamma(z_{n1})}{\lambda_1}-\gamma(z_{n1})x_n\right]=0 => \lambda_1 = \frac{N_1}{\sum_{n=1}^{N}\gamma(z_{n1})x_n}\ \text{(answer)}$$

Q3) Laplace approximation

Approximate the Gamma$(x|a,b)$ distribution with parameters a and b using the Laplace approximation, i.e., the approximating distribution is a Gaussian centered at the mode of the original distribution. Parameters a and b are known constants. Hint: use $E(x) = -\log Gamma(x|a,b)$ as the starting point. (6p)

Step 1: Derive the gradient $-\nabla\log Gamma(x,a|b)$ and the Hessian $\mathbf{H} = -\nabla\nabla\log Gamma(x,a|b)$ needed for the Laplace approximation.

We have:

$$Gamma(x|a,b) = \frac{b^a}{\Gamma(a)}x^{a-1}\exp(-bx)$$

$$\log\text{Gamma}(x\mid a,b) = a\log b-\log\Gamma(a)+(a-1)\log x-bx$$

$$=> E(x) = -\log\text{Gamma}(x\mid a,b) = -a\log b+\log\Gamma(a)-(a-1)\log x+bx$$

The gradient of the energy for the Laplace approximation is:

$$\nabla E = -\nabla\log Gamma(x|a,b) = \frac{\partial}{\partial x}E(x)$$
$$= \frac{\partial}{\partial x}[-a\log b+\log\Gamma(a)-(a-1)\log x+bx] = \frac{1-a}{x}+b$$

The Hessian of the energy for the Laplace approximation is:

$$\nabla E = -\nabla\nabla\log Gamma(x|a,b) = \frac{\partial^2}{\partial^2 x}E(x)$$
$$= \frac{\partial}{\partial x}\left[\frac{1-a}{x}+x\right] = -\frac{1-a}{x^2} = \frac{a-1}{x^2}$$

Step 2: Find the mode of the Gamma distribution, which is the solution of the equation $\nabla E(x) = 0$

$$\nabla E = 0 => \frac{1-a}{x} + b = 0 => \hat{x} = \frac{a-1}{b}$$

Then, substituting the mode $\hat{x}$ into the Hessian, we get:

$$\nabla\nabla E(\hat{x}) = \frac{a-1}{\hat{x}^2} = \frac{b^2}{a-1}$$

Step 3: Given $\hat{x}$, the Laplace approximation is given by

$$q(x) = \mathcal{N}(x|\mathbf{m}, \mathbf{S}), \qquad \mathbf{S} = \mathbf{H}^{-1}(\boldsymbol{\theta})$$

where the mean $\mathbf{m} = \hat{x}$ is the mode/mean of the approximating Gaussian distribution and the covariance matrix $\mathbf{S}$ is the inverse Hessian of $E(x)$ evaluated at the point $\hat{x}$

Therefore, the Laplace approximation for the Gamma distribution is:

$$q(x) = \mathcal{N}\left(x \mid \frac{a-1}{b}, \frac{a-1}{b^2}\right)$$

Laplace approximation of posterior distribution

- In general, for any posterior $p(\mathbf{w}|\alpha, \mathcal{D})$ it holds that

$$p(\mathbf{w}|\alpha, \mathcal{D}) \propto \exp(-E(\mathbf{w})), \quad E(\mathbf{w}) = -\log p(\mathbf{w}|\alpha, \mathcal{D}).$$

1. Approximate $E(\mathbf{w})$ by a 2nd order Taylor polynomial $\widetilde{E}(\mathbf{w})$ at the minimum $\overline{\mathbf{w}}$

$$\widetilde{E}(\mathbf{w}) = E(\overline{\mathbf{w}}) + \frac{1}{2}(\mathbf{w}-\overline{\mathbf{w}})^T H_{\overline{\mathbf{w}}}(\mathbf{w}-\overline{\mathbf{w}})$$

   (Note, this is quadratic in $\mathbf{w}$.)
2. Obtain a Gaussian approximation $q(\mathbf{w}|\alpha, \mathcal{D})$:

$$p(\mathbf{w}|\alpha, \mathcal{D}) \approx q(\mathbf{w}|\alpha, \mathcal{D}) \propto \exp(-\widetilde{E}(\mathbf{w}))$$

- For logistic regression,

$$E(\mathbf{w}) = \frac{\alpha}{2}\mathbf{w}^T\mathbf{w} - \sum_{i=1}^{N} \log \sigma(\mathbf{w}^T\mathbf{h}_i), \quad \mathbf{h}_i \equiv (2c_i - 1)\mathbf{x}_i.$$

Laplace approximation in practice

- In practice:
  - Find the minimum $\overline{\mathbf{w}}$ of $E(\mathbf{w})$ analytically (root of the derivative) or by numerical optimization, e.g. Newton's method:

$$\mathbf{w}^{new} = \mathbf{w} - \mathbf{H}_w^{-1} \nabla E$$

  - When converged, compute the Hessian $H_{\overline{\mathbf{w}}}$ of $E(\mathbf{w})$ at $\overline{\mathbf{w}}$.
  - The posterior approximation is

$$q(\mathbf{w}|\alpha, \mathcal{D}) = N(\mathbf{w}|\mathbf{m}, \mathbf{S}), \quad \mathbf{m} = \overline{\mathbf{w}}, \quad \mathbf{S} = \mathbf{H}_{\overline{\mathbf{w}}}^{-1}.$$

- Reminder: if $f \equiv f(x_1, \ldots, x_n)$

$$H_f = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \vdots & & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{pmatrix}$$

Q4) Variational Bayes

Suppose you are given data $(y_n, \mathbf{x}_n)$, where $y_n \in \mathbb{R}$ and $x_n \in \mathbb{R}^2$ for all $n = 1, \ldots, N$. We model this using a linear regression model

$$y_n = ax_{n1} + bx_{n2} + \epsilon_n, \quad n = 1, \ldots, N,$$

where

$$\epsilon_n \overset{i.i.d}{\sim} N(0, 1).$$

Prior distributions for the parameters are

$$a \sim N(0, 1), \text{ and}$$
$$b \sim N(0, 1).$$

Assume a variational distribution $q(a, b) = q(a)q(b)$ for the parameters of the model, where the factors are assumed to be of the form

$$q(a) = N(a|\mu_a, \sigma_a^2)$$
$$q(b) = N(b|\mu_b, \sigma_b^2).$$

Derive the variational update for factor $q(a)$. (6p)

Update of factor $q(a)$. Note that this exercise doesnt need the latent variable representation of the model. We can directly use the joint distribution of the model.

Step 1: Write down the log joint distribution of the model based on the priors and likelihoods

$$p(\mathbf{x}, \mathbf{y}, a, b) = p(\mathbf{y}|\mathbf{x}, a, b) \log p(a) \log p(b)$$

$$=> \log p(\mathbf{x}, \mathbf{y}, a, b) = \sum_{n=1}^{N} [\log p(y_n|x_n, a, b)] + \log p(a) + \log p(b)$$

The variance of the error is also the variance of the likelihood for $\mathbf{y}$

Substituting in the expressions for the likelihood and prior distributions, we get:

$$\log p(\mathbf{x}, \mathbf{y}, a, b) = \sum_{n=1}^{N} [\log \mathcal{N}(y_n|ax_{n1} + bx_{n2}, 1)] + \log \mathcal{N}(a|0, 1) + \log \mathcal{N}(b|0, 1)$$

Step 2: To derive the variational update for the factor $q(a)$ q(a), we can use the coordinate ascent variational inference (CAVI) algorithm. This involves optimizing the ELBO/minimizes the KL divergence with respect to one factor at a time while holding the others fixed. This is done by calculating the expectation of the logarithm with respect to the variational distributions excluding the current one in consideration, which is $q(a)$

$\log q^*(a) = E_{q(b)}[p(\mathbf{x}, \mathbf{y}, a, b)]$

=> $\log q^*(a) = E_{q(b)}[\sum_{n=1}^{N} [\log p(y_n|x_n, a, b)] + \log p(a) + \log p(b)]$

We need to keep only the terms dependent on a. The rest terms are constant with respect to this factor can be added to the constant "C"

=> $\log q^*(a) = E_{q(b)}[\log p(a)] + E_{q(b)}[\log p(\mathbf{y}|\mathbf{x}, a, b)] + C$

=> $\log q^*(a) = \log p(a) + E_{q(b)}[\log p(\mathbf{y}|\mathbf{x}, a, b)] + C$ (E.q 1)

According to the exercise, we have $p(a) = \mathcal{N}(0, 1)$ as the prior.

=> $\log p(a) = \log \mathcal{N}(a|0, 1) = -\frac{1}{2} \log 2\pi - \frac{1}{2} \frac{(a-0)^2}{1} = -\frac{1}{2} \log 2\pi - \frac{1}{2} a^2$

We can drop the term that is independent of a

=> $\log p(a) = -\frac{1}{2} a^2 + C$ (E.q 2)

Additionally, we have

$$E_{q(b)}[\log p(\mathbf{y}|\mathbf{x},a,b)] = E_{q(b)}\left[\sum_{n=1}^{N}\left(-\frac{1}{2}\log 2\pi - \frac{1}{2}(y_n - ax_{n1} - bx_{n2})^2\right)\right]$$

Dropping all terms not depending on a, we have:

$$E_{q(b)}[\log p(\mathbf{y}|\mathbf{x},a,b)] = E_{q(b)}\left[\sum_{n=1}^{N}\left(-\frac{1}{2}(-2ax_{n1}y_n + a^2x_{n1}^2 + 2ax_{n1}bx_{n2})\right)\right]$$

$$E_{q(b)}[\log p(\mathbf{y}|\mathbf{x},a,b)] = E_{q(b)}\left[\sum_{n=1}^{N}\left(ax_{n1}y_n - \frac{1}{2}a^2x_{n1}^2 - ax_{n1}bx_{n2}\right)\right]$$

$E_{q(b)}[\log p(\mathbf{y}|\mathbf{x},a,b)] = \sum_{n=1}^{N}\left(ax_{n1}y_n - \frac{1}{2}a^2x_{n1}^2 - ax_{n1}E_{q(b)}\left[b\right]x_{n2}\right)$, where $E_{q(b)}\left[b\right] = 0$ in the prior

=> $E_{q(b)}[\log p(\mathbf{y}|\mathbf{x},a,b)] = \sum_{n=1}^{N}\left(ax_{n1}y_n - \frac{1}{2}a^2x_{n1}^2\right)$ (E.q 3)

Step 3: Plugging (2)(3) into equation (1), we have:

$$\log q^*(a) = -\frac{1}{2}a^2 + \sum_{n=1}^{N}\left(ax_{n1}y_n - \frac{1}{2}a^2x_{n1}^2\right) + C$$

=> $\log q^*(a) = -\frac{1}{2}a^2 + a\sum_{n=1}^{N}x_{n1}y_n - \frac{1}{2}a^2\sum_{n=1}^{N}x_{n1}^2 + C$

=> $\log q^*(a) = -\frac{1}{2}a^2(\sum_{n=1}^{N}x_{n1}^2 + 1) + a\sum_{n=1}^{N}x_{n1}y_n + C$

=> $\log q^*(a) = -\frac{1}{2}a^2(\sum_{n=1}^{N}x_{n1}^2 + 1) + a\sum_{n=1}^{N}x_{n1}y_n + C$

Step 4: Figuring out the closed form solution for the variational update for $q(a)$, if it happens that the prior and the likelihood are conjugate. In this case, both prior and likelihood are Gaussian, so the posterior is also Gaussian.

Completing the square form $-\frac{1}{2}x^TAx + b^Tx$

If $\log q^*(a) \propto -\frac{1}{2}x^TAx + b^Tx$ => $q(a) = \mathcal{N}(a|m,S)$

where $\mathbf{S} = A^{-1}$ and $\mathbf{m} = A^{-1}b$

=> $\log q^*(a) \propto -\frac{1}{2}a^2(\sum_{n=1}^{N}x_{n1}^2 + 1) + a\sum_{n=1}^{N}x_{n1}y_n$

Thus we have the final update for the factor q(a) as:

$q(a) = \mathcal{N}(a|m_a, s_a^2)$

where

$m_a = s_a^2(\sum_{n=1}^{N} x_{n1} y_n)$

and

$s_a^2 = (\sum_{n=1}^{N} x_{n1}^2 + 1)^{-1}$

Q5) Miscellaneous

Briefly (max. 4 sentences each) explain the terms/concepts and their usage/relevance in the context of the course (2p each).

1. Kullback-Leibler divergence
2. ML-II
3. Bayes factor

The Kullback-Leibler (KL) divergence is a measure of the difference between two probability distributions. It is commonly used in machine learning to optimize probabilistic models and to compare the accuracy of different models. It is particularly useful in Bayesian methods for model selection and optimization, and in deep learning for optimizing neural networks.

ML-II (Marginal Likelihood Maximization): In Bayesian inference, the marginal likelihood (also known as evidence) is used to compare different models and select the best one. Maximizing the marginal likelihood involves integrating out the model parameters, which can be difficult or impossible in closed form. ML-II is a technique that maximizes an approximation of the marginal likelihood using iterative optimization methods, such as expectation-maximization (EM) or variational inference (VI). This approach is widely used in model selection and hyperparameter tuning in machine learning, and has been shown to outperform other techniques such as cross-validation in some cases.

The Bayes factor is a measure of the evidence in favor of one hypothesis over another, given some observed data. It is used in Bayesian model selection to compare the relative merits of different models, and to determine which model is more likely to have generated the observed data. The Bayes factor can be used to compare models with different numbers of parameters and can help avoid overfitting.

Distribution reference

$$N(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \quad \text{(Gaussian)}$$

$$\text{Gamma}(x|a,b) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx}, \quad a>0, b>0, x>0 \quad \text{(Gamma)}$$

$$\text{Exp}(x|\lambda) = \lambda e^{-\lambda x}, \quad x \in [0,\infty), \quad \lambda > 0 \quad \text{(Exponential)}$$

Exam 2019

Q1) Bayes' rule

A) Suppose a distribution factorizes according to the graph shown in Fig. 1, and the domains of the variables are as follows: $dom(X) = \{x_1, x_2\}$, $dom(Y) = \{y_1, y_2\}$, $dom(Z) = \{z_1, z_2\}$. In addition, the following conditional probabilities are known:

$$p(x_2) = 0.8, \quad p(z_1) = p(z_2) = 0.5, \quad p(y_2|x_2, z_1) = 0.9, \quad p(y_2|x_2, z_2) = 0.2.$$

Compute $p(z_2|y_2, x_2)$. (3p)

B) Briefly explain *conjugate priors* and their relevance and usage in the context of the course. (3p)

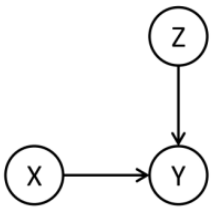


Figure 1

Based on the information provided, we can use Bayes' rule to compute $p(z_2|y_2; x_2)$.

$$p(z_2|y_2, x_2) = \frac{p(x_2, y_2|z_2)p(z_2)}{p(x_2, y_2)} = \frac{p(y_2|x_2, z_2)p(x_2)p(z_2)}{p(x_2|y_2)p(x_2)} = \frac{p(y_2|x_2, z_2)p(z_2)}{p(x_2|y_2)}$$

We know that $p(z_1) = p(z_2) = 0.5$ and $p(y_2|x_2, z_2) = 0.2$

We can also compute p(y2|x2) using the law of total probability:

$$p(y_2|x_2) = p(y_2|x_2, z_1)p(z_1) + p(y_2|x_2, z_2)p(z_2) = 0.9 * 0.5 + 0.2 * 0.5 = 0.55$$

Plugging in the values:

$$p(z_2|y_2, x_2) = \frac{0.2 * 0.5}{0.55} = 0.1818$$

B) Briefly explain *conjugate priors* and their relevance and usage in the context of the course. (3p)

Conjugate priors are probability distributions that have a special relationship with the likelihood function of a given model. Specifically, when the prior distribution is conjugate to the likelihood, the posterior distribution will have the same functional form as the prior distribution. This can greatly simplify the process of updating the distribution as new data is observed.

Conjugate priors are particularly useful in the context of Bayesian machine learning and probabilistic modeling because they allow for efficient and exact Bayesian inference. Rather than computing the posterior distribution directly using Bayes' rule, which may be intractable for complex models, one can simply update the parameters of the conjugate prior distribution using the observed data. This avoids the need for computationally expensive numerical integration or approximation methods.

Moreover, conjugate priors often have interpretable parameters that can be used to encode prior knowledge or beliefs about the model. For example, in Bayesian linear regression, a conjugate prior such as the normal-inverse-gamma distribution can be used to incorporate knowledge about the mean and variance of the regression coefficients.

Overall, conjugate priors provide a powerful and flexible tool for Bayesian inference in machine learning and probabilistic modeling.

Q2) Bayesian networks

A) Are the following statements true or false for the graph in Fig. 2? Justify your answer by specifying the paths between the variables and the blocking variables (if any). (correct answer and justification: 1p per question).

1. C and G are d-separated by $\{B, D\}$.
2. A and C are d-separated by $\oslash$.
3. A and D are d-separated by B.

Definition 20 (d-connection, d-separation)**.** If G is a directed graph in which $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Z}$ are disjoint sets of vertices, then $\mathcal{X}$ and $\mathcal{Y}$ are d-connected by $\mathcal{Z}$ in G if and only if there exists an undirected path U between some vertex in $\mathcal{X}$ and some vertex in $\mathcal{Y}$ such that for every collider C on U, either C or a descendent of C is in $\mathcal{Z}$, and no non-collider on U is in $\mathcal{Z}$.

$\mathcal{X}$ and $\mathcal{Y}$ are d-separated by $\mathcal{Z}$ in G if and only if they are not d-connected by $\mathcal{Z}$ in G.

One may also phrase this as follows. For every variable $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, check every path U between x and y. A path U is said to be *blocked* if there is a node w on U such that either

1. w is a collider and neither w nor any of its descendants is in $\mathcal{Z}$.
2. w is not a collider on U and w is in $\mathcal{Z}$.

If all such paths are blocked then $\mathcal{X}$ and $\mathcal{Y}$ are d-separated by $\mathcal{Z}$.

If the variable sets $\mathcal{X}$ and $\mathcal{Y}$ are d-separated by $\mathcal{Z}$, they are independent conditional on $\mathcal{Z}$ in all probability distributions such a graph can represent.

1. C and G are d-separated by Z={B; D}. True. There are three paths from C to G
   - C, B, E, G and B is not a collider on path and B is in Z => This path is blocked
   - C, D, E, G and G is not a collider on path and G is in Z => This path is blocked
   - C, D, F, E, G with same reason above, plus F is a collider on path and F is not in Z => This path is also blocked. All paths are blocked => C and G are d-separated by Z={B; D}
2. A and C are d-separated by Z={∅}. True. There are three paths from A to C
   - A, B, C and B is a collider on path and B is not in Z => This path is blocked
   - A, B, E, D, C and E is a collider on path and E is not in Z => This path is blocked
   - A, B, E, F, D, C and F is a collider on path and F is not in Z => This path is blocked. All paths are blocked => A and C are d-separated by Z={∅}
3. A and D are d-separated by {B}. False. There are three paths from A to D
   - A, B, C, D and B is a collider on path and B is in Z => This path is not blocked
   - A, B, E, D and E is a collider on path and E is not in Z => This path is blocked
   - A, B, E, F, D and F is a collider on path and F is not in Z => This path is blocked. Exist one path that is not blocked => A and D are not d-separated by {B}

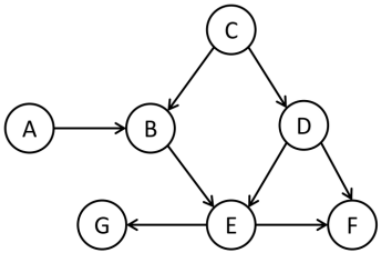


Figure 2

B) How many graphs are there which are Markov equivalent to the graph in Figure 2? Justify your answer. (3p)

Definition: Two graphs are Markov equivalent if they both represent the same set of conditional independence statements. Define the skeleton of a graph by removing the directions on the arrows. Define an immorality in a DAG as a configuration of three nodes, X,Y,Z such that Z is a child of both X and Y, with X and Y not directly connected, so Z is a collider. Two DAGs represent the same set of independence assumptions (they are Markov equivalent) if and only if they have the same skeleton and the same set of immoralities

Having the same skeleton here means their undirected graph versions are similar, or isomorphic. So we can simply reuse this graph above but with different directions of some edges to obtain an equivalent Markov version.

Now, we can obtain all immoralities in this graph by identifying the colliders, and check whether their parents are connected or not. The colliders in the graph are B, E and F

For B, its parents are A and C. They are not connected, so A→B←C is one skeleton

For E, its parents are B and D. They are not connected, so B→E←D is a skeleton

For F, its parents are D and E, but they are connected. So Therefore D→F←E is not a skeleton

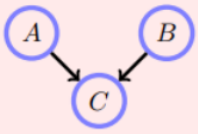
To obtain a Markov equivalent graph, we can simply alter any number of directions of any other edges not contained inside these skeletons. However, we have to pay attention to the fact that by flipping a direction, it must not create a cycle in the graph, because the Bayesian network is DAG. Additionally, flipping that direction does not create a new collider such that its parents are not connected. Under these constraints, In other words, these edges are fixed and cannot be considered for flipping:

A→B, C→B, B→E,D→E. Now the edges left considered for flipping are
E→G, C→D, D→F, E→F

Flipping E→G as G→E => Creates a new collider E where parents G and B are not connected. => This flipping is invalid
Flipping C→D as D→C. Does not create cycle nor collider => This flipping is valid
Flipping D→F as F→D. Creates a cycle D-E-F => This flipping is invalid
Flipping E→F as F→E. Creates a new collider E where parents B and F are not connected => This flipping is invalid

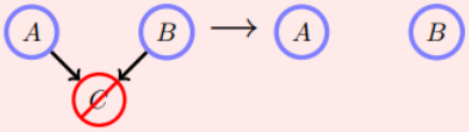
Therefore, there is only 1 Markov equivalent graph to figure 2, which is flipping C→D as D→C (answer)

Definition 21 (Some properties of Belief Networks).

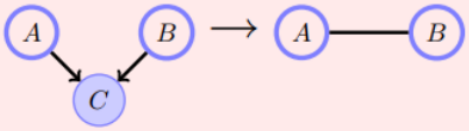


$$p(A, B, C) = p(C|A, B)p(A)p(B) \tag{3.3.26}$$

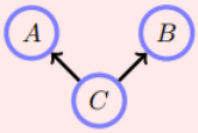
A and B are (unconditionally) independent : $p(A, B) = p(A)p(B)$.
A and B are conditionally dependent on C : $p(A, B|C) \neq p(A|C)p(B|C)$.



Marginalising over C makes A and B independent.

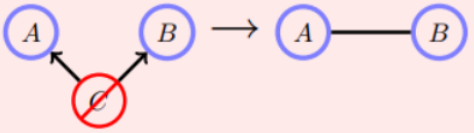


Conditioning on C makes A and B (graphically) dependent.

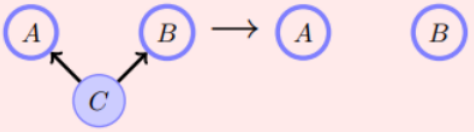


$$p(A, B, C) = p(A|c)p(B|C)p(C) \tag{3.3.27}$$

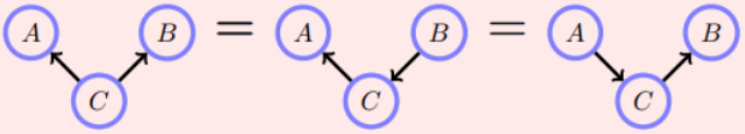
A and B are (unconditionally) dependent : $p(A, B) \neq p(A)p(B)$.
A and B are conditionally independent on C : $p(A, B|C) = p(A|C)p(B|C)$.



Marginalising over C makes A and B (graphically) dependent.



Conditioning on C makes A and B independent.



Definition 22 (Markov Equivalence). Two graphs are Markov equivalent if they both represent the same set of conditional independence statements.

Define the *skeleton* of a graph by removing the directions on the arrows. Define an *immorality* in a DAG as a configuration of three nodes, A,B,C such that C is a child of both A and B, with A and B not directly connected. Two DAGs represent the same set of independence assumptions (they are *Markov equivalent*) if and only if they have the same skeleton and the same set of immoralities [74].

Q3) Variational Bayes

Suppose our data $\mathbf{x} = (x_1, \ldots, x_N)$ consist of N observations drawn independently from the normal distribution

$$x_i \sim N(\mu, \tau^{-1}), \text{ for } i = 1, \ldots, N.$$

We assume the following prior on the parameters

$$p(\mu) = N(\mu|\mu_0, \lambda_0^{-1})$$
$$p(\tau) = \text{Gamma}(\tau|a_0, b_0).$$

Derive the variational update for factor $q(\tau)$, when we assume that the posterior distribution $p(\mu, \tau|\mathbf{x})$ is approximated using a factorized distribution $q(\mu, \tau) = q(\mu)q(\tau)$. You can assume that the current factor for μ is

$$q(\mu) = N(\mu|\mu_*, \sigma_*^2).$$

(6p)

Hint 1: $Var(X) = E(X^2) - E(X)^2$.

Hint 2: The Gamma prior is conjugate here.

Step 1: Write down the log of the joint distribution of all the variables in the model:

$$\log p(\mathbf{x}, \mu, \tau) = \sum_{i=1}^{N} [\log p(x_i|\mu, \tau)] + \log p(\mu) + \log p(\tau).$$

$$\log p(\mathbf{x}, \mu, \tau) = \sum_{i=1}^{N} \left[\log \mathcal{N}(x_i|\mu, \tau^{-1})\right] + \log \mathcal{N}(\mu|\mu_0, \lambda_0^{-1}) + \log \text{Gamma}(\tau|a_0, b_0).$$

We have

$$\log \mathcal{N}(x_i|\mu, \tau^{-1}) = \log \left[(2\pi\tau^{-1})^{-1/2} \exp(-\frac{1}{2}(x_i - \mu)^2(\tau^{-1})^{-1})\right]$$
$$= \log \left[(2\pi)^{-1/2}\tau^{1/2} \exp(-\frac{1}{2}\tau(x_i - \mu)^2)\right]$$
$$= -\frac{1}{2}\log(2\pi) + \frac{1}{2}\log\tau - \frac{1}{2}\tau(x_i - \mu)^2$$

$$\log \mathcal{N}(\mu|\mu_0, \lambda_0^{-1}) = \log \left[(2\pi\lambda_0^{-1})^{-1/2} \exp(-\frac{1}{2}(\mu - \mu_0)^2(\lambda_0^{-1})^{-1}) \right]$$
$$= \log \left[(2\pi)^{-1/2} \lambda_0^{1/2} \exp(-\frac{1}{2}\lambda_0(\mu - \mu_0)^2) \right]$$
$$= -\frac{1}{2}\log(2\pi) + \frac{1}{2}\log\lambda_0 - \frac{1}{2}\lambda_0(\mu - \mu_0)^2$$

$$\log \mathrm{Gamma}(\tau|a_0, b_0) = \log \left[\frac{{b_0}^{a_0}}{\Gamma(a_0)} \tau^{a_0 - 1} \exp(-b_0\tau) \right] = a_0 \log b_0 - \log\Gamma(a_0) + (a_0 - 1)\log\tau - b_0\tau$$

Step 2: Substitute the expressions for the joint distribution:

$$\log p(\mathbf{x}, \mu, \tau) = \sum_{i=1}^{N} [-\frac{1}{2}\log(2\pi) + \frac{1}{2}\log\tau - \frac{1}{2}\tau(x_i - \mu)^2] - \frac{1}{2}\log(2\pi) + \frac{1}{2}\log\lambda_0$$
$$-\frac{1}{2}\lambda_0(\mu - \mu_0)^2 + a_0 \log b_0 - \log\Gamma(a_0) + (a_0 - 1)\log\tau - b_0\tau$$

$$=> \log p(\mathbf{x}, \mu, \tau) = -\frac{N}{2}\log(2\pi) + \frac{N}{2}\log\tau - \frac{\tau}{2}\sum_{i=1}^{N}(x_i - \mu)^2 - \frac{1}{2}\log(2\pi) + \frac{1}{2}\log\lambda_0$$
$$-\frac{1}{2}\lambda_0(\mu - \mu_0)^2 + a_0 \log b_0 - \log\Gamma(a_0) + (a_0 - 1)\log\tau - b_0\tau$$

Step 3: Take the expectation of this expression with respect to $q(\mu)$:

$$\mathbb{E}_{q(\mu)}[\log p(\mathbf{x}, \mu, \tau)] = -\frac{N}{2}\log(2\pi) + \frac{N}{2}\log\tau - \frac{\tau}{2}\left(\sum_{i=1}^{N} x_i^2 - 2\mathbb{E}_{q(\mu)}[\mu]\sum_{i=1}^{N} x_i + N\mathbb{E}_{q(\mu)}[\mu^2]\right)$$
$$-\frac{1}{2}\log(2\pi) + \frac{1}{2}\log\lambda_0 - \frac{1}{2}\lambda_0(\mathbb{E}_{q(\mu)}[\mu^2] - 2\mu_0\mathbb{E}_{q(\mu)}[\mu] + \mu_0^2) + a_0 \log b_0 - \log\Gamma(a_0) + (a_0 - 1)\log\tau - b_0\tau$$

Step 4: Identify the terms that do not depend on τ and add them to the constant C:

$$\mathbb{E}_{q(\mu)}[\log p(\mathbf{x}, \mu, \tau)] = \frac{N}{2}\log\tau - \frac{\tau}{2}(\sum_{i=1}^{N} x_i^2 - 2\mathbb{E}_{q(\mu)}[\mu]\sum_{i=1}^{N} x_i + N\mathbb{E}_{q(\mu)}[\mu^2]) + (a_0 - 1)\log\tau - b_0\tau + C$$

According to Hint 1: $Var(X) = E(X^2) - E(X)^2$ => $\mathbb{E}_{q(\mu)}[\mu^2] = \mathbb{E}_{q(\mu)}[\mu]^2 + Var(\mu) = \mu_*^2 + \sigma_*^2$

Additionally, $\mathbb{E}_{q(\mu)}[\mu] = \mu_*$ by definition

$$\mathbb{E}_{q(\mu)}[\log p(\mathbf{x}, \mu, \tau)] = \left(\frac{N}{2} + a_0 - 1\right)\log\tau - \tau\left(b_0 + \frac{1}{2}\left(\sum_{i=1}^{N} x_i^2 - 2\mu_*\sum_{i=1}^{N} x_i + N(\mu_*^2 + \sigma_*^2)\right)\right) + C$$

Use these terms to derive an update for $\log q^*(\tau)$:

$$\log q^*(\tau) \propto \left(\frac{N}{2} + a_0 - 1\right) \log \tau - \tau \left(b_0 + \frac{1}{2}\left(\sum_{i=1}^{N} x_i^2 - 2\mu_* \sum_{i=1}^{N} x_i + N(\mu_*^2 + \sigma_*^2)\right)\right)$$

$$=> q^*(\tau) \propto \exp\left(\left(\frac{N}{2} + a_0 - 1\right) \log \tau - \tau \left(b_0 + \frac{1}{2}\left(\sum_{i=1}^{N} x_i^2 - 2\mu_* \sum_{i=1}^{N} x_i + N(\mu_*^2 + \sigma_*^2)\right)\right)\right)$$

Let the terms defined as:

$$a_N = a_0 + \frac{N}{2}$$

and

$$b_N = b_0 + \frac{1}{2}\left(\sum_{i=1}^{N} x_i^2 - 2\mu_* \sum_{i=1}^{N} x_i + N(\mu_*^2 + \sigma_*^2)\right)$$

$=> q^*(\tau) \propto \exp\left((a_N - 1)\log \tau - b_N \tau\right) = \exp\left(\log[\tau^{(a_N - 1)} \exp(-b_N \tau)]\right) = \tau^{a_N - 1} \exp(-b_N \tau)$

Recognize that this resembles Gamma distribution:

$$q^*(\tau) \propto \text{Gamma}(\tau | a_N, b_N) = \frac{{b_N}^{a_N}}{\Gamma(a_N)} \tau^{a_N - 1} \exp(-b_N \tau)$$

according to Hint 2, the prior is conjugate to the likelihood, so the posterior update of $q^*(\tau)$ is also a Gamma distribution

where

$$a_N = a_0 + \frac{N}{2}$$

and

$$b_N = b_0 + \frac{1}{2}\left(\sum_{i=1}^{N} x_i^2 - 2\mu_* \sum_{i=1}^{N} x_i + N(\mu_*^2 + \sigma_*^2)\right)$$

Q4) Black-box variational inference

Assume that N observations x_n, $n = 1, \ldots, N$ have been generated from the model in Fig. 3 with some conditional distributions $p(\lambda_1)$, $p(\lambda_2)$, $p(\lambda_3|\lambda_2)$, $p(z_n|\lambda_1)$, $p(x_n|z_n, \lambda_3)$. Assume that the variational approximation is

$$q(\lambda_1, \lambda_2, \lambda_3, z_1, \ldots, z_n) = q(\lambda_1|\theta_1)q(\lambda_2|\theta_2)q(\lambda_3|\theta_3)\prod_{n=1}^{N} q(z_n|\eta_n),$$

where $\theta_1, \theta_2, \theta_3, \eta_1, \ldots, \eta_N$ are variational parameters.

A) Write and simplify the formula to calculate the ELBO for the model in Figure 3. (2p)

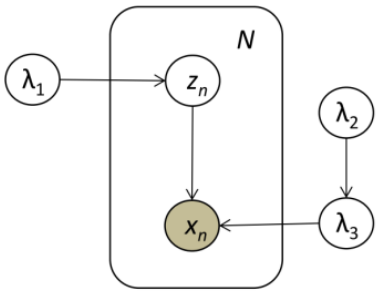


Figure 3

The derivation of the VB algorithm was based on minimizing $KL(q||p)$ in

$$\log p(x) = \mathcal{L}(q) + KL(q||p)$$

When conjugate priors and exponential family distributions are used, we can compute the variational lower bound $\mathcal{L}(q)$ directly, which is the general formula for the ELBO:

$$\mathcal{L}(q) = \int q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})} d\mathbf{Z} = E_q[\log p(\mathbf{X}, \mathbf{Z})] - E_q[\log q(\mathbf{Z})]$$

where $\mathbf{Z}$ is a generic notation that includes all unobservables. On the other hand, $\mathbf{X}$ is a notation for the observables.

In this case, $\mathbf{X} = \mathbf{x}$ and $\mathbf{Z} = \{\mathbf{z}, \lambda_1, \lambda_2, \lambda_3\}$

The joint distribution $p(\mathbf{x}, \mathbf{z}, \lambda_1, \lambda_2, \lambda_3)$ is defined from Figure (3) as:

$$p(\mathbf{x}, \mathbf{z}, \lambda_1, \lambda_2, \lambda_3) = p(\lambda_1)p(\lambda_2)p(\lambda_3|\lambda_2) \prod_{n=1}^{N} p(z_n|\lambda_1)p(x_n|z_n, \lambda_3)$$

$$=> \log p(\mathbf{x}, \mathbf{z}, \lambda_1, \lambda_2, \lambda_3) = \log p(\lambda_1) + \log p(\lambda_2) + \log p(\lambda_3|\lambda_2) + \sum_{n=1}^{N} [\log p(z_n|\lambda_1) + \log p(x_n|z_n, \lambda_3)]$$

The joint distribution of the variational approximation $q(\mathbf{z}, \lambda_1, \lambda_2, \lambda_3)$ is defined as:

$$q(\mathbf{z}, \lambda_1, \lambda_2, \lambda_3) = q(\lambda_1|\theta_1)q(\lambda_2|\theta_2)q(\lambda_3|\theta_3) \prod_{n=1}^{N} q(z_n|\eta_n)$$

$$=> \log q(\mathbf{z}, \lambda_1, \lambda_2, \lambda_3) = \log q(\lambda_1|\theta_1) + \log q(\lambda_2|\theta_2) + \log q(\lambda_3|\theta_3) + \sum_{n=1}^{N} \log q(z_n|\eta_n)$$

The ELBO in this exercise is given as:

$\mathcal{L}(q) = E_q[\log p(\mathbf{x}, \mathbf{z}, \lambda_1, \lambda_2, \lambda_3)] - E_q[\log q(\mathbf{z}, \lambda_1, \lambda_2, \lambda_3)]$

=> $\mathcal{L}(q) = E_q[\log p(\lambda_1) + \log p(\lambda_2) + \log p(\lambda_3|\lambda_2) + \sum_{n=1}^{N} [\log p(z_n|\lambda_1) + \log p(x_n|z_n, \lambda_3)]] - E_q[\log q(\lambda_1|\theta_1) + \log q(\lambda_2|\theta_2) + \log q(\lambda_3|\theta_3) + \sum_{n=1}^{N} \log q(z_n|\eta_n)]$

=> $\mathcal{L}(q) = E_{q(\lambda_1)}[\log p(\lambda_1)] + E_{q(\lambda_2)}[\log p(\lambda_2)] + E_{q(\lambda_1)q(\lambda_2)}[\log p(\lambda_3|\lambda_2)] + \sum_{n=1}^{N} \left(E_{q(\lambda_1)q(z_n)}[\log p(z_n|\lambda_1)] + E_{q(x_n)q(z_n)q(\lambda_2)}[\log p(x_n|z_n, \lambda_3)]\right) - E_{q(\lambda_1)q(\theta_1)}[\log q(\lambda_1|\theta_1)] - E_{q(\lambda_2)q(\theta_2)}[\log q(\lambda_2|\theta_2)] - E_{q(\lambda_3)q(\theta_3)}[\log q(\lambda_3|\theta_3)] - \sum_{n=1}^{N} E_{q(z_n)q(\eta_n)}[\log q(z_n|\eta_n)]$

Variational lower bound (ELBO)

- The derivation of the VB algorithm was based on minimizing $KL(q||p)$ in

$$\log p(\mathbf{x}) = \mathcal{L}(q) + KL(q||p)$$

- When conjugate priors and exponential family distributions are used, we can compute the variational lower bound $\mathcal{L}(q)$ directly

$$\mathcal{L}(q) = \int q(\mathbf{z}) \log \left\{ \frac{p(\mathbf{x}, \mathbf{z})}{q(\mathbf{z})} \right\} d\mathbf{z}$$

- Computing $\mathcal{L}(q)$ gives:
  1. alternative way to define the factor updates by maximizing $\mathcal{L}(q)$.
  2. simple check of the VB algorithm - $\mathcal{L}(q)$ should never decrease.
  3. criterion to monitor convergence.
  4. an estimate of $\log p(x)$ to be used in **model selection**

Simple example: computing the ELBO

- The model:

$$p(x_n|\theta,\tau) = (1-\tau)N(x_n|0,1) + \tau N(x_n|\theta,1), \;\; n = 1,\ldots,N.$$

Prior:

$$\tau \sim Beta(\alpha_0,\alpha_0) \qquad \theta \sim N(0,\beta_0^{-1})$$

- After factorizing $\log p(\mathbf{x},\mathbf{z},\tau,\theta)$, ELBO can be written as:

$$\begin{aligned}\mathcal{L}(q) = {} & E_{q(\tau)}[\log p(\tau)] + E_{q(\theta)}[\log p(\theta)] + E_{q(\mathbf{z})q(\tau)}[\log p(\mathbf{z}|\tau)] \\ & + E_{q(\mathbf{z})q(\theta)}[\log p(\mathbf{x}|\mathbf{z},\theta)] - E_{q(\mathbf{z})}[\log q(\mathbf{z})] - E_{q(\tau)}[\log q(\tau)] \\ & - E_{q(\theta)}[\log q(\theta)].\end{aligned}$$

B) Explain the idea of black-box variational inference, and compare it with the 'standard' variational inference. What are the advantages and disadvantages? (2p)

Black-box variational inference (BBVI) is a type of variational inference (VI) algorithm that does not require the explicit derivation of the gradient of the log evidence lower bound (ELBO) with respect to the variational parameters. Instead, BBVI treats the probabilistic model as a black box, and uses Monte Carlo methods to estimate the gradient of the ELBO with respect to the variational parameters. This allows BBVI to be applied to a wider range of probabilistic models, including models with intractable likelihoods or models with complex dependencies between variables.

In contrast, standard variational inference requires the derivation of the gradient of the ELBO with respect to the variational parameters, which can be difficult or impossible for complex models. Standard VI also assumes that the variational distribution belongs to a certain family of distributions, such as the mean-field family, which can limit its flexibility.

The advantage of BBVI is that it can be applied to a wider range of models and can provide more flexible variational approximations. Additionally, BBVI can be more computationally efficient than standard VI for certain models. However, BBVI can be less stable than standard VI and may require more tuning of hyperparameters. It can also be more difficult to diagnose convergence issues with BBVI.

Black-box variational inference is a form of variational inference that allows for quick application to many models with little additional derivation. It is based on a stochastic optimization of the variational objective, where the noisy gradient is computed from Monte Carlo samples from the

variational distribution1. This method can be used with non-standard parameterizations for distributions that are easier to reason about.

In contrast, standard variational inference typically requires significant model-specific analysis to derive a variational inference algorithm. This can hinder and deter practitioners from quickly developing and exploring a variety of models for a problem at hand.

The advantage of black-box variational inference is that it allows for faster exploration of a wide space of models. It has been shown to reach better predictive likelihoods much faster than sampling methods. However, one disadvantage is that it may not always provide as accurate an approximation as model-specific variational inference algorithms.

In summary, black-box variational inference provides a trade-off between speed and accuracy, allowing for quick exploration of many models at the cost of potentially less accurate approximations.

C) Using generic notation, the gradient of the ELBO can be written as:

$$\nabla_\lambda L = E_{q(z|\lambda)}[\nabla_\lambda \log q(z|\lambda)(\log p(x,z) - \log q(z|\lambda))]. \quad (1)$$

Write and simplify the following terms in Equation (1): i) $\log p(x,z)$, ii) $\log q(z|\lambda)$, $\nabla_\lambda \log q(z|\lambda)$ for the model specified in Figure 3. (2p)

Q5) EM algorithm

Consider a simple factor analysis model:

$$\mathbf{x}_n \sim N_2(\mathbf{w}z_n, \sigma^2 I), \quad n = 1, \ldots, N,$$
$$z_n \sim N(0,1), \quad n = 1, \ldots, N,$$

where $\mathbf{x}_n \in R^2$ and $z_n \in R$ for all $n = 1, \ldots, N$. Parameters of the model are the loading matrix (a vector in this case), $\mathbf{w} \in R^2$, and variance $\sigma^2 \in R$.

A) Derive and simplify the complete data log-likelihood. (2p)

B) It can be shown that the posterior distribution $p(z_n|\mathbf{x}_n, \mathbf{w}_0, \sigma_0^2)$, where $\mathbf{w}_0, \sigma_0^2$ are current estimates of the parameters, is a Gaussian $N(z_n|\mu_n, \sigma_z^2)$ with certain μ_n and σ_z^2. Derive formulas for μ_n and σ_z^2. (2p)

C) Derive the Q function needed in the E step of the EM algorithm. (2p)
Hint 1: You can solve C even if you did not solve B, i.e., the solution to C can be given using μ_n and σ_z^2.
Hint 2: Completing the square.

Q5: EM

X_n

Complete data log likelihood:

$$\log(p(X, z \mid w, \sigma)) = \sum^{n} \log p(x_n, z_n \mid w, \sigma) = \sum^{n} \log p(x_n \mid z_n, w, \theta) \cdot p(z_n)$$

$$= \sum^{n} [\log p(x_n \mid z_n, w, \theta) + \log p(z_n)]$$

$$= \sum^{n} \log \mathcal{N}_2(x_n \mid w z_n, \sigma^2 I) + \log \mathcal{N}(0, 1)$$

$$\propto \sum^{n} \log \frac{1}{\sigma} - \frac{1}{2\sigma^2}(x_n - w z_n)^T (x_n - w z_n) + -\frac{z_n^2}{2}$$

dont need to did n't drop other step

b)

$$p(z_n \mid x_n, w_0, \sigma_0^2) \propto p(z_n)\, p(x_n \mid z_n, w_0, \sigma_0^2)$$

$$= -\frac{1}{2\sigma^2}\left(x_n^T x_n - 2 x_n^T w z_n + z_n w^T w z_n - \frac{z_n^2}{2}\right)$$

$$= -\frac{1}{2}\left(z_n^2 w^T w \frac{1}{\sigma^2} - 2\left[x_n^T w \sigma^{-2}\right] z_n + \frac{z_n^2}{2}\right)$$

$$= -\frac{1}{2}\left(z_n^2 \left(w^T w \frac{1}{\sigma^2} + \frac{1}{2}\right) - \frac{2\left[x_n^T w \sigma^2\right] z_n \left(w^T w \sigma^2 + 1\right)}{w^T w \sigma^2 + 1}\right)$$

$\frac{1}{\sigma^2}$ ↑ nhầm dấu rồi

$$= -\frac{1}{2}\left(w^T w \sigma^2 + 1\right)\left(z_n - \frac{x_n^T w \sigma^2}{w^T w \sigma^2 + 1}\right)^2$$

$$\approx \mathcal{N}\left(z \mid \frac{x_n^T w \sigma^2}{w^T w \sigma^2 + 1}, \left(w^T w \sigma^2 + 1\right)^{-1}\right)$$

$\frac{x_n^T w \sigma^2}{w^T w \sigma^2 + 1} = \mu_n$; $\left(w^T w \sigma^2 + 1\right)^{-1} = \sigma_2$

c) $Q(w, \sigma^2 \mid w_0, \sigma_0^2) = \mathbb{E}_{z \mid x, w_0, \sigma_0^2}\left[\log - \frac{1}{2\sigma^2}(x - w_0 z_n)^T (x_n - w_0 z_n) - \frac{z_n^2}{2}\right]$

expand and replace z_n with μ_z, z_n^2 with $\mu^2 + \sigma_z^2$

Distribution reference

$$N(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \quad \text{(Gaussian)}$$

$$N_k(x|\mu,\Sigma) = (2\pi)^{-\frac{k}{2}} |\Sigma|^{-\frac{1}{2}} e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)} \quad \text{(Multivariate Gaussian)}$$

$$N_k(x|\mu,\sigma^2 I) = (2\pi)^{-\frac{k}{2}}\sigma^{-k}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^T(x-\mu)\right\} \quad \text{(MVN with diagonal covariance)}$$

$$\text{Gamma}(x|a,b) = \frac{b^a}{\Gamma(a)}x^{a-1}e^{-bx}, \quad a>0, b>0, x>0$$

Exam 2018

1) Bayesian networks

A) Are the conditional independence statements below always 'true' for a Bayesian network with structure shown in Figure 1? Justify your answer by specifying paths between the variables and the blocking variables (if any). (correct answer and justification: 1.5p per question).

1. $x_2 \perp\!\!\!\perp x_6 \mid x_5, x_1$
2. $x_2 \perp\!\!\!\perp x_6 \mid x_5, x_3, x_7$

B) In this question you must model a problem with 4 binary variables: G ('gray'), V ('Vancouver'), R ('rain') and S ('sad'). Consider a Bayesian network for these variables with structure and conditional distributions as shown in Figure 2. Write down an expression for $P(S=1|V=1)$ in terms of $\alpha, \beta, \gamma, \delta$. (3p).

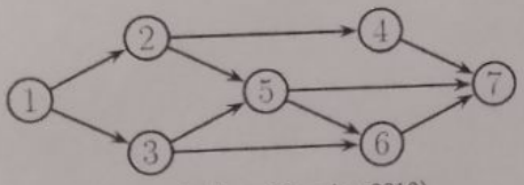


Figure 1 (from Murphy, 2012)

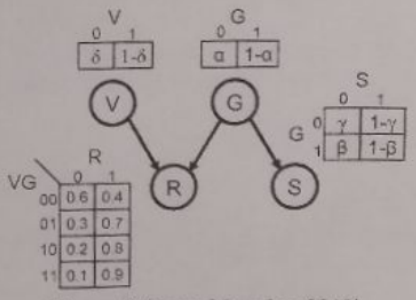


Figure 2 (from Murphy, 2012)

2) EM algorithm

Consider N observations x_n, $n = 1, \ldots, N$, assumed to be i.i.d. from a mixture of Poisson distributions:

$$p(x_n|\pi,\lambda) = \sum_{k=1}^{K}\pi_k \text{Poisson}(x_n|\lambda_k).$$

Represent the model using latent variables and derive the E step of the expectation maximization algorithm, which could be used to learn the maximum likelihood estimates for the parameters $\pi = (\pi_1, \ldots, \pi_K)^T$ and $\lambda = (\lambda_1, \ldots, \lambda_K)^T$. (6p)

3) Laplace approximation

Approximate the Beta distribution with parameters a and b, $\text{Beta}(x|a,b)$, using the Laplace approximation, i.e., the approximating distribution is a Gaussian centered at the mode of the original distribution. Parameters a and b are known constants, and you can assume that $a > 1$, and $b > 1$, such that the Beta distribution has a mode in the interval $(0, 1)$. Hint: use $E(x) = -\log Beta(x|a,b)$ as the starting point. (6p)

Step 1: Derive the gradient $-\nabla \log Beta(x, a|b)$ and the Hessian $\mathbf{H} = -\nabla\nabla \log Beta(x, a|b)$ needed for the Laplace approximation.

We have:

$$Beta(x|a,b) = \frac{x^{a-1}(1-x)^{b-1}}{B(a,b)} = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}$$

$$\log \text{Beta}(x \mid a,b) = \log\Gamma(a+b) - \log\Gamma(a) - \log\Gamma(b) - (a-1)\log x + (b-1)\log(1-x)$$

$$=> E(x) = -\log \text{Beta}(x \mid a,b) = -\log\Gamma(a+b) + \log\Gamma(a) + \log\Gamma(b) - (a-1)\log x - (b-1)\log(1-x)$$

The gradient of the energy for the Laplace approximation is:

$$\nabla E = -\nabla \log Beta(x|a,b) = \frac{\partial}{\partial x} E(x)$$
$$= \frac{\partial}{\partial x}[-\log\Gamma(a+b) + \log\Gamma(a) + \log\Gamma(b) - (a-1)\log x - (b-1)\log(1-x)] =$$
$$-\frac{a-1}{x} + \frac{b-1}{1-x} = \frac{1-a}{x} + \frac{1-b}{x-1}$$

The Hessian of the energy for the Laplace approximation is:

$$\nabla E = -\nabla\nabla \log Beta(x|a,b) = \frac{\partial^2}{\partial^2 x} E(x)$$
$$= \frac{\partial}{\partial x}\left[\frac{1-a}{x} + \frac{1-b}{x-1}\right] = \frac{\partial}{\partial x}\left[(1-a)\frac{1}{x} + (1-b)\frac{1}{x-1}\right]$$
$$= (1-a)(-\frac{1}{x^2}) + (1-b)(-\frac{1}{(x-1)^2}) = \frac{a-1}{x^2} + \frac{b-1}{(x-1)^2}$$

Step 2: Find the mode of the Beta distribution, which is the solution of the equation $\nabla E(x) = 0$

$$\nabla E = 0 => \frac{1-a}{x} + \frac{1-b}{x-1} = 0 => \hat{x} = \frac{a-1}{a+b-2}$$

Then, substituting the mode $\hat{x}$ into the Hessian, we get:

$$\nabla\nabla E(\hat{x}) = \frac{a-1}{\hat{x}^2} + \frac{b-1}{(\hat{x}-1)^2} = \frac{a+b-2}{a-1} + \frac{a+b-2}{b-1} = \frac{(a+b-2)^2}{(a-1)(b-1)}$$

Step 3: Given $\hat{x}$, the Laplace approximation is given by

$$q(x) = \mathcal{N}(x|\mathbf{m}, \mathbf{S}), \qquad \mathbf{S} = \mathbf{H}^{-1}(\boldsymbol{\theta})$$

where the mean $\mathbf{m} = \hat{x}$ is the mode/mean of the approximating Gaussian distribution and the covariance matrix $\mathbf{S}$ is the inverse Hessian of $E(x)$ evaluated at the point $\hat{x}$

Therefore, the Laplace approximation for the Beta distribution is:

$$q(x) = \mathcal{N}\left(x \mid \frac{a-1}{a+b-2}, \frac{(a-1)(b-1)}{(a+b-2)^2}\right)$$

4) Variational Bayes

Suppose you are given data $(y_n, \mathbf{x}_n)$, where $y_n \in \mathbb{R}$ and $x_n \in \mathbb{R}^2$ for all $n = 1, \ldots, N$. We model this using a linear regression model

$$y_n = ax_{n1} + bx_{n2} + \epsilon_n, \quad n = 1, \ldots, N,$$

where

$$\epsilon_n \overset{i.i.d}{\sim} N(0, 1).$$

Prior distributions for the parameters are

$$a \sim N(0, 1), \text{ and}$$
$$b \sim N(0, 1).$$

Assume a variational distribution $q(a, b) = q(a)q(b)$ for the parameters of the model, where the factors are assumed to be of the form

$$q(a) = N(a\ \mu_a, \sigma_a^2)$$
$$q(b) = N(b|\mu_b, \sigma_b^2).$$

Derive the variational update for factor $q(a)$. (6p)

<Same as Q4 June 2021>

5) Edward

A) Write Edward code for Model and Inference descriptions for the regression model used in Question 4. (3p)
B) Briefly explain the idea of black-box variational inference and how it differs from the 'traditional' variational inference. What are the strengths and weaknesses of the two approaches? (3p)

<Does not need to solve this exercise>

Distribution reference

$$N(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \quad \text{(Gaussian)}$$

$$N_k(x|\mu,\Sigma) = (2\pi)^{-\frac{k}{2}} |\Sigma|^{-\frac{1}{2}} e^{-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)} \quad \text{(Multivariate Gaussian)}$$

$$\text{Uniform}(x|a,b) = \begin{cases} 1/(b-a), & \text{if } x \in [a,b] \\ 0, & \text{otherwise} \end{cases}$$

$$\text{Exp}(x|\lambda) = \lambda e^{-\lambda x}, \quad x \in [0,\infty), \quad \lambda > 0 \quad \text{(Exponential)}$$

$$\text{Gamma}(x|a,b) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx}, \quad a > 0, b > 0, x > 0$$

$$\text{Poisson}(x|\lambda) = e^{-\lambda}\frac{\lambda^x}{x!}, \quad x = 0,1,2,\ldots, \quad \lambda > 0$$

$$\text{Beta}(x|\alpha,\beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}, \quad x \in (0,1), \alpha > 0, \beta > 0, B(\alpha,\beta) \text{ is the 'beta function'.}$$

Exam 2015

1) Belief (Bayesian) networks

A) True or false? Justify your answer briefly, in max 2 sentences. (correct answer and justification: 1p per question).

1. C and E are d-separated by $\{A, F\}$ in Fig. 1.
2. A and G are d-separated by $\{C, E\}$ in Fig. 1.
3. The Markov equivalence class to which the graph in Fig. 2 belongs has three members.

B) Answer the questions (1.5p each)

1. Factorize the probability distribution $p_{A,B,C,D}(a,b,c,d)$ according to the graph in Fig. 2.
2. Write down the formula to compute $p_{C|A,B,D}(c|a,b,d)$ for the distribution represented by the graph in Fig. 2.

Fig. 1

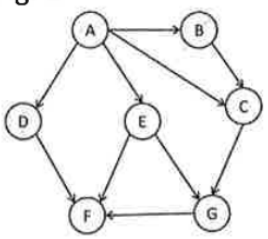
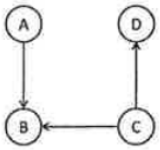


Fig. 2



2) EM algorithm

Consider N observations x_n, $n = 1, \ldots, N$, from a two-component mixture model of exponential distributions

$$p(x_n|\theta, \lambda_1, \lambda_2) = \theta \operatorname{Exp}(x_n|\lambda_1) + (1 - \theta) \operatorname{Exp}(x_n|\lambda_2).$$

Represent the model using latent variables and derive the E and M steps of the expectation maximization algorithm to learn the maximum likelihood estimates of the parameters $(\theta, \lambda_1, \lambda_2)$.

<Solved in 2021 June Q2)

3) Variational approximation

A) Compute the Kullback–Leibler divergence $KL(q, p)$ between $q(x)$ =Uniform$(x|a, b)$ and $p(x) = N(x|0, 1)$.
B) Approximate the Gaussian $N(0, 1)$ distribution using a variational approximation with approximating distribution $q(x)$ =Uniform$(x|a, b)$. The distribution q has two parameters, a and b, which you have to optimize.

(3) Variational approximation

(A) Compute the Kullback—Leibler divergence $KL(q, p)$ between $q(x) = \mathbf{Uniform}(x|a, b)$ and $p(x) = N(x|0, 1)$.

The Kullback-Leibler divergence between two probability distributions $p(x)$ and $q(x)$ is defined as:

$$KL(q||p) = \int q(x) \log \frac{q(x)}{p(x)} dx$$

In this case, $q(x) = \mathbf{Uniform}(x \mid a, b)$ and $p(x) = \mathcal{N}(x \mid 0, 1)$, so we have:

$$KL(q||p) = \int_a^b \frac{1}{b-a} \log \left(\frac{1}{b-a} \left(\frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2}) \right)^{-1} \right) dx$$

$$KL(q||p) = \log(b-a) + \log(\sqrt{2\pi}) + \frac{1}{b-a} \int_a^b \frac{x^2}{2} dx$$

$$KL(q||p) = \log(b-a) + \log(\sqrt{2\pi}) + \frac{1}{2(b-a)} [\frac{x^3}{3}]_a^b$$

$$KL(q||p) = \log(b-a) + \log(\sqrt{2\pi}) + \frac{b^3 - a^3}{6(b-a)}$$

$$KL(q||p) = \log(b-a) + \log(\sqrt{2\pi}) + \frac{(b-a)(b^2 + ab + a^2)}{6(b-a)}$$

$$KL(q||p) = \log(b-a) + \log(\sqrt{2\pi}) + \frac{b^2 + ab + a^2}{6}$$ (answer)

4) Gibbs sampling

Consider the factor analysis model

$$\begin{aligned}
\mathbf{x}_n &\sim \mathcal{N}_D\left(\mathbf{W}\mathbf{z}_n, \mathrm{diag}(\boldsymbol{\psi})^{-1}\right), \quad n = 1, \ldots, N \\
\psi_d &\sim \mathrm{Gamma}(a, b), \quad d = 1, \ldots, D \\
\mathbf{W}_k &\sim \mathcal{N}_D(\mathbf{0}, \alpha\mathbf{I}), \quad k = 1, \ldots, K \\
\mathbf{z}_n &\sim \mathcal{N}_K(\mathbf{0}, \mathbf{I}), \quad n = 1, \ldots, N,
\end{aligned}$$

where $\mathbf{W}_k$ denotes the loadings for the kth factor and ψ_d^{-1} is the specific noise variance of the dth observed variable. Furthermore let D denote the number of observed variables (i.e. $\mathbf{x}_n \in R^D$), N the number of data points, and K the number of factors in the model. $\mathrm{diag}(\boldsymbol{\psi})$ is a diagonal matrix with elements $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_D)^T$ on the diagonal.

A) Write down pseudo-code for the Gibbs sampler to generate samples from the posterior distribution $p(\boldsymbol{\psi}, \mathbf{W}, \mathbf{z}|\mathbf{x})$, where we have denoted $\mathbf{z} = (z_1, \ldots, z_n)$ and $\mathbf{x} = (x_1, \ldots, x_n)$. The exact forms of the distributions are *not* required here.

B) Derive the conditional distribution $p(\psi_d|\boldsymbol{\psi}_{-d}, \mathbf{W}, \mathbf{z}, \mathbf{x})$ required in the Gibbs sampler. Here $\boldsymbol{\psi}_{-d}$ denotes vector $\boldsymbol{\psi}$ from which the dth element has been removed. *Hints*: start by writing the likelihood proportionally s.t. all terms not dependent on ψ_d have been discarded. Note that a multivariate Gaussian with a diagonal covariance matrix can be expressed as a product of univariate Gaussian distributions.

<No need to solve this exercise>

5) Miscellaneous

Briefly (max. 4 sentences each) explain the terms/concepts and their usage/relevance in the context of the course (1.5p each).

1. proposal distribution
2. rotation invariance in the factor analysis models
3. marginal likelihood
4. Markov chain Monte Carlo (MCMC)

Distribution reference

$$\begin{aligned}
N(x|\mu, \sigma^2) &= \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \quad \text{(Gaussian)} \\
N_k(x|\mu, \Sigma) &= (2\pi)^{-\frac{k}{2}} |\Sigma|^{-\frac{1}{2}} e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)} \quad \text{(Multivariate Gaussian)} \\
\mathrm{Uniform}(x|a, b) &= \begin{cases} 1/(b-a), & \text{if } x \in [a, b] \\ 0, & \text{otherwise} \end{cases} \quad \text{(Uniform)} \\
\mathrm{Exp}(x|\lambda) &= \lambda e^{-\lambda x}, \quad x \in [0, \infty), \quad \lambda > 0 \quad \text{(Exponential)} \\
\mathrm{Gamma}(x|a, b) &= \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx}, \quad a > 0, b > 0, x > 0 \quad \text{(Gamma)}
\end{aligned}$$