Assignment 1E, 2025

DD2434 Machine Learning, Advanced Course

Bruno Carchia   Riccardo Alfonso Cerrone

carchia@kth.se   cerrone@kth.se

November 23, 2025

Group # 60

1.1   Dependencies in a Directed Graphical Model

1.1.1 Yes.

1.1.2 No.

1.1.3 No.

1.1.4 No.

1.1.5 Yes.

1.1.6 No.

1.2   Inference in Bayesian Networks

1.2.7   Fish Classification

Suppose the fish was caught on December 20, with

$$\mathbb{P}(X_1) = (0.5, 0, 0, 0.5)$$

corresponding to the four seasons (winter, spring, summer, autumn). The lightness has not been measured, but it is known that the fish is thin.

We want to compute the posterior distributions

$$\mathbb{P}(X_2 \mid X_4 = \text{thin})$$

in order to classify the fish as salmon or sea bass.

Using the Bayes rule, we have

$$\mathbb{P}(X_2 \mid X_4 = \text{thin}) = \frac{\mathbb{P}(X_4 = \text{thin} \mid X_2) \cdot \mathbb{P}(X_2)}{\mathbb{P}(X_4 = \text{thin})}$$

Where:

- $\mathbb{P}(X_4 = \text{thin} \mid X_2)$ is given by the corresponding CPT

- $\mathbb{P}(X_2) = \sum_{x_1} \mathbb{P}(X_2, X_1 = x_1) = \sum_{x_1} \mathbb{P}(X_2 \mid X_1 = x_1) \cdot \mathbb{P}(X_1 = x_1)$ can be computed using the given CPT and the prior on X_1

- $\mathbb{P}(X_4 = \text{thin})$ is a normalizing constant and for this reason we can ignore it in the computation.

At the end, we have:

$$\mathbb{P}(X_2 \mid X_4 = \text{thin}) \propto \mathbb{P}(X_4 = \text{thin} \mid X_2) \cdot \sum_{x_1} \mathbb{P}(X_2 \mid X_1 = x_1) \cdot \mathbb{P}(X_1 = x_1)$$

The second step is to compute the above expression for both values of X_2 (salmon and sea bass) and then compare the results to classify the fish.

$$\mathbb{P}(X_2 = \text{salmon} \mid X_4 = \text{thin}) \propto 0.65 \cdot (0.5 \cdot 0.88 + 0 \cdot 0.32 + 0 \cdot 0.42 + 0.5 \cdot 0.78)$$
$$\propto 0.65 \cdot (0.44 + 0 + 0 + 0.39)$$
$$\propto 0.65 \cdot 0.83$$
$$\propto 0.5395$$

$$\mathbb{P}(X_2 = \text{sea bass} \mid X_4 = \text{thin}) \propto 0.06 \cdot (0.5 \cdot 0.12 + 0 \cdot 0.68 + 0 \cdot 0.58 + 0.5 \cdot 0.22)$$
$$\propto 0.06 \cdot (0.06 + 0 + 0 + 0.11)$$
$$\propto 0.65 \cdot 0.17$$
$$\propto 0.0102$$

For the classification we compare the two results. As said before, we can ignore the normalizing constant because it is the same for both, so it can be seen as a good evaluation score.

$$0.5395 > 0.0102$$

so we classify the fish as **salmon**.

1.2.8   Season Inference

Suppose the fish is thin $X_4 = t$ and of medium lightness $X_3 = m$, we want to understand which season is more likely to have caught the fish X_1.

In other words we want to compute the following posterior distribution

$$\mathbb{P}(X_1 = x_1 \mid X_3 = m, X_4 = t)$$

Using the Bayes rule we have

$$\mathbb{P}(X_1 = x_1 \mid X_3 = m, X_4 = t) = \frac{\mathbb{P}(X_1 = x_1, X_3 = m, X_4 = t)}{\mathbb{P}(X_3 = m, X_4 = t)}$$

where the denominator can be seen as a normalizing constant that for our purposes we can ignore. Consequently, we have the following

$$\mathbb{P}(X_1 = x_1 \mid X_3 = \text{medium}, X_4 = \text{thin}) \propto \mathbb{P}(X_1 = x_1, X_3 = m, X_4 = t)$$

In order to apply the data given by the problem, we apply the chain rule of probability to the right hand side

$$\mathbb{P}(X_1 = x_1, X_3 = m, X_4 = t) = \mathbb{P}(X_1 = x_1) \cdot \mathbb{P}(X_3 = m \mid X_1 = x_1) \cdot \mathbb{P}(X_4 = t \mid X_3 = m, X_1 = x_1)$$

Now we can compute each term:

Term 1 - $\mathbb{P}(X_1 = x_1)$:   is given by the prior.

Term 2 - $\mathbb{P}(X_3 = m \mid X_1 = x_1)$:   it can be computed as follows knowing that for X_3 , X_2 is enough and X_1 is independent of X_3 given X_2

$$\begin{aligned}
\mathbb{P}(X_3 = m \mid X_1 = x_1) &= \sum_{x_2} \mathbb{P}(X_3 = m, X_2 = x_2 \mid X_1 = x_1) \\
&= \sum_{x_2} \mathbb{P}(X_3 = m \mid X_1 = x_1, X_2 = x_2) \, \mathbb{P}(X_2 = x_2 \mid X_1 = x_1) \\
&= \sum_{x_2} \mathbb{P}(X_3 = m \mid X_2 = x_2) \, \mathbb{P}(X_2 = x_2 \mid X_1 = x_1)
\end{aligned}$$

Disclaimer: the derivation of Term 2 is not strictly necessary for the computations that follow.

Term 3 - $\mathbb{P}(X_4 = \mathbf{t} \mid X_3 = \mathbf{m}, X_1 = x_1)$: it needs to be computed as follows knowing that for X_4, X_2 is enough and X_1/X_3 is independent from X_4 given X_2:

$$\mathbb{P}(X_4 = \mathrm{t} \mid X_3 = \mathrm{m}, X_1 = x_1) = \sum_{x_2} \mathbb{P}(X_4 = \mathrm{t}, X_2 = x_2 \mid X_3 = \mathrm{m}, X_1 = x_1)$$

$$= \sum_{x_2} \mathbb{P}(X_4 = \mathrm{t} \mid X_2 = x_2, X_3 = \mathrm{m}, X_1 = x_1) \, \mathbb{P}(X_2 = x_2 \mid X_3 = \mathrm{m}, X_1 = x_1)$$

$$= \sum_{x_2} \mathbb{P}(X_4 = \mathrm{t} \mid X_2 = x_2) \, \mathbb{P}(X_2 = x_2 \mid X_3 = \mathrm{m}, X_1 = x_1)$$

$$= \sum_{x_2} \mathbb{P}(X_4 = \mathrm{t} \mid X_2 = x_2) \, \frac{\mathbb{P}(X_3 = \mathrm{m} \mid X_2 = x_2) \, \mathbb{P}(X_2 = x_2 \mid X_1 = x_1)}{\mathbb{P}(X_3 = \mathrm{m} \mid X_1 = x_1)}$$

We can build the final expression for the approximation of $\mathbb{P}(X_1 = x_1 \mid X_3 = \mathrm{m}, X_4 = \mathrm{t})$ substituting the three terms in the equation where $\mathbb{P}(X_3 = \mathrm{m} \mid X_1 = x_1)$ will cancel out:

$$\mathbb{P}(X_1 = x_1) \cdot \sum_{x_2} \mathbb{P}(X_3 = \mathrm{m} \mid X_2 = x_2) \cdot \mathbb{P}(X_2 = x_2 \mid X_1 = x_1) \cdot \mathbb{P}(X_4 = \mathrm{t} \mid X_2 = x_2)$$

At this point we can compute the above expression for each value of X_1 (the four seasons) and then compare the results to find the most likely season.

- For $X_1 = $ winter:

$$0.25 \cdot (0.34 \cdot 0.88 \cdot 0.65 + 0.12 \cdot 0.12 \cdot 0.06)$$
$$= 0.048836$$

- For $X_1 = $ spring:

$$0.25 \cdot (0.34 \cdot 0.32 \cdot 0.65 + 0.12 \cdot 0.68 \cdot 0.06)$$
$$= 0.018904$$

- For $X_1 = $ summer:

$$0.25 \cdot (0.34 \cdot 0.42 \cdot 0.65 + 0.12 \cdot 0.58 \cdot 0.06)$$
$$= 0.024249$$

- For $X_1 = $ autumn:

$$0.25 \cdot (0.34 \cdot 0.78 \cdot 0.65 + 0.12 \cdot 0.22 \cdot 0.06)$$
$$= 0.043491$$

Comparing the results we have that the most likely season is **winter**.

1.3   CAVI

1.3.11   Exact posterior derivation

We know that the model is defined in this assignment is defined as follows

$$X_n \mid \mu, \tau \sim \mathcal{N}\left(\mu, \frac{1}{\tau}\right), \qquad (\mu, \tau) \sim \mathcal{N}\Gamma(\mu_0, \lambda_0, \alpha_0, \beta_0)$$

Using the Bayes rule we can write the posterior as:

$$\mathbb{P}(\theta \mid D) = \frac{\mathbb{P}(D \mid \theta) \cdot \mathbb{P}(\theta)}{\mathbb{P}(D)} \propto \mathbb{P}(D \mid \theta) \cdot \mathbb{P}(\theta)$$

where:

- $P(\theta) = P(\mu, \tau)$ is the prior distribution

  Knowing that:

$$\mathbb{P}(\tau) \sim \mathrm{Gamma}(\alpha_0, \beta_0) \qquad \mathbb{P}(\mu \mid \tau) \sim \mathcal{N}\left(\mu_0, \frac{1}{\lambda_0 \tau}\right)$$

$$\mathbb{P}(\mu, \tau) = \mathbb{P}(\tau) \cdot \mathbb{P}(\mu \mid \tau) = \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} \sqrt{\frac{\lambda_0 \tau}{2\pi}} \tau^{\alpha_0 - 1} e^{-\beta_0 \tau} e^{-\frac{\lambda_0 \tau}{2}(\mu - \mu_0)^2}$$

- $\mathbb{P}(D \mid \theta)$ is the likelihood of the data

$$\mathbb{P}(D \mid \theta) = \prod_{n=1}^{N} f_\theta(x_n) = \prod_{n=1}^{N} \sqrt{\frac{\tau}{2\pi}} e^{-\frac{\tau}{2}(x_n - \mu)^2}$$

We can now write the log-posterior ignoring the terms that are constants with respect to μ and τ

$$\log P(\mu, \tau \mid X) = \log\left(\prod_{n=1}^{N} P(x_n \mid \tau) P(\mu, \tau)\right) = \sum_{n=1}^{N} \log P(x_n \mid \mu, \tau) + \log P(\mu, \tau)$$

$$\propto \sum_{n=1}^{N} \left[\frac{1}{2}\log\tau - \frac{1}{2}\log(2\pi) - \frac{\tau}{2}(x_n - \mu)^2\right] + \left[\frac{1}{2}\log\left(\frac{\lambda_0\tau}{2\pi}\right) + (\alpha_0 - 1)\log\tau - \beta_0\tau - \frac{\lambda_0\tau}{2}(\mu - \mu_0)^2\right]$$

$$\propto \frac{N}{2}\log\tau - \frac{\tau}{2}\sum_{n=1}^{N}(x_n - \mu)^2 + \left(\alpha_0 - \frac{1}{2}\right)\log\tau - \beta_0\tau - \frac{\lambda_0\tau}{2}(\mu - \mu_0)^2$$

$$\propto \frac{N}{2}\log\tau - \frac{\tau}{2}\sum_{n=1}^{N}(x_n^2 - 2x_n\mu + \mu^2) + \left(\alpha_0 - \frac{1}{2}\right)\log\tau - \beta_0\tau - \frac{\lambda_0\tau}{2}(\mu^2 - 2\mu\mu_0 + \mu_0^2)$$

$$\propto \left(\frac{N}{2} + \alpha_0 - \frac{1}{2}\right)\log\tau - \beta_0\tau - \frac{\tau}{2}\left[\mu^2(N + \lambda_0) - 2\mu\left(\sum_{n=1}^{N}x_n + \lambda_0\mu_0\right) + \sum_{n=1}^{N}x_n^2 + \lambda_0\mu_0^2\right]$$

$$\propto \left(\frac{N}{2} + \alpha_0 - \frac{1}{2}\right) \log \tau - \beta_0 \tau - \frac{\tau}{2}(N+\lambda_0)\left[\mu^2 - \frac{2\mu}{N+\lambda_0}\left(\sum_{n=1}^{N} x_n + \lambda_0 \mu_0\right) + \frac{1}{N+\lambda_0}\left(\sum_{n=1}^{N} {x_n}^2 + \lambda_0 \mu_0^2\right)\right]$$

We can define C as:

$$C = \frac{1}{N+\lambda_0}\sum_{n=1}^{N}\left({x_n}^2 + \lambda_0 \mu_0^2\right)$$

Continuing from the previous expression we have:

$$\propto \left(\frac{N}{2} + \alpha_0 - \frac{1}{2}\right) \log \tau - \beta_0 \tau - \frac{\tau}{2}(N+\lambda_0)[\mu^2 - \frac{2\mu}{N+\lambda_0}(\sum_{n=1}^{N} x_n + \lambda_0 \mu_0) + C]$$

$$\propto \left(\frac{N}{2} + \alpha_0 - \frac{1}{2}\right) \log \tau - \tau\left(\beta_0 + \frac{C}{2}(N+\lambda_0)\right) - \frac{\tau}{2}(N+\lambda_0)\left[\mu^2 - \frac{2\mu}{N+\lambda_0}\left(\sum_{n=1}^{N} x_n + \lambda_0 \mu_0\right)\right]$$

We can define D as:

$$D = \frac{2\mu}{N+\lambda_0}\left(\sum_{n=1}^{N} x_n + \lambda_0 \mu_0\right)$$

And in order to complete the square we can add and subtract $\frac{D^2\tau(N+\lambda_0)}{2}$

$$\propto \left(\frac{N}{2} + \alpha_0 - \frac{1}{2}\right) \log \tau - \tau\left(\beta_0 + \frac{C}{2}(N+\lambda_0)\right) - \frac{\tau}{2}(N+\lambda_0)\left[\mu^2 - \frac{2\mu}{N+\lambda_0}D\right] + \frac{D^2\tau(N+\lambda_0)}{2} - \frac{D^2\tau(N+\lambda_0)}{2}$$

$$= \left(\frac{N}{2} + \alpha_0 - \frac{1}{2}\right) \log \tau - \tau\left(\beta_0 + \frac{C}{2}(N+\lambda_0) - \frac{N+\lambda_0}{2}D^2\right) - \frac{\tau}{2}(N+\lambda_0)[\mu^2 - 2\mu D + D^2]$$

$$= \left(\frac{N}{2} + \alpha_0 - \frac{1}{2}\right) \log \tau - \tau\left(\beta_0 + \frac{C}{2}(N+\lambda_0) - \frac{N+\lambda_0}{2}D^2\right) - \frac{\tau}{2}(N+\lambda_0)[\mu - D]^2$$

We can rewrite it as:

$$\log P(\mu, \tau \mid X) \propto \left(\alpha_0^* - \frac{1}{2}\right) \log \tau - \beta_0^* \tau - \frac{\tau \lambda_0^*}{2}(\mu - \mu_0^*)^2$$

We recognize the kernel of a Normal-Gamma distribution, so the posterior is distributed as:

$$P(\mu, \tau \mid X) \sim \mathcal{N}\Gamma(\mu_0^*, \lambda_0^*, \alpha_0^*, \beta_0^*)$$

with updated parameters:

$$\alpha_0^* = \alpha_0 + \frac{N}{2}$$

$$\lambda_0^* = N + \lambda_0$$

$$\mu_0^* = D = \frac{\lambda_0 \mu_0 + \sum x_n}{\lambda_0 + N}$$

$$\beta_0^* = \beta_0 + \frac{N+\lambda_0}{2}(C - D^2) = \beta_0 + \frac{1}{2}\left[\sum x_n^2 + \lambda_0 \mu_0^2 - \lambda_0^* \mu_0^{*2}\right]$$

1.3.12 VI algorithm implementation

Derivation of CAVI updates

$$\log q^*(\tau) = \mathbb{E}_{q(\mu)}\big[\log p(\tau, X \mid \mu)\big]$$

$$= \mathbb{E}_{q(\mu)}\Big[(a_0 - \tfrac{1}{2})\log\tau - \beta_0\tau - \frac{\lambda_0}{2}(\mu - \mu_0)^2 + \frac{N}{2}\log\tau - \frac{\tau}{2}\sum_{i=1}^{N}(x_i - \mu)^2\Big]$$

$$= \big(a_0 - \tfrac{1}{2} + \tfrac{N}{2}\big)\log\tau \;-\; \tau\Big(\beta_0 + \frac{\lambda_0}{2}\mathbb{E}_{q(\mu)}[(\mu - \mu_0)^2] + \frac{1}{2}\mathbb{E}_{q(\mu)}\Big[\sum_{i=1}^{N}(x_i - \mu)^2\Big]\Big)$$

$$= \log\tau\,\big(a_0 + \tfrac{N}{2} - \tfrac{1}{2}\big) - \tau\Big(\beta_0 + \frac{\lambda_0}{2}\mathbb{E}_{q(\mu)}[(\mu - \mu_0)^2] + \frac{1}{2}\mathbb{E}_{q(\mu)}\sum_{i=1}^{N}(x_i - \mu)^2\Big)$$

$$q^*(\tau) \propto \tau^{a_0 + \frac{N}{2} - \frac{1}{2} - 1}\,\exp\Bigg(-\tau\Big[\beta_0 + \frac{\lambda_0}{2}\mathbb{E}_{q(\mu)}[(\mu - \mu_0)^2] + \frac{1}{2}\mathbb{E}_{q(\mu)}\sum_{i=1}^{N}(x_i - \mu)^2\Big]\Bigg)$$

$$q(\tau) \sim \mathrm{Gam}(\alpha_N, \beta_N) \tag{1}$$

$$\alpha_N = \alpha_0 + \frac{N+1}{2} \tag{2}$$

$$\beta_N = b_0 + \frac{\mathbb{E}_{q(\mu)}}{2}\Big[\lambda_0(\mu - \mu_0)^2 + \sum_{n=1}^{N}(x_i - \mu)^2\Big] \tag{3}$$

$$\log q^*(\mu) = \mathbb{E}_{q(\tau)}\big[\log p(\mu, X \mid \tau)\big]$$

$$= \mathbb{E}_{q(\tau)}\Big[-\frac{\lambda_0\tau}{2}(\mu - \mu_0)^2 - \frac{\tau}{2}\sum_{i=1}^{N}(x_i - \mu)^2\Big]$$

$$= -\frac{\mathbb{E}_{q(\tau)}[\tau]}{2}\Big[\lambda_0(\mu - \mu_0)^2 + \sum_{i=1}^{N}(x_i - \mu)^2\Big].$$

$$q(\mu) \sim \mathcal{N}(\mu_N, \lambda_N) \tag{4}$$

$$\mu_N = \frac{\lambda_0\mu_0 + \sum x_n}{\lambda_0 + N} \tag{5}$$

$$\lambda_N = \mathbb{E}_{q(\tau)}[\tau](\lambda_0 + N) \tag{6}$$

Derivation of the ELBO

We now derive the Evidence Lower Bound by expanding each term in $\mathbb{E}_q[\log p(D, \mu, \tau)]$ and $\mathbb{E}_q[\log q(\mu, \tau)]$. All expectations will be computed using the variational distributions $q(\mu)$ in (4) and $q(\tau)$ in (1). In particular, note that:

$$\mathbb{E}_q[\mu] = \mu_N \quad \text{from (5)}, \qquad \mathbb{E}_q[\mu^2] = \frac{1}{\lambda_N} + \mu_N^2 \quad \text{from (4)},$$

and, analogously for the Gamma distribution in (1),

$$\mathbb{E}_q[\tau] = \frac{\alpha_N}{\beta_N}, \qquad \mathbb{E}_q[\log \tau] = \psi(\alpha_N) - \log \beta_N.$$

We start from the ELBO definition:

$$\begin{aligned}
\mathcal{L} &= \mathbb{E}_q\left[\log \frac{\mathbb{P}(D, \mu, \tau)}{q(\mu, \tau)}\right] \\
&= \mathbb{E}_q[\log \mathbb{P}(D, \mu, \tau)] - \mathbb{E}_q[\log q(\mu, \tau)] \\
&= \textcolor{red}{\mathbb{E}_q[\log \mathbb{P}(\mu \mid \lambda_0, \mu_0, \tau)]} + \textcolor{blue}{\mathbb{E}_q[\log \mathbb{P}(\tau \mid \alpha_0, b_0)]} + \textcolor{orange}{\mathbb{E}_q[\log \mathbb{P}(D \mid \mu, \tau)]} \\
&\quad - \textcolor{magenta}{\mathbb{E}_q[\log q(\mu)]} - \textcolor{magenta}{\mathbb{E}_q[\log q(\tau)]}
\end{aligned}$$

The joint log-density separates into prior terms for μ and τ plus the likelihood term for the dataset D.

The following sections compute each coloured term separately.

Term 1: $\mathbb{E}_q[\log p(\mu \mid \lambda_0, \mu_0, \tau)]$ We use the Normal prior over μ, whose log-density depends on $\log \tau$ and on the quadratic term $(\mu - \mu_0)^2$. Substituting the expressions and applying expectations using the moments of $q(\mu)$ from (4), we obtain:

$$\begin{aligned}
\textcolor{red}{\mathbb{E}_q[\log \mathbb{P}(\mu \mid \lambda_0, \mu_0, \tau)]} &= \mathbb{E}_q\left[\frac{1}{2}\log \tau + \frac{1}{2}\log \frac{\lambda_0}{2\pi} - \frac{1}{2}\lambda_0 \tau (\mu - \mu_0)^2\right] \\
&= \frac{1}{2}\log \frac{\lambda_0}{2\pi} + \frac{1}{2}\mathbb{E}_q[\log \tau] - \frac{1}{2}\lambda_0 \mathbb{E}_q[\tau]\left(\mathbb{E}_q[\mu^2] - 2\mu_0 \mathbb{E}_q[\mu] + \mu_0^2\right)
\end{aligned}$$

Term 2: $\mathbb{E}_q[\log p(\tau \mid \alpha_0, b_0)]$ Using the Gamma prior, whose log-density is linear in $\log \tau$ and τ, and substituting expectations using the parameters from (2)-(3):

$$\begin{aligned}
\textcolor{blue}{\mathbb{E}_q[\log \mathbb{P}(\tau \mid \alpha_0, b_0)]} &= \mathbb{E}_q\left[\alpha_0 \log b_0 - \log \Gamma(\alpha_0) + (\alpha_0 - 1)\log \tau - b_0 \tau\right] \\
&= \alpha_0 \log b_0 - \log \Gamma(\alpha_0) + (\alpha_0 - 1)\mathbb{E}_q[\log \tau] - b_0 \mathbb{E}_q[\tau]
\end{aligned}$$

Term 3: $\mathbb{E}_q[\log p(D \mid \mu, \tau)]$ The likelihood factorizes over the observations and contains terms depending on $(x_n - \mu)^2$. Using $\mathbb{E}_q[(x_n - \mu)^2]$ computed from (4), we obtain:

$$
\begin{aligned}
\mathbb{E}_q[\log \mathbb{P}(D \mid \mu, \tau)] &= \mathbb{E}_q\left[\log \left(\prod_{n=1}^{N} \mathbb{P}(x_n \mid \mu, \tau) \right) \right] = \mathbb{E}_q\left[\sum_{n=1}^{N} \log \mathbb{P}(x_n \mid \mu, \tau) \right] \\
&= \sum_{n=1}^{N} \mathbb{E}_q\left[\log \left(\sqrt{\frac{\tau}{2\pi}}\, e^{-\frac{\tau}{2}(x_n-\mu)^2} \right) \right] \\
&= \sum_{n=1}^{N} \mathbb{E}_q\left[\frac{1}{2}\log \tau - \frac{1}{2}\log(2\pi) - \frac{\tau}{2}(x_n-\mu)^2 \right] \\
&= \frac{N}{2}\mathbb{E}_q[\log \tau] - \frac{N}{2}\log(2\pi) - \frac{1}{2}\mathbb{E}_q[\tau]\sum_{n=1}^{N} \mathbb{E}_q\left[(x_n-\mu)^2\right] \\
&= \frac{N}{2}\mathbb{E}_q[\log \tau] - \frac{N}{2}\log(2\pi) - \frac{1}{2}\mathbb{E}_q[\tau]\left(\sum_{n=1}^{N} x_n^2 - 2\mathbb{E}_q[\mu]\sum_{n=1}^{N} x_n + N\mathbb{E}_q[\mu^2] \right)
\end{aligned}
$$

Term 4: $\mathbb{E}_q[\log q(\mu)]$ The entropy term of the Gaussian $q(\mu)$ in (4) yields:

$$
\begin{aligned}
\mathbb{E}_q[\log q(\mu)] &= \mathbb{E}_q\left[\log \left(\sqrt{\frac{\lambda_N}{2\pi}}\, e^{-\frac{\lambda_N}{2}(\mu-\mu_N)^2} \right) \right] \\
&= \mathbb{E}_q\left[\frac{1}{2}\log \frac{\lambda_N}{2\pi} - \frac{\lambda_N}{2}(\mu - \mu_N)^2 \right] \\
&= \frac{1}{2}\log \frac{\lambda_N}{2\pi} - \frac{\lambda_N}{2}\mathbb{E}_q\left[(\mu - \mu_N)^2\right] \\
&= \frac{1}{2}\log \frac{\lambda_N}{2\pi} - \frac{\lambda_N}{2}\cdot\frac{1}{\lambda_N} \\
&= \frac{1}{2}\left(\log \frac{\lambda_N}{2\pi} - 1 \right)
\end{aligned}
$$

Term 5: $\mathbb{E}_q[\log q(\tau)]$ Finally, for the Gamma variational distribution in (1):

$$
\begin{aligned}
\mathbb{E}_q[\log q(\tau)] &= \mathbb{E}_q\left[\alpha_N \log \beta_N - \log \Gamma(\alpha_N) + (\alpha_N - 1)\log \tau - \beta_N \tau \right] \\
&= \alpha_N \log \beta_N - \log \Gamma(\alpha_N) + (\alpha_N - 1)\mathbb{E}_q[\log \tau] - \beta_N \mathbb{E}_q[\tau]
\end{aligned}
$$

This completes the derivation of the ELBO, expressed entirely in terms of the variational parameters μ_N, λ_N, α_N, β_N, which correspond directly to the CAVI updates given in (4)-(3), and μ_0, λ_0, α_0, b_0.

Conclusions