Advanced Probabilistic Methods

Marshal Sinaga - marshal.sinaga@aalto.fi

Last update: March 25, 2024

This note aims to cover a few examples of probabilistic machine-learning methods. The primary references are Bayesian Reasoning and Machine Learning by David Barber and CS-E4820 by Pekka Martinen. Ideally, this note will be updated regularly until April 16, 2024.

1 Variational Bayes for Simple Model

Suppose we have N independent observations $\mathbf{x} = (x_1, \cdots, x_N)$ from a two-component mixture of univariate Gaussian distributions.

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

that is with probability $1-\tau$ the observation x_n is generated from the first component $N(x_n|0,1)$, and with probability τ from the second component $N(x_n|\theta,1)$. The model 1 has two unknown parameters (τ, θ), the mixture coefficient and the mean of the second component.

The goal is to carry out a full Bayesian analysis via mean-field variational Bayesian approximation. We place the following priors on the unknown parameters.

$$\begin{aligned}\tau &\sim Beta(\alpha_0, \alpha_0)\\ \theta &\sim N(0, \beta_0^{-1})\end{aligned}$$

We formulate the model using latent variables $\mathbf{z} = (z_1, \cdots, z_N)$, which explicitly specify the component responsible for generating observation x_n. In detail,

$$z_n = (z_{n1}, z_{n2})^\top = \begin{cases} (1,0)^\top & x_n \text{ is from } N(x_n|0,1) \\ (0,1)^\top & x_n \text{ is from } N(x_n|\theta,1) \end{cases}$$

and place a prior on the latent variables

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

The likelihood in the latent variable model is given by

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

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

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

and the log joint distribution can correspondingly written as

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

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

Update factor $q(\mathbf{z})$ To compute the updated distribution $q^*(\mathbf{z})$, we first compute the expectation of the log of the joint distribution over all other unknowns in the model.

$$\begin{aligned}
&\log q^*(\mathbf{z}) = \mathbb{E}_{\tau,\theta}[\log p(\mathbf{x}, \mathbf{z}, \tau, \theta)] \\
&= \mathbb{E}_{\tau}[\log p(\mathbf{z}|\tau) + \mathbb{E}_{\theta}[\log p(\mathbf{x}|\mathbf{z}, \theta)]] + \text{ const} \\
&= \mathbb{E}_{\tau}[\sum_{n=1}^{N} z_{n2} \log \tau + z_{n1} \log(1-\tau)] + \mathbb{E}_{\theta}[\sum_{n=1}^{N} z_{n1} \log N(x_n|0,1) + z_{n2} \log N(x_n|\theta,1)] + \text{const} \\
&= \sum_{n=1}^{N} z_{n2}\mathbb{E}_{\tau}[\log \tau] + z_{n1}\mathbb{E}_{\tau}[\log(1-\tau)] + \sum_{n=1}^{N} z_{n1} \log N(x_n|0,1) + z_{n2}\mathbb{E}_{\theta}[\log N(x_n|\theta,1)] + \text{const} \\
&= \sum_{n=1}^{N} z_{n1} \left(\mathbb{E}_{\tau}[\log(1-\tau)] - \frac{1}{2}\log 2\pi - \frac{1}{2}x_n^2 \right) + \sum_{n=1}^{N} z_{n2} \left(\mathbb{E}_{\tau}[\log \tau] - \frac{1}{2}\log 2\pi - \frac{1}{2}\mathbb{E}_{\theta}[(x_n-\theta)^2] \right) + \text{const} \\
&= \sum_{n=1}^{N} z_{n1} \log \rho_{n1} + z_{n2} \log \rho_{n2} + \text{const} \qquad (2)
\end{aligned}$$

Where we have defined ρ_{n1} and ρ_{n2} for all n as follows

$$\log \rho_{n1} = \mathbb{E}_{\tau}[\log(1-\tau)] - \frac{1}{2}\log 2\pi - \frac{1}{2}x_n^2 \qquad (3)$$

$$\log \rho_{n2} = \mathbb{E}_{\tau}[\log \tau] - \frac{1}{2}\log 2\pi - \frac{1}{2}\mathbb{E}_{\theta}[(x_n-\theta)^2] \qquad (4)$$

By exponentiating both sides of Equation 2, we obtain

$$q^*(\mathbf{z}) \propto \prod_{n=1}^{N} \prod_{k=1}^{2} \rho_{nk}^{z_{nk}}$$

which can be normalized to make a proper distribution

$$q^*(\mathbf{z}) = \prod_{n=1}^{N} \prod_{k=1}^{2} r_{nk}^{z_{nk}}$$

where

$$r_{nk} = \frac{\rho_{nk}}{\sum_{j=1}^{2} \rho_{nj}}$$

Note that computing r_{nk} requires $\mathbb{E}_\tau[\log\tau], \mathbb{E}_\tau[\log(1-\tau)]$, and $\mathbb{E}_\theta[(x_n-\theta)^2]$, where the expectations are computed over the distribution $q(\tau)$ and $q(\theta)$, which will be derived next.

Update factor $q(\tau)$

$$\begin{aligned}
&\log q^*(\tau) = \mathbb{E}_{\mathbf{z},\theta}[\log p(\mathbf{x},\mathbf{z},\tau,\theta)] \\
&= \log p(\tau) + \mathbb{E}_{\mathbf{z}}[\log p(\mathbf{z}|\tau)] + \text{const} \\
&= \log p(\tau) + \sum_{n=1}^{N} \mathbb{E}_{z_n}[z_{n2}]\log\tau + \mathbb{E}_{z_n}[z_{n1}]\log(1-\tau) + \text{const} \\
&= \log p(\tau) + \sum_{n=1}^{N} r_{n2}\log\tau + r_{n1}\log(1-\tau) + \text{const} \\
&= \log \tau^{\alpha_0-1} + \log(1-\tau)^{\alpha_0-1} + \sum_{n=1}^{N} \log\tau^{r_{n2}} + \log(1-\tau)^{r_{n1}} + \text{const} \qquad (5) \\
&= \log \tau^{\sum_{n=1}^{N} r_{n2}+\alpha_0-1} + \log(1-\tau)^{\sum_{n=1}^{N} r_{n1}+\alpha_0-1} + \text{const} \qquad (6)
\end{aligned}$$

We exponentiate and recognize the exponentiated form as

$$q^*(\tau) = Beta(\tau|N_2+\alpha_0, N_1+\alpha_0)$$

We exponentiate and recognize the exponentiated form as

$$q^*(\tau) = Beta(\tau|N_2+\alpha_0, N_1+\alpha_0)$$

i.e., τ has $Beta(a,b)$ with $a = N_2+\alpha_0$ and $b = N_1+\alpha_0$, where $N_k = \sum_{n=1}^{N} r_{nk}$ for $k = 1, 2$. Using this distribution, we get the following formulas for the terms required when updating $q(\mathbf{z})$

$$\mathbb{E}_\tau[\log\tau] = \psi(N_2+\alpha_0) - \psi(N_1+N_2+2\alpha_0) \qquad (7)$$

$$\mathbb{E}_\tau[\log(1-\tau)] = \psi(N_1+\alpha_0) - \psi(N_1+N_2+2\alpha_0) \qquad (8)$$

where ψ is the digamma function. Formulas above follow from the basic property of Beta distribution and the fact that if $\tau \sim Beta(a,b)$ then $1-\tau \sim Beta(b,a)$

Update factor $q(\theta)$

$$
\begin{aligned}
&\log q^*(\theta) = \mathbb{E}_{\tau,\mathbf{z}}[\log p(\mathbf{x}, \mathbf{z}, \tau, \theta)] \\
&= \log p(\theta) + \mathbb{E}_{\mathbf{z}}[\log p(\mathbf{x}|\mathbf{z}, \theta)] + \text{const} \\
&= -\frac{1}{2}\log \beta_0^{-1} - \frac{\beta_0}{2}\theta^2 + \mathbb{E}_{\mathbf{z}}\left[\sum_{n=1}^{N} z_{n1}\left(-\frac{1}{2}x_n^2\right) + z_{n2}\left(-\frac{1}{2}(x_n-\theta)^2\right)\right] + \text{const} \\
&= -\frac{1}{2}\log \beta_0^{-1} - \frac{\beta_0}{2}\theta^2 + \sum_{n=1}^{N} \mathbb{E}_{z_n}[z_{n1}]\left(-\frac{1}{2}x_n^2\right) + \mathbb{E}_{z_n}[z_{n2}]\left(-\frac{1}{2}(x_n-\theta)^2\right) + \text{const} \\
&= -\frac{1}{2}\log \beta_0^{-1} - \frac{\beta_0}{2}\theta^2 + \sum_{n=1}^{N} r_{n1}\left(-\frac{1}{2}x_n^2\right) + r_{n2}\left(-\frac{1}{2}(x_n-\theta)^2\right) + \text{const} \\
&= -\frac{\beta_0}{2}\theta^2 + \sum_{n=1}^{N} -\frac{r_{n2}}{2}\left(x_n^2 - 2x_n\theta + \theta^2\right) + \text{const} \\
&= -\frac{1}{2}\left(\left(\beta_0 + \sum_{n=1}^{N} r_{n2}\right)\theta^2 + \sum_{n=1}^{N} r_{n2}x_n^2 - 2\theta\sum_{n=1}^{N} r_{n2}x_n\right) + \text{const} \\
&= -\frac{\beta_0 + \sum_{n=1}^{N} r_{n2}}{2}\left(\theta - \frac{1}{\beta_0 + \sum_{n=1}^{N} r_{n2}}\sum_{n=1}^{N} r_{n2}x_n\right)^2 + \text{const} \qquad (9)
\end{aligned}
$$

Again, we exponentiate both sides of 9 and recognize this as

$$q^*(\theta) = N(\theta|m_2, \beta_2^{-1}) \qquad (10)$$

with

$$\beta_2 = \beta_0 + N_2 \quad \text{and} \quad m_2 = \beta_2^{-1}N_2\bar{x}_2$$

where we have defined

$$\bar{x}_2 = \frac{1}{N_2}\sum_{n=1}^{N} r_{n2}x_n$$

We can use the distribution 10 to compute $\mathbb{E}_\theta[(x_n-\theta)^2]$, needed when updating $q(\mathbf{z})$:

$$
\begin{aligned}
\mathbb{E}_\theta[(x_n-\theta)^2] &= \mathbb{E}_\theta[(x_n - m_2 + m_2 - \theta)^2] \\
&= (x_n - m_2)^2 + 2(x_n - m_2)\mathbb{E}[m_2 - \theta] + \mathbb{E}[(m_2-\theta)^2] \\
&= (x_n - m_2)^2 + \beta_2^{-1} \qquad (11)
\end{aligned}
$$

The overall VB algorithm is obtained by cycling through updating:

- The responsibilities r_{nk} using formulas 3, 4, 5
- The terms 11 needed when computing the responsibilities
- The term 7 and 8 needed when computing the responsibilities

2 Derivation of ELBO for the Simple Model

Recall that variational inference is based on the decomposition.

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

where $q(\mathbf{Z})$ is any approximation to the posterior distribution $p(\mathbf{Z}|\mathbf{X})$ of the unobserved variables $\mathbf{Z}$ in the model, given the observed variables $\mathbf{X}$. The goal of the variational inference algorithm is to maximize the evidence lower bound (ELBO) $\mathcal{L}(q)$, or equivalently minimize the KL-divergence $\mathrm{KL}[q|p]$ between the approximation and the true posterior. Here, we show how to compute the ELBO for the "simple model" derived earlier. Briefly, the model is

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

The latent variable representation is given by

$$p(\mathbf{x}|\mathbf{z},\theta) = \prod_{n=1}^{N} N(x_n|0,1)^{z_{n1}} N(x_n|\theta,1)^{z_{n2}} \tag{12}$$

and

$$p(\mathbf{z}|\tau) = \prod_{n=1}^{N} \tau^{z_{n2}}(1-\tau)^{z_{n1}} \tag{13}$$

Priors are specified as follows

$$p(\tau) = Beta(\tau|\alpha_0,\beta_0) \propto \tau^{\alpha_0-1}(1-\tau)^{\alpha_0-1}$$
$$p(\theta) = N(\theta|0,\beta_0^{-1}) \propto \exp\left(-\frac{\beta_0}{2}\theta^2\right)$$

The logarithm of the joint distribution can be written as:

$$\log p(\mathbf{x},\mathbf{z},\tau,\theta) = \log p(\tau) + \log p(\theta) + \log p(\mathbf{z}|\tau) + \log p(\mathbf{x}|\mathbf{z},\theta) \tag{14}$$

We assume the mean-field approximation.

$$p(\mathbf{z},\tau,\theta|\mathbf{x}) \approx q(\tau)q(\theta)\prod_{n} q(z_n) \tag{15}$$

Assume that currently, we have factors.

$$q(z_n|r_{n1},r_{n2}) = Categorical(z_n|r_{n1},r_{n2}) = r_{n1}^{z_{n1}} r_{n2}^{z_{n2}} \tag{16}$$
$$q(\tau) = Beta(\tau|\alpha_\tau,\beta_\tau) \tag{17}$$
$$q(\theta) = N(\theta|m_2,\beta_2^{-1}) \tag{18}$$

where $r_{n1}, r_{n2}, n = 1,\cdots,N, \alpha_\tau, \beta_\tau, m_2, \beta_2$ are so-called variational parameters, i.e., parameters that specify the exact distribution of the factor. The general formula of ELBO is given by

$$\begin{aligned}\mathcal{L}(q) &= \int q(\mathbf{Z}) \frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})} d\mathbf{Z} \\ &= \mathbb{E}_q[\log p(\mathbf{X}, \mathbf{Z})] - \mathbb{E}[\log q(\mathbf{Z})] \end{aligned} \tag{19}$$

where $\mathbf{Z}$ is a generic notation that includes all unobservables. We then rewrite ELBO as follows:

$$\begin{aligned}\mathcal{L}(q) &= \mathbb{E}_{q(\tau)q(\theta)q(\mathbf{z})}[\log p(\mathbf{x}, \mathbf{z}, \tau, \theta)] - \mathbb{E}_{q(\tau)q(\theta)q(\mathbf{z})}[\log q(\tau)q(\theta)q(\mathbf{z})] \\ &= \mathbb{E}_{q(\tau)q(\theta)q(\mathbf{z})}[\log p(\tau) + \log p(\theta) + \log p(\mathbf{z}|\tau) + \log p(\mathbf{x}|\mathbf{z}, \theta)] \\ &\quad - \mathbb{E}_{q(\tau)q(\theta)q(\mathbf{z})}[\log q(\tau) + \log q(\theta) + \log q(\mathbf{z})] \\ &= \mathbb{E}_{q(\tau)}[\log p(\tau)] + \mathbb{E}_{q(\theta)}[\log p(\theta) + \mathbb{E}_{q(\tau)q(\mathbf{z})}[\log p(\mathbf{z}|\tau)] + \mathbb{E}_{q(\mathbf{z})q(\theta)}[\log p(\mathbf{x}|\mathbf{z}, \theta) \\ &\quad - \mathbb{E}_{q(\mathbf{z})}[\log q(\mathbf{z})] - \mathbb{E}_{q(\tau)}[\log q(\tau)] - \mathbb{E}_{q(\theta)}[\log q(\theta)] \end{aligned} \tag{20}$$

As with the simple model, all seven terms in formula 20 can be computed analytically when conjugate priors are used. Below, we consider each of these terms. The ELBO can be computed simply by plugging each derived term into Equation 20. In these derivations, we will occasionally discard some terms that do not depend on the variational parameters, as our purpose of deriving the ELBO is to monitor the convergence of the VB algorithm, and those terms are constant across the iterations.

1st term in Equation 20:

$$\begin{aligned}\mathbb{E}_{q(\tau)}[\log p(\tau)] &= \mathbb{E}_{q(\tau)}[(\alpha_0 - 1)\log\tau + (\alpha_0 - 1)\log(1-\tau)] \\ &= (\alpha_0 - 1)\mathbb{E}_{q(\tau)}[\log\tau] + (\alpha_0 - 1)\mathbb{E}_{q(\tau)}[\log(1-\tau)] \\ &= (\alpha_0 - 1)[\psi(\alpha_\tau) - \psi(\alpha_\tau + \beta_\tau)] + (\alpha_0 - 1)[\psi(\beta_\tau) - \psi(\alpha_\tau + \beta_\tau)]\end{aligned}$$

2nd term in Equation 20:

$$\begin{aligned}\mathbb{E}_{q(\theta)}[\log p(\theta)] &= \mathbb{E}_{q(\theta)}\left[-\frac{\beta_0}{2}\theta^2\right] \\ &= -\frac{\beta_0}{2}(\mathbb{V}[\theta] + \mathbb{E}[\theta]^2) \\ &= -\frac{\beta_0}{2}\left(\beta_2^{-1} + m_2^2\right)\end{aligned}$$

3rd term in Equation 20:

$$\begin{aligned}
\mathbb{E}_{q(\tau)q(\mathbf{z})}[\log p(\mathbf{z}|\tau)] &= \sum_{n=1}^{N} \mathbb{E}_{q(\tau)q(z_n)}[\log p(z_n|\tau)] \\
&= \sum_{n=1}^{N} \mathbb{E}_{q(\tau)q(z_n)}[z_{n2}\log\tau + z_{n1}\log(1-\tau)] \\
&= \sum_{n=1}^{N} \mathbb{E}_{q(z_n)}[z_{n2}]\mathbb{E}_{q(\tau)}[\log\tau] + \mathbb{E}_{q(z_n)}[z_{n1}]\mathbb{E}_{q(\tau)}[\log(1-\tau)] \\
&= \sum_{n=1}^{N} r_{n2}[\psi(\alpha_\tau) - \psi(\alpha_\tau+\beta_\tau)] + r_{n1}[\psi(\beta_\tau) - \psi(\alpha_\tau+\beta_\tau)]
\end{aligned}$$

4th term in Equation 20:

$$\begin{aligned}
\mathbb{E}_{q(\mathbf{z})q(\theta)}[\log p(\mathbf{x}|\mathbf{z},\theta)] &= \sum_{n=1}^{N} \mathbb{E}_{q(z_n)q(\theta)}\left[-\frac{z_{n1}}{2}(\log 2\pi + x_n^2) - \frac{z_{n2}}{2}(\log 2\pi + (x_n-\theta)^2)\right] \\
&= \sum_{n=1}^{N} -\mathbb{E}_{q(z_n)}[z_{n1}]\mathbb{E}_{q(\theta)}\left[\frac{1}{2}(\log 2\pi + x_n^2)\right] - \mathbb{E}_{q(z_n)}[z_{n2}]\mathbb{E}_{q(\theta)}\left[\frac{1}{2}(\log 2\pi + (x_n-\theta)^2)\right] \\
&= -\frac{1}{2}\sum_{n=1}^{N} r_{n1}\log 2\pi - \frac{1}{2}\sum_{n=1}^{N} r_{n1}x_n^2 - \frac{1}{2}\sum_{n=1}^{N} r_{n2}\log 2\pi - \frac{1}{2}\sum_{n=1}^{N} r_{n2}\mathbb{E}_{q(\theta)}[(x_n-\theta)^2] \\
&= -\frac{1}{2}\sum_{n=1}^{N}(r_{n1}+r_{n2})\log 2\pi - \frac{1}{2}\sum_{n=1}^{N} r_{n1}x_n^2 - \frac{1}{2}\sum_{n=1}^{N} r_{n2}(\mathbb{E}_{q(\theta)}[(x_n-\theta)]^2 + \mathbb{V}[x_n-\theta]) \\
&= -\frac{1}{2}\sum_{n=1}^{N}(r_{n1}+r_{n2})\log 2\pi - \frac{1}{2}\sum_{n=1}^{N} r_{n1}x_n^2 - \frac{1}{2}\sum_{n=1}^{N} r_{n2}((x_n-\mathbb{E}_{q(\theta)}[\theta])]^2 + \mathbb{V}[\theta]) \\
&= -\frac{N}{2}\log(2\pi) - \frac{1}{2}\sum_{n=1}^{N} r_{n1}x_n^2 - \frac{1}{2}\sum_{n=1}^{N} r_{n2}((x_n-m_2)^2 + \beta_2^{-1})
\end{aligned}$$

5th term in Equation 20:

$$\begin{aligned}
\mathbb{E}_{q(\mathbf{z})}[\log q(\mathbf{z})] &= \sum_{n=1}^{N} \mathbb{E}_{q(z_n)}[z_{n1}\log r_{n1} + z_{n2}\log r_{n2}] \\
&= \sum_{n=1}^{N} r_{n1}\log r_{n1} + r_{n2}\log r_{n2}
\end{aligned}$$

6th term in Equation 20:

$$\mathbb{E}_{q(\tau)}[\log q(\tau)] = \log\frac{\Gamma(\alpha_\tau+\beta_\tau)}{\Gamma(\alpha_\tau)\Gamma(\beta_\tau)} + (\alpha_\tau-1)\psi(\alpha_\tau) + (\beta_\tau-1)\psi(\beta_\tau) - (\alpha_\tau+\beta_\tau-2)\psi(\alpha_\tau+\beta_\tau)$$

This is just the negative entropy of $Beta(\alpha_\tau, \beta_\tau)$

7th term in Equation 20:
By the definition of the negative entropy of normal distribution, we obtain

$$\mathbb{E}_{q(\theta)}[\log q(\theta)] = -\frac{1}{2}\log(2\pi e\beta_2^{-1})$$

3 Variational Bayes for a Factor Analysis

The data set consists of D-dimensional vectors $\mathbf{x}_n \in \mathbb{R}^D$, for $n = 1, \ldots, N$. We model the data using factor analysis with K-dimensional factors $\mathbf{z}_n \in \mathbb{R}^K$. In detail, the model is specified as follows:

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

Here, $\mathbf{W}$ is a $D \times K$ factor loading matrix and $\mathbf{w}_d$ is the dth row of $\mathbf{W}$ written as a column vector. Parameter ψ_d^{-1} is the variance for the dth dimension in the observed data and $\mathrm{diag}(\psi)$ denotes a diagonal matrix with elements $\psi = (\psi_1, \ldots, \psi_D)^T$ on the diagonal.

We approximate the posterior $p(\psi, \mathbf{Z}, \mathbf{W}|\mathbf{X})$ using the mean-field approximation:

$$q(\Theta) = \prod_{d=1}^{D} q(\mathbf{w}_d) \prod_{n=1}^{N} q(\mathbf{z}_n) \prod_{d=1}^{D} q(\psi_d).$$

The goal is deriving the update factor $q(\mathbf{z}_n)$ and $q(\mathbf{w}_d)$, respectively. Initially, write the logarithm of the joint distribution, $\log p(\psi, \mathbf{Z}, \mathbf{W}, \mathbf{X})$ as follows:

$$\begin{aligned}
\log p(\psi, \mathbf{Z}, \mathbf{W}, \mathbf{X}) &= \log p(\mathbf{X}|\mathbf{W}, \mathbf{Z}, \psi) + \log p(\mathbf{W}) + \log p(\mathbf{Z}) + \log p(\psi) \\
&= \sum_{n=1}^{N} \log p(\mathbf{x}_n|\mathbf{W}, \mathbf{z}_n, \psi) + \sum_{d=1}^{D} \log p(\mathbf{w}_d) + \sum_{n=1}^{N} \log p(\mathbf{z}_n) + \sum_{d=1}^{D} \log \psi_d
\end{aligned}$$

Update factor $q(\mathbf{z}_n)$

$$
\begin{aligned}
\log q^*(\mathbf{z}_n) &= \mathbb{E}_{\mathbf{W},\mathbf{Z}_{\setminus n},\psi}[\log p(\mathbf{x}_n|\mathbf{W},\mathbf{z}_n,\psi) + \log p(\mathbf{z}_n)] + const \\
&= \mathbb{E}_{\mathbf{W},\psi}[\log p(\mathbf{x}_n|\mathbf{W},\mathbf{z}_n,\psi)] + \log p(\mathbf{z}_n) + const \\
&= \mathbb{E}_{\mathbf{W},\psi}[-\frac{1}{2}(\mathbf{x}_n - \mathbf{W}\mathbf{z}_n)^\top \mathrm{diag}(\psi)(\mathbf{x}_n - \mathbf{W}\mathbf{z}_n)] - \frac{1}{2}\mathbf{z}_n^\top \mathbf{z}_n + const \\
&= \mathbb{E}_{\mathbf{W},\psi}[-\frac{1}{2}\left(\mathbf{x}_n^\top \mathrm{diag}(\psi)\mathbf{x}_n + \mathbf{z}_n^\top \mathbf{W}^\top \mathrm{diag}(\psi)\mathbf{W}\mathbf{z}_n - 2\mathbf{z}_n^\top \mathbf{W}^\top \mathrm{diag}(\psi)\mathbf{x}_n\right) - \frac{1}{2}\mathbf{z}_n^\top \mathbf{z}_n] + const \\
&= -\frac{1}{2}\left(\mathbb{E}_{\mathbf{W},\psi}[\mathbf{z}_n^\top (\mathbf{W}^\top \mathrm{diag}(\psi)\mathbf{W} + \mathbf{I})\mathbf{z}_n] - 2\mathbb{E}_{\mathbf{W},\psi}[\mathbf{z}_n^\top \mathbf{W}^\top \mathrm{diag}(\psi)\mathbf{x}_n]\right) + const \\
&= -\frac{1}{2}\left(\mathbf{z}_n^\top \left(\sum_{d=1}^{D} \mathbb{E}_\psi[\psi_d]\mathbb{E}_{\mathbf{W}}[\mathbf{w}_d\mathbf{w}_d^\top] + \mathbf{I}\right)\mathbf{z}_n - 2\mathbf{z}_n^\top \mathbb{E}_{\mathbf{W}}[\mathbf{W}]^\top \mathbb{E}_\psi[\mathrm{diag}(\psi)]\mathbf{x}_n\right) + const \\
&= -\frac{1}{2}\left(\mathbf{z}_n^\top \left(\sum_{d=1}^{D} \langle\psi_d\rangle\langle\mathbf{w}_d\mathbf{w}_d^\top\rangle + \mathbf{I}\right)\mathbf{z}_n - 2\mathbf{z}_n^\top \langle\mathbf{W}^\top\rangle \mathrm{diag}(\langle\psi\rangle)\mathbf{x}_n\right) + const
\end{aligned}
$$

Let

$$
\begin{aligned}
\mathbf{K}_n &= \left(\sum_{d=1}^{D} \langle\psi_d\rangle\langle\mathbf{w}_d\mathbf{w}_d^\top\rangle + \mathbf{I}\right)^{-1} \\
\hat{\mu}_n &= \langle\mathbf{W}^\top\rangle \mathrm{diag}(\langle\psi\rangle)\mathbf{x}_n
\end{aligned}
$$

By completing the square, we obtain

$$
\log q^*(\mathbf{z}_n) = -\frac{1}{2}\left(\mathbf{z}_n - \mathbf{K}_n\hat{\mu}_n\right)^\top \mathbf{K}_n^{-1}(\mathbf{z}_n - \mathbf{K}_n\hat{\mu}_n)
$$

It implies that $q^*(\mathbf{z}_n) = \mathcal{N}(\mathbf{z}_n|\mu_n, \mathbf{K}_n)$ with $\mu_n = \mathbf{K}_n\hat{\mu}_n$.

Update factor $q(\mathbf{w}_d)$

$$
\begin{aligned}
\log q^*(\mathbf{w}_d) &= \mathbb{E}_{\mathbf{Z},\mathbf{W}_{\setminus d},\psi}\left[\sum_{n=1}^{N}\log p(\mathbf{x}_n|\mathbf{W},\mathbf{z}_n,\psi)\right] + \log p(\mathbf{w}_d) + const \\
&= \mathbb{E}_{\mathbf{Z},\mathbf{W}_{\setminus d},\psi}\left[\sum_{n=1}^{N}\sum_{d=1}^{D} -\frac{1}{2}(\mathbf{x}_{nd} - \mathbf{w}_d^\top \mathbf{z}_n)^\top \psi_d (\mathbf{x}_{nd} - \mathbf{w}_d^\top \mathbf{z}_n)\right] - \frac{1}{2\alpha}\mathbf{w}_d^\top \mathbf{w}_d + const \\
&= \mathbb{E}_{\mathbf{Z},\mathbf{W}_{\setminus d},\psi}\left[\sum_{n=1}^{N} -\frac{1}{2}(\mathbf{x}_{nd}^\top \psi_d \mathbf{x}_{nd} + \mathbf{z}_n^\top \mathbf{w}_d \psi_d \mathbf{w}_d^\top \mathbf{z}_n - 2\mathbf{z}_n^\top \mathbf{w}_d \psi_d \mathbf{x}_{nd})\right] - \frac{1}{2\alpha}\mathbf{w}_d^\top \mathbf{w}_d + const \\
&= \mathbb{E}_{\mathbf{Z},\mathbf{W}_{\setminus d},\psi}\left[\sum_{n=1}^{N} -\frac{1}{2}(\mathbf{z}_n^\top \mathbf{w}_d \psi_d \mathbf{w}_d^\top \mathbf{z}_n - 2\mathbf{z}_n^\top \mathbf{w}_d \psi_d \mathbf{x}_{nd})\right] - \frac{1}{2\alpha}\mathbf{w}_d^\top \mathbf{w}_d + const \\
&= \mathbb{E}_{\mathbf{Z},\mathbf{W}_{\setminus d},\psi}\left[-\frac{1}{2}\left(\mathbf{w}_d^\top \left(\psi_d \sum_{n=1}^{N}\mathbf{z}_n \mathbf{z}_n^\top\right)\mathbf{w}_d - 2\mathbf{w}_d^\top \psi_d \sum_{n=1}^{N}\mathbf{z}_n \mathbf{x}_{nd}\right)\right] - \frac{1}{2\alpha}\mathbf{w}_d^\top \mathbf{w}_d + const \\
&= \mathbb{E}_{\mathbf{Z},\mathbf{W}_{\setminus d},\psi}\left[-\frac{1}{2}\left(\mathbf{w}_d^\top \left(\psi_d \sum_{n=1}^{N}\mathbf{z}_n \mathbf{z}_n^\top + \alpha^{-1}\mathbf{I}\right)\mathbf{w}_d - 2\mathbf{w}_d^\top \psi_d \sum_{n=1}^{N}\mathbf{z}_n \mathbf{x}_{nd}\right)\right] + const \\
&= -\frac{1}{2}\left(\mathbf{w}_d^\top \left(\mathbb{E}_\psi[\psi_d]\sum_{n=1}^{N}\mathbb{E}_{\mathbf{Z}}\left[\mathbf{z}_n \mathbf{z}_n^\top\right] + \alpha^{-1}\mathbf{I}\right)\mathbf{w}_d - 2\mathbf{w}_d^\top \mathbb{E}_\psi[\psi_d] \sum_{n=1}^{N}\mathbb{E}_{\mathbf{Z}}[\mathbf{z}_n]\mathbf{x}_{nd}\right) + const \\
&= -\frac{1}{2}\left(\mathbf{w}_d^\top \left(\langle\psi_d\rangle \sum_{n=1}^{N}\langle \mathbf{z}_n \mathbf{z}_n^\top\rangle + \alpha^{-1}\mathbf{I}\right)\mathbf{w}_d - 2\mathbf{w}_d^\top \langle\psi_d\rangle \sum_{n=1}^{N}\langle\mathbf{z}_n\rangle \mathbf{x}_{nd}\right) + const
\end{aligned}
$$

Let

$$
\begin{aligned}
\mathbf{K}_d &= \left(\sum_{n=1}^{N}\langle\psi_d\rangle\langle \mathbf{z}_n \mathbf{z}_n^\top\rangle + \alpha^{-1}\mathbf{I}\right)^{-1} \\
\hat{\mu}_d &= \sum_{n=1}^{N}\langle\psi_d\rangle\langle\mathbf{z}_n\rangle \mathbf{x}_{nd}
\end{aligned}
$$

By completing the square, we obtain

$$
\log q^*(\mathbf{w}_d) = -\frac{1}{2}\left(\mathbf{w}_d - \mathbf{K}_d \hat{\mu}_d\right)^\top \mathbf{K}_d^{-1}\left(\mathbf{w}_d - \mathbf{K}_d \hat{\mu}_d\right)
$$

It implies that $q^*(\mathbf{w}_d) = \mathcal{N}(\mathbf{w}_d|\mu_d, \mathbf{K}_d)$ with $\mu_d = \mathbf{K}_d\hat{\mu}_d$.

4 Bayesian Linear Regression with Stochastic Variational Inference

The model is defined as follows:

$$
\begin{aligned}
&y_i \sim \mathcal{N}(w_0 + w_1 x_i, \sigma_l^2), \quad x_i \in \mathbb{R}, \sigma_l = 0.4, i = 1, \ldots, N \\
&\mathbf{w} \sim \mathcal{N}(0, \alpha^2 I).
\end{aligned}
$$

Given data $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$, we are interested in the posterior distribution $p(\mathbf{w}|\mathcal{D})$ which we approximate using mean-field approximation:

$$p(\mathbf{w}|\mathcal{D}) \approx q(\mathbf{w}) = \prod_{d=0}^{1} q(w_d) = \prod_{d=0}^{1} \mathcal{N}(w_d|\mu_d, \sigma_d^2)$$

That is, we model each w_d as an independent Gaussian with mean μ_d and σ_d^2 and use SVI to optimize them such that:

$$\hat{\lambda} = \text{argmin}_\lambda \text{KL}[q(\mathbf{w})|p(\mathbf{w}|\mathcal{D})] \tag{21}$$

$$= \text{argmin}_\lambda \underbrace{\mathbb{E}_{q_\lambda(\mathbf{w})}\left[-\log p(\mathcal{D}|\mathbf{w})\right] + \text{KL}\left[q(\mathbf{w})|p(\mathbf{w})\right]}_{Loss = -ELBO} + c. \tag{22}$$

Here, the variational parameters are denoted by $\lambda = \{(\mu_d, \sigma_d), i = 0, 1\}$. The first term of the ELBO is the expected log-likelihood, which will be estimated using a pathwise estimator, and the second term is the KL between the approximate posterior $q_\lambda(\mathbf{w})$ and the prior $p(\mathbf{w})$ that can be derived analytically in this case.

4.1 Negative Log-likelihood

$$\begin{aligned}
-\log p(\mathcal{D}|\mathbf{w}) &= -\log \prod_{i=1}^{N} p(y_i|x_i, \mathbf{w}, \sigma_l^2) \\
&= -\sum_{i=1}^{N} \log p(y_i|x_i, \mathbf{w}, \sigma_l^2) \\
&= \frac{1}{2\sigma_l^2} \sum_{i=1}^{N} \left(y_i - (w_1 x_i + w_0)\right)^2 + const
\end{aligned}$$

Let MSE defined as follows

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^{N} \left(y_i - w_1 x_i + w_0\right)^2$$

Thus, we are able to rewrite $-\log p(\mathcal{D}|\mathbf{w})$ as

$$-\log p(\mathcal{D}|\mathbf{w}) = \frac{N}{2\sigma_l^2} \text{MSE} + const$$

4.2 Deriving KL-divergence

$$
\begin{aligned}
\mathrm{KL}\left[q(\mathbf{w})|p(\mathbf{w})\right] &= \mathbb{E}_{q(\mathbf{w})}\left[\frac{\log q(\mathbf{w})}{p(\mathbf{w})}\right] = \mathbb{E}_{q(\mathbf{w})}\left[\log q(\mathbf{w}) - \log p(\mathbf{w})\right] \\
&= \mathbb{E}_{q(\mathbf{w})}\left[\sum_{d=0}^{1} \log q(w_d) - \sum_{d=0}^{1} p(w_d)\right] \\
&= \sum_{d=0}^{1} \mathbb{E}_{q(\mathbf{w})}[\log q(w_d)] - \mathbb{E}_{p(\mathbf{w})}[\log p(w_d)] \\
&= \sum_{d=0}^{1} \mathbb{E}_{q(w_d)}[\log q(w_d)] - \mathbb{E}_{p(w_d)}[\log p(w_d)] \\
&= \sum_{d=0}^{1} -H_{q(w_d)}(w_d) + H_{p(w_d)}(w_d) \\
&= \log(2\pi e\alpha^2) + \frac{1}{2}\sum_{d=0}^{1} -\log(2\pi e\sigma_d^2)
\end{aligned}
$$

with $H_{q(w_d)}$ and $H_{p(w_d)}$ denote the entropy w.r.t $q(w_d)$ and $p(w_d)$, respectively.