Indian Institute of Technology Kanpur
CS698X: Probabilistic Modelling and Inference, 18–19

ASSIGNMENT 3

Student Name: Gurpreet Singh
Roll Number: 150259
Date: March 3, 2018

Question 1

1.1. Effect of Prior

The above prior facilitates sparse learning for the weight parameters, however the sparsity is only induced for some of the parameters, and the "intensity" of the sparsity is of two types, one for which the precision is high, and one for which it is lower.

The given prior somewhat classifies the parameters into two categories based on their importance, and contribution to the regression model.

1.2. EM Algorithm

1.2.1. Posterior over Latent Variables

We need to find the posterior over the latent variables, here $\mathbf{w}$. Using Bayes rule, we know

$$\mathbb{P}\left[\mathbf{w} \,\middle|\, \mathbf{y}, \mathbf{X}, \sigma^2, \boldsymbol{\gamma}\right] \quad \propto \quad \mathbb{P}\left[\mathbf{y} \,\middle|\, \mathbf{X}, \mathbf{w}, \sigma^2\right] \mathbb{P}\left[\mathbf{w} \,\middle|\, \sigma^2, \boldsymbol{\gamma}\right]$$

Since the likelihood model and prior are both simply gaussians, we can write

$$\begin{aligned} \textbf{Likelihood:} \qquad \mathbb{P}\left[\mathbf{y} \,\middle|\, \mathbf{X}, \mathbf{w}, \sigma^2\right] &= \mathcal{N}\left(\mathbf{y} \,\middle|\, \mathbf{X}\mathbf{w}, \sigma^2 \mathbf{I}_N\right) \\ \textbf{Prior:} \qquad \mathbb{P}\left[\mathbf{w} \,\middle|\, \sigma^2, \boldsymbol{\gamma}\right] &= \mathcal{N}\left(\mathbf{w} \,\middle|\, \mathbf{0}, \sigma^2 \mathbf{K}\right) \end{aligned}$$

where $\mathbf{K} = \operatorname{diag}\left(\kappa_{\gamma_1}, \kappa_{\gamma_2} \ldots \kappa_{\gamma_D}\right)$ is a diagonal covariance matrix.

Since this is simply a linear regression setting, with different regularization weights or prior variance for different dimensions of the weight vector, we can use Completing the Squares trick to find the posterior, which will be as follows

$$\mathbb{P}\left[\mathbf{w} \,\middle|\, \mathbf{y}, \mathbf{X}, \sigma^2, \boldsymbol{\gamma}\right] \quad = \quad \mathcal{N}\left(\mathbf{w} \,\middle|\, \boldsymbol{\mu}_w, \boldsymbol{\Sigma}_w\right) \tag{1}$$

where

$$\begin{aligned} \boldsymbol{\Sigma}_w &= \sigma^2 \left(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \mathbf{K}^{-1}\right)^{-1} \\ \boldsymbol{\mu}_w &= \frac{1}{\sigma^2} \boldsymbol{\Sigma}_w \mathbf{X}^{\mathrm{T}} \mathbf{y} \end{aligned}$$

1.2.2. Expectation of CLL

We can now write the Complete Data Log-Likelihood $\left(\log \mathbb{P}\left[\mathbf{w}, \mathbf{y} \,\middle|\, \mathbf{X}, \sigma^2, \boldsymbol{\gamma}\right]\right)$ term as follows

$$\begin{aligned}
\log \mathbb{P}\left[\mathbf{w}, \mathbf{y} \,\middle|\, \mathbf{X}, \sigma^2, \boldsymbol{\gamma}\right] &= \log \mathbb{P}\left[\mathbf{y} \,\middle|\, \mathbf{X}, \mathbf{w}, \sigma^2\right] + \log \mathbb{P}\left[\mathbf{w} \,\middle|\, \sigma^2, \boldsymbol{\gamma}\right] \\
&= -\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\mathbf{w})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\mathbf{w}) - \frac{N+D}{2}\log\left(2\pi\sigma^2\right) - \\
&\qquad - \frac{1}{2\sigma^2}\mathbf{w}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{w} - \sum_{d=1}^{D}\frac{1}{2}\log\left(\kappa_{\gamma_d}\right)
\end{aligned}$$

Therefore, we can write the Expected CLL ($\mathbb{E}\left[\,\mathrm{CLL}\,\right]$) as

$$\mathbb{E}\left[\,\mathrm{CLL}\,\right] = \mathop{\mathbb{E}}_{\mathbf{w}\,|\,\mathbf{y}}\left[\log \mathbb{P}\left[\mathbf{w}, \mathbf{y} \,\middle|\, \mathbf{X}, \sigma^2, \boldsymbol{\gamma}\right]\right]$$

$$\begin{aligned}
\implies \mathbb{E}\left[\,\mathrm{CLL}\,\right] &= -\frac{1}{2\sigma^2}\left(\mathbf{y}^{\mathrm{T}}\mathbf{y} - 2\mathbf{y}^{\mathrm{T}}\mathbf{X}\mathbb{E}\left[\,\mathbf{w}\,\right] + \operatorname{Tr}\left(\left(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \mathbf{K}^{-1}\right)\mathbb{E}\left[\,\mathbf{w}\mathbf{w}^{\mathrm{T}}\,\right]\right)\right) - \\
&\qquad - \frac{N+D}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2}\sum_{d=1}^{D}\log\left(\kappa_{\gamma_d}\right)
\end{aligned} \tag{2}$$

Now, we need to compute the expectations required in the above expression. However, we need to note that the expectation is with respect to known values of σ^2 and $\boldsymbol{\gamma}$. Since the posterior of $\mathbf{w}$ is simply a Gaussian, we can directly write the expectation as follows

$$\begin{aligned}
\mathbb{E}\left[\,\mathbf{w}\,\right] &= \boldsymbol{\mu}_w \\
&= \left(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \mathbf{K}^{-1}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} \\
\mathbb{E}\left[\,\mathbf{w}\mathbf{w}^{\mathrm{T}}\,\right] &= \boldsymbol{\Sigma}_w + \boldsymbol{\mu}_w\boldsymbol{\mu}_w{}^{\mathrm{T}} \\
&= \sigma^2\left(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \mathbf{K}^{-1}\right)^{-1} + \mathbf{y}^{\mathrm{T}}\mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \mathbf{K}^{-1}\right)^{-2}\mathbf{X}^{\mathrm{T}}\mathbf{y}
\end{aligned}$$

1.2.3. Maximization Step

In order to compute the MAP, we need to find the posterior with respect to $\mathbb{E}\left[\,CLL\,\right]$. Therefore, we can write the MAP estimates as follows

$$\left\{\sigma^2, \boldsymbol{\gamma}, \theta\right\}_{\mathrm{MAP}} = \mathop{\arg\max}_{\sigma^2, \boldsymbol{\gamma}, \theta}\ \mathbb{E}\left[\,\mathrm{CLL}\,\right] + \log \mathbb{P}\left[\sigma^2, \boldsymbol{\gamma}, \theta\right]$$

Note. We might want to compute the MAP by computing the expectation of the log-posterior over the parameters $\left(\mathbb{P}\left[\sigma^2, \boldsymbol{\gamma}, \theta \,\middle|\, \mathbf{y}, \mathbf{w}, \mathbf{X}\right]\right)$, rather than computing the expectation over CLL, however since the prior term of the parameters is independent of $\mathbf{w}$ and the denominator (of the posterior) is independent of the parameters, we will end up with the same expression as given above.

The prior of the parameters is as follows

$$\begin{aligned}\mathbb{P}\left[\,\sigma^2, \boldsymbol{\gamma}, \theta\,\right] &= \mathbb{P}\left[\,\sigma^2\,\right] \prod_{d=1}^{D} \left\{\mathbb{P}\left[\,\gamma_d \,\middle|\, \theta\,\right]\right\} \mathbb{P}\left[\,\theta\,\right] \\ \implies \log \mathbb{P}\left[\,\sigma^2, \boldsymbol{\gamma}, \theta\,\right] &= \log \mathbb{P}\left[\,\sigma^2\,\right] + \sum_{d=1}^{D} \log \mathbb{P}\left[\,\gamma_d \,\middle|\, \theta\,\right] + \log \mathbb{P}\left[\,\theta\,\right]\end{aligned}$$

From the question, we know

$$\begin{aligned}\log \mathbb{P}\left[\,\sigma^2\,\right] &= -\left(\frac{\nu}{2} + 1\right) \log\left(\sigma^2\right) - \frac{\nu\lambda}{2\sigma^2} + constants \\ \log \mathbb{P}\left[\,\gamma_d \,\middle|\, \theta\,\right] &= \gamma_d \log(\theta) + (1 - \gamma_d) \log(1 - \theta) \\ \log \mathbb{P}\left[\,\theta\,\right] &= (a_0 - 1) \log(\theta) + (b_0 - 1) \log(1 - \theta)\end{aligned}$$

Now, we can use the standard MAP estimation to find the MAP estimates of the parameters.

1. **Update of σ^2**

$$\begin{aligned}0 &= \frac{\delta\left(\mathbb{E}\left[\,\text{CLL}\,\right] + \log \mathbb{P}\left[\,\sigma^2, \boldsymbol{\gamma}, \theta\,\right]\right)}{\delta\left(\sigma^2\right)} \\ &= \frac{1}{2\sigma^4}\left(\mathbf{y}^{\mathrm{T}}\mathbf{y} - 2\mathbf{y}^{\mathrm{T}}\mathbf{X}\mathbb{E}\left[\,\mathbf{w}\,\right] + \,\mathrm{Tr}\left(\left(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \mathbf{K}^{-1}\right)\mathbb{E}\left[\,\mathbf{w}\mathbf{w}^{\mathrm{T}}\,\right]\right)\right) - \frac{N + D}{2\sigma^2} \\ &\qquad - \frac{1}{\sigma^2}\left(\frac{\nu}{2} + 1\right) + \frac{\nu\lambda}{2\sigma^4} \\ \implies \sigma^2 &= \frac{\mathbf{y}^{\mathrm{T}}\mathbf{y} - 2\mathbf{y}^{\mathrm{T}}\mathbf{X}\mathbb{E}\left[\,\mathbf{w}\,\right] + \,\mathrm{Tr}\left(\left(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \mathbf{K}^{-1}\right)\mathbb{E}\left[\,\mathbf{w}\mathbf{w}^{\mathrm{T}}\,\right]\right) + \nu\lambda}{N + D + \nu + 2}\end{aligned}$$

2. **Update of $\gamma_d \,\middle|\, \theta$**

Since $\gamma_d \in \{0, 1\}$, instead of differentiating, we can directly write (ignoring terms independent of γ_d, when known)

$$\begin{aligned}\gamma_d &= \underset{\gamma'_d \in \{0,1\}}{\arg\max}\ \mathbb{E}\left[\,\text{CLL}\,\right] + \log \mathbb{P}\left[\,\sigma^2, \boldsymbol{\gamma}, \theta\,\right] \\ &= \underset{\gamma'_d \in \{0,1\}}{\arg\max}\ -\frac{1}{2\sigma^2\kappa_{\gamma'_d}}\mathbb{E}\left[\,\mathbf{w}\mathbf{w}^{\mathrm{T}}\,\right]_{d,d} - \frac{1}{2}\log\left(\kappa_{\gamma'_d}\right) + \gamma'_d \log(\theta) + (1 - \gamma'_d) \log(1 - \theta)\end{aligned}$$

3. **Update of θ**

$$\begin{aligned}0 &= \frac{\delta\left(\mathbb{E}\left[\,\text{CLL}\,\right] + \log \mathbb{P}\left[\,\sigma^2, \boldsymbol{\gamma}, \theta\,\right]\right)}{\delta(\theta)} \\ &= \frac{1}{\theta}\left(\sum_{d=1}^{D} \gamma_d + a_0 - 1\right) - \frac{1}{1 - \theta}\left(\sum_{d=1}^{D} \{1 - \gamma_d\} + b_0 - 1\right) \\ \implies \theta &= \frac{\sum_{d=1}^{D} \gamma_d + a_0 - 1}{D + a_0 + b_0 - 2}\end{aligned}$$

We can now write the complete EM algorithm as given in Algorithm 1

```
Algorithm 1: EM Algorithm
Input: Input data $\mathbf{y}$ and $\mathbf{X}$ and hyper-paramters $a_0$, $b_0$, $\nu$, $\lambda$, $v_0$ and $v_1$
Output: Approximate Posterior $q(\mathbf{W},\mathbf{Z},\boldsymbol{\pi})$
Steps:
  1. Initialize $\left\{\boldsymbol{\gamma},\sigma^2,\theta\right\}$ as $\left\{\boldsymbol{\gamma},\sigma^2,\theta\right\}^{(0)}$
  2. For $t=0,1\ldots T-1$
     (a) Update the posterior of $\mathbf{w}$ as
           $\mathbb{P}\left[\mathbf{w}^{(t+1)}\,\middle|\,\mathbf{y},\mathbf{X},{\sigma^2}^{(t)},\boldsymbol{\gamma}^{(t)}\right] = \mathcal{N}\left(\mathbf{w}^{(t)}\,\middle|\,\boldsymbol{\mu}_w^{(t+1)},\boldsymbol{\Sigma}_w^{(t+1)}\right)$
         where
           $\boldsymbol{\Sigma}_w^{(t+1)} = {\sigma^2}^{(t)}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}+{\mathbf{K}^{(t)}}^{-1}\right)^{-1}$
           $\boldsymbol{\mu}_w^{(t+1)} = \frac{1}{{\sigma^2}^{(t)}}\boldsymbol{\Sigma}_w^{(t+1)}\mathbf{X}^{\mathrm{T}}\mathbf{y}$
     (b) Update the expectations as
           $\mathbb{E}\left[\,\mathbf{w}\,\right]^{(t+1)} = \boldsymbol{\mu}_w^{(t+1)}$
           $\mathbb{E}\left[\,\mathbf{w}\mathbf{w}^{\mathrm{T}}\,\right]^{(t+1)} = \boldsymbol{\Sigma}_w^{(t+1)}+\boldsymbol{\mu}_w^{(t+1)}{\boldsymbol{\mu}_w^{(t+1)}}^{\mathrm{T}}$
     (c) Upadate the parameters as
           ${\sigma^2}^{(t+1)} = \frac{\mathbf{y}^{\mathrm{T}}\mathbf{y}-2\mathbf{y}^{\mathrm{T}}\mathbf{X}\mathbb{E}\left[\,\mathbf{w}\,\right]^{(t+1)}+\ \mathrm{Tr}\left(\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}+{\mathbf{K}^{(t)}}^{-1}\right)\mathbb{E}\left[\,\mathbf{w}\mathbf{w}^{\mathrm{T}}\,\right]^{(t+1)}\right)+\nu\lambda}{N+D+\nu+2}$
           $\theta^{(t+1)} = \frac{\sum_{d=1}^{D}\gamma_d^{(t)}+a_0-1}{D+a_0+b_0-2}$
           $\gamma_d^{(t+1)} = \underset{\gamma_d\in\{0,1\}}{\arg\max}\ \ \gamma_d\log\left(\theta^{(t+1)}\right)+(1-\gamma_d)\log\left(1-\theta^{(t+1)}\right)$
                 $-\ \frac{1}{2{\sigma^2}^{(t+1)}\kappa_{\gamma_d}}\mathbb{E}\left[\,\mathbf{w}\mathbf{w}^{\mathrm{T}}\,\right]_{d,d}^{(t+1)}-\frac{1}{2}\log\left(\kappa_{\gamma_d}\right)$
  3. Return $\left\{\boldsymbol{\gamma},\sigma^2,\theta\right\}=\left\{\boldsymbol{\gamma},\sigma^2,\theta\right\}^{(T)}$ and $\mathbb{P}\left[\mathbf{w}\,\middle|\,\mathbf{y},\mathbf{X},{\sigma^2}^{(T-1)},\boldsymbol{\gamma}^{(T-1)}\right]$
```

Question 2

Using the Mean-Field Assumption, we have

$$\mathbb{P}\left[\mu,\tau \,\middle|\, \mathbf{x}\right] \quad = \quad q(\mu,\tau) \quad = \quad q(\mu)q(\tau)$$

We can assume that $q(\mu)$ forms a Gaussian distribution and $q(\tau)$ forms a Gamma distribution. This is in fact what we would obtain from inspection. Therefore, we assume

$$\begin{aligned} q(\mu) &= \mathcal{N}\left(\mu \,\middle|\, \mu_N, \lambda_N^{-1}\right) \\ q(\tau) &= \text{Gamma}\left(\tau \,\middle|\, a_N, b_N\right) \end{aligned}$$

Now from definition, we can write the ELBO as follows

$$\mathcal{L}(q) \quad = \quad \int q(\mu,\tau)\ \left(\log \mathbb{P}\left[\mathbf{x} \,\middle|\, \mu,\tau\right] + \log \mathbb{P}\left[\mu,\tau\right] - \log q(\mu,\tau)\right)\ \mathrm{d}\mu\ \mathrm{d}\tau$$

Since the likelihood is a Gaussian and we are assuming the prior to be a Normal-Gamma, we have

$$\begin{aligned} \mathbb{P}\left[\mathbf{x} \,\middle|\, \mu,\tau\right] &= \prod_{n=1}^{N} \mathcal{N}\left(x_n \,\middle|\, \mu, \tau^{-1}\right) \\ \mathbb{P}\left[\mu,\tau\right] &= \mathcal{N}\left(\mu \,\middle|\, \mu_0, (\lambda_0\tau)^{-1}\right)\ \text{Gamma}\left(\tau \,\middle|\, a_0, b_0\right) \end{aligned}$$

We can partially differentiate for every unknown parameter ($\{\mu_N, \lambda_N, a_N, b_N\}$)

$$\frac{\delta\left(\mathcal{L}(q)\right)}{\delta\left(\mu_N\right)} \quad = \quad \int \frac{\delta\left(q(\mu)\right)}{\delta\left(\mu_N\right)} q(\tau) \left(\log \mathbb{P}\left[\mathbf{x} \,\middle|\, \mu,\tau\right] + \log \mathbb{P}\left[\mu,\tau\right] - \log q(\mu,\tau)\right) - \frac{\delta\left(q(\mu)\right)}{\delta\left(\mu_N\right)} q(\tau)\ \mathrm{d}\mu\ \mathrm{d}\tau$$

$$\begin{aligned} \mathcal{L}(q) \quad = \quad & \int q(\mu)q(\tau)\left(-\frac{\tau}{2}\sum_{n=1}^{N}\left(x_n-\mu\right)^2 - \frac{N}{2}\log\left(2\pi\right) + \frac{N}{2}\log\left(\tau\right)\right)\ \mathrm{d}\mu\ \mathrm{d}\tau \\ + & \int q(\mu)q(\tau)\left(-\frac{\lambda_0\tau}{2}\left(\mu-\mu_0\right)^2 - \frac{1}{2}\log\left(2\pi\right) + \frac{1}{2}\log\left(\lambda_0\tau\right)\right)\ \mathrm{d}\mu\ \mathrm{d}\tau \\ + & \int q(\mu)q(\tau)\left(a_0\log\left(b_0\right) - \log\left(\Gamma(a_0)\right) + (a_0-1)\log\left(\tau\right) - b_0\tau\right)\ \mathrm{d}\mu\ \mathrm{d}\tau \\ - & \int q(\mu)q(\tau)\left(-\frac{\lambda_N}{2}\left(\mu-\mu_N\right)^2 - \frac{1}{2}\log\left(2\pi\right) + \frac{1}{2}\log\left(\lambda_N\right)\right)\ \mathrm{d}\mu\ \mathrm{d}\tau \\ - & \int q(\mu)q(\tau)\left(a_N\log\left(b_N\right) - \log\left(\Gamma(a_N)\right) + (a_N-1)\log\left(\tau\right) - b_N\tau\right)\ \mathrm{d}\mu\ \mathrm{d}\tau \end{aligned}$$

We can simplify this as follows

$$\begin{aligned} \mathcal{L}(q) \quad = \quad & -\frac{\mathbb{E}\left[\tau\right]}{2}\sum_{n=1}^{N}\mathbb{E}\left[\left(x_n-\mu\right)^2\right] - \frac{\lambda_0\mathbb{E}\left[\tau\right]}{2}\mathbb{E}\left[\left(\mu-\mu_0\right)^2\right] + \frac{N+1+2a_0-2a_N}{2}\mathbb{E}\left[\log\left(\tau\right)\right] \\ + & \frac{\lambda_N}{2}\mathbb{E}\left[\left(\mu-\mu_N\right)^2\right] - (b_0-b_N)\mathbb{E}\left[\tau\right] - \frac{1}{2}\log\left(\lambda_N\right) - a_N\log\left(b_N\right) + \log\left(\Gamma(a_N)\right) \\ + & \ \textit{constants} \end{aligned}$$

Therefore, we need to compute the expectations with respect to $q(\mu,\tau)$ in order to complete the ELBO.

1. Computing $\mathbb{E}\left[\,\log(\tau)\,\right]$

   Note. Proof taken from here

   From the definition of expectation, we have

$$
\begin{aligned}
\mathbb{E}\left[\,\log(\tau)\,\right] &= \int q(\tau)\log(\tau)\ \mathrm{d}\tau \\
&= \int \frac{(b_N)^{a_N}}{\Gamma(a_N)}\tau^{a_N-1}e^{-b_N\tau}\log(\tau)\ \mathrm{d}\tau \\
&= \int \frac{(b_N)^{a_N}}{\Gamma(a_N)}\tau^{a_N-1}e^{-b_N\tau}\log(b_N\tau)\ \mathrm{d}\tau - \log(b_N) \\
&= \int \frac{1}{\Gamma(a_N)}\tau^{a_N-1}e^{-\tau}\log(\tau)\ \mathrm{d}\tau - \log(b_N) \\
&= \frac{1}{\Gamma(a_N)}\frac{\mathrm{d}}{\mathrm{d}a_N}\int \tau^{a_N-1}e^{\tau}\ \mathrm{d}\tau - \log(b_N) \\
&= \frac{\Gamma'(a_N)}{\Gamma(a_N)} - \log(b_N) \\
&= \psi(a_N) - \log(b_N)
\end{aligned}
$$

   where $\psi(.)$ is the digamma function.

2. Computing $\mathbb{E}\left[\,\tau\,\right]$

   Since this is simply the first moment for a random variable sampled from a Gamma distribution, we can directly write this expectation as

$$
\mathbb{E}\left[\,\tau\,\right] = \frac{a_N}{b_N}
$$

3. Computing $\mathbb{E}\left[\,(\mu-\theta)^2\,\right]$

   Using properties of Gaussian, we can write,

$$
\begin{aligned}
\mathbb{E}\left[(\mu-\theta)^2\right] &= \mathbb{E}\left[\mu^2 - 2\theta\mu + \theta^2\right] \\
&= \frac{1}{\lambda_N} + \mu_N^2 - 2\theta\mu_N + \theta^2 \\
&= \frac{1}{\lambda_N} + (\mu_N - \theta)^2
\end{aligned}
$$

Hence, we can now compute the ELBO in terms of $\{\mu_N, \lambda_N, a_N, b_N\}$ as follows

$$
\begin{aligned}
\mathcal{L}(q) = &- \frac{a_N}{2b_N}\left(N\frac{1}{\lambda_N} + \sum_{n=1}^{N}(x_n - \mu_N)^2\right) \\
&- \frac{\lambda_0 a_N}{2b_N}\left(\frac{1}{\lambda_N} + (\mu_N - \mu_0)^2\right) \\
&+ \frac{N + 1 + 2a_0 - 2a_N}{2}\left(\psi(a_N) - \log(b_N)\right) \\
&- (b_0 - b_N)\frac{a_N}{b_N} - \frac{1}{2}\log(\lambda_N) - a_N\log(b_N) \\
&+ \log(\Gamma(a_N)) + \mathit{constants}
\end{aligned}
$$

Since we assume the Mean-Field Assumption holds, we estimate the parameters iteratively, *i.e.* we first assume some value of $\{\mu_N, \lambda_N\}$ and update $\{a_N, b_N\}$ and vice-versa.

2.1. Variational Inference

2.1.1. Update of $q(\tau)$

Since we are maximizing w.r.t. to a_N and b_N, therefore we have

$$\begin{aligned}
0 &= \frac{\delta\left(\mathcal{L}(a_N, b_N, \mu'_N, \lambda'_N)\right)}{\delta\left(a_N\right)} \\
&= -\frac{1}{2b_N}\left(\frac{N+\lambda_0}{\lambda'_N} + \sum_{n=1}^{N}\left(x_n - \mu'_N\right)^2 + \lambda_0\left(\mu'_N - \mu_0\right)^2 + 2b_0\right) \\
&\quad + \frac{N+1+2a_0-2a_N}{2}\psi'(a_N) + 1
\end{aligned}$$

$$\implies \frac{N+1+2a_0-2a_N}{2}\psi'(a_N) = \frac{\alpha}{2b_N} \tag{3}$$

where

$$\alpha = \frac{N+\lambda_0}{\lambda'_N} + \sum_{n=1}^{N}\left(x_n - \mu'_N\right)^2 + \lambda_0\left(\mu'_N - \mu_0\right)^2 + 2b_0 - 2b_N$$

Similarly for b_N

$$\begin{aligned}
0 &= \frac{\delta\left(\mathcal{L}(a_N, b_N, \mu'_N, \lambda'_N)\right)}{\delta\left(b_N\right)} \\
&= \frac{a_N\alpha}{2b_N^2} - \frac{N+1+2a_0-2a_N}{2b_N}
\end{aligned}$$

$$\implies \frac{a_N\alpha}{2b_N} = N+1+2a_0-2a_N \tag{4}$$

Since $\psi'(a_N) > 1$, using equations 3 and 4, we get

$$\alpha = 0 = N+1+2a_0-2a_N$$

Hence, we have the updates for a_N and b_N as follows

$$\begin{aligned}
a_N &= a_0 + \frac{N+1}{2} \\
b_N &= b_0 + \frac{N+\lambda_0}{2\lambda'_N} + \frac{1}{2}\sum_{n=1}^{N}\left(x_n - \mu'_N\right)^2 + \frac{\lambda_0}{2}\left(\mu'_N - \mu_0\right)^2
\end{aligned}$$

2.1.2. Update of $q(\mu)$

Similar to previous section, we find the updates of μ_N and λ_N by equating the respective partial derivatives to 0.

$$\begin{aligned}
0 &= \frac{\delta\left(\mathcal{L}(a'_N, b'_N, \mu_N, \lambda_N)\right)}{\delta\left(\mu_N\right)} \\
&= \frac{a'_N}{2b'_N}\sum_{n=1}^{N}(x_n - \mu_N) + \frac{\lambda_0 a'_N}{2b'_N}(\mu_0 - \mu_N) \\
\implies \mu_N &= \frac{\sum_{n=1}^{N} x_n + \lambda_0\mu_0}{N + \lambda_0}
\end{aligned}$$

$$\begin{aligned}
0 &= \frac{\delta\left(\mathcal{L}(a'_N, b'_N, \mu_N, \lambda_N)\right)}{\delta\left(\lambda_N\right)} \\
&= \frac{N a'_N}{2b'_N\lambda_N^2} + \frac{\lambda_0 a'_N}{2b'_N\lambda_N^2} - \frac{1}{2\lambda_N} \\
\implies \lambda_N &= \frac{a'_N}{b'_N}(N + \lambda_0)
\end{aligned}$$

$$\begin{aligned}
\mu_N &= \frac{\sum_{n=1}^{N} x_n + \lambda_0\mu_0}{N + \lambda_0} \\
\lambda_N &= \frac{a'_N}{b'_N}(N + \lambda_0)
\end{aligned}$$

We can now write the final Variation Inference algorithm, which is given in algorithm 2.

Algorithm 2: Variation Inference for Univariate Gaussian

```
Input: Input data $\mathbf{x}$
Output: Approximate Posterior $q(\mu, \tau)$
Steps:
  1. Initialize some values of $a_N$ and $b_N$ as $a_N^{(0)}$ and $b_N^0$
  2. Compute $\overline{x} = \frac{1}{N} \sum_{n=1}^{N} x_n$
  3. For $t = 0, 1 \ldots T - 1$
     (a) Update the parameters of $q(\mu)$ as
           $\mu_N^{(t+1)} = \frac{N\overline{x} + \lambda_0 \mu_0}{N + \lambda_0}$
           $\lambda_N^{(t+1)} = \frac{a_N^{(t)}}{b_N^t} (N + \lambda_0)$
     (b) Update the parameters of $q(\tau)$ as
           $a_N^{(t+1)} = a_0 + \frac{N+1}{2}$
           $b_N^{(t+1)} = b_0 + \frac{N + \lambda_0}{2\lambda_N^{(t+1)}} + \frac{1}{2} \sum_{n=1}^{N} \left( x_n - \mu_N^{(t+1)} \right)^2 + \frac{\lambda_0}{2} \left( \mu_0 - \mu_N^{(t+1)} \right)^2$
  4. Return $q(\mu, \tau) = q(\mu) q(\tau) = \mathcal{N}\left( \mu \,|\, \mu_N^{(T)}, {\lambda_N^{(T)}}^{-1} \right) \text{Gamma}\left( \tau \,\middle|\, a_N^{(T)}, b_N^{(T)} \right)$
```

Question 3

Black Box Variational Inference extends the idea of Variational Inference **??** which aims at approximating a "difficult" distribution using a family of simpler distributions, that is closest to the distribution in hand.

Variational Inference, although provides a novel approach to approximating difficult to compute distributions, suffers from intractability in generic problems. That is, for specific classes of models where the conditional distributions have a convenient form, it is possible to do Variational Inference using Alternating Maximization, however, for cases where this is not true, model-specific approaches are required to obtain a closed form solution.

Black Box Variational Inference aims to tackle this problem by rewriting the gradient of the optimization objective as an expectation and approximate the gradient using Monte Carlo Estimation techniques. The objective is therefore reduced to optimizing an objective, therefore allowing us to extend the optimization to Stochastic Methods.

3.1. Black Box Variational Inference

Similar to vanilla Variational Inference, the objective to maximize is the Evidence Lower BOund (ELBO), which is given as

$$\mathcal{L}(\boldsymbol{\lambda}) \quad \triangleq \quad \underset{q_{\boldsymbol{\lambda}}}{\mathbb{E}} \left[\log p(\mathbf{X}, \mathbf{Z}) - \log q(\mathbf{Z}) \right]$$

where $p(\mathbf{Z} \,|\, \mathbf{X})$ is the distribution to be approximated and $q_{\boldsymbol{\lambda}}(\mathbf{Z})$ is the family of distributions which serve as the hypothesis space for the approximation distribution, with $\boldsymbol{\lambda}$ denoting the parameters for the hypothesis space.

The ELBO is optimized using Stochastic Gradient Ascent, with the gradient estimated as an expectation with respect to the proposal distribution, given as below

$$\nabla_{\boldsymbol{\lambda}} \mathcal{L}(\boldsymbol{\lambda}) \quad = \quad \underset{q_{\boldsymbol{\lambda}}}{\mathbb{E}} \left[\nabla_{\boldsymbol{\lambda}} \log q(\mathbf{Z} \,|\, \boldsymbol{\lambda}) \Big(\log p(\mathbf{X}, \mathbf{Z}) - \log q(\mathbf{Z} \,|\, \boldsymbol{\lambda}) \Big) \right]$$

The derivation of this equation was discussed in class, and therefore I do not present it here. We can now build an "unbiased" estimator of this gradient using Monte Carlo Estimation, which essentially requires us to sample from the proposal distribution and estimate the gradient as below

$$\nabla_{\boldsymbol{\lambda}} \mathcal{L}(\boldsymbol{\lambda}) \quad \approx \quad \frac{1}{\mathrm{S}} \sum_{\mathbf{Z}_s \in \mathcal{S}} \nabla_{\boldsymbol{\lambda}} \log q(\mathbf{Z}_s \,|\, \boldsymbol{\lambda}) \Big(\log p(\mathbf{X}, \mathbf{Z}_s) - \log q(\mathbf{Z}_s \,|\, \boldsymbol{\lambda}) \Big)$$

where $\mathrm{S} = |\, \mathcal{S} \,|$ and $\mathcal{S} \sim q_{\lambda}^{\mathrm{S}}$ is sampled i.i.d. from the proposal distribution.

The method outlined above depends only on the individual components of the model, however remain independent of the structure of the underlying model. This is useful as it allows to plug in different models without changing the inherent structure of the inference model.

3.2. Controlling the Variance

As noted by the authors, the variance of the Monte Carlo Estimator can be very large, therefore the step size would have to be kept very small in order to guarantee convergence, therefore increasing the number of times steps to be taken to converge. The authors, therefore, give two techniques in order to control the variance of the estimator, while keeping the expectation constant, which is what is required.

1. **Rao-Blackwellization** Rao-Blackwellization reduces the variance of a random variable by replacing it with its conditional expectation with respect to a subset of the variables. The authors Rao-Blackwellize the estimator for each component without needing to compute model-specific integrals.

   Rao-Blackwellization, in its simplest form states that $\mathbb{E}\left[\, J(X,Y) \,\right] = \mathbb{E}\left[\, \widehat{J}(X) \,\right]$ where $\widehat{J}(X) = \mathbb{E}\left[\, J(X,Y) \,\middle|\, X \,\right]$. Therefore

$$\mathrm{Var}\left(\widehat{J}(X)\right) = \mathrm{Var}\left(J(X,Y)\right) - \mathbb{E}\left[\left(J(X,Y) - \widehat{J}(X)\right)^2\right]$$

   Hence, $\widehat{J}(X)$ has a lower variance than $J(X,Y)$ and is therefore a lower variance estimator than $J(X,Y)$. Applying this to the problem of estimating the ELBO derivative, where if we consider Mean-Field Assumption, then we have

$$q(\mathbf{Z} \,|\, \boldsymbol{\lambda}) \quad = \quad \prod_{m=1}^{M} q(\mathbf{Z}_m \,|\, \boldsymbol{\lambda}_m)$$

   We can therefore write the gradient with respect to $\boldsymbol{\lambda}_m$ as an iterated conditional expectation as

$$\nabla_{\boldsymbol{\lambda}_m} \mathcal{L} \quad = \quad \mathop{\mathbb{E}}_{q_{\boldsymbol{\lambda}_m}} \left[\nabla_{\boldsymbol{\lambda}_m} \log q(\mathbf{Z}_m \,|\, \boldsymbol{\lambda}_m) \left(\log p_m(\mathbf{X}, \mathbf{Z}_m) - \log q(\mathbf{Z}_m \,|\, \boldsymbol{\lambda}_m) \right) \right]$$

   Therefore, it is possible to construct a Monte Carlo Estimator for computing the gradient with respect to each paramter set $\boldsymbol{\lambda}_m$ under the mean-field assumption as

$$\nabla_{\boldsymbol{\lambda}_m} \mathcal{L} \quad \approx \quad \frac{1}{\mathrm{S}} \sum_{\mathbf{Z}_s \in \mathcal{S}} \nabla_{\boldsymbol{\lambda}_m} \log q(\mathbf{Z}_s \,|\, \boldsymbol{\lambda}_m) \left(\log p_m(\mathbf{X}, \mathbf{Z}_s) - \log q(\mathbf{Z}_s \,|\, \boldsymbol{\lambda}_m) \right)$$

   where $\mathrm{S} = |\, \mathcal{S} \,|$ and $\mathcal{S} \sim q^{\mathrm{S}}(\mathbf{Z}_m \,|\, \boldsymbol{\lambda}_m)$

2. **Control Variates** A control variate is a family of functions with equivalent expectation, *i.e.* for a function $f(.)$, the control variate of this function could be written as

$$\widehat{f}(z)$$

   This can be used in BBVI by choosing a, given h, such that the variance of $\widehat{f}(z)$ is the lowest. The authors note that for a given h, the a such that $\widehat{f}(z)$ has minimum variance is as follows

$$a^* \quad = \quad \frac{\mathrm{Cov}\,(f, h)}{\mathrm{Var}\,(h)}$$

   In case of BBVI, the functions f and h is chosen for each component (in case of Mean-Field VI) as follows

$$\begin{aligned} f_m(\mathbf{Z}) \quad &= \quad \nabla_{\boldsymbol{\lambda}_m} \log q(\mathbf{Z}_m \,|\, \boldsymbol{\lambda}_m) \left(\log p(\mathbf{X}, \mathbf{Z}_m) - \log q(\mathbf{Z}_m \,|\, \boldsymbol{\lambda}_m) \right) \\ h_m(\mathbf{Z}) \quad &= \quad \nabla_{\boldsymbol{\lambda}_m} \log q(\mathbf{Z}_m \,|\, \boldsymbol{\lambda}_m) \end{aligned}$$

   and the value of a^* is approximated as

$$\widehat{a}_m^* = \frac{\sum_{d=1}^{D_m} \hat{\mathrm{Cov}}\left(f_m^d, h_m^d\right)}{\hat{\mathrm{Var}}\left(h_m^d\right)} \tag{5}$$

   where the covariance and variance is computed using the Monte Carlo samples itself, and f_m^d and h_m^d denote d$^{\mathrm{th}}$ dimension of the functions respectively, corresponding to each dimension of the parameter $\boldsymbol{\lambda}_m$.

The authors have also shown some extensions to the model using adaptive learning rates with Adagrad, as well as stochastic optimization using a sample of the data points in order to improve convergence. The overall method of BBVI provides good convergence and allows models to be learned without applying model specific approximations in order to achieve a closed form solution.

Question 4

4.1. Generative Story

$$
\begin{aligned}
\pi_k &\sim \text{Beta}\left(\alpha/K, 1\right) && \text{for } k = 1 \ldots K \\
z_{nk} &\sim \text{Bernoulli}\left(\pi_k\right) && \text{for } n = 1 \ldots N,\ k = 1 \ldots K \\
\mathbf{w}_k &\sim \mathcal{N}\left(\mathbf{0}, \sigma_w^2 \mathbf{I}_D\right) && \text{for } k = 1 \ldots K \\
\mathbf{x}_n &\sim \mathcal{N}\left(\mathbf{W}\mathbf{z}_n, \sigma_x^2 \mathbf{I}_D\right) && \text{for } n = 1 \ldots N
\end{aligned}
$$

4.2. Complete Data Likelihood

We represent the $\boldsymbol{\Theta}$ as the tuple of all latent variables, *i.e.* $(\mathbf{W}, \mathbf{Z}, \boldsymbol{\pi})$, and $\boldsymbol{\theta}$ as the tuple of all hyperparameters (known), *i.e.* $\left(\sigma_w^2, \sigma_x^2, \alpha\right)$.

We can write the Complete Data Likelihood as

$$
\begin{aligned}
\mathbb{P}\left[\mathbf{X}, \boldsymbol{\Theta} \,|\, \boldsymbol{\theta}\right] &= \mathbb{P}\left[\mathbf{X} \,|\, \mathbf{W}, \mathbf{Z}, \sigma_x^2\right] \mathbb{P}\left[\mathbf{W} \,|\, \sigma_w^2\right] \mathbb{P}\left[\mathbf{Z} \,|\, \boldsymbol{\pi}\right] \mathbb{P}\left[\boldsymbol{\pi} \,|\, \alpha\right] \\
\implies \log \mathbb{P}\left[\mathbf{X}, \boldsymbol{\Theta} \,|\, \boldsymbol{\theta}\right] &= \log \mathbb{P}\left[\mathbf{X} \,|\, \mathbf{W}, \mathbf{Z}, \sigma_x^2\right] + \log \mathbb{P}\left[\mathbf{W} \,|\, \sigma_w^2\right] + \log \mathbb{P}\left[\mathbf{Z} \,|\, \boldsymbol{\pi}\right] + \log \mathbb{P}\left[\boldsymbol{\pi} \,|\, \alpha\right]
\end{aligned}
$$

We can now easily compute the log-probabilities individually. Note that we can treat all the terms that are not dependent on $\boldsymbol{\Theta}$. Therefore

$$
\begin{aligned}
\log \mathbb{P}\left[\mathbf{X} \,|\, \mathbf{W}, \mathbf{Z}, \sigma_x^2\right] &= -\frac{1}{2\sigma_x^2} \sum_{n=1}^{N} \left(\mathbf{x}_n - \mathbf{W}\mathbf{z}_n\right)^{\mathrm{T}} \left(\mathbf{x}_n - \mathbf{W}\mathbf{z}_n\right) + \textit{constants} \\
\log \mathbb{P}\left[\mathbf{W} \,|\, \sigma_w^2\right] &= -\frac{1}{2\sigma_w^2} \sum_{k=1}^{K} {\mathbf{w}_k}^{\mathrm{T}} \mathbf{w}_k + \textit{constants} \\
\log \mathbb{P}\left[\mathbf{Z} \,|\, \boldsymbol{\pi}\right] &= \sum_{n=1}^{N} \log \mathbb{P}\left[\mathbf{z}_n \,|\, \boldsymbol{\pi}\right] \\
&= \sum_{n=1}^{N} \sum_{k=1}^{K} \left\{z_{nk} \log\left(\pi_k\right) + \left(1 - z_{nk}\right) \log\left(1 - \pi_k\right)\right\} \\
\log \mathbb{P}\left[\boldsymbol{\pi} \,|\, \alpha\right] &= \left(\frac{\alpha}{K} - 1\right) \sum_{k=1}^{K} \log\left(\pi_k\right) + \textit{constants}
\end{aligned}
$$

4.3. Variational Inference

We can now compute the updates for the latent variables, where the posterior approximation $q(\mathbf{W}, \mathbf{Z}, \boldsymbol{\pi})$ is assumed to folow the Mean-Field Assumption, *i.e.*

$$
q(\mathbf{W}, \mathbf{Z}, \boldsymbol{\pi}) = \prod_{k=1}^{K} \left\{ \prod_{n=1}^{N} q(z_{nk}) \cdot q(\mathbf{w}_k) q(\pi_k) \right\}
$$

We can therefore compute the update of each latent variable in an iterative fashion

4.3.1. Update of $q(\mathbf{Z})$

We do not actually update the value of $q(\mathbf{Z})$, but update $q(z_{nk})$ for each $n = 1 \dots N$, $k = 1 \dots K$. We know from the Mean-Feild Variational Inference updates

$$\begin{aligned}
\log q(z_{nk}) \quad &= \quad \underset{\mathbf{W}, \bar{z}_{nk}, \boldsymbol{\pi}}{\mathbb{E}} \left[\log \mathbb{P} \left[\mathbf{X}, \boldsymbol{\Theta} \,\middle|\, \boldsymbol{\theta} \right] \right] \\
&= \quad \mathbb{E} \left[\log \mathbb{P} \left[\mathbf{X} \,\middle|\, \mathbf{W}, \mathbf{Z}, \sigma_x^2 \right] \right] + \mathbb{E} \left[\log \mathbb{P} \left[\mathbf{W} \,\middle|\, \sigma_w^2 \right] \right] \\
&\quad\quad + \ \mathbb{E} \left[\log \mathbb{P} \left[\mathbf{Z} \,\middle|\, \boldsymbol{\pi} \right] \right] + \mathbb{E} \left[\log \mathbb{P} \left[\boldsymbol{\pi} \,\middle|\, \alpha \right] \right] \\
&= \quad -\frac{1}{2\sigma_x^2} \mathbb{E} \left[z_{nk}^2 \mathbf{w}_k{}^{\mathrm{T}} \mathbf{w}_k - 2 z_{nk} \mathbf{w}_k{}^{\mathrm{T}} \left(\mathbf{x}_n - \sum_{\substack{k'=1 \\ k' \neq k}}^{K} z_{nk'} \mathbf{w}_{k'} \right) \right] \\
&\quad\quad + \ \mathbb{E} \left[z_{nk} \log (\pi_k) + (1 - z_{nk}) \log (1 - \pi_k) \right] + \mathit{constants}
\end{aligned}$$

Note that since z_{nk} is a Bernoulli variable, we know $z_{nk}^2 = z_{nk}$. Due to mean-field assumption, we can write this as

$$\begin{aligned}
\log q(z_{nk}) \quad &= \quad z_{nk} \Bigg\{ \mathbb{E} \left[\log (\pi_k) \right] - \mathbb{E} \left[\log (1 - \pi_k) \right] - \frac{1}{2\sigma_x^2} \mathbb{E} \left[\mathbf{w}_k{}^{\mathrm{T}} \mathbf{w}_k \right] \\
&\quad\quad + \ \frac{1}{\sigma_x^2} \mathbb{E} \left[\mathbf{w}_k{}^{\mathrm{T}} \right] \left(\mathbf{x}_n - \sum_{\substack{k'=1 \\ k' \neq k}}^{K} \mathbb{E} \left[z_{nk'} \right] \mathbb{E} \left[\mathbf{w}_{k'} \right] \right) \Bigg\} + \mathit{constants}
\end{aligned}$$

Hence, $q(z_{nk})$ is a Bernoulli distribution, given as follows

$$q(z_{nk}) \quad = \quad \text{Bernoulli} \left(z_{nk} \,\middle|\, \widehat{\pi}_{nk} \right) \tag{6}$$

where

$$\begin{aligned}
\widehat{\pi}_{nk} \quad &= \quad \frac{\exp (\gamma_{nk})}{1 + \exp (\gamma_{nk})} \\
\gamma_{nk} \quad &= \quad \mathbb{E} \left[\log (\pi_k) \right] - \mathbb{E} \left[\log (1 - \pi_k) \right] - \frac{1}{2\sigma_x^2} \mathbb{E} \left[\mathbf{w}_k{}^{\mathrm{T}} \mathbf{w}_k \right] \qquad (7) \\
&\quad\quad + \ \frac{1}{\sigma_x^2} \mathbb{E} \left[\mathbf{w}_k{}^{\mathrm{T}} \right] \left(\mathbf{x}_n - \sum_{\substack{k'=1 \\ k' \neq k}}^{K} \mathbb{E} \left[z_{nk'} \right] \mathbb{E} \left[\mathbf{w}_{k'} \right] \right)
\end{aligned}$$

4.3.2. Update of $q(\mathrm{W})$

$$\begin{aligned}
\log q(\mathbf{w}_k) &= \underset{\overline{\mathbf{w}}_k, \mathbf{Z}, \boldsymbol{\pi}}{\mathbb{E}} \left[\log \mathbb{P}\left[\mathbf{X}, \boldsymbol{\Theta} \,\middle|\, \boldsymbol{\theta} \right] \right] \\
&= \mathbb{E}\left[\frac{1}{\sigma_x^2} {\mathbf{w}_k}^{\mathrm{T}} \sum_{n=1}^{N} z_{nk} \left(\mathbf{x}_n - \sum_{\substack{k'=1 \\ k' \neq k}}^{K} z_{nk'} \mathbf{w}_{k'} \right) - \left(\frac{1}{2\sigma_x^2} \sum_{n=1}^{N} z_{nk}^2 + \frac{1}{2\sigma_w^2} \right) {\mathbf{w}_k}^{\mathrm{T}} \mathbf{w}_k \right] \\
&\quad + \; constants
\end{aligned}$$

Since we can represent this as a gaussian, we say

$$q(\mathbf{w}_k) \quad = \quad \mathcal{N}\left(\mathbf{w}_k \,\middle|\, \widehat{\boldsymbol{\mu}}_k, \widehat{\boldsymbol{\Sigma}}_k \right) \tag{8}$$

where

$$\begin{aligned}
\widehat{\boldsymbol{\Sigma}}_k &= \left(\frac{1}{\sigma_x^2} \sum_{n=1}^{N} \mathbb{E}\left[\, z_{nk} \,\right] + \frac{1}{\sigma_w^2} \right)^{-1} \mathbf{I}_D \\
\widehat{\boldsymbol{\mu}}_k &= \frac{1}{\sigma_x^2} \cdot \widehat{\boldsymbol{\Sigma}}_k \left\{ \sum_{n=1}^{N} \mathbb{E}\left[\, z_{nk} \,\right] \left(\mathbf{x}_n - \sum_{\substack{k'=1 \\ k' \neq k}}^{K} \mathbb{E}\left[\, z_{nk'} \,\right] \mathbb{E}\left[\, \mathbf{w}_{k'} \,\right] \right) \right\}
\end{aligned}$$

4.3.3. Update of $q(\pi)$

$$\begin{aligned}
\log q(\pi_k) &= \underset{\mathbf{W}, \mathbf{Z}, \overline{\pi}_k}{\mathbb{E}} \left[\log \mathbb{P}\left[\mathbf{X}, \boldsymbol{\Theta} \,\middle|\, \boldsymbol{\theta} \right] \right] \\
&= \mathbb{E}\left[\sum_{n=1}^{N} z_{nk} \log\left(\pi_k\right) + \left(1 - z_{nk}\right) \log\left(1 - \pi_k\right) + \left(\frac{\alpha}{K} - 1 \right) \log\left(\pi_k\right) \right] + constants
\end{aligned}$$

Clearly, $q(\pi_k)$ is a Beta distribution. Hence, we can write

$$q(\pi_k) \quad = \quad \mathrm{Beta}\left(\widehat{\alpha}_k, \widehat{\beta}_k \right) \tag{9}$$

where

$$\begin{aligned}
\widehat{\alpha}_k &= \frac{\alpha}{K} + \sum_{n=1}^{N} \mathbb{E}\left[\, z_{nk} \,\right] \\
\widehat{\beta}_k &= 1 + N - \sum_{n=1}^{N} \mathbb{E}\left[\, z_{nk} \,\right]
\end{aligned}$$

Now, we can write the expectation terms for all variables.

4.3.4. Expectations

$$
\begin{aligned}
\mathbb{E}\left[\, z_n k \,\right] &= \widehat{\pi}_{nk} \\
\mathbb{E}\left[\, \mathbf{w}_k \,\right] &= \widehat{\boldsymbol{\mu}}_k \\
\mathbb{E}\left[\, {\mathbf{w}_k}^{\mathrm{T}} \mathbf{w}_k \,\right] &= \mathbb{E}\left[\, \mathrm{Tr}\left(\mathbf{w}_k {\mathbf{w}_k}^{\mathrm{T}} \right) \right] \\
&= \mathrm{Tr}\left(\widehat{\boldsymbol{\Sigma}}_k + \widehat{\boldsymbol{\mu}}_k \widehat{\boldsymbol{\mu}}_k^{\mathrm{T}} \right) \\
&= \mathrm{Tr}\left(\widehat{\boldsymbol{\Sigma}}_k \right) + \widehat{\boldsymbol{\mu}}_k^{\mathrm{T}} \widehat{\boldsymbol{\mu}}_k \\
\mathbb{E}\left[\, \log\left(\pi_k\right) \right] &= \psi\left(\widehat{\alpha}\right) - \psi\left(\widehat{\alpha}_k + \widehat{\beta}_k\right) \\
\mathbb{E}\left[\, \log\left(1 - \pi_k\right) \right] &= \psi\left(\widehat{\beta}\right) - \psi\left(\widehat{\alpha}_k + \widehat{\beta}_k\right)
\end{aligned}
$$

I have referred to StackOverflow for the expectation of $\log\left(\pi_k\right)$. For the last equality, we just need to realize that if π is a beta distribution with parameters α and β, then $1 - \pi$ is also a beta distribution with parameters β and α.

I have given the final and complete algorithm in Algorithm 3.

```
**Algorithm 3**: Variation Inference

**Input:** Input data $\mathbf{X}$ and hyper-paramters $\alpha$, $\sigma_x^2$ and $\sigma_w^2$
**Output:** Approximate Posterior $q(\mathbf{W}, \mathbf{Z}, \boldsymbol{\pi})$
**Steps:**
  1. Initialize $q(z_{nk}) = \text{Bernoulli}\left(z_{nk} \,\middle|\, \widehat{\pi}_{nk}^{(0)}\right)$ for all $n \in [\,N\,]$ and $k \in [\,K\,]$
  2. For $t = 0, 1 \ldots T-1$
     (a) Update $q(\boldsymbol{\pi})$ as
         $\forall\, k \in [\,K\,], \quad \widehat{\alpha}_k^{(t+1)} = \frac{\alpha}{K} + \sum_{n=1}^{N} \widehat{\pi}_{nk}^{(t)}$
         $\widehat{\beta}_k^{(t+1)} = N + 1 - \sum_{n=1}^{N} \widehat{\pi}_k^{(t)}$
         $q(\pi_k) = \text{Beta}\left(\widehat{\alpha}_k^{(t+1)}, \widehat{\beta}_k^{(t+1)}\right)$
     (b) Update $q(\mathbf{W})$ as
         i. For $n = 1 \ldots N$, compute
            $\boldsymbol{\delta}_n = \sum_{k=1}^{K} \widehat{\pi}_{nk}^{(t)} \cdot \widehat{\boldsymbol{\mu}}_k^{(t)}$
         ii. For $k = 1 \ldots K$, do
            $\forall\, n \in [\,N\,], \quad \boldsymbol{\delta}_n = \boldsymbol{\delta}_n - \widehat{\pi}_{nk}^{(t)} \cdot \widehat{\boldsymbol{\mu}}_k^{(t)}$
            $\widehat{\boldsymbol{\Sigma}}_k^{(t+1)} = \left(\frac{1}{\sigma_x^2} \sum_{n=1}^{N} \widehat{\pi}_{nk}^{(t)} + \frac{1}{\sigma_w^2}\right)^{-1} \mathbf{I}_D$
            $\widehat{\boldsymbol{\mu}}_k^{(t+1)} = \frac{1}{\sigma_x^2} \cdot \widehat{\boldsymbol{\Sigma}}_k^{(t+1)} \left\{ \sum_{n=1}^{N} \widehat{\pi}_{nk}^{(t)} \left(\mathbf{x}_n - \boldsymbol{\delta}_n\right) \right\}$
            $q(\mathbf{w}_k) = \mathcal{N}\left(\widehat{\boldsymbol{\mu}}_k^{(t+1)}, \widehat{\boldsymbol{\Sigma}}_k^{(t+1)}\right)$
            $\forall\, n \in [\,N\,], \quad \boldsymbol{\delta}_n = \boldsymbol{\delta}_n + \widehat{\pi}_{nk}^{(t)} \cdot \widehat{\boldsymbol{\mu}}_k^{(t+1)}$
     (c) Update $q(\mathbf{Z})$ as
         $\forall\, n \in [\,N\,], k \in [\,K\,], \quad \widehat{\pi}_{nk} = \frac{\exp\left(\gamma_{nk}\right)}{1 + \exp\left(\gamma_{nk}\right)}$
         $q(z_{nk}) = \text{Bernoulli}\left(\widehat{\pi}_{nk}^{(t+1)}\right)$
         where $\gamma_{nk}$ is computed as in equation 7
  3. Return $q(\mu, \tau) = q(\mu) q(\tau) = \mathcal{N}\left(\mu \,\middle|\, \mu_N^{(T)}, {\lambda_N^{(T)}}^{-1}\right) \text{Gamma}\left(\tau \,\middle|\, a_N^{(T)}, b_N^{(T)}\right)$
```