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

Question 1

Part A

The above prior facilitates sparse learning for the weight parameters.

Part B

Posterior over Latent Variables

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

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

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

$$\textbf{Likelihood:} \qquad \mathbb{P}\left[\mathbf{y} \mid \mathbf{X}, \mathbf{w}, \sigma^2\right] \quad = \quad \mathcal{N}\left(\mathbf{y} \mid \mathbf{X}\mathbf{w}, \sigma^2 \mathbf{I}_N\right)$$

$$\textbf{Prior:} \qquad \mathbb{P}\left[\mathbf{w} \mid \sigma^2, \boldsymbol{\gamma}\right] \quad = \quad \mathcal{N}\left(\mathbf{w} \mid \mathbf{0}, \sigma^2 \mathbf{K}\right)$$

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

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

where

$$\boldsymbol{\Sigma}_w \quad = \quad \sigma^2 \left(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \mathbf{K}^{-1}\right)^{-1}$$

$$\boldsymbol{\mu}_w \quad = \quad \frac{1}{\sigma^2} \boldsymbol{\Sigma}_w \mathbf{X}^{\mathrm{T}}\mathbf{y}$$

Expectation of CLL

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

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

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

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

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

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

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

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

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

Note. We might want to compute the MAP by computing the expectation of the log-posterior over the parameters $\left(\mathbb{P}\left[\sigma^2, \boldsymbol{\gamma}, \theta \mid \mathbf{y}, \mathbf{w}, \mathbf{X} \right] \right)$, 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

$$\mathbb{P}\left[\sigma^2, \boldsymbol{\gamma}, \theta\right] \;=\; \mathbb{P}\left[\sigma^2\right] \prod_{d=1}^{D} \left\{ \mathbb{P}\left[\gamma_d \mid \theta\right] \right\} \mathbb{P}\left[\theta\right]$$

$$\implies \log \mathbb{P}\left[\sigma^2, \boldsymbol{\gamma}, \theta\right] \;=\; \log \mathbb{P}\left[\sigma^2\right] + \sum_{d=1}^{D} \log \mathbb{P}\left[\gamma_d \mid \theta\right] + \log \mathbb{P}\left[\theta\right]$$

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

$$0 \;=\; \frac{\delta\left(\mathbb{E}\left[\operatorname{CLL}\right] + \log \mathbb{P}\left[\sigma^2, \boldsymbol{\gamma}, \theta\right]\right)}{\delta\left(\sigma^2\right)}$$

$$=$$

<++>

<++>

<++>

Question 2

Using the Mean-Field Assumption, we have

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

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

$$q(\mu) \quad = \quad \mathcal{N}\left(\mu \mid \mu_N, \lambda_N^{-1}\right)$$

$$q(\tau) \quad = \quad \text{Gamma}\left(\tau \mid a_N, b_N\right)$$

Now from definition, we can write the ELBO as follows

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

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

$$\mathbb{P}\left[\mathbf{x} \mid \mu, \tau\right] \quad = \quad \prod_{n=1}^{N} \mathcal{N}\left(x_n \mid \mu, \tau^{-1}\right)$$

$$\mathbb{P}\left[\mu, \tau\right] \quad = \quad \mathcal{N}\left(\mu \mid \mu_0, (\lambda_0 \tau)^{-1}\right) \text{Gamma}\left(\tau \mid a_0, b_0\right)$$

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

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

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

We can simplify this as follows

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

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

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

   Note. Proof taken from here

   From the definition of expectation, we have

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

   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

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

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.

Update of $q(\tau)$

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

<++>

<++>

<++>