Homework 2

JD Gallego Posada, *University of Amsterdam* 21/09/2016

1 MAP Solution for Linear Regression

1. Write down the likelihood $p(\mathbf{t}|\mathbf{w})$ using a) a product over N and b) in vector/matrix form.

Define $\mathbf{t} = \begin{bmatrix} t_1 \\ \vdots \\ t_N \end{bmatrix}$, $\mathbf{w} = \begin{bmatrix} w_0 \\ \vdots \\ w_{M-1} \end{bmatrix}$, $\mathbf{\Sigma} = \beta^{-1}\mathbf{I}$, and $\mathbf{\Phi} = \begin{bmatrix} \rightharpoonup \boldsymbol{\phi}_1^T \leftharpoonup \\ \vdots \\ \rightharpoonup \boldsymbol{\phi}_N^T \leftharpoonup \end{bmatrix}$.

$$
\begin{aligned}
p(\mathbf{t}|\mathbf{\Phi}, \mathbf{w}, \beta) &= \prod_{i=1}^{N} p(t_i|\phi_i, \mathbf{w}, \beta) = \prod_{i=1}^{N} \mathcal{N}(t_i|\mathbf{w}^T\boldsymbol{\phi}_i, \beta^{-1}) \\
&= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}}\sqrt{\beta}\exp\left(-\frac{\beta}{2}(t_i - \mathbf{w}^T\boldsymbol{\phi}_i)^2\right) \\
&= \left(\frac{\sqrt{\beta}}{\sqrt{2\pi}}\right)^N \exp\left(-\frac{\beta}{2}\sum_{i=1}^{N}\left(t_i - \mathbf{w}^T\boldsymbol{\phi}_i\right)^2\right)
\end{aligned}
$$

Note that $(\mathbf{t} - \mathbf{\Phi}\mathbf{w})^T(\mathbf{t} - \mathbf{\Phi}\mathbf{w}) = \|\mathbf{t} - \mathbf{\Phi}\mathbf{w}\|^2 = \sum_{i=1}^{N}\left(t_i - \mathbf{w}^T\boldsymbol{\phi}_i\right)^2 = \sum_{i=1}^{N}\left(t_i - \boldsymbol{\phi}_i^T\mathbf{w}\right)^2$. Thus,

$$
p(\mathbf{t}|\mathbf{\Phi}, \mathbf{w}, \beta) = \frac{1}{(2\pi)^{N/2}}\frac{1}{|\mathbf{\Sigma}|^{1/2}}\exp\left(-\frac{1}{2}(\mathbf{t} - \mathbf{\Phi}\mathbf{w})^T\mathbf{\Sigma}^{-1}(\mathbf{t} - \mathbf{\Phi}\mathbf{w})\right) = \mathcal{N}(\mathbf{t}|\mathbf{\Phi}\mathbf{w}, \mathbf{\Sigma})
$$

2. Write down the prior $p(\mathbf{w}|\alpha)$. Compute its log.

$$
p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \mathbf{I}/\alpha) = \frac{\alpha^{M/2}}{(2\pi)^{M/2}}\exp\left(-\frac{\alpha}{2}\mathbf{w}^T\mathbf{w}\right) = \frac{\alpha^{M/2}}{(2\pi)^{M/2}}\exp\left(-\frac{\alpha}{2}\sum_{i=0}^{M-1}w_i^2\right)
$$

$$
\log p(\mathbf{w}|\alpha) = \frac{M}{2}\log(\alpha) - \frac{M}{2}\log(2\pi) - \frac{\alpha}{2}\mathbf{w}^T\mathbf{w} = \frac{M}{2}\log(\alpha) - \frac{M}{2}\log(2\pi) - \frac{\alpha}{2}\sum_{i=0}^{M-1}w_i^2
$$

3. Write down an expression for the posterior over $\mathbf{w}$.

$$
\begin{aligned}
p(\mathbf{w}|\mathbf{t}, \mathbf{\Phi}, \alpha, \beta) &= \frac{p(\mathbf{t}|\mathbf{\Phi}, \mathbf{w}, \beta)\, p(\mathbf{w}|\alpha)}{\displaystyle\int_{\mathbb{R}^M} p(\mathbf{t}|\mathbf{\Phi}, \mathbf{w}', \beta)\, p(\mathbf{w}'|\alpha)\, d\mathbf{w}'} \\
&= \frac{\mathcal{N}(\mathbf{t}|\mathbf{\Phi}\mathbf{w}, \mathbf{\Sigma})\,\mathcal{N}(\mathbf{w}|\mathbf{0}, \mathbf{I}/\alpha)}{\displaystyle\int_{\mathbb{R}^M}\mathcal{N}(\mathbf{t}|\mathbf{\Phi}\mathbf{w}'), \mathbf{\Sigma})\,\mathcal{N}(\mathbf{w}'|\mathbf{0}, \mathbf{I}/\alpha)\, d\mathbf{w}'}
\end{aligned}
$$

4. Compute the log-posterior, both for the a) and b) likelihood forms from above. Collect everything that does not depend on $\mathbf{w}$ into a constant C. What parts of the previous expression do not depend on $\mathbf{w}$? Why is finding the MAP much simpler than finding the full posterior distribution?

Note that the evidence in $p(\mathbf{w}|\mathbf{t}, \mathbf{\Phi}, \alpha, \beta)$ is not a function of $\mathbf{w}$, thus we can treat it as a constant I w.r.t. $\mathbf{w}$. This causes the computation of the MAP estimator $\mathbf{w}_{\text{MAP}}$ to be easier than the full posterior distribution, since taking partial derivative of the log-posterior w.r.t. $\mathbf{w}$ eliminates I, whose calculation is typically hard.

$$\log p(\mathbf{t}|\mathbf{\Phi}, \mathbf{w}, \beta) = \frac{N}{2}\log\left(\frac{\beta}{2\pi}\right) - \frac{\beta}{2}\sum_{i=1}^{N}\left(t_i - \mathbf{w}^T\boldsymbol{\phi}_i\right)^2 = \frac{N}{2}\log\left(\frac{\beta}{2\pi}\right) - \frac{\beta}{2}\left(\mathbf{t} - \mathbf{\Phi}\mathbf{w}\right)^T\left(\mathbf{t} - \mathbf{\Phi}\mathbf{w}\right)$$

$$\log p(\mathbf{w}|\alpha) = \frac{M}{2}\log(\alpha) - \frac{M}{2}\log(2\pi) - \frac{\alpha}{2}\sum_{i=0}^{M-1} w_i^2 = \frac{M}{2}\log(\alpha) - \frac{M}{2}\log(2\pi) - \frac{\alpha}{2}\mathbf{w}^T\mathbf{w}$$

$$\begin{aligned}
\log p(\mathbf{w}|\mathbf{t}, \mathbf{\Phi}, \alpha, \beta) &= \log p(\mathbf{t}|\mathbf{\Phi}, \mathbf{w}, \beta) + \log p(\mathbf{w}|\alpha) - \log I \\
&= -\frac{\alpha}{2}\mathbf{w}^T\mathbf{w} - \frac{\beta}{2}\left(\mathbf{t} - \mathbf{\Phi}\mathbf{w}\right)^T\left(\mathbf{t} - \mathbf{\Phi}\mathbf{w}\right) + C \\
&= -\frac{\alpha}{2}\sum_{i=0}^{M-1} w_i^2 - \frac{\beta}{2}\sum_{i=1}^{N}\left(t_i - \mathbf{w}^T\boldsymbol{\phi}_i\right)^2 + C
\end{aligned}$$

where $C = \frac{N}{2}\log\left(\frac{\beta}{2\pi}\right) + \frac{M}{2}\log(\alpha) - \frac{M}{2}\log(2\pi) - \log I$

5. Solve for $\mathbf{w}_{MAP}$ by a) taking the derivative of the log-posterior with respect to $\mathbf{w}$, b) setting it to 0, and c) solving for $\mathbf{w}$. Do this for both forms of likelihood.

Vector form

$$\begin{aligned}
\frac{\partial}{\partial \mathbf{w}}\log p(\mathbf{w}|\mathbf{t}, \mathbf{\Phi}, \alpha, \beta) = -\alpha\mathbf{w}^T + \beta\left(\mathbf{t} - \mathbf{\Phi}\mathbf{w}\right)^T\mathbf{\Phi} &= 0 \\
-\alpha\mathbf{w}^T + \beta\mathbf{t}^T\mathbf{\Phi} - \beta\mathbf{w}^T\mathbf{\Phi}^T\mathbf{\Phi} &= 0 \\
\mathbf{w}^T\left(\alpha\mathbf{I} + \beta\mathbf{\Phi}^T\mathbf{\Phi}\right) &= \beta\mathbf{t}^T\mathbf{\Phi} \\
\left(\alpha\mathbf{I} + \beta\mathbf{\Phi}^T\mathbf{\Phi}\right)\mathbf{w} &= \beta\mathbf{\Phi}^T\mathbf{t} \\
\mathbf{w}_{\text{MAP}} &= \beta\left(\alpha\mathbf{I} + \beta\mathbf{\Phi}^T\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^T\mathbf{t}
\end{aligned}$$

Scalar form

Note: $(\boldsymbol{\phi}_i)_j$ means the j-th component of vector $\boldsymbol{\phi}_i$.

$$\begin{aligned}
\frac{\partial}{\partial w_j}\log p(\mathbf{w}|\mathbf{t}, \mathbf{\Phi}, \alpha, \beta) = -\frac{\alpha}{2}\frac{\partial}{\partial w_j}\sum_{i=0}^{M-1} w_i^2 - \frac{\beta}{2}\frac{\partial}{\partial w_j}\sum_{i=1}^{N}\left(t_i - \sum_{k=0}^{M-1} w_k\left(\boldsymbol{\phi}_i\right)_k\right)^2 &= 0 \\
-\frac{\alpha}{2}2w_j - \frac{\beta}{2}\sum_{i=1}^{N}\frac{\partial}{\partial w_j}\left(t_i^2 - 2t_i\left(\sum_{k=0}^{M-1} w_k\left(\boldsymbol{\phi}_i\right)_k\right) + \left(\sum_{k=0}^{M-1} w_k\left(\boldsymbol{\phi}_i\right)_k\right)^2\right) &= 0 \\
-w_j\alpha - \frac{\beta}{2}\sum_{i=1}^{N}\left(-2t_i\left(\boldsymbol{\phi}_i\right)_j + 2\left(\sum_{k=0}^{M-1} w_k\left(\boldsymbol{\phi}_i\right)_k\right)\left(\boldsymbol{\phi}_i\right)_j\right) &= 0 \\
-w_j\alpha + \beta\sum_{i=1}^{N} t_i\left(\boldsymbol{\phi}_i\right)_j - \beta\sum_{i=1}^{N}\sum_{k=0}^{M-1} w_k\left(\boldsymbol{\phi}_i\right)_k\left(\boldsymbol{\phi}_i\right)_j &= 0 \\
w_j\alpha + \beta\sum_{k=0}^{M-1} w_k\sum_{i=1}^{N}\left(\boldsymbol{\phi}_i\right)_k\left(\boldsymbol{\phi}_i\right)_j &= \beta\sum_{i=1}^{N} t_i\left(\boldsymbol{\phi}_i\right)_j
\end{aligned}$$

Consider the expansion of equation $\left(\alpha\mathbf{I} + \beta\mathbf{\Phi}^T\mathbf{\Phi}\right)\mathbf{w} = \beta\mathbf{\Phi}^T\mathbf{t}$:

$$\begin{bmatrix} \alpha & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \alpha \end{bmatrix} \begin{bmatrix} w_0 \\ \vdots \\ w_{M-1} \end{bmatrix} + \beta \begin{bmatrix} (\boldsymbol{\phi}_1)_0 & & (\boldsymbol{\phi}_N)_0 \\ & \cdots & \\ (\boldsymbol{\phi}_1)_{M-1} & & (\boldsymbol{\phi}_N)_{M-1} \end{bmatrix} \begin{bmatrix} (\boldsymbol{\phi}_1)_0 & & (\boldsymbol{\phi}_1)_{M-1} \\ & \cdots & \\ (\boldsymbol{\phi}_N)_0 & & (\boldsymbol{\phi}_N)_{M-1} \end{bmatrix} \begin{bmatrix} w_0 \\ \vdots \\ w_{M-1} \end{bmatrix} =$$

$$\beta \begin{bmatrix} (\boldsymbol{\phi}_1)_0 & & (\boldsymbol{\phi}_N)_0 \\ & \cdots & \\ (\boldsymbol{\phi}_1)_{M-1} & & (\boldsymbol{\phi}_N)_{M-1} \end{bmatrix} \begin{bmatrix} t_1 \\ \vdots \\ t_N \end{bmatrix}$$

Let's analyse the j-th row of this system of equations:

$$\begin{bmatrix} \\ \alpha w_j \\ \\ \end{bmatrix} + \beta \begin{bmatrix} \\ (\boldsymbol{\phi}_1)_j(\boldsymbol{\phi}_1)_0 + \cdots + (\boldsymbol{\phi}_N)_j(\boldsymbol{\phi}_N)_0 & \cdots & (\boldsymbol{\phi}_1)_j(\boldsymbol{\phi}_1)_{M-1} + \cdots + (\boldsymbol{\phi}_N)_j(\boldsymbol{\phi}_N)_{M-1} \\ \\ \end{bmatrix} \begin{bmatrix} w_0 \\ \vdots \\ w_{M-1} \end{bmatrix} =$$

$$\beta \begin{bmatrix} \\ (\boldsymbol{\phi}_1)_j t_1 + \cdots + (\boldsymbol{\phi}_N)_j t_n \\ \\ \end{bmatrix}$$

$$\begin{bmatrix} \\ \alpha w_j \\ \\ \end{bmatrix} + \beta \begin{bmatrix} \\ w_0\left((\boldsymbol{\phi}_1)_j(\boldsymbol{\phi}_1)_0 + \cdots + (\boldsymbol{\phi}_N)_j(\boldsymbol{\phi}_N)_0\right) + \cdots + w_{M-1}\left((\boldsymbol{\phi}_1)_j(\boldsymbol{\phi}_1)_{M-1} + \cdots + (\boldsymbol{\phi}_N)_j(\boldsymbol{\phi}_N)_{M-1}\right) \\ \\ \end{bmatrix} =$$

$$\beta \begin{bmatrix} \\ (\boldsymbol{\phi}_1)_j t_1 + \cdots + (\boldsymbol{\phi}_N)_j t_n \\ \\ \end{bmatrix}$$

Which is equivalent to the equation:

$$w_j\alpha + \beta\sum_{k=0}^{M-1} w_k \sum_{i=1}^{N} (\boldsymbol{\phi}_i)_k (\boldsymbol{\phi}_i)_j = \beta\sum_{i=1}^{N} t_i (\boldsymbol{\phi}_i)_j$$

Since this equation is valid for all $j \in \{0, 1, \cdots, M-1\}$, we have a system of M equations equivalent to the matrix form presented earlier. Thus having the same solution $\mathbf{w}_{\text{MAP}}$.

6. Bonus - What is the role this basis function? Why should we avoid placing the same penalty/prior for this basis? Rewrite $p(\mathbf{w})$ so that the first basis function has its own prior/penalty.

This first basis function accounts for the intercept of the model, i.e. a shift in the $\mathbf{t}$ dimension. Note that penalizing this intercept would generate an undesired dependence on the origin chosen for $\mathbf{t}$, because adding a constant to the target vector would not simply be reflected by an equivalent shift in the predictions since the *bias* term would also be penalized. For instance, consider the problem of predicting temperature given input $\mathbf{X}$ for targets $\mathbf{t}_C$ measured in Celsius and $\mathbf{t}_K$ measured in Kelvins, for which $\mathbf{t}_K = 273.15 + \mathbf{t}_C$.

$$p(\mathbf{w}|\alpha_0, \alpha_1) = \mathcal{N}(w_0|0, \alpha_0^{-1})\mathcal{N}(\mathbf{w}_{[1..M-1]}|\mathbf{0}, \alpha_1^{-1}\mathbf{I})$$

2 Probability Distributions, Likelihoods, and Estimators

Question 2.1

For each of the probability distributions above, write down their normalizing constants. Interpretation: constant k s.t $\int_{\Omega} kf(x)dx = 1 = \sum_{x\Omega} kp(x)$ where f gathers only the functional terms directly depending on the random variable.

Distribution	Normalizing Constant	Distribution	Normalizing Constant
Bernoulli	1	Beta	$\frac{\Gamma(\theta_1+\theta_0)}{\Gamma(\theta_1)\Gamma(\theta_0)}$
Poisson	$e^{-\theta}$	Gamma	$\frac{\theta_1^{\theta_0}}{\Gamma(\theta_0)}$
Gaussian	$\frac{1}{\sqrt{2\pi\theta_1}}$		

Question 2.2

For this problem we are given $\{r_t\}_{t=1}^{365}$ and assume $p(r_t|\rho) = \text{Bernoulli}(r_t|\rho)$. Note that the pmf of the Bernoulli distribution can be states as $p(r_t|\rho) = \rho^{r_t}(1-\rho)^{1-r_t}$.

1. What is the likelihood for a single observation?

$$p(r_t|\rho) = \rho^{r_t}(1-\rho)^{1-r_t}$$

For the entire set of observations?

$$p(\{r_t\}_{t=1}^{N}|\rho) = \prod_{t=1}^{N} \rho^{r_t}(1-\rho)^{1-r_t} = \rho^{\sum_{t=1}^{N} r_t}(1-\rho)^{N-\sum_{t=1}^{N} r_t}$$

2. Write the log-likelihood for the entire set of observations.

$$\log p(\{r_t\}_{t=1}^{N}|\rho) = \log(\rho)\sum_{t=1}^{N} r_t + \log(1-\rho)\left(N - \sum_{t=1}^{N} r_t\right)$$

3. Solve for the MLE of ρ. Do it in general (with symbols for counts n_0, n_1 for days without and with rain) and for this specific case (plug-in the numbers).

Let $n_1 = \sum_{t=1}^{N} r_t$ and $n_0 = N - n_1$. Then, the likelihood and log-likelihood can be stated as:

$$p(\{r_t\}_{t=1}^{N}|\rho) = \rho^{n_1}(1-\rho)^{n_0} \qquad \log p(\{r_t\}_{t=1}^{N}|\rho) = n_1\log(\rho) + n_0\log(1-\rho)$$

$$\frac{\partial}{\partial\rho}\log p(\{r_t\}_{t=1}^{N}|\rho) = \frac{n_1}{\rho} - \frac{n_0}{1-\rho} = 0 \quad \rightarrow \quad \rho_{\text{MLE}} = \frac{n_1}{n_1+n_0} = \frac{n_1}{N}$$

In this case, $\rho_{\text{MLE}} = 217/365$.

4. Assume a Beta prior for ρ with parameters a and b. What is the MAP for ρ? 5. Write the form of the posterior distribution for ρ? You do not need to solve it analytically. 6. (Optional) Solve for the posterior distribution analytically.

$$p(\rho|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\rho^{a-1}(1-\rho)^{b-1}$$

$$p(\rho|\{r_t\}_{t=1}^{N}, a, b) = \frac{p(\{r_t\}_{t=1}^{N}|\rho)\,p(\rho|a,b)}{p(\{r_t\}_{t=1}^{N})} = \frac{\rho^{n_1}(1-\rho)^{n_0}\cancel{\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}}\rho^{a-1}(1-\rho)^{b-1}}{\cancel{\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}}\int_0^1 \rho'^{(n_1+a-1)}(1-\rho')^{(n_0+b-1)}d\rho'}$$

Note that the integral in the denominator is the integral over the domain of a Beta distribution with variable ρ' and parameters $n_1 + a$ and $n_0 + b$ without its normalization constant. Thus, this integral must be equivalent to the inverse of the normalization constant of that distribution, namely, $\dfrac{\Gamma(n_1+a)\Gamma(n_0+b)}{\Gamma(n_1+a+n_0+b)}$.

Then, we have that the posterior distribution for ρ is:

$$p(\rho|\{r_t\}_{t=1}^N, a, b) = \frac{\Gamma(n_1+a+n_0+b)}{\Gamma(n_1+a)\Gamma(n_0+b)}\rho^{n_1+a-1}(1-\rho)^{n_0+b-1} = \text{Beta}(\rho|n_1+a, n_0+b)$$

$$\log p(\rho|\{r_t\}_{t=1}^N, a, b) = \log\left(\frac{\Gamma(n_1+a+n_0+b)}{\Gamma(n_1+a)\Gamma(n_0+b)}\right) + (n_1+a-1)\log(\rho) + (n_0+b-1)\log(1-\rho)$$

$$\frac{\partial}{\partial\rho}\log p(\rho|\{r_t\}_{t=1}^N, a, b) = \frac{n_1+a-1}{\rho} + \frac{n_0+b-1}{1-\rho} = 0 \quad \rightarrow \quad \rho_{\text{MAP}} = \frac{n_1+a-1}{n_1+a-1+n_0+b-1}$$

Question 2.3

For this problem we are given $d_1, \ldots, d_{14} = 4, 7, 3, 0, 2, 2, 1, 5, 4, 4, 3, 3, 2, 3$ and assume $p(d_t|\lambda) = \text{Poisson}(d_t|\lambda)$. Note that the sum of $\{d_t\} = 43$.

1. What is the likelihood for a single observation?

$$p(d_t|\lambda) = \frac{\lambda^{d_t}}{d_t!}e^{-\lambda}$$

For the entire set of observations?

$$p(\{d_t\}_{t=1}^N|\rho) = \prod_{t=1}^N \frac{\lambda^{d_t}}{d_t!}e^{-\lambda} = \frac{\lambda^{\sum_{t=1}^N d_t}\, e^{-N\lambda}}{\prod_{t=1}^N d_t!}$$

2. Write the log-likelihood for the entire set of observations.

$$\log p(\{d_t\}_{t=1}^N|\rho) = \left(\sum_{t=1}^N d_t\right)\log(\lambda) - \lambda N - \sum_{t=1}^N \log(d_t!)$$

3. Solve for the MLE of λ. Do it in general and for this specific case (plug-in the numbers).

$$\frac{\partial}{\partial\lambda}\log p(\{d_t\}_{t=1}^N|\lambda) = \frac{\sum_{t=1}^N d_t}{\lambda} - N = 0 \quad \rightarrow \quad \lambda_{\text{MLE}} = \frac{\sum_{t=1}^N d_t}{N}$$

In this case, $\rho_{\text{MLE}} = 43/14$.

4. Assume a Gamma prior for λ with parameters a and b. What is the MAP estimate of λ? 5. Write the form of the posterior distribution for λ? You do not need to solve it analytically. 6. (Optional) Solve for the posterior distribution analytically.

$$p(\lambda|a, b) = \frac{b^a}{\Gamma(a)}\lambda^{a-1}e^{-b\lambda}$$

$$p(\lambda|\{d_t\}_{t=1}^N, a, b) = \frac{p(\{d_t\}_{t=1}^N|\lambda)\, p(\lambda|a, b)}{p(\{r_t\}_{t=1}^N)} = \frac{\dfrac{\lambda^{\sum_{t=1}^N d_t}\, e^{-N\lambda}}{\cancel{\prod_{t=1}^N d_t!}}\dfrac{\cancel{b^a}}{\cancel{\Gamma(a)}}\lambda^{a-1}e^{-b\lambda}}{\dfrac{\cancel{b^a}}{\cancel{\Gamma(a)}}\dfrac{1}{\cancel{\prod_{t=1}^N d_t!}}\displaystyle\int_0^\infty \lambda'^{\left(a+\sum_{t=1}^N d_t - 1\right)}e^{-(N+b)\lambda'}d\lambda'}$$

Note that the integral in the denominator is the integral over the domain of a Gamma distribution with variable λ' and parameters $\theta_0 = a + \sum_{t=1}^{N} d_t$ and $\theta_1 = N + b$ without its normalization constant. Thus, this integral must be equivalent to the inverse of the normalization constant of that distribution, namely, $\dfrac{\Gamma(a + \sum_{t=1}^{N} d_t)}{(N+b)^{\left(a+\sum_{t=1}^{N} d_t\right)}}$.

Then, we have that the posterior distribution for λ is:

$$p(\lambda|\{d_t\}_{t=1}^{N}, a, b) = \frac{(N+b)^{\left(a+\sum_{t=1}^{N} d_t\right)}}{\Gamma(a + \sum_{t=1}^{N} d_t)} \lambda^{\left(a+\sum_{t=1}^{N} d_t - 1\right)} e^{-(N+b)\lambda} = \text{Gamma}(\lambda | a + \sum_{t=1}^{N} d_t, N+b)$$

$$\log p(\lambda|\{d_t\}_{t=1}^{N}, a, b) = \log\left(\frac{(N+b)^{\left(a+\sum_{t=1}^{N} d_t\right)}}{\Gamma(a + \sum_{t=1}^{N} d_t)}\right) + \left(a + \sum_{t=1}^{N} d_t - 1\right)\log(\lambda) - (N+b)\lambda$$

$$\frac{\partial}{\partial \lambda}\log p(\lambda|\{d_t\}_{t=1}^{N}, a, b) = \frac{a + \sum_{t=1}^{N} d_t - 1}{\lambda} - (N+b) = 0 \quad \rightarrow \quad \lambda_{\text{MAP}} = \frac{a + \sum_{t=1}^{N} d_t - 1}{N+b}$$

Question 2.4

For this problem we are given $\{l_n\}_{n=1}^{N}$ and we know that $n \in \{D_0\}$ are the indices for the disease free patients and $n \in \{D_1\}$ are the indices for the patients with the disease. Besides we assume $p(l) = \pi_0 p(l|d=0) + \pi_1 p(l|d=1)$, where $\pi_0 = 0.999$ and $\pi_1 = 0.001$ and both $p(l|d=0)$ and $p(l|d=1)$ are conditional Gaussian distributions.

1. Write down the likelihood of the observations as a product over N level recordings. 2. Write down the likelihood as a product over the likelihoods for $\{D_0\}$ and $\{D_1\}$.

$$\begin{aligned}
p(\{(l_n, d_n)\}_{n=1}^{N} | \mu_0, \sigma_0^2, \mu_1, \sigma_1^2) &= \prod_{n=1}^{N} \left(\pi_0 \mathcal{N}(l_n|\mu_0, \sigma_0^2)\right)^{[d_n=0]} \left(\pi_1 \mathcal{N}(l_n|\mu_1, \sigma_1^2)\right)^{[d_n=1]} \\
&= \left(\prod_{n \in D_0} \pi_0 \mathcal{N}(l_n|\mu_0, \sigma_0^2)\right) \left(\prod_{n \in D_1} \pi_1 \mathcal{N}(l_n|\mu_1, \sigma_1^2)\right) \\
&= \pi_0^{|D_0|} (2\pi)^{-|D_0|/2} (\sigma_0^2)^{-|D_0|/2} \exp\left(-\frac{1}{2\sigma_0^2} \sum_{n \in D_0} (l_n - \mu_0)^2\right) \cdot \\
&\quad \pi_1^{|D_1|} (2\pi)^{-|D_1|/2} (\sigma_1^2)^{-|D_1|/2} \exp\left(-\frac{1}{2\sigma_1^2} \sum_{n \in D_1} (l_n - \mu_1)^2\right)
\end{aligned}$$

3. Compute the log-likelihood.

$$\begin{aligned}
\log p(\{(l_n, d_n)\}_{n=1}^{N} | \mu_0, \sigma_0^2, \mu_1, \sigma_1^2) &= |D_0| \log(\pi_0) - \frac{|D_0|}{2}\log(2\pi) - \frac{|D_0|}{2}\log(\sigma_0^2) - \frac{1}{2\sigma_0^2}\sum_{n \in D_0}(l_n - \mu_0)^2 \\
&\quad + |D_1| \log(\pi_1) - \frac{|D_1|}{2}\log(2\pi) - \frac{|D_1|}{2}\log(\sigma_1^2) - \frac{1}{2\sigma_1^2}\sum_{n \in D_1}(l_n - \mu_1)^2
\end{aligned}$$

4. Find the MLE for μ_0 and σ_0^2. Assume we can do the same for μ_1 and σ_1^2.

$$\frac{\partial}{\partial \mu_0} \log p(\{l_n\}_{n=1}^{N} | \mu_0, \sigma_0^2, \mu_1, \sigma_1^2) = \frac{1}{2\sigma_0^2} \sum_{n \in D_0} 2(l_n - \mu_0) = 0 \quad \rightarrow \quad \mu_{0MLE} = \frac{\sum_{n \in D_0} l_n}{|D_0|}$$

$$\frac{\partial}{\partial \sigma_0^2} \log p(\{l_n\}_{n=1}^N | \mu_0, \sigma_0^2, \mu_1, \sigma_1^2) = -\frac{|D_0|}{2\sigma_0^2} + \frac{\sum_{n \in D_0} (l_n - \mu_0)^2}{2} (\sigma_0^2)^{-2} = 0 \quad \rightarrow \quad \sigma_{0MLE}^2 = \frac{\sum_{n \in D_0} (l_n - \mu_0)^2}{|D_0|}$$

5. To make a prediction, solve for $p(d = 1|l_\star)$, where $l_\star$ is a level recorded for a new patient. 6. Reduce your solution to have the form of a sigmoid.

$$\begin{aligned}
p(d = 1|l_\star) &= \frac{p(l_\star|d = 1)\, p(d = 1)}{p(l_\star)} \\
&= \frac{\pi_1\, p(l_\star|d = 1)}{\pi_1\, p(l_\star|d = 1) + \pi_0\, p(l_\star|d = 0)} \\
&= \frac{1}{1 + \dfrac{\pi_0\, p(l_\star|d = 0)}{\pi_1\, p(l_\star|d = 1)}} \\
&= \frac{1}{1 + \dfrac{\pi_0\, \mathcal{N}(l_\star|\mu_{0MLE}, \sigma_{0MLE}^2)}{\pi_1\, \mathcal{N}(l_\star|\mu_{1MLE}, (\sigma_1^2)_{MLE})}}
\end{aligned}$$

Remember that for every z, we have $z = e^{-\log(\frac{1}{z})}$. Then, define $a(l_\star)$ as:

$$a(l_\star) = \log\left(\frac{\pi_1\, \mathcal{N}(l_\star|\mu_{1MLE}, (\sigma_1^2)_{MLE})}{\pi_0\, \mathcal{N}(l_\star|\mu_{0MLE}, \sigma_{0MLE}^2)}\right)$$

Finally, we can restate the probability $p(d = 1|l_\star)$ as:

$$p(d = 1|l_\star) = \frac{1}{1 + e^{-a(l_\star)}}$$