Assignment 2
Machine Learning 1, Fall 2016

Dana Kianfar
University of Amsterdam

January 6, 2018

1 MAP solution for Linear Regression

Question 1

1. Given that our data is i.i.d we can write the likelihood in form (a) as follows.

$$
\begin{aligned}
p(\mathcal{D} \mid \boldsymbol{\theta}) &= \prod_{n=1}^{N} p\left(t_n | \phi_n, \mathbf{w}, \beta\right) \\
&= \prod_{n=1}^{N} \mathcal{N}\left(t_n | \mathbf{w}^T \phi_n, \beta^{-}1\right) \\
&= \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\beta^{-1}}} \exp\left(\frac{-\left(t_n - \mathbf{w}^T\phi_n\right)^2}{2\beta^{-1}}\right) \\
&= \underbrace{\left(\frac{\beta}{2\pi}\right)^{\frac{N}{2}}}_{A_1} \exp\left(\underbrace{\frac{-\beta}{2}\sum_{n=1}^{N}\left(t_n - \mathbf{w}^T\phi_n\right)^2}_{A_2}\right)
\end{aligned}
$$

We can then rewrite A_2 as follows.

$$
A_2 = \frac{-1}{2}\sum_{n=1}^{N}\left(t_n - \mathbf{w}^T\phi_n\right)\beta\left(t_n - \mathbf{w}^T\phi_n\right)
$$

Where $A_2 \in \mathbb{R}, \mathbf{w} \in \mathbb{R}^{M\times 1}, \phi_n \in \mathbb{R}^{M\times 1}, t_n \in \mathbb{R}$.

We can achieve the same scalar result for A_2 by using dot products as follows.

$$
A_2 = \frac{-1}{2}\left(\mathbf{t} - \Phi\mathbf{w}\right)^T \Sigma^{-1}\left(\mathbf{t} - \Phi\mathbf{w}\right)
$$

Where:

$\Sigma \in \mathbb{R}^{N\times N} = \beta^{-1}\mathbb{I}_N$: is the co-variance matrix
$\Phi \in \mathbb{R}^{N\times M}$: is the design matrix composed of $\left(\phi_1, \phi_2, \cdots, \phi_n\right)^T$
$\mathbf{t} \in \mathbb{R}^{N\times 1}$: is the target vector composed of $\left(t_1, t_2, \cdots, t_n\right)^T$
$\mathbb{I}_N \in \mathbb{R}^{N\times N}$: is the identity matrix.

The term A_1 does not change as $|\Sigma| = \beta^{-N}$. Therefore we can rewrite $p(\mathcal{D} \mid \boldsymbol{\theta})$ in form (b) as follows.

$$
\begin{aligned}
p(\mathcal{D} \mid \boldsymbol{\theta}) &= \left(\frac{1}{|\Sigma|^{\frac{1}{2}} (2\pi)^{\frac{N}{2}}}\right) \exp\left(\frac{-1}{2} (\mathbf{t} - \Phi\mathbf{w})^T \Sigma^{-1} (\mathbf{t} - \Phi\mathbf{w})\right) \\
&= \left(\frac{\beta}{2\pi}\right)^{\frac{N}{2}} \exp\left(\frac{-1}{2} (\mathbf{t} - \Phi\mathbf{w})^T \Sigma^{-1} (\mathbf{t} - \Phi\mathbf{w})\right) \\
&= \mathcal{N}(\mathbf{t}|\Phi\mathbf{w}, \Sigma)
\end{aligned}
$$

2.

$$
\begin{aligned}
p(\mathbf{w}|\alpha) &= \mathcal{N}\left(\mathbf{w}|\vec{0}, \alpha^{-1}\mathbb{I}\right) \\
&= \left(\frac{\alpha}{2\pi}\right)^{\frac{M}{2}} \exp\left(\frac{-1}{2} \left(\mathbf{w} - \vec{0}\right)^T \mathbf{S}^{-1} \left(\mathbf{w} - \vec{0}\right)\right) \\
&= \left(\frac{\alpha}{2\pi}\right)^{\frac{M}{2}} \exp\left(\frac{-1}{2} \mathbf{w}^T \mathbf{S}^{-1} \mathbf{w}\right)
\end{aligned}
$$

Where:

$\mathbf{S} \in \mathbb{R}^{M\times M} = \alpha^{-1}\mathbb{I}_M$: is the co-variance matrix

$\mathbb{I}_M \in \mathbb{R}^{M\times M}$: is the identity matrix.

We compute its log as follows.

It holds that $S^{-1} = \left(\alpha^{-1}\mathbb{I}_M\right)^{-1} = \alpha\mathbb{I}_M$ and $\mathbf{w}^T\mathbb{I}_M\mathbf{w} = \mathbf{w}^T\mathbf{w}$.

$$
\begin{aligned}
\log\left(p(\mathbf{w}|\alpha)\right) &= \frac{M}{2}\left(\log\alpha - \log(2\pi)\right) - \frac{1}{2}\mathbf{w}^T\mathbf{S}^{-1}\mathbf{w} \\
&= \frac{M}{2}\left(\log\alpha - \log(2\pi)\right) - \frac{\alpha}{2}\mathbf{w}^T\mathbf{w}
\end{aligned}
$$

3. The posterior is calculated as follows.

$$
\textit{posterior} = \frac{\textit{likelihood} \times \textit{prior}}{\textit{evidence}}
$$

We replace each term by its mathematical notation as follows.

$$
p(\mathbf{w}|\mathbf{t}, \Phi, \beta, \alpha) = \frac{p(\mathbf{t}|\Phi, \mathbf{w}, \beta) \cdot p(\mathbf{w}|\alpha)}{p(\mathbf{t})}
$$

Where:

$p(\mathbf{w}|\mathbf{t}, \Phi, \beta, \alpha)$: is the posterior

$p(\mathbf{t}|\Phi, \mathbf{w}, \beta)$: is the likelihood

$p(\mathbf{w}|\alpha)$: is the prior

$p(\mathbf{t})$: is the evidence

We expand further.

$$
p(\mathbf{w}|\mathbf{t}, \Phi, \beta, \alpha) = \frac{p(\mathbf{t}|\Phi, \mathbf{w}, \beta) \cdot p(\mathbf{w}|\alpha)}{\int_\Theta p(\mathbf{t}|\Phi, \mathbf{w}', \beta) \cdot p(\mathbf{w}'|\alpha)\, d\mathbf{w}'}
$$

Where Θ is the domain of the p.d.f of $\mathbf{w}$. This is a consequence of a Bayesian setting, where $\mathbf{w}$ is a random variable and all it's possible values are considered simultaneously.

4. Form (a)

$$p\left(\mathbf{w}|\mathbf{t},\Phi,\beta,\alpha\right)=\frac{\beta}{2}\sum_{n=1}^{N}\left(t_n-\mathbf{w}^T\phi_n\right)^2-\frac{\alpha}{2}\mathbf{w}^T\mathbf{w}+\frac{N}{2}\left(\log\beta-\log\left(2\pi\right)\right)$$
$$+\frac{M}{2}\left(\log\alpha-\log\left(2\pi\right)\right)-\log\left(p\left(\mathbf{t}\right)\right)$$
$$=\frac{\beta}{2}\sum_{n=1}^{N}\left(t_n-\mathbf{w}^T\phi_n\right)^2-\frac{\alpha}{2}\mathbf{w}^T\mathbf{w}+\mathbf{C}$$

Form (b)

$$\log p\left(\mathbf{w}|\mathbf{t},\Phi,\beta,\alpha\right)=\frac{-1}{2}\left(\mathbf{t}-\Phi\mathbf{w}\right)^T\Sigma^{-1}\left(\mathbf{t}-\Phi\mathbf{w}\right)-\frac{\alpha}{2}\mathbf{w}^T\mathbf{w}+\frac{N}{2}\left(\log\beta-\log\left(2\pi\right)\right)$$
$$+\frac{M}{2}\left(\log\alpha-\log\left(2\pi\right)\right)-\log\left(p\left(\mathbf{t}\right)\right)$$
$$=\frac{-1}{2}\left(\mathbf{t}-\Phi\mathbf{w}\right)^T\Sigma^{-1}\left(\mathbf{t}-\Phi\mathbf{w}\right)-\frac{\alpha}{2}\mathbf{w}^T\mathbf{w}+\mathbf{C}$$

With an optimization goal over $\mathbf{w}$, we can simply ignore the denominator of the posterior as it does not depend on $\mathbf{w}$. Therefore by maximizing the numerator over $\mathbf{w}$, we are in fact maximizing the entire posterior distribution. Estimating the MAP is easier than working with the full posterior distribution as the MAP estimator ignores the denominator of the posterior $p(\mathbf{t})$ which complex and computationally intractable.

5. Form (a)

$$\frac{\partial\log p\left(\mathbf{w}|\mathbf{t},\Phi,\beta,\alpha\right)}{\partial\mathbf{w}}=\frac{\beta}{2}\sum_{n=1}^{N}(-2)\left(t_n-\mathbf{w}^T\phi_n\right)\phi_n-\frac{\alpha}{2}\cdot 2\cdot\mathbf{w}$$
$$=-\beta\sum_{n=1}^{N}\left(t_n\phi_n-\mathbf{w}\phi_n^T\phi_n\right)-\alpha\mathbf{w}$$
$$=-\beta\sum_{n=1}^{N}t_n\phi_n-\beta\sum_{n=1}^{N}\mathbf{w}\phi_n^T\phi_n-\alpha\mathbf{w}$$
$$\frac{\partial\log p\left(\mathbf{w}|\mathbf{t},\Phi,\beta,\alpha\right)}{\partial\mathbf{w}}=0$$
$$-\beta\sum_{n=1}^{N}t_n\phi_n+\beta\sum_{n=1}^{N}\mathbf{w}\phi_n^T\phi_n-\alpha\mathbf{w}=0$$
$$\left(\beta\sum_{n=1}^{N}\phi_n^T\phi_n-\alpha\right)\mathbf{w}=\beta\sum_{n=1}^{N}t_n\phi_n$$
$$\mathbf{w}=\frac{\sum_{n=1}^{N}t_n\phi_n}{\left(\sum_{n=1}^{N}\phi_n^T\phi_n-\frac{\alpha}{\beta}\right)}$$

Form (b)

$$
\begin{aligned}
\frac{\partial \log p(\mathbf{w}|\mathbf{t},\Phi,\beta,\alpha)}{\partial \mathbf{w}} &= \frac{\partial\left(\frac{-1}{2}\left((\mathbf{t}-\Phi\mathbf{w})^T\Sigma^{-1}(\mathbf{t}-\Phi\mathbf{w})+\alpha\mathbf{w}^T\mathbf{w}\right)+\mathbf{C}\right)}{\partial \mathbf{w}} \\
&= \frac{\partial\left(\frac{-1}{2}\left(\mathbf{t}^T\Sigma^{-1}\mathbf{t}-\mathbf{t}^T\Sigma^{-1}\Phi\mathbf{w}-\mathbf{w}^T\Phi^T\Sigma^{-1}\mathbf{t}+\mathbf{w}^T\Phi^T\Sigma^{-1}\Phi\mathbf{w}+\alpha\mathbf{w}^T\mathbf{w}\right)\right)}{\partial \mathbf{w}} \\
&= \frac{-1}{2}\left(-\Phi^T\Sigma^{-1}\mathbf{t}-\Phi^T\Sigma^{-1}\mathbf{t}+2\Phi^T\Sigma^{-1}\Phi\mathbf{w}+2\alpha\mathbf{w}\right) \\
&= \frac{-1}{2}\left(-2\Phi^T\Sigma^{-1}\mathbf{t}+2\Phi^T\Sigma^{-1}\Phi\mathbf{w}+2\alpha\mathbf{w}\right) \\
&= \Phi^T\Sigma^{-1}\mathbf{t}-\Phi^T\Sigma^{-1}\Phi\mathbf{w}-\alpha\mathbf{w} \\
&= \beta\Phi^T\mathbf{t}-\left(\beta\Phi^T\Phi-\alpha\mathbb{I}_M\right)\mathbf{w} \\
\frac{\partial \log p(\mathbf{w}|\mathbf{t},\Phi,\beta,\alpha)}{\partial \mathbf{w}} &= 0 \\
\beta\Phi^T\mathbf{t}-\left(\beta\Phi^T\Phi-\alpha\mathbb{I}_M\right)\mathbf{w} &= 0 \\
\left(\beta\Phi^T\Phi-\alpha\mathbb{I}_M\right)\mathbf{w} &= \beta\Phi^T\mathbf{t} \\
\mathbf{w} &= \left(\beta\Phi^T\Phi-\alpha\mathbb{I}_M\right)^{-1}\beta\Phi^T\mathbf{t} \\
\mathbf{w} &= \left(\Phi^T\Phi-\frac{\alpha}{\beta}\mathbb{I}_M\right)^{-1}\Phi^T\mathbf{t}
\end{aligned}
$$

6. The first term of the coefficient vector $\mathbf{w}_0$ accounts for the distance of our predictions from the origin. The assignment of $\phi_0 = 1$ allows the model to fit the intercept along the y-axis. This is the so called bias term of the model. It is simply a shift along the y-axis. Penalizing this term would make the optimization depend on the origin of $\mathbf{t}$, and as a result the fitted model would try to fit the data points while passing through the origin. The variance is not the same in the w_0 coefficient as the rest of the vector. In some applications, data is normalized and re-scaled in which case all data points are centered on the origin. We may however give w_0 its own regularization parameter by separating it from the joint distribution of $\mathbf{w}$ in our prior $p(\mathbf{w}|\alpha)$ as follows.

$$p(\mathbf{w}|\alpha,\alpha') = p(\mathbf{w}'|\alpha)\cdot p(w_0|\alpha')$$

Where $\mathbf{w} \in \mathbb{R}^M$ and $\mathbf{w}' = (w_1, w_2, \cdots, w_{M-1})^T$.

2 Probability distributions, likelihoods and estimators

Question 2.1

1. The normalizing constant for the Bernoulli distribution is 1.

$$\sum_{x=0}^{1}\theta^{[x=1]}(1-\theta)^{[x=0]} = \theta + 1 - \theta = 1$$

The normalizing constant for the Beta distribution is $\frac{\Gamma(\theta_1+\theta_0)}{\Gamma(\theta_0)\Gamma(\theta_1)}$.

$$\int_{x\in[0,1]}\frac{\Gamma(\theta_1+\theta_0)}{\Gamma(\theta_0)\Gamma(\theta_1)}x^{\theta_1-1}(1-x)^{\theta_0-1}dx = \frac{\Gamma(\theta_1+\theta_0)}{\Gamma(\theta_0)\Gamma(\theta_1)}\int_{x\in[0,1]}x^{\theta_1-1}(1-x)^{\theta_0-1}dx = 1$$

The normalizing constant for the Poisson distribution is $e^{-\theta}$.

$$\sum_{x=0} \frac{\theta^x}{x!} e^{-\theta} = e^{-\theta} \sum_{x=0} \frac{\theta^x}{x!} = 1$$

The normalizing constant for the Gamma distribution is $\frac{\theta_1^{\theta_0}}{\Gamma(\theta_0)}$.

$$\int_{x \in \mathbb{R}_{\geq 0}} \frac{\theta_1^{\theta_0}}{\Gamma(\theta_0)} \cdot x^{\theta_0 - 1} e^{-\theta_1 x} dx = \frac{\theta_1^{\theta_0}}{\Gamma(\theta_0)} \int_{x \in \mathbb{R}_{\geq 0}} x^{\theta_0 - 1} e^{-\theta_1 x} dx = 1$$

The normalizing constant for the Gaussian distribution is $\frac{1}{\sqrt{2\pi\theta_1}}$.

$$\int_{x \in \mathbb{R}} \frac{1}{\sqrt{2\pi\theta_1}} \cdot e^{\frac{-(x-\theta_0)^2}{2\theta_1}} dx = \frac{1}{\sqrt{2\pi\theta_1}} \int_{x \in \mathbb{R}} e^{\frac{-(x-\theta_0)^2}{2\theta_1}} dx = 1$$

Question 2.2

1. Single observation

$$p(r_t|\rho) = Bernoulli(r_t|\rho) = \rho^{[r_t=1]}(1-\rho)^{[r_t=0]}$$

Entire set of observations. Dataset is iid.

$$\begin{aligned} p(\mathbf{r}|\rho) = \prod_{n=1}^{365} Bernoulli(r_t|\rho) &= \prod_{n=1}^{365} \rho^{[r_t=1]}(1-\rho)^{[r_t=0]} \\ &= \rho^{217}(1-\rho)^{365-217} \\ &= \rho^{217}(1-\rho)^{148} \end{aligned}$$

2.

$$\log(p(\mathbf{r}|\rho)) = \log(\rho^{217}(1-\rho)^{148}) = 217 \log \rho + 148 \log(1-\rho)$$

3.

$$\begin{aligned} \frac{\partial \log(p(\mathbf{r}|\rho))}{\partial \rho} &= \frac{\partial(n_1 \log \rho + n_0 \log(1-\rho))}{\partial \rho} \\ &= \frac{n_1}{\rho} - \frac{n_0}{1-\rho} \\ \frac{\partial \log(p(\mathbf{r}|\rho))}{\partial \rho} &= 0 \\ \frac{n_1(1-\rho) - n_0\rho}{\rho(1-\rho)} &= 0 \\ n_1(1-\rho) - n_0\rho &= 0 \\ -(n_1+n_0)\rho &= -n_1 \\ \rho_{ML} &= \frac{n_1}{n_1+n_0} \end{aligned}$$

For the problem at hand, we plug in the given numbers as follows.

$$\begin{aligned} \rho_{ML} &= \frac{n_1}{n_1+n_0} \\ &= \frac{217}{365} \end{aligned}$$

4.

$$
\begin{aligned}
p(\rho|\mathbf{r},a,b) &= \frac{p(\mathbf{r}|\rho)\cdot p(\rho|a,b)}{p(\mathbf{r})} \\
\log p(\rho|\mathbf{r},a,b) &= \log p(\mathbf{r}|\rho) + \log p(\rho|a,b) - \log p(\mathbf{r}) \\
&= \log\left(\rho^{n_1}(1-\rho)^{n_0}\right) + \log\left(\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\cdot\rho^{a-1}(1-\rho)^{b-1}\right) - \log p(\mathbf{r}) \\
&= n_1\log\rho + n_0\log(1-\rho) + \log\left(\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\right) + (a-1)\log\rho + (b-1)\log(1-\rho) \\
&- \log p(\mathbf{r}) \\
&= \log\rho\,(n_1+a-1) + \log(1-\rho)\,(n_0+b-1) + C
\end{aligned}
$$

Where constant C is the sum of terms that don't depend on ρ.

$$
\begin{aligned}
\frac{\partial \log p(\rho|\mathbf{r},a,b)}{\partial\rho} &= \frac{n_1+a-1}{\rho} - \frac{n_0+b-1}{1-\rho} \\
\frac{\partial \log p(\rho|\mathbf{r},a,b)}{\partial\rho} &= 0 \\
\frac{n_1+a-1}{\rho} - \frac{n_0+b-1}{1-\rho} &= 0 \\
\frac{n_1+a-1}{\rho} &= \frac{n_0+b-1}{1-\rho} \\
\frac{1-\rho}{\rho} &= \frac{n_0+b-1}{n_1+a-1} \\
\frac{1}{\rho} - 1 &= \frac{n_0+b-1}{n_1+a-1} \\
\frac{1}{\rho} &= \frac{n_0+b-1+n_1+a-1}{n_1+a-1} \\
\rho_{MAP} &= \frac{n_1+a-1}{n_0+b-1+n_1+a-1} \\
&= \frac{n_1+a-1}{n_0+n_1+b+a-2}
\end{aligned}
$$

Using the information provided in this problem, we have the following.

$$
\rho_{MAP} = \frac{216+a}{363+b+a}
$$

5.

$$
\begin{aligned}
p(\rho|\mathbf{r},a,b) &= \frac{p(\mathbf{r}|\rho)\cdot p(\rho|a,b)}{\int_0^1 p(\mathbf{r}|\rho')\cdot p(\rho')\cdot d\rho'} \\
&= \frac{\rho^{n_1}(1-\rho)^{n_0}\cdot\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\cdot\rho^{a-1}(1-\rho)^{b-1}}{\int_0^1 \rho'^{n_1}(1-\rho')^{n_0}\cdot\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\cdot\rho'^{a-1}(1-\rho')^{b-1}\cdot d\rho'} \\
&= \frac{\rho^{n_1+a-1}(1-\rho)^{n_0+b-1}}{\int_0^1 \rho'^{n_1+a-1}(1-\rho')^{n_0+b-1}\cdot d\rho'}
\end{aligned}
$$

6.

$$\int_0^1 p(\rho|\mathbf{r},a,b)\cdot d\rho = \int_0^1 \frac{\rho^{n_1+a-1}(1-\rho)^{n_0+b-1}}{\underbrace{\int_0^1 \rho'^{n_1+a-1}(1-\rho')^{n_0+b-1}\cdot d\rho'}_{\text{C}}}\cdot d\rho = 1$$

C is a constant w.r.t to ρ. The RHS has taken the form of a Beta distribution. Therefore, we have the following.

$$\int_0^1 \rho'^{n_1+a-1}(1-\rho')^{n_0+b-1}\cdot d\rho' = \int_0^1 \rho^{n_1+a-1}(1-\rho)^{n_0+b-1}\cdot d\rho$$

Furthermore, we can express the posterior by reformulating the denominator C as the normalizing constant of a Beta distribution. We have expressed the posterior as a new Beta distribution with parameters $\theta_1 = a + n_1$ and $\theta_0 = b + b_0$.

$$\int_0^1 p(\rho|\mathbf{r},a,b)\cdot d\rho = \frac{\Gamma(n_1+a)\Gamma(n_0+b)}{\Gamma(n_0+n_1+a+b)}\rho^{n_1+a-1}(1-\rho)^{n_0+b-1}$$

Question 2.3

1. Estimate for a single observation

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

For the entire set of observations, we have the following. T is the total number of observation points, which in this problem is 14. We also define $\mathbf{d} = \{d_t\}_1^T$.

$$\begin{aligned} p(\mathbf{d}|\lambda) \sim Poisson\left(\mathbf{d}|\lambda\right) &= \prod_{t=1}^{T}\frac{\lambda^{d_t}}{d_t!}e^{-\lambda} \\ &= \frac{e^{-T\lambda}\cdot\lambda^{\left(\sum_{t=1}^T d_t\right)}}{\prod_{t=1}^T d_t!} \end{aligned}$$

2.

$$\begin{aligned} \log p(\mathbf{d}|\lambda) &= -T\lambda + \log\lambda\cdot\sum_{t=1}^T d_t - \sum_{t=1}^T \log d_t! \\ &= -T\lambda + \log\lambda\cdot\sum_{t=1}^T d_t - \sum_{t=1}^T\sum_{i=1}^{d_t} i \\ &= -T\lambda + \log\lambda\cdot\sum_{t=1}^T d_t - \sum_{t=1}^T \frac{d_t\left(d_t+1\right)}{2} \\ &= -T\lambda + \log\lambda\cdot\sum_{t=1}^T d_t - \sum_{t=1}^T \frac{d_t^2+d_t}{2} \end{aligned}$$

3. General Case. We define $t_1 = \sum_{t=1}^{T} d_t$.

$$\begin{aligned}
\frac{\partial \log p\left(\mathbf{d}|\lambda\right)}{\partial \lambda} &= -T + \frac{t_1}{\lambda} \\
\frac{\partial \log p\left(\mathbf{d}|\lambda\right)}{\partial \lambda} &= 0 \\
-T + \frac{t_1}{\lambda} &= 0 \\
\frac{t_1}{\lambda} &= T \\
\lambda_{ML} &= \frac{t_1}{T}
\end{aligned}$$

Specific Case. We have $t_1 = \sum_{t=1}^{T} d_t = 43$ and $T = 14$.

$$\lambda_{ML} = \frac{t_1}{T} = \frac{43}{14}$$

4. We apply Bayes theorem to find the posterior distribution given the prior for λ.

$$\begin{aligned}
p\left(\lambda|\mathbf{d}, a, b\right) &= \frac{p\left(\mathbf{d}|\lambda\right) p\left(\lambda|a, b\right)}{\underbrace{\int_{\lambda' \in \mathbb{R}_{\geq 0}} p\left(\mathbf{d}|\lambda'\right) p\left(\lambda'|a, b\right) d\lambda'}_{C}} \\
&= \frac{\frac{e^{-T\lambda} \cdot \lambda^{t_1}}{\prod_{t=1}^{T} d_t!} \cdot \frac{b^a \lambda^{a-1} e^{-b\lambda}}{\Gamma(a)}}{C} \\
&= \frac{\frac{e^{-(T+b)\lambda} \cdot \lambda^{(a-1+t_1)} \cdot b^a}{\prod_{t=1}^{T} d_t! \Gamma(a)}}{C} \\
\log p\left(\lambda|\mathbf{d}, a, b\right) &= \log p\left(\mathbf{d}|\lambda\right) + \log p\left(\lambda|a, b\right) - \log C \\
&= -(T+b)\lambda + \log \lambda \cdot (a - 1 + t_1) - \sum_{t=1}^{T} \log d_t! - a \log b - \log\left(\Gamma\left(a\right)\right) + \log C \\
\frac{\partial \log p\left(\lambda|\mathbf{d}, a, b\right)}{\partial \lambda} &= -(T+b) + \frac{a - 1 + t_1}{\lambda} \\
\frac{\partial \log p\left(\lambda|\mathbf{d}, a, b\right)}{\partial \lambda} &= 0 \\
-(T+b) + \frac{a - 1 + t_1}{\lambda} &= 0 \\
\frac{a - 1 + t_1}{\lambda} &= T + b \\
\lambda_{MAP} &= \frac{a - 1 + t_1}{T + b}
\end{aligned}$$

5.

$$
\begin{aligned}
p(\lambda|\mathbf{d}, a, b) &= \frac{p(\mathbf{d}|\lambda)\, p(\lambda|a, b)}{\int_{\lambda' \in \mathbb{R}_{\geq 0}} p(\mathbf{d}|\lambda')\, p(\lambda'|a, b)\, d\lambda'} \\
&= \frac{\frac{e^{-(T+b)\lambda} \cdot \lambda^{(a-1+t_1)} \cdot b^a}{\prod_{t=1}^{T} d_t! \Gamma(a)}}{\int_{\lambda' \in \mathbb{R}_{\geq 0}} \frac{e^{-(T+b)\lambda'} \cdot \lambda'^{(a-1+t_1)} \cdot b^a}{\prod_{t=1}^{T} d_t! \Gamma(a)} d\lambda'} \\
&= \frac{e^{-(T+b)\lambda} \cdot \lambda^{(a-1+t_1)}}{\underbrace{\int_{\lambda' \in \mathbb{R}_{\geq 0}} e^{-(T+b)\lambda'} \cdot \lambda'^{(a-1+t_1)} d\lambda'}_{C}}
\end{aligned}
$$

C is a constant w.r.t to λ. The RHS has taken the form of a Gamma distribution. Given that the integral of the posterior over λ is equal to one (as its a valid p.d.f), we can conclude that the denominator C is the normalizing constant and reformulate the posterior as follows. The result is a Gamma distribution with parameters $\theta_0 = t_1 + a$ and $\theta_1 = T + b$.

$$
p(\lambda|\mathbf{d}, a, b) = \frac{(T+b)^{t_1+a} \cdot \lambda^{t_1+a-1} \cdot e^{-(T+b)\lambda}}{\Gamma(a + t_1)}
$$

Question 2.4

1.

$$
p(\mathbf{l}|\mu_0, \sigma_0, \mu_1, \sigma_1) = \prod_{n=1}^{N} \left(\pi_0 \cdot \mathcal{N}(l_n|\mu_0, \sigma_0)^{[n \in \mathcal{D}_0]}\right) \cdot \left(\pi_1 \cdot \mathcal{N}(l_n|\mu_1, \sigma_1)^{[n \in \mathcal{D}_1]}\right)
$$

2.

$$
p(\mathbf{l}|\mu_0, \sigma_0, \mu_1, \sigma_1) = \prod_{n \in \mathcal{D}_0} (\pi_0 \cdot \mathcal{N}(l_n|\mu_0, \sigma_0)) \cdot \prod_{n \in \mathcal{D}_1} (\pi_1 \cdot \mathcal{N}(l_n|\mu_1, \sigma_1))
$$

3.

$$
\begin{aligned}
\log p(\mathbf{l}|\mu_0, \sigma_0, \mu_1, \sigma_1) &= \sum_{n \in \mathcal{D}_0} \log(\pi_0 \cdot \mathcal{N}(l_n|\mu_0, \sigma_0)) + \sum_{n \in \mathcal{D}_1} \log(\pi_1 \cdot \mathcal{N}(l_n|\mu_1, \sigma_1)) \\
&= -\frac{|\mathcal{D}_0|}{2}\left(-2\log \pi_0 + \log 2\pi + \log \sigma_0^2\right) - \frac{1}{2} \sum_{n \in \mathcal{D}_0} \frac{(l_n - \mu_0)^2}{\sigma_0^2} \\
&+ -\frac{|\mathcal{D}_1|}{2}\left(-2\log \pi_1 + \log 2\pi + \log \sigma_1^2\right) - \frac{1}{2} \sum_{n \in \mathcal{D}_1} \frac{(l_n - \mu_1)^2}{\sigma_1^2}
\end{aligned}
$$

4. We solve for μ_0.

$$\frac{\partial \log p(\mathbf{l}|\mu_0,\sigma_0,\mu_1,\sigma_1)}{\partial \mu_0} = 0 - \frac{1}{2}\sum_{n\in\mathcal{D}_0}(-1)(2)\frac{l_n-\mu_0}{\sigma_0^2}$$

$$= \sum_{n\in\mathcal{D}_0}\frac{l_n-\mu_0}{\sigma_0^2}$$

$$\frac{\partial \log p(\mathbf{l}|\mu_0,\sigma_0,\mu_1,\sigma_1)}{\partial \mu_0} = 0$$

$$\sum_{n\in\mathcal{D}_0} l_n - \mu_0 = 0$$

$$|\mathcal{D}_0|\mu_0 = \sum_{n\in\mathcal{D}_0} l_n$$

$$\mu_0 = \frac{1}{|\mathcal{D}_0|}\sum_{n\in\mathcal{D}_0} l_n$$

And now for σ_0.

$$\frac{\partial \log p(\mathbf{l}|\mu_0,\sigma_0,\mu_1,\sigma_1)}{\partial \sigma_0} = \frac{-|\mathcal{D}_0|}{\sigma_0} + \sigma_0^{-3}\sum_{n\in\mathcal{D}_0}(l_n-\mu_0)^2$$

$$\frac{-|\mathcal{D}_0|}{\sigma} = -\sigma_0^{-3}\sum_{n\in\mathcal{D}_0}(l_n-\mu_0)^2$$

$$|\mathcal{D}_0|\sigma_0^2 = \sum_{n\in\mathcal{D}_0}(l_n-\mu_0)^2$$

$$\sigma_0^2 = \frac{1}{|\mathcal{D}_0|}\sum_{n\in\mathcal{D}_0}(l_n-\mu_0)^2$$

5. See below

6.

$$p(d=1|l_*,\mu_0,\sigma_0,\mu_1,\sigma_1) = \frac{\mathcal{N}(l_*|\mu_1,\sigma_1)\cdot\pi_1}{\sum_{i=0}^{1}\mathcal{N}(l_*|\mu_i,\sigma_i)\cdot\pi_i}$$

$$= \frac{1}{1+\underbrace{\frac{\mathcal{N}(l_*|\mu_0,\sigma_0)\cdot\pi_0}{\mathcal{N}(l_*|\mu_1,\sigma_1)\cdot\pi_1}}_{Z}}$$

$$Z = \frac{\frac{\pi_0}{\sigma_0}}{\frac{\pi_1}{\sigma_1}}\exp\left(\frac{-1}{2}\left(\frac{(l*-\mu_0)^2}{\sigma_0^2}-\frac{(l*-\mu_1)^2}{\sigma_1^2}\right)\right)$$

$$= \exp\left(\log\left(\frac{\sigma_1\mu_0}{\sigma_0\mu_1}\right)-\frac{1}{2}\left(\frac{(l*-\mu_0)^2}{\sigma_0^2}-\frac{(l*-\mu_1)^2}{\sigma_1^2}\right)\right)$$

$$= \exp(-a(l))$$

Where $a(l) = -\log\left(\frac{\sigma_1\mu_0}{\sigma_0\mu_1}\right)+\frac{1}{2}\left(\frac{(l*-\mu_0)^2}{\sigma_0^2}-\frac{(l*-\mu_1)^2}{\sigma_1^2}\right)$.

$$p(d=1|l_*,\mu_0,\sigma_0,\mu_1,\sigma_1) = \frac{1}{1+\exp(-a(l))}$$