Solutions to Homework Set Four

ECE 271A
Electrical and Computer Engineering
University of California San Diego
Nuno Vasconcelos

1.

a) The main difference with respect to what we have seen so far is that, in the regression problem everything is conditioned on the knowledge of x. That is, we have

$$P_{\mathbf{z}|\mathbf{x},\theta}(\mathbf{z}|\mathbf{x},\theta) = \mathcal{G}(\mathbf{\Phi}(\mathbf{x})\theta, \mathbf{\Sigma}).$$

We break down $\mathbf{T}$ into the $\mathbf{x}$ and $\mathbf{z}$ components, i.e. $\mathbf{T} = (\mathbf{T_z}, \mathbf{T_x})$, $\mathcal{D}_x$ ($\mathcal{D}_y$) being a sample of the random variable $\mathbf{T_x}$ ($\mathbf{T_y}$). Hence, for the posterior we have

$$\begin{aligned} P_{\theta|\mathbf{T}}(\theta|\mathcal{D}) &= P_{\theta|\mathbf{T_x},\mathbf{T_z}}(\theta|\mathcal{D}_x, \mathcal{D}_z) \\ &= \frac{P_{\mathbf{T_z}|\theta,\mathbf{T_x}}(\mathcal{D}_z|\theta,\mathcal{D}_x)P_{\theta|\mathbf{T_x}}(\theta|\mathcal{D}_x)}{\int P_{\mathbf{T_z}|\theta,\mathbf{T_x}}(\mathcal{D}_z|\theta,\mathcal{D}_x)P_{\theta|\mathbf{T_x}}(\theta|\mathcal{D}_x)}d\theta \\ &= \frac{P_{\mathbf{T_z}|\theta,\mathbf{T_x}}(\mathcal{D}_z|\theta,\mathcal{D}_x)P_{\theta}(\theta)}{\int P_{\mathbf{T_z}|\theta,\mathbf{T_x}}(\mathcal{D}_z|\theta,\mathcal{D}_x)P_{\theta}(\theta)}d\theta, \end{aligned}$$

and, therefore,

$$\begin{aligned} P_{\theta|\mathbf{T}}(\theta|\mathcal{D}) &\propto \exp\left\{-\frac{1}{2}\left[(\mathbf{z}-\mathbf{\Phi}\theta)^T\mathbf{\Sigma}^{-1}(\mathbf{z}-\mathbf{\Phi}\theta) + \theta^T\mathbf{\Gamma}^{-1}\theta\right]\right\} \\ &\propto \exp\left\{-\frac{1}{2}\left[\theta^T(\mathbf{\Phi}^T\mathbf{\Sigma}^{-1}\mathbf{\Phi} + \mathbf{\Gamma}^{-1})\theta - 2\theta^T\mathbf{\Phi}^T\mathbf{\Sigma}^{-1}\mathbf{z}\right]\right\}. \end{aligned}$$

This is the same as

$$\begin{aligned} P_{\theta|\mathbf{T}}(\theta|\mathcal{D}) &\propto \exp\left\{-\frac{1}{2}\left[(\theta-\mu_\theta)^T\mathbf{\Sigma}_\theta^{-1}(\theta-\mu_\theta)\right]\right\} \\ &\propto \exp\left\{-\frac{1}{2}\left[\theta^T\mathbf{\Sigma}_\theta^{-1}\theta - 2\theta^T\mathbf{\Sigma}_\theta^{-1}\mu_\theta\right]\right\} \end{aligned}$$

when

$$\begin{aligned} \mathbf{\Sigma}_\theta^{-1} &= \mathbf{\Phi}^T\mathbf{\Sigma}^{-1}\mathbf{\Phi} + \mathbf{\Gamma}^{-1} \\ \mu_\theta &= \mathbf{\Sigma}_\theta\mathbf{\Phi}^T\mathbf{\Sigma}^{-1}\mathbf{z} \\ &= (\mathbf{\Phi}^T\mathbf{\Sigma}^{-1}\mathbf{\Phi} + \mathbf{\Gamma}^{-1})^{-1}\mathbf{\Phi}^T\mathbf{\Sigma}^{-1}\mathbf{z}. \end{aligned}$$

It follows that

$$P_{\theta|\mathbf{T}}(\theta|\mathcal{D}) = \mathcal{G}\left(\theta, (\mathbf{\Phi}^T\mathbf{\Sigma}^{-1}\mathbf{\Phi} + \mathbf{\Gamma}^{-1})^{-1}\mathbf{\Phi}^T\mathbf{\Sigma}^{-1}\mathbf{z}, (\mathbf{\Phi}^T\mathbf{\Sigma}^{-1}\mathbf{\Phi} + \mathbf{\Gamma}^{-1})^{-1}\right).$$

For the predictive distribution we have

$$P_{z|\mathbf{T},x}(z|\mathcal{D},x) = \int \mathcal{G}(z, \phi^T\theta, \sigma(x)^2)\mathcal{G}(\theta, \mu_\theta, \mathbf{\Sigma}_\theta)d\theta$$

where $\sigma^2(x)$ is the variance associated with the particular value of x. This integral can be decomposed into

$$P_{z|\mathbf{T},x}(z|\mathcal{D},x) = \int_{\mu_z} \mathcal{G}(z,\mu_z,\sigma(x)^2) \int_{\{\theta|\phi^T\theta=\mu_z\}} \mathcal{G}(\theta,\mu_\theta,\mathbf{\Sigma}_\theta) d\theta d\mu_z.$$

Since the second term is the integral of a Gaussian outside a line, it is just the Gaussian resulting from the projection onto that line (i.e. the marginal distribution along the line), leading to

$$\begin{aligned} P_{z|\mathbf{T},x}(z|\mathcal{D},x) &= \int_{\mu_z} \mathcal{G}(z,\mu_z,\sigma(x)^2)\mathcal{G}(\mu_z,\phi^T\mu_\theta,\phi^T\mathbf{\Sigma}_\theta\phi) d\mu_z \\ &= \mathcal{G}(z,\phi^T\mu_\theta,\sigma(x)^2+\phi^T\mathbf{\Sigma}_\theta\phi) \end{aligned}$$

where we have also used the standard trick of the convolution as a sum of independent distributions on the second step.

An alternative way to solve this is to consider an intermediate random variable

$$\xi = \phi^T(x)\theta. \tag{1}$$

Note that, given x, this is a deterministic transformation of θ. Note, further, that the random variables are related by

$$z = \xi + \epsilon$$

and the predictive distribution for z can be written as

$$\begin{aligned} P_{z|\mathbf{T},x}(z|\mathcal{D},x) &= \int P_{z|\xi,\mathbf{T},x}(z|\xi,\mathcal{D}) P_{\xi|x,\mathbf{T}}(\xi|\mathcal{D}) d\xi \\ &= \int \mathcal{G}(z,\phi^T\theta,\sigma(x)^2) P_{\xi|x,\mathbf{T}}(\xi|\mathcal{D}) d\xi. \end{aligned}$$

Since $\theta|\mathbf{T}$ is Gaussian of parameters $\mu_\theta, \mathbf{\Sigma}_\theta$, it follows from (1) that $\xi|\mathbf{T},x$ is Gaussian of parameters $\phi(x)^T\mu_\theta$ and variance $\phi(x)^T\mathbf{\Sigma}_\theta\phi(x)$. Hence,

$$\begin{aligned} P_{z|\mathbf{T},x}(z|\mathcal{D},x) &= \int \mathcal{G}(z,\xi,\sigma(x)^2)\mathcal{G}(\xi,\phi(x)^T\mu_\theta,\phi(x)^T\mathbf{\Sigma}_\theta\phi(x)) d\xi \\ &= \mathcal{G}(z,\phi^T\mu_\theta,\sigma(x)^2+\phi^T\mathbf{\Sigma}_\theta\phi) \end{aligned}$$

where we have again used the convolution trick.

b) The MAP estimate is just the mean of the posterior, i.e.

$$\theta_{MAP} = (\mathbf{\Phi}^T\mathbf{\Sigma}^{-1}\mathbf{\Phi} + \mathbf{\Gamma}^{-1})^{-1}\mathbf{\Phi}^T\mathbf{\Sigma}^{-1}\mathbf{z},$$

and differs from the weighted least squares estimate only through the introduction of the term $\mathbf{\Gamma}^{-1}$ in the equation above. The role of this term is to regularize the solution. This can be easily seen by making $\mathbf{\Gamma}$ diagonal of the form

$$\mathbf{\Gamma} = \begin{bmatrix} \alpha\mathbf{I}_k & \mathbf{0} \\ \mathbf{0} & \beta\mathbf{I}_{K-k} \end{bmatrix}$$

in which case

$$\mathbf{\Gamma}^{-1} = \begin{bmatrix} \frac{1}{\alpha}\mathbf{I}_k & \mathbf{0} \\ \mathbf{0} & \frac{1}{\beta}\mathbf{I}_{K-k} \end{bmatrix}.$$

By controlling α and β we can then change the relative importance of the terms in $(\mathbf{\Phi}^T\mathbf{\Sigma}^{-1}\mathbf{\Phi}+\mathbf{\Gamma}^{-1})^{-1}$. For example, by making α large and β very small we can make $\mathbf{\Phi}^T\mathbf{\Sigma}^{-1}\mathbf{\Phi}+\mathbf{\Gamma}^{-1}$ "look" diagonal outside

of the first $k \times k$ entries, with a very large value in each diagonal entry after the position (k, k). This means that $(\mathbf{\Phi}^T\mathbf{\Sigma}^{-1}\mathbf{\Phi} + \mathbf{\Gamma}^{-1})^{-1}$ will be approximately equal to $(\mathbf{\Phi}^T\mathbf{\Sigma}^{-1}\mathbf{\Phi})^{-1}$ for the first $k \times k$ entries and basically equal to zero elsewhere which, in turn, implies that the contribution of the last $n - k$ terms of the vector $\mathbf{\Phi}^T\mathbf{\Sigma}^{-1}\mathbf{z}$ to the solution will be negligible.

Why is this important? Well, if you write down the expression for the entries of $\mathbf{\Phi}^T\mathbf{\Sigma}^{-1}\mathbf{\Phi}$ and $\mathbf{\Phi}^T\mathbf{\Sigma}^{-1}\mathbf{z}$ you will see that they go from low-order polynomial terms in x to high-order ones. For example, assuming $\mathbf{\Sigma} = \mathbf{I}$, we have

$$\mathbf{\Phi}^T\mathbf{\Sigma}^{-1}\mathbf{\Phi} = \mathbf{\Phi}^T\mathbf{\Phi} = \begin{bmatrix} \sum_i 1 & \dots & \sum_i x_i^K \\ \sum_i x_i & \dots & \sum_i x_i^{K+1} \\ & \vdots & \\ \sum_i x_i^K & \dots & \sum_i x_i^{2K} \end{bmatrix}$$

and

$$\mathbf{\Phi}^T\mathbf{\Sigma}^{-1}\mathbf{z} = \mathbf{\Phi}^T\mathbf{z} = \begin{bmatrix} \sum_i z_i \\ \sum_i x_i z_i \\ \vdots \\ \sum_i x_i^K z_i \end{bmatrix}.$$

Hence, the procedure above would make the contribution of the higher-order polynomial terms very small, therefore effectively reducing the degree of the regression model from K to k. Hence, the advantage of including a non-zero $\mathbf{\Gamma}^{-1}$ is that it allows us to control the complexity of our model explicitly. This is not possible in the standard weighted least squares formulation. Of course, in practice we would not want a "all or nothing" solution as above. But we could, for example, have a diagonal $\mathbf{\Gamma}$ where each entry controls the importance of the associated polynomial term in the regression model.

c) A polynomial of degree 25 is going to be extremely "wiggly" and it is hard to believe that it will make good predictions outside the training set. Hence, I would bias the solution against a polynomial of such a high-degree. Similarly, I would favor solutions that have low-order since most phenomena in nature do not require a polynomial of order superior to 3 or 4. Hence, my strategy would be to make the diagonal entries of $\mathbf{\Gamma}$ monotonically decreasing. Probably, I would make the entries corresponding to really high-orders, e.g. above 15, very close to zero. This would favor low-order polynomials.

In terms of the bias/variance trade-off, this would increase the bias and decrease the variance. To see that it would increase the bias it suffices to note that I would no longer be able to model a polynomial of degree 25 exactly. That is, the model would loose some of its expressive power, which is fine because it has too much. To see that the variance would decrease, it suffices to see that the parameters obtained with two small training sets would not vary as wildly as before (when fitting a high-order polynomial a change of one single point can lead to significant change of the parameter values, but for a line this is unlikely to happen). This is the reason why the procedure is called regularization. We are making the model less flexible and increasing its inertia. It is like gluing a piece of cloth onto a piece of cardboard, when you are trying to put up a banner. The cloth can still bend, but not as much as before (larger bias). On the other hand, if there is some wind (noise), the cloth is not going to create folds in response to it (less variance).

2. The exponential family and conjugate priors.

a) For a training set $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$,

$$\begin{aligned} P_{\mathbf{T}|\theta}(\mathcal{D}|\theta) &= \prod_{i=1}^{n} P_{\mathbf{X}|\theta}(\mathbf{x}_i|\theta) \\ &= \prod_{i=1}^{n} f(\mathbf{x}_i) g(\theta) e^{\phi(\theta)^T \mathbf{u}(\mathbf{x}_i)} \\ &= [g(\theta)]^n \left[\prod_{i=1}^{n} f(\mathbf{x}_i)\right] \exp\left\{\phi(\theta)^T \sum_{i=1}^{n} \mathbf{u}(\mathbf{x}_i)\right\}. \end{aligned}$$

The normalization constant is $[g(\theta)]^n$.

b) For a likelihood in the exponential family and a prior of the form

$$P_\theta(\theta) \propto g(\theta)^\eta e^{\phi(\theta)^T \nu}$$

the posterior is

$$\begin{aligned} P_{\theta|\mathbf{T}}(\theta|\mathcal{D}) &\propto [g(\theta)]^n \left[\prod_{i=1}^{n} f(\mathbf{x}_i)\right] \exp\left\{\phi(\theta)^T \sum_{i=1}^{n} \mathbf{u}(\mathbf{x}_i)\right\} g(\theta)^\eta \exp\left\{\phi(\theta)^T \nu\right\} \quad (2) \\ &\propto [g(\theta)]^{n+\eta} \left[\prod_{i=1}^{n} f(\mathbf{x}_i)\right] \exp\left\{\phi(\theta)^T \left(\nu + \sum_{i=1}^{n} \mathbf{u}(\mathbf{x}_i)\right)\right\} \quad (3) \\ &\propto [g(\theta)]^{n+\eta} \exp\left\{\phi(\theta)^T \left(\nu + \sum_{i=1}^{n} \mathbf{u}(\mathbf{x}_i)\right)\right\} \quad (4) \\ &= \frac{[g(\theta)]^{n+\eta} \exp\left\{\phi(\theta)^T \left(\nu + \sum_{i=1}^{n} \mathbf{u}(\mathbf{x}_i)\right)\right\}}{\int [g(\theta)]^{n+\eta} \exp\left\{\phi(\theta)^T \left(\nu + \sum_{i=1}^{n} \mathbf{u}(\mathbf{x}_i)\right)\right\} d\theta} \quad (5) \end{aligned}$$

Hence, the posterior has the same form of the prior with n replaced by $n + \eta$ and $\sum_{i=1}^{n} \mathbf{u}(\mathbf{x}_i)$ replaced by $\nu + \sum_{i=1}^{n} \mathbf{u}(\mathbf{x}_i)$. This shows, not only that $P_\theta(\theta)$ is *the conjugate prior for the exponential family*, but that the result of "propagating" this prior through the likelihood function is highly intuitive. In fact, the posterior is equivalent to the prior but incorporating the information provided by the training set. It leads to an intuitive interpretation of the prior parameters, namely

- η is a *virtual* number of samples that are added to the training set, leading to an extended training set of size $n + \eta$.
- ν is the value that is added, by this additional set of virtual samples, to the sufficient statistic.

This interpretation can be most clearly seen by comparing the MAP with ML estimate of θ. While for ML we have

$$\begin{aligned} \hat{\theta}_{ML} &= \arg\max_\theta P_{\mathbf{T}}(\mathcal{D}; \theta) \\ &= \arg\max_\theta [g(\theta)]^n \left[\prod_{i=1}^{n} f(\mathbf{x}_i)\right] \exp\left\{\phi(\theta)^T \left(\sum_{i=1}^{n} \mathbf{u}(\mathbf{x}_i)\right)\right\} \\ &= \arg\max_\theta [g(\theta)]^n \exp\left\{\phi(\theta)^T \left(\sum_{i=1}^{n} \mathbf{u}(\mathbf{x}_i)\right)\right\} \end{aligned}$$

for MAP

$$\begin{aligned}\hat{\theta}_{MAP} &= \arg\max_{\theta} P_{\theta|\mathbf{T}}(\theta|\mathcal{D}) \\ &= \arg\max_{\theta}[g(\theta)]^{n+\eta}\exp\left\{\phi(\theta)^T\left(\nu+\sum_{i=1}^{n}\mathbf{u}(\mathbf{x}_i)\right)\right\}.\end{aligned}$$

Hence, the *MAP estimate is exactly the same as the ML one, but on a extended training set, augmented with the virtual samples introduced by the inclusion of the prior.*

c)

i) Show that the likelihoods are in the exponential family. We start by the **Bernoulli**, whose likelihood can be written as

$$\begin{aligned}P_{\mathbf{T}|\theta}(\mathcal{D}|\theta) &= \theta^{\sum_i x_i}(1-\theta)^{n-\sum_i x_i} \\ &= \exp\left\{\left(\sum_i x_i\right)\log\theta+\left(n-\sum_i x_i\right)\log(1-\theta)\right\} \\ &= \exp\left\{n\log(1-\theta)+\log\frac{\theta}{(1-\theta)}\sum_i x_i\right\} \\ &= (1-\theta)^n\exp\left\{\log\frac{\theta}{(1-\theta)}\sum_i x_i\right\}\end{aligned}$$

and clearly is is the exponential family with

$$g(\theta)=1-\theta,\quad f(x)=1,\quad \phi(\theta)=\log\frac{\theta}{(1-\theta)}\quad\text{and}\quad u(x)=x.$$

For the **Poisson**

$$\begin{aligned}P_{\mathbf{T}|\theta}(\mathcal{D}|\theta) &= e^{-n\theta}\theta^{\sum_i x_i}\frac{1}{\prod_i x_i!} \\ &= [e^{-\theta}]^n\left[\prod_i\frac{1}{x_i!}\right]\exp\left\{\left(\sum_i x_i\right)\log\theta\right\}\end{aligned}$$

and we have an exponential density with

$$g(\theta)=e^{-\theta},\quad f(x)=\frac{1}{x!},\quad \phi(\theta)=\log\theta\quad\text{and}\quad u(x)=x.$$

For the **exponential**

$$P_{\mathbf{T}|\theta}(\mathcal{D}|\theta) = \theta^n\exp\{-\theta\sum_i x_i\}$$

and (as expected) we have an exponential density with

$$g(\theta)=\theta,\quad f(x)=1,\quad \phi(\theta)=-\theta\quad\text{and}\quad u(x)=x.$$

Finally, for the **Normal** of known mean μ

$$P_{\mathbf{T}|\theta}(\mathcal{D}|\theta) = \left(\sqrt{\frac{\theta}{2\pi}}\right)^n\exp\left\{-\frac{\theta}{2}\sum_i(x_i-\mu)^2\right\}$$

and we have an exponential density with

$$g(\theta) = \sqrt{\frac{\theta}{2\pi}}, \quad f(x) = 1, \quad \phi(\theta) = -\frac{\theta}{2} \quad \text{and} \quad \mathbf{u}(x) = (x-\mu)^2.$$

ii) Given that all likelihoods are in the exponential family, it suffices to show that the priors are functions of the form discussed in **b)** to show that they are conjugate priors.

The **Beta** distribution is

$$\begin{aligned} P_\theta(\theta) &\propto \theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &= \theta^{\alpha-1}(1-\theta)^{\beta-1+(\alpha-1)-(\alpha-1)} \\ &= \left(\frac{\theta}{1-\theta}\right)^{\alpha-1}(1-\theta)^{\beta+\alpha-2} \\ &= (1-\theta)^{\beta+\alpha-2}\exp\left\{(\alpha-1)\log\left(\frac{\theta}{1-\theta}\right)\right\} \\ &= (1-\theta)^{\eta}\exp\left\{\nu\log\left(\frac{\theta}{1-\theta}\right)\right\} \end{aligned}$$

with $\eta = \beta + \alpha - 2$ and $\nu = \alpha - 1$. Noting that the **Bernoulli** is the exponential distribution with

$$g(\theta) = 1-\theta, \quad \text{and} \quad \phi(\theta) = \log\frac{\theta}{1-\theta},$$

it follows from **b)** that the **Beta** distribution is a conjugate prior for the **Bernoulli** likelihood function.

The **Gamma** distribution can be written as

$$\begin{aligned} P_\theta(\theta) &\propto \theta^{\alpha-1}e^{-\beta\theta} \\ &= [e^-\theta]^{\beta}e^{(\alpha-1)\log\theta} \\ &= [e^-\theta]^{\eta}e^{\nu\log\theta} \end{aligned}$$

with $\eta = \beta$ and $\nu = \alpha - 1$. Noting that the **Poisson** is the exponential distribution with

$$g(\theta) = e^{-\theta}, \quad \text{and} \quad \phi(\theta) = \log\theta,$$

it follows from **b)** that the **Gamma** distribution is a conjugate prior for the **Poisson** likelihood function.

The **Gamma** distribution can also be written as

$$\begin{aligned} P_\theta(\theta) &\propto \theta^{\alpha-1}e^{-\beta\theta} \\ &= \theta^{\eta}e^{-\nu\theta} \end{aligned}$$

with $\eta = \alpha - 1$ and $\nu = \beta$. Noting that the **exponential** is the exponential distribution with

$$g(\theta) = \theta, \quad \text{and} \quad \phi(\theta) = -\theta,$$

it follows from **b)** that the **Gamma** distribution is a conjugate prior for the **exponential** likelihood function.

Finally, the **Gamma** distribution can be written as

$$\begin{aligned} P_\theta(\theta) &\propto \left(\frac{\theta}{2\pi}\right)^{\alpha-1} e^{-\beta\theta} \\ &= \left(\frac{\theta}{2\pi}\right)^{\frac{\eta}{2}} e^{-\frac{\theta}{2}\nu} \end{aligned}$$

with $\eta = 2(\alpha-1)$, and $\nu = 2\beta$. Noting that the **Normal** of known mean is the exponential distribution with

$$g(\theta) = \sqrt{\frac{\theta}{2\pi}} \quad \text{and} \quad \phi(\theta) = -\frac{\theta}{2}$$

it follows from **b)** that the **Gamma** distribution is a conjugate prior for the **Normal** likelihood function (when the mean is known).

iii) We have already seen what the general expression for the posterior is in (5). Here we simply have to plug-in the appropriate values of η and ν, as computed above. This leads to the following distributions

$$\begin{aligned}
\textbf{Bernoulli} \quad P_{\theta|\mathbf{T}}(\theta|\mathcal{D}) &\propto (1-\theta)^{n+\beta+\alpha-2} \exp\left\{\log\frac{\theta}{1-\theta}\left(\sum_i x_i + \alpha - 1\right)\right\} \\
&= (1-\theta)^{n+\beta+\alpha-2}\left(\frac{\theta}{1-\theta}\right)^{(\sum_i x_i+\alpha-1)} \\
&= (1-\theta)^{n+\beta-1-\sum_i x_i}\theta^{\sum_i x_i+\alpha-1} \\
&= \textbf{Beta}(\alpha + \sum_i x_i, \beta + (n - \sum_i x_i)) \\
\textbf{Poisson} \quad P_{\theta|\mathbf{T}}(\theta|\mathcal{D}) &\propto e^{-\theta(n+\beta)} \exp\left\{\left(\sum_i x_i + \alpha - 1\right)\log\theta\right\} \\
&= e^{-(n+\beta)\theta}\theta^{\sum_i x_i+\alpha-1} \\
&= \textbf{Gamma}(\alpha + \sum_i x_i, \beta + n) \\
\textbf{Exponential} \quad P_{\theta|\mathbf{T}}(\theta|\mathcal{D}) &\propto \theta^{n+\alpha-1} e^{-\theta(\sum_i x_i+\beta)} \\
&= \textbf{Gamma}(\alpha + n, \beta + \sum_i x_i) \\
\textbf{Normal} \quad P_{\theta|\mathbf{T}}(\theta|\mathcal{D}) &\propto \left(\sqrt{\frac{\theta}{2\pi}}\right)^{n+2(\alpha-1)} \exp\left\{-\frac{\theta}{2}\left(\sum_i (x_i-\mu)^2 + 2\beta\right)\right\} \\
&\propto \theta^{\frac{n}{2}+\alpha-1} e^{-\theta\left[\frac{1}{2}\sum_i (x_i-\mu)^2+\beta\right]} \\
&= \textbf{Gamma}(\alpha + \frac{n}{2}, \beta + \frac{1}{2}\sum_i (x_i-\mu)^2)
\end{aligned}$$

iv) For the **Bernoulli** case the sufficient statistic is the number of tosses in $\mathcal{D}$ for which the outcome was "1". The result of the propagation is the addition of $\beta+\alpha-2$ virtual samples to the training set, of which $\alpha-1$ are "1"s and $\beta-1$ are "0"s. This establishes a very intuitive interpretation of the parameters of the **Beta** prior. The **Poisson** models the number of arrivals in a given time interval. The sufficient statistic is the sum of all numbers of arrivals, which is equivalent to n times the average number of arrivals per observation. The result of the propagation is the addition of β virtual observations that

total $\alpha - 1$ arrivals. This, once again, gives an intuitive interpretation to the prior parameters: the prior encourages the arrival rate θ to become closer to the value $\frac{\alpha-1}{\beta}$. The **exponential** models the inter-arrival times of a Poisson process. The sufficient statistic is the sum of all observed times, which is equivalent to n times the average observed inter-arrival time. The result of the propagation is the addition $\alpha - 1$ virtual intervals with a cumulative inter-arrival time of β time intervals. Similarly to the the case of the **Poisson** likelihood, the prior encourages the arrival rate θ to become closer to the value $\frac{\alpha-1}{\beta}$. Note how the introduction of the conjugate prior maintains the consistency between the **Poisson** and **exponential** processes. For the **Normal** of known mean, the sufficient statistic is n times the sample variance. The result of the propagation is the addition $2(\alpha - 1)$ virtual samples that add 2β to the sample variance. Hence the prior encourages the sample variance to become closer to the value $\frac{\beta}{\alpha-1}$.

From these examples, we see that the selection of the prior parameters steers the posterior parameters towards a given value. This makes it significantly more intuitive to set up the prior. For example, it is usually much more intuitive to say "I believe that the variance of this Gaussian should be close to 1" than to say "I believe that this **Gamma** prior has parameters that satisfy $\beta = \alpha - 1$". Be careful however not to over-do this. Remember that the prior reflects your beliefs before the data is observed and it should not be interpreted as a hack that allows you to directly influence the posterior to be what you want, irrespectively of the data. This is definitely not what Bayesian inference is about and will get your paper rejected by any serious reviewer that understands Bayesian estimation.