# Exercise Sheet 9

Robin Schmidt
Probabilistic Inference & Learning

December 14, 2018

# Exponential Families

## 1. Famous Exponential Families

### (a)

*Proof.* The Bernoulli distribution is an exponential family

$$\begin{aligned}
p(x|f) &= f^x \cdot (1-f)^{1-x} \\
&= exp\left(\log(f^x \cdot (1-f)^{1-x})\right) \\
&= exp\left(\log(f^x) + \log((1-f)^{1-x})\right) \\
&= exp\left(x \cdot \log(f) + (1-x) \cdot \log(1-f)\right) \\
&= exp\left(x \cdot \log(f) + \log(1-f) - x \cdot \log(1-f)\right) \\
&= exp\left(x \cdot \log(\frac{f}{1-f}) + \log(1-f)\right)
\end{aligned}$$

We get the following parameters:

$$\begin{aligned}
\phi(x) &= [x] \\
w &= \log(\frac{f}{1-f}) \\
-\log Z(f) &= \log(1-f) \\
\log Z(f) &= -\log(1-f) \\
Z(f) &= e^{-\log(1-f)} = (1-f)^{-1} = \frac{1}{1-f}
\end{aligned}$$

Now if we want to compute an explicity one possible choice for sufficient statstics $\phi$, the natural parameters $w$ and the partition function $Z(w)$ we can just set $f = 0.5$ and $x = 1$ to get:

$$\begin{aligned}
\phi(x) &= [1] \\
w &= \log(\frac{0.5}{1-0.5}) = \log(1) = 0 \\
Z(f) = Z(w) &= \frac{1}{1-0.5} = 2
\end{aligned}$$

If we solve $w$ for $f$ we get:

$$\begin{aligned}
w &= \log(\frac{f}{1-f}) \\
e^w &= \frac{f}{1-f} \\
\frac{e^w}{1} &= \frac{f}{1-f} \\
\frac{1}{e^w} &= \frac{1-f}{f} \\
\frac{1}{e^w} &= \frac{1}{f} - \frac{f}{f} \\
\frac{1}{e^w} &= \frac{1}{f} - 1 \\
\frac{1}{e^w} + 1 &= \frac{1}{f} \\
\frac{1+e^w}{e^w} &= \frac{1}{f} \\
f &= \frac{e^w}{1+e^w}
\end{aligned}$$

Now we can use that in $\log Z(f)$ to write it in terms of the natural parameters and get:

$$\begin{aligned}
\log Z(w) &= -\log(1-f) \\
&= -\log(1 - \frac{e^w}{1+e^w}) \\
&= -\log(\frac{1+e^w}{1+e^w} - \frac{e^w}{1+e^w}) \\
&= -\log(\frac{1}{1+e^w}) \\
&= -\left(\log(1) - \log(1+e^w)\right) \\
&= \log(1+e^w)
\end{aligned}$$

$\square$

**(b)**

*Proof.* The Gamma distribution is an exponential family

$$
\begin{aligned}
p(x|\alpha, \beta) &= \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \\
&= exp\left(\log\left(\frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}\right)\right) \\
&= exp\left(\log\left(x^{\alpha-1} e^{-\beta x}\right) + \log\left(\frac{\beta^\alpha}{\Gamma(\alpha)}\right)\right) \\
&= exp\left(-\beta x + (\alpha - 1)\log(x) + \log\left(\frac{\beta^\alpha}{\Gamma(\alpha)}\right)\right)
\end{aligned}
$$

We get the following parameters:

$$
\begin{aligned}
\phi(x) &= \begin{bmatrix} x \\ \log(x) \end{bmatrix} \\
w &= \begin{bmatrix} -\beta \\ (\alpha - 1) \end{bmatrix} \\
-\log Z(\alpha, \beta) &= \log\left(\frac{\beta^\alpha}{\Gamma(\alpha)}\right) \\
\log Z(\alpha, \beta) &= -\log\left(\frac{\beta^\alpha}{\Gamma(\alpha)}\right) = -(\alpha \log(\beta) - \log(\Gamma(\alpha))) = \log(\Gamma(\alpha)) - \alpha \log(\beta) \\
Z(\alpha, \beta) &= e^{-\log\left(\frac{\beta^\alpha}{\Gamma(\alpha)}\right)} = \left(\frac{\beta^\alpha}{\Gamma(\alpha)}\right)^{-1} = \frac{\Gamma(\alpha)}{\beta^\alpha}
\end{aligned}
$$

Now if we want to compute an explicity one possible choice for sufficient statstics $\phi$, the natural parameters $w$ and the partition function $Z(w)$ we can just set $\alpha = 1$, $\beta = 1$ and $x = 1$ to get:

$$
\begin{aligned}
\phi(1) &= \begin{bmatrix} 1 \\ 0 \end{bmatrix} \\
w &= \begin{bmatrix} -1 \\ 0 \end{bmatrix} \\
Z(\alpha, \beta) = Z(w) &= \frac{\Gamma(1)}{1^1} = 1
\end{aligned}
$$

If we solve $w$ for $\alpha$ and $\beta$ each dependent on either $w_1$ or $w_2$ we get:

$$
\begin{aligned}
\beta &= -w_1 \\
\alpha &= w_2 + 1
\end{aligned}
$$

Now we can substitute these findings in $\log Z(\alpha, \beta)$ to write it in terms of the natural parameters and recieve:

$$
\begin{aligned}
\log Z(w) &= \log(\Gamma(\alpha)) - \alpha \log(\beta) \\
&= \log(\Gamma(w_2 + 1)) - (w_2 + 1)\log(-w_1)
\end{aligned}
$$

$\square$

**(c)**

*Proof.* The Dirichlet distribution is an exponential family

$$
\begin{aligned}
p(x|\alpha) &= \frac{1}{B(\alpha)} \prod_{i=1}^{n} x_i^{\alpha_i - 1} \\
&= exp\left(\log\left(\frac{1}{B(\alpha)} \prod_{i=1}^{n} x_i^{\alpha_i - 1}\right)\right) \\
&= exp\left(\log\left(\frac{1}{B(\alpha)}\right) + \log\left(\prod_{i=1}^{n} x_i^{\alpha_i - 1}\right)\right) \\
&= exp\left(\log\left(\frac{1}{B(\alpha)}\right) + \sum_{i=1}^{n} \log\left(x_i^{\alpha_i - 1}\right)\right) \\
&= exp\left(\log\left(\frac{1}{B(\alpha)}\right) + \sum_{i=1}^{n} (\alpha_i - 1) \cdot \log\left(x_i\right)\right) \\
&= exp\left(\sum_{i=1}^{n} (\alpha_i - 1) \cdot \log\left(x_i\right) + \log\left(\frac{1}{B(\alpha)}\right)\right)
\end{aligned}
$$

We get the following parameters:

$$
\begin{aligned}
\phi(x) &= \begin{bmatrix} \log x_1 \\ \vdots \\ \log x_n \end{bmatrix} \\
w &= \begin{bmatrix} \alpha_1 - 1 \\ \vdots \\ \alpha_n - 1 \end{bmatrix} \\
-\log Z(\alpha) &= \log\left(\frac{1}{B(\alpha)}\right) \\
\log Z(\alpha) &= -\log\left(\frac{1}{B(\alpha)}\right) = -(\log(1) - \log(B(\alpha))) = \log(B(\alpha)) \\
Z(\alpha) &= e^{-\log\left(\frac{1}{B(\alpha)}\right)} = B(\alpha)
\end{aligned}
$$

Now if we want to compute an explicity one possible choice for sufficient statstics $\phi$, the natural parameters $w$ and the partition function $Z(w)$ we can just set $\alpha = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$ and $x = \begin{bmatrix} \frac{1}{2} \\ \vdots \\ \frac{1}{2^n} \end{bmatrix}$ to get:

$$\phi(x) = \begin{bmatrix} \log \frac{1}{2} \\ \vdots \\ \log \frac{1}{2^n} \end{bmatrix}$$
$$w = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}$$
$$Z(\alpha) = Z(w) = B\left( \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \right)$$

If we solve $w$ for $\alpha_1$ to $\alpha_n$ each dependent on $w_1$ to $w_n$ we get:

$$\alpha = \begin{bmatrix} w_1 + 1 \\ \vdots \\ w_n + 1 \end{bmatrix}$$

Now we can substitute these findings in $\log Z(\alpha)$ to write it in terms of the natural parameters and recieve:

$$\begin{aligned} \log Z(w) &= \log(B(\alpha)) \\ &= \log(B(w + 1)) \end{aligned}$$

Keep in mind, that it is possible to write $B(\alpha)$ in terms of the Gamma function, but for the sake of simplicity, this doesn't get used in this exercise. $\square$

**(d)**

*Proof.* The multivariate Gaussian distribution is an exponential family

$$
\begin{aligned}
p(x|\mu,\Sigma) &= \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}\exp\left(-\frac{1}{2}(x-\mu)^\top\Sigma^{-1}(x-\mu)\right)\\
&= exp\left(\log\left(\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}\exp\left(-\frac{1}{2}(x-\mu)^\top\Sigma^{-1}(x-\mu)\right)\right)\right)\\
&= exp\left(-\log\left((2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}\right)+\log\left(\exp\left(-\frac{1}{2}(x-\mu)^\top\Sigma^{-1}(x-\mu)\right)\right)\right)\\
&= exp\left(-\log\left((2\pi)^{\frac{d}{2}}\right)-\log\left(|\Sigma|^{\frac{1}{2}}\right)-\frac{1}{2}(x-\mu)^\top\Sigma^{-1}(x-\mu)\right)\\
&= exp\left(-\frac{d}{2}\cdot\log(2\pi)-\frac{1}{2}\cdot\log(|\Sigma|)-\frac{1}{2}(x-\mu)^\top\Sigma^{-1}(x-\mu)\right)\\
&= exp\left(-\frac{1}{2}\left(d\cdot\log(2\pi)+\log(|\Sigma|)+(x-\mu)^\top\Sigma^{-1}(x-\mu)\right)\right)\\
&= exp\left(-\frac{1}{2}\left(d\cdot\log(2\pi|\Sigma|)+(x-\mu)^\top\Sigma^{-1}(x-\mu)\right)\right)\\
&= exp\left(-\frac{1}{2}\left(d\cdot\log(2\pi|\Sigma|)+x\Sigma^{-1}x^\top-x\Sigma^{-1}\mu^\top-x^\top\Sigma^{-1}\mu+\mu^\top\Sigma^{-1}\mu\right)\right)\\
&= exp\left(-\frac{1}{2}\left(d\cdot\log(2\pi|\Sigma|)+x\Sigma^{-1}x^\top-2\mu^\top\Sigma^{-1}x+\mu^\top\Sigma^{-1}\mu\right)\right)
\end{aligned}
$$

From this point on we need to use the relationship between the Frobenius inner product and the vectorizing operator which is given by:

$$
\begin{aligned}
x^\top\Sigma^{-1}x &= \Sigma^{-1} : xx^\top\\
&= \operatorname{vec}\left(\Sigma^{-1}\right)^\top\operatorname{vec}\left(xx^\top\right)\\
\mu^\top\Sigma^{-1}x &= \left(\Sigma^{-1}\mu\right)^\top x
\end{aligned}
$$

This leads to the expression:

$$
p(x|\mu,\Sigma) = exp\left(-\frac{1}{2}\left(d\cdot\log(2\pi|\Sigma|)+\operatorname{vec}\left(\Sigma^{-1}\right)^\top\operatorname{vec}\left(xx^\top\right)-2\left(\Sigma^{-1}\mu\right)^\top x+\mu^\top\Sigma^{-1}\mu\right)\right)
$$

We get the following parameters if we keep in mind that $-\frac{1}{2}$ is still outside of the brackets, which influences the weights $w$:

$$
\begin{aligned}
\phi(x) &= \begin{bmatrix} x \\ \operatorname{vec}\left(xx^\top\right) \end{bmatrix} \\
w &= \begin{bmatrix} \left(\Sigma^{-1}\mu\right)^\top \\ -\frac{1}{2}\operatorname{vec}\left(\Sigma^{-1}\right) \end{bmatrix} \\
-\log\ Z(\mu,\Sigma) &= -\frac{1}{2}d\cdot\log\left(2\pi|\Sigma|\right) - \frac{1}{2}\mu^\top\Sigma^{-1}\mu \\
\log\ Z(\mu,\Sigma) &= \frac{1}{2}d\cdot\log\left(2\pi|\Sigma|\right) + \frac{1}{2}\mu^\top\Sigma^{-1}\mu \\
Z(\mu,\Sigma) &= exp\left(\frac{1}{2}d\cdot\log\left(2\pi|\Sigma|\right)\right)\cdot exp\left(\frac{1}{2}\mu^\top\Sigma^{-1}\mu\right) \\
&= \left(2\pi|\Sigma|\right)^{\frac{d}{2}}\cdot exp\left(\frac{1}{2}\mu^\top\Sigma^{-1}\mu\right)
\end{aligned}
$$

If we solve $w$ for $\mu$ and $\Sigma$ we get:

$$
\begin{aligned}
w_2 &= -\frac{1}{2}\Sigma^{-1} \\
\Sigma &= (-2w_2)^{-1} \\
w_1 &= \left(\Sigma^{-1}\mu\right)^\top \\
\mu &= -\frac{w_1^\top}{2w_2}
\end{aligned}
$$

Now we can substitute these findings in $\log Z(\mu,\Sigma)$ to write it in terms of the natural parameters and recieve:

$$
\begin{aligned}
\log\ Z(w) &= \frac{1}{2}d\cdot\log\left(2\pi|(-2w_2)^{-1}|\right) + \frac{1}{2}(-\frac{w_1^\top}{2w_2})^\top((-2w_2)^{-1})^{-1}(-\frac{w_1^\top}{2w_2}) \\
&= \frac{1}{2}d\cdot\log\left(-4\pi\frac{1}{|w_2|}\right) + (-\frac{w_1}{2w_2})(-w_2)(-\frac{w_1^\top}{2w_2}) \\
&= \frac{1}{2}d\cdot\log\left(-4\pi\frac{1}{|w_2|}\right) - \frac{w_1w_2w_1^\top}{4w_2^3} \\
&= -\frac{1}{2}d\log\left(-4\pi|w_2|\right) - \frac{w_1w_2w_1^\top}{4w_2^3}
\end{aligned}
$$

Now if we want to compute an explicity one possible choice for sufficient statstics $\phi$, the natural parameters $w$ and the partition function $Z(w)$ we can just set $\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, $\mu =$

$\begin{bmatrix} 1 \\ 1 \end{bmatrix}$ and $x = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ to get:

$$\begin{aligned}
Z(\mu, \Sigma) = Z(w) &= (2\pi|\Sigma|)^{\frac{d}{2}} \cdot exp\left(\frac{1}{2}\mu^\top \Sigma^{-1} \mu\right) \\
&= (2\pi)^{\frac{d}{2}} \cdot e^{\frac{1}{2}} \\
w_1 &= \begin{bmatrix} 1 \\ 1 \end{bmatrix}^\top \\
w_2 &= \begin{bmatrix} -\frac{1}{2} \\ 0 \\ 0 \\ -\frac{1}{2} \end{bmatrix} \\
\phi_1 &= \begin{bmatrix} 1 \\ 1 \end{bmatrix} \\
\phi_2 &= \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}
\end{aligned}$$

□

## 2. Maximum Likelihood Inference

$$\frac{1}{m} \sum_i \phi(x_i) = \nabla_w \log Z(w)$$

### (a)

If we recall $\log Z(w)$ in terms of the natural parameters already computed in exercise 1 (for detailed computation please see exercise 1) we get:

$$\begin{aligned}
w_{ML} &= \nabla_w \log Z(w) \\
&= \nabla_w \log(1 + e^w) \\
&= \frac{e^w}{e^w + 1}
\end{aligned}$$

Since here $w = \log\left(\frac{f}{1-f}\right)$ we can rewrite this in terms of the standard parameters instead of the natural parameters, which gives:

$$\begin{aligned} w_{ML} &= \frac{e^{\log\left(\frac{f}{1-f}\right)}}{e^{\log\left(\frac{f}{1-f}\right)}+1} \\ &= \frac{\frac{f}{1-f}}{\frac{f}{1-f}+1} \\ &= \frac{f}{1-f}\cdot\frac{1-f}{f+1-f} \\ &= \frac{f-f^2}{1-f} \end{aligned}$$

**(d)**

If we recall $\log Z(w)$ in terms of the natural parameters already computed in exercise 1 (for detailed computation please see exercise 1) we get:

$$\begin{aligned} w_{ML} &= \nabla_w \log Z(w) \\ &= \nabla_w \left(-\frac{1}{2}d\log\left(-4\pi|w_2|\right) - \frac{w_1 w_2 w_1^\top}{4w_2^3}\right) \\ &= \begin{bmatrix} \frac{\partial}{\partial w_1}\left(-\frac{1}{2}d\log\left(-4\pi|w_2|\right) - \frac{w_1 w_2 w_1^\top}{4w_2^3}\right) \\ \frac{\partial}{\partial w_2}\left(-\frac{1}{2}d\log\left(-4\pi|w_2|\right) - \frac{w_1 w_2 w_1^\top}{4w_2^3}\right) \end{bmatrix} \\ &= \begin{bmatrix} -(w_2 + w_2^\top)w_1\cdot(4w_2^3)^{-1} \\ \frac{\partial}{\partial w_2}\left(-\frac{1}{2}d\log\left(-4\pi|w_2|\right) - \frac{w_1 w_2 w_1^\top}{4w_2^3}\right) \end{bmatrix} \end{aligned}$$

## 3. Conjugate Prior Inference

### (a)

$$
\begin{aligned}
p(x_i|\sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right) \\
&= exp\left(\log\left(\frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)\right)\right) \\
&= exp\left(\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) + \log\left(exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)\right)\right) \\
&= exp\left(-\log\left(\sqrt{2\pi\sigma^2}\right) - \frac{(x_i-\mu)^2}{2\sigma^2}\right) \\
&= exp\left(-\log\left(\sqrt{2\pi\sigma^2}\right) - \left(\frac{x_i^2 - 2x_i\mu + \mu^2}{2\sigma^2}\right)\right) \\
&= exp\left(-\log\left(\sqrt{2\pi\sigma^2}\right) - \left(\frac{x_i^2}{2\sigma^2} - \frac{2x_i\mu}{2\sigma^2} + \frac{\mu^2}{2\sigma^2}\right)\right) \\
&= exp\left(-\frac{1}{2\sigma^2}x_i^2 + \frac{\mu}{\sigma^2}x_i + \frac{\mu^2}{2\sigma^2} - \log\left(\sqrt{2\pi\sigma^2}\right)\right)
\end{aligned}
$$

We get the following parameters:

$$
\begin{aligned}
\phi(x) &= \begin{bmatrix} x^2 \\ x \end{bmatrix} \\
w &= \begin{bmatrix} -\frac{1}{2\sigma^2} \\ \frac{\mu}{\sigma^2} \end{bmatrix} \\
-\log Z(\mu, \sigma^2) &= \frac{\mu^2}{2\sigma^2} - \log\left(\sqrt{2\pi\sigma^2}\right) \\
\log Z(\mu, \sigma^2) &= -\frac{\mu^2}{2\sigma^2} + \log\left(\sqrt{2\pi\sigma^2}\right)
\end{aligned}
$$

We can now solve $w$ for $\mu$ and $\sigma$ to get:

$$
\begin{aligned}
w_1 &= -\frac{1}{2\sigma^2} \\
\sigma^2 &= -\frac{1}{2w_1} \\
w_2 &= \frac{\mu}{\sigma^2} \\
\mu &= w_2\sigma^2 = -\frac{w_2}{2w_1}
\end{aligned}
$$

This leads to:

$$
\begin{aligned}
\log Z(w) &= -\frac{\left(-\frac{w_2}{2w_1}\right)^2}{2\left(-\frac{1}{2w_1}\right)} + \log\left(\sqrt{2\pi\left(-\frac{1}{2w_1}\right)}\right) \\
&= -\frac{\frac{w_2^2}{4w_1^2}}{\frac{1}{w_1}} + \log\left(\sqrt{2\pi\left(-\frac{1}{2w_1}\right)}\right) \\
&= -\frac{w_2^2}{4w_1} + \log\left(\sqrt{2\pi\left(-\frac{1}{2w_1}\right)}\right)
\end{aligned}
$$

### (b)

Using the formulas shown in the lecture for exponential families, which are given by:

$$
\begin{aligned}
x \sim p_w(x|w) &= \exp\left[\phi(x)^\top w - \log Z(w)\right] \\
p_\alpha(w|\alpha,\nu) &= \exp\left[\begin{pmatrix} \alpha \\ \nu \end{pmatrix}^\top \begin{pmatrix} w \\ \log Z(W) \end{pmatrix} - \log F(\alpha,\nu)\right] \\
p_\alpha(w|\alpha,\nu)\prod_{i=1}^{n} p_w\left(x_i|w\right) &\propto p_\alpha(w|\alpha + \sum_i \phi\left(x_i\right), \nu + n)
\end{aligned}
$$

Applying them for our use case gets us to:

$$
\begin{aligned}
p(w|x_1,\ldots,x_m) &= \mathcal{G}(w;\alpha,\beta)\prod_{i=1}^{m} p_w\left(x_i|w=\sigma^{-2}\right) \propto \mathcal{G}(w|\underbrace{\alpha + \sum_i \phi\left(x_i\right)}_{\alpha_m}, \underbrace{\nu + m}_{\beta_m}) \\
\mathcal{G}(w;\alpha,\beta) &= exp\left[\begin{bmatrix} x \\ \log(x) \end{bmatrix}^\top \begin{bmatrix} -\beta \\ (\alpha-1) \end{bmatrix} - \left(\log(\Gamma(w_2+1)) - (w_2+1)\log(-w_1)\right)\right] \\
p_w\left(x_i|w=\sigma^{-2}\right) &= exp\left[\begin{bmatrix} x^2 \end{bmatrix}\begin{bmatrix} \sigma^{-2} \end{bmatrix} - \left(-\frac{w_2^2}{4w_1} + \log\left(\sqrt{2\pi\left(-\frac{1}{2w_1}\right)}\right)\right)\right]
\end{aligned}
$$

This yields the following parameters:

$$
\begin{aligned}
\alpha_m &= x + \sum_i x_i^2 \\
\beta_m &= \log(x) + m
\end{aligned}
$$