Comp 0086 Assignment 1

Alexandra Maria Proca (SN: 20047328)

October 21, 2020

Question 1.

I will begin by proving that the derivative of the log partition is equal to the expectation of the sufficient statistic, as this will be used to find the E[T(x)] for some of the distributions in this problem.

$The\ first\ derivative\ of\ the\ log\ partition\ A(\theta)\ is\ equal\ to\ the\ expected\ value\ of\ the\ sufficient\ statistic.$
$The\ total\ probability\ of\ the\ exponential\ family\ from\ -\infty\ to\ \infty = 1:$
$\int_{-\infty}^{\infty} f(\mathbf{x})g(\theta)e^{\phi(\theta)^{\mathrm{T}}\mathbf{T}(\mathbf{x})}\frac{\partial}{\partial x} = 1$
$g(\theta)\ can\ be\ rewritten\ as\ \mathrm{e}^{\log g(\theta)},\ which\ then\ becomes\ \mathrm{e}^{-A(\theta)}\ (A(\theta) = -\log g(\theta)).\ A(\theta)\ is\ defined\ as\ the\ log\ partition\ in\ terms\ of\ the\ conventional\ parameters.\ The\ exponential\ family\ can\ be\ rewritten\ to\ include\ it:$
$\int_{-\infty}^{\infty} f(\mathbf{x})e^{\phi(\theta)^{\mathrm{T}}\mathbf{T}(\mathbf{x})-A(\theta)}\frac{\partial}{\partial x} = 1$
$which\ can\ be\ rewritten\ in\ terms\ of\ A(\theta):$
$\int_{-\infty}^{\infty} f(\mathbf{x})e^{\phi(\theta)^{\mathrm{T}}\mathbf{T}(\mathbf{x})}e^{-A(\theta)}\frac{\partial}{\partial x}$
$= e^{-A(\theta)}\int_{-\infty}^{\infty} f(\mathbf{x})e^{\phi(\theta)^{\mathrm{T}}\mathbf{T}(\mathbf{x})}\frac{\partial}{\partial x} = 1$
$\log\left(e^{-A(\theta)}\int_{-\infty}^{\infty} f(\mathbf{x})e^{\phi(\theta)^{\mathrm{T}}\mathbf{T}(\mathbf{x})}\frac{\partial}{\partial x}\right) = \log 1$
$\log e^{-A(\theta)} + \log\int_{-\infty}^{\infty} f(\mathbf{x})e^{\phi(\theta)^{\mathrm{T}}\mathbf{T}(\mathbf{x})}\frac{\partial}{\partial x} = 0$
$-A(\theta) + \log\int_{-\infty}^{\infty} f(\mathbf{x})e^{\phi(\theta)^{\mathrm{T}}\mathbf{T}(\mathbf{x})}\frac{\partial}{\partial x} = 0$
$A(\theta) = \log\int_{-\infty}^{\infty} f(\mathbf{x})e^{\phi(\theta)^{\mathrm{T}}\mathbf{T}(\mathbf{x})}\frac{\partial}{\partial x}$
$From\ here,\ the\ first\ derivative\ of\ the\ log\ partition\ can\ be\ found\ as\ the\ expected\ value\ of\ the\ sufficient\ statistic,\ \mathbb{E}[\mathbf{T}(x)].$
$\frac{\partial A(\theta)}{\partial\theta} = \left(\log\int_{-\infty}^{\infty} f(\mathbf{x})e^{\phi(\theta)^{\mathrm{T}}\mathbf{T}(\mathbf{x})}\frac{\partial}{\partial x}\right)\frac{\partial}{\partial\theta}$
$Taking\ the\ derivative\ with\ respect\ to\ \theta,$
$= \frac{\left(\int\limits_{-\infty}^{\infty} f(\mathbf{x})e^{\phi(\theta)^{\mathrm{T}}\mathbf{T}(\mathbf{x})}\frac{\partial}{\partial x}\right)\frac{\partial}{\partial\theta}}{\int_{-\infty}^{\infty} f(\mathbf{x})e^{\phi(\theta)^{\mathrm{T}}\mathbf{T}(\mathbf{x})}\frac{\partial}{\partial x}}$
$= \frac{\int\limits_{-\infty}^{\infty} f(\mathbf{x})e^{\phi(\theta)^{\mathrm{T}}\mathbf{T}(\mathbf{x})}\mathbf{T}(\mathbf{x})\frac{\partial}{\partial x}}{\int_{-\infty}^{\infty} f(\mathbf{x})e^{\phi(\theta)^{\mathrm{T}}\mathbf{T}(\mathbf{x})}\frac{\partial}{\partial x}}$
$Because\ A(\theta) = \log\int_{-\infty}^{\infty} f(\mathbf{x})e^{\phi(\theta)^{\mathrm{T}}\mathbf{T}(\mathbf{x})}\frac{\partial}{\partial x},\ e^{A(\theta)} = \int_{-\infty}^{\infty} f(\mathbf{x})e^{\phi(\theta)^{\mathrm{T}}\mathbf{T}(\mathbf{x})}\frac{\partial}{\partial x}\ .\ Replacing\ the\ denominator\ with\ e^{A(\theta)},$
$= \frac{\int\limits_{-\infty}^{\infty} f(\mathbf{x})e^{\phi(\theta)^{\mathrm{T}}\mathbf{T}(\mathbf{x})}\mathbf{T}(\mathbf{x})\frac{\partial}{\partial x}}{e^{A(\theta)}}$
$= \int\limits_{-\infty}^{\infty} f(\mathbf{x})e^{\phi(\theta)^{\mathrm{T}}\mathbf{T}(\mathbf{x})-A(\theta)}\mathbf{T}(\mathbf{x})\frac{\partial}{\partial x}$
$f(\mathbf{x})e^{\phi(\theta)^{\mathrm{T}}\mathbf{T}(\mathbf{x})-A(\theta)}\ is\ the\ exponential\ family.\ Thus,$
$= \int\limits_{-\infty}^{\infty} p(\mathbf{x}|\theta)\mathbf{T}(\mathbf{x})\frac{\partial}{\partial x}$
$This\ is\ the\ definition\ of\ the\ expectation,\ \mathbb{E}[\mathbf{T}(\mathbf{x})] = \sum_{n=1}^{N} p(\mathbf{x}_i|\theta)\mathbf{T}(\mathbf{x}_i) = \int\limits_{-\infty}^{\infty} p(\mathbf{x}|\theta)\mathbf{T}(x)\frac{\partial}{\partial x}$
$Thus,$
$\frac{\partial A(\theta)}{\partial\theta} = \mathbb{E}[\mathbf{T}(x)]$

Multivariate Normal

a.

$p(\mathbf{x}|\mu,\Sigma) = |2\pi\Sigma|^{-1/2}e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}$
$= |2\pi\Sigma|^{-1/2}e^{-\frac{1}{2}\left(\boldsymbol{x}^{\mathrm{T}}-\boldsymbol{\mu}^{\mathrm{T}}\right)\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}$
$= |2\pi\Sigma|^{-1/2}e^{-\frac{1}{2}\left(\boldsymbol{x}^{\mathrm{T}}-\boldsymbol{\mu}^{\mathrm{T}}\right)\left(\Sigma^{-1}\boldsymbol{x}-\Sigma^{-1}\boldsymbol{\mu}\right)}$
$= |2\pi\Sigma|^{-1/2}e^{-\frac{1}{2}\left(\boldsymbol{x}^{\mathrm{T}}\Sigma^{-1}\boldsymbol{x}-\boldsymbol{x}^{\mathbf{T}}\boldsymbol{\Sigma}^{-\mathbf{1}}\boldsymbol{\mu}-\boldsymbol{\mu}^{\mathrm{T}}\Sigma^{-1}\boldsymbol{x}+\boldsymbol{\mu}^{\mathrm{T}}\Sigma^{-1}\boldsymbol{\mu}\right)}$
Because of the general rule that $\boldsymbol{x}^{\mathrm{T}}A\boldsymbol{y} = \boldsymbol{y}^{\mathrm{T}}A\boldsymbol{x}$, $\boldsymbol{x}^{\mathrm{T}}\Sigma^{-1}\boldsymbol{\mu} = \boldsymbol{\mu}^{\mathrm{T}}\Sigma^{-1}\boldsymbol{x}$ *and they can be summed.*
$= |2\pi\Sigma|^{-1/2}e^{-\frac{1}{2}\left(\boldsymbol{x}^{\mathrm{T}}\Sigma^{-1}\boldsymbol{x}-\mathbf{2}\boldsymbol{\mu}^{\mathrm{T}}\Sigma^{-1}\boldsymbol{x}+\boldsymbol{\mu}^{\mathrm{T}}\Sigma^{-1}\boldsymbol{\mu}\right)}$
$= |2\pi\Sigma|^{-1/2}e^{\boldsymbol{\mu}^{\mathrm{T}}\Sigma^{-1}\boldsymbol{x}-\frac{1}{2}\boldsymbol{x}^{\mathrm{T}}\Sigma^{-1}\boldsymbol{x}-\frac{\mathbf{1}}{\mathbf{2}}\boldsymbol{\mu}^{\mathrm{T}}\Sigma^{-1}\boldsymbol{\mu}}$
$= |2\pi\Sigma|^{-1/2}e^{-\frac{\mathbf{1}}{\mathbf{2}}\boldsymbol{\mu}^{\mathrm{T}}\Sigma^{-1}\boldsymbol{\mu}}e^{\boldsymbol{\mu}^{\mathrm{T}}\Sigma^{-1}\boldsymbol{x}-\frac{1}{2}\boldsymbol{x}^{\mathrm{T}}\Sigma^{-1}\boldsymbol{x}}$
Because $\boldsymbol{x}$ *is a vector and* Σ^{-1} *is a square matrix,*
$= |2\pi\Sigma|^{-1/2}e^{-\frac{\mathbf{1}}{\mathbf{2}}\boldsymbol{\mu}^{\mathrm{T}}\Sigma^{-1}\boldsymbol{\mu}}e^{\boldsymbol{\mu}^{\mathrm{T}}\Sigma^{-1}\boldsymbol{x}-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n}\boldsymbol{x}_i\Sigma_{ij}^{-1}\boldsymbol{x}_j}$
$= |2\pi\Sigma|^{-1/2}e^{-\frac{\mathbf{1}}{\mathbf{2}}\boldsymbol{\mu}^{\mathrm{T}}\Sigma^{-1}\boldsymbol{\mu}}e^{\boldsymbol{\mu}^{\mathrm{T}}\Sigma^{-1}\boldsymbol{x}+\mathrm{Tr}\left(-\frac{1}{2}\boldsymbol{x}^{\mathrm{T}}\Sigma^{-1}\boldsymbol{x}\right)}$
Due to the cyclic property of traces,
$= |2\pi\Sigma|^{-1/2}e^{-\frac{\mathbf{1}}{\mathbf{2}}\boldsymbol{\mu}^{\mathrm{T}}\Sigma^{-1}\boldsymbol{\mu}}e^{\boldsymbol{\mu}^{\mathrm{T}}\Sigma^{-1}\boldsymbol{x}+\mathrm{Tr}\left(-\frac{1}{2}\Sigma^{-1}\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\right)}$
Σ^{-1} *and* $\boldsymbol{x}\boldsymbol{x}^{T}$*can be vectorized so that their dot product equals their trace*
$= |2\pi\Sigma|^{-1/2}e^{-\frac{\mathbf{1}}{\mathbf{2}}\boldsymbol{\mu}^{\mathrm{T}}\Sigma^{-1}\boldsymbol{\mu}}e^{\boldsymbol{\mu}^{\mathrm{T}}\Sigma^{-1}\boldsymbol{x}+vec\left(-\frac{1}{2}\Sigma^{-1}\right)^{\mathrm{T}}vec\left(\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\right)}$
$= |2\pi\Sigma|^{-1/2}e^{-\frac{\mathbf{1}}{\mathbf{2}}\boldsymbol{\mu}^{\mathrm{T}}\Sigma^{-1}\boldsymbol{\mu}}e^{\begin{bmatrix}\boldsymbol{\mu}^{\mathrm{T}}\Sigma^{-1} & vec\left(-\frac{1}{2}\Sigma^{-1}\right)^{\mathrm{T}}\end{bmatrix}\begin{bmatrix}\boldsymbol{x}\\ vec\left(\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\right)\end{bmatrix}}$
Given the form of the exponential family, $p(x|\theta) = f(x)g(\theta)e^{\phi(\theta)^{\mathrm{T}}\mathbf{T}(x)}$,
the natural parameters are
$\phi(\theta) = \begin{bmatrix}\boldsymbol{\mu}^{\mathrm{T}}\Sigma^{-1}\\ vec\left(-\frac{1}{2}\Sigma^{-1}\right)^{\mathrm{T}}\end{bmatrix}$
and the sufficient statistic is
$\mathbf{T}(\mathbf{x}) = \begin{bmatrix}\boldsymbol{x}\\ vec\left(\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\right)\end{bmatrix}$

b.

$\mathbb{E}[\mathbf{T}(x)] = \mathbb{E}\begin{bmatrix}\boldsymbol{x}\\ vec\left(\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\right)\end{bmatrix}$
$\mathbb{E}[\boldsymbol{x}] = \int_{-\infty}^{\infty}|2\pi\Sigma|^{-1/2}e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}\boldsymbol{x}dx$
$= |2\pi\Sigma|^{-1/2}\int_{-\infty}^{\infty}e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}\boldsymbol{x}dx$
Let $\boldsymbol{y} = \boldsymbol{x}-\boldsymbol{\mu}$
$= |2\pi\Sigma|^{-1/2}\int_{-\infty}^{\infty}e^{-\frac{1}{2}(\boldsymbol{y})^{\mathrm{T}}\Sigma^{-1}(\boldsymbol{y})}(\boldsymbol{y}+\boldsymbol{\mu})dy$
Because the integral goes from $-\infty$ *to* ∞, *the* $\boldsymbol{y}$ *term in* $(\boldsymbol{y}+\boldsymbol{\mu})$ *will go to* 0 *from symmetry* $(-y+y=0)$
$= |2\pi\Sigma|^{-1/2}\int_{-\infty}^{\infty}e^{-\frac{1}{2}(\boldsymbol{y})^{\mathrm{T}}\Sigma^{-1}(\boldsymbol{y})}\boldsymbol{\mu}dy$
$(\boldsymbol{y})^{\mathrm{T}}\Sigma^{-1}(\boldsymbol{y})$ *is an even function. Thus by integrating,*
$= \boldsymbol{\mu}$

$\mathbb{E}\left[vec\left(\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\right)\right]$
For all $\left(\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\right)_{ij}$*, the expectation is equal to that at* $\mathbb{E}\left[\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\right]_{ij}$*. Thus,*
$= \mathbb{E}\left[\left(\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\right)_{ij}\right]$
$= \mathbb{E}[x_i x_j]$
It is known that $cov[X,Y] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$*. Thus the expression can be rewritten as*
$= cov[x_i, x_j] + \mathbb{E}[x_i]\mathbb{E}[x_j]$
Because the covariance matrix at ij *is the covariance between* x_i *and* x_j*, and because* $\mathbb{E}[\boldsymbol{x}] = \boldsymbol{\mu}$*,*
$= \Sigma_{ij} + \mu_i \mu_j$
Thus for all ij*,*

$$\mathbb{E}\left[vec\left(\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\right)\right] = \begin{bmatrix} \Sigma_{11} + \mu_1\mu_1 \\ \Sigma_{12} + \mu_1\mu_2 \\ \vdots \\ \Sigma_{DD} + \mu_D\mu_D \end{bmatrix} = vec\left(\Sigma + \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}}\right)$$

$$\mathbb{E}[\mathbf{T}(x)] = \begin{bmatrix} \boldsymbol{\mu} \\ vec\left(\Sigma + \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}}\right) \end{bmatrix}$$

Binomial

a.

$$p(x|p) = \binom{N}{x} p^x (1-p)^{(N-x)}$$

It is known that for any $x, x = e^{\log(x)}$

$$= e^{\log\left[\binom{N}{x} p^x (1-p)^{(N-x)}\right]}$$

$$= e^{\log\binom{N}{x} + \log(p^x) + \log\left((1-p)^{(N-x)}\right)}$$

$$= e^{\log\binom{N}{x} + x\log(p) + (N-x)\log(1-p)}$$

$$= e^{\log\binom{N}{x} + x\log(p) + (N-x)\log(1-p)}$$

$$= e^{\log\binom{N}{x} + x\log(p) + N\log(1-p) - x\log(1-p)}$$

The terms that do not contain both x *and* p *can be removed from the exponent*

$$= \binom{N}{x} (1-p)^N e^{x\log(p) - x\log(1-p)}$$

$$= \binom{N}{x} (1-p)^N e^{x\log\left(\frac{p}{1-p}\right)}$$

$$= \binom{N}{x} (1-p)^N e^{\left[\log\left(\frac{p}{1-p}\right)\right]^{\mathrm{T}}[x]}$$

Given the form of the exponential family, $p(x|\theta) = f(x)g(\theta)e^{\phi(\theta)^{\mathrm{T}}\mathbf{T}(x)}$*,*
the natural parameter is

$$\phi(\theta) = \left[\log\left(\frac{p}{1-p}\right)\right]$$

and the sufficient statistic is

$$\mathbf{T}(x) = [x]$$

b.

$\mathbb{E}[\mathbf{T}(x)] = \mathbb{E}[x] = \sum_{x=0}^{N} x \binom{N}{x} p^x (1-p)^{N-x}$

$When\ x = 0,\ the\ first\ term\ is\ 0,\ so\ it\ can\ be\ removed\ from\ the\ summation$

$= \sum_{x=1}^{N} x \binom{N}{x} p^x (1-p)^{N-x}$

$= \sum_{x=1}^{N} x \frac{N!}{(N-x)!x!} p^x (1-p)^{N-x}$

$= \sum_{x=1}^{N} x \frac{N(N-1)!}{(N-x)!(x)(x-1)!} p \left(p^{x-1}\right) (1-p)^{N-x}$

$= \sum_{x=1}^{N} x \frac{Np}{x} \frac{(N-1)!}{(N-x)!(x-1)!} p^{x-1} (1-p)^{N-x}$

$= Np \sum_{x=1}^{N} \frac{(N-1)!}{(N-x)!(x-1)!} p^{x-1} (1-p)^{N-x}$

$Let\ M = N-1\ and\ y = x-1.\ Then,$

$= Np \sum_{x=1}^{N} \frac{M!}{(M-y)!y!} p^y (1-p)^{M-y}$

$Taking \sum_{x=1}^{N} \frac{M!}{(M-y)!y!} p^y (1-p)^{M-y}\ as\ the\ binomial\ probability\ mass\ function\ from\ 0\ to\ M\ then$
$by\ definition:$

$\sum_{x=1}^{N} \frac{M!}{(M-y)!y!} p^y (1-p)^{M-y} = 1$

$Plugging\ back\ in,\ the\ equation\ becomes\ Np(1)$

$\mathbb{E}[\mathbf{T}(x)] = Np$

Multinomial

a.

$p(\boldsymbol{x}|\boldsymbol{p}) = \frac{N!}{x_1!x_2!\ldots x_n!} \prod_{d=1}^{D} p_d^{x_d}$

$It\ is\ known\ that\ for\ any\ x, x = e^{\log(x)}$

$= \frac{N!}{x_1!x_2!\ldots x_n!} e^{\log\left(\prod_{d=1}^{D} p_d^{x_d}\right)}$

$= \frac{N!}{x_1!x_2!\ldots x_n!} e^{\sum_{d=1}^{D} \log\left(p_d^{x_d}\right)}$

$= \frac{N!}{x_1!x_2!\ldots x_n!} e^{\sum_{d=1}^{D} x_d \log p_d}$

$= \frac{N!}{x_1!x_2!\ldots x_n!} e^{\sum_{d=1}^{D} (\log p_d)(x_d)}$

$$= \frac{N!}{x_1!x_2!\ldots x_n!} e^{\begin{bmatrix}\log p_1 & \log p_2 & \ldots & \log p_D\end{bmatrix}\begin{bmatrix}x_1 \\ x_2 \\ \vdots \\ x_D\end{bmatrix}}$$

Given the form of the exponential family, $p(x|\theta) = f(x)g(\theta)e^{\phi(\theta)^{\mathrm{T}}\mathbf{T}(x)}$,
the natural parameter is

$$\phi(\theta) = \begin{bmatrix} \log p_1 \\ \log p_2 \\ \vdots \\ \log p_D \end{bmatrix}$$

and the sufficient statistic is

$$\mathbf{T}(x) = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_D \end{bmatrix}$$

b.

$$\mathbb{E}[\mathbf{T}(x)] = \mathbb{E}\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_D \end{bmatrix}$$

For all x_d,

$$\mathbb{E}[x_d] = \sum_{x_1, x_2, \ldots, x_D}^{N} x_d \tfrac{N!}{x_1!x_2!\ldots x_D!} p_1^{x_1} p_2^{x_2} \ldots p_D^{x_D}$$

x_d *is removed from the summation and displayed as a separate summation. The second summation in the line is over all* x *not including* x_d

$$= \sum_{x_d}^{N} x_d \sum_{x_1, \ldots x_{d-1}, x_{d+1} \ldots, x_D}^{N} \tfrac{N(N-1)!}{x_1!x_2!\ldots x_d(x_d-1)!\ldots x_D!} p_1^{x_1} p_2^{x_2} \ldots p_d p_d^{x_d-1} \ldots p_D^{x_D}$$

$$= Np_d \sum_{x_1, \ldots x_{d-1}, x_{d+1} \ldots, x_D}^{N} \tfrac{(N-1)!}{x_1!x_2!\ldots(x_d-1)!\ldots x_D!} p_1^{x_1} p_2^{x_2} \ldots p_d^{x_d-1} \ldots p_D^{x_D}$$

Let $M = \{x \in N, x \neq x_d\}$ *and* $x_i = x_d - 1$. *Then,*

$$= Np_d \sum_{x_1, \ldots, x_D}^{M} \tfrac{M!}{x_1!x_2!\ldots(x_i)!\ldots x_D!} p_1^{x_1} p_2^{x_2} \ldots p_d^{x_i} \ldots p_D^{x_D}$$

Taking $\sum_{x_1, \ldots, x_D}^{M} \tfrac{M!}{x_1!x_2!\ldots(x_i)!\ldots x_D!} p_1^{x_1} p_2^{x_2} \ldots p_d^{x_i} \ldots p_D^{x_D}$ *as the multinomial probability mass function from* 0 *to* M *then by definition :*

$$\sum_{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_D}^{M} \tfrac{M!}{x_1!x_2!\ldots(x_i)!\ldots x_D!} p_1^{x_1} p_2^{x_2} \ldots p_d^{x_i} \ldots p_D^{x_D} = 1$$

Plugging back in, the equation becomes $Np_d(1)$

$$\mathbb{E}[x_d] = Np_d \textit{ for all } x_d \textit{ and thus, } \mathbb{E}[\mathbf{T}(x)] = \begin{bmatrix} Np_1 \\ Np_2 \\ \vdots \\ Np_D \end{bmatrix}.$$

Poisson

a.

$p(x|\mu) = \frac{\mu^x e^{-\mu}}{x!}$
$It\ is\ known\ that\ for\ any\ x,\ x = e^{\log(x)}$
$= \frac{1}{x!}e^{-\mu}e^{\log \mu^x}$
$= \frac{1}{x!}e^{-\mu}e^{x \log \mu}$
$= \frac{1}{x!}e^{-\mu}e^{(\log \mu)^{\mathrm{T}} x}$
$Given\ the\ form\ of\ the\ exponential\ family,\ p(x|\theta) = f(x)g(\theta)e^{\phi(\theta)^{\mathrm{T}}\mathbf{T}(x)},$
$the\ natural\ parameter\ is$
$\phi(\theta) = \begin{bmatrix}\log \mu\end{bmatrix}$
$and\ the\ sufficient\ statistic\ is$
$\mathbf{T}(x) = \begin{bmatrix}x\end{bmatrix}$

b.

$\mathbb{E}[\mathbf{T}(x)] = \mathbb{E}[x] = \sum\limits_{i=0}^{n} x_i \frac{e^{-\mu}\mu^{x_i}}{x_i!}$
$When\ x = 0,\ the\ first\ term\ is\ 0,\ so\ it\ can\ be\ removed\ from\ the\ summation$
$= \sum\limits_{i=1}^{n} x_i \frac{e^{-\mu}\mu^{x_i}}{x_i!}$
$= \sum\limits_{i=1}^{n} x_i \frac{e^{-\mu}\mu\mu^{x_i-1}}{x_i(x_i-1)!}$

$= \mu e^{-\mu}\sum\limits_{i=1}^{n} \frac{\mu^{x_i-1}}{(x_i-1)!}$
$Let\ y_j = x_i - 1\ and\ j = i - 1.\ Then,$
$= \mu e^{-\mu}\sum\limits_{j=0}^{n} \frac{\mu^{y_j}}{(y_j)!}$
$Taylor's\ Theorem\ states\ that\ \sum\limits_{k=0}^{\infty}\frac{z^k}{k!} = e^z.\ Therefore,\ as\ n\ approaches\ \infty,$
$= \mu e^{-\mu}e^{\mu}$
$= \mu$
$\mathbb{E}[\mathbf{T}(x)] = \mu$

Beta

a.

$p(x|\alpha, \beta) = \frac{1}{\mathrm{B}(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1}$
$It\ is\ known\ that\ for\ any\ x, x = e^{\log(x)}$
$= \frac{1}{\mathrm{B}(\alpha,\beta)}e^{\log x^{\alpha-1} + \log(1-x)^{\beta-1}}$
$= \frac{1}{\mathrm{B}(\alpha,\beta)}e^{(\alpha-1)\log x + (\beta-1)\log(1-x)}$
$= \frac{1}{\mathrm{B}(\alpha,\beta)}e^{\alpha \log x - \log x + \beta \log(1-x) - \log(1-x)}$
$= \frac{1}{\mathrm{B}(\alpha,\beta)}e^{-\log x - \log(1-x)}e^{\alpha \log x + \beta \log(1-x)}$
$= \frac{1}{x(1-x)}\left(\frac{1}{\mathrm{B}(\alpha,\beta)}\right)e^{\begin{bmatrix}\alpha & \beta\end{bmatrix}\begin{bmatrix}\log x \\ \log(1-x)\end{bmatrix}}$
$Given\ the\ form\ of\ the\ exponential\ family,\ p(x|\theta) = f(x)g(\theta)e^{\phi(\theta)^{\mathrm{T}}\mathbf{T}(x)},$
$the\ natural\ parameter\ is$
$\phi(\theta) = \begin{bmatrix}\alpha \\ \beta\end{bmatrix}$
$and\ the\ sufficient\ statistic\ is$
$\mathbf{T}(x) = \begin{bmatrix}\log x \\ \log(1-x)\end{bmatrix}$

b.

As shown previously, the derivative of the log partition function equates to the expectation of the sufficient statistic. To derive the log partition,

$p(x|\alpha,\beta) = \frac{1}{x(1-x)}\left(\frac{1}{\mathrm{B}(\alpha,\beta)}\right) e^{\begin{bmatrix}\alpha & \beta\end{bmatrix}\begin{bmatrix}\log x \\ \log(1-x)\end{bmatrix}}$

$= \frac{1}{x(1-x)} e^{\log\left(\frac{1}{\mathrm{B}(\alpha,\beta)}\right)} e^{\begin{bmatrix}\alpha & \beta\end{bmatrix}\begin{bmatrix}\log x \\ \log(1-x)\end{bmatrix}}$

$= \frac{1}{x(1-x)} e^{\begin{bmatrix}\alpha & \beta\end{bmatrix}\begin{bmatrix}\log x \\ \log(1-x)\end{bmatrix} - \left(-\log\left(\frac{1}{\mathrm{B}(\alpha,\beta)}\right)\right)}$

$= \frac{1}{x(1-x)} e^{\begin{bmatrix}\alpha & \beta\end{bmatrix}\begin{bmatrix}\log x \\ \log(1-x)\end{bmatrix} - \log \mathrm{B}(\alpha,\beta)}$

The log partition of the beta exponential family is found to be $A(\alpha,\beta) = \log \mathrm{B}(\alpha,\beta)$.

$\mathbb{E}[\mathbf{T}(x)] = \begin{bmatrix}\frac{\partial A(\alpha,\beta)}{\partial\alpha} \\ \frac{\partial A(\alpha,\beta)}{\partial\beta}\end{bmatrix}$

$\frac{\partial A(\alpha,\beta)}{\partial\alpha} = \log \mathrm{B}(\alpha,\beta)\frac{\partial}{\partial\alpha}$

$= \frac{1}{\mathrm{B}(\alpha,\beta)}\left(\mathrm{B}(\alpha,\beta)\frac{\partial}{\partial\alpha}\right)$

$\mathrm{B}(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$. *Thus,*

$= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\left(\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}\frac{\partial}{\partial\alpha}\right)$

The derivative of $\Gamma(x)$ *is* $\Gamma(x)\psi(x)$. *The equation becomes*

$= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\left(\frac{\Gamma(\alpha+\beta)\Gamma(\alpha)\psi(\alpha)\Gamma(\beta) - \Gamma(\alpha)\Gamma(\beta)\Gamma(\alpha+\beta)\psi(\alpha+\beta)}{(\Gamma(\alpha+\beta))^2}\right)$

Simplifying the fractions yields

$= \psi(\alpha) - \psi(\alpha+\beta)$

Similarly with respect to β,

$\frac{\partial A(\alpha,\beta)}{\partial\beta} = \log \mathrm{B}(\alpha,\beta)\frac{\partial}{\partial\beta}$

$= \frac{1}{\mathrm{B}(\alpha,\beta)}\left(\mathrm{B}(\alpha,\beta)\frac{\partial}{\partial\beta}\right)$

$= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\left(\frac{\Gamma(\alpha+\beta)\Gamma(\beta)\psi(\beta)\Gamma(\alpha) - \Gamma(\alpha)\Gamma(\beta)\Gamma(\alpha+\beta)\psi(\alpha+\beta)}{(\Gamma(\alpha+\beta))^2}\right)$

Simplifying the fraction yields

$= \psi(\beta) - \psi(\alpha+\beta)$

$\mathbb{E}[\mathbf{T}(x)] = \begin{bmatrix}\psi(\alpha) - \psi(\alpha+\beta) \\ \psi(\beta) - \psi(\alpha+\beta)\end{bmatrix}$

Gamma

a.

$p(x|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x}$

It is known that for any $x, x = e^{\log(x)}$

$= \frac{\beta^\alpha}{\Gamma(\alpha)}e^{\log x^{\alpha-1}}e^{-\beta x}$

$= \frac{\beta^\alpha}{\Gamma(\alpha)}e^{(\alpha-1)\log x - \beta x}$

$= \frac{\beta^\alpha}{\Gamma(\alpha)}e^{\alpha\log x - \log x - \beta x}$

$= \frac{\beta^\alpha}{\Gamma(\alpha)}e^{-\log x}e^{\alpha\log x - \beta x}$

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

Given the form of the exponential family, $p(x|\theta) = f(x)g(\theta)e^{\phi(\theta)^{\mathrm{T}}\mathbf{T}(x)}$,
the natural parameter is
$\phi(\theta) = \begin{bmatrix} \alpha \\ -\beta \end{bmatrix}$
and the sufficient statistic is
$\mathbf{T}(x) = \begin{bmatrix} \log x \\ x \end{bmatrix}$

b.

As shown previously, the derivative of the log partition function equates to the expectation of the sufficient statistic. To derive the log partition,

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

The log partition of the gamma exponential family is found to be $A(\alpha,\beta) = \log\frac{\Gamma(\alpha)}{\beta^\alpha}$.

$\mathbb{E}[\mathbf{T}(x)] = \begin{bmatrix}\frac{\partial A(\alpha,\beta)}{\partial\alpha} \\ \frac{\partial A(\alpha,\beta)}{\partial\beta}\end{bmatrix}$
$\frac{\partial A(\alpha,\beta)}{\partial\alpha} = \log\frac{\Gamma(\alpha)}{\beta^\alpha}\frac{\partial}{\partial\alpha}$
$= \frac{\beta^\alpha}{\Gamma(\alpha)}\left(\frac{\Gamma(\alpha)}{\beta^\alpha}\frac{\partial}{\partial\alpha}\right)$
$= \frac{\beta^\alpha}{\Gamma(\alpha)}\left(\frac{\beta^\alpha\Gamma(\alpha)\psi(\alpha) - \Gamma(\alpha)(\beta^\alpha)\log(\beta)}{(\beta^\alpha)^2}\right)$
$= \psi(\alpha) - \log(\beta)$

$\frac{\partial A(\alpha,\beta)}{\partial\beta} = \log\frac{\Gamma(\alpha)}{\beta^\alpha}\frac{\partial}{\partial\beta}$
$= (\log\Gamma(\alpha) - \alpha\log\beta)\frac{\partial}{\partial\beta}$
$= -\frac{\alpha}{\beta}$
$\mathbb{E}[\mathbf{T}(x)] = \begin{bmatrix}\psi(\alpha) - \log(\beta) \\ -\frac{\alpha}{\beta}\end{bmatrix}$

Dirichlet

a.

$p(\boldsymbol{x}|\boldsymbol{\alpha}) = \frac{\Gamma\left(\sum\limits_{d=1}^{D}\alpha_d\right)}{\prod\limits_{d=1}^{D}\Gamma(\alpha_d)}\prod_{d=1}^{D}x_d^{\alpha_d-1}$

It is known that for any $x, x = e^{\log(x)}$

$= \frac{\Gamma\left(\sum\limits_{d=1}^{D}\alpha_d\right)}{\prod\limits_{d=1}^{D}\Gamma(\alpha_d)}e^{\log\left(\prod\limits_{d=1}^{D}x_d^{\alpha_d-1}\right)}$

$= \frac{\Gamma\left(\sum\limits_{d=1}^{D}\alpha_d\right)}{\prod\limits_{d=1}^{D}\Gamma(\alpha_d)}e^{\sum\limits_{d=1}^{D}\log x_d^{\alpha_d-1}}$

$$= \frac{\Gamma\left(\sum_{d=1}^{D} \alpha_d\right)}{\prod_{d=1}^{D} \Gamma(\alpha_d)} e^{\sum_{d=1}^{D} (\alpha_d - 1) \log x_d}$$

$$= \frac{\Gamma\left(\sum_{d=1}^{D} \alpha_d\right)}{\prod_{d=1}^{D} \Gamma(\alpha_d)} e^{\sum_{d=1}^{D} (\alpha_d \log x_d - \log x_d)}$$

$$= \frac{1}{\prod_{d=1}^{D} x_d} \left(\frac{\Gamma\left(\sum_{d=1}^{D} \alpha_d\right)}{\prod_{d=1}^{D} \Gamma(\alpha_d)} \right) e^{\begin{bmatrix}\alpha_1 & \alpha_2 & \dots & \alpha_d\end{bmatrix} \begin{bmatrix}\log x_1 \\ \log x_2 \\ \vdots \\ \log x_D\end{bmatrix}}$$

Given the form of the exponential family, $p(x|\theta) = f(x)g(\theta)e^{\phi(\theta)^{\mathrm{T}}\mathbf{T}(x)}$,
the natural parameter is

$$\phi(\theta) = \begin{bmatrix}\alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_D\end{bmatrix}$$

and the sufficient statistic is

$$\mathbf{T}(x) = \begin{bmatrix}\log x_1 \\ \log x_2 \\ \vdots \\ \log x_D\end{bmatrix}$$

b.

As shown previously, the derivative of the log partition function equates to the expectation of the sufficient statistic. To derive the log partition,

$$p(\boldsymbol{x}|\boldsymbol{\alpha}) = \frac{1}{\prod_{d=1}^{D} x_d} \left(\frac{\Gamma\left(\sum_{d=1}^{D} \alpha_d\right)}{\prod_{d=1}^{D} \Gamma(\alpha_d)} \right) e^{\begin{bmatrix}\alpha_1 & \alpha_2 & \dots & \alpha_d\end{bmatrix} \begin{bmatrix}\log x_1 \\ \log x_2 \\ \vdots \\ \log x_D\end{bmatrix}}$$

$$= \frac{1}{\prod_{d=1}^{D} x_d} e^{\log \Gamma\left(\sum_{d=1}^{D} \alpha_d\right) - \log\left(\prod_{d=1}^{D} \Gamma(\alpha_d)\right)} e^{\begin{bmatrix}\alpha_1 & \alpha_2 & \dots & \alpha_d\end{bmatrix} \begin{bmatrix}\log x_1 \\ \log x_2 \\ \vdots \\ \log x_D\end{bmatrix}}$$

$$= \frac{1}{\prod_{d=1}^{D} x_d} e^{\log \Gamma\left(\sum_{d=1}^{D} \alpha_d\right) - \sum_{d=1}^{D}} \log \Gamma(\alpha_d) e^{\begin{bmatrix}\alpha_1 & \alpha_2 & \dots & \alpha_d\end{bmatrix} \begin{bmatrix}\log x_1 \\ \log x_2 \\ \vdots \\ \log x_D\end{bmatrix}}$$

$$= \frac{1}{\prod_{d=1}^{D} x_d} e^{\begin{bmatrix}\alpha_1 & \alpha_2 & \dots & \alpha_d\end{bmatrix}\begin{bmatrix}\log x_1 \\ \log x_2 \\ \vdots \\ \log x_D\end{bmatrix} - \left(\sum_{d=1}^{D} \log\Gamma(\alpha_d) - \log\Gamma\left(\sum_{d=1}^{D}\alpha_d\right)\right)}$$

The log partition of the dirichlet exponential family is found to be $A(\alpha) = \sum_{d=1}^{D} \log\Gamma(\alpha_d) - \log\Gamma\left(\sum_{d=1}^{D}\alpha_d\right)$.

$\mathbb{E}[\mathbf{T}(x)] = \left[\frac{\partial A(\alpha)}{\partial \alpha}\right]$

Solving for all α_i, $\frac{\partial A(\alpha)}{\partial \alpha_i}$

$= \left(\sum_{d=1}^{D} \log\Gamma(\alpha_d) - \log\Gamma\left(\sum_{d=1}^{D}\alpha_d\right)\right)\frac{\partial}{\partial \alpha_i}$

Because all $\alpha_d \neq \alpha_i$ *are constants, their derivatives with respect to* α_i *are* 0. *Thus, the equation can be simplified to*

$= (\log\Gamma(\alpha_i))\frac{\partial}{\partial \alpha_i} - \left(\frac{1}{\Gamma\left(\sum_{d=1}^{D}\alpha_d\right)}\left(\Gamma\left(\sum_{d=1}^{D}\alpha_d\right)\frac{\partial}{\partial \alpha_i}\right)\right)$

$= \frac{1}{\Gamma(\alpha_i)}\Gamma(\alpha_i)\psi(\alpha_i) - \left(\frac{1}{\Gamma\left(\sum_{d=1}^{D}\alpha_d\right)}\left(\Gamma\left(\sum_{d=1}^{D}\alpha_d\right)\psi\left(\sum_{d=1}^{D}\alpha_d\right)\right)\right)$

$= \psi(\alpha_i) - \psi\left(\sum_{d=1}^{D}\alpha_d\right)$

$$\mathbb{E}[\mathbf{T}(x)] = \begin{bmatrix}\psi(\alpha_1) - \psi\left(\sum_{d=1}^{D}\alpha_d\right) \\ \psi(\alpha_2) - \psi\left(\sum_{d=1}^{D}\alpha_d\right) \\ \vdots \\ \psi(\alpha_D) - \psi\left(\sum_{d=1}^{D}\alpha_d\right)\end{bmatrix}$$

Question 2.

$\hat{\theta}_{ML} = \underset{\theta \in \mathcal{T}}{\arg\max} p(\boldsymbol{x}|\theta)$

$= \left(\prod_{i=1}^{n} g(\theta)f(x_i)e^{\theta^{\mathrm{T}}\mathbf{T}(x_i)}\right)\nabla_\theta$

Taking the log likelihood,

$= \log\left(\left(\prod_{i=1}^{n} g(\theta)f(x_i)\right)e^{\theta^{\mathrm{T}}\mathbf{T}(x_i)}\right)\nabla_\theta$

$= \left(\sum_{i=1}^{n} \log g(\theta) + \log f(x_i) + \theta^{\mathrm{T}}\mathbf{T}(x_i)\right)\nabla_\theta$

$= \left(n\log g(\theta) + \sum_{i=1}^{n}\left(\log f(x_i) + \theta^{\mathrm{T}}\mathbf{T}(x_i)\right)\right)\nabla_\theta$

$= n\frac{g(\theta)\nabla_\theta}{g(\theta)} + \sum_{i=1}^{n}\mathbf{T}(x_i)$

$= \frac{g(\theta)\nabla_\theta}{g(\theta)} + \frac{1}{n}\sum_{i=1}^{n}\mathbf{T}(x_i) \ = \ 0$

$-\frac{g(\theta)\nabla_\theta}{g(\theta)} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{T}(x_i)$

Question 3.

a.

A multivariate Gaussian would not be appropriate because the image pixels are discrete and binary values of 0 and 1, whereas a multivariate Gaussian would model a continuous distribution of values with a mean between 0 and 1. Binary values are not continuous and are either valued 0 or 1 with probability p or 1-p; a Gaussian is inadequate for portraying a binary distribution of pixel values.

b.

$$P(\boldsymbol{x}|\boldsymbol{p}) = \prod_{d=1}^{D} p_d^{x_d}(1-p_d)^{(1-x_d)}$$

$$\hat{p}_{ML} = \underset{p \in \mathcal{T}}{\arg\max}\ P(\boldsymbol{x}|\boldsymbol{p})$$
$$= \left(\prod_{i=1}^{N}\prod_{d=1}^{D} p_d^{x_d^i}(1-p_d)^{\left(1-x_d^i\right)}\right)\nabla_{\hat{p}} = 0$$
$$Taking\ the\ log\ likelihood,\ and\ solving\ for\ all\ p_j,$$
$$\mathcal{LL}(p_j) = \log\left(\prod_{i=1}^{N}\prod_{d=1}^{D} p_d^{x_d^i}(1-p_d)^{\left(1-x_d^i\right)}\right)dp_j$$
$$= \sum_{i=1}^{N}\sum_{d=1}^{D}\left(\log p_d^{x_d^i} + \log(1-p_d)^{\left(1-x_d^i\right)}\right)dp_j$$
$$= \sum_{i=1}^{N}\sum_{d=1}^{D}\left(x_d^i \log p_d + \left(1-x_d^i\right)\log(1-p_d)\right)dp_j$$
$$Because\ all\ p_d \neq p_j\ are\ constants,\ their\ derivatives\ with\ respect\ to\ p_j\ are\ 0.\ Thus,\ the\ equation\ can\ be\ simplified\ to$$
$$= \sum_{i=1}^{N}\left(\frac{x_j^i}{p_j} - \frac{1-x_j^i}{1-p_j}\right)$$
$$= \frac{N_j}{p_j} - \frac{N-N_j}{1-p_j} = 0$$
$$where\ N_j\ equals\ the\ sum\ of\ x^{i\prime}s\ at\ j\ (in\ this\ example,\ the\ sum\ of\ the\ value\ of\ pixel\ j\ over\ all\ images).$$
$$Solving\ for\ p_j,$$
$$\frac{N_j}{p_j} = \frac{N-N_j}{1-p_j}$$
$$N_j - N_j p_j = N p_j - N_j p_j$$
$$p_j = \frac{N_j}{N}$$
$$\hat{p}_{ML} = \begin{bmatrix} \frac{N_1}{N} \\ \frac{N_2}{N} \\ \vdots \\ \frac{N_D}{N} \end{bmatrix}$$

c.

$$By\ Bayes'\ Theorem,$$
$$P(\boldsymbol{p}|\boldsymbol{x}) = \frac{P(\boldsymbol{x}|\boldsymbol{p})P(\boldsymbol{p})}{P(\boldsymbol{x})}$$
$$= \frac{\left(\prod_{i=1}^{N}\left(\prod_{d=1}^{D} p_d^{x_d^i}(1-p_d)^{\left(1-x_d^i\right)}\right)\right)\frac{1}{\mathrm{B}(\alpha,\beta)}p_j^{\alpha-1}(1-p_j)^{\beta-1}}{P(\boldsymbol{x})}$$
$$Because\ P(\boldsymbol{x})\ is\ constant,\ it\ can\ be\ removed\ and\ the\ maximum\ a\ posteriori\ value\ of\ p\ will\ still\ be\ found.$$
$$Taking\ the\ log\ posterior\ and\ solving\ for\ p_j,$$
$$\hat{p}_{MAP} = \underset{p \in \mathcal{T}}{\arg\max} P(\boldsymbol{p}|\boldsymbol{x})$$
$$= \log\left(\left(\prod_{i=1}^{N}\left(\prod_{d=1}^{D} p_d^{x_d^i}(1-p_d)^{\left(1-x_d^i\right)}\right)\right)\frac{1}{\mathrm{B}(\alpha,\beta)}p_j^{\alpha-1}(1-p_j)^{\beta-1}\right)dp_j = 0$$

$$= \left(\left(\sum_{i=1}^{N} \left(\sum_{d=1}^{D} \log\left(p_d^{x_d^i}\right) + \log(1-p_d)^{\left(1-x_d^i\right)} \right) \right) - \log \mathrm{B}(\alpha,\beta) + \log p_j^{\alpha-1} + \log(1-p_j)^{\beta-1} \right) dp_j$$

$$= \left(\left(\sum_{i=1}^{N} \left(\sum_{d=1}^{D} x_d^i \log(p_d) + \left(1-x_d^i\right)\log(1-p_d) \right) \right) - \log \mathrm{B}(\alpha,\beta) + (\alpha-1)\log p_j + (\beta-1)\log(1-p_j) \right) dp_j$$

Because all $p_d \neq p_j$ are constants, their derivatives with respect to p_j are 0. Thus, the equation can be simplified to

$$= \sum_{i=1}^{N} \left(\frac{x_j^i}{p_j} - \frac{1-x_j^i}{1-p_j} \right) + \frac{\alpha-1}{p_j} - \frac{\beta-1}{1-p_j}$$

$$= \frac{N_j}{p_j} - \frac{N-N_j}{1-p_j} + \frac{\alpha-1}{p_j} - \frac{\beta-1}{1-p_j}$$

where N_j equals the sum of x^i's at j (in this example, the sum of the value of pixel j over all images).

$$= \frac{N_j+\alpha-1}{p_j} - \frac{N-N_j+\beta-1}{1-p_j} = 0$$

Solving for p_j,

$$\frac{N_j+\alpha-1}{p_j} = \frac{N-N_j+\beta-1}{1-p_j}$$

$$N_j + \alpha - 1 - N_j p_j - \alpha p_j + p_j = N p_j - N_j p_j + \beta p_j - p_j$$

$$\alpha p_j - 2p_j + \beta p_j + N p_j = N_j + \alpha - 1$$

$$p_j(\alpha - 2 + \beta + N) = N_j + \alpha - 1$$

$$p_j = \frac{N_j+\alpha-1}{\alpha-2+\beta+N}$$

$$\hat{p}_{MAP} = \begin{bmatrix} \dfrac{N_1+\alpha-1}{\alpha-2+\beta+N} \\ \dfrac{N_2+\alpha-1}{\alpha-2+\beta+N} \\ \vdots \\ \dfrac{N_D+\alpha-1}{\alpha-2+\beta+N} \end{bmatrix}$$

d.

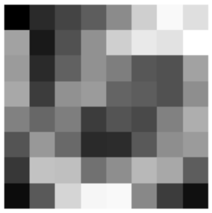
```
import numpy as np
from matplotlib import pyplot as plt
Y = np.loadtxt('binarydigits.txt')
N, D = Y.shape
MLParam = []
for d in range(D):
    N_d = 0
    for n in range(N):
        if Y[n-1][d-1] == 1:
            N_d += 1
    MLParam.append(N_d/N)

print(MLParam)

plt.figure()
plt.imshow(np.reshape(MLParam, (8,8)),
           interpolation="None",
           cmap='gray')
plt.axis('off')
```

```
[0.0, 0.13, 0.21, 0.29, 0.43, 0.64, 0.77, 0.69, 0.5, 0.08, 0.25, 0.45, 0.64, 0.72, 0.7, 0.79,
0.48, 0.13, 0.3, 0.45, 0.39, 0.27, 0.25, 0.5, 0.52, 0.19, 0.45, 0.48, 0.31, 0.29, 0.25, 0.44,
0.4, 0.32, 0.39, 0.19, 0.26, 0.23, 0.4, 0.54, 0.26, 0.47, 0.33, 0.13, 0.14, 0.28, 0.44, 0.48,
0.17, 0.6, 0.59, 0.35, 0.44, 0.57, 0.52, 0.29, 0.04, 0.28, 0.66, 0.76, 0.77, 0.42, 0.19, 0.0
5]
```

```
(-0.5, 7.5, 7.5, -0.5)
```

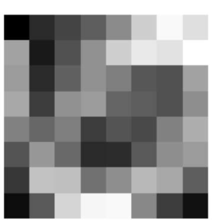


e.

```
import numpy as np
from matplotlib import pyplot as plt
Y = np.loadtxt('binarydigits.txt')
N, D = Y.shape
MAPParam = []
alpha = 3
beta = 3
for d in range(D):
    p_d = 0
    N_d = 0
     for n in range(N):
         if Y[n-1][d-1] == 1:
             N_d += 1
    p_d = (N_d + alpha - 1)/(alpha-2+beta+N)
    MAPParam.append(p_d)
print(MAPParam)
plt.figure()
plt.imshow(np.reshape(MAPParam, (8,8)),
           interpolation="None",
           cmap='gray')
plt.axis('off')
```

```
[0.019230769230769232, 0.14423076923076922, 0.22115384615384615, 0.2980769230769231, 0.432692
3076923077, 0.6346153846153846, 0.7596153846153846, 0.6826923076923077, 0.5, 0.09615384615384
616, 0.25961538461538464, 0.4519230769230769, 0.6346153846153846, 0.7115384615384616, 0.69230
76923076923, 0.7788461538461539, 0.4807692307692308, 0.14423076923076922, 0.3076923076923077,
0.4519230769230769, 0.3942307692307692, 0.27884615384615385, 0.25961538461538464, 0.5, 0.5192
307692307693, 0.20192307692307693, 0.4519230769230769, 0.4807692307692308, 0.317307692307692
3, 0.2980769230769231, 0.25961538461538464, 0.4423076923076923, 0.40384615384615385, 0.326923
0769230769, 0.3942307692307692, 0.20192307692307693, 0.2692307692307692, 0.2403846153846154,
0.40384615384615385, 0.5384615384615384, 0.2692307692307692, 0.47115384615384615, 0.336538461
53846156, 0.14423076923076922, 0.15384615384615385, 0.28846153846153844, 0.4423076923076923,
0.4807692307692308, 0.18269230769230768, 0.5961538461538461, 0.5865384615384616, 0.3557692307
692308, 0.4423076923076923, 0.5673076923076923, 0.5192307692307693, 0.2980769230769231, 0.057
692307692307696, 0.28846153846153844, 0.6538461538461539, 0.75, 0.7596153846153846, 0.4230769
230769231, 0.20192307692307693, 0.0673076923076923]
```

```
(-0.5, 7.5, 7.5, -0.5)
```



In this example, the MAP may be better than the ML estimate because there is a good prior for the data. Knowing that the pixel values are binary, a beta prior is adequate for the MAP and we can incorporate the prior knowledge into the parameter estimation with MAP. In contrast, the ML estimate does not incorporate any prior and thus the representation of binary data will not be accounted for in the estimate.

Question 4.

a.

$P(M_i|D) = \frac{P(D|M_i)P(M_i)}{P(D)}$
Because all three models are equally likely a priori,
$P(M_i) = \frac{1}{3}$
$P(D) = \sum_{i=1}^{3} P(D|M_i)P(M_i)$
$= \sum_{i=1}^{3} P(D|M_i)\left(\frac{1}{3}\right)$
$= \left(\frac{1}{3}\right)\sum_{i=1}^{3} P(D|M_i)$
$P(M_i)$ *and* $P(D)$ *are the same for all three models. In order to calculate the relative probability of the three models,* $P(D|M_i)$ *must be solved for.*
Solving for $P(D|M_1)$*, where* M_1 *is the model where all* D *components are generated from a Bernoulli distribution with* $p_d = 0.5$,
$P(D|M_1) = \int P(D|\theta_1, M_1)P(\theta_1|M_1)d\theta_1$
Because there is only one possible value of parameter p_d *for model* M_1, $P(\theta_1|M_1) = 1$
$P(D|M_1) = \int \left(\left(\prod_{i=1}^{N}\prod_{d=1}^{D} p_d^{x_d^i}(1-p_d)^{(1-x_d^i)}\right)(1)\right)dp_d$
Because p_d *is the same value over all pixels and images, the product of* p_d *and* $1-p_d$ *over all pixels and images can be written as a summation in the exponent,*

$= \int \left(p_d^{\sum_{i=1}^{N}\sum_{d=1}^{D} x_d^i}(1-p_d)^{\sum_{i=1}^{N}\sum_{d=1}^{D}(1-x_d^i)} \right) dp_d$
Because there is only one value for p_d*, the point mass is at* 0.5 *and equal to* 1*. The integral in the expression can be removed and* p_d *can be replaced with its value of* 0.5.
$= 0.5^{\sum_{i=1}^{N}\sum_{d=1}^{D} x_d^i}(1-0.5)^{\sum_{i=1}^{N}\sum_{d=1}^{D}(1-x_d^i)}$
$= 0.5^{\sum_{i=1}^{N}\sum_{d=1}^{D} x_d^i}0.5^{\sum_{i=1}^{N}\sum_{d=1}^{D}(1-x_d^i)}$
$= 0.5^{\sum_{i=1}^{N}\sum_{d=1}^{D} x_d^i+(1-x_d^i)}$
$= 0.5^{\sum_{i=1}^{N}\sum_{d=1}^{D} 1}$
$= 0.5^{ND}$
$P(D|M_1) = 0.5^{ND}$

b.

Solving for $P(D|M_2)$*, where* M_2 *is the model where all* D *components are generated from Bernoulli distributions with unknown, but identical,* p_d,
$P(D|M_2) = \int P(D|\theta_2, M_2)P(\theta_2|M_2)d\theta_2$
Because all D *components have a* p_d *with the same value for model* M_2, $P(\theta_2|M_2) = 1$

$P(D|M_2) = \int_0^1 \left(\left(\prod_{i=1}^{N} \prod_{d=1}^{D} p_d^{x_d^i} (1-p_d)^{\left(1-x_d^i\right)} \right) (1) \right) dp_d$

Because p_d *is the same value over all pixels and images, the product of* p_d *and* $1-p_d$ *over all pixels and images can be written as a summation in the exponent,*

$= \int_0^1 \left(p_d^{\sum_{i=1}^{N} \sum_{d=1}^{D} x_d^i} (1-p_d)^{\sum_{i=1}^{N} \sum_{d=1}^{D} \left(1-x_d^i\right)} \right) dp_d$

This expression is the beta function $B(\alpha,\beta) = \int_0^1 \left(p^{\alpha-1}(1-p)^{\beta-1}\right) dp$,

where $\alpha = \left(\sum_{i=1}^{N} \sum_{d=1}^{D} x_d^i \right) + 1 = N_d^i + 1$, N_d^i *being the sum of all pixel values over all images*

and $\beta = \left(\sum_{i=1}^{N} \sum_{d=1}^{D} \left(1-x_d^i\right) \right) + 1 = ND - N_d^i + 1$, ND *being the product of the number of images and number of pixels per image.*

Therefore,

$P(D|M_2) = B\left(N_d^i + 1, ND - N_d^i + 1\right)$

c.

Solving for $P(D|M_3)$, *where* M_3 *is the model where each component is Bernoulli distributed with separate, unknown* p_d,

$P(D|M_3) = \int P(D|\theta_3, M_3) P(\theta_3|M_3) d\theta_3$

Because all D *components have separate, unknown* p_d *for model* M_3, *the expression must be integrated over all* p_d

$\int \ldots \int \left(\prod_{i=1}^{N} \prod_{d=1}^{D} p_d^{x_d^i} (1-p_d)^{\left(1-x_d^i\right)} \right) dp_1 dp_2 ... dp_D$

Because p_d *is the same value for a particular pixel over all images, the product of* p_d *and* $1-p_d$ *over all images can be written as a summation in the exponent,*

$\int \ldots \int \left(\prod_{d=1}^{D} p_d^{\sum_{i=1}^{N} x_d^i} (1-p_d)^{\sum_{i=1}^{N} \left(1-x_d^i\right)} \right) dp_1 dp_2 ... dp_D$

Because the probability distribution of p *is uniform, a particular* p_d *can be solved for and generalized over all* p

$= \int_0^1 \left(\prod_{d=1}^{D} p_d^{\sum_{i=1}^{N} x_d^i} (1-p_d)^{\sum_{i=1}^{N} \left(1-x_d^i\right)} \right) dp_d$

$= \prod_{d=1}^{D} \int_0^1 \left(p_d^{\sum_{i=1}^{N} x_d^i} (1-p_d)^{\sum_{i=1}^{N} \left(1-x_d^i\right)} \right) dp_d$

Again, the expression is the beta function $B(\alpha,\beta) = \int_0^1 \left(p^{\alpha-1}(1-p)^{\beta-1}\right) dp$,

where $\alpha = \left(\sum_{i=1}^{N} x_d^i \right) + 1 = N_d + 1$, N_d *being the sum of all values at a particular pixel d over all images*

and $\beta = \left(\sum_{i=1}^{N} \left(1-x_d^i\right) \right) + 1 = N - N_d + 1$

Therefore,

$P(D|M_3) = \prod_{d=1}^{D} B(N_d + 1, N - N_d + 1)$

Finally, now that all $P(D|M_i)$ have been solved for, $P(D)$ can be calculated.

$P(D) = \left(\frac{1}{3}\right) \sum_{i=1}^{3} P(D|M_i)$

$= \left(\frac{1}{3}\right) \left(0.5^{ND} + B\left(N_d^i + 1, ND - N_d^i + 1\right) + \prod_{d=1}^{D} B(N_d + 1, N - N_d + 1)\right)$

In order to rescale the probabilities in python so that they do not become 0, $\log P(D|M_i)$ *will be calculated in the code because* $P(D|M_i)$ *can be rewritten as* $e^{\log P(D|M_i)}$ *and solved for equivalently.*

```
import numpy as np
from matplotlib import pyplot as plt
import math

Y = np.loadtxt('binarydigits.txt')
N, D = Y.shape

LogLikelihood1=N*D*math.log(0.5)
LogLikelihood3 = 0

N_i_d = 0
for d in range(D):
    N_d = 0
    for n in range(N):
        if Y[n-1][d-1] == 1:
            N_d += 1
            N_i_d += 1
    alpha3=N_d+1
    beta3=N-N_d+1
    LogLikelihood3=LogLikelihood3+(math.lgamma(alpha3) + math.lgamma(beta3) - math.lgamma(a
alpha2=N_i_d+1
beta2=(N*D)-N_i_d+1
LogLikelihood2=(math.lgamma(alpha2) + math.lgamma(beta2) - math.lgamma(alpha2+beta2))
print("logP(D|M_1): ", LogLikelihood1)
print("logP(D|M_2): ", LogLikelihood2)
print("logP(D|M_3): ", LogLikelihood3)
```

```
logP(D|M_1):  -4436.14195558365
logP(D|M_2):  -4283.721342577344
logP(D|M_3):  -3851.1957439211315
```

Using the values found for $\log P(D|M_1)$, $\log P(D|M_2)$, *and* $\log P(D|M_3)$ *and simplifying, substituting* $P(D|M_i)$ *with* $e^{\log(P(M_i\}D))}$

$P(M_i|D) = \frac{P(D|M_i)P(M_i)}{P(D)}$

$P(M_1|D) = \frac{P(D|M_1)P(M_1)}{P(D)}$

$= \frac{e^{\log P(D|M_1)}\left(\frac{1}{3}\right)}{\frac{1}{3}\left(e^{\log P(D|M_1)} + e^{\log P(D|M_2)} + e^{\log P(D|M_2)}\right)}$

$= \frac{e^{-4436.142}\left(\frac{1}{3}\right)}{\left(\frac{1}{3}\right)(e^{-4436.142} + e^{-4283.721} + e^{-3851.196})}$

$= \frac{e^{-4436.142}}{e^{-4436.142} + e^{-4283.721} + e^{-3851.196}}$

Dividing the numerator and denominator by the numerator,
$= \frac{1}{1+e^{152.421}+e^{584.946}}$
$= \frac{1}{1.0935\times10^{254}}$
≈ 0

$P(M_2|D) = \frac{P(D|M_2)P(M_2)}{P(D)}$
$= \frac{e^{-4283.721}(\frac{1}{3})}{(\frac{1}{3})(e^{-4436.142}+e^{-4283.721}+e^{-3851.196})}$
Dividing the numerator and denominator by the numerator,
$= \frac{1}{e^{-152.421}+1+e^{432.525}}$
$= \frac{1}{6.3738\times10^{-67}+1+6.9698\times10^{187}}$
$= \frac{1}{6.9698\times10^{187}}$
≈ 0

$P(M_3|D) = \frac{P(D|M_3)P(M_3)}{P(D)}$
$= \frac{e^{-3851.196}(\frac{1}{3})}{(\frac{1}{3})(e^{-4436.142}+e^{-4283.721}+e^{-3851.196})}$
$= \frac{e^{-3851.196}}{e^{-4436.142}+e^{-4283.721}+e^{-3851.196}}$
Dividing the numerator and denominator by the numerator,
$= \frac{1}{e^{-584.946}+e^{-432.525}+1}$
$= \frac{1}{9.1449\times10^{-255}+1.4348\times10^{-188}+1}$
$\approx \frac{1}{1}$
$= 1$
Thus, $P(M_1|D) \approx 0$
$P(M_2|D) \approx 0$
$P(M_3|D) \approx 1$
and model 3, *where each component is Bernoulli distributed with separate, unknown* p_d *is the most probable model given the data.*

Question 5.

a.

For all eigenvectors $\boldsymbol{x}$ *of* A, $A\boldsymbol{x} = \lambda\boldsymbol{x}$.
If $B = A + cI$, *where* I *is the identity matrix and* $c \in \mathbb{R}$, *then for all eigenvectors* $\boldsymbol{x}$ *of* A,
$B\boldsymbol{x} = (A + cI)\boldsymbol{x}$
$= A\boldsymbol{x} + cI\boldsymbol{x}$
$= \lambda\boldsymbol{x} + c\boldsymbol{x}$
$= (\lambda + c)\boldsymbol{x}$
This means that
$B\boldsymbol{x} = (\lambda + c)\boldsymbol{x}$
and thus, by the definition of an eigenvalue, $\lambda + c$ *are eigenvalues of* B *for all eigenvalues* λ *and eigenvectors* x *of* A, *meaning that* B *has eigenvalues* $\lambda_1 + c,\ \lambda_2 + c, \ldots, \lambda_n + c$.

b.

All possible linear combinations of v *and* w *can be represented using scalars* b *and* c :
$A(bv + cw)$
$= Abv + Acw$
$= bAv + cAw$
Because v *and* w *are eigenvectors of* A,
$= b\lambda v + c\lambda w$
$= \lambda(bv + cw)$

Let the vector $z = bv + cw$. *Then,*
$= \lambda z$
and
$A(bv + cw) = Az = \lambda z$
z is an eigenvector of A by the definition of an eigenvector. Its eigenvalue is the same eigenvalue as v and w : λ.

Question 6.

a.

Optimization function : $f(x, y) = x + 2y$
Constraint : $y^2 + xy = 1$
Using the formulae for the Lagrange multiplier
$\nabla f + \lambda \nabla g = 0$
$g(x, y) = 0$
a system of equations is found :
$g(x, y) = y^2 + xy - 1 = 0 \qquad (1)$
$\frac{\partial(x+2y)}{\partial x} + \lambda \frac{\partial\left(y^2+xy-1\right)}{\partial x} = 0$
$\frac{\partial(x+2y)}{\partial y} + \lambda \frac{\partial\left(y^2+xy-1\right)}{\partial y} = 0$
the latter two simplify to
$1 + \lambda y = 0 \qquad (2)$
$2 + \lambda(2y + x) = 0 \qquad (3)$
Solving for λ *in* (2),
$\lambda = -\frac{1}{y}$
Substituting for λ *in* (3),
$2 - \frac{1}{y}(2y + x) = 0$
$2 - 2 - \frac{x}{y} = 0$
$\frac{x}{y} = 0$
$x = 0$
Substituting for x in (1),
$y^2 + 0 - 1 = 0$
$y^2 = 1$
$y = \pm 1$
$\lambda = \pm 1$
The local extrema of $f(x, y) = x + 2y$ *with constraint* $g(x, y) = y^2 + xy - 1$ *are* $(0, 1)$, *where* $\lambda = -1$, *and* $(0, -1)$, *where* $\lambda = 1$.

b.

The function $x = \ln(a)$ *can be raised to the e,*
$e^x = a$
Moving all terms to one side,
$f(x, a) = e^x - a = 0$
Given the update equation in Newton's algorithm,
$x_{n+1} = x_n - \frac{f(x,a)}{f'(x,a)}$
the derived function $f(x, a) = e^x - a$ *can be substituted,*
$x_{n+1} = x_n - \frac{e^x - a}{e^x}$

Question 7.

a.

b.

$R_A(x) = \frac{x^{\mathrm{T}}Ax}{x^{\mathrm{T}}x}$
Because any vector $x \in \mathbb{R}^n$ *can be represented as*
$x = \sum_{i=1}^{n} \left(\xi_i^{\mathrm{T}} x\right) \xi_i,$ $R_A(x)$ *can be rewritten as*

$$= \frac{\left(\sum_{i=1}^{n}(\xi_i^{\mathrm{T}}x)\xi_i\right)^{\mathrm{T}} A \left(\sum_{i=1}^{n}(\xi_i^{\mathrm{T}}x)\xi_i\right)}{\left(\sum_{i=1}^{n}(\xi_i^{\mathrm{T}}x)\xi_i\right)^{\mathrm{T}}\left(\sum_{i=1}^{n}(\xi_i^{\mathrm{T}}x)\xi_i\right)}$$

For eigenvector ξ_1 *with the maximum eigenvalue of* A, λ_1 ,

$$= \frac{\left(\sum_{i=1}^{n}(\xi_i^{\mathrm{T}}\xi_1)\xi_i\right)^{\mathrm{T}} A \left(\sum_{i=1}^{n}(\xi_i^{\mathrm{T}}\xi_1)\xi_i\right)}{\left(\sum_{i=1}^{n}(\xi_i^{\mathrm{T}}\xi_1)\xi_i\right)^{\mathrm{T}}\left(\sum_{i=1}^{n}(\xi_i^{\mathrm{T}}\xi_1)\xi_i\right)}$$

Because all eigenvectors of A *are orthogonal, the product* $\left(\xi_i^{\mathrm{T}}\xi_1\right)\xi_i$ *of all* $\xi_i \neq \xi_1$ *is* 0 *and thus*
$\sum_{i=1}^{n} \left(\xi_i^{\mathrm{T}}\xi_1\right)\xi_i = \left(\xi_1^{\mathrm{T}}\xi_1\right)\xi_1 = \xi_1$
Thus,
$= \frac{\xi_1{}^{\mathrm{T}}A\xi_1}{\xi_1{}^{\mathrm{T}}\xi_1}$
$= \frac{\xi_1{}^{\mathrm{T}}\lambda_1\xi_1}{\xi_1{}^{\mathrm{T}}\xi_1}$
$= \lambda_1\left(\frac{\xi_1{}^{\mathrm{T}}\xi_1}{\xi_1{}^{\mathrm{T}}\xi_1}\right)$
$= \lambda_1$
Since the eigenvalues of A *are enumerated by decreasing size and* $R_A(\xi_i) = \lambda_i$, λ_1*is the largest possible eigenvalue and the maximum value for* $R_A(x)$ *for all* $x \in \mathbb{R}^n$. *Thus,* $R_A(x)$ *can be at most* λ_1 *and* $R_A(x) \leq \lambda_1$.

c.

Let $\xi_j \in \mathbb{R}^n$ *be an eigenvector of* A *not contained in* $\{\xi_1, ...\xi_k\}$, *the span of all eigenvectors of* A *with maximum eigenvalue* λ_1, *and let* ξ_j *have an eigenvalue* $\lambda_j \neq \lambda_1$ *where each* λ_i *is unique and enumerated by decreasing size such that* $\lambda_1 > \lambda_2 > \ldots > \lambda_n$.
Then, $R_A(\xi_j) = \frac{\xi_j{}^{\mathrm{T}}A\xi_j}{\xi_j{}^{\mathrm{T}}\xi_j}$
$= \frac{\xi_j{}^{\mathrm{T}}\lambda_j\xi_j}{\xi_j{}^{\mathrm{T}}\xi_j}$
$= \lambda_j\frac{\xi_j{}^{\mathrm{T}}\xi_j}{\xi_j{}^{\mathrm{T}}\xi_j}$
$= \lambda_j$
By definition, $\lambda_j \neq \lambda_1$ *and* λ_1 *is the maximum eigenvalue,* $\lambda_1 > \lambda_2 > \ldots > \lambda_n$ *and maximum possible value of* $R_A(x)$ *for* $x \in \mathbb{R}^n$. *Therefore, for all* $\xi_j \notin \{\xi_1, ...\xi_k\}$ *with eigenvalue* λ_1,
$R_A(\xi_j) = \lambda_j < \lambda_1$.