COMP0086

Yee Jie Thay, 20129714

Assignment 01

Unsupervised & Probabilistic Learning

2020-10-21

yee.thay.20@ucl.ac.uk

1. Statistics and Distributions

   (a) Natural parameters and sufficient statistics of 7 distributions below

   Exponential family can be expressed in the below format

$$p(x|\theta) = f(x)g(\theta)e^{\phi(\theta)^T\mathbf{T}(x)}$$

   i. Multivariate Normal

   For Multivariate Normal Distribution where $\mathbf{x}$ is a vector of $n \times 1$, $\boldsymbol{\mu}$ is a vector of $n \times 1$ and $\boldsymbol{\Sigma}$ is a positive definite symmetric matrix of $n \times n$

$$\begin{aligned}
p(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) &= \frac{1}{(2\pi)^{\frac{n}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}}exp(-\frac{1}{2}(\mathbf{x}^T-\boldsymbol{\mu}^T)\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}))\\
&= \frac{1}{(2\pi)^{\frac{n}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}}exp(-\frac{1}{2}(\mathbf{x}^T-\boldsymbol{\mu}^T)(\boldsymbol{\Sigma}^{-1}\mathbf{x}-\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}))\\
&= \frac{1}{(2\pi)^{\frac{n}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}}exp(-\frac{1}{2}(\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}-\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}-\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}+\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}))\\
&= \frac{1}{(2\pi)^{\frac{n}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}}exp(-\frac{1}{2}(\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}-2\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}+\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}))\\
&= \frac{1}{(2\pi)^{\frac{n}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}}exp(-\frac{1}{2}(\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x})+\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})exp(-\frac{(\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})}{2})\\
&= \frac{1}{(2\pi)^{\frac{n}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}}exp(-\frac{(\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})}{2})exp(-\frac{\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}}{2}+\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})\\
&= \left(\frac{1}{(2\pi)^{\frac{n}{2}}}\right)\left(\frac{1}{|\boldsymbol{\Sigma}|^{\frac{1}{2}}}exp(-\frac{(\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})}{2})\right)exp(\frac{-\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}}{2}+\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})\\
&= \left(\frac{1}{(2\pi)^{\frac{n}{2}}}\right)\left(\frac{1}{|\boldsymbol{\Sigma}|^{\frac{1}{2}}}exp(-\frac{(\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})}{2})\right)exp(\frac{-\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}}{2})exp(\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})\\
&= \left(\frac{1}{(2\pi)^{\frac{n}{2}}}\right)\left(\frac{1}{|\boldsymbol{\Sigma}|^{\frac{1}{2}}}exp(-\frac{(\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})}{2})\right)exp(\frac{tr(-\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x})}{2})exp(\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})
\end{aligned}$$

(Real number's trace is itself)

$$= \left(\frac{1}{(2\pi)^{\frac{n}{2}}}\right)\left(\frac{1}{|\boldsymbol{\Sigma}|^{\frac{1}{2}}}exp(-\frac{(\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})}{2})\right)exp\left(\frac{tr(-\boldsymbol{\Sigma}^{-1}\mathbf{x}\mathbf{x}^T)}{2}\right)exp\left(\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right)$$

(Cyclic nature of trace)

   Thus, from here we can see that

$$\begin{aligned}
f(x) &= \frac{1}{(2\pi)^{\frac{n}{2}}}\\
g(\theta) &= (\frac{1}{|\boldsymbol{\Sigma}|^{\frac{1}{2}}}exp(-\frac{(\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})}{2}))\\
\phi(\theta) &= (\frac{-\boldsymbol{\Sigma}^{-1}}{2},\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})\\
\mathbf{T}(x) &= (\mathbf{x}\mathbf{x}^T,\mathbf{x})
\end{aligned}$$

   ii. Binomial

   For Binomial Distribution where p is the probability of one event whilst $1-p$ is the probability of the other event and N is the number of trials

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

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

(Transform using exponential of its natural logarithm)

$$= \binom{N}{x} exp(x\, ln(p) + (N-x)\, ln(1-p))$$

$$= \binom{N}{x} exp(x\, ln(p) + N\, ln(1-p) - x\, ln(1-p))$$

$$= \binom{N}{x} exp(x\, (ln(p) - ln(1-p)) + N\, ln(1-p))$$

$$= \binom{N}{x} exp(N\, ln(1-p)) exp(x\, (ln(\frac{p}{1-p}))$$

Thus, from here we can see that

$$f(x) = \binom{N}{x}$$

$$g(\theta) = exp(Nln(1-p))$$

$$\phi(\theta) = ln(\frac{p}{1-p})$$

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

iii. Multinomial

For Multinomial Distribution where $\mathbf{p}$ is a vector of $(D-1) \times 1$, $\mathbf{x}$ is a vector of $D \times 1$, N is the number of trials such that $\sum_{i=1}^{D} x_i = N$ and $\sum_{i=1}^{D} p_i = 1$

$$p(x|N, \mathbf{p}) = \frac{N!}{x_1!x_2!...x_D!} \prod_{i=1}^{D} p_i^{x_i}$$

$$= \frac{N!}{x_1!x_2!...x_D!} exp(ln(\prod_{i=1}^{D} p_i^{x_i}))$$

(Transform using exponential of its natural logarithm)

$$= \frac{N!}{x_1!x_2!...x_D!} exp(ln(\prod_{i=1}^{D} p_i^{x_i}))$$

$$= \frac{N!}{x_1!x_2!...x_D!} exp(\sum_{i=1}^{D} x_i ln(p_i))$$

$$= \frac{N!}{x_1!x_2!...x_D!} exp\left(\sum_{i=1}^{D-1} (x_i ln(p_i)) + x_D ln(p_D)\right)$$

$$= \frac{N!}{\prod_{i=1}^{D} x_i!} exp\left(\sum_{i=1}^{D-1} x_i ln(p_i) + \left(N - \sum_{i=1}^{D-1} x_i\right) ln\left(1 - \sum_{i=1}^{D-1} p_i\right)\right)$$

$$= \frac{N!}{\prod_{i=1}^{D} x_i!} exp\left(\sum_{i=1}^{D-1} x_i ln(p_i) + \left(N - \sum_{i=1}^{D-1} x_i\right) ln\left(p_D\right)\right)$$

$$= \frac{N!}{\prod_{i=1}^{D} x_i!} exp\left(\sum_{i=1}^{D-1} x_i ln\left(\frac{p_i}{p_D}\right)\right) exp\left(Nln\left(p_D\right)\right)$$

$$= \frac{N!}{\prod_{i=1}^{D} x_i!} exp\left(Nln\left(p_D\right)\right) exp\left(\sum_{i=1}^{D-1} x_i ln\left(\frac{p_i}{p_D}\right)\right)$$

Thus, from here we can see that

$$f(x) = \frac{N!}{\prod_{i=1}^{D} x_i!}$$

$$g(\theta) = exp\left(Nln\left(p_D\right)\right)$$
$$\phi(\theta) = ln\left(\frac{\mathbf{p}}{p_D}\right)$$
$$\mathbf{T}(x) = \mathbf{x}$$

iv. Poisson

For Poisson Distribution,

$$\begin{aligned}
p(x|\mu) &= \frac{\mu^x\, e^{-\mu}}{x!} \\
&= \frac{1}{x!}\, exp(ln(\mu^x\, e^{-\mu})) \qquad \text{(Transform using exponential of its natural logarithm)} \\
&= \frac{1}{x!}\, exp(x\, ln(\mu) - \mu) \\
&= \frac{1}{x!}\, exp(-\mu) exp(x\, ln(\mu))
\end{aligned}$$

Thus, from here we can see that

$$\begin{aligned}
f(x) &= \frac{1}{x!} \\
g(\theta) &= exp(-\mu) \\
\phi(\theta) &= ln(\mu) \\
\mathbf{T}(x) &= x
\end{aligned}$$

v. Beta

For Beta Distribution where $B(\alpha, \beta)$ is $\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ and $\Gamma(n) = (n-1)!$

$$\begin{aligned}
p(x|\alpha, \beta) &= \frac{1}{B(\alpha, \beta)} x^{(\alpha-1)}(1-x)^{(\beta-1)} \\
&= \frac{1}{B(\alpha, \beta)} exp(ln(x^{(\alpha-1)}(1-x)^{(\beta-1)})) \\
&\qquad \text{(Transform using exponential of its natural logarithm)} \\
&= \frac{1}{B(\alpha, \beta)} exp((\alpha-1)ln(x) + (\beta-1)ln(1-x)) \\
&= \frac{1}{B(\alpha, \beta)} exp(\alpha\, ln(x) - ln(x) + \beta\, ln(1-x) - ln(1-x)) \\
&= \frac{1}{B(\alpha, \beta)} exp(-(ln(x) + ln(1-x)) + \alpha\, ln(x) + \beta\, ln(1-x)) \\
&= \frac{1}{B(\alpha, \beta)} exp(-ln(x(1-x)) + \alpha\, ln(x) + \beta\, ln(1-x)) \\
&= \frac{1}{x(1-x)} \frac{1}{B(\alpha, \beta)} exp(\alpha\, ln(x) + \beta\, ln(1-x))
\end{aligned}$$

Thus, from here we can see that

$$\begin{aligned}
f(x) &= \frac{1}{x(1-x)} \\
g(\theta) &= \frac{1}{B(\alpha, \beta)} \qquad \text{(where we note that } B(.) = \tfrac{(\alpha-1)!(\beta-1)!}{(\alpha+\beta-1)!}\text{)} \\
\phi(\theta) &= (\alpha, \beta) \\
\mathbf{T}(x) &= (ln(x), ln(1-x))
\end{aligned}$$

vi. Gamma

For Gamma Distribution where $\Gamma(n) = (n-1)!$

$$\begin{aligned}
p(x|\alpha,\beta) &= \frac{\beta^\alpha}{\Gamma(\alpha)} x^{(\alpha-1)} e^{(-\beta x)} \\
&= \frac{1}{\Gamma(\alpha)} exp(ln(\beta^\alpha\, x^{(\alpha-1)} e^{(-\beta x)})) \\
&\qquad \text{(Transform using exponential of its natural logarithm)} \\
&= \frac{1}{\Gamma(\alpha)} exp(\alpha ln(\beta) + (\alpha-1)ln(x) + (-\beta x)) \\
&= \frac{e^{\alpha ln(\beta)}}{\Gamma(\alpha)} exp((\alpha-1)ln(x) + (-\beta)x)
\end{aligned}$$

Thus, from here we can see that

$$\begin{aligned}
f(x) &= 1 \\
g(\theta) &= \frac{e^{\alpha ln(\beta)}}{\Gamma(\alpha)} \\
\phi(\theta) &= (\alpha - 1, -\beta) \\
\mathbf{T}(x) &= (ln(x), x)
\end{aligned}$$

vii. Dirichlet

For Dirichlet Distribution where $\Gamma(n) = (n-1)!$

$$\begin{aligned}
p(x|\alpha) &= \frac{\Gamma(\sum_{d=1}^D \alpha_d)}{\prod_{d=1}^D \Gamma(\alpha_d)} \prod_{d=1}^D x_d^{\alpha_d - 1} \\
&= \Gamma(\sum_{d=1}^D \alpha_d) exp(ln((\prod_{d=1}^D \Gamma(\alpha_d))^{-1} \prod_{d=1}^D x_d^{\alpha_d-1})) \\
&\qquad \text{(Transform using exponential of its natural logarithm)} \\
&= exp(ln(\Gamma(\sum_{d=1}^D \alpha_d)) - \sum_{d=1}^D ln(\Gamma(\alpha_d))) exp(\sum_{d=1}^D (\alpha_d - 1) ln(x_d))
\end{aligned}$$

Thus, from here we can see that

$$\begin{aligned}
f(x) &= 1 \\
g(\theta) &= exp(ln(\Gamma(\sum_{d=1}^D \alpha_d)) - \sum_{d=1}^D ln(\Gamma(\alpha_d))) \\
\phi(\theta) &= (\alpha_1 - 1, \alpha_2 - 1, ..., \alpha_D - 1) \\
\mathbf{T}(x) &= (ln(x_1), ln(x_2), ..., ln(x_D))
\end{aligned}$$

(b) Using the general form of exponential families, we note the below

$$p(x|\theta) = f(x)g(\theta)e^{\phi(\theta)^T \mathbf{T}(x)}$$

Given that this is normalised such that it sums to 1

$$\int p(x|\theta)dx = \int f(x)g(\theta)e^{\phi(\theta)^T\mathbf{T}(x)}dx = 1 \tag{1.0}$$

$$\begin{aligned}
g(\theta)^{-1} &= \int f(x)e^{\phi(\theta)^T\mathbf{T}(x)}dx \\
A(\theta) &= -ln(g(\theta)) \\
&= ln(\int f(x)e^{\phi(\theta)^T\mathbf{T}(x)}dx) \\
&= ln(Q(\theta)) \qquad \text{(Set } Q(\theta) = \textstyle\int f(x)e^{\phi(\theta)^T\mathbf{T}(x)}dx\text{)}
\end{aligned}$$

$$\begin{aligned}
\frac{dA}{d(\phi(\theta))} &= \frac{1}{Q(\theta)} \times \frac{dQ(\theta)}{d(\phi(\theta))} = \frac{Q'(\theta)}{Q(\theta)} \\
&= \frac{\int \mathbf{T}(x) f(x) e^{\phi(\theta)^T \mathbf{T}(x)} dx}{\int f(x) e^{\phi(\theta)^T \mathbf{T}(x)} dx} \\
&= \frac{\int \mathbf{T}(x) f(x) g(\theta) e^{\phi(\theta)^T \mathbf{T}(x)} dx}{\int f(x) g(\theta) e^{\phi(\theta)^T \mathbf{T}(x)} dx} \\
&\quad \text{(Adding constant } g(\theta) \text{ to both numerator and denominator)} \\
&= \frac{\int \mathbf{T}(x) f(x) g(\theta) e^{\phi(\theta)^T \mathbf{T}(x)} dx}{1} && \text{(From 1.0)} \\
&= \mathbf{E}(\mathbf{T}(x))
\end{aligned}$$

With that in mind, we can easily apply the solution to the various distribution based on the exponential family form

i. Multivariate Normal

For Multivariate Normal Distribution, we know that

$$\begin{aligned}
f(x) &= \frac{1}{(2\pi)^{\frac{n}{2}}} \\
g(\theta) &= (\frac{1}{|\mathbf{\Sigma}|^{\frac{1}{2}}} exp(-\frac{(\boldsymbol{\mu}^T \mathbf{\Sigma}^{-1} \boldsymbol{\mu})}{2})) \\
\phi(\theta) = \boldsymbol{\eta} &= (\mathbf{\Sigma}^{-1}\boldsymbol{\mu}, \frac{-\mathbf{\Sigma}^{-1}}{2}) \\
\mathbf{T}(x) &= (\mathbf{x}, \mathbf{x}\mathbf{x}^T) \\
\boldsymbol{\eta}_2 &= \frac{-\mathbf{\Sigma}^{-1}}{2} && (1.1) \\
\mathbf{\Sigma}^{-1} &= -2\boldsymbol{\eta}_2 \\
(\mathbf{\Sigma}^{-1})^{-1} &= -\frac{1}{2}\boldsymbol{\eta}_2^{-1} \\
\mathbf{\Sigma} &= -\frac{1}{2}\boldsymbol{\eta}_2^{-1} && (1.2) \\
\boldsymbol{\eta}_1 &= \mathbf{\Sigma}^{-1}\boldsymbol{\mu} && (1.3) \\
\mathbf{\Sigma}\boldsymbol{\eta}_1 &= \mathbf{\Sigma}\mathbf{\Sigma}^{-1}\boldsymbol{\mu} \\
\boldsymbol{\mu} &= \mathbf{\Sigma}\boldsymbol{\eta}_1 \\
\boldsymbol{\mu} &= -\frac{1}{2}\boldsymbol{\eta}_2^{-1}\boldsymbol{\eta}_1 && \text{(Using results from 1.12, we get 1.4)} \\
A(\theta) &= -ln(g(\theta)) \\
&= -ln(\frac{1}{|\mathbf{\Sigma}|^{\frac{1}{2}}} exp(-\frac{(\boldsymbol{\mu}^T \mathbf{\Sigma}^{-1} \boldsymbol{\mu})}{2})) \\
&= \frac{1}{2} ln(|\mathbf{\Sigma}|) + \frac{(\boldsymbol{\mu}^T \mathbf{\Sigma}^{-1} \boldsymbol{\mu})}{2} \\
&= \frac{1}{2} ln(|(-\frac{1}{2}\boldsymbol{\eta}_2^{-1})|) + \frac{((-\frac{1}{2}\boldsymbol{\eta}_2^{-1}\boldsymbol{\eta}_1)^T(-\frac{1}{2}\boldsymbol{\eta}_2^{-1})^{-1}(-\frac{1}{2}\boldsymbol{\eta}_2^{-1}\boldsymbol{\eta}_1))}{2} \\
&= \frac{1}{2} ln(|(-2\boldsymbol{\eta}_2)^{-1}|) - \frac{1}{4}(\boldsymbol{\eta}_2^{-1}\boldsymbol{\eta}_1)^T \boldsymbol{\eta}_2 \boldsymbol{\eta}_2^{-1}\boldsymbol{\eta}_1 \\
&= \frac{1}{2} ln(\frac{1}{|(-2\boldsymbol{\eta}_2)|}) - \frac{1}{4}(\boldsymbol{\eta}_1^T \boldsymbol{\eta}_2^{-1}\boldsymbol{\eta}_1) \\
&= \frac{1}{2}(ln(1) - ln(|(-2\boldsymbol{\eta}_2)|)) - \frac{1}{4}(\boldsymbol{\eta}_1^T \boldsymbol{\eta}_2^{-1}\boldsymbol{\eta}_1) \\
&= \frac{-1}{2} ln(|(-2\boldsymbol{\eta}_2)|) - \frac{1}{4}(\boldsymbol{\eta}_1^T \boldsymbol{\eta}_2^{-1}\boldsymbol{\eta}_1)
\end{aligned}$$

$$
\begin{aligned}
\frac{\partial A(\theta)}{\partial \boldsymbol{\eta_1}} &= \frac{-1}{2}\boldsymbol{\eta}_2^{-1}\boldsymbol{\eta}_1 \\
&= \frac{-1}{2}(\frac{-\boldsymbol{\Sigma}^{-1}}{2})^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \\
&= \boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \\
&= \boldsymbol{\mu} \\
\frac{\partial A(\theta)}{\partial \boldsymbol{\eta_2}} &= \frac{-1}{4}(\boldsymbol{\eta}_1^T(-\boldsymbol{\eta}_2^{-1}\boldsymbol{\eta}_2^{-1})\boldsymbol{\eta}_1) - \frac{1}{2}(\boldsymbol{\eta}_2^{-1})^T \\
& \qquad \text{(Equation 57 and 59 in Matrix Cook Book)} \\
&= \frac{1}{4}(\boldsymbol{\Sigma^{-1}}\boldsymbol{\mu})^T(-2\boldsymbol{\Sigma})(-2\boldsymbol{\Sigma})(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}) + (\boldsymbol{\Sigma}) \\
&= \boldsymbol{\mu}^T(\boldsymbol{\Sigma}^{-1})^T\boldsymbol{\Sigma}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + (\boldsymbol{\Sigma}) \\
&= \boldsymbol{\mu}^T\boldsymbol{\mu} + (\boldsymbol{\Sigma})
\end{aligned}
$$

$$\mathbf{E}(\mathbf{T}(x)) = (\boldsymbol{\mu}, \boldsymbol{\mu}^T\boldsymbol{\mu} + \boldsymbol{\Sigma})$$

ii. Binomial

For Binomial Distribution, we know that

$$
\begin{aligned}
g(\theta) &= exp(Nln(1-p)) \\
\phi(\theta) &= ln(\frac{p}{1-p}) \\
\mathbf{T}(x) &= (x) \\
A(\theta) &= -ln(g(\theta)) \\
&= -ln(exp(Nln(1-p))) \\
&= -Nln(1-p) \\
\phi(\theta) &= ln(\frac{p}{1-p}) \\
e^{\phi(\theta)} &= \frac{p}{1-p} \\
(1-p)e^{\phi(\theta)} &= p \\
e^{\phi(\theta)} &= p(1+e^{\phi(\theta)}) \\
p &= \frac{e^{\phi(\theta)}}{1+e^{\phi(\theta)}} \\
\mathbf{E}(\mathbf{T}(x)) &= \frac{dA(\theta)}{dp}\frac{dp}{d\phi(\theta)} = \frac{N}{1-p}\frac{dp}{d\phi(\theta)} \\
&= \frac{N}{1-p} \times \frac{e^{\phi(\theta)}}{(1+e^{\phi(\theta)})^2} \\
&= \frac{N}{1-p} \times \frac{\frac{p}{1-p}}{(1+\frac{p}{1-p})^2} \\
&= \frac{N}{1-p} \times \frac{\frac{p}{1-p}}{\frac{1}{(1-p)^2}} \\
&= \frac{N}{1-p} \times \frac{p(1-p)^2}{1-p} \\
&= Np
\end{aligned}
$$

iii. Multinomial

For Multinomial Distribution, we know that

$$\begin{aligned}
f(x) &= \frac{N!}{\prod_{i=1}^{D} x_i!} \\
g(\theta) &= exp\left(Nln\left(p_D\right)\right) \\
\phi(\theta) = \boldsymbol{\eta} &= ln\left(\frac{\mathbf{p}}{p_D}\right) \\
\mathbf{p} &= p_D e^{\boldsymbol{\eta}} \\
\mathbf{T}(x) &= \mathbf{x} \\
A(\theta) &= -ln(g(\theta)) \\
&= -ln(exp\left(Nln\left(p_D\right)\right)) \\
&= -Nln\left(1 - \sum_{i=1}^{D-1} p_i\right) && \text{(from } p_D = 1 - \textstyle\sum_{i=1}^{D-1} p_i\text{)} \\
\mathbf{E}(\mathbf{T}(x)) &= \frac{\partial A(\theta)}{\partial \mathbf{p}} \frac{\partial \mathbf{p}}{\partial \eta} \\
&= \frac{N}{p_D}\left(p_D e^{\boldsymbol{\eta}}\right) \\
&= N e^{\boldsymbol{\eta}} \\
&= \frac{N}{p_D}\mathbf{p}
\end{aligned}$$

iv. Poisson

For Poisson Distribution, we know that

$$\begin{aligned}
f(x) &= \frac{1}{x!} \\
g(\theta) &= exp(-\mu) \\
\phi(\theta) = \eta &= ln(\mu) \\
\mathbf{T}(x) &= x
\end{aligned}$$

Expressing in natural parameters

$$\begin{aligned}
\mu &= e^{\eta} \\
A(\theta) &= -ln(g(\theta)) \\
&= -ln(exp(-\mu)) \\
&= \mu \\
&= e^{\eta} \\
\mathbf{E}(\mathbf{T}(x)) &= \frac{dA(\theta)}{d\eta} \\
&= e^{\eta} \\
&= e^{ln(\mu)} \\
&= \mu
\end{aligned}$$

v. Beta

For Beta Distribution, we know that

$$\begin{aligned}
f(x) &= \frac{1}{x(1-x)} \\
g(\theta) &= \frac{1}{B(\alpha, \beta)} && \text{(where we note that } B(.) = \tfrac{(\alpha-1)!(\beta-1)!}{(\alpha+\beta-1)!}\text{)} \\
\phi(\theta) = \boldsymbol{\eta} &= (\alpha, \beta) \\
\mathbf{T}(x) &= (ln(x), ln(1-x))
\end{aligned}$$

Expressing in natural parameters

$$\begin{aligned}
\eta_1 &= \alpha => \alpha = \eta_1 \\
\eta_2 &= \beta => \beta = \eta_2 \\
A(\theta) &= -ln(g(\theta)) \\
&= -ln(\frac{1}{B(\alpha,\beta)}) \\
&= ln(B(\alpha,\beta)) \\
&= ln(\frac{(\alpha-1)!(\beta-1)!}{(\alpha+\beta-1)!}) \\
&= ln(\frac{(\Gamma(\alpha))(\Gamma(\beta))}{\Gamma(\alpha+\beta)} \\
&= ln(\Gamma(\alpha)) + ln(\Gamma(\beta)) - ln(\Gamma(\alpha+\beta)) \\
&= ln(\Gamma(\eta_1)) + ln(\Gamma(\eta_2)) - ln(\Gamma(\eta_1+\eta_2)) \\
\mathbf{E}(\mathbf{T}_1(x)) &= \frac{dA(\theta)}{d\eta_1} \\
&= \frac{\Gamma^{'}(\eta_1)}{\Gamma(\eta_1)} - \frac{\Gamma^{'}(\eta_1+\eta_2)}{\Gamma(\eta_1+\eta_2)} \\
&= \frac{\Gamma^{'}(\alpha)}{\Gamma(\alpha)} - \frac{\Gamma^{'}(\alpha+\beta)}{\Gamma(\alpha+\beta)} \\
\mathbf{E}(\mathbf{T}_2(x)) &= \frac{dA(\theta)}{d\eta_2} \\
&= \frac{\Gamma^{'}(\eta_2)}{\Gamma(\eta_2)} - \frac{\Gamma^{'}(\eta_1+\eta_2)}{\Gamma(\eta_1+\eta_2)} \\
&= \frac{\Gamma^{'}(\beta)}{\Gamma(\beta)} - \frac{\Gamma^{'}(\alpha+\beta)}{\Gamma(\alpha+\beta)}
\end{aligned}$$

vi. Gamma

For Gamma Distribution, we know that

$$\begin{aligned}
f(x) &= 1 \\
g(\theta) &= \frac{e^{\alpha ln(\beta)}}{\Gamma(\alpha)} \\
\phi(\theta) = \boldsymbol{\eta} &= (\alpha-1, -\beta) \\
\mathbf{T}(x) &= (ln(x), x)
\end{aligned}$$

Expressing in natural parameters

$$\begin{aligned}
\eta_1 &= \alpha - 1 => \alpha = \eta_1 + 1 \\
\eta_2 &= -\beta => \beta = -\eta_2 \\
A(\theta) &= -ln(g(\theta)) \\
&= -ln(\frac{e^{\alpha ln(\beta)}}{\Gamma(\alpha)}) \\
&= -\alpha ln(\beta) + ln\Gamma(\alpha) \\
&= ln\Gamma(\eta_1+1) - (\eta_1+1)ln(-\eta_2) \\
\mathbf{E}(\mathbf{T}_1(x)) &= \frac{dA(\theta)}{d\eta_1} \\
&= \frac{\Gamma^{'}(\eta_1+1)}{\Gamma(\eta_1+1)} - ln(-\eta_2)
\end{aligned}$$

$$
\begin{aligned}
&= \ln(\frac{\alpha}{\beta}) \\
\mathbf{E}(\mathbf{T}_2(x)) &= \frac{dA(\theta)}{d\eta_2} \\
&= \frac{\eta_1 + 1}{-\eta_2} \\
&= \frac{\alpha - 1 + 1}{\beta} \\
&= \frac{\alpha}{\beta}
\end{aligned}
$$

vii. Dirichlet

For Dirichlet Distribution, we know that

$$
\begin{aligned}
f(x) &= 1 \\
g(\theta) &= exp(ln(\Gamma(\sum_{d=1}^{D} \alpha_d)) - \sum_{d=1}^{D} ln(\Gamma(\alpha_d))) \\
\phi(\theta) = \boldsymbol{\eta} &= (\alpha_1 - 1, \alpha_2 - 1, ..., \alpha_D - 1) \\
\mathbf{T}(x) &= (ln(x_1), ln(x_2), ..., ln(x_D))
\end{aligned}
$$

Expressing in natural parameters

$$
\begin{aligned}
\boldsymbol{\eta} &= \boldsymbol{\alpha} - 1 => \boldsymbol{\alpha} = \boldsymbol{\eta} + 1 \\
A(\theta) &= -ln(g(\theta)) \\
&= -ln\left(exp\left(ln\Gamma(\sum_{d=1}^{D} \alpha_d)\right) - \sum_{d=1}^{D} ln\Gamma(\alpha_d)\right) \\
&= -ln\Gamma(\sum_{d=1}^{D} \alpha_d) + \sum_{d=1}^{D} ln\Gamma(\alpha_d) \\
&= -ln\Gamma(\sum_{d=1}^{D} \eta_d + 1) + \sum_{d=1}^{D} ln\Gamma(\eta_d + 1) \\
\mathbf{E}(\mathbf{T}(x)) &= \frac{dA(\theta)}{d\boldsymbol{\eta}} \\
&= -\frac{\Gamma^{'}(\sum_{d=1}^{D} \eta_d)}{\Gamma(\sum_{d=1}^{D} \eta_d)} + \frac{\Gamma^{'}(\boldsymbol{\eta})}{\Gamma(\boldsymbol{\eta})}
\end{aligned}
$$

2. (a) Multivariate Normal

Express the maximum-likelihood value of the parameters for Multivariate Normal Distribution above

$$
\begin{aligned}
p(x|\theta) &= f(x)g(\theta)e^{\phi(\theta)^T \mathbf{T}(x)} \\
p(\mathbf{x}|\mu, \mathbf{\Sigma}) &= \frac{1}{(2\pi)^{\frac{d}{2}} |\mathbf{\Sigma}|^{\frac{1}{2}}} exp\left(-\frac{1}{2}(\mathbf{x}^T - \mu^T)\mathbf{\Sigma}^{-1}(\mathbf{x}\mu)\right)
\end{aligned}
$$

Assuming conditional IID,

$$
P(\mathbf{x}|\mu, \mathbf{\Sigma}) = \prod^{N}(\frac{1}{(2\pi)^{\frac{d}{2}} |\mathbf{\Sigma}|^{\frac{1}{2}}} exp\left(-\frac{1}{2}(\mathbf{x}^T - \mu^T)\mathbf{\Sigma}^{-1}(\mathbf{x}\mu)\right))
$$

The log-likelihood function is

$$
ln(P(\mathbf{x}|\mu, \mathbf{\Sigma})) = \frac{-N\, d\, ln(2\pi)}{2} - \frac{Nln(|\mathbf{\Sigma}|)}{2} - \frac{1}{2}\left((\mathbf{x}^T - \mu^T)\mathbf{\Sigma}^{-1}(\mathbf{x} - \mu)\right)
$$

Differentiating with respect to μ,

$$\begin{aligned}
\frac{\partial\left(ln(P(\mathbf{x}|\mu,\boldsymbol{\Sigma}))\right)}{\partial\mu} &= \frac{\partial\left(\frac{-N\,d\,ln(2\pi)}{2} - \frac{Nln(|\boldsymbol{\Sigma}|)}{2} - \frac{1}{2}\left((\mathbf{x}^T-\mu^T)\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\mu)\right)\right)}{\partial\mu} \\
&= (\mathbf{x}-\mu)^T\boldsymbol{\Sigma}^{-1} \\
&= \sum_{i=1}^{N}(x_i-\mu)^T\boldsymbol{\Sigma}^{-1}
\end{aligned}$$

Setting this to zero and noting that $\boldsymbol{\Sigma}^{-1}$ cannot be 0,

$$\begin{aligned}
n\mu &= \sum_{i=1}^{n}(x_i) \\
\mu_{ML} &= \frac{1}{n}\sum_{i=1}^{n}(x_i)
\end{aligned}$$

Similarly for $\boldsymbol{\Sigma}$,

$$\frac{\partial\left(lnP(\mathbf{x}|\mu,\boldsymbol{\Sigma})\right)}{\partial\boldsymbol{\Sigma}} = \frac{\partial\left(\frac{-N\,d\,ln(2\pi)}{2} - \frac{Nln(|\boldsymbol{\Sigma}|)}{2} - \frac{1}{2}\left((\mathbf{x}^T-\mu^T)\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\mu)\right)\right)}{\partial\boldsymbol{\Sigma}}$$

Using Equation 57 in Matrix Cook Book and scalar of trace is itself, we can simplify

$$\begin{aligned}
&= \frac{\partial\Big(\frac{-N\,d\,ln(2\pi)}{2} - \frac{Nln(|\boldsymbol{\Sigma}|)}{2} - \big(\frac{1}{2}tr((\mathbf{x}-\mu)(\mathbf{x}-\mu)^T\boldsymbol{\Sigma}^{-1})\Big)}{\partial\boldsymbol{\Sigma}} \\
&= \frac{N}{2}\Sigma - \frac{1}{2}\sum_{i=1}^{N}(x_i-\mu)(x_i-\mu)^T
\end{aligned}$$

Setting above to zero,

$$\Sigma_{ML} = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)(x_i-\mu)^T$$

(b) Binomial

Express the maximum-likelihood value of the mean parameter for Binomial Distribution

$$\begin{aligned}
p(x|\theta) &= f(x)g(\theta)e^{\phi(\theta)^T\mathbf{T}(x)} \\
p(x|N,p) &= \binom{N}{x}p^x(1-p)^{N-x}
\end{aligned}$$

Assuming conditional IID,

$$P(\mathbf{x}|N,p) = \prod^{N}\binom{N}{x}p^x(1-p)^{N-x}$$

The log-likelihood function is

$$ln(P(x|N,p)) = N\left(lg(\binom{N}{x}) + xln(p) + (N-x)ln(1-p)\right)$$

Differentiating with respect to p,

$$\begin{aligned}
\frac{d(ln(P(x|N,p)))}{dp} &= \frac{d\left(N\left(lg(\binom{N}{x}) + xln(p) + (N-x)ln(1-p)\right)\right)}{dp} \\
&= N\left(\frac{x}{p} - \frac{N-x}{1-p}\right)
\end{aligned}$$

Setting this to zero,

$$\begin{aligned}
\frac{x}{p} &= \frac{N-x}{1-p} \\
x - xp &= Np - xp \\
p_{ML} &= \frac{x}{N}
\end{aligned}$$

(c) Multinomial

Express the maximum-likelihood value of the mean parameter for Multinomial Distribution

$$p(x|\theta) = f(x)g(\theta)e^{\phi(\theta)^T\mathbf{T}(x)}$$

$$p(x|N,\mathbf{p}) = \frac{N!}{x_1!x_2!...x_D!}\prod_{d=1}^{D} p_d^{x_d}$$

The log-likelihood function is

$$ln(P(x|N,\mathbf{p})) = ln(N!) - \sum_{d=1}^{D} ln(x_d) + \sum_{d=1}^{D} x_d ln(p_d)$$

However we need to take into account the constraint where $\sum_{d=1}^{D} p_d = 1$, thus applying the method of Lagrenge multipliers,

$$ln(P(x|N,\mathbf{p})) = ln(N!) - \sum_{d=1}^{D} ln(x_d) + \sum_{d=1}^{D} x_{id} ln(p_d) - \lambda(\sum_{d=1}^{D} p_d - 1)$$

Differentiating with respect to $\mathbf{p}$,

$$\frac{\partial(ln(P(\mathbf{x}|N,\mathbf{p})))}{\partial\mathbf{p}} = \left(\frac{1}{\mathbf{p}} \circ \mathbf{x}\right) - \lambda$$

Differentiating with respect to λ,

$$\frac{\partial(ln(P(\mathbf{x}|N,\mathbf{p})))}{\partial\lambda} = -\sum_{d=1}^{D} p_d + 1$$

Setting these to zero,

$$\mathbf{p} = \frac{\mathbf{x}}{\lambda}$$

$$\sum_{d=1}^{D} p_d = 1$$

For each p_d in $\mathbf{p}$, $p_d = \frac{x_d}{\lambda}$ and thus, $\frac{1}{\lambda}\sum_{d=1}^{D}(x_d) = 1$

$$\sum_{d=1}^{D}(x_d) = \lambda = N \qquad \text{(by definition of Multinomial model)}$$

$$\mathbf{p} = \frac{\mathbf{x}}{N}$$

(d) Poisson

Express the maximum-likelihood value of the mean parameter for Poisson Distribution

$$p(x|\theta) = f(x)g(\theta)e^{\phi(\theta)^T\mathbf{T}(x)}$$

$$p(x|\mu) = \frac{\mu^x\, e^{-\mu}}{x!}$$

Assuming conditional IID,

$$P(\mathbf{x}|\mu) = \prod^{N} \frac{\mu^x\, e^{-\mu}}{x!}$$

The log-likelihood function is

$$ln(P(\mathbf{x}|\mu)) = \sum_{i=1}^{N} (-ln(x_i!) + x_i ln(\mu) - \mu)$$

Differentiating with respect to μ,

$$\begin{aligned}\frac{d(ln(P(\mathbf{x}|\mu)))}{d\mu} &= \frac{d\left(\sum_{i=1}^{N}(-ln(x_i!) + x_i ln(\mu) - \mu)\right)}{d\mu} \\ &= \sum_{i=1}^{N}\left(\frac{x}{\mu} - 1\right) \\ &= N - \frac{1}{\mu}\sum_{i=1}^{N} x_i\end{aligned}$$

Setting this to zero,

$$\sum_{i=1}^{N} x_i = N\mu$$

$$\mu_{ML} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

(e) Beta

Express the maximum-likelihood value of the mean parameter for Beta Distribution

$$p(x|\theta) = f(x)g(\theta)e^{\phi(\theta)^T \mathbf{T}(x)}$$

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

Assuming conditional IID,

$$P(\mathbf{x}|\alpha,\beta) = \prod^{N} \frac{1}{B(\alpha,\beta)}x^{(\alpha-1)}(1-x)^{(\beta-1)}$$

The log-likelihood function is

$$ln(P(\mathbf{x}|\alpha,\beta) = \sum_{i=1}^{N}(-lnB(\alpha,\beta) + (\alpha-1)lnx_i + (\beta-1)ln(1-x_i))$$

Differentiating with respect to α,

$$\frac{\partial(ln(P(\mathbf{x}|\alpha,\beta)))}{\partial\alpha} = \frac{\partial\left(\sum_{i=1}^{N}(ln\Gamma(\alpha+\beta) - ln\Gamma(\alpha) - ln\Gamma(\beta) + (\alpha-1)lnx_i + (\beta-1)ln(1-x_i))\right)}{\partial\alpha}$$

$$= N\frac{\Gamma^{'}(\alpha+\beta)}{\Gamma^(\alpha+\beta)} + N\frac{\Gamma^{'}(\alpha)}{\Gamma^(\alpha)} + \sum_{i=1}^{N} lnx_i$$

Differentiating with respect to β,

$$\frac{\partial(ln(P(\mathbf{x}|\alpha,\beta)))}{\partial\beta} = \frac{\partial\left(\sum_{i=1}^{N}(ln\Gamma(\alpha+\beta) - ln\Gamma(\alpha) - ln\Gamma(\beta) + (\alpha-1)lnx_i + (\beta-1)ln(1-x_i))\right)}{\partial\beta}$$

$$= N\frac{\Gamma^{'}(\alpha+\beta)}{\Gamma^(\alpha+\beta)} + N\frac{\Gamma^{'}(\beta)}{\Gamma^(\beta)} + \sum_{i=1}^{N} ln(1-x_i)$$

Setting these to zero,

$$0 = N\frac{\Gamma^{'}(\alpha+\beta)}{\Gamma^(\alpha+\beta)} + N\frac{\Gamma^{'}(\alpha)}{\Gamma^(\alpha)} + \sum_{i=1}^{N} x_i$$

$$0 = N\frac{\Gamma^{'}(\alpha+\beta)}{\Gamma^(\alpha+\beta)} + N\frac{\Gamma^{'}(\beta)}{\Gamma^(\beta)} + \sum_{i=1}^{N} ln(1-x_i)$$

Believe, this can only be solved numerically and there is no closed form analytical solutions

(f) Gamma

Express the maximum-likelihood value of the mean parameter for Gamma Distribution

$$p(x|\theta) = f(x)g(\theta)e^{\phi(\theta)^T \mathbf{T}(x)}$$

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

Assuming conditional IID,

$$P(\mathbf{x}|\alpha,\beta) = \prod^{N} \frac{\beta^\alpha}{\Gamma(\alpha)}x^{(\alpha-1)}e^{(-\beta x)}$$

The log-likelihood function is

$$ln(P(\mathbf{x}|\alpha,\beta)) = \sum_{i=1}^{N}(\alpha ln\beta - ln\Gamma(\alpha) + (\alpha-1)lnx - \beta x)$$

Differentiating with respect to α,

$$\begin{aligned}
\frac{\partial(ln(P(\mathbf{x}|\alpha,\beta)))}{\partial\alpha} &= \frac{\partial\left(\sum_{i=1}^{N}(\alpha ln\beta - ln\Gamma(\alpha) + (\alpha-1)lnx - \beta x)\right)}{\partial\alpha} \\
&= \sum_{i=1}^{N}\left(ln\beta - \frac{\Gamma^{'}(\alpha)}{\Gamma^{(}\alpha)} + lnx_i\right) \\
&= Nln\beta - N\frac{\Gamma^{'}(\alpha)}{\Gamma^{(}\alpha)} + \sum_{i=1}^{N} lnx_i
\end{aligned}$$

Differentiating with respect to β,

$$\begin{aligned}
\frac{\partial(ln(P(\mathbf{x}|\alpha,\beta)))}{\partial\beta} &= \frac{\partial\left(\sum_{i=1}^{N}(\alpha ln\beta - ln\Gamma(\alpha) + (\alpha-1)lnx - \beta x)\right)}{\partial\beta} \\
&= \sum_{i=1}^{N}\left(\frac{\alpha}{\beta} - x_i\right) \\
&= N\frac{\alpha}{\beta} - \sum_{i=1}^{N} x_i
\end{aligned}$$

Setting these to zero,

$$0 = N\frac{\alpha}{\beta} - \sum_{i=1}^{N} x_i => \frac{\alpha}{\beta} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$0 = Nln\beta - N\frac{\Gamma^{'}(\alpha)}{\Gamma^{(}\alpha)} + \sum_{i=1}^{N} lnx_i$$

Believe, this can only be solved numerically and there is no closed form analytical solutions

(g) Dirichlet

Express the maximum-likelihood value of the mean parameter for Dirichlet Distribution

$$p(x|\theta) = f(x)g(\theta)e^{\phi(\theta)^T\mathbf{T}(x)}$$

$$p(x|\boldsymbol{\alpha}) = \frac{\Gamma(\sum_{d=1}^{D}\alpha_d)}{\prod_{d=1}^{D}\Gamma(\alpha_d)}\prod_{d=1}^{D} x_d^{\alpha_d-1}$$

Assuming conditional IID,

$$P(\mathbf{x}|\boldsymbol{\alpha}) = \prod^{N}\left(\frac{\Gamma(\sum_{d=1}^{D}\alpha_d)}{\prod_{d=1}^{D}\Gamma(\alpha_d)}\prod_{d=1}^{D} x_d^{\alpha_d-1}\right)$$

The log-likelihood function is

$$ln(P(\mathbf{x}|\boldsymbol{\alpha})) = \sum_{i=1}^{N}\left(ln(\Gamma(\sum_{d=1}^{D}\alpha_d)) - \sum_{d=1}^{D} ln\Gamma(\alpha_d) + \sum_{d=1}^{D}(\alpha_d - 1)ln(x_{id})\right)$$

Differentiating with respect to $\boldsymbol{\alpha}$,

$$\begin{aligned}
\frac{d(ln(P(\mathbf{x}|\boldsymbol{\alpha})))}{d\boldsymbol{\alpha}} &= \frac{d\left(\sum_{i=1}^{N}\left(ln(\Gamma(\sum_{d=1}^{D}\alpha_d)) - \sum_{d=1}^{D} ln\Gamma(\alpha_d) + \sum_{d=1}^{D}(\alpha_d - 1)ln(x_{id})\right)\right)}{d\boldsymbol{\alpha}} \\
&= \sum_{i=1}^{N}\left(\frac{\Gamma^{'}(\sum_{d=1}^{D}\alpha_d))}{\Gamma^{(}\sum_{d=1}^{D}\alpha_d))} - \frac{\Gamma^{'}(\alpha_d)}{\Gamma^{(}\alpha_d)} + \sum_{d=1}^{D} lnx_{id}\right) \\
&= N\frac{\Gamma^{'}(\sum_{d=1}^{D}\alpha_d))}{\Gamma^{(}\sum_{d=1}^{D}\alpha_d))} - N\frac{\Gamma^{'}(\alpha_d)}{\Gamma^{(}\alpha_d)} + \sum_{i=1}^{N}\sum_{d=1}^{D} lnx_{id}
\end{aligned}$$

Setting these to zero,

$$0 = \frac{\Gamma^{'}(\sum_{d=1}^{D}\alpha_d))}{\Gamma^{(}\sum_{d=1}^{D}\alpha_d))} - \frac{\Gamma^{'}(\alpha_d)}{\Gamma^{(}\alpha_d)} + \sum_{i=1}^{N}\sum_{d=1}^{D} lnx_{id}$$

Believe, this can only be solved numerically and there is no closed form analytical solutions

3. (a) Explain why a multivariate Gaussian would not be an appropriate model for this data set of images.

   i. Each pixel in the data set image $x_d^{(n)}$ has only 2 states. It is either 1 or 0. Thus, multivariate gaussian which is a continuous distribution will not be a good model. This is especially because the probability of one state of a pixel is directly related to the probability of it being in the other state.

(b) What is the equation for the maximum likelihood (ML) estimate of p? Note that you can solve for p directly

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

Above shows the probability of one image. Given that we have N images that are independent and identically distributed samples, we will need $\prod_i^N P(\mathbf{x}|\mathbf{p})$. Thus, the log-likelihood function is

$$ln(\boldsymbol{P}(\mathbf{x}|\mathbf{p})) = \sum_{i=1}^{N}\left(\sum_{d=1}^{D}(x_{id}lnp_d + (1-x_{id})ln(1-p_d))\right)$$

Differentiating with respect to $\mathbf{p}$,

$$\frac{\partial(ln(\boldsymbol{P}(\mathbf{x}|\mathbf{p})))}{\partial\mathbf{p}} = \frac{\partial\left(\sum_{i=1}^{N}\left(\sum_{d=1}^{D}(x_{id}lnp_d + (1-x_{id})ln(1-p_d))\right)\right)}{\partial\mathbf{p}}$$

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

(where $\mathbf{x}$ and $\mathbf{p}$ are both $D \times 1$ vector and the division in $\frac{\mathbf{x}}{\mathbf{p}}$ denotes element wise division $(\frac{x_1}{p_1}, ..., \frac{x_d}{p_d})^T)$)

Setting these to zero,

$$\sum_{i=1}^{N}\left(\frac{\mathbf{x}_i}{\mathbf{p}} - \frac{1-\mathbf{x}_i}{1-\mathbf{p}}\right) = 0$$

$$\sum_{i=1}^{N}(\frac{1-\mathbf{x}_i}{1-\mathbf{p}}) = \sum_{i=1}^{N}\frac{\mathbf{x}_i}{\mathbf{p}}$$

$$(1-\mathbf{p}) \circ \sum_{i=1}^{N}\mathbf{x}_i = (\mathbf{p}) \circ \sum_{i=1}^{N}(1-\mathbf{x}_i) \qquad \text{(where } \circ \text{ is the Hadamard product)}$$

$$\sum_{i=1}^{N}\mathbf{x}_i = N\mathbf{p}$$

$$\mathbf{p} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i$$

Thus, we can see the maximum likelihood estimator for each parameter is the mean of $\mathbf{x}$

(c) What is the maximum a posteriori (MAP) estimate of p?

Using Likelihood as

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

and Prior as

$$P(p_d) = \frac{1}{B(\alpha,\beta)}p_d^{\alpha-1}(1-p_d)^{\beta-1}$$

$$P(\mathbf{p}) = \prod_{d=1}^{D} P(p_d)$$

The Posterior will be

$$P(\mathbf{p}|\mathbf{x}) \propto P(\mathbf{x}|\mathbf{p}) \times P(\mathbf{p})$$

Applying log transformation

$$ln(\boldsymbol{P}(\mathbf{p}|\mathbf{x})) \propto \sum_{i=1}^{N}\left(\sum_{d=1}^{D}(x_{id}lnp_d + (1-x_{id})ln(1-p_d))\right) +$$
$$\sum_{i=1}^{N}\left(\sum_{d=1}^{D}(-ln(B(\alpha,\beta) + (\alpha-1)ln(p_d) + (\beta-1)ln(1-p_d))\right)$$

Differentiating with respect to $\mathbf{p}$,

$$\frac{\partial(ln(P(\mathbf{p}|\mathbf{x})))}{\partial\mathbf{p}} = \sum_{i=1}^{N}\left(\sum_{d=1}^{D}\left(\frac{x_{id}}{p_d} - \frac{1-x_{id}}{1-p_d} + \frac{\alpha-1}{p_d} - \frac{\beta-1}{1-p_d}\right)\right)$$
$$= \sum_{i=1}^{N}\left(\frac{\mathbf{x}_i}{\mathbf{p}} - \frac{1-\mathbf{x}_i}{1-\mathbf{p}} + \frac{\alpha-1}{\mathbf{p}} - \frac{\beta-1}{1-\mathbf{p}}\right)$$

(where $\mathbf{x}$ and $\mathbf{p}$ are both $D \times 1$ vector and the division in $\frac{\mathbf{x}}{\mathbf{p}}$ denotes element wise division $(\frac{x_1}{p_1}, ..., \frac{x_d}{p_d})^T)$)

Setting these to zero,

$$0 = \sum_{i=1}^{N}\left(\frac{\mathbf{x}_i}{\mathbf{p}} - \frac{1-\mathbf{x}_i}{1-\mathbf{p}} + \frac{\alpha-1}{\mathbf{p}} - \frac{\beta-1}{1-\mathbf{p}}\right)$$

$$(1-\mathbf{p}) \circ \left(\alpha - 1 + \sum_{i=1}^{N}(\mathbf{x}_i)\right) = \mathbf{p} \circ \left(\beta - 1 + \sum_{i=1}^{N}(1-\mathbf{x}_i)\right)$$

$$\alpha - 1 + \sum_{i=1}^{N}(\mathbf{x}_i) = \mathbf{p} \circ \left(\beta - 1 + \sum_{i=1}^{N}(1-\mathbf{x}_i) + \alpha - 1 + \sum_{i=1}^{N}(\mathbf{x}_i)\right)$$

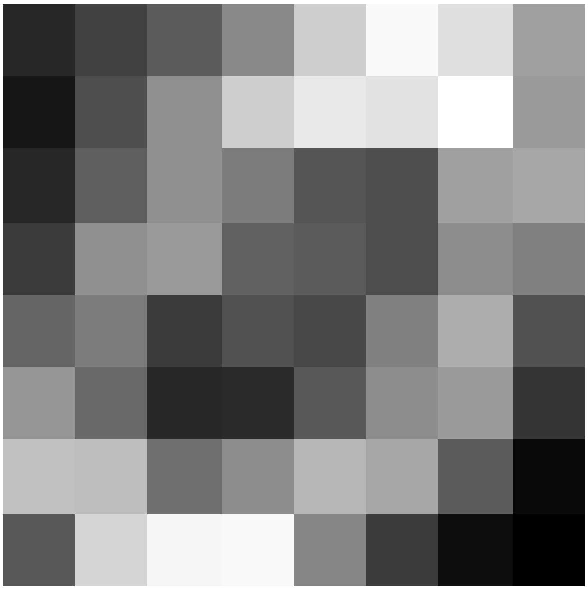
$$\alpha - 1 + \sum_{i=1}^{N}(\mathbf{x}_i) = \mathbf{p} \circ (\beta - 1 + \sum_{i=1}^{N} 1 + \alpha - 1)$$

$$\mathbf{p} = \frac{\alpha - 1 + \sum_{i=1}^{N}(\mathbf{x}_i)}{(\beta + \alpha - 2 + N)}$$

Thus, we can see the maximum a posteriori estimator for each parameter is the weighted mean of $\mathbf{x}$

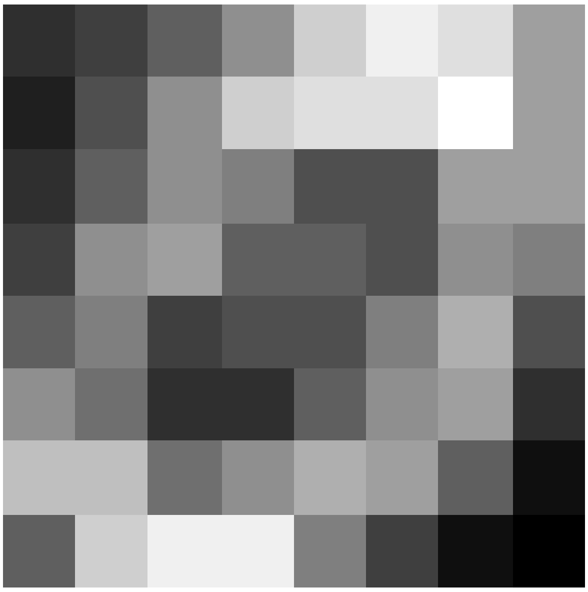
(d) Include a listing of your code within your submission, and a visualisation of the learned parameter vector as an image. (You may use Matlab, Octave or Python)

```python
import numpy as np
from matplotlib import pyplot as plt
Y = np.loadtxt('binarydigits.txt')

# My code as per below
X = Y.T
# Create log likelihood function that we will look to maximise (as positive as possible)
# This looks to tackle each parameter by itself and look through all the 100 data points of each parameter and then optimise p such as to maximise the log likelihood equation given by sum(ln_prob_bern).  Given this is a binary data set, it is not suitable to use gradient descent unless we use some sort of sigmoid function with it.  Alternatively, we can solve it analytically through the use of Y.sum(axis=0)/100 where we find the mean of each parameter as its maximum likelihood estimator
ln_prob_bern = lambda x,p,alpha,beta: x* np.log(p) + (1.0-x)*np.log(1.-p)
def iterate_p_for_params(log_likelihood_func,X,alpha=0,beta=0):
    output = []
    for param in X:
        holder = []
        for increment in range(1,1000):
            p = increment/1000.0
            new = sum(log_likelihood_func(param,p,alpha,beta))
            # Initialise the value of p
            if not holder:
                holder=[new,p]
            # Retain the value of p that maximises the function
            if new>holder[0]:
                holder=[new,p]
        output.append(holder)
    return output

ML = iterate_p_for_params(ln_prob_bern,X)

plt.figure()
plt.imshow(np.reshape(list(zip(*ML))[1], (8,8)),
           interpolation="None",
           cmap=’gray’)
plt.axis(’off’)
```



(e) Modify your code to learn MAP parameters with $\alpha = \beta = 3$. Show the new learned parameter vector for this data set as an image. Explain why this might be better or worse than the ML estimate

i. Below is the new learned parameter vector for this data set as an image.



ii. Visually both images are very similar in terms of where they have given relatively higher probabilities to with regards to the parameters. However if one specifically look at the differences between the exact parameter values, one will notice that there is significant absolute differences that stands out.

  A. E.g. Parameter 64 where ML has given a minimal probability of 0.01 (I used 1% increment). MAP estimator came in at 0.4. Given there are only 100 samples, it has resulted in the relative impact of the Beta distribution being more outsized.

iii. I believe that the MAP estimator does no worse than ML though it is hard to come to a conclusive argument for N=100

  A. This is because the 2 parameter vectors we have produced for both estimators are very similar relatively (i.e. the relative intra-probability of the 64 parameters are similar for both estimators) even though the absolute difference in probability estimated by the 2 methods are significantly different

iv. However, there is also a caveat to this. If the prior is differs drastically from the "truth", it will take a lot of data to eventually overwhelm it. The pre-conceived notion from the prior will only be slowly eliminated with more data and thus, there is a realistic chance that we suffer from tunnel vision. I.e. We used parameters for our estimations based on what we are looking out for and not what the actual data is suggesting to us.

4. Find the expressions needed to calculate the (relative) probability of the three different models and calculate the posterior probabilities of each of the three models having generated the data in binarydigits.txt.
If we use a model with all D components generated from a Bernoulli distribution with $p_d = 0.5$, we get the below

$$\boldsymbol{P}_{0.5}(\mathbf{p}) = \prod_{i=1}^{N} \left(\prod_{d=1}^{D} 0.5^{x_{id}} 0.5^{(1-x_{id})} \right) = 0.5^{ND}$$

$$* = 0.5^{6400}$$

Similarly if we use a model for Bernoulli distributions with unknown, but identical probability

$$\boldsymbol{P}_{\hat{p}}(\mathbf{p}) = \frac{2495}{6400}^{2495} \frac{3905}{6400}^{3905}$$

If we use a model for Bernoulli distributions with separate, unknown $\tilde{p_d}$, it will be as follow below

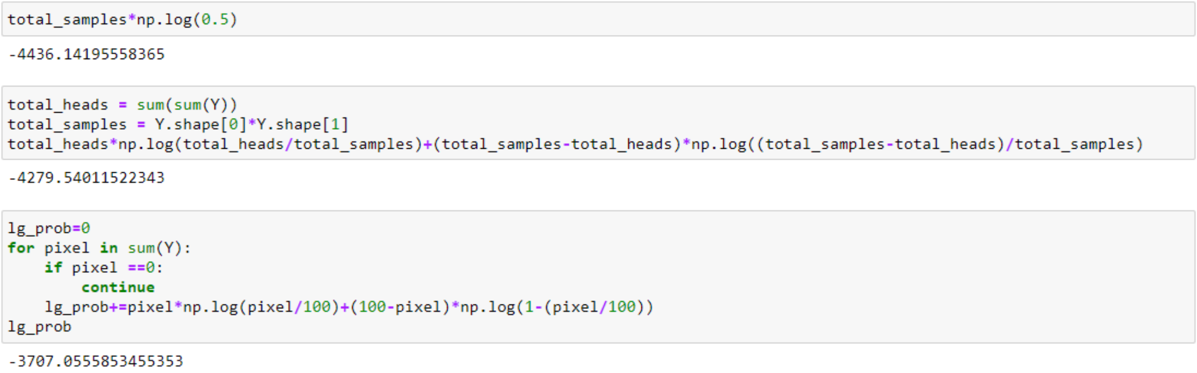
$$\boldsymbol{P}_{\tilde{p_d}}(\mathbf{p}) = \prod_{i}^{N} (p_d{}^{x_i}(1-p_d)^{D-x_i})$$

Assuming all three models are equally likely a priori, I calculated the posteriors for all 3 models where $\hat{p}$ and $\tilde{p_d}$ are both drawn from a uniform random distribution.

I note that the posterior probabilities of each of the three models having generated the data in binarydigits.txt as being proportional to the probability below. The ln prob of Model A,B and C are -4436.14,-4279.54 and -3707.05 respectively, doing exponential transformation back to get probabilities,

$$\boldsymbol{Posterior}_{.5}(\mathbf{p}) = e^{-4436.14} * \frac{1}{3}$$
$$\boldsymbol{Posterior}_{\hat{p}}(\mathbf{p}) = e^{-4279.54} * \frac{1}{3}$$
$$\boldsymbol{Posterior}_{\tilde{p_d}}(\mathbf{p}) = e^{-3707.05} * \frac{1}{3}$$

We note that $\boldsymbol{P}_{\tilde{p_d}}(\mathbf{p})$ is more than a magnitude of e^{500} higher than that of the other 2 models. Thus the relative posterior will have moved from equal probability to each of the three models into one where Model A and Model B gets 0% probability each.



```
total_samples*np.log(0.5)
```

```
-4436.14195558365
```

```
total_heads = sum(sum(Y))
total_samples = Y.shape[0]*Y.shape[1]
total_heads*np.log(total_heads/total_samples)+(total_samples-total_heads)*np.log((total_samples-total_heads)/total_samples)
```

```
-4279.54011522343
```

```
lg_prob=0
for pixel in sum(Y):
    if pixel ==0:
        continue
    lg_prob+=pixel*np.log(pixel/100)+(100-pixel)*np.log(1-(pixel/100))
lg_prob
```

```
-3707.0555853455353
```

5. Let A be a symmetric $n \times n$-matrix, with eigenvalues $\lambda_1, ..., \lambda_n$

(a) Show that the matrix $B = A + cI$, where I is the identity matrix and $c \in \mathbb{R}$, has eigenvalues $\lambda_1 + c, ..., \lambda_n + c$

$$\begin{aligned} (A + cI)\boldsymbol{x}_i &= A\boldsymbol{x}_i + cIx \\ &= \lambda_i \boldsymbol{x}_i + c\boldsymbol{x}_i \\ &= (\lambda_i + c)\boldsymbol{x}_i && \text{(given that } \lambda_i \text{ is an eigenvalue that is a scalar)} \end{aligned}$$

By extrapolation, we can implement this for every $\boldsymbol{x}_i$ which is the eigenvector associated to λ_i. Thus the eigen values of matrix B will be $\lambda_1 + c, ..., \lambda_n + c$

(b) Suppose $\boldsymbol{v}$ and $\boldsymbol{w}$ are eigenvectors of $\boldsymbol{A}$, both with the same eigenvalue λ. Show that any linear combination of $\boldsymbol{v}$ and $\boldsymbol{w}$ is again an eigenvector of $\boldsymbol{A}$. What is its eigenvalue?

Given that $\boldsymbol{v}$ and $\boldsymbol{w}$ are eigenvectors of $\boldsymbol{A}$ with the same eigenvalue of λ, show that any linear combination of v and w is again an eigenvector of $\boldsymbol{A}$. What is its eigenvalue?this means that

$$\begin{aligned} A\boldsymbol{v} &= \lambda\boldsymbol{v}; A\boldsymbol{w} = \lambda\boldsymbol{w} \\ c\lambda\boldsymbol{v} + d\lambda\boldsymbol{w} &= cA\boldsymbol{v} + (d)A\boldsymbol{w} && \text{(where } c \in \mathbb{R} \text{ and } d \in \mathbb{R} \text{ and are scalars)} \end{aligned}$$

$$\lambda(c\boldsymbol{v} + d\boldsymbol{w}) = A(c\boldsymbol{v} + d\boldsymbol{w}) \qquad \text{(Given the distributive nature of matrix)}$$

If we let $\boldsymbol{z} = c\boldsymbol{v} + d\boldsymbol{w}$, we note that $\boldsymbol{z}$ is a linear combination of $\boldsymbol{v}$ and $\boldsymbol{w}$

$$A\boldsymbol{z} = \lambda\boldsymbol{z}$$

This satisfies the condition that $\boldsymbol{z}$ is an eigenvector of $\boldsymbol{A}$ and thus we can see that an linear combination of $\boldsymbol{v}$ and $\boldsymbol{w}$ is also an eigenvector of $\boldsymbol{A}$ and the eigenvalue of this linear combination is λ

6. Optimization

(a) Find the local (!) extrema of the function $f(x, y) := x + 2y$ subject to the constraint $y^2 + xy = 1$

Maximise $f(x, y) := x + 2y$ subject to the constraint $y^2 + xy = 1$

$$\begin{aligned} f(x, y, \lambda) &= x + 2y - \lambda(y^2 + xy - 1) \\ &= x + 2y - \lambda y^2 - \lambda xy + \lambda \end{aligned}$$

Differentiating them by x, y and λ and setting them to zero, we get the system of equations below

$$\frac{\partial f(x, y, \lambda)}{\partial x} = 1 - 2\lambda y = 0 \tag{6.0}$$

$$\frac{\partial f(x, y, \lambda)}{\partial y} = 2 - 2\lambda y - \lambda x = 0 \tag{6.1}$$

$$\frac{\partial f(x, y, \lambda)}{\partial \lambda} = -(y^2 + xy - 1) = 0 \tag{6.2}$$

$$\lambda y = 1 \qquad \text{(From 6.1, we get 6.3)}$$

$$\lambda x = 2 - 2 = 0 \qquad \text{(From 6.1 and 6.3, we get 6.4)}$$

We note that $x = 0$ from 6.4 as $\lambda \neq 0$ as seen from 6.3

$$y^2 = 1 \qquad \text{(From 6.2 and 6.4 with } x = 0\text{)}$$

$$y = 1 => \lambda = 1 \qquad \text{(from 6.3)}$$

$$y = -1 => \lambda = -1 \qquad \text{(from 6.3)}$$

Evaluating $f(x, y)$ at $(x, y) = (0, 1)$ and $(x, y) = (0, -1)$, we get

$$f(x, y) = 2; f(x, y) = -2$$

Thus, we have the solutions for the system of equations at $(x, y, \lambda) = [(0, 1, 1), (0, -1, -1)]$.

(b) Suppose we have a numerical routine to evaluate the exponential function exp(x). How can we compute the function ln(a), for a given $a \in \mathbb{R}_+$, using Newton's method?

i. Derive a function $f(x, a)$ to which Newton's method can be applied to find x such that $x = ln(a)$

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

To evaluate $x = ln(a)$, we first transform the equation into the form of $f(x, a) = 0$. Thus, we set

$$\begin{aligned} x &= ln(a) \\ e^x &= a \\ => f(x, a) &= e^x - a = 0 \end{aligned}$$

ii. Specify the update equation $x_{n+1} = ...$ in Newton's algorithm for this problem.

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$
$$x_{n+1} = x_n - \frac{e^{x_n} - a}{e^{x_n}}$$
$$x_{n+1} = x_n - 1 + \frac{a}{e^{x_n}}$$

7. Use the extreme value theorem of calculus (recall: a continuous function on a compact domain attains its maximum and minimum) to show that $sup_{x \in \mathbb{R}^n}$ that $R_{A(x)}$ is attained.

(a) Extreme Value Theorem states that if f is continuous on a closed interval $[a, b]$, then f has both a minimum and a maximum on the interval

(b) Show that $R_A(x) \leq \lambda_1$

$$\boldsymbol{A}\xi_i = \lambda_i \xi_i \tag{7.0}$$
$$\mathbf{x} = \sum_{i=1}^{n} \xi_i^T \mathbf{x} \xi_i$$
$$\boldsymbol{A}\mathbf{x} = \boldsymbol{A} \sum_{i=1}^{n} \xi_i^T \mathbf{x} \xi_i$$
$$= \sum_{i=1}^{n} (\xi_i^T \mathbf{x}) \boldsymbol{A} \xi_i \qquad \text{(As } (\xi_i^T \mathbf{x}) \text{ is a scalar)}$$
$$= \sum_{i=1}^{n} (\xi_i^T \mathbf{x}) \lambda_i \xi_i \qquad \text{(From 7.0)}$$
$$= \sum_{i=1}^{n} \lambda_i \xi_i^T \mathbf{x} \xi_i$$
$$R_A(\mathbf{x}) = \frac{\mathbf{x}^T \boldsymbol{A} \mathbf{x}}{\mathbf{x}^T \mathbf{x}}$$
$$= \frac{\mathbf{x}^T \sum_{i=1}^{n} \lambda_i \xi_i^T \mathbf{x} \xi_i}{\mathbf{x}^T \mathbf{x}}$$

(Note For each i and j element in ξ vector, $\langle \xi_i, \xi_j \rangle = 1$ when $k = j$ and $\langle \xi_i, \xi_j \rangle = 0$ when $i \neq$j)

$$= \frac{\sum_{i=1}^{n} \lambda_i x_i^2}{\sum_{i=1}^{n} x_i^2}$$

Assume $R_A(x) > \lambda_1$ and we bring $\sum_{i=1}^{n} x_i^2$ to the RHS,

$$\lambda_1 x_1^2 + ... + \lambda_n x_n^2 > \lambda_1 x_1^2 + ... + \lambda_1 x_n^2$$

However, this will never hold as $\lambda_1 > ... > \lambda_n$. Thus, we have proven that $R_A(x)$ will never be more than λ_1 so, by contradiction, will always be less than or equal to λ_1.