1 Moment-generating function of the sufficient statistic

The probability density or mass function of an exponential family member takes the following form

$$P(x|\theta) = f(x)g(\theta)e^{\phi(\theta)^\top \mathsf{T}(x)}$$

where θ is the conventional parameter, $\phi(\theta)$ is the natural parameter, and $\mathsf{T}(x)$ is the sufficient statistic.
We can also write the probability density function in terms of the natural parameter $\phi(\theta) \equiv \eta$,

$$P(x|\theta) = f(x)e^{\eta^\top \mathsf{T}(x) - A(\eta)}$$

where $A(\eta) = -\ln g(\theta)$ is called the log-partition function.
We show that taking derivative of the log-partition function with respect to the natural parameter generates moments of the sufficient statistic. First, we have that a probability density function always integrate to 1

$$e^{-A(\eta)} \int_x f(x)e^{\eta^\top \mathsf{T}(x)} dx = 1$$

We take derivative of both sides with respect to η and obtain

$$\begin{aligned}
&\frac{de^{-A(\eta)}}{d\eta} \int_x f(x)e^{\eta^\top \mathsf{T}(x)} dx + e^{-A(\eta)} \int_x f(x) \frac{de^{\eta^\top \mathsf{T}(x)}}{d\eta} dx = 0 \\
\Rightarrow \quad & -\frac{dA(\eta)}{d\eta} e^{-A(\eta)} \int_x f(x)e^{\eta^\top \mathsf{T}(x)} dx + e^{-A(\eta)} \int_x f(x)\mathsf{T}(x)e^{\eta^\top \mathsf{T}(x)} dx = 0
\end{aligned}$$

We notice that the first term is the derivative of the log-partition function and the second term is the expectation of the sufficient statistic

$$-\frac{dA(\eta)}{d\eta} \int_x f(x)e^{\eta^\top \mathsf{T}(x)} e^{-A(\eta)} dx + \int_x f(x)\mathsf{T}(x)e^{\eta^\top \mathsf{T}(x)} e^{-A(\eta)} dx = -\frac{dA(\eta)}{d\eta} + \langle \mathsf{T}(x) \rangle = 0$$

Therefore, we arrive at

$$\langle \mathsf{T}(x) \rangle = \frac{dA(\eta)}{d\eta}$$

A more general fact is that higher-order derivatives of the log-partition function generate higher-order moments of the sufficient statistics.

2 Multivariate normal

A multivariate normal distribution is expressed as

$$\begin{aligned}
\mathbf{x} &\sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\
P(\mathbf{x}) &= |2\pi\boldsymbol{\Sigma}|^{-1/2} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}
\end{aligned}$$

The probability density function can be written as

$$\begin{aligned}
P(\mathbf{x}) &= |2\pi\boldsymbol{\Sigma}|^{-1/2} \exp\left[-\frac{1}{2}\left(\mathbf{x}^\top\Sigma^{-1}\mathbf{x} - \mathbf{x}^\top\Sigma^{-1}\boldsymbol{\mu} - \boldsymbol{\mu}^\top\Sigma^{-1}\mathbf{x} + \boldsymbol{\mu}^\top\Sigma^{-1}\boldsymbol{\mu}\right)\right] \\
&= |2\pi\boldsymbol{\Sigma}|^{-1/2} e^{-\frac{1}{2}\boldsymbol{\mu}^\top\Sigma^{-1}\boldsymbol{\mu}} \exp\left[\boldsymbol{\mu}^\top\Sigma^{-1}\mathbf{x} - \frac{1}{2}\mathbf{x}^\top\Sigma^{-1}\mathbf{x}\right]
\end{aligned}$$

Hence, the base measure f, the normalizer g, the natural parameters ϕ, the sufficient statistic T are

$$\begin{aligned}
f(\mathbf{x}) &= 1 \\
g(\boldsymbol{\mu}, \Sigma) &= |2\pi\boldsymbol{\Sigma}|^{-1/2} e^{-\frac{1}{2}\boldsymbol{\mu}^\top\Sigma^{-1}\boldsymbol{\mu}} \\
\phi(\boldsymbol{\mu}, \Sigma) &= \begin{bmatrix} \Sigma^{-1}\boldsymbol{\mu} \\ -\frac{1}{2}\mathsf{vec}(\Sigma^{-1}) \end{bmatrix} \\
\mathsf{T}(\mathbf{x}) &= \begin{bmatrix} \mathbf{x} \\ \mathsf{vec}(\mathbf{x}\mathbf{x}^\top) \end{bmatrix}
\end{aligned}$$

We denote the natural parameters as

$$\boldsymbol{\eta}_1 = \Sigma^{-1}\boldsymbol{\mu}, \quad \boldsymbol{\eta}_2 = -\frac{1}{2}\Sigma^{-1}$$

The log-partition function can be expressed in terms of the natural parameters as

$$\begin{aligned}\boldsymbol{\mu} &= -\frac{1}{2}\boldsymbol{\eta}_2^{-1}\boldsymbol{\eta}_1 \\ \Sigma &= -\frac{1}{2}\boldsymbol{\eta}_2^{-1}\end{aligned}$$

Take the negative logarithm of $g(\boldsymbol{\mu}, \Sigma)$ and re-expressing it with $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2$, we get $A(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2)$.

$$A(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2) = -\ln g(\theta) = \frac{D}{2}\ln 2\pi + \frac{1}{2}\ln\left|-\frac{1}{2}\boldsymbol{\eta}_2^{-1}\right| - \frac{1}{4}\boldsymbol{\eta}_1^{\mathsf{T}}\boldsymbol{\eta}_2^{-1}\boldsymbol{\eta}_1$$

Taking derivatives of $A(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2)$ gives us the expectation of the sufficient statistics

$$\begin{aligned}\langle \mathbf{x} \rangle &= \frac{dA(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2)}{d\boldsymbol{\eta}_1} = -\frac{1}{2}\boldsymbol{\eta}_2^{-1}\boldsymbol{\eta}_1 = \boldsymbol{\mu} \\ \langle \mathbf{x}\mathbf{x}^{\mathsf{T}} \rangle &= \frac{dA(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2)}{d\boldsymbol{\eta}_2} \\ &= \frac{1}{2}\frac{d\ln\left|-\frac{1}{2}\boldsymbol{\eta}_2^{-1}\right|}{d(-\frac{1}{2}\boldsymbol{\eta}_2^{-1})}\frac{d(-\frac{1}{2}\boldsymbol{\eta}_2^{-1})}{d\boldsymbol{\eta}_2} - \frac{1}{4}\frac{\boldsymbol{\eta}_1^{\mathsf{T}}\boldsymbol{\eta}_2^{-1}\boldsymbol{\eta}_1}{d\boldsymbol{\eta}_2} \\ &= \frac{1}{2}(-2\boldsymbol{\eta}_2)\frac{1}{2}\boldsymbol{\eta}_2^{-1}\boldsymbol{\eta}_2^{-1} + \frac{1}{4}\boldsymbol{\eta}_2^{-1}\boldsymbol{\eta}_1\boldsymbol{\eta}_1^{\mathsf{T}}\boldsymbol{\eta}_2^{-1} \\ &= \Sigma + \boldsymbol{\mu}\boldsymbol{\mu}^{\mathsf{T}}\end{aligned}$$

In computing derivatives, we used several matrix calculus facts:

$$\begin{aligned}\frac{d}{da}a^{\mathsf{T}}Ma &= 2Ma \\ \frac{d}{dM}a^{\mathsf{T}}M^{-1}a &= -M^{-\mathsf{T}}aa^{\mathsf{T}}M^{-\mathsf{T}} \\ \frac{d}{dM}\ln|M| &= M^{-\mathsf{T}}\end{aligned}$$

where $a \in \mathbb{R}^D, M \in \mathbb{R}^{D\times D}$. The Matrix Cookbook is recommended for looking up such matrix calculus facts.

3 Binomial

A binomial distribution is expressed as

$$\begin{aligned}x &\sim \mathsf{Binom}(p) \\ P(x) &= \binom{N}{x}p^x(1-p)^{(N-x)}\end{aligned}$$

The probability function can be equivalently written as

$$\begin{aligned}P(x) &= \binom{N}{x}p^x(1-p)^{(N-x)} \\ &= \binom{N}{x}\exp\left[x\ln p + (N-x)\ln(1-p)\right] \\ &= \binom{N}{x}\exp\left[x\ln\frac{p}{1-p} + N\ln(1-p)\right] \\ &= \binom{N}{x}e^{N\ln(1-p)}\exp\left(x\ln\frac{p}{1-p}\right)\end{aligned}$$

Hence, the base measure f, the normalizer g, the natural parameters ϕ, the sufficient statistic T are

$$\begin{aligned}
f(x) &= \binom{N}{x} \\
g(p) &= e^{N \ln(1-p)} \\
\phi(p) &= \ln \frac{p}{1-p} \\
\mathsf{T}(x) &= x
\end{aligned}$$

The expectation of the sufficient statistics can be computed as

$$\begin{aligned}
\langle x \rangle &= -\frac{d \ln g(p)}{dp} \frac{dp}{d\phi(p)} \\
&= -\frac{dN \ln(1-p)}{dp} \frac{dp}{d \ln \frac{p}{1-p}} \\
&= \frac{N}{1-p} \frac{1}{\frac{1}{p} + \frac{1}{1-p}} \\
&= Np
\end{aligned}$$

4 Multinomial

A multinomial distribution is expressed as

$$\begin{aligned}
\mathbf{x} &\sim \mathsf{Multinom}(\mathbf{p}) \\
P(\mathbf{x}) &= \frac{N!}{x_1! x_2! \dots x_D!} \prod_{d=1}^{D} p_d^{x_d}
\end{aligned}$$

The probability function can be equivalently written as

$$\begin{aligned}
P(\mathbf{x}) &= \frac{N!}{x_1! x_2! \dots x_D!} \prod_{d=1}^{D} p_d^{x_d} \\
&= \frac{N!}{x_1! x_2! \dots x_D!} \exp\left(\sum_{d=1}^{D} x_d \ln p_d \right)
\end{aligned}$$

Hence, the base measure f, the normalizer g, the natural parameters ϕ, the sufficient statistic T are

$$\begin{aligned}
f(\mathbf{x}) &= \frac{N!}{x_1! x_2! \dots x_D!} \\
g(\mathbf{p}) &= 1 \\
\phi(\mathbf{p}) &= \begin{bmatrix} \ln p_1 \\ \ln p_2 \\ \vdots \\ \ln p_{D-1} \end{bmatrix} \\
\mathsf{T}(\mathbf{x}) &= \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{D-1} \end{bmatrix}
\end{aligned}$$

Note that the sufficient statistic is $(D-1)$-dimensional, not D-dimensional. This is because we have the constraint $\sum_{d=1}^{D} x_d = N$, and can always infer the D-th dimension if given $(D-1)$ dimensions. The same argument applies to the natural parameters because there is also a constraint $\sum_{d=1}^{D} p_d = 1$.

The expectation of the sufficient statistics can be computed using the definition of expectation

$$
\begin{aligned}
\langle x_d \rangle &= \sum_{x_d=0}^{N} x_d P(\mathbf{x}) \\
&= \sum_{x_d=1}^{N} x_d \frac{N!}{x_1! x_2! \cdots x_D!} p_1^{x_1} p_2^{x_2} \cdots p_D^{x_D} \\
&= \sum_{x_d=1}^{N} x_d \frac{N}{x_d} p_d \cdot \frac{(N-1)!}{x_1! \cdots (x_d-1)! \cdots x_D!} p_1^{x_1} \cdots p_d^{x_d-1} \cdots p_D^{x_D} \\
&= N p_d \sum_{x_d=1}^{N} \frac{(N-1)!}{x_1! \cdots (x_d-1)! \cdots x_D!} p_1^{x_1} \cdots p_d^{x_d-1} \cdots p_D^{x_D} \\
&= N p_d \sum_{x_d=0}^{N-1} \frac{(N-1)!}{x_1! x_2! \cdots x_D!} p_1^{x_1} p_2^{x_2} \cdots p_D^{x_D} \\
&= N p_d
\end{aligned}
$$

where we used an equality from the multinomial expansion

$$
1 = (p_1 + p_2 + \cdots + p_D)^{N-1} = \sum_{x_d=0}^{N-1} \frac{(N-1)!}{x_1! x_2! \cdots x_D!} p_1^{x_1} p_2^{x_2} \cdots p_D^{x_D}
$$

5 Poisson

A Poisson distribution is expressed as

$$
\begin{aligned}
x &\sim \mathsf{Poisson}(\mu) \\
P(x) &= \frac{\mu^x e^{-\mu}}{x!}
\end{aligned}
$$

The probability function can be equivalently written as

$$
\begin{aligned}
P(x) &= \frac{\mu^x e^{-\mu}}{x!} \\
&= \frac{1}{x!} e^{-\mu} e^{\ln \mu^x} \\
&= \frac{1}{x!} e^{-\mu} e^{x \ln \mu}
\end{aligned}
$$

Hence, the base measure f, the normalizer g, the natural parameters ϕ, the sufficient statistic T are

$$
\begin{aligned}
f(x) &= \frac{1}{x!} \\
g(\mu) &= e^{-\mu} \\
\phi(\mu) &= \ln \mu \\
\mathsf{T}(x) &= x
\end{aligned}
$$

The expectation of the sufficient statistics can be computed as

$$
\langle x \rangle = -\frac{d \ln g(\mu)}{d\mu} \frac{d\mu}{d\phi(\mu)} = -\frac{d \ln e^{-\mu}}{d\mu} \frac{d\mu}{d \ln \mu} = \mu
$$

6 Beta

A Beta distribution is expressed as

$$
\begin{aligned}
x &\sim \mathsf{Beta}(\alpha, \beta) \\
P(x) &= \frac{1}{B(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1}
\end{aligned}
$$

The probability function can be equivalently written as

$$\begin{aligned}
P(x) &= \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1} \\
&= \frac{1}{B(\alpha,\beta)} e^{(\alpha-1)\ln x} e^{(\beta-1)\ln(1-x)} \\
&= \frac{1}{B(\alpha,\beta)} e^{(\alpha-1)\ln x + (\beta-1)\ln(1-x)}
\end{aligned}$$

Hence, the base measure f, the normalizer g, the natural parameters ϕ, the sufficient statistic T are

$$\begin{aligned}
f(x) &= 1 \\
g(\alpha,\beta) &= \frac{1}{B(\alpha,\beta)} \\
\phi(\alpha,\beta) &= \begin{bmatrix} \alpha-1 \\ \beta-1 \end{bmatrix} \\
\mathsf{T}(x) &= \begin{bmatrix} \ln x \\ \ln(1-x) \end{bmatrix}
\end{aligned}$$

The expectation of the sufficient statistic is

$$\langle \mathsf{T}(x) \rangle = \begin{bmatrix} \langle \ln x \rangle \\ \langle \ln(1-x) \rangle \end{bmatrix} = -\frac{d \ln g(\alpha,\beta)}{d\phi(\alpha,\beta)} = \frac{d \ln B(\alpha,\beta)}{d \begin{bmatrix} \alpha-1 \\ \beta-1 \end{bmatrix}} = \begin{bmatrix} \frac{d \ln B(\alpha,\beta)}{d\alpha} \\ \frac{d \ln B(\alpha,\beta)}{d\beta} \end{bmatrix}$$

We can write the Beta function in terms of Gamma functions

$$B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

The expectation can thus be computed as

$$\begin{aligned}
\langle \ln x \rangle &= \frac{d \ln \Gamma(\alpha)}{d\alpha} - \frac{d \ln \Gamma(\alpha+\beta)}{d\alpha} = \psi(\alpha) - \psi(\alpha+\beta) \\
\langle \ln(1-x) \rangle &= \frac{d \ln \Gamma(\beta)}{d\alpha} - \frac{d \ln \Gamma(\alpha+\beta)}{d\alpha} = \psi(\beta) - \psi(\alpha+\beta)
\end{aligned}$$

where the digamma function is defined as the logarithmic derivative of the gamma function $\psi(z) = \frac{d}{dz} \ln \Gamma(z)$.

7 Gamma

A Gamma distribution is expressed as

$$\begin{aligned}
x &\sim \mathsf{Gamma}(\alpha,\beta) \\
P(x) &= \frac{\beta^\alpha}{\Gamma(\alpha)} x^{(\alpha-1)} e^{-\beta x}
\end{aligned}$$

The probability function can be equivalently written as

$$\begin{aligned}
P(x) &= \frac{\beta^\alpha}{\Gamma(\alpha)} x^{(\alpha-1)} e^{-\beta x} \\
&= \frac{\beta^\alpha}{\Gamma(\alpha)} e^{(\alpha-1)\ln x} e^{-\beta x} \\
&= \frac{\beta^\alpha}{\Gamma(\alpha)} e^{(\alpha-1)\ln x - \beta x}
\end{aligned}$$

Hence, the base measure f, the normalizer g, the natural parameters ϕ, the sufficient statistic T are

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

The expectation of the sufficient statistics can be computed as

$$\langle \mathsf{T}(x)\rangle = \begin{bmatrix}\langle \ln x\rangle \\ \langle x\rangle\end{bmatrix} = -\frac{d\ln g(\alpha,\beta)}{d\phi(\alpha,\beta)} = \frac{d\left[-\alpha\ln\beta + \ln\Gamma(\alpha)\right]}{d\begin{bmatrix}\alpha-1 \\ -\beta\end{bmatrix}} = \begin{bmatrix}-\ln\beta + \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} \\ \frac{\alpha}{\beta}\end{bmatrix} = \begin{bmatrix}-\ln\beta + \psi(\alpha) \\ \frac{\alpha}{\beta}\end{bmatrix}$$

8 Dirichlet

A Dirichlet distribution is expressed as

$$\begin{aligned}
\mathbf{x} &\sim \mathsf{Dirichlet}(\boldsymbol{\alpha}) \\
P(x) &= \frac{\Gamma\left(\sum_{d=1}^{D}\alpha_d\right)}{\prod_{d=1}^{D}\Gamma(\alpha_d)}\prod_{d=1}^{D}x_d^{\alpha_d-1}
\end{aligned}$$

The probability function can be equivalently written as

$$\begin{aligned}
P(x) &= \frac{\Gamma\left(\sum_{d=1}^{D}\alpha_d\right)}{\prod_{d=1}^{D}\Gamma(\alpha_d)}\prod_{d=1}^{D}x_d^{\alpha_d-1} \\
&= \frac{\Gamma\left(\sum_{d=1}^{D}\alpha_d\right)}{\prod_{d=1}^{D}\Gamma(\alpha_d)}\exp\left[\sum_{d=1}^{D}(\alpha_d-1)\ln x_d\right]
\end{aligned}$$

Hence, the base measure f, the normalizer g, the natural parameters ϕ, the sufficient statistic T are

$$\begin{aligned}
f(\mathbf{x}) &= 1 \\
g(\boldsymbol{\alpha}) &= \frac{\Gamma\left(\sum_{d=1}^{D}\alpha_d\right)}{\prod_{d=1}^{D}\Gamma(\alpha_d)} \\
\phi(\boldsymbol{\alpha}) &= \begin{bmatrix}\alpha_1-1 \\ \alpha_2-1 \\ \vdots \\ \alpha_D-1\end{bmatrix} \\
\mathsf{T}(\mathbf{x}) &= \begin{bmatrix}\ln x_1 \\ \ln x_2 \\ \vdots \\ \ln x_D\end{bmatrix}
\end{aligned}$$

The expectation of the sufficient statistics can be computed as

$$\langle \mathsf{T}(x)\rangle = \begin{bmatrix}\langle \ln x_1\rangle \\ \langle \ln x_2\rangle \\ \vdots \\ \langle \ln x_D\rangle\end{bmatrix} = -\frac{d\ln g(\boldsymbol{\alpha})}{d\phi(\boldsymbol{\alpha})}$$

Equivalently,

$$\begin{aligned}
\langle \ln x_d\rangle &= -\frac{d\ln g(\boldsymbol{\alpha})}{d\alpha_d} \\
&= \frac{d}{d\alpha_d}\left[-\ln\Gamma\left(\sum_{d=1}^{D}\alpha_d\right) + \sum_{d=1}^{D}\ln\Gamma(\alpha_d)\right] \\
&= \frac{d}{d\alpha_d}\left[-\ln\Gamma\left(\sum_{d=1}^{D}\alpha_d\right) + \ln\Gamma(\alpha_d)\right] \\
&= -\psi\left(\sum_{d=1}^{D}\alpha_d\right) + \psi(\alpha_d)
\end{aligned}$$