

Lecture Notes 7: Exponential Family

Professor: Zhihua Zhang

Definition 7.1 (Statistic). Given random variables(vectors) X_1, \dots, X_n with respect to sets of possible values $\mathcal{X}_1, \mathcal{X}_2, \dots, \mathcal{X}_n$, respectively. A random vector $\mathbf{t}_n : \mathcal{X}_1 \times \mathcal{X}_2 \times \dots \times \mathcal{X}_n \rightarrow \mathbb{R}^{k(n)}$ is called a $k(n)$ dimensional statistic.

Example 7.1. $t_n(X_1, \dots, X_n) = (X_1, \dots, X_n)$.

Example 7.1 is the simplest statistic, usually we want to achieve data reduction by statistic, i.e., $k(n) < n$. Sometimes $k(n)$ are independent with n .

Example 7.2.

$\mathbf{t}_n = \frac{1}{n}(X_1 + \dots + X_n), k(n) = 1$

$\mathbf{t}_n = [n, (X_1 + \dots + X_n), (X_1^2 + \dots + X_n^2)], k(n) = 3$, the zero order, first and second moment.

$\mathbf{t}_n = [n, \text{median}(X_1, \dots, X_n)]$, the median.

$\mathbf{t}_n = \max\{X_1, \dots, X_n\} - \min\{X_1, \dots, X_n\}$, the range.

Definition 7.2 (Sufficient Statistic). The sequence $\mathbf{t}_1, \mathbf{t}_2, \dots, \mathbf{t}_n$ is a sufficient statistic for X_1, X_2, \dots, X_n if for $n \geq 1$, the joint density for X_1, X_2, \dots, X_n given θ has the form

$$p(x_1, x_2, \dots, x_n | \theta) = h_n(\mathbf{t}_n, \theta) g(x_1, x_2, \dots, x_n)$$

for some function $h_n \geq 0, g > 0$.

Theorem 7.1. The sequence $\mathbf{t}_1, \mathbf{t}_2, \dots, \mathbf{t}_n$ is sufficient for infinitely exchangeable X_1, X_2, \dots if and only if for any $n \geq 1$, the density $p(x_1, x_2, \dots, x_n | \theta, \mathbf{t}_n)$ is independent of θ .

Proof. For any $\mathbf{t}_n = t_n(X_1, X_2, \dots, X_n)$,

$$p(x_1, x_2, \dots, x_n | \theta) = p(x_1, x_2, \dots, x_n | \theta, \mathbf{t}_n) p(\mathbf{t}_n, \theta)$$

If $p(x_1, x_2, \dots, x_n | \theta, \mathbf{t}_n)$ is independent of θ , then $p(x_1, x_2, \dots, x_n | \theta, \mathbf{t}_n)$ is g , $p(\mathbf{t}_n, \theta)$ is h_n . So \mathbf{t}_n is a sufficient statistic.

If \mathbf{t}_n is sufficient, then $p(x_1, x_2, \dots, x_n | \theta) = h_n(\mathbf{t}_n, \theta) g(x_1, x_2, \dots, x_n), h_n \geq 0, g > 0$. Taking integral on both sides, we have

$$\int_{\{\mathbf{t}_n(x_1, \dots, x_n) = \mathbf{t}_n\}} p(x_1, \dots, x_n | \theta) dx_1 \dots dx_n = \int_{\{\mathbf{t}_n(x_1, \dots, x_n) = \mathbf{t}_n\}} h_n(\mathbf{t}_n, \theta) g(x_1, \dots, x_n) dx_1 \dots dx_n$$

Note that $h_n(\mathbf{t}_n, \theta)$ in the right side is unrelated to the integral, $\int g(\mathbf{x}) d\mathbf{x}$ can be seemed as a function of \mathbf{t}_n , denoted by $G(\mathbf{t}_n)$ and $\int p(\mathbf{x} | \theta) d\mathbf{x}$ can be seemed as $p(\mathbf{t}_n | \theta)$. Hence, we have

$$\begin{aligned} p(\mathbf{t}_n | \theta) &= h_n(\mathbf{t}_n, \theta) G(\mathbf{t}_n) \\ \implies h_n(\mathbf{t}_n, \theta) &= \frac{p(\mathbf{t}_n | \theta)}{G(\mathbf{t}_n)} \end{aligned}$$

So,

$$\begin{aligned} p(x_1, x_2, \dots, x_n | \theta) &= \frac{p(\mathbf{t}_n | \theta)}{G(\mathbf{t}_n)} g(x_1, x_2, \dots, x_n) \\ \implies p(x_1, x_2, \dots, x_n | \theta, \mathbf{t}_n) &= \frac{p(x_1, x_2, \dots, x_n | \theta)}{p(\mathbf{t}_n | \theta)} = \frac{g(x_1, x_2, \dots, x_n)}{G(\mathbf{t}_n)} \end{aligned}$$

Thus we can see $p(x_1, x_2, \dots, x_n | \theta, \mathbf{t}_n)$ is independent with θ . □

Example 7.3 (Bernoulli Distribution). For a Bernoulli sequence X_1, \dots, X_n ,

$$\begin{aligned} p(x_1, \dots, x_n) &= \int_0^1 p(x_1, \dots, x_n | \theta) dF(\theta) \\ &= \int_0^1 \prod_{i=1}^n B(x_i | \theta) dF(\theta) \\ &= \int_0^1 \theta^{S_n} (1 - \theta)^{n - S_n} dF(\theta) \end{aligned}$$

where $S_n = x_1 + \dots + x_n$. So,

$$p(x_1, \dots, x_n | \theta) = \theta^{S_n} (1 - \theta)^{n - S_n}$$

Let $\mathbf{t}_n = [n, S_n]$, $p(x_1, \dots, x_n | \theta)$ can be factorized into $h_n = \theta^{S_n} (1 - \theta)^{n - S_n}$ and $g = 1$. So \mathbf{t}_n is the sufficient statistic of Bernoulli distribution.

Example 7.4 (Normal Distribution).

$$\begin{aligned} p(x_1, \dots, x_n | \mu, \lambda) &= \prod_{i=1}^n \left(\frac{\lambda}{2\pi} \right)^{\frac{1}{2}} \exp\left(-\frac{\lambda}{2} (x_i - \mu)^2\right) \\ &= \left(\frac{\lambda}{2\pi} \right)^{\frac{n}{2}} \exp\left(-\frac{\lambda}{2} \sum_{i=1}^n (x_i - \mu)^2\right) \\ &= \left(\frac{\lambda}{2\pi} \right)^{\frac{n}{2}} \exp\left(-\frac{\lambda}{2} [n(\bar{x} - \mu) + nS_n^2]\right) \end{aligned}$$

where $\bar{x}_n = \frac{1}{n} \sum_{i=1}^n x_i$, $S_n^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}_n)^2$. So the sufficient statistic of normal distribution can be $[n, \bar{X}_n, S_n^2]$. Note that the sufficient statistic is not unique, for example, $[n, \bar{X}_n, \frac{1}{n} \sum_{i=1}^n X_i^2]$ is also sufficient statistic of normal distribution.

7.1 Exponential Family

Definition 7.3 (one-parameter exponential family). A p.d.f or p.m.f $p(x | \theta)$, labelled by $\theta \in \Theta \subseteq \mathbb{R}$ is said to belong to one-parameter exponential family if it is of the form

$$p(x | \theta) = f(x) g(\theta) \exp(c \cdot \phi(\theta) h(x))$$

where $g^{-1}(\theta) = \int f(x) \exp(c \cdot \phi(\theta) h(x)) dx < \infty$ is a regularization factor. Denoted by $E_f(f, g, h, \phi, c, \theta)$.

Definition 7.4. The family is called regular if $\mathcal{X}, (X \in \mathcal{X})$ does not depend on θ , otherwise is called non-regular.

Proposition 7.1 (Sufficient statistic for E_f). If $X_1, \dots, X_n \in \mathcal{X}$ is an exchangeable sequence such that given regular $E_f(X|f, g, h, \phi, c, \theta)$,

$$p(x_1, \dots, x_n) = \int_{\theta} \prod_{i=1}^n E_f(x_i|f, g, h, \phi, c) dF(\theta)$$

for some $dF(\theta)$. Then $\mathbf{t}_n = t_n(X_1, \dots, X_n) = [n, h(X_1) + \dots + h(X_n)]$ is sufficient statistic.

Example 7.5 (Bernoulli Distribution).

$$\begin{aligned} p(x|\theta) &= \theta^x (1 - \theta)^{1-x}, \quad x \in \{0, 1\}, \theta \in [0, 1] \\ &= (1 - \theta) \left(\frac{\theta}{1 - \theta} \right)^x \\ &= (1 - \theta) \exp(x \ln \frac{\theta}{1 - \theta}) \end{aligned}$$

So, $f(x) = 1, g(\theta) = 1 - \theta, c = 1, h(x) = x, \phi(\theta) = \ln \frac{\theta}{1 - \theta}$.

Example 7.6 (Poisson Distribution).

$$\begin{aligned} p(x|\theta) &= \frac{\theta^x \cdot e^{-\theta}}{x!} \\ &= \frac{1}{x!} \exp(-\theta) \exp(x \ln \theta) \end{aligned}$$

So, $f(x) = \frac{1}{x!}, g(\theta) = e^{-\theta}, c = 1, h(x) = x, \phi(\theta) = \ln \theta$.

Example 7.7 (Normal Distribution with Unknown Variance).

$$\begin{aligned} p(x|\theta) &= N(x|0, \sigma^2) \\ &= \left(\frac{1}{2\pi\sigma^2} \right)^{\frac{1}{2}} \exp\left(-\frac{x^2}{2\sigma^2}\right) \\ &= \left(\frac{1}{2\pi} \right)^{\frac{1}{2}} \theta^{-\frac{1}{2}} \exp\left(-\frac{x^2}{2\theta}\right) \end{aligned}$$

So, $f(x) = \left(\frac{1}{2\pi}\right)^{\frac{1}{2}}, g(\theta) = \theta^{-\frac{1}{2}}, c = -\frac{1}{2}, h(x) = x^2, \phi(\theta) = \theta^{-1}$.

Example 7.8 (Uniform Distribution, non-regular).

$$p(x|\theta) = U(x|[0, \theta]) = \frac{1}{\theta}$$

So, $f(x) = 1, g(\theta) = \theta^{-1}, c = 1, h(x) = 0, \phi(\theta) = 0$. Since \mathcal{X} is $[0, \theta]$, related to θ , so it is non regular.

$$\begin{aligned} f_X(x_1, \dots, x_n) &= \frac{1}{\theta} \mathbf{1}_{\{0 \leq x_1 \leq \theta\}} \cdots \frac{1}{\theta} \mathbf{1}_{\{0 \leq x_n \leq \theta\}} \\ &= \frac{1}{\theta^n} \mathbf{1}_{\{0 \leq \min\{x_i\}\}} \mathbf{1}_{\{\max\{x_i\} \leq \theta\}} \end{aligned}$$

where $\mathbf{1}\{\dots\}$ is the indicator function. So the sufficient statistic $\mathbf{t}_n = [n, \max\{x_i\}]$.

Definition 7.5 (k-parameters exponential family). A p.d.f or p.m.f $p(x|\theta)$, $x \in \mathcal{X}$, which is labelled by $\theta \in \Theta \subseteq \mathbb{R}$ is said to belong to k-parameters exponential family if it is of the form

$$p(x|\theta) = f(x)g(\theta) \exp \left(\sum_{j=1}^k c_j \cdot \phi_j(\theta) h_j(x) \right)$$

Denoted by $E_{f_k}(x|f, g, h, \phi, c, \theta)$.

Proposition 7.2 (Sufficient statistic for E_{f_k}). If $X_1, \dots, X_n \in \mathcal{X}$ is an exchangeable sequence such that given regular $E_{f_k}(X|f, g, h, \phi, c, \theta)$,

$$p(x_1, \dots, x_n|\theta) = \prod_{i=1}^n E_{f_k}(x_i|f, g, h, \phi, c, \theta)$$

Then $\mathbf{t}_n = t_n(X_1, \dots, X_n) = [n, \sum_{i=1}^n h_1(X_i), \dots, \sum_{i=1}^n h_k(X_i)]$ is sufficient statistic of X_1, \dots, X_n .

Example 7.9 (Normal Distribution with Unknown Mean and Variance). Let $\theta = [\mu, \lambda]$,

$$\begin{aligned} p(x|\theta) &= N(x|\mu, \lambda) \\ &= \left(\frac{\lambda}{2\pi}\right)^{\frac{1}{2}} \exp\left(-\frac{\lambda}{2}(x - \mu)^2\right) \\ &= \left(\frac{1}{2\pi}\right)^{\frac{1}{2}} \lambda^{\frac{1}{2}} \exp\left(-\frac{\lambda}{2}\mu^2\right) \exp(\lambda\mu x - \frac{1}{2}\lambda x^2) \end{aligned}$$

So, $g(\theta) = \lambda^{\frac{1}{2}} \exp(-\frac{\lambda}{2}\mu^2)$, $c_1 = 1, c_2 = -\frac{1}{2}$, $\phi_1(\theta) = \lambda\mu, \phi_2(\theta) = \lambda$, $h_1(x) = x, h_2(x) = x^2$.

7.2 Canonical(Natural) Exponential Family

The p.d.f of exponential family can be rewritten into another form:

$$p(\mathbf{y}|\varphi) = cef(\mathbf{y}|a, b, \varphi) = a(\mathbf{y}) \exp(\mathbf{y}^T \varphi - b(\varphi))$$

where $\mathbf{y} = (y_1, \dots, y_k), \phi = (\varphi_1, \dots, \varphi_k)$. Comparing to the previous form, we can see $y_i = h_i(x), \varphi_i = c_i \phi(\theta)$.

Proposition 7.3 (moments of cef). For \mathbf{y} in definition of cef, we have

$$E[\mathbf{y}|\varphi] = \int \mathbf{y} a(\mathbf{y}) \exp(\mathbf{y}^T \varphi - b(\varphi)) d\mathbf{y}$$

Since $\int a(\mathbf{y}) \exp(\mathbf{y}^T \varphi - b(\varphi)) d\mathbf{y} = 1$, taking derivation on both sides.

$$\begin{aligned} & \int a(\mathbf{y}) \exp(\mathbf{y}^T \varphi - b(\varphi)) (\mathbf{y} - \nabla_{\varphi} b(\varphi)) d\mathbf{y} = 0 \\ \implies & \int a(\mathbf{y}) \exp(\mathbf{y}^T \varphi - b(\varphi)) \mathbf{y} d\mathbf{y} = \int a(\mathbf{y}) \exp(\mathbf{y}^T \varphi - b(\varphi)) \nabla_{\varphi} b(\varphi) d\mathbf{y} \\ \implies & E[\mathbf{y}] = \nabla_{\varphi} b(\varphi) \end{aligned}$$

Example 7.10 (Poisson Distribution).

$$e^{-\lambda} \frac{\lambda^x}{x!} = \frac{1}{x!} \exp(x \log \lambda - \lambda) = \frac{1}{x!} \exp(x\theta - e^{\theta}), \lambda = e^{\theta}$$

So, $E[\mathbf{y}] = \nabla_{\varphi} b(\varphi) = \lambda$

Theorem 7.2. If $X = (X_1, \dots, X_n)$ is random variable from a regular exponential family distribution such that

$$p(\mathbf{x}|\theta) = \prod_{i=1}^n f(x_i) [g(\theta)]^n \exp \left(\sum_{j=1}^k c_j \phi_j(\theta) \sum_{i=1}^n h_j(x_i) \right).$$

Then the conjugate family for θ has the form

$$p(\mathbf{x}|\tau) = [K(\tau)]^{-1} [g(\theta)]^{\tau_0} \exp \left(\sum_{j=1}^k c_j \phi_j(\theta) \tau_j \right)$$

where $k(\tau) = \int_{\theta} [g(\theta)]^{\tau} \exp \left(\sum_{j=1}^k c_j \phi_j(\theta) \tau_j \right) d\theta < \infty$

Example 7.11 (Bernoulli Likelihood).

$$\begin{aligned} p(\mathbf{x}|\theta) &= \prod_{i=1}^n \theta^{x_i} (1 - \theta)^{(1-x_i)} \\ &= (1 - \theta)^n \exp \left(\left(\log \frac{\theta}{1 - \theta} \right) \sum_{i=1}^n x_i \right) \end{aligned}$$

So,

$$\begin{aligned} p(\theta|\tau) &\propto (1 - \theta)^{\tau_0} \exp \left(\log \frac{\theta}{1 - \theta} \tau_1 \right) \\ &\propto (1 - \theta)^{\tau_0} \left(\frac{\theta}{1 - \theta} \right)^{\tau_1} \\ &\propto \theta^{\tau_1} (1 - \theta)^{\tau_0 - \tau_1} \end{aligned}$$

Hence, the conjugate prior of Bernoulli distribution is beta distribution.

Example 7.12 (Poisson Likelihood).

$$\begin{aligned} p(\mathbf{x}|\theta) &= \prod_{i=1}^n \frac{\theta^{x_i} \exp(-\theta)}{x_i!} \\ &= \prod_{i=1}^n (x_i!)^{-1} \exp(-n\theta) \exp(\log \theta \sum_{i=1}^n x_i) \end{aligned}$$

So,

$$\begin{aligned} p(\theta|\tau) &\propto \exp(-\tau_0\theta) \exp(\tau_1 \log \theta) \\ &\propto \theta^{\tau_1} \exp(-\tau_0\theta) \end{aligned}$$

Hence, the conjugate prior of Poisson distribution is gamma distribution.

Example 7.13 (Normal Likelihood). Let $\theta = (\mu, \lambda)$

$$\begin{aligned} p(\mathbf{x}|\theta) &= \prod_{i=1}^n \left(\frac{\lambda}{2\pi} \right)^{\frac{1}{2}} \exp \left(-\frac{\lambda}{2} (x_i - \mu)^2 \right) \\ &= (2\pi)^{-\frac{n}{2}} \left[\lambda^{\frac{1}{2}} \exp(-\frac{\lambda}{2} \mu^2) \right]^n \exp(\mu \lambda \sum_{i=1}^n x_i - \frac{\lambda}{2} \sum_{i=1}^n x_i^2) \end{aligned}$$

So,

$$\begin{aligned} p(\theta|\tau) &\propto \left[\lambda^{\frac{1}{2}} \exp(-\frac{\lambda}{2} \mu^2) \right]^{\tau_0} \exp(\mu \lambda \tau_1 - \frac{\lambda}{2} \tau_2) \\ &\propto \lambda^{\frac{\tau_0}{2}} \exp(-\frac{\tau_0 \lambda}{2} \mu^2) \exp(\mu \lambda \tau_1 - \frac{\lambda}{2} \tau_2) \\ &\propto \lambda^{\frac{\tau_0}{2}} \exp \left(-\frac{\tau_0 \lambda}{2} \left(\mu - \frac{\tau_1}{\tau_0} \right)^2 \right) \exp \left(\frac{\lambda \tau_1^2}{2\tau_0} \right) \exp \left(-\frac{\lambda \tau_2}{2} \right) \\ &\propto (\tau_0 \lambda)^{\frac{1}{2}} \exp \left(-\frac{\tau_0 \lambda}{2} \left(\mu - \frac{\tau_1}{\tau_0} \right)^2 \right) \exp \left(\frac{\lambda \tau_1^2}{2\tau_0} \right) \exp \left(-\frac{\lambda \tau_2}{2} \right) \lambda^{\frac{\tau_0}{2}} (\tau_0 \lambda)^{-\frac{1}{2}} \\ &\propto (\tau_0 \lambda)^{\frac{1}{2}} \exp \left(-\frac{\tau_0 \lambda}{2} \left(\mu - \frac{\tau_1}{\tau_0} \right)^2 \right) \exp \left(-\frac{\lambda}{2} \left(\tau_2 - \frac{\tau_1^2}{\tau_0} \right) \right) \tau_0^{-\frac{1}{2}} \lambda^{\frac{\tau_0+1}{2}-1} \end{aligned}$$

Note that $(\tau_0 \lambda)^{\frac{1}{2}} \exp \left(-\frac{\tau_0 \lambda}{2} \left(\mu - \frac{\tau_1}{\tau_0} \right)^2 \right)$ can be seemed as a normal prior, $p(\mu|\lambda, \tau)$ and $\exp \left(-\frac{\lambda}{2} \left(\tau_2 - \frac{\tau_1^2}{\tau_0} \right) \right) \tau_0^{-\frac{1}{2}} \lambda^{\frac{\tau_0+1}{2}-1}$ can be seemed as a gamma prior, $p(\lambda|\tau)$.

Theorem 7.3 (posterior). $p(\theta|x, \tau) = p(\theta|\tau + \mathbf{t}_n(x))$, where $\tau + \mathbf{t}_n(x) = (\tau_0 + n, \tau_1 + \sum_{i=1}^n h_1(x_i), \dots, \tau_k + \sum_{i=1}^n h_k(x_i))$.