Useful theorems and concepts

Saki Kuzushima

2019-06-27

Contents

Sum of exponential random variables is a gamma random variable

Theorem: If $X_i \overset{i.i.d}{\sim} Expo(\lambda)$, $i = 1...n$, then $Y = \sum_{i=1}^n X_i \sim Gamma(n, \lambda)$.

Proof:

First, recall that if $X \sim Expo(\lambda)$,

$$M_X(t) = \frac{\lambda}{\lambda - t}, \quad (t < \lambda)$$

if $Z \sim Gamma(\alpha, \beta)$,

$$M_Z(t) = (\frac{1}{1 - \beta t})^{-\alpha}$$

Observe that

$$\begin{aligned}
M_Y(t) &= \mathbb{E}[e^{tY}] \\
&= \mathbb{E}[e^{t\sum_{i=1}^n X_i}] \\
&= \mathbb{E}[\prod_{i=1}^n e^{tX_i}] \\
&= \prod_{i=1}^n \mathbb{E}[e^{tX_i}] \\
&= \prod_{i=1}^n \frac{\lambda}{\lambda - t} \\
&= (\frac{\lambda}{\lambda - t})^n \\
&= (\frac{\lambda - t}{\lambda})^{-n} \\
&= (1 - \frac{t}{\lambda})^{-n}
\end{aligned}$$

This is an mgf of Gamma distribution. So, $Y \sim Gamma(n, \lambda)$.

Linear combination of Gaussian random variables is also Gaussian

Theorem: if $X \sim N(\mu, \sigma^2)$, and $Y = aX + b$, $(a \neq 0)$, then $Y \sim N(a\mu + b, a^2\sigma^2)$.

Proof: Recall that if $X \sim N(\mu, \sigma^2)$

$$M_X(t) = e^{\mu t + \frac{\sigma^2}{2} t^2}$$

Then,

$$\begin{aligned}
M_Y(t) &= \mathbb{E}[e^{t(aX+b)}] \\
&= \mathbb{E}[e^{taX+tb}] \\
&= e^{tb}\mathbb{E}[e^{taX}] \\
&= e^{tb}M_X(ta) \\
&= e^{tb}e^{\mu t + \frac{\sigma^2}{2} t^2} \\
&= exp(tb + t\mu a + \frac{\sigma^2}{2} t^2) \\
&= exp((a\mu + b)t + \frac{a^2\sigma^2}{2} t^2)
\end{aligned}$$

This is also an mgf of Gaussian. $Y \sim N(a\mu + b, a^2\sigma^2)$.

Distribution of maximum values

Theorem: If $X_1 ... X_n \overset{i.i.d}{\sim} f_X(x)$, and $Y \equiv max\{X_1 ... X_n\}$, then $f_Y(y) = nF_X^{n-1}(y) f_X(y)$.

Proof:

$$\begin{aligned}
F_Y(y) &= P(Y \leq y) \\
&= P(max\{X_1 ... X_n\} \leq y) \\
&= P(X_1 \leq y) ... P(X_n \leq y) \\
&= F_X(y) ... F_X(y) \\
&= F_X^n(y)
\end{aligned}$$

Then, because $f_Y(y) = \frac{\partial}{\partial y} F_Y(y)$,

$$f_Y(y) = nF_X^{n-1}(y) f_X(y)$$

The Dirichlet distribution belongs to the exponential family

The fact that the Dirichlet distribution belongs to the exponential family is useful. While the expression of the Dirichlet in terms of the exponential family is widely available, the derivation is not. The following sketches the conversion from the usual expression of the pdf of Dirichlet distribution to the form in terms of the exponential family.

The pdf of the Dirichlet distribution is

$$p(\theta|\alpha) = \frac{1}{B(\alpha)} \prod_{i=1}^{K} \theta_i^{\alpha_i - 1} \tag{1}$$

where $\theta = [\theta_1....\theta_K]^\intercal$, and $\sum_{i=1}^K \theta_i = 1$, $\theta_i \geq 0, \forall i \in \{1, 2, ..., K\}$.

The exponential family has the following general pdf

$$p(\theta|\eta) = h(\theta)\exp(\eta^\intercal t(\theta) - A(\eta)) \tag{2}$$

where $t(\theta)$ is the sufficient statistic, η is called the natural parameter, $A(\eta)$ is the log normalization factor, and $h(x)$ is the base measure.

It is known that the Dirichlet distribution belongs to the exponential family, and the parameters are

- The natural parameter: $\eta = \alpha = [\alpha_1...\alpha_K]^\intercal$
- The sufficient statistic: $t(\theta) = \log\theta = \log[\theta_1, ...\theta_K]^\intercal$.
- The log normalization factor: $A(\eta) = \sum_{i=1}^K \log\Gamma(\eta_i) - \log\Gamma(\sum_{i=1}^K \eta_i)$
- The base measure: $h(x) = \frac{1}{\prod_{i=1}^K}\theta_i$

The following sketches why the parameters are expressed in these ways.

First, recall that $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$). Then the pdf of the Dirichlet distribution is

$$\begin{aligned}
p(\theta|\alpha) &= \frac{1}{B(\alpha)}\prod_{i=1}^K \theta_i^{\alpha_i - 1} \\
&= \frac{\Gamma(\sum_{i=1}^K \alpha_i)}{\prod_{i=1}^K \Gamma(\alpha_i)} \cdot \frac{\prod_{i=1}^K \theta_i^{\alpha_i}}{\prod_{i=1}^K \theta_i}
\end{aligned}$$

Taking the exponential and log,

$$\begin{aligned}
p(\theta|\alpha) &= \frac{\Gamma(\sum_{i=1}^K \alpha_i)}{\prod_{i=1}^K \Gamma(\alpha_i)} \cdot \frac{\prod_{i=1}^K \theta_i^{\alpha_i}}{\prod_{i=1}^K \theta_i} \\
&= \frac{1}{\prod_{i=1}^K \theta_i}\exp\Big[\log\frac{\Gamma(\sum_{i=1}^K \alpha_i)}{\prod_{i=1}^K \Gamma(\alpha_i)}\prod_{i=1}^K \theta_i^{\alpha_i}\Big] \\
&= \frac{1}{\prod_{i=1}^K \theta_i}\exp\Big[\log\Gamma(\sum_{i=1}^K \alpha_i) - \log\prod_{i=1}^K \Gamma(\alpha_i) + \log\prod_{i=1}^K \theta_i^{\alpha_i}\Big] \\
&= \frac{1}{\prod_{i=1}^K \theta_i}\exp\Big[\log\Gamma(\sum_{i=1}^K \alpha_i) - \sum_{i=1}^K \log\Gamma(\alpha_i) + \sum_{i=1}^K \alpha_i\log\theta_i\Big] \\
&= \frac{1}{\prod_{i=1}^K \theta_i}\exp\Big[\alpha^\intercal\log\theta - \big\{\sum_{i=1}^K \log\Gamma(\alpha_i) - \log\Gamma(\sum_{i=1}^K \alpha_i)\big\}\Big]
\end{aligned}$$

Comparison of this expression with the general pdf of the exponential family yields the parameters listed above.

The expectation of the sufficient statistic of the exponential family is the derivative of the log normalizing factor

Note that this is useful in the derivation of the original variational inference for LDA (in Blei et al 2003).

Theorem:

$$\mathbb{E}_\theta[t(\theta)] = \frac{\partial}{\partial \eta} A(\eta) \tag{3}$$

Proof:

Using the fact that the pdf must integrate to one, we first express $a(\eta)$ by the other parameters.

$$\begin{aligned}
1 &= \int_\theta p(\theta|\eta) d\theta = \int_\theta h(\theta) \exp(\eta\ t(\theta) - A(\eta)) d\theta \\
\iff 1 &= \frac{1}{\exp A(\eta)} \int_\theta h(\theta) \exp(\eta\ t(\theta)) d\theta \\
\iff \exp A(\eta) &= \int_\theta h(\theta) \exp(\eta\ t(\theta)) d\theta \\
\iff A(\eta) &= \log \int_\theta h(\theta) \exp(\eta\ t(\theta)) d\theta
\end{aligned}$$

Let $g(\eta) = \frac{1}{\exp A(\eta)}$. Then this is also equivalent to

$$1 = g(\eta) \int_\theta h(\theta) \exp(\eta\ t(\theta)) d\theta \tag{4}$$

Taking the derivative of Eq (3) with respect to η,

$$\begin{aligned}
0 &= \frac{\partial}{\partial \eta} \Big[g(\eta) \int_\theta h(\theta) \exp(\eta\ t(\theta)) d\theta \Big] \\
&= g'(\eta) \int_\theta h(\theta) \exp(\eta\ t(\theta)) d\theta + g(\eta) \frac{\partial}{\partial \eta} \int_\theta h(\theta) \exp(\eta\ t(\theta)) d\theta \\
&= \frac{g'(\eta)}{g(\eta)} + g(\eta) \int_\theta h(\theta) \exp(\eta\ t(\theta)) t(\theta) d\theta \\
&= \frac{g'(\eta)}{g(\eta)} + \int_\theta g(\eta) h(\theta) \exp(\eta\ t(\theta)) t(\theta) d\theta \\
&= \frac{g'(\eta)}{g(\eta)} + \mathbb{E}_\theta[t(\theta)]
\end{aligned}$$

i.e.

$$\mathbb{E}_\theta[t(\theta)] = -\frac{g'(\eta)}{g(\eta)} = -\frac{\partial}{\partial \eta} log(g(\eta)) \tag{5}$$

because $\frac{\partial}{\partial \eta} log(g(\eta)) = \frac{1}{g(\eta)} \cdot \frac{\partial}{\partial \eta} g(\eta)$.

Because $g(\eta) = \frac{1}{\exp(A\eta))} = -\exp(A(\eta))$, Eq (4) is equivalent to

$$\mathbb{E}[t(\theta)] = -\frac{\partial}{\partial \eta} log(g(\eta)) = \frac{\partial}{\partial \eta} A(\eta) \tag{6}$$

KL divergence

to be updated

Mutual Information

The mutual information measures the dependence of two random variables, and defined as follows.

Let X and Y be two (discrete) random variables, then the mutual information is

$$I(X;Y) = D_{KL}(P_{XY})||P_X \otimes P_Y) = \sum_y \sum_x P_{XY}(x,y) \log\Big(\frac{P_XY(x,y)}{P_X(x)P_Y(y)}\Big) \tag{7}$$

where D_{KL} is KL divergence. Likewise the normalized mutual information (NMI) is

$$\frac{I(X,Y)}{H(X)H(Y)} \tag{8}$$

where $H(\cdot)$ is entropy.

The relationship between beta and gamma function

Theorem:

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

where

$$B(\alpha,\beta) = \int_{x=0}^{\infty} x^{\alpha-1}(1-x)^{\beta-1}dx$$
$$\Gamma(\alpha) = \int_{x=0}^{\infty} e^{-x}x^{\alpha-1}dx$$

Proof:

$$\begin{aligned}\Gamma(\alpha)\Gamma(\beta) &= \int_{x=0}^{\infty} e^{-x}x^{\alpha-1}dx \int_{y=0}^{\infty} e^{-y}y^{\beta-1}dy \\ &= \int_{x=0}^{\infty}\int_{y=0}^{\infty} e^{-x}x^{\alpha-1}e^{-y}y^{\beta-1}\ dydx \\ &= \int_{x=0}^{\infty}\int_{y=0}^{\infty} e^{-(x+y)}x^{\alpha-1}y^{\beta-1}\ dydx\end{aligned}$$

Let $x = uw$ and $y = u(1-w)$. Then $u = x+y$ and $w = \frac{x}{x+y}$. Because $0 \le x < \infty$ and $0 \le x < \infty$. We can rewrite w

$$w = \frac{x}{x+y} = (\frac{x+y}{x})^{-1} = (1+\frac{y}{x})^{-1}$$

Because $0 \le \frac{y}{x} < \infty$, $0 \le (1+\frac{y}{x})^{-1} < 1$.

By the change of variables,

$$\Gamma(\alpha)\Gamma(\beta) = \int_{u=0}^{\infty}\int_{w=0}^{\infty} e^{-u}(uw)^{\alpha-1}(u(1-w))^{\beta-1}|J|\ dudw$$

where J is the determinant of Jaccobian.

$$J = \begin{bmatrix} w & u \\ 1-w & -u \end{bmatrix} = -u$$

Therefore,

$$
\begin{aligned}
\Gamma(\alpha)\Gamma(\beta) &= \int_{u=0}^{\infty}\int_{w=0}^{\infty} e^{-u}(uw)^{\alpha-1}(u(1-w))^{\beta-1}|J|\ dudw \\
&= \int_{u=0}^{\infty}\int_{w=0}^{\infty} e^{-u}(uw)^{\alpha-1}(u(1-w))^{\beta-1}u\ dudw \\
&= \int_{u=0}^{\infty}\int_{w=0}^{\infty} e^{-u}u^{\alpha+\beta-1}w^{\alpha-1}(1-w)^{\beta-1}\ dudw \\
&= \int_{u=0}^{\infty} e^{-u}u^{\alpha+\beta-1}\ du \int_{w=0}^{\infty} w^{\alpha-1}(1-w)^{\beta-1}\ dw \\
&= \Gamma(\alpha+\beta)B(\alpha,\beta)
\end{aligned}
$$

The expected value of a random variable with beta distribution

Theorem:

Let X be a random varible with a beta distribution.

$$X \sim Beta(\alpha,\beta)$$

Then

$$\mathbb{E}[X] = \frac{\alpha}{\alpha+\beta}$$

Proof:

$$
\begin{aligned}
\mathbb{E}[X] &= \int_{x=0}^{1} x\frac{1}{B(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1}\ dx && (9) \\
&= \frac{1}{B(\alpha,\beta)}\int_{x=0}^{1} x^{\alpha}(1-x)^{\beta-1}\ dx && (10) \\
&= \frac{B(\alpha+1,\beta)}{B(\alpha,\beta)} && (11)
\end{aligned}
$$

The last line follows because any pdf integrates to 1

$$
\begin{aligned}
1 &= \int_{x=0}^{1} \frac{1}{B(\alpha+1,\beta)}x^{\alpha}(1-x)^{\beta-1}\ dx && (12) \\
\iff B(\alpha+1,\beta) &= \int_{x=0}^{1} x^{\alpha}(1-x)^{\beta-1}\ dx && (13)
\end{aligned}
$$

Using the fact that

$$B(\alpha,\beta) = \frac{\Gamma(\alpha),\Gamma(\beta)}{\Gamma(\alpha,\beta)} \tag{14}$$

and

$$\Gamma(\alpha+1) = \alpha\Gamma(\alpha) \tag{15}$$

The equation (11) can be simplified

$$
\begin{aligned}
\frac{B(\alpha+1,\beta)}{B(\alpha,\beta)} &= \frac{\Gamma(\alpha+1)\Gamma(\beta)}{\Gamma(\alpha+\beta+1)}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} && (16)\\
&= \frac{\alpha\Gamma(\alpha)\Gamma(\beta)}{(\alpha+\beta)\Gamma(\alpha+\beta)}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} && (17)\\
&= \frac{\alpha}{\alpha+\beta} && (18)
\end{aligned}
$$