Exponential, Poisson and Gamma

Suppose on average every $1/\lambda$ hours, a Stochastic train arrives at the Random station. Further we assume the waiting time between two adjacent train follows independent $\mathsf{Ex}(\lambda)$. Next we show that

- $N(t)$ = the number of trains arrived during the time interval $[0, t]$ (from time 0 to t hours) $\sim \mathsf{Po}(\lambda t)$.
- X_k = waiting time (in hours) for the kth train $\sim \mathsf{Ga}(k, \beta = 1/\lambda)$. In particular, $X_1 \sim \mathsf{Ga}(1, \beta = 1/\lambda) = \mathsf{Ex}(\lambda)$.

Let's set our arrival time at the station as time 0, and let X_1 = waiting time for the first train afterwards. Does it make sense to model X_1 by $\mathsf{Ex}(\lambda)$, since the assumption that the waiting time between two adjacent trains follows $\mathsf{Ex}(\lambda)$ seems to only imply that $X_1 + Y \sim \mathsf{Ex}(\lambda)$, where Y denotes the time between the arrival time of the previous train and our arrival time? The answer is Yes, and it's related to the memoryless property of the Exponential distribution.

Find the distribution of $X_2 = X_1 + W$ where X_1 and W are independent and follow $\mathsf{Ex}(\lambda)$. Using the convolution formula, we have

$$\begin{aligned} f_{X_2}(x) &= \int_{\infty}^{\infty} f_{X_1}(x_1) \times f_W(x - x_1) dx_1 \\ &= \int_0^x \lambda e^{-\lambda x_1} \times \lambda e^{-\lambda(x-x_1)} dx_1 \\ &= \int_0^x \lambda^2 e^{-\lambda x} dx_1 = \lambda^2 e^{-\lambda x} \int_0^x dx_1 = \lambda^2 x e^{-\lambda x}, \end{aligned}$$

where the integration range for x_1 is $(0, x)$ because 1) $x_1 > 0$ (the original constraint on x_1) and 2) $x - x_1 > 0$ ($w = x - x_1$ and $w > 0$).

Find the distribution of $X_3 = X_2 + W$ where $f_{X_2}(x)$ is given above and W follows an independent $\mathsf{Ex}(\lambda)$. Using the convolution formula, we have

$$\begin{aligned} f_{X_3}(x) &= \int_{\infty}^{\infty} f_{X_2}(x_2) \times f_W(x - x_2) dx_2 \\ &= \int_0^x \lambda^2 x_2 e^{-\lambda x_2} \times \lambda e^{-\lambda(x-x_2)} dx_2 \\ &= \int_0^x \lambda^3 x_2 e^{-\lambda x} dx_2 = \lambda^3 e^{-\lambda x} \frac{x^2}{2}. \end{aligned}$$

We can prove by induction that the pdf for X_k is

$$f_{X_k}(x) = \frac{\lambda^k x^{k-1}}{(k-1)!} e^{-\lambda x}. \tag{1}$$

Suppose (1) holds for $k-1$ and then use the convolution formula to show that (1) also holds for $X_k = X_k + W$ where W follows an independent $\mathsf{Ex}(\lambda)$. The distribution (1) is also known as the Erlang distribution, a special case of the Gamma distribution.

Show that $N(t) \sim \mathsf{Po}(\lambda t)$, that is, for any integer $k \geq 0$,

$$\mathbb{P}\big(N(t)=k\big) = \frac{(\lambda t)^k}{k!} e^{-\lambda t}. \tag{2}$$

1. When $k=0$,
$$\mathbb{P}\big(N(t)=0\big) = \mathbb{P}(X_1 > t) = e^{-\lambda t} = \frac{(\lambda t)^0}{0!} e^{-\lambda t}.$$

2. For any $k > 0$,
$$\begin{aligned} \mathbb{P}\big(N(t)=k\big) &= \mathbb{P}(X_k \leq t \text{ AND } X_{k+1} > t) \\ &= \mathbb{P}(X_k \leq t \text{ AND } X_k + W > t). \end{aligned}$$
Using the joint pdf for (X_k, W) and the appropriate integration region, we have
$$\begin{aligned} \mathbb{P}\big(N(t)=k\big) &= \int_0^t \frac{\lambda^k x^{k-1}}{(k-1)!} e^{-\lambda x} \Big(\int_{t-x}^{\infty} \lambda e^{-\lambda w} dw \Big) dx \\ &= \int_0^t \frac{\lambda^k x^{k-1}}{(k-1)!} e^{-\lambda x} e^{-\lambda(t-x)} dx \\ &= \frac{\lambda^k}{(k-1)!} e^{-\lambda t} \int_0^t x^{k-1} dx = \frac{\lambda^k t^k}{(k)!} e^{-\lambda t}. \end{aligned}$$

The Gamma Distribution

$$\begin{aligned} X &\sim \mathsf{Ga}(\alpha, \beta = 1/\lambda) \\ \alpha > 0 &: \text{ shape parameter ; } \quad \beta > 0 : \text{ scale parameter.} \end{aligned}$$

- pdf
$$f(x) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}, \quad x > 0;$$
OR
$$f(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}, \quad x > 0.$$
where $\Gamma(\cdot)$ is called the *Gamma* function. Recall some results on Gamma functions:
$$\Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1), \quad \Gamma(n) = (n-1)!, \quad \Gamma(1) = 1, \ \Gamma(\frac{1}{2}) = \sqrt{\pi}.$$

- mean and variance
$$\mathbb{E}(X) = \alpha\beta, \quad \mathrm{Var}(X) = \alpha\beta^2.$$

- mgf
$$M(t) = \frac{1}{(1-\beta t)^{\alpha}}, \quad t < \frac{1}{\beta}.$$
- $\mathsf{Ex}(\lambda) = \mathsf{Ga}(1, 1/\lambda)$.
- $\chi^2(r) = \mathsf{Ga}(r/2, 2)$. Recall the mgf of $\chi^2(r)$ is given by $(1-2t)^{r/2}$.

  $X \sim \mathsf{Ga}(\alpha, \beta)$ where α is an integer, then $2Y/\beta \sim \chi^2(2\alpha)$. This can be proved easily by computing the mgf of $2Y/\beta$.
- Scaling
$$s\mathsf{Ga}(\alpha, \beta) = \mathsf{Ga}(\alpha, s\beta)$$
- Additivity of two *independent* Gammas that have the same scale parameter:
$$\mathsf{Ga}(\alpha_1, \beta) + \mathsf{Ga}(\alpha_2, \beta) = \mathsf{Ga}(\alpha_1 + \alpha_2, \beta).$$

  Especially,
  - $X \sim \mathsf{Ex}(\lambda)$ and $Y \sim \mathsf{Ex}(\lambda)$, then $X + Y \sim \mathsf{Ga}(2, 1/\lambda)$;
  - $X \sim \chi^2(r_1)$ and $Y \sim \chi^2(r_2)$, then $X + Y \sim \chi^2(r_1 + r_2)$.

Jensen's Inequality

g is a **convex** function if
$$tg(t_1) + (1-t)g(t_2) \geq g\big(tx_1 + (1-t)x_2\big),$$
for any $t \in [0, 1]$, and $x_1, x_2, tx_1 + (1-t)x_2$ are all in the support of g.

If g is a convex on an open interval I and X is a random variable whose support is contained in I, then
$$\mathbb{E}\big[g(X)\big] \geq g(\mathbb{E}X),$$
provided that both expectations exist. If g is strictly convex, then the equality holds only when X is a constant random variable. Recall that the equality holds if g is a linear function.

Use Jensen's inequality to show the following results.

- $\mathbb{E}(X^2) \geq (\mathbb{E}X)^2$ (of course, this result can also be obtained from $\mathrm{Var}(X) \geq 0$.)
- $\mathbb{E}e^{tX} \geq e^{t\mathbb{E}(X)}$, therefore, $M_X(t) \geq e^{t\mu}$.
- $\mathbb{E}\big(\frac{1}{X}\big) \geq \frac{1}{\mathbb{E}(X)}$ for a positive random variable X.
- $\mathbb{E}[\ln X] \leq \ln \mathbb{E}(X)$ for a positive random variable X.

(Feel free the skip the remaining material on KL divergence.) How to measure the distance between two distributions/random variables? The *Kullback-Leibler* divergence between two distributions is defined to be

$$\begin{cases} \sum_x p_1(x) \log \frac{p_1(x)}{p_2(x)}, & \text{for two discrete distributions;} \\ \int_{-\infty}^{\infty} f_1(x) \log \frac{f_1(x)}{f_2(x)} dx. & \text{for two continuous distributions,} \end{cases}$$

For example, the KL divergence between $\mathsf{Bern}(p_1)$ and $\mathsf{Bern}(p_2)$ is

$$p_1 \log \frac{p_1}{p_2} + (1-p_1) \log \frac{1-p_1}{1-p_2},$$

and the one between $\mathsf{N}(\mu_1, \sigma^2)$ and $\mathsf{N}(\mu_2, \sigma^2)$ is

$$\begin{aligned} & \int \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu_1)^2}{2\sigma^2}\right\} \log \frac{e^{-(x-\mu_1)^2/(2\sigma^2)}}{e^{-(x-\mu_2)^2/(2\sigma^2)}} dx \\ = & \int \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu_1)^2}{2\sigma^2}\right\} \left[-\frac{(x-\mu_1)^2}{2\sigma^2} + \frac{(x-\mu_2)^2}{2\sigma^2}\right] dx \\ = & \mathbb{E}_X \frac{1}{2\sigma^2}\left[(X-\mu_2)^2 - (X-\mu_1)^2\right] \quad X \sim \mathsf{N}(\mu_1, \sigma^2) \\ = & \frac{1}{2\sigma^2}\left[(\mu_1-\mu_2)^2 + \sigma^2 - \sigma^2\right] \\ = & \frac{(\mu_1-\mu_2)^2}{2\sigma^2} \end{aligned}$$

Note that the KL divergence is not necessarily symmetric, so it is not a distance metric in the strict sense, but like any other distance metric, the KL divergence is always non-negative and equals zero if and only the two arguments (two distributions) are the same, which can be proved using Jensen's inequality. For the continuous case,

$$\begin{aligned} \int_{-\infty}^{\infty} f_1(x) \log \frac{f_1(x)}{f_2(x)} dx &= \mathbb{E}_{X\sim f_1} \log \frac{f_1(X)}{f_2(X)} \\ &= \mathbb{E}_{X\sim f_1}\Big[-\log \frac{f_2(X)}{f_1(X)}\Big] \\ &\geq -\log\Big(\mathbb{E}_{X\sim f_1} \frac{f_2(X)}{f_1(X)}\Big) \\ &= -\log\Big(\int f_1(x) \frac{f_2(x)}{f_1(x)} dx\Big) \\ &= -\log\Big(\int f_2(x) dx\Big) = -\log 1 = 0, \end{aligned}$$

where $\mathbb{E}_{X\sim f_1}$ means averaging out X with respect to the pdf $f_1(x)$ and we have used the Jensen's inequality on $g(x) = -\log(x)$.

Example 1.10.4 (on p.71): Let $a_1, \ldots, a_n$ be a set of positive numbers and let $a_+ = \sum_i a_i$. Show that

$$\Big(a_1 \cdot a_2 \cdots a_n\Big)^{1/n} \le \frac{a_+}{n}, \tag{3}$$

i.e., geometric mean $\le$ arithmetic mean.

This can be proved by the non-negativity of the KL divergence. Consider two discrete distributions over $1, 2, \ldots, n$ with

$$\mathbb{P}(X = j) = \frac{1}{n}, \quad \mathbb{P}(Y = j) = \frac{a_j}{a_+} \quad j = 1, \ldots, n.$$

The KL divergence between the two random variables (i.e., two distributions) X and Y is

$$\begin{aligned} & \sum_{i=1}^n \frac{1}{n} \log \frac{1/n}{a_i/a_+} = \sum_{i=1}^n \frac{1}{n} \log \frac{a_+/n}{a_i} \\ = & \sum_{i=1}^n \frac{1}{n} \log \frac{a_+}{n} - \sum_{i=1}^n \frac{1}{n} \log a_i \\ = & \log \frac{a_+}{n} - \log \Big(a_1 \cdot a_2 \cdots a_n\Big)^{1/n}, \end{aligned}$$

which is always ≥ 0, therefore leads to (3).