MATH 170B LECTURE NOTE 1: ADDITIONAL TOPICS IN RANDOM VARIABLES

HANBAEK LYU

1. Random variable as a function of another random variable

We have studied some of the fundamental RVs such as Bernoulli, Binomial, geometric and Poisson for discrete RVs and uniform, exponential, and normal for continuous RVs. Since different RVs can be used to model different situations, it is desirable to enlarge our vocabulary of RVs. A nice way to doing so is to compose a RV with a function to get a new RV.

1.1. **Functions of a single RV.** Suppose X is a RV with a known CDF. We may define a new random variable $Y = g(X)$ for a function $g:\mathbb{R}\to\mathbb{R}$. Can we derive CDF of Y and also its PDF (or PMF if X is discrete)? If we can solve the CDF of $Y = g(X)$, $\mathbb{P}(g(X)\le x)$, and recognize it as the CDF of some known RV, then we can identify what RV Y is.

Example 1.1 (Exponential from uniform)**.** Let $X \sim \text{Uniform}([0,1))$ and fix a constant $\lambda > 0$. Define a random variable $Y = -\frac{1}{\lambda}\log(1-X)$. Then $Y \sim \text{Exp}(\lambda)$. To see this, we calculate the CDF of Y as below:

$$\begin{aligned}\mathbb{P}(Y\le y) &= \mathbb{P}\left(-\frac{1}{\lambda}\log(1-X)\le y\right) &(1)\\ &= \mathbb{P}\left(\log(1-X)\ge -\lambda y\right) &(2)\\ &= \mathbb{P}\left(1-X\ge e^{-\lambda y}\right) &(3)\\ &= \mathbb{P}\left(X\le 1-e^{-\lambda y}\right) &(4)\\ &= (1-e^{-\lambda y})\mathbf{1}(y\ge 0). &(5)\end{aligned}$$

Hence the CDF of Y is that of an exponential RV with rate λ. Note that the third equality above uses the fact that exponential function is an increasing function. ▲

Remark 1.2. In fact, this is how a computer generates an exponential RV: it first samples a uniform RV X from $[0,1)$, and then outputs $-\frac{1}{\lambda}\log(1-X)$. So the computer does not need to know the exponential distribution in order to generate exponential RVs.

In general, we can at least describe the PDF of $Y = g(X)$ if we know the PDF of X. See the following example for an illustration.

Example 1.3. Let X be a RV with PDF f_X. Define $Y = X^2$. Then

$$\begin{aligned}F_Y(y) = \mathbb{P}(Y\le y) = \mathbb{P}(X^2\le y) &= \mathbb{P}(-\sqrt{y}\le X\le\sqrt{y})\mathbf{1}(y\ge 0) &(6)\\ &= \mathbf{1}(y\ge 0)\int_{-\sqrt{y}}^{\sqrt{y}} f_X(t)\,dt. &(7)\end{aligned}$$

By fundamental theorem of calculus and chain rule, we can differentiate the last expression by y and get

$$f_Y(y) = \frac{dF_Y(y)}{dy} = \mathbf{1}(y\ge 0)\left(f_X(\sqrt{y})\frac{1}{2\sqrt{y}} + f_X(-\sqrt{y})\frac{1}{2\sqrt{y}}\right). \tag{8}$$

▲

A particularly simple but useful instance is when g is a linear function. We first record a general observation.

Proposition 1.4 (Linear transform). *Let X be a RV with PDF f_X. Fix constants $a, b \in \mathbb{R}$ with $a > 0$, and define a new RV $Y = aX + b$. Then*

$$f_{aX+b}(y) = \frac{1}{|a|} f_X((y-b)/a). \tag{9}$$

Proof. First suppose $a > 0$. Then

$$F_Y(y) = \mathbb{P}(Y \le y) = \mathbb{P}(aX + b \le y) = \mathbb{P}(X \le (y-b)/a) = \int_{-\infty}^{(y-b)/a} f_X(t)\,dt. \tag{10}$$

By differentiating the last integral by y, we get

$$f_Y(y) = \frac{dF_Y(y)}{dy} = \frac{1}{a} f_X((y-b)/a). \tag{11}$$

For $a < 0$, a similar calculation shows

$$F_Y(y) = \mathbb{P}(Y \le y) = \mathbb{P}(aX + b \le y) = \mathbb{P}(X \ge (y-b)/a) = \int_{(y-b)/a}^{\infty} f_X(t)\,dt, \tag{12}$$

so we get

$$f_Y(y) = \frac{dF_Y(y)}{dy} = -\frac{1}{a} f_X((y-b)/a). \tag{13}$$

This shows the assertion. □

Example 1.5 (Linear transform of normal RV is normal). Let $X \sim N(\mu, \sigma^2)$ and fix constants $a, b \in \mathbb{R}$ with $a \ne 0$. Define a new RV $Y = aX + b$. Then since

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \tag{14}$$

by Proposition 1.4, we have

$$f_Y(y) = \frac{1}{\sqrt{2\pi(a\sigma)^2}} \exp\left(-\frac{(y-b-a\mu)^2}{2(a\sigma)^2}\right). \tag{15}$$

Notice that this is the PDF of a normal RV with mean $a\mu + b$ and variance $(a\sigma)^2$. In particular, if we take $a = 1/\sigma$ and $b = \mu/\sigma$, then $Y = (X-\mu)/\sigma \sim N(0,1)$, the standard normal RV. This is called *standardization* of normal RV. ▲

Exercise 1.6 (Linear transform of exponential RV). Let $X \sim \text{Exp}(\lambda)$ and fix constants $a, b \in \mathbb{R}$ with $a \ne 0$. Show that

$$f_{aX+b}(x) = \frac{\lambda}{|a|} e^{-\lambda(x-b)/a} \mathbf{1}((x-b)/a > 0). \tag{16}$$

Is $aX + b$ always an exponential RV?

If we compare Example 1.1 and Proposition 1.4 against Example 1.3, we see the invertibility of the function g makes the computation of $F_{g(X)}$ much cleaner. Moreover, if we inspect the formula (9) more closely, we see that the constant factor $1/|a|$ is in fact $1/|g'(x)|$ and $(y-b)/a$ is the inverse function of g, where $g(x) = ax + b$. This leads us to the following observation.

Proposition 1.7 (invertible transform)**.** *Let X be a RV with PDF f_X. Let $g : \mathbb{R} \to \mathbb{R}$ be a differentiable and invertible function. Then we have*

$$f_{g(X)}(y) = (g^{-1})'(y) f_X(g^{-1}(y)) = \frac{1}{|g'(g^{-1}(y))|} f_X(g^{-1}(y)). \tag{17}$$

Proof. Since g is invertible, g is either strictly increasing or strictly decreasing. First suppose the former, so $g' > 0$ everywhere. Then

$$F_{g(X)}(y) = \mathbb{P}(g(X) \le y) = \mathbb{P}(X \le g^{-1}(y)) = \int_{-\infty}^{g^{-1}(y)} f_X(t)\,dt. \tag{18}$$

Recall that since $g^{-1}(g(x)) = x$, by chain rule we have $(g^{-1})'(g(x)) \cdot g'(x) = 1$. If we write $y = g(x)$, then

$$(g^{-1})'(y) = 1/g'(g^{-1}(y)). \tag{19}$$

Hence differentiating (20) gives (17).

Second, suppose g is strictly decreasing so $g' < 0$ everywhere. Then

$$F_{g(X)}(y) = \mathbb{P}(g(X) \le y) = \mathbb{P}(X \ge g^{-1}(y)) = \int_{g^{-1}(y)}^{\infty} f_X(t)\,dt, \tag{20}$$

so differentiating by y and using (19) also gives (17), as desired. □

Exercise 1.8 (Cauchy from uniform)**.** Let $X \sim \text{Uniform}((-\pi/2, \pi/2))$. Define $Y = \tan(X)$.

(i) Show that $d\tan(y)/dy = \sec^2(y)$.

(ii) Show that $1 + \tan^2(y) = \sec^2(y)$. (Hint: draw a right triangle with angle y)

(iii) Recall that arctan is the inverse function of tan. Show that $\arctan(t)$ is strictly increasing and differentiable. Furthermore, show that

$$\frac{d}{dt}\arctan(t) = \frac{1}{1+t^2}. \tag{21}$$

(iv) Show that Y is a standard Cauchy random variable, that is,

$$f_Y(y) = \frac{1}{\pi(1+y^2)}. \tag{22}$$

Remark 1.9. The expectation of Cauchy random variables are not well-defined. We say the expectation of a continuous RV with PDF $f_X(x)$ is well-defined if

$$\int_{-\infty}^{\infty} |x| f_X(x)\,dx < \infty. \tag{23}$$

But for the standard Cauchy distribution,

$$\int_{-\infty}^{\infty} |x| f_X(x)\,dx = 2\int_{0}^{\infty} |x| f_X(x)\,dx = \frac{2}{\pi}\int_0^\infty \frac{x}{1+x^2}\,dx \approx \frac{2}{\pi}\int_0^\infty \frac{x}{x^2}\,dx = \frac{2}{\pi}\int_0^\infty \frac{1}{x}\,dx = \infty. \tag{24}$$

In such situation, we say the distribution is 'heavy-tailed'.

1.2. **Functions of two random variables.** We can also cook up a RV from two random variables. Namely, if X, Y are RVs and g is a two variable function $g : \mathbb{R}^2 \to \mathbb{R}$, then $Z = g(X, Y)$ is a random variable.

Example 1.10. Let $X_1 \sim \text{Exp}(\lambda_1)$ and $X_2 \sim \text{Exp}(\lambda_2)$ and suppose they are independent. Define $Y = \max(X_1, X_2)$. To calculate its CDF, note that

$$\mathbb{P}(Y \le y) = \mathbb{P}(\max(X_1, X_2) \le y) = \mathbb{P}(X_1 \le y \text{ and } X_2 \le y) = \mathbb{P}(X_1 \le y)\mathbb{P}(X_2 \le y), \tag{25}$$

where the last equality uses the independence between X_1 and X_2. Using the CDF of exponential RV,

$$\mathbb{P}(Y \le y) = (1 - e^{-\lambda_1 y})(1 - e^{-\lambda_2 y})\mathbf{1}(y \ge 0). \tag{26}$$

Differentiating by y, we get the PDF of Y

$$\begin{aligned} f_Y(y) &= \lambda_1 e^{-\lambda_1 y}(1 - e^{-\lambda_2 y}) + (1 - e^{-\lambda_1 y})\lambda_2 e^{-\lambda_2 y} && (27) \\ &= \lambda_1 e^{-\lambda_1 y} + \lambda_2 e^{-\lambda_2 y} - (\lambda_1 + \lambda_2)e^{-(\lambda_1+\lambda_2)y}. && (28) \end{aligned}$$

▲

Exercise 1.11. Let $X_1 \sim \text{Exp}(\lambda_1)$ and $X_2 \sim \text{Exp}(\lambda_2)$ and suppose they are independent. Define $Y = \min(X_1, X_2)$. Show that $Y \sim \text{Exp}(\lambda_1 + \lambda_2)$. (Hint: Compute $\mathbb{P}(Y \ge y)$.)

Exercise 1.12. Let $X, Y \sim \text{Uniform}([0,1])$ be independent uniform RVs. Define $Z = X + Y$. Observe that the pair (X, Y) is uniformly distributed over the unit square $[0,1]^2$. So

$$\mathbb{P}(Z \le z) = \mathbb{P}(X + Y \le z) = \text{Area of the region } \{(x, y) \in [0,1]^2 \,|\, x + y \le z\}. \tag{29}$$

(i) Draw a picture shows that

$$\mathbb{P}(Z \le z) = \begin{cases} z^2/2 & \text{if } 0 \le z \le 1 \\ 1 - (2-z)^2/2 & \text{if } 1 \le z \le 2 \\ 0 & \text{otherwise.} \end{cases} \tag{30}$$

(ii) Conclude that

$$f_Z(z) = \begin{cases} z & \text{if } 0 \le z \le 1 \\ 2 - z & \text{if } 1 \le z \le 2 \\ 0 & \text{otherwise.} \end{cases} \tag{31}$$

1.3. **Sums of independent RVs – Convolution.** When two RVs X and Y are independent and if the new random variable Z is their sum $X + Y$, then the distribution Z is given by the *convolution* of PMFs (or PDFs) of each RV. The idea should be clear from the following baby example.

Example 1.13 (Two dice)**.** Roll two dice independently and let their outcome be recorded by RVs X and Y. Note that both X and Y are uniformly distributed over $\{1,2,3,4,5,6\}$. So the pair (X, Y) is uniformly distributed over the (6×6) integer grid $\{1,2,3,4,5,6\} \times \{1,2,3,4,5,6\}$. In other words,

$$\mathbb{P}((X, Y) = (x, y)) = \mathbb{P}(X = x)\mathbb{P}(Y = y) = \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36}. \tag{32}$$

Now, what is the distribution of the sum $Z = X + Y$? Since each point (x, y) in the grid is equality probable, we just need to count the number of such points on the line $x + y = z$, for each value of z. In other words,

$$\mathbb{P}(X+Y=z)=\sum_{x=1}^{6}\mathbb{P}(X=x)\mathbb{P}(Y=z-x). \tag{33}$$

This is easy to compute from the following picture: For example, $\mathbb{P}(X+Y=7)=6/36=1/6$. ▲

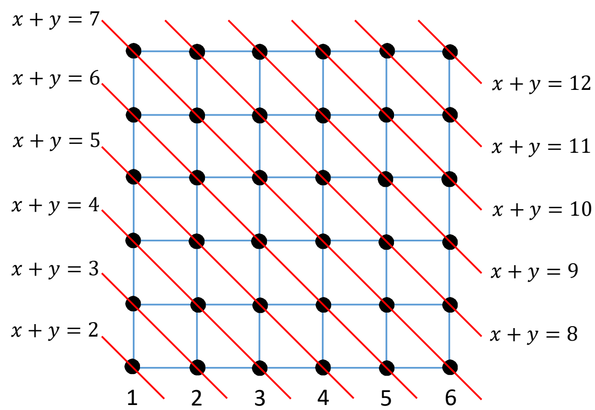

FIGURE 1. Probability space for two dice and lines on which sum of the two are constant.

Proposition 1.14 (Convolution of PMFs)**.** *Let X, Y be two independent integer-valued RVs. Let $Z = X + Y$. Then*

$$\mathbb{P}(Z=z)=\sum_{x}\mathbb{P}(X=x)\mathbb{P}(Y=z-x). \tag{34}$$

Proof. Note that the pair (X, Y) is distributed over $\mathbb{Z}^2$ according to the distribution

$$\mathbb{P}(X=x,\,Y=y)=\mathbb{P}(X=x)\mathbb{P}(Y=y), \tag{35}$$

since X and Y are independent. Hence in order to get $\mathbb{P}(Z=z)=\mathbb{P}(X+Y=z)$, we need to add up all probabilities of the pairs (x, y) over the line $x + y = z$. If we first fix the values of x, then y should take value $z - x$. Varying the range of x, we get (34). □

Exercise 1.15 (Sum of ind. Poisson RVs is Poisson)**.** Let $X \sim \text{Poisson}(\lambda_1)$ and $Y \sim \text{Poisson}(\lambda_2)$ be independent Poisson RVs. Show that $X + Y \sim \text{Poisson}(\lambda_1 + \lambda_2)$.

For the continuous case, a similar observation should hold as well. Namely, we should be integrating all the probabilities of the pair (X, Y) at points (x, y) along the line $x + y = z$ in order to get the probability density $f_{X+Y}(z)$. We will show this in the following proposition using Fubini's theorem and change of variables.

Proposition 1.16 (Convolution of PDFs)**.** *Let X, Y be two independent RVs with PDFs f_X and f_Y, respectively. Then the RV $Z := X + Y$ has PDF*

$$f_Z(z)=\int_{-\infty}^{\infty}f_X(x)f_Y(z-x)\,dx. \tag{36}$$

Proof. As usual, we begin with computing the CDF of Z. Note that since X, Y are independent, the pair (X, Y) is distributed over the plane $\mathbb{R}^2$ according to the distribution

$$f_{X,Y}(x, y) = f_X(x) f_Y(y). \tag{37}$$

So we can write the probability $\mathbb{P}(Z \le z)$ as the following double integral

$$\mathbb{P}(Z \le z) = \int_{x+y\le z} f_X(x) f_Y(y)\, dy\, dx \tag{38}$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{z-x} f_X(x) f_Y(y)\, dy\, dx, \tag{39}$$

where for the second inequality we have used Fubini's theorem. Next, make a change of variable $t = x + y$. Then $y = t - x$ and $dy = dt$, so

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{z} f_X(x) f_Y(t-x)\, dt\, dx. \tag{40}$$

Swapping the order of dt and dx by using Fubini one more time,

$$= \int_{-\infty}^{z} \int_{-\infty}^{\infty} f_X(x) f_Y(t-x)\, dx\, dt = \int_{-\infty}^{z} g(t)\, dt, \tag{41}$$

where we have written the inner integral as a function of t. By differentiating with respect to z, we get

$$f_Z(z) = g(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z-x)\, dx. \tag{42}$$

□

Example 1.17. Let $X, Y \sim N(0,1)$ be independent standard normal RVs. Let $Z = X + Y$. We will show that $Z \sim N(0,2)$ using the convolution formula. Recall that X and Y have the following PDFs:

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \qquad f_Y(y) = \frac{1}{\sqrt{2\pi}} e^{-y^2/2} \tag{43}$$

By taking convolution of the above PDFs, we have

$$f_Z(z) = \int_{-\infty}^{\infty} \left(\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)\right)\left(\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(z-x)^2}{2}\right)\right) dx \tag{44}$$

$$= \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp\left(-\frac{x^2}{2} - \frac{(z-x)^2}{2}\right) dx \tag{45}$$

$$= \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp\left(-x^2 + xz - \frac{z^2}{2}\right) dx \tag{46}$$

$$= \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp\left(-\left(x - \frac{z}{2}\right)^2 - \frac{z^2}{4}\right) dx \tag{47}$$

$$= \frac{1}{\sqrt{4\pi}} e^{-z^2/4} \int_{-\infty}^{\infty} \frac{1}{\sqrt{\pi}} \exp\left(-\left(x - \frac{z}{2}\right)^2\right) dx = \frac{1}{\sqrt{4\pi}} e^{-z^2/4}, \tag{48}$$

where we have recognized the integrand in the line as the PDF of $N(-z/2, 1/2)$ so that the integral is 1. Since the last expression is the PDF of $N(0,2)$, it follows that $Z \sim N(0,2)$. ▲.

The following example generalizes the observation we made in the previous example.

Example 1.18 (Sum of ind. normal RVs is normal)**.** Let $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$ be independent normal RVs. We will see that $Z = X + Y$ is again a normal random variable with distribution $N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$. The usual convolution computation for this is pretty messy (c.f., wikipedia article). Instead let's save some work by using the fact that normal distributions are preserved under linear transform (Exercise 1.5). So instead of X and Y, we may consider $X' := (X - \mu_1)/\sigma_1$ and $Y' := (Y - \mu_1)/\sigma_1$ (It is important to note that we must use the same linear transform here for X and Y). Then $X' \sim N(0,1)$, and $Y' \sim N(\mu, \sigma^2)$ where $\mu = (\mu_2 - \mu_1)/\sigma_1$ and $\sigma = \sigma_2/\sigma_1$. Now it suffices to show that $Z' := X' + Y' \sim N(\mu, 1 + \sigma^2)$ (see the following exercise for details).

To compute the convolution of the corresponding normal PDFs:

$$f_Z(z) = \int_{-\infty}^{\infty} \left(\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)\right)\left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(z-x-\mu)^2}{2\sigma^2}\right)\right) dx \tag{49}$$

$$= \frac{1}{2\pi\sigma} \int_{-\infty}^{\infty} \exp\left(-\frac{x^2}{2} - \frac{(z-x-\mu)^2}{2\sigma^2}\right) dx. \tag{50}$$

At this point, we need to 'complete the square' for x for the bracket inside the exponential as below:

$$\frac{x^2}{2} + \frac{(z-x-\mu)^2}{2\sigma^2} = \frac{1}{2\sigma^2}\left(\sigma^2 x^2 + (x + \mu - z)^2\right) \tag{51}$$

$$= \frac{1+\sigma^2}{2\sigma^2}\left(x^2 + \frac{2(\mu - z)x}{1+\sigma^2} + \frac{(\mu-z)^2}{1+\sigma^2}\right) \tag{52}$$

$$= \frac{1+\sigma^2}{2\sigma^2}\left[\left(x + \frac{(\mu-z)}{1+\sigma^2}\right)^2 + \frac{(\mu-z)^2}{1+\sigma^2} - \frac{(\mu-z)^2}{(1+\sigma^2)^2}\right] \tag{53}$$

$$= \frac{1+\sigma^2}{2\sigma^2}\left[\left(x + \frac{(\mu-z)}{1+\sigma^2}\right)^2 + \frac{(z-\mu)^2}{1+\sigma^2}\frac{\sigma^2}{1+\sigma^2}\right] \tag{54}$$

$$= \frac{1+\sigma^2}{2\sigma^2}\left(x + \frac{(\mu-z)}{1+\sigma^2}\right)^2 + \frac{(z-\mu)^2}{2(1+\sigma^2)}. \tag{55}$$

Now rewriting (50),

$$f_Z(z) = \frac{1}{2\pi\sigma}\int_{-\infty}^{\infty} \exp\left(-\frac{(z-\mu)^2}{2(1+\sigma^2)}\right) \exp\left(-\frac{1+\sigma^2}{2\sigma^2}\left(x + \frac{(\mu-z)}{1+\sigma^2}\right)^2\right) dx \tag{56}$$

$$= \frac{1}{\sqrt{2\pi(1+\sigma^2)}} \exp\left(-\frac{(z-\mu)^2}{2(1+\sigma^2)}\right) \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\frac{\sigma^2}{1+\sigma^2}}} \exp\left(-\frac{\left(x + \frac{(\mu-z)}{1+\sigma^2}\right)^2}{\frac{2\sigma^2}{1+\sigma^2}}\right) dx \tag{57}$$

$$= \frac{1}{\sqrt{2\pi(1+\sigma^2)}} \exp\left(-\frac{(z-\mu)^2}{2(1+\sigma^2)}\right), \tag{58}$$

where we have recognized the integral after second equality as that of the PDF of a normal RV with mean $\frac{z-\mu}{1+\sigma^2}$ and variance $\frac{\sigma^2}{1+\sigma^2}$. Hence $Z' \sim N(\mu, 1+\sigma^2)$, as desired. ▲

Exercise 1.19. Let X, Y be independent RVs and fix constants $a > 0$ and $b \in \mathbb{R}$.

(i) Show that $X + Y$ is a normal RV if and only if $(aX + b) + (aY + b)$ is so.

(ii) Show that $X + Y$ is a normal RV, then $X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$, where $\mu_1 = \mathbb{E}(X)$, $\mu_2 = \mathbb{E}(Y)$, $\sigma_1 = \mathrm{Var}(X)$, and $\sigma_1^2 = \mathrm{Var}(Y)$.

Exercise 1.20 (Sum of i.i.d. Exp is Erlang)**.** Let $X_1, X_2, \cdots, X_n \sim \mathrm{Exp}(\lambda)$ be independent exponential RVs.

(i) Show that $f_{X_1+X_2}(z) = \lambda^2 z e^{-\lambda z}\mathbf{1}(z \ge 0)$.
(ii) Show that $f_{X_1+X_2+X_3}(z) = 2^{-1}\lambda^3 z^2 e^{-\lambda z}\mathbf{1}(z \ge 0)$.
(iii) Let $S_n = X_1 + X_2 + \cdots + X_n$. Use induction to show that $S_n \sim \text{Erlang}(n,\lambda)$, that is,

$$f_{S_n}(z) = \frac{\lambda^n z^{n-1} e^{-\lambda z}}{(n-1)!}. \tag{59}$$

2. Covariance and Correlation

2.1. **Covariance.** When two RVs X and Y are independent, we know that the pair (X,Y) is distributed according to the product distribution $\mathbb{P}((X,Y)=(x,y)) = \mathbb{P}(X=x)\mathbb{P}(Y=y)$ and we can say a lot of things about their sum, difference, product, maximum, etc. For instance, the expectation of their product is the product of their expectations:

Exercise 2.1. Let X and Y be two independent RVs. Show that $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$.

But what if they are not independent? Then their joint distribution $\mathbb{P}((X,Y)=(x,y))$ can be very much different from the product distribution $\mathbb{P}(X=x)\mathbb{P}(Y=y)$. Covariance is the quantity that measures the 'average disparity' between the true joint distribution $\mathbb{P}((X,Y)=(x,y))$ and the product distribution $\mathbb{P}(X=x)\mathbb{P}(Y=y)$.

Definition 2.2 (Covariance)**.** Given two RVs X and Y, their *covariance* is denoted by $\text{Cov}(X,Y)$ and is defined by

$$\text{Cov}(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y). \tag{60}$$

We say X and Y are *correlated* (resp., *uncorrelated*) if $\text{Cov}(X,Y) \ne 0$ (resp., $\text{Cov}(X,Y) = 0$).

Exercise 2.3. Show the following.
(i) $\text{Cov}(X,X) = \text{Var}(X)$.
(ii) $\text{Cov}(X,Y) = \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))]$.

Exercise 2.4. Show that two RVs X and Y are uncorrelated if they are independent.

Example 2.5 (Uncorrelated but dependent)**.** Two random variables can be uncorrelated but still be dependent. Let (X,Y) be a uniformly sampled point from the unit circle in the 2-dimensional plane. Parameterize the unit circle by $S^1 = \{(\cos\theta, \sin\theta) \,|\, 0 \le \theta < 2\pi\}$. Then we can first sample a uniform angle $\Theta \sim \text{Uniform}([0,2\pi))$, and then define $(X,Y) = (\cos\Theta, \sin\Theta)$. Recall from your old memory that

$$\sin 2t = 2\cos t \sin t. \tag{61}$$

Now

$$\begin{aligned}
\mathbb{E}(XY) &= \mathbb{E}(\cos\Theta\sin\Theta) && (62)\\
&= \frac{1}{2}\mathbb{E}(\sin 2\Theta) && (63)\\
&= \frac{1}{2}\int_0^{2\pi} \sin 2t\,dt && (64)\\
&= \frac{1}{2}\left[-\frac{1}{2}\cos 2t\right]_0^{2\pi} = 0. && (65)
\end{aligned}$$

On the other hand,

$$\mathbb{E}(X) = \mathbb{E}(\cos\Theta) = \int_0^{2\pi} \cos t \, dt = 0 \tag{66}$$

and likewise $\mathbb{E}(Y) = 0$. This shows $\text{Cov}(X, Y) = 0$, so X and Y are uncorrelated. However, they satisfy the following deterministic relation

$$X^2 + Y^2 = 1, \tag{67}$$

so clearly they cannot be independent. ▲

So if uncorrelated RVs can be dependent, what does the covariance really measure? It turns out, $\text{Cov}(X, Y)$ measures the 'linear tendency' between X and Y.

Example 2.6 (Linear transform)**.** Let X be a RV, and define another RV Y by $Y = aX + b$ for some constants $a, b \in \mathbb{R}$. Let's compute their covariance using linearity of expectation.

$$\text{Cov}(X, Y) = \text{Cov}(X, aX + b) \tag{68}$$
$$= \mathbb{E}(aX^2 + bX) - \mathbb{E}(X)\mathbb{E}(aX + b) \tag{69}$$
$$= a\mathbb{E}(X^2) + b\mathbb{E}(X) - \mathbb{E}(X)(a\mathbb{E}(X) + b) \tag{70}$$
$$= a[\mathbb{E}(X^2) - \mathbb{E}(X)^2] \tag{71}$$
$$= a\text{Var}(X). \tag{72}$$

Thus, $\text{Cov}(X, aX + b) > 0$ if $a > 0$ and $\text{Cov}(X, aX + b) < 0$ if $a < 0$. In other words, if $\text{Cov}(X, Y) > 0$, then X and Y tend to be large at the same time; if $\text{Cov}(X, Y) > 0$, then Y tends to be small if X tends to be large. ▲

From the above example, it is clear that why the x- and y-coordinates of a uniformly sampled point from the unit circle are uncorrelated – they have no linear relation!

Exercise 2.7 (Covariance is symmetric and bilinear)**.** Let X and Y be RVs and fix constants $a, b \in \mathbb{R}$. Show the following.

(i) $\text{Cov}(aX + b, Y) = a\text{Cov}(X, Y)$.
(ii) $\text{Cov}(X + Z, Y) = \text{Cov}(X, Y) + \text{Cov}(Z, Y)$.
(iii) $\text{Cov}(X, Y) = \text{Cov}(Y, X)$.

Next, let's say four RVs X, Y, Z, and W are given. Suppose that $\text{Cov}(X, Y) > \text{Cov}(Z, W) > 0$. Can we say that 'the positive linear relation' between X and Y is stronger than that between Z and W? Not quite.

Example 2.8. Suppose X is a RV. Let $Y = 2X$, $Z = 2X$, and $W = 4X$. Then

$$\text{Cov}(X, Y) = \text{Cov}(X, 2X) = 2\text{Var}(X), \tag{73}$$

and

$$\text{Cov}(Z, W) = \text{Cov}(2X, 4X) = 8\text{Var}(X). \tag{74}$$

But $Y = 2X$ and $W = 2Z$, so the linear relation between the two pairs should be same. ▲

So to compare the magnitude of covariance, we first need to properly normalize covariance so that the effect of fluctuation (variance) of each coordinate is not counted: then only the correlation between the two coordinates will contribute. This is captured by the following quantity.

Definition 2.9 (Correlation coefficient)**.** Given two RVs X and Y, their *correlation coefficient* $\rho(X,Y)$ is defined by

$$\rho(X,Y)=\frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}\,Y}}. \tag{75}$$

Example 2.10. Suppose X is a RV and fix constants $a,b\in\mathbb{R}$. Then

$$\rho(X,aX+b)=\frac{a\mathrm{Cov}(X,X)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(aX+b)}}=\frac{a\,\mathrm{Var}(X)}{\sqrt{\mathrm{Var}(X)}\sqrt{a^2\,\mathrm{Var}(X)}}=\frac{a}{|a|}=\mathrm{sign}(a). \tag{76}$$

▲

Exercise 2.11 (Cauchy-Schwarz inequality)**.** Let X,Y are RVs. Suppose $\mathbb{E}(Y^2)>0$. We will show that the 'inner product' of X and Y is at most the product of their 'magnitudes'

(i) For any $t\in\mathbb{R}$, show that

$$\mathbb{E}[(X-tY)^2]=t^2\mathbb{E}(Y^2)-2t\mathbb{E}(XY)+\mathbb{E}(X^2) \tag{77}$$

$$=\mathbb{E}(Y^2)\left(t-\frac{\mathbb{E}(XY)}{\mathbb{E}(Y^2)}\right)^2+\frac{\mathbb{E}(X^2)\mathbb{E}(Y^2)-\mathbb{E}(XY)^2}{\mathbb{E}(Y^2)}. \tag{78}$$

Conclude that

$$0\le\mathbb{E}\left[\left(X-\frac{\mathbb{E}(XY)}{\mathbb{E}(Y^2)}Y\right)^2\right]=\frac{\mathbb{E}(X^2)\mathbb{E}(Y^2)-\mathbb{E}(XY)^2}{\mathbb{E}(Y^2)}. \tag{79}$$

(ii) Show that a RV Z satisfies $\mathbb{E}(Z^2)=0$ if and only if $\mathbb{P}(Z=0)=1$.

(iii) Show that

$$\mathbb{E}(XY)\le\sqrt{\mathbb{E}(X^2)}\sqrt{\mathbb{E}(Y^2)}, \tag{80}$$

where the equality holds if and only if

$$\mathbb{P}\left(X=\frac{\mathbb{E}(XY)}{\mathbb{E}(Y^2)}Y\right)=1. \tag{81}$$

Exercise 2.12. Let X,Y are RVs such that $\mathrm{Var}(Y)>0$. Let $\bar{X}=X-\mathbb{E}(X)$ and $\bar{Y}=Y-\mathbb{E}(Y)$.

(i) Use (79) to show that

$$0\le\mathbb{E}\left[\left(\bar{X}-\frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(Y)}\bar{Y}\right)^2\right]=\mathrm{Var}(X)\left(1-\rho(X,Y)^2\right). \tag{82}$$

(ii) Show that $|\rho(X,Y)|\le 1$.

(iii) Show that $|\rho(X,Y)|=1$ if and only if $\bar{X}=a\bar{Y}$ for some constant $a\neq 0$.

2.2. **Variance of sum of RVs.** Let X,Y be RVs. If they are not necessarily independent, what is the variance of their sum? Using linearity of expectation, we compute

$$\mathrm{Var}(X+Y)=\mathbb{E}[(X+Y)^2]-\mathbb{E}(X+Y)^2 \tag{83}$$

$$=\mathbb{E}[X^2+Y^2+2XY]-(\mathbb{E}(X)+\mathbb{E}(Y))^2 \tag{84}$$

$$=[\mathbb{E}(X^2)-\mathbb{E}(X)^2]+[\mathbb{E}(Y^2)-\mathbb{E}(Y)^2]+2[\mathbb{E}(XY)-\mathbb{E}(X)\mathbb{E}(Y)] \tag{85}$$

$$=\mathrm{Var}(X)+\mathrm{Var}(Y)+2\mathrm{Cov}(X,Y). \tag{86}$$

Note that $\mathrm{Cov}(X,Y)$ shows up in this calculation. We can push this computation for sum of more than just two RVs.

Proposition 2.13. *For RVs $X_1, X_2 \cdots, X_n$, we have*

$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{Var}(X_i) + 2 \sum_{1\le i,j\le n} \mathrm{Cov}(X_i, X_j). \tag{87}$$

Proof. By linearity of expectation, we have

$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \mathbb{E}\left[\left(\sum_{i=1}^{n} X_i\right)^2\right] - \left(\mathbb{E}\left[\sum_{i=1}^{n} X_i\right]\right)^2 \tag{88}$$

$$= \mathbb{E}\left[\sum_{1\le i,j\le n} X_i X_j\right] - \sum_{1\le i,j\le n} \mathbb{E}(X_i)\mathbb{E}(X_j) \tag{89}$$

$$= \left[\sum_{1\le i,j\le n} \mathbb{E}(X_i X_j)\right] - \sum_{1\le i,j\le n} \mathbb{E}(X_i)\mathbb{E}(X_j) \tag{90}$$

$$= \sum_{1\le i,j\le n} \left[\mathbb{E}(X_i X_j) - \mathbb{E}(X_i)\mathbb{E}(X_j)\right] \tag{91}$$

$$= \sum_{1\le i\le n} [\mathbb{E}(X_i X_i) - \mathbb{E}(X_i)\mathbb{E}(X_i)] + \sum_{1\le i\ne j\le n} \left[\mathbb{E}(X_i X_j) - \mathbb{E}(X_i)\mathbb{E}(X_j)\right] \tag{92}$$

$$= \sum_{1\le i\le n} \mathrm{Var}(X_i) + 2 \sum_{1\le i<j\le n} \left[\mathbb{E}(X_i X_j) - \mathbb{E}(X_i)\mathbb{E}(X_j)\right] \tag{93}$$

$$= \sum_{1\le i\le n} \mathrm{Var}(X_i) + 2 \sum_{1\le i<j\le n} \mathrm{Cov}(X_i, X_j). \tag{94}$$

□

Exercise 2.14. Show that for independent RVs $X_1, X_2, \cdots, X_n$, we have

$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{Var}(X_i). \tag{95}$$

Example 2.15 (Number of fixed point in a random permutation)**.** Suppose n people came to a party and somehow the host mixed up their car keys and gave them back completely randomly at the end of the party. Let X_i be a RV, which takes value 1 if person i got the right key and 0 otherwise. Let $N_n = X_1 + X_2 + \cdots + X_n$ be the total number of people who got their own keys back. We will show that $\mathbb{E}(N_n) = \mathrm{Var}(N_n) = 1$.

First, we observe that each $X_i \sim \mathrm{Bernoulli}(1/n)$. So we know that $\mathbb{E}(X_i) = 1/n$ and $\mathrm{Var}(X_i) = \mathbb{E}(X_i^2) - \mathbb{E}(X_i)^2 = \mathbb{E}(X_i) - \mathbb{E}(X_i)^2 = n^{-1} - n^{-2} = (n-1)/n^2$. Clearly X_i's are not independent: If the first person got the key number 2, then the second person will never get the right key.

A very important fact is that the linearity of expectation holds regardless of dependence (c.f. Exercise 1.8 in Note 0), so

$$\mathbb{E}[N_n] = \mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{E}(X_i) = \sum_{i=1}^{n} \frac{1}{n} = 1. \tag{96}$$

On the other hand, to compute the covariance, let's take a look at $\mathbb{E}(X_1 X_2)$. Note that if the first person got her key, then the second person gets his key with probability $1/(n-1)$. So

$$\mathbb{E}(X_1 X_2) = 1 \cdot \mathbb{P}(X_1 = 1,\, X_2 = 1) = \mathbb{P}(X_1 = 1)\mathbb{P}(X_2 = 1 \,|\, X_1 = 1) = \frac{1}{n} \cdot \frac{1}{n-1}. \tag{97}$$

Hence we can compute their covariance:

$$\mathrm{Cov}(X_1, X_2) = \mathbb{E}(X_1X_2) - \mathbb{E}(X_1)\mathbb{E}(X_2) = \frac{1}{n(n-1)} - \frac{1}{n^2} = \frac{n-(n-1)}{n^2(n-1)} = \frac{1}{n^2(n-1)}. \tag{98}$$

Since there is nothing special about the pair (X_1, X_2), we get

$$\mathrm{Var}(N_n) = \sum_{i=1}^{n} \mathrm{Var}(X_i) + 2 \sum_{1\le i,j\le n} \mathrm{Cov}(X_i, X_j) \tag{99}$$

$$= \sum_{i=1}^{n} \frac{n-1}{n^2} + 2 \sum_{1\le i,j\le n} \frac{1}{n^2(n-1)} \tag{100}$$

$$= \frac{n-1}{n} + 2\binom{n}{2}\frac{1}{n^2(n-1)} \tag{101}$$

$$= \frac{n-1}{n} + 2\frac{n(n-1)}{2!}\frac{1}{n^2(n-1)} \tag{102}$$

$$= \frac{n-1}{n} + \frac{1}{n} = 1. \tag{103}$$

So in the above example, we have shown $\mathbb{E}(N) = \mathrm{Var}(N) = 1$. Does this ring a bell? If $X \sim$ Poisson(1), then $\mathbb{E}(X) = \mathrm{Var}(N) = 1$ (c.f. Exercise 1.21 in Note 0). So is N_n somehow related to the Poisson RV with rate 1? In the following two exercises, we will show that N_n approximately follows Poisson(1) if n is large. ▲

Exercise 2.16 (Derangements)**.** In reference to Example 2.15, let D_n be the total number of arrangements of n keys so that no one gets the correct key.

(i) Show that the total number of arrangements of n keys is $n!$.
(ii) Show that there are $(n-1)!$ arrangements where person 1 got the right key.
(iii) Show that there are $(n-2)!$ arrangements where person 1 and 2 got the right key.
(iv) Show that there are $(n-k)!$ arrangements where person $i_1, i_2, \cdots, i_k$ got the right key.
(v) By using inclusion-exclusion, show that

$$D_n = n! - \binom{n}{1}(n-1)! + \binom{n}{2}(n-2)! - \binom{n}{3}(n-3)! + \cdots + (-1)^n\binom{n}{n}(n-n)! \tag{104}$$

$$= n!\left(1 - \frac{1}{1!} + \frac{1}{2!} - \frac{1}{3!} + \cdots + (-1)^n\frac{1}{n!}\right) \to \frac{1}{e} \quad \text{as } n\to\infty. \tag{105}$$

Exercise 2.17. Let $N_n = X_1 + X_2 + \cdots + X_n$ be as in Example 2.15.

(i) Use Exercise 2.16 to show that for each $1 \le k \le n$,

$$\mathbb{P}(N_n = k) = \binom{n}{k}\frac{D_{n-k}}{n!} \tag{106}$$

$$= \frac{n!}{k!(n-k)!}\frac{(n-k)!}{n!}\left(1 - \frac{1}{1!} + \frac{1}{2!} - \frac{1}{3!} + \cdots + (-1)^{n-k}\frac{1}{(n-k)!}\right) \tag{107}$$

$$= \frac{1}{k!}\left(1 - \frac{1}{1!} + \frac{1}{2!} - \frac{1}{3!} + \cdots + (-1)^{n-k}\frac{1}{(n-k)!}\right). \tag{108}$$

(ii) Conclude that

$$\lim_{n\to\infty} \mathbb{P}(N_n = k) = \frac{e^{-1}}{k!} = \mathbb{P}(\mathrm{Poisson}(1) = k). \tag{109}$$

Remark 2.18. Recall that Poisson(1) can be obtained from Binomial(n,p) where $p=1/n$, for large n (c.f. Example 1.20 in Note 0). In other words, the sum of n independent Bernoulli$(1/n)$ RVs is distributed approximately as Poisson(1). In the key arrangement problem in Example 2.15, note that the correlation coefficient between X_i and X_j is very small:

$$\rho(X_i,X_j)=\frac{\mathrm{Cov}(X_i,X_j)}{\sqrt{\mathrm{Var}(X_j)}\sqrt{\mathrm{Var}(X_j)}}=\frac{n^2}{n^2(n-1)^2}=\frac{1}{(n-1)^2}. \tag{110}$$

So it's kind of make sense that X_i's are almost independent for large n, so $N_n\sim$ Poisson(1) approximately for large n.

3. Conditional expectation and variance

3.1. **Conditional expectation.** Let X,Y be discrete RVs. Recall that the expectation $\mathbb{E}(X)$ is the 'best guess' on the value of X when we do not have any prior knowledge on X. But suppose we have observed that some possibly related RV Y takes value y. What should be our best guess on X, leveraging this added information? This is called the *conditional expectation of X given $Y=y$*, which is defined by

$$\mathbb{E}[X|Y=y]=\sum_x x\mathbb{P}(X=x|Y=y). \tag{111}$$

This best guess on X given $Y=y$, of course, depends on y. So it is a function in y. Now if we do not know what value Y might take, then we omit y and $\mathbb{E}[X|Y]$ becomes a RV, which is called the *conditional expectation of X given Y*.

Example 3.1. Suppose we have a biased coin whose probability of heads is itself random and is distributed as $Y\sim$ Uniform$([0,1])$. Let's flip this coin n times and let X be the total number of heads. Given that $Y=y\in[0,1]$, we know that X follows Binomial(n,y). In general, $X|U\sim$ Binomial(n,Y). So $\mathbb{E}[X|Y=y]=ny$, and $\mathbb{E}[X\,|\,Y]=nY$. Hence as a random variable, $\mathbb{E}[X|Y]=nY\sim$ Uniform$([0,n])$. So the expectation of $\mathbb{E}[X|Y]$ is the mean of Uniform$([0,n])$, which is $n/2$. This value should be the true expectation of X. ▲

The above example suggests that if we first compute the conditional expectation of X given $Y=y$, and then average this value over all choice of y, then we should get the actual expectation of X. Justification of this observation is based on the following fact

$$\mathbb{P}(Y=y\,|\,X=x)\mathbb{P}(X=x)=\mathbb{P}(X=x,\,Y=y)=\mathbb{P}(X=x\,|\,Y=y)\mathbb{P}(Y=y). \tag{112}$$

That is, if we are interested in the event that $(X,Y)=(x,y)$, then we can either first observe the value of X and then Y, or the other way around.

Proposition 3.2 (Iterated expectation)**.** *Let X,Y be discrete RVs. Then* $\mathbb{E}(X)=\mathbb{E}[\mathbb{E}[X|Y]]$.

Proof. We are going to write the iterated expectation $\mathbb{E}[\mathbb{E}[X|Y]]$ as a double sum and swap the order of summation (Fubini's theorem, as always).

$$\mathbb{E}[\mathbb{E}[X|Y]]=\sum_y \mathbb{E}[X|Y=y]\mathbb{P}(Y=y) \tag{113}$$

$$=\sum_y\left(\sum_x x\mathbb{P}(X=x|Y=y)\right)\mathbb{P}(Y=y) \tag{114}$$

$$= \sum_{y}\sum_{x} x\mathbb{P}(X = x|Y = y)\mathbb{P}(Y = y) \tag{115}$$

$$= \sum_{y}\sum_{x} x\mathbb{P}(X = x,\, Y = y) \tag{116}$$

$$= \sum_{x}\sum_{y} x\mathbb{P}(Y = y|X = x)\mathbb{P}(X = x) \tag{117}$$

$$= \sum_{x} x\left(\sum_{y}\mathbb{P}(Y = y|X = x)\right)\mathbb{P}(X = x) \tag{118}$$

$$= \sum_{x} x\mathbb{P}(X = x) = \mathbb{E}(X). \tag{119}$$

□

Remark 3.3. Here is an intuitive reason why the iterated expectation works. Suppose you want to make the best guess $\mathbb{E}(X)$. Pretending you know Y, you can imporove your guess to be $E(X\,|\,Y)$. Then you admit that you didn't know anything about Y and average over all values of Y. The result is $\mathbb{E}[\mathbb{E}[X\,|\,Y]]$, and this should be the same best guess on X when we don't know anything about Y.

All our discussions above hold for continuous RVs as well: We simply replace the sum by integral and PMF by PDF. To summarize how we compute the iterated expectations when we condition on discrete and continuous RV:

$$\mathbb{E}[\mathbb{E}[X\,|\,Y]] = \begin{cases} \sum_{y}\mathbb{E}[X\,|\,Y = y]\mathbb{P}(Y = y) & \text{if } Y \text{ is discrete} \\ \int_{-\infty}^{\infty}\mathbb{E}[X\,|\,Y = y]f_Y(y)\,dy & \text{if } Y \text{ is continuous.} \end{cases} \tag{120}$$

Exercise 3.4 (Iterated expectation for probability)**.** Let X, Y be RVs.

(i) For any $x \in \mathbb{R}$, show that $\mathbb{P}(X \le x) = \mathbb{E}[\mathbf{1}(X \le x)]$.

(ii) By using iterated expectation, show that

$$\mathbb{P}(X \le x) = \mathbb{E}[\mathbb{P}(X \le x\,|\,Y)], \tag{121}$$

where the expectation is taken over for all possible values of Y.

Example 3.5 (Example 3.1 revisited)**.** Let $Y \sim \text{Uniform}([0,1])$ and $X \sim \text{Binomial}(n, Y)$. Then $X|Y = y \sim \text{Binomial}(n, y)$ so $\mathbb{E}[X|Y = y] = ny$. Hence

$$\mathbb{E}[X] = \int_0^1 \mathbb{E}[X\,|\,Y = y]f_Y(y)\,dy = \int_0^1 ny\,dy = n/2. \tag{122}$$

▲

Example 3.6. Let $X_1 \sim \text{Exp}(\lambda_1)$ and $X_2 \sim \text{Exp}(\lambda_2)$ be independent exponential RVs. We will show that

$$\mathbb{P}(X_1 < X_2) = \frac{\lambda_1}{\lambda_1 + \lambda_2} \tag{123}$$

using the iterated expectation. Using iterated expectation for probability,

$$\mathbb{P}(X_1 < X_2) = \int_0^{\infty} \mathbb{P}(X_1 < X_2\,|\,X_1 = x_1)\lambda_1 e^{-\lambda_1 x_1}\,dx_1 \tag{124}$$

$$= \int_0^{\infty} \mathbb{P}(X_2 > x_1)\lambda_1 e^{-\lambda_1 x_1}\,dx_1 \tag{125}$$

$$= \lambda_1 \int_0^{\infty} e^{-\lambda_2 x_1} e^{-\lambda_1 x_1}\,dx_1 \tag{126}$$

$$= \lambda_1 \int_0^\infty e^{-(\lambda_1+\lambda_2)x_1}\,dx_1 = \frac{\lambda_1}{\lambda_1+\lambda_2}. \tag{127}$$

▲

Exercise 3.7. Consider a post office with two clerks. Three people, A, B, and C, enter simultaneously. A and B go directly to the clerks, and C waits until either A or B leaves before he begins service. Let X_A be the time that A spends at a register, and define X_B and X_B similarly. Compute the probability $\mathbb{P}(X_A > X_B + X_C)$ that A leaves the post office after B and C, in the following scenarios:

(a) the service time for each clerk is exactly (nonrandom) ten minutes?
(b) the service times are i, independently with probability $1/3$ for $i \in \{1,2,3\}$?
(c) the service times are independent exponential variables with mean $1/\mu$?

Exercise 3.8. Suppose we have a stick of length L. Break it into two pieces at a uniformly chosen point and let X_1 be the length of the longer piece. Break this longer piece into two pieces at a uniformly chosen point and let X_2 be the length of the longer one. Define $X_3, X_4, \cdots$ in a similar way.

(i) Let $U \sim \text{Uniform}([0,L])$. Show that X_1 takes values from $[L/2, L]$, and that $X_1 = \max(U, L-U)$.
(ii) From (i), deduce that for any $L/2 \le x \le L$, we have

$$\mathbb{P}(X_1 \ge x) = \mathbb{P}(U \ge x \text{ or } L-U \ge x) = \mathbb{P}(U \ge x) + \mathbb{P}(U \le L-x) = \frac{2(L-x)}{L}. \tag{128}$$

Conclude that $X_1 \sim \text{Uniform}([L/2, L])$. What is $\mathbb{E}[X_1]$?

(iii) Show that $X_2 \sim \text{Uniform}([x_1/2, x_1])$ conditional on $X_1 = x_1$. Using iterated expectation, show that $\mathbb{E}[X_2] = (3/4)^2 L$.
(iv) In general, show that $X_{n+1} \,|\, X_n \sim \text{Uniform}([X_n/2, X_n])$. Conclude that $\mathbb{E}[X_n] = (3/4)^n L$.

3.2. **Conditional expectation as an estimator.** We introduced the conditional expectation $\mathbb{E}[X \,|\, Y = y]$ as the best guess on X given that $Y = y$. Such a 'guess' on a RV is called an *estimator*. Let's first take a look at two extremal cases, where observing Y gives absolutely no information on X or gives everything.

Example 3.9. Let X and Y be independent discrete RVs. Then knowing the value of Y should not yield any information on X. In other words, given that $Y = y$, the best guess of X should still be $\mathbb{E}(X)$. Indeed,

$$\mathbb{E}(X|Y=y) = \sum_{x=0}^{n} x\,\mathbb{P}(X = x \,|\, Y = y) = \sum_{x=0}^{n} x\,\mathbb{P}(X = x) = \mathbb{E}(X). \tag{129}$$

On the other hand, given that $X = x$, the best guess on X is just x, since the RV X has been revealed and there is no further randomness. In other words,

$$\mathbb{E}(X|X=x) = \sum_{z=0}^{n} z\,\mathbb{P}(X = z \,|\, X = x) = \sum_{x=0}^{n} x\mathbf{1}(z = x) = x. \tag{130}$$

▲

Exercise 3.10. Let X, Y be discrete RVs. Show that for any function $g : \mathbb{R} \to \mathbb{R}$,

$$\mathbb{E}[Xg(Y) \,|\, Y] = g(Y)\mathbb{E}[X \,|\, Y]. \tag{131}$$

We now observe some general properties of the conditional expectation as an estimator.

Exercise 3.11. Let X, Y be RVs and denote $\hat{X} = \mathbb{E}[X|Y]$, meaning that $\hat{X}$ is an estimator of X given Y. Let $\tilde{X} = \hat{X} - X$ be the *estimation error*.

(i) Show that $\hat{X}$ is an *unbiased* estimator of X, that is, $\mathbb{E}(\hat{X}) = \mathbb{E}(X)$.
(ii) Show that $\mathbb{E}[\hat{X}|Y] = \hat{X}$. Hence knowing Y does not improve our current best guess $\hat{X}$.
(iii) Show that $\mathbb{E}[\tilde{X}] = 0$.
(iv) Show that $\text{Cov}(\hat{X}, \tilde{X}) = 0$. Conclude that

$$\text{Var}(X) = \text{Var}(\hat{X}) + \text{Var}(\tilde{X}). \tag{132}$$

3.3. **Conditional variance.** As we have defined conditional expectation, we could define the variance of a RV X given that anther RV Y takes a particular value. Recall that the (unconditioned) variance of X is defined by

$$\text{Var}(X) = \mathbb{E}[(X - \mathbb{E}(X))^2]. \tag{133}$$

Note that there are two places where we take expectation. Given Y, we should improve both expectations so the *conditional variance of* X *given* Y *is defined by*

$$\text{Var}(X \,|\, Y) = \mathbb{E}[(X - \mathbb{E}[X \,|\, Y])^2 \,|\, Y]. \tag{134}$$

Proposition 3.12. *Let* X *and* Y *be RVs. Then*

$$\text{Var}(X \,|\, Y) = \mathbb{E}[X^2 \,|\, Y] - \mathbb{E}[X \,|\, Y]^2. \tag{135}$$

Proof. Using linearity of conditional expectation and the fact that $\mathbb{E}[X \,|\, Y]$ is not random given Y,

$$\begin{aligned}
\text{Var}(X \,|\, Y) &= \mathbb{E}[X^2 - 2X\mathbb{E}[X \,|\, Y] + \mathbb{E}[X \,|\, Y]^2 \,|\, Y] && (136)\\
&= \mathbb{E}[X^2 \,|\, Y] - \mathbb{E}[2X\mathbb{E}[X \,|\, Y] \,|\, Y] + \mathbb{E}[\mathbb{E}[X \,|\, Y]^2 \,|\, Y] && (137)\\
&= \mathbb{E}[X^2 \,|\, Y] - \mathbb{E}[X \,|\, Y]\mathbb{E}[2X \,|\, Y] + \mathbb{E}[X \,|\, Y]^2\mathbb{E}[1 \,|\, Y] && (138)\\
&= \mathbb{E}[X^2 \,|\, Y] - 2\mathbb{E}[X \,|\, Y]^2 + \mathbb{E}[X \,|\, Y]^2 && (139)\\
&= \mathbb{E}[X^2 \,|\, Y] - \mathbb{E}[X \,|\, Y]^2. && (140)
\end{aligned}$$

□

The following exercise explains in what sense the conditional expectation $\mathbb{E}[X \,|\, Y]$ is the best guess on X given Y, and that the minimum possible mean squared error is exactly the conditional variance $\text{Var}(X \,|\, Y)$.

Exercise 3.13. Let X, Y be RVs. For any function $g : \mathbb{R} \to \mathbb{R}$, consider $g(Y)$ as an estimator of X. Let $\mathbb{E}_Y[(X - g(Y))^2 \,|\, Y]$ be the *mean squared error.*

(i) Show that

$$\begin{aligned}
\mathbb{E}_Y[(X - g(Y))^2 \,|\, Y] &= \mathbb{E}_Y[X^2 \,|\, Y] - 2g(Y)\mathbb{E}_Y[X \,|\, Y] + g(Y)^2 && (141)\\
&= (g(Y) - \mathbb{E}_Y(X \,|\, Y))^2 + \mathbb{E}_Y[X^2 \,|\, Y] - \mathbb{E}_Y[X \,|\, Y]^2 && (142)\\
&= (g(Y) - \mathbb{E}_Y(X \,|\, Y))^2 + \text{Var}(X \,|\, Y). && (143)
\end{aligned}$$

(ii) Conclude that the mean squared error is minimized when $g(Y) = \mathbb{E}_Y[X \,|\, Y]$ and the global minimum is $\text{Var}(X \,|\, Y)$.

Next, we study how we can decompose the variance of X by conditioning on Y.

Proposition 3.14 (Law of total variance)**.** *Let* X *and* Y *be RVs. Then*

$$\text{Var}(X) = \mathbb{E}(\text{Var}(X \,|\, Y)) + \text{Var}(\mathbb{E}[X \,|\, Y]). \tag{144}$$

Proof. Using previous result, iterated expectation, and linearity of expectation, we have

$$\begin{aligned}
\mathrm{Var}(X) &= \mathbb{E}(X^2) - (\mathbb{E}(X))^2 && (145)\\
&= \mathbb{E}_Y(\mathbb{E}(X^2|Y)) - (\mathbb{E}_Y(\mathbb{E}(X|Y)))^2 && (146)\\
&= \mathbb{E}_Y(\mathrm{Var}(X|Y) + (\mathbb{E}(X|Y))^2) - (\mathbb{E}_Y(\mathbb{E}(X|Y)))^2 && (147)\\
&= \mathbb{E}_Y(\mathrm{Var}(X|Y)) + \left[\mathbb{E}_Y(\mathbb{E}(X|Y))^2) - (\mathbb{E}_Y(\mathbb{E}(X|Y)))^2\right] && (148)\\
&= \mathbb{E}_Y(\mathrm{Var}(X|Y)) + \mathrm{Var}_Y(\mathbb{E}(X|Y)). && (149)
\end{aligned}$$

□

Here is a handwavy explanation on why the above is true. Given Y, we should measure the fluctuation of $X\,|\,Y$ from the conditional expectation $\mathbb{E}[X\,|\,Y]$, and this is measured as $\mathrm{Var}(X\,|\,Y)$. Since we don't know Y, we average over all Y, giving $\mathbb{E}(\mathrm{Var}(X\,|\,Y))$. But the reference point $\mathbb{E}[X\,|\,Y]$ itself varies with Y, so we should also measure its own fluctuation by $\mathrm{Var}(\mathbb{E}[X\,|\,Y])$. These fluctuations add up nicely like Pythagorian theorem because $\mathbb{E}[X\,|\,Y]$ is an optimal estimator so that these two fluctuations are 'orthogonal'.

Exercise 3.15. Let X, Y be RVs. Write $\bar{X} = \mathbb{E}[X\,|\,Y]$ and $\tilde{X} = X - \mathbb{E}[X\,|\,Y]$ so that $X = \bar{X} + \tilde{X}$. Here $\bar{X}$ is the estimate of X given Y and $\tilde{X}$ is the estimation error.

(i) Using Exercise 3.11 (iii) and iterated expectation, show that

$$\mathbb{E}[\tilde{X}^2] = \mathrm{Var}(\mathbb{E}[X\,|\,Y]). \tag{150}$$

(ii) Using Exercise 3.11 (iv), conclude that

$$\mathrm{Var}(X) = \mathbb{E}(\mathrm{Var}(X\,|\,Y)) + \mathrm{Var}(\mathbb{E}[X\,|\,Y]). \tag{151}$$

Example 3.16. Let $Y \sim \mathrm{Uniform}([0,1])$ and $X \sim \mathrm{Binomial}(n, Y)$. Since $X\,|\,Y = y \sim \mathrm{Binomial}(n, y)$, we have $\mathbb{E}[X\,|\,Y = y] = ny$ and $\mathrm{Var}(X\,|\,Y = y) = ny(1-y)$. Also, since $Y \sim \mathrm{Uniform}([0,1])$, we have

$$\mathrm{Var}(\mathbb{E}[X\,|\,Y]) = \mathrm{Var}(nY) = \frac{n^2}{12}. \tag{152}$$

So by iterated expectation, we get

$$\mathbb{E}(X) = \mathbb{E}_Y(\mathbb{E}[X\,|\,Y]) = \int_0^1 ny\,dy = \frac{n}{2}. \tag{153}$$

On the other hand, by law of total variance,

$$\begin{aligned}
\mathrm{Var}(X) &= \mathbb{E}(\mathrm{Var}(X\,|\,Y)) + \mathrm{Var}(\mathbb{E}(X\,|\,Y)) && (154)\\
&= \int_0^1 ny(1-y)\,dy + \mathrm{Var}(nY) && (155)\\
&= n\left[\frac{y^2}{2} - \frac{y^3}{3}\right]_0^1 + \frac{n^2}{12} && (156)\\
&= \frac{n^2}{12} + \frac{n}{6}. && (157)
\end{aligned}$$

▲

In fact, we can figure out the entire distribution of the binomial variable with uniform rate using conditioning, not just its mean and variance.

Exercise 3.17. Let $Y \sim \mathrm{Uniform}([0,1])$ and $X \sim \mathrm{Binomial}(n, Y)$ as in Exercise 3.16.

(i) Use iterated expectation for probability to write

$$\mathbb{P}(X=k) = \binom{n}{k}\int_0^1 y^k (1-y)^{n-k}\,dy. \tag{158}$$

(ii) Write $A_{n,k} = \int_0^1 y^k(1-y)^{n-k}\,dy$. Use integration by parts and show that

$$A_{n,k} = \frac{k}{n-k+1} A_{n,k-1}. \tag{159}$$

for all $1 \le k \le n$. Conclude that for all $0 \le k \le n$,

$$A_{n,k} = \frac{1}{\binom{n}{k}} \frac{1}{n+1}. \tag{160}$$

(iii) Conclude that $X \sim \text{Uniform}(\{0, 1, \cdots, n\})$.

Exercise 3.18 (Exercise 3.8 continued)**.** Let $X_1, X_2, \cdots, X_n$ be as in Exercise 3.8.

(i) Show that $\text{Var}(X_1) = L^2/48$.
(ii) Show that $\text{Var}(X_2) = (7/12)\,\text{Var}(X_1) + (1/48)\mathbb{E}(X_1)^2$.
(iii) Show that $\text{Var}(X_{n+1}) = (7/12)\,\text{Var}(X_n) + (1/48)\mathbb{E}(X_n)^2$ for any $n \ge 1$.
(iv) Using Exercise 3.8, show the following recursion on variance holds:

$$\text{Var}(X_{n+1}) = \frac{7}{12}\text{Var}(X_n) + \frac{1}{48}\left(\frac{9}{16}\right)^n L^2. \tag{161}$$

Furthermore, compute $\text{Var}(X_2)$ and $\text{Var}(X_3)$.
(v)* Let $A_n = \left(\frac{16}{9}\right)^n \text{Var}(X_n)$. Show that A_n's satisfy

$$A_{n+1} + L^2 = \left(\frac{28}{27}\right)(A_n + L^2). \tag{162}$$

(vi)* Show that $A_n = \left(\frac{28}{27}\right)^{n-1}(A_1 + L^2) - L^2$ for all $n \ge 1$.
(vii)* Conclude that

$$\text{Var}(X_n) = \left[\left(\frac{7}{12}\right)^n - \left(\frac{9}{16}\right)^n\right] L^2. \tag{163}$$

4. Transforms of RVs

In this section, we will see how we associate a function $M_X(t)$ to each RV X and how we can understand X by looking at $M_X(t)$ instead. The advantage is that now we can use powerful tools from calculus and analysis (e.g., differentiation, integral, power series, Taylor expansion, etc.) to study RVs.

4.1. Moment generating function.

Let X be a RV. Consider a new RV $g(X) = e^{tX}$, where t is a real-valued parameter we inserted for a reason to be clear soon. A classic point of view of studying X is to look at its *moment generating function* (MGF), which is the expectation $\mathbb{E}[e^{tX}]$ of the RV e^{tX}.

Example 4.1. Let X be a discrete RV with PMF

$$\mathbb{P}(X=x) = \begin{cases} 1/2 & \text{if } x=2 \\ 1/3 & \text{if } x=3 \\ 1/6 & \text{if } x=5. \end{cases} \tag{164}$$

Its MGF is

$$\mathbb{E}[e^{tX}] = \frac{e^{2t}}{2} + \frac{e^{3t}}{3} + \frac{e^{5t}}{6}. \tag{165}$$

▲

Here is a heuristic for why we might be interested in the MGF of X. Recall the Taylor expansion of the exponential function e^s:

$$e^s = 1 + \frac{s}{1!} + \frac{s^2}{2!} + \frac{s^3}{3!} + \cdots. \tag{166}$$

Plug in $s = tX$ and get

$$e^{tX} = 1 + \frac{X}{1!}t + \frac{X^2}{2!}t^2 + \frac{X^3}{3!}t^3 + \cdots. \tag{167}$$

Taking expectation and using its 'linearity', this gives us

$$\mathbb{E}[e^{tX}] = 1 + \frac{\mathbb{E}[X]}{1!}t + \frac{\mathbb{E}[X^2]}{2!}t^2 + \frac{\mathbb{E}[X^3]}{3!}t^3 + \cdots. \tag{168}$$

Notice that the right hand side is a power series in variable t, and the kth moment $\mathbb{E}[X^k]$ of X shows up in the coefficient of the kth order term t^k. In other words, by simply taking the expectation of e^{tX}, we can get all higher moments of X. In this sense, the MGF $\mathbb{E}[e^{tX}]$ generates all moments of X, hence we call its name 'moment generating function'.

As you might have noticed, the equation (168) needs more justification. For example, what if $\mathbb{E}[X^3]$ is infinity? Also, can we really use linearity of expectation for a sum of infinitely many RVs as in the right hand side of (167)? We will get to this theoretical point later, and for now let's get ourselves more familiar to MGF computation.

Example 4.2 (Bernoulli RV)**.** Let $X \sim \text{Bernoulli}(p)$. Then

$$\mathbb{E}[e^{tX}] = e^t p + e^0(1-p) = 1 - p + e^t p. \tag{169}$$

▲

Example 4.3 (Poisson RV)**.** Let $X \sim \text{Poisson}(\lambda)$. Then using the Taylor expansion of the exponential function,

$$\mathbb{E}[e^{tX}] = \sum_{k=0}^{\infty} e^{kt}\frac{\lambda^k e^{-\lambda}}{k!} = e^{-\lambda}\sum_{k=0}^{\infty}\frac{(e^t\lambda)^k}{k!} = e^{-\lambda}e^{\lambda e^t} = e^{\lambda(e^t-1)}. \tag{170}$$

Exercise 4.4 (Geometric RV)**.** Let $X \sim \text{Geom}(p)$. Show that

$$\mathbb{E}[e^{tX}] = \frac{pe^t}{1-(1-p)e^t}. \tag{171}$$

Example 4.5 (Uniform RV)**.** Let $X \sim \text{Uniform}([a,b])$. Then

$$\mathbb{E}[e^{tX}] = \int_a^b e^{tx}\frac{1}{b-a}\,dx = \frac{1}{b-a}\left[\frac{e^{tx}}{t}\right]_a^b = \frac{e^{bt}-e^{at}}{t(b-a)}. \tag{172}$$

▲

Example 4.6 (Exponential RV)**.** Let $X \sim \text{Exp}(\lambda)$. Then

$$\mathbb{E}[e^{tX}] = \int_0^\infty e^{tx} \lambda e^{-\lambda x}\, dx = \lambda \int_0^\infty e^{(t-\lambda)x}\, dx. \tag{173}$$

Considering two cases when $t < \lambda$ and $t \geq \lambda$, we get

$$\mathbb{E}[e^{tX}] = \begin{cases} \frac{\lambda}{\lambda - t} & \text{if } t < \lambda \\ \infty & \text{if } t \geq \lambda. \end{cases} \tag{174}$$

▲

Example 4.7 (Standard normal RV)**.** Let $X \sim N(0,1)$. Then

$$\mathbb{E}[e^{tX}] = \int_{-\infty}^{\infty} e^{tx} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2/2 + tx}\, dx. \tag{175}$$

By completing square, we can write

$$-\frac{x^2}{2} + tx = -\frac{1}{2}(x^2 - 2tx) = \frac{1}{2}(x-t)^2 + \frac{t^2}{2}. \tag{176}$$

So we get

$$\mathbb{E}[e^{tX}] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(x-t)^2/2} e^{t^2/2}\, dx = e^{t^2/2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-(x-t)^2/2}\, dx. \tag{177}$$

Notice that the integrand in the last expression is the PDF of a normal RV with distribution $N(t,1)$. Hence the last integral equals 1, so we conclude

$$\mathbb{E}[e^{tX}] = e^{t^2/2}. \tag{178}$$

▲

Exercise 4.8 (MGF of linear transform)**.** Let X be a RV and a, b be constants. Let $M_X(t)$ be the MGF of X. Then show that

$$\mathbb{E}[e^{t(aX+b)}] = e^{bt} M_X(at). \tag{179}$$

Exercise 4.9 (Standard normal)**.** Let $X \sim N(\mu, \sigma^2)$ and $Z \sim N(0,1)$. Using the fact that $\mathbb{E}[e^{tZ}] = e^{t^2/2}$ and Exercise 4.9, show that

$$\mathbb{E}[e^{tY}] = e^{\sigma^2 t^2/2 + t\mu}. \tag{180}$$

4.2. **Two important theorems about MGFs.** The power series expansion (168) of MGF may not be valid in general. The following theorem gives a sufficient condition for which such an expansion is true. We omit its proof in this lecture.

Theorem 4.10. *Let X be a RV. Suppose there exists a constant $h > 0$ such that $\mathbb{E}[e^{tX}] < \infty$ for all $x \in (-h, h)$. Then the kth moment $\mathbb{E}[X^k]$ exists for all $k \geq 0$ and there exists a constant $\varepsilon > 0$ such that for all $t \in (-\varepsilon, \varepsilon)$,*

$$\mathbb{E}[e^{tX}] = \sum_{k=0}^{\infty} \frac{\mathbb{E}[X^k]}{k!} t^k. \tag{181}$$

For each RV X, we say its MGF *exists* whenever the hypothesis of the above theorem holds. One of the consequence of the above theorem is that we can access its kth moment by taking kth derivative of its MGF and evaluating at $t = 0$.

Exercise 4.11. Suppose the MGF of a RV X exists. Then show that for each integer $k \geq 0$,

$$\frac{d^k}{dt^k}\mathbb{E}[e^{tX}]\Big|_{t=0} = \mathbb{E}[X^k]. \tag{182}$$

Example 4.12 (Poisson RV)**.** Let $X \sim \text{Poisson}(\lambda)$. In Example 4.3, we have computed

$$\mathbb{E}[e^{tX}] = e^{\lambda(e^t-1)} \qquad \forall t \in \mathbb{R}. \tag{183}$$

Differentiating by t and evaluating at $t = 0$, we get

$$\mathbb{E}[X] = \frac{d}{dt}e^{\lambda(e^t-1)}\Big|_{t=0} = e^{\lambda(e^t-1)}\lambda e^t\Big|_{t=0} = \lambda. \tag{184}$$

We can also compute its second moment as

$$\mathbb{E}[X^2] = \frac{d^2}{dt^2}e^{\lambda(e^t-1)}\Big|_{t=0} = \frac{d}{dt}\lambda e^{\lambda(e^t-1)+t}\Big|_{t=0} = \lambda e^{\lambda(e^t-1)+t}(\lambda e^t+1)\Big|_{t=0} = \lambda(\lambda+1). \tag{185}$$

This also implies that

$$\text{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \lambda(\lambda+1) - \lambda^2 = \lambda. \tag{186}$$

▲

Example 4.13 (Exponential RV)**.** Let $X \sim \text{Exp}(\lambda)$. Our calculation in Example 4.6 implies that

$$\mathbb{E}[e^{tX}] = \frac{\lambda}{\lambda - t} \qquad t \in (-\lambda, \lambda). \tag{187}$$

We can compute the first and second moment of X:

$$\mathbb{E}[X] = \frac{d}{dt}\frac{\lambda}{\lambda-t}\Big|_{t=0} = \frac{\lambda}{(\lambda-t)^2}\Big|_{t=0} = \frac{1}{\lambda} \tag{188}$$

$$\mathbb{E}[X^2] = \frac{d^2}{dt^2}\frac{\lambda}{\lambda-t}\Big|_{t=0} = \frac{d}{dt}\frac{\lambda}{(\lambda-t)^2}\Big|_{t=0} = \frac{2\lambda}{(\lambda-t)^3}\Big|_{t=0} = \frac{2}{\lambda^2}. \tag{189}$$

In fact, by recognizing $\lambda/(\lambda - t)$ as a geometric series,

$$\mathbb{E}[e^{tX}] = \frac{1}{1-t/\lambda} = 1 + (t/\lambda) + (t/\lambda)^2 + (t/\lambda)^3 + \cdots \tag{190}$$

$$= 1 + \frac{1!/\lambda}{1!}t + \frac{2!/\lambda^2}{2!}t^2 + \frac{3!/\lambda^3}{3!}t^3 + \cdots. \tag{191}$$

Hence by comparing with (181), we conclude that $\mathbb{E}[X^k] = k!/\lambda^k$ for all $k \geq 0$. ▲

The second theorem for MGFs is that they determine the distribution of RVs. This will be critically used later in the proof of the central limit theorem.

Theorem 4.14. *Let X, Y, and X_n for $n \geq 1$ be RVs whose MGFs exist.*

(i) *(Uniqueness) Suppose $\mathbb{E}[e^{tX}] = \mathbb{E}[e^{tY}]$ for all sufficiently small t. Then $\mathbb{P}(X \leq s) = \mathbb{P}(Y \leq s)$ for all $s \in \mathbb{R}$.*

(ii) *(Continuity) Suppose $\lim_{n\to\infty}\mathbb{E}[e^{tX_n}] = \mathbb{E}[e^{tX}]$ for all sufficiently small t and that $\mathbb{E}[e^{tX}]$ is continuous at $t = 0$. Then $\mathbb{P}(X_n \leq s) \to \mathbb{P}(X \leq s)$ for all s such that $\mathbb{P}(X \leq x)$ is continuous at $x = s$.*

4.3. **MGF of sum of independent RVs.** One of the nice properties of MGFs is the following factorization for sums of independent RVs.

Proposition 4.15. *Let X, Y be independent RVs. Then*

$$\mathbb{E}[e^{t(X+Y)}] = \mathbb{E}[e^{tX}]\mathbb{E}[e^{tY}]. \tag{192}$$

If you believe that the RVs e^{tX} and e^{tY} are independent, then the proof of the above result is one-line:

$$\mathbb{E}[e^{t(X+Y)}] = \mathbb{E}[e^{tX}e^{tY}] = \mathbb{E}[e^{tX}]\mathbb{E}[e^{tY}]. \tag{193}$$

In general, it is a special case of the following result.

Proposition 4.16. *Let X, Y be independent RVs. Then for any integrable functions $g_1, g_2 : \mathbb{R} \to \mathbb{R}$, we have*

$$\mathbb{E}[g_1(X)g_2(Y)] = \mathbb{E}[g_1(X)]\mathbb{E}[g_2(Y)]. \tag{194}$$

Proof. If X, Y are continuous RVs,

$$\begin{aligned}
\mathbb{E}[g_1(X)g_2(Y)] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g_1(x)g_2(y)f_{X,Y}(x,y)\,dx\,dy && (195)\\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g_1(x)g_2(y)f_X(x)f_Y(y)\,dx\,dy && (196)\\
&= \int_{-\infty}^{\infty} g_1(x)f_X(x)\left(\int_{-\infty}^{\infty} g_2(y)f_Y(y)\right)dy && (197)\\
&= \mathbb{E}[g_2(Y)]\int_{-\infty}^{\infty} g_1(x)f_X(x)\,dx && (198)\\
&= \mathbb{E}[g_1(X)]\mathbb{E}[g_2(Y)]. && (199)
\end{aligned}$$

For discrete RVs, use summation and PMF instead of integral and PDF. □

Exercise 4.17 (Binomial RV)**.** Let $X \sim \text{Binomial}(n, p)$. Use the MGF of Bernoulli RV and Proposition 4.15 to show that

$$\mathbb{E}[e^{tX}] = (1 - p + e^t p)^n. \tag{200}$$

Example 4.18 (Sum of independent Poisson RVs)**.** Let $X_1 \sim \text{Poisson}(\lambda_1)$ and $X_2 \sim \text{Poisson}(\lambda_2)$ be independent Poisson RVs. Let $Y = X_1 + X_2$. Using Exercise 4.3, we have

$$\mathbb{E}[e^{tY}] = \mathbb{E}[e^{tX_1}]\mathbb{E}[e^{tX_2}] = e^{(\lambda_1+\lambda_2)(e^t-1)}. \tag{201}$$

Notice that the last expression is the MGF of a Poisson RV with rate $\lambda_1 + \lambda_2$. By the Uniqueness of MGF (Theorem 4.14 (i)), we conclude that $Y \sim \text{Poisson}(\lambda_1 + \lambda_2)$. ▲

Exercise 4.19 (Sum of independent normal RVs)**.** Let $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$ be independent normal RVs.

(i) Show that $\mathbb{E}[e^{t(X_1+X_2)}] = \exp[(\sigma_1^2 + \sigma_2^2)t^2/2 + t(\mu_1 + \mu_2)]$.

(ii) Conclude that $X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

4.4. **Sum of random number of independent RVs.** Suppose $X_1, X_2, \cdots$ are independent and identically distributed (i.i.d.) RVs and let N be another independent RV taking values in nonnegative integers (e.g., Binomial). For a new RV Y by

$$Y = X_1 + X_2 + \cdots + X_N. \tag{202}$$

Note that we are summing a random number of X_i's, so there are two sources of randomness that determines Y. As usual, we use conditioning to study such RVs. For instance,

$$\mathbb{E}[Y \mid N = n] = \mathbb{E}[X_1 + \cdots + X_n] = \mathbb{E}[X_1] + \cdots + \mathbb{E}[X_n] = n\mathbb{E}X_1 \tag{203}$$

$$\mathrm{Var}(Y \mid N = n) = \mathrm{Var}(X_1 + \cdots + X_n) = \mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n) = n\,\mathrm{Var}(X_1). \tag{204}$$

Hence iterated expectation gives

$$\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y|N]] = \mathbb{E}[N\mathbb{E}[X_1]] = \mathbb{E}[X_1]\mathbb{E}[N]. \tag{205}$$

On other other hand, law of total variance gives

$$\mathrm{Var}(Y) = \mathbb{E}[\mathrm{Var}(Y \mid N)] + \mathrm{Var}(\mathbb{E}[Y \mid N]) \tag{206}$$

$$= \mathbb{E}[N\,\mathrm{Var}(X_1)] + \mathrm{Var}(N\mathbb{E}[X_1]) \tag{207}$$

$$= \mathrm{Var}(X_1)\mathbb{E}[N] + \mathbb{E}[X_1]^2\,\mathrm{Var}(N). \tag{208}$$

Furthermore, can we also figure out the MGF of Y? After all, MGF is an expectation so we can also get it by iterated expectation. First we compute the conditional version. Denoting $M_X(t) = \mathbb{E}[e^{tX}]$,

$$\mathbb{E}[e^{tY} \mid N = n] = \mathbb{E}[e^{t(X_1+\cdots+X_n)}] = \mathbb{E}[e^{tX_1} \cdots e^{tX_n}] \tag{209}$$

$$= \mathbb{E}[e^{tX_1}] \cdots \mathbb{E}[e^{tX_n}] = \mathbb{E}[e^{tX_1}]^n = M_{X_1}(t)^n \tag{210}$$

$$= e^{n \log M_{X_1}(t)}. \tag{211}$$

The last line is the trick here. Now the iterated expectation gives

$$\mathbb{E}[e^{tY}] = \mathbb{E}[\mathbb{E}[e^{tY} \mid N]] = \mathbb{E}[e^{(\log M_{X_1}(t))N}]. \tag{212}$$

Note that the last expression is nothing but the MFG of N evaluated at $\log M_{X_1}(t)$ instead of t. Hence

$$\mathbb{E}[e^{tY}] = M_N(\log M_{X_1}(t)). \tag{213}$$

Let us summarize what have obtained so far.

Proposition 4.20. *Let $X_1, X_2, \cdots$ be i.i.d. RVs and let N be another independent RV which takes values from nonnegative integers. Let $Y = \sum_{k=0}^N X_k$. Denote the MGF of any RV Z by $M_Z(t)$. Then we have*

$$\mathbb{E}[Y] = \mathbb{E}[N]\mathbb{E}[X_[1]] \tag{214}$$

$$\mathrm{Var}[Y] = \mathrm{Var}(X_1)\mathbb{E}[N] + \mathbb{E}[X_1]^2\,\mathrm{Var}(N) \tag{215}$$

$$M_Y(t) = M_N(\log M_{X_1}(t)). \tag{216}$$

Example 4.21. Let $X_i \sim \mathrm{Exp}(\lambda)$ for $i \geq 0$ and let $N \sim \mathrm{Poisson}(\lambda)$. Suppose all RVs are independent. Define $Y = \sum_{k=1}^N X_i$. Then

$$\mathbb{E}[Y] = \mathbb{E}[N]\mathbb{E}[X_1] = \lambda/\lambda = 1, \tag{217}$$

$$\mathrm{Var}(Y) = \mathrm{Var}(X_1)\mathbb{E}[N] + \mathbb{E}[X_1]^2\,\mathrm{Var}(N) = \frac{\lambda}{\lambda^2} + \frac{\lambda}{\lambda^2} = \frac{2}{\lambda}. \tag{218}$$

On other hand, recall that $M_{X_1}(t) = \frac{\lambda}{\lambda - t}$ and $M_N(t) = e^{\lambda(e^t - 1)}$. Hence

$$\mathbb{E}[e^{tY}] = e^{\lambda(\exp(\log\frac{\lambda}{\lambda-t})-1)} = e^{\lambda(\frac{\lambda}{\lambda-t}-1)} = e^{\frac{\lambda t}{\lambda - t}}. \tag{219}$$

So we know everything about Y. Knowing the MGF of Y, we could get all the moments of Y. For instance,

$$\mathbb{E}[Y] = \frac{d}{dt} e^{\frac{\lambda t}{\lambda - t}} \Big|_{t=0} = e^{\frac{\lambda t}{\lambda - t}} \frac{\lambda(\lambda - t) + \lambda t}{(\lambda - t)^2} \Big|_{t=0} = 1. \tag{220}$$

▲

Exercise 4.22. Let $X_1, X_2, \cdots$ be i.i.d. RVs and let N be another independent RV which takes values from nonnegative integers. Let $Y = \sum_{k=0}^{N} X_k$. Denote the MGF of any RV Z by $M_Z(t)$. Using the fact that $M_Y(t) = M_N(\log M_{X_1}(t))$, derive

$$\mathbb{E}[Y] = \mathbb{E}[N]\mathbb{E}[X_1], \tag{221}$$

$$\mathrm{Var}[Y] = \mathrm{Var}(X_1)\mathbb{E}[N] + \mathbb{E}[X_1]^2 \mathrm{Var}(N). \tag{222}$$

Example 4.23. Let $X_i \sim \mathrm{Exp}(\lambda)$ for $i \geq 0$ and let $N \sim \mathrm{Geom}(p)$. Let $Y = \sum_{k=1}^{N} X_k$. Suppose all RVs are independent. Recall that

$$M_{X_1}(t) = \frac{\lambda}{\lambda - t}, \qquad M_N(t) = \frac{pe^t}{1 - (1-p)e^t}. \tag{223}$$

Hence

$$M_Y(t) = \frac{p\frac{\lambda}{\lambda - t}}{1 - (1-p)\frac{\lambda}{\lambda - t}} = \frac{p\lambda}{(\lambda - t) - \lambda(1-p)} = \frac{p\lambda}{p\lambda - t}. \tag{224}$$

Notice that this is the MGF of an $\mathrm{Exp}(p\lambda)$ variable. Thus by uniqueness, we conclude that $Y \sim \mathrm{Exp}(p\lambda)$. If you remember, sum of k independent $\mathrm{Exp}(\lambda)$ RVs were not an exponential RV (its distribution is $\mathrm{Erlang}(k, \lambda)$. See Exercise 1.19 in Note 1). But as we have seen in this example, if you sum a random number of independent exponentials, they could by exponential again. ▲

Exercise 4.24. Let $X_i \sim \mathrm{Geom}(q)$ for $i \geq 0$ and let $N \sim \mathrm{Geom}(p)$. Suppose all RVs are independent. Let $Y = \sum_{k=0}^{N} X_k$.

(i) Show that the MGF of Y is given by

$$\mathbb{E}[e^{tY}] = \frac{pqe^t}{1 - (1 - pq)e^t}. \tag{225}$$

(ii) Conclude that $Y \sim \mathrm{Geom}(pq)$.