Lecture 2 Random Variables [*]

Tiejun Li

1 A Crash Course on Basic Concepts

1.1 Discrete Examples

We will concentrate on the elementary and intuitive aspects of probability here. In the discrete case, the function $P(X)$ is called the probability mass function (pmf).

- Bernoulli distribution:
$$P(X) = \begin{cases} p, & X = 1, \\ q, & X = 0. \end{cases}$$
where $p > 0, q > 0, p + q = 1$. The mean and variance are
$$\mathbb{E}X = p, \mathrm{Var}(X) = pq.$$
If $p = q = \frac{1}{2}$, it is the well-known fair-coin tossing game.

- Binomial distribution $B(n, p)$:

  n independent experiments of Bernoulli distribution X_k, $X := X_1 + \ldots + X_n$, then
$$P(X = k) = C_n^k p^k q^{n-k}.$$
The mean and variance are
$$\mathbb{E}X = np, \mathrm{Var}(X) = npq.$$

- Multinomial distribution $M(p_1, \ldots, p_r)$:

  Multinomial distribution is a simple generalization of binomial distribution, in which each trial results in exactly one of some fixed number r possible outcomes with probability $p_1, p_2, \ldots, p_r$, where
$$\sum_{i=1}^{r} p_i = 1, \quad 0 \le p_i \le 1, \; i = 1, \ldots, r,$$

[*] School of Mathematical Sciences, Peking University, Beijing 100871, P.R. China

and we have n independent trials. Let the random variables X_i indicate the number of times the i-th outcome was observed over the n trials. $X = (X_1, \ldots, X_r)$ follows a multinomial distribution with parameters n and p, where $p = (p_1, \ldots, p_r)$.

The pmf of the multinomial distribution is:

$$P(X_1 = x_1, \ldots, X_r = x_r) = \frac{n!}{x_1! \cdots x_r!} p_1^{x_1} \cdots p_r^{x_r}, \quad n = x_1 + \cdots + x_r.$$

The mean, variance and covariance are

$$\mathbb{E}(X_i) = np_i, \quad \mathrm{Var}(X_i) = np_i(1 - p_i), \quad \mathrm{Cov}(X_i, X_j) = -np_ip_j \ (i \neq j).$$

- Poisson distribution:

  The number X of radiated particles in a fixed time τ obeys

  $$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda},$$

  where λ is the average number of radiated particles each time. The mean and variance are

  $$\mathbb{E}X = \lambda, \mathrm{Var}(X) = \lambda.$$

  Poisson distribution may be viewed as the limit of binomial distribution (the law of rare events)

  $$C_n^k p^k q^{n-k} \longrightarrow \frac{\lambda^k}{k!} e^{-\lambda} \quad (n \to \infty, np = \lambda).$$

  Poisson distribution can also describe the spatial distribution of randomly scattered points. For example, Let A be a set in R^2. $X_A(\omega)$ be the number of points in A. If the points are uniformly distributed on the plane, and suppose the scattering density is λ (mean number of points per area), then X_A has Poisson distribution

  then

  $$\lambda = \text{ area of } A \times \text{ number of points/area.}$$

  X_A has Poisson distribution

  $$P(X_A = n) = \frac{(\lambda \cdot \mathrm{meas}(A))^n}{n!} e^{-\lambda \cdot \mathrm{meas}(A)}.$$

- Geometric probability.

  Probability = Ratio of areas

  Special case of continuous examples — uniform distribution.

Example 1. *Maxwell-Boltzmann, Bose-Einstein, Fermi-Dirac statistics.*

Suppose there are n particles and N bins, where $N > n$.

1. *Given n bins, what is the probability that each bin has one particle? (Boson)*
2. *What is the probability that there exist n bins such that each bin has exactly one particle? (Fermion, Pauli exclusion principle)*

In statistical physics the classical particles are distinguishable. If they satisfy the Pauli exclusion principle, then they are subject to Maxwell-Boltzmann statistis. The quantum particles are indistinguishable. If they satisfy the Pauli exclusion principle, then they are subject to Fermi-Dirac statistis (Fermions). If they do not satisfy the Pauli exclusion principle, then they are subject to Bose-Einstein statistis (Bosons). Distinguishable particles that are subject to the exclusion principle do not occur in physics.

The whole picture is as follows:

	Distinguishable balls (classical)	*Undistinguishable balls (quantum)*
Without exclusion	N^n *(Maxwell-Boltzmann)*	C^n_{N+n-1} *(Bose-Einstein)*
With exclusion	P^n_N	C^n_N *(Fermi-Dirac)*

1.2 Continuous Examples

In continuous case, the function $p(x)$ is called the probability density function (pdf).

- Uniform distribution $\mathcal{U}[0,1]$:

$$p(x) = \begin{cases} 1 & \text{if } x \in [0,1] \\ 0 & \text{otherwise} \end{cases}$$

The mean and variance are

$$\mathbb{E}X = \frac{1}{2}, \mathrm{Var}(X) = \frac{1}{12}.$$

- Exponential distribution:$(\lambda > 0)$

$$p(x) = \begin{cases} 0 & \text{if } x < 0 \\ \lambda e^{-\lambda x} & \text{if } x \geq 0 \end{cases}$$

The mean and variance are

$$\mathbb{E}X = \frac{1}{\lambda}, \mathrm{Var}(X) = \frac{1}{\lambda^2}.$$

Waiting time for continuous time Markov process also has exponential distribution, where λ is the rate of the process.

- Normal distribution(Gaussian distribution)($N(0,1)$):

  $$p(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$$

  or more generally $N(\mu,\sigma)$

  $$p(x)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

  where μ is the mean (expectation), σ^2 is the variance.

  High dimensional case ($N(\mu,\Sigma)$)

  $$p(x)=\frac{1}{(2\pi)^{n/2}(\det\Sigma)^{1/2}}e^{-(\boldsymbol{X}-\mu)^T\Sigma^{-1}(\boldsymbol{X}-\mu)}$$

  where μ is the mean, Σ is a symmetric positive definite matrix, which is the covariance matrix of $\boldsymbol{X}$. $\det\Sigma$ is the determinant of Σ. More general high dimensional normal distribution is defined with characteristic functions $g(t)=\exp\Big(i\mu\cdot\boldsymbol{t}-\frac{1}{2}\boldsymbol{t}'\Sigma\boldsymbol{t}\Big)$.

  Remark 1. *In 1D case, the normal distribution $N(np,npq)$ may be viewed as the limit of the Binomial distribution $B(n,p)$ when n is large. This is the famous De Moivre-Laplace limit theorem. It is a special case of the central limit theorem (CLT). Notice that*

  $$\frac{B(n,p)-np}{\sqrt{npq}}\longrightarrow N(0,1)\ \textit{as}\ n\to\infty.$$

  Remark 2. *In 1D case, the normal distribution $N(\lambda,\lambda)$ may be viewed as the limit of the Poisson distribution Poisson(λ) when λ is large. Notice the simple fact that the sum of two independent Poisson(λ) and Poisson(μ) is Poisson$(\lambda+\mu)$ (why?), we can decompose Poisson(λ) into the sum of n i.i.d. Poisson(λ/n), we have*

  $$\frac{Poisson(\lambda)-\lambda}{\sqrt{\lambda}}\longrightarrow N(0,1)\ \textit{when}\ \lambda\ \textit{is large.}$$

1.3 Probability Space

- σ-algebra $\mathcal{F}$

  $\mathcal{F}$ is a collection of subsets of Ω:

  1. $\Omega\in\mathcal{F}$;
  2. If $A\in\mathcal{F}$, then $\bar{A}=\Omega\backslash A\in\mathcal{F}$;
  3. If $A_1,A_2,\cdots,A_n,\cdots\in\mathcal{F}$, then $\bigcup_{j=1}^{\infty}A_j\in\mathcal{F}$.

  Here $(\Omega,\mathcal{F})$ is called a measurable space.

- Probability measure P

  1. (Positive) $\forall A \in \mathcal{F}, P(A) \geq 0$;
  2. (Countably additive) If $A_1, A_2, \cdots \in \mathcal{F}$, and they are disjoint, then $P(\bigcup_{j=1}^{\infty} A_j) = \sum_{j=1}^{\infty} P(A_j)$;
  3. (Normalization) $P(\Omega) = 1$.

- Probability space — Triplet $(\Omega, \mathcal{F}, P)$

  1. Random variable: a measurable function $X : \Omega \to R$.
  2. Distribution(or law): a probability measure μ on R defined for any set $B \subset R$ by
  $$\mu(B) = \text{Prob}(X \in B) = P\{\omega \in \Omega : X(\omega) \in B\}.$$
  3. Probability density function(pdf): an integrable function $p(x)$ on R such that for any set $B \subset R$,
  $$\mu(B) = \int_B p(x)dx.$$
  4. Mean (expectation):
  $$\mathbb{E}f(X) = \int_\Omega f(X(\omega))P(d\omega) = \int_R f(x)d\mu(x) = \int_R f(x)p(x)dx.$$
  5. Variance:
  $$\text{Var}(X) = \mathbb{E}(X - \mathbb{E}X)^2 = \mathbb{E}X^2 - (\mathbb{E}X)^2.$$
  6. p-th moment: $\mathbb{E}|X|^p$.
  7. Covariance:
  $$\text{Cov}(X, Y) = \mathbb{E}(X - \mathbb{E}X)(Y - \mathbb{E}Y).$$
  8. Independence:
  $$\mathbb{E}f(X)g(Y) = \mathbb{E}f(X)\mathbb{E}g(Y).$$
  for all continuous functions f and g.

1.4 Notions of Convergence

Probability space $(\Omega, \mathcal{F}, P)$, $\{X_n\}$ — a sequence of random variables, μ_n — the distribution of X_n. X — another random variable with distribution μ.

Definition 1. *(Almost sure convergence) X_n converges to X almost surely as $n \to \infty$, ($X_n \to X$, a.s.) if*
$$P\{\omega \in \Omega, \quad X_n(\omega) \to X(\omega)\} = 1$$

Definition 2. *(Convergence in probability) X_n converges to X in probability if for any $\epsilon > 0$,*

$$P\{\omega | X_n(\omega) - X(\omega)| > \epsilon\} \to 0$$

as $n \to +\infty$.

Definition 3. *(Convergence in distribution) X_n converges to X in distribution ($X_n \xrightarrow{d} X$) (i.e. $\mu_n \rightharpoonup \mu$ or $\mu_n \xrightarrow{d} \mu$, weak convergence), if for any bounded continuous function f*

$$\mathbb{E}f(X_n) \to \mathbb{E}f(X)$$

Definition 4. *(Convergence in L^p) If $X_n, X \in L^p$, and*

$$\mathbb{E}|X_n - X|^p \to 0.$$

If $p = 1$, that is convergence in mean; if $p = 2$, that is convergence in mean square.

Relation:

$$\begin{tikzcd}
\text{Almost sure convergence} \arrow[r, harpoon, shift left] & \text{Converge in probability} \arrow[l, harpoon, shift left, "\text{subsequence}"] \arrow[r] & \text{Converge in distribution} \\
& L^p \text{ convergence} \arrow[u, Rightarrow] &
\end{tikzcd}$$

1.5 Conditional Expectation

Let X and Y be two discrete random variables with joint probability

$$p(i, j) = \mathbb{P}(X = i, Y = j).$$

The *conditional probability* that $X = i$ given that $Y = j$ is given by

$$p(i|j) = \frac{p(i, j)}{\sum_i p(i, j)} = \frac{p(i, j)}{\mathbb{P}(Y = j)}$$

if $\sum_i p(i, j) > 0$ and conventionaly taken to be zero if $\sum_i p(i, j) = 0$. The natural definition of the *conditional expectation* of $f(X)$ given that $Y = j$ is

$$\mathbb{E}(f(X)|Y = j) = \sum_i f(i)p(i|j). \tag{1}$$

The axiomatic definition of the conditional expectation $Z = E(X|\mathcal{G})$ is defined with respect to a sub-σ-algebra $\mathcal{G} \subset \mathcal{F}$ as follows.

Definition 5 (Conditional expectation)**.** *For any random variable X with $\mathbb{E}|X| < \infty$, the condition expectation Z of X given $\mathcal{G}$ is defined as* *(i) Z is a random variable which is measurable with respect to $\mathcal{G}$;*

(ii) for any set $A \in \mathcal{G}$,

$$\int_A Z(\omega)\mathbb{P}(d\omega) = \int_A X(\omega)\mathbb{P}(d\omega).$$

The existence of $Z = E(X|\mathcal{G})$ comes from the Radon-Nikodym theorem by considering the measure μ on $\mathcal{G}$ defined by $\mu(A) = \int_A X(\omega)\mathbb{P}(d\omega)$ (see [3]). One can easily find that μ is absolutely continuous with respect to the measure $\mathbb{P}|_{\mathcal{G}}$, the probability measure confined in $\mathcal{G}$. Thus Z exists and is unique up to the almost sure equivalence in $\mathbb{P}|_{\mathcal{G}}$.

Theorem 1 (Properties of conditional expectation)**.** *Suppose X, Y are random variables with $\mathbb{E}|X|, \mathbb{E}|Y| < \infty$, $a, b \in \mathbb{R}$. Then*

(i) $\mathbb{E}(aX + bY|\mathcal{G}) = a\mathbb{E}(X|\mathcal{G}) + b\mathbb{E}(Y|\mathcal{G})$

(ii) $\mathbb{E}(\mathbb{E}(X|\mathcal{G})) = \mathbb{E}(X)$

(iii) $\mathbb{E}(X|\mathcal{G}) = X$, if X is $\mathcal{G}$-measurable

(iv) $\mathbb{E}(X|\mathcal{G}) = \mathbb{E}X$, if X is independent of $\mathcal{G}$

(v) $\mathbb{E}(XY|\mathcal{G}) = Y\mathbb{E}(X|\mathcal{G})$, if Y is $\mathcal{G}$-measurable

(vi) $\mathbb{E}(X|\mathcal{G}) = \mathbb{E}(\mathbb{E}(X|\mathcal{G})|\mathcal{H})$ for the sub-σ-algebras $\mathcal{G} \subset \mathcal{H}$.

Lemma 1 (Conditional Jensen's inequality)**.** *Let X be a random variable such that $\mathbb{E}|X| < \infty$ and $\phi : \mathbb{R} \to \mathbb{R}$ is a convex function such that $\mathbb{E}|\phi(X)| < \infty$. Then*

$$\mathbb{E}(\phi(X)|\mathcal{G}) \geq \phi(\mathbb{E}(X|\mathcal{G})). \tag{2}$$

The readers may be referred to [4] for the details of the proof.

For the conditional expectation of a random variable X with respect to another random variable Y, it is natural to define it as

$$\mathbb{E}(X|Y) := \mathbb{E}(X|\mathcal{G}) \tag{3}$$

where $\mathcal{G}$ is the σ-algebra $Y^{-1}(\mathcal{B})$ generated by Y.

To realize the equivalence between the abstract definition (3) and (1) when Y only takes finitely discrete values, we suppose the following decomposition

$$\Omega = \bigcup_{j=1}^{n} \Omega_j$$

and $\Omega_j = \{\omega : Y(\omega) = j\}$. Then the σ-algebra $\mathcal{G}$ is simply the sets of all possible unions of Ω_j. The measurability of conditional expectation $\mathbb{E}(X|Y)$ with respect to $\mathcal{G}$ means $E(X|Y)$

takes constant on each Ω_j, which exactly corresponds to $E(X|Y=j)$ as we will see. By definition, we have

$$\int_{\Omega_j} \mathbb{E}(X|Y)\mathbb{P}(d\omega) = \int_{\Omega_j} X(\omega)\mathbb{P}(d\omega) \tag{4}$$

which implies

$$\mathbb{E}(X|Y) = \frac{1}{\mathbb{P}(\Omega_j)} \int_{\Omega_j} X(\omega)\mathbb{P}(d\omega). \tag{5}$$

This is exactly $\mathbb{E}(X|Y=j)$ in (1) when $f(X)=X$ and X also takes discrete values.

The conditional expectation has the following important property as the optimal approximation in L^2 norm among all of the Y-measurable functions.

Proposition 1. *Let $g(Y)$ be any measurable function of Y, then*

$$\mathbb{E}(X - \mathbb{E}(X|Y))^2 \le \mathbb{E}(X - g(Y))^2. \tag{6}$$

Proof. We have

$$\begin{aligned}\mathbb{E}(X-g(Y))^2 = &\,\mathbb{E}(X-E(X|Y))^2 + \mathbb{E}(E(X|Y)-g(Y))^2 \\ &+ 2\mathbb{E}\Big[(X-E(X|Y)(E(X|Y)-g(Y))\Big].\end{aligned}$$

and

$$\begin{aligned}&\mathbb{E}\Big[(X-\mathbb{E}(X|Y)(\mathbb{E}(X|Y)-g(Y))\Big] \\ =&\mathbb{E}\Big[\mathbb{E}\big[(X-\mathbb{E}(X|Y)(E(X|Y)-g(Y))|Y\big]\Big] \\ =&\mathbb{E}\Big[(\mathbb{E}(X|Y)-\mathbb{E}(X|Y))(E(X|Y)-g(Y))\Big]=0\end{aligned}$$

by properties (ii),(iii) and (v) in Theorem 1. The proof is done. □

2 Characteristic Function

The *characteristic function* of a random variable X or its distribution μ is defined as

$$f(\xi) = \mathbb{E}e^{i\xi X} = \int_{\mathbb{R}} e^{i\xi x}\mu(dx). \tag{7}$$

Proposition 2. *The characteristic function has the following properties:*

1. $\forall \xi \in \mathbb{R}$, $|f(\xi)| \le 1$, $f(\xi) = \overline{f(-\xi)}$, $f(0)=1$;

2. f *is uniformly continuous on* $\mathbb{R}$;

3. $f^{(n)}(0) = i^n \mathbb{E}X^n$ *provided* $\mathbb{E}|X|^n < \infty$.

Proof. The proof of statements 1 and 3 are straightforward. The second statement is valid by

$$\begin{aligned}|f(\xi_1) - f(\xi_2)| = |\mathbb{E}(e^{i\xi_1 X} - e^{i\xi_2 X})| = |\mathbb{E}(e^{i\xi_1 X}(1 - e^{i(\xi_2-\xi_1)X}))| \\ \leq \mathbb{E}|1 - e^{i(\xi_2-\xi_1)X}|.\end{aligned}$$

Dominated convergence theorem concludes the proof. □

Example 2. *The characteristic functions of some typical distributions are as below.*

1. *Bernoulli distribution:* $f(\xi) = q + pe^{i\xi}$.
2. *Binomial distribution* $B(n,p)$*:* $f(\xi) = (q + pe^{i\xi})^n$.
3. *Poisson distribtion* $\mathcal{P}(\lambda)$*:* $f(\xi) = e^{\lambda(e^{i\xi}-1)}$.
4. *Exponential distribution* $\mathcal{E}xp(\lambda)$*:* $f(\xi) = (1 - \lambda^{-1}i\xi)^{-1}$.
5. *Normal distribution* $N(\mu, \sigma^2)$*:* $f(\xi) = \exp\left(i\mu\xi - \frac{\sigma^2\xi^2}{2}\right)$.

The following important theorem gives an explicit characterization of the weak convergence of probability measures based on their characteristic functions, which is a key in proving central limit theorem later.

Theorem 2 (Lévy's continuity theorem)**.** *Let* $\{\mu_n\}_{n\in\mathbb{N}}$ *be a sequence of probability measures, and* $\{f_n\}_{n\in\mathbb{N}}$ *be their corresponding characteristic functions. Assume that*

1. f_n *converges everywhere on* $\mathbb{R}$ *to a limiting function* f*.*
2. f *is continuous at* $\xi = 0$*.*

Then there exists a probability distribution μ *such that* $\mu_u \xrightarrow{d} \mu$*. Moreover* f *is the characteristic function of* μ*.*

Conversely, if $\mu_n \xrightarrow{d} \mu$*, where* μ *is some probability distribution then* f_n *converges to* f *uniformly in every finite interval, where* f *is the characteristic function of* μ*.*

For a proof, see [4].

As in Fourier transforms, one can also define the inverse transform

$$\rho(x) = \frac{1}{2\pi}\int_{\mathbb{R}} e^{-i\xi x} f(\xi) d\xi.$$

An interesting question arises as to when this gives the density of a probability measure. To answer this we define

Definition 6. *A function f is called positive semi-definite if for any finite set of values $\{\xi_1, \dots, \xi_n\}$, $n \in \mathbb{N}$, the matrix $(f(\xi_i - \xi_j))_{i,j=1}^n$ is positive semi-definite, i.e.*

$$\sum_{i,j} f(\xi_i - \xi_j) v_i \bar{v}_j \geq 0, \tag{8}$$

for any $v_1, \dots, v_n \in \mathbb{C}$.

Theorem 3 (Bochner's Theorem). *A function f is the characteristic function of a probability measure if and only if it is a positive semi-definite and continuous at 0 with $f(0) = 1$.*

Proof. We only gives the necessity part. Suppose f is a characteristic function, then

$$\sum_{i,j=1}^n f(\xi_i - \xi_j) v_i \bar{v}_j = \int_{\mathbb{R}} \Big| \sum_{i=1}^n v_i e^{i\xi_i x} \Big|^2 \mu(dx) \geq 0. \tag{9}$$

The sufficiency part is difficult and the readers may be referred to [4]. □

3 Generating function

For discrete R.V. taking integer values, the generating function has the central importance

$$G(x) = \sum_{k=0}^{\infty} P(k) x^k.$$

One immediately has the formula:

$$P(k) = \frac{1}{k!} G^{(k)}(x) \Big|_{x=0}.$$

Definition 7. *Define the convolution of two sequences $\{a_k\}$, $\{b_k\}$ as $\{c_k\} = \{a_k\} * \{b_k\}$, the components are defined as*

$$c_k = \sum_{j=0}^{k} a_j b_{k-j}.$$

Theorem 4. *Consider two independent R.V. X and Y with PMF*

$$P(X = j) = a_j, \quad P(Y = k) = b_k$$

and $\{c_k\} = \{a_k\} * \{b_k\}$. Suppose the generating functions are $A(x)$, $B(x)$ and $C(x)$, respectively, then the generating function of $X + Y$ is $C(x)$.

Some generating functions:

- Bernoulli distribution: $G(x) = q + px$.
- Binomial distribution: $G(x) = (q + px)^n$.
- Poisson distribution: $G(x) = e^{-\lambda + \lambda x}$.

4 Moment Generating Function and Cumulants

The moment generating function of a random variable X is defined for all values of t by

$$M(t) = \mathbb{E}e^{tX} = \begin{cases} \displaystyle\sum_x p(x)e^{tx}, & X \text{ is discrete-valued} \\ \displaystyle\int_{\mathbb{R}} p(x)e^{tx}dx, & X \text{ is continuous} \end{cases} \tag{10}$$

provided that e^{tX} is integrable. It is obvious $M(0) = 1$.

Once $M(t)$ can be defined, one can show $M(t) \in C^\infty$ in its domain and its relation to the nth moments

$$M^{(n)}(t) = \mathbb{E}(X^n e^{tX}) \text{ and } \mu_n := \mathbb{E}X^n = M^{(n)}(0),\ n \in \mathbb{N}. \tag{11}$$

This gives

$$M(t) = \sum_{n=0}^{\infty} \mu_n \frac{t^n}{n!}, \tag{12}$$

which tells why $M(t)$ is called the moment generating function.

Theorem 5. *Denote $M_X(t), M_Y(t)$ and $M_{X+Y}(t)$ the moment generating functions of random variables X, Y and $X+Y$, respectively. If X, Y are independent, we have*

$$M_{X+Y}(t) = M_X(t)M_Y(t). \tag{13}$$

The proof is straightforward.

The following moment generating functions of typical random variables can be obtained by direct calculations.

(a) Binomial distribution: $M(t) = (pe^t + 1 - p)^n$.

(b) Poisson distribution: $M(t) = \exp[\lambda(e^t - 1)]$.

(c) Exponential distribution: $M(t) = \lambda/(\lambda - t)$ for $t < \lambda$.

(d) Normal distribution $N(\mu, \sigma^2)$: $M(t) = \exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right)$.

The cumulant generating function $K(t)$ is defined based on $M(t)$ by

$$K(t) = \ln M(t) = \ln \mathbb{E}e^{tX} = \sum_{n=1}^{\infty} \kappa_n \frac{t^n}{n!}. \tag{14}$$

With such definition, we have the cumulants $\kappa_0 = 0$ and

$$\kappa_n = K^{(n)}(0), \quad n \in \mathbb{N}. \tag{15}$$

The moment generating function is not so powerful as the characteristic function since the integrable condition is usually too strong for many random variables. Under similar consideration, we can also define another type of cumulant generating function $H(t)$ as

$$H(t) = \ln \mathbb{E} e^{itX} = \sum_{n=1}^{\infty} \kappa_n \frac{(it)^n}{n!}.$$

All of the definitions above can be extended to random vectors without difficulty. In this circumstance, we have

$$M(\boldsymbol{t}) = \mathbb{E} e^{\boldsymbol{t}\cdot\boldsymbol{X}}, \quad \boldsymbol{t} \in \mathbb{R}^d$$

and correspondingly the moments

$$\mu_{\boldsymbol{k}} = \mathbb{E}(X_1^{k_1} \cdots X_d^{k_d}) = \frac{\partial^{|\boldsymbol{k}|} M}{\partial t_1^{k_1} \cdots \partial t_d^{k_d}}(\boldsymbol{0}), \quad \boldsymbol{k} = (k_1, \dots, k_d) \in \mathbb{N}^d,$$

where $|\boldsymbol{k}| := \sum_{j=1}^d k_j$ is the order of multi-index $\boldsymbol{k}$. The relation between $M(\boldsymbol{t})$ and $\mu_{\boldsymbol{k}}$ is simply

$$M(\boldsymbol{t}) = \sum_{k_1=0}^{\infty} \cdots \sum_{k_d=0}^{\infty} \mu_{\boldsymbol{k}} \frac{t_1^{k_1} \cdots t_d^{k_d}}{k_1! \cdots k_d!}. \tag{16}$$

The $K(\boldsymbol{t}), H(\boldsymbol{t})$ can be defined similarly, and the corresponding cumulants are defined by

$$\kappa_{\boldsymbol{k}} = \frac{\partial^{|\boldsymbol{k}|} K}{\partial t_1^{k_1} \cdots \partial t_d^{k_d}}(\boldsymbol{0}), \quad \boldsymbol{k} = (k_1, \dots, k_d) \in \mathbb{N}^d,$$

and

$$K(\boldsymbol{t}) = \sum_{k_1=0}^{\infty} \cdots \sum_{k_d=0}^{\infty} \kappa_{\boldsymbol{k}} \frac{t_1^{k_1} \cdots t_d^{k_d}}{k_1! \cdots k_d!}.$$

It is straightforward to verify the relations

$$\begin{aligned} \mu_X = \kappa_X, \quad \mu_{XY} &= \kappa_{XY} + \mu_X \mu_Y, \\ \mu_{XYZ} = \kappa_{XYZ} + \mu_X \kappa_{YZ} &+ \mu_Y \kappa_{XZ} + \mu_Z \kappa_{XY} + \mu_X \mu_Y \mu_Z, \end{aligned}$$

and so on. The general relation between μ and κ for scalar X is left as an exercise.

For the multi-variate normal distribution $N(\boldsymbol{\mu}, \Sigma)$ we obtain

$$M(\boldsymbol{t}) = \exp\Big(\boldsymbol{\mu}\cdot\boldsymbol{t} + \frac{1}{2}\boldsymbol{t}^T \Sigma \boldsymbol{t}\Big), \quad K(\boldsymbol{t}) = \boldsymbol{\mu}\cdot\boldsymbol{t} + \frac{1}{2}\boldsymbol{t}^T \Sigma \boldsymbol{t}. \tag{17}$$

Note that only the cumulants κ_n with order $n \leq 2$ survive for Gaussian distributions. This property can be utilized to prove the useful Wick's theorem (see Exercise 6).

The moment and cumulant generating functions have explicit meaning in statistical physics, in which

$$Z(\beta) = \mathbb{E} e^{-\beta E}, \quad F(\beta) = -\beta^{-1} \ln Z(\beta)$$

are called *partition function* and *Helmholtz free energy*, respectively. Here $\beta = (k_B T)^{-1}$ is the inverse temperature, which is just a physical constant. They can be connected to M and K by

$$Z(\beta) = M_X(-\beta), \quad F(\beta) = -\beta^{-1} K_X(-\beta)$$

if X is taken as E, the energy of the system.

5 Borel-Cantelli Lemma

Let $\{A_n\}$ be a sequence of events, $A_n \in \mathcal{F}$. Define

$$\begin{aligned} \limsup_{n\to\infty}(A_n) &= \{\omega \in \Omega, \quad \omega \in A_n \text{ infinitely often (i.o.)}\} \\ &= \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k \end{aligned}$$

Lemma 2. *(First Borel-Cantelli Lemma) If* $\sum_{n=1}^{\infty} P(A_n) < +\infty$*, then* $P(\limsup_{n\to\infty} A_n) = P\{\omega : \omega \in A_n, i.o.\} = 0$.

Proof. $P\{\bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k\} \le P\{\bigcup_{k=n}^{\infty} A_k\} \le \sum_{k=n}^{\infty} P(A_k)$ for any n, but the last term goes to 0, as $n \to \infty$.

As an example of the application of this result, we prove

Lemma 3. *Let* $\{X_n\}$ *be a sequence of identically distributed (not necessarily independent) random variables, such that* $\mathbb{E}|X_n| < +\infty$*. Then*

$$\lim_{n\to\infty} \frac{X_n}{n} = 0 \qquad a.s.$$

The proof of this relies on another useful fact.

Lemma 4. *(Chebyshev Inequality) Let* X *be a random variable such that* $\mathbb{E}|X|^k < +\infty$*, for some integer* k*. Then*

$$P\{|X| > \lambda\} \le \frac{1}{\lambda^k}\mathbb{E}|X|^k$$

for any positive constant λ*.*

Proof. For any $\lambda > 0$,

$$\begin{aligned} \mathbb{E}|X|^k &= \int_{-\infty}^{\infty} |x|^k d\mu \ge \int_{|X|\ge\lambda} |X|^k d\mu \\ &\ge \lambda^k \int_{|X|\ge\lambda} d\mu = \lambda^k P\{|X| \ge \lambda\}. \end{aligned}$$

Proof of Lemma 3. For any $\epsilon > 0$, define

$$\begin{aligned}
A_n &= \{\omega \in \Omega : \left|\frac{X_n(\omega)}{n}\right| > \epsilon\} \\
\sum_n P(A_n) &= \sum_n P\{|X_n| > n\epsilon\} \\
&= \sum_n \sum_{k=n} P\{k\epsilon < |X_n| < (k+1)\epsilon\} \\
&= \sum_k kP\{k\epsilon < |X_n| < (k+1)\epsilon\} \\
&\leq \frac{1}{\epsilon}\mathbb{E}|X| < +\infty
\end{aligned}$$

Therefore if we define

$$B_\epsilon = \{\omega \in \Omega, \qquad \omega \in A_n \text{ i.o.}\}$$

then $P(B_\epsilon) = 0$. Let $B = \bigcup_{n=1}^\infty B_{\frac{1}{n}}$. Then $P(B) = 0$, and

$$\lim_{n\to\infty} \frac{X_n(\omega)}{n} = 0, \quad \text{if } \omega \notin B.$$

Lemma 5. *(Second Borel-Cantelli Lemma) If $\sum_{n=1}^\infty P(A_n) = +\infty$, and A_n are mutually independent, then*

$$P\{\omega \in \Omega, \quad \omega \in A_n \ i.o.\} = 1$$

6 Homeworks

- HW1. Prove the second Borel-Cantelli Lemma.
- HW2. Prove that if $X \sim \mathcal{P}(\lambda)$, $Y \sim \mathcal{P}(\mu)$ and X is independent of Y, then $X + Y \sim \mathcal{P}(\lambda + \mu)$.
- HW3. Suppose $X \sim \mathcal{P}(\lambda)$, $Y \sim \mathcal{P}(\mu)$ are two independent Poisson random variables and the sum $X + Y = N$ is fixed. Then the conditional distribution of X (or Y) is a Binomial distribution with parameter $n = N$ and $p = \lambda/(\lambda + \mu)$ (or $p = \mu/(\lambda + \mu)$).
- HW4. Prove the following statements:
  1. (Memoryless property of exponential distribution) Suppose $X \sim \mathcal{E}(\lambda)$, prove that
  $$\text{Prob}(X > s + t | X > s) = \text{Prob}(X > t) \quad \text{for all } s, t > 0.$$
  2. Let X be a random variable such that
  $$\text{Prob}(X > s + t) = \text{Prob}(X > s)\text{Prob}(X > t) \quad \text{for all } s, t > 0,$$
  prove that there exists $\lambda > 0$ such that $X \sim E(\lambda)$.

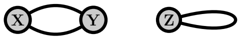
- HW5. (**Wick's theorem**) For multi-variate Gaussian random variables $(X_1, X_2, \ldots, X_n)$ with mean 0, utilize (17) and (16) to prove

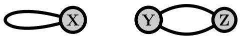
$$\mathbb{E}(X_1 X_2 \cdots X_k) = \begin{cases} \sum\prod \mathbb{E}(X_i X_j), & k \text{ is even}, \\ 0, & k \text{ is odd}, \end{cases}$$

where the notation $\sum\prod$ means summing of products over all possible partitions of $X_1, \ldots, X_k$ into pairs, e.g. for (X,Y,Z) is jointly Gaussian we obtain

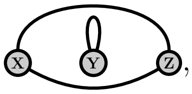
$$\begin{aligned} \mathbb{E}(X^2Y^2Z^2) =& (\mathbb{E}X^2)(\mathbb{E}Y^2)(\mathbb{E}Z^2) + 2(\mathbb{E}YZ)^2\mathbb{E}X^2 + 2(\mathbb{E}XY)^2\mathbb{E}Z^2 + 2(\mathbb{E}XZ)^2\mathbb{E}Y^2 \\ &+ 8(\mathbb{E}XY)(\mathbb{E}YZ)(\mathbb{E}XZ). \end{aligned} \tag{18}$$

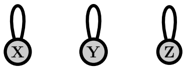
Each term in (18) can be schematically mapped to some graph as below

$(\mathbb{E}XY)^2\mathbb{E}Z^2 \longmapsto$ 

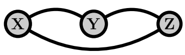
, $(\mathbb{E}YZ)^2\mathbb{E}X^2 \longmapsto$ 

,

$(\mathbb{E}XZ)^2\mathbb{E}Y^2 \longmapsto$ 

, $(\mathbb{E}X^2)(\mathbb{E}Y^2)(\mathbb{E}Z^2) \longmapsto$ 

,

$(\mathbb{E}XY)(\mathbb{E}YZ)(\mathbb{E}XZ) \longmapsto$ 

.

And the coefficient of each term is the combinatorial number for generating the corresponding schematic combinations. This is essentially the so-called Feynman diagrams.

- HW6. Suppose that the events A_n are mutually independent with $\text{Prob}(\cup_n A_n) = 1$ and $\text{Prob}(A_n) < 1$ for each n. Prove that $\text{Prob}\{A_n \ i.o.\} = 1$.

- HW7. Numerically investigate the limit process

$$\text{Binomial} \longrightarrow \text{Poisson} \longrightarrow \text{Normal distribution}$$

with MATLAB. Find the suitable parameter regime that the limit holds.

References

[1] Renguan Wang, An Introduction to Probability Theory, Peking University Press, 1998. (in chinese)

[2] B. Kisacanin, Mathematical problems and proofs: combinatorics, number theory, and geometry, Kluwer Academic Pub., New York, 2002.

[3] Billingsley. *Probability and measure.* John Wiley and Sons, New York, 1979.

[4] K.L. Chung. *A course in probability theory.* Academic Press, third edition, 2001.