PROBABILITY 3 REVISION NOTES

AARON SMITH
(REVISED BY MÁRTON BALÁZS)

1. Random Variables; Convolutions

1.1. **Definition.** A random variable X is *discrete* if there are countably many possible values X can take.

1.2. **Definition.** A random variable X is (absolutely) *continuous* if for all sets $A \subseteq \mathbb{R}$ ("of practical interest"/measurable) we have that

$$\mathbb{P}(X \in A) = \int_A f(x)dx,$$

with a function f called the density of X.

1.1. Discrete random variables.

1.3. **Example** (Integer valued random variable)**.** $X \in \mathbb{Z}$; we define $p(k) := \mathbb{P}(X = k)$ with the properties $p(k) \geq 0$ for all $k \in \mathbb{Z}$ and $\sum_{k\in\mathbb{Z}} p(k) = 1$. We define the expectation $\mathbb{E}X = \sum_{k\in\mathbb{Z}} kp(k)$ and the nth moment to be $\mathbb{E}X^n = \sum_{k\in\mathbb{Z}} k^n p(k)$. In general, $\mathbb{E}g(X) = \sum_{k\in\mathbb{Z}} g(k)p(k)$ for a function g on integers. In the above we assumed that the sums exist.

1.4. **Definition** (Binomial distribution)**.** $X \sim \text{Binomial}(n, p)$; $n \in \mathbb{N}$; $p \in [0, 1]$ with mass function

$$p(k) = \binom{n}{k} p^k q^{n-k}$$

(here and often in the sequel $q = 1 - p$; notice that the binomial coefficient is only non-zero for $0 \leq k \leq n$).

- *Meaning*: X is the number of successes in n independent trials; each is a success with probability p.
- Mass function? By Newton's Binomial Theorem: $(a + b)^n = \sum_{k=0}^n \binom{n}{k} a^n b^{n-k}$.
- Expectation and Variance? X is the sum of n independent Bernoulli(p) random variables i.e. $X \stackrel{d}{=} \sum_{i=1}^n X_i$ where $X_i \sim$ iid. Bernoulli(p); hence $\mathbb{E}X = np$; $\text{Var}X = npq$.

1.5. **Definition** (Bernoulli distribution)**.** $X \sim \text{Bernoulli}(p)$; $p \in [0, 1]$ where Bernoulli(p) $\stackrel{d}{=}$ Binomial$(1, p)$; hence has mass function

$$p(0) = q, \qquad p(1) = p.$$

- *Intuition*: There is a success $X = 1$ with probability p and a failure $X = 0$ with probability q.
- Mass function? $p + q = 1$.
- Expectation and Variance? $\mathbb{E}X = p$; $\mathbb{E}X^2 = p$ so $\text{Var}X = \mathbb{E}(X - \mathbb{E}X)^2 = \mathbb{E}X^2 - [\mathbb{E}X]^2 = p - p^2 = pq$.

1.6. **Definition** (Geometric distribution)**.** $X \sim \text{Geometric}(p)$ or $X \sim \text{OptGeom}(p)$ if X is the number of trials until the first success; each happening independently and with probability $p \in [0, 1]$.

$Y \sim \text{PessGeom}(p)$ if Y is the number of failures in a sequence of independent trials (each a success with probability p) before the first success.

$$p_X(k) = q^{k-1}p; \quad k = 1, 2, \ldots \qquad\qquad p_Y(k) = q^k p; \quad k = 0, 1, \ldots.$$

- *Remark*: $Y \stackrel{d}{=} X - 1$ since the number of failures is one less than the number of trials needed to see the first success.

- Mass function? Sum of geometric series: $\sum_{k=a}^{b} cr^k = cr^a \frac{1-r^{b-(a-1)}}{1-r}$. Hence

$$\sum_{k=1}^{\infty} p_X(k) = \sum_{k=1}^{\infty} q^{k-1}p = \frac{p}{1-q} = 1.$$

  For the Pessimistic Geometric $\sum_{k=0}^{\infty} p_Y(k) = \sum_{k=0}^{\infty} q^k p = \frac{p}{1-q} = 1$.

- Expectation and Variance? Use the formula for sum of geometric series (and the differentiated version). $\mathbb{E}X = \frac{1}{p}$; $\mathrm{Var}X = \frac{q}{p^2}$.

- $Y \stackrel{d}{=} X - 1 \implies \mathbb{E}Y = \mathbb{E}X - 1 = \frac{1}{p} - 1 = \frac{q}{p}$; $\mathrm{Var}Y = \mathrm{Var}X = \frac{q}{p^2}$.

1.7. **Definition** (Negative Binomial distribution)**.** $X \sim \mathrm{NegBin}(r, p)$; $r \in \mathbb{N}$; $p \in [0, 1]$. X is the number of independent trials (success with probability p) until the rth success.

$$p(k) = \binom{k-1}{r-1} q^{k-r} p^r; \quad k = r, r+1, \ldots.$$

- Need r successes contributing p^r; $k - r$ failures contributing q^{k-r} multiplied by the $\binom{k-1}{r-1}$ ways of rearranging the first $r - 1$ successes within the first $k - 1$ positions.

- $\mathrm{NegBin}(1, p) \stackrel{d}{=} \mathrm{OptGeom}(p)$.

- Mass function? First change the summation variable to get

$$\sum_{k=r}^{\infty} p(k) = \sum_{k=r}^{\infty} \binom{k-1}{r-1} q^{k-r} p^r = p^r \sum_{m=0}^{\infty} \binom{m+r-1}{r-1} q^m = p^r \sum_{m=0}^{\infty} \binom{m+r-1}{m} q^m.$$

  Next, use the general definition of a binomial coefficient $\binom{\alpha}{m}$ for real α and non-negative integer m to write

$$\begin{aligned}\binom{-r}{m} &= \frac{(-r)\cdot(-r-1)\cdots(-r-m+1)}{m!} \\ &= (-1)^m \frac{(r+m-1)\cdot(r+m-2)\cdots r}{m!} = (-1)^m \binom{m+r-1}{m}.\end{aligned}$$

  Thus,

$$\sum_{k=r}^{\infty} p(k) = p^r \sum_{m=0}^{\infty} \binom{-r}{m} (-q)^m = p^r \sum_{m=0}^{\infty} \binom{-r}{m} (-q)^m \cdot 1^{-r-m} = p^r \cdot (1-q)^{-r} = 1$$

  by Newton's Binomial Theorem, also valid in this more general form. As you see there is a good reason to call these distributions Negative Binomial.

- Expectation and Variance? Show later that sum of r independent Geometric(p) distributions $\implies$ $\mathbb{E}X = \frac{r}{p}$ and $\mathrm{Var}X = \frac{rq}{p^2}$.

1.8. **Definition** (Poisson distribution)**.** $X \sim \mathrm{Poisson}(\lambda)$ is the limiting distribution of $\mathrm{Binomial}(n, p(n))$ as $n \to \infty$ and $\lim_{n\to\infty} np(n) = \lambda$ [see later].

$$p(k) = \frac{e^{-\lambda}\lambda^k}{k!}; \quad k = 0, 1, \ldots.$$

- *Intuition*: X is the number of events which happen at rate λ.

- Mass function? By definition of exponential function as a sum.

- Expectation and Variance? $\mathbb{E}X = \mathrm{Var}X = \lambda$.

1.9. **Definition** (Hypergeometric distribution)**.** X has Hypergeometric distribution if it has mass function

$$p(k) = \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}} = \frac{\binom{n}{k}\binom{N-n}{m-k}}{\binom{N}{m}}; \qquad k = (n+m-N)^+, \ldots, \min\{n, m\}.$$

- *Intuition*: There are N deer; m are tagged; we catch n of them. X is the number of the caught deer which are tagged, i.e., the Hypergeometric distribution is the size of an intersection of two subsets of a population size N. What is the probability that the number of caught deer which are tagged is k?

  There are $\binom{N}{n}$ possible ways of catching the n deer.

  There are $\binom{m}{k}$ ways of choosing the k from all m tagged deer.

  There are $\binom{N-m}{n-k}$ ways of choosing the $n-k$ untagged from the total $N-m$ untagged deer.

- Mass function? Clearly from construction.

- Expectation and Variance? Each of the n captures has a probability $\frac{m}{N}$ of being a tagged deer. Let $\mathbb{1}_i$ be the indicator that the ith capture is a tagged deer. Then by linearity of expectation

$$\mathbb{E}X = \mathbb{E}\sum_{i=1}^{n} \mathbb{1}_i = \sum_{i=1}^{n} \frac{m}{N} = \frac{mn}{N}.$$

1.2. **Independence.**

1.10. **Definition.** X and Y are independent if for all $A, B \subseteq \mathbb{R}$ measurable,

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B).$$

For integer valued random variables, this is equivalent to $p_{X,Y}(n,m) = p_X(n)p_Y(m)$ for all $n,\ m$.

1.3. **Convolution of integer valued random variables.**

X and Y independent integer valued random variables. What is the mass function of $X+Y$? Define $p_{X+Y}(k) := \mathbb{P}(X+Y=k)$ then

$$p_{X+Y}(k) = \mathbb{P}(\{X+Y=k\}) = \mathbb{P}\Big(\bigcup_{i=-\infty}^{\infty} (\{X=k-i\} \cap \{Y=i\})\Big).$$

Since $\bigcup_{i=-\infty}^{\infty}(\{X=k-i\}\cap\{Y=i\})$ is a disjoint partition of $\{X+Y=k\}$ (Why? $\{Y=i\}$ is a disjoint partition of Ω so $\{X=k-i\}\cap\{Y=i\}$ is still a disjoint set over i, but now it has union $\{X+Y=k\}$), we have that

1.11. **Definition** (Convolution of mass functions)**.**

$$p_{X+Y}(k) = \sum_{i=-\infty}^{\infty} \mathbb{P}(\{X=k-i\}\cap\{Y=i\}) = \sum_{i=-\infty}^{\infty} p_X(k-i)p_Y(i).$$

by independence of X and Y.

1.12. *Remark.* This sum converges since $\sum_{i=-\infty}^{\infty} p_X(k-i)p_Y(i) \le \sum_{i=-\infty}^{\infty} p_Y(i) = 1$.

1.13. **Theorem** (Convolution of binomials)**.**

$$\text{Binomial}(n,p) * \text{Binomial}(m,p) \overset{d}{=} \text{Binomial}(n+m,p).$$

Proof. We can easily see this by considering the binomials as sums of Bernoullis. Indeed, $X \overset{d}{=} \sum_{i=1}^{n} X_i$; $X_i \overset{\text{i.i.d.}}{\sim}$ Bernoulli(p) and $Y \overset{d}{=} \sum_{i=1}^{m} Y_i$; $Y_i \overset{\text{i.i.d.}}{\sim}$ Bernoulli(p) and as the two sets of Bernoullis are also independent, $X+Y \overset{d}{=} \sum_{i=1}^{n} X_i + \sum_{i=1}^{m} Y_i \overset{d}{=} \sum_{i=1}^{m+n} Z_i$ where $Z_i \overset{\text{i.i.d.}}{\sim}$ Bernoulli(p) meaning $X+Y \sim$ Binomial$(n+m,p)$.

Using convolutions?

$$\begin{aligned}\sum_{i=-\infty}^{\infty} p_X(k-i)p_Y(i) &= \sum_{i=-\infty}^{\infty} \binom{n}{k-i} p^{k-i}q^{n-(k-i)} \binom{m}{i} p^i q^{m-i} \\ &= p^k q^{n+m-k} \sum_{i=-\infty}^{\infty} \binom{n}{k-i}\binom{m}{i} \\ &= \binom{n+m}{k} p^k q^{n+m-k} \sum_{i=-\infty}^{\infty} \frac{\binom{n}{k-i}\binom{m}{i}}{\binom{n+m}{k}}.\end{aligned}$$

This sum is equal to 1 since it is the sum of the mass function of a Hypergeometric distribution with $n+m$ deer; m are tagged and we capture k. □

1.14. **Theorem** (Convolution of Poissons)**.**

$$\text{Poisson}(\lambda) * \text{Poisson}(\mu) \overset{d}{=} \text{Poisson}(\lambda+\mu).$$

Proof.

$$\sum_{i=-\infty}^{\infty} p_X(k-i)p_Y(i) = \sum_{i=0}^{k} \frac{e^{-\lambda}\lambda^{k-i}}{(k-i)!}\frac{e^{-\mu}\mu^i}{i!} = e^{-(\lambda+\mu)} \sum_{i=0}^{k} \frac{\lambda^{k-i}}{(k-i)!}\frac{\mu^i}{i!} = \frac{e^{-(\lambda+\mu)}}{k!} \underbrace{\sum_{i=0}^{k} \binom{k}{i} \lambda^{k-i}\mu^i}_{=(\lambda+\mu)^k}$$

by Newton's binomial theorem. □

1.15. **Lemma** (Pascal's identity)**.** *The binomial coefficients can be arranged as in Pascal's triangle*

$$\binom{n}{r} = \binom{n-1}{r-1} + \binom{n-1}{r}.$$

Analytic proof.

$$\binom{n-1}{r-1} + \binom{n-1}{r} = \frac{(n-1)!}{(n-r)!(r-1)!} + \frac{(n-1)!}{(n-r-1)!r!} = \frac{(n-1)!r}{(n-r)!r!} + \frac{(n-1)!(n-r)}{(n-r)!r!} = \frac{n!}{(n-r)!r!}.$$ □

Combinatorial proof. Consider choosing r items from n where one of them is marked. Ways of choosing with marked in selection $= \binom{n-1}{i-1}$; ways of choosing without marked in selection $= \binom{n-1}{i}$. □

1.16. **Example.** By induction we have the following

(a) $\text{Geometric}(p) * \text{Geometric}(p) \overset{d}{=} \text{NegBin}(2,p)$;

(b) $\text{NegBin}(r,p) * \text{Geometric}(p) \overset{d}{=} \text{NegBin}(r+1,p)$;

(c) $*^r\text{Geometric}(p) \overset{d}{=} \text{NegBin}(r,p)$;

(d) $\text{NegBin}(r,p) * \text{NegBin}(s,p) \overset{d}{=} \text{NegBin}(r+s,p)$.

Proof. It is clear that (c) follows from (a) and (b) by induction and (d) follows from (c). Let $X, Y \overset{\text{i.i.d.}}{\sim} \text{Geometric}(p)$. Then

$$\sum_{i=-\infty}^{\infty} p_X(k-i)p_Y(i) = \sum_{i=1}^{k-1} q^{k-i-1}pq^{i-1}p = p^2q^{k-2}(k-1);$$

hence $X+Y \sim \text{NegBin}(2,p)$. This proves (a) and forms the basis step for (c).

Now let $X \sim \text{Geometric}(p)$ and $Y \sim \text{NegBin}(r,p)$. Then

$$\sum_{i=-\infty}^{\infty} p_X(k-i)p_Y(i) = \sum_{i=r}^{k-1} q^{k-i-1}p\binom{i-1}{r-1}q^{i-r}p^r = q^{k-(r+1)}p^{r+1}\sum_{i=r}^{k-1}\binom{i-1}{r-1}.$$

Hence it suffices to check that $\sum_{i=r}^{k-1}\binom{i-1}{r-1} = \binom{k-1}{r}$. We use Pascal's identity.

$$\sum_{i=r}^{k-1}\binom{i-1}{r-1} = \sum_{i=r}^{k-1}\left[\binom{i}{r} - \binom{i-1}{r}\right] = \binom{k-1}{r} - \binom{r-1}{r} = \binom{k-1}{r}.$$

(c) now follows and thus (d) is proved. □

1.4. **Continuous random variables; convolutions thereof.**

Let X and Y be independent continuous random variables. What is the distribution of $X+Y$?

$$F_{X+Y}(a) = \mathbb{P}(X+Y \le a) = \int\int_{\{x+y\le a\}} f(x,y)dxdy = \int_{-\infty}^{\infty}\int_{-\infty}^{a-y} f_X(x)dx f_Y(y)dy;$$

hence we have that

1.17. **Definition** (Convolution of distribution functions)**.**

$$F_{X+Y}(a) = \int_{-\infty}^{\infty} F_X(a-y)f_Y(y)dy.$$

Furthermore, we can differentiate to find the convolution of density functions.

$$\frac{d}{da}F_{X+Y}(a) = \frac{d}{da}\int_{-\infty}^{\infty} F_X(a-y)f_Y(y)dy = \int_{-\infty}^{\infty}\frac{d}{da}F_X(a-y)f_Y(y)dy = \int_{-\infty}^{\infty} f_X(a-y)f_Y(y)dy.$$

1.18. **Definition** (Convolution of density functions)**.**

$$f_{X+Y}(a) = \int_{-\infty}^{\infty} f_X(a-y)f_Y(y)dy.$$

1.19. **Definition** (Uniform distribution)**.** $X \sim \text{Uniform}(a,b)$ if X is equally likely to "fall anywhere" between a and b.

$$f(x) = \begin{cases} \frac{1}{b-a}, & \text{if } x \in [a,b], \\ 0, & \text{otherwise.} \end{cases}$$

- Density function? $\int_a^b \frac{1}{b-a}dx = 1$.
- Expectation and Variance? $\mathbb{E}X = \frac{a+b}{2}$; $\text{Var}X = \frac{(b-a)^2}{12}$.

1.20. **Example.** Convolution of two independent Uniform$(0,1)$ random variables.

$$f_{X+Y}(a) = \int_{-\infty}^{\infty} f_X(a-y) f_Y(y) dy = \int_{-\infty}^{\infty} \mathbb{1}_{\{0\le a-y\le 1\}} \mathbb{1}_{\{0\le y\le 1\}} dy = \int_{-\infty}^{\infty} \mathbb{1}_{\{a-1\le y\le a\}} \mathbb{1}_{\{0\le y\le 1\}} dy$$

$$= \begin{cases} 0 & a<0, \\ \int_0^a dy & a\in[0,1), \\ \int_{a-1}^1 dy & a\in[1,2), \\ 0 & a\ge 2. \end{cases} = \begin{cases} 0 & a<0, \\ a & a\in[0,1), \\ 2-a & a\in[1,2), \\ 0 & a\ge 2. \end{cases} \implies X+Y \text{ has triangular distribution.}$$

1.21. **Definition** (Exponential distribution)**.** $X \sim$ Exponential(λ); $\lambda \ge 0$.

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{if } x>0, \\ 0 & \text{otherwise.} \end{cases}$$

- Distribution? $F(x) = 1-e^{-\lambda x}$ for $x>0$ and 0 otherwise.
- Density function? $\int_0^\infty \lambda e^{-\lambda x} dx = -e^{-\lambda x}|_0^\infty = 1$.
- Expectation and Variance? Integration by parts, $\mathbb{E}X = \frac{1}{\lambda}$; $\mathrm{Var}X = \frac{1}{\lambda^2}$.

1.22. **Definition** (Normal distribution)**.** $X \sim \mathcal{N}(\mu, \sigma^2)$; $\mu \in \mathbb{R}$; $\sigma > 0$.

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

- Distribution? Defined equal to $\Phi(x)$ in the standard case $\mu=0$, $\sigma=1$.
- Density function? Define $I := \int_{-\infty}^{\infty} \varphi(x) dx$ where φ is the density of $Z \sim \mathcal{N}(0,1)$.

  $$I^2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \varphi(x)\varphi(y) dxdy = \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\frac{x^2+y^2}{2}} dxdy = \frac{1}{2\pi}\int_0^{2\pi}\int_0^{\infty} re^{-\frac{r^2}{2}} drd\theta = 1$$

  and hence $I=1$. By making a substitution we can verify it is a density function for general normal distribution.
- If $X \sim \mathcal{N}(\mu,\sigma^2)$ then $\frac{X-\mu}{\sigma} \overset{d}{=} Z$ where $Z \sim \mathcal{N}(0,1)$. Indeed let $Z = \frac{X-\mu}{\sigma}$ then

  $$F_Z(z) = \mathbb{P}(Z\le z) = \mathbb{P}\left(\frac{X-\mu}{\sigma} \le z\right) = \mathbb{P}(X \le \sigma z+\mu) = F_X(\sigma z+\mu).$$

  Hence

  $$f_Z(z) = \frac{d}{dz}F_Z(z) = \frac{d}{dz}F_X(\sigma z+\mu) = f_X(\sigma z+\mu)\times\sigma = \varphi(z).$$

- Expectation and Variance? $\mathbb{E}Z = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} xe^{-\frac{x^2}{2}} dx = 0$ as odd function integrated over the real line. $\mathbb{E}Z^2 = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} x^2 e^{-\frac{x^2}{2}} dx = -\frac{1}{\sqrt{2\pi}} xe^{-\frac{x^2}{2}}|_{-\infty}^{\infty} + \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{x^2}{2}} dx = 1$ and so $\mathrm{Var}Z = 1$.

  Since $X = \sigma Z + \mu$, $\mathbb{E}X = \mu$; $\mathrm{Var}X = \sigma^2$.

1.23. **Theorem** (Convolution of Normals)**.**

$$\mathcal{N}(\mu,\sigma^2) * \mathcal{N}(\nu,\tau^2) \overset{d}{=} \mathcal{N}(\mu+\nu, \sigma^2+\tau^2).$$

Proof. Let $X \sim \mathcal{N}(\mu,\sigma^2)$ and $Y \sim \mathcal{N}(\nu,\tau^2)$. First consider the case where $\mu=\nu=0$.

$$f_{X+Y}(a) = \int_{-\infty}^{\infty} f_X(a-y) f_Y(y) dy = \frac{1}{2\pi\sigma\tau}\int_{-\infty}^{\infty} e^{-\frac{(a-y)^2}{2\sigma^2}} e^{-\frac{y^2}{2\tau^2}} dy = \frac{1}{2\pi\sigma\tau}\int_{-\infty}^{\infty} e^{-\frac{1}{2}*} dy,$$

where

$$\begin{aligned}
* &= \frac{(a-y)^2}{\sigma^2} + \frac{y^2}{\tau^2} \\
&= \frac{y^2}{\sigma^2} + \frac{y^2}{\tau^2} - \frac{2ay}{\sigma^2} + \frac{a^2}{\sigma^2} \\
&= \frac{\sigma^2+\tau^2}{\sigma^2\tau^2}\left(y^2 - \frac{2a\tau^2}{\sigma^2+\tau^2}y\right) + \frac{a^2}{\sigma^2} \\
&= \frac{\sigma^2+\tau^2}{\sigma^2\tau^2}\left(y - \frac{a\tau^2}{\sigma^2+\tau^2}\right)^2 + \frac{a^2}{\sigma^2} - \frac{\sigma^2+\tau^2}{\sigma^2\tau^2}\left(\frac{a\tau^2}{\sigma^2+\tau^2}\right)^2 \\
&= \frac{\sigma^2+\tau^2}{\sigma^2\tau^2}\left(y - \frac{a\tau^2}{\sigma^2+\tau^2}\right)^2 + \frac{a^2}{\sigma^2} - \frac{a^2\tau^2}{\sigma^2(\sigma^2+\tau^2)} \\
&= \frac{\sigma^2+\tau^2}{\sigma^2\tau^2}\left(y - \frac{a\tau^2}{\sigma^2+\tau^2}\right)^2 + \frac{a^2(\sigma^2+\tau^2) - a^2\tau^2}{\sigma^2(\sigma^2+\tau^2)} \\
&= \frac{\sigma^2+\tau^2}{\sigma^2\tau^2}\left(y - \frac{a\tau^2}{\sigma^2+\tau^2}\right)^2 + \frac{a^2}{\sigma^2+\tau^2}.
\end{aligned}$$

Substituting this into the above we get that

$$f_{X+Y}(a) = \frac{1}{2\pi\sigma\tau}\int_{-\infty}^{\infty} e^{-\frac{1}{2}\left[\frac{\sigma^2+\tau^2}{\sigma^2\tau^2}\left(y-\frac{a\tau^2}{\sigma^2+\tau^2}\right)^2\right]}dy \times e^{-\frac{1}{2}\frac{a^2}{\sigma^2+\tau^2}}.$$

Let $x = \frac{\sqrt{\sigma^2+\tau^2}}{\sigma\tau}\left(y - \frac{a\tau^2}{\sigma^2+\tau^2}\right)$, then

$$f_{X+Y}(a) = \frac{1}{2\pi\sigma\tau}\int_{-\infty}^{\infty} e^{-\frac{x^2}{2}}\frac{\sigma\tau}{\sqrt{\sigma^2+\tau^2}}dx \times e^{-\frac{1}{2}\frac{a^2}{\sigma^2+\tau^2}} = \frac{1}{\sqrt{2\pi(\sigma^2+\tau^2)}}e^{-\frac{1}{2}\frac{a^2}{\sigma^2+\tau^2}},$$

and so $X+Y \sim \mathcal{N}(0, \sigma^2+\tau^2)$. Now for general μ and ν,

$$X+Y = X-\mu+Y-\nu+\mu+\nu.$$

$X-\mu \sim \mathcal{N}(0,\sigma^2)$ and $Y-\nu \sim \mathcal{N}(0,\tau^2)$ hence, by the above, $X-\mu+Y-\nu \sim \mathcal{N}(0,\sigma^2+\tau^2)$. Moreover, the result follows. □

1.24. **Definition** (Cauchy distribution)**.** Let $b \in \mathbb{R}$; $a > 0$. Consider a torch hanging at height a above b on the real axis. The torch points at an angle α from the downwards vertical where $\alpha \sim \text{Uniform}(-\frac{\pi}{2}, \frac{\pi}{2})$. Let

$$\tan(\alpha) = \frac{X-b}{a},$$

so X is the position the light beam hits the real axis. We say $X \sim \text{Cauchy}(b,a)$. What is the distribution/density function of X?

$$\begin{aligned}
F_X(x) = \mathbb{P}(X \le x) = \mathbb{P}\left(\tan(\alpha) \le \frac{x-b}{a}\right) = \mathbb{P}\left(\alpha \le \arctan\frac{x-b}{a}\right) &= \frac{1}{\pi}\left[\frac{\pi}{2} + \arctan\left(\frac{x-b}{a}\right)\right] \\
&= \frac{1}{2} + \frac{1}{\pi}\arctan\left(\frac{x-b}{a}\right).
\end{aligned}$$

$$f_X(x) = \frac{d}{dx}F_X(x) = \frac{1}{\pi}\frac{1}{1+(\frac{x-b}{a})^2}\frac{1}{a} = \frac{1}{\pi}\frac{a}{a^2+(x-b)^2}.$$

1.25. *Remark.* We say X has standard Cauchy distribution if $X \sim \text{Cauchy}(0,1)$ i.e., $a=1$, $b=0$. Making the necessary substitutions above,

$$F(x) = \frac{1}{2} + \frac{1}{\pi}\arctan(x) \qquad \text{and} \qquad f(x) = \frac{1}{\pi}\frac{1}{1+x^2}.$$

1.26. *Remark.* The Cauchy distribution has no mean or variance. This is because $\mathbb{E}X = O(\int_{-\infty}^{\infty}\frac{1}{x}dx)$, which is undefined.

1.27. **Proposition.** *Let $X \sim \text{Cauchy}(b, a)$. Then $\lambda X + \mu \sim \text{Cauchy}(\lambda b + \mu, |\lambda| a)$. In particular, if Z is standard Cauchy, then $aZ + b \sim \text{Cauchy}(b, a)$ for $a > 0$.*

Proof. Assume that $\lambda > 0$ such that $\lambda x + \mu$ is an increasing function of x. Then

$$F_{\lambda X+\mu}(x) = \mathbb{P}(\lambda X + \mu \leq x) = \mathbb{P}\left(X \leq \frac{x-\mu}{\lambda}\right) = \frac{1}{2} + \frac{1}{\pi} \arctan\left(\frac{x-\mu-b\lambda}{a\lambda}\right).$$

Differentiating with respect to x,

$$f_{\lambda X+\mu}(x) = \frac{1}{\pi}\frac{1}{a\lambda}\frac{1}{1+(\frac{x-\mu-b\lambda}{a\lambda})^2} = \frac{1}{\pi}\frac{a\lambda}{(a\lambda)^2 + (x-\mu-b\lambda)^2}. \qquad \square$$

1.28. **Lemma** (Cauchy's residue theorem)**.** *Let U be a simply connected (any closed curve in U can be continuously shrunk to a point inside the set) open domain in $\mathbb{C}$; $z_1, z_2, \ldots, z_n \in \mathbb{C}$; and f holomorphic (differentiable) on $U \backslash \{z_1, \ldots, z_n\}$. If γ is a closed curve orientated positively (interior on left) in U then*

$$\oint_\gamma f(z)dz = 2\pi i \sum_{i=1}^{n} \text{Res}(f; z_i).$$

Note that if f has Laurent series $f(z) = \sum_{n=-\infty}^{\infty} c_n (z - z_i)^n$ then we define $\text{Res}(f; z_i) = c_{-1}$.

1.29. **Theorem** (Convolution of Cauchys)**.**

$$\text{Cauchy}(b_X, a_X) * \text{Cauchy}(b_Y, a_Y) \stackrel{d}{=} \text{Cauchy}(b_X + b_Y, a_X + a_Y).$$

Proof. Let $X \sim \text{Cauchy}(b_X, a_X)$ and $Y \sim \text{Cauchy}(b_Y, a_Y)$. Then

$$f_{X+Y}(z) = \int_{-\infty}^{\infty} f_X(z-x) f_Y(x) dx = \int_{-\infty}^{\infty} \frac{1}{\pi}\frac{a_X}{a_X^2 + (z-x-b_X)^2}\frac{1}{\pi}\frac{a_Y}{a_Y^2 + (x-b_Y)^2} dx = \lim_{R\to\infty} \int_{-R}^{R} g(x)dx.$$

We extend the integral into the upper half plane, denoted γ and denote by φ the semicircular path joining R to $-R$ in the upper half plane. Then

$$\oint_\gamma g(x)dx = \oint_\varphi g(x)dx + \int_{-R}^{R} g(x)dx \implies f_{X+Y}(z) = \lim_{R\to\infty} \oint_\gamma g(x)dx - \lim_{R\to\infty} \oint_\varphi g(x)dx.$$

Along φ, $|g(x)| \leq |\frac{C}{x^2}| = \frac{C}{R^2}$ for some $C > 0$. Hence

$$\left| \oint_\varphi g(x)dx \right| \leq \oint_\varphi |g(x)|dx \leq \oint_\varphi \frac{C}{R^2} dx = \frac{\pi C}{R} \to 0 \implies \lim_{R\to\infty} \oint_\varphi g(x)dx = 0.$$

Thus it remains to compute the integral over the semicircle and the real line between $-R$ and R. Since this curve γ is a closed curve, we can use Cauchy's residue theorem. What are the singularities of

$$g(x) = \frac{1}{\pi^2}\frac{a_X}{a_X^2 + (z-x-b_X)^2}\frac{a_Y}{a_Y^2 + (x-b_Y)^2}?$$

- $a_X^2 + (z - x - b_X)^2 = 0$ when $z - x - b_X = \pm i a_X$ if and only if $x = z - b_X \pm i a_X$.
- $a_Y^2 + (x - b_Y)^2 = 0$ when $x - b_Y = \pm i a_Y$ if and only if $x = b_Y \pm i a_Y$.

Note that the two residues in the upper half plane are $z - b_X + ia_X$ and $b_Y + ia_Y$. We have that $g(x)$ can be rewritten as

$$\begin{aligned} g(x) &= \frac{a_X a_Y}{\pi^2} \frac{1}{[x-(z-b_X+ia_X)][x-(z-b_X-ia_X)][x-(b_Y+ia_Y)][x-(b_Y-ia_Y)]} \\ &= \frac{a_X a_Y}{\pi^2} \frac{1}{[x-z+b_X-ia_X][x-z+b_X+ia_X][x-b_Y-ia_Y][x-b_Y+ia_Y]}. \end{aligned}$$

Now we calculate residues. For a simple pole y, $\operatorname{Res}(g; y) = \lim_{x\to y}(x-y)g(x)$. Hence

$$\begin{aligned} \operatorname{Res}(g; z-b_X+ia_X) &= \frac{a_X a_Y}{\pi^2} \frac{1}{[x-z+b_X+ia_X][x-b_Y-ia_Y][x-b_Y+ia_Y]}\Bigg|_{x=z-b_X+ia_X} \\ &= \frac{a_X a_Y}{\pi^2} \frac{1}{2ia_X \underbrace{[z-(b_X+b_Y)+i(a_X-a_Y)]}_{:=A} \underbrace{[z-(b_X+b_Y)+i(a_X+a_Y)]}_{:=B}} \\ &= \frac{a_Y}{2i\pi^2 AB}. \end{aligned}$$

$$\begin{aligned} \operatorname{Res}(g; b_Y+ia_Y) &= \frac{a_X a_Y}{\pi^2} \frac{1}{[x-z+b_X-ia_X][x-z+b_X+ia_X][x-b_Y+ia_Y]}\Bigg|_{x=b_Y+ia_Y} \\ &= \frac{a_X a_Y}{\pi^2} \frac{1}{\underbrace{[-z+(b_X+b_Y)-i(a_X-a_Y)]}_{=-A} \underbrace{[-z+(b_X+b_Y)+i(a_X+a_Y)]}_{=-\overline{B}} 2ia_Y} \\ &= \frac{a_X}{2i\pi^2 A\overline{B}}. \end{aligned}$$

Hence by Cauchy's residue theorem, we have that

$$\begin{aligned} f_{X+Y}(z) &= 2\pi i\big[\operatorname{Res}(g; z-b_X+ia_X) + \operatorname{Res}(g; b_Y+ia_Y)\big] \\ &= 2\pi i\left[\frac{a_Y}{2i\pi^2 AB} + \frac{a_X}{2i\pi^2 A\overline{B}}\right] \\ &= \frac{1}{\pi}\left[\frac{a_Y}{AB} + \frac{a_X}{A\overline{B}}\right] \\ &= \frac{1}{\pi}\frac{a_Y\overline{B} + a_X B}{A|B|^2}. \end{aligned}$$

We simplify the numerator.

$$\begin{aligned} a_Y\overline{B} + a_X B &= a_Y\big[z-(b_X+b_Y)-i(a_X+a_Y)\big] + a_X\big[z-(b_X+b_Y)+i(a_X+a_Y)\big] \\ &= (a_X+a_Y)\big[z-(b_X+b_Y)+i(a_X-a_Y)\big] \\ &= (a_X+a_Y)A. \end{aligned}$$

Furthermore,

$$f_{X+Y}(z) = \frac{1}{\pi}\frac{a_X+a_Y}{|B|^2} = \frac{1}{\pi}\frac{a_X+a_Y}{(a_X+a_Y)^2 + \big[z-(b_X+b_Y)^2\big]}.$$ □

1.30. *Remark.* Let X and Y be independent standard Cauchy random variables. Then by the theorem above, $X+Y \sim \text{Cauchy}(0,2)$. By the transformation of a standard Cauchy $2X \sim \text{Cauchy}(0,2)$. This is highly unusual – we would not expect two times a random variable to have the same distribution as the sum of two independent copies.

Indeed this implies that we cannot have a finite variance: $\operatorname{Var}(X+Y) = 2\operatorname{Var}X$ but $\operatorname{Var}(2X) = 4\operatorname{Var}X$ which cannot be equal for a finite positive variance.

Furthermore, if $X_i \overset{\text{i.i.d.}}{\sim} \text{Cauchy}(0,1)$ then

$$X_1 + \cdots + X_n \sim \text{Cauchy}(0,n) \quad \Longrightarrow \quad \frac{X_1+\cdots+X_n}{n} \sim \text{Cauchy}(0,1)$$

and we see that the strong law of large numbers does not hold. This is because a Cauchy distribution does not have finite mean.

1.31. **Definition** (Gamma function, Gamma distribution)**.** We define the Gamma function $\Gamma : \mathbb{R}_+ \to \mathbb{R}$ via

$$\Gamma(z) = \int_0^\infty x^{z-1}e^{-x}dx \qquad (z > 0).$$

We have $\Gamma(1) = \int_0^\infty e^{-x}dx = 1$ (density of Exponential(1) distribution). Furthermore,

$$\Gamma(z) = \int_0^\infty x^{z-1}e^{-x}dx = \underbrace{-x^{z-1}e^{-x}\big|_0^\infty}_{=0} + \int_0^\infty (z-1)x^{z-2}e^{-x}dx = (z-1)\Gamma(z-1).$$

Moreover, for $n \in \mathbb{Z}_+$, $\Gamma(n) = (n-1)\Gamma(n-1) = \cdots = (n-1)!\Gamma(1) = (n-1)!$.

We can define the Gamma(n, λ); $\lambda > 0$, $n \in \mathbb{N}$, and the general Gamma(t, λ) $(t > 0)$ distributions with respective density functions

$$f(x) = \begin{cases} \frac{\lambda^n}{(n-1)!}x^{n-1}e^{-\lambda x} & x > 0 \\ 0 & \text{otherwise,} \end{cases} \qquad \text{and} \qquad f(x) = \begin{cases} \frac{\lambda^t}{\Gamma(t)}x^{t-1}e^{-\lambda x} & x > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Notice from the above working that these are exactly the same but with $t = n \in \mathbb{N}$.

- Density function? By the definition of Gamma function and making the substitution $y = \lambda x$,

$$\int_0^\infty \frac{\lambda^t}{\Gamma(t)}x^{t-1}e^{-\lambda x}dx = \frac{\lambda^t}{\Gamma(t)}\int_0^\infty \left(\frac{y}{\lambda}\right)^{t-1} e^{-y}\frac{1}{\lambda}dy = \frac{\lambda^t}{\Gamma(t)}\frac{\Gamma(t)}{\lambda^t} = 1.$$

- Expectation and Variance? By the properties of the Gamma function.

$$\mathbb{E}X = \int_0^\infty x\frac{\lambda^t}{\Gamma(t)}x^{t-1}e^{-\lambda x}dx = \frac{\lambda^t}{\Gamma(t)}\int_0^\infty \left(\frac{y}{\lambda}\right)^{t} e^{-y}\frac{1}{\lambda}dy = \frac{\Gamma(t+1)}{\lambda\Gamma(t)} = \frac{t}{\lambda}.$$

$$\mathbb{E}X^2 = \int_0^\infty x^2\frac{\lambda^t}{\Gamma(t)}x^{t-1}e^{-\lambda x}dx = \frac{\lambda^t}{\Gamma(t)}\int_0^\infty \left(\frac{y}{\lambda}\right)^{t+1} e^{-y}\frac{1}{\lambda}dy = \frac{\Gamma(t+2)}{\lambda^2\Gamma(t)} = \frac{t(t+1)}{\lambda^2}.$$

$$\mathrm{Var}X = \frac{t(t+1)}{\lambda^2} - \frac{t^2}{\lambda^2} = \frac{t}{\lambda^2}.$$

1.32. **Theorem** (Convolution of Exponentials)**.**

(a) $*^n\text{Exponential}(\lambda) \stackrel{d}{=} \text{Gamma}(n, \lambda)$*;*

(b) $\text{Gamma}(n, \lambda) * \text{Gamma}(m, \lambda) \stackrel{d}{=} \text{Gamma}(n + m, \lambda)$ *for* $m, n \in \mathbb{N}$*;*

(c) *In general,* $\text{Gamma}(t, \lambda) * \text{Gamma}(s, \lambda) \stackrel{d}{=} \text{Gamma}(t + s, \lambda)$.

Proof. First (a), which we prove by induction. By comparison of densities, Exponential$(\lambda) = $ Gamma$(1, \lambda)$. Now we aim to show that Gamma$(n, \lambda) *$ Exponential$(\lambda) = $ Gamma$(n + 1, \lambda)$. Indeed,

$$f_*(a) = \int_{-\infty}^\infty f_X(a-x)f_Y(x)dx = \int_0^a \lambda e^{-\lambda(a-x)}\frac{\lambda^n}{(n-1)!}x^{n-1}e^{-\lambda x}dx = \frac{\lambda^{n+1}}{(n-1)!}e^{-\lambda a}\int_0^a x^{n-1}dx$$
$$= \frac{\lambda^{n+1}}{(n-1)!}e^{-\lambda a}\frac{1}{n}x^n\big|_0^a = \frac{\lambda^{n+1}}{n!}a^n e^{-\lambda a}.$$

This proves (a). (b) is a corollary of (a), noting that Gamma$(n, \lambda) *$ Gamma$(m, \lambda) = *^{m+n}$Exponential(λ). No proof of (c) given. □

1.33. **Definition** (χ^2 distribution)**.** Let $X \sim \mathcal{N}(0, \sigma^2)$. Then $X^2 \sim \chi^2_1$. How is this defined?

$$F_{\chi^2_1}(a) = \mathbb{P}(X^2 \le a) = \mathbb{P}(-\sqrt{a} \le X \le \sqrt{a}) = \mathbb{P}(-\sqrt{a}/\sigma \le Z \le \sqrt{a}/\sigma) = \Phi(\sqrt{a}/\sigma) - (1 - \Phi(\sqrt{a}/\sigma))$$
$$= 2\Phi(\sqrt{a}/\sigma) - 1.$$

Differentiating with respect to a to find the density,

$$f_{\chi^2_1}(a) = \frac{d}{da}\left[2\Phi(\sqrt{a}/\sigma) - 1\right] = 2\varphi\left(\frac{\sqrt{a}}{\sigma}\right)\frac{1}{2\sigma\sqrt{a}} = \frac{1}{\sqrt{2\pi a\sigma^2}}e^{-\frac{a}{2\sigma^2}}.$$

I claim that this is equal to $\text{Gamma}(t,\lambda)$ for some λ and some t. In order to find these we compare terms. Set

$$\frac{1}{\sqrt{2\pi a\sigma^2}}e^{-\frac{a}{2\sigma^2}} \equiv \frac{\lambda^t}{\Gamma(t)}a^{t-1}e^{-\lambda a}.$$

Clearly then we must have

- $\lambda \equiv 1/2\sigma^2$;
- $t \equiv 1/2$.

Thus it remains to show that $\Gamma(1/2) = \sqrt{\pi}$. Since these are two proper densities, this is surely the case. Moreover,

$$X^2 \sim \chi^2_1 \stackrel{d}{=} \text{Gamma}\left(\frac{1}{2\sigma^2}, \frac{1}{2}\right).$$

1.34. **Theorem** (Convolution of χ^2_1)**.** *Let $X_1, X_2, \ldots \sim \mathcal{N}(0,\sigma^2)$ be independent. Then $X_i^2 \sim \chi^2_1$. Then $X_i^2 \sim \text{Gamma}(1/2, 1/2\sigma^2)$ and therefore*

$$X_1^2 + \cdots + X_n^2 \sim \text{Gamma}(n/2, 1/2\sigma^2).$$

When $\sigma = 1$ we call this a χ^2_n distribution.

1.35. *Remark.* By the strong law of large numbers and since a $\text{Gamma}(n,\lambda)$ is the sum of n Exponential(λ) distributions; each with mean $1/\lambda$,

$$\frac{\text{Gamma}(n,\lambda)}{n} \xrightarrow{\text{a.s.}} \frac{1}{\lambda}.$$

Similarly, by the central limit theorem,

$$\frac{\lambda X - n}{\sqrt{n}} \xrightarrow{d} \mathcal{N}(0,1).$$

2. The Poisson Process

2.1. **Definition** (Poisson process)**.** A Poisson process with *rate* $\lambda > 0$ is a continuous time counting process $N(t)$ such that

(a) $N(t)$ is of time homogeneous and independent increments;

(b) $N(t) = \#\{\text{marks in the interval } [0,t)\} \sim \text{Poisson}(\lambda t)$.

2.2. **Theorem.** *Let $N(t)$ be a counting process where the time between events are independent* Exponential(λ) *so that the time until the nth event T_n is a* $\text{Gamma}(n,\lambda)$ *random variable, then $N(t)$ is a Poisson process.*

Proof. Note that $\mathbb{P}(N(t) \geq n) = \mathbb{P}(T_n < t)$. For $k \geq 1$,

$$\begin{aligned}
\mathbb{P}(N(t) = k) &= \mathbb{P}(N(t) \geq k) - \mathbb{P}(N(t) \geq k+1) \\
&= \mathbb{P}(T_k < t) - \mathbb{P}(T_{k+1} < t) \\
&= \int_0^t \frac{\lambda^k}{(k-1)!} x^{k-1} e^{-\lambda x} dx - \underbrace{\int_0^t \frac{\lambda^{k+1}}{k!} x^k e^{-\lambda x} dx}_{u=x^k, \frac{dv}{dx}=e^{-\lambda x}} \\
&= \int_0^t \frac{\lambda^k}{(k-1)!} x^{k-1} e^{-\lambda x} dx + \frac{\lambda^{k+1}}{k!} \frac{x^k e^{-\lambda x}}{\lambda}\Big|_0^t - \int_0^t \frac{\lambda^{k+1}}{k!} \frac{k x^{k-1} e^{-\lambda x}}{\lambda} dx \\
&= \frac{(\lambda t)^k e^{-\lambda x}}{k!}.
\end{aligned}$$

For $k = 0$ it is trivial. Hence $N(t) \sim \text{Poisson}(\lambda t)$; as required. □

3. Generating functions

3.1. **Definition** (Probability generating function)**.** We define the probability generating function of a random variable X to be the function $P : \mathbb{R} \to \mathbb{R}$, defined by

$$P(s) = \mathbb{E}s^X.$$

Throughout this section our random variables will be non-negative integer valued.

3.2. **Proposition** (Properties of generating functions)**.**

(a) $P(1) = \mathbb{P}(X < \infty)$. *If X has a proper distribution then this is 1, otherwise $P(1) < 1$.*

Proof. Indeed $P(s)\big|_{s=1} = \sum_{n=0}^{\infty} s^n p(n)\big|_{s=1} = \sum_{n=0}^{\infty} p(n)$. □

(b) *The radius of convergence $R \geq 1$.*

Proof. By (a), $P(s) < \infty$ for all $|s| < 1$. Hence $R \geq 1$. □

(c) $P(0) = \mathbb{P}(X = 0)$.

Proof. Indeed $P(s)\big|_{s=0} = \sum_{n=0}^{\infty} s^n p(n)\big|_{s=0} = p(0) + \sum_{n=1}^{\infty} s^n p(n)\big|_{s=0} = p(0)$. □

(d) $(\frac{d}{ds})^k P(s)\big|_{s=0} = k!p(k)$. *In particular, the distribution of a random variable is uniquely determined by its probability generating function.*

Proof.

$$\begin{aligned}
\left(\frac{d}{ds}\right)^k P(s) &= \left(\frac{d}{ds}\right)^k \sum_{n=0}^{\infty} s^n p(n) = \sum_{n=0}^{\infty} \left(\frac{d}{ds}\right)^k s^n p(n) = \sum_{n=0}^{\infty} n(n-1)\cdots(n-k+1)s^{n-k}p(n) \\
&= \underbrace{\sum_{n=0}^{k-1} n(n-1)\cdots(n-k+1)s^{n-k}p(n)}_{\text{all terms } =0} + k!p(k) + \sum_{n=k+1}^{\infty} n(n-1)\cdots(n-k+1)s^{n-k}p(n).
\end{aligned}$$

Evaluating at $s = 0$ yields the result. Thus there is a one-to-one correspondence between $P(s)$ and $\{p(n)\}_{n=0}^{\infty}$. Indeed $p(k)$ is equal to the kth coefficient of the Taylor series of P expanded about 0. □

(e) $(\frac{d}{ds})^k P(s)\big|_{s=1} = \mathbb{E}\big[X(X-1)\cdots(X-k+1)\big]$. *We call this the kth factorial moment.*

Proof. As above,

$$(\frac{d}{ds})^k P(s)\big|_{s=1} = \sum_{n=0}^{\infty} n(n-1)\cdots(n-k+1)s^{n-k}p(n)\big|_{s=1} = \mathbb{E}\big[X(X-1)\cdots(X-k+1)\big]. \qquad \square$$

3.3. **Theorem.** *A continuous function $P : [0,1) \to \mathbb{R}$ is of the form $P(s) = \sum_{n=0}^{\infty} p(n)s^n$ with $p(n) \geq 0$ if and only if for all $n \geq 0$ and for all $s \in (0,1)$, $P^{(n)}(s)$ exists and is non-negative. Moreover, P is the generating function of a probability distribution if $P(1) = 1$ such that $\sum_{n=0}^{\infty} p(n) = 1$; thus the coefficients form a probability distribution on $\mathbb{Z}_+$.*

3.4. **Proposition.** *Let X and Y be independent non-negative integer valued random variables and let P_X and P_Y be the generating functions of X and Y respectively. Then*

$$P_{X+Y}(s) = P_X(s)P_Y(s).$$

Proof.

$$P_{X+Y}(s) = \mathbb{E}(s^{X+Y}) = \mathbb{E}(s^X s^Y) = \mathbb{E}s^X \mathbb{E}s^Y = P_X(s)P_Y(s).$$

The penultimate equality follows from the independence of X and Y. $\square$

We can use this result to determine the distribution of a convolution of two independent random variables – this saves using the convolution formula.

3.5. **Example.** $X \sim \text{Poisson}(\lambda)$ and $Y \sim \text{Poisson}(\mu)$ independent. Then

$$P_X(s) = \mathbb{E}s^X = \sum_{n=0}^{\infty} s^n \frac{e^{-\lambda}\lambda^n}{n!} = e^{-\lambda}e^{\lambda s} = e^{\lambda(s-1)}.$$

Hence the convolution has generating function

$$P_{X+Y}(s) = P_X(s)P_Y(s) = e^{(\lambda+\mu)(s-1)},$$

which in turn implies that $X + Y \sim \text{Poisson}(\lambda + \mu)$.

3.6. **Theorem** (Random number of summands)**.** *Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables and let Y be a non-negative integer valued random variable, jointly independent of the X_i's. Let $Z := \sum_{i=1}^{Y} X_i$, then*

(a) *$P_Z(s) = P_Y[P_{X_1}(s)]$; that is, $P_Z = P_Y \circ P_{X_1}$;*

(b) *$\mathbb{E}Z = \mathbb{E}Y \cdot \mathbb{E}X_1$;*

(c) *$\text{Var}Z = [\mathbb{E}X_1]^2\text{Var}Y + \mathbb{E}Y\text{Var}X_1$.*

Proof. For (a), we use the law of iterated expectation (aka. Tower Rule or Law of Total Expectations), properties of the conditional expectation and independence.

$$P_Z(s) = \mathbb{E}s^{\sum_{i=1}^{Y} X_i} = \mathbb{E}\left[\mathbb{E}\left(\prod_{i=1}^{Y} s^{X_i}\middle| Y\right)\right] = \mathbb{E}\left[\prod_{i=1}^{Y}\mathbb{E}(s^{X_i}|Y)\right] = \mathbb{E}P_{X_1}(s)^Y = P_Y(P_{X_1}(s)).$$

For (b) use that $P'(1) = \mathbb{E}X$ and the chain rule.

$$P'_Z(s) = P'_Y(P_{X_1}(s)) \times P'_{X_1}(s).$$

Evaluate at $s = 1$ and use that $P_{X_1}(1) = 1$. For (c) note that

$$P_Z''(s) = P_Y'(P_{X_1}(s)) \times P_{X_1}''(s) + P_{X_1}'(s) \times P_Y''(P_{X_1}(s))P_{X_1}'(s).$$

Hence

$$\mathbb{E}[Z(Z-1)] = P_Z''(1) = \mathbb{E}Y\mathbb{E}[X_1(X_1-1)] + (\mathbb{E}X_1)^2\mathbb{E}[Y(Y-1)].$$

Moreover,

$$\begin{aligned}
\mathrm{Var}Z &= \mathbb{E}Z^2 - (\mathbb{E}Z)^2 \\
&= \mathbb{E}[Z(Z-1)] + \mathbb{E}Z - (\mathbb{E}Z)^2 \\
&= \mathbb{E}Y\mathbb{E}[X_1(X_1-1)] + (\mathbb{E}X_1)^2\mathbb{E}[Y(Y-1)] + \mathbb{E}X_1\mathbb{E}Y - (\mathbb{E}X_1\mathbb{E}Y)^2 \\
&= \mathbb{E}Y\mathbb{E}X_1^2 - \mathbb{E}Y\mathbb{E}X_1 + (\mathbb{E}X_1)^2\mathbb{E}Y^2 - (\mathbb{E}X_1)^2\mathbb{E}Y + \mathbb{E}X_1\mathbb{E}Y - (\mathbb{E}X_1\mathbb{E}Y)^2 \\
&= [\mathbb{E}X_1]^2\mathrm{Var}Y + \mathbb{E}Y\mathrm{Var}X_1.
\end{aligned}$$

□

3.7. **Example** (Bernoulli thinning of Poisson process)**.** Let $Y \sim \text{Poisson}(\lambda)$. Each arrival is independently a success with probability p, i.e., let $X_i \sim$ iid. Bernoulli(p), jointly independent of Y, and let $Z = \sum_{i=1}^{Y} X_i$ be the number of successes. Then $P_{X_1}(s) = ps + q$ and so

$$P_Z(s) = e^{\lambda(ps+q-1)} = e^{\lambda p(s-1)}$$

and hence $Z \sim \text{Poisson}(\lambda p)$.

4. Galton-Watson (Branching) Process

4.1. **Definition.** Let $\{Z_{n,j}\}_{n,j=1}^{\infty}$ have non-negative integer distribution, be iid and have generating function $P(s)$. Let

- $Z_0 = 1$;
- $Z_n = Z_{n,1} + Z_{n,2} + \cdots + Z_{n,Z_{n-1}}$.

Then Z_n is the population size at time n. $Z_{n,j}$ is the number of offspring the jth member of generation $n-1$ has.

4.2. **Proposition.**

$$P_{Z_n} = \circ^n P.$$

Proof. Follows from previous on generating functions of random summands:

$$P_{Z_n}(s) = (P_{Z_{n-1}} \circ P)(s) = \cdots = (P_{Z_0} \circ \underbrace{P \circ \cdots \circ P}_{\times n})(s) = (\circ^n P)(s)$$

since $Z_0 = 1$ implies $P_{Z_0}(s) = s$. □

4.3. **Definition.** Let π be the probability that the population goes extinct, i.e., that at some point in time all members of the previous generation have no offspring. If $\pi < 1$ then with some positive probability the population never dies out.

4.4. **Theorem** (Probability of extinction)**.** *Assume $p(0) = \mathbb{P}(Z_{n,j} = 0) > 0$; $p(0) < 1$. Define $m := \mathbb{E}Z_{n,j}$. Then (if we assume the offspring distribution is proper),*

$$\pi = \begin{cases} 1 & \text{if } m < 1 \text{ (subcritical case)}, \\ 1 & \text{if } m = 1 \text{ (critical case)}, \\ s_0 & \text{if } m > 1 \text{ (supercritical case)}, \end{cases}$$

where s_0 is such that $P(s_0) = s_0$; $0 < s_0 < 1$.

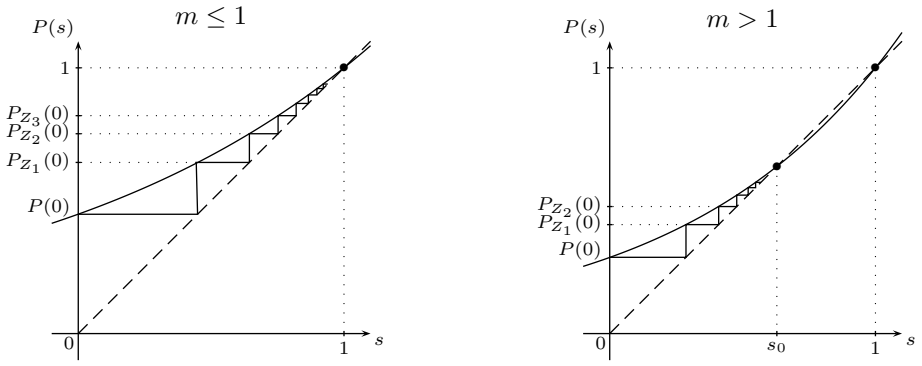
Proof. Note that $\{Z_n = 0\}$ is an increasing sequence of events; hence

$$\pi = \mathbb{P}\Big(\bigcup_{n=1}^{\infty}\{Z_n = 0\}\Big) = \lim_{n\to\infty}\mathbb{P}(Z_n = 0).$$

But $\mathbb{P}(Z_n = 0) = P_{Z_n}(0) = (\circ^n P)(0)$. Graphical properties:

- P is convex on $[0,1]$ (in fact all generating functions are, as they are positive linear combinations of convex functions);
- $P(1) = 1$;
- $P'(1) = m$;
- $P(0) = \mathbb{P}(Z_{n,j} = 0)$;
- by the theorem not proved on generating functions $P^{(k)}(s)$ exists and is non-negative for all k and all $s \in (0,1)$.

Hence we sketch the generating function P for the two cases $m \le 1$ and $m > 1$. □



4.5. *Remark.* When $m < 1$ [SUBCRITICAL] we have exponential decay in the limit. When $m = 1$ [CRITICAL] we might not.

4.6. **Corollary.** *From the graphs can also see that for all $s \in (0,1)$*

$$(\circ^n P)(s) = P_{Z_n}(s) \to \begin{cases} 1 & \text{if } m \le 1, \\ s_0 & \text{if } m > 1. \end{cases}$$

Furthermore, for all $k > 0$,

$$\mathbb{P}(Z_n = k) = \frac{1}{k!}\left(\frac{d}{ds}\right)^k P_{Z_n}(s)\Big|_{s=0} \to 0.$$

No formal proof, but intuitive since P_{Z_n} converges to a constant function.

4.7. **Example** (Pessimistic geometric offspring distribution)**.** $Z_{n,j} \sim \text{PessGeom}(p)$. $\mathbb{E}Z_{n,j} = \frac{1}{p} - 1 = \frac{q}{p}$. Fix $\frac{q}{p} > 1$ i.e. $q > p$ so that we are in the SUPERCRITICAL case – this is the case when there is no extinction with positive probability.

$$P(s) = \mathbb{E}s^{Z_{n,j}} = \sum_{i=0}^{\infty} q^i p s^i = \frac{p}{1-qs}.$$

Solve for s_0 such that $P(s_0) = s_0$.

$$\frac{p}{1-qs_0} = s_0 \iff qs_0^2 - s_0 + p = 0 \iff s_0 = \frac{1 \pm \sqrt{1-4pq}}{2q}.$$

Note that $1 - 4pq = 1 - 4p(1-p) = 1 - 4p + 4p^2 = (1-2p)^2$. Hence

$$s_0 = \frac{1 \pm |1-2p|}{2q} = \frac{1 \pm (1-2p)}{2q}$$

as $|\cdot|$ becomes irrelevant with $\pm$.

$$+ \implies s_0 = 1, \qquad\qquad - \implies s_0 = \frac{p}{q}.$$

By the theorem, the extinction probability is the smallest root. Since $p/q < 1$, $\pi = p/q$. Furthermore, the survival probability

$$\theta = 1 - \pi = 1 - \frac{p}{q} = 1 - \frac{p - 1 + 1}{1 - p} = 2 - \frac{1}{1 - p}$$

for $p < 1/2$ and for 0 for $p \geq 1/2$ [CRITICAL and SUBCRITICAL cases].

In fact, we can check this in the CRITICAL case $p = q$,

$$P(s) = \frac{p}{1 - qs} = \frac{1}{2 - s},$$

$$P(P(s)) = \frac{1}{2 - \frac{1}{2-s}} = \frac{2 - s}{3 - 2s},$$

$$P(P(P(s))) = \frac{1}{2 - \frac{2-s}{3-2s}} = \frac{3 - 2s}{4 - 3s}.$$

Conjecture that

$$(\circ^n P)(s) = \frac{n - (n-1)s}{(n+1) - ns}.$$

Suppose that it is true for n. Then we show it is for $n + 1$.

$$(\circ^{n+1} P)(s) = P((\circ^n P)(s)) = \frac{1}{2 - \frac{n-(n-1)s}{(n+1)-ns}} = \frac{(n+1) - ns}{2(n+1) - 2ns - n + (n-1)s} = \frac{(n+1) - ns}{(n+2) - (n+1)s}.$$

Hence

$$\mathbb{P}(Z_n = 0) = P_{Z_n}(0) = (\circ^n P)(0) = \frac{n}{n+1} \to 1.$$

So $\pi = \lim_{n\to\infty} \mathbb{P}(Z_n = 0) = 1$. Note that $\mathbb{P}(Z_n > 0) = 1 - \frac{n}{n+1} \sim \frac{1}{n}$ and hence the decay is power law and not exponential as in the SUBCRITICAL case.

4.8. **Theorem.** *Let X be the total number of individuals that ever existed in a branching process. Let X have generating function Q. Then*

$$Q(s) = \left(\frac{s}{P(s)}\right)^{-1} \longleftarrow \textit{ function inverse!}$$

Proof. Z_1 is the number of offspring the initial member of the population ($Z_0 = 1$) has. Hence we may view the subsequent members of the population as members of Z_1 i.i.d. branching processes. Since all the branching processes are i.i.d., each of these has X members also (in distribution, not deterministically). Hence

$$X \stackrel{d}{=} 1 + \sum_{i=1}^{Z_1} X_i \qquad \text{where} \qquad X_i \stackrel{d}{=} X, \text{ iid., and independent of } Z_1.$$

Hence

$$Q(s) = \mathbb{E}s^X = s\mathbb{E}s^{\sum_{i=1}^{Z_1} X_i} = sP(Q(s)).$$

This comes from generating function of a random summand (earlier). Q is an increasing function on $[0, 1]$ hence is bijective and so Q^{-1} exists on $[0, 1]$. Let $y = Q(s)$. Then

$$y = sP(y) \iff s = \frac{y}{P(y)} \implies Q^{-1}(y) = \frac{y}{P(y)}.$$

□

5. Simple random walks on $\mathbb{Z}$

5.1. **Definition.** Let

$$X_i = \begin{cases} +1 & \text{w.p. } p, \\ -1 & \text{w.p. } q. \end{cases}$$

be independent and

$$S_n = \sum_{i=1}^{n} X_i$$

be the position of the walker at time n having started at 0.

5.2. **Definition** (Level 1 hitting time)**.** Define $\tau^+ = \inf\{n \geq 1 : S_n = 1\}$ to the first time the random walk reaches +1. We call this the level 1 hitting time. Let τ^+ have generating function P^+.

5.3. **Theorem** (Level 1 hitting time)**.**

$$P^+(s) = \frac{1 - \sqrt{1 - 4pqs^2}}{2qs}.$$

$$\mathbb{P}(\textit{walker ever reaches level 1}) = \mathbb{P}(\tau^+ < \infty) = \begin{cases} \frac{p}{q} & \textit{if } p < q \iff p < \frac{1}{2}, \\ 1 & \textit{if } p \geq q \iff p \geq \frac{1}{2}. \end{cases}$$

$$\mathbb{E}\tau^+ = \begin{cases} \infty & \textit{if } p \leq q \iff p \leq \frac{1}{2}, \\ \frac{1}{p-q} & \textit{if } p > q \iff p > \frac{1}{2}. \end{cases}$$

Hence $\mathbb{E}\tau^+ \to \infty$ *as* $p \searrow \frac{1}{2}$.

Proof.

$$\begin{aligned} P^+(s) &= \mathbb{E}s^{\tau^+} \\ &= \mathbb{E}(\underbrace{s^{\tau^+}|X_1 = 1}_{\text{Here } \tau^+ = 1})\mathbb{P}(X_1 = 1) + \mathbb{E}(\underbrace{s^{\tau^+}|X_1 = -1}_{(\tau^+|X_1=-1) \overset{d}{=} 1 + \tau_1^+ + \tau_2^+ \text{ where } \tau_i^+ \overset{d}{=} \tau^+ \text{ and independent}})\mathbb{P}(X_1 = -1) \\ &= sp + s[P^+(s)]^2 q. \end{aligned}$$

By the quadratic formula

$$P^+(s) = \frac{1 \pm \sqrt{1 - 4pqs^2}}{2qs}.$$

We cannot take the solution with a + since this blows up for $s \to 0$. This proves the first part.

Now we compute $\mathbb{P}(\tau^+ < \infty)$.

$$\mathbb{P}(\tau^+ < \infty) = P^+(1) = \frac{1 - \sqrt{1 - 4pq}}{2q} = \frac{1 - |1 - 2p|}{2q} = \begin{cases} \frac{1-(1-2p)}{2q} = \frac{p}{q} & \text{when } 1 - 2p > 0 \iff p < \frac{1}{2} \\ \frac{1+(1-2p)}{2q} = 1 & \text{when } 1 - 2p \leq 0 \iff p \geq \frac{1}{2}. \end{cases}$$

Hence if $p \geq q$, $p \geq 1/2$ the walk reaches +1 almost surely. If $p < q$, $p < 1/2$ then the walk reaches +1 with probability $p/q < 1$.

For the final part, if $p < q$ then $\mathbb{P}(\tau^+ = \infty) > 0$ and so $\mathbb{E}\tau^+ = \infty$.

Now suppose $p \geq q$. Then

$$\begin{aligned}
\mathbb{E}\tau^+ &= \frac{d}{ds}P^+(s)\Big|_{s=1} \\
&= \frac{2qs \times (-\frac{1}{2})(1-4pqs^2)^{-1/2}(-8pqs) - (1-\sqrt{1-4pqs^2}) \times 2q}{4q^2s^2}\Big|_{s=1} \\
&= \frac{2p}{|1-2p|} - \frac{1-|1-2p|}{2q} \\
&= \frac{2p}{2p-1} - \frac{2-2p}{2q} \\
&= \frac{2p}{2p-1} - 1 \\
&= \frac{1}{2p-1} \\
&= \frac{1}{p-q}.
\end{aligned}$$

Note that this $\nearrow \infty$ as $p \searrow \frac{1}{2}$. □

5.4. **Theorem** (Level -1 hitting time)**.** *Let τ^- be the first time the walker hits level -1. Then by simply reversing the roles of p and q we have*

$$P^-(s) = \frac{1-\sqrt{1-4pqs^2}}{2ps}.$$

$$\mathbb{P}(\textit{walker ever reaches level -1}) = \mathbb{P}(\tau^- < \infty) = \begin{cases} \frac{q}{p} & \text{if } q < p \iff q < \frac{1}{2}, \\ 1 & \text{if } q \geq p \iff q \geq \frac{1}{2}. \end{cases}$$

$$\mathbb{E}\tau^- = \begin{cases} \infty & \text{if } q \leq p \iff q \leq \frac{1}{2}, \\ \frac{1}{q-p} & \text{if } q > p \iff q > \frac{1}{2}. \end{cases}$$

Hence $\mathbb{E}\tau^- \to \infty$ as $q \searrow \frac{1}{2}$.

5.5. **Theorem** (First return time to 0)**.** *Let $\tau^0 = \inf\{n \geq 1 : S_n = 0\}$. Let P^0 be the generating function of τ_0. Then*

$$P^0(s) = 1 - \sqrt{1-4pqs^2}.$$

Furthermore

$$\mathbb{P}(\tau^0 < \infty) = \begin{cases} 2p & \text{if } p \leq q, \\ 2q & \text{if } q < p. \end{cases}$$

Proof.

$$\begin{aligned}
P^0(s) &= \mathbb{E}s^{\tau^0} \\
&= \mathbb{E}(s^{\tau^0}|X_1 = 1)\mathbb{P}(X_1 = 1) + \mathbb{E}(s^{\tau^0}|X_1 = -1)\mathbb{P}(X_1 = -1) \\
&= psP^-(s) + qsP^+(s) \\
&= ps\frac{1-\sqrt{1-4pqs^2}}{2ps} + qs\frac{1-\sqrt{1-4pqs^2}}{2qs} \\
&= 1 - \sqrt{1-4pqs^2}.
\end{aligned}$$

For the next part,

$$\mathbb{P}(\tau^0 < \infty) = P^0(1) = 1 - |1-2p| = 1 - |p-q| = \begin{cases} 2p & \text{if } p \le q, \\ 2q & \text{if } q < p. \end{cases}$$

Finally, when $p \neq q$ then $\tau^0 = \infty$ with some positive probability and therefore $\mathbb{E}\tau^0 = \infty$. When $p = q = 1/2$,

$$\frac{d}{ds}P^0(s)\Big|_{s=1} = \frac{d}{ds}\Big[1 - \sqrt{1-s^2}\Big]\Big|_{s=1} = -\frac{1}{2}(1-s^2)^{-1/2}(-2s)\Big|_{s=1} = \infty \qquad \square$$

In all cases the expected time to return to 0 is infinite.

- $p = q \implies \tau^0 < \infty$ almost surely, we say NULL RECURRENT.
- $p \neq q \implies \tau^0 = \infty$ with some positive probability, we say TRANSIENT.

5.6. *Remark* (Relation to branching process). Let a branching process have offspring distribution

$$Z_{n,j} = \begin{cases} 0 & \text{w.p. } q \\ 2 & \text{w.p. } p. \end{cases}$$

Then $P(s) = q + ps^2$. Let Q be the generating function of the total number whoever existed. Then

$$\begin{aligned} Q(s) &= sP(Q(s)) \\ &= s \times [q + pQ(s)^2] \qquad \iff \qquad psQ(s)^2 - Q(s) + qs = 0. \end{aligned}$$

By the quadratic formula

$$Q(s) = \frac{1 \pm \sqrt{1 - 4pqs^2}}{2ps}.$$

Again we take the $-$ so that it doesn't blow up when $s \to 0$. Hence X is equal in distribution to the level -1 hitting time. Why? Each birth can be viewed as increasing the number of branches by ± 1, the same as a simple random walk.

6. Limit distributions, Continuity Theorem (using generating functions)

6.1. **Definition.** If X_n is a sequence of non-negative integer valued random variables then we say that they *converge in distribution* (weakly) to X if for all $k \ge 0$,

$$\mathbb{P}(X_n = k) \to \mathbb{P}(X = k) \qquad \text{as} \qquad n \to \infty.$$

6.2. **Theorem** (Continuity theorem)**.** *Let X_n be a sequence of non-negative integer valued and finite (proper) random variables. Then*

$$\lim_{n\to\infty} \mathbb{P}(X_n = k) =: p(k)$$

exists for all $k \ge 0$ if and only if for all $s \in (0,1)$

$$\lim_{n\to\infty} P_{X_n}(s) =: P(s)$$

exists. In this case, $P(s) = \sum_{n=0}^{\infty} s^n p(k)$.

Proof. Suppose that $\lim_{n\to\infty}\mathbb{P}(X_n = k) = p(k)$ as $n \to \infty$. Let $P(s)$ be the generating function of the limiting distribution.

$$\begin{aligned}
|P_{X_n}(s) - P(s)| &= \left|\sum_{k=0}^{\infty}(\mathbb{P}(X_n = k) - p(k))s^k\right| \\
&\leq \sum_{k=0}^{\infty}|\mathbb{P}(X_n = k) - p(k)|s^k \\
&= \sum_{k=0}^{M}|\mathbb{P}(X_n = k) - p(k)|\underbrace{s^k}_{<1} + \sum_{k=M+1}^{\infty}\overbrace{|\mathbb{P}(X_n = k) - p(k)|}^{\leq 1 \text{ as worst case scenario}}s^k \\
&\leq \sum_{k=0}^{M}|\mathbb{P}(X_n = k) - p(k)| + \sum_{k=M+1}^{\infty}s^k.
\end{aligned}$$

Since the sum on the right is convergent, the tail converges to 0. Hence there exists M such that the right hand term is less than a fixed $\varepsilon > 0$. So

$$|P_{X_n}(s) - P(s)| \leq \sum_{k=0}^{M}|\mathbb{P}(X_n = k) - p(k)| + \varepsilon.$$

As $\lim_{n\to\infty}\mathbb{P}(X_n = k) = p(k)$, there exists sufficiently large n such that the remaining sum is less than ε. Hence we have the desired convergence.

Now suppose that for all $s \in (0, 1)$ we have $\lim_{n\to\infty} P_{X_n}(s) = P(s)$. Write $p_n(k) = \mathbb{P}(X_n = k)$.

(1) Since $(p_n(1))$ is a sequence in compact $[0, 1]$ there exists a subsequence $(p_{1,n}(1))$ of $(p_n(1))$ such that $p_{1,n}(1)$ converges.

(2) Since $(p_{1,n}(2))$ is a sequence in compact $[0, 1]$ there exists a subsequence $(p_{2,n}(2))$ of $(p_{1,n}(2))$ such that $p_{2,n}(2)$ converges.

$\vdots$

(k) Since $(p_{k-1,n}(k))$ is a sequence in compact $[0, 1]$ there exists a subsequence $(p_{k,n}(k))$ of $(p_{k-1,n}(k))$ such that $p_{k,n}(k)$ converges.

Then $p_{n,n}(k)$ converges for all k. Why?

$$(p_{n,n}(k)) = (p_{1,1}(k), \ldots, \underbrace{p_{k,k}(k), p_{k+1,k+1}(k), \ldots}_{\text{subsequence of } p_{k,k}(k), \text{ which converges}}).$$

Hence there exists a sequence n_j such that $\lim_{j\to\infty} p_{n_j}(k)$ exists.

Fact: If every convergent subsequence of a bounded sequence converges to the same limit, then the sequence itself converges and to the same limit as the subsequences.

Proof of the fact. By contradiction, suppose the sequence is not convergent. Then its liminf and limsup are finite but different, and there are respective subsequences that converge to these, which contradicts the assumption.

Fix n_j as above and let another (increasing) sequence n'_j be such that $\lim_{j\to\infty} p_{n'_j}(k)$ exists. *Claim*:

$$\lim_{j\to\infty} p_{n'_j}(k) = \lim_{j\to\infty} p_{n_j}(k).$$

Indeed, this completes the proof since by the fact above, the whole sequence therefore converges to the same limit as the subsequence $p_{n_j}(k)$ for all $k \geq 0$.

Proof of the claim. For all $s \in (0, 1)$ as $\lim_{n\to\infty} P_{X_n}(s) = P(s)$ we have that $P_{X_{n_j}}(s) \to P(s)$ and $P_{X_{n'_j}}(s) \to P(s)$. But

$$P_{X_{n_j}}(s) \to \sum_{k=0}^{\infty}\Big[\lim_{j\to\infty} p_{n_j}(k)\Big]s^k \qquad P_{X_{n'_j}}(s) \to \sum_{k=0}^{\infty}\Big[\lim_{j\to\infty} p_{n'_j}(k)\Big]s^k.$$

Hence

$$\sum_{k=0}^{\infty}\Big[\lim_{j\to\infty} p_{n_j}(k)\Big]s^k = \sum_{k=0}^{\infty}\Big[\lim_{j\to\infty} p_{n'_j}(k)\Big]s^k$$

by the uniqueness of limits in $\mathbb{R}$. Moreover, the limits agree by the uniqueness of moment generating functions. □

6.3. **Theorem** (Poisson Approximation of Binomial)**.** *Let $X_n \sim \text{Binomial}(n, p(n))$ such that $np(n) \to \lambda$ as $n \to \infty$. Then $X_n \xrightarrow{d} X$ where $X \sim \text{Poisson}(\lambda)$.*

Proof. By the continuity theorem if the generating functions converge then the distribution functions converge. Note that

$$P_{X_n}(s) = \sum_{r=0}^{n}\binom{n}{r}p^r q^{n-r}s^r = (q+ps)^n = (1+p(n)(s-1))^n = \left(1+\frac{np(n)(s-1)}{n}\right)^n \to e^{\lambda(s-1)}. \qquad \square$$

6.4. **Theorem** (Law of rare events)**.** *Consider the triangular array $\{X_{n,k} : n \in \mathbb{N}, 1 \le k \le n\}$ which looks like*

$$\begin{array}{llll} \{ & X_{1,1}, & & \\ & X_{2,1}, & X_{2,2}, & \\ & X_{3,1}, & X_{3,2}, & X_{3,3}, \quad \ldots\} \end{array}$$

Let $X_{n,k} \sim \text{Bernoulli}(p_{n,k})$ be independent and such that

- *$\max_{1\le k\le n} p_{n,k} \to 0$ as $n \to \infty$, and,*
- *$\sum_{k=1}^{n} p_{n,k} \to \lambda \in (0,\infty)$ as $n \to \infty$, that is, the expected number of successes in each row goes to λ,*

then $\sum_{k=1}^{n} X_{n,k} \xrightarrow{d} \text{Poisson}(\lambda)$.

Proof. Note that this is a generalisation of the previous theorem, where in that case each $X_{n,k} \sim \text{Bernoulli}(p)$. We show that the generating function converges to that of a Poisson then invoke the continuity theorem which then implies that it converges weakly to a Poisson random variable.

$$P_n(s) = \mathbb{E}s^{\sum_{k=1}^{n} X_{n,k}} = \prod_{k=1}^{n} P_{X_{n,k}}(s) = \prod_{k=1}^{n}(1-p_{n,k}+p_{n,k}s) = \prod_{k=1}^{n}(1-(1-s)p_{n,k}).$$

Taking log of both sides,

$$\log P_n(s) = \sum_{k=1}^{n}\log(1-(1-s)p_{n,k}).$$

Note that

$$\log(1-x) = -\sum_{n=1}^{\infty}\frac{x^n}{n} = -x - \sum_{n=2}^{\infty}\frac{x^n}{n},$$

and therefore for $x > 0$

$$\log(1-x) < -x$$

and

$$\log(1-x) > -x - \sum_{n=2}^{\infty}x^n \underset{\text{for } x<1}{=} -x - x^2\frac{1}{1-x} \underset{\text{for } x<1/2}{>} -x-2x^2.$$

Hence for $x \in (0, 1/2)$,

$$-x-2x^2 < \log(1-x) < -x.$$

Plugging this in above we get

$$(s-1)\lambda \leftarrow -\sum_{k=1}^{n}(1-s)p_{n,k} - \sum_{k=1}^{n}2(1-s)^2p_{n,k}^2 < \log P_n(s) < -\sum_{k=1}^{n}(1-s)p_{n,k} = \sum_{k=1}^{n}(s-1)p_{n,k} \to (s-1)\lambda.$$

On the left this comes from the fact that

$$\sum_{k=1}^{n}p_{n,k}^2 \leq \sum_{k=1}^{n}\max_{1\leq j\leq n}p_{n,j}p_{n,k} \to 0 \times \lambda. \qquad \square$$

7. Weak law of large numbers

7.1. **Theorem** (Markov's inequality)**.** *Let $X \geq 0$ be a random variable; $c > 0$. Then*

$$\mathbb{P}(X \geq c) \leq \frac{\mathbb{E}X}{c}.$$

Proof. Note that $X \geq c\mathbb{1}_{\{X\geq c\}}$. Taking expectations, the result follows immediately. $\square$

7.2. **Theorem** (Chebyshev's inequality)**.** *Let X (not necessarily non-negative) be a random variable with $\mathrm{Var}X = \sigma^2 < \infty$; $c > 0$. Then $\mathbb{E}X$ exists and is finite and*

$$\mathbb{P}(|X - \mathbb{E}X| \geq c) \leq \frac{\mathrm{Var}X}{c^2}.$$

Proof. Using Markov's,

$$\mathbb{P}(|X-\mathbb{E}X| \geq c) = \mathbb{P}(|X-\mathbb{E}X|^2 \geq c^2) \leq \frac{\mathbb{E}|X-\mathbb{E}X|^2}{c^2} = \frac{\mathrm{Var}X}{c^2}. \qquad \square$$

7.3. **Theorem** (Chernoff's inequality)**.** *Let X be a (not necessarily non-negative) random variables; $c \in \mathbb{R}$. Then*

$$\mathbb{P}(X \geq c) \leq \inf_{\lambda>0} e^{-\lambda c}M_X(\lambda).$$

Proof. By Markov's

$$\mathbb{P}(X \geq c) = \mathbb{P}(e^{\lambda X} \geq e^{\lambda c}) \leq \frac{\mathbb{E}e^{\lambda X}}{e^{\lambda c}} = e^{-\lambda c}M_X(\lambda).$$

Since this holds for all $\lambda > 0$ (for which $e^{\lambda x}$ is an increasing function) we can take infinum. $\square$

7.4. **Theorem** (Weak law of large numbers)**.** *Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables with finite mean $\mathbb{E}X_1 = \mu < \infty$. Then*

$$\frac{1}{n}\sum_{i=1}^{n}X_i \xrightarrow{P} \mu.$$

Proof. This proof requires an assumption that $\mathbb{E}X^2 < \infty$ also. The theorem is true without this assumption and is a corollary of the strong law of large numbers (later).

Since $\mathbb{E}(\frac{1}{n}\sum_{i=1}^n X_i) = \mu$ and by Chebyshev's,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n X_i - \mu\right| \geq \varepsilon\right) \leq \frac{\mathrm{Var}(\frac{1}{n}\sum_{i=1}^n X_i)}{\varepsilon^2} = \frac{\frac{1}{n^2}\times n\sigma^2}{\varepsilon^2} \to 0. \qquad \square$$

8. Stirling's formula, de Moirve-Laplace Central Limit Theorem

This bit is available as a printed pdf on the unit homepage.

9. Measure Theory

9.1. **Definition** (Power set)**.** The power set $\mathcal{P}(\Omega)$ of a set Ω is the set of all of its subsets $\mathcal{P}(\Omega) = \{A : A \subseteq \Omega\}$.

9.2. **Definition** (Algebra)**.** Let Ω be a set. A set $\mathcal{A} \subseteq \mathcal{P}(\Omega)$ (i.e., $\mathcal{A}$ is a set of subsets of Ω) is an *algebra* if

(a) $\Omega \in \mathcal{A}$;

(b) For all $A, B \in \mathcal{A}$, $A \cup B \in \mathcal{A}$;

(c) For all $A \in \mathcal{A}$, $\Omega \backslash A \in \mathcal{A}$.

9.3. **Definition** (Finitely additive measure)**.** A set function $\mu : \mathcal{A} \to [0, \infty]$ defined on an algebra $\mathcal{A}$ is a *finitely additive measure* on $\mathcal{A}$ if for all disjoint $A, B \in \mathcal{A}$ (i.e. $A \cap B = \emptyset$), $\mu(A \cup B) = \mu(A) + \mu(B)$.

9.4. **Definition** (σ-algebra)**.** Let Ω be a set. A set $\mathcal{F} \subseteq \mathcal{P}(\Omega)$ ($\mathcal{F}$ is a set of subsets of Ω) is a *σ-algebra* if

(a) $\Omega \in \mathcal{F}$;

(b) For any countable collection of sets $A_1, A_2, A_3, \ldots \in \mathcal{F}$, $\bigcup_{n=1}^\infty A_n \in \mathcal{F}$;

(c) For all $A \in \mathcal{F}$, $\Omega \backslash A \in \mathcal{F}$.

Note that (b) and (c) imply that for any countable collection of sets $A_1, A_2, A_3, \ldots \in \mathcal{F}$, $\bigcap_{n=1}^\infty A_n \in \mathcal{F}$ by De Morgan's law.

If Ω is a set and $\mathcal{F}$ a σ-algebra on Ω then we say that $(\Omega, \mathcal{F})$ is a *measurable space*.

9.5. **Definition** (Measure)**.** A set function $\mu : \mathcal{F} \to [0, \infty]$ is a *measure* on $\mathcal{F}$ if it is σ-additive. That is to say, for any countable collection of sets $A_1, A_2, A_3, \ldots \in \mathcal{F}$ which is pairwise disjoint ($A_i \cap A_j = \emptyset$ for all $i \neq j$), we have $\mu(\bigcup_{n=1}^\infty A_n) = \sum_{n=1}^\infty \mu(A_n)$.

9.6. **Definition.** A *probability measure* is a measure by the above definition $\mathbb{P} : \mathcal{F} \to [0, \infty]$ with the boundary condition $\mathbb{P}(\Omega) = 1$. Note that this implies $\mathbb{P}(\emptyset) = 0$ by the σ-additivity property.

9.7. **Theorem** (Continuity of probability and more)**.** *Let $\mathcal{A}$ be an algebra and $\mathbb{P} : \mathcal{A} \to [0, \infty]$ a finitely additive measure such that $\mathbb{P}(\Omega) = 1$. Then the following definitions are equivalent.*

(a) *$\mathbb{P}$ is a measure, that is, it satisfies the σ-additivity property;*

(b) *If A_n is an increasing sequence in $\mathcal{A}$, i.e., $A_n \subseteq A_{n+1}$ and $A_n \in \mathcal{A}$ for all $n \in \mathbb{N}$ then whenever $\bigcup_{n=1}^\infty A_n \in \mathcal{A}$ (this is not certain as $\mathcal{A}$ is not necessarily a σ-algebra) we have*

$$\lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}\left(\bigcup_{n=1}^\infty A_n\right) = \mathbb{P}(\lim_{n\to\infty} A_n),$$

where we define $\bigcup_{n=1}^\infty A_n =: \lim_{n\to\infty} A_n$;

(c) *If A_n is a decreasing sequence in $\mathcal{A}$, i.e., $A_{n+1} \subseteq A_n$ and $A_n \in \mathcal{A}$ for all $n \in \mathbb{N}$ then whenever $\bigcap_{n=1}^\infty A_n \in \mathcal{A}$ we have*

$$\lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}\left(\bigcap_{n=1}^\infty A_n\right) = \mathbb{P}(\lim_{n\to\infty} A_n),$$

where we define $\bigcap_{n=1}^{\infty} A_n =: \lim_{n\to\infty} A_n$*;*

(d) *If* A_n *is a decreasing sequence with* $\bigcap_{n=1}^{\infty} A_n = \lim_{n\to\infty} A_n = \emptyset$ *then* $\lim_{n\to\infty} \mathbb{P}(A_n) = 0 (= \mathbb{P}(\emptyset))$.

Proof. We first show that (a) implies (b). Let $B_1 = A_1$, $B_n = A_n \backslash A_{n-1}$ for all $n \geq 2$. Then $\{B_n\}_{n=1}^{\infty}$ satisfies $\bigcup_{n=1}^{\infty} B_n = \bigcup_{n=1}^{\infty} A_n$ and $B_i \cap B_j = \emptyset$ for all $i \neq j$. So $\{B_n\}_{n=1}^{\infty}$ forms a disjoint partition of $\bigcup_{n=1}^{\infty} A_n$. By the σ-additivity assumption in (a) we have that

$$\mathbb{P}(\bigcup_{n=1}^{\infty} A_n) = \mathbb{P}(\bigcup_{n=1}^{\infty} B_n) = \sum_{n=1}^{\infty} \mathbb{P}(B_n) = \lim_{n\to\infty} \sum_{k=1}^{n} \mathbb{P}(B_k) = \lim_{n\to\infty} \mathbb{P}(\bigcup_{k=1}^{n} B_k) = \lim_{n\to\infty} \mathbb{P}(A_n).$$

Now we show (b) implies (c). Since A_n is a decreasing sequence, $\Omega \backslash A_n$ is an increasing sequence. Moreover,

$$\mathbb{P}(\bigcap_{n=1}^{\infty} A_n) = 1 - \mathbb{P}(\bigcup_{n=1}^{\infty} \Omega \backslash A_n) = 1 - \lim_{n\to\infty} \mathbb{P}(\Omega \backslash A_n) = \lim_{n\to\infty} \mathbb{P}(A_n).$$

(c) implies (d) trivially, it is simply a special case. It remains to show that (d) implies (a), i.e., that if A_n is a decreasing sequence with $\bigcap_{n=1}^{\infty} A_n = \lim_{n\to\infty} A_n = \emptyset$ then $\lim_{n\to\infty} \mathbb{P}(A_n) = 0$ implies $\mathbb{P}$ has the σ-additivity property.

Take any disjoint family of sets $\{A_k\}_{k=1}^{\infty}$. Then by finite additivity,

$$\sum_{k=1}^{\infty} \mathbb{P}(A_k) = \lim_{n\to\infty} \sum_{k=1}^{n} \mathbb{P}(A_k) = \lim_{n\to\infty} \mathbb{P}(\bigcup_{k=1}^{n} A_k) = \lim_{n\to\infty} [\mathbb{P}(\bigcup_{k=1}^{\infty} A_k) - \mathbb{P}(\bigcup_{k=n+1}^{\infty} A_k)].$$

Now $\bigcup_{k=n+1}^{\infty} A_k$ is a decreasing sequence in n where $\bigcap_{n=1}^{\infty} \bigcup_{k=n+1}^{\infty} A_k = \emptyset$. This is because if $\omega \in \bigcup_{k=n+1}^{\infty} A_k$ then $\omega \in A_N$ for unique N (by disjointness of the family). Hence ω is not in the intersection of all tail unions. By (c), $\mathbb{P}(\bigcup_{k=n+1}^{\infty} A_k) \to 0$. Moreover, σ-additivity holds. □

9.8. *Remark.* If A_n is an increasing sequence then $\mathbb{P}(A_n)$ is an increasing sequence. Indeed this is because $A_n \subseteq A_{n+1}$ implies that $A_{n+1} = A_{n+1} \backslash A_n \cup A_n$. By the σ-additivity of $\mathbb{P}$ we have that $\mathbb{P}(A_{n+1}) = \mathbb{P}(A_{n+1} \backslash A_n) + \mathbb{P}(A_n) \geq \mathbb{P}(A_n)$. Moreover $\mathbb{P}(A_n) \nearrow \mathbb{P}(\lim_{n\to\infty} A_n)$.

This applies similarly to decreasing sequences. If A_n is decreasing then $\mathbb{P}(A_n)$ is a decreasing sequence and $\mathbb{P}(A_n) \searrow \mathbb{P}(\lim_{n\to\infty} A_n)$.

9.9. **Definition** (Probability space)**.** A *probability space* is a triplet $(\Omega, \mathcal{F}, \mathbb{P})$ where Ω is any set, $\mathcal{F}$ is a σ-algebra on Ω and $\mathbb{P}$ is a probability measure, that is a measure on $\mathcal{F}$ satisfying σ-additivity and normalised such that $\mathbb{P}(\Omega) = 1$.

9.10. **Lemma.** *Let* $A \subseteq \mathcal{P}(\Omega)$*, i.e. A is a collection of subsets of* Ω*. Then there exists a smallest algebra* $\alpha(A)$ *and a smallest* σ*-algebra* $\sigma(A)$ *which contains every set of A. We refer to these as the algebra and the* σ*-algebra generated by A respectively.*

Sketch proof. Let $\alpha(A)$ and $\sigma(A)$ be intersections of all algebras (respectively σ-algebras) which contain A. It is easy to check that this too is an algebra (respectively σ-algebra). □

9.11. **Definition.** On $\mathbb{R}$ we define the algebra

$$\mathcal{A} := \left\{ \bigcup_{k=1}^{n} (a_k, b_k] : 1 \leq n < \infty; a_1 < b_1 < a_2 < b_2 < \cdots ; a_k, b_k \in \mathbb{R} \cup \{-\infty, \infty\} \right\}$$

which is the set of all finite unions of subintervals. Note that $\mathcal{A}$ is not a σ-algebra. Indeed $(0, 1 - 1/n] \in \mathcal{A}$ for all $n \in \mathbb{N}$, however $\cup_{n=1}^{\infty} (0, 1 - 1/n] = (0, 1) \notin \mathcal{A}$. We define the Borel σ-algebra $\mathcal{B}(\mathbb{R}) := \sigma(\mathcal{A})$ to be the smallest σ-algebra which contains all finite unions of subintervals of the reals. The members of $\mathcal{B}(\mathbb{R})$ are called the Borel measurable sets of $\mathbb{R}$ and it is left as a remark that these contain any sets of suitable interest.

Similarly we define $\mathcal{B}(\mathbb{R}^n)$ to be the σ-algebra generated by all rectangle sets in $\mathbb{R}^n$ and $\mathcal{B}(\mathbb{R}^\infty)$ to be the σ-algebra generated by all cylinder sets of $\mathbb{R}^\infty$. Note that $\mathbb{R}^\infty$ is the set of all real valued sequences and a cylinder set is of the form $\{(x_1,\dots,x_n) : n \in \mathbb{N}; x_1 \in (a_1,b_1]; \dots x_n \in (a_n,b_n]; a_k < b_k \in \mathbb{R}\cup\{-\infty,\infty\}\}$. This again contains all sets of suitable interest. To name a few we mention

- $\{(x_n) : \lim_{n\to\infty} x_n$ exists and finite$\}$;
- $\{(x_n) : \sup_{n\in\mathbb{N}} x_n > a\}$;
- $\{(x_n) : \liminf_{n\to\infty} x_n > a\}$.

We can also define $\mathcal{B}(\mathbb{R}^T)$ for some uncountable set T (think time, reals). This contains all functions $X(t)$ although we may possibly have to restrict attention to càdlàg functions; these are continuous on the right and have a limit from the left.

9.12. **Theorem** (Carathéodory)**.** *Let $\mathcal{A}$ be an algebra on Ω; $\mathcal{F} = \sigma(\mathcal{A})$ the σ-algebra generated by $\mathcal{A}$. If μ is a σ-additive measure on $\mathcal{A}$ then there exists a unique extension of μ to the measurable space $(\Omega,\mathcal{F})$.*

9.13. **Definition** (Lebesgue measure)**.** Let $\mathcal{A}$ be the algebra (on $[0,1]$) containing disjoint intervals of the form $(a_1,b_1]\cup(a_2,b_2]\cup\dots\cup(a_k,b_k]\subset[0,1]$. Define the measure μ_0 such that

$$\mu_0((a_1,b_1]\cup(a_2,b_2]\cup\dots\cup(a_k,b_k]) = \sum_{i=1}^{k}(b_i-a_i).$$

It can be shown that μ_0 is σ-additive. Then by Carathéodory's extension theorem there exists a unique measure μ on the σ-algebra generated by $\mathcal{A}$, $\sigma(\mathcal{A})$ such that $\mu = \mu_0$ on $\mathcal{A}$. We define this μ to be the Lebesgue measure on $([0,1],\mathcal{B}[0,1])$.

9.14. **Definition.** A function $X : (\Omega,\mathcal{F}) \to \mathbb{R}$ is *$\mathcal{F}$-measurable* if for all $B \in \mathcal{B}(\mathbb{R})$, $X^{-1}(B) \in \mathcal{F}$. That is, $\mathcal{F}$ is rich enough so that for any Borel measurable set B the set of $\omega \in \Omega$ such that $X(\omega) \in B$ is in $\mathcal{F}$.

9.15. **Definition.** A *random variable* is an $\mathcal{F}$-measurable function $X : (\Omega,\mathcal{F},\mathbb{P}) \to \mathbb{R}$ where $\mathbb{P}$ is a probability measure on $\mathcal{F}$. Note that we require X to be $\mathcal{F}$-measurable (sometimes denoted $X \in m\mathcal{F}$) such that

$$\mathbb{P}(X\in B) = \mathbb{P}(X(\omega)\in B) = \mathbb{P}(\{\omega\in\Omega : X(\omega)\in B\}) = \mathbb{P}(X^{-1}(B)).$$

This is well defined since for all $B \in \mathcal{B}(\mathbb{R})$, $X^{-1}(B) \in \mathcal{F}$.

The *(cumulative) distribution function* of X is defined as

$$F : \mathbb{R} \to [0,1], \qquad F(x) = \mathbb{P}(\{X \le x\}) = \mathbb{P}(\{\omega\in\Omega : X(\omega)\le x\}).$$

This makes perfect sense according to the above as $\{\omega\in\Omega : X(\omega)\le x\} = X^{-1}\big((-\infty,x]\big) \in \mathcal{F}$.

9.16. **Corollary.** *Let X be a random variable. Then the distribution function of X is right continuous for all $x \in X(\Omega)$, i.e.*

$$F(\xi) \searrow F(x) \quad as \quad \xi \searrow x.$$

Proof. Let $x_n \searrow x$. Define $A_n = \{\omega\in\Omega : X(\omega)\le x_n\}$. Since x_n is a decreasing sequence, A_n is a decreasing sequence of events. Furthermore, $\lim_{n\to\infty} A_n = \{\omega\in\Omega : X(\omega)\le x\}$. Then by the continuity of probability, $F(x_n) = \mathbb{P}(A_n) \searrow \mathbb{P}(\lim_{n\to\infty} A_n) = F(x)$. □

9.17. **Corollary.** *Let X have distribution function F. Then*

$$\lim_{\xi\nearrow x} F(\xi) = \mathbb{P}(X < x).$$

Proof. Let x_n be a sequence such that $x_n \nearrow x$. Then define $A_n = \{\omega \in \Omega : X(\omega) \leq x_n\}$. Then $\lim_{n\to\infty} F(x_n) = \lim_{n\to\infty} \mathbb{P}(X \leq x_n) = \lim_{n\to\infty} \mathbb{P}(A_n)$. Since x_n is an increasing sequence, $A_n \subseteq A_{n+1}$, so A_n is an increasing sequence. Hence $\lim_{n\to\infty} A_n = \bigcup_n A_n$ which is the event for at least one n, $X(\omega) \leq x_n$, which is true if and only if $X(\omega) < x$. Hence $\lim_{n\to\infty} F(x_n) = \mathbb{P}(\lim_{n\to\infty} A_n) = \mathbb{P}(X < x)$. □

9.18. **Example.** Let $\mathbb{P}((a,b]) = F(b) - F(a)$ where F is a cumulative distribution function. Then F is continuous from the right. This measure is σ-additive on the algebra of sub-intervals of $\mathbb{R}$. Furthermore, it extends to the Lebesgue-Stieltjes measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, the measurable space $\mathbb{R}$ with the Borel σ-algebra on $\mathbb{R}$.

Intuition: To every distribution function there is an associated measure on $\mathbb{R}$.

9.19. *Remark* (Linking a random variable with specified distribution to a probability space). Let F be a distribution function, suppose for now that it is a continuous and strictly increasing distribution. Then consider the probability space $([0,1], \mathcal{B}[0,1], \mu)$ where μ is the Lebesgue measure (which generalises the notion of length). Note that here μ is also a probability measure.

Let $X : \Omega \to \mathbb{R}$ be defined by $X(\omega) = F^{-1}(\omega)$. Claim: X is $\mathcal{B}[0,1]$ measurable and X has distribution function given by F. Indeed for all $A \in \mathcal{B}(\mathbb{R})$ we have that $X^{-1}(A) = \{\omega \in [0,1] : X(\omega) \in A\} = \{\omega \in [0,1] : F^{-1}(\omega) \in A\} = F(A)$.

Furthermore $\mu(X \in (a,b]) = \mu(\{\omega \in \Omega : X(\omega) \in (a,b]\}) = \mu(\{\omega \in \Omega : F^{-1}(\omega) \in (a,b]\}) = \mu(F(a), F(b)] = F(b) - F(a)$. The penultimate equality follows from the increasing property of F. Hence X has distribution function given by X.

Also claim that if X is stochastically dominated by Y then on this space we have that $X(\omega) \leq Y(\omega)$ for all $\omega \in \Omega$. X stoch. dom by Y implies that $F_Y(x) \leq F_X(x)$ for all $x \in \mathbb{R}$. We claim that this implies $F_X^{-1}(\omega) \leq F_Y^{-1}(\omega)$ for all ω.

Suppose not, and there exists ω such that $F_Y^{-1}(\omega) < F_X^{-1}(\omega)$. Let $a := F_Y^{-1}(\omega)$. Then

$$F_X(a) = F_X(F_Y^{-1}(\omega)) < F_X(F_X^{-1}(\omega)) = \omega = F_Y(a).$$

The inequality follows from the fact that F_X is increasing. This contradicts X stoch. dom by Y.

In the case where the distribution function is not continuous or not strictly increasing we define

$$X^-(\omega) = \inf\{x \in \mathbb{R} : F(x) \geq \omega\} \quad \text{and} \quad X^+(\omega) = \inf\{x \in \mathbb{R} : F(x) > \omega\}$$

to be the first values for which $F(x)$ is greater than of equal to (respectively strictly greater than) ω. Each of these has distribution F and are equal almost surely.

9.20. **Theorem.** *For all random variables X there exists a sequence of simple random variables $X_1, X_2, \ldots$ such that $|X_n| \leq |X|$ almost surely and for all $\omega \in \Omega$ we have $X_n(\omega) \to X(\omega)$ as $n \to \infty$. A simple random variable is of the form*

$$X_n = \sum_{k=1}^{n} x_k \mathbb{1}_{A_k}$$

for a sequence of sets $A_k \in \mathcal{F}$, $x_k \in \mathbb{R}$. That is, a simple random variable is constant on finitely many measurable sets. If $X \geq 0$ for all $\omega \in \Omega$ then this sequence of simple random variables can be chosen to satisfy $X_n(\omega) \nearrow X(\omega)$.

No proof but some intuition. In the discrete case this is obvious. There exists a finite partition of $\mathcal{F}$ for which X is constant on each element; if not, then $X \notin m\mathcal{F}$. If this finest partition is finite, then X is itself a simple random variable. If not (but the finest partition is countable), then define a sequence of simple random variables, where X_n is the simple random variable taking constant values on the first n sets in the finest partition. Then $\lim_{n\to\infty} \sum_{k=1}^{n} x_k \mathbb{1}_{A_k}(\omega) = X(\omega)$. Futhermore, if X is non-negative then $\sum_{k=1}^{n} x_k \mathbb{1}_{A_k}(\omega) \leq \sum_{k=1}^{n+1} x_k \mathbb{1}_{A_k}(\omega)$ for all $\omega \in \Omega$ and the sequence increases to $X(\omega)$.

In the uncountable case and non-discrete X we consider an analogue of this intuition.

9.21. **Lemma.** *If $X_1, X_2, \dots$ are random variables and for all $\omega \in \Omega$ we have that $\lim_{n\to\infty} X_n(\omega)$ exists then the function $X(\omega) := \lim_{n\to\infty} X_n(\omega)$ is also a random variable. If X and Y are random variables then so are $X+Y$, XY, X/Y and $\varphi(X)$ for all measurable φ.*

9.22. **Definition** (Expectation)**.** We define expectation of a random variable in steps.

(1) We define $\mathbb{E}\mathbb{1}_A = \mathbb{P}(A)$ for all $A \in \mathcal{F}$.

(2) For a simple random variable $X_n = \sum_{k=1}^n x_k \mathbb{1}_{A_k}$ we say $\mathbb{E}X_n = \sum_{k=1}^n x_k \mathbb{P}(A_k)$.

(3) For a non-negative random variable X there exists a sequence of simple random variables X_n such that $X_n \nearrow X$. Then we define $\mathbb{E}X = \lim_{n\to\infty} \mathbb{E}X_n$. Note that $\mathbb{E}X_n \nearrow \mathbb{E}X$.

(4) For a general random variable we let $X = X^+ - X^-$ for some non-negative random variables X^+ and X^-. We then define $\mathbb{E}X = \mathbb{E}(X^+ - X^-) = \mathbb{E}X^+ - \mathbb{E}X^-$.

Remark. An equivalent way of describing expectations of real-valued random variables is integrating w.r.t. the Lebesgue-Stieltjes measure introduced in Example 9.18:

$$\mathbb{E}X = \int_{\mathbb{R}} x\, dF(x), \qquad \text{or} \qquad \mathbb{E}g(X) = \int_{\mathbb{R}} g(x)\, dF(x),$$

where F is the distribution function of the random variable X. Intuitively,

$$\sum_i g(x_i)(F(x_{i+1}) - F(x_i)) \to \int_{\mathbb{R}} g(x) dF(x)$$

with a proper refining of the mesh $\{x_i\}$. The increment of F in the sum approximates the probability that X is close to the point x_i.

10. Toolbox

10.1. **Lemma** (The First Borel-Cantelli Lemma)**.** *Let $A_1, A_2, \dots$ be a sequence of events such that $\sum_{n=1}^\infty \mathbb{P}(A_n) < \infty$, then*

$$\mathbb{P}\Big(\bigcap_{n=1}^\infty \bigcup_{k=n}^\infty A_k\Big) = 0.$$

10.2. *Remark* (Interpretation). $\sum_{n=1}^\infty \mathbb{P}(A_n)$ is the expected number of A_ns which are true because

$$\mathbb{E}\sum_{n=1}^\infty \mathbb{1}_{A_n} = \sum_{n=1}^\infty \mathbb{P}(A_n).$$

So the expected number of A_n to occur is finite.

$\bigcup_{k=n}^\infty A_k$ is the event that after n, at least one of $A_n, A_{n+1}, \dots$ occurs. Thus $\bigcap_{n=1}^\infty \bigcup_{k=n}^\infty A_k$ is the event that for all n, one of the A_k, $k \geq n$ occurs. So $\bigcap_{n=1}^\infty \bigcup_{k=n}^\infty A_k$ is the event that infinitely many occur.

Notation. $\bigcap_{n=1}^\infty \bigcup_{k=n}^\infty A_k = \lim_{n\to\infty} \bigcup_{k=n}^\infty A_k = \limsup_{n\to\infty} A_n$.

Borel-Cantelli 1 in English: if the expected number of A_n to occur is finite, then the probability infinitely many occur is 0; or finitely many occur a.s.

Proof. $\bigcup_{k=n}^\infty A_k \subseteq \bigcup_{k=n-1}^\infty A_k$ so $\bigcup_{k=n}^\infty A_k$ is a decreasing sequence in n. Thus

$$\mathbb{P}\Big(\bigcap_{n=1}^\infty \bigcup_{k=n}^\infty A_k\Big) = \lim_{n\to\infty} \mathbb{P}\Big(\bigcup_{k=n}^\infty A_k\Big) \leq \lim_{n\to\infty} \sum_{k=n}^\infty \mathbb{P}(A_k) = 0,$$

since the tail of a convergent sum converges to 0. □

10.3. **Lemma** (The Second Borel-Cantelli Lemma)**.** *Let $A_1, A_2, \dots$ be a sequence of independent events such that $\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty$, then*

$$\mathbb{P}\Big(\bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k \Big) = 1.$$

10.4. *Remark* (Interpretation). Partial converse to Borel-Cantelli lemma 1, but requires that the events are independent.

$\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty$ so the expected number of A_n to occur is infinite.

Borel-Cantelli 2 in English: if the expected number of independent events A_n to occur is infinite, then infinitely many occur a.s.

Special case: if an event occurs with constant probability p, then $\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty$ and hence the event occurs a.s. and infinitely often.

Proof. By De Morgan's laws,

$$\mathbb{P}\Big(\bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k \Big) = 1 - \mathbb{P}\Big(\Big[\bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k \Big]^c \Big) = 1 - \mathbb{P}\Big(\bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} A_k^c \Big).$$

Now $\bigcap_{k=n}^{\infty} A_k^c \subseteq \bigcap_{k=n+1}^{\infty} A_k^c$ so it is an increasing sequence. Thus

$$\mathbb{P}\Big(\bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k \Big) = 1 - \mathbb{P}\Big(\bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} A_k^c \Big) = 1 - \lim_{n\to\infty} \mathbb{P}\Big(\bigcap_{k=n}^{\infty} A_k^c \Big) = 1 - \lim_{n\to\infty} \prod_{k=n}^{\infty} \mathbb{P}(A_k^c),$$

by independence. Furthermore, using that for all $x \in \mathbb{R}$, $1 - x \le e^{-x}$,

$$\begin{aligned} \mathbb{P}\Big(\bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k \Big) &= 1 - \lim_{n\to\infty} \prod_{k=n}^{\infty} \mathbb{P}(A_k^c) = 1 - \lim_{n\to\infty} \prod_{k=n}^{\infty} [1 - \mathbb{P}(A_k)] \\ &\ge 1 - \lim_{n\to\infty} \prod_{k=n}^{\infty} e^{-\mathbb{P}(A_k)} = 1 - \lim_{n\to\infty} \underbrace{e^{-\sum_{k=n}^{\infty} \mathbb{P}(A_k)}}_{=0}, \end{aligned}$$

since the tail of any infinite sum is also infinite. Hence $\mathbb{P}(\bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k) \ge 1$, which implies equality. □

10.5. **Theorem** (Monotone Convergence Theorem)**.** *Let $X, Y, X_1, X_2, \dots$ be random variables. Then,*

(a) *If $X_n \ge Y$ for all n; $\mathbb{E}Y > -\infty$ and $X_n \nearrow X$ as $n \to \infty$ then $\mathbb{E}X_n \to \mathbb{E}X$.*

(b) *If $X_n \le Y$ for all n; $\mathbb{E}Y < \infty$ and $X_n \searrow X$ as $n \to \infty$ then $\mathbb{E}X_n \to \mathbb{E}X$.*

Proof. Proof of (a) only; (b) follows similarly.

Suppose $Y \equiv 0$, i.e., $\forall \omega \in \Omega$, $Y(\omega) = 0$. Then as $X_n \ge 0$ (seen in the measure theory section) for each X_k there exists a sequence of simple (that is, constant on finitely many measurable sets) random variables $X_k^{(n)}$

such that $X_k^{(n)} \nearrow X_k$ as $n \to \infty$.

$$\begin{array}{ccccccccc} X_1 & \leq & X_2 & \leq & X_3 & \leq & \cdots & \leq & X \\ \vee| & & \vee| & & \vee| & & & & \\ \vdots & & \vdots & & \vdots & & & & \\ \vee| & & \vee| & & \vee| & & & & \\ X_1^{(3)} & & X_2^{(3)} & & X_3^{(3)} & & & & \\ \vee| & & \vee| & & \vee| & & & & \\ X_1^{(2)} & & X_2^{(2)} & & X_3^{(2)} & & & & \\ \vee| & & \vee| & & \vee| & & & & \\ X_1^{(1)} & & X_2^{(1)} & & X_3^{(1)} & & & & \end{array}$$

Define $Z^{(n)} := \max_{1 \leq j \leq n} X_j^{(n)}$. That is, $Z^{(n)}$ is the maximum value of the first n terms in the nth row from the bottom in the table above.

Properties of $Z^{(n)}$:

- For all $1 \leq k \leq n$ we have $X_k^{(n)} \leq Z^{(n)} \leq X_n$. The first inequality follows immediately from the definition of $Z^{(n)}$, it is simply the maximum of such values, and is hence an upper bound. The second inequality follows from chasing the column up within which the maximum lies and then across to the value X_n. Formally, for some $1 \leq k \leq n$, $Z^{(n)} = X_k^{(n)} \leq X_k \leq X_n$.
- $Z^{(n-1)} \leq Z^{(n)}$. Why? $Z^{(n-1)} = \max_{1 \leq j \leq n-1} X_j^{(n-1)} \leq \max_{1 \leq j \leq n-1} X_j^{(n)} \leq Z^{(n)}$. The inequality follows from the fact that for all $j \in \mathbb{N}$ we have $X_j^{(n-1)} \leq X_j^{(n)}$, and then we maximise over a larger domain.

Define $Z := \lim_{n \to \infty} Z^{(n)}$, which exists because $Z^{(n)}$ is an increasing sequence (the limit may possibly be infinite).

Since for all $1 \leq k \leq n$ we have $X_k^{(n)} \leq Z^{(n)} \leq X_n$, taking $n \to \infty$ we see that

$$\lim_{n \to \infty} X_k^{(n)} \leq \lim_{n \to \infty} Z^{(n)} \leq \lim_{n \to \infty} X_n \implies X_k \leq Z \leq X \underbrace{\implies}_{k \to \infty} Z = X.$$

Note that, since the $Z^{(n)}$s are simple (indeed they are a maximum of simple random variables), by the definition of expectation of a limit simple random variables,

$$\mathbb{E}X = \mathbb{E}Z = \mathbb{E} \lim_{n \to \infty} Z^{(n)} = \lim_{n \to \infty} \mathbb{E}Z^{(n)} \leq \lim_{n \to \infty} \mathbb{E}X_n.$$

Thus it remains to show that $\mathbb{E}X \geq \lim_{n \to \infty} \mathbb{E}X_n$. Since $X_n \nearrow X$ we have that $X_n \leq X$ for all n, which implies that $\mathbb{E}X_n \leq \mathbb{E}X$. Hence,

$$\lim_{n \to \infty} \mathbb{E}X_n \leq \mathbb{E}X.$$

In the case where $Y \not\equiv 0$ then we repeat the above analysis with $X_n - Y$ which is a non-negative random variable. □

10.6. **Corollary.** *If $X_n \geq 0$ then $\mathbb{E} \sum_{n=1}^{\infty} X_n = \sum_{n=1}^{\infty} \mathbb{E}X_n$.*

Proof. By definition we have $\sum_{n=1}^{\infty} X_n = \lim_{N \to \infty} \sum_{n=1}^{N} X_n$. Define $Y_N = \sum_{n=1}^{N} X_n$. Since $X_n \geq 0$, $Y_N \geq 0$ for all N; Y_N is an increasing sequence; and $Y_N \nearrow Y := \sum_{n=1}^{\infty} X_n$. Moreover, by the monotone convergence theorem, $\lim_{N \to \infty} \mathbb{E}Y_N = \mathbb{E}Y = \mathbb{E} \sum_{n=1}^{\infty} X_n$. □

10.7. **Theorem** (Fatou's Lemma)**.** *Let $X_1, X_2, \ldots; Y$ be random variables such that $X_n \geq Y$ and $\mathbb{E}Y > -\infty$. Then*

$$\mathbb{E} \liminf_{n \to \infty} X_n \leq \liminf_{n \to \infty} \mathbb{E}X_n.$$

10.8. *Remark* (Acronym). A good way to remember this is that we always write in the $\leq$ direction and together it initialises 'Elle'.

Proof. Define $Z_n := \inf_{m\geq n} X_m$. Then Z_n is an increasing sequence; indeed $\inf_{m\geq n} X_m \leq \inf_{m\geq n+1} X_m$ since the infinum is over a larger domain. Furthermore, $Z_n \nearrow Z := \liminf_{n\to\infty} X_n$; this follows from the fact that Z_n is increasing and by definition

$$\liminf_{n\to\infty} X_n = \lim_{n\to\infty} (\inf_{m\geq n} X_m) = \lim_{n\to\infty} Z_n.$$

Now $Z_n = \inf_{m\geq n} X_m \geq Y$ as $X_m \geq Y$ for all $m \in \mathbb{N}$. Thus we are in good shape to apply the monotone convergence theorem:

$$\lim_{n\to\infty} \mathbb{E}Z_n = \mathbb{E}Z = \mathbb{E}\liminf_{n\to\infty} X_n$$

But on the left-hand-side, as the limit exists it is equal to the lim inf. Now $Z_n = \inf_{m\geq n} X_m \leq X_n$ and thus

$$\mathbb{E}\liminf_{n\to\infty} X_n = \lim_{n\to\infty} \mathbb{E}Z_n = \liminf_{n\to\infty} \mathbb{E}Z_n \leq \liminf_{n\to\infty} \mathbb{E}X_n. \qquad \square$$

10.9. **Theorem** (Dominated Convergence Theorem)**.** *Let $X, Y, X_1, X_2, \ldots$ be random variables such that $|X_n| \leq Y$; $\mathbb{E}Y < \infty$; and $X_n \xrightarrow{a.s.} X$ as $n \to \infty$. Then*

(a) $\mathbb{E}|X| < \infty$;

(b) $\mathbb{E}X_n \to \mathbb{E}X$ *as* $n \to \infty$;

(c) $\mathbb{E}|X_n - X| \to 0$.

Proof. To prove (a) note that $X_n \xrightarrow{\text{a.s.}} X$ as $n \to \infty$ implies that $|X_n| \xrightarrow{\text{a.s.}} |X|$ as $n \to \infty$ (mod is a continuous function). Furthermore, since $|X_n| \leq Y$, we have that $|X| = \lim_{n\to\infty} |X_n| \leq Y$ almost surely.

To prove (b) we construct the following chain of inequalities

$$\mathbb{E}X = \mathbb{E}\lim_{n\to\infty} X_n = \mathbb{E}\liminf_{n\to\infty} X_n \underset{*}{\leq} \liminf_{n\to\infty} \mathbb{E}X_n \leq \limsup_{n\to\infty} \mathbb{E}X_n \underset{**}{\leq} \mathbb{E}\limsup_{n\to\infty} X_n = \mathbb{E}\lim_{n\to\infty} X_n = \mathbb{E}X,$$

where $*$ follows from Fatou's lemma and $**$ follows from Fatou's lemma on $-X_n$; indeed we have the relation $\liminf_{n\to\infty}(-X_n) = -\limsup_{n\to\infty} X_n$. Thus equality holds throughout the chain and we conclude that

$$\underbrace{\liminf_{n\to\infty} \mathbb{E}X_n = \limsup_{n\to\infty} \mathbb{E}X_n}_{\implies = \lim_{n\to\infty} \mathbb{E}X_n} = \mathbb{E}X.$$

Since the lim inf and lim sup agree, $\lim_{n\to\infty} \mathbb{E}X_n$ exists and is equal to $\mathbb{E}X$. Thus

$$\lim_{n\to\infty} \mathbb{E}X_n = \mathbb{E}X (\underset{\star}{=} \mathbb{E}\lim_{n\to\infty} X_n),$$

where $\star$ follows from the fact that $\mathbb{E}\lim_{n\to\infty} X_n$ is also an element in the chain. This proves (b).

Now to prove (c) we note that $|X_n - X| \leq |X_n| + |X| \leq 2Y$. Then we repeat the analysis above with $|X_n - X|$ and bounding random variable $Y \equiv 2Y$. $\square$

10.10. **Theorem** (Cauchy-Schwarz Inequality)**.** *If X and Y are random variables with $\mathbb{E}X^2 < \infty$ and $\mathbb{E}Y^2 < \infty$ then*

$$\mathbb{E}|XY| \leq \sqrt{\mathbb{E}X^2 \mathbb{E}Y^2}.$$

Proof. A corollary of Hölder's inequality — later. $\square$

10.11. **Corollary.** *Let X be a random variable. Then X has finite variance if and only if X has finite second moment.*

Proof. Let $\mathbb{E}X^2 < \infty$. Then $\mathrm{Var}X = \mathbb{E}X^2 - (\mathbb{E}X)^2 \leq \mathbb{E}X^2 < \infty$. (Indeed $\mathbb{E}X^2 < \infty$ implies $\mathbb{E}X \leq \mathbb{E}|X| < \infty$ by Cauchy-Schwarz.) If a random variable has infinite mean, then it has infinite second moment and thus infinite variance. Thus finite variance implies finite mean. These together implies finite second moment. □

10.12. **Theorem** (Jensen's inequality)**.** *If $g : \mathbb{R} \to \mathbb{R}$ is convex and X is a random variable with $\mathbb{E}|X| < \infty$ then*

$$g(\mathbb{E}X) \leq \mathbb{E}g(X).$$

10.13. *Remark.* This is easy to remember from the definition of convexity: Any line segment between two points on the curve lies entirely above the curve i.e., for $0 \leq \alpha = 1 - \beta \leq 1$, $g(\alpha x_1 + \beta x_2) \leq \alpha g(x_1) + \beta g(x_2)$. Then we extend this definition in the natural way.

Proof. Since g is convex the tangent to the curve at all points is a lower bound for the curve. Hence for all $x_0 \in \mathbb{R}$ there exists λ such that $g(x) \geq g(x_0) + \lambda(x - x_0)$.

Hence $g(X) \geq g(x_0) + \lambda(X - x_0)$; in particular, for $x_0 = \mathbb{E}X \in \mathbb{R}$,

$$g(X) \geq g(\mathbb{E}X) + \lambda(X - \mathbb{E}X).$$

Note that λ is a constant which depends on $\mathbb{E}X$ only (the gradient of the tangent is dependent only on the position on the curve x_0) and thus it is constant with respect to the expectation of X. Therefore

$$\mathbb{E}g(X) \geq \mathbb{E}g(\mathbb{E}X) + \mathbb{E}[\lambda(X - \mathbb{E}X)] = g(\mathbb{E}X) + \lambda[\mathbb{E}X - \mathbb{E}X] = g(\mathbb{E}X).$$ □

10.14. **Theorem** (Ljapunov's Inequality)**.** *For all $0 < s < t$,*

$$(\mathbb{E}|X|^s)^{\frac{1}{s}} \leq (\mathbb{E}|X|^t)^{\frac{1}{t}}.$$

Proof.

$$(\mathbb{E}|X|^s)^{\frac{1}{s}} = [(\mathbb{E}|X|^s)^{\frac{t}{s}}]^{\frac{1}{t}} \underset{\star\star}{\leq} [\mathbb{E}(|X|^s)^{\frac{t}{s}}]^{\frac{1}{t}} = (\mathbb{E}|X|^t)^{\frac{1}{t}},$$

where $\star\star$ follows from Jensen's inequality on the function $(\,\cdot\,)^{\frac{t}{s}} : \mathbb{R} \to \mathbb{R}$; which is convex since $t > s$ and therefore $\frac{t}{s} > 1$. □

10.15. **Theorem** (Hölder's inequality)**.** *Let $p, q > 1$ be such that $\frac{1}{p} + \frac{1}{q} = 1$. If the random variables X and Y are such that $\mathbb{E}|X|^p < \infty$ and $\mathbb{E}|Y|^q < \infty$ then*

$$\mathbb{E}|XY| \leq (\mathbb{E}|X|^p)^{\frac{1}{p}} \cdot (\mathbb{E}|Y|^q)^{\frac{1}{q}}.$$

10.16. *Remark* (Cauchy-Schwarz). Note that with $p = q = 2$, Hölder's inequality is precisely the Cauchy-Schwarz inequality.

Proof. Since $\log : \mathbb{R} \to \mathbb{R}$ is a concave function and $\frac{x}{p} + \frac{y}{q}$ is a convex combination of x and y we have that

$$\log\left(\frac{x}{p} + \frac{y}{q}\right) \geq \frac{\log(x)}{p} + \frac{\log(y)}{q} \implies \frac{x}{p} + \frac{y}{q} \geq x^{\frac{1}{p}} \cdot y^{\frac{1}{q}},$$

since the exponential function is increasing. Let $x = \frac{|X|^p}{\mathbb{E}|X|^p}$ and $y = \frac{|Y|^q}{\mathbb{E}|Y|^q}$. Then

$$\frac{1}{p}\frac{|X|^p}{\mathbb{E}|X|^p} + \frac{1}{q}\frac{|Y|^q}{\mathbb{E}|Y|^q} \geq \frac{|X|}{(\mathbb{E}|X|^p)^{\frac{1}{p}}} \cdot \frac{|Y|}{(\mathbb{E}|Y|^q)^{\frac{1}{q}}}.$$

Taking expectations of both sides,

$$1 \geq \frac{\mathbb{E}|XY|}{(\mathbb{E}|X|^p)^{\frac{1}{p}}(\mathbb{E}|Y|^q)^{\frac{1}{q}}}.$$ □

10.17. **Definition** (L^p-norm)**.** Let X be a random variable with finite pth absolute moment. We denote $\|X\|_p := (\mathbb{E}|X|^p)^{\frac{1}{p}}$ for $p \geq 1$ and call it the L^p-norm of X. It is a norm on the space of $\mathcal{F}$-measurable functions $X : \Omega \to \mathbb{R}$.

In this notation the previous results reduce to the following.

- Ljapunov: for $0 < s < t$, $\|X\|_s \leq \|X\|_t$;
- Cauchy-Schwarz: $\mathbb{E}|XY| \leq \|X\|_2\|Y\|_2$;
- Hölder: for $p, q > 1$ such that $\frac{1}{p} + \frac{1}{q} = 1$, $\mathbb{E}|XY| \leq \|X\|_p\|Y\|_q$.

10.18. **Theorem.** *$\|\cdot\|_p$, $p \geq 1$ defines a true norm on the space of $\mathcal{F}$-measurable functions $X : \Omega \to \mathbb{R}$ with finite pth moment.*

Proof. Clearly for all $X \in m\mathcal{F}$, $\|X\|_p \geq 0$. Moreover, equality holds if and only if $X = 0$ almost surely. Let $\lambda \in \mathbb{R}$. Then $\|\lambda X\|_p = (\mathbb{E}|\lambda X|^p)^{\frac{1}{p}} = (\lambda^p\mathbb{E}|X|^p)^{\frac{1}{p}} = \lambda\|X\|_p$. Thus it remains to prove the triangle inequality. See Minkowski inequality below. □

10.19. **Theorem** (Minkowski's Inequality)**.** *For any random variables X and Y and for all $p \geq 1$, the triangle inequality for the L^p-norm holds, i.e.,*

$$\|X + Y\|_p \leq \|X\|_p + \|Y\|_p.$$

Proof. If either (or both) $\|X\|_p = \infty$ or $\|Y\|_p = \infty$ then the inequality holds trivially. Hence suppose that $\|X\|_p < \infty$ and $\|Y\|_p < \infty$.

For the case $p = 1$ this is trivial and follows immediately from the triangle inequality of the mod.

Now consider the case $p > 1$. Define $F(x) = (a + x)^p - 2^{p-1}(a^p + x^p)$, $x > 0$; where $a > 0$ is some constant. This has the derivative

$$F'(x) = p(a + x)^{p-1} - 2^{p-1}px^{p-1},$$

and so F is stationary at $x = a$. Furthermore,

$$F'(x) > 0 \iff p(a + x)^{p-1} - 2^{p-1}px^{p-1} > 0 \iff \left(\frac{a + x}{2x}\right)^{p-1} > 1 \iff x < a.$$

Similarly, $F'(x) = 0$ if and only if $x = a$ and $F'(x) < 0$ if and only if $x > a$. Thus F is an increasing function for $x < a$; reaching a global maximum at $x = a$ and then decreasing for $x > a$. Therefore, $F(x) \leq F(a) = 0$ for all $x \in \mathbb{R}$. We therefore have the inequality

$$(a + x)^p \leq 2^{p-1}(a^p + x^p)$$

for all $a > 0$, $x > 0$, $p > 1$. Applying this:

$$|X + Y|^p \leq (|X| + |Y|)^p \leq 2^{p-1}(|X|^p + |Y|^p).$$

Taking expectations,

$$\mathbb{E}|X + Y|^p \leq 2^{p-1}(\mathbb{E}|X|^p + \mathbb{E}|Y|^p) < \infty^\star$$

since we assumed that both $\|X\|_p < \infty$ and $\|Y\|_p < \infty$. This verifies that $\|X + Y\|_p < \infty$, that is, $X + Y \in L^P(\Omega)$. Now we prove the Minkowski inequality.

$$\mathbb{E}|X + Y|^p = \mathbb{E}(|X + Y||X + Y|^{p-1}) \underset{*}{\leq} \mathbb{E}\big[\big(|X| + |Y|\big)|X + Y|^{p-1}\big] = \mathbb{E}(|X||X + Y|^{p-1}) + \mathbb{E}(|Y||X + Y|^{p-1}).$$

Above, $*$ follows from the triangle inequality on $\mathbb{R}$. Let q be such that $\frac{1}{p}+\frac{1}{q}=1$; this implies $q=p/(p-1)$. By Hölder's inequality

$$\mathbb{E}(|X||X+Y|^{p-1}) \leq (\mathbb{E}|X|^p)^{\frac{1}{p}}(\mathbb{E}|X+Y|^{(p-1)q})^{\frac{1}{q}} = (\mathbb{E}|X|^p)^{\frac{1}{p}}(\underbrace{\mathbb{E}|X+Y|^p}_{<\infty \text{ by } \star})^{\frac{1}{q}} = \|X\|_p \|X+Y\|_p^{\frac{p}{q}},$$

$$\mathbb{E}(|Y||X+Y|^{p-1}) \leq (\mathbb{E}|Y|^p)^{\frac{1}{p}}(\mathbb{E}|X+Y|^{(p-1)q})^{\frac{1}{q}} = (\mathbb{E}|Y|^p)^{\frac{1}{p}}(\mathbb{E}|X+Y|^p)^{\frac{1}{q}} = \|Y\|_p \|X+Y\|_p^{\frac{p}{q}}.$$

Plugging this into the above yields that,

$$\underbrace{\mathbb{E}|X+Y|^p}_{=\|X+Y\|_p^p} \leq \big(\|X\|_p + \|Y\|_p\big)\|X+Y\|_p^{\frac{p}{q}}.$$

Dividing through by $\|X+Y\|_p^{\frac{p}{q}} (\geq 0)$ and noting that $p-p/q=1$ by the definition of q (multiply by p), this is precisely the Minkowski inequality. □

11. Modes of Convergence

11.1. **Definition** (Modes of convergence)**.** Let $X_1, X_2, X_3, \ldots$ be a sequence of random variables. We say that

(a) X_n converges *weakly* (or *in distribution*) to X, written $X_n \xrightarrow{w} X$ or $X_n \xrightarrow{d} X$ or $X_n \Rightarrow X$, if and only if for every bounded and continuous function f on the image of X we have $\mathbb{E}f(X_n) \to \mathbb{E}f(X)$.

Note that this is equivalent to $F_{X_n}(x) \to F_X(x)$ (the distribution functions for X_n and X respectively) for all x for which $F_X(x)$ is continuous.

Remark. We do not require a common probability space, only that the distribution functions converge.

(b) X_n converges *in probability* to X, written $X_n \xrightarrow{P} X$, if for all $\varepsilon > 0$, $\mathbb{P}(|X_n - X| \geq \varepsilon) \to 0$ as $n \to \infty$.

That is, the subset of the sample space Ω for which the difference between X_n and X is larger than arbitrary $\varepsilon > 0$ has probability going to 0.

Remark. Note that here we do require X_n and X to be on a common probability space (i.e. we can couple the outcome of X_n and X for a realisation $\omega \in \Omega$).

(c) X_n converges *almost surely* (or *strongly*) to X, written $X_n \xrightarrow{\text{a.s.}} X$ if and only if $\mathbb{P}(X_n \to X) = 1$.

Note that $X_n \to X$ is an event that can be interpreted as follows. On a realisation $\omega \in \Omega$, we define a sequence of real numbers $X_1(\omega), X_2(\omega), \ldots$ and $X(\omega)$. If $X_n(\omega) \to X(\omega)$ then the event $X_n \to X$ has occurred. We say $X_n \xrightarrow{\text{a.s.}} X$ if the set $\{\omega \in \Omega : X_n(\omega) \to X(\omega)\}$ has probability 1. Note that this is a very strong stipulation.

(d) X_n converges *in* L^p, written $X_n \xrightarrow{L^p} X$, if and only if $\|X_n - X\|_p \to 0$ as $n \to \infty$, or equivalently, if and only if $\mathbb{E}|X_n - X|^p \to 0$ as $n \to \infty$.

Remark. By Ljapunov's inequality, $\|X\|_s \leq \|X\|_t$ for all $0 < s < t$. Hence for $0 < s < t$, $X_n \xrightarrow{L^t} X$ implies that $X_n \xrightarrow{L^s} X$.

11.2. **Theorem.**

(a) *$X_n \xrightarrow{a.s.} X$ if and only if for all $\varepsilon > 0$, $\mathbb{P}(\sup_{k \geq n} |X_k - X| \geq \varepsilon) \to 0$ as $n \to \infty$.*

(b) *X_n is Cauchy almost surely if and only if for all $\varepsilon > 0$, $\mathbb{P}(\sup_{k,\ell \geq n} |X_k - X_\ell| \geq \varepsilon) \to 0$ as $n \to \infty$. This is also equivalent to for all $\varepsilon > 0$, $\mathbb{P}(\sup_{k \geq 0} |X_{n+k} - X_n| \geq \varepsilon) \to 0$.*

11.3. *Remark.* Note that (a) is like a 'boosted' version of convergence in probability, where we require that all points onwards from n are within ε of X.

Proof. (a). Define the events $A_k^m := \{\omega \in \Omega : |X_k - X| \geq \frac{1}{m}\}$ and $A^m := \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k^m$, which is the event that $|X_k - X| \geq \frac{1}{m}$ for infinitely many k.

Note that $X_n \not\to X$ if for some $m \in \mathbb{N}$, A^m occurs; that is, for some $m > 0$, $|X_k - X| \geq \frac{1}{m}$ for infinitely many k. The event that this happens for at least one m is $\bigcup_{m=1}^{\infty} A^m$. Thus $X_n \xrightarrow{\text{a.s.}} X$ if and only if $\mathbb{P}(\bigcup_{m=1}^{\infty} A^m) = 0$. Since $\mathbb{P}(A^m) \leq \mathbb{P}(\bigcup_{m=1}^{\infty} A^m) \leq \sum_{m=1}^{\infty} \mathbb{P}(A^m)$, $\mathbb{P}(\bigcup_{m=1}^{\infty} A^m) = 0$ if and only if $\mathbb{P}(A^m) = 0$ for all $m \in \mathbb{N}$. $\mathbb{P}(A^m) = 0$ if and only if

$$0 = \mathbb{P}\Bigg(\bigcap_{n=1}^{\infty} \underbrace{\bigcup_{k=n}^{\infty} A_k^m}_{\star} \Bigg) = \lim_{n\to\infty} \mathbb{P}\Bigg(\underbrace{\bigcup_{k=n}^{\infty} A_k^m}_{\star\star} \Bigg) = \lim_{n\to\infty} \mathbb{P}(\sup_{k\geq n} |X_k - X| \geq 1/m).$$

$\star$ defines a decreasing sequence. $\star\star$ is the event that for some $k \geq n$ we have $|X_k - X| \geq \frac{1}{m}$, which occurs if and only if the supremum, $\sup_{k\geq n} |X_k - X| \geq 1/m$.

As the above m is arbitrarily large, the proof is complete.

To prove (b), we repeat exactly the same analysis with the event $B_{k,\ell}^m = \{\omega \in \Omega : |X_k - X_\ell| \geq \frac{1}{m}\}$. □

11.4. **Theorem.** *The four types of convergence above can be partially ordered in strength.*

(a) $X_n \xrightarrow{a.s.} X \implies X_n \xrightarrow{P} X$.

(b) $X_n \xrightarrow{L^p} X \implies X_n \xrightarrow{P} X$.

(c) $X_n \xrightarrow{P} X \implies X_n \xrightarrow{d} X$.

Proof. To prove (a) we use the previous theorem, that is, that $X_n \xrightarrow{\text{a.s.}} X$ if and only if for all $\varepsilon > 0$, $\mathbb{P}(\sup_{k\geq n} |X_k - X| \geq \varepsilon) \to 0$ as $n \to \infty$. The event $|X_n - X| \geq \varepsilon$ implies that $\sup_{k\geq n} |X_k - X| \geq \varepsilon$, and thus

$$\mathbb{P}(|X_n - X| \geq \varepsilon) \leq \mathbb{P}(\sup_{k\geq n} |X_k - X| \geq \varepsilon) \to 0,$$

which implies $X_n \xrightarrow{P} X$.

To prove (b) we use Markov's inequality, which says that for a non-negative random variable Z, $\mathbb{P}(Z \geq c) \leq \mathbb{E}Z/c$. Then for all $p > 0$,

$$\mathbb{P}(|X_n - X| \geq \varepsilon) = \mathbb{P}(|X_n - X|^p \geq \varepsilon^p) \leq \frac{\mathbb{E}|X_n - X|^p}{\varepsilon^p} = \frac{\|X_n - X\|_p^p}{\varepsilon^p} \to 0$$

as $n \to \infty$ since $X_n \xrightarrow{L^p}$ if and only if $\|X_n - X\|_p \to 0$.

Now we prove (c) which says that convergence in probability implies weak convergence (in distribution). This is the most difficult to prove. Fix f bounded and continuous such that $|f(x)| \leq c$ (♠) for some $c > 0$. We want to show that $\mathbb{E}f(X_n) \to \mathbb{E}f(X)$, or equivalently that $|\mathbb{E}f(X_n) - \mathbb{E}f(X)| \to 0$. Fix $\varepsilon > 0$.

Assuming X has a proper distribution, i.e., that X is finite almost surely, there exists $N > 0$ such that

$$(\star) \qquad \mathbb{P}(|X| > N) \leq \frac{\varepsilon}{4c}.$$

$[-2N, 2N]$ is compact, and thus f is uniformly continuous on $[-2N, 2N]$. Hence there exists $\delta > 0$ such that for all $x, y \in [-2N, 2N]$ with $|x - y| \leq \delta$ we have

$$(\dagger) \qquad |f(x) - f(y)| \leq \frac{\varepsilon}{2}.$$

Define $\mathbb{E}(Z; A) = \mathbb{E}(Z\mathbb{1}_A)$. Then for a disjoint partition of the sample space $A_1, A_2, \ldots, A_n$, we may write $\mathbb{E}Z = \mathbb{E}(Z; A_1) + \mathbb{E}(Z; A_2) + \cdots + \mathbb{E}(Z; A_n)$. This is because $\mathbb{E}(Z; A_1) + \mathbb{E}(Z; A_2) + \cdots + \mathbb{E}(Z; A_n) =$

$\mathbb{E}[Z(\mathbb{1}_{A_1}+\cdots+\mathbb{1}_{A_n})]$ by linearity of expectation and because $(\mathbb{1}_{A_1}+\cdots+\mathbb{1}_{A_n})(\omega)=1$ for all $\omega\in\Omega$ by the disjointness of the partition.

$$\begin{aligned}
|\mathbb{E}f(X_n)-\mathbb{E}f(X)| &= |\mathbb{E}[f(X_n)-f(X)]| \\
&\le \mathbb{E}|f(X_n)-f(X)| && \text{Jensen's inequality on } |\cdot| \\
&= \mathbb{E}\Big[\underbrace{|f(X_n)-f(X)|}_{\le\frac{\varepsilon}{2}\impliedby\dagger};\underbrace{|X_n-X|\le\delta;|X|\le N}_{\implies X_n,X\in[-2N,2N],\mathbb{1}\le 1}\Big] \\
&+ \mathbb{E}\Big[\underbrace{|f(X_n)-f(X)|}_{\spadesuit\implies\le 2c};\underbrace{|X_n-X|\le\delta}_{\mathbb{1}\le 1};|X|>N\Big] \\
&+ \mathbb{E}\Big[\underbrace{|f(X_n)-f(X)|}_{\spadesuit\implies\le 2c};|X_n-X|>\delta\Big] \\
&\le \frac{\varepsilon}{2}+2c\underbrace{\mathbb{P}(|X|>N)}_{\star\implies\le\frac{\varepsilon}{4c}}+2c\mathbb{P}(|X_n-X|>\delta) \\
&\le \varepsilon+2c\mathbb{P}(|X_n-X|>\delta).
\end{aligned}$$

Since $X_n\xrightarrow{P}X$, there exists $N>0$ such that $n\ge N$ implies that $\mathbb{P}(|X_n-X|>\delta)\le\frac{\varepsilon}{2c}$. Thus $n\ge N$ implies that

$$|\mathbb{E}f(X_n)-\mathbb{E}f(X)|\le\varepsilon+2c\mathbb{P}(|X_n-X|>\delta)\le 2\varepsilon. \qquad \square$$

12. Strong Law of Large Numbers

12.1. **Theorem** (Strong law of large numbers (SLLN)). *Let $X_1,X_2,\dots$ be independent and identically distributed random variables with finite mean $\mu<\infty$. Then*

$$\frac{1}{n}\sum_{i=1}^{n}X_i\xrightarrow{a.s.}\mu.$$

12.2. **Corollary** (Weak law of large numbers (WLLN)). *Let $X_1,X_2,\dots$ be independent and identically distributed random variables with finite mean $\mu<\infty$. Then*

$$\frac{1}{n}\sum_{i=1}^{n}X_i\xrightarrow{P}\mu.$$

Proof. This follows immediately from the strong law of large numbers, since almost sure convergence implies convergence in probability. $\square$

The rest of this section is work towards proving the SLLN.

12.3. **Theorem** (Kolmogorov's inequality). *Suppose we have $X_1,X_2,\dots,X_n$ independent (not necessarily identically distributed) with $\mathbb{E}X_i=0$ and $\mathrm{Var}X_i<\infty$ for all $i=1,2,\dots,n$. Then,*

(a) *for all $\varepsilon>0$,*

$$\mathbb{P}(\max_{1\le k\le n}|S_k|\ge\varepsilon)\le\frac{\mathbb{E}S_n^2}{\varepsilon^2},$$

where $S_k=\sum_{i=1}^{k}X_i$.

(b) *If in addition* $|X_i| \leq c$ *almost surely for all* $i = 1, \ldots, n$ *and some* $c > 0$ *then*

$$\mathbb{P}(\max_{1\leq k\leq n} |S_k| \geq \varepsilon) \geq 1 - \frac{(c+\varepsilon)^2}{\mathbb{E}S_n^2}.$$

Proof. We start with (a). Define $A := \{\omega \in \Omega : \max_{1\leq k\leq n} |S_k(\omega)| \geq \varepsilon\}$ as the event that $\max_{1\leq k\leq n} |S_k(\omega)| \geq \varepsilon$ and let $A_k := \{\omega \in \Omega : |S_i(\omega)| < \varepsilon \text{ for } i = 1, 2, \ldots, k-1; |S_k(\omega)| \geq \varepsilon\}$ for $k = 1, \ldots, n$ to be the event that k is the first value for which $|S_i(\omega) > \varepsilon|$.

Then the A_k are disjoint and $\bigcup_{k=1}^n A_k$ is the event that at least one of $|S_k(\omega)|$ for $k = 1, \ldots, n$ is larger than or equal to ε. Hence $A = \bigcup_{k=1}^n A_k$.

Note that by the disjointness of the A_k,

$$\mathbb{E}S_n^2 \geq \mathbb{E}(S_n^2; A) = \sum_{k=1}^n \mathbb{E}(S_n^2; A_k).$$

Now we focus on $\mathbb{E}(S_n^2; A_k)$.

$$\begin{aligned}
\mathbb{E}(S_n^2; A_k) &= \mathbb{E}\big[(S_k + X_{k+1} + \cdots + X_n)^2; A_k\big] \\
&= \mathbb{E}(S_k^2; A_k) + 2\mathbb{E}\big[S_k(\underbrace{X_{k+1} + \cdots + X_n}_{\text{independent of } S_k, A_k}); A_k\big] + \underbrace{\mathbb{E}\big[(X_{k+1} + \cdots + X_n)^2; A_k\big]}_{\geq 0} \\
&\geq \mathbb{E}(S_k^2; A_k) + 2\mathbb{E}(S_k; A_k)\underbrace{\mathbb{E}(X_{k+1} + \cdots + X_n)}_{=0} \\
&= \mathbb{E}(S_k^2; A_k).
\end{aligned}$$

Thus $\mathbb{E}(S_n^2; A_k) \geq \mathbb{E}(S_k^2; A_k)$, but $\mathbb{E}(S_k^2; A_k) \geq \mathbb{E}(\varepsilon^2; A_k) = \varepsilon^2\mathbb{P}(A_k)$. The inequality holds since for $\omega \in \Omega$ for which $\mathbb{1}_{A_k}$ is non-zero, $|S_k(\omega)| \geq \varepsilon$. Furthermore,

$$\mathbb{E}S_n^2 \geq \sum_{k=1}^n \mathbb{E}(S_n^2; A_k) \geq \sum_{k=1}^n \mathbb{E}(S_k^2; A_k) \geq \varepsilon^2 \sum_{k=1}^n \mathbb{P}(A_k) = \varepsilon^2\mathbb{P}(A),$$

since $A = \bigcup_{k=1}^n A_k$ and the partition is disjoint. Rearranging gives (a).

Now we prove (b); to do so we use the same definition of A and A_k above. First we note that

$$\mathbb{E}(S_n^2; A) = \mathbb{E}S_n^2 - \mathbb{E}(S_n^2; A^c) \underset{*}{\geq} \mathbb{E}S_n^2 - \varepsilon^2\mathbb{P}(A^c) = \mathbb{E}S_n^2 - \varepsilon^2 + \varepsilon^2\mathbb{P}(A).$$

The inequality $*$ follows from that the $\omega \in \Omega$ for which $\mathbb{1}_{A^c}$ is non-zero is also the $\omega \in \Omega$ for which $\max_{1\leq k\leq n} |S_k(\omega)| < \varepsilon$ and therefore $S_n(\omega) < \varepsilon$ and thus $\mathbb{E}(S_n^2; A^c) \leq \mathbb{E}(\varepsilon^2; A^c) = \varepsilon^2\mathbb{P}(A^c)$.

On the other hand,

$$\begin{aligned}
\mathbb{E}(S_n^2; A_k) &= \mathbb{E}\big[(S_k + X_{k+1} + \cdots + X_n)^2; A_k\big] \\
&= \mathbb{E}(S_k^2; A_k) + \underbrace{2\mathbb{E}\big[S_k(X_{k+1} + \cdots + X_n); A_k\big]}_{=0, \text{ as above}} + \mathbb{E}\big[(X_{k+1} + \cdots + X_n)^2; A_k\big] \\
&= \mathbb{E}(S_k^2; A_k) + \mathbb{E}\big[(S_n - S_k)^2; A_k\big].
\end{aligned}$$

Hence

$$\mathbb{E}(S_n^2; A) = \sum_{k=1}^n \mathbb{E}(S_n^2; A_k) = \sum_{k=1}^n \mathbb{E}(S_k^2; A_k) + \sum_{k=1}^n \mathbb{E}\big[(S_n - S_k)^2; A_k\big].$$

The event A_k implies (by definition) that $|S_{k-1}(\omega)| < \varepsilon$. Therefore $|S_k| = |X_k + S_{k-1}| \leq |X_k| + |S_{k-1}| \leq c + \varepsilon$ since X_k is bounded by c almost surely. Moreover,

$$(\star) \qquad \mathbb{E}(S_k^2; A_k) \leq (c+\varepsilon)^2\mathbb{P}(A_k).$$

Now $S_n - S_k = X_{k+1} + \cdots + X_n$ which is independent of A_k. Thus $\mathbb{E}\big[(S_n - S_k)^2; A_k\big] = \mathbb{E}(S_n - S_k)^2\mathbb{P}(A_k)$. By the independence of the X_i, $\mathbb{E}(X_iX_j) = \mathbb{E}X_i\mathbb{E}X_j = 0$. Hence

$$(\star\star) \qquad \mathbb{E}\big[(S_n - S_k)^2; A_k\big] = \mathbb{E}(S_n - S_k)^2\mathbb{P}(A_k) = \mathbb{E}(X_{k+1} + \cdots + X_n)^2\mathbb{P}(A_k) = \sum_{j=k+1}^{n} \mathbb{E}X_j^2\mathbb{P}(A_k)$$

as all cross terms cancel by the above.

Moreover,

$$\begin{aligned}
\mathbb{E}(S_n^2; A) &= \sum_{k=1}^{n} \mathbb{E}(S_k^2; A_k) + \sum_{k=1}^{n} \mathbb{E}\big[(S_n - S_k)^2; A_k\big] \\
&\leq \sum_{k=1}^{n} \underbrace{(c+\varepsilon)^2\mathbb{P}(A_k)}_{\text{by } \star} + \sum_{k=1}^{n} \overbrace{\underbrace{\sum_{j=k+1}^{n} \mathbb{E}X_j^2\mathbb{P}(A_k)}_{\leq \sum_{j=1}^{n} \mathbb{E}X_j^2\mathbb{P}(A_k)}}^{\text{by } \star\star} \\
&\leq (c+\varepsilon)^2\mathbb{P}(A) + \underbrace{\sum_{j=1}^{n} \mathbb{E}X_j^2}_{=\mathbb{E}S_n^2}\mathbb{P}(A) = \big[(c+\varepsilon)^2 + \mathbb{E}S_n^2\big]\mathbb{P}(A).
\end{aligned}$$

Combining the two inequalities we have found we have that $\big[(c+\varepsilon)^2 + \mathbb{E}S_n^2\big]\mathbb{P}(A) \geq \mathbb{E}S_n^2 - \varepsilon^2 + \varepsilon^2\mathbb{P}(A)$. Now we simply rearrange. The inequality implies that

$$\begin{aligned}
\big[(c+\varepsilon)^2 - \varepsilon^2 + \mathbb{E}S_n^2\big]\mathbb{P}(A) \geq \mathbb{E}S_n^2 - \varepsilon^2 \implies \mathbb{P}(A) &\geq \frac{\mathbb{E}S_n^2 - \varepsilon^2}{(c+\varepsilon)^2 - \varepsilon^2 + \mathbb{E}S_n^2} \\
&= 1 - \frac{(c+\varepsilon)^2}{\underbrace{(c+\varepsilon)^2 - \varepsilon^2}_{>0} + \mathbb{E}S_n^2} \\
&\geq 1 - \frac{(c+\varepsilon)^2}{\mathbb{E}S_n^2}.
\end{aligned}$$

□

12.4. **Theorem** (Kolmogorov-Khinchin)**.** *Let $X_1, X_2, \ldots$ be independent (not necessarily identically distributed) random variables with $\mathbb{E}X_i = 0$ for all $i \in \mathbb{N}$. Then*

(a) $\sum_{n=1}^{\infty} \mathbb{E}X_n^2 < \infty \implies \sum_n X_n$ *converges almost surely;*

(b) *If $|X_n| \leq c$ almost surely and $\sum_n X_n$ converges almost surely then $\sum_{n=1}^{\infty} \mathbb{E}X_n^2 < \infty$.*

Proof. In order to prove (a) we show that S_n is Cauchy almost surely. Recall that S_n is Cauchy almost surely if and only if for all $\varepsilon > 0$, $\mathbb{P}(\sup_{k\geq 1} |S_{n+k} - S_n| \geq \varepsilon) \to 0$.

$$\begin{aligned}
\mathbb{P}(\sup_{k\geq 1} |S_{n+k} - S_n| \geq \varepsilon) &\underset{*}{=} \lim_{N\to\infty} \mathbb{P}(\max_{1\leq k\leq N} |S_{n+k} - S_n| \geq \varepsilon) \\
&\underset{\dagger}{\leq} \lim_{N\to\infty} \frac{\sum_{j=n+1}^{n+N} \mathbb{E}X_j^2}{\varepsilon^2} = \frac{1}{\varepsilon^2} \sum_{j=n+1}^{\infty} \mathbb{E}X_j^2 \to 0
\end{aligned}$$

since it is the tail of a convergent sum (by assumption).

$*$ is allowed since $\max_{1\leq k\leq N} |S_{n+k} - S_n| \geq \varepsilon$ is an increasing sequence of events in N. $\dagger$ follows from the first part of Kolmogorov inequality on $S_{n+k} - S_n = X_{n+1} + \cdots + X_{n+k} := T_k$ where

$$\mathbb{E}T_k^2 = \sum_{j=n+1}^{n+N} \mathbb{E}X_j^2$$

since cross terms have zero-expectation. Thus S_n is Cauchy almost surely. By the completeness of $\mathbb{R}$, S_n converges almost surely, which proves (a).

Now for (b). If S_n converges almost surely then it is Cauchy almost surely. Again, using that theorem from before, S_n is Cauchy almost surely if and only if for all $\varepsilon > 0$, $\mathbb{P}(\sup_{k\geq 1} |S_{n+k} - S_n| \geq \varepsilon) \to 0$. Then for all sufficiently large n,

$$\frac{1}{2} > \mathbb{P}(\sup_{k\geq 1} |S_{n+k} - S_n| \geq \varepsilon) \underset{\ddagger}{\geq} 1 - \frac{(c+\varepsilon)^2}{\sum_{j=n+1}^{\infty} \mathbb{E}X_j^2}.$$

‡ follows from the second part of Kolmogorov's inequality and by the same working with limits above. Rearranging this inequality gives that

$$\sum_{j=n+1}^{\infty} \mathbb{E}X_j^2 \leq 2(c+\varepsilon)^2,$$

and hence it converges (for all n sufficiently large). Moreover, $\sum_{j=1}^{\infty} \mathbb{E}X_j^2$ converges since it is the sum of a finite sum (up to $n+1$) and an (infinite) convergent sum. □

12.5. **Theorem** (Toeplitz lemma)**.** *Let (a_n) be a sequence with $a_n \geq 0$ for all n and $b_n := \sum_{i=1}^{n} a_i > 0$ (thus at least $a_1 > 0$ strictly). Suppose $b_n \nearrow \infty$ as $n \to \infty$. If $x_n \to x$ as $n \to \infty$, then*

$$\frac{1}{b_n} \sum_{i=1}^{n} a_i x_i \to x \text{ as } n \to \infty.$$

12.6. **Example.** If $a_i = 1$ for all $i \in \mathbb{N}$ then $b_n = n \nearrow \infty$. Then if $x_n \to x$, $\frac{1}{n}\sum_{i=1}^{n} x_i \to x$. This is also known as Cesàro's lemma.

Proof. Fix $\varepsilon > 0$ and choose n_0 such that $n \geq n_0$ implies that $|x_n - x| < \varepsilon/2$. Then fix $n_1 \geq n_0$ such that

$$\frac{1}{b_{n_1}} \sum_{i=1}^{n_0} a_i |x_i - x| < \varepsilon/2.$$

We can do this since we treat $\sum_{i=1}^{n_0} a_i|x_i - x|$ as fixed and since $b_n \nearrow \infty$ we simply need to choose n_1 large enough.

Then for all $n \geq n_1 \geq n_0$, (writing $x = (\frac{1}{b_n}\sum_{i=1}^{n} a_i)x$)

$$\begin{aligned}
\left| \frac{1}{b_n} \sum_{i=1}^{n} a_i x_i - x \right| &\leq \frac{1}{b_n} \sum_{i=1}^{n} a_i |x_i - x| \\
&\leq \frac{1}{b_{n_1}} \sum_{i=1}^{n_0} a_i |x_i - x| + \frac{1}{b_n} \sum_{i=n_0+1}^{n} a_i |x_i - x| \\
&< \frac{\varepsilon}{2} + \underbrace{\frac{1}{b_n} \sum_{i=n_0+1}^{n} a_i}_{\leq 1} \frac{\varepsilon}{2} \leq \varepsilon.
\end{aligned}$$

□

12.7. **Theorem** (Kronecker's lemma)**.** *Let $b_n > 0$ for $n \in \mathbb{N}$; $b_n \nearrow \infty$ as $n \to \infty$. If (x_n) is a sequence such that $\sum_n x_n$ converges then*

$$\frac{1}{b_n} \sum_{i=1}^{n} b_i x_i \to 0.$$

Proof. The proof of Kronecker's lemma is slightly more involved than that of the Toeplitz lemma.

Let $b_0 := 0$; $s_0 := 0$ and $s_n := \sum_{i=1}^n x_i$.

$$\sum_{i=1}^n b_i x_i = \sum_{i=1}^n b_i(s_i - s_{i-1}) \underset{*}{=} b_n s_n - \underbrace{b_0 s_0}_{=0} - \sum_{i=1}^n s_{i-1}(b_i - b_{i-1}).$$

The equality $*$ is known as *summation by parts* and is the analogue of integration by parts for sums. It is easily verified by comparing terms:

- On the left we have $+b_1 s_1, +b_2 s_2, \ldots, +b_n s_n$.
- On the left we have $-b_1 s_0, -b_2 s_1, \ldots, -b_n s_{n-1}$.

It is easy to see that these terms appear on the right hand side. Moreover,

$$\frac{1}{b_n}\sum_{i=1}^n b_i x_i = s_n - \frac{1}{b_n}\sum_{i=1}^n s_{i-1}\underbrace{(b_i - b_{i-1})}_{:=a_i}.$$

Then $a_i \geq 0$ for all i since b_i increasing implies $a_i = b_i - b_{i-1} \geq 0$. Furthermore, $\sum_{i=1}^n a_i = \sum_{i=1}^n (b_i - b_{i-1}) = b_n$. Then we are in good shape to apply Toeplitz lemma,

$$\lim_{n\to\infty}\frac{1}{b_n}\sum_{i=1}^n b_i x_i = \lim_{n\to\infty} s_n - \lim_{n\to\infty}\frac{1}{b_n}\sum_{i=1}^n a_i s_{i-1} = s - s = 0. \qquad \square$$

12.8. **Theorem** (Kolmogorov)**.** *Let $X_1, X_2, \ldots$ be independent (not necessarily identically distributed) with finite variance $\mathrm{Var}X_n < \infty$ for all $n \in \mathbb{N}$. Let $b_n \nearrow \infty$ as $n \to \infty$; $b_n > 0$ such that*

$$\sum_{n=1}^\infty \frac{\mathrm{Var}X_n}{b_n^2} < \infty \quad \textit{then} \quad \frac{S_n - \mathbb{E}S_n}{b_n} \xrightarrow{a.s.} 0.$$

12.9. **Example.** When $b_n = n$ with X_n i.i.d. random variables then Kolmogorov's theorem is the strong law of large numbers with the stipulation that the random variables must additionally have a finite second moment.

Proof. We rewrite the sequence in order that we may use Kronecker's lemma.

$$\frac{S_n - \mathbb{E}S_n}{b_n} = \frac{1}{b_n}\sum_{i=1}^n (X_i - \mathbb{E}X_i) = \frac{1}{b_n}\sum_{i=1}^n b_i \frac{X_i - \mathbb{E}X_i}{b_i}.$$

By Kronecker's lemma, to show that this converges to 0 in n, it suffices to show that $\sum_i \frac{X_i - \mathbb{E}X_i}{b_i}$ converges. Let

$$Y_i = \frac{X_i - \mathbb{E}X_i}{b_i} \implies \sum_{i=1}^\infty \mathbb{E}Y_i^2 = \sum_{i=1}^\infty \frac{\mathrm{Var}X_i}{b_i^2} < \infty.$$

Thus applying Kolmogorov-Khinchin to Y_i we have that $\sum_{n=1}^\infty Y_n < \infty$ almost surely. $\square$

12.10. **Lemma.** *Let X be a random variable such that $X \geq 0$ almost surely. Then*

$$\sum_{n=1}^\infty \mathbb{P}(X \geq n) \leq \mathbb{E}X \leq 1 + \sum_{n=1}^\infty \mathbb{P}(X \geq n).$$

Proof. Note that if $Y \geq 0$ a.s. is an integer valued random variable, then $\mathbb{E}Y = \sum_{n=1}^\infty \mathbb{P}(Y \geq n)$, thus equality holds on the left and strict inequality on the right. Why is this true in the integer valued case?

$$\sum_{n=1}^\infty \mathbb{P}(Y \geq n) = \sum_{n=1}^\infty \sum_{k=n}^\infty \mathbb{P}(Y = k) = \sum_{k=1}^\infty \sum_{n=1}^k \mathbb{P}(Y = k) = \sum_{k=1}^\infty k\mathbb{P}(X = k) = \mathbb{E}Y.$$

Now consider the general case. We can certainly approximate X by its floor: $\lfloor X \rfloor \leq X \leq \lfloor X \rfloor + 1$. But the above holds for $\lfloor X \rfloor$, and hence

$$\sum_{n=1}^{\infty} \mathbb{P}(\lfloor X \rfloor \geq n) = \mathbb{E}\lfloor X \rfloor \leq \mathbb{E}X \leq \mathbb{E}\lfloor X \rfloor + 1 = 1 + \sum_{n=1}^{\infty} \mathbb{P}(\lfloor X \rfloor \geq n).$$

But since n is an integer, $X \geq n$ if and only if $\lfloor X \rfloor \geq n$, which completes the proof. □

12.11. **Theorem** (Strong law of large numbers)**.** *Let $X_1, X_2, \ldots$ be independent and identically distributed random variables with finite mean $\mu < \infty$. Then*

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{a.s.} \mu.$$

Proof. Assume without loss of generality that $\mu = 0$ (if non-zero then the theorem applies to $X_i \equiv X_i - \mu$ which has zero-mean).

By the previous lemma,

$$\infty > \mathbb{E}|X_1| \geq \sum_{n=1}^{\infty} \mathbb{P}(|X_1| \geq n) = \sum_{n=1}^{\infty} \mathbb{P}(|X_n| \geq n)$$

by the fact that the X_i are identically distributed. Let $A_n := \{\omega \in \Omega : |X_n(\omega)| \geq n\}$. Then $\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty$. Hence by the first Borel-Cantelli lemma, $\mathbb{P}(A_n \, i.o.) = 0$. (Note that this statement did not require the independence of the X_i.) So $A_n = \{|X_n| \geq n\}$ happens finitely many times *almost surely.*

Thus we define the adjusted random variable $\widetilde{X}_n$ where

$$\widetilde{X}_n := \begin{cases} X_n, & \text{if } |X_n| < n, \\ 0, & \text{if } |X_n| \geq n. \end{cases}$$

Since we previously concluded by Borel-Cantelli 1 that $|X_n| \geq n$ only finitely many times almost surely, then the new sequence $\widetilde{X}_n$ is equal to X_n except for only finitely many n's almost surely. Therefore

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{\text{a.s.}} 0 \iff \frac{1}{n}\sum_{i=1}^{n} \widetilde{X}_i \xrightarrow{\text{a.s.}} 0.$$

This equivalence holds only because $\widetilde{X}_n$ changes only *finitely* many terms (almost surely).

We calculate the limiting expectation of $\widetilde{X}_n$. Now $\mathbb{E}\widetilde{X}_n = \mathbb{E}(X_n \mathbb{1}_{\{|X_n|<n\}})$ since $\widetilde{X}_n = X_n \mathbb{1}_{\{|X_n|<n\}}$ by defintion. Thus $\mathbb{E}\widetilde{X}_n = \mathbb{E}(X_n; |X_n| < n)$. By the i.i.d. property of the X_i, $\mathbb{E}\widetilde{X}_n = \mathbb{E}(X_1; |X_1| < n)$. Note that

- $|X_1 \mathbb{1}_{\{|X_1|<n\}}| \leq |X_1|$ a.s.;
- $\mathbb{E}|X_1| < \infty$ (by the hypothesis);
- $X_1 \mathbb{1}_{\{|X_1|<n\}} \xrightarrow{\text{a.s.}} X_1$ as $n \to \infty$.

Hence we are in good shape to apply the Dominated Convergence Theorem to the sequence $X_1 \mathbb{1}_{\{|X_1|<n\}}$ and conclude that

$$\mathbb{E}\widetilde{X}_n = \mathbb{E}(X_1; |X_1| < n) = \mathbb{E}(X_1 \mathbb{1}_{\{|X_1|<n\}}) \to \mathbb{E}X_1 = \mu = 0.$$

Hence we have shown that $\mathbb{E}\widetilde{X}_n \to 0$ as $n \to \infty$. Hence

$$(\dagger) \qquad \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\widetilde{X}_i \to 0$$

by Toeplitz's lemma. Note that in general $\mathbb{E}\widetilde{X}_i \neq 0$. Why? Consider X_1 a continuous random variable for the purpose of illustration. Then

$$\mathbb{E}\widetilde{X}_n = \mathbb{E}(X_1 \mathbb{1}_{\{|X_1|<n\}}) = \int_{-\infty}^{\infty} x\mathbb{1}_{\{|x|<n\}} f_{X_1}(x)dx = \int_{-n}^{n} x f_{X_1}(x)dx \neq \mathbb{E}X = 0.$$

We want to show that $\frac{1}{n}\sum_{i=1}^n \widetilde{X}_i \xrightarrow{\text{a.s.}} 0$. By †,

$$\frac{1}{n}\sum_{i=1}^n \widetilde{X}_i \xrightarrow{\text{a.s.}} 0 \iff \frac{1}{n}\sum_{i=1}^n (\widetilde{X}_i - \mathbb{E}\widetilde{X}_i) \xrightarrow{\text{a.s.}} 0 \iff \frac{1}{n}\sum_{i=1}^n i\frac{\widetilde{X}_i - \mathbb{E}\widetilde{X}_i}{i} \xrightarrow{\text{a.s.}} 0.$$

Thus by Kronecker's lemma it suffices to show that $\sum_i \frac{\widetilde{X}_i - \mathbb{E}\widetilde{X}}{i}$ converges almost surely. By Kolmogorov-Khinchin (since X_i are independent, $\widetilde{X}_i$ are independent and therefore $(\widetilde{X}_i - \mathbb{E}\widetilde{X}_i)/i$ are independent) this is true if

$$\sum_{i=1}^{\infty} \mathbb{E}\left[\frac{\widetilde{X}_i - \mathbb{E}\widetilde{X}}{i}\right]^2 = \sum_{i=1}^{\infty} \frac{\operatorname{Var}\widetilde{X}_i}{i^2} < \infty.$$

In order to show this we perform some algebraic manipulation.

$$\begin{aligned}
\sum_{n=1}^{\infty} \frac{\operatorname{Var}\widetilde{X}_n}{n^2} &= \sum_{n=1}^{\infty} \frac{\mathbb{E}\widetilde{X}_n^2 - [\mathbb{E}\widetilde{X}_n]^2}{n^2} \\
&\leq \sum_{n=1}^{\infty} \frac{\mathbb{E}\widetilde{X}_n^2}{n^2} \\
&= \sum_{n=1}^{\infty} \frac{\mathbb{E}(X_n^2; |X_n| < n)}{n^2} \\
&= \sum_{n=1}^{\infty} \frac{1}{n^2}\sum_{k=1}^{n} \mathbb{E}(X_n^2; k-1 \leq |X_n| < k) && \because \text{events form a disjoint partition} \\
&= \sum_{n=1}^{\infty} \frac{1}{n^2}\sum_{k=1}^{n} \mathbb{E}(X_1^2; k-1 \leq |X_1| < k) && \because X_i \text{ are i.i.d.} \\
&= \sum_{k=1}^{\infty} \mathbb{E}(X_1^2; k-1 \leq |X_1| < k)\sum_{n=k}^{\infty} \frac{1}{n^2} && \text{By switching the order of the sum}
\end{aligned}$$

Let us work on $\sum_{n=k}^{\infty} \frac{1}{n^2}$. For $k = 1$ this is simply equal to $\frac{\pi^2}{6} < 2$. For $k \geq 2$,

$$\sum_{n=k}^{\infty} \frac{1}{n^2} \leq \int_{k-1}^{\infty} \frac{1}{x^2}dx = -\frac{1}{x}\Big|_{k-1}^{\infty} = \frac{1}{k-1} \leq \frac{2}{k}.$$

Note that the final inequality holds only in the case $k \geq 2$. Hence for all $k \in \mathbb{N}$, $\sum_{n=k}^{\infty} \frac{1}{n^2} \leq \frac{2}{k}$. Making this substitution,

$$\begin{aligned}
\sum_{n=1}^{\infty} \frac{\operatorname{Var}\widetilde{X}_n}{n^2} &\leq \sum_{k=1}^{\infty} \mathbb{E}(\underbrace{X_1^2}_{<k|X_1| \Longleftarrow}; k-1 \leq |X_1| < k)\frac{2}{k} \\
&< 2\sum_{k=1}^{\infty} \mathbb{E}(|X_1|; k-1 \leq |X_1| < k) \\
&= 2\mathbb{E}|X_1| < \infty && \because \text{disjoint partition}
\end{aligned}$$

□

12.12. **Theorem** (SLLN in reverse)**.** *Let $X_1, X_2, \ldots$ be an iid sequence such that*

$$\frac{1}{n}\sum_{k=1}^{n} X_k \xrightarrow{a.s.} c$$

for some $c < \infty$. Then $\mathbb{E}X_k = c$.

Proof. Indeed

$$\frac{X_k}{k} = \frac{S_k - S_{k-1}}{k} = \frac{S_k}{k} - \frac{k-1}{k}\frac{S_{k-1}}{k-1} \xrightarrow{\text{a.s.}} c - 1 \times c = 0.$$

Then

$$\mathbb{P}(|X_k| > k \text{ } i.o.) = \mathbb{P}(|X_k/k| > 1 \text{ } i.o.) = 0.$$

By independence and Borel-Cantelli Lemma 2,

$$\sum_{k=1}^{\infty} \mathbb{P}(|X_n| > n) < \infty \implies \sum_{k=1}^{\infty} \mathbb{P}(|X_1| > n) < \infty \implies \mu := \mathbb{E}|X_1| < \infty$$

by a few lemmas ago. Since the mean is finite we can apply the SLLN,

$$\frac{1}{n}\sum_{k=1}^{n} X_k \xrightarrow{\text{a.s.}} \mu \implies \mu = \mathbb{E}X_k = c. \qquad \square$$

12.13. *Remark.* Does SLLN hold when $\mathbb{E}X_1^- < \infty$ but $\mathbb{E}X_1^+ = \infty$ ($\implies \mathbb{E}X_1 = \infty$)? Yes. Why? Define

$$S_n^c = \sum_{i=1}^{\infty} X_i \mathbb{1}_{\{X_i \le c\}}.$$

SLLN applies to $X_i\mathbb{1}_{\{X_i \le c\}}$ as this is now bounded by c, preventing the unboundedness issue. Note that

$$\liminf_{n\to\infty} \frac{S_n}{n} \ge \liminf_{n\to\infty} \frac{S_n^c}{n} = \mathbb{E}(X_i \mathbb{1}_{\{X_i \le c\}}) \nearrow \mathbb{E}X_1 = \infty.$$

13. Characteristic Functions

13.1. **Definition.** The characteristic function of a random variable X is the function $\varphi : \mathbb{R} \to \mathbb{C}$ is defined by

$$\varphi(t) = \mathbb{E}e^{itX}.$$

We write, see the remark on Lebesgue-Stieltjes itegration after Definition 9.22,

$$\varphi(t) = \int_{-\infty}^{\infty} e^{itX} dF(x)$$

where $F(x)$ is the distribution function of X. Note that if X is continuous then

$$\int_{-\infty}^{\infty} e^{itX} dF(x) = \int_{-\infty}^{\infty} e^{itX} f(x) dx$$

and if X is discrete then

$$\int_{-\infty}^{\infty} e^{itX} dF(x) = \sum_n e^{itx_n} p(x_n).$$

13.2. *Remark.* The characteristic function always exists.

13.3. **Proposition** (Properties of characteristic functions)**.**

(a) $\varphi : \mathbb{R} \to \mathbb{C}$ *is uniformly continuous.*

(b) $\varphi(0) = 1$*;* $|\varphi(t)| \le 1$ *for all* $t \in \mathbb{R}$*. Furthermore,* φ *satisfies one of*

(i) $|\varphi(t)| < 1$ *for all* $t \ne 0$*;*

(ii) *there exists* $\lambda \in \mathbb{R}$*,* $\lambda > 0$*, such that* $|\varphi(t)| < 1$ *for* $0 < t < \lambda$ *and* $|\varphi(\lambda)| = 1$*.*

In this case, φ *is periodic in* $|\cdot|$ *with period* λ*, that is* $|\varphi(t + n\lambda)| = |\varphi(t)|$*, and* X *has lattice distribution, that is, there exists* $b \in \mathbb{R}$ *such that*

$$\mathbb{P}\left(X \in \left\{b + \frac{2\pi k}{\lambda} : k \in \mathbb{Z}\right\}\right) = 1.$$

(iii) $|\varphi(t)| = 1$ *for all* $t \in \mathbb{R}$.

In this case, $\varphi(t) = e^{ibt}$ *for some* $b \in \mathbb{R}$ *and* $X \equiv b$ *almost surely.*

(c) *For* $a, b \in \mathbb{R}$ *fixed,* $\varphi_{aX+b}(t) = e^{ibt}\varphi_X(at)$.

(d) $\overline{\varphi_X(t)} = \varphi_{-X}(t) = \varphi_X(-t)$.

(e) *There exist random variables* X *and* Y *of different distribution such that* $\varphi_X(t) = \varphi_Y(t)$ *for* $t \in [a, b]$ *for some* $-\infty < a < b < \infty$.

However, if $\varphi_X(t) = \varphi_Y(t)$ *for all* $t \in \mathbb{R}$ *then* X *and* Y *are of the same distribution. In this sense, the characteristic function uniquely determines the distribution of* X.

(f) *If* X *and* Y *are independent then*

$$\varphi_{X+Y}(t) = \varphi_X(t)\varphi_Y(t).$$

Proof. (a). Note that using Jensen's inequality

$$|\varphi(t+h) - \varphi(t)| = |\mathbb{E}e^{i(t+h)X} - \mathbb{E}e^{itX}| = |\mathbb{E}e^{itX}(e^{ihX} - 1)| \leq \mathbb{E}\underbrace{|e^{itX}|}_{=1}|e^{ihX} - 1|.$$

Left with something which is uniform in t. $|e^{ihX} - 1| \xrightarrow{\text{a.s.}} 0$ as $h \to 0$ and $|e^{ihX} - 1| \leq 2$ hence we can apply DCT $\implies \mathbb{E}|e^{ihX} - 1| \to \mathbb{E}0 = 0$.

(b). Clearly $\varphi(0) = \mathbb{E}e^0 = 1$. Using Jensen's inequality on $|\cdot|$,

$$|\varphi(t)| = |\mathbb{E}e^{itX}| \leq \mathbb{E}|e^{itX}| = 1.$$

Clearly only one of (i), (ii) or (iii) can happen since $|\varphi(t)| \leq 1$ and $\varphi(0) = 1$.

(ii). We prove that if there exists $\lambda \in \mathbb{R}$, $\lambda > 0$, such that $|\varphi(t)| < 1$ for $0 < t < \lambda$ and $|\varphi(\lambda)| = 1$ then φ has period λ and X has lattice distribution.

If $|\varphi(\lambda)| = 1$ then $\varphi(\lambda) = e^{ib\lambda}$ for some $b \in \mathbb{R}$. Hence

$$1 = \varphi(\lambda)e^{-ib\lambda} = \mathbb{E}e^{i\lambda(X-b)} = \mathbb{E}\big[\cos(\lambda(X-b)) + i\sin(\lambda(X-b))\big].$$

So $\mathbb{E}\sin(\lambda(X-b)) = 0$ and $\mathbb{E}\cos(\lambda(X-b)) = 1$.

$\implies \cos(\lambda(X-b)) = 1$ and $\sin(\lambda(X-b)) = 0$ almost surely.

Why? Because $\cos \leq 1$ so all mass is concentrated on values of X for which $\cos(\lambda(X-b)) = 1$.

$\implies \lambda(X-b) = 2\pi k$, $k \in \mathbb{Z}$ almost surely.

$\implies X \in \{b + 2\pi k/\lambda : k \in \mathbb{Z}\}$ almost surely; so X has lattice distribution.

In this case we show that φ is periodic. As X has a lattice distribution we can define

$$p_k := \mathbb{P}(X = b + 2\pi k/\lambda) \qquad \text{where} \qquad \sum_{k\in\mathbb{Z}} p_k = 1.$$

Then

$$\varphi(t) = \mathbb{E}e^{itX} = \sum_{k\in\mathbb{Z}} p_k e^{it(b+\frac{2\pi k}{\lambda})} = e^{itb}\sum_{k\in\mathbb{Z}} p_k e^{it\frac{2\pi k}{\lambda}}.$$

On the other hand,

$$|\varphi(t+\lambda)| = \left|\mathbb{E}e^{i(t+\lambda)X}\right| = \left|\sum_{k\in\mathbb{Z}} p_k e^{i(t+\lambda)(b+\frac{2\pi k}{\lambda})}\right| = \left|e^{i(t+\lambda)b}\sum_{k\in\mathbb{Z}} p_k e^{it\frac{2\pi k}{\lambda}}\underbrace{e^{i2\pi k}}_{=1}\right| = \left|e^{i\lambda b}\right|\left|e^{itb}\sum_{k\in\mathbb{Z}} p_k e^{it\frac{2\pi k}{\lambda}}\right|$$
$$= |\varphi(t)|.$$

Note that if φ is of period λ then it is also of period $n\lambda$. So there is a smallest λ for which φ has period.

Note that if X has a lattice distribution then any finer lattice will also contain all of the probability, however some points will have zero-mass. Seek to find the λ which makes these gaps the widest.

With this λ we claim that $|\varphi(t)| < 1$ for $0 < t < \lambda$.

Let λ be the smallest λ such that

$$\mathbb{P}\left(X \in \left\{b + \frac{2\pi k}{\lambda} : k \in \mathbb{Z}\right\}\right) = 1.$$

Then

$$|\varphi(t)| = \left|\sum_{k\in\mathbb{Z}} p_k e^{it(b+\frac{2\pi k}{\lambda})}\right|.$$

Indeed for $t = \lambda$,

$$\left|\sum_{k\in\mathbb{Z}} p_k e^{i\lambda(b+\frac{2\pi k}{\lambda})}\right| = |e^{i\lambda b}|\left|\sum_{k\in\mathbb{Z}} p_k e^{i2\pi k}\right| = 1.$$

For $t < \lambda$,

$$\left|\sum_{k\in\mathbb{Z}} p_k e^{it(b+\frac{2\pi k}{\lambda})}\right| = |e^{itb}|\left|\sum_{k\in\mathbb{Z}} p_k e^{i2\pi k\frac{t}{\lambda}}\right|.$$

Since $t/\lambda < 1$ the exponential sees others than a full revolution and so is not equal to 1 for some k. That implies that the sum cannot be one in mod either as we are adding up numbers of different complex phases but with mods summing up to one. However, suppose that $p_k = 0$ for where we don't have a full revolution (making the sum still 1), then there exists a finer lattice distribution, a contradiction. Hence by the triangle inequality this is less than 1.

(iii). If $|\varphi(\lambda)| = 1$ for all $\lambda \in \mathbb{R}$ then by the same proof as start of (ii), for all $\lambda \in \mathbb{R}$

$$X \in \left\{b_\lambda + \frac{2\pi k}{\lambda} : k \in \mathbb{Z}\right\}$$

almost surely. Fix a λ_1. Then on λ_1's lattice there exists at least 1 point with positive mass. Thus for any other λ, its lattice must have a point which coincides with this point. So all mass is on this point, call it b, and $X = b$ almost surely.

$$\varphi(t) = \mathbb{E}e^{itX} = e^{itb}.$$

(c). $\varphi_{aX+b}(t) = \mathbb{E}e^{it(aX+b)} = e^{itb}\mathbb{E}e^{i(at)X} = e^{itb}\varphi_X(at)$.

(d). $\overline{\varphi_X(t)} = \overline{\mathbb{E}e^{itX}} = \mathbb{E}\overline{e^{itX}} = \mathbb{E}e^{-itX}$.

(e). The uniqueness part will follow from the inversion formula (see later).

(f). $\varphi_{X+Y}(t) = \mathbb{E}e^{it(X+Y)} = \mathbb{E}e^{itX}\mathbb{E}e^{itY}$; separation of expectation follows from independence. □

13.4. **Theorem** (Bochner)**.** *A continuous function w is the characteristic function of a distribution if and only if*

- $w(0) = 1$*;*
- w *is positive definite, that is, for all $t_1, \ldots, t_n \in \mathbb{R}$, for all $z_1, \ldots, z_n \in \mathbb{C}$*

$$\sum_{j,k} w(t_j - t_k) z_j \overline{z_k} \geq 0.$$

Proof. Only in the forward direction. Let w be a characteristic function, then we already know that $w(0) = 1$. Furthermore,

$$\begin{aligned}\sum_{j,k} w(t_j - t_k) z_j \overline{z_k} = \sum_{j,k} \mathbb{E} e^{i(t_j - t_k)X} z_j \overline{z_k} &= \mathbb{E}\left[\sum_j e^{it_j} z_j \sum_k e^{-it_k}\overline{z_k}\right] \\ &= \mathbb{E}\left[\sum_j e^{it_j} z_j \overline{\sum_k e^{it_k} z_k}\right] \\ &= \mathbb{E}\left|\sum_j e^{it_j} z_j\right|^2 \geq 0.\end{aligned}$$

□

13.5. **Theorem.** *If $\varphi(t)$ is differentiable k times at $t = 0$ then the* $\begin{cases} kth & \text{if } k \text{ even,} \\ k-1st & \text{if } k \text{ odd} \end{cases}$ *moment is finite and for all* $j \leq \begin{cases} k & \text{if } k \text{ even,} \\ k-1 & \text{if } k \text{ odd} \end{cases}$

$$\mathbb{E}X^j = (-i)^j \left(\frac{d}{dt}\right)^j \varphi(t)\Big|_{t=0}.$$

Proof. None given. □

13.6. **Theorem.** *If X has all finite moments define*

$$R = \liminf_{k\to\infty} \left(\frac{|\mathbb{E}X^k|}{k!}\right)^{-\frac{1}{k}}.$$

Then R is the radius of convergence for φ. That is, φ extends analytically to $\mathbb{C}$ within radius R and for $t \in \mathbb{R}$ with $|t| < R$ we have that

$$\varphi(t) = \sum_{k=0}^{\infty} \frac{i^k}{k!} t^k \mathbb{E}X^k.$$

Intuitively this follows because of the Taylor expansion of the exponential, and for all t within this radius the sum and expectation can be switched.

If $R = \infty$ then φ takes the form above and can be completely constructed from its moments, which in turn implies that the distribution can be reconstructed in this case.

In general it is a nontrivial question, known as the *moment problem*, whether the moments completely determine a distribution. The $R = \infty$ case above is a positive example.

Proof. None given. □

13.7. **Example.** If $X \sim \text{Cauchy}(0, 1)$ then $\mathbb{E}X^k$ does not exists for all k. $\varphi(t) = e^{-|t|}$ which is not analytic.

13.8. **Example** (Is this a characteristic function?)**.** Use the properties above to determine.

$\sin(t)$? No. Because $\sin(0) = 0$.

$\cos(t)$? Well

$$\cos(t) = \frac{e^{it} + e^{-it}}{2}$$

so it is the characteristic function of a discrete random variable which takes values ± 1 with equal probability.

$\frac{1}{2}(1+\cos(t))$? Well

$$\frac{1}{2}(1+\cos(t)) = \frac{1}{2} + \frac{e^{it}}{4} + \frac{e^{-it}}{4}$$

which will be the characteristic function of a random variable which is 0 w.p. 1/2 and ± 1 with probability 1/4.

Any convex combination of characteristic functions is a characteristic function! Let $I = 1$ with probability p and $I = 0$ with probability $q = 1 - p$, independent of random variables X and Y. Let $Z = IX + (1-I)Y$. Then $Z = X$ with probability p and Y with probability q. By the partition theorem

$$\varphi_Z(t) = \mathbb{E}e^{itZ} = p\mathbb{E}e^{itX} + q\mathbb{E}e^{itY} = p\varphi_X(t) + q\varphi_Y(t).$$

e^{-t^4}? $\frac{d^2}{dt^2}e^{-t4}|_{t=0} = 0$ hence $\mathbb{E}X^2 = 0 \implies X = 0$ almost surely. Then $\varphi(t) = 1$ – contradiction.

13.9. **Lemma** (Fubini's theorem)**.** *The order of integration can be swapped if the double integrable of the absolute value is finite.*

13.10. **Lemma** (Jensen's inequality for integration)**.** *Let f be a convex function and g an integrable function, then*

$$f\left(\frac{1}{b-a}\int_a^b g(x)dx\right) \leq \frac{1}{b-a}\int_a^b f(g(x))dx.$$

Proof. Let $X \sim \text{Uniform}(a,b)$, then $\frac{1}{b-a}\int_a^b g(x)dx = \mathbb{E}g(X)$. By Jensen's inequality for expectation

$$f\left(\int_a^b \frac{1}{b-a}g(x)dx\right) = f(\mathbb{E}g(X)) \leq \mathbb{E}f(g(X)) = \int_a^b \frac{1}{b-a}f(g(x))dx.$$

□

The example to be used below is $f(\cdot) = |\cdot|$. In this case $b-a$ cancels out, and

$$\left|\int_a^b g(x)dx\right| \leq \int_a^b |g(x)|dx.$$

13.11. **Theorem** (Inversion formula)**.** *If $\varphi(t)$ is the characteristic function of a probability measure μ then*

(a) *for every $a < b \in \mathbb{R}$*

$$\mu(a,b) + \frac{1}{2}\mu\{a\} + \frac{1}{2}\mu\{b\} = \lim_{c\to\infty}\frac{1}{2\pi}\int_{-c}^{c}\frac{e^{-ita} - e^{-itb}}{it}\varphi(t)dt.$$

(b) *if $\int_{-\infty}^{\infty}|\varphi(t)|dt < \infty$ then F is absolutely continuous, that is, X has continuous distribution and*

$$f(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}\varphi(t)dt.$$

(c) *Similarly, $\mu\{a\} = \lim_{c\to\infty}\frac{1}{2c}\int_{-c}^{c} e^{-ita}\varphi(t)dt$.*

(d) *if φ is the characteristic function of an integer valued random variable then*

$$p(k) = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{-itk}\varphi(t)dt.$$

Proof. (a). Define

$$\phi_c = \frac{1}{2\pi}\int_{-c}^{c} \frac{e^{-ita} - e^{-itb}}{it}\varphi(t)dt = \frac{1}{2\pi}\int_{-c}^{c} \frac{e^{-ita} - e^{-itb}}{it}\int_{-\infty}^{\infty} e^{itx}dF(x)dt.$$

By Fubini's theorem, if

$$\frac{1}{2\pi}\int_{-c}^{c}\int_{-\infty}^{\infty}\left|\frac{e^{-ita} - e^{-itb}}{it}e^{itx}\right|dF(x)dt < \infty$$

then swapping the order of integration is permitted. Note that

$$\begin{aligned}\left|\frac{e^{-ita} - e^{-itb}}{it}e^{itx}\right| = \left|\frac{e^{-ita} - e^{-itb}}{it}\right| &= \left|\int_a^b e^{-itx}dx\right| \\ &\leq \int_a^b |e^{-itx}|dx \quad \longleftarrow \text{ Jensen's for integrals} \\ &= b - a.\end{aligned}$$

Therefore,

$$\frac{1}{2\pi}\int_{-c}^{c}\int_{-\infty}^{\infty}\left|\frac{e^{-ita} - e^{-itb}}{it}e^{itx}\right|dF(x)dt \leq \frac{1}{2\pi}\int_{-c}^{c}\underbrace{\int_{-\infty}^{\infty}(b-a)dF(x)}_{\int_{-\infty}^{\infty}dF(x)=1}dt = 2c(b-a) < \infty.$$

Hence we can swap. Therefore,

$$\phi_c = \int_{-\infty}^{\infty}\underbrace{\frac{1}{2\pi}\int_{-c}^{c}\frac{e^{-ita} - e^{-itb}}{it}e^{itx}dt}_{:=\psi_c}dF(x).$$

$$\psi_c = \frac{1}{2\pi}\int_{-c}^{c}\frac{e^{-ita} - e^{-itb}}{it}e^{itx}dt = \frac{1}{2\pi}\int_{-c}^{c}\frac{e^{it(x-a)}}{it} - \frac{e^{it(x-b)}}{it}dt.$$

Note that when we integrate e^{it}/it over some interval $[-c, c]$ it is the same as integrating $e^{it}/2it - e^{-it}/2it$ (symmetry) hence (using that $\int_{-\infty}^{\infty}\frac{\sin(u)}{u}du = \pi$), assuming $a < b$,

$$\begin{aligned}\psi_c &= \frac{1}{2\pi}\int_{-c}^{c}\frac{\sin(t(x-a))}{t}dt - \frac{1}{2\pi}\int_{-c}^{c}\frac{\sin(t(x-b))}{t}dt \\ &= \frac{1}{2\pi}\int_{-c(x-a)}^{c(x-a)}\frac{\sin(u)}{u}du - \frac{1}{2\pi}\int_{-c(x-b)}^{c(x-b)}\frac{\sin(u)}{u}du \\ &\underset{c\to\infty}{=} \begin{cases} a < b < x & \frac{1}{2\pi}\pi - \frac{1}{2\pi}\pi = 0, \\ a < x < b & \frac{1}{2\pi}\pi - \frac{1}{2\pi}(-\pi) = 1, \\ x < a < b & \frac{1}{2\pi}(-\pi) - \frac{1}{2\pi}(-\pi) = 0, \\ x = a & 0 - \frac{1}{2\pi}(-\pi) = \frac{1}{2}, \\ x = b & \frac{1}{2\pi}\pi - 0 = \frac{1}{2}. \end{cases}\end{aligned}$$

The various cases come from the positivity of the limits. Can we swap limit and integral? $|\psi_c| \leq k$ uniformly in x and c so yes, by Dominated Convergence.

$$\lim_{c\to\infty}\phi_c = \lim_{c\to\infty}\int_{-\infty}^{\infty}\psi_c dF(x) = \int_{-\infty}^{\infty}\lim_{c\to\infty}\psi_c dF(x) = \mu(a,b) + \frac{1}{2}\mu\{a\} + \frac{1}{2}\mu\{b\}.$$

(b). Let $f(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}\varphi(t)dt$ and show that this is the density.

$$\int_b^a f(x)dx = \frac{1}{2\pi}\int_b^a\int_{-\infty}^{\infty} e^{-itx}\varphi(t)dtdx \longleftarrow \text{Fubini } \int_b^a\int_{-\infty}^{\infty}|e^{-itx}\varphi(t)|dtdx = (b-a)\underbrace{\int_{-\infty}^{\infty}|\varphi(t)|dt}_{<\infty}$$

$$= \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_b^a e^{-itx}dx\varphi(t)dt$$
$$= \frac{1}{2\pi}\int_{-\infty}^{\infty}\frac{e^{-ita}-e^{-itb}}{it}\varphi(t)dt$$
$$= \lim_{c\to\infty}\frac{1}{2\pi}\int_{-c}^{c}\frac{e^{-ita}-e^{-itb}}{it}\varphi(t)dt$$
$$= \mu(a,b) + \frac{1}{2}\mu\{a\} + \frac{1}{2}\mu\{b\} \implies \text{continuity.}$$

Note that by continuity we have that $\mu\{a\} = \mu\{b\} = 0$. With this the above display proves (b).

(c). Similarly as (a):

$$\frac{1}{2c}\int_{-c}^{c} e^{-ita}\varphi(t)dt = \frac{1}{2c}\int_{-c}^{c} e^{-ita}\int_{-\infty}^{\infty} e^{itx}dF(x)dt$$
$$= \int_{-\infty}^{\infty}\frac{1}{2c}\int_{-c}^{c} e^{it(x-a)}dtdF(x). \quad \longleftarrow \text{Fubini}$$

When $x = a$,

$$\psi_c(x) := \frac{1}{2c}\int_{-c}^{c} e^{it(x-a)}dt = \frac{1}{2c}\int_{-c}^{c} 1dt = 1,$$

whereas for $x \neq a$

$$\psi_c(x) := \frac{1}{2c}\int_{-c}^{c} e^{it(x-a)}dt = \frac{e^{ic(x-a)}-e^{-ic(x-a)}}{2ic(x-a)} = \frac{\sin(c(x-a))}{c(x-a)} \to 0$$

as $c \to \infty$. Also, $|\psi_c(x)| \leq 1$ uniformly in c and x, therefore Dominated Convergence applies and

$$\lim_{c\to\infty}\frac{1}{2c}\int_{-c}^{c} e^{-ita}\varphi(t)dt = \lim_{c\to\infty}\int_{-\infty}^{\infty}\psi_c(x)dF(x) = \int_{-\infty}^{\infty}\lim_{c\to\infty}\psi_c(x)dF(x) = \mu\{a\}.$$

(d).

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} e^{-itk}\varphi(t)dt = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{-itk}\int_{-\infty}^{\infty} e^{itx}dF(x)dt$$

$$\int_{-\pi}^{\pi}\int_{-\infty}^{\infty}|e^{-itk}e^{itx}|dF(x)dt = 2\pi < \infty \longrightarrow \quad = \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\pi}^{\pi} e^{it(x-k)}dtdF(x).$$

Similarly to the previous calculation, for $x = k$

$$\psi(x) := \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{it(x-k)}dt = \frac{1}{2\pi}\int_{-\pi}^{\pi} 1dt = 1,$$

whereas for any other integer x,

$$\psi(x) := \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{it(x-k)}dt = \frac{e^{i\pi(x-k)}-e^{-i\pi(x-k)}}{2\pi i(x-k)} = \frac{\sin(\pi(x-k))}{\pi(x-k)} = 0,$$

and we are not interested in non-integer x as the distribution is integer-valued (dF has only discrete masses at integers). Plugging this back we have

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} e^{-itk}\varphi(t)dt = \int_{-\infty}^{\infty}\psi(x)dF(x) = \mu\{k\} = p(k).$$

□

13.12. **Corollary** (of (c)). *If $\varphi(t) \to 0$ as $|t| \to \infty$ then μ has no point masses (converse not true).*

Proof. Let X and Y have characteristic function φ and be independent. Then

$$\varphi_{X-Y}(t) = \mathbb{E}e^{it(X-Y)} = \varphi(t)\overline{\varphi(t)} = |\varphi(t)|^2.$$

By the inversion formula (c),

$$\mathbb{P}(X=Y) = \mathbb{P}(X-Y=0) = \lim_{c\to\infty}\frac{1}{2c}\int_{-c}^{c} e^{-it0}|\varphi(t)|^2dt = \lim_{c\to\infty}\frac{1}{2c}\int_{-c}^{c}|\varphi(t)|^2dt.$$

Claim that this limit is 0. Fix $\varepsilon > 0$. As $|\varphi(t)|^2 \to 0$ as $|t| \to \infty$ there exists $\delta > 0$ such that $|t| > \delta$ implies $|\varphi(t)|^2 < \varepsilon$. Note that for all $|t| \le \delta$ we have $|\varphi(t)|^2 \le 1$ as a requirement of φ (properties above).

For all $c > \delta$,

$$\begin{aligned}\frac{1}{2c}\int_{-c}^{c}|\varphi(t)|^2dt &= \frac{1}{2c}\int_{-c}^{-\delta}|\varphi(t)|^2dt + \frac{1}{2c}\int_{-\delta}^{\delta}|\varphi(t)|^2dt + \frac{1}{2c}\int_{\delta}^{c}|\varphi(t)|^2dt\\ &\le \frac{(c-\delta)\varepsilon}{2c} + \frac{2\delta}{2c} + \frac{(c-\delta)\varepsilon}{2c} \to \frac{\varepsilon}{2} + 0 + \frac{\varepsilon}{2} = \varepsilon.\end{aligned}$$

So $\mathbb{P}(X=Y) = 0$. Let $a \in \mathbb{R}$ be arbitrary. Then $\mathbb{P}(X=a)^2 = \mathbb{P}(X=a)\mathbb{P}(Y=a) \le \mathbb{P}(X=Y) = 0$. □

14. Weak convergence, Central limit theorem

We start with technicalities to prove the Prokhorov's Theorem, then the Continuity Lemma which connects the convergence of characteristic functions to that of distributions. The Central Limit Theorem will then be an easy consequence of this Lemma.

14.1. **Definition** (Relatively compact)**.** A family $\mathcal{P}$ of distributions is *relatively compact* if any sequence $(\mu_n) \subseteq \mathcal{P}$ has a subsequence which converges weakly to a probability distribution.

14.2. **Example.** $F_n = \mathbb{1}_{\{X>n\}}$ is not relatively compact since any subsequence converges to 0, which is not a probability distribution.

14.3. **Definition** (Tight)**.** A sequence of distributions (μ_n) is *tight* if for all $\varepsilon > 0$ there exists a compact set K (that is bounded and closed in $\mathbb{R}^n$) such that

$$\sup_{n\in\mathbb{N}}\mu_n(\Omega\backslash K) \le \varepsilon.$$

14.4. **Definition** (Generalised distribution function)**.** A *generalised distribution function* is a function $G : \mathbb{R} \to [0,1]$ which is

- non-decreasing;
- continuous from the right;
- $0 \le \lim_{x\to-\infty} G(x) \le \lim_{x\to\infty} G(x) \le 1$.

14.5. **Lemma.** *If (G_n) is a sequence of generalised distribution functions then there exists a subsequence n_k and a generalised distribution function G such that for all $x \in \mathbb{R}$ for which G is continuous*

$$\lim_{k\to\infty} G_{n_k}(x) = G(x).$$

Proof. Let T be a countably dense subset of $\mathbb{R}$, e.g. $\mathbb{Q}$. Then $T = \{x_1, x_2, \ldots\}$, $\overline{T} = \mathbb{R}$.

(1) As $G_n(x_1) \in [0,1]$, G_n has a subsequence $G_{1,n}$ such that $G_{1,n}(x_1) \to g_1$.

(2) As $G_{1,n}(x_2) \in [0,1]$, $G_{1,n}$ has a subsequence $G_{2,n}$ such that $G_{2,n}(x_2) \to g_2$.

(3) As $G_{2,n}(x_3) \in [0,1]$, $G_{2,n}$ has a subsequence $G_{3,n}$ such that $G_{3,n}(x_3) \to g_3$.

⋮

Hence we have a subsequence $G_{n,n}$ which converges for all $x \in T$, namely $G_{n,n}(x_k) \to g_k$. Define

$$G_T(x_k) = \lim_{n\to\infty} G_{n,n}(x_k) = g_k.$$

We now wish to extend G_T onto the whole of $\mathbb{R}$. Define, for $x \in \mathbb{R}$,

$$G(x) = \inf\{G_T(y) : y \in T, y > x\}.$$

Claim: $G(x)$ is non-decreasing. Let $x_1 \le x_2$. Then

$$G(x_1) = \inf\{G_T(y) : y \in T, y > x_1\} \le \inf\{G_T(y) : y \in T, y > x_2\} = G(x_2).$$

Claim: $G(x)$ is right continuous. Fix $x \in \mathbb{R}$ and let (x_n) be any sequence such that $x_n \searrow x$. Well as x_n is decreasing

$$\inf\{G_T(y) : y \in T, y > x_1\} \ge \inf\{G_T(y) : y \in T, y > x_2\} \ge \cdots \ge \inf\{G_T(y) : y \in T, y > x_n\}$$

and therefore this sequence $G(x_n)$ has a limit which we need to show equals $G(x)$. It is clear from monotonicity that $\lim_{n\to\infty} G(x_n) \ge G(x)$, so by contradiction suppose

$$\lim_{n\to\infty} G(x_n) > G(x) = \inf\{G_T(y) : y \in T, y > x\}.$$

Then there must be a $T \ni y > x$ with $G_T(y) < \lim_n G(x_n)$, which is a contradiction as for large n, $x < x_n < y$ occurs and T is dense in $\mathbb{R}$.

Claim: The third property of generalised distribution functions is immediate from $0 \le G_T(y) \le 1$ for all $y \in T$.

It remains to show that $G_{n,n}(x) \to G(x)$ where G continuous. Fix z_0 such that G is continuous at z_0. Take $y \in T$ with $y > z_0$. Because $G_{n,n}$ is a generalised distribution function ergo non-decreasing,

$$\limsup_{n\to\infty} G_{n,n}(z_0) \le \limsup_{n\to\infty} G_{n,n}(y) = G_T(y).$$

Taking the infinum over $y \in T$, $y > z_0$,

$$\limsup_{n\to\infty} G_{n,n}(z_0) \le \inf\{G_T(y) : y \in T, y > z_0\} = G(z_0).$$

Take $z_1 < z_0$, since T is dense in $\mathbb{R}$ there exists $y \in T$ such that $z_1 < y < z_0$. By definition $G(z_1) \le G_T(y)$ so

$$G(z_1) \le G_T(y) = \lim_{n\to\infty} G_{n,n}(y) = \liminf_{n\to\infty} G_{n,n}(y) \le \liminf_{n\to\infty} G_{n,n}(z_0).$$

By the continuity at z_0, $G(z_0) = \lim_{z_1 \nearrow z_0} G(z_1)$ and therefore

$$G(z_0) \le \liminf_{n\to\infty} G_{n,n}(z_0).$$

All in all

$$G(z_0) \le \liminf_{n\to\infty} G_{n,n}(z_0) \le \limsup_{n\to\infty} G_{n,n}(z_0) \le G(z_0)$$

which implies the result. □

14.6. **Theorem** (Prokhorov)**.** *If Ω is a complete separable (contains a countably dense subset) metric space (e.g. $\mathbb{R}$) then a family of distributions $\mathcal{P}$ is relatively compact if and only if it is tight.*

Proof. We give the proof for $\Omega = \mathbb{R}$. Suppose $\mathcal{P}$ is tight. Let F_n be a sequence of distributions in $\mathcal{P}$. Let F_{n_k} be a subsequence which converges to a generalised distribution function G (exists by the previous lemma). We show that tightness implies that the limiting distribution is proper.

Fix $\varepsilon > 0$, $a < b \in \mathbb{R}$ such that

$$\sup_{n\in\mathbb{N}} \mu_n(\mathbb{R}\backslash[a,b]) \le \varepsilon$$

which can be done as the sequence μ_n is tight.

Fact. If $f : \mathbb{R} \to \mathbb{R}$ is an increasing function then the set of points for which f is discontinuous is at most countable.

Proof of the fact. Let $A \subseteq \mathbb{R}$ be the set of points for which f is discontinuous. For $x \in A$, $f(x^-) < f(x^+)$ since f is increasing. Hence there exists $q \in \mathbb{Q}$ such that $f(x^-) < q < f(x^+)$. Define the function $g : A \to \mathbb{Q}$, $x \mapsto q$. This construction is not unique, but we claim that for any selection $g(x) \in \mathbb{Q}$ under this procedure is injective. Indeed, fix $x, y \in A$ such that $x \neq y$ and without loss of generality assume $x < y$. Then $f(x^+) \leq f(y^-)$ (equality option possible if f constant between x and y). Moreover

$$f(x^-) < g(x) < f(x^+) \leq f(y^-) < g(y) < f(y^+)$$

and so $g(x) \neq g(y)$. So g is injective. So A is at most countable.

Thus there exists $a' < a < b < b'$ such that G is continuous at a' and b'. Then for all k

$$1 - \varepsilon \leq \mu_{n_k}[a, b] \leq \mu_{n_k}(a', b'] = F_{n_k}(b') - F_{n_k}(a') \to G(a') - G(b')$$

as $k \to \infty$ as G is continuous at these points. Taking $b' \to \infty$ and $a' \to -\infty$ we see that for all $\varepsilon > 0$

$$\lim_{b' \to \infty} G(b') - \lim_{a' \to \infty} G(a') \geq 1 - \varepsilon \implies = 1.$$

Conversely, suppose that (μ_n) is relatively compact. For a contradiction, suppose (μ_n) is *not* tight.

Then there exists $\varepsilon > 0$ such that for all $K \subset \Omega$ compact, $\sup_{n \in \mathbb{N}} \mu_n(\Omega \backslash K) > \varepsilon$.

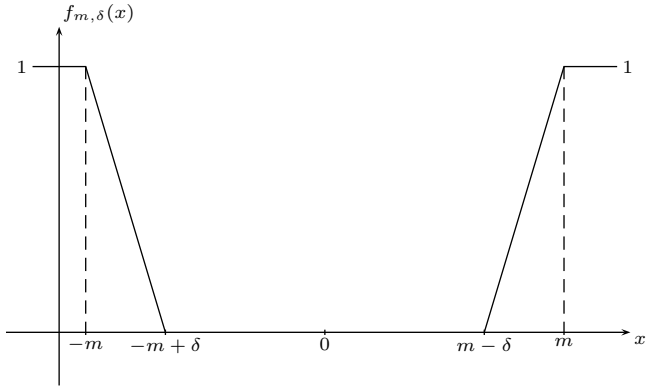
$\implies \exists \varepsilon > 0, \forall K$ compact, $\exists \mu_n$ such that $\mu_n(\mathbb{R} \backslash K) > \varepsilon$.

$\implies \exists \varepsilon > 0, \forall m, \exists \mu_n$ such that $\mu_n(\mathbb{R} \backslash [-m, m]) > \varepsilon$.

Hence there exists a subsequence μ_{n_m} of μ_n such that for all m, $\mu_{n_m}(\mathbb{R} \backslash [-m, m]) > \varepsilon$.

Since μ_n is relatively compact there exists a further subsequence $\mu_{n_{m_k}} \xrightarrow{w} \mu$ for some probability measure μ.

For all $m, \delta > 0$ define $f_{m,\delta}$ by



Clearly $\mu_{n_{m_k}}(\mathbb{R} \backslash [-m, m]) = \int_{\mathbb{R} \backslash [-m,m]} f_{m,\delta}(x) d\mu_{n_{m_k}}(x)$ as on here it's equal to 1. Hence

$$\mu_{n_{m_k}}(\mathbb{R} \backslash [-m, m]) \leq \int_{\mathbb{R}} f_{m,\delta}(x) d\mu_{n_{m_k}}(x) \overset{k \to \infty}{\longrightarrow} \int_{\mathbb{R}} f_{m,\delta}(x) d\mu(x) \overset{\delta \searrow 0}{\searrow} \int_{\mathbb{R}} \mathbb{1}_{\mathbb{R} \backslash [-m,m]}(x) d\mu(x) = \mu(\mathbb{R} \backslash [-m, m]).$$

The first convergence follows from the definition of weak convergence (convergence of expectation of continuous bounded function, the Lebesgue integral w.r.t. the measure is the expectation).

Since for all m we have $\mu_{n_m}(\mathbb{R} \backslash [-m, m]) > \varepsilon$, the weakly convergent subsequence also satisfies this.

$$\varepsilon \leq \mu_{n_{m_k}}(\mathbb{R} \backslash [-m_k, m_k]) \quad \implies \quad \varepsilon \leq \limsup_{k \to \infty} \mu_{n_{m_k}}(\mathbb{R} \backslash [-m_k, m_k]).$$

m_k is a strictly increasing sequence and hence for sufficiently large k, $m_k \geq m$. So eventually, $[-m, m] \subseteq [-m_k, m_k] \implies \mathbb{R} \backslash [-m_k, m_k] \subseteq \mathbb{R} \backslash [-m, m] \implies \mu_{n_{m_k}}(\mathbb{R} \backslash [-m_k, m_k]) \leq \mu_{n_{m_k}}(\mathbb{R} \backslash [-m, m])$.

Hence

$$\varepsilon \leq \limsup_{k \to \infty} \mu_{n_{m_k}}(\mathbb{R} \backslash [-m_k, m_k]) \leq \limsup_{k \to \infty} \mu_{n_{m_k}}(\mathbb{R} \backslash [-m, m]) \leq \mu(\mathbb{R} \backslash [-m, m]) \overset{m \to \infty}{\longrightarrow} 0.$$

The last inequality follows from the construction of $f_{m,\delta}$.

Contradiction as $\varepsilon > 0$. Hence (μ_n) is tight. $\square$

14.7. **Lemma** (1)**.** *Let (μ_n) be a tight sequence. If every subsequence of (μ_n) which is weakly convergent converges weakly to the same limit μ then $\mu_n \xrightarrow{w} \mu$.*

Proof. Assume that $\mu_n \overset{w}{\not\to} \mu$. Then there exists a bounded and continuous function f such that

$$\int f(x)d\mu_n(x) \not\to \int f(x)d\mu(x).$$

By definition of convergence in reals, there exists $\varepsilon > 0$ and a subsequence $\mu_{n'}$ such that

$$\left| \int f(x)d\mu_{n'}(x) - \int f(x)d\mu(x) \right| \geq \varepsilon \qquad (\star)$$

for all n'.

By Prokhrov's theorem, as (μ_n) is tight, it is relatively compact. Hence every sequence in (μ_n) contains a subsequence which is weakly convergent to a probability measure. Hence there exists a subsequence n'' of n' which is weakly convergent to a probability distribution. By the assumption, $\mu_{n''} \xrightarrow{w} \mu$ (all weakly convergent subsequences $\xrightarrow{w} \mu$). Hence

$$\int f(x)d\mu_{n''}(x) \to \int f(x)d\mu(x).$$

This contradicts $(\star)$. □

14.8. **Lemma** (2)**.** *Let (μ_n) be a tight sequence on $\mathbb{R}$. Let μ_n have characteristic function φ_n, that is $\varphi_n(t) = \int e^{itx}d\mu_n(x)$. Then μ_n converges weakly if and only if for all $t \in \mathbb{R}$ $\lim_{n\to\infty} \varphi_n(t)$ exists.*

Proof. Let (μ_n) be tight.

Suppose $\mu_n \xrightarrow{w} \mu$. Then for all bounded continuous f,

$$\int f(x)d\mu_n(x) \to \int f(x)d\mu(x).$$

In particular as e^{itx} is bounded (by 1) and continuous

$$\varphi_n(t) = \int e^{itx}d\mu_n(x) \to \int e^{itx}d\mu(x) = \varphi(t).$$

Conversely suppose that $\lim_{n\to\infty} \varphi_n(t)$ exists. As (μ_n) is tight, it is relatively compact by Prokhorov. Hence there exists a subsequence $\mu_{n'} \xrightarrow{w} \mu$ for some probability measure μ.

Suppose that $\mu_n \not\to \mu$ (leads to contradiction).

If $\mu_{n'}$ is the only weakly convergent subsequence then $\mu_n \xrightarrow{w} \mu$ by lemma 1. So by the negation of lemma 1, there exists another subsequence of n (not of n') say n'' such that $\mu_{n''} \xrightarrow{w} \nu \neq \mu$.

As $\lim_{n\to\infty} \varphi_n(t)$ exists, all subsequences of $\varphi_n(t)$ converge to the same limit. Therefore

$$\varphi_\nu(t) = \lim_{n''\to\infty} \varphi_{n''}(t) = \lim_{n\to\infty} \varphi_n(t) = \lim_{n'\to\infty} \varphi_{n'}(t) = \varphi_\mu(t).$$

Since characteristic functions are unique to measures we have a contradiction. □

14.9. **Lemma** (3)**.** *There exists $k > 0$ such that for all $a > 0$*

$$\int_{|x|\geq\frac{1}{a}} dF(x) \leq \frac{k}{a}\int_0^a \big[1 - \mathrm{Re}\varphi(t)\big]dt.$$

Proof.

$$\begin{aligned}
\frac{1}{a}\int_0^a \big[1-\mathrm{Re}\varphi(t)\big]dt &= \frac{1}{a}\int_0^a \left[1-\int_{-\infty}^{\infty}\cos(tx)dF(x)\right]dt \\
&= \frac{1}{a}\int_0^a\int_{-\infty}^{\infty}\big[1-\cos(tx)\big]dF(x)dt \quad \longleftarrow \text{as } \int dF(x)=1 \\
\text{Fubini} \longrightarrow &= \frac{1}{a}\int_{-\infty}^{\infty}\int_0^a \big[1-\cos(tx)\big]dt\,dF(x) \\
&= \frac{1}{a}\int_{-\infty}^{\infty}\left[t-\frac{\sin(tx)}{x}\right]_{t=0}^{a}dF(x) \\
&= \frac{1}{a}\int_{-\infty}^{\infty}\left[t-\frac{\sin(tx)}{x}\right]_{t=0}^{a}dF(x) \\
&= \frac{1}{a}\int_{-\infty}^{\infty} a-\frac{\sin(ax)}{x}dF(x) \\
&= \int_{-\infty}^{\infty}\underbrace{1-\frac{\sin(ax)}{ax}}_{\geq 0}dF(x) \\
&\geq \int_{|ax|\geq 1}\underbrace{1-\frac{\sin(ax)}{ax}}_{>0}dF(x) \\
&\geq \int_{|ax|\geq 1}\frac{1}{k}dF(x)
\end{aligned}$$

Reducing the size of the domain of integration makes the value smaller since the integrand is nonnegative everywhere. Since we never integrate over $x=0$ on the reduced domain, for which the integrand is equal to 0, the integrand is strictly greater than 0. Hence there exists k such that the integrand is strictly greater than k^{-1}. □

14.10. **Theorem** (Continuity lemma)**.** *Let (μ_n) be a sequence of distributions and (φ_n) the corresponding sequence of characteristic functions.*

(1) *If $\mu_n \xrightarrow{w} \mu$ then $\varphi_n(t)\to\varphi(t)$ as $n\to\infty$.*

(2) *If for all $t\in\mathbb{R}$, $\varphi(t)=\lim_{n\to\infty}\varphi_n(t)$ exists and is continuous at $t=0$ then φ is the characteristic function of a probability measure μ and $\mu_n\xrightarrow{w}\mu$.*

Proof. (1) is trivial by the definition of weak convergence, noting that $t\mapsto e^{itx}$ is a bounded continuous function.

To show (2) we

(1) Show that μ_n is tight;

(2) Invoke lemma 2 (which is the result when tightness is *assumed*).

By lemma 3,

$$\mu_n\Big(\mathbb{R}\backslash\big(-\frac{1}{a},\frac{1}{a}\big)\Big)=\int_{|x|\geq\frac{1}{a}}d\mu_n(x)\leq\frac{k}{a}\int_0^a\big[1-\mathrm{Re}\varphi_n(t)\big]dt.$$

Aim to apply DCT. Bounded? $|1-\mathrm{Re}\varphi_n(t)|\leq 1+|\mathrm{Re}\varphi_n(t)|\leq 1+|\varphi_n(t)|\leq 2$ (application of Jensen's inequality, above). Furthermore $1-\mathrm{Re}\varphi_n(t)\to 1-\mathrm{Re}\varphi(t)$ for all $t\in\mathbb{R}$ by the assumption. By dominated convergence

$$\mu_n\Big(\mathbb{R}\backslash\big(-\frac{1}{a},\frac{1}{a}\big)\Big)\leq\frac{k}{a}\int_0^a\big[1-\mathrm{Re}\varphi_n(t)\big]dt\xrightarrow{n\to\infty}\frac{k}{a}\int_0^a\big[1-\mathrm{Re}\varphi(t)\big]dt\xrightarrow{a\to 0}0.$$

To prove the last convergence note that φ continuous at $t=0$ implies that $\lim_{t\to 0}\varphi(t)=\varphi(0)=1$. Fix $\varepsilon>0$. Then there exists $\delta>0$ such that $|t|<\delta$ implies $|1-\varphi(t)|<\varepsilon$.

So $|1-\mathrm{Re}\varphi(t)|=|\mathrm{Re}(1-\varphi(t))|=\sqrt{|1-\varphi(t)|^2-|\mathrm{Im}(1-\varphi(t))|^2}\le|1-\varphi(t)|<\varepsilon$ too. Moreover, for any $a<\delta$,

$$\left|\frac{k}{a}\int_0^a\big[1-\mathrm{Re}\varphi(t)\big]dt\right|\le\frac{k}{a}\int_0^a|1-\mathrm{Re}\varphi(t)|dt\le\frac{k\varepsilon}{a}\int_0^a dt=k\varepsilon.$$

So the integral can be made arbitrarily small.

So for fixed ε there exists sufficiently small a such that $\mu_n(\mathbb{R}\backslash(-1/a,1/a))<\varepsilon$. Choose any compact set K such that $(-1/a,1/a)\subset K$, then $\mu_n(\mathbb{R}\backslash K)<\mu_n(\mathbb{R}\backslash(-1/a,1/a))<\varepsilon$. Then $\sup_{n\in\mathbb{N}}\mu_n(\mathbb{R}\backslash K)\le\varepsilon$. Hence (μ_n) is tight.

Applying lemma 2 gives the result. □

14.11. **Lemma.** *If $\mathbb{E}|X|^n<\infty$ for some $n\ge1$ then as $t\to0$*

$$\varphi(t)=\sum_{k=0}^{n}\frac{(it)^k}{k!}\mathbb{E}X^k+\frac{(it)^n}{n!}o(1).$$

Proof. By the Taylor expansion of exp we have for some random Θ_1,Θ_2; $|\Theta_1|\le1,|\Theta_2|\le1$,

$$\begin{aligned}\varphi(t)&=\mathbb{E}e^{itX}\\&=\mathbb{E}\left[\sum_{k=0}^{n-1}\frac{(itX)^k}{k!}+\frac{(itX)^n}{n!}\big(\cos(\Theta_1tX)+i\sin(\Theta_2tX)\big)\right]\\&=\mathbb{E}\left[\sum_{k=0}^{n}\frac{(itX)^k}{k!}+\frac{(itX)^n}{n!}\big(\cos(\Theta_1tX)+i\sin(\Theta_2tX)-1\big)\right]\\&=\sum_{k=0}^{n}\frac{(it)^k}{k!}\mathbb{E}X^k+\frac{(it)^n}{n!}\mathbb{E}\left[X^n\big(\cos(\Theta_1tX)+i\sin(\Theta_2tX)-1\big)\right].\end{aligned}$$

Note that $|\cos(\Theta_1tX)+i\sin(\Theta_2tX)-1|\le3$ and goes to 0 as $t\to0$ hence by DCT the whole expectation goes to 0 as $t\to0$ (as $\mathbb{E}X^n<\infty$) so equal to $o(1)$. □

14.12. **Theorem** (Weak law of large numbers)**.** *Let (X_n) be an i.i.d. sequence of random variables with $\mathbb{E}X_i=m$ finite. Then*

$$\frac{1}{n}\sum_{k=1}^{n}X_k\xrightarrow{P}m.$$

Proof. Show converges weakly and then that it converges in probability.

Let $S_n=\sum_{k=1}^n X_k$. Then

$$\varphi_{\frac{S_n}{n}}(t)=\varphi_{S_n}\left(\frac{t}{n}\right)=\prod_{i=1}^{n}\varphi_{X_i}\left(\frac{t}{n}\right)=\left[\varphi_{X_1}\left(\frac{t}{n}\right)\right]^n=\left(1+\frac{itm}{n}+\frac{ito(1)}{n}\right)^n\to e^{itm}.$$

Hence $\frac{1}{n}\sum_{k=1}^n X_k\xrightarrow{w}m$. The following fact implies the result.

Fact. Weak convergence and convergence in probability are equivalent when the limit is a constant.

Proof of the fact. Suppose $X_n\xrightarrow{w}c$. Then

$$\mathbb{P}(|X_n-c|>\varepsilon)\le\mathbb{P}(X_n\le c-\varepsilon)+\mathbb{P}(X_n>c+\varepsilon)=F_n(c-\varepsilon)+1-F_n(c+\varepsilon)\to0.$$

Conversely suppose $X_n \xrightarrow{P} c$. Then

$$F_n(c-\varepsilon)+1-F_n(c+\varepsilon) \le \mathbb{P}(X_n \le c-\varepsilon)+\mathbb{P}(X_n \ge c+\varepsilon) = \mathbb{P}(|X_n-c| \ge \varepsilon) \to 0,$$

or $\limsup_n(F_n(c-\varepsilon)+1-F_n(c+\varepsilon)) \le 0$, $\liminf_n(F_n(c+\varepsilon)-F_n(c-\varepsilon)) \ge 1$ which forces $\lim_n F_n(c+\varepsilon)=1$ and $\lim_n F_n(c-\varepsilon)=0$ by $0 \le F_n(c+\varepsilon),\ F_n(c-\varepsilon) \le 1$. □

14.13. **Theorem** (Global central limit theorem)**.** *Let (X_n) be an i.i.d. sequence of random variables with $\mathbb{E}|X_k|^2 < \infty$ ($\implies \mathbb{E}|X_k| < \infty$ by Cauchy-Schwarz or Lyapunov). Put $\mathbb{E}X_i = m$, $\mathrm{Var}X_i = \sigma^2$ and $S_n = \sum_{k=1}^n X_k$. Then*

$$\frac{S_n - nm}{\sigma\sqrt{n}} \xrightarrow{w} \mathcal{N}(0,1).$$

Proof. By the continuity lemma we show $\mathbb{P}(\frac{S_n-nm}{\sigma\sqrt{n}} \le a) \to \Phi(a)$.

Case 1. $m=0$, $\sigma=1$.

$$\varphi_{\frac{S_n}{\sqrt{n}}} = \varphi_{S_n}\left(\frac{t}{\sqrt{n}}\right) = \prod_{k=1}^n \varphi_{X_i}\left(\frac{t}{\sqrt{n}}\right) = \left[\varphi_{X_i}\left(\frac{t}{\sqrt{n}}\right)\right]^n = \left[1+\underbrace{\frac{itm}{\sqrt{n}}}_{m=0} - \frac{t^2}{2n}\underbrace{\mathbb{E}X^2}_{=1} - \frac{t^2}{2n}o(1)\right]^n \to e^{-\frac{t^2}{2}},$$

the Standard Normal characteristic function.

Case 2. General m and σ.

$$\frac{S_n-nm}{\sigma\sqrt{n}} = \frac{\sum_{k=1}^n \frac{X_k-m}{\sigma}}{\sqrt{n}}.$$

Each term has mean 0 and variance 1 so the previous case applies. □

14.14. **Theorem** (Poisson central limit theorem)**.** *Let $X \sim \mathrm{Poisson}(\lambda)$. Then*

$$\frac{X-\lambda}{\sqrt{\lambda}} \xrightarrow{w} \mathcal{N}(0,1).$$

Proof. Can prove using characteristic functions (homework) or using a sneaky trick and the central limit theorem.

Note $X \overset{d}{=} \sum_{k=1}^{\lfloor\lambda\rfloor} X_k$ where $X_k \sim \mathrm{Poisson}(\lambda/\lfloor\lambda\rfloor)$. Then use central limit theorem. □

14.15. **Theorem** (Cramèr-Berry-Essèen)**.** *Let X_i be iid, $\mathbb{E}X_i = 0$; $\mathbb{E}|X_i|^2 = \sigma^2$ and $\mathbb{E}|X_i|^3 < \infty$. Then there exists $c \in [\frac{1}{\sqrt{2\pi}}, 0.8)$ such that*

$$\sup_a \left|\mathbb{P}\left(\frac{S_n}{\sigma\sqrt{n}} < a\right) - \Phi(a)\right| \le \frac{c\mathbb{E}|X_1|^3}{\sigma^3\sqrt{n}}.$$

No proof given.

14.16. **Example.** $X_i \sim \mathrm{Bernoulli}(1/2)$. Note that this isn't centered! Let $\widetilde{X}_i = X_i - \mathbb{E}X_i$. Then $\mathbb{E}\widetilde{X}_i = 0; \mathrm{Var}\widetilde{X}_i = \mathrm{Var}X = 1/4$. Take $n=10$ i.i.d. copies of these. On one hand, $\mathbb{P}(S_{10} \le 6) = 0.8281$ by direct calculation with the Binomial$(10,1/2)$ mass function. On the other hand, the CLT gives

$$\mathbb{P}(S_{10} \le 6) = \mathbb{P}(S_{10} \le 6.5) = \mathbb{P}\Big(\frac{S_{10}-10\cdot\frac{1}{2}}{\frac{1}{2}\cdot\sqrt{10}} \le \frac{6.5-10\cdot\frac{1}{2}}{\frac{1}{2}\cdot\sqrt{10}}\Big) \simeq \Phi\Big(\frac{6.5-10\cdot\frac{1}{2}}{\frac{1}{2}\cdot\sqrt{10}}\Big) \simeq 0.8389.$$

The difference is $0.8389 - 0.8281 = 0.0108$. The Cramèr-Berry-Essèen bound would be, with $\mathbb{E}\widetilde{X}_i^3 = 1/8$, $\frac{0.8/8}{1/8\cdot\sqrt{10}} \simeq 0.25$ for this difference.

14.17. **Theorem** (Local central limit theorem)**.** *X_i iid, $\mathbb{E}X_i = 0$; $\mathrm{Var}X_i = \sigma^2 < \infty$ with X_i absolutely continuous with bounded density function. Then*

$$\lim_{n\to\infty} \sup_x \left| \frac{d}{dx}\mathbb{P}\left(\frac{S_n}{\sigma\sqrt{n}} \le x\right) - \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \right| = 0.$$

That is, the densities become close.

No proof given.

14.18. **Theorem** (Lindeberg central limit theorem)**.** *Let $\{X_{n,m} : n \in \mathbb{N}, 1 \le m \le N_n\}$ be a set of random variables such that*

- $\mathbb{E}X_{n,m} = 0$;
- $\mathrm{Var}X_{n,m} = \sigma_{n,m}^2 < \infty$;
- $S_n := \sum_{m=1}^{N_n} X_{n,m}$;
- *For fixed n, $X_{n,1}, X_{n,2}, \ldots, X_{n,N_n}$ are independent;*
- $\sigma_n^2 := \mathrm{Var}S_n = \sum_{m=1}^{N_n} \sigma_{n,m}^2$;
- *The Lindeberg condition holds, that is, for all $\varepsilon > 0$,*

$$\lim_{n\to\infty} \frac{1}{\sigma_n^2} \sum_{m=1}^{N_n} \mathbb{E}(X_{n,m}^2; |X_{n,m}| > \varepsilon\sigma_n) = 0.$$

Then $\mathbb{P}(S_n/\sigma_n \le a) \to \Phi(a)$ as $n \to \infty$.

No proof given.

14.19. *Remark* (Intuition behind the Lindeberg condition). Fix $\varepsilon > 0$.

$$\begin{aligned}
\frac{\sigma_{n,m}^2}{\sigma_n^2} &= \frac{\mathbb{E}X_{n,m}^2}{\sigma_n^2} \\
&= \frac{\mathbb{E}(X_{n,m}^2; |X_{n,m}| > \varepsilon\sigma_n) + \mathbb{E}(X_{n,m}^2; |X_{n,m}| \le \varepsilon\sigma_n)}{\sigma_n^2} \\
&\le \frac{\mathbb{E}(X_{n,m}^2; |X_{n,m}| > \varepsilon\sigma_n) + \varepsilon^2\sigma_n^2}{\sigma_n^2} \\
&= \frac{1}{\sigma_n^2}\mathbb{E}(X_{n,m}^2; |X_{n,m}| > \varepsilon\sigma_n) + \varepsilon^2.
\end{aligned}$$

The Lindeberg condition states that this quantity stays small, in other words no individual variance $\sigma_{n,m}^2$ plays a macroscopically visible role in the sum σ_n^2.