Notes on PSTAT 213

Haosheng Zhou

Sept, 2022

Contents

Simple Random Walk

$S_n = X_1 + ... + X_n$ is a SRW with X_i *i.i.d.* starting from $S_0 = 0$. T_b denotes the first hitting time of S_n to b, $\mathbb{P}(X_i = 1) = p, \mathbb{P}(X_i = -1) = q, p + q = 1$.

Theorem 1. *(Hitting Time Theorem)* $\forall b \neq 0$ *such that* $\frac{n+b}{2} \in \{0, 1, ..., n\}$,

$$\mathbb{P}(T_b = n) = \frac{|b|}{n}\mathbb{P}(S_n = b) = \frac{|b|}{n}\binom{n}{\frac{n+b}{2}} p^{\frac{n+b}{2}} q^{\frac{n-b}{2}} \ (n \geq 1) \tag{1}$$

Proof. Prove by counting paths. It's obvious that if $p = q$, each path consisting of points (t, S_t) $(t = 0, 1, ..., n)$ has same probability of appearing. Now p, q are not necessarily the same, so if a fixed path has a moving upward and $n - a$ moving downward, the probability of appearing is just

$$\frac{\binom{n}{a}}{2^n} p^a q^{n-a} \tag{2}$$

If now a path hits b at time n for the first time, it should first hit b at time n, which means that there are $\frac{n+b}{2}$ going upward and $\frac{n-b}{2}$ going downward. Each path that hits b at time n has same probability of appearing, which is $\frac{\binom{n}{\frac{n+b}{2}}}{2^n} p^{\frac{n+b}{2}} q^{\frac{n-b}{2}}$. As a result, the problem reduces to counting the number of all paths within those paths that have also hit b between time 0 to n.

We do a translation for all the paths such that now we start at $(0, -b)$ and want to count the number of paths that ends at $(n, 0)$ but has also hit 0 in between. This count is just the sum of the number of paths that starts at $(0, -b)$ and ends at $(n - 1, 1)$ but has also hit 0 in between and the number of paths that starts at $(0, -b)$ and ends at $(n - 1, -1)$ but has also hit 0 in between. Assume WLOG that $b > 0$, notice that the first count is

$$\binom{n-1}{\frac{n+b}{2}} \tag{3}$$

and the second count is due to reflection principle that it's just the number of paths that starts at $(0, b)$ and ends at $(n - 1, -1)$ which is

$$\binom{n-1}{\frac{n+b}{2}} \tag{4}$$

As a result, the sum should be

$$2\binom{n-1}{\frac{n+b}{2}} \tag{5}$$

The number of path that starts at $(0, b)$ and ends at $(n, 0)$ is

$$\binom{n}{\frac{n+b}{2}} \tag{6}$$

So if a path is conditioned on already starting at $(0,0)$ and ending at (n,b), it has probability of hitting b in between as

$$\frac{2\binom{n-1}{\frac{n+b}{2}}}{\binom{n}{\frac{n+b}{2}}} = \frac{n-b}{n} \tag{7}$$

if a path is conditioned on already starting at $(0,0)$ and ending at (n,b), it has probability of not hitting b in between as

$$\frac{b}{n} \tag{8}$$

That's why $\mathbb{P}(T_b = n) = \frac{b}{n}\mathbb{P}(S_n = b)$ for $b > 0$ and the theorem is proved. The similar proof holds for $b < 0$. □

Remark. *If we want to know the distribution of T_0, we also have to lift the time at 0 to the time at 1 (consider whether S_1 is 1 or -1) since reflection can't be applied for when the path starts or ends at 0.*

Theorem 2. *Set the **maximum process** $M_n = \max_{0\le k\le n} S_k$ for symmetric SRW S_n, then*

$$\forall r \ge 1, \mathbb{P}(M_n \ge r, S_n = v) = \begin{cases} \mathbb{P}(S_n = v) & v \ge r \\ \mathbb{P}(S_n = 2r - v) & v < r \end{cases} \tag{9}$$

Proof. If $v \ge r$, then $\mathbb{P}(M_n \ge r, S_n = v) = \mathbb{P}(S_n = v)$ naturally.

For the other case, let's count the number of paths. The number of paths from $(0,0)$ to $(n, 2r-v)$ is

$$\binom{n}{\frac{n+2r-v}{2}} \tag{10}$$

The number of paths from $(0,0)$ to (n,v) that has hit r in between is equal to the number of paths from $(0,-r)$ to $(n, v-r)$ that has hit 0 in between. By reflection principle, this is just the number of paths from $(0,r)$ to $(n, v-r)$, which is

$$\binom{n}{\frac{n+v-2r}{2}} \tag{11}$$

same to the count above, so it's proved.

Another Proof:

Since the SRW is Markov process and $T_r < \infty$ *a.s.*, strong Markov property tells us

$$S_n^{T_r} = S_{n+T_r} - S_{T_r} = S_{n+T_r} - r \tag{12}$$

is also a SRW and is independent of $\mathscr{F}_{T_r}$.

Let's then do calculations:

$$\mathbb{P}(M_n \geq r, S_n = v) = \mathbb{P}(T_r \leq n, S_n = v) \tag{13}$$
$$= \mathbb{P}\left(T_r \leq n, S^{T_r}_{n-T_r} = v - r\right) \tag{14}$$
$$= \mathbb{P}\left(T_r \leq n, -S^{T_r}_{n-T_r} = v - r\right) \tag{15}$$

the last step is due to the fact that $T_r \in \mathscr{F}_{T_r}, S^{T_r}_n \overset{d}{=} -S^{T_r}_n$ and that $S^{T_r}_n$ is independent of $\mathscr{F}_{T_r}$.

$$\mathbb{P}(M_n \geq r, S_n = v) = \mathbb{P}\left(T_r \leq n, -S^{T_r}_{n-T_r} = v - r\right) \tag{16}$$
$$= \mathbb{P}(T_r \leq n, S_n = 2r - v) \tag{17}$$
$$= \mathbb{P}(S_n = 2r - v) \tag{18}$$

□

Remark. *By this reflection principle, we see that for $r \geq 0$,*

$$\mathbb{P}(M_n \geq r) = \sum_{v=-n,-n+2,\ldots,n} \mathbb{P}(M_n \geq r, S_n = v) \tag{19}$$
$$= \sum_{v<r} \mathbb{P}(S_n = 2r - v) + \sum_{v \geq r} \mathbb{P}(S_n = v) \tag{20}$$
$$= \mathbb{P}(S_n = r) + \mathbb{P}(S_n \geq r + 1) + \mathbb{P}(S_n = 2r + n) + \ldots + \mathbb{P}(S_n = 2r - r + 1) \tag{21}$$
$$= \mathbb{P}(S_n = r) + \mathbb{P}(S_n \geq r + 1) + \mathbb{P}(S_n \geq r + 1) \tag{22}$$
$$= \mathbb{P}(S_n = r) + 2\mathbb{P}(S_n \geq r + 1) \tag{23}$$

that's why we get

$$\mathbb{P}(M_n = r) = \mathbb{P}(M_n \geq r) - \mathbb{P}(M_n \geq r + 1) \tag{24}$$
$$= \mathbb{P}(S_n = r) + 2\mathbb{P}(S_n \geq r + 1) - \mathbb{P}(S_n = r + 1) - 2\mathbb{P}(S_n \geq r + 2) \tag{25}$$
$$= \mathbb{P}(S_n = r) + 2\mathbb{P}(S_n = r + 1) - \mathbb{P}(S_n = r + 1) \tag{26}$$
$$= \mathbb{P}(S_n = r) + \mathbb{P}(S_n = r + 1) \tag{27}$$

To calculate probability like $\mathbb{P}(M_8 = 6)$, just use the formula to get

$$\mathbb{P}(M_8 = 6) = \mathbb{P}(S_8 = 6) + \mathbb{P}(S_8 = 7) \tag{28}$$
$$= \frac{\binom{8}{1}}{2^8} = \frac{1}{32} \tag{29}$$

Generating Function of SRW

0 Hitting Time

Now in the general setting, p probability going upward and q going downward with $p+q=1$. Now

$$p_0(n) = \mathbb{P}(S_n = 0) \tag{30}$$

and

$$f_0(n) = \mathbb{P}(S_1 \neq 0, ..., S_{n-1} \neq 0, S_n = 0) \tag{31}$$

where $f_0(n)$ gives the probability mass of first hitting time T_0. There respective generating functions are denoted

$$P_0(s) = \sum_{n=0}^{\infty} p_0(n) s^n \tag{32}$$

$$F_0(s) = \sum_{n=0}^{\infty} f_0(n) s^n \tag{33}$$

$$\tag{34}$$

then since SRW is Markov, use the Markov property w.r.t. 1 unit of time translation to get

$$p_0(0) = 1, f_0(0) = 0 \tag{35}$$

$$\forall n \geq 1, p_0(n) = \mathbb{P}(S_n = 0) \tag{36}$$

$$= \sum_{k=1}^{n} \mathbb{P}(T_0 = k)\, \mathbb{P}(S_n = 0 | T_0 = k) \tag{37}$$

$$= \sum_{k=1}^{n} \mathbb{P}(T_0 = k)\, \mathbb{P}(S_{n-k} = 0) \tag{38}$$

$$= \sum_{k=1}^{n} f_0(k) p_0(n-k) \tag{39}$$

to compare the coefficient, proved that

$$P_0(s) = 1 + P_0(s) F_0(s) \tag{40}$$

Note that

$$P_0(s) = \sum_{n=0}^{\infty} \mathbb{P}(S_n = 0)\, s^n \tag{41}$$

$$= \sum_{n=0,2,\ldots} \binom{n}{\frac{n}{2}} (pq)^{\frac{n}{2}} s^n \tag{42}$$

$$= \sum_{n=0}^{\infty} \binom{2n}{n} (pqs^2)^n \tag{43}$$

$$= \sum_{n=0}^{\infty} \frac{(2n-1)!!2^n n!}{n!n!} (pqs^2)^n \tag{44}$$

$$= \sum_{n=0}^{\infty} (-4)^n \binom{-\frac{1}{2}}{n} (pqs^2)^n \tag{45}$$

$$= (1 - 4pqs^2)^{-\frac{1}{2}} \tag{46}$$

by the Taylor series.

As a result, plug in to get

$$F_0(s) = \frac{P_0(s) - 1}{P_0(s)} \tag{47}$$

$$= 1 - (1 - 4pqs^2)^{\frac{1}{2}} \tag{48}$$

From this generating function, we can investigate whether T_0 is almost surely finite or has finite expectation for general SRW. It's easy to see that

$$\mathbb{P}(T_0 < \infty) = \sum_{n=1}^{\infty} \mathbb{P}(T_0 = n) = F_0(1) = 1 - |p - q| \tag{49}$$

as a result, $T_0 < \infty$ *a.s.* **if and only if** $p = \frac{1}{2}$.

Taking derivative for $F_0(s)$ to get

$$F_0'(s) = 4pqs(1 - 4pqs^2)^{-\frac{1}{2}} \tag{50}$$

$$\mathbb{E}(T_0 \cdot \mathbb{I}_{T_0<\infty}) = F_0'(1) = \frac{4pq}{|p - q|} \tag{51}$$

as a result, $\mathbb{E}(T_0 \cdot \mathbb{I}_{T_0<\infty}) < \infty$ **if and only if** $p = \frac{1}{2}$.

In the context above, we investigate all generating functions of the stopping time T_0 which is the hitting time of 0. One can notice that actually this gives us the generating function of the i-th hitting time to 0, denoted T_0^i. By

Markov property,

$$\mathbb{P}\left(T_0^i = k\right) = \sum_{j=0}^{k} \mathbb{P}\left(T_0^{i-1} = j\right) \cdot \mathbb{P}\left(T_0^i = k | T_0^{i-1} = j\right) \tag{52}$$

$$= \sum_{j=0}^{k} \mathbb{P}\left(T_0^{i-1} = j\right) \cdot \mathbb{P}\left(T_0 = k - j\right) \tag{53}$$

so if we denote the generating function of T_0^i by $F_0^i(s)$, then

$$F_0^i(s) = \sum_{k=0}^{\infty} \mathbb{P}\left(T_0^i = k\right) \cdot s^k \tag{54}$$

$$= \sum_{k=0}^{\infty} \sum_{j=0}^{k} \mathbb{P}\left(T_0^{i-1} = j\right) \cdot \mathbb{P}\left(T_0 = k - j\right) \cdot s^k \tag{55}$$

$$= F_0^{i-1}(s) \cdot F_0(s) \tag{56}$$

$$= [F_0(s)]^i \tag{57}$$

it's then easy to see that

$$\mathbb{P}\left(T_0^i < \infty\right) = F_0^i(1) = [F_0(1)]^i = [1 - |p - q|]^i \tag{58}$$

so **SRW is recurrent if and only if** $p = \frac{1}{2}$. Naturally, let's investigate whether SRW is null recurrent when $p = \frac{1}{2}$.

$$\mathbb{E}(T_0^i \cdot \mathbb{I}_{T_0^i < \infty}) = \frac{d}{ds} F_0^i(s)|_{s=1} \tag{59}$$

$$= i[F_0(1)]^{i-1} \cdot F_0'(1) \tag{60}$$

$$= i[1 - |p - q|]^{i-1} \cdot \frac{4pq}{|p - q|} \tag{61}$$

so all states in SRW is null recurrent when $p = \frac{1}{2}$, which indicates a natural conclusion that there's no stationary distribution for symmetric SRW.

1 Hitting Time

One might find that generating functions for T_0 tells us nothing about the information of other hitting times, e.g. T_1. To get $F_1(s)$ as the generating function of T_1, we need to apply Markov property

$$\forall n > 1, \mathbb{P}\left(T_1 = n\right) = \mathbb{P}\left(T_1 = n | X_1 = 1\right) \cdot \mathbb{P}\left(X_1 = 1\right) + \mathbb{P}\left(T_1 = n | X_1 = -1\right) \cdot \mathbb{P}\left(X_1 = -1\right) \tag{62}$$

$$= q \cdot \mathbb{P}\left(T_1 = n | X_1 = -1\right) = q \cdot \mathbb{P}\left(T_2 = n - 1\right) \tag{63}$$

and it's obvious that $\mathbb{P}(T_1 = 1) = p$. To connect $F_1(s)$ with $F_2(s)$, it's natural to think of Markov property once more. Similar to what we have done for the i-th hitting time to 0, let's denote $F_i(1)$ as the generating function of T_i, the first hitting time to $i \geq 1$

$$\mathbb{P}(T_i = n) = \sum_{k=0}^{n} \mathbb{P}(T_i = n | T_1 = k) \cdot \mathbb{P}(T_1 = k) \tag{64}$$

$$= \sum_{k=0}^{n} \mathbb{P}(T_{i-1} = n - k) \cdot \mathbb{P}(T_1 = k) \tag{65}$$

here the strong Markov property is applied when $T_1 < \infty$ *a.s.* w.r.t. $\mathscr{F}_{T_1}$, note that when $T_1 = \infty$, $T_i = \infty$ so such equation still holds. This is telling us that getting the generating function of T_1 is equivalent to getting the generating function of any hitting time T_i

$$F_i(s) = [F_1(s)]^i \tag{66}$$

Return to the previous question on $F_1(s)$, this provides connection between $\mathbb{P}(T_1 = n)$ and $\mathbb{P}(T_2 = n - 1)$ that

$$F_1(s) = ps + \sum_{k=2}^{\infty} q \cdot \mathbb{P}(T_2 = k - 1) s^k \tag{67}$$

$$= ps + qs \cdot F_2(s) \tag{68}$$

$$= ps + qs \cdot [F_1(s)]^2 \tag{69}$$

solve this quadratic equation w.r.t. $F_1(s)$ to get

$$F_1(s) = \frac{1 \pm \sqrt{1 - 4pqs^2}}{2qs} \tag{70}$$

notice that any generating function shall satisfy $F_1(0) = 0$, so we only take one appropriate root as the generating function

$$F_1(s) = \frac{1 - \sqrt{1 - 4pqs^2}}{2qs} \tag{71}$$

naturally, one might calculate the quantity of one's interest that

$$\mathbb{P}(T_1 < \infty) = F_1(1) = \frac{1 - |p - q|}{2q} = \begin{cases} 1 & p \geq q \\ \frac{p}{q} & p < q \end{cases} \tag{72}$$

$$\mathbb{E}(T_1 \cdot \mathbb{I}_{T_1 < \infty}) = F_1'(1) = \frac{2p}{|p - q|} - \frac{1}{2q} + \frac{|p - q|}{2q} = \begin{cases} \frac{1}{p - q} & p > q \\ \frac{p}{q} \frac{1}{q - p} & p < q \\ \infty & p = q \end{cases} \tag{73}$$

in the more general case,

$$\mathbb{P}\left(T_i < \infty\right) = F_i(1) = \begin{cases} 1 & p \geq q \\ \left(\frac{p}{q}\right)^i & p < q \end{cases} \tag{74}$$

$$\mathbb{E}(T_i \cdot \mathbb{I}_{T_i<\infty}) = F_i'(1) = i[F_1(1)]^{i-1} \cdot F_1'(1) = \begin{cases} \frac{i}{p-q} & p > q \\ \left(\frac{p}{q}\right)^i \frac{i}{q-p} & p < q \\ \infty & p = q \end{cases} \tag{75}$$

as a result, $\mathbb{E}(T_i | T_i < \infty) = \frac{i}{|p-q|}$ holds generally.

A slight generalization is still the j-th hitting time to i, denoted T_i^j. To get its generating function $F_i^j(s)$, notice that

$$\mathbb{P}\left(T_i^j = n\right) = \sum_{k=0}^{n} \mathbb{P}\left(T_i^1 = k\right) \cdot \mathbb{P}\left(T_i^j = n | T_i^1 = k\right) \tag{76}$$

$$= \sum_{k=0}^{n} \mathbb{P}\left(T_i^1 = k\right) \cdot \mathbb{P}\left(T_0^{j-1} = n - k\right) \tag{77}$$

by Markov property, since after hitting i for the first time we are restarting the SRW from i and hitting 0 after restarting is equivalent to hitting i from the very start. As a result, $F_i^j(s) = F_i(s) \cdot F_0^{j-1}(s)$, by previous proofs, we know that $F_i(s) = [F_1(s)]^i$ and $F_0^{j-1}(s) = [F_0(s)]^{j-1}$, so

$$F_i^j(s) = [F_1(s)]^i \cdot [F_0(s)]^{j-1} \tag{78}$$

One is also able to calculate the probability and expectations one care about.

$$\mathbb{P}\left(T_i^j < \infty\right) = F_i^j(1) = [F_1(1)]^i \cdot [F_0(1)]^{j-1} \tag{79}$$

$$= \left(\frac{1 - |p-q|}{2q}\right)^i \cdot (1 - |p-q|)^{j-1} \tag{80}$$

and for the expectation

$$\mathbb{E}(T_i^j \cdot \mathbb{I}_{T_i^j<\infty}) = \frac{d}{ds} F_i^j(s)\Big|_{s=1} \tag{81}$$

$$= i[F_1(1)]^{i-1} \cdot F_1'(1) \cdot [F_0(1)]^{j-1} + [F_1(1)]^i \cdot (j-1)[F_0(1)]^{j-2} \cdot F_0'(1) \tag{82}$$

Remark. *The only important thing here is the **Markov property**. By selecting appropriate translation of time, one can always transform all j-th hitting time problems into the first hitting time of 0 and 1.*

Gambler's Ruin

Now for a general SRW, consider the exit time instead of the hitting time. Assume now the SRW starts at x and $T_{a,b}$ denotes the stopping time when SRW hits either a or b with $a < x < b$. It's quite clear that $T_{a,b} = T_a \wedge T_b$. This is telling us that if $p > q$ then $T_b < \infty$ *a.s.*, if $p < q$ then $T_a < \infty$ *a.s.*, if $p = q$ then $T_a, T_b < \infty$ *a.s.*. As a result, $T_{a,b} < \infty$ *a.s.* is almost surely finite.

As a result, a natural question to ask is that what's the probability that the SRW is exiting from a. Since $T_{a,b} < \infty$ *a.s.*,

$$\mathbb{P}_x\left(S_{T_{a,b}} = a\right) + \mathbb{P}_x\left(S_{T_{a,b}} = b\right) = 1 \tag{83}$$

where $\mathbb{P}_x$ means the probability measure of the SRW starting from x. Set

$$r(x) = \mathbb{P}_x\left(S_{T_{a,b}} = a\right) \tag{84}$$

and apply the Markov property to consider the first step

$$\begin{aligned} r(x) &= p \cdot \mathbb{P}_x\left(S_{T_{a,b}} = a | X_1 = 1\right) + q \cdot \mathbb{P}_x\left(S_{T_{a,b}} = a | X_1 = -1\right) && (85)\\ &= p \cdot \mathbb{P}_{x+1}\left(S_{T_{a,b}} = a\right) + q \cdot \mathbb{P}_{x-1}\left(S_{T_{a,b}} = a\right) && (86)\\ &= p \cdot r(x+1) + q \cdot r(x-1) && (87) \end{aligned}$$

here $\mathbb{P}_x\left(S_{T_{a,b}} = a | X_1 = 1\right) = \mathbb{P}_{x+1}\left(S_{T_{a,b}} = a\right)$ is due to the fact that we can stop the SRW at time 1 and restart it as if it starts from $x+1$ at time 0. The boundary condition is $r(a) = 1, r(b) = 0$.

Use the characteristic equation to solve the recurrence relationship:

$$\begin{aligned} \lambda &= p\lambda^2 + q && (88)\\ \lambda &= 1 \ or \ \frac{q}{p} && (89) \end{aligned}$$

we have to discuss whether $p = q$ since there might be roots with multiplicity.

If $p = q$, $\lambda = 1$ has multiplicity 2 so

$$r(x) = (C_1 x + C_2) \cdot 1^x \tag{90}$$

for some constant C_1, C_2, plug in boundary condition to solve out

$$C_1 = \frac{1}{a-b}, C_2 = -\frac{b}{a-b} \tag{91}$$

so we conclude

$$\mathbb{P}_x\left(S_{T_{a,b}} = a\right) = \frac{x-b}{a-b} \tag{92}$$

when $p = q = \frac{1}{2}$ in the symmetric case.

Now if $p \neq q$, there are two different roots and

$$r(x) = C_1 \cdot 1^x + C_2 \cdot \left(\frac{q}{p}\right)^x \tag{93}$$

for some constant C_1, C_2, plug in boundary condition to solve out

$$C_1 = -\frac{\left(\frac{q}{p}\right)^b}{\left(\frac{q}{p}\right)^a - \left(\frac{q}{p}\right)^b}, C_2 = \frac{1}{\left(\frac{q}{p}\right)^a - \left(\frac{q}{p}\right)^b} \tag{94}$$

so we conclude

$$\mathbb{P}_x\left(S_{T_{a,b}} = a\right) = \frac{\left(\frac{q}{p}\right)^x - \left(\frac{q}{p}\right)^b}{\left(\frac{q}{p}\right)^a - \left(\frac{q}{p}\right)^b} \tag{95}$$

when $p \neq q$ in the asymmetric case.

Law of Arcsine

The law of arcsine describes the asymptotic distribution of the **last hitting time to 0 and the overall time above 0** for **symmetric SRW**. The setting of the problem is that the last hitting time to 0 in time interval $[0, 2n]$ is defined as

$$L_{2n} = \sup\{m \leq 2n : S_m = 0\} \tag{96}$$

note that if the time is not bounded above, such random variable would not even be a stopping time (prove using strong Markov property by contradiction). Consider $0 \leq \frac{L_{2n}}{2n} \leq 1$, we would prove that such quotient has the law of arcsine (SRW starts from 0).

Let's start by observing that

$$\forall 0 \leq k \leq n, k \in \mathbb{N}, \mathbb{P}\left(L_{2n} = 2k\right) = \mathbb{P}\left(S_{2k} = 0\right) \cdot \mathbb{P}\left(L_{2n} = 2k | S_{2k} = 0\right) \tag{97}$$

$$= \mathbb{P}\left(S_{2k} = 0\right) \cdot \mathbb{P}\left(S_1 \neq 0, S_2 \neq 0, ..., S_{2n-2k} \neq 0\right) \tag{98}$$

by Markov property that we stop SRW at time $2k$ and restart it as if it starts from 0 at time 0. Due to former calculations, $\mathbb{P}\left(S_1 \neq 0, S_2 \neq 0, ..., S_{2n-2k} \neq 0\right) = \mathbb{P}\left(S_{2n-2k} = 0\right)$ so

$$\forall 0 \leq k \leq n, k \in \mathbb{N}, \mathbb{P}\left(L_{2n} = 2k\right) = \mathbb{P}\left(S_{2k} = 0\right) \cdot \mathbb{P}\left(S_{2n-2k} = 0\right) \tag{99}$$

now since

$$\mathbb{P}(S_{2k}=0)\cdot\mathbb{P}(S_{2n-2k}=0)=\frac{\binom{2k}{k}\cdot\binom{2n-2k}{n-k}}{2^{2n}} \tag{100}$$

$$\sim\frac{\sqrt{2k}\sqrt{(2n-2k)}}{2\pi k(n-k)}\ (n\to\infty) \tag{101}$$

$$=\frac{1}{\pi}\frac{1}{\sqrt{k(n-k)}}\ (n\to\infty) \tag{102}$$

by Stirling's formula, as a result, if $\frac{k}{n}\to x\ (n\to\infty)$

$$n\cdot\mathbb{P}(L_{2n}=2k)\to\frac{1}{\pi}\frac{1}{\sqrt{x(1-x)}} \tag{103}$$

which provides the main thought of the law of arcsine

$$\forall 0<a\le b<1,\mathbb{P}\left(a\le\frac{L_{2n}}{2n}\le b\right)\to\int_a^b\frac{1}{\pi}\frac{1}{\sqrt{x(1-x)}}\,dx\ (n\to\infty) \tag{104}$$

the details can be verified by proving $\frac{1}{\pi}\frac{1}{\sqrt{x(1-x)}}$ is the uniform limit on any compact set $[a,b]$. The "arcsine" comes from the fact that

$$\int_a^b\frac{1}{\pi}\frac{1}{\sqrt{x(1-x)}}\,dx=\frac{2}{\pi}\arcsin\sqrt{x}\Big|_{(a,b)}=\frac{2}{\pi}\arcsin\sqrt{b}-\frac{2}{\pi}\arcsin\sqrt{a} \tag{105}$$

Remark. *The law of arcsine is interesting if we think of the following bet that we are having 0 money at first and by tossing the coin we can get 1 or -1 for the same probability $\frac{1}{2}$, which means that this is a totally fair bet.*

However, by the law of arcsine,

$$\mathbb{P}\left(a\le\frac{L_{2n}}{2n}\le\frac{1}{2}\right)=\frac{1}{2}-\frac{2}{\pi}\arcsin\sqrt{a}\to\frac{1}{2}\ (a\to0,n\to\infty) \tag{106}$$

$$\mathbb{P}(L_{2n}\le n)\to\frac{1}{2}\ (n\to\infty) \tag{107}$$

which means that if we are keeping betting until time $2n$ where n is a large enough time, we have $\frac{1}{2}$ probability seeing that we are always having positive amount of money or negative amount of money after time n. So the asymptotic behavior of this fair bet model is now clear. If we are keeping betting until time $2n$ where n is a large enough time, we have $\frac{1}{4}$ probability of becoming a "winner", who always enjoys positive return in the latter half of the bet; we have $\frac{1}{4}$ probability of becoming a "loser", who always suffers from negative return in the latter half of the bet; we have $\frac{1}{2}$ probability of becoming a "normal person", whose return fluctuates up and down around 0.

This is telling us that even in totally fair games, the accumulation in time matters and presents ***concentration*** *phenomenon. This can be seen from the density $\frac{1}{\pi}\frac{1}{\sqrt{x(1-x)}}$ that the likelihood is lowest at $\frac{1}{2}$ but goes to ∞ at $0,1$, which means that extreme values of $\frac{L_{2n}}{2n}$ are far more likely to be observed (either never hits 0 or always hits 0).*

Eventually, one might notice that for a symmetric SRW starting from 0, the overall time it spends above 0 also has the law of arcsine.

$$\pi_{2n} = \#\{(t, S_t) : 0 \le t \le 2n, S_t \ge 0\} \tag{108}$$

be the overall time during $[0, 2n]$ such that SRW takes positive values. Then

$$\forall 0 < a \le b < 1, \mathbb{P}\left(a \le \frac{\pi_{2n}}{2n} \le b\right) \to \int_a^b \frac{1}{\pi} \frac{1}{\sqrt{x(1-x)}}\, dx \ (n \to \infty) \tag{109}$$

One might notice that actually $\pi_{2n} \overset{d}{=} L_{2n}$, the reason is that we can break up the event according to when the SRW first hits 0 and whether the SRW before the first hitting time to 0 is positive or negative

$$\mathbb{P}(\pi_{2n} = 2k) = \sum_{m=1}^{n} \mathbb{P}(\pi_{2n} = 2k | T_0 = 2m, S_{0\to T_0} \ge 0) \cdot \mathbb{P}(T_0 = 2m, S_{0\to T_0} \ge 0) \tag{110}$$

$$+ \sum_{m=1}^{n} \mathbb{P}(\pi_{2n} = 2k | T_0 = 2m, S_{0\to T_0} \le 0) \cdot \mathbb{P}(T_0 = 2m, S_{0\to T_0} \le 0) \tag{111}$$

$$= \frac{1}{2} \sum_{m=1}^{k} \mathbb{P}(\pi_{2n} = 2k | T_0 = 2m, S_{0\to T_0} \ge 0) \cdot \mathbb{P}(T_0 = 2m) \tag{112}$$

$$+ \frac{1}{2} \sum_{m=1}^{n-k} \mathbb{P}(\pi_{2n} = 2k | T_0 = 2m, S_{0\to T_0} \le 0) \cdot \mathbb{P}(T_0 = 2m) \tag{113}$$

$$= \frac{1}{2} \sum_{m=1}^{k} \mathbb{P}(\pi_{2n-2m} = 2k - 2m) \cdot \mathbb{P}(T_0 = 2m) + \frac{1}{2} \sum_{m=1}^{n-k} \mathbb{P}(\pi_{2n-2m} = 2k) \cdot \mathbb{P}(T_0 = 2m) \tag{114}$$

by Markov property. When the segment before $T_0 = 2m$ is positive, we just need another $2k - 2m$ to be positive in the remaining $2n - 2m$ time by restarting the SRW from 0. When the segment before $T_0 = 2m$ is negative, there's no contribution to π_{2n}, so we still need $2k$ to be positive in the remaining $2n - 2m$ time by restarting the SRW from 0.

Now notice that

$$\mathbb{P}(\pi_{2n} = 2n) = \mathbb{P}(S_1, ..., S_{2n} \ge 0) \tag{115}$$

$$= 2\mathbb{P}(S_1, ..., S_{2n} > 0) \tag{116}$$

$$= \mathbb{P}(S_1, ..., S_{2n} \ne 0) \tag{117}$$

$$= \mathbb{P}(S_{2n} = 0) \tag{118}$$

where the second equation comes from the reflection principle and the last equation is the property we have proved.

Now apply backward induction, the conclusion

$$\mathbb{P}\left(\pi_{2n}=2k\right)=\mathbb{P}\left(S_{2k}=0\right)\cdot\mathbb{P}\left(S_{2n-2k}=0\right) \tag{119}$$

holds for $k=n$. Assume that it's true for $k+1,k+2,...,n$, let's see whether it's true for k

$$\mathbb{P}\left(\pi_{2n}=2k\right)=\frac{1}{2}\sum_{m=1}^{k}\mathbb{P}\left(\pi_{2n-2m}=2k-2m\right)\cdot\mathbb{P}\left(T_0=2m\right)+\frac{1}{2}\sum_{m=1}^{n-k}\mathbb{P}\left(\pi_{2n-2m}=2k\right)\cdot\mathbb{P}\left(T_0=2m\right) \tag{120}$$

$$=\frac{1}{2}\sum_{m=1}^{k}\mathbb{P}\left(S_{2k-2m}=0\right)\cdot\mathbb{P}\left(S_{2n-2k}=0\right)\cdot\mathbb{P}\left(T_0=2m\right)+\frac{1}{2}\sum_{m=1}^{n-k}\mathbb{P}\left(S_{2k}=0\right)\cdot\mathbb{P}\left(S_{2n-2m-2k}=0\right)\cdot\mathbb{P}\left(T_0=2m\right) \tag{121}$$

$$=\frac{1}{2}\mathbb{P}\left(S_{2n-2k}=0\right)\mathbb{P}\left(S_{2k}=0\right)+\frac{1}{2}\mathbb{P}\left(S_{2k}=0\right)\mathbb{P}\left(S_{2n-2k}=0\right) \tag{122}$$

$$=\mathbb{P}\left(S_{2k}=0\right)\cdot\mathbb{P}\left(S_{2n-2k}=0\right) \tag{123}$$

where we used another Markov property that $\mathbb{P}\left(S_{2k}=0\right)=\sum_{m=1}^{k}\mathbb{P}\left(T_0=2m\right)\cdot\mathbb{P}\left(S_{2k-2m}=0\right)$. As a result, we have proved that

$$\pi_{2n}\overset{d}{=}L_{2n} \tag{124}$$

so the law of arcsine also holds.

Branching Process

The branching process Z_n is defined as

$$Z_0 = 1 \tag{125}$$

$$Z_n = X_1^n + ... + X_{Z_{n-1}}^n \tag{126}$$

where X_j^i denotes the j-th child in the i-th generation and all X_j^i are *i.i.d.* and all Z_k are independent of X_j^i.

This models the evolution of population. Z_k is the number of people in the k-th generation and each person in the k-th generation gives birth to some number of children which forms the $k+1$-th generation. Due to this natural formation, the distribution of X_j^i is called the **offspring distribution**, characterizing the number of possible children each person might have.

One of the most important thing to care about in the branching process is whether the population will extinct. Denote $\eta = \mathbb{P}(\exists n, Z_n = 0)$ as the **extinction probability**. It's then quite easy to see that this probability has something to do with μ, the mean of offspring distribution.

Generating Functions

Now assume we know about the generating function $G(s)$ of the offspring distribution. The generating function of Z_n, which is denoted $G_n(s)$ has the following form

$$G_n(s) = \mathbb{E}s^{Z_n} = \mathbb{E}[\mathbb{E}(s^{Z_n}|Z_{n-1})] \tag{127}$$

$$= \mathbb{E}[\mathbb{E}(s^{X_1^n+...+X_{Z_{n-1}}^n}|Z_{n-1})] \tag{128}$$

$$= \mathbb{E}[(\mathbb{E}s^{X_1^n})^{Z_{n-1}}] \tag{129}$$

$$= G_{n-1}(\mathbb{E}s^{X_1^n}) \tag{130}$$

$$= G_{n-1}(G(s)) \tag{131}$$

As a result, the generating function of Z_n is just derived by iterating $G(s)$ under composition for n times, denoted $G^{(n)}(s)$. The generating function for Z_n is just

$$G_n(s) = G^{(n)}(s) \tag{132}$$

Extinction Probability

Actually, when $\mu < 1$, the extinction is very easy to see by Markov inequality.

Theorem 3. ***(Subcritical Phase)*** *When $\mu < 1$, branching process extincts, i.e. $\eta = 1$.*

Proof. By Markov inequality,

$$\mathbb{P}(Z_n > 0) = \mathbb{P}(Z_n \geq 1) \leq \mathbb{E}Z_n \tag{133}$$

the expectation can be calculated

$$\mathbb{E}Z_n = \mathbb{E}[\mathbb{E}(Z_n|Z_{n-1})] = \mathbb{E}\mu Z_{n-1} = \mu \cdot \mathbb{E}Z_{n-1} \tag{134}$$

easily conclude that $\mathbb{E}Z_n = \mu^n$. When $\mu < 1$, $\mu^n \to 0$ $(n \to \infty)$. This is telling us that $\mathbb{P}(Z_n > 0) \to 0$ $(n \to \infty)$ so for fixed constant k, $\mathbb{P}(\forall n, Z_n > 0) \leq \mathbb{P}(Z_k > 0) \to 0$ $(k \to \infty)$ and we have proved that $\eta = 1$. □

When $\mu = 1$, there are two cases. If the offspring distribution has probability mass 1 at 1, then it's obvious that $\forall n, Z_n = 1$ so the process can't extinct. This is the trivial case. For the non-trivial case, we can prove that the extinction always happens.

Theorem 4. ***(Critical Phase)*** *When $\mu = 1$ and for X following the offspring distribution $\mathbb{P}(X = 1) < 1$, then branching process extincts, i.e. $\eta = 1$.*

Proof. The martingale property provides much more convenience in the proof of this theorem. First notice that Z_n is itself a martingale (generally $\frac{Z_n}{\mu^n}$ is a martingale) and it's non-negative. According to MG convergence theorem,

$$Z_n \overset{a.s.}{\to} Z_\infty \ (n \to \infty) \tag{135}$$

since Z_n can only take integer values, so do Z_∞ since the convergence is almost surely. This is telling us that it must be the case that $Z_n = Z$ *a.s.* eventually. In other words, there exists N such that for $\forall n > N$, $Z_n = Z$ *a.s.*.

For large enough n,

$$\mathbb{P}(Z_n = 0) = \sum_{k=0}^{\infty} \mathbb{P}(Z_{n-1} = k) \cdot \mathbb{P}(Z_n = 0|Z_{n-1} = k) \tag{136}$$

$$= \sum_{k=1}^{\infty} \mathbb{P}(Z_{n-1} = k) \cdot \mathbb{P}(Z_n = 0|Z_{n-1} = k) + \mathbb{P}(Z_{n-1} = 0) \tag{137}$$

$$= \sum_{k=1}^{\infty} \mathbb{P}(Z_{n-1} = k) \cdot [\mathbb{P}(X_1^n = 0)]^k + \mathbb{P}(Z_{n-1} = 0) \tag{138}$$

since the k people on the $n-1$-th generation are independent.

Taking the limit $n \to \infty$ to find

$$\mathbb{P}(Z_\infty = 0) = \sum_{k=1}^{\infty} \mathbb{P}(Z_\infty = k) \cdot [\mathbb{P}(X_1^n = 0)]^k + \mathbb{P}(Z_\infty = 0) \tag{139}$$

telling us that $\forall k \geq 1, \mathbb{P}(Z_\infty = k) = 0$. This is because (i): $\mathbb{P}(X_1^n = 0)$ does not depend on k and n is the probability mass at 0 of the offspring distribution. (ii): the offspring distribution has mean $\mu = 1$ but has point mass $p_1 < 1$ at

1, so if $p_0 = 0$, then $\sum_{k=0}^{\infty} kp_k = p_1 + 2p_2 + ... = \sum_{k\geq 1} p_k + \sum_{k\geq 2} p_k + ... > 1$ since $p_2 + p_3 + ... > 0$. As a result, we proved that $\mathbb{P}(X_1^n = 0) > 0$ and the conclusion above is correct.

Since $\forall k \geq 1, \mathbb{P}(Z_\infty = k) = 0$, we conclude that $Z_\infty = 0$ *a.s.* and $\eta = 1$ extinction almost surely happens.

□

For the most interesting case where $\mu > 1$, we would be able to solve out η as the fixed point of the generating function of the offspring distribution $G(s)$ (note that $G(1) = 1$ so 1 is a trivial fixed point).

Theorem 5. ***(Supercritical Phase)*** *For $\mu > 1$, $G(s)$ has fixed point $s^* \in [0, 1)$ and such fixed point is unique. Moreover, $\eta = s^*$ is the extinction probability.*

Proof. Now first consider the behavior of $G(s)$ on the interval $[0, 1)$. It's clear that $G(0) = p_0 \geq 0, G(1) = 1$, and for p_k denoting the probability masses of the offspring distribution,

$$G'(s) = \sum_{k=0}^{\infty} ks^{k-1}p_k \geq 0 \tag{140}$$

$$G'(1) = \mu > 1 \tag{141}$$

$$G''(s) = \sum_{k=0}^{\infty} k(k-1)s^{k-2}p_k > 0 \tag{142}$$

geometrically, $G(s)$ is strictly convex on $[0, 1]$ and the slope at 1 is larger than 1. Note that $G(s)$ is also smooth starting from p_0 and ending at 1. As a result, there would exist an intersection of the graph of $y = G(s)$ and the line $y = s$. To make this statement rigorous, consider

$$\forall h \in [0, 1), G(1) - G(1-h) = \int_{1-h}^{1} G'(s)\, ds = G'(\xi) \cdot h \tag{143}$$

for some $\xi \in [1 - h, 1]$. Setting $h \to 0$ to find $G'(\xi) \to \mu$ so $G(1 - h) \sim 1 - \mu h$ $(h \to 0)$. This is saying that at the points near $s = 1$, $G(1 - h) < 1 - h$. By the continuity of $G(s)$ and the intermediate value theorem, there must exist a fixed point on $[0, 1)$.

The strict convexity of G is telling us that if $s^* \in [0, 1)$ is the fixed point, then

$$\forall s \in (s^*, 1), G(s) < s \tag{144}$$

notice that $G''(s) > 0$ is strictly positive. To see this

$$G''(s) = \sum_{k=2}^{\infty} k(k-1)s^{k-2}p_k \tag{145}$$

$$\exists k \geq 2, p_k > 0 \tag{146}$$

because $\mu > 1$. Apply the equivalent condition for strict convexity to get

$$\forall s \in (s^*, 1), \frac{G(1) - G(s^*)}{1 - s^*} < \frac{G(1) - G(s)}{1 - s} \tag{147}$$

$$\forall s \in (s^*, 1), 0 < \frac{1 - G(s)}{1 - s} \tag{148}$$

$$\forall s \in (s^*, 1), G(s) < s \tag{149}$$

thus we have proved the uniqueness of such fixed point.

At last, prove that $s^* = \eta$ is just the extinction probability. To see this, let's first notice that

$$\{Z_1 = 0\} \subset \{Z_2 = 0\} \subset ... \subset \{Z_n = 0\} \subset ... \tag{150}$$

$$\eta = \mathbb{P}(\exists n, Z_n = 0) = \lim_{k \to \infty} \mathbb{P}(Z_k = 0) \tag{151}$$

let's denote $\theta_k = \mathbb{P}(Z_k = 0)$ the probability that extinction already happens at time k. To figure out the recurrence relationship of such θ_k,

$$\theta_k = \mathbb{P}(Z_k = 0) = \sum_{j=0}^{\infty} \mathbb{P}(Z_1 = j) \cdot \mathbb{P}(Z_k = 0 | Z_1 = j) \tag{152}$$

$$= \sum_{j=0}^{\infty} p_j \cdot (\mathbb{P}(Z_{k-1} = 0))^j = \sum_{j=0}^{\infty} p_j \cdot \theta_{k-1}^j = G(\theta_{k-1}) \tag{153}$$

since the descendants of the j children at generation 1 are independent and extinction at for each person of generation 1 the probability of seeing extinction at time k is the probability of seeing extinction after time $k-1$ starting from generation 0 (Markov property). This relationship is very important since it's showing us the fixed point iteration form.

Now the only objective is to prove that the fixed point iteration converges to the fixed point $\lim_{k \to \infty} \theta_k = s^*$. Since G is and increasing function, the series θ_k is increasing and is upper bounded by 1, so the limit exists $\theta_k \to \eta$ $(k \to \infty)$. By induction, it's easy to see that $\forall k, \theta_k \leq s^*$ so $\eta \leq s^*$. However, since G is continuous, setting $k \to \infty$ on both sides of $\theta_k = G(\theta_{k-1})$ gives

$$\eta = G(\eta) \tag{154}$$

by the existence and uniqueness of the fixed point s^* in the interval $[0, 1)$, we conclude that $\eta = s^*$.

□

Remark. *The key point here is* ***the fixed point iteration*** $\theta_k = G(\theta_{k-1})$ ***for*** $\theta_k = \mathbb{P}(Z_k = 0)$***, the probability that extinction already happens at time*** k *and the perspective that* ***the extinction probability*** η ***is just the limit of*** θ_k*, due to the structure of the branching process that if extinction happens at a time, it will last forever.*

Remark. *To get* $\theta_2 = \mathbb{P}(Z_2 = 0)$*, the probability of already seeing extinction at time 2, there are two ways to do*

this.

The first way is to use the generating function.

$$G_2(s) = \mathbb{P}(Z_2 = 0) + \mathbb{P}(Z_2 = 1) \cdot s + \mathbb{P}(Z_2 = 2) \cdot s^2 + ... \tag{155}$$

by setting $s = 0$, one get $\theta_2 = G_2(0)$.

The second way is to use the recurrence relationship we just mentioned.

$$\theta_2 = G(\theta_1) = G(G(\theta_0)) = G(G(0)) \tag{156}$$

However, one shall notice that these two ways are giving the same answer since the generating function for Z_n is just the n-th iterated composition of $G(s)$, so $G(G(0)) = G_2(0)$.

From this perspective, the extinction probability η is just the limit of the constant term $G_n(0)$ in the power series expansion of $G_n(s)$.

Total Progeny

The **total progeny** is defined as

$$T = \sum_{n=0}^{\infty} Z_n \tag{157}$$

the overall number of people in the branching process. It's easy to see that if $\eta = 1$, then $T < \infty$ *a.s.*, otherwise (T has positive probability taking the value ∞. Due to this fact, the generating function of the total progeny is defined as

$$G_T(s) = \mathbb{E}\left(s^T \cdot \mathbb{I}_{T<\infty}\right) \tag{158}$$

Theorem 6. ***(Generating Function of Total Progeny)***

$$\forall s \in [0, 1), G_T(s) = s \cdot G(G_T(s)) \tag{159}$$

Proof. Tear apart the expectation w.r.t. the value of Z_1 to get

$$G_T(s) = \sum_{k=0}^{\infty} \mathbb{P}(Z_1 = k) \cdot \mathbb{E}(s^T \cdot \mathbb{I}_{T<\infty} | Z_1 = k) \tag{160}$$

now under the condition that $Z_1 = k$, $T = 1 + T_1 + ... + T_k$ where T_j denotes the total progeny of the descendants of the j-th person in the first generation

$$G_T(s) = \sum_{k=0}^{\infty} p_k \cdot s \cdot \mathbb{E}(s^{T_1}...s^{T_k} \cdot \mathbb{I}_{T_1<\infty}...\mathbb{I}_{T_k<\infty} | Z_1 = k) \tag{161}$$

by the independence in the branching process, $T_1, ..., T_k, Z_1$ are independent and $T_1, ..., T_k$ are identically distributed, so

$$G_T(s) = \sum_{k=0}^{\infty} p_k \cdot s \cdot \mathbb{E}(s^{T_1}...s^{T_k} \cdot \mathbb{I}_{T_1<\infty}...\mathbb{I}_{T_k<\infty}) \tag{162}$$

$$= s \cdot \sum_{k=0}^{\infty} p_k \cdot \mathbb{E}(s^{T_1} \cdot \mathbb{I}_{T_1<\infty})...\mathbb{E}(s^{T_k} \cdot \mathbb{I}_{T_k<\infty}) \tag{163}$$

$$= s \cdot \sum_{k=0}^{\infty} p_k \cdot [G_{T_1}(s)]^k \tag{164}$$

$$= s \cdot G(G_{T_1}(s)) \tag{165}$$

at last notice that $T \overset{d}{=} T_1$ since the branching process starting from generation 0 with 1 person has the same distribution as the branching process starting from generation 1 with 1 person, so the distribution of the total progeny in these two cases has to be the same. So we conclude that

$$G_T(s) = s \cdot G(G_T(s)) \tag{166}$$

□

Remark. *By noticing the continuity of G_T and taking $s \to 1^-$, one may find that*

$$G_T(1) = G(G_T(1)) \tag{167}$$

when $\eta = 1$, it's obvious that $G_T(1) = \mathbb{P}(T < \infty) = 1$. When $\eta < 1$, however, $G_T(1) < 1$ and is the fixed point of the generating function $G(s)$. By previous proofs, we have argued that the fixed point of $G(s)$ in $[0, 1)$ exists and is uniquely η, so we conclude that $G_T(1) = \mathbb{P}(T < \infty) = \eta$ ***is another perspective for deriving the extinction probability****.*

Theorem 7. ***(Expected Total Progeny)*** *If $\mu \geq 1$, $\mathbb{E}T = \infty$ while if $\mu < 1$, $\mathbb{E}T = \frac{1}{1-\mu}$.*

Proof. Since $\mathbb{E}Z_n = \mu^n$, apply the monotone convergence theorem to see

$$\mathbb{E}T = \mathbb{E}\sum_{n=0}^{\infty} Z_n = \sum_{n=0}^{\infty} \mathbb{E}Z_n = \sum_{n=0}^{\infty} \mu^n = \frac{1}{1-\mu} \tag{168}$$

when $\mu < 1$ and is ∞ if $\mu = 1$. □

Branching Process and Random Walk

First, we establish the breadth-first search of traversing through all nodes of a tree. This provides us a way to uniquely label all the nodes in a tree with integers. In detail, label the root node as 1 and the leftest child of the root node as 2, the second leftest child of the root node as 3, etc. The spirit is that whenever a node has other

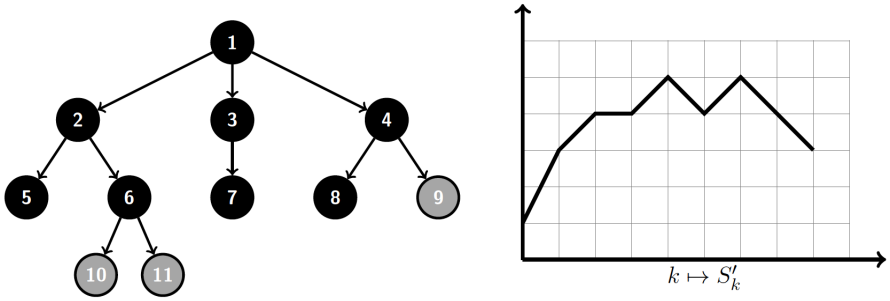

Figure 1: The Breadth-first Search of a Tree and the Construction of S'_n
Node 2 has depth 1, node 6 has depth 2. For all nodes of depth 3, their integer labels are $5, 6, 7, 8, 9$ from left to right, consecutive integers. Such labeling is done based on the breadth-first search of a tree.

unexplored nodes of the same generation (or say, the same level), traverse through the nodes on the same level prior to traversing through its descendants. As a result, under the breadth-first search of a tree, the integer assigned to each node is just the time that node is visited. It's easy to observe that for nodes in the same generation, the integer assigned to them are consecutive integers from left to right and for nodes in the different generations, the closer the distance between that node and the root node (called the depth) is, the less its integer label is. Refer to Figure 1 for the graphical illustration.

After labelling all nodes in a tree, we can now define two processes S'_n and S_n and show their connections with each other. **In the following context, exploration always means exploring the tree in the breadth-first search order.**

Let's first define S'_n and this requires us to define X'_n in prior. X'_n **is defined as the number of children node** n **has**. Here notice the **difference between children and descendants**. In Figure 1, node $10, 11$ are the descendants of node 2 but are not the children of node 2. Node 6 is the children and the descendants of node 2. One node's children have to be its descendants but not vice versa. To explain the definition of X'_n more clearly, we still refer to Figure 1. When the exploration comes to node 1, it has 3 children (node 2,3,4) so $X'_1 = 3$. When the exploration comes to node 2, it has 2 children (node 5,6) so $X'_2 = 2$. When the exploration comes to node 5, it has no children, so $X'_5 = 0$.

S'_n is defined through X'_n in the following way

$$S'_0 = 1 \tag{169}$$

$$S'_i = S'_{i-1} + X'_i - 1 = X'_1 + ... + X'_i - (i-1) \tag{170}$$

one might be a little bit confused here since the explanation and the meaning of S'_n is unclear. Let's first accept this

setting to calculate some of the values of S'_n.

$$\begin{cases} S'_0 = 1 \\ X'_1 = 3, S'_1 = 1 + 3 - 1 = 3 \\ X'_2 = 2, S'_2 = 3 + 2 - 1 = 4 \\ X'_3 = 1, S'_3 = 4 + 1 - 1 = 4 \\ X'_4 = 2, S'_4 = 4 + 2 - 1 = 5 \end{cases} \tag{171}$$

This is giving us the correct result which is consistent with the right subplot in Figure 1 showing the trajectory of S'_n. So what's the meaning of such S'_n? Actually S'_n **is the number of unexplored nodes within all nodes on the same level as node** n **and all the children of the explored nodes when the exploration comes to node** n. To see this in Figure 1, node 2 has 2 nodes on the same level (node 3,4) and 2 children (node 5,6), all of those 4 nodes are not explored prior to node 2, so $S'_2 = 4$ (the children of node 1 are node 2,3,4, where node 2 is being visited and node 3,4 have already been counted). For node 6, it has 4 nodes on the same level (node 5,7,8,9) and 2 children (node 10,11), only node 5 among all those 6 nodes has already been visited prior to the visit to node 6, so $S'_6 = 4 + 2 - 1 = 5$ (node 5 has no children and the children of all other explored nodes are either explored or counted). For node 8, it has 1 node on the same level (node 9) and no children, node 9 has not been explored yet. However, the two children (node 10,11) of the explored node 6 are not explored yet and not counted towards, so $S'_8 = 1 + 2 = 3$.

Remark. *There is another explanation for* S'_n*, which is shorter in words but easier to be mistaken.* S'_n ***is actually the number of unexplored nodes when the exploration comes to node*** n***, when the branching tree is built up dynamically****. Here the important point is that we do not know the structure of the tree at the beginning and we count the number of unexplored nodes as we draw out the branching tree simultaneously.*

For example, in figure 1, when we explore node 1, it immediately gives birth to 3 children and the tree only contains node 1,2,3,4 for the time being, so there are 3 unexplored nodes in the tree (node 2,3,4), then $S'_1 = 3$*. When we explore node 2, it immediately gives birth to 2 children and the tree only contains node 1,2,3,4,5,6 for the time being, so there are 4 unexplored nodes in the tree (node 3,4,5,6), then* $S'_2 = 4$*. The point is that when a node is explored, it immediately gives birth to its children but so its children appear in the tree but its grandchildren won't appear.*

Now we have managed to construct X'_n, S'_n based on a fixed realization of the branching process. Let's copy the definition of X'_n and assign it to X_n, copy the definition of S'_n and assign it to S_n.It's clear that X_n, S_n depend on the realization of the branching process. For a given branching process, whenever a realization of the branching process is fixed, the correspondent process X_n, S_n are also fixed.

In order to relate branching process to random walk, we want to create a replication for X_n, S_n. So how shall we do that? Recall that random walk is a stochastic process with *i.i.d.* increments. In our construction of S_n the $X_i - 1$ are just the increments, and they are *i.i.d.* due to the definition of the branching process! So why don't we

set X'_n to be *i.i.d.* random variables following the offspring distribution! By doing this, the S'_n defined as

$$S'_0 = 1, S'_i = X'_1 + ... + X'_i - (i-1),\ X_1, ..., X_i \sim \{p_k\}_{k\geq 0}\ i.i.d. \tag{172}$$

should be a random walk. In detail, S'_n **is a random walk starting from 1 with *i.i.d.* increments following the distribution of ξ where**

$$\mathbb{P}(\xi = k) = p_{k+1}\ (k = -1, 0, 1, ...) \tag{173}$$

note that here $\xi \overset{d}{=} O - 1$ where O is a random variable following the offspring distribution. To see this, write S'_n in the form of the sum of increments that

$$S'_n - S'_0 = \sum_{k=1}^{n} (X'_k - 1) \tag{174}$$

we would expect to see the connection between S_n, S'_n.

Remark. *One might be a little bit confused with the difference between S_n and S'_n. Just to clarify, S_n* ***is actually a "function" of the branching process*** *Z_n. For each realization of Z_n, we can always draw out a tree capturing all the branching happening and S_n is just constructed based on such a tree in the way that X_n denotes the number of children of node n and*

$$S_i = X_1 + ... + X_i - (i-1) \tag{175}$$

but the important point is that X_n, S_n totally depends on Z_n, in other words, $X_n, S_n \in \sigma(Z_n, n \in \mathbb{N})$.

In contrast, ***the S'_n here is totally independent of the branching process*** *Z_n since X'_n has nothing to do with the realization of the branching process. To construct S'_n, we just sample all i.i.d. random variables X'_n independent of the branching process and construct S'_n by*

$$S'_i = X'_1 + ... + X'_i - (i-1) \tag{176}$$

In the context above, we are using S'_n to show the construction of S_n just to remind the readers of the motivation and of the fact that those two processes are intuitively independent copies of each other and share the same structure (proved below).

Define the stopping time as the first hitting time to 0 of the random walk

$$T' = \inf\{n : S'_n = 0\} = \inf\{n : X'_1 + ... + X'_n = n - 1\} \tag{177}$$

where $T' = \infty$ if the inf does not exist. Once again, we notify the reader here that T' has nothing to do with the branching process Z_n. To be consistent with previous context, we still use T to denote the total progeny of the branching process. Then one would naturally expect that $\{S_i\}_{i\leq T} \overset{d}{=} \{S'_i\}_{i\leq T'}$.

Remark. *It's easy for one to notice that X_i' can only take values $0, 1, ...$ since it follows the offspring distribution. As a result, the increment of S_i' is $X_i' - 1$ which is at least -1. This is telling us that*

$$S_{T'}' = 0 \ a.s. \tag{178}$$

note that S_n totally depends on the branching process, so S_n is defined at all time before the total progeny T and $S_T = 0$ a.s..

A good question to ask here is that why the definition of T' and the truncation of the process S_n' at time T' are necessary. We know that $S_n \geq 0$ a.s. is always non-negative since it's defined based on the tree the branching process has generated. However $S_n' \geq 0$ a.s. does not hold. Since X_i' are $i.i, d.$ random variables, it is completely possible for us to observe $X_1' = X_2' = X_3' = 0$ with positive probability (if $p_0 > 0$). The consequence is that $S_2' = -1 < 0$. If we do not choose to stop the random walk S_n' before it hits negative values, it will be impossible for us to state anything similar to $\{S_i\} \overset{d}{=} \{S_i'\}$ because

$$\forall n, S_n \geq 0 \ a.s. \tag{179}$$

$$\exists n, \mathbb{P}(S_n' < 0) > 0 \tag{180}$$

that's why the stopping time T' is necessary for the theorem we want to state.

The same reasoning holds to show that stopping the process S_n at the total progeny T is necessary. Otherwise $\forall n > T, S_n = 0$ a.s. can never have the same finite-dimensional distribution as that of any non-trivial random walk.

Theorem 8. ***(Connection between Branching Process and Random Walk)***

$$\{S_i\}_{i \leq T} \overset{d}{=} \{S_i'\}_{i \leq T'} \tag{181}$$

Proof. Prove by induction. $S_0' = S_0 = 1$ by definition.

If now $(S_0, ..., S_{i-1}) \overset{d}{=} (S_0', ..., S_{i-1}')$, let's first consider the case where $S_{i-1} = 0$. If $S_{i-1} = 0$, this means that node $i-1$ is already the rightest node among all nodes having the same level and all children of the explored nodes have already been explored prior to node $i-1$. This is equivalent to saying $T = i - 1$ is already at the place of truncation and the theorem is true.

The similar reasoning holds if $S_{i-1}' = 0$ which means that $T' = i - 1$ is at the place of truncation.

For the case where $S_{i-1} > 0, S_{i-1}' > 0$, we don't have to care about the truncation of stopping time any longer. Pick the unexplored node i which is counted towards S_{i-1}. The number of the children of node i is denoted X_i and is independent of $(S_0, ..., S_{i-1})$ (by construction of S_n, node i is unexplored yet so its children won't be counted towards any of $S_0, ..., S_{i-1}$), of course X_i follows the offspring distribution.

$$\forall k_0, ..., k_i \in \mathbb{N}, \mathbb{P}(S_0 = k_0, ..., S_i = k_i) = \mathbb{P}(S_0 = k_0, ..., S_{i-1} = k_{i-1}, X_i = k_i - k_{i-1} + 1) \tag{182}$$

$$= \mathbb{P}(S_0 = k_0, ..., S_{i-1} = k_{i-1}) \cdot \mathbb{P}(X_i = k_i - k_{i-1} + 1) \tag{183}$$

$$= \mathbb{P}\left(S'_0 = k_0, ..., S'_{i-1} = k_{i-1}\right) \cdot \mathbb{P}(X'_i = k_i - k_{i-1} + 1) \tag{184}$$

$$= \mathbb{P}\left(S'_0 = k_0, ..., S'_{i-1} = k_{i-1}, X'_i = k_i - k_{i-1} + 1\right) \tag{185}$$

$$= \mathbb{P}(S'_0 = k_0, ..., S'_i = k_i) \tag{186}$$

so $(S_0, ..., S_i) \overset{d}{=} (S'_0, ..., S'_i)$ by the *i.i.d.* increment structure of random walk S'_n, and the theorem is proved.

□

Remark. *Note that since the total progeny T satisfies the following property that*

$$T \overset{a.s.}{=} \inf\{n : S_n = 0\} \tag{187}$$

as stated in the proof above, and due to the theorem above, one can immediately conclude that

$$T \overset{d}{=} T' \tag{188}$$

in other words, ***the total progeny in branching process has the same distribution as the first hitting time to 0 of the random walk S'_n.***

We have to realize that the value of such connection lies in the fact that existent results in random walk may help us get deeper results in the branching process. For example, results for the first hitting time T' of random walk may be applied to discover new results for the total progeny T in the branching process.

Branching Process Conditional on Extinction

In the following context, we will be discovering only distributional results, so equality in distribution suffices. As a result, we will not distinguish between $X_n \& X'_n$, $S_n \& S'_n$ and $T \& T'$ on the event that the time is less than the number of total progeny. Such operations are justified by the last theorem we have proved above.

Denote

$$H = (X_1, ..., X_T) \tag{189}$$

as the **history of the process** X_n up to time T where H can be of infinite length if $T = \infty$. When we fix the value of the total progeny as $T = t$, $(x_1, ..., x_t) \in \mathbb{R}_+^t$ is a possible value taken by H if and only if

$$s_i = x_1 + ... + x_i - (i-1) \; s.t. \begin{cases} \forall i < t, s_i > 0 \\ s_t = 0 \end{cases} \tag{190}$$

the exploration among the nodes in the tree stops if and only if the time reaches the realization of the total progeny. As a result,

$$\mathbb{P}\left(H=(x_1,...,x_t)\right)=\mathbb{P}\left(X_1=x_1,...,X_t=x_t\right)=\prod_{i=1}^{t}p_{x_i} \tag{191}$$

since $X_1,...,X_t$ *i.i.d.* follows the offspring distribution. This is saying that **given total progeny, the probability of the appearance of all possible X_n paths is known**.

Notice that the extinction in branching process is equivalent to the finiteness of total progeny. So such argument will be very useful to look into the branching process conditional on extinction, where the history can only take values as finite-dimensional vectors.

Theorem 9. ***(Duality Principle for Branching Process)*** *For offspring distribution $\{p_k\}_{k\geq 0}$, define $\{p'_k\}_{k\geq 0}=\eta^{k-1}p_k$ to be its **conjugate offspring distribution**. The branching process with offspring distribution $\{p_k\}_{k\geq 0}$ conditional on extinction has the same distribution as the branching process with the conjugate offspring distribution $\{p'_k\}_{k\geq 0}$.*

Proof. Consider the history H of the process X_n related to the branching process Z_n with offspring distribution $\{p_k\}_{k\geq 0}$. Conditional on extinction, $T<\infty$ *a.s.* so all possible history values are finite-dimensional vectors. Fix the total progeny $T=t<\infty$ for $\forall t\in\mathbb{N}$ and denote E as the event of extinction, apply Bayes theorem

$$\begin{aligned}\mathbb{P}\left(H=(x_1,...,x_t)|E\right)&=\frac{\mathbb{P}\left(E|H=(x_1,...,x_t)\right)\cdot\mathbb{P}\left(H=(x_1,...,x_t)\right)}{\mathbb{P}\left(E\right)} &(192)\\ &=\frac{\mathbb{P}\left(H=(x_1,...,x_t)\right)}{\eta} &(193)\\ &=\frac{\prod_{i=1}^{t}p_{x_i}}{\eta} &(194)\end{aligned}$$

Compute the same probability for another branching process Z'_n with offspring distribution $\{p'_k\}_{k\geq 0}$,

$$\begin{aligned}\mathbb{P}\left(H'=(x_1,...,x_t)\right)&=\prod_{i=1}^{t}p'_{x_i}=\prod_{i=1}^{t}\eta^{x_i-1}p_{x_i} &(195)\\ &=\frac{\eta^{\sum_{i=1}^{t}x_i-(t-1)}\prod_{i=1}^{t}p_{x_i}}{\eta} &(196)\\ &=\frac{\eta^{s_t}\prod_{i=1}^{t}p_{x_i}}{\eta} &(197)\\ &=\frac{\prod_{i=1}^{t}p_{x_i}}{\eta} &(198)\end{aligned}$$

since $S_t=0$ when total progeny is $T=t$, so when $S_t=s_t$ is compatible with $T=t$, $\eta^{s_t}=1$ must hold. So we have

proved that

$$\mathbb{P}\left(H=(x_1,...,x_t)|E\right)=\mathbb{P}\left(H'=(x_1,...,x_t)\right) \tag{199}$$

and the theorem is proved.

□

Remark. *By using the connection with branching process and random walk, we have proved the fact that* ***the branching process conditional on extinction is in distribution still a branching process****, which is otherwise very hard to prove.*

Note that

$$\sum_{k=0}^{\infty}p'_k=\sum_{k=0}^{\infty}\eta^{k-1}p_k=\frac{\sum_{k=0}^{\infty}\eta^k p_k}{\eta}=\frac{G(\eta)}{\eta}=\frac{\eta}{\eta}=1 \tag{200}$$

so the conjugate offspring distribution is exactly a probability distribution.

Theorem 10. ***(Extinction with Large Total Progeny for Supercritical Phase)*** *If the mean of offspring distribution $\mu>1$, then*

$$\mathbb{P}\left(k\le T<\infty\right)\le\frac{e^{-Ik}}{1-e^{-I}} \tag{201}$$

where rate I is given by

$$I=\sup_{t\le 0}\left\{t-\log\mathbb{E}[e^{tX}]\right\}>0 \tag{202}$$

notice that the condition that $\forall t\in\mathbb{R},\mathbb{E}[e^{tX}]<\infty$ is ***unnecessary*** *here, so this result holds for a wide range of offspring distribution.*

Proof. Fix the total progeny $T=t<\infty$ for $\forall t\in\mathbb{N}$, so $S_t=0$ and according to the definition of such random walk S_t, we know that $X_1+...+X_t=t-1\le t$

$$\mathbb{P}\left(k\le T<\infty\right)\le\sum_{t=k}^{\infty}\mathbb{P}\left(T=k\right)=\sum_{t=k}^{\infty}\mathbb{P}\left(S_t=0\right) \tag{203}$$

$$\le\sum_{t=k}^{\infty}\mathbb{P}\left(X_1+...+X_t\le t\right) \tag{204}$$

according to the Chernoff bound (concentration inequality), since $\mathbb{E}X_1=\mu>1$ we get

$$\mathbb{P}\left(X_1+...+X_t\le t\right)\le e^{-tI} \tag{205}$$

to conclude that

$$\mathbb{P}\left(k \leq T < \infty\right) \leq \sum_{t=k}^{\infty} e^{-tI} = \frac{e^{-kI}}{1 - e^{-I}} \tag{206}$$

□

The Chernoff bound we are using is stated as:

Theorem 11. ***(Chernoff Concentration Bound)*** *X_i are i.i.d. random variables, then $\forall a \geq \mathbb{E}X_1$, exists a rate $I(a)$ such that*

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i \geq na\right) \leq e^{-nI(a)} \tag{207}$$

where the rate is defined as

$$I(a) = \sup_{t \geq 0} \left\{ta - \log \mathbb{E}[e^{tX_1}]\right\} \tag{208}$$

On the other hand, for $\forall a \leq \mathbb{E}X_1$, exists a rate $I(a)$ such that

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i \leq na\right) \leq e^{-nI(a)} \tag{209}$$

where the rate is defined as

$$I(a) = \sup_{t \leq 0} \left\{ta - \log \mathbb{E}[e^{tX_1}]\right\} \tag{210}$$

Proof. Let's only prove for the case where $\forall a \geq \mathbb{E}X_1$. Apply Markov inequality for $\forall t \geq 0$ to get

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i \geq na\right) = \mathbb{P}\left(e^{t\sum_{i=1}^{n} X_i} \geq e^{tna}\right) \leq e^{-tna} \cdot \mathbb{E}[e^{t\sum_{i=1}^{n} X_i}] \tag{211}$$

here the trick is to leave this constant t to be unspecified now and optimize it later to get the tightest bound. The upper bound, due to *i.i.d.* property, becomes

$$e^{-tna} \cdot \left(\mathbb{E}[e^{tX_1}]\right)^n \tag{212}$$

now minimize this bound $e^{-tna} \cdot \left(\mathbb{E}[e^{tX_1}]\right)^n$ w.r.t. $t \geq 0$ to get the tightest bound

$$e^{-n\sup_{t\geq 0}\left\{ta - \log \mathbb{E}[e^{tX_1}]\right\}} = e^{-nI(a)} \tag{213}$$

the other case can be proved by setting X_i as $-X_i$.

□

Remark. *Consider the behavior of* $f(t) = ta - \log \mathbb{E}[e^{tX_1}]$ *in the case where* $a \geq \mathbb{E}X_1$*. One would see*

$$f'(t) = a - \frac{\mathbb{E}[X_1 e^{tX_1}]}{\mathbb{E}[e^{tX_1}]} \tag{214}$$

$$f'(0) = a - \mathbb{E}X_1 \geq 0 \tag{215}$$

so the function is increasing at $t = 0$*, also notice that*

$$f''(t) = -\frac{\mathbb{E}[X_1^2 e^{tX_1}]\mathbb{E}[e^{tX_1}] - \mathbb{E}[X_1 e^{tX_1}]\mathbb{E}[X_1 e^{tX_1}]}{\left(\mathbb{E}[e^{tX_1}]\right)^2} \leq 0 \tag{216}$$

by Cauchy-Schwarz applied for the numerator. This is telling us that $f(t)$ *is concave and the supreme of* $ta - \log \mathbb{E}[e^{tX_1}]$ *cannot be achieved by negative* t *except* $a = \mathbb{E}X_1$*. When* $a > \mathbb{E}X_1$ *is strictly larger, the tail probability of i.i.d. sum is always exponential decaying and we can write*

$$I(a) = \sup_t \left\{ ta - \log \mathbb{E}[e^{tX_1}] \right\} \tag{217}$$

where the supreme is taken for $\forall t \in \mathbb{R}$*.*

Now return to the extinction probability with large total progeny. We have some remarks for the theorem that

Remark. *Now* $\mu > 1$ *is strictly larger, so the supreme in* I *can be taken for* $\forall t \in \mathbb{R}$ *instead of* $t \leq 0$*.*

Also notice that ***the supreme of*** I ***is taken at strictly negative*** $t < 0$ *because the derivative* $f'(0) = 1 - \mu < 0$*. As a result, we conclude that*

$$I > 0 \tag{218}$$

I ***is strictly positive*** *since* $f(0) = 0, f'(0) = 1 - \mu < 0$ *and* ***the condition that*** $\forall t, \mathbb{E}[e^{tX}] < \infty$ ***is also unnecessary****.*

Remark. *The extinction with large total progeny seems to be difficult to estimate, but by* ***exploiting the implicit*** *i.i.d.* ***increment structure in the branching process****, one can greatly simplify such problem by using concentration inequalities.*

An interesting implication of this theorem is that by setting $k \to \infty$*, one would see that the larger the total progeny is, the less likely the extinction happens, and* ***the extinction probability decays exponentially fast w.r.t. the total progeny size****.*

Law of Total Progeny

To get the law of total progeny, notice that in the random walk interpretation, we have proved that total progeny has the same distribution as the first hitting time to 0 of the random walk S_n. As a result, the problems turns into

finding the distribution of the first hitting time of a random walk. The crucial hitting time theorem stated below (which is a more general one) may help.

Theorem 12. ***(Hitting Time Theorem)*** *For random walk S_n starting from $S_0 = k \in \mathbb{N}$ with i.i.d. increments Y_i such that Y_i only takes integer values and $Y_i \geq -1$ a.s., denote the first hitting time to 0 of S_n as H_0, then*

$$\forall n \geq 1, \mathbb{P}(H_0 = n) = \frac{k}{n}\mathbb{P}(S_n = 0) \tag{219}$$

Proof. Prove by induction on n. When $n = 1$, if $k > 1$ or $k = 0$ then both sides are 0 since the increment cannot be smaller than -1; if $k = 1$, then $\mathbb{P}(H_0 = 1) = \mathbb{P}(Y_1 = -k) = \mathbb{P}(S_1 = 0)$ so when $n = 1$ the conclusion holds.

Now assume the conclusion holds for $n - 1$, let's prove it for n. When $k = 0$, the case if trivial so we assume that $k \geq 1$.

$$\mathbb{P}_k(H_0 = n) = \sum_{p=-1}^{\infty} \mathbb{P}_k(H_0 = n | Y_1 = p) \cdot \mathbb{P}(Y_1 = p) \tag{220}$$

$$= \sum_{p=-1}^{\infty} \mathbb{P}_{k+p}(H_0 = n-1) \cdot \mathbb{P}(Y_1 = p) \tag{221}$$

by the Markov property where $\mathbb{P}_k$ denotes the probability measure if the random walk starts from k. By induction hypothesis, since $k + p \geq 0$,

$$\mathbb{P}_k(H_0 = n) = \sum_{p=-1}^{\infty} \frac{k+p}{n-1}\mathbb{P}_{k+p}(S_{n-1} = 0) \cdot \mathbb{P}(Y_1 = p) \tag{222}$$

by Markov property once more, $\mathbb{P}_{k+p}(S_{n-1} = 0) = \mathbb{P}_k(S_n = 0 | Y_1 = p)$, so the law of total probability tells us

$$\mathbb{P}_k(H_0 = n) = \sum_{p=-1}^{\infty} \frac{k+p}{n-1}\mathbb{P}_k(S_n = 0 | Y_1 = p) \cdot \mathbb{P}(Y_1 = p) \tag{223}$$

$$= \frac{1}{n-1}\left[k \sum_{p=-1}^{\infty} \mathbb{P}_k(S_n = 0 | Y_1 = p) \cdot \mathbb{P}(Y_1 = p) + \sum_{p=-1}^{\infty} p \cdot \mathbb{P}_k(S_n = 0 | Y_1 = p) \cdot \mathbb{P}(Y_1 = p)\right] \tag{224}$$

$$= \frac{1}{n-1}\left[k \cdot \mathbb{P}_k(S_n = 0) + \mathbb{E}_k(Y_1 \cdot \mathbb{I}_{S_n = 0})\right] \tag{225}$$

$$= \frac{1}{n-1}\mathbb{P}_k(S_n = 0)\left[k + \mathbb{E}_k(Y_1 | S_n = 0)\right] \tag{226}$$

now we only have to prove that $\mathbb{E}_k(Y_1 | S_n = 0) = -\frac{k}{n}$ to complete the whole proof.

Due to the fact that $Y_1, ..., Y_n$ *i.i.d.*, it's easy to see that $\forall 1 \leq i < j \leq n, \mathbb{E}_k(Y_i | S_n = 0) = \mathbb{E}_k(Y_j | S_n = 0)$. This leads to the fact that

$$n \cdot \mathbb{E}_k(Y_1 | S_n = 0) = \mathbb{E}_k\left(\sum_{i=1}^{n} Y_i \Big| S_n = 0\right) = \mathbb{E}_k\left(S_n - S_0 \Big| S_n = 0\right) = -k \tag{227}$$

this calculation ends the proof.

□

Remark. *This hitting time theorem is very interesting since it only requires the increments of the random walk to be any discrete probability distribution taking values no less than -1, which includes a large variety of random walk models. This is a generalization of the hitting time theorem for SRW.*

Apply the connection between branching process and random walk to find

Theorem 13. ***(Law of Total Progeny)*** *$\forall k \geq 1, k \in \mathbb{N}$, $T_1, ..., T_k$ are i.i.d. random variables following the same distribution as the total progeny T, then*

$$\forall n \geq 1, \mathbb{P}(T_1 + ... + T_k = n) = \frac{k}{n}\mathbb{P}(X_1 + ... + X_n = n - k) \tag{228}$$

where X_i is the number of children of node i in the branching process as defined above.

Proof. We know that the random walk related to the branching process has the following definition

$$S_n = 1 + (X_1 - 1) + ... + (X_n - 1) \tag{229}$$

where $X_1, ..., X_n$ are *i.i.d.* and follow the offspring distribution.

Now $T_1, ..., T_k$ are *i.i.d.* random variables following the same distribution as the total progeny T, but notice that we have proved the fact that the total progeny T has the same distribution as the first hitting time to 0 of S_n. Denote $H_0^1, ..., H_0^k$ as the *i.i.d.* copies of the first hitting time to 0 of S_n. Then

$$T_1 \overset{d}{=} H_0^1 \tag{230}$$

$$T_1 + ... + T_k \overset{d}{=} H_0^1 + ... + H_0^k \tag{231}$$

however, this is still not easy to deal with! Is there a way for us to turn $H_0^1 + ... + H_0^k$ into a single first hitting time to 0 of a new random walk? Actually we can do that since now S_n starts from 1. If we shift the path of S_n downward by 1 unit to get a new random walk R_n such that $R_0 = 0$, then H_0^i would be equal distributionally to the first hitting time to -1 of R_n. As a result, denote τ_{-k} as the first hitting time to $-k$ of R_n, then by Markov property,

$$H_0^1 + ... + H_0^k \overset{d}{=} \tau_{-k} \tag{232}$$

notice at last that by shifting the path of R_n upward by k unit to get a new random walk W_n starting from $W_0 = k$, the distribution of τ_{-k} is equal to the distribution of the first hitting time to 0 of W_n, denoted τ_0^W. The relationship of S_n and W_n is easy to see since they are only different by translations

$$W_n = S_n - 1 + k = X_1 + ... + X_n - n + k \tag{233}$$

Apply the hitting time theorem for τ_0^W to know

$$\mathbb{P}\left(\tau_0^W = n\right) = \frac{k}{n}\mathbb{P}\left(W_n = 0\right) = \frac{k}{n}\mathbb{P}\left(X_1 + ... + X_n = n - k\right) \tag{234}$$

we conclude that

$$\mathbb{P}\left(T_1 + ... + T_k = n\right) = \mathbb{P}\left(\tau_0^W = n\right) = \frac{k}{n}\mathbb{P}\left(X_1 + ... + X_n = n - k\right) \tag{235}$$

which ends the proof. □

Remark. *$T_1 + ... + T_k$ can be interpreted as **the total progeny of the branching process with** $Z_0 = k$, i.e. k people at generation 0 generating k independent branching trees. By setting $k = 1$ in the theorem, one can conclude that*

$$\mathbb{P}\left(T = n\right) = \frac{1}{n}\mathbb{P}\left(X_1 + ... + X_n = n - 1\right) \tag{236}$$

*is **the law of total progeny**. Since $X_1, ..., X_n$ i.i.d. follow the offspring distribution, the distribution of $X_1 + ... + X_n$ can be figured out easily with the help of characteristic functions.*

Poisson Branching Process

Now we consider the case where the offspring distribution is $P(\lambda)$ the Poisson distribution. We mainly apply all results we have derived above to see why Poisson branching process is special.

The generating function of the offspring distribution is

$$G(s) = \sum_{k=0}^{\infty} s^k \frac{\lambda^k}{k!} e^{-\lambda} = e^{s\lambda - \lambda} \tag{237}$$

extinction probability in the supercritical phase ($\lambda > 1$) is now given by

$$\eta = e^{\eta\lambda - \lambda} \tag{238}$$

impossible to solve out the closed form.

The duality principle now tells us that

$$p_k' = \eta^{k-1} p_k = \frac{1}{\eta}\frac{(\eta\lambda)^k}{k!} e^{-\lambda} \tag{239}$$

$$= e^{\lambda - \eta\lambda}\frac{(\eta\lambda)^k}{k!} e^{-\lambda} \tag{240}$$

$$= \frac{(\eta\lambda)^k}{k!} e^{-\eta\lambda} \tag{241}$$

so **the Poisson branching process conditional on extinction is still a Poisson branching process with offspring distribution** $P(\eta\lambda)$ which is really intuitive!

The **distribution of total progeny**, as we have just stated, is

$$\mathbb{P}(T=n)=\frac{1}{n}\mathbb{P}(X_1+...+X_n=n-1) \tag{242}$$

$$\tag{243}$$

since $X_1+...+X_n\sim P(n\lambda)$. To get the exact order of growth of the total progeny as $n\to\infty$,

Theorem 14. ***(Asymptotic Growth of Total Progeny in Poisson Branching Process)*** *For Poisson branching process with offspring distribution $P(\lambda)$,*

$$\mathbb{P}(T=n)=\frac{1}{\lambda\sqrt{2\pi n^3}}e^{-n(\lambda-1-\log\lambda)}\left[1+O\left(\frac{1}{n}\right)\right] \tag{244}$$

In particular, when it's the critical phase, i.e. $\lambda=1$,

$$\mathbb{P}(T=n)=\frac{1}{\sqrt{2\pi n^3}}\left[1+O\left(\frac{1}{n}\right)\right] \tag{245}$$

Proof. The proof is simple calculations applying the Stirling formula for the factorial, let's first deal with the case where $\lambda=1$

$$\frac{1}{n}\frac{n^{n-1}}{(n-1)!}e^{-n}=\frac{n^{n-1}}{\sqrt{2\pi n}\left(\frac{n}{e}\right)^n[1+O(\frac{1}{n})]}e^{-n} \tag{246}$$

$$=\frac{1}{\sqrt{2\pi n^3}[1+O(\frac{1}{n})]} \tag{247}$$

it's proved.

Notice that in the general case

$$\mathbb{P}(T=n)=\frac{1}{n}\frac{(n\lambda)^{n-1}}{(n-1)!}e^{-n\lambda} \tag{248}$$

$$=\lambda^{n-1}e^{-n\lambda+n}\cdot\frac{1}{n}\frac{n^{n-1}}{(n-1)!}e^{-n} \tag{249}$$

$$=\frac{1}{\lambda}e^{-n\lambda+n+n\log\lambda}\cdot\frac{1}{n}\frac{n^{n-1}}{(n-1)!}e^{-n} \tag{250}$$

use the conclusion when $\lambda=1$, it's also proved.

□

Remark. *Note that $f(\lambda) = \lambda - 1 - \log\lambda$ is on the exponential with*

$$f(1) = 0 \tag{251}$$

$$f'(\lambda) = 1 - \frac{1}{\lambda} \tag{252}$$

*so **when** $\lambda \neq 1$, $f(\lambda) > 0$ **and** $\mathbb{P}(T = n)$ **is always exponentially decaying in the size of progeny**.*

The interesting thing lies in the critical phase when $\lambda = 1$, where

$$\mathbb{P}(T = n) \sim Cn^{-\frac{3}{2}} \ (n \to \infty) \tag{253}$$

this is telling us that such T is in the domain of attraction of non-negative α-stable law with $\alpha = \frac{3}{2}$.

In other words, for a Poisson branching process with $Z_0 = k, \lambda = 1$, denote the total progeny in the tree of each ancestor as $T_1, ..., T_k$, then they are i.i.d. so we conclude that

$$\exists a_k, b_k \in \mathbb{R}, s.t. \ \frac{T_1 + ... + T_k - a_k}{b_k} \to Y \ (k \to \infty) \tag{254}$$

*where Y has the one-sided α-stable law with $\alpha = \frac{3}{2}$. **For Poisson branching process with k ancestors in the critical phase, the total progeny follows a normalized one-sided $\frac{3}{2}$-stable law when the number of ancestors is large enough.***

As one would expect to see for a single random variable, the Poisson distribution is the limit distribution of the binomial distribution. In detail, a $B(n, p)$ random variable with $np = \lambda$ would have limit distribution $P(\lambda)$ if $n \to \infty$. Naturally, we want to generalize such kind of conclusions in the context of branching process, i.e. Poisson branching process should be able to be approximated by binomial branching process.

In order to achieve this objective, we have to introduce the coupling technique for binomial and Poisson distribution. For any two random variables X, Y we call the random vector $(\hat{X}, \hat{Y})$ as a **coupling** of X, Y if the marginal of $\hat{X}$ has the same distribution as X and the marginal of $\hat{Y}$ has the same distribution as Y. There are many kinds of coupling techniques, for example, independent copy coupling used in the proof of the convergence of Markov Chain to its stationary distribution, maximum coupling maximizing the total variation $TV(X, Y)$ etc. Different coupling techniques have their respective usages. Here we introduce the coupling technique to deal with binomial and Poisson random variables. This result is the direct consequence of **the Poisson limit theorem**.

Theorem 15. ***(Coupling for Poisson Limit Theorem)*** *Fix $n \in \mathbb{N}$, let $I_1, ..., I_n$ be independent random variables and $I_i \sim B(1, p_i)$, set $\lambda = \sum_{i=1}^n p_i$ and $X = I_1 + ... + I_n, Y \sim P(\lambda)$. Then there exists coupling $(\hat{X}, \hat{Y})$ of X, Y such that*

$$\mathbb{P}\left(\hat{X} \neq \hat{Y}\right) \leq \sum_{i=1}^n p_i^2 \tag{255}$$

Proof. We have to make use of the additivity of Poisson random variables. Set $J_i \sim P(p_i)$ and $J_1, ..., J_n$ be indepen-

dent, denote

$$p_{i,x} = \mathbb{P}(I_i = x) = p_i \cdot \mathbb{I}_{x=1} + (1-p_i) \cdot \mathbb{I}_{x=0} \tag{256}$$

$$q_{i,x} = \mathbb{P}(J_i = x) = \frac{p_i^x}{x!} e^{-p_i} \tag{257}$$

Intuitively, we would expect to first construct the connection between I_i, J_i and then construct the coupling for X, Y since both have the structure as independent sums. That's why we take $(\hat{I}_i, \hat{J}_i)$ as the **maximal coupling** of I_i, J_i for each fixed i. By doing this and noticing the definition of total variation that $TV(I_i, J_i) = \inf_{(\hat{I}_i, \hat{J}_i)} \mathbb{P}\left(\hat{I}_i \neq \hat{J}_i\right)$ and such inf can always be achieved, we have

$$\mathbb{P}\left(\hat{I}_i \neq \hat{J}_i\right) = TV(I_i, J_i) \tag{258}$$

$$= \frac{1}{2} \sum_{k=0}^{\infty} |p_{i,k} - q_{i,k}| \tag{259}$$

$$= \frac{1}{2} \left(|p_{i,0} - q_{i,0}| + |p_{i,1} - q_{i,1}| + \sum_{k=2}^{\infty} q_{i,k} \right) \tag{260}$$

$$= p_i - p_i e^{-p_i} \tag{261}$$

$$= p_i(1 - e^{-p_i}) \leq p_i^2 \tag{262}$$

Now we construct $\hat{X}, \hat{Y}$ based on those couplings in the following way:

$$\hat{X} = \sum_{i=1}^{n} \hat{I}_i \tag{263}$$

$$\hat{Y} = \sum_{i=1}^{n} \hat{J}_i \tag{264}$$

immediately conclude that since $\hat{I}_1, ..., \hat{I}_n$ are independent and follows $\hat{I}_i \sim B(1, p_i)$, $\hat{X} \overset{d}{=} X$. By the additivity of independent Poisson random variable, $\hat{Y} \overset{d}{=} Y$ gives a coupling of X, Y. (We just need to ensure that the random vectors $(\hat{I}_1, \hat{J}_1), ..., (\hat{I}_n, \hat{J}_n)$ are independent, which is an easy to satisfy)

$$\mathbb{P}\left(\hat{X} \neq \hat{Y}\right) \leq \sum_{i=1}^{n} \mathbb{P}\left(\hat{I}_i \neq \hat{J}_i\right) \leq \sum_{i=1}^{n} p_i^2 \tag{265}$$

completes the proof.

□

One can directly prove the following Poisson limit theorem based on this coupling argument.

Theorem 16. ***(Poisson Limit Theorem)*** *For triangular random variable array $X_{i,j}$ where the random variables*

in each row are independent and each entry $X_{i,j}$ can only take non-negative integer values with

$$\mathbb{P}(X_{i,j}=1)=p_{i,j},\mathbb{P}(X_{i,j}\geq 2)=\varepsilon_{i,j} \tag{266}$$

now if the following conditions hold

$$\begin{cases}\sum_{m=1}^{n} p_{n,m}\to\lambda\ (n\to\infty)\\ \max_{1\leq m\leq n} p_{n,m}\to 0\ (n\to\infty)\\ \sum_{m=1}^{n}\varepsilon_{n,m}\to 0\ (n\to\infty)\end{cases} \tag{267}$$

then for random variable $S_n=X_{n,1}+...+X_{n,n}$, $S_n\xrightarrow{d}P(\lambda)\ (n\to\infty)$.

Proof. Just notice the fact that for discrete random variables X_n,Y, total variation $TV(X_n,Y)\to 0\ (n\to\infty)$ implies convergence in distribution $X_n\overset{d}{=}Y\ (n\to\infty)$.

Now first we want to get rid of the case where $X_{i,j}\geq 2$. Denote $X'_{i,j}\sim B(1,p_{i,j})$ and $S'_n=X'_{n,1}+...+X'_{n,n}$, then

$$\mathbb{P}(S_n\neq S'_n)\leq\sum_{m=1}^{n}\varepsilon_{n,m}\to 0\ (n\to\infty) \tag{268}$$

since $S_n,S'_n\in\mathbb{Z}$, this is telling us that proving $S'_n\xrightarrow{d}P(\lambda)\ (n\to\infty)$ suffices.

Use the theorem above for coupling to construct $(\hat{X}_n,\hat{Y}_n)$ such that $\hat{X}_n=X'_{n,1}+...+X'_{n,n}=S'_n$ and $\hat{Y}_n\sim P(\sum_{m=1}^{n}p_{n,m})$ to find that when n is large enough, denote $Y\sim P(\lambda)$, then

$$TV(\hat{Y}_n,Y)\to 0\ (n\to\infty) \tag{269}$$

so

$$TV(S_n,Y)\leq TV(S_n,\hat{Y}_n)+TV(\hat{Y}_n,Y) \tag{270}$$

$$\leq\sum_{m=1}^{n}p_{n,m}^2+TV(\hat{Y}_n,Y)\to 0\ (n\to\infty) \tag{271}$$

to see the reason, notice that

$$\sum_{m=1}^{n}p_{n,m}^2\leq\max_{1\leq m\leq n}p_{n,m}\cdot\sum_{m=1}^{n}p_{n,m}\to 0\ (n\to\infty) \tag{272}$$

concludes the whole proof.

□

Remark. *This Poisson limit theorem can also be proved very easily by considering the convergence of characteristic*

function. However, this coupling proof not only proves the result, but also shows us the convergence rate in the sense of total variation.

After the preparation, we will use coupling to prove the approximation of Poisson branching process.

Theorem 17. ***(Poisson Branching Process Approximated by Binomial Branching Process)***

For binomial branching process with offspring distribution $B(n,p)$ and the Poisson branching process with offspring distribution $P(\lambda)$. Assume $\lambda = np$, then

$$\forall k \geq 1, \mathbb{P}_{n,p}\left(T \geq k\right) = \mathbb{P}_{\lambda}\left(T^* \geq k\right) + e_n(k) \tag{273}$$

where

$$|e_n(k)| \leq \frac{\lambda^2}{n} \sum_{s=1}^{k-1} \mathbb{P}_{\lambda}\left(T^* \geq s\right) \leq \frac{k\lambda^2}{n} \tag{274}$$

Here $\mathbb{P}_{n,p}$ denotes the distribution of binomial branching process with parameter n, p, $\mathbb{P}_{\lambda}$ denotes the distribution of Poisson branching process with parameter λ and T is the total progeny in binomial branching process while T^* is the total progeny in Poisson branching process. (All processes related to Poisson branching process will contain $*$)

Proof. Use the coupling argument to find the coupling for X_i, X_i^* where $X_i \sim B(n, \frac{\lambda}{n}), X_i^* \sim P(\lambda)$. Note that X_i is n independent sums of Bernoulli r.v. and $\lambda = n \cdot \frac{\lambda}{n}, X_i^* \sim P(\lambda)$ satisfies the condition for the coupling argument. So there always exists coupling (Y_i, Y_i^*) of X_i, X_i^* such that

$$\forall i, \mathbb{P}\left(Y_i \neq Y_i^*\right) \leq \sum_{i=1}^{n} \frac{\lambda^2}{n^2} = \frac{\lambda^2}{n} \tag{275}$$

recall that X_i stands for the number of children of node i in binomial branching tree and X_i^* stands for the number of children of node i in Poisson branching tree. We would be able to get $(Y_1, Y_1^*), ..., (Y_n, Y_n^*), ...$ infinitely many coupling pairs, and we will make sure that those pairs are independent when constructing them.

The main thought of the remaining proof is to bound

$$|\mathbb{P}_{n,p}\left(T \geq k\right) - \mathbb{P}_{\lambda}\left(T^* \geq k\right)| \leq \max\left\{\mathbb{P}\left(T \geq k, T^* < k\right), \mathbb{P}\left(T^* \geq k, T < k\right)\right\} \tag{276}$$

and prove that both probabilities on the RHS are bounded above by $\frac{\lambda^2}{n} \sum_{s=1}^{k-1} \mathbb{P}_{\lambda}\left(T^* \geq s\right)$.

Let's only look into $\mathbb{P}\left(T \geq k, T^* < k\right)$, the split of such probability depends on the coupling.

$$\mathbb{P}\left(T \geq k, T^* < k\right) \leq \sum_{s=1}^{k-1} \mathbb{P}\left(\forall i \leq s-1, Y_i = Y_i^*, Y_s \neq Y_s^*, T \geq k\right) \tag{277}$$

split the event $T^* < k$ according to the first time $Y_s \neq Y_s^*$. Such s must exist since by the connection of branching process with random walk, $\exists t < k, Y_1^* + ... + Y_t^* = t - 1$ but $\nexists t < k, Y_1 + ... + Y_t = t - 1$.

Then

$$\mathbb{P}\left(T \geq k, T^* < k\right) \leq \sum_{s=1}^{k-1} \mathbb{P}\left(\forall i \leq s-1, Y_i = Y_i^*, Y_s \neq Y_s^*, T \geq k\right) \tag{278}$$

$$\leq \sum_{s=1}^{k-1} \mathbb{P}\left(T^* \geq s, Y_s \neq Y_s^*\right) \tag{279}$$

$$= \sum_{s=1}^{k-1} \mathbb{P}\left(T^* \geq s\right) \mathbb{P}\left(Y_s \neq Y_s^*\right) \tag{280}$$

$$\leq \frac{\lambda^2}{n} \sum_{s=1}^{k-1} \mathbb{P}\left(T^* \geq s\right) \tag{281}$$

since $T \geq k$ implies $T^* \geq s$ and $T^* \geq s$ is independent of $Y_s \neq Y_s^*$ since whether total progeny is larger than s only depends on $Y_1^*, ..., Y_{s-1}^*$. So the theorem is proved.

□

Remark. *This theorem is telling us that* ***when the condition*** $np = \lambda$ ***holds, the tail probability of the two total progeny are close to each other and the gap is controlled by*** $\frac{k\lambda^2}{n}$ ***which shrinks to 0 when*** $n \to \infty$. *So one can approximate Poisson branching process well enough by an appropriate binomial branching process.*

Markov Chain

The Markov chain X_n in the discrete time is defined as

$$\mathbb{P}\left(X_{n+1}|\mathscr{F}_n\right) = \mathbb{P}\left(X_{n+1}|X_n\right) \tag{282}$$

which means that the distribution of X_{n+1} knowing X_n only depends on X_n but does not depend on any past history before X_n. To write it specifically,

$$\forall i_j \in \mathbb{R}, , \mathbb{P}\left(X_{n+1} = i|X_n = i, X_{n-1} = i_{n-1}, ..., X_0 = i_0\right) = \mathbb{P}\left(X_{n+1} = i|X_n = i\right) \tag{283}$$

The discussion of Markov chain starts from the rigorous statement of the Markov property, which has been used several times in the previous context without a rigorous definition. Define S as the set in which X_n take values, it's called the **state set**. We can describe $X_n = i$ by saying that the Markov chain is in state i at time n. Here, we only consider the **time homogeneous** Markov Chain in **discrete time**, where at different times the transition probability among states is the same. The transition probability between different states is defined as

$$p_{ij} = \mathbb{P}\left(X_{n+1} = j|X_n = i\right) \tag{284}$$

the probability transiting from state i to state j. It's easy to see that if $|S| < \infty$, i.e. the Markov Chain only has finitely many possible states, all transition probabilities can be formed as a matrix denoted P (which fails if there are infinitely many possible states).

It's easy to see that only having transition probability is not enough to fix the distribution of the whole Markov chain, we still need an **initial distribution** $X_0 \sim \mu$ so that the distribution of the whole chain is fixed. However, we would need to consider the probability space where the whole Markov chain is located. To build such a probability space, we just need μ and all transition probabilities, this provides us with a consistent family of finite-dimensional distribution. By Kolmogorov's extension theorem, we can argue that there exists a probability space

$$(\Omega_0, \mathscr{F}_\infty, \mathbb{P}_\mu) \tag{285}$$

$$\begin{cases} \Omega_0 = \{(\omega_0, \omega_1, ...)\} \\ \mathbb{P}_\mu(A) = \int \mathbb{P}_x(A)\, \mu(dx) \end{cases} \tag{286}$$

where Ω_0 is the sample space and each sample point $\omega = (\omega_0, \omega_1, ...)$ is the possible value of a whole Markov chain with $X_n(\omega) = \omega_n$ as the state of the Markov chain at time n. The probability measure on this whole space $\mathbb{P}_\mu$ depends on the initial distribution and the transition probability. The probability measure of the event A for such Markov chain with initial distribution μ is just the probability measure of the event A for such Markov chain assuming it starts from $X = x$ and taking expectation w.r.t. X under measure μ.

By setting up the probability space for the whole Markov chain, one just have to define the **shifting operator**

that shifts the sample trajectory of the Markov chain left by time n.

$$\theta_n(\omega_0, \omega_1, ...) = (\omega_n, \omega_{n+1}, ...) \tag{287}$$

an easy observation is that such θ_n is actually an operator mapping from $\theta_n : \Omega_0 \to \Omega_0$. As a result, for random variable Y which is the r.v. in the probability space of the whole Markov chain,

$$Y : \Omega_0 \to \mathbb{R}, Y \circ \theta_n : \Omega_0 \to \mathbb{R} \tag{288}$$

by taking composition with the shifting operator, we get a new random variable as if the time has been shifted for n units.

Remark. *For the example above, let's assuming that Y has no measurability issues and denotes the indicator such that the Markov chain hits 0 at time n, then*

$$Y(\omega) = \mathbb{I}_{\omega_n = 0}(\omega_0, \omega_1, ...) \tag{289}$$

by applying the shifting operator,

$$Y \circ \theta_n(\omega) = \mathbb{I}_{[\theta_n(\omega)]_n = 0}(\theta_n(\omega)) \tag{290}$$

$$= \mathbb{I}_{\omega_{2n} = 0}(\omega_n, \omega_{n+1}, ...) \tag{291}$$

so $Y \circ \theta_n$ is actually the indicator such that Markov Chain starts at time n from X_n and hits 0 at time $2n$. As a result, the event is translated in time by n units.

Now let's state the Markov property

Theorem 18. ***(Markov Property)*** *$Y : \Omega_0 \to \mathbb{R}$ is bounded measurable, then*

$$\mathbb{E}_\mu(Y \circ \theta_m | \mathscr{F}_m) = \mathbb{E}_{X_m} Y \tag{292}$$

Proof. To prove the property relevant to conditional expectation, firstly it's obvious that $\mathbb{E}_{X_m} Y \in \mathscr{F}_m$. So we just need to prove that

$$\forall A \in \mathscr{F}_m, \mathbb{E}_\mu(Y \circ \theta_m \cdot \mathbb{I}_A) = \mathbb{E}_{X_m}(Y \mathbb{I}_A) \tag{293}$$

First prove it for the simple case of Y and A where

$$Y(\omega) = g_0(\omega_0) \cdot ... \cdot g_n(\omega_n) \tag{294}$$

$$A = \{\omega : \omega_0 \in A_0, ..., \omega_m \in A_m\} \tag{295}$$

where $A_0, ..., A_m \subset \mathbb{R}$ are Borel measurable and $g_0, ..., g_n$ are bounded measurable functions.

From the distribution of the Markov chain depending on the transition probability and the initial distribution,

$$\mathbb{E}_\mu(Y \circ \theta_m \cdot \mathbb{I}_A) = \mathbb{E}_\mu(g_0(\omega_m) \cdot ... \cdot g_n(\omega_{m+n}) \cdot \mathbb{I}_A) \tag{296}$$

$$= \int_{A_0} \mu(dx_0) \int_{A_1} p(x_0, dx_1) ... \int_{A_m} p(x_{m-1}, dx_m) g_0(x_m) \int p(x_m, dx_{m+1}) g_1(x_{m+1}) ... \tag{297}$$

$$\int p(x_{m+n-1}, dx_{m+n}) g_n(x_{m+n}) \tag{298}$$

$$= \mathbb{E}_\mu \left[\mathbb{E}_{X_m}(g_0(X_0) ... g_n(X_n) \cdot \mathbb{I}_A) \right] \tag{299}$$

where p is the Markov transition kernel $p(\omega, A)$ such that for each fixed $\omega \in \Omega_0$ it's a measure on $(\Omega_0, \mathscr{F}_\infty)$ and for each measurable set A it's a measurable function. Actually, $p(X_n, B) = \mathbb{P}(X_{n+1} \in B | X_n)$. Here we are also using the Fubini theorem to tear the integral into two parts, the part $dx_0 ... dx_{m-1}$ and the part $dx_m ... dx_{m+n}$.

Now to prove the property in general case, use $\pi - \lambda$ theorem. Collect all A such that the property above holds to form $\mathscr{H}$. One can find that $\mathscr{F}_m$ is generated by measurable rectangles as the product of m measurable sets which forms a π-system. By proving $\mathscr{H}$ is a λ-system, the property holds for $\forall A \in \mathscr{F}_m$.

Similarly, we have assumed that Y has the form of the product of bounded measurable functions. Collect all Y such that the property above holds (for $\forall A \in \mathscr{F}_m$ as proved) to form a set $\mathscr{G}$. Such $\mathscr{G}$ contains all indicators, is closed under addition and scaling and is closed under increasing limit (bounded limit) of non-negative function approximation. By the monotone class theorem, $\mathscr{G}$ contains all bounded measurable functions and the property is proved. □

Remark. *The Markov property is just telling us that the conditional expectation of a m units time-shifted random variable when the Markov Chain starts from time 0 is just the expectation of the random variable as if we stop the Markov chain at time m and restart it.*

In other words, the Markov property is the ***"memoryless property of a process"****. On knowing X_m, one forgets all that has already happened and can restart the Markov chain whenever one wants.*

The Markov property can be extended to the strong Markov property with simple partition on the value of the stopping time. Now τ is a stopping time so $\forall n \in \mathbb{N}, \{\tau \leq n\} \in \mathscr{F}_n$, now since we are in the discrete time setting, τ must be a discrete r.v. The associated **stopped sigma field** is defined as

$$\mathscr{F}_\tau = \{A : \forall n \in \mathbb{N}, A \cap \{\tau \leq n\} \in \mathscr{F}_n\} \tag{300}$$

contains all events such that one can tell whether the event happens at time n when the stopping time has already been realized. So the occurrence of these events are known as soon as the stopping time is realized. The shift operator is defined in the random way that

$$\theta_\tau = \sum_{n=0}^{\infty} \theta_n \cdot \mathbb{I}_{\tau=n} \tag{301}$$

note that since the stopping time τ may have positive probability of taking ∞, we have to add an extra definition

that when $\{\tau = \infty\}$ happens, the action of θ_τ is to map every sample point ω to an element Δ specified in Ω_0 just for the notation purpose.

Theorem 19. ***(Strong Markov Property)*** *$Y_n : \Omega_0 \to \mathbb{R}$ is measurable and $\exists M, \forall n, |Y_n| \leq M$, then*

$$\mathbb{E}_\mu(Y_\tau \circ \theta_\tau | \mathscr{F}_\tau) = \mathbb{E}_{X_\tau} Y_\tau \tag{302}$$

on the event $\{\tau < \infty\}$.

The proof is just breaking down the stopping time w.r.t. the value it's taking so it's easy to write out.

We focus on the applications of the Markov property and let's first prove the Chapman-Kolmogorov equation. In the following context, use $\mathbb{P}_x$ to denote the distribution of the Markov chain given that it starts from $X_0 = x$.

Theorem 20. ***(Chapman-Kolmogorov equation)*** *For Markov Chain X_n,*

$$\mathbb{P}_x\left(X_{m+n} = z\right) = \sum_y \mathbb{P}_x\left(X_m = y\right) \cdot \mathbb{P}_y\left(X_n = z\right) \tag{303}$$

Proof. It's easy to see that we shall stop the process at time m and restart it again

$$\mathbb{P}_x\left(X_{m+n} = z\right) = \sum_y \mathbb{P}_x\left(X_m = y\right) \cdot \mathbb{P}\left(X_{m+n} = z | X_m = y\right) \tag{304}$$

notice that the conditional probability

$$\mathbb{P}\left(X_{m+n} = z | X_m = y\right) = \mathbb{E}(\mathbb{I}_{X_{m+n}=z} | \mathscr{F}_m)|_{X_m = y} \tag{305}$$

and let's set $Y(\omega) = \mathbb{I}_{\omega_n = z}(\omega)$ with time to be shifted by m units. By Markov property,

$$\mathbb{E}(\mathbb{I}_{X_{m+n}=z} | \mathscr{F}_m) = \mathbb{E}_{X_m} Y = \mathbb{P}_{X_m}\left(X_n = z\right) \tag{306}$$

and thus we conclude that

$$\mathbb{P}\left(X_{m+n} = z | X_m = y\right) = \mathbb{P}_y\left(X_n = z\right) \tag{307}$$

and the theorem is proved. □

Remark. *Assume Markov chain has finitely many states so there exists transition matrix P, then it has non-negative entries with each row sum up to 1. Denote $P^{(n)}$ as the n-step transition matrix with*

$$P_{ij}^{(n)} = \mathbb{P}\left(X_n = j | X_0 = i\right) \tag{308}$$

the Chapman-Kolmogorov equation is just telling us that $P^{(n)} = P^n$ is the power of matrix P.

Similarly, by denoting $\mu^{(n)}$ as the distribution of X_n under the condition that $X_0 \sim \mu$, it's teliing us that

$$\mu^{(n)} = \mu P^n \tag{309}$$

is the action of n-step transition matrix on the initial distribution since

$$\mathbb{P}_\mu(X_n = j) = \sum_i \mathbb{P}_\mu(X_0 = i) \cdot \mathbb{P}(X_n = x | X_0 = i) \tag{310}$$

$$= \sum_i \mathbb{P}_\mu(X_0 = i) \cdot P_{ij}^{(n)} \tag{311}$$

Recurrence and Transience

A state $i \in S$ is called **recurrent** if

$$\mathbb{P}_i(T_i < \infty) = 1 \tag{312}$$

where T_i is the first hitting time of i and is ≥ 1 $T_i = \inf\{n \geq 1 : X_n = i\}$. So for recurrent state, departing from this states one can always go back to this state in finite time. If a state is not recurrent then it's **transient**.

Here we can still take the generating function approach and define

$$P_{ij}(s) = \sum_{n=0}^{\infty} p_{ij}(n) s^n \tag{313}$$

$$F_{ij}(s) = \sum_{n=0}^{\infty} f_{ij}(n) s^n \tag{314}$$

where $p_{ij}(n)$ is the probability to transit from state i to state j in exactly n steps and $f_{ij}(n)$ is the probability that $T_j = n$ with the Markov chain starting from state i at time 0.

$$p_{ij}(n) = \mathbb{P}(X_n = j | X_0 = i) \tag{315}$$

$$f_{ij}(n) = \mathbb{P}_i(T_j = n) \tag{316}$$

it's natural to think of applying Markov property to get the recurrence relationship of those two generating functions.

Theorem 21. *The recurrence relationship differs according to the starting state i and ending state j as*

$$P_{ii}(s) = 1 + F_{ii}(s) P_{ii}(s) \tag{317}$$

$$P_{ij}(s) = F_{ij}(s) P_{jj}(s) \ (i \neq j) \tag{318}$$

Proof. Consider the transition probability and apply the **first visit decomposition** w.r.t. the realization of T_j

$$p_{ij}(0) = \delta_{ij} \tag{319}$$

$$\forall n \geq 1, p_{ij}(n) = \mathbb{P}_i\left(X_n = j\right) \tag{320}$$

$$= \sum_{k=0}^{n} \mathbb{P}_i\left(T_j = k\right) \cdot \mathbb{P}\left(X_n = j | T_j = k\right) \tag{321}$$

$$= \sum_{k=0}^{n} \mathbb{P}_i\left(T_j = k\right) \cdot \mathbb{P}_j\left(X_{n-k} = j\right) \tag{322}$$

$$= \sum_{k=0}^{n} f_{ij}(k) \cdot p_{jj}(n-k) \tag{323}$$

as a result, when $i \neq j, p_{ij}(0) = 0$, we have proved that

$$P_{ij}(s) = F_{ij}(s) P_{jj}(s) \tag{324}$$

this is due to the strong Markov property. To look into the detail, we are actually setting shift operator θ_{T_j} with $T_j < \infty$ *a.s.* (this can be proved, one can also choose to multiply by $\mathbb{I}_{T_j < \infty}$ on both sides) and

$$Y_k(\omega) = \mathbb{I}_{\omega_{n-k}=j}(\omega) \tag{325}$$

that's why

$$\mathbb{E}(Y_{T_j} \circ \theta_{T_j} | \mathscr{F}_{T_j}) = \mathbb{P}\left(X_n = j | \mathscr{F}_{T_j}\right) \tag{326}$$

$$\mathbb{E}_{X_{T_j}} Y_{T_j} = \mathbb{P}_{X_{T_j}}\left(X_{n-T_j} = j\right) \tag{327}$$

and $\mathbb{P}\left(X_n = j | T_j = k\right) = \mathbb{P}_j\left(X_{n-k} = j\right)$.

Now consider the case where $i = j$, so $P_{ij}(s) = 1$ and

$$P_{ii}(s) - 1 = F_{ii}(s) P_{ii}(s) \tag{328}$$

so the other case is also proved.

□

Theorem 22. *State j is recurrent if and only if $\sum_{n=0}^{\infty} p_{jj}(n) = \infty$. If this is true, then for $\forall i$ if there exists some p such that $f_{ij}(p) > 0$, then $\sum_{n=0}^{\infty} p_{ij}(n) = \infty$.*

State j is transient if and only if $\sum_{n=0}^{\infty} p_{jj}(n) < \infty$. If this is true, then $\forall i, \sum_{n=0}^{\infty} p_{ij}(n) < \infty$.

Proof. Since now $P_{jj}(s) = 1 + P_{jj}(s) F_{jj}(s)$, we know

$$P_{jj}(s) = \frac{1}{1 - F_{jj}(s)} \tag{329}$$

so $P_{jj}(1) = \infty$ if and only if $F_{jj}(1) = 1$. Notice that

$$F_{jj}(1) = \mathbb{P}_j\left(T_j < \infty\right) \tag{330}$$

$$P_{jj}(1) = \sum_{n=0}^{\infty} p_{jj}(n) \tag{331}$$

so it's proved for the recurrent and transient case.

If j is recurrent, for any i such that there exists p such that $f_{ij}(p) > 0$, then

$$\sum_{n=0}^{\infty} p_{ij}(s) = P_{ij}(1) = F_{ij}(1)P_{jj}(1) \tag{332}$$

now we know that $F_{ij}(1) \geq f_{ij}(p) > 0$ and $P_{jj}(1) = \infty$, so the conclusion is proved. □

Remark. *If state j is transient, then $\sum_{n=0}^{\infty} p_{jj}(n) < \infty$, so it's obvious that $p_{jj}(n) \to 0$ $(n \to \infty)$.*

If one define $N_{ij}(n)$ as the overall times the Markov chain hits state j no later than time n starting from state i, then by definition

$$N_{ij}(n) = \left(\sum_{k=1}^{n} \mathbb{I}_{X_k=j}\right)\Big|_{X_0=i} \tag{333}$$

then by monotone convergence theorem

$$\mathbb{E}N_{ij}(\infty) = \sum_{k=1}^{\infty} \mathbb{P}_i\left(X_k = j\right) = \sum_{k=1}^{\infty} p_{ij}(k) \tag{334}$$

the series stands for ***the expectation of the time of visit to j starting from i.***

From a different perspective, one might be able to find that recurrence is actually very natural and easy to understand. Note that **state j is recurrent if and only if**

$$\mathbb{P}_j\left(T_j < \infty\right) = 1 \tag{335}$$

let's denote T_j^k as the k-th hitting time to j of the Markov chain, by Markov property applied for $\mathscr{F}_{T_j^1}$,

$$\mathbb{P}_i\left(T_j^k < \infty\right) = \mathbb{P}_i\left(T_j^1 < \infty\right) \cdot \left[\mathbb{P}_j\left(T_j^1 < \infty\right)\right]^{k-1} \tag{336}$$

from this result, one can conclude that state j is recurrent implies

$$\forall k \geq 1, \mathbb{P}_j\left(T_j^k < \infty\right) = 1 \tag{337}$$

and this is equivalent to saying that

$$\mathbb{P}\left(X_n = j \ i.o.\right) = 1 \tag{338}$$

almost surely the Markov chain hits state j for infinitely many times, and this implies $\mathbb{E}N_{jj}(\infty) = \infty$. By noticing the fact that

$$\mathbb{P}_j\left(T_j^k \leq n\right) = \mathbb{P}\left(N_{jj}(n) \geq k\right) \tag{339}$$

$$\mathbb{E}N_{jj}(\infty) = \sum_{k=1}^{\infty} \mathbb{P}\left(N_{jj}(\infty) \geq k\right) \tag{340}$$

it's natural that **state j is recurrent if and only if the expectation of the time of visit from j to j is infinite**.

Recurrent states can be separated into two kinds of states: **positive recurrent** and **null recurrent states**. Positive recurrent state j has

$$\mathbb{E}_j T_j < \infty \tag{341}$$

while null recurrent state j has

$$\mathbb{E}_j T_j = \infty \tag{342}$$

actually the names of the definition has something to do with the existence of the stationary distribution which will be mentioned in the later context.

Concentration Inequality

Concentration inequalities are tools that describe the phenomenon of a function of independent random variables taking values concentrated near its expectation. The easiest example is the Chebyshev inequality where $X_1, ..., X_n$ *i.i.d.* with finite second moments, then

$$\forall \varepsilon > 0, \mathbb{P}\left(\left|\frac{S_n}{n} - \mathbb{E}X_1\right| \geq \varepsilon\right) \leq \frac{Var(X_1)}{n\varepsilon^2} \tag{343}$$

so the value of $\frac{S_n}{n}$ concentrates near its expectation with the tail probability decaying at the rate $\frac{1}{n}$. However, this is typically not a tight bound and does not work well in many cases, that's why we need to develop more concentration inequalities.

We have already proved the following Chernoff concentration bound above. The idea of this bound is that for random variables with finite exponential moments, one always apply Markov inequality $\mathbb{P}\left(\sum_{i=1}^n X_i \geq na\right) \leq \frac{\mathbb{E}e^{\lambda \sum_{i=1}^n X_i}}{e^{\lambda na}} = e^{n \log \mathbb{E}e^{\lambda X_1} - \lambda na}$ and then specify the $\lambda > 0$ by taking inf to get the tightest bound.

Theorem 23. ***(Chernoff Concentration Bound)*** *X_i are i.i.d. random variables, then $\forall a \geq \mathbb{E}X_1$, exists a rate $I(a)$ such that*

$$\mathbb{P}\left(\sum_{i=1}^n X_i \geq na\right) \leq e^{-nI(a)} \tag{344}$$

where the rate is defined as

$$I(a) = \sup_{t \geq 0}\left\{ta - \log \mathbb{E}[e^{tX_1}]\right\} \tag{345}$$

On the other hand, for $\forall a \leq \mathbb{E}X_1$, exists a rate $I(a)$ such that

$$\mathbb{P}\left(\sum_{i=1}^n X_i \leq na\right) \leq e^{-nI(a)} \tag{346}$$

where the rate is defined as

$$I(a) = \sup_{t \leq 0}\left\{ta - \log \mathbb{E}[e^{tX_1}]\right\} \tag{347}$$

Apply the Chernoff bound for Gaussian random variables to get the following bound

Theorem 24. ***(Gaussian Chernoff Bound)*** *$X_1, ..., X_n \sim N(0, \sigma^2)$ i.i.d., then*

$$\forall a \geq \mathbb{E}X_1, \mathbb{P}\left(\sum_{i=1}^n X_i \geq na\right) \leq e^{-n\frac{a^2}{2\sigma^2}} \tag{348}$$

Proof. Notice that $\mathbb{E}e^{\lambda X_1} = e^{\frac{1}{2}\sigma^2\lambda^2}$ so $a\lambda - \log \mathbb{E}e^{\lambda X_1} = a\lambda - \frac{\sigma^2\lambda^2}{2}$ and the sup is taken at $\lambda^* = \frac{a}{\sigma^2}$. Plug it back to

see that the tail probability bound is $e^{-n\frac{a^2}{2\sigma^2}}$.

□

We start to see the exponential decaying in the tail probability when the random variable is Gaussian. Similar bound also works for sub-Gaussian random variables. If $\exists \mu \in \mathbb{R}, \sigma > 0, \mathbb{E}e^{\lambda(X-\mu)} \leq e^{\frac{\sigma^2\lambda^2}{2}}$, we call this X **sub-Gaussian** with parameter σ.

Theorem 25. ***(Sub-Gaussian Chernoff Bound)*** *$X_1, ..., X_n$ are independent and sub-Gaussian with parameters $\sigma_1, ..., \sigma_n$, $\mathbb{E}X_i = \mu_i$, then*

$$\forall a > 0, \mathbb{P}\left(\sum_{i=1}^{n}(X_i - \mu_i) \geq a\right) \leq e^{-\frac{a^2}{2\sum_{i=1}^{n}\sigma_i^2}} \tag{349}$$

Proof. Apply Markov inequality to see that

$$\forall \lambda > 0, \forall a > 0, \mathbb{P}\left(\sum_{i=1}^{n}(X_i - \mu_i) \geq a\right) \leq \mathbb{E}e^{\lambda\sum_{i=1}^{n}(X_i-\mu_i)}e^{-\lambda a} \tag{350}$$

$$\leq e^{\frac{\lambda^2}{2}\sum_{i=1}^{n}\sigma_i^2 - \lambda a} \tag{351}$$

minimize the bound w.r.t. λ to see that minimum is achieved when $\lambda^* = \frac{a}{\sum_{i=1}^{n}\sigma_i^2}$ and the inequality is proved. □

By changing the Gaussian condition into sub-Gaussian, now we do not require identical distribution but only requires independency. We proceed now to find a family of random variables which is sub-Gaussian to apply this bound. It's natural to see that all bounded random variables shall be sub-Gaussian.

Theorem 26. ***(Hoeffding's Lemma)*** *Any bounded random variable X, $a \leq X \leq b$ a.s., $\mathbb{E}X = \mu$, then $Var(X) \leq \frac{(b-a)^2}{4}$ and $\mathbb{E}e^{\lambda(X-\mu)} \leq e^{\frac{\lambda^2(b-a)^2}{8}}$.*

Proof. Since $\left|X - \frac{a+b}{2}\right| \leq \frac{b-a}{2}, Var(X) = Var\left(X - \frac{a+b}{2}\right) \leq \mathbb{E}\left(X - \frac{a+b}{2}\right)^2 \leq \frac{(b-a)^2}{4}$.

To prove the other conclusion, denote $\psi_X(\lambda) = \log \mathbb{E}_{\mathbb{P}}e^{\lambda X}$ so we want to build estimation for $\psi_X(\lambda)$. Consider applying the change of measure that

$$\frac{d\mathbb{P}_\lambda}{d\mathbb{P}} = e^{\lambda X - \psi_X(\lambda)} = \frac{e^{\lambda X}}{\mathbb{E}_{\mathbb{P}}e^{\lambda X}} > 0 \tag{352}$$

so since $\mathbb{E}_{\mathbb{P}}e^{\lambda X - \psi_X(\lambda)} = 1$ we know that $\mathbb{P}_\lambda$ is also a probability measure but $\mathbb{P}_\lambda([a,b]) = \int_{[a,b]} e^{\lambda X - \psi_X(\lambda)}\, d\mathbb{P} = 1$ since $a \leq X \leq b$ *a.s.* and all probability mass of $\mathbb{P}_\lambda$ is concentrated on $[a,b]$. (In the following context, we label the expectation and variance w.r.t. two different measures)

As a result, any random variable Z under probability measure $\mathbb{P}_\lambda$ has variance less than $\frac{(b-a)^2}{4}$ (since $Var_{\mathbb{P}_\lambda}(Z) = Var_{\mathbb{P}_\lambda}(Z\mathbb{I}_{Z\in[a,b]}) \leq \frac{(b-a)^2}{4}$) and

$$\psi_X''(\lambda) = \frac{d}{d\lambda}\frac{\mathbb{E}_{\mathbb{P}}Xe^{\lambda X}}{\mathbb{E}_{\mathbb{P}}e^{\lambda X}} = \frac{\mathbb{E}_{\mathbb{P}}X^2e^{\lambda X}\mathbb{E}_{\mathbb{P}}e^{\lambda X} - \mathbb{E}_{\mathbb{P}}^2Xe^{\lambda X}}{\mathbb{E}_{\mathbb{P}}^2e^{\lambda X}} \tag{353}$$

find the structure of RN-derivative in this expression that

$$\forall \lambda > 0, \psi_X''(\lambda) = e^{-\psi_X(\lambda)} \mathbb{E}_{\mathbb{P}} X^2 e^{\lambda X} - e^{-2\psi_X(\lambda)} \mathbb{E}_{\mathbb{P}}^2 X e^{\lambda X} \tag{354}$$

$$= \mathbb{E}_{\mathbb{P}_\lambda} X^2 - \mathbb{E}_{\mathbb{P}_\lambda}^2 X \tag{355}$$

$$= Var_{\mathbb{P}_\lambda} X \leq \frac{(b-a)^2}{4} \tag{356}$$

note that $\psi_X(0) = 0, \psi_X'(0) = \mathbb{E}_{\mathbb{P}} X = \mu$, by Taylor expansion,

$$\psi_X(\lambda) = \psi_X(0) + \psi_X'(0)\lambda + \frac{\psi_X''(\xi)}{2}\lambda^2 \tag{357}$$

$$\leq \mu\lambda + \frac{(b-a)^2}{8}\lambda^2 \tag{358}$$

where $\exists \xi, 0 < \xi < \lambda$ and $\mathbb{E}e^{\lambda X} \leq e^{\mu\lambda + \frac{(b-a)^2}{8}\lambda^2}$. □

Remark. *There is also a way to prove this fact by Jensen's inequality but the change of measure technique is more fundamental and essential. Note that this bound is actually tight since for X taking values ± 1 with probability $\frac{1}{2}$ each, $VarX = \frac{(1+1)^2}{2} = 1$.*

Hoeffding's lemma tells us that all bounded random variables that take values in $[a, b]$ is sub-Gaussian with parameter $\sigma = \frac{b-a}{2}$. The Chernoff bound applied for bounded random variables gives the following concentration bound.

Theorem 27. ***(Hoeffding's Inequality for Independent Bounded Random Variables)*** *$X_1, ..., X_n$ are independent bounded random variables with $a_i \leq X_i \leq b_i$ a.s. and $\mathbb{E}X_i = \mu_i$, then*

$$\forall a > 0, \mathbb{P}\left(\sum_{i=1}^n (X_i - \mu_i) \geq a\right) \leq e^{-\frac{2a^2}{\sum_{i=1}^n (b_i - a_i)^2}} \tag{359}$$

Proof. Apply the Chernoff bound for sub-Gaussian random variables to see X_i is sub-Gaussian with parameter $\sigma_i = \frac{b_i - a_i}{2}$, so

$$\forall a > 0, \mathbb{P}\left(\sum_{i=1}^n (X_i - \mu_i) \geq a\right) \leq e^{-\frac{a^2}{2\sum_{i=1}^n \frac{(b_i - a_i)^2}{4}}} = e^{-\frac{2a^2}{\sum_{i=1}^n (b_i - a_i)^2}} \tag{360}$$

□

Remark. *Apply Hoeffding's inequality for random signs X_i taking values ± 1 with value $\frac{1}{2}$ each to see that $b_i - a_1 = 2, \mu_i = 0$ so*

$$\forall a > 0, \mathbb{P}\left(\sum_{i=1}^n X_i \geq a\right) \leq e^{-\frac{a^2}{2n}} \tag{361}$$

which gives us ***Hoeffding's inequality for random signs****.*

Remark. *Notice that the other half of Hoeffding's inequality is of course true, so*

$$\forall a>0, \mathbb{P}\left(\sum_{i=1}^{n}(X_i-\mu_i)\leq a\right)\leq e^{-\frac{2a^2}{\sum_{i=1}^{n}(b_i-a_i)^2}} \tag{362}$$

and combine two parts to see that

$$\forall a>0, \mathbb{P}\left(\left|\sum_{i=1}^{n}(X_i-\mu_i)\right|\geq a\right)\leq 2e^{-\frac{2a^2}{\sum_{i=1}^{n}(b_i-a_i)^2}} \tag{363}$$

so

$$\forall a>0, \mathbb{P}\left(|S_n-\mathbb{E}S_n|\geq a\right)\leq 2e^{-\frac{2a^2}{\sum_{i=1}^{n}(b_i-a_i)^2}} \tag{364}$$

gives the tail probability bound for both left and right tails of the sum S_n. As a result, in most cases we only need to bound the right tail and actually both tails can be bounded simultaneously.

One can slightly relax the independence condition in Hoeffding's inequality to get the following Azuma's inequality. However, it's still required that the process is a martingale.

Theorem 28. ***(Azuma's Inequality)*** *$\{X_n\}_{n\geq 0}$ is a MG with bounded increments $A_k\leq X_k-X_{k-1}\leq B_k$ a.s., then*

$$\forall a>0, \mathbb{P}\left(X_n-X_0\geq a\right)\leq e^{-\frac{2a^2}{\sum_{i=1}^{n}(B_i-A_i)^2}} \tag{365}$$

Proof. Start from Markov inequality and Hoeffding's lemma (it also holds for conditional expectation) that

$$\begin{aligned}
\forall a>0, \mathbb{P}\left(X_n-X_0\geq a\right) &= \mathbb{P}\left(\sum_{i=1}^{n}(X_i-X_{i-1})\geq a\right) && (366)\\
&\leq \mathbb{E}e^{\lambda\sum_{i=1}^{n}(X_i-X_{i-1})}e^{-\lambda a} && (367)\\
&= \mathbb{E}\left[\mathbb{E}\left(e^{\lambda\sum_{i=1}^{n}(X_i-X_{i-1})}\middle|X_0,...,X_{n-1}\right)\right]e^{-\lambda a} && (368)\\
&= \mathbb{E}\left[e^{\lambda\sum_{i=1}^{n-1}(X_i-X_{i-1})}\mathbb{E}\left(e^{\lambda(X_n-X_{n-1})}\middle|X_0,...,X_{n-1}\right)\right]e^{-\lambda a} && (369)\\
&\leq \mathbb{E}\left[e^{\lambda\sum_{i=1}^{n-1}(X_i-X_{i-1})}\right]e^{\frac{(B_n-A_n)^2\lambda^2}{8}-\lambda a} && (370)\\
&\leq \cdots && (371)\\
&\leq e^{\frac{\sum_{i=1}^{n}(B_i-A_i)^2\lambda^2}{8}-\lambda a} && (372)
\end{aligned}$$

so minimize w.r.t. $\forall\lambda>0$ to see that the tightest bound is taken when $\lambda^*=\frac{4a}{\sum_{i=1}^{n}(B_i-A_i)^2}$ and the inequality is proved. □

Remark. *Apply Azuma's inequality for symmetric simple random walk $\{S_n\}$ starting from 0 to see that each increment is between $[-1,1]$ so*

$$\forall a>0, \mathbb{P}\left(S_n \geq a\right) \leq e^{-\frac{a^2}{2n}} \tag{373}$$

and this tells us that

$$\mathbb{P}\left(S_n \geq \sqrt{2n\log n}\right) \leq \frac{1}{n} \tag{374}$$

$$\mathbb{P}\left(S_n \geq \sqrt{2n\log\log n}\right) \leq \frac{1}{\log n} \tag{375}$$

$$\mathbb{P}\left(S_n \geq \sqrt{2n\log n + 4n\log\log n}\right) \leq \frac{1}{n\log^2 n} \tag{376}$$

so by Borel-Cantelli, we can conclude that $\mathbb{P}\left(S_n \geq \sqrt{2n\log n + 4n\log\log n}\ i.o.\right) = 0$, so almost surely eventually $S_n < \sqrt{2n\log n + 4n\log\log n}$ and this is consistent with the law of iterated logarithm.

The easiest general concentration inequality for any Lipschitz function $F:\mathbb{R}^m \to \mathbb{R}$ is established for Gaussian random variables.

Theorem 29. ***(Concentration of Gaussian Measure for Lipschitz Function)*** *$X_1, ..., X_n \sim N(0,1)$ i.i.d. with $F:\mathbb{R}^n \to \mathbb{R}$ Lipschitz with Lipschitz constant L then*

$$\forall a>0, \mathbb{P}\left(|F(X) - \mathbb{E}F(X)| \geq a\right) \leq 2e^{-\frac{2a^2}{\pi^2 n L^2}} \tag{377}$$

Proof. WLOG, assume $\mathbb{E}F(X)=0$ and $L \leq 1$, so $||\nabla F||_\infty \leq 1$. (Can always consider $\frac{F}{L}$ with Lipschitz constant less than 1. Since F Lipschitz, it's almost everywhere differentiable and the magnitude of any partial derivative cannot exceed the Lipschitz constant) By Markov inequality,

$$\forall a>0, \mathbb{P}\left(F(X) \geq a\right) \leq \mathbb{E}e^{\lambda F(X)}e^{-\lambda a} \tag{378}$$

so we only have to calculate $\mathbb{E}e^{\lambda F(X)}$.

If $Y = (Y_1, ..., Y_n) \overset{d}{=} X$ and Y is independent of X, then by Jensen's inequality, $\mathbb{E}e^{-\lambda F(Y)} \geq e^{-\lambda \mathbb{E}F(Y)} = 1$, so

$$\mathbb{E}e^{\lambda F(X)} \leq \mathbb{E}e^{\lambda[F(X)-F(Y)]} \tag{379}$$

makes the difference $F(X) - F(Y)$ appear. Consider

$$F(X) - F(Y) = \int_0^{\frac{\pi}{2}} \frac{d}{d\theta} F(\cos\theta \cdot Y + \sin\theta \cdot X)\, d\theta \tag{380}$$

and denote $Z_\theta = \cos\theta \cdot Y + \sin\theta \cdot X$ to find that Z_θ is independent of $\frac{d}{d\theta}Z_\theta = -\sin\theta \cdot Y + \cos\theta \cdot X$ and both $Z_\theta, \frac{d}{d\theta}Z_\theta$ have distribution $N(0, I_n)$ (since it's an orthogonal transformation).

Apply Jensen's inequality again for the uniform probability measure on $\left(0, \frac{\pi}{2}\right)$ to see

$$\mathbb{E}e^{\lambda[F(X)-F(Y)]} = \mathbb{E}e^{\frac{2}{\pi}\int_0^{\frac{\pi}{2}} \frac{\pi\lambda}{2}\frac{d}{d\theta}F(Z_\theta)\,d\theta} \tag{381}$$

$$\leq \frac{2}{\pi}\mathbb{E}\int_0^{\frac{\pi}{2}} e^{\frac{\pi\lambda}{2}\frac{d}{d\theta}F(Z_\theta)}\,d\theta \tag{382}$$

$$= \frac{2}{\pi}\int_0^{\frac{\pi}{2}} \mathbb{E}e^{\frac{\pi\lambda}{2}\nabla F(Z_\theta)\cdot\frac{d}{d\theta}Z_\theta}\,d\theta \tag{383}$$

let's calculate the expectation involved to find

$$\mathbb{E}e^{\frac{\pi\lambda}{2}\nabla F(Z_\theta)\cdot\frac{d}{d\theta}Z_\theta} \leq \mathbb{E}e^{\frac{\pi\lambda}{2}||\nabla F(Z_\theta)||_\infty\cdot||\frac{d}{d\theta}Z_\theta||_1} \tag{384}$$

$$\leq \mathbb{E}e^{\frac{\pi\lambda}{2}||\frac{d}{d\theta}Z_\theta||_1} \tag{385}$$

$$\leq \left[\mathbb{E}e^{\frac{\pi\lambda}{2}|Z|}\right]^n = e^{\frac{\pi^2 n}{8}\lambda^2} \tag{386}$$

where $Z \sim N(0,1)$.

So the one-sided probability bound is given by $e^{\frac{\pi^2 n}{8}\lambda^2-\lambda a}$ and the tightest bound is given by $\lambda^* = \frac{4a}{\pi^2 n}$, the theorem is proved since the other side can also be bounded by the same bound. □

Remark. *The theorem tells us that for Lipschitz actions on independent Gaussian random variables the action values concentrate in the interval centered at the expectation with radius $\sqrt{n}$. This Gaussian concentration inequality is not that interesting since there's an n on the denominator of the tail probability. When $n \to \infty$, we can see that the bound becomes looser and becomes trivial eventually, so there is dimensional dependency shown in this inequality. However, we can actually get a much stronger result that is independent of the dimension.*

Theorem 30. ***(Gaussian Concentration, Tsirelson-Ibragimov-Sudakov)*** *$X_1, ..., X_n \sim N(0,1)$ i.i.d. with $F : \mathbb{R}^n \to \mathbb{R}$ Lipschitz with Lipschitz constant L then*

$$\forall a > 0, \mathbb{P}\left(|F(X) - \mathbb{E}F(X)| \geq a\right) \leq 2e^{-\frac{a^2}{2L^2}} \tag{387}$$

The proof of this theorem makes use of the log-Sobolev inequality applied for Gaussian measure that $Ent(F^2) \leq 2\mathbb{E}||\nabla F(X)||_2^2$ so $\log \mathbb{E}e^{\lambda F(X)} \leq \frac{\lambda^2 L^2}{2}$, combining with the Chernoff bound we can see that this theorem holds. This concentration inequality is much more powerful since it's **independent of the dimension**.

Stein's Method

Stein's method provides another characterization of Gaussian random variables. The advantage of this method is that it does not depend on characteristic function, can be generalized to other distributions and gives the rate of convergence. Stein's method starts from the following characterization of Gaussian random variables.

Motivation for Gaussian Approximation

Theorem 31. ***(Stein's Lemma)*** *If $Z \sim N(0,1)$ then $\forall f \in C^1$ such that $\mathbb{E}|Zf(Z)| < \infty, \mathbb{E}|f'(Z)| < \infty$, $\mathbb{E}Zf(Z) = \mathbb{E}f'(Z)$. Conversely, if $\forall f$ bounded continuous and piecewise C^1 such that $\mathbb{E}|Zf(Z)| < \infty, \mathbb{E}|f'(Z)| < \infty$, $\mathbb{E}Zf(Z) = \mathbb{E}f'(Z)$, then $Z \sim N(0,1)$.*

Proof. If $Z \sim N(0,1)$, denote φ as standard Gaussian PDF so

$$\mathbb{E}f'(Z) = \int_{\mathbb{R}} f'(z)\varphi(z)\,dz \tag{388}$$

$$= f \cdot \varphi\Big|_{-\infty}^{+\infty} - \int_{\mathbb{R}} f(z)\varphi'(z)\,dz \tag{389}$$

notice that since $\mathbb{E}|Zf(Z)| < \infty$, $zf(z)\varphi(z) \to 0$ $(z \to \infty)$ so $f \cdot \varphi\Big|_{-\infty}^{+\infty} = 0$ and notice that for standard Gaussian PDF, it's true that the **dual equation** $\varphi'(z) = -z\varphi(z)$ is true, so

$$\mathbb{E}f'(Z) = \int_{\mathbb{R}} zf(z)\varphi(z)\,dz = \mathbb{E}Zf(Z) \tag{390}$$

Conversely, consider the following ODE

$$f'(x) - xf(x) = g(x) \tag{391}$$

it's true that $f_g(x) = e^{\frac{x^2}{2}} \int_{-\infty}^{x} g(t)e^{-\frac{t^2}{2}}\,dt$ is the solution. By taking $g(x) = \mathbb{I}_{x \le x_0} - \Phi(x_0)$ for $\forall x_0 \in \mathbb{R}$, we see that f_g is bounded continuous and is C^1 except at x_0. So $\mathbb{E}g(Z) = 0$ and $\forall x_0 \in \mathbb{R}, \mathbb{P}(Z \le x_0) = \Phi(x_0)$ so $Z \sim N(0,1)$. □

Remark. *Stein's lemma characterizes normality by saying that it's equivalent to having*

$$\mathbb{E}Zf(Z) = \mathbb{E}f'(Z) \tag{392}$$

hold for a class of functions f. Actually, this is also true for $W \sim N(\mu, \sigma^2)$ since $\frac{W-\mu}{\sigma} \sim N(0,1)$ so by setting $h(x) = f\left(\frac{x-\mu}{\sigma}\right)$,

$$\mathbb{E}\frac{W-\mu}{\sigma} f\left(\frac{W-\mu}{\sigma}\right) = \mathbb{E}f'\left(\frac{W-\mu}{\sigma}\right) \tag{393}$$

$$\mathbb{E}\frac{W-\mu}{\sigma} h(W) = \sigma\mathbb{E}h'(W) \tag{394}$$

so the characterization of $W \sim N(\mu, \sigma^2)$ *is given by*

$$\mathbb{E}(W-\mu)f(W) = \sigma^2 \mathbb{E}f'(W) \tag{395}$$

to hold for a class of functions f*.*

The useful point of Stein's method is that it's natural to think about that if $\mathbb{E}[Zf(Z) - f'(Z)]$ is not zero but is close to 0, then we should be able to argue that Z is close to a standard Gaussian. Now the problem is that we have to build up the distance between two probability measure to describe what it means by saying "Z is close to a standard Gaussian". Actually, many distances between probability distribution has the form as $d(\mathbb{P}, \mathbb{Q}) = \sup_{h \in \mathscr{H}} |\mathbb{E}_{X\sim\mathbb{P}} h(X) - \mathbb{E}_{Y\sim\mathbb{Q}} h(Y)|$ for $\mathscr{H}$ as a class of functions (test functions). Intuitively, the distance between two probability measures is formed as the maximum possible difference between the expected value under the image of test functions.

By taking $\mathscr{H} = \left\{\mathbb{I}_{(-\infty,z)}(x), \forall z \in \mathbb{R}\right\}$, the **Kolmogorov's distance** is

$$||\mathbb{P} - \mathbb{Q}||_\infty = \sup_{z\in\mathbb{R}} |F_X(z) - F_Y(z)| = ||F_X - F_Y||_\infty \tag{396}$$

where F_X is the CDF of $\mathbb{P}$ and F_Y is the CDF of $\mathbb{Q}$.

By taking $\mathscr{H} = \{h \ Lipschitz\}$, the **Wasserstein distance** is

$$||\mathbb{P} - \mathbb{Q}||_1 = \int_{\mathbb{R}} |F_X(z) - F_Y(z)|\, dz = ||F_X - F_Y||_1 \tag{397}$$

where F_X is the CDF of $\mathbb{P}$ and F_Y is the CDF of $\mathbb{Q}$ and the Wasserstein distance is characterized by the coupling of $\mathbb{P}, \mathbb{Q}$.

By taking $\mathscr{H} = \{h = \mathbb{I}_A, \forall A \in \mathscr{F} \ measurable\}$, the **total variation** is

$$TV(\mathbb{P}, \mathbb{Q}) = \sup_{A\in\mathscr{F}} |\mathbb{P}(A) - \mathbb{Q}(A)| \tag{398}$$

where F_X is the CDF of $\mathbb{P}$ and F_Y is the CDF of $\mathbb{Q}$.

So now our objective is to estimate the distance between the distribution of W and the distribution of $Z \sim N(0,1)$ by estimating $|\mathbb{E}h(W) - \mathbb{E}h(Z)|$ for $h \in \mathscr{H}$. We shall take the similar approach as we have done previously to consider the ODE

$$f'(x) - xf(x) = h(x) - \mathbb{E}h(Z) \tag{399}$$

and the solution is given by $f_h(x) = e^{\frac{x^2}{2}} \int_{-\infty}^{x} [h(t) - \mathbb{E}h(Z)] e^{-\frac{t^2}{2}}\, dt$ so we see that

$$\mathbb{E}h(W) - \mathbb{E}h(Z) = \mathbb{E}[f_h'(W) - W f_h(W)] \tag{400}$$

and the distance between two probability distributions can be estimated by calculating $\mathbb{E}[f_h'(W) - W f_h(W)]$.

To prove that W is asymptotically standard Gaussian (which depends on n), one typically proves that $|\mathbb{E}[f_h'(W) - Wf_h(W)]| \to 0$ $(n \to \infty)$ with $f_h(x) = e^{\frac{x^2}{2}} \int_{-\infty}^{x} [h(t) - \mathbb{E}h(Z)]e^{-\frac{t^2}{2}}\, dt$ given and h depends on the distance between probability measures one is using.

Central Limit Theorem

Let's say we want to prove Lindeberg CLT using Stein's method. So $X_1, ..., X_n$ *i.i.d.* with $\mathbb{E}X_i = 0, Var(X_i) = 1$ and it's natural to form $W = \frac{S_n}{\sqrt{n}}$ and we want to prove that $W \xrightarrow{d} N(0,1)$ $(n \to \infty)$.

By the scheme we have discussed above, it's equivalent to estimating

$$\mathbb{E}[f_h'(W) - Wf_h(W)] \tag{401}$$

where $f_h(x) = e^{\frac{x^2}{2}} \int_{-\infty}^{x} [h(t) - \mathbb{E}h(Z)]e^{-\frac{t^2}{2}}\, dt$. Of course, we don't want to directly figure out f_h so we want to use the Taylor expansion of f_h. As a result, it's natural to write W into the form with perturbation. The trick is to use **"leave-one-out"** scheme meaning that we set $X_i' = \frac{X_i}{\sqrt{n}}$ so $W = \sum_{i=1}^n X_i'$ and $W_i = \sum_{j \neq i} X_j' = W - X_i'$ so $W = W_i + X_i'$ is convenient to apply Taylor expansion

$$f_h(W) = f_h(W_i) + f_h'(W_i)X_i' + R_i \tag{402}$$

where $R_i = \frac{f_h''(\xi_i)}{2}(X_i')^2$ for some ξ_i between W and W_i. Now we see that

$$\begin{aligned}
\mathbb{E}Wf_h(W) &= \sum_{i=1}^n \mathbb{E}X_i' f_h(W) && (403)\\
&= \sum_{i=1}^n \mathbb{E}X_i' f_h(W_i) + \sum_{i=1}^n \mathbb{E}(X_i')^2 f_h'(W_i) + \sum_{i=1}^n \mathbb{E}X_i' R_i && (404)\\
&= \sum_{i=1}^n \mathbb{E}X_i' \mathbb{E}f_h(W_i) + \sum_{i=1}^n \mathbb{E}(X_i')^2 \mathbb{E}f_h'(W_i) + \sum_{i=1}^n \mathbb{E}X_i' R_i && (405)\\
&= \frac{1}{n}\sum_{i=1}^n \mathbb{E}f_h'(W_i) + R && (406)
\end{aligned}$$

where $R = \sum_{i=1}^n \mathbb{E}X_i' R_i$ is the remainder and the calculation is based on the fact that W_i is independent of X_i'.

As a result, we see that

$$\mathbb{E}[f_h'(W) - Wf_h(W)] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[f_h'(W) - f_h'(W_i)] + R \tag{407}$$

and the remainder term R can be further estimated. Here we introduce the following technical lemma.

Lemma 1. *If h is absolute continuous, then $||f_h''||_\infty \leq 2||h'||_\infty$. (Refer to Normal Approximation by Stein's Method for proof)*

We see that by the lemma

$$R = \sum_{i=1}^{n} \mathbb{E}(X_i')^3 \frac{f_h''(\xi_i)}{2} \tag{408}$$

$$|R| \leq \sum_{i=1}^{n} \mathbb{E}|X_i'|^3 \frac{||f_h''||_\infty}{2} \leq ||h'||_\infty \sum_{i=1}^{n} \mathbb{E}|X_i'|^3 \tag{409}$$

and

$$\left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[f_h'(W) - f_h'(W_i)] \right| \leq \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}|f_h'(W) - f_h'(W_i)| \tag{410}$$

$$= \frac{1}{n} \sum_{i=1}^{n} \mathbb{E} \left| \int_{W_i}^{W} f_h''(t)\, dt \right| \tag{411}$$

$$\leq \frac{||f_h''||_\infty}{n} \sum_{i=1}^{n} \mathbb{E}\,|W - W_i| \tag{412}$$

$$\leq \frac{2||h'||_\infty}{n} \sum_{i=1}^{n} \mathbb{E}|X_i'| \tag{413}$$

$$= \frac{2||h'||_\infty}{n} \sqrt{n}\mathbb{E}|X_i| \tag{414}$$

$$\leq \frac{2||h'||_\infty}{\sqrt{n}} \tag{415}$$

combine those two parts to see that

$$|\mathbb{E}h(W) - \mathbb{E}h(Z)| = |\mathbb{E}[f_h'(W) - W f_h(W)]| \tag{416}$$

$$\leq \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[f_h'(W) - f_h'(W_i)] \right| + |R| \tag{417}$$

$$\leq ||h'||_\infty \left(\frac{2}{\sqrt{n}} + \sum_{i=1}^{n} \mathbb{E}|X_i'|^3 \right) \tag{418}$$

we have proved the following Stein's method for the sum of *i.i.d.* random variables

Theorem 32. ***(Stein's Method for** i.i.d. **Sum)*** *$X_1, ..., X_n$ i.i.d. with $\mathbb{E}X_i = 0, Var(X_i) = 1$ and $W = \frac{S_n}{\sqrt{n}}, Z \sim N(0,1)$, then for any absolute continuous function h with $||h'||_\infty < \infty$,*

$$|\mathbb{E}h(W) - \mathbb{E}h(Z)| \leq ||h'||_\infty \left(\frac{2}{\sqrt{n}} + \sum_{i=1}^{n} \mathbb{E}|X_i'|^3 \right) \tag{419}$$

if all quantities are finite.

Remark. *The theorem above already provides us with some kind of Berry-Esseen bound. Since here we are requiring*

h to be absolute continuous, it's impossible to take h as indicators which provides link to CDF. However, it provides a good connection with the Wasserstein distance and it's possible to relax the condition of h to get more useful bounds. Note that the term $\sum_{i=1}^{n} \mathbb{E}|X_i'|^3$ is a bad term that grows at order n and from the derivation this term only disappears when $\mathbb{E}(X_1')^3 = 0$, which is a too strong condition to hold. This is the main reason why such bound is not that interesting.

This simple example has already shown the power of Stein's method that it is able to imply the rate of convergence under different distances without using c.f. Not depending on c.f. will be an advantage if we are dealing with weak convergence on more general objects like graphs and networks.

Remark. *For Poisson random variables $Z \sim P(\lambda)$, Stein's method gives the characterization $\forall f, \mathbb{E}Zf(Z) = \lambda \mathbb{E}f(Z+1)$ thus similar approach can be taken to prove the Poisson CLT for independent Bernoulli random variables with the rate of convergence measured by the total variation distance.*

Large Deviation Theory

Single Random Variable

The law of large numbers enables us to comment on the limiting behavior of *i.i.d.* sums, CLT enables us to approximate $\mathbb{P}(S_n \geq a\sqrt{n})$ for S_n to be *i.i.d.* sum of square integrable random variables with mean 0 and variance 1. Concentration inequalities provide exponential upper bounds for $\mathbb{P}(S_n \geq a)$ where S_n is the sum of independent random variables. Large deviation theory focuses on bounding $\mathbb{P}(|S_n - \mathbb{E}S_n| > an^\alpha)$ for $\alpha > \frac{1}{2}$ and $a \in \mathbb{R}$. Here "large" refers to the fact that the deviation is of order n^α strictly larger than $\sqrt{n}$, the order of deviation in CLT and it's different from concentration inequality since large deviation theory tells us the asymptotic order of the probability instead of just providing an upper bound so there's finer estimations.

Let's first discuss the large deviation result for **a single random variable Y where $\mathbb{E}Y < 0, \mathbb{P}(Y > 0) > 0$ with $\exists \delta > 0, \forall t \in (-\delta, \delta), M_Y(t) < \infty$ so all moments exist**. This setting is adopted generally since for any random variable X with all moments exist, we want to consider $\mathbb{P}(X - \mathbb{E}X > a)$ for some $a > 0$. By specifying $Y = X - \mathbb{E}X - a$, we see that $\mathbb{E}Y = -a < 0$ and $\mathbb{P}(Y > 0) = \mathbb{P}(X - \mathbb{E}X > a) > 0$.

Theorem 33. ***(Large Deviation for Single R.V.)*** *$\mathbb{E}Y < 0, \mathbb{P}(Y > 0) > 0$ and $\exists \delta > 0, \forall |t| < \delta, M_Y(t) < \infty$, then consider $\rho \overset{def}{=} \inf_{t>0} M_Y(t) \overset{def}{=} M_Y(\tau)$ such that the change of measure $\frac{dF_Z(y)}{dF_Y(y)} = \frac{e^{\tau y}}{\rho}$ gives the distribution of the new r.v. Z, so*

$$\mathbb{E}Z = 0, \mathbb{E}Z^2 \overset{def}{=} s^2 > 0 \tag{420}$$

and

$$\mathbb{P}(Y \geq 0) = \rho e^{-\theta}, 0 \leq \theta \leq \frac{\tau s}{\mathbb{P}(Z \geq 0)} - \log \mathbb{P}(Z \geq 0) \tag{421}$$

Proof. Since $\mathbb{E}\frac{e^{\tau Y}}{\rho} = \frac{M_Y(\tau)}{\rho} = 1$, we know that F_Z is a probability measure on $\mathbb{R}$ so it has to be induced by some random variable Z and since $\frac{e^{\tau y}}{\rho} > 0$, F_Z, F_Y are absolute continuous w.r.t. each other. It's quite obvious that the MGF of Z is given by

$$M_Z(t) = \mathbb{E}e^{tZ} = \int_{\mathbb{R}} e^{ty} \frac{e^{\tau y}}{\rho} \, dF_Y(y) = \frac{1}{\rho} \mathbb{E}e^{(t+\tau)Y} = \frac{1}{\rho} M_Y(t + \tau) \tag{422}$$

notice that $M_Y'(0) = \mathbb{E}Y < 0$ and $\forall t, M_Y''(t) = \mathbb{E}Y^2 e^{tY} > 0$ since $\mathbb{P}(Y > 0) > 0$. So $M_Y(t)$ is continuous, strictly decreasing at 0 and is always convex so ρ is well-defined as the minimum of such function and it must be attained at some point $\tau > 0$. Since $M_Y(0) = 1$ and $M_Y'(0) < 0$, we know $0 < \rho < 1$.

Now Z must also have all moments to be finite since it has finite MGF in some neighborhood of 0, so $\mathbb{E}Z = M_Z'(0) = \frac{1}{\rho} M_Y'(\tau) = 0$ since τ is the minimizer of $M_Y(t)$ so the derivative vanishes. On the other hand, $\mathbb{E}Z^2 = M_Z''(0) = \frac{1}{\rho} M_Y''(\tau) > 0$ is strictly positive so it won't degenerate.

Now let's give estimations of $\mathbb{P}(Y \geq 0)$ based on this new r.v. Z.

$$\mathbb{P}(Y \geq 0) = \int_{\mathbb{R}} \mathbb{I}_{[0,\infty)}(y)\, dF_Y(y) = \rho \int_{\mathbb{R}} e^{-\tau y} \mathbb{I}_{[0,\infty)}(y)\, dF_Z(y) = \rho \cdot \mathbb{E} e^{-\tau Z} \mathbb{I}_{Z \geq 0} \tag{423}$$

so $\theta = -\log \mathbb{E} e^{-\tau Z} \mathbb{I}_{Z \geq 0}$.

By Markov inequality, $\forall t > 0, \mathbb{P}(Y \geq 0) \leq \mathbb{E} e^{tY} = M_Y(t)$, so by taking inf w.r.t. $t > 0$ on both sides, we see that $\mathbb{P}(Y \geq 0) \leq \rho$ and this tells us $\theta \geq 0$.

For the other bound, let's try to figure out a lower bound of $\mathbb{E} e^{-\tau Z} \mathbb{I}_{Z \geq 0}$. Apply Jensen's inequality for conditional expectation and Cauchy Schwarz to get

$$\begin{aligned}
\mathbb{E} e^{-\tau Z} \mathbb{I}_{Z \geq 0} &= \mathbb{E}(e^{-\tau Z} | Z \geq 0) \cdot \mathbb{P}(Z \geq 0) && (424)\\
&\geq e^{-\tau \mathbb{E}(Z | Z \geq 0)} \cdot \mathbb{P}(Z \geq 0) && (425)\\
&= e^{-\frac{\tau}{\mathbb{P}(Z \geq 0)} \mathbb{E}(Z \mathbb{I}_{Z \geq 0})} \cdot \mathbb{P}(Z \geq 0) && (426)\\
&\geq e^{-\frac{\tau}{\mathbb{P}(Z \geq 0)} \sqrt{\mathbb{E} Z^2 \cdot \mathbb{P}(Z \geq 0)}} \cdot \mathbb{P}(Z \geq 0) && (427)\\
&= e^{-\frac{s\tau}{\sqrt{\mathbb{P}(Z \geq 0)}}} \cdot \mathbb{P}(Z \geq 0) && (428)
\end{aligned}$$

so $\theta \leq \frac{s\tau}{\sqrt{\mathbb{P}(Z \geq 0)}} - \log \mathbb{P}(Z \geq 0) \leq \frac{s\tau}{\mathbb{P}(Z \geq 0)} - \log \mathbb{P}(Z \geq 0)$.

□

Remark. *Note that we have actually proved that* $0 \leq \theta \leq \frac{s\tau}{\sqrt{\mathbb{P}(Z \geq 0)}} - \log \mathbb{P}(Z \geq 0)$ *which is* ***a tighter bound****. When we are using this large deviation bound in practice, we often still need a lower bound of* $\mathbb{P}(Z \geq 0)$*, but this is easier since* Z *has nicer property than* Y*.*

The upper bound of $\mathbb{P}(Z \geq 0)$ is always easy since one can apply Markov inequality or Chernoff bound to relate this probability with $M_Z(t)$. However, the lower bound of $\mathbb{P}(Z \geq 0)$ is not that obvious. Luckily, here Z is a random variable with all moments to be finite so the moment estimation provides a lower bound of this probability.

Theorem 34. ***(Moment Lower Bound of*** $\mathbb{P}(Z \geq 0)$***)*** *If* $\mathbb{E}Z = 0, \mathbb{E}Z^2 = s^2, \mathbb{E}Z^4 = \xi^4 > 0$*, then* $\mathbb{P}(Z \geq 0) \geq \frac{s^4}{4\xi^4}$.

Proof. The proof is some simple applications of Holder's inequality and in order to match with the norm of random variables, split Z into positive and negative parts $Z = Z_+ - Z_-$, since $\mathbb{E}Z = 0$, we know $\mathbb{E}Z_+ = \mathbb{E}Z_-$. The trick lies in treating the moments of Z_+, Z_- in different ways

$$\begin{aligned}
\mathbb{E} Z_+^2 &= \mathbb{E} Z^2 \mathbb{I}_{Z \geq 0} \leq \sqrt{\mathbb{E} Z^4 \cdot \mathbb{P}(Z \geq 0)} = \xi^2 \sqrt{\mathbb{P}(Z \geq 0)} && (429)\\
\mathbb{E} Z_-^2 &= \mathbb{E} Z_-^{\frac{4}{3}} Z_-^{\frac{2}{3}} \leq (\mathbb{E} Z_-)^{\frac{2}{3}} (\mathbb{E} Z_-^4)^{\frac{1}{3}} = \xi^{\frac{4}{3}} (\mathbb{E} Z_+)^{\frac{2}{3}} && (430)
\end{aligned}$$

in order to make the upper bound of $\mathbb{E}Z_+^2, \mathbb{E}Z_-^2$ have the same power in $\mathbb{P}(Z \geq 0)$, it's necessary to have $[\mathbb{P}(Z \geq 0)]^{\frac{3}{4}}$ in the upper bound of $\mathbb{E}Z_+$, so it's natural to apply Holder's inequality for conjugate $4, \frac{4}{3}$ to get

$$\mathbb{E} Z_+ = \mathbb{E} Z \mathbb{I}_{Z \geq 0} \leq (\mathbb{E} Z^4)^{\frac{1}{4}} [\mathbb{P}(Z \geq 0)]^{\frac{3}{4}} = \xi [\mathbb{P}(Z \geq 0)]^{\frac{3}{4}} \tag{431}$$

and we see

$$s^2 = \mathbb{E}Z^2 = \mathbb{E}Z_+^2 + \mathbb{E}Z_-^2 \leq \xi^2\sqrt{\mathbb{P}(Z \geq 0)} + \xi^2\sqrt{\mathbb{P}(Z \geq 0)} = 2\xi^2\sqrt{\mathbb{P}(Z \geq 0)} \tag{432}$$

this gives the bound that $\mathbb{P}(Z \geq 0) \geq \frac{s^4}{4\xi^4}$. □

Sum of i.i.d. Random Variables

Let's consider large deviation bound for **the sum of** *i.i.d.* **random variables** $S_n = X_1 + ... + X_n$ **where** $\mathbb{E}X_i = 0$ **and** $\exists \delta > 0, \forall |t| < \delta, M_{X_1}(t) < \infty$ **with** $\mathbb{P}(X_1 > a) > 0$ **for some** $a > 0$. As what we have stated above, "large" refers to the fact that we consider deviations of order strictly larger than $\sqrt{n}$, here we only consider $\mathbb{P}(S_n > na)$, the deviation of order n. Cramer's large deviation theorem gives estimation for this probability

Theorem 35. ***(Cramer's Large Deviation Theorem)*** *For $S_n = X_1 + ... + X_n$ as i.i.d. sum with $\mathbb{E}X_1 = 0$ and the MGF of X_1 finite in some neighborhood of 0 and $\mathbb{P}(X > a) > 0$ for some $a > 0$, then*

$$[\mathbb{P}(S_n > na)]^{\frac{1}{n}} \to e^{-\psi(a)} \ (n \to \infty) \tag{433}$$

where $\psi(a) = -\inf_{t>0}\{\log[e^{-at}M_{X_1}(t)]\} > 0$.

Proof. Recall the Chernoff bound for the sum of *i.i.d.* random variables, it directly tells us

$$\forall \lambda > 0, \mathbb{P}(S_n > na) \leq \mathbb{E}e^{\lambda S_n}e^{-\lambda na} = [\mathbb{E}e^{\lambda X_1}]^n e^{-\lambda na} = [e^{-a\lambda}M_{X_1}(t)]^n \tag{434}$$

so minimize w.r.t. $\lambda > 0$ to get the tightest bound

$$[\mathbb{P}(S_n > na)]^{\frac{1}{n}} \leq \inf_{\lambda > 0}\left\{e^{-a\lambda}M_{X_1}(t)\right\} = e^{-\psi(a)} \tag{435}$$

so one direction is proved.

For the other direction, let's prove $\liminf_{n\to\infty}\frac{1}{n}\log\mathbb{P}\left(\frac{S_n}{n} > a\right) \geq -\psi(a)$. We prove this statement by proving a stronger statement that for any open set $G \subset \mathbb{R}$, $\liminf_{n\to\infty}\frac{1}{n}\log\mathbb{P}\left(\frac{S_n}{n} \in G\right) \geq -\inf_{x\in G}\psi(x)$. To prove this, just need to prove the statement that $\forall a \in G, \exists \delta > 0, (a-\delta, a+\delta) \subset G$ and $\liminf_{n\to\infty}\frac{1}{n}\log\mathbb{P}\left(a - \delta < \frac{S_n}{n} < a + \delta\right) \geq -\psi(a)$.

Here we make use of the Fenchel conjugate structure of $\psi(a)$ (stated in remark) that $\exists \lambda = \lambda(a) > 0$ such that $a\lambda - \Lambda(\lambda) = \psi(a), \Lambda'(\lambda) = a$. The Laplacian transform naturally leads to the change of measure again, so set

$$\frac{dF_\lambda(x)}{dF(x)} = e^{x\lambda - \Lambda(\lambda)} > 0 \tag{436}$$

with F as the CDF of X_1 and F_λ as another measure on the real line. Now $\mathbb{E}e^{X_1\lambda - \Lambda(\lambda)} = \frac{M_{X_1}(\lambda)}{M_{X_1}(\lambda)} = 1$, so F_λ is a

probability measure induced by some random variable Y_1. Notice that

$$\mathbb{E}Y_1 = \mathbb{E}X_1 e^{X_1\lambda - \Lambda(\lambda)} = \frac{d}{d\lambda}\log \mathbb{E}e^{X_1\lambda} = \Lambda'(\lambda) = a \tag{437}$$

now since we want to consider the probability $\mathbb{P}(S_n > na)$ and S_n is the sum of *i.i.d.* copies of X_1, it's natural to think about $T_n = Y_1 + ... + Y_n$ as the sum of *i.i.d.* copies of Y_1. Now we see that

$$\frac{1}{n}\log \mathbb{P}(n(a-\delta) < S_n < n(a+\delta)) \tag{438}$$

$$= \frac{1}{n}\log \int \mathbb{I}_{\sum_{i=1}^n x_i \in (n(a-\delta), n(a+\delta))}\, dF(x_1)... dF(x_n) \tag{439}$$

$$= \frac{1}{n}\log \int \mathbb{I}_{\sum_{i=1}^n x_i \in (n(a-\delta), n(a+\delta))} \prod_{i=1}^n e^{-x_i\lambda + \Lambda(\lambda)}\, dF_\lambda(x_1)... dF_\lambda(x_n) \tag{440}$$

$$= \frac{1}{n}\log \mathbb{E}\left(\mathbb{I}_{\sum_{i=1}^n Y_i \in (n(a-\delta), n(a+\delta))} \prod_{i=1}^n e^{-Y_i\lambda + \Lambda(\lambda)} \right) \tag{441}$$

$$= \frac{1}{n}\log \mathbb{E}\left(\prod_{i=1}^n e^{-Y_i\lambda + \Lambda(\lambda)} \Big| n(a-\delta) < T_n < n(a+\delta) \right) + \frac{1}{n}\log \mathbb{P}(n(a-\delta) < T_n < n(a+\delta)) \tag{442}$$

$$\geq \frac{1}{n}\log \mathbb{P}(n(a-\delta) < T_n < n(a+\delta)) + \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left(\Lambda(\lambda) - Y_i\lambda \Big| a - \delta < \frac{T_n}{n} < a + \delta \right) \tag{443}$$

by Jensen's inequality applied for conditional expectation. Note that now $Y_1, ..., Y_n$ are *i.i.d.* random variables with $\mathbb{E}Y_1 = a$, so by SLLN we have

$$\frac{T_n}{n} \overset{a.s.}{\to} a \ (n \to \infty) \tag{444}$$

from this one directly know that

$$\frac{1}{n}\log \mathbb{P}(n(a-\delta) < T_n < n(a+\delta)) \to 0 \ (n \to \infty) \tag{445}$$

we only investigate the other term to find

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left(\Lambda(\lambda) - Y_i\lambda \Big| a - \delta < \frac{T_n}{n} < a + \delta \right) \tag{446}$$

$$= \Lambda(\lambda) - \lambda\mathbb{E}\left(\frac{T_n}{n} \Big| a - \delta < \frac{T_n}{n} < a + \delta \right) \tag{447}$$

$$\geq -(-\Lambda(\lambda) + \lambda a) - \lambda\delta \geq -\psi(a) - \lambda\delta \tag{448}$$

by the Frenchel conjugate structure of $\psi(a)$. Set $\delta \to 0$ to conclude the proof. □

Remark. *The Cramer's lerge deviation bound is actually telling us that for i.i.d. sum with each of the random*

variable to have a nice enough distribution (finite MGF in neighborhood of 0), the ***Chernoff bound is actually asymptotically tight!***

If we denote $\Lambda(t) = \log M_{X_1}(t)$ *as the log-Laplacian transform, then* $\Lambda^*(a) = \sup_t \{at - \Lambda(t)\} = \psi(a)$ *is actually* ***the Frenchel conjugate of log-Laplacian transform (rate function)*** *evaluated at* a*, from the property of Frenchel conjugate, it's easy to see that* $\psi(a)$ *is convex and l.s.c. in* a *and it's also positive here.*

Remark. *One can see easily that the theorem above can be generalized to work in* $G \subset \mathbb{R}^d$*. Actually,*

$$\lim_{n\to\infty} \left[\mathbb{P}\left(\frac{S_n}{n} \in G\right)\right]^{\frac{1}{n}} = e^{-\inf_{x\in G}\psi(x)} \tag{449}$$

if $X_1, ..., X_n$ *are taking values in* $\mathbb{R}^d$*.*

The following theorem solves the

Theorem 36. ***(Chernoff's Large Deviation Theorem)*** $X_1, ..., X_n$ *are i.i.d. with* $\mathbb{E}X_1 = \mu < 0, \mathbb{P}(X_1 > 0) > 0$ *and* $\exists \delta > 0, \forall |t| < \delta, M_{X_1}(t) < \infty$*. Use the same notation* $\rho \overset{def}{=} \inf_{t>0}\{M_{X_1}(t)\} = \inf_{t\in\mathbb{R}}\{M_{X_1}(t)\}$*, then*

$$[\mathbb{P}(X_1 + ... + X_n \geq 0)]^{\frac{1}{n}} \to \rho \ (n \to \infty) \tag{450}$$

Proof. Consider $Y_i = X_i - \mu$ as *i.i.d.* random variables with $M_{Y_1}(t)$ finite in a neighborhood of 0, with $\mathbb{E}Y_1 = 0, \mathbb{P}(Y_1 > -\mu) > 0$ so by Cramer's large deviation theorem,

$$[\mathbb{P}(X_1 + ... + X_n \geq 0)]^{\frac{1}{n}} = [\mathbb{P}(Y_1 + ... + Y_n \geq -n\mu)]^{\frac{1}{n}} \to e^{\inf_{t>0}\left\{\log e^{\mu t} M_{Y_1}(t)\right\}} \tag{451}$$

where $\inf_{t>0}\{\log e^{\mu t} M_{Y_1}(t)\} = \inf_{t>0}\{\log M_{X_1}(t)\} = \log \rho$, so the theorem is proved.

One can refer to another proof of this theorem directly using the large deviation bounds for single random variable.(But one has to figure out the auxiliary random variable U_n for $T_n = X_1 + ... + X_n$ and show that $\frac{\theta_n}{n} \to 0$ where $\mathbb{P}(T_n \geq 0) = \rho_n e^{-\theta_n}$ by CLT)

□

Examples

As an example of the application, consider applying large deviation result for $S_n \sim B(n, p)$. Since $S_n \overset{d}{=} X_1 + ... + X_n$ where $X_1, ..., X_n \sim B(1, p)$ *i.i.d.*, we know that

$$\mathbb{P}(S_n \geq na) = \mathbb{P}\left(\sum_{i=1}^{n}(X_i - a) \geq 0\right) \tag{452}$$

so we set $Y_i = X_i - a$ as *i.i.d.* random variables with $\mathbb{E}Y_i = p - a$. Note that this probability is only non-trivial when $0 < a < 1$, so combined with the condition of the Chernoff large deviation bound, we can estimate the probability

for $a \in (p, 1)$. The only thing to do now is to calculate the limit

$$\rho = \inf_{t>0} \{M_{Y_1}(t)\} \tag{453}$$

notice that $M_{Y_1}(t) = e^{-at}(1-p) + e^{t(1-a)}p$, so inf is taken when $t^* = \log \frac{a(1-p)}{p(1-a)}$ so $\rho = \frac{p^a(1-p)^{1-a}}{a^a(1-a)^{1-a}}$. We write the result in the log form that

$$\forall p < a < 1, \frac{1}{n} \log \mathbb{P}(S_n \geq na) \to a \log \frac{p}{a} + (1-a) \log \frac{1-p}{1-a} \ (n \to \infty) \tag{454}$$

which has something to do with the binary entropy function.

The second example has something to do with hypothesis testing for **Bernoulli sequence model**. Now we have a lot of samples $X_1, ..., X_n$ *i.i.d.* coming from population $B(1, p)$ with p unknown. There are two hypothesis, $H_0 : p = p_0, H_1 : p = p_1$ $(p_0 < p_1)$ and now we want to do the testing with all admissible decision rules restricted to those of the form $\delta_a, p_0 < a < p_1$. $\delta_a(X_1, ..., X_n)$ rejects H_0 on observing $S_n \geq na$ and rejects H_1 otherwise. Now we want to find an a such that the test is minimax and we want to know the asymptotic error probability when $n \to \infty$.

Consider type I error $\mathbb{P}(S_n \geq na | H_0)$, by large deviation theorem, we know that

$$\mathbb{P}(S_n \geq na | H_0) \sim e^{-nK(a,p_0)} \tag{455}$$

where $K(a, p) = a \log \frac{a}{p} + (1-a) \log \frac{1-a}{1-p}$. Similarly, type II error has the asymptotic order

$$\mathbb{P}(S_n < na | H_1) \sim e^{-nK(a,p_1)} \tag{456}$$

so in order to get a minimax test, we just need to have constant risk function, so we set the probability of type I and type II error to be the same $K(a, p_0) = K(a, p_1)$ to solve out the best a

$$a(p_0, p_1) = \frac{\log \frac{1-p_0}{1-p_1}}{\log \frac{1-p_0}{1-p_1} + \log \frac{p_1}{p_0}} \tag{457}$$

under such a, the asymptotic error probability is given by

$$K(a(p_0, p_1), p) \sim e^{-n \frac{(p_1-p_0)^2}{8p(1-p)}} \ (n \to \infty) \tag{458}$$