# Notes on Elementary Stochastic Processes

based on MAT 135B

Cheng, Feng

## Discrete Markov Chains

### Introduction

Definition.

$$P(X_{n+1} = i_{n+1} \mid X_n = i_n, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0) = P(X_{n+1} = i_{n+1} \mid X_0 = i_0)$$

(We are only considering homogeneous Markov chains here, whose one-step transitional probabilities are unrelated to the values of $n$.)

Three elements: the state space $S$, the transitional probabilities represented by the stochastic matrix $P$, and the initial distribution $\alpha$.

The stochastic matrix $P$ is given by the transitional probabilities $p_{ij}$ from state $i$ to state $j$ ($i, j \in S$) in a single step. It is obvious that the row sums of $P$ is equal to 1.

There is a one-to-one correspondence between a Markov chain with initial distribution $\alpha$ and transitional stochastic matrix $P$ and a sequence $\{X_n\}_{n\geq 0}$ satisfying

$$P(X_0 = i_0, \ldots, X_n = i_n) = \alpha_{i_0} p_{i_0 i_1} \ldots p_{i_{n-1} i n}$$

We also have the extended Markov property and the strong Markov property. Both are very useful tools in the theory of Markov chains. The former states that the future event $F$ depends solely on the current state $\{X_n = i\}$, regardless of the past history $H$. The latter says that under finite stopping time $T$ (such that the event $\{T = n\}$ is determined solely in terms of $X_0, X_1, \ldots, X_n$), the sequence of events $\mathbf{X}'$ starting from time $T$ is a new Markov chain with the same transitional probabilities and is unrelated to the history before $T$.

The fact that $n$-step transition probabilities can simply be represented by the $n$-th power of the transition matrix $P$ is due to the Chapman-Kolmogorov equations:

$$p_{ij}^{m+n} = \sum_{k \in S} p_{ik}^m p_{kj}^n,$$

which break down the calculation of $n$-step transition probabilities.

It is customary to use a row vector to represent the initial distribution $\alpha$. We may then calculate the distribution at time $n$ directly by $\alpha P^n$, a row vector as well.

## Classification of States

We say $j$ is *accessible* from $i$ (denoted by $i \to j$) if $p_{ij}^n > 0$ for some $n \geq 0$, i.e., state $j$ can be visited from state $i$ in a finite number of steps. $i \leftrightarrow j$ means state $i$ and state $j$ are mutually accessible, that the two states *communicate*.

It is not hard to show that $\leftrightarrow$ is an equivalence relation, which implies a Markov chain can be reduced to *communicating classes*, all of which contain states that communicate with one another. These individual classes have many useful properties. An *irreducible* Markov chain has itself as the only communicating class.

Here we introduce the notion of recurrence and transience. A state $i$ is *recurrent* if it is certain that $i$ will return to $i$ after a finite number of steps. A state is *transient* if it is not recurrent.

If we define the first-passage time of state $i$ as

$$T_i = \inf\{n \geq 1 : X_n = i\},$$

then the state $i$ is recurrent if $P(T_i < \infty) = 1$, and it is transient if $P(T_i = \infty) > 0$. (Note that if the set $\{n \geq 1 : X_n = i\}$ is empty, then we let $T_i = \infty$. Random variables similar to $T_i$ in stochastic processes are customarily defined to also include $\infty$.)

Alternatively, if we simply define $f_i$ to be the return probability of state $i$, $f_i = 1$ means $i$ is recurrent, and $f_i < 1$ means $i$ is transient.

Writing the expectation of $V_i$, the number of returns to $i$, in terms of indicators with respect to $\{V_i \geq k\}$ (by the strong Markov property) and $\{X_n = i\}$, we can show that

$$\sum_{n=1}^{\infty} p_{ii}^n = \infty$$

if $i$ is recurrent, and

$$\sum_{n=1}^{\infty} p_{ii}^n < \infty$$

if $i$ is transient.

*Remark.* $V_i + 1$ here actually follows Geometric$(1 - f_i)$ for $f_i \neq 1$. Furthermore, we can use $P(V_i \geq k) = f_i^k$ for any $k \in \mathbf{Z}^+$ (which we get from the calculation above) to show that $P(V_i = \infty) = 1$ for any recurrent $i$, and $P(V_i < \infty) = 1$ for any transient $i$.

In summary, recurrence/transience can be characterized in terms of the first-passage time $T_i$, the sum of transition probabilities $\sum_n p_{ii}^n$, or the number of returns

$$V_i = |\{n \geq 1 : X_n = i\}|.$$

It should be noted that finite first-passage time $T_i$ does not imply $E(T_i)$ is finite as well, though the contrary is true by Markov's inequality. This tells us we need a stronger kind of recurrence. The

*mean recurrence time* is the expected first-passage time

$$m_i = E(T_i) = \sum_{n=1}^{\infty} nP(T_i = n).$$

When $i$ is transient, namely $P(T_i = \infty) > 0$, $E(T_i)$ diverges to $\infty$. On the other hand, when $i$ is recurrent, namely $P(T_i < \infty) = 1$, if $E(T_i) < \infty$, $i$ is *positive recurrent*, while if $E(T_i) = \infty$, $i$ is *null recurrent*.

As it turns out, both recurrence and positive recurrence are class properties, i.e., all states in a communicating class should be simultaneously (not) recurrent and simultaneously (not) positive recurrent. To show recurrence/transience as a class property, one uses the Chapman-Kolmogorov equations and $\sum_n p_{ii}^n = \infty$ and $< \infty$. The most direct way of showing positive recurrence as a class property involves invariant distribution, which we will mention later.

The *period* $d_i$ of a state $i$ is given by the gcd of all possible steps needed to return to $i$. Formally, we write $d_i = \gcd\{n : p_{ii}^n > 0\}$. The state $i$ is *aperiodic* if $d_i = 1$, and is *periodic* if $d_i > 1$. Using the Chapman-Kolmogorov equations, it is not hard to show that period is also a class property.

Here we provide a summary of some elementary facts useful in judging the properties of states and classes in a Markov chain:

A subset $S_0$ of the state space $S$ is *closed* if $p_{ij} = 0$ for any $i \in S$ and state $j$ not in $S$. Using ordinary language, this means $S_0$ can only be accessible within itself.

- Consider a closed subset $S_0$ with finitely many states. Starting from an arbitrary state in $S_0$, we can write all the visits into a sequence of random variables. There must exist a state that is visited an infinite number of times. This implies every such $S_0$ has at least one recurrent state.
- If we know some state $j$ is accessible from a recurrent state $i$, in reverse $i$ will also be accessible from $j$ because the probability of never visiting $j$ from $i$ ($p_{ij}^{\infty}$) is 0. Thus, the two states communicate. It follows that the set of all states accessible from any recurrent state $i$ of a Markov chain forms its recurrent class that we have discussed above.
- Therefore, any recurrent class $C$ is a closed subset of states since any state in $C$ cannot have access to, or communicate with, any state outside $C$. In particular, in reverse, if a communicating class is closed and **finite-state**, then it contains a recurrent state. Thus, this class is recurrent. (If the class has infinitely many states, then this does not necessarily hold. The asymmetric random walk on $\mathbf{Z}$ has $\mathbf{Z}$ as the only class, which is actually transient.)

Thus, in a finite-state communicating class, there is an iff relationship between the class being recurrent and the class being closed. In the general context of finite-state Markov chains, we may now determine recurrence/transience solely based on whether the class is closed or not. This is indeed very useful.

Note that all these facts concerning a single class or a closed subset of states could be directly applied to irreducible Markov chains, quite obviously.

### Limiting Probabilities and Invariant Distribution

Many of the theorems that occur in this section assumes irreducibility and positive recurrence of the chain. Before looking at the theorems (ideas of which are simple but the proofs are complicated), remember that a finite irreducible Markov chain is necessarily positive recurrent (we skip the proof). This is oftentimes the problem setting we will encounter.

First, we define the *invariant distribution* (or *stationary distribution*). The invariant distribution is invariant under the passage of time; formally speaking, if we take the convention of writing distributions into a row vector $\pi$ over the state space $S$ with sum 1, the invariant distribution needs to satisfy $\pi P = \pi$. ($\pi$ here is the left row eigenvector for eigenvalue 1 of $P$.)

The upcoming theorems all focus on irreducible Markov chains. Now we introduce the fundamental theorem (the proof of which is completely skipped because of its technicality). In an irreducible chain, an invariant distribution $\pi$ exists iff all states are positive recurrent iff we know some state is positive recurrent, and the $\pi$ is **uniquely** given by

$$\pi_i = \frac{1}{m_i},$$

the reciprocal of the mean recurrence time of state $i$. This tells us that the invariant probability times the mean recurrence time at state $i$ should be 1. Also, irreducible Markov chains allow either none or only one invariant distribution.

This implies that as long as we have found one invariant distribution $\pi$ [by the Markov chain being symmetric or doubly stochastic (which we will cover soon) over all states, e.g., without referring to $m_i$ or even the formula $\pi P = \pi$], it is the only invariant distribution desired.

It follows from our intuition that the invariant probability $\pi_i$ might as well be interpreted as the proportion of time spent at this state $i$. It turns out that this is true in the limit; we can prove formally via the weak law of large numbers that under the same assumption of irreducibility and positive recurrence,

$$\frac{V_i(n)}{n} \xrightarrow{P} \frac{1}{m_i} = \pi_i \quad \text{as} \quad n \to \infty,$$

where $V_i(n)$ represents the number of times state $i$ is visited before time $n$, regardless of the initial distribution of $X_0$.

If we add the further condition of aperiodicity, then for any $i, j \in S$,

$$\lim_{n\to\infty} p_{ij}^n = \pi_j.$$

This is the so-called convergence to equilibrium theorem, which implies that the transition probability from any **arbitrary** state $i$ to a fixed $j$, in the long run, will become the invariant distribution of the state $j$. The proof of this strong theorem requires the *coupling* technique. (Note in particular that the previous theorem is concerned with the convergence of the random variable $\frac{V_i(n)}{n}$ in probability,

while this theorem is the convergence of a sequence of numbers $p_{ij}^n$.)

If we extend this theorem to Markov chains in general with period $d$, we need to modify our conclusion as follows: for any $i, j \in S$, there exists a particular remainder $0 \leq r \leq d-1$ such that

$$\lim_{m\to\infty} p_{ij}^{md+r} = d\pi_j = \frac{d}{m_j}.$$

In addition, for the $n$'s that are not congruent to $r$ modulo $d$, $p_{ij}^n = 0$. The reason behind this is that $i$ can only reach $j$ in congruent numbers of steps modulo $d$. The vague intuition behind this formula is that since $j$ may only be visited once in $d$ steps, the probability of visiting state $j$ among every $d$ steps is given $d$ times the "weight" of $j$,

$$\pi_j = \frac{1}{m_j}.$$

The actual proof of this is quite involved.

Another point we want to address here is the invariant distribution for a doubly stochastic Markov chain. A Markov chain is *doubly stochastic* if each column sum of the transition matrix is also 1, i.e., for all $j \in S$, $\sum_{i\in S} p_{ij} = 1$.

As a matter of fact, if an irreducible Markov chain is doubly stochastic over a finite state space $S$, it follows that the unique invariant distribution is uniform over all states:

$$\pi = \frac{1}{|S|}\begin{bmatrix}1 & \cdots & 1\end{bmatrix}.$$

To show this, note that $\begin{bmatrix}1 & \cdots & 1\end{bmatrix} P = \begin{bmatrix}1 & \cdots & 1\end{bmatrix}$ because the column sums of $P$ are all 1. We then just have to "normalize" the row vector to row sum 1.

Informally speaking, a Markov chain is reversible if the sequence of random variables $X_0, X_1, \ldots$ running under the invariant distribution $\pi$, when looking backward from any time $t$, is what might occur in the original forward chain with the same probability. To describe this formally, the new sequence of random variables $\mathbf{Y}$ should share the same p.m.f. with the original chain $\mathbf{X}$:

$$\{Y_0 \coloneqq X_t = i_0, Y_1 \coloneqq X_{t-1} = i_1, \ldots, Y_t \coloneqq X_0 = i_t\} \stackrel{D}{=} \{X_0 = i_0, X_1 = i_1, \ldots, X_t = i_t\}.$$

As a result, we cannot tell whether the reversible chain is running forward or backward by looking at the sequence of $t$ random variables.

If the forward chain has its invariant distribution $\pi$ as the initial distribution, it is not hard to show that the backward sequence $Y_0, Y_1, \ldots, Y_t$ is a Markov chain with initial distribution $\pi$ as well, and with backward transition probabilities given by

$$P(Y_{m+1} = X_k = i \,|\, Y_m = X_{k+1} = j) = \frac{\pi_i p_{ij}}{\pi_j}.$$

The fact that the backward sequence is also Markov rephrases reversibility to saying that the

forward and backward chains share the same transition probabilities. Here comes the actual definition:

Let **X** be an **irreducible** Markov chain with $X_0$ having the invariant distribution $\pi$. If its time-reversal **Y** has the same transition matrix $P$ as **X** does, or simply

$$\pi_i p_{ij} = \pi_j p_{ji} \text{ for all } i, j \in S,$$

then we call **X** *reversible*. (The equations above are known as the *detailed balance equations*.)

*Remark.* Reversibility is by assumption restricted to irreducible positive recurrent Markov chains. This is because we only want to determine the reversibility of a chain based on *only* one invariant distribution. Also, the detailed balance equations allows only one solution.

Next, we show that in an irreducible chain, if a distribution $\lambda$ satisfies the detailed balance equations $\lambda_i p_{ij} = \lambda_j p_{ji}$, then $\lambda$ is the unique invariant distribution:

$$\sum_{i \in S} \lambda_i p_{ij} = \sum_{i \in S} \lambda_j p_{ji} = \lambda_j \sum_{i \in S} p_{ji} = \lambda_j \implies \lambda P = \lambda.$$

This property gives us an easy criterion for finding the invariant distribution. A direct application of this appears in the *random walk on weighted graphs*. Given the probability from $i$ to $j$

$$p_{ij} = \frac{w_{ij}}{\sum_k w_{ik}}$$

based on the proportion of weight, let

$$\lambda_i = \frac{\sum_k w_{ik}}{\sum_{i,k} w_{ik}}$$

be the proportion of the the weight out from $i \in S$ over the total weight. We now have

$$\lambda_i p_{ij} = \frac{w_{ij}}{\sum_{i,k} w_{ik}} = \lambda_j p_{ji}.$$

When the weighted graph is connected, it is irreducible and thus $\lambda$ is our desired invariant distribution of the chain. The weighted graph can have self-edges because $p_{ii}$ does not affect the detailed balance equations.

If we simply take all the edge weights to be 1, it follows that $\pi_i = \frac{\deg(v_i)}{2|E|}$.

## Branching Processes

The branching process is a stochastic process that emulates the spread of a family name. There is only a single individual at time 0, represented by $X_0 = 1$. Every individual $W_j$ in any generation has an i.i.d. *offspring distribution*, given by the p.m.f.

$$p_i \coloneqq P(\text{number of offspring} = i), i = 0, 1, 2, \ldots.$$

The offspring from the individual $W$ will appear in the next generation. The trivial case $p_i = 1$ for some $i$ is mostly ignored.

It should be noted that a branching process is discrete and non-negative integer-valued. This suggests that we should use the method of probability generating function to investigate the properties of branching processes. Every $W_j$ in generation $n-1$ is i.i.d. with p.g.f. $G(s)$, and thus we may apply the p.g.f. random sum formula to get

$$G_n(s) = G_{n-1}(G(s)),$$

where the subscript tells the generation the p.g.f. is about. Since $G_0(s) = s$, $G_n$ becomes the $n$-th iterate of $G$ for all non-negative integers $n$.

It follows that $E(X_n) = G_n'(1) = G_{n-1}'(G(1))G'(1) = G_{n-1}'(1)E(W) = E(X_{n-1})\mu$, which gives us the formula $E(X_n) = \mu^n$, if we denote $E(W)$ by $\mu$.

This indicates that when $\mu < 1$, $E(X_n) \to 0$; when $\mu = 1$, $E(X_n) \to 1$; and when $\mu > 1$; $E(X_n) \to \infty$ as $n \to \infty$.

Define $e_n = P(\{X_n = 0\})$, the extinction probability at $n$-th generation. Since $\{X_n = 0\}$ is an increasing sequence of events, the ultimate extinction probability

$$e = P(\cup_{n=1}^{\infty}\{X_n = 0\}) = \lim_{n\to\infty} e_n.$$

This extinction probability $e$ is actually the smallest non-negative solution to $G(x) = x$. To prove this, first note $G_n(0) = P(X_n = 0) = e_n$ and recall $G_n(s) = G(G_{n-1}(s))$ because $G_n(s)$ is the $n$-th iterate of $G$. Thus,

$$e_n = G(e_{n-1}) \text{ for all } n \in \mathbf{N}.$$

Take $n \to \infty$ on both sides, we have $e = \lim_{n\to\infty} G(e_n)$. $G(x)$ is uniformly convergent on $[0, 1]$ and is thus continuous on $[0, 1]$. Therefore, since all $e_n \in [0, 1]$, by the preservation of sequential limit under continuity, $e = \lim_{n\to\infty} G(e_n) = G(e)$.

Beyond the fact that $e$ is a solution to $G(x) = x$, we want to show $e$ is the smallest non-negative solution. Suppose $t$ is any non-negative solution to $G(x) = x$. Since the coefficients of $G(x)$ are

non-negative, $G$ is a non-decreasing function, we have the following iteration:

$$e_1 = G(0) \leq G(t) = t, e_2 = G(1) \leq G(t) = t, \ldots, e_n = G(e_{n-1}) \leq G(t) = t \ldots.$$

We now have $e_n$ is bounded above by $t$ for all $n$, and thus the limit $e \leq t$.

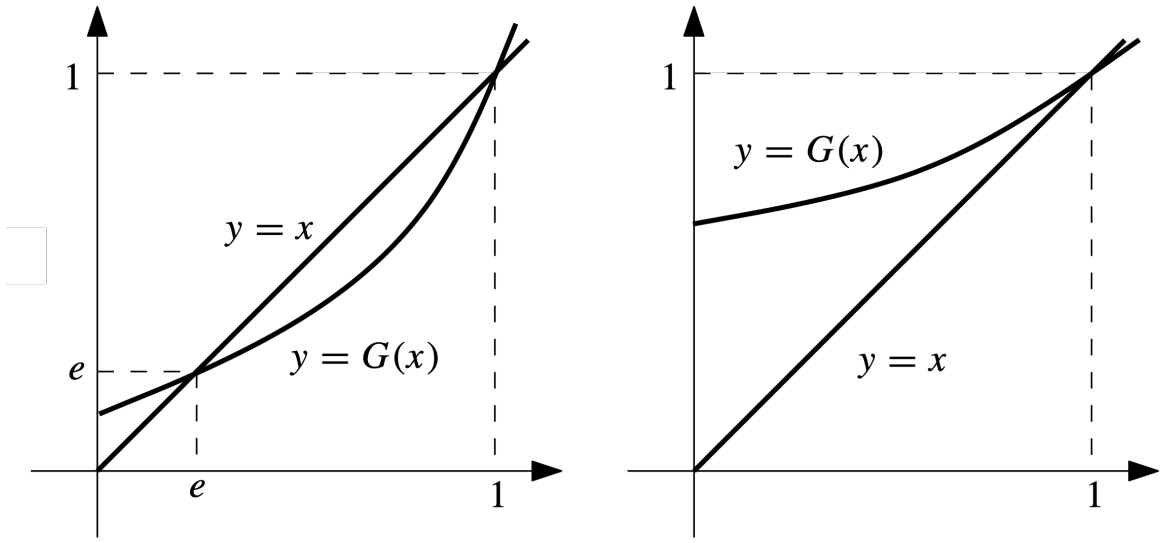
Furthermore, the following theorem will tell us that extinction necessarily occurs ($e = 1$) iff $\mu \leq 1$ (if we exclude the case $p_1 = 1$, which let $G(x) = x$ for all $x$). Intuitively the mean offspring $\mu = 1$ is the most likely threshold of extinction. To show this, first of all restricting to $[0, 1]$, we have $G$ being continuous, non-decreasing, and concave-up (by considering the 1st and 2nd derivative).

The easiest way to "see" the result is to look at the two cases

$$\mu = G'(1) > 1 \quad \text{and} \quad G'(1) \leq 1$$

graphically. Remember that $G(x) = x$ always has solution 1.

Case I on the left allows another solution in $[0, 1)$, and Case II allows 1 as the only solution in $[0, 1]$. See the screenshot below from Grimmett and Welsh.




Thus, $\mu \leq 1$ iff $e = 1$, i.e., the branching process ultimately goes extinct. (If we want to prove this more rigorously, we might need to show that a concave-up function intersects a linear function with slope 1 at most twice, with the first intersection having $G^{'}(x) < 1$ and the second having $G^{'}(x) > 1$.)

## Poisson Processes

A *counting process* $N(t)$ for $t \geq 0$ is a continuous stochastic process such that

- $N(t)$ takes in non-negative integers for each $t$;
- $N(t)$ is non-decreasing;
- $N(t)$ is right-continuous, i.e., $N(t) - N(s)$ represents the increment of $N(t)$ in the interval $(s, t]$.

A *Poisson process with arrival rate* $\lambda$ is the particular counting process that "follows" the Poisson distribution:

- $N(0) = 0$;
- $(s_1, t_1] \cap (s_2, t_2] = \emptyset$, then $N(t_1) - N(s_1)$ and $N(t_2) - N(s_2)$ are independent random variables. (independent increments)
- The number of events in any interval of length $t$ follows Poisson$(\lambda t)$. (stationary increments dependent solely on $t$)

Condition 2 and 3 is explicitly

$$P(N(t+h) - N(t) = k) = e^{-\lambda h} \frac{(\lambda h)^k}{k!}$$

for arbitrary $t, h$, and $k \geq 0$.

Similar to what we have done in the theory of Markov chains, different characterizations of stochastic processes may lead to interesting interpretations and properties. As we will soon find out, Poisson processes establish the bridge between the discrete Poisson distribution and the continuous exponential distribution, whose p.m.f. and p.d.f. typically lacked intuition when they were first introduced in a probability class. (As a matter of fact, the exponential distribution is the only continuous distribution with the *lack-of-memory* property, and thus serves as the basis for continuous-time stochastic processes with independent increments.)

Regarding the 2nd and 3rd condition, we have the equivalent infinitesimal definition that, for any $t \geq 0$, $N(t)$ follows the equation that for very small positive $h$,

$$P(N(t+h) - N(t) = 1) = \lambda h + o(h) \quad \text{and} \quad P(N(t+h) - N(t) = 0) = 1 - \lambda h + o(h).$$

To illustrate why the two definitions are the same, recall how the Poisson distribution may be interpreted as infinite coin flips. Thus, the coin with success probability $\lambda h$ to increase $N$ by 1 in all intervals with very small length $h$ is what approximates the Poisson process. Note that when $h \to 0$, the $o(h)$ in the two expressions above will vanish. The rigorous proof of the equivalence between the two definitions requires differential equations and is omitted here.

Both theorems below are easily proven following the infinitesimal definition.

The superposition of two Poisson processes $N_1(t)$ and $N_2(t)$ with rate $\lambda_1$ and $\lambda_2$ by letting $N(t) = N_1(t) + N_2(t)$ is a new Poisson process with rate $\lambda_1 + \lambda_2$. If events in a Poisson process with rate $\lambda$ has an independent probability $p$ of being type 1, then the counting process of type 1 is a Poisson process with rate $\lambda p$, and the counting process of the remaining events is an *independent* Poisson with rate $\lambda(1-p)$.

Equally important is the characterization using arrival and interarrival times. We define the $n$-th arrival time $S_n$ as $\inf\{t \geq 0 : N(t) = n\}$.

It turns out that interarrival times $T_1 \coloneqq S_1 - S_0, T_2 \coloneqq S_2 - S_1, \ldots$ between two adjacent events in a Poisson process are all independent and follow Exponential($\lambda$). To show this, first we have

$$P(T_1 > t) = P(N(t) = 0) = e^{-\lambda t},$$

which tells us that $T_1 \sim \text{Exponential}(\lambda)$. We can proceed to conclude about $T_2, T_3, \ldots$ by induction, following

$$P(T_{n+1} > t \,|\, T_n = t_n, \ldots, T_1 = t_1) = P(0 \text{ arrival in } (s_n, s_n + t]) = e^{-\lambda t},$$

since the interval $(s_n, s_n + t]$ is of length $t$. (Here $n \geq 1$ and $s_n = t_1 + t_2 + \cdots + t_n$, given arbitrary $t_i$'s.)

In general, we can show that the Poisson process is a continuous-time Markov chain, and conditioning on any stopping time $T < \infty$ (e.g. the $n$-th arrival time $S_n = \sum_{i=1}^{n} T_i$ above), $N(T + t) - N(T)$ is a new Poisson process with the same rate $\lambda$, independent of $N$ prior to time $T$. The interarrival result above also follows "directly" from this strong Markov property because all $t_i$'s are arbitrary, as we have mentioned above.

It is now clear that $E(S_n) = \sum_{i=1}^{n} E(T_i) = n/\lambda$. Furthermore, since $S_n$ is the sum of $n$ i.i.d. Exponential($\lambda$) random variables, it follows the *Gamma distribution* with density

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

We may get this expression above by differentiating both sides of

$$P(S_n > t) = P(N(t) < n) = \sum_{j=0}^{n-1} e^{-\lambda t} \frac{(\lambda t)^j}{j!}.$$

The last result we discuss here is the relationship between the uniform distribution and the Poisson process. As it turns out, conditioning on the event $N(t) = n$, we have that the $n$ arrival times $S_1, S_2, \ldots, S_n$ in the interval $[0, t]$ are distributed as the order statistics of $n$ independent Uniform($[0, t]$) random variables.

We briefly sketch why this is true. The key to this is that

$$\begin{aligned} f(s_1, s_2, \dots s_n \mid N(t) = n) &= \frac{f(s_1, s_2, \dots s_n, n)}{P(N(t) = n)} \\ &= \frac{f(s_1) f(s_2 - s_1) \dots f(s_n - s_{n-1}) P(s_{n+1} > t)}{P(N(t) = n)} \\ &= \frac{\lambda e^{-\lambda s_1} \lambda e^{-\lambda(s_2 - s_1)} \dots \lambda e^{-\lambda(s_n - s_{n-1})} e^{-\lambda(t - s_n)}}{e^{-\lambda t} (\lambda t)^n / n!} = \frac{n!}{t^n}, \end{aligned}$$

the exact joint density of the order statistics corresponding to $n$ independent Uniform$([0, t])$ random variables.