# Applications to Queueing Theory

Introduction to Stochastic Processes (Erhan Cinlar)
Ch. 6.5, 6.6



## Applications to Queueing Theory: M/G/1 Queue

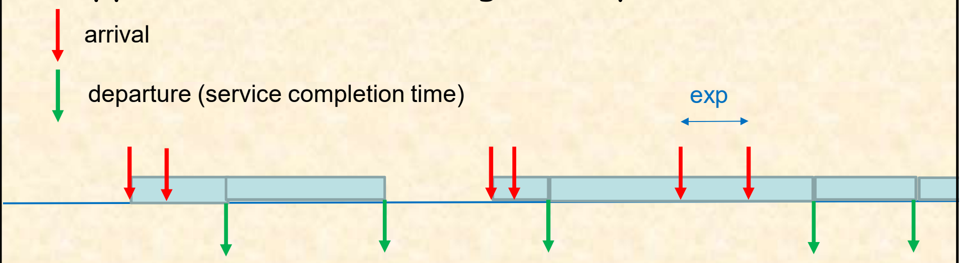
M/G/1:

Arrival Process: **M**emoryless (Poisson arrival or exponential (geometric) interarrivals

Service Process: **G**enerally-distributed service times

Number of servers: **1**

## Applications to Queueing Theory: M/G/1 Queue




X(t): Number of customers in the system (queue and under service)

Consider a specific subset of times $\{ t_e \}$ only. That means that we embed X(t) on times $\{ t_e \}$ . We do not look at X(t) at times other than in $\{ t_e \}$.

$X(t_e)$ is the process X(t) embedded at times $\{ t_e \}$.

X(t) is not a MC. Why?

If $\{ t_e \}$ = { times of customer departure }, then $X(t_e)$ is a MC. Why?



## Applications to Queueing Theory: M/G/1 Queue

$N_t(\omega)$ : number of arrivals during the time interval $[0,t]$.

$Z_1(\omega), Z_2(\omega),...$: service times of customers who depart first, second, ...

$Y_t(\omega)$ : number of customers in the system (waiting or being served at time $t$ )

Assumptions:

- ♣ $N = \{N_t; t \geq 0\} \sim P(a)$
- ♣ $Z_1, Z_2, ...$ i.i.d. $\sim \phi$

- Consider the future of $Y$ from a time $T$ of a departure onward.
- Define $X_n$ as the number of customers in the system just after the instant of the $n^{\text{th}}$ departure. ( $X_n$ is a SP embedded at departure times)

**Theorem:** $X$ is a MC with the transition matrix

Distribution of arrivals over a service time

$$P = \begin{pmatrix} q_0 & q_1 & q_2 & q_3 & \cdots \\ q_0 & q_1 & q_2 & q_3 & \cdots \\ & q_0 & q_1 & q_2 & \cdots \\ & & q_0 & q_1 & \cdots \\ & & & q_0 & \cdots \\ & & & & \ddots \end{pmatrix}, \quad q_k = \int_0^\infty \frac{e^{-at}(at)^k}{k!} d\phi(t), \quad k = 0,1,...$$

Proof: We need to show $P\{X_{n+1}=j \mid X_0,\dots,X_n\} = P\{X_{n+1}=j \mid X_n\}$

$$P\{X_{n+1}=j \mid X_n=i\} = \begin{cases} q_j & i=0, j\geq 0 \\ q_{j+1-i} & i>0, j\geq i-1 \\ 0 & \text{otherwise} \end{cases}$$

- Let $T$ the time of the $n^{\text{th}}$ departure.
- Let $Z=Z_{n+1}$ the service time of the $n+1$ customer.

Then, $X_{n+1} = \begin{cases} X_n + (N_{T+Z}-N_T)-1, & X_n>0 \\ N_{S+Z}-N_S, & X_n=0 \end{cases}$ (S:arrival time of the n+1 customer )

Using Poisson properties: $P\{N_{T+Z}-N_T=k \mid X_0,\dots X_n;T\} = P\{N_Z=k\}$

Distribution of arrivals over a service time

$$q_k = P\{N_Z=k\} = E\left[P\{N_Z=k \mid Z\}\right] = E\left[\frac{e^{-aZ}(aZ)^k}{k!}\right] = \int_0^\infty \frac{e^{-at}(at)^k}{k!} d\phi(t)$$

- $i=0$ $\quad P\{X_{n+1}=j \mid X_n=0\} = P\{N_{S+Z}-N_S=j\} = P\{N_Z=j\} = q_j$
- $i>0$ $\quad P\{X_{n+1}=j \mid X_n=i\} = P\{N_{T+Z}-N_T=j+1-i\}$

$$= P\{N_Z=j+1-i\} = \begin{cases} q_{j+1-i}, & j\geq i-1 \\ 0, & j<i-1 \end{cases}$$



## Applications to Queueing Theory: M/G/1 Queue

The MC $X$ is irreducible and aperiodic. If

$$r = E[N_Z] = aE[Z] = ab$$

Mean number of arrivals over a mean service time

Then (intuitively based on queue evolution/growth, also rigorously proven)

- If $r>1$ all states are transient
- If $r<1$ all states are recurrent non-null.
- If $r=1$ all states are recurrent null

Notation:

$r_k = 1-q_0-\cdots-q_k$ (prob arrivals over a service time exceed $k$; summing them we get $r$, next)

$r = r_0 + r_1 + \cdots = (q_1+q_2+q_3+\cdots)+(q_2+q_3+\cdots)+(q_3+\cdots)+\cdots$

$= q_1 + 2q_2 + 3q_3 + \cdots$ (this is the definition of the mean $r$ of the distr of arrivals over a service time)

**Proposition:** The chain $X$ is recurrent non-null aperiodic if and only if $r<1$.

Proof: We need to show that

$$\pi = \pi \cdot P, \qquad \pi \cdot 1 = 1$$

$$\left.\begin{aligned} \pi_0 &= \pi_0 q_0 + \pi_1 q_0 \\ \pi_1 &= \pi_0 q_1 + \pi_1 q_1 + \pi_2 q_0 \\ \pi_2 &= \pi_0 q_2 + \pi_1 q_2 + \pi_2 q_1 + \pi_3 q_0 \\ &\vdots \end{aligned}\right\} \Rightarrow \begin{aligned} \pi_1 q_0 &= \pi_0 r_0 \\ \pi_2 q_0 &= \pi_0 r_1 + \pi_1 r_1 \\ \pi_3 q_0 &= \pi_0 r_2 \pi_1 r_2 + \pi_2 r_1 \\ &\vdots \end{aligned}$$

## Applications to Queueing Theory: M/G/1 Queue

Summing all equations ($q_0 = 1 - r_0$, $r = r_0 + r_1 + r_2 + \cdots$)

$$(1-r_0)\cdot\sum_{j=1}^{\infty}\pi_j = \pi_0 r + (r - r_0)\sum_{j=1}^{\infty}\pi_j$$

If $r < 1$, then we obtain $\sum_{j=1}^{\infty}\pi_j = \frac{r}{1-r}\pi_0 \Rightarrow \sum_{j=0}^{\infty}\pi_j = \frac{1}{1-r}\pi_0$

The condition $\pi \cdot 1 = 1$ is satisfied with $\pi_0 = 1 - r$

**Theorem:** The limits $\pi(j) = \lim_{n\to\infty} P^n(i,j)$ exist $\forall j \in E$ and are independent of the initial state $i$.

- If $r \geq 1$, then $\pi(j) = 0, \quad \forall j$.
- If $r < 1$, then

$$\begin{aligned}\pi(0) &= 1-r \\ \pi(1) &= (1-r)\frac{r_0}{q_0} \\ &\vdots \\ \pi(j+1) &= (1-r)\sum_{k=1}^{j}\left(\frac{1}{q_0}\right)^{k+1}\sum_{\mathbf{a}\in S_{jk}} r_{a_1} r_{a_2}\cdots r_{a_k}\end{aligned}$$

where $S_{jk}$ is the set of all $k$-tuples $\mathbf{a} = (a_1,\ldots,a_k)$ of integers $a_i \geq 1$ with $a_1 + \cdots a_k = j$



## Applications to Queueing Theory: M/G/1 Queue

<u>More on the recurrent non-null case</u>

Having the limiting distributions, we can compute $E[X_n]$, $Var(X_n)$ etc., in the limit $n \to \infty$.

Instead, we could also proceed as follows, without using the limiting distributions:

$$X_{n+1} = X_n + M_n - U_n$$

where

$$U_n = 1 - 1_0(X_n)$$

$M_n$ is the number of arrivals during the $n+1$ th service.

$$\lim_{n\to\infty} E[U_n] = 1 - \lim_{n\to\infty} E[1_0(X_n)] = 1 - \lim_{n\to\infty} P\{X_n = 0\} = 1 - \pi(0) = r = a\cdot b$$

$$E[M_n] = \mathbf{r{=}a\cdot b}$$

$$E[M_n^2] = E\left[E[N_Z^2 \mid Z]\right] = E\left[aZ + a^2 Z^2\right] = a\cdot b + a^2 c^2$$

$$c^2 = E[Z^2] = \int_0^{\infty} t^2 d\phi(t)$$

V(X)=σ²=E(X-E(X))²=E(X²)-E(X)² => E(X²) = E(X)² +V(X)

## Applications to Queueing Theory: M/G/1 Queue

$$X_{n+1}^2 = X_n^2 + M_n^2 + U_n^2 + 2X_nM_n - 2X_nU_n - 2M_nU_n$$

But

- $U_n^2 = U_n$ ($U_n$ takes values 1, 0)
- $X_nU_n = X_n$ (If $X_n > 0$, then $U_n = 1$, else if $X_n = 0$, then $U_n = 0$)

so that,

$$X_{n+1}^2 = X_n^2 + M_n^2 + U_n + 2X_nM_n - 2X_n - 2M_nU_n$$

Taking expectations of both sides we obtain

$$E[X_{n+1}^2] = E[X_n^2] + E[M_n^2] + E[U_n] + 2E[X_n]E[M_n] - 2E[X_n] - 2E[M_n]E[U_n]$$

and by letting $n \to \infty$

$$0 = ab + a^2c^2 + ab + 2qab - 2q - 2a^2b^2$$

where $$q = \lim_{n\to\infty} E[X_n] = ab + \frac{a^2c^2}{(2-2ab)}$$

Knowing the statistics of $X_n$ we can find the statistics of $V_n$, ($W_n$), as $n \to \infty$

$$V_n = W_n + Z_n$$

$V_n$ ($W_n$) is the total (waiting) time spent in the system by the $n^{th}$ customer.

($X_n$ is equal to arrivals between time of arrival and time of departure of the $n^{th}$ customer, under FIFO)



<u>What if $r \geq 1$?</u>

Consider $f_k(j)$ the probability starting from state $k+j$, the MC $X$ never enters in the set $\{0,1,...,k\}$

$f_k(j)$ is the maximal solution of the system $h = Q \cdot h, \quad 0 \leq h \leq 1$

where $Q$ is the matrix obtained from $P$ by deleting all rows and columns corresponding to the states $\{0,1,...,k\}$.

$$Q = \begin{pmatrix} q_1 & q_2 & q_3 & \cdots \\ q_0 & q_1 & q_2 & \cdots \\ & q_0 & q_1 & \cdots \\ & & & \ddots \end{pmatrix}$$

$Q$ does not depend on $k$, therefore $f_k(j) = f_0(j)$ for all $j,k$.

**<u>Lemma:</u>** The probability that $X$ never enters $\{0,1,...,k\}$ starting from $k+j$ is the same as the probability $f(j)$ that $X$ never enters $0$ starting from $j$.

**<u>Theorem:</u>** Let $f(j)$ be the probability that the queue, starting with $j$ customers never becomes empty. Then,

$$f(j) = 1 - \beta^j, \qquad j = 1,2,...$$

where $\beta$ is the smallest number in [0,1] satisfying $\beta = q_0 + q_1\beta + q_2\beta^2 + \cdots$

The $\beta$ is strictly less than one if and only if the traffic intensity $r > 1$. Therefore, $X$ is transient if and only if $r > 1$.

# Applications to Queueing Theory: G/M/1 Queue

G/M/1:

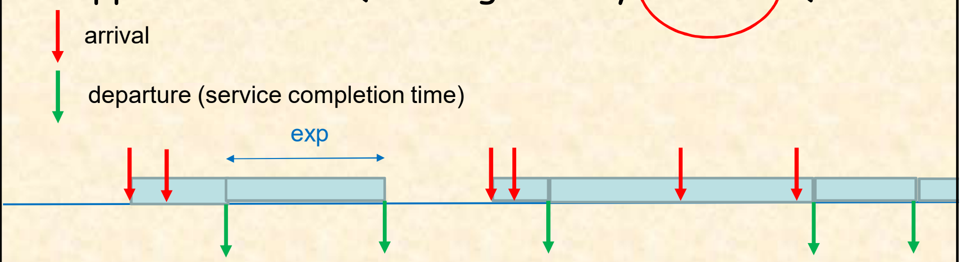
Arrival Process: **G**enerally-distributed arrival times

Service Process: **M**emoryless (exponential (geometric) service times)

Number of servers: **1**



# Applications to Queueing Theory: G/M/1 Queue




X(t): Number of customers in the system (queue and under service)

Consider a specific subset of times { $t_e$ } only. That means that we embed X(t) on times { $t_e$ } . We do not look at X(t) an times other than in { $t_e$ }.

$X(t_e)$ is the process X(t) embedded at times { $t_e$ }.

X(t) is not a MC.

If { $t_e$ } = { times of customer **arrival** }, then $X(t_e)$ is a MC.

## Applications to Queuing Theory: G/M/1 Queue

Exponentially distributed service times  exp($a$)
i.i.d. interarrival times  $\phi$.

In this case $q_n = \int_0^\infty \frac{e^{-at}(at)^n}{n!} d\phi(t)$ ← Distribution of services over an interarrival time

is the probability that the server completes exactly $n$ services during an interarrival time (provided that there are that many customers).

Define: $r_n = q_{n+1} + q_{n+2} + \cdots$

$$r = \sum_{n=1}^{\infty} nq_n = r_0 + r_1 + r_2 + \cdots$$

$r$ is the expected number of services which the server is capable of completing during an iterarrival time. It can be proved that

- $r \geq 1$ Server can keep up with arrivals (recurrent)
- $r < 1$ Queue size increases to infinity (transient)

If $X_n^\varepsilon$ is the number of customers present in the system just before the time $T_n$ of the $n^{\text{th}}$ arrival, then

**Theorem:** $X^\varepsilon = \{X_n^\varepsilon; n \in N\}$ is a MC with $E = \{0,1,2,\ldots\}$, $P^\varepsilon = \begin{pmatrix} r_0 & q_0 & & & \\ r_1 & q_1 & q_0 & & \\ r_2 & q_2 & q_1 & q_0 & \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$



## Applications to Queuing Theory: G/M/1 Queue

Proof: Let $M_{n+1}$ be the number of services completed during the $n+1^{\text{th}}$ interarrival time $[T_n, T_{n+1})$. Then,

$$X_{n+1}^\varepsilon = X_n^\varepsilon + 1 - M_{n+1}$$

But $M_{n+1}$ is conditionally independent of the past history before $T_n$ given the present number $X_n^\varepsilon$. If $Z = T_{n+1} - T_n$

$$P\{M_{n+1} = k \mid X_n^\varepsilon, Z\} = \begin{cases} \frac{e^{-aZ}(aZ)^k}{k!} & X_n^\varepsilon + 1 > k \\ \sum_{m=k}^{\infty} \frac{e^{-aZ}(aZ)^m}{m!} & X_n^\varepsilon + 1 = k \quad (*) \\ 0 & \text{otherwise} \end{cases}$$

Taking expectations with respect to $Z$, which is independent of $X_n^\varepsilon$, we obtain

$$P\{M_{n+1} = k \mid X_n^\varepsilon = i\} = \begin{cases} q_k & k \leq i \\ r_{k-1} & k = i+1 \\ 0 & \text{otherwise} \end{cases}$$

Equation $X_{n+1}^\varepsilon = X_n^\varepsilon + 1 - M_{n+1}$ and the previous one provide matrix $P^\varepsilon$

# Applications to Queuing Theory: G/M/1 Queue

**Theorem:** $X^{\varepsilon}$ is recurrent non-null if and only if $r > 1$. If $r > 1$,

$$\pi^{\varepsilon}(j) = \lim_{n\to\infty} P^{\varepsilon\, n}(i,j) = \lim_{n\to\infty} P^{\varepsilon}\{X_n^{\varepsilon} = j \mid X_0^{\varepsilon} = i\}$$

and

$$\pi^{\varepsilon}(j) = (1-\beta)\beta^j, \qquad j = 0,1,2,...$$

where $\beta$ is the unique number satisfying

$$\beta = q_0 + q_1\beta + q_2\beta^2 + \cdots$$

If $r \le 1$ then $\pi^{\varepsilon}(j) = 0$ for all $j$.

Proof: $X^{\varepsilon}$ is recurrent non-null if and only if

$$\nu = \nu \cdot P^{\varepsilon}, \qquad \nu \cdot 1 = 1$$

has a solution.

$$\begin{array}{lcccccccccc}
\nu_0 & = & & & q_1\nu_0 & + & q_2\nu_0 & + & q_3\nu_0 & + & \cdots \\
 & & & & & + & q_2\nu_1 & + & q_3\nu_1 & + & \cdots \\
 & & & & & & & + & q_3\nu_2 & + & \cdots \\
\nu_1 & = & q_0\nu_0 & + & q_1\nu_1 & + & q_2\nu_2 & + & q_3\nu_3 & + & \cdots \\
\nu_2 & = & q_0\nu_1 & + & q_1\nu_2 & + & q_2\nu_3 & + & q_3\nu_4 & + & \cdots
\end{array}$$



# Applications to Queuing Theory: G/M/1 Queue

$$\begin{array}{lcccccccccc}
\nu_0 & = & & & q_1\nu_0 & + & q_2\nu_0 & + & q_3\nu_0 & + & \cdots \\
 & & & & & + & q_2\nu_1 & + & q_3\nu_1 & + & \cdots \\
 & & & & & & & + & q_3\nu_2 & + & \cdots \\
\nu_1 & = & q_0\nu_0 & + & q_1\nu_1 & + & q_2\nu_2 & + & q_3\nu_3 & + & \cdots \\
\nu_2 & = & q_0\nu_1 & + & q_1\nu_2 & + & q_2\nu_3 & + & q_3\nu_4 & + & \cdots
\end{array}$$

Let $f(j) = \nu_0 + \cdots + \nu_{j-1}, \quad j = 1,2,....$ Then,

$$\left\{\begin{array}{lcccccccc}
f(1) & = & q_1 f(1) & + & q_2 f(2) & + & q_3 f(3) & + & \cdots \\
f(2) & = & q_0 f(1) & + & q_1 f(2) & + & q_2 f(3) & + & \cdots \\
f(3) & = & & & q_0 f(2) & + & q_1 f(3) & + & \cdots
\end{array}\right\} \Rightarrow f = Q \cdot f$$

We are interested in a solution satisfying

$$\lim_{j\to\infty} f(j) = \sum_{j=0}^{\infty} \nu_j = 1$$

$Q$ was obtained from $P$ by deleting $0^{th}$ row and column. Such an $f$ exists if and only if $X$ is transient which means that $r > 1$. In this case $f(j) = 1 - \beta^j$. Solving for $\nu$ we obtain

$$\nu_0 = f(1) = 1 - \beta, \qquad \nu_1 = f(2) - f(1) = (1-\beta)\beta, ...$$

## Applications to Queuing Theory: G/M/1 Queue

**Theorem:** $X^\varepsilon$ is transient if and only if $r<1$. If $r<1$, the probability $f^\varepsilon(j)$ that the queue starting with $j$ customers never becomes empty is given by

$$f^\varepsilon(j)=\pi(0)+\pi(1)+\cdots+\pi(j), \qquad j=1,2,...$$

where the $\pi(j)$ are those found in the M/G/1 case.

Proof:

- $f^\varepsilon$ is the solution to the system $h=Q^\varepsilon h$, $0\le h\le 1$.
- $Q^\varepsilon$ is the matrix obtained from $P^\varepsilon$ by deleting the $0^{\text{th}}$ row and column.

The equations for $h=Q^\varepsilon h$ are ( $f^\varepsilon(j)=h_j$ )

$$\begin{aligned} h_1 &= q_0h_2 + q_1h_1 \\ h_2 &= q_0h_3 + q_1h_2 + q_2h_1 \\ h_3 &= q_0h_4 + q_1h_3 + q_2h_2 + q_3h_1 \\ &\vdots \end{aligned}$$

If we define ….



If we define $\pi_0=q_0h_1$, $\pi_1=(1-q_0)h_1$, and let

$$\pi_j=h_j-h_{j-1}, \qquad j=2,3,...$$

then the first of the previous equations along with $\pi_0=q_0h_1$, implies the equations

$$\begin{aligned} \pi_0 &= q_0\pi_0+q_0\pi_1 \\ \pi_1 &= q_1\pi_0+q_1\pi_1+q_0\pi_2 \end{aligned}$$

and subtracting the equation for $h_{j-1}$ from the one for $h_j$ yields

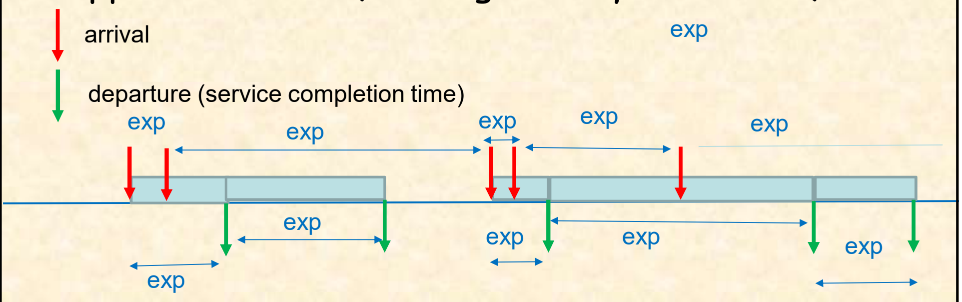
$$\begin{aligned} \pi_2 &= q_2\pi_0+q_2\pi_1+q_1\pi_2+q_0\pi_3 \\ \pi_3 &= q_3\pi_0+q_3\pi_1+q_2\pi_2+q_1\pi_3+q_0\pi_4 \end{aligned}$$

In other words, $\pi$ satisfies $\pi=\pi P$ with $P$ the transition matrix in the M/G/1 case, and we are interested in the solution

$$\pi=\pi P, \qquad \sum_j \pi_j=\lim_j h_j=1$$

- Such a solution exists if and only if $r<1$.
- The solution $\pi$ is connected to $h$ by the relation $h_j=\pi_0+\cdots+\pi_j$

# Applications to Queueing Theory: M/M/1 Queue




X(t): Number of customers in the system (queue and under service)

X(t) is a MC. Why?



# Special case M/M/1

We can consider this queue as a special case of M/G/1 or G/M/1. In the sequel we use G/M/1. Now the interarrival distribution is given by:

$$\phi(t) = 1 - e^{-\lambda t}, \qquad t \geq 0$$

To compute the limiting distribution of $X^{\varepsilon}$ (queue size just before the $n^{\text{th}}$ arrival, we find first $\beta$, where

$$\beta = \sum_{k=0}^{\infty} q_k \beta^k = \sum_{k=0}^{\infty} \beta^k \int_0^{\infty} \frac{e^{-at}(at)^k}{k!} \lambda e^{-\lambda t} dt = \int_0^{\infty} e^{-at(1-\beta)} \lambda e^{-\lambda t} dt = \frac{\lambda}{\lambda + a - a\beta}$$

The previous equation becomes $\beta = \dfrac{\lambda}{\lambda + a - a\beta}$ or $(1-\beta)(\lambda - ab) = 0$

with solutions $\beta = 1$ and $\boxed{\beta = \frac{\lambda}{a}}$. When $r = \frac{a}{\lambda} > 1$, the smallest solution is $\beta = \dfrac{\lambda}{a}$

So we have $\lim_{n\to\infty} P\{X_n^{\varepsilon} = j\} = \left(1 - \dfrac{\lambda}{a}\right)\left(\dfrac{\lambda}{a}\right)^j, \qquad j = 0,1,\ldots$

It turns out that $\boxed{\lim_{t\to\infty} P\{Y_t = j\} = \left(1 - \dfrac{\lambda}{a}\right)\left(\dfrac{\lambda}{a}\right)^j,}$ $j = 0,1,\ldots$ for the queue size $Y_t$ at time $t$.

and $\lim_{n\to\infty} P\{X_n = j\} = \left(1 - \dfrac{\lambda}{a}\right)\left(\dfrac{\lambda}{a}\right)^j, \qquad j = 0,1,\ldots$ for the queue size $X_n$ just after the $n^{\text{th}}$ departure.