general random walk. No new theoretical concepts are introduced,[17] but merely a terminology for a short and intuitive description of the process $\{S_n\}$. For example, if I is any interval (or other set), the event $\{S_n \in I\}$ is called a *visit* to I, and the study of the successive visits to a given interval I reveals important characteristics of the fluctuations of $S_1, S_2, \ldots$. The index n will be interpreted as time parameter and we shall speak of the "epoch n." In this section we describe some striking features of random walks in terms of the successive record values. The usefulness of the results will be shown by the applications in section 9. A second (independent) approach is outlined in section 10.

Imbedded Renewal Processes

A record value occurs at epoch $n > 0$ if

$$S_n > S_j \qquad j = 0, 1, \ldots, n-1. \tag{8.2}$$

Such indices may not exist for a given sample path; if they do exist they form a finite or infinite ordered sequence. It is therefore legitimate to speak of the first, second, ..., occurrence of (8.2). Their epochs are again random variables, but possibly defective. With these preparations we are now in a position to introduce the important random variables on which much of the analysis of random walks will be based.

Definition. *The* kth *(ascending) ladder index is the epoch of the* kth *occurrence of* (8.2). *The* kth *ladder height is the value of* S_n *at the* kth *ladder epoch. (Both random variables are possibly defective.)*

The descending ladder variables are defined in like manner with the inequality in (8.2) *reversed.*[18]

The term *ascending* will be treated as redundant and used only for emphasis or clarity.

In the graph of a sample path $(S_0, S_1, \ldots)$ the ladder points appear as the points where the graph reaches an unprecedented height (record value). Figure 1 represents a random walk $\{S_n\}$ drifting to $-\infty$ with the last positive term at $n = 31$. The 5 ascending and 18 descending ladder points are indicated by ● and ○, respectively. For a random walk with Cauchy variables see figure 2. (page 204)

[17] Sample spaces of infinite random walks were considered also in volume 1, but there we had to be careful to justify notions such as "probability of ruin" by the obvious limiting processes. Now these obvious passages to the limit are justified by measure theory. (See IV,6.)

[18] Replacing the defining strict inequalities by $\geq$ and $\leq$ one gets the *weak* ladder indices. This troublesome distinction is unnecessary when the underlying distribution is continuous. In figure 1 weak ladder points are indicated by the letter w.

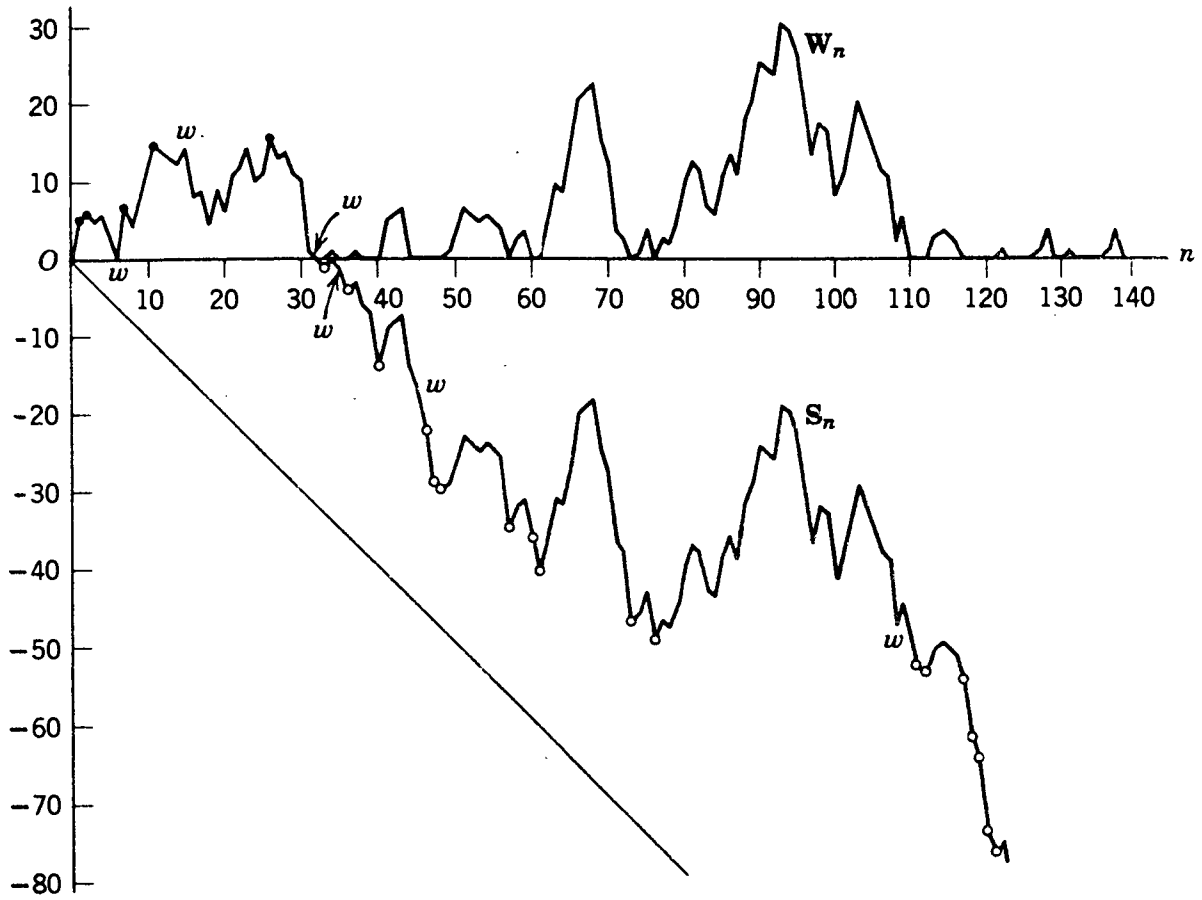

Figure 1. *Random Walk and the Associated Queuing Process.* The variables $\mathbf{X}_n$ of the random walk $\{\mathbf{S}_n\}$ have expectation -1 and variance 16. Ascending and descending ladder points are indicated by ● and ○, respectively. The seventh ladder point is (26, 16) and represents with high probability the maximum of the entire random walk.

[The letter w indicates where a record value is assumed for a second or third time; these are the *weak* ladder points defined by (8.2) when the strict inequality is replaced by $\geq$.]

Throughout the graph $\mathbf{S}_n$ *exceeds its expected value* $-n$. In fact, $n = 135$ *is the first index such that* $\mathbf{S}_n \leq -n$ (namely $\mathbf{S}_{135} = -137$). This accords with the fact that the expectation of such n is infinite.

The variables $\mathbf{X}_n$ are of the form $\mathbf{X}_n = \mathscr{B}_n - \mathscr{A}_n$, where the variables $\mathscr{B}_n$ and $\mathscr{A}_n$ are mutually independent and uniformly distributed over 1, 3, 5, 7, 9 and 2, 4, 6, 8, 10, respectively. In example 9(a) the variable $\mathbf{W}_n$ represents the total waiting time of the nth customer if the interarrival times assume the values 2, 4, 6, 8, 10 with equal probabilities while the service times equal 1, 3, 5, 7, or 9, each with probability $\frac{1}{5}$. The distribution of $\mathbf{X}_n$ attributes probability $(5 - k)/25$ to the points $\pm 2k - 1$, where $k = 0, 1, 2, 3, 4$.

Example. (a) In the "ordinary" random walk F has the atoms 1 and -1 with weights p and q. The ascending ladder variables are defective if $q > p$, the defect p/q [see **1**; XI,(3.9)]. The kth ladder height necessarily equals k and for this reason volume **1** mentions only ladder epochs. The kth ladder index is the epoch of the *first visit* to the point k. Its distribution

was found in 1; XI,4(d) and in the special case $p = \frac{1}{2}$ already in tneorem 2 of 1; III,4.

The *first* ladder index $\mathscr{T}_1$ is the epoch of the first entry into $\overline{0, \infty}$, and the *first* ladder height $\mathscr{H}_1$ equals $\mathbf{S}_{\mathscr{T}_1}$. The continuation of the random walk beyond epoch $\mathscr{T}_1$ is a probabilistic replica of the entire random walk. Given that $\mathscr{T}_1 = n$, the occurrence of a second ladder index a an epoch $k > n$ depends only on $\mathbf{X}_{n+1}, \ldots, \mathbf{X}_k$, and hence the number of trials between the first ladder index and the second is a random variable $\mathscr{T}_2$ which is independent of $\mathscr{T}_1$ and has the same distribution. In this way it is seen more generally that the kth *ladder index and the* kth *ladder height may be written in the form*

$$\mathscr{T}_1 + \cdots + \mathscr{T}_k, \qquad \mathscr{H}_1 + \cdots + \mathscr{H}_k$$

where the $\mathscr{T}_j$ *and* $\mathscr{H}_j$ *are mutually independent random variables distributed, respectively, as* $\mathscr{T}_1$ *and* $\mathscr{H}_1$. In other words, the ladder indices and heights form (possibly terminating) *renewal processes.*

For terminating processes it is intuitively obvious that $\mathbf{S}_n$ drifts to $-\infty$, and with probability one $\mathbf{S}_n$ reaches a finite maximum. The next section will show that the ladder variables provide a powerful tool for the analysis of a class of processes of considerable practical interest.

Example. (b) *Explicit expressions.* Let F have the density defined by

$$\frac{abe^{ax}}{a+b} \text{ if } x < 0; \qquad \frac{abe^{-bx}}{a+b} \text{ if } x > 0. \tag{8.3}$$

This random walk has the rare distinction that all pertinent distributions can be calculated explicitly. It is of great interest in queuing theory because f is the convolution of two exponential densities concentrated on $\overline{0, \infty}$ and $\overline{-\infty, 0}$, respectively. This means that $\mathbf{X}_j$ may be written as the difference $\mathbf{X}_j = \mathscr{B}_j - \mathscr{A}_j$ of two *positive exponentially distributed random variables.* Without loss of generality we assume $a \leq b$.

The *ascending ladder height* $\mathscr{H}_1$ *has the density* ae^{-bx}; this variable is defective and its defect equals $(b-a)/b$. The *ascending ladder epoch* $\mathscr{T}_1$ has the generating function $b^{-1}p(s)$ where

$$2p(s) = a + b - \sqrt{(a+b)^2 - 4abs}. \tag{8.4}$$

The defect is again $(b-a)/b$.

The descending ladder height $\mathscr{H}_1^-$ has density ae^{ax} for $x < 0$, the descending ladder epoch $\mathscr{T}_1^-$ has the generating function $a^{-1}p(s)$. In the special case $a = b$ it reduces to $1 - \sqrt{1-s}$, and this generating function is familiar from ordinary random walks (or coin tossing). [For proofs and other results see XII,4–5 and XVIII,3. See also example 4 e).]

9. THE QUEUING PROCESS

An incredibly voluminous literature[19] has been devoted to a variety of problems connected with servers, storage facilities, waiting times, etc. Much progress has been made towards a unification, but the abundance of small variants obscures the view so that it is difficult to see the forest for the trees. The power of new and general methods is still underrated. We begin by a formal introduction of a stochastic process defined by a recursive scheme that at first sight appears artificial. Examples will illustrate the wide applicability of the scheme; later on we shall see that sharp results can be obtained by surprisingly simple methods. (See XII,5.)

Definition 1. *Let* $\mathbf{X}_1, \mathbf{X}_2, \ldots$ *be mutually independent random variables with a common (proper) distribution* F. *The induced queuing process is the sequence of random variables* $\mathbf{W}_0, \mathbf{W}_1, \ldots$ *defined recursively by* $\mathbf{W}_0 = 0$ *and*

$$\mathbf{W}_{n+1} = \begin{cases} \mathbf{W}_n + \mathbf{X}_{n+1} & \text{if } \mathbf{W}_n + \mathbf{X}_{n+1} \geq 0 \\ 0 & \text{if } \mathbf{W}_n + \mathbf{X}_{n+1} \leq 0 \end{cases} \tag{9.1}$$

In short, $\mathbf{W}_{n+1} = (\mathbf{W}_n + \mathbf{X}_{n+1}) \cup 0$.

For an illustration see figure 1.

Examples. (*a*) *The one-server queue.* Suppose that "customers" arrive at a "server" the arrivals forming a proper renewal process with *inter-arrival times*[20] $\mathscr{A}_1, \mathscr{A}_2, \ldots$ (the epochs of arrivals are $0, \mathscr{A}_1, \mathscr{A}_1 + \mathscr{A}_2, \ldots$ and the customers are labeled $0, 1, 2, \ldots$). With the nth customer there is associated a *service time* $\mathscr{B}_n$, and we assume that the $\mathscr{B}_n$ are independent of the arrivals and of each other and subject to a common distribution. The server is either "free" or "busy"; it is free at the initial epoch 0. The

[19] For references consult the specialized books listed in the bibliography. It would be difficult to give a brief outline of the development of the subject with a proper assignment of credits. The most meritorious papers responsible for new methods are now rendered obsolete by the progress which they initiated. [D. V. Lindley's integral equation of queuing theory (1952) is an example.] Other papers are noteworthy by their treatment of (sometimes very intricate) special problems, but they find no place in a skeleton survey of the general theory. On the whole, the prodigal literature on the several subjects emphasizes examples and variants at the expense of general methods. An assignment of priorities is made difficult also by the many duplications. [For example, the solution of a certain integral equation occurs in a Stockholm thesis of 1939 where it is credited to unpublished lectures by Feller in 1934. This solution is now known under several names.] For the history see two survey papers by D. G. Kendall of independent interest: *Some problems in the theory of queues*, and *Some problems in the theory of dams*, J. Roy. Statist. Soc. Series B vol. 13 (1951) pp. 151–185, and vol. 19 (1957) pp. 207–233.

[20] Normally the interarrival times will be constant or exponentially distributed but it is fashionable to permit arbitrary renewal processes; see footnote 14 to section 7.

sequel is regulated by the following rule. If a customer arrives at an epoch where the server is free, his service commences without delay. Otherwise he joins a waiting line (queue) and the server continues uninterruptedly to serve customers in the order of their arrival[21] until the waiting line disappears and the server becomes "free." By *queue length* we mean the number of customers present including the customer being served. The *waiting time* $\mathbf{W}_n$ of the nth customer is the time from his arrival to the epoch where his service *commences*; the total time spent by the customer at the server is $\mathbf{W}_n + \mathscr{B}_n$. (For example, if the first few service times are 4, 4, 1, 3, ... and the interarrival times are 2, 3, 2, 3, ..., customers number 1, 2, ... *join* queues of length 1, 1, 2, 1, ..., respectively, and have waiting times 2, 3, 2, 2, ...).

To avoid trite ambiguities such as when a customer arrives at the epoch of another's departure we shall assume that the distributions A and B of the variables $\mathscr{A}_n$ and $\mathscr{B}_n$ are continuous. Then the queue length at any epoch is well defined.

We proceed to devise a scheme for calculating the waiting times $\mathbf{W}_n$ recursively. By definition customer number 0 arrives at epoch 0 at a free server and so his waiting time is $\mathbf{W}_0 = 0$. Suppose now that the nth customer arrives at epoch t and that we know his waiting time $\mathbf{W}_n$. His service time commences at epoch $t + \mathbf{W}_n$ and terminates at epoch $t + \mathbf{W}_n + \mathscr{B}_n$. The *next* customer arrives at time $t + \mathscr{A}_{n+1}$. He finds the server free if $\mathbf{W}_n + \mathscr{B}_n < \mathscr{A}_{n+1}$ and has a waiting time $\mathbf{W}_{n+1} = \mathbf{W}_n + \mathscr{B}_n - \mathscr{A}_{n+1}$ if this quantity is ≥ 0. In other words, *the sequence* $\{\mathbf{W}_n\}$ *of waiting times coincides with the queuing process induced by the independent random variables*

$$\mathbf{X}_n = \mathscr{B}_{n-1} - \mathscr{A}_n, \qquad n = 1, 2, \ldots \tag{9.2}$$

(*b*) *Storage and inventories.* For an intuitive description we use water *reservoirs* (and dams), but the model applies equally to other storage facilities or inventories. The content depends on the input and the output. The input is due to supplies by rivers and rainfall, the output is regulated by demand except that this demand can be satisfied only when the reservoir is not empty.

Consider now the water contents[22] $0, \mathbf{W}_1, \mathbf{W}_2, \ldots$ at selected epochs $0, \tau_1, \tau_2, \ldots$. Denote by $\mathbf{X}_n$ the actual supply minus the theoretical (ideal)

[21] This "queue discipline" is totally irrelevant to queue length, duration of busy periods, and similar problems. Only the individual customer feels the effect of the several disciplines, among which "first come first served," "first come last served" and "random choice" are the extremes. The whole picture would change if departures were permitted.

[22] For simplicity we start with an empty reservoir. An adjustment to arbitrary initial conditions causes no difficulties [see example (*c*)].

demand during τ_{n-1}, τ_n and let us pretend that all changes are instantaneous and concentrated at the epochs $\tau_1, \tau_2, \ldots$. We start with $\mathbf{W}_0 = 0$ at epoch 0. In general the change $\mathbf{W}_{n+1} - \mathbf{W}_n$ should equal $\mathbf{X}_{n+1}$ except when the demand exceeds the contents. For this reason the $\mathbf{W}_n$ must satisfy (9.1) and so the *successive contents are subject to the queuing process induced by* $\{\mathbf{X}_k\}$ provided the theoretical net changes $\mathbf{X}_k$ are independent random variables with a common distribution.

The problem (for the mathematician if not for the user) is to find conditions under which the $\mathbf{X}_k$ will appear as independent variables with a common distribution F and to find plausible forms for F. Usually the τ_k will be equidistant or else a sample from a Poisson process, but it suffices for our purposes to assume that the τ_k form a *renewal process* with interarrival times $\mathscr{A}_1, \mathscr{A}_2, \ldots$. The most frequently used models fall into one of the following two categories:

(i) The input is at a constant rate c, the demand $\mathscr{B}_n$ arbitrary. Then $\mathbf{X}_n = c\mathscr{A}_n - \mathscr{B}_n$. We must suppose this $\mathbf{X}_n$ to be independent of the "past" $\mathbf{X}_1, \ldots, \mathbf{X}_{n-1}$. (The usual assumption that $\mathscr{A}_n$ and $\mathscr{B}_n$ be independent is superfluous: there is no reason why the demand $\mathscr{B}_n$ should not be correlated with the duration $\mathscr{A}_n$.)

(ii) The output is at a constant rate, the input arbitrary. The description is the same with the roles of $\mathscr{A}_n$ and $\mathscr{B}_n$ reversed.

(*c*) *Queues for a shuttle train.*[23] A shuttle train with r places for passengers leaves a station every hour on the hour. Prospective passengers appear at the station and wait in line. At each departure the first r passengers in line board the train, and the others remain in the waiting line. We suppose that the number of passengers arriving between successive departures are independent random variables $\mathscr{A}_1, \mathscr{A}_2, \ldots$ with a common distribution. Let $\mathbf{W}_n$ be the number of passengers in line just after the nth departure, and assume for simplicity $\mathbf{W}_0 = 0$. Then $\mathbf{W}_{n+1} = \mathbf{W}_n + \mathscr{A}_{n+1} - r$ if this quantity is positive, and $\mathbf{W}_{n+1} = 0$ otherwise. Thus $\mathbf{W}_n$ *is the variable of a queuing process* (9.1) *generated by the* random walk with variables $\mathbf{X}_n = \mathscr{A}_n - r$. ▶

We turn to a description of the queuing process $\{\mathbf{W}_n\}$ in terms of the random walk generated by the variables $\mathbf{X}_k$. As in section 8 we put $\mathbf{S}_0 = 0$, $\mathbf{S}_n = \mathbf{X}_1 + \cdots + \mathbf{X}_n$ and adhere to the notation for the ladder variables. For ease of description we use the terminology appropriate for the server of example (*a*).

[23] P. E. Boudreau, J. S. Griffin Jr., and Mark Kac, *An elementary queuing problem*, Amer. Math. Monthly, vol. 69 (1962) pp. 713–724. The purpose of this paper is didactic, that is, it is written for outsiders without knowledge of the subject. Although a different mode of description is used, the calculations are covered by those in example XII,4(*c*).

Define ν as the subscript for which $S_1 \geq 0, S_2 \geq 0, \ldots, S_{\nu-1} \geq 0$, but $S_\nu < 0$. In this situation customers number $1, 2, \ldots, \nu-1$ had positive waiting times $W_1 = S_1, \ldots, W_{\nu-1} = S_{\nu-1}$, and customer number ν was the first to find the server free (the first lucky customer). At the epoch of his arrival the process starts from scratch as a replica of the whole process. Now ν is simply the index of the first negative sum, that is, ν is the first descending ladder index, and we denote it consistently by $\mathscr{T}_1^-$. We have thus reached the *first conclusion: The descending ladder indices correspond to the lucky customers who find the server free.* Put differently, the epochs of arrival of the lucky customers constitute a renewal process with recurrence times distributed as $\mathscr{T}_1^-$.

In practical cases the variable $\mathscr{T}_1^-$ must not be defective, for its defect p would equal the probability that a customer never finds the server free and with probability one there would be a last lucky customer followed by an unending queue. It will turn out that $\mathscr{T}_1^-$ is proper whenever $E(\mathscr{B}_k) < E(\mathscr{A}_k)$.

Suppose now that customer number $\nu - 1$ arrives at epoch τ. His waiting time was $W_{\nu-1} = S_{\nu-1}$ and so the epoch of his departure is $\tau + W_{\nu-1} + \mathscr{B}_{\nu-1}$. The first lucky customer (number ν) arrives at epoch $\tau + \mathscr{A}_\nu$ when the server was free for

$$\mathscr{A}_\nu - W_{\nu-1} - \mathscr{B}_{\nu-1} = -S_{\nu-1} - X_\nu = -S_\nu$$

time units. But by definition S_ν is the first descending ladder height $\mathscr{H}_1^-$. As the process starts from scratch we have reached the *second conclusion: The durations of the free periods are independent random variables with the same distribution as* $-\mathscr{H}_1^-$ (the recurrence time for the descending ladder heights). In other words, customer number $\mathscr{T}_1^- + \cdots + \mathscr{T}_r^-$ is the rth customer who finds the server free. At the epoch of his arrival the server has been free for $-\mathscr{H}_r^-$ time units.

It should now be clear that between successive ladder epochs *the segments of the graph for the queuing process* $\{W_n\}$ *are congruent to those for the random walk* but displayed vertically so as to start at a point of the time axis (figure 1). To describe this analytically denote for the moment by $[n]$ the *last* descending ladder index $\leq n$; in other words, $[n]$ is a (random) index such that $[n] \leq n$ and

$$S_{[n]} \leq S_j \qquad j = 0, 1, \ldots, n. \tag{9.3}$$

This defines $[n]$ uniquely with probability 1 (the distribution of X_i being continuous). Clearly

$$W_n = S_n - S_{[n]}. \tag{9.4}$$

This relation leads to the most important conclusion if we look at the

variables $\mathbf{X}_1, \ldots, \mathbf{X}_n$ in *reverse order*. Put for abbreviation $\mathbf{X}_1' = \mathbf{X}_n, \ldots, \mathbf{X}_n' = \mathbf{X}_1$. The partial sums of these variables are

$$\mathbf{S}_k' = \mathbf{X}_1' + \cdots + \mathbf{X}_k' = \mathbf{S}_n - \mathbf{S}_{n-k},$$

and (9.4) shows that the maximal term of the sequence $0, \mathbf{S}_1', \ldots, \mathbf{S}_n'$ has subscript $n - [n]$ and equals $\mathbf{W}_n$. But the distribution of $(\mathbf{X}_1', \ldots, \mathbf{X}_n')$ is identical with that of $(\mathbf{X}_1, \ldots, \mathbf{X}_n)$. We have thus the basic

Theorem.[24] *The distribution of the queuing variable* $\mathbf{W}_n$ *is identical with the distribution of the random variable*

$$\mathbf{M}_n = \max [0, \mathbf{S}_1, \ldots, \mathbf{S}_n] \tag{9.5}$$

in the underlying random walk $\{\mathbf{X}_k\}$.

The consequences of this theorem will be discussed in chapter XII. Here we show that it permits us to reduce certain ruin problems to queuing processes despite the dissimilarity of the appearance.

Example. (*d*) *Ruin problems.* In section 5 ruin was defined as the event that $\mathbf{X}(t) > z + ct$ for some t where $\mathbf{X}(t)$ is the variable of a compound Poisson process with distribution (4.2). Denote the epochs of the successive jumps in this process by $\tau_1, \tau_2, \ldots$. If ruin occurs at all it occurs also at some epoch τ_k and it suffices therefore to consider the probability that $\mathbf{S}_n = \mathbf{X}(\tau_n) - c\tau_n > z$ for some n. But by the definition of a compound Poisson process $\mathbf{X}(\tau_n)$ is the sum of n independent variables $\mathbf{Y}_k$ with the common distribution F, while τ_n is the sum of n independent exponentially distributed variables $\mathscr{A}_k$. Accordingly we are in effect dealing with the random walk generated by the variables $\mathbf{X}_k = \mathbf{Y}_k - c\mathscr{A}_k$ whose probability *density* is given by the convolution

$$\frac{\alpha}{c} \int_x^\infty e^{\alpha(x-y)/c} F\{dy\}. \tag{9.6}$$

Ruin occurs iff in the random walk the event $\{\mathbf{S}_n \geq z\}$ *takes place for some n.* To find the probability of ruin amounts therefore to finding the distributions of the variables $\mathbf{W}_n$ in the associated queuing process.

(*e*) *A numerical illustration.* The most important queuing process arises when *the interarrival and service times are exponentially distributed* with expectations $1/a$ and $1/b$, respectively, where $a < b$. From the characteristics of this process described in example 8(*b*), one can conclude that the *waiting time of the* nth *customer has a limit distribution* W *with an atom of*

[24] Apparently first noticed by F. Pollaczek in 1952 and exploited (in a different context) by F. Spitzer, *The Wiener–Hopf equation whose kernel is a probability density*, Duke Math. J., vol. 24 (1957) pp. 327–344. For Spitzer's proof see problem 21.