Analysis of dependence metrics for queueing processes
Es-Saghouani, A.

Citation for published version (APA):
Chapter 1

Introduction

Having applications in everyday situations that we face, queueing theory is an appealing part of applied probability. Queueing theory started to attract interest from mathematicians at the beginning of the last century, when a paper of Erlang [49] on the theory of telephone traffic appeared. In this pioneering work, Erlang showed that the number of telephone calls made during a given interval of time, assuming random origination of the calls, follows a Poisson distribution. In a second paper, which is generally considered as the more important one, Erlang [50] studied blocking probabilities when a fixed number of telephone lines are available, and proved what has become the famous Erlang loss formula.

Nowadays, queueing theory has turned into an indispensable tool for modeling various real life phenomena. For instance, it is used in the study of production and storage systems, where the objective is to regulate the demand so as to achieve a storage of desirable level, but also in communication networks, e.g., telephone networks, Internet and computer systems where one aims at guaranteeing a given level of service. In other application domains models are used that are closely related to queueing systems, for instance in insurance risk. There one of the objectives is the study of the ruin probability of an insurance company, given some initial capital.

A description of a simple queueing system is as follows. Suppose that we have a service station that processes work fed into the system. If traffic arrives at a faster rate than it can be served, then the unfinished work can be stored in a buffer, which, for ease, is assumed to have infinite storage capacity. Mathematically such a system can be modeled as follows. Let $X \equiv \{X(t) : t \in \mathbb{R}\}$ be the stochastic process describing the way the input arrives to the service station, and let $A(s, t) = X(t) - X(s)$ denote the amount of traffic entering between time epochs $s$ and $t$ ($s \leq t$). Furthermore, the system is emptied at a constant rate $c > 0$, meaning that $ct$ units of work can be potentially processed in any time window of length $t$. This naturally defines a workload process, denoted henceforth by $Q \equiv \{Q(t) : t \in \mathbb{R}\}$, representing the amount of work in storage at time $t$. In the literature, the process $Q$ is also known as the buffer content process or the virtual waiting time process. Throughout this thesis we will use these terms interchangeably.

Obviously, the most interesting situation is the one where the state of the system alternates between busy periods and idle periods, rather than being essentially
Introduction

To make sure that the system alternates between busy and idle periods, we should require that the average rate at which the work arrives at the system is less than the rate at which the system is emptied. Assuming that the input process $X$ has stationary increments, $A(s, t) := X(t) - X(s)$ for $s \leq t$, by which we mean that $A(s, t)$ and $A(s + u, t + u)$ have the same probability distribution for any $u$, then it can easily be shown that the mean of the increments $E[A(s, t)]$ is linear in $t - s$. In this case we can write $EA(s, t) = \varpi \cdot (t - s)$, with $\varpi = EA(0, 1)$. Then the criterion just described above reduces to requiring that

$$\varpi < c.$$  \hspace{1cm} (1.1)

If Condition (1.1) is satisfied, then the queueing system is said to be stable and we will refer to (1.1) as the stability condition. Indeed, if $\varpi > c$, then the content of the queue will eventually grow beyond any bound.

In this monograph the main stochastic process that we will be interested in, is the content process $\{Q(t) : t \in \mathbb{R}\}$. In Section 1.1, we will see that the workload $Q(t)$ of the queue at time $t$ can be formulated as a functional of the net input process $\{A(s, t) - c(t - s) : s \leq t\}$. Assuming stationarity of the increments $A(s, t)$ of the input process, and imposing condition (1.1), it can be shown that the distributions of $Q(t)$ have a weak limit as $t \to \infty$, and we let $Q_e$ denote the stationary (or steady-state, or equilibrium) workload. Most of the research concerning the process $Q$ has dealt with the stochastic properties of the steady-state workload $Q_e$. For special input processes explicit results are available for the steady-state distribution $\mathbb{P}(Q_e \leq q)$, or sometimes for the Laplace-Stieltjes transform $Ee^{-sQ_e}$, $s \geq 0$, of the stationary workload $Q_e$. In contrast, transient behavior of the system (i.e., the time-dependent case where one studies $\mathbb{P}(Q(t) \leq q)$ given the state of the system at time 0) is considerably less explored, and only in relatively few situations one has succeeded in deriving explicit results.

The main objective of this thesis, as its title indicates, is to investigate the dependence structure of the workload process $Q$. Specifically, we will investigate to what extent the dependence structure of the input process is inherited by the workload process. Throughout this thesis we will assume that the system is already in stationarity at time $t = 0$. Under this assumption we will study two types of metrics capturing the dependence structure of the workload process $Q$.

- The covariance function $R(t)$, to be defined in (1.9), is a measure that gives insight in the speed at which the correlation between $Q(0)$ and $Q(t)$ for $t \geq 0$ vanishes as $t$ grows large.

- A second measure that will be extensively studied, is the measure $r(T|p, q)$, to be defined in (1.13), featuring the joint probability of the workload exceeding some thresholds $p$ at time 0 and $q$ at time $T$. 

continuously busy. To make sure that the system alternates between busy and idle periods, we should require that the average rate at which the work arrives at the system is less than the rate at which the system is emptied. Assuming that the input process $X$ has stationary increments, $A(s, t) := X(t) - X(s)$ for $s \leq t$, by which we mean that $A(s, t)$ and $A(s + u, t + u)$ have the same probability distribution for any $u$, then it can easily be shown that the mean of the increments $E[A(s, t)]$ is linear in $t - s$. In this case we can write $EA(s, t) = \varpi \cdot (t - s)$, with $\varpi = EA(0, 1)$. Then the criterion just described above reduces to requiring that

$$\varpi < c.$$  \hspace{1cm} (1.1)

If Condition (1.1) is satisfied, then the queueing system is said to be stable and we will refer to (1.1) as the stability condition. Indeed, if $\varpi > c$, then the content of the queue will eventually grow beyond any bound.

In this monograph the main stochastic process that we will be interested in, is the content process $\{Q(t) : t \in \mathbb{R}\}$. In Section 1.1, we will see that the workload $Q(t)$ of the queue at time $t$ can be formulated as a functional of the net input process $\{A(s, t) - c(t - s) : s \leq t\}$. Assuming stationarity of the increments $A(s, t)$ of the input process, and imposing condition (1.1), it can be shown that the distributions of $Q(t)$ have a weak limit as $t \to \infty$, and we let $Q_e$ denote the stationary (or steady-state, or equilibrium) workload. Most of the research concerning the process $Q$ has dealt with the stochastic properties of the steady-state workload $Q_e$. For special input processes explicit results are available for the steady-state distribution $\mathbb{P}(Q_e \leq q)$, or sometimes for the Laplace-Stieltjes transform $Ee^{-sQ_e}$, $s \geq 0$, of the stationary workload $Q_e$. In contrast, transient behavior of the system (i.e., the time-dependent case where one studies $\mathbb{P}(Q(t) \leq q)$ given the state of the system at time 0) is considerably less explored, and only in relatively few situations one has succeeded in deriving explicit results.

The main objective of this thesis, as its title indicates, is to investigate the dependence structure of the workload process $Q$. Specifically, we will investigate to what extent the dependence structure of the input process is inherited by the workload process. Throughout this thesis we will assume that the system is already in stationarity at time $t = 0$. Under this assumption we will study two types of metrics capturing the dependence structure of the workload process $Q$.

- The covariance function $R(t)$, to be defined in (1.9), is a measure that gives insight in the speed at which the correlation between $Q(0)$ and $Q(t)$ for $t \geq 0$ vanishes as $t$ grows large.

- A second measure that will be extensively studied, is the measure $r(T|p, q)$, to be defined in (1.13), featuring the joint probability of the workload exceeding some thresholds $p$ at time 0 and $q$ at time $T$. 

continuously busy. To make sure that the system alternates between busy and idle periods, we should require that the average rate at which the work arrives at the system is less than the rate at which the system is emptied. Assuming that the input process $X$ has stationary increments, $A(s, t) := X(t) - X(s)$ for $s \leq t$, by which we mean that $A(s, t)$ and $A(s + u, t + u)$ have the same probability distribution for any $u$, then it can easily be shown that the mean of the increments $E[A(s, t)]$ is linear in $t - s$. In this case we can write $EA(s, t) = \varpi \cdot (t - s)$, with $\varpi = EA(0, 1)$. Then the criterion just described above reduces to requiring that

$$\varpi < c.$$  \hspace{1cm} (1.1)

If Condition (1.1) is satisfied, then the queueing system is said to be stable and we will refer to (1.1) as the stability condition. Indeed, if $\varpi > c$, then the content of the queue will eventually grow beyond any bound.

In this monograph the main stochastic process that we will be interested in, is the content process $\{Q(t) : t \in \mathbb{R}\}$. In Section 1.1, we will see that the workload $Q(t)$ of the queue at time $t$ can be formulated as a functional of the net input process $\{A(s, t) - c(t - s) : s \leq t\}$. Assuming stationarity of the increments $A(s, t)$ of the input process, and imposing condition (1.1), it can be shown that the distributions of $Q(t)$ have a weak limit as $t \to \infty$, and we let $Q_e$ denote the stationary (or steady-state, or equilibrium) workload. Most of the research concerning the process $Q$ has dealt with the stochastic properties of the steady-state workload $Q_e$. For special input processes explicit results are available for the steady-state distribution $\mathbb{P}(Q_e \leq q)$, or sometimes for the Laplace-Stieltjes transform $Ee^{-sQ_e}$, $s \geq 0$, of the stationary workload $Q_e$. In contrast, transient behavior of the system (i.e., the time-dependent case where one studies $\mathbb{P}(Q(t) \leq q)$ given the state of the system at time 0) is considerably less explored, and only in relatively few situations one has succeeded in deriving explicit results.

The main objective of this thesis, as its title indicates, is to investigate the dependence structure of the workload process $Q$. Specifically, we will investigate to what extent the dependence structure of the input process is inherited by the workload process. Throughout this thesis we will assume that the system is already in stationarity at time $t = 0$. Under this assumption we will study two types of metrics capturing the dependence structure of the workload process $Q$.

- The covariance function $R(t)$, to be defined in (1.9), is a measure that gives insight in the speed at which the correlation between $Q(0)$ and $Q(t)$ for $t \geq 0$ vanishes as $t$ grows large.

- A second measure that will be extensively studied, is the measure $r(T|p, q)$, to be defined in (1.13), featuring the joint probability of the workload exceeding some thresholds $p$ at time 0 and $q$ at time $T$. 

continuously busy. To make sure that the system alternates between busy and idle periods, we should require that the average rate at which the work arrives at the system is less than the rate at which the system is emptied. Assuming that the input process $X$ has stationary increments, $A(s, t) := X(t) - X(s)$ for $s \leq t$, by which we mean that $A(s, t)$ and $A(s + u, t + u)$ have the same probability distribution for any $u$, then it can easily be shown that the mean of the increments $E[A(s, t)]$ is linear in $t - s$. In this case we can write $EA(s, t) = \varpi \cdot (t - s)$, with $\varpi = EA(0, 1)$. Then the criterion just described above reduces to requiring that

$$\varpi < c.$$  \hspace{1cm} (1.1)

If Condition (1.1) is satisfied, then the queueing system is said to be stable and we will refer to (1.1) as the stability condition. Indeed, if $\varpi > c$, then the content of the queue will eventually grow beyond any bound.
1.1. Workload process

The remainder of this introductory chapter is organized as follows. Section 1.1 describes the workload process in both discrete-time and continuous-time queueing systems. In Section 1.2, we recall some basic properties of the covariance function and give some motivations for its study. Then, in the same section we will introduce and motivate the choice of the alternative dependence metric $R(T|p, q)$. Sections 1.3, 1.4 and 1.5 present basic background results on the stochastic processes considered as input processes, namely Gaussian, Lévy and Markov modulated fluid processes, respectively. The overview and contributions of the thesis are given in Section 1.6.

1.1 Workload process

For ease of exposition, let us begin by considering a discrete-time queueing system. Suppose that at discrete time epochs $t = 1, 2, \ldots$, the amount of traffic entering the system at time $t$ is given by the random quantity $Y(t)$. Define the stochastic process $\{X(t) : t = 0, 1, \ldots\}$ as follows: $X(0) = 0$ and for $t \geq 1$, $X(t) = Y(1) + Y(2) + \ldots + Y(t)$. Thus $X(t)$ is the amount of traffic arrived to the system in time epochs 1 up to $t$. Let the queue be emptied at a constant rate $c > 0$, and assume in addition that the queue was empty at time epoch $t = 0$. The workload process $\{Q(t) : t \geq 0\}$ induced by the process $\{Y(t) : t \geq 1\}$ is defined by

$$Q(0) = 0,$$

and for $t \geq 0$

$$Q(t+1) = \begin{cases} Q(t) + Y(t+1) - c & \text{if } Q(t) + Y(t+1) - c \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

(1.2)

In other words: the process $\{Q(t) : t \geq 0\}$ is given by the following recurrence relation, which is known as the Lindley recursion:

$$Q(0) = 0, \quad Q(t+1) = \max \{Q(t) + Y(t+1) - c, 0\}, \quad t \geq 0.$$  

(1.3)

Solving the recurrence relation (1.3) shows that

$$Q(t) = \max_{s=0,1,\ldots,t} (X(t) - X(s) - c(t-s)) = \max_{s=0,1,\ldots,t} (A(s,t) - c(t-s)).$$

Since we assumed that the increments $A(s,t)$ of the process $X$ are stationary and that the stability condition (1.1) is satisfied, it can be shown that

$$Q_e \overset{d}{=} \sup_{s \in \mathbb{N}_0} (A(-s, 0) - cs),$$

(1.4)

where $\overset{d}{=} \text{ denotes equality in distribution, and } A(-s, 0)$ is to be interpreted as the amount of traffic entering the system in time epochs $-s$ up to $-1$. This result is often attributed to Reich [100].
Now consider the continuous-time case. Suppose that the input of a queueing system is a stochastic process \( \{X(t) : t \in \mathbb{R}\} \) with stationary increments \( A(s, t) \). As before, we assume that the queue is emptied at a positive deterministic rate \( c > 0 \). With \( Q(0) \geq 0 \), the continuous-time analogue of (1.3) is obtained by reflecting \( \{Q(0) + A(s, t) - c(t - s) : s \leq t\} \) at zero. More specifically, for \( t \geq 0 \), the workload \( Q(t) \) at time \( t \) is given by

\[
Q(t) = A(0, t) - ct + \max\left(Q(0), -\inf_{0 \leq s \leq t} (A(0, s) - cs)\right), \tag{1.5}
\]

which, using straightforward calculus, can be rewritten as

\[
Q(t) = \max\left(\sup_{0 \leq s \leq t} (A(s, t) - c(t - s)), Q(0) + A(0, t) - ct\right). \tag{1.6}
\]

Under the stability condition (1.1), it can be shown, see for instance [69], that there exists a stationary stochastic process defined by

\[
Q(t) = \sup_{-\infty < s \leq t} (A(s, t) - c(t - s)), \quad -\infty < t < \infty, \tag{1.7}
\]

satisfying (1.6). The steady-state workload is then given by

\[
Q_e \overset{d}{=} \sup_{s \geq 0} (A(-s, 0) - cs), \tag{1.8}
\]

where \( A(-s, 0) \) is to be interpreted as the amount of traffic generated during the time interval \([-s, 0)\). Observe that (1.8) is the continuous counterpart of (1.4).

### 1.2 Dependence metrics

Consider the stationary workload process \( \{Q(t) : t \geq 0\} \) defined by (1.7). By stationarity, its mean \( \mathbb{E}Q(t) \) and variance \( \mathbb{V}a\mathbb{r}Q(t) \) are constant in \( t \), and will be denoted henceforth by \( \mu \) and \( v \) respectively. Moreover, by stationarity \( \mathbb{E}Q(s)Q(t) \) depends only on the difference \( t - s \). Furthermore, assuming the variance \( v \) is finite, it can be shown, using the Cauchy-Schwarz inequality, that also for every \( s \) and \( t \), \( \mathbb{E}Q(s)Q(t) \) is finite. Therefore, the covariance function \( R(\cdot) \) and the correlation function \( r(\cdot) \) are well-defined, and are given by, respectively,

\[
R(t) := \mathbb{C}ov(Q(0), Q(t)) = \mathbb{E}Q(0)Q(t) - \mathbb{E}Q(0)\mathbb{E}Q(t) = \mathbb{E}Q(0)Q(t) - \mu^2 \tag{1.9}
\]

\[
r(t) := \mathbb{C}orr(Q(0), Q(t)) = \frac{\mathbb{C}ov(Q(0), Q(t))}{\sqrt{\mathbb{V}a\mathbb{r}Q(0)}\sqrt{\mathbb{V}a\mathbb{r}Q(t)}} = \frac{R(t)}{v}. \tag{1.10}
\]

There are several reasons for studying the covariance function \( R(\cdot) \) of the workload process \( Q \), see for instance [102].
• In the first place, interesting stochastic properties of the process $Q$ can be captured via its covariance function. Indeed, for instance $L^2$-continuity (that is, quadratic mean continuity), -differentiability and -integrability of the workload process $Q$ depend upon similar properties of its covariance function.

• Moreover, knowledge of the covariance function $R(\cdot)$ is important to assess the variance of estimators for the mean workload $\mu = E_{Q_e}$, for instance when one considers estimators of the type

$$\hat{\mu} = \frac{1}{T} \int_0^T Q(t)dt.$$ 

This can be seen as follows. First notice that $\hat{\mu}$ is an unbiased estimator of $\mu$ ($E\hat{\mu} = \mu$) if the queue was in stationarity at time 0. To assess its precision we have to evaluate its standard error, which can be calculated from

$$\text{Var}\hat{\mu} = \frac{1}{T^2} \int_0^T \int_0^T \text{Cov}(Q(s), Q(t))dsdt = \frac{2}{T^2} \int_0^T \int_0^T R(u)du dt.$$ 

This explains the interest in studying the covariance function.

• Furthermore, application of spectral analysis to queueing theory requires the determination of the correlation function as the Fourier transform of the spectral density. Indeed, if the correlation function $r(\cdot)$ is continuous at $t = 0$, and hence at every point $t \in [0, \infty)$, then it can be represented as

$$r(t) = \int_{-\infty}^{\infty} e^{itx} dF(x),$$  \hspace{1cm} (1.11)$$

where $F(\cdot)$ is real, strictly increasing and bounded. $F(\cdot)$ is called the spectral distribution of the process $Q$. Since $Q$ is a real-valued process, $r(t)$ will then be real and (1.11) yields

$$r(t) = \int_{-\infty}^{\infty} \cos(tx)dF(x) = F(0) + 2 \int_0^{\infty} \cos(tx)dF(x).$$  \hspace{1cm} (1.12)$$

For further reading on the covariance function of stationary stochastic processes, we refer the reader to Cramér and Leadbetter [34].

In this thesis we also consider an alternative measure for the study of the dependence structure of the workload process. Assuming the queue is already in stationarity at time 0 and with $Q_e$ denoting the stationary workload, we define for $p, q$ and $T > 0$

$$R(T|p, q) := \frac{P(Q(0) > p, Q(T) > q)}{P(Q_e > p)P(Q_e > q)}.$$  \hspace{1cm} (1.13)$$
Observe that ‘the more independent’ the events \( \{ Q(0) > p \} \) and \( \{ Q(T) > q \} \) are, the more \( R(T|p, q) \) approaches the value 1. In this sense, the metric \( R(T|p, q) \) can be seen as a measure that describes the dependence of the events \( \{ Q(0) > p \} \) and \( \{ Q(T) > q \} \). Moreover, the measure \( R(T|p, q) \) can be related to the covariance of the corresponding indicator functions. Indeed, considering the indicator functions \( 1_{\{ Q(0) > p \}} \) and \( 1_{\{ Q(T) > q \}} \) of the events \( \{ Q(0) > p \} \) and \( \{ Q(T) > q \} \), respectively, it is easily seen that

\[
\text{Cov}(1_{\{ Q(0) > p \}}, 1_{\{ Q(T) > q \}}) = P(Q_e > p)P(Q_e > q) (R(T|p, q) - 1).
\]

A further motivation for the choice of this dependence metric and its relation to mixing properties will be given in Chapter 2.

Now that we have defined the performance metrics of our interest, we proceed by briefly describing the three classes of input models that we consider in this monograph.

### 1.3 Gaussian processes

A real-valued stochastic process \( \{ X(t) : t \in \mathbb{R} \} \) is called Gaussian, if all its finite dimensional distributions are multivariate normal distributions, i.e., for all \( 1 \leq n < \infty \), and \( t_1 < t_2 < \ldots < t_n \), the random vector \( (X(t_1), X(t_2), \ldots, X(t_n)) \) has a multivariate normal distribution. In particular, for each \( t \), the random variable \( X(t) \) has a normal distribution with some mean \( m(t) \) and variance \( v(t) \).

Furthermore, a Gaussian process is completely characterized by its mean function \( m(t) = E(X(t)) \) and covariance function \( \Gamma(s, t) = \text{Cov}(X(s), X(t)) \). Notice that every covariance function is nonnegative definite in the sense that, for any finite set of indices \( \{ t_1, \ldots, t_k \} \), and \( \alpha_1, \ldots, \alpha_k \) arbitrary numbers, \( \Gamma(s, t) \) satisfies

\[
\sum_{i,j=1}^{k} \alpha_i \Gamma(t_i, t_j) \alpha_j \geq 0.
\]

(1.14)

It follows that, given any function \( \Gamma(s, t) \) satisfying (1.14), we can always construct a Gaussian process having \( \Gamma(s, t) \) as its covariance function, see for instance [34, pp. 80-82].

In this thesis, we only consider Gaussian processes with stationary increments \( A(s, t) = X(t) - X(s) \) for \( s \leq t \). For this class of Gaussian processes it can be easily verified that

\[
\Gamma(s, t) = \frac{1}{2} (v(t) + v(s) - v(t - s)).
\]

(1.15)
As indicated by (1.15), the class of Gaussian processes with stationary increments is extremely rich. Indeed, if we define $\Gamma(s, t)$ with $v(t)$ a continuous positive function, such that $\Gamma(s, t)$ is nonnegative definite, then we can construct a corresponding Gaussian process $X$ with stationary increments.

Some examples of Gaussian processes, that will be studied in this thesis, are:

- **Fractional Brownian motion**, characterized by the following variance function
  \[ v(t) = |t|^{2H}, \quad H \in [0, 1] \text{ and } t \in \mathbb{R}; \]  \hspace{1cm} (1.16)
  where $H$ is the so-called Hurst parameter. Observe that for $H = 1/2$ we have Brownian motion.

- **The integrated Ornstein-Uhlenbeck process** having the variance function
  \[ v(t) = |t| - 1 + e^{-|t|}, \quad t \in \mathbb{R}. \]  \hspace{1cm} (1.17)

For this class of input processes with $\varpi < c$, there exist no explicit formulae for the steady-state probability distribution of the stationary workload $Q_e$, except for Brownian motion input. For Brownian motion input, where $v(t) = |t|$, it is well-known that the stationary workload $Q_e$ is exponentially distributed with parameter $2(c - \varpi)$, see for instance [61]. For other Gaussian processes, one has found approximations of $\mathbb{P}(Q_e > B)$, the probability that the workload exceeds some threshold $B > 0$, in different asymptotic regimes.

A simple (and remarkably good) approximation of $\mathbb{P}(Q_e > B)$ is given by

\[ \mathbb{P}(Q_e > B) \approx \exp \left\{ - \inf_{s \geq 0} I(s, B|c, \varpi) \right\}; \]  \hspace{1cm} (1.18)

where

\[ I(s, B|c, \varpi) := \frac{1}{2} \frac{(B + (c - \varpi)s)^2}{v(s)}, \quad B, s > 0. \]  \hspace{1cm} (1.19)

It can be checked that for Brownian motion input, Relation (1.18) is exact. More details can be found in [57, 80] and the references therein.

For further reading on Gaussian processes, see [5, 6, 64], and for the use of Gaussian processes as input in queueing models, we refer to [4, 80] and the references therein.

### 1.4 Lévy processes

A Lévy process $X \equiv \{ X(t) : t \in \mathbb{R} \}$ is a stochastic process possessing the following properties:
1. Introduction

(i) \( X(0) = 0 \) with probability one,

(ii) \( X \) has independent increments, i.e., for \( s \leq t \), \( X(t) - X(s) \) is independent of \( \{X(u) : u \leq s\} \),

(iii) \( X \) has stationary increments, i.e., for \( s \leq t \), \( X(t) - X(s) \) is equal in distribution to \( X(t - s) \),

(iv) the sample-paths of \( X \) are almost surely right continuous with left limits.

Lévy processes are intimately related to the class of infinitely divisible distributions, which are completely characterized by the Lévy-Khintchine formula: a probability distribution \( F \) is infinitely divisible if and only if its characteristic exponent

\[
\Psi(\vartheta) := \log \int_{\mathbb{R}} e^{i\vartheta x} F(dx), \ \vartheta \in \mathbb{R},
\]

takes the following expression, known as the Lévy-Khintchine representation,

\[
\Psi(\vartheta) = i\delta \vartheta - \frac{1}{2} \sigma^2 \vartheta^2 + \int_{\mathbb{R}} \left( e^{i\vartheta x} - 1 - i\vartheta x 1_{|x|<1} \right) \Pi(dx).
\] (1.20)

Here \( \delta \in \mathbb{R}, \ \sigma^2 \geq 0 \) and \( \Pi \) is a Lévy measure, i.e., a measure concentrated on \( \mathbb{R} \setminus \{0\} \) satisfying

\[
\int_{\mathbb{R}} \min(1, x^2) \Pi(dx) < \infty.
\] (1.21)

A Lévy process \( X \) has the property that for all \( t \), \( \mathbb{E}e^{i\vartheta X(t)} = e^{t\Psi(\vartheta)} \), where \( \Psi(\cdot) \) is the characteristic exponent of \( X(1) \), which has an infinitely divisible distribution. Within the class of Lévy processes we distinguish the following three classes. If the measure \( \Pi \) gives no mass to the negative half line, i.e., \( \Pi(-\infty, 0) = 0 \), which means that the process has no negative jumps, we say that the Lévy process \( X \) is spectrally positive. On the other hand, if the measure \( \Pi \) is concentrated on \(( -\infty, 0) \), i.e., \( \Pi(0, \infty) = 0 \), which means that the process has no positive jumps, we say that the Lévy process \( X \) is spectrally negative. Lévy processes which have monotone sample-paths almost surely are called subordinators.

In this thesis special attention is paid to the class of spectrally-positive Lévy processes. For this class we usually use Laplace transforms instead of characteristic functions. We define the Laplace exponent

\[
\varphi(\vartheta) := \log \mathbb{E}e^{-\vartheta X(1)}, \ \vartheta \geq 0,
\]

which is given by

\[
\varphi(\vartheta) = \delta \vartheta + \frac{1}{2} \sigma^2 \vartheta^2 + \int_{(0, \infty)} \left( e^{-\vartheta x} - 1 + \vartheta x 1_{|x|<1} \right) \Pi(dx),
\] (1.22)
with $\delta$, $\sigma^2$ and $\Pi$ as in (1.20).

Examples of Lévy processes are Brownian motion (for which $\Pi \equiv 0$), the compound Poisson process (for which $\delta = \sigma^2 = 0$), and $\alpha$-stable Lévy processes with characteristic exponent given by

$$\Psi(\vartheta) = -|\vartheta|^\alpha \left(1 - i\beta \tan \left(\frac{\pi \alpha}{2}\right) \text{sgn}(\vartheta)\right),$$

(1.23)

where $\alpha \in (0, 1) \cup (1, 2)$ and $\beta \in [-1, 1]$. For further reading on Lévy processes, we refer to [20, 22, 55, 74, 106].

Results on the characterization of the Laplace transform of the stationary workload $Q_e$ for queues fed by spectrally one-sided input processes are due to Zolotarev [113]. However, in general it is not feasible to determine the steady-state distribution by explicit inversion of the Laplace transform. Therefore, one may then resort to study the asymptotics of $\mathbb{P}(Q_e > B)$ as $B \to \infty$. Under different assumptions on the moment generating function $\mathbb{E} e^{\vartheta X(1)}$ of $X(1)$ asymptotics of $\mathbb{P}(Q_e > B)$ as $B \to \infty$ have been determined. For heavy-tailed input, i.e., $\mathbb{E} e^{\vartheta X(1)} = \infty$ for all $\vartheta > 0$, we refer to for instance [13, 47]. For light-tailed input, i.e., $\mathbb{E} e^{\vartheta X(1)} < \infty$ for some $\vartheta > 0$, asymptotic results can be found in e.g. [21, 59].

1.5 Markov modulated fluid input processes

In this section we describe a Markov modulated fluid input process. Let $J \equiv \{J(t) : t \geq 0\}$ be an irreducible continuous-time Markov chain defined on a finite state space $\mathcal{E} = \{1, 2, \ldots, N\}$. Let $(\pi_i, i \in \mathcal{E})$ denote the stationary distribution of the Markov chain $J$. Further, let $\{r_i : i \in \mathcal{E}\}$ be a finite set of real numbers. The process $\{A(s, t) : 0 \leq s \leq t\}$ defined by

$$A(s, t) = \int_s^t r_{J(u)}du,$$  

(1.24)

is called a Markov modulated process. This type of input processes is widely used in, e.g., manufacturing and communication networks.

Under the stability condition, which for this class of input processes is given by

$$\sum_{i=1}^{N} r_i \pi_i < 0,$$

the steady-state distribution of $(Q_e, J_e)$ exists and can be explicitly derived. For more background on this model (and several variants) we refer to for instance [72, 103, 107].
1.6 Overview and contributions

In this section we give a short overview of the remainder of this thesis. In the next two chapters of this thesis we consider Gaussian queues, that is, queues fed by Gaussian processes, such as fractional Brownian motion (fBm) and the integrated Ornstein-Uhlenbeck (iOU) process. As already mentioned in Section 1.3, for these input processes no explicit expression for the distribution of the stationary workload is known. Hence it is unfeasible to analyze the covariance function $R(t)$ of the stationary version of the workload process (as this would even require the knowledge of $E Q(0) Q(t)$, the joint moment function). Therefore, we will analyze the dependence structure of the workload process by considering the metric $R(T|p, q)$ defined in (1.13). In Chapter 2, based on [53], we analyze the behavior of the metric $R(T|p, q)$ for $T \to \infty$ under the so-called many-sources scaling. That is, we assume that the input of the queue is an aggregation of $n$ i.i.d. Gaussian processes. To keep the queue stable we scale the service rate by $n$ and in addition we scale also the buffer with $n$. We denote the resulting workload process by $Q^n$. We focus on two special cases, viz. fBm and iOU. For large values of $T$, we study rough, logarithmic asymptotics of our dependence metric associated with $Q^n$ as $n \to \infty$. Relying on (the generalized version of) Schilder’s theorem, we are able to characterize its decay. The main result of this chapter is that, at least for the special cases considered, the dependence structure of the input process essentially carries over to the workload process (in the asymptotic regime that we have chosen and in terms of our specific notion of dependence, viz., the metric $R(T|p, q)$).

Chapter 3, based on [38], is devoted to the analysis of transient characteristics of Gaussian queues under the so-called large buffer regime. More specifically, we determine the logarithmic asymptotics of $\mathbb{P}(Q(0) > pB, Q(TB) > qB)$ as $B \to \infty$ and hence the logarithmic asymptotics of $R(TB|pB, qB)$ as $B \to \infty$ can also be determined. For any pair $(p, q)$ three regimes can be distinguished:

(A) For small values of $T$, either of the events $\{Q(0) > pB\}$ and $\{Q(TB) > qB\}$ will essentially imply the other. More specifically: if $p > q$ then the event $\{Q(TB) > qB\}$ ‘comes for free’, and if $q > p$, then the event $\{Q(0) > pB\}$ ‘comes for free’.

(B) There is an intermediate range of values of $T$ for which it is to be expected that both $\{Q(0) > pB\}$ and $\{Q(TB) > qB\}$ are ‘tight’ (in the sense that none of them implies the other with overwhelming probability), but that the time epochs 0 and $T$ tend to lie in the same busy period.

(C) Finally, for large values of $T$ still both events are ‘tight’, but now they occur in different busy periods with overwhelming probability.
For the short-range dependent case explicit calculations are presented, whereas for the long-range dependent case structural results are proven.

The Lévy-driven queue, that is, a queue with Lévy input, is studied in Chapters 4 and 5. In Chapter 4, which is based on [51], we consider a queue fed by a spectrally positive Lévy process. For this class of Lévy processes the Laplace transform of the stationary workload is known, see Zolotarev [113]. Using this result and a result taken from [65], we are able to derive the Laplace transform

$$
\rho(\vartheta) = \int_0^\infty r(t)e^{-\vartheta t}dt
$$

of the correlation function $r(t)$ of the stationary workload process $\{Q(t) : t \geq 0\}$. This expression allows us to prove structural properties of $r(\cdot)$. More specifically, we prove that the correlation function is positive, decreasing, and convex, relying on the machinery of completely monotone functions. We also show that $r(\cdot)$ can be represented as the complementary distribution function of a specific random variable. These results are used to compute the asymptotics of $r(t)$, for $t$ large, for the cases of light-tailed and heavy-tailed Lévy input.

In Chapter 5, based on [37], we consider a queue fed by a general Lévy process. In this case we will study the metric $R(T_B|p_B, q_B)$ for various types of functions $T_B$. The main focus will be on refined exact asymptotics (rather than rough logarithmic asymptotics) of rare event probabilities of the type

$$
P(Q(0) > p_B, Q(T_B) > q_B),
$$

for given positive numbers $p, q$, and a positive deterministic function $T_B$. The following contributions are then made.

- We first identify conditions on the function $T_B$ under which the probability of interest is dominated by the ‘most demanding event’, in the sense that it is asymptotically equivalent to $P(Q_e > \max\{p, q\}B)$ for $B$ large. These conditions essentially reduce to $T_B$ being sublinear (i.e., $T_B/B \to 0$ as $B \to \infty$).

- A second condition on $T_B$ is derived under which the probability of interest essentially ‘decouples’, in that $P(Q(0) > pB, Q(T_B) > qB)$ is asymptotically equivalent to $P(Q_e > pB)P(Q_e > qB)$ for $B$ large. For various models considered in the literature this ‘decoupling condition’ reduces to requiring that $T_B$ is superlinear (i.e., $T_B/B \to \infty$ as $B \to \infty$). This is not true for certain ‘heavy-tailed’ cases, for instance the situations in which the Lévy input process corresponds to an $\alpha$-stable process, or to a compound Poisson process with regularly varying job sizes, in which the ‘decoupling condition’ reduces to $T_B/B^2 \to \infty$. For these input processes we also establish the asymptotics of the probability under consideration for $T_B$ increasing superlinearly but subquadratically.

Moreover, for light-tailed input, special attention is paid to the case where $T_B$ is a linear function in $B$, that is we suppose that $T_B = RB$, for some $R > 0$. We derive intuitively appealing asymptotics, by intensively relying on sample-path large deviations results. The regimes obtained in this case, can be interpreted in terms of most
likely paths to overflow.

Chapter 6 is based on [52] and deals with a Markov-fluid-driven queue, that is, a queue fed by a Markov modulated process. For this model we consider two important metrics, viz., the busy period of the system and the covariance function of the stationary workload process, which we capture in terms of their Laplace transforms. Relying on sample-path large deviations, we identify the logarithmic asymptotics of the probability that the busy period lasts longer than $t$, as $t \to \infty$.

In Chapter 7 we consider a discrete-time model with a general input process $X$ having stationary increments. More specifically, we consider the workload process of a queue operating in discrete time, focusing on the (multivariate) distribution of the workloads at different points in time. In a many-sources framework exact asymptotics are determined, relying on large deviations results for the sample means of multivariate random variables.

Each chapter of this monograph can be read independently of the other chapters, that is, all chapters are essentially self-contained. Furthermore, every chapter starts with an introduction where a historical account of the literature is given. For the key quantities we have used a consistent notation.