# Large Deviations for Gaussian Queues

## Modelling Communication Networks

**Michel Mandjes**
*Korteweg-de Vries Institute for Mathematics,*
*University of Amsterdam, The Netherlands*

BICENTENNIAL
1807
WILEY
2007
BICENTENNIAL

John Wiley & Sons, Ltd

# Large Deviations for Gaussian Queues

# Large Deviations for Gaussian Queues

## Modelling Communication Networks

**Michel Mandjes**
*Korteweg-de Vries Institute for Mathematics,*
*University of Amsterdam, The Netherlands*

BICENTENNIAL
1807
WILEY
2007
BICENTENNIAL

John Wiley & Sons, Ltd

Copyright © 2007 John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester,
West Sussex PO19 8SQ, England

Telephone (+44) 1243 779777

Email (for orders and customer service enquiries): cs-books@wiley.co.uk
Visit our Home Page on www.wileyeurope.com or www.wiley.com

This publication is designed to provide accurate and authoritative information in regard to the subject matter covered. It is sold on the understanding that the Publisher is not engaged in rendering professional services. If professional advice or other expert assistance is required, the services of a competent professional should be sought.

***Other Wiley Editorial Offices***

John Wiley & Sons Inc., 111 River Street, Hoboken, NJ 07030, USA

Jossey-Bass, 989 Market Street, San Francisco, CA 94103-1741, USA

Wiley-VCH Verlag GmbH, Boschstr. 12, D-69469 Weinheim, Germany

John Wiley & Sons Australia Ltd, 42 McDougall Street, Milton, Queensland 4064, Australia

John Wiley & Sons (Asia) Pte Ltd, 2 Clementi Loop #02-01, Jin Xing Distripark, Singapore 129809

John Wiley & Sons Canada Ltd, 6045 Freemont Blvd, Mississauga, Ontario, L5R 4J3, Canada

Wiley also publishes its books in a variety of electronic formats. Some content that appears
in print may not be available in electronic books.

Anniversary Logo Design: Richard J. Pacifico

***British Library Cataloguing in Publication Data***

A catalogue record for this book is available from the British Library

ISBN-13: 978-0-470-01523-0

Typeset in 10/12 Times by Laserwords Private Limited, Chennai, India
Printed and bound in Great Britain by TJ International, Padstow, Cornwall
This book is printed on acid-free paper responsibly manufactured from sustainable forestry
in which at least two trees are planted for each one used for paper production.

# Contents

# Preface and acknowledgments

In the spring of 2001, Gaussian queues started to attract my attention. At that moment, I was working on a number of problems on networks of queues as well as queues operating under nonstandard scheduling disciplines–which turned out to be substantially harder to analyze than the classical single first-in first-out queue. What a mathematician does when he wishes to analyze this kind of hard problem is, to find a modelling framework that is 'clean' enough to enable (to some extent) explicit solutions, yet general enough to cover all interesting scenarios (in my case, all relevant arrival processes including the important class of long-range dependent inputs). At some point I came across a series of papers by Petteri Mannersalo and Ilkka Norros (VTT Research, Espoo, Finland) on Gaussian queues, and it turned out that the Gaussian input model they considered combined these attractive properties. I started to learn more about it, and it led to a very nice collaboration with Petteri and Ilkka (in the mean time, their annual spring-visit to Amsterdam has almost become a tradition...).

Coincidentally, more or less at the same time Krzysztof Dębicki took a leave of absence from the University of Wrocław, Poland, and became a colleague of mine at CWI in Amsterdam. He had been working already for a while on Gaussian queues, and I think it is fair to consider him as a real expert on their tail asymptotics. It was the start of a fruitful and very pleasant collaboration. Needless to say that this book would not have been the same without Petteri, Ilkka, and Krzys, and I would like to thank them for this.

The idea of writing a book on Gaussian queues came up in 2004. By then, I had been involved in a series of papers on the large deviations of Gaussian queues, and started to realize that they make up a nice and coherent body of theory. The Gaussian machinery proved to be a powerful tool, and I am convinced that it will be extremely useful for many other queuing problems as well. During the last years of research, my attention shifted somewhat from the asymptotics of Gaussian queues to their applications in communication networks; due to the generality and versatility of the Gaussian paradigm, it enabled the development of effective guidelines for several highly relevant networking problems.

To illustrate the strengths of the concept of Gaussian queues, I have laid much emphasis on applications. The book contains four application-related chapters, in which Quality-of-Service differentiation (through Generalized Processor Sharing), link dimensioning, and bandwidth trading are dealt with.

I have attempted to make the book as much self-contained as possible, so that it becomes accessible for a broader audience; unfortunately, it turned out that it is nearly impossible to write the book in such a way that it does not require any prior knowledge on standard probabilistic concepts.

The book could serve as a textbook for undergraduate students in mathematics (and 'mathematics-related' disciplines); as argued above, a rigorous background in probability is required. The book also targets graduate students in engineering, in particular, computer science and electrical engineering with interest in networking applications. It could also serve as a reference book for senior researchers both in academia and at telecommunication labs – particularly the application-part may be interesting to practice-oriented scientists.

My PhD students Abdelghafour Es-Saghouani, Pascal Lieshout, and Frank Roijers read earlier drafts of the manuscript, and provided me with many helpful suggestions. Much of the material presented in this book was drawn from my papers on Gaussian queues. These papers were not just my work – I would like to thank, besides the names I already mentioned above, the other co-authors: Hans van den Berg, Nam Kyoo Boots, Sem Borst, Ton Dieker, Krishnan Kumaran, Aiko Pras, Maya Ramakrishnan, Aleksandr Stolyar, Miranda van Uitert, Pieter Venemans, and Alan Weiss.

The first draft of this book was written while I had a 1-month leave at the University of Melbourne, Australia. This place proved to be a wonderful working environment, and I would like to thank Peter Taylor for hosting me, and Tony Krzesinski (University of Stellenbosch, South Africa) for being the ideal office mate during this period. I would also like to thank the publisher, Wiley Chichester, for its support – in particular Susan Barclay, Kelly Board, Wendy Hunter, Simon Lightfoot, and Kathryn Sharples. I should also mention Laserwords, Chennai, and in particular Gopika, who were responsible for the typesetting job (and I realize that I have probably been the most demanding author ever).

Work presented in this book has been partly carried out in the framework of the projects Equip (in the programme 'Netwerken', funded by the Netherlands Organization for Scientific Research, NWO), Logica (in the programme 'Open Competitie', also funded by NWO), Equanet (in the programme 'ICT Doorbraak', funded by Dutch Ministry of Economic Affairs through its agency Senter/Novem), and Bricks (in the programme 'Bsik', funded by several Dutch ministries).

I dedicate this book to my wife Miranda, for her continuous love and support, and our daughter Chloe.

Amsterdam, November 2007,
*Michel Mandjes*

# Chapter 1

# Introduction

*Performance Analysis, Queuing Theory, Large Deviations*. Performance analysis of communication networks is the branch of applied probability that deals with the evaluation of the level of efficiency the network achieves, and the level of (dis)satisfaction enjoyed by its users. Clearly, there is a broad variety of measures that characterize these two aspects. Focusing on the efficiency of the use of network resources, one could think of the throughput, i.e., the rate at which the network effectively works–in the case of a single network element, this could be the rate (in terms of, say, bits per second) at which traffic leaves. Another option is to use a relative measure, such as the utilization, commonly defined as the ratio of the throughput and the available service speed of the network element. Also the (dis)utility experienced by users can be expressed by a broad variety of measures. Realizing that at any network element traffic can be stored in a buffer when the input rate temporarily exceeds the available service rate, it seems justified to study performance indicators that describe the delay incurred when passing the network node. Buffers have a finite size, so there is the possibility of losing traffic, and as a result the fraction of traffic lost becomes a relevant metric.

Performance analysis is a probabilistic discipline, as the main underlying assumption is that user behavior is inherently *random*, and therefore described by a statistical model. This statistical model defines the probabilistic properties of the arrival process (or, input process) of traffic at the network. Traffic could arrive in a smooth way, but highly irregular patterns also occur; in the latter case, communication engineers call the arrival process *bursty*.

Justified by the above description of network elements as storage systems, we could model a communication network as a network of *queues*; at any node traffic arrives, is stored if it cannot be handled immediately, and is served. Performance analysis often relies heavily on results from the theory that describes the performance of these queues, i.e., *queuing theory*. A key element of performance analysis

*Large deviations for Gaussian queues* M. Mandjes

is the characterization of the impact of ‘user parameters’ on the performance offered by the network (how is the delay affected by the arrival rate? what is the impact of increased variability of the input traffic? etc.). On the other hand, one often studies the sensitivity of the performance in the system parameters (what is the impact of the buffer size on the loss probability? how does the service speed affect the mean delay? etc.)

A substantial part of the defined performance metrics relates to *rare events*. Often network engineers have the target to design the system such that the loss probability is below, say, $10^{-6}$. Another common objective is that the probability that the delay is larger than some predefined excessive value is of the same order. This explains why we heavily rely on a subdomain of probability theory that exclusively focuses on the analysis of rare events: *large deviations theory*. This theory has a long history, but has been applied intensively for performance analysis purposes only during the last, say, two decades.

*Traffic management, dimensioning*. Once one is capable of evaluating the performance of a static situation (i.e., calculating performance metrics for a given arrival process and given network characteristics), the next step is often to choose the set of design parameters such that a certain condition is met, or such that some objective function is optimized. For instance, a requirement imposed upon the network element could be that just (on average) a fraction $\epsilon$ of the incoming traffic is lost. Evidently, when increasing the buffer size $\textsc{b}$, the loss probability decreases, and therefore it is legitimate to ask for which minimal $\textsc{b}$ the loss probability is at most $\epsilon$. Of course, there is often a cost incurred when increasing $\textsc{b}$. As a result one could imagine that one should maximize an objective function that consists of a ‘utility part’, minus a ‘cost part’, where both parts increase in $\textsc{b}$. Selecting an appropriate value for $\textsc{b}$ is usually called *buffer dimensioning*; similarly the choice of a suitable service speed is referred to as *link rate dimensioning* (or, shortly, link dimensioning).

On the other hand, knowledge of the static situation enables the computation of conditions on the arrival process (both in terms of average input traffic rate and the variability of the arrival process) under which the network can offer some required performance level:

- In this way, one could develop mechanisms that decide what the maximum number of users is such that the mean delay stays within some predefined bound; such a mechanism is usually called *admission control*. To implement an admission control, one needs to be able to characterize the so-called admissible region, which is, in a situation of two classes of users, the combination of all numbers of users of both classes $(n_1, n_2)$ for which for both classes the performance requirement is met.

- Also, insight into the static situation may tell us how to ‘smooth’ traffic (i.e., decrease the variability of the arrival process), such that the traffic stream becomes more ‘benign’, and the loss probability in some target queue can meet some set requirement (a technique known as *traffic shaping*). Traffic shaping is usually done by inserting an additional queue between the traffic sources and the target queue that is emptied at a service rate $c'$ that is lower than the peak rate of the original stream (but higher than the service rate $c$ of the target queue). Then the traffic stream arriving at the second queue is smoother than the original traffic stream, and therefore easier to handle, but this is at the expense of introducing additional delay.

This traffic shaping example explains the interest in *tandem queues*, i.e., systems of queues in series (in which the output of the first queue feeds into the second queue). In such a situation one would, for instance, like to dimension the shaping rate: given a buffer $b$ and service rate $c$ in the target queue, how should one choose the shaping rate $c'$ to ensure that the loss probability in the target queue is below $\epsilon$ (where it assumed that the shaper queue has a relatively big buffer).

In the literature, the set of control measures that affect the network's efficiency or the user's (dis-)satisfaction is often called *traffic management*. Clearly, dimensioning is a traffic management action that relates to a relatively long timescale: one can choose a new value for the buffer size or the link rate only at a very infrequent rate; the process of updating the resource capacities is known as the *planning cycle*. Mechanisms like admission control serve to control fluctuations of the offered traffic at a relatively short timescale: admission control is done on the timescale that new users arrive (and hence the decision to accept or reject a new user has to be done essentially in real time).

*Performance differentiation*. We have described above the situation in which we wished to guarantee some performance requirement that is uniform across users; for instance, all users should be offered the same maximum loss probability. In practice, however, all applications have their own specific performance requirements. Think of a voice user, who tolerates a substantial amount of loss (up to the order of a few percents, if certain codecs are used) but whose delay is critical, versus a data user, who has very stringent requirements with respect to loss, but is less demanding with respect to delay. Of course one could treat all traffic in the same fashion, e.g., by using first-in-first-out (FIFO) queues; clearly, to meet the performance requirements of all users, the requirement of the most stringent users should be satisfied. Such an approach will, however, inevitably lead to a waste of resources, and therefore one has developed queuing disciplines that actively discriminate. An example of such a scheme is the (two-class) *priority queue*, in which one class has strict priority over another class. The high-priority class does not 'see' the low-priority class, so its performance can be evaluated as in the FIFO case. The low-priority class, however, sees a fluctuating service capacity, and therefore its performance is considerably harder to analyze.

Strict priority has the intrinsic drawback of 'starvation', i.e., the low-priority class can be excluded from service for relatively long periods of time (namely, the periods in which the high-priority class uses all the bandwidth). To avoid this starvation effect, one could guarantee the low-priority class at least some minimal service rate. This thought led to the idea of *generalized processor sharing* (GPS). In GPS, both classes have their own queue. Class $i$ can always use a fraction $\phi_i$ of the total service rate $C$ (where $\phi_1 + \phi_2 = 1$). If one of the classes does not use its full capacity, then the remaining capacity is allocated to the other class (thus making the service discipline work conserving). Note that the priority queue is a special case of GPS (choose $\phi_i = 1$ to give class $i$ strict priority over the other class). One of the crucial engineering questions here is, for two user classes with given traffic arrival processes and performance targets, how should the weights be set?

*Scope of this book.* In view of the above, one could say that traffic management is all about the interrelationship between

(N) the network traffic offered (not only in terms of the average imposed load, but also in terms of its fluctuations, summarized in a certain arrival process);

(R) the amount of network resources available (link capacity, buffers, etc.);

(P) the performance level achieved.

With this interrelationship in mind, we conclude that there are three indispensable prerequisites for appropriate traffic management.

In the first place, we should have accurate traffic models at our disposal (i.e., N). Part A of this book is devoted to a class of models that has proven to be suitable in the context of communication networks: Gaussian traffic processes. An interesting feature of this class is that it is highly versatile, as it covers a broad class of correlation structures. We introduce this class and provide a number of generic properties. Then we explain why Gaussian models are likely to be an adequate statistical descriptor, and how this can be empirically verified. We also present a number of standard Gaussian models that are used throughout the book.

Secondly, we show in part B how to assess the performance of the network, for a given Gaussian traffic model, and for given amounts of available resources (i.e., (N,R) $\mapsto$ P). In other words, we analyze Gaussian queues, i.e., queues with Gaussian input. It turns out that only for a very limited subclass of inputs exact analysis is possible, and this explains why we resort to asymptotics. We present and explain several asymptotic results. Emphasis is on the so-called many-sources framework, which is an asymptotic regime in which the number of users grows large (where the traffic streams generated by these users have more or less similar statistical properties), and where the resources are scaled accordingly. Single queues are relatively easy to deal with in this framework, but we also focus on problems that are significantly harder, such as the analysis of a tandem queue, and a queue operating under GPS.

The final subject of the book is how these Gaussian queues can be used for traffic management purposes. Essentially, these problems all amount to questions of the type (N,P) $\mapsto$ R: given a traffic model and some performance target, how much resources are needed? Specific attention will be paid to link dimensioning in the single queue, the weight setting problem in generalized processor sharing, and bandwidth trading.

## Bibliographical notes

This book focuses on large deviations for Gaussian queues, with applications to communication networking. There is a vast body of related literature, which we will cite at several occasions. Here, we briefly list a number of textbooks that can be used as background.

The literature on performance analysis is vast, and the key journals include *IEEE/ACM Transactions on Networking*, *Computer Networks*, and *Performance Evaluation*. A textbook that gives an excellent survey on performance evaluation techniques is by Roberts *et al.* [253], albeit with a focus on somewhat out-of-date technologies. We also recommend the book by Kurose and Ross [167], and the classical book by Bertsekas and Gallager [32].

There are several strong textbooks on queuing theory – without attempting to provide an exhaustive list, here we mention the books by Baccelli and Brémaud [17], Cohen [52], Prabhu [246], and Robert [250]. The beautiful survey by Asmussen [13] deserves some special attention, as it gives an excellent account of the state of the art on many topics in queuing theory. The leading journal in queuing is *Queueing Systems*, but there are many nice articles scattered over several other journals (including *Advances in Applied Probability*, *Journal of Applied Probability* and *Stochastic Models*).

During the last two decades a number of books on large deviations appeared with a focus on applications in performance and networking. In this context we mention the book by Bucklew [42] as a nice introduction to large deviations and the underlying intuition. The book by Shwartz and Weiss [267] is technically considerably more demanding, but the reader's efforts pay off when working through a beautiful series of appealing examples. Interestingly, Chang [46] connects deterministic network calculus methods with large deviations techniques. The book that is perhaps most related to the present book is Ganesh, O'Connell, and Wischik [109]. Also there the emphasis is on the application of large-deviations techniques in a queuing setting, albeit without focusing on Gaussian inputs, and without applying it (explicitly) in a communication networks context.

Apart from these books, there are a number of books on large deviations, but without a focus on queuing. Ellis [91] approaches large deviations from the angle of statistical mechanics, whereas in Dupuis and Ellis [87] control-theoretic elements appear. Perhaps the most complete, rigorous introductory book is by Dembo and Zeitouni [72]. Other useful textbooks include Deuschel and Stroock [75] and den Hollander [132]. Articles on large deviations appear in a broad variety of journals; besides the Applied Probability journals mentioned above, this also includes *Stochastic Processes and their Applications* and *Annals of Applied Probability*.

# Part A: Gaussian traffic and large deviations

The first part of this book is of an introductory nature. It defines the basic concepts used throughout the book, by focusing on two topics. First, we introduce the notion of Gaussian traffic, and we argue why, under rather general circumstances, the Gaussian model offers an accurate description of network traffic. The second topic is large deviations: our main tool to probabilistically analyze rare events.

*Large deviations for Gaussian queues* M. Mandjes

# Chapter 2

# The Gaussian source model

In the introduction we argued that the traffic model is one of the cornerstones of performance evaluation of communication networks: it is a crucial building block in queuing models. This chapter introduces the class of traffic models that is studied in this book: the Gaussian traffic model. The main goal is to (qualitatively) argue why we feel that the Gaussian model is particularly suitable for modeling traffic streams in communication networks.

In Section 2.1, we identify a number of ‘desirable properties’ that a network traffic model should obey. After introducing some notation and some preliminaries on Gaussian random variables in Section 2.2, we define the Gaussian traffic model in Section 2.3. A few generic examples of Gaussian inputs are described in Section 2.4, viz. fractional Brownian motion and integrated Ornstein–Uhlenbeck; in this section also, a number useful concepts are introduced, most notably that of long-range dependence. Section 2.5 describes a number of other Gaussian source models that are used throughout the book; they are the Gaussian counterpart of ‘classical’ input models. Finally in Section 2.6 we return to the (qualitative) requirements stated in Section 2.1, and show that some Gaussian models are very well in line with these.

## 2.1 Modeling network traffic

There is a huge collection of traffic processes available from the literature, each having its own features. In general, an arrival process is an infinitely dimensional object $(A(t), t \in \mathbb{R})$, where $A(t)$ denotes the amount of traffic generated in time interval $[0, t)$, for $t > 0$; $(A(t), t \in \mathbb{R})$ is sometimes referred to as the *cumulative work process*. It is noted that $A(-t)$ is to be interpreted as the negative of the amount of traffic generated in $(-t, 0]$. Usually it is assumed that an arrival process

*Large deviations for Gaussian queues* M. Mandjes

is nondecreasing (as traffic cannot be negative); later we argue that this assumption is merely a technicality, and that it is, in many situations, not necessary. We also define, for $s < t$, the work that has arrived in time window $[s, t)$ as $A(s, t)$ (so that $A(s, t) = A(t) - A(s)$).

It is clear that some arrival processes fit better to 'real' network traffic than others. In this section, we list a number of properties that network traffic usually obeys. In the sequel, we use these properties as the motivation for our choice of Gaussian traffic models.

1. *Stationarity*. In quite general circumstances, it can be assumed that the traffic arrival process is stationary, at least over periods up to, say, one or more hours (over longer periods this usually does not hold–during a period of a day we often see the common diurnal pattern). In mathematical terms this stationarity means that the cumulative process $(A(t), t \in \mathbb{R})$ has *stationary increments*, i.e., in distribution,

$$A(s, t) = A(0, t - s) \quad \text{for all} \quad s < t.$$

In other words, the distribution of $A(s, t)$ is determined only by the *length* of the corresponding time interval (i.e., $t - s$), and not by the *position* of the interval.

2. *High aggregation level*. A common feature in modern communication networks is that the input stream of each node usually consists of the superposition of a large number of individual streams. It is clear that this is the case at the core links of networks, where the resources are shared by thousands of users, but even at the access of communication networks, the contributions of a substantial number of users is aggregated (at least in the order of tens).

Besides the fact that one can safely focus on relatively high aggregates, one can also argue that the behavior of a substantial part of the user population can be assumed homogeneous (i.e., obeying the same statistical law), as many users run the same applications over the network. In any case, one can usually subdivide the user population into a number of classes, within which the users are (nearly) homogeneous.

3. *Extreme irregularity of the traffic rate*. Measurements often indicate that the traffic rate exhibits extreme irregularity at a wide variety of timescales, including very small timescales. This 'burstiness' could be modeled by imposing the requirement that the instantaneous traffic rate process

$$R(t) := \lim_{s \uparrow t} \frac{A(s, t)}{t - s}$$

behaves irregularly (it could have nondifferentiable trajectories, for instance).

4. *Strong positive correlations on a broad range of timescales*. In the early 1990s, measurements, most notably those performed at Bellcore and at AT&T, gave the impression that network traffic exhibits significant positive correlation

on a broad range of timescales. A long series of measurement studies, performed at various networking environments, followed. These studies yielded convincing empirical evidence that network traffic is, under rather general circumstances, *long-range dependent*. We choose to postpone giving a precise definition of this notion, but on an intuitive level it means that the variance of $A(t)$ grows superlinearly in $t$. Traditional 'Markovian' traffic models had the implicit underlying assumption that this variance grows, at least for $t$ large, linearly, and hence those models could not be used anymore.

## 2.2 Notation and preliminaries on Gaussian random variables

Let us give a few remarks on the notation used throughout this book, which we have tried to keep as light as possible. Here we introduce a number of general concepts.

- We denote the mean and variance of a random variable $X$ by $\mathbb{E}X$ and $\mathbb{V}\mathrm{ar}X$, respectively; recall that $\mathbb{V}\mathrm{ar}X = \mathbb{E}(X^2) - (\mathbb{E}X)^2$. Also,

  $$\mathbb{C}\mathrm{ov}(X, Y) := \mathbb{E}(XY) - (\mathbb{E}X)(\mathbb{E}Y)$$

  denotes the covariance between $X$ and $Y$.

- $X =_{\mathrm{d}} Y$ indicates that the random variables $X$ and $Y$ are equally distributed.

- $X =_{\mathrm{d}} \mathcal{N}(\mu, \sigma^2)$ means that $X$ is normally distributed with mean $\mu$ and variance $\sigma^2$. In other words, $X$ has density

  $$f_{\mu,\sigma^2}(x) := \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left(\frac{(x-\mu)^2}{\sigma^2}\right)\right).$$

  We often use the notation $\Phi_{\mu,\sigma^2}(x)$ for the distribution function of a $\mathcal{N}(\mu, \sigma^2)$-distributed random variable:

  $$\Phi_{\mu,\sigma^2}(x) := \int_{-\infty}^{x} f_{\mu,\sigma^2}(y)\,\mathrm{d}y.$$

  This distribution function can be rewritten in terms of the distribution function of a standard normal random variable:

  $$\Phi_{\mu,\sigma^2}(x) = \mathbb{P}\left(\mathcal{N}_{\mu,\sigma^2} < x\right) = \mathbb{P}\left(\frac{\mathcal{N}_{\mu,\sigma^2} - \mu}{\sigma} < \frac{x-\mu}{\sigma}\right) = \Phi_{0,1}\left(\frac{x-\mu}{\sigma}\right).$$

  Throughout the book, we often abbreviate $\Phi_{0,1}(\cdot)$ as $\Phi(\cdot)$.

- We say that $X$ has a multivariate normal distribution (of dimension $d \in \mathbb{N}$) with $d$-dimensional mean vector $\mu$ and (nonsingular) $d \times d$ covariance matrix $\Sigma$ if $X$ has density

  $$f(x_1, \ldots, x_d) = \frac{1}{\sqrt{2\pi \det(\Sigma)}} \exp\left(-\frac{1}{2}(x - \mu)\mathrm{T}\Sigma^{-1}(x - \mu)\right),$$

  where $\det(\Sigma)$ denotes the (non-zero) determinant of the matrix $\Sigma$.

- In the setting of multivariate normal random variables, conditional distributions can be given through elegant formulas. Consider for ease the case of $(X, Y)$ being bivariate normal. The random variable $(X \mid Y = y)$, for some $y \in \mathbb{R}$ is then (univariate) normal with mean

  $$\mathbb{E}(X \mid Y = y) = \mathbb{E}X + \frac{\mathbb{C}\mathrm{ov}(X, Y)}{\mathbb{V}\mathrm{ar}Y} \cdot (y - \mathbb{E}Y) \tag{2.1}$$

  and variance

  $$\mathbb{V}\mathrm{ar}(X \mid Y = y) = \mathbb{V}\mathrm{ar}X - \frac{(\mathbb{C}\mathrm{ov}(X, Y)^2}{\mathbb{V}\mathrm{ar}Y}; \tag{2.2}$$

  notice that, interestingly, the conditional variance does not depend on the condition (i.e., the value $y$).

## 2.3 Gaussian sources

Now that we have, from Section 2.1, a list of properties that our traffic model should satisfy, this section introduces a versatile class of arrival processes: the so-called Gaussian sources. The aim of the remainder of this chapter is to show that certain types of Gaussian sources meet all the properties identified above.

For a Gaussian source, the entire probabilistic behavior of the cumulative work process can be expressed in terms of a mean traffic rate and a variance function. The mean traffic rate $\mu$ is such that $\mathbb{E}A(s, t) = \mu \cdot (t - s)$, i.e., the amount of traffic generated is proportional to the length of the interval. The variance function $v(\cdot)$ is such that $\mathbb{V}\mathrm{ar}A(s, t) = v(t - s)$; in particular $\mathbb{V}\mathrm{ar}A(t) = v(t)$.

**Definition 2.3.1 Gaussian source.** *$A(\cdot)$ is a Gaussian process with stationary increments, if for all $s < t$,*

$$A(s, t) =_{\mathrm{d}} \mathcal{N}(\mu \cdot (t - s), v(t - s)).$$

*We say that $A(\cdot)$ is a* Gaussian source. *We call a Gaussian source* centered *if, in addition, $\mu = 0$.*

The fact that the sources introduced in Definition 2.3.1 have stationary increments is an immediate consequence of the fact that the distribution of $A(s,t)$ just depends on the length of the time window (i.e., $t-s$), and not on its position.

The variance function $v(\cdot)$ fully determines the correlation structure of the Gaussian source. This can be seen as follows. First notice, assuming for ease $0<s<t$, that $\mathbb{C}\mathrm{ov}(A(s),A(t)) = \mathbb{V}\mathrm{ar}A(s) + \mathbb{C}\mathrm{ov}(A(0,s),A(s,t))$. Then, using the standard property that

$$\mathbb{V}\mathrm{ar}A(0,t) = \mathbb{V}\mathrm{ar}A(0,s) + 2\,\mathbb{C}\mathrm{ov}(A(0,s),A(s,t)) + \mathbb{V}\mathrm{ar}A(s,t),$$

we find the useful relation

$$\Gamma(s,t) := \mathbb{C}\mathrm{ov}(A(s),A(t)) = \frac{1}{2}(v(t)+v(s)-v(t-s)).$$

Indeed, knowing the variance function, we can compute all covariances. In particular $(A(s_1),\ldots,A(s_d))\mathrm{T}$ is distributed $d$-variate normal, with mean vector $(\mu s_1,\ldots,\mu s_d)\mathrm{T}$ and covariance matrix $\Sigma$, whose $(i,j)$th entry reads

$$\Sigma_{ij} = \Gamma(s_i,s_j), \quad i,j=1,\ldots,d.$$

The class of Gaussian sources with stationary increments is extremely rich, and this intrinsic richness is best illustrated by the multitude of possible choices for the variance function $v(\cdot)$. In fact, one could choose any function $v(\cdot)$ that gives rise to a positive semi-definite covariance function:

$$\sum_{s,t\in S} \alpha_s \mathbb{C}\mathrm{ov}(A(s),A(t))\alpha_t \geq 0,$$

for all $S \subseteq \mathbb{R}$, and $\alpha_s \in \mathbb{R}$ for all $s \in S$. The following example shows a way to find out whether a specific function can be a variance function.

**Exercise 2.3.2** Consider increasing variance functions $v(\cdot)$ of the type $v(t)=t^\alpha$. Prove that necessarily $\alpha \in (0,2]$.

**Solution.** Notice that the fact that $v(\cdot)$ be increasing yields that $\alpha > 0$. The fact that $v(\cdot)$ should correspond to a positive semi-definite covariance function rules out $\alpha > 2$. Cauchy–Schwartz inequality implies that the correlation coefficient lies between $-1$ and 1; when applying this to $A(0,t)$ and $A(t,2t)$ we obtain

$$\frac{\mathbb{C}\mathrm{ov}(A(0,t),A(t,2t))}{\mathbb{V}\mathrm{ar}A(0,t)} \in [-1,1],$$

or, after simple algebraic manipulations, $v(2t)/2v(t) \in [0,2]$. Now inserting $v(t)=t^\alpha$ immediately leads to $\alpha \leq 2$.

In fact, the above reasoning indicates that $\alpha \in (1, 2)$ corresponds to positive correlation, where the extreme situation $\alpha = 2$ can be regarded as 'perfect positive correlation' ($A(ft) = fA(t)$, for some $f \geq 0$); then the correlation coefficient is 1. On the other hand, $\alpha = 1$ relates to the complete lack of correlation (i.e., $\mathbb{C}\mathrm{ov}(A(0, s), A(s, t)) = 0$, with $s < t$); in this case the correlation coefficient equals 0. Finally, if $\alpha \in (0, 1)$, then there is negative correlation. ◇

## 2.4 Generic examples–long-range dependence and smoothness

This section highlights two basic classifications of Gaussian sources. These classifications can be illustrated by means of two generic types of Gaussian sources, which we also introduce in this section and which will be used throughout the book. The first directly relates to Exercise 2.3.2.

**Definition 2.4.1 Fractional Brownian motion (or fBm).** *A* fractional Brownian motion *source has variance function $v(\cdot)$ characterized by $v(t) = t^{2H}$, for an $H \in (0, 1)$. We call $H$ the* Hurst parameter.

The case with $H = 1/2$ is known as (ordinary) *Brownian motion.* In this case it is well known that the increments are independent, which is in line with the lack of correlation observed in Exercise 2.3.2.

**Definition 2.4.2 Integrated Ornstein–Uhlenbeck (or iOU).** *An* integrated Ornstein–Uhlenbeck *source has variance function $v(\cdot)$ characterized by $v(t) = t - 1 + e^{-t}$.*

*Long-range dependence.* The first way of classifying Gaussian sources relates to the correlation structure on long timescales: we are going to distinguish between srd sources and lrd sources.

To this end, we first introduce the notion of correlation on timescale $t$, for intervals of length $\epsilon$. With $t \gg \epsilon > 0$, it is easily seen that

$$\mathbb{C}(t, \epsilon) := \mathbb{C}\mathrm{ov}(A(0, \epsilon), A(t, t + \epsilon)) = \frac{1}{2}(v(t + \epsilon) - 2v(t) + v(t - \epsilon)).$$

For $\epsilon$ small, and $v(\cdot)$ twice differentiable, this looks like $\epsilon^2 v''(t)/2$. This argument shows that the 'intensity of the correlation' is expressed by the second derivative of $v(\cdot)$: 'the more convex (concave, respectively) $v(\cdot)$ at timescale $t$, the stronger the positive (negative) dependence between traffic sent 'around time 0' and traffic sent 'around time $t$'.

The above observations can be illustrated by using the generic processes fBm and iOU. As $v''(t) = (2H)(2H - 1)t^{2H-2}$, we see that for fBm the correlation is positive when $H > \frac{1}{2}$ (the higher the $H$, the stronger this correlation; the larger the $t$, the weaker this correlation), and negative when $H < \frac{1}{2}$ (the lower the $H$, the stronger this correlation; the larger the $t$, the weaker this correlation), in line with the findings of Exercise 2.3.2. It is readily checked that for iOU $v''(t) = e^{-t}$. In other words: the correlation is positive, and decreasing in $t$.

Several processes could exhibit positive correlation, but the intensity of this correlation can vary dramatically; compare the (fast!) exponential decay of $v''(t)$ for iOU traffic with the (slow!) polynomial decay of $v''(t)$ for fBm traffic. The following definition gives a classification.

**Definition 2.4.3 Long-range dependence.** *We call a traffic source* long-range dependent *(lrd), when the covariances* $\mathbb{C}(k,1)$ *are nonsummable:*

$$\sum_{k=1}^{\infty} \mathbb{C}(k,1) = \infty,$$

*and* short-range dependent *(srd) when this sum is finite.*

Turning back to the case of fBm, with variance function given by $v(t) = t^{2H}$, it is easily checked that

$$\lim_{k\to\infty} \frac{\mathbb{C}(k,1)}{k^{2H-2}} = \frac{1}{2} \cdot \lim_{k\to\infty} \frac{(1+1/k)^{2H} - 2 + (1-1/k)^{2H}}{1/k^2} = \frac{1}{2} \cdot v''(1).$$

This entails that we have to check whether $k^{2H-2}$ is summable or not. We conclude that Gaussian sources with this variance function are lrd iff $2H > 1$, i.e., whenever they belong to the positively correlated case.

For iOU we have that

$$\mathbb{C}(k,1) = \frac{1}{2}\left(e^{-k-1} - 2e^{-k} + e^{-k+1}\right),$$

which is summable. This implies that, according to Definition 2.4.3, iOU is srd.

*Smoothness.* A second criterion to classify Gaussian processes is based on the level of smoothness of the sample paths. We coin the following definition.

**Definition 2.4.4 Smoothness.** *We call a Gaussian source* smooth *if, for any* $t > 0$,

$$\lim_{\epsilon\downarrow 0} \frac{\mathbb{C}\mathrm{ov}(A(0,\epsilon), A(t,t+\epsilon))}{\sqrt{\mathbb{V}\mathrm{ar}(A(0,\epsilon))\mathbb{V}\mathrm{ar}(A(t,t+\epsilon))}} = \lim_{\epsilon\downarrow 0} \frac{\mathbb{C}(t,\epsilon)}{v(\epsilon)} \neq 0,$$

*and* nonsmooth *otherwise.*

An fBm source is nonsmooth, as is readily verified:

$$\lim_{\epsilon\downarrow 0} \frac{\mathbb{C}(t,\epsilon)}{v(\epsilon)} = \lim_{\epsilon\downarrow 0} \frac{1}{2}\epsilon^{2-2H} v''(t) = 0,$$

for any $t > 0$ and $H \in (0, 1)$. On the other hand, the iOU source is smooth, as, for any $t > 0$, applying that $2v(\epsilon)/\epsilon^2 \to 1$ as $\epsilon \downarrow 0$,

$$\lim_{\epsilon \downarrow 0} \frac{\mathbb{C}(t, \epsilon)}{v(\epsilon)} = v''(t) = e^{-t} > 0.$$

Generally speaking, one could say that Gaussian sources are smooth if there is a notion of a *traffic rate*. The following exercise gives more insight into this issue.

**Exercise 2.4.5** Determine the distribution of the instantaneous traffic rate process

$$R(t) := \lim_{s \uparrow t} \frac{A(s, t)}{t - s},$$

for iOU; $t > 0$. And what happens in case of fBm?

**Solution.** For (centered) iOU, $R(t)$ has a normal distribution with mean 0 and variance

$$\lim_{s \uparrow t} \mathbb{V}\mathrm{ar} \left( \frac{A(s, t)}{t - s} \right) = \lim_{s \uparrow t} \frac{\mathbb{V}\mathrm{ar} A(t - s)}{(t - s)^2} = \frac{1}{2}.$$

It can also be verified that $\mathbb{C}\mathrm{ov}(R(0), R(t)) = e^{-t}$. In other words, the rate process is again Gaussian.

In case of fBm, $\mathbb{V}\mathrm{ar} R(t) = \infty$ for all $t > 0$. ◇

The above findings can be interpreted in the following, more concrete way. One could say that for fBm there is full independence of the direction in which the process is moving (in line with the nondifferentiable nature of the sample-paths for fBm). For iOU there is some positive dependence between the rates (but this dependence vanishes fast, viz. at an exponential rate).

## 2.5 Other useful Gaussian source models

The example of fBm in the previous section may have left the impression that nonsmoothness is a necessary condition for long-range dependence. This is by no means true. In this section we present a number of other useful Gaussian processes, including one that is smooth and lrd at the same time. The common feature of these processes is that they are the so-called *Gaussian counterpart* of well-known, classical (non-Gaussian) arrival processes. In this context, we say that an arrival process $(A(t), t \in \mathbb{R})$ (with stationary increments) has the Gaussian counterpart $(\overline{A}(t), t \in \mathbb{R})$ if $\overline{A}(\cdot)$ is Gaussian and, in addition, $\mathbb{E}A(t) = \mathbb{E}\overline{A}(t)$ and $\mathbb{V}\mathrm{ar}A(t) = \mathbb{V}\mathrm{ar}\overline{A}(t)$ for all $t$. Generally speaking, $\overline{A}(\cdot)$ inherits the correlation structure of $A(\cdot)$.

*A. The Gaussian counterpart of the Poisson stream.* Let jobs of size 1 arrive according to a Poisson process with rate $\lambda > 0$. Then it is well known that $A(t)$, denoting the amount of traffic generated in an interval of length $t$, has a Poisson distribution with mean $\lambda t$ (and hence also variance $\lambda t$). As a consequence, the Gaussian counterpart of this model is (a scaled version of) Brownian motion.

*B. The Gaussian counterpart of the M/G/∞ input model.* In the M/G/∞ input model, jobs arrive according to a Poisson process of rate $\lambda$, stay in the system during some random time $D$ (where the durations of the individual jobs constitute a sequence of i.i.d. random variables), and transmit traffic at rate, say, 1 while in the system. Clearly, this model does not correspond to a Gaussian source, but of course we could approximate it by a Gaussian source with the same mean and correlation structure.

It follows immediately that the mean rate $\mu$ of this Gaussian counterpart should be chosen as $\mu = \lambda \mathbb{E}D$, assuming that $D$ has a finite mean, say $\delta$. The variance is somewhat harder to compute. With $A(t)$ denoting the amount of traffic generated by the M/G/∞ input model in an interval of length $t$, we have that $A(t)$ can be decomposed into the contribution of the $M$ jobs that were already present at the start of the interval (say, time 0), and the contribution of the $N_t$ jobs arriving in $(0, t)$:

$$A(t) =_{\rm d} \sum_{i=0}^{M} X_i(t) + \sum_{i=0}^{N_t} Y_i(t); \tag{2.3}$$

here $X_i(t)$ is the amount of traffic generated by an arbitrary job that was present at time 0, whereas $Y_i(t)$ denotes the amount of traffic of an arbitrary job arriving in $(0, t)$. Observe that these two sums in the right-hand side of Equation (2.3) are independent. In more detail,

- standard theory on M/G/∞ queues says that $M$ has a Poisson distribution with mean $\lambda\delta$, and the $X_i(t)$ are i.i.d. (as a random variable $X(t)$), independently of $M$;

- on the other hand, $N_t$ is Poisson with mean $\lambda t$, and, again, the $Y_i(t)$ are i.i.d. (as some random variable $Y(t)$), independently of $N_t$.

Recalling the property that, with $Z_i$ i.i.d. (as a random variable $Z$), and $N$ Poisson with mean $\nu$ (independent of the $Z_i$), we have that

$$\mathbb{V}\text{ar}\left(\sum_{i=1}^{N} Z_i\right) = \nu\, \mathbb{V}\text{ar} Z,$$

we find that we are left with finding $\mathbb{V}\text{ar}X(t)$ and $\mathbb{V}\text{ar}Y(t)$.

To this end, we wonder how $X(t)$ and $Y(t)$ are distributed. Let $f_D(\cdot)$ $(F_D(\cdot))$ be the density (distribution function) of $D$. In the sequence we need the notion of the so-called *integrated tail* (or residual lifetime) of $D$, which is a random variable that we denote by $D^r$. The density of this $D^r$ is given by [13, Ch. 5]

$$f_{D^r}(t) = \frac{1 - F_D(t)}{\delta};$$

the fact that $\mathbb{E}D \equiv \delta = \int_0^\infty (1 - F_D(t))\,\mathrm{d}t$ implies that this is indeed a density. $F_{D^r}(\cdot)$ is the distribution function corresponding to $f_{D^r}(\cdot)$.

First consider $X(t)$. It is standard result that the remaining time the job stays in the system for a time that has density $f_{D^r}(\cdot)$. If this remaining time is $s < t$, then $X(t) = s$' otherwise it is $t$. In other words, the $k$th moment of $X(t)$ equals

$$\mathbb{E}(X(t))^k = \int_0^t s^k f_{D^r}(s)\,\mathrm{d}s + t^k(1 - F_{D^r}(t)).$$

This immediately yields a formula for $\mathbb{V}\mathrm{ar}X(t)$.

Then focus on $Y(t)$. It is well known that the epoch the job arrives is uniformly distributed on $(0, t)$. With a similar reasoning as above, where the variable $u$ corresponds to the epoch the job arrives, we find that

$$\mathbb{E}(Y(t))^k = \int_0^t \frac{1}{t}\left((t-u)^k(1 - F_D(t-u)) + \int_0^{t-u} s^k f_D(s)\,\mathrm{d}s\right)\mathrm{d}u.$$

Again, this enables computation of $\mathbb{V}\mathrm{ar}Y(t)$.

Taking all terms together yields

$$\begin{aligned}\mathbb{V}\mathrm{ar}A(t) = \lambda\delta &\left(\int_0^t s^2 f_{D^r}(s)\,\mathrm{d}x + t^2(1 - F_{D^r}(t))\right)\\ &+ \lambda\left(\int_0^t (t-u)^2(1 - F_D(t-u))\,\mathrm{d}u\right.\\ &\left.+ \int_0^t \int_u^t (s-u)^2 f_D(s-u)\,\mathrm{d}s\,\mathrm{d}u\right); \qquad (2.4)\end{aligned}$$

see also [197].

**Exercise 2.5.1** (i) Let $D$ be exponentially distributed with mean $\delta$. Show that

$$\mathbb{V}\mathrm{ar}A(t) = 2\lambda\delta^3\left(\frac{t}{\delta} - 1 + \exp\left(-\frac{t}{\delta}\right)\right). \qquad (2.5)$$

Compare this with the variance function of iOU, and conclude that iOU is (up to a scaling) the Gaussian counterpart of the M/G/∞ input model with exponential jobs.

(ii) Let $D$ have a Pareto distribution, i.e., $F_D(t) = 1 - (1/(1+t))^\alpha$. To have that $\delta < \infty$, we assume $\alpha > 1$. Verify that

$$v(t) = \frac{2\lambda}{(3-\alpha)(2-\alpha)(1-\alpha)} \left(1 - (t+1)^{3-\alpha} + (3-\alpha)t\right), \tag{2.6}$$

with $\alpha = (1+\delta)/\delta$, excluding $\delta = 1$ or $\frac{1}{2}$.

**Solution.** The claims follow after straightforward calculus. ◇

**Exercise 2.5.2** (i) Show that the Gaussian counterpart of the M/G/$\infty$ input model is smooth.

(ii) Assuming $D$ exponential, is the Gaussian counterpart lrd? Assuming $D$ as Pareto, is the Gaussian counterpart lrd?

**Solution.** (i) Using standard rules for differentiation of integrals,

$$v'(t) = \lambda\delta \cdot 2t\,(1 - F_{D^r}(t)) + \lambda \int_0^t 2(t-u)(1 - F_D(t-u))\,\mathrm{d}u$$

and hence $v'(0) = 0$. Similarly,

$$\begin{aligned} v''(t) &= 2\lambda\delta\,(1 - F_{D^r}(t) - t f_{D^r}(t)) + 2\lambda \int_0^t (1 - F_D(t-u) - (t-u) f_D(t-u))\,\mathrm{d}u \\ &= 2\lambda \int_t^\infty (1 - F_D(s))\,\mathrm{d}s. \end{aligned}$$

Hence, $v''(0) = 2\lambda\delta < \infty$. As a result,

$$\lim_{\epsilon \downarrow 0} \frac{\mathbb{C}(t,\epsilon)}{v(\epsilon)} = \lim_{\epsilon \downarrow 0} \frac{\epsilon^2 v''(t)/2}{\epsilon^2 v''(0)} > 0.$$

(ii) When $D$ is exponential, the Gaussian counterpart is srd (recall that we know from Exercise 2.5.1 that it has the same correlation structure as iOU). From Equation (2.6) it is readily seen that for Pareto $D$ the Gaussian counterpart is lrd iff $\alpha \in (1,2)$. For these parameters, $\mathbb{V}\mathrm{ar} A(t)$ grows superlinearly in $t$, whereas for $\alpha \geq 2$, it grows essentially linearly. ◇

*C. The Gaussian counterpart of the purely periodic stream.* Many networking applications inject information packets at a (more or less) constant rate. Perhaps the most prominent example is *voice* traffic: sound is put into equally sized packets, which are fed into the network after constant time intervals (say, every $D$ units of time). In fact, the only random element in this traffic model is the *phasing*. More precisely, when considering the first interval of length $D$, say $[0, D)$, it is obvious that exactly one packet will arrive; the epoch that this packet arrives, however, is uniformly distributed over $[0, D)$. From then on, the process is deterministic; if the packet arrived at $t \in [0, D)$, then packets also arrive at epoch $t + iD$, with integer $i$.

Now consider the Gaussian counterpart of this arrival process. It is not hard to check that (if packets have size 1) $\mu = 1/D$. Observe that $A(t)$ equals either $\lfloor t/D \rfloor$ or $\lceil t/D \rceil$. Note that for $t \in [0, D)$, $A(t)$ is 1 (0) with probability $t/D$ ($1 - t/D$, respectively). This entails that

$$\mathbb{V}\mathrm{ar} A(t) = (t \bmod D)(1 - t \bmod D),$$

where $t \bmod D$ denotes the fractional part of $t/D$, i.e., $t \bmod D = t/D - \lfloor t/D \rfloor$. We conclude that the Gaussian counterpart of the purely period stream has a variance function that equals (up to some scaling) $t(1 - t)$, repeated periodically.

With $B(t)$ denoting a Brownian motion, the process (defined for $t \in [0, 1]$)

$$\overline{B}(t) = (B(t) \mid B(1) = 0)$$

is usually referred to as a *Brownian bridge*. Using the formula (2.2) for conditional variances given in Section 2.2, we find that

$$\mathbb{V}\mathrm{ar}\overline{B}(t) = \mathbb{V}\mathrm{ar}B(t) - \frac{(\mathbb{C}\mathrm{ov}(B(t), B(1))^2}{\mathbb{V}\mathrm{ar}B(1)} = t - t^2 = t(1 - t);$$

Here, that $\mathbb{C}\mathrm{ov}(B(t), B(1)) = t$ is also used. Observe that the shape of $\mathbb{V}\mathrm{ar}\overline{B}(t)$ makes sense: (i) for small $t$ it looks like the variance of $B(t)$ (i.e., $t$), as the effect of the condition $B(1) = 0$ will still be minimal; (ii) for $t$ close to 1 it goes to 0.

We conclude from the above that the Gaussian counterpart of the periodic stream is a scaled Brownian bridge, repeated periodically.

*D. Two-timescale models.* In many applications in communication networks, a purely periodic traffic model is too crude. Often sources alternate between an on-mode and an off-mode, where during on-times packets are transmitted (as before) periodically. The typical example where such a model can be used is *voice with silence suppression*: a codec recognizes whether there is a voice signal to transmit or not, and if there is, then packets are generated in a periodic fashion. This type of models is called *two-timescale models*, as it combines a model that relates to the very short timescale (i.e., the timescale of the transmission of packets, which is in the order of, say, tens of milliseconds) with the timescale of a user alternating between an active and silent mode (in the order of seconds).

As said, within the on-time, packets (of fixed size, say, 1) are sent periodically; for ease, we normalize time such that the corresponding packet interarrival time is 1. The on-times (or *bursts*) are random variables with length $T_{\mathrm{on}}$, which is a natural number. The off-times (or *silence* periods) are random variables with length $T_{\mathrm{off}}$, which is also integer valued. We emphasize that–although $T_{\mathrm{on}}$ and $T_{\mathrm{off}}$ only attain values in $\mathbb{N}$−the arrival process is *not* a discrete-time process, as the 'phase' of the source is uniformly distributed on $[0, 1)$.

Whereas in the examples A, B, and C we computed the variance function of $A(t)$ in a direct way, we now follow an indirect approach. We compute the so-called *moment generating function* (mgf) of $A(t)$, i.e.,

$$\alpha_t(\theta) := \mathbb{E}\exp(\theta A(t)),$$

for $t > 0$ and $\theta \in \mathbb{R}$. From this mgf, the mean $\mathbb{E}A(t)$ (which equals $\mu t$) can be found through differentiation with respect to $\theta$ and letting $\theta$ go to 0:

$$\mathbb{E}A(t) = \left.\left(\frac{\partial}{\partial\theta}\alpha_t(\theta)\right)\right|_{\theta=0}.$$

Similarly, the second moment of $A(t)$ can be found by taking the second derivative, so that we find for the variance

$$\mathbb{V}\mathrm{ar}A(t) = \left.\left(\frac{\partial^2}{\partial\theta^2}\alpha_t(\theta)\right)\right|_{\theta=0} - \left(\left.\left(\frac{\partial}{\partial\theta}\alpha_t(\theta)\right)\right|_{\theta=0}\right)^2.$$

For ease we assume that $T_{\mathrm{on}}$ is geometric with probability $p$, i.e.,

$$\mathbb{P}(T_{\mathrm{on}} = k) = (1-p)^{k-1}p$$

(and even then the *residual* on-time has this distribution, due to the memoryless property of the geometric distribution). The off-time $T_{\mathrm{off}}$ (and the residual off-time) is assumed to be geometrically distributed with probability $q$.

Define, for $t > 0$, $t_m := t \bmod 1$. It is clear that $A(t_m) = 0$ when the source is off, whereas $A(t_m) = 1$ (0) with probability $t_m$ ($1 - t_m$, respectively) when the source is on. In other words,

$$\mathbb{E}_1 e^{\theta A(t_m)} = t_m e^\theta + (1 - t_m), \qquad \mathbb{E}_0 e^{\theta A(t_m)} = 1,$$

with $\mathbb{E}_i(\cdot)$, for $i = 0, 1$, defined in a self-evident way. This leaves us with the task of computing $\mathbb{E}_i e^{\theta A(T)}$ for integer $T$. A straightforward argument immediately gives us the system of coupled difference equations

$$\mathbb{E}_1 e^{\theta A(T)} = (1-p)e^\theta \mathbb{E}_1 e^{\theta A(T-1)} + p\mathbb{E}_0 e^{\theta A(T-1)},$$
$$\mathbb{E}_0 e^{\theta A(T)} = qe^\theta \mathbb{E}_1 e^{\theta A(T-1)} + (1-q)\mathbb{E}_0 e^{\theta A(T-1)}.$$

This system can be solved in a standard way. Let $\lambda_+(\theta)$ and $\lambda_-(\theta)$ be the eigenvalues of the matrix

$$\begin{pmatrix} (1-p)e^\theta & p \\ qe^\theta & 1-q \end{pmatrix},$$