M/G/infinity transience, and its applications to overload detection

M.R.H. Mandjes, P. Zuraniewski

PNA-E0910

Centrum Wiskunde & Informatica

Centrum Wiskunde & Informatica (CWI) is the national research institute for Mathematics and Computer Science. It is sponsored by the Netherlands Organisation for Scientific Research (NWO).
CWI is a founding member of ERCIM, the European Research Consortium for Informatics and Mathematics.

CWI's research has a theme-oriented structure and is grouped into four clusters. Listed below are the names of the clusters and in parentheses their acronyms.

Probability, Networks and Algorithms (PNA)

Software Engineering (SEN)

Modelling, Analysis and Simulation (MAS)

Information Systems (INS)

P.O. Box 94079, 1090 GB Amsterdam (NL)
Science Park 123, 1098 XG Amsterdam (NL)
Telephone +31 20 592 9333
Telefax +31 20 592 4199

ISSN 1386-3711

M/G/∞ transience, and its applications to overload detection

Michel Mandjes[*] and Piotr Żuraniewski[†]

May 14, 2009

Abstract

When controlling communication networks, it is of crucial importance to have procedures that are capable of checking whether there are unanticipated load changes. In this paper we develop techniques for detecting such load changes, in a setting in which each connection consumes roughly the same amount of bandwidth (with VoIP as a leading example). For the situation of exponential holding times an explicit analysis can be performed in a large-deviations regime, leading to approximations of the test statistic of interest (and, in addition, to results for the transient of the M/M/∞ queue, which are of independent interest). This procedure being applicable to exponential holding times only, and also being numerically rather involved, we then develop an approximate procedure for general holding times. In this procedure we record the number of trunks occupied at equidistant points in time $\Delta, 2\Delta, \ldots$, where Δ is chosen sufficiently large to safely assume that the samples are independent; this procedure is backed by results on the transient of the M/G/∞ queue, thus complementing earlier results on relaxation times. The validity of the testing procedures is demonstrated through an extensive set of numerical experiments.

2000 Mathematics Subject Classification: 60K25 60F05 90B22
Keywords and Phrases: M/G/infinity, overload detection, large deviations

[*] M. Mandjes (email: mmandjes@science.uva.nl) is with Korteweg-de Vries Institute, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, Amsterdam, the Netherlands. He is also affiliated to CWI, P.O. Box 94079, 1090 GB Amsterdam, the Netherlands, and EURANDOM, Eindhoven, the Netherlands; part of this work was done while he was at Stanford University, Stanford, CA 94305, US.

[†] P. Żuraniewski (email: piotr.zuraniewski@agh.edu.pl) is with Department of Applied Mathematics, AGH University of Science and Technology, Kraków, Poland; part of this work was done while he was at Korteweg-de Vries Institute, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, Amsterdam, the Netherlands.

1 Introduction

When sizing communication networks, probably still the most frequently used tool is the celebrated *Erlang loss formula*, dating back to the early 1900s. This formula was originally developed for computing the blocking probability for circuit-switched (for instance voice) traffic sharing a trunk group of size, say $C \in \mathbb{N}$, and could be used for *dimensioning* purposes: it enables the selection of a value of C such that it is guaranteed that the blocking probability is below some tolerable level ε. Despite the fact that the formula has been around for a rather long time, it is still a cornerstone when resolving dimensioning issues, owing to its general applicability and its explicit, manageable form. Also, it is important to notice that it can in principle be used in any setting in which each connection requires (roughly) the same amount of bandwidth (which can then be normalized to 1). As a consequence, it is also applicable for technologies that are currently used, for instance in the context of voice-over-IP (VoIP).

In more detail, the Erlang loss formula is based on the (realistic) assumption of a Poisson arrival stream of flows (say, with intensity λ, expressed in Hertz or s^{-1}). The call durations are independent and identically distributed, with mean $1/\mu$ (in s), and the load ϱ is defined as the unit-less number $\varrho := \lambda/\mu$. If there are C lines available, the probability of blocking in this model is

$$p(C \mid \varrho) := \left(\frac{\varrho^C}{C!}\right) \bigg/ \left(\sum_{c=0}^{C} \frac{\varrho^c}{c!}\right).$$

Importantly, this formula shows that for dimensioning the trunk group, no information on λ and μ is needed apart from their ratio $\varrho = \lambda/\mu$. Observe that no assumption on the distribution of the call holding times was imposed; the above formula applies for all holding-time distributions with mean $1/\mu$.

We denote by $\bar{\varrho}$ the maximum load ϱ such that $p(C \mid \varrho)$ is below some predefined tolerance ε. The underlying queueing model is often referred to as the M/G/C/C queue; a useful (rough) approximation of $p(C \mid \varrho)$ is the probability that the number of busy servers in the corresponding infinite-server queue, that is, M/G/∞, exceeds C.

Above we described a rudimentary dimensioning procedure, but when operating a network one has to constantly check the validity of the input assumptions the dimensioning decision was based upon. More concretely, one has to check whether the load ϱ has not reached the maximum allowable load $\bar{\varrho}$. Clearly, if the load has increased beyond $\bar{\varrho}$, measures have to be taken to deal with the overload, perhaps by rerouting the excess calls, or, on a longer timescale, by increasing the available capacity.

This motivates why it is of crucial importance to design procedures to (statistically) assess whether the load has changed. In statistical terms we would call this a 'changepoint detection problem' [13]: from observations of the number of lines used, we wish to infer whether a load change has taken place. Also, one would like to know *when* the change has occurred; then an alarm can be issued that triggers traffic management measures (overload control, such as rerouting, or temporary adaptations of the amount of bandwidth available to the calls).

Empirical guidelines for the problem described above have been developed in e.g., [9], but there is a clear need for more rigorously supported procedures. Without aiming to give an exhaustive overview, we mention here related work on a fractal model [17], and also [5,15,16]. An application of the celebrated cusum technique [13] in the networking domain can be found in [7], see also [11]. Several valuable contributions to the changepoint detection problem are due to Tartakovsky and co-authors, cf. [14]. We finally mention that, interestingly, in other application areas the same type of problems play a crucial role, see e.g. [3].

The main contributions of the present paper are the following.

- We first consider the case in which the call durations have an exponential distribution. We show how a likelihood-based cusum-type of test can be set up. The crucial complication is that the number of trunks occupied does *not* constitute a sequence of i.i.d. random variables (as there will be dependence between subsequent observations). Therefore the 'traditional' cusum result does not apply here, and a new approach had to be developed.

Setting up our test requires knowledge of the transient probabilities in the corresponding M/M/∞ system. We first show how, in a large-deviations setting, these transient probabilities can be determined. These have interesting features, such as a so-called bifurcation, as in [12]. The test also requires the computation of the probability that a sum of likelihoods exceeds some threshold. We show how this can be done, relying on calculus-of-variations techniques.

- The findings above being only applicable to the case of exponentially distributed call durations, and given the high numerical complexity of the resulting procedure, we then look for an alternative approach that works for the M/G/∞ in general, and that requires substantially less computational effort.

  We explain how classical changepoint-detection techniques can be used here. These classical techniques rely on the assumption of independent observations (where the observations correspond to samples of the number of calls in progress, at equidistant points in time, say $\Delta, 2\Delta, \ldots$). This independence assumption is clearly not fulfilled in our model, at least not formally, but evidently for Δ sufficiently large the dependence will have a minor impact. We develop new estimates on the relaxation time of the M/G/∞ queue, which tell us how large Δ should be in order to be able to safely assume independence.

- We then show how accurately the proposed procedures can detect overload. This we do through a series of simulation experiments. Special attention is paid to the trade-off between the detection ratio and the false alarm rate. The experiments indicate that our procedure, after some tuning, provides a powerful technique for changepoint detection.

We have organized the paper as follows. In Section 2 we present our model and some preliminaries, and define our goal in terms of a changepoint detection problem. Section 3 presents a framework for changepoint detection for the M/M/∞ model, whereas Section 4 presents the approximate analysis for the M/G/∞ model. The last section is devoted to numerical experimentation.

2 Model, preliminaries, and goals

In this section we describe the goals of the paper, and the underlying mathematical model. Our analysis will be based on the M/G/∞ queue, that is, a service system in which calls arrive according to a Poisson process (with rate, say, λ), where it is assumed that the call durations form an i.i.d. sequence $B_1, B_2, \ldots$, and infinitely many servers. With $1/\mu$ denoting the mean value of a generic call duration B, the load of the system is defined as $\varrho := \lambda/\mu$. It is well-known that the stationary distribution of the number of calls simultaneously present, say Y, is Poisson with mean ϱ:

$$\mathbb{P}(Y = k) = \frac{\varrho^k}{k!} e^{-\varrho}. \tag{1}$$

Also the transient distribution of this system can be dealt with fairly explicitly. Suppose that $Y(t)$ denotes the number of trunks occupied at time t, and assuming that the queue is in stationarity at time 0, the following decomposition applies. Conditioning on $Y(0) = k$, with '$=_{\rm d}$' denoting equality in distribution, we have that

$$Y(t) =_{\rm d} \mathbb{B}{\rm in}(k, p_t) + \mathbb{P}{\rm ois}(\lambda t q_t), \tag{2}$$

where $\mathbb{B}{\rm in}(k, p)$ denotes a binomial random variable with parameters k and p, and $\mathbb{P}{\rm ois}(\lambda)$ as Poisson random variable with mean λ; in addition, the binomial and Poisson random variables in the right-hand side of (2) are independent. Here, p_t is the probability that an arbitrary call that is present at time 0 is still present at time t, which can be computed as

$$p_t = \mathbb{P}(B^\star > t) = \frac{1}{\mathbb{E}B}\int_t^\infty \mathbb{P}(B > s)\mathrm{d}s,$$

where $B^\star$ denotes the excess life-time distribution of B. Likewise, q_t is the probability that an arbitrary call that arrives in $(0, t]$ is still present at time t; using the fact that the arrival epoch of such an arbitrary call is uniformly distributed on $(0, t]$, conditioning on the arrival epoch $s \in (0, t]$ yields that

$$q_t = \int_0^t \frac{1}{t}\mathbb{P}(B > t - s)\mathrm{d}s = \int_0^t \frac{1}{t}\mathbb{P}(B > s)\mathrm{d}s = \frac{\mathbb{E}B}{t} \cdot \mathbb{P}(B^\star \leq t).$$

Observe that the mean of the Poissonian term in the right-hand side of (2), $\lambda t q_t$, equals $\varrho\mathbb{P}(B^\star < t)$.
It is readily verified that the correlation coefficient of $Y(0)$ and $Y(t)$ equals

$$\mathbb{C}\mathrm{orr}(Y(0), Y(t)) = \mathbb{P}(B^\star > t);$$

here it is used that $Y(0)$ has a Poisson distribution with mean ϱ.

As mentioned in the introduction, the goal of the paper is to detect changes in the load imposed on a M/G/∞ queue. More specifically, with ϱ the load imposed on the queueing resource, and $\bar{\varrho}$ the maximum allowable load (in order to meet a given performance criterion, for instance in terms of a blocking probability), we want to test whether all samples correspond to load ϱ (which we associate with hypothesis H_0), or whether there has been a changepoint within the data set, such that before the changepoint the data were in line with load ϱ, and after the changepoint with $\bar{\varrho}$ (which is hypothesis H_1).

3 Analysis for M/M/∞

In this section we consider the case that the calls are i.i.d. samples from an exponential distribution with mean $1/\mu$; the model is then known as M/M/∞. We consider the discrete-time Markovian model describing the dynamics of the number of trunks occupied, by recording the continuous-time process at the embedded epochs at which this number changes.
Let, for $i = 1, 2, \ldots$, $Y_i := \sum_{j=1}^i X_i$, where the probabilities $\mathbb{P}(X_i = \pm 1 \mid Y_{i-1})$ are defined through, for given numbers λ_m and μ_m,

$$(X_i \mid Y_{i-1} = m) = \left\{ \begin{array}{rl} 1 & \text{with probability } \lambda_m \\ -1 & \text{with probability } \mu_m = 1 - \lambda_m. \end{array} \right.$$

As mentioned above, in this section we consider assume that the dynamics of the number of trunks occupied are described by the M/M/∞ model, i.e.,

$$\lambda_m \equiv \lambda_m(\varrho) = \frac{\lambda}{\lambda + m\mu} = \frac{\varrho}{\varrho + m},$$

with $\varrho := \lambda/\mu$. We consider the model with an infinite number of trunks available; then the (steady-state) probability of C calls present can be used as an approximation of the blocking probability in the model with C lines.
In this section, our analysis relies on applying the so-called *many-flows scaling*. Under this scaling the load is renormalized by n (that is, we replace $\varrho \mapsto n\varrho$), and at the same time the number of trunks is inflated by a factor n, as motivated in [12, Ch. 12]. It effectively means that we can use *large-deviations theory* to asymptotically (large n) determine the distribution of the number of calls simultaneously present. Under this scaling the steady-state number of calls present has a Poisson distribution with mean $n\varrho$:

$$\mathbb{P}(Y = k) = \frac{(n\varrho)^k}{k!}e^{-n\varrho},$$

which means that a straightforward application of Stirling's formula yields the following expression for the exponential decay rate of $\mathbb{P}(Y = \lfloor n\beta \rfloor)$:

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}(Y = \lfloor n\beta \rfloor) = -\varrho + \beta + \beta \log\left(\frac{\varrho}{\beta}\right) =: \xi(\beta);$$

here we recognize the large-deviations rate function of the Poisson distribution [12, Example 1.13]. Using Cramér's theorem, we also have that the probability $\mathbb{P}(Y \geq n\beta)$ has the same exponential decay rate.

Goal: changepoint. We want to test whether there is a 'changepoint', that is, during our observation period the load parameter ϱ (which we let correspond to the probability model $\mathbb{P}$) changes into $\bar{\varrho} \neq \varrho$ (the model $\mathbb{Q}$). More formally, we consider the following (multiple) hypotheses. Recall that X_i is the sequence of observed steps (which have value 1 or -1).

H_0: $(X_i)_{i=1}^n$ is distributed according to the above described birth-death chain with parameter ϱ.

H_1: For some $\delta \in \{1/n, 2/n, \ldots, (n-1)/n\}$, it holds that $(X_i)_{i=1}^{\lfloor n\delta \rfloor}$ is distributed according to the birth-death chain with parameter ϱ, whereas $(X_i)_{i=\lfloor n\delta \rfloor+1}^n$ is distributed according to the birth-death chain with parameter $\bar{\varrho} \neq \varrho$.

Inspired by the Neyman-Pearson lemma, see e.g. [2, Ch. V.E and Appendix E], we consider the following likelihood-ratio test statistic:

$$\max_{\delta\in[0,1)} \left(\frac{1}{n} \sum_{i=\lfloor n\delta \rfloor+1}^{n} L_i - \varphi(\delta) \right), \text{ with } L_i := \log \frac{\mathbb{Q}(X_i \mid Y_{i-1})}{\mathbb{P}(X_i \mid Y_{i-1})},$$

for some function $\varphi(\cdot)$ we will specify later.
To enable statistical tests, we wonder what the probability is, under H_0, that the above test statistic is larger than 0. For reasons of tractability, we consider in this section its exponential decay rate (asymptotic in the scaling parameter n):

$$\eta(\varphi) := \lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\max_{\delta\in[0,1)} \left(\frac{1}{n} \sum_{i=\lfloor n\delta \rfloor+1}^{n} L_i - \varphi(\delta) \right) > 0 \right). \tag{3}$$

Another option that we will treat in detail, is to explicitly take into account information on the number of calls present at time 0:

$$\eta(\varphi \mid \beta_0) := \lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\max_{\delta\in[0,1)} \left(\frac{1}{n} \sum_{i=\lfloor n\delta \rfloor+1}^{n} L_i - \varphi(\delta) \right) > 0 \,\middle|\, Y_0 = n\beta_0 \right). \tag{4}$$

We first decompose of the exponential decay rate (4) as follows. We define

$$\begin{aligned}
\eta(\varphi, \delta \mid \beta_0) &:= \lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\frac{1}{n} \sum_{i=\lfloor n\delta \rfloor+1}^{n} L_i > \varphi(\delta) \,\middle|\, Y_0 = n\beta_0 \right); \\
\bar{\eta}(\varphi, \delta \mid \beta_\delta) &:= \lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\frac{1}{n} \sum_{i=\lfloor n\delta \rfloor+1}^{n} L_i > \varphi(\delta) \,\middle|\, Y_{\lfloor n\delta \rfloor} = n\beta_\delta \right); \\
\xi(\beta_\delta \mid \beta_0) &:= \lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(Y_{\lfloor n\delta \rfloor} = n\beta_\delta \,\middle|\, Y_0 = n\beta_0 \right).
\end{aligned}$$

Standard large-deviations argumentation (e.g., principle of the largest term) yields that

$$\begin{aligned}
\eta(\varphi \mid \beta_0) &= \sup_{\delta\in[0,1)} \eta(\varphi, \delta \mid \beta_0) \\
&= \sup_{\delta\in[0,1)} \sup_{\beta_\delta>0} \left(\xi(\beta_\delta \mid \beta_0) + \bar{\eta}(\varphi, \delta \mid \beta_\delta) \right),
\end{aligned}$$

and also

$$\eta(\varphi) = \sup_{\delta\in[0,1)} \sup_{\beta_\delta>0} \left(\xi(\beta_\delta) + \bar{\eta}(\varphi,\delta \mid \beta_\delta)\right).$$

The decay rate of the transient probabilities, that is, $\xi(\beta_\delta \mid \beta_0)$, will be analyzed in Section 3.1, whereas the decay rate of the exceedance probabilities $\bar{\eta}(\varphi,\delta \mid \beta_\delta)$ (which we will sometimes refer to as 'likelihood probabilities') will be addressed in Section 3.2.

3.1 Transient probabilities

To analyze the decay rate of $\mathbb{P}(Y_{\lfloor n\delta\rfloor} = n\beta_\delta \mid Y_0 = n\beta_0)$, we rely on *Slow Markov Walk* theory [2, Ch. IV.C]. As this technique has been described in detail in [2] we restrict ourselves to sketching the main steps. Then we show how to apply this theory to determine the transient probabilities $\xi(\beta_\delta \mid \beta_0)$.

Slow Markov Walk. A prominent role in Slow Markov Walk theory is played by the so-called 'local large deviations rate function', which is given by

$$\begin{aligned} I_x(u) &= \sup_\theta \left(\theta u - \log\left(e^\theta \lambda_{nx}(n\varrho) + e^{-\theta}(1-\lambda_{nx}(n\varrho))\right)\right) \\ &= \sup_\theta \left(\theta u - \log\left(e^\theta \frac{\varrho}{\varrho+x} + e^{-\theta}\frac{x}{\varrho+x}\right)\right). \end{aligned}$$

Intuitively reasoning, $I_x(u)$ measures the 'effort the process has to make' (per time unit), starting in state x, to move into direction u. It is readily verified that the optimizing θ is given by

$$\theta^\star \equiv \theta^\star_x(u) = \frac{1}{2}\log\left(\frac{x}{\varrho}\cdot\frac{1+u}{1-u}\right); \tag{5}$$

if $\theta^\star$ is positive (negative) the process has to 'speed up' ('slow down') to be moving into direction u.
The purpose of Slow Markov Walk theory is to determine the exponential decay rates of the empirical mean process $n^{-1}\cdot Y_{\lfloor nt\rfloor}$ to be in a certain set, or close to a given function f. Loosely speaking, it says that

$$\lim_{n\to\infty}\frac{1}{n}\log\mathbb{P}\left(\frac{1}{n}\cdot Y_{\lfloor nt\rfloor}\approx f(t), t\in[0,\delta)\right) = -\int_0^\delta I_{f(t)}(f'(t))\mathrm{d}t;$$

sometimes the right-hand side of the previous display is referred to as the 'cost' of the path f in the interval $[0,\delta)$. In this sense, we can determine also the 'average path' of $Y_{\lfloor nt\rfloor}/n$, which is the path with *zero* cost: it consists of pairs $(f(t), f'(t))$ for which $I_{f(t)}(f'(t)) = 0$, or, put differently, $\theta^\star_{f(t)}(f'(t)) = 0$. It can be calculated that this average path is given through the differential equation

$$\frac{f(t)}{\varrho}\cdot\frac{1+f'(t)}{1-f'(t)} = 1, \quad\text{or}\quad f'(t) = \frac{\varrho - f(t)}{\varrho + f(t)}; \tag{6}$$

This path converges to the 'mean' $f(\infty) = \varrho$ as $t\to\infty$, as was expected.
Inserting $\theta^\star$, as given in (5), into the objective function $I_x(u)$, we find after tedious computations:

$$\begin{aligned} I_x(u) &= \frac{1}{2}u\log\left(\frac{1+u}{1-u}\right) + \frac{1}{2}u\log\left(\frac{x}{\varrho}\right) - \frac{1}{2}\log\left(\frac{x\varrho}{(\varrho+x)^2}\right) \\ &\quad -\log 2 + \frac{1}{2}\log(1-u) + \frac{1}{2}\log(1+u). \end{aligned}$$

Determining the decay rate of $\xi(\beta_\delta \mid \beta_0)$. We can now reduce the search for the decay rate $\xi(\beta_\delta \mid \beta_0)$ to a variational problem. Slow Markov Walk theory says that

$$\xi(\beta_\delta \mid \beta_0) = -\inf_{f\in\mathscr{A}}\int_0^\delta I_{f(t)}(f'(t))\mathrm{d}t,$$

where the set $\mathscr{A}$ consists of all paths f such that $f(0) = \beta_0$ and $f(\delta) = \beta_\delta$. This variational problem can be solved by applying elementary results from calculus of variations; see for instance [12, Appendix C]. The optimizing path is characterized by the so-called DuBois-Reymond equation [12, Eq. (C.3)]:

$$I_{f(t)}(f'(t)) - f'(t) \cdot \frac{\partial}{\partial u} I_x(u)\bigg|_{x=f(t),u=f'(t)} = K,$$

or, equivalently,

$$\log\left((1-(f'(t))^2) \cdot \frac{(\varrho + f(t))^2}{4f(t)\varrho}\right) = K;$$

the K is to determined later on (and essentially serves as a 'degree-of-freedom', to be chosen such that $f(\delta) = \beta_\delta$). After some elementary algebraic manipulations we find the ordinary differential equation

$$f'(t) = \pm\sqrt{1 - e^K \cdot \frac{4f(t)\varrho}{(\varrho + f(t))^2}}; \tag{7}$$

notice that for $K = 0$ we retrieve the 'average path' (6), as expected.
Unfortunately, the above differential equation allows only for an indirect solution, but, interestingly, we can explicitly find the inverse of the solution (that is, t in terms of f, rather than f in terms of t), as follows. Recalling that $\varrho + f(t) > 0$, by separating variables we obtain

$$t = \pm \int \frac{(\varrho + f)}{\sqrt{f^2 + (2\varrho - 4\varrho e^K)f + \varrho^2}}\, \mathrm{d}f.$$

We are going to solve this differential equation by applying Abel's theorem: for the natural number ℓ and $W_\ell(x)$, $V_{\ell-1}(x)$ being polynomials of degree ℓ and $\ell - 1$ respectively, we have

$$\int \frac{W_\ell(x)}{\sqrt{ax^2+bx+c}}\, \mathrm{d}x = V_{\ell-1}(x)\sqrt{ax^2+bx+c} + \int \frac{K}{\sqrt{ax^2+bx+c}}\, \mathrm{d}x.$$

Differentiation of the above expression leads to

$$\frac{W_\ell(x)}{\sqrt{ax^2+bx+c}} = V'_{\ell-1}(x)\sqrt{ax^2+bx+c} + V_{\ell-1}(x)\frac{2ax+b}{2\sqrt{ax^2+bx+c}} + \frac{K}{\sqrt{ax^2+bx+c}},$$

which, after multiplication by $2\sqrt{ax^2+bx+c}$, results in a polynomial equation that enables the computation of the coefficients of $V_{\ell-1}(x)$, as well as the constant K:

$$2W_\ell(x) = 2V'_{\ell-1}(x)(ax^2+bx+c) + V_{\ell-1}(x)(2ax+b) + 2K.$$

For our differential equation we obtain (noticing that $\ell = 1$), with $b_\varrho := 2\varrho - 4\varrho e^K$, after solving the polynomial equation,

$$\int \frac{(\varrho + f)}{\sqrt{f^2 + b_\varrho f + \varrho^2}}\, \mathrm{d}f = \sqrt{f^2 + b_\varrho f + \varrho^2} + \int \frac{2\varrho e^K}{\sqrt{f^2 + b_\varrho f + \varrho^2}}\, \mathrm{d}f.$$

This eventually yields

$$t = \pm\left(\sqrt{f^2 + b_\varrho f + \varrho^2} + 2\varrho e^K \log\left(f + \varrho - 2\varrho e^K + \sqrt{f^2 + b_\varrho f + \varrho^2}\right) + \gamma\right), \tag{8}$$

where γ is chosen such that the boundary condition, i.e., $f(0) = \beta_0$, is met.

Numerical evaluation. To obtain path the $f(t)$, for a given value of K, (8) needs to be solved, but obviously there are alternatives. One could for instance solve the differential equation (7) iteratively starting in $f(0) = \beta_0$,

by applying techniques of the Runge-Kutta type. There are numerical difficulties, though, as the path may be horizontal at some point between 0 and δ (so that the most straightforward numerical procedures do not work). Below we comment in greater detail on possible ways to solve the differential equation.
A second step the is then to find a value of K such that indeed $f_K^\star(\delta) = \beta_\delta$. Notice that, because we can move up or down by just 1, we have to require that $\beta_\delta \in [\max\{0, \beta_0 - \delta\}, \beta_0 + \delta]$. Once we have found the optimal path (having taken into account the condition $f_K^\star(\delta) = \beta_\delta$), say the path $f^\star(\cdot)$, we can (numerically) evaluate

$$\int_0^\delta I_{f^\star(t)}((f^\star)'(t))\mathrm{d}t,$$

thus finding the decay rate $\xi(\beta_\delta \mid \beta_0)$.
As indicated, we proceed by making a few observations that enable the numerical evaluation of the decay rate $\xi(\beta_\delta \mid \beta_0)$.

- If $\beta_0 < \varrho$ and $\beta_\delta > \varrho$ or vice versa the above differential equation can, for any given K, be numerically solved in a straightforward fashion, because the path will be *monotone.* More precisely, one can rely on well-known Runge-Kutta techniques, starting in $f(0) = \beta_0$. By varying the value of K, we can then find the path that is at β_δ at time δ.

  Analysis analogous to [12, Ch. 12] reveals the following properties. (i) Suppose $\beta_0 < \varrho < \beta_\delta$. Then the K that is such that $f_K(\delta) = \beta_\delta$ is *negative*. The above iterative Runge-Kutta scheme can be used, with for instance a bisection loop that selects the right $K < 0$. (ii) If $\beta_\delta < \varrho < \beta_0$, the optimal path is the time-reversed of the path that starts in β_δ and ends in β_0. This means that the optimal path can be identified as under (i), i.e., starting in $\beta_\delta < \varrho$, and ending in $\beta_0 > \varrho$; note that the decay rate differs, though (but can be determined by numerically evaluating the integral over the local rate function along the resulting path).

- Problems may arise, however, when the optimal path may have derivative 0 at some point in $(0, \delta)$. This is typically the case when β_0 and β_δ or both smaller or larger than ϱ, and δ is at the same time relatively large (as then the optimal path is such that the number of trunks occupied, starting from β_0, is first 'pulled' towards ϱ, and then 'pushed back' into the direction of β_δ). Interestingly, for given $K > 0$, one can compute the value f_K of $f(s)$ at the point s for which $f'(s) = 0$. It turns out that

  $$f_K = \varrho\left(2e^K - 1 \pm 2\sqrt{e^{2K} - e^K}\right);$$

  elementary arguments show that we have to take the $-$-sign ($+$-sign) when β_0 and β_δ are both smaller (larger) than ϱ. Also, it is readily verified that for $K = 0$ one obtains $f_K = \varrho$, and for $K \to \infty$ in the $-$-branch $f_K \to 0$ and in the $+$-branch $f_K \to \infty$. The solution has, as in [12, Section 12.5], a *bifurcation point*: for small δ (say, δ smaller than some critical timescale T) the path will typically be monotone ($K < 0$), whereas for larger δ (i.e., $\delta > T$) the path will have slope 0 for some point between 0 and δ ($K > 0$). There is no explicit expression for the timescale T avalable, but we can identify a timescale $T^- < T$ such that for any smaller δ the path will be monotone, as follows.

  First observe that we can explicitly solve (6) to obtain, for a constant γ:

  $$t = \pm(-2\varrho\log(\varrho + f(t)) - f(t) + \gamma);$$

  unfortunately we cannot invert this relation (thus obtaining $f(t)$ as function of t explicitly). Mimicking the argumentation in [12, Section 12.5], we find T^- by imposing $f(T^-) = \beta_0$, while γ is determined through $f(0) = \beta_\delta$ (here, again, time-reversibility properties are applied). We thus arrive at, for obvious reasons using the absolute value,

  $$T^- = \left|2\varrho\log\left(\frac{\varrho - \beta_0}{\varrho - \beta_\delta}\right) + \beta_0 - \beta_\delta\right|.$$

  We arrive at the following conclusion:

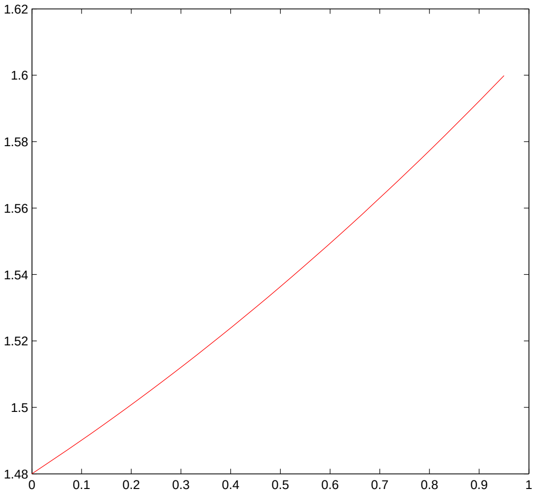

Figure 1: Monotone path.

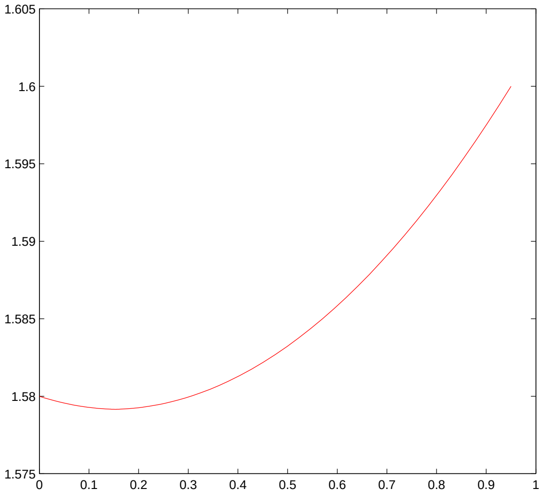

Figure 2: Nonmonotone path.

- For $\delta < T^-$ the path is monotone, we have $K < 0$, and we can use the method described for the case $\beta_0 < \varrho < \beta_\delta$. Use the $+$-branch of the differential equation.
- For $\delta > T^-$ one should realize that $\delta > T^-$ is just a necessary but not sufficient condition for a non-monotone optimal path to occur, as argued in [12, Section 12.5]; for $\delta \in [T^-, \infty)$ close to T^- still monotone paths come out. The bifurcation point T can be determined empirically.

Figures 1-2 serve as examples, and show the paths for specific monotone and non-monotone cases. In both cases β_0 as well as β_δ are larger than ϱ, so there is a bifurcation point T. In the left graph, $\delta < T$ and hence the path is monotone, whereas in the right graph $\delta > T$ and hence the path has a minimum in $(0, \delta)$. In both figures $\delta = 0.95$, $\varrho = 1.05$, and $\beta_\delta = 1.6$, but in the left panel $\beta_0 = 1.48$, whereas in the right panel $\beta_0 = 1.58$. The paths have been found by applying Runge-Kutta techniques.

3.2 Likelihood probabilities

In this section we analyze the decay rate $\bar{\eta}(\varphi, \delta \mid \beta_\delta)$, using the same methodology as in Section 3.1. As the line of reasoning is very similar to the one followed in Section 3.1, we just sketch the basic steps.
First observe that we can shift time so that we obtain

$$\bar{\eta}(\varphi, \delta \mid \beta_\delta) = \lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\frac{1}{n} \sum_{i=1}^{\lfloor n(1-\delta) \rfloor} L_i > \varphi(1-\delta) \,\middle|\, Y_0 = n\beta_\delta \right);$$

if this is indeed a large deviation probability, then we can replace the inequality '$> \varphi(1-\delta)$' by an equality '$= \varphi(1-\delta)$'. We again want to use Slow Markov Walk theory, in that we wish to evaluate

$$\bar{\eta}(\varphi, \delta \mid \beta_\delta) = -\inf_{f \in \mathscr{B}} \int_0^{1-\delta} I_{f(t)}(f'(t))\mathrm{d}t,$$

where $\mathscr{B}$ are the paths (with $f(0) = \beta_\delta$) such that

$$\lim_{n\to\infty} \frac{1}{n} Y_{\lfloor nt \rfloor} = f(t),$$

for $t \in [0, \delta)$, implies that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{\lfloor n(1-\delta)\rfloor} L_i = \varphi(1-\delta).$$

Let us characterize the paths with this property. To this end, first rewrite

$$g_f(\delta) := \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{\lfloor n(1-\delta)\rfloor} L_i = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{(1-\delta)/\varepsilon} \sum_{i=n(k-1)\varepsilon+1}^{nk\varepsilon} L_i.$$

For i in $\{n(k-1)\varepsilon+1, \ldots, nk\varepsilon\}$ we have that

$$\frac{\mathbb{Q}(X_i \mid Y_{i-1})}{\mathbb{P}(X_i \mid Y_{i-1})} = \frac{\bar{\varrho}}{\varrho} \cdot \frac{\varrho + f(k\varepsilon)}{\bar{\varrho} + f(k\varepsilon)} + O(\varepsilon)$$

if $X_i = 1$ and

$$\frac{\mathbb{Q}(X_i \mid Y_{i-1})}{\mathbb{P}(X_i \mid Y_{i-1})} = \frac{\varrho + f(k\varepsilon)}{\bar{\varrho} + f(k\varepsilon)} + O(\varepsilon)$$

if $X_i = -1$. Let $U_{k,n}$ be the number of steps upwards in $\{n(k-1)\varepsilon+1, \ldots, nk\varepsilon\}$, and $D_{k,n}$ the number of steps downwards. Then trivially $U_{k,n} + D_{k,n} = n\varepsilon$, but on the other hand $U_{k,n} - D_{k,n} = n\varepsilon f'(k\varepsilon) + O(\varepsilon^2)$. From these relations we can solve $U_{k,n}$ and $D_{k,n}$. We end up with

$$\sum_{k=1}^{(1-\delta)/\varepsilon} \left(\frac{\varepsilon}{2} f'(k\varepsilon) \log\left(\frac{\bar{\varrho}}{\varrho}\right) - \frac{\varepsilon}{2} \log\left(\frac{\bar{\varrho}}{\varrho}\right) + \varepsilon \log\left(\frac{\varrho + f(k\varepsilon)}{\bar{\varrho} + f(k\varepsilon)}\right) + O(\varepsilon^2) \right).$$

Letting $\varepsilon \downarrow 0$, we obtain

$$\begin{aligned} g_f(\delta) &= \int_0^{1-\delta} \frac{1}{2} \log\left(\frac{\bar{\varrho}}{\varrho}\right) \cdot f'(t)\,\mathrm{d}t - \frac{1-\delta}{2} \log\left(\frac{\bar{\varrho}}{\varrho}\right) + \int_0^{1-\delta} \log\left(\frac{\varrho + f(t)}{\bar{\varrho} + f(t)}\right) \mathrm{d}t \\ &= h_f(\delta) + \int_0^{1-\delta} \log\left(\frac{\varrho + f(t)}{\bar{\varrho} + f(t)}\right) \mathrm{d}t, \end{aligned}$$

with

$$h_f(\delta) := \frac{1}{2} \log\left(\frac{\bar{\varrho}}{\varrho}\right) \cdot (f(1-\delta) - f(0)) - \frac{1-\delta}{2} \log\left(\frac{\bar{\varrho}}{\varrho}\right).$$

Hence we are left with the following variational problem, with Lagrange multiplier L:

$$\inf_{f\in\mathcal{B}} \left(\int_0^{1-\delta} \left(I_{f(t)}(f'(t)) - L \log\left(\frac{\varrho + f(t)}{\bar{\varrho} + f(t)}\right) \right) \mathrm{d}t - L h_f(\delta) \right). \tag{9}$$

The DuBois-Reymond equation reads

$$I_{f(t)}(f'(t)) - f'(t) \cdot \frac{\partial}{\partial u} I_x(u) \Big|_{x=f(t),u=f'(t)} - L \log\left(\frac{\varrho + f(t)}{\bar{\varrho} + f(t)}\right) = K,$$

which reduces to

$$f'(t) = \pm \sqrt{1 - e^K \left(\frac{\varrho + f(t)}{\bar{\varrho} + f(t)}\right)^L \frac{4 f(t) \varrho}{(\varrho + f(t))^2}}.$$

We again need to numerically solve this, under $f(0) = \beta_\delta$. K and L should be chosen such that $g_f(\delta) = \varphi(1-\delta)$ and (9) is minimal. In more detail, a procedure could be the following. For given K, L, solve the differential equation, to obtain the path $f^\star_{K,L}(\cdot)$. For given L, determine the $K \equiv K(L)$ such that $g_{f^\star_{K,L}}(\delta) = \varphi(\delta)$. Then minimize, over L,

$$\int_0^{1-\delta} I_{f^\star_{K(L),L}(t)}((f^\star_{K(L),L})'(t))\mathrm{d}t.$$

It is clear, however, that such procedures are, from a numerical standpoint, in general quite involved. A substantial simplification can be achieved by approximating the functions involved by appropriate polynomial functions (cf. Ritz method).

3.3 Discussion

Now that we have derived in Sections 3.1 and 3.2 expressions for the decay rate of interest, it remains to select an appropriate function $\varphi(\cdot)$. We can choose, for a given value of β_0, $\varphi(\cdot)$ such that $\eta(\varphi, \delta \mid \beta_0) \equiv \alpha$ for all $\delta \in [0, 1)$. As argued in [2, Ch. V.E], this choice gives the best type-II error rate performance.

The procedure described above is a natural counterpart for the 'usual' changepoint detection procedures that were designed for i.i.d. increments; importantly, we recall the fact that in our model the increments are dependent made it necessary to develop a new method. The most significant drawbacks of the above procedure are: (i) it only applies to the case of exponentially distributed call durations; (ii) its computational complexity is high. In the next section we present an approach with is somewhat more crude, but overcomes these two problems.

4 Analysis for M/G/∞

In this section we present an approach to do changepoint detection in an M/G/∞ queue. Clearly, the observations $Y(0), Y(\Delta), Y(2\Delta), \ldots$ are *not* independent; remember from Section 2 that the correlation coefficient between $Y(0)$ and $Y(\Delta)$ is given by $\mathbb{P}(B^\star > \Delta)$. It is evident, however, that this dependence is negligible for Δ sufficiently large. In Section 4.1 we analyze how large Δ should be to be able to safely assume independence – as a useful by-product, we derive insight into the so-called relaxation times in the M/G/∞ queue (which can be interpreted as a measure of the speed of convergence to the stationary distribution). Then Section 4.2 describes a changepoint detection procedure, which again relies on Slow Markov Walk theory [2, Ch. IV.C]; however, where we used this framework for *dependent* observations in Section 3, we now focus on the case in which the observations are i.i.d. (and sampled from a Poisson distribution).

4.1 Transient probabilities

We first focus on the question: for a given number of calls present at time 0, how fast does the (transient) distribution of the number of calls present at time t, converge to stationary distribution (1)? This speed of convergence is often referred to as *relaxation time*, cf. Kingman [6] in the setting of an M/G/1 queue, and results for various queueing systems by Blanc and van Doorn [1]. We also refer to recent results on the relaxation time for the Erlang-loss system [4].

To this end, we identify a function $u_{k,\ell}(\cdot)$ such that

$$\lim_{t\to\infty} \frac{\mathbb{P}(Y(t) = \ell \mid Y(0) = k) - \mathbb{P}(Y = \ell)}{u_{k,\ell}(t)} = 1. \tag{10}$$

We first observe that that, due to (2),

$$\mathbb{P}(Y(t) = \ell \mid Y(0) = k) = \sum_{m=0}^{\min\{k,\ell\}} \mathbb{P}(\mathbb{B}\mathrm{in}(k, p_t) = m)\, \mathbb{P}(\mathbb{P}\mathrm{ois}(\lambda t q_t) = \ell - m).$$

Take the term corresponding to $m = 0$ in the summation in the right-hand side of the previous display, and subtract $\mathbb{P}(Y = \ell)$, to obtain, recalling that $\lambda t q_t = \varrho\, \mathbb{P}(B^\star < t) = \varrho - \varrho\, \mathbb{P}(B^\star > t)$,

$$\varrho e^{-\varrho} \left(\frac{\varrho^{\ell-1}}{(\ell-1)!} - \frac{\varrho^\ell}{\ell!} \right) \cdot \mathbb{P}(B^\star > t) \cdot (1 + o(1))$$

as $t \to \infty$; here we used that

$$\lim_{t\to\infty} \frac{f(\varrho) - f(\varrho(1 - \mathbb{P}(B^\star > t)))}{\varrho \mathbb{P}(B^\star > t)} = f'(\varrho).$$

Now we focus on the term corresponding to $m = 1$, which obeys

$$ke^{-\varrho}\frac{\varrho^{\ell-1}}{(\ell-1)!}\cdot\mathbb{P}(B^\star > t)\cdot(1+o(1)).$$

Finally observe that the terms corresponding to $m \geq 2$ are $o(\mathbb{P}(B^\star > t))$. Combining the above findings, we conclude that (10) indeed applies, with

$$u_{k,\ell}(t) = U_{k,\ell}\cdot\mathbb{P}(B^\star > t),\;\text{ with } U_{k,\ell} := \left(\varrho e^{-\varrho}\left(\frac{\varrho^{\ell-1}}{(\ell-1)!} - \frac{\varrho^{\ell}}{\ell!}\right) + ke^{-\varrho}\frac{\varrho^{\ell-1}}{(\ell-1)!}\right).$$

Suppose is our goal is to enforce 'approximate independence' between $Y(0)$ and $Y(t)$ by choosing t sufficiently large that for all $k, \ell \in \{0, \dots, C\}$ we have that $\mid U_{k,\ell} \mid \cdot\mathbb{P}(B^\star > t) < \varepsilon_{\max}$. Observe that

$$\max_{k,\ell\in\{0,\dots,C\}} U_{k,\ell} \leq (\varrho + k)e^{-\varrho}\frac{\varrho^{\ell-1}}{(\ell-1)!}.$$

Now we can make use of the fact that the mode of the Poisson distribution lies roughly at ϱ, or, more precisely,

$$\max_{i=0,1,\dots} e^{-\varrho}\frac{\varrho^i}{i!} \leq g(\varrho_m) := e^{-\varrho_m}\frac{\varrho^{\varrho_m}}{\varrho_m!},$$

with $\varrho_m := \lfloor\varrho\rfloor$ if ϱ is non-integer and ϱ else. We conclude that $U_{k,\ell} \leq (\varrho + C)g(\varrho)$ for all $k, \ell \in \{0, \dots, C\}$. Likewise,

$$\min_{k,\ell\in\{0,\dots,C\}} U_{k,\ell} \geq -\varrho e^{-\varrho}\frac{\varrho^{\ell+1}}{\ell!} \geq -\varrho g(\varrho).$$

Using these bounds it is trivial to choose t such that $\mid U_{k,\ell} \mid \cdot\mathbb{P}(B^\star > t) < \varepsilon_{\max}$ for all $k, \ell \in \{0, \dots, C\}$.

4.2 Changepoint detection procedure

As described above, we can now choose Δ so large that $u_{k,\ell}(\Delta) < \varepsilon_{\max}$, for all $k, \ell \in \{1, \dots, C\}$ and $\varepsilon_{\max}$ some given small positive number. In this way we enforced 'approximate independence', thus justifying the use of procedures for i.i.d. observations, as in [2, Section VI.E].

Goal: changepoint. Again, we wish to detect a changepoint, that is, during our observation period the load parameter ϱ (which we let again correspond to the probability model $\mathbb{P}$) changes into $\bar\varrho \neq \varrho$ (the model $\mathbb{Q}$). More formally, we consider the following (multiple) hypotheses. Let $Y_i := Y(i\Delta)$ be the sequence of observations of the number of calls present at time $i\Delta$.

- H_0: $(Y_i)_{i=1}^n$ are distributed according to a Poisson random variable with parameter ϱ.
- H_1: For some $\delta \in \{1/n, 2/n, \dots, (n-1)/n\}$, it holds that $(Y_i)_{i=1}^{\lfloor n\delta\rfloor}$ is distributed according to a Poisson random variable with parameter ϱ, whereas $(Y_i)_{i=\lfloor n\delta\rfloor+1}^n$ is distributed according to Poisson random variable with parameter $\bar\varrho \neq \varrho$.

Again, in view of the Neyman-Pearson lemma, we consider the following likelihood-ratio test statistic:

$$\max_{\delta\in[0,1)}\left(\frac{1}{n}\sum_{i=\lfloor n\delta\rfloor+1}^{n} L_i - \varphi(\delta)\right),\;\text{ with }\; L_i := \log\frac{\mathbb{Q}(Y_i)}{\mathbb{P}(Y_i)} = e^{\varrho-\bar\varrho}\left(\frac{\bar\varrho}{\varrho}\right)^{Y_i},$$

for some function $\varphi(\cdot)$ we will provide later. If the test statistic is larger than 0, we reject H_0.
We can now use the machinery of [2, Section VI.E] to further specify this test. We first introduce the moment generating function and its Legendre transform:

$$\begin{aligned} M(\vartheta) &= \sum_{k=0}^{\infty}\left(\frac{\bar\varrho^k}{k!}e^{-\bar\varrho}\right)^{\vartheta}\left(\frac{\varrho^k}{k!}e^{-\bar\varrho}\right)^{1-\vartheta} = e^{-\bar\varrho\vartheta-\varrho(1-\vartheta)}\sum_{k=0}^{\infty}\frac{(\bar\varrho^{\vartheta}\varrho^{1-\vartheta})^k}{k!} = e^{-\varrho}e^{(\varrho-\bar\varrho)\vartheta}\exp\left(\varrho\left(\frac{\bar\varrho}{\varrho}\right)^{\vartheta}\right); \\ I(u) &= \sup_{\vartheta}(\vartheta u - \log M(\vartheta)) = \vartheta^\star(u)\,u - \log M(\vartheta^\star(u)), \end{aligned}$$

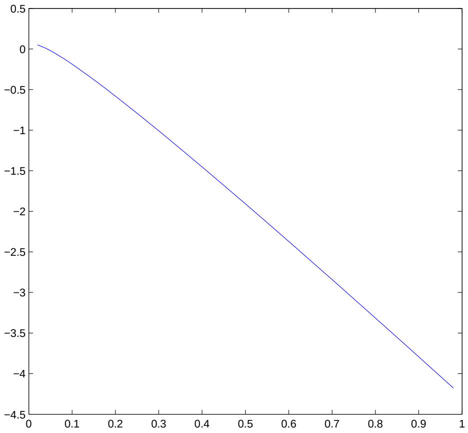

Figure 3: $\varphi(\cdot)$ for $n = 50$, $\varrho = 200$, $\bar{\varrho} = 250$, $\alpha = 0.05$.

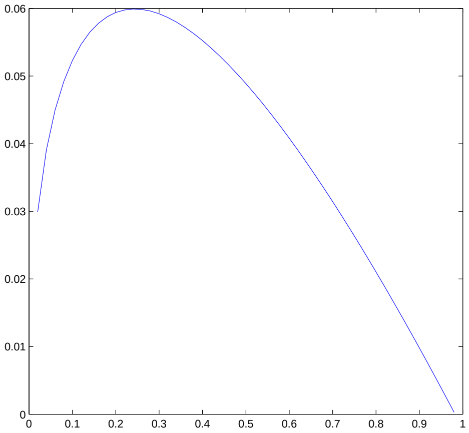

Figure 4: $\varphi(\cdot)$ for $n = 50$, $\varrho = 200$, $\bar{\varrho} = 210$, $\alpha = 0.05$.

where the optimizing $\vartheta(u)$ equals

$$\vartheta^\star(u) = \frac{\log(u + \bar{\varrho} - \varrho) - \log(\varrho \log(\bar{\varrho}/\varrho))}{\log(\bar{\varrho}/\varrho)}.$$

From [2, Section VI.E, Eqn. (46)–(48)], we can compute the decay rate of issuing an alarm under H_0, for a given threshold function $\varphi(\cdot)$:

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\max_{\delta\in[0,1)}\left(\frac{1}{n}\sum_{i=\lfloor n\delta\rfloor+1}^{n} L_i - \varphi(\delta)\right) > 0\right) = \max_{\delta\in[0,1)} \lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\frac{1}{n}\sum_{i=\lfloor n\delta\rfloor+1}^{n} L_i > \varphi(\delta)\right)$$

$$= \max_{\delta\in[0,1)} \delta \cdot \lim_{n\to\infty} \frac{1}{n\delta} \log \mathbb{P}\left(\frac{1}{n\delta}\sum_{i=1}^{\lfloor n\delta\rfloor} L_i > \frac{\varphi(1-\delta)}{\delta}\right) = \max_{\delta\in[0,1)} \delta I\left(\frac{\varphi(1-\delta)}{\delta}\right).$$

In order to get an essentially uniform alarm rate, we can choose $\varphi(\cdot)$ by requiring that

$$\delta I\left(\frac{\varphi(1-\delta)}{\delta}\right) = \alpha^\star, \tag{11}$$

where $\alpha^\star = -\log\alpha/n$; here α is a measure for the likelihood of false alarms (for instance 0.05). Unfortunately, $\varphi(\cdot)$ cannot be solved in closed form, but it can be obtained numerically in a straightforward way (using a standard bisection procedure).

5 Numerical evaluation

In this section we present the results of our numerical experimentation. We first focus on testing the procedures proposed for M/G/∞ in Section 4.2 (for which we could rely on a fairly explicit characterization of the threshold function $\varphi(\cdot)$), and then shift to the setting of Section 3.

Changepoint detection for nearly independent Poisson samples. As mentioned above, we start by presenting numerical results for the setting of Section 4.2. Figures 3 and 4 show the shape of the threshold function $\varphi(\cdot)$. These are computed by numerically solving Equation (11). We observe that $\varphi(\cdot)$ is for $\varrho = 200$ and $\bar{\varrho} = 250$ nearly a straight line.

- *Experiment A1.* We then consider the situation that Y_1 up to Y_{100} are sampled from the evolution of the M/M/∞ queue, at epochs Δ, $2\Delta, \ldots, 100\Delta$; $Y_0 = Y(0)$ is sampled according to the equilibrium distribution (1). For these first 100 observations we chose $\lambda = 200$ and $\mu = 1$, leading to $\varrho = 200$. Then Y_{101} up to Y_{200} are generated in an analogous way, but now with $\lambda = 250$, and hence $\bar{\varrho} = 250$. Assuming a maximum allowable blocking probability of 0.1%, the value $\bar{\varrho} = 250$ corresponds with $C = 291$ lines. It is easily verified that choosing $\Delta = 10$ makes sure that $\mid U_{k,\ell} \mid \cdot \mathbb{P}(B^\star > t) < \varepsilon_{\max}$, for an $\varepsilon_{\max}$ of 0.01, using the procedures developed in Section 4.1.

  We take windows of length 50, that is, we test whether H_0 should be rejected based on data points $Y_i, \ldots, Y_{i+49}$, for $i = 1$ up to 151. The first window in which the influence of $\bar{\varrho}$ is noticeable is therefore window 52. 500 runs are performed. Figure 5 shows the detection ratio as a function of the window id. It indeed hardly shows false alarms up to id 52, and then the detection ratio grows to 1 quite rapidly, as desired.

  Clearly, from window 101 on all observations have been affected by the load change. For window i between 52 and 101, one could (within the window that consists of 50 observations) detect a load change at the earliest at the $(101 - i)$-th observation — this is what could be called the 'true changepoint'; in addition, we call the ratio of $101 - i$ and the window length 50, which is a number between 0 and 1, the 'true delta', in line with the meaning of δ in Section 4.2. Figure 6 provides insight into the spread of the time of detection. It shows that the detection takes place always somewhat later than the true changepoint (as could be expected, as it takes a few observations to 'gather enough statistical evidence'), but the delay is fairly short. In 50% of the cases the delay is less than 8 observations, in 75% less than 12 observations, as can be seen from the graph.

- *Experiment A2.* In Experiment A1 we instantaneously changed ϱ into $\bar{\varrho}$ (which is the value tested against). We now study the effect of a load change to a value $\hat{\varrho} < \bar{\varrho}$. The main question is: despite the fact that $\hat{\varrho}$ is not the value of the load we test against, do we still detect a load change?

  The experiment is done in a similar fashion as Experiment A1: there is a load change from $\varrho = 200$ to $\hat{\varrho} \in \{201, \ldots, 250\}$ from time 100Δ on, and we test against $\bar{\varrho} = 250$. Figure 7 shows that values of $\hat{\varrho}$ up to 225 are hardly detected. For $\hat{\varrho}$ larger than 235 in at least 50% of the runs an alarm has been issued. Only for $\hat{\varrho}$ larger than 245 the changepoint has been detected with high probability (more than, say 85%).

- *Experiment A3.* In Experiment A1 we instantaneously changed ϱ into $\bar{\varrho}$, but a next question is what happens when ϱ *gradually* increases to $\bar{\varrho}$. We performed the same experiment as in Experiment A1, but now the load first has value 200, then starts to increase from observation 76 on in a linear way, to reach value 250 at observation 125 (and hence only from window id 76 on part of the observations were under $\bar{\varrho}$). Figures 8 and 9 are the counterparts of Figures 5 and 6. Compared to Figure 5, the detection ratio in Figure 8 is considerably less steep; note that this could be expected from Experiment A2, as we saw there that only if the load is close to $\bar{\varrho}$ load changes are detected.

Changepoint detection for jump process of M/M/∞. The function $\varphi(\cdot)$ can be determined by executing the computations proposed in Section 3, but due to their intrinsic complexity we chose for the obvious alternative of determining it empirically (that is, by simulation). For $\varrho = 200$ and $\bar{\varrho} = 250$, we thus obtained the curve shown in Figure 10.
We performed the following experiments:

- *Experiment B1.* We consider the following setting, in which we start with Y_0 having a Poisson distribution with mean ϱ, then sample 2500 times according to the measure $\mathbb{P}$, and then 2500 times according to $\mathbb{Q}$. The window size has length 2000. It means that from window id 501 on the measure $\mathbb{Q}$ has impact on the test statistic. Figure 11 shows that we indeed detect the changepoint after window id 1501, but the plot

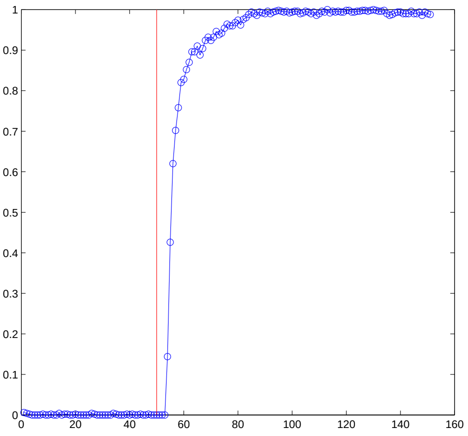

Figure 5: Detection ratio Exp. A1.

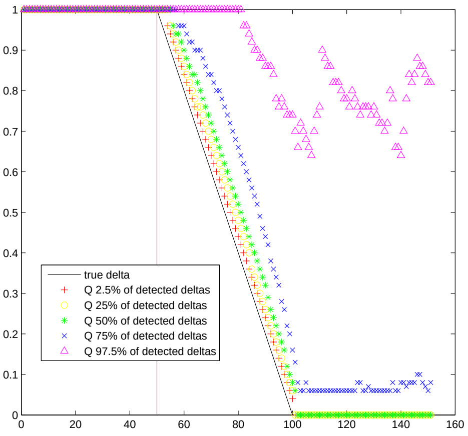

Figure 6: Detection epoch Exp. A1.

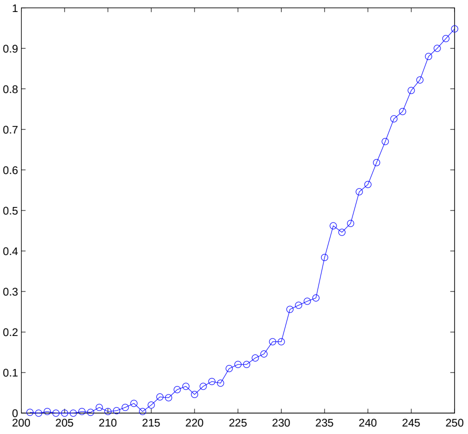

Figure 7: Detection ratio Exp. A2.

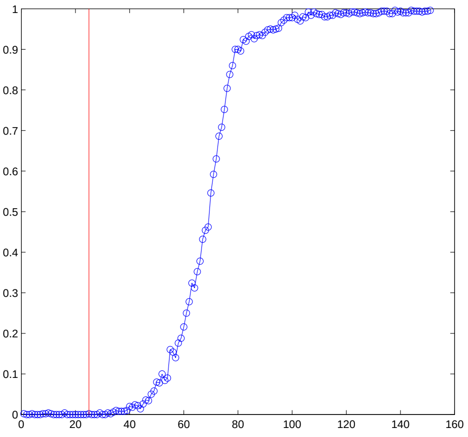

Figure 8: Detection ratio Exp. A3.

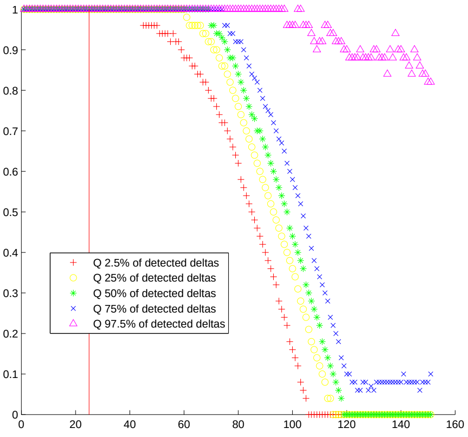

Figure 9: Detection epoch Exp. A3.

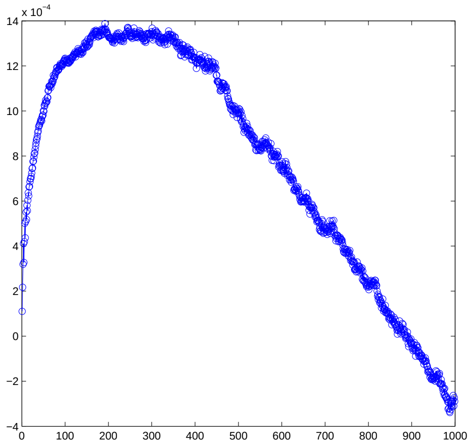

Figure 10: $\varphi(\cdot)$ for $\varrho = 200$, $\bar{\varrho} = 250$, $\alpha = 0.05$.

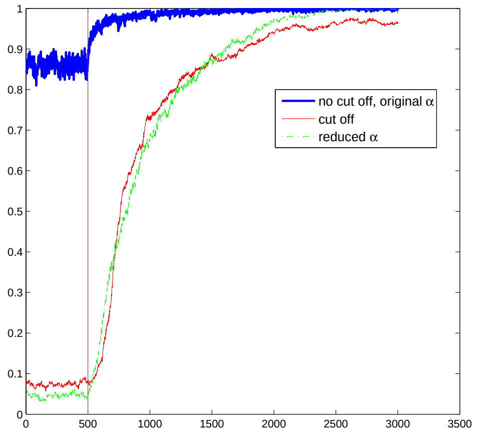

Figure 11: Detection ratio Exps. B1, B2, B3.

also shows that the chance of false alarms (that is, alarms before window id 501) is substantially higher than the 5% that was aimed for, viz. about 87%. The main reason for this phenomenon is that $\varphi(\cdot)$ was (empirically) determined by making

$$\mathbb{P}\left(\sum_{i=\lfloor n\delta\rfloor+1}^{n} L_i > \varphi(\delta)\right) = \alpha \tag{12}$$

for all $\delta \in [0,1)$. The probability of an alarm under H_0, however, is

$$\mathbb{P}\left(\exists \delta \in [0,1) : \sum_{i=\lfloor n\delta\rfloor+1}^{n} L_i > \varphi(\delta)\right),$$

which is evidently larger than probability (12). Apparently this difference can be quite large (although all probabilities involved have the same exponential decay rate, when n grows large).

- *Experiment B2.* There are several ways to make reduce the fraction of false positives. We first consider the effect of changing the criterion

$$\exists \delta \in [0,1) : \sum_{i=\lfloor n\delta\rfloor+1}^{n} L_i > \varphi(\delta) \quad \text{into} \quad \exists \delta \in [0,\delta_{\max}) : \sum_{i=\lfloor n\delta\rfloor+1}^{n} L_i > \varphi(\delta).$$

  Figure 11 shows what happens when imposing such a 'cut off'; we consider the case $\delta_{\max} = 0.9$. We see that the fraction of false alarms is indeed reduced to a number close to 5%, but it is clearly at the expense of detecting load changes after window id 501.

- *Experiment B3.* We now study an alternative to imposing a 'cut off', viz., using an α' which is smaller α. In Figure 11 we considered the detection ratio for $\alpha' = 0.15\%$. Note that this requires redetermination of the function $\varphi(\cdot)$. The effect is very similar to that of Experiment B2: reduction of the false alarm rate, at the expense of loss of detection. It seems that it tuning either $\delta_{\max}$ or α' is necessary to control the false alarm rate.

- *Experiment B4.* In this experiment we start at $Y_0 = 200$ and simulate the first 2500 slots under ϱ, and the last 2500 under $\bar{\varrho}$. The window length is 2000. We imposed a $\delta_{\max}$ of 0.8. In Figure 12 we show the empirical cumulative distribution function of the first epoch that an alarm is issued. Interestingly, its

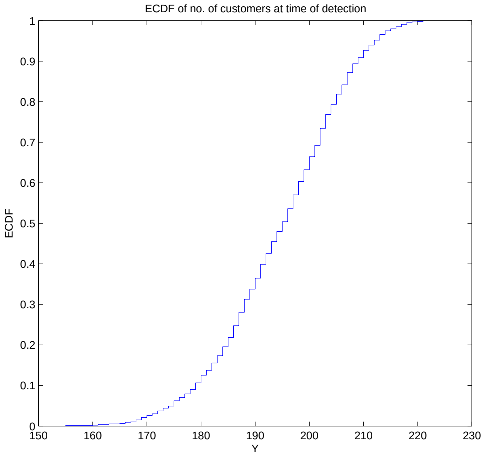

Figure 12: Empirical distribution function of detection epoch Exp. B4.

mean is 236, that is, considerably less than 250. In other words: the alarm is early detected in that the 'new equilibrium', that is, 250, has not been reached yet. This aspect is a significant advantage of this approach over the approach of Section 4.

6 Concluding remarks and discussion

In this paper we developed procedures that are capable of detecting load changes, in a setting in which each connection consumes roughly the same amount of bandwidth (think of VoIP). For the situation that the holding times are exponentially distributed, we designed a testing procedure, relying on large-deviations results. In passing, we found results for the transient of the $\mathrm{M}/\mathrm{M}/\infty$ queue, which are of independent interest.

We observed that this testing framework is applicable to exponential holding times only, and in addition it is numerically rather demanding. We therefore developed an approximative procedure for general holding times. In this procedure we record the number of trunks occupied at equidistant points in time $\Delta, 2\Delta, \ldots$, and we then rely on existing results for sequences of i.i.d. random variables. (Approximate) independence is enforced by choosing Δ sufficiently large; this was made precise by applying new results on the relaxation time of the $\mathrm{M}/\mathrm{G}/\infty$ queue.

The last part of the paper was devoted to numerical experimentation. It was shown that the procedures that we developed were, after elementary tuning, capable of tracking load changes. Special attention was paid to managing the trade-off between the detection ratio and the false alarm rate.

Future research. In this paper we considered traffic generated by applications that require per connection (roughly) the same amount of bandwidth. A next step would be to consider the same problem, but now in a setting where the aggregate traffic stream is the result of many streaming and elastic users. An approach could be to model the traffic process under H_0 by a Gaussian process [8, 10], and to develop changepoint detection procedures for Gaussian processes. Observe, however, that we again have to resolve the issue of dependence between the observations.

A second issue for future research relates to applying the procedures developed in this paper in a real network. This requires extensive evaluation of the testing machinery with real traces.

Acknowledgments

The authors thank COST action TMA (IC0703) for financial support.

References

[1] J. Blanc and E. van Doorn (1984). Relaxation times for queueing systems. In: J.W. de Bakker, M. Hazewinkel, J.K. Lenstra (eds.), *Mathematics and Computer Science*. CWI Monograph 1, North-Holland, Amsterdam, the Netherlands, pp. 139–162.

[2] J. Bucklew. *Large Deviation Techniques in Decision, Simulation, and Estimation.* Wiley, New York, NY, United States, 1990.

[3] J. Chen and A. Gupta (1997). Testing and locating variance changepoints with application to stock prices. *Journal of the American Statistical Association,* Vol. 92, pp. 739–747.

[4] E. van Doorn and A. Zeifman (2009). On the speed of convergence to stationarity of the Erlang loss system. To appear in: *Queueing Systems.*

[5] L. Ho, D. Cavuto, S. Papavassiliou, and A. Zawadzki (2000). Adaptive/automated detection of service anomalies in transaction-oriented WANS: Network analysis, algorithms, implementation, and deployment. *IEEE Journal of Selected Areas in Communications,* Vol. 18, pp. 744–757.

[6] J. Kingman (1962). On queues in which customers are served in random order. *Proceedings of the Cambridge Philosophical Society*, Vol. 58, pp. 79–91.

[7] S.-Y. Lin, J.-C. Liu, and W. Zhao (2007). Adaptive CUSUM for anomaly detection and its application to detect shared congestion. *Technical Report, Texas A&M University,* TAMU-CS-TR-2007-1-2.

[8] M. Mandjes (2007). *Large Deviations of Gaussian Queues.* Wiley, Chichester, UK.

[9] M. Mandjes, I. Saniee, and A. Stolyar (2005). Load characterization, overload prediction, and load anomaly detection for voice over IP traffic. *IEEE Transactions on Neural Networks,* Vol. 16, pp. 1019–1028.

[10] R. van de Meent, M. Mandjes, and A. Pras (2006). Gaussian traffic everywhere? *Proc. 2006 IEEE International Conference on Communications,* Istanbul, Turkey.

[11] G. Münz and G. Carle (2008). Application of forecasting techniques and control charts for traffic anomaly detection. *Proc. 19th ITC Specialist Seminar on Network Usage and Traffic,* Berlin, Germany.

[12] A. Shwartz and A. Weiss (1995). *Large Deviations for Performance Analysis*. Chapman and Hall, London, United Kingdom

[13] D. Siegmund (1985). *Sequential Analysis*. Springer-Verlag, Berlin, Germany.

[14] A. Tartakovsky and V. Veeravalli (2004). Changepoint detection in multichannel and distributed systems with applications. In: N. Mukhopadhyay, S. Datta and S. Chattopadhyay (eds.), *Applications of Sequential Methodologies*. Marcel Dekker, New York, USA, pp. 331–363.

[15] M. Thottan and C. Ji (1998). Proactive anomaly detection using distributed intelligent agents. *IEEE Network,* Vol. 12, pp. 21–27.

[16] M. Thottan and C. Ji (2003). Anomaly detection in IP networks. *IEEE Transactions on Signal Processing,* Vol. 51, pp. 2191–2204.

[17] P. Żuraniewski and D. Rincón (2006). Wavelet transforms and change-point detection algorithms for tracking network traffic fractality. *Proc. NGI 2006,* pp. 216–223.

Centrum Wiskunde & Informatica (CWI) is the national research institute for mathematics and computer science in the Netherlands. The institute's strategy is to concentrate research on four broad, societally relevant themes: earth and life sciences, the data explosion, societal logistics and software as service.

Centrum Wiskunde & Informatica (CWI) is het nationale onderzoeksinstituut op het gebied van wiskunde en informatica. De strategie van het instituut concentreert zich op vier maatschappelijk relevante onderzoeksthema's: aard- en levenswetenschappen, de data-explosie, maatschappelijke logistiek en software als service.

Bezoekadres:
Science Park 123
Amsterdam

Postadres:
Postbus 94079, 1090 GB Amsterdam
Telefoon 020 592 93 33
Fax 020 592 41 99
info@cwi.nl
www.cwi.nl

CWI
Centrum Wiskunde & Informatica