1 Fluid Queue Models for Observed Long-range Dependence in Telecommunication Data

Helmut Gogl*, Michael Greiner*, Manfred Jobmann* & Claudia Klüppelberg†

Zusammenfassung

This contribution introduces briefly subexponential distributions which constitute a large class of heavy-tailed distributions. We investigate their (sometimes desastrous) influence within teletraffic models. Some of these models are reviewed w.r.t. their capability to capture certain stylised features of the data, such as variability of arrival rates, heavy-tailedness of on- and off-periods, and long-range dependence in teletraffic transmission. Then we demonstrate some of the above effects in an explorative data analysis of Munich Universities' intranet data.

1.1 Background and Terminology

Recent measurements of traffic both on local and wide area communications networks have shown some extraordinary behaviour which proves critical for understanding the performance of broad-band networks: the data collected (e.g. packets on Ethernet networks) at Bellcore [Leland et al., 1994], frames from Variable-Bit-Rate (VBR) video service ([Beran et al., 1995], [Garrett & Willinger, 1994]), FTP data connections, NNTP, and WWW arrivals in wide area traffic show enormous *variability of arrival rates* indicating that a Poisson process may be an insufficient model for packet traffic, see [Paxson & Floyd, 1995].

Moreover, time series of teletraffic data show a *long-range dependence effect*, meaning that the current state of a time series has a strong dependency on the remote past. Definitions vary from author to author, but a commonly accepted definition in a covariance stationary time series is that a process (X_n) has long-range dependence, if the correlation coefficients $\mathrm{corr}(X_0, X_n)$ decrease to 0 at a rate slower than exponential. Admittedly, many authors even require that the autocorrelation coefficients are not summable, but we want some more flexibility in modeling. The exponential as a reference rate is motivated by the fact that for linear models as for instance causal and invertible ARMA (autoregressive-moving average) models the correlation coefficients decrease to 0 exponentially fast, hence long-range dependence in the above sense cannot be modeled by such traditional models.

*Institut für Informatik, Technische Universität München, Germany
E-mail: m.greiner@ieee.org, {gogl, jobmann}@in.tum.de
†Zentrum Mathematik, Technische Universität München, Germany
E-mail: cklu@ma.tum.de

Various models have been suggested to capture these effects observed in teletraffic data. They range from traditional queueing models to sophisticated on/off models ([Heath et al., 1996a], [Heath et al., 1996b], [Jelenković & Lazar, 1997]), shot noise models ([Kurtz, 1996]), Markov modulated queues ([Jelenković & Lazar, 1996], [Jelenković & Lazar, 1998]), and fractional Brownian motion ([Leland et al., 1994], [Taqqu & Levy, 1986], [Willinger et al., 1997]).

The aim of this article is

- to clarify the various notions of heavy-tailed distributions as used in the queueing and network area,
- to describe the consequences of subexponential input distributions to the distributional behaviour of the output processes,
- to discuss possible models where heavy-tailed or dependent input may explain the observed long-range dependence in teletraffic data.

Similar material can be found in a more detailed paper [Greiner et al., 1999] by the same authors. This paper is organised as follows. In Section 1.2 we summarise various notions and properties of heavy-tailed distributions, the outer frame being built by the class of subexponential distributions.

In Section 1.3 we indicate what disasters heavy-tailed input can result in classical queueing models. Such models have been taken as basis for more sophisticated models in teletraffic data transmission. For instance, buffer sizes correspond to workload processes. We want to gain some qualitative insight into the effect of heavy tails on performance measures like waiting time distribution and queue length.

In Section 1.4 we discuss some models within the queueing context which have been suggested as appropriate models for teletraffic data. We derive certain performance measures within such models. Section 1.5 concludes the paper with an explorative data analysis of Munich Universities' intranet data, measured at a network access point of the German Broadband Research Network (B-WiN).

1.2 Subexponential distributions

In an intuitive approach we consider heavy-tailed distributions as ideal models to allow for very large values in a sample from such a distribution. It seems to be common agreement that the tail of a heavy-tailed distribution function (df) decreases more slowly than any exponential tail, i.e. for a heavy-tailed random variable (rv) X

$$P(X > x)e^{\varepsilon x} \to \infty\,, \quad x \to \infty\,,$$

for any positive ε. This class includes for instance *Pareto, lognormal*, and *heavy-tailed Weibull distributions*. In certain applications, in particular in queueing theory, some more structure for the distribution tail is needed, which leads to the definition of subexponential distributions.

In this section we summarise definitions and properties of subexponential dfs concentrating on those properties we shall need later in the paper. For further properties see [Baccelli et al., 1999] in this volume. A more complete account on subexponential dfs can be found in [Embrechts et al., 1997] or the review article [Goldie & Klüppelberg, 1998], from which results are quoted freely.

We present two defining properties of subexponential dfs. The first, more analytic one, is motivated by the Pollaczek–Khinchin formula (1.7) below, while the second probabilistic one provides a more intuitive interpretation of subexponentiality.

Definition 1.2.1 (Subexponential distribution function)
Let $(X_i)_{i\in\mathrm{N}}$ be iid positive rvs with df F such that $F(x)<1$ for all $x>0$. Denote

$$\overline{F}(x) = 1 - F(x)\,,\quad x\ge 0\,,$$

the tail of F and

$$\overline{F}^{n*}(x) = 1 - F^{n*}(x) = P(X_1+\cdots+X_n > x)\,,\quad x\ge 0\,,$$

the tail of the n–fold convolution of F. F is a subexponential df ($F\in\mathcal{S}$) *if one of the following equivalent conditions holds:*

(a) $\lim_{x\to\infty}\dfrac{\overline{F}^{n*}(x)}{\overline{F}(x)} = n$ *for some (all)* $n\ge 2$,

(b) $\lim_{x\to\infty}\dfrac{P(X_1+\cdots+X_n > x)}{P(\max(X_1,\ldots,X_n)>x)} = 1$ *for some (equivalently all)* $n\ge 2$. 2

Remark 1.2.2
(i) Definition 1.2.1(b) provides a physical interpretation of subexponentiality: the sum of n iid subexponential rvs is likely to be large if and only if their maximum is. This accounts for extremely large values in a subexponential sample.

(ii) In order to interchange limits and infinite sums or limits and integrals we need the following uniform bound for the quotient in Definition 1.2.1(a). For every $\varepsilon>0$ there exists some positive constant $K(\varepsilon)$ such that

$$\frac{\overline{F}^{n*}(x)}{\overline{F}(x)} \le K(\varepsilon)(1+\varepsilon)^n$$

holds for all $n\in\mathrm{N}$ and $x>0$.

(iii) Definition 1.2.1(b) demonstrates the heavy–tailedness of subexponential dfs. It is further substantiated by the implications

$$F\in\mathcal{S} \implies \lim_{x\to\infty}\frac{\overline{F}(x+y)}{\overline{F}(x)} = 1 \quad \forall\, y\in\mathrm{R} \tag{1.1}$$

$$\implies \overline{F}(x)/e^{-\varepsilon x}\to\infty\,,\quad x\to\infty\,,\quad \forall\,\varepsilon>0\,. \tag{1.2}$$

$$\implies \int_0^\infty e^{\varepsilon x}\,dF(x) = \infty \quad \forall\,\varepsilon>0 \tag{1.3}$$

Property (1.2) accounts for the name subexponential df: the tail of F decreases more slowly than any exponential tail. Property (1.3) shows that subexponential dfs have no exponential moments. This prevents any method being applicable that requires the existence of exponential moments.

(iv) An important (though much smaller) subclass of $\mathcal{S}$ is the class of *dfs with regularly varying tail.* We write $\overline{F} \in \mathcal{R}(-\alpha)$ if

$$\lim_{x\to\infty} \frac{\overline{F}(tx)}{\overline{F}(x)} = t^{-\alpha}, \quad t > 0.$$

Since $\overline{F}$ is non-increasing, the index $\alpha \in [0,\infty)$. $\overline{F} \in \mathcal{R}(-\alpha)$ is equivalent to $\overline{F}(x) = x^{-\alpha}L(x)$ for some slowly varying function L; we write $L \in \mathcal{R}(0)$. Examples for L are constants, functions converging to a constant, logarithms, or iterated logarithms. If $\overline{F} \in \mathcal{R}(-\alpha)$ for $\alpha < 1$, then F has infinite mean; if $\alpha < 2$, then F has infinite variance. The class of regularly varying functions has the advantage to allow the application of Abel-Tauber theorems, quite a common tool in applied probability. Unfortunately, there is no characterisation of a subexponential distribution in terms of its Laplace transforms apart from (1.3). 2

The class of df given by property (1.1) is important in its own right and defines a special class.

Definition 1.2.3 *The df F of a positive rv X such that $F(x) < 1$ for all $x > 0$ belongs to the class $\mathcal{L}$ if*

$$\lim_{x\to\infty} P(X - x > y | X > x) = \lim_{x\to\infty} \frac{\overline{F}(x+y)}{\overline{F}(x)} = 1 \quad \forall\, y \in \mathrm{R}. \tag{1.4}$$

2

For positive y this is the df of the overshoot over a threshold x. For the class $\mathcal{L}$ this overshoot degenerates, i.e. it becomes infinite.

Define for a positive rv X with df F having finite mean μ its *equilibrium distribution* (or *integrated tail distribution*) by

$$F_I(x) = \frac{1}{\mu}\int_0^x \overline{F}(y)\,dy, \quad x \geq 0. \tag{1.5}$$

The following result is a consequence of the property (1.4).

Corollary 1.2.4 *If $F \in \mathcal{L}$, then*

$$\lim_{x\to\infty} \frac{\overline{F}(x)}{\overline{F}_I(x)} = 0. \qquad \blacksquare$$

We shall need the following result on convolution closure of $\mathcal{S}$. In the sequel $f(x) \sim g(x)$, $x \to \infty$, means that $\lim_{x\to\infty} f(x)/g(x) = 1$.

Theorem 1.2.1 *Let $F \in \mathcal{S}$ and $\overline{G}_i(x) \sim c_i\overline{F}(x)$ for $i = 1, 2$, where $c_i \in [0,\infty)$. Then $\overline{G_1 * G_2}(x) \sim (c_1 + c_2)\overline{F}(x)$.* $\blacksquare$

We conclude the section with some examples; more are to be found in [Embrechts et al., 1997] or [Goldie & Klüppelberg, 1998].

Example 1.2.5 (Subexponential dfs)
(i) If $\overline{F}(x) \sim cx^{-\alpha}$, $x \to \infty$, for some $\alpha \geq 0$ and $c > 0$, then F is called *power-tailed df*. Notice that Pareto distributions themselves belong to this class.

(ii) If $\overline{F}(x) = x^{-\alpha}L(x)$ (i.e. regularly varying) for some $\alpha \geq 0$ and $L \in \mathcal{R}(0)$, then F is also called *Pareto-like*.

(iii) The Weibull-like df with tail $\overline{F}(x) \sim r(x)\exp(-x^{\beta})$ for $\beta \in (0,1)$ and regularly varying function r is subexponential.

(iv) The lognormal df with density $f(x) = \frac{1}{\sqrt{2\pi}\sigma x}\exp(-(\ln x - \mu)^2/(2\sigma^2))$ is subexponential. 2

1.3 Classical queueing models and subexponentials

The first papers to recognise the importance of subexponential dfs in queueing theory were [Cohen, 1973], [Pakes, 1975], and [Smith, 1972].

We consider an M/G/1 queue with arrival rate $\lambda > 0$, service time df F having finite mean μ and equilibrium df $F_I(x) = \int_0^x \overline{F}(y)\,dy/\mu$. We assume that the queue is stable, i.e. its traffic intensity $\rho = \lambda\mu < 1$.

Denote for $n \in \mathrm{N}_0$ by W_n the waiting time of the nth customer. Then the sequence (W_n) satisfies Lindley's equation which is given by the following recursion:

$$W_{n+1} = (W_n + X_n - U_{n+1})^+\,, \quad n = 0,1,2,\ldots, W_0 = 0\,, \tag{1.6}$$

where X_n is the service time of the nth customer and $U_{n+1} = T_{n+1} - T_n$ is the interarrival time between nth and $(n+1)$st customer. It can be shown (see e.g. [Feller, 1971] or [Resnick, 1992]) that

$$W_n \stackrel{d}{=} \left(\max_{0\leq k\leq n}\sum_{i=0}^{k}(X_i - U_{i+1})\right)^+\,, \quad n = 0,1,2,\ldots\,,$$

and

$$E(X_i - U_{i+1}) = \mu - \lambda^{-1} = \lambda^{-1}(\lambda\mu - 1) < 0\,, \quad i = 0,1,2,\ldots\,.$$

Then W_n is distributed as the maximum of a random walk with negative drift. Hence $W_n \stackrel{\text{a.s.}}{\to} W_\infty$ where W_∞ is a finite rv with df $W(t)$, $t \geq 0$. For (U_n) iid exponential rvs the stationary waiting time distribution $W(t)$, $t \geq 0$, is given by the Pollaczek-Khinchin formula:

$$W(t) = (1-\rho)\sum_{n=0}^{\infty}\rho^n F_I^{n*}(t)\,, \quad t \geq 0\,, \tag{1.7}$$

where $F_I^{0*} = I_{[0,\infty)}$ is the df of Dirac (unit) measure at 0. In this representation ρF_I is the *ladder height df* of the embedded random walk. The infinite series on the rhs of (1.7) defines a defective renewal measure ($\rho F_I(x) \to \rho < 1$ as $x \to \infty$), and the corresponding renewal process is transient: the sequence of renewals (ladder heights) eventually stops, and at each ladder height $1-\rho$ is the probability of termination then and there. This is a consequence of the negative drift of the embedded random walk, which is ensured by $\rho < 1$. For details see [Feller, 1971], Section VI.9.

We rewrite formula (1.7) in terms of the tails,

$$\overline{W}(t) = (1-\rho)\sum_{n=1}^{\infty} \rho^n \overline{F}_I^{n*}(t)\,, \quad t \geq 0\,.$$

Dividing both sides by $\overline{F}_I(t)$, we see that Definition 1.2.1(a) yields an asymptotic estimate for $\overline{W}(t)$ provided that one can safely interchange the limit and the infinite sum. This is ensured by Remark 1.2.2(ii) and Lebesgue's dominated convergence theorem.

It turns out that this is not just a consequence of subexponentiality, but is characterised by it, as follows from the following theorems (see [Embrechts & Veraverbeke, 1982]).

Theorem 1.3.1 (Stationary waiting time in the M/G/1 queue)

$$W \in \mathcal{S} \iff F_I \in \mathcal{S} \iff \lim_{t\to\infty} \frac{\overline{W}(t)}{\overline{F}_I(t)} = \frac{\rho}{1-\rho}\,. \qquad \blacksquare$$

This theorem can be generalised to a GI/G/1 queue, where the arrival process is an arbitrary renewal process.

Theorem 1.3.2 (Stationary waiting time in the GI/G/1 queue)

$$W \in \mathcal{S} \iff F_I \in \mathcal{S} \implies \lim_{t\to\infty} \frac{\overline{W}(t)}{\overline{F}_I(t)} = \frac{\rho}{1-\rho}\,. \qquad \blacksquare$$

The next question to ask is how and when high workloads (i.e. large buffer contents or buffer overflows) happen in such a classical queueing system like M/G/1 or GI/G/1.

The *waiting time process* or *workload process* $(V_t)_{t\geq 0}$ denotes the sum of all service times (whole or remaining) in the system.

Important information about a queueing system with heavy-tailed service time can be gained by considering high excursions of the workload process $(V_t)_{t\geq 0}$ after a buffer overflow happened. This would be the traffic being lost or to be stored elsewhere. For mathematical details and proofs we refer to [Asmussen & Klüppelberg, 1997].

Another quantity of interest is the *queue length in system* (stationary number of customers in system), we denote it by L. In an M/G/1 queue, under FIFO (first in first out) and when the nth customer's sojourn time D_n in the system (total time spent in the system from arrival to departure) is independent of future interarrival times, then distributional Little's law holds, meaning that

$$L \stackrel{d}{=} N_D\,, \tag{1.8}$$

where (N_t) denotes a time stationary version of the renewal counting process (with first arrival time distributed according to the equilibrium distribution) and D denotes the stationary sojourn time. Notice that $D = W_\infty + X$ (independent sum) is the sum of the stationary waiting time and the service time. If the service time X is subexponential, then by Corollary 1.2.4 the tail of F_I is heavier than the tail of F. Hence W_∞ dominates X in the sum and (see e.g. Theorem 5.1 of [Goldie & Klüppelberg, 1998])

$$P(D > x) \sim P(W_\infty > x)\,, \quad x \to \infty\,.$$

For heavy-tailed service times, the following result has been proved in [Asmussen et al., 1998].

Theorem 1.3.3 *Consider an* M/G/1 *queue with arrival rate* $\lambda > 0$ *and traffic intensity* $\rho < 1$. *Denote the service time df* F *and assume that the equilibrium df* $F_I \in \mathcal{S}$. *Let* W_∞ *denote the stationary waiting time. Assume that*

$$\lim_{x\to\infty} \frac{\overline{F}(xe^{y/\sqrt{x}})}{\overline{F}(x)} = 1\,, \qquad \textit{locally uniformly in } y \in \mathrm{R}\,. \tag{1.9}$$

Then the stationary queue length L *satisfies*

$$P(L > k) \sim P(\lambda W_\infty > k) \sim \frac{\rho}{1-\rho}\overline{F}_I(k/\lambda)\,, \quad k \to \infty\,. \tag{1.10}$$

■

The extra condition (1.9) is a tail condition guaranteeing that the tail decreases more slowly than the Weibull tail $\exp(-\sqrt{x})$, hence the result holds for any $\overline{F} \in \mathcal{R}(-\alpha)$ for $\alpha > 1$, lognormal df, and Weibull distributions with tail $\overline{F}(x) \sim \exp(-x^\beta)$ for $\beta < 0.5$. In these cases the queue length becomes large only by a large service time. The Poisson arrivals do not contribute substantially to the queue length (only via the arrival rate in ρ).

When the service time is lighter than the tail of a Weibull distribution with parameter $\beta = 0.5$, the number of customers arriving comes into the picture as well. Then the combination of the number of customers and the likely large service time makes the queue-length large, for explicit formulas see [Asmussen et al., 1998].

1.4 Long-range dependence and heavy tails in teletraffic data

In recent years the question has been raised whether classical queueing and network models may not be too simplistic to model teletraffic networks. The heavy tails and dependence structure exhibited in explorative data analysis of teletraffic data cannot always be explained in the frame of such classical models, where heavy-tailed output is only possible by heavy-tailed input as we have seen in the preceding section. However, quite a variety of models have a regeneration structure leading to Lindley's equation (1.6) for certain quantities of interest and hence, as in Section 1.3, to asymptotic results.

As an example consider the Asynchronous Transfer Mode (ATM) based broadband networks with statistical multiplexing (SMUX). Most of the multiplexed entities are calls originating from various sources. In order to operate properly, each of these calls has to satisfy some quality of service requirements (QoS). QoS requirements are usually bounds on performance measures characterising the dynamic behaviour of the multiplexed traffic. The most basic model of a SMUX is an infinite buffer single server queue with a work conserving scheduler. The fundamental performance measure is the tail of the stationary waiting time distribution $P(W_\infty > x)$.

Numerous investigations have shown that the arrival processes that arise in ATM networks (like voice and video) have a very complex statistical structure; an especially troublesome characteristic is the high dependency. Modeling of this high dependency usually leads to analytically very complex statistical characteristics, typically making the associated evaluation of the queue length distribution intractable. However, because of the stringent QoS requirements in ATM, in particular the tail of the stationary waiting time distribution in the domain of very small probabilities is needed.

1.4.1 A simple on/off-model

A simple physical model is based on a sequence of points $(T_n)_{n\in\mathrm{N}}$ which constitute a stationary renewal process, i.e. $P(T_1 > t) = \overline{F}_I(t)$, $t \geq 0$, is the equilibrium df to F, where F is the df of all interarrival times after the first one. Assume further that the interarrival times are heavy-tailed in the sense that $F \in \mathcal{L}$, which implies that $\overline{F}(x) = o(\overline{F}_I(x))$ (see Corollary 1.2.4).

Suppose that each of the interarrivals is either an on-period, where traffic is transmitted, or an off-period, where no traffic is transmitted. Assume furthermore that each interarrival is randomly chosen as on- or off-period by a Markov mechanism. Define the continuous time fluid process $(A_t)_{t\geq 0}$ as being 1 during an on-period, and 0 during an off-period. The process (A_t) is sometimes called a "Markov chain embedded in a stationary renewal" process. This process gives rise to long-range dependence, as is seen from the following argument. Long-range dependence is defined by the property that $\mathrm{corr}(A_0, A_t)$ decreases to 0 more slowly than exponentially. Observe that for large values of t, $\mathrm{corr}(A_0, A_t)$ is linearly related to the probability that the activity period that covers 0 is still going on at t. The resulting distribution of the residual activity period has df F_I, i.e. has by Corollary 1.2.4 a heavier tail than the usual interarrival times. Consequently the autocorrelation function decreases like $\overline{F}_I$, implying long-range dependence. This can be made precise as follows.

Theorem 1.4.1 *Let $(T_n)_{n\in\mathrm{N}}$ be a stationary renewal point process whose interarrival times have df $F \in \mathcal{L}$ and $F_I \in \mathcal{S}$. Let furthermore $(A_t)_{t\geq 0}$ be the embedded Markov chain with state space $E = \{0,1\}$, transition probabilities p_{ij} for $i,j = 0,1$, and stationary distribution π_i for $i = 0,1$. We consider the stationary version for $(A_t)_{t\geq 0}$. Then*

$$\mathrm{cov}(A_0, A_t) \quad \sim \quad \mathrm{var}(A_0)\overline{F}_I(t)\,, \quad t \to \infty\,, \tag{1.11}$$

where F_I is the equilibrium df of F.

Proof The proof is a special case of the argument given in [Jelenković & Lazar, 1998], cf. [Greiner et al., 1999]. ■

If F has regularly varying tail, then we can apply Karamata's theorem and obtain the following.

Corollary 1.4.1 *Let $\overline{F}(x) = x^{-\alpha}L(x)$, $x \geq 0$, for $\alpha > 1$ and $L \in \mathcal{R}(0)$. Then*

$$\operatorname{cov}(A_0, A_t) \sim \operatorname{var}(A_0)\frac{1}{(\alpha-1)\mu}\, t^{-(\alpha-1)}L(t)\,, \quad t \to \infty\,.$$

For $\alpha \in (1,2)$ the autocorrelation function is not integrable. ■

This model is a simplified idealisation. It assumes that the tails for both, the on- and off-periods, have the same relative heaviness. As [Willinger et al., 1997] point out, this may not be consistent with telecommunication data. However, the example illustrates in a simple way how heavy tails can induce long-range dependence.

1.4.2 On/off models with different on-time and off-time distributions

In this section we present a generalisation of the above model with alternating on- and off-periods. We follow the presentation in [Heath et al., 1996a]. The non-negative iid rvs $(\xi_{\text{on}}, \xi_n)_{n\in\mathbb{N}_0}$ represent the on-periods, and the non-negative iid rvs $(\eta_{\text{off}}, \eta_n)_{n\in\mathbb{N}_0}$ the off-periods. On- and off-periods are assumed to be independent, the on-periods have common df F_{on}, the off-periods have common df F_{off}, both have finite mean μ_{on} and μ_{off} and we set $\mu = \mu_{\text{on}} + \mu_{\text{off}}$.

There exists a stationary renewal process with interarrival times distributed as $\xi_{\text{on}} + \eta_{\text{on}}$. This means that each renewal point is the starting point of an on-period, and each interarrival time consists of exactly one on- and one off-period. In a stationary version of the process we see in 0 either an on-period or an off-period. If we see an on-period, then an off-period follows before the renewal point T_1. If we see an off-period, then the renewal point T_1 follows immediately after this off-period. To capture the time interval $[0, T_1)$ we define independent rvs ξ_I, η_I, and B independent of $(\eta_{\text{off}}, \xi_n, \eta_n)_{n\in\mathbb{N}}$, where ξ_I has df $F_{\text{on},I}(x) = (1/\mu_{\text{on}}) \int_0^x \overline{F}_{\text{on}}(y)dy$, η_I has df $F_{\text{off},I}(x) = (1/\mu_{\text{off}}) \int_0^x \overline{F}_{\text{off}}(y)dy$, and B is a Bernoulli rv with success probability $P(B=1) = \mu_{\text{on}}/\mu$. Then the stationary situation is modeled by

$$T_1 = B(\xi_I + \eta_{\text{off}}) + (1-B)\eta_I\,.$$

The corresponding on/off process $(A_t)_{t\geq 0}$ which is equal to 1 if t falls in an on-period and 0 if t falls in an off-period can be defined in terms of $(T_n)_{n\in\mathbb{N}}$ as follows:

$$A_t = B I_{\{[0,\xi_I)\}}(t) + \sum_{n=1}^{\infty} I_{[T_n, T_n+\xi_{n+1})}(t)\,, \quad t > 0\,.$$

Thus if $t \geq T_1$

$$A_t = \begin{cases} 1 & \text{if } T_n \leq t < T_n + \xi_{n+1}\,, \\ 0 & \text{if } T_n + \xi_{n+1} \leq t < T_{n+1}\,, \end{cases}$$

while for $t \in [0, T_1)$ we have

$$A_t = \begin{cases} 1 & \text{if } B = 1 \text{ and } 0 \leq t < \xi_I\,, \\ 0 & \text{otherwise}\,. \end{cases}$$

With this construction, (A_t) is strictly stationary ((A_t) inherits the stationarity from the stationary renewal sequence (T_n)). Moreover, $P(A_t = 1) = \mu_{\rm on}/\mu$. To see this, write

$$P(A_t = 1) = EA_t = P(B = 1)P(\xi_I > t) + \sum_{n=1}^{\infty} P(T_n \le t < T_n + \xi_{n+1})\,. \qquad (1.12)$$

Recall that the renewal function of the stationary sequence (T_n) is equal to

$$U(t) = \sum_{n=1}^{\infty} P(T_n \le t) = \frac{t}{\mu}\,, \quad t > 0\,.$$

Now we can evaluate the infinite sum in (1.12) as

$$\begin{aligned}\sum_{n=1}^{\infty}\int_0^t \overline{F}_{\rm on}(t-u)dP(T_n \le u) &= \int_0^t \overline{F}_{\rm on}(t-u)dU(u)\\ = \frac{1}{\mu}\int_0^t \overline{F}_{\rm on}(t-u)du &= \frac{\mu_{\rm on}}{\mu}F_{{\rm on},I}(t)\,, t > 0\,.\end{aligned}$$

Hence,

$$EA_t = \frac{\mu_{\rm on}}{\mu}(P(\xi_I > t) + P(\xi_I \le t)) = \frac{\mu_{\rm on}}{\mu}\,.$$

The main theorem in [Heath et al., 1996a] describes the autocovariance function of the process $(A_t)_{t\ge 0}$.

Theorem 1.4.2 *Assume that* $\overline{F}_{\rm on}(t) = t^{-\alpha}L(t)$, $t \ge 0$, *where* $L \in \mathcal{R}(0)$ *and* $\alpha \in (1,2)$. *Assume also that* $\overline{F}_{\rm off}(t) = o(\overline{F}_{\rm on}(t))$ *as* $t \to \infty$ *and that* $\xi_{\rm on} + \eta_{\rm off}$ *is non-degenerate. Then*

$$\mathrm{cov}(A_0, A_t) \sim \frac{\mu_{\rm off}^2}{(\alpha-1)\mu^3} t^{-(\alpha-1)}L(t)\,, \quad t \to \infty\,. \qquad \blacksquare$$

The proof is rather technical and even a sketch of it would go beyond the scope of this paper. The essential argument relies on the rate of convergence in Smith's key renewal theorem for heavy-tailed interarrival times.

Define

$$A_t^* = \int_0^t A_u du\,, \quad t \ge 0\,,$$

the cumulative input to the system up to time t. Since $EA_t = \mu_{\rm on}/\mu$, we have $A_t^*/t \stackrel{\rm a.s.}{\to} \mu_{\rm on}/\mu$ as $t \to \infty$. Assume the system has a constant release rate $r > 0$ if the buffer is not empty. For stability we require $\mu_{\rm on}/\mu < r < 1$ (recall that 1 is the input rate of traffic into the system). If the buffer is empty, we set $r = 0$.

Define the *buffer contents process* $(V_t)_{t\ge 0}$ (which corresponds to the workload process) by the stochastic differential equation

$$dV_t = dA_t^* - r(V_t)dt\,, \qquad (1.13)$$

for given initial rv V_0. During an on-period traffic enters at net rate $1-r$ and during an off-period the buffer content is released at rate r. The (T_n) are regeneration times of the contents process $(V_t)_{t\geq 0}$ which is stationary ergodic. Consider the change in the buffer between T_n and T_{n+1}. We see that

$$V_{T_{n+1}} = (V_{T_n} + (1-r)\xi_{n+1} - r\eta_{n+1})^+ , \quad n \in \mathrm{N}_0 , \tag{1.14}$$

where the increments have mean

$$E\big((1-r)\xi_{n+1} - r\eta_{n+1}\big) = (1-r)\mu_{\rm on} - r\mu_{\rm off} = \mu_{\rm on} - r\mu < 0 .$$

Hence V_{T_n} satisfies Lindley's equation (1.6) and the limit variable is determined by

$$V_{T_n} \stackrel{d}{\to} V_{T_\infty} = \left(\max_{n\geq 0} \sum_{i=0}^{n} ((1-r)\xi_{n+1} - r\eta_{n+1}) \right)^+ . \tag{1.15}$$

In a subexponential regime the tail behaviour of the stationary waiting time distribution is given by Theorem 1.3.2. The rvs $(1-r)\xi_n$ correspond to the service times in this theorem. The constant ρ is given by

$$\rho \;=\; \frac{\text{rate of the arrival process}}{\text{rate of the service process}} \;=\; \frac{(1-r)\mu_{\rm on}}{r\mu_{\rm off}} .$$

Now we can reformulate Theorem 1.3.2 in our context and obtain the following result.

Proposition 1.4.2 *Let W denote the df of V_{T_∞} and set $\rho = (1-r)\mu_{\rm on}/(r\mu_{\rm off})$. Then*

$$W \in \mathcal{S} \iff F_{{\rm on},I} \in \mathcal{S} \Longrightarrow \overline{W}(x) \sim \frac{\rho}{1-\rho}\overline{F}_{{\rm on},I}\left(\frac{x}{1-r}\right) , \quad x \to \infty . \qquad \blacksquare$$

If $F_{\rm on}$ has regularly varying tail, then we can apply Karamata's theorem and obtain the following.

Corollary 1.4.3 *Let $\overline{F}_{\rm on}(x) = x^{-\alpha}L(x)$, $x \geq 0$, for $\alpha > 1$ and $L \in \mathcal{R}(0)$. Then*

$$\overline{W}(x) \sim \frac{\rho}{1-\rho}\frac{(1-r)^{\alpha-1}}{\mu_{\rm on}(\alpha-1)}x^{-(\alpha-1)}L(x) =: bx^{-(\alpha-1)}L(x) , \quad x \to \infty . \qquad \blacksquare$$

The buffer content process has its cycle maxima not at the points (T_n), but at the points $(T_n + \xi_n)$, hence the following result is not surprising. It shows that the distributional limit of V_t, the buffer contents at some time t, has an even heavier tail than that of V_{T_n}, the buffer contents at renewal points.

Proposition 1.4.4 *Let $\overline{F}_{\rm on}(x) = x^{-\alpha}L(x)$, $x \geq 0$, for $\alpha > 1$ and $L \in \mathcal{R}(0)$. Define b as in Corollary 1.4.3. Then $V_t \stackrel{d}{\to} V_\infty$ and*

$$P(V_\infty > x) \sim \left(b + \frac{(1-r)^{\alpha-1}}{\mu(\alpha-1)}\right) x^{-(\alpha-1)}L(x) , \quad x \to \infty . \qquad \blacksquare$$

Remark 1.4.5
(i) The buffer contents process as above has already been considered in [Boxma, 1996], also in the context of telecommunication traffic. He requires (as [Heath et al., 1996a] do) regularly varying tails for F_{on}. Boxma's model is more general in the sense that he allows for several ($N \in \mathbb{N}$) sources transmitting data to one buffer. But as he states, the analytical treatment becomes then rather complicated. Apart from the case of one source, he treats some examples for $N > 1$.
(ii) The asymptotic distribution of cycle maxima in a more general, namely subexponential, queueing setting has been derived in [Asmussen, 1996]. 2

1.5 An explorative data analysis

In this section we want to discuss whether real telecommunication traffic justifies the application of on/off models as introduced in section 1.4 as source models. In particular, does real telecommunication traffic exhibit bursty arrival patterns such that power-tailed or Pareto-like distributions of the on/off periods are appropriate? The traffic we consider in the sequel was observed at a network access point (Customer Service Switch) of the Munich Universities' intranet to the German Broadband Research Network (B-WiN) at the Leibniz Research and Supercomputing Center of the Bavarian Academy of Sciences. It consists of pure IP (Internet Protocol) data traffic restricted to the network applications/protocols HTTP, FTP, NNTP, SMTP, and TELNET.

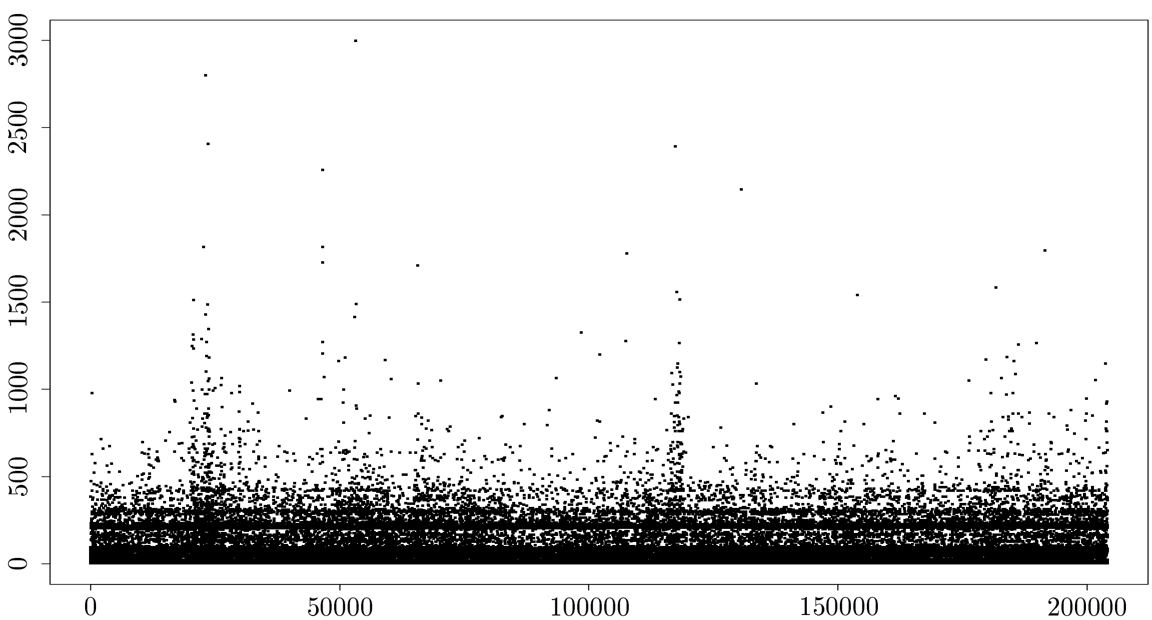

Abbildung 1.1: Length of on-periods in μs; i.e. lengths of cell bursts, extracted from a total of 1 690 730 ATM cells. This comprises 247,995 IP packets which were captured within approximately two minutes in the afternoon of December 23, 1997.

From that data traffic the lengths of on-periods were extracted with the burst analysis described in [Gogl et al., 1999] in this volume (see there and also [Gogl, 1998] for more details on how the data were collected). A plot of the lengths of the successive on-periods is presented in Figure 1.1. The scatterness of the data and in particular the frequency and size of the peaks (large on-periods) suggest that further investigation may reveal heavy-tailedness of the on-periods.

Various graphical tools are presented in the literature to investigate the tail behavior of a distribution. For details we refer to Section 6.2 of [EMBRECHTS ET AL., 1997]. For such methods applied to telecommunication data see [RESNICK, 1997].

In order to check for infinite variance of the on- and off-periods, respectively, we start with applying the *limit distribution (LD) test* (see e.g. [CROVELLA ET AL., 1998]) to our datasets: each dataset (X_i) of size N is aggregated over blocks $k = 1, 2, \ldots, \lfloor \frac{N}{m} \rfloor$ of size m with increasing m, i.e.

$$X_k^{(m)} = \sum_{i=(k-1)m+1}^{km} X_i \,.$$

If the tail probabilities of the aggregated datasets, plotted on log-scale, decline exponentially for increasing m, the original dataset follows a finite-variance distribution (due to the CLT). On the other hand, if the slopes α of the tail probability graphs remain (roughly) constant and $\alpha < 2$, the original dataset follows an infinite-variance distribution. This statement is based on the central limit theorem. Under fairly general conditions, sums of finite variance rvs converge properly normalised to a standardnormal distribution, whereas sums of infinite variance rvs converge to a stable distribution. For large m one expects to see exponential decay for the distribution tail of the sum, whose distributional limit is the normal distribution, whereas the distribution tail of the sum with stable limit decreases like a power tail with index $\alpha \in (0, 2)$. In Figure 1.2 the LD-plots of the on- and off-periods are shown. The upper graph for the on-periods starts for $m = 1$ with a linear slope, which for larger m turns into a concave function. This could be explained by data having regularly varying tail, however with index $\alpha > 2$. The linear function for $m = 1$ is based on the property of power-tailed distributions, namely

$$\lim_{x\to\infty} \frac{d \log(\overline{F}(x))}{d \log x} = -\alpha \,.$$

For larger m the aggregation takes over showing the normal limit distribution. The lower graph for the off-periods exhibits even for $m = 1$ a light-tailed distribution. Hence we may conclude that both, the on- and off-periods have finite variance, the on-periods are obviously lighter tailed than the off-periods.

In order to reach at a clearer picture we furthermore apply the *mean excess function* which is another excellent explorative method to discriminate distribution tails. For a non-negative rv X the mean excess function is defined as

$$e(u) = E(X - u | X > u)\,, \quad u > 0\,.$$

This function is constant for exponential X, it tends to 0 for light-tailed dfs and to ∞ for heavy-tailed dfs. The Pareto distribution is characterised by a linearly increasing mean excess function. More precisely, if $P(X > x) = (1 + \frac{x}{\alpha\beta})^{-\alpha}$ for $\alpha > 1$, $\beta \in \mathrm{R}$, then

$$e(u) = \frac{\alpha\beta + u}{\alpha - 1}\,.$$

Figure 1.3 shows the empirical mean excess function of the on- and off-periods of the above traffic. A Pareto tail (or regularly varying tail) seems to fit well to the on-periods, whereas the off-periods have apparently a lighter tail than the on-periods.

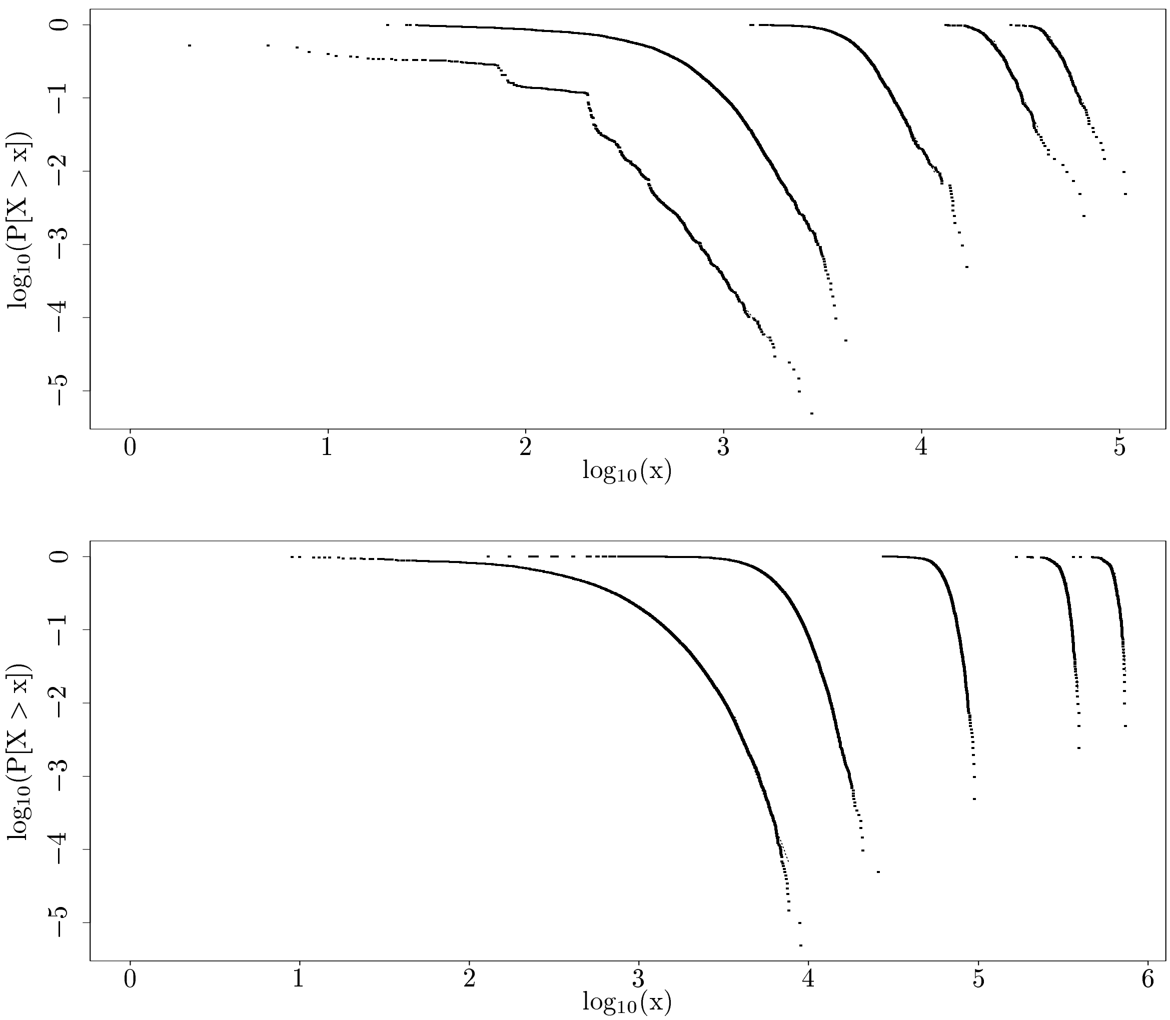

Abbildung 1.2: LD plots of the lengths of the on-periods (upper graph) and off-periods (lower graph) on log-scale to the basis 10 for $m = 1, 10, 100, 500,$ and 1000.

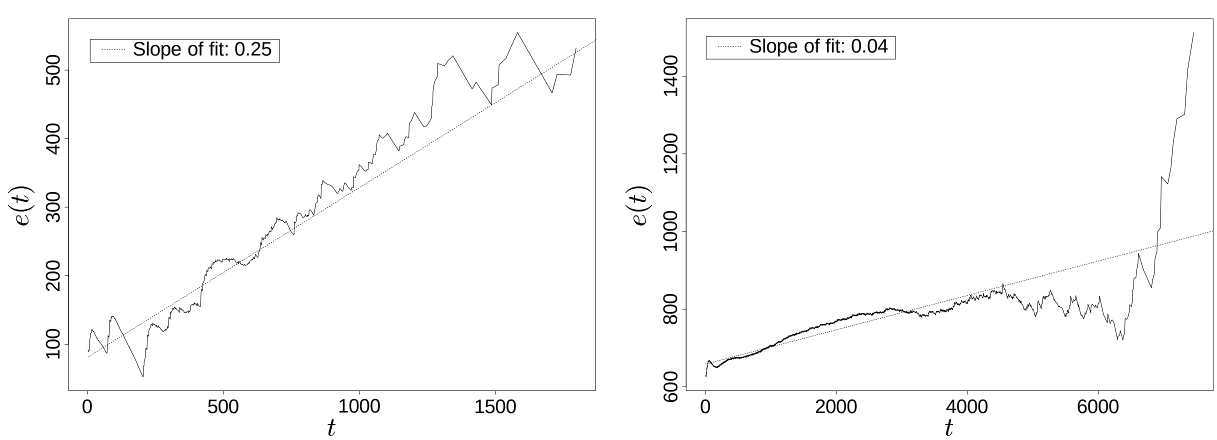

Abbildung 1.3: Mean excess function of the on-periods (lhs) and off-periods (rhs).

This indicates that an on/off model as presented in Section 1.4.1 may be appropriate for our data. Notice that for the moment we ignore the clusters around the peaks and assume that the lengths of on-periods can be modeled by iid rvs. We investigate the corresponding fluid process $(A_t)_{t\geq 0}$, which is defined as 1 during an on-period and 0 during an off-period.

Finally, we estimate the autocorrelation function of this process for our data. The result is shown in Figure 1.4, where we plot the estimated autocorrelation function from lag 5 to lag 7 000. We only start from lag 5 to make the function visible at larger lags (of course the estimated autocorrelation at lag 0 is equal to 1). Comparing the estimated autocorrelation function to the 95% confidence bounds, we clearly see a long-range dependence effect.

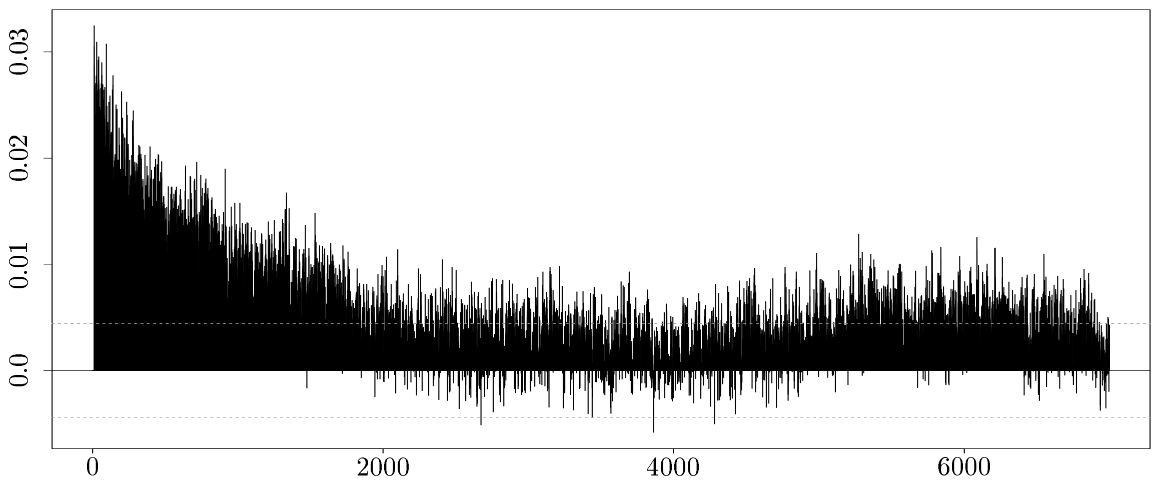

Abbildung 1.4: The empirical autocorrelation function of $(A_t)_{t\geq 0}$ from lag 5 up to lag 7 000.

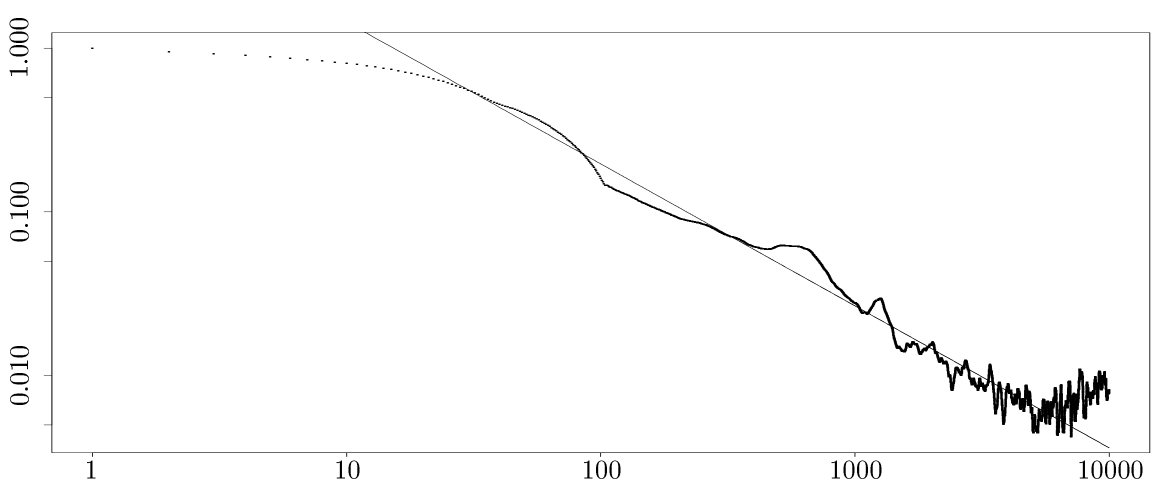

Abbildung 1.5: The log-log plot of the empirical autocovariance function of (A_t). Both axes have been log-transformed to the basis 10.

In Figure 1.5 we plot the estimated autocorrelation function on log-scale (to the basis 10). Now we estimate the slope of the curve by linear regression for the lags in the interval

(20, 7 000). The slope is estimated by −0.82 which – accordingly to Theorem 1.4.2 – means $\alpha = 1.82$, a value fairly lower than in the previous tests. Hence we estimate

$$\widehat{\mathrm{corr}(A_0, A_t)} = c\, t^{-0.82} ,$$

for some positive constant c. Notice that for this model we have estimated an autocorrelation function, which decays hyperbolically and which is not integrable over t.

We should like to say that this first very crude model fitting is by no means the end of the story, on the contrary, we only consider it as a promising start. The data exhibit many structural features which have to be explored in much greater detail. We only mention here the clusters around the peaks in Figure 1.1. Whether these and other data from the measurements of [Gogl, 1998] exhibit the much discussed and important feature of todays network traffic, namely *self-similarity*, needs more careful investigations and will be done in future work. But remarkable enough, it is widely agreed upon that on/off models are very appropriate as source models for self-similar network traffic: the superposition of many strictly alternating independent and identically distributed on/off sources with heavy-tailed distributions with infinite variance results in self-similar aggregate traffic ([Taqqu et al., 1997]). It is shown by the contributions [Lipsky & Schwefel, 1999a] and [Lipsky & Schwefel, 1999b] in this volume that – besides having this theoretical soundness – on/off models are additionally computationally very tractable and versatile.

1.6 References

[Asmussen, 1996] Asmussen, S. (1996). Rare events in the presence of heavy tails. In *Stochastic Networks, Stability and Rare Events*, pages 197–214. Lecture Notes in Statistics, Springer, New York.

[Asmussen & Klüppelberg, 1997] Asmussen, S. and Klüppelberg, C. (1997). Stationary M/G/1 excursions in the presence of heavy tails. *J. Appl. Probab.*, 34:208–212.

[Asmussen et al., 1998] Asmussen, S., Klüppelberg, C., and Sigman, K. (1998). Tail behaviour of queue length via independent sampling. Preprint 1998. Center for Mathematical Sciences, Munich University of Technology, D-80290 Munich, Germany. http://www-m4.mathematik.tu-muenchen.de/m4/Papers/papers.html.

[Baccelli et al., 1999] Baccelli, F., Schlegel, S., and Schmidt, V. (1999). Asymptotics of Tandem Queues with Subexponential Service Times. In Greiner, M. and Jobmann, M., editors, *Stochastic Modeling of High-Speed Networks*. CS Press, München.

[Beran et al., 1995] Beran, J., Sherman, R., Taqqu, M., and Willinger, W. (1995). Long-range dependence in variable-bit-rate video traffic. *IEEE Transactions on Communications*, 43.

[Boxma, 1996] Boxma, O. J. (1996). Fluid queues and regular variation. *Performance Evaluation*, 29:699–712.

[Cohen, 1973] Cohen, J. W. (1973). Some results on regular variation for distributions in queueing and fluctuation theory. *J. Appl. Probab.*, 10:343–353.

[Crovella et al., 1998] Crovella, M. E., Taqqu, M. S., and Bestavros, A. (1998). Heavy-Tailed Probability Distributions in the World Wide Web. In Adler, R. J., Feldman, R. E., and Taqqu, M. S., editors, *A Practical Guide to Heavy Tails: Statistical Techniques and Applications*, pages 3–25, Boston. Birkhäuser.

[Embrechts et al., 1997] Embrechts, P., Klüppelberg, C., and Mikosch, T. (1997). *Modelling Extremal Events for Insurance and Finance.* Springer, Berlin.

[Embrechts & Veraverbeke, 1982] Embrechts, P. and Veraverbeke, N. (1982). Estimates for the probability of ruin with special emphasis on the possibility of large claims. *Insurance: Mathematics and Economics*, 1:55–72.

[Feller, 1971] Feller, W. (1971). *An Introduction to Probability Theory and its Applications, Vol. II.* Wiley, New York.

[Garrett & Willinger, 1994] Garrett, M. W. and Willinger, W. (1994). Analysis, Modeling and Generation of Self-Similar VBR Video Traffic. In *ACM SigComm'94*. ACM, London.

[Gogl, 1998] Gogl, H. (1998). Data acquisition and data analysis in ATM based networks. Technical report, Institut für Informatik, Technische Universiät München, Germany.
http://www8.informatik.tu-muenchen.de/forschung/leistung/ATM_Project/atm_index.html.

[Gogl et al., 1999] Gogl, H., Greiner, M., and Jobmann, M. (1999). Measurement and Statistical Analysis of Traffic Streams in ATM-based Networks. In Greiner, M. and Jobmann, M., editors, *Stochastic Modeling of High-Speed Networks.* CS Press, München.

[Goldie & Klüppelberg, 1998] Goldie, C. M. and Klüppelberg, C. (1998). Subexponential Distributions. In Adler, R. J., Feldman, R. E., and Taqqu, M. S., editors, *A Practical Guide to Heavy Tails: Statistical Techniques and Applications*, pages 435–460, Boston. Birkhäuser.

[Greiner et al., 1999] Greiner, M., Jobmann, M., and Küppelberg, C. (1999). Telecommunication Traffic, Queueing Models, and Subexponential Distributions. Invited paper, to appear in *Queueing Systems (QUESTA)*.

[Heath et al., 1996a] Heath, D., Resnick, S., and Samorodnitsky, G. (1996a). Heavy Tails and Long Range Dependence in On/Off Processes and Associated Fluid Models. Technical Report #1144, School of Operations Research and Industrial Engineering, Cornell University.

[Heath et al., 1996b] Heath, D., Resnick, S., and Samorodnitsky, G. (1996b). Patterns of Buffer Overflow in a Class of Queues with Long Memory in the Input Stream. To appear in The Annals of Applied Probability.

[Jelenković & Lazar, 1996] Jelenković, P. R. and Lazar, A. A. (1996). A network multiplexer with multiple time scale and subexponential arrivals. In *Stochastic Networks, Stability and Rare Events*, pages 215–235. Lecture Notes in Statistics, Springer, New York.

[Jelenković & Lazar, 1997] Jelenković, P. R. and Lazar, A. A. (1997). Asymptotic Results for Multiplexing Subexponential On-Off Sources. Submitted for publication to Advances in Applied Probability.

[Jelenković & Lazar, 1998] Jelenković, P. R. and Lazar, A. A. (1998). Subexponential asymptotics of a markov-modulated random walk with queueing applications. *J. Appl. Probab.*, 35. To appear.

[Kurtz, 1996] Kurtz, T. (1996). Limit theorems for workload input models. In Kelly, F. P., Zachary, S., and Ziedins, I., editors, *Stochastic Networks: Theory and Applications*. Royal Statistical Society Lecture Note Series 4.

[Leland et al., 1994] Leland, W. E., Taqqu, M. S., Willinger, W., and Wilson, D. V. (1994). On the self-similiar nature of ethernet traffic (extended version). *IEEE/ACM Transactions on Networking*, 2(1):1–15.

[Lipsky & Schwefel, 1999a] Lipsky, L. and Schwefel, H.-P. (1999a). N-Burst Processes as Analytic Models of Self-Similar Traffic in Telecommunications Systems – I. Introduction and LAQT Background. In Greiner, M. and Jobmann, M., editors, *Stochastic Modeling of High-Speed Networks*. CS Press, München.

[Lipsky & Schwefel, 1999b] Lipsky, L. and Schwefel, H.-P. (1999b). N-Burst Processes as Analytic Models of Self-Similar Traffic in Telecommunications Systems – II. Specific System Models and Performance Results. In Greiner, M. and Jobmann, M., editors, *Stochastic Modeling of High-Speed Networks*. CS Press, München.

[Pakes, 1975] Pakes, A. G. (1975). On the tails of waiting time distributions. *J. Appl. Probab.*, 12:555–564.

[Paxson & Floyd, 1995] Paxson, V. and Floyd, S. (1995). Wide area traffic: The failure of poisson modeling. *IEEE/ACM Transactions on Networking*, 3(3):226–244.

[Resnick, 1992] Resnick, S. I. (1992). *Adventures in Stochastic Processes*. Birkhäuser, Boston.

[Resnick, 1997] Resnick, S. I. (1997). Heavy tail modelling in teletraffic data. *Ann. Statist.*, 25:1805–1848. With discussion.

[Smith, 1972] Smith, W. L. (1972). On the tails of queueing time distributions. Technical report, Mimeo Series No. 830, Dept. of Statistics, University of North Carolina, Chapel Hill.

[Taqqu & Levy, 1986] Taqqu, M. S. and Levy, J. (1986). Using renewal processes to generate long-range dependence and high variability. In Eberlein, E. and Taqqu, M. S., editors, *Dependence in Probability and Statistics*, pages 73–89. Birkhäuser, Boston.

[Taqqu et al., 1997] Taqqu, M. S., Willinger, W., and Sherman, R. (1997). Proof of a Fundamental Result in Self-Similar Traffic Modeling. *Computer Communication Review*, 27:5–23.

[Willinger et al., 1997] Willinger, W., Taqqu, M. S., Sherman, R., and Wilson, D. V. (1997). Self-Similarity Through High-Variability: Statistical Analysis of Ethernet LAN Traffic at the Source Level. *IEEE/ACM Transactions on Networking*, 5(1):71–86.