FITTING MIXTURES OF EXPONENTIALS TO LONG-TAIL DISTRIBUTIONS TO ANALYZE NETWORK PERFORMANCE MODELS[1]

by

Anja Feldmann[2] and Ward Whitt[3]

AT&T Laboratory – Research

December 16, 1996

Abstract

Traffic measurements from communication networks have shown that many quantities characterizing network performance have long-tail probability distributions, i.e., with tails that decay more slowly than exponentially. File lengths, call holding times, scene lengths in MPEG video streams, and intervals between connection requests in Internet traffic all have been found to have long-tail distributions, being well described by distributions such as the Pareto and Weibull. It is known that long-tail distributions can have a dramatic effect upon performance, e.g., long-tail service-time distributions cause long-tail waiting-time distributions in queues, but it is often difficult to describe this effect in detail, because performance models with component long-tail distributions tend to be difficult to analyze. We address this problem by developing an algorithm for approximating a long-tail distribution by a hyperexponential distribution (a finite mixture of exponentials). We first prove that, in principle, it is possible to approximate distributions from a large class, including the Pareto and Weibull distributions, arbitrarily closely by hyperexponential distributions. Then we develop a specific fitting algorithm. Our fitting algorithm is recursive over time scales, starting with the largest time scale. At each stage, an exponential component is fit in the largest remaining time scale and then the fitted exponential component is subtracted from the distribution. Even though a mixture of exponentials has an exponential tail, it can match a long-tail distribution in the regions of primary interest when there are enough exponential components. When a good fit is achieved, the approximating hyperexponential distribution inherits many of the difficulties of the original long-tail distribution; e.g., it is still difficult to obtain reliable estimates from simulation experiments. However, some difficulties are avoided; e.g., it is possible to solve some queueing models that could not be solved before. We give examples showing that the fitting procedure is effective, both for directly matching a long-tail distribution and for predicting the performance in a queueing model with a long-tail service-time distribution.

---

[1] An abbreviated version of this paper will be presented at *IEEE INFOCOM'97*, Kobe, Japan, April 1997

[2] Room 2C-312, AT&T Laboratories, Murray Hill, NJ 07974-0636 email: anja@research.att.com
www: "http://www.research.att.com/~anja"

[3] Room 2C-178, AT&T Laboratories, Murray Hill, NJ 07974-0636, email: wow@research.att.com

1. Introduction

A major challenge for engineering the emerging high-speed integrated-services communication networks is to develop models that can realistically capture the performance effects of the complex traffic that will be offered to and carried by these networks. Evidence of traffic complexity appears in many forms, such as in the long-range dependence and self-similarity found in the statistical analysis of traffic measurements (e.g., Leland et al. [35]). There is also strong evidence of important phenomena at several different time scales (e.g., Montgomery and de Veciana [41]).

The complexity revealed by these traffic measurements have led some to suggest that this traffic cannot be analyzed by available traffic models. However, we contend that available traffic models can represent remarkably complex behavior. Most comparisons between traffic models and traffic data have been made with rather weak strawmen, such as the simple Poisson process or the batch Poisson process. A good example of a more powerful traffic model is the *Markovian arrival process* (MAP) or its extension, the *batch Markovian arrival process* (BMAP), also known as the virtual Markovian point process, see Lucantoni [36], Chapter 5 of Neuts [43], and Andersen et al. [4]. The potential power of a MAP is dramatically demonstrated by a theoretical result due to Asmussen and Koole [7]. They proved that any stationary point process can be approximated arbitrarily closely by a MAP. (The meaning of "close" is defined in Section 2.)

This is not to say that there are no difficulties. It is challenging to analyze models with elaborate MAPs and BMAPs constructed to capture complex traffic behavior, but new effective computational schemes are being developed, e.g., [15], [37]. This is also not to say that new models should not be sought and examined. However, the main theme of this paper is that there is more that we can do with the tools at hand than might be expected.

In this paper we focus on one phenomenon that seems to underlie much of the observed traffic complexity: long-tail probability distributions. Let F be a *cumulative distribution function* (cdf) and let the associated *complementary cdf* (ccdf) be $F^c(t) = 1 - F(t)$. We say that a cdf F (or its associated ccdf F^c) has a *long tail* (also known as fat tail or heavy tail) if the ccdf F^c decays more slowly than exponentially, i.e., if

$$e^{\gamma t} F^c(t) \to \infty \quad \text{as} \quad t \to \infty \quad \text{for all } \gamma > 0 \ . \tag{1.1}$$

In contrast, we say that cdf F has a *short tail* if its ccdf F^c decays exponentially, i.e., if there exists some $\gamma > 0$ such that

$$e^{\gamma t} F^c(t) \to 0 \quad \text{as } t \to \infty \ . \tag{1.2}$$

Neither (1.1) nor (1.2) describes the actual decay rates of the ccdf's well; they are intended for general classification. A typical long-tail cdf might have a *power tail*, i.e.,

$$F^c(t) \sim \alpha t^{-\beta} \quad \text{as} \quad t \to \infty \ , \tag{1.3}$$

where α and β are positive constants and $f(t) \sim g(t)$ as $t \to \infty$ means that $f(t)/g(t) \to 1$ as $t \to \infty$, whereas a typical short-tail cdf might have *bounded support* ($F^c(t) = 0$ for some t) or an *exponential tail*, i.e.,

$$F^c(t) \sim \alpha e^{-\eta t} \quad \text{as} \quad t \to \infty \tag{1.4}$$

for positive constants α and η.

Two familiar long-tail distributions are the Pareto distribution and the Weibull distribution. One form of the *Pareto distribution*, which we refer to as Pareto(a,b), has ccdf

$$F^c(t) = (1 + bt)^{-a} \tag{1.5}$$

for positive parameters a and b; see p. 233 of Johnson and Kotz [31]. One form of the *Weibull distribution*, which we refer to as Weibull(c, a), has ccdf

$$F^c(t) = e^{-(t/a)^c} \tag{1.6}$$

for positive parameters a and c; see Chapter 20 of [31]. From (1.5) it is easy to see that the Pareto ccdf in (1.5) has a power tail and so always has a long tail. The Weibull ccdf in (1.6) has a long tail (but not a power tail) according to definition (1.1) if $c < 1$, and we will only consider that case.

There has been a long history traffic measurements, but the identification of long-tail distributions has been a major theme in recent years; see Pawlita [44]. Marshall and Morgan [38] note that the empirical distributions of local-area network traffic have longer tails than an exponential distribution. Meier-Hellstern et al. [39] observed high variability in their interarrival times of packets that seems best described with long-tail distributions. The analysis of a large dataset of local area Internet IP traffic collected at Bellcore showed that traffic is highly variable over several time scales. Measurements of source on and off times (high and low activity times) of individual network sources within the Bellcore dataset have indicated long-tail distributions (Leland et al. [35]), and Willinger, et al. [59] have proved that such long-tailed on and off times for individual sources can explain the self-similarity in the aggregate traffic.

Paxson [45, 46] and Paxson and Floyd [47] find that long-tail distributions yield statistically better models for the tail behavior of durations, number of bytes, and burst bytes of ftp connections on the Internet. Feldmann [22, 23] has shown that the intervals between connection requests in Internet traffic have long-tail distributions. Cáceres et. al [14] present further evidence of long-tail distributions in Internet traffic. Recent analysis by Crovella and Bestavros [18] of the durations of world wide web transfers have led to scrutinizing the file length distribution on file servers. Both distributions have been found to be long-tailed. Mogul's [40] investigation of a very busy world-wide-web server indicates that interarrival times of accesses have long tails. Jelenković et al. [30] find that the lengths of scenes in MPEG video streams have a long-tail distribution. Izquierdo and Reeves [27] show that the number of cells in VBR encoded video sequences has a long-tail distribution. Even telephone call holding-time distributions have been found to be long-tailed; e.g., see Bolotin [12] and Duffy et al. [21].

The accumulated evidence is clear: many important probability distributions associated with network traffic have long tails. Moreover, it is known that long-tail distributions can have a dramatic impact upon network performance. For example, in 1973 Cohen [17] showed that the steady-state waiting-time distribution in a single-server queue with unlimited waiting space inherits the long-tail property of a service-time distribution with a power tail. For more recent work in this direction, see [1, 6, 16, 19, 20, 25, 29]. However, the impact of a long-tail distribution depends on the context and requires careful analysis. For example, in the single-server queue, large delays are caused by large service times and short interarrival times, e.g., see [55, 1, 16]. In some distributions, long tails imply that small values are more likely too, but exceptionally long interarrival times by themselves typically do not cause large delays.

Not only are long-tail distributions prevalent and important, but they are difficult to analyze. For example, even the relatively simple $M/G/1$ queue is difficult to analyze when the service-time distribution is Pareto. Abate et al. [1] calculate performance measures for the $GI/G/1$ queue when the general interarrival-time and service-time distributions are long-tailed using numerical transform inversion [2], but it is necessary to have the Laplace transforms of these distributions, and there evidently is no convenient expression for the Laplace transforms of the Pareto and Weibull distributions.

Our main contribution in this paper is to point out that it is possible to approximate long-tail probability distributions by convenient short-tail probability distributions, so that available performance models can be effectively analyzed and so that the effect of the long-tail distribution upon performance can be determined. (We do *not* claim that the long-tail distribution has no effect.) Moreover, we develop a remarkably simple algorithm for constructing suitable approximating distributions for a large class of long-tail distributions. The class of long-tail distributions that can be approximated by the method developed here includes the Pareto and Weibull distributions in (1.5) and (1.6) as special cases.

Although at first it may be surprising that long-tail distributions can be approximated by short-tail distributions, there is a simple explanation in the notion of time scale. In almost all network performance settings, the distribution of interest only matters through its values in some finite interval $[t_1, t_2]$. For t_1 sufficiently small and t_2 sufficiently large, the precise form of the distribution outside the interval $[t_1, t_2]$ should not matter. (Because of the nature of time scales, it is usually appropriate to measure time logarithmically. Thus, we might have $t_1 = 10^{-a}$ and $t_2 = 10^b$ for appropriate constants a and b.) The main point is that, in principle, it should be possible to approximate any long-tail distribution by a short-tail distribution. A simple way to do this is to truncate the distribution at the points t_1 and t_2 and assign the negligible probabilities of the intervals $[0, t_1)$ and (t_2, ∞) to the points t_1 and t_2, respectively. Although this produces a short-tail distribution that captures the essential behavior of the original long-tail distribution, it may not be a convenient approximation.

Here we consider hyperexponential distributions as approximating distributions. A *hyperexponential* (H_k) distribution is a mixture of k exponentials for some k, i.e., the ccdf has the form

$$H^c(t) = \sum_{i=1}^{k} p_i e^{-\lambda_i t} \tag{1.7}$$

where $p_i \geq 0$ for all i and $p_1 + \ldots + p_k = 1$. Our fitting algorithm fits a hyperexponential distribution to a given long-tail distribution, aiming to be accurate over a finite interval $[t_1, t_2]$ for suitably small t_1 and suitably large t_2.

Given data that might be well described by either a Pareto distribution or a hyperexponential distribution, we would usually prefer the Pareto distribution for a simple description because it provides a more parsimonious description. The H_k distribution in (1.7) has $2k - 1$ parameters, whereas the Pareto distribution has only 2. Statistical estimation also tends to work better when there are fewer parameters.

We primarily suggest replacing long-tail distributions such as the Pareto distribution by hyperexponential distributions, because performance models tend to be easier to analyze when component distributions in the model are hyperexponential. One reason is that hyperexponential distributions are special phase-type distributions, which have been found to make performance models more tractable; see Neuts [42]. Another reason that we might choose hyperexponential distributions is because they have simple Laplace transforms. the Laplace transform of the density h of the ccdf H^c in (1.7) and the Laplace-Stietjes Transform of the cdf H is

$$\hat{h}(s) = \int_0^\infty e^{-st} h(t) dt = \int_0^\infty e^{-st} dH(t) = \sum_{i=1}^{k} \frac{p_i \lambda_i}{\lambda_i + s}. \tag{1.8}$$

The explicit Laplace transform (1.8) makes it possible to analyze many performance models by numerical transform inversion, e.g., see [1, 2, 15, 16, 36, 37]. For these numerical transform inversion algorithms, having a relatively large number of phases (e.g., 10 or 100) presents no serious difficulty. We will illustrate this advantage by considering the $M/G/1$ queue with a long-tail service-time distribution. We have no difficulty calculating the steady-state waiting-time distribution in the $M/G/1$ queue by numerical transform inversion after making the hyperexponential approximation.

We also show that hyperexponential distributions make it easier to obtain Markov stochastic processes, which tend to be far easier to analyze than non-Markov stochastic processes. In particular, in

Section 8 we show that hyperexponential approximations can help analyze superpositions of independent on-off sources, where each source sends input at a constant rate (fluid) or as a Poisson process when it is on. If the on or off periods have long-tail distributions, then the aggregate input model tends to be intractable, but if the on and off periods of each source have hyperexponential distributions, then the aggregate input becomes a Markov-modulated fluid or Poisson process, for which there are effective algorithms. Unfortunately, however, this representation is not totally satisfactory, because the Markovian state space becomes larger when the number of exponential components in a mixture increases. Hence, if there are many sources, the state space of the approximating aggregate input model may be so large that analysis remains difficult. Nevertheless, the approximation is a step towards tractable models. If there are only a few source, then the model can now be solved, whereas it could not be solved before.

Once a hyperexponential fit is contemplated, there are many ways to proceed, such as a least squares fit using a mathematical program. A natural alternative is the expectation-maximization (EM) algorithm, which is an iterative procedure that minimizes the Kullback-Leibler "distance"; see Asmussen, Nerman and Olsson [8], Turin [51] and references therein. A difficulty with the EM algorithm is that the iteration can be slow when there are many parameters. The EM algorithm can be enhanced significantly if a good starting point can be provided. In preliminary experiments we have found that our algorithm is also useful to quickly provide a good starting point for the EM algorithm, but we do not discuss those experiments here.

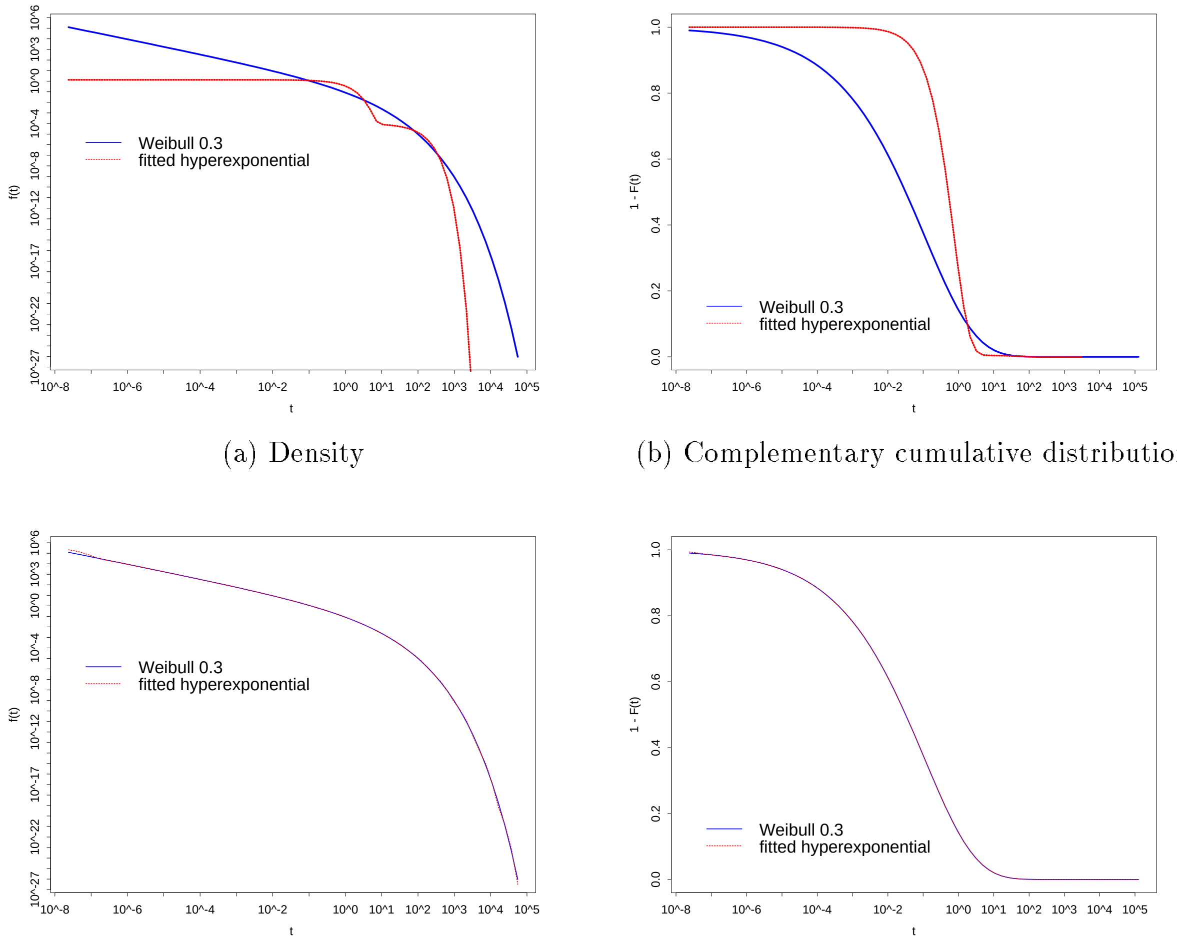
We intend to compare various fitting schemes in a future paper. In this paper we present a simple recursive scheme, based on the notion of time scales. We recursively fit starting in the largest time scale that matters and successively reduce the time scale. We start by fitting a weighted exponential $p_1 e^{-\lambda_1 t}$ to the tail of the given ccdf. Since we focus on the tail, λ_1^{-1} should be suitably large. Then we subtract this weighted exponential from the original ccdf and fit a second weighted exponential $p_2 e^{-\lambda_2 t}$ to the new tail where $\lambda_2^{-1} < \lambda_1^{-1}$. Since the exponential ccdf's are short tailed, it should be possible to choose the second exponential component so that it is negligible further out in the region where the first exponential $p_1 e^{-\lambda_1 t}$ was fit. We describe the algorithm in more detail and discuss previous related work in Section 4. To illustrate right away, we consider an example.

1.1. Example

Suppose that we consider a Weibull distribution as in (1.6) with exponent $c = 0.3$ and a chosen so that the distribution has mean 1. (That makes $a = 9.26053$.) Since c is close to 0, this Weibull distribution is strongly long-tailed. This is partly reflected by its next two moments, which are $m_2 = 29.2$ and $m_3 = 4481$. However, the first three moments do not nearly capture the full long-tail effect. To illustrate, we first consider fitting an H_2 distribution (a mixture of two exponentials, which has three parameters) to the Weibull distribution by matching the first three moments. A three-moment matching algorithm for the H_2 fit is given on p. 136 of Whitt [54]. the resulting H_2 parameters are $p_1 = 0.00501$, $\lambda_1 = 0.019$, and $\lambda_2 = 1.355$. The approximating H_2 density and ccdf are compared to their Weibull counterparts in Figure 1 (a), (b). It is obvious that the fit is quite poor, even though the H_2 distribution has the same first three moments.

In contrast, the density and ccdf of an H_k fit obtained by our algorithm in Section 4 is shown in (c) and (d) of Figure 1. The fit is so good that it is hard to see two curves in (c) and (d). This H_k fit has $k = 20$ exponentials. The three moments of the approximating H_{20} distribution are $m_1 = 1.0060$, $m_2 = 30.6$, and $m_3 = 4640$. The parameters of H_{20} are given in Table 1.

By this example, we do not mean to imply that 20 exponentials are necessarily required to produce a satisfactory approximation of this Weibull distribution, but this number certainly seems to be sufficient for almost all network performance applications.



(a) Density

(b) Complementary cumulative distribution

(c) Density

(d) Complementary cumulative distribution

Figure 1: A comparison between the Weibull(0.3, 9.261) density and ccdf with hyperexponential approximations. This example shows the difference in the quality of fit between matching three moments (a), (b) and applying our algorithm (c), (d).

An attractive feature of our algorithm is that it does not depend on the moments. Therefore, it can be used even if the moments do not exist or are not known. However, it is useful to calculate the first few moments of the original and the approximating distributions to help judge the quality of the fit.

1.2. Organization of the paper

Here is how the rest of this paper is organized. In Section 2 we discuss robustness of performance models. We refer to some of the evidence indicating that if a component probability distribution in a performance model is well approximated by another, then the performance measures of interest will be suitably close. We also give a precise meaning for "close." In Section 3 we rigorously prove that it is possible to approximate many long-tail distributions by hyperexponential distributions. We identify a

	Parameters of the algorithm fit		
i	p_i	λ_i	$1/\lambda_i$
1	0.013457	23678496	4.2232e-08
2	0.007274	3103042	3.2226e-07
3	0.011161	728019	1.3736e-06
4	0.017063	170717	5.8576e-06
5	0.025935	40004.9	2.4997e-05
6	0.039055	9366.24	0.000107
7	0.057927	2190.72	0.000456
8	0.083791	511.995	0.001953
9	0.116197	119.685	0.008355
10	0.149927	28.064	0.035632
11	0.170885	6.643	0.15052
12	0.157748	1.607	0.62230
13	0.103040	0.405	2.47202
14	0.039333	0.108	9.23375
15	0.006820	0.031	31.8780
16	0.000384	0.0099	101.303
17	4.34e-06	0.0033	300.270
18	4.52e-09	0.0012	850.935
19	1.22e-13	0.00042	2361.97
20	1.13e-20	0.00015	6517.72

Table 1: Parameters of the approximating H_{20} cdf of Example 1.1.

class of distributions containing many long-tail distributions, including Pareto and Weibull, for which arbitrarily close hyperexponential approximations can be made.

We present our recursive algorithm for constructing approximating hyperexponential distributions in Section 4. Some readers might wish to skip the more theoretical Sections 2 and 3 and go directly to the algorithm. In Section 5 we explain when the algorithm should be effective. Then we present several examples in Section 6.

In Section 7 we investigate how our fitting algorithm is related to fitting probability distributions to data. We show through simulation experiments that, consistent with intuition, it is usually much better to fit a long-tail distribution with only a few parameters to the data and then afterwards apply our algorithm to the long-tail distribution in order to obtain a high-order hyperexponential approximation than it is to apply our algorithm directly to the empirical distribution generated from the data.

In Section 8 we show how hyperexponential distributions can help analyze the aggregate input from the superposition of on-off sources. Finally, we state our conclusions in Section 9.

2. The Robustness of Performance Models

Since we intend to approximate component distributions in performance models by other distributions, it is important that the performance models be robust to such changes. As a specific example, we will consider approximating long-tail service-time distributions by hyperexponential distributions in the $GI/G/1$ queue. (The $GI/G/1$ queue is just one example; There are many possible applications of hyperexponential approximations besides the $GI/G/1$ queue). The $GI/G/1$ queue has a single server, unlimited waiting room and interarrival times and service times coming from independent sequences of independent and identically distributed random variables with general distributions. If we approximate the given general interarrival-time and service-time distributions by other distributions, then we want descriptive performance measures such as the steady-state waiting-time distribution also to be approximately what it would be with the original interarrival-time and service-time distributions. Fortunately, such robustness, stability or continuity properties have been established for performance models, e.g., see Section VIII.5 of Asmussen [5], Section 21 of Borovkov [13], Kalashnikov and Rachev [32] and Whitt [52], [53].

Even though robustness results have been established, care is needed because the robustness results do not hold unconditionally. The robustness depends upon what we mean by "close" and upon regularity conditions. For probability distributions on the real line (or, more generally, on a metric space) it is customary to use the notion of weak convergence, as in Billingsley [11]. In that framework, we say that a sequence of probability measures $\{P_n : n \geq 1\}$ converges to a probability measure P, and write $P_n \Rightarrow P$, if

$$\int f dP_n \to \int f dP \quad \text{as} \quad n \to \infty \tag{2.1}$$

for all bounded continuous real-valued functions f. On the real line the probability measures P_n and P are characterized by cdf's F_n and F, e.g., $F(t) = P((-\infty, t])$. Then convergence of probability measures $P_n \Rightarrow P$ as $n \to \infty$ is equivalent to convergence of cdf's in the form

$$F_n(t) \to F(t) \quad \text{as} \quad n \to \infty \tag{2.2}$$

for all points t that are continuity points of the limiting cdf F, which we denote by $F_n \Rightarrow F$. For continuous cdf's, a metric associated with this convergence is the uniform metric

$$m(F_1, F_2) = \sup_t |F_1(t) - F_2(t)| \ . \tag{2.3}$$

(For further discussion, see the introduction to [11].) For random variables (or more general random elements) X_n and X distributed as P_n and P, respectively, we say that X_n converges in distribution to X and write $X_n \Rightarrow X$ as $n \to \infty$ if $P_n \Rightarrow P$ as $n \to \infty$.

With this background, we can state a robustness theorem for the $GI/G/1$ queue due to Borovkov [13] p. 118. A random variable is said to be proper if it is finite with probability one.

Theorem 2.1. (Borovkov). *Consider a sequence of $GI/G/1$ queueing models indexed by n with interarrival times, service times and steady-state waiting-time distributed as $U^{(n)}$, $V^{(n)}$ and $W^{(n)}$, respectively. Consider a prospective limiting $GI/G/1$ model with corresponding random variables U, V and W. If $EV^{(n)} < EU^{(n)}$ for all n, $EV < EU$, $U_n \Rightarrow U$, $V_n \Rightarrow V$ and $EV_n \to EV$ as $n \to \infty$, then $W^{(n)}, n \geq 1$, and W are proper random variables and $W_n \Rightarrow W$ as $n \to \infty$.*

The condition $EV^{(n)} < EU^{(n)}$ in Theorem 2.1 is needed in order to ensure that the n^{th} model is stable, i.e., that a proper steady-state waiting-time $W^{(n)}$ exists. An important point in Theorem 2.1 is that we also need to assume that the limiting system is stable ($EV < EU$), that the mean service times converge ($EV_n \to EV$), and that the limiting mean is necessarily finite ($EV < \infty$ since $EV < EU$). We need to assume that $EV_n \to EV$ as $n \to \infty$, because convergence in distribution does not imply convergence of moments. As a secondary point, note that there is no requirement that the mean interarrival times $EU^{(n)}$ and EU be finite or that $EU^{(n)} \to EU$ as $n \to \infty$.

If we also want convergence of moments, i.e., $E(W^{(n)k}) \to E(W^k) < \infty$ as $n \to \infty$, then we need to assume corresponding convergence and finiteness of one higher service-time moment; i.e., it is necessary and sufficient to have $EV^{(n)(k+1)} \to EV^{k+1} < \infty$ as well as the other conditions of Theorem 2.1. This can be deduced from Theorem 2.2 on p. 185 of Asmussen [5] and its proof.

To illustrate how we can apply Theorem 2.1, suppose that a $GI/G/1$ queueing system of interest has a generic service time V with a Pareto distribution as in (1.5). In the next section we will show that, without imposing any moment conditions, we can approximate the Pareto distribution of V arbitrarily closely by a hyperexponential distribution as in (1.7); i.e., for each n we can let $V^{(n)}$ have a hyperexponential distribution (where the number of component exponentials depends on n) and have $V^{(n)} \Rightarrow V$ as $n \to \infty$.

We would like to deduce that $W^{(n)} \Rightarrow W$ for the waiting-times in the associated $GI/G/1$ models. (Assume that the interarrival-time distribution is fixed.) However, we cannot draw this conclusion without the extra conditions in Theorem 2.1. The crucial extra condition is that $EV < \infty$; for the $GI/G/1$ application we must require that the Pareto distribution have a finite mean. If $EV = \infty$, then the approximation procedure will fail, but if $EV < \infty$, then it will work. It turns out that we can choose the approximating distributions so that $EV^{(n)} \to EV$ as $n \to \infty$, and we need to do so, but we also need to require that $EV < \infty$ and $EV < EU$ as well. However, with such extra conditions, approximating component distributions can achieve the desired result. The remaining questions are only the practical ones: How many exponentials are needed before the distribution of $V^{(n)}$ is suitably close to the distribution of V? And how do we actually find a good approximating distribution?

2.1. Example

To illustrate the robustness of the queueing model, we consider the Weibull distribution in Example 1.1 as a service-time distribution in the $M/G/1$ queue (having an exponential interarrival-time distribution). We let the arrival rate (and thus the traffic intensity) be 0.75. We focus on the steady-state waiting-time ccdf $P(W > t)$. In addition to the three-moment H_2 fit and the H_{20} fit by our algorithm in Section 4, we consider a simple exponential fit obtained by matching only the mean.

We compare numerical results (calc) for the $M/H_2/1$ and $M/M/1$ models to simulations (exp) of the $M/H_2/1$ and $M/W/1$ models in (a) and (b) of Figure 2 (W stands for Weibull). In contrast, we compare numerical results for the $M/H_{20}/1$ and $M/M/1$ models to simulations of the $M/H_{20}/1$ and $M/W/1$ models in (c) and (d) of Figure 2. In all cases, the steady-state waiting-time ccdf is displayed, with the y-axis being in log scale in (b) and (d).

The $M/M/1$ model is appealing, because the steady-state waiting-time cdf for it is available in closed form (a simple exponential plus an atom at the origin), but it yields a remarkably poor approximation. Clearly the service-time distribution beyond its mean matters greatly. The $M/H_2/1$ numerical results could be obtained in several ways; we used numerical transform inversion [2]. The simulations were based on a time interval of 5.3×10^6, which corresponds to about 4×10^6 arrivals.

From (a) and (b) of Figure 2, we see that the $M/H_2/1$ approximation for the waiting-time ccdf is much better than the H_2 approximation for the W service-time distribution directly. This reflects the extensive experience showing that approximations based on two moments of the interarrival-time and service-time distributions can be quite effective [57]. However, even though the $M/H_2/1$ approximation might be good enough for some engineering applications, the $M/H_{20}/1$ approximation in (c) and (d) is far better. This is perhaps more evident from Table 2, which displays the ccdf values for the W, H_2 and H_{20} cases. The relative errors are substantial for small and very large values.

3. Complete Monotonicity

To have a good theoretical basis for approximating one distribution by another, it is appropriate to consider what is possible. From this perspective, it is important to note that every hyperexponential distribution has a decreasing probability density functions (pdf) and possibly an atom at 0. Thus, hyperexponential distributions cannot capture departures from this structure, such as atoms away from 0 or a non-monotone pdf.

On the other hand, there is a large class of distributions (necessarily with monotone pdf's) which

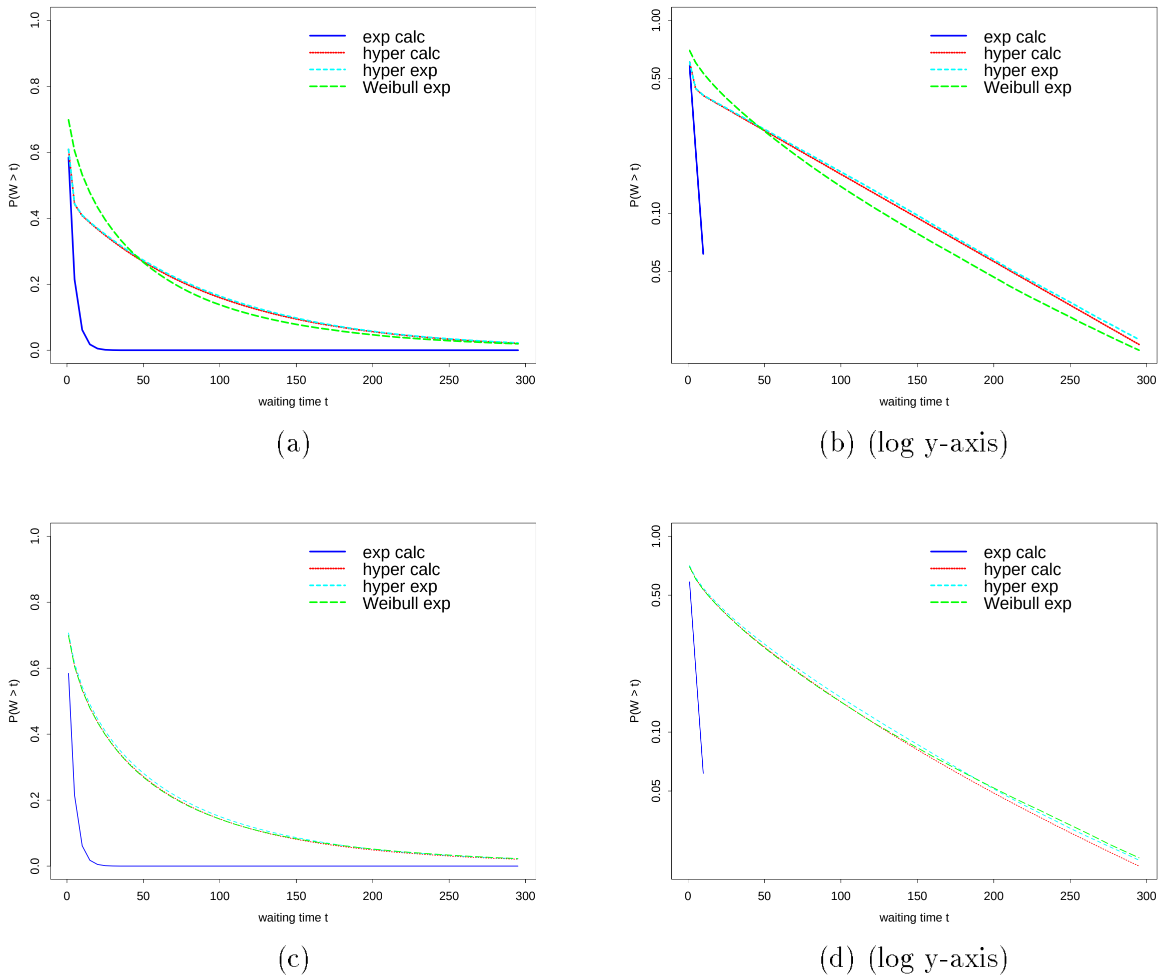


Figure 2: The steady-state $M/G/1$ waiting-time ccdf $P(W > t)$ with a Weibull(.3, 9.261) service-time distribution having mean = 1. the numerical results (calc) are for the model with the approximating hyperexponential and exponential service-time distributions. The simulations (exp) are for the model with the Weibull and the approximating hyperexponential distribution. Parts (a) and (b) contain the H_2 fit by matching the first three moments, while parts (c) and (d) contain the H_{20} fit by the algorithm in Section 4. Parts (b) and (d) are the same as parts (a) and (c), respectively, but with the y-axis in log scale.

can be approximately arbitrarily closely by hyperexponentials. The nice class of probability distributions are those with completely monotone pdf's. A probability density function (pdf) f is said to be *completely monotone* if all derivatives of f exist and

$$(-1)^n f^{(n)}(t) \geq 0 \quad \text{for all} \quad t > 0 \quad \text{and} \quad n \geq 1 \; ; \tag{3.1}$$

see p. 66 of Keilson [33] and p. 439 of Feller [24]. the link between completely monotone pdf's and mixtures of exponential pdf's is provided by Bernstein's [10] theorem. (see [24]).

Waiting-time	Simulated Weibull	Algorithm fit Calc. Hyper. Exp.	Moment fit Calc. Hyper. Exp.
1	0.6983	0.6991	0.6087
5	0.6046	0.6063	0.4429
10	0.5323	0.5347	0.4079
15	0.4777	0.4808	0.3866
20	0.4332	0.4368	0.3669
25	0.3957	0.3995	0.3483
30	0.3634	0.3673	0.3306
35	0.3350	0.3391	0.3138
40	0.3096	0.3141	0.2979
45	0.2869	0.2917	0.2828
70	0.2017	0.2073	0.2179
125	0.1031	0.1068	0.1229
135	0.0924	0.0956	0.1107
145	0.0828	0.0857	0.0998
155	0.0745	0.0770	0.0899
165	0.0670	0.0693	0.0810
185	0.0545	0.0566	0.0658
205	0.0444	0.0465	0.0534
225	0.0364	0.0385	0.0434
235	0.0333	0.0351	0.0391
245	0.0303	0.0320	0.0352

Table 2: A comparison of numerical results for the steady-state waiting-time ccdf $P(W > t)$ in the $M/H_2/1$ and $M/H_{20}/1$ models with simulation results for the $M/W/1$ model of Example 2.1.

Theorem 3.1. (Bernstein) *Every completely monotone pdf f is a mixture of exponential pdf's, i.e.,*

$$f(t) = \int_0^\infty \lambda e^{-\lambda t} dG(\lambda) , \quad t \geq 0 , \tag{3.2}$$

for some proper cdf G.

We call G in (3.2) the *spectral* cdf. (Then the support of G is called the *spectrum.* The *support* of G is the set of all t for which $G(t+\epsilon) - G(t-\epsilon) > 0$ for all $\epsilon > 0$.) Of course, the spectral cdf G appearing in (3.2) is a general cdf; it need not have finite support. (A cdf G has finite support if it has a probability mass function attaching probabilities p_i to n points t_i with $p_1 + \ldots + p_n = 1$ for some n.) However, cdf's with finite support are dense in the family of all cdf's (using the standard mode of convergence in (2.1) and (2.2)). Hence, Theorem 3.1 implies the following result.

Theorem 3.2. *If F is a cdf with a completely monotone pdf, then there are hyperexponential cdf's $F^{(n)}$, $n \geq 1$, i.e., cdf's of the form*

$$F^{(n)}(t) = \sum_{i=1}^{k_n} p_{ni}(1 - e^{-\lambda_{ni} t}) , \quad t \geq 0 , \tag{3.3}$$

with $\lambda_{ni} \leq \infty$ and $p_{n1} + \ldots p_{nk_n} = 1$, such that $F^{(n)} \Rightarrow F$ as $n \to \infty$.

Theorems 3.1 and 3.2 are important for approximating long-tail distributions because many long-tail pdf's are completely monotone. For example, by differentiating (and using mathematical induction), it is easy to see that the pdf's of the Pareto distribution in (1.5) and the Weibull distribution with $a < 1$ in (1.6) are completely monotone. For the Pareto distribution, Harris [26] directly showed that the spectral cdf is gamma. (This is an easy calculation; see [26] or [31].)

The gamma pdf with shape parameter less than 1 is also completely monotone. The Pareto mixture of exponentials (PME) distribution considered in [1] is also completely monotone, because it directly satisfies (3.2). The PME distribution is convenient because its Laplace transform is available. (See Section 6.3 below.) Other methods for constructing long-tail distributions with convenient Laplace transforms are described in [3].

In order to approximate a completely monotone cdf F having spectral cdf G by a hyperexponential distribution (a finite mixture of exponentials), it suffices to approximate the spectral cdf G by a spectral cdf $G^{(n)}$ with finite support. One concrete way is to choose $n+1$ points t_i with $0 = t_0 < t_1 < \ldots < t_n = \infty$ and let $p_{ni} = G(t_i) - G(t_{i-1})$ and $\lambda_{ni} = (t_i + t_{i-1})/2$, $1 \leq i \leq n$. This makes $\lambda_{nn} = \infty$, so that $F^{(n)}$ has an atom of size p_{nn} at 0. By letting the successive sets $T_n \equiv \{t_0, \ldots, t_n\}$ become dense in the finite interval $[0, t]$ for every t, we achieve the desired result as $n \to \infty$. To have the successive approximations be refinements of the previous ones, we can let the subsets T_n be nested, i.e., we can also have $T_n \subseteq T_{n+1}$ for all n.

We might also want the means of $F^{(n)}$ to be nondecreasing. We can achieve that property by changing the definition of λ_{ni} to $\lambda_{ni} = t_{ni}$. However, this choice tends to produce worse approximations. Given the spectral cdf G, more elaborate fitting procedures are also possible. The essential idea is to choose a cdf $G^{(n)}$ with finite support approximating G.

It is sometimes convenient to represent a completely monotone pdf in a different way, in particular, as

$$f(t) = \int_0^\infty \lambda^{-1} e^{-t/\lambda} dH(\lambda) \ , \quad t \geq 0 \ , \tag{3.4}$$

instead of as in (3.2). We call the H in (3.4) the *mixing* cdf. If the spectral cdf G and the mixing cdf H in (3.2) and (3.4) have pdf's g and h, then they are related by

$$h(t) = t^{-2} g(t^{-1}) \ , \quad t \geq 0 \ . \tag{3.5}$$

The mixing representation (3.4) is convenient for working with moments. If $m_k(H)$ and $m_k(F)$ are the k^{th} moments of H and F, respectively, then from (3.4) it follows that

$$m_k(F) = m_k(H) k! \ , \quad k \geq 1 \ . \tag{3.6}$$

Hence, if we choose $H^{(n)}$ to be a cdf with finite support approximating H, where $H^{(n)}$ has the same first k moments as H, then the associated approximating hyperexponential distribution $F^{(n)}$ with mixing distribution $H^{(n)}$ will have the same first k moments as the cdf F with mixing cdf H (defined by (3.4).) With this structure, it is possible to identify certain extremal (bounding) hyperexponential pdf's among all completely monotone pdf's with given first k moments; e.g., queueing applications are discussed in [55].

Paralleling Theorem 3.2, it is possible to show that *any* cdf on the non-negative real line can be approximated arbitrarily closely in the sense of Section 2 by a phase-type cdf (which includes the hyperexponential distribution as special case), as in Chapter 2 of Neuts [42]. The EM algorithm is a way to fit phase type distributions [8].

4. The Recursive Fitting Procedure

In this section we specify the recursive procedure for fitting a hyperexponential (H_k) cdf H to a given cdf F on the nonnegative real line. We think of the original cdf as being a long-tail distribution such as Pareto or Weibull with exponent less than one. We think of the cdf F as having a monotone probability density function (pdf) f, but we do not require it. We discuss conditions under which the procedure should be effective in Section 5.

The H_k distribution has ccdf (1.7) and associated pdf

$$h(t) = \sum_{i=1}^{k} p_i \lambda_i e^{-\lambda_i t} , \quad t \geq 0 , \tag{4.1}$$

where $\sum_{i=1}^{k} p_i = 1$, $\lambda_i > 0$ and $p_i > 0$ for all i. Clearly the H_k pdf is monotone.

Without loss of generality, let the exponential parameters λ_i in (4.1) be labeled so that $\lambda_1 < \ldots < \lambda_k$. Then the higher indexed components have tails which decay more rapidly. Our idea is to fit the H_k components recursively, starting with the pair (λ_1, p_1) and then proceeding to (λ_2, p_2) and so forth. If λ_2 is sufficiently greater than λ_1, then $\sum_{i=2}^{k} e^{-\lambda_i t}$ should be negligible compared to $p_1 e^{-\lambda_1 t}$ for t sufficiently large (in the tail). This should enable us to choose the pair (p_1, λ_1) without being concerned about the other H_k parameter values. We then subtract the component $p_1 e^{-\lambda_1 t}$ from both $H^c(t)$ and $F^c(t)$ and fit the second component to the remaining tail. If again λ_3 is sufficiently greater than λ_2, then $\sum_{i=3}^{k} e^{-\lambda_i t}$ should be negligible compared to $p_2 e^{-\lambda_2 t}$ for t sufficiently large, and we can fit the pair (λ_2, p_2) without being concerned about the other H_k parameters.

After deriving this recursive fitting procedure, we learned that the general recursive estimation procedure actually has a long history, being know as Prony's [48] method; see p. 114 of Turin [50]. In that context, we contribute by showing when the recursive fitting procedure should be effective (Sections 3 and 5 here) and by applying it to approximate long-tail distributions.

Here is the procedure: we first choose the number k of exponential components and k arguments where we will match quantiles: $0 < c_k < c_{k-1} < \ldots < c_1$. We assume that the ratios c_i/c_{i+1} are sufficiently large; e.g., we could have $c_i = c_1 10^{-(i-1)}$ for $2 \leq i \leq k$. Let b be such that $1 < b < c_i/c_{i+1}$ for all i; e.g., with $c_i = c_1 10^{-(i-1)}$ we could have $b = 2$.

We choose λ_1 and p_1 to match the ccdf $F^c(t)$ at the arguments c_1 and bc_1; i.e., we solve the two equations

$$p_1 e^{-\lambda_1 c_1} = F^c(c_1) \tag{4.2}$$

and

$$p_1 e^{-\lambda_1 b c_1} = F^c(bc_1) \tag{4.3}$$

for p_1 and λ_1, assuming that $c_1, b, F^c(c_1)$ and $F^c(bc_1)$ are known, obtaining

$$\lambda_1 = \frac{1}{(b-1)c_1} \ln(F^c(c_1)/F^c(bc_1)) \tag{4.4}$$

and

$$p_1 = F^c(c_1) e^{\lambda_1 c_1} . \tag{4.5}$$

With this procedure, we are assuming that λ_i will be sufficiently larger than λ_1 for all $i \geq 2$ that the final approximation will satisfy

$$\sum_{i=1}^{k} p_i e^{-\lambda_i t} \approx p_1 e^{-\lambda_1 t} \quad \text{for} \quad t \geq c_1 .$$

We have no guarantee that this property will hold, but the accuracy can be checked when the fit is complete. (See Section 5 for further discussion and Section 6 for examples.)

Next, for $2 \leq i \leq k$, let

$$F_i^c(c_i) = F_{i-1}^c(c_i) - \sum_{j=1}^{i-1} p_j e^{-\lambda_j c_i} \tag{4.6}$$

and

$$F_i^c(bc_i) = F_{i-1}^c(bc_i) - \sum_{j=1}^{i-1} p_j e^{-\lambda_j bc_i}, \tag{4.7}$$

where $F_1^c(t) = F^c(t)$. Then proceed as above, letting

$$p_i e^{-\lambda_i c_i} = F_i^c(c_i) \tag{4.8}$$

and

$$p_i e^{-\lambda_i bc_i} = F_i^c(bc_i)\ , \tag{4.9}$$

to obtain

$$\lambda_i = \frac{1}{(b-1)c_i} \ln(F_i^c(c_i)/F_i^c(bc_i)) \tag{4.10}$$

and

$$p_i = F_i^c(c_i)e^{\lambda_i c_i} \tag{4.11}$$

for $2 \le i \le k-1$. Finally, for the last parameter pair (λ_k, p_k), we require that

$$p_k = 1 - \sum_{j=1}^{k-1} p_j \tag{4.12}$$

and

$$p_k e^{-\lambda_k c_k} = F_k^c(c_k)\ , \tag{4.13}$$

where $F_k^c(c_k)$ is defined in (4.6), so that

$$\lambda_k = \frac{1}{c_k} \ln(p_k/F_k^c(c_k))\ . \tag{4.14}$$

Assuming that we obtain probability weights ($p_i > 0$ for all i), and that the parameters λ_i are well separated, we should obtain a good fit. Assuming that we obtain probability weights, the procedure produces an H_k ccdf H^c that is larger than the original ccdf F^c at the matching points, i.e.,

$$H^c(c_i) > F^c(c_i)\ , \quad 1 \le i \le k\ , \tag{4.15}$$

and

$$H^c(bc_i) > F^c(bc_i)\ , \quad 1 \le i \le k-1\ . \tag{4.16}$$

However, if F^c is a long-tail distribution, then there will be a t_0 such that

$$F^c(t) \ge H^c(t) \quad \text{for all} \quad t \ge t_0\ . \tag{4.17}$$

Hence, it is important to choose c_1 sufficiently large that t_0 is beyond the region of interest.

Our implementation of the algorithm in software allows the user to proceed interactively, choosing new parameter settings as desired, after looking at tables and graphs of the results. The standard approach is to specify k, c_1, c_k, and b. Then the algorithm chooses the remaining c_i such that the ratio of c_i/c_{i+1} is constant and proceeds with the fitting procedure. An available alternative is to specify one point at a time, start with the pair (c_i, b_i), inspect the preliminary result, and continue by choosing the next pair (c_{i+1}, b_{i+1}).

When we are done, we calculate several moments of the H_k distribution via

$$m_j(H_k) = j! \sum_{i=1}^{k} p_i/\lambda_i^j \tag{4.18}$$

and compare them to the moments of F if they are available. As numerical measures of achieved fitting accuracy, we compute the absolute and relative errors of the ccdf and cdf. For both, the cdf and the ccdf, the absolute error is

$$AE(F,t) = |H^c(t) - F^c(t)| = |H(t) - F(t)|. \tag{4.19}$$

A relative error for both the cdf and ccdf is

$$RE(F,t) = \frac{|H^c(t) - F^c(t)|}{\min\{F(t), F^c(t)\}}. \tag{4.20}$$

We graphically display these errors as functions of t over any requested interval (l, u). We calculate the curves by considering points whose logarithms are evenly spaced over (l, u).

To illustrate, we display the absolute and relative errors of the H_2 and the H_{20} fits to the Weibull distribution in Example 1.1 in Figure 3. It turns out that the H_{20} fit was done with $c_k = 10^{-7}$ and $c_1 = 9 \times 10^4$. Since c_1 is not large, there are somewhat large relative errors for the H_{20} cdf in the region $10^2 - 10^5$. However, the ccdf values in this region are very small, e.g., $F^c(10^2) = 4.254e-4$, $F^c(10^3) = 1.878e-7$, $F^c(10^4) = 3.794e-14$, and $F^c(10^5) = 1.667e-27$.

We also calculate the maximum absolute and relative errors over any desired subinterval (l, u), e.g., for the ccdf

$$AE(F^c, l, u) = \sup_{l \le t \le u} AE(F^c, t) \tag{4.21}$$

and

$$RE(F^c, l, u) = \sup_{l \le t \le u} RE(F^c, t) \tag{4.22}$$

where the supremum is estimated by calculating the maximum over many points whose logarithms are evenly spaced in (l, u).

5. When Should the Procedure Work?

In this section we discuss conditions under which the fitting procedure in Section 4 should be effective. In particular, we point out that the procedure is natural for distributions with *decreasing failure rate* (DFR). to see this, note that the fitting formula for λ_i in (4.10) can be rewritten as

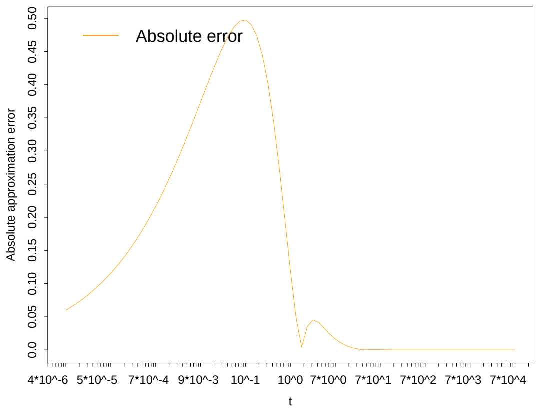
$$\lambda_i = -\frac{\ln(F_i^c(bc_i) - \ln(F_i^c(c_i)}{bc_i - c_i} . \tag{5.1}$$

As $b \to 1$, formula (5.1) approaches

$$\lambda_i = -\frac{d}{dt}\ln(F^c(t)|_{t=c_i} = \frac{f(c_i)}{F^c(c_i)} = r(c_i) , \tag{5.2}$$

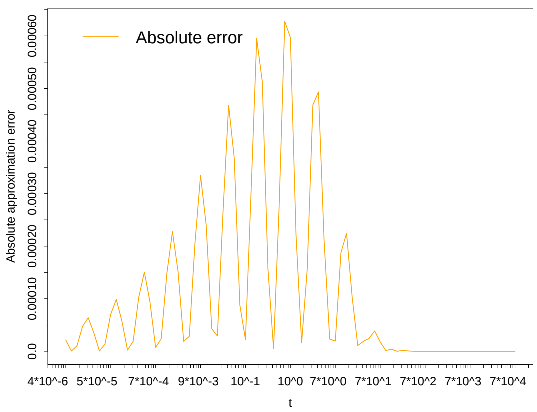
which is the *hazard rate function* (or failure rate function) associated with the ccdf F^c evaluated at c_i; e.g., see Barlow and Proschan [9]. Indeed, we could consider (4.10) replaced with (5.2), but (4.10) seems more robust.

The idea in the procedure of Section 4 is to have λ_i be significantly less than λ_{i+1} for all i. In order to have λ_i be less than λ_{i+1} for all i, it is natural to require that the ccdf $F^c(t)$ be DFR. This is equivalent to having $F^c(t)$ be *log-convex*. A sufficient condition for the ccdf $F^c(t)$ to be log-convex is for the pdf $f(t)$ to be log-convex; see p. 73 of Keilson [33]. Since mixtures of log-convex pdf's are log-convex (Theorem 5.4c on p. 66 of Keilson [33]), all completely monotone pdf's are log-convex. Hence all completely monotone pdf's are DFR.

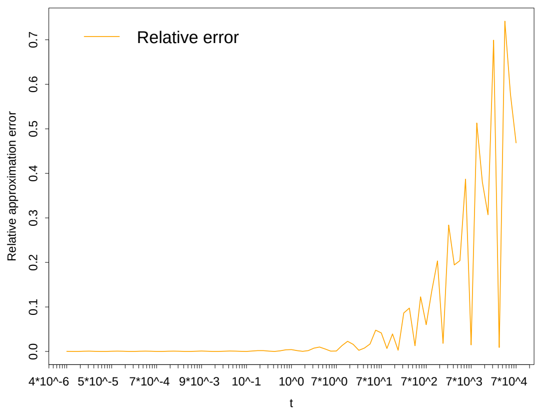


(a) Absolute error

(b) Relative error



(c) Absolute error



(d) Relative error

Figure 3: The relative and absolute errors of the H_2 and the H_{20} fits to the Weibull(0.3, 9.261) distribution from Example 1.1.

In summary, our algorithm is natural for completely monotone pdf's such as the Pareto and Weibull distributions (see Section 3) and, more generally, for DFR pdf's. However, by the same reasoning, our algorithm is inappropriate for increasing failure rate (IFR) distributions. For example, our algorithm does not work for the uniform distribution, i.e., when $F(t) = t/b, 0 \leq t \leq b$, and $F(t) = 1, t \geq b$, which clearly has a very short tail. Since many long-tail distributions are DFR, our algorithm has substantial applicability.

Even though many long-tail distributions are DFR, many others are not. Indeed, the long-tail property (1.1) is unaltered by changing the probability distribution on any initial interval $[0, t]$. Thus, the long-tail property does not nearly guarantee the DFR property.

6. Examples

In this section we give several examples showing how a hyperexponential distribution can be fit to a long-tail distribution with the algorithm described in Section 4. Besides presenting further examples for the Weibull distribution, we give approximations for two Pareto distributions, and a PME distribution [1]. The definitions of the first two distributions are given in the Introduction.

6.1. Weibull Distribution

We start with a Weibull cdf that is only moderately long-tailed, having parameters $c = 0.6$, $a = 0.66464$, and mean $= 1$. Figure 4 (a)–(d) show the results of fitting a hyperexponential distribution with 6 exponentials to this Weibull distribution. The parameters of the fitted H_6 distribution are given in Table 3. In Figure 4 the curves almost coincide for the cdf and the ccdf. In the density one can detect a small deviation around $10 - 100$ and above 250, but overall the fit looks very good. Only the plot of the hazard (density divided by ccdf) reveals that the fit is not precise. But still the fit is reasonably close at least from 10^{-3} to 10^{1}, 4 orders of magnitude.

From Table 4 we see how the algorithm matches the moments of the Weibull distribution. There is a 2% error in the mean, but almost an 30% error in the second moment. If these approximate moments are not deemed close enough, then a new fit can be considered with more exponentials.

Table 5 shows how much each exponential term contributes to each of the first three moments. This information is quite revealing. Although the probability of the fifth exponential term is quite small, namely 0.068, this term contributes substantially to the higher two moments. For any specific distribution, this information allows us to judge if the range of the c values is appropriate or should be expanded or reduced. The probability parameters help to decide if the range of c values cover all desired time scales.

In Figure 4 (e) and (f) we display results for the steady-state waiting time in the $M/G/1$ queue for various service-time distributions. As in Example 2.1, we let the interarrival-time distribution be exponential and the arrival rate (and traffic intensity) be 0.75. We display simulation results for the Weibull and the approximating hyperexponential service-time distributions, and we display numerical results for the same hyperexponential distribution and the exponential service-time distributions (with the same mean). All simulations are based on a time period of 5.3×10^{6}, which corresponds to about 4×10^{6} arrivals. The numerical results for the $M/H_6/1$ model are obtained by numerical transform inversion [2].

Figure 4 (a)–(d) indicate that the hyperexponential distribution with 6 exponentials is a good approximation to the Weibull distribution with $c = 0.6$. Accordingly, it is no surprise that Figure 4 (e), (f) show that the ccdf's of the steady state waiting-time in the $M/G/1$ queue with the Weibull and hyperexponential service-time distributions are very close. Indeed the simulations of both the original Weibull distribution and the fitted hyperexponential distribution basically coincide with the analytical curve for the fitted hyperexponential distribution until the simulation error dominates the waiting-time probability. As in Example 2.1, the exponential approximation is not good.

Next we consider fitting a hyperexponential distribution to the Weibull(.3, 9.261) distribution we considered in the Example 1.1. This Weibull distribution has a much longer tail and spans more orders of magnitude than the previous example. Therefore one might want to consider a larger number of exponentials to obtain a good fit over more time scales (as we did in Example 1.1). Yet we might only be interested in a few time scales. The next two examples show how a fit with a smaller number of exponentials, lets say 4, might satisfy such a need.

Figure 5 gives two examples of such fits. Parts (a), (c), and (e) show the density, the hazard, and the ccdf of the first fit based on $c_k = 0.001$ and $c_1 = 90$, while parts (b), (d), and (f) show the same for the second fit based on $c_k = 1$ and $c_1 = 2000$. Both fits look better than the simple three-moment fit shown in the Example 1.1, but neither is nearly as good as the H_{20} fit there. A comparison between the two fits shows that the first one matches the original distribution better in the range from 10^{-8} to 10^{-1}, while the second one matches the original better in the range from 10^1 to 10^4. This corresponds loosely to the values chosen for c_k and c_1. The result of the different emphasis is that the second hyperexponential distribution matches the moments of the original distribution better than the first one (shown in Table 6), while the first hyperexponential distribution approximates the density, the hazard, and the ccdf more accurately over the plotted range in Figure 5 (a), (c), (e). Therefore, depending on the application, either fit may be preferable. For the analysis of the waiting-times of the $M/G/1$ queue with Weibull distributed service time, the fit from Figure 5 (b), (d), (f) is better suited since the calculations and simulations are sensitive to deviations in the tail. The approximation of the waiting-time ccdf, shown in Figure 6, is reasonable.

Increasing the number of exponentials to 20, leads to a fit that is good for more than 13 orders of magnitude (Figures 1 and 7). Even in the hazard plot of Figure 7 (b), the differences between the fitted hyperexponential distribution and the Weibull distribution are minimal. Indeed, 20 exponentials should be an overkill for almost all applications. For example, for the $M/G/1$ queue, Figure 2 (c), (d) shows that the curves for the waiting-times of the simulation results and the analytical results are very close.

	Parameters of the hyperexp. distribution		
i	p_i	λ_i	$1/\lambda_i$
1	0.029931	676.178	0.001479
2	0.093283	38.709	0.025834
3	0.332195	4.274	0.233977
4	0.476233	0.761	1.313542
5	0.068340	0.248	4.031035
6	0.000018	0.097	10.29943

Table 3: Parameters of the H_6 cdf fit to a Weibull$(.6, 0.665)$ distribution.

Moment	Weibull	hyperexponential
1	1	0.981
2	3.091	3.905
3	24.96	33.48

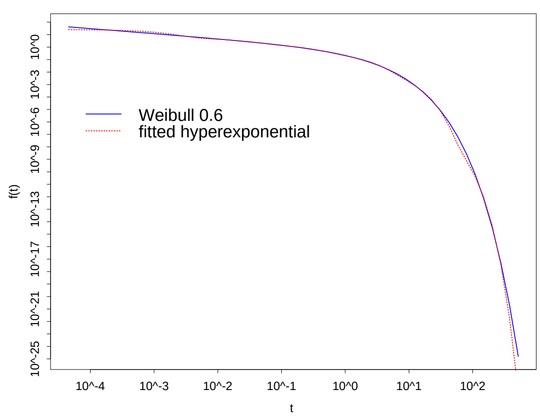
Table 4: Moments of the original Weibull(.6, 0.665) distribution and the H_6 fit.

term	1st moment	2nd moment	3rd moment
1	0.00004	1.3e-07	5.8e-10
2	0.00241	0.00012	9.6-e06
3	0.07773	0.03637	0.02553
4	0.62555	1.64338	6.47593
5	0.27548	2.22096	26.8583
6	0.00018	0.00376	0.11614
sum	0.98140	3.90459	33.4759

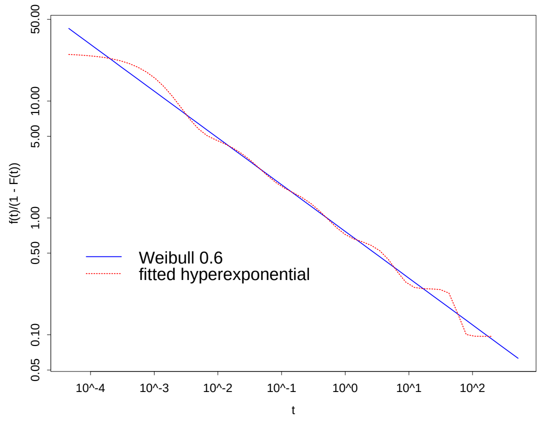
Table 5: Contributions of the individual exponential terms in the H_6 fit to its first three moments.

Moment	Weibull	hyperexponential 1	hyperexponential 2
1	1.00	0.67	1.01
2	29.24	8.80	25.10
3	4480.63	1161.40	3547.00

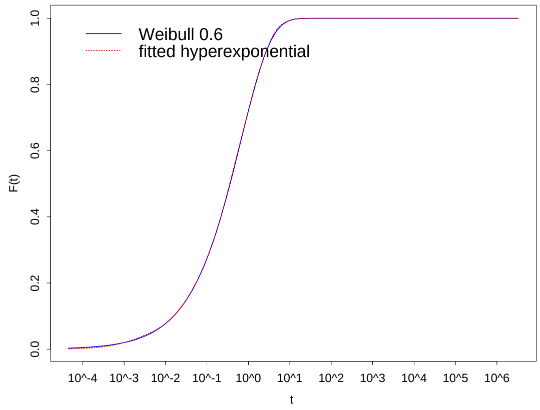
Table 6: Moments of the original Weibull(.3, 9.261) distribution and the two fitted H_4 cdf's.



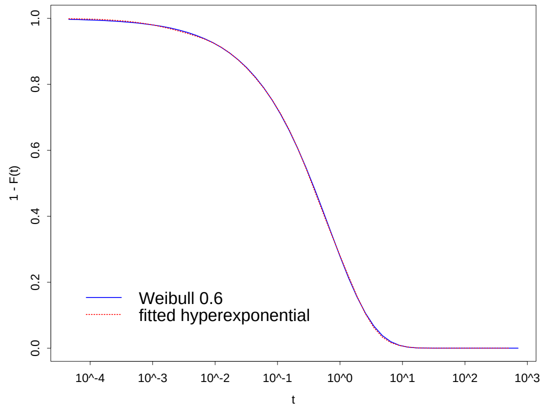
(a) Density



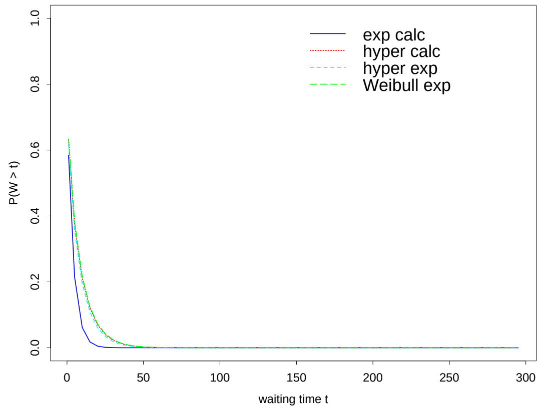
(b) Hazard



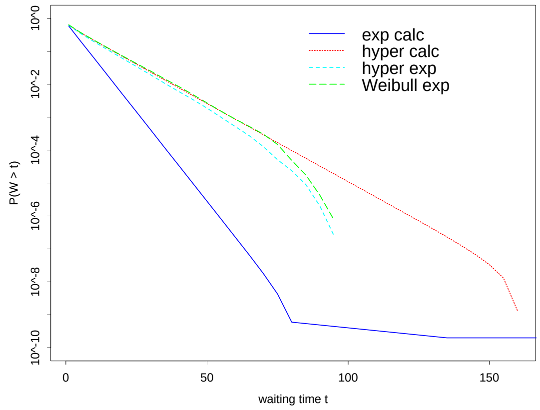
(c) Cumulative distribution



(d) Complementary cumulative distribution



(e) $M/G/1$ queue probabilities

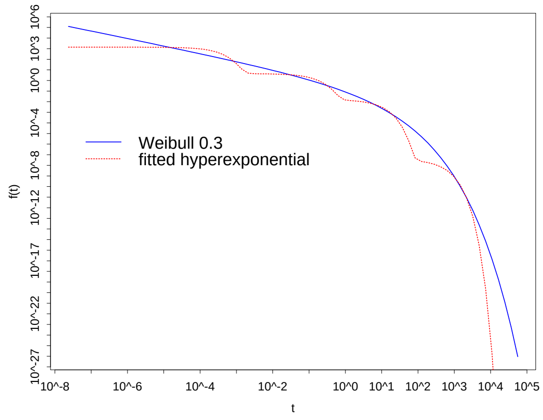


(f) $M/G/1$ queue probabilities (log y-axis)

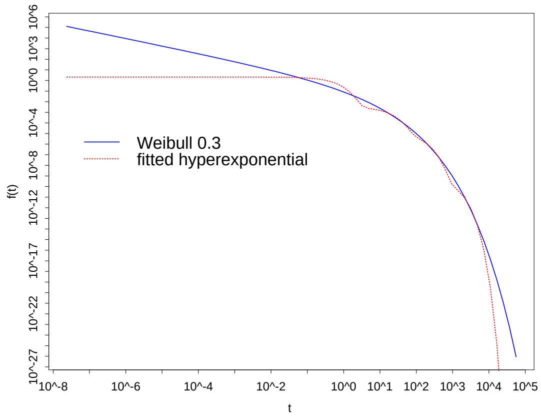
Figure 4: Parts (a), (b), (c), and (d) show H_6 fit to a Weibull(.6, 0.665) distribution. The algorithm used $c_k = 0.001$ and $c_1 = 120$. Parts (e), (f) give a comparison of numerical results and simulations of the steady-state $M/G/1$ waiting-time ccdf for the same Weibull distribution and the H_6 fit by the algorithm in Section 4. Part (f) is the same as part (e), but with the y-axis in log scale.

First fit

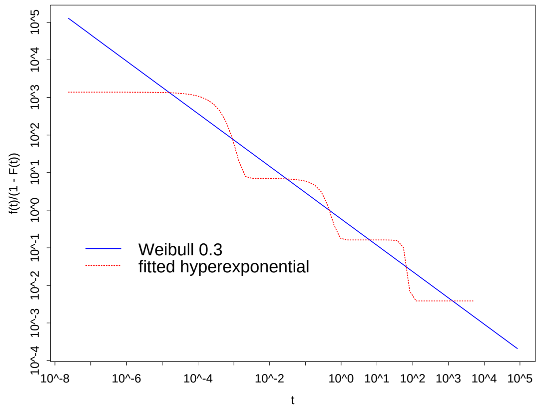
Second fit



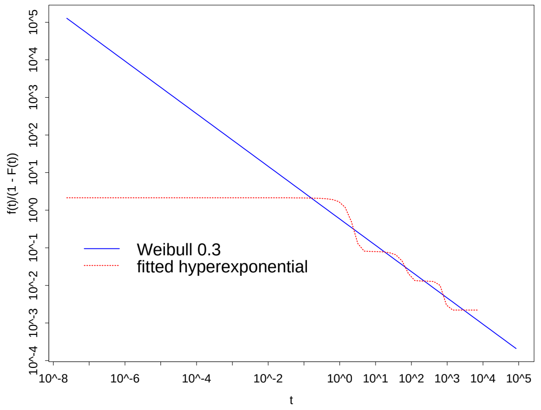
(a) Density



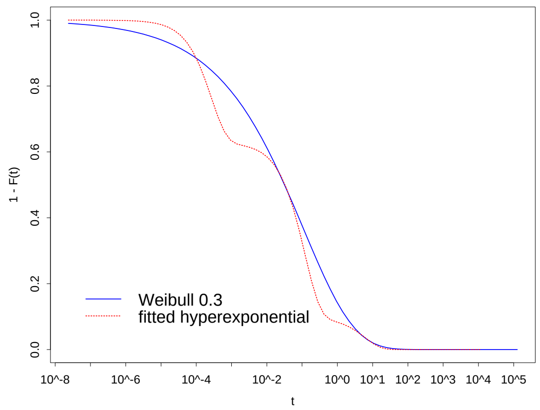
(b) Density



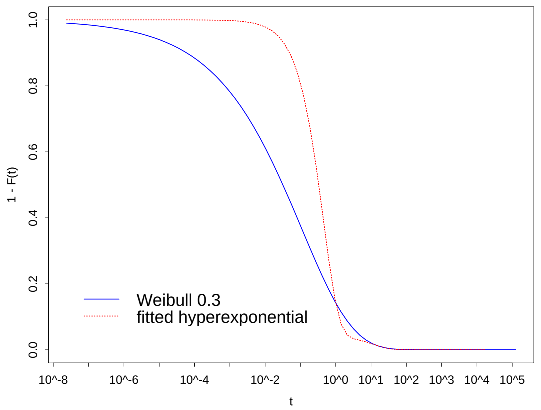
(c) Hazard



(d) Hazard

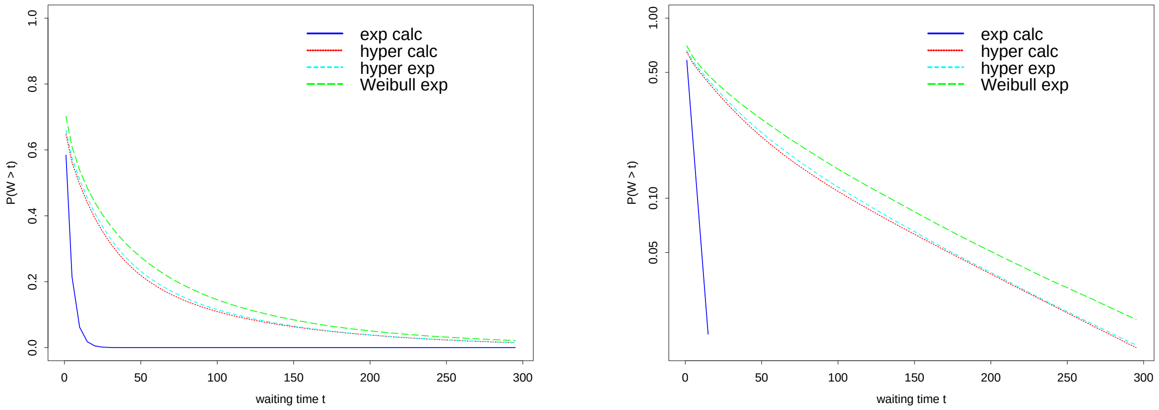


(e) Complementary cumulative distribution



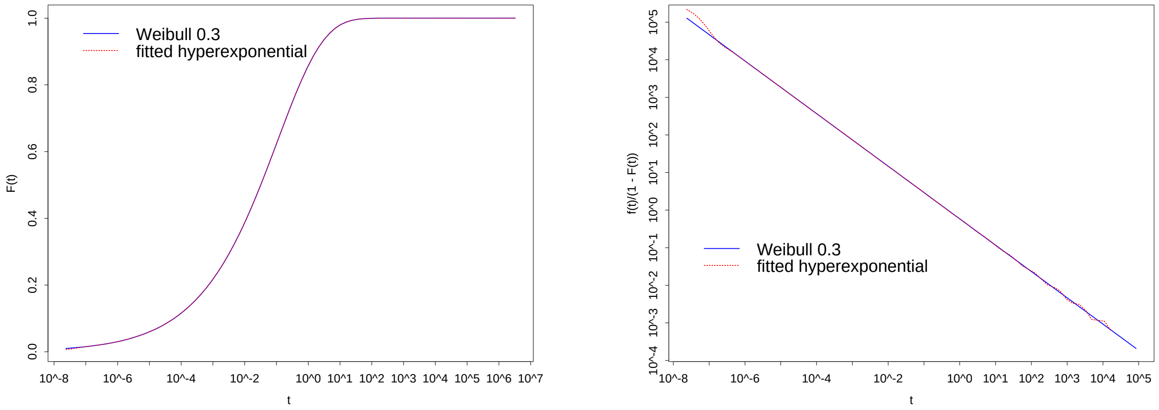
(f) Complementary cumulative distribution

Figure 5: (a), (c), (e) H_4 fit to a Weibull(.3, 9.261) distribution using $c_k = 0.001$ and $c_1 = 90$. (b), (d), (f) H_4 fit to a Weibull(.3, 9.261) distribution using $c_k = 1$ and $c_1 = 2000$.



(a) (b) (log y-axis)

Figure 6: The steady-state $M/G/1$ waiting-time ccdf with a Weibull(.3, 9.261) service-time distribution and the H_4 fit by the algorithm in Section 4 with $c_k = 1$ and $c_1 = 2000$. Part (b) is the same as part (a), but with the y-axis in log scale.



(a) Cumulative distribution (b) Hazard

Figure 7: H_{20} fit to a Weibull(.3, 9.261) distribution using $c_k = 10^{-7}$ and $c_1 = 90,000$.

6.2. Pareto Distribution

The Pareto distribution (defined in (1.5)) is very challenging because it can have infinite moments. Indeed, as its parameter a approaches 0, more and more moments become infinite. In this section we discuss the fitting of two different Pareto distributions: one without third moment, having $a = 2.2$, and one with neither second nor third moment, having $a = 1.2$. Figures 8 and 9 show the results of hyperexponential fits using 13 and 14 exponential terms, respectively. Visually, both fits look very good for the 12 orders of magnitude covered by the plots. Table 7 gives the first three moments of the distributions, while Table 8 gives the hyperexponential parameters. From Table 7, it is apparent that the first three moments of the hyperexponential distribution fit the finite moments of the Pareto distributions reasonable well. The infinite moments are approximated by values in the order of 10^6 to 10^{14}.

Even though we are approximating very long-tail distributions with short-tail distributions, we are not eliminating all problems associated with such long-tail distributions. Instead, the approximation gives us the opportunity to transfer some difficulties from the domain of long-tail distributions to the more familiar domain of hyperexponential distributions.

To illustrate this point, consider the data in Table 9. Table 9 shows how much each exponential term of the hyperexponential approximation contributes to the first three moments. A difficulty in dealing with long-tail distributions is that large values (e.g., long service times) occur with non-negligible probability and therefore contribute substantially to the moments. The same is true for the hyperexponential distributions that approximate the Pareto distributions. For example, the total probability associated with terms 7 through 13 for the Pareto(2.2, 0.83) distribution is only 1.135×10^{-7}, yet the total contribution of these exponentials to the second moment is 3.07 or 26.7% overall. For the Pareto(1.2, 5) distribution, the total probability associated with exponentials 10 to 14 is only 3.76×10^{-7}, but these exponentials are crucial for the approximation, contributing a total of 0.082 or 8.32% to the mean of the distribution and 2.8×10^7 to the second moment. Indeed, these terms largely determine the values of the second and third moments.

So far, the application we have used to demonstrate the goodness of fit in the approximation has been the probability distribution of the waiting-time in the $M/G/1$ queue. Given that we are now considering distributions with large variance, extra care is needed on the experimental (simulation) part of this evaluation. Let $\bar{X}_n$ be the sample mean from a random sample of size n from either the Pareto distribution or the fitted hyperexponential distribution. The sample mean converges to the mean of the distribution as $n \to \infty$ by the law of large numbers, but the variance of the sample mean is proportional to the variance of the distribution (and inversely proportional to the size of the sample). While this is no major issue for the Pareto(2.2, 0.83) distribution and its fitted hyperexponential distribution, this is a concern for the Pareto(1.2, 5) distribution. In this case the Pareto distribution has an infinite variance and the fitted hyperexponential distribution has a very large variance. Hence it is very difficult to obtain a sample mean that is close to the mean of the sampled distribution. At the very least, this implies that the sample size has to be very large. Indeed it may be meaningless to compare the simulation results to the calculated results if the problem is sensitive to the mean, which is the case for service-time distributions in queueing models. As shown in [56], obtaining good simulation estimates of queueing characteristics becomes increasingly difficult as service-time variability increases. Moreover, the approach to steady state gets very slow, so that it may be more appropriate to consider the transient behavior of the queueing system.

Nevertheless, Figures 10 and 11 show both the simulation and the analytical results for the waiting-time probabilities of an $M/G/1$ queue with these two Pareto distributions. For the Pareto(2.2, 0.83) distribution, Figure 10 shows that the simulation results for the Pareto and the fitted hyperexponential distributions are reasonably close to each other. Figure 10 (b) shows the results of 5 independent replications each based on 4×10^6 arrivals for both distributions. The differences of the curves are within the simulation error. Also note that the curve for the analytical evaluation of the hyperexponential distribution is well covered by the simulation results.

Paralleling Figure 10, Figure 11 shows the results of simulations for the Pareto(1.2, 5) distribution. Figure 11 demonstrates the pitfall of simulations with distributions that have large second moments. Even in the normal scale, the 5 different simulation runs span a wide range. The location of each individual curve is highly dependent on the sample mean. Since all sample means were less than 1, it is not surprising that the curve corresponding to the calculated waiting-time probabilities dominates all simulation results.

Since the Pareto(1.2, 5) service-time distribution has infinite variance, the $M/G/1$ busy period has infinite variance, from which it is possible to deduce that the variance of the empirical distribution of the first n waiting times, at any time, multiplied by n, has variance growing faster than n. Hence, in this problem there is long-range dependence. Consistent with this observation, Figure 11 shows that the high variability is reflected in the hyperexponential approximation.

Moment	Pareto $a = 2.2$	hyperexp.	Pareto $a = 1.2$	hyperexp.
1	1	1.006	1	0.986
2	11	11.49	∞	2.8e+06
3	∞	3.7e+07	∞	8.9e+14

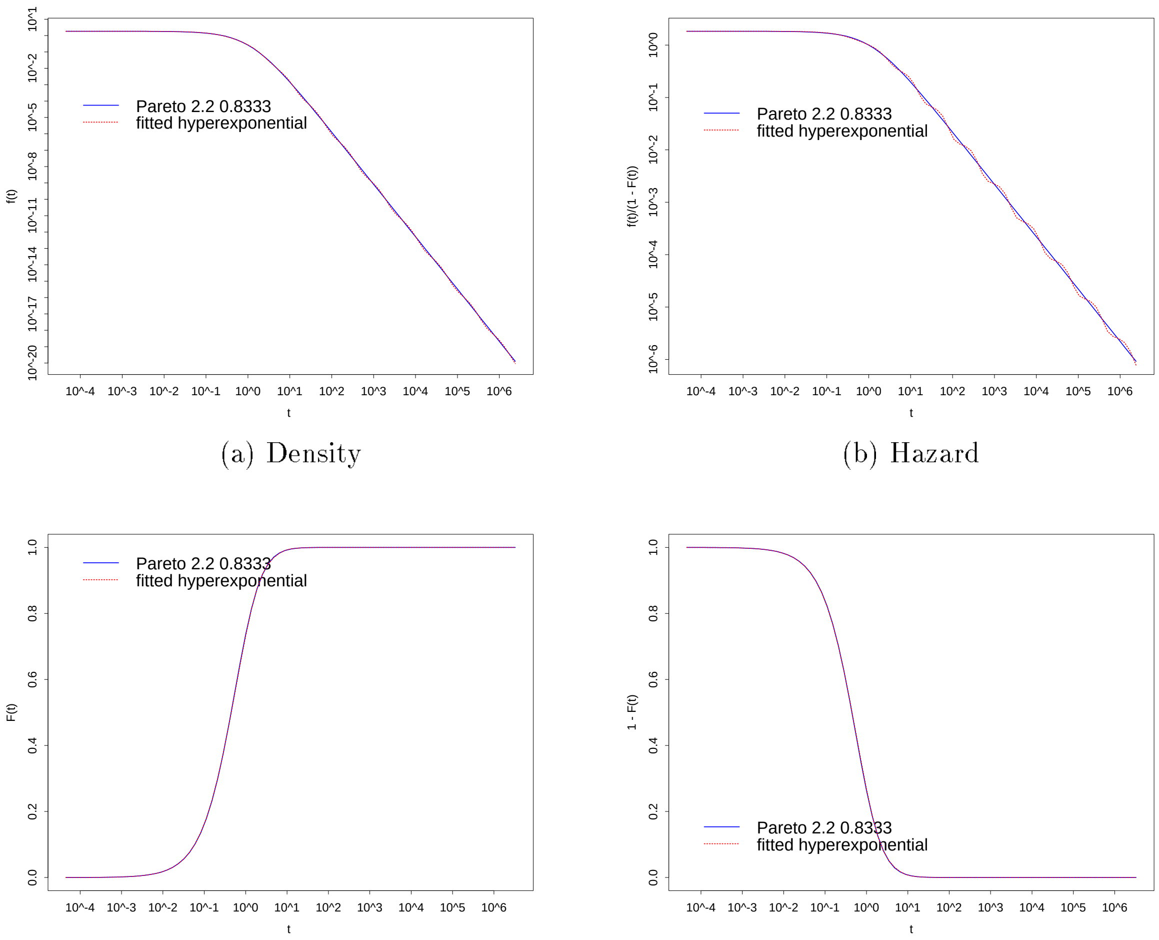
Table 7: Moments of the two original Pareto distributions and the two fitted hyperexponential distributions.

	Parameters of the fit to Pareto $a = 2.2$			Parameters of fit to Pareto $a = 1.2$		
i	p_i	λ_i	$1/\lambda_i$	p_i	λ_i	$1/\lambda_i$
1	0.193963	4.491	0.222677	0.089437	23.304	0.042910
2	0.651199	1.422	0.703442	0.533823	6.516	0.153472
3	0.147814	0.371	2.698616	0.307218	1.546	0.646659
4	0.006832	0.076	13.21435	0.059768	0.306	3.263373
5	0.000188	0.014	70.49069	0.008462	0.057	17.51902
6	4.61e-06	0.003	382.9488	0.001122	0.01	95.28793
7	1.11e-07	0.0005	2087.592	0.000147	0.002	519.5631
8	2.65e-09	8.8e-5	11387.48	1.92e-05	3.5e-4	2834.259
9	6.35e-11	1.6e-5	62126.17	2.50e-06	6.5e-5	15463.11
10	1.52e-12	2.9e-6	339032.8	3.27e-07	1.2e-5	8.44e+04
11	3.63e-14	5.4e-7	1.85e+07	4.27e-08	2.2e-6	4.61e+05
12	8.51e-16	9.7e-8	1.03e+07	5.56e-09	3.9e-7	2.54e+06
13	1.72e-17	1.5e-8	6.56e+08	7.18e-10	6.8e-8	1.47e+07
14				8.37e-11	8.3e-9	1.20e+08

Table 8: Parameters of the fitted hyperexponential distributions for two Pareto distributions.

	Pareto $a = 2.2$ moments			Pareto $a = 1.2$ moments		
# Terms	1st	2nd	3rd	1st	2nd	3rd
1	0.04319	0.01924	0.01285	0.00384	0.00033	4e-05
2	0.45808	0.64447	1.36003	0.08193	0.02515	0.01158
3	0.39889	2.15292	17.4297	0.19867	0.25694	0.49845
4	0.09028	2.38586	94.5830	0.19505	1.27301	12.4629
5	0.01322	1.86381	394.144	0.14825	5.19454	273.010
6	0.00176	1.35080	1551.86	0.10689	20.3700	5823.03
7	0.00023	0.96519	6044.79	0.0763	79.2857	1.24e+05
8	3.0e-05	0.68786	2.35e+04	0.05437	308.184	2.62e+06
9	3.9e-06	0.49000	9.13e+04	0.03873	1197.71	5.56e+07
10	5.1e-07	0.34911	3.55e+05	0.02759	4656.70	1.18e+09
11	6.7e-08	0.24923	1.39e+06	0.01968	18154.4	2.51e+10
12	8.8e-09	0.18078	5.59e+06	0.01414	7.19e+04	5.48e+11
13	1.1e-09	0.14825	2.91e+07	0.01058	3.12e+05	1.38e+13
14				0.01006	2.42e+06	8.72e+14

Table 9: Contributions of the individual exponential terms in the approximating hyperexponential distribution to the first three moments, for two Pareto distributions.



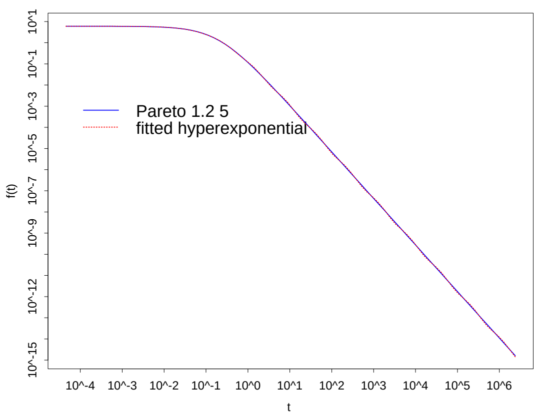
(a) Density

(b) Hazard

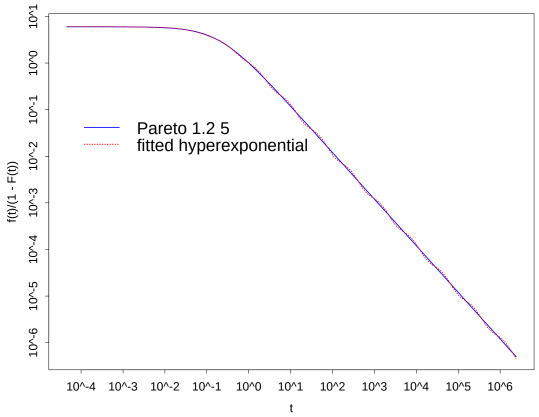
(c) Cumulative distribution

(d) Complementary cumulative distribution

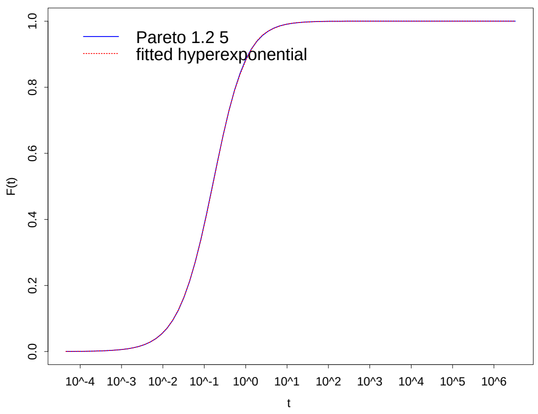
Figure 8: H_{13} fit to a Pareto(2.2, 0.83) distribution having mean 1 using $c_k = 0.1438$ and $c_1 = 10^7$.



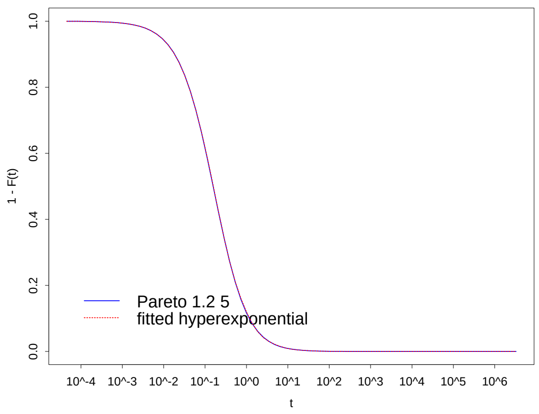
(a) Density



(b) Hazard

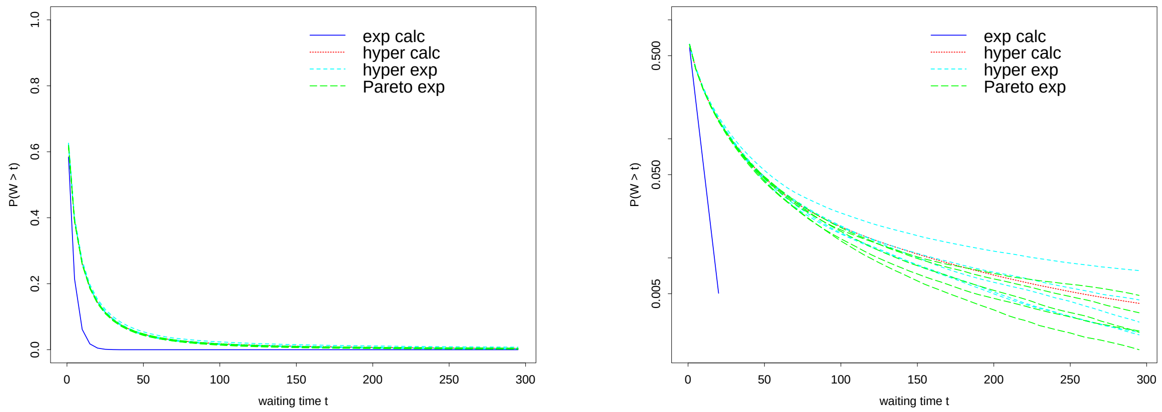


(c) Cumulative distribution



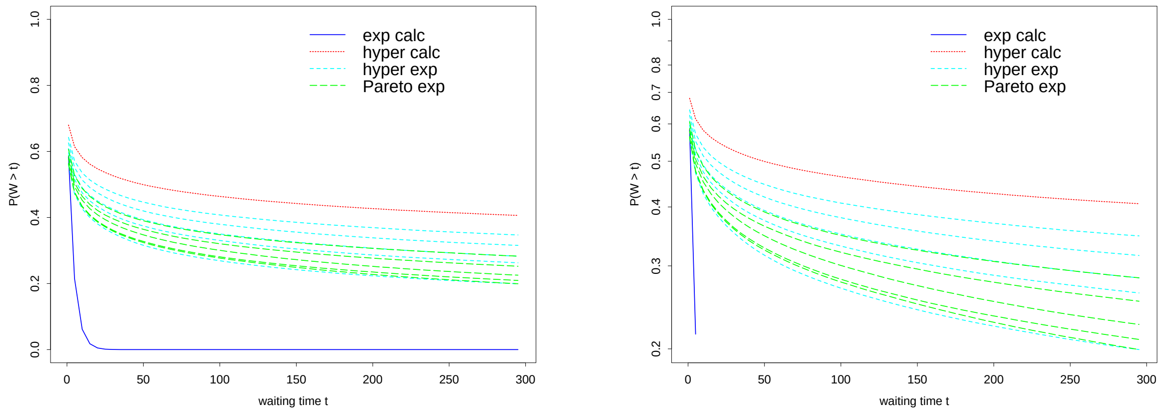
(d) Complementary cumulative distribution

Figure 9: H_{14} fit to a Pareto(1.2, 5) distribution having mean 1 using $c_k = 0.0264$ and $c_1 = 10^7$.



(a) (b) (log y-axis)

Figure 10: A comparison of numerical results and simulations of the steady-state $M/G/1$ waiting-time ccdf with a Pareto(2.2, 0.83) service-time distribution and the H_{13} fit using $c_k = 0.1438$ and $c_1 = 10^7$. Part (b) is the same as part (a), but with the y-axis in log scale.



(a) (b) (log y-axis)

Figure 11: A comparison of numerical results and simulations of the steady-state $M/G/1$ waiting-time ccdf with a Pareto(1.2, 5) service-time distribution having mean = 1 and the H_{14} fit using $c_k = 0.0264$ and $c_1 = 10^7$. Part (b) is the same as part (a), but with the y-axis in log scale.

6.3. Pareto mixtures of exponentials

Abate et al. [1] introduced the Pareto mixture of exponential (PME) distributions to study queues with long-tail service-time distributions. A PME pdf can be expressed as

$$f_r(t) = \int_{(r-1)/r}^{\infty} g_r(y) y^{-1} e^{-t/y} dy \ , \tag{6.1}$$

where $g_r(t)$ is a Pareto pdf on the interval $[(r-1)/r, \infty)$ of the form

$$g_r(t) = r\left(\frac{(r-1)}{r}\right)^r t^{-(r+1)} \ , \quad t \geq (r-1)/r \ . \tag{6.2}$$

We refer to a PME distribution with parameter r as PME(r). Since a PME pdf is constructed as a mixture of exponentials, it is completely monotone and thus DFR.

PME distributions are convenient to use in queueing examples because they have relatively convenient Laplace transforms. In general,

$$\hat{f}_r(s) = r\left(\frac{r-1}{r}\right)^r \int_0^{r/(r-1)} \frac{x^r}{s+x} dx \ . \tag{6.3}$$

Moreover, for $r = k$ or $k + 0.5$ for integer k, $\hat{f}_r(s)$ can be expressed in closed form, e.g.,

$$\hat{f}_3(s) = 1 - s + \frac{4}{3}s^2 - \frac{8}{9}s^3 \ln\left(1 + \frac{3}{2s}\right) \tag{6.4}$$

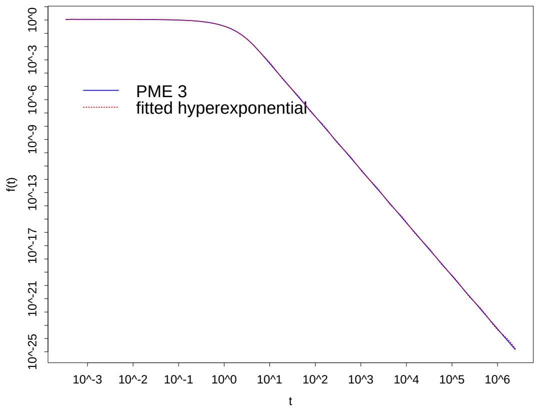
and

$$\hat{f}_{2.5}(s) = 1 - s + \frac{9s^2}{5} - 5(0.60)^{2.5} \ \operatorname{Arctan}(\sqrt{5/3s}) \ ; \tag{6.5}$$

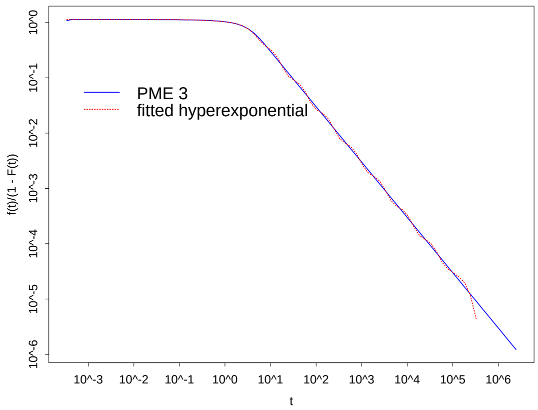
see Section 2 of [1].

This implies that it is possible to analyze the waiting time distribution of the $M/G/1$ queue if the service times are chosen from a PME distribution. Therefore the PME distribution is a good distribution to calibrate the performance of the fitting algorithm described in Section 4. Figure 12 shows the result of fitting a hyperexponential distribution with 10 exponentials to a PME(3) distribution. The parameters of the fitted hyperexponential distribution are shown in Table 10. Given all the other examples, it is no surprise that the fit is excellent. Only from the density and the hazard plots can we see that the hyperexponential distribution is only an approximation of the PME distribution. Note that the errors in the approximate H_k density and hazard occur outside the fitting interval (c_k, c_1).

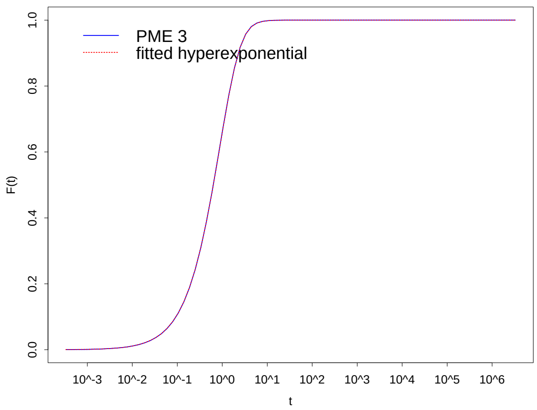
Figure 13 shows the analytical results of the waiting times of the $M/G/1$ queue for both the PME distribution and the fitted hyperexponential distribution. Particularly impressive is Figure 13 (b), where the ccdf waiting-time values are plotted in log scale. As in previous examples, the exponential service-time cdf chosen to match the mean yields a very poor approximation for the waiting-time ccdf. However, the numerical results for the waiting-time ccdf with the PME and the fitted H_{13} service-time cdf's are nearly identical, confirming that deviations seen previously in Figures 10 and 11 are due to simulation errors. This figure also illustrates the limitations of simulation. The waiting time probabilities, calculated from a simulation with the fitted hyperexponential distribution as service time distribution, deviate substantially from the analytical results for values larger than 200. The reason is obviously that the number of simulated arrivals is too small for this high level of variability.



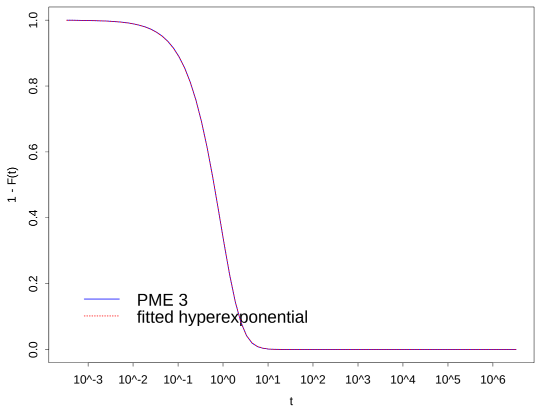
(a) Density



(b) Hazard



(c) Cumulative distribution

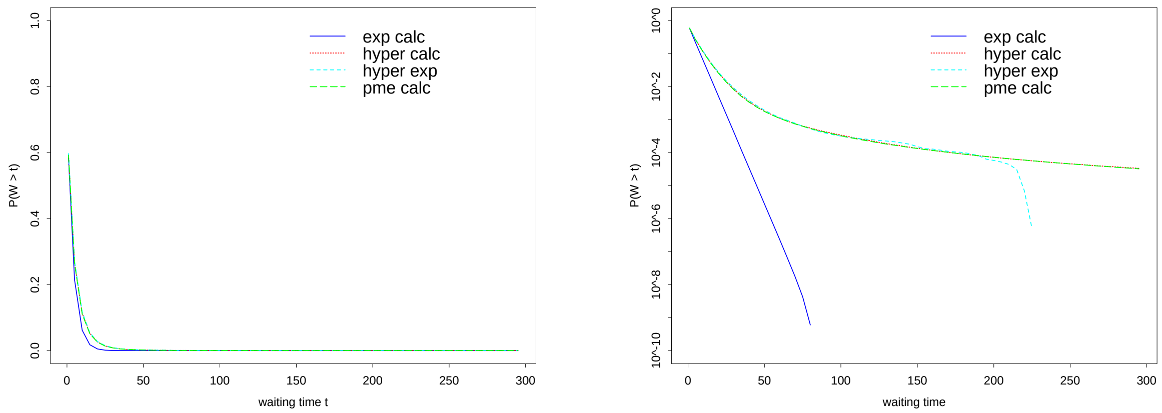


(d) Complimentary cumulative distribution

Figure 12: H_{12} fit to a PME(3) distribution using $c_k = 0.464$ to $c_1 = 10^6$.

	Parameters of the hyperexp. distribution		
i	p_i	λ_i	$1/\lambda_i$
1	0.055338	2.45024	0.40812
2	0.869780	1.11421	0.89749
3	0.073470	0.39385	2.53906
4	0.001386	0.10475	9.54655
5	2.60e-05	0.02782	35.9422
6	4.87e-07	0.00739	135.321
7	9.12e-09	0.00196	509.495
8	1.71e-10	0.00052	1918.61
9	3.20e-12	1.38e-4	7231.38
10	5.93e-14	3.65e-5	27385.9
11	1.05e-15	9.40e-6	106430
12	1.42e-17	2.08e-6	480898

Table 10: Parameters of the H_{12} cdf fit to a PME(3) distribution.



(a) (b) (log y-axis)

Figure 13: The steady-state $M/G/1$ waiting-time ccdf with a PME(3) service-time distribution and the H_{12} fit by the algorithm in Section 4 using $c_k = 0.464$ and $c_1 = 10^6$. Part (b) is the same as part (a), but with the y-axis in log scale.

7. Fitting a Hyperexponential Distribution to Data

Besides using the fitting algorithm to fit a hyperexponential distribution to another distribution, we can also use the fitting algorithm to fit a hyperexponential distribution to data. In this case the empirical ccdf obtained from the data replaces the ccdf of the initial probability distribution in the algorithm.

However, we would suggest caution when applying our algorithm directly to data. Our experience is that it is usually much better to first fit a suitable long-tail probability distribution with only a few parameters to the data, and then afterwards apply our algorithm to fit a multi-parameter hyperexponential distribution to the long-tail distribution. By this two-step procedure, we usually are able to obtain a good multi-parameter hyperexponential fit to data.

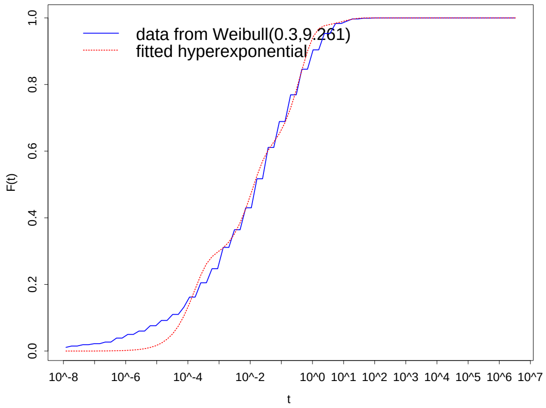
To illustrate, we consider a simulation experiment in which we try to fit a probability distribution to a sample of 1000 points drawn from the Weibull(.3, 9.261) distribution considered in Example 1.1 having unit mean. Even though the sample size is not very large, it is large enough to obtain a good fit to the two-parameter Weibull distribution using the maximum likelihood estimator (see p. 255 of Johnson and Kotz [31]). The Weibull parameters achieved from one sample were $c = 0.3016$ and $a = 9.369$ (yielding a mean of 0.96532). Since the estimated values of c and a are close to the original parameters, our algorithm applied to the fitted Weibull distribution can produce an excellent H_{20} approximation to the original Weibull distribution. For this experiment, the original Weibull distribution, the fitted Weibull distribution and the H_{20} fit to the fitted Weibull distribution are all very close, just as in Figure 1.

In contrast, we consider what happens when we apply our hyperexponential fitting algorithm directly to the data. Since the sample is not large, the range of the empirical ccdf is limited. Thus, it is not possible to directly apply our algorithm with many exponential terms. We show what happens with 4 exponentials. Figure 14 (a), (b) show how the fitted hyperexponential distribution matches the experimental cdf and ccdf. Figure 14 (c), (d) compare the fitted hyperexponential distribution to two Weibull distributions: the original Weibull distribution and the fitted Weibull distribution. Although the fits in Figure 14 look quite good, the pictures are deceptive, because the small and large values are not matched well. To illustrate, the moments are not matched well, as can be seen from Table 11. This can be explained in part by the fact that the sample moments of the data are not very close to the moments of the sampled distribution.

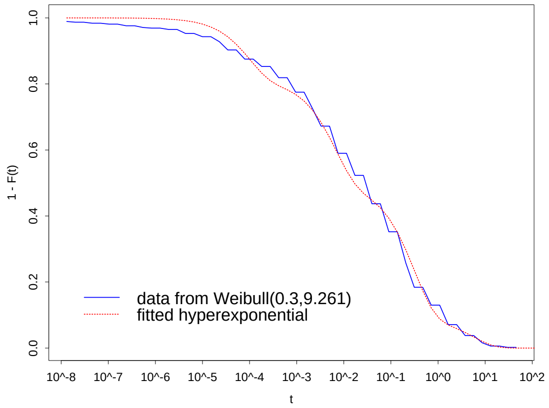
The experiment we have considered is somewhat biased, because we considered a hyperexponential fit to Weibull data. If we know in advance that the data is generated from the Weibull distribution, then using a statistical estimation procedure tailored to the Weibull distribution evidently should be good. It is less clear with an unknown data source. However, regardless of the data source, our fitting procedure is not designed to treat data. It does not address the statistical problems of the estimation. However, our procedure might well be applied effectively after some initial smoothing of the data, but that approach remains to be explored.

Moment	Weibull	Data	hyperexp.	fitted Weibull
1	1.00	0.93	0.44	0.97
2	29.24	24.72	5.26	26.64
3	4480.60	1591.75	145.24	3820.47

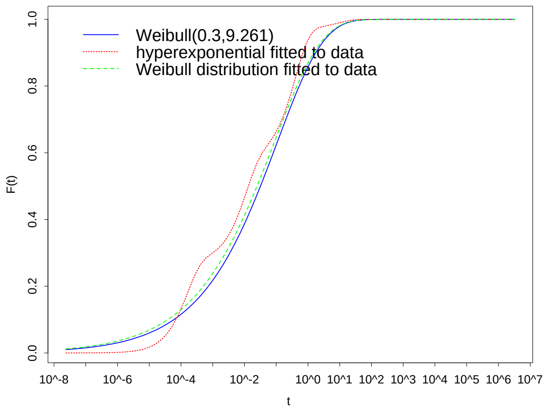
Table 11: Moments of the original Weibull distribution, the data sample, the fitted H_4 distribution, and fitted Weibull distribution.



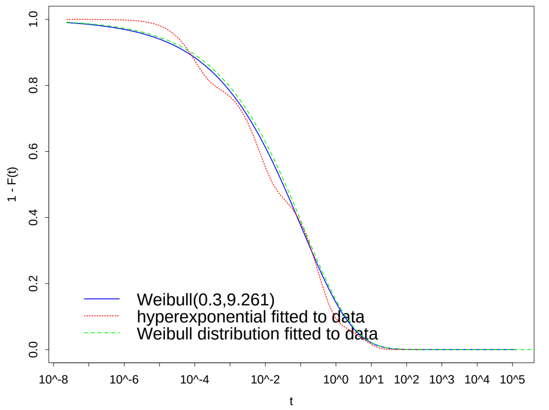
(a) Cumulative distribution



(b) Complementary cumulative distribution



(c) Cumulative distribution



(d) Complementary cumulative distribution

Figure 14: H_4 fit to the empirical cdf from a sample of size 1000 from a Weibull(.3, 9.261) distribution using $c_k = 0.0001$ and $c_1 = 5$.

8. Making Markov-Modulated On-Off Sources

A commonly considered model for sources in communication networks is the on-off model (e.g., [4, 15, 16, 28, 37, 49, 59]). In the basic on-off source model, the on and off periods come from independent sequences of i.i.d. random variables, with the on periods having cdf F_1 and the off periods having another common cdf F_2. During the on period there is input according to a Poisson process, a deterministic fluid process or some other stochastic process, and in the off period there is no input.

The special case in which F_1 and F_2 are exponentially distributed is especially convenient to analyze, because then the process indicating whether the source is active (on) or not (off) is Markov. Moreover, then the superposition of multiple independent sources of this kind is a Markov-modulated input process, with the state of the underlying Markov chain specifying whether each source is on or off. If the input during the on period of each source is a Poisson process, then the aggregate (superposition) process is a Markov-modulated Poisson process (MMPP). If the input in the on period of each source is a fluid process, then the aggregate input process is a Markov modulated rate process (MMRP). The input rate in any Markov chain state is then the sum of the rates for all the sources that are on in that state.

However, data from actual communication networks indicates that the on-period and off-period cdf's F_1 and F_2 often actually have long tails [59]. Unfortunately, this property makes the aggregate input process difficult to analyze directly. However, if we can fit the on-time and off-time cdf's F_1 and F_2 to hyperexponential distributions, then the aggregate input process can again be represented as a Markov modulated input process. To see this, let the on and off times for one source have ccdf's

$$F_1^c(t) = \sum_{i=1}^{k} p_i e^{-\lambda_i t} , \quad t > 0 \qquad \text{and} \qquad F_2^c(t) = \sum_{j=1}^{m} q_j e^{-\mu_j t} , \quad t > 0 ,$$

respectively. We let the underlying continuous-time Markov chain have $k+m$ states, with state i, $1 \le i \le k$, corresponding to the source being on with the component exponential having parameter λ_i, and state i, $k+1 \le i \le m$, corresponding to the process being off with the component exponential parameter having parameter μ_{k-i}. From state i, $1 \le i \le k$, the process transitions to state $k+j$, $1 \le j \le m$, with intensity $\lambda_k q_j$; from state $k+j$, $1 \le j \le m$, the process transitions to state i, $1 \le i \le k$, with intensity $\mu_j p_i$; and all other possible transitions have 0 intensity.

In order to treat the superposition process, the underlying Markov chain is the product of all the component Markov chains. If the Markov chain for source i has $k_i + m_i$ states, then the number of states in the Markov chain for the aggregate input process containing n component sources is

$$\prod_{i=1}^{n} (k_i + m_i) \; .$$

Clearly the number of states in the Markov chain underlying the aggregate process can be very large. This will occur when n, k_i or m_i are large. Since the long-tail property may lead to relatively large k_i and m_i, it clearly causes the Markov modulated model to become more difficult to analyze. Nevertheless, the H_k fit brings the model into the domain of existing algorithms. For example, algorithms for calculating the transient and steady-state performance characteristics in the $MMPP/G/1$ queue have been developed by Choudhury, Lucantoni and Whitt [15], [37]. (The $MMPP/G/1$ queue is a special case of the $BMAP/G/1$ queue.)

9. Conclusions

In this paper we have developed an effective simple algorithm for approximating a large class of probability distributions with monotone densities by hyperexponential distributions (Section 4). We have given examples showing that the algorithm is effective for approximating Pareto and Weibull distributions (Sections 1, 2, and 6). We have shown that the algorithm should be effective for distributions with decreasing failure rate, and should not be used for distributions with increasing failure rate (Section 5). We have proved that, in principle, completely monotone pdf's (all of which have decreasing failure rate) can be approximated arbitrarily closely by hyperexponential pdf's, and that as a result (under extra regularity conditions) the associated waiting-time distribution in a $GI/G/1$ queue with a completely monotone service-time distribution can be approximated arbitrarily closely by the waiting-time distribution in the associated $GI/G/1$ queue with the approximating hyperexponential service-time distribution (Sections 2 and 3). Since many long-tail distributions are completely monotone, these results serve as a theoretical foundation for approximating long-tail distributions by hyperexponential distributions. Since phase-type probability distributions are dense in the family of all probability distributions, by the same reasoning, they are rich enough to approximate any distribution, if enough phases are allowed. The EM algorithm is a candidate fitting algorithm for general phase-type distribution.

We believe that hyperexponential approximations of long-tail distributions can be useful, but they do not remove all difficulties. If a good fit is done, then the high variability of the long-tail distribution will be inherited by the approximating hyperexponential distribution. This high variability can make precise estimation by computer simulation difficult, as we saw in some of the examples in Section 6. In Section 8 we showed that hyperexponential approximations can make models of the superpositions of on-off sources more tractable, but since the state space of the Markovian environment process may be large, the approximating aggregate input process can still be difficult to analyze. However, we did see that the hyperexponential approximation makes it possible to calculate steady-state performance distributions in the $M/G/1$ queue with a long-tail service-time distribution by numerical transform inversion. The same technique applies to the more general $BMAP/G/1$ queue and other performance models.

We have emphasized that our fitting algorithm is intended to approximate one probability distribution by another, and not to fit a probability distribution directly to data (Section 7). In some circumstances our algorithm could be used to fit a hyperexponential distribution to an empirical distribution (histogram) obtained from data, but our algorithm is not designed for that purpose. Indeed, in simulation experiments with long-tail data, we found that much better fits are obtained by first fitting a long-tail distribution with very few parameters (e.g., 2) to the data and then applying our algorithm to obtain a hyperexponential distribution.

Finally, the algorithm presented here is only one of many possible fitting algorithms. We intend to compare alternative fitting algorithms in a future paper.

Acknowledgment

We thank our colleague William Turin for helpful discussions about the EM algorithm and the history of recursive estimation, including the reference to de Prony (1795) [48].

References

[1] J. Abate, G. L. Choudhury and W. Whitt, Waiting-time tail probabilities in queues with long-tail service-time distributions, *Queueing Systems* 16 (1994) 311–338.

[2] J. Abate and W. Whitt, The Fourier-series method for inverting transforms of probability distributions, *Queueing Systems* 10 (1992) 5–88.

[3] J. Abate and W. Whitt, An operational calculus for probability distributions via Laplace transforms, *Adv. Appl. Prob.* 28 (1996) 75–113.

[4] A. T. Andersen, A. Jensen and B. F. Nielsen, Modelling and performance study of packet-traffic with self-similar characteristics over several time-scales with Markovian arrival processes (MAP) *Twelfth Nordic Teletraffic Seminar, NTS12* (1995) 269–283.

[5] S. Asmussen, *Applied Probability and Queues*, Wiley, New York, 1987.

[6] S. Asmussen, L. F. Henriksen and C. Klüppelberg, Large claims approximations for risk processes in a Markovian environment, *Stoch. Proc. Appl.* 54 (1994) 29–43.

[7] S. Asmussen and G. Koole, Marked point processes as limits of Markovian arrival streams, *J. Appl. Prob.* 30 (1993) 365–372.

[8] S. Asmussen, O. Nerman and M. Olsson, Fitting phase type distributions via the EM algorithm, *Scand. J. Statist.* 23 (1996) 419–441.

[9] R. E. Barlow and F. Proschan, *Statistical Theory of Reliability and Life Testing*, Holt, Rinehart and Winston, New York, 1975.

[10] S. N. Bernstein, Sur les fonctions absolument montones, Acta Math. 51 (1928) 1–66.

[11] P. Billingsley, *Convergence of Probability Measures*, Wiley, New York, 1968.

[12] V. A. Bolotin, Modeling call holding time distributions for CCS network design and performance analysis. *IEEE J. Sel. Areas Commun.* 12 (1994) 433–438.

[13] A. A. Borovkov, *Stochastic Processes in Queueing Theory*, Springer-Verlag, New York, 1976.

[14] R. Cáceres, P. B. Danzig, S. Jamin, and D. J. Mitzel. Characteristics of wide-area TCP/IP conversations. *Computer Communication Review* 21 (1991).

[15] G. L. Choudhury, D. M. Lucantoni and W. Whitt, Squeezing the most out of ATM, *IEEE Trans. Commun.* 44 (1996) 203–217.

[16] G. L. Choudhury and W. Whitt, Long-tail buffer-content distributions in broadband networks, *Perf. Eval.*, to appear.

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

[18] M. E. Crovella and A. Bestavros. Self-similarity in World Wide Web traffic - evidence and possible causes. *Proceedings of Sigmetrics'96* (1996) 160-169.

[19] N. G. Duffield, Economies of scale in queues with sources having power-law large deviations scalings, *J. Appl. Prob.*, to appear.

[20] N. G. Duffield and N. O'Connell, Large deviations and overflow probabilities for the general single-server queue, with applications, *Math. Proc. Camb. Phil. Soc.*, 118 (1995) 363–374.

[21] D. E. Duffy, A. E. McIntosh, M. Rosenstein and W. Willinger, Statistical analysis of CCSN/SST traffic data from working CCS subnetworks. *IEEE J. Sel. Areas Commun.* 12 (1994) 544–551.

[22] A. Feldmann. *On-line Call Admission for High-Speed Networks.* PhD thesis, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, 1995.

[23] A. Feldmann, Modeling characteristics of TCP connections. AT&T Laboratories, 1996.

[24] W. Feller, *An Introduction to Probability Theory and Its Applications*, Wiley, New York, 1971.

[25] D.P. Gaver and P.A. Jacobs, Waiting times when service times are stable laws: tamed and wild, Dept. of Operations Research, Naval Postgraduate School, Monterey, CA, 1995.

[26] C. M. Harris, The Pareto distribution as a queue service distribution, *Opns. Res.* 16 (1968) 307–313.

[27] M. R. Izquierdo and D. R. Reeves, Statistical characterization of MPEG VBR video at the slice layer, *Proceedings of the SPIE - The International Society for Optical Engineering* (1995) 268–279.

[28] R. Jain and S.A. Routhier. Packet trains: measurements and a new model for computer network traffic. *IEEE Journal on Selected Areas in Communications* 4 (1986) 986–995.

[29] P. R. Jelenković and A. A. Lazar, Subexponential asymptotics of a Markov-modulated $G/G/1$ queue, *J. Appl. Prob.*, to appear.

[30] P. R. Jelenković, A. A. Lazar and N. Semret, The effect of multiple time scales and subexponentiality of MPEG video streams on queueing behavior, Dept. of Electrical Engineering, Columbia University, 1996.

[31] N. L. Johnson and S. Kotz, *Distributions in Statistics, Continuous Univariate Distributions*, Wiley, New York, 1970.

[32] V. V. Kalashnikov and S. T. Rachev, *Mathematical Methods for Construction of Queueing Models*, Wadsworth & Brooks/Cole, Pacific Grove, CA, 1990.

[33] J. Keilson, *Markov Chain Models — Rarity and Exponentiality*, Springer-Verlag, New York, 1979.

[34] W. E. Leland, M. S. Taqqu, W. Willinger and D. V. Wilson, On the self-similar nature of Ethernet traffic. *ACM/SIGCOMM Computer Communications Review*, 23 (1993) 183–193.

[35] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson. On the self-similar nature of Ethernet traffic. *IEEE/ACM Transactions on Networking* 2 (1994) 1–15.

[36] D. M. Lucantoni, The $BMAP/G/1$ queue: a tutorial, in *Models and Techniques for Performance Evaluation of Computer and Communication Systems*, L. Doniatiello and R. Nelson eds., Springer-Verlag, New York (1993) 330-358.

[37] D. M. Lucantoni, G. L. Choudhury and W. Whitt, The transient $BMAP/G/1$ queue, *Stochastic Models* 10 (1994) 145–182.

[38] W. T. Marshall and S. P. Morgan. Statistics of mixed data traffic on a local area network. *Computer Networks and ISDN Systems* (1985) 185–195.

[39] K. Meier-Hellstern, P. E. Wirth, Y.-L. Yan, and D. A. Hoeflin. Traffic models for ISDN data users: Office automation application. In *Proceedings of the 13th ITC* (1991) 167–172.

[40] J. Mogul. Network behavior of a busy web server and its clients. Technical Report 95/5, Digital Equipment Corp. Western Research Laboratory, 1995.

[41] M. Montgomery and G. de Veciana. On the relevance of time scales in performance oriented traffic characterizations. *IEEE INFOCOM'96* 513–520.

[42] M. F. Neuts, *Matrix-Geometric Solutions in Stochastic Models*, The Johns Hopkins University Press, Baltimore, 1981.

[43] M. F. Neuts, *Structured Stochastic Matrices of $M/G/1$ Type and Their Applications*, Marcel Dekker, New York, 1989.

[44] P. Pawlita, Two decades of data traffic measurements: a survey of published results, experiences and applicability, *Proc. 12th Int. Teletraffic Congress, Torino, Italy*, (1988) 5.2.A.5.

[45] V. Paxson. Empirically derived analytic models of wide-area TCP connections: extended report. Technical Report LBL-34086, Lawrence Berkeley Laboratory, 1993.

[46] V. Paxson. Empirically derived analytic models of wide-area TCP connections. *IEEE/ACM Transactions on Networking* 2 (1994) 316–336.

[47] V. Paxson and S. Floyd. Wide-area traffic: the failure of Poisson modeling. *IEEE/ACM Transactions on Networking* 3 (1995) 226–244.

[48] R. de Prony, Essai experimentale et analytique, *J. Ecole Polytechnique* (1795) 24–76.

[49] S. Robert and J. Y. Le Boudec. Can self-similar traffic be modelled by Markovian processes? *Broadband Communications, IZS'96 (Bernhard Plattner, ed.)* Springer-Verlag (1996) 119–130.

[50] W. Turin, *Performance Analysis of Digital Transmission Systems*, Computer Science Press, New York, 1990.

[51] W. Turin, Fitting probabilistic automata via the EM algorithm, *Stochastic Models* 12 (1996) 405–424.

[52] W. Whitt, The continuity of queues, *Adv. Appl. Prob.* 6 (1974) 175–183.

[53] W. Whitt, Continuity of generalized semi-Markov processes, *Math. Oper. Res.* 5 (1980) 494–501.

[54] W. Whitt, Approximating a point process by a renewal process, I: two basic methods, *Opns. Res.* 30 (1982) 125–147.

[55] W. Whitt, On approximations for queues, III: mixtures of exponential distributions, *AT&T Bell Lab. Tech. J.* 63 (1984) 163–175.

[56] W. Whitt, Planning queueing simulations, *Management Sci.* 35 (1989) 1341–1366.

[57] W. Whitt, Approximations for the $GI/G/m$ Queue, *Production and Operations Management* 2 (1993) 114–161.

[58] W. Willinger, M. S. Taqqu, W. E. Leland and D. V. Wilson, Self-similarity in high-speed packet traffic: analysis and modeling of Ethernet traffic measurements, *Statistical Science* 10 (1995) 67–85.

[59] W. Willinger, M. S. Taqqu, R. Sherman and D. V. Wilson, Self similarity through high variability: statistical analysis of Ethernet LAN traffic at the source level. *SIGCOMM'95* 100–113.