Fitting World-Wide Web Request Traces with the EM-Algorithm

Rachid El Abdouni Khayari[a], Ramin Sadre[a] and Boudewijn Haverkort[a]

[a] Laboratory for Performance Evaluation and Distributed Systems
RWTH Aachen, Department of Computer Science
D-52056 Aachen, Germany

ABSTRACT

In recent years, several studies have shown that network traffic exhibits the property of *self-similarity*. Traditional (Poissonian) modelling approaches have been shown not to be able to describe this property and generally lead to the underestimation of interesting performance measures. Crovella and Bestavros[1] have shown that network traffic that is due to World Wide Web transfers shows characteristics of self-similarity and they argue that this can be explained by the heavy-tailedness of many of the involved distributions. Considering these facts, developing methods which are able to handle self-similarity and heavy-tailedness is of great importance for network capacity planing purposes.

In this paper we discuss two methods to fit hyper-exponential distributions to data sets which exhibit heavy-tails. One method is taken from the literature and shown to fall short. The other, new method, is shown to perform well in a number of case studies.

Keywords: World wide web, heavy-tailed distributions, hyper-exponential distributions, ML-fitting, EM-fitting, queueing analysis, traffic characterisation

1. INTRODUCTION

Over the last decade, extensive traffic measurements have shown the presence of properties such as *self-similarity*, *fractality* and *long-range dependency* in network traffic. The seminal paper by Leland *et al.*,[2] showed self-similarity in Ethernet traffic; later, similar effects were shown to exist in wide area network traffic, signaling traffic, and in multimedia and video traffic. Also, it has been have shown that ignoring these effects in the analysis of queueing systems leads in general to undervaluation of important performance measures.[3] Additionally, studies have shown that the presence of these properties is generally correlated with the presence of *heavy-tailed distributions* (HTDs).

Various efforts have been persued to develop appropriate traffic models to evaluate the performance of systems und self-similar traffic.[4–6] In the sequel we will focus on the approach put forward by Feldmann and Whitt[5] (the FW approach). The FW-approach proposes a method to fit a hyper-exponential distribution to a given HTD. Although this method is fast, it requires an explicit representation of an HTD, for which either a Weibull or a Pareto distribution can be used. However, as will be shown below, often the measurements to be fitted do not suit a Weibull or a Pareto distribution, so that the final HED obtained does not describe the measurements well. The FW-approach is illustrated in the upper-half of Figure 1.

To avoid the use of an intermediate HTD, we have decided to directly fit a HED to the measured data via the Expectation Maximization (EM-) algorithm. This is described in detail in the paper; see also the lower half of Figure 1.

To validate the new fitting approach, we use both a large trace from the RWTH proxy server and a well-known NASA trace and fit the requested object-size distribution. We both study the fit itself, as well as performance results obtained when using the fitted distributions in a discrete-event simulation of an M|G|1 queue for different utilisations (also in Figure 1; the right-most comparison).

This paper is further organised as follows. We will give some background on HTDs in Section 2. Then, we describe the FW-approach in Section 3. The new fitting approach is discussed in Section 4. We validate (and compare) the new approach in Section 5. The paper is concluded in Section 6 with some final remarks. Two appendices, on statistical fitting procedures, are included to make the paper self-contained.

Further author information: `http://www-lvs.informatik.rwth-aachen.de/`

Internet Performance and Control of Network Systems II, Robert D. van der Mei, Frank Huebner-Szabo de Bucs, Editors, Proceedings of SPIE Vol. 4523 (2001)

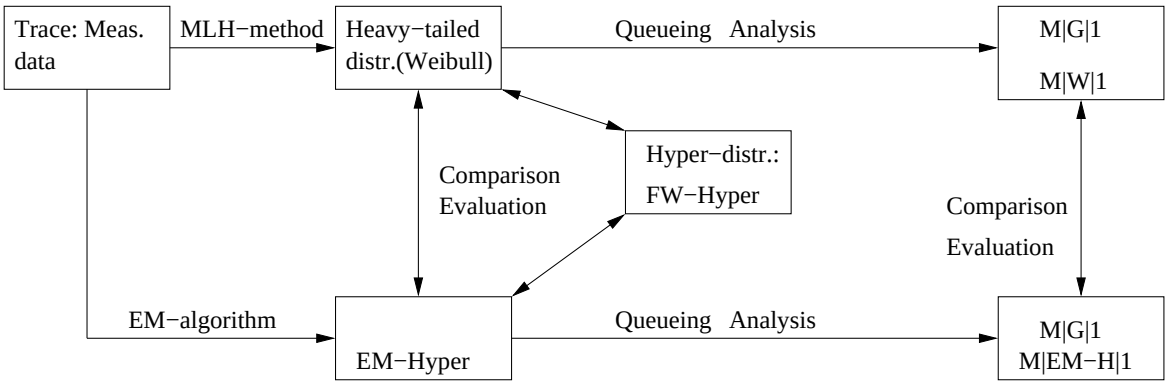

Figure 1. Graphical representation of the two fitting procedures

2. HEAVY-TAILED DISTRIBUTIONS

Self-similarity in network traffic has been explained by the fact that many of the involved distributions, e.g., of file sizes, are heavy-tailed. In a HTD, the complementary cumulative distribution function F^c decays more slowly than exponentially, i.e., $e^{\gamma t}F^c(t) \to \infty$ as $t \to \infty$ for all $\gamma > 0$. For a random variable X, distributed according to some HTD, we typically have:

$$P[X > x] \sim x^{-\alpha}, \qquad x \to \infty, \qquad 0 < \alpha < 2. \tag{1}$$

Heavy-tailed distributions have an infinite variance. The degree of the heavy-tailedness is given by the value of the shape parameter α.

The following table gives some characteristics of two well-known HTDs, namely the Pareto and the Weibull distribution (in case the stated conditions are not met, the expectation and/or variance do not exist)[7]:

Name	Density $f(x)$	Expectation	Variance
Pareto	$ak^a x^{-(a+1)}$	$\frac{ak}{a-1}$, for $a > 1$	$\frac{ak^2}{(a-2)(a-1)^2}$, for $a > 2$
Weibull	$\frac{b}{a^b}x^{b-1}e^{-(\frac{x}{a})^b}$, for $a > 0$ and $b > 0$	$\frac{a}{b}\Gamma(1/b)$	$\frac{a^2}{b^2}\{2b\Gamma(2/b) - [\Gamma(1/b)]^2\}$

3. APPROXIMATION OF HTDS WITH HEDS

In this section we present the FW-approach towards approximating HTDs with HEDs.

3.1. Description of the method

In the FW-approach, it is assumed that an HTD is given in an explicit form. How this HTD is obtained from, for instance, measurement data, is not described by Feldmann and Whitt; see our comments below. Provided that an explicit representation of the HTD $F(x)$ is available, an I-phase HED distribution of the form

$$H(x) = 1 - \sum_{i=1}^{I} c_i e^{-\lambda_i x} \tag{2}$$

is fitted to F. Note that, for $I \to \infty$, one can represent *any* distribution, with squared coefficient of variation at least 1, with completely monotone probability density function arbitrary close by hyper-exponentials. However, it is shown[5] that with values of I up to 20, HTDs can approximate Weibull and Pareto distributions for large ranges of x. For given HTD $F(x)$ and number of phases I the FW-approach operates as follows:

1. Choose quantiles $0 < q_I < q_{I-1} < \ldots < q_1$ with sufficiently large q_i/q_{i+1}, e.g., $q_i/q_{i+1} \approx 10$ (for $i = 1, \cdots, I-1$). Furthermore, let b be such that $1 < b < q_i/q_{i+1}$ for all i.

2. In I iterative steps, the parameters for the phases in the HED are computed. We start with setting $j := 1$ and $F_1^c(x) = 1 - F(x)$.

3. In the j-th phase, we compute c_j and λ_j by solving the equations

$$\begin{aligned} c_j e^{-\lambda_j q_j} &= F_j^c(q_j), \\ c_j e^{-\lambda_j b q_j} &= F_j^c(b q_j), \end{aligned}$$

yielding

$$\begin{aligned} \lambda_j &= \frac{1}{(b-1)q_j} \ln\left(\frac{F_j^c(q_j)}{F_j^c(b q_j)}\right), \\ c_j &= F_j^c(q_j) e^{\lambda_j q_j}. \end{aligned}$$

4. Repeat step 3 for $j = 2, \ldots, I-1$ where

$$\begin{aligned} F_i(q_i) &= F_{i-1}(q_i) - c_{i-1} e^{-\lambda_{i-1} q_i}, \\ F_i(b q_i) &= F_{j-i}(b q_i) - c_{j-i} e^{-\lambda_{i-1} b q_i}. \end{aligned}$$

5. Finally, for the last phase I we find: $c_I := 1 - \sum_{j=1}^{I-1} c_j$ and λ_I follows from $c_I e^{-\lambda_I q_I} = F_I^c(q_I)$.

The complexity of the algorithm is $\mathcal{O}(I)$. However, since the algorithm cannot be applied directly to measurement data, the costs of an algorithm, like the maximum likelihood (ML) algorithm[8] (see Appendix A) to fit the measurements to an explicit HTD must be considered as well.

3.2. Application and validation

When applying the FW-approach to find object-size distribution from the log-files used in our case study (for a detailed description of the traces and its statistical parameters, see Section 5), we found that the typically employed HTDs, like Pareto and Weibull do not describe the object size distributions well. Both distributions fit the tail of emperical measurement distribution well, but fail to fit the head properly. For example, a Weibull distribution whose first and second moment have been fitted to the data, results in a median that is half the median found in the data (the median is located in the head). Hence, even when the FW-approach does give a good fit with respect to a given HTD, if the provided HTD does not describe the data well, then the finally fitted HED does not describe the measurements well, too.

Feldmann and Whitt[5] point out that it might be possible to extend their approach so it can be directly applied to measurement data. They also warned that the algorithm, at least without extension, is not designed to directly treat data but might well be applied after some initial smoothing of the data. In fact, our studies have shown that the smoothing is absolute necessary, since otherwise the algorithm is too sensitive to the location of the quantiles q_i. Furthermore, the quality of the approximation heavily depends on the quality of the smoothing.

4. FITTING DIRECTLY TO HEDS

The Expectation Maximization (EM) algorithm is a well-known algorithm to fit measurements to distributions[9–12]; a detailed description can be found in Appendix B. The EM-algorithm works iteratively and does require neither an intermediate HTD nor heuristics. Below, we outline the method in general, and then specialise it to the case where the distribution function to fit to is a HED.

4.1. General approach

Given measurement data $x_1, \ldots, x_N$, we search the parameters $\underline{c} = (c_1, \ldots, c_I)$ and $\underline{\theta} = (\underline{\theta_1}, \ldots, \underline{\theta_I})$ of a distribution with density function

$$p(x|(\underline{c}, \underline{\theta})) = \sum_{i=1}^{I} c_i \cdot p(x|\underline{\theta_i}) \tag{3}$$

so that this distribution "best" fits the distribution of the measurement data. The density in (3) is a convex combination of basic density functions $p(x|\underline{\theta_i})$ parameterized by $\underline{\theta_i}$ with weights $c_i \geq 0$ and $\sum_{i=1}^{I} c_i = 1$. For instance, for Weibull distributions as basic distributions $\underline{\theta_i}$ would be (a_i, b_i), and for exponential distributions we would have $\underline{\theta_i} = (\lambda_i)$.

Now, let $\alpha = (\underline{c}, \underline{\theta})$ and $\hat{\alpha} = (\underline{\hat{c}}, \underline{\hat{\theta}})$ be two sets of parameters for the density p. The EM-algorithm defines a new distribution with density function

$$h(i|x_n, \underline{\theta_i}) = \frac{c_i \cdot p(x_n|\underline{\theta_i})}{p(x_n|\alpha)}, \tag{4}$$

as well as the following function

$$Q(\alpha, \hat{\alpha}) = \sum_{n=1}^{N} \sum_{i=1}^{I} h(i|x_n, \underline{\theta_i}) \cdot \log\left(\hat{c}_i \cdot p(x_n|\alpha) h(i|x_n, \alpha)\right). \tag{5}$$

The function Q provides a quality criterion for α and $\hat{\alpha}$: it says how much better the density function $p(x|\hat{\alpha})$ fits the measurement data than the density function $p(x|\alpha)$.

The EM algorithm proceeds iteratively: starting from an initial parameter set $\alpha = (\underline{c}, \underline{\theta})$, it computes a new parameter set $\alpha' = (\underline{c'}, \underline{\theta'})$ which maximizes $Q(\alpha, \alpha')$, that is, with α' such that $Q(\alpha, \hat{\alpha} := \alpha')$ is maximized, we improve the fit the most. This α' is used as starting point for the next iteration. The algorithm stops when $\alpha \approx \alpha'$ (see below). To find the next value α', the EM algorithm has to solve the equation system (possibly non-linear):

$$\frac{\partial Q}{\partial \alpha'} = 0 \Rightarrow \frac{\partial Q}{\partial \underline{\theta}_1} = 0, \cdots, \frac{\partial Q}{\partial \underline{\theta}_I} = 0. \tag{6}$$

Using Lagrange multipliers (with auxiliary condition $\sum_{i=1}^{I} c_i = 1$), the new weights are given by:

$$c_i' \quad = \quad \frac{1}{N} \sum_{n=1}^{N} \frac{c_i p(x_n|\underline{\theta_i})}{p(x_n|\alpha)}. \tag{7}$$

In general, the non-linear equation system (6) is difficult to solve. However, in case we take HEDs as basic densities, it is feasible. We discuss this in the next section.

4.2. Specialisation to HEDs

We now take HEDs as basic distribution functions: $p(x|\lambda_i) = \lambda_i e^{-\lambda_i x}$. Equation (6) yields

$$\frac{\partial Q}{\partial \lambda_i'} = 0 \quad \Rightarrow \quad \sum_{n=1}^{N} h(i|x_n, \lambda_i) \cdot \frac{\partial}{\partial \lambda_i'} \log\left(c_i' \cdot p(x_n|i, \lambda_i')\right) = 0. \tag{8}$$

$$\Rightarrow \quad \sum_{n=1}^{N} h(i|x_n, \lambda_i) \cdot \frac{\partial}{\partial \lambda_i'} \log\left(c_i' \cdot \lambda_i' \cdot e^{-\lambda_i' x_n}\right) = 0. \tag{9}$$

$$\Rightarrow \quad \sum_{n=1}^{N} h(i|x_n, \lambda_i) \cdot \frac{\partial}{\partial \lambda_i'} \left(\log c_i' + \log \lambda_i' - \lambda_i' x_n\right) = 0. \tag{10}$$

$$\Rightarrow \quad \sum_{n=1}^{N} h(i|x_n, \lambda_i) \cdot (1/\lambda_i' - x_n) = 0. \tag{11}$$

$$\Rightarrow \quad \frac{\sum_{n=1}^{N} h(i|x_n, \lambda_i)}{\lambda_i'} = \sum_{n=1}^{N} h(i|x_n, \lambda_i') \cdot x_n. \tag{12}$$

$$\Rightarrow \quad \lambda_i' = \frac{\sum_{n=1}^{N} h(i|x_n, \lambda_i)}{\sum_{n=1}^{N} h(i|x_n, \lambda_i) \cdot x_n}. \tag{13}$$

The EM-algorithm now takes the following form:

1. Select an appropriate number of distributions I and select start values λ_i and c_i $(i = 1, \cdots, I)$, as well as a positive real error boundary value ϵ (with $p(x_n|\lambda_i) = \lambda_i e^{-\lambda_i x_n}$).
2. Compute for $i := 1$ to I:
   (a) $p(i|x_n, \lambda_i) = \frac{c_i \cdot p(x_n|\lambda_i)}{p(x_n)}$
   (b) $c_i' = \frac{1}{N} \sum_{n=1}^{N} p(i|x_n, \lambda_i)$
   (c) $\lambda_i' = \frac{\sum_{n=1}^{N} p(i|x_n, \lambda_i)}{\sum_n p(i|x_n, \lambda_i) \cdot x_n}$
3. Return to step 2 with $c_i := c_i'$ and $\lambda_i := \lambda_i'$ until the difference between c_i and c_i' and/or the difference between λ_i and λ_i' for all i is smaller than the boundary value ϵ.

In the above algorithm we have preset the number of phases I (also called the number of centers). In a different variant of the EM-algorithm, this number does not have to be preset, but is computed on-the-fly, thus yielding a number of phases that is large enough to describe the required HTD, yet as small as possible to keep the fitted HED small.[10,13]

Regarding computational complexity, the EM-algorithm is an iterative algorithm where the complexity of each iteration is $\mathcal{O}(N \cdot I)$, with N the number of measurement samples and I the number of centers. A problem with the EM-algorithm is the fact that it is difficult to predict the number of iterations needed to reach a given precision of the result.[9] However, in our experiments, good results generally have been obtained within 5–10 iterations. Additionally, it should be noted that even for a case study with N well over 17 million (see below) and $I = 6$, one iteration took approximately 1 minute on a standard personal computer.

5. APPLICATION

5.1. Statistics of the measurement data

Crovella and Bestavros[1] have shown that network traffic that is due to WWW transfers can show characteristics that are consistent with self-similarity and that this can be explained by the heavy-tailedness of many of the involved distributions. Since traffic originating from `HTTP` transfers amounts to at least 85% of all Internet traffic,[14] understanding its nature is crucial.

RWTH trace

Early 2000, we collected the access logs of the RWTH Aachen proxy server. The logs comprise the description of about 115 million `HTTP` and `FTP` requests made over a period of 54 days. After some preprocessing and filtering, about 17.3 million requests of interest remained. We studied the sizes of the objects requested by the clients. Figure 2 shows the complementary distribution of the object sizes as log-log plot. Some statistics from these traces:

$$\begin{aligned} \text{minimum object size} &= 118 \text{ (byte)} \\ \text{maximum object size} &= 10^7 \text{ (byte)} \\ \text{expected object size} &= 6664 \text{ (byte)} \\ \text{median object size} &= 2638 \text{ (byte)} \\ \text{squared coefficient of variation} &= 6.12 \end{aligned}$$

As can be observed from both Figure 2 and the statistics listed above, the distribution function decays much slower than a negative exponential distribution and is clearly heavy tailed. In particular, we also observe a median much smaller than the mean.

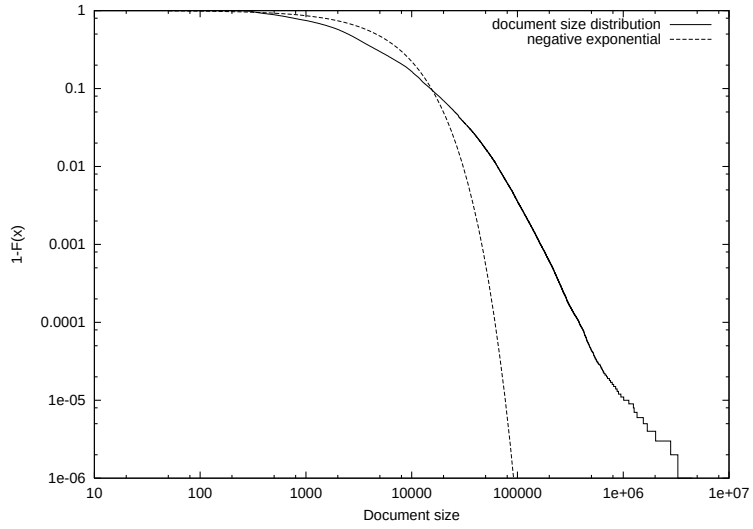

Figure 2. RWTH trace: Complementary log-log plot of document size distribution

	RWTH trace	Weibull	FW:H_5	FW:H_{10}	FW:H_{20}	EM:H_5	EM:H_{10}
$E[X]$	6663.69	6663.69	7025.25	5097.09	4977.77	6663.69	6663.69
CV^2	6.12	6.11	1.949	4.20152	4.49	6.16 (0.3%)	6.22 (1.6%)
$E[X^3]$	$2.748 \cdot 10^{14}$	$4.04 \cdot 10^{13}$	–	–	–	$2.59 \cdot 10^{14}$ (5.5%)	$2.91 \cdot 10^{14}$ (5.8%)
median	2638	1322	–	–	–	2638 (<0.1%)	2530 (4%)

Table 1. RWTH trace: Comparison of statistics of the measurement data, a fitted HTD and two fitted HEDs

NASA trace

The next trace we use for validation was first presented and evaluated in 1996 by Arlitt and Williamson.[15] It consists of about 3.1 million requests collected at the web server of the Kennedy Space Center. As in the RWTH trace, the size distribution of the requested objects in the NASA trace is clearly heavy tailed yielding a high coefficient of variation and an expectation much larger than the median:

$$\begin{aligned} \text{minimum object size} &= 3 \text{ (byte)} \\ \text{maximum object size} &= 6.823 \cdot 10^6 \text{ (byte)} \\ \text{expected object size} &= 20744.9 \text{ (byte)} \\ \text{median object size} &= 4142 \text{ (byte)} \\ \text{squared coefficient of variation} &= 13.3853 \end{aligned}$$

5.2. Matching HTDs

RWTH trace

In Table 1 we shows how well the two algorithms match the moments and median of the measurement data from the RWTH-trace. As can be observed, the fitted Weibull distribution (using the ML approach) fails to fit the third moment and the median of the data set well. The three hyperexponentials fitted with the FW-approach do not match the expectation and squared coefficient of variation of the trace. We did not compute their third moment and median. In contrast, both HEDs fitted with the new algorithm do fit these two statistics well. The relative error (written in parenthesis) is below 6% for all statistics.

	NASA trace	Weibull	FW:H_5	FW:H_{10}	FW:H_{20}	EM:H_5	EM:H_{10}
$E[X]$	20744.9	20793	169950	46829.5	12475	20744.9	20744.9
CV^2	13.38	13.44	1.006	1.2068	11.24	13.67 (2.1%)	13.67 (2.1%)
$E[X^3]$	$5.02 \cdot 10^{15}$	$6.73 \cdot 10^{15}$	–	–	–	$5.74 \cdot 10^{15}$ (12.5%)	$5.74 \cdot 10^{15}$ (12.5%)
median	4142	1764	–	–	–	3847 (7.6%)	3847 (7.6%)

Table 2. NASA trace: Comparison of statistics of the measurement data, a fitted HTD and two fitted HEDs

first-order queueing performance				
ρ	0.67		0.83	
$E[N]$	5.47		15.83	
$E[W]$	48,047		1,199,958	
second-order queueing performance				
measure	c_W^2	RE (%)	c_W^2	RE (%)
trace	10.40 (9.5%)		5.19 (2.3%)	
Weibull	2.89 (3.3%)	72	1.71 (1.6%)	67.1
H_5	10.56 (13%)	15	4.68 (5.0%)	9.8
H_{10}	11.38 (15%)	9.4	5.41 (9.7%)	4.2

Table 3. RWTH trace: Queueing performance for the measurement data, a fitted HTD and two fitted HEDs

NASA trace

The matching results for the NASA trace are shown in Table 2. They are comparable to those obtained for the RWTH trace. However, it seems that the higher degree of heavy tailedness of the NASA trace results in larger errors for the matched HTDs.

5.3. Embedding HTDs in queueing models

In the second validation step we used the fitted distribution for the RWTH-trace as service-time distribution in an M|G|1 queue (modelling a proxy server). Using a discrete-event simulator, we studied mean queueing measures for two different service loads ($\rho = 2/3$ and $\rho = 5/6$). These are depicted in the upper half of Table 3. As far as mean queueing performance is concerned, the results for the fitted Weibull, the fitted HEDs and the measurement data (in a trace-driven simulation) are all the same (we show these results only once). However, in the lower half of Table 2 we show a number of higher-order queueing statistics for the four different distributions. We show, again for the two levels of utilisation, the squared coefficient of variation of the waiting time, the 95% confidence intervals relative to the mean (in parentheses), as well as the relative errors RE in percent, defined as $\mathrm{RE}(x, y) = |(x - y)/y| \cdot 100\%$. We observe that the relative errors for the fitted HEDs are the smallest. Furthermore, a higher number of phases in the HED does not necessarily lead to a better fit when measured in terms of queueing performance. Finally, the confidence interval is the smallest when using the Weibull fit. It seems that the Weibull fit does not describe the statistical properties in the trace well; it seems to be too deterministic.

6. CONCLUSIONS AND FINAL REMARKS

In this paper we have presented a new, direct way of fitting HEDs to measurement data that describes HTDs. In comparison to a previous approach, the new method may be computationally less attractive (it has a higher complexity) but its results are far more satisfying. Furthermore, the new method does not require any intermediate distribution function (form) to be chosen.

At the same time, the proposed fitting procedure allows us to classify events (in the case study: objects), that is, a fitted HED tells us that with probability c_i, an object "belongs to" class i of which the mean length is $1/\lambda_i$. We are currently experimenting with such a classification scheme, in order to make scheduling and caching decisions in world-wide web servers.

APPENDIX A. THE MAXIMUM-LIKELIHOOD METHOD

Let $p(x|\underline{v})$ be the distribution density function of a distribution parameterized by the vector $\underline{v} = (v_1, \ldots, v_M)$. Given measurement data $x_1, \ldots, x_N$, we search a value for $\underline{v}$ (a so-called *parameter estimator*) so that the distribution with density $p(x|\underline{v})$ "best" fits the distribution of the measurement data. In the maximum-likelihood (ML) method the quality of the fitting is expressed by the *likelihood-function*:

$$L(x_1, \cdots, x_n|\underline{v}) = \prod_{i=1}^{N} p(x_i|\underline{v}). \tag{14}$$

The goal is to maximize L by choosing an appropriate value for $\underline{v}$. This can be achieved by solving the following set of equations:

$$\frac{\partial L}{\partial v_1} = 0, \cdots, \frac{\partial L}{\partial v_M} = 0, \tag{15}$$

Instead of (15), one often solves

$$\frac{\partial \log L}{\partial \log v_1} = 0, \cdots, \frac{\partial \log L}{\partial \log v_M} = 0, \tag{16}$$

since the log-function transforms the product into a sum.

Examples

- The **exponential** distribution with $p(x|\lambda) = \lambda e^{-\lambda x}$ yields

  $$L(x_1, \cdots, x_N|\lambda) = \lambda^N e^{-\lambda(\sum_{i=1}^{N} x_i)}$$

  with "maximizing" parameter estimator

  $$\lambda = \frac{N}{\sum_{i=1}^{N} x_i}.$$

- For the **Weibull** distribution with $p(x|a,b) = \frac{b}{a^b} x^{b-1} e^{-(x/a)^b}$, we find

  $$L(x_1, \ldots x_N|a,b) = (\frac{b}{a^b})^b \prod_{i=1}^{N} x_i^{b-1} e^{-\left(\sum_{i=1}^{N} (x_i/a)^b\right)}$$

  which yields

  $$a = \left(\frac{\sum_{i=1}^{N} x_i^b}{N}\right)^{1/b} \qquad \text{and} \qquad \frac{1}{b} + \frac{1}{N}\sum_{i=1}^{N} \log x_i - \frac{\sum_{i=1}^{N} (x_i)^b \log(x_i)}{\sum_{i=1}^{n} (x_i)^b} = 0.$$

  This system of equations should be solved iteratively, e.g., with the Newton iteration method.

APPENDIX B. THE EM METHOD

The EM (Expectation Maximization) method is an extension of the ML method (see Appendix A). Again measurement data $x_1, \ldots, x_N$ and a distribution to fit are given. In the EM method the density function is the weighted sum of one or more "basic" densities $p(x|\underline{\theta_i})$, i.e.,

$$p(x|\alpha) = \sum_{i=1}^{I} c_i \cdot p(x|\underline{\theta_i})$$

where $\alpha = (\underline{c}, \underline{\theta})$ is the parameter of the composed distribution with $\underline{c} = (c_1, \ldots, c_I)$, and $\underline{\theta} = (\underline{\theta_1}, \ldots, \underline{\theta_I})$. For the mixture weights c_i, we assume $c_i \geq 0$ for $i = 1, \ldots, I$ and $\sum_{i=1}^{I} c_i = 1$.

As in the ML method we search a parameter estimator α that maximizes the likelihood function $L(x_1,\ldots,x_N|\alpha)$. Alternatively, we can define the optimal estimator as the estimator $\hat{\alpha}$ that maximizes the difference $D(\alpha,\hat{\alpha}) = \log L(x_1,\ldots,x_N|\hat{\alpha}) - \log L(x_1,\ldots,x_N|\alpha)$ for any other estimator α. Developing D we obtain

$$D(\alpha,\hat{\alpha}) = \sum_{n=1}^{N} \log p(x_n|\hat{\alpha}) - \sum_{n=1}^{N} \log p(x_n|\alpha) = \sum_{n=1}^{N} \log \frac{p(x_n|\hat{\alpha})}{p(x_n|\alpha)} \tag{17}$$

$$= \sum_{n=1}^{N} \underbrace{\sum_{y} h(y|x_n,\alpha)}_{=1} \log \frac{p(x_n|\hat{\alpha})}{p(x_n|\alpha)} \tag{18}$$

In the last equation we have introduced a so-called *hidden variable* y with distribution $h(y|x_n,\alpha)$. Using y we can define a two-dimensional density function $h(x_n,y|\alpha) = p(x_n|\alpha)\cdot h(y|x_n,\alpha)$ and we obtain (by the use of the law of conditional probabilities):

$$D(\alpha,\hat{\alpha}) = \sum_{n=1}^{N}\sum_{y} h(y|x_n,\alpha) \log\left(\frac{h(x_n,y|\hat{\alpha})}{h(x_n,y|\alpha)}\frac{h(y|x_n,\alpha)}{h(y|x_n,\hat{\alpha})}\right) \tag{19}$$

$$= \sum_{n=1}^{N}\sum_{y} h(y|x_n,\alpha) \log\frac{h(x_n,y|\hat{\alpha})}{h(x_n,y|\alpha)} + \sum_{n=1}^{N}\sum_{y} h(y|x_n,\alpha) \log\frac{h(y|x_n,\alpha)}{h(y|x_n,\hat{\alpha})}. \tag{20}$$

It can be shown that the second additive term $\sum_{n=1}^{N}\sum_{y} h(y|x_n,\alpha)\log\frac{h(y|x_n,\alpha)}{h(y|x_n,\hat{\alpha})}) \geq 0$. The proof of this statement is simple: since it holds $\log(t) \leq t-1$, for $t>0$, we have:

$$\sum_{y} p(y)\cdot\frac{p(y)}{q(y)} = -\sum_{y} p(y)\cdot\log\frac{q(y)}{p(y)} \geq -\sum_{y} p(y)\left(\frac{q(y)}{p(y)}-1\right) = -\sum_{y} q(y) - \sum_{y} p(y) = 1-1 = 0. \tag{21}$$

From this it follows that:

$$D(\alpha,\hat{\alpha}) \geq \sum_{n=1}^{N}\sum_{y} h(y|x_n,\alpha)\log\frac{h(x_n,y|\hat{\alpha})}{h(x_n,y|\alpha)},$$

or, by defining $Q(\alpha,\hat{\alpha}) = \sum_{n=1}^{N}\sum_{y} h(y|x_n,\alpha)\log h(x_n,y|\hat{\alpha})$:

$$D(\alpha,\hat{\alpha}) \geq Q(\alpha,\hat{\alpha}) - Q(\alpha,\alpha). \tag{22}$$

The EM algorithm uses the following iteration scheme to find the optimal estimator α:

1. Choose an initial estimator α
2. Calculate a better estimator $\alpha' := argmax_{\hat{\alpha}}\{Q(\alpha,\hat{\alpha})\}$
3. Continue iteration with $\alpha := \alpha'$ if $|\alpha-\alpha'| > \epsilon$.

To compute α' in step 2, we first have to define the hidden variable y and its density function. We choose $y \in \{1,\ldots,I\}$ with density function

$$h(y|x_n,\alpha) = \frac{c_y \cdot p(x_n|\underline{\theta_y})}{p(x_n|\alpha)}.$$

Now, $\alpha' = (\underline{c'},\underline{\theta'})$ can be computed by solving the non-linear equation:

$$\frac{\partial Q}{\partial \underline{\theta'_i}} = \sum_{n=1}^{N} h(i|x_n,\underline{\theta_i})\frac{\delta}{\delta\underline{\theta'_i}}\log(c'_i\cdot p(x_n|\underline{\theta'_i})) = 0, \qquad i=1,\cdots,I. \tag{23}$$

Using a Lagrange multiplier, one obtains[12]

$$c_i' = \frac{1}{N}\sum_{n=1}^{N} h(i|x_n,\underline{\theta_i}) = \frac{1}{N}\sum_{n=1}^{N}\frac{c_i \cdot p(x_n|\underline{\theta_i})}{p(x_n|\alpha)}, \qquad i = 1,\cdots,I. \tag{24}$$

The following table gives an overview over some interesting distributions and their explicite solutions for $\underline{\theta'}$. We have used the results from the second row (exponential densities) to fit HEDs, other basic densities, e.g., Gauss or lognormal can be successful applied as well.

Distribution	Parameters $(\underline{\theta_i})$	$p(x_n\|\underline{\theta_i})$	Iterations
Exponential	(λ_i)	$\lambda_i e^{-\lambda_i x_n}$	$\lambda_i' = \frac{\sum_{n=1}^{N} p(i\|x_n,\underline{\theta_i})}{\sum_{n=1}^{N} p(i\|x_n,\underline{\theta_i})\cdot x_n}$
Gauss	(μ_i,σ_i)	$\frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{(x_n-\mu_i)^2}{2\sigma_i^2}}$	$\mu_i' = \frac{\sum_{n=1}^{N} p(i\|x_n,\underline{\theta_i})\cdot x_n}{\sum_{n=1}^{N} p(i\|x_n,\underline{\theta_i})}$, $\sigma'^2_i = \frac{\sum_{n=1}^{N} p(i\|x_n,\underline{\theta_i})\cdot(x_n-\mu_i')^2}{\sum_{n=1}^{N} p(i\|x_n,\underline{\theta_i})}$
Lognormal	(μ_i,σ_i)	$\frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{(\ln x_n-\mu_i)^2}{2\sigma_i^2}}$	$\mu_i' = \frac{\sum_{n=1}^{N} p(i\|x_n,\underline{\theta_i})\cdot \ln x_n}{\sum_{n=1}^{N} p(i\|x_n,\underline{\theta_i})}$, $\sigma'^2_i = \frac{\sum_{n=1}^{N} p(i\|x_n,\underline{\theta_i})\cdot(\ln x_n-\mu_i')^2}{\sum_{n=1}^{N} p(i\|x_n,\underline{\theta_i})}$

REFERENCES

1. M. E. Crovella and A. Bestavros, "Self-similarity in world wide web traffic: Evidence and possible causes," *IEEE/ACM Transactions on Networking* **5**(6), pp. 835–846, 1997.
2. W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson, "On the self-similar nature of Ethernet traffic," in *Proc. ACM SIGCOMM '93*, vol. 23 of *Computer Communications Review*, pp. 183–193, Oct. 1993.
3. V. Paxson and S. Floyd, "Wide-area traffic: The failure of Poisson modeling," *IEEE/ACM Transactions on Networking* **3**, pp. 226–244, June 1995.
4. D. Brocker, "Messung und Modellierung komplexer Verkehrsstrukturen in Hochgeschwindigkeitsnetzen," Master's thesis, RWTH-Aachen, Lehr- und Forschungsgebiet Informatik 4 (Verteilte Systeme), Germany, 1998.
5. A. Feldmann and W. Whitt, "Fitting mixtures of exponentials to long-tail distributions to analyze network performance models," *Performance Evaluation* **31**, pp. 245–258, 1998.
6. B. Friis, "Modelling long-range dependent and heavy-tailed phenomena by matrix analytic methods," in *Advances in Algorithmic Methods for Stochastic Models*, G. Latouche and P. Taylor, eds., pp. 265–278, Notable Publications, Inc., 2000.
7. R. Jain, *The Art of Computer Systems Performance Analysis — Techniques for Experimental Design, Measurement, Simulation, and Modeling*, John Wiley & Sons, 1991.
8. A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society* **39(B)**, pp. 1–38, 1977.
9. S. Asmussen and O. Nerman, "Fitting phase-type distributions via the EM algorithm," in *Symposium i Anvendt Statistik, Copenhagen*, pp. 335–346, 1991.
10. J. Dahmen, K. Beulen, and H. Ney, "A mixture density based approach for object recognition for image retrieval," 6^{th} *International RIAO Conference on Content-Based Multimedia Information Access* , pp. 1632–1647, 2000.
11. T. Ryden, "Statistical estimation for markov-modulated poisson processes and markovian arrival processes," in *Advances in Algorithmic Methods for Stochastic Models*, G. Latouche and P. Taylor, eds., pp. 329–350, Notable Publications, Inc., 2000.
12. E. G. Schukat-Talamazzini, *Automatische Spracherkennung — Grundlagen, statistische Modelle und effiziente Algorithmen*, Friedr. Vieweg & Sons, 1995.
13. Y. Linde, A. Buzo, and R. Gray, "An algorithm for vector quantizer design," *IEEE Transactions on Communications* **28**(1), pp. 84–95, 1980.
14. C. R. Cunha, A. Bestavros, and M. E. Crovella, "Characteristics of WWW Client-based Traces," Tech. Rep. TR-95-010, Computer Science Department, Boston University, 1995.
15. M. F. Arlitt and C. L. Williamson, "Internet web servers: Workload chracterization and performance implications," *IEEE/ACM Transactions on Networking* **5**, pp. 631–645, October 1997.