Fitting Heavy-Tailed Distributions via the EM-Algorithm

Ramin Sadre, Rachid El Abdouni Khayari, Boudewijn R. Haverkort
RWTH Aachen, Department of Computer Science
Laboratory for Performance Evaluation and Distributed Systems
D-52056 Aachen

July 4, 2000

Abstract

In recent years several studies have shown that network traffic has the property of *self-similarity* [6, 4]. In [4], Crovella and Bestavros have shown that network traffic that is due to World Wide Web transfers can show characteristics that are consistent with self-similarity. They argue that this can be explained by the heavy-tailedness of many of the involved distributions, i.e., it has been found that the WWW document size as well as the user think times between issueing requests are heavy-tailed. Traffic originating from `HTTP` transfers amounts to at least 85% [3] of all Internet traffic; so understanding its nature is crucial.

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 in a period of 54 days. After some preprocessing and filtering, about 17 million requests of interest have remained. We studied the sizes of the objects requested by the clients and found, as expected, that the object size distribution is heavy-tailed.

In order to synthesize WWW-request size traces, for instance for usage in a discrete-event simulation, one generally tries to fit a heavy-tailed distribution to the measured data. One often chooses Pareto or Weibull distributions, not only because they are heavy-tailed but also because they only depend on two parameters and therefore are easy to fit by well-known algorithms like the maximum-likelihood algorithm [1].

However, heavy-tailed distributions cannot be used so easily for analytical or numerical evaluation studies. Analytical methods are often restricted to the use of Poisson or other "Markovian" distributions, which are not heavy-tailed. Hence, not all important properties of the system can be considered, which generally lead to the underestimation of interesting performance measures. To overcome this problem, one tries to approximate the Weibull or Pareto distribution by analytically tractable distributions. This has been done for hyper-exponential distributions by Feldmann and Whitt (FW) [2]. The algorithm developed in [2] works best when the distribution to be fitted is known, i.e., when an explicit expression for the distribution is available. Unfortunately, our

studies have shown that neither Weibull nor Pareto distributions fit well to our measurement data. Both distributions fit the tail of the empirical 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). One may overcome this problem by choosing a hybrid distribution but this would introduce auxiliary parameters for the distribution.

Hence, we have decided to directly fit a hyper-exponential distribution to the measurement data via the EM-algorithm [1] without any intermediate stage of a heavy-tailed distribution. To evaluate the results we have compared the first three moments and the median of the fitted distribution with the corresponding ones of the empirical distribution (i.e. the measurement data). Our experiments have shown that the first moment is exactly matched by the EM-algorithm and that the relative error for the second and the third moment and for the median is below 6%.

Finally, we have studied the complexity of the EM-algorithm and the FW-algorithm assuming a hyper-exponential distribution with M phases to be fitted. The FW-algorithm is very fast — its complexity is linear in M — however, it requires explicitly an heavy-tailed distribution to fit to, of which the construction time complexity depends on N (the number of measurement samples). The EM-algorithm is an iterative algorithm where the complexity of each iteration is of order $N \cdot M$. The general deficit of 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 [5]. However, in our experiments, good results generally have been obtained after 5 to 10 iterations. Additionally, it should be noted that even for $N = 17,000,000$ and $M = 6$, one iteration takes approximately one minute only on a standard personal computer.

References

[1] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39(B):1–38, 1977.

[2] A. Feldmann and W. Whitt. Fitting mixtures of exponentials to long-tail distributions to analyze network performance models. *Performance Evaluation*, 31:245–258, 1998.

[3] A. Bestravos C.R. Cunha and M. Crovella. Characteristics of www client-based traces. Technical Report BU-CS-95-010, Computer Science Departement Boston University, 1995.

[4] M. E. Crovella and A. Bestavros. Self-similarity in world wide web traffic: Evidence and possible causes. *IEEE/ACM Transactions on Networking*, 5(6):835–846, 1997.

[5] S. Asmussen and O. Nerman. Fitting phase-type distributions via the EM algorithm. In *Symposium i Anvendt Statistik, Copenhagen*, pages 335–346, 1991.

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