Fitting Network Traffic to Phase-Type Bounds

Massieh Kordi Boroujeny, Brian L. Mark, and Yariv Ephraim
Dept. of Electrical and Computer Engineering
George Mason University
Fairfax, Virginia, U.S.A.
mkordibo@gmu.edu, bmark@gmu.edu, yephraim@gmu.edu

2020 54th Annual Conference on Information Sciences and Systems (CISS) 978-1-7281-4085-8/20/$31.00 ©2020 IEEE 10.1109/CISS48834.2020.1570617448

***Abstract*—Network traffic is difficult to characterize due to its random, bursty nature. Even if a traffic source could be fit to a stochastic model with reasonable accuracy, analysis of end-to-end network performance metrics for such traffic models is generally intractable. In prior work, an approach to characterize traffic burstiness using a bound based on the class of phase-type distributions was proposed. Such phase-type bounds could be applied in conjunction with stochastic network calculus to derive probabilistic end-to-end delay bounds for a traffic stream. In this paper, we focus on the problem of estimating a tight phase-type burstiness bound for a given traffic trace. We investigate a method based on least squares and another based on the expectation-maximization algorithm. Our numerical results compare the two approaches in the scenario of a heavy-tailed M/G/1 queue. We find that while both methods are viable approaches for deriving phase-type bounds on traffic burstiness, the least squares approach performs better, particularly when a tail limit is imposed.**[1]

***Index Terms*—stochastic network calculus, traffic burstiness bound, phase-type distribution, least squares, heavy-tailed queue.**

I. Introduction

Providing performance guarantees in communication networks is an important, yet challenging task due to the bursty nature of variable bit rate traffic streams and the difficulty of modeling their interactions within a network. To address this issue, we adopt the approach of stochastic network calculus, in which traffic burstiness bounds are used to compute end-to-end network performance bounds. In particular, we adopt the "generalized" stochastic bounded burstiness (gSBB) concept proposed in [8], [19], which is closely related to the stochastically bounded burstiness (SBB) concept proposed earlier in [14]. In the gSBB and SBB frameworks, the burstiness of a given traffic source is bounded by functions satisfying certain properties. The traffic burstiness bounds can be propagated through the network via stochastic network calculus theorems, which result in end-to-end network performance bounds.

In earlier work [9], [10], we proposed the use of a particular class of bounding functions for applying the gSBB and SBB traffic burstiness bounds based on phase-type distributions. The class of phase-type distributions has the important property of being dense in the family of all distributions with non-negative support. This property implies, in theory, that any given traffic source can be bounded arbitrarily tightly with a phase-type bound. In practice, there is a trade-off between computational complexity of the phase-type model and tightness of the phase-type bound. The tightness of the bound will also depend on the characteristics of the particular traffic source and the amount of traffic data available.

In the present paper, we develop and investigate methods for estimating phase-type bounds for a given traffic stream. We develop a method to estimate a phase-type bound for a given traffic source using a least squares criterion. We compare the least squares approach with another method proposed in [10] based on the expectation-maximization (EM) algorithm. In both cases, the bound is obtained by offering the traffic stream to a constant service rate queue and fitting the observed virtual workload distribution to a phase-type distribution. We consider special cases of the phase-type distribution, including 1) a mixture of exponential distributions, also known as a hyper-exponential distribution; 2) a mixture of Erlang distributions, also known as a hyper-Erlang distribution; and 3) an acyclic canonical form of the phase-type distribution. We demonstrate the various methods in a case study involving a heavy-tailed queue. For this system, the least squares approach achieves the tightest phase-type bound when the underlying phase-type distribution has the form of a hyperexponential distribution.

The remainder of the paper is organized as follows. In Section II, we review the phase-type bounds developed in [10]. In Section III, we highlight properties of phase-type distributions that we use to estimate phase-type bounds. In Section IV, we develop a direct method based on a least squares criterion to estimate a phase-type bound for a given traffic source. In Section V, we provide a numerical example demonstrating the proposed method to an M/G/1 heavy-tailed queue. Concluding remarks are provided in Section VI.

II. Phase-Type Bounded Traffic

The concept of phase-type bounded traffic is defined as follows [10].

Definition 1. A traffic process $R(t)$ is *phase-type bounded* with upper rate ρ and bounding parameter $(A, \boldsymbol{\pi}, \mathbf{Q}, T)$ if

$$\mathsf{P}\{W(t) \geq \sigma\} \leq A\boldsymbol{\pi}e^{\mathbf{Q}\sigma}\mathbf{1}, \tag{1}$$

for all $t \geq 0$ and all $\sigma \in (0, T]$. Here, $\mathbf{1}$ is a column vector of all ones of appropriate dimension, $A \geq 0$, $T > 0$, and $W(t)$

[1]This work was supported in part by the U.S. National Science Foundation under Grant CCF-1717033.

978-1-7281-4085-8/20/$31.00 ©2020 IEEE

is the virtual workload of a constant rate queue with service rate ρ and input traffic R, defined by

$$W(t) := \max_{0 \le s \le t} \left\{ \int_s^t R(\tau) d\tau - \rho(t-s) \right\}, \qquad (2)$$

and $(\boldsymbol{\pi}, \mathbf{Q})$ represents the parameter of a phase-type distribution (see Section III-A). The integral in (2) represents the amount of traffic arriving to the queue in the interval $(s, t]$.

The phase-type traffic bound is a particular case of generalized stochastically bounded burstiness (gSBB), which was developed in [8], [19] and is defined as follows.

Definition 2 (gSBB). A traffic process R is said to have generalized stochastically bounded burstiness (gSSB) with upper rate ρ and bounding function $f(\sigma) \in \mathcal{BF}$ if, for all $t \ge 0$ and all $\sigma \ge 0$,

$$\mathsf{P}\{W(t) \ge \sigma\} \le f(\sigma), \qquad (3)$$

where $\mathcal{BF}$ is defined as the family of positive, non-increasing functions.

In [10], it was shown that the phase-type bound defined above is closed with respect to stochastic network calculus theorems based on the gSBB concept. We also refer to the phase-type traffic bound as *tail-limited gPHBB* because it may be considered a special case of gSBB with an additional constraint on the length of the tail of the distribution to be considered. The concept of gSBB is closely related to the Stochastically Bounded Burstiness (SBB) concept introduced in [14], which in turn generalizes the Exponentially Bounded Burstiness (EBB) concept proposed earlier in [18].

III. Phase-type Distribution

In this section, we briefly review the phase-type distribution and two special forms, namely, the hyperexponential distribution and the canonical form 1, which are used in Section IV.

A. Definition

The phase-type distribution is defined in terms of a Markov chain $X = \{X(t) : t \ge 0\}$ with state space $E = \{1, 2, \ldots, M, M+1\}$, where states $1, 2, \ldots, M$ are transient states and $M+1$ is an absorbing state. The generator of X has the form [1]

$$\begin{pmatrix} \mathbf{Q} & \mathbf{q} \\ \mathbf{0} & 0 \end{pmatrix}, \qquad (4)$$

where $\mathbf{Q} = [q_{ij} : i, j = 1, \ldots, M]$ is an $n \times n$ matrix such that q_{ij} is the transition rate from state i to state j and $\mathbf{q} = -\mathbf{Q}\mathbf{1}$ is an $M \times 1$ column vector. Define $\pi_i = \mathsf{P}(X(0) = i)$ for $i = 1, \ldots, M+1$ and the vector $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_M)$. Hence, the initial distribution of X is given by $(\boldsymbol{\pi}, \pi_{M+1})$, where π_{M+1} is the probability that the chain starts in the absorbing state. Let $\tau := \inf\{t \ge 0 : X(t) = M+1\}$ be the time until absorption of the Markov process X. The random variable τ is said to be phase-type with parameter $(\boldsymbol{\pi}, \mathbf{Q})$. In this case,

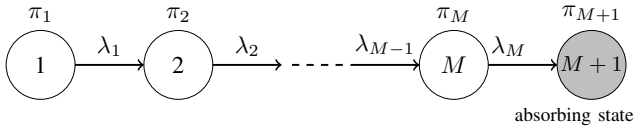

Fig. 1. Canonical Form 1(CF1) of an acyclic phase-type random variable

the cumulative distribution function and survival function of τ are given, respectively, by

$$F(\sigma) = 1 - \boldsymbol{\pi} e^{\mathbf{Q}\sigma} \mathbf{1}, \qquad (5)$$

$$S(\sigma) = \mathsf{P}(\tau > \sigma) = 1 - F(\sigma) = \boldsymbol{\pi} e^{\mathbf{Q}\sigma} \mathbf{1}, \quad \sigma \ge 0. \qquad (6)$$

The parameter of a phase-type distribution consists of $M^2 + M$ independent scalar components. The class of phase-type distributions has the important property of being dense in the family of distributions of non-negative random variables; i.e., the distribution of any random variable taking values in $[0, \infty)$ can be approximated arbitrarily closely by a phase-type distribution [17, Theorem 5.2]. In addition, phase-type distributions are mathematically tractable and form a closed set with respect to operations such as convolutions or mixtures.

B. Hyperexponential distribution

The family of phase-type distributions includes mixtures of Erlang distributions, also known as hyper-Erlang distributions. Although a special case of the phase-type, the family of hyper-Erlang distributions is also dense in the set of all distributions with non-negative support. The class of hyper-Erlang distributions in turn contains the class of mixtures of exponential distributions, also known as hyperexponential distributions. Although the class of hyperexponential distributions does not have the denseness property of the class of phase-type distributions or the class of hyper-Erlang distributions, investigations in [5] have shown that monotonically decreasing densities can be well-represented using hyperexponential distributions. In the context of phase-type traffic bounds, the density of the virtual workload is typically monotonically decreasing. Thus, in this paper we focus on the simpler class of hyperexponential distributions rather than the class of hyper-Erlang distributions.

A hyperexponential distribution with M mixture components is parameterized by the exponential rates given by $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_M)$ and the mixture probabilities given by $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_M)$. Thus, the parameter of a hyperexponential distribution with M exponential mixture components consists of $2M$ parameters. In this case, the survival function is given by

$$S_{\mathrm{he}}(\sigma; \boldsymbol{\pi}, \boldsymbol{\lambda}) = \sum_{i=1}^{M} \pi_i e^{-\lambda_i \sigma}, \quad \sigma \ge 0, \qquad (7)$$

where the dependence on the parameter $(\boldsymbol{\pi}, \boldsymbol{\lambda})$ is shown explicitly in the notation on the left-hand side.

C. Canonical form 1 distribution

By restricting $\mathbf{Q}$ to an upper-triangular matrix, a so-called acyclic phase-type distribution is obtained. Like the general

phase-type distributions and the hyper-Erlang distributions, the family of acyclic phase-type distributions is also dense in the set of distributions with non-negative support [4]. In [4] it is shown every acyclic phase-type random variable can be represented in *canonical form 1* (CF1) such that the associated generating Markov chain can be represented as in Fig. 1, where

$$\pi_i \geq 0, \quad i = 1, 2, \ldots, M+1; \qquad \sum_{i=1}^{M+1} \pi_i = 1$$
$$\lambda_M \geq \lambda_{M-1} \geq \ldots \geq \lambda_2 \geq \lambda_1 \geq 0, \tag{8}$$

and the generator matrix has the form

$$\mathbf{Q} = \begin{bmatrix} -\lambda_1 & \lambda_1 & 0 & \ldots & 0 \\ 0 & -\lambda_2 & \lambda_2 & \ldots & 0 \\ \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & -\lambda_{M-1} & \lambda_{M-1} \\ 0 & 0 & \ldots & 0 & -\lambda_M \end{bmatrix} \tag{9}$$

when there is no mass at absorbing state $\pi_{M+1} = 0$. If there is no probability mass at absorbing state, this representation of the CF1 phase-type distribution consists of $2M$ parameter components. the Laplace transform of the cumulative distribution function of the CF1 random variable is given by [2]

$$M_\tau(s) = \pi_{M+1} + \frac{N(s)}{D(s)}, \quad D(s) = \prod_{i=1}^{M}(s + \lambda_i), \tag{10}$$

The following relation is useful for computing the matrix exponential [12]:

$$e^{\mathbf{Q}} = \mathbf{V} e^{\mathbf{D}} \mathbf{V}^{-1}, \tag{11}$$

where $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m]$ is the eigenvector matrix of $\mathbf{Q}$ and $\mathbf{D} = \mathrm{diag}\{d_1, d_2, \ldots, d_M\}$ is the diagonal matrix of the corresponding eigenvalues. If the λ_i's are all distinct ($i = 1, 2, \ldots, M$), then it can be shown that eigenvalues of $\mathbf{Q}$ are given by $\{-\lambda_1, -\lambda_2, \ldots, -\lambda_M\}$. Therefore we have $\mathbf{D} = \mathrm{diag}\{-\boldsymbol{\lambda}\}$. The assumption of distinct λ_i's is not a major limitation of the CF1 phase-type distribution. Therefore, the survival function in (6) can be written as

$$S_{\mathrm{cf}}(\sigma; \boldsymbol{\pi}, \boldsymbol{\lambda}) = \boldsymbol{\pi} e^{\mathbf{Q}\sigma} \mathbf{1} = \boldsymbol{\pi} \mathbf{V} e^{\mathbf{D}t} \mathbf{V}^{-1} \mathbf{1}, \quad \sigma \geq 0. \tag{12}$$

IV. Least Squares Method

In this section, we formulate an optimization method to fit a given traffic trace to a general phase-type distribution, a hyperexponential distribution, and a CF1 distribution. The method is formulated to obtain an upper bound on the tail of the empirical traffic distribution while minimizing the squared error. Hence, this approach is referred to as a least squares method. Our approach leverages the interpretation of $W(t)$ in Definition 1 (phase-type bounded traffic) as the virtual workload in a constant rate server queue.

A. General phase-type distribution

In Definition 1, the tail probability $P\{W(t) \geq \sigma\}$ is upper bounded by a scalar multiple of the survival function, $S(t)$, of a general-phase type distribution, as given in (6). The error of the bound is given by

$$g(\sigma; A, \boldsymbol{\pi}, \mathbf{Q}) = A\boldsymbol{\pi} e^{\mathbf{Q}\sigma} \mathbf{1} - P\{W(t) \geq \sigma\}, \tag{13}$$

for $\sigma \in (0, T]$. We can assume that $\boldsymbol{\pi}_{M+1} = 0$, since no bound is imposed on $P\{W(t) = 0\}$; hence, $\boldsymbol{\pi}\mathbf{1} = 1$. Further, by defining $\boldsymbol{\alpha} = A\boldsymbol{\pi}$, the error can be re-parameterized as

$$g(\sigma; \boldsymbol{\alpha}, \mathbf{Q}) = \boldsymbol{\alpha} e^{\mathbf{Q}\sigma} \mathbf{1} - P\{W(t) \geq \sigma\}, \tag{14}$$

for $\sigma \in (0, T]$. The squared error over the interval $(0, T]$, normalized by the length of the interval T, can be written as

$$h(\boldsymbol{\alpha}, \mathbf{Q}) := \frac{1}{T} \int_0^T g^2(\sigma; \boldsymbol{\alpha}, \mathbf{Q}) \, d\sigma. \tag{15}$$

We assume the tail limit T and the number of elements M in the probability vector $\boldsymbol{\pi}$ are given. The problem of finding phase-type bound parameter $(A, \boldsymbol{\pi}, \mathbf{Q}, T)$ for a given $T > 0$ to fit a given traffic trace is then formulated as the following semi-infinitely constrained optimization problem:

$$\begin{aligned} &\min_{\boldsymbol{\alpha}, \mathbf{Q}} h(\boldsymbol{\alpha}, \mathbf{Q}) \\ &\text{subject to :} \\ &\qquad g(\sigma; \boldsymbol{\alpha}, \mathbf{Q}) \geq 0, \quad \forall \sigma \in (0, T], \\ &\qquad \boldsymbol{\alpha} \geq \mathbf{0}, \end{aligned} \tag{16}$$

where $\mathbf{Q}$ is constrained to be the transition rate matrix parameter of a phase-type distribution. We then set $A = \boldsymbol{\alpha}\mathbf{1}$ and $\boldsymbol{\pi} = \boldsymbol{\alpha}/A$. Problem (16) is difficult to solve for a general phase-type distribution. Thus, we consider special cases, in particular, the hyperexponential and acyclic phase-type distributions.

B. Hyperexponential distribution

For the hyperexponential distribution, the survival function is given by (7). The error in the phase-type bound for the hyperexponential distribution can be written as

$$g_{\mathrm{he}}(\sigma; \boldsymbol{\alpha}, \boldsymbol{\lambda}) = \sum_{i=1}^{M} \alpha_i e^{-\lambda_i \sigma} - P\{W(t) \geq \sigma\}, \tag{17}$$

for $\sigma \in (0, T]$ and the corresponding average squared error is given by

$$h_{\mathrm{he}}(\boldsymbol{\alpha}, \boldsymbol{\lambda}) := \frac{1}{T} \int_0^T g_{\mathrm{he}}^2(\sigma; \boldsymbol{\alpha}, \boldsymbol{\lambda}) \, d\sigma. \tag{18}$$

For the hyperexponential distribution, Problem (16) reduces to

$$\begin{aligned} &\min_{\boldsymbol{\alpha}, \boldsymbol{\lambda}} h_{\mathrm{he}}(\boldsymbol{\alpha}, \boldsymbol{\lambda}) \\ &\text{subject to :} \\ &\qquad g_{\mathrm{he}}(\sigma; \boldsymbol{\alpha}, \boldsymbol{\lambda}) \geq 0, \quad \forall \sigma \in (0, T], \\ &\qquad \boldsymbol{\alpha} \geq \mathbf{0}, \quad \boldsymbol{\lambda} > \mathbf{0}. \end{aligned} \tag{19}$$

As in the general case, we set $A = \boldsymbol{\alpha}\mathbf{1}$ and $\boldsymbol{\pi} = \boldsymbol{\alpha}/A$.

C. CF1 phase-type distribution

For the CF1 phase-type random variable the optimization problem formulation is given by

$$\begin{aligned}&\min_{\boldsymbol{\alpha},\boldsymbol{\lambda}} h_{\rm cf}(\boldsymbol{\alpha},\boldsymbol{\lambda})\\ &\text{subject to :}\\ &\quad g_{\rm cf}(\sigma;\boldsymbol{\alpha},\boldsymbol{\lambda})\geq 0,\quad \forall\sigma\in(0,T],\\ &\quad \boldsymbol{\alpha}\geq\mathbf{0},\quad \lambda_M>\lambda_{M-1}>\cdots>\lambda_1>0,\end{aligned}\tag{20}$$

where

$$g_{\rm cf}(\sigma;\boldsymbol{\alpha},\boldsymbol{\lambda})=\boldsymbol{\alpha}\mathbf{V}e^{\mathbf{D}\sigma}\mathbf{V}^{-1}\mathbf{1}-\mathsf{P}\{W(t)\geq\sigma\},\tag{21}$$

and

$$h_{\rm cf}(\boldsymbol{\alpha},\boldsymbol{\lambda}):=\frac{1}{T}\int_0^T g_{\rm cf}^2(\sigma;\boldsymbol{\alpha},\boldsymbol{\lambda})\;d\sigma.\tag{22}$$

D. Numerical optimization method

Problems (19) and (20) are semi-infinitely constrained optimization problems. By relaxing the upper bound constraint on the interval $\sigma\in(0,T]$, these formulations become simpler constrained optimization problems. The constrained optimization formulation for the hyperexponential distribution is simply

$$\min_{\boldsymbol{\alpha}\geq\mathbf{0},\boldsymbol{\lambda}>\mathbf{0}}\;h_{\rm he}(\boldsymbol{\alpha},\boldsymbol{\lambda}).\tag{23}$$

For the CF1 distribution, the constrained optimization formulation is

$$\min_{\substack{\boldsymbol{\alpha}\geq\mathbf{0}\\ \lambda_M>\cdots>\lambda_1>0}}\;h_{\rm cf}(\boldsymbol{\alpha},\boldsymbol{\lambda}).\tag{24}$$

A solution obtained from the solving the simpler constrained optimization formulation can be made into an upper bound by multiplying $\boldsymbol{\alpha}$ by a scalar $a>1$. Then the parameter $(a\boldsymbol{\alpha},\boldsymbol{\lambda})$ can be applied as the initial point for the semi-infinitely constrained formulation. Alternatively, the solution $(a\boldsymbol{\alpha},\boldsymbol{\lambda})$ may itself be used directly if it gives a sufficiently tight bound.

In our numerical studies we have used the `fmincon` function in MATLAB [11] for constrained optimization, which is based on an interior-point algorithm [16], and `fseminf` in MATLAB for semi-infinite constrained optimization. In [7], a problem formulation similar to Problem (20) without the upper bound constraint was used for phase-type fitting, but their optimization method was based on the Frank-Wolfe algorithm [6].

We can improve the speed and accuracy of the optimization algorithms by providing the partial derivatives of the objective function with respect to the parameter components. For the CF1 phase-type and hyperexponential distributions, the derivative is relatively simple. For the hyperexponential distribution, we have

$$\frac{\partial h_{\rm he}(\boldsymbol{\alpha},\boldsymbol{\lambda})}{\partial\alpha_i}=\frac{2}{T}\int_0^T e^{-\lambda_i\sigma}g_{\rm he}(\sigma;\boldsymbol{\alpha},\boldsymbol{\lambda})\;\mathrm{d}\sigma,\tag{25}$$

$$\frac{\partial h_{\rm he}(\boldsymbol{\alpha},\boldsymbol{\lambda})}{\partial\lambda_i}=-\frac{2}{T}\int_0^T \alpha_i\sigma e^{-\lambda_i\sigma}g_{\rm he}(\sigma;\boldsymbol{\alpha},\boldsymbol{\lambda})\;\mathrm{d}\sigma,\tag{26}$$

Algorithm 1 Least squares method for hyperexponential.

Input: Input traffic; ρ, T, M; threshold ϵ
Input: Initial point for $(\boldsymbol{\alpha},\boldsymbol{\lambda})$
Output: $A,\boldsymbol{\pi},\boldsymbol{\lambda}$

1: Feed traffic stream to a server with rate ρ and compute empirical $\mathsf{P}\{W(t)\geq\sigma\}$ for $\sigma\in(0,T]$.
2: Compute partial derivatives (25)–(26).
3: Compute $(\boldsymbol{\alpha},\boldsymbol{\lambda})$ by solving (23) using `fmincon`.
4: Find the smallest $a>1$ such that

$$g_{\rm he}(\sigma;a\boldsymbol{\alpha},\boldsymbol{\lambda})\geq 0\quad\forall\sigma\in(0,T].$$

5: $\boldsymbol{\alpha}\leftarrow a\boldsymbol{\alpha}$
6: **if** $h_{\rm he}(\boldsymbol{\alpha},\boldsymbol{\lambda})>\epsilon$ **then**
7: Compute $(\boldsymbol{\alpha},\boldsymbol{\lambda})$ by solving (19) using `fseminf`.
8: **end if**
9: $A\leftarrow\boldsymbol{\alpha}\mathbf{1}$; $\boldsymbol{\pi}\leftarrow\boldsymbol{\alpha}/A$
10: **return** A, $\boldsymbol{\pi},\boldsymbol{\lambda}$

for $i=1,\ldots,M$. For the CF1 phase-type distribution, we have

$$\frac{\partial h_{\rm cf}(\boldsymbol{\alpha},\boldsymbol{\lambda})}{\partial\alpha_i}=\frac{2}{T}\int_0^T \mathbf{e}_i\mathbf{V}e^{\mathbf{D}\sigma}\mathbf{V}^{-1}g_{\rm cf}(\sigma;\boldsymbol{\alpha},\boldsymbol{\lambda})\mathrm{d}\sigma,\tag{27}$$

where $\mathbf{e}_i$ is a M-element row vector with $\mathbf{e}_i(j)=0$ for $j=1,2,\ldots,M,\ j\neq i$ and $\mathbf{e}_i(i)=1$. Similarly, we have

$$\frac{\partial h_{\rm cf}(\boldsymbol{\alpha},\boldsymbol{\lambda})}{\partial\lambda_i}=-\frac{2}{T}\int_0^T \sigma e^{-\lambda_i\sigma}\boldsymbol{\alpha}\mathbf{V}\mathbf{e}_i^T\mathbf{e}_i\mathbf{V}^{-1}\mathbf{1}g_{\rm cf}(\sigma;\boldsymbol{\alpha},\boldsymbol{\lambda})\mathrm{d}\sigma.\tag{28}$$

In computing the squared error, the integration need not be calculated using an equidistant set of points of the interval $(0,T]$. In fact, the survival probability densities encountered in communication networks mostly have almost fixed-slope linear tails for large values of σ. Therefore, the integration can be done on a logarithmic-scale equally-distant set of points on $[0,T]$, with a much smaller number of points.

The least squares method for the hyperexponential distribution is summarized in Algorithm 1. In Steps 3 and 7, the current value of $(\boldsymbol{\alpha},\boldsymbol{\lambda})$ is used as the initial point for executing the MATLAB functions `fmincon` and `fseminf`, respectively. In Step 6, the average squared error $h_{\rm he}(\boldsymbol{\alpha},\boldsymbol{\lambda})$ is compared against a threshold ϵ. If the threshold is exceeded, a better parameter $(\boldsymbol{\alpha},\boldsymbol{\lambda})$ is obtained by solving the semi-infinitely constrained optimization problem (19) using `fseminf` with the current value of $(\boldsymbol{\alpha},\boldsymbol{\lambda})$ as the initial point. We omit a formal description of the least squares method for the CF1 distribution, since it is very similar to Algorithm 1.

V. Case Study

In this section we investigate the performance of the least squares method of Section IV in characterizing a traffic stream with Poisson process arrivals and heavy-tailed packet lengths, denoted as M/G/1 heavy-tailed traffic. This traffic model was used previously in [10] to characterize the traffic using an EM-based algorithm. We compare the results derived by least

squares method with the result derived using the EM-based algorithm.

A. M/G/1 heavy-tailed queue

We adopt the model of the heavy-tailed M/G/1 queue in [3]. In this model, packets arrive to the queue according to a Poisson process with rate λ. The service time, denoted by $\tau_{\boldsymbol{\theta}}$, depends on a gamma-distributed random variable $\boldsymbol{\theta}$. The conditional probability density function of $\tau_{\boldsymbol{\theta}}$ given $\boldsymbol{\theta} = \theta$ is

$$\mathsf{P}\{\tau_{\boldsymbol{\theta}} < t \mid \boldsymbol{\theta} = \theta\} = 1 - \delta \left(\frac{\theta}{\theta + t}\right)^{v}, \tag{29}$$

where $1 < v < 2$, $0 < \delta \leq 1$, and the density of $\boldsymbol{\theta}$ is given by

$$f_{\boldsymbol{\theta}}(\theta) = \frac{s^{2-v}}{\Gamma(2-v)} \theta^{1-v} e^{-s\theta}, \tag{30}$$

where $s > 0$ is a constant and $\Gamma(\cdot)$ is the gamma function. For this service time random variable, $\tau_{\boldsymbol{\theta}}$, we have

$$\beta := E\{\tau_{\boldsymbol{\theta}}\} = \frac{2-v}{v-1}\frac{\delta}{s} \tag{31}$$

Thus, the utilization factor of the queue is given by $\rho = \lambda\beta$. For stability of the queue we must have $\rho < 1$. For the particular case $v = 3/2$, the cumulative distribution function of $\tau_{\boldsymbol{\theta}}$ is shown in [3] to have the form

$$\mathsf{P}\{\tau_{\boldsymbol{\theta}} \leq t\} = 1 + \delta \left[\frac{2\sqrt{st}}{\sqrt{\pi}} - (1+2st)e^{st}\mathrm{erfc}(\sqrt{st})\right], \tag{32}$$

where the complementary error function is defined by

$$\mathrm{erfc}(x) = \frac{2}{\sqrt{\pi}} \int_{x}^{\infty} e^{-u^2} \mathrm{d}u. \tag{33}$$

We consider the packet lengths as being equal to the service times in the M/G/1 queue with constant service rate 1. Then the virtual workload at the time of a packet arrival will be equal to the waiting time of the packet (including its own service time) in the adopted M/G/1 queue. Let $W(t)$ denote the virtual workload of the queue, as defined in (2) and let t_n denote the arrival time of the nth packet to the queue, $n = 1, 2, \ldots$. Then $W(t_n)$ is equivalent to the waiting time of the nth packet arriving to the queue. The distribution of the stationary waiting time

$$W := \lim_{n\to\infty} W(t_n)$$

for the M/G/1 queue (with $v = 3/2$) is given in [3, Eq. (1.8)] as follows:

$$\begin{aligned}\mathsf{P}\{W \leq \sigma\} = 1 &- \frac{1+\sqrt{\rho}}{2}\sqrt{\rho}e^{(1-\sqrt{\rho})^2 s\sigma} \cdot \mathrm{erfc}\left[(1-\sqrt{\rho})\sqrt{s\sigma}\right] \\ &+ \frac{1-\sqrt{\rho}}{2}\sqrt{\rho}e^{(1-\sqrt{\rho})^2 s\sigma} \cdot \mathrm{erfc}\left[(1+\sqrt{\rho})\sqrt{s\sigma}\right]. \end{aligned} \tag{34}$$

The result (34) applies to the waiting time in equilibrium or steady-state. On the other hand, for a traffic stream to be phase-type bounded according to Definition 1 the bound should be valid for all $t \geq 0$. We next argue that if the bound applies to the steady-state waiting time distribution, then it applies to the virual workload distribution for all $t \geq 0$. A random variable X is *stochastically smaller* than random variable Y, denoted as $X \leq_{st} Y$, if $\mathsf{P}\{X > x\} \leq \mathsf{P}\{Y > x\}$ for all x [13]. According to the following theorem, the waiting times of a sequence of customers arriving to a GI/G/1 queue are stochastically monotonically increasing.

Theorem 1. (see [15, Theorem 5.1.1]) Consider a GI/G/1 queue such that the stationary waiting time distribution exists. If $W(0) \leq_{st} W(t_1)$, then

$$W(t_1) \leq_{st} W(t_2) \leq_{st} \ldots \leq_{st} W. \tag{35}$$

TABLE I
PHASE-TYPE BOUND PERFORMANCE FOR M/G/1 HEAVY-TAILED TRAFFIC

	Method	Obj. func. value($\times T$)	Log-likelihood($\times 10^7$)	M
1	**LS(he): Prob.** (23)	0.2676	-3.3434	30
2	**LS(he): Prob.** (19)	0.2051	-3.3434	30
3	**LS(CF1): Prob.** (24)	0.9991	-3.3814	5
4	**LS(CF1) Prob.** (20)	0.9975	-3.3813	5
5	**EM(hyper-Erlang) [10]**	0.5351	-3.2277	30

In all of the cases we have considered, the traffic is stationary and ergodic and the queue is stable since the utilization factor $\rho < 1$. This guarantees existence of a stationary waiting time distribution, such that Theorem 1 applies. In this case, (35) implies that $W(t) \leq_{st} W(t_n)$ for all $t \leq t_n$ and $n \geq 1$, and hence, $W(t) \leq_{st} W$ for all $t \geq 0$. Therefore, if $\mathsf{P}\{W \geq \sigma\}$ is bounded by a function $f(\sigma) = A\boldsymbol{\pi}e^{\mathbf{Q}\sigma}\mathbf{1}$, the same bound applies to $\mathsf{P}\{W(t) \geq \sigma\}$ for all $t \geq 0$. A similar argument was used in [8] to justify the gSBB bound in a discrete-time setting.

B. Numerical example

Next, we characterize the M/G/1 heavy-tailed traffic described in the previous section by hyperexponential and CF1 phase-type bounds using the least squares (LS) method. We compare the results with phase-type bounds obtained using the EM-based algorithm proposed in [10].

In our case study we assume $v = 1.5$ and set $s = \delta = 1$ and $\lambda = 0.5$. Then, $\beta = 1$ according to (31) and $\rho = 0.5$, which implies that $\mathsf{P}\{W = 0\} = 1 - \rho = 0.5$. In this example we have chosen $T = 5 \times 10^6$. In applying the EM-based method, 10^7 samples were generated using (34). We have compared the objective function values resulting from the different methods. The integration required to compute the squared error was done using 10^4 equidistant points on a logarithmic scale in the interval $(0, T]$.

In Table I, the squared error and log-likelihood values obtained using five different methods are shown. Methods 1 and 3 involve constrained optimization only followed by multiplication of the vector $\boldsymbol{\alpha}$ by a scalar $a > 1$ such that an upper bound is obtained. Method 2 corresponds to Algorithm 1, while Method 4 is a similar algorithm using the CF1 phase-type distribution. In estimating phase-type bounds with the proposed least squares method, we have

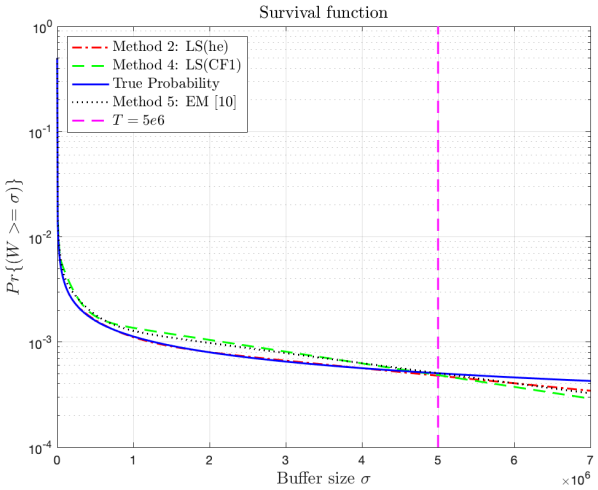

Fig. 2. Phase-type bounds and tail probability for M/G/1 heavy-tailed traffic.

used hyperexponential distributions with $M = 30$ mixture components (Methods 1 and 2) and CF1 phase-type with $M = 5$ (Methods 3 and 4). Further improvement of the bound was not obtained by increasing the order of the CF1 phase-type distribution. Method 5 is an EM-based approach proposed in [10] using a hyper-Erlang distribution with $M = 30$ Erlang mixture components. The EM-based algorithm results in the special case of a hyperexponential distribution. This result is expected, since a distribution with a monotonically decreasing probability density function, as in our case, can be well approximated using a mixture of exponentials [5].

As can be seen in Table I, for the least squares approaches, going from the constrained optimization to the semi-infinitely constrained formulation results in a tighter bound, as expected. The best result is achieved by using Method 2. The bound derived by the EM-based algorithm (Method 5), results in the highest likelihood, which is expected, since the EM algorithm aims to maximize the likelihood of the samples. However, Method 5 results in a looser bound on the tail probability of the virtual workload. Note that the EM approach does not take into account the tail limit T.

The (empirical) survival function of the virtual workload corresponds to the curve labelled "true probability" in Fig. 2. Here, the stationary virtual workload distribution is heavy-tailed and therefore cannot be bounded for all σ by a bound in the form of a phase-type survival function. Nevertheless, we can bound this distribution by a phase-type bound with a finite number of phases over a time interval $[0, T]$. In practice, the virtual workload cannot grow without bound, since physical buffers are always of finite size. The phase-type bounds obtained using Methods 2, 4, and 5 are shown in Fig. 2. Over the interval $(0, T]$, Method 2 clearly results in the tightest phase-type bound.

VI. Conclusion

We developed a method for characterizing a given traffic stream in terms of phase-type bounds. This characterization can be used in conjunction with stochastic network calculus results to bound end-to-end performance measures such as delay and virtual workloads in the network. A least squares criterion was used to formulate the problem of finding phase-type bounds as an optimization problem using two special forms of the phase-type distribution: the hyperexponential and canonical form 1 distributions. The best results for our numerical example were obtained using the hyperexponential distribution. Compared to an EM-based method proposed in [10], the least squares approach produced tighter bounds. The least squares method described in this paper is an offline approach, but it could be developed into an online method by considering a sufficiently large moving window of the most recent traffic samples.

References

[1] M. Bladt and B. Nielsen, *Matrix-Exponential Distributions in Applied Probability*. New York, NY: Springer, 2017.
[2] A. Bobbio and A. Cumani, "ML estimation of the parameters of a PH distribution in triangular canonical form," *Comput. Perform. Eval.*, pp. 33–46, Jan. 1992.
[3] O. J. Boxma and J. W. Cohen, "The M/G/1 queue with heavy-tailed service time distribution," *IEEE J. Sel. Areas Commun.*, vol. 16, no. 5, pp. 749–763, Jun. 1998.
[4] A. Cumani, "On the canonical representation of homogeneous Markov processes modelling failure - time distributions," *Microelectron. Reliab.*, vol. 22, no. 3, pp. 583–602, 1982.
[5] A. Feldmann and W. Whitt, "Fitting mixtures of exponentials to long-tail distributions to analyze network performance models," in *Proc. IEEE INFOCOM*, vol. 3, Apr. 1997, pp. 1096–1104.
[6] M. Frank and P. Wolfe, "An algorithm for quadratic programming," *Nav. Res. Logist.*, vol. 3, no. 1-2, pp. 95–110, 1956.
[7] A. Horváth and M. Telek, "PhFit: A general phase-type fitting tool," in *Proc. 12th Perform. TOOLS*, Apr. 2002, pp. 82–91.
[8] Y. Jiang, Q. Yin, Y. Liu, and S. Jiang, "Fundamental calculus on generalized stochastically bounded bursty traffic for communication networks," *Comput. Netw.*, vol. 53, no. 12, pp. 2011–2021, Aug. 2009.
[9] M. Kordi Boroujeny, Y. Ephraim, and B. L. Mark, "Phase-type bounds on network performance," in *Proc. Conf. Inf. Sci. Sys. (CISS)*, Princeton, NJ, Mar. 2018, pp. 1–6.
[10] M. Kordi Boroujeny, B. L. Mark, and Y. Ephraim, "Tail-limited phase-type burstiness bounds for network traffic," in *Proc. Conf. Inf. Sci. Sys. (CISS)*, Baltimore, MD, Mar. 2019, pp. 1–6.
[11] *Optimization Toolbox User's Guide*, R2018b, The MathWorks Inc., Natick, MA, 2018.
[12] C. Moler and C. Van Loan, "Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later," *SIAM Review*, vol. 45, no. 1, pp. 3–49, 2003.
[13] S. Ross, *Stochastic Processes*, 2nd ed. John Wiley & Sons, 1996.
[14] D. Starobinski and M. Sidi, "Stochastically bounded burstiness for communication networks," *IEEE Trans. Inf. Theory*, vol. 46, no. 1, pp. 206–212, Jan. 2000.
[15] D. Stoyan, *Comparison Methods for Queues and Other Stochastic Models*. John Wiley & Sons, 1983.
[16] R. Waltz, J. Morales, J. Nocedal, and D. Orban, "An interior algorithm for nonlinear optimization that combines line search and trust region steps," *Math. Program.*, vol. 107, no. 3, pp. 391–408, Jul. 2006.
[17] R. W. Wolff, *Stochastic Modeling and the Theory of Queues*. New Jersey: Prentice-Hall, 1989.
[18] O. Yaron and M. Sidi, "Performance and stability of communication networks via robust exponential bounds," *IEEE/ACM Trans. Netw.*, vol. 1, no. 3, pp. 372–385, Jun. 1993.
[19] Q. Yin, Y. Jiang, S. Jiang, and P. Y. Kong, "Analysis on generalized stochastically bounded bursty traffic for communication networks," in *Proc. IEEE Local Comput. Netw. (LCN)*, Nov. 2002, pp. 141–149.