Accelerating Performance Inference over Closed Systems by Asymptotic Methods

Giuliano Casale*
Department of Computing
Imperial College London, UK
g.casale@imperial.ac.uk

ABSTRACT

Recent years have seen a rapid growth of interest in exploiting monitoring data collected from enterprise applications for automated management and performance feedbacks. In spite of this trend, even simple performance inference problems involving queueing-theoretic formulas often incur computational bottlenecks, for example upon computing likelihoods in models of batch systems. Motivated by this issue, we revisit the solution of multiclass closed queueing networks, which are popular models used to describe batch and distributed applications with parallelism constraints.

We first prove that the normalizing constant of the equilibrium state probabilities of a closed model can be reformulated in an exact manner as a multidimensional integral over the unit simplex. This gives as a by-product the first exact expressions for the multiclass normalizing constant that are both tractable and explicit. We then derive a novel method based on cubature rules to efficiently evaluate the proposed integral form in small and medium-sized models. For large models, we propose novel asymptotic expansions and Monte Carlo sampling methods to efficiently and accurately approximate normalizing constants and likelihoods. We illustrate the resulting accuracy gains in problems involving optimization-based inference.

Categories and Subject Descriptors

C.4 [**Performance of Systems**]: Modeling Techniques

Keywords

Performance inference, queueing theory, closed system

1. INTRODUCTION

During the last decade there has been a growing trend among enterprises to exploit large volumes of monitoring data for availability and performance management [18]. While activities such as capacity planning have been traditionally carried out by human experts, a large number of data-driven application performance management products are now available on the software market to support human operators in managing the lifecycle of large production systems.

A common issue faced in data centers is that the number of monitored systems tends to be large and thus human operators can supply to the application performance management software only a limited amount of information on the internals of each system. Statistical methods may therefore be used by the software to identify a good predictive model for a system. Model inference techniques commonly used in machine learning and Bayesian statistics can support this task by providing automated algorithms for performance model parameterization and model selection [37, 38]. For instance, inference methods can leverage queueing theory to represent the probabilistic relationships between some performance metrics of a distributed system. Applications of these techniques to performance engineering are several, recent examples include queueing parameter estimation in the presence of resource contention [43], inference of performance models with latent data [44], and performance design under uncertainties based on posterior distributions [17]. A comprehensive bibliography of this area can be found in [37].

In spite of this trend, we show that performance inference remains a difficult problem when considering closed models, which are often used to describe batch systems and distributed applications with parallelism limits. Inference on closed models can rely on product-form queueing network theory, since equilibrium probability formulas are known exactly [7] and such models can therefore be treated analytically. However, a major drawback with these models is that the computation of likelihoods requires to determine an expensive normalizing constant appearing in the equilibrium probabilities.

Prior work has proposed methods to exactly compute normalizing constants using recursive algorithms [10, 29, 39, 15], generating functions [9, 24], and moment-based methods [11]. Furthermore, methods based on Laplace transform inversion [13], asymptotic expansions [35, 31], and Monte Carlo integration [41] have led to inexpensive approximations for large models. In spite of these developments, we show that existing techniques are fairly inefficient when it comes to their use in inference problems. We find in particular that likelihood maximization programs are either too expensive or return largely inaccurate solutions, depending on the method used to approximate the normalizing constant. This effectively hinders the application of likelihood-based methods to closed systems.

Motivated by these observations, we revisit the computational theory of closed product-form queueing networks. We first obtain novel exact formulas that express the normalizing constant as an integral of a product of powers of linear forms over the unit simplex. This result leads to novel asymptotic expansions for the normalising constant based on Laplace's method [25]. This approach requires all nodes to have asymptotically large queues as the population size grows. We ensure this property through a novel scaling that adds

*This research has been partially funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 644869 (DICE). The author wishes to thank Tony Field for helpful discussions during preparation of this work.

ACM SIGMETRICS '17 Urbana-Champaign, IL, US
ACM ISBN 978-1-4503-2138-9.
DOI: 10.1145/1235

at every node a set of permanently residing jobs. We also obtain a novel Monte Carlo integration scheme that enables the efficient sampling of the normalizing constant to the desired precision. Lastly, we derive novel explicit solutions for the normalizing constant in terms of combinatorial summations, which are to the best of our knowledge the first tractable expressions in the multiclass setting that are also explicit. Using a large scale validation on thousands of maximum likelihood estimation problems, we demonstrate that the proposed techniques tackle the computational limitations of state-of-the-art methods.

The rest of the paper is organized as follows. Section 2 introduces the reference model for closed systems and surveys related work. Section 3 gives novel exact theoretical results concerning the solution of closed product-form networks. Section 4 develops asymptotic expansions and Monte Carlo integration methods. Section 5 investigates the accuracy of the methods. Lastly, Section 6 summarizes results and concludes the paper.

2. BACKGROUND

2.1 Notation

The reference model is a product-form closed queueing network with M nodes and R job classes [7]. Let N_r be the number of jobs in class r and let $\boldsymbol{N} = (N_1, \ldots, N_R)$, $N = \sum_r N_r$. The first $K \leq M$ nodes have a single-server and the remaining ones are infinite server nodes. Matrix $\boldsymbol{\theta} = [\theta_{kr}]$ collects the demands placed by class-r jobs at node k, i.e., the product of the mean number of visits by the mean service time for that class at the node. We denote by $\sigma_r = \sum_{k=K+1}^{M} \theta_{kr}$ the sum of the class-r demands at the infinite server nodes.

In the case of a product-form network of processor sharing and infinite server nodes with exponential service times the model maps to a Markov process with state space

$$\mathcal{S}_M = \{\boldsymbol{n} \in \mathbb{N}^{MR} \mid n_{kr} \geq 0, \textstyle\sum_{i=1}^{M} n_{ir} = N_r\}$$

where $\boldsymbol{n} = (\boldsymbol{n}_1, \ldots, \boldsymbol{n}_M)$, $\boldsymbol{n}_k = (n_{k1}, \ldots, n_{kR})$, and n_{kr} is the number of class-r jobs at node k. The equilibrium distribution for this process is

$$\pi(\boldsymbol{n}) = \frac{C(\boldsymbol{n})}{G_{\boldsymbol{\theta}}(\boldsymbol{N})} \prod_{k=1}^{M} \prod_{r=1}^{R} \theta_{kr}^{n_{kr}} \qquad \boldsymbol{n} \in \mathcal{S}_M \tag{1}$$

where $G_{\boldsymbol{\theta}}(\boldsymbol{N})$ is a normalizing constant over $\mathcal{S}_M$ and we define $C(\boldsymbol{n}) = \prod_{k=1}^{K} n_k! / \prod_{i=1}^{M} \boldsymbol{n}_i!$, where for vector $\boldsymbol{v} = (v_1, \ldots, v_n)$ we define $\boldsymbol{v}! = \prod_{j=1}^{n} v_j!$ and $v = v_1 + \ldots + v_n$. By the given definitions the normalizing constant may be written as

$$G_{\boldsymbol{\theta}}(\boldsymbol{N}) = \sum_{\boldsymbol{n} \in \mathcal{S}_M} C(\boldsymbol{n}) \prod_{k=1}^{M} \prod_{r=1}^{R} \theta_{kr}^{n_{kr}} \tag{2}$$

It is possible to show that the same expression holds also for nodes with other service disciplines and service time distributions [7].

2.2 Computational methods: state-of-the-art

Since the number of states of the model grows as $\mathcal{O}(N^{MR})$ with the job population, it is usually infeasible to obtain $G_{\boldsymbol{\theta}}(\boldsymbol{N})$ by direct summation over $\mathcal{S}_M$. Many alternative computational methods have been defined over a time span of four decades, we limit to give a high-level review, pointing to the references for details.

2.2.1 *Exact methods*

The classic exact computational methods for $G_{\boldsymbol{\theta}}(\boldsymbol{N})$ are the multiclass convolution algorithm (CA) [39] and RECAL [15], which feature respectively $\mathcal{O}(N^R)$ and $\mathcal{O}(N^K)$ time and space requirements. Such polynomial complexities quickly become intractable in applications. Other exact algorithms with similar complexities may be found in [40, 9, 20, 24]. The method of moments (MoM) [11] lowers the requirements to $\mathcal{O}(N^2 \log N)$ time and $\mathcal{O}(N \log N)$ space using a recursive system of linear equations, applicable under some regularity conditions on $\boldsymbol{\theta}$. This method becomes computationally demanding as M and R grow simultaneously and solution times in large models are of the order of minutes, and this is too expensive for optimization-based performance inference. Summarizing, exact methods for $G_{\boldsymbol{\theta}}(\boldsymbol{N})$ exist, but they are hardly applicable to performance inference problems due to their cost. Experiments illustrating these limitations are shown in Section 2.3.

2.2.2 *Approximate methods*

Approximate mean-value analysis (AMVA) algorithms [42] provide accurate estimates of mean performance measures and are $\mathcal{O}(1)$ with respect to job populations. Yet the focus on mean performance metrics is restrictive since inference algorithms typically require a probability model such as (1), for example to express prior distributions on parameters or to infer an optimal parameterization using likelihood maximization [38, 43]. AMVA does not apply in these cases as it cannot compute likelihoods.

Flow-equivalent aggregation methods are alternative approximation methods that allow to efficiently aggregate a subnetwork into a single node with state-dependent service rates [12]. Unfortunately flow-equivalent aggregation is tractable only in single-class models.

Such limitations are addressed by specialized approximations for $G_{\boldsymbol{\theta}}(\boldsymbol{N})$, which include Monte Carlo integration, numerical methods, and asymptotic expansions. Monte Carlo integration methods are first introduced in [41], based on the integral form [35]

$$G_{\boldsymbol{\theta}}(\boldsymbol{N}) = \frac{1}{\boldsymbol{N}!} \int_{\mathbb{R}_+^K} e^{-y} \prod_{r=1}^{R} \left(\sigma_r + \sum_{k=1}^{K} \theta_{kr} y_k \right)^{N_r} d\boldsymbol{y} \tag{3}$$

where $\boldsymbol{N}! = \prod_{r=1}^{R} N_r!$, $\mathbb{R}_+^K = \{\boldsymbol{y} \in \mathbb{R}^K | \boldsymbol{y} \geq \boldsymbol{0}\}$, with $\boldsymbol{0} = (0, \ldots, 0)$ and $y = \sum_k y_k$. Expression (3) is obtained by replacing the $n_k!$ terms in $C(\boldsymbol{n})$ with the integral form of the gamma function [35] and by repeated application of the multinomial theorem

$$\left(\sum_{k=1}^{K} x_k \right)^{V} = V! \sum_{\substack{\boldsymbol{v} \geq \boldsymbol{0} \\ v = V}} \prod_{i=1}^{K} \frac{x_i^{v_i}}{v_i!} \tag{4}$$

where $\boldsymbol{v} \in \mathbb{N}^K$. The integral form (3) can be efficiently evaluated using importance sampling [41], normally requiring $J = 10^5 - 10^7$ samples to approximate $G_{\boldsymbol{\theta}}(\boldsymbol{N})$ with low variance. Computing millions of samples is acceptable for evaluating individual models, but places an excessive overhead for optimization, as it is not too dissimilar from the cost of simulation. Moreover, the variance of the Monte Carlo estimators for $G_{\boldsymbol{\theta}}(\boldsymbol{N})$ can adversely affect the identification of the search direction [45].

Numerical methods for $G_{\boldsymbol{\theta}}(\boldsymbol{N})$ include Laplace transform inversion (LTI) [13] and ODE-based methods (TE) [45]. LTI allows for arbitrary approximation accuracy, but still incurs a significant computational cost. For instance, [13, p. 967] provides an example where LTI requires $\approx 10^{12}$ operations on a model with $K = 64$ and $R = 9$, which is beyond the acceptable cost for a single iteration of an optimization program. Instead the approximations proposed in this paper scale efficiently to models of this size. TE is $\mathcal{O}(1)$, but numerically delicate, since its ODEs involve the derivatives of $G_{\boldsymbol{\theta}}(\boldsymbol{N})$. The method quickly becomes difficult to apply in large models due to the rapid growth of $G_{\boldsymbol{\theta}}(\boldsymbol{N})$ that affects numerical precision.

2.2.3 *Asymptotic expansions*

Asymptotic expansions for $G_{\boldsymbol{\theta}}(\boldsymbol{N})$ are $\mathcal{O}(1)$ with respect to a growth in job populations and thus capable of accelerating optimization. Expansions appear in [35, 27, 32, 33, 31] and are discussed below. Other asymptotic methods exist but they are not relevant to the present work as they either focus only on single-class models [19] or study asymptotic values of mean-value performance metrics [2, 6, 8, 26, 28], whereas we focus here on computing likelihoods in multiclass systems.

PANACEA (PAN) [35] is applicable only to models with infinite servers and in normal usage, i.e., where resources are lightly utilized so that $\max_i \alpha_i < 1$, where $\alpha_i = \sum_r N_r \theta_{ir} \sigma_r^{-1}$, $i = 1, \ldots, K$. Normal usage conditions tend to be restrictive in applications, where the analysis of heavy-load behaviors is of practical importance.

The ray method [27] (RAY) is an approximation method for PDEs which combined with singular perturbation theory provides an approximation for $G_{\boldsymbol{\theta}}(\boldsymbol{N})$ under an increasing number of nodes and a simultaneous scaling of the service demands. We extensively compare throughout this paper against this method.

Saddle-point approximation (SPA) provides asymptotic expansions of contour integrals arising from the generating function of $G_{\boldsymbol{\theta}}(\boldsymbol{N})$. A limitation of SPA is that it is derived for small models only [32, 33, 31].

As we show in Section 4, our asymptotic expansions are based on Laplace's method [25], which may be seen as a specialization of the saddle-point method for real integrals. This differs substantially from the SPA method which applies to contour integrals in the complex domain and leads to rather different expressions for $G_{\boldsymbol{\theta}}(\boldsymbol{N})$.

2.3 Motivating example: demand estimation from state samples

To illustrate the limitations of existing computational techniques, we compare the above methods in a likelihood maximization application. Assume to measure a set of L state samples, $\boldsymbol{n}^{(l)} \in \mathcal{S}_M$, $l = 1, \ldots, L$. We seek for a maximum likelihood estimator (MLE) for the demand matrix $\boldsymbol{\theta}$. In practice, problems of this kind arise during model calibration, where one seeks for an optimal parameterization and the $\boldsymbol{n}^{(l)}$ samples represent system state measurements. Several other estimators exist [43], our goal here is to present the challenges arising in inference problems that rely on likelihoods formulated over closed systems. Likelihood-based estimation offers a number of advantages over other estimation techniques, for example it can cope with missing and aggregate data through expectation-maximization [38].

We assume $\sigma_r = 0$ and knowledge of $\theta_r = \sum_k \theta_{kr}$, $\forall r$, i.e., the end-to-end response time of a single class-r request when $N = 1$. From (1) the log-likelihood of $\boldsymbol{\theta}$ is given by

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{l=1}^{L} \log C(\boldsymbol{n}^{(l)}) + L \sum_k \sum_r \widetilde{Q}_{kr} \log \theta_{kr} - L \log G_{\boldsymbol{\theta}}(\boldsymbol{N}) \tag{5}$$

where $\widetilde{Q}_{kr} = \sum_l n_{kr}^{(l)}/L$ is the measured mean queue-length of class r at node k. The first term is neglected upon optimizing over $\boldsymbol{\theta} \geq \mathbf{0}$. Note that the cost of computing $\mathcal{L}(\boldsymbol{\theta})$ in this case is dominated by $\log G_{\boldsymbol{\theta}}(\boldsymbol{N})$.

We consider (5) for a model with $M = K = 3$, $R = 3$, $\boldsymbol{N} = (N, N, N)/3$, and $\theta_{kr} = k * r$ and seek for a (local) MLE that maximizes $\mathcal{L}(\boldsymbol{\theta})/L$ subject to $\boldsymbol{\theta} \geq \mathbf{0}$. We also let $L \to \infty$ by using in place of the $\widetilde{Q}_{kr}$ the exact mean queue-lengths computed by mean-value analysis [40]. We apply MATLAB's fmincon interior point algorithm, and calculate $G_{\boldsymbol{\theta}}(\boldsymbol{N})$ at each iteration using one among CA, RECAL, MoM, TE, or RAY. We also use Monte Carlo integration (MCI) with AMVA-based initialization [45]. The assumptions underpinning PAN and SPA are not met on this example: PAN requires infinite servers; SPA is not available for models with 3 nodes and 3 classes or larger. TE also requires infinite servers, but we can set $\sigma_r = 10^{-8}$, $\forall r$; a similar perturbation cannot be used with PAN since the method also requires normal usage. Lastly, we include in the experiment a variant of (5) where we neglect the normalizing constant by setting $\log G_{\boldsymbol{\theta}}(\boldsymbol{N}) = 0$. This variant is denoted by NOG and corresponds to the log-likelihood formula for a product-form open queueing network with demands $\boldsymbol{\theta}$. Each chosen method is initialized at the same point, sampled from a uniform distribution. We set a timeout of 10 minutes, after which fmincon returns after completing the running iteration.

Table 1: Demand estimation results. T = timeout (10 min). Runtimes are rounded up to the nearest integer. On this example, memory requirements are negligible for all methods.

$M = K = 3, R = 3$	Abs. Perc. Error (%)			Time (seconds)		
$N_r = N/R =$	*2*	*20*	*40*	*2*	*20*	*40*
CA (No timeout)	0.3	0.0	0.0	2	419	3413
CA	0.3	0.0	40.3	2	419	T
MCI3	90.9	138.9	115.1	5	5	7
MCI6	135.9	91.1	118.3	185	171	137
MoM	15.4	51.2	71.0	T	T	T
NOG	38.6	92.2	96.1	1	1	8
RAY	82.1	72.8	59.4	6	3	8
RECAL	0.3	31.7	76.5	3	T	T
TE-2	38.7	92.2	96.1	143	174	138
TE-3	35.2	91.3	95.6	T	T	T

Experiments are run on a quad-core desktop computer. Table 1 shows execution times with $N_r = N/R = 2, 20, 40$ jobs and the absolute percentage errors of the returned demands with respect to the true $\boldsymbol{\theta}$. The suffixes for MCI and TE are the number of samples J and the scale of the ODE step size τ, respectively, e.g., MCI3 has $J = 10^3$ and TE-2 has $\tau = 10^{-2}$.

The *CA (No timeout)* method is exact and thus provides an upper bound on achievable accuracy on this instance and it is obtained by running the CA exact method until termination of the optimization. This baseline is required since the problem (5) is non-convex, thus the choice of the initial point affects the achievable accuracy.

We note that all methods incur a considerable degradation of accuracy and running times as the population N grows. Some methods, such as TE, have a similar or worse performance than NOG, which ignores the normalizing constant. CA is the best among the exact methods, but its execution times grow quickly and on larger models are infeasible. Among existing approximations, the RAY asymptotic expansion achieves the best results, although the errors remain high, around 59%-82%. However, computational times are very scalable. A similar conclusion applies to MCI with a small number of samples. This motivates us to further investigate into asymptotic expansion and Monte Carlo integration methods. We remark that on this example the expansion proposed later in Section 4 achieves less than 0.7% absolute percentage error in all the three cases, with runtimes between $2s$ and $4s$. A validation on a broader set of instances is presented in Section 5.

3. EXACT RESULTS

In order to inexpensively approximate the normalizing constant, we first derive novel integral expressions for $G_{\boldsymbol{\theta}}(\boldsymbol{N})$. This derivation leads to novel numerical approximations and provides a theoretical baseline for developing asymptotic results.

3.1 Integral form over the unit simplex

We first derive an exact integral form for the normalizing constant in networks without infinite servers.

THEOREM 1. *In a multiclass closed queueing network with K single-server nodes*

$$G_{\boldsymbol{\theta}}(\boldsymbol{N}) = \frac{(N+K-1)!}{\boldsymbol{N}!} \int_{\Delta_K} \prod_{r=1}^{R} \left(\sum_{k=1}^{K} \theta_{kr} u_k \right)^{N_r} d\boldsymbol{u} \quad (6)$$

where $\Delta_K = \{\boldsymbol{u} \in \mathbb{R}^K \mid u_i \geq 0, \sum_i u_i = 1\}$ is the unit simplex.

PROOF. The multinomial theorem (4) implies that for any set of real numbers $(a_1, \ldots, a_R)$ and variables $\boldsymbol{t} = (t_1, \ldots, t_R)$

$$a_1^{N_1} a_2^{N_2} \cdots a_R^{N_R} = \frac{1}{N!} \frac{\partial^{N_1} \cdots \partial^{N_R}}{\partial t_1^{N_1} \cdots \partial t_R^{N_R}} \left(\sum_{i=1}^{R} a_i t_i \right)^N \quad (7)$$

Since $\sigma_r = 0$ we apply (7) to the product in the integrand of (3) with $a_r = \sum_k \theta_{kr} y_k$, finding after exchanging the order of differentiation and integration

$$\begin{aligned} G_{\boldsymbol{\theta}}(\boldsymbol{N}) &= \frac{1}{\boldsymbol{N}!} \frac{\partial^{N_1} \cdots \partial^{N_R}}{\partial t_1^{N_1} \cdots \partial t_R^{N_R}} \int_{\mathbb{R}_+^K} \frac{e^{-y}}{N!} \left(\sum_{k=1}^{K} \theta_k y_k \right)^N d\boldsymbol{y} \\ &= \frac{1}{\boldsymbol{N}!} \frac{\partial^{N_1} \cdots \partial^{N_R}}{\partial t_1^{N_1} \cdots \partial t_R^{N_R}} g_{\boldsymbol{t}\boldsymbol{\theta}}(N) \end{aligned} \quad (8)$$

where $g_{\boldsymbol{t}\boldsymbol{\theta}}(N)$ is the normalizing constant of a single-class model with demands $\theta_k = \sum_r \theta_{kr} t_r$, $k = 1, \ldots, K$, and the last passage follows by (3). We prove in Appendix A the following equivalence

$$g_{\boldsymbol{t}\boldsymbol{\theta}}(N) = [\theta_1, \ldots, \theta_K] x^{N+K-1} \quad (9)$$

where the right-hand side is the divided difference[1] of x^{N+K-1} relatively to the interpolation points $\theta_1, \ldots, \theta_K$. We can then apply the Hermite-Genocchi formula [3], which is a classic integral form for divided differences

$$[\theta_1, \ldots, \theta_K] f(x) = \int_{\Delta_K} f^{(K-1)}(\theta_1 u_1 + \ldots + \theta_K u_K) d\boldsymbol{u} \quad (10)$$

where $f^{(K-1)}(x)$ is the $(K-1)$th derivative of $f(x)$ and Δ_K is the unit simplex. Here we set $f(x) = x^{N+K-1}$ and show in Appendix B that (10) holds in this case also under nondistinct demands. Using (9)-(10) in (8) we then get

$$G_{\boldsymbol{\theta}}(\boldsymbol{N}) = \frac{(N+K-1)!}{N!\boldsymbol{N}!} \int_{\Delta_K} \frac{\partial^{N_1} \cdots \partial^{N_R}}{\partial t_1^{N_1} \cdots \partial t_R^{N_R}} \left(\sum_{k=1}^{K} \theta_k u_k \right)^N d\boldsymbol{u}$$

We now plug in $\theta_k = \sum_r \theta_{kr} t_r$ and apply (7) to the integrand, finding (6) after simplifications. □

Theorem 1 provides a novel integral form for the multiclass normalizing constant, with a simplified integrand compared to (3) and defined over a compact domain. The similarities of the integrands are due to the use in both proofs of the multinomial theorem. Moreover it can be shown that (3) follows from (6) by Laplace transformation if the simplex is expressed in a suitable parametric form [34].

The proof of Theorem 1 draws connections between queueing networks and interpolation theory and provides an argument to obtain explicit solutions, as we show in the next section.

[1]For a given function $f(x)$, divided differences extend the notion of forward difference of $f(x)$ to a set of interpolation points arbitrarily located in the domain of $f(x)$. We point to [3, 36] for further details.

3.2 Explicit solutions

While our interest is on deriving approximations, novel exact computational formulas may also be obtained from Theorem 1. Such expressions are not used throughout due to their cost, but they appear of theoretical interest due to the lack of expressions for the multiclass normalizing constant that are both tractable and explicit [9, 31].

COROLLARY 1. *The normalizing constant of a closed multiclass queueing network is given by*

$$G_{\boldsymbol{\theta}}(\boldsymbol{N}) = \sum_{\boldsymbol{0} \leq \boldsymbol{t} \leq \boldsymbol{N}} \frac{(-1)^{N-t}}{\boldsymbol{N}!} \prod_{r=1}^{R} \binom{N_r}{t_r} g_{\boldsymbol{t}\boldsymbol{\theta}}(N) \quad (11)$$

where $\boldsymbol{t} = (t_1, \ldots, t_R)$, $t = \sum_r t_r$, and $g_{\boldsymbol{t}\boldsymbol{\theta}}(N)$ is the normalizing constant of a single-class model with demands $\theta_k = \sum_r t_r \theta_{kr}$.

PROOF. We consider finite differences [36], where (7) is replaced by [4]

$$a_1^{N_1} a_2^{N_2} \cdots a_R^{N_R} = \sum_{\boldsymbol{0} \leq \boldsymbol{t} \leq \boldsymbol{N}} \frac{(-1)^{N-t}}{N!} \prod_{s=1}^{R} \binom{N_s}{t_s} \left(\sum_{r=1}^{R} t_r a_r \right)^N \quad (12)$$

The result follows by using (12) on the product in the integrand of (6) and recognizing $g_{\boldsymbol{t}\boldsymbol{\theta}}(N)$ in the resulting expression. □

Computing $g_{\boldsymbol{t}\boldsymbol{\theta}}(N)$ explicitly using the closed-form formulas in [9, Eq. 3.12] implies that (11) requires $\mathcal{O}(N^R)$ time and $\mathcal{O}(1)$ space. For instance, for a single-class model with distinct demands the normalizing constant can be computed in $\mathcal{O}(1)$ as [30, 9]

$$g_{\boldsymbol{t}\boldsymbol{\theta}}(N) = \sum_{k=1}^{K} \frac{\theta_k^{N+K-1}}{\prod_{i \neq k} (\theta_k - \theta_i)} \quad (13)$$

The next corollary derives the second explicit formula for $G_{\boldsymbol{\theta}}(\boldsymbol{N})$.

COROLLARY 2. *The normalizing constant of a closed multiclass queueing network model can be expressed as*

$$G_{\boldsymbol{\theta}}(\boldsymbol{N}) = \sum_{\substack{\boldsymbol{h} \geq \boldsymbol{0}: \\ h \leq N}} \frac{(-1)^{N-h}}{\boldsymbol{N}!} \binom{N+K-1}{N-h} \prod_{r=1}^{R} \left(\sum_{i=1}^{K} h_i \theta_{ir} \right)^{N_r} \quad (14)$$

where $\boldsymbol{h} \in \mathbb{N}^K$ and $h = \sum_i h_i$.

PROOF. Observe that the specialization of (1) to single-class models is

$$g_{\boldsymbol{t}\boldsymbol{\theta}}(N) = \sum_{\boldsymbol{m} \in \mathcal{S}_K} \prod_{k=1}^{K} \theta_k^{m_k} \quad (15)$$

Using (12) in (15) with $a_k = \theta_k$ and the definition of $\mathcal{S}_K$ yields

$$g_{\boldsymbol{t}\boldsymbol{\theta}}(N) = \sum_{\substack{\boldsymbol{m} \geq \boldsymbol{0}: \\ m = N}} \sum_{\boldsymbol{0} \leq \boldsymbol{h} \leq \boldsymbol{m}} \frac{(-1)^{N-h}}{N!} \prod_{j=1}^{K} \binom{m_j}{h_j} \left(\sum_{i=1}^{K} \sum_{r=1}^{R} h_i t_r \theta_{ir} \right)^N$$

with $h = \sum_i h_i$. Plugging the last formula into (11) and using (12) to eliminate the dependence on $\boldsymbol{t}$ gives after rearranging terms

$$G_{\boldsymbol{\theta}}(\boldsymbol{N}) = \sum_{\substack{\boldsymbol{m} \geq \boldsymbol{0}: \\ m = N}} \sum_{\boldsymbol{0} \leq \boldsymbol{h} \leq \boldsymbol{m}} \frac{(-1)^{N-h}}{\boldsymbol{N}!} \prod_{j=1}^{K} \binom{m_j}{h_j} \prod_{r=1}^{R} \left(\sum_{i=1}^{K} h_i \theta_{ir} \right)^{N_r}$$

We can here use a single summation on $\boldsymbol{h} \geq \mathbf{0}$, $h \leq N$, noting that

$$\sum_{\substack{\boldsymbol{m} \geq \boldsymbol{h}: \\ m=N}} \prod_{i=1}^{K} \binom{m_i}{h_i} = \binom{N+K-1}{N-h} \tag{16}$$

This identity can be proved by first rewriting the expression in terms of $\boldsymbol{v} = \boldsymbol{m} - \boldsymbol{h} \geq \mathbf{0}$ and then iteratively applying a corollary of Vandermonde's convolution [21, Eq. (3)]. □

This explicit form requires $\mathcal{O}(N^K)$ time and $\mathcal{O}(1)$ space, which makes it preferable to (11) on models with many classes, but a small number of nodes.

To the best of our knowledge, (11) and (14) are the only exact and tractable algebraic expressions for $G_{\boldsymbol{\theta}}(\boldsymbol{N})$ with a $\mathcal{O}(1)$ space requirement. This improves over the space complexities of classic recursive algorithms such as CA and RECAL, while retaining the same time complexities. which may be useful in cases where one wants to solve several models in parallel without incurring into memory bottlenecks. In practice, the above expressions are applicable only to models where $H = \min(K, R)$ is small (e.g., $H \leq 4$), typically with up to a few tens of jobs. Moreover, due to the large magnitude of the terms, multi-precision arithmetic should be used to avoid numerical issues upon computing (11) and (14). The techniques developed later do not suffer these problems and help in approximating larger models.

3.3 Infinite server nodes

Consider now a model where the first K nodes are single-server queues and the remaining $M - K$ nodes are infinite servers. The following corollary generalizes the integral form.

COROLLARY 3. *In a model including infinite server nodes*

$$G_{\boldsymbol{\theta}}(\boldsymbol{N}) = \int_{\Delta_K} \int_{v=0}^{+\infty} \frac{e^{-v} v^{K-1}}{\boldsymbol{N}!} \prod_{r=1}^{R} \left(\sigma_r + v \sum_{k=1}^{K} \theta_{kr} u_k \right)^{N_r} dv \, d\boldsymbol{u} \tag{17}$$

PROOF. We first plug the definition of $\sigma_r = \sum_{i=K+1}^{M} \theta_{kr}$ in (17) and then use the multinomial theorem to write

$$G_{\boldsymbol{\theta}}(\boldsymbol{N}) = \int_{\Delta_K} \int_{v=0}^{+\infty} e^{-v} v^{n+K-1} \sum_{\boldsymbol{n} \in \mathcal{S}_M} \prod_{i=1}^{M} \prod_{r=1}^{R} \frac{\theta_{ir}^{n_{ir}}}{\boldsymbol{n}_i!} \prod_{k=1}^{K} u_k^{n_k} dv \, d\boldsymbol{u}$$

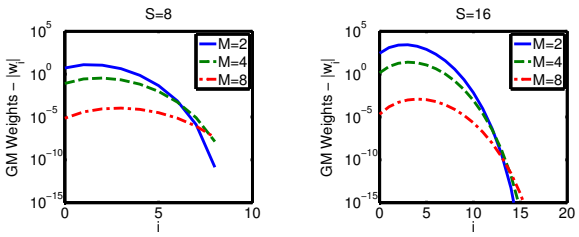
with $n = \sum_{k=1}^{K} n_k$ and $n_k = \sum_r n_{kr}$. Note that $\int_{v=0}^{+\infty} e^{-v} v^{n+K-1} = (n+K-1)!$. We then obtain (2) by applying the closed-form expression of the Dirichlet integral, see (36) in Appendix B. □

Note that in the case $\sigma_r = 0$ the above proof readily provides an alternative derivation of (6). Compared to this elementary argument the proof of (6) provides a strategy to obtain explicit solutions.

3.4 Numerical evaluation

Cubature rules are interpolatory formulas that approximate a multidimensional integral by computing the integrand at a finite set of points and typically best suited for smooth integrands [14]. For polynomial integrands, cubature rules may also allow the exact evaluation of the integral, if interpolation occurs at a large enough set of points. An advantage of (6) over (3) is that it expresses $G_{\boldsymbol{\theta}}(\boldsymbol{N})$ as an integral of a polynomial over the simplex, making it suitable for application of cubature rules.

We focus here on Grundmann-Möller (GM) cubature rules, which are tailored to the exact and approximate integration of polynomials



(a) $S = 8$ (b) $S = 16$

Figure 1: Grundmann-Möller weights w_i for models with $M = K$ single-server nodes

over the simplex [5]. Applying directly the definition of GM cubature rule of degree $2S+1$ to (6) leads to the following expression [5]

$$G_{\boldsymbol{\theta}}(\boldsymbol{N}) = \frac{(N+K-1)!}{\boldsymbol{N}!} \sum_{i=0}^{S} w_i \sum_{\substack{\boldsymbol{b} \geq \mathbf{0}: \\ b=S-i}} \prod_{r=1}^{R} \left(\sum_{j=1}^{K} \frac{(2b_j+1)\theta_{jr}}{(2S+K-2i)} \right)^{N_r} \tag{18}$$

where $\boldsymbol{b} \in \mathbb{N}^K$, $b = \sum_{i=1}^{K} b_i$, and the weights are

$$w_i = (-1)^i 2^{-2S} \frac{(2S+K-2i)^{2S+1}}{i!(2S+K-i)!}$$

The number of points in the rule (18) is $L = \binom{K+S}{S}$, thus worst-case complexity is $\mathcal{O}(S^K)$ time and $\mathcal{O}(1)$ space.

If the integrand is a multivariate polynomial of degree N, then a GM cubature rule of degree $S = \lceil (N-1)/2 \rceil$ returns the exact value of the integral in $\mathcal{O}((N/2)^K)$ time and $\mathcal{O}(1)$ space [5]. This is indeed the case for both (6) and (17). In the case without infinite servers, this is evident since the integrand is a product of linear forms. We now show that the same conclusion holds in models with infinite servers. Let $G^{\boldsymbol{u}}_{\boldsymbol{\theta}}(\boldsymbol{N})$ be the normalizing constant for a model composed of an infinite server node with demand σ_r and K identical single-server nodes having class-r demand $\tilde{\theta}_r = \sum_k \theta_{kr} u_k$. We have the following result.

PROPOSITION 1.

$$\int_{v=0}^{+\infty} \frac{e^{-v} v^{K-1}}{\boldsymbol{N}!} \prod_{r=1}^{R} \left(\sigma_r + v \sum_{k=1}^{K} \theta_{kr} u_k \right)^{N_r} dv = \Gamma(K) G^{\boldsymbol{u}}_{\boldsymbol{\theta}}(\boldsymbol{N})$$

PROOF. Let $\boldsymbol{n} \in \mathbb{N}^R$, $n = \sum_r n_r$. Using [11, Thm. 2] we get

$$G^{\boldsymbol{u}}_{\boldsymbol{\theta}}(\boldsymbol{N}) = \sum_{\mathbf{0} \leq \boldsymbol{n} \leq \boldsymbol{N}} \binom{n+K-1}{n} n! \prod_{r=1}^{R} \frac{\tilde{\theta}_r^{n_r}}{n_r!} \cdot \frac{\sigma_r^{N_r - n_r}}{(N_r - n_r)!}$$

$$= \sum_{\mathbf{0} \leq \boldsymbol{n} \leq \boldsymbol{N}} \frac{\Gamma(n+K)}{\Gamma(K)} \prod_{r=1}^{R} \frac{\tilde{\theta}_r^{n_r}}{n_r!} \cdot \frac{\sigma_r^{N_r - n_r}}{(N_r - n_r)!}$$

We now plug the integral expression of $\Gamma(n+K)$ and the statement follows by the multinomial theorem and the definition of $\tilde{\theta}_r$. □

Since $G^{\boldsymbol{u}}_{\boldsymbol{\theta}}(\boldsymbol{N})$ is a normalizing constant, by definition it is a multivariate polynomial of degree N in $\boldsymbol{u}$. The theorem therefore confirms that the inner integral in (17) is a polynomial of degree N. Thus a GM rule with $S = \lceil (N-1)/2 \rceil$ returns the exact value of $G_{\boldsymbol{\theta}}(\boldsymbol{N})$ also in the presence of infinite servers. Therefore, similarly to (11) and (14), also (18) provides an exact expression for $G_{\boldsymbol{\theta}}(\boldsymbol{N})$ that is both explicit and tractable.

Using smaller values of S trades accuracy for speed. In particular it is possible to approximate $G_{\boldsymbol{\theta}}(\boldsymbol{N})$ by truncation of the outer summation of (18) thanks to the rapid decay of the weights w_i.

This is illustrated in Figure 1. For large enough i, the weights quickly and monotonically decrease, thus a few outer iterations of (18) are often sufficient to return a good approximation. As we show later, GM rules perform very well on small and medium-sized inference problems. However, as the model size grows, the asymptotic expansions introduced next become more efficient.

4. ASYMPTOTIC EXPANSION

4.1 Preliminaries

We now exploit the geometry of the unit simplex to derive an asymptotic expansion for $G_{\boldsymbol{\theta}}(\boldsymbol{N})$. Our approach first applies a logistic transformation to the integration variables, which is a classic method for mapping integrands defined over the n-dimensional simplex to $\mathbb{R}^n$ [1]. We find after this transformation that the integrand becomes increasingly peaked at a unique point in the interior of the integration domain as N grows, satisfying the conditions for Laplace's method [25].

A technical requirement to apply this method is that all queue-lengths grow asymptotically large as $N \to +\infty$, a property which is violated by non-bottleneck nodes. To address this issue, we introduce a novel scaling where we also slowly increase a population of jobs that permanently reside at a single node each. That is, we introduce K additional classes, each with population ϵN, $\epsilon > 0$, where the ith class is composed by jobs that self-loop at node i, placing a unit service demand at each visit. In this way we are considering the perturbed normalizing constant

$$G^{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{N}) = \frac{(\eta_\epsilon(N)-1)!}{\boldsymbol{N}!(\epsilon N!)^K} \int_{v=0}^{+\infty} e^{-v} v^{K(1+\epsilon N)-1} \times \int_{\Delta_K} \prod_{i=1}^{K} u_i^{\epsilon N} \prod_{r=1}^{R} \left(\sigma_r + v \sum_{k=1}^{K} \theta_{kr} u_k \right)^{N_r} d\boldsymbol{u}\, dv \quad (19)$$

where $\eta_\epsilon(N) = N + K(1+\epsilon N)$. Note that

$$\lim_{\epsilon \to 0} G^{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{N}) = G_{\boldsymbol{\theta}}(\boldsymbol{N})$$

Our asymptotic expansion is proved for an auxiliary function $I_\epsilon(N)$, which uniquely defines $G^{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{N})$. This function is defined as follows. First, we allow for real values of ϵ by replacing where needed factorials in $G^{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{N})$ with the gamma function $\Gamma(\cdot)$. Without loss of generality, we then normalize service demands to range in $[0, 1]$. That is, we set

$$G^{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{N}) = \frac{\Gamma(\eta_\epsilon(N))}{\boldsymbol{N}!(\Gamma(1+\epsilon N))^K} I_\epsilon(N) \prod_{r=1}^{R} \alpha_r^{N_r} \quad (20)$$

where $\alpha_r = \sigma_r + \max_i \theta_{ir}$ and we define the auxiliary function

$$I_\epsilon(N) = \frac{1}{\Gamma(\eta_\epsilon(N))} \int_{v=0}^{+\infty} e^{-v} v^{K(1+\epsilon N)-1} \times \int_{\Delta_K} \prod_{i=1}^{K} u_i^{\epsilon N} \prod_{r=1}^{R} \left(\tilde{\sigma}_r + v \sum_{k=1}^{K} \tilde{\theta}_{kr} u_k \right)^{N_r} d\boldsymbol{u}\, dv \quad (21)$$

with $\tilde{\theta}_{kr} = \alpha_r^{-1} \theta_{kr}$, and $\tilde{\sigma}_r = \sum_{k=K+1}^{M} \tilde{\theta}_{kr}$. From now on, we focus on $I_\epsilon(N)$ and to simplify notation use θ_{kr} and σ_r in place of $\tilde{\theta}_{kr}$ and $\tilde{\sigma}_r$, subject to $\theta_{kr} \leq 1$ and $\sigma_r \leq 1$. Moreover, we assume that the ratios $\beta_r = N_r/N$ remain constant while increasing N.

4.2 Laplace's method

We first obtain the asymptotic approximation for $I_\epsilon(N)$ at $\hat{\boldsymbol{u}}$ in a model with single-server nodes only. This method requires to verify a set of well-known analytical conditions [25]. We here verify a slightly stronger set of assumptions. After showing that

$$I_\epsilon(N) = \int_{\mathbb{R}^{K-1}} e^{-N h_N(\boldsymbol{x})} d\boldsymbol{x} \quad (22)$$

for smooth and infinitely differentiable $h_N(\boldsymbol{x})$, having constant order with respect to N and bounded derivatives, we prove the validity of Laplace's method $\forall N > 0$ by showing that there exist a $\epsilon_N > 0$ such that $\forall \epsilon \geq \epsilon_N$:

- Condition 1: $I_\epsilon(N)$ exists and it is finite;
- Condition 2: $h_N(\boldsymbol{x})$ attains a unique stationary point in the interior of the integration domain of (22);
- Condition 3: the Hessian of $-N h_N(\boldsymbol{x})$ has a positive determinant at its stationary point.

Under these conditions it is possible to apply Laplace's method [25], which provides a $\mathcal{O}(N^{-1})$ asymptotic approximation. Higher-order expansions may also be considered, but their computational cost grows quickly with the number of nodes in the model [25].

THEOREM 2. *In a closed network without infinite servers, for all $N > 0$ there exists an $\epsilon_N > 0$ such that $\forall \epsilon \geq \epsilon_N$*

$$I_\epsilon(N) = \sqrt{\frac{(2\pi)^{K-1}}{\det(\boldsymbol{A}_\epsilon)}} \prod_{r=1}^{R} \left(\sum_{k=1}^{K} \theta_{kr} \hat{u}_k \right)^{N_r} \prod_{i=1}^{K} \hat{u}_i^{1+\epsilon N} + \mathcal{O}(N^{-1}) \quad (23)$$

where $\hat{\boldsymbol{u}}$ is the unique solution in Δ_K of the system of nonlinear equations

$$u_i = \eta_\epsilon^{-1}(N) \left(1 + \epsilon N + \sum_{r=1}^{R} \xi_r(\boldsymbol{u}) \theta_{ir} u_i \right) \quad i = 1, \ldots, K-1 \quad (24)$$

with $\xi_r(\boldsymbol{u}) = N_r (\sum_{k=1}^{K} \theta_{kr} u_k)^{-1}$, and where $\det(\boldsymbol{A}_\epsilon) > 0$ with $\boldsymbol{A}_\epsilon = [A_{ij}]$ having entries

$$A_{ij} = \begin{cases} -\sum_{\substack{h=1 \\ h \neq i}}^{K} A_{ih} & i = j \\ \left(\sum_{r=1}^{R} \frac{\xi_r^2(\hat{\boldsymbol{u}})}{N_r} \theta_{ir} \theta_{jr} - \eta_\epsilon(N) \right) \hat{u}_i \hat{u}_j & i \neq j \end{cases} \quad (25)$$

for $i, j = 1, \ldots, K-1$.

The last result provides by (20) an approximation for $G^{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{N})$. The role of the ϵ parameter is to ensure that the stationary point of the integrand of (6) belongs to the interior of the integration domain and that $\det(\boldsymbol{A}_\epsilon) > 0$. For models with a finite N, the first condition holds irrespective of the value of ϵ, which is needed only asymptotically. Thus in applications it is sufficient to choose an ϵ such that $\det(\boldsymbol{A}_\epsilon) > 0$ and approximations with smaller values of ϵ are expected to return more accurate results. We also show that, for a sufficiently large ϵ, matrix $\boldsymbol{A}$ is positive definite. This property is used later to develop a Monte Carlo integration method.

Lastly, we note that (23) is a product. This is highly beneficial in applications, since we can avoid numerical difficulties associated to the normalizing constant by directly computing $\log G_{\boldsymbol{\theta}}(\boldsymbol{N})$. As a result, across thousands of models that we have solved in the numerical validation, we have never experienced numerical issues with (23). On the contrary, normalizing constant methods based on summations such as CA or (18) eventually fail on large models due to floating-point range exceptions and round-off errors.

4.3 Proof of Theorem 2

The result follows by proving the assumptions for the Laplace's method. Since $\sigma_r = 0$, (21) here simplifies to

$$I_\epsilon(N) = \int_{\Delta_K} \prod_{i=1}^{K} u_i^{\epsilon N} \prod_{r=1}^{R} \left(\sum_{k=1}^{K} \theta_{kr} u_k \right)^{N_r} d\boldsymbol{u} \tag{26}$$

We first apply an additive logistic transformation [1]

$$u_i = \begin{cases} e^{x_i} \left(1 + \sum_{k=1}^{K-1} e^{x_k}\right)^{-1} & i = 1, \dots, K-1 \\ \left(1 + \sum_{k=1}^{K-1} e^{x_k}\right)^{-1} & i = K \end{cases} \tag{27}$$

with Jacobian

$$\left| \frac{\partial \boldsymbol{u}}{\partial \boldsymbol{x}} \right| = \prod_{i=1}^{K-1} e^{x_i} \left(1 + \sum_{k=1}^{K-1} e^{x_k}\right)^{-K}$$

This yields (22) with

$$-N h_N(\boldsymbol{x}) = \sum_{i=1}^{K-1} (1+\epsilon N) x_i + \sum_{r=1}^{R} N_r \log \left(\theta_{Kr} + \sum_{k=1}^{K-1} \theta_{kr} e^{x_k} \right) - \eta_\epsilon(N) \log \left(1 + \sum_{k=1}^{K-1} e^{x_k} \right)$$

Note that $h_N(\boldsymbol{x})$ is smooth and infinitely differentiable. Moreover, $h_N(\boldsymbol{x})$ has a constant order with respect to N and its partial derivatives are also smooth and bounded at all orders. We are now ready to verify the conditions for Laplace's method given in Section 4.2.

4.3.1 Condition 1: existence and finiteness

Since the logistic transformation does not affect existence and finiteness, it is sufficient to verify these properties on the initial integral (26). Existence follows since Δ_K is a finite domain and the integrand of (26) exists at all points of Δ_K. $I_\epsilon(N)$ is also finite for all $\epsilon > 0$ and $N > 0$ since $\theta_{kr} \leq 1$ and the domain Δ_K has a constant volume as N grows.

4.3.2 Condition 2: unique stationary point

Condition 2 is verified as follows. We set $\nabla h_N(\boldsymbol{x}) = \boldsymbol{0}$ and use the inverse transformation of (27) to express the result over Δ_K. The transformation is given by [1]

$$\hat{\boldsymbol{x}} = (\log \hat{u}_1 / \hat{u}_K, \dots, \log \hat{u}_{K-1} / \hat{u}_K) \tag{28}$$

and yields the system of nonlinear equations (24). We now show that this system admits a unique solution $\hat{\boldsymbol{u}}$. Moreover we also show that if $\epsilon > 0$ then $\hat{\boldsymbol{u}} \in \Delta^\epsilon_{K,N} \subset \Delta_K$ for all $N > 0$, with

$$\Delta^\epsilon_{K,N} = \left\{ \boldsymbol{u} | \boldsymbol{u} \in \Delta_K, u_i \geq (1+\epsilon N) \eta_\epsilon^{-1}(N), \forall i \right\}$$

Mapping back $\hat{\boldsymbol{u}}$ to $\mathbb{R}^{K-1}$ using the logistic transformation (27), this implies that the stationary point of $h_N(\boldsymbol{x})$ is in the interior of the integration domain.

We begin by proving existence of a solution in $\Delta^\epsilon_{K,N}$. Let us consider a point $\boldsymbol{u}^{(n)} \in \Delta_K$, $n \in \mathbb{N}$, and the continuous mapping

$$u_i^{(n+1)} = \eta_\epsilon^{-1}(N) \left(1 + \epsilon N + \sum_{r=1}^{R} \xi_r(\boldsymbol{u}^{(n)}) \theta_{ir} u_i^{(n)} \right) \tag{29}$$

for $i = 1, \dots, K-1$. Since Δ_K is convex, non-empty and compact, (29) has a fixed point $\hat{\boldsymbol{u}} \in \Delta_K$ and this must also be a solution of (24) by definition. We now show by contradiction that (24) has no solution in $\Delta_K \setminus \Delta^\epsilon_{K,N}$. Assume that a solution $\boldsymbol{u} \in \Delta_K \setminus \Delta^\epsilon_{K,N}$ exists. Then there exists a node i such that $0 \leq u_i < (1+\epsilon N)\eta_\epsilon^{-1}(N)$. Plugging $\boldsymbol{u}$ into (24) makes the ith equation infeasible, since $\sum_{r=1}^{R} \xi_r(\hat{\boldsymbol{u}}) \theta_{ir} u_i \geq 0$ implies $u_i \geq (1+\epsilon N)\eta_\epsilon^{-1}(N)$, against the assumptions.

To prove uniqueness, we focus on $\Delta^\epsilon_{K,N}$, since no solutions exist outside this sub-domain. Consider the nonlinear program

$$\min_{\boldsymbol{u} \in \Delta^\epsilon_{K,N}} \sum_{i=1}^{K-1} \frac{u_i^2}{2} - \eta_\epsilon^{-1}(N) \left((1+\epsilon N) \sum_{i=1}^{K-1} u_i + \sum_{r=1}^{R} N_r \log \left(\sum_{k=1}^{K} \theta_{kr} u_k \right) \right) \tag{30}$$

with first-order Karush-Kuhn-Tucker (KKT) conditions

$$\begin{aligned} -u_i + \eta_\epsilon^{-1}(N) \left((1+\epsilon N) + \sum_{r=1}^{R} \xi_r \theta_{ir} u_i \right) &= \lambda + \mu_i \\ \sum_{k=1}^{K} u_k &= 1 \\ u_i \mu_i &= 0 \\ \eta_\epsilon(N) u_i &\geq (1+\epsilon N) \\ \mu_i &\geq 0 \end{aligned}$$

for all $i = 1, \dots, K-1$. A feasible solution in $\Delta^\epsilon_{K,M}$ requires $\mu_i = 0$, $\forall i$. It is possible to verify that the objective is strictly convex over $\Delta^\epsilon_{K,N}$, being the sum of functions that are convex and strictly convex over this domain. Thus the KKT conditions admit a unique solution, which must be $\hat{\boldsymbol{u}}$ since this is feasible for $\lambda = 0$. Since the solutions to the above KKT conditions include all the solutions of (24) in $\Delta^\epsilon_{K,N}$, we conclude that (24) has a unique solution $\hat{\boldsymbol{u}}$ in $\Delta^\epsilon_{K,N}$.

4.3.3 Condition 3: positive Hessian determinant

Let $\boldsymbol{A}_\epsilon = [A_{ij}], i, j = 1, \dots, K-1$, be the Hessian of $-N h_N(\boldsymbol{x})$ evaluated at the stationary point $\hat{\boldsymbol{x}}$. Computing $\boldsymbol{A}_\epsilon$ by the definition we obtain (25). We now prove that for all $N > 0$ there exists an $\epsilon_N > 0$ such that $\boldsymbol{A}_\epsilon$ is positive definite for all $\epsilon \geq \epsilon_N$.

From (24) we have $\lim_{\epsilon \to +\infty} \hat{u}_i = K^{-1}, \forall i$ which implies that $\lim_{\epsilon \to +\infty} A_{ij} < 0$, $i \neq j$. Thus, there exist an ϵ_N such that $\forall \epsilon > \epsilon_N$, $-\boldsymbol{A}_\epsilon$ has positive off-diagonal entries. This implies that

$$\boldsymbol{Q} = - \begin{bmatrix} \boldsymbol{A}_\epsilon & \boldsymbol{a}_K^T \\ \boldsymbol{a}_K & A_{KK} \end{bmatrix}$$

with $\boldsymbol{a}_K = (A_{1K}, \dots, A_{K-1,K})$, is an irreducible infinitesimal generator. Being $-\boldsymbol{A}_\epsilon$ the principal sub-matrix of an irreducible infinitesimal generator, it is negative definite and therefore $\boldsymbol{A}_\epsilon$ is positive definite, implying $\det(\boldsymbol{A}_\epsilon) > 0, \forall \epsilon > \epsilon_N$.

4.3.4 Final expression

To conclude the proof of Theorem 2 we can apply Laplace's method in $\mathbb{R}^{K-1}$ to obtain the expansion

$$I_\epsilon(N) = \sqrt{\frac{(2\pi)^{K-1}}{\det(\boldsymbol{A}_\epsilon)}} e^{-N h_N(\hat{\boldsymbol{x}})} + \mathcal{O}(N^{-1})$$

The final expression (23) follows by the expression of $h_N(\hat{\boldsymbol{x}})$ and the inverse transformation (28).

4.4 Extensions

4.4.1 Models with infinite servers

When one or more nodes become empty, the asymptotic stationary point of $h_N(\boldsymbol{x})$ does not lie anymore in the interior of the

integration domain. In the presence of infinite servers, it does not seem possible to scale parameters in (22) to prevent this situation in the asymptotic regime, where either infinite servers or single-server nodes will become a bottleneck. In the case without infinite servers, the perturbation ϵ ensures that this situation does not happen, but the same method cannot prevent infinite servers from becoming empty. In this case we apply Laplace's method as a sub-asymptotic heuristic. The sub-asymptotic method amounts to fitting the integrand of (22) to a multivariate normal density and using the resulting closed-form expressions to approximate $I_\epsilon(N)$, for $N < \infty$. The method also requires to approximate the indefinite integral in (17).

The heuristic follows a very similar argument as in the case without infinite servers, thus we just give a sketch. We first apply the change of variable $v = e^{x_0}$ and the logistic transformation (27) so that we now have a K-dimensional integral

$$I_\epsilon(N) = \int_{\mathbb{R}^K} b_N e^{-N h_N(\boldsymbol{x})} d\boldsymbol{x}$$

with $\boldsymbol{x} = (x_1, \ldots, x_{K-1}, x_0)$, $b_N = (\Gamma(\eta_\epsilon(N)))^{-1}$, and

$$\begin{aligned} -N h_N(\boldsymbol{x}) &= -e^{x_0} + K(1+\epsilon N) x_0 \\ &+ \sum_{r=1}^{R} N_r \log\left(\sigma_r + e^{x_0}\theta_{Kr} + e^{x_0}\sum_{k=1}^{K-1}(\sigma_r + \theta_{kr} e^{x_k})\right) \\ &+ (1+\epsilon N)\sum_{k=1}^{K-1} x_k - \eta_\epsilon(N)\log\left(1 + \sum_{k=1}^{K-1} e^{x_k}\right) \end{aligned}$$

Setting $\nabla h_N(\boldsymbol{x}) = 0$ and simplifying terms, we find that the stationary point is written in terms of the original integration variables as the solution $(\hat{\boldsymbol{u}}, \hat{v})$ of the system

$$\begin{aligned} u_i &= \eta_\epsilon^{-1}(N)\left(1 + \epsilon N + \sum_{r=1}^{R}\xi_r(\boldsymbol{u}, v)(\sigma_r + v\theta_{ir})u_i\right) \quad \forall i \neq K \\ v &= \eta_\epsilon(N) + 1 - \sum_{r=1}^{R}\xi_r(\boldsymbol{u}, v)\sigma_r \end{aligned} \tag{31}$$

where $\boldsymbol{u} \in \Delta_K$, $\xi_r(\boldsymbol{u}, v) = N_r(\sigma_r + v\sum_{k=1}^{K}\theta_{kr}u_k)^{-1}$. The formula for v uses that $\sum_{r=1}^{R}\xi_r(\boldsymbol{u}, v)(\sigma_r + v\sum_{k=1}^{K}\theta_{kr}u_k) = N$.

Explicit formulas for the entries of $\boldsymbol{A}_\epsilon$ are obtained by computing the Hessian matrix of $-N h_N(\boldsymbol{x})$ expressed in terms of the variables $(\hat{\boldsymbol{u}}, \hat{v})$. Define $\hat{\theta}_{kr} = \sigma_r + \hat{v}\theta_{ir}$, $\forall k, r$, then the entries of $\boldsymbol{A}_\epsilon = [A_{ij}]$ are the second-order partial derivatives evaluated at $\boldsymbol{x} = \hat{\boldsymbol{x}}$. Since $\boldsymbol{A}$ is a symmetric matrix, its entries are entirely specified by the following expressions

$$\begin{aligned} A_{ii} &= -\sum_{j\neq i} A_{ij} \\ A_{ij} &= \sum_{r=1}^{R}\left(\frac{\xi_r^2(\hat{\boldsymbol{u}}, \hat{v})}{N_r}\hat{\theta}_{kr}\hat{\theta}_{jr} - \eta_\epsilon(N)\right)\hat{u}_i\hat{u}_j \qquad i \neq j \\ A_{00} &= -\hat{v} + \hat{v}\sum_{r=1}^{R}\frac{\xi_r^2(\hat{\boldsymbol{u}}, \hat{v})}{N_r}\left[\sum_{k=1}^{K}\hat{\theta}_{kr}\hat{u}_k - \hat{v}\left(\sum_{k=1}^{K-1}\hat{u}_k\theta_{kr}\right)\right]\left(\sum_{k=1}^{K}\theta_{kr}\hat{u}_k\right) \\ A_{i0} &= -\hat{v}\hat{u}_i\left[\sum_{r=1}^{R}\frac{\xi_r^2(\hat{\boldsymbol{u}}, \hat{v})}{N_r}\hat{\theta}_{ir}\left(\sum_{k=1}^{K}\theta_{kr}\hat{u}_k\right) - \sum_{r=1}^{R}\xi_r(\hat{\boldsymbol{u}}, \hat{v})\theta_{ir}\right] \quad i \neq 0 \end{aligned}$$

for all $i, j = 1, \ldots, K-1$. The knowledge of $(\hat{\boldsymbol{u}}, \hat{v})$ and $\boldsymbol{A}_\epsilon$ enables a Laplace-type approximation for (17)

$$\begin{aligned} I_\epsilon(N) \approx \frac{e^{-\hat{v}}\hat{v}^{K(1+\epsilon N)}}{\Gamma(\eta_\epsilon(N))}\sqrt{\frac{(2\pi)^K}{\det(\boldsymbol{A}_\epsilon)}} \\ \times \prod_{r=1}^{R}\left(\sigma_r + \hat{v}\sum_{k=1}^{K}\theta_{kr}\hat{u}_k\right)^{N_r}\prod_{k=1}^{K}\hat{u}_k^{1+\epsilon N} \end{aligned} \tag{32}$$

where the exponent of $\hat{v}$ includes the contribution of the Jacobian, and we have used the transformations (28) and $x_0 = \log v$. Equation (32) can be readily applied to approximating models with infinite servers. The recommendations for the optimal selection of the ϵ parameter given for the case without infinite servers apply also here.

4.4.2 *Combining Laplace's method with AMVA*

In some inference problems, measurements for both system state and mean performance metrics are available. In this case, in addition to likelihood-based inference, one may require that the estimated model also matches the empirical values of some mean performance metrics. This typically requires to run an AMVA algorithm alongside the likelihood maximization algorithm.

In this section we argue that a more efficient way to optimize these models is to heuristically calibrate $\hat{\boldsymbol{u}}$ using the results of mean-value analysis. This effectively doubles the speed of the optimization, since the same AMVA fixed-point iteration can be used both to calculate likelihoods and constraints on mean measures. Furthermore, our validation results shown later indicate that this heuristic calibration can deliver slightly more accurate results than (26), suggesting that some sub-asymptotic components of $G_{\boldsymbol{\theta}}(\boldsymbol{N})$ not captured by Laplace's method can be recovered through AMVA. A limitation of this calibration is that no formal guarantee is in place to ensure that $\det(\boldsymbol{A}_\epsilon) > 0$ on all instances. However, no problematic instance in this sense is observed throughout the numerical validation. In degenerate cases where $\det(\boldsymbol{A}_\epsilon) \leq 0$ one may try to heuristically increase ϵ in order to resolve this issue.

Let us now observe that, as $N \to \infty$, (24) tends to

$$u_i = \frac{\epsilon}{1+K\epsilon} + \sum_{r=1}^{R}\hat{\xi}_r(\hat{\boldsymbol{u}})\theta_{ir}u_i \quad i = 1, \ldots, K \tag{33}$$

where $\hat{\xi}_r(\boldsymbol{u}) = \beta_r(1+K\epsilon)^{-1}(\sum_{k=1}^{K}\theta_{kr}u_k)^{-1}$. As $\epsilon \to 0$ the last expression coincides with the expression of the mean-value analysis queue-length equations when u_i is the total queue-length at node i normalized by N and $N \to \infty$. This suggests an alternative heuristic to determine the position of the point $\hat{\boldsymbol{u}}$ at which we instantiate the Laplace's method. Instead of using (24) this may be chosen as $\hat{q}_i = \sum_r q_{ir}(\boldsymbol{N})/N$, $\forall i$, where $q_{ir}(\boldsymbol{N})$ is the mean queue-length of class r at node i in a model with population $\boldsymbol{N}$, which can be efficiently approximated using AMVA. In order to reduce computational burden it is therefore sufficient to replace $\hat{\boldsymbol{u}}$ with $\hat{\boldsymbol{q}} = (\hat{q}_1, \ldots, \hat{q}_M)$.

4.4.3 *Monte Carlo integration*

The applicability of Laplace's method indicates that the integral $I_\epsilon(N)$ may be approximated using a multivariate normal distribution centered at $\hat{\boldsymbol{u}}$, with covariance matrix $\boldsymbol{A}_\epsilon^{-1}$. The resulting normal distribution is non-degenerate if $\boldsymbol{A}_\epsilon$ is positive definite. In situations where asymptotic expansions are expensive or inaccurate, one may thus apply an importance sampling method to $I_\epsilon(N)$ using samples from a multivariate normal. Let $\boldsymbol{x} \sim \mathcal{N}(\hat{\boldsymbol{x}}, \boldsymbol{A}_\epsilon^{-1})$ where $\hat{\boldsymbol{x}}$ is obtained by applying the inverse of the logistic transformation (27) to $\hat{\boldsymbol{u}}$, i.e., $\hat{x}_i = \log(\hat{u}_i/\hat{u}_K)$, $\forall i \neq K$. Denote by $\boldsymbol{x}^{(j)}$ the jth

Table 2: Small and medium models - deterministic methods

$K, R \in \{2,4,6\}$	MAPE (%)		
$N/R =$	*2*	*4*	*8*
CUB1	0.0	0.4	8.9
CUB3	0.0	0.4	7.4
CUB5	0.0	0.0	0.7
CUB7	0.0	0.0	0.1
LE	25.7	25.2	24.1
LE-A	26.7	23.3	18.3
RAY	362.0	142.7	96.5

Table 3: Small and medium models - randomized methods

$K, R \in \{2,4,6\}$	MAPE (%)		
$N/R =$	*2*	*4*	*8*
LS1	23.1	21.1	31.6
LS2	12.1	12.9	12.1
LS3	6.0	7.0	7.4
MCI1	28.9	41.7	55.5
MCI2	9.3	13.0	18.1
MCI3	3.0	4.1	5.3

sample drawn, out of a total of J. We have the importance sampling estimator

$$I_\epsilon(N) \approx J^{-1} \sum_{j=1}^{J} b_N e^{-N h_N(\boldsymbol{x}^{(j)})} \phi^{-1}(\boldsymbol{x}^{(j)}) \tag{34}$$

where ϕ is the normal density function and b_N and h_N are as in Section 4.3. From basic Monte Carlo integration theory, this expression converges to $I_\epsilon(N)$ as $O(J^{-1/2})$ under a growing number of samples J. The same result can be applied to models with infinite servers, using the formulas in Section 4.4.1.

5. NUMERICAL RESULTS

5.1 Algorithms

In this section we assess accuracy and speed of the proposed methods. We distinguish the proposed algorithms in two groups:

- *Deterministic methods*, such as the asymptotic expansions, which return the same answer in successive invocations on the same model and therefore are suitable for use within deterministic optimization programs. We include in this group RAY and the asymptotic expansion (22), referred to as the *logistic expansion* (LE), and the heuristic variant of LE calibrated with AMVA, denoted by LE-A. We also consider in this group the cubature rules given in (18) with $S = \{1,3,5,7\}$ and denote, e.g., by CUB5 a cubature rule with $S = 5$. We have also experimented with TE, CA, and MoM, but computational times are far larger than those of the other methods and incompatible with the scale of the experimental validation, which encompasses thousands of optimization programs.
- *Randomized methods*, which use sampling to achieve the desired accuracy in return for an increased effort. We include in this group MCI and the Monte Carlo integration method in (34), which we call *logistic sampling* (LS). Monte Carlo methods are instantiated with $J \in \{10^1, 10^2, 10^3\}$ samples, e.g., MCI2 stands for MCI with $J = 10^2$, and similarly LS3 has $J = 10^3$.

For deterministic methods, we are interested in assessing both accuracy and ability to guide optimization-based search. For randomized methods, we verify accuracy as the number of samples grows. Remarks on the implementations are as follows:

- In LE we use fixed-point iteration to solve (29), setting the convergence tolerance on the 1-norm of $\hat{\boldsymbol{u}}$ to $\tau = 10^{-10}$. The initial point has $u_i = 1/K$, $\forall i$. We also set $\epsilon = 10^{-10}$. Out of the thousands of models solved, none failed to converge and none required to increase ϵ beyond its initial value.
- LE-A is implemented using the Bard-Schweitzer AMVA [42] for determining the point $\hat{\boldsymbol{q}}$ with a $\delta = 10^{-6}$ convergence tolerance on the 1-norm of the mean queue-lengths. Also in this case none of the models failed to converge.
- For cases where $\boldsymbol{A}_\epsilon$ is not positive definite, LS is instantiated by increasing ϵ in steps of $10^{-3}N$, until obtaining a positive definite matrix. In small and medium-sized models, this calibration is not normally required. However, on large models where the entries of $\boldsymbol{A}$ are small, very few increments of ϵ are normally sufficient to address the issue. In the random validation on large models, this calibration is required on 44% of the instances and occurs prior to computing (34). The computational cost of the calibration is negligible.

5.2 Methodology

An important issue for the validation methodology is that for large models $G_{\boldsymbol{\theta}}(\boldsymbol{N})$ cannot be computed exactly due to the large cost of the exact algorithms. Thus we first carry out a validation on small and medium-sized models where the normalizing constant can be obtained exactly. Afterwards, we report a similar validation on larger models where we estimate $G_{\boldsymbol{\theta}}(\boldsymbol{N})$ by Monte Carlo integration with a large number of samples (10^7). In the large-scale setting, we attempt to compensate the variance of the estimator of the normalizing constant by assessing percentage error with respect to the scale, i.e., $\log G_{\boldsymbol{\theta}}(\boldsymbol{N})$. Note that on most large-scale models the order of the normalizing constant is the dominant factor in the likelihood expressions.

The validation does not include mean performance metrics. This is because their computation can be performed very efficiently using AMVA methods [42]. Our methods are instead proposed for the accurate computation of likelihoods and probabilities, which require $G_{\boldsymbol{\theta}}(\boldsymbol{N})$ and are still difficult to compute in practice.

The computational times to run LE and LE-A inside optimization programs are very small, typically a fraction of a second. This is due to the rapid convergence of the fixed-point algorithms used to determine the location of the stationary point. For example, on the largest model with $K = R = 64$ and $N = 4096$ LE takes $26s$ to find at the first iteration the stationary point, but just $2s$ for a new prediction after a 10^{-6} increment of $\boldsymbol{\theta}$, provided that the fixed point equations (29) are re-initialized at the previously-found stationary point. When the model size is decreased to $K = R = 8$ the first solution requires just $0.21s$, while successive updates about $0.015s$. Since the optimization-based study considers an identical timeout for all methods, we do not provide details on the running times of individual algorithms. Lastly, we remark that space complexity is negligible and does not grow significantly with the model size. This is because all approximation methods considered throughout have $\mathcal{O}(1)$ space complexity as the population sizes N grow, with N typically being the largest model parameter.

5.3 Computing a single normalizing constant

5.3.1 *Small and medium-sized models*

We consider randomly-generated models with $K \in \{2,4,6\}$ nodes, $R \in \{2,4,6\}$ classes, and where each class has the same

Table 4: Small and medium models with infinite server nodes - deterministic methods

$K, R \in \{2, 4, 6\}$	MAPE (%)			
σ_r	$0.1\theta_{max}$	θ_{max}	$10\theta_{max}$	$100\theta_{max}$
PAN	N/A	N/A	26.2	0.1
LE	7.1	4.8	2.9	1.5

Table 5: Large models - deterministic methods

$K, R \in \{16, 32, 64\}$	MAPE (%)					
$N/R =$	*2*	*4*	*8*	*16*	*32*	*64*
CUB1	0.0	0.1	0.4	1.1	2.5	2.9
LE	1.8	1.0	0.7	0.4	0.4	0.5
LE-A	1.8	1.0	0.7	0.4	0.4	0.4
RAY	14.5	4.8	1.4	0.5	0.3	0.4

number of jobs equal to $N/R = \{2, 4, 8\}$. Thus the largest model in this group has 6 nodes, 6 classes and 48 jobs. We use less jobs than in the motivating example in Table 1 since we now consider models with $R = 6$ classes that are much more expensive to solve exactly. For any given triplet (K, R, N), we solve 100 random instances, for a total of 2700 models. In each instance, demands are generated at random in $[0, 1]$. $G_{\boldsymbol{\theta}}(\boldsymbol{N})$ is computed exactly using the CA algorithm.

Note that RAY is the only method that incurs failures during execution. This occurs on 9 out of 2700 models and it is due to a singular determinant in its expression. Indeed, RAY does not provide correctness guarantees in the sub-asymptotic setting [27]. We count as a failure a run that either stops due to excessive memory requirements, or that returns a 0, NaN, $\pm\infty$, or a complex value for $G_{\boldsymbol{\theta}}(\boldsymbol{N})$ due to numerical issues.

Tables 2 and 3 give the *mean absolute percentage error* (MAPE) on $G_{\boldsymbol{\theta}}(\boldsymbol{N})$ for deterministic and randomized methods. The results indicate that the CUB dominates all other methods, including the randomized ones. Execution times of CUB on these models are in the order of a few milliseconds, making this method preferable in small and medium-sized models. As expected, asymptotic expansions incur smaller errors as the number of jobs grows. The Monte Carlo integration methods, MCI and LS, are instead more accurate with fewer jobs. However, increasing the number of samples in both methods quickly lowers errors to the desired level.

In order to better understand the differences between MCI and LS, we have investigated how the accuracy of the two methods varies under increasing number of nodes K or number of classes R. MCI decreases errors as K increases, whereas it performs worse under increasing R. For example, MCI1 goes from 35.4% to 48.2% as the number of classes goes from 2 to 6, whereas it decreases errors from 57.4% to 30.1% when the number of nodes grows of the same amount. Conversely, LS is rather insensitive to R, with LS1 error lying between 22.5% and 26.4%, but the method incurs larger errors as K grows, with LS1 going from 12.5% with 2 nodes to 38.9% with 6 nodes. This increased error is due to the larger number of integration dimensions in (22), which requires a larger N value to deliver a similar level of accuracy. Similar trends are seen also in large models and suggest that the two methods can complement each other, preferring MCI on models with many nodes and LS on models with several classes.

Table 6: Large models - randomized methods

$K, R \in \{16, 32, 64\}$	MAPE (%)					
$N/R =$	*2*	*4*	*8*	*16*	*32*	*64*
LS1	0.9	0.6	1.4	2.4	2.3	1.8
LS2	0.5	0.5	1.3	2.2	2.2	1.8
LS3	0.3	0.4	1.2	2.1	2.1	1.7
MCI1	0.1	0.1	0.2	0.3	1.5	6.1
MCI2	0.0	0.0	0.1	0.1	0.4	3.3
MCI3	0.0	0.0	0.0	0.0	0.2	1.6

5.3.2 *Large-scale models*

We now consider a similar setup as in the previous experiment, but with K, R ranging in $\{16, 32, 64\}$. The number of jobs ranges in $N/R = \{2, 4, 8, 16, 32, 64\}$. Thus the largest models have 64 nodes, 64 classes and 4096 jobs. As explained before, in this setting it is difficult to obtain the exact value of the normalizing constant, thus we compare against MCI with 10^7 samples and focus on matching $\log G_{\boldsymbol{\theta}}(\boldsymbol{N})$. The MAPE on $\log G_{\boldsymbol{\theta}}(\boldsymbol{N})$ may be seen as the percentage error introduced in log-likelihoods such as (5).

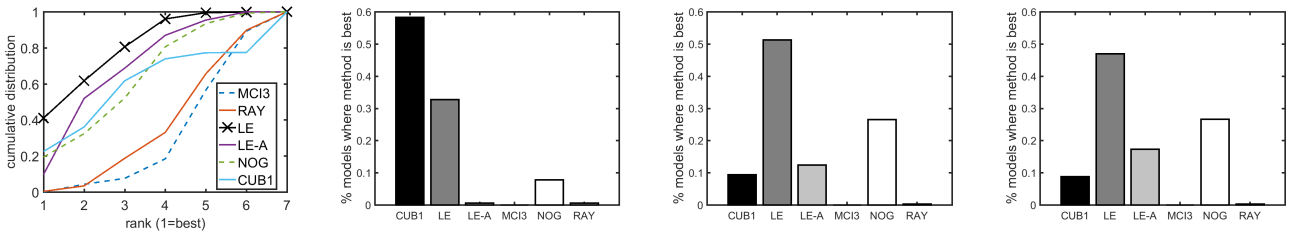
Tables 5 and 6 present the results of these experiments. Since CUB3, CUB5 and CUB7 fail on over 98% of the large instances due to floating-point range exceptions, the corresponding entries are omitted from the table. The results indicate that CUB1 remains the best method under small population sizes, however as N grows the asymptotic expansions become the most accurate. The RAY method is the least accurate in light load, but has a similar accuracy to the other methods in high-load. This is indeed the regime that matches the assumptions for the scaling used in RAY [27]; we have noted however that on models with a large number of classes, but a few queues, RAY is less accurate than LE and LE-A, which is consistent with the fact that the scaling used in [27] assumes a growing number of nodes. For example, going from $K = 16$ to $K = 64$ nodes RAY improves its error from an average of 6.82% to 1.28%. On the opposite, when the number of classes is increased from $R = 16$ to $R = 64$, RAY goes from an average error of 0.99% to 7.22%. In the same ranges for nodes and classes, LE and LE-A have narrow error bands between 0.46% and 1.23% average error.

5.3.3 *Models with infinite server nodes*

We have repeated the experiments in Section 5.3.1 on models with a infinite server node, focusing on the validation of the heuristic given in Section 4.4.1. Since with infinite servers RAY is no longer applicable, we have validated LE against the PANACEA (PAN) asymptotic expansion [35]. We have set in the experiments an identical think time on all classes equal to $\sigma_r \in \{0.1\theta_{max}, \theta_{max}, 10\theta_{max}, 100\theta_{max}\}$, where $\theta_{max} = \max_r \max_{k=1,\ldots,K} \theta_{kr}$. Results are shown in Table 4. PAN fails on all models with think time $\sigma_r = 0.1\theta_{max}$ and $\sigma_r = \theta_{max}$ due to violation of the normal usage assumption; it instead returns a 26.2% MAPE with $\sigma_r = \theta_{max}$, and an error less than 0.1% with $\sigma_r = 10\theta_{max}$. LE returns a valid solution in all cases, with decreasing errors for increasing σ_r values. Thus, LE appears generally more robust than PAN, which is preferable only in very lightly loaded models.

5.4 Optimization programs

We now compare the methods against the likelihood maximization problem (5) for service demand estimation, focusing on deterministic methods. For the sake of illustration of the limited performance of randomized methods in this setting, we also include MCI3 in the validation. We consider problems with $K, R \in \{2, 4, 8, 16, 32\}$ and populations with $N/R \in \{2, 20, 40\}$ jobs per



(a) $N/R \in \{2, 20, 40\}$ (b) $N/R = 2$ (c) $N/R = 20$ (d) $N/R = 40$

Figure 2: Optimization programs: experimental results

Table 7: Percentage of failed runs (1125 models).

Method	Total	$N/R = 2$	$N/R = 20$	$N/R = 40$
CUB1	22.4	0.0	26.9	40.3
RAY	8.8	2.7	16.0	7.7

class. Each experiment is carried out with the same procedure described in Section 2.3, in particular setting a timeout of $T = 10$ minutes. We repeat the same experiment 15 times randomizing demands, solving 1125 optimization programs for each method. Upon detecting an invalid normalizing constant the interior point method returns a failure. We also mark as failed all the runs returning demands that do not satisfy the constraint $\boldsymbol{\theta} \geq \mathbf{0}$.

5.4.1 *Metric*

As mentioned in Section 2.3, a critical issue in the analysis of the results is that (5) is non-convex, thus the choice of the initial points affects the relative error on the $\boldsymbol{\theta}$ estimate, irrespective of the quality of the approximation of $G_{\boldsymbol{\theta}}(\boldsymbol{N})$. This is addressed in Section 2.3 by comparing results against an exact method, run without timeout. Here we cannot apply the same approach due to the size of the models and also MCI with a large number of samples ($J = 10^7$) fails since the variance of the estimator adversely affects the search direction of the interior-point method. Thus, on most models it does not seem possible to determine and compare the methods based on an absolute accuracy metric.

To cope with this problem, we use the same initial point for all the methods and compare them relatively to each other. For each model, we rank methods based on the absolute percentage error from the true value of $\boldsymbol{\theta}$. In this way, a method that achieves the best possible estimator given the initial point, will be ranked first, irrespective of the magnitude of the error that depends on the initial point and the local optimum found. Methods that return the same demands are assigned the same rank.

5.4.2 *Results*

Results are given in Figure 2. We include in the study also the NOG method, which neglects the normalizing constant. In Figures 2(b)-(d) we show how frequently each model is ranked best for a given population level N. The results indicate that LE outperforms all the other methods and it is slightly better than LE-A. However, in models with a small number of jobs CUB is preferable, which is consistent with the observations in Section 5.3.1. The fair performance of NOG is explained by the fact that asymptotically the closed network approaches an open network, where the arrival rate intensity matches the cumulative departure rate from the bottleneck nodes. The methods proposed in this paper remain preferable to NOG, as they are the best ones in most models. This is evident in Figure 2(a), which indicates that LE is the best method among the considered ones. The figure shows that about 41% of the times LE is the best method, and in about 62% of the cases it ranks second, typically behind CUB1, NOG, or LE-A. Table 7 reports statistics on the number of failed models, which occur only for CUB as the load grows and for RAY, similarly to what seen for small and medium-sized models. Methods not shown in the table do not incur failures.

5.4.3 *Single-class models*

Lastly, for completeness we include results concerning inference in single-class models, which also arise in practice. For such models we do not study the computation of a single normalizing constant since exact $\mathcal{O}(1)$ expressions are available, e.g. (13). We instead consider likelihood maximization with $K \in \{2, 4, 8, 16, 32\}$ nodes, $R = 1$ class, $N = \{2, 20, 40\}$, and single-class demands $\theta_k = k/K$. The initial guess for the demand matrix is $\theta_k = 1/K$, $\forall k$. We include in the study the exact-order asymptotic (EOA) formula recently proposed in [19].

Generally speaking, exact methods such as CA are very fast on single-class models, and return a tiny error, on average just 0.058%. However, CA complexity is $\mathcal{O}(N)$, thus in models with very large populations N approximations may be of interest. In the above study, LE returns a MAPE of just 0.28%. The other methods are instead rather inaccurate, with a MAPE of 28.1% for CUB1, 26.0% for RAY, 24.8% for LE-A, and 66.1% for EOA. This overall indicates that LE is fit for use also in single-class problems, although exact methods such as CA are likely sufficient in practice.

5.5 Summary

Summarizing, the results indicate the following main properties for the proposed algorithms:

- On small models, CUB dominates all other algorithms.
- On large models, one should choose CUB if the population is small, or otherwise prefer LE. The same criteria applies to optimization programs involving likelihoods.
- Among randomized methods, LS is the best on models with many classes, whereas MCI is best with several nodes.

6. CONCLUSION

This paper has shown that performance inference over closed systems faces computational hurdles. If the model is described by a closed product-form network, we have addressed such issues by novel asymptotic expansions and Monte Carlo sampling methods for the normalizing constant of state probabilities. Future research may further investigate the implications of the integral form (6) for the exact theory of normalizing constants, for example on models with multi-server and load-dependent nodes.

7. REFERENCES

[1] J. Aitchison. The statistical analysis of compositional data. *J. Royal Stat. Society. Series B.*, 139–177, 1982.

[2] J. Anselmi, P. Cremonesi. A unified framework for the bottleneck analysis of multiclass queueing networks. *Perform. Eval.*, 67(4):218-234, 2010.

[3] K. E. Atkinson. *An Introduction to Numerical Analysis*. John Wiley & Sons, 2nd ed., 1989.

[4] V. Baldoni, *et al*. How to integrate a polynomial over a simplex. *Math. of Computation*, 80(273):297–325, 2011.

[5] A. Grundmann, H.M. Möller. Invariant integration formulas for the n-simplex by combinatorial methods. *SIAM Journal on Numerical Analysis*, 15(2):282–290, 1978.

[6] G. Balbo, G. Serazzi. Asymptotic analysis of multiclass closed queueing networks: Multiple bottlenecks. *Perform. Eval.*, 30(3):115–152, 1997.

[7] F. Baskett, *et al.*. Open, closed, and mixed networks of queues with different classes of customers. *JACM*, 22:248–260, 1975.

[8] A. W. Berger, L. M. Bregman, and Y. Kogan. Bottleneck analysis in multiclass closed queueing networks and its application. *QUESTA*, 31(3-4):217–237, 1999.

[9] A. Bertozzi, J. McKenna. Multidimensional residues, generating functions, and their application to queueing networks. *SIAM Review*, 35(2):239–268, 1993.

[10] J. P. Buzen. Computational algorithms for closed queueing networks with exponential servers. *Comm. of the ACM*, 16(9):527–531, 1973.

[11] G. Casale. Exact analysis of performance models by the method of moments. *Perf. Eval.*, 68(6):487–506, 2011.

[12] K. M. Chandy, U. Herzog, L. Woo. Parametric Analysis of Queueing Networks. *IBM J. Res. Dev.*, 19(1):36–42, 1975.

[13] G L. Choudhury, K. K. Leung, and W. Whitt. Calculating normalization constants of closed queuing networks by numerically inverting their generating functions. *JACM*, 42(5):935–970, 1995.

[14] R. Cools. An encyclopaedia of cubature formulas. *Journal of Complexity*, 19:445–453, 2003.

[15] A. E. Conway, N. D. Georganas. RECAL - A new efficient algorithm for the exact analysis of multiple-chain closed queueing networks. *JACM*, 33(4):768–791, 1986.

[16] C. de Boor. Divided differences. *Surveys in Approximation Theory*, 1:46–49, 2005.

[17] L. Etxeberria, *et al.*. Performance-based Selection of Software and Hardware Features under Parameter Uncertainty. *Proc. of QoSA*, 23–32, 2014.

[18] C. Fletcher. Innovation Insight for Algorithmic IT Operations Platforms. Gartner report G00296380, 24 March 2016.

[19] D. K. George, C. H. Xia, and M. S. Squillante. Exact-order asymptotic analysis for closed queueing networks. *J. Applied Probability*, 49(2):503–520, 2012.

[20] A. I. Gerasimov. On normalizing constants in multiclass queueing networks. *Oper. Res.*, 43(4):704–711, 1995.

[21] H. W. Gould. Some Generalizations of Vandermonde's Convolution. *The American Mathematical Monthly*, 63(2):84–91, 1956.

[22] J. J. Gordon. The evaluation of normalizing constants in closed queueing networks. *Oper. Res.*, 38(5):863–869, 1990.

[23] P. G. Harrison. On normalizing constants in queueing networks. *Oper. Res.*, 33(2):464–468, 1985.

[24] P. G. Harrison, T. T. Lee. A new recursive algorithm for computing generating functions in closed queueing networks. In *Proc. of IEEE MASCOTS*, 223–230. IEEE Press, 2004.

[25] R. E. Kass, L. Tierney, and J. B. Kadane. The validity of posterior expansions based on Laplace's method. *Bayesian and likelihood methods in stat. and econ.*, 7:473, 1990.

[26] F. P. Kelly, L. Massoulié, and N. S. Walton. Resource pooling in congested networks: proportional fairness and product form. *QUESTA*, 63(1-4):165–194, 2009.

[27] C. Knessl, C. Tier. Asymptotic expansions for large closed queueing networks with multiple job classes. *IEEE Trans. Computers*, 41(4):480–488, 1992.

[28] C. Knessl, C. Tier. Asymptotic approximations and bottleneck analysis in product form queueing networks with large populations. *Perf. Eval.*, 33(4):219–248, 1998.

[29] H. Kobayashi. A computational algorithm for queue distributions via the Pólya theory of enumeration. *Perf. of Computer Systems*, North-Holland, 1979, pp. 79-88.

[30] E. Koenigsberg. Cyclic queues. *Operational Research Quarterly*, 9, 1:22–35, 1958.

[31] Y. Kogan. Asymptotic expansions for probability distributions in large loss and closed queueing networks. *Perform. Eval. Rev.*, 29(3):25–27, Dec. 2001.

[32] Y. Kogan, M. Shenfild. Asymptotic solution of generalized multiclass Engset model. In *Proc. of ITC*, 1239–1249, 1994.

[33] Y. Kogan, A. Yakovlev. Asymptotic analysis for closed multichain queueing networks with bottlenecks. *QUESTA*, 23:235–258, 1996.

[34] J. B. Lasserre, E. S. Zeron. A Laplace transform algorithm for the volume of a convex polytope. *JACM*, 48(6), 2001.

[35] J. McKenna, D. Mitra. Asymptotic expansions and integral representations of moments of queue lengths in closed Markovian networks. *JACM*, 31(2):346–360, 1984.

[36] L. M. Milne-Thomson. *The calculus of finite differences*. Mac Millan, London, 1933.

[37] Y. Nazarathy, P. K. Pollett. *Parameter and State Estimation in Queues and Related Stochastic Models*. Technical Report, School of Math. and Physics, Univ. of Queensland, Nov 2012.

[38] Y. Pawitan. *In All Likelihood - Statistical Modeling and Inference Using Likelihood.* Oxford Science, 2001.

[39] M. Reiser, H. Kobayashi. Queueing networks with multiple closed chains. *IBM J. Res. Dev.*, 19(3):283–294, 1975.

[40] M. Reiser, S. S. Lavenberg. Mean-value analysis of closed multichain queueing networks. *JACM*, 27(2):312–322, 1980.

[41] K.W. Ross, D.H.K. Tsang and J. Wang. Monte carlo summation and integration applied to multiclass queueing networks. *JACM*, 41(6):1110–1135, 1994.

[42] P. Schweitzer. Approximate Analysis of Multiclass Closed Networks of Queues. In *Proc. of the Int'l Conf. on Stoch. Control and Optim.*, 25–29, 1979.

[43] S. Spinner, *et al.* Evaluating approaches to resource demand estimation. *Perf. Eval.*, 92:51–71, 2015.

[44] C. Sutton, M. Jordan. Bayesian inference for queueing networks and modeling of internet services. *Annals of Applied Stat.*, 5(1), 254–282, 2011.

[45] W. Wang, G. Casale, C. Sutton. A Bayesian Approach to Parameter Inference in Queueing Networks. *ACM TOMACS*, 27(1), 2016.

APPENDIX

A. DIVIDED DIFFERENCES

We show that $g_{\boldsymbol{t\theta}}(N) = [\theta_1, \dots, \theta_K]x^{N+K-1}$ holds for arbitrary demands $\theta_1 \leq \theta_2 \leq \cdots \leq \theta_K$. In the case of distinct demands, the result follows by [36, Sec. 1.31], [5]. Otherwise note that divided differences may be recursively defined as follows [16]

$$[\theta_1, \dots, \theta_K]f(x) = \begin{cases} \frac{f^{(K-1)}(\theta_1)}{(K-1)!} & \text{if } \theta_1 = \theta_K \\ \frac{[\theta_2,\dots,\theta_K]f(x) - [\theta_1,\dots,\theta_{K-1}]f(x)}{\theta_K - \theta_1} & \text{otherwise} \end{cases}$$

where $f^{(n)}(x)$ denotes the nth derivative of $f(x)$.

First note the following relation [23]

$$\begin{aligned} g_{\boldsymbol{t\theta}}(N) &= \frac{\theta_K g^{-1}_{\boldsymbol{t\theta}}(N) - \theta_1 g^{-K}_{\boldsymbol{t\theta}}(N)}{\theta_K - \theta_1} \\ &= \frac{g^{-1}_{\boldsymbol{t\theta}}(N+1) - g^{-K}_{\boldsymbol{t\theta}}(N+1)}{\theta_K - \theta_1} \end{aligned} \tag{35}$$

where $g^{-i}_{\boldsymbol{t\theta}}$ refers to a network without node i and we have used that

by convolution $g^{-1}_{\boldsymbol{t\theta}}(N+1) = g^{-1-K}_{\boldsymbol{t\theta}}(N+1) + \theta_K g^{-1}_{\boldsymbol{t\theta}}(N)$ and $g^{-K}_{\boldsymbol{t\theta}}(N+1) = g^{-1-K}_{\boldsymbol{t\theta}}(N+1) + \theta_1 g^{-K}_{\boldsymbol{t\theta}}(N)$, see e.g. [10].

We can now show that the recursive definition of divided differences is a special instance of (35). Note that if $f(x) = x^{N+K-1}$ then the case $\theta_1 \neq \theta_K$ readily matches (35). Upon reaching $\theta_1 = \theta_K$ in (35) we have a model with balanced demands, in which case [22, Eq. 19]: $g_{\boldsymbol{t\theta}}(N) = \binom{N+K-1}{K-1}\theta_1^N$. The last expression matches $\frac{f^{(K-1)}(\theta_1)}{(K-1)!}$ when $f(x) = x^{N+K-1}$.

B. HERMITE-GENOCCHI

We want to show that if $f(x) = x^{N+K-1}$ the Hermite-Genocchi theorem (10) holds also under nondistinct demands, i.e.

$$[\theta_1, \dots, \theta_K]x^{N+K-1} = \frac{(N+K-1)!}{N!}\int_{\Delta_K} (\theta_1 u_1 + \dots + \theta_K u_K)^N d\boldsymbol{u}$$

for arbitrary $\theta_1, \dots, \theta_K$. Using the multinomial theorem (4)

$$[\theta_1, \dots, \theta_K]x^{N+K-1} = (N+K-1)!\int_{\Delta_K} \sum_{\substack{\boldsymbol{n}\geq 0:\\ n=N}} \prod_{i=1}^{K} \frac{\theta_i^{n_i} u_i^{n_i}}{n_i!} d\boldsymbol{u}$$

By the normalizing constant formula of the Dirichlet distribution

$$\int_{\Delta_K} \prod_{k=1}^{K} u_k^{n_k} d\boldsymbol{u} = \frac{\prod_{k=1}^{K} n_k!}{(\sum_{k=1}^{K} n_k + K - 1)!} \tag{36}$$

we thus find the identity proved in Appendix A

$$[\theta_1, \dots, \theta_K]x^{N+K-1} = \sum_{\substack{\boldsymbol{n}\geq 0:\\ n=N}} \prod_{i=1}^{K} \theta_i^{n_i} = g_{\boldsymbol{t\theta}}(N)$$

where we have noted that the sum is the specialization of (1) to single-class models.