Accelerating Performance Inference over Closed Systems by Asymptotic Methods

Giuliano Casale
Imperial College London
London SW7 2AZ, 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 analysis. 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 exactly as a multidimensional integral over the unit simplex. This gives as a by-product novel explicit expressions for the multiclass normalizing constant. We then derive a 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.

ACM Reference format:
Giuliano Casale. 2017. Accelerating Performance Inference over Closed Systems by Asymptotic Methods. In *Proceedings of SIGMETRICS '17, June 5–9, 2017, Urbana-Champaign, IL, USA, ,* 1 pages.
DOI: http://dx.doi.org/10.1145/3078505.3078514

1 MODEL

We consider closed queueing networks with K single-server nodes, $M - K$ infinite server nodes, R job classes, and product-form solution [1]. Let N_r be the number of jobs in class r and let $\boldsymbol{N} = (N_1, \dots, N_R)$, $N = \sum_r N_r$. Let $\boldsymbol{\theta} = [\theta_{kr}]$ collect the service demands placed by class-r jobs at node k. The model has state space $\mathcal{S}_M = \{\boldsymbol{n} \in \mathbb{N}^{MR} \mid n_{kr} \geq 0, \sum_{k=1}^{M} n_{kr} = N_r\}$, where n_{kr} is the number of class-r jobs at node k. The normalizing constant of the state probabilities is given by [1]

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

We study how to efficiently obtain $G_{\boldsymbol{\theta}}(\boldsymbol{N})$, motivated by the problem of computing likelihoods for model parameterization, model selection, and statistical inference.

2 MAIN RESULTS

For networks without infinite servers, we show in [2] that an exact integral form for the normalizing constant is given by

$$G_{\boldsymbol{\theta}}(\boldsymbol{N}) = \frac{(N+K-1)!}{N_1! \cdots N_R!} \int_{\Delta_K} \prod_{r=1}^{R} \left(\sum_{k=1}^{K} \theta_{kr} u_k \right)^{N_r} d\boldsymbol{u} \tag{2}$$

where $\Delta_K = \{\boldsymbol{u} \in \mathbb{R}^K \mid u_i \geq 0, \sum_i u_i = 1\}$ is the unit simplex. Stemming from this novel integral form, [2] proposes exact and approximate computational methods for $G_{\boldsymbol{\theta}}(\boldsymbol{N})$. For example, using cubature rules [3], (2) can be interpolated, either exactly or approximately, as

$$G_{\boldsymbol{\theta}}(\boldsymbol{N}) = \frac{(N+K-1)!}{\prod_{s=1}^{R} N_s!} \sum_{i=0}^{S} w_i \sum_{\substack{\boldsymbol{b} \geq \boldsymbol{0}: \\ b = S-i}} \prod_{r=1}^{R} \left(\sum_{j=1}^{K} \frac{(2b_j+1)\theta_{jr}}{(2S+K-2i)} \right)^{N_r} \tag{3}$$

where S specifies the interpolation degree, $\boldsymbol{b} \in \mathbb{N}^K$, $b = \sum_{i=1}^{K} b_i$, and $w_i = (-1)^i 2^{-2S} (2S+K-2i)^{2S+1} / (i!(2S+K-i)!)$. Expression (3) requires $O(S^K)$ time and $O(1)$ space as N grows, and can be computed exactly by setting $S = \lceil (N-1)/2 \rceil$.

Further, following a logistic transformation of the integrand of (2), it is possible to apply Laplace's method and obtain a $O(N^{-1})$ asymptotic expansion for $G_{\boldsymbol{\theta}}(\boldsymbol{N})$. This requires to solve a system of nonlinear equations, similar to the queue-length equations used in mean-value analysis. The derivation of the expansion also yields a Monte Carlo sampling method, which trades computational cost for accuracy in computing $G_{\boldsymbol{\theta}}(\boldsymbol{N})$. The above results extend to networks with infinite server nodes.

A numerical validation in [2] shows that cubature rules are very effective on small models, whereas the proposed asymptotic expansion is the most effective method on large models. Monte Carlo sampling is found to improve state-of-the-art algorithms in the case of models with several classes.

REFERENCES

[1] F. Baskett, K. M. Chandy, R. R. Muntz, F. G. Palacios. Open, closed, and mixed networks of queues with different classes of customers. *JACM*, 22:248–260, 1975.
[2] G. Casale. Accelerating Performance Inference over Closed Systems by Asymptotic Methods. In *Proc. ACM Meas. Anal. Comput. Syst. (POMACS)* 1:1, June 2017. Preprint: https://spiral.imperial.ac.uk/handle/10044/1/43431. DOI: http://dx.doi.org/10.1145/3084445.
[3] A. Grundmann, H.M. Möller. Invariant integration formulas for the n-simplex by combinatorial methods. *SIAM J. on Numerical Analysis*, 15(2):282–290, 1978.

This research has been partially funded by the European Union's Horizon 2020 research and innovation programme (644869) and by a UK EPSRC grant (EP/L00738X/1).

SIGMETRICS '17, June 5–9, 2017, Urbana-Champaign, IL, USA

DOI: http://dx.doi.org/10.1145/3078505.3078514