Perturbation analysis of two queues with random time-limited polling

M. Saxena[a], S. Kapodistria[b] and R. Núñez Queija[c,d]
[a,b] *Eurandom and Department of Mathematics and Computer Science*
Eindhoven University of Technology, The Netherlands
[a]m.mayank@tue.nl, [b]s.kapodistria@tue.nl
[c]*Korteweg-de Vries Institute, University of Amsterdam, The Netherlands*
[d]*CWI, Amsterdam, The Netherlands*

March 22, 2019

Abstract

Perturbation analysis has proven to be a fruitful technique in the analysis of several multi-dimensional Markov models. In order to successfully apply perturbation analysis, one must find an appropriate scaling parameter. This leaves the approach with a large degree of freedom. In this paper we explore the use of perturbation analysis to determine the stationary distribution of a specific polling model with random time-limited service. We discuss four candidate scaling approaches to analyze the polling model and describe the difficulties that must be overcome to construct a computationally feasible algorithm.
Keywords: Singular perturbation, analytic expansion, time-limited polling.

1 Introduction

Multi-dimensional Markov models, and specifically in the analysis of queueing systems, often prove to be hard to analyze. In this paper we focus on the characterization of the stationary distribution of a specific polling model with random time-limited service, by means of perturbation analysis. Perturbation analysis was successfully used in a wide range of applications. Depending on the application field, the specific modeling approach and the object of interest, different terminology has been used for variations of perturbation analysis, including aggregation-disaggregation analysis [8], decomposability [6], and time-scale separation and quasi-stationary analysis [1, 4, 5, 10]. In these papers the focus is essentially on the leading term in perturbation analysis. A modern treatise of analytic expansions of transient and stationary measures can be found in Avrachenkov et al. [3] In the queueing theory literature the power series algorithm, see e.g.[9], is a variation of perturbation analysis that has been very successful.

In this paper we investigate the use of singular perturbation for computation of the stationary distribution of a two-dimensional Markov model of a specific polling model with time-limited service. In previous work [7] we have established that the target stationary distribution satisfies a boundary value problem, which we have not been able to solve. Our aim is to use perturbation

analysis to develop a computation scheme for the stationary distribution. We use four different parameter scalings and for each of these we discuss the difficulties that are encountered when used for numerical computations.

The structure of the paper is as follows: We first describe the model in Section 2. A brief explanation of perturbation analysis in the context of our model follows in Section 3. In Section 4 we discuss one of the scaling approaches in some detail, while we sketch three other scalings in Section 5. We conclude in Section 6.

2 Model description and notation

We consider a two-queue polling model. Customers arrive to queues $i = 1, 2$, according to independent Poisson processes with rates λ_i. There is a single server that serves both queues. The service times of customers in both queues are exponentially distributed with rates μ_1 and μ_2, respectively. The server spends an exponentially distributed time with rate c_i, $i = 1, 2$ at queue i, after which it instantaneously switches to the other queue (there is no switch-over time).
Stability condition. The necessary and sufficient stability condition is [7]

$$\frac{\lambda_1}{\mu_1} < \frac{c_2}{c_1 + c_2} \text{ and } \frac{\lambda_2}{\mu_2} < \frac{c_1}{c_1 + c_2}. \tag{1}$$

With $Q_i(t)$ we denote the queue length at time t of queue $i, i = 1, 2$, and $S(t) \in \{1, 2\}$ is the queue where the server is serving at time t. The stochastic process $\{Q(t) = (Q_1(t), Q_2(t), S(t)) : t \geq 0\}$ with state space $\{\mathcal{S} = \{(n_1, n_2, i) : n_1, n_2 \in \mathbb{N}_0, i = 1, 2\}$ is a CTMC. Let

$$\pi(n_1, n_2, i) = \lim_{t\to\infty} \mathbb{P}(Q_1(t) = n_1, Q_2(t) = n_2, S(t) = i),$$

$$\text{and } \pi((n_1, n_2(|i) = \lim_{t\to\infty} \mathbb{P}(Q_1(t) = n_1, Q_2(t) = n_2|S(t) = i).$$

The equilibrium equations for the defined model are: for $n_1, n_2 \geq 0$,

$$\begin{aligned}(\lambda_1 + \lambda_2 + c_1 + \mu_1 1_{(n_1\geq 1)})\pi(n_1, n_2, 1) &= \mu_1\pi(n_1 + 1, n_2, 1) + 1_{(n_2\geq 1)}\lambda_2\pi(n_1, n_2 - 1, 1) \\ &\quad + 1_{(n_1\geq 1)}\lambda_1\pi(n_1 - 1, n_2, 1) + c_2\pi(n_1, n_2, 2). \qquad (2)\\ (\lambda_1 + \lambda_2 + c_2 + \mu_2 1_{(n_2\geq 1)})\pi(n_1, n_2, 2) &= \mu_2\pi(n_1, n_2 + 1, 2) + 1_{(n_1\geq 1)}\lambda_1\pi(n_1 - 1, n_2, 2) \\ &\quad + 1_{(n_2\geq 1)}\lambda_2\pi(n_1, n_2 - 1, 2) + c_1\pi(n_1, n_2, 1). \qquad (3)\end{aligned}$$

A direct analytic computation of the joint queue length distribution (or its probability generating function (PGF)) turns out to be as challenging as the workload analysis presented in [7, Sec. 6]. In what follows we explore the use of parametric perturbation under different scaling schemes, and also discuss the advantages and disadvantages of each particular scheme.

3 Perturbation analysis

We perturb some of the parameters of the model by scaling them by a common parameter ε. There is a rich choice in which parameters to perturb. We demonstrate how this choice affects the nature and the complexity of the underlying solution. Furthermore, for each specific choice, we

demonstrate how to derive the leading term, and how to obtain recursive schemes to compute all following terms.

More precisely, we scale some parameters (to be specified for each of the four options) by ε and reflect the dependence of the perturbed model on ε in our notation, for example $\pi((n_1, n_2, \varepsilon)|i)$ for the conditional queue-length probabilities. Assuming the perturbed queue length distribution has a Taylor series expansion

$$\pi((n_1, n_2, \varepsilon)|i) = \sum_{k=0}^{\infty} \varepsilon^k \pi^k((n_1, n_2)|i), \quad i = 1, 2, \tag{4}$$

we are then left with the computation of the coefficients in this expansion. The choice of scaling should be such that the limiting distribution, as $\varepsilon \to 0$ can be determined, and that $\pi^k((n_1, n_2)|i)$ can be determine recursively from the equilibrium equations (2) and (3) after the scaling. The original joint queue length distribution can be retrieved by setting $\varepsilon = 1$.
In this article we consider four different scaling schemes. The first scaling scheme is the least trivial and is discussed in more detail than the other schemes.

4 Scaling 1: queue 1 with short visits and infrequent arrivals

We scale the rates λ_1, μ_1 and c_2 as $\lambda_1 \to \varepsilon^2\lambda_1$, $\mu_1 \to \varepsilon\mu_1$ and $c_2 \to \varepsilon c_2$. The parameters that are not perturbed are μ_2, λ_2 and c_1. We will see that this choice directly leads to recursions for $\pi^k((n_1, n_2)|i), i = 1, 2$ with an explicit leading term when $\varepsilon \to 0$. The quadratic term plays an important role; in the scalings to be discussed later, we will see that without the quadratic term, the recursions have a more complex form.

Substituting the perturbed parameters into (2) and (3) and rearranging the terms we obtain

$$\begin{aligned}
\varepsilon\pi((n_1, n_2, \varepsilon)|1) =& \varepsilon\frac{\lambda_2}{c_1 + \lambda_2}\pi((n_1, n_2 - 1, \varepsilon)|1)1_{(n_2\geq 1)} + \varepsilon\frac{c_1}{c_1 + \lambda_2}\pi((n_1, n_2, \varepsilon)|2) \\
&+ \varepsilon^2\frac{\mu_1}{c_1 + \lambda_2}[\pi((n_1 + 1, n_2, \varepsilon)|1) - \pi((n_1, n_2, \varepsilon)|1)1_{(n_1\geq 1)}] \\
&- \varepsilon^3\frac{\lambda_1}{c_1 + \lambda_2}[\pi((n_1, n_2, \varepsilon)|1) - \pi((n_1 - 1, n_2, \varepsilon)|1)1_{(n_1\geq 1)}], \qquad (5) \\
\pi((n_1, n_2 + 1, \varepsilon)|2) =& \left[1_{(n_2\geq 1)} + \frac{\lambda_2}{\mu_2}\right]\pi((n_1, n_2, \varepsilon)|2) - \frac{\lambda_2}{\mu_2}\pi((n_1, n_2 - 1, \varepsilon)|2)1_{(n_2\geq 1)} \\
&- \varepsilon\frac{c_2}{\mu_2}\left[\pi((n_1, n_2, \varepsilon)|1) - \pi((n_1, n_2, \varepsilon)|2)\right] \\
&+ \varepsilon^2\frac{\lambda_1}{\mu_2}\left[\pi((n_1, n_2, \varepsilon)|2) - \pi((n_1 - 1, n_2, \varepsilon)|2)1_{(n_1\geq 1)}\right]. \qquad (6)
\end{aligned}$$

Substituting (4) and computing the coefficient of ε^{k+1} from (5) and ε^k from (6), yields the required

recursions of the perturbed queue length distributions, for $k \geq 0$,

$$\begin{aligned}\pi^k((n_1,n_2)|1) =& \frac{\lambda_2}{c_1+\lambda_2}\pi^k((n_1,n_2-1)|1)1_{(n_2\geq 1)} + \frac{c_1}{c_1+\lambda_2}\pi^k((n_1,n_2)|2) \\ &+ \frac{\mu_1}{c_1+\lambda_2}\Big[\pi^{k-1}((n_1+1,n_2)|1) - \pi^{k-1}((n_1,n_2)|1)1_{(n_1\geq 1)}\Big]1_{(k\geq 1)} \\ &- \frac{\lambda_1}{c_1+\lambda_2}\left[\pi^{k-2}((n_1,n_2)|1) - \pi^{k-2}((n_1-1,n_2)|1)1_{(n_1\geq 1)}\right]1_{(k\geq 2)}, \qquad (7)\\ \pi^k((n_1,n_2+1)|2) =& \left[1_{(n_2\geq 1)} + \frac{\lambda_2}{\mu_2}\right]\pi^k((n_1,n_2)|2) - \frac{\lambda_2}{\mu_2}\pi^k((n_1,n_2-1)|2)1_{(n_2\geq 1)} \\ &- \frac{c_2}{\mu_2}\left[\pi^{k-1}((n_1,n_2)|1) - \pi^{k-1}((n_1,n_2)|2)\right]1_{(k\geq 1)} \\ &+ \frac{\lambda_1}{\mu_2}\left[\pi^{k-2}((n_1,n_2)|2) - \pi^{k-2}((n_1-1,n_2)|2)1_{(n_1\geq 1)}\right]1_{(k\geq 2)}. \qquad (8)\end{aligned}$$

To iterate the recursions (7) and (8) we need $\pi^k((n_1,0)|2)$. The recursion for $\pi^k((n_1,0)|2))$ is derived from (8). Let $\pi^k((n_1,\cdot)|2)$ denote the coefficient of ε^k of the marginal queue length distribution $\pi((n_1,\cdot,\varepsilon)|2)$. Multiplying (8) by n_2 and then summing over n_2 from 0 to ∞, we obtain

$$\begin{aligned}\pi^k((n_1,0)|2) =& \left(1-\frac{\lambda_2}{\mu_2}\right)\pi^k((n_1,\cdot)|2) - \frac{c_2}{\mu_2}\Big[\sum_{n_2=1}^{\infty} n_2\pi^{k-1}((n_1,n_2)|1) - \sum_{n_2=1}^{\infty} n_2\pi^{k-1}((n_1,n_2)|2)\Big]1_{(k\geq 1)} \\ &+ \frac{\lambda_1}{\mu_2}\Big[\sum_{n_2=1}^{\infty} n_2\pi^{k-2}((n_1,n_2)|2) - \sum_{n_2=1}^{\infty} n_2\pi^{k-2}((n_1-1,n_2)|2)1_{(n_1\geq 1)}\Big]1_{(k\geq 2)}. \qquad (9)\end{aligned}$$

There is still an unknown term $\pi^k((n_1,\cdot)|2)$ in the above equation, which is the coefficient of ε^k of the steady-state marginal queue length distribution $\pi^k((n_1,\cdot,\varepsilon)|2)$, which is known [7, Sec. 3] and leads to the recursion

$$\begin{aligned}\pi^k((n_1,\cdot)|2) =& \frac{\lambda_1 c_1}{\mu_1 c_2}\pi^k((n_1-1,\cdot)|2)1_{(n_1\geq 1)} + \left[\left(1-\frac{\lambda_1 c_1}{\mu_1 c_2}\right)1_{(k=0)} - \frac{\lambda_1}{\mu_1}1_{(k=1)}\right]1_{(n_1=0)} \\ &- \frac{\lambda_1}{c_2}\pi^{k-1}((n_1,\cdot)|2)1_{(k\geq 1)} + \frac{\lambda_1}{\mu_1}\left(1+\frac{\mu_1}{c_2}\right)\pi^{k-1}((n_1-1,\cdot)|2)1_{(n_1\geq 1)}1_{(k\geq 1)} \\ &+ \frac{\lambda_1^2}{\mu_1 c_2}\Big[\pi^{k-2}((n_1-1,\cdot)|2)1_{(n_1\geq 1)} - \pi^{k-2}((n_1-2,\cdot)|2)1_{(n_1\geq 2)}\Big]1_{(k\geq 2)}. \qquad (10)\end{aligned}$$

Finally by combining (8), (9) and (10), we get a complete recursion for $\pi^k((n_1,n_2)|2)$, as well as a recursion for $\pi^k((n_1,n_2)|1)$ from (7).

Computation of the first terms of the recursions. Calculating the results from (8), (9) for $k=0$ yields

$$\pi^0((n_1,n_2)|2) = \left(1-\frac{\lambda_1 c_1}{\mu_1 c_2}\right)\left(\frac{\lambda_1 c_1}{\mu_1 c_2}\right)^{n_1}\left(1-\frac{\lambda_2}{\mu_2}\right)\left(\frac{\lambda_2}{\mu_2}\right)^{n_2}. \qquad (11)$$

Setting $k=0$ in (7), iterating it for n_2 gives

$$\pi^0((n_1,n_2)|1) = \frac{c_1}{c_1+\lambda_2}\left(\frac{\mu_2}{c_1+\lambda_2}\right)^{n_2}\frac{1-\left(\frac{c_1+\lambda_2}{\mu_2}\right)^{n_2+1}}{1-\left(\frac{c_1+\lambda_2}{\mu_2}\right)}\pi^0((n_1,n_2)|2). \qquad (12)$$

We have now arrived at an explicit recursion for the coefficients of the expansion. Unfortunately the above analysis does not come with a guarantee on the numerical stability if we wish to take $\varepsilon \to 1$. This drawback is common to many applications of perturbation analysis. Obtaining theoretical lower bounds on the radius of convergence of the expansion often lead to very crude bounds (and therefore of little use), see e.g. [2, p. 846]. For our model we conducted numerical experiments, which suggest that the radius of convergence in general does not extend to $= 1$. A better understanding of the radius of convergence is therefore needed to make this approach useful for approximations.

5 Three alternative scalings

In this section we briefly address three other scaling approaches and - similar to the previous scaling - we discuss the challenges encountered.

5.1 Scaling 2

In this scaling, we scale the rates $\lambda_1, \lambda_2, \mu_2$ and c_2 as $\varepsilon\lambda_1, \varepsilon\lambda_2, \varepsilon\mu_2$ and εc_2. We first focus on the initial solution as $\varepsilon \to 0$ in (4). In this limit, one can observe that the server is always serving the second queue, which implies that the second queue behaves as an M/M/1 queueing system and also independent of the first queue. Hence the joint queue length distribution of the system is the product of the distribution of the first and second queue as $\varepsilon \to 0$. The queue length distribution of the separate queue can be computed with some effort. Next step is to derive the other term of series expansion (4). The recursions for those terms can be obtained from (2) and (3) using (4), though it is not obvious to see how the initial solutions can be used to obtain $\pi^k((n_1, n_2)|i)$ for $k \geq 1$. But it does not exclude the possibility of using the initial solutions to derive the other terms in a non-trivial way.

5.2 Scaling 3

In this scaling, we scale the residing times c_i as εc_i for $i = 1, 2$. This scaling is an interesting scaling because in the limit $\varepsilon \to 0$, the considered model behaves as a two-dimensional Markov modulated fluid queues. Unfortunately, the computations the limiting distribution as $\varepsilon \to 0$ turns out to be as challenging as the workload analysis presented in [7, Sec. 6]. This could be a topic for further investigation.

5.3 Scaling 4

In this scaling, we perturb the service and arrival rates as $\varepsilon\mu_i$ and $\varepsilon\lambda_i$ for $i = 1, 2$, respectively. This scaling has been discussed in detail by us in [7, Sec. 7]. In this paper, we propose a singular perturbation analysis for the calculation of the joint queue length distribution of the perturbed Markov chain. We have shown that the steady-state joint queue length distribution is written as a Taylor series expression in terms of ε, whose coefficients form a geometric sequence, that can be used for both exact and numerical calculations. Furthermore, we have shown that there exists a computationally stable updating formula [7, Eqn. (7.15)] for the calculation of the perturbed

steady-state joint queue length distribution. The only issue in this scaling is that, to approximate the joint distribution numerically, one needs to solve a large system of equations, for which we need to truncate the state space.

6 Conclusion

Perturbation analysis has proven to be a fruitful technique in the analysis of several multi-dimensional Markov models. In order to successfully apply perturbation analysis, one must find an appropriate scaling parameter in general. In this paper we explored the use of perturbation analysis to determine the stationary distribution of a specific polling model with random time-limited service. Exact analysis of our specific model has proven extremely challenging, thus making it natural to resort to perturbation analysis. We discussed four candidate scaling approaches to analyze the polling model. In all four cases, we succeeded to reduce the complexity of the original problem, but were not able to establish a numerically reliable computation scheme. For each of the four scaling approaches we described the difficulties that must be overcome to accomplish this goal. Our work is meant to serve as an intermediate step in various directions, to eventually facilitate reliable computations of stationary performance measures.

References

[1] Eitan Altman, Damien Artiges, and Karim Traore. On the integration of best-effort and guaranteed performance services. *European Transactions on Telecommunications*, 10(2):125–134, 1999.

[2] Eitan Altman, Konstantin E Avrachenkov, and Rudesindo Núñez-Queija. Perturbation analysis for denumerable markov chains with application to queueing models. *Advances in Applied Probability*, 36(3):839–853, 2004.

[3] K.E. Avrachenkov, J.A. Filar, and P.G. Howlett. *Analytic Perturbation Theory and Its Applications*. Other titles in applied mathematics. Society for Industrial and Applied Mathematics, 2013.

[4] Thomas Bonald and Alexandre Proutière. On performance bounds for the integration of elastic and adaptive streaming flows. In *ACM SIGMETRICS Performance Evaluation Review*, volume 32, pages 235–245. ACM, 2004.

[5] Thomas Bonald and Alexandre Proutière. On stochastic bounds for monotonic processor sharing networks. *Queueing Systems*, 47(1-2):81–106, 2004.

[6] Pierre Jacques Courtois. *Decomposability: queueing and computer system applications*. Academic Press, 1977.

[7] Mayank Saxena, Onno Boxma, Stella Kapodistria, and Rudesindo Núnez Queija. Two queues with random time-limited polling. *Probability and Mathematical Statistics*, 37(2):257–289, 2017.

[8] Herbert A Simon and Albert Ando. Aggregation of variables in dynamic systems. *Econometrica: journal of the Econometric Society*, pages 111–138, 1961.

[9] W.B. Van den Hout. *The Power Series Algorithm, A Numerical Approach to Markov Processes*. PhD. Thesis Tilburg University, 1996.

[10] Gijs Van Kessel, Rudesindo Núnez-Queija, and Sem Borst. Differentiated bandwidth sharing with disparate flow sizes. In *Proceedings IEEE 24th Annual Joint Conference of the IEEE Computer and Communications Societies.*, volume 4, pages 2425–2435. IEEE, 2005.