12th IFAC/IEEE Workshop on Discrete Event Systems
Cachan, France. May 14-16, 2014

Series expansion techniques for fast evaluation of acyclic finite-capacity queueing networks

Dieter Fiems * **Koen De Turck** *

* *Department of Telecommunications and Information Processing. Ghent University. St-Pietersnieuwstraat 41, 9000 Gent, Belgium (e-mail: {df,kdeturck}@telin.ugent.be).*

Abstract: An efficient numerical evaluation technique is presented for families of finite Markov chains indexed by a parameter ϵ, that possess a monotonicity property for the chain at $\epsilon = 0$. The monotonicity allows for retrieving the series expansion of the steady state solution in $\epsilon = 0$ in $O(NM)$ operations, where N is the size of the state space and M is the number of terms in the series expansion. We then apply the method to a layered queueing network with four queues, three queues in tandem and one queue for the servers operating this tandem queue.

Keywords: Taylor-series expansions, Markov chains, queueing systems, performance evaluation, light-traffic.

1. INTRODUCTION

Let us assume that there exists a ϵ_m such that for every $0 \leq \epsilon < \epsilon_m$, $C_\epsilon = A + \epsilon B$ is the generator matrix of a finite Markov chain $\{X_\epsilon(t)\}_{t>0}$. Assuming that $X_0(t)$ has at most one ergodic class, it is well known that the steady state probability vector $\pi(\epsilon)$ of the chain is analytic in $\epsilon = 0$. That is, there is an expansion of the form,

$$\sum_{n=0}^{\infty} \pi_n \epsilon^n ,$$

which converges to $\pi(\epsilon)$ in a neighbourhood of 0. As $\pi(\epsilon)(A + \epsilon B) = 0$, one immediately find that the terms of this expansion adhere,

$$\pi_0 A = 0 , \quad \pi_{n+1} A = \pi_n B .$$

By the normalisation condition, the elements of π_0 and π_n ($n \in \mathbb{N}^+$) sum up to 1 and 0, respectively. Hence, π_0 is the normalised Perron-Frobenius eigenvector of the generator matrix A and,

$$\pi_{n+1} = \pi_n B A^{\#} ,$$

where $A^{\#} = (e'\pi_0 - A)^{-1} + e'\pi_0$ is the generalised inverse of A (with e' a column vector of ones).

When the state-space of the family of Markov chains is relatively small, the steady-state solution of the Markov chain can be obtained by direct methods. In such a case one may calculate the terms of the series expansion to assess the sensitivity to some parameter ϵ of the steady state solution, and associated performance measures. In contrast, when the state-space of the Markov chain is very large, say of size M, a direct calculation of $\pi(\epsilon)$ is not feasible for $\epsilon > 0$ as this takes $O(M^3)$ operations. At first sight, the series expansion does not allow for mitigating the computational burden of the large state space as the equations for the series expansions have an equal numerical complexity. However, if we assume that the chain at $\epsilon = 0$ only contains transitions in one direction (either up or down), the matrix A is triangular and the numerical complexity reduces to $O(M^2)$ at most per iteration. For many practical applications, the number of possible transitions from a state is far smaller than the size of the state space and does not scale with the size of the state space, such that the numerical complexity for calculating the first N terms of the expansion is only $O(MN)$ in total. This shows that it is possible to calculate the terms of the series expansion fast if one constrains A. Assessing the series expansion approach as an accurate performance analysis tool is the subject of this paper.

Series expansion methods go by different names, including perturbation techniques, the power series method and light-traffic analysis. While the naming is not absolute, perturbation methods are mainly motivated by sensitivity analysis of the results with respect to some system parameter. The case where the perturbation does not preserve the class-structure of the non-perturbed chain — the so-called singular perturbations — has received much attention in literature (Altman et al., 2004; Laserre, 1994). Recent research also addresses the accuracy of perturbations by means of perturbation bounds (Heidergott et al., 2009, 2010). The power series method transforms a Markov chain of interest in a set of Markov chains parametrised by a variable γ. For $\gamma = 0$, the chain is not only easily solved, but one can also obtain the series expansion in γ. For $\gamma = 1$ one gets the original Markov chain such that the series expansion can be used to approximate the solution of the original Markov chain, provided the convergence region of the series expansion includes $\gamma = 1$ (van den Hout, 1996). Finally, light-traffic analysis often corresponds to a series expansion in the arrival rate at a queue. For an overview on the technique of series expansions in stochastic systems, we further refer the reader to the surveys in (Błaszczyszyn et al., 1995) and (Kovalenko, 1994).

To assess series expansion techniques as a numerical evaluation tool for families of Markov chains with a large (but finite) state space, we here focus on a queueing network where customers of some queues act as servers for other queues (Perel and Yechiali, 2010). Such queueing systems can e.g. be motivated by applications in wireless networks with limited connectivity. In such networks, wireless terminals rely on other wireless terminals to relay information to a base station. That is, some wireless terminals act as servers for other wireless terminals. The tandem network introduced in section 2 not only has a large state space, but also allows for investigating different types of series expansion techniques. Two types of expansions are considered. Section 3 considers a perturbation as explained above, where the matrix A is indeed triangular. Mitigating the triangularity assumption, section 4 considers a perturbation where A is block triangular.

Using perturbation as an analysis technique is not new, both light-traffic analysis and the power series methods (cfr. super) being prime examples of such techniques. The use of perturbation as a numerical analysis technique for queueing systems – say, in the spirit of matrix-analytic methods for queueing systems as explored by Neuts and others — is fairly recent. The technique has been successfully applied to coupled queueing systems (De Cuypere et al., 2014), processor sharing queues (Fiems and De Turck, 2013), epidemic processes (De Cuypere et al., 2013) and in the context of opportunistic scheduling (Evdokimova et al., 2014).

2. QUEUEING MODEL

We consider the Markovian queueing network of figure 1. There are three finite-capacity queues in tandem, say queue Q_1 to Q_3, and one queue Q_s of servers. Let λ_i and C_i, $i \in \mathcal{K} = \{1, 2, 3, s\}$, denote the input rate in Q_i and the queue capacity of Q_i, respectively. Further, let μ_s be the output rate of Q_s. Service in the queues of the tandem is delivered by the customers in Q_s. Let μ_i be the service rate delivered to queue i by a single server such that the rate from queue i is $\mu_i x_s$ when there are x_s customers in the server queue. Moreover, we assume blocking: there is no service in Q_1 when Q_2 is full, and no service in queue 2 when Q_3 is full. We have here chosen a very simple scheme to divide server capacity amongst the queues, but the method outlined below also allows for studying more complex schemes. For example, the available servers can be divided amongst the queues according to discriminatory or generalised processor sharing disciplines, we can assume that a distinct server can only attend one queue at a time, etc.

Let $\mathcal{C}_i = \{0, 1, \ldots, C_i\}$ such that the state space of the queueing system under consideration is $\mathcal{C} = \mathcal{C}_1 \times \mathcal{C}_2 \times \mathcal{C}_3 \times \mathcal{C}_s$. Let $\mathbf{x} = (x_1, x_2, x_3, x_s) \in \mathcal{C}$, and let $\mathbf{e}_i = [\mathbb{1}_{\{j=i\}}]_{,j\in\mathcal{K}}$ (for $i \in \mathcal{K}$), we then get the following balance equation,

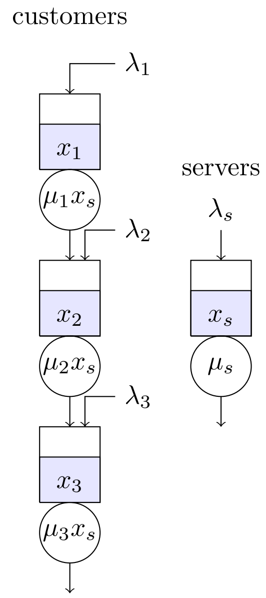

Fig. 1. Queueing system where the customers in the server queue are serving the customers in the tandem queue.

$$\begin{aligned}\pi(\mathbf{x})\Big(\sum_{i\in\mathcal{K}} \lambda_i \mathbb{1}_{\{x_i<C_i\}} + \mu_s \mathbb{1}_{\{x_s>0\}} \\ + \mu_1 x_s \mathbb{1}_{\{x_1>0, x_2<C_2\}} + \mu_2 x_s \mathbb{1}_{\{x_2>0, x_3<C_3\}} \\ + \mu_3 x_s \mathbb{1}_{\{x_3>0\}}\Big) = \sum_{i\in\mathcal{K}} \pi(\mathbf{x}-\mathbf{e}_i)\lambda_i \\ + \pi(\mathbf{x}+\mathbf{e}_s)\mu_s + \pi(\mathbf{x}+\mathbf{e}_1-\mathbf{e}_2)\mu_1 x_s \\ + \pi(\mathbf{x}+\mathbf{e}_2-\mathbf{e}_3)\mu_2 x_s + \pi(\mathbf{x}+\mathbf{e}_3)\mu_3 x_s\,. \quad (1)\end{aligned}$$

for $\mathbf{x} \in \mathcal{C}$ and where we assumed $\pi(\mathbf{x}) = 0$ for $\mathbf{x} \notin \mathcal{C}$ to simplify notation.

3. PERTURBATION I

As a first example, we scale the arrival rates in the customer queues and the service rate in the server queue: $\lambda_i = \epsilon\alpha_i$, $i \in \{1, 2, 3\}$ and $\mu_s = \epsilon\beta_s$, and consider the series expansion around $\epsilon = 0$. For $\epsilon = 0$, there are no arrivals in Q_1 to Q_3 and no departures from Q_s. Hence, there is a single absorbing state $(0, 0, 0, C_s)$ and the framework of regular perturbation applies.

Let $\pi_n(\mathbf{x})$ denote the nth term in the series expansion in λ_s of $\pi(\mathbf{x})$,

$$\pi(\mathbf{x}) = \sum_{n=0}^{\infty} \pi_n(\mathbf{x})\epsilon^n\,.$$

Plugging the expansion in the balance equation (1) and comparing terms in ϵ^n yields,