*Adv. Appl. Prob.* **42,** 1102–1125 (2010)
*Printed in Northern Ireland*


# APPROXIMATIONS TO QUASI-BIRTH-AND-DEATH PROCESSES WITH INFINITE BLOCKS

NIGEL BEAN,* *University of Adelaide*
GUY LATOUCHE,** *Université Libre de Bruxelles*

**Abstract**

The numerical analysis of quasi-birth-and-death processes rests on the resolution of a matrix-quadratic equation for which efficient algorithms are known when the matrices have finite order, that is, when the number of phases is finite. In this paper we consider the case of infinitely many phases from the point of view of theoretical convergence of truncation and augmentation schemes, and we develop four different methods. Two methods rely on forced transitions to the boundary. In one of these methods, the transitions occur as a result of the truncation itself, while in the other method, they are artificially introduced so that the augmentation may be chosen to be as natural as possible. Two other methods rely on forced transitions within the same level. We conclude with a brief numerical illustration.

*Keywords:* Quasi-birth-and-death process; infinite-dimensional matrix; truncation and augmentation; approximations; matrix-analytic method

2010 Mathematics Subject Classification: Primary 60J10; 60J22

## 1. Introduction

Take a quasi-birth-and-death (QBD) process with infinitely many phases: this is a two-dimensional process $\{X_n = (L_n, \varphi_n)\}_{n\geq 0}$ on the state space $\mathbb{N} \times S$ with $|S| = \infty$ and with transition matrix

$$\boldsymbol{P} = \begin{bmatrix} \boldsymbol{B} & \boldsymbol{A}_1 & \boldsymbol{0} & \cdots \\ \boldsymbol{A}_{-1} & \boldsymbol{A}_0 & \boldsymbol{A}_1 & \ddots \\ \boldsymbol{0} & \boldsymbol{A}_{-1} & \boldsymbol{A}_0 & \ddots \\ \vdots & \ddots & \ddots & \ddots \end{bmatrix}. \tag{1.1}$$

The blocks $\boldsymbol{B}$, $\boldsymbol{A}_{-1}$, $\boldsymbol{A}_0$, and $\boldsymbol{A}_1$ are nonnegative matrices with infinitely many rows and columns. The components $L$ and $\varphi$ of $X$ are respectively called the level and the phase. For the sake of simplicity, we assume that $S$ is the set of strictly positive integers.

The stationary probability vector $\boldsymbol{\pi}$ is decomposed as $\boldsymbol{\pi} = (\boldsymbol{\pi}_n)_{n\geq 0}$, with each subvector $\boldsymbol{\pi}_n = (\pi_{n,j})_{j\in S}$ having infinitely many components. It is known (see [6, Chapter 6]) that, if the QBD is positive recurrent then its stationary probability distribution has the matrix-geometric form $\boldsymbol{\pi}_n^\top = \boldsymbol{\pi}_0^\top \boldsymbol{R}^n$ for $n \geq 0$, where $\boldsymbol{R}$, called the rate matrix, is the minimal nonnegative solution of the equation

$$\boldsymbol{Y} = \boldsymbol{A}_1 + \boldsymbol{Y}\boldsymbol{A}_0 + \boldsymbol{Y}^2\boldsymbol{A}_{-1} \tag{1.2}$$

Received 26 January 2010; revision received 30 June 2010.
* Postal address: Department of Applied Mathematics, University of Adelaide, SA 5005, Australia.
Email address: nigel.bean@adelaide.edu.au
** Postal address: Département d'Informatique, Université Libre de Bruxelles, Campus Plaine, CP 212, Boulevard du Triomphe, B-1050 Bruxelles, Belgium.

and $\boldsymbol{\pi}_0$ is the unique solution of

$$\boldsymbol{x}^\top(\boldsymbol{B} + \boldsymbol{R}\boldsymbol{A}_{-1}) = \boldsymbol{x}^\top, \tag{1.3}$$

normalized by $\boldsymbol{x}^\top \sum_{\nu\geq 0} \boldsymbol{R}^\nu \mathbf{1} = 1$.

In some circumstances, the transitions between the levels 0 and 1 are different from those between the levels $n$ and $n+1$ for $n \geq 1$. Then, the transition matrix has the structure

$$\boldsymbol{P} = \begin{bmatrix} \boldsymbol{B} & \boldsymbol{B}_1 & \mathbf{0} & \cdots \\ \boldsymbol{B}_{-1} & \boldsymbol{A}_0 & \boldsymbol{A}_1 & \ddots \\ \mathbf{0} & \boldsymbol{A}_{-1} & \boldsymbol{A}_0 & \ddots \\ \vdots & \ddots & \ddots & \ddots \end{bmatrix}, \tag{1.4}$$

the stationary distribution is given by $\boldsymbol{\pi}_n = \boldsymbol{\pi}_1 \boldsymbol{R}^{n-1}$ for $n \geq 1$, with $\boldsymbol{R}$ as above, and the initial subvectors are given by

$$[\boldsymbol{\pi}_0^\top, \boldsymbol{\pi}_1^\top] \begin{bmatrix} \boldsymbol{B}_0 & \boldsymbol{B}_1 \\ \boldsymbol{B}_{-1} & \boldsymbol{A}_0 + \boldsymbol{R}\boldsymbol{A}_{-1} \end{bmatrix} = [\boldsymbol{\pi}_0^\top, \boldsymbol{\pi}_1^\top],$$
$$\boldsymbol{\pi}_0^\top \mathbf{1} + \boldsymbol{\pi}_1^\top \sum_{\nu\geq 0} \boldsymbol{R}^\nu \mathbf{1} = 1.$$

If the number of phases is finite then it would be easy to solve (1.2) by, for example, the logarithmic or cyclic reduction algorithms described in [6, Chapter 8] and [1, Chapter 7]. However, when the number of phases is infinite, we may at best compute a finite-dimensional approximation of $R$ and $\boldsymbol{\pi}_0$, and the question is how to do this.

The problem of approximating the stationary distribution of Markov chains with an infinite state space and without a particular structure is not new. Often, one starts from the *truncated* matrix ${}_k\boldsymbol{P}$ formed from a subset of the elements of the transition matrix $\boldsymbol{P}$. As matrix ${}_k\boldsymbol{P}$ is substochastic, the problem is to transform it into a stochastic matrix ${}_k\tilde{\boldsymbol{P}}$ in such a way that the stationary distribution ${}_k\boldsymbol{\pi}$ of ${}_k\tilde{\boldsymbol{P}}$ converges to that of $\boldsymbol{P}$ as $k$ tends to $\infty$; this is called the *augmentation* step.

From a computational perspective, there is no difficulty with QBD processes evolving in an infinite state space, because there is no restriction on the values for the level: it is the number $|S|$ of different values for the phase which has to be finite, and sufficiently small, for calculations to be feasible. Thus, it is on the phase dimension that we need to impose a constraint, and we will define the truncated matrices ${}_k\boldsymbol{P}$ as

$${}_k\boldsymbol{P} = \begin{bmatrix} {}_k\boldsymbol{B} & {}_k\boldsymbol{A}_1 & \mathbf{0} & \cdots \\ {}_k\boldsymbol{A}_{-1} & {}_k\boldsymbol{A}_0 & {}_k\boldsymbol{A}_1 & \ddots \\ \mathbf{0} & {}_k\boldsymbol{A}_{-1} & {}_k\boldsymbol{A}_0 & \ddots \\ \vdots & \ddots & \ddots & \ddots \end{bmatrix}, \tag{1.5}$$

where ${}_k\boldsymbol{B}$, ${}_k\boldsymbol{A}_{-1}$, ${}_k\boldsymbol{A}_0$, and ${}_k\boldsymbol{A}_1$ are the $k$ by $k$ northwest corners of the matrices $\boldsymbol{B}$, $\boldsymbol{A}_{-1}$, $\boldsymbol{A}_0$, and $\boldsymbol{A}_1$ respectively. In Section 2 we show with simple probabilistic arguments that the rate matrices ${}_k\boldsymbol{R}$ of the truncated QBDs automatically converge from below to the rate matrix of the original process. This falls short of our aim, since the stationary distribution depends not only on $\boldsymbol{R}$, but also on the density at level 0. It does, however, point to the important role that the *structure* of the QBD may play, even though, as we show in Example 4.1, the structure by itself is not enough.

https://doi.org/10.1239/aap/1293113153 Published online by Cambridge University Press

Let us temporarily forget about the QBD structure of our processes and let $\boldsymbol{P}$ be the transition matrix of an arbitrary Markov chain on the integers. We assume that it is irreducible and positive recurrent, so that its stationary probability vector $\boldsymbol{\pi}$ is well defined. Gibson and Seneta [3] took ${}_k\boldsymbol{P}$ to be the $k$ by $k$ northwest corner of $\boldsymbol{P}$ and discussed various ways of constructing a stochastic matrix ${}_k\tilde{\boldsymbol{P}}$.

With *last column* augmentation, the missing mass on each row of ${}_k\boldsymbol{P}$ is added to the last column; thus, ${}_k\tilde{\boldsymbol{P}} = {}_k\boldsymbol{P} + (\boldsymbol{I} - {}_k\boldsymbol{P})\mathbf{1} \cdot \boldsymbol{e}_k^\top$, where $\boldsymbol{I}$ is the identity matrix, $\mathbf{1}$ is a vector of 1s, and $\boldsymbol{e}_k$ is the vector of size $k$ with a 1 in $k$th position and 0s elsewhere. The resulting sequence of stationary probability vectors, unfortunately, does not always converge: one interesting case is given in [3], which inspired our Example 4.1.

In *fixed column* augmentation, ${}_k\tilde{\boldsymbol{P}} = {}_k\boldsymbol{P} + (\boldsymbol{I} - {}_k\boldsymbol{P})\mathbf{1} \cdot \boldsymbol{f}_m^\top$, where $\boldsymbol{f}_m$ has size $k$, with 0s everywhere except for a 1 in position $m$ fixed once and for all; often, $m = 1$, in which case we speak of *first column* augmentation. With fixed column augmentation, the sequence $\{{}_k\boldsymbol{\pi}\}_{k \geq m}$ of stationary distributions does converge to $\boldsymbol{\pi}$ (see [3, Theorem 3.1]).

A useful property is for $\boldsymbol{P}$ to be a *Markov matrix*, that is, for $\boldsymbol{P}$ to have one column bounded away from 0. Assuming without loss of generality that this is the first column, we have $P_{i,1} \geq \epsilon > 0$ for some $\epsilon$ and all $i$. In this case, any augmentation leads to a sequence of stationary distributions which converge to $\boldsymbol{\pi}$ (see [3, Theorem 2.1]).

In fact, convergence of the sequence of approximating stationary distributions has been related to the behavior of expected first passage times to finite subsets of states in [11] and also in [4], where it was most clearly shown. We illustrate this in Sections 3 and 5. In Section 3 we adapt to QBDs the simplest 'augment the first column' scheme and we force a transition to state $(0, 1)$ whenever the process attempts to move to some state $(n, i)$ with $i \geq k + 1$. We show how Theorem 1 of [4], presented herein as Theorem 3.1, is immediately applicable. In Section 5 we transform the QBD in such a way that its transition matrix becomes a Markov matrix. To do this, we introduce an artificial event called a 'catastrophe', which has the effect of restarting the process in the state $(0, 1)$. Here, again, one easily shows convergence of the modified processes.

Naturally, we would prefer to modify the process in such a way as to better preserve its dynamics, instead of repeatedly restarting the system from scratch. We show in Section 6 that it is actually possible to preserve the QBD structure of the transition matrix, and merely to modify the phase. If the process attempts to move from $(n, i)$ to some phase $j$ with $j > k$, we force it to move to phase 1 in the same level, or in one of its neighboring levels, $n - 1$ or $n + 1$. Compared to the augmentations of Sections 3 and 5, this clearly reduces the disruption to the behavior of the process.

As we show in Theorem 6.1, this gives a sequence of approximations which converges under suitable conditions to the stationary distribution of the original process. The proof is based on an extension of the arguments in [4] and [11] to a Markov regenerative process with an infinite regenerative set of states. Our last scheme is a blend of those in Sections 5 and 6: we introduce an artificial event (as in Section 5) which affects the phase only (as in Section 6). We conclude in Section 7 with a few numerical examples to illustrate and compare the various procedures.

We restrict our attention to QBDs in discrete time, noting that continuous-time QBDs may be treated in exactly the same manner if the transition rates are uniformly bounded, by using their uniformized discrete-time version. We assume without loss of generality (see [7]) that the

https://doi.org/10.1239/aap/1293113153 Published online by Cambridge University Press

matrix $\boldsymbol{P}$ in (1.1) is irreducible and that the doubly infinite matrix

$$\boldsymbol{P}_D = \begin{bmatrix} \ddots & \ddots & \ddots & \vdots & \iddots \\ \ddots & \boldsymbol{A}_0 & \boldsymbol{A}_1 & \boldsymbol{0} & \cdots \\ \ddots & \boldsymbol{A}_{-1} & \boldsymbol{A}_0 & \boldsymbol{A}_1 & \ddots \\ \cdots & \boldsymbol{0} & \boldsymbol{A}_{-1} & \boldsymbol{A}_0 & \ddots \\ \iddots & \vdots & \ddots & \ddots & \ddots \end{bmatrix} \tag{1.6}$$

is irreducible as well. Matrices with the structure of (1.1) will be called *QBD matrices* and will be characterized by their blocks $\boldsymbol{A}_{-1}$, $\boldsymbol{A}_0$, $\boldsymbol{A}_1$, and $\boldsymbol{B}$. Finally, in order to avoid technical difficulties, we assume that the augmented matrices are such that they have a unique stationary distribution.

## 2. Truncation of QBDs

The matrix $R$ in (1.2) may also be expressed as the series $\boldsymbol{R} = \boldsymbol{A}_1 \sum_{\nu\geq 0} \boldsymbol{U}^\nu$, with $\boldsymbol{U} = \boldsymbol{A}_0 + \boldsymbol{A}_1\boldsymbol{G}$, where $\boldsymbol{G}$ is the stochastic matrix of first passage probabilities from level 1 to level 0: $G_{ij} = \mathrm{P}[\theta < \infty,\ \varphi_\theta = j \mid X_0 = (1, i)]$ for $i$ and $j$ in $S$, with $\theta = \inf\{n \geq 1\colon L_n = 0\}$ being the first return time to level 0.

The series $\sum_{\nu\geq 0} \boldsymbol{U}^\nu$ always converges and its limit is equal to $(\boldsymbol{I} - \boldsymbol{U})^{-1}$ when $|S|$ is finite, while it is the minimal nonnegative solution of $(\boldsymbol{I} - \boldsymbol{U})\boldsymbol{X} = \boldsymbol{I}$ when $|S|$ is infinite.

The matrix $G$ is the minimal nonnegative solution of

$$\boldsymbol{Y} = \boldsymbol{A}_{-1} + \boldsymbol{A}_0\boldsymbol{Y} + \boldsymbol{A}_1\boldsymbol{Y}^2, \tag{2.1}$$

and it is related to $\boldsymbol{U}$ by $\boldsymbol{G} = \sum_{\nu\geq 0} \boldsymbol{U}^\nu \boldsymbol{A}_{-1}$, so that we may rewrite (1.3) as

$$\boldsymbol{x}^\top(\boldsymbol{B} + \boldsymbol{A}_1\boldsymbol{G}) = \boldsymbol{x}^\top. \tag{2.2}$$

Finally, with $N_j = \sum_{1\leq n\leq\theta} 1\{X_n = (1, j)\}$, where $1\{\cdot\}$ denotes the indicator function, being the number of visits to $(1, j)$ before the first return to level 0, $\boldsymbol{R}$ can be interpreted as a matrix of expected number of visits:

$$R_{ij} = \mathrm{E}[N_j \mid X_0 = (0, i)]$$

for all $i$ and $j$ in $S$.

Now, the process ${}_kX$ with transition matrix (1.5) exactly follows the evolution of the process $X$ with transition matrix (1.1) as long as the phase takes values between 1 and $k$, and it ceases to evolve as soon as the phase becomes strictly greater than $k$, that is, $X_n = {}_kX_n$ for $n < T_k$, where $T_k = \inf\{n \geq 0\colon \varphi_n \geq k + 1\}$. Thus, it is easy to see that the matrix ${}_k\boldsymbol{G}$, which we define as

$${}_kG_{ij} = \mathrm{P}[\theta < \infty,\ \theta < T_k,\ \varphi_\theta = j \mid X_0 = (1, i)] \qquad \text{for } i, j \leq k, \tag{2.3}$$

is the minimal nonnegative solution of

$$\boldsymbol{Y} = {}_k\boldsymbol{A}_{-1} + {}_k\boldsymbol{A}_0\boldsymbol{Y} + {}_k\boldsymbol{A}_1\boldsymbol{Y}^2,$$

and that ${}_kG_{ij}$ converges to $G_{ij}$ from below as $k \to \infty$ for any given $i$ and $j$, since $\lim_{k\to\infty} T_k = \infty$ almost surely (a.s.) for any given $X_0$.

https://doi.org/10.1239/aap/1293113153 Published online by Cambridge University Press

Similarly, the rate matrix

$$ {}_k\boldsymbol{R} = {}_k\boldsymbol{A}_1(\boldsymbol{I} - {}_k\boldsymbol{A}_0 - {}_k\boldsymbol{A}_1\,{}_k\boldsymbol{G})^{-1} \tag{2.4} $$

is also the minimal nonnegative solution of the equation

$$ \boldsymbol{Y} = {}_k\boldsymbol{A}_1 + \boldsymbol{Y}\,{}_k\boldsymbol{A}_0 + \boldsymbol{Y}^2\,{}_k\boldsymbol{A}_{-1} $$

and is such that ${}_kR_{ij} = \mathrm{E}[{}_kN_j \mid X_0 = (0, i)]$ for $i, j \le k$, where ${}_kN_j = \sum_{1\le n\le \min(\theta, T_k)} 1\{X_n = (1, j)\}$ is the number of visits to $(1, j)$ before the first return to level 0, or before the first passage time to a phase at least equal to $k+1$, whichever comes first. We immediately conclude that $\lim_{k\to\infty} {}_kR_{ij} = R_{ij}$ from below for any given $i$ and $j$. Thus, the truncation defined in (1.5) automatically provides us with a sequence of matrices which converges to $\boldsymbol{R}$. This, however, is only one of the ingredients of the matrix-geometric distribution, and we lack the equivalent of (2.2). If we used ${}_k\boldsymbol{P}$ as such, we would obtain the boundary matrix ${}_k\boldsymbol{B} + {}_k\boldsymbol{A}_{1k}\boldsymbol{G}$; however, since this matrix is substochastic, the system $\boldsymbol{x}^\top = \boldsymbol{x}^\top({}_k\boldsymbol{B} + {}_k\boldsymbol{A}_{1k}\boldsymbol{G})$ does not have a solution.

In the following sections we use a range of approaches to ensure a stochastic boundary condition.

## 3. First column augmentation

In the first approach, we apply one of the principles mentioned in the introduction and augment ${}_k\boldsymbol{P}$ by putting the missing mass on the column corresponding to the state $(0, 1)$. This gives ${}_k\tilde{\boldsymbol{P}} = {}_k\boldsymbol{P} + (\boldsymbol{I} - {}_k\boldsymbol{P})\boldsymbol{1}\cdot\boldsymbol{\eta}_1^\top$, where $\boldsymbol{\eta}_1^\top = [\boldsymbol{e}_1^\top, \boldsymbol{0}^\top, \boldsymbol{0}^\top, \ldots]$ and $\boldsymbol{e}_1^\top = [1, 0, \ldots, 0]$ is a vector of $k$ components. The matrix ${}_k\tilde{\boldsymbol{P}}$ has the structure

$$ {}_k\tilde{\boldsymbol{P}} = \begin{bmatrix} {}_k\tilde{\boldsymbol{B}} & {}_k\boldsymbol{A}_1 & \boldsymbol{0} & \boldsymbol{0} & \cdots \\ {}_k\tilde{\boldsymbol{B}}_{-1} & {}_k\boldsymbol{A}_0 & {}_k\boldsymbol{A}_1 & \boldsymbol{0} & \ddots \\ {}_k\boldsymbol{C} & {}_k\boldsymbol{A}_{-1} & {}_k\boldsymbol{A}_0 & {}_k\boldsymbol{A}_1 & \ddots \\ {}_k\boldsymbol{C} & \boldsymbol{0} & {}_k\boldsymbol{A}_{-1} & {}_k\boldsymbol{A}_0 & \ddots \\ \vdots & \vdots & \ddots & \ddots & \ddots \end{bmatrix}, \tag{3.1} $$

where

$$ {}_k\boldsymbol{C} = (\boldsymbol{I} - {}_k\boldsymbol{A}_{-1} - {}_k\boldsymbol{A}_0 - {}_k\boldsymbol{A}_1)\boldsymbol{1}\cdot\boldsymbol{e}_1^\top, $$
$$ {}_k\tilde{\boldsymbol{B}}_{-1} = {}_k\boldsymbol{A}_{-1} + {}_k\boldsymbol{C}, $$
$$ {}_k\tilde{\boldsymbol{B}} = {}_k\boldsymbol{B} + (\boldsymbol{I} - {}_k\boldsymbol{B} - {}_k\boldsymbol{A}_1)\boldsymbol{1}\cdot\boldsymbol{e}_1^\top. $$

The matrix ${}_k\tilde{\boldsymbol{P}}$ in (3.1) is a Markov chain of GI/M/1 type (see [8, Theorem 1.2.1]) and its stationary distribution has the form

$$ {}_k\boldsymbol{\pi}_n = {}_k\boldsymbol{\pi}_0({}_k\boldsymbol{R})^n, $$

where ${}_k\boldsymbol{R}$ is as defined in (2.4) and ${}_k\boldsymbol{\pi}_0$ is given by the system

$$ {}_k\boldsymbol{\pi}_0^\top({}_k\tilde{\boldsymbol{B}} + {}_k\boldsymbol{R}\,{}_k\tilde{\boldsymbol{B}}_{-1} + {}_k\boldsymbol{R}(\boldsymbol{I} - {}_k\boldsymbol{R})^{-1}\,{}_k\boldsymbol{C}) = {}_k\boldsymbol{\pi}_0^\top, \tag{3.2} $$
$$ {}_k\boldsymbol{\pi}_0^\top(\boldsymbol{I} - {}_k\boldsymbol{R})^{-1}\boldsymbol{1} = 1. $$

Using the special structure of (3.1), we may express the boundary equation (3.2) in a more transparent manner: assume that the process starts in level $n$ for some $n \ge 2$ and define $\tau$ as

https://doi.org/10.1239/aap/1293113153 Published online by Cambridge University Press

the first passage time to any lower level—this will be level $n-1$ or level 0. It is easy to verify that ${}_k\boldsymbol{G}$ defined in (2.3) gives the probability that the process goes down to level $n-1$ before level 0, independently of $n \geq 2$. The matrix ${}_k\boldsymbol{H}$ defined as

$$ {}_kH_{ij} = \mathrm{P}[\tau < \infty,\ X_\tau = (0, j) \mid X_0 = (n, i)] $$

gives the probability of jumping to level zero without visiting level $n-1$, independently of $n \geq 2$. For $X_0$ in level 1, since ${}_k\tilde{\boldsymbol{B}}_{-1} = {}_k\boldsymbol{A}_{-1} + {}_k\boldsymbol{C}$, it easily follows that $\mathrm{P}[\tau < \infty,\ X_\tau = (0, j) \mid X_0 = (1, i)] = {}_kG_{ij} + {}_kH_{ij}$. In this manner, we see that (3.2) may be expressed as

$$ {}_k\boldsymbol{\pi}_0^\top({}_k\tilde{\boldsymbol{B}} + {}_k\boldsymbol{A}_1({}_k\boldsymbol{G} + {}_k\boldsymbol{H})) = {}_k\boldsymbol{\pi}_0^\top. $$

By Lemma A.1 in Appendix A,

$$ {}_k\boldsymbol{H} = (\boldsymbol{I} - {}_k\boldsymbol{A}_0 - {}_k\boldsymbol{A}_{1k}\boldsymbol{G} - {}_k\boldsymbol{A}_1)^{-1}{}_k\boldsymbol{C}, $$

so that ${}_k\boldsymbol{\pi}_0$, and the whole distribution ${}_k\boldsymbol{\pi}$, are completely determined once ${}_k\boldsymbol{G}$ is known.

The question of convergence is rapidly dealt with. Consider general transition matrices $\boldsymbol{P}$ and $\{{}_k\tilde{\boldsymbol{Q}}\}_{k\in\mathbb{N}}$ on the same state space, with the property that $\lim_{k\to\infty}({}_k\tilde{\boldsymbol{Q}})_{ij} = P_{ij}$ for all states $i$ and $j$. Denote by $Z$ and ${}_k\tilde{Z}$ the processes with transition matrices $\boldsymbol{P}$ and ${}_k\tilde{\boldsymbol{Q}}$, respectively, and assume that they are all defined on the same probability space. Assume that $Z$ and ${}_k\tilde{Z}$ are regenerative, with cycle times $C$ and $\tilde{C}_k$, respectively. Define $T_k$ to be the first time when ${}_k\tilde{Z}$ differs from $Z$. Theorem 1 of [4], stated below, is expressed in terms of a queueing system, but it is quite general.

**Theorem 3.1.** ([4, Theorem 1].) *If, for all sufficiently large $k$,*

$$ 1\{\tilde{C}_k < T_k\} = 1\{C < T_k\} \quad \textit{and} \quad \tilde{C}_k 1\{\tilde{C}_k < T_k\} = C1\{C < T_k\} \quad \textit{a.s.}, $$

*and if*

$$ \lim_{k\to\infty} \mathrm{E}[\tilde{C}_k 1\{\tilde{C}_k \geq T_k\}] = 0, $$

*then $\lim_{k\to\infty}({}_k\boldsymbol{\pi})_i = \boldsymbol{\pi}_i$ for all $i$.*

In the sequel, we write $\lim_{k\to\infty} {}_k\boldsymbol{\pi} = \boldsymbol{\pi}$ as a shorthand notation to indicate that convergence holds for every fixed state.

Sufficient conditions are that a nonnegative random variable $V$ exists, with $\mathrm{E}[V] < \infty$, such that $\mathrm{E}[\tilde{C}_k 1\{\tilde{C}_k \geq T_k\}] \leq \mathrm{E}[(V + C)1\{C \geq T_k\}]$, or such that $\tilde{C}_k \leq V + C$ a.s.

Now, consider the Markov chains $X$ and ${}_k\tilde{X}$ with transition matrices (1.1) and (3.1), respectively. In order to apply Theorem 3.1, we expand the matrix on the right-hand side of (3.1) so that it is defined for all the states of the original QBD, with $({}_k\tilde{\boldsymbol{P}})_{(n,i),(n',i')} = 0$ if $i$ or $i'$, or both, is strictly greater than $k$.

Without loss of generality, we may assume that the initial state is $(0, 1)$. Then, the QBDs are regenerative processes with cycle durations $C$ and $\tilde{C}_k$ equal to the return time to $(0, 1)$. The first time when they differ, $T_k$, is the first passage time of $X$ to any of the states in $\{(n, j)\colon n \geq 0,\ j \geq k+1\}$; by our construction,

- if $C < T_k$ then $\tilde{C}_k = C$,
- if $C > T_k$ then $\tilde{C}_k = T_k$,

so that $\tilde{C}_k \leq C$ a.s. As mentioned above, this is a sufficient condition to apply Theorem 3.1 and to conclude that $\lim_{k\to\infty} {}_k\boldsymbol{\pi} = \boldsymbol{\pi}$.

https://doi.org/10.1239/aap/1293113153 Published online by Cambridge University Press

## 4. Block augmentation

First column augmentation has the obvious drawback that whenever the process attempts to move beyond phase $k$, the whole system is cleaned up and restarts. Of course, we would prefer to augment the transition matrix ${}_k\boldsymbol{P}$ in a manner that better preserves the process dynamics. In particular, we would prefer to modify individual blocks so as to keep the homogeneous QBD structure of (1.1). For instance, we might use

$$ {}_k\tilde{\boldsymbol{P}} = \begin{bmatrix} {}_k\tilde{\boldsymbol{B}} & {}_k\tilde{\boldsymbol{A}}_1 & \boldsymbol{0} & \cdots \\ {}_k\tilde{\boldsymbol{A}}_{-1} & {}_k\tilde{\boldsymbol{A}}_0 & {}_k\tilde{\boldsymbol{A}}_1 & \ddots \\ \boldsymbol{0} & {}_k\tilde{\boldsymbol{A}}_{-1} & {}_k\tilde{\boldsymbol{A}}_0 & \ddots \\ \vdots & \ddots & \ddots & \ddots \end{bmatrix}, \tag{4.1} $$

with the blocks

$$ {}_k\tilde{\boldsymbol{A}}_{-1} = {}_k\boldsymbol{A}_{-1}, \qquad {}_k\tilde{\boldsymbol{A}}_1 = {}_k\boldsymbol{A}_1, \quad \text{and} \quad {}_k\tilde{\boldsymbol{A}}_0 = {}_k\boldsymbol{A}_0 + \operatorname{diag}({}_k\boldsymbol{d}), \tag{4.2} $$

where $\operatorname{diag}(\cdot)$ is a diagonal matrix and

$$ {}_k\boldsymbol{d} = (\boldsymbol{I} - {}_k\boldsymbol{A}_{-1} - {}_k\boldsymbol{A}_0 - {}_k\boldsymbol{A}_1)\boldsymbol{1}. \tag{4.3} $$

With this choice of augmentation, we merely disable events which would move the system beyond phase $k$.

Alternatively, we might allow the system to move as close as possible to the forbidden phase and use the augmented blocks

$$ {}_k\tilde{\boldsymbol{A}}_\nu = {}_k\boldsymbol{A}_\nu + \boldsymbol{a}_\nu^{(k)} \cdot \boldsymbol{e}_k^\top, \tag{4.4} $$

with $(\boldsymbol{a}_\nu^{(k)})_i = \sum_{j \geq k+1} (\boldsymbol{A}_\nu)_{ij}$ for $i \leq k$; note that $\boldsymbol{a}_{-1}^{(k)} + \boldsymbol{a}_0^{(k)} + \boldsymbol{a}_1^{(k)} = {}_k\boldsymbol{d}$. This corresponds to accepting the change of level and replacing transitions from phase $i$ to phase $j$ by transitions from $i$ to $\min(j, k)$.

The problem with such schemes is that convergence is not guaranteed, as can be seen in the example below, inspired from [3].

**Example 4.1.** We choose $\lambda$ and $\mu$ such that $\mu > \lambda > 0$ and $\lambda + \mu < 1$, we define $\boldsymbol{A}_{-1} = \mu \boldsymbol{I}$, $\boldsymbol{A}_1 = \lambda \boldsymbol{I}$, and $\boldsymbol{A}_0 = (1 - \lambda - \mu)\boldsymbol{A}$, where $\boldsymbol{A}$ is an irreducible stochastic matrix, and we choose $\boldsymbol{B} = \boldsymbol{A}_{-1} + \boldsymbol{A}_0$.

In this setup, the value of the level clearly has no influence on the transition of the phase and vice versa; thus, the QBD possesses level-phase independence (see [9]) and the stationary distribution has a product form given by $(1 - \rho)\rho^n \boldsymbol{\alpha}$, where $\rho = \lambda/\mu$ and $\boldsymbol{\alpha}$ is the stationary vector of $\boldsymbol{A}$. The algebraic proof of this statement is given in Appendix B.

Now, let us take the transition matrix

$$ \boldsymbol{A} = \begin{bmatrix} q_1 & p_1 & 0 & \cdots \\ q_2 & 0 & p_2 & \ddots \\ q_3 & 0 & 0 & \ddots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} $$

https://doi.org/10.1239/aap/1293113153 Published online by Cambridge University Press

from [3], with $q_i = 1 - p_i$ for all $i$ and $p_i$ defined as

$$p_i = \begin{cases} \frac{1}{4} & \text{if } i = 1, \\ \dfrac{(i-1)^4}{((i-1)^2-1)(i+1)^2} & \text{for } i \equiv 1 \bmod N,\ i \neq 1, \\ \left(\dfrac{i}{i+1}\right)^2 & \text{for } i \equiv 2, \ldots, N-1 \bmod N, \\ 1 - \dfrac{1}{i^2} & \text{if } i \equiv 0 \bmod N, \end{cases}$$

for some fixed $N \geq 3$.

If, as suggested above, we augment the last column of each block (which is equivalent here to augmenting the diagonal), then ${}_k\tilde{\boldsymbol{A}}_{-1} = \mu \boldsymbol{I}$, ${}_k\tilde{\boldsymbol{A}}_1 = \lambda \boldsymbol{I}$, and ${}_k\tilde{\boldsymbol{A}}_0 = (1 - \lambda - \mu)_k\tilde{\boldsymbol{A}}$, where ${}_k\tilde{\boldsymbol{A}} = {}_k\boldsymbol{A} + (\boldsymbol{I} - {}_k\boldsymbol{A})\boldsymbol{1} \cdot \boldsymbol{e}_k^\top$. The stationary distribution of this QBD is given by ${}_k\boldsymbol{\pi}_n = (1-\rho)\rho^n {}_k\boldsymbol{\alpha}$, where ${}_k\boldsymbol{\alpha}$ is the stationary probability vector of ${}_k\tilde{\boldsymbol{A}}$. It was shown in [3] that ${}_k\boldsymbol{\alpha}$ does not converge to $\boldsymbol{\alpha}$ as $k \to \infty$, so ${}_k\boldsymbol{\pi}$ does not converge to $\boldsymbol{\pi}$ either.

In Section 6 we will apportion the missing mass on the first column of the three blocks, as opposed to the first column of the matrix ${}_k\boldsymbol{P}$ itself, without specifying a priori in which of the three. Then,

$${}_k\tilde{\boldsymbol{B}} = {}_k\boldsymbol{B} + ({}_k\boldsymbol{\delta}_{-1} + {}_k\boldsymbol{\delta}_0) \cdot \boldsymbol{e}_1^\top \tag{4.5}$$

and

$${}_k\tilde{\boldsymbol{A}}_\nu = {}_k\boldsymbol{A}_\nu + {}_k\boldsymbol{\delta}_\nu \cdot \boldsymbol{e}_1^\top \quad \text{for } \nu = -1, 0, 1, \tag{4.6}$$

where ${}_k\boldsymbol{\delta}_\nu \geq \boldsymbol{0}$, ${}_k\boldsymbol{\delta}_{-1} + {}_k\boldsymbol{\delta}_0 + {}_k\boldsymbol{\delta}_1 = {}_k\boldsymbol{d}$, and ${}_k\boldsymbol{d}$ is defined in (4.3). One possibility is to relocate the missing mass in the same block, in which case ${}_k\boldsymbol{\delta}_\nu = \boldsymbol{a}_\nu^{(k)}$, but circumstances may suggest other choices, as we show in Example 7.1.

Before examining the augmented blocks (4.5) and (4.6), however, we consider in the next section a scheme which allows us to use *any* block augmentation.

## 5. Artificial restart

As mentioned in the introduction, any augmentation will do if the transition matrix is *Markov*. This motivates us to suggest a procedure whereby we would perturb the original QBD in a way which guarantees convergence. In the theorem below, ${}_k\tilde{\boldsymbol{P}}$ is obtained from $\boldsymbol{P}$ by the usual truncation on the first $k$ phases, and augmentation by block augmentation without further constraints.

We can now use these ideas to prove the following theorem.

**Theorem 5.1.** *Consider the transition matrix $\boldsymbol{P}$ given by (1.1), with stationary distribution $\boldsymbol{\pi}$, and the stochastic QBD matrix ${}_k\tilde{\boldsymbol{P}}$ with blocks ${}_k\tilde{\boldsymbol{A}}_\nu \geq {}_k\boldsymbol{A}_\nu$ for $\nu = -1, 0, 1$ and ${}_k\tilde{\boldsymbol{B}} \geq {}_k\boldsymbol{B}$. Consider the transition matrix*

$${}_k\tilde{\boldsymbol{Q}}(\gamma) = \gamma \boldsymbol{e} \cdot \boldsymbol{\eta}_1^\top + (1-\gamma)_k\tilde{\boldsymbol{P}} \tag{5.1}$$

*and denote its stationary distribution by ${}_k\boldsymbol{\pi}(\gamma)$. We have*

$$\lim_{\gamma \to 0} \lim_{k \to \infty} {}_k\boldsymbol{\pi}(\gamma) = \boldsymbol{\pi}.$$

https://doi.org/10.1239/aap/1293113153 Published online by Cambridge University Press

**Remark 5.1.** Before we prove the theorem, we briefly comment on it. With the matrix ${}_k\tilde{\boldsymbol{P}}$ being obtained from $\boldsymbol{P}$ without constraint, it is possible that its stationary distribution does not converge to $\boldsymbol{\pi}$ as $k$ goes to $\infty$. Transformation (5.1) creates a Markov matrix, at the cost of introducing an artificial event, which we term a *catastrophe* and which occurs with probability $\gamma$, independently of the state of the system. This has the benefit that $\lim_{k\to\infty} {}_k\boldsymbol{\pi}(\gamma)$ is equal to the stationary distribution $\boldsymbol{\pi}(\gamma)$ of the matrix

$$\boldsymbol{Q}(\gamma) = \gamma \boldsymbol{e} \cdot \boldsymbol{\eta}_1^{\top} + (1-\gamma)\boldsymbol{P}. \tag{5.2}$$

We then show that $\boldsymbol{\pi}(\gamma)$ tends to $\boldsymbol{\pi}$ as $\gamma$ tends to 0.

*Proof of Theorem 5.1.* Without loss of generality, we assume that the initial state is (0, 1). Then, the processes with transition matrices ${}_k\tilde{\boldsymbol{Q}}(\gamma)$ and $\boldsymbol{Q}(\gamma)$ are regenerative processes with cycle times $\tilde{C}_k$ and $C$, respectively, equal to the return time to (0, 1). The first epoch when they differ is the first passage time $T_k$ to any of the states in $\{(n, j)\colon n \geq 0,\ j \geq k+1\}$. The random variables $\tilde{C}_k$ and $C$ are bounded by the interval of time $V$ until a catastrophe occurs; $V$ is geometrically distributed and has a finite expectation. By Theorem 3.1, this is a sufficient condition for $\lim_{k\to\infty} {}_k\boldsymbol{\pi}(\gamma)$ to be equal to $\boldsymbol{\pi}(\gamma)$.

Now, the stationary probability $\pi_{(0,1)}(\gamma)$ is the inverse of the expected cycle time $\mathrm{E}[C]$, and $C = \min(V, D)$, where $D$ is the cycle time for the process with transition probability matrix $\boldsymbol{P}$, independent of $V$. Thus,

$$\mathrm{E}[C] = \sum_{\nu\geq 0} \mathrm{P}[C > \nu] = \sum_{\nu\geq 0} (1-\gamma)^{\nu}\,\mathrm{P}[D > \nu] = \frac{1-\Delta(1-\gamma)}{\gamma},$$

where $\Delta(\cdot)$ is the generating function of $D$. Therefore,

$$\pi_{(0,1)}(\gamma) = \frac{\gamma}{1-\Delta(1-\gamma)}$$

and

$$\lim_{\gamma\to 0}\lim_{k\to\infty} {}_k\pi_{(0,1)}(\gamma) = \lim_{\gamma\to 0} \pi_{(0,1)}(\gamma) = \lim_{\gamma\to 0} \frac{\gamma}{1-\Delta(1-\gamma)} = \frac{1}{\Delta'(1)} = \pi_{(0,1)}.$$

By [11, Theorem 3.1], this concludes the proof.

It is easy to verify that $\pi_{(0,1)}(\gamma)$ actually decreases to $\pi_{(0,1)}$ as $\gamma$ decreases to 0.

## 6. First phase augmentation

The augmentation scheme defined in (4.5) and (4.6) imposes a somewhat lesser perturbation of the system dynamics, by augmenting on the first phase without forcing the process to move all the way to level 0. With this strategy, the most natural resynchronization of the QBDs is no longer on the single state (0, 1) but on the infinite set $K = \{(n, 1)\colon n \geq 0\}$. We need, therefore, to extend the results in [4] in two ways, in order to accommodate both the Markov renewal structure of the process and the fact that $|K| = \infty$. Before doing so in Theorem 6.1, we need the following technical lemma.

**Lemma 6.1.** *Consider a positive recurrent QBD with transition matrix $\boldsymbol{P}$ given in (1.1), such that the doubly infinite matrix $\boldsymbol{P}_D$ given in (1.6) is irreducible. Denote by $C$ the first return time to any state in $K = \{(n, 1) : n \geq 0\}$. Then there exists $\Omega$ such that $\mathrm{E}[C \mid X_0 = (i, 1)] \leq \Omega$ independently of $i$.*

https://doi.org/10.1239/aap/1293113153 Published online by Cambridge University Press

*Proof.* Denote by $X$ and $X^*$ the processes with transition matrix $\boldsymbol{P}$ and $\boldsymbol{P}_D$, respectively. We may define them on the same probability space and choose $X_0 = X_0^*$ so that their paths will coincide at least until the first passage to level 0, provided that the initial level is positive. Denote by $C^*$ the first return time of $X^*$ to $K$ and by $\Theta$ the joint first passage time of $X$ and $X^*$ to level 0, and observe that $C^* = C$ if $C \le \Theta$. Now,

$$\begin{aligned}\mathrm{E}[C \mid X_0 = (i,1)] &= \mathrm{E}[C1\{C < \Theta\} \mid X_0 = (i,1)] + \mathrm{E}[C1\{C \ge \Theta\} \mid X_0 = (i,1)] \\ &= \mathrm{E}[C^*1\{C^* < \Theta\} \mid X_0^* = (i,1)] + \mathrm{E}[C1\{C \ge \Theta\} \mid X_0 = (i,1)].\end{aligned}$$

The second term in the equation above tends to 0 as $i$ goes to $\infty$ because $\Theta$ tends to $\infty$ and $C$ is finite a.s. Thus,

$$\begin{aligned}\lim_{i\to\infty} \mathrm{E}[C \mid X_0 = (i,1)] &= \lim_{i\to\infty} \mathrm{E}[C^*1\{C^* < \Theta\} \mid X_0^* = (i,1)] \\ &= \lim_{i\to\infty} \mathrm{E}[C^* \mid X_0^* = (i,1)] \\ &= \mathrm{E}[C^* \mid X_0^* = (0,1)] \\ &< \infty,\end{aligned}$$

where the second equality is justified by the fact that $\Theta \to \infty$ as $i \to \infty$, and the third equality is justified by the fact that $\mathrm{E}[C^* \mid X_0^* = (i,1)]$ is independent of $i$, the behavior of $X^*$ being completely homogeneous in the levels.

The sequence $\{\mathrm{E}[C \mid X_0 = (i,1)]\}$ being finite and converging, it is uniformly bounded.

**Theorem 6.1.** *Consider a sequence $\{{}_k\tilde{Z}\}$ of QBD processes with transition matrix ${}_k\tilde{\boldsymbol{P}}$, where the blocks of ${}_k\tilde{\boldsymbol{P}}$ are defined in (4.5) and (4.6), and assume that each process is irreducible and positive recurrent. Denote by ${}_k\boldsymbol{\pi}$ the stationary distribution of ${}_k\tilde{Z}$, and denote its rate matrix by ${}_k\tilde{\boldsymbol{R}}$.*

*If there exists $\rho < 1$ such that $\mathrm{sp}({}_k\tilde{\boldsymbol{R}}) < \rho$ for large enough $k$, where $\mathrm{sp}(\cdot)$ is the spectral radius, then $\lim_{k\to\infty} {}_k\boldsymbol{\pi} = \boldsymbol{\pi}$.*

*Proof.* As in the proof of [4, Theorem 1], we assume that all the processes are defined on the same probability space. We interpret them as semi-regenerative processes, with regeneration times at the visit epochs to the states in $K$, and we denote by $C$ and $\tilde{C}_k$ the first passage time to $K$ by the processes $X$ and ${}_k\tilde{Z}$, respectively.

The embedded semi-Markov processes have kernels $\boldsymbol{H}(t)$ and ${}_k\tilde{\boldsymbol{H}}(t)$, where

$$\begin{aligned}H_{ij}(t) &= \mathrm{P}[C \le t,\ L(C) = j \mid X_0 = (i,1)], \\ {}_k\tilde{H}_{ij}(t) &= \mathrm{P}[\tilde{C}_k \le t,\ L(\tilde{C}_k) = j \mid {}_k\tilde{Z}_0 = (i,1)],\end{aligned}$$

and, by [2, Theorem 10.4.9],

$$ {}_k\boldsymbol{\pi}_{nj} = \frac{1}{{}_k\boldsymbol{\nu}^\top {}_k\boldsymbol{m}} \sum_{i\ge 0} {}_k\nu_i \sum_{u\ge 0} {}_kF_{in}(j,u), \tag{6.1}$$

where ${}_k\boldsymbol{\nu}$ is the stationary distribution of ${}_k\tilde{\boldsymbol{H}}(\infty)$, ${}_km_i = \mathrm{E}[\tilde{C}_k \mid {}_k\tilde{Z}_0 = (i,1)]$ is the expected duration of an inter-regeneration interval, and ${}_kF_{in}(j,u) = \mathrm{P}[{}_k\tilde{Z}_u = (n,j),\ \tilde{C}_k \ge u \mid {}_k\tilde{Z}_0 = (i,1)]$ is the distribution of the process during an inter-regeneration interval.

In order to prove from (6.1) that the sequence ${}_k\boldsymbol{\pi}$ converges, we need to prove that the sequence ${}_k\boldsymbol{\nu}$ converges, which seems to be at least as difficult. However, the vector ${}_k\boldsymbol{\nu}$ appears in both the numerator and the denominator, which will simplify things considerably.

https://doi.org/10.1239/aap/1293113153 Published online by Cambridge University Press

The sequence ${}_k\boldsymbol{\nu}$, being bounded, has accumulation points and our strategy will be to show that they all lead to the same result. Let us, therefore, fix $j_0$ arbitrarily and choose $\sigma$ to be any sequence such that $\lim_{k\in\sigma} {}_k\nu_{j_0}$ exists. By [11, Lemma 2.1], there exists a constant $c$, $0 < c \leq 1$, which may depend on the sequence $\sigma$, such that

$$\lim_{k\in\sigma} {}_k\nu_j = c\nu_j \quad \text{for all } j.$$

Let $T_k$ be the first passage time of $X$ to any of the states of the form $(n, j)$ with $j \geq k+1$. By construction, $X_t = {}_k\tilde{Z}_t$ for all $t < T_k$ and $\tilde{C}_k = \min(C, T_k)$, which implies by Lemma 6.1 that

$$ {}_km_i \leq \mathrm{E}[C \mid X_0 = (i, 1)] \leq \Omega \quad \text{uniformly in } i \text{ and } k. \tag{6.2}$$

Now, in the special case where $j = 1$, (6.1) reduces to ${}_k\pi_{n1} = {}_k\nu_n / {}_k\boldsymbol{\nu}^\top {}_k\boldsymbol{m}$. Indeed, the sum $\sum_{u\geq 0} {}_kF_{in}(1, u)$ is equal to 0 for $i \neq n$, for the process cannot spend any time in $(n, 1)$, starting from $(i, 1)$, *before* returning to $K$; for $i = n$, this sum is equal to 1, the unit of time spent in $X_0 = (n, 1)$ before moving to $X_1$.

We apply this to the state $(i, 1)$ and write ${}_k\nu_i = ({}_k\boldsymbol{\nu}^\top {}_k\boldsymbol{m}) {}_k\pi_{i1}$. By (6.2), ${}_km_i \leq \Omega$, so that ${}_k\boldsymbol{\nu}^\top {}_k\boldsymbol{m} \leq \Omega$ as well. Furthermore, ${}_k\pi_{i,1} = {}_k\boldsymbol{\pi}_0^\top ({}_k\tilde{\boldsymbol{R}})^i \boldsymbol{e}_1^\top$, and its asymptotic decay rate $\mathrm{sp}({}_k\tilde{\boldsymbol{R}})$ is bounded by $\rho < 1$ by assumption, so that ${}_k\pi_{i,1} = O(\rho^i)$.

We have therefore shown that there exists a constant $\chi$ such that ${}_k\nu_i \leq \chi\rho^i$, independently of $k$. As $\rho < 1$, the series of upper bounds converges and we obtain, by the dominated convergence theorem,

$$\lim_{k\in\sigma} {}_k\boldsymbol{\nu}^\top {}_k\boldsymbol{m} = c \sum_{i\geq 0} \nu_i \lim_{k\in\sigma} {}_km_i, \tag{6.3}$$

provided that $\lim_{k\in\sigma} {}_k\boldsymbol{m}$ exists.

It is clear from the construction of the processes ${}_k\tilde{Z}$ that

$$\begin{aligned} \tilde{C}_k < T_k \quad &\Longleftrightarrow \quad C < T_k \quad \Longrightarrow \quad \tilde{C}_k = C, \\ \tilde{C}_k \geq T_k \quad &\Longleftrightarrow \quad C \geq T_k \quad \Longrightarrow \quad \tilde{C}_k = T_k, \end{aligned}$$

and we may write

$$\begin{aligned} {}_km_i &= \mathrm{E}[\tilde{C}_k 1\{\tilde{C}_k < T_k\} \mid {}_k\tilde{Z}_0 = (i, 1)] + \mathrm{E}[\tilde{C}_k 1\{\tilde{C}_k \geq T_k\} \mid {}_k\tilde{Z}_0 = (i, 1)] \\ &= \mathrm{E}[C1\{C < T_k\} \mid X_0 = (i, 1)] + \mathrm{E}[T_k 1\{C \geq T_k\} \mid X_0 = (i, 1)]. \end{aligned} \tag{6.4}$$

The second term in (6.4) is bounded above by $\mathrm{E}[C1\{C \geq T_k\} \mid X_0 = (i, 1)]$, and this upper bound decreases to 0, since $C$ is finite a.s. and $T_k$ increases to $\infty$ with $k$. The first term monotonically converges to $m_i = \mathrm{E}[C1\{C < \infty\} \mid X_0 = (i, 1)]$. This allows us to conclude from (6.3) that

$$\lim_{k\in\sigma} {}_k\boldsymbol{\nu}^\top {}_k\boldsymbol{m} = c\boldsymbol{\nu}^\top \boldsymbol{m}.$$

We apply exactly the same argument to the numerator of (6.1). In order to justify the various exchanges of limiting and summation operators, we rely on the fact that the series may be written as

$$\begin{aligned} \sum_{u\geq 0} {}_kF_{in}(j, u) &= \mathrm{E}\left[\sum_{u\geq 0} 1\{{}_k\tilde{Z}_u = (n, j),\ \tilde{C}_k \geq u\} \,\middle|\, {}_k\tilde{Z}_0 = (i, 1)\right] \\ &= \mathrm{E}\left[\sum_{0\leq u\leq \tilde{C}_k} 1\{{}_k\tilde{Z}_u = (n, j)\} \,\middle|\, {}_k\tilde{Z}_0 = (i, 1)\right] \\ &\leq \mathrm{E}[\tilde{C}_k \mid {}_k\tilde{Z}_0 = (i, 1)], \end{aligned}$$

https://doi.org/10.1239/aap/1293113153 Published online by Cambridge University Press

and we finally obtain $\lim_{k\in\sigma}{}_k\boldsymbol{\pi} = \boldsymbol{\pi}$. The limit being independent of $\sigma$, this concludes the proof.

The assumption that the transition matrices are irreducible for all $k$ is not very restrictive, what really matters is that there is a unique irreducible class of positive recurrent states, at least for infinitely many values of $k$.

**Remark 6.1.** It is not clear how restrictive is the assumption of the existence of $\rho$. It seems likely that a sufficient condition is that $\kappa(\boldsymbol{R}) < 1$, where $\kappa(\cdot)$ is the convergence norm (see [10, Section 6.1]), a condition which should hold in general. Our argument is as follows. Since the processes $\{{}_k\tilde{Z}\}$ are positive recurrent with finitely many phases, we know that $\mathrm{sp}({}_k\tilde{\boldsymbol{R}}) < 1$, but we need more. If we decompose the number $N_j$ of visits to $(1, j)$ under the taboo of level 0 as the sum ${}_kN_j + {}_kN'_j$ of visits before and after time $T_k$, respectively, then we may write ${}_k\tilde{\boldsymbol{R}} = {}_k\boldsymbol{R} + {}_k\boldsymbol{R}'$, where

$$ {}_kR'_{ij} = \mathrm{E}\Bigg[ \sum_{\min(\theta,T_k)+1\le n\le\theta} 1\{{}_k\tilde{Z}_n = (1, j)\} \;\Bigg|\; {}_k\tilde{Z}_0 = (0, i) \Bigg]. $$

Since $\theta$ is finite a.s. and $\lim_{k\to\infty} T_k = \infty$, the limit $\lim_{k\to\infty}{}_kR'_{ij}$ is equal to 0. Furthermore, as shown at the end of Section 2, ${}_k\boldsymbol{R}$ converges from below to $\tilde{\boldsymbol{R}}$ so that, by [5, Lemma 2.3], $\mathrm{sp}({}_k\boldsymbol{R})$ converges from below to $\kappa(\boldsymbol{R})$. All the ingredients are therefore present to suggest that the sequence $\mathrm{sp}({}_k\tilde{\boldsymbol{R}})$ would converge to $\kappa(\boldsymbol{R})$, which is at most equal to 1, since the original QBD is positive recurrent. In the cases where $\kappa(\boldsymbol{R})$ is strictly less than 1, $\mathrm{sp}({}_k\tilde{\boldsymbol{R}})$ would be bounded away from 1 for large enough $k$. We do not however have a formal proof.

**Remark 6.2.** We may also combine the principle of augmentation on the first phase with the catastrophe from Section 5; the effect of a catastrophe being limited here to bringing the process to phase one, possibly with a minor change of level.

We define the QBD matrix $\boldsymbol{P}^*$ with blocks

$$ \boldsymbol{B}^* = (\boldsymbol{\delta}^*_{-1} + \boldsymbol{\delta}^*_0)\cdot\boldsymbol{e}_1^\top \quad\text{and}\quad \boldsymbol{A}^*_\nu = \boldsymbol{\delta}^*_\nu\cdot\boldsymbol{e}_1^\top \quad\text{for } \nu = -1, 0, 1, $$

where $\boldsymbol{\delta}^*_\nu \ge \boldsymbol{0}$, $\boldsymbol{\delta}^*_{-1} + \boldsymbol{\delta}^*_0 + \boldsymbol{\delta}^*_1 = \boldsymbol{1}$, and $\boldsymbol{e}_1$ a vector of infinitely many components with the first component equal to 1 and all other components equal to 0. We see from every state $(n, i)$ that the system directly moves to some state in $K$, although there is some degree of freedom in defining the probabilities $(\boldsymbol{\delta}^*_{-1})_{(n,i)}$, $(\boldsymbol{\delta}^*_0)_{(n,i)}$, and $(\boldsymbol{\delta}^*_1)_{(n,i)}$ of moving to $(n-1, 1)$, $(n, 1)$, and $(n+1, 1)$, respectively.

Next, we define the QBD matrices $\boldsymbol{Q}^*(\gamma) = \gamma\boldsymbol{P}^* + (1-\gamma)\boldsymbol{P}$, $0 \le \gamma \le 1$, which will here play a role similar to that of the matrix $\boldsymbol{Q}(\gamma)$ in (5.2). We take a sequence of QBD matrices ${}_k\tilde{\boldsymbol{P}}$ without any assumption on the type of augmentation and define the sequence

$$ {}_k\boldsymbol{Q}^*(\gamma) = \gamma{}_k\boldsymbol{P}^* + (1-\gamma){}_k\tilde{\boldsymbol{P}}, \tag{6.5} $$

where ${}_k\boldsymbol{P}^*$ is obtained by performing a $k$ by $k$ truncation of the blocks of $\boldsymbol{P}^*$; we denote by ${}_k\boldsymbol{\pi}^*(\gamma)$ the stationary distribution of ${}_k\boldsymbol{Q}^*(\gamma)$.

It is easy to duplicate the proof of Theorem 6.1 and to show that, under a similar set of conditions, $\lim_{\gamma\to0}\lim_{k\to\infty}{}_k\boldsymbol{\pi}^*(\gamma) = \boldsymbol{\pi}$. The first step is a nearly verbatim repetition of the proof of Theorem 6.1: we rely on the fact that $\tilde{C}_k(\gamma)$, the first return to $K$, is bounded by $V$, the time until the first catastrophe and show that $\lim_{k\to\infty}{}_k\boldsymbol{\pi}^*(\gamma) = \boldsymbol{\pi}^*(\gamma)$ for a fixed $\gamma$. In the second step we prove that $\lim_{\gamma\to0}\boldsymbol{\pi}^*(\gamma) = \boldsymbol{\pi}$, and here again, we repeat the argument in

https://doi.org/10.1239/aap/1293113153 Published online by Cambridge University Press

Theorem 6.1; the first return time $C(\gamma)$ to $K$ is bounded by $C$, so that we may use Lemma 6.1 to obtain a uniform bound on the expectation of $C(\gamma)$.

The problem is that we need some reasonably simple constraint on the rate matrices ${}_k\boldsymbol{R}^*(\gamma)$ and $\boldsymbol{R}^*(\gamma)$ of the QBDs ${}_k\boldsymbol{Q}^*(\gamma)$ and $\boldsymbol{Q}^*(\gamma)$, respectively, in order to allow for the exchange of limit and summation as in (6.3). We might argue, as in Remark 6.1, that $\kappa(\boldsymbol{R}) < 1$ is likely to be a sufficient condition, but this remains to be proved. We return to this question in the next section.

## 7. Illustrations

We take three examples of QBDs for which we know the exact stationary distribution, so that we are able to compare the successive approximations to the exact values. Our measure of distance between $\boldsymbol{\pi}$ and ${}_k\boldsymbol{\pi}$ is

$$\varepsilon_k = \sum_{(n,i)\in\mathcal{E}_k} |\pi_{n,i} - {}_k\pi_{n,i}| + \sum_{(n,i)\notin\mathcal{E}_k} \pi_{n,i},$$

where $\mathcal{E}_k = \{(n,i)\colon n \geq 0, 1 \leq i \leq k\}$. The measure $\varepsilon_k$ is the $L_1$ norm of the difference between $\boldsymbol{\pi}$ and a vector obtained by expanding ${}_k\boldsymbol{\pi}$ so that it is defined over the whole state space of the original QBD, with components set to 0 for states not in $\mathcal{E}_k$. Following [12], we note that

$$\varepsilon_k = 2 \sum_{(n,i)\in\mathcal{E}_k^+} ({}_k\pi_{n,i} - \pi_{n,i}) \geq 2\left(1 - \sum_{(n,i)\in\mathcal{E}_k} \pi_{n,i}\right), \tag{7.1}$$

where $\mathcal{E}_k^+$ is the set of states in $\mathcal{E}_k$ for which ${}_k\pi_{n,i} > \pi_{n,i}$. The rightmost expression in (7.1) is particularly useful since this lower bound provides us with a benchmark against which to judge the quality of any approximation ${}_k\boldsymbol{\pi}$.

**Example 7.1.** (*The discrete-time M/PH/1 queue.*) Consider a discrete-time queue, where arrivals occur according to a Bernoulli process and service times have a PH$(\boldsymbol{\tau}, \boldsymbol{T})$ representation of order $m < \infty$; we denote by $\lambda$ the probability of an arrival at any given time.

The M/PH/1 queue is usually analyzed as the process $\{(L_n, \varphi_n),\ n \geq 0\}$, where $L_n$ is the number of customers in the system at time $n$ and, for $L_n \geq 1$, $\varphi_n$ is the phase at time $n$ of the customer being served; if the queue is empty then the value of the phase is irrelevant. The transition matrix $\boldsymbol{P}^*$ has the structure (1.4) with

$$\begin{aligned} \boldsymbol{B}_{-1} &= (1-\lambda)\boldsymbol{t}, & \boldsymbol{B}_0 &= 1-\lambda, & \boldsymbol{B}_1 &= \lambda\boldsymbol{\tau}, \\ \boldsymbol{A}_{-1} &= (1-\lambda)\boldsymbol{t}\cdot\boldsymbol{\tau}, & \boldsymbol{A}_0 &= \lambda\boldsymbol{t}\cdot\boldsymbol{\tau} + (1-\lambda)\boldsymbol{T}, & \boldsymbol{A}_1 &= \lambda\boldsymbol{T}, \end{aligned}$$

where $\boldsymbol{t} = \mathbf{1} - \boldsymbol{T}\mathbf{1}$. Its stationary distribution $\boldsymbol{\omega}$ is given by

$$\boldsymbol{\omega}_1 = \omega_0\lambda\boldsymbol{\tau}\boldsymbol{V}, \qquad \boldsymbol{\omega}_n = \boldsymbol{\omega}_1(R^*)^{n-1} \quad \text{for } n \geq 1,$$

when it exists, where $\boldsymbol{V} = (\boldsymbol{I} - \lambda\mathbf{1}\cdot\boldsymbol{\tau} - (1-\lambda)\boldsymbol{T})^{-1}$, $\boldsymbol{R}^* = \lambda\boldsymbol{T}\boldsymbol{V}$, and $\omega_0 = (1 + \lambda\boldsymbol{\tau}\boldsymbol{V}(\boldsymbol{I} - \boldsymbol{R}^*)^{-1}\mathbf{1})^{-1}$.

We define the residual service time representation as $\{(L_n, r_n),\ n \geq 0\}$, where $L_n$ is, as before, the number of customers in the queue and $r_n$ is the residual service time for the customer in service at time $n$. This is also a QBD, since the number of customers varies at most by one unit, but it has infinitely many phases if the service time is unbounded, as we now assume. The

https://doi.org/10.1239/aap/1293113153 Published online by Cambridge University Press

transition matrix $\boldsymbol{P}$ has the structure (1.4) with

$$\begin{aligned} \boldsymbol{B}_{-1} &= (1-\lambda)\boldsymbol{e}_1, & \boldsymbol{B}_0 &= 1-\lambda, & \boldsymbol{B}_1 &= \lambda\boldsymbol{q}, \\ \boldsymbol{A}_{-1} &= (1-\lambda)\boldsymbol{e}_1 \cdot \boldsymbol{q}, & \boldsymbol{A}_0 &= \lambda\boldsymbol{e}_1 \cdot \boldsymbol{q} + (1-\lambda)\boldsymbol{M}, & \boldsymbol{A}_1 &= \lambda\boldsymbol{M}, \end{aligned}$$

where $q_j = \boldsymbol{\tau} T^{j-1}\boldsymbol{t}$ for $j \geq 1$ and

$$\boldsymbol{M} = \begin{bmatrix} 0 & 0 & 0 & \cdots \\ 1 & 0 & 0 & \ddots \\ 0 & 1 & 0 & \ddots \\ \vdots & \ddots & \ddots & \ddots \end{bmatrix}.$$

By conditioning on $\varphi_n$ we readily see that

$$\begin{aligned} \mathrm{P}[L_n = k,\, r_n = j] &= \sum_{1 \leq i \leq m} \mathrm{P}[L_n = k,\, \varphi_n = i]\, \mathrm{P}[r_n = j \mid L_n = k,\, \varphi_n = i] \\ &= \sum_{1 \leq i \leq m} \mathrm{P}[L_n = k,\, \varphi_n = i](T^{j-1}\boldsymbol{t})_i, \end{aligned}$$

so that the stationary distribution $\boldsymbol{\pi}$ of $\boldsymbol{P}$ is given by $\pi_0 = \omega_0$, and $\pi_{k,j} = \boldsymbol{\omega}_k \boldsymbol{T}^{j-1}\boldsymbol{t}$ for $k \geq 1$ and all $j$. Knowing this allows us to compare the exact distribution to the approximations ${}_k\boldsymbol{\pi}$ obtained by the various procedures.

In our two examples, $(\boldsymbol{\tau}, \boldsymbol{T})$ is a mixture of two geometric distributions:

$$\boldsymbol{\tau} = [\alpha, 1-\alpha], \qquad \boldsymbol{T} = \begin{bmatrix} p_1 & 0 \\ 0 & p_2 \end{bmatrix}$$

(note that our geometric distributions start at 1, not 0 as is often the case). The factorial moments are given by $k!\,\boldsymbol{\tau}(\boldsymbol{I}-\boldsymbol{T})^{-k}\boldsymbol{T}^{k-1}\mathbf{1}$ for $k \geq 1$ (see [6, Section 2.5]), so that

$$\mathrm{E}[S] = \frac{\alpha}{1-p_1} + \frac{1-\alpha}{1-p_2}$$

and

$$\mathrm{E}[S(S-1)] = \frac{2\alpha p_1}{(1-p_1)^2} + \frac{2(1-\alpha)p_2}{(1-p_2)^2}.$$

We fix the mean $\mu$ and the squared coefficient of variation $c^2 = \sigma^2/\mu^2$, and we chose the parameters so that each phase contributes one half of the mean:

$$\frac{\alpha}{1-p_1} = \frac{1-\alpha}{1-p_2} = \frac{\mu}{2}. \tag{7.2}$$

Making use of this relation, we find that

$$\mathrm{var}(S) = \frac{\mu^2}{2\alpha} + \frac{\mu^2}{2(1-\alpha)} - \mu - \mu^2,$$

which leads to

$$2\alpha(1-\alpha) = \frac{\mu^2}{(c^2+1)\mu^2 + \mu}.$$

The right-hand side of this is given and so we easily solve for $\alpha$; the values of $p_1$ and $p_2$ immediately follow from (7.2).

https://doi.org/10.1239/aap/1293113153 Published online by Cambridge University Press

TABLE 1.

| Case | $c$ | $\alpha$ | $p_1$ | $p_2$ |
|---|---|---|---|---|
| 1 | $\sqrt{2}$ | 0.206 63 | 0.979 34 | 0.920 66 |
| 2 | 2 | 0.111 43 | 0.988 86 | 0.911 14 |

We consider two M/PH/1 queues with $\lambda = 0.045$ and $\mu = 20$. In the first case $c = \sqrt{2}$ and in the second case $c = 2$. The parameter values are given in Table 1, rounded to five decimal places.

Figure 1 gives a general impression of how the various techniques behave in the first case, Figure 2 is the log version. We have plotted the lower bound for visual reference and we note that it decreases quite slowly: the truncation parameter has to be greater than 200 before the distance between $\boldsymbol{\pi}$ and ${}_k\boldsymbol{\pi}$ can possibly drop below $10^{-2}$. The different curves are marked as follows. *First column* refers to the procedure analyzed in Section 3, with transition matrix given by (3.1). For *last phase*, the augmented blocks in the transition matrix (4.1) are given by (4.4). This corresponds to putting the missing mass on its last column; in effect, the service time is bounded by $k$. The same blocks are used as ${}_k\tilde{\boldsymbol{P}}$ in (5.1) for *AR first column*, which identifies the results of the procedure in Section 5, and as ${}_k\tilde{\boldsymbol{P}}$ in (6.5) for *AR first phase*, which refers to the approach briefly described in Remark 6.2. Finally, *first phase* is the method which is the object of Theorem 6.1. The augmented matrices in (4.1) are of the form (4.5) and (4.6) with ${}_k\boldsymbol{\delta}_{-1} = 0$, ${}_k\boldsymbol{\delta}_0 = \boldsymbol{a}_{-1}^{(k)}$, and ${}_k\boldsymbol{\delta}_1 = \boldsymbol{a}_0^{(k)}$; note that $\boldsymbol{a}_1^{(k)} = \boldsymbol{0}$. This illustrates the fact that the missing mass on each row may be arbitrarily allocated among the first columns of the three blocks. Here, it has the effect of forcing a resampling of the service time distribution whenever the service length is too long.

We observe that first column and AR first column always converge more slowly (and less accurately in the AR case) than first phase and AR first phase, respectively, and that last phase appears to be the most accurate approximation for this model. Furthermore, Figure 2 shows that the three schemes first column, first phase, and last phase converge linearly.

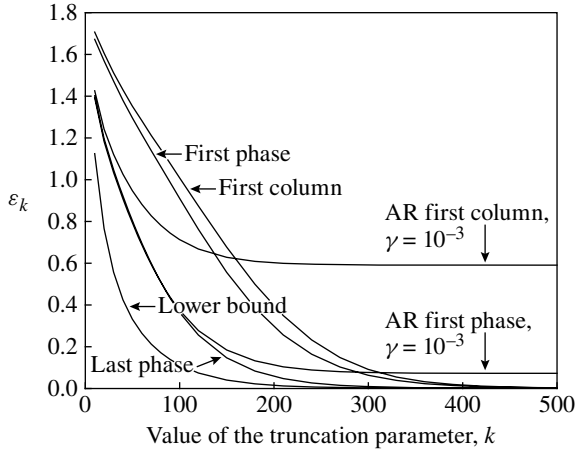


FIGURE 1: Plot of the error $\varepsilon_k$ against the truncation parameter, $k$, for the M/PH/1 model using a service distribution with a coefficient of variation of $\sqrt{2}$; AR is the abbreviation for 'artificial restart'.

https://doi.org/10.1239/aap/1293113153 Published online by Cambridge University Press

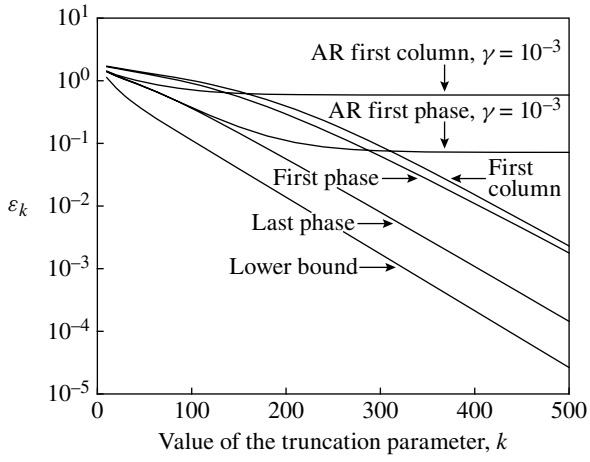


FIGURE 2: Plot of $\log_{10} \varepsilon_k$ against the truncation parameter, $k$, for the M/PH/1 model using a service distribution with a coefficient of variation of $\sqrt{2}$.

The main characteristic of the AR schemes is most apparent in Figure 2: as $k$ increases, the error initially diminishes, until it plateaus and remains nearly constant. This is because the Poisson process of catastrophes has a parameter independent of $k$ and because its influence is masked for small values of $k$; there is, however, a point where it becomes the major effect.

More details are given in Figure 3, where we plot the errors for the two AR schemes, with $\gamma$ taken as $10^{-3}$, $10^{-4}$, and $10^{-5}$. First, we observe that to augment on the first phase is one order of magnitude better in terms of the error than to augment on the first column of the whole matrix, a clear sign that it is better to preserve the dynamic structure as much as possible. Furthermore, the smaller the value of $\gamma$, the longer the linear convergence persists before the plateau, and, hence, the more accurate the answer. Actually, with $\gamma = 10^{-5}$, AR first phase is nearly as good as last phase until $k = 300$ and better than first phase until $k = 500$ at least. This

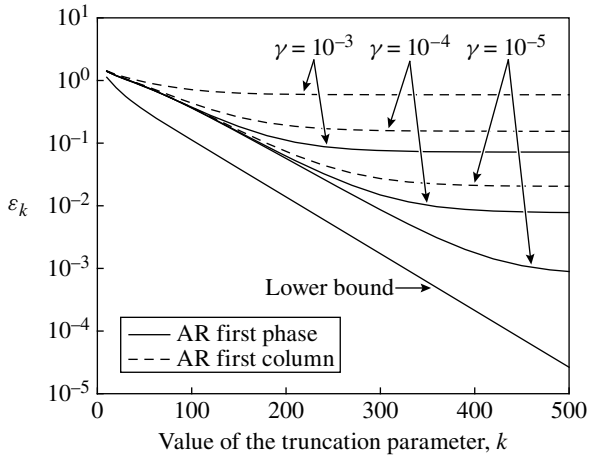


FIGURE 3: Plot of $\log_{10} \varepsilon_k$ against the truncation parameter, $k$, for various AR methods and the M/PH/1 model using a service distribution with a coefficient of variation of $\sqrt{2}$.

https://doi.org/10.1239/aap/1293113153 Published online by Cambridge University Press

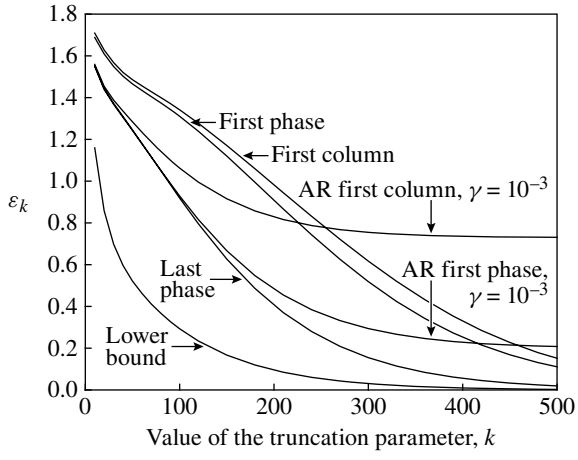


FIGURE 4: Plot of the error $\varepsilon_k$ against the truncation parameter, $k$, for the M/PH/1 model using a service distribution with a coefficient of variation of 2.

points to the advantage of the AR first phase approach, which is demonstrably convergent, and which allows us to augment the various blocks in a manner which is well suited to the model at hand. Figure 3 also makes it clear that, for any *given* $\gamma$, the distribution eventually converges to the wrong distribution.

Not surprisingly, the quality of the approximation depends on the service time distribution. In Figure 4 we plot the error against the truncation parameter, $k$, for an M/PH/1 queue with $c = 2$. Convergence is much slower here than in Figure 1, but all the qualitative features remain the same.

The question which remains to be answered is whether the sufficient condition of Theorem 6.1 is satisfied or not. Define $\{p_n : n \geq 0\}$ as the stationary marginal distribution for the number of customers in the queue, and define the series $p(z) = \sum_{n\geq 0} z^n p_n$, $\boldsymbol{R}(z) = \sum_{n\geq 0} z^n \boldsymbol{R}^n$, and $\boldsymbol{R}^*(z) = \sum_{n\geq 0} z^n (\boldsymbol{R}^*)^n$. Since $p_n = \boldsymbol{\omega}_n \mathbf{1} = \boldsymbol{\pi}_n \mathbf{1}$, we have

$$p(z) = \pi_0 + z\boldsymbol{\pi}_1 \boldsymbol{R}(z)\mathbf{1} = \omega_0 + z\boldsymbol{\omega}_1 \boldsymbol{R}^*(z)\mathbf{1}. \tag{7.3}$$

The number $m$ of phases is finite, so the radius of convergence of $p(z)$ and that of $\boldsymbol{R}^*(z)$ are equal. We denote this common radius of convergence by $c^*$, and we note that $c^* = 1/\mathrm{sp}(\boldsymbol{R}^*)$, with $\mathrm{sp}(\boldsymbol{R}^*) < 1$ since the QBD process $\{(L_n, \varphi_n)\}$ is positive recurrent.

By the first equality in (7.3), we conclude that the series $\boldsymbol{R}(z)$ converges for $z < c^*$, so that the convergence radius $c$ of $\boldsymbol{R}$ is at least equal to $c^*$, and the convergence norm $\kappa(\boldsymbol{R}) = 1/c \leq 1/c^* = \mathrm{sp}(\boldsymbol{R}^*) < 1$. We argued in Remark 6.1 that this should be sufficient to ensure that $\mathrm{sp}({}_k\tilde{\boldsymbol{R}})$ is bounded away from 1 and that we can apply the first phase and AR first phase techniques legitimately. As a verification, we show in Figure 5 that $\mathrm{sp}({}_k\boldsymbol{R})$ does converge, from below, to the spectral radius of the rate matrix $\boldsymbol{R}^*$ associated with the usual representation of the M/PH/1 queue.

**Example 7.2.** (*Tandem queues.*) We consider the Jackson network made of two queues in tandem: customers arrive at queue 1 according to a Poisson process with rate $\lambda$, after receiving a first exponential service with parameter $\mu_1$, and they move to queue 2 where they receive a second exponential service with parameter $\mu_2$. Both buffers have infinite capacity.

https://doi.org/10.1239/aap/1293113153 Published online by Cambridge University Press

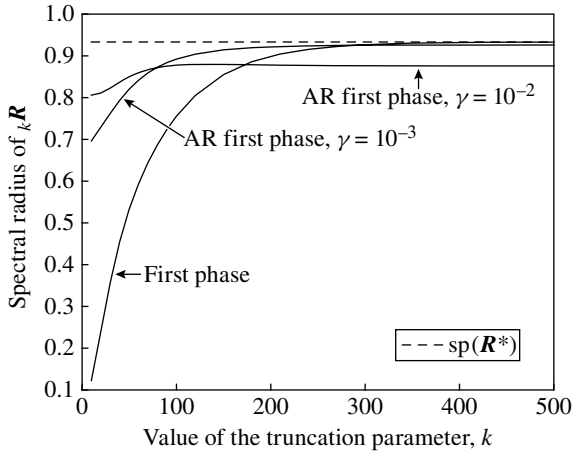


FIGURE 5: Plot of the spectral radius of $_k\boldsymbol{R}$ against the truncation parameter, $k$, for the M/PH/1 model using a service distribution with a coefficient of variation of $\sqrt{2}$.

We follow [5] and represent the system as a QBD $\{(L(t), \varphi(t))\}$, where the level is the number of customers in queue 2, while the phase is the number of customers in queue 1. Since the transition rates are uniformly bounded by $\lambda + \mu_1 + \mu_2$, we may use instead the discrete-time Markov chain $\{(L_n, \varphi_n)\}$ obtained by uniformization, which has the same stationary distribution. With this representation, it is well known that

$$\pi_{n,j} = (1 - \rho_2)(1 - \rho_1)\rho_2^n \rho_1^j \tag{7.4}$$

for $n, j \geq 0$, where $\rho_1 = \lambda/\mu_1$ and $\rho_2 = \lambda/\mu_2$, provided that $\rho_1$ and $\rho_2$ are both strictly less than 1.

It was shown in [5, Lemma 4.8, Theorem 4.9] that $\kappa(\boldsymbol{R}) = \rho_2$ if $\mu_1 > \mu_2$ (in which case queue 2 is the bottleneck), and $\kappa(\boldsymbol{R}) = \eta \leq \rho_2$ if $\mu_1 < \mu_2$ (queue 1 is the bottleneck), where $\eta$ is the unique root in (0, 1) of the equation $-\lambda - \mu_1 - \mu_2(1 - z) + 2\sqrt{\lambda\mu_1/z} = 0$. This contrasts somewhat with (7.4), where we see that the decay rate of $\pi_{n,j}$ with respect to the level $n$ is always $\rho_2$.

Furthermore, the authors of [5] considered the truncated and augmented QBD (4.1), where the blocks are given by (4.2) or, equivalently, by (4.4) in this example. They showed that $\text{sp}(_k\tilde{R})$ converges to $\kappa(R)$. We verify in Figures 6 and 7 that $\text{sp}(_k\tilde{\boldsymbol{R}})$ converges from below to $\kappa(\boldsymbol{R})$ for the first phase and AR first phase techniques, so we can apply these legitimately; for the AR techniques, the augmented blocks are given by (4.2).

Consider the situation where $\lambda = 8$ and $\mu_1 = 10 > \mu_2 = 9$; we see in Figure 6 that the spectral radii of the truncated QBDs converge, from below, to $\rho_2 = \frac{8}{9}$ as expected. Figure 8 shows extremely fast convergence for all techniques, including the last phase augmentation, and similar qualitative features to those observed for the M/PH/1 example.

In our second example, $\lambda = 8$ and $\mu_1 = 9 < \mu_2 = 10$. Figure 7 shows, again, that the spectral radii converge from below; here the limit is $\eta \approx 0.7869 < \rho_2 = \frac{8}{10}$. Figure 9 shows similar qualitative features to those observed earlier. However, first phase and first column show less difference, relatively.

It is worth drawing attention to a difference between [5] and our paper. Here, we are concerned with the convergence of $_k\pi_{(n,j)}$ to $\pi_{(n,j)}$ for a *given* $n$ and $j$. It is actually much

https://doi.org/10.1239/aap/1293113153 Published online by Cambridge University Press

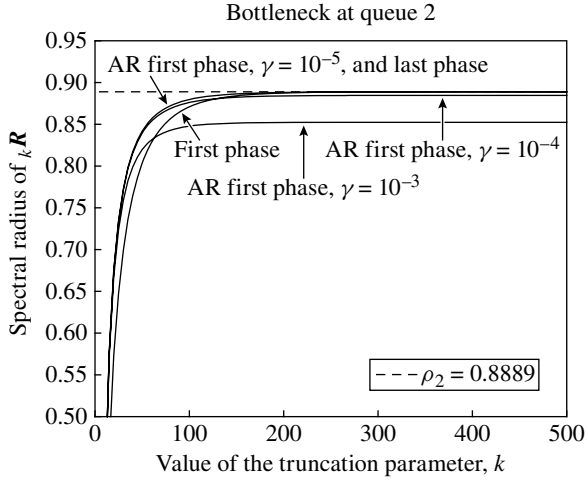


FIGURE 6: Plot of the spectral radius of ${}_k\boldsymbol{R}$ against the truncation parameter, $k$, for the tandem queue example where $\mu_1 = 10$ and $\mu_2 = 9$.

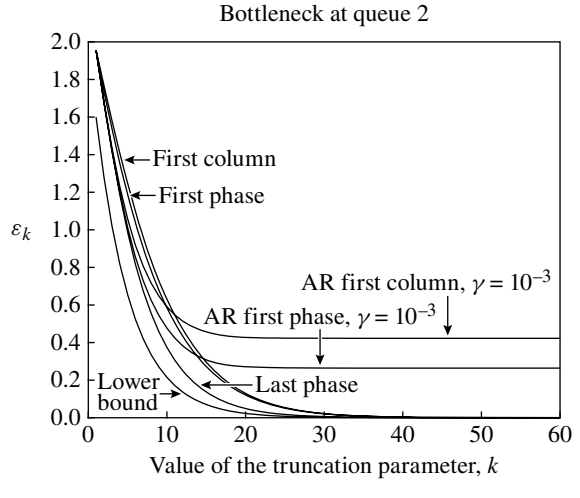


FIGURE 7: Plot of the error $\varepsilon_k$ against the truncation parameter, $k$, for the tandem queue example where $\mu_1 = 10$ and $\mu_2 = 9$.

easier to prove that, for any given $(n, i)$ and $(m, j)$, the ratio ${}_k\pi_{(n,i)}/{}_k\pi_{(m,j)}$ converges to $\pi_{(n,i)}/\pi_{(m,j)}$ as $k \to \infty$, even for last column augmentation; see [3, Lemma 2.1]. In particular, $\lim_{k\to\infty} {}_k\pi_{(n+1,i)}/{}_k\pi_{(n,i)} = \rho_2$ for any given $n$, no matter how large, and for any $i$.

The analysis in [5] concerns the *decay rate* $\lim_{n\to\infty} {}_k\pi_{(n+1,i)}/{}_k\pi_{(n,i)}$ and it was shown there that this limit converges to $\kappa(\boldsymbol{R})$ as $k \to \infty$, which is different from $\rho_2$ if $\mu_1 < \mu_2$. In short, we may have

$$\lim_{k\to\infty}\lim_{n\to\infty}\frac{{}_k\pi_{(n+1,i)}}{{}_k\pi_{(n,i)}} \neq \lim_{n\to\infty}\frac{\pi_{(n+1,i)}}{\pi_{(n,i)}} = \lim_{n\to\infty}\lim_{k\to\infty}\frac{{}_k\pi_{(n+1,i)}}{{}_k\pi_{(n,i)}},$$

another case of limits not being interchangeable.

https://doi.org/10.1239/aap/1293113153 Published online by Cambridge University Press

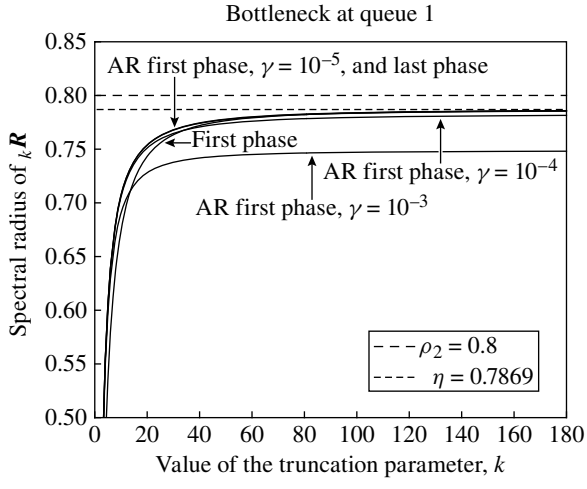


FIGURE 8: Plot of the spectral radius of ${}_k\boldsymbol{R}$ against the truncation parameter, $k$, for the tandem queue example where $\mu_1 = 9$ and $\mu_2 = 10$.

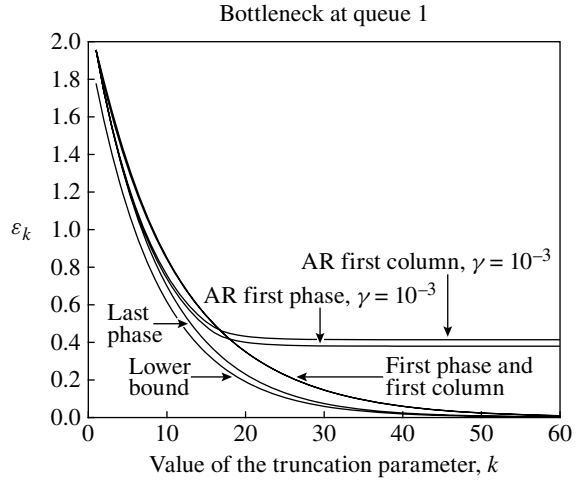


FIGURE 9: Plot of the error $\varepsilon_k$ against the truncation parameter, $k$, for the tandem queue example where $\mu_1 = 9$ and $\mu_2 = 10$.

**Example 7.3.** (*Example 4.1 (continued).*) We return to the QBD of Example 4.1. As shown in Section 5, that QBD is level-phase independent, and we easily see that $\mathrm{sp}({}_k\tilde{\boldsymbol{R}}) = \rho < 1$ for all $k$ and any scheme that augments the diagonal block only. In the figures to follow, we have chosen $N = 25$, and $\lambda = 0.18$ and $\mu = 0.2$, so that $\rho = 0.9$.

Figure 10 shows us that last phase augmentation really does diverge, as we stated in Section 4, since $\varepsilon_k$ stays above 0.75 when $k$ is an integer multiple of $N$. We also note that first phase does converge, and that we actually cannot tell the difference between first phase and the lower bound; first column also converges but significantly more slowly. Also, when $k$ is not an integer multiple of $N$, last phase is also indistinguishable from the lower bound.

https://doi.org/10.1239/aap/1293113153 Published online by Cambridge University Press

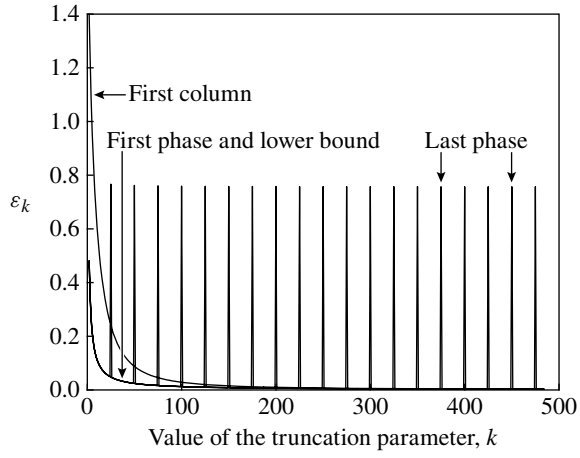


FIGURE 10: Plot of the error $\varepsilon_k$ against the truncation parameter, $k$, for the Gibson and Seneta example with $N = 25$.

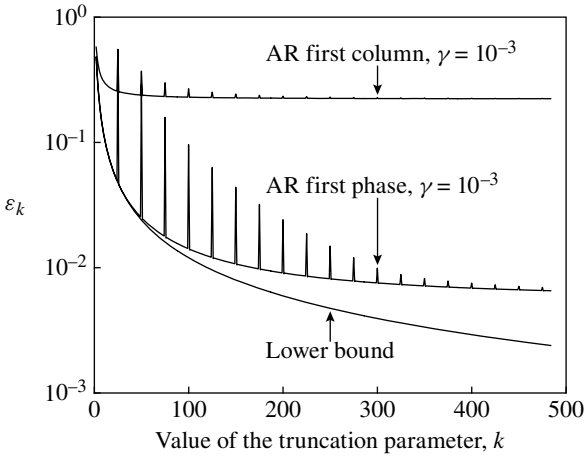


FIGURE 11: Plot of $\log_{10} \varepsilon_k$ against the truncation parameter, $k$, for the Gibson and Seneta example with $N = 25$.

Figure 11 shows, on a log scale, how the AR methods perform; the artificial restart mechanism has the effect of damping the oscillations due to last phase augmentation, at least for larger values of $k$, when the restart mechanism gets a chance to activate before the QBD reaches phase $k$. This is obtained at the cost of converging to the wrong distribution, with AR first phase converging more accurately than AR first column.

In Figure 12 we show in greater detail the influence of the value of $\gamma$ on the performance of the AR first phase method. For larger values of $\gamma$, control over the fluctuations is achieved for much smaller values of $k$; for example, when $\gamma = 10^{-2}$, no fluctuations after $k = 50$ are visible, while, for $\gamma = 10^{-3}$, this is more like $k = 250$. However, the larger the value of $\gamma$, the greater the effect of the perturbation due to artificial restarts, and so the greater the difference between $\boldsymbol{\pi}^*(\gamma)$ and $\boldsymbol{\pi}$. To show this, we plot both $\varepsilon_k$ and its lower bound for each value of $\gamma$; the lower bound is clearly visible for $\gamma = 10^{-2}$ only, and, for $\gamma < 10^{-4}$, this difference is unobservable.

https://doi.org/10.1239/aap/1293113153 Published online by Cambridge University Press

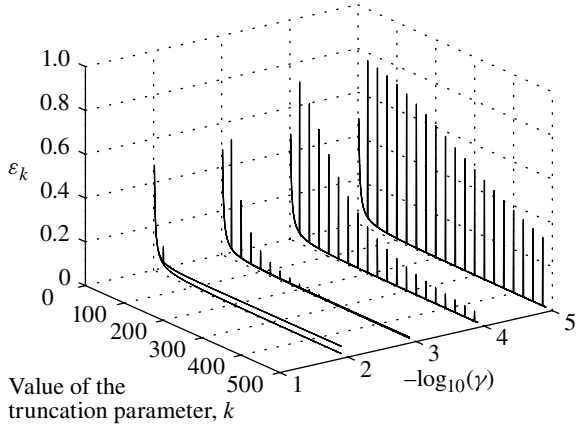


FIGURE 12: Plot of the error $\varepsilon_k$ against the truncation parameter, $k$, and $\log_{10}(\gamma)$ for the Gibson and Seneta example with $N = 25$ using the AR first phase technique as $\gamma$ varies.

## 8. Conclusion

In conclusion, in addition to presenting for QBDs the analysis of two existing truncation and augmentation approximation procedures, we have developed two new demonstrably convergent schemes that are specialized to QBD-type structures. We have also proven, through the introduction of AR mechanisms, that it may be useful to consider more complex schemes than simple truncation and augmentation.

We have demonstrated through examples that our two new schemes out-perform their general-purpose counterparts, and in the case of the AR schemes, this is by quite a margin; this confirms the usefulness of keeping as much of the structure as possible.

Finally, we have complemented the results of Kroese *et al.* [5] by showing that the stationary distribution of the truncated and augmented system may converge to the exact limit even if the decay rate does not.

## Appendix A. Direct jumps to 0

Consider a QBD with direct jumps to level 0, with finitely or infinitely many phases, and transition matrix

$$\boldsymbol{P} = \begin{bmatrix} \boldsymbol{B}_0 & \boldsymbol{A}_1 & \boldsymbol{0} & \boldsymbol{0} & \cdots \\ \boldsymbol{B}_{-1} & \boldsymbol{A}_0 & \boldsymbol{A}_1 & \boldsymbol{0} & \ddots \\ \boldsymbol{C} & \boldsymbol{A}_{-1} & \boldsymbol{A}_0 & \boldsymbol{A}_1 & \ddots \\ \boldsymbol{C} & \boldsymbol{0} & \boldsymbol{A}_{-1} & \boldsymbol{A}_0 & \ddots \\ \vdots & \vdots & \ddots & \ddots & \ddots \end{bmatrix}.$$

Define the first passage probability matrices $\boldsymbol{G}$ and $\boldsymbol{H}$ as

$$G_{ij} = \mathrm{P}[\tau < \infty,\ X_\tau = (n-1, j) \mid X_0 = (n, i)],$$
$$H_{ij} = \mathrm{P}[\tau < \infty,\ X_\tau = (0, j) \mid X_0 = (n, i)],$$

for all $i$, $j$ and $n \geq 2$, where $\tau$ is the first passage time to level $n-1$ or below.

https://doi.org/10.1239/aap/1293113153 Published online by Cambridge University Press

**Lemma A.1.** *The matrix $\boldsymbol{G}$ is the minimal nonnegative solution of (2.1) and the matrix $\boldsymbol{H}$ is the minimal nonnegative solution of*

$$\boldsymbol{X} = \boldsymbol{C} + \boldsymbol{A}_0\boldsymbol{X} + \boldsymbol{A}_1\boldsymbol{G}\boldsymbol{X} + \boldsymbol{A}_1\boldsymbol{X}, \tag{A.1}$$

*which can be expressed as*

$$\boldsymbol{H} = [\boldsymbol{I} - (\boldsymbol{A}_0 + \boldsymbol{A}_1\boldsymbol{G} + \boldsymbol{A}_1)]^{-1}\boldsymbol{C} \tag{A.2}$$

*if $|S| < \infty$.*

*Proof.* The statement about $\boldsymbol{G}$ is well known, and given here for the sake of completeness. Simple probabilistic arguments show that $\boldsymbol{H}$ is a solution of (A.1), which we write as

$$\boldsymbol{X} = \boldsymbol{C} + \boldsymbol{A}_1\boldsymbol{X} + \boldsymbol{U}\boldsymbol{X}, \tag{A.3}$$

where $\boldsymbol{U} = \boldsymbol{A}_0 + \boldsymbol{A}_1\boldsymbol{G}$. We need to prove that $\boldsymbol{H} = \boldsymbol{X}_{\min}$, the minimal nonnegative solution of (A.3).

The matrix $\boldsymbol{U}$ is substochastic and the series $\boldsymbol{N}_{\boldsymbol{U}} = \sum_{\nu\geq 0}(\boldsymbol{U})^{\nu}$ converges, so that we may rewrite (A.3) as $\boldsymbol{X} = \boldsymbol{N}_{\boldsymbol{U}}(\boldsymbol{C} + \boldsymbol{A}_1 X) + \lim_{\nu\to\infty}\boldsymbol{U}^{\nu}\boldsymbol{X}$. Now, consider the equation

$$\boldsymbol{Y} = \boldsymbol{N}_{\boldsymbol{U}}\boldsymbol{C} + \boldsymbol{N}_{\boldsymbol{U}}\boldsymbol{A}_1.$$

Its minimal nonnegative solution is $\boldsymbol{Y}_{\min} = \sum_{\nu\geq 0}(\boldsymbol{N}_{\boldsymbol{U}}\boldsymbol{A}_1)^{\nu}\boldsymbol{N}_{\boldsymbol{U}}\boldsymbol{C}$, provided that the series converges. We may also write that

$$\boldsymbol{Y}_{\min} = \sum_{\nu\geq 1}\sum_{k_1,k_2,\ldots,k_\nu\geq 0}\boldsymbol{U}^{k_1}\boldsymbol{A}_1\boldsymbol{U}^{k_2}\boldsymbol{A}_1\cdots\boldsymbol{A}_1\boldsymbol{U}^{k_\nu}\boldsymbol{C},$$

where the product $[\boldsymbol{U}^{k_1}\boldsymbol{A}_1\boldsymbol{U}^{k_2}\boldsymbol{A}_1\cdots\boldsymbol{U}^{k_{\nu-1}}\boldsymbol{A}_1]$ is empty for $\nu = 1$, and equal to $\boldsymbol{I}$, by convention, in that case. The general term in the series above is the probability that, without violating the taboo of the levels 0 to $n-1$, the process returns $k_1$ times to level $n$, then moves to level $n+1$, where it will return $k_2$ times without returning to level $n$, and then moves to level $n+2$, where it will return $k_3$ times without returning to level $n+1$, etc., until it reaches level $n+\nu$, from which it will drop to level 0 after having made $k_\nu$ returns.

The interpretation of $\boldsymbol{Y}_{\min}$ is therefore

$$(Y_{\min})_{ij} = \sum_{\nu\geq 0}\mathrm{P}[\tau < \infty,\ X_\tau = (0, j),\ L_{\tau-1} = n + \nu \mid X_0 = (n, i)],$$

which shows that the series converges, and that $\boldsymbol{Y}_{\min} = \boldsymbol{H}$. We can easily verify by direct substitution that $\boldsymbol{Y}_{\min}$ is a solution of (A.3) and that $\boldsymbol{X} \geq \boldsymbol{Y}_{\min}$ for any other nonnegative solution $\boldsymbol{X}$, which concludes the proof of the first statement. It is then a simple matter to prove (A.2) when $|S| < \infty$.

## Appendix B. Level-phase independence

Assume that $\boldsymbol{A}_1 = \lambda\boldsymbol{I}$, $\boldsymbol{A}_{-1} = \mu\boldsymbol{I}$, and $\boldsymbol{A}_0 = (1-\lambda-\mu)\boldsymbol{A}$, where $\mu > \lambda > 0$ and $\lambda + \mu < 1$, and that $\boldsymbol{B} = \boldsymbol{A}_{-1} + \boldsymbol{A}_0$. By [6, Lemma 6.3.2], $\boldsymbol{\pi}_n^{\top} = \boldsymbol{\alpha}^{\top}(\boldsymbol{I} - \boldsymbol{R})\boldsymbol{R}^n$ for all $n$, where $\boldsymbol{\alpha}^{\top}$ is the stationary vector of $\boldsymbol{A} = \boldsymbol{A}_{-1} + \boldsymbol{A}_0 + \boldsymbol{A}_1$ and $\boldsymbol{R}$ is the minimal nonnegative

https://doi.org/10.1239/aap/1293113153 Published online by Cambridge University Press

solution of (1.2). To prove that $\boldsymbol{\pi}_n^\top = (1-\rho)\rho^n\boldsymbol{\alpha}^\top$, we need to show that $\boldsymbol{\alpha}^\top$ is an eigenvector of $\boldsymbol{R}$ and that the corresponding eigenvalue is $\rho$. The equation for $\boldsymbol{R}$ is written in this case as

$$\boldsymbol{R} = \lambda\boldsymbol{A} + (1-\lambda-\mu)\boldsymbol{R}\boldsymbol{A} + \mu\boldsymbol{R}^2\boldsymbol{A}$$

and may be solved by functional iteration. This shows that $\boldsymbol{R}$ is of the form $\boldsymbol{R} = s(\boldsymbol{A})$ for some function $s(z) = \sum_{\nu\geq 0} s_\nu z^\nu$, so $\boldsymbol{\alpha}^\top$ is an eigenvector of $\boldsymbol{R}$ and $\boldsymbol{\alpha}^\top\boldsymbol{R} = s(1)\boldsymbol{\alpha}^\top$.

It is readily seen that $s_\nu \geq 0$ for all $\nu$, so $s(z)$ is, for $z \geq 0$, the minimal nonnegative solution of

$$s(z) = \lambda + (1-\lambda-\mu)zs(z) + \mu s^2(z).$$

This minimal solution is

$$s(z) = \frac{1-(1-\lambda-\mu)z - \sqrt{(1-(1-\lambda-\mu)z)^2 - 4\lambda\mu}}{2\mu},$$

and it is easy to verify that $s(1) = \rho$, which completes the proof.

## Acknowledgements

The authors gratefully acknowledge the support of the Australian Research Council through Discovery Project DP0770388. The second author is also pleased to acknowledge the support of the Fonds de la Recherche Scientifique - FNRS, Belgium.

## References

[1] Bini, D. A., Latouche, G. and Meini, B. (2005). *Numerical Methods for Structured Markov Chains*. Oxford University Press.
[2] Çinlar,E. (1975). *Introduction to Stochastic Processes*. Prentice-Hall, Englewood Cliffs, NJ.
[3] Gibson, D. and Seneta, E. (1987). Augmented truncation of infinite stochastic matrices. *J. Appl. Prob.* **24,** 600–608.
[4] Heyman, D. P. and Whitt, W. (1989). Limits for queues as the waiting room grows. *Queueing Systems* **5,** 381–392.
[5] Kroese, D. P., Scheinhardt, W. R. W. and Taylor, P. G. (2004). Spectral properties of the tandem Jackson network, seen as a quasi-birth-and-death process. *Ann. Appl. Prob.* **14,** 2057–2089.
[6] Latouche, G. and Ramaswami, V. (1999). *Introduction to Matrix Analytic Methods in Stochastic Modeling*. SIAM, Philadelphia PA.
[7] Latouche, G. and Taylor, P. G. (2002). Truncation and augmentation of level-independent QBD processes. *Stoch. Proces. Appl.* **99,** 53–80.
[8] Neuts, M. F. (1981). *Matrix-Geometric Solutions in Stochastic Models*. Johns Hopkins University Press, Baltimore, MD.
[9] Ramaswami, V. and Taylor, P. G. (1996). An operator-analytic approach to product-form networks. *Commun. Statist. Stoch. Models* **12,** 121–142.
[10] Seneta, E. (1980). Computing the stationary distribution for infinite Markov chains. *Linear Algebra Appl.* **34,** 259–267.
[11] Wolf, D. (1980). Approximation of the invariant probability measure of an infinite stochastic matrix. *Adv. Appl. Prob.* **12,** 710–726.
[12] Zhao, Y. Q. and Liu, D. (1996). The censored Markov chain and the best augmentation. *J. Appl. Prob.* **33,** 623–629.