Adv. Appl. Prob. **40,** 1157–1173 (2008)
Printed in Northern Ireland

STOCHASTIC AND SUBSTOCHASTIC SOLUTIONS FOR INFINITE-STATE MARKOV CHAINS WITH APPLICATIONS TO MATRIX-ANALYTIC METHODS

WINFRIED K. GRASSMANN,* *University of Saskatchewan*
JAVAD TAVAKOLI,** *University of British Columbia Okanagan*

Abstract

This paper deals with censoring of infinite-state banded Markov chains. Censoring involves reducing the time spent in states outside a certain set of states to 0 without affecting the number of visits within this set. We show that, if all states are transient, there is, besides the standard censored Markov chain, a nonstandard censored Markov chain which is stochastic. Both the stochastic and the substochastic solutions are found by censoring a sequence of finite transition matrices. If all matrices in the sequence are stochastic, the stochastic solution arises in the limit, whereas the substochastic solution arises if the matrices in the sequence are substochastic. We also show that, if the Markov chain is recurrent, the only solution is the stochastic solution. Censoring is particularly fruitful when applied to quasi-birth-and-death (QBD) processes. It turns out that key matrices in such processes are not unique, a fact that has been observed by several authors. We note that the stochastic solution is important for the analysis of finite queues.

Keywords: Infinite-state Markov chain; matrix-analytic method; nonrecurrent Markov chain; queueing

2000 Mathematics Subject Classification: Primary 60J10; 60K25

1. Introduction

For an infinite-state, discrete-time Markov chain, much information can be gained by censoring. Censoring involves ignoring the sample function of a Markov chain when it is outside of a certain subset E of states, without affecting the visits when the process is in this subset. Stated differently, the process is embedded into E. It is important to note that censoring can be used to provide a probabilistic interpretation to Gaussian elimination. In this paper we assume that the states are the natural numbers, starting at 1, and we will typically censor the states above a certain state n.

Censoring also allows us to deal with transient Markov chains which have no equilibrium probabilities: following Zhao *et al.* [24], we pick a base state, say state 1, and, for each state i of the Markov chains, we use v_i, the expected number of visits to state i in a cycle, where a cycle starts before entering the base state and ends before the base state is entered the next time. This measure obviously also exists if all states are transient, and it can take the place of the equilibrium probabilities which are only available if the process is positive recurrent. If the Markov chain is positive recurrent, the variables v_i, $i = 1, 2, \ldots$, are proportional to the

Received 30 April 2008; revision received 24 October 2008.

* Postal address: Department of Computer Science, University of Saskatchewan, 101 Science Court, Saskatoon SK S7N 5C9, Canada. Email address: grassman@cs.usask.ca

** Postal address: Department of Mathematics, Statistics and Physics, University of British Columbia Okanagan, 3333 University Way, Kelowna, Canada V1V 1V7. Email address: javad.tavakoli@ubc.ca

https://doi.org/10.1239/aap/1231340168 Published online by Cambridge University Press

equilibrium probabilities. Moreover, this measure fits very naturally with the methodology of censoring.

Clearly, censoring is also applicable for finite or infinite substochastic matrices. Here, as later in this paper, the term *substochastic* means *strictly substochastic*, that is, stochastic matrices are excluded.

Kemeny *et al.* [11] pointed out that inverses of infinite-dimensional matrices are not unique, and this affects the uniqueness of a censored Markov chain. Independently, researchers in matrix-analytic methods also found that certain key matrices used in this area are nonunique [1], [9], [14], [17]. As shown in [6], censoring provides the basis of matrix-analytic methods, and, hence, the nonuniqueness of certain matrices in matrix-analytic methods follows from the nonuniqueness of inverses in infinite-state Markov chains.

We will mainly consider discrete-time Markov chains, but as we will point out, the derivation also applies to continuous-time Markov chains. For instance, instead of the number of visits between two consecutive visits to a base state, we have to consider the expected time spent in state i between two consecutive visits to the base state, and divide this number by the expected time spent uninterruptedly in the base state.

To avoid complications, we assume that all states communicate, which implies that all states are either recurrent or transient. We also assume that the transition matrices are banded, but the theory should also be applicable in the case in which the matrices are *weakly banded*, that is, where a bandwidth can be chosen such that the probability of a transition outside the band is less than $\varepsilon > 0$. Now, let P be the transition matrix of an infinite-state, discrete-time Markov chain with states $1, 2, 3, \dots$, let ${}_N P$ be a sequence of finite transition matrices of size N, and let $N \to \infty$. We require that the transition matrices, ${}_N P$, have the same transition probabilities as the original matrix, P, except for states close to N. Of course, since the matrices ${}_N P$ are finite, all transition probabilities involving any state $k > N$ are removed. The transition probabilities close to N can essentially be chosen arbitrarily, as long as ${}_N P$ is stochastic or substochastic, and as long as all states communicate. Now consider the matrices ${}_N P^{(n)}$, which are obtained from the matrices ${}_N P$ by censoring above state n, and consider the existence of and the meaning of the limit of ${}_N P^{(n)}$ as $N \to \infty$. Here the limits are to be understood pointwise. Since the transition probabilities of the ${}_N P$ close to N may differ, we can construct many sequences $\{{}_N P,\ N > 0\}$, but if the process is recurrent, all these sequences approach the same limit. However, if the process is transient, there are typically exactly two limits. If, in the transient case, all the ${}_N P$ matrices are stochastic, so is the limit of ${}_N P^{(n)}$, but if the matrices are substochastic, a different limit arises. Hence, there is typically a stochastic limit, which is different from the substochastic limit. In the literature, people emphasize the substochastic limit. However, the stochastic limit turns out to be important for the analysis of queueing systems with traffic intensities above 1, but with a finite waiting room N. In this case, the resulting Markov chain has a stochastic transition matrix for all N, even as $N \to \infty$. Hence, in this case, it is the stochastic solution that is of interest, not the substochastic solution.

We have to exclude sequences of matrices where some members are stochastic while other members are substochastic, because in this case, there may be a subsequence which is stochastic and another subsequence which is substochastic. We could, of course, allow a mixture of stochastic and substochastic matrices for the first few elements of the sequence $\{{}_N P\}$, but we will not do this because no additional insight would be gained.

If a process has only transient states, or if it has absorbing states, then it no longer has nonzero equilibrium probabilities, but it still has nonzero expected visits in a cycle. On the other hand, if all states of the original Markov chain P are transient, but ${}_N P$ is stochastic, then

https://doi.org/10.1239/aap/1231340168 Published online by Cambridge University Press

${}_NP$ has nonzero equilibrium probabilities. However, as we let N go to ∞, it takes longer and longer to reach the equilibrium when starting in any state of sufficient distance from N, and in the limit, it takes an infinite amount of time to reach equilibrium.

The outline of this paper is as follows. In Section 2 we discuss censoring, and in Section 3 we apply this theory to transient and recurrent Markov chains. In Section 4 we show the connection with matrix-analytic methods, and in Section 5 we expand these considerations by using eigenvalues. Since the main emphasis of this paper involves connecting different theories, we will review some results that have appeared in the literature. This means that this paper is partially a review article, in the sense that it provides a new perspective to known results.

2. Censoring

Consider a discrete-time Markov chain $\{X_1, X_2, X_3, \dots\}$, where the X_n, $n \geq 1$, can assume only the values $1, 2, 3, \dots, N$, where N may be finite or infinite. We partition the states into two sets, namely from 1 to n and from $n + 1$ to N. If we partition the transition matrix P conformally, we obtain

$$P = \begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix}.$$

After censoring all states above n, we obtain a reduced Markov chain $P^{(n)}$, which becomes, according to [11],

$$P^{(n)} = P_{11} + P_{12} Q P_{21}, \tag{1}$$

where

$$Q = \sum_{\nu=0}^{\infty} P_{22}^{\nu}.$$

As shown in [11], $(Q)_{ij}$ gives the expected number of visits to state $j > n$ before returning to a state $k \leq n$, given that the process starts in state $i > n$. Also, $(P_{12}Q)_{ij}$ gives the expected number of visits to state $j > n$ before returning to a state $k \leq n$, given that the process starts in state $i \leq n$.

Note that, if I is the identity matrix, Q is the inverse of $I - P_{22}$, because

$$(I - P_{22})Q = (I - P_{22})(I + P_{22} + P_{22}^2 + \cdots) = I - P_{22} + P_{22} - P_{22}^2 + P_{22}^2 - \cdots = I.$$

For finite N, the inverse is unique, and (1) really represents block elimination. In the case of infinite matrices, Kemeny *et al.* [11, p. 5] have shown that $I - P_{22}$ can have more than one inverse. We note, however, that Kemeny *et al.* used the inverse with the smallest entries.

Let the original process be $\{X_1, X_2, X_3, \dots\}$, and let $\{X_1^{(n)}, X_2^{(n)}, X_3^{(n)}, \dots\}$ be the process censored at level n. Hence, if m_ν indicates the epoch where X_m is less than or equal to n for the νth time then $X_\nu^{(n)} = X_{m_\nu}$. This establishes a one-to-one correspondence of the visits to any state $i \leq n$ in the processes $\{X_m,\ m \geq 1\}$ and $\{X_\nu^{(n)},\ \nu \geq 1\}$, that is, the number of visits to state $i \leq n$ is not affected by censoring. In particular, if the number of visits to state i in the original process is finite, so is the number of visits in the censored process. It follows that, if state i is transient in the original Markov chain, it is also transient in the censored Markov chain, and if the number of visits is infinite in the original chain, so is the number of visits in the censored Markov chain. Therefore, if the process $\{X_m,\ m \geq 1\}$ is recurrent, so is the process $\{X_\nu^{(n)},\ \nu \geq 1\}$, and vice-versa. A similar result follows if $\{X_m,\ m \geq 1\}$ is transient. If the number of states is finite, and if all states communicate, recurrent Markov chains

https://doi.org/10.1239/aap/1231340168 Published online by Cambridge University Press

have stochastic transition matrices and transient Markov chains have substochastic matrices, provided that all the rows and columns corresponding to absorbing states are deleted. Hence, for such Markov chains, censoring converts stochastic matrices into stochastic matrices and substochastic matrices into substochastic matrices.

As mentioned in the introduction, we use the expected number of visits, v_i, to state i, $i = 1, 2, \dots$, between two consecutive visits of a particular state, say state 1. Specifically, we start counting visits immediately after state 1 is entered, and we stop immediately before state 1 is entered again. It follows that $v_1 = 1$. Note that, for any noncensored state i, the variable v_i keeps its value when the process is censored at n, provided that $i \leq n$. To find the values of v_i for all $i > 1$, we follow Zhao *et al.* [24] and use (1) with $n = 1$, in which case $P_{12}Q$ becomes a vector indicating the expected number of visits to states $2, 3, 4, \dots$ between two consecutive visits to state 1. If v^- is defined to be $[v_2, v_3, \dots]$, we have, since $QP_{22} = P_{22} + P_{22}^2 + \cdots = Q - I$,

$$v^- = P_{12}Q = P_{12} + P_{12}QP_{22} = P_{12} + v^- P_{22}.$$

Since $v_1 = 1$, this expands to

$$v_j = p_{1j} + \sum_{i=2}^{\infty} v_i p_{ij} = \sum_{i=1}^{\infty} v_i p_{ij}, \qquad j > 1. \tag{2}$$

This proves that the v_j, $j \geq 1$, satisfy the equilibrium equations, except for $j = 1$. In fact, if P is substochastic, (2) will fail for $j = 1$. On the other hand, if P is stochastic, (2) also holds for $j = 1$, which implies that the equilibrium vector $\pi = [\pi_i,\ i = 1, 2, \dots]$ can be found from the vector $v = [v_1, v_2, \dots]$ as

$$\pi_i = \frac{v_i}{\sum_{j=1}^{\infty} v_j}.$$

From the definition of censoring we conclude that, for any state $i \leq n$, the probability of ever returning to this particular state is not affected by censoring at n. Hence, once a Markov chain is reduced to a single state, say state 1, then we can conclude that $p_{11}^{(1)}$, the only element of $P^{(1)}$, provides the probability that a process starting in state 1 will return to state 1 again. If $p_{11}^{(1)} = 1$, the process is recurrent, and if $p_{11}^{(1)} < 1$, the process is transient. In other words, $1 - p_{11}^{(1)}$ is the probability of escape and $1/(1 - p_{11}^{(1)})$ is the expected number of visits to state 1 from now to ∞ if starting in state 1.

If state 1 is transient, there will be a last visit to state 1. After this visit, other states are visited, and these are counted when calculating v_i. This follows from the equation

$$P_{12}Q = P_{12} + P_{12}P_{22} + P_{12}P_{22}^2 + \cdots.$$

To find $P^{(n)}$, we can eliminate one variable at a time, rather than through block elimination. We apply (1) with N finite and $n = N - 1$ to obtain, provided that $p_{NN} \neq 1$,

$$p_{ij}^{(N-1)} = p_{ij} + \frac{p_{iN}p_{Nj}}{1 - p_{NN}}. \tag{3}$$

This equation indicates that in order to go from i to j in the reduced Markov chain we can either go from i to j directly or we can go from i to N, and from there to j. It follows that, if there is a path from i to j in the original Markov chain described by P, there is also a path from i to j in the reduced Markov chain described by $P^{(N-1)}$. Hence, the application of (3), which we refer to as state reduction, will preserve paths.

https://doi.org/10.1239/aap/1231340168 Published online by Cambridge University Press

It is clear that the transition matrix $P^{(N-1)}$ can be reduced by using (3), except that N is replaced by $N-1$. This reduction continues, and to express this, we write

$$p_{ij}^{(n-1)} = p_{ij}^{(n)} + \frac{p_{in}^{(n)} p_{nj}^{(n)}}{1 - p_{nn}^{(n)}}. \tag{4}$$

This reduction process stops when only state 1 is left. At this point, we set $v_1 = 1$. To find the v_n, $n > 1$, we use

$$v_n = \sum_{i=1}^{n-1} v_i \frac{p_{in}^{(n)}}{1 - p_{nn}^{(n)}}, \qquad n = 2, 3, \ldots, N.$$

This relation is true because state reduction essentially amounts to Gaussian elimination; we leave it to the reader to fill in the details. Since we assumed that all states communicate, $p_{nn}^{(n)}$, $n > 1$, cannot be equal to 1.

Even though this is somewhat off the topic of this paper, we note that the method works even for decomposable Markov chains. In this case, $p_{nn}^{(n)} = 1$ as soon as all recurrent states of the communicating class of this chain are found. Any state i with $p_{in} > 0$ is transient and can be deleted. The transition matrix of the remaining Markov chain can then be solved separately.

If the transition matrix is stochastic, $1 - p_{nn}^{(n)}$ must be equal to $\sum_{j=1}^{n-1} p_{nj}^{(n)}$, and (4) can be written as

$$p_{ij}^{(n-1)} = p_{ij}^{(n)} + \frac{p_{in}^{(n)} p_{nj}^{(n)}}{\sum_{k=1}^{n-1} p_{nk}^{(n)}}. \tag{5}$$

This approach has been called the Grassmann–Taksar–Heyman (GTH) method; for details, see [18], [19], and [20]. It is known [5] that the GTH method ensures stability by avoiding subtractions. In our context, it is important to note that, after each step of the GTH algorithm, the matrix $P^{(n-1)} = [p_{ij}^{(n-1)}]$ is a stochastic matrix.

3. Stochastic and substochastic limits for censored Markov chains

In this section we assume that the Markov chain $\{X_t\}$ is banded, that is, there is a value b such that

$$\mathrm{P}\{|X_{t+1} - X_t| > b\} = 0. \tag{6}$$

In addition to this we require that there exist positive values m and q such that the probability of reaching j from i in m steps or less is at least q. This makes it impossible that, as i and j go to ∞, with $|i - j| \le b$, the shortest path between these two states becomes infinite. If this condition is met, we say that i and j *communicate strongly*. Thus, we require that all states of P communicate strongly.

We now consider sequences of finite-dimensional matrices starting at $n+1$, say ${}_{n+1}P$, ${}_{n+2}P, \ldots, {}_{N}P, \ldots$, where N is the dimension of ${}_{N}P$. We require that these matrices satisfy the following conditions.

1. They are banded with bandwidth b.
2. For all N, $({}_{N}P)_{ij} = (P)_{ij}$ for $1 \le i, j < N - b$.
3. All states of ${}_{N}P$ communicate strongly.
4. The row sums for all rows at or below $N - b$ must be equal to 1.

https://doi.org/10.1239/aap/1231340168 Published online by Cambridge University Press

There are many sequences $\{{}_NP,\ N > n\}$ that satisfy conditions 1–4. In particular, we can truncate the matrix P at N, that is, remove all the rows and columns of P above N without changing the entries $p_{ij},\ i, j \le N$, to obtain ${}_NP$. If all states are communicating, truncation will always lead to a substochastic matrix: truncating at N can result in a stochastic matrix only if no state above N can be reached from the states at or below N. As an alternative to the truncation, we can do the cut as before, except that column N is changed in such a way that the row sums are equal to 1. In the first case we obtain a sequence of substochastic matrices, and in the second case we obtain a sequence of stochastic matrices. Of course, there are other methods to obtain matrices ${}_NP$ satisfying conditions 1–4, leading to what we call different *types* of sequences. As mentioned earlier, we assume that either all members of the sequence are stochastic or that they are all substochastic. The substochastic matrices will be denoted by ${}_N\underline{P}$, and the stochastic matrices will be denoted by ${}_N\overline{P}$. To avoid the ${}_N\underline{P}$ converging toward a stochastic matrix, we require that, for all sufficiently large N, ${}_N\underline{P}$ has at least one row with a row sum less than $1 - p$ with $p > 0$, where p is independent of N.

Equation (1) can be applied to any ${}_NP$. We will also use the preceding subscript N for the values corresponding to P_{22}, Q, and other entities we will need. In other words, ${}_NP_{22}$ is the value of the matrix corresponding to P_{22} in (1), and a similar relation exists between ${}_NQ$ and Q, as well as between ${}_NP^{(n)}$ and $P^{(n)}$. If we want to stress that ${}_NP$ is a stochastic matrix, we will denote the corresponding censored matrix by ${}_N\overline{P}^{(n)}$. We also use ${}_N\overline{Q}$ and other entities with overbars. Similarly, for the substochastic case, we use ${}_N\underline{P}^{(n)}$, ${}_N\underline{Q}$, and so on.

We will make extensive use of limits. In the case of finite-dimensional matrices, these limits must always be understood pointwise. In the case of infinite matrices we will avoid limits where possible because they lead to mathematical intricacies we do not want to deal with here (see, e.g. [11, p. 33]). In most cases we can cast the problem such that only finite-dimensional matrices are needed. For instance, though the matrix Q may be infinite, we only need its first b rows and columns, and using only these rows and columns leads to a finite matrix, say Q'. In this way, whether or not ${}_NQ'$ has a (pointwise) limit in some context is a well-defined question. In the case of ${}_N\overline{P}$ we use the following convention: if we use $\overline{P}$ without the preceding subscript then we implicitly assume that $\overline{P}$ is equal to ${}_N\overline{P}$ with sufficiently large N. For instance, if we need a state M such that, for every state above M, a particular statement holds, we implicitly assume that $N >> M$.

Our main result is that, if the Markov process is recurrent, every type of sequence $\{{}_NP^{(n)},\ N > n\}$ of matrices satisfying conditions 1–4 has the limit $P^{(n)}$, that is,

$$\lim_{N\to\infty} {}_N\underline{P}^{(n)} = \lim_{N\to\infty} {}_N\overline{P}^{(n)} = P^{(n)}.$$

However, in the case where P is a transient Markov chain, all types of sequences of substochastic matrices $\{{}_N\underline{P}^{(n)}\}$ have the same substochastic limit $P^{(n)}$, and all sequences of stochastic matrices have the same stochastic limit, which will be denoted by $\overline{P}^{(n)}$. Since, according to our definition, a substochastic matrix cannot be stochastic, the two limits must be different. Of course, in both cases we assume that conditions 1–4 are satisfied.

Regarding the recurrent case, we have the following theorem. We presented a similar theorem earlier [8] (see also [23]), but the following proof is more rigorous than the one in [8].

Theorem 1. *If P is recurrent then sequences $\{{}_NP^{(n)},\ N > n\}$ of all types satisfying conditions 1–4 converge pointwise to the same limit as $N \to \infty$, that is,*

$$\lim_{N\to\infty} {}_NP^{(n)} = P^{(n)}.$$

https://doi.org/10.1239/aap/1231340168 Published online by Cambridge University Press

Proof. Since the process is recurrent, the probability that the process never returns to a state $k \leq n$ is 0, which implies that the time T until the return has a proper distribution. Hence, we can find, for each $\varepsilon > 0$, a t such that $\mathrm{P}\{T > t\} < \varepsilon$. If the process is banded with bandwidth b then, during time t, we cannot reach any state above $n + bt$, that is,

$$\mathrm{P}\{n + bt + 1 \text{ reached before return to } k \leq n \mid T \leq t\} = 0.$$

Let $d_{ij}^{(s)} = (P_{12} P_{22}^{s} P_{21})_{ij}$. Clearly,

$$d_{ij}^{(s)} = \mathrm{P}\{X(s+2) = j,\ X(\tau) > n,\ \tau = 2, 3, \ldots, s+1 \mid X_1 = i\}.$$

Hence,

$$\begin{aligned} d_{ij} &= \sum_{s=0}^{\infty} d_{ij}^{(s)} \\ &= \sum_{s=0}^{t} d_{ij}^{(s)} + \varepsilon \\ &= \sum_{s=0}^{t} \mathrm{P}\{X(s+2) = j,\ n < X(\tau) < n + bt + 1 \mid X_1 = i\} + \varepsilon. \end{aligned}$$

These probabilities are the same for P and any ${}_N P$ with $N > n + bt$. Since (1) implies that $p_{ij}^{(n)} = p_{ij} + d_{ij}$, the result follows.

As stated earlier, in the transient case, ${}_N\underline{P}^{(n)}$ and ${}_N\overline{P}^{(n)}$ have different limits. We have the following theorem.

Theorem 2. *Let P be the transition matrix of a strongly communicating transient Markov chain, and suppose that P has a bandwidth of b, as defined in (6). Let $\{{}_N\underline{P}\}$ be a sequence that, for sufficiently large N, consists of substochastic matrices satisfying conditions 1–4. For all types of such sequences, the limit of the corresponding censored Markov chain is equal to $P^{(n)}$, provided that, for sufficiently large N, there exists a $p > 0$ such that every ${}_N\underline{P}$ has at least one row i, $i = N - b + 1, N - b + 2, \ldots, N$, with a row sum less than $1 - p$.*

Proof. First, we assume that ${}_N\underline{P}$ is a truncated version of $P = [p_{ij},\ i, j \geq 1]$, that is, ${}_N\underline{P} = [p_{ij},\ i, j \leq N]$. We also define ${}_N\tilde{P}$ to be the infinite-dimensional matrix obtained by padding ${}_N\underline{P}$ with 0s, that is, ${}_N\tilde{P} = [\tilde{p}_{ij},\ i, j = 1, 2, \ldots]$, with $\tilde{p}_{ij} = p_{ij}$ for $i, j \leq N$ and $\tilde{p}_{ij} = 0$ for $i > N$ or $j > N$. We now use (1) to find $P^{(n)}$ and ${}_N\underline{P}^{(n)}$. It is clear that, except for the padding 0s, (1) gives the same value for ${}_N\underline{P}^{(n)}$, whether we use ${}_N\underline{P}$ or ${}_N\tilde{P}$. Since the entries of ${}_N\tilde{P}$ are nondecreasing with N, ${}_N\underline{P}^{(n)}$ given by (1) is nondecreasing in N. Also, the entries of ${}_N\underline{P}^{(n)}$, being probabilities, are bounded. It follows that in the case where P is truncated, ${}_N\underline{P}^{(n)}$ converges to $P^{(n)}$ as $N \to \infty$.

To deal with the general case, we show that we can always find an M such that the transition probabilities of states above M have only a negligible effect after censoring at $n < M$. Given such a value M, let ${}_M\hat{P} = [\hat{p}_{ij}]$ be an infinite-state Markov chain such that, for $i, j \leq M$, $\hat{p}_{ij} = p_{ij}$. For $i > M$ or $j > M$, the probabilities $\hat{p}_{ij}$ must be chosen such that ${}_M\hat{P}$ remains banded with all states strongly communicating, and at least one row of ${}_M\hat{P}$ above M must have a row sum less than or equal to $1 - p$. Otherwise, these probabilities can be chosen arbitrarily. We now prove that, for each $\varepsilon > 0$, there exists an M such that the entries of ${}_M\hat{P}^{(n)}$ differ

https://doi.org/10.1239/aap/1231340168 Published online by Cambridge University Press

from the corresponding entries of $P^{(n)}$ by at most ε. To prove this, we define d_{ij} to be equal to $(P_{12}QP_{21})_{ij}$, as we did in Theorem 1, and we write

$$d_{ij} = d_{ij}^{M-} + d_{ij}^{M+}.$$

Here, d_{ij}^{M-} is the probability that a sample path starting in state $i \leq n$ will reach a state above n and, after returning to a state at or below n for the first time, the sample path will hit state j, and that during this excursion, no state above M is visited. The probability d_{ij}^{M+} must be interpreted in a similar fashion, except that at least one state above M is visited.

Now it is enough to show that, for each $\varepsilon > 0$, we can find an M such that $d_{ij}^{M+} \leq \varepsilon$. Obviously, d_{ij}^{M+} is less than or equal to the probability of ever returning to any state at or below n once a state above M is visited. Therefore, if we can find an M for each $\varepsilon > 0$ such that d_{ij}^{M+} is less than ε, then we ensure that d_{ij} and with it $p_{ij}^{(n)}$ change by less than ε. To construct such an M, let $\phi_{i'}$ be the probability of absorption once in state i', $M - b < i' \leq M$, without returning to any state at or below $M - b$, and let ψ be an upper limit of the probability that, once a state at or below $M - b$ is reached, a state at or below n is reached before returning to a state above $M - b$. We set $\phi = \max_{i'}\{\phi_{i'}\}$ and $\overline{\phi} = \max_{i'}\{(1 - \phi_{i'})\}$. Now consider an excursion, that is a process starting in state j', $M - b < j' \leq M$, visiting states at or below $M - b$ before returning to a state in $(M - b, M]$, at which time the excursion ends. The probability of having exactly ν such excursions before absorption is bounded by $\overline{\phi}^{\nu}\phi$. In each excursion, the probability of visiting a state at or below n has an upper bound ψ, and the probability of having ν excursions without visiting any state at or below n is $(1 - \psi)^{\nu}$. Therefore, the probability of visiting a state at or below n at least once has an upper bound of $1 - (1 - \psi)^{\nu}$. If $\psi \to 0$, this bound approaches $\nu\psi$. Hence, the probability of visiting a state $j \leq n$ is bounded by

$$\sum_{\nu=1}^{\infty} \overline{\phi}^{\nu}\phi\nu\psi = \frac{\overline{\phi}\phi\psi}{(1 - \overline{\phi})^2}.$$

The probability d_{ij}^{M+} is less than this expression because, if we do not return to any state at or below n, we will not return to the state $j \leq n$. Hence, if we find an M such that $\overline{\phi}\phi\psi/(1 - \overline{\phi})^2 < \varepsilon$, we have accomplished what we set out to do. This leads to

$$\psi < \frac{\varepsilon(1 - \overline{\phi})^2}{\overline{\phi}\phi}.$$

To find an M such that ψ satisfies this condition, note that, owing to our assumption, eventual absorption is certain, and before absorption, a state in the range $(M - b, M]$ must be visited. Hence, If T is the length of the time interval between the first moment a state $i' \leq M - b$ is visited and the first moment a state j' in $(M - b, M]$ is visited, then T must have a proper distribution. Therefore, for each ε_1, there exists a t such that $\mathrm{P}\{T > t\} < \varepsilon_1$. In a time interval of length t or less we cannot decrease the present state i by more than tb. Hence, to go from M to $n - b + 1$, which is the worst case, we need at least $(M - n - 1)/b + 1$ steps. Thus, in order to have $d_{ij}^{M+} < \varepsilon$, M must be chosen such that

$$\mathrm{P}\left\{T > \frac{M}{b} + 1\right\} < \frac{\varepsilon(1 - \overline{\phi})^2}{\overline{\phi}\phi},$$

and this completes the proof.

https://doi.org/10.1239/aap/1231340168 Published online by Cambridge University Press

The stochastic (nonstandard) solution is also unique, as Theorem 3, below, indicates. To show this, we need the following lemma together with its proof.

Lemma 1. *If $M \geq n + (m + \nu)b$ then the probability that there are no values $\tau < \tau^*$ and $\nu < \nu^*$ satisfying $Y_\tau = Y'_\nu$ is less than $(1-q)^\nu$. Here, τ^* is the smallest value satisfying $Y_. \leq n$ and ν^* is the corresponding time for the process $Y'_.$.*

Proof. First consider a state i, $n < i \leq M$. If τ is the first time $Y'_. \leq i$ then $0 \leq i - Y'_\tau \leq b$. Hence, the probability that $Y'_{\tau+u} = i$, $u \leq m$, is at least q, provided that no state at or below n is reached first. Note that the lowest point $Y'_{\tau+\nu}$, $\nu < u$, can reach is above $i - mb$. Hence, no state at or below n is reached if $M \geq n + (m + \nu)b$. This completes the proof.

Theorem 3. *Let P be the transition matrix of a transient Markov chain with bandwidth b and strongly communicating states. Consider a sequence of any type of stochastic matrices $_N\overline{P}$ satisfying conditions 1–4. Then, the corresponding sequences $\{_N\overline{P}^{(n)}\}$ all have the limit $\overline{P}^{(n)}$.*

Proof. For simplicity, we use the matrix $\overline{P}$ which, as defined earlier, is really the matrix $_N\overline{P}$, where N is sufficiently large. We will show that, for each $\varepsilon > 0$, there exists an M such that, if entries of the transition matrix $\overline{P}$ above M are changed in such a way that the matrix remains stochastic, the entries of the censored version of $\overline{P}$ change by at most ε. Since the process is recurrent, both before and after the change, the return to a state at or below n is assured. To find M such that $\overline{P}^{(n)}$ changes by at most ε, we need to consider the processes starting at $k \leq n$ and ending at $j \leq n$. Clearly, all sample functions of this process reaching no state above M will not be affected by the changes above row M. The sample functions that do reach states above M must return to a state in the range $(M - b, M]$ on their way down. Hence, we only need to concern ourselves with what happens after a state i in $(M - b, M]$ is reached. We now show that, for sufficiently large M, the probability of returning to a particular state j, $j \leq n$, is essentially the same whether the process starts in i or $i' \neq i$, $M - b < i, i' \leq M$. To show this, consider the two processes Y_t and Y'_t with $Y_1 = i$ and $Y'_1 = i'$, where both processes are governed by the transition matrix $\overline{P}$. The strong Markov property implies that if there exist values t_1 and t_2 such that $Y_{t_1} = Y'_{t_2}$, then $Y_{\tau+t_1}$ is stochastically indistinguishable from $Y'_{\tau+t_2}$ for $\tau \geq 0$. If the states of the matrix $\overline{P}$ are strongly communicating (that is, if there exist values m and $q > 0$ such that there exists a $\nu \leq m$ for which $p_{ij}^{[\nu]} > q$, $|i - j| \leq b$), we can choose an M for each $\varepsilon_1 > 0$ such that the event $Y_{t_1} = Y'_{t_2}$ will occur before reaching $j \leq n$ with a probability greater than $1 - \varepsilon_1$.

Since $\overline{P}$ is banded, there must be a τ for each interval given by $(M - kb, M - (k-1)b]$, $k = 1, 2, \dots$, such that $M - kb < Y_\tau \leq M - (k-1)b$. Let i_k be the value assumed by Y_t while in the kth interval for the first time. Clearly, for each i_k, there exists a τ satisfying $Y'_\tau \leq i_k$ for the first time, and, hence, a probability q that the process $\{Y'_r\}$ reaches i_k. If $M \geq \nu b + n$, there are ν values of i_k, $k = 1, 2, \dots, \nu$, above state n. However, before encountering i_k, the process $\{Y'_.\}$ could potentially fall below $i_k - mb$. To ensure that this does not occur, we apply Lemma 1 to set $M = (\nu + m)b$, where ν is chosen such that $(1-q)^\nu < \varepsilon$. This completes the proof.

In the Markov chain with the transition matrix $_N\overline{P}$, all states are recurrent and they, therefore, have equilibrium probabilities. However, as $N \to \infty$, it takes longer and longer to reach this equilibrium when starting in a state less than some fixed value k. To see this, using arguments similar to the one used to prove Theorem 1, we can find a value k^*, which is independent of N for sufficiently large N, such that the probability of visiting a state $i \leq N - k^*$ is less than ε. Hence, the equilibrium probabilities of states $i \leq N - k^*$ must be small. If we start with a state

https://doi.org/10.1239/aap/1231340168 Published online by Cambridge University Press

below k then, obviously, it takes at least $(N-k^*-k)/b$ steps to reach k^*, and if N goes to ∞, $(N-k^*-k)/b$ goes to ∞ as well. Hence, even though an equilibrium exists, it may take very long to reach it when starting from a given state.

We claim that, in transient Markov chains, the stochastic solution is possibly more meaningful than the substochastic solution. To see this, consider a queueing system with finite waiting room. No matter how large the waiting room, the system is recurrent, even in the case where the arrival rate exceeds the service rate. In these cases, from a certain size of the waiting room onward, certain features of the solution may be independent of the buffer size, and it then makes sense to let the buffer size go to ∞, and this implies that N, the number of states, also goes to ∞.

If the stochastic solution is useful then we need algorithms to obtain it. One way to do this is to start with a sufficiently large value of N and apply (4). Unfortunately, if, through rounding, the sum across the row becomes slightly less than 1, the matrix becomes substochastic rather than stochastic and, as a consequence, the solution will tend toward the substochastic solution $\underline{P}^{(n)}$ rather than $\overline{P}^{(n)}$. To avoid this, we can use (5) which, in contrast to (4), will always assure that the sum across the row remains 1. Hence, when (5) is used, no numerical instability is expected. This agrees with the findings of [5], [18], and [19] regarding the stability of the GTH method.

To show the implications of Theorems 1, 2, and 3, consider the M/M/1 queue with an arrival rate of λ and a service rate of μ. After uniformization, we obtain $p=\lambda/(\lambda+\mu)$ for the probability that the line increases and $1-p$ for the probability that the line decreases. Of course, if the system is empty, the line remains at 0 with probability $1-p$. Theorem 1 indicates that, if the process is recurrent ($p<0.5$) and if N is sufficiently large, the entries of the transition matrix $P^{(n)}$ will change by less than ε if transition probabilities above $N-b$ are changed. On the other hand, Theorem 2 indicates that, for $p>0.5$, we obtain the substochastic matrix $\underline{P}^{(n)}$ if the matrices $_NP$ are all substochastic, and Theorem 3 indicates that we obtain the stochastic matrix $_N\overline{P}^{(n)}$ if the matrices $_NP$ are all stochastic.

We note that $p_{nn}^{(n)}$ is the only entry of the reduced matrix affected by (4), and because of this, we define $d_n=p_{nn}^{(n)}$. Since $p_{n-1\,n-1}^{(n-1)}=0$, (4) yields

$$d_{n-1}=\frac{p(1-p)}{1-d_n},\qquad n=N,N-1,\dots,1. \tag{7}$$

This equation has two fixed points, namely $d_n=p$ and $d_n=1-p$. The solution resulting in a stochastic transition matrix is $d_n=p$.

Note that, since there are exactly two fixed points, one of them must be unstable. If both were stable, there would have to exist one third unstable fixed point, separating the range of d_N going to the two different stable fixed points. This confirms our observation that, if one of the row sums is below 1, even if only by a tiny amount, we obtain the substochastic matrix $\underline{P}^{(n)}$.

We ran recursion (7) in Microsoft Excel® for different values of p and initial values d_N. For $p=0.4$, d_n always converged toward $p=0.4$, and it did not matter whether we started with $d_N=0$ or $d_N=p$, as was to be expected due to Theorem 1. If $p=0.6$, the Markov chain is transient, and we still obtain the stochastic solution if we set $d_N=0.6$. Indeed, d_n remained 0.6 in Excel. However, starting with $d_N=0.599\,99$ made (7) converge to 0.4 rather than 0.6. The variable v_i for the two cases is also different: if d_i approaches 0.4, $v_0=v_1=\cdots=v_N=1$, whereas, if d_1 becomes 0.6, $v_i=1.5^i$, as the reader may verify. This second solution leads to the standard solution for queues with finite waiting rooms. Note that $d_0>0.6$ should be avoided because this causes d_n to increase until $1-d_n\le 0$, and this may cause subtractive cancellation or even a division by 0.

https://doi.org/10.1239/aap/1231340168 Published online by Cambridge University Press

Of course, using the GTH algorithm, we ensure that the system remains stochastic. Using (5), $1 - d_n$ in (7) is replaced by the sum across the row, that is, by $1 - p$, and we trivially find $d_n = p$, even if $p > 0.5$

The best way to visualize the situation given by Theorems 1, 2, and 3 is by thinking of a gas in a container which is subject to Brownian motion, which is a continuous Markov chain in time and space. However, for our analogy, we can think of this continuous chain being discretized. The variables v_i, $i \geq 1$, are then the expected number of molecules in state i in relation to the molecules in state 1. We look at containers of finite height N, and we increase the height. If the gas in the container is heavier than air, say it is CO_2, then, as we let N go to ∞, hardly any gas will escape, even if the container is open. Though, theoretically, CO_2 will eventually diffuse, this diffusion is so small that we almost have an equilibrium, as indicated by the vector v. This situation corresponds to Theorem 1. On the other hand, Theorem 2 deals with a container filled with a gas that is lighter than air, say H_2, which is open at the top. In this case, the gas will obviously drift out of the container, and there is no meaningful equilibrium, that is, the only equilibrium is the state in which all the gas has disappeared. Of course, if the container is closed at the top, Theorem 3 applies, and any gas lighter than air accumulates at the top. Hence, the gas can be expected to become denser as one goes to the top of the container, and the same can be expected to hold in Markov chains. We will use the eigenvalues given in Section 5 to investigate this phenomenon in the case of quasi-birth-and-death (QBD) processes.

Note that an equilibrium is eventually reached for any finite N, though not in any finite time if N is infinite. If the container is open, and the gas is lighter than air, then no equilibrium is reached because the gas drifts out of the container, and this reflects the substochastic solution. A closed container, on the other hand, will prevent any drift. In fact, since a drift involves a change in expectations, there cannot be a drift in any system in equilibrium. However, there can be a potential drift. For CO_2, for instance, the potential drift would be to the bottom of the container and, for H_2, it would be toward the top of the container. Generally, the potential drift is defined as the drift that will arise once the boundaries are removed, that is, once the container is opened both at the top and at the bottom.

4. Applications to matrix-analytic methods

We now apply the theory derived in the previous section to matrix-analytic methods, including the M/G/1 paradigm, the GI/M/1 paradigm, and the QBD process. In all these paradigms, there are levels, ranging from 0 to ∞, and phases, ranging from 1 to M, except that in level 0, the phases range from 1 to M_0. Level 0 will be called the *boundary level*. Also, unless the process enters or leaves the boundary level, the transition probabilities are independent of the level.

In the GI/M/1 paradigm we can go at most one level up, and in the M/G/1 paradigm we can go at most one level down. In the QBD process we can change the level by at most one, which makes the QBD process satisfy both the M/G/1 and the GI/M/1 paradigms. Even though the GI/M/1 paradigm allows us to go down any number of levels, for practical reasons, we often require that we can go down only D levels (see, e.g. [12]). In this case, we can convert the process into a QBD process by reblocking, that is, we can combine D levels into a single level. Hence, the analysis of the QBD process covers most cases of interest.

The analysis of matrix-analytic methods was initiated by Wallace [21], but its main proponent was Neuts, who coined the words M/G/1 and GI/M/1 paradigm. For details of Neuts' work on the GI/M/1 paradigm, see [16] and, for the M/G/1 paradigm, see [17]. For a discussion of the QBD process, see [13]. The first time censoring was applied to matrix-analytic methods was

https://doi.org/10.1239/aap/1231340168 Published online by Cambridge University Press

in a paper by Grassmann and Heyman [6]. In fact, these authors extended the paradigms to what they called the GI/G/1 paradigm, a paradigm that allows several steps up or down. We note that most results derived for the M/G/1 and the GI/M/1 paradigms can be derived from the results of [6]. In particular, a result given by Latouche [12, Algorithm U, p. 238] follows directly by specializing equations (26a) and (26b) of [6] to the GI/M/1 paradigm. Only the notation is different. Latouche did mention the result of Kemeny *et al.* [11], but he did not use these results to derive algorithm U.

Note that in [6] block elimination (or will we say block censoring) was used. The state-by-state elimination was only described later in [7]. Of course, the end results of block elimination and state-by-state elimination are the same, but the latter allows us to exploit special structures, such as tridiagonal matrices, leading to faster algorithms.

Here we restrict our attention to the QBD process. By definition, this process has a transition matrix of the following form:

$$A = \begin{bmatrix} A_{00} & A_{01} & 0 & 0 & \dots & \dots & 0 \\ A_{10} & A_1 & A_0 & 0 & \dots & \dots & 0 \\ 0 & A_2 & A_1 & A_0 & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \end{bmatrix}.$$

For simplicity, we assume that $A_0 + A_1 + A_2$ is a nondecomposable stochastic matrix. In many cases, $M_0 = M$, with $A_{01} = A_0$ and $A_{10} = A_2$.

To decide whether the system is recurrent, we use the potential drift, which, according to [13], can be obtained as follows. First we find a row vector $\hat{\pi}$ by solving the following equations:

$$\hat{\pi} = \hat{\pi}(A_0 + A_1 + A_2), \qquad \hat{\pi}e = 1.$$

Here, e is a column vector with all entries equal to 1. The potential drift is now given by

$$\text{potential drift toward higher levels } = \hat{\pi}A_0e - \hat{\pi}A_2e.$$

It is known that, if the potential drift is negative, the QBD process is positive recurrent, if it is 0, it is null recurrent, and if it is positive, it is transient. This is what we would expect from the gas analogy introduced earlier.

Let $A^{(n)}$ be the matrix A with all levels above n censored. As we apply (1), we find that, for levels 1 to $n-1$, entries of A and $A^{(n)}$ coincide. In fact, the lower-right corner of $A^{(n)}$ becomes

$$\begin{matrix} 0 & \dots & 0 & A_2 & A_1 & A_0 \\ 0 & \dots & 0 & 0 & A_2 & Y \end{matrix}.$$

After eliminating level n, this yields

$$Y = A_1 + A_0(I - Y)^{-1}A_2. \tag{8}$$

Theorem 1 implies that if the Markov chain is recurrent then the value of Y can be obtained with arbitrary precision by cutting the Markov chain at a level N, and using censoring. Theorems 2 and 3 imply that Y is no longer unique when the Markov chain has a potential drift toward ∞, that is, there are two matrices Y satisfying (8), one corresponding to the stochastic solution and the other corresponding to the substochastic solution.

https://doi.org/10.1239/aap/1231340168 Published online by Cambridge University Press

To find the (unique) matrix Y when the potential drift is toward level 0, we can start with a value Y_0, where Y_0 can be chosen arbitrarily, and recursively determine (see, e.g. [8])

$$Y_{n+1} = A_1 + A_0(I - Y_n)^{-1}A_2. \tag{9}$$

We continue applying this recursion until some prescribed stopping criterion is satisfied. At this point, we can obtain $A^{(2)}$, the matrix involving only level 0 and 1, as

$$A^{(2)} = \begin{bmatrix} A_{00} & A_{01} \\ A_{10} & Y \end{bmatrix}.$$

Censoring level 1 as well leads to the transition matrix $A_{00} + A_{01}(I - Y)^{-1}A_{10}$. The $v_i = [v_{i,1}, v_{i,2}, \dots, v_{i,M}]$, $i > 0$, can now be found as follows. We solve

$$v_0 = v_0(A_{00} + A_{01}(I - Y)^{-1}A_{10})$$

for v_0, using $v_{0,1} = 1$ also. To find the v_i, $i > 0$, let $R = A_0(I - Y)^{-1}$. We then have

$$\begin{aligned} v_1 &= v_0 A_{01}(I - Y)^{-1}, \\ v_i &= v_{i-1}R, \qquad i > 1. \end{aligned} \tag{10}$$

Note that if (9) requires N iterations to find Y with a given precision then the finite matrix with $N + 1$ levels, and the same value of Y_0 in its lower-right corner, will lead to exactly the same approximate value for Y once level 1 is reached. However, the back substitution step is different because $v_i = v_{i-1}A_0(I - Y_{N-i})^{-1}$ is used rather than $v_i = v_{i-1}A_0(I - Y)^{-1} = v_{i-1}R$, as in (10).

We now compare our approach with the classical treatment of the QBD process. By specializing the GI/M/1 paradigm to the QBD process, we find that, according to [16],

$$R = A_0 + RA_1 + R^2A_2. \tag{11}$$

In fact, the matrix R in (11) is identical to the matrix R in (10). This follows from $R = A_0(I - Y)^{-1}$ and (8), which implies that $Y = A_1 + RA_2$. Hence,

$$A_0 = R(I - Y) = R(I - A_1 - RA_2),$$

and this is equivalent to (11).

In the M/G/1 paradigm, we use the matrix G. In the context of a QBD process, this matrix is defined as

$$G = A_2 + A_1G + A_0G^2.$$

It turns out that $G = (I - Y)^{-1}A_2$. The proof can be conducted along similar lines as the corresponding proof for R.

Latouche and Taylor [14] obtained G as follows. They started with some initial value G_0 and recursively calculated values $G_1, G_2, \dots$ using the following relation:

$$G_{n+1} = (I - A_1 - A_0G_n)^{-1}A_2. \tag{12}$$

This recursion is essentially identical to (9). To see this, note that $I - A_1 - A_0G_n$ corresponds to $I - Y_n$ and G_n corresponds to $(I - Y_n)^{-1}A_2$. With these conventions, (9) and (12) become identical. The advantage of using (9) is that the connection to the normal elimination procedure

https://doi.org/10.1239/aap/1231340168 Published online by Cambridge University Press

becomes clear. Indeed, to find Y_{n+1} from Y_n, we can eliminate one state after the other by using (4) until the level in question is eliminated. This is particularly advantageous when A_0 and A_2 are triangular, which occurs frequently when the matrix in question is banded. Of course, since the two methods are identical, this can also be used in (12), but this is not obvious.

If the QBD process is transient, that is, if there is a potential drift away from the origin, Theorem 3 indicates that there are two solutions for Y, one making $A^{(n)}$ stochastic and the other one making $A^{(n)}$ substochastic. The effect of this is that there are two solutions for G in cases where the drift is away from the origin. The existence of two solutions has been observed by Neuts [17], and discussed by Gail *et al.* [1], He and Neuts [9], and Latouche and Taylor [14]. To show how this relates to our approach, note that G_n is a stochastic matrix if and only if $A^{(n)}$ is stochastic, or, equivalently, if $(A_2 + Y_n)e = e$. To see this, note that $G_n = (I - Y_n)^{-1}A_2$ implies that $A_2 = (I - Y_n)G_n$ and

$$0 = (A_2 + Y_n - I)e = ((I - Y_n)G_n + (Y_n - I))e.$$

Premultiplying the right-hand side by $(I - Y_n)^{-1}$ leads to $(G_n - I)e = 0$, which proves that G is stochastic if and only if $(A_2 + Y)e = e$, as claimed. Hence, the stochastic solution makes G stochastic and the substochastic solution makes G substochastic.

If we are interested in finite but long queues, we need the version of Y that makes $A^{(n)}$ stochastic. Latouche and Taylor [14] showed that the stochastic solution is numerically unstable if the potential drift is away from the origin. This, of course, matches exactly with our discussion following Theorem 3. In fact, in the case of the QBD process, this can be proved mathematically because, according to Gail *et al.* [2], there are exactly two relevant matrices G, and, hence, one of them must be unstable. In fact, when using the GTH method, we are always sure to obtain a stochastic solution, and the algorithm becomes stable, even if the process has a potential drift away from level 0. Hence, the stochastic version of $A^{(n)}$ can always be found reliably.

5. Eigenvalues in QBD processes

The v_i corresponding to the stochastic solution behave quite differently from the ones in the substochastic solution. In fact, the v_i tend to increase with i in the stochastic solution, while they tend to constant values in substochastic solutions. To show this, we use eigenvalues. Specifically, we define the matrix polynomial $A(\gamma) = A_0 + \gamma(A_1 - I) + \gamma^2 A_2$, and we consider the vectors s^i and the scalars γ_i satisfying $0 = s^i A(\gamma_i)$. This problem is really a generalized eigenvalue problem, with the eigenpairs (s^i, γ_i). Generally, there exist $2M$ eigenvalues, counting multiplicities (see, e.g. [4]). Also, as is standard in matrix polynomials, there may be eigenvalues that are infinite (see, e.g. [4] or [22]).

Since $A(1) = A_0 + A_1 - I + A_2$ and $\hat{\pi}A(1) = 0$, $\gamma = 1$ is clearly an eigenvalue. Unless $A_0 + A_1 + A_2$ is decomposable, a case we want to exclude, the geometric multiplicity of this eigenvalue is 1. We now prove that the algebraic multiplicity of $\gamma = 1$ is greater than 1 if and only if $\hat{\pi}(A_2 - A_0) = 0$. The condition for an algebraic multiplicity greater than 1 is that, for $\gamma = 1$, both $\hat{\pi}A(\gamma) = 0$ and $\hat{\pi}A'(\gamma) = 0$, where $A'(\gamma)$ is the derivative of $A(\gamma)$ with respect to γ. Since, for $\gamma \neq 0$, $\hat{\pi}A(\gamma)$ has the same 0s as $\hat{\pi}A(\gamma)/\gamma$, we take the derivative of $\hat{\pi}A(\gamma)/\gamma$, and the result follows easily. Since $\hat{\pi}(A_0 - A_2)$ is the drift away from the origin, we have proved that γ has an algebraic multiplicity greater than 1 if and only if the process is null recurrent.

https://doi.org/10.1239/aap/1231340168 Published online by Cambridge University Press

The location of the other eigenvalues can be conveniently obtained from the following equation (see [15]):

$$A(\gamma) = (R - I\gamma)(Y - I)(G\gamma - I). \tag{13}$$

To prove (13), we expand its right-hand side to yield

$$A(\gamma) = -R(Y - I) + \gamma(Y - I - R(Y - I)G) - \gamma^2(Y - I)G,$$

and it is now easily verified that

$$\begin{gathered} -R(Y - I) = A_0, \\ Y - I - R(Y - I)G = Y - I - RA_2 = A_1 - I, \\ -(Y - I)G = A_2. \end{gathered}$$

Equation (13) implies that γ is an eigenvalue of $A(\gamma)$ if it is either an eigenvalue of R or the reciprocal of an eigenvalue of G. If G has eigenvalues of 0, the corresponding eigenvalues of $A(\gamma)$ are infinite. It is also clear that the factors $R - I\gamma$ and $G\gamma - I$ must each have M eigenvalues. (See also [3] for related results.) If the potential drift is toward 0 then the eigenvalues of R are inside the unit circle, and G is stochastic, which implies that the eigenvalues of the factor $G\gamma - I$ are outside or on the unit circle. As pointed out by Gail *et al.* [1], there are cases where G is periodic, which means that G can have several eigenvalues on the unit circle, but these cases rarely arise in practical situations. If the potential drift is away from 0 then it is known that in the substochastic solution R has a spectral radius of 1 and G is substochastic. Hence, all eigenvalues of $G\gamma - I$ are outside the unit circle. In the stochastic solution of nonrecurrent QBD processes, G is stochastic, that is, $G\gamma - I$ has to have an eigenvalue of 1. To make G stochastic, the eigenvalue $\gamma = 1$ which belonged to R in the substochastic solution now goes to G, and in exchange, G has to give up one of its eigenvalues, with the effect that R now obtains an eigenvalue greater than 1. Hence, the matrix R corresponding to the stochastic solution is no longer the minimal solution of (11).

We note that, once the eigenvalues and eigenvectors of a matrix are known, the matrix can be reconstructed provided we have the full set of eigenvectors (see, e.g. [10, pp. 46–47]). Hence, to find R, we identify all eigenvalues γ_i belonging to R, together with the corresponding eigenvectors s^i, and we have

$$R = S^{-1}\Lambda S, \tag{14}$$

where $\Lambda = \operatorname{diag}(\gamma_i)$ and S is the matrix with row i equal to s^i. In the case where the set of eigenvectors is less than M, Jordan forms must be used, but we will not discuss this further. Because of (14), v_n can be obtained as

$$v_n = v_1\Lambda^{n-1}S.$$

If γ_1 is the largest eigenvalue of R, it dominates the solution, and if s^1 is the corresponding eigenvector, we have

$$v_n \approx cs^1\gamma_1^n.$$

Here, c is a factor of proportionality which is immaterial to the following discussion. If the potential drift is toward level 0, the largest eigenvalue of R is less than 1 and v_n decreases eventually. If, on the other hand, the potential drift is away from 0, and the substochastic solution is used, the largest eigenvalue of R is $\gamma_1 = 1$, and the corresponding eigenvector is $s^1\hat{\pi}$. Hence, the v_n will reach constant values proportional to $\hat{\pi}$. This is what we would expect

https://doi.org/10.1239/aap/1231340168 Published online by Cambridge University Press

because, as the sample function drifts to ∞, it visits every level approximately equally often. If, on the other hand, the stochastic solution is used then the v_n increase for sufficiently large n. This matches with our earlier comparison involving a gas lighter than air in a closed container: this gas will accumulate at the top of the container.

If the potential drift is 0 then a simple limit argument demonstrates that both R and G have an eigenvalue of 1. In this case, there is only the stochastic solution, but at the same time, the v_n asymptotically approach the value of $\hat{\pi}$, multiplied by some constant.

For any stochastic solution, irrespective if the potential drift is to 0 or to ∞, we have to require that the largest eigenvalue of R must lead to an actual drift of 0 in equilibrium. In other words, if γ_1 is the largest eigenvalue of R, and if s^1 is the corresponding eigenvector, we must have

$$s^1 A_0 e = s^1 \gamma_1 A_2 e.$$

We tested this for a limited number of processes and found that it holds in all the cases considered, including stochastic solutions of transient Markov chains. In fact, in our examples, we found that all eigenpairs of R corresponding to a stochastic solution satisfy this equality.

As pointed out in [3], there are many solutions of (11). In fact, we can select any M out of the $2M$ eigenpairs and form R. However, according to [1], at most two of these solutions have a stochastic meaning in our context.

6. Conclusions and extensions

The main result of this paper is that in cases of a potential drift away from the origin we have two different censored Markov chains and, as a consequence, two solutions of v_n. There is a stochastic solution, which must be used when dealing with large but finite queues, and a substochastic solution. It was shown that in QBD processes with a potential drift to ∞, and for sufficiently large i, the v_i essentially remain constant in the substochastic solution, but they increase in the stochastic solution. This result also applies when considering a gas lighter than air in a closed container. This gas would also collect at the top of the container.

Our analysis was carried out for discrete-time Markov chains, but the results also apply for continuous-time Markov chains. In this case, the variables v_i indicate the expected times in state i as a multiple of the expected time in state 1. Also, when going to the continuous case, some of the formulae change slightly, while others remain the same. In particular, (1) remains unchanged, but with Q being equal to an integral of a matrix exponential. Equation (5) remains unchanged. When applying the equivalent of (8), one must replace A_1 by A_1^*, a matrix with negative diagonal elements, and a similar change affects Y. If we write Y^* for this new matrix, we obtain

$$Y^* = A_1^* + A_0(-Y^*)^{-1}A_2.$$

Note that this equation can also be used in the discrete case, provided that we set $A_1^* = A_1 - I$ and $Y^* = Y - I$. Hence, all arguments, after obvious modifications, go through for continuous-time Markov chains, as we would expect because of uniformization.

Acknowledgements

This research was supported by the Natural Sciences and Engineering Research Council of Canada. We would also like to express our sincere thanks to the anonymous referee for his/her careful report, which greatly helped to improve this paper.

References

[1] Gail, H. R., Hantler, S. L. and Taylor, B. A. (1994). Solution of the basic matrix equation for M/G/1 and G/M/1 type Markov chains. *Stoch. Models* **10,** 1–43.

[2] Gail, H.R., Hantler, S. L. and Taylor, B. A. (1996). Spectral analysis of M/G/1 and G/M/1 type Markov chains. *Adv. Appl. Prob.* **28,** 114–165.

[3] Gail, H. R., Hantler, S. L. and Taylor, B. A. (2000). Use of characteristic roots for solving infinite state Markov chains. In *Computational Probability*, ed. W. K. Grassmann, Kluwer, Boston, MA, pp. 205–254.

[4] Gohberg, I.. Lancaster, P. and Rodman, L. (1982). *Matrix Polynomials*. Academic Press, New York.

[5] Grassmann, W. K. (1993). Rounding errors in certain algorithms involving Markov chains. *ACM Trans. Math. Software* **19,** 496–508.

[6] Grassmann, W. K. and Heyman, D. P. (1990). Equilibrium distribution of block-structured Markov chains with repeating rows. *J. Appl. Prob.* **27,** 557–576.

[7] Grassmann, W. K. and Heyman, D. P. (1993). Computation of steady-state probabilities for infinite-state Markov chains with repeating rows. *ORSA J. Comput.* **5,** 292–303.

[8] Grassmann, W. K. and Stanford, D. A. (2000). Matrix analytic methods. In *Computational Probability*, ed. W. K. Grassmann, Kluwer, Boston, MA, pp. 153–202.

[9] He, Qi-Ming and Neuts, M. F. (2001). On the convergence and limits of certain matrix sequences arising in quasi-birth-and-death Markov chains. *J. Appl. Prob.* **38,** 519–541.

[10] Horn, R. A. and Johnson, C. R. (1985). *Matrix Analysis*. Cambridge University Press.

[11] Kemeny, J. G., Snell, J. L. and Knapp, A. W. (1966). *Denumerable Markov Chains*. Van Nostrand, Princeton, NJ.

[12] Latouche, G. (1992). Algorithms for infinite Markov chains with repeating columns. In *Linear Algebra, Markov Chains, and Queueing Models* (IMA Vols Math. Appl. **48**), Springer, Heidelberg, pp. 231–265.

[13] Latouche, G. and Ramaswami, V. (1999). *Introduction to Matrix Analytic Methods in Stochastic Modelling*. SIAM, Philadelphia, PA.

[14] Latouche, G. and Taylor, P. (2002). Truncation and augmentation of level-independent QBD processes. *Stoch. Process. Appl.* **99,** 53–80.

[15] Naoumov, V. (1996). Matrix-multiplicative approach to quasi-birth-and-death processes analysis. In *Matrix-Analytic Methods in Stoch. Models*, Marcel Dekker, New York, pp. 87–106.

[16] Neuts, M. F. (1981). *Matrix-Geometric Solutions in Stochastic Models*. Johns Hopkins University Press, Baltimore, MD.

[17] Neuts, M. F. (1989). *Structured Stochastic Matrices of* M/G/1 *Type and Their Applications*. Marcel Dekker, New York.

[18] O'Cinneide, C. A. (1993). Entrywise perturbabtion theory and error analysis for Markov chains. *Numerische Mathematik* **65,** 109–120.

[19] O'Cinneide, C. A. (1996). Relative error bounds of the LU decomposition via the GTH algorithm. *Numerische Mathematik* **73,** 507–519.

[20] Stewart, W. (1994). *An Introduction to the Numerical Solution of Markov Chains*. Princeton University Press.

[21] Wallace, V. (1969). The solution of quasi birth and death processes arising from multiple access computer systems. Doctoral thesis, University of Michigan.

[22] Wilkinson, J. S. (1965). *The Algebraic Eigenvalue Problem*. Clarendon Press, Oxford.

[23] Zhao, Y. Q. (2000). Censoring technique in studying block-structured Markov chains. In *Advances in Agorithmic Methods for Stochastic Models*, eds G. Latouch and P. G. Taylor, Notable Publications, Neshanic Station, NJ, pp. 417–433.

[24] Zhao, Y. Q., Braun, W. J. and Li, W. (1999). Northwest corner and banded matrix approximation to a countable Markov chain. *Naval Res. Logistics* **46,** 187–197.

https://doi.org/10.1239/aap/1231340168 Published online by Cambridge University Press