Journal of the Operations Research
Society of Japan
Vol. 40, No. 3, September 1997

NECESSARY AND SUFFICIENT CONDITIONS FOR GLOBAL GEOMETRIC CONVERGENCE OF BLOCK GAUSS-SEIDEL ITERATION ALGORITHM APPLIED TO MARKOV CHAINS

Ushio Sumita
International University of Japan

Nobuko Igaki
Tezukayama University

(Received March 22, 1995; Revised April 7, 1997)

Abstract Convergence properties of the block Gauss-Seidel algorithm applied to ergodic Markov chains are discussed in this paper. This algorithm is one of the most prevalent methods for computing ergodic probability vectors of large-scale Markov chains. We will provide necessary and sufficient conditions for global geometric convergence of this algorithm. To apply this algorithm, the state space of a Markov process is decomposed into mutually exclusive and exhaustive lumps. The convergence properties depend on this lumping. It is also shown that there exists at least one set of lumps, for any ergodic stochastic matrix, which assures geometric convergence of the algorithm.

1. Introduction

Markov chain modeling plays an important role in the performance analysis of computer / communication and manufacturing systems. Accordingly, the development of efficient algorithms for computing the ergodic probability vector for large-scale Markov chains constitutes an essential core of computational probability. An important class of computational methods for large-scale Markov chains consists of iterative algorithms. Instead of computing the exact ergodic probability vector directly, these algorithms generate a sequence of vectors, which upon convergence produces the desired result. The prevalent algorithmic methods in this class include the block Gauss-Seidel iteration algorithm, see e.g.,[1], the iterative aggregation-disaggregation algorithm developed by Takahashi [9] and Takahashi and Takami [10], and the replacement process approach developed by Sumita and Rieders [7, 8].

One difficulty arising from these iterative algorithms is the problem of convergence. To the authors' best knowledge, little is known about global convergence theorem and speed of convergence for such iterative algorithms except the direct successive substitution method. Rare exceptions are a local convergence theorem due to Cao and Stewart [2] for Takahashi's aggregation-disaggregation algorithm and the modified aggregation-disaggregation algorithm by Schweitzer [6] where the geometric convergence is enforced by utilizing the direct successive substitution method whenever appropriate. More recently, Mitra and Tsoucas [5] have examined convergence properties of point Gauss-Seidel relaxations.

The purpose of this paper is to provide necessary and sufficient conditions for global geometric convergence of the block Gauss-Seidel algorithm applied to ergodic stochastic matrices. We show that, under certain conditions, this algorithm converges geometrically starting with an arbitrary initial positive vector. The geometric factor governing the speed of convergence is given by the magnitude of the second maximum eigenvalue of a stochastic matrix constructed from the original stochastic matrix. As to show the other side of the coin, we also provide a necessary and sufficient condition under which the iterative algorithm does not converge for

This paper has been partially supported by the Nippon Telegram and Telephone Research Fund.

any initial probability vector.

In the block Gauss-Seidel algorithm, the state space of Markov process is decomposed into mutually exclusive and exhaustive lumps. The convergence properties depend on this decomposition. We also show that there exists at least one decomposition for which the block Gauss-Seidel algorithm always converges geometrically with appropriate rearrangement of states. These results naturally extend the results of Mitra and Tsoucas [5] to the block form. Since many other iterative algorithms are related to the block Gauss-Seidel algorithm directly or indirectly, the main result of this paper may provide an important stepping stone for understanding the convergence property of other iterative algorithms.

2. The Block Gauss-Seidel Algorithm for Large-Scale Markov Chains

Let $\{J(k) : k \geq 0\}$ be a discrete time Markov chain on $V = \{1, 2, \ldots, N\}$ governed by a stochastic matrix $\boldsymbol{A}$. Throughout the paper we assume that the Markov chain is ergodic with the ergodic probability vector $\boldsymbol{\pi}^{\mathrm{T}}$ so that

$$\boldsymbol{\pi}^{\mathrm{T}} = \boldsymbol{\pi}^{\mathrm{T}}\boldsymbol{A} \ . \tag{2.1}$$

Let L be a decomposition of the state space V into mutually exclusive and exhaustive lumps defined by

$$L = \{L(1), L(2), \cdots, L(M)\} \ , \quad V = \bigcup_{n=1}^{M} L(n) \ , \quad \text{and} \quad L(m) \bigcap L(n) = \emptyset \text{ for } m \neq n \ . \tag{2.2}$$

Throughout the paper, we assume that, whenever a decomposition L is given, the states are rearranged in nondecreasing order of lumps, i.e.,

$$\{1, 2, \cdots, N\} \to \{\underbrace{i_{L(1)1}, \cdots, i_{L(1)\alpha(1)}}_{L(1)}, \underbrace{i_{L(2)1}, \cdots, i_{L(2)\alpha(2)}}_{L(2)}, \cdots, \underbrace{i_{L(M)1}, \cdots, i_{L(M)\alpha(M)}}_{L(M)}\} \tag{2.3}$$

where $\alpha(m)$ denotes the number of states in $L(m)$, $m = 1, 2, \cdots, M$, and $i_{L(m)j}$ denote the j-th state in the lump $L(m), j = 1, \cdots, \alpha(m)$. Inside each lump, one can choose any order of states.

For $m, n \in \{1, 2, \cdots, M\}$, submatrices of $\boldsymbol{A}$ are written as $\boldsymbol{A}_{L(m)L(n)} = [a_{ij}]_{i \in L(m), j \in L(n)}$, and subvectors of $\boldsymbol{\pi}^{\mathrm{T}}$ are written as $\boldsymbol{\pi}^{\mathrm{T}}_{L(m)} = [\pi_j]_{j \in L(m)}$, etc. For convenience, we also define

$$H = \{1, 2, \cdots, M\}; \ G(m-) = \{1, 2, \cdots, m-1\}; \text{ and } G(m+) = \{m+1, \cdots, M\} \tag{2.4}$$

where $G(1-) = G(M+) = \emptyset$. From (2.1), one easily sees the block-wise balance equations given by

$$\begin{aligned} \boldsymbol{\pi}^{\mathrm{T}}_{L(m)} \ = \ & \boldsymbol{\pi}^{\mathrm{T}}_{L(m)}\boldsymbol{A}_{L(m)L(m)} + \sum_{r \in G(m-)} \boldsymbol{\pi}^{\mathrm{T}}_{L(r)}\boldsymbol{A}_{L(r)L(m)} \\ & + \sum_{r \in G(m+)} \boldsymbol{\pi}^{\mathrm{T}}_{L(r)}\boldsymbol{A}_{L(r)L(m)}, \quad m \in H, \end{aligned} \tag{2.5}$$

where the summation is ignored whenever the corresponding index set is empty.

It is well known that for any ergodic stochastic matrix $\boldsymbol{A}$, in the direct successive substitution method

$$\boldsymbol{p}^{\mathrm{T}}(k) = \boldsymbol{p}^{\mathrm{T}}(k-1)\boldsymbol{A} \ , \quad k = 1, 2, \cdots, \tag{2.6}$$

starting with any probability vector $\boldsymbol{p}^{\mathrm{T}}(0) > \mathbf{O}^{\mathrm{T}}$, $\boldsymbol{p}^{\mathrm{T}}(k)$ always converges geometrically to $\boldsymbol{\pi}^{\mathrm{T}}$ as $k \to \infty$. In this context, corresponding to (2.5), (2.6) is rewritten as

$$(2.7)\quad \begin{aligned} \boldsymbol{p}^{\mathrm{T}}_{L(m)}(k) &= \boldsymbol{p}^{\mathrm{T}}_{L(m)}(k-1)\boldsymbol{A}_{L(m)L(m)} + \sum_{r\in G(m-)} \boldsymbol{p}^{\mathrm{T}}_{L(r)}(k-1)\boldsymbol{A}_{L(r)L(m)} \\ &\quad + \sum_{r\in G(m+)} \boldsymbol{p}^{\mathrm{T}}_{L(r)}(k-1)\boldsymbol{A}_{L(r)L(m)}, \quad m \in H, \quad k = 1,2,\cdots, \end{aligned}$$

where $\boldsymbol{p}^{\mathrm{T}}_{L(m)}(k) = [p_j(k)]_{j\in L(m)}$ is a subvector of the k-th iteration's vector $\boldsymbol{p}^{\mathrm{T}}(k)$. The block Gauss-Seidel iteration algorithm applied to Equation (2.1) exploits block-based computations by utilizing updated information to the point of lump m computation as much as possible. In place of (2.7), one considers

$$(2.8)\quad \begin{aligned} \boldsymbol{\pi}^{\mathrm{T}}_{L(m)}(k) &= \boldsymbol{\pi}^{\mathrm{T}}_{L(m)}(k)\boldsymbol{A}_{L(m)L(m)} + \sum_{r\in G(m-)} \boldsymbol{\pi}^{\mathrm{T}}_{L(r)}(k)\boldsymbol{A}_{L(r)L(m)} \\ &\quad + \sum_{r\in G(m+)} \boldsymbol{\pi}^{\mathrm{T}}_{L(r)}(k-1)\boldsymbol{A}_{L(r)L(m)}, \quad m \in H, \quad k = 1,2,\cdots, \end{aligned}$$

where this equation is solved for $\boldsymbol{\pi}^{\mathrm{T}}_{L(m)}(k)$. For computing $\boldsymbol{\pi}^{\mathrm{T}}_{L(m)}(k)$ in the k−th iteration, $\boldsymbol{\pi}^{\mathrm{T}}_{L(r)}(k)$ for $r \in G(m-)$, already computed in the k−th iteration, are used, while $\boldsymbol{\pi}^{\mathrm{T}}_{L(r)}(k-1)$ for $r \in G(m+)$ are employed from the previous iteration. From the uniqueness of the solution $\boldsymbol{\pi}^{\mathrm{T}}$ satisfying (2.1), it can be readily seen that the iterative algorithm based on (2.8) produces $\boldsymbol{\pi}^{\mathrm{T}}$ if it converges. Hence, the following iterative algorithm for computing the ergodic probability vector of $\boldsymbol{A}$ can be constructed. A column vector having all components equal to 1 is denoted by $\mathbf{1}$.

Algorithm 1 (The Block Gauss-Seidel Algorithm)

Input : an ergodic stochastic matrix $\boldsymbol{A}$, an $N-$dimensional positive vector $\boldsymbol{\pi}^{\mathrm{T}}(0) > \mathbf{O}^{\mathrm{T}}$, a lumping L as given in (2.2) and an accuracy level $\varepsilon > 0$.

Output : the ergodic probability vector $\boldsymbol{\pi}^{\mathrm{T}}$ of $\boldsymbol{A}$, if it converges.

[1] Set $m = 1$ and $k = 1$.

[2] Loop : Solve the linear equation (2.8) for $\boldsymbol{\pi}^{\mathrm{T}}_{L(m)}(k)$.

[3] If $m < M$, set $m = m + 1$ and go to Loop.

[4] If $\|\boldsymbol{\pi}^{\mathrm{T}}(k) - \boldsymbol{\pi}^{\mathrm{T}}(k-1)\| > \varepsilon$, set $k = k + 1$, $m = 1$, and go to Loop.

[5] Set $\boldsymbol{\pi}^{\mathrm{T}} = \boldsymbol{\pi}^{\mathrm{T}}(k)/\{\boldsymbol{\pi}^{\mathrm{T}}(k)\mathbf{1}\}$ and stop.

It should be noted that the block Gauss-Seidel algorithm is scale free in that $\boldsymbol{\pi}^{\mathrm{T}}_{L(m)}(k)$ can be scaled up or down by an arbitrary factor at an arbitrary stage of the algorithm, provided that the same factor is employed for all $\boldsymbol{\pi}^{\mathrm{T}}_{L(m)}(k)$, $m \in H$. Furthermore, the algorithm manipulates only nonnegative numbers if the successive substitution is used for solving (2.8). There are no divisions involved in the algorithm except the final normalization.

3. A Necessary and Sufficient Condition for Global Geometric Convergence of the Block Gauss-Seidel Algorithm

In this section, we show a necessary and sufficient condition under which Algorithm 1 converges geometrically starting with an arbitrary N-dimensional vector $\boldsymbol{\pi}^{\mathrm{T}}(0) > \mathbf{0}^{\mathrm{T}}$. For convenience, the block lower diagonal matrix $\boldsymbol{A}_L$, the block diagonal matrix $\boldsymbol{A}_D$ and the block upper diagonal matrix $\boldsymbol{A}_U$ are defined as

$$(3.1)\quad \boldsymbol{A}_L = [\delta_{\{m>n\}}\boldsymbol{A}_{L(m)L(n)}];\ \boldsymbol{A}_D = [\delta_{\{m=n\}}\boldsymbol{A}_{L(m)L(n)}];\ \boldsymbol{A}_U = [\delta_{\{m<n\}}\boldsymbol{A}_{L(m)L(n)}],$$

where $\delta_{\{\mathrm{P}\}} = 1$ if the statement P holds and $\delta_{\{\mathrm{P}\}} = 0$ otherwise. We note that

$$(3.2)\quad \boldsymbol{A} = \boldsymbol{A}_L + \boldsymbol{A}_D + \boldsymbol{A}_U \ .$$

Theorem 3.1 *Let $\{\boldsymbol{\pi}^{\mathrm{T}}(k) : k = 1, 2, \ldots\}$ be a sequence of vectors generated by Algorithm 1 starting with $\boldsymbol{\pi}^{\mathrm{T}}(0) > \mathbf{0}^{\mathrm{T}}$. Then one has*

$$(3.3)\quad \boldsymbol{\pi}^{\mathrm{T}}(k) = \boldsymbol{\pi}^{\mathrm{T}}(k-1)\boldsymbol{B} \ , \quad k = 1, 2, \cdots,$$

where $\boldsymbol{B}$ is given by

$$(3.4)\quad \boldsymbol{B} = \boldsymbol{A}_L[\boldsymbol{I} - (\boldsymbol{A}_D + \boldsymbol{A}_U)]^{-1} \ .$$

Proof : We first note that Equation (2.8) for $m \in H$ can be rewritten as

$$(3.5)\quad \boldsymbol{\pi}^{\mathrm{T}}(k) = \boldsymbol{\pi}^{\mathrm{T}}(k)\boldsymbol{A}_D + \boldsymbol{\pi}^{\mathrm{T}}(k)\boldsymbol{A}_U + \boldsymbol{\pi}^{\mathrm{T}}(k-1)\boldsymbol{A}_L \ .$$

Since $\boldsymbol{A}$ is ergodic, $\boldsymbol{A}_L$ and $\boldsymbol{A}_U$ are nonzero matrices. Consequently, $\boldsymbol{A}_D + \boldsymbol{A}_U$ is strictly substochastic and $[\boldsymbol{I} - (\boldsymbol{A}_D + \boldsymbol{A}_U)]$ is nonsingular. The theorem now follows from (3.5) by solving for $\boldsymbol{\pi}^{\mathrm{T}}(k)$. ■

We note that $\boldsymbol{B}$ is a nonnegative square matrix. From (2.1) and (3.2) , it can be readily seen that

$$(3.6)\quad \boldsymbol{\pi}^{\mathrm{T}}\boldsymbol{A}_L = \boldsymbol{\pi}^{\mathrm{T}}[\boldsymbol{I} - (\boldsymbol{A}_D + \boldsymbol{A}_U)] \ ; \ \boldsymbol{A}_L\mathbf{1} = [\boldsymbol{I} - (\boldsymbol{A}_D + \boldsymbol{A}_U)]\mathbf{1} \ .$$

We are now in a position to prove the following theorem. The spectral radius of a square matrix $\boldsymbol{Y}$ is denoted by $\rho(\boldsymbol{Y})$.

Theorem 3.2 *Let $\boldsymbol{A}$ be any ergodic stochastic matrix having the ergodic probability vector $\boldsymbol{\pi}^{\mathrm{T}}$. Let $\boldsymbol{B}$ be as defined in* (3.4) *. One then has :*

$$(3.7)\quad \rho(\boldsymbol{B}) = 1 \ \text{ with } \ \boldsymbol{\pi}^{\mathrm{T}} = \boldsymbol{\pi}^{\mathrm{T}}\boldsymbol{B} \ \text{ and } \ \boldsymbol{B}\boldsymbol{u} = \boldsymbol{u} \ \text{ where } \ \boldsymbol{u} = \boldsymbol{A}_L\mathbf{1}/\{\boldsymbol{\pi}^{\mathrm{T}}\boldsymbol{A}_L\mathbf{1}\} \ ;$$

$$(3.8)\quad \boldsymbol{B} = \boldsymbol{u}\boldsymbol{\pi}^{\mathrm{T}} + \triangle_B \ ,$$

$$(3.9)\quad \boldsymbol{\pi}^{\mathrm{T}}\triangle_B = \mathbf{0}^{\mathrm{T}} \ ; \ \triangle_B\boldsymbol{u} = \mathbf{0} \ ; \ \boldsymbol{\pi}^{\mathrm{T}}\boldsymbol{u} = 1 \ , \ \text{and}$$

$$(3.10)\quad \rho(\triangle_B) \leq 1 \ .$$

Proof : We first note from (3.2) , (3.4) and (3.6) that

$$\boldsymbol{\pi}^{\mathrm{T}}\boldsymbol{B} = \boldsymbol{\pi}^{\mathrm{T}}\boldsymbol{A}_L[\boldsymbol{I} - (\boldsymbol{A}_D + \boldsymbol{A}_U)]^{-1} = \boldsymbol{\pi}^{\mathrm{T}}[\boldsymbol{I} - (\boldsymbol{A}_D + \boldsymbol{A}_U)][\boldsymbol{I} - (\boldsymbol{A}_D + \boldsymbol{A}_U)]^{-1} = \boldsymbol{\pi}^{\mathrm{T}} \ ,$$

and

$$\begin{aligned}(\boldsymbol{\pi}^{\mathrm{T}}\boldsymbol{A}_L\mathbf{1})\boldsymbol{B}\boldsymbol{u} &= \boldsymbol{A}_L[\boldsymbol{I} - (\boldsymbol{A}_D + \boldsymbol{A}_U)]^{-1}\boldsymbol{A}_L\mathbf{1} \\ &= \boldsymbol{A}_L[\boldsymbol{I} - (\boldsymbol{A}_D + \boldsymbol{A}_U)]^{-1}[\boldsymbol{I} - (\boldsymbol{A}_D + \boldsymbol{A}_U)]\mathbf{1} = \boldsymbol{A}_L\mathbf{1} = (\boldsymbol{\pi}^{\mathrm{T}}\boldsymbol{A}_L\mathbf{1})\boldsymbol{u} \ .\end{aligned}$$

Hence $\boldsymbol{\pi}^{\mathrm{T}}$ and $\boldsymbol{u}$ are the left- and the right-eigenvectors of $\boldsymbol{B}$ associated with eigenvalue one, satisfying $\boldsymbol{\pi}^{\mathrm{T}}\boldsymbol{u} = 1$.

Since $\boldsymbol{B}$ is a nonnegative square matrix, one sees from the generalized Perron-Frobenius theorem (see Theorem I* of Debreu and Herstein [3]) that $\boldsymbol{B}$ has the maximum eigenvalue $\lambda \geq 0$ associated with the left eigenvector $\boldsymbol{x}^{\mathrm{T}} \geq \mathbf{O}^{\mathrm{T}}$ and the right eigenvector $\boldsymbol{y} \geq \mathbf{O}$ satisfying $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{y} = 1$. From $\boldsymbol{\pi}^{\mathrm{T}} = \boldsymbol{\pi}^{\mathrm{T}}\boldsymbol{B}$ and $\boldsymbol{B}\boldsymbol{u} = \boldsymbol{u}$, one has $\lambda \geq 1$. Suppose $\lambda > 1$. From (3.4), one sees that

$$\boldsymbol{B}[\boldsymbol{I} - (\boldsymbol{A}_D + \boldsymbol{A}_U)] = \boldsymbol{A}_L \quad . \tag{3.11}$$

Since $\lambda\boldsymbol{x}^{\mathrm{T}} = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{B}$, Equation (3.11) then leads to $\lambda\boldsymbol{x}^{\mathrm{T}}[\boldsymbol{I} - (\boldsymbol{A}_D + \boldsymbol{A}_U)] = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}_L$, or equivalently,

$$\lambda\boldsymbol{x}^{\mathrm{T}}[\boldsymbol{I} - \hat{\boldsymbol{A}}(\lambda)] = \mathbf{O}^{\mathrm{T}} \quad , \tag{3.12}$$

where

$$\hat{\boldsymbol{A}}(\lambda) = \boldsymbol{A}_D + \boldsymbol{A}_U + \frac{1}{\lambda}\boldsymbol{A}_L \quad . \tag{3.13}$$

Since $\boldsymbol{A}_L$ is a nonzero matrix, $\hat{\boldsymbol{A}}(\lambda)$ is strictly substochastic for $\lambda > 1$. Hence $[\boldsymbol{I} - \hat{\boldsymbol{A}}(\lambda)]$ is nonsingular and Equation (3.12) has a unique solution $\boldsymbol{x}^{\mathrm{T}} = \mathbf{O}^{\mathrm{T}}$. This contradicts the fact that $\boldsymbol{x}^{\mathrm{T}}$ is the left eigenvector associated with λ and $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{y} = 1$. Hence $\lambda = \rho(\boldsymbol{B}) = 1$. From the generalized Perron-Frobenius theorem, $\boldsymbol{B}$ has a spectral decomposition as given in (3.8) satisfying (3.9) and (3.10). This completes the proof. ■

Theorem 3.1 implies that

$$\boldsymbol{\pi}^{\mathrm{T}}(k) = \boldsymbol{\pi}^{\mathrm{T}}(0)\boldsymbol{B}^k \quad , \qquad k = 1, 2, \cdots . \tag{3.14}$$

Hence, the next theorem immediately follows from (3.14) and Theorem 3.2.

Theorem 3.3 *Let $\boldsymbol{A}$ be any ergodic stochastic matrix having the ergodic probability vector $\boldsymbol{\pi}^{\mathrm{T}}$ and define a matrix $\boldsymbol{B}$ as in* (3.4). *Also let $\triangle_B$ be as in Theorem 3.2 . Then Algorithm 1 applied to $\boldsymbol{A}$ converges geometrically starting with any positive vector $\boldsymbol{\pi}^{\mathrm{T}}(0)$ if and only if $\rho(\triangle_B) < 1$. Furthermore, the geometric factor governing the speed of convergence is given by $\rho(\triangle_B)$.*

Proof : From (3.7) and (3.8), one sees that a dyadic matrix $\boldsymbol{u}\boldsymbol{\pi}^{\mathrm{T}}$ is idempotent and the two matrices $\boldsymbol{u}\boldsymbol{\pi}^{\mathrm{T}}$ and $\triangle_B$ are matrix orthogonal to each other. It follows from (3.8) that

$$\boldsymbol{B}^k = \boldsymbol{u}\boldsymbol{\pi}^{\mathrm{T}} + \triangle_B{}^k \quad , \qquad k = 1, 2, \cdots . \tag{3.15}$$

Substituting (3.15) into (3.14), one finds that $\boldsymbol{\pi}^{\mathrm{T}}(k) = \{\boldsymbol{\pi}^{\mathrm{T}}(0)\boldsymbol{u}\}\boldsymbol{\pi}^{\mathrm{T}} + \boldsymbol{\pi}^{\mathrm{T}}(0)\triangle_B{}^k$. Since $\rho(\triangle_B) < 1$, the second term converges geometrically to $\mathbf{O}^{\mathrm{T}}$ as k goes to infinity, where the geometric factor $\rho(\triangle_B)$ is the second maximum of the absolute eigenvalues of $\boldsymbol{B}$. The positive constant $\boldsymbol{\pi}^{\mathrm{T}}(0)\boldsymbol{u}$ is eliminated in [5] of Algorithm 1. ■

4. Existence of Lumping for Assuring Geometric Convergence of the Block Gauss-Seidel Algorithm

We have seen that the block Gauss-Seidel algorithm converges geometrically if and only if $\rho(\triangle_B) < 1$. However, $\boldsymbol{B}$ is clearly decomposable from (3.4) and the problem of seeing when $\rho(\triangle_B) < 1$ holds is nontrivial. This difficulty can be conquered by introducing a stochastic matrix $\boldsymbol{S}$ which inherits the convergence structure of $\boldsymbol{B}$.

Theorem 4.1 *Let $\boldsymbol{A}$ be any ergodic stochastic matrix having the ergodic probability vector $\boldsymbol{\pi}^{\mathrm{T}}$ and define a matrix $\boldsymbol{S}$ by*

$$\boldsymbol{S} = [\boldsymbol{I} - (\boldsymbol{A}_D + \boldsymbol{A}_U)]^{-1}\boldsymbol{A}_L \quad . \tag{4.1}$$

Then the following statements hold.

$\boldsymbol{S}$ is a stochastic matrix with $\boldsymbol{\mu}^{\mathrm{T}} = \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{S}$ where $\boldsymbol{\mu}^{\mathrm{T}} = \boldsymbol{\pi}^{\mathrm{T}}\boldsymbol{A}_L/\{\boldsymbol{\pi}^{\mathrm{T}}\boldsymbol{A}_L\mathbf{1}\}$; (4.2)

$$\boldsymbol{S} = \mathbf{1}\boldsymbol{\mu}^{\mathrm{T}} + \triangle_S \quad , \tag{4.3}$$

$$\boldsymbol{\mu}^{\mathrm{T}}\triangle_S = \mathbf{0}^{\mathrm{T}} \;\; ; \;\; \triangle_S\mathbf{1} = \mathbf{0} \;\; ; \;\; \boldsymbol{\mu}^{\mathrm{T}}\mathbf{1} = 1 \quad , \quad \text{and} \tag{4.4}$$

$$\rho(\triangle_S) = \rho(\triangle_B) \quad . \tag{4.5}$$

Proof : From the definition in (4.1) , the matrix $\boldsymbol{S}$ is clearly nonnegative. We see from (4.1) and (3.6) that

$$\boldsymbol{S}\mathbf{1} = [\boldsymbol{I} - (\boldsymbol{A}_D + \boldsymbol{A}_U)]^{-1}\boldsymbol{A}_L\mathbf{1} = [\boldsymbol{I} - (\boldsymbol{A}_D + \boldsymbol{A}_U)]^{-1}[\boldsymbol{I} - (\boldsymbol{A}_D + \boldsymbol{A}_U)]\mathbf{1} = \mathbf{1} \quad . \tag{4.6}$$

Hence $\boldsymbol{S}$ is stochastic. Furthermore,

$$\begin{aligned}(\boldsymbol{\pi}^{\mathrm{T}}\boldsymbol{A}_L\mathbf{1})\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{S} &= \boldsymbol{\pi}^{\mathrm{T}}\boldsymbol{A}_L[\boldsymbol{I} - (\boldsymbol{A}_D + \boldsymbol{A}_U)]^{-1}\boldsymbol{A}_L \\ &= \boldsymbol{\pi}^{\mathrm{T}}[\boldsymbol{I} - (\boldsymbol{A}_D + \boldsymbol{A}_U)][\boldsymbol{I} - (\boldsymbol{A}_D + \boldsymbol{A}_U)]^{-1}\boldsymbol{A}_L = (\boldsymbol{\pi}^{\mathrm{T}}\boldsymbol{A}_L\mathbf{1})\boldsymbol{\mu}^{\mathrm{T}} \quad ,\end{aligned}$$

proving the first statement. From the generalized Perron-Frobenius theorem, $\boldsymbol{S}$ has a spectral decomposition as given in (4.3) satisfying (4.4). As in (3.15), one sees that

$$\boldsymbol{S}^k = \mathbf{1}\boldsymbol{\mu}^{\mathrm{T}} + \triangle_S{}^k \quad , \quad k = 1, 2, \cdots . \tag{4.7}$$

From (3.4) and (4.1), one obtain

$$\boldsymbol{B}^k = \boldsymbol{A}_L\boldsymbol{S}^{k-1}[\boldsymbol{I} - (\boldsymbol{A}_D + \boldsymbol{A}_U)]^{-1} \quad , \tag{4.8}$$

so that $\boldsymbol{B}^k$ converges if and only if $\boldsymbol{S}^k$ converges.

Substituting (3.15) and (4.7) into (4.8),

$$\triangle_B{}^k = \boldsymbol{A}_L\triangle_S{}^{k-1}[\boldsymbol{I} - (\boldsymbol{A}_D + \boldsymbol{A}_U)]^{-1} \quad . \tag{4.9}$$

Using l_1 Norm, it follows from (4.9) that

$$\left\{\frac{\rho(\triangle_B)}{\rho(\triangle_S)}\right\}^k \leq \| \boldsymbol{A}_L \| \cdot \| [\boldsymbol{I} - (\boldsymbol{A}_D + \boldsymbol{A}_U)]^{-1} \| \, \rho(\triangle_S)^{-1} \quad , \tag{4.10}$$

for arbitrary k ; $k = 1, 2, \cdots$. One then has

$$\frac{\rho(\triangle_B)}{\rho(\triangle_S)} \leq 1 \quad . \tag{4.11}$$

On the other hand, starting with equation

$$\boldsymbol{S}^k = [\boldsymbol{I} - (\boldsymbol{A}_D + \boldsymbol{A}_U)]^{-1}\boldsymbol{B}^{k-1}\boldsymbol{A}_L \quad , \tag{4.12}$$

similar to (4.8), we have

$$\frac{\rho(\triangle_S)}{\rho(\triangle_B)} \leq 1 \quad . \tag{4.13}$$

Statement (4.5) then follows from (4.11) and (4.13), completing the proof. ■

Let $\{N(k) : k \geq 0\}$ be the discrete time Markov chain on V governed by $\boldsymbol{S}$. From the definition of $\boldsymbol{S}$ in (4.1), this Markov chain can be constructed from the original Markov chain $\{J(k) : k \geq 0\}$ governed by $\boldsymbol{A}$ in the following manner. Let $N(0) = J(0)$. Following a sample path of $\{J(k) : k \geq 0\}$, find $k^*(k)$, $k = 1, 2, \cdots$, satisfying $k^*(0) = 0$ and

$$(4.14) \quad k^*(k) = k^*(k-1) + min\{t : J(k^*(k-1) + t - 1) \in L(n), J(k^*(k-1) + t) \in L(m), \text{ for some } n > m\}.$$

Then set $N(k) = J(k^*(k)), k = 1, 2, \cdots$. In words, $\{N(k) : k \geq 0\}$ can be constructed by masking out all transitions of $\{J(k) : k \geq 0\}$ between two consecutive "downward" transitions among lumps. From this observation, or algebraically from (4.1) , one sees that $\boldsymbol{S}$ has zero columns under the subset Z of V defined by

$$(4.15) \quad Z = \{j : j \in L(m) \text{ for some } m \in H \text{ and } a_{i,j} = 0 \text{ for all } i \in L(n) \text{ with } n > m\}.$$

We note that $Z \supset L(M)$. Let $W = V \setminus Z$. Then there exists a permutation matrix $\boldsymbol{X}$ such that

$$(4.16) \quad \boldsymbol{X}\boldsymbol{S}\boldsymbol{X}^{-1} = \begin{bmatrix} \boldsymbol{S}_{WW} & \mathbf{0} \\ \boldsymbol{S}_{ZW} & \mathbf{0} \end{bmatrix}$$

The next theorem provides an alternative form of the necessary and sufficient condition given in Theorem 3.3.

Theorem 4.2 *Let $\boldsymbol{A}$ be any ergodic stochastic matrix having the ergodic probability vector $\boldsymbol{\pi}^{\mathrm{T}}$. Then Algorithm 1 converges geometrically for an arbitrary initial positive vector $\boldsymbol{\pi}^{\mathrm{T}}(0)$ if and only if $\boldsymbol{S}_{WW}$ given in* (4.16) *is aperiodic.*

Proof : Since the eigenvalue structure is invariant under similarity transformation, the theorem holds from Theorem 3.3, (4.5) of Theorem 4.1, and (4.16) . ■

It should be noted that Theorem 4.2 does not require the matrix $\boldsymbol{S}$ to be ergodic. For Algorithm 1 to converge geometrically, the Markov chain $\{N(k) : k \geq 0\}$ cannot have any closed set of states with any periodicity but still may have transient states for which stationary probabilities would be zero. The next theorem is immediate from Theorem 4.2.

Theorem 4.3 *Let $\boldsymbol{A}$ be any ergodic stochastic matrix having the ergodic probability vector $\boldsymbol{\pi}^{\mathrm{T}}$. Then Algorithm 1 does not converge for any initial positive vector $\boldsymbol{\pi}^{\mathrm{T}}(0)$ except for $\boldsymbol{\pi}^{\mathrm{T}}$ if and only if $\{N(k) : k \geq 0\}$ has a closed set of states with common periodicity.*

Theorems 4.2 and 4.3 suggest that, given an ergodic stochastic matrix $\boldsymbol{A}$, Algorithm 1 may converge for a certain set of lumps but may not converge for another set of lumps. This point is illustrated through the following example.

Example 1 We consider a Markov chain $\{J(k) : k \geq 0\}$ on $V = \{1, 2, 3, 4\}$ governed by an ergodic stochastic matrix $\boldsymbol{A}$ given by

$$(4.17) \quad \boldsymbol{A} = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1/2 & 1/2 \end{bmatrix} .$$

It can be readily seen that the Markov chain has the ergodic vector

$$\boldsymbol{\pi}^{\mathrm{T}} = [1/5,\ 1/5,\ 1/5,\ 2/5] \ . \tag{4.18}$$

Case A : L(1)={ 1 } , L(2)={ 2 } , L(3)={ 3 } , L(4)={ 4 }
W={ 1, 2, 3 } , and Z={ 4 }
In this case, $\boldsymbol{S}$ and $\boldsymbol{S}_{WW}$ computed from (4.1) and (4.16) respectively become

$$\boldsymbol{S} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \ , \quad \text{and} \quad \boldsymbol{S}_{WW} = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \ . \tag{4.19}$$

Hence $\{N(k); \geq 0\}$ has a closed set of states W with periodicity one, and the point Gauss-Seidel iteration algorithm does not converge for any positive initial vector, except $\boldsymbol{\pi}^{\mathrm{T}}$.

Case B : L(1)={ 1 , 2} , L(2)={ 3 } , L(3)={ 4 }
W={ 2, 3 } , and Z={ 1, 4 }
Here $\boldsymbol{S}$ and $\boldsymbol{S}_{WW}$ are computed from (4.1) and (4.16) respectively as

$$\boldsymbol{S} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \ , \quad \text{and} \quad \boldsymbol{S}_{WW} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \ . \tag{4.20}$$

Again $\{N(k) : k \geq 0\}$ has a closed set of states W with periodicity one, and the block Gauss-Seidel iteration algorithm does not converge for any positive initial vector, except $\boldsymbol{\pi}^{\mathrm{T}}$, for this lumping.

Case C : L(1)={ 3, 4 } , L(2)={ 2 } , L(3)={ 1 }
We rearrange the states $\{1,2,3,4\} \rightarrow \{4,3,2,1\}$, so that the states are decomposed in non-decreasing order of lump numbers. Then, with W={ 4 } , and Z={ 3, 2, 1 } , $\boldsymbol{S}$ and $\boldsymbol{S}_{WW}$ are computed from (4.1) and (4.16) respectively as

$$\boldsymbol{S} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \ , \quad \text{and} \quad \boldsymbol{S}_{WW} = [1] \ . \tag{4.21}$$

Since state 1 in W is positive recurrent and states 2, 3 and 4 in Z are transient, the block Gauss-Seidel iteration algorithm converges geometrically for any positive initial vector. Indeed, since $\boldsymbol{S}$ itself is idempotent in this case, the algorithm converges in one step.

If each lump contains only one state, then the block Gauss-Seidel Algorithm reduces to point Gauss-Seidel Algorithm. It is known that point Gauss-Seidel Algorithm always converges if the states are rearranged properly for any ergodic Markov chain (See [4] pp228-232). The next theorem shows that one can always construct a set of lumps, where at least one lump contains more than two states, so that the block Gauss-Seidel algorithm converges geometrically for any positive initial vector.

Theorem 4.4 *For any ergodic stochastic matrix $\boldsymbol{A}$, there exists at least one set of lumps for which Algorithm 1 converges geometrically starting with an arbitrary positive vector.*

Proof : Choose an arbitrary state $i \in V$. Since the original Markov chain $\{J(k) : k \geq 0\}$ is ergodic, there exists at least one path

$$i = i(1) \to i(2) \to \cdots \to i(T+1) = i \ , \quad i(k) \neq i(j) \text{ for } k, j \in \{1, 2, \cdots T\}, \tag{4.22}$$

such that the probability of returning to state i along this path starting from state i is positive. Then construct a set of lumps in such a way that $i(t) \in L(t), t = 1, \cdots, T$. Other lumps can be constructed arbitrarily and with this lumping, it is clear that the constructed Markov chain $\{N(k) : k \geq 0\}$ has a positive probability for self-transition at state i. Hence at least one of the diagonal elements of $\boldsymbol{S}_{WW}$ is positive and therefore is aperiodic. The theorem now follows from Theorem 4.2. ■

Note that one can make the number of lumps strictly less than N. Actually some of adjacent states on the path $i = i(1) \to i(2) \to \cdots \to i(T)$ can belong to one lump, unless all states in this path belong to one lump. This provides that the lumps are constructed in nondecreasing order toward state $i(T)$.

The following example illustrates Theorem 4.4.

Example 2 We consider the same Markov chain given in Example 1, and focus on a returning path $3 \to 2 \to 1 \to 4 \to 3$. Along this path, various lumpings can be constructed for which geometric convergence of the algorithm is assured by Theorem 4.4. For example,

$$L(1) = 3, L(2) = 2, L(3) = 1, L(4) = 4; \tag{4.23}$$

$$L(1) = \{3, 2\}, L(2) = \{1, 4\}; \tag{4.24}$$

$$L(1) = \{3\}, L(2) = \{2, 1\}, L(3) = \{4\}; \tag{4.25}$$

and

$$L(1) = \{3, 2, 1\}, L(2) = \{4\}, \text{etc.} \tag{4.26}$$

Here we pick up the lumping (4.25) . With rearrangement of the states given as { 1, 2, 3, 4 } → { 3, 1, 2, 4 }, the resulting stochastic matrix $\boldsymbol{A}$ is given by

$$\boldsymbol{A} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 1/2 & 0 & 0 & 1/2 \end{bmatrix} . \tag{4.27}$$

From this matrix, $\boldsymbol{S}$ and $\boldsymbol{S}_{WW}$ are computed from (4.1) and (4.16) respectively as

$$\boldsymbol{S} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} , \quad \text{and} \quad \boldsymbol{S}_{WW} = [1]. \tag{4.28}$$

In Case B of Example 1, one has $L(1) = \{1, 2\}, L(2) = \{3\}, L(3) = \{4\}$ and Algorithm 1 does not converge. However, in this example, by merely changing the lump numbers via rearrangement, Algorithm 1 converges for any positive initial vector.

The block Gause-Seidel algorithm is often applied to compute ergodic probability solutions of queueing systems. The next is an example of applying Theorem 4.4 to a queueing system $M/G/1(N)$.

Example 3 We consider an $M/G/1$ queueing system with finite waiting room, arrival rate λ and service time distribution function $F(x)$. Let $X(k)$ be the number of customers upon departure of the k-th customer. Then $\{X(k) : k \geq 0\}$ is a discrete Markov chain on the state space $V = \{0, 1, 2, \cdots N\}$, where N is the capacity of customers in system, governed by a stochastic matrix

$$\boldsymbol{P} = \begin{bmatrix} q_0 & q_1 & q_2 & \cdots & q_{N-1} & \sum_{n=N}^{\infty} q_n \\ q_0 & q_1 & q_2 & \cdots & q_{N-1} & \sum_{n=N}^{\infty} q_n \\ 0 & q_0 & q_1 & \cdots & q_{N-2} & \sum_{n=N-1}^{\infty} q_n \\ 0 & 0 & q_0 & \cdots & q_{N-3} & \sum_{n=N-2}^{\infty} q_n \\ \vdots & & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & q_0 & \sum_{n=1}^{\infty} q_n \end{bmatrix} \quad ; \quad q_n = \int_0^{\infty} \frac{(\lambda t)^n}{n!} e^{-\lambda t} dF(t) \ , \quad n = 0, 1, 2, \cdots .$$

If we choose numbers $\beta(1), \beta(2), \cdots, \beta(m)$ such that $0 \leq \beta(1) < \beta(2) < \cdots < \beta(m) = N$, and construct lumping $\{L(k)\}_{k=1}^{m}$ as

$$(4.29) \qquad \begin{cases} L(1) &= \{\ i \mid 0 \leq i \leq \beta(1)\ \}, \\ L(k) &= \{\ i \mid \beta(k-1)+1 \leq i \leq \beta(k)\ \}, \quad \text{for} \quad k = 2, 3, \cdots, m, \end{cases}$$

then Algorithm 1 converges for this lumping. To show this, we take an arbitrary number j ; $1 \leq j \leq m-1$, and consider a returning path $\beta(j) \to \beta(j)+1 \to \beta(j)$. From (4.29), $\beta(j) \in L(j), \beta(j)+1 \in L(j+1)$. Hence, Algorithm 1 converges by Theorem 4.4.

Acknowledgment

The authors wish to thank Julian Keilson and Paul Schweitzer for helpful discussions on the subject. The authors also wish to thank anonymous referees for construction comments.

References

[1] Blum, E.K. *Numerical Analysis and Computation : Theory and Practice,* Addison-Wesley, California, 1972.

[2] Cao, W. and Stewart, W.T., "Iterative Aggregation/Disaggregation Techniques for Nearly Uncoupled Markov Chains," *Journal of the ACM,* Vol.32 (1985), 702-719.

[3] Debreu, G. and Herstein, I. N., "Nonnegative Square Matrices," *Econometrica,* Vol.21 (1953), 597-607.

[4] Heyman, D.P. and Sobel, M.J. (eds.) *Handbooks in Operations Research and Management Science,* Vol.2, *Stochastic Models,* Elsevier Science Publishers, North-Holland, 1991.

[5] Mitra, D. and Tsoucas, P., "Convergence of Relaxations for Numerical Solutions of Stochastic Problems," *Mathematical Computer Performance and Reliability,* (Courtois and Hordijk, eds.) North Holland, Amsterdam, 119-133.

[6] Schweitzer, P., "Aggregation Methods for Large Markov Chains," *Mathematical Computer Performance and Reliability,* (Courtois and Hordijk, eds.) North Holland, Amsterdam, 275-286.

[7] Sumita, U. and Rieders, M., "A New Algorithm for Computing the Ergodic Probability Vector for Large Markov Chains : Replacement Process Approach," *Probability in the Engineering and Information Sciences,* Vol.4 (1990), 89-116.

[8] Sumita, U. and Rieders, M., "Application of the Replacement Process Approach for Computing the Ergodic Probability Vector of Large Scale Row-Continuous Markov Chains," *Journal of Operations Research Society of Japan,* Vol.33, No.4 (1990), 279-307.

[9] Takahashi, Y., "A Lumping Method for Numerical Calculations of Stationary Distributions of Markov Chains", Res. Rep. No. B-18, Dept. of Information Sciences, Tokyo Institute of Technology, 1975.

[10] Takahashi, Y. and Takami, Y., "A Numerical Method for the Steady-State Probabilities of a GI/G/c Queueing System in a General Case," *Journal of Operations Research Society of Japan,* Vol.19, No.2 (1976), 147-157.

Ushio SUMITA
Graduate School of International Management
International University of Japan
Yamato, Minami Uonuma, Niigata 942-72, Japan
E-mail : sumita@iuj.ac.jp

Nobuko IGAKI
Department of Management Information Systems
and Decision Sciences
Tezukayama University
Tezukayama 7-1-1, Nara 631 Japan
E-mail : igaki@tezukayama-u.ac.jp