Numer. Math. 65, 109–120 (1993)

Numerische Mathematik

Entrywise perturbation theory and error analysis for Markov chains

Colm Art O'Cinneide*

School of Industrial Engineering, Grissom Hall, Purdue University, West Lafayette, IN 47907-1287, USA

Received September 5, 1992

Summary. Grassmann, Taksar, and Heyman introduced a variant of Gaussian elimination for computing the steady-state vector of a Markov chain. In this paper we prove that their algorithm is stable, and that the problem itself is well-conditioned, in the sense of entrywise relative error. Thus the algorithm computes each entry of the steady-state vector with low relative error. Even the small steady-state probabilities are computed accurately. The key to our analysis is to focus on entrywise relative error in both the data and the computed solution, rather than making the standard assessments of error based on norms. Our conclusions do not depend on any condition numbers for the problem.

Mathematics Subject Classification (1991): 65F05, 65G05, 15A51, 60J10, 60J27

1. Introduction

A fundamental problem in computational probability is that of finding the steady-state distribution of a finite-state, discrete-time, irreducible Markov chain. This is equivalent to finding the steady-state vector of an irreducible stochastic matrix. See Seneta [9] for the basic facts and terminology of stochastic matrices and Markov chains. We formally state

Problem I. The data consists of an irreducible stochastic matrix P of order n. The problem is to compute the steady-state vector $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_n)$, which is determined by

$$\boldsymbol{\pi} P = \boldsymbol{\pi}, \qquad \sum_{j=1}^{n} \pi_j = 1 \,.$$

In [4], Grassmann, Taksar, and Heyman introduced a variant of Gaussian elimination for this problem. Their method, which is now known as the GTH

* This work was supported by NSF under grants DMS-9106207 and DDM-9203134

algorithm, is the subject of this paper. Empirical evidence that the GTH algorithm computes steady-state probabilities with small relative error has been documented in [4, 6]. The GTH algorithm involves no subtractions, and therefore loss of significant digits due to cancellation is ruled out completely. Grassmann, Taksar, and Heyman attributed the accuracy of their method to the absence of subtraction. The analysis presented here supports this.

The key result of this paper is that Problem I above is particularly well conditioned, in the sense of *entrywise relative error*. This is to say, if small relative errors are introduced into the entries of P, then only small relative errors result in the entries of π. Theorem 1 in the next section contains the formal statement. This strengthens a closely related result of Seneta ([9], Theorem 7.2 and its corollary). An important consequence of Theorem 1 is that the steady-state vector of an irreducible stochastic matrix is close to that of its floating-point approximation in the sense of entrywise relative error. (Minor annoyances such as the floating-point matrix failing to be stochastic are dealt with in Sect. 2.) In view of this conditioning property, we should seek an algorithm that computes the steady-state vector with small entrywise relative error. We prove in Sect. 4 that the GTH algorithm of [4] achieves this. It has been noted [5, 6, 14, 16] that Gaussian elimination may produce large relative errors when used to solve Problem I. A quick summary of our results is that the relative error in the entries of the steady-state vector computed by GTH is bounded by roughly $9n^2\boldsymbol{u}$, $\boldsymbol{u}$ being the unit roundoff, when a small fraction of the computation is done in double precision. Due to statistical effects, the actual error is expected to be much smaller than this bound.

The accurate computation of the small entries of a steady-state vector is a real concern in stochastic modeling. We illustrate this with an example from telecommunications. Consider the problem of modeling packetized data arriving at a statistical multiplexer in a communication network [17]. A server reads an address from each incoming packet, and then directs the packet to the appropriate outgoing channel. Since the server takes time to read the address of the packet, some buffering is needed to prevent loss of other arriving packets. The choice of buffer size is an important design issue. For switching in the Broadband Integrated Services Digital Networks (B-ISDN's) of the future [17], certain applications, such as transmission of high-definition television signals, demand a packet loss rate of no more than about one in a billion (10^{-9}) [2]. In the model, this packet loss rate is the steady-state probability that the buffer is full. For the model to be useful, we must be able to calculate this small probability. Theorem 2 below guarantees that the GTH algorithm with double precision arithmetic computes each component of the steady-state vector with low relative error for models with on the order of 1000 states.

Our analysis assumes nothing beyond irreducibility about the structure of the Markov chain, unlike [12–14, 16] where the nearly-uncoupled case is discussed. It involves no growth factors or condition numbers. The former is to be expected, as $P - I$ is (weakly) diagonally dominant [3], but the latter is surprising. Our analysis is also unusual in that it is not based on norms, and in that it assures good entrywise relative accuracy of the computed solution. The existing literature is generally based on norms and absolute error [1, 5, 8, 16], and does not yield entrywise relative error bounds for the computed solution.

2. Conditioning

When a stochastic matrix is stored electronically in floating-point form, the machine matrix may not be stochastic. This is partly because the row sums of the machine matrix may not equal 1, due to rounding. This difficulty may be avoided by removing the redundancy in the full matrix P, and recording only the off-diagonal entries as data. However, there is still the possibility that the machine data will not represent a stochastic matrix, since the sum of the off-diagonal entries of a row may exceed 1 due to rounding, this implying a negative diagonal entry. Such perturbations are not problematic, and we incorporate them into the analysis by slightly extending the scope of Problem I. We now describe the extended problem.

Let $G = (g_{ij})$ be a *generator* of order n. That is to say G has nonnegative off-diagonal entries and row sums 0. We further assume that G is irreducible, which is to say that $G + \lambda I$ is an irreducible nonnegative matrix for $\lambda > 0$ sufficiently large. (I denotes the identity matrix of the appropriate order.) Such matrices G are in fact the *continuous-time* analogues of the P's of Problem I: they are the generators of irreducible, continuous-time Markov chains. The extended problem is to find the steady-state vector of G.

Problem II. The data is the *off-diagonal* entries g_{ij}, $i \neq j$, of an irreducible order-n generator G. The problem is to compute the steady-state vector $\boldsymbol{\pi}$ of G, which is determined by the equations

$$\sum_{i=1, i\neq j}^{n} \pi_i g_{ij} = \pi_j \left(\sum_{i=1, i\neq j}^{n} g_{ji} \right), \qquad \sum_{j=1}^{n} \pi_j = 1 .$$

An alternative way of writing the first n equations here is $\boldsymbol{\pi} G = 0$. We have chosen to write them in a way that emphasizes the fact that only the off-diagonal entries of G are needed. To cast Problem I in the form of Problem II, we need only take p_{ij}, $i \neq j$, as our data g_{ij}, $i \neq j$. The matrix G is then $P - I$. Thus Problem I is a special case of Problem II.

For completeness, we state a simple and well-known lemma on irreducible generators, which is needed below.

Lemma 1. *Let the order-n irreducible generator G be partitioned as*

$$G = \begin{pmatrix} -\alpha & \boldsymbol{w}^{\mathrm{T}} \\ \boldsymbol{v} & B \end{pmatrix}, \tag{2.1}$$

where B is a square matrix of order $n-1$. Then B is invertible, $F = -B^{-1}$ is a nonnegative matrix, and the vector

$$\boldsymbol{a} = (\mathbf{1}; \boldsymbol{w}^{\mathrm{T}} F) \tag{2.2}$$

is positive and satisfies $\boldsymbol{a}G = \mathbf{0}$.

Of course, the steady-state vector $\boldsymbol{\pi}$ of G in the lemma is simply $\boldsymbol{a}/s$ where $s = a_1 + a_2 + \cdots + a_n$. In (2.1) we are using the convention that vectors are column vectors by default, so that row vectors such as $\boldsymbol{w}^{\mathrm{T}}$ are indicated with a transpose. Conventions in probability lead us to make exceptions for left eigenvectors of G such as $\boldsymbol{\pi}$ and $\boldsymbol{a}$, which are row vectors. On another point of notation, the partition (2.1) allows us to refer to the entries of G without using

complicated subscripts. For example, we prefer $\boldsymbol{w}^{\mathrm{T}} = (w_2, w_3, \ldots, w_n)$ to $(g_{12}, g_{13}, \ldots, g_{1n})$, although they mean the same thing.

Theorem 1 below is the good-conditioning property of Problem II promised in the Introduction. It includes the observation that the machine data for Problem II will always represent a true irreducible generator; Problem I was not robust under rounding errors in this sense.

Theorem 1. *Let g_{ij}, $i \neq j$, be the off-diagonal entries of an irreducible generator G of order n, and suppose that the numbers $\tilde{g}_{ij}$, $i \neq j$, satisfy*

$$K_{\mathrm{L}} g_{ij} \leqq \tilde{g}_{ij} \leqq K_{\mathrm{U}} g_{ij}, \quad i \neq j\,, \tag{2.3}$$

where $0 < K_{\mathrm{L}} \leqq 1 \leqq K_{\mathrm{U}}$. Then $\tilde{g}_{ij}$, $i \neq j$, are the off-diagonal entries of an irreducible generator $\tilde{G}$, and the steady-state vectors $\boldsymbol{\pi}$ and $\tilde{\boldsymbol{\pi}}$ of G and $\tilde{G}$ satisfy

$$\left(\frac{K_{\mathrm{L}}}{K_{\mathrm{U}}}\right)^n \pi_j \leqq \tilde{\pi}_j \leqq \left(\frac{K_{\mathrm{U}}}{K_{\mathrm{L}}}\right)^n \pi_j, \quad j = 1, 2, \ldots, n\,, \tag{2.4}$$

and

$$\left(\frac{K_{\mathrm{L}}}{K_{\mathrm{U}}}\right)^n \frac{\pi_j}{\pi_k} \leqq \frac{\tilde{\pi}_j}{\tilde{\pi}_k} \leqq \left(\frac{K_{\mathrm{U}}}{K_{\mathrm{L}}}\right)^n \frac{\pi_j}{\pi_k}, \quad j, k = 1, 2, \ldots, n\,. \tag{2.5}$$

Proof. $\tilde{G}$ is defined uniquely by setting its off-diagonal entries to be the $\tilde{g}_{ij}$'s, and its diagonal entries to be the negatives of the off-diagonal row sums. Inequality (2.3) implies that $\tilde{g}_{ij} \geqq 0$ for $i \neq j$, since $K_{\mathrm{L}} \geqq 0$, and therefore $\tilde{G}$ is a generator. Irreducibility of a generator is determined by the pattern of zeros and nonzeros in its off-diagonal entries, and, since this pattern is the same for G and $\tilde{G}$ (because in fact $K_{\mathrm{L}} > 0$), it follows that $\tilde{G}$ is irreducible.

Suppose now that G and $\tilde{G}$ differ only in their first row, and that (2.3) holds. When we partition $\tilde{G}$ as in (2.1), the matrix B is the same as for G, while the vector $\boldsymbol{w}^{\mathrm{T}}$ is different. We denote the new $\boldsymbol{w}^{\mathrm{T}}$ vector by $\tilde{\boldsymbol{w}}^{\mathrm{T}}$. Since $F = -B^{-1}$ is the same for both generators, we may identify left eigenvectors of eigenvalue zero for G and $\tilde{G}$ according to (2.2) as

$$\boldsymbol{a} = (1; \boldsymbol{w}^{\mathrm{T}} F) \qquad \text{and} \qquad \tilde{\boldsymbol{a}} = (1; \tilde{\boldsymbol{w}}^{\mathrm{T}} F)\,.$$

By definition, $a_1 = \tilde{a}_1 = 1$. For the other components of $\tilde{a}$, with f_{ij} denoting the (i,j)-entry of F, we deduce from (2.3) that

$$\tilde{a}_j = \sum_{i=2}^{n} \tilde{w}_i f_{ij} \leqq K_{\mathrm{U}} \sum_{i=2}^{n} w_j f_{ij} = K_{\mathrm{U}} a_j, \quad j > 1\,,$$

and

$$\tilde{a}_j = \sum_{i=2}^{n} \tilde{w}_i f_{ij} \geqq K_{\mathrm{L}} \sum_{i=2}^{n} w_j f_{ij} = K_{\mathrm{L}} a_j, \quad j > 1\,.$$

Combining these we have

$$K_{\mathrm{L}} a_j \leqq \tilde{a}_j \leqq K_{\mathrm{U}} a_j, \quad j = 1, 2, \ldots, n\,. \tag{2.6}$$

With $s = \sum_1^n a_j$ and $\tilde{s} = \sum_1^n \tilde{a}_j$, it follows that

$$K_{\mathrm{L}} s \leqq \tilde{s} \leqq K_{\mathrm{U}} s\,. \tag{2.7}$$

As $\pi_j = a_j/s$ and $\tilde{\pi}_j = \tilde{a}_j/\tilde{s}$, we deduce from (2.6) and (2.7) that

$$\left(\frac{K_L}{K_U}\right)\pi_j \leqq \tilde{\pi}_j \leqq \left(\frac{K_U}{K_L}\right)\pi_j, \quad j = 1, 2, \ldots, n . \tag{2.8}$$

Similarly, as $\pi_j/\pi_k = a_j/a_k$ and $\tilde{\pi}_j/\tilde{\pi}_k = \tilde{a}_j/\tilde{a}_k$, we deduce from (2.6) that

$$\left(\frac{K_L}{K_U}\right)\frac{\pi_j}{\pi_k} \leqq \frac{\tilde{\pi}_j}{\tilde{\pi}_k} \leqq \left(\frac{K_U}{K_L}\right)\frac{\pi_j}{\pi_k}, \quad j, k = 1, 2, \ldots, n . \tag{2.9}$$

Theorem 1 now follows upon changing the rows of G into those of $\tilde{G}$ one row at a time, each time invoking the bounds given by (2.8) and (2.9). □

When K_L and K_U are both close to 1, hypothesis (2.3) implies that the off-diagonal entries of G and $\tilde{G}$ are close in the sense of relative error. To clearly relate Theorem 1 to the conditioning of Problem II, we give a special case.

Corollary 1. *Let g_{ij}, $i \neq j$, be the off-diagonal entries of an irreducible generator G of order n, and let $\tilde{g}_{ij}$ be close to g_{ij} for $i \neq j$ in the sense of entrywise relative error:*

$$|\tilde{g}_{ij} - g_{ij}| \leqq \varepsilon g_{ij}, \quad i \neq j ,$$

where $0 \leqq \varepsilon < 1$. Then $\tilde{g}_{ij}$, $i \neq j$, are the off-diagonal entries of an irreducible generator $\tilde{G}$, and the steady-state vectors π and $\tilde{\pi}$ of G and $\tilde{G}$ satisfy

$$\frac{|\tilde{\pi}_j - \pi_j|}{\pi_j} \leqq \left(\frac{1+\varepsilon}{1-\varepsilon}\right)^n - 1 = 2n\varepsilon + O(\varepsilon^2), \quad j = 1, 2, \ldots, n .$$

Proof. The hypothesis is equivalent to (2.3) with $K_L = 1 - \varepsilon$ and $K_U = 1 + \varepsilon$. The conclusion follows from (2.4) by routine manipulations. □

If $\tilde{g}_{ij}$ denotes the floating-point approximation to the real number g_{ij}, then the hypothesis of Corollary 1 holds with $\varepsilon = \boldsymbol{u}$, the unit roundoff in floating-point arithmetic. The conclusion is that the machine data has a steady-state vector which is within an entrywise relative error of only about $2n\boldsymbol{u}$ of the true steady-state vector.

We digress to give a simple probabilistic insight into these results, for the reader familiar with Markov chains. Divide the path of the Markov chain with generator G into regenerative cycles according to visits to state 1. Classify these cycles according to the state first visited after state 1. Thus a cycle of type $j \geqq 2$ begins with a visit to state 1 followed by a visit to state j. The relative frequency of cycles of type j is proportional to $g_{1j} = w_j$. Thus, small relative changes in the w_j's will result in small changes in the relative frequencies of the different types of cycles, and so the steady-state probabilities will experience small relative changes.

Theorem 7.2 and its corollary in Seneta [9] contain the idea of Theorem 1. However, Theorem 1 allows larger relative perturbations on the diagonal than Seneta's results do, but gives essentially the same conclusion. Theorem 5 of Meyer and Stewart [8] characterizes the sensitivity of the steady-state vector to perturbations in the entries of a stochastic matrix P, but does not explicitly treat relative perturbations of the kind considered here. Insensitivity of the steady-state vector to small relative perturbations in the data is a theme of [11–13, 15, 16], but some block structure is assumed. Normwise perturbation analysis [1, 8, 13] leads to

consideration of the norm of the Drazin inverse of G as a measure of the conditioning of Problem II. The conclusions of this analysis suffer from the drawbacks of norms mentioned at the end of the Introduction.

As we stated in the Introduction, the GTH algorithm brings to fruition the possibility engendered in Theorem 1 of computing the steady-state vector with good entrywise relative error. In the next section we describe the GTH algorithm. We remark that Gaussian elimination with floating-point arithmetic is an unstable algorithm for Problem II. This is because, while Theorem 1 assures us that the solution is well determined by the floating-point representation of the data, Gaussian elimination may introduce substantial errors into the computed solution [6, 5]. See [14, 16], where this point is illustrated in the light of a conditioning property for the nearly-uncoupled case.

We close this section with an example showing that perturbations of the approximate magnitude of the bounds in (2.4) and (2.5) are possible under (2.3). We adopt the setting of Corollary 1, as it is easy to work with. Consider the order-n generator G_ε defined by

$$\begin{pmatrix} -1-\varepsilon & 1+\varepsilon & 0 & \cdots & 0 & 0 \\ (M-1)(1-\frac{M+1}{M-1}\varepsilon) & -M(1-\varepsilon) & 1+\varepsilon & \cdots & 0 & 0 \\ (M-1)(1-\frac{M+1}{M-1}\varepsilon) & 0 & -M(1-\varepsilon) & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ (M-1)(1-\frac{M+1}{M-1}\varepsilon) & 0 & 0 & \cdots & -M(1-\varepsilon) & 1+\varepsilon \\ M(1-\varepsilon) & 0 & 0 & \cdots & 0 & -M(1-\varepsilon) \end{pmatrix},$$

where $M, \varepsilon > 0$. The maximal entrywise relative perturbation in the generator G_ε compared to G_0 is $(M+1)\varepsilon/(M-1)$. The steady-state probabilities are given by

$$\pi_j(\varepsilon) = C\left(\frac{1+\varepsilon}{M(1-\varepsilon)}\right)^{j-1}, \quad j = 1, 2, \ldots, n,$$

where C is determined to make the sum 1. Let us take εsmall and M large, in such a way that $M = \varepsilon^{-2}$. Then $C = 1 + O(\varepsilon^2)$, so that C has no impact on the second non-zero term in the expansion of $\pi_j(\varepsilon)$ in powers of ε. Moreover, the maximal entrywise relative perturbation $(M+1)\varepsilon/(M-1)$ becomes $\varepsilon + O(\varepsilon^2)$, and so the hypothesis of Corollary 1 is satisfied (ignoring higher order terms in ε) with $G = G_0$ and $\tilde{G} = G_\varepsilon$. A little calculation gives

$$\frac{|\pi_j(\varepsilon) - \pi_j(0)|}{\pi_j(0)} = 2(j-1)\varepsilon + O(\varepsilon^2).$$

For ε small, the maximal relative perturbation in the steady-state vector is in its n-th component, and is $2(n-1)\varepsilon + O(\varepsilon^2)$. Thus the bound $2n\varepsilon + O(\varepsilon^2)$ of Corollary 1 is nearly best possible. The same example shows that the bounds (2.4) and (2.5) are nearly best possible.

3. The GTH algorithm

Let us point out some of the features that may be exploited in solving Problem II. More details may be found in [5]. The question is to find a nonzero vector $\boldsymbol{a}$ with $\boldsymbol{a}G = \mathbf{0}$, where G is an irreducible generator. Let us factorize G in the form LU, by Gaussian elimination, where U is unit upper triangular and L is lower triangular. In doing this, no pivoting will be necessary [3], and the last column of L will be 0. As U is invertible, the system $\boldsymbol{a}G = \mathbf{0}$ reduces to the triangular system $\boldsymbol{a}L = \mathbf{0}$. This is solved by setting $a_n = 1$ and back-substituting. There results a left eigenvector $\boldsymbol{a}$ of G with eigenvalue zero, which only has to be scaled to identify the steady-state vector. Note that only one triangular system has to be solved.

Consider the LU-factorization of the generator G, carried out without pivoting. This produces a series of matrices of decreasing order $G = G^{(1)}, G^{(2)}, G^{(3)}, \ldots$, where $G^{(k)}$ denotes the matrix to the south-east of the k-th pivot entry (and including that pivot entry), just before the k-th Gauss transformation [7] is applied. These may be defined inductively as follows. Having defined $G^{(k)}$, we partition it as

$$G^{(k)} = \begin{pmatrix} -\alpha_k & \boldsymbol{w}_k^{\mathrm{T}} \\ \boldsymbol{v}_k & B_k \end{pmatrix},$$

where B_k is of order $n - k$. We define $G^{(k+1)}$ by

$$G^{(k+1)} = B_k + \frac{\boldsymbol{v}_k \boldsymbol{w}_k^{\mathrm{T}}}{\alpha_k}. \tag{3.1}$$

It is easily verified that $G^{(k+1)}$ inherits the property of being a generator from $G^{(k)}$. Therefore we have

$$\alpha_k = \sum_{j=k+1}^{n} w_{kj}, \quad k = 1, 2, \ldots, n, \tag{3.2}$$

where w_{kj} is the $(j-k)$-th entry of the vector $\boldsymbol{w}_k$.

Back-substitution now proceeds as follows. We set $a_n = 1$, and, for $k = n-1, n-2, \ldots, 1$, calculate a_k according to

$$a_k = \frac{1}{\alpha_k} \sum_{j=k+1}^{n} a_j v_{jk}, \tag{3.3}$$

v_{jk} being the $(j-k)$-th entry of the vector $\boldsymbol{v}_k$. It remains only to scale the a_j's to find the π_j's. To do this, set

$$s = a_1 + a_2 + \cdots + a_n, \qquad \text{and} \qquad \pi_j = a_j / s, \quad j = 1, 2, \ldots, n. \tag{3.4}$$

The GTH algorithm follows formulas (3.1) – (3.4). Let us describe it in steps.

GTH ALGORITHM

For $k = 1, 2, \ldots, n-1$

Step 1: Calculate α_k according to (3.2).

Step 2: Calculate the off-diagonal entries of (3.1).

Set $a_n = 1$.

For $k = n-1, n-2, \ldots, 1$

Step 3: Calculate a_k according to (3.3).

Step 4: Calculate the π_j's according to (3.4).

The simple Step 1 is what distinguishes GTH from standard Gaussian elimination. It allows one to avoid computing the diagonals in (3.1) by subtraction, and so ensures that no cancellation occurs in the algorithm.

The matrix $G^{(k)}$ has a probabilistic interpretation. It is the generator for the underlying Markov chain observed only while it is in the reduced state space $\{k, k+1, \ldots, n\}$. For this reason, the GTH algorithm is sometimes called the State-Reduction Algorithm, although this interpretation applies whether or not (3.2) is used to compute the pivots. For more on the probabilistic interpretation of Gaussian elimination, see [4].

4. Error analysis

If one attempts a simple forward error analysis of Gaussian elimination, rounding errors appear to grow exponentially. This is because no account is being taken of the conditioning of the problem. Backward error analysis [18] allows one to separate the conditioning of the problem from the stability of the algorithm. Problem II has been shown to be well-conditioned irrespective of the data, and so a forward error analysis of GTH may be possible. Here we give such an analysis. We prove that the computed steady-state probabilities are accurate to a relative error of $O(n^3)\boldsymbol{u}$, where $\boldsymbol{u}$ is the unit roundoff in floating-point arithmetic. For example, with $\boldsymbol{u} = 5 \times 10^{-14}$, a typical value for double-precision arithmetic, our analysis guarantees that for a 1000-state Markov chain every steady-state probability, *no matter how small*, is computed with about 4 correct digits (see Theorem 2). For a chain with 10,000 states, one accurate digit is guaranteed. Statistical effects in rounding error accumulation lead us to expect much better accuracy.

Our analysis also shows how the accuracy of the algorithm might be improved at very little cost. We find that the rounding errors in computing the pivots as sums in (3.2) are responsible for the leading-order term in our error bound. If these pivots are accumulated in double precision (or at least higher precision), we find that the computed steady-state probabilities are provably accurate to a relative error of only $O(n^2)\boldsymbol{u}$. This compares quite well with the best possible accuracy, which is about $2n\boldsymbol{u}$ according to Theorem 1 and the example following it. It is achieved at an additional cost of only $O(n^3)$ double-precision additions, which is negligible compared to the overall cost of $O(n^3)$ operations.

In what follows, a "hat" will indicated a value computed in floating-point arithmetic. In assessing floating-point errors, care must be taken to fully exploit nonnegativity and the absence of subtraction. The main consequence of this is that the result of a sequence of floating-point operations, including additions, will have good relative accuracy. The reason for this is, in essence, that if each term in a sum of nonnegative numbers is approximated to a relative error of ε, then the (exact) sum of the approximations will have a relative error of at most ε also. Following Appendix 3 of [10], we write $\langle k \rangle$ for a quotient of the form

$$\frac{(1+\varepsilon_1)(1+\varepsilon_2)\cdots(1+\varepsilon_a)}{(1+\delta_1)(1+\delta_2)\cdots(1+\delta_b)},$$

where each ε_i and δ_i is no greater in magnitude than the unit roundoff $\boldsymbol{u}$, and $a + b = k$. As explained in [10], these symbols provide a convenient way of keeping

track of relative-error bounds. For example, we have the useful formalism $\langle k_1\rangle\langle k_2\rangle = \langle k_1 + k_2\rangle$. Moreover, assuming that $k\boldsymbol{u} \leqq 0.1$, we have

$$\langle k\rangle = 1 + \varepsilon, \quad \text{where } |\varepsilon| \leqq 1.06k\boldsymbol{u} . \tag{4.1}$$

See [7, 10, 18] for the basics of rounding-error analysis. Our first result is

Theorem 2. *For every instance of Problem II of order n with floating-point data, the GTH algorithm computes the solution with accuracy characterized by*

$$\hat{\pi}_j = \langle 2\phi(n) + n\rangle \pi_j, \quad j = 1, 2, \ldots, n ,$$

where $\phi(n) = (1/3)(2n^3 + 6n^2 - 8n)$. *If* $(2\phi(n) + n)\boldsymbol{u} \leqq 0.1$, *then*

$$\frac{|\hat{\pi}_j - \pi_j|}{\pi_j} \leqq 1.06(2\phi(n) + n)\boldsymbol{u}, \quad j = 1, 2, \ldots, n .$$

Proof. It simplifies things a little to first address the accuracy of the left eigenvector a computed at Step 3 of GTH, rather than the ultimate solution π. Recall that a_j is π_j/π_n, so that $a_n = 1$. In this proof, *we refer to a left eigenvector of a generator whose last entry is* 1 *as an a-vector.* We construct a function ϕ such that for any instance of Problem II of order m in floating-point form, when its a-vector $\boldsymbol{a}$ is computed by GTH in floating-point arithmetic with unit roundoff $\boldsymbol{u}$, the computed a-vector is entrywise within a factor of $\langle\phi(m)\rangle$ of the true a-vector:

$$\hat{a}_j = \langle\phi(m)\rangle a_j, \quad j = 1, 2, \ldots, m . \tag{4.2}$$

This implies, upon taking account of the additional inaccuracy in the $\hat{\pi}_j$'s due to the $m - 1$ additions and one division in Step 4 of GTH, that

$$\hat{\pi}_j = \langle 2\phi(m) + m\rangle \pi_j ,$$

from which Theorem 2 follows by (4.1).

It remains to prove (4.2) for a suitable function ϕ. The proof is by induction on m. We may define $\phi(1)$ to be zero, as GTH assigns the correct value 1 to a_1 when $n = 1$. The induction hypothesis is that we have assigned values to $\phi(1), \phi(2), \ldots, \phi(n-1)$, for some integer $n > 1$, in such a way that (4.2) holds for $m = 1, 2, \ldots, n - 1$. We show how to assign a value to $\phi(n)$ so that (4.2) continues to hold for $m = n$. In this way, ϕ is defined so that (4.2) holds for all m.

Having set out our induction hypothesis, let G be an irreducible floating-point generator of order n. To minimize subscripting, we partition G using the notation of (2.1). Step 1 of GTH is to compute the pivot α according to equation (3.2). The sum of $n - 1$ nonnegative floating-point numbers is computed with a relative error characterized by

$$\hat{\alpha} = \langle n - 2\rangle \alpha . \tag{4.3}$$

Next, according to Step 2, the off-diagonal entries of $G^{(2)}$ are computed. A superscript (2) will identify quantities associated with $G^{(2)}$. Entrywise, the formula is

$$g_{ij}^{(2)} = g_{ij} + \frac{v_i w_j}{\alpha}, \quad i \neq j, \ i, j = 2, 3, \ldots, n . \tag{4.4}$$

As α is computed with the error described in (4.3), and three more rounding errors are introduced in the multiplication, division, and addition needed to compute

$g_{ij}^{(2)}$, we see that the relative error in computing the off-diagonal entries of $G^{(2)}$ is characterized by

$$\hat{g}_{ij}^{(2)} = \langle n+1 \rangle g_{ij}^{(2)}, \quad i \neq j . \tag{4.5}$$

Of course, we are using the fact that everything is nonnegative.

Now we have shown that the off-diagonal entries of $\hat{G}^{(2)}$ are within a factor of $\langle n+1 \rangle$ of those of $G^{(2)}$. Thus inequality (2.3) holds with K_U and K_L being the upper and lower bounds on the range of $\langle n+1 \rangle$, which are $(1-u)^{-(n+1)}$ and $(1-u)^{n+1}$, respectively. Therefore, by inequality (2.5) of Theorem 1, the true a-vector $\boldsymbol{a}^{(2)} = (a_2, a_3, \ldots, a_n)$ of $G^{(2)}$ is entrywise within a factor of $\langle 2(n-1)(n+1) \rangle = \langle 2(n^2-1) \rangle$ of the true a-vector of $\hat{G}^{(2)}$. Note that in applying (2.5) $G^{(2)}$ is of order only $n-1$.

The GTH algorithm continues now by simply applying the GTH algorithm itself to the order-$(n-1)$ generator $\hat{G}^{(2)}$. This allows us to argue informally as follows. In the preceding paragraph, we saw that the true a-vector $\boldsymbol{a}^{(2)}$ for $G^{(2)}$ is close to that of $\hat{G}^{(2)}$. Since the latter matrix is of order only $n-1$, the induction hypothesis assures us that the true a-vector of $\hat{G}^{(2)}$ is computed accurately. Thus the computed a-vector of $\hat{G}^{(2)}$ is close to the true a-vector of $G^{(2)}$. Let us repeat this argument, but now quantifying the accuracy. From the preceding paragraph, the true a-vector of $G^{(2)}$ is entrywise within a factor of $\langle 2(n^2-1) \rangle$ of the true a-vector of $\hat{G}^{(2)}$. By the induction hypothesis, the computed a-vector for $\hat{G}^{(2)}$ is entrywise within a factor of $\langle \phi(n-1) \rangle$ of the true a-vector of $\hat{G}^{(2)}$. Combining these statements, we find that the a-vector $\boldsymbol{a}^{(2)}$ of $G^{(2)}$ is computed accurate entrywise to a factor of $\langle \phi(n-1) + 2(n^2-1) \rangle$.

As $\boldsymbol{a} = (a_1; \boldsymbol{a}^{(2)})$, it remains only to assess the error in computing a_1 from the back-substitution equation (3.3) (Step 3 of GTH). The accuracy inherited by $\hat{a}_1$ is easily seen to be characterized by a factor of $\langle \phi(n-1) + 2(n^2-1) + 2n - 2 \rangle$, taking account of the $n-1$ multiplications, $n-2$ additions, and the division by the inaccurate $\hat{\alpha}$ of (4.3). From this we conclude that the following choice of $\phi(n)$ preserves (4.2) for $m = n$ as long as $\phi(n-1)$ does for $m = n-1$:

$$\phi(n) = \phi(n-1) + 2n^2 + 2n - 4 .$$

With the initial condition $\phi(1) = 0$ this determines an allowable choice of $\phi(n)$ for all integers $n \geqq 1$:

$$\phi(n) = \tfrac{1}{3}(2n^3 + 6n^2 - 8n) . \tag{4.6}$$

We have proved (4.2) with this choice of ϕ, and Theorem 2 follows. □

We show below that the accuracy of the computed pivot $\hat{\alpha}$ is critical to our overall assessment of the accuracy of the computed steady-state vector. There are several considerations which would suggest either that (4.3) is an overestimate of error in $\hat{\alpha}$, or that this error can be reduced cheaply. For example

- If we calculate the sum in (3.2) in double precision, or at least in somewhat greater precision, $\hat{\alpha}$ will be accurate to a factor of $\langle 1 \rangle$. Since only about $n^2/2$ additions are expended in computing these pivots throughout the GTH algorithm, as against $O(n^3)$ operations overall, the extra cost of high-precision computation of the pivots is negligible.
- By summing the w's in (3.2) in pairs repeatedly, the pivot will be computed accurate to a factor of $\langle \log_2 n + 1 \rangle$, compared to the $\langle n-2 \rangle$ of (4.3).

Summing in increasing order of magnitude is another accuracy-improving device.

- Statistical effects in the accumulation of rounding errors in (3.2) suggest that the error factor in $\hat{\alpha}$ will be closer to $\langle\sqrt{n}\rangle$ than to $\langle n-2\rangle$.
- If the generator satisfies $g_{ij}=0$ for $j>i+K$, with $K \ll n$, as is sometimes the case in modeling, then at most $K-1$ nonzero entries must be added for each pivot, giving an error characterized by a factor of $\langle K-2\rangle$. This applies to the block-tridiagonal examples of Heyman [6].

We now repeat the analysis leading to Theorem 2, with the assumption that the pivot sums in (3.2) are accumulated in double precision. Our induction hypothesis is that the GTH algorithm, with this improvement, computes the a-vector of any floating-point generator of order $m<n$ with accuracy characterized by

$$\hat{a}_j=\langle\psi(m)\rangle a_j, \quad j=1,2,\ldots,m, \tag{4.7}$$

where $\psi(1),\psi(2),\ldots,\psi(n-1)$ are given values. Our goal is to identify a value for $\psi(n)$ such that (4.7) continues to hold for $m=n$. The only difference in this analysis is in our assessment of the error in the pivot $\hat{\alpha}$, which, in contrast to (4.3), is now

$$\hat{\alpha}=\langle 1\rangle\alpha .$$

A routine assessment of rounding error for (4.4) shows that the relative error in computing the off-diagonal entries of $G^{(2)}$ at Step 2 of GTH is characterized by

$$\hat{g}_{ij}^{(2)}=\langle 4\rangle g_{ij}^{(2)}, \quad i\neq j . \tag{4.8}$$

Therefore, by inequality (2.5) with $K_{\mathrm{L}}=(1-u)^4$ and $K_{\mathrm{U}}=(1-u)^{-4}$, the true a-vector $\boldsymbol{a}^{(2)}=(a_2,a_3,\ldots,a_n)$ of $G^{(2)}$ is entrywise within a factor $\langle 8(n-1)\rangle$ of the true a-vector of $\hat{G}^{(2)}$. Note that $G^{(2)}$ is of order $n-1$ in applying (2.5).

By the induction hypothesis, the computed a-vector for $\hat{G}^{(2)}$ is entrywise within a factor of $\langle\psi(n-1)\rangle$ of the true a-vector of $\hat{G}^{(2)}$. It follows that $a^{(2)}=(a_2,a_3,\ldots,a_n)$ is computed accurate entrywise to a factor of $\psi\mathbf{t}(n-1)+8(n-1)\rangle$. It remains to assess the error in computing a_1 by back-substitution according to (3.3). The relative error inherited by $\hat{a}_1$ is characterized by a factor of $\langle\psi(n-1)+8(n-1)+n+1\rangle$, by routine error analysis. From this we conclude that the following choice of $\psi(n)$ preserves (4.7) for $m=n$ as long as $\psi(n-1)$ does for $m=n-1$:

$$\psi(n)=\psi(n-1)+9n-7 .$$

With the initial condition $\psi(1)=0$ this determines $\psi(n)$ for $n\geq 1$:

$$\psi(n)=\tfrac{1}{2}(9n^2-5n-4) . \tag{4.9}$$

We have proved (4.7) for all m with this function ψ. Arguing as we did just following (4.2), we summarize our results as

Theorem 3. *For every instance of Problem II of order n with floating-point data, the GTH algorithm with pivots totaled in double precision computes the solution with accuracy characterized by*

$$\tilde{\pi}_j=\langle 2\psi(n)+n\rangle\pi_j, \quad j=1,2,\ldots,n ,$$

where $\psi(n)$ *is given by* (4.9). *If* $(2\psi(n) + n)\boldsymbol{u} \leqq 0.1$, then

$$\frac{|\hat{\pi}_j - \pi_j|}{\pi_j} \leqq 1.06(2\psi(n) + n)\boldsymbol{u} \leqq 9.54n^2\boldsymbol{u} .$$

In Sect. 2 we saw that the relative error due to initial roundoff is in the worst case about $2n\boldsymbol{u}$. The bound given by Theorem 3 is greater by a factor of only about $5n$. Under the weight of $O(n^3)$ single-precision operations, this worst-case accuracy seems very good.

Acknowledgements. The author thanks D.P. Heyman and G.W. Stewart for their comments, and for providing copies of their recent papers on this subject.

References

1. Barlow, J.L. (1992): Error bounds for the computation of null-vectors with applications to Markov chains. Proceedings of the IMA Workshop on Linear Algebra, Markov Chains, and Queuing Models, University of Minnesota (to appear)
2. Blondia, C. (1991): Finite-capacity vacation models with non-renewal input. J. Appl. Probab. **28**, 174–197
3. Funderlic, R.E., Neumann, M., Plemmons, R.J. (1982): LU decompositions of generalized diagonally dominant matrices. Numer. Math. **40**, 57–69
4. Grassmann, W.K., Taksar, M.I., Heyman, D.P. (1985): Regenerative analysis and steady-state distributions for Markov chains. Oper. Res. **33**(5), 1107–1116
5. Harrod, W.J., Plemmons, R.J. (1984): Comparison of some direct methods for computing stationary distributions of Markov chains. SIAM J. Sci. Stat. Comput. **5**(2), 453–469
6. Heyman, D.P. (1987): Further comparisons of direct methods for computing stationary distributions of Markov chains. SIAM J. Alg. Disc. Meth. **8**(2), 226–232
7. Golub, G.H., Van Loan, C.F. (1989): Matrix computations. Johns Hopkins University Press, Baltimore
8. Meyer, C.D., Stewart, G.W. (1988): Derivatives and perturbations of eigenvectors. SIAM J. Numer. Anal. **25**(3), 679–691
9. Seneta, E. (1981): Nonnegative matrices and Markov chains. Springer, Berlin Heidelberg New York
10. Stewart, G.W. (1973): Introduction to matrix computations. Academic Press, New York
11. Stewart, G.W. (1983): Computable error bounds for aggregated Markov chains. J. Assoc. Comput. Mach. **30**(2), 271–285
12. Stewart, G.W. (1984): On the structure of nearly uncoupled Markov chains. In: Iazeolla, G.G., Courtois, P.J., Hordijk A. ed., Mathematical Computer Performance and Reliability, pp. 287–302. Elsevier, North-Holland
13. Stewart, G.W. (1990): On the sensitivity of nearly uncoupled Markov chains. In: Stewart, W.J. ed., Numerical Solution of Markov Chains. Marcel Dekker, New York
14. Stewart, G.W. (1992): Rounding error, perturbation theory, and Markov chains. To appear in Proceedings of the IMA Workshop on Linear Algebra, Markov Chains, and Queuing Models, University of Minnesota
15. Stewart, G.W. (1992): On the perturbation of Markov chains with nearly transient states. Technical Report. Department of Computer Science and Institute for Advanced Computer Studies, University of Maryland
16. Stewart, G.W., Zhang, G. (1991): On a direct method for the solution of nearly uncoupled Markov chains. Numer. Math. **59**, 1–11
17. Walrand, J. (1991): Communication networks: a first course. Irwin, Boston
18. Wilkinson, J.H. (1963): Rounding errors in algebraic processes. Prentice-Hall, Englewood Cliffs, NJ