SOMEWHAT STOCHASTIC MATRICES

BRANKO ĆURGUS AND ROBERT I. JEWETT

Abstract. The standard theorem for stochastic matrices with positive entries is generalized to matrices with no sign restriction on the entries. The condition that column sums be equal to 1 is kept, but the positivity condition is replaced by a condition on the distances between columns.

1. Introduction.

The notion of a Markov chain is ubiquitous in linear algebra and probability books. In linear algebra a Markov chain is a sequence $\{\mathbf{x}_k\}$ of vectors defined recursively by a specified vector $\mathbf{x}_0$, a square matrix P and the recursion $\mathbf{x}_k = P\mathbf{x}_{k-1}$ for $k = 1, 2, \ldots$. That is, $\mathbf{x}_k = P^k\mathbf{x}_0$. Natural probabilistic restrictions are imposed on $\mathbf{x}_0$ and P. It is assumed that $\mathbf{x}_0$ is a *probability vector*; that is, its entries are nonnegative and add up to 1. It is assumed that P is a *stochastic matrix*; that is, it is a square matrix whose columns are probability vectors. The original version of the main theorem about Markov chains appears in Markov's paper [2]. In the language of linear algebra it reads:

> *Suppose that P is a stochastic matrix with all positive entries. Then there exists a unique probability vector $\mathbf{q}$ such that $P\mathbf{q} = \mathbf{q}$. If $\{\mathbf{x}_k\}$ is a Markov chain determined by P, then it converges to $\mathbf{q}$.*

More generally, the same conclusion holds for a stochastic matrix P for which P^s has all positive entries for some positive integer s. All elementary linear algebra textbooks that we examined state this theorem. None give a complete proof. Partial proofs, or intuitive explanations of the theorem's validity, are always based on knowledge about the matrix's eigenvalues and eigenvectors. This argument becomes sophisticated when the matrix is not diagonalizable.

What these proofs leave obscure is a certain contractive property of a stochastic matrix already observed by Markov. Of course, this contractive property is explored in research papers and some advanced books. However, the relative simplicity of the underlining idea gets lost in the technical details of an advanced setting. We feel that this contractive property deserves to be popularized. We use it here to provide a direct proof of a theorem which is more general than the one stated above.

We consider real square matrices A whose columns add up to 1. Such a matrix we call a *somewhat stochastic* matrix. The probabilistic condition that all entries be nonnegative is dropped. Instead of assuming that all entries of A^s are positive, we make an assumption about distances between the columns of A^s. This assumption leads to a contractive property of a matrix that yields convergence. This and other definitions are given next.

2010 *Mathematics Subject Classification.* Primary: 15B51. Secondary: 15B48, 60J10.

Key words and phrases. stochastic matrix, Markov chain, generalized stochastic matrix, contractivity coefficient.

2. Definitions.

All numbers in this note are real, except in Example 3. All matrices are square and will be denoted by upper case letters. All vectors, except for $\mathbb{1}$, are column vectors and will be denoted by bold lower case letters. All entries of the row vector $\mathbb{1}$ are equal to 1,

$$\mathbb{1} = [1 \ \ 1 \ \ \cdots \ \ 1].$$

This row vector helps to express conditions that were already mentioned and will appear repeatedly. The equation $\mathbb{1}A = \mathbb{1}$ says that all column sums of A are equal to 1. And $\mathbb{1}\mathbf{x} = 1$ says that the entries of a column vector $\mathbf{x}$ sum to 1, while $\mathbb{1}\mathbf{x} = 0$ says that the entries of a column vector $\mathbf{x}$ sum to 0. We use the standard notation $\mathbb{N}$ for the set of positive integers and $\mathbb{R}$ for the set of real numbers.

For a vector $\mathbf{x}$ with entries $x_1, \dots, x_n$ we set

$$\|\mathbf{x}\| := \sum_{j=1}^{n} |x_j|.$$

Notice that the distance $\|\mathbf{x}-\mathbf{y}\|$ between vectors $\mathbf{x}$ and $\mathbf{y}$ associated with this norm is the n-dimensional version of the Manhattan distance.

Consider an $n \times n$ matrix A with columns $\mathbf{a}_1, \dots, \mathbf{a}_n$ and entries a_{ij}. For the purpose of the next two definitions, we think of the columns of a matrix as points in $\mathbb{R}^n$. In this way the concept of a diameter of a set is applied to a matrix as follows:

$$\operatorname{diam} A = \max_{1 \le i,j \le n} \|\mathbf{a}_i - \mathbf{a}_j\|.$$

Next we define

$$\operatorname{var} A = \frac{1}{2} \operatorname{diam} A = \max_{1 \le i,j \le n} \frac{1}{2} \sum_{l=1}^{n} |a_{li} - a_{lj}|.$$

We call this quantity the *column variation* of a matrix A. The idea of using that quantity is due to Markov [2, Section 5]. In [2], for fixed i, j the quantity $\frac{1}{2}\sum_{l=1}^{n} |a_{li} - a_{lj}|$ is not given explicitly as half the sum of the absolute values of the real numbers $a_{li} - a_{lj}$, but rather as the sum of the positive terms in this list. Since Markov considered only stochastic matrices, for which the sum of all terms in this list is 0, the quantity he used coincides with the variation. For more on Markov's work see [4]. The column and row variation appear in research literature under various names; see [1, Section 3.3].

Recalling the original theorem about a Markov chain stated in our first paragraph, we will show that the inequality

$$\|P^k \mathbf{x}_0 - \mathbf{q}\| \le (\operatorname{var} P)^k \|\mathbf{x}_0 - \mathbf{q}\| \qquad (1)$$

holds for all k. Furthermore, for a stochastic matrix P with all positive entries it turns out that $\operatorname{var} P < 1$. This strict contractive property of a stochastic matrix with all positive entries implies convergence of the Markov chain $\{P^k \mathbf{x}_0\}$.

3. The column variation of a matrix.

The first step towards a proof of inequality (1) is the proposition that follows. To repeat, our results do not require entries to be nonnegative. Not only that, but in this proposition A could be a rectangular matrix with complex entries. However, the assumption that the entries of the vector $\mathbf{y}$ are real is essential, as is shown in Example 3.

Proposition. *Let A be an $n \times n$ real matrix and let $\mathbf{y}$ be an $n \times 1$ vector with real entries such that $\mathbb{1}\mathbf{y} = 0$. Then*

$$\|A\mathbf{y}\| \leq (\operatorname{var} A)\|\mathbf{y}\|. \tag{2}$$

Proof. We will use the common notation for the positive and negative part of a real number t: $t^+ = \max\{t, 0\}$ and $t^- = \max\{-t, 0\}$. Clearly $t^+, t^- \geq 0$ and $t = t^+ - t^-$ and $|t| = t^+ + t^-$.

Let A be an $n \times n$ matrix with columns $\mathbf{a}_1, \ldots, \mathbf{a}_n$ and let $\mathbf{y} \in \mathbb{R}^n$ be such that $\mathbb{1}\mathbf{y} = 0$.

The inequality (2) is obvious if $\mathbf{y} = \mathbf{0}$. Assume that $\mathbf{y} \neq \mathbf{0}$. Set $\mathbf{z} = \bigl(2/\|\mathbf{y}\|\bigr)\mathbf{y}$. Then, (2) is equivalent to

$$\|A\mathbf{z}\| \leq (\operatorname{var} A)\|\mathbf{z}\| \qquad \text{with} \qquad \|\mathbf{z}\| = 2. \tag{3}$$

Clearly $\mathbb{1}\mathbf{z} = 0$. Let $z_1, \ldots, z_n$ be the entries of $\mathbf{z}$. Then we have

$$2 = \|\mathbf{z}\| = \sum_{j=1}^{n} |z_j| = \sum_{j=1}^{n} \bigl(z_j^+ + z_j^-\bigr) = \sum_{j=1}^{n} z_j^+ + \sum_{j=1}^{n} z_j^-$$

and

$$0 = \sum_{j=1}^{n} z_j = \sum_{j=1}^{n} \bigl(z_j^+ - z_j^-\bigr) = \sum_{j=1}^{n} z_j^+ - \sum_{j=1}^{n} z_j^-.$$

From the last two displayed relations we deduce that

$$\sum_{j=1}^{n} z_k^+ = \sum_{j=1}^{n} z_k^- = 1. \tag{4}$$

Using again the notation introduced at the beginning of the proof we get

$$A\mathbf{z} = \sum_{j=1}^{n} z_j \mathbf{a}_j = \sum_{j=1}^{n} z_j^+ \mathbf{a}_j - \sum_{i=1}^{n} z_i^- \mathbf{a}_i. \tag{5}$$

Since $A\mathbf{z}$ is represented in (5) as a difference of two convex combinations of the columns of A, the inequality $\|A\mathbf{z}\| \leq \operatorname{diam} A$ follows from the geometrically clear fact that a set has the same diameter as its convex hull. However, we use (4) and (5) to continue with an algebraic argument:

$$\begin{aligned} A\mathbf{z} &= \sum_{j=1}^{n} \left(\sum_{i=1}^{n} z_i^- \right) z_j^+ \mathbf{a}_j - \sum_{i=1}^{n} \left(\sum_{j=1}^{n} z_j^+ \right) z_i^- \mathbf{a}_i \\ &= \sum_{j=1}^{n} \sum_{i=1}^{n} z_j^+ z_i^- \bigl(\mathbf{a}_j - \mathbf{a}_i\bigr). \end{aligned}$$

Consequently,

$$\begin{aligned} \|A\mathbf{z}\| &\leq \sum_{j=1}^{n} \sum_{i=1}^{n} z_j^+ z_i^- \bigl\|\mathbf{a}_j - \mathbf{a}_i\bigr\| && \text{(by the triangle inequality and } z_j^+, z_j^- \geq 0\text{)} \\ &\leq (\operatorname{diam} A) \sum_{k=1}^{n} z_k^+ \sum_{j=1}^{n} z_j^- && \text{(by definition of } \operatorname{diam} A\text{)} \\ &= 2(\operatorname{var} A) && \text{(by (4) and definition of } \operatorname{var} A\text{)} \\ &= (\operatorname{var} A)\, \|\mathbf{z}\| && \text{(since } \|\mathbf{z}\| = 2\text{).} \end{aligned}$$

This completes the proof of (3) and the theorem is proved. □

4. Powers of somewhat stochastic matrices.

Let A be a somewhat stochastic matrix, that is, A is square and real and $\mathbb{1}A = \mathbb{1}$. The equalities

$$\mathbb{1}A^k = (\mathbb{1}A)A^{k-1} = \mathbb{1}A^{k-1} = \cdots = (\mathbb{1}A)A = \mathbb{1}A = \mathbb{1}$$

show that any power A^k of A is somewhat stochastic. Furthermore, if $\mathbb{1}\mathbf{y} = 0$, then $\mathbb{1}A^k\mathbf{y} = \mathbb{1}\mathbf{y} = 0$ for all positive integers k. This property of a somewhat stochastic matrix A allows us to repeatedly apply the proposition to powers of A. Assuming that $\mathbb{1}\mathbf{y} = 0$ and, for the sake of simplicity, setting $c = \operatorname{var} A$ we have

$$\|A^k\mathbf{y}\| = \|A(A^{k-1}\mathbf{y})\| \leq c\,\|A^{k-1}\mathbf{y}\| \leq \cdots \leq c^{k-1}\|A\mathbf{y}\| \leq c^k\|\mathbf{y}\|. \tag{6}$$

Now we are ready to state and prove the main result.

Theorem. *Let A be an $n\times n$ somewhat stochastic matrix. Assume that there exists $s \in \mathbb{N}$ such that $\operatorname{var}(A^s) < 1$. Then:*

(a) *there exists a unique $\mathbf{q} \in \mathbb{R}^n$ such that*

$$A\mathbf{q} = \mathbf{q} \qquad \text{and} \qquad \mathbb{1}\mathbf{q} = 1,$$

(b) *if $\mathbf{x}$ is such that $\mathbb{1}\mathbf{x} = 1$, then the sequence $\{A^k\mathbf{x}\}$ converges to $\mathbf{q}$ as k tends to $+\infty$.*

Proof. The assumption that $\mathbb{1}A = \mathbb{1}$ means that 1 is an eigenvalue of $A^\top$, the transpose of A. Since $A^\top$ and A have the same eigenvalues, there exists a real nonzero vector $\mathbf{v}$ such that $A\mathbf{v} = \mathbf{v}$.

Let s be a positive integer such that $\operatorname{var}(A^s) < 1$ and set $c = \operatorname{var}(A^s)$. Clearly $A^s\mathbf{v} = \mathbf{v}$. If $\mathbb{1}\mathbf{v} = 0$, then the proposition yields

$$\|\mathbf{v}\| = \|A^s\mathbf{v}\| \leq c\|\mathbf{v}\| < \|\mathbf{v}\|.$$

This is a contradiction. Therefore $\mathbb{1}\mathbf{v} \neq 0$. Setting $\mathbf{q} = \bigl(\mathbb{1}\mathbf{v}\bigr)^{-1}\mathbf{v}$ provides a vector whose existence is claimed in (a). To verify uniqueness, let $\mathbf{p}$ be another such vector. Then $\mathbb{1}(\mathbf{p} - \mathbf{q}) = 0$, $A^s(\mathbf{p} - \mathbf{q}) = \mathbf{p} - \mathbf{q}$, and, by the proposition,

$$\|\mathbf{p} - \mathbf{q}\| = \|A^s(\mathbf{p} - \mathbf{q})\| \leq c\|\mathbf{p} - \mathbf{q}\|.$$

Consequently, $\mathbf{p} - \mathbf{q} = 0$, since $0 \leq c < 1$.

Let $k \in \mathbb{N}$ be such that $k > s$ and assume $\mathbb{1}\mathbf{y} = 0$. By the division algorithm there exist unique integers j and r such that $k = sj + r$ and $r \in \{0, \ldots, s-1\}$. Here $j > (k/s) - 1 > 0$. Now we apply (6) to the matrix A^s and vector $A^r\mathbf{y}$. We obtain

$$\|A^k\mathbf{y}\| = \|(A^s)^jA^r\mathbf{y}\| \leq c^j\|A^r\mathbf{y}\|.$$

Consequently, for all $k > s$ we have

$$\|A^k\mathbf{y}\| \leq c^{(k/s)-1}\max_{0\leq r<s}\|A^r\mathbf{y}\|. \tag{7}$$

Let $\mathbf{x} \in \mathbb{R}^n$ be such that $\mathbb{1}\mathbf{x} = 1$. Then $\mathbb{1}(\mathbf{x} - \mathbf{q}) = 0$. Substituting $\mathbf{y} = \mathbf{x} - \mathbf{q}$ in (7) yields

$$\bigl\|A^k\mathbf{x} - \mathbf{q}\bigr\| = \bigl\|A^k(\mathbf{x} - \mathbf{q})\bigr\| \leq c^{(k/s)-1}\max_{1\leq r<s}\bigl\|A^r(\mathbf{x} - \mathbf{q})\bigr\|.$$

Now, since $0 \leq c < 1$, we get $A^k\mathbf{x} \to \mathbf{q}$ as $k \to +\infty$. This proves (b) and completes the proof. □

The standard theorem about Markov chains is a special case of the theorem.

Corollary. *Let P be an $n \times n$ stochastic matrix. Assume that there exists $s \in \mathbb{N}$ such that all entries of P^s are positive. Then:*

(a) *there exists a unique probability vector $\mathbf{q} \in \mathbb{R}^n$ such that $P\mathbf{q} = \mathbf{q}$,*

(b) *if $\mathbf{x}$ is a probability vector, then the sequence $\{P^k\mathbf{x}\}$ converges to $\mathbf{q}$ as k tends to $+\infty$.*

Proof. To apply the theorem, we will prove that $\operatorname{var}(P^s) < 1$. For $i, j \in \{1, \ldots, n\}$, denote by b_{ij} the entries of P^s which, by assumption, are positive. Next, notice that for positive numbers a and b we have $|a - b| < a + b$. Therefore, for arbitrary i, j we have

$$\sum_{l=1}^{n} |b_{li} - b_{lj}| < \sum_{l=1}^{n} (b_{li} + b_{lj}) = 2.$$

This proves that the distance between arbitrary columns of P^s is less then 2. Consequently, $\operatorname{diam}(P^s) < 2$ and hence $\operatorname{var}(P^s) < 1$. Now we apply the theorem for convergence. The proofs of the remaining claims are standard. □

The theorem can be restated in terms of the powers of A. This follows from the following equivalency.

Let A be an arbitrary square matrix and let $\mathbf{q}$ be a vector such that $\mathbb{1}\mathbf{q} = 1$. Denote by Q the square matrix each of whose columns is equal to $\mathbf{q}$, that is, $Q = \mathbf{q}\mathbb{1}$. Then the following two statements are equivalent:

(i) if $\mathbf{x}$ is such that $\mathbb{1}\mathbf{x} = 1$, then the sequence $\{A^k\mathbf{x}\}$ converges to $\mathbf{q}$ as k tends to $+\infty$,

(ii) the powers A^k tend to Q as k tends to $+\infty$.

Assume (i) and let $\mathbf{e}_1, \ldots, \mathbf{e}_n$ be the vectors of the standard basis of $\mathbb{R}^n$. Then the j-th column of A^k is $A^k\mathbf{e}_j$. By (i), $A^k\mathbf{e}_j$ converges to $\mathbf{q}$ as k tends to $+\infty$. This proves (ii).

Now assume (ii) and let $\mathbf{x}$ be a vector with $\mathbb{1}\mathbf{x} = 1$. Then $A^k\mathbf{x}$ converges to $Q\mathbf{x} = (\mathbf{q}\mathbb{1})\mathbf{x} = \mathbf{q}(\mathbb{1}\mathbf{x}) = \mathbf{q}$. This proves (i).

In fact, Q is a projection onto the span of $\mathbf{q}$. To see this, calculate $Q^2 = (\mathbf{q}\mathbb{1})(\mathbf{q}\mathbb{1}) = \mathbf{q}(\mathbb{1}\mathbf{q})\mathbb{1} = \mathbf{q}\,1\,\mathbb{1} = \mathbf{q}\mathbb{1} = Q$ and $Q\mathbf{x} = \mathbf{q}(\mathbb{1}\mathbf{x}) = (\mathbb{1}\mathbf{x})\mathbf{q}$ for any $\mathbf{x} \in \mathbb{R}^n$.

5. Examples.

We conclude the note with three examples.

Example 1. The matrix

$$A = \frac{1}{5}\begin{bmatrix} 0 & 2 & -4 \\ -1 & -1 & 0 \\ 6 & 4 & 9 \end{bmatrix}.$$

is somewhat stochastic. The largest distance between two columns is between the second and the third, and it equals $12/5$. Therefore, $\operatorname{var} A = 6/5 > 1$. But

$$A^2 = \frac{1}{25}\begin{bmatrix} -26 & -18 & -36 \\ 1 & -1 & 4 \\ 50 & 44 & 57 \end{bmatrix}$$

and $\operatorname{var}(A^2) = 18/25 < 1$. Hence, the theorem applies. For this matrix, $\mathbf{q} = \begin{bmatrix} -2 & \frac{1}{3} & \frac{8}{3} \end{bmatrix}^\top$.

Example 2. Consider the following three kinds of stochastic matrices:

$$A = \begin{bmatrix} 1 & + & + \\ 0 & + & + \\ 0 & 0 & + \end{bmatrix}, \quad B = \begin{bmatrix} + & + & 0 \\ + & 0 & + \\ 0 & + & + \end{bmatrix}, \quad \text{and} \quad C = \begin{bmatrix} 0 & + & 0 \\ 0 & 0 & 1 \\ 1 & + & 0 \end{bmatrix}.$$

Here we use $+$ for positive numbers.

Since A is upper triangular, all its powers are upper triangular, so no power of A has all positive entries. Thus, the standard theorem does not apply. However, directly from the definition it follows that $\operatorname{var} A < 1$, so the theorem applies. Also, $\mathbf{q} = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}^\top$.

The matrix B is not positive, but $\operatorname{var} B < 1$; so our theorem applies. Also, the standard theorem applies here as well since B^2 is positive.

The first five powers of C are:

$$\begin{bmatrix} 0 & + & 0 \\ 0 & 0 & 1 \\ 1 & + & 0 \end{bmatrix}, \quad \begin{bmatrix} 0 & 0 & + \\ 1 & + & 0 \\ 0 & + & + \end{bmatrix}, \quad \begin{bmatrix} + & + & 0 \\ 0 & + & + \\ + & + & + \end{bmatrix}, \quad \begin{bmatrix} 0 & + & + \\ + & + & + \\ + & + & + \end{bmatrix}, \quad \text{and} \quad \begin{bmatrix} + & + & + \\ + & + & + \\ + & + & + \end{bmatrix}.$$

The variation of the first two matrices is 1, while $\operatorname{var}(C^3) < 1$. The first positive power of C is C^5.

Example 3. In this example we consider matrices with complex entries. Let $\omega = (-1 + i\sqrt{3})/2$. Then $1, \omega$, and $\overline{\omega}$ are the cube roots of unity. So, $1 + \omega + \overline{\omega} = 0$, $\overline{\omega}\omega = 1$, $\omega^2 = \overline{\omega}$ and $\overline{\omega}^2 = \omega$.

Consider one vector and two matrices:

$$\mathbf{v} = \begin{bmatrix} 1 \\ \omega \\ \overline{\omega} \end{bmatrix}, \quad A = \frac{1}{3} \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}, \quad B = \frac{1}{3} \begin{bmatrix} 1 & \overline{\omega} & \omega \\ \omega & 1 & \overline{\omega} \\ \overline{\omega} & \omega & 1 \end{bmatrix}.$$

We calculate $A\mathbf{v} = \mathbf{0}$, $B\mathbf{v} = \mathbf{v}$ and $\operatorname{var} B = \sqrt{3}/2$. Since $\|B\mathbf{v}\| > \sqrt{3}/2\|\mathbf{v}\|$, the matrix B and the vector $\mathbf{v}$ provide an example which shows that the conclusion of the proposition may not hold with complex entries.

A linear combination of A and B shows that the restriction to real numbers cannot be dropped in the theorem. Let γ be a complex number and set

$$C = A + \gamma B.$$

Since $A^2 = A$, $AB = BA = 0$ and $B^2 = B$, we have

$$C^k = A + \gamma^k B.$$

The matrix A is stochastic with variation 0, while $\mathbb{1}B = 0$ and $\operatorname{var} B = \sqrt{3}/2$. Hence, $\mathbb{1}C = \mathbb{1}$, that is, C is somewhat stochastic with complex entries. Also,

$$\operatorname{var} C = \operatorname{var}(\gamma B) = |\gamma| \sqrt{3}/2.$$

Therefore, if $1 < |\gamma| < 2/\sqrt{3}$, then $\operatorname{var} C < 1$, but the sequence $\{C^k\}$ diverges, as we can see from the formula for C^k.

Finally, we mention that the vector $\mathbf{v}$ together with the vectors $\mathbf{u} = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix}^\top$ and $\mathbf{w} = \begin{bmatrix} 1 & \overline{\omega} & \omega \end{bmatrix}^\top$ form an orthogonal basis for the complex inner product space $\mathbb{C}^3$, that A is the orthogonal projection onto the span of $\mathbf{u}$, and that B is the orthogonal projection onto the span of $\mathbf{v}$.

References

[1] I. Ipsen and T. Selee, Ergodicity coefficients defined by vector norms. SIAM J. Matrix Anal. Appl. **32** (2011) 153–200.

[2] A. A. Markov, Extension of the law of large numbers to dependent quantities (in Russian), Izvestiia Fiz.-Matem. Obsch. Kazan Univ., (2nd Ser.), **15** (1906), 135–156; also in [3, pp.339–361]; English translation [5, Section 11].

[3] A. A. Markov, *Selected works. Theory of numbers. Theory of probability.* (Russian) Izdat. Akad. Nauk SSSR, Leningrad, 1951.

[4] E. Seneta, Markov and the creation of Markov chains, in A. N. Langville and W. J. Stewart, eds., MAM 2006: Markov Anniversary Meeting, Boson Books, Raleigh, NC, 2006, 1–20.

[5] O. B. Sheynin, ed., *Probability and Statistics. Russian Papers.* Selected and translated by Oscar Sheynin. NG Verlag, Berlin, 2004; also available as item 5 at `http://www.sheynin.de/download.html`

Department of Mathematics, Western Washington University, Bellingham, Washington 98225, USA `curgus@wwu.edu` `http://faculty.wwu.edu/curgus`

Department of Mathematics, Western Washington University, Bellingham, Washington 98225, USA