On the Sensitivity of Nearly Uncoupled Markov Chains

G. W. Stewart

1. Introduction

A nearly uncoupled Markov chain (NUMC) is a discrete chain whose matrix $\mathbf{P}$ of transition probabilities is almost block diagonal. More precisely, the states of a NUMC with k blocks can be ordered so that its transition matrix assumes the form

$$\mathbf{P} = \mathbf{D} + \mathbf{E} \equiv \begin{pmatrix} \mathbf{D}_{11} & \mathbf{E}_{12} & \cdots & \mathbf{E}_{1k} \\ \mathbf{E}_{21} & \mathbf{D}_{22} & \cdots & \mathbf{E}_{1k} \\ \vdots & \vdots & & \vdots \\ \mathbf{E}_{k1} & \mathbf{E}_{k2} & \cdots & \mathbf{D}_{kk} \end{pmatrix}, \tag{1.1}$$

where the elements of the off-diagonal blocks $\mathbf{E}_{ij}$ are small. Chains of this kind are used to model systems in which certain groups of states are loosely coupled to one another. Since the system spends a relatively large amount of time in a group before passing on to another, it seems natural that a NUMC would achieve steady states within groups rather quickly and a steady state between groups more slowly. This behavior was first noted by Simon and Ando [10] and has been the subject of much subsequent research (see, e.g., [2, 11, 1, 6]).

One practical difficulty with NUMCs is that it is difficult to determine the off-diagonal elements accurately, at least empirically. This is because transitions between blocks are rare events. Consequently, the behavior of the chain must be observed over a long period to estimate the elements of $\mathbf{E}$. It is therefore important to have some idea of the sensitivity of $\boldsymbol{\pi}^{\mathrm{T}}$ to perturbation in the elements of $\mathbf{P}$. The purpose of this paper is describe the factors that make $\boldsymbol{\pi}^{\mathrm{T}}$ sensitive to such perturbation.

In the next section we will review the perturbation theory for the left Perron vector[1] of an irreducible Markov chain. The following section consists of a review the theory of NUMCs. We will then apply the results of the two sections to determine the sensitivity of $\boldsymbol{\pi}^{\mathrm{T}}$ to perturbations in $\mathbf{P}$. This is principally an expository

[1] We will use the term Perron vectors to refer to the positive eigenvectors, left and right, of an irreducible nonnegative matrix. Since a NUMC is a perforce aperiodic, it has only one eigenvalue of magnitude one and its left Perron vector is a steady-state vector.

survey of the problem. Complete bounds with proofs will appear in a subsequent paper.

Throughout this paper, we will assume that $\mathbf{P}$ is an irreducible stochastic matrix of order n, partitioned as in (1.1). For simplicity we will take $k = 3$ in displayed formulas, the general case being an obvious extension. The vector $\boldsymbol{\pi}^{\mathrm{T}}$ will denote the unique positive left Perron vector of $\mathbf{P}$ normalized so that its components sum to one. The corresponding right eigenvector of $\mathbf{P}$, whose elements are all one, will be written $\mathbf{1}$.

The symbol $\|\cdot\|$ will denote the Euclidean vector norm and the subordinate matrix norm defined by

$$\|\mathbf{A}\| = \max_{\|\mathbf{x}\|=1} \|\mathbf{A}\mathbf{x}\|.$$

Since we will be concerned with the behavior of NUMCs as $\mathbf{E}$ approaches zero, we will set

$$\epsilon = \|\mathbf{E}\|.$$

2. Perturbation Theory for Irreducible Markov Chains

In this section we will be concerned with the following problem. Let $\mathbf{P}$ be an irreducible stochastic matrix (here we do not assume that the associated chain is a NUMC) with left Perron vector $\boldsymbol{\pi}^{\mathrm{T}}$. Let

$$\tilde{\mathbf{P}} = \mathbf{P} + \mathbf{F}$$

be an irreducible, stochastic matrix with left Perron vector $\tilde{\boldsymbol{\pi}}^{\mathrm{T}}$. The problem is to find a bound on $\|\tilde{\boldsymbol{\pi}}^{\mathrm{T}} - \boldsymbol{\pi}^{\mathrm{T}}\|$.

The perturbation theory for the steady state vectors of markov chains was initiated by Schweitzer [9], and has been developed in various forms since (e.g., see [5, 3, 4, 8]). Here we present it in a form that emphasizes its interaction with the transient behavior of the chain.

We begin with a decomposition of $\mathbf{P}$.

Theorem 2.1. *The matrix* $\mathbf{P}$ *can be written in the form*

$$\mathbf{P} = (\mathbf{1}\ \mathbf{U})\mathrm{diag}(1, \mathbf{B}) \begin{pmatrix} \boldsymbol{\pi}^{\mathrm{T}} \\ \mathbf{V}^{\mathrm{T}} \end{pmatrix} = \mathbf{1}\boldsymbol{\pi}^{\mathrm{T}} + \mathbf{U}\mathbf{B}\mathbf{V}^{\mathrm{T}} \equiv \mathbf{S} + \mathbf{T}, \qquad (2.1)$$

where

$$(\mathbf{1}\ \mathbf{U})^{-1} = \begin{pmatrix} \boldsymbol{\pi}^{\mathrm{T}} \\ \mathbf{V}^{\mathrm{T}} \end{pmatrix}$$

and none of the eigenvalues of the matrix $\mathbf{B}$ *are one. Moreover,* $\|\mathbf{V}\| = 1$ *and* $\|\mathbf{U}\| \leq \sqrt{n}$.

The nonzero eigenvalues of $\mathbf{T}$ are the same as the eigenvalues of $\mathbf{B}$. If the chain is aperiodic, these eigenvalues are all less than one in magnitude. Hence $\lim_{\sigma\to\infty} \mathbf{T}^{\sigma} \to 0$, and the rate at which it converges determines the rate at which the chain converges to the STEADY-STATE MATRIX $\mathbf{S}$ For this reason, we call $\mathbf{T}$ the TRANSIENT MATRIX of the system. Note that both $\mathbf{S}$ and $\mathbf{T}$ are independent of the choice of $\mathbf{U}$ and $\mathbf{V}$.

This decomposition can be used to characterize the perturbed Perron vector $\tilde{\boldsymbol{\pi}}^{\mathrm{T}}$.

Theorem 2.2. *Let* $\mathbf{P}$ *have the decomposition (2.1) and let*

$$\tilde{\mathbf{B}} = \mathbf{V}^{\mathrm{H}}\tilde{\mathbf{P}}\mathbf{U}.$$

Then $\mathbf{I} - \tilde{\mathbf{B}}$ *is nonsingular and*

$$\tilde{\boldsymbol{\pi}}^{\mathrm{T}} = \boldsymbol{\pi}^{\mathrm{T}} + \boldsymbol{\pi}^{\mathrm{T}}\mathbf{F}\mathbf{U}(\mathbf{I} - \tilde{\mathbf{B}})^{-1}\mathbf{V}^{\mathrm{T}}. \tag{2.2}$$

Since $\mathbf{V}^{\mathrm{T}}\mathbf{1} = 0$, the vector $\tilde{\boldsymbol{\pi}}^{\mathrm{T}}$ defined by (2.2) is properly normalized. Moreover,

$$\frac{\|\tilde{\boldsymbol{\pi}}^{\mathrm{T}} - \boldsymbol{\pi}^{\mathrm{T}}\|}{\|\boldsymbol{\pi}^{\mathrm{T}}\|} \leq \|\mathbf{U}(\mathbf{I} - \tilde{\mathbf{B}})^{-1}\mathbf{V}^{\mathrm{T}}\|\|\mathbf{F}\|.$$

Now for $\mathbf{F}$ sufficiently small, $(\mathbf{I} - \tilde{\mathbf{B}})^{-1} \cong (\mathbf{I} - \mathbf{B})^{-1}$. Hence if we set

$$\mathbf{T}^{\natural} = \mathbf{U}(\mathbf{I} - \mathbf{B})^{-1}\mathbf{V}^{\mathrm{T}}$$

(Schweitzer [9] calls this the *fundamental matrix*), we may assert that

The condition of $\boldsymbol{\pi}^{\mathrm{T}}$ *is* $\|\mathbf{T}^{\natural}\|$.

Another way of putting this result is to observe that if an eigenvalues of $\mathbf{T}$ is near one, then $\mathbf{T}^{\natural}$ must be large. In other words, a slowly converging Markov chain is an ill-conditioned Markov chain. This means that we can expect troubles with NUMCs, which can have very slow transients.

3. The Theory of NUMCs

In this section we shall consider an irreducible NUMCin the form (1.1). The basic fact about such chains is that the transient matrix can be decomposed into a fast transient and a slow transient. To state the result precisely, we must introduce some notation.

First, note that as $\epsilon \to 0$, the matrices $\mathbf{D}_{ii}$ approach stochastic matrices. Hence each has an eigenvalue approaching one. We will define the set $\mathcal{L}_i$ be the set consisting of remaining eigenvalues of $\mathbf{D}_{ii}$.

Second, let

$$\boldsymbol{\pi}^{\mathrm{T}} = (\boldsymbol{\pi}_1^{\mathrm{T}}\ \boldsymbol{\pi}_2^{\mathrm{T}}\ \boldsymbol{\pi}_3^{\mathrm{T}})$$

where the partitioning is conformal with (1.1). Since $\boldsymbol{\pi}^{\mathrm{T}} > 0$, none of the components in the above partitioning is zero. Hence we may set

$$\boldsymbol{\pi}^{\mathrm{T}} = (v_1\overline{\boldsymbol{\pi}}_1^{\mathrm{T}}\ v_2\overline{\boldsymbol{\pi}}_2^{\mathrm{T}}\ v_3\overline{\boldsymbol{\pi}}_3^{\mathrm{T}}). \tag{3.1}$$

where the v_i are chosen so that the sums of the components of the $\overline{\boldsymbol{\pi}}_i$ are one. Note that $v_1+v_2+v_3 = 1$. We will call the numbers v_i the COUPLING COEFFICIENTS of the NUMC.

We are now in a position to establish the fundamental result on NUMCs

Theorem 3.1. *Let* $\mathbf{P}$ *be an irreducible* NUMC *and suppose that as* $\epsilon \to 0$*:*

1. *The coupling coefficients* v_i *are uniformly bounded away from zero.*
2. *The sets* $\mathcal{L}_i$ *are uniformly bounded away from one.*

Then there are matrices

$$\mathbf{X} = (\mathbf{1}\ \mathbf{X}_{\mathrm{s}}\ \mathbf{X}_{\mathrm{f}})$$

and

$$\mathbf{Y}^{\mathrm{T}} = \begin{pmatrix} \boldsymbol{\pi}^{\mathrm{T}} \\ \mathbf{Y}_{\mathrm{s}}^{\mathrm{T}} \\ \mathbf{Y}_{\mathrm{f}}^{\mathrm{T}} \end{pmatrix}$$

with $\mathbf{X}^{-1} = \mathbf{Y}^{\mathrm{T}}$ *such that*

$$\mathbf{Y}^{\mathrm{T}}\mathbf{P}\mathbf{X} = \mathrm{diag}(1, \mathbf{B}_{\mathrm{s}}, \mathbf{B}_{\mathrm{f}})$$

or equivalently

$$\mathbf{P} = \mathbf{1}\boldsymbol{\pi}^{\mathrm{T}} + \mathbf{X}_{\mathrm{s}}\mathbf{B}_{\mathrm{s}}\mathbf{Y}_{\mathrm{s}}^{\mathrm{T}} + \mathbf{X}_{\mathrm{f}}\mathbf{B}_{\mathrm{f}}\mathbf{Y}_{\mathrm{f}}^{\mathrm{T}} \equiv \mathbf{S} + \mathbf{T}_{\mathrm{s}} + \mathbf{T}_{\mathrm{f}}.$$

Moreover the eigenvalues of $\mathbf{B}_\mathrm{s}$ *are bounded below by* $1-O(\epsilon)$ *and the eigenvalues of* $\mathbf{B}_\mathrm{f}$ *are uniformly bounded away from one.*

It is easy to see that

$$\mathbf{P}^\sigma = \mathbf{1}\boldsymbol{\pi}^\mathrm{T} + \mathbf{X}_\mathrm{s}\mathbf{B}_\mathrm{s}^\sigma\mathbf{Y}_\mathrm{s}^\mathrm{T} + \mathbf{X}_\mathrm{f}\mathbf{B}_\mathrm{f}^\sigma\mathbf{Y}_\mathrm{f}^\mathrm{T} = \mathbf{S} + \mathbf{T}_\mathrm{s}^\sigma + \mathbf{T}_\mathrm{f}^\Sigma.$$

Since the nonzero eigenvalues of $\mathbf{T}_\mathrm{s}$ are near one while the eigenvalues of $\mathbf{T}_\mathrm{f}$ are bounded away from one, $\mathbf{T}_\mathrm{s}^\sigma$ approaches zero more slowly than $\mathbf{T}_\mathrm{f}^\sigma$. Hence they decompose the chain into slow and fast transients. The matrix $\mathbf{T} = \mathbf{T}_\mathrm{s} + \mathbf{T}_\mathrm{f}$ is the transient matrix introduced in the last section.

It will turn out that it is the slow transient that will control the sensitivity of the steady state vector. Consequently, we must have some way of approximating the matrix $\mathbf{B}_\mathrm{s}$. The following procedure gives a good approximation.

Let

$$\mathbf{Q} = \begin{pmatrix} \mathbf{1} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{1} \end{pmatrix}$$

and

$$\mathbf{R}^\mathrm{T} = \begin{pmatrix} \boldsymbol{\pi}_1^\mathrm{T} & \mathbf{0}^\mathrm{T} & \mathbf{0}^\mathrm{T} \\ \mathbf{0}^\mathrm{T} & \boldsymbol{\pi}_2^\mathrm{T} & \mathbf{0}^\mathrm{T} \\ \mathbf{0}^\mathrm{T} & \mathbf{0}^\mathrm{T} & \boldsymbol{\pi}_3^\mathrm{T} \end{pmatrix}, \tag{3.2}$$

Then it is easy to verify that

1. $\mathbf{1} = \mathbf{Q}\mathbf{1}$,
2. $\boldsymbol{\pi}^\mathrm{T} = \mathbf{v}^\mathrm{T}\mathbf{R}^\mathrm{T}$.

These two equations are the expressions for the right and left eigenvectors of $\mathbf{P}$.

Let us now produce an approximation for the slow transient. To do this, we introduce the COUPLING MATRIX

$$\mathbf{C} = \mathbf{R}^\mathrm{T}\mathbf{P}\mathbf{Q} = \begin{pmatrix} \boldsymbol{\pi}_1^\mathrm{T}\mathbf{D}_{11}\mathbf{1} & \boldsymbol{\pi}_1^\mathrm{T}\mathbf{E}_{12}\mathbf{1} & \boldsymbol{\pi}_1^\mathrm{T}\mathbf{E}_{13}\mathbf{1} \\ \boldsymbol{\pi}_2^\mathrm{T}\mathbf{E}_{21}\mathbf{1} & \boldsymbol{\pi}_2^\mathrm{T}\mathbf{D}_{22}\mathbf{1} & \boldsymbol{\pi}_2^\mathrm{T}\mathbf{E}_{23}\mathbf{1} \\ \boldsymbol{\pi}_3^\mathrm{T}\mathbf{E}_{31}\mathbf{1} & \boldsymbol{\pi}_3^\mathrm{T}\mathbf{E}_{32}\mathbf{1} & \boldsymbol{\pi}_3^\mathrm{T}\mathbf{D}_{33}\mathbf{1} \end{pmatrix}. \tag{3.3}$$

It is easy to see that $\mathbf{C}$ is stochastic:

$$\mathbf{C}\mathbf{1} = \mathbf{R}^\mathrm{T}\mathbf{P}\mathbf{Q}\mathbf{1} = \mathbf{R}^\mathrm{T}\mathbf{P}\mathbf{1} = \mathbf{R}^\mathrm{T}\mathbf{1} = \mathbf{1},$$

Moreover, its steady-state vector is $\mathbf{v}^{\mathrm{T}}$ (the vector of coupling coefficients):

$$\mathbf{v}^{\mathrm{T}}\mathbf{C} = \mathbf{v}^{\mathrm{T}}\mathbf{R}^{\mathrm{T}}\mathbf{P}\mathbf{Q} = \boldsymbol{\pi}^{\mathrm{T}}\mathbf{P}\mathbf{Q} = \boldsymbol{\pi}^{\mathrm{T}}\mathbf{Q} = \mathbf{v}^{\mathrm{T}}.$$

By Theorem 2.1, we can find matrices $\mathbf{U}$, $\mathbf{V}$, and $\widetilde{\mathbf{B}}_{\mathrm{s}}$ such that

$$(\mathbf{1}\ \mathbf{U})^{-1} = \begin{pmatrix} \mathbf{v}^{\mathrm{T}} \\ \mathbf{V}^{\mathrm{T}} \end{pmatrix}$$

and

$$\mathbf{C} = (\mathbf{1}\ \mathbf{U})\mathrm{diag}(1, \widetilde{\mathbf{B}}_{\mathrm{s}}) \begin{pmatrix} \mathbf{v}^{\mathrm{T}} \\ \mathbf{V}^{\mathrm{T}} \end{pmatrix} = \mathbf{1}\mathbf{v}^{\mathrm{T}} + \mathbf{U}\widetilde{\mathbf{B}}_{\mathrm{s}}\mathbf{V}^{\mathrm{T}}.$$

Since $\mathbf{C}$ is stochastic and by (3.3) its off-diagonal elements are $O(\epsilon)$, we have $\mathbf{C} = I + O(\epsilon)$. Since $\mathbf{V}^{\mathrm{T}}\mathbf{U} = I$ and $\widetilde{\mathbf{B}}_{\mathrm{s}} = \mathbf{V}^{\mathrm{T}}\mathbf{C}\mathbf{U}$, it follows from the boundedness of $\mathbf{U}$ and $\mathbf{V}$ that $\widetilde{\mathbf{B}}_{\mathrm{s}} = \mathbf{I} + O(\epsilon)$. Thus the eigenvalues of $\widetilde{\mathbf{B}}_{\mathrm{s}}$ behave as we would expect those of a slow transient to behave. In fact, it is essentially the slow transient.

Theorem 3.2. *Under the hypotheses of Theorem 3.1*

$$\mathbf{B}_{\mathrm{s}} = \widetilde{\mathbf{B}}_{\mathrm{s}} + O(\epsilon^2).$$

4. The Perturbation of NUMCs

We are now ready to combine the results from the last two sections. Specifically, let $\widetilde{\mathbf{P}} = \mathbf{P} + \mathbf{F}$ be a perturbation of a NUMC that satisfies regularity conditions one and two. From equation (2.2), we see that the problem of assessing the effects of $\mathbf{F}$ on the steady state vector amounts to finding the matrix $\mathbf{T}^{\natural}$.

We have seen that for a typical NUMC the transient matrix decomposes into a slow transient and a fast transient. Specifically, the transient matrix $\mathbf{T}$ is given by

$$\mathbf{T} = \mathbf{T}_{\mathrm{s}} + \mathbf{T}_{\mathrm{f}}.$$

Moreover, it is easily verified from the definitions of $\mathbf{T}_{\mathrm{s}}$ and $\mathbf{T}_{\mathrm{f}}$ that if we set

$$\mathbf{T}_{\mathrm{i}}^{\natural} = \mathbf{X}_{\mathrm{i}}(\mathbf{I} - \mathbf{T}_{\mathrm{i}})^{-1}\mathbf{Y}_{\mathrm{i}}, \qquad \mathrm{i} = \mathrm{s}, \mathrm{f}$$

then

$$\mathbf{T}^{\natural} = \mathbf{T}_{\mathrm{s}}^{\natural} + \mathbf{T}_{\mathrm{f}}^{\natural}$$

Consequently the condition of $\boldsymbol{\pi}^{\mathrm{H}}$ is given by

$$\|\mathbf{T}^{\natural}\| \leq \|\mathbf{T}_{\mathrm{s}}^{\natural}\| + \|\mathbf{T}_{\mathrm{f}}^{\natural}\|. \tag{4.1}$$

Since $\mathbf{T}_{\mathrm{f}}$ is bounded and its eigenvalues are bounded away from one, the second term in (4.1) is bounded as $\epsilon \to 0$.

The first term is another story. It is equal to $\|\mathbf{X}_{\mathrm{s}}(\mathbf{I}-\mathbf{B}_{\mathrm{s}})^{-1}\mathbf{Y}_{\mathrm{s}}^{\mathrm{T}}\|$. Now we have noted in the last section that $\mathbf{B}_{\mathrm{s}}$ is equal to $\mathbf{I}+O(\epsilon)$. It follows that

$$\|\mathbf{T}_{\mathrm{s}}^{\natural}\| \geq O(\epsilon^{-1}).$$

In other words,

> *the condition of $\boldsymbol{\pi}^{\mathrm{T}}$ increases at least in inverse proportion to the size of $\mathbf{E}$.*

This negative result is perhaps disappointing, but it accords with common sense. If the condition were bounded, we could find a value of $\|\mathbf{F}\|$ for which $\boldsymbol{\pi}^{\mathrm{T}}$ is satisfactorily accurate no matter what the value of ϵ. In particular if ϵ were less than this value of $\|\mathbf{F}\|$, we could set $\mathbf{E}$ to zero and still get an accurate steady state vector — which is obvious nonsense.

It should be stressed that the perturbations introduced by the slow transient are by no means arbitrary. A more detailed analysis shows the vectors $\overline{\boldsymbol{\pi}}_1^{\mathrm{T}}$ are quite stable, whereas the coupling coefficients v_i are sensitive to perturbations in $\mathbf{P}$. Again this accords with our intuition about NUMCs.

The numerical assessment of the condition of $\boldsymbol{\pi}^{\mathrm{T}}$ is not difficult. Most aggregation procedures require one to compute and approximation to the coupling matrix $\mathbf{C}$ and the coupling coefficients. From there it is a small step to compute an approximation to the matrix $\mathbf{B}_{\mathrm{s}}$ and estimate the norm of $(\mathbf{I}-\mathbf{B}_{\mathrm{s}})^{-1}$, which can be done by any of a number of well known techniques [7].

References

[1] W. L. Cao and W. J. Stewart (1985). "Iterative Aggregation/Disaggregation Techniques for Nearly Uncoupled Markov Chains." *Journal of the Association for Computing Machinery*, 32, 702–719.

[2] P. J. Courtois (1977). *Decomposability.* Academic Press, New York.

[3] P. J. Courtois and P. Semal (1984). "Error Bounds for the Analysis by Decomposition of Non-Negative Matrices." In G. Iazeolla, P. J. Courtois, and A. Hordijk, editors, *Mathematical Computer Performance and Reliability*, pages 287–302, North Holland. Elsevier.

[4] G. H. Golub and C. D. Meyer (1985). "Using the QR Factorization and Group Inversion to Compute, Differentiate, and Estimate the Sensitivity of Stationary Probabilities for Markov Chains." *SIAM Journal on Algebraic and Discrete Methods*, 7, 273–281.

[5] M. Haviv and L. van der Heyden (1984). "Perturbation Bounds for the Stationary Probabilities of a Finite Markov Chain." *Advances in Applied Probability*, 16, 804–818.

[6] Moshe Haviv (1987). "Aggregration/Disaggregation Methods for Computing the Stationary Distribution of a Markov Chain." *SIAM Journal on Numerical Analysis*, 24, 952–966.

[7] N. J. Higham (1987). "A Surevey of Condition Number Estimation for Triangular Matrices." *SIAM Review*, 29, 575–596.

[8] C. Meyer and G. W. Stewart (1988). "Derivatives and Perturbations of Eigenvectors." *SIAM Journal on Numerical Analysis*, 25, 679–691.

[9] P. J. Schweitzer (1968). "Perturbation Theory and Finite Markov Chains." *Journal of Applied Probability*, 5, 401–413.

[10] H. A. Simon and A. Ando (1961). "Aggregation of Variables in Dynamic Systems." *Econometrica*, 29, 111–138.

[11] G. W. Stewart (1984). "On the Structure of Nearly Uncoupled Markov Chains." In G. Iazeolla, P. J. Courtois, and A. Hordijk, editors, *Mathematical Computer Performance and Reliability*, pages 287–302, North Holland. Elsevier.