University of Maryland College Park

Institute for Advanced Computer Studies TR–94–77
Department of Computer Science TR–3306

On Markov Chains with Sluggish Transients*

G. W. Stewart†

June, 1994

ABSTRACT

In this note it is shown how to construct a Markov chain whose subdominant eigenvalue does not predict the decay of its transient.

*This report is available by anonymous ftp from `thales.cs.umd.edu` in the directory `pub/reports`.

†Department of Computer Science and Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742. This work was supported in part by the National Science Foundation under grant CCR 9115568.

On Markov Chains with Sluggish Transients

G. W. Stewart

ABSTRACT

In this note it is shown how to construct a Markov chain whose subdominant eigenvalue does not predict the decay of its transient.

1. Introduction

Let $\mathbf{P}$ be the transition matrix of an finite, irreducible, aperiodic Markov chain, and let $\mathbf{e}$ be the vector whose components all are one. Since the row sums of $\mathbf{P}$ are one, $\mathbf{Pe} = \mathbf{e}$; i.e., $\mathbf{e}$ is a right eigenvector of $\mathbf{P}$ corresponding to the eigenvalue one. From the theory of nonnegative matrices, this eigenvalue is simple and is strictly larger in magnitude than the remaining eigenvalues. Moreover, there is a unique, positive, left eigenvector $\boldsymbol{\pi}^{\mathrm{T}}$ satisfying $\boldsymbol{\pi}^{\mathrm{T}}\mathbf{P} = \boldsymbol{\pi}$ and $\boldsymbol{\pi}^{\mathrm{T}}\mathbf{e} = 1$. The vector $\boldsymbol{\pi}^{\mathrm{T}}$ is the steady-state vector for the chain whose transition matrix is $\mathbf{P}$.

If we set

$$\mathbf{T} = \mathbf{P} - \mathbf{e}\boldsymbol{\pi}^{\mathrm{T}},$$

then the eigenvalues of $\mathbf{P}$ and $\mathbf{T}$ are the same, except that $\mathbf{T}$ has an eigenvalue of zero corresponding to the eigenvalue one of $\mathbf{P}$. Now it is easily verified that

$$\mathbf{P}^k = \mathbf{e}\boldsymbol{\pi}^{\mathrm{T}} + \mathbf{T}^k.$$

Since the eigenvalues of $\mathbf{T}$ are less than one in magnitude, $\lim_{k\to\infty} \mathbf{T}^k = 0$, or equivalently

$$\lim_{k\to\infty} \mathbf{P}^k = \mathbf{e}\boldsymbol{\pi}^{\mathrm{T}}.$$

In other words, the powers of the transition matrix converge to a matrix whose rows are the steady-state vector.

These facts are well known. In this note we will be concerned with the rate of convergence of the powers of $\mathbf{P}$. The folklore says that convergence is proportional ρ^k, where ρ is the magnitude of the largest eigenvalue of the transient matrix $\mathbf{T}$. A justification for this rule of thumb is the equation

$$\lim_{k\to\infty} \|\mathbf{T}^k\|^{\frac{1}{k}} = \rho,$$

which holds for any consistent matrix norm. In plain words, if we average the rate of convergence over k iterations, then with increasing k the average approaches ρ.

Unfortunately, this asymptotic result does not tell us what happens in the short run. And even in the long run, it is possible for $\|\mathbf{T}^k\|/\rho^k$ to diverge to infinity. It might be conjectured that the transition matrices of Markov chains are so structured that pathological cases cannot occur. As it turns out, certain pathologies do seem to be excluded. But, as we shall see, a Markov chain can have a sluggish transient that falls behind the the decay of its subdominant eigenvalue.

Since our purpose is to produce counter examples, it is not necessary to analyze the problem in full generality. Consequently, in the next section we analyze the behavior of the powers of a 2×2 Jordan block. In §3 we show how to construct a Markov chain with a sluggish transient. The note concludes with some general observations. Throughout the paper $\|\mathbf{A}\|$ will denote the spectral norm of $\mathbf{A}$; that is, the square root of the largest eigenvalue of $\mathbf{A}^{\mathrm{T}}\mathbf{A}$.

2. Two by Two Jordan Blocks

The advantage of considering a 2×2 Jordan block is that it exhibits pathologies found in larger blocks yet is simple enough to analyze. We will write our block, somewhat unusually, in the form

$$\mathbf{J} = \rho \begin{pmatrix} 1 & \sigma \\ 0 & 1 \end{pmatrix},$$

where $0 \le \rho < 1$ and $\sigma > 0$. It is easy to see that

$$\mathbf{J}^k = \rho^k \begin{pmatrix} 1 & k\sigma \\ 0 & 1 \end{pmatrix}. \tag{2.1}$$

The first conclusion we can draw from this expression is that for large k

$$\|\mathbf{J}^k\| \cong k\rho^k\sigma. \tag{2.2}$$

Thus the ratio $\|\mathbf{J}^k\|/\rho^k$ approaches infinity, a possibility alluded to in the introduction.

From (2.2) it follows that for k large enough

$$\frac{\|\mathbf{J}^{k+1}\|}{\|\mathbf{J}^k\|} \cong \rho\left(1 + \frac{1}{k}\right). \tag{2.3}$$

Thus the *local rate of convergence* is $\rho\left(1 + \frac{1}{k}\right)$. At first glance, it might seem that the factor $1 + \frac{1}{k}$, which approaches one, represents an insignificant perturbation of ρ. However, when ρ is near one, it is important in two respects.

First, the analysis is asymptotic, and for the factor to represent a decrease we must have

$$k > \frac{\rho}{1-\rho}. \tag{2.4}$$

For example if $\rho = 0.99$, then k will have to be greater than 99. We will return to this point later.

Second, even when k satisfies (2.4), the small change introduced by the factor $1 + \frac{1}{k}$ has an inordinate effect on the number of iteration required to reduce $\|\mathbf{J}^k\|$. To see why, note that if a process is converging to zero as α^k, then for $\beta > 1$

$$N_\beta = -\frac{1}{\log_\beta \alpha} \tag{2.5}$$

is the number of iterations required for a reduction by a factor of β^{-1}. Let us call N_β the *β-fold reduction time.* Now if $\rho = 1 - \epsilon$, then from (2.3) and (2.5) [with generous use of the approximation $\ln(1+\eta) = \eta + O(\eta^2)$] we find that the local e-fold reduction time is

$$N_e \cong \frac{k}{k\epsilon - 1}.$$

Here is a table of the local e-fold reduction time for $\epsilon = 0.01$.

k	N_e
200	200
300	150
400	133
500	125

In this case, the local e-fold reduction time converges very slowly — in fact harmonically — to its asymptotic value of 100.

Up to this point we have considered the asymptotic behavior of $\|\mathbf{J}^k\|$ with k restricted to satisfy (2.4). The reason for the restriction is that for small k the element $k\rho^k\sigma$ of $\mathbf{J}^k$ can increase. But whether this increase induces a corresponding increase in $\|\mathbf{J}^k\|$ depends on the size of σ. If σ is large enough, $\|\mathbf{J}^k\|$ will show an initial increase and then turn around and begins to decrease at the local rate given by (2.3). On the other hand, when σ is small, $\|\mathbf{J}\|$ decreases monotonically, perhaps at first at a rate near that of ρ^k, but asymptotically at the local rate. It is worth while to determine the critical value of σ that separates the two regimes.

It follows directly from the definition of the spectral norm that

$$\left\| \begin{pmatrix} 1 & \gamma \\ 0 & 1 \end{pmatrix} \right\|^2 = 1 + \gamma \left(\frac{\gamma + \sqrt{4+\gamma^2}}{2} \right) = 1 + \gamma + O(\gamma^2). \tag{2.6}$$

Figure 2.1: Modes of decay: $\rho = 0.99$, $\sigma = 0.035, 0.02, 0.005, 0$

Moreover, if $\rho = 1 - \epsilon$, where ϵ is small, then for small k the square of ρ^k is $\rho^{2k} \cong 1 - 2k\epsilon$. Setting $\gamma = k\sigma$ in (2.6), we have from (2.1) that

$$\|\mathbf{J}^k\|^2 \cong (1 - 2k\epsilon)(1 + k\sigma) \cong 1 - 2k\epsilon + k\sigma.$$

It follows that if $-2k\epsilon + k\sigma = 0$ or

$$\sigma = 2\epsilon = 2(1 - \rho),$$

then initially the norms $\|\mathbf{J}\|^k$ are approximately stationary.

Figure 2.1 plots $\log_{10}(\|\mathbf{J}\|^k)$ for $\rho = 0.99$ and for σ supercritical (0.035), critical (0.02), subcritical (0.005), and pure (0.00). Note the very different behaviors.

3. A Sluggish Markov Chain

In this section, we will show how to imbed a 2×2 Jordan block in a 3×3 Markov chain. The algorithm goes as follows.

1. Choose ρ and σ and generate the matrix
$$\mathbf{D} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1-\rho & -\rho\sigma \\ 0 & 0 & 1-\rho \end{pmatrix}.$$
2. Generate a random positive vector $\boldsymbol{\pi}^{\mathrm{T}}$, normalized so that $\boldsymbol{\pi}^{\mathrm{T}}\mathbf{e} = 1$.
3. Generate a random 3×2 matrix $\mathbf{U}$ such that $\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{I}$ and $\mathbf{U}^{\mathrm{T}}\mathbf{e} = 0$.
4. Set $\mathbf{W}^{\mathrm{T}} = (\boldsymbol{\pi} \;\; \mathbf{U})$.
5. Calculate $\mathbf{Q} = \mathbf{W}^{-1}\mathbf{D}\mathbf{W}$.
6. If $\mathbf{Q}$ does not have positive diagonal elements and negative off-diagonal elements, go to step 2.
7. Set $\mathbf{P} = \mathbf{I} - \mathbf{Q}$.

Here are some comments on the algorithm. For details consult the matlab implementation in the appendix to this paper.

To generate $\mathbf{U}$, a random normal matrix 3×2 is generated and its columns are orthogonalized against $\mathbf{e}$ and each other and then normalized. The QR decomposition can be used to accomplish the orthonormalization.

The first column of $\mathbf{W}^{-1}$ is $\mathbf{e}$. For by the definition of the inverse, it must satisfy $\boldsymbol{\pi}^{\mathrm{T}}\mathbf{w}_1^{(-1)} = 1$ and $\mathbf{U}^{\mathrm{T}}\mathbf{w}_1^{(-1)} = 0$. By the construction of $\boldsymbol{\pi}$ and $\mathbf{U}$, $\mathbf{e}$ is the only vector satisfying these equations.

Since $\mathbf{W}\mathbf{Q}\mathbf{W}^{-1} = \mathbf{D}$, the first row $\boldsymbol{\pi}^{\mathrm{T}}$ of $\mathbf{W}$ is a left eigenvector with eigenvalue zero and the first column $\mathbf{e}$ of $\mathbf{W}^{-1}$ is a right eigenvector with eigenvalue zero.

It is necessary to check the sign pattern of $\mathbf{Q}$ since the procedure is not guaranteed to produce a generator. Strictly speaking, it is also necessary to check that the diagonal elements of $\mathbf{Q}$ are less than one, so that $\mathbf{P} = \mathbf{I} - \mathbf{Q}$ is stochastic. However, with the small ρ and σ we use here, that is not an issue.

Figure 3.1: $\log_{10} \|\mathbf{P}^k - \mathbf{e}\boldsymbol{\pi}^{\mathrm{T}}\|$ for $\rho = 0.99$, $\sigma = 0.01$

In a typical run with $\rho = 0.99$ and $\sigma = 0.01$, the procedure yielded the following matrix:

$$\mathbf{P} = \begin{pmatrix} 0.99425020008452 & 0.00100182554309 & 0.00474797437240 \\ 0.01148687116416 & 0.98848878048060 & 0.00002434835524 \\ 0.00040017460421 & 0.00233880596091 & 0.99726101943488 \end{pmatrix}. \tag{3.1}$$

Figure 3.1 plots the common logarithm of the residual norm $\|\mathbf{P}^k - \mathbf{e}\boldsymbol{\pi}^{\mathrm{T}}\|$. The lower line is the plot of the residual norm for $\sigma = 0$, which effectively represents a pure decay. The sluggish chain quickly settles into its local convergence mode, and the two lines diverge.

Figure 3.2 plots the local 10-fold reduction time. After an initial increase it begins to converge slowly to its asymptotic value of about 130.

Figure 3.2: Local 10-fold reduction time for $\rho = 0.99$, $\sigma = 0.01$

4. Discussion

The above example shows that sluggish Markov chains exist and that their behavior differs markedly from the behavior predicted by the subdominant eigenvalue. Although we have presented only one example, it is typical of the other examples generated by our algorithm. The chief difference from example to example is in the behavior during the initial iterations, where the hump in the local 10-fold reduction time varies in size and can even be absent. In any event, the interested reader can try out the matlab code in the appendix.

It might be objected that the structure of the matrix (3.1) is rather special: it is a perturbation of the identity matrix. However, such are the matrices that approximate the slow transient of a nearly completely decomposable Markov chain

[1]. Otherwise, the matrix itself gives no hint of its sluggish transient.

The value $\sigma = 0.01$ used in the example is decidedly subcritical. In fact, for $\sigma = 0.012$ (still subcritical) our algorithm was unable to generate a legitimate transition matrix, even after five thousand repetitions. This suggests that Markov of chains of critical or supercritical sluggishness may not exist. The conjecture is worth further study.

References

[1] P.-J. Courtois. *Decomposability.* Academic Press, New York, 1977.

Appendix: Matlab Code

```
e = [1;1;1];
d = [0, 0, 0; 0, 1-rho, -rho*sigma; 0, 0, 1-rho];
i = 0;
while(1==1)
  i=i+1
  [u,r] = qr([e,randn(3,2)]);
  pi  = rand(1,3);
  pi = pi/(pi*e);
  w = [pi; u(:,2:3)'];
  wi = inv(w);
  q=wi*d*w;
  if (q(1,1)>0 & q(2,2)>0 & q(3,3)>0 & q(1,2)<0 & q(1,3)<0 ...
       & q(2,1)<0 & q(2,3)<0 & q(3,1)<0 & q(3,2)<0),
     break;
  end
end
p = eye(3)-q
clear x;
clear y;
for i=1:500
   x(i) = log10(norm(p^i-e*pi));
end
for i=1:499
   y(i) = 1/(x(i)-x(i+1));
end
plot(x)
```