Markov Chains with Memory, Tensor Formulation, and the Dynamics of Power Iteration

Sheng-Jhih Wu[a], Moody T. Chu[b,1]

[a]*Center for Advanced Statistics and Econometrics Research, School of Mathematical Sciences, Soochow University, Suzhou, China.*
[b]*Department of Mathematics, North Carolina State University, Raleigh, NC 27695-8205.*

Abstract

A Markov chain with memory is no different from the conventional Markov chain on the product state space. Such a Markovianization, however, increases the dimensionality exponentially. Instead, Markov chain with memory can naturally be represented as a tensor, whence the transitions of the state distribution and the memory distribution can be characterized by specially defined tensor products. In this context, the progression of a Markov chain can be interpreted as variants of power-like iterations moving toward the limiting probability distributions. What is not clear is the makeup of the "second dominant eigenvalue" that affects the convergence rate of the iteration, if the method converges at all. Casting the power method as a fixed-point iteration, this paper examines the local behavior of the nonlinear map and identifies the cause of convergence or divergence. As an application, it is found that there exists an open set of irreducible and aperiodic transition probability tensors where the Z-eigenvector type of power iterates fail to converge.

Keywords: Markov chain with memory, transition probability tensor, stationary distribution, power method, rate of convergence, second dominant eigenvalue
2000 MSC: 15A18, 15A51, 15A69, 60J99

1. Introduction

A Markov chain is a stochastic process $\{X_t\}_{t=0}^{\infty}$ over a finite state space S, where the conditional probability distribution of future states in the process depends upon the present or past states. The classical "Markov property" specifies that the probability of transition to the next state depends only on the probability of the current state. That is, among the states $s_i \in S$, the model assumes that

$$\Pr(X_{t+1} = s_{t+1} | X_t = s_t, \ldots, X_2 = s_2, X_1 = s_1) = \Pr(X_{t+1} = s | X_t = s_t).$$

For simplicity, identify the states as $S = \{1, 2, \ldots, n\}$ and assume that the chain is time homogeneous. Then a transition probability matrix $P = [p_{ij}]$ defined by

$$p_{ij} := \Pr(X_{t+1} = i \mid X_t = j) \tag{1}$$

is independent of t and column stochastic. The above process is, generally characterized as memoryless[2], is a well studied subject.

There are situations where the data sequence does depend on past values. As can be expected, the additional history of memory often has the advantage of offering a more precise predictive value. By bringing more memory into the random process, we can build a higher order Markov model. Interesting applications

Email addresses: szwu@suda.edu.cn (Sheng-Jhih Wu), chu@math.ncsu.edu (Moody T. Chu)
[1]This research was supported in part by the National Science Foundation under grant DMS-1316779.
[2]Strictly speaking, it should be called a chain with memory 1 based on the definition (2).

include packet video traffic in larger buffers [25], finance risk management [17, 18, 27], wind turbine design [24], alignment of DNA sequences or long-range correlated dynamic systems [19, 20, 28], growth of polymer chains [6, 12], cloud data mining [9, 26], and many others [16]. A Markov chain with memory m is a process satisfying

$$\Pr(X_{t+1} = s_{t+1}|X_t = s_t, \ldots, X_1 = s_1) = \Pr(X_{t+1} = s_{t+1}|X_t = s_t, \ldots, X_{t-m+1} = s_{t-m+1}) \quad (2)$$

for all $t \geq m$. By defining

$$Y_t = (X_t, X_{t-1}, ..., X_{t-m+1}) \quad (3)$$

and by taking the ordered m-tuples of X values as its product state space, it is easy to see that the chain $\{Y_t\}$ with suitable starting values satisfies the Markov property. In principle, upon exploiting the underlying structure, the transition process can be analyzed with the classical theory for memoryless Markov chain. Note, however, that the size of the aggregated chain, also known as the Markovianization, is considerably larger — of dimension n^{m-1}. Though mathematically equivalent, basic tasks such as bookkeeping multi-states and other associated operations will be fairly tedious[3].

In recent years higher-order tensor analysis have become an effective way to address high-throughput and multi-dimensional data by different disciplines. Markov chain with memory fits naturally such a tensor formulation. Assuming again time homogeneity, a Markov chain with memory $m-1$ can be conveniently represented via the order-m tensor $\mathcal{P} = [p_{i_1 i_2 \ldots i_m}]$ defined by

$$p_{i_1 i_2 \ldots i_m} := \Pr(X_{t+1} = i_1|X_t = i_2, \ldots, X_{t-m+2} = i_m), \quad (4)$$

where $\mathcal{P}$ is called a transition probability tensor. Necessarily we have the properties that $0 \leq p_{i_1 i_2 \ldots i_m} \leq 1$ and that

$$\sum_{i_1=1}^{n} p_{i_1 i_2 \ldots i_m} = 1 \quad (5)$$

for every fixed $(m-1)$-tuple $(i_2, \ldots, i_m)$. What is most interesting is that the transitions among the states as well as the history of memory can be characterized by specially defined tensor products. Our goal in this paper is to recast such a process under the tensor formulation. In particular, we are interested in understanding the dynamics of the transition to the stationary distribution and the associated 2-phase power iteration scheme in the context of tensor operations.

While some classical results in matrix theory can be extended naturally to tensors, there are cases where the nonlinearity of tensors makes the generalization far more cumbersome. The notion of eigenvalue is one such incident. Depending on the applications, there are several ways to mull over how an eigenvalue of a tensor should be defined [3, 13, 14, 23]. Markov chain with memory and the associated transition probability tensor can serve as a practical model for exploring the following two notions of eigenvalues and their implications:

1. The classical concept of eigenvalues when characterizing the evolution of the joint probability mass functions.
2. The notion of Z-eigenvalue [23] when approximating the evolution of the state probability distribution.

In this context, we study the role of the "second dominant eigenvalue" in such a dynamics of a Markov chain with memory. We also intend to address some practical issues arisen from a recent discussion in [12] which proposes to short cut the computation of the stationary state distribution by approximating the stationary joint probability mass function. These issues include whether the assumption used in proposing the Z-eigenvector computation is statistically justifiable and the anatomy of the true cause that affects the rate of convergence. The tool we are about to develop gives some insight into this limiting behavior. It is possible to generalize our framework to other types of eigenvalues for tensors, e.g., the so called H-eigenvalues [21].

[3] See an example of vectorizing a Markov chain with memory 2 in Section 3.

For demonstration, we choose to concentrate only on the application to the transition probability tensors in this presentation.

This paper is organized as follows. We begin in Section 2 with some basic properties of transition probability tensors. We review two types of dynamics necessarily involved in a Markov chain with memory, each of which entails a particular kind of tensor product. The evolution of the joint probability mass function itself follows a scheme similar to the conventional power method, whereas finding the stationary probability distributions of the state vector requires a 2-phase iteration. In Sections 3, we argue that an appropriate rearrangement of the transition probability tensor reveals the proper cause of convergence for this classical type of evolution. In Section 4 we address some concerns arisen from the recent notion of approximating the stationary distribution by the dominant Z-eigenvector. We identify the true makeup of the "second" dominant eigenvalue in the tensor setting. Most importantly, we show that the convergence of this shortcut type of power method proposed in [12] is not always guaranteed by counter examples. Included in the Appendix is the local analysis in a similar spirit for matrices, which probably offers an alternative explanation of convergence for the classical power method .

2. Dynamics involved in a Markov chain with memory

Needless to say, a critical ingredient in the Markov process with memory $m-1$ is the joint probability mass function of state variables $X_t, \dots X_{t-m+2}$ over S at time t, denoted as

$$\Pi_{t,t-1,\dots,t-m+2} = [\pi^{(t)}_{i_2\dots i_m}], \tag{6}$$

where

$$\pi^{(t)}_{i_2\dots i_m} := \Pr(X_t = i_2, \dots, X_{t-m+2} = i_m).$$

Note that $\Pi_{t,t-1,\dots,t-m+2}$ is an order-$(m-1)$ tensor whose entries are nonnegative and satisfy the identity

$$\sum_{i_2,\dots,i_m=1}^{n} \pi^{(t)}_{i_2\dots i_m} = 1. \tag{7}$$

With appropriate ordering, the joint probability mass function $\Pi_{t,t-1,\dots,t-m+2}$ is simply the typical state distribution of the Markovianized $(m-1)$-tuple $(X_t, \dots, X_{t-m+2})$.

The definition of conditional probability naturally dictates that in a Markov chain with memory the probability distribution of the next state X_{t+1} based on memory $X_t, \dots X_{t-m+2}$ should be calculated in the following way which naturally defines a kind of tensor product, denoted by the symbol $\circledast_1$.

Lemma 2.1. *Let the column vector $\mathbf{x}^{(t+1)}$ denote the probability distribution of the variable X_{t+1} over the state space S. Then*

$$\mathbf{x}^{(t+1)} = \mathcal{P} \circledast_1 \Pi_{t,t-1,\dots,t-m+2} := [\langle p_{i_1,:}, \Pi_{t,t-1,\dots,t-m+2}\rangle] \in \mathbb{R}^n, \tag{8}$$

where $p_{i_1,:}$ denotes the i_1-th facet in the 1st direction of $\mathcal{P}$ and $\langle\cdot,\cdot\rangle$ is the Frobenius inner product generalized to multi-dimensional arrays.

We claim that entries of the next joint probability mass function $\Pi_{t+1,t,\dots,t-m+3} = [\pi^{(t+1)}_{i_1\dots i_{m-1}}]$ of state variables $X_{t+1}, \dots X_{t-m+3}$ are given as follows, which defines another kind of tensor product. Once this calculation is complete, it allows the chain to continue evolving.

Lemma 2.2. *Given the joint probability mass function $\Pi_{t,t-1,\dots,t-m+2}$, if X_{t+1} is obtained from the Markov chain with memory $X_t, \dots X_{t-m+2}$, then the entries of $\Pi_{t+1,t,\dots,t-m+3}$ are given by*

$$\pi^{(t+1)}_{i_1\dots i_{m-1}} = \sum_{i_m=1}^{n} p_{i_1 i_2\dots i_m} \pi^{(t)}_{i_2\dots i_m}, \quad i_1, \dots, i_{m-1} = 1, \dots, n. \tag{9}$$

Proof. Observe that

$$\Pi_{t+1,t,\ldots,t-m+3} = \sum_{i_m=1}^{n} \Pr(X_{t+1}, X_t, \ldots, X_{t-m+3}, X_{t-m+2} = i_m)$$

$$= \sum_{i_m=1}^{n} \Pr(X_{t+1}|X_t, \ldots, X_{t-m+3}, X_{t-m+2} = i_m) \Pr(X_t, \ldots X_{t-m+3}, X_{t-m+2} = i_m). \quad (10)$$

The expression (9) is simply the case when $X_{t+1} = i_1$, $X_t = i_2$, …, $X_{i-m+3} = i_{m-1}$. □

Note that the operation required in (9) is different from the usual mode-m tensor product defined in the literature [10]. For convenience, denote this process for transiting the joint probability mass function by the symbol

$$\Pi_{t+1,t,\ldots,t-m+3} = \mathcal{P} \boxdot \Pi_{t,t-1,\ldots,t-m+2}. \quad (11)$$

As an demonstration of this operator $\boxdot$, we can rewrite the relation (9) for a Markov chain with memory 2 in the matrix form as

$$\Pi_{t+1,t} = [\mathcal{P}(:,1,:)\Pi_{t,t-1}(1,:)^\top, \ldots, \mathcal{P}(:,n,:)\Pi_{t,t-1}(n,:)^\top] \quad (12)$$

where $\mathcal{P}(:,j,:) \in \mathbb{R}^{n\times n}$ and $\Pi_{t,t-1}(j,:)$ stand for the j-th facet in the 2nd direction of $\mathcal{P}$ and the j-th row of $\Pi_{t,t-1}$, respectively. The summation over the index i_3 is included in the matrix-to-vector multiplication. In contrast, for a Markov chain with memory 1 (the so called memoryless case), the products $\circledast_1$ and $\boxdot$ are identical and $\Pi_{t+1} = \mathbf{x}^{(t+1)}$.

Our understanding thus far is derived from elementary probability theory. It implies that, for a Markov chain with memory greater than 1, there are two dynamics involved in the evolving process. One is the multiplication in the form of (8) for the transition of state distribution. The other is the multiplication in the form of (11) for the transition of the joint probability mass function. To characterize the limiting behavior of the state distribution $\{\mathbf{x}^{(t)}\}$, it is necessary to understand the limiting behavior of the joint probability mass function $\{\Pi_{t,t-1,\ldots,t-m+2}\}$, and vice versa. Although the approach is standard, we find little discussion in the literature that analyzes these processes directly in the setting of tensors [8]. In the following, we demonstrate that the tensor notation is convenient for arguing the dynamics behavior of Markov chains.

3. Power iteration for joint probability mass function

In this section, we elaborate further on the limiting behavior of the joint probability mass function $\{\Pi_{t,t-1,\ldots,t-m+2}\}$. It will be instructive if we first consider the Markov chain with memory 2, which gives rise to an order-3 probability transition tensor. The discussion can readily be extended to the general case.

For a Markov chain with memory 2 to move forward, we have to process two sequences of probability distributions hand by hand. At the time step t, we have a distribution $\Pi_{t,t-1} = [\pi_{i_2 i_3}]$ for memory (X_t, X_{t-1}) over $S \times S$. Then the next state X_{t+1} over S based on this memory has a distribution $\mathbf{x}^{(t+1)}$ defined by the tensor product

$$\mathbf{x}^{(t+1)} = \mathcal{P} \circledast_1 \Pi_{t,t-1}. \quad (13)$$

In the meantime, the memory distribution is also evolved into

$$\Pi_{t+1,t} = \mathcal{P} \boxdot \Pi_{t,t-1}. \quad (14)$$

We use Figure 1 to depict the interaction and the involvement of the two distinct tensor multiplications. First, given memory (X_t, X_{t-1}), the probability of being moved into state i_1 at step $t+1$ is given by the double sum

$$x_{i_i}^{(t+1)} = \sum_{i_2,i_3=1}^{n} p_{i_1 i_2 i_3} \pi_{i_2 i_3}^{(t)}. \quad (15)$$

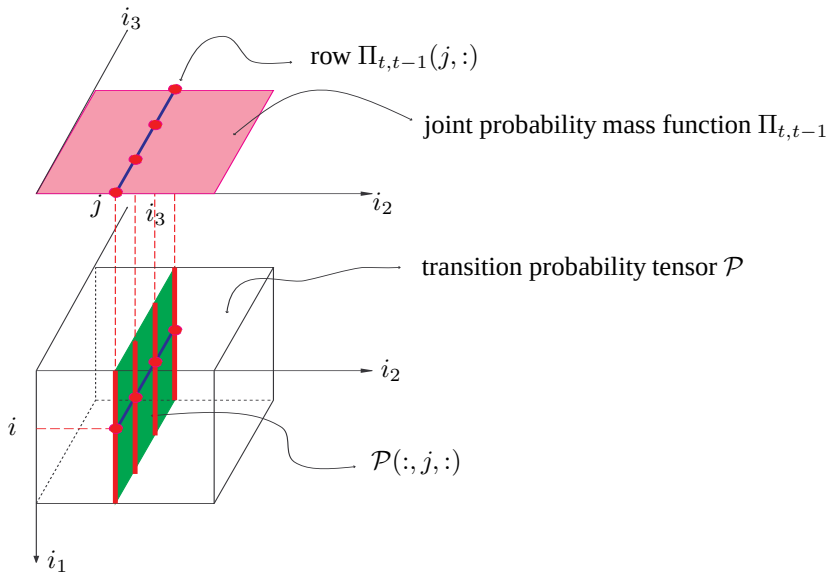

Figure 1: Update of joint probability mass function $\Pi_{t+1,t}$ from $\Pi_{t,t-1}$.

Plotting the matrix $\Pi_{t,t-1}$ as the separated horizontal (magenta) plane above the cubic box representing the transition probability tensor $\mathcal{P}$, the mechanism for computing the distribution $\mathbf{x}^{(t+1)}$ for the state variable X_{t+1} can be thought of as taking the Frobenius inner product of the matrix $\Pi_{t,t-1}$ with each (horizontal) cross section of the tensor $\mathcal{P}$ in the 1st direction. Similarly, the probability of having memory $X_{t+1} = i_1, X_t = i_2$ at the step $t+1$ is given by

$$\pi^{(t+1)}_{i_1 i_2} = \sum_{i_3=1}^{n} p_{i_1 i_2 i_3} \pi^{(t)}_{i_2 i_3}, \tag{16}$$

which is the inner product of the i_2-th (blue) row of $\Pi_{t,t-1}$ with each (blue) row of the vertical (green) cross section of the tensor $\mathcal{P}$ in the 2nd direction.

It is easy to observe the relationship that

$$x^{(t+1)}_{i_1} = \sum_{i_2=1}^{n} \pi^{(t+1)}_{i_1 i_2}. \tag{17}$$

So the limiting behavior of the sequence $\{\mathbf{x}^{(t+1)}\}$ follows from that of the sequence $\{\Pi_{t+1,t}\}$. Indeed, such a "row sum" relationship holds in general, which we state as follows. So, for limiting behavior, it suffices to first understand the dynamics of the iteration (11).

Lemma 3.1. *For a Markov chain with memory $m-1$, the state probability distribution is related to the joint probability mass distribution via*

$$x^{(t+1)}_{i_1} = \sum_{i_2,\ldots,i_{m-1}=1}^{n} \pi^{(t+1)}_{i_1 i_2 \ldots i_{m-1}}. \tag{18}$$

Proof. By definition, the order-$(m-1)$ tensor $\Pi_{t+1,t,\ldots,t-m+3}$ stands for the joint probability mass distribution of $(X_{t+1}, X_t, \ldots, X_{t-m+3})$, the relationship (18) is simply the marginal distribution of the variable X_{t+1}. □

For the case of memory 2, we may rewrite this updating mechanism (16) for each fixed i_2 in columns via a sequence of matrix-to-vector multiplications

$$\Pi_{t+1,t}(:, i_2) = \mathcal{P}(:, i_2, :)\Pi_{t,t-1}(i_2, :)^{\top}, \quad i_2 = 1, \ldots n. \tag{19}$$

At first glance, such a scheme seems to be the familiar power method[4] applied to the matrix $\mathcal{P}(:, i_2, :)$. The subtle difference is at the "transpose" involved in (19). In order to repeat the "matrix-to-column" multiplication for a fixed i_2, we must know every other "rows" of $\Pi_{t+1,t}$ which is not available until the entire sequence of multiplications in the form of (19) has been completed. In other words, only by treating the operation $\boxdot$ in (14) as a whole "tensor-to-tensor" multiplication, we may treat the iteration as a power method. This is not an ordinary power iteration.

We certainly can recast the power-like iteration (14) in the usual context of matrix operations as follows. This should manifests the complication of the "second dominant eigenvalue" of the order-3 tensor $\mathcal{P}$. See also [8] for a similar discussion. Let $\mathbf{vec}(M)$ denote the conventional vectorization of the matrix M by stacking its columns into a single column vector. Let C be the $n^2 \times n^2$ permutation matrix that does the swapping of indices

$$(j-1)n + i \to (i-1)n + j, \quad 1 \le i, j \le n.$$

Also, let B denote the $n^2 \times n^2$ block diagonal matrix whose i_2-th diagonal block is precisely the $n \times n$ matrix $\mathcal{P}(:, i_2, :)$. Then the operation $\boxdot$ is equivalent to the matrix-to-vector multiplication

$$\mathbf{vec}(\Pi_{t+1,t}) = BC\mathbf{vec}(\Pi_{t,t-1}). \tag{20}$$

The scheme (20) is exactly the power method applied to the $n^2 \times n^2$ matrix $\mathcal{A} := BC$. It is not difficult to check that $\mathcal{A}$ has the block structure

$$\mathcal{A} = \left[\begin{array}{cccc|cccc|c|cccc} \mathcal{P}(:,1,1) & 0 & \dots & 0 & \mathcal{P}(:,1,2) & 0 & \dots & 0 & \dots & \mathcal{P}(:,1,n) & 0 & \dots & 0 \\ 0 & \mathcal{P}(:,2,1) & & 0 & 0 & \mathcal{P}(:,2,2) & & 0 & \dots & 0 & \mathcal{P}(:,2,n) & & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \mathcal{P}(:,n,1) & 0 & 0 & \dots & \mathcal{P}(:,n,2) & \dots & 0 & 0 & \dots & \mathcal{P}(:,n,n) \end{array}\right],$$

where each $\mathcal{P}(:, i, j)$ is a column vector in $\mathbb{R}^n$. By (5), $\mathcal{A}$ is itself column stochastic. By (7), $\mathbf{vec}(\Pi_{t,t-1})$ is itself a distribution vector. This is one way to "unfold" an order-3 transition probability tensor $\mathcal{P}$ into a column stochastic matrix $\mathcal{A}$. From this point on, the following results follow from what we already know about the conventional power method applied to the matrix $\mathcal{A}$.

Lemma 3.2. *Suppose that $\mathcal{P}$ is the transition probability tensor of a Markov chain with memory 2. Assume that the Perron root $\lambda_1(\mathcal{A}) = 1$ of the corresponding $\mathcal{A}$ is simple. Then, starting with any generic initial memory distribution $\Pi_{0,-1}$, the following statements hold.*

1. *The convergence of the joint probability mass functions generated by (14) is guaranteed.*
2. *The limit point $\widetilde{\Pi}$ of joint probability mass functions is the de-vectorization of the normalized dominant eigenvector of $\mathcal{A}$ under the 1-norm.*
3. *The stationary distribution $\widetilde{\mathbf{x}}$ of the states under this Markov chain (13) with memory 2 exists and is the row sum of $\widetilde{\Pi}$.*
4. *The rate of convergence is the modulus of the second dominant eigenvalue of the matrix $\mathcal{A}$.*

Though we shall not carry out the "unfolding" explicitly, the above argument is generalizable to Markov chains with memory higher than 2. One quick way to look at this situation is to regard the Markov process of $\mathcal{P}$ acting on the joint probability mass function via the multiplication $\boxdot$ defined by (11) as a linear map from the space $\mathbb{T}^{m-1}$ of order-$(m-1)$ tensors to $\mathbb{T}^{m-1}$ itself. Any finite dimensional linear relationship can always be expressed in terms of a matrix-to-vector multiplication. Therefore, the convergence of the sequence $\{\Pi_{t,t-1,\dots,t-m+2}\}$ generated by the iteration (11) can be guaranteed. The tedious work is to construct the corresponding column stochastic matrix $\mathcal{A}$, once we specify how the tensor $\Pi_{t,t-1,\dots,t-m+2}$ is to be flattened into a vector. The latter depends on how the multi-dimensional states $X_t, \dots, X_{t-m+2}$ are to be ordered. When all the details are done, then the "second dominant eigenvalue" of the resulting matrix $\mathcal{A}$ determines the rate of convergence.

[4] For a quick overview of the power method and its convergence, see the Appendix in Section 6.

Thus far, we have argued that the Markov evolution for the joint probability mass function naturally induces a power-like iteration under the multiplication $\boxdot$. The corresponding map $\mathcal{P} : \mathbb{T}^{m-1} \to \mathbb{T}^{m-1}$ is a linear transformation. In this context, the notion of eigenvalue for the tensor $\mathcal{P}$ should be defined in exactly the same way as we usually do for square matrices, barring the peculiar operation $\boxdot$ for multiplication. Such an approach is not always the one adopted in the literature. For instance, we ponder upon the so called Z-eigenvector and compare it with the dominant eigenvector under $\boxdot$ in the next section.

4. Power iteration for Z-eigenvector computation

The relationship (8) characterizes the dynamics of the probability distribution $\mathbf{x}^{(t+1)}$ in terms of $\Pi_{t,t-1,\ldots,t-m+2}$ which itself evolves according to (11). This is by far the most formal way to describe the actual evolution of the state distribution for a Markov chain with memory. By continuity, the stationary distribution of the states satisfies the same relationship (8) with the limiting joint probability function of (11). For the latter, we have postulated its existence through the standard argument for the power method in the preceding section.

Recently it has been proposed to circumvent the 2-phase evolution process by assuming directly that a limiting joint probability distribution of the high-order Markov chain is the Kronecker product of its limiting probability distribution [12]. The rationale is that if the sequence $\{\mathbf{x}^{(t)}\}$ has ever reached a stationary distribution $\widetilde{\mathbf{x}}$ over S, then it seems reasonable to assume that the limiting joint probability mass function be of the form

$$\lim_{t\to\infty} \Pi_{t,t-1,\ldots,t-m+2} = \underbrace{\widetilde{\mathbf{x}} \otimes \widetilde{\mathbf{x}} \otimes \ldots \otimes \widetilde{\mathbf{x}}}_{m-1 \text{ times}}. \tag{21}$$

Under this assumption, it is deduced from (8) that the stationary distribution $\widetilde{\mathbf{x}}$ should satisfy the equation

$$\mathcal{P} \circledast_1 \mathbf{z} \otimes \mathbf{z} \otimes \ldots \otimes \mathbf{z} = \mathbf{z} \tag{22}$$

which is conveniently abbreviated as

$$\mathcal{P}\mathbf{z}^{m-1} = \mathbf{z} \tag{23}$$

in the literature. The solution to (23) is called the Z-eigenvector associated with, in this case, the unit Z-eigenvalue of $\mathcal{P}$ [2, 13, 14, 23]. It can be shown that a solution to (23) does exist and that entries of any such a solution are all positive, if $\mathcal{P}$ is irreducible [12, Theorem 2.2]. Under some additional conditions on $\mathcal{P}$, it even can be shown that the solution is unique [4, 12].

In addition to Z-eigenvalues, there are other ways to define eigenvalues for a given tensor [13, 14, 23]. Accordingly, a variety of methods has been proposed for computing eigenpairs of a tensor [5, 12, 15, 21, 30, 31]. Perhaps the simplest means for finding the Z-eigenvector $\widetilde{\mathbf{z}}$ in (23) is an iterative scheme[5] of the form

$$\mathbf{z}_{k+1} := \mathcal{P}\mathbf{z}_k^{m-1}, \tag{24}$$

where the starting point $\mathbf{z}_0$ is an arbitrary probability vector. Note that each $\mathbf{z}_{k+1}$ remains to be a probability vector under exact arithmetic[6]. Under some mild conditions, the sequence $\{\mathbf{z}_k\}$ does converge linearly to a solution of (23).

While studying the nonlinear equation (23) and the dynamics of power-like iteration (24) is of mathematical interest in its own right, we want to point out that there are serious issues associated with the assumption (21) for Markov chains with memory.

[5] Though $\mathbf{z}_k$ of remains to be a probability vector, it does not have the same meaning as $\mathbf{x}^{(t)}$ which represents the distribution of the random variable X_t at step t. We thus use different notations.

[6] Z-eigenvectors are not scaling invariant for a general tensor. So care must be taken when performing the normalization which is an essential part of a power method. For our applications, all iterates are automatically of unit length in 1-norm, so this normalization is not needed in exact arithmetic.

- Such an assumption inadvertently implies that the limiting distribution of memory is a symmetric tensor and is of tensor rank one, which by our numerical experiments with (11) is not the case in general. A consequential fallout is that the stationary distribution $\widetilde{\mathbf{x}}$ from the real Markov chain (8) does not satisfy (23) at all.
- One might think of $\widetilde{\mathbf{z}}$ satisfying (23) as a certain kind of approximation to the true stationary distribution $\widetilde{\mathbf{x}}$ of the Markov chain with memory. Still, in our numerical experiments, we find that $\widetilde{\mathbf{z}}^{m-1}$ is not the best rank-1 tensor approximation to the limiting distribution $\widetilde{\Pi}$.
- Unlike the "almost sure" convergence of power iteration (14), (20), or even the general (11) described in the preceding section for the joint probability mass function, we find that the set of transition probability tensors for which the power-like method (24) fails to converge is nonempty and, more importantly, has a nonzero measure in the ambient space.

More details will be given in the subsequent discussion.

Still, the problem of finding the Z-eigenvectors of a general tensor $\mathcal{P}$, whereas the equation (23) is only a special case, remains challenging and interesting. See, for example, an interesting discourse in [2] on counting the number of Z-eigenvalues. In the next two subsections we offer two results that might help advance the understanding when restricted to Markov chains with memory. First, we investigate what part of a transition probability tensor $\mathcal{P}$ affects the rate of convergence of the power-like iteration (24), if it converges at all. Second, we analyze a situation of $\mathcal{P}$ where the iteration does not converge at all.

4.1. Attribute of the second dominant eigenvalue

For a square matrix A, it is known that its second dominant eigenvalue affects convergence of the power method. We are curious to know whether there is a similar notion of the second dominant eigenvalue of the tensor $\mathcal{P}$.

By casting such a power-like method for the dominant Z-eigenvector as a fixed-point iteration, we gain some insight into the cause of convergence or divergence for Z-eigenvector computation [7]. In the following, we work specifically on the transition probability tensor $\mathcal{P}$. With slight modification to take into account that Z-eigenvectors are not scaling invariant, the approach can be extended to general tensors. Our main point is to show that for a power-like iteration on tensors the second eigenvalue comes into play in a far more complicated way.

Let $\triangle^{n-1}$ denote the standard simplex in $\mathbb{R}^n$, that is,

$$\triangle^{n-1} = \{\mathbf{z} \in \mathbb{R}^n | z_i \geq 0, \text{ and } \sum_{i=1}^{n} z_i = 1\}. \tag{25}$$

Define the map $\mathbf{f} : \mathbb{R}^n \to \triangle^{n-1}$ by

$$\mathbf{f}(\mathbf{z}) = \frac{\mathcal{P}\mathbf{z}^{m-1}}{\langle \mathcal{P}\mathbf{z}^{m-1}, \mathbf{1} \rangle}, \tag{26}$$

whenever the denominator is not zero. Note that $\mathbf{f}|_{\triangle^{n-1}} = \mathcal{P}\mathbf{z}^{m-1}$ maps $\triangle^{n-1}$ into itself. By the Brouwer fixed-point theorem, there exists at least one point $\widetilde{\mathbf{z}} \in \triangle^{n-1}$ such that $\mathbf{f}(\widetilde{\mathbf{z}}) = \widetilde{\mathbf{z}}$. We are interesting in knowing how fast the iteration (24) converges to such a fixed point, if it converges at all.

We have already introduced one kind of tensor product $\circledast_1$ in (8), namely,

$$\mathcal{P} \circledast_1 \underbrace{\mathbf{z} \otimes \cdots \otimes \mathbf{z}}_{m-1 \text{ times}} = \mathcal{P}\mathbf{z}^{m-1} := \left[\sum_{i_2,\ldots,i_m=1}^{n} p_{\nu_1 i_2 \ldots i_m} x_{i_2} \cdots x_{i_m} \right]_{\nu_1=1}^{n}, \tag{27}$$

[7] The same technique offers an interesting base-free argument for analyzing the conventional power method applied to matrices. Readers might want to read the Appendix first to see how such a local analysis plays out without the burden of dealing with multi-dimensional arrays.

where the subscript in $\circledast_1$ indicates that the first index in $\mathcal{P}$ is excluded from the summation. This tensor product ends up with a column vector whose entries, for convenience, are indexed by ν_1. In a similar way, we now introduce another kind of tensor product $\circledast_{1\ell}$ defined by

$$\mathcal{P} \circledast_{1\ell} \underbrace{\mathbf{z} \otimes \cdots \otimes \mathbf{z}}_{m-2 \text{ times}} := \left[\sum_{i_2,\ldots,\widehat{i_\ell},\ldots,i_m=1}^{n} p_{\nu_1 i_2 \ldots \nu_\ell \ldots i_m} x_{i_2} \cdots \widehat{x_{i_\ell}} \cdots x_{i_m} \right]_{\nu_1,\nu_\ell=1}^{n}, \tag{28}$$

where $\widehat{i_\ell}$ means that quantities associated with this particular index are taken out from the remaining list. The double subscript in $\circledast_{1\ell}$ indicates that the 1-st and the ℓ-th indices in $\mathcal{P}$ are excluded from the summation. This product results in an $n \times n$ matrix whose entries are double indexed by integers (ν_1, ν_ℓ). It is easy to verify that for any given $\mathbf{h} \in \mathbb{R}^n$ we can write

$$\mathcal{P} \circledast_1 \mathbf{z}^{\ell-2} \otimes \mathbf{h} \otimes \mathbf{z}^{m-\ell} = \left(\mathcal{P} \circledast_{1\ell} \mathbf{z}^{m-2}\right) \mathbf{h}, \tag{29}$$

where the right-hand side is a matrix-to-vector multiplication.

We now calculate the Jacobian matrix $D\mathbf{f}(\mathbf{z})$. First, the Fréchet derivative $\mathbf{f}'$ at $\mathbf{z} \in \triangle^{n-1}$ acting on an arbitrary $\mathbf{h} \in \mathbb{R}^n$ is easy to obtain by the generalized Leibniz product rule,

$$\left(\mathcal{P}\mathbf{z}^{m-1}\right)'.\mathbf{h} = \mathcal{P} \circledast_1 \mathbf{h} \otimes \mathbf{z}^{m-2} + \mathcal{P} \circledast_1 \mathbf{z} \otimes \mathbf{h} \otimes \mathbf{z}^{m-3} + \ldots + \mathcal{P} \circledast_1 \mathbf{z}^{m-2} \otimes \mathbf{h}. \tag{30}$$

By using (29), we can represent the action of the derivative operator in terms of matrix-to-vector multiplication:

$$D\mathbf{f}(\mathbf{z})\mathbf{h} = \left(\frac{(\sum_{\ell=2}^{m} \mathcal{P} \circledast_{1\ell} \mathbf{z} \otimes \cdots \otimes \mathbf{z}) \langle \mathcal{P}\mathbf{z}^{m-1}, \mathbf{1} \rangle - \mathcal{P}\mathbf{z}^{m-1}\mathbf{1}^\top (\sum_{\ell=2}^{m} \mathcal{P} \circledast_{1\ell} \mathbf{z} \otimes \cdots \otimes \mathbf{z})}{\langle \mathcal{P}\mathbf{z}^{m-1}, \mathbf{1} \rangle^2} \right) \mathbf{h} \tag{31}$$

and thus retrieve the Jacobian information. In particular, at a fixed point $\widetilde{\mathbf{z}} \in \triangle^{n-1}$, the equation (23) is satisfied and the corresponding Jacobian matrix is reduced to the matrix

$$D\mathbf{f}(\widetilde{\mathbf{z}}) = (I - \widetilde{\mathbf{z}}\mathbf{1}^\top) \underbrace{\left(\sum_{\ell=2}^{m} \mathcal{P} \circledast_{1\ell} \widetilde{\mathbf{z}} \otimes \cdots \otimes \widetilde{\mathbf{z}} \right)}_{\Omega}. \tag{32}$$

Lemma 4.1. *For generic transition probability tensor $\mathcal{P}$, the spectrum of the Jacobian matrix $D\mathbf{f}(\widetilde{\mathbf{z}})$ is composed of zero and those eigenvalues of the matrix Ω whose moduli are strictly less than $m-1$.*

Proof. Clearly, each term $\mathcal{P} \otimes_{1\ell} \widetilde{\mathbf{z}} \otimes \cdots \otimes \widetilde{\mathbf{z}}$ in the summation for Ω is itself a column stochastic matrix. Observe further that

$$\Omega\widetilde{\mathbf{z}} = \left(\sum_{\ell=2}^{m} \mathcal{P} \circledast_{1\ell} \widetilde{\mathbf{z}} \otimes \cdots \otimes \widetilde{\mathbf{z}} \right) \widetilde{\mathbf{z}} = \sum_{\ell=2}^{m} \mathcal{P}\widetilde{\mathbf{z}}^{m-1} = (m-1)\widetilde{\mathbf{z}}. \tag{33}$$

Thus $\lambda_1 = m-1$ is the dominant eigenvalue of Ω with the right eigenvector $\widetilde{\mathbf{z}}$. It follows from (32) that zero is an eigenvalue of the Jacobian $D\mathbf{f}(\widetilde{\mathbf{z}})$ with $\widetilde{\mathbf{z}}$ as the corresponding right eigenvector.

Suppose $\mathbf{w}_i \in \mathbb{C}^n$ is a left eigenvector of Ω with eigenvalues $\lambda_i \in \mathbb{C}$, $i = 2, \ldots, n$. Without loss of generality, we may assume that Ω is a positive matrix generically. By the Perron-Frobenius theorem, the Perron root λ_1 is unique and $|\lambda_i| < m-1$, $i = 2, \ldots n$. It follows that $\mathbf{w}_i^\top \widetilde{\mathbf{z}} = 0$ and, hence,

$$\mathbf{w}_i^\top D\mathbf{f}(\widetilde{\mathbf{z}}) = \mathbf{w}_i^\top (I - \widetilde{\mathbf{z}}\mathbf{1}^\top)\Omega = \mathbf{w}^\top \Omega \quad = \quad \lambda_i \mathbf{w}_i^\top. \tag{34}$$

So $(\lambda_i, \mathbf{w}_i)$ is a left eigenpair of $D\mathbf{f}(\widetilde{\mathbf{z}})$. In other words, if the transition probability tensor $\mathcal{P}$ is generic in the sense that the corresponding matrix Ω is positive, then the spectrum of the Jacobian matrix $D\mathbf{f}(\widetilde{\mathbf{z}})$ is $\{0, \lambda_2, \ldots, \lambda_n\}$. □

Let $\lambda_2(\Omega)$ denote the second largest eigenvalue in modulus of the matrix Ω. Then, by an argument parallel to that outlined in the Appendix, we draw the following conclusion.

Theorem 4.1. *Assuming that the transition probability tensor $\mathcal{P}$ is generic in the sense that the corresponding Ω is positive, then the limiting behavior of the iteration by the power method (24), if the iteration converges at all, has the rate of convergence $|\lambda_2(\Omega)|$ which must be less than 1.*

We summarize our observations as follows. The evolution of state distributions in a memoryless Markov chain is equivalent to the conventional power method applied to the probability transition matrix P defined in (1) directly. If P is positive, then the second dominant eigenvalue of the matrix P alone determines the rate of convergence to the stationary distribution. Likewise, the evolution of joint probability mass functions in a Markov chain with memory induces a power method in the form (14) applied to a transition probability tensor $\mathcal{P}$ defined in (4). It is the second dominant eigenvalue of the flattened matrix $\mathcal{A}$, which depends on $\mathcal{P}$ only, that determines the rate of convergence to a limiting joint probability mass function and, hence, to the stationary distribution of the states. In contrast, if the power-like method (24) is applied to the same transition probability tensor $\mathcal{P}$, then it is the second dominant eigenvalue of the matrix Ω that affects the convergence to a solution $\widetilde{\mathbf{z}}$ of the equation (23). Recall that Ω is defined by

$$\Omega := \sum_{\ell=2}^{m} \mathcal{P} \circledast_{1\ell} \widetilde{\mathbf{z}}^{m-2} \tag{35}$$

which involves a summation over the products of different facets of $\mathcal{P}$ with the fixed point $\widetilde{\mathbf{z}}$. Such a combination is far more complicated than the matrix case. Such an understanding of the cause governing the iteration (24) is interesting and is probably new.

4.2. Examples of divergence

We have already pointed out that $\lambda_1(\Omega) = m-1$ and, for convergence, it is necessary that $|\lambda_2(\Omega)| < 1$. It immediately becomes suspicious that the two dominant eigenvalues of Ω from a given $\mathcal{P}$ can always be so widely separated. In this section, we give a family of examples of a transition probability tensor showing that $|\lambda_2(\Omega)| > 1$ and hence the power-like iteration (24) does not converge.

Consider an order-3 transition probability tensor $\mathcal{P}$ over $S = \{1, 2\}$ with probabilities depicted in Figure 2, where $0 \le a, b, c, d \le 1$ and assume $a + d \ne b + c$. Denote its dominant Z-eigenvector as

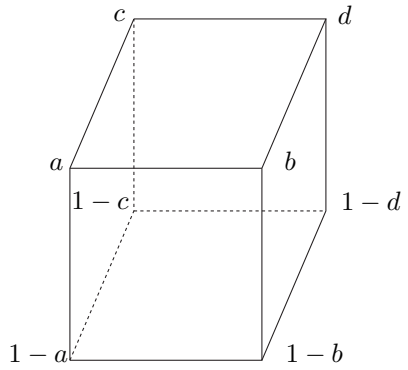

Figure 2: Order-3 transition probability tensor over 2 states.

$\widetilde{\mathbf{z}} = [z, 1-z]^\top$. Then the equation (23) is equivalent to the quadratic equation

$$(a - b - c + d)z^2 + (b + c - 2d - 1)z + d = 0$$

whose two real solutions are trivially

$$z = \frac{2d + 1 - b - c \pm \sqrt{(b + c - 1)^2 + 4d(1 - a)}}{2(a - b - c + d)}. \tag{36}$$

Given values of a, b, c, d, we are interested in the root satisfying $0 \le z \le 1$. By our theory, the corresponding Ω is given by

$$\Omega = \mathcal{P} \otimes_{12} \widetilde{\mathbf{z}} + \mathcal{P} \otimes_{13} \widetilde{\mathbf{z}} \quad = \quad \begin{bmatrix} b+c+(2a-b-c)z & 2d+(b+c-2d)z \\ (-2a+b+c)z+2-b-c & (2d-b-c)z-2d+2 \end{bmatrix} \tag{37}$$

which has eigenvalues 2 and $b + c - 2d + 2(a - b - c + d)z$. Thus the second eigenvalue of Ω is

$$1 \pm \sqrt{(b+c-1)^2 + 4d(1-a)},$$

depending on which z is used.

As an example, take $a = 0$ and $b = c = d = 1$. Then $z = \frac{-1+\sqrt{5}}{2}$ and $\lambda_2 = 1 - \sqrt{5}$. In this case, therefore, the power-like iteration cannot generate the limiting stationary distribution vector $\widetilde{\mathbf{z}}$ because $|\lambda_2| > 1$. In fact, our numerical experiment indicates that the iterates generated by the power-like method will have two accumulation points $[1, 0]^\top$ and $[0, 1]^\top$ and that the iterations move back and forth between these two points. The dominant eigenvector $\widetilde{\mathbf{z}}$ is repelling equilibrium. By continuity, we see that a small perturbation of a, b, c, d while keeping them positive will not change the fact that the corresponding λ_2 has modulus larger than 1. This observation suffices to establish the following result.

Theorem 4.2. *There exists an open set of positive transition probability tensors with nonzero measure for which the power-like iteration (24) will not converge.*

For instance[8], take $a = \epsilon$ and $b = c = d = 1 - \epsilon$. Then the corresponding $\Omega(\epsilon)$ has its second eigenvalue $1 - \sqrt{8\epsilon^2 - 12\epsilon + 5} < -1$ for all $0 \le \epsilon < \frac{3-\sqrt{7}}{4}$.

4.3. Deviation from true stationary distribution

Even if the given transition probability tensor $\mathcal{P}$ is such that the iteration (24) does converge to the dominant Z-eigenvector, we question the rationale of the assumption (21). Now that we understand that a true Markov chain with memory should evolve with a dynamics for the sequence of vectors $\{\mathbf{x}^{(t)}\}$ in the sense of (8) and a dynamics for the sequence of tensors $\{\Pi_{t,t-1,\ldots,t-m+2}\}$ in the sense of (11), we perform some numerical simulations to investigate whether there is a statistically significant deviation between results based on this assumption and those from the true Markov process.

Denote the limiting joint probability mass function, the stationary distribution, and the dominant Z-eigenvector by $\widetilde{\Pi}$, $\widetilde{\mathbf{x}}$, and $\widetilde{\mathbf{z}}$, respectively. To simulate the general behavior of these quantities, we have to try out large samples of Markov chains. It will be sufficiently informative to consider the Markov chain with memory 2. Toward this goal, the columns (in the sense of (5)) of the order-3 transition probability tensor $\mathcal{P}$ can be thought of as coming from a uniform distribution over the simplex Δ^{n-1}.

Lemma 4.2. *Let $\mathcal{P}$ be a random order-3 tensor with independent and identically distributed columns from the simplex Δ^{n-1}. Then the random vectors $\widetilde{\mathbf{x}}$ and $\widetilde{\mathbf{z}}$ have the same expected value*

$$\mathbb{E}(\widetilde{\mathbf{x}}) = \mathbb{E}(\widetilde{\mathbf{z}}) = \left[\frac{1}{n}, \ldots, \frac{1}{n}\right]^\top. \tag{38}$$

Also,

$$\mathbb{E}(\widetilde{\Pi}) = \frac{1}{n^2}\mathbb{1}, \tag{39}$$

where $\mathbb{1}$ is the $n \times n$ matrix with all ones.

[8] For this example, the value γ defined by formula (2.2) in [12] is equal to 2, but we do not see convergence. This is in contrast to the assertion of Theorem 3.1 in [12].

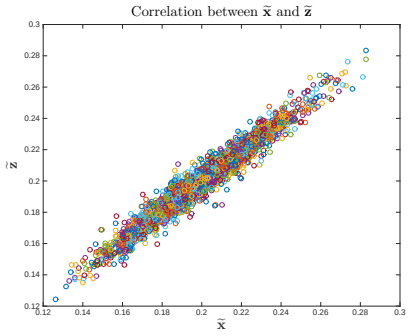

Figure 3: Plots of $\widetilde{\mathbf{z}}$ versus $\widetilde{\mathbf{x}}$ over 500 randomly generated order-3 transition probability tensors $\mathcal{P}$. Colors represent different rows in the vectors.

Proof. Each of $\widetilde{\Pi}$, $\widetilde{\mathbf{x}}$, and $\widetilde{\mathbf{z}}$ depends on $\mathcal{P}$ and, therefore, is itself a random variable of some distributions. We need not specify the exact distributions of $\widetilde{\Pi}$, $\widetilde{\mathbf{x}}$, and $\widetilde{\mathbf{z}}$. However, because any column permutation of $\mathcal{P}$ leads to the same distribution of $\mathcal{P}$, any row permutation of the vector $\widetilde{\mathbf{x}}$ (or $\widetilde{\mathbf{z}}$) leads to the same distribution of $\widetilde{\mathbf{x}}$ (or $\widetilde{\mathbf{z}}$). All entries of the vector $\widetilde{\mathbf{x}}$ (or $\widetilde{\mathbf{z}}$) must share the same contribution. This symmetry implies (38).

Together with the additional fact that any "row" permutation of $\mathcal{P}$ leads to the same distribution of $\mathcal{P}$, we conclude that all entries of the limiting joint probability mass function $\widetilde{\Pi}$ share the same contribution and, hence, (39) follows. □

The proof for (38) does not rely on the order of the tensor. We may thus say that the dominant Z-eigenvector $\widetilde{z}$ is always an unbiased estimator of the stationary distribution $\widetilde{x}$ of the Markov chain with memory. However, for Markov chains with memory 2, we obtain the difference

$$\mathbb{E}(\widetilde{\mathbf{x}}^2) - \mathbb{E}(\widetilde{\Pi}) = \mathbb{E}(\widetilde{\mathbf{x}}\widetilde{\mathbf{x}}^\top) - \mathbb{E}(\widetilde{\mathbf{x}})\mathbb{E}(\widetilde{\mathbf{x}})^\top = \mathbb{E}(\widetilde{\mathbf{x}}\widetilde{\mathbf{x}}^\top) - \mathbb{E}(\widetilde{\mathbf{z}})\mathbb{E}(\widetilde{\mathbf{z}})^\top = \mathrm{cov}(\widetilde{\mathbf{x}}) \tag{40}$$

which clearly indicates the assumption (21) for a Markov chain with memory 2 is off by an average of the amount $\mathrm{cov}(\widetilde{x})$. For Markov chains with memory higher than 2, the difference between $\mathbb{E}(\widetilde{\mathbf{x}}^{m-1})$ and $\mathbb{E}(\widetilde{\Pi})$ is algebraically more complicated.

As an illustration, we randomly generate 500 test data over $\Delta^4 \subset \mathbb{R}^5$. Each data set includes one order-3 transition probability tensor $\mathcal{P}$ and two starting distribution vectors $\mathbf{x}_{-1}$ and $\mathbf{x}_0$. Entries in the data are generated independently from the identical uniform distribution over the interval $[0, 1]$ and then are normalized accordingly to meet the stochastic requirements. Let the order-2 tensor $\Pi_{0,-1} = \mathbf{x}_{-1} \otimes \mathbf{x}_0$ represent the joint probability mass function for the starting memory. After going through the iterative processes described in the preceding sections, we gather the limiting points of these test data and compare.

We first plot the correlation between $\widetilde{\mathbf{x}}$ and $\widetilde{\mathbf{z}}$. Each circle "o" in Figure 3 with one specific color represents one pair of entries $(\widetilde{x}_i, \widetilde{z}_i)$, $i = 1, \ldots, 5$, from the limiting distributions $\widetilde{\mathbf{x}}$ and $\widetilde{\mathbf{z}}$ of the same order-3 tensor $\mathcal{P}$. We already know that all entries of $\widetilde{\mathbf{x}}$ (and $\widetilde{\mathbf{z}}$) share the same distribution, so it does not matter if we plot the regressions of all entries together. Table 1 summarizes the regression values R, slopes M, and z-intercepts B for each row. The high R values suggest that it is reasonable to assume that $\widetilde{\mathbf{x}}$ and $\widetilde{\mathbf{z}}$ are linearly correlated.

	$(\widetilde{x}_1, \widetilde{z}_1)$	$(\widetilde{x}_2, \widetilde{z}_2)$	$(\widetilde{x}_3, \widetilde{z}_3))$	$(\widetilde{x}_4, , \widetilde{z}_4)$	$(\widetilde{x}_5, , \widetilde{z}_5)$
R	0.9790	0.9729	0.9748	0.9727	0.9750
M	0.9343	0.9332	0.9414	0.9384	0.9488
B	0.0129	0.0134	0.0115	0.0128	0.0102

Table 1: Regression values R, slopes M, and z-intercepts

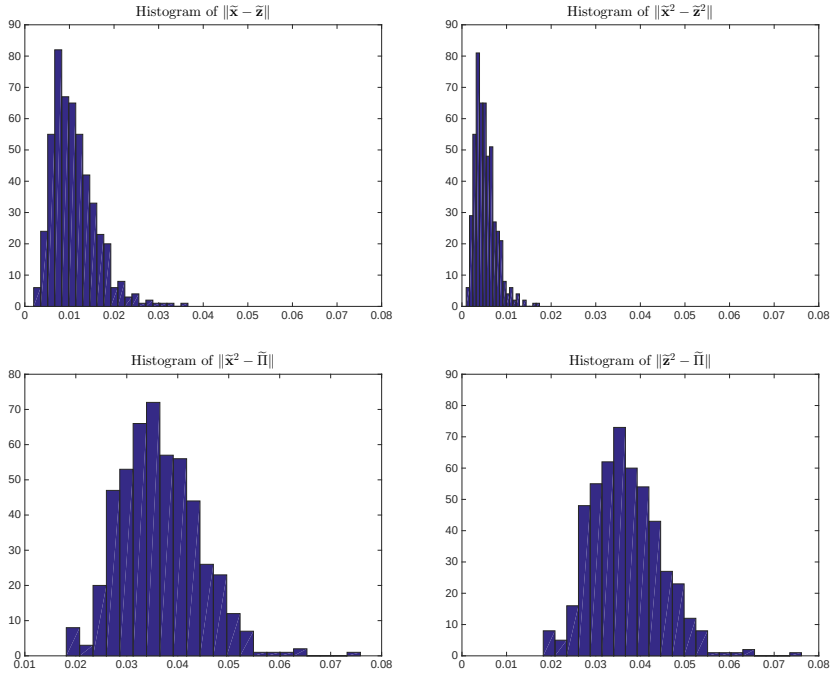

Figure 4: Comparisons between the stationary distribution $\widetilde{\mathbf{x}}$ and the dominant Z-eigenvector $\widetilde{\mathbf{z}}$, and the corresponding distributions of memory.

Regarding $\widetilde{\mathbf{x}}$ and $\widetilde{\mathbf{z}}$ as the output from two distinct procedures of the same input $\mathcal{P}$, it might be curious to know how they differentiate from each other case by case. We thus plot the histograms of the 2-norms $\|\widetilde{\mathbf{x}} - \widetilde{\mathbf{z}}\|$, $\|\widetilde{\mathbf{x}}^2 - \widetilde{\Pi}\|$, $\|\widetilde{\mathbf{z}}^2 - \widetilde{\Pi}\|$, and $\|\widetilde{\mathbf{x}}^2 - \widetilde{\mathbf{z}}^2\|$ out of the 500 random tests in Figure 4 over the same scale. Without delving into rigorous statistical testing, we can see that the variations $\|\widetilde{\mathbf{z}} - \widetilde{\mathbf{x}}\|$ and $\|\widetilde{\mathbf{z}}^2 - \widetilde{\mathbf{x}}^2\|$ shown in the upper drawing does suggest that the two stationary distributions $\widetilde{\mathbf{z}}$ and $\widetilde{\mathbf{x}}$ might be called statistically close [1], which is also suggested by the closeness to one of the regression slops M and to zero of the z-intercepts B in Table 1. However, the variations in the lower drawing indicate that the difference between the true limiting joint probability mass function $\widetilde{\Pi}$ and the assumed limiting joint probability mass function $\widetilde{\mathbf{z}}^2$ by (21) is statistically more significant. In fact, in the case of Markov chains with memory 2, we have argued that the difference is averaged at the amount $\mathrm{cov}(\widetilde{\mathbf{x}})$.

Furthermore, based on the same experiment with 500 random data above, we observe that every limiting joint probability mass function $\widetilde{\Pi}$ is of full matrix rank, whereas the matrix $\widetilde{\mathbf{z}}^2$ is of rank 1. We notice that one singular value of $\widetilde{\Pi}$ is always significantly larger than the other four singular values. See Figure 5. However, we have also checked that the matrix $\widetilde{\mathbf{z}}^2$ is by no means the best rank-1 approximation to $\widetilde{\Pi}$ in the sense of minimizing $\|\widetilde{\Pi} - \mathbf{z}^2\|$ subject to $\mathbf{z} \in \Delta^4$. Barring this rank difference, it remains open for interpretation of whether $\widetilde{\mathbf{z}}$ can really be used as a reasonable approximation to $\widetilde{\mathbf{x}}$ in general. If yes, to what degree of statistical closeness is it acceptable?

5. Conclusions

A Markov chain with memory has a natural representation as a transition probability tensor. The Markov process involves the progression of two types of distributions. The probability distribution of the states evolves as a tensor product $\circledast_1$ in the way defined in (8), whereas the joint probability mass distribution of the memory evolves as another tensor product $\boxdot$ in the way defined in (9). The latter is essentially the same as the conventional power method, whereas the stationary distribution of the states can be obtained from the

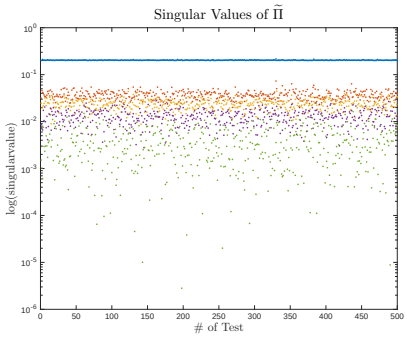

Figure 5: Logarithmic plot of singular values of the stationary joint probability mass function $\widetilde{\Pi}$.

"row sum" of the stationary distribution of the memory. We therefore conclude that for a generic Markov chain with memory the distribution of the states does converge to a stationary distribution.

We also provide a statistical basis that in average the difference between the true stationary distribution $\widetilde{\mathbf{x}}$ and the Z-eigenvector $\widetilde{\mathbf{z}}$ calculated from the assumed shortcut memory (21) is statistically indistinguishable. However, such an assumption on memory is questionable as is evidenced by (40). We further demonstrate by a family of transition probability tensors that the power-like iteration (24) for Z-eigenvector calculation may fail to converge.

In all cases, we propose a general approach by casting any of the power-like iterative schemes as a fixed-point iteration and draw conclusion on the limiting behavior of such an iterative method via the spectrum of the associated Jacobian matrix. The insight obtained from local analysis on the particular matrix Ω defined in (32) and the effect of its second dominant eigenvalue is perhaps new.

6. Appendix: Convergence of the power iteration for matrices

Suppose that a given matrix $A \in \mathbb{R}^{n\times n}$ has a dominant eigenvalue λ_1 in the sense that its spectrum satisfies $|\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq \ldots \geq |\lambda_n|$. It is well known that the sequence $\{\mathbf{x}_k\}$ generated from the iterative scheme

$$\left\{ \begin{array}{rcl} \mathbf{w}_{k+1} & := & A\mathbf{x}_k, \\ \mathbf{x}_{k+1} & := & \frac{\mathbf{w}_{k+1}}{\|\mathbf{w}_{k+1}\|}, \end{array} \right. \tag{41}$$

converges to the unit eigenvector $\mathbf{v}_1$ associated with λ_1. This procedure, known as the power method, has been the most rudimentary means for eigenvalue computation. Though the power method is not effective per se, its fundamental principle sheds light on more advanced methods. For example, the Rayleigh quotient iteration which is a variation of the shifted inverse power method continues to play an integral role due to its rapid convergence [22], whereas the shifted QR algorithm which can be interpreted as an application of the power method on subspaces with reorthogonalization is modern day's power horse for eigenvalue computation [29]. In certain cases, the power method remains to be useful for computing the eigenvector associated with the dominant eigenvalue of a matrix. One such instance is in the application of Markov chain analysis where the stationary distribution $\boldsymbol{\pi}$ satisfying $P\boldsymbol{\pi} = \boldsymbol{\pi}$ for a column stochastic matrix P is needed. Recall that the value 1 is universally the dominant eigenvalue of any stochastic matrix, so $\boldsymbol{\pi}$ is the dominant right eigenvector.

The ratio $|\frac{\lambda_2}{\lambda_1}|$ is widely recognized as the convergence rate of a power method, whence defining the important role of the second dominant eigenvalue λ_2 in eigenvalue computation. The search engine Google, for example, exploits this knowledge by introducing a perturbation to force a bound on λ_2 of its hyperlink matrix which is stochastic. The power method is then employed to approximate the corresponding stationary distribution $\boldsymbol{\pi}$, known as the PageRank, to help rank the relative importance of a particular web page [11].

A typical way in numerical linear algebra to argue the rate of convergence of the power method (41) for a matrix A is to assume the existence of a basis of eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$[9]. Upon expanding the starting vector

$$\mathbf{x}_0 = \sum_{i=1}^{n} c_i \mathbf{v}_i$$

in terms of the basis, the iterate $\mathbf{x}_k$ can then be expressed as

$$\mathbf{x}_k = \frac{c_1 \lambda_1^k \left(\mathbf{v}_1 + \sum_{i=2}^n c_i \left(\frac{\lambda_i}{\lambda_1} \right)^k \mathbf{v}_i \right)}{\left\| c_1 \lambda_1^k \left(\mathbf{v}_1 + \sum_{i=2}^n c_i \left(\frac{\lambda_i}{\lambda_1} \right)^k \mathbf{v}_i \right) \right\|}.$$

Hence, we see that the non-essential quantities decay at a rate of approximately $|\frac{\lambda_2}{\lambda_1}|$. Such a loose argument is conceptually acceptable, but can hardly be generalized to tensors because the tensor space may not have a basis of eigenvectors. An alternative argument is to use fixed-point theory. We have done so for tensors in Section 4.1. We now demonstrate how it applied to matrices.

Let $\mathbb{S}^{n-1}$ denote the unit sphere in $\mathbb{R}^n$. Without loss of generality, suppose A is nonsingular. Define a map $\mathbf{f} : \mathbb{S}^{n-1} \to \mathbb{S}^{n-1}$ by

$$\mathbf{f}(\mathbf{x}) = \frac{A\mathbf{x}}{\|A\mathbf{x}\|_2} \tag{42}$$

where the normalization by the 2-norm is only for convenience. The power method can be cast as the fixed-point iteration

$$\mathbf{x}_{k+1} = \mathbf{f}(\mathbf{x}_k). \tag{43}$$

Since $\mathbf{f}$ is a continuous function mapping from a compact set into itself, by the Brouwer fixed-point theorem, there is a point $\widetilde{\mathbf{x}} \in \mathbb{S}^{n-1}$ such that $\mathbf{f}(\widetilde{\mathbf{x}}) = \widetilde{\mathbf{x}}$. In particular, by switching the sign if necessary, we may assume that the dominant unit eigenvector $\mathbf{v}_1$ is one such a fixed point. We now describe the local behavior of $\mathbf{f}$ nearby $\mathbf{v}_1$.

For $\mathbf{x}_k$ sufficiently near $\mathbf{v}_1$, we have the linear approximation

$$\mathbf{x}_{k+1} - \mathbf{v}_1 = \mathbf{f}(\mathbf{x}_k) - \mathbf{f}(\mathbf{v}_1) \approx D\mathbf{f}(\mathbf{v}_1)(\mathbf{x}_k - \mathbf{v}_1), \tag{44}$$

where it is easy to see that the Jacobian matrix of $\mathbf{f}$ is given by

$$D\mathbf{f}(\mathbf{x}) = \frac{A}{\|A\mathbf{x}\|_2} - \frac{A\mathbf{x}\mathbf{x}^\top A^\top A}{\|A\mathbf{x}\|_2^3}. \tag{45}$$

It follows that at $\mathbf{v}_1$ we have

$$D\mathbf{f}(\mathbf{v}_1) = \frac{1}{|\lambda_1|}(I - \mathbf{v}_1\mathbf{v}_1^\top)A. \tag{46}$$

Obviously, $\mathbf{v}_1^\top D\mathbf{f}(\mathbf{v}_1) = 0$. Let $\mathbf{w}_i \in \mathbb{C}^n$ be any eigenvector of $A^\top$ associated with eigenvalue $\lambda_i \in \mathbb{C}$, $i = 2, \ldots n$. Then it is known that $\mathbf{w}_i^\top \mathbf{v}_1 = 0$ since $\lambda_i \neq \lambda_1$. Thus $\mathbf{w}_i^\top D\mathbf{f}(\mathbf{v}_1) = \frac{\lambda_i}{|\lambda_1|}\mathbf{w}_i^\top$. In all, we make the following conclusion.

Lemma 6.1. *The spectrum of the Jacobian matrix $D\mathbf{f}(\mathbf{v}_1)$ is precisely $\left\{0, \frac{\lambda_2}{|\lambda_1|}, \frac{\lambda_3}{|\lambda_1|}, \ldots, \frac{\lambda_n}{|\lambda_1|}\right\}$.*

[9] In case that the matrix A is defective, some arguments can still be made. See, for example, detailed discussions in the classic book [7]. The local analysis presented in this section, however, does not require such a basis.

As a simple demonstration, consider the generic case that the matrix $Df(\mathbf{v}_1)$ has a spectral decomposition $D\mathbf{f}(\mathbf{v}_1) = U^{-1}\Lambda U$. Then by (44) we can write

$$U(\mathbf{x}_{k+1} - \mathbf{v}_1) \approx \Lambda U(\mathbf{x}_k - \mathbf{v}_1), \tag{47}$$

implying that

$$\|U(\mathbf{x}_{k+1} - \mathbf{v}_1)\|_\infty \approx \left|\frac{\lambda_2}{\lambda_1}\right| \|U(\mathbf{x}_k - \mathbf{v}_1)\|_\infty. \tag{48}$$

It is in this sense that once $\mathbf{x}_k$ is sufficiently close to $\mathbf{v}_1$, then $\mathbf{x}_{k+1}$ is even closer and that the rate of linear convergence is given by the ratio $|\frac{\lambda_2}{\lambda_1}|$.

References

[1] T. BATU, L. FORTNOW, R. RUBINFELD, W. D. SMITH, AND P. WHITE, *Testing that distributions are close*, in 41st Annual Symposium on Foundations of Computer Science (Redondo Beach, CA, 2000), IEEE Comput. Soc. Press, Los Alamitos, CA, 2000, pp. 259–269.

[2] D. CARTWRIGHT AND B. STURMFELS, *The number of eigenvalues of a tensor*, Linear Algebra Appl., 438 (2013), pp. 942–952.

[3] K. CHANG, L. QI, AND T. ZHANG, *A survey on the spectral theory of nonnegative tensors*, Numer. Linear Algebra Appl., 20 (2013), pp. 891–912.

[4] K. C. CHANG AND T. ZHANG, *On the uniqueness and non-uniqueness of the positive z-eigenvector for transition probability tensors*, J. Math. Anal. Appl., 408 (2013), pp. 525–540.

[5] Z. CHEN, L. QI, Q. YANG, AND Y. YANG, *The solution methods for the largest eigenvalue (singular value) of nonnegative tensors and convergence analysis*, Linear Algebra Appl., 439 (2013), pp. 3713–3733.

[6] H. G. DIAZ, R. MOLINA, AND E. URIARTE, *Stochastic molecular descriptors for polymers. 1. modelling the properties of icosahedral viruses with 3d-markovian negentropies*, Polymer, 45 (2004), pp. 3845 – 3853.

[7] D. K. FADDEEV AND V. N. FADDEEVA, *Computational methods of linear algebra*, Translated by Robert C. Williams, W. H. Freeman and Co., San Francisco-London, 1963.

[8] D. F. GLEICH, L.-H. LIM, AND Y. YU, *Multilinear PageRank*, SIAM J. Matrix Anal. Appl., 36 (2015), pp. 1507–1541.

[9] Y. HUA, X. LIU, AND H. JIANG, *ANTELOPE: a semantic-aware data cube scheme for cloud data center networks*, IEEE Trans. Comput., 63 (2014), pp. 2146–2159.

[10] T. G. KOLDA AND B. W. BADER, *Tensor decompositions and applications*, SIAM Rev., 51 (2009), pp. 455–500.

[11] A. N. LANGVILLE AND C. D. MEYER, *Google's PageRank and beyond: the science of search engine rankings*, Princeton University Press, Princeton, NJ, 2012. Paperback edition of the 2006 original.

[12] W. LI AND M. K. NG, *On the limiting probability distribution of a transition probability tensor*, Linear Multilinear Algebra, 62 (2014), pp. 362–385.

[13] L.-H. LIM, *Singular values and eigenvalues of tensors: A variational approach*, in Proceedings of 1st IEEE International Workshop on Computational Advances of Multi-Tensor Adaptive Processing (CAMSAP), Puerto Vallarta, December 13–15 2005, pp. 129–132.

[14] L.-H. LIM, M. K. NG, AND L. QI, *The spectral theory of tensors and its applications*, Numer. Linear Algebra Appl., 20 (2013), pp. 889–890.

[15] Y. LIU, G. ZHOU, AND N. F. IBRAHIM, *An always convergent algorithm for the largest eigenvalue of an irreducible nonnegative tensor*, J. Comput. Appl. Math., 235 (2010), pp. 286–292.

[16] I. L. MACDONALD AND W. ZUCCHINI, *Hidden Markov and other models for discrete-valued time series*, vol. 70 of Monographs on Statistics and Applied Probability, Chapman & Hall, London, 1997.

[17] R. S. MAMON AND R. J. ELLIOTT, eds., *Hidden Markov models in finance*, International Series in Operations Research & Management Science, 104, Springer, New York, 2007.

[18] ——, eds., *Hidden Markov models in finance*, International Series in Operations Research & Management Science, 209, Springer, New York, 2014. Further developments and applications. Vol. II.

[19] S. S. MELNYK, O. V. USATENKO, AND V. A. YAMPOL′SKII, *Memory functions of the additive Markov chains: applications to complex dynamic systems*, Phys. A, 361 (2006), pp. 405–415.

[20] S. L. NARASIMHAN, J. A. NATHAN, AND K. P. N. MURTHY, *Can coarse-graining introduce long-range correlations in a symbolic sequence?*, EPL (Europhysics Letters), 69 (2005), p. 22.

[21] M. NG, L. QI, AND G. ZHOU, *Finding the largest eigenvalue of a nonnegative tensor*, SIAM J. Matrix Anal. Appl., 31 (2009), pp. 1090–1099.

[22] B. N. PARLETT, *The Rayleigh quotient iteration and some generalizations for nonnormal matrices*, Math. Comp., 28 (1974), pp. 679–693.

[23] L. QI, *Eigenvalues and invariants of tensors*, J. Math. Anal. Appl., 325 (2007), pp. 1363–1377.

[24] A. E. RAFTERY, *A model for high-order Markov chains*, J. Roy. Statist. Soc. Ser. B, 47 (1985), pp. 528–539.

[25] O. ROSE, *A memory markov chain model for vbr traffic with strong positive correlations*, tech. report, Institute of Computer Science, University of Wüzburg, 1999. downloadable at http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.8.9230&rep=rep1&type=pdf.

[26] E. ROSOLOWSKY, *Statistical analyses of data cubes*, in Statistical challenges in modern astronomy V, vol. 209 of Lecture Notes in Statist., Springer, New York, 2013, pp. 367–382.

[27] V. SOLOVIEV, V. SAPTSIN, AND D. CHABANENKO, *Markov Chains application to the financial-economic time series prediction*, ArXiv e-prints, (2011).

[28] O. V. USATENKO, V. A. YAMPOL'SKII, K. E. KECHEDZHY, AND S. S. MEL'NYK, *Symbolic stochastic dynamical systems viewed as binary N -step markov chains*, Phys. Rev. E, 68 (2003), p. 061107.

[29] D. S. WATKINS, *Understanding the* QR *algorithm*, SIAM Rev., 24 (1982), pp. 427–440.

[30] L. ZHANG AND L. QI, *Linear convergence of an algorithm for computing the largest eigenvalue of a nonnegative tensor*, Numer. Linear Algebra Appl., 19 (2012), pp. 830–841.

[31] G. ZHOU, L. QI, AND S.-Y. WU, *Efficient algorithms for computing the largest eigenvalue of a nonnegative tensor*, Front. Math. China, 8 (2013), pp. 155–168.