ON THE SECOND DOMINANT EIGENVALUE AFFECTING THE POWER METHOD FOR TRANSITION PROBABILITY TENSORS
DRAFT AS OF May 23, 2014

MOODY T. CHU * AND SHENG-JHIH WU†

Abstract. It is known that the second dominant eigenvalue of a matrix determines the convergence rate of the power method. Though ineffective for general eigenvalue computation, the power method has been of practical usage for computing the stationary distribution of a stochastic matrix. For a Markov chain with memory m, the transition "matrix" becomes an order-m tensor. Under suitable assumptions, the same power method has been used to compute the limiting probability distribution of a transition probability tensor. What is not clear is what affects the convergence rate of the iteration, if the method converges at all. Casting the power method as a fixed-point iteration, we examine the local behavior of the nonlinear map, identify the cause of convergence or divergence, and provide a family of counterexamples showing that even if the transition probability tensor is irreducible and aperiodic, the power iteration may fail to converge.

AMS subject classifications. 15A18, 15A51, 15A69, 60J99

Key words. power method, Markov chain with memory, transition probability tensors, rate of convergence, stochastic matrices, stationary distributions, second dominant eigenvalue of tensors

1. Introduction. 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 can be proved that the sequence $\{\mathbf{x}_k\}$ generated from the iterative scheme

$$\left\{ \begin{array}{rcl} \mathbf{w}_{k+1} & := & A\mathbf{x}_k, \\ \mathbf{x}_{k+1} & := & \frac{\mathbf{w}_{k+1}}{\|\mathbf{w}_{k+1}\|}, \end{array} \right. \tag{1.1}$$

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 [14], 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 [16].

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 $\boldsymbol{\pi}P = \boldsymbol{\pi}$ for a row stochastic matrix P is needed. Recall 1 is universally the dominant eigenvalue of any stochastic matrix, so $\boldsymbol{\pi}$ is the dominant left eigenvector.

It is a well known fact that the second dominant eigenvalue λ_2 of A affects the convergence behavior of the sequence $\{\mathbf{x}_k\}$. It is typically said that converges rate is the ratio $|\frac{\lambda_2}{\lambda_1}|$. An important application that exploits this knowledge is the search engine Google. First, the very large-scale web hyperlink matrix H is slightly modified via a binary column vector $\mathbf{a}$ to remove dangles (caused by dead end pages) and keep it row stochastic. Then the rank-1 matrix $E := \frac{\mathbf{1}\mathbf{1}^\top}{n}$ with $\mathbf{1} = [1, 1, \ldots, 1]^\top$ is added through the combination

$$G(c) = \alpha \left(H + \frac{\mathbf{a}\mathbf{1}^\top}{n} \right) + (1-\alpha)E,$$

where $\alpha \in (0, 1)$, so that $G(\alpha)$ is row stochastic, irreducible, and aperiodic, which ensures a unique and positive stationary distribution. The power method is thus employed to approximate the corresponding $\boldsymbol{\pi}$

*Department of Mathematics, North Carolina State University, Raleigh, NC 27695-8205, USA. (chu@math.ncsu.edu.) This research was supported in part by the National Science Foundation under grants DMS-1014666 and DMS-1316779.

†Institute of Mathematics, Academia Sinica, 6F, Astronomy-Mathematics Building, No.1, Sec. 4, Roosevelt Road, Taipei 10617, Taiwan. (sjw@math.sinica.edu.tw).

which is known as the PageRank in the internet community [7]. Because H is of size billions by billions, even a straightforward task such as matrix-to-vector multiplications is overwhelming. We can afford to perform only finitely many times of iterations. The role of the second dominant eigenvalue therefore becomes important. It is an exercise to show that $|\lambda_2| = \alpha$ [12]. It has been said that Google uses $\alpha = .85$, whence the PageRank with be within 10^{-4} accuracy by about 60 iterations, regardless of the size of the matrix H.

As the demand for large data by different disciplines increases immensely in recent years, considerable attention has been turned to higher-order tensors. 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 instance. Depending on the applications, there are several ways to mull over how an eigenvalue of a tensor should be defined [3, 9, 10, 15]. In this paper, we use Markov chain with memory as a model to rationalize two concepts of eigenvalues.

1. The classical concept of eigenvalues when a transition probability tensor is used to derive the evolution of the joint probability density functions.
2. The so called Z-eigenvalue when a transition probability tensor is used to derive the evolution of the state probability distribution.

Our interest is motivated by recent applications of the power method to compute the limiting probability distribution vector associated with a Markov chain with memory [8]. Two questions arise and we intend to offer a clarification in this presentation. First, the assumption used to propose the Z-eigenvector computation is dubious and unjustifiable as our theory will show. Second, even if the proposed scheme of dominant Z-eigenvector is acceptable, what affects its rate of convergence is still not clear. Our theory gives insight into this limiting behavior. Under a different context, it is possible that the largest of the so called H-eigenvalues becomes desirable [13]. Our framework of analysis can easily be generalized.

This paper is organized as follows. In order to understand the effect of the second dominant eigenvalue, we recast the power method applied to a matrix a fixed-point iteration and perform a local analysis in Section 2. This approach is somewhat different from the traditional way to prove the convergence of power method, but allows us to generalize the analysis for the power method to higher-order tensors. For our applications, we concentrate on transition probability tensors which naturally arise from Markov chain with memory. This connection is briefly reviewed in Section 3. A Markov chain with memory requires two types of dynamics to reflect the evolution of the probability distributions for both the state and the memory. The dynamics naturally translates into two types of power methods in order to find the probability distributions of the stead-state vector and the joint probability density function, respectively. These iterative schemes and their limiting behavior are discussed in Sections 3.2 and 4, respectively. One of our main contributions is at identifying the true origination of the "second" dominant eigenvalue in the tensor setting. As a consequence, we find that the convergence of the power method proposed in [8] is not always guaranteed and we illustrate this by counter examples.

2. Convergence of power method for matrices. A typical way to argue the rate of convergence of the power method for a matrix A is by expanding the starting vector

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

in terms of the basis of eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$. 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.

We now offer an alternative argument. 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

$$\mathbf{f}(\mathbf{x}) = \frac{A\mathbf{x}}{\|A\mathbf{x}\|_2} \tag{2.1}$$

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

$$\mathbf{x}_{k+1} = \mathbf{f}(\mathbf{x}_k). \tag{2.2}$$

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

$$\mathbf{x}_{k+1} - \mathbf{v}_1 = \mathbf{f}(\mathbf{x}_k) - \mathbf{f}(\mathbf{v}_1) \approx D\mathbf{f}(\mathbf{v}_1)(\mathbf{x}_k - \mathbf{v}_1), \tag{2.3}$$

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

$$D\mathbf{f}(\mathbf{x}) = \frac{A}{\|A\mathbf{x}\|_2} - \frac{A\mathbf{x}\mathbf{x}^\top A^\top A}{\|A\mathbf{x}\|_2^3}. \tag{2.4}$$

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

$$D\mathbf{f}(\mathbf{v}_1) = \frac{1}{|\lambda_1|}(I - \mathbf{v}_1\mathbf{v}_1^\top)A. \tag{2.5}$$

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

Let $D\mathbf{f}(\mathbf{v}_1) = U^{-1}\Lambda U$ be the spectral decomposition of the matrix $Df(\mathbf{v}_1)$. Then by (2.3) we can write

$$U(\mathbf{x}_{k+1} - \mathbf{v}_1) \approx \Lambda U(\mathbf{x}_k - \mathbf{v}_1), \tag{2.6}$$

implying that

$$\|U(\mathbf{x}_{k+1} - \mathbf{v}_1)\|_\infty \approx \left|\frac{\lambda_2}{\lambda_1}\right| \|U(\mathbf{x}_k - \mathbf{v}_1)\|_\infty. \tag{2.7}$$

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 $|\frac{\lambda_2}{\lambda_1}|$.

3. Transition probability tensors. A typical Markov chain is a stochastic process of random variables $\{X_t\}_{t=0}^\infty$ over a finite state space S, where the conditional probability distribution of future states in the process depends only upon the present state, but not on the sequence of events that preceded it. That is, among the states $s_i \in S$, we have the Markov property

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

Without loss of generality, we may identify the states as $S = \{1, 2, \ldots, n\}$. Assume further that the chain is time homogeneous. Then a transition probability matrix $P = [p_{ij}]$ defined by

$$p_{ij} := \Pr(X_{t+1} = j \mid X_t = i) \tag{3.1}$$

is independent of t and row stochastic because, given the current state $X_t = i$, the next step X_{t+1} must assume one of the states in S. The above process is generally characterized as memoryless[1].

There are situations where the data sequence does depend on past values. A number of interesting applications are mentioned in [8] with references which we shall not repeat here. The list includes wind turbine design, alignment of DNA sequences, and growth of polymer chains due to steric effect. To model this kind of phenomenon, a Markov chain with memory m is a process satisfying

$$\Pr(X_{t+1} = s_{t+1}|X_t = s_t, \ldots, X_1 = s_1) = \Pr(X_{t+1} = s_{t+1}|X_t = s_t, \ldots, X_{t-m+1} = s_{t-m+1}) \quad (3.2)$$

for all $t \geq m$. Assume 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

$$p_{i_1 i_2 \ldots i_m} := \Pr(X_{t+1} = i_1|X_t = i_2, \ldots, X_{t-m+2} = i_m). \quad (3.3)$$

Necessarily, $0 \leq p_{i_1 i_2 \ldots i_m} \leq 1$ and for every fixed $(m-1)$-tuple $(i_2, \ldots, i_m)$ we have

$$\sum_{i_1=1}^{n} p_{i_1 i_2 \ldots i_m} = 1. \quad (3.4)$$

$\mathcal{P}$ is called a transition probability tensor[2].

3.1. Markov chain process. The dynamics of a Markov chain with memory goes as follows. Let the joint probability density function of state variable $X_t, \ldots X_{t-m+2}$ over S at time t be denoted as

$$\Pi_{t,t-1,\ldots,t-m+2} = [\pi^{(t)}_{i_2 \ldots i_m}], \quad (3.5)$$

where

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

Note that $\Pi_{t,t-1,\ldots,t-m+2}$ is an order-(m-1) tensor with the property

$$\sum_{i_2,\ldots,i_m=1}^{n} \pi^{(t)}_{i_2 \ldots i_m} = 1. \quad (3.6)$$

The probability distribution of X_{t+1}, denoted by the column vector $\mathbf{x}^{(t+1)}$, can be calculated as the tensor product

$$\mathbf{x}^{(t+1)} = \mathcal{P} \circledast_1 \Pi_{t,t-1,\ldots,t-m+2} := [\langle p_{i_1,:}, \Pi_{t,t-1,\ldots,t-m+2} \rangle] \in \mathbb{R}^n, \quad (3.7)$$

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. In order to carry on this evolution, we need to compute the next joint probability density function $\Pi_{t+1,t,\ldots,t-m+3} = [\pi^{(t+1)}_{i_1 \ldots i_{m-1}}]$. Toward this, we observe

$$\begin{aligned} \Pi_{t+1,t,\ldots,t-m+3} &= \sum_{i_m=1}^{n} \Pr(X_{t+1}, X_t, \ldots, X_{t-m+3}, X_{t-m+2} = i_m) \\ &= \sum_{i_m=1}^{n} \Pr(X_{t+1}|X_t, \ldots, X_{t-m+3}, X_{t-m+2} = i_m) \Pr(X_t, \ldots X_{t-m+3}, X_{t-m+2} = i_m). \end{aligned} \quad (3.8)$$

[1] Strictly speaking, based on the definition (3.2), it should be called a chain with memory 1.

[2] In the case of a chain with memory 1, note that $\mathcal{P}$ is column stochastic.

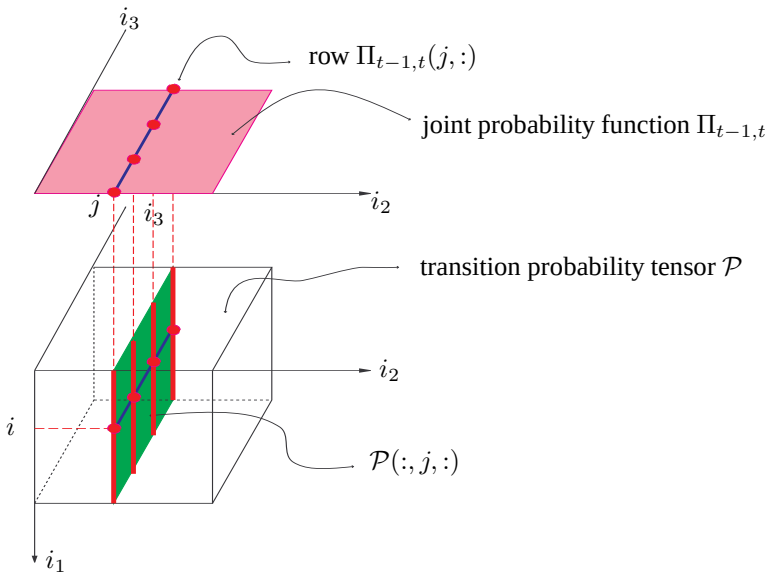

FIGURE 3.1. *Update of joint probability function* $\Pi_{t+1,t}$ *from* $\Pi_{t,t-1}$.

More specifically, we obtain an entry-wise expression

$$\pi^{(t+1)}_{i_1\dots i_{m-1}} = \sum_{i_m=1}^{n} p_{i_1 i_2 \dots i_m} \pi^{(t)}_{i_2\dots i_m} \tag{3.9}$$

and the process can move on.

For a Markov chain with memory 2, the relation (3.9) can be expressed as

$$\Pi_{t+1,t} = [\mathcal{P}(:,1,:)\Pi_{t,t-1}(1,:)^\top, \dots, \mathcal{P}(:,n,:)\Pi_{t,t-1}(n,:)^\top] \tag{3.10}$$

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.

Note that the operation specified in (3.9) is not the usual mode-p tensor product defined in the literature [6]. We denote this transition of the joint probability density function by the symbol

$$\Pi_{t+1,t,\dots,t-m+3} = \mathcal{P} \boxdot \Pi_{t,t-1,\dots,t-m+2}. \tag{3.11}$$

The process (3.11) for the joint probability density function is analogous to that of (3.7) for state distribution. Indeed, it is necessary to understand the limiting behavior of the joint probability density function before the limiting behavior of the sequence $\{\mathbf{x}^{(t)}\}$ can be characterized. So far as we know, an analysis based directly on these processes has not been completely available yet. We shall present some partial results in this paper.

3.2. Power method for joint probability density functions. In this subsection, we elaborate further on the limiting behavior of the joint probability density function. It will be instructive if we first consider the Markov chain with memory 2. With the drawing in Figure 3.1, we might have a better grasp of the dynamics.

As the step t evolves, we have two sequences of probability distributions to be processed hand by hand. First, we have a distribution $\Pi_{t,t-1} = [\pi_{i_2 i_3}]$ of memories (X_t, X_{t-1}) over $S \times S$. Second, a distribution $\mathbf{x}^{(t+1)}$ for X_{t+1} over S based on this memory is defined via

$$\mathbf{x}^{(t+1)} = \mathcal{P} \circledast_1 \Pi_{t,t-1}. \tag{3.12}$$

In the meantime, the memory is also evolved into

$$\Pi_{t+1,t} = \mathcal{P} \boxdot \Pi_{t,t-1}. \tag{3.13}$$

The mechanism for computing the distribution $\mathbf{x}^{(t+1)}$ for the random variable X_{t+1} can be thought of as taking the Frobenius inner product of the matrix $\Pi_{t,t-1}$ (plotted as the separated horizontal plane at the top of Figure 3.1) with each (horizontal) cross section of the tensor $\mathcal{P}$ in the 1st direction. That is, the probability of being moved into state i_1 at step $t+1$ is given by

$$x_{i_i}^{(t+1)} = \sum_{i_2=1}^{n} \sum_{i_3=1}^{n} p_{i_1 i_2 i_3} \pi_{i_2 i_3}^{(t)}. \tag{3.14}$$

Similarly, the probability of having memory $X_{t+1} = i_1, X_t = i_2$ at the step $t+1$ is given by

$$\pi_{i_1 i_2}^{(t+1)} = \sum_{i_3=1}^{n} p_{i_1 i_2 i_3} \pi_{i_2 i_3}^{(t)}, \tag{3.15}$$

which is the inner product of the i_2-th row of $\Pi_{t,t-1}$ (plotted as the horizontal bar in the plane $\Pi_{t,t-1}$ of Figure 3.1) with each row of the (vertical) cross section of the tensor $\mathcal{P}$ in the 2nd direction. Clearly, we have the relationship

$$x_{i_1}^{(t+1)} = \sum_{i_2=1}^{n} \pi_{i_1 i_2}^{(t+1)}. \tag{3.16}$$

Because each term $\pi_{i_1 i_2}^{(t+1)}$ in (3.16) is nonnegative, we see that $\mathbf{x}^{(t+1)}$ converges if and only if $\Pi_{t+1,t}$ converges as t goes to infinity. If suffices to consider the limiting behavior of the iteration (3.15). For each fixed i_2, we may rewrite this updating mechanism in the matrix-to-vector multiplication form

$$\Pi_{t+1,t}(:, i_2) = \mathcal{P}(:, i_2, :)\Pi_{t,t-1}(i_2, :)^{\top}. \tag{3.17}$$

This scheme is *not* exactly the ordinary power method applied to the matrix $\mathcal{P}(:, i_2, :)$ because to execute the iteration (3.17) we must know the i_2-th row of $\Pi_{t,t-1}$. In other words, the iterations (3.17) under the operation $\boxdot$ must be carried out simultaneously for all $i_2 = 1, \ldots, n$.

We can rewrite the iteration (3.13) in the following way. Let $\mathbf{vec}(M)$ denote the vectorization of the matrix M formed by stacking the columns of M into a single column vector and C the $n^2 \times n^2$ permutation matrix that does the index swapping

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

Also, let B be 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

$$\mathbf{vec}(\Pi_{t+1,t}) = BC\mathbf{vec}(\Pi_{t,t-1}), \tag{3.18}$$

which is exact 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

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

Note that, by (3.4), $\mathcal{A}$ is itself column stochastic. Also, by (3.6), $\mathbf{vec}(\Pi_{t,t-1})$ is itself a distribution vector. From this point on, everything follows from our understanding about the conventional power method. We summarize the results as follows.

THEOREM 3.1. *Suppose that $\mathcal{P}$ is the transition probability tensor of a Markov chain with memory 2 such that the Perron-Frobenius eigenvalue $\lambda_1(\mathcal{A}) = 1$ of the corresponding $\mathcal{A}$ is simple. Then, starting with any generic initial memory distribution $\Pi_{0,-1}$, it is true that*

1. *The convergence of the joint probability density functions generated by (3.13) is guaranteed.*
2. *The rate of convergence is the second dominant eigenvalue $|\lambda_2(\mathcal{A})|$ of the matrix $\mathcal{A}$.*
3. *The limit point $\widetilde{\Pi}$ of joint probability density functions is the de-vectorization of the normalized dominant eigenvector of $\mathcal{A}$ under the 1-norm.*
4. *The stationary distribution $\widetilde{\mathbf{x}}$ of the states under this Markov chain (3.12) with memory 2 exists and is the row sum of the limiting joint probability density function.*

3.3. Higher memory. For Markov chains with memories higher than 2, the convergence of the sequence $\{\Pi_{t,t-1,\ldots,t-m+2}\}$ generated by (3.11) can still be guaranteed because, with some tedious rearrangement, the linear relationship (3.9) can still be written in terms of a matrix-to-vector multiplication. The difficulty is at the complicated description of the corresponding matrix $\mathcal{A}$ whose structure depends upon how the tensor $\Pi_{t,t-1,\ldots,t-m+2}$ is vectorized.

It is worth noting that in the above context, the Markov process $\mathcal{P}$ acting on the joint probability density function is really a linear map from the space $\mathbb{T}^{m-1}$ of order-$(m-1)$ tensors to $\mathbb{T}^{m-1}$. As such, the notion of eigenvalue for $\mathcal{P}$ is no different from the ordinary notion of eigenvalue for square matrices.

4. Power method for Z-eigenvector computation. The proceeding discussion is a formal way to calculate the probability distribution $\mathbf{x}^{(t+1)}$ based on the relationship (3.7). The stationary distribution of the states follows from the same relationship only after the limiting joint probability function is known. One alternative approach to circumvent this complication is to assume that a limiting joint probability distribution of the high-order Markov chain is the Kronecker product of its limiting probability distribution. The rationale is that, if the sequence $\{\mathbf{x}^{(t)}\}$ has reached a stationary distribution $\widetilde{\mathbf{x}}$ over S, then it is reasonable to think of the steady-state probability among states are independent of each other. As such, a joint probability density function should be of the form

$$\lim_{t\to\infty} \Pi_{t,t-1,\ldots,t-m+2} = \underbrace{\widetilde{\mathbf{x}} \otimes \widetilde{\mathbf{x}} \otimes \ldots \otimes \widetilde{\mathbf{x}}}_{m-1 \text{ times}}. \tag{4.1}$$

Thus, referring back to (3.7), the stationary distribution $\widetilde{\mathbf{x}}$ should satisfy the equation

$$\mathcal{P} \circledast_1 \mathbf{z} \otimes \mathbf{z} \otimes \ldots \otimes \mathbf{z} = \mathbf{z}, \tag{4.2}$$

which is conveniently abbreviated as

$$\mathcal{P}\mathbf{z}^{m-1} = \mathbf{z} \tag{4.3}$$

in the literature and its solution is called the Z-eigenvector associated with the Z-eigenvalue 1 [9, 10, 15, 2]. It can be shown that solutions of (4.3) do exist and that all the entries of solutions are positive, if $\mathcal{P}$ is irreducible [8, Theorem 2.2]. Under some additional conditions on $\mathcal{P}$, the solution is even unique [8, Theorem 2.4].

We do want to point out that the assumption (4.1) is dubious as it inadvertently implies that the limiting distribution of memories is of rank one. We shall give numerical evidence to show that it is not the case in general. Consequently, the stationary distribution $\widetilde{\mathbf{x}}$ from (3.7) does not satisfy (4.3). Regardless, solving the equation (4.3) is of mathematical interest in its own right. Thus, it might be worth continuing to discuss the power method applied to $\mathcal{P}$ in the sense of (4.3) for computing the dominant Z-eigenvector which we denote by $\widetilde{\mathbf{z}}$. We should also point out that Z-eigenvectors are not scaling invariant. So cares 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 does not cause a concern.

Quite a few methods have been proposed for computing eigenpairs of a tensor [4, 5, 8, 11, 13, 17, 18, 19], depending on which definition of eigenpairs is of interest. Motivated by (1.1), an iterative scheme[3]

$$\mathbf{z}_{k+1} := \mathcal{P}\mathbf{z}_k^{m-1} \tag{4.4}$$

starting with a prescribed probability vector $\mathbf{z}_0$, is perhaps the simplest means for finding the Z-eigenvector $\widetilde{\mathbf{z}}$ in (4.3). Note that each $\mathbf{z}_{k+1}$ remains to be a probability vector under exact arithmetic. The sequence $\{\mathbf{z}_k\}$ does converge linearly to a solution of (4.3) under certain conditions.

Since the power method for matrices is affected by the second dominant eigenvalue of the underlying matrix A, we are curious to know what part of $\mathcal{P}$ affects the rate of convergence of (4.4), if it converges at all. By casting such a power 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. In the following, we employ a technique similar to the preceding section to analyze the power iteration for tensors. We specifically work on the transition probability tensor $\mathcal{P}$, though the idea works in general. Our main point is to show that for tensors the "second eigenvalue" comes into play in a far more complicated way.

4.1. Attribute of the second dominant eigenvalue. Let $\triangle^{n-1}$ denote the standard simplex in $\mathbb{R}^n$, that is,.

$$\triangle^{n-1} = \{\mathbf{z} \in \mathbb{R}^n | x_i \geq 0, \text{ and } \sum_{i=1}^{n} x_i = 1\}. \tag{4.5}$$

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

$$\mathbf{f}(\mathbf{z}) = \frac{\mathcal{P}\mathbf{z}^{m-1}}{\langle \mathcal{P}\mathbf{z}^{m-1}, \mathbf{1}\rangle}, \tag{4.6}$$

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, so 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 (4.4) converges to such a fixed point.

We have already introduced one kind of tensor product $\circledast_1$ in (3.7), namely,

$$\mathcal{P} \circledast_1 \underbrace{\mathbf{z} \otimes \cdots \otimes \mathbf{z}}_{m-1 \text{ times}} = \mathcal{P}\mathbf{z}^{m-1} := \left[\sum_{i_2,\ldots,i_m=1}^{n} p_{\nu_1 i_2,\ldots i_m} x_{i_2} \cdots x_{i_m} \right]_{\nu_1=1}^{n}, \tag{4.7}$$

where the subscript in $\circledast_1$ indicates that the first index in $\mathcal{P}$ is excluded from the summation. We identify this first index by the dummy variable ν_1, so this product ends up with a column vector. Likewise, we now introduce another kind of tensor product $\circledast_{1\ell}$ that will occur in the following analysis. Specifically,

$$\mathcal{P} \circledast_{1\ell} \underbrace{\mathbf{z} \otimes \cdots \otimes \mathbf{z}}_{m-2 \text{ times}} := \left[\sum_{i_2,\ldots,\widehat{i_\ell},\ldots,i_m=1}^{n} p_{\nu_1 i_2 \ldots \nu_\ell \ldots i_m} x_{i_2} \cdots \widehat{x_{i_\ell}} \cdots x_{i_m} \right]_{\nu_1,\nu_\ell=1}^{n}, \tag{4.8}$$

where $\widehat{i_\ell}$ means that quantities associated with this particular index are taken out from the remaining list. Note that the double subscript in $\circledast_{1\ell}$ indicates that the first 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 identified by the double index (ν_1, ν_ℓ). It is easy to verify that the important relationship

$$\left(\mathcal{P} \circledast_{1,\ell} \mathbf{z}^{m-2}\right) \mathbf{h} = \mathcal{P} \circledast_1 \mathbf{z}^{\ell-2} \otimes \mathbf{h} \otimes \mathbf{z}^{m-\ell} \tag{4.9}$$

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

holds for any give $\mathbf{h} \in \mathbb{R}^n$. When $\ell = 3$, for example, we have

$$\mathcal{P} \circledast_1 \mathbf{z} \otimes \mathbf{h} \otimes \mathbf{z} \otimes \cdots \otimes \mathbf{z} := (\mathcal{P} \circledast_{13} \mathbf{z} \otimes \cdots \otimes \mathbf{z})\mathbf{h}.$$

Similar to the local analysis developed earlier for matrices, our first task for tensors is to calculate the Jacobian matrix $D\mathbf{f}(\mathbf{z})$. Toward this goal, the Fréchet derivative $\mathbf{f}'$ at $\mathbf{z} \in \triangle^{n-1}$ acting on an arbitrary $\mathbf{h} \in \mathbb{R}^n$ is somewhat easier to manipulate by the generalized Leibniz product rule,

$$\left(\mathcal{P}\mathbf{z}^{m-1}\right)'.\mathbf{h} = \mathcal{P} \circledast_1 \mathbf{h} \otimes \mathbf{z}^{m-2} + \mathcal{P} \circledast_1 \mathbf{z} \otimes \mathbf{h} \otimes \mathbf{z}^{m-3} + \ldots + \mathcal{P} \circledast_1 \mathbf{z}^{m-2} \otimes \mathbf{h}. \tag{4.10}$$

By using (4.9), we can represent the action of the derivative operator in terms of matrix-to-vector multiplication:

$$D\mathbf{f}(\mathbf{z})\mathbf{h} = \left(\frac{(\sum_{\ell=2}^m \mathcal{P} \circledast_{1\ell} \mathbf{z} \otimes \cdots \otimes \mathbf{z})\langle \mathcal{P}\mathbf{z}^{m-1}, \mathbf{1}\rangle - \mathcal{P}\mathbf{z}^{m-1}\mathbf{1}^\top (\sum_{\ell=2}^m \mathcal{P} \circledast_{1\ell} \mathbf{z} \otimes \cdots \otimes \mathbf{z})}{\langle \mathcal{P}\mathbf{z}^{m-1}, \mathbf{1}\rangle^2} \right) \mathbf{h} \tag{4.11}$$

and thus retrieve the Jacobian information.

At a fixed point $\widetilde{\mathbf{z}} \in \triangle^{n-1}$, the equation (4.3) is satisfied and the corresponding Jacobian matrix is reduced to the matrix

$$D\mathbf{f}(\widetilde{\mathbf{z}}) = (I - \widetilde{\mathbf{z}}\mathbf{1}^\top) \underbrace{\left(\sum_{\ell=2}^m \mathcal{P} \circledast_{1\ell} \widetilde{\mathbf{z}} \otimes \cdots \otimes \widetilde{\mathbf{z}} \right)}_{\Omega}. \tag{4.12}$$

Each term $\mathcal{P} \otimes_{1\ell} \widetilde{\mathbf{z}} \otimes \cdots \otimes \widetilde{\mathbf{z}}$ in the summation for Ω is itself column stochastic. Furthermore,

$$\Omega\widetilde{\mathbf{z}} = \left(\sum_{\ell=2}^m \mathcal{P} \circledast_{1\ell} \widetilde{\mathbf{z}} \otimes \cdots \otimes \widetilde{\mathbf{z}} \right) \widetilde{\mathbf{z}} = \sum_{\ell=2}^m \mathcal{P}\widetilde{\mathbf{z}}^{m-1} = (m-1)\widetilde{\mathbf{z}}, \tag{4.13}$$

showing that $\lambda_1 = m - 1$ is the dominant eigenvalue of Ω with the right eigenvector $\widetilde{\mathbf{z}}$. It follows that $D\mathbf{f}(\widetilde{\mathbf{z}})$ an eigenvalue 0.

The Jacobian matrix in (4.12) is analogous to that in (2.5). In particular, the matrix Ω in (4.12) plays the same role as the matrix A in (2.5). Suppose $\mathbf{w}_i \in \mathbb{C}^n$ is an eigenvector of $\Omega^\top$ with eigenvalues $\lambda_i \in \mathbb{C}$, $i = 2, \ldots, n$. If Ω is positive, the we have $|\lambda_i| < m - 1$ and $\mathbf{w}^\top\widetilde{\mathbf{z}} = 0$. It follows that

$$\mathbf{w}^\top D\mathbf{f}(\widetilde{\mathbf{z}}) = \mathbf{w}^\top(I - \widetilde{\mathbf{z}}\mathbf{1}^\top)\Omega = \mathbf{w}^\top\Omega = \lambda_i\mathbf{w}^\top, \tag{4.14}$$

implying that the Jacobian matrix $D\mathbf{f}(\widetilde{\mathbf{z}})$ has eigenvalues 0 and those of Ω with modulus less than $m-1$. We have thus reached the following conclusion.

THEOREM 4.1. *The limiting behavior of the iteration by the power method (4.4) is determined by the second dominant eigenvalue of the matrix Ω defined in (4.12). If the iteration converges at all, then the rate of convergence is $|\lambda_2(\Omega)|$.*

In the power method for matrices, the second dominant eigenvalue of the matrix A alone affects the limiting behavior. In the power method (4.4) for the tensors, it is the second dominant eigenvalue of the matrix Ω that affects the convergence. Take notice of the summation in (4.12) for defining the matrix Ω. Such a combination by running ℓ through different facets of $\mathcal{P}$ is far more complicated. Being able to pinpoint this cause of convergence or divergence is an interesting result in its own right.

4.2. Examples of divergence. We have already pointed out that $\lambda_1(\Omega) = m - 1$. For convergence, we need $|\lambda_2(\Omega)| < 1$. It becomes suspicious that the gap between these two dominant eigenvalues can be always so wide. In this section, we give a family of examples of a transition probability tensor showing that $|\lambda_2(\Omega)| > 1$ and hence the power method does not converge.

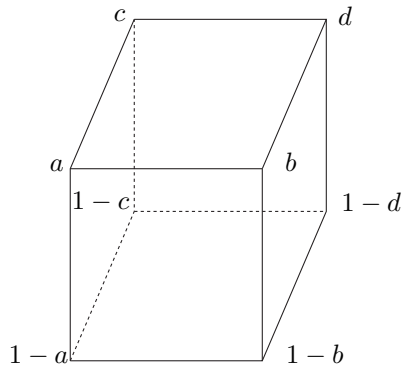

FIGURE 4.1. *Order-3 transition probability tensor over 2 states.*

Consider an order-3 transition probability tensor $\mathcal{P}$ over $S = \{1, 2\}$ depicted in Figure 4.1 where $0 \leq a, b, c, d \leq 1$ and $a + d \neq b + c$. Write its dominant eigenvector $\widetilde{\mathbf{z}} = [z, 1 - z]^\top$. Then the equation (4.3) 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

$$z = \frac{2d + 1 - b - c \pm \sqrt{(b + c - 1)^2 + 4d(1 - a)}}{2(a - b - c + d)}. \tag{4.15}$$

Depending on the values of a, b, c, d, we are interested in the root satisfying $0 \leq z \leq 1$. The corresponding Ω is given by

$$\Omega = \mathcal{P} \otimes_{12} \widetilde{\mathbf{z}} + \mathcal{P} \otimes_{13} \widetilde{\mathbf{z}} = \begin{bmatrix} b + c + (2a - b - c)z & 2d + (b + c - 2d)z \\ (-2a + b + c)z + 2 - b - c & (2d - b - c)z - 2d + 2 \end{bmatrix} \tag{4.16}$$

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 a numerical 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 the power method cannot generate the limiting stationary distribution vector $\widetilde{\mathbf{z}}$ because $|\lambda_2| > 1$. Indeed, our numerical experiment indicates that the iterates generated by the power 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.

In fact, using continuity argument, we can see that there exists a set of positive transition probability tensors with nonzero measure for which the power method will not converge. For instance, 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 $0 \leq \epsilon < \frac{3-\sqrt{7}}{4}$.

4.3. Deviation from true stationary distribution. We have suggested earlier that although the dominant Z-eigenvector computation of a transition probability tensor $\mathcal{P}$ is of mathematical interest, to rationalize its application via the assumption (4.1) might need further justification. In this subsection, we give numerical evidence to show the deviation of results based on this assumption from the true Markov process.

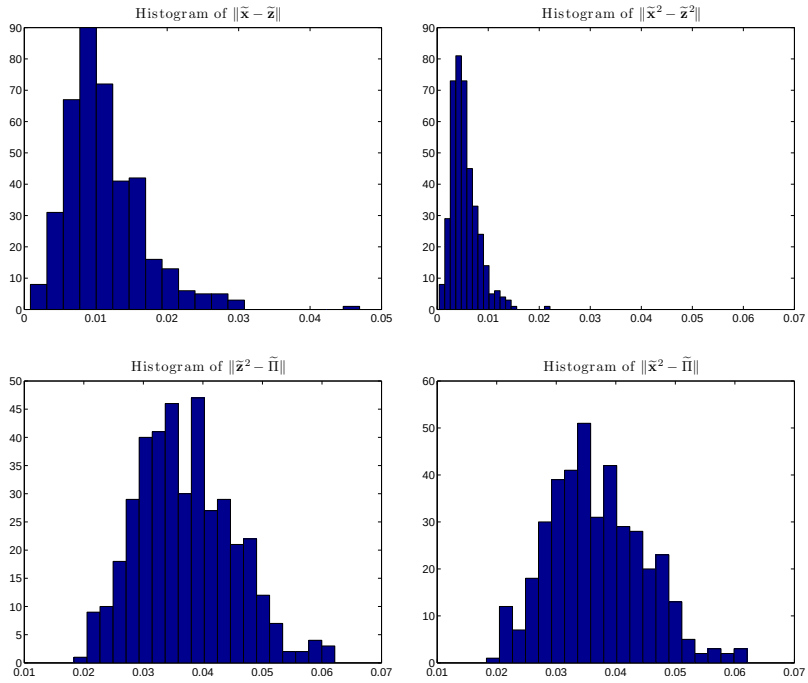

FIGURE 4.2. *Comparisons between the stationary distribution $\widetilde{\mathbf{z}}$ and the dominant Z-eigenvector $\widetilde{\mathbf{x}}$, and the corresponding distributions of memories.*

We randomly generate 400 test data over $\mathbb{R}^5$. Each data set includes one order-3 transition probability tensor $\mathcal{P}$, two starting distribution vectors $\mathbf{x}_{-1}$ and $\mathbf{x}_0$, and one order-2 tensor $\Pi_{0,-1} = \mathbf{x}_{-1} \otimes \mathbf{x}_0$ representing the joint probability density function for the starting memory. 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. After going through the calculation, the limiting joint probability density function is denoted by $\widetilde{\Pi}$, the stationary distribution by $\widetilde{\mathbf{x}}$, and dominant Z-eigenvector by $\widetilde{\mathbf{z}}$. We compare the histograms of $\|\widetilde{\mathbf{x}} - \widetilde{\mathbf{z}}\|$, $\|\widetilde{\mathbf{x}}^2 - \widetilde{\Pi}\|$, $\|\widetilde{\mathbf{z}}^2 - \widetilde{\Pi}\|$, and $\|\widetilde{\mathbf{x}}^2 - \widetilde{\mathbf{z}}^2\|$, all measured in the 2-norm. The results are plotted in Figure 4.2. As can be seen, the variations $\|\widetilde{\mathbf{z}} - \widetilde{\mathbf{x}}\|$ and $\|\widetilde{\mathbf{z}}^2 - \widetilde{\mathbf{x}}^2\|$ shown in the upper drawing does seem to suggest that the two distributions $\widetilde{\mathbf{z}}$ and $\widetilde{\mathbf{x}}$ might be called statistically close [1]. However, the variations in the lower drawing indicate that the difference between $\widetilde{\Pi}$ and $\widetilde{\mathbf{z}}^2$ is statistically more significant.

5. Conclusions. In this paper, we cast the power method as fixed-point iteration of some properly defined mappings. Limiting behavior of iterates generated by the power method can be understood from the spectrum of the corresponding Jacobian matrices at the fixed-point. When the power method is applied to a matrix A, the ratio $\left|\frac{\lambda_2}{\lambda_1}\right|$ does show up as the dominant eigenvalue of the Jacobian matrix, reconfirming the known fact that the second dominant eigenvalue λ_2 of A affects the rate of convergence. When the power method is applied to a transition probability tensor of order m, two types of power methods are involved — one iterates on the joint probability density functions, which is essentially the same as the conventional power method for computing the dominant eigenvector; and the other iterates on the distribution vectors, which amounts to computing the dominant Z-eigenvector. We identify two specially structured matrices $\mathcal{A}$ and Ω for the two power methods, respectively. The second dominant eigenvalue will affect the local behavior nearby a fixed-point of the corresponding power method. In contrast to the matrices, it is possible that the second dominant eigenvalue of Ω has modulus greater than one and, hence, the fixed-point is a repeller.

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

[7] 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.

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

[9] 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.

[10] 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.

[11] 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.

[12] C. MEYER, *Matrix analysis and applied linear algebra*, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2000. With 1 CD-ROM (Windows, Macintosh and UNIX) and a solutions manual (iv+171 pp.).

[13] 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.

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

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

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

[17] 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.

[18] 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.

[19] ———, *On the largest eigenvalue of a symmetric nonnegative tensor*, Numer. Linear Algebra Appl., 20 (2013), pp. 913–928.