Automatica 48 (2012) 2512–2519

Contents lists available at SciVerse ScienceDirect

Automatica

journal homepage: www.elsevier.com/locate/automatica

Brief paper

Oja's algorithm for graph clustering, Markov spectral decomposition, and risk sensitive control☆

V. Borkar [a], S.P. Meyn [b,1]

[a] Department of Electrical Engineering, Indian Institute of Technology, Powai, Mumbai 400076, India

[b] Department of Electrical and Computer Engineering and the Coordinated Sciences Laboratory, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

ARTICLE INFO

Article history:
Received 1 October 2010
Received in revised form
19 October 2011
Accepted 3 January 2012
Available online 24 July 2012

Keywords:
Graph algorithms
Oja's algorithm
Stochastic approximation
Markov chains
Spectral theory of Markov chains
Multiplicative ergodic theory
Risk sensitive control

ABSTRACT

Given a positive definite matrix M and an integer $N_m \geq 1$, Oja's subspace algorithm will provide convergent estimates of the first N_m eigenvalues of M along with the corresponding eigenvectors. It is a common approach to principal component analysis. This paper introduces a normalized stochastic-approximation implementation of Oja's subspace algorithm, as well as new applications to the spectral decomposition of a *reversible* Markov chain. Recall that this means that the stationary distribution satisfies the detailed balance equations (Meyn & Tweedie, 2009). Equivalently, the statistics of the process in steady state do not change when time is reversed. Stability and convergence of Oja's algorithm are established under conditions far milder than that assumed in previous work. Applications to graph clustering, Markov spectral decomposition, and multiplicative ergodic theory are surveyed, along with numerical results.

1. Introduction

Spectral decomposition is a classical approach to model reduction for systems that are complex due to dimension or randomness. This technique is known as principal component analysis or the Karhunen–Loève decomposition, depending on the context (Hyvärinen, 1999; Jolliffe, 2002; Loève, 1978). The same technique has been developed more recently for network decomposition (Nadler, Lafon, Coifman, & Kevrekidis, 2006; Schölkopf, Smola, & Müller, 1998; Weiss, 1999), which in particular provides an appealing alternative to the min-cut max-flow theorem.

Given a symmetric $N \times N$ matrix w, its spectral decomposition amounts to the computation of its N real eigenvalues and corresponding eigenvectors. In the Karhunen–Loève decomposition the matrix w is a covariance matrix, and the decomposition leads to a representation of a stationary process as a moving-average of white noise. In the graph clustering problem the elements of this matrix represent positive edge weights: $w_{ij} = w_{ji}$ is the weight of the link connecting nodes i and j. The first decomposition of a connected graph is obtained by computation of the eigenvector corresponding to the second eigenvalue. It can be shown that the eigenvector possesses positive and negative entries, and this sign structure is used to define a generalized network cut in Nadler et al. (2006), Schölkopf et al. (1998) and Weiss (1999).

Oja's subspace algorithm is an approach to computation of the leading eigenvalues and eigenvectors of the matrix w (Chen, Hua, & Yan, 1998; Oja, 1982; Sikora & Skarbek, 2009). Fix an integer $N_m \leq N$, and let $m(t)$ denote an $N \times N_m$ matrix whose columns are intended to approximate an N_m-dimensional eigenspace corresponding to the N_m largest of the N eigenvalues of w. A deterministic version of Oja's algorithm is expressed as the polynomial differential equation:

$$\frac{d}{dt}m(t) = [I - m(t)m^{\mathrm{T}}(t)]wm(t), \tag{1}$$

☆ S.M. was partially supported by the National Science Foundation under grant ECS-0523620, and by AFOSR grant FA9550-09-1-0190. V.B. was supported in part by a J. C. Bose Fellowship of Dept. of Science and Technology, Govt. of India, and a grant from General Motors India Science Lab. The material in this paper was partially presented at the ValueTools'08 Proceedings of the 3rd International Conference on Performance Evaluation Methodologies and Tools, October 20–24, 2008, Athens, Greece. This paper was recommended for publication in revised form by Associate Editor Valery Ugrinovskii under the direction of Editor Ian R. Petersen. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of their sponsors.

E-mail addresses: borkar.vs@gmail.com (V. Borkar), meyn@illinois.edu (S.P. Meyn).

[1] Tel.: +1 217 244 1782; fax: +1 217 333 3600.

0005-1098/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.automatica.2012.05.016

where $m(0)$ is given as the initial condition. If the matrix w is positive definite then the analysis of Chen et al. (1998) establishes convergence of $\boldsymbol{m}$ for almost every initial conditions.

This paper introduces a normalized implementation of Oja's algorithm that is also a multidimensional generalization of the one-dimensional algorithm of Krasulina (Krasulina, 1970). Stability and convergence of the normalized algorithm are established under conditions far milder than that assumed in previous work. Applications to graph clustering, as well as new applications to the spectral decomposition of a reversible Markov chain are surveyed.

In the following section, we introduce the stochastic approximation algorithm, and present the main result establishing convergence of the algorithm. Applications to spectral graph theory are surveyed in Section 3, and Section 4 contains extensions of the algorithm to compute the spectrum of a reversible Markov chain. Section 5 shows connections to multiplicative ergodic theory and risk-sensitive control where the interest is in the top eigenvalue and eigenvector. Examples are contained in Section 6, and conclusions may be found in Section 7.

2. Stochastic approximation and Oja's algorithm

Oja's 1985 paper (Oja & Karhunen, 1985) introduces a stochastic approximation algorithm based on the o.d.e. (1). Suppose that $\boldsymbol{X}$ is an $\mathbb{R}^n$-valued stationary process with covariance matrix $w = \mathsf{E}[X(t)X(t)^{\mathrm{T}}]$. We can express Oja's stochastic approximation algorithm as the matrix recursion:

$$M(n+1) - M(n) = a(n)\big[I - M(n)M^{\mathrm{T}}(n)\big]\widehat{W}(n)M(n), \tag{2}$$

where $\widehat{W}(n) = X(n)X^{\mathrm{T}}(n)$, and $a(n)$ is a decreasing parameter—the step-size for the algorithm (Borkar, 2008b). Specific assumptions will be imposed later. Almost sure convergence to the appropriate dimensional dominant eigenspace was established by applying stochastic approximation techniques that were available at the time. These techniques require Lipschitz continuity of the right hand side of the recursion in the variable $M(n)$, which is violated in this recursion. This issue is addressed in Oja and Karhunen (1985) and in Sikora and Skarbek (2009), by imposing additional conditions on $\boldsymbol{X}$.

The lack of Lipschitz continuity presents problems even in deterministic approximations of (1) in discrete time. One such algorithm is introduced in Yi, Ye, Lv, and Tan (2005) through sampling the o.d.e. to obtain the deterministic recursion,

$$m(n+1) - m(n) = a(n)\big[I - m(n)m^{T}(n)\big]wm(n). \tag{3}$$

While convergence is established for the deterministic algorithm, the proof is complex. Complexity is due in large part to the cubic nonlinearity seen here just as in the stochastic approximation algorithm.

To obtain an algorithm that satisfies the Lipschitz continuity and thereby place the algorithm within the framework of Borkar (2008a,b) and Borkar and Meyn (2000) we introduce a normalization. The normalized o.d.e. is given by

$$\begin{aligned} \frac{d}{dt}m(t) &= a(t)\big[I - m(t)m^{T}(t)\big]wm(t), \\ a(t) &= [1 + \operatorname{trace}(m(t)m(t)^{\mathrm{T}})]^{-1}. \end{aligned} \tag{4}$$

The right hand side of the differential equation is Lipschitz in the variable $m(t)$. Solutions to this differential equation are simply time-scaled versions of the solutions to (1). In particular, from each initial condition the set of limit points are identical.

The stochastic approximation algorithm considered in this paper is again of the form (2) in which the gain sequence is modified as in the o.d.e. (4), with an additional scaling as follows:

$$a(n) = b(n)(1 + \operatorname{trace}(M(n)M(n)^{\mathrm{T}}))^{-1}. \tag{5}$$

It is assumed throughout that the following assumptions hold for the sequence $\boldsymbol{b} = \{b(n) : n \geq 0\}$. It is non-negative, with

$$\begin{aligned} &\sum_{n=0}^{\infty} b(n) = \infty, \qquad \sum_{n=0}^{\infty} b(n)^2 < \infty, \\ &\sup_{n\geq 0}\left(\frac{\sum_{k\geq n} b(k)^2}{b(n)}\right) < \infty. \end{aligned} \tag{6}$$

An example is $b(n) = (1+n)^{-1}$, $n \geq 0$.

Under these conditions the algorithm is stable. To guarantee consistency we modify the algorithm slightly through the introduction of white noise:

$$\begin{aligned} M(n+1) - M(n) = a(n)\big(&\big[I - M(n)M^{\mathrm{T}}(n)\big] \\ &\times \widehat{W}(n)M(n) + \xi(n+1)\big), \end{aligned} \tag{7}$$

where $\boldsymbol{\xi}$ is an i.i.d. $N(0, I)$ sequence. Proposition 2.1 states that this recursion shares the best possible convergence properties observed in the o.d.e. (1). While the deterministic algorithm can become trapped in an arbitrary eigenspace of w, the stochastic algorithm (7) is strongly consistent from each initial condition.

While the above result is stated for i.i.d. $\boldsymbol{X}$, Proposition 2.1 extends to ergodic Markov $\boldsymbol{X}$ as well, see, e.g., Corollary 8 and Theorem 9, p. 74–75, of Borkar (2008b).

Proposition 2.1. *Consider the algorithms (2) or (7), where $\boldsymbol{a}$ is given in (5), and with $\boldsymbol{b}$ satisfying the conditions in (6). Suppose that the process $\boldsymbol{X}$ is i.i.d., with covariance $w > 0$, and that it is independent of the i.i.d. $N(0, I)$ sequence $\boldsymbol{\xi}$.*

Then, the following conclusions hold for each initial $M(0)$:

(i) Stability*: For either of the algorithms (2) or (7),*

$$\limsup_{n\to\infty} \|M(n)\| < \infty \quad a.s.$$

(ii) Convergence*: For the algorithm (7), with probability one, any limit point $M(\infty)$ of the sequence of matrices $\{M(n)\}$ has columns that lie in the eigenspace spanned by the first m eigenvalues of w.*

Proof. First we establish that the solutions to either stochastic approximation recursion are bounded a.s. by applying Theorem 7 of Borkar (2008b, Chapter 3) (see also Borkar & Meyn, 2000). This result constructs an "o.d.e. at infinity" that approximates the behavior of the recursion for large initial conditions. Based on the recursion (2) or (7) we obtain the o.d.e.,

$$\frac{d}{dt}m^{\infty}(t) = -\left[\frac{m^{\infty}(t)m^{\infty T}(t)}{\operatorname{trace}(m^{\infty}(t)m^{\infty}(t)^{\mathrm{T}})}\right]wm^{\infty}(t), \tag{8}$$

where $m^{\infty}(0) \in \mathbb{R}^{N\times N_m}$ is given as initial condition. Define the real valued function $V\colon \mathbb{R}^{N\times N_m} \to \mathbb{R}_+$ as the quadratic:

$$V(m) := \operatorname{trace}(m^{\mathrm{T}}wm), \quad m \in \mathbb{R}^{N\times N_m}.$$

Under the positivity assumption on w this function vanishes only when m is identically zero. This property combined with the following drift condition implies that V serves as a Lyapunov function:

$$\frac{d}{dt}V(m^{\infty}(t)) = -2\left[\frac{\operatorname{trace}([m^{\infty T}(t)wm^{\infty}(t)]^2)}{\operatorname{trace}(m^{\infty}(t)m^{\infty}(t)^{\mathrm{T}})}\right] < 0,$$