The ODE Method and Spectral Theory of Markov Operators

Jianyi Huang[1], Ioannis Kontoyiannis[2], and Sean P. Meyn[1]

[1] University of Illinois at Urbana-Champaign, Urbana IL 61801, USA
[2] Brown University, Box F, 182 George St., Providence, RI 02912, USA

Abstract. We give a development of the ODE method for the analysis of recursive algorithms described by a stochastic recursion. With variability modeled via an underlying Markov process, and under general assumptions, the following results are obtained:

(i) Stability of an associated ODE implies that the stochastic recursion is stable in a strong sense when a gain parameter is small.
(ii) The range of gain-values is quantified through a spectral analysis of an associated linear operator, providing a non-local theory, even for nonlinear systems.
(iii) A second-order analysis shows precisely how variability leads to sensitivity of the algorithm with respect to the gain parameter.

All results are obtained within the natural operator-theoretic framework of geometrically ergodic Markov processes.

1 Introduction

Stochastic approximation algorithms and their variants are commonly found in control, communication and related fields. Popularity has grown due to increased computing power, and the interest in various 'machine learning' algorithms [6,7,12]. When the algorithm is linear, then the error equations take the following linear recursive form,

$$X_{t+1} = [I - \alpha M_t]\, X_t + W_{t+1}, \qquad (1)$$

where $\boldsymbol{X} = \{X_t\}$ is an error sequence, $\boldsymbol{M} = \{M_t\}$ is a sequence of $k \times k$ random matrices, $\boldsymbol{W} = \{W_t\}$ is a random "disturbance" or "noise", $\alpha \geq 0$ is a fixed constant, and I is the $k \times k$ identity matrix.

An important example is the LMS (least mean square) algorithm. Consider the discrete linear time-varying model,

$$y(t) = \theta(t)^T \phi(t) + n(t), \quad t \geq 0, \qquad (2)$$

where $\{y(t)\}$ and $\{n(t)\}$ are the sequences of (scalar) observations and noise, respectively. The vector-valued processes $\theta(t) = [\theta_1(t), \ldots, \theta_k(t)]^T$ and $\phi(t) = [\phi_1(t), \ldots, \phi_k(t)]^T$, $t \geq 0$, denote the k-dimensional regression vector and time varying parameters, respectively. These will be taken to be functions of an underlying Markov chain in the analysis that follows.

The LMS algorithm is given by the recursion

$$\hat{\theta}(t+1) = \hat{\theta}(t) + \alpha\phi(t)e(t), \tag{3}$$

where $e(t) := y(t) - \hat{\theta}(t)^T\phi(t)$, and the parameter $\alpha \in (0,1]$ is the *step size*. Hence,

$$\tilde{\theta}(t+1) = (I - \alpha\phi(t)\phi(t)^T)\tilde{\theta}(t) + [\theta(t+1) - \theta(t) - \alpha\phi(t)n(t)]\,, \tag{4}$$

where $\tilde{\theta}(t) \triangleq \theta(t) - \hat{\theta}(t)$. This is of the form (1) with $M_t = \phi(t)\phi(t)^T$, $W_{t+1} = \theta(t+1) - \theta(t) - \alpha\phi(t)n(t)$, and $X_t = \tilde{\theta}(t)$.

On iterating (1) we obtain the representation,

$$\begin{aligned} X_{t+1} &= (I - \alpha M_t)\, X_t + W_{t+1} \\ &= (I - \alpha M_t)\,[(I - \alpha M_{t-1})\, X_{t-1} + W_t] + W_{t+1} \\ &= \prod_{i=t}^{0}(I - \alpha M_i)X_0 + \prod_{i=t}^{1}(I - \alpha M_i)W_1 + \cdots + (I - \alpha M_t)W_t + W_{t+1}. \end{aligned} \tag{5}$$

From the last expression it is clear that the matrix products $\prod_{i=t}^{s}(I - \alpha M_i)$ play an important role in the behavior of (1).

Properties of products of random matrices are of interest in a wide range of fields. Application areas include numerical analysis [15,34], statistical physics [9,10], recursive algorithms [11,5,27,17], perturbation theory for dynamical systems [1], queueing theory [23], and even botany [30]. Seminal results are contained in [3,13,29,28].

A complementary and popular research area concerns the eigenstructure of *large* random matrices (see e.g. [33,16] for a recent application to capacity of communication channels). Although the results of the present paper do not address these issues, they provide justification for simplified models in communication theory, leading to bounds on the capacity for time-varying communication channels [24].

The relationship with dynamical systems theory is particularly relevant to the issues addressed here. Consider a *nonlinear* dynamical system described by the equations,

$$X_{t+1} = X_t - f(X_t, \Phi_{t+1}) + W_{t+1}\,, \tag{6}$$

where $\boldsymbol{\Phi} = \{\Phi_t\}$ is an ergodic Markov process, evolving on a state space X, and $f\colon \mathbb{R}^k \times \mathsf{X} \to \mathbb{R}^k$ is smooth and Lipschitz continuous. For this nonlinear model we can construct a random linear model of the form (1) to address many interesting issues. Viewing the initial condition $\gamma = X_0 \in \mathbb{R}^k$ as a continuous variable, we write $X_t(\gamma)$ as the resulting state trajectory, and consider the sensitivity matrix,

$$S_t = \frac{\partial}{\partial\gamma}X_t(\gamma), \qquad t \geq 0\,.$$

From (6) we have the linear recursion,

$$S_{t+1} = [I - M_{t+1}]S_t, \tag{7}$$

where $M_{t+1} = \nabla_x f\,(X_t, \Phi_{t+1})$, $t \ge 0$. If $\boldsymbol{S} = \{S_t\}$ is suitably stable then the same is true for the nonlinear model, and we find that trajectories couple to a steady state process $\boldsymbol{X}^* = \{X_t^*\}$:

$$\lim_{t\to\infty} \|X_t(\gamma) - X_t^*\| = 0\,.$$

These ideas are related to issues developed in Section 3.

The traditional analytic technique for addressing the stability of (6) or of (1) is the *ODE method* of [22]. For linear models, the basic idea is that, for small values of α, the behavior of (1) should mimic that of the linear ODE,

$$\frac{d}{dt}\gamma_t = -\alpha\overline{M}\gamma_t + \overline{W}\,, \tag{8}$$

where $\overline{M}$ and $\overline{W}$ are steady-state means of M_t and W_t, respectively. To obtain a finer performance analysis one can instead compare (1) to the linear diffusion model,

$$d\Gamma_t = -\alpha\overline{M}\Gamma_t + dB_t, \tag{9}$$

where $\boldsymbol{B} = \{B_t\}$ is a Brownian Motion.

Under certain assumptions one may show that, if the ODE (8) is stable, then the stochastic model (1) is stable in a statistical sense for a range of small α, and comparisons with (9) are possible under still stronger assumptions (see e.g. [4,8,21,20,14] for results concerning both linear and nonlinear recursions).

In [27] an alternative point of view was proposed where the stability verification problem for (1) is cast in terms of the spectral radius of an associated discrete-time semigroup of linear operators. This approach is based on the functional analytic setting of [26], and analogous techniques are used in the treatment of multiplicative ergodic theory and spectral theory in [2,18,19]. The main results of [27] may be interpreted as a significant extension of the ODE method for linear recursions.

Our present results give a unified treatment of both the linear and nonlinear models treated in [27] and [8], respectively.[1] Utilizing the operator-theoretic framework developed in [18] also makes it possible to offer a transparent treatment, and also significantly weaken the assumptions used in earlier results.

We provide answers to the following questions:

(i) For what range of $\alpha > 0$ is the random linear system (1) L_2-stable, in the sense that $\mathsf{E}_x[\|X_t\|^2]$ is bounded in t for any initial condition $\Phi_0 = x \in \mathsf{X}$?

[1] Our results are given here with only brief proof outlines; a more detailed and complete account is in preparation.

(ii) What does the averaged model (8) tell us about the behavior of the original stochastic model?

(iii) What is the impact of variability on *performance* of recursive algorithms?

2 Linear Theory

In this section we develop stability theory and structural results for the linear model (1) where $\alpha \geq 0$ is a fixed constant.

It is assumed that an underlying Markov chain $\boldsymbol{\Phi}$, with general state-space X, governs the statistics of (1) in the sense that $\boldsymbol{M}$ and $\boldsymbol{W}$ are functions of the Markov chain:

$$M_t = m(\Phi_t), \quad W_t = w(\Phi_t), \qquad t \geq 0\,. \tag{10}$$

We assume that the entries of the $k \times k$-matrix valued function m are bounded functions of $x \in \mathsf{X}$. Conditions on the vector-valued function w are given below.

We begin with some basic assumptions on $\boldsymbol{\Phi}$, required to construct a linear operator with useful properties.

2.1 Some spectral theory

We assume throughout that the Markov chain $\boldsymbol{\Phi}$ is *geometrically ergodic* [25,18]. This is equivalent to assuming the validity of the following conditions:

Irreducibility & aperiodicity: There exists a σ-finite measure ψ on the state space X such that, for any $x \in \mathsf{X}$ and any measurable $A \subset \mathsf{X}$ with $\psi(A) > 0$,

$$P^t(x, A) := \mathsf{P}\{\Phi_t \in A \mid \Phi(0) = x\} > 0, \quad \text{for all sufficiently large } t > 0.$$

Minorization: There exists a non-empty set $C \in \mathcal{B}(\mathsf{X})$, a non-zero, positive measure ν on $\mathcal{B}(\mathsf{X})$, and $t_0 \geq 1$ satisfying

$$P^{t_0}(x, A) \geq \nu(A) \qquad x \in C,\ A \in \mathcal{B}(\mathsf{X}).$$

In this case, the set C and the measure ν are called *small.*

Geometric drift: There exists a *Lyapunov function* $V : \mathsf{X} \to [1, \infty)$, $\gamma < 1$, $b < \infty$, a small set C, and small measure ν, satisfying

$$PV\,(x) := \int P(x, dy) V(y) \leq \gamma V(x) + b\mathbb{I}_C(x), \qquad x \in \mathsf{X} \tag{11}$$

Under these assumptions it is known that there is a unique invariant probability measure π, and the underlying distributions of $\boldsymbol{\Phi}$ converge to π geometrically fast, in total-variation norm. Moreover, in (11) we may assume without loss of generality that $\pi(V^2) := \int V^2(x)\pi(dx) < \infty$. For a detailed development of geometrically ergodic Markov processes see [25,26,18].

Let L_∞^V denote the set of measurable *vector-valued* functions $g\colon \mathsf{X} \to \mathbb{C}^k$ satisfying

$$\|g\|_V := \sup_{x\in\mathsf{X}} \frac{\|g(x)\|}{V(x)} < \infty\,,$$

where $\|\cdot\|$ is the Euclidean norm on $\mathbb{C}^k$, and $V\colon \mathsf{X} \to [1,\infty)$ is the Lyapunov function as above. For a linear operator $\mathcal{L}\colon L_\infty^V \to L_\infty^V$ we define the induced operator norm via

$$|\!|\!|\mathcal{L}|\!|\!|_V := \sup \|\mathcal{L}f\|_V / \|f\|_V$$

where the supremum is over all non-zero $f \in L_\infty^V$. We say that $\mathcal{L}$ is a bounded linear operator if $|\!|\!|\mathcal{L}|\!|\!|_V < \infty$, and its spectral radius is then given by

$$\xi := \lim_{t\to\infty} \left(|\!|\!|\mathcal{L}^t|\!|\!|\right)^{1/t} \tag{12}$$

The *spectrum* $S(\mathcal{L})$ of the linear operator $\mathcal{L}$ is

$$S(\mathcal{L}) := \{z \in \mathbb{C} : (Iz - \mathcal{L})^{-1} \text{ does not exist as a bdd linear operator on } L_\infty^V\}.$$

If $\mathcal{L}$ is a finite matrix, its spectrum is just the collection of all its eigenvalues. Generally, for the linear operators considered in this paper, the dimension of $\mathcal{L}$ and its spectrum will be infinite.

The family of linear operators $\mathcal{L}_\alpha\colon L_\infty^V \to L_\infty^V$, $\alpha \in \mathbb{R}$, that will be used to analyze the recursion (1) are defined by

$$\begin{aligned} \mathcal{L}_\alpha f(x) &:= \mathsf{E}\left[(I - \alpha m(\Phi_1))^{\scriptscriptstyle T} f(\Phi_1) \mid \Phi_0 = x\right] \\ &= \mathsf{E}_x\left[(I - \alpha M_1)^{\scriptscriptstyle T} f(\Phi_1)\right]\,, \end{aligned} \tag{13}$$

and we let ξ_α denote the spectral radius of $\mathcal{L}_\alpha$. The motivation for (13) comes from the representation (5), and the following expression for the iterates of this semigroup:

$$\mathcal{L}_\alpha^t f\,(x) = \mathsf{E}_x\left[(I - \alpha M_1)^{\scriptscriptstyle T} \cdots (I - \alpha M_t)^{\scriptscriptstyle T} f(\Phi_t)\right]\,, \qquad t \ge 1\,. \tag{14}$$

The transpose ensures that the matrices are multiplied in order consistent with (5).

We assume throughout the paper that $m\colon \mathsf{X} \to \mathbb{R}^{k\times k}$ is a bounded function. Under these conditions we obtain the following result as in [27].

Theorem 1. *There exists* $\alpha_0 > 0$ *such that for* $\alpha \in (0, \alpha_0)$, $\xi_\alpha < \infty$, *and* $\xi_\alpha \in S(\mathcal{L}_\alpha)$. □

To ensure that the recursion (1) is stable in the mean, it is sufficient that the spectral radius satisfy $\xi_\alpha < 1$. Under this condition it is obvious that the mean $\mathsf{E}[X_t]$ is uniformly bounded in t (see (14)). The following result summarizes additional conclusions obtained below.

Theorem 2. *Suppose that the eigenvalues of* $\overline{M} := \int m(x)\,\pi(dx)$ *are all positive, and that* $w^2 \in L_\infty^V$, *where the square is interpreted component-wise. Then, there exists a bounded open set* $O \in \mathbb{R}$ *containing* $(0, \alpha_0)$, *where* α_0 *is given in Theorem 1, such that:*

(i) *For all* $\alpha \in O$ *we have* $\xi_\alpha < 1$, *and for any initial condition* $\Phi_0 = x \in \mathsf{X}$, $X_0 = \gamma \in \mathbb{R}^k$,

$$\mathsf{E}_x[\|X_t\|^2] \to \tau_\alpha^2, \quad \textit{geometrically fast, as } t \to \infty,$$

for a finite constant τ_α^2.

(ii) *If* $\boldsymbol{\Phi}$ *is stationary, then for* $\alpha \in O$ *there exists a stationary process* $\boldsymbol{X}^\alpha$ *such that for any initial condition* $\Phi_0 = x \in \mathsf{X}$, $X_0 = \gamma \in \mathbb{R}^k$,

$$\mathsf{E}_x[\|X_t(\gamma) - X_t^\alpha\|^2] \to 0, \quad \textit{geometrically fast, as } t \to \infty.$$

(iii) *If* $\alpha \notin \overline{O}$ *and the noise* $\boldsymbol{W}$ *is i.i.d. with a positive definite covariance matrix, then* $\mathsf{E}_x[\|X_t\|^2]$ *is unbounded.*

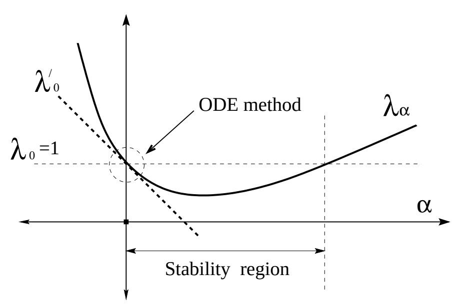

Fig. 1. The graph shows how $\lambda_\alpha := \xi_\alpha$ varies with α. When α is close to 0, Theorem 4 below implies that the ODE (8) determines stability of the algorithm since it determines whether or not $\xi_\alpha < 1$. A second-derivative formula is also given in Theorem 4: If λ_0'' is large, then the range of α for stability will be correspondingly small.

Proof Outline for Theorem 2 Starting with (5), we may express the expectation $\mathsf{E}_x\left[X_{t+1}^T X_{t+1}\right]$ as a sum of terms of the form,

$$\mathsf{E}_x\left[W_j^T\Bigl(\prod_{i=t}^{j}(I-\alpha M_i)\Bigr)^T\Bigl(\prod_{i=t}^{k}(I-\alpha M_i)\Bigr)W_k\right],\qquad j,k=0,\dots,t\,. \quad (15)$$

For simplicity consider the case $j=k$. Taking conditional expectations at time j, one can then express the expectation (15) as

$$\operatorname{trace}\Bigl(\mathsf{E}_x\Bigl[(\mathcal{Q}_\alpha^{t-j}h\,(\Phi_j))w(\Phi_j)w(\Phi_j)^T\Bigr]\Bigr)$$

where $\mathcal{Q}_\alpha$ is defined in (20), and $h\equiv I_{k\times k}$. We define O as the set of α such that the spectral radius of this linear operator is strictly less than unity. Thus, for $\alpha\in O$ we have, for some $\eta_\alpha<1$,

$$\operatorname{trace}\Bigl((\mathcal{Q}_\alpha^{t-j}h\,(y))w(y)w(y)^T\Bigr)=O(V(y)^2e^{-\eta_\alpha(t-j)}),\qquad \Phi_j=y\in\mathsf{X}.$$

Similar reasoning may be applied for arbitrary k,j, and this shows that $\mathsf{E}_x[\|X_t\|^2]$ is bounded in $t\ge 0$ for any deterministic initial conditions $\Phi_0=x\in\mathsf{X}$, $X_0=\gamma\in\mathbb{R}^k$.

To construct the stationary process X^α we apply backward coupling as presented in [32]. Consider the system starting at time $-n$, initialized at $\gamma=0$, and let $X_t^{\alpha,n}$, $t\ge -n$, denote the resulting state trajectory. We then have for all $n,m\ge 1$,

$$X_t^{\alpha,m}-X_t^{\alpha,n}=\Bigl(\prod_{i=t}^{0}(I-\alpha M_i)\Bigr)[X_0^{\alpha,m}-X_0^{\alpha,n}],\qquad t\ge 0\,,$$

which implies convergence in L_2 to a stationary process: $X_t^\alpha:=\lim_{n\to\infty}X_t^{\alpha,n}$, $t\ge 0$. We can then compare to the process initialized at $t=0$,

$$X_t^\alpha-X_t(\gamma)=\Bigl(\prod_{i=t}^{0}(I-\alpha M_i)\Bigr)[X_0^\alpha-X_0(\gamma)],\qquad t\ge 0\,,$$

and the same reasoning as before gives (ii). □

2.2 Spectral decompositions

Next we show that $\lambda_\alpha:=\xi_\alpha$ is in fact an eigenvalue of $\mathcal{L}_\alpha$ for a range of $\alpha\sim 0$, and we use this fact to obtain a multiplicative ergodic theorem. The maximal eigenvalue λ_α in Theorem 3 is a generalization of the Perron-Frobenius eigenvalue; c.f. [31,18].

Theorem 3. *Suppose that the eigenvalues $\{\lambda_i(\overline{M})\}$ of $\overline{M}$ are distinct, then*

(i) *There exists* $\varepsilon_0 > 0, \delta_0 > 0$ *such that the linear operator* $\mathcal{L}_z$ *has exactly* k *distinct eigenvalues* $\{\lambda_{1,z}, \ldots, \lambda_{k,z}\} \subset S(\mathcal{L}_z)$ *within the restricted range,*

$$B_1(\delta_0) = \{\lambda \in S(\mathcal{L}_z) : |\lambda - 1| < \delta_0\},$$

whenever z *lies in the open ball* $B_0(\varepsilon_0) := \{z \in \mathbb{C} : |z| < \varepsilon_0\}$. *The* i*th eigenvalue* $\lambda_{i,z}$ *is an analytic function of* z *on this domain for each* i.

(ii) *For* $z \in B(\varepsilon_0)$ *there are associated eigenfunctions* $\{h_{1,z}, \ldots, h_{k,z}\} \subset L_\infty^V$ *and eigenmeasures* $\{\mu_{1,z}, \ldots, \mu_{k,z}\}$ *satisfying*

$$\mathcal{L}_z h_{i,z} = \lambda_{i,z} h_{i,z}, \qquad \mu_{i,z}\mathcal{L}_z = \lambda_{i,z}\mu_{i,z}\,.$$

Moreover, for each i, $x \in \mathsf{X}$, $A \in \mathcal{B}(\mathsf{X})$, $\{h_{i,z}(x), \mu_{i,z}(A)\}$ *are analytic on* $B(\varepsilon_0)$.

(iii) *Suppose moreover that the eigenvalues* $\{\lambda_i(\overline{M})\}$ *are real. Then we may take* $\varepsilon_0 > 0$ *sufficiently small so that* $\{\lambda_{i,\alpha}, h_{i,\alpha}, \mu_{i,\alpha}\}$ *are real for* $\alpha \in (0, \varepsilon_0)$. *The maximal eigenvalue* $\lambda_\alpha := \max_i \lambda_{i,\alpha}$ *is equal to* ξ_α, *and the corresponding eigenfunction and eigenmeasure may be scaled so that the following limit holds,*

$$\lambda_\alpha^{-t}\mathcal{L}_\alpha^t \to h_\alpha \otimes \mu_\alpha, \qquad t \to \infty,$$

where the convergence is in the V*-norm.*
In fact, there exists $\delta_0 > 0$ *and* $b_0 < \infty$ *such that for any* $f \in L_\infty^V$ *the following limit holds:*

$$\left|\lambda_\alpha^{-t}\mathsf{E}_x\Big[\Big(\prod_{i=1}^t (I - \alpha M_i)\Big)^{T} f(\Phi_t)\Big] - h_\alpha(x)\mu_\alpha(f)\right| \le b_0 e^{-\delta_0 t} V(x)\,.$$

Proof. The linear operator $\mathcal{L}_0$ possesses a k-dimensional eigenspace corresponding to the eigenvalue $\lambda_0 = 1$. This eigenspace is precisely the set of constant functions, with a corresponding basis of eigenfunctions given by $\{e^i\}$, where e^i is the ith basis element in $\mathbb{R}^k$. The k-dimensional set of vector-valued eigenmeasures $\{\pi^i\}$ given by $\pi^i = {e^i}^T\pi$ spans the set of all eigenmeasures with eigenvalue $\lambda_{0,i} = 1$.

Consider the rank-k linear operator $\Pi\colon L_\infty^V \to L_\infty^V$ defined by $\Pi f := \pi(f)$. This is equivalently expressed as

$$\Pi f\,(x) = (\pi(f_1), \ldots, \pi(f_k))^{T} = \big[\sum e^i \otimes \pi^i\big] f, \qquad f \in L_\infty^V.$$

It is obvious that $\Pi\colon L_\infty^V \to L_\infty^V$ is a rank-k linear operator, and for $\alpha = 0$ we have from the V-uniform ergodic theorem of [25],

$$\mathcal{L}_0^t - \Pi = [\mathcal{L}_0 - \Pi]^t \to 0, \qquad t \to \infty,$$

where the convergence is in norm, and hence takes place exponentially fast. It follows that the spectral radius of $(\mathcal{L}_0 - \Pi)$ is strictly less than unity. By standard arguments it follows that, for some $\varepsilon_0 > 0$, the spectral radius of $\mathcal{L}_z - \Pi$ is also strictly less than unity. The results then follow as in Theorem 3 of [19]. ☐

Conditions under which the bound $\xi_\alpha < 1$ is satisfied are given in Theorem 4, where we also provide formulae for the derivatives of λ_α:

Theorem 4. *Suppose that the eigenvalues $\{\lambda_i(\overline{M})\}$ are real and distinct, then the maximal eigenvalue $\lambda_\alpha = \xi_\alpha$ satisfies:*

(i) $\frac{d}{d\alpha}\lambda_\alpha\Big|_{\alpha=0} = -\lambda_{\min}(\overline{M})$.

(ii) *The second derivative is given by,*

$$\frac{d^2}{d\alpha^2}\lambda_\alpha\Big|_{\alpha=0} = 2\sum_{l=0}^{\infty} v_0^T \mathsf{E}_\pi[(M_0 - \overline{M})(M_{l+1} - \overline{M})]r_0\,,$$

where r_0 is a right eigenvector of $\overline{M}$ corresponding to $\lambda_{\min}(\overline{M})$, and v_0 is the left eigenvector, normalized so that $v_0^T r_0 = 1$.

(iii) *Suppose that $m(x) = m^T(x)$, $x \in \mathsf{X}$, then we may take $v_0 = r_0$ in (ii), and the second derivative may be expressed,*

$$\frac{d^2}{d\alpha^2}\lambda_\alpha\Big|_{\alpha=0} = \mathrm{trace}\,(\Gamma - \Sigma),$$

where an Γ is the Central Limit Theorem covariance for the stationary vector-valued stochastic process $F_k = [M_k - \overline{M}]v_0$, and $\Sigma = \mathsf{E}_\pi[F_k F_k^T]$ is its variance [25].

Proof. To prove (i), we differentiate the eigenfunction equation $\mathcal{L}_\alpha h_\alpha = \lambda_\alpha h_\alpha$ to obtain

$$\mathcal{L}_\alpha{}' h_\alpha + \mathcal{L}_\alpha h_\alpha' = \lambda_\alpha' h_\alpha + \lambda_\alpha h_\alpha'. \tag{16}$$

Setting $\alpha = 0$ then gives a version of *Poisson's equation*,

$$\mathcal{L}_0' h_0 + P h_0' = \lambda_0' h_0 + h_0', \tag{17}$$

where $\mathcal{L}_0' h_0 = \mathsf{E}_x\left[-m(\Phi_1)^T h_0(\Phi_1)\right]$. An application of Theorem 3 (ii) shows that $h_0' \in L_\infty^V$, which justifies an integration of both sides of (17) with respect to the invariant probability π to obtain

$$\mathsf{E}_\pi\left[-m(\Phi_1)^T\right] h_0 = -\overline{M}^T h_0 = \lambda_0' h_0.$$

This shows that λ_0' is an eigenvalue of $-\overline{M}$, and h_0 is an associated eigenvector for $\overline{M}^T$. It follows that $\lambda_0' = -\lambda_{\min}(\overline{M})$ by maximality of λ_α.

We note that Poisson's equation (17) combined with equation (17.39) of [25] implies the formula,

$$h_0'(x) = \mathsf{E}_\pi[h_0'(\Phi(0))] - \sum_{l=0}^{\infty} \mathsf{E}_x[(M_{l+1} - \overline{M})^T] h_0\,. \tag{18}$$

To prove (ii) we consider the second-derivative formula,

$$\mathcal{L}''_\alpha h_\alpha + 2\mathcal{L}'_\alpha h'_\alpha + \mathcal{L}_\alpha h''_\alpha = \lambda_\alpha h''_\alpha + 2\lambda'_\alpha h'_\alpha + \lambda''_\alpha h_\alpha.$$

Evaluating these expressions at $\alpha = 0$ and integrating with respect to π then gives the steady state expression,

$$\lambda''_0 h_0 = -2\mathsf{E}_\pi[(M_1 + \lambda'_0)h'_0(\Phi_1)]. \tag{19}$$

In deriving this identity we have used the expressions,

$$\mathcal{L}'_0 f\,(x) = \mathsf{E}_x[M_1 f(\Phi_1)], \quad \mathcal{L}''_0 f\,(x) = 0, \qquad f \in L^V_\infty,\ x \in \mathsf{X}.$$

This combined with (19) gives the desired formula since we may take $v_0 = h_0$ in (ii).

To prove (iii) we simply note that in the symmetric case the formula in (ii) becomes,

$$\lambda''_0 = \sum_{k\neq 0} \mathsf{E}_\pi[\|F_k\|^2] = \operatorname{trace}(\Gamma - \Sigma)\,. \qquad \square$$

2.3 Second-order statistics

In order to understand the second-order statistics of $\boldsymbol{X}$ it is convenient to introduce another linear operator $\mathcal{Q}_\alpha$ as follows,

$$\begin{aligned}\mathcal{Q}_\alpha f(x) &= \mathsf{E}\left[(I - \alpha m(\Phi_1))^{\mathsf{T}} f(\Phi_1)(I - \alpha m(\Phi_1)) | \Phi_0 = x\right] \\ &= \mathsf{E}_x\left[(I - \alpha M_1)^{\mathsf{T}} f(\Phi_1)(I - \alpha M_1)\right],\end{aligned} \tag{20}$$

where the domain of $\mathcal{Q}_\alpha$ is the collection of matrix-valued functions $f\colon \mathsf{X} \to \mathbb{C}^{k\times k}$. When considering $\mathcal{Q}_\alpha$ we redefine L^V_∞ accordingly. It is clear that $\mathcal{Q}_\alpha : L^V_\infty \to L^V_\infty$ is a bounded linear operator under the geometric drift condition and the boundedness assumption on m.

Let ξ^Q_z denote the spectral radius of $\mathcal{Q}_\alpha$. We can again argue that ξ^Q_z is smooth in a neighborhood of the origin, and the following follows as in Theorem 4:

Theorem 5. *Assume that the eigenvalues of $\overline{M}$ are real and distinct. Then there exists $\varepsilon_0 > 0$ such that for each $z \in B(\varepsilon_0)$ there exists an eigenvalue $\eta_z \in \mathbb{C}$ for $\mathcal{Q}_z$ satisfying $|\eta_z| = \xi^Q_z$, and η_α is real for real $\alpha \in (0, \varepsilon_0)$. The eigenvalue η_z is smooth on $B(\varepsilon_0)$ and satisfies,*

$$\eta'_0(\mathcal{Q}) = -2\lambda_{\min}(\overline{M}).$$

PROOF. This is again based on differentiation of the eigenfunction equation given by $\mathcal{Q}_\alpha h_\alpha = \eta_\alpha h_\alpha$, where η_α and h_α are the eigenvalue and matrix-valued eigenfunction, respectively. Taking derivatives on both sides gives

$$\mathcal{Q}_\alpha{}' h_\alpha + \mathcal{Q}_\alpha h'_\alpha = \eta'_\alpha h_\alpha + \eta_\alpha h'_\alpha \tag{21}$$

where $\mathcal{Q}'_0 h_0 = \mathsf{E}_x\left[-m(\Phi_1)^T h_0(\Phi_1) - h_0(\Phi_1) m(\Phi_1)\right]$. As before, we then obtain the steady-state expression,

$$\mathsf{E}_\pi\left[-m(\Phi_1)^T h_0 - h_0 m(\Phi_1)\right] = -\overline{M}^T h_0 - h_0 \overline{M} = \eta'_0 h_0. \tag{22}$$

And, as before, we may conclude that $\eta'_0 = 2\lambda'_0 = -2\lambda_{\min}(\overline{M})$. □

2.4 An illustrative example

Consider the discrete-time, linear time-varying model

$$y_t = \theta_t^T \phi_t + n_t, \qquad t \geq 0\,, \tag{23}$$

where $\boldsymbol{y} = \{y_t\}$ is a sequence of scalar observations, $\boldsymbol{n} = \{n_t\}$ is a noise process, $\boldsymbol{\theta} = \{\theta_t\}$ is the sequence of k-dimensional regression vectors, and $\boldsymbol{\phi}=\{\phi_t\}$ are k-dimensional time-varying parameters. In this section we illustrate the results above using the LMS (least mean square) parameter estimation algorithm,

$$\widehat{\theta}_{t+1} = \widehat{\theta}_t + \alpha \phi_t e_t,$$

where $\boldsymbol{e} = \{e_t\}$ is the error sequence, $e_t := y_t - \widehat{\theta}_t^T \phi_t$, $t \geq 0$.

As in the Introduction, writing $\widetilde{\theta}_t = \theta_t - \widehat{\theta}_t$ we obtain

$$\tilde{\theta}_{t+1} = (I - \alpha \phi_t \phi_t^T)\tilde{\theta}_t + [\theta_{t+1} - \theta_t - \alpha \phi_t n_t]\;.$$

This is of the form (1) with $X_t = \tilde{\theta}_t$, $M_t = \phi_t \phi_t^T$ and $W_{t+1} = \theta_{t+1} - \theta_t - \alpha \phi_t n_t$.

For the sake of simplicity and to facilitate explicit numerical calculations, we consider the following special case: We assume that $\boldsymbol{\phi}$ is of the form $\phi_t = (s_t, s_{t-1})^T$, where the sequence $\boldsymbol{s}$ is Bernoulli ($s_t = \pm 1$ with equal probability) and take $\boldsymbol{n}$ to be an i.i.d. noise sequence.

In analyzing the random linear system we may ignore the noise $\boldsymbol{n}$ and take $\boldsymbol{\Phi} = \boldsymbol{\phi}$. This is clearly geometrically ergodic since it is an ergodic, finite state space Markov chain, with four possible states. In fact, $\boldsymbol{\Phi}$ is geometrically ergodic with Lyapunov function $V \equiv 1$. In the case $k = 2$, viewing $h \in L_\infty^V$ as a real vector, the eigenfunction equation for $\mathcal{L}_\alpha$ becomes

$$\mathcal{L}_\alpha h_\alpha = \frac{1}{2}\begin{bmatrix} A_1 & A_0 & A_2 & A_0 \\ A_1 & A_0 & A_2 & A_0 \\ A_0 & A_2 & A_0 & A_1 \\ A_0 & A_2 & A_0 & A_1 \end{bmatrix} h_\alpha = \lambda_\alpha h_\alpha \tag{24}$$

where $A_0 = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$, $A_1 = \begin{bmatrix} 1-\alpha & -\alpha \\ -\alpha & 1-\alpha \end{bmatrix}$, $A_2 = \begin{bmatrix} 1-\alpha & \alpha \\ \alpha & 1-\alpha \end{bmatrix}$.

In this case, we have the following local behavior:

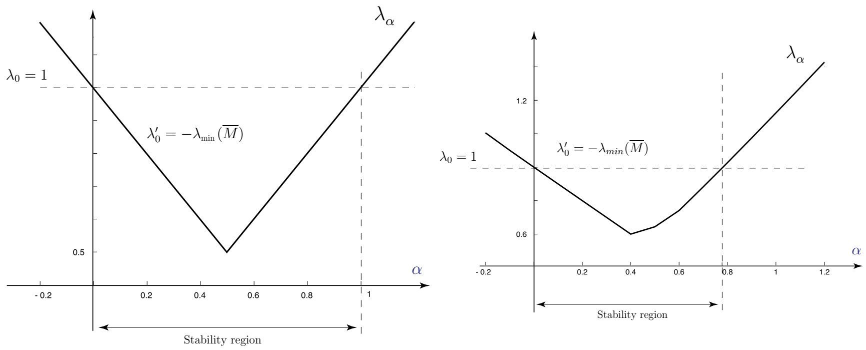

Fig. 2. The figure on the left shows the Perron-Frobenius eigenvalue $\lambda_\alpha = \xi_\alpha$ for the LMS model with $\phi_t = (s_t, s_{t-1})^T$. The figure on the right shows the case where $\phi_t = (s_t, s_{t-1}, s_{t-2})^T$. In both cases, the sequence $\boldsymbol{s}$ is i.i.d. Bernoulli.

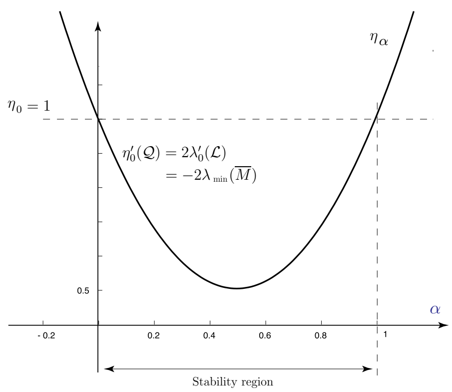

Fig. 3. The maximal eigenvalues $\eta_\alpha = \xi_\alpha^Q$ are piecewise quadratic in α in the case where $\phi_t = (s_t, s_{t-1})^T$ with $\boldsymbol{s}$ as above.

Theorem 6. *In a neighborhood of* 0*, the spectral radii of* $\mathcal{L}_\alpha$*,* $\mathcal{Q}_\alpha$ *satisfy*

$$\frac{d}{d\alpha}\xi_\alpha\Big|_{\alpha=0} = -\lambda_{\min}(\overline{M}) = -1; \qquad \frac{d}{d\alpha}\xi_\alpha^Q\Big|_{\alpha=0} = -2\lambda_{\min}(\overline{M}) = -2$$

$$\frac{d^n}{d\alpha^n}\xi_\alpha\Big|_{\alpha=0} = 0, n \geq 2; \qquad \frac{d^n}{d\alpha^n}\xi_\alpha^Q\Big|_{\alpha=0} = 0, n \geq 3.$$

So λ_α *and* η_α *are linear and quadratic around* 0*, respectively.*

Proof. This follows from differentiating the respective eigenfunction equations. Here we only show the proof for operator $\mathcal{Q}$; the proof for operator $\mathcal{L}$ is similar.

Taking derivatives on both sides of the eigenfunction equation for $\mathcal{Q}_\alpha$ gives,

$$\mathcal{Q}_\alpha{}' h_\alpha + \mathcal{Q}_\alpha h'_\alpha = \eta'_\alpha h_\alpha + \eta_\alpha h'_\alpha \tag{25}$$

Setting $\alpha = 0$ gives a version of *Poisson's equation*,

$$\mathcal{Q}'_0 h_0 + \mathcal{Q} h'_0 = \eta'_0 h_0 + \eta_0 h'_0 \tag{26}$$

Using the identities of h_0 and $\mathcal{Q}'_0 h_0 = \mathsf{E}_x[-M_1^T h_0 - h_0 M_1]$, we obtain the steady state expression

$$\overline{M}^T h_0 + h_0 \overline{M} = -\eta'_0 h_0. \tag{27}$$

Since $\overline{M} = I$, we have $\eta'_0 = -2$. Now, taking the 2nd derivatives on both sides of (25) gives,

$$\mathcal{Q}''_\alpha h_\alpha + 2\mathcal{Q}'_\alpha h'_\alpha + \mathcal{Q}_\alpha h''_\alpha = \eta''_\alpha h_\alpha + 2\eta'_\alpha h'_\alpha + \eta_\alpha h''_\alpha. \tag{28}$$

Letting $\alpha = 0$ and considering the steady state, we obtain

$$2\overline{M}^T h_0 \overline{M} - 2\mathsf{E}_\pi[M_1^T h'_0 + h'_0 M_1] = \eta''_0 h_0 + 2\eta'_0 \mathsf{E}_\pi[h'_0]. \tag{29}$$

Poisson's equation (26) combined with equation (27) and equation (17.39) of [25] implies the formula,

$$\begin{aligned} h'_0(x) &= \mathsf{E}_\pi(h'_0) + \textstyle\sum_{l=0}^\infty \mathsf{E}_x[-M_{l+1}^T h_0 - h_0 M_{l+1} - \eta'_0 h_0] \\ &= \mathsf{E}_\pi(h'_0) + \textstyle\sum_{l=0}^\infty \mathsf{E}_x[(\overline{M} - M_{l+1})^T h_0 + h_0(\overline{M} - M_{l+1})]. \end{aligned} \tag{30}$$

So, from $\overline{M} = I$, $\eta'_0 = -2$ and (29) we have $\eta''_0 = 2$. In order to show η_α is quadratic near zero, we take the 3rd derivative on both sides of (28) and consider the steady state at $\alpha = 0$,

$$\mathcal{Q}'''_0 h_0 + 3\mathcal{Q}''_0 h'_0 + 3\mathcal{Q}'_0 h''_0 + \mathcal{Q}_0 h'''_0 = \eta'''_0 h_0 + 3\eta''_0 h'_0 + 3\eta'_0 h''_0 + \eta_0 h'''_0. \tag{31}$$

With equation (17.39) of [25] and $\eta'_0 = -2$ and $\eta''_0 = 2$, we can show $\eta'''_0 = 0$ and $\eta_0^{(n)} = 0$ for $n > 3$, hence η_α is quadratic around 0. □

3 Nonlinear models

We now turn to the nonlinear model shown in (6). We take the special form,

$$X_{t+1} = X_t - \alpha[f(X_t, \Phi_{t+1}) + W_{t+1}]\,, \tag{32}$$

We continue to assume that $\boldsymbol{\Phi}$ is geometrically ergodic, and that $W_t = w(\Phi_t)$, $t \geq 0$, with $w^2 \in L_\infty^V$. The associated ODE is given by

$$\frac{d}{dt}\gamma_t = \overline{f}(\gamma_t), \tag{33}$$

where $\overline{f}(\gamma) = \int f(\gamma, x)\, \pi(dx)$, $\gamma \in \mathbb{R}^k$.

We assume that $\overline{W} = \mathsf{E}_\pi[W_1] = 0$, and the following conditions are imposed on f. The function $\overline{f}_\infty$ appearing in Condition (N1) may be used to construct an ODE that approximates the behavior of (32) when the initial condition is very large.

(N1) The function f is Lipschitz, and there exists a function $\overline{f}_\infty : \mathbb{R}^k \to \mathbb{R}^k$ such that

$$\lim_{r\to\infty} r^{-1}\overline{f}(r\gamma) = \overline{f}_\infty(\gamma), \qquad \gamma \in \mathbb{R}^k.$$

Furthermore, the origin in $\mathbb{R}^k$ is an asymptotically stable equilibrium point for the ODE,

$$\frac{d}{dt}\gamma_t^\infty = \overline{f}_\infty(\gamma_t^\infty). \tag{34}$$

(N2) There exists $b_f < \infty$ such that $\sup_{\gamma\in\mathbb{R}^k} \|f(\gamma, x) - \overline{f}(\gamma)\|^2 \le b_f V(x),\ x \in \mathsf{X}$.

(N3) There exists a unique stationary point x^* for the ODE (33) that is a globally asymptotically stable equilibrium.

Define the absolute error by

$$\varepsilon_t := \|X_t - x^*\|, \qquad t \ge 0. \tag{35}$$

The following result is an extension of Theorem 1 of [8] to Markov models:

Theorem 7. *Assume that (N1)–(N3) hold. Then there exists $\varepsilon_0 > 0$ such that for any $0 < \alpha < \varepsilon_0$:*

(i) *For any $\delta > 0$, there exists $b_1 = b_1(\delta) < \infty$ such that*

$$\limsup_{n\to\infty} \mathsf{P}(\varepsilon_n \ge \delta) \le b_1\alpha.$$

(ii) *If the origin is a globally exponentially asymptotically stable equilibrium for the ODE (33), then there exists $b_2 < \infty$ such that for every initial condition $\Phi_0 = x \in \mathsf{X}$, $X_0 = \gamma \in \mathbb{R}^k$,*

$$\limsup_{n\to\infty} \mathsf{E}[\varepsilon_n^2] \le b_2\alpha.$$

Proof Outline for Theorem 7 The continuous-time process $\{x_t^o : t \ge 0\}$ is defined to be the interpolated version of $\boldsymbol{X}$ given as follows: Let $T_j = j\alpha$, $j \ge 0$, and define $x^o(T_j) = \alpha X_j$, with $\boldsymbol{x}^o$ defined by linear interpolation on the remainder of $[T_j, T_{j+1}]$ to form a piecewise linear function. Using geometric ergodicity we can bound the error between $\boldsymbol{x}^o$ and solutions to the ODE (33) as in [8], and we may conclude that the joint process $(\boldsymbol{X}, \boldsymbol{\Phi})$ is geometrically ergodic with Lyapunov function $V_2(\gamma, x) = \|\gamma\|^2 + V(x)$. □

We conclude with an extension of Theorem 2 describing the behavior of the sensitivity process $\boldsymbol{S}$.

Theorem 8. *Assume that (N1)–(N3) hold, and that the eigenvalues of the matrix* $\overline{M}$ *have strictly positive real part, where*

$$\overline{M} := \nabla \overline{f}\,(x^*)\,.$$

Then there exists $\varepsilon_1 > 0$ *such that for any* $0 < \alpha < \varepsilon_1$*, the conclusions of Theorem 7 (ii) hold, and, in addition:*

(i) *The spectral radius* ξ_α *of the random linear system (7) describing the evolution of the sensitivity process is strictly less than one.*

(ii) *There exists a stationary process* $\boldsymbol{X}^\alpha$ *such that for any initial condition* $\Phi_0 = x \in \mathsf{X}$, $X_0 = \gamma \in \mathbb{R}^k$,

$$\mathsf{E}_x[\|X_t - X_t^\alpha\|^2] \to 0, \qquad t \to \infty\,.$$

References

1. L. Arnold. *Random dynamical systems*. Springer-Verlag, Berlin, 1998.
2. S. Balaji and S.P. Meyn. Multiplicative ergodicity and large deviations for an irreducible Markov chain. *Stochastic Process. Appl.*, 90(1):123–144, 2000.
3. R. Bellman. Limit theorems for non-commutative operations. I. *Duke Math. J.*, 21, 1954.
4. Michel Benaïm. Dynamics of stochastic approximation algorithms. In *Séminaire de Probabilités, XXXIII*, pages 1–68. Springer, Berlin, 1999.
5. Albert Benveniste, Michel Métivier, and Pierre Priouret. *Adaptive algorithms and stochastic approximations*. Springer-Verlag, Berlin, 1990. Translated from the French by Stephen S. Wilson.
6. D.P. Bertsekas and J. Tsitsiklis. *Neuro-Dynamic Programming*. Atena Scientific, Cambridge, Mass, 1996.
7. B. Bharath and V. S. Borkar. Stochastic approximation algorithms: overview and recent trends. *Sādhanā*, 24(4-5):425–452, 1999. Chance as necessity.
8. V.S. Borkar and S.P. Meyn. The O.D.E. Method for Convergence of Stochastic Approximation and Reinforcement Learning. *SIAM J. Control Optim.*, 38:447–69, 2000.
9. P. Bougerol. Limit theorem for products of random matrices with Markovian dependence. In *Proceedings of the 1st World Congress of the Bernoulli Society, Vol. 1 (Tashkent, 1986)*, pages 767–770, Utrecht, 1987. VNU Sci. Press.
10. A. Crisanti, G. Paladin, and A. Vulpiani. *Products of random matrices in statistical physics*. Springer-Verlag, Berlin, 1993.
11. O. Dabeer and E. Masry. The LMS adaptive algorithm: Asymptotic error analysis. In *Proceedings of the 34th Annual Conference on Information Sciences and Systems, CISS 2000*, pages WP1–6 – WP1–7, Princeton, NJ, March 2000.
12. Paul Fischer and Hans Ulrich Simon, editors. *Computational learning theory*, Berlin, 1999. Springer-Verlag. Lecture Notes in Artificial Intelligence.
13. H. Furstenberg and H. Kesten. Products of random matrices. *Ann. Math. Statist*, 31:457–469, 1960.
14. László Gerencsér. Almost sure exponential stability of random linear differential equations. *Stochastics Stochastics Rep.*, 36(2):91–107, 1991.

15. R. Gharavi and V. Anantharam. Structure theorems for partially asynchronous iterations of a nonnegative matrix with random delays. *Sādhanā*, 24(4-5):369–423, 1999. Chance as necessity.
16. S.V. Hanly and D. Tse. Multiaccess fading channels. II. Delay-limited capacities. *IEEE Trans. Inform. Theory*, 44(7):2816–2831, 1998.
17. J. A. Joslin and A. J. Heunis. Law of the iterated logarithm for a constant-gain linear stochastic gradient algorithm. *SIAM J. Control Optim.*, 39(2):533–570 (electronic), 2000.
18. I. Kontoyiannis and S.P. Meyn. Spectral theory and limit theorems for geometrically ergodic Markov processes. *Submitted*, 2001. Also presented at the 2001 INFORMS Applied Probability Conference, NY, July, 2001.
19. I. Kontoyiannis and S.P. Meyn. Spectral theory and limit theorems for geometrically ergodic Markov processes. Part II: Empirical measures & unbounded functionals. *Preprint*, 2001.
20. H. J. Kushner. *Approximation and weak convergence methods for random processes, with applications to stochastic systems theory.* MIT Press, Cambridge, MA, 1984.
21. H.J. Kushner and G. Yin. *Stochastic approximation algorithms and applications.* Springer-Verlag, New York, 1997.
22. L. Ljung. On positive real transfer functions and the convergence of some recursive schemes. *IEEE Trans. Automatic Control*, AC-22(4):539–551, 1977.
23. Jean Mairesse. Products of irreducible random matrices in the $(\max, +)$ algebra. *Adv. in Appl. Probab.*, 29(2):444–477, 1997.
24. M. Medard, S.P. Meyn, and J. Huang. Capacity benefits from channel sounding in Rayleigh fading channels. *INFOCOM (submitted)*, 2001.
25. S. P. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability.* Springer-Verlag, London, 1993.
26. S.P. Meyn and R.L. Tweedie. Computable bounds for geometric convergence rates of Markov chains. *Ann. Appl. Probab.*, 4(4):981–1011, 1994.
27. G.V. Moustakides. Exponential convergence of products of random matrices, application to the study of adaptive algorithms. *International Journal of Adaptive Control and Signal Processing*, 2(12):579–597, 1998.
28. V. I. Oseledec. Markov chains, skew products and ergodic theorems for "general" dynamic systems. *Teor. Verojatnost. i Primenen.*, 10:551–557, 1965.
29. V. I. Oseledec. A multiplicative ergodic theorem. Characteristic Ljapunov, exponents of dynamical systems. *Trudy Moskov. Mat. Obšč.*, 19:179–210, 1968.
30. J. B. T. M Roerdink. Products of random matrices or "why do biennials live longer than two years? *CWI Quarterly*, 2:37–44, 1989.
31. E. Seneta. *Non-negative Matrices and Markov Chains.* Springer-Verlag, New York, Second edition, 1980.
32. H. Thorisson. *Coupling, stationarity, and regeneration.* Springer-Verlag, New York, 2000.
33. D. Tse and S.V. Hanly. Multiaccess fading channels. I. Polymatroid structure, optimal resource allocation and throughput capacities. *IEEE Trans. Inform. Theory*, 44(7):2796–2815, 1998.
34. Divakar Viswanath. Random Fibonacci sequences and the number $1.13198824\ldots$. *Math. Comp.*, 69(231):1131–1155, 2000.