ALGORITHMIC ANALYSIS OF STOCHASTIC MODELS
The Changing Face of Mathematics

Ramanujan Endowment Lecture
at Anna University, Chennai, India - December, 2000

V. Ramaswami
AT&T Labs, Research
180 Park Avenue, E-233
Florham Park, NJ 07932, USA
email: vramaswami@att.com

Abstract

The progress in computing and communications technologies made in the last quarter of the past Century has not only ushered in the "Information Age," but it has also influenced the basic sciences, including mathematics, in fundamental ways. Thanks to the significantly increased computing power, mathematicians can now augment classical techniques of analysis, proof and solution with an algorithmic approach in a manner that enables the consideration of more complex models with wider applicability, and also obtain results with greater practical value to engineering. While providing powerful tools to the mathematician, the technologies, nevertheless, are also posing new challenges and problems, and opening many new vistas for further mathematical research.

Among the areas exemplifying all these, a notable one is algorithmic methods for stochastic models based on "the matrix (operator) analytic method." This lecture is an overview of those methods.

1 INTRODUCTION

We have witnessed in the last 25 years major advances in the fields of computing and communications. These have increased our ability to effect, with remarkable speed and without much manual labor, a large number of computations with high levels of precision. This is having a major impact on mathematical sciences. Most notable among these are: (a) the increasing ability to consider complex models that do not lend themselves to "explicit, closed form" solutions, but are needed to reflect systems faithfully; (b) the greater level of acceptance of an implementable algorithm as a solution, and, sometimes, even a greater preference for it over a formula solution achieved at the expense of generality and ease of use; and (c) the ability to provide solutions in a form more appropriate for practical use or for implementation in devices. Yet, technology has not only added greater power to mathematics, but it is also posing newer challenges by way of new problems for solution, thereby serving as an engine driving more innovation in the mathematical disciplines as well. Good illustrations of all these assertions may be found in the story of algorithmic methods for stochastic models that also gained significant momentum beginning in the mid Seventies.

It is important that we delineate the scope of our discussion clearly, particularly so since the algorithmic approach, in its generality, is a vast topic that is not new to mathematics or to probability and statistics. Our subject matter is the set of tools that go by the name of "matrix analytic methods" (now, a possible misnomer due to generalizations in the operator contexts) that were initiated by M.F. Neuts [20], [21], serve as a powerful framework to analyze large classes of stochastic processes in a unified manner, and find numerous applications in many areas such as performance analysis of computer and communication systems. This area of research came about due to practical needs in certain queueing applications for more complex models than were typically covered in the literature and, more importantly, for numerical solutions to guide a variety of engineering decisions. It now has a vast literature of its own that continues to grow in many interesting directions. Thus, what is to follow is but an overview of the central concepts, with a sprinkling of historical notes to underscore the general themes discussed in the opening paragraph. For a detailed discussion and a large number of references we refer the reader to [20] [21], [15], [31], [34].

2 CLASSICAL MODELS

The context for matrix analytic methods is Markov chains and Markov renewal processes, the work horses of applied probability. In modeling a stochastic process such as a queue, inventory or storage, water level in a dam, population, etc., these are used directly, or appear as embedded processes at certain selected time points (such as arrival or departure epochs in a queue, epochs of replenishment of inventory, etc.). The context offers many interesting topics for analysis: various first passage times (e.g., time to emptiness of the dam or extinction of the population), derived quantities such as waiting time distribution of arrivals to a queue, etc. Matrix analytic methods provide a unified framework for solving a host of such problems through iterative and stable algorithms.

To keep our discussion manageable, we begin with the simple case of a discrete time, discrete state space Markov chain $\{X_n : n \geq 0\}$ on the state space $\{0, 1, \ \ldots \ \}$ with transition matrix P and concentrate on the issues concerning its steady state distribution $\boldsymbol{\pi} = (\pi_0, \pi_1, \ldots)$, which we shall assume to exist. (We hasten to emphasize, however, that the theory addresses many topics besides steady state distributions, and its scope is much more general than Markov chains.) It is well-known that under the assumptions of aperiodicity and irreducibility of the Markov chain, what we are seeking is the solution to the set of linear equations

$$\boldsymbol{\pi} = \boldsymbol{\pi} P, \quad \boldsymbol{\pi}\mathbf{1} = 1, \tag{1}$$

where $\mathbf{1}$ is a column vector of 1's and $\boldsymbol{\pi}$ is the row vector of stationary probabilities). Even in the case when P is finite-dimensional, the computation of $\boldsymbol{\pi}$ can be non-trivial, and in the infinite dimensional case, certainly, not much can be done by way of computations without additional assumptions. Thus, the natural thing to do is to impose some additional conditions on P, or equivalently on $\{X_n\}$.

Historically, there have been three types of matrices P for which the problems are so simple that they are considered even in a first course in stochastic processes. These three are displayed in the equations below:

$$P = \begin{bmatrix} b_0 & a_0 & 0 & 0 & \dots \\ a_2 & a_1 & a_0 & 0 & \dots \\ 0 & a_2 & a_1 & a_0 & \dots \\ 0 & 0 & a_2 & a_1 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \tag{2}$$

$$P = \begin{bmatrix} b_0 & b_1 & b_2 & b_3 & \dots \\ a_0 & a_1 & a_2 & a_3 & \dots \\ 0 & a_0 & a_1 & a_2 & \dots \\ 0 & 0 & a_0 & a_1 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \tag{3}$$

$$P = \begin{bmatrix} b_0 & a_0 & 0 & 0 & \dots \\ b_1 & a_1 & a_0 & 0 & \dots \\ b_2 & a_2 & a_1 & a_0 & \dots \\ b_3 & a_3 & a_2 & a_1 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \tag{4}$$

Readers will recognize the above three transition matrices as those arising in connection with the birth-and-death process, the $M/G/1$ queue, and the $GI/M/1$ queue respectively.

Although it is not noted thus explicitly in the literature, much of the simplicity of models of the above type comes from their structure embodying certain "skip-free" properties and "spatial homogeneity." Indeed, it is these that allow one to reduce the problem of characterizing $\boldsymbol{\pi}$ to the consideration of the familiar equation $z = \bar{a}(z)$, where $\bar{a}(z) = \sum z^n a_n$, and to obtain a recursion for π_i in terms of the (minimal) root $z \in (0, 1]$ of that equation—see, e.g., Karlin & Taylor [12], or Takács [37]. Let us be specific about the properties we refer to here:

Note first that the chain corresponding to the matrix in equation (3) is "skip-free to the left"—i.e., from any state n, it cannot enter any of the states $n-2$, $n-3$, $\ldots$ in one step. This is reflected in the upper Hessenberg structure of the matrix in the right of equation (3). Furthermore, it has the spatial homogeneity property: for $n \geq 1$, the transition probability $P(n, n-1+j)$ depends only on j, a fact reflected in the repeating-with-a-shift structure of the rows of P.

Turning to the chain of equation (4), note that the process is "skip-free to the right"—i.e., from any n, it cannot enter $n+2, n+3, \ldots$ in one step—and that it has the spatial homogeneity property: for $n \geq 1$, the transition probability $P(n-1+j, n)$ depends only on j. The first fact is reflected in the lower Hessenberg structure of P, and the second in the repeating-with-a-shift structure of the columns of P.

Of course, the chain of equation (2) falls in the intersection of both the above classes; it is skip-free in both directions and also has the spatial homogeneity property.

3 MATRIX METHODS - EARLY WORK

In this section, we shall discuss the generalizations of (2)-(4) considered by Neuts in the early work on matrix-analytic methods and the attending results that parallel those of the "scalar" case. It is convenient to discuss each model class separately to show the parallels; that also agrees with the way the subject area developed historically.

3.1 M/G/1 Type Chains

Considered first by Neuts [16] were "Markov chains of the M/G/1 type." These are Markov chains whose transition matrices have the structure shown in equation (3) but, unlike in the simple case discussed in the literature, the quantities a_i and b_i are now square matrices of a fixed size $m < \infty$. It turns out—see Neuts [21]—that the examples falling under this two-dimensional class are indeed numerous and cover many models of particular interest to applications. (A simple example of such a model is the chain embedded at departure epochs in a $E_m/G/1$ queue, a queue with general service times for which the interarrival times are distributed as a sum of m *iid* exponentially distributed random variables; we get block matrices because in addition to considering the queue length at departure, we also need to consider the "phase" of the arrival process to get a Markov chain; the process here is, of course, two-dimensional.) Let us examine how the generalization proceeded.

In the scalar case, it is well-known that the starting point of the analysis is the consideration of the quantity g which is the probability that starting in 1, the chain eventually visits 0. For the steady state distribution to exist, it is necessary that g be equal to one. A conditioning on the first step of the chain along with an exploitation of spatial homogeneity would give us the equation

$$g = \sum_{0}^{\infty} a_n g^n, \tag{5}$$

and hence the interest in the equation $z = \bar{a}(z)$ mentioned earlier. Once, it is determined that the model is indeed ergodic, one can compute π_i recursively—see [21].

Certainly, Neuts was not the first to consider matrix generalization of the M/G/1 queue, and one can find discussions of special instances of the matrix case, such as the $E_m/G/1$ queue, in the literature. What had been attempted in the literature by way of generalization was, however, something quite different. As a generalization of the equation $z = \bar{a}(z)$ of the scalar case, the classical approach considered the determinental equation $|zI - \bar{a}(z)| = 0$, where $\bar{a}(z)$ is the matrix generating function of the matrices a_i, and attempted to characterize quantities of interest in terms of the roots z_i of that determinental equation and the corresponding (generalized) eigenvector(s) of the matrices $\bar{a}(z_i)$. We shall not elaborate on that approach except to note that it immediately landed one in difficult problems concerning the nature of those roots and eigenvectors, computation in complex arithmetic, etc., and that the approach of Neuts helped to avoid those difficulties; see Neuts [17] and Ramaswami [24] for some details.

In the matrix case, we have a Markov chain $\{(X_n, J_n)\}$ with state space $\{(i,j) : i \geq 0,\ 1 \leq j \leq m\}$, and the block partitioned matrix structure as in equation (3) is obtained by partitioning the state space into "levels" $\boldsymbol{i} = \{(i,j) : 1 \leq j \leq m\}$. For queueing models of this type, the level $\boldsymbol{i}$ is usually the set of states with the number of customers in the system equal to i, and j is an auxiliary variable such as the phase of the arrival process.

The starting point of the analysis of Neuts is to consider the first passage probabilities from level $\mathbf{1}$ to the level $\mathbf{0}$. These are the natural generalization of the quantity g in the scalar case. Specifically, let g_{ij} denote the probability that, starting in the state $(1,i)$, the chain eventually enters the level $\mathbf{0}$ and that the specific state of level $\mathbf{0}$ entered is $(0,j)$; and let g be the $m \times m$ matrix of elements g_{ij}. An argument conditioning on the first step of the chain shows that the matrix g satisfies the (matrix) equation (5), and it is not too hard to prove that g is also the minimal nonnegative solution of that equation. Note that g is a substochastic matrix in general, and that g must be stochastic in the ergodic case. Parenthetically, we note that the the roots z_i and the eigenvectors obtained in the classical approach can be shown—see [30]—to be none other than the eigenvalues and eigenvectors of the matrix g.

Remarkably, most quantities of interest for models of this type can be expressed in terms of the matrix g, and thus the computation of g becomes the most important step in the analysis. To illustrate this, consider the computation of the steady state probability vectors $\boldsymbol{\pi}_i$ with elements π_{ij}, where π_{ij} is the steady state probability of the state (i,j). We observe that $\boldsymbol{\pi}_0$, up to a multiplicative constant, is given by the steady state distribution of the embedded Markov chain at visits to level $\mathbf{0}$, and that embedded chain has transition matrix $\hat{P} = \sum_0^\infty b_n g^n$; in other words, once g is computed, the computation of $\boldsymbol{\pi}_0$, at least up to a multiplicative constant, becomes a simple (finite dimensional) linear problem. Once this is done, one can then compute the remaining terms recursively using a simple relationship obtained by Ramaswami [26]:

$$\boldsymbol{\pi}_{n+1} = \left[\boldsymbol{\pi}_0 \bar{b}_{n+1} + \sum_{i=1}^{n} \boldsymbol{\pi}_i \bar{a}_{n+2-i}\right] (I - \bar{a}_1)^{-1}, \tag{6}$$

where $\bar{a}_n = \sum_{k=n}^\infty a_k g^{k-n}$ and $\bar{b}_n = \sum_{k=n}^\infty b_k g^{k-n}$. Incidentally, the above recursion was obtained using a purely probabilistic argument considering the censored chains on the levels up to level $\boldsymbol{n+1}$, and avoids the potential slow convergence problems of iterative methods like Gauss-Seidel techniques. Indeed, its derivation was motivated by convergence problems encountered in the analysis of a meteor scatter communication system [9], a queue with server break-downs, whose very nature required the computation of a large number of sub-vectors $\boldsymbol{\pi}_n$—an instance of technology applications goading mathematics to step up.

The computation of the matrix g was initially accomplished by Neuts using simple iterative methods based on successive substitutions starting with the zero matrix:

$$g_0 = 0, \qquad g_{n+1} = \sum_{k=0}^\infty a_k g_n^k. \tag{7}$$

It is easy to demonstrate that the above iterative scheme provides a monotonically increasing (entry-wise) sequence of matrices converging to the desired matrix g.

Clearly, the equation (5) establishes g as a fixed point of the (contraction) map $f(x) = \sum a_n x^n$ defined on the set of nonnegative substochastic matrices x. In fact, one can establish the existence of a fixed point from standard fixed point theorems of functional analysis for contraction maps and also apply many results known from the general literature to that equation; see Neuts [16] and Ramaswami [25] for some details. There are two key findings from the cited references that deserve mention here: (a) the simple iterative method of equation (7) has only "linear convergence" and can become quite slow, particularly for models where the decay of the steady state distribution is not very fast; (b) standard techniques of acceleration, e.g., of the Newton type, often end up costing more in computation time and memory. These facts posed some serious mathematical challenges; their resolution will be discussed in a subsequent section.

3.2 GI/M/1 type chains

Having considered the matrix generalization of the $M/G/1$ queue, it was but natural that Neuts [18], [19] moved on to consider block partitioned matrices of the form in equation (4). An example of such a model is the embedded Markov chain in the $GI/E_m/1$ queue, the queue with renewal arrivals and an Erlang service time of order m; the class itself has numerous special cases [20].

Once again, we denote the Markov chain by $\{(X_n, J_n)\}$ and states by (i, j), $i \geq 0$, $1 \leq j \leq m$. The transition matrix of this Markov chain is assumed to be of the form (4) where the elements in the right side are now $m \times m$ matrices. We assume the chain to be ergodic and denote the steady state probabilities by π_{ij}. We also use the notion of levels and partitioning by levels as done in the M/G/1 case.

In the scalar case, i.e., $m = 1$, it is well-known that the steady state distribution π_n is geometric. Indeed, $\pi_n = \pi_0 r^n$, where r is the minimal solution in $(0, 1)$ of the equation

$$r = \sum_{n=0}^{\infty} r^n a_n. \tag{8}$$

Thus, it is natural to ask if in the matrix case, one would get for the steady state distribution $\boldsymbol{\pi}_n$ a matrix-geometric structure—i.e., could we write $\boldsymbol{\pi}_n = \boldsymbol{\pi}_0 r^n$ for some matrix r? The main result of [18] was to demonstrate that it indeed is the case, and that the matrix r (called "rate matrix") is the minimal nonnegative solution of the matrix equation (8). Since $\boldsymbol{\pi}_n \to \mathbf{0}$ as $n \to \infty$, the matrix r must have spectral radius (largest eigenvalue) less than one—a condition analogous to $r < 1$ in the scalar case. It was also shown that r can be obtained as the limit of the monotonically converging sequence of matrices defined by the recursions:

$$r_0 = 0, \quad r_{n+1} = \sum_{k=0}^{\infty} r_n^k a_k. \tag{9}$$

In a subsequent paper [19], Neuts also gave a probabilistic interpretation for r: r_{ij} is the expected number of visits to $(n+1, j)$ during a return to level $\boldsymbol{n}$, given that the chain starts in (n, i).

Several things are worth recalling here: (a) Just as in the scalar case, the derivation of the matrix-geometric result was by a trial-and-error scheme; (b) the simple iterations of equation (9) also have only linear convergence and could suffer from slow convergence, and typically so when $r \approx 1$; (c) although r is nonnegative and $sp(r) < 1$, it is not true that r is a stochastic or substochastic matrix. Finally, unlike that for g, the function $\sum r^n a_n$ is not a contraction map, and standard results on such maps do not apply directly [25].

3.3 Quasi Birth-and-Death Processes

The matrix analogue of the simple birth and death process is a Markov chain $\{(X_n, J_n)\}$ on the two-dimensional state space $\{(i, j) : i \geq 0, \ 1 \leq j \leq m\}$ and transition matrix of the form given in equation (2), where b_0 and a_i are matrices of order m. A canonical example is, of course, a birth and death process in a Markovian environment, where the birth and death rates depend on the environment state J. Hence such a process goes by the name Quasi-birth-and-death process (QBD). Clearly, this class of chains is a special case of both the $M/G/1$ type and the $GI/M/1$ type chains we discussed above, and either methodology applies. We have both the g matrix and the r matrix, with the same interpretations, but they now satisfy the matrix quadratic equations:

$$g = a_2 + a_1 g + a_0 g^2, \quad r = a_0 + r a_1 + r^2 a_2. \tag{10}$$

Note that since the quantities involved are matrices, the terms of the products may not be interchanged.

Once again, many interesting questions arise from the above. (a) Given the closeness in appearance of the equations for r and g, is there a relationship between them? (b) How special is the QBD case? (c) Can the fact that the equations involve only matrix-quadratics (and not an infinite sum as in the general cases considered earlier) lead to better methods?

Thus, while providing a powerful unified approach to a wide class of models, the matrix-analytic results also generated a host of other interesting questions: How general can the structure be made and the qualitative flavor of the results maintained? How can one speed up computations? What are the implications of the matrix-geometric result for derived quantities? It has taken several decades since the discovery of the above mentioned matrix-analytic results to answer some of these interesting questions, and much further questions do remain open. Those are what we discuss next.

For much of the early work on matrix-analytic methods and for numerous interesting examples to which they apply, we refer the reader to the two monographs of Neuts [21], [20]. A delineation of the theory as it stands today is presented in [15] using mainly Quasi Birth and Death Processes; that work, however, limits itself mostly to the core methodology and does not delve in detail into applications and examples.

4 OPERATOR UNDERPINNINGS

We noted that the early work on matrix analytic methods by Neuts restricted itself to finite dimensional blocks—i.e., to the case where the second co-ordinate J assumed only a finite set of values. Note that even with this, J could itself be multi-dimensional, and the requirement is simply that the totality of values assumed by J be finite.

The reasons for restricting to finite dimension were many: (a) The approach was by and large *ad hoc*; (b) computations were the main thrust of the approach; and (c) in proving the results, a heavy reliance was made on Perron-Frobenius theory for finite dimensional, nonnegative matrices. Nevertheless, it was conjectured very early on by Neuts that the approach was much more general and should have operator analogues in the general cases; the connection with nonlinear maps and fixed point theorems of functional analysis were highly suggestive of this.

Following a conjecture by Neuts, Tweedie [39] demonstrated soon that the matrix-geometric solution extends in a natural manner to an "operator geometric" form when one allows the second coordinate J to be in an arbitrary (Polish) space, but retains for the transition kernel the structure in (4); as an example, Tweedie considered the $GI/G/1$ queueing model at arrival epochs, with the elapsed service time of the customer in service playing the role of the phase variable. The techniques, however, continued to have an *ad hoc*—"let us guess and verify"— flavor, and did not really provide deep insights into wherefrom the geometric structure arises. Although Tweedie's work dates quite early in the story of these methods, for a long while, extensions to the operator case were not pursued further on account of perceived technical difficulties and doubts about their computational merit. Some recent work, however, shows that some of these fears may indeed be exaggerated, and we discuss some key developments below.

4.1 Level Crossing Arguments

In a tutorial paper on matrix-analytic methods, Ramaswami [31] demonstrated that the implication of a geometric structure by the the skip-free-to-the-right property and spatial homogeneity becomes obvious once we use a sample path argument using the level crossing approach. Furthermore, the techniques of proof can be entirely probabilistic and elementary (based on renewal theory for terminating renewal processes). The true merit of this approach lies in the fact that these arguments hold good even in the case when J is no longer finite-dimensional. Let us quickly recap the main ideas of [31].

Consider the chain of equation (4). Denote by $P(0, n+1; t)$ the matrix of probabilities of being in level $\boldsymbol{n+1}$ at time t given that the process starts in level $\mathbf{0}$ at time 0; we have a matrix since we need to keep track of the phase (J) at time 0 as well as at time t. Since the chain is skip-free-to- the-right, starting in level $\mathbf{0}$, it cannot be at level $\boldsymbol{n+1}$ at time t without being at level $\boldsymbol{n}$ at least once before time t. Indeed, there should be a unique time point $1 \leq k \leq t-1$ such that $X_k = n$ and $X_i > n$ for $i = k+1, \ldots, t$. Thus, we can write

$$P(0, n+1; t) = \sum_{k=1}^{t-1} P(0, n; k) Q(n, n+1; t-k), \tag{11}$$

where $Q(n, n+1; s)$ is the taboo probability that, having started in level $\boldsymbol{n}$ at time 0, the chain is at level $\boldsymbol{n+1}$ at time s avoiding the levels $\mathbf{0}, \ldots, \boldsymbol{n}$ in intermediate steps. Clearly, the successive visits to $\boldsymbol{n+1}$ avoiding lower levels form a Markov renewal process; that Markov renewal process is also terminating if the chain is irreducible. This implies, by taking the limit as $t \to \infty$ in equation (11), that the steady state probabilities must satisfy the relation

$$\boldsymbol{\pi}_{n+1} = \boldsymbol{\pi}_n \sum_{t=1}^{\infty} Q(n, n+1; t). \tag{12}$$

The matrix-geometric structure $\boldsymbol{\pi}_{n+1} = \boldsymbol{\pi}_n r$ is immediate by using the spatial homogeneity property which implies that

$$Q(n, n+1; t) \equiv Q(0, 1; t),$$

and therefore we may write $\sum_t Q(n, n+1; t) = r$ as a quantity independent of n. As mentioned earlier, the beauty of the above argument is that it does not require J to be finite-dimensional.

4.2 Continuum of Phases

Extending the methods to the case when the phase variable (J) assumes infinitely many values is not only a matter of mathematical curiosity, but also of practical value. In many practical queueing systems, particularly in the high speed communications area using the internet protocols, there is evidence—see [23], [1]—of quantities such as delay distributions exhibiting heavy tails and of traffic processes exhibiting long range dependence and self-similarity over many time scales. The matrix-geometric result $\boldsymbol{\pi}_n = \boldsymbol{\pi}_0 r^n$ with a finite dimensional matrix r automatically implies geometric decay of the steady state distribution. For queueing models, this would in turn imply an exponentially decaying tail for waiting time distributions and certain hitting time distributions of interest. These imply that the models covered by the finite dimensional matrix analytic theory may be inadequate for certain applications of current interest.

In using matrix-analytic methods in such contexts, the trend in the literature has been to approximate complex distributions and processes with finite dimensional phase type distributions; see [11] for an example. Typically, these use increasingly larger and larger number of phases to effect better approximations, by pushing the exponential asymptotics farther out. A more direct method may be to construct infinite dimensional models up front. Although a truncation will be needed for numerical computations, the general mathematical framework could possibly provide guidelines for such truncations and for developing useful asymptotic results. Coming to using a continuum of values for the variable J, its use is predicated on our belief that continuous models are much more convenient from a mathematical modeling perspective, particularly in dealing with phenomena such as self-similarity.

The extension to a countable of number of phases is almost immediate and does not change the results much. However, using a continuum of phases changes the mathematical setting significantly. Leaving aside the measure theoretic aspects, operationally, we need to replace transition matrices by transition kernels, and matrix multiplication by composition integrals of the form

$$ab(x, y) = \int a(x, z) b(z, y)\, dz. \tag{13}$$

Thus, when J_n is allowed to assume a continuum of values, a skip-free-upward Markov chain $\{(X_n, J_n)\}$ has a transition kernel of the form in equation (4) with the major difference that the a_i, b_i are now transition kernels; in these, we interpret $a_i(y, z)$ and $b_i(y, z)$ as the probability densities of a change to the value z from the value y of the phase variable J along with a jump of appropriate size in the value of the level variable X.

We have shown in [22] that the level crossing argument, carried out in the context of a suitably defined semi-Markov process, will yield an operator-geometric form for the steady state density $\pi_n(y)$, the steady state density of the state (n, y). Specifically,

$$\pi_n(y) = \int \pi_{n-1}(z) r(z, y)\, dz, \tag{14}$$

where the kernel $r(z, y)$ is the minimal nonnegative solution of the equation (8) with matrix multiplications being replaced by composition integrals as defined in equation (13).

A key result we demonstrated in [22] is that for sufficiently smooth kernels a_n that can be expanded in terms of an orthonormal family of functions, the kernel r also has a similar expansion, and that the coefficients of the expansion for r can be computed using steps similar to those in the matrix case. While the matrices in such computations do not have some key properties (nonnegativity and substochasticity), yet it appears possible to effect accurate computations using the more recent algorithms that limit the number of steps in the iterations significantly. Unfortunately, this work has not attracted enough attention, and there is significant scope for developing useful models and analysis along the directions given there.

From the perspective of theory, the extension of the results to the countable case has already yielded some interesting results. In Taylor & Ramaswami [29] we demonstrated how product form results for networks of queues come from certain special properties of the operator in the operator-geometric characterization of the steady state distribution. That also helped to discover some new product form networks. Nevertheless, not much else is known in this area at the present time.

A systematic examination of the infinite dimensional extensions, and particularly the continuous case, would be beneficial to application areas such as high speed networking. Although we used classical special functions in our work, interesting are the possibilities of using others such as wavelet functions for the orthonormal expansions, particularly in the context of modeling self-similarity and scaling behaviors. This is an open area for research where hardly anything has been done yet.

4.3 Continuum of Levels

There is obviously an analogous theory for two-dimensional Markov processes in continuous time that exhibit skip-free properties in at least one direction. The literature covers such processes as well.

In the continuous time context, one may also wish to consider the case where the level variable is defined over a continuum, say $[0, \infty)$. The skip-free properties now are simply appropriate continuity conditions on the paths of the coordinate X; and, modulo technicalities, the level crossing arguments continue to hold.

The continuous analogue of the geometric sequence is the exponential function. It is therefore natural for one to conjecture that models of this type have a "matrix exponential" steady state distribution with "a generator" that can be obtained as the solution of a nonlinear matrix equation. Such extensions of the matrix analytic methods can be found in Sengupta [36] and Ramaswami [28], [33]; the level crossing idea is again the fundamental tool.

While space here does not allow a detailed presentation of results of the matrix exponential type, we do wish to make some comments, particularly in the case of the so-called "Markov-modulated fluid flow" model of [33].

The fluid model of [33] is a Markov model $\{(X_t, J_t)\}$, in which J_t is a continuous time Markov chain assuming a finite set of values, and X_t changes continuously at constant rate α_i in those intervals for which $J_t = i$. Such models are used, for example, in the context of high speed communications, where the flow of information bits is modeled as a continuous fluid, and the Markovian environment either depicts some aspect of the system such as the number of active connections or is a simple modeling artifice to capture randomly fluctuating changes in traffic rates.

Traditional techniques of analysis of such models were based on partial differential equations obtained from the Chapman-Kolmogorov equations—see [2]. Due to singularities, the numerical solution of those equations is unstable. The methods based on the operator analytic approach in [33], however, yield the steady state distribution in a matrix exponential form in terms of a generator matrix that can be computed using stable algorithms. Indeed, the computation of the generator matrix has been reduced in [33] to the computation of the matrix g for a suitably defined Quasi-birth-and-death process in *discrete* time with a *discrete* state space. We refer the reader to [33] for the details of the analysis and the implications of the results.

We conjecture that the advantages of our new approach to fluid flow models will go beyond the development of stable algorithms for numerical computation. At present, there is no elegant and tractable way in which one can incorporate phenomena such as heavy tails and self-similarity in the context of fluid models. Even as a formal solution, the traditional approach based on differential equations stops with the finite dimensional case. The level crossing analysis which is at the heart of our approach, however, extends naturally, to the case where the phase space is infinite dimensional; and combined with techniques similar to those in [22], it may actually provide a tractable way to extend the fluid models in a manner useful for the cited applications. Once again, nothing has yet been done in this direction, and this is a worthwhile area for much further research.

4.4 General Skip-free Processes

Of the two assumptions that underlie the structures discussed thus far, spatial homogeneity (the repeating row or column structure) could be relaxed. This allows for consideration of such examples as state dependent queues, where transition rates change with "level;" the "shorter of the two queues problem" is a simple example. From applications perspective, there is certainly considerable interest in such extensions since in many practical systems, some form of state dependent control of the process is exercised (e.g., different levels of congestion control triggered at various queue sizes) that may nullify spatial homogeneity without affecting the skip-free nature. With the skip-free structure still available, one can invoke the level crossing argument and extend many results of the matrix analytic type—see [31], [30], [29], [8]. What is lost in relaxing the spatial homogeneity assumption is the ability to reduce everything to the solution of a single nonlinear matrix equation; we get a nonlinear matrix equation for each level separately. Although some formulas have been generalized to such a general case (e.g., the "logarithmic reduction method" in [29]), at this level of generality, from a computational perspective, not much more can be done. The work of Bright and Taylor [8], however, serves to show that there is scope for developing results that hold good across large subclasses of problems. In addition to the many open algorithmic questions related to these, there are also interesting mathematical issues related to tail behavior of steady state distributions, convergence rates, etc., that are wide open.

5 UNITY OF THE STRUCTURES

This section is a brief interlude to demonstrate the interplay between purely mathematical and algorithmic work in this area. They relate to two important developments related to the various structures considered thus far.

5.1 Duality of the Structures

Historically, the methodologies for skip-free-upward and skip-free-downward processes were developed in parallel and independently of each other. However, the striking similarities in many of the results, particularly the nonlinear matrix equations (5) and (8) were begging some result unifying the two paradigms. Motivated by some equations in a work of Asmussen on a special case, Ramaswami [27] obtained some results that accomplished this through an argument that, in its generality, was later clarified by him and Asmussen [4] as rooted in time reversal.

First of all, note that at first sight it appears that one can obtain an equation of either form (5), (8) from the other by matrix transposition. However, simple transposition takes one out of the probabilistic framework altogether. The matrices a_n in these equations are not arbitrary, but nonnegative matrices such that $\sum_n a_n$ is row stochastic; transposition could destroy the row stochasticity property.

What was obtained in [27] was algebraically a simple fact. Consider the common case where $a = \sum_n a_n$, the matrix of phase transitions, is irreducible so that a has a steady state probability vector $\boldsymbol{\xi}$. Let Δ denote a diagonal matrix with $\boldsymbol{\xi}$ on the main diagonal.

Given a g satisfying equation (5), it is trivial to verify that the matrix $r = \Delta^{-1} g' \Delta$ satisfies the equation $r = \sum_n r^n c_n$, an equation of the type (8), with $c_n = \Delta^{-1} a_n' \Delta$ and that the matrices c_n are nonnegative and such that $\sum_n c_n$ is row stochastic. Thus, associated with the g of each skip-free downward chain, there is an associated r of a suitably defined skip-free upward chain. A similar result clearly holds the other way too relating each r to an appropriately defined matrix g.

Readers familiar with time reversal for Markov chains will not be surprised to learn that the above are rooted in the notions of duality and time reversal. A first passage interval from level $\boldsymbol{n+1}$ to level $\boldsymbol{n}$ in a skip-free downward process, when looked at backwards in time, is after all an interval that takes one from $\boldsymbol{n}$ to $\boldsymbol{n+1}$ in the dual where upward and downward movements in levels have been reversed and states below level $\boldsymbol{n}$ are taboo; the dual is, of course, a process skip-free upward. These results are elementary in the scalar case and are well-known—see Takacs [38]; but they are quite non-trivial in the general case, since we not only have to switch upward and downward jumps of the level process, but we also have to simultaneously reverse the phase process at the same time.

From a theoretical perspective, the duality results demonstrated the unity of the two structures that underly the matrix analytic approach. Besides this, they have had many other uses. They are useful for proving, using only elementary arguments, some interesting results such as the phase type nature of waiting time distributions in many queues [28]. They have been used recently to develop algorithms in other contexts such as spatial point process models [35] of phase type. They are also fundamental to the argument that reduces the analysis of the fluid flow model to that of a discrete time QBD [33].

Yet, this theoretical result, born out of mathematical curiosity, is not without algorithmic value either. We noted earlier that while the matrix g is always substochastic, nothing much can be said about r except that it is nonnegative. The duality result allows us to transform the computation of r always to an equivalent problem of computing g, wherein the iterates remain in a bounded set and make the algorithms behave in predictable ways.

5.2 Generality of the QBD

Recently, Ramaswami [32] demonstrated that *any* discrete time, discrete state space Markov chain can be obtained as an embedded chain in a QBD (with possibly level dependent, infinite dimensional blocks). He used this to deduce the equations for the general structures from those of the QBD, showing thereby that QBDs and the quadratic equations are both fundamental and very general in the study of Markov chains.

Stripped of all notation and technicalities, the idea of [32] is extremely simple: given a Markov chain that is not skip-free, for each step introduce an additional variable that tracks the number of levels by which the chain must jump and pretend as though there are intermediate steps (in the enlarged process) wherein this number gets depleted by one in each step and the level variable gets changed only by ± 1. The original process is equivalent to the embedded Markov chain obtained at epochs when the newly introduced variable assumes the value zero.

This work, which at first sight, appears to be an exercise to satisfy some mathematical curiosities, however, has recently turned out to show value in algorithmic contexts as well. Specifically, the technique of converting to a QBD has led to quite efficient algorithms for much more general classes of problems —see [7]. Also, researchers are currently using these ideas to develop efficient algorithms for solving general matrix polynomial equations [D. Bini - personal communication].

6 ALGORITHMS

Having discussed the scope of the methods in terms of the structures to which they apply and the theoretical underpinnings, we are now ready to recap some of the major developments in the algorithmic arena.

6.1 Simple Iterations

We noted in Section 3 that the simple iterative techniques of (7) and (9) have only linear convergence and could bog down for models such as nearly saturated queues. Thus, there has been significant interest in the literature from early on in finding faster procedures for solving the nonlinear matrix equations of interest—see [25]. Concurrently, there have also been attempts to interpret the steps of various algorithms in a probabilistic sense, mostly as a matter of curiosity. Important among these is a piece of work by Latouche [13] that developed an alternate version of the simple iteration that lends itself to nice probabilistic interpretations.

In [13], Latouche considered the Quasi-birth-and-death process of equation (2), and introduced an additional matrix u defined as follows: u_{ij} is the conditional probability, starting in the state $(1, i)$, of returning to the level **1** in the state $(1, j)$ avoiding the lower level **0** in intermediate steps. Recall that for this model, g is the matrix of probabilities of an eventual first passage from level **1** to level **0**. With the definition of u as stated, it is easy to show that u and g satisfy the equations

$$u = a_1 + a_0 g, \quad g = (I - u)^{-1} a_2. \tag{15}$$

Of these, the first equation follows by a simple conditioning on the first step of the chain, and the second by considering the number of returns to level **1** before the first visit to level **0**. These naturally suggest the recursions:

$$g_0 = 0; \quad u_n = a_1 + a_0 g_n; \quad g_{n+1} = (I - u_n)^{-1} a_2 \tag{16}$$

Latouche showed that the matrices u_n and g_n are (element-wise) monotonically increasing in n and converge as $n \to \infty$ to the matrices u and g respectively. Furthermore, he showed that u_n and g_{n+1} are the respective first passage probabilities under the taboo condition that during the passage interval no visit can be made to levels above $\boldsymbol{n}$.

There are some interesting findings that come out immediately from the above approach. With each iteration in the algorithm, we are grabbing paths that span exactly one additional level. For this reason, the algorithm may be called a "linear algorithm." The number of steps needed to get an accuracy of ϵ is N_ϵ, the smallest index for which a level has probability less than ϵ of being entered during a first passage to level $\mathbf{0}$. Naturally, if a model is such that a very large level $\boldsymbol{n}$ will be visited in the first passage to level $\mathbf{0}$ from level $\mathbf{1}$ (as for example in a nearly saturated queue where busy periods may see large build ups of the queue), then a large number of steps would be required.

Thus, the above variant of the simple iteration scheme in equation (5) helps to show the difficulties associated with simple iterative schemes in a probabilistically clear manner. One of the major advantages of the algorithmic approach has been the ability it provides to interpret the steps of the analysis directly in terms of the dynamic behavior of the model.

6.2 Quadratically Convergent Algorithms

For a long while, the difficulties with convergence (in certain cases like saturated queues) went ignored in the belief that these were only of pathological interest. Waking us all up came then a paper by Daigle and Lucantoni [10] that considered a QBD model of a voice-data application and demonstrated the unacceptable performance of the simple iterative scheme and many of its known variants even for some practical problems. This paper accelerated the quest for an implementable, quadratically convergent algorithm.

Let us now turn to the QBD and recall that the simple iteration process suffers slow convergence, when for large values of n, the probability of visiting level $\boldsymbol{n}$ cannot be ignored. Thus, in these cases, the decay parameter of the steady state distribution $\boldsymbol{\pi}_n = \boldsymbol{\pi}_0 r^n$, which of course is the largest eigenvalue of r, must be close to unity. This leads one to ask the natural question: why not then attempt to compute r^2 (or some higher power of r) directly, because once we know r^2 we can, after all, get r from a simple linear equation—recall (8). Latouche and Ramaswami [14], effectively turned this idea into a quadratically convergent algorithm (that also avoided the computational burden of the Newton method.) Following are the key ideas used.

Taking note of the duality and the advantage of considering g, the methods were developed in [14] in terms of g. Writing g in terms of g^2 gives

$$g = \hat{a}_2 + \hat{a}_0 g^2, \tag{17}$$

where $\hat{a}_i = (I - a_1)^{-1} a_i$. Now, one observes that the matrix g^2 is the g-matrix of the QBD obtained at visits to even numbered levels $\{\mathbf{0}, \mathbf{2}, \mathbf{4}, \dots\}$, and we can apply a similar argument to write g^2 in terms of g^4. Continuing this process *ad infinitum,* it is possible to obtain an "explicit" formula for g as:

$$g = \sum_{k\geq 0} \left(\prod_{0\leq i\leq k-1} a_0^{[i]} \right) a_2^{[k]}, \tag{18}$$

where the matrices appearing in the right side can be computed recursively as:

$$a_i^{[0]} = (I - a_1)^{-1} a_i,$$

and

$$a_i^{[k+1]} = \left(I - u^{[k]}\right)^{-1} \left(a_i^{[k]}\right)^2,$$

with

$$u^{[k]} = a_0^{[k]} a_2^{[k]} + a_2^{[k]} a_0^{[k]}.$$

With a little effort, it is easy to demonstrate that the matrices $a_i^{[k]}$ appearing in (18) are indeed the blocks of the embedded QBD at level transitions when the levels visited are of the form $n \times \mathbf{2^k}$ for some n. This enables an interpretation of the sum in (18) truncated at N as an approximation to the matrix of first passage probabilities obtained under taboo of the levels above level $\mathbf{2^N}$.

Thus, the new algorithm based on successive truncations of the sum in (18) captures in its n-th step all paths that go up to level $\mathbf{2^n}$. That is what gives quadratic convergence. In fact, it is proved in [14] that the rate of convergence is η^{2^n}, where η is the largest eigenvalue of the matrix r. To grasp the practical meaning of this result, note that if under this new algorithm, a certain number N of iterations is needed to attain an accuracy of ϵ, then the process can cross level $\mathbf{2^N}$ with a probability greater than ϵ; in other words, the algorithm can handle even processes subject to enormous excursions to very high levels.

The above work naturally lead to the consideration of the more general models, and soon, numerical analysts Bini and Meini [5], [6] obtained a set of quadratically convergent algorithms for the general $M/G/1$ type chains using cyclic reduction methods for Hessenberg matrices. Curiously, they found that for the QBD process, their algorithms, modulo minor differences, were equivalent to the algorithm of [14]. To complete the story, recent work cited in Section 5.2 has shown that even in the general case, it is indeed equivalent to the algorithm obtained in [14].

These in toto have served to establish the matrix/operator analytic approach as a viable solution method for a very large class of stochastic models.

7 CONCLUDING REMARKS

The story of matrix/operator analytic methods for stochastic models illustrates the major impact the progress in computing and communications technologies is having on mathematical disciplines. For us dealing with stochastic modeling, the increased ability to compute has not only enabled the examination of models in greater generality, but it has also provided ways to model more faithfully, to analyze models in a unified manner, and to give results in a usable form. As in other instances, here too, while computing has been a great aid to mathematical analysis, it certainly does not replace it. The questions have changed, or rather the set of questions has become larger, but the questions remain fundamentally mathematical in nature. Also, the technologies which are aiding mathematics are simultaneously goading mathematical research further. Indeed, the interplay among mathematics, computing and applications is ever increasing as are the challenges for mathematics and the mathematician.

References

[1] P. Abry & D. Veitch: Wavelet analysis of long-range dependent traffic, *IEEE Trans. Inf. Theory*, 44, 2–15, 1998.

[2] D. Anick, D. Mitra & M.M. Sondhi: Stochastic theory of a data handling system with multiple sources, *Bell System Tech. J.,* 61, 1871–1894, 1982.

[3] S. Asmussen: *Applied Probability and Queues*, Wiley, 1987.

[4] S. Asmussen & V. Ramaswami: Probabilistic interpretations of some duality results for the matrix paradigms in queueing theory, *Stochastic Models*, 6, 715–734, 1990.

[5] D. Bini & B. Meini: On the solution of a nonlinear matrix equation arising in queueing problems, *SIAM J. Matrix Anal. Appl.*, 17, 906–926, 1996.

[6] D. Bini & B. Meini: Improved cyclic reduction for solving queueing problems, *Numerical Algorithms*, 15, 57–74, 1997.

[7] D. A. Bini, B. Meini & V. Ramaswami: Analyzing M/G/1 paradigms through QBDs: the role of the block structure in computing the matrix G, in *Advances in Algorithmic Methods for Stochastic Models*, G. Latouche & P. Taylor (Eds.), Notable Publications, Inc., NJ, 2000.

[8] L.W. Bright & P.G. Taylor: Calculating the equilibrium distribution of level dependent quasi birth and death processes, *Stochastic Models*, 11, 497–525, 1995.

[9] Y. Chandramouli, M.F. Neuts & V. Ramaswami: A queueing model for meteor burst packet communication systems, *IEEE Trans. Comm.*, 37, 1024–1030, 1989.

[10] J.N. Daigle & D.M. Lucantoni: Queueing systems having phase dependent arrival and service rates, in *Numerical Solutions of Markov Chains*, William J. Stewart (Ed.), 161–202, Marcel Dekker, NY, 1991.

[11] A. Horváth & M. Telek: Approximating heavy tailed behaviour with phase type distributions, in *Advances in Algorithmic Methods for Stochastic Models*, G. Latouche & P.G. Taylor (Eds.), 191–214, Notable Publications, Inc., NJ, 2000.

[12] S. Karlin & H.M. Taylor: *A first course in stochastic processes*, Academic Press, NY, 1975.

[13] G. Latouche: Algorithms for infinite Markov chains with repeating columns, *IMA Workshop on Linear Algebra, Markov Chains and Queueing Models,* January 1992.

[14] G. Latouche & V. Ramaswami: A logarithmic reduction algorithm for quasi birth and death processes, *J. Appl. Prob.*, 30, 650–674, 1993.

[15] G. Latouche & V. Ramaswami: *Introduction to matrix analytic methods in stochastic modelling*, SIAM & ASA, 1999.

[16] M.F. Neuts: The Markov renewal branching process, in *Proc. of the Conference on Mathematical Methodology in the Theory of Queues, Kalamazoo, Michigan*, NY, Springer-Verlag, 1–21, 1974.

[17] M.F. Neuts: Queues solvable without Rouché's theorem, *Opns. Res.,* 27, 767-781, 1978.

[18] M.F. Neuts: Markov chains with applications in queueing theory, which have a matrix-geometric invariant vector, *Adv. Appl. Prob.*, 185–212, 1978.

[19] M.F. Neuts: The probabilistic significance of the rate matrix in matrix-geometric invariant vectors, *J. Appl. Prob.*, 17, 291–296, 1980.

[20] M.F. Neuts: *Matrix-geometric solutions in stochastic models, An Algorithmic approach*, The Johns Hopkins University Press, Baltimore, MD, 1981.

[21] M.F. Neuts: *Structured stochastic matrices of M/G/1 type and their applications*, Marcel Dekker, NY, 1989.

[22] B.F. Nielsen & V. Ramaswami: A computational framework for quasi birth and death processes with a continuous phase variable, in *Teletraffic Contributions for the Information Age: Proceedings of the 15th International Teletraffic Congress*, V. Ramaswami & P.E. Wirth (Eds.), Elsevier, 477–486, 1997.

[23] K. Park & W. Willinger (Eds.): *Self-Similar Network Traffic and Performance Evaluation*, Wiley, 2000.

[24] V. Ramaswami: The busy period of queues which have a matrix geometric steady state probability vector, *Opsearch*, 19, 238–261, 1982.

[25] V. Ramaswami: Nonlinear matrix equations in applied probability - solution techniques and open problems, SIAM Review, 30, 256–263, 1988.

[26] V. Ramaswami: A stable recursion for the steady state vector in Markov chains of M/G/1 type, *Stochastic Models*, 4, 183–188, 1988.

[27] V. Ramaswami: A duality theorem for the matrix paradigms in queueing theory, *Stochastic Models*, 6, 151–161, 1990.

[28] V. Ramaswami: From the matrix geometric to the matrix exponential, *QUESTA*, 6, 222–260, 1990.

[29] V. Ramaswami & P.G. Taylor: An operator-analytic approach to product-form networks, *Stochastic Models*, 12, 121–142, 1996.

[30] V. Ramaswami & P.G. Taylor: Some properties of the rate operators in level dependent quasi birth and death processes with a countable number of phases, *Stochastic Models*, 12, 143–164, 1996.

[31] V. Ramaswami: Matrix analytic methods: a tutorial overview with some extensions and new results, in *Matrix analytic methods in Stochastic Models*, S.R. Chakravarthy & A.S.Alfa (Eds.), Marcel Dekker, NY, 261–295, 1997.

[32] V. Ramaswami: The generality of the quasi birth and death process, in *Advances in Matrix Analytic Methods for Stochastic Models*, A.S. Alfa & S. Chakravarthy (Eds.), Notable Publications, Inc., NJ, 1998.

[33] V. Ramaswami: Matrix analytic methods for stochastic fluid flows, in *Teletraffic Engineering in a Competitive World, ITC-16 Proceedings*, P. Key & D. Smith (Eds.), Elsevier, 1019–1030, 1999.

[34] V. Ramaswami: The surprising reach of the matrix analytic approach, Conference on Stochastic Processes & Applications, Cochin University, India, December 1999; *to appear in the Proceedings.*

[35] M-A. Remiche: *Isotropic Phase Planar Point Processes: Analysis and Applications to Cellular Mobile Telecommunication*, Ph.D. Dissertation, Université Libre de Bruxelles, Brussels, 2000.

[36] B. Sengupta: Markov processes whose steady state distribution is matrix exponential with an application to the GI/G/1 queue, *Adv. Appl. Prob.*, 21, 159–180, 1989.

[37] L. Takács: *Introduction to the theory of queues*, Oxford University Press, NY, 1962.

[38] L. Takács: *Combinatorial Methods in the Theory of Stochastic Processes*, Wiley, NY, 1967.

[39] R.L. Tweedie: Operator-geometric stationary distribution for Markov chains, with applications to queueing models, *Adv. Appl. Prob.*, 14, 368–391, 1982.