The Physics of Matrix-Analytic Algorithms

Peter Taylor

Department of Mathematics and Statistics,
The University of Melbourne

Australian Research Council
Centre of Excellence for Mathematics
and Statistics of Complex Systems

Matrix-Analytic Methods

- In the 1970s and 1980s Marcel Neuts proposed a class of techniques for analysing Markov chains with block-structured transition matrices that have become known as matrix-analytic methods.
- More recently, other classes of models have been analysed using similar techniques.
- The interaction of mathematical analysis and physical insight has played an important role in the development of results in this area.
- There is an emphasis on computability of performance measures and, in particular, on algorithmic development.

Matrix-Analytic Methods

I shall discuss physical interpretations of Matrix-Analytic algorithms in the context of

- block-structured continuous-time Markov chains - specifically chains of GI/M/1-type, M/G/1-type and Quasi-Birth-and-Death processes (QBDs), and
- stochastic fluid models.

In each case, my emphasis will be on understanding the nature of sample paths that are taken into account at different stages of the algorithm.

Block-structured Markov chains

A discrete-time *Markov chain of $GI/M/1$-type* has a two-dimensional state space. The first dimension is countably-infinite and the second dimension is finite. When the chain is in state (k, i), we say that it is in *level k* and *phase i*. With a suitable ordering of the states, the transition matrix can be written in the form

$$P = \begin{bmatrix} \tilde{A}_1 & A_0 & 0 & 0 & \cdots \\ \tilde{A}_2 & A_1 & A_0 & 0 & \cdots \\ \tilde{A}_3 & A_2 & A_1 & A_0 & \cdots \\ \tilde{A}_4 & A_3 & A_2 & A_1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

Block-structured Markov chains

The terminology *Markov chain of $GI/M/1$-type* comes from the fact that the embedded Markov chain generated by a $GI/M/1$ queue observed at arrival points has a transition matrix of the form

$$P_G = \begin{bmatrix} \tilde{a}_1 & a_0 & 0 & 0 & \cdots \\ \tilde{a}_2 & a_1 & a_0 & 0 & \cdots \\ \tilde{a}_3 & a_2 & a_1 & a_0 & \cdots \\ \tilde{a}_4 & a_3 & a_2 & a_1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

where a_k is the probability that there are k services during an inter-arrival interval and $\tilde{a}_k = \sum_{\ell=k}^{\infty} a_\ell$.

Block-structured Markov chains

A discrete-time *Markov chain of $M/G/1$-type* has a state space of identical structure. Rather than being block skip-free to the right, it is block skip-free to the left, so that its transition matrix can be written in the form

$$P = \begin{bmatrix} \tilde{A}_1 & \tilde{A}_2 & \tilde{A}_3 & \tilde{A}_4 & \cdots \\ A_0 & A_1 & A_2 & A_3 & \cdots \\ 0 & A_0 & A_1 & A_2 & \cdots \\ 0 & 0 & A_0 & A_1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

Block-structured Markov chains

As for Markov chains of $GI/M/1$-type, the terminology comes from the fact that the embedded Markov chain generated by a $M/G/1$ queue observed at departure points has a transition matrix of the form

$$P = \begin{bmatrix} \tilde{a}_1 & \tilde{a}_2 & \tilde{a}_3 & \tilde{a}_4 & \cdots \\ a_0 & a_1 & a_2 & a_3 & \cdots \\ 0 & a_0 & a_1 & a_2 & \cdots \\ 0 & 0 & a_0 & a_1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

where a_k is the probability that there are k arrivals during a service time and $\tilde{a}_k$ is the probability that the first service time in a busy period will finish with $k-1$ customers waiting in the queue. In this simple case, $\tilde{a}_k = a_{k-1}$.

Block-structured Markov chains

From a modelling point of view, the second dimension has many uses. For example, it can be used to denote

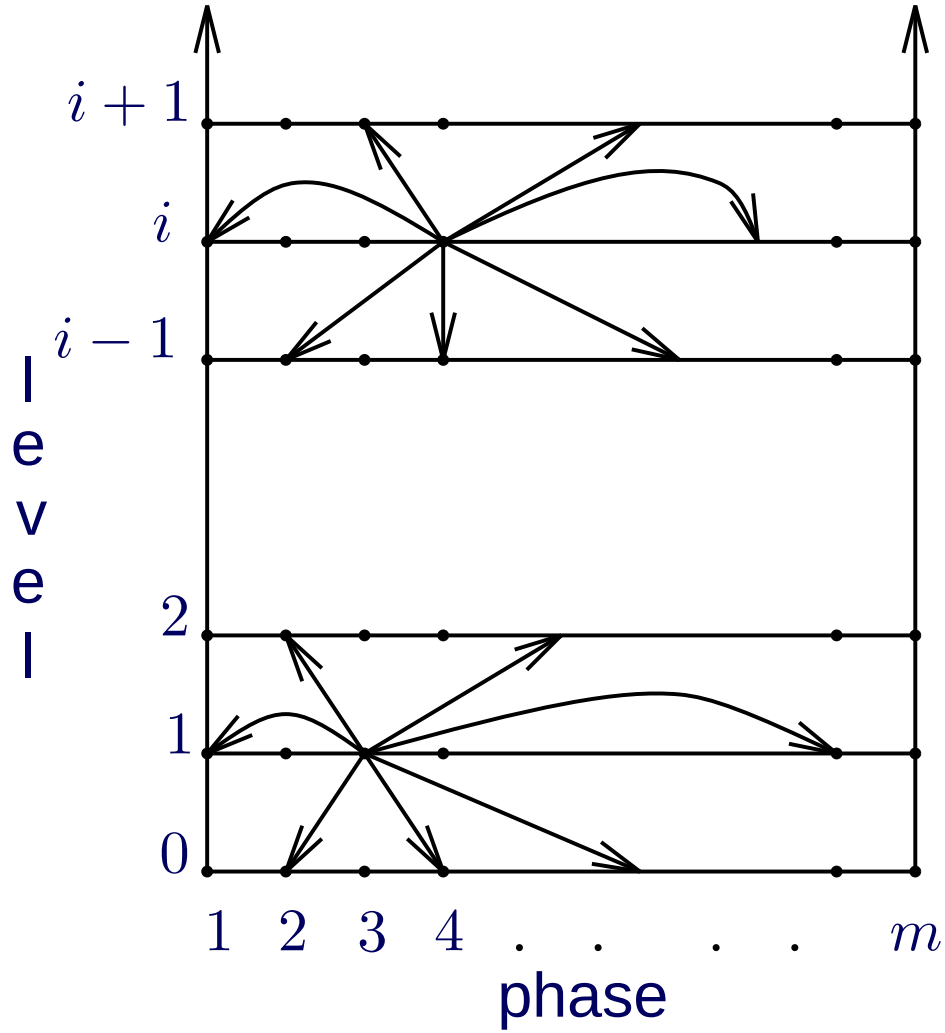
- the state of an independently-moving environment,
- the number of transmitting sources,
- the progress of one or more phase-type random variables,
- the number of customers in an associated queue,
- the underlying state of a hidden Markov chain model,
- information about customer mix,
- etc.

Block-structured Markov chains

Markov chains that are both of $GI/M/1$-type and $M/G/1$-type are known as *Quasi-Birth-and-Death Processes (QBDs)*. Their transition matrices can be written in the form

$$P = \begin{bmatrix} \tilde{A}_1 & \tilde{A}_0 & 0 & 0 & \cdots \\ A_2 & A_1 & A_0 & 0 & \cdots \\ 0 & A_2 & A_1 & A_0 & \cdots \\ 0 & 0 & A_2 & A_1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

A Quasi-Birth-and-Death Process



Chains of $GI/M/1$-type

For a discrete-time chain of $GI/M/1$-type, let $\boldsymbol{x}$ be the solution to

$$\boldsymbol{x}\left[\sum_{k=0}^{\infty} A_k\right] = \boldsymbol{x}.$$

Then the chain is positive recurrent, null recurrent or transient according as

$$\boldsymbol{x}A_0\boldsymbol{e}' - \boldsymbol{x}\left[\sum_{k=2}^{\infty}(k-1)A_k\right]\boldsymbol{e}',$$

is less than, equal to or greater than zero.

Chains of $GI/M/1$-type

Write the stationary distribution of a discrete-time positive recurrent chain of $GI/M/1$-type as $\boldsymbol{\pi} = (\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \ldots)$. Then there exists a matrix R such that

$$\boldsymbol{\pi}_n = \boldsymbol{\pi}_0 R^n.$$

The vector $\boldsymbol{\pi}_0$ satisfies

$$\boldsymbol{\pi}_0 \left[\sum_{k=0}^{\infty} R^k \tilde{A}_{k+1} \right] = \boldsymbol{\pi}_0.$$

This is the well-known *matrix-geometric form* of the stationary distribution.

Chains of $GI/M/1$-type

The matrix R is the minimal nonnegative solution to the matrix equation

$$\sum_{k=0}^{\infty} R^k A_k = R.$$

The (i,j)th entry of the matrix R is the expected number of visits to phase j of level 1 before first return to level 0 conditional on the process starting in phase i of level 0.

In general, R has spectral radius which is less than or equal to one, and the chain is positive recurrent if and only if the spectral radius of R is less than one.

Chains of $M/G/1$-type

Chains of $M/G/1$-type do not have a matrix-geometric stationary distribution. To derive the stationary distribution, we use the fact that $\boldsymbol{\pi} = (\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_n)$ is proportional to the stationary distribution of the finite-state Markov chain with transition matrix

$$P = \begin{bmatrix} \tilde{A}_1 & \tilde{A}_2 & \tilde{A}_3 & \cdots & \tilde{A}_n & \sum_{k=0}^{\infty} \tilde{A}_{k+n+1} G^k \\ A_0 & A_1 & A_2 & A_3 & \cdots & \vdots \\ 0 & A_0 & A_1 & A_2 & \cdots & \vdots \\ 0 & 0 & A_0 & A_1 & \cdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \ddots & \sum_{k=0}^{\infty} A_{k+2} G^k \\ 0 & 0 & 0 & \cdots & A_0 & \sum_{k=0}^{\infty} A_{k+1} G^k \end{bmatrix}.$$

Chains of $M/G/1$-type

The (i,j)th entry of the matrix G is the probability that the chain hits level $k-1$ in finite time, and does so in phase j, given that it starts in phase i of level k.

Clearly G is substochastic and it is stochastic if and only if the chain is recurrent.

Elementary arguments show that the matrix G is the minimal nonnegative solution to the matrix equation

$$\sum_{k=0}^{\infty} A_k G^k = G.$$

Chains of $M/G/1$-type

G is a matrix of probabilities, rather than a matrix of expected values (as R is). As such, it is a 'nicer' object to work with. Furthermore, for a QBD, the matrix R can be written in terms of the matrix G via the relation

$$R = A_2 \left[I - A_1 - A_0 G \right]^{-1}$$

and, for a chain of $GI/M/1$-type, the matrix R can be written in terms of the matrix G for the *dual chain* of $M/G/1$-type. For this reason, we concentrate on algorithms for calculating G.

To keep the notation simple, I shall discuss the QBD special case.

Calculating the matrix G

In the QBD special case, G is the minimal nonnegative solution to the matrix quadratic equation

$$A_2 + A_1 G + A_0 G^2 = G.$$

This equation has an analytic solution only in a few special cases. In general, we have to resort to numerical solution.

A simple procedure

For an irreducible QBD, A_1 is invertible. So, an obvious first approach to solving this equation is to transform it into a fixed-point equation:

$$\begin{aligned}(I - A_1)G &= A_2 + A_0G^2 \\ \Rightarrow G &= (I - A_1)^{-1}\left[A_2 + A_0G^2\right]\end{aligned}$$

and use the iterative procedure

$$G_{n+1} = (I - A_1)^{-1}\left[A_2 + A_0G_n^2\right]$$

with $G_0 = 0$.

A simple procedure

Neuts showed that, with this iteration, G_n does converge to G.

Furthermore, except when the QBD is null-recurrent, this convergence is *linear*. That is, there exists a constant $\gamma \in (0, 1)$ such that

$$\limsup_{n \to \infty} ||G_n - G||^{1/n} = \gamma.$$

Physical interpretations

The type of question that we shall be interested in is

Can we give a physical interpretation to the nth iterate of procedures such as the one described above?

For Neuts' original iteration, this question has not had a precise answer until recently. I shall say something about it at the conclusion of the talk.

In general, to understand physical interpretations of the type that I shall discuss here, we need to know about *censoring*.

Censoring

Consider an irreducible, finite-state discrete-time Markov chain whose states are partitioned into two sets E_1 and E_2. This induces a partitioning of its transition matrix T so that

$$T = \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix}.$$

The stationary distribution $\pi = (\pi_1, \pi_2)$ that satisfies $\pi T = \pi$ also satisfies

$$\pi_1 = \pi_1 \left[T_{11} + T_{12} \left(I - T_{22} \right)^{-1} T_{21} \right]$$

with

$$\pi_2 = \pi_1 T_{12} \left(I - T_{22} \right)^{-1}.$$

Censoring

Note that

$$(I - T_{22})^{-1} = \sum_{k=0}^{\infty} T_{22}^k$$

and so

$$\pi_1 \left[T_{11} + T_{12} \left(I - T_{22}\right)^{-1} T_{21}\right] = \pi_1 \left[T_{11} + T_{12} \left[\sum_{k=0}^{\infty} T_{22}^k\right] T_{21}\right]$$

and so we can interpret π_1 as the stationary distribution of the discrete-time Markov chain, censored so that it is observed only when it is in E_1.

Similar comments can be made in the case where E_2 is infinite as long as $\sum_{k=0}^{\infty} T_{22}^k$ converges elementwise, which is the case when it leaves E_2 with probability one.

Censoring

In fact, we can say more:

It is not just the case that π_1 is the stationary distribution of the discrete-time Markov chain, censored so that it is observed only when it is in E_1, but

$$\left[T_{11} + T_{12} \left[\sum_{k=0}^{\infty} T_{22}^k \right] T_{21} \right]$$

is the transition matrix of this chain. This is true even if the matrix is substochastic, in which case there is a positive probability that it may leave E_1 and not return.

Censoring

We can also observe that the (i, j)th entry of

$$[I - T_{22}]^{-1} T_{21} = \left[\sum_{k=0}^{\infty} T_{22}^{k}\right] T_{21}$$

is the probability that the Markov chain first enters E_1 in state j given that it started in state i of E_2.

Another Linear Algorithm

Above, I claimed that it was hard to give a physical interpretation of Neuts' original algorithm.

We can, however, easily give a physical interpretation for a related algorithm due to Latouche.

Write the basic equation for G in a different way:

$$\begin{aligned}(I - A_1 - A_0 G)G &= A_2 \\ \Rightarrow G &= (I - A_1 - A_0 G)^{-1} A_2,\end{aligned}$$

and use the iteration

$$G_{n+1} = (I - A_1 - A_0 G_n)^{-1} A_2,$$

with $G_0 = 0$.

Another Linear Algorithm

The matrix $G_1 = (I - A_1)^{-1} A_2 = \left[\sum_{k=0}^{\infty} A_1\right] A_2$, whose (i, j)th entry is the probability that the chain hits level $k-1$ in finite time, and does so in phase j, given that it starts in phase i of level k *and given that it never reaches level* $k+1$.

We can use induction to show that the (i, j)th entry of G_n is the probability that the chain hits level $k-1$ in finite time, and does so in phase j, given that it starts in phase i of level k *and given that it never reaches level* $k+n$.

Thus, the successive iterates of this algorithm have the same physical interpretation as that of the matrix G, but with a linearly increasing taboo level.

A Quadratic Algorithm

In (1993), Latouche and Ramaswami proposed the *logarithmic-reduction algorithm*. This works by evaluating the closed-form expression

$$G = \sum_{\ell=0}^{\infty} \left[\prod_{i=0}^{\ell-1} U^i \right] D^\ell.$$

A Quadratic Algorithm

where the matrices U^ℓ and D^ℓ satisfy the recursion

$$U^{\ell+1} = \left[I - U^\ell D^\ell - D^\ell U^\ell\right]^{-1} \left[U^\ell\right]^2$$

and

$$D^{\ell+1} = \left[I - U^\ell D^\ell - D^\ell U^\ell\right]^{-1} \left[D^\ell\right]^2,$$

with $U^0 = (I - A_1)^{-1} A_0$ and $D^0 = (I - A_1)^{-1} A_2$.

A Quadratic Algorithm

If we let

$$\hat{G}_n = \sum_{\ell=0}^{n} \left[\prod_{i=0}^{\ell-1} U^i \right] D^\ell.$$

then, except when the QBD is null-recurrent, $\hat{G}_n$ converges to G *quadratically*. That is, there exists a constant $\gamma \in (0,1)$ such that

$$\limsup_{n\to\infty} ||\hat{G}_n - G||^{1/2^n} = \gamma.$$

A Quadratic Algorithm

This algorithm also has a neat physical interpretation in terms of taboo probabilities.

The (i, j)th entry of the matrix $\hat{G}_n$ is the probability that the QBD will first enter level $k-1$ in phase j and does not visit any level higher than $k + 2^{n+1} - 2$ in between, given that it starts in phase i of level k.

Notice that the taboo level increases exponentially fast in terms of the number of iterates, which is consistent with the quadratic convergence of the algorithm.

A Quadratic Algorithm

The matrices $U^0 = (I - A_1)^{-1} A_0$ and $D^0 = (I - A_1)^{-1} A_2$ are, respectively the transition matrices of the discrete-time QBD derived from the original discrete-time QBD by observing it at the time points at which it changes level.

We can use induction to see that

$$U^{\ell+1} = \left[I - U^\ell D^\ell - D^\ell U^\ell\right]^{-1} \left[U^\ell\right]^2$$

and

$$D^{\ell+1} = \left[I - U^\ell D^\ell - D^\ell U^\ell\right]^{-1} \left[D^\ell\right]^2,$$

are the transition matrices of the discrete-time QBD derived from the original discrete-time QBD by observing it at the time points at which it hits levels of the form $k + m \times 2^{\ell+1}$.

A Quadratic Algorithm

So the summand on the right hand side of

$$G = \sum_{\ell=0}^{\infty} \left[\prod_{i=0}^{\ell-1} U^i \right] D^\ell.$$

takes into account sample paths that hit levels $k+1, k+3, \ldots, k+2^\ell - 1$, all with level $k-1$ taboo, and then hits level $k-1$ with level $k+2^{\ell+1}-1$ taboo.

Further Work

The numerical analysis community has become interested in these problems.

In 1995, Bini and Meini adapted the *Cyclic Reduction Algorithm* to the calculation of G for processes of M/G/1 type. This algorithm uses a similar censoring idea to the logarithmic reduction algorithm, but organises the calculations slightly differently.

A number of speed-up features, such as transforming the matrices to move eigenvalues away from the unit circle and using Fast Fourier Transforms are now included in implementations. Benny Van Houdt maintains a web-site with state-of-the-art Matlab code.

Newton's Method

In deriving Latouche's linearly-convergent algorithm, we used the fact that

$$A_2 + A_1 G + A_0 G^2 = G$$

is equivalent to

$$G = (I - A_1 - A_0 G)^{-1} A_2.$$

One thing that we can do is apply Newton's method to the solution of this equation. We would expect this to lead to a quadratically-convergent algorithm

Newton's Method

We obtain the sequence

$$G_{\mathrm{N}}^{(n+1)} - U^{(n)} A_0 G_{\mathrm{N}}^{(n+1)} U^{(n)} A_2 = U^{(n)} A_2 - U^{(n)} A_0 G_{\mathrm{N}}^{(n)} U^{(n)} A_2 \quad (\dagger)$$

where

$$U^{(n)} = (I - A_1 - A_0 G_{\mathrm{N}}^{(n)})^{-1}.$$

and

$$G_{\mathrm{N}}^{(0)} = 0.$$

The difficult part of implementing this is solving the Stein equation (†) above for $G_{\mathrm{N}}^{(n+1)}$.

Newton's Method

In 1994, it was shown by Latouche that, for any initial matrix $G_{\mathrm{N}}^{(0)}$ with $0 \leq G_{\mathrm{N}}^{(0)} \leq G$, the sequence $G_{\mathrm{N}}^{(n)}$ converges monotonically and quadratically to G.

By transforming (†) into a standard linear system by concatenating the columns of $G_{\mathrm{N}}^{(n+1)}$ and writing the coefficient matrix as a direct sum involving $U^{(n)}A_0$ and $U^{(n)}A_2$, Latouche provided an algorithm for evaluating the sequence of matrices $\{G_{\mathrm{N}}^{(n)}\}$.

Using this transformation, he showed that each iteration of the algorithm has a complexity of order $O(m^6)$.

Newton's Method

Latouche tested Newton's algorithm against the linearly-convergent algorithm presented above and found that, while Newton's algorithm required up to an order of magnitude fewer iterations, it could take up to an order of magnitude longer in terms of computer time to calculate G to within a given tolerance.

Since then, it would be fair to say that the conventional wisdom in the matrix-analytic community is that the complexity of each iteration of Newton's method makes it uncompetitive with other algorithms. This attitude was only reinforced by the later discovery of the quadratically-convergent logarithmic-reduction algorithm.

Newton's Method

However, in 1992, Gardiner, Laub, Amato and Moler had provided a $O(m^3)$ algorithm for solving the Stein equation (†). This motivated us to revisit the question of how useful Newton's method is in this context.

We were also interested in the question of whether we can give a physical interpretation for Newton's method, in a similar vein to the physical interpretations discussed above for the linear and logarithmic reduction algorithms.

Numerical experience

- For a QBD with $m = 20$, Newton's method took 13 iterations and .17 seconds of CPU time. the logarithmic-reduction algorithm took 23 iterations but only .13 seconds of CPU time.
- For an example deliberately constructed to "favour" the logarithmic reduction algorithm with $m = 6$, Newton's method took 11 iterations and .11 seconds of CPU time. the logarithmic-reduction algorithm took 11 iterations but only .07 seconds of CPU time.
- For another example with $m = 6$, Newton's method took 5 iterations and .05 seconds of CPU time. the logarithmic-reduction algorithm took 5 iterations but only .03 seconds of CPU time.

Physical Interpretation

As with the methods discussed above, the iterates $G_{\mathrm{N}}^{(n)}$ in Newton's Method contain the probabilities of certain sets of sample paths that start in level k and end in level $k-1$. In the physical description that we gave above for the linear and quadratically-convergent algorithms, these sets were defined in terms of taboo levels.

In order to understand Newton's method, we need to look at the sample paths in a different way.

Physical Interpretation

Denote the set of sample paths taken into account in $G_{\mathrm{N}}^{(n)}$ by $\Psi^{(n)}$.

The matrices $U^{(n)}A_0$ and $U^{(n)}A_2$ contain the probabilities of sets of sample paths that start in level k and end in levels $k+1$ and $k-1$ respectively. Denote these sets of sample paths by $\Phi_0^{(n)}$ and $\Phi_2^{(n)}$ *shifted to level* k.

Physical Interpretation

We have $U^{(0)} = (I - A_1)^{-1}$ and

$$\begin{aligned} G_{\mathrm{N}}^{(1)} &= U^{(0)} A_2 + U^{(0)} A_0 G_{\mathrm{N}}^{(1)} U^{(0)} A_2 \\ &= \sum_{\ell=1}^{\infty} \left((I - A_1)^{-1} A_0 \right)^{\ell-1} \left((I - A_1)^{-1} A_2 \right)^{\ell} \end{aligned}$$

where the second equation follows by repeatedly inserting the left hand side into the right hand side.

Physical Interpretation

So $\Psi^{(1)}$ accounts for sample paths that start in level one, increase to some level ℓ, possibly remaining in any level along the way but never dropping back, and then decrease to level $k-1$, again possibly remaining in any level but never increasing.

Thus, the sample paths in $\Psi^{(1)}$ are those that have a “single peak", no matter how high.

Sample Paths in $\Psi^{(1)}$

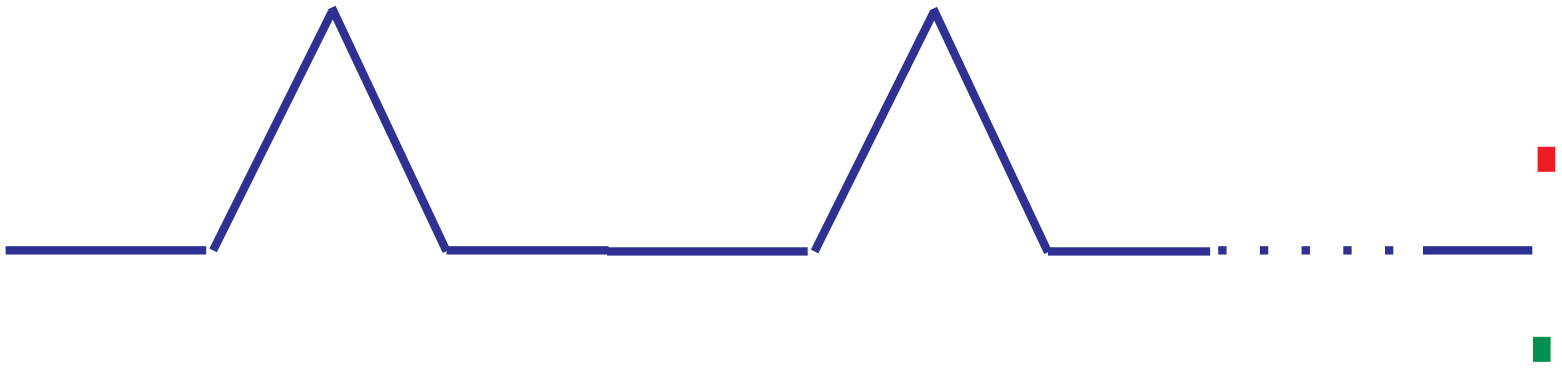
Physical Interpretation

The sets $\Phi_0^{(1)}$ and $\Phi_2^{(1)}$ contain sample paths taken into account by the matrices $U^{(1)}A_0$ and $U^{(1)}A_2$. We have

$$U^{(1)} = (I - A_1 - A_0 G_{\mathrm{N}}^{(1)})^{-1},$$

so $\Phi_0^{(1)}$ and $\Phi_2^{(1)}$ consist of sample paths that have any number of transitions between states at level k or “single peak" excursions from level k back to itself, followed respectively by a single transition to level $k+1$ and level $k-1$.

Sample Paths in $\Phi_0^{(1)}$ and $\Phi_2^{(1)}$



Physical Interpretation

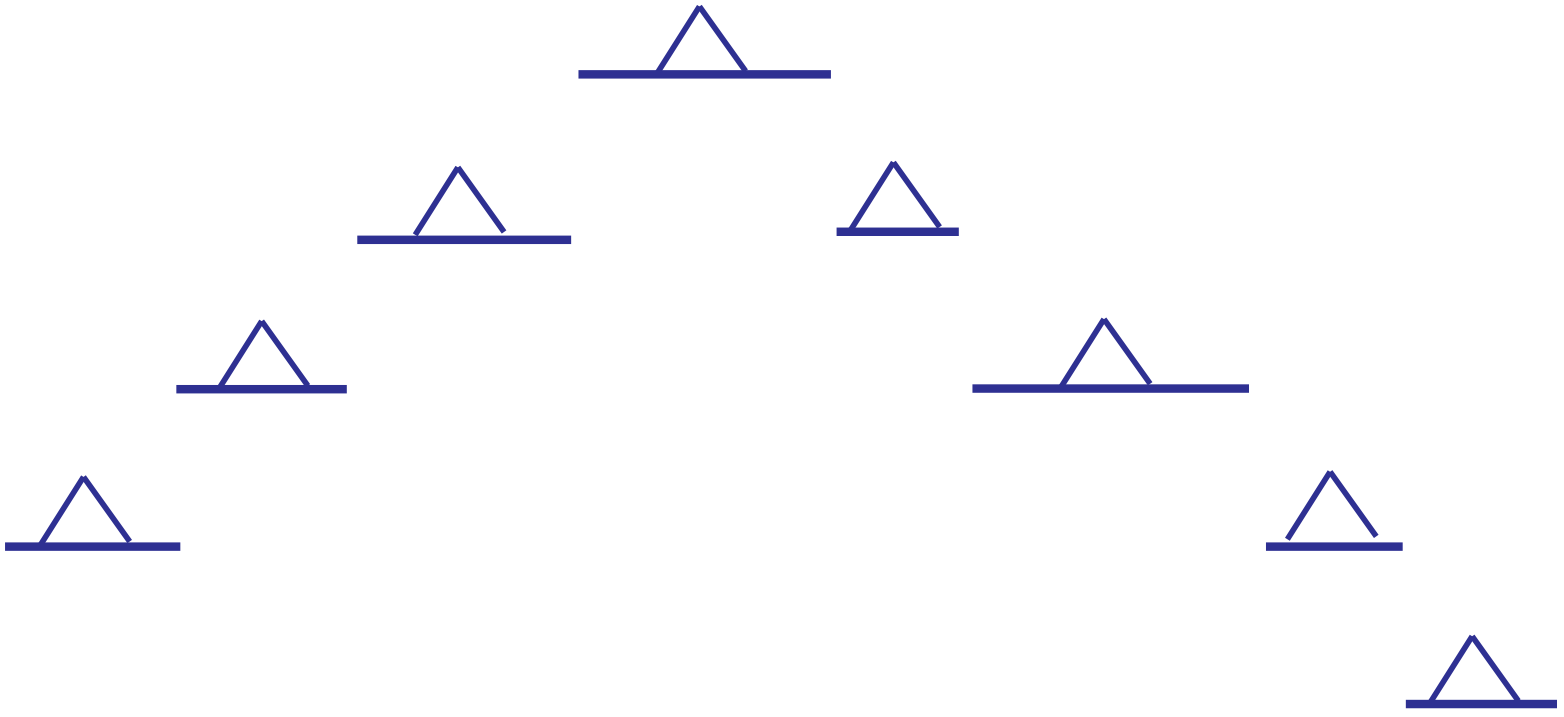
$$G_{\mathrm{N}}^{(2)} - G_{\mathrm{N}}^{(1)}$$

$$= \sum_{\ell=1}^{\infty} \left(U^{(1)} A_0\right)^{\ell-1} \left(U^{(1)} A_2\right)^{\ell} - \sum_{\ell=1}^{\infty} \left(U^{(1)} A_0\right)^{\ell} G_{\mathrm{N}}^{(1)} \left(U^{(1)} A_2\right)^{\ell}.$$

So, sample paths in $\Psi^{(2)}$ but not in $\Psi^{(1)}$ are made up of a succession of sample paths that either stay at the same level or have “single peak" excursions upward, each shifted one level higher, up to some some level ℓ, whereupon a succession of sample paths that either stay at the same level or have “single peak" excursions upward, each shifted one level lower occurs, until the process drops to level $k-1$.

The subtraction of the second term on the right hand side of equation ensures that paths are not counted multiple times.

Sample Paths in $\Psi^{(2)}$ but not in $\Psi^{(1)}$



Physical Interpretation

For general n,

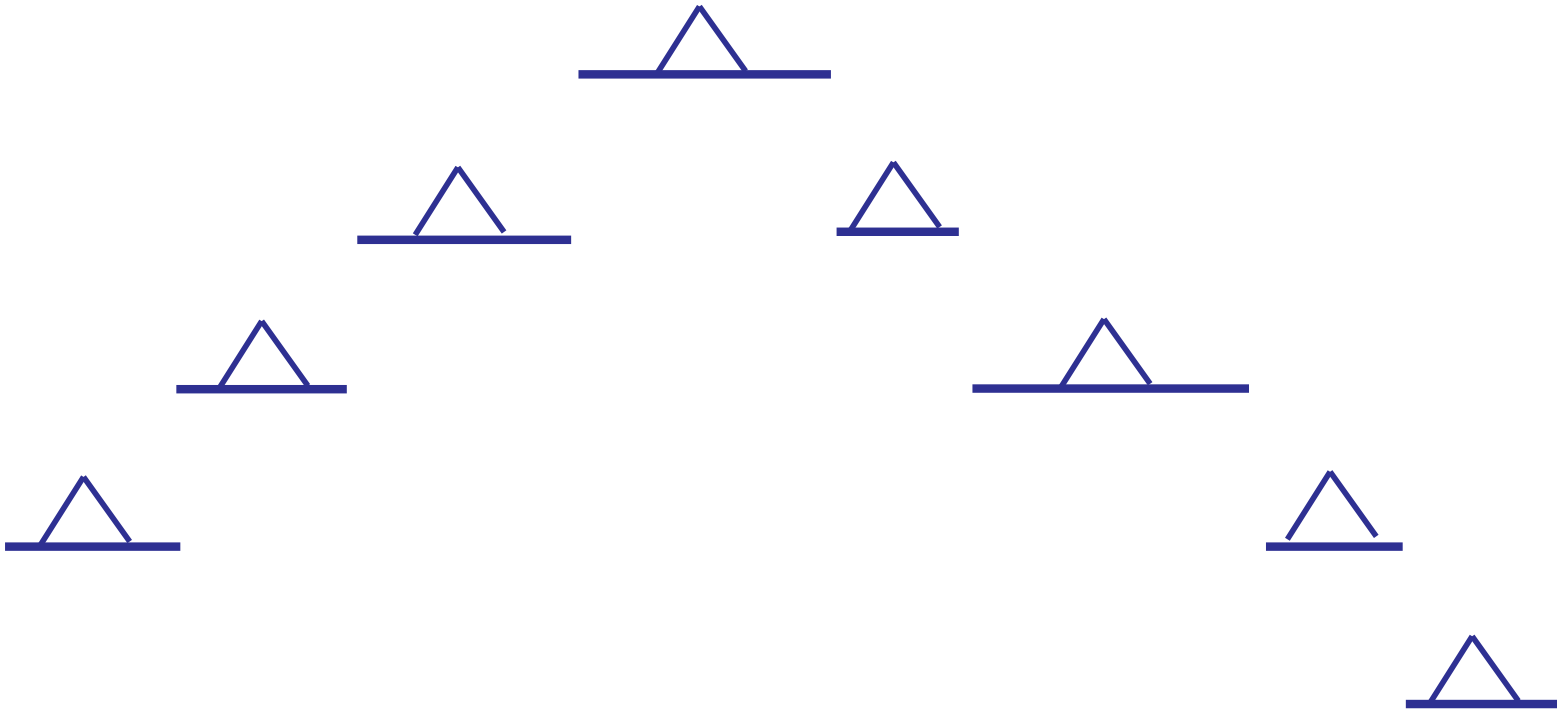
$$G_{\mathrm{N}}^{(n+1)} - G_{\mathrm{N}}^{(n)}$$

$$= \sum_{\ell=1}^{\infty} \left(U^{(n)} A_0\right)^{\ell-1} \left(U^{(n)} A_2\right)^{\ell} - \sum_{\ell=1}^{\infty} \left(U^{(n)} A_0\right)^{\ell} G_{\mathrm{N}}^{(n)} \left(U^{(n)} A_2\right)^{\ell}.$$

So, sample paths in $\Psi^{(n+1)}$ but not in $\Psi^{(n)}$ are made up of a succession of sample paths in $\Phi_0^{(n)}$, each shifted one level higher, up to some some level ℓ, whereupon a succession of sample paths in $\Phi_2^{(n)}$ occurs, each shifted one level lower, until the process drops to level $k-1$.

The subtraction of the second term on the right hand side of equation ensures that paths are not counted multiple times.

Sample Paths in $\Psi^{(n+1)}$ but not in $\Psi^{(n)}$



Physical Interpretation

So, instead of progressively including more sample paths by relaxing a taboo level, Newton's method progressively includes more and more complicated sample paths.

This happens in a 'fractal' way: basic units of paths at one iteration are the sets of paths that were accounted for in the previous iteration.

Very complicated paths are taken into account within a few iterations, which intuitively supports the fact that Newton's Method converges quadratically.

Stochastic Fluid Models

- Kosten (late 1970s).
- Anick, Mitra and Sondhi (1982).
- Rogers (1994).
- Asmussen (1995).
- Ramaswami (1999).
- da Silva Soares and Latouche (2002).
- Bean, O'Reilly and Taylor (2004).

In these models, the rate c_i at which the level of a fluid increases, or decreases, is governed by the state i of an underlying continuous-time Markov chain. The parameters c_i can be positive, negative or zero.

Our Fluid Model

Again is a two-dimensional process, $\{(X(t), \varphi(t)) : t \in \mathcal{R}^+\}$, where

- $X(t) \in \mathcal{R}^+$ is the *level*,
- $\varphi(t) \in \mathcal{S} = \mathcal{S}_1 \cup \mathcal{S}_2$ is the *phase*,
- the phase process $\{\varphi(t) : t \in \mathcal{R}^+\}$ is an irreducible, finite, continuous-time Markov chain with generator $\mathcal{T}$, and
- the net rate of input c_i to the infinite fluid buffer, when $\varphi(t)$ is in state i, is equal to 1 for $i \in \mathcal{S}_1$ and -1 for $i \in \mathcal{S}_2$.

The assumption of unit increase and decrease rates is without loss of generality for the performance measures that I shall be discussing here.

Analysing Fluid Flow Models

We would like to know

- whether the process is stable,
- when it is stable, the stationary distribution,
- statistics of sojourn times in various sets,
- the probabilities of return to the initial level in each of the phases, and
- the distribution of the hitting time on the initial level conditional on the initial phase.

Analysing Fluid Flow Models

Various methods have been used:

1. Spectral methods.
2. Wiener-Hopf factorisation.
3. Discretising the fluid variable and working with the corresponding discrete-state environment, which is a quasi-birth-and-death process.
4. Using the physics of the process directly within the fluid flow environment.

Analysing Fluid Flow Models

The return probabilities Ψ_{ij} that the chain hits level 0 in finite time, and does so in phase $j \in \mathcal{S}_2$, given that it starts in phase $i \in \mathcal{S}_i$ at level 0 are fundamental in the sense that many other performance measures can be calculated from them. Store them in a matrix Ψ.

Then, for example, Ramaswami showed that, for $x > 0$, the stationary density of a positive recurrent stochastic fluid model is

$$[\pi_1(x), \pi_2(x)] = \alpha \left[e^{Kx}, e^{Kx}\Psi\right]$$

where

$$K = [T_{11} + \Psi T_{21}] .$$

Return Times to the Initial Level

Let $\theta(x) = \inf\{t > 0 : X(t) = x\}$ be the first passage time to level x in the process $(X(t), \varphi(t))$.

For all $i \in \mathcal{S}_1$ and $j \in \mathcal{S}_2$, let

$$\Psi_{ij} = P[\theta(z) < \infty \, , \, \phi(\theta(z)) = j \mid X(0) = z, \phi(0) = i],$$

and let $\Psi = [\Psi_{ij}]$.

In physical terms, Ψ_{ij} is the probability that, starting from level z in phase $i \in \mathcal{S}_1$, the process $(X(t), \varphi(t))$ first returns to level z in finite time and does so in phase $j \in \mathcal{S}_2$, while avoiding levels below z.

Return Times to the Initial Level

The matrix of return probabilities Ψ is the matrix that we want to solve for. Partitioning T according to $\mathcal{S} = \mathcal{S}_1 \cup \mathcal{S}_2$ so that

$$\mathcal{T} = \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix},$$

it can be shown that Ψ is the minimal nonnegative solution of the nonsymmetric algebraic Riccati equation,

$$T_{11}\Psi + \Psi T_{22} + \Psi T_{21}\Psi + T_{12} = \mathbf{0}.$$

Algorithms

By the simple action of writing the above equation in a different form and then introducing an iteration, several algorithms for Ψ can be obtained. For example, we can write

$$T_{11}\Psi + \Psi T_{22} = -T_{12} - \Psi T_{21}\Psi.$$

and then devise an iteration by setting $\Psi_0 = \mathbf{0}$, and using

$$(1) \quad T_{11}\Psi_{n+1} + \Psi_{n+1}T_{22} = -T_{12} - \Psi_n T_{21}\Psi_n,$$

to calculate subsequent values of Ψ_n.

Algorithms

In a similar manner we can write

$$(2) \quad T_{11}\Psi_{n+1} + \Psi_{n+1}(T_{22} + T_{21}\Psi_n) = -T_{12},$$
$$(3) \quad (T_{11} + \Psi_n T_{21})\Psi_{n+1} + \Psi_{n+1}T_{22} = -T_{12},$$
$$(4) \quad (T_{11} + \Psi_n T_{21})\Psi_{n+1} + \Psi_{n+1}(T_{22} + T_{21}\Psi_n) = -T_{12} + \Psi_n T_{21}\Psi_n,$$

all with $\Psi_0 = \mathbf{0}$.

Algorithms

All of these are equations of type $AX + XB = C$. These can be solved using the Bartels-Stewart algorithm (1972), the Hessenberg-Schur method (Golub, Nash and van Loan (1979)) or by using the function 'lyap' in the Matlab Control System Toolbox.

The first and fourth of the above iterations were analysed by Guo (2001).

Our focus has been on providing a physical interpretation of these and other algorithms.

Physical Interpretation

In order to gain insight into the physical interpretations of the above algorithms for calculating Ψ, we need the following result (Bwidehatia and Rosenthal (1997)).

Whenever the spectra of the matrices A and B are contained in the open left half plane, the linear form

$$AX + XB = -C$$

is equivalent to the integral form

$$X = \int_0^\infty e^{Ay} C e^{By} dy.$$

Physical Interpretation

So

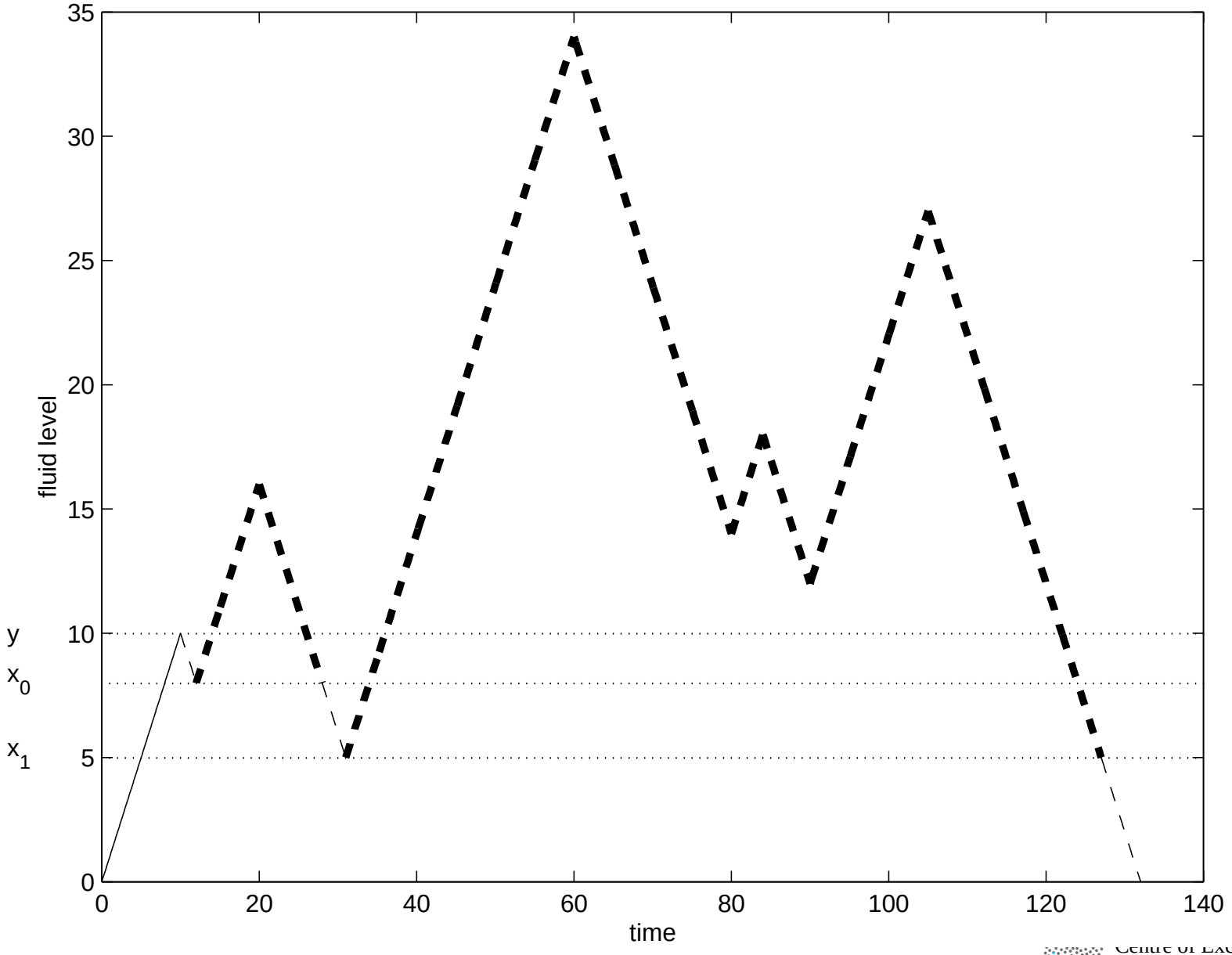
$$T_{11}\Psi + \Psi T_{22} = -T_{12} - \Psi T_{21}\Psi$$

is equivalent to

$$\Psi = \int_0^\infty e^{T_{11}y} T_{12} e^{T_{22}y} dy + \int_0^\infty e^{T_{11}y} \Psi T_{21} \Psi e^{T_{22}y} dy$$

and we can interpret y in the first integral as the level of transition from $\mathcal{S}_1$ to $\mathcal{S}_2$ in a 'single peak' path, and y in the second integral as the minimum level at which a transition from $\mathcal{S}_2$ to $\mathcal{S}_1$ occurs in any more complicated path.

A Fluid Sample Path



Physical Interpretation

As we saw with the block-structured Markov chains, at each iteration of the algorithms given above, the matrix Ψ_n records the total probability mass of certain sample paths that return to the initial level, but not others.

We can understand the physical interpretation of each algorithm by examining the set of sample paths Ω_n that are included at iteration n.

Physical Interpretation

The iteration

$$T_{11}\Psi_{n+1} + \Psi_{n+1}T_{22} = -T_{12} - \Psi_n T_{21}\Psi_n,$$

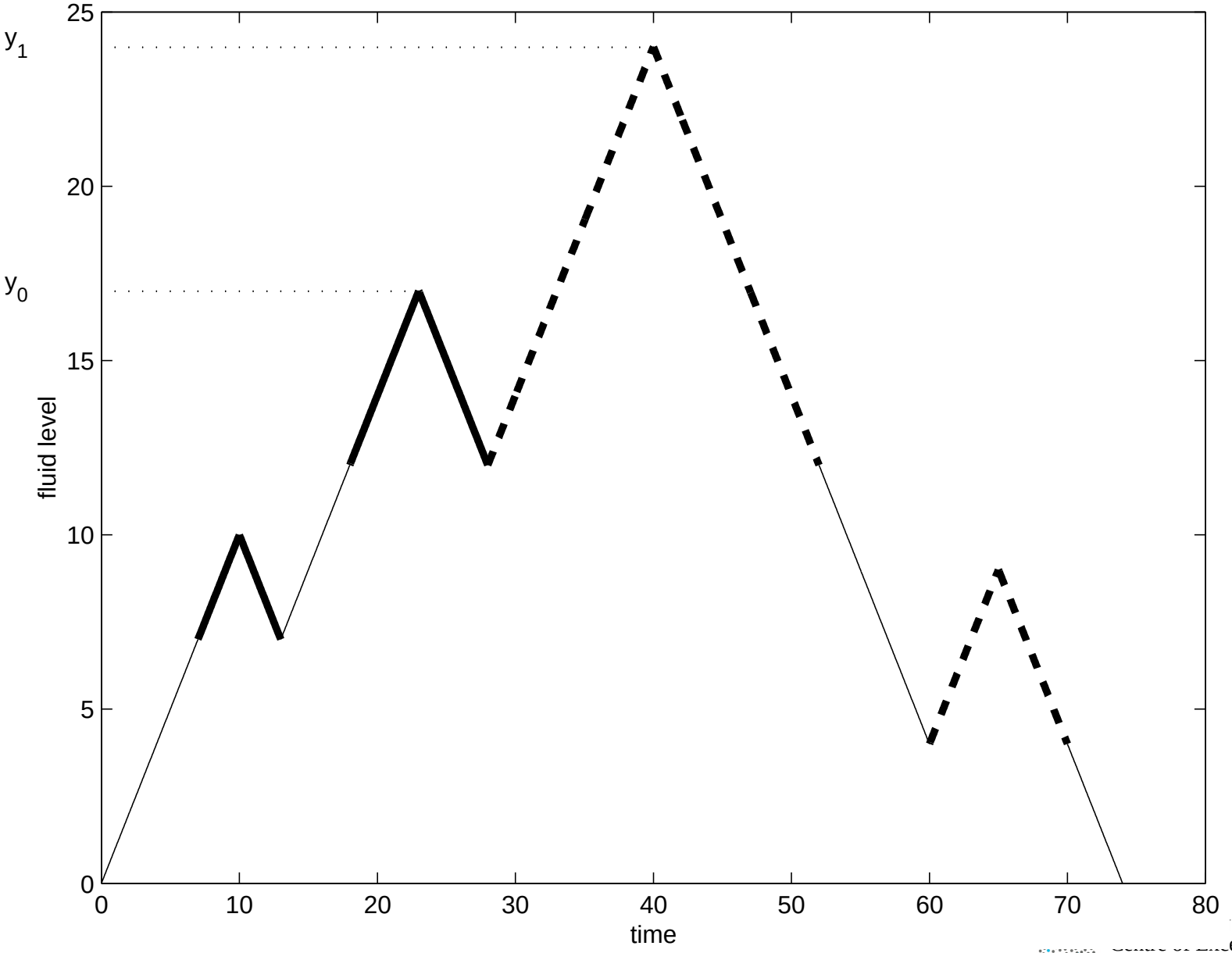
is equivalent to

$$\Psi_{n+1} = \int_0^\infty e^{T_{11}y}T_{12}e^{T_{22}y}dy + \int_0^\infty e^{T_{11}y}\Psi_n T_{21}\Psi_n e^{T_{22}y}dy.$$

We see that Ω_{n+1} contains

- the set of single peak paths, and
- the set of paths in which excursions above the minimum level at which a transition from $\mathcal{S}_2$ to $\mathcal{S}_1$ occurs must be in Ω_n.

A Path in Ω_4



Physical Interpretation

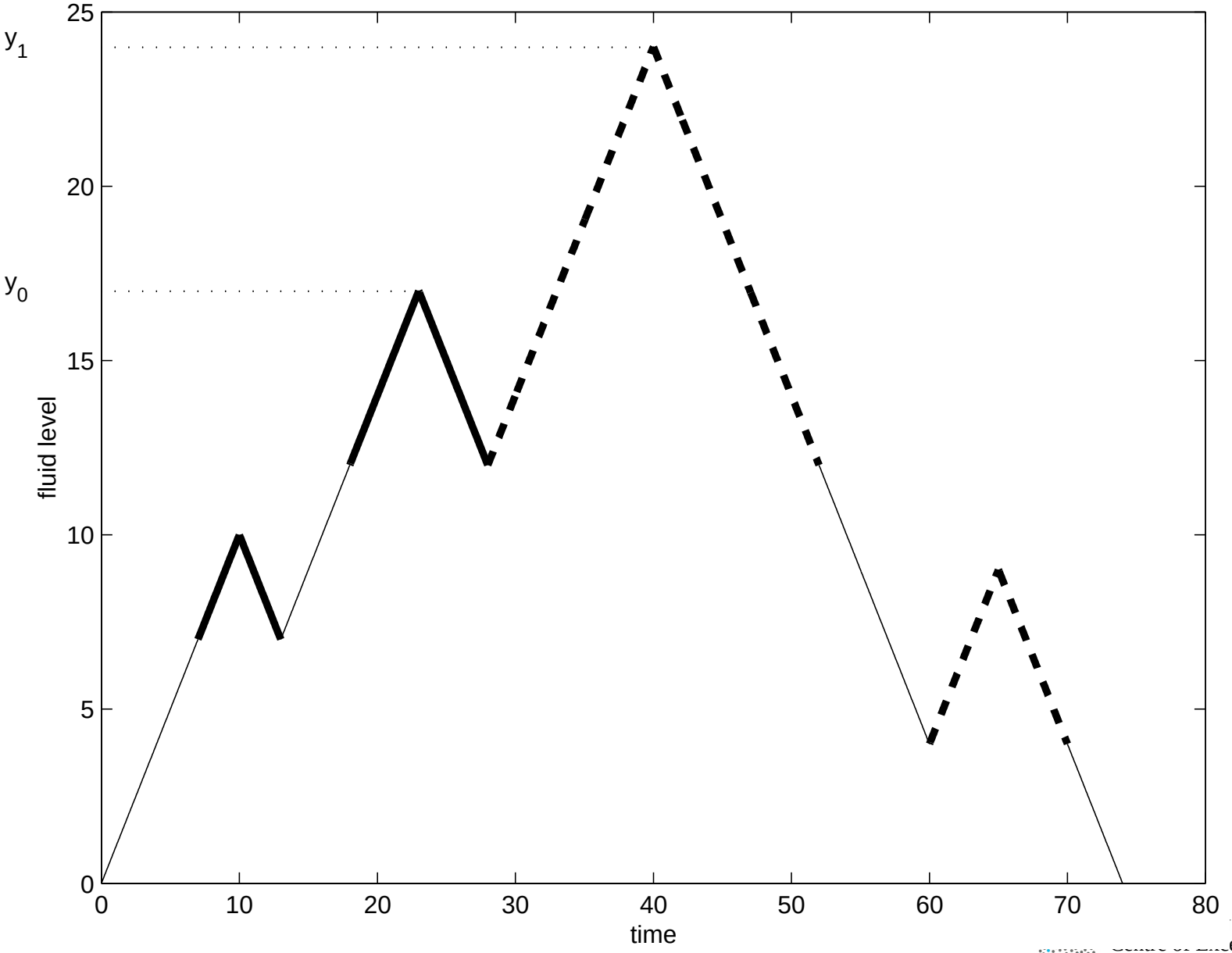
In the other algorithms, the ‘upwards’ and/or ‘downwards’ processes can be more complex. The most complicated algorithm is Algorithm 4, which is actually an implementation of Newton’s Method:

$$\Psi_{n+1} = \int_0^\infty e^{(T_{11}+\Psi_n T_{21})y}(T_{12} - \Psi_n T_{21}\Psi_n)e^{(T_{22}+T_{21}\Psi_n)y}dy$$

in which both the ‘upward’ and ‘downward’ paths can involve excursions from a level back to itself.

This iteration is quadratically convergent, unlike the other algorithms, which are linearly convergent.

A Path in Newton's Ω_2



Physical Interpretation

We can also give physical interpretations of Asmussen's algorithm and of Ramswami's transformation to a QBD model.

Once the latter transformation is made, any standard algorithm for calculating the matrix G in a QBD can be used. In particular the quadratically-convergent logarithmic reduction (Latouche and Ramaswami (1993)) and cyclic reduction (Bini and Meini (1996)) algorithms can be used.

In addition to Newton's method, these 'QBD Methods' are the only quadratically-convergent algorithms for stochastic fluid models that we know of.

Back to Neuts' Algorithm

I mentioned earlier that it has been difficult to give a physical interpretation for Neuts' original procedure

$$G_{n+1} = (-A_1)^{-1} \left[A_2 + A_0 G_n^2 \right]$$

with $G_0 = 0$.

Let Ξ_n be the set of sample paths starting in level k and finishing in level $k-1$ that are taken into account G_n. Then it was known that no sample path in Ξ_n reaches level $k+n+1$. Also, sample paths in Ξ_n have at most 2^n transitions at which the level increases.

However these two restrictions still don't fully define Ξ_n.

Back to Neuts' Algorithm

Bean, Kontoleon and Taylor (2008) were able to establish a correspondence between the sample paths accounted for in the Neuts algorithm and the tree topologies that are generated by the nth iteration of the *Depth Algorithm*.

The Depth algorithm is an algorithm that we proposed for the calculation of the extinction probabilities in a Markovian Binary Tree, which is a special case of a continuous-time multitype branching process.

It is not the best available algorithm for this purpose (see the work of Sophie Hautphenne for this), but it is interesting to see the connection.