The power-series algorithm. A numerical approach to Markov processes

van den Hout, W.B.

Publication date:
1996

CentER
for
Economic Research

The Power-Series Algorithm

A Numerical Approach to Markov Processes

W.B. van den Hout

Tilburg University

The Power-Series Algorithm

A Numerical Approach to Markov Processes

The Power-Series Algorithm

A Numerical Approach to Markov Processes

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de Katholieke Universiteit Brabant, op gezag van de rector magnificus, prof. dr. L.F.W. de Klerk, in het openbaar te verdedigen ten overstaan van een door het college van dekanen aangewezen commissie in de aula van de Universiteit op woensdag 27 maart 1996 om 16.15 uur door

WILLEM BERNHARD VAN DEN HOUT

geboren op 23 mei 1968
te Helmond

PROMOTOR: Prof.dr.ir. O.J. Boxma
COPROMOTOR: Dr. J.P.C. Blanc

Thou maintainest my lot.
The lines are fallen unto me in pleasant places;
yea, I have a goodly heritage.

PSALMS 16:5,6

Gij zelf bestendigt wat het lot mij toewees.
De meetsnoeren vielen mij in liefelijke dreven,
Ja, mijn erfdeel bekoort mij.

PSALM 16:5,6

Woord vooraf

Graag wil ik bij het afronden van dit proefschrift allen bedanken die mij bij de totstandkoming ervan hebben geholpen. Dit zijn er velen. Een aantal zal ik hier in het bijzonder noemen.

Allereerst wil ik mijn begeleider en copromotor Hans Blanc hartelijk bedanken voor de deskundige en zorgvuldige dagelijkse begeleiding en de aangename samenwerking. De vrijheid die ik hierin heb gekregen waardeer ik zeer. Mijn promotor Onno Boxma bedank ik van harte voor zijn uitstekende begeleiding, voor zijn persoonlijke betrokkenheid en de plezierige gesprekken. Ook de andere leden van de promotie-commissie ben ik erkentelijk: Prof.dr.ir. J.W. Cohen, Prof.dr. F.A. van der Duyn Schouten, Prof.dr. I. Mitrani en Prof.dr. M.H.C. Paardekoper. I would like to thank Isi Mitrani especially for giving me the opportunity to visit the University of Newcastle and for making my stay in England a pleasant one.

Met veel plezier heb ik gewerkt aan de Katholieke Universiteit Brabant. In het bijzonder wil ik daarvoor bedanken de secretariële ondersteuning, collega-AIO Rob van der Mei en kamergenoot Jean-Jacques Herings. Ook de Stichting Mathematisch Centrum stond mij bij mijn onderzoek ter zijde, met financiële hulp van de Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO). Ik ben hen hiervoor zeer erkentelijk. Met name dhr. Wim Aspers wil ik bedanken voor de professionele en prettige samenwerking.

Bovenal wil ik mijn familie en vrienden bedanken. Wat zij me de afgelopen vier jaar hebben gegeven en geleerd is vele malen waardevoller dan wat ik in een proefschrift kan neerschrijven.

Wilbert van den Hout
februari 1996, Leiden

Contents

Chapter 1

Introduction

Although the theory of Markov processes is well-developed, the numerical analysis of large-scale Markov processes remains a difficult problem. The Power-Series Algorithm (PSA) is a method to compute performance measures for such processes. Over the past ten years, numerous papers have been written with successful applications of the PSA to various Markov processes, in particular to Markov processes arising in queueing theory.

The objective of this thesis is twofold. The first objective is to improve the theoretical basis of the PSA. Analyticity of the performance measures in light traffic will be proved for a wide class of models and remedies will be found for previous imperfections of the algorithm. The second objective is to extend the applicability of the PSA. The class of models to which the PSA can be applied will be enlarged and the PSA will be adapted for transient analysis. The applications included in this thesis are not meant as a thorough numerical analysis of new models. They are illustrations of the behaviour of the method in difficult situations and of its wide applicability and flexibility.

1.1 Steady-state analysis of Markov processes

The main part of this thesis is devoted to the problem of finding the steady-state distribution of an ergodic continuous-time Markov process on a countable state space. This steady-state distribution can be found by solving the so-called balance equations. However, in practice, these linear equations can only be solved either if the state space of the problem is small enough or if the problem has some special structure.

If the state space is small, the balance equations can be solved using direct methods, for example LU decomposition by Gaussian elimination [59] or the GTH algorithm [64]. The increased speed and memory capacity of modern computers make it possible to apply these methods to ever larger problems. However, the computation time of direct methods grows fast in the size of the problem. Therefore, they are less efficient for larger problems. New methods have been developed to overcome this problem, like the iterative Gauss-Seidel and successive overrelaxation or the projection methods [132]. These methods can often take advantage of the fact that the balance equations are sparse [14]. If the problem is still too large, the problem needs to be reduced to a manageable size. Infinite state spaces must necessarily be truncated. Sometimes, bounds on the thus introduced error can be obtained [51,78]. Much research has been devoted to nearly-completely decomposability [44] and aggregation/disaggregation methods. Recent developments on many of these topics are described in the proceedings and monograph by Stewart [130,131,132].

In queueing theory, problems often have some special structure. Therefore, it may be possible to analyze the Markov process by other methods than solving the balance equations. Then the problem of large state spaces could be avoided. Sometimes infinite state spaces are even an advantage, because they result in simple mathematical expressions. For example, many one-dimensional Markov processes with a finite supplementary space can be analyzed with the matrix-geometric method [102,113] and spectral expansion [53,110]. Several two-dimensional models can be analyzed using methods based on generating functions, like the boundary-value method [42] and the uniformization technique [54]. Multidimensional models can only be analyzed in some cases, such as when the joint distribution has a product form [85,50]. A special structure of the transition probabilities allows for application of the compensation approach [5,6,77]. Discrete Fourier transforms have been successfully applied to some polling models [96,97]. Also, the recent developments in the numerical inversion of generating functions [2] allow for the numerical solution of many one-dimensional models and some multidimensional models like resource-sharing and special polling models.

Another approach to study Markov processes is to consider limit behaviour. For many Markov processes, a notion of heavy and light traffic can be defined. This can, for example, be based on arrival rates or the number of queues. Interpolation methods can be used to combine heavy and light-traffic results. In heavy traffic and with an appropriate normalization, many processes behave like a reflected Brownian motion or a related process on a continuous state space [46,70]. Especially light traffic approaches have received considerable attention lately. Some of these approaches will be discussed in section 1.1.2, as they are closely related to the PSA.

1.1.1 The Power-Series Algorithm

The short and incomplete overview above illustrates that there is a great diversity of methods to analyze Markov processes, each with its own merits and drawbacks. In this wide field of methods, the PSA can best be classified as a light-traffic method. It is not restricted to any subclass of Markov processes, but it aims to be an efficient method for Markov processes with a multidimensional state space.

The main idea of the PSA is to consider the steady-state distribution of the Markov process as a function of a system parameter. Consider a queueing system with several queues and Poisson arrival processes at all queues. Multiply all arrival rates by ρ and normalize the arrival rates in such a way that the model is stable for $\rho \in [0, 1)$. For small values of ρ the system is in light traffic and will be relatively empty, for large values of ρ it is in heavy traffic and busy. Clearly, the steady-state probabilities are functions of ρ. Assume that these functions are analytic in light traffic, so that they are determined by the coefficients of the power-series expansion in ρ, around $\rho = 0$. The balance equations also depend on ρ. Substituting the power-series expansions of the steady-state probabilities in the balance and normalization equations renders equalities of analytic functions. Analytic functions are only equal if all coefficients of the power-series expansions are identical. This renders equations that allow for the recursive calculation of these coefficients.

This idea was first used by Beneš in 1965 (see chapter 8 in [16]). He considers a physical communication system consisting of a set of terminals, a control unit which processes the information needed to set up calls, and a connecting network through which calls are switched between terminals. The principal problem treated is the calculation of the grade of service, as measured by the probability of blocking. The considered processes have a finite state space. From this, Beneš proves that the steady-state probabilities are rational functions of the arrival rate and therefore analytic. He provides recursive calculation schemes to calculate the power-series expansions, not only in light traffic.

The idea was independently rediscovered by Keane, about twenty years later. In a paper by Hooghiemstra, Keane and Van de Ree [73], it is applied to a model in which a single exponential processor distributes its capacity over a number of Poisson arrival streams. Contrary to the model considered by Beneš, the state space is infinite. The total number of customers in the system behaves like a single $M/M/1$ queue. Because of this structure, analyticity of the steady-state probabilities can be proved in light traffic. In a later paper [15], the coefficients of the expansions were obtained explicitly for the symmetric two-queue model. Since then, the PSA has been developed further, mainly by Blanc. It was applied to general queueing models with a quasi birth-death structure, such as queues in parallel [20], for instance the shortest-queue model [19,24], and various polling models [27,23]. The introduction of the epsilon algorithm greatly improved the convergence properties [22]. In a paper on the $BMAP/PH/1$ queue [74], the PSA was

extended to queues with more general arrival processes and service time distributions. Because of the (finite) batch arrivals, the process is no longer a quasi birth-death process. For an overview of these applications, see [25]. In all these papers, the transition rates of the arrival process are multiplied by ρ and normalized such that the system is stable for $\rho \in [0,1)$. The parameter ρ can thus be interpreted as the load of the system. Unfortunately, this light-traffic approach is not generally applicable. For example, only networks can be analyzed that have finite batch arrivals and a feed-forward structure (see section 2.8.5). In this thesis a more general approach will be proposed to overcome these limitations.

Koole [87] shows that the PSA can be applied to general Markov processes with a single recurrent class. Each state x in the state space is assigned a level $\ell(x)$. The transition rate from state x to state y is then multiplied by $\rho^{\ell(y)-\ell(x)}$, for all states x, y such that $\ell(x) < \ell(y)$. Koole provides sufficient conditions under which the coefficients of the steady-state probabilities can be calculated recursively. For any Markov process with a single recurrent class, level functions exist such that these sufficient conditions are satisfied. Different level functions lead to different algorithms, but it is not clear how the level function can best be chosen to obtain efficient algorithms. The queueing applications above assign to each state the level equal to the total number of customers. Koole provides numerical results for a fork-join model and a bounded Petri net.

Instead of as a light-traffic method, the PSA can also be interpreted as a homotopy method: the transition rates of the original Markov process are transformed with a parameter γ, such that for $\gamma = 1$ the transformed process is the original process and the asymptotic process for γ in a neighbourhood of $\gamma = 0$ is easy to analyze. Then the information from the problem near $\gamma = 0$ can be used to solve the problem at $\gamma = 1$. This approach will be followed in this thesis. It has the disadvantage that the parameter γ may not have a physical interpretation, but it allows for a more general transformation and less restrictive models.

The major advantage of the PSA is its flexibility. Except for the assumption that the process is Markovian, very little special structure needs to be assumed. More complicated balance equations lead to a more complex algorithm, but rarely give rise to theoretical problems. Other advantages are that the main idea is simple and that no advanced mathematical procedures are needed for integration or inversion, for example. It *is* essential that sophisticated extrapolation methods are used to make the PSA applicable also for intermediate and quite heavy traffic, but these methods can be used routinely. A disadvantage of the PSA is that it is quite sensitive to extreme parameter values and that it suffers from the curse of dimensionality because it directly depends on the balance equations. Also it has been impossible so far to find useful error bounds. The credibility of the results is established by the convergence of the obtained power series and by checking known characteristics of the model, like the probability of an empty system, marginal distributions or conservation laws. For a more detailed discussion of

the characteristics of the PSA, see section 2.3.5.

1.1.2 Other light-traffic approaches

Reiman and Simon [118,119,125] consider Poisson-driven queueing networks and obtain expansions of performance measures from a sample path argument, restricting the total number of arrivals in the sample path. Because of the complexity of the approach only a few coefficients can be calculated, which they combine with heavy traffic limits. Analyticity of Poisson-driven stochastic systems was proved by Zazanis [139].

The MacLaurin series approach by Gong and Hu [60] for the *GI/G/1* queue and by Zhu and Li [140] for the Markov-modulated *G/G/1* queue is quite similar to the approach of the PSA. Starting from the Lindley equation instead of the balance equations, they directly obtain the expansions of the moments of the system time and the delay, without computing the queue-length distribution. Let $\mathcal{A}$ and $\mathcal{S}$ be generic interarrival and service times, and $\mathcal{T}$ and $\mathcal{W}$ the steady-state sojourn and waiting times of a customer. The distributions of these variables are related by

$$\mathcal{T} \stackrel{d}{=} \mathcal{W} + \mathcal{S} \stackrel{d}{=} \max\{\ 0,\ \mathcal{T} - \mathcal{A}\ \} + \mathcal{S}. \tag{1.1}$$

When $\mathcal{S} \doteq \theta\mathcal{X}$ with $\mathcal{X}$ independent of θ, then the moments of $\mathcal{T}$ and $\mathcal{W}$ are functions of θ. The power-series expansions of these functions can be obtained from (1.1). In [61], Padé approximation is used to improve the convergence of these power series. Analyticity of $GI/G/1$ queues in light traffic is shown by Hu [79], for analytic interarrival-time distributions. This approach allows for non-Markovian (but not general) interarrival and service times. The complexity of the algorithm is comparable to the complexity of the PSA.

Similar ideas are used by Blaszczyszyn, Frey and Schmidt [30] and by Baccelli and Schmidt [11]. The first paper analyzes Markov-modulated multi-server queues, and the second considers systems that have a so-called Poisson-driven (max,+)-linear structure [9]. Such systems can model non-Markovian stochastic Petri nets in the class of event graphs. Examples are queueing models like fork-join networks, tandem queues and synchronized queueing networks. But also manufacturing models, such as Kanban networks and Job-Shop systems. High-order moments can be calculated because of the use of higher-order moment measures and Palm distributions of general marked point processes. Similar light-traffic approaches to analyze risk processes can be found in [31,56].

There are also several papers on optimal control that obtain the optimal policy in light traffic. Here, light-traffic analysis is not used as a numerical method. To determine which policy is optimal, one only needs to obtain the first coefficient of the power-series expansion of the criterion function at which the policies differ. This coefficient can often be determined symbolically. The approach has been applied to the problem of repairman allocation [126], load-balancing problems [82,83,84,88,99] and the order of servers for

tandem queues [67]. In [29], the symbolic and numerical approach were combined to find the optimal Bernoulli service discipline of cyclic polling systems. A general framework is provided by Koole and Passchier [89].

1.2 Transient analysis of Markov processes

In chapter 3, the PSA will be adapted to analyze the transient distribution of Markov processes. In practice, systems are rarely in steady state. This can be either because *i*) the system is not ergodic, so it will never reach steady state, or because *ii*) the system started in a situation different from steady state and not enough time has passed to eliminate the influence of the initial condition, or because *iii*) one or more of the parameters of the system vary over time. The question whether a system is ergodic or not can often be answered with techniques that are less computational than the techniques for finding the steady-state or transient distribution [72,101,109]. But if a steady-state distribution has been calculated, the system must necessarily be ergodic. On this fact, the ergodicity condition in appendix A is based. If a system is ergodic, then it is interesting to know how fast the system will approach the steady-state situation. The relaxation time is a measure of the time needed to reach steady state [17,18,21,26,41]. If the relaxation time is small, then it will not be necessary to study the transient distribution.

Like for the steady-state analysis, there are many different methods to study the transient behaviour of Markov processes. If the process is homogeneous with generator Q, then the transient distribution $\pi(t)$ at time t is determined by the differential equations $\pi'(t) = \pi(t)Q$ with solution $\pi(t) = \pi(0)\exp(Qt)$. Most general-purpose methods are based on one of these characterizations. The main problem in calculating transient distributions is stiffness. This occurs when the Markov process has transition rates of different orders of magnitude, and it is especially troublesome when t is large.

There are many ways to calculate the matrix exponential $\exp(Qt)$, but none of them is completely satisfactory when the model is stiff or the state space is large [4,111]. A promising new method is based on Krylov subspaces [117,123]. Explicitly computing the exponential of a matrix is avoided by Jensen's method [81,68,62,63,47,48,36]. This method will be described in section 3.1.2. It does not handle stiffness very well, but this can be partly solved by using steady-state detection [40,104]. Some of the general methods to solve the ordinary differential equation (ODE) $\pi'(t) = \pi(t)Q$ are especially designed to handle stiffness [132]. Malhorta, Muppala and Trivedi [104] provide a comparison for TR-BDF2 (the trapezoidal rule with second order backward difference) and implicit Runge-Kutta methods. They conclude that these ODE solvers are only preferable when the model is extremely stiff.

The transient distribution of a non-homogeneous process with generator $Q(t)$ is deter-

mined by the differential equation $\pi'(t) = \pi(t)Q(t)$. General ODE-solvers can be used to solve this equation. The closed-form solution is no longer the exponential of a matrix and Jensen's method or similar methods can not be used. The usual way to overcome this problem is to reduce the non-homogeneous Markov process to a homogeneous Markov process by considering small enough time intervals [120,45,49]. The idea is to assume that the generator $Q(t)$ is constant on these intervals, so that on each interval the process can be treated as a homogeneous process. This approach is exact if the generator is piece-wise constant and it will be a good approximation if $Q(t)$ is continuous and the intervals are small enough.

Nowadays, the need for non-homogeneous models has decreased because of the homogeneous models that are available. For example, the variability of peak hours can very well be modelled by a Markov-Modulated Poisson Process or a Markovian Arrival Process. These can not model peak hours with an exactly deterministic duration, but can approximate any arrival process arbitrarily close [8]. If both the arrival and the service process vary with time, a random environment can be used [113].

Several approximations for non-homogeneous processes are available. Often non-homogeneous processes are cyclic. Then a common, cautious, approach is to analyze the worst-case homogeneous Markov process, for example based on busy-hour parameters. Another approximation is the pointwise-stationary approximation, obtained by assuming that the transient distribution is constantly equal to the steady-state distribution of the current generator [66,105]. This approximation will be accurate if the generator $Q(t)$ changes slowly, compared to the relaxation times. For more sophisticated approximations, see [45,52,94,114,121].

In the previous section 1.1 on steady-state analysis, it was observed that the special structure of queueing models could be exploited to design more efficient solution procedures. This appears to be more difficult for the transient analysis. Exceptions are some one-dimensional birth-death processes [86,92,116]. Especially the $M/M/1$ queue has received ample attention, both in the homogeneous [1,3,10,43,90,95] and in the non-homogeneous case [105,122]. For the homogeneous $BMAP/G/1$, two-dimensional transforms are available, with one transform variable relating to time. These can be inverted numerically [37,103]. The non-homogeneous Erlang loss model was studied in [32,45]. For networks of queues, only the infinite-server case seems to be tractable [17,33,69,80,93,106], because different customers do not influence each other.

1.3 Contents of the thesis

This thesis consists of two main parts, chapters 2 and 3. Chapter 2 provides and studies the PSA to analyze the *steady-state* distribution of continuous-time Markov processes. It starts with a simple example to illustrate the main idea of the method. The method

that will be proposed is a generalization of the light-traffic approach used in all previous papers. It has the advantage that it can be used to analyze very general networks of queues, as done in [75] and section 2.8 of this thesis. These networks have a Multi-queue Markovian Arrival Process ($MMAP$), Markovian Service Process (MSP) and Markovian routing. The $MMAP$ is a network generalization of the Batch Markovian Arrival process ($BMAP$). The MSP is also very similar to the $BMAP$, but for the service process instead of the arrival process. Both the $MMAP$ and the MSP were not introduced before, probably because this type of networks is too general to be analyzed by other methods. For some models, the transformation parameter has a physical interpretation, like the load of the system or a blocking probability. A recursive algorithm is derived to obtain the coefficients of the power-series expansions of the steady-state probabilities, but also of other performance measures and their derivatives. The algorithm is extensively studied and remedies are found for several previous imperfections of the algorithm. It is shown for a wide class of models that the steady-state probabilities are analytic functions of γ. So far, this could only be proved for very special models. Since analyticity is the basic assumption of the PSA, this greatly improves the theoretical basis of the PSA.

Chapter 3 tries to use the ideas of the PSA to calculate the *transient* distribution of a continuous-time Markov process. First, Jensen's method is described and it is shown that this method can be made uniformly convergent over time when information about the steady-state distribution is used. The PSA for transient analysis is a generalization of Jensen's method. It will turn out that this generalization does not lead to efficient algorithms for homogeneous processes. However, it does provide interesting theoretical results. For non-homogeneous processes there is reason to believe that the algorithm can be useful, but this has not yet been tested numerically.

The appendices contain supplementary information on Markov processes, analytic functions and extrapolation methods. Appendix A gives a short overview of definitions from the theory of Markov processes. Also, it provides sufficient conditions for ergodicity of Markov processes that are less restrictive than existing conditions in the literature. It is based on the existence of a non-null solution to the balance equations, but does not require that this solution is positive or that the Markov process is uniformizable. Appendix B provides an overview of some parts of the theory of analytic functions. Knowledge of these results is essential for a good understanding of the theoretical background of the PSA. Appendix C describes some extrapolation methods. They are indispensable for an efficient implementation of the PSA.

1.4 Notation

For reference, some of the notation used in this thesis is reviewed here. Throughout the thesis, double numbers will be used to refer to formulae, theorems, tables and figures.

The first number indicates the chapter or appendix. References to formulae will be between brackets. The abbreviation LHS will be used for the *left-hand side* of a formula and RHS for the *right-hand side*.

Vectors

Unless indicated otherwise, vectors will be row vectors. The vector o is a vector of zeros and e a *column* vector of ones. The unit vectors e_s are row vectors of zeros except for component s, which equals 1. All these vectors have appropriate size. If the size of the vector is S, then $e_0 = e_{S+1} = o$.

The plus and minus signs are used as superscripts of vectors to denote the element-wise maximum and minimum of a vector with the zero vector: $x^- \leq o \leq x^+$ and $x = x^- + x^+$. Inequalities of vectors are true if they are true element-wise.

The operator Δ can be applied to elements of vectors and will denote the first order difference with respect to the subindex: $\Delta x_k = x_{k+1} - x_k$. The ℓ-th order difference is recursively defined by $\Delta^1 = \Delta$ and $\Delta^\ell = \Delta\Delta^{\ell-1}$ for $\ell \geq 2$. When applied to a vector of size K, Δx denotes the vector of size $K-1$ with the k-th element equal to Δx_k.

Matrices

Matrices will be printed in capitals, the elements of matrices in the corresponding small letters with subindices: $A = [a_{ij}]$. The matrix O is a matrix of zeros and I the usual unit matrix, both with appropriate size. The matrices I_ℓ are the unit matrices of size ℓ.

The operator $\otimes$ denotes the Kronecker product and $\odot$ the Hadamard (or element-wise) product:

$$A \otimes B \doteq \begin{pmatrix} a_{11}B & \dots & a_{1n}B \\ \vdots & & \vdots \\ a_{m1}B & \dots & a_{mn}B \end{pmatrix}, \quad A \odot B \doteq \begin{pmatrix} a_{11}b_{11} & \dots & a_{1n}b_{1n} \\ \vdots & & \vdots \\ a_{m1}b_{m1} & \dots & a_{mn}b_{mn} \end{pmatrix}.$$

The exponential of a matrix A is the matrix formally defined by

$$\exp(A) \doteq \sum_{k \geq 0} \frac{1}{k!} A^k. \tag{1.2}$$

When a scalar is added to a matrix, this is an abbreviation for adding the scalar to all diagonal elements: $A + \alpha \doteq A + \alpha I$. The matrix A^* will denote the matrix that is equal to A, but with the first column removed. The notation $[x, A^*]$ will be used for the matrix that is equal to A, but with the first column replaced by the column vector x. The superscripts d and o will be used to denote the diagonal and off-diagonal parts of a matrix: $A = A^d + A^o$, with A^d a diagonal matrix.

Norms of matrices

Throughout this thesis, the maximal-absolute-row-sum norm will be used:

$$\|A\| \doteq \sup_i \sum_j |a_{ij}| \,. \tag{1.3}$$

Notice that $\|e\| = 1$, because e is a column vector, whereas $\|e^T\|$ is equal to the size of the vector. The triangular inequality $\|A + B\| \leq \|A\| + \|B\|$ holds and the norm is submultiplicative: $\|AB\| \leq \|A\|\|B\|$. The norm of the exponential of a matrix is at most equal to the exponential of the norm of the matrix: $\|\exp(A)\| \leq \exp(\|A\|)$. Finally, $\|A^*\| \leq \|A\|$ and $\|[x, A^*]\| \leq \|x\| + \|A\|$.

The maximal-absolute-row-sum norm (1.3) is used for convenience. Applied to probabilistic quantities, it renders much nicer results than for example the spectral norm. If A is a stochastic matrix or row vector, then $\|A\| = 1$. Also, this norm is very convenient when A is a generator, that is if A is non-negative except for the non-positive diagonal elements and if $Ae \leq o$ (see page 104 of appendix A). Then the maximal-absolute-row-sum norm has a close relation to the maximal-absolute-element norm

$$\sigma(A) \doteq \sup_{i,j} |a_{ij}| \,. \tag{1.4}$$

If A is a generator, then $\sigma(A) = \sup_i |a_{ii}|$ and the following inequalities hold:

$$\begin{aligned} \sigma(A) \leq \|A\| &\leq 2\sigma(A), \\ \|A + \alpha\| \leq \sigma(A) &+ |\sigma(A) - \alpha|, \qquad \text{for all } \alpha \in \mathbb{R}. \end{aligned}$$

If A is an *honest* generator, that is $Ae = o$, then the second and third inequalities hold with equality.

On finite matrices of equal size, all norms are equivalent: for any two different norms $\|.\|_1$ and $\|.\|_2$, positive constants α, β exist such that the inequalities $\alpha\|A\|_1 \leq \|A\|_2 \leq \beta\|A\|_1$ hold for any matrix A. On infinite matrices, this is no longer true. As a result of this, depending on the norm that is used, a sequence of matrices may converge or not. For example, consider the maximal-absolute-column-sum norm $\|A\|_1 = \sup_j \sum_i |a_{ij}| = \|A^T\|$ and the maximal-absolute-row-sum norm $\|A\|_2 = \|A\|$. Let x_k be the *row* vector with the first k elements equal to $1/k$ and the other elements equal to zero. Then $\|x_k\|_1 = 1/k$ and $\|x_k\|_2 = 1$, for all $k \geq 0$. So, according to the first norm the sequence of vectors x_k converges to the zero vector o. But, according to the second norm, the sequence does not converge to o nor to any other vector. Row vectors will be used more frequently than column vectors in this thesis. For row vectors, convergence according to the maximal-absolute-row-sum norm (1.3) implies convergence according to any norm $\|.\|_3$ such that $\sup_i \|e_i\|_3 < \infty$. This is true for the normally used norms, like the maximal-absolute-column-sum, maximal-absolute-element and the spectral norm.

Sets and functions

Sets will be denoted by calligraphic symbols (just as stochastic variables) and # denotes the number of elements in a set. The indicator function $1(\omega)$ is equal to one if expression ω is true and zero otherwise. The operators $\lceil . \rceil$ and $\lfloor . \rfloor$ denote rounding up and down:

$$\lceil x \rceil \doteq \min \{ \ z \in \mathbb{Z} \mid z \geq x \}, \qquad \lfloor x \rfloor \doteq \max \{ \ z \in \mathbb{Z} \mid z \leq x \},$$

for all $x \in \mathbb{R}$.

For any $\ell \in \mathbb{N}$, the set '$\mathcal{O}\left(x^{\ell}\right)$, for $x \downarrow 0$' will denote the set of all functions $f : \mathbb{C} \to \mathbb{C}$ satisfying

$$\exists_{\delta>0,\ M\geq 0} \text{ such that } |x| < \delta \Rightarrow |f(x)| \leq M|x|^{\ell}.$$

Similarly, for any $\alpha \geq 0$, the set '$\mathcal{O}\left(e^{-\alpha x}\right)$, for $x \to \infty$' denotes the set of all functions $f : \mathbb{R} \to \mathbb{R}$ satisfying

$$\exists_{D,M\geq 0} \text{ such that } x > D \Rightarrow |f(x)| \leq Me^{-\alpha x}.$$

The notation $O\left(x^{\ell}\right)$ and $O\left(e^{-\alpha x}\right)$ will be used to denote arbitrary functions from these sets of functions.

Chapter 2

The PSA for steady-state analysis

The PSA for steady-state analysis is best illustrated by a very simple example. Consider the two-state continuous-time Markov process (CTMP) in figure 2.1. The transition rate from state 1 to state 2 is equal to 1 and the transition rate back from state 2 to state 1 is equal to 2. This Markov process is ergodic and the steady-state distribution can be

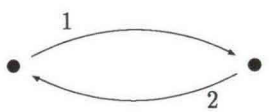

Figure 2.1: The original process

calculated from the balance and normalization equations

$$1\ \pi_1 = 2\ \pi_2,$$
$$\pi_1 + \pi_2 = 1.$$

The first equation reflects that in steady state the rate at which the process leaves a state is equal to the entrance rate to that state. The second equation sets the total probability equal to one. Now, suppose that solving this set of two linear equations with two unknowns would be a task too difficult to accomplish. One could then try to solve the above problem by considering the more general Markov process in figure 2.2. The

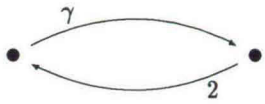

Figure 2.2: The transformed process

transition rate from state 2 to state 1 is still equal to 2, but the upward rate is replaced by γ. The steady-state distribution of this new process can be found from the balance and normalization equations

$$\begin{aligned} \gamma\ \pi_1 &= 2\ \pi_2, \\ \pi_1 + \pi_2 &= 1. \end{aligned}$$

The solution is now a function of γ, so at first sight this does not simplify the problem at all. However, suppose that the steady-state probabilities are *analytic* functions of γ, at $\gamma = 0$. Then the steady-state probabilities can be represented by power series in γ:

$$\begin{aligned} \pi_1(\gamma) &= \sum_{r=0}^{\infty} \gamma^r u_{1r}, \\ \pi_2(\gamma) &= \sum_{r=0}^{\infty} \gamma^r u_{2r}. \end{aligned}$$

Substituting this in the balance and normalization equations renders

$$\begin{aligned} \sum_{r=0}^{\infty} \gamma^{r+1} u_{1r} &= 2 \sum_{r=0}^{\infty} \gamma^r u_{2r}, \\ \sum_{r=0}^{\infty} \gamma^r \left[u_{1r} + u_{2r}\right] &= 1. \end{aligned}$$

Analytic functions are only identical if all corresponding coefficients of their power-series expansions are identical. So in the equalities above, the constants on either side of the equality sign must be equal and also the linear coefficients, the quadratic coefficients, and so on. This renders the equalities

$$\begin{array}{lll} 0 = u_{20}, & u_{1,r-1} = 2\ u_{2r}, & \text{for all } r \geq 1, \\ u_{10} + u_{20} = 1, & u_{1r} + u_{2r} = 0, & \text{for all } r \geq 1. \end{array}$$

These equations allow for the recursive calculation of the coefficients of the power-series expansions:

r	0		1		2		3		4		5
u_{1r}	1		$-\frac{1}{2}$		$\frac{1}{4}$		$-\frac{1}{8}$		$\frac{1}{16}$		
	$\uparrow$	$\searrow$	$\uparrow$	$\searrow$	$\uparrow$	$\searrow$	$\uparrow$	$\searrow$	$\uparrow$	$\searrow$	$\uparrow$
u_{2r}	0		$\frac{1}{2}$		$-\frac{1}{4}$		$\frac{1}{8}$		$-\frac{1}{16}$		

The closed-form solution is

$$\begin{aligned} u_{1r} &= \left(-\tfrac{1}{2}\right)^r, && \text{for all } r \geq 0, \\ u_{2r} &= -\left(-\tfrac{1}{2}\right)^r 1(r>0), && \text{for all } r \geq 0. \end{aligned}$$

Truncating the power series after the R-th coefficient renders the approximations

$$\begin{aligned} \tilde{\pi}_1(\gamma) &= \sum_{r=0}^{R} \gamma^r \left(-\tfrac{1}{2}\right)^r = \frac{1-\left(-\frac{1}{2}\gamma\right)^{R+1}}{1+\frac{1}{2}\gamma}, \\ \tilde{\pi}_2(\gamma) &= -\sum_{r=1}^{R} \gamma^r \left(-\tfrac{1}{2}\right)^r = \frac{\frac{1}{2}\gamma+\left(-\frac{1}{2}\gamma\right)^{R+1}}{1+\frac{1}{2}\gamma}. \end{aligned} \tag{2.1}$$

To solve the original problem in figure 2.1, the power-series expansions can be evaluated at $\gamma = 1$:

R	0	1	2	3	4	5
$\tilde{\pi}_1(1)$	1.00000	0.50000	0.75000	0.62500	0.68750	0.65625
$\tilde{\pi}_2(1)$	0.00000	0.50000	0.25000	0.37500	0.31250	0.34375

These approximations converge to $\pi_1 = \frac{2}{3}$ and $\pi_2 = \frac{1}{3}$, when $R \to \infty$. So, without explicitly solving any sets of equations, the original problem can be solved to any degree of accuracy by calculating enough coefficients of the power-series expansions and evaluating these expansions at $\gamma = 1$.

Application of the PSA always essentially follows this procedure: introduce a transformation parameter γ, assume that the steady-state probabilities are analytic in γ, substitute the power-series expansions in the balance and normalization equations and equate the coefficients of corresponding powers of γ. To illustrate possible difficulties, notice that the power-series expansions (2.1) only converge for $R \to \infty$ if $|\gamma| < 2$. They converge slowly if γ is close to 2 and diverge if $\gamma \geq 2$. However, there are methods to overcome these problems. In section 2.6 and appendix C, a number of methods is discussed. For example, the epsilon algorithm would greatly improve the convergence properties in this example. It would guess that the obtained truncated power series come from the functions $\frac{2}{2+\gamma}$ and $\frac{\gamma}{2+\gamma}$. The truncation level $R = 2$ would then suffice to find the correct steady-state distribution for any $\gamma \geq 0$.

The rest of this chapter deals with the justification of the above approach and the extension to general Markov processes. In section 2.1 the notation of the general Markov process is introduced, in section 2.2 that of the transformed process. In section 2.3, the recursive algorithm is derived to find the power-series expansions of the steady-state probabilities, performance measures and derivatives of the transformed process. This section also contains a discussion of the characteristics of the algorithm, compared to other methods. Some general remarks on analyticity of Markov processes will be made

in section 2.4. For the transformed Markov process considered by the PSA, section 2.5 provides sufficient conditions to ensure convergence of the power-series expansions, for small values of the transformation parameter. Section 2.6 studies what can still go wrong if these conditions are satisfied and section 2.7 discusses what may happen if they are not satisfied. Finally, in section 2.8, the algorithm is applied to a wide class of queueing networks and some different ways of applying the PSA are compared. Several definitions of concepts from the theory of Markov processes can be found in appendix A.

2.1 The Markov process

The PSA is especially well suited for multidimensional Markov processes. The notation will explicitly reflect this multidimensional nature. Continuous-time Markov processes $\{(\mathcal{N}_t, \mathcal{I}_t); t \geq 0\}$ will be considered on state space $\Omega = \mathbb{N}^S \times \{1, \ldots, I\}$. This state space consists of the S-dimensional vectors of natural numbers and a finite supplementary space. In a queueing context, S could be the number of queues, $\mathcal{N}_t$ the queue-length process and $\mathcal{I}_t$ a supplementary variable to model for example non-exponentiality of the arrival and service processes. This thesis was written with the application to queueing models in mind. Therefore, queueing terminology will be used, although the considered Markov process need not be a queueing process. Many multidimensional Markov processes can be rephrased as queueing processes.

In principle, all transitions and transition rates are allowed:

$$(n, i) \rightarrow (n + b, j) \quad \text{with rate} \quad \alpha_{bj}(n, i), \tag{2.2}$$

for all (n, i), $(n + b, j) \in \Omega$. The transition rates are indexed by the *change* b of the queue-length variable, and not the *new value* $n + b$, which is more usual. The reason for this is that the transformed process presented in the next section will explicitly depend on b and not on $n + b$.

In this chapter, the problem will be considered of finding the steady-state distribution of the CTMP $\{(\mathcal{N}_t, \mathcal{I}_t); t \geq 0\}$, with performance measures and derivatives. The vector $(\mathcal{N}, \mathcal{I})$ will denote random variables with the same distribution as the steady-state distribution. To ensure that the problem is well defined, it will be assumed that the Markov process is non-instantaneous and ergodic. Then the steady-state probabilities, defined by

$$p(n, i) \doteq \lim_{t \to \infty} \mathbb{P}\{\ (\mathcal{N}_t, \mathcal{I}_t) = (n, i)\ \} = \mathbb{P}\{\ (\mathcal{N}, \mathcal{I}) = (n, i)\ \},$$

exist for all $(n, i) \in \Omega$ and are independent of the initial conditions $(\mathcal{N}_0, \mathcal{I}_0)$. They are uniquely determined by the balance and normalization equations

$$\begin{aligned} p(n, i)\ \bar{\alpha}(n, i) &= \sum_{(n-b,j) \in \Omega} p(n - b, j)\ \alpha_{bi}(n - b, j), \quad \text{for all } (n, i) \in \Omega, \\ \sum_{(n,i) \in \Omega} p(n, i) &= 1, \end{aligned} \tag{2.3}$$

with $\bar{\alpha}(n,i)$ equal to the total transition rate out of state (n,i):

$$\bar{\alpha}(n,i) \doteq \sum_{(n+b,j)\in\Omega} \alpha_{bj}(n,i), \quad \text{for all } (n,i) \in \Omega.$$

To suppress the variable for the supplementary space, it will be convenient to write the balance and normalization equations in matrix notation:

$$\begin{aligned} p(n)\bar{A}(n) &= \sum_{n-b\in\mathbb{N}^S} p(n-b)A_b(n-b), \quad \text{for all } n \in \mathbb{N}^S, \\ \sum_{n\in\mathbb{N}^S} p(n)e &= 1, \end{aligned} \tag{2.4}$$

with

$$\begin{aligned} p(n) &= [\ p(n,i)\ ], \\ A_b(n) &= [\ \alpha_{bj}(n,i)\ ], \\ \bar{A}(n) &= [\ \bar{\alpha}(n,i)\ 1(i=j)\ ], \end{aligned}$$

for $n, n+b \in \mathbb{N}^S$ and with e a column vector of ones. The i-th element of the row vector $p(n)$ is the steady-state probability of state (n,i). The (i,j)-th element of the matrix $A_b(n)$ is the transition rate from state (n,i) to state $(n+b,j)$ and the matrix $\bar{A}(n)$ is the diagonal matrix with the i-th diagonal element equal to the total departure rate from state (n,i).

2.2 The transformed Markov process

The set of balance equations (2.4) can in general be infinite and hence difficult to solve. The PSA aims to be an efficient way of solving it. The PSA transforms the original Markov process with a parameter γ in such a way that for $\gamma = 1$ the transformed process is equal to the original process. The steady-state probabilities of the transformed process can be regarded as a function of γ. If the transformation is chosen in an appropriate way, they will be analytic functions of γ at $\gamma = 0$ and the coefficients of the power-series expansions can be calculated recursively. The steady-state distribution of the original process is then found by evaluating these power series at $\gamma = 1$. When the transformation parameter γ has a physical interpretation, a range of values $\gamma \in [0, \gamma^*)$ will be of interest.

Not any transformation will do. In the example in the introduction of this chapter, if not only the upward but also the downward transition rate had been multiplied by γ, then the coefficients could not have been calculated recursively. In this section one particular transformation will be proposed and, in section 2.5, sufficient conditions will be derived under which this particular transformation is appropriate. This transformation is not necessarily the best or the only possible choice. For some Markov processes it may not work or other transformations may work better. In sections 2.7.5 and 2.8.5, alternative transformations will be considered. However, for many quasi birth-death processes and many queueing applications the transformation proposed here seems to be the most

natural transformation.

The transformed Markov process for arbitrary values of the transformation parameter $\gamma \in [0,1]$ will be called the γ-process. The 0-process is the γ-process with $\gamma = 0$. The 1-process is the γ-process with $\gamma = 1$, which will be equal to the original process. To specify the γ-process, define the following subsets of $\mathbb{Z}^S$:

$$\mathcal{Z}_< \doteq \left\{ b \in \mathbb{Z}^S \;\middle|\; be < 0 \right\},$$

$$\mathcal{Z}_\geq \doteq \left\{ b \in \mathbb{Z}^S \;\middle|\; be \geq 0 \right\}.$$

Transitions $(n,i) \to (n+b,j)$ with $b \in \mathcal{Z}_<$ decrease the total queue length and will be called downward transitions. Those with $b \in \mathcal{Z}_\geq$ will be called upward transitions. Also, the vectors b themselves will be called downward or upward transitions, even though strictly speaking they are vectors and not transitions of the process. A selfloop is a transition with $b = o$, which possibly changes the supplementary variable but does not change the queue length.

In the γ-process, the downward transitions $b \in \mathcal{Z}_<$ will be the same as in the original Markov process. The upward transitions will be transformed. For each upward transition $b \in \mathcal{Z}_\geq$, define the number

$$r_b \doteq \begin{cases} be, & \text{if } be > 0, \\ 1, & \text{if } be = 0. \end{cases}$$

This number is equal to the increase of the total queue length caused by transition b if this increase is positive and equal to 1 if the total queue length remains constant. In the transformed process, each upward transition $(n,i) \to (n+b,j)$ is split into two transitions:

$$\begin{array}{lllll} (n,i) & \to & (n+b,j) & \text{with rate} & \gamma^{r_b} \quad \alpha_{bj}(n,i), \\ (n,i) & \to & (n+b^-,j) & \text{with rate} & (1-\gamma^{r_b}) \quad \alpha_{bj}(n,i), \end{array} \tag{2.5}$$

for all $b \in \mathcal{Z}_\geq$ and $\gamma \in [0,1]$. The original transition $(n,i) \to (n+b,j)$ is still possible, but the original rate $\alpha_{bj}(n,i)$ is multiplied by γ^{r_b}. The total transition rate from each state is the same as in the original Markov process, but the upward transition b is with probability $(1-\gamma^{r_b})$ replaced by the transition b^-, a downward transition or selfloop. The vector b^- is the element-wise minimum of b and the zero vector (see section 1.4). Hence, b^- is always non-positive and $b^- = o$ if and only if b is non-negative. If both (n,i) and $(n+b,j)$ are in Ω, then so is $(n+b^-,j)$. Some examples can be found in table 2.1, at the end of this section.

If the γ-process is ergodic, the steady-state distribution is uniquely determined by

the balance and normalization equations:

$$\begin{aligned} p(\gamma,n)\bar{A}(n) &= \sum_{b\in\mathcal{Z}_<} & & p(\gamma,n-b) & A_b(n-b) \\ &+ \sum_{b\in\mathcal{Z}_\geq} & \gamma^{r_b} & p(\gamma,n-b) & A_b(n-b) \\ &+ \sum_{b\in\mathcal{Z}_\geq} & (\,1-\gamma^{r_b}\,) & p(\gamma,n-b^-) & A_b(n-b^-), \\ \sum_{n\in\mathbb{N}^S} p(\gamma,n)e &= 1, \end{aligned} \tag{2.6}$$

for $n \in \mathbb{N}^S$ and $\gamma \in [0,1]$. The transitions $b \in \mathcal{Z}_<$ are the same as in (2.4). The transitions $b \in \mathcal{Z}_\geq$ are split into transition b with probability γ^{r_b} and transition b^- with probability $(\,1-\gamma^{r_b}\,)$.The 1-process is the same as the original Markov process, with balance equations (2.4). The 0-process is a Markov process with only downward transitions and selfloops.

The most important aspect of the transformation is that the rates of transitions that increase the total number of customers by r are multiplied by γ^r ($r = be > 0$). As a consequence of this, the order of each steady-state probability is equal to the total number of customers ne in the system:

$$p(\gamma,n) \in \mathcal{O}(\gamma^{ne}), \quad \text{for } \gamma \downarrow 0 \text{ and } n \in \mathbb{N}^S. \tag{2.7}$$

This property will be referred to as the order property and will be proved in section 2.5. Also, sufficient conditions will be given for $p(\gamma,n)$ to be analytic in γ. If these conditions hold, the functions $p(\gamma,n)$ can be represented by their power-series expansions (see (2.8) below). Property (2.7) then implies that the coefficients corresponding to γ^r are zero for all states with more than r customers in the system, so for each fixed r there is only a finite number of non-zero coefficients. Otherwise it would be impossible to calculate the coefficients.

A second aspect of the transformation is that transitions that keep the total number in the system constant ($be = 0$) are multiplied by γ. Without this, in the r-th step of the algorithm one large set of equations would have to be solved with, in general, size $I \times \binom{r+S-1}{S-1}$, while now $\binom{r+S-1}{S-1}$ sets of equations with size I need to be solved, which is usually much easier. This will be illustrated in section 2.8.5.

Finally, the extra transitions b^- are added, for $b \in \mathcal{Z}_\geq$. These transitions are added to extend the class of Markov processes that can be handled. In section 2.3 it is shown that the algorithm is well defined if the 0-process has a single recurrent class consisting of only empty states. Since the extra transitions are non-positive, these extra transitions extend the class of models for which this assumption is satisfied. For example, without these extra transitions only feed-forward networks could be analyzed, while now networks with general Markovian routing can be studied [75]. For an extensive example, see section 2.8.5. If all upward transitions are strictly positive and do not change the

supplementary variable, like in all birth-death processes and some quasi birth-death processes, then all added transitions are selfloops. These can be ignored without changing the behaviour of the process.

Example 2.1 Consider the network in figure 2.3. Customers arrive simultaneously at both queues, according to a Poisson process with rate λ. At queue 1, the service rate is μ_1 and each customer whose service is completed joins queue 2. At queue 2, the service rate is μ_2 and each customer whose service is completed leaves the network.

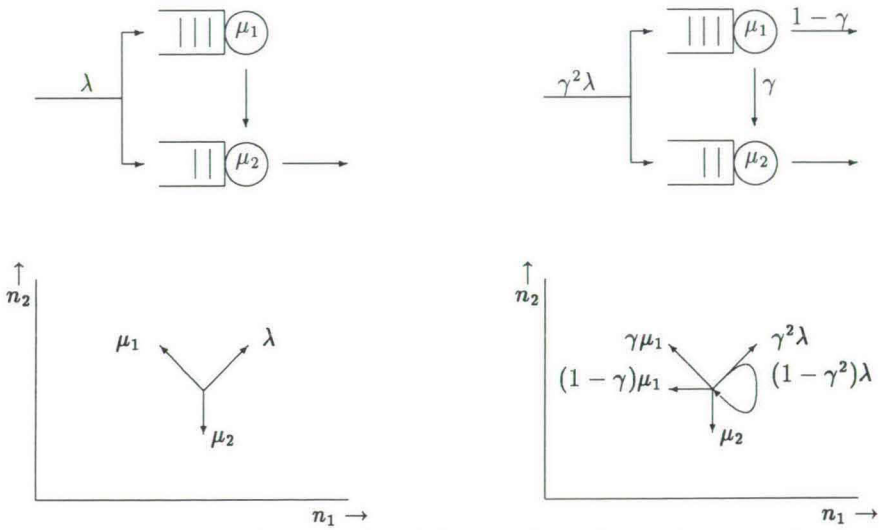

Figure 2.3: The original and the transformed queueing network

An arrival increases the total queue length by two, so in the transformed model the arrival rate is multiplied by γ^2. The associated new transition b^- is a selfloop with rate $(1-\gamma^2)\lambda$. This selfloop does not influence the behaviour of the process, so it can be ignored. A service completion at queue 1 does not change the total queue length, so the rates of these transitions are multiplied by γ. Customers now leave queue 1 to join queue 2 with rate $\gamma\mu_1$. The associated new transition is a departure from the network with rate $(1-\gamma)\mu_1$. A service completion at queue 2 decreases the total queue length, so these transitions are downward and they are not transformed. The parameters are given in table 2.1.

b	$\alpha_b(n)$	r_b	b^-
$\binom{1}{1}$	λ	2	$\binom{0}{0}$
$\binom{-1}{1}$	$\mu_1\ 1(n_1 \geq 1)$	1	$\binom{-1}{0}$
$\binom{0}{-1}$	$\mu_2\ 1(n_2 \geq 1)$	–	–

Table 2.1: The parameters of the network.

□

2.3 The Power-Series Algorithm

In this section the algorithm is derived to analyze the transformed process introduced in the previous section. For the steady-state probabilities, it calculates the coefficients of the power-series expansions in the transformation parameter γ. From these, the power-series expansions of performance measures and derivatives with respect to some model parameter will be obtained. The original process can be analyzed by evaluating the γ-process at $\gamma = 1$. In the derivation of the algorithm, the assumption is made that the steady-state probabilities are analytic in γ and that several orders of summation can be reversed. Sufficient conditions for this are given in section 2.5.

An essential part of an efficient implementation of the PSA is formed by procedures to improve the convergence of the power series, like conformal mappings and the epsilon algorithm. These will be discussed in section 2.6 and appendix C.

2.3.1 The steady-state distribution

For now, assume that the steady-state probabilities are analytic functions of γ at $\gamma = 0$. For a brief overview of the theory of analytic functions, see appendix B.3. Because the steady-state probabilities are analytic, they can be represented by the power-series expansions:

$$p(\gamma, n) = \sum_{r \geq ne} \gamma^r u(r, n), \quad \text{for all } n \in \mathbb{N}^S. \tag{2.8}$$

The summation starts at $r = ne$ because of the order property (2.7). If the γ-process is ergodic, then the steady-state probabilities satisfy the balance and normalization equations. Substituting the expansions in these equations renders equalities between functions of γ. Analytic functions are only identical if all corresponding coefficients of their power-series expansions are identical. This leads to equalities between the coefficients of

the power-series expansions that allow the recursive calculation of the coefficients. The idea of equating coefficients of analytic functions is a powerful idea. For example, it also gave birth to Padé approximation and Jensen's method for the transient analysis of Markov processes.

In more detail, the derivation of the recursive algorithm is as follows. Substituting the power-series expansions (2.8) in the balance and normalization equations (2.6) renders

$$\begin{aligned}
\sum_{r\geq ne} \gamma^r u(r,n)\bar{A}(n) &= \sum_{b\in\mathcal{Z}_<} \sum_{r\geq ne-be} \gamma^r \quad u(r,n-b) \quad A_b(n-b) \\
&+ \sum_{b\in\mathcal{Z}_\geq} \sum_{r\geq ne-be} \gamma^{r_b+r} \quad u(r,n-b) \quad A_b(n-b) \\
&+ \sum_{b\in\mathcal{Z}_\geq} \sum_{r\geq ne-b^-e} (\,1-\gamma^{r_b}\,)\gamma^r \quad u(r,n-b^-) \quad A_b(n-b^-), \\
\sum_{n\in\mathbb{N}^S} \sum_{r\geq ne} \gamma^r u(r,n)e &= 1.
\end{aligned} \tag{2.9}$$

These equations are equalities between functions of γ. If the functions on either side of the equality signs are analytic in γ, then the coefficients of corresponding powers of γ on either side must be equal (see theorem B.7): the constants on either side are equal, but also the linear term, the quadratic term and so on. Equating the r-th order term on either side leads to the equations

$$\begin{aligned}
u(r,n)\bar{A}(n) &= \sum_{b\in\mathcal{Z}_<} \quad u(r,n-b) \quad A_b(n-b) \\
&+ \sum_{b\in\mathcal{Z}_\geq} \quad u(r-r_b,n-b) \quad A_b(n-b) \\
&+ \sum_{b\in\mathcal{Z}_\geq} \quad u(r,n-b^-) \quad A_b(n-b^-) \\
&- \sum_{b\in\mathcal{Z}_\geq} \quad u(r-r_b,n-b^-) \quad A_b(n-b^-), \\
\sum_{ne\leq r} u(r,n)e &= 1(r=0),
\end{aligned} \tag{2.10}$$

for $0 \leq ne \leq r$. Basically, these are the recursive equations of the algorithm. To use them computationally, they need to be rewritten slightly. In the third summation of the RHS, the terms with $b \in \mathbb{N}^S$ must be brought to the left because then $u(r,n-b^-) = u(r,n-o) = u(r,n)$. Some of the terms in the second and fourth summation cancel out, namely those with $b = b^- = o$. With the definitions

$$\begin{aligned}
B(n) &= \bar{A}(n) - \sum_{b\in\mathbb{N}^S} A_b(n), && \text{for all } n \in \mathbb{N}^S, \\
u(r,n) &= o, && \text{if } n \notin \mathbb{N}^S \text{ or } r < ne, \\
\mathcal{Z}_1 &= \mathcal{Z}_\geq \setminus \{o\}, \\
\mathcal{Z}_2 &= \mathcal{Z}_\geq \setminus \mathbb{N}^S,
\end{aligned} \tag{2.11}$$

this leads to the following equalities

$$\begin{aligned} u(r,n)B(n) &= \sum_{b\in\mathcal{Z}_<} u(r,n-b) \quad A_b(n-b) \\ &+ \sum_{b\in\mathcal{Z}_1} u(r-r_b,n-b) \quad A_b(n-b) \\ &+ \sum_{b\in\mathcal{Z}_2} u(r,n-b^-) \quad A_b(n-b^-) \\ &- \sum_{b\in\mathcal{Z}_1} u(r-r_b,n-b^-) \quad A_b(n-b^-), \\ u(r,o)e &= 1(r=0) - \sum_{0<ne\leq r} u(r,n)e, \end{aligned} \tag{2.12}$$

for $n \in \mathbb{N}^S$ and $r \geq ne$.

For the I coefficients of the empty states, there are $I+1$ equations. One of them can be ignored, which comes down to ignoring one of the balance equations. For any matrix A, let A^* denote the matrix that is equal to A but with the first column removed (see section 1.4). Then, ignoring the balance equation of the first empty state $(o,1)$, equation (2.12) for $n = o$ can be reduced to

$$\begin{aligned} u(r,o)B^*(o) &= \sum_{b\in\mathcal{Z}_<} u(r,-b) \quad A_b^*(-b) \\ &+ \sum_{b\in\mathcal{Z}_2} u(r,-b^-) \quad A_b^*(-b^-) \\ &- \sum_{b\in\mathcal{Z}_1} u(r-r_b,-b^-) \quad A_b^*(-b^-), \\ u(r,o)e &= -\sum_{0<ne\leq r} u(r,n)e, \end{aligned} \tag{2.13}$$

for $r \geq 0$. The second summation on the RHS of (2.12) can be omitted here, because it is empty: $-b \notin \mathbb{N}^S$ for any $b \in \mathcal{Z}_1$. For $r = 0$, the equations can be simplified further:

$$\begin{aligned} u(0,o)B^*(o) &= o, \\ u(0,o)e &= 1. \end{aligned} \tag{2.14}$$

Now, all summations on the RHS can be omitted because of the order property (2.7).

These equations (2.12), (2.13) and (2.14) allow for the recursive calculation of all coefficients. They can be calculated in such an order that all coefficients on the RHS are obtained previous to the coefficient $u(r,n)$ on the LHS. The coefficients $u(\tilde{r},\tilde{n})$ on the RHS of both (2.12) and (2.13) satisfy either $\tilde{r} < r$ or they satisfy $\tilde{r} = r$ and $\tilde{n}e > ne$. Because $u(r,n) = o$ whenever $ne > r$, this implies that all coefficients can be calculated recursively for increasing values of r and, for each fixed r, for decreasing values of ne starting with $ne = r$. So, if the expansions of the steady-state probabilities (2.8) are to be calculated up to the coefficients of the R-th power of γ, the following algorithm can be used:

Power-Series Algorithm
calculate $u(0,o)$ from (2.14),
for $r := 1$ to R do
  for $N := r$ down to 1 do
    for all $n \in \mathbb{N}^S$ with $ne = N$ do
      calculate $u(r,n)$ from (2.12),
  calculate $u(r,o)$ from (2.13).

This algorithm is well defined if all sets of equations (2.12), (2.13) and (2.14) have a unique solution. Elementary linear algebra shows that necessary and sufficient for this is the following assumption:

Assumption 0' *The matrices $B(n)$, for $n \neq o$, and $[e, B^*(o)]$ are non-singular.*

Applied to birth-death processes, it reduces to the condition derived in [20]. This algebraic assumption 0' has a simple probabilistic interpretation:

Assumption 0 *The 0-process has a single recurrent class consisting of only empty states, which will eventually be reached from any state in Ω.*

This assumption can often be verified much easier. The equivalence of both assumptions will be proved in lemma 2.1. If there is no supplementary space ($I = 1$), the proof is elementary. Assumption 0' then reduces to the assumption that the scalars $B(n)$ are non-zero for all $n \neq o$. This is the same as assuming that, in the 0-process, each non-empty state has a positive transition rate to other states. Since the 0-process has only downward transitions and selfloops, this means that the empty state is the only absorbing state and will eventually be reached.

Lemma 2.1 *Assumption 0' and assumption 0 are equivalent.*

Proof: The equivalence of both assumptions will be shown by studying the 0-process in more detail. Besides the γ-processes on the complete state space Ω, Markov processes will be considered on finite sets with fixed queue length: $\Omega_n = \{n\} \times \{1, \ldots, I\}$ and $\Omega_n^\Delta = \{n\} \times \{1, \ldots, I, \Delta\}$, where Δ will be an absorbing state.

First, consider the non-empty states. The diagonal elements of $B(n)$ are equal to the rates in the 0-process out of the states in Ω_n to other states in Ω; the non-diagonal elements are equal to minus the rates in the 0-process from states in Ω_n to other states in Ω_n. Therefore, the elements of the vector $B(n)e$ are equal to the total rate in the 0-process of transitions from states in Ω_n to states not in Ω_n. Since the 0-process has no upward transitions, these transitions can only be downward and once the 0-process has left Ω_n it will never return. Aggregate all states not in Ω_n into a single state $\Delta = \Omega \backslash \Omega_n$.

Starting from a state in Ω_n, the 0-process then reduces to a process on the finite state space Ω_n^Δ. Entering Δ corresponds to a downward transition from Ω_n in the 0-process. The process on Ω_n^Δ has the following balance and normalization equations:

$$(\pi, \pi_\Delta) \begin{pmatrix} B(n) & -B(n)e \\ o & 0 \end{pmatrix} = (o, 0), \quad (\pi, \pi_\Delta) \begin{pmatrix} e \\ 1 \end{pmatrix} = 1.$$

The state space is finite, so one balance equation can be ignored. Replace the balance equation of state Δ by the normalization. If and only if $B(n)$ is invertible, the process on Ω_n^Δ has the unique steady-state distribution (π, π_Δ) equal to

$$(\pi, \pi_\Delta) = (o, 1) \begin{pmatrix} B(n) & e \\ o & 1 \end{pmatrix}^{-1} = (o, 1) \begin{pmatrix} B^{-1}(n) & -B^{-1}(n)e \\ o & 1 \end{pmatrix} = (o, 1).$$

The assumption that $B(n)$ is invertible for $n \neq o$ is therefore equivalent to the assumption that, in each Markov process on Ω_n^Δ, the state Δ is the only absorbing class. This in turn is equivalent to the assumption that, in the 0-process, all non-empty states are transient and the empty states will eventually be reached.

Next, consider the 0-process after it has reached an empty state. Once the 0-process is in the set Ω_o, it will never leave Ω_o. On Ω_o, the steady-state distribution is determined by the balance and normalization equations

$$\pi B(o) = o, \quad \pi e = 1. \tag{2.15}$$

This set of equations uniquely determines the steady-state distribution if and only if the Markov process on the finite state space Ω_o has only one recurrent class, but also if and only if the matrix $[e, B^*(o)]$ is invertible. Therefore, assumption 0' and assumption 0 are indeed equivalent. □

2.3.2 Performance measures

Often, one is not only interested in the steady-state distribution, but also in performance measures like means and (co)variances of the queue length distribution or the expectation of some reward or cost function. From the expansions of the steady-state probabilities, the expansions of any such performance measure can be obtained if it can be expressed as the expectation of a function $f : \Omega \to \mathbb{R}$. Let $f(n)$ be the column vector with the i-th element equal to $f(n, i)$. Then

$$\begin{aligned} \mathbb{E}_\gamma \{ f(\mathcal{N}, \mathcal{I}) \} &= \sum_{n \in \mathbb{N}^S} p(\gamma, n) \; f(n) \\ &= \sum_{n \in \mathbb{N}^S} \sum_{r \geq ne} \gamma^r \; u(r, n) \; f(n) \\ &= \sum_{r \geq 0} \gamma^r \; v(r), \end{aligned} \tag{2.16}$$

with

$$v(r) = \sum_{ne \le r} u(r,n) f(n), \quad \text{for all } r \ge 0. \tag{2.17}$$

The reversal of the order of summation in (2.16) is justified under very mild conditions. These will be derived in section 2.5. Examples of performance measures are:

$$\begin{aligned}
\mathbb{P}_\gamma \{ (\mathcal{N}, \mathcal{I}) \in \mathcal{E} \} &= \sum_{r \ge 0} \gamma^r \sum_{ne \le r, 1 \le i \le I} u_i(r,n)\, 1((n,i) \in \mathcal{E}), && \text{for all } \mathcal{E} \subseteq \Omega, \\
\mathbb{P}_\gamma \{ \mathcal{N} = n \} &= \sum_{r \ge ne} \gamma^r u(r,n) e, && \text{for all } n \in \mathbb{N}^S \\
\mathbb{E}_\gamma \left\{ (\mathcal{N}e)^\ell \right\} &= \sum_{r \ge 0} \gamma^r \sum_{ne \le r} (ne)^\ell \; u(r,n) e, && \text{for all } t \ge 0, \\
\mathbb{E}_\gamma \left\{ \mathcal{N}_s^\ell \right\} &= \sum_{r \ge 0} \gamma^r \sum_{ne \le r} n_s^\ell \; u(r,n) e, && \text{for all } 1 \le s \le S,\ t \ge 0, \\
\mathbb{E}_\gamma \{ \mathcal{N}_s \mathcal{N}_t \} &= \sum_{r \ge 0} \gamma^r \sum_{ne \le r} n_s n_t \; u(r,n) e, && \text{for all } 1 \le s, t \le S.
\end{aligned} \tag{2.18}$$

These are the probabilities of a particular subset of states or queue length, the ℓ-th moment of the total queue length or the queue length at queue s, and the cross-product of the queue lengths at queue s and t.

To calculate the r-th coefficient $v(r)$ of the expansion of the expectation, only the r-th coefficients $u(r,n)$ of the steady-state probabilities are needed and these coefficients are non-zero only for the finitely many states with $ne \le r$. So, if assumption 0 is satisfied, then also the power-series expansions of performance measures can be calculated. All that needs to be changed in the algorithm is that after the calculation of all r-th order coefficients of the steady-state probabilities, also the r-th order coefficient $v(r)$ of the performance measures is calculated:

Power-Series Algorithm for performance measures

```
calculate u(0,o) from (2.14),
calculate v(0) from (2.17),
for r := 1 to R do
    for N := r down to 1 do
        for all n ∈ N^S with ne = N do
            calculate u(r,n) from (2.12),
    calculate u(r,o) from (2.13),
    calculate v(r) from (2.17).
```

This renders all the coefficients of the R-th order truncated power-series expansion of the performance measure (2.16).

2.3.3 Derivatives

For optimization purposes one may be interested in calculating the derivatives of the steady-state probabilities or performance measures (see [27,29]). The power-series expansion in γ immediately provides the derivatives with respect to γ. Of course, this is

only interesting if the transformation parameter γ has a physical interpretation. In this section, the PSA will be extended to calculate also derivatives of arbitrary order with respect to some other model parameter ν.

The ℓ-th derivative of a function f with respect to ν will be denoted by $f^{(\ell)}$ and $f^{(0)} \doteq f$. Repeated application of the product rule shows that the ℓ-th derivative of the product of two functions f and g is equal to $[fg]^{(\ell)} = \sum_{k=0}^{\ell} \binom{\ell}{k} f^{(k)} g^{(\ell-k)}$. The derivatives of vectors and matrices will be the element-wise derivatives.

Assume that the derivatives of the steady-state probabilities are analytic functions of γ at $\gamma = 0$ and also in $\mathcal{O}(\gamma^{ne})$, for $\gamma \downarrow 0$:

$$p^{(\ell)}(\gamma, n) = \sum_{r \geq ne} \gamma^r u_\ell(r, n), \quad \text{for all } n \in \mathbb{N}^S. \tag{2.19}$$

The parameter γ is a parameter independent of the system parameter ν. Repeated differentiation of $p(\gamma, n)$ shows that the coefficients of the power-series expansions of the derivatives of the steady-state probabilities are the derivatives of the coefficients of the power-series expansions of the steady-state probabilities:

$$u_\ell(r, n) = u^{(\ell)}(r, n), \quad \text{for all } 0 \leq ne \leq r. \tag{2.20}$$

That is, provided that differentiation and summation can be reversed. This assumption is obvious if $p(\gamma, n)$ is analytic in ν, but it is not easily proved and not generally true.

If the reversal of differentiation and summation is allowed, then the recursive relations that determine the coefficients $u_\ell(r, n)$ can be found by differentiating the recursive relations that determine the $u(r, n) = u_0(r, n)$. Taking the ℓ-th derivative in equations (2.12), (2.13) and (2.14) leads to

$$\begin{aligned} u_\ell(r,n)B(n) \;=\; & - \sum_{k=0}^{\ell-1} \binom{\ell}{k} \; u_k(r,n) \; B^{(\ell-k)}(n) \\ & + \sum_{b \in \mathcal{Z}_<} \sum_{k=0}^{\ell} \binom{\ell}{k} \; u_k(r, n-b) \; A_b^{(\ell-k)}(n-b) \\ & + \sum_{b \in \mathcal{Z}_1} \sum_{k=0}^{\ell} \binom{\ell}{k} \; u_k(r-r_b, n-b) \; A_b^{(\ell-k)}(n-b) \\ & + \sum_{b \in \mathcal{Z}_2} \sum_{k=0}^{\ell} \binom{\ell}{k} \; u_k(r, n-b^-) \; A_b^{(\ell-k)}(n-b^-) \\ & - \sum_{b \in \mathcal{Z}_1} \sum_{k=0}^{\ell} \binom{\ell}{k} \; u_k(r-r_b, n-b^-) \; A_b^{(\ell-k)}(n-b^-), \end{aligned} \tag{2.21}$$

for $n \in \mathbb{N}^S \setminus \{o\}$, $r \geq ne$,

$$\begin{aligned}
u_\ell(r,o)B^*(o) &= -\sum_{k=0}^{\ell-1}\binom{\ell}{k} u_k(r,o) B^{*(\ell-k)}(o) \\
&\quad + \sum_{b\in\mathcal{Z}_<}\sum_{k=0}^{\ell}\binom{\ell}{k} u_k(r,-b) A_b^{*(\ell-k)}(-b) \\
&\quad + \sum_{b\in\mathcal{Z}_2}\sum_{k=0}^{\ell}\binom{\ell}{k} u_k(r,-b^-) A_b^{*(\ell-k)}(-b^-) \\
&\quad - \sum_{b\in\mathcal{Z}_1}\sum_{k=0}^{\ell}\binom{\ell}{k} u_k(r-r_b,-b^-) A_b^{*(\ell-k)}(-b^-), \\
u_\ell(r,o)e &= -\sum_{0<ne\leq r} u_\ell(r,n)e,
\end{aligned} \tag{2.22}$$

for $n = o$, $r > 0$, and

$$\begin{aligned}
u_\ell(0,o)B^*(o) &= -\sum_{k=0}^{\ell-1} u_k(0,o)B^{*(\ell-k)}(o), \\
u_\ell(0,o)e &= 1(\ell = 0).
\end{aligned} \tag{2.23}$$

From the expansions of the derivatives of the steady-state probabilities, the expansions of derivatives of performance measures can be obtained:

$$\begin{aligned}
\mathbb{E}_\gamma^{(\ell)}\{\, f(\mathcal{N},\mathcal{I})\,\} &= \sum_{n\in\mathbb{N}^S} [p(\gamma,n)f(n)]^{(\ell)} \\
&= \sum_{n\in\mathbb{N}^S}\sum_{r\geq ne} \gamma^r\, [u(r,n)f(n)]^{(\ell)} \\
&= \sum_{r\geq 0} \gamma^r \sum_{ne\leq r} [u(r,n)f(n)]^{(\ell)} \\
&= \sum_{r\geq 0} \gamma^r\, v_\ell(r),
\end{aligned} \tag{2.24}$$

with

$$v_\ell(r) = \sum_{ne\leq r}\sum_{k=0}^{\ell}\binom{\ell}{k} u_k(r,n)\, f^{(\ell-k)}(n), \quad \text{for all } r \geq 0, \tag{2.25}$$

for ℓ times differentiable functions $f : \Omega \to \mathbb{R}$.

So if the expansions of the first L derivatives of the steady-state probabilities (2.19) and performance measures (2.24) are to be calculated up to the coefficients of the R-th power of γ, the following algorithm can be used:

Power-Series Algorithm for derivatives

for $\ell := 0$ to L do
  calculate $u_\ell(0,o)$ from (2.23),
  calculate $v_\ell(0)$ from (2.25),
  for $r := 1$ to R do
    for $N := r$ down to 1 do
      for all $n \in \mathbb{N}^S$ with $ne = N$ do
        calculate $u_\ell(r,n)$ from (2.21),
    calculate $u_\ell(r,o)$ from (2.22),
    calculate $v_\ell(r)$ from (2.25).

If the transition rates and functions are L times differentiable, assumption 0 is again necessary and sufficient for all these sets of equations to have a unique solution. So, except for differentiability of each individual transition rate, no extra assumptions need to be made for the algorithm to be well defined.

2.3.4 Memory allocation

The maximal truncation level R of the power series is usually determined by the available memory space to store the calculated coefficients. Therefore, the memory should be used as efficient as possible. This section contains some comments on memory allocation.

The coefficients $u(r,n)$ form an $(S+1)$-dimensional array of vectors, since $r \in \mathbb{N}$ and $n \in \mathbb{N}^S$. Most programming languages only support rectangular data structures. These should not be used. For a truncation level R, all coefficients $u(r,n)$ are calculated with $0 \leq r \leq R$ and $ne \leq r$. Therefore, a rectangular data structure would need to have size $(R+1)^{S+1}$. The number of used elements is much less:

$$\#\left\{ (r,n) \in \mathbb{N}^{S+1} \mid 0 \leq r \leq R \text{ and } ne \leq r \right\} = \binom{S+1+R}{S+1}.$$

The quotient of both is approximately equal to $S!$, so especially with large numbers of queues only a small portion of the available memory is used. A more efficient way of storing the coefficients is by mapping the $(S+1)$-dimensional array onto a one-dimensional array with the mapping

$$C(r,n) = \sum_{s=0}^{S} \binom{s + r - \sum_{t=s+1}^{S} n_t}{s+1}.$$

For an efficient way to calculate these values, see [22]. The coefficients can be calculated in increasing order of $C(r,n)$. Apart from more efficient memory use, this mapping also has the advantage that the data structure does not depend on the number of queues. This simplifies writing a single computer program for models with different numbers of queues.

If the maximal increase of the total queue length by a single transition is finite and one is not interested in all steady-state probabilities, the memory requirements are much smaller. Let B be the size of the largest possible batch arrival:

$$B = \sup_{b \in \mathbb{Z}_{\geq}} \{ be \mid \text{a state } n \text{ exists such that } A_b(n) > 0 \}.$$

The calculation of the r-th order coefficient $v(r)$ of performance measures only calls for the r-th order coefficients $u(r,n)$ of the probabilities (see formula (2.17)). From the recursive equations (2.12) it is clear that to compute coefficient $u(r,n)$, only coefficients $u(s,m)$ are used with $r - B \leq s \leq r$ and $me \leq s$. Therefore, the coefficients of order

smaller than $r - B$ can be removed from memory. The maximal number of coefficients that need to be stored is equal to

$$\#\left\{ (s,m) \in \mathbb{N}^{S+1} \mid R - B \leq s \leq R \text{ and } me \leq s \right\} = \binom{S+R-B}{S} + \ldots + \binom{S+R}{S}. \quad (2.26)$$

Especially if B is small, this is considerably less than $\binom{S+1+R}{S+1}$. The easiest way to implement this idea is not to use the mapping $C(r,n)$, but instead use

$$C^*(r,n) = C(r,n) \text{ modulo } \left[\binom{S+R-B}{S} + \ldots + \binom{S+R}{S}\right].$$

The coefficients are now stored in the array in a cyclic way. The memory space of coefficients that are no longer needed is used for newly calculated coefficients. In specific cases, like quasi birth-death processes and symmetric models, the required memory size can be reduced further (see [22,24]).

2.3.5 Characteristics of the algorithm

According to Moler and Van Loan [111], the effectiveness of an algorithm is determined by the following attributes, listed in decreasing order of importance: generality, reliability, stability, accuracy, efficiency, storage requirements, ease of use and simplicity. In this section, the PSA will be discussed with respect to these various characteristics.

A method is *general* if it can be applied to a wide class of problems. On this characteristic the PSA scores very well. Only mild assumptions need to be made on the transition structure. Assumption 0 is satisfied for many queueing applications. Even if it is not satisfied, the algorithm can often be adjusted in such a way that the PSA is applicable (see section 2.7.5 and [87]). Other methods need to assume far more structure, like for example that there are only two queues or that the network is a product-form network.

Although the PSA does not need to assume much structure on the type of transitions, stiffness does provide limitations on the problems that can be analyzed. Stiffness arises when the transition rates of different transitions are of different orders of magnitude. If the system is stiff or heavily loaded, then the power series are likely to converge slowly or even diverge. Many coefficients will need to be calculated, introducing more round-off errors and leading to large computation times and memory requirements. Extrapolation methods are very helpful, but do not solve the problems completely. Any numerical method will suffer from this problem of stiffness, but it seems to affect the PSA quite strongly.

The generality of the PSA is both caused and limited by the fact that it directly depends on the balance equations. On the one hand, it has the effect that if the balance equations are slightly changed, then also the recursive equations of the PSA are only slightly changed. This makes the PSA very flexible. On the other hand, it also lays the curse of dimensionality on the PSA. Because of the direct dependence on the balance

equations, a large state space will cause large memory requirements for the PSA. Typically, models with up to 4-6 queues can be handled, depending on the structure, the stiffness and the load of the system, on the desired accuracy and on the available time and memory space. For some special models the curse can be lifted. For the symmetric shortest-queue model, up to 30 queues can be analyzed [24].

More general alternatives for the PSA are simulation and other methods for solving the balance equations. Simulation is more flexible and larger systems can be analyzed, also non-Markovian. For moderately sized systems, the PSA is faster and more accurate. Solving the balance equations by some other method is often not a feasible option for multidimensional models. The PSA greatly benefits from the fact that extrapolation methods can be used in a natural way. With the help of these methods, the PSA can be applied to models with quite heavy traffic. The PSA truncates in a very implicit and flexible way. Other methods for solving the balance equations truncate explicitly. The state space needs to be truncated at a level such that the probability mass in the truncated states is negligible. For more heavily loaded systems this often results in truncation levels that are too high. Only for stiff models, solving the balance equations by a stable truncation method will be preferred.

When considering reliability, stability and accuracy, it is important to distinguish between characteristics of a problem and those of a method. A method for inverting matrices can not be blamed for having trouble in inverting nearly singular matrices. Similarly, a method for finding the steady-state distribution of a Markov process can not be blamed for having troubles with stiff Markov processes. Stiffness does provide a good criterion for comparison of different methods.

An algorithm is *reliable* if it provides warnings when errors are introduced. Unfortunately, the PSA does not have error bounds. Without extrapolation methods, the power series often diverge, so these extrapolation methods are indispensable. Apart from convergence, it is usually very difficult to make meaningful statements about the extrapolated series. Therefore, it seems unlikely that useful error bounds will be found in the future. Still, the results turn out to be quite reliable. In all numerical evaluations so far, the obtained (extrapolated) power series either diverged or converged to a reasonable answer. The reliability of the results can be established by comparing the series at different truncation levels. Validation by means of known characteristics of the model has shown that the variation in the results for consecutive truncation levels is a trustworthy indication of the error.

An algorithm is *stable* if it does not introduce more sensitivity to perturbation than is inherent to the underlying problem. An unstable method can be reliable if the instability can be detected. The PSA is not a very stable method. In the paragraph on generality it was already mentioned that it seems more sensitive to stiffness than other methods. However, this is only sufficient reason to discard an algorithm if there are other superior methods. Superior methods may exist for models with special structures

but many models can not be analyzed by methods other than the PSA.

As defined in [111], *accuracy* primarily refers to the error introduced by truncating a series or by terminating iterations. The results of the PSA mostly become more accurate when a higher truncation level is used. However, this can not be guaranteed in general. In the high-order coefficients, the round-off errors become more severe. Using a bilinear mapping reduces this problem, but does not solve it completely. With models that are too stiff, it can be observed that initially the results get more accurate with increasing truncation levels. At a certain level the round-off errors take over and result in a fast divergence of the power series.

It is difficult to evaluate the *efficiency* and the *storage requirements* of the PSA, because it is not known in general what truncation level R can or needs to be used. This number varies from, say, 20 up to 120. Also, these characteristics can not be compared to other methods because other methods are often not available. What *can* be done is to evaluate the efficiency and memory requirements for a fixed number of queues S and truncation level R. The memory requirements and computation time of the PSA grow fast in both S and R. This limits the number of queues and the stiffness that can be handled. For most models, the memory capacity is more restrictive than the computation time.

In the previous section it was shown that the number of coefficients that need to be calculated is equal to $\binom{S+1+R}{S+1}$. The number of coefficients that need to be stored is at most equal to $\binom{S+R-B}{S} + \ldots + \binom{S+R}{S}$. The amount of work to compute a particular coefficient very much depends on the model under consideration. Each coefficient is a linear function of previously calculated coefficients. Therefore, an upper bound on the required number of matrix multiplications for a single coefficient is given by the number of calculated coefficients $\binom{S+1+R}{S+1}$. The total number of matrix multiplications is therefore at most $\binom{S+1+R}{S+1}^2$. This very crude bound is a polynomial in R, with order $2(S+1)$. For most models, however, the number of matrix multiplications to compute a single coefficient is bounded in R. Then the total number of matrix multiplications is a polynomial in R with order $(S+1)$. This is true for all quasi birth-death processes.

Calculating more performance measures hardly increases the memory requirements and computation time, because the coefficients of a performance measure form a one-dimensional array, whereas the coefficients of the steady-state probabilities form a multidimensional array. If L is the order of the highest derivative, then the memory requirements are linear in L and the computation time is quadratic in L.

On *ease of use* and *simplicity*, the PSA compares favourably to many other methods. The main idea is quite straightforward. Except for sophisticated extrapolation methods, no complicated mathematical procedures are used. A thorough understanding of extrapolation methods is advisable, but not absolutely necessary to obtain satisfactory results. It is often sufficient that the user can establish whether a series converges or not.

The reader is challenged to see for him or herself how easy and simple the PSA is, by applying it to the fork-join model [87]. Consider the model with S queues, simultaneous arrivals at all queues according to a Poisson arrival process with rate λ and different exponential service times with rate μ_s at queue s. The balance equations of this Markov process are

$$\left[\lambda + \sum_s \mu_s 1(n_s > 0)\right] p(n) = \lambda p(n-e) + \sum_s \mu_s p(n+e_s),$$

for all $n \in \mathbb{N}^S$ and $\lambda < \min_s \mu_s$. Applying the PSA as described in this chapter comes down to the following. Multiply λ by γ^S. Substitute the power-series expansions $p(\gamma, n) = \sum_{r=ne}^{\infty} \gamma^r u(r,n)$ in the new balance equations and equate coefficients of corresponding powers of γ. This renders recursive relations for the coefficients of all non-empty states. The coefficients of the empty state can be obtained by substituting the power-series expansions in the normalization equation $\sum_n p(\gamma, n) = 1$ and equating corresponding powers of γ. To analyze the original process, evaluate the obtained power series at $\gamma = 1$. Even without extrapolation methods, inspiring results will be obtained.

The considerations above lead to the conclusion that if specific methods are available to analyze a certain class of continuous-time Markov processes, then these methods may well be preferred to the PSA. However, the PSA is a flexible, fairly reliable and easy to use method that can handle many models unmanageable by other methods, provided the dimension of the problem is moderate and stiffness is mild.

2.4 Analyticity of Markov processes

In the transformed process, the transition rates are analytic functions of γ. Under what conditions does the analyticity of the transition rates of a Markov process imply analyticity of the steady-state probabilities? In this section, examples will be given to elucidate this question. In the next section, sufficient conditions will be derived under which the transformed process of the PSA has analytic steady-state probabilities.
If the state-space of an ergodic Markov process is finite and the transition rates are analytic functions of some parameter, then the steady-state probabilities are also analytic in that parameter. This is immediately shown by applying Cramer's rule for the inversion of a matrix to the balance and normalization equations: if $xA = b$, then $x_i = \det(A_i)/\det(A)$, where A_i is the matrix A with the i-th row replaced by b. The determinant of a finite matrix is a finite sum of finite products of elements of the matrix. So, if each element of a matrix is an analytic function of some parameter, then so is the determinant of the matrix. The quotient of two determinants is a rational function, and therefore analytic if the determinant in the denominator is non-zero. If the Markov process is ergodic, the balance and normalization equations are non-singular. Therefore,

the determinant in the denominator is indeed non-zero and the steady-state probabilities are analytic. A formal proof is given in the first paper on the PSA by Hooghiemstra, Keane and Van de Ree [73].

On infinite state spaces, this is no longer true. In the same paper [73], a counter-example by Aaronson and Gilat is given. In the example, the transition rates are analytic functions of a particular parameter, whereas the steady-state distribution is not even continuous in this parameter, let alone analytic. Example 2.2 below is a continuous-time version of this example.

If the steady-state probabilities *are* analytic, they need not be entire functions. The radius of convergence of the power-series expansion of a function is equal to the distance between the origin and the nearest singularity (see theorem B.8). Three examples will be given to illustrate where these singularities can be located. They are all examples of the transformed processes considered by the PSA. In example 2.3, the radius of convergence is infinite because there are no singularities at all. In example 2.4, the singularities are on a circle around the origin. In example 2.5, the singularities can be arbitrarily close to the origin, so the radius of convergence can be arbitrarily small. Usually, the singularities of the steady-state probabilities are equal for all states. The probabilities are related by the balance equations. If the RHS of a balance equation has a singularity at a particular place, then the LHS must also have a singularity there. However, the singularities on the RHS may cancel out. For example, it is not unusual that the probability of an empty system has a much simpler form than the other probabilities, without any singularities. Also, the Pollaczek-Khintchine formula shows that the singularities of the steady-state probabilities need not be singularities of the expected queue length.

Example 2.2 This example is based on the properties of the summation

$$\sum_{n=1}^{\infty} n^{-y}. \tag{2.27}$$

This summation converges for all $y \in \mathbb{C}$ such that $\operatorname{Re} y > 1$. The corresponding analytic function has an analytic continuation to $\mathbb{C} \setminus \{1\}$, which is generally known as the Rieman zeta function $\zeta(y)$. In $y = 1$, it has a simple pole with residue 1, so $\lim_{y \to 0} y\zeta(1+y) = 1$. For more details, see for example [134].

Consider the following CTMP on $\mathbb{N}$. From all non-empty states n, the only possible transition is down to state $n-1$, with rate 1. From the empty state, transitions are possible to all non-empty states, with rates

$$\alpha_n(x) = 2^{-n} + x^2 n^{-2-x^2}, \quad \text{for all } n \geq 1.$$

These rates are real and positive for all $x \in \mathbb{R}$. The total departure rate from the empty state is

$$\bar{\alpha}(x) = \sum_{n=1}^{\infty} \alpha_n(x) = \sum_{n=1}^{\infty} 2^{-n} + x^2 \sum_{n=1}^{\infty} n^{-2-x^2} = 1 + x^2 \zeta(2+x^2).$$

Let $p(x,n)$ denote the steady-state probability of state $n \in \mathbb{N}$ for a given value of $x \in \mathbb{R}$. The balance equations of the Markov process are

$$\begin{aligned} \bar{\alpha}(x)\, p(x,0) &= p(x,1), \\ p(x,n) &= \alpha_n(x)\, p(x,0) + p(x,n+1), \quad \text{for all } n \geq 1. \end{aligned}$$

It is easily checked that the normalized solution to these equations is given by

$$\begin{aligned} p(x,0) &= \left[1 + \sum_{k=1}^{\infty} k\, \alpha_k(x)\right]^{-1}, \\ p(x,n) &= p(x,0) \sum_{k=n}^{\infty} \alpha_k(x), \quad \text{for all } n \geq 1. \end{aligned}$$

If $x = 0$, then the probability of the empty state is

$$p(0,0) = \left[1 + \sum_{k=1}^{\infty} k\, 2^{-k}\right]^{-1} = [1+2]^{-1} = \tfrac{1}{3}.$$

On the other hand, if x is positive but small, then

$$\begin{aligned} \lim_{x \downarrow 0} p(x,0) &= \lim_{x \downarrow 0} \left[1 + \sum_{k=1}^{\infty} k\, 2^{-k} + x^2 \sum_{k=1}^{\infty} k^{-1-x^2}\right]^{-1} \\ &= \lim_{x \downarrow 0} [1 + 2 + x^2 \zeta(1+x^2)]^{-1} = [1+2+1]^{-1} = \tfrac{1}{4}. \end{aligned}$$

Therefore, $p(x,0)$ is neither right-continuous nor analytic in x at $x = 0$, even though the transition rates *are* analytic. □

Example 2.3 There are systems for which the power-series expansions of the steady-state probabilities are entire functions. Consider the $M/M/1$ queue with arrival rate γ and service rate μ. The steady-state probabilities of this queue are equal to

$$p(\gamma,n) = (\gamma\mu^{-1})^n (1 - \gamma\mu^{-1}), \quad \text{for all } n \geq 0.$$

These are finite polynomials in γ, so the steady-state probabilities are entire functions of γ, even though the system is only ergodic if $0 \leq \gamma < \mu$.

More generally, for an $M^X/GE_J/1$ queue, it can be shown by studying the balance equations that the steady-state probabilities of the γ-process are finite polynomials in γ [74]. The arrival process M^X has exponential interarrival times and batch arrivals with finite maximal batch size. The GE_J service time distribution is a generalized Erlang distribution, i.e. the convolution of J independent, not necessarily identical, exponential distributions. □

Example 2.4 Consider again the $M/M/1$ model but with at most m customers in the system. The steady-state probabilities are now

$$p(\gamma,n) = \left(\gamma\mu^{-1}\right)^n \frac{1 - \gamma\mu^{-1}}{1 - (\gamma\mu^{-1})^{m+1}}, \quad \text{for all } 0 \leq n \leq m. \tag{2.28}$$

These steady-state probabilities have (removable) singularities where the denominator is zero, that is at $\gamma = \mu \exp\left(\frac{k}{m+1} 2\pi i\right)$. Therefore, the power series converge if $0 \leq \gamma < \mu$, whereas the system is ergodic for all $\gamma \geq 0$. This is quite the opposite of the behaviour of the open system in the previous example. □

Example 2.5 Consider the following single server queue. The service-time distribution is exponential with rate μ. Customers arrive with interarrival times that have independent identical hyper-exponential distributions. The rate is equal to either α_1 or α_2, with probability π_1 and π_2, respectively. Choose the mean interarrival time equal to 1: $\frac{\pi_1}{\alpha_1} + \frac{\pi_2}{\alpha_2} = 1$. At arrival, a customer is either accepted to the queue with probability γ or rejected with probability $(1-\gamma)$. This process is the γ-process of the $H_2/M/1$ queue. The load of the queue is γ/μ. Provided that $\mu > 1$, the queue is stable for all $\gamma \in [0,1]$.

The Laplace-Stieltjes transform of the hyper-exponential interarrival-time distribution is

$$\phi(s) = \frac{\alpha_1\alpha_2 + \bar{\alpha}s}{(\alpha_1 + s)(\alpha_2 + s)},$$

with $\bar{\alpha} = \pi_1\alpha_1 + \pi_2\alpha_2$. In the γ-process, each arriving customer is rejected with probability $(1-\gamma)$. Therefore, the γ-process is also a $G/M/1$ queue, but the interarrival time consists of a geometric number of original interarrival times. By conditioning on this number, the Laplace-Stieltjes transform can be shown to be equal to

$$\psi(s) = \frac{\gamma\phi(s)}{1 - (1-\gamma)\phi(s)}.$$

According to standard $G/M/1$ theory (see section II.3.2 in [41]), the steady-state queue-length distribution is equal to

$$p(\gamma, n) = (1 - \gamma/\mu)\, 1(n = 0) + (\gamma/\mu)\,(1-r)r^{n-1} 1(n > 0), \quad \text{for all } n \geq 0. \tag{2.29}$$

Here, r is the solution in the interval (0,1) of the equation $\psi\left(\mu(1-r)\right) = r$:

$$r = \frac{1}{2\mu}\left[\alpha_1 + \alpha_2 + \mu - (1-\gamma)\bar{\alpha} - \sqrt{(\alpha_1 + \alpha_2 + \mu - (1-\gamma)\bar{\alpha})^2 - 4\gamma(\alpha_1\alpha_2 + \mu\bar{\alpha})}\right].$$

This solution r is a function of γ. The only singularities of this function in γ are the branch points where the root is zero, that is at

$$\gamma_{1,2} = \frac{1}{\bar{\alpha}^2}\left[\sqrt{(\bar{\alpha} - \alpha_1)(\bar{\alpha} - \alpha_2)} \pm \sqrt{\alpha_1\alpha_2 + \mu\bar{\alpha}}\right]^2.$$

Since $\bar{\alpha}$ is in between α_1 and α_2, the first root is the root of a negative number. The second root is the root of a positive number. The absolute value of both branch points is equal to

$$|\gamma_{1,2}| = \frac{-(\bar{\alpha} - \alpha_1)(\bar{\alpha} - \alpha_2) + \alpha_1\alpha_2 + \mu\bar{\alpha}}{\bar{\alpha}^2} = \frac{\mu + \pi_2\alpha_1 + \pi_1\alpha_2}{\pi_1\alpha_1 + \pi_2\alpha_2}.$$

This can be made arbitrarily small. Choose $\alpha_2 = \alpha_1^{-1}$. Then $\pi_1 = \frac{\alpha_1}{\alpha_1+1} = 1 - \pi_2$ and the expression reduces to $(\mu+1)\frac{\alpha_1}{\alpha_1^2-\alpha_1+1}$. This is small if α_1 is large.

The steady-state queue-length probabilities (2.29) are finite polynomials in r, so they all have singularities in γ where r has singularities in γ. The Markov process has only two supplementary states for each queue length. Therefore, for each n, each singularity of r must also be a singularity of at least one of the steady-state probabilities $p(\gamma, n, 1)$ and $p(\gamma, n, 2)$. Singularities close to the origin typically occur when the model is stiff, that is when it has parameters that are of different orders of magnitude like α_1 and α_2 here. □

2.5 Analyticity in the transformation parameter of the PSA

The basic assumption of the PSA is that the steady-state probabilities and performance measures of the γ-process are analytic functions of γ at $\gamma = 0$. The example 2.2 with the zeta function, shows that this assumption is by no means obvious. The theorems in this section show that it *is* justified under certain conditions. For models where γ can be interpreted as a measure of the load of the system, the theorems are light-traffic theorems.

Initially, extrapolation methods like conformal mapping and the epsilon algorithm were not used. Therefore, it was very important that the steady-state probabilities were analytic in γ, since this would imply direct convergence of the power-series expansions. The model in the first paper on the PSA [73] does not have singularities in the unit disk. Therefore, the power-series expansions converge for all $|\gamma| < 1$, that is in both light and heavy traffic. Example 2.5, with singularities arbitrarily close to the origin, shows that this is not generally true. The radius of convergence can be very small. Except for some special models, it does not seem likely that general conditions can be obtained that guarantee analyticity on the entire complex unit disk. Maybe, it is possible to prove analyticity not only at $\gamma = 0$, but for all γ in the real unit interval $[0, 1]$. This would be theoretically interesting and seems reasonable for many models. However, it would not really improve the theoretical foundation of the PSA as a numerical method. The radius of convergence of the power-series expansions is determined by the nearest singularity, irrespective of whether this singularity is real or complex valued. And because only a finite number of coefficients of the expansion around the origin is available, the analytic continuations along the unit interval can not be obtained. Analyticity at $\gamma = 0$ *is* important, because it underlies the derivation of the algorithm. Also, it is a necessary condition for most extrapolation methods. These methods lessen the importance of analyticity in heavy traffic, because even divergent power-series expansions can often be made to converge to the correct value.

Define the vector-functions $q_\ell(\gamma, n)$ as the functions determined by the produced power series for the steady-state probabilities:

$$q_\ell(\gamma, n) \doteq \sum_{r \geq ne} \gamma^r u_\ell(r, n), \quad \text{for all } n \in \mathbb{N}^S. \tag{2.30}$$

Also, define $E_\ell(\gamma)$ as the functions determined by the produced power series for the performance measures:

$$E_\ell(\gamma) \doteq \sum_{r \geq 0} \gamma^r v_\ell(r). \tag{2.31}$$

Theorem 2.1 shows that assumptions 1 and 1' are both sufficient conditions for convergence of the power series (2.30), for γ small enough. Similarly, theorem 2.2 shows that assumption 2 is sufficient for convergence of the power series (2.31).

Convergence does not immediately imply convergence to the right value. Do the produced power series have any connection with the γ-process? By definition, the functions $q_0(\gamma, n)$ satisfy the balance and normalization equations. From this and assumption 3, theorem 2.3 shows that indeed the $q_0(\gamma, n)$ are equal to the steady-state probabilities $p(\gamma, n)$. Together with the absolute convergence of the power-series expansion of $E_0(\gamma)$ this immediately implies that $E_0(\gamma)$ is equal to the steady-state performance measure $\mathbb{E}_\gamma\{\ f(\mathcal{N}, \mathcal{I})\ \}$. So, under assumptions 1 or 1', 2 and 3, both the steady-state distribution and the performance measures are analytic functions of γ.

For the derivatives (of order 1 or higher) things are less clear. It is difficult to justify the reversal of differentiation and summation used to obtain (2.20). Assumptions 1 or 1' and assumption 2 do imply convergence of the obtained power series, but not necessarily to the right value. And convergence around $\gamma = 0$ does not imply that the power series can be extrapolated to obtain correct results at $\gamma = 1$. No characterization of the derivatives exists that can be used to check convergence to the right value. Example 2.2 shows that differentiability of the transition rates does not imply differentiability of the steady-state distribution. If the original process is ergodic for a particular value $\nu = \nu_0$, then it may not even be ergodic on any neighbourhood of ν_0, as can be seen from the third part of example 2.6 on page 65. Nevertheless, if the obtained power series do converge then the right value seems to be the most likely candidate for the limit value. Derivatives have been calculated with success in several applications [27,29].

Define for all $n \in \mathbb{N}^S \backslash \{o\}$ and $\ell \geq 0$:

$$\begin{aligned} c_\ell(n) \doteq & \quad \|B^{(\ell)}(n)B^{-1}(n)\| \quad 1(\ell > 0) \\ & + \sum_{b \in \mathcal{Z}_<} \|A_b^{(\ell)}(n-b)B^{-1}(n)\| + \sum_{b \in \mathcal{Z}_1} \|A_b^{(\ell)}(n-b)B^{-1}(n)\| \\ & + \sum_{b \in \mathcal{Z}_2} \|A_b^{(\ell)}(n-b^-)B^{-1}(n)\| + \sum_{b \in \mathcal{Z}_1} \|A_b^{(\ell)}(n-b^-)B^{-1}(n)\|, \end{aligned}$$

$$\begin{aligned} c_\ell(o) \doteq & \quad \|[o, B^{*(\ell)}(o)][e, B^*(o)]^{-1}\| \quad 1(\ell > 0) \\ & + \sum_{b \in \mathcal{Z}_<} \|[o, A_b^{*(\ell)}(-b)][e, B^*(o)]^{-1}\| \\ & + \sum_{b \in \mathcal{Z}_2} \|[o, A_b^{*(\ell)}(-b^-)][e, B^*(o)]^{-1}\| + \sum_{b \in \mathcal{Z}_1} \|[o, A_b^{*(\ell)}(-b^-)][e, B^*(o)]^{-1}\|. \end{aligned} \tag{2.32}$$

These constants are the norms of the ℓ-th order matrices on the RHS of the recurrence relations (2.21) and (2.22), after multiplication by $B^{-1}(n)$ and $[e, B^*(o)]^{-1}$. The used matrix norm is the maximal-absolute-row-sum norm (1.3). The definition for the empty states is more complicated because the balance equation of the first empty state is replaced by the normalization constraint.

Theorem 2.1 shows that assumption 1 is sufficient to ensure convergence in a neighbourhood of $\gamma = 0$ of the power series produced by the PSA for steady-state probabilities and their derivatives with respect to the system parameter ν (see page 26), up to order L:

Assumption 1 *The transition rates are L times differentiable and*

$$\sup_{n \in \mathbb{N}^S} c_\ell(n) < \infty, \quad \textit{for all } 0 \leq \ell \leq L.$$

For the interpretation of this assumption, consider the case without supplementary space ($I = 1$). Then all matrices on the RHS of (2.32) are positive scalars. The norm of a positive scalar is the scalar itself. For $\ell = 0$, the $c_0(n)$ are sums of terms $A_b(.)B^{-1}(n)$. The $A_b(.)$ correspond to transition rates into state n, whereas $B^{-1}(n)$ is the inverse of the transition rate out of state n. Multiplied by the appropriate steady-state probabilities and powers of γ, the balance equations of the γ-process state that the rate-in is equal to the rate-out (see (2.6)). Assumption 1 requires that without these multiplications the total rate-in is not too large compared to the rate-out, so it is related to a stability condition. For $\ell > 0$, the assumption requires that the derivatives of both the rate-in and the rate-out are not too large, compared to the rate-out.

A sufficient condition for assumption 1 to be true is that the state-dependence of the balance equations is limited. If there is only a finite number of essentially different balance equations, there is also only a finite number of different values of $c_\ell(n)$. Then the supremum is the maximum of these values, which is necessarily finite because the original Markov process is non-instantaneous. More formally, this sufficient condition is described in the next assumption:

Assumption 1' *The transition rates are L times differentiable and a finite set $\mathcal{S} \subset \mathbb{N}^S$ and function $f : \mathbb{N}^S \to \mathcal{S}$ exist such that*

$$A_b(n) = A_b(f(n)), \quad \textit{for all } n, n+b \in \mathbb{N}^S.$$

The assumption requires that the possible transitions from any state $n \notin \mathcal{S}$ are identical to the transitions from state $f(n) \in \mathcal{S}$. In that case, the number of essentially different balance equations is at most the number of states in $\mathcal{S}$. It is easily shown that assumption 1' implies assumption 1:

$$\sup_{n \in \mathbb{N}^S} c_\ell(n) = \sup_{n \in \mathbb{N}^S} c_\ell(f(n)) = \sup_{n \in \mathcal{S}} c_\ell(n) < \infty.$$

Assumption 1' is more restrictive than assumption 1. For example, it excludes all models that are not uniformizable. However, it is readily checked and satisfied in most queueing models. For example, consider a network of S queues where the arrival and routing process do not depend on the queue lengths and the service process at a particular queue only depends on whether that queue is empty or not (like in example 2.1). Then assumption 1' is satisfied, with $\mathcal{S} = \{0,1\}^S$ and $f_s(n) = \min\{1, n_s\}$ for all $1 \leq s \leq S$. Similar models with a finite number m_s of servers at queue s are also included. In that case $\mathcal{S} = \prod_{s=1}^{S} \{0, \ldots, m_s\}$ and $f_s(n) = \min\{m_s, n_s\}$ for all $1 \leq s \leq S$. Models with infinitely many servers do not satisfy assumption 1' but often *do* satisfy assumption 1, because $A_b(n-b)$, $A_b(n-b^-)$ and $B(n)$ usually grow about equally fast in n for these models. The model in the third part of example 2.6, to be presented on page 65, satisfies neither assumption 1' nor assumption 1.

In theorem 2.1, lemma 2.2 will be used to find a geometric bound on the coefficients, which proves analyticity for small γ. The lemma shows that a solution to a certain set of inequalities is bounded by any solution to the same set but with reversed inequalities.

Lemma 2.2 *For all $i \geq 0$, let $g_i : \mathbb{R}^i \to \mathbb{R}$ be a non-decreasing function. If*

$$\begin{aligned} x_i &\leq g_i(x_0, \ldots, x_{i-1}), \quad \textit{for all } i \geq 1, \\ y_i &\geq g_i(y_0, \ldots, y_{i-1}), \quad \textit{for all } i \geq 1 \end{aligned}$$

and $x_0 \leq y_0$, then

$$x_i \leq y_i, \quad \textit{for all } i \geq 0.$$

Proof: Suppose that $x_i \leq y_i$, for some $j \geq 0$ and all $0 \leq i \leq j$. This is true for $j = 0$. Then,

$$x_{j+1} \leq g_{j+1}(x_0, \ldots, x_j) \leq g_{j+1}(y_0, \ldots, y_j) \leq y_{j+1}.$$

Therefore, $x_i \leq y_i$ for all $0 \leq i \leq j+1$ and by induction for all $i \geq 0$. □

Theorem 2.1 *Under assumptions 0 and 1, and in a neighbourhood of $\gamma = 0$, the functions $q_\ell(\gamma, n)$ are analytic in γ for all $n \in \mathbb{N}^S$, $0 \le \ell \le L$.*

Proof: Scalar sequences $\bar{u}_\ell(r,n)$ will be obtained, for all $n \in \mathbb{N}^S$ and $0 \le \ell \le L$, such that

$$\|u_\ell(r,n)\| \le \bar{u}_\ell(r,n), \quad \text{for all } r \ge ne, \tag{2.33}$$

and such that the series

$$\sum_{r \ge ne} \gamma^r \bar{u}_\ell(r,n) \tag{2.34}$$

converge in a neighbourhood of $\gamma = 0$. If such convergent majorants exist, the power series produced by the algorithm for steady-state probabilities and their derivatives are absolutely convergent and analytic in the convergence region of the majorant (see theorem B.1 and the definition of analyticity on page 112).

Define

$$c_\ell \doteq \sup_{n \in \mathbb{N}^S} c_\ell(n), \quad \text{for all } 0 \le \ell \le L.$$

By assumption 1, the values of all c_ℓ are finite. Multiplying equation (2.21) by $B^{-1}(n)$, taking norms and using the triangular inequality and submultiplicativity leads to

$$\begin{aligned}
\|u_\ell(r,n)\| \le & \sum_{k=0}^{\ell-1} \binom{\ell}{k} \|u_k(r,n)\| \|B^{(\ell-k)}(n)B^{-1}(n)\| \\
+ & \sum_{b \in \mathcal{Z}_<} \sum_{k=0}^{\ell} \binom{\ell}{k} \|u_k(r,n-b)\| \|A_b^{(\ell-k)}(n-b)B^{-1}(n)\| \\
+ & \sum_{b \in \mathcal{Z}_1} \sum_{k=0}^{\ell} \binom{\ell}{k} \|u_k(r-r_b,n-b)\| \|A_b^{(\ell-k)}(n-b)B^{-1}(n)\| \\
+ & \sum_{b \in \mathcal{Z}_2} \sum_{k=0}^{\ell} \binom{\ell}{k} \|u_k(r,n-b^-)\| \|A_b^{(\ell-k)}(n-b^-)B^{-1}(n)\| \\
+ & \sum_{b \in \mathcal{Z}_1} \sum_{k=0}^{\ell} \binom{\ell}{k} \|u_k(r-r_b,n-b^-)\| \|A_b^{(\ell-k)}(n-b^-)B^{-1}(n)\| \\
\le & \left\{ \sum_{k=0}^{\ell} \binom{\ell}{k} c_\ell \right\} \max\{ \max_{0 \le k \le \ell-1} \|u_k(r,n)\|, \\
& \max_{b \in \mathcal{Z}_<,\ 0 \le k \le \ell} \|u_k(r,n-b)\|, \\
& \max_{b \in \mathcal{Z}_1,\ 0 \le k \le \ell} \|u_k(r-r_b,n-b)\|, \\
& \max_{b \in \mathcal{Z}_2,\ 0 \le k \le \ell} \|u_k(r,n-b^-)\|, \\
& \max_{b \in \mathcal{Z}_1,\ 0 \le k \le \ell} \|u_k(r-r_b,n-b^-)\| \},
\end{aligned}$$

for $n \in \mathbb{N}^S \setminus \{o\}$, $r \ge ne$. The second inequality follows from replacing all the norms of coefficients by the maximum of all these norms and from the definition of the $c_\ell(n)$ and c_ℓ. The recurrence relations for the empty states (2.22) can be written in a single

matrix equation:

$$
\begin{aligned}
u_\ell(r,o)[e,B^*(o)] \;=\; & -\sum_{k=0}^{\ell-1}\binom{\ell}{k}\; u_k(r,o) && [o,B^{*(\ell-k)}(o)] \\
& +\sum_{b\in\mathcal{Z}_<}\sum_{k=0}^{\ell}\binom{\ell}{k}\; u_k(r,-b) && [o,A_b^{*(\ell-k)}(-b)] \\
& +\sum_{b\in\mathcal{Z}_2}\sum_{k=0}^{\ell}\binom{\ell}{k}\; u_k(r,-b^-) && [o,A_b^{*(\ell-k)}(-b^-)] \\
& -\sum_{b\in\mathcal{Z}_1}\sum_{k=0}^{\ell}\binom{\ell}{k}\; u_k(r-r_b,-b^-) && [o,A_b^{*(\ell-k)}(-b^-)] \\
& -\sum_{0<ne\le r} u_\ell(r,n) && [e,O^*],
\end{aligned}
\tag{2.35}
$$

for $r>0$. The matrix O^* is an I by $(I-1)$ matrix of zeros. Multiplying by $[e,B^*(o)]^{-1}$ and taking norms renders

$$
\begin{aligned}
\|u_\ell(r,o)\| \le & \left\{\sum_{k=0}^{\ell}\binom{\ell}{k}c_\ell\right\}\max\{\max_{0\le k\le \ell-1}\|u_k(r,o)\|, \\
& \max_{b\in\mathcal{Z}_<,\ 0\le k\le\ell}\|u_k(r,-b)\|, \\
& \max_{b\in\mathcal{Z}_2,\ 0\le k\le\ell}\|u_k(r,-b^-)\|, \\
& \max_{b\in\mathcal{Z}_1,\ 0\le k\le\ell}\|u_k(r-r_b,n-b^-)\|\ \} \\
+ & \|[e,O^*][e,B^*(o)]^{-1}\|\sum_{0<ne\le r}\|u_\ell(r,n)\|,
\end{aligned}
\tag{2.36}
$$

for $r>0$. Here, it is used that $\|[o,A^*]\|\le\|A\|$, for any matrix A. Let the numbers $\bar{u}_\ell(r,n)$ be such that they satisfy the reversed inequalities:

$$
\begin{aligned}
\bar{u}_\ell(r,n) \ge & \left\{\sum_{k=0}^{\ell}\binom{\ell}{k}c_k\right\}\max\{\max_{0\le k\le \ell-1}\bar{u}_k(r,n), \\
& \max_{b\in\mathcal{Z}_<,\ 0\le k\le\ell}\bar{u}_k(r,n-b), \\
& \max_{b\in\mathcal{Z}_1,\ 0\le k\le\ell}\bar{u}_k(r-r_b,n-b), \\
& \max_{b\in\mathcal{Z}_2,\ 0\le k\le\ell}\bar{u}_k(r,n-b^-), \\
& \max_{b\in\mathcal{Z}_1,\ 0\le k\le\ell}\bar{u}_k(r-r_b,n-b^-)\ \}, \\
\bar{u}_\ell(r,o) \ge & \left\{\sum_{k=0}^{\ell}\binom{\ell}{k}c_k\right\}\max\{\max_{0\le k\le \ell-1}\bar{u}_k(r,o), \\
& \max_{b\in\mathcal{Z}_<,\ 0\le k\le\ell}\bar{u}_k(r,-b), \\
& \max_{b\in\mathcal{Z}_2,\ 0\le k\le\ell}\bar{u}_k(r,-b^-), \\
& \max_{b\in\mathcal{Z}_1,\ 0\le k\le\ell}\bar{u}_k(r-r_b,-b^-)\ \} \\
+ & \|[e,O^*][e,B^*(o)]^{-1}\|\sum_{0<ne\le r}\bar{u}_\ell(r,n), \\
\bar{u}_\ell(0,o) \ge & \|u_\ell(0,o)\|,
\end{aligned}
\tag{2.37}
$$

for $r\ge ne\ge 1$ and $0\le\ell\le L$. Because all coefficients are calculated sequentially, lemma 2.2 then shows that

$$\|u_\ell(r,n)\|\le\bar{u}_\ell(r,n),\quad \text{for all } 0\le ne\le r.$$

The three inequalities (2.37) are indeed satisfied by the sequences

$$\bar{u}_\ell(r,n) = C_{0,\ell}\, C_1^{\ell-ne}\, C_2^{1(n=o)}\, C_3^r, \tag{2.38}$$

for $n \in \mathbb{N}^S$, $r \geq ne$, and with

$$\begin{aligned}
C_{0,0} &\doteq \|u_0(0,o)\| = 1, \\
C_{0,\ell} &\doteq \max\{\, \|u_\ell(0,o)\|,\ C_{0,\ell-1}\,\}, \quad \text{for all } 1 \leq \ell \leq L, \\
C_1 &> \max\left\{ \sum_{k=0}^{L} \binom{L}{k} c_k,\ 1 \right\}, \\
C_2 &\doteq 1 + \|[e, O^*][e, B^*(o)]^{-1}\| \left(1 - C_1^{-1}\right)^{-S}, \\
C_3 &\doteq C_1^2 C_2.
\end{aligned} \tag{2.39}$$

That $C_{0,0} = \|u_0(0,o)\| = 1$, is because $u_0(0,o)$ is a distribution, as can be seen from comparing (2.14) and (2.15). The constants $C_{0,\ell}$ are non-decreasing in ℓ. The bounds $\bar{u}_\ell(r,n)$ are the product of three factors. The first factor $C_{0,\ell}C_1^\ell$ only depends on ℓ. It is non-decreasing in ℓ, so the maxima in the first and second inequality of (2.37) are all attained for the largest value of k. The second factor $C_1^{-ne}\ C_2^{1(n=o)}$ only depends on n and the third factor C_3^r only depends on r.

The first inequality in (2.37) holds because the following five inequalities hold:

$$\begin{aligned}
\left\{ \sum_{k=0}^{\ell} \binom{\ell}{k} c_k \right\} \max_{0 \leq k \leq \ell-1} \bar{u}_k(r,n) &\leq C_1\, \bar{u}_{\ell-1}(r,n) \\
&\leq \bar{u}_\ell(r,n), \\
\left\{ \sum_{k=0}^{\ell} \binom{\ell}{k} c_k \right\} \max_{b \in \mathcal{Z}_<,\ 0 \leq k \leq \ell} \bar{u}_k(r, n-b) &\leq C_1 \max_{be<0} \bar{u}_\ell(r, n-b) \\
&= \bar{u}_\ell(r,n) \max_{be<0} C_1^{1+be} \\
&= \bar{u}_\ell(r,n), \\
\left\{ \sum_{k=0}^{\ell} \binom{\ell}{k} c_k \right\} \max_{b \in \mathcal{Z}_1,\ 0 \leq k \leq \ell} \bar{u}_k(r-r_b, n-b) &\leq C_1 \max_{be \geq 0} \bar{u}_\ell(r-r_b, n-b) \\
&= \bar{u}_\ell(r,n) \max_{be \geq 0} C_1^{1+be} C_2^{1(n=b)} C_3^{-r_b} \\
&\leq \bar{u}_\ell(r,n) \max_{be \geq 0} (C_1^2 C_2)^{r_b} C_3^{-r_b} \\
&= \bar{u}_\ell(r,n), \\
\left\{ \sum_{k=0}^{\ell} \binom{\ell}{k} c_k \right\} \max_{b \in \mathcal{Z}_2,\ 0 \leq k \leq \ell} \bar{u}_k(r, n-b^-) &\leq C_1 \max_{be \geq 0,\ b \notin \mathbb{N}^S} \bar{u}_\ell(r, n-b^-) \\
&= \bar{u}_\ell(r,n) \max_{be \geq 0,\ b \notin \mathbb{N}^S} C_1^{1+b^-e} \\
&= \bar{u}_\ell(r,n), \\
\left\{ \sum_{k=0}^{\ell} \binom{\ell}{k} c_k \right\} \max_{b \in \mathcal{Z}_1,\ 0 \leq k \leq \ell} \bar{u}_k(r-r_b, n-b^-) &\leq C_1 \max_{be \geq 0} \bar{u}_\ell(r-r_b, n-b^-) \\
&= \bar{u}_\ell(r,n) \max_{be \geq 0} C_1^{1+b^-e} C_3^{-r_b} \\
&\leq \bar{u}_\ell(r,n),
\end{aligned}$$

for all $r \geq ne \geq 1$.

The second inequality in (2.37) is checked in two steps. In a similar way as for the first inequality, it can be shown that

$$\left\{\sum_{k=0}^{\ell}\binom{\ell}{k}c_k\right\}\left\{\begin{array}{ll}\max\limits_{0\leq k\leq \ell-1} & \bar{u}_k(r,o),\\ \max\limits_{b\in Z_<,\ 0\leq k\leq \ell} & \bar{u}_k(r,-b),\\ \max\limits_{b\in Z_2,\ 0\leq k\leq \ell} & \bar{u}_k(r,-b^-),\\ \max\limits_{b\in Z_1,\ 0\leq k\leq \ell} & \bar{u}_k(r-r_b,-b^-)\end{array}\right\} \leq C_{0,\ell}C_1^{\ell}C_3^{r}\,. \tag{2.40}$$

Also, it is true that

$$\begin{aligned}\|[e,O^*][e,B^*(o)]^{-1}\| \sum_{0<ne\leq r}\bar{u}_\ell(r,n) &\leq \|[e,O^*][e,B^*(o)]^{-1}\|\ C_{0,\ell}C_1^{\ell}C_3^{r}\sum_{n\in\mathbb{N}^S}C_1^{-ne}\\ &= \|[e,O^*][e,B^*(o)]^{-1}\|\ C_{0,\ell}C_1^{\ell}C_3^{r}\left(1-C_1^{-1}\right)^{-S}.\end{aligned} \tag{2.41}$$

Here, it is used that $C_1 > 1$. Together, (2.40) and (2.41) prove the second inequality in (2.37), because the sum of the RHSs of (2.40) and (2.41) is equal to $\bar{u}_\ell(r,o)$:

$$C_{0,\ell}C_1^{\ell}C_3^{r} + \|[e,O^*][e,B^*(o)]^{-1}\|\ C_{0,\ell}C_1^{\ell}C_3^{r}\left(1-C_1^{-1}\right)^{-S} = C_{0,\ell}C_1^{\ell}C_2C_3^{r} = \bar{u}_\ell(r,o),$$

for all $r \geq 1$.

Finally, the third inequality in (2.37) holds because

$$\|u_\ell(0,o)\| \leq C_{0,\ell} \leq C_{0,\ell}C_1^{\ell}C_2 = \bar{u}_\ell(0,o).$$

Therefore, the solution (2.38) satisfies all inequalities in (2.37) and is a majorant as claimed in (2.33).

What remains to be proved is that the solution (2.38) converges in a neighbourhood of $\gamma = 0$, as claimed in (2.34). This is indeed true, because the solution is geometric in r and converges if $|\gamma| < C_3^{-1}$. Therefore, $\bar{u}_\ell(r,n)$ is a convergent majorant, which proves the theorem. □

The lower bound on the radius of convergence C_3^{-1} is uniform for all $n \in \mathbb{N}^S$. Bounds on the truncation error can be constructed from the bounds on the coefficients:

$$\begin{aligned}\left\|\sum_{r\geq R+1}\gamma^r u_\ell(r,n)\right\| &\leq \sum_{r\geq R+1}\gamma^r\bar{u}_\ell(r,n)\\ &\leq \sum_{r\geq R+1}\gamma^r C_{0,\ell}C_1^{\ell-ne}\ C_2^{1(n=o)}\ C_3^{r}\\ &= C_{0,\ell}C_1^{\ell-ne}\ C_2^{1(n=o)}\ \frac{(\gamma C_3)^{R+1}}{1-\gamma C_3},\end{aligned}$$

for $|\gamma| < C_3^{-1}$. Unfortunately, the bound C_3^{-1} is usually very small. It can be slightly improved by making it depend on ℓ instead of on L. The majorant (2.38) is valid for all

fixed $L \geq 0$. So, in the definition of C_1 in (2.39), L can be replaced by ℓ. Then C_1 and C_3 are both increasing functions of ℓ and independent of L.

Even with this improvement, the lower bound on the radius of convergence is too small to have any real practical importance. For the case without supplementary space, it will be shown that the lower bound can be at most 0.113401. This will be done by maximizing C_3^{-1} over the possible values of $S \geq 1$ and $C_1 > 1$. If $I = 1$, then $C_2 = 1 + \left(1 - C_1^{-1}\right)^{-S}$. This is increasing in S, so $C_3^{-1} = [C_1^2 C_2]^{-1}$ is decreasing in S for all fixed C_1. At $S = 1$, C_3^{-1} is equal to $(C_1 - 1)\,(2C_1 - 1)^{-1}\,C_1^{-2}$. This is maximal at $C_1 = \left[7 + \sqrt{17}\right]/8 \approx 1.39039$, with value 0.113401. So, even if the steady-state probabilities are entire functions of γ, like for the $M/M/1$ and $M/M/\infty$ queues, the obtained lower bound on the radius of convergence is very small. This unfavourable behaviour is caused by the fact that C_1 needs to be at least 1, for C_2 to be well defined in (2.39).

Convergence of the power series for the steady-state probabilities with their derivatives does not immediately imply convergence for the performance measures and their derivatives. For this, the additional assumption 2 on the function $f : \Omega \to \mathbb{R}$ from section 2.3.2 is sufficient:

Assumption 2 *The function f is L times differentiable and finite constants $F_\ell, G_\ell \geq 0$ exist such that*

$$f^{(\ell)}(n,i) \leq F_\ell\, G_\ell^{ne}, \quad \textit{for all } (n,i) \in \Omega \textit{ and } 0 \leq \ell \leq L.$$

The assumption requires that the function $f(n,i)$ and its derivatives grow at most exponentially in n, which is a weak condition. The moments and covariances in the examples (2.18) on page 26 are all bounded by a polynomial in n, so the assumption is satisfied for $\ell = 0$. Since these examples do not depend on any system parameter ν at all, the derivatives are all identically zero. Therefore, assumption 2 is satisfied for all $\ell \geq 0$.

In theorem 2.2, analyticity of the power-series expansions for performance measures and their derivatives (2.31) is shown by proving absolute convergence (see theorem B.1), using the bounds for the steady-state probabilities found in theorem 2.1.

Theorem 2.2 *Under assumptions 0, 1 and 2, and in a neighbourhood of $\gamma = 0$, the functions $E_\ell(\gamma)$ are analytic in γ for all $0 \leq \ell \leq L$.*

Proof: These functions are analytic because their power-series expansions (2.31) with

coefficients defined by (2.25) are absolutely convergent:

$$\begin{aligned}
\sum_{r\geq 0}\gamma^r\,|v_\ell(r)| &= \sum_{r\geq 0}\gamma^r\left|\sum_{ne\leq r}\sum_{k=0}^{\ell}\binom{\ell}{k}u_k(r,n)\;f^{(\ell-k)}(n)\right| \\
&\leq \sum_{r\geq 0}\gamma^r\sum_{ne\leq r}\sum_{k=0}^{\ell}\binom{\ell}{k}\|u_k(r,n)\|\;\|f^{(\ell-k)}(n)\| \\
&\leq \sum_{r\geq 0}\gamma^r\sum_{ne\leq r}\sum_{k=0}^{\ell}\binom{\ell}{k}\;C_{0,k}C_1^{k-ne}C_2C_3^r\;F_{\ell-k}G_{\ell-k}^{ne} \\
&= \sum_{k=0}^{\ell}\binom{\ell}{k}C_{0,k}C_1^kC_2F_{\ell-k}\sum_{n\in\mathbb{N}^S}\left(C_1^{-1}G_{\ell-k}\right)^{ne}\sum_{r\geq ne}(\gamma C_3)^r \\
&= \sum_{k=0}^{\ell}\binom{\ell}{k}\frac{C_{0,k}C_1^kC_2F_{\ell-k}}{1-\gamma C_3}\sum_{n\in\mathbb{N}^S}\left(\gamma C_1^{-1}C_3G_{\ell-k}\right)^{ne} \\
&= \sum_{k=0}^{\ell}\binom{\ell}{k}\frac{C_{0,k}C_1^kC_2F_{\ell-k}}{1-\gamma C_3}\left(1-\gamma C_1^{-1}C_3G_{\ell-k}\right)^{-S}<\infty,
\end{aligned}$$

if $|\gamma|<C_3^{-1}$ and $|\gamma|<C_1(C_3G_\ell)^{-1}$, for all $0\leq\ell\leq L$. The first inequality follows from the triangular inequality and submultiplicativity. In the second inequality, assumption 2 and bound (2.38) are used. □

The functions $q_0(\gamma,n)$ would be the desired steady-state probabilities $p(\gamma,n)$ if they satisfy the balance equations of the γ-process and if the γ-process is ergodic. They satisfy the balance equations if the reversals of the order of summation were justified in the derivation of the PSA in section 2.3. The ergodicity of the γ-process will be obvious in many applications. The γ-process is such that upward transitions in the original process are replaced by downward transitions or selfloops. So, if the original process is ergodic, then the γ-process will usually also be ergodic for all γ in [0,1]. However, this is not true in general as will be illustrated by the second part of example 2.6 on page 65. Both for the reversal of the order of summations and for the ergodicity for small values of γ, the following additional assumption is required:

Assumption 3 *Finite constants $F,G\geq 0$ exist such that*

$$\bar{\alpha}(n,i)\leq F\;G^{ne},\quad \textit{for all } (n,i)\in\Omega.$$

This assumption requires that in the original process, and therefore in each γ-process, the total transition rate out of state (n,i) grows at most exponentially in n. Since the growth rate G is allowed to be arbitrarily large, this assumption is very weak. For example, it is much weaker than assumption 1' or the assumption that the Markov process is uniformizable. Theorem 2.3 will show that assumption 3 is sufficient to ensure that indeed $q_0(\gamma,n)=p(\gamma,n)$, for all $n\in\mathbb{N}^S$ and for γ small enough. In the technical proof of the theorem, the usual ergodicity theorems can not be used because it is not known in advance that the functions $q_0(\gamma,n)$ are non-negative and because the process need not be

uniformizable. The theorems A.1 and A.2 in appendix A do not make these assumptions.

Theorem 2.3 *Under assumptions 0, 1 and 3, and in a neighbourhood of $\gamma = 0$, the γ-process is ergodic with steady-state probabilities $q_{0i}(\gamma, n)$ for all $(n, i) \in \Omega$.*

Proof: The case $\gamma = 0$ will be considered separately. Assumption 0 ensures that the process will always end up in the single finite recurrent class of empty states. Comparison of (2.14) and (2.15) shows that the steady-state probabilities are equal to $q_{0i}(0, n) = u_{0i}(0, o)1(n = o)$, for $(n, i) \in \Omega$. Therefore, the theorem holds for $\gamma = 0$. For $\gamma > 0$ but sufficiently small, it will be checked in the rest of the proof whether the γ-process and the functions $q_{0i}(\gamma, n)$ satisfy the conditions (A.4), (A.5) and (A.6) of theorem A.2 in appendix A.

The 1-process was assumed to be irreducible and non-instantaneous. For $\gamma > 0$, each transition in the 1-process is also possible in the γ-process, so the γ-process is also irreducible. The total transition rate in the γ-process from each state is constant in γ (or non-decreasing in γ if selfloops are ignored) and therefore the γ-process is also non-instantaneous for all $\gamma \in (0, 1]$. As a consequence of assumption 3, the following inequality holds:

$$\sum_{b \in \mathcal{Z}} \|A_b(n)\| \;\leq\; \max_{1 \leq i \leq I} \bar{\alpha}(n, i) \;\leq\; F\, G^{ne}, \tag{2.42}$$

for any $\mathcal{Z} \subseteq \mathbb{Z}^S$ and $n \in \mathbb{N}^S$.

That the balance equations (A.4) are satisfied can be shown as follows. Rearranging equations (2.12) for $n \neq o$, renders

$$\begin{aligned} u_0(r, n)\bar{A}(n) \;=\; & \sum_{b \in \mathcal{Z}_<} u_0(r, n - b) \quad A_b(n - b) \\ + & \sum_{b \in \mathcal{Z}_\geq} u_0(r - r_b, n - b) \quad A_b(n - b) \\ + & \sum_{b \in \mathcal{Z}_\geq} u_0(r, n - b^-) \quad A_b(n - b^-) \\ - & \sum_{b \in \mathcal{Z}_\geq} u_0(r - r_b, n - b^-) \quad A_b(n - b^-), \end{aligned}$$

for $r \geq ne$. Multiplying both sides by γ^r, summing over $r \geq ne$ and changing the order of the summations on the RHS renders

$$\begin{aligned} \sum_{r \geq ne} \gamma^r u_0(r, n)\bar{A}(n) \;=\; & \sum_{b \in \mathcal{Z}_<} \sum_{r \geq ne} \gamma^r \; u_0(r, n - b) \quad A_b(n - b) \\ + & \sum_{b \in \mathcal{Z}_\geq} \gamma^{r_b} \sum_{r \geq ne - r_b} \gamma^r \; u_0(r, n - b) \quad A_b(n - b) \\ + & \sum_{b \in \mathcal{Z}_\geq} \sum_{r \geq ne} \gamma^r \; u_0(r, n - b^-) \quad A_b(n - b^-) \\ - & \sum_{b \in \mathcal{Z}_\geq} \gamma^{r_b} \sum_{r \geq ne - r_b} \gamma^r \; u_0(r, n - b^-) \quad A_b(n - b^-). \end{aligned} \tag{2.43}$$

Since $u_0(r,n) = 0$ for all $r < ne$, this is equivalent to

$$\begin{aligned} q_0(\gamma,n)\bar{A}(n) &= \sum_{b\in\mathcal{Z}_<} \qquad\qquad q_0(\gamma,n-b) \quad A_b(n-b) \\ &+ \sum_{b\in\mathcal{Z}_\geq} \quad \gamma^{r_b} \quad q_0(\gamma,n-b) \quad A_b(n-b) \\ &+ \sum_{b\in\mathcal{Z}_\geq} \quad (\,1-\gamma^{r_b}\,) \quad q_0(\gamma,n-b^-) \quad A_b(n-b^-). \end{aligned}$$

This coincides with the balance equations of the γ-process (2.6) for the non-empty states. The only operation in the derivation above that may not be allowed is the reversal of the order of summation (just as in the reverse derivation that leads from (2.9) to (2.12)). Changing the order of summations is justified when assumption 2 is satisfied, because then the four individual terms on the RHS of (2.43) are all absolutely convergent for γ small enough (see theorem B.2 in appendix B). In the first term, the order of the summations over r and b can be reversed because:

$$\begin{aligned} &\sum_{b\in\mathcal{Z}_<}\sum_{r\geq ne} \gamma^r \|u_0(r,n-b)\| \; \|A_b(n-b)\| \\ &\leq \sum_{b\in\mathbb{Z}^S}\sum_{r\geq ne-be} \gamma^r \; C_1^{-ne+be} C_2 C_3^r \; FG^{ne-be} \; 1(n-b\in\mathbb{N}^S) \\ &= \tfrac{C_2F}{1-\gamma C_3} \sum_{b\in\mathbb{Z}^S} \left(\gamma\, C_1^{-1}C_3G\right)^{(n-b)e} 1(n-b\in\mathbb{N}^S) \\ &= \tfrac{C_2F}{1-\gamma C_3} \sum_{n\in\mathbb{N}^S} \left(\gamma C_1^{-1}C_3G\right)^{ne} \\ &= \tfrac{C_2F}{1-\gamma C_3} \left(1-\gamma C_1^{-1}C_3G\right)^{-S}, \end{aligned}$$

if $|\gamma| < C_3^{-1}$ and $|\gamma| < C_1(C_3G)^{-1}$. The bound (2.38) was used and also inequality (2.42) with $\mathcal{Z} = \{b\}$. In the second term of (2.43) the order of summations can be reversed because the summation over b has a finite number of terms. There is only a finite number of upward transitions that can lead to state n:

$$\#\left\{\, b\in\mathcal{Z}_\geq \;\middle|\; n-b\in\mathbb{N}^S \,\right\} = \binom{ne+S}{S}.$$

In the third term of (2.43) the order of summations can also be reversed. This will be shown by conditioning on the value of b^-:

$$\begin{aligned} &\sum_{b\in\mathcal{Z}_\geq}\sum_{r\geq ne} \gamma^r \|u_0(r,n-b^-)\| \; \|A_b(n-b^-)\| \\ &= \sum_{d\in-\mathbb{N}^S}\sum_{\substack{b\in\mathcal{Z}_\geq \\ b^-=d}}\sum_{r\geq ne} \gamma^r \|u_0(r,n-b^-)\| \; \|A_b(n-b^-)\| \\ &\leq \sum_{d\in-\mathbb{N}^S} \left\{ \sum_{\substack{b\in\mathcal{Z}_\geq \\ b^-=d}} \|A_b(n-d)\| \right\} \sum_{r\geq ne-de} \gamma^r \|u_0(r,n-d)\| \\ &\leq \sum_{d\in-\mathbb{N}^S} FG^{ne-de} \sum_{r\geq ne-de} \gamma^r C_1^{-ne+de} C_2 C_3^r \\ &= \tfrac{FC_2}{1-\gamma C_3} \sum_{d\in-\mathbb{N}^S} \left(\gamma C_1^{-1}C_3G\right)^{ne-de} \\ &= \tfrac{FC_2}{1-\gamma C_3} \left(\gamma C_1^{-1}C_3G\right)^{ne} \left(1-\gamma C_1^{-1}C_3G\,\right)^{-S}, \end{aligned}$$

if $|\gamma| < C_3^{-1}$ and $|\gamma| < C_1(C_3G)^{-1}$. Here, inequality (2.42) was used with n replaced by $n-d$ (which is fixed) and with $\mathcal{Z} = \{\ b \in \mathcal{Z}_{\geq} \mid b^- = d\ \}$. That in the fourth term of (2.43) the order of summations can be reversed can be shown in a similar way as for the third term. A similar approach applied to equations (2.13) and (2.14) also leads to the balance equations of all empty states, except state $(o,1)$. This completes the part of the proof that shows that condition (A.4) holds, with the omitted state denoted by k equal to state $(o,1)$.

That the normalization equation (A.5) is satisfied can be shown as follows. Rearranging the normalization parts of equations (2.13) and (2.14) renders

$$\sum_{0\leq ne\leq r} u_0(r,n)e = 1(r=0),$$

for $r \geq 0$. Multiplying both sides by γ^r, summing over $r \geq 0$ and changing the order of the summations over r and n renders

$$\sum_{n\in\mathbb{N}^S}\sum_{r\geq ne} \gamma^r u_0(r,n)e = \sum_{(n,i)\in\Omega} q_{0i}(\gamma,n) = 1.$$

Here, the order of summations can be reversed because again the power series are absolutely convergent (see theorem B.2 in appendix B):

$$\begin{aligned}\sum_{n\in\mathbb{N}^S}\sum_{r\geq ne} \gamma^r |u_0(r,n)e| &\leq \sum_{n\in\mathbb{N}^S}\sum_{r\geq ne} \gamma^r C_1^{-ne} C_2 C_3^r \\ &= \tfrac{C_2}{1-\gamma C_3} \sum_{n\in\mathbb{N}^S} \left(\gamma C_1^{-1} C_3\right)^{ne} \\ &= \tfrac{C_2}{1-\gamma C_3} \left(\ 1-\gamma C_1^{-1} C_3\ \right)^{-S},\end{aligned} \tag{2.44}$$

if $|\gamma| < C_3^{-1}$ and $|\gamma| < C_1 C_3^{-1}$.

Finally, condition (A.6) is satisfied because

$$\begin{aligned}\sum_{(n,i)\in\Omega} |q_{0i}(\gamma,n)|\bar{\alpha}(n,i) &= \sum_{n\in\mathbb{N}^S} \|q_0(\gamma,n)\|\bar{\alpha}(n,i) \\ &\leq \sum_{n\in\mathbb{N}^S}\sum_{r\geq ne} \gamma^r\ C_1^{-ne} C_2 C_3^r\ FG^{ne} \\ &= \tfrac{FC_2}{1-\gamma C_3} \sum_{n\in\mathbb{N}^S} \left(\gamma C_1^{-1} C_3 G\right)^{ne} \\ &= \tfrac{FC_2}{1-\gamma C_3} \left(\ 1-\gamma C_1^{-1} C_3 G\ \right)^{-S} < \infty,\end{aligned}$$

if $|\gamma| < C_3^{-1}$ and $|\gamma| < C_1(C_3G)^{-1}$. With this it has been shown that the γ-process and the functions $q_{0i}(\gamma,n)$ satisfy all conditions of theorem A.2 in appendix A, which finishes the proof of theorem 2.3. □

2.6 Extrapolation methods

Even if all assumptions of the previous section are satisfied, the PSA may still not work. All that is guaranteed by the assumptions, is that the algorithm is well defined and that

the power-series converge for small enough values of γ. For larger γ, particularly at $\gamma = 1$, the power-series may converge very slowly or diverge, as was shown in examples 2.4 and 2.5. If convergence is slow, then many coefficients need to be calculated, leading to large computation times and memory requirements, but also to round-off errors. In appendix C some general extrapolation methods are described to improve the convergence of power series. In this section it will be studied how these methods can be used to improve the convergence of the power series produced by the PSA.

2.6.1 Bilinear mapping

In appendix C.1, it is shown how bilinear mappings can be used to enlarge the radius of convergence of power series. There, the coefficients of the new power series are calculated from the coefficients of the original power series. However, when applying the bilinear mapping to the PSA, the new coefficients can also be calculated directly. This makes the algorithm numerically more stable, because the mapping is such that the new coefficients have a smaller growth rate than the original coefficients. Also, it reduces round-off errors. The steady-state distribution will be considered not as a function of γ, but as a function of θ, with

$$\theta = \frac{H\gamma}{1+G\gamma}, \quad \gamma = \frac{\theta}{H-G\theta}, \quad \text{with } 0 \leq G \leq H-1. \tag{2.45}$$

This mapping maps the interval $\left[0, \frac{H-1}{G}\right)$ onto itself. If γ can be interpreted as a measure of the load of the system and the system is stable for $\gamma \in [0, \gamma^*)$, then it is natural to choose $H = 1+G\gamma^*$. All interesting values of γ are then in the interval $\left[0, \frac{H-1}{G}\right) = [0, \gamma^*)$. If γ has no physical interpretation and one is interested in $\gamma = 1$, numerical experiments have shown that the choice $H = 1 + 1.1G$ often yields satisfactory results. The choice of G depends on the particular model. If the model is quite regular, then G can be small (≤ 2), otherwise it needs to be large. Unfortunately, the speed of convergence often decreases with G (whereas the radius of convergence increases with G).

The use of the mapping will be illustrated by application to general 1-dimensional Markov processes without supplementary space. Generalization to multidimensional processes with supplementary space is straightforward. Let B ($\leq \infty$) be the maximal size of upward transitions:

$$\alpha_b(n) = 0, \quad \text{for all } b > B \text{ and } n \in \mathbb{N}.$$

In the γ-process, the upward transition rates $\alpha_b(n)$ are multiplied by γ^b. The added transitions are selfloops with total rate $\sum_{b=1}^{B}\left(1-\gamma^b\right)\alpha_b(n)$, for all $n \in \mathbb{N}$. The balance equations of the γ-process are

$$\left[\sum_{b=-n}^{-1} \alpha_b(n) + \sum_{b=1}^{B} \gamma^b \alpha_b(n)\right] p(\gamma, n)$$
$$= \sum_{b=-\infty}^{-1} \alpha_b(n-b)p(\gamma, n-b) + \sum_{b=1}^{B} \gamma^b \alpha_b(n-b)p(\gamma, n-b),$$

for all $n \in \mathbb{N}$. If the steady-state probabilities $p(\gamma, n)$ are analytic in γ at $\gamma = 0$, then the steady-state probabilities as functions of θ are analytic in θ at $\theta = 0$, also with order n (see appendix C.1). Consequently, they can be represented by the power-series expansion in θ:

$$\tilde{p}(\theta, n) \doteq p\left(\frac{\theta}{H - G\theta}, n\right) = \sum_{r=n}^{\infty} \theta^r \tilde{u}(r, n), \quad \text{for all } n \in \mathbb{N}. \tag{2.46}$$

Replacing γ in the balance equations renders

$$\begin{aligned} &\left[\sum_{b=-n}^{-1} \alpha_b(n) + \sum_{b=1}^{B} \left(\tfrac{\theta}{H-G\theta}\right)^b \alpha_b(n)\right] \tilde{p}(\theta, n) \\ &\quad = \sum_{b=-\infty}^{-1} \alpha_b(n-b)\, \tilde{p}(\theta, n-b) + \sum_{b=1}^{B} \left(\tfrac{\theta}{H-G\theta}\right)^b \alpha_b(n-b)\, \tilde{p}(\theta, n-b). \end{aligned} \tag{2.47}$$

First assume that B is finite. Multiplying by $(H - G\theta)^B$ removes all fractions:

$$\begin{aligned} &\left[\sum_{b=-n}^{-1} (H - G\theta)^B \alpha_b(n) + \sum_{b=1}^{B} \theta^b (H - G\theta)^{B-b} \alpha_b(n)\right] \tilde{p}(\theta, n) \\ &\quad = \sum_{b=-\infty}^{-1} (H - G\theta)^B \quad \alpha_b(n-b) \quad \tilde{p}(\theta, n-b) \\ &\quad + \sum_{b=1}^{B} \theta^b (H - G\theta)^{B-b} \quad \alpha_b(n-b) \quad \tilde{p}(\theta, n-b). \end{aligned}$$

Applying Newton's binomial formula renders

$$\begin{aligned} &\left[\sum_{b=-n}^{-1} \sum_{k=0}^{B} \tbinom{B}{k} H^{B-k} (-G\theta)^k \alpha_b(n) + \sum_{b=1}^{B} \sum_{k=0}^{B-b} \tbinom{B-b}{k} \theta^b H^{B-b-k} (-G\theta)^k \alpha_b(n)\right] \tilde{p}(\theta, n) \\ &\quad = \sum_{b=-\infty}^{-1} \sum_{k=0}^{B} \tbinom{B}{k} H^{B-k} (-G\theta)^k \quad \alpha_b(n-b) \quad \tilde{p}(\theta, n-b) \\ &\quad + \sum_{b=1}^{B} \sum_{k=0}^{B-b} \tbinom{B-b}{k} \theta^b H^{B-b-k} (-G\theta)^k \quad \alpha_b(n-b) \quad \tilde{p}(\theta, n-b). \end{aligned}$$

Dividing by H^B and substituting the power series expansions (2.46) and equating corresponding powers of θ renders the recursion

$$\begin{aligned} &\left[\sum_{b=-n}^{-1} \alpha_b(n)\right] \tilde{u}(r, n) = \\ &\quad - \sum_{b=-n}^{-1} \sum_{k=1}^{B} \tbinom{B}{k} H^{-k} (-G)^k \quad \alpha_b(n) \quad \tilde{u}(r-k, n) \\ &\quad - \sum_{b=1}^{B} \sum_{k=0}^{B-b} \tbinom{B-b}{k} H^{-b-k} (-G)^k \quad \alpha_b(n) \quad \tilde{u}(r-b-k, n) \\ &\quad + \sum_{b=-r+n}^{-1} \sum_{k=0}^{B} \tbinom{B}{k} H^{-k} (-G)^k \quad \alpha_b(n-b) \quad \tilde{u}(r-k, n-b) \\ &\quad + \sum_{b=1}^{B} \sum_{k=0}^{B-b} \tbinom{B-b}{k} H^{-b-k} (-G)^k \quad \alpha_b(n-b) \quad \tilde{u}(r-b-k, n-b). \end{aligned} \tag{2.48}$$

In particular cases these expressions can greatly simplify, especially for birth-death processes.

If the maximal size of the upward transitions B is infinite, then multiplying by $[H - G\theta]^B$ in (2.47) is not possible. Instead, the following identity can be used:

$$\left(\frac{\theta}{H - G\theta}\right)^b = \sum_{k=0}^{\infty} \binom{b-1+k}{b-1} H^{-b-k} G^k \theta^{b+k}.$$

Substitution in (2.47) also removes all fractions:

$$\begin{aligned}
&\left[\sum_{b=-n}^{-1} \alpha_b(n) + \sum_{b=1}^{\infty} \sum_{k=0}^{\infty} \tbinom{b-1+k}{b-1} H^{-b-k} G^k \theta^{b+k} \alpha_b(n)\right] \tilde{p}(\theta, n) \\
&\quad = \sum_{b=-\infty}^{-1} \qquad\qquad\qquad\qquad \alpha_b(n-b) \quad \tilde{p}(\theta, n-b) \\
&\quad + \sum_{b=1}^{n} \quad \sum_{k=0}^{\infty} \tbinom{b-1+k}{b-1} H^{-b-k} G^k \theta^{b+k} \quad \alpha_b(n-b) \quad \tilde{p}(\theta, n-b).
\end{aligned}$$

Equating corresponding powers of θ now renders

$$\begin{aligned}
&\left[\sum_{b=-n}^{-1} \alpha_b(n)\right] \tilde{u}(r, n) = \\
&\quad - \sum_{b=1}^{r-n} \quad \sum_{k=0}^{r-b-n} \tbinom{b-1+k}{b-1} H^{-b-k} G^k \quad \alpha_b(n) \qquad \tilde{u}(r-b-k, n) \\
&\quad + \sum_{b=-r+n}^{-1} \qquad\qquad\qquad\qquad \alpha_b(n-b) \quad \tilde{u}(r, n-b) \\
&\quad + \sum_{b=1}^{n} \quad \sum_{k=0}^{r-n} \tbinom{b-1+k}{b-1} H^{-b-k} G^k \quad \alpha_b(n-b) \quad \tilde{u}(r-b-k, n-b).
\end{aligned}$$

This recursion can also be used if B is finite. It renders the same coefficients as recursion (2.48), but requires slightly more calculations.

2.6.2 Value and pole extrapolation

The extrapolation of values and poles as described in appendix C.2 depends very much on the application. In many queueing models on an infinite state space, an upper bound γ^* exists such that the system is stable for $\gamma \in [0, \gamma^*)$. See, for instance, example 2.3 with $\gamma^* = \mu$. In that case the steady-state probabilities will all have value zero at $\gamma = \gamma^*$. Also, it is not unusual that the k-th moment of the queue length has a k-th order pole at $\gamma = \gamma^*$. See, for instance, the Pollaczek-Khintchine formula for the $M/G/1$ queue. The residue of this pole can either be obtained from heavy-traffic analysis or be estimated in the way suggested in appendix C.2. If the mapping is applied first, with the recommended choice $H = 1 + G\gamma^*$, then the zeros and poles at $\gamma = \gamma^*$ will remain there. Especially the pole extrapolation is very effective in improving the accuracy of the results obtained by the PSA.

2.6.3 The epsilon, theta and Levin algorithms

Extrapolation methods, as described in appendix C.3, are often very effective in improving the convergence speed of the power series produced by the PSA. Experience has shown that the epsilon algorithm is usually more effective than the others. This would be in accordance with the observation in [128] that the epsilon algorithm is preferable unless the series is alternating or logarithmically convergent, which is rarely the case. Usually, the series have irregular sign patterns and linear convergence. However, since the structure of the power series produced by the PSA is not known in advance, it is advisable to use several algorithms. A satisfactory choice is to use the epsilon and theta algorithms and Levin's u-transform. Also, the relative variation criterion to choose between the algorithms works well (see (C.14) with $\kappa = 5$), applied to all even columns of the epsilon and theta algorithms and the subsequent estimates of Levin's u-transform. This is the procedure that was used for the network examples in section 2.8.4.

The procedure described above can be used routinely, but not as a black box. Studying the original power series and the different extrapolated series provides valuable information about the reliability of the results. If possible, the results should be validated, using known characteristics like for example the probability of an empty system, marginal distributions or conservation laws.

If applicable, first the mapping and value or pole extrapolation should be applied. The new series thus obtained can be analyzed by the epsilon, theta and Levin algorithms. In theory, the main diagonal of the epsilon algorithm is invariant under the mapping. That is, if $f(\gamma) = g(\theta)$ with γ and θ related by (2.45), then $[L/L]_f(\gamma) = [L/L]_g(\theta)$ (for definitions, see appendix C.3). However, usually the growth rate of the mapped sequence is smaller, so applying the mapping will still be useful to make the PSA numerically more stable. Also, the main diagonal does not always contain the better approximants. In the $M/M/m$ model, for example, the structure of the steady-state probabilities with less than m customers is different from the structure of those with more than m customers. Therefore, the $[L + \lceil m/2 \rceil / L - \lceil m/2 \rceil]$ Padé approximant of the mean queue length will be better than the $[L/L]$ approximant because the first m partial sums are not used. And since the non-diagonal approximants are calculated anyway, one might as well consider all columns of the epsilon table.

2.7 What if...

In this section, it will be studied what happens if the assumptions of the PSA are not satisfied. First, the assumptions of section 2.1 will be considered, namely the assumptions that $\Omega = \mathbb{N}^S \times \{1, \ldots, I\}$ and that the original process is irreducible, non-instantaneous and ergodic. Next, it will be considered what happens if assumption 0 in section 2.3 is not satisfied, that is if the 0-process has several recurrent classes. Finally, assumptions 1, 2 and 3 from section 2.5 will be studied. The problems that arise when the assumptions

are not satisfied will be discussed and possible remedies will be illustrated by small size Markov processes. Unless stated otherwise, only one assumption at a time is assumed to be not satisfied. In the figures illustrating the examples, the process on the left is the original process, and the process on the right is the γ-process. All transitions without indication of the transition rate have rate equal to 1.

2.7.1 What if the state space is incomplete?

What if Ω is not equal to $\mathbb{N}^S \times \{1, \ldots, I\}$, so $\Omega^c \doteq \mathbb{N}^S \times \{1, \ldots, I\} \backslash \Omega$ is non-empty? This usually does not cause any real problems. All steady-state probabilities in Ω^c are zero, and so are the coefficients of the power-series expansions. In the example in the introduction of this chapter 2, the state space $\Omega = \{1, 2\}$ was only a small part of $\mathbb{N}$ and all states in Ω^c could simply be ignored.

However, problems may arise when the added transitions make jumps to states in Ω^c. For example, consider the process on $\Omega = \{(0,0), (1,1), (2,0)\}$ in figure 2.4. In the

Figure 2.4

original process, the state (1,0) could not be reached, so no departure rate was defined. In the γ-process, this state *can* be reached and it is absorbing. Therefore, assumption 0 is not satisfied and the PSA can not be applied in this form. However, it is not difficult to

Figure 2.5

find a remedy for this problem. Add state (0,1) to the state space Ω and add a transition to the original process from state (1,0) to state (0,0) with rate 1, like in figure 2.5. This new process has the same steady-state distribution as the original process and it can be analyzed with the PSA since assumption 0 is satisfied. This remedy can be used in general: from all states in Ω^c that can be reached in the γ-process, add a transition to a

state in Ω. These new transitions should be non-positive, otherwise the 0-process will not be irreducible so assumption 0 will be violated. For numerical stability, the transition rates of the new transitions should be of the same order of magnitude as other rates of the Markov process.

Sometimes, it may even be an advantage when the state space is not equal to $\mathbb{N}^S \times \{1, \ldots, I\}$, namely when the state space is finite. In that case, the steady-state probabilities are rational functions of γ (see section 2.4). For rational functions it is known that the epsilon algorithm needs only a finite number of coefficients to obtain the correct value. The finiteness can also give the opportunity to use the PSA for heavy-

Figure 2.6

traffic analysis. As a simple example, consider the $M/M/1/3$ queue. Figure 2.6 is the standard light-traffic approach. Because the number of states is finite, the order of the states can be reversed. Then, each downward transition is multiplied by γ, instead of each upward transition (see figure 2.7). Now, small γ correspond to low service capacity, so the approach is a heavy-traffic approach.

Figure 2.7

2.7.2 What if the original process is reducible?

There are different types of reducibility. The harmless type is when the state space of the original process consists of a recurrent class and transient states that eventually lead to this recurrent class. Then the steady-state distribution of the original process is well defined and all transient states have steady-state probability zero. Examples are the process in figure 2.7 in the previous section and the process on $\Omega = \{0, 1, 2\}$ in figure 2.8 below. The transient states of the original process can still be transient in the γ-process for all $\gamma \in [0, 1]$, like state 2 of the process in figure 2.8. The coefficients of the power-series expansions of the steady-state probabilities of such transient states are all zero. On the other hand, the transient states of the original process may also be recurrent in the γ-process, like state (1,0) in the process of figure 2.7. Then, for

Figure 2.8

such a transient state, the coefficients are non-zero but the power-series expansion of the steady-state probability evaluated at $\gamma = 1$ is equal to zero. In either case, the problem is well defined and the PSA will work as long as assumption 0 is satisfied.

Another type of reducibility is when the original process has transient states that do not eventually lead to a recurrent class, like the process on $\mathbb{N}$ in figure 2.9. From

Figure 2.9

each state $n \in \mathbb{N}$, the only possible transition is to state $n + 1$. Now, the PSA will not work because in the 0-process each state is absorbing. Therefore, assumption 0 is not satisfied. This is not surprising, since the original process is not ergodic. No solution procedure can be blamed for not being able to solve an ill-defined problem.

The final type of reducibility is when the original process has more than one recurrent class and the process will eventually reach one of the classes. Again, this is not really a problem since the problem of finding the steady-state distribution is ill defined. Usually, the γ-process will also have several recurrent classes, so assumption 0 will not be satisfied and the PSA will not work. In some cases, the extra downward transitions of the γ-process will make it irreducible. Consider the process on $\Omega = \{n \in \mathbb{N}^2 | \; ne \leq 3\}$ in figure 2.10 below. In the original process, all states $n \in \Omega$ with $ne \leq 2$ are transient,

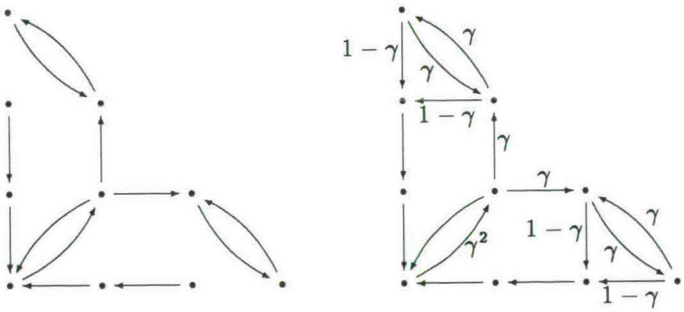

Figure 2.10

and the process will eventually end up in either the recurrent class $\{(0,3),(1,2)\}$ or the

recurrent class $\{(2,1),(3,0)\}$. The long-run distribution depends on the initial distribution, so the steady-state distribution is ill defined. In the γ-process, the states in the two previously recurrent classes have new transitions back to the previously transient states and all states form a single recurrent class. Assumption 0 is satisfied, so the power-series expansions of the steady-state probabilities of the γ-process can be calculated and evaluated at $\gamma = 1$. This renders the distribution with probability mass $\frac{1}{4}$ in each of the states (0,3), (1,2), (2,1) and (3,0). This is the correct steady-state distribution if the original process started in one of the transient states, or if it started in either of the recurrent classes with probability $\frac{1}{2}$. So, even though the original problem is ill defined, the PSA does come up with a solution and this solution corresponds to a particular choice of the initial distribution of the process. Of course this example is contrived, but it does show that a successful application of the PSA does not necessarily imply that the original problem was well defined.

2.7.3 What if the original process has instantaneous states?

Instantaneous states are states with infinite total transition rate. This can be either because individual transition rates are infinite or because the total transition rate is a divergent sum of finite individual transition rates. In either case, solving the sets of recursive equations in (2.12), (2.13) and (2.14) will generally cause problems.

When in each state at most a finite number of individual transition rates is infinite, the instantaneous states can often be removed from the state space. Usually, probabilities can then be specified such that the instantaneous state (m,k) leads to state $(m+b,j)$ with probability $\pi_{bj}(m,k)$. Then the instantaneous state can be removed by changing the transition rates of the other states:

$$a'_{bj}(n,i) = a_{bj}(n,i) + a_{m-n,k}(n,i)\frac{\pi_{n+b-m,j}(m,k)}{1-\pi_{ok}(m,k)}, \tag{2.49}$$

for all $(n,i) \in \Omega\backslash\{(m,k)\}$. Instead of via the instantaneous state, the process immediately makes a transition to the new state. The new rate is the direct transition rate plus the rate of the transition via the instantaneous state. For the removal of several instantaneous states, possibly with cycles, see [65] or also [63,39].

When all individual transition rates are finite but their sum is infinite, removing the instantaneous states will not be possible because the distribution π can not be specified. However, because all individual transition rates are finite, the only possible problems in solving the recursive equations of the PSA caused by instantaneous states can come from the infinity of $B(.) = \bar{A}(.) - \sum_{b\in\mathbb{N}^S} A_b(.)$. If these matrices are all finite, then the power-series expansions can be calculated without any problems. This is always true in the one-dimensional case ($S = 1$). In the multidimensional case without supplementary space ($S > 1,\ I = 1$) it is equivalent to the condition that for each state the sum of all the rates of transitions $b \notin \mathbb{N}^S$ is finite. Still, even if all matrices $B(.)$ are indeed finite,

this only guarantees that all calculations can be carried out and all coefficients are finite. Convergence of the obtained power series is likely to give problems.

2.7.4 What if the original process is not ergodic?

If the original process is not ergodic but assumptions 0, 1, 2, and 3 are satisfied, then the power-series expansions of the steady-state probabilities as a function of γ are well-defined analytic functions for small γ. The question is what happens if these power series are evaluated at $\gamma = 1$.

Figure 2.11

Consider the $M/M/1$ queue with arrival rate λ and service rate μ. All assumptions are satisfied here, except that the original process is ergodic only if $\lambda < \mu$. The steady-state probabilities of the γ-process and the first two moments are equal to

$$\begin{aligned}
p(\gamma, n) &= \left(1-\gamma\tfrac{\lambda}{\mu}\right)\left(\gamma\tfrac{\lambda}{\mu}\right)^n, \quad \text{for all } n \geq 0, \\
\mathbb{E}_\gamma\{\,\mathcal{N}\,\} &= \left(1-\gamma\tfrac{\lambda}{\mu}\right)\textstyle\sum_{n\geq 0} n\left(\gamma\tfrac{\lambda}{\mu}\right)^n = \gamma\tfrac{\lambda}{\mu}\left(1-\gamma\tfrac{\lambda}{\mu}\right)^{-1}, \\
\mathbb{E}_\gamma\{\,\mathcal{N}^2\,\} &= \left(1-\gamma\tfrac{\lambda}{\mu}\right)\textstyle\sum_{n\geq 0} n^2\left(\gamma\tfrac{\lambda}{\mu}\right)^n = \left(\gamma\tfrac{\lambda}{\mu}\right)\left(1+\gamma\tfrac{\lambda}{\mu}\right)\left(1-\gamma\tfrac{\lambda}{\mu}\right)^{-2}.
\end{aligned}$$

The power-series expansions of the steady-state probabilities are finite polynomials, so they converge at $\gamma = 1$. The power-series expansions of both moments converge at $\gamma = 1$ only if the original process is ergodic. However, with the techniques from appendix C, they can easily be made to converge at $\gamma = 1$, even if the original process is not ergodic. In the non-ergodic case, the power-series expansions of the steady-state probabilities evaluated at $\gamma = 1$ render negative values. Also the first moment renders a negative value, but the second moment is always positive.

The $M/M/1$ queue is typical for the general case. If the process is not ergodic, the power-series can not converge to a distribution. By construction, if they converge then they satisfy the balance equations and sum up to 1. Therefore, they must either diverge or converge to values that are not non-negative. If only particular performance measures are computed this convergence to incorrect values may not be noticed, like the second moment in the $M/M/1$ case. So, carefulness is in order.

2.7.5 What if assumption 0 is not satisfied?

If the original process is irreducible, then so are the γ-processes for $\gamma \in (0, 1)$, since all original transitions can still be made (but with different rates). Therefore, if the γ-process

is ergodic for $\gamma \in (0,1)$, then the steady-state distribution is uniquely determined by the balance and normalization equations. However, in the 0-process all upward transitions vanish. As a result of this, assumption 0 may not be satisfied since the 0-process may have several recurrent classes. Then the steady-state distribution is no longer uniquely determined by the balance and normalization equations. Since the PSA is based on the information of the γ-process at $\gamma = 0$, it will not be applicable if the problem at $\gamma = 0$ is ill behaved. Nevertheless, the algorithm can often be applied after some modifications. Four methods will be discussed in this section. The first approach removes states from the original process. The second approach removes coefficients from the recurrence relations. The third derives extra information from a continuity assumption, but is not generally applicable. The fourth approach changes the transformation. Which method is best suited depends on the particular application.

The first two approaches are illustrated by the same small-size example. Both approaches can be used in many cases, but how easily they can be applied depends on the particular model. Consider the Markov process in figure 2.12. For the original process, the balance

Figure 2.12: A troublesome process

equations of the non-empty states and the normalization equation are

$$p(1,0) = p(0,0), \quad p(1,1) = p(1,0), \quad p(0,1) = p(1,1),$$
$$p(0,0) + p(1,0) + p(1,1) + p(0,1) = 1.$$

Standard application of the PSA multiplies the arrival rates from state (0,0) to state(1,0) and from (1,0) to (1,1) by γ. For the γ-process, the balance and normalization equations are

$$\gamma p(\gamma,1,0) = \gamma p(\gamma,0,0), \quad p(\gamma,1,1) = \gamma p(\gamma,1,0), \quad p(\gamma,0,1) = p(\gamma,1,1),$$
$$p(\gamma,0,0) + p(\gamma,1,0) + p(\gamma,1,1) + p(\gamma,0,1) = 1.$$

The recurrence relations for the power-series expansions of the steady-state probabilities of the original model are therefore

$$u(r,1,0) = u(r,0,0), \quad u(r,1,1) = u(r-1,1,0), \quad u(r,0,1) = u(r,1,1),$$
$$u(r,0,0) + u(r,1,0) + u(r,1,1) + u(r,0,1) = 1(r=0),$$

for $r \geq 0$, with $u(-1,1,0) = 0$. The usual order of calculation is for increasing values of r and decreasing values of $n_1 + n_2$. This is not possible here: to calculate coefficient $u(r,1,0)$, coefficient $u(r,0,0)$ is needed, but this coefficient has not yet been calculated. The coefficients $u(r,0,0)$ and $u(r,1,0)$ need to be calculated simultaneously by solving a small set of equations. Assumption 0 is not satisfied: for $\gamma = 0$, both state (0,0) and state (1,0) are absorbing. Also, the order property (2.7) is not satisfied.

The first approach to solve this problem changes the original problem, such that the new problem can be analyzed with the PSA. The problem is caused by state (1,0), so the obvious cure is to remove this state from the process. Substituting the balance equation $p(1,0) = p(0,0)$ and choosing a different normalization renders the new untransformed process in figure 2.13.

Figure 2.13: The process with state (1,0) removed

The balance and normalization equations are

$$\tilde{p}(1,1) = \tilde{p}(1,0), \quad \tilde{p}(0,1) = \tilde{p}(1,1),$$
$$\tilde{p}(0,0) + \tilde{p}(1,1) + \tilde{p}(0,1) = 1.$$

The only recurrent state of the 0-process of this new Markov process is state (0,0), so the PSA can be applied to find $\tilde{p}(\gamma,0,0) = \frac{1}{1+2\gamma^2}$ and $\tilde{p}(\gamma,1,1) = \tilde{p}(\gamma,0,1) = \frac{\gamma^2}{1+2\gamma^2}$. Evaluated at $\gamma = 1$, this renders the distribution with probability mass $\frac{1}{3}$ in all three states. The solution to the original problem can then be found by finding the probability of the removed states, which will in general require solving a set of equations, and renormalizing:

$$p(0,0) = \frac{\tilde{p}(0,0)}{1+\tilde{p}(0,0)}, \quad p(1,0) = \frac{\tilde{p}(0,0)}{1+\tilde{p}(0,0)},$$
$$p(1,1) = \frac{\tilde{p}(1,1)}{1+\tilde{p}(0,0)}, \quad p(0,1) = \frac{\tilde{p}(0,1)}{1+\tilde{p}(0,0)}.$$

This finally renders the distribution with probability mass $\frac{1}{4}$ in all four states.

The second approach is essentially very similar to the previous one. The first approach removes the states that cause problems from the balance equations. The second approach removes the coefficients of these states from the recurrence relations. The coefficients of state (1,0) can be removed by substituting $u(r,1,0) = u(r,0,0)$:

$$u(r,1,1) = u(r-1,0,0), \quad u(r,0,1) = u(r,1,1),$$
$$2\,u(r,0,0) + u(r,1,1) + u(r,0,1) = 1(r=0),$$

for $r \geq 0$, with $u(-1,0,0) = 0$. Now, the coefficients can be calculated recursively in the usual order, followed by the calculation of the coefficient of state (1,0). This renders $p(\gamma,0,0) = \frac{1}{2+2\gamma}$, $p(\gamma,1,1) = p(\gamma,0,1) = \frac{\gamma}{2+2\gamma}$ and $p(\gamma,1,0) = \frac{1}{2+2\gamma}$. Evaluated at $\gamma = 1$, this renders the distribution with probability mass $\frac{1}{4}$ in all four states. The advantage of this approach over the first is that the original process is not altered and no renormalization is needed. For a more extensive discussion of this approach, see [28]. There, it is applied to polling models in which the cyclic server is allowed to rest at the queues when the system is empty.

The third approach is not generally applicable. When it is applicable, it will usually be preferred over the other approaches. For $\gamma > 0$ the process is irreducible, because the original process is irreducible. For $\gamma = 0$, all upward transitions vanish, which can lead to violation of assumption 0. When there are indeed several recurrent classes, the solution of the balance equations is no longer unique. Often, the distribution $p(0,.) \doteq \lim_{\gamma \downarrow 0} p(\gamma,.)$ will be one of the distributions that satisfy the balance equations. How can this distribution be obtained? Sometimes, it is possible to find extra equations that are true for $\gamma > 0$, but not necessarily for $\gamma = 0$. The third approach is to require that the extra equations are also true for $\gamma = 0$, hoping that this reduces the number of solutions to 1. If so, then this unique steady-state distribution at $\gamma = 0$ is the solution that makes the steady-state probabilities right-continuous at $\gamma = 0$.

As an example, consider the $M/H_J/1$ model. At a queue, customers arrive and are served by a single server. The interarrival time has an exponential distribution with rate λ. The service-time distribution is hyper-exponential: with probability π_j it has rate μ_j, for all $1 \leq j \leq J$. Let M be the diagonal matrix with the service rates on the diagonal and π the row vector of initial probabilities. Then the mean service time is equal to $\mu^{-1} = \pi M^{-1} e$. The service phase is chosen right after the departure of the preceding customer. In the γ-process, the arrival rate is multiplied by γ. The γ-process is ergodic for γ in the interval $(0, \lambda^{-1}\mu)$ and the balance and normalization equations are

$$\begin{aligned} &p(\gamma,n)\ [\gamma\lambda + M1(n>0)] = p(\gamma,n-1)\gamma\lambda + p(\gamma,n+1)Me\pi, \quad \text{for all } n \geq 0, \\ &\sum_{n=0}^{\infty} p(\gamma,n)e = 1, \end{aligned}$$

with $p(\gamma,-1) \equiv 0$. Replacing the probabilities by their power-series expansions and equating corresponding powers of γ renders

$$\begin{aligned} u(r,n)M &= -u(r-1,n)\lambda + u(r-1,n-1)\lambda + u(r,n+1)Me\pi, && \text{for all } r \geq n \geq 1, \\ u(r,0)e &= 1(r=0) - \sum_{n=1}^{r} u(r,n)e, && \text{for all } r \geq 0, \end{aligned} \tag{2.50}$$

with $u(r,n) = o$ if $n = -1$ or $r < n$. Unless $J = 1$, assumption 0 is not satisfied because all empty states are absorbing if $\gamma = 0$. Therefore, the coefficients for the empty states can not be calculated by the above equations (2.50). The balance equations provide no

useful information and the normalization equation only determines their sum. However, extra equations can be obtained here. If $\gamma > 0$ and the system is empty, then the service phase is equal to j with probability π_j. Therefore:

$$p(\gamma, 0) = p(\gamma, 0)e\pi, \quad \text{if } \gamma > 0. \tag{2.51}$$

If $\gamma = 0$, the process is reducible and the steady-state distribution is ill-defined. Nevertheless, suppose that (2.51) is also true for $\gamma = 0$. Substituting the power-series expansions in (2.51) and equating coefficients of corresponding powers of γ renders the additional equations

$$u(r, 0) = u(r, 0)e\pi, \quad \text{for all } r \geq 0.$$

Together with the second equation in (2.50), this extra information allows for the recursive calculation of all the empty coefficients. More extensive examples are given in the network model of section 2.8 and the papers [74,75]. Other sources of additional equations can be independence or symmetry properties of the model or for example conservation laws.

The fourth approach is to use a different transformation. According to assumption 0, problems arise when the 0-process has several recurrent classes. The solution to the problem could be to use a different transformation such that assumption 0 *is* satisfied. One way to accomplish this is to prevent that certain transitions vanish at $\gamma = 0$. For example, the problems with the process in figure 2.12 are caused by the fact that the transition from state (1,0) to state (1,1) vanishes when $\gamma = 0$. This can be avoided by not transforming this transition and keeping the transition rate equal to 1 in the γ-process. Another way to satisfy assumption 0 is to add more transitions to the γ-process such that the 0-process has only one recurrent class consisting of all empty states. For instance, this can be done by adding the following transitions to the γ-process:

$$\begin{array}{lll} (n,i) \rightarrow (\ [\ n - e^T\ ]^+, i\ ) & \text{with rate} \quad \delta(1-\gamma), & \text{for all } 1 \leq i \leq I \text{ and } n \neq o, \\ (o,i) \rightarrow (\ o, i \bmod I + 1\ ) & \text{with rate} \quad \delta(1-\gamma), & \text{for all } 1 \leq i \leq I, \end{array} \tag{2.52}$$

for some fixed $\delta > 0$. The vector x^+ denotes the element-wise maximum of the zero vector o and x (see section 1.4). This way, all non-empty states have a transition to a state with less customers, so for $\gamma = 0$ the non-empty states can not be recurrent. For the empty states, a set of cyclic transitions is added so that all the empty states form one recurrent class. The γ-process with $\gamma = 1$ is still equal to the original process. Setting up the balance equations and the recurrence relations for the coefficients of the power-series expansions shows that indeed for this extended transformation all coefficients can be calculated recursively with an algorithm similar to the algorithm described before. With the extra transitions suggested in (2.52), assumption 0 is satisfied for any Markov process. However, it may not be the most efficient way to add transitions. For specific models, other ways may yield power series with better convergence properties.

In the example below, a number of different ways to add transitions is compared. The different transformations are tested on a multiprocessor model with breakdowns and repairs. Let $\mathcal{S}$ be the number of available servers (processors) and $\mathcal{C}$ the number of customers (jobs). Servers arrive at rate η. Each server departs at rate ξ, the total departure rate is $\mathcal{S}\xi$ and the number of available servers behaves like an $M/M/\infty$ queue. Customers arrive at rate λ. They are served by the available servers at rate μ, so the departure rate is $\mu \min\{\mathcal{C}, \mathcal{S}\}$. The mean number of available servers is $\frac{\eta}{\xi}$ and the system is stable if $\lambda < \mu\frac{\eta}{\xi}$. This model was analyzed by Ettl and Mitrani [53] using the spectral expansion method. This method compares favourably with the PSA because it is less sensitive to extreme parameter values, more specifically larger numbers of customers served per server. A disadvantage of the method is that it can only be applied to one-dimensional problems, so the state space needs to be truncated by limiting the number of servers.

In the transformed process of the PSA, the arrival rates of both servers and customers is multiplied by γ. Since both the number of servers and the number of customers increase linearly with γ, the transformed process is ergodic for all $\gamma > 0$, provided the original system is ergodic. The balance equations of this birth-death process are

$$\begin{aligned}
&[\gamma\eta + \xi s + \gamma\lambda + \mu \min\{s,c\}]\, p(s,c) \\
&\quad = \gamma\eta \;\; p(s-1,c) \;+\; \xi(s+1) \;\; p(s+1,c) \\
&\quad + \gamma\lambda \;\; p(s,c-1) \;+\; \mu \min\{s,c+1\} \;\; p(s,c+1),
\end{aligned}$$

for all $(s,c) \in \mathbb{N}^2$. The parameter values $\eta = \xi = \lambda = 1$ and $\mu = 2$ were chosen. On average, a server serves only a single customer. All states with no servers are absorbing if $\gamma = 0$, so assumption 0 is not satisfied. The following ways to add transitions will be compared:

	transition	at rate	from states
transformation 1:	$(s,c) \to (s-1,c-1)^+$,	$\delta(1-\gamma)$,	$s \geq 0,\ c \geq 0$,
transformation 2:	$(s,c) \to (s,c-1)$,	$\delta(1-\gamma)$,	$s \geq 0,\ c \geq 1$,
transformation 3:	$(s,c) \to (s,c-1)$,	$\delta(1-\gamma)$,	$s = 0,\ c \geq 1$,
transformation 4:	$(s,c) \to (s,c-1)$,	$\delta(1-\gamma)^2$,	$s \geq 0,\ c \geq 1$,

all with the same value of δ. The first transformation is as suggested in the general approach (2.52). The process is no longer a birth-death process. In the interior of the state space the new transitions are of the same kind, but not on the boundary. For example, state (s,0) can be entered by an extra transition not only from (s+1,1) but also from (s+1,0). In the second transformation, only the number of customers is decreased. The third one decreases only the number of customers when there are no servers, so only the states that violate assumption 0 are altered. In these first three transformations, the transition rate becomes negative for $\gamma > 1$. The fourth transformation may be better, because this system is well-defined for all $\gamma \geq 0$. Setting up the new balance equations, substituting the power-series expansions and equating corresponding powers

of γ renders recursive algorithms. The obtained power series for the probability that there are no customers in the queue are evaluated at truncation levels 5, 10, 20, 40, 60, 80 and 100. These power series do not converge at $\gamma = 1$, but they *do* after application of the procedure as described in section 2.6.3. The results were validated by calculating the marginal distribution of the number of servers, which should be a Poisson distribution with mean $\frac{\eta}{\xi}$.

	5	10	20	40	60	80
Transformation 1	0.80508	0.37926	0.28239	0.28968	0.28943	0.28942
Transformation 2	1.11703	1.11336	0.28895	0.28945	0.28942	0.28942
Transformation 3	1.15225	0.22784	0.28992	0.28908	0.28942	0.28943
Transformation 4	5.14300	0.52296	-46651.7	0.25399	0.25244	0.28924

Table 2.2: The four transformations, all with $\delta = 1$.

All approaches converge to the answer 0.28942, the second transformation faster than the others. This indicates that the extra transitions should be added in such a way that the distortion is as evenly as possible (transformation 1 versus 2) and that changing the process evenly is more important than making less alterations (transformation 2 versus 3). This is probably because the epsilon algorithm can remove the influence of the extra transitions better if the transformation is more evenly. Transformation 4 requires more calculations and yields power series that are more irregular. Next, the value of δ will be varied for the second transformation.

	5	10	20	40	60	80
$\delta = 0.25$	32.14995	-0.205918	-1.313E11	0.363251	0.290917	0.289349
$\delta = 1$	1.117034	1.113363	0.288949	0.289448	0.289424	0.289424
$\delta = 4$	0.276481	0.201189	0.278093	0.289260	0.289424	0.289424
$\delta = 16$	0.734703	0.547376	0.311964	0.291300	0.292388	0.288278

Table 2.3: Transformation 2, for different values of δ.

The results show that the precise value of δ is not very important, but it should be in the order of magnitude of the other transition rates.

2.7.6 What if assumptions 1, 2 or 3 are not satisfied?

Assumptions 1, 2 and 3 are sufficient conditions for convergence. Contrary to assumption 0, if they are not satisfied the PSA can still be applied. In that case, there is no guarantee that the obtained power series converge. Example 2.6 shows that if assumption 1 is not satisfied, the γ-process can indeed be ill behaved, even for small γ. Assumptions 2 and 3 are very weak. If they are not satisfied, then usually the original process and the PSA are either both ill behaved or both well behaved. Example 2.7 is an example where assumption 2 is not satisfied and both are ill behaved. In example 2.8, assumption 3 is not satisfied but both are well behaved.

Example 2.6 Consider the Markov process on $\mathbb{N}^2$ illustrated in figure 2.14. When no rate is indicated, the rate equals 1. This process always ends in the finite cycle $(0,0) \to (3,1) \to (2,0) \to (1,1) \to (0,0)$, so the steady-state distribution exists. The transition-diagram of the corresponding γ-process (without the selfloops) is given in figure 2.15. For $\gamma = 0$, the process will always end up in the origin, so assumption 0 is satisfied. Assumption 1 is satisfied if $\sup_k \lambda_k < \infty$ ($c_0(2k+2,0) = 2+\lambda_k$, for $k \geq 1$). No moments will be considered, so assumption 2 does not apply. Assumption 3 is satisfied if λ_k grows at most exponentially in k.

When the γ-process is in a state $(2k,0)$, with $k \geq 1$, then in two steps it will go either up to state $(2k+2,0)$ or down to state $(2k-2,0)$. The probability of going up is equal to $\pi_k(\gamma) = (1-\gamma)\ \lambda_k\gamma^3(1+\lambda_k\gamma^3)^{-1}$, which is the probability of going first to state $(2k-1,0)$ and then to state $(2k+2,0)$.

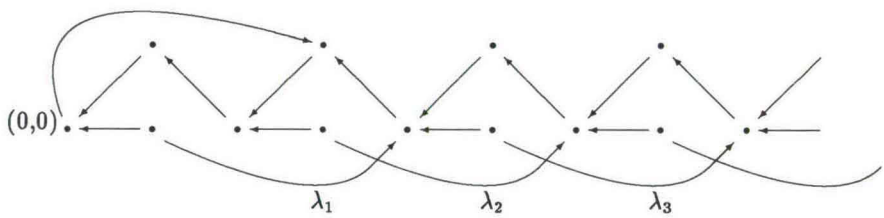

Figure 2.14: The original process

First, consider the case $\lambda_k \equiv \lambda < 2048/27$. Then, going down is more likely than going up: $\pi_k(\gamma) < \frac{1}{2}$, for all $k \geq 1$ and γ in [0,1]. Therefore, the γ-process is ergodic for all γ in [0,1]. Applying the PSA will give no problems.

Next, take $\lambda_k \equiv \lambda \geq 2048/27$. Then the equation $\pi_k(\gamma) = \frac{1}{2}$ has solutions γ_1 and γ_2 with $0 < \gamma_1 \leq \frac{3}{8} \leq \gamma_2 < \frac{1}{2}$. The γ-process is ergodic for $\gamma \in [0,\gamma_1) \cup (\gamma_2,1]$, null-recurrent for $\gamma \in \{\gamma_1,\gamma_2\}$ and transient for $\gamma \in (\gamma_1,\gamma_2)$. Therefore, the γ-process is ergodic at $\gamma = 1$ (as was assumed in section 2.1) and in a neighbourhood of $\gamma = 0$ (as was proved in theorem 2.3), but it is not ergodic for all values of γ in between. Applying the PSA will render functions that will be negative on $\gamma \in (\gamma_1,\gamma_2)$, but the techniques in appendix C can be used to obtain convergence at $\gamma = 1$.

Finally, take $\lambda_k = k$, for $k \geq 1$. Then $\lim_{k\to\infty} \pi_k(\gamma) = 1-\gamma > \frac{1}{2}$, for all γ in $(0,\frac{1}{2})$. On the other hand, $\pi_k(\gamma) < 1-\gamma \leq \frac{1}{2}$, for all $k \geq 1$ and γ in $[\frac{1}{2},1]$. Therefore, the γ-process is ergodic for $\gamma \in \{0\} \cup [\frac{1}{2},1]$ and transient for $\gamma \in (0,\frac{1}{2})$. That the γ-process is not ergodic in a neighbourhood of $\gamma = 0$ is not in contradiction with theorem 2.2, because assumption 1 is not satisfied. Application of the PSA will not be successful here.

In this example, the transient behaviour of the γ-process can be avoided by not considering the original process as a process on $\mathbb{N}^2$ but as a process on $\mathbb{N} \times \{0,1\}$. On

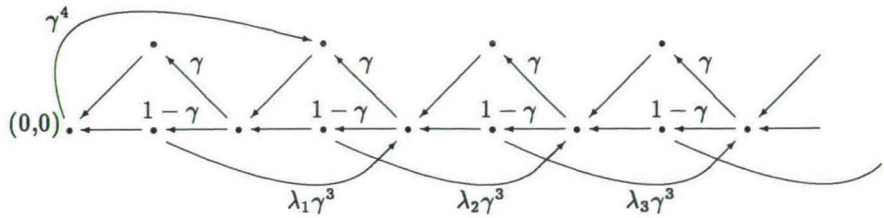

Figure 2.15: The transformed process on $\mathbb{N}^2$

$\mathbb{N}^2$, the transition $(2k,0) \to (2k-1,1)$ is an upward transition and hence redirected to the transition $(2k,0) \to (2k-1,0)$. From the state $(2k-1,0)$, the process has a high probability to go up to state $(2k+2,0)$. This way, replacing upward transitions by downward transitions results in more visits to states from which large upward transitions are likely. Considered as a process on $\mathbb{N} \times \{0,1\}$, the transition $(2k,0) \to (2k-1,1)$ is downward instead of upward, and is therefore not redirected. □

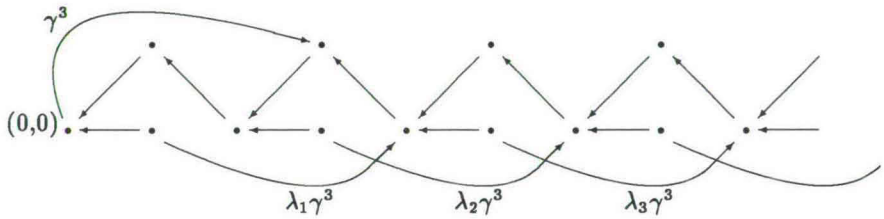

Figure 2.16: The transformed process on $\mathbb{N} \times \{0,1\}$

Example 2.7 Moments of functions that grow faster than geometric in the queue-length will usually be infinite even for small γ, because the queue-length distribution often has geometric tails. Consider again the $M/M/1$ queue from section 2.7.4. The function $f(n) = n!$ grows faster than exponentially in n, so it does not satisfy assumption 2. The expectation for the original process is infinite. By the Cauchy-Hadamard theorem B.3, the radius of convergence of the expansion

$$\mathbb{E}_\gamma \{ \mathcal{N}! \} = \left(1 - \gamma\frac{\lambda}{\mu}\right) \sum_{n\geq 0} n! \left(\gamma\frac{\lambda}{\mu}\right)^n$$

is equal to

$$\left[\limsup_{n\to\infty} \left| n! \left(\frac{\lambda}{\mu}\right)^n \right|^{1/n} \right]^{-1} = \frac{\mu}{\lambda} \left[\limsup_{n\to\infty} (n!)^{1/n} \right]^{-1} = \frac{\mu}{\lambda} [\ \infty\ ]^{-1} = 0.$$

Therefore, the expectation is not analytic in γ, not even for small γ. □

Example 2.8 Consider the birth-death process on $\Omega = \mathbb{N}$ in figure 2.17. From state $n \in \Omega$, the upward transition rate is equal to 1 and the downward transition rate is equal to $n!$. Assumption 3 is not satisfied here. Still, the γ-process is ergodic for any $\gamma > 0$.

Figure 2.17

The faster than exponential growth of the transition rates is no problem because the upward rates are finite. The PSA can be applied without any problem. □

2.8 Networks of queues

To illustrate the generality of the PSA it will be applied to a very wide class of network models. Networks of queues are usually difficult to analyze, both analytically and numerically. Except for product-form networks and some special two-queue models, very few results are available. The class of models that will be introduced in this section is too general for analysis by other methods and the arrival and service processes that are used have not been introduced before. The arrival process will be a Multi-queue Markovian Arrival Process ($MMAP$), which is a multi-queue generalization of the Batch Markovian Arrival Process ($BMAP$) introduced by Lucantoni [102]. On top of the ability of the $BMAP$ to model dependencies between interarrival times and batch sizes, the $MMAP$ can also model all kinds of dependencies between arrivals at the different queues, like fork and round-robin arrivals. At each queue the service process is a Markovian Service Process (MSP). It includes for example set-up times, sequences of phase-type distributions and multi-server queues. The routing of customers is Markovian, which includes a large variety of network structures.

Because of the 'curse of dimensionality', the number of queues in the networks must necessarily be moderate. Networks of up to 4 or 5 queues can be analyzed if the algorithm is programmed carefully, extrapolation methods are employed and the parameters of the model are not too extreme. With a good user interface to determine the parameters for a particular model, the PSA provides a means to easily evaluate many different models. The wide range of models that can be analyzed makes the PSA a valuable aid for studying the interaction between queues and for developing and testing approximations of performance measures and heuristics.

2.8.1 The network process

The number of queues in the network will be denoted by S. Each queue has an unbounded capacity. Finite buffers could be handled, but would require specifying the blocking process.

In the examples described below, a phase-type distribution has generator T, initial distribution α and $T^0 \doteq -Te$ (conform Neuts [113], but without probability mass at zero). Such a phase-type distribution has probability distribution $F(x) = 1 - \alpha \exp(Tx)e$ and density function $F'(x) = \alpha \exp(Tx)T^0$, for $x \geq 0$. The k-th non-central moment is equal to $(-1)^k k! \left(\alpha T^{-k} e\right)$. The class of phase-type distributions includes the Erlang and hyper-exponential distributions as well as finite mixtures of these.

Multi-queue Markovian Arrival Process

The arrival process is a Multi-queue Markovian Arrival Process. This very general arrival process can model all kinds of interactions between the arrivals at the different queues. It is far more general than the models usually considered in the literature and renders intractable results for other methods. It has an underlying irreducible Markov process with J_0 states. In this underlying process, a transition $j \to h$ is made with rate α_{jh} $(1 \leq j, h \leq J_0 < \infty)$. The S-dimensional vectors $b_1, \ldots, b_M$ are the possible batch arrivals, with $b_m \in \mathbb{N}^S \backslash \{o\}$ for all $1 \leq m \leq M \leq \infty$ and $b_0 \doteq o$. A transition $j \to h$ in the underlying process causes an arrival of batch b_m with probability q_{mjh}.

$$\begin{array}{ll} A \doteq [\alpha_{jh}], & \bar{A} \doteq \operatorname{diag}(Ae), \\ Q_m \doteq [q_{mjh}], & \sum_{m=0}^{M} Q_m = ee^T, \\ A_m \doteq A \odot Q_m = [\alpha_{jh} q_{mjh}], & \sum_{m=0}^{M} A_m = A. \end{array}$$

The pure $MMAP$ $\{(\mathcal{N}_t, \mathcal{J}_t),\ t \geq 0\}$ on state space $\mathbb{N}^S \times \{1, \ldots, J_0\}$ is identical to the $BMAP$ if $S = 1$ and $b_m = m$ $(0 \leq m \leq \infty)$. It then has generator

$$Q^{MMAP} = \begin{pmatrix} A_0 - \bar{A} & A_1 & A_2 & \cdots \\ O & A_0 - \bar{A} & A_1 & \cdots \\ O & O & A_0 - \bar{A} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

Like in a pure birth process, only upward transitions are possible. Lucantoni [102] lists a number of special cases of the $BMAP$, like the Poisson process, Markov-modulated Poisson processes, PH-renewal processes and processes with correlated batch arrivals.

If each queue has an independent $BMAP$, this can be modeled as an $MMAP$ for the network. Other special cases of $MMAP$s are:
1) Independent Poisson arrivals with rate λ_s at queue s:

$$\begin{array}{l} J_0 = 1, \quad A_0 - \bar{A} = -\sum_{s=1}^{S} \lambda_s, \\ b_s = e_s, \quad A_s = \lambda_s, \quad \text{for all } 1 \leq s \leq S = M. \end{array}$$

2) Round-robin arrivals, an arrival at queue s is followed by an arrival at queue $s+1$ with the interarrival time exponentially distributed with rate λ_s:

$$\begin{aligned} &J_0 = S, \quad A_0 - \bar{A} = -\mathrm{diag}(\lambda), \\ &b_s = e_s, \quad A_s = \lambda_s e_s e_{s+1}^T, \quad \text{for all } 1 \le s \le S = M. \end{aligned}$$

3) Fork arrivals, simultaneous arrivals at each queue with independent phase-type interarrival times (with ℓ phases):

$$\begin{aligned} &J_0 = \ell, \quad A_0 - \bar{A} = T, \\ &b_1 = e, \quad A_1 = T^0\alpha, \quad M = 1. \end{aligned}$$

Markovian Service Process

The service processes at all queues are independent Markovian Service Processes. Like the $MMAP$, the MSP is more general than the service processes usually considered in the literature. An MSP has an underlying Markov process with J states and the transition rates are allowed to depend on the number of customers n at that queue. A transition $j \to h$ is made with rate $\beta_{jh}(n)$ and such a transition causes a service completion of ℓ customers with probability $r_{\ell jh}(n)$ $(1 \le j, h \le J < \infty;\ 0 \le \ell \le n \le \infty)$. Which ℓ customers leave the queue will not be specified because it is not essential for the queue-length process (but it is for the waiting times).

$$\begin{aligned} &B(n) \doteq [\beta_{jh}(n)], && \bar{B}(n) \doteq \mathrm{diag}(B(n)e), \\ &R_\ell(n) \doteq [r_{\ell jh}(n)], && \textstyle\sum_{\ell=0}^n R_\ell(n) = ee^T, \\ &B_\ell(n) \doteq B(n) \odot R_\ell(n) = [\beta_{jh}(n) r_{\ell jh}(n)], && \textstyle\sum_{\ell=0}^n B_\ell(n) = B(n). \end{aligned}$$

A pure MSP $\{(\mathcal{N}_t, \mathcal{J}_t),\ t \ge 0\}$ on state space $\mathbb{N} \times \{1, \ldots, J\}$ has generator

$$Q^{MSP} = \begin{pmatrix} B_0(0) - \bar{B}(0) & O & O & \cdots \\ B_1(1) & B_0(1) - \bar{B}(1) & O & \cdots \\ B_2(2) & B_1(2) & B_0(2) - \bar{B}(2) & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

Like in a pure death process, only downward transitions are possible. A number of assumptions will be made about this service process. First, it is assumed that all non-empty states are transient, so from any initial state the empty states will eventually be reached. Furthermore, it is assumed that when the MSP reaches the empty states, it enters state $(0, j)$ with known probability ϕ_j, where it remains. For this it is sufficient (but not necessary) that

$$B(0) = O, \quad B_\ell(\ell) = B_\ell(\ell) e\phi, \quad \text{for all } \ell \ge 1.$$

For a queue, this implies that the MSP is idle if and only if the queue is empty and during an idle period the MSP is in state j with probability ϕ_j. These two assumptions

are essential for the algorithm (to satisfy assumption 0 on page 24 and derive equation (2.58)).

Two further assumptions are made that are not necessary for the algorithm but that do greatly simplify the balance equations. The first is that customers are not served in batches:

$$R_\ell(n) = O, \quad \text{for all } n \geq 0 \text{ and } \ell \geq 2.$$

Without this assumption a more complicated routing process would need to be defined. The second assumption is that an arriving customer does not cause a change in the service process.

Despite these assumptions, the remaining class of service processes is still a very rich class and includes the examples of MSPs listed below. Here, the vectors e_1 and e_2 are the unit vectors of size 2.

1) Independent phase-type service-time distributions:

$$\begin{aligned} &B_0(n) - \bar{B}(n) = T, && \text{for all } n \geq 1, \\ &B_1(n) = T^0\alpha, && \text{for all } n \geq 1, \\ &\phi = \alpha. \end{aligned}$$

2) As 1), but with set-up times: after each idle period the first service time has initial distribution α_1, all other service times have initial distribution α_2:

$$\begin{aligned} &B_0(n) - \bar{B}(n) = T, && \text{for all } n \geq 1, \\ &B_1(1) = T^0\alpha_1, && \\ &B_1(n) = T^0\alpha_2, && \text{for all } n \geq 2, \\ &\phi = \alpha_1. \end{aligned}$$

Any pair of phase-type distributions $(\tilde{T}_1, \tilde{\alpha}_1)$ and $(\tilde{T}_2, \tilde{\alpha}_2)$ can be modeled by a single generator T with two different initial distributions α_1 and α_2, by taking T block diagonal ($T = e_1e_1^T \otimes \tilde{T}_1 + e_2e_2^T \otimes \tilde{T}_2$, $\alpha_1 = e_1 \otimes \tilde{\alpha}_1$ and $\alpha_2 = e_2 \otimes \tilde{\alpha}_2$).

Examples 1 and 2 are special cases of sequences of phase-type distributions $\{(T_\ell, \alpha_\ell),\ \ell \geq 1\}$. Because the number of phases J of the MSP is finite, such sequences must, after a number of set-up distributions, start repeating itself, either in a deterministic or in a probabilistic sense. Because the MSP starts anew at the beginning of each busy period, also mixtures of sequences are possible. This could be used to model for example a situation where at the beginning of each busy period, either a fast or a slow server is chosen. Other examples of MSPs are multi-server queues:

3) c identical exponential servers with rate μ:

$$\begin{aligned} &B_0(n) - \bar{B}(n) = -\mu \min\{c, n\}, && \text{for all } n \geq 1, \\ &B_1(n) = \mu \min\{c, n\}, && \text{for all } n \geq 1, \\ &\phi = 1. \end{aligned}$$

4) c identical phase-type servers:

$$\begin{array}{ll} B_0(n) - \bar{B}(n) = \sum_{s=1}^{n} I_{s-1} \otimes T \otimes I_{c-s}, & \text{for all } 1 \leq n \leq c, \\ B_0(n) - \bar{B}(n) = \sum_{s=1}^{c} I_{s-1} \otimes T \otimes I_{c-s}, & \text{for all } c < n, \\ B_1(n) = [(\sum_{s=1}^{n} I_{s-1} \otimes T^0 \otimes I_{n-s})(I_{n-1} \otimes \alpha)] \otimes I_{c-n}, & \text{for all } 1 \leq n \leq c, \\ B_1(n) = \sum_{s=1}^{c} I_{s-1} \otimes T^0 \alpha \otimes I_{c-s}, & \text{for all } c < n, \\ \phi = \alpha \otimes \ldots \otimes \alpha. & \end{array}$$

Here, I_s is a unit-matrix with size ℓ^s for $0 \leq s \leq c$, where ℓ is the number of phases of the phase-type distribution. The transitions are defined such that if there are no waiting customers in the queue ($n \leq c$), then the first n servers are active and the other $c - n$ servers are idle; when server s completes service, then the customers at servers $s+1, \ldots, n$ move to servers $s, \ldots, n-1$, continuing service in the same phase. Server n becomes idle, with the service phase distributed according to α. This way, no variables need to be added to keep track of which servers are active, and when a new customer arrives, service can be started without changing the state of the MSP.

Modelling non-identical servers would require an extra phase for each server to indicate whether it is active or not. Then, an arriving customer would need to activate a server, which violates the assumption that arrivals of customers can not change the state of the MSP. The number of phases of the MSP is equal to ℓ^c, so infinite-server queues can only be modeled in the exponential case.

Markovian routing process

The routing is Markovian: after service completion at queue s the customer joins queue t with probability χ_{st} and leaves the network with probability χ_{s0} ($1 \leq s, t \leq S$):

$$X \doteq [\chi_{st}], \quad \chi_0 \doteq [\chi_{s0}], \quad Xe + \chi_0 = e.$$

Markovian network process

The above described arrival, service and routing processes determine the network process $\{(\mathcal{N}_t, \mathcal{J}_t),\ t \geq 0\}$ on state space

$$\Omega \doteq \left\{(n, j) \middle| n \in \mathbb{N}^S, 1 \leq j_s \leq J_s \text{ for all } 0 \leq s \leq S\right\}.$$

The state $(n, j) \in \Omega$ denotes that there are n_s customers at queue s, the arrival process is in state j_0 and the service process at queue s is in state j_s ($1 \leq s \leq S$). To introduce matrix notation, it is convenient to map the $(2S+1)$-dimensional state space Ω onto the $(S+1)$-dimensional state space

$$\bar{\Omega} \doteq \left\{(n, i) \middle| n \in \mathbb{N}^S, 1 \leq i \leq I\right\},$$

where $I = J_0 \times \ldots \times J_S$. This can be done 'lexicographically' with the mapping

$$i(j) = 1 + \sum_{s=0}^{S} (j_s - 1)\bar{J}_{s+1},$$

where $\bar{J}_s = J_s \times \ldots \times J_S$ for $0 \leq s \leq S$ and $\bar{J}_S + 1 = 1$. The reverse mapping is

$$j_s(i) = 1 + \left[(i-1) \bmod \bar{J}_s\right] \operatorname{div} \bar{J}_{s+1}, \quad \text{for all } 0 \leq s \leq S.$$

This mapping determines the network process $\{(\mathcal{N}_t, \mathcal{I}_t),\ t \geq 0\}$ on state space $\bar{\Omega}$. If the network is stable, the steady-state probabilities of this process

$$p_i(n) \doteq \lim_{t\to\infty} \mathbb{P}\left\{(\mathcal{N}_t, \mathcal{I}_t) = (n, i)\right\}$$

exist for all $(n, i) \in \bar{\Omega}$. They are independent of the initial state $(\mathcal{N}_0, \mathcal{I}_0)$ and uniquely determined by the balance and normalization equations. For any matrix A, let double brackets denote the Kronecker product

$$[\![A]\!]_s \doteq I_{J_0 \times \ldots \times J_{s-1}} \otimes A \otimes I_{J_{s+1} \times \ldots \times J_S}, \quad \text{for all } 0 \leq s \leq S.$$

Then the balance equations are

$$\begin{aligned} p(n)\left\{ \left[\!\left[\bar{A} - A_0\right]\!\right]_0 + \sum_{s=1}^{S} \left[\!\left[\bar{B}_s(n_s) - B_{s0}(n_s) - \chi_{ss} B_{s1}(n_s)\right]\!\right]_s \right\} & \\ = \sum_{m=1}^{M} p(n - b_m) \quad [\![A_m]\!]_0 & \\ + \sum_{s=1}^{S} \sum_{\substack{t=0 \\ t \neq s}}^{S} \chi_{st} \ p(n + e_s - e_t) \ [\![B_{s1}(n_s + 1)]\!]_s, & \end{aligned} \tag{2.53}$$

for $n \in \mathbb{N}^S$, with $p(n) = o$ if $n \notin \mathbb{N}^S$. The matrices A_0 and $B_{s0}(n_s)$ on the LHS correspond to changes in the arrival and service processes without arrival or service completion. The matrix $\chi_{ss} B_{s1}(n_s)$ corresponds to the event that a customer joins the same queue again, which does not change the queue lengths. The first expression on the RHS corresponds to an arrival and the second to a service completion followed by either a departure from the network ($t = 0$) or a transition to another queue ($t \neq 0, s$).

2.8.2 The transformed network process

Applying the PSA as described in section 2.3 to the networks described above, comes down to transforming the arrival and routing process of the network as follows. Let $r_m = b_m e$ denote the number of customers in batch b_m, for $1 \leq m \leq M$. Replace the probability matrices Q_m by

$$\begin{aligned} Q_m(\gamma) &= \gamma^{r_m} Q_m, \quad \text{for all } 1 \leq m \leq M, \\ Q_0(\gamma) &= Q_0 + \sum_{m=1}^{M} (1 - \gamma^{r_m}) Q_m = ee^T - \sum_{m=1}^{M} \gamma^{r_m} Q_m. \end{aligned}$$

The probability of an arrival of r customers is multiplied by γ^r, and the remaining probability mass is added to the probability of no arrival, so $\sum_{m=0}^{M} Q_m(\gamma) = ee^T$ for $\gamma \in [0,1]$ and $Q_m(1) = Q_m$ for $1 \leq m \leq M$. For smaller γ, less arrivals occur on average at each queue and for $\gamma = 0$ no arrivals occur at all. This transformation comes down to rejecting arriving customers. Larger batches are more likely to be rejected.

Let X^d denote the diagonal matrix with the same diagonal as the routing matrix X, and X^o the off-diagonal part of X, so $X^d + X^o = X$. In the transformed network process, the routing probabilities X and χ_0 are replaced by

$$X(\gamma) = X^d + \gamma X^o, \quad \chi_0(\gamma) = \gamma\chi_0 + (1-\gamma)(I - X^d)e.$$

The probability to go from queue s to queue t, with $t \neq s$, is multiplied by γ, and the remaining probability mass is added to the probability to leave the network, so $X(\gamma)e + \chi_0(\gamma) = e$, for all $\gamma \in [0,1]$, and $X(1) = X$, $\chi_0(1) = \chi_0$. For smaller γ, the customers on average visit less queues, because after each service completion they leave the network with higher probability. For $\gamma = 0$, customers only visit a single queue, possibly several times.

The arrival rates at the queues from outside the network are equal to the elements of the vector

$$\lambda(\gamma) = \sum_{m=1}^{M} \gamma^{r_m}\, b_m\, \xi A_m e, \tag{2.54}$$

where ξ is the steady-state distribution of the Markov process underlying the $MMAP$, which can be calculated from $\xi(\bar{A} - A) = o, \xi e = 1$. The arrival rates both from outside the network and from the other queues are equal to the elements of the vector

$$\begin{aligned} \nu(\gamma) &= \lambda(\gamma)\,[I - X(\gamma)]^{-1} \\ &= \lambda(\gamma)(I - X^d)^{-1} \left\{ \sum_{r=0}^{\infty} \gamma^r \left[X^o (I - X^d)^{-1} \right]^r \right\}. \end{aligned} \tag{2.55}$$

This power series converges and since it has only non-negative coefficients, $\nu(\gamma)$ is increasing in γ: for larger γ there are more arrivals and customers leave the network less often. The service process at each queue does not depend on γ. From this it is easily seen that if the original network is stable, the transformed network is also stable for all γ in [0,1]. The steady-state probabilities are, up to a multiplicative constant, uniquely determined by the balance equations:

$$\begin{aligned} & p(\gamma,n) \left\{ \left[\!\left[\bar{A} - A \right]\!\right]_0 + \sum_{s=1}^{S} \left[\!\left[\bar{B}_s(n_s) - B_{s0}(n_s) - \chi_{ss} B_{s1}(n_s) \right]\!\right]_s \right\} \\ = & \sum_{m=1}^{M} \gamma^{r_m} \{p(\gamma, n - b_m) - p(\gamma,n)\} \left[\!\left[A_m \right]\!\right]_0 \\ + & \sum_{s=1}^{S} \sum_{\substack{t=1 \\ t \neq s}}^{S} \gamma\chi_{st} \{p(\gamma, n + e_s - e_t) - p(\gamma, n + e_s)\} \left[\!\left[B_{s1}(n_s+1) \right]\!\right]_s \\ + & \sum_{s=1}^{S} (1 - \chi_{ss}) \, p(\gamma, n + e_s) \left[\!\left[B_{s1}(n_s+1) \right]\!\right]_s, \end{aligned} \tag{2.56}$$

for $n \in \mathbb{N}^S$, with $p(\gamma, n) = o$ if $n \notin \mathbb{N}^S$.

2.8.3 The Power-Series Algorithm

The transformed model does not satisfy assumption 0 from section 2.3. All non-empty states are transient, but the empty states consist of $J_1 \times \ldots \times J_S$ recurrent classes, since the states of the service processes do not change when there are no customers. Below, it will be shown that the third approach from section 2.7.5 can be used to overcome this problem. Assumption 1 is satisfied if the state dependence of the service processes at all queues is limited. A sufficient condition is that a constant $N \geq 0$ exists such that

$$B_{s0}(n) = B_{s0}(N) \text{ and } B_{s1}(n) = B_{s1}(N), \quad \text{for all } n \geq N \text{ and } 1 \leq s \leq S.$$

In that case, assumption 1' is satisfied with set $\mathcal{S} = \{0, \ldots, N\}^S$ and function $f_s(n) = \min\{N, n_s\}$. This assumption is very weak and immediately implies assumptions 1 and 3. Assumption 2 is satisfied for all usual performance measures, like the moments and (co)variances of the queue length (2.18). So under very weak conditions, the steady-state probabilities and performance measures are analytic functions of γ at $\gamma = 0$ and they can be represented by their power-series expansions

$$p(\gamma, n) = \sum_{r=ne}^{\infty} \gamma^r u(r, n), \quad \text{for all } n \in \mathbb{N}^S. \tag{2.57}$$

The transformed network process is such that the coefficient vectors $u(r, n)$ of these power-series expansions can be calculated recursively by the PSA. This will be shown first for the empty states, and then for the non-empty states.

The $MMAP$ is not influenced by the queue-length and the service processes. Summing the steady-state probabilities of the network over all possible queue lengths and states of the service processes must therefore render the steady-state distribution of the arrival process:

$$\sum_{n \in \mathbb{N}^S} p(\gamma, n) \left[I_{J_0} \otimes e \right] = \xi,$$

where e is a vector of ones with size $J_1 \times \ldots \times J_S$. When the network is empty, the states of the service processes at all queues are distributed according to the initial distributions ϕ_s:

$$p(\gamma, o) = \{p(\gamma, o) \left[I_{J_0} \otimes e \right]\} \otimes \phi,$$

where $\phi = \phi_1 \otimes \ldots \otimes \phi_S$. Combining both renders

$$p(\gamma, o) = \left\{ \xi - \sum_{n > o} p(\gamma, n) \left[I_{J_0} \otimes e \right] \right\} \otimes \phi.$$

Substituting the power-series expansions (2.57) and equating the coefficients of corresponding powers of γ on either side of the equality sign shows that the coefficients of the expansions of the empty states $p(\gamma, o)$ satisfy

$$\begin{aligned} u(0,o) &= \xi \otimes \phi, \\ u(r,o) &= -\left\{ \sum_{0<ne\leq r} u(r,n)\,[I_{J_0} \otimes e] \right\} \otimes \phi, \quad \text{for all } r \geq 1. \end{aligned} \tag{2.58}$$

Notice that $u(0,o)e = 1$, so for $\gamma = 0$ all probability mass is at the empty states.

Substituting the power-series expansions (2.57) in the balance equations (2.56) and equating the coefficients of corresponding powers of γ on either side of the equality sign, shows that the coefficients of the power-series expansions of the non-empty states satisfy the following recurrence relations:

$$\begin{aligned} & u(r,n)\left\{ \left[\!\left[\bar{A} - A\right]\!\right]_0 + \sum_{s=1}^{S} \left[\!\left[\bar{B}_s(n_s) - B_{s0}(n_s) - \chi_{ss} B_{s1}(n_s)\right]\!\right]_s \right\} \\ &= \sum_{m=1}^{M} \{u(r-r_m, n-b_m) - u(r-r_m, n)\} \quad [\![A_m]\!]_0 \\ &+ \sum_{s=1}^{S} \sum_{\substack{t=1 \\ t\neq s}}^{S} \chi_{st} \{u(r-1, n+e_s-e_t) - u(r-1, n+e_s)\} \quad [\![B_{s1}(n_s+1)]\!]_s \\ &+ \sum_{s=1}^{S} (1-\chi_{ss}) \quad u(r, n+e_s) \quad [\![B_{s1}(n_s+1)]\!]_s, \end{aligned} \tag{2.59}$$

for $n \in \mathbb{N}^S$, $r \geq ne$, and with $u(r,n) = o$ if $r < ne$ or $n \notin \mathbb{N}^S$. At the end of section 2.3.1, the equivalence of assumptions 0' and 0 was shown. This implies that the matrix on the LHS,

$$\left[\!\left[\bar{A} - A\right]\!\right]_0 + \sum_{s=1}^{S} \left[\!\left[\bar{B}_s(n_s) - B_{s0}(n_s) - \chi_{ss} B_{s1}(n_s)\right]\!\right]_s,$$

is invertible for all $n \in \mathbb{N}^S \backslash \{o\}$, since all non-empty states are transient for $\gamma = 0$. The coefficients $u(\tilde{r}, \tilde{n})$ on the RHS either have $\tilde{r} < r$ or have $\tilde{r} = r$ and $\tilde{n}e > ne$. All coefficients $u(\tilde{r}, \tilde{n})$ with $\tilde{n}e > \tilde{r}$ are zero because of the order property (2.7). Together, this implies that the coefficients of the expansions of the steady-state probabilities up to the R-th power of γ can be calculated recursively, for increasing values of r and, for each fixed r, for decreasing values of ne, starting with $ne = r$:

Power-Series Algorithm

Calculate $u(0,o)$ from (2.58),
for $r = 1, \ldots, R$ do
    for $N = r, \ldots, 1$ do
        for all $n \in \mathbb{N}^S$ with $ne = N$ do
            Calculate $u(r,n)$ from (2.59),
    Calculate $u(r,o)$ from (2.58).

The storage requirements of the algorithm can be substantially reduced if the maximal batch size $\bar{r} = \sup_m r_m$ is finite. From (2.59) it can be seen that in step r of the algorithm, coefficients $u(\tilde{r}, \tilde{n})$ with $\tilde{r} < r - \bar{r}$ are no longer needed to calculate the remaining coefficients. The same is true for coefficients with $\tilde{r} = r - \bar{r}$ and $\tilde{n}e > ne + 1$, because customers leave one at a time.

The MSP at queue s depends only on the queue length at queue s. The state dependence could be made more general, not only for the service processes but also for the arrival and routing processes. The state dependence of the service and routing processes must be such that, for $\gamma = 0$, all non-empty states of the transformed network process are transient, so that eventually the network will be totally empty. Then the service processes must be stopped and the distribution ϕ over the service phases must be known (but ϕ need not be the Kronecker product $\phi_1 \otimes \ldots \otimes \phi_S$). The Markov process underlying the $MMAP$ must be state independent to calculate ξ, but the probability matrices Q_m *can* be state dependent. This way, the coefficients of the empty states can still be calculated from (2.58) and the coefficients of the non-empty states can be calculated from (2.59) if the parameters are replaced by the state dependent parameters.

2.8.4 Examples

Two examples are presented to illustrate the flexibility and strength of the PSA. The first example below considers the optimal order of queues in series. The second example shows that the total number of customers in cyclic networks depends mainly on the first moment of the service-time distributions. The mapping and pole extrapolation were not used because the power series were regular enough to obtain convergence by means of only the epsilon algorithm, which performed better than the theta and Levin algorithms (with the relative variation criterion and $\kappa = 5$).

Optimal order of queues in series

An important design problem in queueing theory is how to order queues in series. How can the mean total queue length, or equivalently the mean sojourn time, be minimized for given arrival process and service-time distributions?

Figure 2.18

Exact analysis is in general very difficult, even for only 2 queues. Whitt [136] proposes a heuristic based on the approximation of the departure process of each queue by a renewal process, characterized by the first two moments of the renewal interval. Greenberg and Wolff [67] proposed a heuristic based on light-traffic behaviour and gave some examples where both heuristics did not give the same solution. They warned that extreme caution must be used in applying approximations to develop design procedures and stated that a heuristic based only on mean and squared coefficient of variation of the

various distributions cannot be expected to work well. However, they did not indicate how large the difference in performance of both suggested solutions would be. Using simulation, Suresh and Whitt [133] later reported that when the heuristics disagree, the difference is negligible. This conclusion is supported by the example below, now using the PSA.

Consider the following model. According to a Poisson process with rate λ, customers arrive to obtain service from two servers in series. Both servers have an Erlang(2) service-time distribution, one with mean 1 and the other with mean 4. Should the customers first visit the fast or the slow server and does the optimal order depend on the arrival rate λ ? According to Whitt [136] the optimal order is to visit the fast server first; Greenberg and Wolff [67] suggest that, in light traffic, the slower server should be visited first. In the table below, the expected total number of customers is shown for both orders and different loads. The value of μ_s is the mean service time at queue s. The indicated load is the load of the slower server and corresponds to arrival rates 0.4, 1.2, 2.0, 2.8 and 3.6.

μ_1, μ_2	$\rho = 0.1$	$\rho = 0.3$	$\rho = 0.5$	$\rho = 0.7$	$\rho = 0.9$
1, 4	0.1337	0.4745	1.009	2.118	7.231
4, 1	0.1335	0.4734	1.005	2.111	7.218

Table 2.4: Mean total queue length for different orders and loads.

To visit the slower server first is better for all loads, but clearly the difference is negligible. Numerical experiments indicate that this conclusion remains valid when the interarrival time distribution is Erlang or hyper-exponential, when the service time distributions are both exponential or hyper-exponential or when the means of the service-time distributions are taken further apart, just as long as the service time distributions have equal coefficient of variation. If the interarrival times and the service times are all exponential, then the system is a product-form network and the total queue length is independent of the order of the queues. In fact, this insensitivity holds for arbitrary arrival processes if the service time distributions are all exponential [135] or all deterministic [57].

Convergence of the power series was slowest for the model with $\rho = 0.9$ and the fast server first. In the table below the original series $V_R(1) \doteq \sum_{r=0}^{R} v_r$ and the series after applying the epsilon algorithm $\epsilon[V_R(1)]$ are shown.

R	1	5	10	20	40	60
$V_R(1)$	0.2250	1.766	3.245	4.877	6.410	6.945
$\epsilon[V_R(1)]$	0.2250	1.766	7.362	7.232	7.231	7.231

Table 2.5: Approximations at different truncation levels.

The original series seems to converge monotonically, but after applying the epsilon algorithm, convergence is much faster. Convergence is usually slower if the load of the

original network is higher or if the parameters of the model are more extreme. For example, hyper-exponential distributions result in slower convergence than Erlang distributions. In general, convergence can only be guaranteed for light traffic and careful analysis of the behaviour of the power series is essential to validate the results.

Insensitivity for higher moments

Consider the following model. Customers arrive according to a process that is a mixture of single arrivals and fork arrivals. At each queue, single customers arrive according to independent Poisson processes, all with rate λ_1. According to another independent Poisson process with rate λ_2, customers arrive simultaneously at all queues:

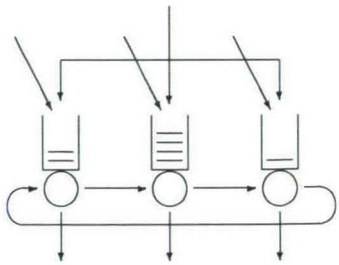

Figure 2.19

$$\begin{array}{lll} I = 1, & \alpha_{11} = S\lambda_1 + \lambda_2, & \\ M = S+1, & b_m = e_m \ (1 \le m \le S), & b_{S+1} = e, \\ q_{0,1,1} = 0, & q_{m,1,1} = \frac{\lambda_1}{S\lambda_1+\lambda_2} \ (1 \le m \le S), & q_{S+1,1,1} = \frac{\lambda_2}{S\lambda_1+\lambda_2}. \end{array}$$

The routing is such that, after service completion at a queue, customers either leave the network with probability p, or go to the next queue with probability $1-p$:

$$\chi_{st} = \begin{cases} p & \text{if } t = 0, \\ 1-p & \text{if } t = (s \bmod S) + 1, \\ 0 & \text{otherwise.} \end{cases}$$

Depending on the coefficient of variation, the service-time distributions at the various queues are either Erlang, exponential or hyper-exponential with balanced means. In table 2.6 below, the probability of an empty network and the mean queue lengths are given for four different models with 3 queues, $\lambda_1 = \lambda_2 = 0.09, p = 0.2$ and $\mu_1 = \mu_2 = \mu_3 = 1$. This way, all queues have identical load $\rho_1 = \rho_2 = \rho_3 = 0.9$. The difference between the four models is in the variance σ_s^2 of the service-time distributions at the queues, but their sum is equal to 3 for all models.

$\sigma_1^2, \sigma_2^2, \sigma_3^2$	$\mathbb{P}\{\mathcal{N} = o\}$	$\mathbb{E}\{\mathcal{N}_1\}$	$\mathbb{E}\{\mathcal{N}_2\}$	$\mathbb{E}\{\mathcal{N}_3\}$	$\mathbb{E}\{\mathcal{N}e\}$
$1, 1, 1$	0.0033	10.28	10.28	10.28	30.84
$2, \frac{1}{2}, \frac{1}{2}$	0.0038	12.95	9.928	7.760	30.63
$1\frac{1}{2}, 1, \frac{1}{2}$	0.0035	11.39	10.97	8.426	30.78
$\frac{1}{2}, 1, 1\frac{1}{2}$	0.0035	9.141	9.701	11.95	30.79

Table 2.6: Performance measures for different variances.

The convergence of the power series of $\mathbb{E}\{\mathcal{N}e\}$ in the second model was poorest:

R	1	5	10	20	40	60
$V_R(1)$	0.2700	2.007	4.326	8.387	14.79	19.37
$\epsilon[V_R(1)]$	0.2700	−0.2523	0.1951	32.41	30.59	30.63

Table 2.7: Approximations at different truncation levels.

Again, both series seem to converge, but more coefficients need to be calculated to stabilize than in table 2.5.

It can be seen that the mean queue length of each queue is increasing in the variance of the service-time distribution of both the queue itself and the preceding queue. From the last column it can be seen that the expected total number of customers in the network is approximately equal for all four models. For $p = 1$, this follows immediately from the Pollaczek-Khintchine formula, because then all queues are $M/G/1$ queues with identical load and mean service time, so:

$$\mathbb{E}\{\mathcal{N}e\} = S\rho_1 + \frac{\rho_1^2}{2(1-\rho_1)}\left(1 + \frac{1}{\mu_1^2}\sum_{s=1}^{S}\sigma_s^2\right).$$

The variances of the service-time distributions were chosen such that their sum is equal for all four models. The individual variances and the shapes of the distributions vary. Numerical experiments indicate that the property that the mean total queue length is mainly determined by the *sum* of the variances also holds for more general models, namely for networks with a symmetric arrival process, generalized-cyclic routing and equal loads at the different queues. Here, a symmetric arrival process can have an arbitrary underlying Markov process. If b_m is a possible arrival then each permutation of b_m is also a possible arrival and if $b_{\tilde{m}}$ is a permutation of b_m, then $Q_{\tilde{m}} = Q_m$. More intuitively, a symmetric arrival process is such that at each arrival of a batch, an arrival of any permutation of this batch would have been equally likely. A generalized-cyclic routing matrix X is a routing matrix such that

$$\chi_{s,t} = \chi_{(s \bmod S)+1,(t \bmod S)+1}, \quad \text{for all } 1 \leq s, t \leq S.$$

For example, if $S = 4$, then a generalized-cyclic routing matrix is of the form

$$X = \begin{pmatrix} p_1 & p_2 & p_3 & p_4 \\ p_4 & p_1 & p_2 & p_3 \\ p_3 & p_4 & p_1 & p_2 \\ p_2 & p_3 & p_4 & p_1 \end{pmatrix}.$$

If the arrival process is symmetric and the routing generalized-cyclic, then the loads at the queues are identical if the mean service times are identical. For such networks the following hypothesis can be formulated: *for networks of $./G/1$ queues, with a symmetric arrival process, generalized-cyclic routing and equal loads at all queues, the expected total*

number of customers in the network is mainly determined by the sum of the variances of the service-time distributions, and not so much by their shapes. Of course such a hypothesis could never be proved by the PSA, but it can be used to evaluate various 'randomly' chosen models and models that are likely to be counter-examples.

2.8.5 Alternative transformations

A disadvantage of the transformation described in section 2.8.2 is that the transformation parameter γ does not always have a physical interpretation. For the first model with queues in series in figure 2.8.4, γ can be interpreted as the probability to accept an arriving customer. In the cyclic model in figure 2.8.4, customers from the different Poisson processes are accepted with different probabilities. In general, the interpretation is not obvious. In this section two alternative transformations will be considered. These alternatives have the advantage that the transformation parameter can be interpreted as the load of the network. However, they can only be applied to a subclass of the networks and the second approach is computationally less attractive.

The usual definition of the load of a single-server queue is the probability that the server at this queue is busy. It is not obvious how to generalize this into a scalar measure of the load of a network of queues. The maximal load of the individual queues seems more reasonable than the average load or the probability that any of the servers is busy.

Although a good definition of the load of a network is not obvious, the transformation parameter γ is not a good measure. The load of the individual queues does increase with γ, as can be seen from the arrival rates in formula (2.55). However, because the transition rates of arrivals of larger batches of customers are multiplied by higher powers of γ, the dependence is non-linear. A more reasonable measure of the load can be obtained when a different transformation is applied. In the approach above, the batch-size probabilities are transformed. A more straightforward approach would be to only multiply all the transition rates A of the Markov process underlying the $MMAP$ by γ. The new arrival process $MMAP_\gamma$ is equal to the original arrival process, but on a different time scale. For larger values of γ the arrival process moves faster, so there are more arrivals. The vector of arrivals per time unit is now equal to

$$\nu(\gamma) = \gamma \left[\sum_{m=1}^{M} b_m \, \xi A_m e \right] [I - X]^{-1} , \tag{2.60}$$

instead of formula (2.55). The dependence on γ is now linear and much simpler. Since the service process is independent of γ, the load of each queue is also linear in γ. Let γ^* be such that all queues, and therefore also the network, are stable for $0 \leq \gamma < \gamma^*$ and unstable for $\gamma \geq \gamma^*$. Then the load of the network process with arrival process $MMAP_\gamma$ could be defined as γ/γ^*, the reciprocal of the factor with which the arrival process must

be made faster to make the network unstable. For this type of models this is identical to the maximal load of the individual queues.

The difference between the transformation suggested in section 2.8.2 and the time-scale transformation suggested above is best illustrated by applying both to the $GI/G/1$ queue. Both transformed processes are again $GI/G/1$ queues, but now with different interarrival-time distributions. The transformation in section 2.8.2 would accept each arriving customer with probability γ and reject the customer with probability $1-\gamma$. The transformed interarrival-time distribution is the convolution of a geometric number of original interarrival-time distributions. With the time-scale transformation, the transformed interarrival-time distribution has the same shape as the original distribution, but with a different scale parameter: the mean interarrival time is divided by γ. Both approaches are identical for the $M/G/1$ queue.

So, why use the complicated transformation from section 2.8.2? Because, in general, just changing the time scale of the arrival process does not lead to a recursive algorithm for the computation of the coefficients of the steady-state distribution of the transformed process. It renders a recursive algorithm only if the maximal batch size $B = \sup_m(b_m e)$ is finite and if the network is a feed-forward network. It was used in [74] for the analysis of the $BMAP/PH/1$ queue with finite maximal batch size. With the time-scale transformation, the balance equations of the transformed process are

$$\begin{aligned} p(\gamma,n)&\left\{\gamma\left[\!\left[\bar{A}-A_0\right]\!\right]_0+\sum_{s=1}^{S}\left[\!\left[\bar{B}_s(n_s)-B_{s0}(n_s)-\chi_{ss}B_{s1}(n_s)\right]\!\right]_s\right\}\\ &= \sum_{m=1}^{M} \gamma \quad p(\gamma,n-b_m) \quad [\![A_m]\!]_0\\ &+ \sum_{s=1}^{S}\sum_{\substack{t=0\\ t\neq s}}^{S} \chi_{st} \quad p(\gamma,n+e_s-e_t) \quad [\![B_{s1}(n_s+1)]\!]_s, \end{aligned}$$

for $n \in \mathbb{N}^S$, and with $p(\gamma,n) = o$ if $n \notin \mathbb{N}^S$. The difference with the balance equations (2.53) of the original process is that the transitions of the arrival process are multiplied by γ. Replacing the probabilities by their power-series expansions and equating the coefficients of corresponding powers of γ renders the recursive equations

$$\begin{aligned} u(r,n)\sum_{s=1}^{S}\left[\!\left[\bar{B}_s(n_s)-B_{s0}(n_s)-\chi_{ss}B_{s1}(n_s)\right]\!\right]_s &= -u(r-1,n)\left[\!\left[\bar{A}-A_0\right]\!\right]_0\\ + \sum_{m=1}^{M} \quad u(r-1,n-b_m) \quad [\![A_m]\!]_0&\\ + \sum_{s=1}^{S}\sum_{\substack{t=0\\ t\neq s}}^{S} \chi_{st} \quad u(r,n+e_s-e_t) \quad [\![B_{s1}(n_s+1)]\!]_s&, \end{aligned}$$

for $n \in \mathbb{N}^S$ and $r \geq ne$, and with $u(r,n) = o$ if $r < ne$ or $n \notin \mathbb{N}^S$. The order of the steady-state probabilities is equal to the number of transitions in the $MMAP$ required to reach the particular state:

$$p(\gamma,n) = O\left(\gamma^{\lceil (ne)/B\rceil}\right), \quad \text{for all } \gamma \downarrow 0 \text{ and } n \in \mathbb{N}^S, \tag{2.61}$$

where $\lceil x \rceil$ denotes x rounded upward (see section 1.4). If the steady-state probabilities are analytic, then the initial coefficients are zero:

$$p(\gamma, n) = \sum_{r=\lceil (ne)/M \rceil}^{\infty} \gamma^r u(r, n), \quad \text{for all } n \in \mathbb{N}^S.$$

If the maximal batch size B were infinite, then all steady-state probabilities of non-empty states would be in $\mathcal{O}(\gamma)$, for $\gamma \downarrow 0$. The power-series expansions would all have a linear coefficient $u(1,.)$ and it would be impossible to calculate this infinite number of coefficients. If the maximal batch size is finite, then $u(r,n) = 0$ if $ne > rB$. The number of non-zero vectors of r-th order coefficients is at most

$$\#\left\{n \in \mathbb{N}^S \,|ne \leq rB\right\} = \binom{Br+S}{S}. \tag{2.62}$$

The coefficient $u(r,n)$ is a function of coefficients $u(r-1,.)$ but also of coefficients $u(r, n+e_s-e_t)$ if $\chi_{st} > 0$. For each r and fixed total queue length, a set of linear equations needs to be solved. The number of variables in each set of equations is $\binom{Br+S-1}{S-1}$. This increases fast with r, which hampers the calculation of the coefficients unless the set of equations has some special structure that makes it easy to solve. Such a special structure exists if the network is a feed-forward network. Then the queues can be ordered such that $\chi_{st} = 0$ for all $s < t$. For fixed r, first calculate the coefficients with only customers at the last queue, then the coefficients with only customers at the last two queues and so on. The coefficients of the empty state are calculated in a similar way as before.

If the network is not feed-forward and the time-scale transformation is used, then sets of equations of increasing size need to be solved. This can be prevented by transforming the routing process as well. This can be done as follows:

$$X(\sigma) = X^d + \sigma X^o, \quad \chi_0(\sigma) = \sigma\chi_0 + (1-\sigma)(I - X^d)e,$$

with $0 \leq \sigma \leq 1$. This is the same transformation as in section 2.8.2, but with a transformation parameter σ, different from γ. For smaller values of σ, customers will sooner leave the network instead of visiting other stations. Together with the time-scale transformation of the $MMAP$, this process has balance equations

$$\begin{aligned}
p(\gamma,\sigma,n)&\left\{\gamma\left[\!\left[\bar{A} - A_0\right]\!\right]_0 + \sum_{s=1}^{S}\left[\!\left[\bar{B}_s(n_s) - B_{s0}(n_s) - \chi_{ss}B_{s1}(n_s)\right]\!\right]_s\right\} \\
= &\sum_{m=1}^{M} \quad \gamma \quad p(\gamma,\sigma,n-b_m) \quad [\![A_m]\!]_0 \\
+ &\sum_{s=1}^{S}\sum_{\substack{t=1\\t\neq s}}^{S} \quad \sigma\chi_{st} \quad \{p(\gamma,\sigma,n+e_s-e_t) \\
& \qquad\qquad\qquad -p(\gamma,\sigma,n+e_s)\} \quad [\![B_{s1}(n_s+1)]\!]_s \\
+ &\sum_{s=1}^{S} \quad (1-\chi_{ss}) \quad p(\gamma,\sigma,n+e_s) \quad [\![B_{s1}(n_s+1)]\!]_s,
\end{aligned}$$

for $n \in \mathbb{N}^S$, and with $p(\gamma, \sigma, n) = o$ if $n \notin \mathbb{N}^S$. The order as a function of γ is the same as in formula (2.61). If the steady-state probabilities are analytic functions of both γ and σ they can be represented by the two-dimensional power-series expansions

$$p(\gamma, \sigma, n) = \sum_{r=\lceil (ne)/M \rceil}^{\infty} \sum_{s=0}^{\infty} \gamma^r \sigma^s u(r, s, n), \quad \text{for all } n \in \mathbb{N}^S.$$

Substituting this in the balance equations and equating equal powers of both γ and σ renders the recursive equations

$$\begin{aligned} u(r,s,n) \sum_{s=1}^{S} \left[\!\left[\bar{B}_s(n_s) - B_{s0}(n_s) - \chi_{ss} B_{s1}(n_s) \right]\!\right]_s &= -u(r-1,s,n) \left[\!\left[\bar{A} - A_0 \right]\!\right]_0 \\ + \sum_{m=1}^{M} & \quad u(r-1,s,n-b_m) \quad [\![A_m]\!]_0 \\ + \sum_{s=1}^{S} \sum_{\substack{t=1 \\ t \neq s}}^{S} & \quad \chi_{st} \quad \{u(r,s-1,n+e_s-e_t) \\ & \quad -u(r,s-1,n+e_s)\} \quad [\![B_{s1}(n_s+1)]\!]_s \\ + \sum_{s=1}^{S} & \quad (1-\chi_{ss}) \quad u(r,s,n+e_s) \quad [\![B_{s1}(n_s+1)]\!]_s, \end{aligned}$$

for $n \in \mathbb{N}^S$, $r \geq ne$ and $s \geq 0$, and with $u(r,s,n) = o$ if $r < ne$, $s < 0$ or $n \notin \mathbb{N}^S$. The coefficients can be calculated for increasing values of r and s. For each fixed r and s, the coefficients can be calculated in decreasing order of ne, starting from $ne = Br$.

From the above, the following conclusions can be drawn. The time-scale transformation applies to feed-forward networks with finite maximal batch size. The major advantage of this approach is the fact that, after normalization, the transformation parameter γ can be interpreted as the usual measure for the load of the system. If the arrival process consists of Poisson processes at each queue, then this approach is identical to the approach in section 2.8.2. The Poisson processes are allowed to be state dependent, but an arrival is not allowed to coincide with a change of the supplementary variable. If the two approaches are not identical, then it is not clear which approach is preferable. The analyticity results in section 2.5 can also be obtained for the time-scale transformation. It has the advantage of the interpretation. However, if the maximal batch size is larger than 1, it has the disadvantage that for each r the number of non-zero r-th order coefficients is much larger: $\binom{Br+S}{S}$ instead of $\binom{r+S}{S}$.

The double transformation described above can also be used for non-feed-forward networks. It has the advantage that, after normalization, γ can be interpreted as the load of the system. The parameter σ is a measure of the extent to which customers travel between the different queues. However, the advantage of this interpretation is less convincing than for the feed-forward networks. And the double transformation has major disadvantages. Because it involves double power series, the number of coefficients that need to be calculated will be much larger. Also the methods to accelerate the

convergence of double power series are less effective than those for univariate power series (see appendix C.4).

The disadvantage of the double power series can be avoided by using the same transformation parameter γ to transform also the routing process. The approach is then very similar to the approach in section 2.8.2. However, the maximal batch size must be finite, γ no longer has a physical interpretation and the number of non-zero coefficients increases faster with r if the maximal batch size is larger than 1. Therefore the approach in section 2.8.2 will normally be preferable.

Chapter 3

The PSA for transient analysis

The transient analysis of Markov processes is generally considered to be more difficult than the steady-state analysis. Even for very simple processes like the $M/M/1$ queue, it is not easy to describe the transient behaviour. In the previous chapter, it was shown that the PSA can be useful for steady-state analysis. In the present chapter, an attempt will be made to apply the same ideas to the transient analysis, both homogeneous and non-homogeneous. This is done by considering the transient distribution of Markov processes with initial distribution and transition rates that are analytic functions of a model parameter γ. This parameter can have any physical interpretation.

Unfortunately, for homogeneous Markov processes it will turn out that the ideas of the PSA do not lead to efficient numerical procedures. For the steady-state analysis, direct methods would require the inversion of a matrix which is a complicated operation for large matrices. Instead of this, the PSA uses a large number of simple matrix multiplications. For the transient analysis, existing direct methods already require only simple operations. The PSA merely increases the amount of work to be done. Computationally, the PSA is only helpful if one is explicitly interested in the transient distribution as a function of some model parameter γ. Also, the PSA provides interesting theoretical insights. The power-series expansions of the transient distribution as a function of both time and γ are obtained in closed form. With this closed form, it is shown that the transient distribution is an analytic function of both time and γ. For finite state spaces this is obvious, but not for infinite state spaces.

By choosing the model parameter γ equal to time, non-homogeneous Markov pro-

cesses can be considered. This provides a numerically stable way to compute the transient distribution and easily shows that if the transition rates are analytic functions of time, then also the transient distribution is an analytic function of time.

This chapter is organized in the following way. First, in section 3.1, direct methods will be discussed to find the transient distribution of Markov processes. It will be shown that what is usually called Jensen's method can be improved if knowledge about the steady-state distribution is available. The extended method has convergence uniform over time. A framework is described that includes the Taylor expansion, Jensen's method and the extended method. This framework is generalized in section 3.2 to find the transient distribution *as a function of both time and* γ. Section 3.3 shows that the obtained power-series expansions converge. By choosing γ equal to time, a method for non-homogeneous Markov processes is obtained in section 3.4.

In the previous chapter on the steady-state analysis using the PSA, a particular transformation of the transition rates was considered. The notation explicitly reflected the multidimensional nature of this transformation. The framework of the present chapter on the transient analysis allows for general transformations. Consequently, the state-space need not reflect the multidimensional nature. It can be one-dimensional, since any countable state space can be mapped onto the non-negative natural numbers. If the state space is large or infinite, it may be necessary to truncate the state space. This will be ignored here. By aggregating all truncated states into a single absorbing state, bounds on the truncation error can be obtained [107,108].

3.1 Direct methods

Let $\{\mathcal{X}_t, t \geq 0\}$ be an honest, uniformizable and homogeneous CTMP on the countable state space Ω with initial distribution ϕ, generator Q and $q_i = -q_{ii}$ for all $i \in \Omega$. The transient distribution $\pi(t)$ at time t is determined by the forward differential system

$$\pi'(t) = \pi(t)Q, \quad \pi(0) = \phi. \tag{3.1}$$

The formal solution is $\pi(t) = \phi \exp(tQ)$ (see appendix A).

3.1.1 The Taylor-series expansion

The Taylor-series expansion of the transient distribution $\pi(t) = \phi \exp(tQ)$ is equal to $\sum_{k=0}^{\infty} \frac{t^k}{k!} \phi Q^k$. An approximation can be found by truncating this expansion:

$$\begin{aligned} \pi_K(t) &= \sum_{k=0}^{K} \frac{t^k}{k!} u_k, \\ u_0 &= \phi, \quad u_k = u_{k-1}Q = \phi Q^k, \quad \text{for all } k \geq 1. \end{aligned}$$

The calculation of this approximation involves negative numbers because the diagonal of Q is negative, which can cause severe round-off problems.

The transient distribution is analytic in t, but the power-series expansion does not converge *uniformly* in t. This can be shown as follows. Since Q is uniformizable, $\|Q\| = 2\sup_i q_i$ is finite. The truncation error is bounded by

$$\|\pi(t) - \pi_K(t)\| \leq \sum_{k=K+1}^{\infty} \frac{t^k}{k!}\|\phi Q^k\| \leq \sum_{k=K+1}^{\infty} \frac{t^k}{k!}\|Q\|^k.$$

The final summation is the tail of the power-series expansion of $\exp(t\|Q\|)$. For any fixed $t \geq 0$, this error can be made arbitrarily small by choosing K large enough. Therefore, $\pi(t)$ is an entire function of t and the expansion converges point-wise, that is for each fixed t. Convergence is not uniform: for any $K \geq 0$ fixed, the approximation is a finite polynomial in t, so it diverges for $t \to \infty$, whereas $\pi(t)$ converges to $\bar{\pi}$. In other words, for each fixed truncation level K, the Taylor approximation is bad for large enough values of time.

3.1.2 Jensen's method

The use of negative numbers is avoided by Jensen's method [81]. In this approach, dummy transitions are introduced from each state i to itself with rate $q - q_i$, with $q \geq \sup_i q_i$. This way, the transition rate from each state is equal to q, and a transition to state j is made with probability $(q_{ij}/q) + 1(i = j)$. Because only dummy transitions are introduced, the transient distribution of this new process is equal to the transient distribution of the original process and it can be found by conditioning on the number of transitions up to time t. Because the transition rate is q, uniform over all states, the number of transitions has a Poisson distribution with mean qt. Conditioning on the number of transitions leads to the approximation

$$\begin{aligned} &\pi_K(t) = \sum_{k=0}^{K} e^{-qt}\frac{(qt)^k}{k!}v_k, \\ &v_0 = \phi, \quad v_k = v_{k-1}[Q/q + I] = \phi[Q/q + I]^k, \quad \text{for all } k \geq 1. \end{aligned} \tag{3.2}$$

The vector v_k is the distribution after k uniformized transitions. The method only involves positive numbers and no subtractions, which makes it quite stable.

Like with the Taylor-series expansion, the approximation does converge point-wise, but not uniformly. For any $t \geq 0$ fixed, analyticity of $\pi(t)$ again guarantees convergence for $K \to \infty$, with truncation error

$$\|\pi(t) - \pi_K(t)\| = \sum_{k=K+1}^{\infty} e^{-qt}\frac{(qt)^k}{k!}\|v_k\| = \sum_{k=K+1}^{\infty} e^{-qt}\frac{(qt)^k}{k!}.$$

Equality holds because all quantities are positive. The error is equal to the probability mass in the tail of the Poisson distribution with mean qt. This implies that, to obtain

a certain accuracy, the truncation point K is $O(qt)$. If the state space is small, this can be reduced to $O(\sqrt{qt})$, by not only truncating the summation to the right but also to the left : $\pi_{L,K}(t) = \sum_{k=L}^{K} e^{-qt}\frac{(qt)^k}{k!}v_k$ (see [47]). For K fixed and $t \to \infty$, e^{-qt} will decrease to 0 faster than the finite polynomial in t diverges. For large t, Jensen's approximation will thus be close to 0 instead of $\bar{\pi}$. Therefore:

$$\lim_{K\to\infty} \sup_{t\geq 0} \|\pi(t) - \pi_K(t)\| \geq \lim_{K\to\infty} \lim_{t\to\infty} \|\pi(t) - \pi_K(t)\| = \lim_{K\to\infty} \|\bar{\pi} - 0\| = 1,$$

so convergence is still not uniform. For each fixed truncation level K, the Jensen approximation is bad for large enough values of time.

3.1.3 Using steady-state information

Assume that the CTMP $\{\mathcal{X}_t, t \geq 0\}$ is ergodic with steady-state distribution $\bar{\pi}$, determined by $\bar{\pi}Q = 0$, $\bar{\pi}e = 1$. This implies that the discrete-time Markov chain (DTMC) at (dummy) transition moments $\{\mathcal{Y}_k, k \geq 0\}$ with transition matrix $[Q/q + I]$ is also ergodic, for large enough q. This can be seen as follows. According to the first Foster criterion [55], sufficient for ergodicity of the DTMC is the existence of a finite solution to the balance equations, together with aperiodicity. The distribution $\bar{\pi}$ is finite and satisfies the balance equations $\bar{\pi}[Q/q + I] = \bar{\pi}$. Aperiodicity holds if $q > \inf_{i\in\Omega} q_i$. So, if not all q_i are identical, then $q = \sup_{i\in\Omega} q_i$ is large enough to make the DTMC ergodic. Otherwise, q should be chosen slightly larger.

Without loss of generality, assume that the DTMC $\{\mathcal{Y}_k, k \geq 0\}$ is ergodic with steady-state distribution $\bar{\pi}$, satisfying $\bar{\pi}[Q/q + I] = \bar{\pi}$. Since v_k is the distribution of the DTMC after k transitions, it will converge to the steady-state distribution $\bar{\pi}$. Actually, the v_k are the consecutive estimates of $\bar{\pi}$ produced by the power method to find the eigenvector of the largest eigenvalue [112]. If $\bar{\pi}$ is known in advance, then it can be used as an estimate of the uncalculated v_k for large k. This leads to the following approximation:

$$\begin{aligned}
\pi_K(t) &= \sum_{k=0}^{K} e^{-qt}\frac{(qt)^k}{k!}v_k + \sum_{k=K+1}^{\infty} e^{-qt}\frac{(qt)^k}{k!}\bar{\pi} \\
&= \sum_{k=0}^{K} e^{-qt}\frac{(qt)^k}{k!}v_k + \left[1 - \sum_{k=0}^{K} e^{-qt}\frac{(qt)^k}{k!}\right]\bar{\pi} \qquad (3.3)\\
&= \bar{\pi} + \sum_{k=0}^{K} e^{-qt}\frac{(qt)^k}{k!}[v_k - \bar{\pi}].
\end{aligned}$$

The second expression shows that this approximation will be a good approximation of $\pi(t)$ for small values of t, since it is close to Jensen's approximation. The third expression shows that it will also be a good approximation for large values of t, since it is close to $\bar{\pi}$.

In fact, this approximation of the transient distribution was already introduced by Jensen [81] in the first paper on this subject. It seems to have been forgotten since then

and the term Jensen's method is usually associated with the approach in the previous section. In this chapter, the term standard Jensen method will be used for the method described in the previous section and the term extended Jensen method for the method that uses the steady-state distribution.

Analyticity again guarantees convergence for any fixed $t \geq 0$, now with error bound

$$\begin{aligned} \|\pi(t) - \pi_K(t)\| &\leq \sum_{k=K+1}^{\infty} e^{-qt} \frac{(qt)^k}{k!} \|v_k - \bar{\pi}\| \\ &= \sum_{k=K+1}^{\infty} e^{-qt} \frac{(qt)^k}{k!} \left\| [v_K - \bar{\pi}][Q/q + I]^{k-K} \right\| \\ &\leq \|v_K - \bar{\pi}\| \sum_{k=K+1}^{\infty} e^{-qt} \frac{(qt)^k}{k!}. \end{aligned} \tag{3.4}$$

For the last inequality, it is used that $[Q/q + I]$ is a stochastic matrix, so it has norm $\|Q/q + I\| = 1$. In the standard Jensen approximation, the accuracy is obtained by making the summation small, which requires the calculation of $O(qt)$ coefficients. Now, accuracy can also come from the fact that v_K converges to $\bar{\pi}$. This convergence is independent of t and geometric in K. To obtain a certain maximal error ϵ, the truncation level K is $O(\log \epsilon)$. Convergence will be slow if the subdominant eigenvalue of $[Q/q + I]$ is close to 1. Because the convergence of v_K is independent of t, the new approximation has uniform convergence:

$$\begin{aligned} \lim_{K\to\infty} \sup_{t\geq 0} \|\pi(t) - \pi_K(t)\| &\leq \lim_{K\to\infty} \|v_K - \bar{\pi}\| \sup_{t\geq 0} \sum_{k=K+1}^{\infty} e^{-qt} \frac{(qt)^k}{k!} \\ &= \lim_{K\to\infty} \|v_K - \bar{\pi}\| = 0. \end{aligned}$$

So, contrary to the Taylor and the standard Jensen approximation, the truncation level K can be chosen such that the approximation has a specified accuracy for all values of time. It seems that this uniform convergence has not been established before in the literature.

The extended Jensen method is quite similar to the steady-state-detection method [40,104]. This method is reported to render considerable computational savings when the Markov process is stiff. Basically, it uses v_K as an estimate of $\bar{\pi}$ and checks the convergence. If v_K has converged enough, then approximation (3.3) is used with $\bar{\pi}$ replaced by the estimate v_K. A disadvantage of this method is that no exact error bounds are available. Based on the geometric convergence of v_k it is possible to give approximate error bounds. It can easily be shown that the difference between the extended Jensen approximation and the steady-state-detection approximation is equal to the error bound

of the extended Jensen method in (3.4):

$$\begin{aligned}\left\|\left\{\bar{\pi}+\sum_{k=0}^{K} e^{-qt}\frac{(qt)^k}{k!}[v_k-\bar{\pi}]\right\}-\left\{v_K+\sum_{k=0}^{K} e^{-qt}\frac{(qt)^k}{k!}[v_k-v_K]\right\}\right\| \\ = \left\|[\bar{\pi}-v_K]+\sum_{k=0}^{K} e^{-qt}\frac{(qt)^k}{k!}[v_K-\bar{\pi}]\right\| \\ = \|\bar{\pi}-v_K\|\sum_{k=K+1}^{\infty} e^{-qt}\frac{(qt)^k}{k!}.\end{aligned}$$

This immediately implies that the error of the steady-state-detection method is at most $2\,\|\bar{\pi}-v_K\|$ and that it also has uniform convergence. The obvious advantage of the steady-state-detection method is that it does not require prior knowledge of $\bar{\pi}$. However, if v_k converges slowly it may be worth the computational effort to find better estimates of the steady-state distribution, either based on the computed distributions v_k or by some other method like solving the (truncated) balance equations or the PSA. The similarity of both methods and the good performance on stiff models of the steady-state-detection method shows that also the extended Jensen method will work well on stiff models.

Often, one is not interested in the transient distribution $\pi(t)$, but only in transient expectations $f(t)=\pi(t)f$, with f a column vector. Then, knowledge of the steady-state distribution $\bar{\pi}$ is not needed, but only knowledge of the steady-state expectation $\bar{f}=\bar{\pi}f$:

$$f_K(t)=\pi_K(t)f=\bar{f}+\sum_{k=0}^{K} e^{-qt}\frac{(qt)^k}{k!}\left[v_k f-\bar{f}\right].$$

Like in the steady-state-detection method, if $\bar{f}$ is not known it can be replaced by the estimate $v_K f$. In this scalar case, better estimates of $\bar{f}$ can easily be found by applying an extrapolation method to the series $v_0 f, v_1 f, \ldots, v_K f$.

3.1.4 A general framework

The three expansions of the previous sections all fall in the class of expansions of the form

$$\pi(t)=\bar{w}+e^{-qt}\sum_{k}\frac{t^k}{k!}w_k. \tag{3.5}$$

For $q=0$, this reduces to the Taylor expansion. For $q=\sup_i q_i$, it is the standard Jensen method if $\bar{w}=o$ and it is the extended Jensen method if $\bar{w}=\bar{\pi}$. The summation in (3.5) is the power-series expansion of $e^{qt}\,[\pi(t)-\bar{w}]$. This is an entire function since $\pi(t)$ is an entire function. Thus, the summation converges.

A recursive algorithm to calculate the coefficients of expansion (3.5) will be derived by equating corresponding powers of time. This is similar to how Jensen's method was derived. Substituting the expansion in the forward differential system (3.1) renders

$$e^{-qt}\sum_{k=0}^{\infty}\frac{1}{k!}[kt^{k-1}-qt^k]w_k=\bar{w}Q+e^{-qt}\sum_{k=0}^{\infty}\frac{t^k}{k!}w_kQ,\quad \bar{w}+w_0=\phi,$$

or

$$\sum_{k=0}^{\infty} \frac{t^k}{k!} w_{k+1} = \sum_{k=0}^{\infty} \frac{(qt)^k}{k!} \bar{w} Q + \sum_{k=0}^{\infty} \frac{t^k}{k!} w_k [Q+q], \quad \bar{w} + w_0 = \phi.$$

Equating coefficients of corresponding powers of t leads to the following approximation:

$$\begin{aligned} \pi_K(t) &= \bar{w} + e^{-qt} \sum_{k=0}^{K} \frac{t^k}{k!} w_k, \\ w_0 &= \phi - \bar{w}, \\ w_k &= q^{k-1} \bar{w} Q + w_{k-1}[Q+q], \quad \text{for all } k \geq 1. \end{aligned} \tag{3.6}$$

In general, this recursion involves negative numbers. However, it has closed-form solution

$$w_k = \phi [Q+q]^k - \bar{w} q^k, \quad \text{for } k \geq 0. \tag{3.7}$$

From this, recursions can easily be obtained that avoid negative numbers, by splitting it into two positive parts.

An approximation $\pi_K(t)$ is a good approximation of $\pi(t)$ if K can be small for fixed t and if for fixed K it behaves well for large t. For small values of time, all approximations in the general framework are very similar. If the approximation must converge to the right limit $\bar{\pi}$ for $t \to \infty$ and each $K \geq 0$, then q should be positive and $\bar{w} = \bar{\pi}$. The relaxation time $\mathcal{R}$ is the smallest R such that $\|\pi(t) - \bar{\pi}\| \in \mathcal{O}\left(e^{-t/R}\right)$, for $t \to \infty$. If $\pi_K(t)$ must not only converge to $\bar{\pi}$ but must also have the right relaxation time $\mathcal{R}$ for each fixed K, then q should be chosen equal to $\mathcal{R}^{-1}$ and $\bar{w} = \bar{\pi}$. The matrix $[Q + \mathcal{R}^{-1}]$ will usually not be non-negative which would introduce numerical instability. Numerical experiments show that the choice $q = \sup_i q_i$ renders better results, especially for intermediate values of t. Actually, the only interesting special cases of the general framework are those discussed in the previous sections. The general framework discussed here will be helpful in section 3.2 for applying the PSA, since all three special cases can be considered simultaneously.

3.1.5 Comparison

Consider an $M/M/1$ queue with service rate 1.0 and arrival rate 0.5. The queue is initially empty. The steady-state probability of an empty queue is 0.5. Figure 3.1 shows the probability that the queue is empty for $t \leq 20$. The true curve is obtained by the extended Jensen method with $K = 100$ (the maximal error for $t \leq 20$ is close to machine precision and the maximal error for any $t \geq 0$ is at most $\|\bar{\pi} - v_{100}\| = 0.0000311$). The approximations are all obtained by truncating the expansions at $K = 10$ and with $q = 1.5$. The SS-detection curve is the curve obtained by the steady-state-detection method. The dotted lines are the error bounds for Taylor's method, and both the standard and the extended Jensen method.

As expected, Taylor's approximation diverges and the standard Jensen approximation converges to the wrong limit 0. The steady-state-detection method performs better,

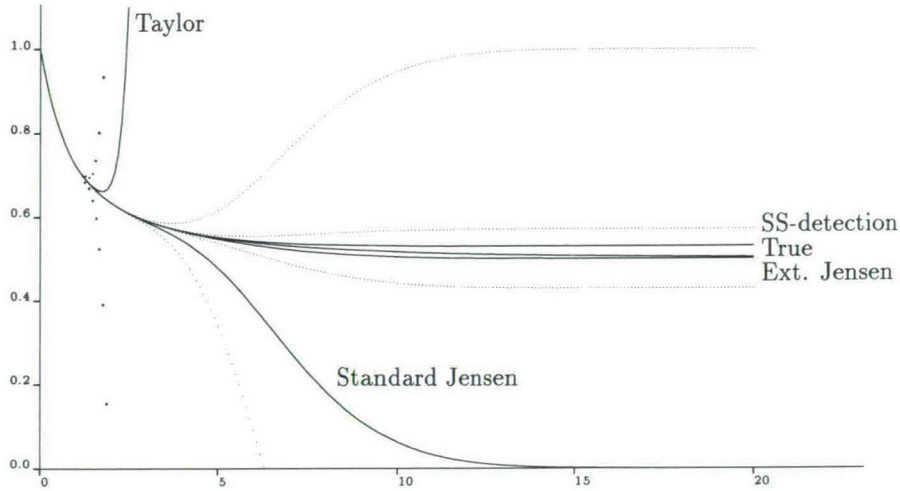

Figure 3.1: Probability of an empty queue for $t \leq 20$

but after 10 (dummy) transitions the probability of an empty queue is equal to 0.530, so it has not converged to 0.5 yet. Therefore, it is less accurate for large values of t. Nevertheless, even with this estimate of the steady-state value the results are considerably more accurate than the standard Jensen approximation. The extended Jensen method with the correct steady-state value is accurate for large values of time. More careful study shows that it converges to 0.5 too fast.

For the $M/M/1$ queue above, the obtained accuracy for a fixed value of K was considered. In [104], the value of K required to obtain a certain precision is considered. This is done for the $M/M/1/m$ model with arrival rate $\lambda = 9$, service rate $\mu = 10$, buffer size $m = 50$ and $q = 19$. The queue is initially empty and the distribution at time $t = 1000$ is calculated. For this model, the steady-state distribution can easily be calculated (see example 2.4).

Figure 3.2 shows the required truncation level K to obtain different levels of accuracy. The number of coefficients needed by the standard Jensen method is only slightly increasing in the required accuracy and roughly equal to 20000 ($> qt = 19000$). Due to the geometric convergence of v_k, the graphs of the extended Jensen method and the steady-state-detection method are nearly straight lines. The steady-state-detection method requires about 20% more coefficients for this model.

From these examples it can be concluded that the steady-state distribution can be used to extend Jensen's method and to obtain convergence uniform over time. If (a good

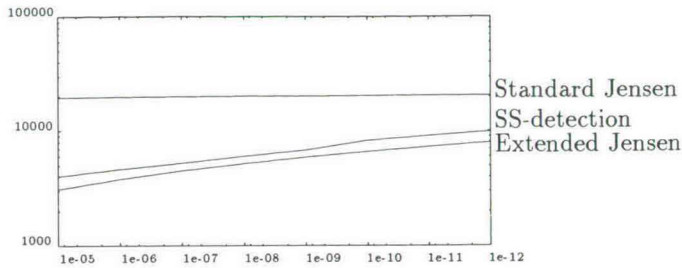

Figure 3.2: Required number of coefficients to obtain prescribed accuracy

estimate of) the steady-state distribution is available or can be calculated without too much effort, the extended Jensen method is preferable. It is similar to the steady-state-detection method, but converges faster and has reliable error bounds.

3.2 The Power-Series Algorithm

Encouraged by the success of the PSA for steady-state analysis, an attempt will be made to extend the PSA to transient analysis. The idea of the PSA for steady-state analysis is to transform the Markov process with a certain transformation parameter γ and analyze the process as a function of this parameter. A particular transformation was used (see section 2.2) and conditions were derived under which the steady-state probabilities are analytic functions of γ. For the transient analysis in the present chapter, not a particular transformation will be considered but the class of all analytic functions $Q(\gamma)$. Also, the initial distribution $\phi(\gamma)$ is allowed to depend on γ. The expansion of the transient distribution $\pi(t,\gamma)$ will be obtained, as a function of both time t and γ. The value $\gamma = 1$ will not be the only value of interest, because γ need not be an artificial parameter. It can be any model parameter, as long as $\phi(\gamma)$ and $Q(\gamma)$ are analytic in this parameter.

In this chapter, the expansion of the product of two functions is repeatedly used. Suppose that $h(x) = f(x)g(x)$ and that $f(x) = \sum_{k=0}^{\infty} x^k f_k$ and $g(x) = \sum_{\ell=0}^{\infty} x^\ell g_\ell$ are both analytic at $x = 0$. Then

$$h(x) = \sum_{k=0}^{\infty}\sum_{\ell=0}^{\infty} x^{k+\ell} f_k g_\ell = \sum_{k=0}^{\infty} x^k \sum_{\ell=0}^{k} f_{k-\ell} g_\ell. \tag{3.8}$$

The coefficients of the expansion of the product of two functions are the convolutions of the coefficients of the two functions. The radius of convergence of the product is equal to the minimum of the individual radii.

Suppose that the initial distribution and the transition rates are analytic functions of γ:

$$\phi(\gamma) = \sum_{r=0}^{\infty} \gamma^r \phi_r, \tag{3.9}$$

$$Q(\gamma) = \sum_{r=0}^{\infty} \gamma^r Q_r. \tag{3.10}$$

The vector function $\phi(\gamma)$ should be a distribution for small enough values of γ, so $\phi_0 e = 1$, $\phi_0 \geq 0$ and $\phi_r e = 0$ for all $r \geq 1$. The matrix function $Q(\gamma)$ should be a generator for small enough values of γ, so $Q_r e = o$ for all $r \geq 0$. For a definition of analyticity of matrix functions, see appendix B.4. If the initial distribution and the transition rates are functions of γ, then the transient distribution is clearly also a function of γ:

$$\pi(t, \gamma) = \phi(\gamma) \exp(tQ(\gamma)).$$

Similar to the expansion (3.5) for the direct methods, the expansion

$$\pi(t, \gamma) = \bar{w}(\gamma) + e^{-tq(\gamma)} \sum_{k=0}^{\infty} \frac{t^k}{k!} w_k(\gamma) \tag{3.11}$$

can be used here. Just as in formula (3.7), the vector functions $w_k(\gamma) = \sum_{r=0}^{\infty} \gamma^r w_{kr}$ are determined by $w_k(\gamma) = \tilde{w}_k(\gamma) - \hat{w}_k(\gamma)$, with

$$\begin{array}{lllllll} \tilde{w}_0(\gamma) & = & \phi(\gamma), & \tilde{w}_k(\gamma) & = & \phi(\gamma)[Q(\gamma) + q(\gamma)]^k & = & \tilde{w}_{k-1}(\gamma)[Q(\gamma) + q(\gamma)], \\ \hat{w}_0(\gamma) & = & \bar{w}(\gamma), & \hat{w}_k(\gamma) & = & \bar{w}(\gamma) q^k(\gamma) & = & \hat{w}_{k-1}(\gamma) q(\gamma), \end{array}$$

for all $k \geq 1$. From this and (3.8), the following approximation and recursion are immediately obtained:

$$\begin{array}{ll} \pi_{KR}(t, \gamma) = \bar{w}(\gamma) + e^{-tq(\gamma)} \sum_{k=0}^{K} \sum_{r=0}^{R} \frac{t^k}{k!} \gamma^r w_{kr}, & \\ w_{kr} = \tilde{w}_{kr} - \hat{w}_{kr}, & \text{for } k, r \geq 0, \\ \tilde{w}_{0r} = \phi_r, \quad \tilde{w}_{kr} = \sum_{s=0}^{r} \tilde{w}_{k-1,s}[Q_{r-s} + q_{r-s}], & \text{for } k \geq 1 \text{ and } r \geq 0, \\ \hat{w}_{0r} = \bar{w}_r, \quad \hat{w}_{kr} = \sum_{s=0}^{r} \hat{w}_{k-1,s} q_{r-s}, & \text{for } k \geq 1 \text{ and } r \geq 0. \end{array} \tag{3.12}$$

With this recursion, the calculation of $\tilde{w}_{kr}$ and $\hat{w}_{kr}$ does not involve negative numbers if $q_r \geq \sup_i |q_{rii}|$ for all $r \geq 0$. Possibly $\bar{w}(\gamma)$ and $exp(-tq(\gamma))$ will need to be approximated as well, but this will be ignored here. The closed-form solution to the recursion (3.12) is

$$\tilde{w}_{kr} = \sum_{s=0}^{r} \phi_{r-s} \sum_{\substack{n \in \mathbb{N}^k \\ ne=s}} \prod_{\ell=1}^{k} [Q_{n_\ell} + q_{n_\ell}], \tag{3.13}$$

$$\hat{w}_{kr} = \sum_{s=0}^{r} \bar{w}_{r-s} \sum_{\substack{n \in \mathbb{N}^k \\ ne=s}} \prod_{\ell=1}^{k} q_{n_\ell}, \tag{3.14}$$

for $k, r \geq 0$, with the definition $\mathbb{N}^0 \doteq \{0\}$. Only formula (3.13) will be proved by induction. The proof of (3.14) is very similar and will be omitted. For $k = 0$, the summation in (3.13) is empty unless $s = 0$. The expression correctly reduces to ϕ_r. Suppose that the closed-form solution is correct for $0 \leq k \leq K-1$. Then

$$\begin{aligned}
\tilde{w}_{Kr} &= \sum_{t=0}^{r} \tilde{w}_{K-1,t} \left[Q_{r-t} + q_{r-t}\right] \\
&= \sum_{t=0}^{r} \left\{ \sum_{s=0}^{t} \phi_{t-s} \sum_{\substack{n \in \mathbb{N}^{K-1} \\ ne=s}} \prod_{\ell=1}^{K-1} \left[Q_{n_\ell} + q_{n_\ell}\right] \right\} \left[Q_{r-t} + q_{r-t}\right] \\
&= \sum_{s=0}^{r} \phi_{r-s} \sum_{t=0}^{s} \sum_{\substack{n \in \mathbb{N}^{K-1} \\ ne=s-t}} \left\{ \prod_{\ell=1}^{K-1} \left[Q_{n_\ell} + q_{n_\ell}\right] \right\} \left[Q_t + q_t\right] \\
&= \sum_{s=0}^{r} \phi_{r-s} \sum_{\substack{n \in \mathbb{N}^{K} \\ ne=s}} \prod_{\ell=1}^{K} \left[Q_{n_\ell} + q_{n_\ell}\right].
\end{aligned}$$

For the third equality, first replace s by $t-s$, then reverse the order of summations, and finally replace s by $r-s$ and t by $r-t$. The fourth equality replaces t by n_K. This completes the proof of (3.13).

The approximation (3.12) can be calculated for any choice of $\bar{w}(\gamma)$ and $q(\gamma)$. The obvious choices are analogous to the direct methods in section 3.1. Then the scalar function $q(\gamma)$ is either identically zero or $q_r = \sup_i |q_{rii}|$ for all $r \geq 0$. The vector function $\bar{w}(\gamma)$ is either identically zero or equal to the steady-state distribution. For these choices the recursion can be simplified, because $\bar{w}(\gamma)Q(\gamma) \equiv o$. In that case

$$\begin{aligned}
&w_0(\gamma) = \phi(\gamma) - \bar{w}(\gamma), \\
&w_k(\gamma) = \left[\phi(\gamma) - \bar{w}(\gamma)\right]\left[Q(\gamma) + q(\gamma)\right]^k = w_{k-1}(\gamma)\left[Q(\gamma) + q(\gamma)\right], \quad \text{for } k \geq 1.
\end{aligned}$$

This renders the recursion

$$\begin{aligned}
w_{0r} &= \phi_r - \bar{w}_r, && r \geq 0, \\
w_{kr} &= \sum_{s=0}^{r} w_{k-1,s}\left[Q_{r-s} + q_{r-s}\right], && r \geq 0 \text{ and } k \geq 1.
\end{aligned} \tag{3.15}$$

Compared to recursion (3.12), this new recursion is simpler and requires slightly less computations, but the number of matrix multiplications remains the same. It is numerically less stable because it involves negative numbers. It has the closed-form solution

$$w_{kr} = \sum_{s=0}^{r} \left[\phi_{r-s} - \bar{w}_{r-s}\right] \sum_{\substack{n \in \mathbb{N}^{k} \\ ne=s}} \prod_{\ell=1}^{k} \left[Q_{n_\ell} + q_{n_\ell}\right],$$

for $k, r \geq 0$.

It may seem that the recursions (3.12) and (3.15) for the transient PSA are computationally much more expensive than the recursion for the direct methods (3.6). This need not be true. Multiplying by all the Q_r, $r \geq 0$, will not be much more work than

multiplying by just Q when the Q_r are more sparse. For example, with the transformation in the previous chapter, the number of non-zero off-diagonal elements of Q is equal to the total number of non-zero off-diagonal elements of the Q_r, $r \geq 0$. Also, it is not unusual that many of the Q_r and q_r are identical. Then the amount of work can be reduced by changing the order of summation and matrix multiplication. So, the number of coefficients that needs to be calculated is much larger than for the direct methods, but the amount of work per coefficient need not be much larger.

Like for the steady-state analysis, power-series expansions of performance measures can easily be obtained from those of the distribution. The transient expectation of the analytic row-vector function

$$f(t,\gamma) = \sum_{k=0}^{\infty}\sum_{r=0}^{\infty} t^k \gamma^r f_{kr}$$

can be obtained from the expansion (3.11) of the column vector $\pi(t,\gamma)$:

$$\pi(t,\gamma)f(t,\gamma) = \bar{w}(\gamma)f(t,\gamma) + e^{-tq(\gamma)} \sum_{k=0}^{\infty}\sum_{r=0}^{\infty} t^k \gamma^r v_{kr},$$

with

$$v_{kr} = \sum_{\ell=0}^{k}\sum_{s=0}^{r} \frac{1}{\ell!} w_{ls} f_{k-\ell,r-s}.$$

Of course, this expression greatly simplifies if $f(t,\gamma)$ is independent of t and/or γ.

Applying the ideas of the PSA to transient analysis has led to bivariate power-series expansions in time t and the model parameter γ. The coefficients of these expansions are tractable, but the amount of work to obtain the coefficients is considerably increased compared to the direct methods. The numerical results of the direct approximation (3.6) correspond to the PSA approximation (3.12) with truncation level $R = \infty$. Thus, the PSA introduces unnecessary truncation errors. For transient methods and bivariate power series, extrapolations methods turn out to be less effective in reducing these truncation errors. The only advantage of the PSA seems to be that the coefficients of the expansions of performance measures need only be calculated once. Subsequently, the obtained power series can be evaluated at many different values of the model parameter γ without much effort. With the direct methods, each value of γ needs to be dealt with separately.

3.3 Analyticity

It will be shown that the transient distribution $\pi(t,\gamma)$ is an analytic function of both time t and the model parameter γ, provided that the initial distribution $\phi(\gamma)$ and the transition rates $Q(\gamma)$ are analytic in γ. The way to prove this will be by proving convergence

of expansion (3.11) for all $t \geq 0$ and for

$$|\gamma| < R(\phi, Q, q, \bar{w}) \doteq \min\{R(\phi), R(Q), R(q), R(\bar{w})\}.$$

Here, $R(\phi), R(Q), R(q)$ and $R(\bar{w})$ are the respective radii of convergence of $\phi(\gamma)$, $Q(\gamma)$, $q(\gamma)$, and $\bar{w}(\gamma)$. For a definition of the radius of convergence of analytic matrix functions, see appendix B.4. Since $q(\gamma)$ and $\bar{w}(\gamma)$ can be chosen in such a way that they are entire functions, this implies the analyticity of $\pi(t, \gamma)$. The expansion (3.11) converges for all values of time and for all values of γ for which the power-series expansions of $\phi(\gamma)$ and $Q(\gamma)$ converge. More than this cannot be expected. If $\phi(\gamma)$ and $Q(\gamma)$ are entire functions of γ, then so is the transient distribution $\pi(t, \gamma)$.

The power series in t obtained by the direct methods in section 3.1 converge under the assumption that the generator Q is uniformizable. Implicitly this is also assumed here. Analyticity of $Q(\gamma)$ implies that for each R with $0 < R < R(Q)$, an α exists such that $\|Q_r\| \leq \alpha R^{-r}$ (see appendix B.4). Consequently,

$$\|Q(\gamma)\| \leq \sum_{r=0}^{\infty} |\gamma|^r \, \|Q_r\| \leq \sum_{r=0}^{\infty} |\gamma|^r \alpha R^{-r} = \frac{\alpha}{1 - |\gamma|/R}, \tag{3.16}$$

for all $|\gamma| < R$. The generator is bounded and uniformizable inside the region of convergence.

Theorem 3.1 *The summation* $\sum_{k=0}^{\infty} \sum_{r=0}^{\infty} \frac{t^k}{k!} \gamma^r w_{kr}$, *with*

$$\begin{aligned} w_{kr} = & \sum_{s=0}^{r} \phi_{r-s} \sum_{\substack{n \in \mathbb{N}^k \\ ne=s}} \prod_{\ell=1}^{k} [Q_{n_\ell} + q_{n_\ell}] \\ & - \sum_{s=0}^{r} \bar{w}_{r-s} \sum_{\substack{n \in \mathbb{N}^k \\ ne=s}} \prod_{\ell=1}^{k} q_{n_\ell}, \qquad \textit{for } k, r \geq 0, \end{aligned}$$

converges for all $t \geq 0$ *and* $|\gamma| < R(\phi, Q, q, \bar{w})$.

Proof: Choose a positive $R < R(\phi, Q, q, \bar{w})$. Then, according to appendix B.4, positive constants α_1, α_2, α_3 and α_4 exist, such that

$$\|\phi_r\| \leq \alpha_1 R^{-r}, \ \|Q_r\| \leq \alpha_2 R^{-r}, \ q_r \leq \alpha_3 R^{-r}, \ \|\bar{w}_r\| \leq \alpha_4 R^{-r},$$

for all $r \geq 0$. This provides a bound on the coefficients:

$$\begin{aligned}
\|w_{kr}\| \leq & \sum_{s=0}^{r} \|\phi_{r-s}\| \sum_{\substack{n \in \mathbb{N}^k \\ ne=s}} \prod_{\ell=1}^{k} \|Q_{n_\ell} + q_{n_\ell}\| & + \sum_{s=0}^{r} \|\bar{w}_{r-s}\| \sum_{\substack{n \in \mathbb{N}^k \\ ne=s}} \prod_{\ell=1}^{k} q_{n_\ell} \\
\leq & \sum_{s=0}^{r} \alpha_1 R^{s-r} \sum_{\substack{n \in \mathbb{N}^k \\ ne=s}} \prod_{\ell=1}^{k} (\alpha_2 + \alpha_3) R^{-n_\ell} & + \sum_{s=0}^{r} \alpha_4 R^{s-r} \sum_{\substack{n \in \mathbb{N}^k \\ ne=s}} \prod_{\ell=1}^{k} \alpha_3 R^{-n_\ell} \\
= & \sum_{s=0}^{r} \alpha_1 (\alpha_2 + \alpha_3)^k R^{s-r} \sum_{\substack{n \in \mathbb{N}^k \\ ne=s}} \prod_{\ell=1}^{k} R^{-n_\ell} & + \sum_{s=0}^{r} \alpha_4 \alpha_3^k R^{s-r} \sum_{\substack{n \in \mathbb{N}^k \\ ne=s}} \prod_{\ell=1}^{k} R^{-n_\ell} \\
\leq & \beta_1 \beta_2^k \sum_{s=0}^{r} R^{s-r} \sum_{\substack{n \in \mathbb{N}^k \\ ne=s}} R^{-ne} \\
= & \beta_1 \beta_2^k R^{-r} \# \left\{ n \in \mathbb{N}^k \;\middle|\; ne \leq r \right\} \\
= & \beta_1 \beta_2^k R^{-r} \binom{r+k}{k},
\end{aligned}$$

with $\beta_1 = \alpha_1 + \alpha_4$ and $\beta_2 = \alpha_2 + \alpha_3$. Consequently,

$$\begin{aligned}
\sum_{k=0}^{\infty} \sum_{r=0}^{\infty} \frac{t^k}{k!} \gamma^r \|w_{kr}\| \leq & \sum_{k=0}^{\infty} \sum_{r=0}^{\infty} \frac{t^k}{k!} \gamma^r \beta_1 \beta_2^k R^{-r} \binom{r+k}{k} \\
= & \beta_1 \sum_{k=0}^{\infty} \frac{(\beta_2 t)^k}{k!} \sum_{r=0}^{\infty} (\gamma/R)^r \binom{r+k}{k} \\
= & \beta_1 \sum_{k=0}^{\infty} \frac{(\beta_2 t)^k}{k!} (1 - \gamma/R)^{-k-1} \\
= & \frac{\beta_1 R}{R - \gamma} \exp\left(\frac{\beta_2 t R}{R - \gamma} \right),
\end{aligned}$$

for all $t \geq 0$ and $|\gamma| < R$. This is true for all positive $R < R(\phi, Q, q, \bar{w})$, which shows that the summation is absolutely convergent for all $t \geq 0$ and $|\gamma| < R(\phi, Q, q, \bar{w})$. □

3.4 Non-homogeneous Markov processes

In section 3.2, the transient distribution $\pi(t, \gamma)$ was obtained for the Markov process with generator $Q(\gamma)$ and initial distribution $\phi(\gamma)$. The parameter γ is allowed to have any physical interpretation, so it can even denote time. Then the generator $Q(t)$ in (3.10) varies over time and the Markov process is non-homogeneous. This observation leads to a promising new algorithm for the analysis of non-homogeneous Markov processes. The derivation in the previous two sections remains valid but the expansion (3.11) reduces to a univariate expansion:

$$\pi(t) = \pi(t, t) = \bar{w}(t) + e^{-tq(t)} \sum_{k=0}^{\infty} t^k u_k, \tag{3.17}$$

with

$$u_k = \sum_{\ell=0}^{k} \frac{1}{\ell!} w_{\ell, k-\ell}. \tag{3.18}$$

The double array of coefficients $w_{k,r}$ can be calculated by the same recursion as in section 3.2. It requires the specification of not only $Q(t)$, but also $\phi(t)$, $\bar{w}(t)$ and $q(t)$. It makes no sense to consider the initial distribution as a function of time, so $\phi(t) \equiv \phi$ (that is $\phi_0 = \phi$ and $\phi_r = o$ for all $r \geq 1$). If it exists, $\bar{w}(t)$ can be chosen equal to the steady-state distribution or the long-run average distribution for periodic models. If these are unknown, $\bar{w}(t)$ will usually be simply the zero vector. It will be difficult to find suitable time-dependent choices for $\bar{w}(t)$. As before, $q(t)$ can be chosen equal to zero or such that $q_r = \sup_i |q_{rii}|$, to make the matrix $[Q_r + q_r]$ non-negative.

The main advantages of the connection between the non-homogeneous case and the transformed homogeneous case is that the closed-form solution applies to both and that theorem 3.1 in the previous section is also valid for the non-homogeneous case. Therefore, it can be concluded that expansion (3.17) converges for small enough t:

Corollary 3.1 *The summation* $\sum_{\ell=0}^{\infty} t^\ell \sum_{k=0}^{\ell} \frac{1}{k!} w_{k,\ell-k}$ *converges for all* $0 \leq t < R(\phi, Q, q, \bar{w})$.

The functions $\phi(t)$, $q(t)$ and $\bar{w}(t)$ can easily be chosen in such a way that $R(\phi, Q, q, \bar{w}) = R(Q)$. However, recursion (3.18) is not a very efficient way to calculate the coefficients of the expansion. A more efficient recursion can be obtained directly from the differential system determining the transient distribution (see appendix A):

$$\pi'(t) = \pi(t)Q(t), \quad \pi(0) = \phi. \tag{3.19}$$

To derive the recursion, first the power-series expansion of $exp(-tq(t)) = \sum_{k=0}^{\infty} \frac{1}{k!} [-tq(t)]^k$ will be obtained. By induction, it can be shown that the expansion of $q^k(t)$ is equal to

$$q^k(t) = \sum_{\ell=0}^{\infty} t^\ell \sum_{\substack{n \in \mathbb{N}^k \\ ne=\ell}} \prod_{i=1}^{k} q_{n_i}.$$

For special choices of $q(t)$ this may simplify considerably. The expansion of $q^k(t)$ leads to the expansion of $exp(-tq(t))$:

$$\begin{aligned} exp(-tq(t)) &= \sum_{k=0}^{\infty} \frac{1}{k!} [-tq(t)]^k \\ &= \sum_{k=0}^{\infty} \frac{(-1)^k}{k!} \sum_{\ell=0}^{\infty} t^{k+\ell} \sum_{\substack{n \in \mathbb{N}^k \\ ne=\ell}} \prod_{i=1}^{k} q_{n_i} \\ &= \sum_{r=0}^{\infty} t^r S_r, \end{aligned}$$

with

$$S_r = \sum_{k=0}^{r} \frac{(-1)^k}{k!} \sum_{\substack{n \in \mathbb{N}^k \\ ne=r-k}} \prod_{i=1}^{k} q_{n_i}.$$

The scalar coefficients S_r can be calculated in a number of calculations that is quadratic in r. Since $\mathbb{N}^0 \doteq \{0\}$, the first constant S_0 is equal to 1. With this, the expansion (3.17) can be written as

$$\pi(t) = \sum_{k=0}^{\infty} t^k \left[\bar{w}_k + \sum_{i=0}^{k} S_i u_{k-i} \right],$$

and the differential system (3.19) as

$$\sum_{k\geq 1} k t^{k-1} \left[\bar{w}_k + \sum_{i=0}^{k} S_i u_{k-i} \right] = \sum_{k=0}^{\infty} t^k \sum_{h=0}^{k} \left[\bar{w}_h + \sum_{i=0}^{h} S_i u_{h-i} \right] Q_{k-h}.$$

Equating the coefficients of t^{k-1} renders

$$u_k = -\left[\bar{w}_k + \sum_{i=1}^{k} S_i u_{k-i} \right] + \frac{1}{k} \sum_{h=0}^{k-1} \left[\bar{w}_h + \sum_{i=0}^{h} S_i u_{h-i} \right] Q_{k-1-h}, \tag{3.20}$$

for all $k \geq 1$ and with $u_0 = \phi$. Since the diagonal of each Q_r is non-positive, this recursion involves negative numbers. On the other hand, it requires less computations than computation by (3.18).

Above, the power-series expansion around $t = 0$ was obtained. The power-series expansion around any $T \geq 0$ can be obtained, starting from the differential system

$$\pi'(T+t) = \pi(T+t)Q(T+t), \quad \pi(T) = \phi. \tag{3.21}$$

The power-series expansion of $Q(T+t)$ around T and the distribution ϕ at time T should be known. The corollary states that the transient distribution is analytic in t, but not necessarily entire: convergence is shown only for $t < R(\phi, Q, q, \bar{w})$. To calculate the distribution at time $T \geq R(\phi, Q, q, \bar{w})$, the procedure can be applied several times if the generator $Q(t)$ is analytic in t for all $t \in [0, T]$. In that case, (3.16) shows that the generator is bounded for all $t \in [0, T]$. Therefore, explosion can not occur before time T and the distribution at time T is well-defined. Also, the generator has a power-series expansion around each $t \in [0, T]$ with positive radius of convergence. The functions $q(t)$ and $w(t)$ can be chosen such that they also have a convergent power-series expansion around each $t \in [0, T]$. This shows that the interval $[0, T]$ can be divided into a finite number of intervals $[0, T_1]$, $[T_1, T_2], \ldots [T_N, T]$ in such a way that the distribution at the end of each interval can be obtained from the distribution at the beginning of the interval. The size of each interval should be smaller than the radius of convergence of the expansions at the beginning of the interval. Choosing the intervals large will decrease the number of intervals, but increase the number of coefficients that need to be calculated for each interval to obtain a certain accuracy.

The method described here is new and seems promising, especially in cases where the computational complexity can be reduced. It is not necessary to consider ever smaller

time intervals to increase the accuracy, like in [120,45]. For a given interval, smaller than the radius of convergence, accuracy can be increased by increasing the number of calculated coefficients. Possibly, both methods could be combined. If the problem is ill-conditioned, the stable formula (3.18) can be used to avoid negative numbers. Otherwise, the more efficient (3.20) can be used.

Appendix A

Markov processes

In this appendix, some concepts from the theory of Markov processes will be reviewed and sufficient conditions for ergodicity of Markov processes will be considered. For irreducible and aperiodic discrete-time Markov chains, the first Foster criterion [55] shows that a sufficient condition for ergodicity is that the balance equations have an absolutely convergent non-null solution. It is not required beforehand that this solution is positive. For continuous-time Markov processes, similar conditions for ergodicity available in the literature do require that the solution is positive or that the process is uniformizable. The PSA provides a way to calculate a solution to the balance equations. Since in general the coefficients of the obtained power series will not be non-negative, there is no guarantee that the calculated solution to the balance equations is non-negative. Therefore, ergodicity conditions that assume that the solution is positive can not be used in section 2.5. Also, the condition that the process is uniformizable is stronger than necessary. The theorems in this appendix do not make these assumptions.

Consider a continuous-time Markov process (CTMP) $\{\mathcal{X}_t;\ t \geq 0\}$ on a countable state space Ω and *transition probability matrix* $P(s,t) = [p_{ij}(s,t)]$ defined by

$$p_{ij}(s,t) \doteq \begin{cases} \mathbf{P}\{\ \mathcal{X}_t = j \mid \mathcal{X}_s = i\ \}, & \text{for all} \quad t > s \geq 0 \quad \text{and} \quad i,j \in \Omega, \\ 1(i=j), & \text{for all} \quad t = s \geq 0 \quad \text{and} \quad i,j \in \Omega. \end{cases}$$

The transition probability matrix is called *standard* if it is right continuous in t at $t = s$ for all $s \geq 0$. Then, the *(infinitesimal) generator* $Q(s) = [q_{ij}(s)]$ is the right-hand

derivative of $P(s,t)$ with respect to t at $t = s$:

$$q_{ij}(s) \doteq \lim_{t \downarrow s} \frac{p_{ij}(s,t) - 1(i=j)}{t-s}, \quad \text{for all } s \geq 0 \text{ and } i,j \in \Omega.$$

The *Chapman-Kolmogorov equation* states that the process after time t is completely determined by the state $\mathcal{X}_t$ at time t: $P(s,r) = P(s,t)P(t,r)$, for all $r > t > s \geq 0$. This can be used to obtain the transition probability matrix from the generator. Subtract $P(s,t)$, divide by $r-t$ and let r approach t. This renders the differential system

$$P'(s,t) = P(s,t)Q(t), \quad P(s,s) = I, \quad \text{for all } t > s \geq 0,$$

where differentiation is with respect to t. If at time $s = 0$ the distribution on Ω is according to ϕ, then the distribution at time t satisfies

$$\pi'(t) = \pi(t)Q(t), \quad \pi(0) = \phi, \quad \text{for all } t \geq 0.$$

This is easily deduced from $\pi(t) = \phi P(0,t)$.

The process is *homogeneous* if $P(s, s+t)$ is independent of s. In this case, define $P(t) \doteq P(0,t)$ and $Q \doteq Q(0)$. The differential system reduces to $\pi'(t) = \pi(t)Q$, with formal solution $\pi(t) = \phi \exp(Qt) = \phi \sum_{k \geq 0} \frac{t^k}{k!} Q^k$. A homogeneous process is called *honest* if the departure rate from each state is equal to the sum of the transition rates to the other states:

$$q_i \doteq -q_{ii} = \sum_{j \in \Omega \setminus \{i\}} q_{ij}, \quad \text{for all } i \in \Omega.$$

The process is *uniformizable* if the departure rates are bounded: $\sup_i q_i < \infty$. A state $i \in \Omega$ is *absorbing* if $q_i = 0$ and *instantaneous* if $q_i = \infty$. The *jump matrix* $R = [r_{ij}]$ is the transition probability matrix of the embedded discrete-time Markov chain (DTMC) at jump moments:

$$r_{ij} \doteq \begin{cases} \dfrac{q_{ij}}{q_i} \; 1(i \neq j), & \text{if } q_i > 0, \text{ for all } i,j \in \Omega, \\ 1(i=j), & \text{if } q_i = 0, \text{ for all } i,j \in \Omega, \end{cases}$$

provided all states are non-instantaneous. Let $R^k = [r_{ij}^k]$ denote the k-th power of the jump matrix, so r_{ij}^k is the probability that the jump chain is in state j after k jumps, starting from state i. The process is *irreducible* if a finite k exists such that $r_{ij}^k > 0$ for all $i,j \in \Omega$, that is if starting from any state any other state can be reached. It is *recurrent* if $\sum_{k \geq 1} r_{ii}^k = \infty$ for any $i \in \Omega$, that is if the expected number of returns to any state is infinite.

A homogeneous process is *ergodic* if it is both irreducible and recurrent and has a *stationary distribution*, that is a distribution π on Ω such that $\pi P(t) = \pi$, for all $t \geq 0$. If a process is ergodic, then the *steady-state distribution*

$$\bar{\pi} \doteq \lim_{t \to \infty} \pi(t) = \lim_{t \to \infty} \phi P(t)$$

exists, independent of the initial distribution ϕ, and

$$\bar{\pi}Q = \lim_{t\to\infty} \pi(t)Q = \lim_{t\to\infty} \pi'(t) = o.$$

If a process is ergodic, the steady-state distribution can be found by solving the *balance equations* $\bar{\pi}Q = o$ and the *normalization equation* $\bar{\pi}e = 1$. The reverse is not true: the existence of a solution π to the balance and normalization equations does not imply that the process is ergodic. Besides the conditions that the process is irreducible, honest and non-instantaneous, additional conditions are required. According to Asmussen it is sufficient that $\pi \geq o$ and $\sum_i \pi_i q_i < \infty$ (section II.4 of [7]). According to Cohen it is sufficient that $\sup_i q_i < \infty$ and $\sum_i |\pi_i| < \infty$ (sections I.3.2 and I.3.4 of [41]). The theorem below shows that the weaker condition $\sum_i |\pi_i| q_i < \infty$ is also sufficient. That the solution is positive is not assumed beforehand but is deduced from the fact that the solution for the embedded jump chain is positive. All these conditions ensure that the process is *non-explosive*, so almost surely only a finite number of transitions occur in each finite time interval. For other recent approaches to ergodicity of homogeneous processes, not based on a solution to the balance equations, see Lindvall [101] and Meyn and Tweedie [109].

Theorem A.1 *An irreducible, honest, non-instantaneous continuous-time Markov process on state space Ω is ergodic if a solution $\pi = [\pi_i]$ exists such that*

$$\pi_i q_i = \sum_{j\in\Omega\setminus\{i\}} \pi_j q_{ji}, \quad \textit{for all } i \in \Omega, \tag{A.1}$$

$$\sum_{i\in\Omega} \pi_i = 1, \tag{A.2}$$

$$\sum_{i\in\Omega} |\pi_i| q_i < \infty. \tag{A.3}$$

In that case π is the steady-state distribution.

Proof: First it is shown that the assumptions imply that the jump chain is ergodic, using theorem I.7.1 in Chung [38]. Since the continuous-time Markov process is irreducible, the jump chain is also irreducible, with period $d \geq 1$ and long-run distribution $\sigma = [\sigma_i]$ with $\sigma_i \doteq \frac{1}{d} \lim_{k\to\infty} r_{ii}^{dk} \geq 0$. Define $\phi = [\phi_i]$ by $\phi_i \doteq \pi_i q_i$. Then the balance equations of the continuous-time Markov process (A.1) and equation (A.3) can be written as

$$\phi_i = \sum_{j\in\Omega} \phi_j r_{ji}, \quad \text{for all } i \in \Omega,$$
$$\sum_{i\in\Omega} |\phi_i| < \infty.$$

This shows that ϕ is a solution to the balance equations of the jump chain and is absolutely convergent. Hence, ϕ is a multiple of the long-run distribution of the jumpchain: $\phi = c\sigma$, with $c \in \mathbb{R}$. Consequently, equation (A.2) shows that $\sum_{i\in\Omega} \pi_i = \sum_{i\in\Omega} \frac{\phi_i}{q_i} =$

$c\sum_{i\in\Omega}\frac{\sigma_i}{q_i}=1$. Since all σ_i are non-negative, c and at least one of the σ_i must be positive. This finally implies that the jump chain is ergodic and that c, σ and ϕ are all strictly positive.

By proposition II.2.4 in Asmussen[7], the ergodicity of the jump chain implies that the continuous-time Markov process is non-explosive. That ϕ is strictly positive implies that also π is strictly positive. From the normalization (A.2) it then follows that π is a probability distribution satisfying the balance equations. Together with the irreducibility and the non-explosiveness, theorem II.4.3 in Asmussen [7] asserts that then the Markov process is ergodic with steady-state distribution π. □

The second theorem is very similar but has slightly weaker conditions. The assumption that the balance equations are satisfied for all states in the state space is relaxed. The balance equations need only be satisfied for all but one state. It is concluded that the balance equations are then satisfied for all states. This requires a reversal of the order of summation, which is justified without making any additional conditions.

Theorem A.2 *An irreducible, honest, non-instantaneous continuous-time Markov process on state space Ω is ergodic if a solution $\pi=[\pi_i]$ exists such that for some $k\in\Omega$:*

$$\pi_i q_i = \sum_{j\in\Omega\backslash\{i\}} \pi_j q_{ji}, \quad \textit{for all } i\in\Omega\backslash\{k\}, \tag{A.4}$$

$$\sum_{i\in\Omega}\pi_i = 1, \tag{A.5}$$

$$\sum_{i\in\Omega}|\pi_i|q_i < \infty. \tag{A.6}$$

In that case π is the steady-state distribution.

Proof: Summing the balance equations (A.4) over all $i\in\Omega\backslash\{k\}$ and reversing the order of summation renders

$$\begin{aligned}\sum_{i\in\Omega\backslash\{k\}}\pi_i q_i &= \sum_{i\in\Omega\backslash\{k\}}\sum_{j\in\Omega\backslash\{i\}}\pi_j q_{ji}\\ &= \sum_{j\in\Omega}\sum_{i\in\Omega\backslash\{j,k\}}\pi_j q_{ji} \quad = \sum_{j\in\Omega}\pi_j q_j - \sum_{j\in\Omega\backslash\{k\}}\pi_j q_{jk}.\end{aligned} \tag{A.7}$$

The reversal is justified because the final summation is absolutely convergent:

$$\sum_{j\in\Omega}|\pi_j|q_j + \sum_{j\in\Omega\backslash\{k\}}|\pi_j|q_{jk} \le 2\sum_{j\in\Omega}|\pi_j|q_j < \infty.$$

Subtracting $\sum_{i\in\Omega}\pi_i q_i$ (which is finite) from (A.7) and reversing the sign, renders the balance equation of state k:

$$\pi_k q_k = \sum_{j\in\Omega\backslash\{k\}}\pi_j q_{jk}.$$

Therefore, the balance equations are satisfied for all states, the assumptions of this theorem imply the assumptions of the previous theorem and the validity of this theorem follows from the previous theorem. □

Appendix B

Analytic functions

In this appendix, results from the theory of analytic functions are reviewed. Without thorough knowledge of these results, a good understanding of the PSA is impossible. None of the results is new. They can be found in good textbooks on complex analysis, like [71,115,129]. The approach followed here will be the approach of Weierstrass, based on power-series. For computational purposes, this is more appropriate than the approach of Riemann, based on complex differentiability.

The primary reasons to include this appendix are to make the thesis more self-contained and to give the reader a concise overview to refresh his or her memory. A secondary reason is that, in many text-books, integration along paths is employed in an early stage. This is avoided here, because it is not useful for the purpose of this thesis. Proofs are included because they are mostly short and instructive.

B.1 Series

Let the series $\sum_{k=0}^{\infty} u_k$ be an infinite sum of complex numbers and let $S_K = \sum_{k=0}^{K} u_k$ be the K-th partial sum of the series. The series is said to *converge* if $S = \lim_{K\to\infty} S_K$ exists and is finite. It converges *absolutely* if $\sum_{k=0}^{\infty} |u_k|$ converges. A *majorant* of the series $\sum_{k=0}^{\infty} u_k$ is a series $\sum_{k=0}^{\infty} v_k$ such that $|u_k| \leq v_k$ for all $k \geq 0$.

Theorem B.1 proves that if a series has a convergent majorant, then it is (absolutely) convergent. It immediately implies that an absolutely convergent series is convergent.

Theorem B.2 provides sufficient conditions for the reversal of summations.

Theorem B.1 (convergent majorant) *If $|u_k| \leq v_k$ for all $k \geq 0$ and the series $\sum_{k=0}^{\infty} v_k$ converges, then the series $\sum_{k=0}^{\infty} u_k$ converges (absolutely).*

Proof: For all $K \geq 0$,

$$0 \leq \left| \sum_{k=K+1}^{\infty} u_k \right| \leq \sum_{k=K+1}^{\infty} |u_k| \leq \sum_{k=K+1}^{\infty} v_k.$$

Because $\sum_{k=0}^{\infty} v_k$ converges, the right-hand side converges to 0 for $K \to \infty$. Therefore, all terms converge to 0 and the series $\sum_{k=0}^{\infty} u_k$ converges (absolutely). □

Theorem B.2 (reversal of summations) *If either of*

$$\sum_{i=0}^{\infty} \left(\sum_{j=0}^{\infty} |u_{ij}| \right) \quad \textit{and} \quad \sum_{j=0}^{\infty} \left(\sum_{i=0}^{\infty} |u_{ij}| \right)$$

converges, then both

$$\sum_{i=0}^{\infty} \left(\sum_{j=0}^{\infty} u_{ij} \right) \quad \textit{and} \quad \sum_{j=0}^{\infty} \left(\sum_{i=0}^{\infty} u_{ij} \right)$$

converge and are equal.

Proof: Let σ be the minimum of the convergent limits in the assumption (in fact, it immediately follows from this theorem B.2 that if either of them converges then both converge and are equal). Then $\sum_{j=0}^{J} |u_{ij}| \leq \sigma$ for all $i, J \geq 0$ and this implies that $\sum_{j=0}^{\infty} |u_{ij}| \leq \sigma$ for all $i \geq 0$. Thus $\sum_{j=0}^{\infty} u_{ij}$ is (absolutely) convergent for all $i \geq 0$. In a similar way, $\sum_{i=0}^{I} |\sum_{j=0}^{\infty} u_{ij}| \leq \sigma$ for all $I \geq 0$ implies that $\sum_{i=0}^{\infty} |\sum_{j=0}^{\infty} u_{ij}| \leq \sigma$, and thus $\sum_{i=0}^{\infty} \left(\sum_{j=0}^{\infty} u_{ij} \right)$ is (absolutely) convergent. Equivalently, it can be shown that $\sum_{j=0}^{\infty} (\sum_{i=0}^{\infty} u_{ij})$ is (absolutely) convergent.

What remains to be shown is that both double series converge to the same limit. For all $K \geq 0$, $\sigma_K = \sum_{i=0}^{K} \sum_{j=0}^{K} |u_{ij}| \leq \sigma$, so σ_K is bounded and increasing in K. Therefore, it converges and for any $\epsilon > 0$ a $K_\epsilon \geq 0$ exists such that

$$\sum_{\substack{(i,j) \in \mathbb{N}^2 \\ \max\{i,j\} > K_\epsilon}} |u_{ij}| < \epsilon.$$

Then $\sum_{i=0}^{K_\epsilon} \sum_{j=0}^{K_\epsilon} u_{ij} = \sum_{j=0}^{K_\epsilon} \sum_{i=0}^{K_\epsilon} u_{ij}$ differs by at most ϵ from both $\sum_{i=0}^{\infty} \sum_{j=0}^{\infty} u_{ij}$ and $\sum_{j=0}^{\infty} \sum_{i=0}^{\infty} u_{ij}$. Therefore, these double series differ at most 2ϵ. □

B.2 Power series

The power series $\sum_{k=0}^{\infty} u_k(z-a)^k$ around centre a has K-th partial sum or truncated power series $S_K(z) = \sum_{k=0}^{K} u_k(z-a)^k$. The power series is said to *converge in* $z \in \mathbb{C}$ if $S(z) = \lim_{K\to\infty} S_K(z)$ exists and is finite. The *radius of convergence* of the power series is the largest r such that the power series converges on the open disk $|z-a| < r$. A power series converges *absolutely in* $z \in \mathbb{C}$ if $\sum_{k=0}^{\infty} |u_k(z-a)^k|$ converges.

The power series converges *uniformly* on $\mathcal{A} \subset \mathbb{C}$, if for any $\epsilon > 0$ an integer K_ϵ exists such that $|S(z) - S_K(z)| \leq \epsilon$ for all $K \geq K_\epsilon$ and all $z \in \mathcal{A}$. Loosely speaking, this implies that the function $S(z)$ can be approximated on $\mathcal{A}$ to any degree of accuracy by the truncated power series $S_K(z)$, provided K is large enough. Notice that K_ϵ is not allowed to depend on the value of z, which is allowed for pointwise convergence.

The Cauchy-Hadamard theorem B.3 proves that the radius of convergence of a power series is equal to

$$R(u) = \left[\limsup_{k\to\infty} |u_k|^{1/k} \right]^{-1}, \tag{B.1}$$

with the convention that $0^{-1} = \infty$ and $\infty^{-1} = 0$. The Weierstrass theorem B.4 proves that a power series is uniformly convergent and continuous on a closed disk with radius slightly smaller than $R(u)$. Cauchy's estimate in theorem B.5 provides a geometric upper bound on the coefficients of a convergent power series. Together with theorem B.1 on absolute convergence this implies that a power series is convergent if and only if it has a geometric majorant. The conversion theorem B.6 shows how a power series around a can be converted to a power series around b.

Theorem B.3 (Cauchy-Hadamard) *The power series $\sum_{k=0}^{\infty} u_k(z-a)^k$ is (absolutely) convergent if $|z-a| < R(u)$ and divergent if $|z-a| > R(u)$.*

Proof: If $|z-a| < R(u)$, the power series is (absolutely) convergent because it has a convergent majorant. If $R(u) = 0$, the set of $z \in \mathbb{C}$ for which convergence needs to be proved is empty. So suppose that $R(u) > 0$ and choose r such that $|z-a| < r < R(u)$. Since $r^{-1} > R(u)^{-1}$, $|u_k|^{1/k}$ will eventually be smaller than r^{-1}. Thus $|u_k| \leq r^{-k}$, say for all $k > K$. Choose $\alpha = \max\left\{1,\ |u_0|,\ |u_1|r, \ldots,\ |u_K|r^K\right\}$. Then $|u_k| \leq \alpha r^{-k}$, for all $k \geq 0$, so

$$\sum_{k=0}^{\infty} \left|u_k(z-a)^k\right| \leq \sum_{k=0}^{\infty} \alpha r^{-k}|z-a|^k = \frac{\alpha}{1-|z-a|/r}.$$

If $|z-a| > R(u)$, then the power series is divergent because the individual terms do not converge to 0. If $R(u) = \infty$, the set of $z \in \mathbb{C}$ for which divergence needs to be proved is empty. So suppose that $R(u) < \infty$ and choose r such that $|z-a| > r > R(u)$. Then $|u_k| > r^{-k}$ infinitely often and therefore

$$\left|u_k(z-a)^k\right| = |u_k||z-a|^k > r^{-k}r^k = 1,$$

infinitely often. □

Theorem B.4 (Weierstrass) *The power series $\sum_{k=0}^{\infty} u_k(z-a)^k$ is uniformly convergent and continuous on $|z-a| \leq r$ for any r such that $0 < r < R(u)$.*

Proof: For any s such that $r < s < R(u)$, an α exists such that $|u_k| \leq \alpha s^{-k}$ for all $k \geq 0$ (see the proof of the previous theorem B.3). Choose $K_\epsilon \geq [\ln(\epsilon/\alpha)+\ln(s/r-1)]/\ln(r/s)$. Then

$$|S(z) - S_K(z)| \leq \sum_{k=K+1}^{\infty} |u_k||z-a|^k \leq \sum_{k=K+1}^{\infty} \alpha s^{-k} r^k = \left(\frac{r}{s}\right)^K \frac{\alpha r}{s-r} \leq \epsilon,$$

for all $K \geq K_\epsilon$ and all $|z-a| \leq r$.

Choose an $\epsilon > 0$, z such that $|z-a| \leq r$ and d such that $|z+d-a| \leq r$. Because of the uniform convergence, a large enough K exists such that

$$|S(z+d) - S_K(z+d)| < \epsilon/3 \quad \text{and} \quad |S_K(z) - S(z)| < \epsilon/3.$$

Since $S_K(z)$ is the sum of a finite number of continuous functions, it is continuous. Therefore, a small enough $\delta > 0$ exists such that

$$|S_K(z+d) - S_K(z)| < \epsilon/3,$$

for all $|d| < \delta$. Combining all three, shows that

$$|S(z+d) - S(z)| \leq \epsilon/3 + \epsilon/3 + \epsilon/3 = \epsilon,$$

for all $|d| < \delta$, so $S(z)$ is continuous. □

Theorem B.5 (Cauchy's estimate) *If $0 < r < R(u)$ and*

$$\alpha = \sup_{|z-a|=r} |S(z)| < \infty,$$

then $|u_k| \leq \alpha r^{-k}$ for all $k \geq 0$.

Proof: Consider first the truncated power series $S_K(z) = \sum_{k=0}^{K} u_k(z-a)^k$, and define

$$\alpha_K = \sup_{|z-a|=r} |S_K(z)|.$$

Also, for a given M with $0 \leq M \leq K$, define

$$T_K(z) = (z-a)^{-M} S_K(z) = u_M + \sum_{\substack{k=0 \\ k \neq M}}^{K} u_k (z-a)^{k-M}.$$

Then

$$\left|T_K\left(a + re^{ni\phi}\right)\right| = r^{-M}\left|S_K\left(a + re^{ni\phi}\right)\right| \leq \alpha_K r^{-M}, \quad \text{for all } n\phi \in \mathbb{R}.$$

If $\phi = 2\pi(K+1)^{-1}$, then

$$\sum_{n=0}^{K} e^{n(k-M)i\phi} = \frac{1 - e^{(k-M)i\phi(K+1)}}{1 - e^{(k-M)i\phi}} = \frac{0}{1 - e^{(k-M)i\phi}} = 0,$$

provided $k - M$ is not a multiple of $K+1$. Thus,

$$\frac{1}{K+1}\sum_{n=0}^{K} T_K\left(a + re^{ni\phi}\right) = u_M + \frac{1}{K+1}\sum_{\substack{k=0\\k\neq M}}^{K} u_k r^{k-M} \sum_{n=0}^{K} e^{n(k-M)i\phi} = u_M.$$

This provides the bound

$$|u_M| \leq \frac{1}{K+1}\sum_{n=0}^{K}\left|T_K\left(a + re^{ni\phi}\right)\right| \leq \frac{1}{K+1}\sum_{n=0}^{K}\alpha_K r^{-M} = \alpha_K r^{-M},$$

for $0 \leq M \leq K$, which proves the theorem for finite power series.

To prove the theorem for the infinite power series $S(z) = \sum_{k=0}^{\infty} u_k(z-a)^k$, choose any $\epsilon > 0$ and $M \geq 0$. Since $r < R(u)$, $S(z)$ is uniformly convergent on $|z-a| \leq r$. Therefore, a $K \geq M$ exists such that $|S(z) - S_K(z)| \leq \epsilon$ for all $|z-a| \leq r$, so $\alpha_K \leq \alpha + \epsilon$. Because of the first part of the proof,

$$|u_M| \leq \alpha_K r^{-M} \leq (\alpha + \epsilon)\, r^{-M}.$$

This is true for any $\epsilon > 0$ and $M \geq 0$, which proves the theorem. □

Theorem B.6 (conversion) *If $|z-b| + |b-a| < R(u)$, then*

$$\sum_{k=0}^{\infty} u_k(z-a)^k = \sum_{k=0}^{\infty} v_k(z-b)^k,$$

with

$$v_k = \sum_{i=k}^{\infty}\binom{i}{k} u_i(b-a)^{i-k}.$$

Proof: That the two power series around a and b are identical, can be shown by reversing the order of summation:

$$\begin{aligned}
\sum_{k=0}^{\infty} v_k(z-b)^k &= \sum_{k=0}^{\infty}\sum_{i=k}^{\infty}\tbinom{i}{k} u_i(b-a)^{i-k}(z-b)^k \\
&= \sum_{i=0}^{\infty} u_i \sum_{k=0}^{i}\tbinom{i}{k}(b-a)^{i-k}(z-b)^k \\
&= \sum_{i=0}^{\infty} u_i\left[(b-a) + (z-b)\right]^i = \sum_{i=0}^{\infty} u_i(z-a)^i.
\end{aligned}$$

This reversal is justified because

$$\begin{aligned}
\sum_{i=0}^{\infty}\sum_{k=0}^{i}\left|u_i\tbinom{i}{k}(b-a)^{i-k}(z-b)^k\right| &= \sum_{i=0}^{\infty}|u_i|\sum_{k=0}^{i}\tbinom{i}{k}|b-a|^{i-k}|z-b|^k \\
&= \sum_{i=0}^{\infty}|u_i|\left(|b-a| + |z-b|\right)^i,
\end{aligned}$$

which converges if $|b-a| + |z-b| < R(u)$. □

Example B.1 Consider the power series

$$\sum_{k=0}^{\infty} kz^k.$$

The power series is (absolutely) convergent on $|z| < 1 = R(u)$, with limit $z(1-z)^{-2}$. The power series is not uniformly convergent on $|z| < 1$, since for any $\epsilon > 0$ and $K \geq 0$, a large enough $x < 1$ exists such that $\sum_{k=K+1}^{\infty} kx^k > \epsilon$. However, the power series is uniformly convergent and continuous on $|z| \leq r$ for all $0 < r < 1$. The coefficients of the power series do not have a constant majorant, but for any $r < 1 = R(u)$ the coefficients are bounded by αr^{-k}, with $\alpha = r(1-r)^{-2}$. The power series around 0 can be converted to a power series around b:

$$\sum_{k=0}^{\infty} kz^k = \sum_{k=0}^{\infty} v_k (z-b)^k$$

with

$$v_k = \sum_{i=k}^{\infty} \binom{i}{k} i\, b^{i-k} = \frac{k+b}{(1-b)^{k+2}}.$$

According to the conversion theorem B.6 this is valid for $|b| + |x-b| < 1$. In fact, the power series around b converges to the same limit function $z(1-z)^{-2}$ for all $|x-b| < |1-b| = R(v)$. So for any $b \neq 1$ it has radius of convergence equal to the distance between b and 1. □

B.3 Analytic functions

Analyticity is an important characteristic of a function, because analytic functions can be approximated arbitrarily close by the truncated power series. There are two equivalent definitions of analyticity, one based on power series and the other on complex differentiability. In the context of this thesis, the first is more useful:

Definition (Analyticity) *The function $f : \mathbb{C} \to \mathbb{C}$ is analytic in $a \in \mathbb{C}$, if a power series $\sum_{k=0}^{\infty} u_k (z-a)^k$ exists, convergent and equal to $f(z)$ for all z in a neighbourhood of a.*

A function is *analytic on* $\mathcal{A} \subset \mathbb{C}$ if it is analytic in all $a \in \mathcal{A}$. An *entire* function is analytic on $\mathbb{C}$. A *regular* point of a function is a point in which the function is analytic. The Weierstrass theorem B.4 implies that a function is continuous in all regular points. The conversion theorem B.6 implies that the set of regular points is an open set.

All points of a function that are not regular are called *singularities*. There are different types of singularities. A singularity a is *isolated* if a has a neighbourhood $\mathcal{A}$, such that the function is analytic on $\mathcal{A}/\{a\}$. An isolated singularity a is *removable* if f can be made analytic on $\mathcal{A}$ by a suitable redefinition of $f(a)$. A non-removable singularity a is an *r-th order pole* if it is a removable singularity of $(x-a)^r f(x)$ and

$\lim_{x\to a}(x-a)^r f(x) = R \neq 0$. The limit R is called the *residue*. An isolated singularity is called *essential* if it is neither removable nor a pole. *Branch points* are examples of non-isolated singularities (see example B.3).

Let $\mathcal{A}$ be a neighbourhood of $a \in \mathbb{C}$, and $\mathcal{B}$ a subset of $\mathcal{A}$ with accumulation point a. If a function $f : \mathcal{A} \to \mathbb{C}$ is analytic on $\mathcal{A}$ and equal to the function $g : \mathcal{B} \to \mathbb{C}$ on $\mathcal{B}$, then f is called an *analytic continuation* of g.

The function f is *complex differentiable* in z if the limit

$$f'(z) = \lim_{d\in\mathbb{C},d\to 0} \frac{f(z+d)-f(z)}{d}$$

exists, independent of the direction from which d approaches 0. According to the definition of analyticity based on complex differentiability, a function f is analytic in a if it is complex differentiable on a neighbourhood of a.

The Taylor theorem B.7 shows that the complex derivative of an analytic function exists and is again an analytic function. The coefficients of the power series in the definition of analyticity are uniquely determined by the complex derivatives of f in a. This unique power series will be called the power-series expansion or Taylor expansion around a of the function f. Theorem B.8 says that the distance between the centre of a power series expansion and the nearest singular point is equal to the radius of convergence of the power-series expansion. Unfortunately, the proof is technical and provides little insight. The uniqueness theorem B.9 shows that two analytic functions that are equal on a set with accumulation point are identical in a neighbourhood of the accumulation point. This implies that the analytic continuation of a function is unique.

Theorem B.7 (Taylor) *If $f : \mathbb{C} \to \mathbb{C}$ is an analytic function and equal to $\sum_{k=0}^{\infty} u_k(z-a)^k$ on $|z-a| < R(u)$, then all complex derivatives of f are analytic and equal to*

$$f^{(n)}(z) = \sum_{k=0}^{\infty} \frac{(n+k)!}{k!} u_{n+k}(z-a)^k, \quad \textit{for all } n \geq 0,$$

on $|z-a| < R(u)$. The coefficients of the power-series expansion around a are uniquely determined by

$$u_n = \frac{1}{n!} f^{(n)}(a), \quad \textit{for all } n \geq 0.$$

Proof: By the conversion theorem B.6, the function f is equal to the power series around b

$$f(z) = \sum_{k=0}^{\infty} \left[\sum_{i=0}^{\infty} \binom{i+k}{k} u_{i+k}(b-a)^i \right] (z-b)^k,$$

if $|z-b| + |b-a| < R(u)$. The difference quotient in b is equal to

$$\frac{f(b+d)-f(b)}{d} = \sum_{k=1}^{\infty} \left[\sum_{i=0}^{\infty} \binom{i+k}{k} u_{i+k}(b-a)^i \right] d^{k-1},$$

if $|d| < R(u) - |b-a|$. It has positive radius of convergence, so it is continuous in $d = 0$. Therefore, the complex derivative exists and is equal to a convergent power series around a:

$$f'(b) = \left.\frac{f(b+d) - f(b)}{d}\right|_{d=0} = \sum_{i=0}^{\infty}(i+1)u_{i+1}(b-a)^i,$$

for all $|b-a| < R(u)$. Repeated application shows that all complex derivatives exist and are equal to a convergent power series around a:

$$f^{(n)}(b) = \sum_{i=0}^{\infty}\frac{(i+n)!}{i!}u_{i+n}(b-a)^i, \quad \text{for all } n \geq 0,$$

for all $|b-a| < R(u)$. Substituting $b = a$ renders $f^{(n)}(a) = n!u_n$, for all $n \geq 0$. □

Theorem B.8 (nearest singularity) *If $f : \mathbb{C} \to \mathbb{C}$ is an analytic function and equal to $\sum_{k=0}^{\infty} u_k(z-a)^k$ on $|z-a| < R(u)$, then f has at least one singular point on the circle $|z-a| = R(u)$.*

Proof: Let $\mathcal{C}$ be the circle $|z-a| = R(u)$ and suppose that all points of $\mathcal{C}$ are regular points of f. This assumption will lead to a contradiction. It implies that each point $c \in \mathcal{C}$ has a convergent power series equal to f on a neighbourhood of c with radius $r(c) > 0$. For any $c \in \mathcal{C}$, let $c' \in \mathcal{C}$ be such that $|c - c'| < r(c)/2$. Then, by the conversion theorem B.6, both $r(c') \geq r(c) - |c-c'|$ and $r(c) \geq r(c') - |c-c'|$. Therefore, $|r(c) - r(c')| \leq |c - c'|$ and $r(c)$ is continuous. It is positive on the compact set $\mathcal{C}$ and attains a minimum $r_1 > 0$. Choose r such that $0 < r < r_1$. The function f is analytic on $|z-a| < R(u) + r_1$, so it is continuous on the compact set $|z-a| \leq R(u) + r$ and

$$\alpha = \sup_{|z-a| \leq R(u)+r} |f(z)|$$

is finite.

Choose r_2 such that $0 < r_2 < r_1 - r$. Choose an arbitrary c_2 on the circle $|z-a| = R(u) - r_2$ and let c_1 be the point in $\mathcal{C}$ such that $|c_1 - c_2| = r_2$. The function f is analytic in c_1 and the power-series expansion of f around c_1 has radius of convergence $r(c_1) \geq r_1$. The function f is also analytic in c_2. Let $\sum_{k=0}^{\infty} v_k(z-c_2)^k$ be the power-series expansion of f around c_2 with radius of convergence $R(v)$. By the conversion theorem B.6, $R(v) \geq r(c_1) - |c_1 - c_2| \geq r_1 - r_2 > r$. Applying the Taylor theorem B.7 and Cauchy's estimate from theorem B.5 to the power-series expansion around c_2 renders

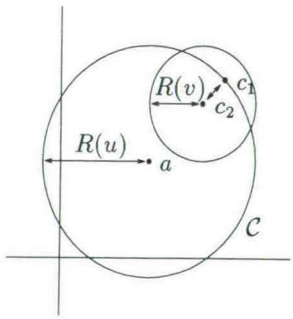

Figure B.1

$$\left|\frac{1}{n!}f^{(n)}(c_2)\right| = |v_n| \leq r^{-n} \sup_{|z-c_2|=r} |f(z)| \leq \alpha r^{-n},$$

for all c_2 in the circle $|z-a| = R(u) - r_2$.

By the Taylor theorem B.7,

$$\frac{1}{n!}f^{(n)}(z) = \sum_{k=0}^{\infty}\binom{n+k}{k}u_{n+k}(z-a)^k,$$

for $|z-a| < R(u)$. Here, applying Cauchy's estimate from theorem B.5 renders

$$\binom{n+k}{k}|u_{n+k}| \leq [R(u) - r_2]^{-k} \sup_{|z-a|=R(u)-r_2}\left|\frac{1}{n!}f^{(n)}(z)\right| \leq [R(u)-r_2]^{-k}\alpha r^{-n}.$$

Multiplying by $[R(u)-r_2]^k r^n$, replacing k by $i-n$, summing over $0 \leq n \leq i$ and letting $r_2 \to 0$ shows that

$$|u_i|\,[R(u)+r]^i \leq \alpha$$

and

$$\limsup_{i\to\infty}|u_i|^{1/i} \leq [R(u)+r]^{-1} < R^{-1}(u).$$

This contradicts the definition of $R(u)$. □

Theorem B.9 (uniqueness) *Let the functions f and g be analytic in a, an accumulation point of $\mathcal{A} \subset \mathbb{C}$. If both functions are identical on $\mathcal{A}$, then they are identical on a neighbourhood of a.*

Proof: Let f and g be analytic on $|z-a| < R$, with power-series expansions $\sum_{k=0}^{\infty} u_k(z-a)^k$ and $\sum_{k=0}^{\infty} v_k(z-a)^k$. Since a is an accumulation point of $\mathcal{A}$, a sequence $\{a_n\}$ converging to a exists such that $a_n \in \mathcal{A}$, $0 < |a_n - a| < R$ and $f(a_n) = g(a_n)$ for all $n \geq 0$.

Both functions are analytic and therefore continuous in a, so

$$u_0 = f(a) = \lim_{n\to\infty} f(a_n) = \lim_{n\to\infty} g(a_n) = g(a) = v_0.$$

Suppose that $u_k = v_k$ for all $0 \leq k \leq K-1$. Then

$$\sum_{k=K}^{\infty} u_k(a_n - a)^k = \sum_{k=K}^{\infty} v_k(a_n - a)^k$$

for all $n \geq 0$. Dividing by $(a_n - a)^K \neq 0$ and letting $n \to \infty$ renders $u_K = v_K$. Therefore, the power-series expansions and thus the functions themselves are identical. □

Example B.2 The function $f(z) = z^2$ is analytic on $\mathbb{C}$, because the power-series expansion $a^2 + 2a(z-a) + (z-a)^2$ converges and equals $f(z)$ for any $z, a \in \mathbb{C}$. The function $g(z) = |z|^2 = z\bar{z}$ is not analytic, as can be seen in the following way. On the real numbers the functions f and g are identical. So if both functions were analytic, they

would be identical on a neighbourhood of any real number. This is not true because g is real, whereas f has non-real values on any open set. Another reason why the function g is not analytic, is because it is complex differentiable in $z = 0$ only:

$$\lim_{r \downarrow 0} \frac{g(z + re^{i\phi}) - g(z)}{re^{i\phi}} = \lim_{r \downarrow 0} ze^{-2i\phi} + \bar{z} + re^{-i\phi} = ze^{-2i\phi} + \bar{z}.$$

Unless z=0, the limit depends on the direction ϕ at which 0 is approached. Therefore, there is not a single point with a neighbourhood on which g is complex differentiable. □

Example B.3 Consider the square-root function. If x is a non-negative real number, then $\sqrt{x}$ is unambiguously defined as the non-negative number such that $(\sqrt{x})^2 = x$. The generalisation to complex values of x is less clear. Let $f : \mathbb{C} \to \mathbb{C}$ be a function such that $f^2(x) = x$ for all $x \in \mathbb{C}$. Then this function maps any point $x = re^{i\phi}$ to either $+\sqrt{r}e^{i\phi/2}$ or $-\sqrt{r}e^{i\phi/2}$. For $0 \leq \psi < 2\pi$, let C_ψ be the set of all complex numbers $\mathbb{C}$ except for the half line starting in the origin at angle ψ with the positive real axis. With appropriate choices of either the plus or the minus sign, the function f can be made continuous on any set C_ψ. However, it can not be made continuous on $\mathbb{C}$. Nor can it be made continuous in the branch point $x = 0$, so surely it is not analytic there.

B.4 Analytic multivariate matrix functions

In the previous section, a scalar function was defined as analytic in a point if it is equal to a convergent power series on a neighbourhood of this point. This can be generalized to multivariate matrix functions $f : \mathbb{C}^\ell \to \mathbb{C}^{m \times n}$. Consider the power series around $a \in \mathbb{C}^\ell$:

$$\sum_{k \in \mathbb{N}^\ell} (z - a)^k U_k, \tag{B.2}$$

with $z \in \mathbb{C}^\ell$, $U_k \in \mathbb{C}^{m \times n}$ for all $k \in \mathbb{N}^\ell$, and $z^k \doteq z_1^{k_1} \times \ldots \times z_\ell^{k_\ell}$. This power series will be called convergent in z if

$$\lim_{i \to \infty} \left\| \sum_{k \in \mathbb{N}^\ell \setminus \mathcal{N}_i} (z - a)^k U_k \right\| = 0,$$

for all sequences of sets $\{o\} \subset \mathcal{N}_0 \subset \mathcal{N}_1 \subset \ldots$ such that $\lim_{i \to \infty} \mathcal{N}_i = \mathbb{N}^\ell$. It is easily shown that the multivariate matrix power series (B.2) converges if the following multivariate scalar power series of positive numbers converges:

$$\sum_{k \in \mathbb{N}^\ell} (z - a)^k \|U_k\|.$$

This is a generalization of absolute convergence. A function $f(z)$ will be called analytic in $a \in \mathbb{C}^\ell$ if it is equal to a convergent power series on a neighbourhood of a. If ℓ, m and

n are finite, like for scalar functions, the concepts of neighbourhood and convergence are unequivocal because all norms are equivalent. If any of ℓ, m and n is infinite, then these concepts depend on the norm that is used. A function can be analytic or not, depending on the norm that is used.

For univariate matrix functions $f : \mathbb{C} \to \mathbb{C}^{m \times n}$, the radius of convergence is equal to

$$R(U) \doteq \left[\limsup_{k \to \infty} \|U_k\|^{1/k} \right]^{-1}. \tag{B.3}$$

This is a generalization of (B.1) and can be proved in a similar way. Consequently, if a univariate matrix function is analytic, then for all $0 < r < R(U)$ an $\alpha > 0$ exists such that $\|U_k\| \leq \alpha r^{-k}$ for all $k \geq 0$.

Appendix C

Extrapolation methods

In the previous appendix analytic functions were discussed, with emphasis on the relation between analyticity of a function and convergence of the power-series expansion. To find the value of an analytic function one can evaluate the power-series expansion. However, convergence may be slow or one may want to evaluate the function outside the convergence region of the power-series expansion. In this appendix, some methods will be discussed to improve convergence properties. This can be either acceleration of convergence or even turning divergent series into convergent series. These method are reviewed here because they are essential for an efficient implementation of the PSA. For a more elaborate overview of extrapolation methods, see the textbooks by Baker and Graves-Morris [12,13] and by Brezinski and Redivo Zaglia [34,35].

An extrapolation method converts a sequence $\{S_k\}_{k\geq 0}$ with limit S, into a new sequence $\{T_k\}_{k\geq 0}$ with limit T. In this thesis, the elements of these sequences will usually be the partial sums of a power series. If the new limit T is equal to S, then the extrapolation method is called *regular*. A regular method is called *accelerating* if the new sequence converges faster to S than the original, that is if $\lim_{k\to\infty}(T_k - S)/(S_k - S) = 0$. As examples, consider the following two transformations:

$$T_k = \frac{S_{k+1} + S_k}{2}, \tag{C.1}$$

and

$$T_k = S_k - \frac{(S_{k+1} - S_k)^2}{S_{k+2} - 2S_{k+1} + S_k}. \tag{C.2}$$

The first linear transform (C.1) is very natural if $S_k - S$ is alternating. Clearly it is regular on any convergent sequence. However, if the original sequence is monotone, then it is not accelerating. The second transform (C.2) is known as Aitken's Δ^2 method. It is especially suitable for sequences that are mainly geometrically convergent. If $S_k = S + \alpha\beta^k$, with $\alpha \neq 0$ and $|\beta| < 1$, then it immediately finds the right limit S. However, the method can not be applied if the original sequence is constant. These examples illustrate that particular methods may work well on certain types of sequences, but poorly on others. In fact, it can be shown that no extrapolation method exists that accelerates all sequences (see section 1.10 in [35]). Therefore, the best method does not exist and it makes sense to use different types of methods.

Extrapolation methods should be used with great caution. Consider the sequence of numbers 1, 0, 1, 0. In an IQ test, the correct guess of the subsequent number would undoubtedly be 1. However, life is no IQ test. Given these first 4 elements, the k-th element of the sequence could have been generated by the equation $(24 - 34k + 15k^2 - 2k^3)/3$. Then the next number would be -7. Or maybe there is no regularity at all and the next number could be any number. Another reason to be careful with extrapolation methods is that sequences often have small errors. For example, consider the case where the original sequence is constant up to rounding errors. Applying an extrapolation method would be trying to find regularity in meaningless noise, possibly with disastrous results.

However, practice has shown that extrapolation methods can be of great help in accelerating convergence. Also in combination with the PSA, they have proved to be very effective. The amount of work necessary to compute the coefficients is rapidly increasing in the number of coefficients. From this, it will be obvious that it is worthwhile to obtain as much information from the computed coefficients as possible.

C.1 Bilinear mapping

The radius of convergence of a power series around the origin is equal to the distance between the origin and the nearest singularity of the function defined by the power series (see theorem B.8). Therefore, the radius of convergence can be enlarged by moving singularities further away. This is possible with the origin preserving bilinear mapping

$$y = \frac{ax}{1+bx}, \quad x = \frac{y}{a-by}, \quad a \geq 1,\ b \geq 0. \tag{C.3}$$

From a power-series expansion in x, the corresponding power-series expansion in y can be obtained :

$$\begin{aligned} S(x) &= \sum_{k=0}^{\infty} x^k u_k \\ &= u_0 + \sum_{k=1}^{\infty} \left[\frac{y}{a-by}\right]^k u_k \\ &= u_0 + \sum_{k=1}^{\infty} \left[\sum_{\ell=0}^{\infty} \binom{\ell+k-1}{k-1} a^{-k-\ell} b^\ell y^{k+\ell}\right] u_k \\ &= u_0 + \sum_{k=1}^{\infty} y^k \sum_{\ell=1}^{k} \binom{k-1}{\ell-1} a^{-k} b^{k-\ell} u_\ell \\ &= \sum_{k=0}^{\infty} y^k v_k \;=\; T(y). \end{aligned}$$

To calculate the k-th coefficient of the expansion in y, only the first k coefficients of the expansion in x are needed. Therefore, the K-th partial sum $T_K(y) = \sum_{k=0}^{K} y^k v_k$ can be obtained from the coefficients of the K-th partial sum $S_K(x) = \sum_{k=0}^{K} x^k u_k$. This is not true for the general bilinear mapping $y = \frac{a+bx}{c+dx}$, also called the Möbius transformation, which is not origin preserving.

Suppose that $S(x)$ is analytic at $x = 0$, with radius of convergence $R(u)$. Then, because of Cauchy's estimate in theorem B.5, the coefficients of the power-series expansion have a geometric bound $|u_k| \leq \alpha r^{-k}$, with $\alpha \geq 0$ and $0 < r < R(u)$. Subsequently, the Cauchy-Hadamard theorem B.3 shows that the radius of convergence $R(v)$ of the power series in y is positive:

$$\begin{aligned} R^{-1}(v) &= \limsup_{k\to\infty} \left|\sum_{\ell=1}^{k} \binom{k-1}{\ell-1} a^{-k} b^{k-\ell} u_\ell\right|^{1/k} \\ &\leq \limsup_{k\to\infty} \left[\sum_{\ell=1}^{k} \binom{k-1}{\ell-1} a^{-k} b^{k-\ell} \alpha r^{-\ell}\right]^{1/k} \\ &= \limsup_{k\to\infty} \frac{b}{a}\left[\frac{\alpha}{br}\left(1+\frac{1}{br}\right)^{k-1}\right]^{1/k} \\ &= \frac{1+br}{ar} \;<\; \infty. \end{aligned}$$

Therefore, also $T(y)$ is analytic at $y = 0$. In a similar way, it can be shown that if $S(x)$ is analytic (singular) at $x = x^* \neq 0$, then $T(y)$ is analytic (singular) at $y = y^* = \frac{ax^*}{1+bx^*} \neq 0$ (provided $x^* \neq -b^{-1}$). In particular, if $S(x)$ has a k-th order pole with residue R at $x = x^*$, then $T(y)$ has a k-th order pole at $y = y^*$:

$$\lim_{y\to y^*} (y-y^*)^k T(y) = \lim_{x\to x^*} \left[\frac{a(x-x^*)}{(1+bx)(1+bx^*)}\right]^k S(x) = \frac{a^k R}{(1+bx^*)^{2k}}.$$

The way the coefficients of the power-series expansion of $T(y)$ are calculated from those of $S(x)$ immediately shows that if the first K coefficients of $S(x)$ are zero, then also the first K coefficients of $T(y)$ are zero. In other words: if $S(x) \in \mathcal{O}\left(x^K\right)$, for $x \downarrow 0$, then also $T(y) \in \mathcal{O}\left(y^K\right)$, for $y \downarrow 0$.

The previous paragraph shows that, if the original power series has positive radius of convergence then so has the new power series. Now, it will be shown that the mapping (C.3) can be used to enlarge the radius of convergence. Define

$$\begin{aligned}
\mathcal{I} &= [0, \tfrac{a-1}{b}), \\
\mathcal{D}_1 &= \left\{ z \in \mathbb{C} \;\middle|\; \left| z - \tfrac{(a-1)^2}{b(2a-1)} \right| < \tfrac{a(a-1)}{b(2a-1)} \right\}, \\
\mathcal{D}_2 &= \left\{ z \in \mathbb{C} \;\middle|\; |z| < \tfrac{a-1}{b} \right\}.
\end{aligned}$$

The mapping maps the interval $\mathcal{I}$ onto itself and maps the disk $\mathcal{D}_1$ onto the disk $\mathcal{D}_2$. If $a > 1$ and $b > 0$, then the disk $\mathcal{D}_1$ is inside the disk $\mathcal{D}_2$. By letting $b \to \infty$ with $\frac{a-1}{b}$ constant, $\mathcal{D}_1$ can be made arbitrarily close to the smallest disk around the interval $\mathcal{I}$. The region $\mathcal{D}_2 \backslash \mathcal{D}_1$ is mapped outside $\mathcal{D}_2$, including all singularities of $S(x)$ in this region. Suppose that $S(x)$ has singularities in $\mathcal{D}_2 \backslash \mathcal{D}_1$, but not in $\mathcal{D}_1$. Then $R(u) < \frac{a-1}{b}$, so the power series in x will diverge for large values of x in $\mathcal{I}$. Yet $R(v) > \frac{a-1}{b}$, so for the y corresponding to these larger values of x, the power series in y does converge. Cauchy's estimate in theorem B.5 shows that the coefficients of the power series have growth rate at most equal to the reciprocal of the distance to the nearest singularity. This indicates that applying the mapping to move singularities further away from the origin will reduce the growth rate of the coefficients.

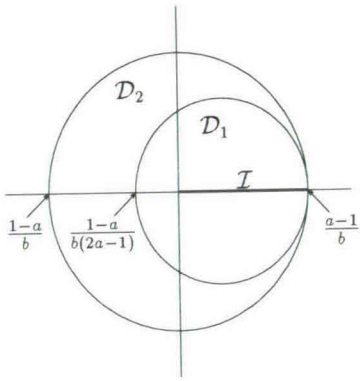

Figure C.1

C.2 Value and pole extrapolation

Possible knowledge about values and poles of the function $f : \mathbb{R} \to \mathbb{R}$ can be used to find better approximations. For example, in the section on multipoint Padé approximations of [13] this information is used to find better Padé approximants. The approach here will be more straightforward. The starting point will be the K-th truncated power-series expansion

$$f_K(x) = \sum_{k=0}^{K} x^k u_k = f(x) + O(x^{K+1}).$$

New approximations $g_K(x)$ of the function $f(x)$ will be obtained that also satisfy $g_K(x) = f(x) + O(x^{K+1})$, but *with* the specified values or poles.

C.2.1 Value extrapolation

Consider first the problem of finding an approximant that has function value y_i at x_i, for $1 \leq i \leq I$. A new polynomial approximant will be obtained with order $K + I$:

$$g_K(x) = f_K(x) + \sum_{i=1}^{I} x^{K+i} v_i.$$

Requiring that $g_K(x_i) = y_i$ for $1 \leq i \leq I$ leads to the set of equations

$$y_i = f_K(x_i) + \sum_{i=1}^{I} {x_i}^{K+i} v_i, \quad \text{for } 1 \leq i \leq I.$$

This can be written as the linear set of equations

$$\begin{pmatrix} 1 & x_1 & \dots & x_1^{I-1} \\ 1 & x_2 & \dots & x_2^{I-1} \\ \vdots & \vdots & & \vdots \\ 1 & x_I & \dots & x_I^{I-1} \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_I \end{pmatrix} = \begin{pmatrix} x_1^{-K-1}\,[y_1 - f_K(x_1)] \\ x_2^{-K-1}\,[y_2 - f_K(x_2)] \\ \vdots \\ x_I^{-K-1}\,[y_I - f_K(x_I)] \end{pmatrix}.$$

The RHS is well-defined if all x_i are non-zero. The matrix in the LHS is a Vandermonde matrix and non-singular if all x_i are different. Then all coefficients $v_1,\ v_2, \dots, v_I$ can be calculated. Both restrictions do not really restrict the applicability.

C.2.2 Pole extrapolation with known residues

Suppose that it is known that the function f has a pole at $x = x_i$ with order r_i and residue R_i, for $1 \leq i \leq I$. For definitions, see section B.3 on analytic functions. Then choose

$$g_K(x) = f_K(x) + \frac{\sum_{j=1}^{I} x^{K+j} v_j}{\prod_{j=1}^{I} (x - x_j)^{r_j}}.$$

This new approximant has poles of the same order at the same places. The unknown coefficients $v_1, \dots, v_I$ are determined by requiring that also the residues are the same:

$$\lim_{x \to x_i} (x - x_i)^{r_i} g_K(x) = \frac{\sum_{j=1}^{I} x_i^{K+j} v_j}{\prod_{j=1,\ j \neq i}^{I} (x_i - x_j)^{r_j}} = R_i.$$

Multiplying by the denominator leads to the set of equations

$$\begin{pmatrix} 1 & x_1 & \dots & x_1^{I-1} \\ 1 & x_2 & \dots & x_2^{I-1} \\ \vdots & \vdots & & \vdots \\ 1 & x_I & \dots & x_I^{I-1} \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_I \end{pmatrix} = \begin{pmatrix} R_1\, x_1^{-K-1} \prod\limits_{j=1,\ j\neq 1}^{I} (x_1 - x_j)^{r_j} \\ R_2\, x_2^{-K-1} \prod\limits_{j=1,\ j\neq 2}^{I} (x_2 - x_j)^{r_j} \\ \vdots \\ R_I\, x_I^{-K-1} \prod\limits_{j=1,\ j\neq I}^{I} (x_I - x_j)^{r_j} \end{pmatrix}.$$

The matrix is again a Vandermonde matrix and the coefficients can be calculated if all x_i are different and non-zero. Notice that the coefficients do not depend on K or the coefficients $u_0, \dots, u_K$.

C.2.3 Pole extrapolation with unknown residue

Sometimes it is known that a function has an r-th order pole at $x = a$, but with unknown residue R. If this pole is the smallest singularity, then an estimate of the residue can be found from the calculated coefficients $u_0, \dots, u_K$. This was implicitly done in [20]. Suppose that, except for the pole in a, $f(x)$ is analytic on $|x| \leq S$ with $S > |a|$. Then $f(x)$ can be written as

$$\begin{aligned} \sum_{k=0}^{\infty} x^k u_k &= \frac{R}{(x-a)^r} + g(x) \\ &= \frac{R}{(x-a)^r} + \sum_{k=0}^{\infty} x^k v_k \\ &= \sum_{k=0}^{\infty} x^k \left[\binom{k+r-1}{r-1} R a^{-k-r} + v_k \right]. \end{aligned}$$

The function $g(x)$ is analytic on $|x| \leq S$. According to Cauchy's estimate in theorem B.5, an α exists such that $|v_k| \leq \alpha S^{-k}$. Therefore,

$$\lim_{k\to\infty} \left| a^{k+r} u_k - \binom{k+r-1}{r-1} R \right| = \lim_{k\to\infty} \left| a^{k+r} v_k \right| \leq \lim_{k\to\infty} \left| a^{k+r} \alpha S^{-k} \right| = 0,$$

and thus

$$\lim_{k\to\infty} \Delta^{r-1} a^{k+r} u_k = \lim_{k\to\infty} \Delta^{r-1} \binom{k+r-1}{r-1} R = R. \tag{C.4}$$

For a definition of the difference operator Δ^{r-1}, see section 1.4. Formula (C.4) can be used to obtain an estimate $\hat{R}$ of R. For example, if the pole is a simple pole and the coefficients u_k have been calculated for $0 \leq k \leq K$, then $\hat{R} = a^{K+1} u_K$ can be used as an estimate of R. If the pole has order 2, then the estimate $\hat{R} = a^{K+1} [a u_K - u_{K-1}]$ can be used. In fact, extrapolation methods can be used to increase the convergence speed of this estimate.

The obtained estimate of the residue can be used in the approach of the previous section C.2.2. This leads to the new approximation

$$g_K(x) = f_K(x) + \frac{\hat{R}\left(\frac{x}{a}\right)^{K+1}}{(x-a)^r}.$$

But better estimates are obtained in the following way. Apart from a way to estimate the residue, formula (C.4) also shows that, for large k, the coefficients u_k will behave like a polynomial in k with order $r-1$:

$$u_k \approx \sum_{s=0}^{r-1} k^s w_s. \tag{C.5}$$

Estimates of the coefficients $w_0, \ldots, w_{r-1}$ can be obtained by requiring that equality holds in formula (C.5), for the last r calculated coefficients $u_{K-r+1}, \ldots, u_K$. This leads to the set of equations

$$\begin{pmatrix} 1 & K-r+1 & \ldots & (K-r+1)^{r-1} \\ 1 & K-r+2 & \ldots & (K-r+2)^{r-1} \\ \vdots & \vdots & & \vdots \\ 1 & K & \ldots & K^{r-1} \end{pmatrix} \begin{pmatrix} w_0 \\ w_1 \\ \vdots \\ w_{r-1} \end{pmatrix} = \begin{pmatrix} u_{K-r+1} \\ u_{K-r+2} \\ \vdots \\ u_K \end{pmatrix}.$$

The matrix is an invertible Vandermonde matrix, so a unique solution exists. If r was chosen too large, then the higher order coefficients will be approximately zero. The obtained coefficients $w_0, \ldots, w_{r-1}$ can be used to estimate all uncalculated coefficients $u_{K+1}, u_{K+2}, \ldots$:

$$\begin{aligned} G_K(x) &= \sum_{k=0}^{K} x^k u_k + \sum_{k=K+1}^{\infty} x^k \sum_{s=0}^{r-1} k^s v_s \\ &= \sum_{k=0}^{K} x^k u_k + \sum_{s=0}^{r-1} v_s \sum_{k=K+1}^{\infty} x^k k^s \end{aligned}$$

The infinite summations can be expressed in closed form:

$$\sum_{k=K+1}^{\infty} x^k k^s = \begin{cases} \frac{x^{K+1}}{(1-x)}, & \text{if } s = 0, \\ \frac{x^{K+1}}{(1-x)^2}\left[(K+1) - Kx\right], & \text{if } s = 1, \\ \frac{x^{K+1}}{(1-x)^3}\left[(K+1)^2 - (2K^2+2K-1)x + K^2x^2\right], & \text{if } s = 2. \end{cases}$$

Similar expressions are available for higher values of s.

C.3 The epsilon, theta and Levin algorithms

This section contains a description of three extrapolation methods. The epsilon algorithm is explained in some detail, based on the close connection with Padé approximants.

The theta and Levin algorithms are only briefly discussed, but in enough detail to implement the algorithms.

The idea of Padé approximation is to approximate a function $f(x) = \sum_{k=0}^{\infty} x^k u_k$ by the quotient of two polynomials, of order say L and M. Like with the procedures in the previous section, singularities can be included in the denominator. The difference is that the singularities are not specified in advance. The polynomials are chosen in such a way that the power-series expansion of their quotient, denoted by $[L/M]_f(x)$, coincides with the power-series expansion of the function, up to coefficient $K = L + M$:

$$[L/M]_f(x) = \frac{\sum\limits_{\ell=0}^{L} x^\ell v_\ell}{\sum\limits_{m=0}^{M} x^m w_m} = \sum_{k=0}^{K} x^k u_k + O\left(x^{K+1}\right). \tag{C.6}$$

Multiplying by the denominator renders

$$\begin{aligned} \sum_{\ell=0}^{L} x^\ell v_\ell &= \sum_{m=0}^{M} \sum_{k=0}^{K} x^{m+k} w_m u_k + O\left(x^{K+1}\right) \\ &= \sum_{\ell=0}^{K} x^\ell \sum_{m=0}^{min\{M,\ell\}} w_m u_{\ell-m} + O\left(x^{K+1}\right). \end{aligned}$$

Equating the corresponding coefficients of the first K powers of x renders the set of equations

$$v_\ell = \sum_{m=0}^{min\{M,\ell\}} w_m u_{\ell-m}, \quad \text{for } 0 \le \ell \le L, \tag{C.7}$$

$$0 = \sum_{m=0}^{min\{M,\ell\}} w_m u_{\ell-m}, \quad \text{for } L+1 \le \ell \le L+M. \tag{C.8}$$

If the normalization $w_0 = 1$ is used, this set of equations is a set of $L + M + 1$ linear equations with the $L+M+1$ unknowns $v_0, \ldots, v_L$ and $w_1, \ldots, w_M$. These equations can be solved by first solving the second set (C.8), which can be written as

$$\begin{pmatrix} u_{L-M+1} & u_{L-M+2} & u_{L-M+3} & \cdots & u_L \\ u_{L-M+2} & u_{L-M+3} & u_{L-M+4} & \cdots & u_{L+1} \\ u_{L-M+3} & u_{L-M+4} & u_{L-M+5} & \cdots & u_{L+2} \\ \vdots & \vdots & \vdots & & \vdots \\ u_L & u_{L+1} & u_{L+2} & \cdots & u_{L+M-1} \end{pmatrix} \begin{pmatrix} w_M \\ w_{M-1} \\ w_{M-2} \\ \vdots \\ w_1 \end{pmatrix} = - \begin{pmatrix} u_{L+1} \\ u_{L+2} \\ u_{L+3} \\ \vdots \\ u_{L+M} \end{pmatrix}, \tag{C.9}$$

with $u_k = 0$ for all $k < 0$. Next, the coefficients $v_0, \ldots, v_L$ can easily be calculated from the first set of equations (C.7).

The set of equations (C.9) is not always non-singular. In other words, it is not always possible to find a Padé approximant with specified degrees of the polynomials.

Take, for example, an analytic function with constant term u_0 equal to 0, so $f(0) = 0$. For any $[0/K]_f(x)$ Padé approximant, either $[0/K]_f(0) \neq 0$ or $[0/K]_f(x) \equiv 0$. So, in general, the defining equation (C.6) will not be satisfied. A more interesting example is the function $f(x) = 1 + x^2$. It is not possible to find coefficients v_0, v_1 and w_0, w_1 such that

$$\frac{v_0 + v_1 x}{w_0 + w_1 x} = 1 + x^2 + O\left(x^3\right).$$

In other words, a $[1/1]_f(x)$ Padé approximant does not exist for this function.

The $[L/0]_f(x)$ Padé approximant is simply the L-th partial sum of the power series: $w_0 = 1$ and $v_\ell = u_\ell$ for all $0 \leq \ell \leq L$. For the $[0/M]_f(x)$ approximant, the set of equations (C.9) is triangular with solution

$$\begin{aligned} w_0 &= 1, \\ w_m &= \sum_{\ell=1}^{m} u_\ell w_{m-\ell}, \quad \text{for } 1 \leq m \leq M. \end{aligned} \tag{C.10}$$

The value of the coefficients does not depend on M. Together with $v_0 = u_0$, this easily renders any of the $[0/M]_f(x)$ Padé approximants.

The anti-diagonals of the matrix in (C.9) contain only identical elements. This special structure makes it possible to calculate Padé approximants more efficiently than with standard equation solvers. One way to do this is by the epsilon algorithm, originated by Shanks [124] and Wynn [137,138]. It does not produce the coefficients of the polynomials in (C.6), but the value of the quotient for a particular value of x. The epsilon algorithm calculates a two-dimensional triangular array:

$$\begin{array}{c|ccccccccc}
 & \epsilon_0^{(-1)} & & \epsilon_2^{(-2)} & & \epsilon_4^{(-3)} & & \epsilon_6^{(-4)} & \cdots \\
\epsilon_{-1}^{(0)} & & \epsilon_1^{(-1)} & & \epsilon_3^{(-2)} & & \epsilon_5^{(-3)} & & \\
\hline
 & \epsilon_0^{(0)} & & \epsilon_2^{(-1)} & & \epsilon_4^{(-2)} & & & \\
\epsilon_{-1}^{(1)} & & \epsilon_1^{(0)} & & \epsilon_3^{(-1)} & & & & \\
 & \epsilon_0^{(1)} & & \epsilon_2^{(0)} & & & & & \\
\epsilon_{-1}^{(2)} & & \epsilon_1^{(1)} & & \ddots & & & & \\
 & \epsilon_0^{(2)} & & & & & & & \\
\epsilon_{-1}^{(4)} & & & & & & & & \\
\vdots & & & & & & & &
\end{array}$$

The elements of this epsilon table are calculated as follows:

The epsilon algorithm

$$\begin{aligned} \epsilon_{-1}^{(k)} &= \epsilon_{2k}^{(-k-1)} = 0, && k \geq 0, \\ \epsilon_0^{(k)} &= \sum_{\ell=0}^{k} x^\ell u_\ell, && k \geq 0, \\ \epsilon_{\ell+1}^{(k)} &= \epsilon_{\ell-1}^{(k+1)} + \left[\epsilon_\ell^{(k+1)} - \epsilon_\ell^{(k)}\right]^{-1}, && \ell \geq 0,\ k \geq -\ell/2. \end{aligned} \tag{C.11}$$

The first column and top row are zero. The second column contains the partial sums. From these all other elements can be calculated. For example, $\epsilon_2^{(0)}$ is calculated from $\epsilon_0^{(1)}$, $\epsilon_1^{(1)}$ and $\epsilon_1^{(0)}$. The elements $\epsilon_{2\ell}^{(k)}$ in the even columns can be shown to be equal to the Padé approximants $[\ell + k/\ell]_f(x)$. The elements in the odd columns are intermediate results with no interesting interpretation. Problems arise when a particular Padé approximant does not exist. Then the denominator in the last equation of (C.11) is zero. For a detailed discussion of such problems, see the textbooks mentioned before [12,13,35].

From the first K coefficients of a power-series expansion, all the approximants $[L/M]_f(x)$ can be calculated with $L + M \leq K$, provided they exist. It is customary to choose the approximant with equal degree of the numerator and denominator ($L = M$), unless there is reason to do otherwise. In that case the elements $\epsilon_\ell^{(k)}$ with $k < 0$ need not be calculated. At the end of this section a different selection procedure will be explained that can also be used to choose between different types of extrapolation methods.

Applying a speed-up procedure to the epsilon algorithm, a new algorithm can be derived called the theta algorithm:

The theta algorithm

$$\begin{aligned}
&\theta_{-1}^{(k)} = \theta_{2k}^{(-k-1)} = 0, && k \geq 0,\\
&\theta_0^{(k)} = \sum_{\ell=0}^{k} x^\ell u_\ell, && k \geq 0,\\
&\theta_{2\ell+1}^{(k)} = \theta_{2\ell-1}^{(k)} + \left[\theta_{2\ell}^{(k+1)} - \theta_{2\ell}^{(k)}\right]^{-1}, && \ell \geq 0,\ k \geq -\ell - 1,\\
&\theta_{2\ell+2}^{(k)} = \theta_{2\ell}^{(k+1)} + \frac{\left[\theta_{2\ell}^{(k+2)} - \theta_{2\ell}^{(k+1)}\right]\left[\theta_{2\ell+1}^{(k+2)} - \theta_{2\ell+1}^{(k+1)}\right]}{\theta_{2\ell+1}^{(k+2)} - 2\theta_{2\ell+1}^{(k+1)} + \theta_{2\ell+1}^{(k)}}, && \ell \geq 0,\ k \geq -\ell - 1.
\end{aligned} \tag{C.12}$$

The difference with the epsilon algorithm is in the rule for the calculation of the even columns. As a result, the number of columns that can be calculated from the same first K coefficients is less. The odd columns are intermediate results, the even columns estimates of S. For an extensive discussion of the theta algorithm, see [35].

A third algorithm is the Levin transform [98,35], defined by

$$L_\ell^{(k)} = \frac{\sum_{i=0}^{\ell}(-1)^i\binom{\ell}{i}(i+k+1)^{\ell-1}\dfrac{S_{k+i}}{R_{k+i}}}{\sum_{i=0}^{\ell}(-1)^i\binom{\ell}{i}(i+k+1)^{\ell-1}\dfrac{1}{R_{k+i}}}.$$

The number R_k is an estimates of the error $S_k - S$. Levin suggests three different estimates, leading to three different versions of the transform:

$$\begin{aligned}
t\text{-transform}: &\quad R_k = S_k - S_{k-1},\\
u\text{-transform}: &\quad R_k = k\,(S_k - S_{k-1}),\\
v\text{-transform}: &\quad R_k = -\frac{(S_k - S_{k-1})\,(S_{k+1} - S_k)}{S_{k+1} - 2S_k + S_{k-1}},
\end{aligned}$$

with $S_{-1} \doteq 0$. The different estimates are designed for different types of series: the t-transform for alternating series, the u-transform for monotonic series and the v-transform is inspired by Aitken's Δ^2 method (C.2).

If the partial sums S_k are available for $0 \le k \le K$, then experience suggests that $L_K^{(0)}$ is the preferred transform. An efficient way of calculation is as follows:

The Levin algorithm

$$\begin{aligned} B_k &= \left(\tfrac{k+1}{K+1}\right)^{K-1} \frac{S_k}{R_k}, && \text{for } 0 \le k \le K, \\ C_k &= \left(\tfrac{k+1}{K+1}\right)^{K-1} \frac{1}{R_k}, && \text{for } 0 \le k \le K, \\ L_K^{(0)} &= \frac{\Delta^K B_0}{\Delta^K C_0}. \end{aligned} \tag{C.13}$$

Notice that the method is not applicable if any of the R_k is zero.

Given these different extrapolation methods, which one is the appropriate choice? Smith and Ford [127,128] compared a number of extrapolation methods, applying them to a wide range of different sequences. Among others, they considered repeated application of Aitken's Δ^2 method, the epsilon and theta algorithms and Levin's t, u and v transforms. None of the methods was uniformly best and each method was best on some sequence. As rules of thumb, they suggest to use

- Levin's u-transform for alternating series, both convergent and divergent,
- The epsilon algorithm for linearly convergent series, monotone divergent series and series with irregular sign patterns,
- The theta algorithm or Levin's u-transform for logarithmically convergent series.

(If a series $\{S_k\}$ is *logarithmically, linearly* or *higher order* convergent, then the limit $\rho = \lim_{k\to\infty} |S_{k+1} - S| \,/\, |S_k - S|$ satisfies $\rho = 1$, $0 < \rho < 1$ or $\rho = 0$, respectively.) There are too many exceptions to these rules of thumb to use them blindly.

Another approach is to use several procedures and select the best. The problem here is how to determine which is best, because the correct value S is unknown. One possible selection criterion is the amount of relative variation in the last κ elements of the extrapolated sequences:

$$\frac{\max\limits_{K-\kappa+1\le k\le K} |T_k| - \min\limits_{K-\kappa+1\le k\le K} |T_k|}{\min\limits_{K-\kappa+1\le k\le K} |T_k|}. \tag{C.14}$$

A warning is in order. If the extrapolated sequence converges fast to the right limit, then the relative variation will soon be small. However, it will also be small if the sequence converges fast to a wrong limit or if it converges slowly but monotonically.

C.4 Multivariate extrapolation methods

The idea of Padé approximation can easily be generalized to multivariate functions [13]. Let $f : \mathbb{R}^S \to \mathbb{R}$ be a function that can be represented by the power-series expansion

$$f(x) = \sum_{k \in \mathbb{N}^S} x^k u_k, \quad \text{with } x^k \doteq x_1^{k_1} \times \ldots \times x_\ell^{k_\ell}$$

Let $\mathcal{K}$, $\mathcal{L}$ and $\mathcal{M}$ be subsets of $\mathbb{N}^S$. Multivariate Padé approximants can be defined as the quotient of truncated multivariate power series such that

$$[\mathcal{K}/\mathcal{L}/\mathcal{M}]_f(x) = \frac{\sum_{l \in \mathcal{L}} x^l v_l}{\sum_{m \in \mathcal{M}} x^m w_m} = \sum_{k \in \mathcal{K}} x^k u_k + \sum_{k \in \mathbb{N}^S \setminus \mathcal{K}} x^k d_k. \tag{C.15}$$

With normalization $w_o = 1$ and $\#\mathcal{L} + \#\mathcal{M} = \#\mathcal{K} + 1$, multiplying by the denominator normally renders a set of linear equations with an equal number of equations and variables. In the univariate case, the choice $\mathcal{K} = \{0, \ldots, L + M\}$, $\mathcal{L} = \{0, \ldots, L\}$ and $\mathcal{M} = \{0, \ldots, M\}$ is obvious. In the multivariate case there is no analogous obvious choice. Also, even for special choices of $\mathcal{K}$, $\mathcal{L}$ and $\mathcal{M}$, the equations generally do not have a simple structure. Complex and lengthy computer programs are required. Due to these additional difficulties, the progress with multivariate Padé approximation has been less satisfactory than in the univariate case.

A less efficient but simple and flexible alternative is to reduce the multivariate case to the univariate. Let $\{\mathcal{N}_\ell\}_{\ell \geq 0}$ be a partition of $\mathbb{N}^S$, with each $\mathcal{N}_\ell$ finite. For a given $x \in \mathbb{R}^S$, define the following function $F : \mathbb{R} \to \mathbb{R}$:

$$F(y) = \sum_{\ell=0}^{\infty} y^\ell v_\ell, \quad \text{with } v_\ell = \sum_{k \in \mathcal{N}_\ell} x^k u_k. \tag{C.16}$$

Then

$$F(1) = \sum_{\ell=0}^{\infty} v_\ell = \sum_{\ell=0}^{\infty} \sum_{k \in \mathcal{N}_\ell} x^k u_k = \sum_{k \in \mathbb{N}^S} x^k u_k = f(x).$$

The partial sums of $F(1)$ correspond to partial sums of $f(x)$ and univariate extrapolation methods can be applied to the partial sums of $F(1)$. The choice of the sets $\{\mathcal{N}_\ell\}$ is not obvious. One possible choice are the diagonal sets

$$\mathcal{N}_\ell = \left\{ n \in \mathbb{N}^S \;\middle|\; \sum_{s=1}^{S} n_s = \ell \right\}.$$

For this choice, Genz [58] has shown that (C.16) is of the type (C.15), but with $\#\mathcal{L} + \#\mathcal{M} < \#\mathcal{K} + 1$. Other possibilities are the square sets

$$\mathcal{N}_\ell = \left\{ n \in \mathbb{N}^S \;\middle|\; \max_{1 \leq s \leq S} n_s = \ell \right\}$$

or asymmetric versions.

Bibliography

[1] Abate, J., W. Whitt, Calculating time-dependent performance measures for the M/M/1 queue, *IEEE Trans. Comm.* **37** (1989), 1102-1104.

[2] Abate, J., W. Whitt, The Fourier-series method for inverting transforms of probability distributions, *Queueing Systems* **10** (1992), 5-88.

[3] Abate, J., M. Kijima, W. Whitt, Decomposition of the M/M/1 transition function, *Queueing Systems* **9** (1991), 323-336.

[4] Abdallah, H., R. Marie, The uniformized power method for transient solutions of Markov processes, *Computers Opns Res.* **20** (1993), 515-526.

[5] Adan, I.J.B.F., J. Wessels, W.H.M. Zijm, Analysis of the symmetric shortest queue problem, *Comm. Statist. Stochastic Models* **6** (1990), 691-713.

[6] Adan, I.J.B.F., *A Compensation Approach for Queueing Problems*, CWI Tract 104, Stichting Mathematisch Centrum, Amsterdam, 1994.

[7] Asmussen, S., *Applied Probability and Queues*, Wiley, Chichester (1987).

[8] Asmussen, S., G. Koole, Marked point processes as limits of Markovian arrival streams, *J. Appl. Prob.* **30** (1993), 365-372.

[9] Baccelli, F., G. Cohen, G.J. Olsder, J.P. Quadrat, *Synchronization and Linearity: An Algebra for Discrete Event Systems*, Wiley & Sons, Chichester, 1992.

[10] Baccelli, F., W.A. Massey, A sample path analysis of the M/M/1 queue, *J. Appl. Prob.* **26** (1989), 418-422.

[11] Baccelli, F., V. Schmidt, Taylor expansions for Poisson driven (max,+)-linear systems, INRIA report 2494, Sophia Antipolis, France, 1995.

[12] Baker, G.A., P. Graves-Morris, *Padé Approximants, Part I: Basic Theory*, Addison-Wesley Publishing Company, London, 1981.

[13] Baker, G.A., P. Graves-Morris, *Padé Approximants, Part II: Extensions and Applications*, Addison-Wesley Publishing Company, London, 1981.

[14] Barker, V.A., Numerical solution of sparse singular systems of equations arising from ergodic Markov chains, *Comm. Statist. Stochastic Models* **5** (1989), 335-381.

[15] Bavinck, H., G. Hooghiemstra, E. de Waard, An application of Gegenbauer polynomials in queueing theory, *J. Comput. Appl. Math.* **49** (1993), 1-10.

[16] Beneš, V.E., *Mathematical Theory of Connecting Networks and Telephone Traffic*, Academic Press, New York, 1965.

[17] Blanc, J.P.C., The transient behaviour of networks with infinite server nodes, in *Performance '84*, ed. E. Gelenbe, Elsevier Science Publishers B.V., North Holland, 1984, 159-174.

[18] Blanc, J.P.C., The relaxation time of two queueing systems in series, *Comm. Statist. Stochastic Models* **1** (1985), 1-16.

[19] Blanc, J.P.C., A note on waiting times in systems with queues in parallel, *J. Appl. Prob.* **24** (1987), 540-546.

[20] Blanc, J.P.C., On a numerical method for calculating state probabilities for queueing systems with more than one waiting line, *J. Comput. Appl. Math.* **20** (1987), 119-125.

[21] Blanc, J.P.C., On the relaxation time of open queueing networks, in *Queueing Theory and its Applications - Liber Amicorum for J.W. Cohen*, eds. O.J. Boxma, R. Syski, CWI Monograph 7, North-Holland, Amsterdam, 1988, 235-259.

[22] Blanc, J.P.C., A numerical approach to cyclic-service queueing models, *Queueing Systems* **6** (1990), 173-188.

[23] Blanc, J.P.C., Performance evaluation of polling systems by means of the power-series algorithm, *Annals of Operations Research* **35** (1992), 155-186.

[24] Blanc, J.P.C., The power-series algorithm applied to the shortest-queue model, *Oper. Res.* **40** (1992), 157-167.

[25] Blanc, J.P.C., Performance analysis and optimization with the power-series algorithm, in *Performance Evaluation of Computer and Communication Systems*, eds. L. Donatiello, R. Nelson, Springer-Verlag, Berlin (1993), 53-80.

[26] Blanc, J.P.C., E.A. van Doorn, Relaxation times for queueing systems, in: *Mathematics and Computer Science*, eds. J.W. de Bakker, M. Hazewinkel, J.K. Lenstra, CWI Monograph 1, North-Holland, Amsterdam (1986), 139-162.

[27] Blanc, J.P.C., R.D. van der Mei, Optimization of polling systems by means of gradient methods and the power-series algorithm, Tilburg University, *Report FEW 575* (1992). (to appear in *INFORMS Journal on Computing* **8**, 1996)

[28] Blanc, J.P.C., R.D. van der Mei, The power-series algorithm applied to polling systems with a dormant server, in: *The Fundamental Role of Teletraffic in the Evolution of Telecommunications Networks*, eds. J. Labetoulle, J.W. Roberts, Elsevier, Amsterdam (1994), 865-874.

[29] Blanc, J.P.C., R.D. van der Mei, Optimization of polling systems with Bernoulli schedules, *Performance Evaluation* **22** (1995), 139-158.

[30] Blaszczyszyn, B., A. Frey, V. Schmidt, Light traffic approximations for Markov-modulated multi-server queues, *Comm. Statist. Stochastic Models* **11** (1995), 423-445.

[31] Blaszczyszyn, B., T. Rolski, Expansions for Markov-modulated systems and approximations of ruin probability, *J. Appl. Prob.* (to appear).

[32] Boucherie, R.J., A note on the transient behaviour of the Engset loss model, *Comm. Statist. Stochastic Models* **9** (1993), 145-156.

[33] Boucherie, R.J., P.G. Taylor, Transient product form distributions in queueing networks, *Research Memorandum 1990-41*, Free University, Department of Econometrics, Amsterdam, The Netherlands, 1990.

[34] Brezinski, C., *History of Continued Fractions and Padé Approximants*, Springer-Verlag, Berlin, 1991.

[35] Brezinski, C., M. Redivo Zaglia, *Extrapolation Methods, Theory and Practice*, North-Holland, Amsterdam, 1991.

[36] Carrasco, J.A., A. Calderón, Regenerative randomization: theory and application examples, in *Proceedings of ACM Sigmetrics '95 / Performance '95*, Ottawa, 1995, 241-252.

[37] Choudhury, G.L., D.M. Lucantoni, W. Whitt, Multidimensional transform inversion with applications to the transient M/G/1 queue, *Ann. Appl. Prob.* **4** (1994), 719-740.

[38] Chung, K.L., *Markov Chains with Stationary Transition Probabilities*, Springer-Verlag, Berlin (1967).

[39] Ciardo, G., J. Muppala, K.S. Trivedi, On the solution of GSPN reward models, *Performance Evaluation* **12** (1991), 237-253.

[40] Ciardo, G., A. Blakemore, P.F. Chimento, J.K. Muppala, K.S. Trivedi, Automated generation and analysis of Markov reward models using stochastic reward nets, in *Linear Algebra, Markov Chains, and Queueing Models*, eds. C.D. Meyer, R.J. Plemmons, Springer-Verlag, New York, 1993, 145-192.

[41] Cohen, J.W., *The Single Server Queue*, North-Holland, Amsterdam, 2nd edition, 1982.

[42] Cohen, J.W., O.J. Boxma, *Boundary Value Problems in Queueing System Analysis*, North-Holland, Amsterdam, 1983.

[43] Conolly, B.W., C. Langaris, On a new formula for the transient state probabilities for M/M/1 queues and computational implications, *J. Appl. Prob.* **30** (1993), 237-246.

[44] Courtois, P.J., *Decomposability, Queueing and Computer System Applications*, Academic Press, New York, 1977.

[45] Davis, J.L., W.A. Massey, W. Whitt, Sensitivity to the service-time distribution in the non-stationary Erlang loss model, *Management Sci.* **41** (1995), 1107-1116.

[46] Dai, J.G., Y. Wang, Nonexistence of Brownian models for certain multiclass queueing networks, *Queueing Systems* **13** (1993), 41-46.

[47] De Souza e Silva, E., H.R. Gail, Calculating availability and performability measures of repairable computer systems using randomization, *J. ACM* **36** (1989), 171-193.

[48] De Souza e Silva, E., H.R. Gail, R.V. Campos, Calculating transient distributions of cumulative reward, in *Proceedings of ACM Sigmetrics '95 / Performance '95*, Ottawa, 1995, 231-240.

[49] Dijk, N.M. van, Uniformization for nonhomogeneous Markov chains, *Oper. Res. Lett.* **12** (1992), 283-291.

[50] Dijk, N.M. van, *Queueing Networks and Product Forms: A Systems Approach*, John Wiley, Chichester, 1993.

[51] Dijk, N.M. van, J. van der Wal, Simple bounds and monotonicity results for finite multi-server exponential tandem queues, *Queueing Systems* **4** (1989), 1-16.

[52] Duda, A., Diffusion approximations for time-dependent queueing systems, *IEEE J. Select. Areas Comm.*, vol. SAC-4, **6** (1986), 905-918.

[53] Ettl, M., I. Mitrani, Applying spectral expansion in evaluating the performance of multiprocessor systems, in *Performance Evaluation of Parallel and Distributed Systems*, eds. O.J. Boxma, G.M. Koole, CWI Tract 105, Amsterdam, 1994, 45-58.

[54] Flatto, L., H.P. McKean, Two queues in parallel, *Comm. Pure Appl. Math.* **30** (1977), 255-263.

[55] Foster, F.G., On the stochastic matrices associated with certain queueing processes, *Ann. Math. Statist.* **24** (1953), 355-360.

[56] Frey, A., V. Schmidt, Taylor-series expansion for multivariate characteristics of risk processes, Insurance: Mathematics and Economics (to appear).

[57] Friedman, H.D., Reduction methods for tandem queuing systems, *Oper. Res.* **13** (1965), 121-131.

[58] Genz, A.C., A nonlinear method for the acceleration of the convergence of multidimensional sequences, *J. Comput. Appl. Math.* **3** (1977), 181-184.

[59] Golub, G.H., C.F. Van Loan, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, 1989.

[60] Gong, W.B., J.Q. Hu, The MacLaurin series for the GI/G/1 queue, *J. Appl. Prob.* **29** (1992), 176-184.

[61] Gong, W.B., S. Nananukul, A. Yan, Padé approximation for stochastic discrete event systems, *Proc. of the 31st Conf. on Decision and Control*, Tucson, Arizona (1992), 3209-3214.

[62] Grassmann, W.K., Transient solutions in Markovian queueing systems, *Computers Opns Res.* **4** (1977), 47-53.

[63] Grassmann, W.K., Finding transient solutions in Markovian event systems through randomization, in *Numerical Solution of Markov Chains*, ed. W.J. Stewart, Marcel Dekker, New York, 1991, 357-371.

[64] Grassmann, W.K., M.I. Taksar, D.P. Heyman, Regenerative analysis and steady state distributions for Markov chains, *Oper. Res.* **33** (1985), 1107-1116.

[65] Grassmann, W.K., Y. Wang, Immediate events in Markov chains, in *Computations with Markov Chains*, ed. W.J. Stewart, Kluwer Academic Publishers, Boston, 1995, 163-176.

[66] Green, L., P. Kolesar, The pointwise stationary approximation for queues with nonstationary arrivals, *Management Sci.* **37** (1991), 84-97.

[67] Greenberg, B.S., R.W. Wolff, Optimal order of servers for tandem queues in light traffic, *Management Science* **34** (1988), 500-508.

[68] Gross, D., D.R. Miller, The randomization technique as a modeling tool and solution procedure for transient Markov processes, *Oper. Res.* **32** (1984), 343-361.

[69] Harrison, J.M., A.J. Lemoine, A note on networks of infinite-server queues, *J. Appl. Prob.* **18** (1981), 561-567.

[70] Harrison, J.M., V. Nguyen, Brownian models of multiclass queueing networks: current status and open problems, *Queueing Systems* **13** (1993), 5-40.

[71] Henrici, P., *Applied and Computational Complex Analysis*, Wiley, New York, 1974.

[72] Heyman, D.P., W. Whitt, The asymptotic behaviour of queues with time-varying arrival rates, *J. Appl. Prob.* **21** (1984), 143-156.

[73] Hooghiemstra, G., M. Keane, S. van de Ree, Power series for stationary distributions of coupled processor models, *SIAM J. Appl. Math.* **48** (1988), 1159-1166.

[74] Hout, W.B. van den, J.P.C. Blanc, Development and justification of the power-series algorithm for *BMAP*-systems, *Comm. Statist. Stochastic Models* **11** (1995), 471-496.

[75] Hout, W.B. van den, J.P.C. Blanc, The power-series algorithm for Markovian queueing networks, in *Computations with Markov Chains*, ed. W.J. Stewart, Kluwer Academic Publishers, Boston, 1995, 321-338.

[76] Hout, W.B. van den, J.P.C. Blanc, The power-series algorithm for a wide class of Markov processes, Tilburg University, *Center Discussion Paper 9487* (1994).

[77] Houtum, G.J. van, I.J.B.F. Adan, J. Wessels, W.H.M. Zijm, The compensation approach for three or more dimensional random walks, in *Operations Research Proceedings 1992*, 342-349, Springer-Verlag, Berlin, 1993.

[78] Houtum, G.J. van, I.J.B.F. Adan, J. Wessels, W.H.M. Zijm, On the precedence relation method for deriving flexible bound models for queueing systems, Eindhoven University of Technology, Dept. of Math. and Comp. Sci., *Memorandum COSOR 94-27*, (1994).

[79] Hu, J.Q., Analyticity of single-server queues in light traffic, *Queueing Systems* **19** (1995), 63-80.

[80] Jackson, R.R.P., P. Asden, A transient solution to the multistage Poisson queueing system with infinite servers, *Oper. Res.* **28** (1980), 618-622.

[81] Jensen, A., Markov chains as an aid in the study of Markov processes, *Skand. Aktuarietidskr.* **36** (1953), 317-336.

[82] Katehakis, M.N., C. Derman, On the maintenance of systems composed of highly reliable components, *Management Sci.* **35** (1989), 551-560.

[83] Katehakis, M.N., A. Levine, A dynamic routing problem - Numerical procedures for light traffic conditions, *Appl. Math. Comp.* **17** (1985), 267-276.

[84] Katehakis, M.N., A. Levine, Allocation of distinguishable servers, *Computers Opns Res.* **13** (1986), 85-93.

[85] Kelly, F.P., *Reversibility and Stochastic Networks*, Wiley, Chichester, 1979.

[86] Kijima, M., The transient solution to a class of Markovian queues, *Computers Math. Applic.* **24** (1992), 17-24.

[87] Koole, G., On the power series algorithm, in *Performance Evaluation of Parallel and Distributed Systems*, eds. O.J. Boxma, G.M. Koole, CWI Tract 105, Amsterdam, 1994, 139-155.

[88] Koole, G., M. Vrijenhoek, Scheduling a repairman in a finite source system, *Technical Report TW-95-03*, Leiden University, 1995. To appear in *ZOR* 1996.

[89] Koole, G., O. Passchier, Optimal control in light traffic Markov decision processes, *Technical Report TW-95-12*, Leiden University, 1995.

[90] Krinik, A., Taylor series solution of the M/M/1 queueing system, *J. Comput. Appl. Math.* **44** (1992), 371-380.

[91] Kroese, D.P., V. Schmidt, Light-traffic analysis for queues with spatially distributed arrivals, *Math. Oper. Res.* (to appear).

[92] Kumar, B.K., P.R. Parthasarathy, M. Sharafali, Transient solution of an M/M/1 queue with balking, *Queueing Systems* **13** (1993), 441-448.

[93] Lebah, M., J. Pellaumail, Transient behavior for some Jackson networks, *Performance Evaluation* **17** (1993), 115-122.

[94] Lee, I-J., E. Roth, A heuristic for the transient expected queue length of Markovian queueing systems, *Oper. Res. Lett.* **14** (1993), 25-27.

[95] Leguesdron, P., J. Pellaumail, G. Rubino, B. Sericola, Transient analysis of the M/M/1 queue, *Adv. Appl. Prob.* **25** (1993), 702-713.

[96] Leung, K.K., Cyclic-service systems with probabilistically-limited service, *IEEE J. Select. Areas Comm.* **9** (1991), 185-193.

[97] Leung, K.K., Cyclic-service systems with non-preemptive, time-limited service, *IEEE Trans. Comm.* **42** (1994), 2521-2524.

[98] Levin, D., Development of non-linear transformations for improving convergence of sequences, *Internat. J. Computer Math., Section B* **3** (1973), 371-388.

[99] Levine, A., D. Finkel, Load balancing in a multi-server queueing system, *Computers Opns Res.* **17** (1990), 17-25.

[100] Lindley, D.V., The theory of queues with a single server, *Proc. Camb. Phil. Soc.* **48** (1952), 277-289.

[101] Lindvall, T., *Lectures on the Coupling Method*, John Wiley & Sons, New York, 1992.

[102] Lucantoni, D.M., New results on the single server queue with a batch Markovian arrival process, *Commun. Statist. Stochastic Models* **7** (1991), 1-46.

[103] Lucantoni, D.M., G.L. Choudhury, W. Whitt, The transient BMAP/G/1 queue, *Comm. Statist. Stochastic Models* **10** (1994), 145-182.

[104] Malhorta, M., J.K. Muppala, K.S. Trivedi, Stiffness-tolerant methods for transient analysis of stiff Markov chains, *Microelectron. Reliab.* **34** (1994), 1825-1841.

[105] Massey, W.A., Asymptotic analysis of the time-dependent M/M/1 queue, *Math. Oper. Res.* **10** (1985), 305-327.

[106] Massey, W.A., W. Whitt, Networks of infinite-server queues with nonstationary Poisson input, *Queueing Systems* **13** (1993), 183-250.

[107] Melamed, B., M. Yadin, Randomization procedures in the computation of cumulative-time distributions over discrete state Markov processes, *Oper. Res.* **32** (1984), 926-944.

[108] Melamed, B., M. Yadin, Numerical computation of sojourn-time distributions in queueing networks, *J. ACM* **31** (1984), 839-854.

[109] Meyn, S.P., R.L. Tweedie, *Markov Chains and Stochastic Stability*, Springer-Verlag, London, 1993.

[110] Mitrani, I., D. Mitra, A spectral expansion method for random walks on semi-infinite strips, IMACS Symposium on Iterative Methods in Linear Algebra, Elsevier Science Publishers B.V., North-Holland, Amsterdam (1992), 141-149.

[111] Moler, C., Van Loan, C., Nineteen dubious ways to compute the exponential of a matrix, *SIAM Review* **20** (1978), 801-836.

[112] Morris, J.L., *Computational Methods in Elementary Numerical Analysis*, John Wiley & Sons, Chichester, 1983.

[113] Neuts, M.F., *Matrix Geometric Solutions in Stochastic Models: an Algorithmic Approach*, The Johns Hopkins University Press, Baltimore, 1981.

[114] Odoni, A.R., E. Roth, An empirical investigation of the transient behavior of stationary queueing systems, *Oper. Res.* **31** (1983), 432-455.

[115] Palka, B.P., *An Introduction to Complex Function Theory*, Springer-Verlag, New York, 1991.

[116] Parthasarathy, P.R., M. Sharafali, Transient solution to the many-server Poisson queue: a simple approach, *J. Appl. Prob.* **26** (1989), 584-594.

[117] Philippe, B., R.B. Sidje, Transient solutions of Markov processes by Krylov subspaces, in *Computations with Markov Chains*, ed. W.J. Stewart, Kluwer Academic Publishers, Boston, 1995, 121-133.

[118] Reiman, M.I., B. Simon, Light traffic limits of sojourn time distributions in Markovian queueing networks, *Comm. Statist. Stochastic Models* **4** (1988), 191-233.

[119] Reiman, M.I., B. Simon, Open queueing systems in light traffic, *Math. Oper. Res.* **14** (1989), 26-59.

[120] Rindos, A., S. Woolet, I. Viniotis, K. Trivedi, Exact methods for the transient analysis of nonhomogeneous continuous time Markov chains, in *Computations with Markov Chains*, ed. W.J. Stewart, Kluwer Academic Publishers, Boston, 1995, 121-133.

[121] Roth, E., A heuristic technique for the transient behavior of Markovian queueing systems, *Oper. Res. Lett.* **3** (1985), 301-305.

[122] Rothkopf, M.H., S.S. Oren, A closure approximation for the nonstationary M/M/s queue, *Management Sci.* **25** (1979), 522-534.

[123] Saad, Y., Analysis of some Krylov subspace approximations to the matrix exponential operator, *SIAM J. Numer. Anal.* **29** (1992), 209-228.

[124] Shanks, D., Non-linear transformations of divergent and slowly convergent sequences, *J. Math. Phys.* **34** (1955), 1-42.

[125] Simon, B., Calculating light traffic limits for sojourn times in open Markovian queueing systems, *Comm. Statist. Stochastic Models* **9** (1993), 213-231.

[126] Smith, D.R., Optimal repairman allocation - Asymptotic results, *Management Sci.* **24** (1978), 665-674.

[127] Smith, D.A., W.F. Ford, Acceleration of linear and logarithmic convergence, *SIAM J. Numer. Anal.* **16** (1979), 223-240.

[128] Smith, D.A., W.F. Ford, Numerical comparison of nonlinear convergence accelerators, *Math. Comp.* **38** (1982), 481-499.

[129] Stewart, I., D. Tall, *Complex Analysis (The Hitchhiker's Guide to the Plane)*, Cambridge University Press, Cambridge, 1983.

[130] Stewart, W.J., *Computations with Markov Chains*, Kluwer Academic Publishers, Boston, 1995.

[131] Stewart, W.J., *Numerical Solution of Markov Chains*, Marcel Dekker, New York, 1991.

[132] Stewart, W.J., *Introduction to the Numerical Solution of Markov Chains*, Princeton University Press, Princeton, New Jersey, 1994.

[133] Suresh, S., W. Whitt, Arranging queues in series: a simulation experiment, *Management Sci.* **36** (1990), 1080-1091.

[134] Titchmarsh, E.C., *The Theory of the Rieman Zeta-function*, Clarendon Press, Oxford, 2nd edition, revised by D.R. Heath-Brown, 1986.

[135] Weber, R.R., The interchangeability of tandem $./M/1$ queues in series, *J. Appl. Prob.* **16** (1979), 690-695.

[136] Whitt, W., The best order for queues in series, *Management Sci.* **31** (1985), 475-487.

[137] Wynn, P., On a device for calculating the $e_m(S_n)$ transformation, *Math. Tables and Aids to Comp.* **10** (1956), 91-96.

[138] Wynn, P., On the convergence and stability of the epsilon algorithm, *SIAM J. Numer. Anal.* **3** (1966), 91-122.

[139] Zazanis, M.A., Analyticity of Poisson-driven stochastic systems, *Adv. Appl. Prob.* **24** (1992), 532-541.

[140] Zhu, Y., H. Li, The MacLaurin expansion for a G/G/1 queue with Markov-modulated arrivals and services, *Queueing Systems* **14** (1993), 125-134.

Samenvatting

Veel computer- en communicatiesystemen kunnen worden bestudeerd met behulp van Markov processen, maar ook bijvoorbeeld verkeers-, voorraad-, produktie- en gezondheidsprocessen. Deze beschrijven de toestand waarin een bepaald proces zich bevindt, de tijd die het proces in die toestand blijft en in welke toestand het proces vervolgens overgaat. Neem bijvoorbeeld in gedachten een wachtrijproces bestaand uit een kruidenier met één kassa. De toestand van het proces is de lengte van de rij voor de kassa. Deze rijlengte kan toenemen door het aansluiten van een klant of, als de rij niet leeg is, afnemen door het vertrek van een klant. Voor de beschrijving van dit proces als een Markov proces zal daarom tenminste bekend moeten zijn hoeveel tijd er gemiddeld verstrijkt tussen de aankomst van opeenvolgende klanten en ook hoe lang de bediende gemiddeld nodig heeft om een klant te helpen.

Bij de analyse van Markov processen kan onderscheid worden gemaakt tussen het lange-termijn (steady-state) gedrag en het korte-termijn (transiënte) gedrag. Deze verschillen doordat het proces begint in een specifieke toestand. Zo zal de rij bij de kassa van de kruidenier vlak na opening eerst leeg zijn. Ook op lange termijn zal het voorkomen dat de rij leeg is, maar aan het begin van de dag weet je zeker dat hij leeg is. Veel studies richten zich op het lange-termijn gedrag omdat dit eenvoudiger te analyseren is. Naast dit verschil in termijn is ook het verschil tussen homogeen en niet-homogeen van belang. Een proces is homogeen wanneer de eigenschappen van het proces gedurende de tijd niet veranderen. Wanneer in de winkel de hele dag gemiddeld evenveel klanten per uur aankomen en de bediende de hele dag even hard werkt dan is er sprake van een homogeen proces. Wanneer er rekening wordt gehouden met de drukte vlak voor sluitingstijd of de lunchpauze van de bediende dan is het proces niet homogeen. Homogene processen zijn aanzienlijk makkelijker te analyseren. Daarom wordt in de praktijk vaak verondersteld dat een proces homogeen is. Dit zal zelden werkelijk zo zijn, maar is gerechtvaardigd

wanneer de veranderingen gedurende de tijd klein zijn.

Het gemak waarmee een Markov proces kan worden geanalyseerd hangt in sterke mate af van het aantal toestanden. Een kleine kruidenier met minder dan 10 klanten in de rij is gemakkelijk te analyseren. Maar het aantal toestanden loopt al op tot een miljoen bij een middelgrote supermarkt met ook minder dan 10 klanten per rij maar met 6 rijen. De complexiteit neemt nog verder toe wanneer wordt geprobeerd om de werkelijkheid beter te benaderen door rekening te houden met piekuren, bedienden die niet allemaal even hard werken, klanten die in groepen aankomen of die van rij veranderen, kassa's die stuk gaan en ga zo maar door.

Het machtreeksalgoritme (power-series algorithm) is een methode die zich, als één van de weinige, speciaal richt op meerdimensionale Markov processen, zoals processen met meerdere wachtrijen. Omdat de geheugencapaciteit van computers beperkt is en de rekentijden van de methode sterk toenemen met het aantal rijen, kunnen over het algemeen modellen met 4 tot 6 rijen worden bestudeerd. De methode kan antwoord geven op vragen van de klanten, zoals 'Hoe groot is de kans dat ik niet hoef te wachten?' en 'Hoe lang moet ik gemiddeld wachten?'. Maar ook op vragen van de winkelier, zoals 'Hoeveel kassa's heb ik nodig' of 'Wat is beter: één lange rij of 6 afzonderlijke rijen?'. Gedurende de afgelopen tien jaar zijn meerdere artikelen geschreven met succesvolle toepassingen van het machtreeksalgoritme op verschillende typen Markov processen, in het bijzonder binnen de wachtrijtheorie. Doel van dit proefschrift is enerzijds de toepasbaarheid verder te vergroten en anderzijds de theoretische onderbouwing te verbeteren.

De indeling van het proefschrift is als volgt. In hoofdstuk 1 wordt een overzicht gegeven van het proefschrift en van de relevante literatuur. De hoofdstukken 2 en 3 beschrijven en bestuderen het machtreeksalgoritme voor respectievelijk de lange-termijn en de korte-termijn analyse van Markov processen. Deze hoofdstukken worden hieronder nader toegelicht. Tenslotte volgen 3 appendices met additionele informatie over Markov processen, analytische functies en extrapolatie methoden. Kennis van deze resultaten is essentieel voor een goed begrip van de theoretische achtergrond van het machtreeksalgoritme en onmisbaar voor een efficiënt gebruik.

Hoofdstuk 2, over de lange-termijn analyse, begint met een eenvoudig voorbeeld om de belangrijkste ideeën van de methode te illustreren. Het lange-termijn gedrag kan worden geanalyseerd door het oplossen van een stelsel vergelijkingen, de zogenaamde evenwichtsvergelijkingen. Echter, wanneer het aantal toestanden van het proces groot is dan is dit stelsel ook groot en meestal lastig op te lossen. Het idee van het machtreeksalgoritme is nu om het proces te verstoren met een transformatieparameter. Het gedrag van het proces is dan afhankelijk van deze parameter. Door de manier waarop het proces wordt verstoord reduceert het lastig op te lossen stelsel tot meerdere eenvoudig op te lossen stelsels. Zo wordt het lange-termijn gedrag van het verstoorde proces verkregen, dat vervolgens kan worden gebruikt om het oorspronkelijke onverstoorde proces te analyseren.

De verstoring kan op verschillende manieren worden aangebracht. De in dit proefschrift voorgestelde methode is algemener dan de verstoring in voorgaande publikaties. Dit heeft tot gevolg dat meer modellen met de methode kunnen worden geanalyseerd. Netwerken van wachtrijen worden over het algemeen lastiger te analyseren naarmate de verschillende wachtrijen meer van elkaar afhankelijk zijn. De ruime toepasbaarheid van het machtreeksalgoritme wordt geïllustreerd door het toe te passen op netwerken met zeer algemene afhankelijkheden tussen de aankomsten bij de rijen en tussen achtereenvolgende bedieningen bij de afzonderlijke rijen. Ook de routes die klanten door het netwerk kunnen volgen zijn algemener. Voorheen konden klanten bijvoorbeeld niet terugkeren naar rijen waar ze al eerder waren geweest. Dit soort netwerken is nu wel analyseerbaar. Voor deze ruime klasse van netwerken kunnen met het machtreeksalgoritme prestatiematen worden bepaald zoals de kans dat het hele netwerk leeg is en het gemiddelde aantal klanten in het netwerk of bij een specifieke rij. Ook kan het machtreeksalgoritme worden gebruikt om netwerken te optimaliseren, dat wil zeggen zodanig te ontwerpen dat bijvoorbeeld de gemiddelde rijlengte zo klein mogelijk is.

De theoretische onderbouwing van de methode wordt ook verbeterd. Wiskundig gesproken is de omweg via het verstoorde proces gerechtvaardigd indien het lange-termijn gedrag op een analytische manier afhangt van de transformatieparameter. Alhoewel analyticiteit een zeer fundamentele aanname is van het machtreeksalgoritme, kon deze veronderstelling voorheen alleen worden bewezen voor enkele zeer specifieke modellen. Dit wordt nu uitgebreid tot een zeer ruime klasse van modellen, die de voorheen gepubliceerde modellen omvat. Ook allerlei andere veronderstellingen van de methode worden geanalyseerd. Er wordt nagegaan wat er kan gebeuren wanneer een model niet aan de veronderstellingen voldoet en er worden aanpassingen van de methode gevonden die eventuele problemen kunnen ondervangen.

Hoofdstuk 3 richt zich op het korte-termijn gedrag van Markov processen. Er wordt een poging gedaan om ook hier de ideeën toe te passen die succesvol bleken te zijn voor de analyse van het lange-termijn gedrag. Dit leidt tot een generalisatie van wat algemeen bekend staat als de methode van Jensen. Deze methode wordt eerst beschreven en er wordt aangetoond dat deze methode vooral goed werkt wanneer hij wordt voorafgegaan door een lange-termijn analyse. Voor homogene Markov processen blijkt de methode van Jensen dan aanzienlijk doelmatiger dan het machtreeksalgoritme. Wel zijn er goede redenen om aan te nemen dat het machtreeksalgoritme gebruikt kan worden voor niet-homogene Markov processen, maar dit is nog niet getest. Ook worden interessante theoretische resultaten verkregen over analyticiteit van het korte-termijn gedrag van Markov processen.

Center for Economic Research, Tilburg University, The Netherlands
Dissertation Series

No.	Author	Title
1	P.J.J. Herings	Static and Dynamic Aspects of General Disequilibrium Theory; ISBN 90 5668 001 3
2*	Erwin van der Krabben	Urban Dynamics: A Real Estate Perspective - An institutional analysis of the production of the built environment; ISBN 90 5170 390 2
3	Arjan Lejour	Integrating or Desintegrating Welfare States? - a qualitative study to the consequences of economic integration on social insurance; ISBN 90 5668 003 x
4	Bas J.M. Werker	Statistical Methods in Financial Econometrics; ISBN 90 5668 002 1
5	Rudy Douven	Policy Coordination and Convergence in the EU; ISBN 90 5668 004 8
6	Arie J.T.M. Weeren	Coordination in Hierarchical Control; ISBN 90 5668 006 4
7	Herbert Hamers	Sequencing and Delivery Situations: a Game Theoretic Approach; ISBN 90 5668 005 6
8	Annemarie ter Veer	Strategic Decision Making in Politics; ISBN 90 5668 007 2
9	Zaifu Yang	Simplicial Fixed Point Algorithms and Applications; ISBN 90 5668 008 0
10	William Verkooijen	Neural Networks in Economic Modelling - An Empirical Study; ISBN 90 5668 010 2
11	Henny Romijn	Acquisition of Technological Capability in Small Firms in Developing Countries; ISBN 90 5668 009 9
12	W.B. van den Hout	The Power-Series Algorithm - A Numerical Approach to Markov Processes; ISBN 90 5668 011 0

* Copies can be ordered from Thesis Publishers, P.O. Box 14791, 1001 LG Amsterdam, The Netherlands, phone + 31 20 6255429; fax: +31 20 6203395; e-mail: thesis@thesis.aps.nl

Stellingen

behorend bij het proefschrift

The Power-Series Algorithm

A Numerical Approach to Markov Processes

van

Wilbert van den Hout

27 maart 1996

I

Het Machtreeks-Algoritme is een flexibele en betrouwbare methode, geschikt voor de analyse van een ruime klasse van meerdimensionale continue-tijd Markov processen waarvoor veelal geen andere methoden beschikbaar zijn. De beperkingen van de methode worden bepaald door de dimensie en de stijfheid van het model.

II

Een continue-tijd Markov proces met generator $Q = [q_{ij}]$ is ergodisch als een $\pi = [\pi_i]$ bestaat, zodanig dat

$$\pi Q = o, \ \pi e = 1 \ \text{ en } \ \sum_i |\pi_i q_{ii}| < \infty.$$

In dat geval is π de evenwichtsverdeling. Het is niet noodzakelijk om op voorhand te eisen dat $\pi \geq 0$ [1] of dat $\sup_i |q_{ii}| < \infty$ [2].

[1] S. Asmussen, *Applied Probability and Queues*, Wiley, Chichester, 1987.

[2] J.W. Cohen, *The Single Server Queue*, North-Holland, Amsterdam, 1969.

III

Jensen's methode [3] voor de transiënte analyse van ergodische continue-tijd Markov processen, ook wel bekend als uniformisatie of randomisatie, is slechts uniform convergent over de tijd wanneer gebruik wordt gemaakt van de evenwichtsverdeling.

[3] A. Jensen, Markov chains as an aid in the study of Markov processes, *Skand. Aktuarietidskr.* 36 (1953), 317-336.

IV

De normalisatieconstante van de produktvorm verdeling van een gesloten Jackson netwerk met N klanten, één bediende per station en belasting ρ_i bij station i $(1 \leq i \leq I)$ is gelijk aan de reciproque van

$$S_1(N, I, \rho) \doteq \sum_{\substack{n \in \mathbb{N}^I \\ |n|=N}} \prod_{i=1}^{I} \rho_i^{n_i}.$$

Voor grote waarden van N is deze uitdrukking lastig te berekenen. Dit kan als volgt worden ondervangen. Een aantal van de belastingen $\rho_1, \ldots, \rho_I$ kan gelijk zijn. Kies J gelijk aan het aantal verschillende belastingen, σ_j $(1 \leq j \leq J)$ gelijk aan de waarden van de verschillende belastingen en m_j $(1 \leq j \leq J)$ gelijk aan het aantal stations waarvan de belasting gelijk is aan σ_j. Dan geldt

$$S_1(N, I, \rho) = S_2(N, J, \sigma, m) = S_3(N, J, \sigma, m),$$

met

$$S_2(N,J,\sigma,m) \doteq \sum_{\substack{n \in \mathbb{N}^J \\ |n|=N}} \prod_{j=1}^{J} \binom{m_j-1+n_j}{n_j} \sigma_j^{n_j},$$

$$S_3(N,J,\sigma,m) \doteq \sum_{j=1}^{J} \sigma_j^N \sum_{\substack{n \in \mathbb{N}^J \\ |n|=m_j-1}} \binom{N+n_j}{n_j} \prod_{\substack{k=1 \\ k \neq j}}^{J} \binom{m_k-1+n_k}{n_k} \left(\frac{\sigma_j}{\sigma_j-\sigma_k}\right)^{m_k} \left(\frac{\sigma_k}{\sigma_k-\sigma_j}\right)^{n_k}.$$

Het aantal termen in de laatste uitdrukking is onafhankelijk van N, waardoor berekening ook voor grote waarden van N eenvoudig is.

V

De hoeveelheid informatie vervat in een enkel toevalsgetal

$$R = \sum_{k=1}^{\infty} 10^{-k} U_k,$$

uniform verdeeld op het interval [0,1] (met U_k uniform verdeeld op $\{0,1,\ldots,9\}$ voor alle $k = 0,1,\ldots$) is gelijk aan de hoeveelheid informatie vervat in een rij toevalsgetallen

$$R_n = \sum_{k=1}^{\infty} 10^{-k} U_{2^n(2k-1)}, \quad \text{met } n = 0,1,\ldots,$$

elk uniform verdeeld op het interval [0,1]. Dit impliceert dat er a priori geen essentieel verschil is tussen enerzijds een deterministische wereld met toevallige begintoestand en anderzijds een wereld met voortdurende toevallige inbreng.

VI

Net als Gods almacht niet beperkt wordt door Zijn deterministische natuurwetten wordt Zijn voorzienigheid niet beperkt door Zijn kanswetten. Beide zijn onderdeel van Zijn voortdurende creatie.

VII

In onze actiegerichte samenleving wordt wachten vooral als hinderlijk ervaren vanwege de onmacht om het wachtproces te beïnvloeden.

VIII

Antwoord op een sollicitatiebrief is een Jobstijding.

IX

Aangezien de doeltreffendheid van medische behandelingen vaak aanzienlijk groter is bij patiënten die zijn opgenomen in een wetenschappelijk onderzoek en betrouwbare kennis van veel medische aandoeningen en behandelingen ontbreekt, is het wenselijk om de voltallige Nederlandse bevolking te betrekken in het eerstvolgende medische onderzoek.

WILBERT VAN DEN

sciences at the University of Amsterdam in 1991. Until October 1995 he was a Ph.D. student at Tilburg University, for the Netherlands Organization for Scientific Research (NWO). His research concentrated on the Power-Series Algorithm for the analysis of queueing systems. Currently he is working at the Medical Decision Making Unit of Leiden University.

The development of computer and communication networks and flexible manufacturing systems has led to new and interesting multidimensional queueing models. The Power-Series Algorithm is a numerical method to analyze and optimize the performance of such models. In this thesis, the applicability of the algorithm is extended. This is illustrated by introducing and analyzing a wide class of queueing networks with very general dependencies between the different queues. The theoretical basis of the algorithm is strengthened by proving analyticity of the steady-state distribution in light traffic and finding remedies for previous imperfections of the method. Applying similar ideas to the transient distribution renders new analyticity results. Various aspects of Markov processes, analytic functions and extrapolation methods are reviewed, necessary for a thorough understanding and efficient implementation of the Power-Series Algorithm.

ISBN 90-5668-0-011-0

CentER Dissertation