SERIE RESEARCH MEMORANDA

THE IMPORTANCE OF BIAS-TERMS FOR ERROR BOUNDS AND COMPARISON RESULTS

Nico M. van Dijk

Research memorandum 1989-36 July '89

VRIJE UNIVERSITEIT
FACULTEIT DER ECONOMISCHE WETENSCHAPPEN
EN ECONOMETRIE
AMSTERDAM

THE IMPORTANCE OF BIAS-TERMS FOR ERROR BOUNDS AND COMPARISON RESULTS

Nico M. van Dijk
Free University, The Netherlands

Abstract

Bias terms of Markov reward structures are shown to play a key-role to conclude error bound or monotonicity results for steady state measures when dealing with approximate systems such as due to

. perturbations
. finite truncations
. system modifications, or
. system comparisons (bounds)

The results are illustrated by a non-product form queueing network example with practical phenomena as blocking, overflow and breakdowns. Monotonicity results and explicit error bounds are hereby established for different approximations.

Contents

1 Introduction

Markov chains are known to be a powerful modeling tool for a variety of practical dynamic situations. Most notably, applications are found in areas of telecommunication (queueing networks, broadcasting, satellite communication), computer performance evaluation (computer networks, parallel programming, store and forward buffering), manufacturing (assembly lines, material handling systems), reliability (maintenance, breakdown analysis), inventory theory and combinations of these as performability analysis of computer systems with breakdowns, error-detections or fault-tolerancy. The modeling of realistic situations, however, may itself already introduce inaccuracies such as by simplifying assumptions and the use of exact parameters.

Traditionally, steady state behaviour is usually the prime interest, such as to estimate a throughput, system efficiency, mean workload, response time or system availability, but also transient analysis becomes more and more important. As closed form exact expressions are available only in a very limited number of practical situations (such as a Jackson network without blocking, a finite reversible material handling system or a simple parallel computer system with Poissonian input), numerical or approixmate computations are frequently employed. As exact numerical computations easily become astronomie while approximate results usually encounter unspecified imprecisions, error bounds, preferably a priori and analytic, for the accuracy or even just its order, are of significant interest.

Generally, several categories of "approximate" results can thus be involved. Let us discuss some of these more detailed.

Perturbations Applications of Markov (reward) chains are usually studied under the assumption that the one-step transition probabilities (and rewards) are known exactly. However, in practice these characteristics are often determined by a few system parmeters, such as the arrival and service rates in a queueing system, which are to be estimated by statistical data. Inaccuracies such as from the confidence interval bounds for the estimated parameters are then naturally involved.

By studying the effect of perturbations in the input data (parameters) of a Markov (reward) process qualitative or sensitivity results with respect to essential system parameters are obtained. In [8] and [13] perturbation results are given for stationary probabilites in the finite case. In [5] and [42] error bounds for approximate undiscounted finite- and discounted infinite-horizon Markov reward structures were derived. The order of these bounds however did not allow a limiting result for the average reward case. To this end, perturbation results from infinite and average reward situations have been developed in [38] and extended to unbounded reward structures in [30] with applications to primarily dimensional queueing systems.

State space truncation In practice, one often encounters large or infinite state spaces such as in an infinite server queue, an open queueing network or a maintenance system with an a priori unbounded lifetime. Truncation of the state space then becomes necessary for computational purposes.

Though the technique of state space truncation is a common feature in practice, theoretical support in terms of orders of accuracy or rates of convergence seems hardly available. Convergence proofs as the truncation size tends to infinity have already been investigated in the early fifties by Savymsakov and were cristallized most notably by Seneta (cf. [15], [16]) with reference to private communications with Kendall. A detailed study of these convergence results as well as an extensive list of related literature can be found in Seneta (1980). In this latter reference, also simple error bounds are provided (cf. theorem 6.4 and its corollary, p.215), but these are just robust bounds which do not secure an order of accuracy.
As a special application, methods based on bounded transition rate functions such as the uniformization method are not generally applicable to infinite systems (e.g. an infinite server queue). By providing state space truncation error bounds such methods can be made applicable in an approximative manner (cf. [25]). Conversely, one might wish to know the accuracy of an infinite modeling of finite systems such as for theoretical purposes, to find bounds (cf. [32], [33], [36]) or to conclude a convergence rate rather than just limiting arguments (cf. [4]) when using approximating finite sequences.

Modifications Explicit steady state expressions are usually obtainable only under strong assumptions that are often unrealistic. For example, a Poissonian unlimited input in a service system is frequently assumed merely for convenient modeling while more realistically inputs are much more likely to be of a large but finite and state dependent nature. By modelling or rather modifying such an input as a Poisson input simple approximate results, such as for the throughput, might be concluded.

As another modification of interest, product form expressions for practical queueing networks typically fail as phenomena such as blocking, dynamic routing, breakdowns and priorities usually destroy necessary partial balance conditions to this end. By slightly modifiying the system protocols or rather the underlying transition structure, however, such conditions can be met from which product form 'approximations' can then be obtained. Simple performance esimates based on product form modifications have so been established for various non-product queueing systems (cf. [28], [31], [37], [40]). Error bounds for these estimates however have not been provided.

Comparison results As a special modification, a system can be compared under two different situations. For example to (i) investigate the effect of enlarging certain system parameters such as a storage or service capacity (cf. [1], [2], [17], [18], [19], [20], [21], [22], [23], [24], [25]), (ii) determine a better protocol such as for dynamic job or server allocation (cf. [3], [17], [38], [41]) or (iii) conclude that transition modifications as product form estimates lead to performance bounds (cf. [21], [28], [31], [37], [39], [40], [46]).

All of such comparison results actually come down to proving monotonicity results. Established comparison or monotonicity prooftechniques as the one-step comparison technique employed in [6], [23], [44], [45] or the related sample path technique in [18], [19], [20], [22], [25], [39], however, do not generally apply (cf. [40]) to this end. The Markov reward prooftechnique that will be illustrated in this paper has already proven to be a fruitful generalization (cf. [28], [29, [32], [33], [36], [37], [40]).

All of the above "approximate results" thus come down to some kind of modification/perturbation of the transition structure and/or a truncation/extension of the state space. This paper will provide a general tool to conclude error bounds for these approximations. More precisely, it will provide a unifying tool from which error bound or monotonicity results can be concluded when dealing with

. perturbations
. finite truncations
. system modifications, or
. system comparisons (bounds).

The key-step to these results turns out to be one and the same: "An estimation of the so-called bias terms for the specific required Markov reward structure in order".

Bias-terms (or fundamental matrices) are common knowledge in Markov decision theory as a key-factor to determine average optimal policies (cf. [11], [26]). They are also known to be directly related to mean first passage times which play a key-role in the conditioning and convergence of numerical procedures to solve steady state equations (cf. [7], [9]). Explicit expressions of passage times, however, can only be obtained in very simple situations such as a simple random walk (cf. [7]).

The crucial point of this paper, in contrast, is that in many concrete situations one can derive explicit bounds for bias-terms by employing an inductive Markov reward prooftechnique. The steps can become technical and complicated when a large number of different types of transitions in a particular state are possible. For rather natural and simple multi-dimensional transition structures though, such as most typically queueing networks, it has already proven to be succesful in a number of situations (cf. [32], [36]). Given the importance of error bounds, this paper primarily aims to advocate this "novel" use of bias-terms.

To this end, several results from [33], [34], and [38] will be combined and extended in a unifying manner. A pilot-example will be analyzed. This example concerns a simple but multi-dimensional queueing network which includes typical practical features as finite capacities, overflow and breakdowns. Monotonicity results and error bounds will be established for a perturbation, state space truncation and system modification. The verification of the necessary conditions and the importance of bias-terms are hereby illustrated.

2 General results

This section will provide the key-theorem from which error bound and monotonicity results of steady state measures of Markov chains can be concluded. First this theorem is presented in section 2.1. Its conditions will be discussed in section 2.2. Next, the transformation to continuous-time Markov chains is given in section 2.3 while in section 2.4 the results are particularized for application to queueing networks.

2.1 Key-theorem

Consider a Markov chain $\{X_t,\ t=0,1,2,\dots\}$ with states ordered at $N=\{1,2,\dots\}$ and one-step transition probability matrix $P=(p(i,j))$. Without loss of generality assume that this Markov chain is irreducible at some set S.

Let $\{\bar{X}_t,\ t=0,1,2,\dots\}$ be another Markov chain with states also ordered at $N=\{1,2,\dots\}$ and one-step transition probability matrix $\bar{P}=(\bar{p}(i,j))$. Again, without loss of generality assume that this Markov chain is irreducible at some set $\bar{S}$. We essentially impose the condition:

$$\bar{S} \subset S \tag{2.1}$$

From now on, we always use an upper bar "-" symbol for an expression concerning the second chain and the symbol "(-)" to indicate that an expression is to be read for both chains. Let operators $\overset{(-)}{T}$ and $\{\overset{(-)}{T}_t \mid t=0,1,2,\dots\}$ on arbitrary functions $f: \overset{(-)}{S} \to R_1$ be defined by

$$\begin{aligned} \overset{(-)}{T} f(i) &= \Sigma_j \overset{(-)}{p}_{ij}\ f(j) \\ \overset{(-)}{T}_{t+1} &= \overset{(-)}{T}\ \overset{(-)}{T}_t, \\ \overset{(-)}{T}_0 &= I \end{aligned} \tag{2.2}$$

and for given function r define functions $\overset{(-)}{V}_N$, $N=0,1,2,\dots$ by:

$$\overset{(-)}{V}_N = \Sigma_{t=0}^{N-1} \overset{(-)}{T}_t r \tag{2.3}$$

In words that is, $\overset{(-)}{V}_N(i)$ represents the total expected reward over N periods when starting in state i at time t=0 and receiving a one step reward $r(X_t)$ at time $t=0,1,\dots,N-1$. Then for any given state $\ell \in \overset{(-)}{S}$:

$$\overset{(-)}{g} = \lim_{N\to\infty} \frac{1}{N} \overset{(-)}{V}_N\ (\ell) \tag{2.4}$$

is the expected average reward where this limit is assumed to exist. The following key-theorem can now be proven. It provides a pair of conditions from which an error bound for the performances of both chains can be concluded.

Theorem 2.1 (Error bound) *Suppose that for some nonnegative function Φ, some initial state $\ell \in \bar{S}$, some constants ε, δ, $\beta > 0$, all $i \in \bar{S}$ and $t \geq 0$:*

$$\left|\Sigma_j [\bar{p}(i,j) - p(i,j)]\ [V_t(j) - V_t(i)]\right| \leq \varepsilon\ \Phi(i) \tag{2.5}$$

$$|(\bar{r}-r)(i)| \leq \delta\ \Phi(i) \tag{2.6}$$

$$\bar{T}_t \Phi(\ell) \leq \beta \tag{2.7}$$

Then

$$|\bar{g}-g| \leq [\varepsilon+\delta]\beta \tag{2.8}$$

Proof As for all t:

$$\overset{(-)}{V}_{t+1} = \overset{(-)}{r} + \overset{(-)}{T} V_t \tag{2.9}$$

by virtue of (2.3), while the transition probabilities $\bar{p}(.,.)$ remain restricted to $\bar{S}\subset S$, for arbitrary $\ell\in\bar{S}$ we can write:

$$\begin{aligned}(\bar{V}_N - V_N)(\ell) &= (\bar{r}-r)(\ell) + (\bar{T}\bar{V}_{N-1} - TV_{N-1})(\ell)\\ &= (\bar{r}-r)(\ell) + (\bar{T}-T)V_{N-1}(\ell) + \bar{T}(\bar{V}_{N-1}-V_{N-1})(\ell)\\ &= \Sigma_{t=0}^{N-1}\ \bar{T}([\bar{r}-r]+[(\bar{T}-T)V_{N-t-1}])(\ell) + \bar{T}(\bar{V}_0-V_0)(\ell),\end{aligned} \tag{2.10}$$

where the latter equality follows by iteration. Now note that the last term in the last right hand side is equal to 0 as $\bar{V}_0(.)=V_0(.)=0$ by definition. Further, for any s and i:

$$(\bar{T}-T)V_s(i)=\Sigma_j[\bar{p}(i,j)-p(i,j)]V_s(j)=\Sigma_j[\bar{p}(i,j)-p(j,i)][V_s(j)-V_s(i)] \tag{2.11}$$

By substituting (2.11) in (2.10), taking absolute values and noting that $\bar{T}_t$ is a monotone operator for all $t \ge 0$, we obtain from (2.5), (2.6) and (2.10):

$$|(\bar{V}_N-V_N)(\ell)| \le [\delta+\varepsilon]\Sigma_{t=0}^{N-1}\bar{T}_t\Phi(\ell) \le[\delta+\varepsilon]\beta N \tag{2.12}$$

Applying (2.4) completes the proof. □

Remark 2.1 Clearly, the conditions (2.5), (2.6) and (2.7) could have been combined in one bounding condition that can be applied directly to (2.10). The present slightly move restrictive conditions are preferred as they appear more practical. In the next theorem, however, which concentrates only on monotonicity, the combination of (2.6) and (2.7) has proven to be most essential in applications (e.g. [28], [37], [40]).

Theorem 2.2 (Monotonicity result) *Suppose that for all* $i\in\bar{S}$ *and* $t \ge 0$:

$$[\bar{r}-r](i) + \Sigma_j[\bar{p}(i,j)-p(i,j)][V_t(j)-V_t(i)] \ge 0 \ (\le 0) \tag{2.13}$$

Then

$$\bar{g} \ge g \ (\bar{g} \le g) \tag{2.14}$$

Proof This follows directly from substituting (2.11) in (2.10) and noting that the operators $\bar{T}_t$ are monotone (i.e. $\bar{T}_t f \ge \bar{T}_t g$ whenever $f \ge g$ componentwise). □

Remark 2.2 (Importance of bias-terms) The crucial step for the above theorems is the simple relation (2.11). This step enables one to transform conditions upon $V_t(.)$ in so-called bias-terms: $V_t(j)-V_t(i)$. While $V_t(.)$ generally grows linearly in t, bias terms for given i and j are generally bounded uniformly in t. More precisely, when r(.) is bounded, say $|r(i)| \leq B$ for all i, by simple Markov reward arguments (cf. [38]) one proves:

$$|V_t(j)-V_t(i)| \leq 2B \min [R_{ij}, R_{ij}] \qquad (2.15)$$

where R_{ij} is the expected number of steps (mean first passage-time, e.g. [7]) to reach state j out of state i. A similar though more technical result in terms of such times can be given also for unbounded rewards (cf. [30]). Most essentially, however, closed form expressions or even simple bounds for such times seem to be limited to simple one-dimensional random walks (cf. [7]). In the next section, we will illustrate how estimates for these bias-terms can be derived in a different analytic manner and most notably also for multi-dimensional applications such as queueing networks (also see step 1 in section 2.2).

Remark 2.3 ((Un)bounded rewards) Note that no conditions are imposed upon the one-step reward function $\bar{r}(.)$ other than that we implicitly assume the average rewards g and $\bar{g}$ to be well-defined. Particularly, unbounded rewards are allowed. For instance, by using a linear one-step reward function r(i)=i we can compute a mean queue length of an infinite system. As a special application of the bounded case, g represents the steady state probability of a set G if we choose:

$$r(i) = \begin{cases} 1 & \text{for } i \in G \\ 0 & \text{otherwise} \end{cases} .$$

Remark 2.4 (State labeling) For expository convenience the states were labeled in a countable manner. Clearly, for multi-dimensional applications such as queueing networks we can always label the states in an appropriate one-dimensional manner. But more conveniently, the results of this section can directly be reread with multi-dimensional states by simply identifying a state with symbols i and j. In section 2.5 this will be applied for queueing networks.

Remark (2.5) (Transient results) Both theorem 2.3 and theorem 2.4 can be extended to transient analysis of reward structures up to an exit time (also see [38]).

2.2 Discussion of conditions

In order to give more practical insight in the conditions and their verifications, this section briefly discusses the steps involved. Hereby, we concentrate on theorem 2.1 only.

First of all, one must typically think of either β or $[\delta+\varepsilon]$ to be small. For convience, let $\Delta(i,j) = \bar{p}(i,j)-p(i,j)$.

Step 1 (Bounded bias-terms). As first and most essential step, one has to find estimates (bounds) $B_{i,j}$ for specific i and j such that for all s:

$$|V_s(j)-V_s(i)| \leq B_{i,j} \tag{2.16}$$

Fortunately, for condition (2.5) this is needed only for all i,j with

$$|\Delta(i,j)| > 0 \tag{2.17}$$

For example, with i the number of jobs in a discrete time birth-death queueing system we only need to consider j=i-1 and j=i+1.

To find $B_{i,j}$, an inductive Markov reward technique based on (2.9) can be applied where T is to be written out in the transition probabilities as per (2.2). The structure of these probabilities and an induction hypothesis upon V_t and the reward r(.) together must prove (2.16) also for s=t+1. This key-step is conceptually straightforward but can be rather complicated in concrete multi-dimensional applications.

Step 2 (Transition differences). Secondly, one has to find out whether the differences in the transition structure $\Delta(.,.)$ are small or just bounded up to a state dependent scaling function $\Phi(.)$. For illustration, think of $\Phi(.)=1$ and consider the following examples.

> **Example 1** Consider a discrete-time birth-death model representing a discrete-time single-server queue with arrival probability α and service probability γ per step (e.g. with probability $\alpha(1-\gamma)$ a job arrives but no job leaves while with probability $\gamma(1-\alpha)$ a job leaves but no job arrives) and consider the same model with arrival probability $\alpha \pm \tau$ (perturbation) as resulting from a statistical confidence interval for estimating α, where τ is to be thought of as being small. Then
>
> $$|\Delta(.,.)| \leq \tau$$
>
> **Example 2** Consider the model as in the example above but now with rejection of arrivals (state space truncation) if upon arrival the number n of jobs present is equal to some limit L. Then with 1(A) an indicator function of event A,
>
> $$|\Delta(.,.)| \leq \alpha \; 1(n=L)$$

Step 3 (Bounding Function Φ). By comparing the transition structures, candidates for an appropriate bounding function $\Phi(.)$ come up naturally. Here one may typically think of polynomial type functions, for example, $\Phi(n)=n$ with n the total number of jobs present in a queueing system. One may thus have various options. As illustration, in example 2 above, condition (2.5) will be satisfied with some constant β resulting from (2.16) and

$$\begin{cases} \varepsilon = \alpha B/L & \text{if } \Phi(.) = n \\ \varepsilon = \alpha B & \text{if } \Phi(.) = 1(n=L) \end{cases}$$

Step 4 (Stability). Which option of $\Phi(.)$ is appropriate will eventually depend on whether we can easily verify (2.7), requiring that its expected value over time remains bounded (stability) by either a small or just a finite number. As illustration, again for example 2 from above, analogously to the standard continuous-time single-server queue one easily shows with $\lambda=\alpha(1-\gamma)$ and $\mu=\gamma(1-\alpha)$

$$\begin{cases} \beta = \beta_1 = 1 & \text{if } \Phi(.) = 1 \\ \beta = \beta_2 = (\lambda/\mu)^L \;/\; \Sigma_{k=0}^{L} (\lambda/\mu)^k & \text{if } \Phi(.) = 1(n=L) \\ \beta = \beta_3 = \Sigma_{k=0}^{L} k(\lambda/\mu)^k \;/\; \Sigma_{k=0}^{L} (\lambda/\mu)^k & \text{if } \Phi(.) = n \end{cases} \qquad \square$$

Summary Roughly speaking, theorem 2.1 can be applicable in a twofold manner given that the bias-terms can be estimated as per (2.16):

(i) By showing that the impact of the difference Δ in the transition structures upon the state-dependent estimates for the bias terms is sufficiently small, such as for example 1 with $\varepsilon=\gamma\beta$ and $\beta=1$ by using $\Phi(.)=1$, or example 2 with $\varepsilon=\tau B/L$ and $\beta=\beta_3$ by using $\Phi(.)=n$.

(ii) By showing that the expected value of the scaling function or the probability of being in states where this difference is significant, is sufficiently small, such as for example 2 with $\varepsilon=\gamma B$ and $\beta=\beta_2$ by using $\Phi(.)=1(n=L)$.

2.3 A special case: Truncation

To illustrate how truncations are covered by the above theorems, consider the special case that for some $L>\infty$:

$$\begin{cases} \bar{p}(i,j) = 0, & j>L,\ i\le L, \\ \bar{p}(i,j) = p(i,j), & j\ne t[i],\ j\le L,\ i\le L \\ \bar{p}(i,t[i]) = p(i,t[i]) + \Sigma_{j>L}\, p(i,j) & i\le L \end{cases} \qquad (2.18)$$

where $t[i]\le L$ is some given "state of truncation" for any $i\le L$. In words that is, all transitions of the original matrix $p(i,j)$ out of state i beyond a certain threshold L are reflected to one and the same state $t[i]$. Condition (2.5) now simply reduces to:

$$\left|\Sigma_{j>L}\, p(i,j)[V_t(j)-V_t(t[i])]\right| \le \varepsilon\, \Phi(i) \qquad (2.19)$$

The fact that different absorption states $t[i]$ for different states i are to be chosen will naturally come in when multi-dimensional applications are transformed in the one-dimensional description given.

Remark 2.6 (Other truncations) The truncation (2.18) is a natural one as it corresponds to the original model as long as the truncation limit L is not exceeded. Clearly, similar conditions can be provided for other types of truncations. For example, rather than letting a transition i→j for all $j>L$ transform into one and the same state $t[i]$, we can also let it transform into different states in a randomized manner.

2.4 Continuous-time Markov chains

Various Markov chain applications, such as most notably in queueing, are of a continuous- rather than discrete-time nature. In order to apply the above results the standard uniformization technique (cf. [12], p. 261, [26], p. 110) is then to be applied provided the transition rates are uniformly bounded. More precisely, let $\{\overset{(-)}{X}_t \mid t=0,1,2,\ldots\}$ be a continuous-time irreducible Markov chain at some set $\overset{(-)}{S}$ with transition rates $\overset{(-)}{q}(i,j)$ such that for some $Q<\infty$ and all $i\in\overset{(-)}{S}$:

$$\Sigma_{j\neq i}\overset{(-)}{q}(i,j) \leq Q \tag{2.20}$$

Let $\overset{(-)}{R}(i)$ be a reward rate, that is the reward per unit of time when the chain is in state i and $\overset{(-)}{G}$ the corresponding average expected reward (per unit of time). The results of section 2.1 then apply by substituting

$$\begin{cases} g = \overset{(-)}{G}/Q \\ \overset{(-)}{r} = \overset{(-)}{R}/Q \\ \overset{(-)}{p}(i,j) = \overset{(-)}{q}(i,j)/Q \quad (j\neq i) \\ \overset{(-)}{p}(i,j) = \overset{(-)}{q}(i,j)/Q + [1-\Sigma_{j\neq i}\overset{(-)}{q}(i,j)/Q] \end{cases} \tag{2.21}$$

The conditions (2.5) and (2.6) are then natural to be satisfied with $\delta=\alpha/Q$ and $\epsilon=\gamma/Q$, where α and γ are to be thought of as small, so that from (2.8) and the above we conclude:

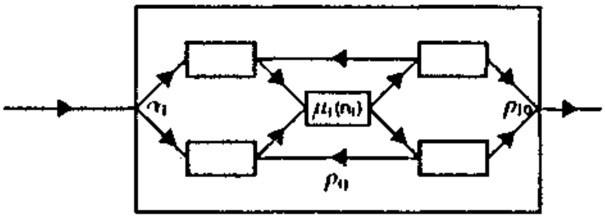
$$|\bar{G}-G| \leq [\alpha+\gamma]\beta \tag{2.22}$$

Remark 2.7 For the unbounded case, i.e. without (2.20), an approximate uniformization can be applied as in [35]. The details are rather technical.

2.5 Special application: Queueing Networks

As queueing networks are an application area of practical interest while both a multi-dimensional (remark 2.4) and special continuous-time (section 2.4) structure of a random walk type nature are to be taken into account, below we particularize our results to queueing network applications.

Model Consider an arbitrary open or closed single class exponential queueing network with N service stations (hereafter called the original model), for example as illustrated below.



The state of the network is described by $\bar{n}=(n_1,\ldots,n_N)$ where n_i is the number of jobs at station i, $i=1,\ldots,N$. By $\bar{n}+e_i$ or $\bar{n}-e_i$ we denote the state equal to $\bar{n}$ except for one job more respectively less at station i, where $\bar{n}-e_i=\bar{n}$ for $n_i=0$, $i=1,\ldots,N$ and where we also allow $i=0$ with the convention that $\bar{n}\pm e_0=\bar{n}$. Consequently, by $\bar{n}-e_i+e_j$ we denote the state equal to $\bar{n}$ with one job moved from station i into station j, where $i=0$ corresponds to an external arrival at station j and $j=0$ to a departure from the system at station i.

Let $q(\bar{n}, \bar{n}-e_i+e_j)$ for $i,j=0,1,\ldots,N$ be the transition rate for a change from state $\bar{n}$ into state $\bar{n}-e_i+\bar{e}_j$, while transition rates for changes not of this form are assumed to be 0. For example, for a standard Jackson network we have

$$q(\bar{n}, \bar{n}-e_i+e_j) = \mu_i \ p_{ij}$$

with μ_i the service state at station i and p_{ij} the routing probability from station i to j, while an additional capacity constraint N_j at station j yielding a reflective blocking (communication protocol) is parametrized by

$$q(\bar{n}, \bar{n}-e_i+e_j) = \mu_i \ p_{ij} \ 1(n_j<N_j)$$

Assumptions

1. The underlying Markov jump process is irreducible at some set S of admissible states $\bar{n}$ with a unique stationary distribution $\pi(.)$.
2. We can choose a finite Q such that

$$Q \geq \sup_{\bar{n}\in S} \Sigma_{i,j} q(\bar{n}, \bar{n}-e_i+e_j)$$

3. For given reward rate $R(\bar{n})$ the measure G is well-defined by

$$G = \Sigma_{\bar{n}} \pi(\bar{n}) R(\bar{n})$$

Approximative model Now consider a modified version of the single class exponential queueing network (hereafter called the modified model) with a description as above, but with $q(\bar{n},\bar{n}-e_i+e_j)$ replaced by $\bar{q}(\bar{n},\bar{n}-e_i+e_j)$, the assumptions 1, 2 and 3 adopted with S, π, r and g replaced by $\bar{S}$, π, $\bar{r}$ and $\bar{g}$, and $\bar{S} \subset S$.

Comparison result In order to compare the models now define the functions $\overset{(-)}{V}_t$ and operators $\overset{(-)}{T}_t$ for functions f: $\overset{(-)}{S} \to R_1$ by

$$\overset{(-)}{V}_{t+1}(\bar{n}) = \overset{(-)}{R}(\bar{n})/Q + \overset{(-)}{T} V_t(\bar{n})$$

and

$$\overset{(-)}{T}_0 f(\bar{n}) = f(\bar{n}), \overset{(-)}{T}_{t+1} f(\bar{n}) = \overset{(-)}{T}(\overset{(-)}{T}_t f)(\bar{n})$$

where

$$\overset{(-)}{T} f(\bar{n}) = \Sigma^N_{i\neq j=0} [\overset{(-)}{q}(\bar{n},\bar{n}-e_i+e_j)/Q] \ f(\bar{n}-e_i+e_j) +$$

$$[\overset{(-)}{q}(\bar{n},\bar{n})/Q + 1-\Sigma^N_{\neq j=0} \ \overset{(-)}{q}(\bar{n},\bar{n}-e_i+e_j)/Q] \ f(\bar{n}) \qquad (2.23)$$

The following application of theorem 2.1 then provides an error bound for the difference $|\bar{G}-G|$ without having to compute the stationary distributions π. Herein, let

$$\Delta(\bar{n},\bar{n}-e_i+e_j) = \bar{q}(\bar{n},\bar{n}-e_i+e_j) - q(\bar{n},\bar{n}-e_i+e_j) \quad (\bar{n}\in\bar{S}) \tag{2.24}$$

Theorem 2.3 (Error bound) *Suppose that for some nonnegative function* Φ, *some initial state* $\bar{\ell}\in\bar{S}$, *some constants* α, γ, β, > 0, *all* $\bar{n}\in\bar{S}$ *and* $t\geq 0$:

$$\left|\Sigma^{N}_{i,j=0}\Delta(\bar{n},\bar{n}-e_i+e_j)[V_t(\bar{n}-e_i+e_j)-V_t(\bar{n})]\right| \leq \alpha\,\Phi(\bar{n}) \tag{2.25}$$

$$|\bar{R}(\bar{n})-R(\bar{n})| \leq \gamma\,\Phi(\bar{n}) \tag{2.26}$$

$$\bar{T}_t\Phi(\bar{\ell}) \leq \beta \tag{2.27}$$

Then

$$|\bar{G}-G| \leq \beta[\alpha+\gamma] \tag{2.28}$$

Proof Directly by applying theorem 2.1 with the substitution (2.21) and $\delta=\alpha/Q$, $\varepsilon=\gamma/Q$ and noting that the difference $\Delta(\bar{n},\bar{n})$ in the transition rates for a change from a state $\bar{n}$ in itself vanishes in (2.5). □

In a similar manner, we conclude from theorem 2.2:

Theorem 2.4 (Monotonicity) *Suppose that for all* $\bar{n}\in\bar{S}$ *and* $t\geq 0$:

$$[\bar{R}-R](\bar{n})+\Sigma^{N}_{i,j=0}\Delta(\bar{n},\bar{n}-e_i+e_j)[V_t(\bar{n}-e_i+e_j)-V_t(\bar{n})] \geq 0 \;\; (\leq 0) \tag{2.29}$$

Then

$$\bar{G} \geq G \quad (\bar{G} \leq G) \tag{2.30}$$

Remark 2.8 The application of theorems 2.1 and 2.2 seems most appropriate for queueing networks as per theorems 2.3 and 2.4 by virtue of the underlying relatively simple random walk type structure of this system. The inductive prooftechnique (see step 1 in section 2.2) for estimating the bias-terms will hereby become more tractable.

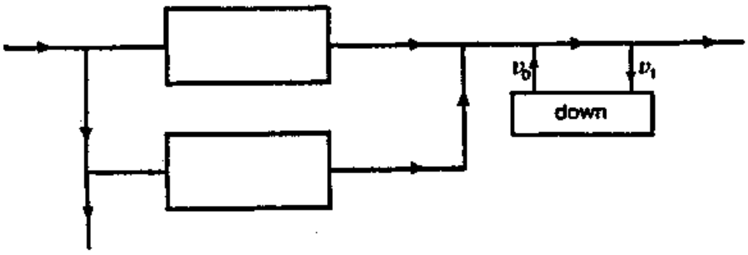
3 A pilot-example: A breakdown overflow model

In this section we will illustrate some possible applications of the results and most of all the way of estimating and the importance of the bias-terms.

To this end, a special queueing network example will be analyzed. Though this example is relatively simple, it has no closed product form expression and serves as a pilot-example for practical aspects as:
- a multi-dimensional structure
- finite constraints and blocking
- dynamic (overflow) routing
- breakdowns

3.1 Model

Description Consider a queueing network of two parallel service stations with capacities for containing at most N_1 and N_2 jobs respectively, where N_1 or N_2 can be infinite. Jobs arrive at the system at a Poisson arrival rate λ. An arriving job first attempts to enter station 1. If station 1 is saturated ($n_1=N_1$), it routes to station 2. If also this station is saturated ($n_1=N_1$ and $n_2=N_2$) it is lost. Further, jobs at station 2 can never switch (back) to station 1. The service rate at station i is $\mu_i(n_i)$ when n_i jobs are present, where we assume $\mu_i(.)$ to be non-decreasing as well as to be bounded, $i=1,2$.



Furthermore, as a special complication the output channel is subject to breakdowns which renders it inoperative for a period of time regardless of whether jobs are present or not. More precisely, this channel is alternatively "up" and "down" for exponential periods with parameters υ_1 and υ_0 respectively. When this channel is down, jobs completing a service are prohibited to leave the system and have to remain at their station. (This latter assumption corresponds to the so-called communication protocol and can be interpreted as if blocked jobs (messages) are to be reserved (retransmitted) or as if the servicing at the stations is stopped as long as the channel is down).

Background The system under consideration has no product-form solution for the steady-state distribution as necessary partial balance conditions to this end are easily shown to fail. Only some marginal explicit results are available.

(i) For the pure overflow case, that is assuming $\upsilon_1=0$ so that breakdowns never occur, the overflow stream is known to be hyperexponential (cf. [41]), so that the overflow station 2 separately can be analyzed as a $G|M|N_2$-system. However, our performance of interest may also depend on station 1.

(ii) For the simple single-station case, that is assuming either $N_1=\infty$ or $N_2=0$, this example is known as an "independent breakdown" model for which, despite its simplicity and independence, a closed form expression seems to be available only for the generating functions (cf. [8], p.103).

Parametrization Let the state $[\bar{n},\theta]$ with $\bar{n}=(n_1,n_2)$ denote the number of jobs n_1 and n_2 at stations 1 and 2 while $\theta=1$ or $\theta=0$ depending upon whether the system (output channel) is "up" or "down" respectively. Set $n=n_1+n_2$. Further, throughout let $1_{\{A\}}$ denote an indicator of an event A, i.e. $1_{\{A\}}=1$ if A is satisfied and $1_{\{A\}}=0$ otherwise.

Note that the results of section 2.4 can be adopted directly with $\bar{n}$ replaced by $[\bar{n},\theta]$ as we can simply interpret θ as the number of jobs n_3, being 1 or 0 at the "up" station in figure 2. From here onward this will be applied throughout without further mentioning.

The transition rates of the above model are now given by

$$q([\bar{n},\theta],\ [\bar{n}',\theta']) =$$

$$\begin{cases} \lambda 1_{\{n_1<N_1\}} & n_1'=n_1+1,\ n_2'=n_2,\ \theta'=\theta=0,1 \\ \lambda 1_{\{n_1=N_1,\ n_2<N_2\}} & n_2'=n_2+1,\ n_1'=n_1,\ \theta'=\theta=0,1 \\ \\ \mu_1(n_1)1_{\{\theta=1\}} & n_1'=n_1-1,\ n_2'=n_2,\ \theta'=1 \\ \mu_2(n_2)1_{\{\theta=1\}} & n_2'=n_2-1,\ n_1'=n_1,\ \theta'=1 \\ \\ \upsilon_0 1_{\{\theta=0\}} & n_1'=n_1,\ n_2'=n_2,\ \theta'=1 \\ \upsilon_1 1_{\{\theta=1\}} & n_1'=n_1,\ n_2'=n_2,\ \theta'=0 \end{cases} \tag{3.1}$$

Clearly, the corresponding Markov chain is irreducible at

$$S = \{[\bar{n},\theta]\,|\,0\leq n_1 \leq N_1,\ 0 \leq n_2 \leq N_2,\ \theta = 0,1\} \tag{3.2}$$

and by assumption we can choose a finite Q such that

$$Q \geq \lambda+\upsilon_0+\upsilon_1+\mu_1(n_1)+\upsilon_2(n_2) \tag{3.3}$$

for all $[\bar{n},\theta]\in S$. Consider the possible reward rates R:

$$\begin{aligned} R &= R_S([\bar{n},\theta]) = 1_{\{\theta=1\}}[\mu_1(n_1)+\mu_2(n_2)] \\ R &= R_j([\bar{n},\theta]) = 1_{\{\theta=1\}}\mu_j(n_j) \quad (j=1,2) \end{aligned} \tag{3.4}$$

Then G representing the average reward as according to section 2.4 is the actual "throughput" of the system (R_S) or of station j (R_j). We wish to compare this throughput G with values $\bar{G}$ and estimate

$$|\bar{G}-G|$$

for values $\bar{G}$ in various "approximate" situations. The key-step to this end is the estimation of the bias-terms as established in the next section.

3.2 Estimation of bias-terms

Let $V_t(.)$ be defined as according to (2.23) with R of any form (3.4) and define

$$\Delta_i V_t([\tilde{n},\theta]) = V_t([\tilde{n}+e_i,\theta])-V_t([\tilde{n},\theta]) \tag{3.5}$$

The following lemma then provides estimates for the bias-terms which will appear to be essential for the application of both theorem 2.1 and theorem 2.2 in various situations.

Lemma 3.1 For both $i=1,2$, all $[\tilde{n}+e_i,\theta]\in S$ and all $t\geq 0$, we have

$$0 \leq \Delta_i V_t([\tilde{n},\theta]) \leq 1 \tag{3.6}$$

We will give the proof for the case $R=R_S$. The proof for the case $R=R_j$ is almost identical and left as a remark (see remark 5.1 below).

Proof The proof will be given by induction to t. Clearly, (3.6) holds for $t=0$ as $V_0(.)=0$ by definition. Suppose that (3.6) holds for $t\leq m$. Then by virtue of (2.23), (3.1) and for convenience writing $h=Q^{-1}$, we obtain for any $[\tilde{n}+e_i,\theta]\in S$:

$$\Delta_i V_m([\tilde{n},\theta]) =$$

$$\Bigl\{h\Bigl[\mu_1(n_1)+\mu_2(n_2)+[\mu_i(n_i+1)-\mu_i(n_i)]\Bigr]\ 1_{\{\theta=1\}} +$$

$$h\ \lambda\ 1_{\{i=1\}}1_{\{n_1+1<N_1\}}\ V_m([\tilde{n}+e_1+e_1,\theta]) +$$

$$h\ \lambda\ 1_{\{i=1\}}1_{\{n_1+1=N_1,\ n_2<N_2\}}\ V_m([\tilde{n}+e_1+e_2,\theta]) +$$

$$h\ \lambda\ 1_{\{i=1\}}1_{\{n_1+1=N_1,\ n_2=N_2\}}\ V_m([\tilde{n}+e_1,\theta]) +$$

$$h\ \lambda\ 1_{\{i=2\}}1_{\{n_1<N_1\}}\ V_m([\tilde{n}+e_2+e_1,\theta]) +$$

$$h\ \lambda\ 1_{\{i=2\}}1_{\{n_1=N_1,\ n_2+1<N_2\}}\ V_m([\tilde{n}+e_2+e_2,\theta]) +$$

$$h\ \lambda\ 1_{\{i=2\}}1_{\{n_1=N_1,\ n_2+1=N_2\}}\ V_m([\tilde{n}+e_2,\theta]) +$$

$$h\ \upsilon_0 1_{\{\theta=0\}}V_m([\tilde{n}+e_i,1]) + h\ \upsilon_1 1_{\{\theta=1\}}V_m([\tilde{n}+e_i,0]) +$$

$$h\ \mu_1(n_1)\ 1_{\{\theta=1\}}V_m([\tilde{n}+e_i-e_1,1]) + h\ \mu_2(n_2)1_{\{\theta=1\}}V_m([\tilde{n}+e_i-e_2,1]) +$$

$$h\ [\mu_i(n_i+1)-\mu_i(n_i)]1_{\{\theta=1\}}V_m([\tilde{n}+e_i-e_i,1]) +$$

$$\Big[1-h\lambda-h\upsilon_\theta-h1_{\{\theta=1\}}[\mu_1(n_1)+\mu_2(n_2)+\mu_i(n_i+1)-\mu_i(n_i)]\Big]\ V_m([\bar{n}+e_i,\theta])\ \Big\}$$

$$-$$

$$\Big\{h[\mu_1(n_1)+\mu_2(n_2)]\ 1_{\{\theta=1\}}\ +$$

$$h\ \lambda\ 1_{\{i=1\}}1_{\{n_1+1<N_1\}}\ V_m([\bar{n}+e_1,\theta])\ +$$

$$h\ \lambda\ 1_{\{i=1\}}1_{\{n_1+1=N_1,\ n_2<N_2\}}\ V_m([\bar{n}+e_1,\theta])\ +$$

$$h\ \lambda\ 1_{\{i=1\}}1_{\{n_1+1=N_1,\ n_2=N_2\}}\ V_m([\bar{n}+e_1,\theta])\ +$$

$$h\ \lambda\ 1_{\{i=2\}}1_{\{n_1<N_1\}}\ V_m([\bar{n}+e_1,\theta])\ +$$

$$h\ \lambda\ 1_{\{i=2\}}1_{\{n_1=N_1,\ n_2+1<N_2\}}\ V_m([\bar{n}+e_2,\theta])\ +$$

$$h\ \lambda\ 1_{\{i=2\}}1_{\{n_1=N_1,\ n_2+1=N_2\}}\ V_m([\bar{n}+e_2,\theta])\ +$$

$$h\ \upsilon_0 1_{\{\theta=0\}}V_m([\bar{n},1])\ +\ h\ \upsilon_1 1_{\{\theta=1\}}V_m([\bar{n},0])\ +$$

$$h\ \mu_1(n_1)\ 1_{\{\theta=1\}}\ V_m([\bar{n}-e_1,1])\ +\ h\ \mu_2(n_2)1_{\{\theta=1\}}V_m([\bar{n}-e_2,1])\ +$$

$$\Big[1-h\lambda-h\upsilon_\theta-h\ 1_{\{\theta=1\}}[\mu_1(n_1)+\mu_2(n_2)]\ V_m([\bar{n},\theta])\ \Big\}$$

$$-$$

$$h[\mu_i(n_i+1)-\mu_i(n_i)]\ 1_{\{\theta=1\}}\ +$$

$$h\ \lambda\ 1_{\{i=1\}}1_{\{n_1+1<N_1\}}\ \Delta_1 V_m([\bar{n}+e_1,\theta])\ +$$

$$h\ \lambda\ 1_{\{i=1\}}1_{\{n_1+1=N_1,\ n_2<N_2\}}\ \Delta_2 V_m([\bar{n}+e_1,\theta])\ +$$

$$h\ \lambda\ 1_{\{i=1\}}1_{\{n_1+1=N_1,\ n_2=N_2\}}\ \Big[V_m([\bar{n}+e_1,\theta])\ -V_m([\bar{n}+e_1,\theta])\Big]\ +$$

$$h\ \lambda\ 1_{\{i=2\}}1_{\{n_1<N_1\}}\ \Delta_2 V_m([\bar{n}+e_1,\theta])\ +$$

$$h\ \lambda\ 1_{\{i=2\}} 1_{\{n_1=N_1,\ n_2+1<N_2\}}\ \Delta_2 V_m([\bar{n}+e_2,\theta]) +$$

$$h\ \lambda\ 1_{\{i=2\}} 1_{\{n_1=N_1,\ n_2+1=N_2\}}\ \Big[V_m([\bar{n}+e_2,\theta]) - V_m([\bar{n}+e_2,\theta])\Big] +$$

$$h\ v_0 1_{\{\theta=0\}} \Delta_i V_m([\bar{n},1]) + h\ v_1 1_{\{\theta=1\}} \Delta_i V_m([\bar{n},0]) +$$

$$h\ \mu_1(n_1)\ 1_{\{\theta=1\}} \Delta_i V_m([\bar{n}-e_1,1]) + h\ \mu_2(n_2) 1_{\{\theta=1\}} \Delta_i V_m([\bar{n}-e_2,\theta]) +$$

$$h\ 1_{\{\theta=1\}} [\mu_i(n_i+1)-\mu_i(n_i)]\ \Big[V_m([\bar{n}+e_i-e_i,1])-V_m([\bar{n},1])\Big] +$$

$$\Big[1-h\lambda-hv_\theta-h1_{\{\theta=1\}}[\mu_1(n_1)+\mu_2(n_2)+\mu_i(n_i+1)-\mu_i(n_i)]\Big]\ \Delta_i V_m([\bar{n},\theta]) \quad (3.7)$$

First of all, note that the fourth, seventh and one but last term in the last expression are equal to 0, while the coefficient in the last term is nonnegative by virtue of (3.2) and $h=Q^{-1}$. By substituting the induction hypothesis $\Delta_i V_m(.) \geq 0$ and recalling that $\mu_i(.)$ is nondecreasing we hereby conclude that $\Delta_i V_{m+1}(.) \geq 0$. To conclude the upper estimate $\Delta_i V_{m+1}(.) \leq 1$, recall that the one but last term is equal to 0 while its coefficient is exactly equal to the additional nonnegative first term $h\ 1_{\{\theta=1\}}[\mu_i(n_i+1)-\mu_i(n_i)]$. By substituing the hypothesis $\Delta_i V_m(.) \leq 1$ and using (3.2) again where $h=Q^{-1}$, we hereby verify also $\Delta_i V_{m+1}(.) \leq 1$. □

Remark 3.1 For the case $R=R_j$ the only difference is the first term in the right hand side of expression (3.7). Here it would become: $h\ 1_{\{j=i\}}[\mu_i(n_i+1)-\mu_i(n_i)]1_{\{\theta=1\}}$. For j=i the arguments thus remain. For $j\neq i$, substitution of the induction hypothesis (3.6) for t=m and (3.2) directly leads to (3.6) for t=m+1.

3.3 Application 1: Perturbation

Reconsider the model of section 3.1 with arrival rate $\bar{\lambda}$ in stead of λ. Let all quantities of section 3.1 be defined accordingly with an upper bar symbol where we assume that $\bar{Q}=Q$ also satisfies (3.2) with $\bar{\lambda}$ (note that this can always be established) and where we use the same reward $\bar{R} = R$.

In order to determine the effect of the perturbation $\lambda \rightarrow \bar{\lambda}$ we will verify the conditions of theorems 2.3 and 2.4. To this end, we conclude from (2.24), (3.1) and with $\Delta([\bar{n},\theta], [\bar{n}-e_i+e_j,\theta']) = 0$ if $\theta'=\theta$ and $i\neq j$:

$$\sum_{\substack{\theta'=0,1\\ i,j=0,1,2}} \Delta([\bar{n},\theta],\ [\bar{n}-e_i+e_j,\theta'])\ [V_t([\bar{n}-e_i+e_j,\theta'])-V_t([\bar{n},\theta])] =$$

$$[\bar{\lambda}-\lambda]\ \Big\{1_{\{n_1<N_1\}}[V_t([\bar{n}+e_1,\theta])-V_t([\bar{n},\theta]) +$$

$$1_{\{n_1=N_1,\,n_2<N_2\}}[V_t([\bar{n}+e_2,\theta])-V_t([\bar{n},\theta])]\Big\} \qquad (3.8)$$

Choosing

$$\Phi([\bar{n},\theta])=1 \qquad ([\bar{n},\theta] \in S) \qquad (3.9)$$

so that (2.27) is satisfied with $\beta=1$ for any initial state $\bar{\ell}=[\bar{n},\theta] \in S$, and recalling that $\bar{R} = R$, by substituing (3.8) in (2.25) and (2.29) using lemma 3.1 and applying theorem 2.3, we obtain:

Result 3.1 With$^{(\bar{G})}$ the total throughput of the system:

$$\bar{G} \geq G\ (\bar{G} \leq G) \quad \text{if} \quad \bar{\lambda} \geq \lambda\ (\bar{\lambda} \leq \lambda) \qquad (3.10)$$

and

$$|\bar{G}-G| \leq |\bar{\lambda}-\lambda| \qquad (3.11)$$

Remark 3.2 Result (3.10) and also (3.11) may seem trivial. However, as per counterintuitive examples such as in [28], [40], the actual throughput can sometimes be increased (decreased) by decreasing (increasing) arrival intensities at particular periods.

3.4 Application 2: Finite truncation

Assume that the second (or overflow) station is an infinite server queue, that is $N_2=\infty$ and $\mu_2(n_2)=n_2$. We are then involved with an infinite system. In order to apply methods (such as numerical) for finite systems, consider the approximate truncated model with $N_2=L$ (and $\mu_2(n_2)=n_2$), where L is some fixed finite number. Further, in this case consider the reward rates $R=R_j$, $j=1,2$ to evaluate the station throughput.

Thus let all expressions of section 3.1 with an upper bar correspond to the truncated model with $N_2=L$ while without bar to $N_2=\infty$, and use $R=\bar{R}=R_j$ for $j=1$ or $j=2$. First of all, note that

$$\bar{S} = \{[\bar{n},\theta]\,|\,0\leq n_1\leq N_1,\ 0\leq n_2\leq L,\ \theta=0,1\} \subset S$$

By comparing the transition rates (3.1) for $N_2=\infty$ (q-model) and $N_2=L$ ($\bar{q}$-model), we conclude from (2.24) and (3.1) that for $[\bar{n},\theta]\in S$:

$$\sum_{\substack{\theta'=0,1\\ i,j=0,1,2}} \Delta([\bar{n},\theta],\ [\bar{n}-e_i+e_j,\theta'])\left[V_t([\bar{n}-e_i+e_j,\theta'])-V_t([\bar{n},\theta])\right]$$

$$= \lambda\ 1_{(n_1=N_1,\ n_2=L\}}\left[V_t([\bar{n}+e_2,\theta])-V_t([\bar{n},\theta])\right] \qquad (3.12)$$

Choosing

$$\Phi([\bar{n},\theta]) = 1_{\{n_1=N_1,\ n_2=L\}} \qquad ([\bar{n},\theta]\in S) \qquad (3.13)$$

and applying lemma 3.1 thus implies that (2.25) is satisfied with $\alpha=\lambda$. As $\bar{R}=R$ at $\bar{S}$, theorem 2.4 thus directly formalizes the intuitively obvious result

$$\bar{G} \le G \qquad (R=\bar{R}=R_j,\ j=1,2) \qquad (3.14)$$

Lemma 3.2 With $\bar{0}=(0,0)$ and $\tau=\bar{\mu}_2(1+\upsilon_1/\upsilon_0)$, we have for all t

$$\bar{T}_t\Phi([\bar{0},1]) \le (\lambda\tau)^L/L! \qquad (3.15)$$

Proof This will only be given in heuristic steps. The details can be formalized but are rather technical and require application of theorem 2.2 itself and an inductive prooftechnique as in the proof of lemma 3.1.

Step 1

Clearly, by directly routing all jobs to the second station, the steady-state probability of L jobs at the second station is estimated from above by that for at least L jobs in an $M|M|\infty|\infty$-system with appropriate service rate $\bar{\mu}_2$ so as to take into account the breakdowns. This tail probability in turn is estimated from above by the right hand side of (3.15).

Step 2

Similarly to lemma 3.1.2 in [30] and based upon the special initial state $[\bar{0},1]$, one can show that

$$\bar{T}_t\Phi([\bar{0},1])$$

is nondecreasing in t. Combination of step 1 and 2 completes the proof. □

Recalling that $R=\bar{R}=R_j$ for j=1, or j=2, we now obtain from theorem 2.3, lemma 3.1 and (3.12):

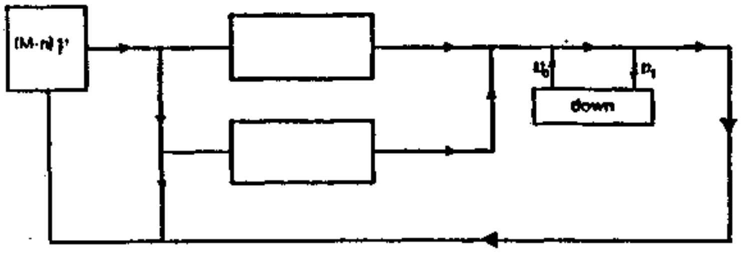
Result 3.2 With $\tau=\mu_2(1+\upsilon_1/\upsilon_0)$, we have

$$|\bar{G}-G| \le \lambda(\lambda\tau)^L/L! \qquad (3.16)$$

Remark 3.3 Note here the special role of the initial state $\bar{\ell}=[\bar{0},1]$ and scaling function Φ. □

3.5 Application 3: Modification (simple throughput bound)

Reconsider the model of section 3.1 in which the Poisson arrivals are replaced by a finite source input of M sources with exponential parameter γ per source. More precisely, that is, when n jobs are present in the system, a job will arrive with arrival intensity $(M-n)\gamma$. Here it is quite natural to assume that M is large while γ is small. Further, as in sector 3.4, let the second (overflow) station be an infinite server station, i.e. $N_2=\infty$ and $\mu_2(n_2)=n_2$ while also $\mu_1(n_1)=n_1$ is assumed. The total number of jobs in the system however can never exceed M.



The throughput $\bar{\lambda}$ of this system cannot be calculated easily. To this end, we will approximate this finite system by the infinite Poisson arrival system of section 3.1 with $N_2=\infty$, $\mu_2(n_2)=n_2$ and

$$\lambda = \gamma M \tag{3.17}$$

Note that this 'approximation' involves in fact both a truncation (or equivalently an infinite extension) and a state-dependent perturbation (arrival rates $(M-n)\gamma$ as opposed to λ).

We aim to show that λ is a simple upper bound and calculate an error bound on its accuracy. To this end, denote the expressions of section 3.1 for the finite source system with an upper bar and for the infinite system (as in section 3.1 itself) without. Hence,

$$\bar{S} = \{[\bar{n},\theta] \mid 0 \le n_1 \le N_1,\ 0 \le n_2 \le N_2;\ n_1+n_2 \le M,\ \theta=0,1\} \subset S$$

Further, use $R = \bar{R} = R_S$ given by (3.4). From the difference in the arrival intensities and with $\gamma=\lambda/M$ as by (3.17) and $n=n_1+n_2$, we obtain for $[\bar{n},\theta]\in\bar{S}$:

$$\sum_{\substack{\theta'=0,1\\ i,j=0,1,2}} \Delta([\bar{n},\theta],\ [\bar{n}-e_i+e_j,\theta'])\left[V_t([\bar{n}-e_i+e_j,\theta'])-V_t([\bar{n},\theta])\right] =$$

$$- n[\lambda/M]\left\{1_{(n_1<N_1)}[V_t[\bar{n}+e_1,\theta])-V_t([\bar{n},\theta])\right.$$

$$\left. + 1_{(n_1=N_1)}[V_t([\bar{n}+e_2,\theta])-V_t([\bar{n},\theta])\right\} \tag{3.18}$$

Choosing,

$$\Phi([\bar{n},\theta]) = n \qquad ([\bar{n},\theta]\in S) \tag{3.19}$$

and applying lemma 3.1 thus implies that (2.25) is satisfied with $\alpha=[\lambda/M]$. As $\bar{R} = R = R_S$, theorem 2.3 can thus be applied if we determine β.

Lemma 3.3 With $\bar{0}=(0,0)$ and W the expected sojourn time of a job in the infinite model, we have

$$\bar{T}_t \; \Phi([\bar{0},1]) \leq \lambda W \tag{3.20}$$

Proof This can be given similar to that of lemma 3.2 in [32], based on showing that the left hand side of (3.20) is non-decreasing in t (as step 2 in lemma 3.2), a bounding argument (as step 1 in lemma 3.1) but now for mean queue lengths and Little's result. □

Recalling that $R = \bar{R} = R_S$ and noting that the sojourn time is standardly estimated from above, we thus obtain from theorems 2.3 and 2.4, relation (3.18), lemma 3.1 and lemma 3.2

Result 3.3

$$0 \leq \lambda-\bar{\lambda} \leq \left[\lambda^2/M\right] \max\,[\mu_1^{-1},\ \mu_2^{-1}] \left[1+\upsilon_1/\upsilon_0\right]$$

Remark 3.4 As in section 3.4, again note the special role of the initial state $\bar{\ell}=[\bar{0},1]$ and scaling function Φ. Particularly, observe that the scaling function is unbounded. □

Evaluation

Various types of "approximate" systems may naturally arise when modeling and/or analyzing a Markov chain model such as representing a queuing network. Most notably, inaccuracies in system input data (perturbations), finite (numerical) approximations of infinite systems (truncations), simplifying transition assumptions for computational purposes, simple bounds (modifications), or comparison of system analogs under different parameters or protocols (comparisons) can be involved. A tool has been provided by which error bounds or comparison results for "approximate" modeling (numerical, bounding, approximating) can be concluded. The key to this tool is the estimation (bounding) of so-called bias-terms for Markov reward structures. An inductive Markov reward technique to this end can be employed. Particularly, this technique applies also to multi-dimensional structures such as of exponential queueing networks and allows practical complicating phenomena as blocking, dynamic routing, machine failures and job-priorities. Further application of this tool and most of all of bias-term estimates seems promising.

References

[1] Adan, I.J.B.F. and Wal, J. van der, (1987), "Monotonicity of the throughput of a closed queueing network in the number of jobs", Memorandum COSOR 87-03, Department of Mathematicsand Computer Science, Eindhoven University of Technology; To appear: *Opns. Res.*

[2] Adan, I.J.B.F. and Wal, J. van der, (1987), "Monotonicity of the throughput of an open queueing network in the interarrival and service times", Memorandum, Department of Mathematics and Computer Science, Eindhoven, University of Technology.

[3] Adan, I.J.B.F. and Wal, J. van der, (1989), "Monotonicity of the throughput in single server production and assembly networks with respect to the buffer sizes."; To appear: Proc. 1th International workshop on queueing systems with blocking.

[4] Barbour, A.D. (1982), "Generalized semi-Markov schemes and open queueing networks", *J. Appl. Prob.*, 469-474.

[5] Hinderer, K. (1978), "On approximate solutions of finite-stage dynamic programs". *Dynamic Programming and its Applications*, ed. M.L. Puterman, Academic Press, New York.

[6] Hordijk, A. and Ridder, Stochastic inequalities.

[7] Kemeny, J.G., Snell, J.L. and Knapp, A.W. (1966), "Denumerable Markov Chains", Van Nostrand, Princeton, N.J.

[8] Jaiswal, N.K. (1968), "Priority Queues", *Academic Press*, New York.

[9] Meyer, C.D. Jr. (1980), "The condition of a finite Markov chain and perturbation bounds for the limiting probabilities.", *SIAM J. Alg. Disc. Math.* 1, 273-283.

[10] Rohlicek, J.R. and Willsky, A.S. (1988), "The reduction of perturbed Markov generators: an algorithm exposing the role of transient states", *J.A.C.M.* 35, 675-696.

[11] Ross, S.M. (1970), "Applied probability models with optimization application", *Holden-Day*, San Francisco.

[12] Ross, S.M. (1984), "Introduction to probability models", *Academic Press*, New York.

[13] Ross, S.M. (1987), "Approximating transition probabilities and mean occupation times in continuous-time Markov chains", *P.E.I.S.* 1, 251-264.

[14] Schweitzer, P.J. (1968), "Perturbation theory and finite Markov chains", *J.Appl. Prob.* 5, 401-413.

[15] Seneta, E. (1967), "Finite approximations to infinite non-negative matrices. Proc. Cambridge, *Phil. Soc.* 63, 983-992.

[16] Seneta, E. (1968), "The principles of truncations in applied probability", *Comm. Math. Univ. Carolina 9*, 533-539.

[17] Seneta, E. (1980), "Non-negative matrices and Markov chains, *Springer Verlag.*

[18] Shanthikumar, J.G. and Yao, D.D. (1986), "The effect of increasing service rates in closed queueing network", *J. Appl. Prob.* 23, 474-483.

[19] Shanthikumar, J.G. and Yao, D.D. (1987), "Stochastic monotonicity of the queue lengths in closed queueing networks, Research Report, University of California, Berkeley, *Opns. Res.* 35, 583-588.

[20] Shanthikumar, J.G. and Yao, D.D. (1987), "General queueing networks: Representation and stochastic monotonicity", Proc. of 26th IEEE Conference on Decision and Control, 1084-1087.

[21] Shanthikumar, J.G. and Yao, D.D. (1988), "Throughput bounds for closed queueing networks with queue-dependent service rates", *Performance Evaluation* 9, 69-78.

[22] Shanthikumar, J.G. and Yao, D.D. (1988), "Monotonicity properties in cyclic queueing networks with finite buffers, Proceedings first international workshop on queueing networks with blocking, North Carolina, may 1988.

[23] Stoyan, D. (1983), "Comparison Method for Queues and other Stochastic models", *Wiley*, New York.

[24] Suri, R. (1985), "A concept of monotõnicity and its characterization for closed queueing networks", *Opns. Res.* 33, 606-624.

[25] Psoucas, P. and Walrand, J. (1983), "Monotonicity of throughput in non-Markovian networks", To appear: *J. Appl. Probability.*

[26] Tijms, H.C. (1986), "Stochastic Modelling and Analysis; A computational Approach, *Wiley*, New York.

[27] Tijms, H.C. and Eikeboom, A.M. (1986), "A simple technique in Markovian control with applications to resource allocation", *Oper. Res. Letters* 1, 25-32.

[28] Van Dijk, N.M. (1988), "Simple bounds for queueing systems with breakdowns", *Performance Evaluation* 8, 117-128.

[29] Van Dijk, N.M. (1988), "A formal proof for the insensitivity of simple bounds for finite multi-server non-exponential tandem queues based on monotonicity results", *Stochastic procc. Appl.* 27, 261-277.

[30] Van Dijk, N.M. (1988), "Perturbation theory for unbounded Markov reward processes with applications to queueing", *Adv. Appl. Prob.* 20, 99-111.

[31] Van Dijk, N.M. (1988), "Simple performance bounds for non-product form queueing networks", Proceedings first international workshop on queueing networks with blocking, North Carolina, may 1989.

[32] Van Dijk, N.M. (1989), "A simple Throughput Bound for Large Closed Queueing Networks With Finite Capacities"; To appear: *Performance Evalutation.*

[33] Van Dijk, N.M. (1989), "Analytic Error Bounds For Approximations of Queueing Networks with an Application to Alternte Routing"; To appear: *J. Austr. Math. Society.*

[34] Van Dijk, N.M. (1989), "Truncation of Markov Chains with applications to queueing", Research report 53, Free University, Amsterdam, Submitted.

[35] Van Dijk, N.M. (1988), "Approximate Uniformization for Continuous-time Markov Chains with an Application to Performability Analysis", Research report 54, Free University, Amsterdam, Submitted.

[36] Van Dijk, N.M. (1988), "Error Bounds For Comparing Open and Closed Queueing Networks with an Application to Performability Analysis", Research report 56, Free University, Amsterdam, Submitted.

[37] Van Dijk, N.M. and Lamond, B.F. (1988), "Bounds for the call congestion of finite single-server exponential tandem queues, *Opns. Res.* 36, 470-477.

[38] Van Dijk, N.M. and Puterman, M.L. (1988), "Perturbation theory for Markov reward processes with applications to queueing systems", *Adv. Appl. Prob.* 20, 79-89.

[39] Van Dijk, N.M., Tsoucas, P. and Walrand, J. (1988), "Simple bounds and monotonicity of the call congestion of finite multiserver delay systems, *Probability in the Engineering and Informational Sciences* 2, 129-138.

[40] Van Dijk, N.M. and Van der Wal, J. (1989), "Simple bounds and monotonicity results for multi-server exponential tandem queues"; To appear: *Queueing Systems*.

[41] Van Doorn, E.A. (1984), "On the overflow process from a finite Markovian queue, *Performance Evalutation* 4, 233-240.

[42] De Waal, P.R. and Van Dijk, N.M. (1988), "Monotonicity of Performance Measures in a Processor Sharing Queue", Research report 51, Free University, Amsterdam.

[43] Whitt, W. (1978), "Approximations of dynamic programs I", *Math. Oper. Res.* 3, 231-243.

[44] Whitt, W. (1981), "Comparing counting processes and queues", *Adv. Appl. Prob.* 13, 207-220.

[45] Whitt, W. (1986), "Stochastic comparisons for non-Markov processes", *Math. Oper. Res.* 11 (4), 608-618.

[46] Yoon, B.S. and Shanthikumar, B.S. (1988), "bounds and approximations for the transient behaviour of continuous time Markov chains", Research report, University of California, Berkeley; To appear: *P.E.I.S.*.