Bounds on the coupling time in acyclic queuing networks

Dopper, J.G.

Jantien G. Dopper

Bounds on the coupling time in acyclic queueing networks

Master thesis
defended on May 30, 2006

Mathematisch Instituut
Universiteit Leiden

Laboratoire ID-IMAG
Grenoble

Contents

Acknowledgement

Je tiens à remercier mes responsables de stage, Bruno Gaujal et Jean-Marc Vincent, pour m'avoir accueillie au sein du laboratoire pour mon stage, et pour leur soutien et pour leur aide pendant mon stage. Ensuite, je tiens a remercier Jérôme Vienne, qui a amélioré et adapte le logiciel `Psi2` à mes besoins. Sans son travail, je n'aurais pas été capable d'effectuer ce projet.

Je tiens à remercier aussi Grégory Mounié qui était toujours là pour réparer ma machine, pour boire un thé et discuter et pour être *chauffeur du taxi* quand je finissait tard le travail.

Je tiens également à remercier tous les membres du laboratoire, car c'est aussi grâce à eux que j'ai passé six mois très agréables a Montbonnot. Ce sont les discussions (parfois très énervées) à la *kfet* qui m'ont ouvert des nouveaux horizons au plan scientifique, mais ce sont aussi les activités plutôt sociales et détentes qui ont fait que les connaissances que j'ai acquises pendant le séjour á Grenoble, ne s'arrêtent pas que aux maths. Tout cela fait que je garde des très bons souvenirs de mon séjour à Grenoble.

Daarnaast bedank ik mijn Leidse begeleider, prof. dr. L.C.M. Kallenberg, zonder wie ik nooit in Grenoble was gekomen.

Tenslotte bedank ik mijn ouders, die mij tijdens mijn (lange) studietraject gesteund hebben.

Chapter 1

Introduction

In daily life, one is often confronted with the phenomenon of queueing: in the supermarket, at the post office or when phoning a company or government institution. While waiting on the phone, a waiting customers sometime is told how many customers are waiting in front of him, or what the mean waiting time is [1]. In order to be able to inform the waiting customers about their mean waiting time, some kind of analysis of the queueing system has to be performed. This master thesis deals with a part of evaluation of the performance of queueing systems, and especially with simulation techniques.

Queueing systems do not only interfere in daily life, but are used in a variety of applications such as the performance evaluation of computer systems and communication networks. In the mathematically modelling of queueing systems, an important tool is Markov chains. One of the main points of interest is the behaviour of the queueing system in the long run. For an irreducible, ergodic (i.e. aperiodic and positive-recurrent) finite-state Markov chain with probability matrix P, this long run behaviour follows the stationary distribution of the chain given by the unique probability vector π which satisfies the linear system $\pi = \pi P$. However, it may be hard to compute this stationary distribution, especially when the finite state space is huge which is frequent in queueing models. In this case, several approaches have been proposed to get samples of the long run behaviour of the system.

The most classical methods are indirect. They consists in first computing an estimation of π and then sample according to this distribution.

Estimating π can be done through efficient *numerical iterative methods* solving the linear system $\pi = \pi P$ [10]. Even if they converge fast, they do not scale when the state space (and thus P) grows. Another approach to estimate π is to use *regenerative simulation* [3, 6] based on the fact that on a trajectory of a Markov chain returning to its original state, the frequency of the visits to each state is steady state distributed. This technique does not suffer from statistical bias but is very sensitive to the return time to the regenerative state. This means that the choice of the initial state is crucial but also that regenerative simulation complexity increases fast with the state space. This is exponential in the number of queues.

There also exist direct techniques to sample states of Markov chain according to its stationary distribution. The classical method is *Monte Carlo simulation*. This

---

[1] In Dutch, this comes down to the the phrases everyone is familiar with like "Er zijn nog twee wachtenden voor u" ("There are two people waiting in front of you") and "De gemiddelde wachttijd bedraagt 6 minuten" ("The mean waiting time is 6 minutes"). Lazy companies however only put on some (often annoying) music, and the only information one gets is that "al onze medewerkers zijn in gesprek, een momentje geduld alstublieft" ("All employees are occupied, so please hang on").

method is based on the fact that for an irreducible aperiodic finite-state Markov chain with initial distribution $\pi^{(0)}$, the distribution $\pi^{(n)}$ of the chain at time n converges to π as n gets very large:

$$\lim_{n\to\infty} \pi^{(n)} = \lim_{n\to\infty} \pi^{(0)} P^n = \pi.$$

After running the Markov chain for a long time, the state of the chain will not depend on the initial state anymore. However, the important question is how long is long enough? That is, when is n sufficiently large such that $|\pi^{(n)} - \pi| \leqslant \epsilon$ for a certain $\epsilon > 0$? Moreover, the samples generated by this method will always be biased.

In 1996, Propp and Wilson [7] solved these problems for Markov chain simulation by proposing an algorithm which returns exact samples of the stationary distribution very fast. The striking difference between Monte Carlo simulation and this new algorithm is that Propp and Wilson do not simulate into the future, but go backwards in time. The main idea is, while going backwards in time, to run several simulations, starting from all states until the output-state at $t = 0$ is the same for all of these. If the output is the same for all runs, we say that the chain has coupled. Because of this coupling property and the backward scheme, this algorithm has been called Coupling From The Past (from now on: CFTP).

When the coupling from the past technique is applicable, one gets in finite time one state with steady-state distribution. Then one can use either one long-run simulation from this state or replicate independently the CFTP algorithm to get a sample of independent steady-state distributed variables. The replication technique has been applied successfully in finite capacity queueing networks with blocking and rejection (very large state-space) [13]. The efficiency of the simulation allows also the estimation of rare events (blocking probability, rejection rate) [12].

One can apply the CFTP technique to finite capacity queueing networks. The aim of this master thesis is to study the simulation time needed to get a stationary sample for finite capacity networks. More precisely, we study the coupling time τ of a CFTP algorithm (*i.e.* the number of steps needed to provide one sample). This coupling time is a random variable and we try to set bounds on the expected coupling time.

The organisation of this thesis is as follows: In Chapter 2 we will introduce the simulation method of CFTP. We will linger on the difference between the backward algorithm and the forward equivalent. In Chapter 3 we introduce queueing networks and explain that CFTP is applicable after uniformization of the system. Furthermore, we will derive some general bounds on the coupling time.

Chapter 4 is dedicated to the analysis of a single queue with a finite capacity and a single server (M/M/1/C queue). Of course, one does not need to run simulations to obtain the stationary distribution of this simple model since they can be easily computed [9, 11]. However, we will need the results on the coupling time of a simple queue in order to construct a bound for acyclic networks. For the M/M/1/C queue, we will derive a recurrence expression for the exact coupling time. Moreover, we will provide some easily calculable bounds which are quite good with respect to the exact coupling time. Finally, we will study the coupling time by a generating function approach, so to set a stochastical bound on the distribution of the coupling time.

In Chapter 5 we construct a bound on the coupling time of a acyclic queueing network, by using the results of the previous chapter. This bound is tested in Chapter 6. We ran several runs of the CFTP algorithm for different acyclic queueing networks. These simulations are carried out with the simulation software `Psi2` which is developed in the ID-IMAG laboratory, Grenoble. Finally, Chapter 7

summarises the results of this thesis and points out topics for further research.

This thesis is the result of a six month internship in the ID-IMAG laboratory in Grenoble, France. In order to make clear for the reader what results are directly related to the research project in Grenoble and what results have been published in literature, we mark every result that is taken from the literature with its reference by [·]. The main outline is that the results of the Chapters 2 and 3 are mainly based on the literature, whereas the Chapters 4, 5 and 6 provide mainly new material.

Chapter 2

The coupling from the past algorithm

The aim of this chapter is introducing a simulation method which is called coupling from the past. This simulation algorithm was introduced by Propp and Wilson [7]. To do this, we will first define a Markov chain in terms of a transition function which is driven by events. Then we will take a look at a coupling method into the future and its restrictions, before turning to coupling from the past in section 2.3.

2.1 Markov chains and simulation

Let $\{X_n\}_{n \in \mathbb{N}}$ be a discrete time Markov chain with a finite state space $\mathcal{S}$ and a transition matrix $P = (p_{i,j})$. In order to run simulations of the Markov chain, we need to specify how to get from X_n to X_{n+1}. The main ingredients for this are events and a transition function.

Definition 2.1. [13] An event e is an application defined on $\mathcal{S}$ that associates to each state $s \in \mathcal{S}$ a new state.

The set of all events is called $\mathcal{E}$, and we suppose that the set $\mathcal{E} = \{e^0, \dots, e^M\}$ is finite. Each event e has a probability $p(e)$ and we suppose that these probabilities are strictly positive for each event $e \in \mathcal{E}$. The events in a Markov chain can be quite natural. For example, in a random walk on $\mathbb{Z}$, the set $\mathcal{E}$ consists of two elements, e^0 and e^1. The event e^0 is a step from s to $s+1$ and the other event e^1 is a step from s to $s-1$ for all $s \in \mathcal{S}$.

The second ingredient is the transition function

$$\phi : \mathcal{S} \times \mathcal{E} \to \mathcal{S},$$

with $\mathbb{P}(\phi(i, e) = j) = p_{i,j}$ for every pair of states $(i, j) \in \mathcal{S}$ and for any $e \in \mathcal{E}$. The function ϕ could be considered as a construction algorithm and e the innovation for the chain. Now, the evolution of the Markov chain is described as a stochastic recursive sequence

$$X_{n+1} = \phi(X_n, e_{n+1}), \tag{2.1}$$

with X_n the state of the chain at time n and $\{e_n\}_{n \in \mathbb{N}}$ an independent and identically distributed sequence of random variables. The sequence of states $\{X_n\}_{n \in \mathbb{N}}$ defined by the recurrence (2.1) is called a trajectory.

To run a simulation using (2.1), we need to specify how to generate the events. A way to generate the events is by using the inverse transformation method [9], p.

644. Let u be distributed uniformly on $[0, 1]$, and define $f : [0, 1] \to \mathcal{E}$ as:

$$f(u) = \begin{cases} e^0 & \text{for } u \in [0, p(e^0)), \\ e^1 & \text{for } u \in [p(e^0), p(e^0) + p(e^1)), \\ \vdots & \vdots \\ e^i & \text{for } u \in \left[\sum_{j=0}^{i-1} p(e^j), \sum_{j=0}^{i} p(e^j)\right), \\ \vdots & \vdots \\ e^M & \text{for } u \in \left[\sum_{j=0}^{M-1} p(e^j), 1\right]. \end{cases} \tag{2.2}$$

Now, when we start in state $s \in \mathcal{S}$, one can run a simulation of a trajectory of a Markov chain from 0 to m by performing Algorithm 1.

Algorithm 1 Forward simulation of a trajectory of length m

---

  n=0;
  s $\leftarrow$ X_0 {choice of initial value X_0}
  repeat
    n=n+1;
    u $\leftarrow$ Random_number; {generation of u_n}
    e $\leftarrow$ $f(u)$; {determination of e_n}
    s $\leftarrow$ $\phi(s, e)$; {determination of state at time n}
  until n=m
  return s

Several Markov chains, each of them being constructed using a different function ϕ, all have the same transition matrix P. The next example illustrates this:

Example 2.1. Suppose we have two states. We will construct three Markov chains, each of them having a transition matrix P with

$$P = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix}.$$
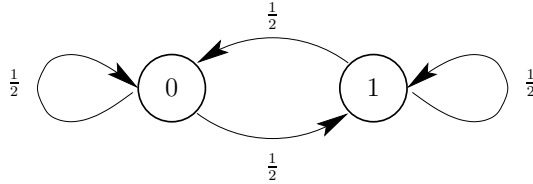
Figure 2.1: *The transition graph of a Markov chain with transition matrix P of Example 2.1.*

1. *Markov chain 1*
   Let the set of events $\mathcal{E}$ consists of two events e^0 and e^1 with $p(e^0) = p(e^1) = 1/2$ and let the function ϕ be defined as:

   $$\phi(s, e^0) = \begin{cases} 0 & \text{for } s = 0, \\ 1 & \text{for } s = 1, \end{cases}$$

   and

   $$\phi(s, e^1) = \begin{cases} 1 & \text{for } s = 0, \\ 0 & \text{for } s = 1. \end{cases}$$

5

This means that if we apply event e^0 , the chain stays in its present state, no matter what its present state is. The event e^1 represents a transition to the other state for both states 0 as 1.

2. *Markov chain 2*

Let the set of events $\mathcal{E}$ again consists of two events e^0 and e^1 with $p\left(e^0\right) = p\left(e^1\right) = 1/2$ and let the function ϕ' be defined as:

$$\phi'\left(s, e^0\right) \quad = \quad \begin{cases} 0 & \text{for } s = 0, \\ 0 & \text{for } s = 1, \end{cases}$$

and

$$\phi'\left(s, e^1\right) \quad = \quad \begin{cases} 1 & \text{for } s = 0, \\ 1 & \text{for } s = 1. \end{cases}$$

Remark that indeed the events are different form the events used for the previous Markov chain. Now, if the present state is state 0 and we apply event e^0, we stay in state 0 just as in the previous Markov chain. However, if we apply event e^0 if we are in state 1, we make a transition to state 0, and this is different from the definition of event e^0 in the first Markov chain.

3. *Markov chain 3*

Let the set of events $\mathcal{E}$ now consist of four events e^0, e^1, e^2 and e^3 with $p\left(e^i\right) = 1/4$ for $i = 0, \ldots, 3$. Let the function ϕ'' be defined as:

$$\phi''\left(s, e^0\right) \quad = \quad \begin{cases} 0 & \text{for } s = 0, \\ 0 & \text{for } s = 1, \end{cases}$$

and

$$\phi''\left(s, e^1\right) \quad = \quad \begin{cases} 0 & \text{for } s = 0, \\ 1 & \text{for } s = 1, \end{cases}$$

and

$$\phi''\left(s, e^2\right) \quad = \quad \begin{cases} 1 & \text{for } s = 0, \\ 0 & \text{for } s = 1, \end{cases}$$

and

$$\phi''\left(s, e^3\right) \quad = \quad \begin{cases} 1 & \text{for } s = 0, \\ 1 & \text{for } s = 1. \end{cases}$$

Now the three Markov chains all have the same transition matrix P. $\triangle$

2.2 Coupling into the future

Let $\phi^{(n)} : S \times \mathcal{E}^n \to S$ denote the function whose output is the state of the chain after n iterations, starting in state $s \in \mathcal{S}$. That is,

$$\phi^{(n)}\left(s, e_{1 \to n}\right) = \phi\left(\ldots \phi\left(\phi\left(s, e_1\right), e_2\right), \ldots, e_n\right). \tag{2.3}$$

This notation can be extended to sets of states. So for a set of states $A \subset \mathcal{S}$ we note

$$\phi^{(n)}(A, e_{1 \to n}) = \left\{ \phi^{(n)}(s, e_{1 \to n}), s \in A \right\}.$$

Assume we run $|\mathcal{S}|$ copies of a Markov chain, and each copy starts in a different state $s \in \mathcal{S}$. Then the number of states that can be attained after n iterations is equal to $\left| \phi^{(n)}(\mathcal{S}, e_{1 \to n}) \right|$. Consider an arbitrary sequence of events $\{e_n\}_{n \in \mathbb{N}}$. Let $a_n^f = \left| \phi^{(n)}(\mathcal{S}, e_{1 \to n}) \right|$. The index f indicates that we are using the forward scheme.

Lemma 2.1. *The sequence of integers $\{a_n^f\}_{n \in \mathbb{N}}$ is non-increasing.*

Proof. The cardinal a_n^f of the image of $\phi^{(n-1)}(\mathcal{S}, e_{1 \to n-1})$ by $\phi(., e_n)$ is less or equal than the cardinal a_{n-1}^f of $\phi^{(n-1)}(\mathcal{S}, e_{1 \to n-1})$. Since

$$\phi^{(n)}(\mathcal{S}, e_{1 \to n}) = \phi\left(\phi^{(n-1)}(\mathcal{S}, e_{1 \to n-1}), e_n \right),$$

the sequence $\{a_n^f\}_{n \in \mathbb{N}}$ is non-increasing. $\qquad \square$

Intuitively this is clear, since the transition function ϕ maps each pair (s, e) on exactly one state. Therefore, when starting with $m \leqslant |\mathcal{S}|$ copies, the number of states one can reach after n iterations cannot exceed m.

Theorem 2.1. *Let ϕ be a transition function on $\mathcal{S} \times \mathcal{E}$. There exists an integer l^* such that*

$$\lim_{n \to +\infty} \left| \phi^{(n)}(\mathcal{S}, e_{1 \to n}) \right| = l^* \text{ almost surely.}$$

Proof. This result is based on the previous lemma and the fact that $\mathcal{S}$ is finite. Consider an arbitrary sequence of events $\{e_n\}_{n \in \mathbb{N}}$. Lemma 2.1 implies that the sequence $\{a_n^f\}_{n \in \mathbb{N}}$ converges to a limit l. Because the sizes of these sets belong to the finite set $\{1, \cdots, |\mathcal{S}|\}$, there exists a $n_0 \in \mathbb{N}$ such that

$$a_{n_0}^f = \left| \phi^{(n_0)}(\mathcal{S}, e_{1 \to n_0}) \right| = l.$$

Consider now l^* the minimum value of l among all possible sequences of events. Then there exists a sequence of events $\{e_n^*\}_{n \in \mathbb{N}}$ and an integer n_0^* such that

$$\left| \phi^{(n)}(\mathcal{S}, e_{1 \to n}^*) \right| = l^* \quad \text{for all } n \geqslant n_0^*.$$

As a consequence of the Borel-Cantelli Lemma, almost all sequences of events $\{e_n\}_{n \in \mathbb{N}}$ include the pattern $e_{1 \to n_0^*}^*$. Consequently, the limit of the cardinality of $\phi^{(n)}(\mathcal{S}, e_{1 \to n})$ is less than or equal to l^*. The minimality of l^* completes the proof. $\qquad \square$

This means that when starting the $|\mathcal{S}|$ copies from the different initial states, after running enough iterations, the set of attainable states will be of size ℓ^*.

Definition 2.2. The system couples if $\ell^* = 1$ with probability 1.

Note that the coupling property of a system ϕ depends only on the structure of ϕ and that the probability measure on $\mathcal{E}$ does not affect the coupling property. The proof of Theorem 2.1 shows that in order to couple with probability 1, it suffices to have at least one sequence of events that ensures coupling. Since the existence of such a sequence depends on the chosen transition function ϕ, the choice of the transition function is important. The next example shows this.

Example 2.2. Recall that the three Markov chains from Example 2.1, constructed by the functions ϕ, ϕ' and ϕ'' respectively, all have the same transition matrix P. These three transition functions are shown in Figure 2.2. In this figure, we see for the three transition functions two intervals $[0,1]$, for state 0 and 1 respectively, and the events. The correspondence between the unit interval and the events is obtained by the method of (2.2). A dashed interval means that the transition function makes a transition to state 0 and a blank interval means that the transition function makes a transition to state 1.
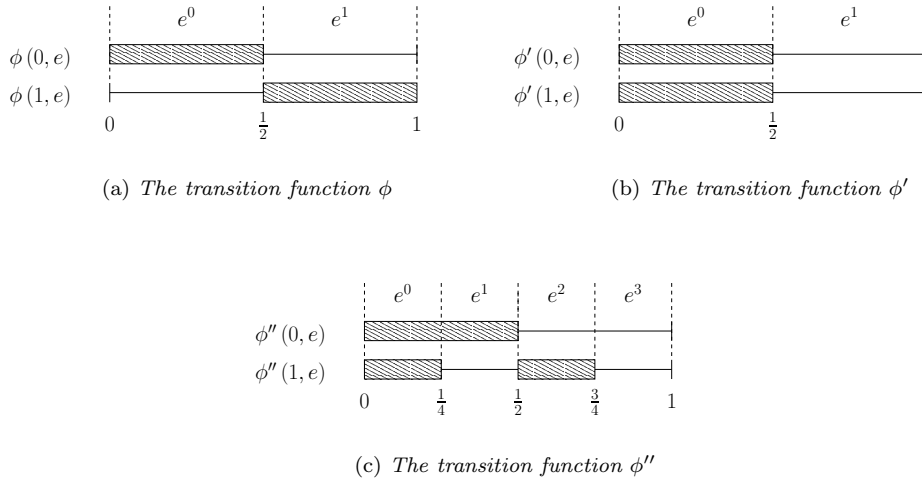

(a) *The transition function ϕ*

(b) *The transition function ϕ'*

(c) *The transition function ϕ''*

Figure 2.2: *Three different transition functions all having the same transition matrix.*

Now, coupling corresponds to having an interval $(a,b) \in [0,1]$ which has the same colour for state 0 as for state 1. So it is clear from Figure 2.2.a. that the system represented by the transition function ϕ can never couple since the dashed intervals do not match. However, if we look at Figure 2.2.b. for function ϕ', we see that every iteration step leads to coupling. Finally, function ϕ'' only assures coupling for the events e^0 and e^3 (Figure 2.2.c). $\triangle$

Definition 2.3. The *forward coupling time* τ^f is a random variable defined by

$$\tau^f = \min\left\{ n \in \mathbb{N} \text{ such that } \left|\phi^{(n)}(\mathcal{S}, e_{1 \to n})\right| = 1 \right\}.$$

Provided that the system couples, the forward coupling time τ^f is almost surely finite. From time τ^f on, all trajectories issued from all initial states at time 0 have collapsed in only one trajectory.

From now on, we suppose the Markov chain $\{X_n\}_{n \in \mathbb{N}}$ is irreducible and aperiodic. Then X_n has a unique stationary distribution. However, $\mathbb{P}(X_{\tau^f} = s) \neq \pi_s$ in general. This means that the distribution of the state where coupling occurs (i,.e the stochastic variable X_{τ^f}) is not the stationary distribution of the Markov chain. The next counterexample, obtained from [5], p. 81, shows that forward simulation does not yield a sample that has the stationary distribution.

Example 2.3 (Counterexample, [5]). Suppose we have a Markov chain with state space $\mathcal{S} = \{0, 1\}$ and transition matrix

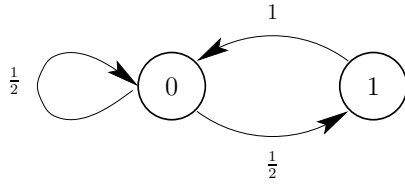$$P = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ 1 & 0 \end{pmatrix}.$$

Figure 2.3: *The transition graph of the Markov chain of Example 2.3*

The stationary distribution π is the solution of $\pi = \pi P$ with the normalization equation $\pi_0 + \pi_1 = 1$. Solving leads to

$$\pi = (\pi_0, \pi_1) = \left(\frac{2}{3}, \frac{1}{3}\right).$$

Let us run two copies of the chain, one starting in state 0 and the other one in state 1, and suppose that they couple at time τ^f. Now we will show that $\mathbb{P}(X_{\tau^f} = 0) \neq \pi_0$. Because of the definition of τ^f, at time $\tau^f - 1$ the two copies cannot be in the same state. Thus at time $\tau^f - 1$, one copy is in state 0 and the other in state 1. Since $p_{10} = 1$, the copy being in state 1 at time $\tau^f - 1$, will be in state 0 at time τ^f. Hence, $\mathbb{P}(X_{\tau^f} = 0) = 1$. and thus coupling always occurs in state 0. This is not in agreement with the stationary distribution π. $\triangle$

Let $e^*_{1 \to n^*_0}$ be a sequence of events that ensures coupling. Then the probability that this sequence occurs equals $p(e^*_1) \cdot p(e^*_2) \cdot \cdots \cdot p(e^*_{n_0})$. If $\tau^f > k \cdot n^*_0$, then the sequence $e^*_{1 \to n^*_0}$ does not appear in the events in the simulation run from time 1 up to time $k \cdot n^*_0$. Then the sequence $e^*_{1 \to n^*_0}$ does not appear in the events used for the simulation from time $i \cdot n^*_0 + 1$ up to time $(i + 1) \cdot n^*_0$ for $i = 0, \ldots, k - 1$. Since this are k independent events, it follows that,

$$\mathbb{P}(\tau^f \geqslant k.n^*_0) \leqslant \left(1 - p(e^*_1).p(e^*_2)\ldots p(e^*_{n_0})\right)^k. \tag{2.4}$$

Thus the existence of some pattern $e^*_{1 \to n^*_0}$ that ensures coupling, guarantees that τ^f is stochastically upper bounded by a geometric distribution.

2.3 Coupling from the past

2.3.1 General coupling from the past

The iteration scheme of (2.3) can be reversed in time as it is usually done in the analysis of stochastic point processes. To do this, one needs to extend the sequence of events to negative indexes. The difference between coupling into the future and coupling from the past is the order of using the events. Using coupling into the future, a trajectory of length n of a single state s is obtained by choosing a sequence of events $e_{1 \to n}$ and applying (2.3). In a simulation run, the first event used is e_1, and the last one is e_n. While using coupling into the past, a trajectory of length n of a single state s is obtained by choosing a sequence of events $e_{-n+1 \to 0}$ and applying:

$$\phi^{(n)}(s, e_{-n+1 \to 0}) = \phi\left(\ldots \phi\left(\phi\left(s, e_{-n+1}\right), e_{-n+2}\right), \ldots, e_0\right).$$

Thus now the the first event used is e_{-n+1}, and the last one is e_0. In other words, coupling into the past adds events at the beginning of the simulation, whereas coupling into the future adds events at the end of the simulation.

Consider an arbitrary sequence $e_{-n+1\to0}$ of events. Analogous to the definition of $\left\{a_n^f\right\}_{n\in\mathbb{N}}$ of the previous section, we define the sequence $\left\{a_n^b\right\}_{n\in\mathbb{N}}$ with $a_n^b = \left|\phi^{(n)}\left(\mathcal{S}, e_{-n+1\to0}\right)\right|$. Now, a_n^b counts the number of possible states at time 0 when applying the sequence $e_{-n+1\to0}$ in a simulation run. By the same reasoning as in Lemma 2.1 and Theorem 2.1, one can show that the sequence $\left\{a_n^b\right\}_{n\in\mathbb{N}}$ is non-increasing and converges to a limit. Consequently, the system couples if the sequence $\left\{a_n^b\right\}_{n\in\mathbb{N}}$ converges almost surely to a set with only one element. Provided that the system couples, there exists a finite time τ^b, the *backward coupling time* almost surely, defined by

$$\tau^b = \min\left\{n \in \mathbb{N}; \text{ such that } \left|\phi^{(n)}\left(\mathcal{S}, e_{-n+1\to0}\right)\right| = 1\right\}.$$

A sequence $\{u_n\}_{n\in\mathbb{Z}}$ is called *stationary* if for every $n = 1, 2, \ldots$ we have

$$(u_0, \ldots, u_n) = (u_k, \ldots, u_{k+n}) \qquad \text{for all } k \in \mathbb{Z}$$

in distribution. Every independent and identically distributed sequence is stationary.

The main result of the backward scheme is the following theorem [7].

Theorem 2.2. *Provided that the system couples, the state when coupling occurs for the backward scheme, is steady state distributed.*

Proof (based on [14]). For all $n \geqslant \tau^b$ and all $s \in \mathcal{S}$ we can split the backward scheme in first a trajectory from time $-n+1$ up to time $-\tau^b$ and then a trajectory from time $-\tau^b + 1$ up to time 0. In doing so, we see that

$$\phi^{(n)}\left(s, e_{-n+1\to0}\right) = \phi^{\left(\tau^b\right)}\left(\phi^{\left(n-\tau^b\right)}\left(s, e_{-n+1\to-\tau^b}\right), e_{-\tau^b+1\to0}\right).$$

By definition of τ^b, for all $s \in \mathcal{S}$ we have $\phi^{\left(\tau^b\right)}\left(s, e_{-\tau^b+1\to0}\right) = s'$ for a certain $s' \in \mathcal{S}$. Therefore,

$$\phi^{\left(\tau^b\right)}\left(\phi^{\left(n-\tau^b\right)}\left(s, e_{-n+1\to-\tau^b}\right), e_{-\tau^b+1\to0}\right) = s',$$

and thus

$$\phi^{(n)}\left(s, e_{-n+1\to0}\right) = \phi^{\left(\tau^b\right)}\left(s, e_{-\tau^b+1\to0}\right), \tag{2.5}$$

for $n \geqslant \tau^b$ and all $s \in \mathcal{S}$.

Let Y denote the state generated by the backward scheme and let a be an arbitrary state. Then

$$\begin{aligned}
\mathbb{P}\left(Y = a\right) &= \mathbb{P}\left(\phi^{\left(\tau^b\right)}\left(s, e_{-\tau^b+1\to0}\right) = a\right), \\
&= \mathbb{P}\left(\bigcup_{n=0}^{\infty}\left\{\phi^{\left(\tau^b\right)}\left(s, e_{-\tau^b+1\to0}\right) = a, \tau^b \leqslant n\right\}\right), \\
&= \mathbb{P}\left(\bigcup_{n=0}^{\infty}\left\{\phi^{(n)}\left(s, e_{-n+1\to0}\right) = a, \tau^b \leqslant n\right\}\right), \\
&= \lim_{n\to\infty}\mathbb{P}\left(\phi^{(n)}\left(s, e_{-n+1\to0}\right) = a\right).
\end{aligned}$$

Since the sequence $\{e_n\}_{n\in\mathbb{Z}}$ is independent and identically distributed, the sequence is stationary and therefore

$$\phi^{(n)}\left(s, e_{-n+1\to0}\right) = \phi^{(n)}\left(s, e_{1\to n}\right)$$

10

in distribution. Hence,

$$\lim_{n\to\infty} \mathbb{P}\left(\phi^{(n)}\left(s, e_{-n+1\to 0}\right) = a\right) = \lim_{n\to\infty} \mathbb{P}\left(\phi^{(n)}\left(s, e_{1\to n}\right) = a\right).$$

Since the Markov chain is irreducible and aperiodic, $\lim_{n\to\infty} \mathbb{P}\left(\phi^{(n)}\left(s, e_{1\to n}\right) = a\right) = \pi_a$. It follows that $\mathbb{P}\left(Y = a\right) = \pi_a$ and thus the value generated by the backward scheme is steady state distributed. $\square$

The above proof goes wrong for coupling into the future. Due to the different order of placing the events in a simulation run, the equivalence of (2.5) for coupling into the future, $\phi^{(n)}\left(e_{1\to n}\right) = \phi^{\left(\tau^f\right)}\left(s, e_{1\to \tau^f}\right)$, does not hold. We can see this as follows: by the definition of τ^f, for all $s \in S$ we have $\phi^{\left(\tau^f\right)}\left(s, e_{1\to \tau^f}\right) = s'$ for a certain $s' \in \mathcal{S}$. Then for all $n \geqslant \tau^f$:

$$\phi^{(n)}\left(e_{1\to n}\right) = \phi^{\left(n-\tau^f\right)}\left(\phi^{\left(\tau^f\right)}\left(s, e_{1\to \tau^f}\right), e_{\tau^f+1\to n}\right) = \phi^{\left(n-\tau^f\right)}\left(s', e_{\tau^f+1\to n}\right).$$

Since $\phi^{\left(n-\tau^f\right)}\left(s', e_{\tau^f+1\to n}\right) \neq s'$ in general, it follows that

$$\phi^{(n)}\left(e_{1\to n}\right) \neq \phi^{\left(\tau^f\right)}\left(s, e_{1\to \tau^f}\right)$$

in general.

From the result of 2.2, a general algorithm (2) sampling the steady state can be constructed.

Algorithm 2 Backward-coupling simulation (general version)

---

    for all $s \in \mathcal{S}$ **do**
      $y(s) \leftarrow s$ {choice of the initial value of the vector y, $n = 0$}
    end for
    repeat
      e $\leftarrow$ Random_event; {generation of e_{-n+1}}
      for all $s \in \mathcal{S}$ **do**
        $y(s) \leftarrow y(\phi(s, e))$;
        {$y(s)$ state at time 0 of the trajectory issued from s at time $-n+1$}
      end for
    until All $y(x)$ are equal
    return $y(x)$

The working of the algorithm is illustrated in Figure 2.4 for a Markov chain with state space $\mathcal{S} = \{0, 1, 2, 3\}$. First, we set $n = 1$ and we generate an event $e_{-n+1} = e_0 \in \mathcal{E}$. Suppose that

$$\begin{cases} \phi\left(0, e_0\right) &=& 0, \\ \phi\left(1, e_0\right) &=& 2, \\ \phi\left(2, e_0\right) &=& 3, \\ \phi\left(3, e_0\right) &=& 1. \end{cases}$$

Since the output is not a single state, we are obliged to do a second run with $n = 2$. Suppose that we generate event e_{-1} and that the sequence $e_{-1\to 0} = e_{-1}, e_0$ delivers

$$\begin{cases} \phi\left(0, e_{-1\to 0}\right) &=& 2, \\ \phi\left(1, e_{-1\to 0}\right) &=& 3, \\ \phi\left(2, e_{-1\to 0}\right) &=& 1, \\ \phi\left(3, e_{-1\to 0}\right) &=& 2. \end{cases}$$
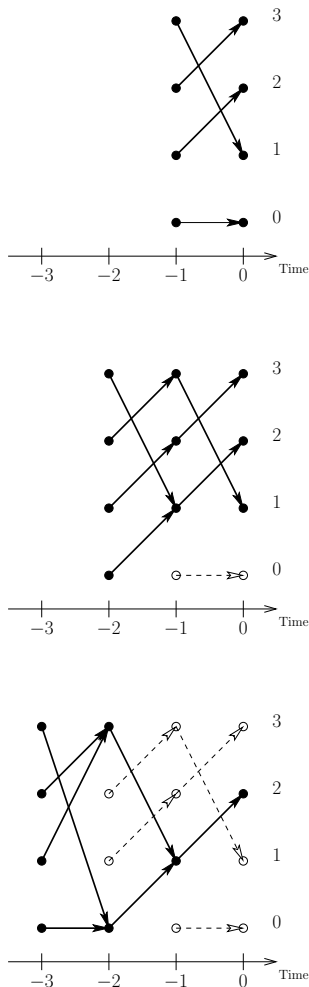
Figure 2.4: *The coupling from the past algorithm applied to a Markov chain with four states. The transitions carried out by the algorithm are in solid lines for every step and the others are dashed.*

Since the chain has not coupled yet, a third iteration is carried out by generating event e_{-2}. Suppose that the sequence $e_{-2 \to 0}$ yields

$$
\begin{cases}
\phi\left(0, e_{-2 \to 0}\right) &=& 2, \\
\phi\left(1, e_{-2 \to 0}\right) &=& 2, \\
\phi\left(2, e_{-2 \to 0}\right) &=& 2, \\
\phi\left(3, e_{-2 \to 0}\right) &=& 2.
\end{cases}
$$

Now the chain has coupled and the output of the algorithm is state 2.

Note that for the iteration at time $-n$ we re-use the sequence of events from time $-n$ up to 0. This means that this sequence of events needs to be stored. One can ask why we cannot take a new sequence of events for every iteration, so as to avoid the storage of events. However, the following example, taken from [5] p. 82, shows that by taking a new sequence of events for every iteration, biased samples are obtained.

Example 2.4. [5] We use again the Markov chain of Example 2.3. We have two

events e^0 and e^1 in the chain and a transition function ϕ with:

$$\phi\left(s, e^0\right) \quad = \quad \begin{cases} 0 & \text{for } s = 0, \\ 0 & \text{for } s = 1, \end{cases}$$

and

$$\phi\left(s, e^1\right) \quad = \quad \begin{cases} 1 & \text{for } s = 0, \\ 0 & \text{for } s = 1. \end{cases}$$

Now we apply the modified algorithm with a new random sequence of events for each iteration. Let Y denote the output of this modified algorithm and τ^b is the backward coupling time. From the results of Example 2.3 it follows that $\mathbb{P}\left(\tau^b = 1\right) = 1/2$ and that $\mathbb{P}\left(Y = 0 \,\middle|\, \tau^b = 1\right) = 1$. When the coupling time τ^b equals 2, the first iteration has not lead to coupling. This happens with probability $1/2$. From time -2, there are four possible sequences of events. Of these four sequences, the three sequences consisting of at least once the event e^0 lead to coupling. Therefore, $\mathbb{P}\left(\tau^b = 2\right) = 1/2 \cdot 3/4 = 3/8$. Of the three coupling sequences of length two, there are two (namely $\{e^0, e^0\}$ and $\{e^1, e^0\}$) which lead to coupling in state 0. Hence, $\mathbb{P}\left(Y = 0 \,\middle|\, \tau^b = 2\right) = 2/3$. Now:

$$\begin{aligned} \mathbb{P}\left(Y = 0\right) \quad &= \quad \sum_{k=0}^{\infty} \mathbb{P}\left(\tau^b = k, Y = 0\right) \\ &\geqslant \quad \mathbb{P}\left(\tau^b = 1, Y = 0\right) + \mathbb{P}\left(\tau^b = 2, Y = 0\right) \\ &= \quad \mathbb{P}\left(\tau^b = 1\right) \mathbb{P}\left(Y = 0 \,\middle|\, \tau^b = 1\right) + \mathbb{P}\left(\tau^b = 2\right) \mathbb{P}\left(Y = 0 \,\middle|\, \tau^b = 2\right) \\ &= \quad \frac{1}{2} \cdot 1 + \frac{3}{8} \cdot \frac{2}{3} = \frac{3}{4} > \frac{2}{3} = \pi_0. \end{aligned}$$

Thus this modified algorithm does not generate samples which are distributed according to the stationary distribution. $\triangle$

The Algorithm 2.4 picks an event e, computes the update function $\phi\left(s, e\right)$ and and adds this step to the trajectory in every loop. This procedure is performed for all $s \in \mathcal{S}$ until coupling at time τ^b. The cost of the algorithm is the number of calls to the transition function ϕ. Therefore, the mean complexity c_ϕ to generate one sample is

$$c_\phi = O\left(\mathbb{E}\left[\tau^b\right] \cdot |\mathcal{S}|\right). \tag{2.6}$$

2.3.2   Monotone coupling from the past

From (2.6) it follows that the coupling time τ^b is of fundamental importance for the efficiency of the sampling algorithm. To improve its complexity, we could try to reduce the factor $|\mathcal{S}|$ or reduce the coupling time. We will first take a look at reducing the factor $|\mathcal{S}|$.

Definition 2.4. A relation $\prec$ on a set S is called a *partial order* if it satisfies the following three properties:

(i) reflexivity: $a \prec a$ for all $a \in S$.

(ii) anti-symmetry: if $a \prec b$ and $b \prec a$ for any $a, b \in S$, then $a = b$

(iii) transitivity: if $a \prec b$ and $b \prec c$ for any $a, b, c \in S$, then $a \prec c$.

Definition 2.5. A transition function $\phi : \mathcal{S} \times \mathcal{E} \to \mathcal{S}$ is called monotone if for every $e \in \mathcal{E}$ and every $x, y \in \mathcal{S}$ with $x \prec y$ we have $\phi\left(x, e\right) \prec \phi\left(y, e\right)$.

13

Suppose that the state space $\mathcal{S}$ is partially ordered by a partial order $\prec$ and denote by MAX and MIN the set of maximal, respectively minimal elements of $\mathcal{S}$ for the partial order $\prec$. Then for every $s \in \mathcal{S}$ there exists a $s_1 \in MIN$ and a $s_2 \in MAX$ such that $s_1 \prec s \prec s_2$. Furthermore, suppose that the transition function ϕ is monotone for each event e. Consequently,

$$\phi(s_1, e) \prec \phi(s, e) \prec \phi(s_2, e),$$

and

$$\phi(s_1, e_{-n+1\to 0}) \prec \phi(s, e_{-n+1\to 0}) \prec \phi(s_2, e_{-n+1\to 0}).$$

Thus it is sufficient to simulate trajectories starting from the maximal and minimal states. Note that when there is only one minimal and only one maximal element in the state space $\mathcal{S}$, the monotonicity property reduces the number of starting points for each iteration from $|\mathcal{S}|$ to 2.

Now, suppose that our system is monotone with $|MAX| = |MIN| = 1$. Then it suffices to run the two copies starting from MAX and MIN. For each iteration we need to run the two copies up to time 0 because we cannot test whether coupling has occurred from the previous iterations. We run the copies until coupling occurs at time τ^b. Then the total number of calls to the transition function ϕ equals

$$2 \cdot \left(1 + 2 + 3 + \cdots + \tau^b\right) = \left(\tau^b\right)^2 + \tau^b, \tag{2.7}$$

where the 2 in front comes from the fact that we run two copies. This means that the mean complexity $c_\phi = O\left(\mathbb{E}\left[\left(\tau^b\right)^2\right]\right)$ for the monotone case.

Now we will show that by smartly choosing the starting points, we can reduce the complexity of the monotone case. So let us take the starting points $-1, -2, -4, \dots$. Then we run the chain until the smallest integer k with $2^k > \tau^b$. By doing so, one can overshoot the real coupling time. However, this is not a problem since we have seen that

$$\phi^{(n)}\left(s, e_{-n+1\to 0}\right) = \phi^{\left(\tau^b\right)}\left(s, e_{-n+1\to 0}\right)$$

for every $n \geqslant \tau^b$. Then the number of calls to ϕ equals

$$2 \cdot \left(1 + 2 + 4 + \cdots + 2^{k-1} + 2^k\right),$$

and by induction one can show that

$$2 \cdot \left(1 + 2 + 4 + \cdots + 2^{k-1} + 2^k\right) < 2^{k+2}. \tag{2.8}$$

Now we will compare the number of iterations of (2.7) with (2.8). By definition of k, we have $2^{k-1} \leqslant \tau^b \leqslant 2^k$. When applying the monotone algorithm without the doubling period, it follows from (2.7) that we need to perform at least $\left(2^{k-1}\right)^2 + 2^{k-1}$ calls to ϕ. When applying the monotone algorithm with a doubling period, one performs less then 2^{k+2} calls. One can easily verify that $2^{k+2} < \left(2^{k-1}\right)^2 + 2^{k-1}$ for $k > 4$. Thus as soon as $k > 4$, the doubling period scheme demands less calls than the one step scheme. This means that as soon as τ^b is bigger than 16, the doubling scheme is more effective in monotonic systems with $|MAX| = |MIN| = 1$.

The monotone version with doubling period of Algorithm (2) is given by Algorithm (3). In this case, we need to store the sequence of events in order to preserve the coherence between the trajectories driven from $MIN \cup MAX$. A realization of the monotone CFTP algorithm with doubling period on a random walk on five states is shown in Figure 2.5. The partial order on this random walk is the ordinary $\leqslant$.

The equivalent of (2.8) in case of general monotony (that is: there exist MAX and MIN, but its size is not necessarily equal to 1), is

$$
\begin{aligned}
(|MAX| + |MIN|)\left(1 + 2 + \cdots + 2^k\right) &\leqslant (|MAX| + |MIN|) \cdot 4 \cdot 2^{k-1}, \\
&\leqslant (|MAX| + |MIN|) \cdot 4 \cdot \mathbb{E}\left[\tau^b\right].
\end{aligned}
$$

Thus this monotone version with doubling period leads to a mean complexity c_ϕ

$$
c_\phi = O\left(\mathbb{E}\left[\tau^b\right] \cdot (|MIN| + |MAX|)\right). \tag{2.9}
$$

Algorithm 3 Backward-coupling simulation (monotone version)

---

  n=1;
  R[n]=Random_event;{array will R stores the sequence of events }
  repeat
    n=2.n;
    for all $s \in MIN \cup MAX$ **do**
      $y(s) \leftarrow s$ {Initialize all trajectories at time $-n$}
    end for
    for i=n downto n/2+1 **do**
      R[i]=Random_event; {generates all events from time $-n+1$ to $\frac{n}{2}+1$}
    end for
    for i=n downto 1 **do**
      for all $s \in MIN \cup MAX$ **do**
        $y(s) \leftarrow \phi(y(s), R[i])$
      end for
    end for
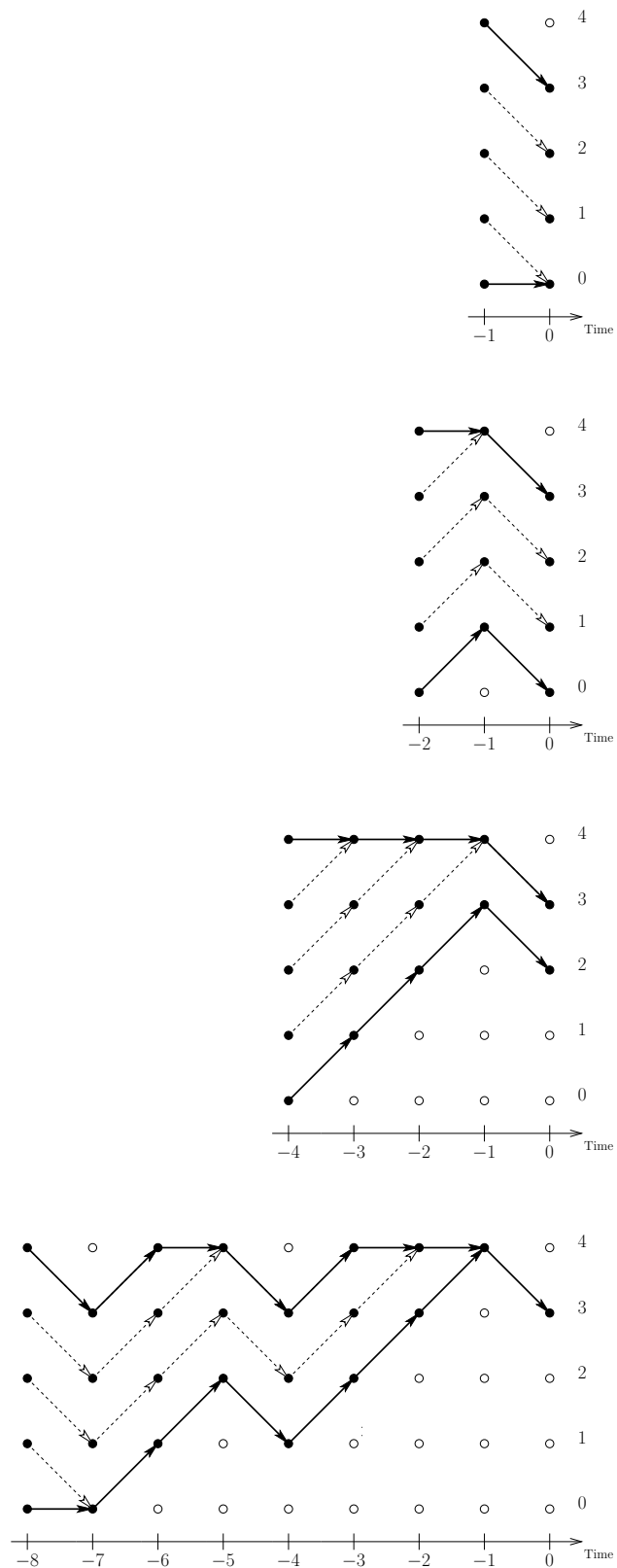 **until** All $y(s)$ are equal
  return $y(s)$

Figure 2.5: *A run of the coupling from the past algorithm on a random walk on $\{0,1,2,3,4\}$. The trajectories starting from the maximal state $MAX = 4$ and the minimal state $MIN = 0$ are in solid lines, whereas the trajectories for all other states are dashed. The output is state 3.*

16

Chapter 3

The CFTP-algorithm in open Markovian queueing networks

In this chapter, we will explain how to apply the CFTP-algorithm on a open network of queues. Furthermore, we will derive some general bounds on the backward coupling time in a network.

3.1 CFTP in a queueing network

Consider an open network Q consisting of $K + 1$ queues $Q_0, \ldots, Q_K$. Each queue Q_i has a finite capacity (with the place at the server included), denoted by C_i, $i = 0, \ldots K$. Thus the state space of a single queue Q_i is $\mathcal{S}_i = \{0, \ldots C_i\}$. Hence, the state space $\mathcal{S}$ of the network is $\mathcal{S} = \mathcal{S}_1 \times \cdots \times \mathcal{S}_K$. The state of the system is described by a vector $s = (s_0, \ldots, s_K)$ with s_i the number of customers in queue Q_i. The state space is partially ordered by the component-wise ordering and so there is a maximal state MAX when all queues are full and a minimal state MIN when all queues are empty.

The network evolves in time due to exogenous customer arrivals from outside of the network and to service completions of customers. After finishing his service at a server, a customer is either directed to another queue by a certain routing policy or leaves the network. A routing policy determines to which queue a customer will go, taking into account the global state of the system. Moreover, the routing policy also decides what happens to a customer if he is directed to a queue the buffer of which is filled with C_i customers. We assume that customers who enter from outside the network to a given queue arrive according to a Poisson process. Furthermore, we suppose that the service times at server i are independent and exponentially distributed with parameter μ_i.

An event in this network is characterized by the origin queue, a list of the destination queues, a routing policy and an enabling condition. A natural enabling condition for the event *end of service* is that there must be at least one customer in the queue. network. As in the preceding chapter, $\mathcal{E} = \{e^0, \ldots, e^M\}$ denotes the finite collection of events of the network. With each event e^i is associated a Poisson process with parameter λ_i. If an event occurs that does not satisfy the enabling condition, the state of the system is unchanged.

Example 3.1. Consider the acyclic queueing network of Figure 3.1 which is characterized by 4 queues and 6 events. These 6 events are characterized by the origin

of customers, one destination queue (where queue Q_{-1} represents the exterior), an enabling condition and the routing policy. In this network, customers who are directed towards a queue which completely filled, are rejected and thus lost for the system. In this network, the service rates are $\mu_0 = \lambda_1 + \lambda_2$, $\mu_1 = \lambda_3$, $\mu_2 = \lambda_4$ and $\mu_3 = \lambda_5$.
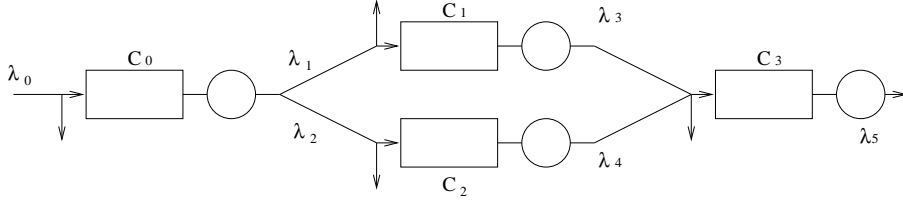
event	rate	origin	destination	enabling condition	routing policy
e^0	λ_0	Q_{-1}	Q_0	none	rejection if Q_0 is full
e^1	λ_1	Q_0	Q_1	$s_0 > 0$	rejection if Q_1 is full
e^2	λ_2	Q_0	Q_2	$s_0 > 0$	rejection if Q_2 is full
e^3	λ_3	Q_1	Q_3	$s_1 > 0$	rejection if Q_3 is full
e^4	λ_4	Q_2	Q_3	$s_2 > 0$	rejection if Q_3 is full
e^5	λ_5	Q_3	Q_{-1}	$s_3 > 0$	none

Figure 3.1: *Acyclic queueing network with rejection.*

For a transition function ϕ we get for the event e^1:

$$\phi(.,e^1) : (s_0, s_1, s_2, s_3) \longmapsto \begin{cases} (s_0 - 1, s_1 + 1, s_2, s_3) & \text{if } s_0 \geqslant 1 \text{ and } s_1 < C_1, \\ (s_0 - 1, s_1, s_2, s_3) & \text{if } s_0 \geqslant 1 \text{ and } s_1 = C_1, \\ (s_0, s_1, s_2, s_3) & \text{if } s_0 = 0. \end{cases}$$

$\triangle$

In addition to monotone functions, we will also define monotone events.

Definition 3.1. An event e is monotone if $\phi(x,e) \prec \phi(y,e)$ for every $x, y \in \mathcal{S}$ with $x \prec y$ and $\prec$ a partial order on $\mathcal{S}$.

Let $N_i : \mathcal{S} \rightarrow \mathcal{S}_i$ with $N_i(s_0, \ldots, s_K) = s_i$. So N_i returns the number of customers in queue Q_i.

Proposition 3.1. *A routing event with rejection if all destinations queues are full, is a monotone event.*

Proof. ([13]) The proof is carried out by examining all possibilities. Let $(x,y) \in \mathcal{S}^2$ such that $x \leqslant y$. Let e be a routing event with origin Q_i and a list of destinations $Q_{j_1}, Q_{j_2}, \ldots, Q_{j_l}$. If $y_i = 0$ then also x_i is and thus the event does not change the states x and y. Hence, $\phi(x,e) = x \leqslant y = \phi(y,e)$. If $y_i > 0$ and $x_i \geqslant 0$, then

$$N_i(\phi(y,e)) = y_i - 1 \geqslant \max\{x_i - 1, 0\} = N_i(\phi(x,e)).$$

Let Q_{j_k} be the first non saturated queue in state y. If the first non saturated queue for state x is strictly before Q_{j_k}, then it follows that $\phi(x,e) \leqslant \phi(y,e)$. If Q_{j_k} is also the first non saturated queue for both state x as y then also $\phi(x,e) \leqslant \phi(y,e)$. Since $x \leqslant y$, these are the only possibilities and this completes the proof. $\square$

Moreover, also other usual events such as routing with blocking and restart and routing with a index policy rule (e.g. join the shortest queue) are monotone events [4, 13].

In order to apply the CFTP algorithm, we construct a discrete-time Markov chain by the uniformizing the system by a Poisson process with rate $\Lambda = \sum_{i=0}^{M} \lambda_i$. One can see this Poisson process as a clock which determines when an event transition takes place. To choose which specific transition actually takes place, the collection $\mathcal{E}$ of events of the network is randomly sampled with

$$p_i = \mathbb{P}\left(\text{event } e^i \text{ occurs}\right) = \frac{\lambda_i}{\Lambda}. \qquad \text{for } i = 0, \dots, M.$$

Proposition 3.2. *The uniformized process has the same stationary distribution as the queueing network, and so does the discrete time Markov chain which is embedded in the uniformized Markov process.*

For a more detailed overview of uniformization and a proof of the above Proposition, see Appendix A. Provided that events are monotone, the CFTP algorithm can be applied to queueing networks to build steady-state sampling of the network. One then only needs to simulate the two trajectories starting from the minimal state MIN and the maximal state MAX. From now on, we consider queueing networks with only monotone events.

3.2 General remarks on the coupling time

To get a first idea of the behaviour of the coupling time τ^b of the CFTP algorithm, we ran the CFTP algorithm on the network of Example 3.1. So we produced samples of coupling time. The parameters used for the simulation are the following:

Capacity of the queues: 10 for every queue Q_i, $i = 0, \dots 3$
Event rates: $\lambda_1 = 1.4$, $\lambda_2 = 0.6$, $\lambda_3 = 0.8$, $\lambda_4 = 0.5$ and $\lambda_5 = 0.4$.

The global input rate λ_0 is varying. The number of samples used to estimate the mean coupling time is 10,000. The result is displayed in Figure 3.2.
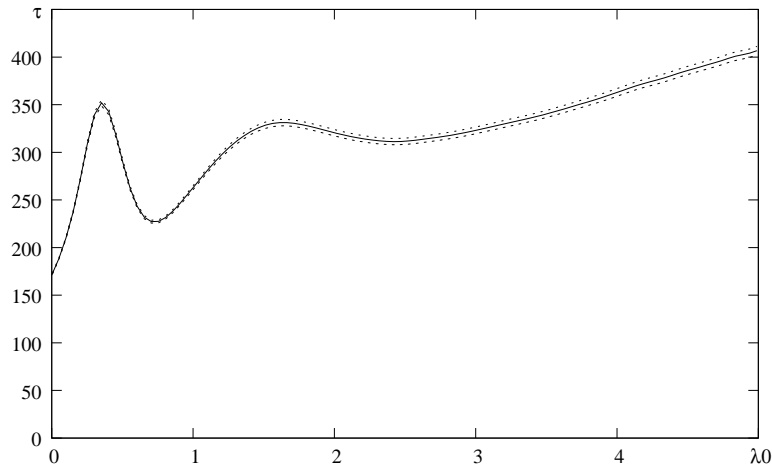

Figure 3.2: *Mean coupling time for the acyclic network of Figure 3.1 when the input rate varies from 0 to 5, with 95% confidence intervals.*

This type of curve is of fundamental importance because the coupling time corresponds to the simulation duration and is involved in the simulation strategy (long run versus replication). These first results can be surprising because they exhibit a strong dependence on parameters values. The aim of this thesis is now to understand more deeply what are the critical values for the network and to build bounds on the coupling time that are non-trivial.

As in section 2.3, τ^b refers to the backward coupling time of the chain, whereas τ^f refers to the forward coupling time.

Proposition 3.3. *The backward coupling time τ^b and the forward coupling time τ^f have the same probability distribution.*

Proof. ([14]). Compute the probability

$$\mathbb{P}(\tau^f > n) = \mathbb{P}\left(\left|\phi^{(n)}\left(\mathcal{S}, e_{1 \to n}\right)\right| > 1\right).$$

Since the process $\{e_n\}_{n \in \mathbb{Z}}$ is stationary, shifting the process to the left leads to

$$\mathbb{P}\left(\left|\phi^{(n)}\left(\mathcal{S}, e_{1 \to n}\right)\right| > 1\right) = \mathbb{P}\left(\left|\phi^{(n)}\left(\mathcal{S}, e_{-n+1 \to 0}\right)\right| > 1\right) = \mathbb{P}\left(\tau^b > n\right).$$

Therefore, τ^f and τ^b have the same distribution. $\qquad\square$

Hence, if we want to make any statement about the probability distribution of the coupling time τ^b of CFTP, we can use the conceptually easier coupling time τ^f. By combining Proposition 3.3 with Inequality 2.4 we see that the existence of a sequence that ensures coupling also guarantees that τ^b is stochastically upper bounded by a geometric distribution.

Definition 3.2. Let τ_i^b denote the *backward coupling time on coordinate i* of the state space. Thus τ_i^b is the smallest n for which

$$\left|\left\{N_i\left(\phi^{(n)}\left(\mathcal{S}, e_{-n+1 \to 0}\right)\right)\right\}\right| = 1.$$

In the same way, we define τ_i^f as the smallest n for which

$$\left|\left\{N_i\left(\phi^{(n)}\left(\mathcal{S}, e_{1 \to n}\right)\right)\right\}\right| = 1.$$

Because coordinate s_i refers to queue Q_i, the random variable τ_i^b (τ_i^f respectively) represents the coupling time from the past (the coupling time into the future respectively) of queue Q_i. Once all queues in the network have coupled, the CFTP algorithm returns one value and hence the chain has coupled. Thus

$$\tau^b = \max_{1 \leqslant i \leqslant K}\{\tau_i^b\} \leqslant_{st} \sum_{i=1}^{K} \tau_i^b. \tag{3.1}$$

By taking expectation and interchanging sum and expectation we obtain:

$$\mathbb{E}\left[\tau^b\right] = \mathbb{E}\left[\max_{1 \leqslant i \leqslant K}\{\tau_i^b\}\right] \leqslant \mathbb{E}\left[\sum_{i=1}^{K} \tau_i^b\right] = \sum_{i=1}^{K} \mathbb{E}\left[\tau_i^b\right]. \tag{3.2}$$

It follows from Proposition 3.3 that τ^b and τ^f have the same distribution. The same holds for τ_i^f and τ_i^b. Hence $\mathbb{E}\left[\tau_i^b\right] = \mathbb{E}\left[\tau_i^f\right]$ and

$$\mathbb{E}\left[\tau^b\right] \leqslant \sum_{i=1}^{K} \mathbb{E}\left[\tau_i^f\right]. \tag{3.3}$$

The bound given in (3.3) is interesting because $\mathbb{E}\left[\tau_i^f\right]$ is sometimes amenable to explicit computations, as will be shown in following chapter. In order to derive those bounds, one may provide yet other bounds, by making the coupling state explicit.

Definition 3.3. The hitting time $h_{j\to k}$ in a Markov chain X_n is defined as

$$h_{j\to k} = \inf_{\mathbb{N}}\{n \text{ such that } X_n = k | X_0 = j\} \text{ with } j,k \in \mathcal{S}.$$

The hitting time $h_{j\to k}$ with $j,k \in \mathcal{S}_i$ is the hitting time of a single queue Q_i of the network. Thus $h_{0\to C_i}$ represents the number of steps it takes a queue Q_i to go from state 0 to state C_i. Suppose that $h_{0\to C_i} = n$ for the sequence of events $e_{1\to n}$. Because of monotonicity of ϕ we have

$$
\begin{aligned}
C_i &= N_i\left(\phi^{(n)}\left(MIN, e_{1\to n}\right)\right) \\
&\leqslant N_i\left(\phi^{(n)}\left(s, e_{1\to n}\right)\right) \\
&\leqslant N_i\left(\phi^{(n)}\left(MAX, e_{1\to n}\right)\right) = C_i,
\end{aligned}
$$

with $s \in \mathcal{S}$, $MIN = (0,\ldots,0) \in \mathcal{S}$ and $MAX = (C_0,\ldots,C_K) \in \mathcal{S}$. Hence, coupling has occurred in Queue Q_i. So $h_{0\to C_i}$ is an upper bound on the forward coupling time τ^f of queue Q_i. The same argumentation holds for $h_{C_i\to 0}$. Thus

$$\mathbb{E}\left[\tau_i^f\right] \leqslant \mathbb{E}\left[\min\{h_{0\to C_i}, h_{C_i\to 0}\}\right]. \tag{3.4}$$

Hence,

$$\mathbb{E}\left[\tau^b\right] \leqslant \sum_{i=1}^{K}\mathbb{E}\left[\tau_i^f\right] \leqslant \sum_{i=1}^{K}\mathbb{E}\left[\min\{h_{0\to C_i}, h_{C_i\to 0}\}\right] \leqslant \sum_{i=1}^{K}\min(\mathbb{E}h_{0\to C_i}, \mathbb{E}h_{C_i\to 0}). \tag{3.5}$$

Chapter 4

Coupling time in the M/M/1/C queue

The M/M/1/C queue is wellknown and there is no need to run simulations to get the stationary distribution since this distribution can quite easily be calculated ([9] p. 487-489 and [11] p. 191). However, in this chapter we will take a look at the distribution of the coupling time in the M/M/1/C queue since this will serve as a building block for establishing bounds on coupling time in acyclic networks. In section 4.1 we focus on the mean coupling time. By linking the coupling time in a new way with hitting times we are able to derive the exact coupling time and easier calculable bounds on the coupling time. In section 4.2, we will explore another approach which is based on formal series to derive bounds on the probability distribution of the coupling time.

4.1 Mean coupling time

First, we will shortly introduce the M/M/1/C queueing model. The M/M/1/C queueing model consists of a single queue with one server. Customers arrive at the queue according to a Poisson process with rate λ and the service time is distributed according to an exponential distribution with parameter μ. In the system there is only place for C customers. So the state space $\mathcal{S} = \{0, \ldots, C\}$. If a customer arrives when there are already C customers in the system, he immediately leaves without entering the queue. After uniformization, we get a discrete time Markov chain which is governed by the events e^a with probability $p = \frac{\lambda}{\lambda + \mu}$ and e^d with probability $q = 1 - p$. Event e^a represents an arrival and event e^d represents end of service with departure of the customer.
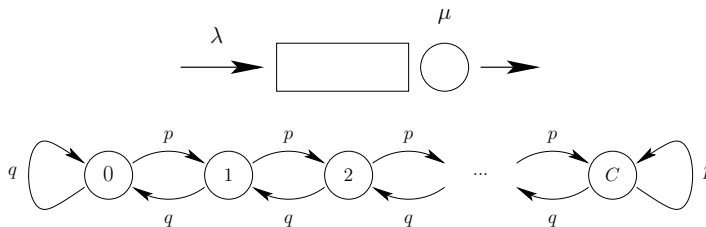

Figure 4.1: *The M/M/1/C queue and the discrete time Markov chain after uniformization*

4.1.1 Exact mean coupling time

The construction of an exact bound on the backward coupling time is based on the next proposition:

Proposition 4.1. *In an M/M/1/C queue we have* $\mathbb{E}\left[\tau^b\right] = \mathbb{E}\left[\min\{h_{0\to C}, h_{C\to 0}\}\right]$

Proof. When applying forward simulation, the chain only can couple in state 0 or state C. This follows since for $r, s \in \mathcal{S}$ with $0 < r < s < C$ we have

$$\phi\left(r, e^a\right) = r + 1 < s + 1 = \phi\left(s, e^a\right),$$

and

$$\phi\left(r, e^d\right) = r - 1 < s - 1 = \phi\left(s, e^d\right).$$

So the chain cannot couple in a state s with $0 < s < C$. Furthermore we have $\phi(C, e^a) = C = \phi(C - 1, e^a)$ and $\phi(0, e^d) = 0 = \phi(1, e^d)$. Hence, forward coupling can only occur in 0 or C. Combining this result with Equation (3.5) leads to $\mathbb{E}\left[\tau^f\right] = \min\{h_{0\to C}, h_{C\to 0}\}$. Since the forward and backward coupling time have the same distribution, it follows that $\mathbb{E}\left[\tau^b\right] = \mathbb{E}\left[\min\{h_{0\to C}, h_{C\to 0}\}\right].$ $\square$

In order to compute $\min\{h_{0\to C}, h_{C\to 0}\}$ we have to run two copies of the Markov chain for a M/M/1/C queue simultaneously. One copy starts in state 0 and the other one starts in state C. We stop when either the chain starting in 0 reaches state C or when the copy starting in state C reaches state 0.

Therefore, let us rather consider a product Markov chain $X(q)$ with state space $\mathcal{S} \times \mathcal{S} = \{(x, y), x = 0, \ldots, C, y = 0, \ldots, C\}$. The Markov chain $X(q)$ is also governed by the two events e^a and e^d and the transition function ψ is:

$$
\begin{aligned}
\psi\left((x, y), e^a\right) &= ((x + 1) \wedge C, (y + 1) \wedge C), \\
\psi\left((x, y), e^d\right) &= ((x - 1) \vee 0, (y - 1) \vee 0).
\end{aligned}
$$

Without any loss of generality, we may assume that $x \leqslant y$. This system corresponds with the *pyramid Markov chain* $X(q)$ displayed in Figure 4.2. Now, running our two copies corresponds with running the product Markov chain starting in state $(0, C)$.

The rest of this section is devoted to establishing a formula for the expected time to reach the base of the pyramid. Although the technique used here (one step analysis) is rather classical, it is interesting to notice how this is related to random walks on the line. This relationship also explains the shifted indices associated to the levels of the pyramid.

Definition 4.1. A state (i, j) of the product Markov chain belongs to level m if $|j - i| = C + 2 - m$.

Then state $(0, C)$ belongs to level 2 and the states $(0, 0)$ and (C, C) belong to level $C + 2$. In Figure 4.2 we see that there are $C + 1$ levels in total. Because of monotonicity of ψ, the level index cannot decrease. Hence, starting at an arbitrary level, the chain will gradually pass all intermediate levels to reach finally level $C + 2$ in state $(0, 0)$ or (C, C). Thus, when starting in state $(0, C)$, the chain will run through all levels to end up at the level $C + 2$. This is also clear from Figure 4.2 and is in accordance with Proposition 4.1.

Definition 4.2. Let $H_{i,j}$ denote the number of steps it takes the product chain starting in (i, j) to reach either state $(0, 0)$ or state (C, C) with $(i, j) \in \mathcal{S} \times \mathcal{S}$.
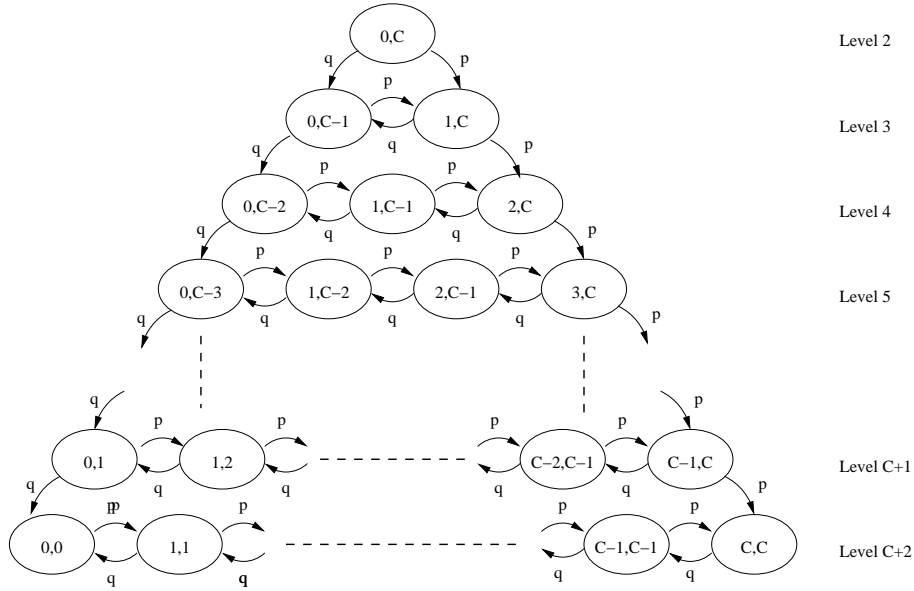
Figure 4.2: *Markov chain $X(q)$ corresponding to $H_{i,j}$*

Thus $H_{i,j}$ represents the hitting time of the coupling states $(0,0)$ and (C,C) (also called absorption time) in a product Markov chain. By definition,

$$H_{0,C} = \min\{h_{0 \to C}, h_{C \to 0}\}.$$

Using a one step analysis, we get the following system of equations for $\mathbb{E}[H_{i,j}]$:

$$\begin{cases} \mathbb{E}[H_{i,j}] &= 1 + p\mathbb{E}[H_{(i+1)\wedge C,(j+1)\wedge C}] + q\mathbb{E}[H_{(i-1)\vee 0,(j-1)\vee 0}] & i \neq j, \\ \mathbb{E}[H_{i,j}] &= 0 & i = j. \end{cases} \quad (4.1)$$

To determine $\mathbb{E}[H_{0,C}]$ we determine the mean time spent on each level and sum up over all levels. Let T_m denote time it takes to reach level $m+1$, starting in level m. Then

$$H_{0,C} = \sum_{m=2}^{C+1} T_m. \quad (4.2)$$

In order to determine T_m, we associate to each level m a random walk R_m on $0, \ldots, m$ with absorbing barriers at state 0 and state m (see Figure 4.3). In the random walk, the probability of going up is p and of going down is $q = 1 - p$. We have the following correspondence between the states of the random walk R_m and the states of $X(q)$:

State 0 of R_m corresponds with state $(0, C - m + 1)$ of $X(q)$,
State i of R_m corresponds with state $(i - 1, C - m + 1 + i)$ of $X(q)$,
$$0 < i < m,$$
State m of R_m corresponds with state $(m - 1, C)$ of $X(q)$.

Now the time spent on level m in $X(q)$ is the same as the time spent in a random walk R_m before absorption. Therefore, in determining T_m, one can use the two following results on random walks, which are also known as ruin problems (see Appendix B).
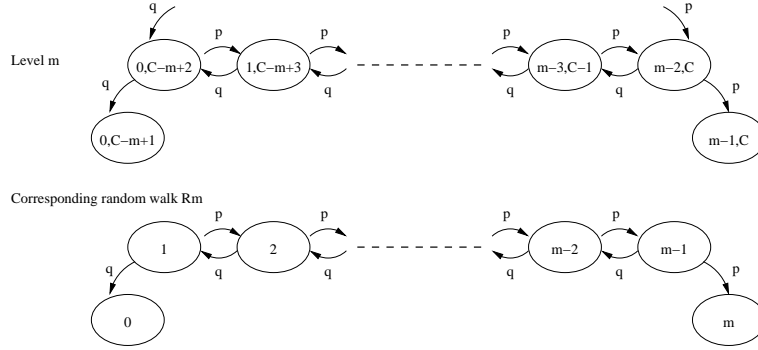
24

Figure 4.3: *Relationship between level m and random walk R_m.*

Lemma 4.1. *Let $\alpha_{i\to 0}^m$ denote the probability of absorption in state 0 of the random walk R_m starting in i. Then:*

$$\alpha_{i\to 0}^m = \begin{cases} \frac{a^m - a^i}{a^m - 1}, & p \neq \frac{1}{2}, \\[2mm] \frac{m-i}{m}, & p = \frac{1}{2}, \end{cases} \tag{4.3}$$

where $a = q/p$.

Lemma 4.2. *Let $\widetilde{T}_i^m$ denote the absorption time of a random walk R_m starting in i. Then:*

$$\mathbb{E}[\widetilde{T}_i^m] = \begin{cases} \frac{i - m(1 - \alpha_{i\to 0}^m)}{q - p}, & p \neq \frac{1}{2}, \\[2mm] i(m-i), & p = \frac{1}{2}. \end{cases} \tag{4.4}$$

Let β_0^m (respectively β_m^m) denote the probability that absorption occurs in 0 (respectively m) in R_m. Then

$$\beta_0^m = \sum_{i=0}^m \alpha_{i\to 0}^m \mathbb{P}\left(R_m \text{ starts in state } i\right), \quad 2 \leqslant m \leqslant C+1 \tag{4.5}$$

and $\beta_m^m = 1 - \beta_0^m$. From the structure of the Markov chain $X(q)$ and the correspondence between $X(q)$ and the random walks, we derive that (see Figure 4.3):

$$\mathbb{P}\left(\text{enter level } m+1 \text{ at } (0, C - m + 1)\right) = \mathbb{P}\left(\text{absorption in 0 in } R_m\right) = \beta_0^m.$$

Now one has:

$$\begin{aligned} \mathbb{E}\left[T_m\right] &= \mathbb{E}[\widetilde{T}_1^m]\beta_0^{m-1} + \mathbb{E}[\widetilde{T}_{m-1}^m]\beta_{m-1}^{m-1} \\ &= \mathbb{E}[\widetilde{T}_1^m]\beta_0^{m-1} + \mathbb{E}[\widetilde{T}_{m-1}^m]\left(1 - \beta_0^{m-1}\right) \\ &= \mathbb{E}[\widetilde{T}_{m-1}^m] + \left(\mathbb{E}[\widetilde{T}_1^m] - \mathbb{E}[\widetilde{T}_{m-1}^m]\right)\beta_0^{m-1}. \end{aligned} \tag{4.6}$$

To complete the calculation, we need to make a distinction between the case with $p = 1/2$ and the case when $p \neq 1/2$. We will first examine the case $p = 1/2$.

Case p = 1/2

$\mathbb{E}[T_m]$ can be calculated explicitly for $p = \frac{1}{2}$. Since the random walk is symmetric, we have $\beta_0^m = \beta_m^m = \frac{1}{2}$. So:

$$\mathbb{E}\left[T_m\right] = \mathbb{E}[\widetilde{T}_1^m]\beta_0^{m-1} + \mathbb{E}[\widetilde{T}_{m-1}^m]\beta_{m-1}^{m-1} = (m-1)\frac{1}{2} + (m-1)\frac{1}{2} = m - 1. \quad (4.7)$$

Hence,

$$\mathbb{E}\left[H_{0,C}\right] = \sum_{m=2}^{C+1}\mathbb{E}\left[T_m\right] = \sum_{m=2}^{C+1}(m-1) = \frac{C^2+C}{2}.$$

Thus we have shown the next result:

Proposition 4.2. *Consider an M/M/1/C queue where the arrival rate λ equals the service rate μ. Then $\mathbb{E}\left[\tau^b\right] = \frac{C^2+C}{2}$.*

Case p $\neq$ 1/2

Since the random walks are not symmetric, we cannot apply the same reasoning as for the case $p = \frac{1}{2}$ to compute β_0^m. Therefore, we will derive a recurrence relation for β_0^m. Entering the random walk R_m corresponds to entering level m in $X(q)$. Since we can only enter level m in the states $(0, C - m + 2)$ and $(m - 2, C)$ this means we can only start the random walk in state 1 or $m - 1$. Therefore (4.5) becomes:

$$\begin{aligned}
\beta_0^m &= \sum_{i=0}^{m}\alpha_{i\to 0}^m\mathbb{P}\left(R_m \text{ starts in state } i\right) \\
&= \alpha_{1\to 0}^m\mathbb{P}\left(R_m \text{ starts in } 1\right) + \alpha_{m-1\to 0}^m\mathbb{P}\left(R_m \text{ starts in } m - 1\right) \\
&= \alpha_{1\to 0}^m\beta_0^{m-1} + \alpha_{m-1\to 0}^m\left(1 - \beta_0^{m-1}\right) \\
&= \alpha_{m-1\to 0}^m + \left(\alpha_{1\to 0}^m - \alpha_{m-1\to 0}^m\right)\beta_0^{m-1} \\
&= \frac{a^m - a^{m-1}}{a^m - 1} + \frac{a^{m-1} - a}{a^m - 1}\beta_0^{m-1}.
\end{aligned}$$

One can easily see that $\beta_0^2 = q$, since the random walk on $0, 1, 2$ starts in state 1 and the first step immediately leads to absorption. This gives the recurrence:

$$\left\{\begin{array}{rcll}
\beta_0^m &=& \frac{a^m - a^{m-1}}{a^m - 1} + \frac{a^{m-1} - a}{a^m - 1}\beta_0^{m-1} & m > 2, \\
\beta_0^2 &=& q.
\end{array}\right. \quad (4.8)$$

Thus we obtain,

Proposition 4.3. *For a M/M/1/C queue, using the foregoing notations,*

$$\mathbb{E}\left[\tau^b\right] = \mathbb{E}\left[H_{0,C}\right] = \sum_{m=2}^{C+1}\left(\mathbb{E}[\widetilde{T}_{m-1}^m] + \left(\mathbb{E}[\widetilde{T}_1^m] - \mathbb{E}[\widetilde{T}_{m-1}^m]\right)\beta_0^{m-1}\right), \quad (4.9)$$

with β_0^m defined by (4.8) and $\mathbb{E}[\widetilde{T}_{m-1}^m]$ and $\mathbb{E}[\widetilde{T}_1^m]$ defined by (4.4).

Thus now Proposition 4.2 and 4.3 deliver expressions which are amenable to explicit calculations. The next proposition says that the case with $p = q$ is a worst case scenario for the coupling time. Intuitively this might be clear, since then one is not driven to the left or the right side of the pyramid of Figure 4.2.

Proposition 4.4. *The coupling time in an M/M/1/C queue is maximal when the input rate λ and the service rate μ are equal.*

Before proving the Proposition, we first proof the following Lemma:

Lemma 4.3. *For $0 \leqslant p < 1/2$ and $1 < C \in \mathbb{R}$ we have,*

$$\frac{(1-p)^C - p^C}{(1-p)^{C+1} - p^{C+1}} < \frac{2C}{C+1}.$$

Proof. (Lemma). Let $f(p) = \frac{(1-p)^C - p^C}{(1-p)^{C+1} - p^{C+1}}$. Observe that $f(0) = 1$ and that

$$
\begin{aligned}
\lim_{p \uparrow \frac{1}{2}} f(p) &= \lim_{p \uparrow \frac{1}{2}} \frac{(1-p)^C}{(1-p)^{C+1}} \cdot \frac{1 - \left(\frac{p}{1-p}\right)^C}{1 - \left(\frac{p}{1-p}\right)^{C+1}} \\
&= \lim_{p \uparrow \frac{1}{2}} \frac{1}{1-p} \cdot \frac{1 + \cdots + \left(\frac{p}{1-p}\right)^{C-1}}{1 + \cdots + \left(\frac{p}{1-p}\right)^{C}} \\
&= \frac{2C}{C+1}.
\end{aligned}
$$

It suffices to show that f has no maximum is the interval $(0, 1/2)$. Therefore, we take a look at the derivative of f:

$$
\begin{aligned}
f'(p) &= \frac{(1-p)^{2C} - p^{2C} + C(1-p)^C p^C \left(\frac{p}{1-p} - \frac{1-p}{p}\right)}{\left((1-p)^{C+1} - p^{C+1}\right)^2} \\
&= \frac{(1-p)^{2C} - p^{2C} + C(2p-1)(1-p)^{C-1} p^{C-1}}{\left((1-p)^{C+1} - p^{C+1}\right)^2}.
\end{aligned}
$$

If there is a maximum in $[0, 1/2)$, then $f'(p) = 0$ for some $p \in (0, 1/2)$. Therefore, take a look at the numerator. If $f' = 0$, then the numerator must be 0 as well. Call the numerator $n(p) = (1-p)^{2C} - p^{2C} + (2p-1)C(1-p)^{C-1} p^{C-1}$, and note that $n(0) = 1$ and $n(1/2) = 0$. Now,

$$
\begin{aligned}
n'(p) &= -2C(1-p)^{2C-1} - 2Cp^{2C-1} - C(C-1)(1-p)^{C-2} p^{C-1}(2p-1) \\
&\quad + C(C-1)(1-p)^{C-1} p^{C-2}(2p-1) + 2C(1-p)^{C-1} p^{C-1}.
\end{aligned}
$$

Observe that for $0 < p < 1/2$,

$$-2C(1-p)^{2C-1} + 2C(1-p)^{C-1} p^{C-1} = 2C(1-p)^{C-1}\left(-(1-p)^C + p^{C-1}\right) < 0,$$

and that also

$$
\begin{aligned}
-C(C-1)(1-p)^{C-2} p^{C-1}(2p-1) + C(C-1)(1-p)^{C-1} p^{C-2}(2p-1) &= \\
(2p-1)C(C-1)(1-p)^{C-2} p^{C-2}((1-p) - p) &< 0
\end{aligned}
$$

Since also $-2Cp^{2C-1} < 0$, it follows that $n'(p) < 0$ for $0 < p < 1/2$. Thus n is decreasing on $(0, 1/2)$ and thus n has no roots in the interval $(0, 1/2)$. Hence, f has no maximum in $(0, 1/2)$ and thus $f(x) < 2C/(C+1)$ for $x \in [0, 1/2)$. $\qquad\square$

Proof. (Proposition 4.4). By definition, $\lambda = \mu$ corresponds to $p = q = 1/2$. The proof holds by induction on C. The result is obviously true when $C = 0$, because whatever q, $\mathbb{E}[H_{0,C}] = 0$.

For $C + 1$, let q be an arbitrary probability with $q > 1/2$ (the case $q < 1/2$ is symmetric). We will compare the expected time for absorption of three Markov
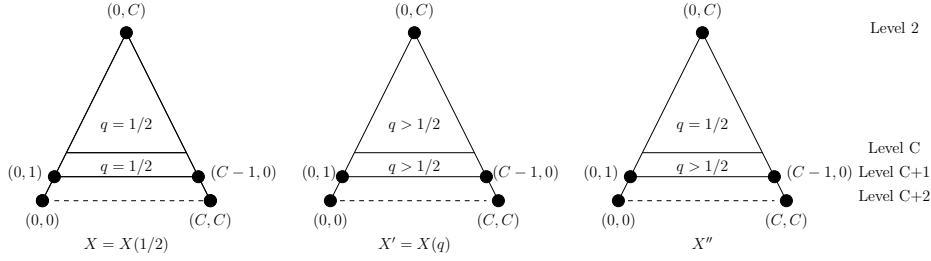
Figure 4.4: *The three different Markov chains X, X' and X''.*

chains. The first one is the Markov chain $X := X(1/2)$ displayed in Figure 4.2, with $q = p = 1/2$. The second one is the Markov chain $X' = X(q)$ displayed in Figure 4.2 and the last one X'' is a mixture between the two previous chains: The first C levels are the same as in X while the last level $(C+1)$ is the same as in X'.

The expected absorption time for the first C levels is the same for X and for X'':

$$\sum_{m=2}^{C} \mathbb{E}T_m = \sum_{m=2}^{C} \mathbb{E}T''_m.$$

By induction, this is larger than for X': we have

$$\sum_{m=2}^{C} \mathbb{E}T_m = \sum_{m=2}^{C} \mathbb{E}T''_m \geqslant \sum_{m=2}^{C} \mathbb{E}T'_m.$$

Therefore, we just need to compare the exit times out of the last level, namely $\mathbb{E}T_{C+1}, \mathbb{E}T'_{C+1}$ and $\mathbb{E}T''_{C+1}$, to finish the proof.

We first compare $\mathbb{E}T_{C+1}$ and $\mathbb{E}T''_{C+1}$. In both cases, the Markov chain enters level $C+1$ in state $(0,1)$ with probability $1/2$.

Equation (4.7) says that $\mathbb{E}T_{C+1} = C$ and Equation (4.4) gives after straightforward computations,

$$\mathbb{E}T''_{C+1} = 1/2 \frac{1 - (C+1)(1 - \alpha_{1 \to 0}^{C+1})}{q - p} + 1/2 \frac{C - (C+1)(1 - \alpha_{C \to 0}^{C+1})}{q - p} \quad (4.10)$$

$$= \frac{C+1}{2(q-p)} \frac{(a^C - 1)(a - 1)}{a^{C+1} - 1} = \frac{C+1}{2} \frac{q^C - p^C}{q^{C+1} - p^{C+1}}. \quad (4.11)$$

It follows from Lemma 4.3 that $\mathbb{E}T''_{C+1} < \frac{C+1}{2} \cdot \frac{2C}{C+1} = C = \mathbb{E}T_{C+1}$. In order to compare $\mathbb{E}T'_{C+1}$ and $\mathbb{E}T''_{C+1}$, let us first show that β_0^m is at least $1/2$, for all $m \geqslant 2$. This is done by an immediate induction on Equation (4.8). If $\beta_0^{m-1} \geqslant 1/2$, then

$$\beta_0^m \geqslant \frac{2a^m - a^{m-1} - a}{2(a^m - 1)}.$$

Now,

$$\frac{2a^m - a^{m-1} - a}{2(a^m - 1)} \geqslant 1/2 \qquad \text{if} \qquad 2a^m - a^{m-1} - a \geqslant a^m - 1,$$

i.e. after recombining the terms, $(a - 1)(a^{m-1} - 1) \geqslant 0$. This is true as soon as $a \geqslant 1$, i.e. as soon as $q \geqslant 1/2$.

To end the proof, it is enough to notice that for the chain X', the expected time to absorption starting in 1, $\mathbb{E}\tilde{T}_1^{m'}$ is smaller than or equal to the expected time to

absorption starting in $m-1$, $\mathbb{E}\tilde{T}^{m'}_{m-1}$ for all m. The difference $\mathbb{E}\tilde{T}^{m'}_{m-1} - \mathbb{E}\tilde{T}^{m'}_{1}$ is

$$\mathbb{E}\tilde{T}^{m'}_{m-1} - \mathbb{E}\tilde{T}^{m'}_{1} = \frac{m-1-m\left(1-\alpha^m_{m-1\to 0}\right)}{q-p} - \frac{1-m\left(1-\alpha^m_{1\to 0}\right)}{q-p} \quad (4.12)$$

$$= \frac{m-2+m\frac{a^m-a^{m-1}}{a^m-1}-m\frac{a^m-a}{a^m-1}}{p\left(a-1\right)} \quad (4.13)$$

$$= \frac{a^m-1}{a^m-1}\cdot\frac{m-2+m\frac{a^m-a^{m-1}}{a^m-1}-m\frac{a^m-a}{a^m-1}}{p\left(a-1\right)} \quad (4.14)$$

$$= \frac{ma^m-ma^{m-1}+ma-m+2a^m+2}{p\left(a^m-1\right)\left(a-1\right)} \quad (4.15)$$

$$= \frac{2m(a-1)\left(\frac{a^{m-1}+1}{2}-\frac{1+a+\cdots+a^{m-1}}{m}\right)}{p\left(a^m-1\right)\left(a-1\right)}. \quad (4.16)$$

$$\quad (4.17)$$

By convexity of $x \mapsto a^x$, we obtain

$$\mathbb{E}\tilde{T}^{m'}_{m-1} - \mathbb{E}\tilde{T}^{m'}_{1} \geqslant 0. \quad (4.18)$$

Thus by setting $m = C+1$ we have $\mathbb{E}\tilde{T}^{C+1'}_{C} \geqslant \mathbb{E}\tilde{T}^{C+1'}_{1}$. Furthermore, note that the random walks associated with level $C+1$ in X' and X'' are the same. Thus $\mathbb{E}\tilde{T}^{C+1'}_{C} = \mathbb{E}\tilde{T}^{C+1''}_{C}$ and $\mathbb{E}\tilde{T}^{C+1'}_{1} = \mathbb{E}\tilde{T}^{C+1''}_{1}$ Combining these observations with (4.6) finally yields:

$$\mathbb{E}T'_{C+1} = \mathbb{E}\tilde{T}^{C+1'}_{C} + \left(\mathbb{E}\tilde{T}^{C+1'}_{1} - \mathbb{E}\tilde{T}^{C+1'}_{C}\right)\beta^C_0 \quad (4.19)$$

$$\leqslant \mathbb{E}\tilde{T}^{C+1'}_{C} \quad (4.20)$$

$$\leqslant \frac{1}{2}\mathbb{E}\tilde{T}^{C+1'}_{C} + \frac{1}{2}\mathbb{E}\tilde{T}^{C+1'}_{1} \quad (4.21)$$

$$= \frac{1}{2}\mathbb{E}\tilde{T}^{C+1''}_{C} + \frac{1}{2}\mathbb{E}\tilde{T}^{C+1''}_{1} \quad (4.22)$$

$$= \mathbb{E}T''_{C+1}. \quad (4.23)$$

Thus we have shown that $\mathbb{E}T'_{C+1} \leqslant \mathbb{E}T''_{C+1} \leqslant \mathbb{E}T_{C+1}$. $\qquad\square$

4.1.2 Explicit bounds

Equation (4.9) provides a quick way to compute the expected backward coupling time $\mathbb{E}\left[\tau^b\right]$ using recurrence equation (4.8). However, it may also be interesting to get a simple closed form for an upper bound for $\mathbb{E}\left[\tau^b\right]$. This can be done using the last inequality in Equation (3.5) that gives an upper bound for $\mathbb{E}\left[\tau^b\right]$ amenable to direct computations:

$$\mathbb{E}\left[\tau^b\right] = \mathbb{E}\left[\min\left\{h_{0\to C}, h_{C\to 0}\right\}\right] \leqslant \min\left\{\mathbb{E}\left[h_{0\to C}\right], \mathbb{E}\left[h_{C\to 0}\right]\right\}. \quad (4.24)$$

Let T_i denote the time to go from state i to $i+1$. Then

$$\mathbb{E}\left[h_{0\to C}\right] = \sum_{i=0}^{C-1}\mathbb{E}\left[T_i\right]. \quad (4.25)$$

To get an expression for $\mathbb{E}\left[T_i\right]$, we condition on the first event. Therefore let $\mathbb{E}\left[T_i|e\right]$ denote the conditional expectation of T_i knowing that the next event is e. Since

$\mathbb{E}[T_i \mid e^a] = 1$ and $\mathbb{E}[T_i \mid e^d] = 1 + \mathbb{E}[T_{i-1}] + \mathbb{E}[T_i]$, conditioning delivers the following recurrent expression for the $\mathbb{E}[T_i]$:

$$
\begin{aligned}
\mathbb{E}[T_i] &= \mathbb{E}[T_i \mid e^d]\mathbb{P}(e^d) + \mathbb{E}[T_i \mid e^a]\mathbb{P}(e^a) \\
&= (1 + \mathbb{E}[T_{i-1}] + \mathbb{E}[T_i])q + p.
\end{aligned}
$$
(4.26)

Solving for $\mathbb{E}[T_i]$ yields

$$
\mathbb{E}[T_i] = \begin{cases} \frac{1}{p} + \frac{q}{p}\mathbb{E}[T_{i-1}] & \text{for } 0 < i < C, \\ \frac{1}{p} & \text{for } i = 0. \end{cases}
$$
(4.27)

By induction one can show that $\mathbb{E}[T_i] = \frac{1}{p}\sum_{k=0}^{i}\left(\frac{q}{p}\right)^k$. Again, we need to distinguish the case $p \neq q$ from the case $p = q$.

Case $p \neq q$

Then $\mathbb{E}[T_i] = \frac{1}{p}\sum_{k=0}^{i}\left(\frac{q}{p}\right)^k = \frac{1-\left(\frac{q}{p}\right)^{i+1}}{p-q}$ and from (4.25) it follows that

$$
\mathbb{E}[h_{0\to C}] = \sum_{i=0}^{C-1}\frac{1-\left(\frac{q}{p}\right)^{i+1}}{p-q} = \frac{C}{p-q} - \frac{q\left(1-\left(\frac{q}{p}\right)^C\right)}{(p-q)^2}.
$$
(4.28)

By reasons of symmetry, we have

$$
\mathbb{E}[h_{C\to 0}] = \frac{C}{q-p} - \frac{p\left(1-\left(\frac{p}{q}\right)^C\right)}{(q-p)^2}.
$$
(4.29)

Case $p = q$

Now $\mathbb{E}[T_i] = \frac{1}{p}\sum_{k=0}^{i}\left(\frac{q}{p}\right)^k = 2(i+1)$, and from (4.25) it follows that

$$
\mathbb{E}[h_{0\to C}] = \sum_{i=0}^{C-1} 2(i+1) = C^2 + C.
$$
(4.30)

By symmetry, also $\mathbb{E}[h_{C\to 0}] = C^2 + C$.

If $p > q$, then $\mathbb{E}[h_{0\to C}] < \mathbb{E}[h_{C\to 0}]$ and because of symmetry, if $p < q$ then $\mathbb{E}[h_{0\to C}] > \mathbb{E}[h_{C\to 0}]$. Since $\left(C^2 + C\right)/2$ is an upper bound corresponding to the critical case $p = q$ on the mean coupling time $\mathbb{E}[\tau^b]$, as shown in Proposition 4.4, one can state:

Proposition 4.5. *The mean coupling time $\mathbb{E}[\tau^b]$ of a M/M/1/C queue with arrival rate λ and service rate μ is bounded using $p = \lambda/(\lambda + \mu)$ and $q = 1 - p$.*
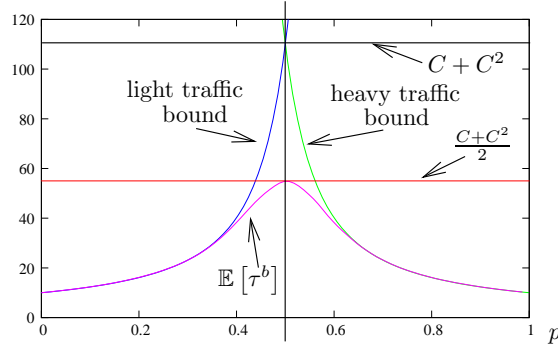*Critical bound:*

$$
\text{for every } p \in [0,1], \quad \mathbb{E}[\tau^b] \leqslant \frac{C^2 + C}{2}.
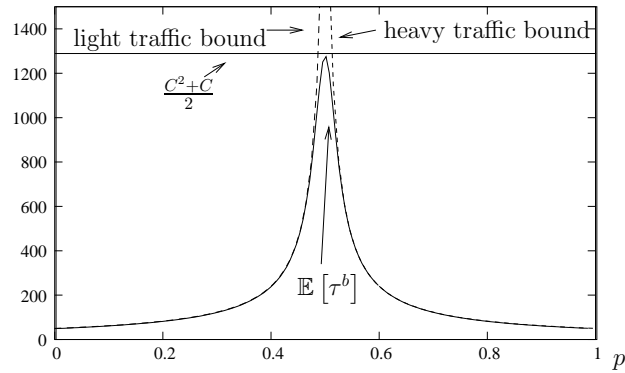$$

Heavy traffic Bound:

$$
\text{if } p > \frac{1}{2}, \quad \mathbb{E}[\tau^b] \leqslant \frac{C}{p-q} - \frac{q\left(1-\left(\frac{q}{p}\right)^C\right)}{(p-q)^2}.
$$

Light traffic bound:

$$if\ p < \frac{1}{2}, \quad \mathbb{E}\left[\tau^b\right] \leqslant \frac{C}{q-p} - \frac{p\left(1 - \left(\frac{p}{q}\right)^C\right)}{(q-p)^2}.$$

(a) *M/M/1/10 queue.*

(b) *M/M/1/50 queue.*

Figure 4.5: *Expected coupling time in an M/M/1/10 and an M/M/1/50 queue when p varies from 0 to 1 and the three explicit bounds given in Proposition 4.5*

Figure 4.5 displays both the exact expected coupling time as given by Equation (4.9) as well as the three explicit bounds given in Proposition 4.5 for a queue with capacity 10 and a queue with capacity 50. Note that the bounds for the M/M/1/10 queue are very accurate under light or heavy traffic ($q < 0.4$ and $q > 0.6$). Then, the ratio is never larger than 1.2. For the M/M/1/50, we see that the discrepancy between the bounds and the real coupling time is even smaller.

Remark 4.1. Note that also the recurrence relation:

$$\mathbb{E}\left[h_{i\to 0}\right] = 1 + p\mathbb{E}\left[h_{(i+1)\wedge C\to 0}\right] + q\mathbb{E}\left[h_{(i-1)\vee 0\to 0}\right]. \tag{4.31}$$

holds for $\mathbb{E}\left[h_{i\to 0}\right]$. Setting $i = C$ and solving leads to the light traffic bound.

4.2 Formal series approach

Another approach to gain understanding of the coupling time of a single queue is a formal series approach. Consider both an $M/M/1$ queue with infinite capacity and an $M/M/1/C$ queue with finite capacity equal to C. Both queues have the same arrival rate λ and service rate μ and suppose for reasons of stability that $\lambda < \mu$. Denote the underlying uniformized discrete Markov chain of the infinite capacity queue by $\{X_n\}_{n \in \mathbb{N}}$ and the underlying chain of the finite capacity queue by $\{X_n^C\}_{n \in \mathbb{N}}$. Let $p = \frac{\lambda}{\lambda + \mu}$ denote the probability of an arrival, and $q = 1 - p$ denote the probability of a departure. Define the hitting time of state 0 as

$$h_0^C = \inf_{\mathbb{N}} \{n : X_n^C = 0\}$$

for X_n^C and as

$$h_0 = \inf_{\mathbb{N}} \{n : X_n = 0\}$$

for X_n. Since the finite capacity queue does not accept arrivals when there are already C customers in the queue, we have $X_n^C \leqslant_{st} X_n$. This implies that h_0^C is stochastically bounded by h_0. In section 4.1.2 we have seen that conditioned on starting in C, $\mathbb{E}\left[h_0^C\right]$ is a rather good bound on the backward coupling time. In this section, we will focus on the conditional distribution of h_0.

Define the formal series of the conditional distribution of the hitting time h_0 as:

$$G(z, x) = \sum_{k=0}^{\infty} \sum_{i=0}^{\infty} z^k x^i \mathbb{P}_i (h_0 = k)$$

with $\mathbb{P}_i (h_0 = k) = \mathbb{P}(h_0 = k | X_0 = i)$.

Our main goal is to obtain a closed expression for G, which can be used for determining the moments of h_0. First we will investigate the structure of $\mathbb{P}_i (h_0 = k)$, since we will use $\mathbb{P}_i (h_0 = k)$ in the computations to deduce a closed form for G. We distinguish five distinctive cases:

1. **Case $i = 0$ and $k > 0$**
   By definition, if there are 0 customers at time 0, the hitting time h_0 is equal to 0. Hence,
   $$\mathbb{P}_i (h_0 = k) = 0.$$

2. **Case $k < i$**
   When there are i customers, the fastest way to reach the state with 0 customers is by i consecutive departures. But we are only allowed to make $k < i$ steps. Hence,
   $$\mathbb{P}_i (h_0 = k) = 0.$$

3. **Case $i = k$**
   In order to reach state 0, there must be $k = i$ consecutive departures. Since the probability of a departure equals q, it follows that
   $$\mathbb{P}_i (h_0 = k) = q^k.$$

4. **Case $i < k$ and $k - i$ uneven**
   In order to reach state 0, there must be for sure i departures. Then there rest $k - i$ steps to take. These $k - i$ events must consist of exactly the same number of arrivals as departures. But since $k - i$ is uneven, this is impossible. Therefore:
   $$\mathbb{P}_i (h_0 = k) = 0.$$

5. **Case $i < k$ and $k - i$ even**

   The same reasoning as in case 4 applies. Since $k - i$ is even this time, we have:

   $$\mathbb{P}_i\left(h_0 = k\right) = q^i q^{\frac{k-i}{2}} p^{\frac{k-i}{2}} W\left(i, k\right),$$

   with $W\left(i, k\right)$ the number of walks starting at i with $h_0 = k$. We will call such a walk an admissible walk. In Figure 4.6 we see all the admissible walks for $i = 1$ and $k = 7$. Remark that all the walks in Figure 4.6 end by a departure. Every admissible walk ends with a departure. Since if it does not, then it ends with an arrival and thus $X_{k-1} = 0$. Then the hitting time is smaller than k which is in contradiction with the definition of k.
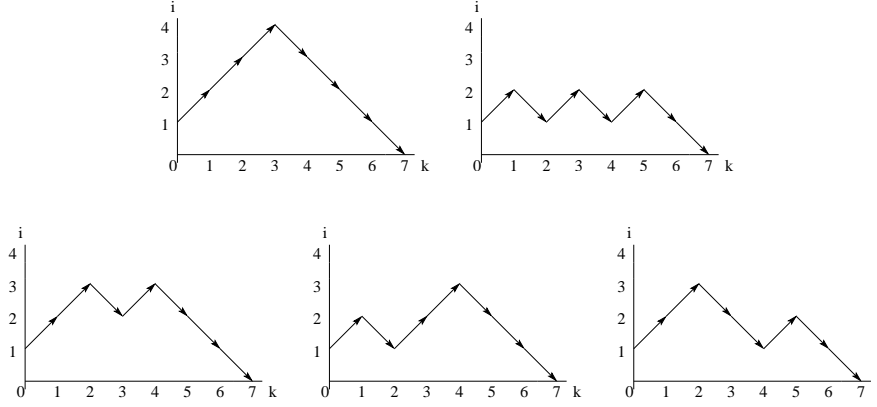

Figure 4.6: *All admissible walks for $i = 1$ and $k = 7$.*

Now we will derive an explicit formula for $W(1, k)$. Since the last step is fixed, we can neglect it in counting all possible walks $W(1, k)$. So, to each admissible walk of length k corresponds a walk of length $k - 1$ by leaving the last step away (see Figure 4.7). This corresponding walk of length $k - 1$ is known to be a Dyck path.

Since $k - 1$ is even, there exists an $n \in \mathbb{N}$ such that $k - 1 = 2n$. Let D_n be the number of Dyck paths of length $2n = k - 1$. By conditioning on the first return to 0, we can derive a recurrence for the D_n. The first return can happen in $2i$ for $1 \leqslant i \leqslant n$. Then the Dyck path is split into two shorter paths, the first of length $2i$ and the other of length $2(n - i)$. Note that the first path, before the return to 0, is composed of a step up, a Dyck path of length $2i - 2$ and a step down. Thus the number of Dyck paths D_n with the first return in $2i$ is $D_{i-1}D_{n-i}$. Since the returns can occur in every $2i$ with $1 \leqslant i \leqslant n$, we obtain

$$D_n = \sum_{i=1}^{n} D_{i-1}D_{n-i} = \sum_{i=0}^{n-1} D_i D_{n-i-1}, \tag{4.32}$$

with initial condition $D_0 = 1$. This reccurence is exactly the recurrence of the Catalan numbers. Hence, the number of Dyck paths of length $k - 1$ is the $k - 1$th Catalan number denoted by C_{k-1} ([8], p. 357-358). Consequently we have:

$$W\left(1, k\right) = C_{k-1} = \binom{k - 1}{(k-1)/2} \frac{1}{\frac{k-1}{2} + 1}. \tag{4.33}$$

33

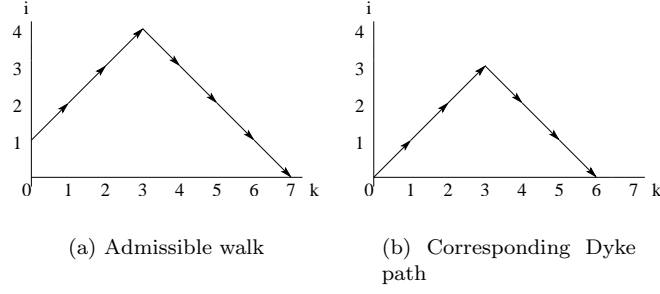(a) Admissible walk   (b) Corresponding Dyke path

Figure 4.7: *To an admissible walk corresponds a Dyck path.*

Thus we can state:

$$
\mathbb{P}_i\left(h_0 = k\right) = \begin{cases}
q^{\frac{k+i}{2}} p^{\frac{k-i}{2}} \binom{k-1}{(k-1)/2} \frac{1}{\frac{k-1}{2}+1} & \text{if } i = 1, k > 1 \text{ and } k - 1 \text{ even,} \\
q^{\frac{k+i}{2}} p^{\frac{k-i}{2}} W(i,k) & \text{if } 1 < i < k \text{ and } k - i \text{ even,} \\
q^k & \text{if } i = k, \\
0 & \text{otherwise.}
\end{cases}
$$

$$(4.34)$$

Now we will develop the formal series $G\left(z, x\right)$ to obtain a closed form. First note that we can split the summation and rewrite the series as:

$$
G\left(z, x\right) = \sum_{k=1}^{\infty} \sum_{i=1}^{\infty} z^k x^i \mathbb{P}_i\left(h_0 = k\right) + \sum_{k=0}^{\infty} z^k \mathbb{P}_0\left(h_0 = k\right) + \sum_{i=1}^{\infty} x^i \mathbb{P}_i\left(h_0 = 0\right).
$$

From (4.34) we derive that

$$
\sum_{k=0}^{\infty} z^k \mathbb{P}_0\left(h_0 = k\right) = 1 \quad \text{and} \quad \sum_{i=1}^{\infty} x^i \mathbb{P}_i\left(h_0 = 0\right) = 0.
$$

It follows that

$$
G\left(z, x\right) = 1 + \sum_{k=1}^{\infty} \sum_{i=1}^{\infty} z^k x^i \mathbb{P}_i\left(h_0 = k\right).
$$

By conditioning on whether the next event is an arrival or a departure we get:

$$
\begin{aligned}
G\left(z, x\right) &= 1 + \sum_{k=1}^{\infty} \sum_{i=1}^{\infty} z^k x^i q \mathbb{P}_{i-1}\left(h_0 = k - 1\right) + \sum_{k=1}^{\infty} \sum_{i=1}^{\infty} z^k x^i p \mathbb{P}_{i+1}\left(h_0 = k - 1\right) \\
&= 1 + qxz \sum_{k=1}^{\infty} \sum_{i=1}^{\infty} z^{k-1} x^{i-1} \mathbb{P}_{i-1}\left(h_0 = k - 1\right) \\
&\qquad\qquad + \frac{pz}{x} \sum_{k=1}^{\infty} \sum_{i=1}^{\infty} z^{k-1} x^{i+1} \mathbb{P}_{i+1}\left(h_0 = k - 1\right) \\
&= 1 + qxzG\left(z, x\right) + \frac{pz}{x} \sum_{k=1}^{\infty} \sum_{i=1}^{\infty} z^{k-1} x^{i+1} \mathbb{P}_{i+1}\left(h_0 = k - 1\right) \\
&= 1 + qxzG\left(z, x\right) + \frac{pz}{x} \sum_{k=0}^{\infty} \sum_{i=2}^{\infty} z^k x^i \mathbb{P}_i\left(h_0 = k\right). \tag{4.35}
\end{aligned}
$$

We observe that

$$\sum_{k=0}^{\infty}\sum_{i=2}^{\infty}z^k x^i \mathbb{P}_i\left(h_0 = k\right) = G\left(z, x\right) - \sum_{k=0}^{\infty}z^k \mathbb{P}_0\left(h_0 = k\right) - \sum_{k=0}^{\infty}z^k x \mathbb{P}_1\left(h_0 = k\right).$$

Combining this with the results of (4.34) yields:

$$\sum_{k=0}^{\infty}\sum_{i=2}^{\infty}z^k x^i \mathbb{P}_i\left(h_0 = k\right) = G\left(z, x\right) - 1 - x\sum_{k=0}^{\infty}z^k \mathbb{P}_1\left(h_0 = k\right). \tag{4.36}$$

It follows from (4.34) that $\mathbb{P}_1\left(h_0 = k\right)$ is zero for all even k. Therefore, in order to obtain $\sum_{k=0}^{\infty}z^k x\mathbb{P}_1\left(h_0 = k\right)$, we set $k = 2m + 1$ and sum over all m:

$$\begin{aligned}
\sum_{k=0}^{\infty}z^k \mathbb{P}_1\left(h_0 = k\right) &= \sum_{m=0}^{\infty}z^{2m+1}\mathbb{P}_1\left(h_0 = 2m + 1\right) \tag{4.37}\\
&= \sum_{m=0}^{\infty}z^{2m+1}q^{m+1}p^m\binom{2m}{m}\frac{1}{m+1}\\
&= qz\sum_{m=0}^{\infty}\left(z^2 pq\right)^m\binom{2m}{m}\frac{1}{m+1}.
\end{aligned}$$

Since the generating function $\sum_{k=0}^{\infty}C_k z^k$ of the Catalan numbers equals (See Appendix B):

$$\sum_{k=0}^{\infty}C_k z^k = \frac{1 - \sqrt{1 - 4z}}{2z},$$

we obtain:

$$\begin{aligned}
\sum_{k=0}^{\infty}z^k \mathbb{P}_1\left(h_0 = k\right) &= qz\sum_{m=0}^{\infty}\left(z^2 pq\right)^2\binom{2m}{m}\frac{1}{m+1} \tag{4.38}\\
&= qz\frac{1 - \sqrt{1 - 4pqz^2}}{2pqz^2} \tag{4.39}\\
&= \frac{1 - \sqrt{1 - 4pqz^2}}{2pz}. \tag{4.40}
\end{aligned}$$

Using this result together with (4.35) and (4.36) yields:

$$G\left(z, x\right) = 1 + qxzG\left(z, x\right) + \frac{pz}{x}\left(G\left(z, x\right) - 1 - x\frac{1 - \sqrt{1 - 4pqz^2}}{2pz}\right).$$

Solving this last equation for $G\left(z, x\right)$ finally returns the closed form:

$$G\left(z, x\right) = \frac{x - 2pz + x\sqrt{1 - 4pqz^2}}{2x - 2qx^2 z - 2pz}. \tag{4.41}$$

Remark 4.2. The method used above is not the only method to obtain $G\left(z, x\right)$. Another approach is the following. The hitting time h_0, knowing that we start in $X_0 = i$, can be vizualised as a walk starting in state i until we reach state 0. This walk is a concatenation of i independently and identically distributed walks W_j with W_j the walk in the M/M/1/∞ queue, starting in state j until state $j - 1$ is reached (see Figure 4.8).
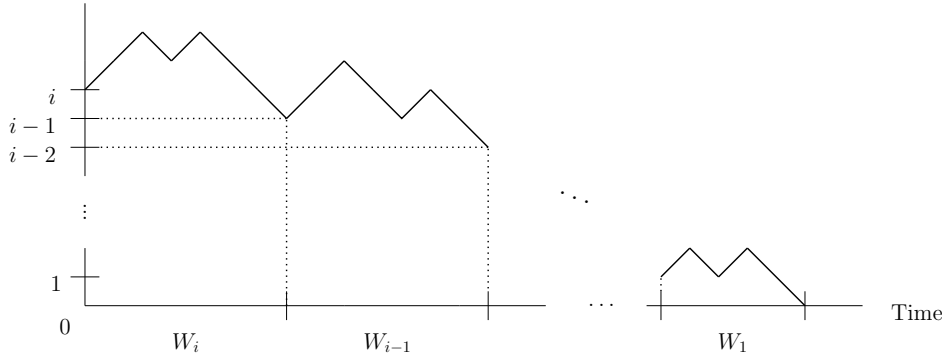
Figure 4.8: *The trajectory from i to 0 in a $M/M/1/\infty$ queue is a concatination of i walks.*

Let the random variable Y_j denote the duration of a walk W_j, i.e. the number of steps it takes to reach state $j - 1$ starting from j. Observe that

$$\mathbb{P}_i\left(h_0 = k\right) = \mathbb{P}\left(\sum_{j=1}^{i} Y_j = k\right),$$

and thus

$$
G(z,x) = \sum_{k=0}^{\infty}\sum_{i=0}^{\infty} z^k x^i \mathbb{P}_i\left(h_0 = k\right), \tag{4.42}
$$

$$
= \sum_{k=0}^{\infty}\sum_{i=0}^{\infty} z^k x^i \mathbb{P}\left(\sum_{j=1}^{i} Y_j = k\right) \tag{4.43}
$$

$$
= \sum_{i=0}^{\infty} x^i \mathbb{E}\left[z^{\sum_{j=1}^{i} Y_j}\right]. \tag{4.44}
$$

The random variables Y_j are independently and identically distributed and therefore we can write:

$$
\mathbb{E}\left[z^{\sum_{j=1}^{i} Y_j}\right] = \mathbb{E}\left[z^{Y_j}\right]^i = G_{Y_j}(z)^i, \tag{4.45}
$$

with $G_{Y_j}(z)$ the generating function of Y_j. Now, combining (4.44) with (4.45) yields:

$$
G(z,x) = \sum_{i=0}^{\infty} x^i G_{Y_j}(z)^i = \frac{1}{1 - x\,G_{Y_j}(z)}. \tag{4.46}
$$

Note that

$$
G_{Y_j}(z) = \sum_{k=0}^{\infty} z^k \mathbb{P}\left(Y_j = k\right) = \frac{1 - \sqrt{1 - 4pqz^2}}{2pz}, \tag{4.47}
$$

where the last equality follows from (4.38) and (4.40). Combining (4.46) with the closed form for $G_{Y_j}(z)$ of (4.47) delivers:

$$
G(z,x) = \frac{2pz}{2pz - x + x\sqrt{1 - 4pqz^2}}. \tag{4.48}
$$

At first sight, this might be surprising, since we already obtained the closed form of (4.41) for $G(z,x)$. However, it is easy algebra to show that

$$
\frac{2pz}{2pz - x + x\sqrt{1 - 4pqz^2}} - \frac{x - 2pz + x\sqrt{1 - 4pqz^2}}{2x - 2qx^2z - 2pz} = 0.
$$

One can use the function $G(z, x)$ to calculate the mean, the variance and the moments of $\mathbb{P}_i(h_0 = k)$. We will illustrate how to obtain the mean and the variance.

The mean

By taking the partial derivative of G with respect to z and then setting $z = 1$ we obtain with the help of Maple software:

$$
\begin{aligned}
\sum_{i=0}^{\infty} x^i \mathbb{E}_i(h_0) &= \left. \frac{\partial G}{\partial z} \right|_{z=1} \\
&= \frac{x}{(1-x)^2 (q-p)},
\end{aligned}
$$

with $\mathbb{E}_i(h_0) = \mathbb{E}(h_0 | X_0 = i)$. Since

$$
\frac{1}{(1-x)^2} = \sum_{i=0}^{\infty} (i+1) x^i, \tag{4.49}
$$

we get for the Taylor series of $x / \left((1-x)^2 (q-p) \right)$ around $x = 0$:

$$
\begin{aligned}
\frac{x}{(1-x)^2 (q-p)} &= \frac{1}{q-p} \left(\sum_{i=0}^{\infty} (i+1) x^{i+1} \right), \\
&= \sum_{i=0}^{\infty} \frac{i}{q-p} x^i.
\end{aligned}
$$

Hence, in a stable M/M/1 queue, we obtain

$$
\mathbb{E}_i[h_0] = \frac{i}{q-p}. \tag{4.50}
$$

By setting $C = i$, we note that this result harmonizes with the bounds of (4.28) and (4.29), since in these bounds the term $\frac{C}{q-p}$ appears.

The variance

In order to obtain the variance of the hitting time h_0, we use the identity

$$
\operatorname{Var}_i[h_0] = \mathbb{E}_i[h_0^2] - \mathbb{E}_i[h_0]^2. \tag{4.51}
$$

It follows from Equation 4.50 that

$$
(\mathbb{E}_i[h_0])^2 = \left(\frac{i}{q-p} \right)^2.
$$

Moreover,

$$
\mathbb{E}_i[h_0^2] = \mathbb{E}_i[h_0^2 - h_0] + \mathbb{E}_i[h_0]. \tag{4.52}
$$

Taking the second partial derivative of G with respect to z and setting $z = 1$ delivers:

$$
\begin{aligned}
\left. \frac{\partial^2 G}{\partial z^2} \right|_{z=1} &= \sum_{k=2}^{\infty} \sum_{i=0}^{\infty} k(k-1) x^i \mathbb{P}_i(h_0 = k) \\
&= \sum_{i=0}^{\infty} x^i \sum_{k=0}^{\infty} k(k-1) \mathbb{P}_i(h_0 = k) \\
&= \sum_{i=0}^{\infty} \mathbb{E}_i[h_0(h_0 - 1)] x^i.
\end{aligned}
$$

37

Thus the i-th coefficient in the Taylor expansion of $\left.\frac{\partial^2 G}{\partial z^2}\right|_{z=1}$ equals $\mathbb{E}_i\left[h_0^2 - h_0\right]$.
The next step is to expand $\left.\frac{\partial^2 G}{\partial z^2}\right|_{z=1}$ as a Taylor series around $x = 0$. With the help
of Maple software we obtain:

$$
\begin{aligned}
\left.\frac{\partial^2 G}{\partial z^2}\right|_{z=1} &= \frac{2\left(4p^2x^2 - 4p^2x - 5px^2 + 3px + x^2\right)}{(q-p)^3 (1-x)^3}, \\
&= \frac{2}{(q-p)^3}\left(\left(4p^2 - 5p + 1\right)\frac{x^2}{(1-x)^3} + \left(3p - 4p^2\right)\frac{x}{(1-x)^3}\right).
\end{aligned}
$$

In developing the series expansion of $\left.\frac{\partial^2 G}{\partial z^2}\right|_{z=1}$ we use the identity

$$
\frac{1}{(1-x)^3} = \sum_{i=0}^{\infty} \frac{i^2 + 3i + 2}{2}x^i.
$$

This delivers:

$$
\begin{aligned}
\left.\frac{\partial^2 G}{\partial z^2}\right|_{z=1} &= \frac{2}{(q-p)^3}\left(\left(4p^2 - 5p + 1\right)\sum_{i=0}^{\infty}\frac{i^2 - i}{2}x^i + \left(3p - 4p^2\right)\sum_{i=0}^{\infty}\frac{i^2 + i}{2}x^i,\right) \\
&= \sum_{i=0}^{\infty}\frac{-8p^2i - 2pi^2 + 8pi + i^2 - i}{(q-p)^3}x^i. \qquad (4.53)
\end{aligned}
$$

By combining (4.52) with (4.50) and (4.53) we get:

$$
\begin{aligned}
\sum_{i=0}^{\infty}\mathbb{E}_i\left[h_0^2\right]x^i &= \sum_{i=0}^{\infty}\frac{-8p^2i - 2pi^2 + 8pi + i^2 - i}{(q-p)^3}x^i + \sum_{i=0}^{\infty}\frac{i}{q-p}x^i, \\
&= \sum_{i=0}^{\infty}\cdot\frac{-4p^2i - 2pi^2 + 4pi + i^2}{(q-p)^3}x^i.
\end{aligned}
$$

This result combined with (4.51) finally yields:

$$
\begin{aligned}
\text{Var}_i\left[h_0\right] &= \frac{-4p^2i - 2pi^2 + 4pi + i^2}{(q-p)^3} - \left(\frac{i}{1-2p}\right)^2, \\
&= \frac{-4p^2i + 4pi}{(q-p)^3}, \\
&= \frac{4p(1-p)i}{(q-p)^3}.
\end{aligned}
$$

It follows that the generating function of the conditional variance $\text{Var}_i\left[h_0\right]$ equals

$$
\sum_{i=0}^{\infty}\frac{4p(1-p)}{(q-p)^3}ix^i
$$

in terms of power series. Since we know the generating function of the conditional
variance in terms of power series, we are able to determine a closed form for the
generating function of the conditional variance. From (4.49), it follows that

$$
\frac{x}{(1-x)^2} = \sum_{i=0}^{\infty}ix^i.
$$

Hence, the generating function of the conditional variance becomes

$$
\frac{4pi(1-p)}{(q-p)^3}\cdot\frac{x}{(1-x)^2}.
$$

Chapter 5

Coupling time in acyclic queueing networks

This chapter is dedicated to the effective computation of a bound of the coupling time in acyclic networks. Acyclicity means that the network does not contain any cycles, i.e. a customer cannot return to a queue he already has visited.

Recall that in Chapter 3 we ran the CFTP algorithm on the acyclic network given in Figure 3.1 on page 18. One may see in Figure 3.2 that the coupling time has a peak when $\lambda_0 = 0.4$. This corresponds to the case when the input rate and service rate in queue Q_3 are equal. This should not be surprising regarding the result for a single queue, which says that the coupling time is maximal when the rates are equal. Then a second peak occurs around $\lambda_0 = 1.4$ when coupling in queue Q_0 is maximal. The rest of the curve shows a linear increase of the coupling time which may suggest an asymptotic linear dependence in λ_0. In this part, an explicit bound on the coupling time which exhibits these two features will be derived.

The result of Section 5.1 concerns an extension of inequality (3.5) to distributions. Then the next section shows how the results for a single M/M/1/C queue can be used for an effective computation of bounds for acyclic networks of queues.

Throughout this chapter, we will illustrate the construction of the bound with a tandem network. The first queue Q_0 has a capacity of 6 and the second queue Q_1 has a capacity of 3 (see Figure 5.1). This network is driven by three events:

event e^1**:** an arrival at queue Q_0.

event e^2**:** an end of service at queue Q_0 and routing to queue Q_1 (provided that the number of customers in Q_0 is strictly positive).

event e^3**:** an end of service at queue Q_1 and departure from the system (provided that the number of customers in Q_1 is strictly positive).

Due to monotonicity, the CFTP-algorithm only needs to be applied to state $MIN = (0, 0)$ and state $MAX = (6, 3)$.
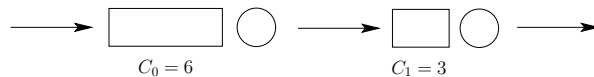

$$C_0 = 6 \qquad C_1 = 3$$

Figure 5.1: *Tandem queueing system.*

5.1 Distribution of the coupling time

In the following, the queues $Q_0, \ldots Q_K$ are numbered according to the topological order of the network. Thus, no event occurring in queue Q_j has any influence on the state of queue Q_i as long as $i < j$.

Proposition 5.1. *The coupling time for an acyclic network is bounded in the stochastic sense by the sum of the forward coupling time of all queues:*

$$\tau^b \leqslant_{st} \tau^f_K + \cdots + \tau^f_0.$$

Proof. The proof is based on the following idea: construct a trajectory of a backward simulation over which the comparison holds. This will imply the stochastic comparison using Strassen's Theorem.

Consider a backward simulation of the network starting at time 0 until coupling occurs for the last queue, at time $-\tau^b_K$. From time $-\tau^b_K$, run a backward simulation until queue Q_{K-1} couples. From time $-\tau^b_K - \tau^b_{K-1}$, run the backward simulation again until queue Q_{K-2} couples. Continue this construction until the first queue has coupled at time $-(\tau^b_K + \cdots + \tau^b_0)$ (see Figure 5.2). Now, on this trajectory, the state in queue Q_0 has coupled between times $-(\tau^b_K + \cdots + \tau^b_0)$ and $-(\tau^b_K + \cdots + \tau^b_1)$. From this time on, Q_0 will remain coupled since no event in other queues may alter its state. The same property holds for queue Q_i between times $-(\tau^b_K + \cdots + \tau^b_i)$ and $-(\tau^b_K + \cdots + \tau^b_{i+1})$, and at time 0, all queues have coupled by acyclicity of the network. Finally, note that the intervals of this simulation are independent of each other so that $\sum_i \tau^b_i = \sum_i \tau^f_i$ in distribution and one gets $\tau^b \leqslant_{st} \tau^f_K + \cdots + \tau^f_0$. $\quad\square$
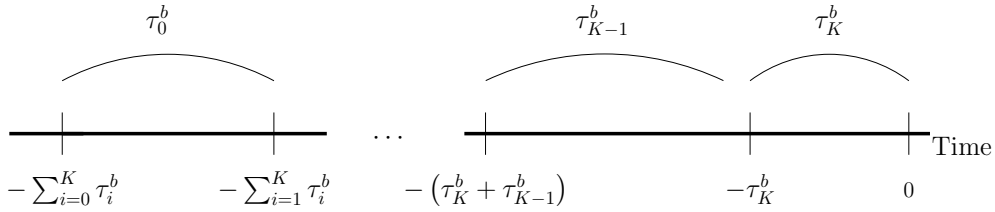

Figure 5.2: *The construction of the proof of Proposition 5.1.*

Note, note that acyclicity is essential in the proof above. For networks with cycles, one would need some kind of association properties of the states of the queues to assess something about the comparison of the distribution of τ^b and the τ^f_i's.

Example 5.1 (Tandem queue). Consider the tandem queueing network. We illustrate the construction of Proposition 5.1 with Figure 5.3. On the vertical axe is set the number of customers in one queue. The solid line represents the evolution of queue Q_0, whereas the dashed line represents the evolution of queue Q_1. Since Q_0 has a capacity equal to 6, the solid line does not exceed the value of 6. For the same reason, the dashed one does not exceed the value of 3. For each queue, two itineraries are depicted: one starting at state 0 and one starting in state C_i. The two trajectories starting in state 0 (one for Q_0 and one for Q_1) form together the evolution of $MIN = (0,0)$, and the two trajectories starting in state 3 and 6 (for Q_0 and Q_1 respectively) form together the evolution of $MAX = (6,3)$.

Figure 5.3.a shows a coupling run for Q_0, and Figure 5.3.b shows a coupling itinerary for queue Q_1. Note that $\tau^b_0 = 12$ and that $\tau^b_1 = 6$. In Figure 5.3.c, one trajectory is constructed from both coupling ensuring itineraries. This composite trajectory leads to coupling.
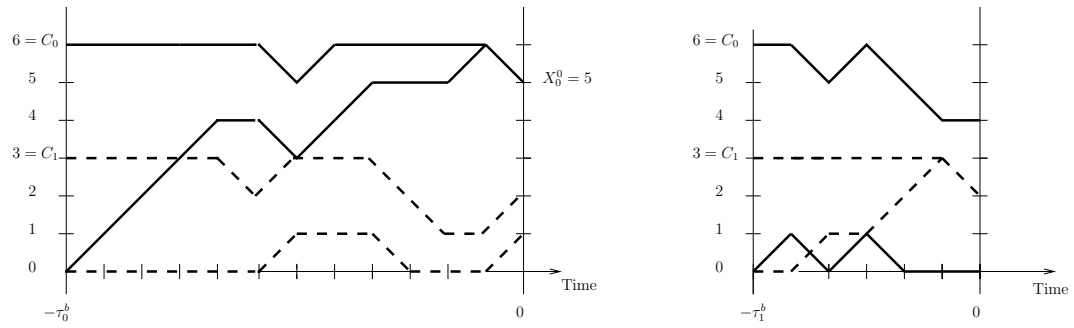
(a) *The simulation of the tandem queue up to coupling of Q_0.*

(b) *The simulation of the tandem network with up to coupling of Q_1.*
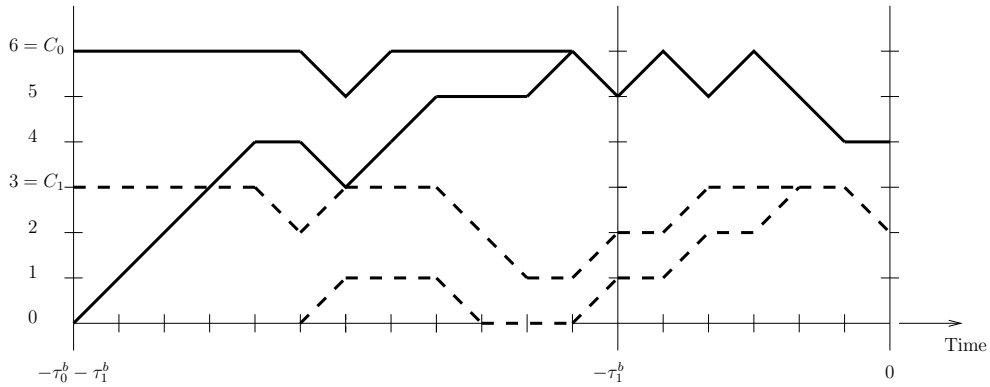
(c) *The composition of the two previous simulations assures global coupling of the network.*

Figure 5.3: *The construction of a bound on the coupling time in a tandem network with $C_0 = 6$ and $C_1 = 3$.*

41

5.2 Upper bound on the coupling time

Here, an acyclic network of M/M/1/C queues with Bernoulli routings is considered. The events here are of only two types:

type 1: Exogenous arrivals. These are Poisson with rate γ_i in queue i.

type 2: Routing of one customer from queue i to queue j after service completion in queue i. These are exponential with rate μ_{ij}.

Queue $K+1$ is a dummy queue representing exits: routing a customer to queue $K+1$ means that the customer exits the network forever. In case of overflow, the new customer trying to enter the full queue is lost. The service rate at queue i is also denoted $\mu_i = \sum_{j=0}^{K+1} \mu_{ij}$.

For the construction of the bound, we will compare the acyclic network with two other models. But first, let us introduce new random variables. The random variable $\tau^b(s_j = x)$ is the backward coupling time of the network, over the set of all initial states with the j-th coordinate equal to x. Namely,

$$\tau^b(s_j = x) = \min \left\{ n \text{ s.t. } \left| \left\{ \phi^{(n)} \left(\mathcal{S} \cap \{s_j = x\}, e_{-n+1}, \ldots, e_0 \right) \right\} \right| = 1 \right\}.$$

Let $\tau_i^b(s_j = x)$ be the backward coupling time on coordinate i given $s_j = x$:

$$\tau_i^b(s_j = x) = \min \left\{ n \text{ s.t. } \left| \left\{ N_i \left(\phi^{(n)} \left(\mathcal{S} \cap \{s_j = x\}, e_{-n+1}, \ldots, e_0 \right) \right) \right\} \right| = 1 \right\},$$

with N_i as defined in Chapter 3.
Since

$$|\{\mathcal{S} \cap \{s_j = x\}\}| \quad < \quad |\mathcal{S}|,$$

we have $\tau^b(s_j = x) \leqslant_{st} \tau^b$ and for all i, $\tau_i^b(s_j = x) \leqslant_{st} \tau_i^b$.

We also have the same notions for forward coupling times:

$$\tau^f(s_j = x) = \min \left\{ n \in \mathbb{N}; \text{ s.t. } \left| \phi^{(n)} \left(\mathcal{S} \cap \{s_j = x\}, e_{1 \to n} \right) \right| = 1 \right\},$$

$\tau_i^f(s_j = x)$ being defined in the same manner, and for hitting times:

$$h_{C_i \to 0}(s_j = x) = \min \left\{ n \in \mathbb{N}; \text{ s. t. } \phi^{(n)} \left(\mathcal{S} \cap \{s_i = C_i, s_j = x\}, e_{1 \to n} \right) \in \mathcal{S} \cap \{s_i = 0\} \right\}.$$

Now one can construct a sequence of $K+1$ backward simulations that ensures coupling in the following way. Let X_j^i denote the state of coupling of queue i after $j+1$ simulations. First simulate the queueing system from the past up to coupling of queue 0. The number of steps is by definition τ_0^b. Queue Q_0 has coupled in a random state X_0^0. Then, run a second backward simulation up to coupling of queue Q_1 given $s_0 = X_0^0$. This simulation takes $\tau_1^b(s_0 = X_0^0)$ steps and the state at time $t = 0$ is X_1^1 for Q_1 and X_0^1 for Q_0.

This construction goes on up to the backward simulation up to coupling of queue Q_K given

$$s_0 = X_0^{K-1}, \quad s_1 = X_1^{K-1}, \quad \ldots \quad , \quad s_{K-1} = X_{K-1}^{K-1}.$$

The last simulation takes

$$\tau_K^b \left(s_0 = X_0^{K-1}, s_1 = X_1^{K-1}, \ldots, s_{K-1} = X_{K-1}^{K-1} \right)$$

steps and the coupling state of Q_K is X_K^K.

Lemma 5.1. *Using the previous construction,*

$$\tau^b \leqslant_{st} \sum_{i=0}^{K} \tau_i^b(s_0 = X_0^{i-1}, \ldots, s_{i-1} = X_{i-1}^{i-1}),$$

and for all i, $(X_0^i, \ldots, X_i^i)$ is steady state distributed for $Q_0, \ldots, Q_i$. Furthermore, for all i,

$$\tau^b \quad \leqslant_{st} \quad \sum_{i=0}^{K} h_{C_i \to 0}(s_0 = X_0^{i-1}, \ldots, s_{i-1} = X_{i-1}^{i-1}).$$

Proof. From the previous sequence of backward simulations one can construct a single simulation by appending them in the reverse order (see Figure 5.4): the backward simulation for queue Q_K preceded by the simulation of Q_{K-1}, and so forth up to the simulation of Q_0. This is a backward simulation of the system (the last state is $(X_0^K, \ldots, X_K^K)$).

A straightforward consequence, using acyclicity, is that $(X_0^i, \ldots, X_i^i)$ is steady state distributed for $Q_0, \ldots, Q_i$ for all i.

Furthermore, one gets in distribution

$$\begin{aligned}
\tau^b \quad &\leqslant_{st} \quad \sum_{i=0}^{K} \tau_i^b(s_0 = X_0^{i-1}, \ldots, s_{i-1} = X_{i-1}^{i-1}) \\
&= \quad \sum_{i=0}^{K} \tau_i^f(s_0 = X_0^{i-1}, \ldots, s_{i-1} = X_{i-1}^{i-1}) \\
&\leqslant_{st} \quad \sum_{i=0}^{K} h_{C_i \to 0}(s_0 = X_0^{i-1}, \ldots, s_{i-1} = X_{i-1}^{i-1}),
\end{aligned}$$

by independence of the variables given the initial states X^{i-1}. $\qquad\square$
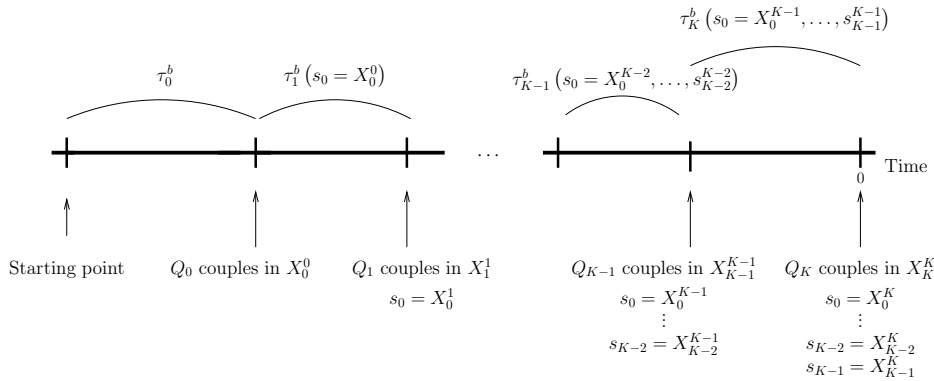

Figure 5.4: *The construction of the proof of Lemma 5.1.*

Example 5.2 (Tandem queue revisited). We will illustrate the construction used in the proof of Lemma 5.1 with the tandem queue example. We use the same events as in Example 5.1. Then the simulation up to coupling of queue Q_0 does not differ from the trajectory leading to coupling of queue Q_0 in Figure 5.3.a. Note that coupling occurs in state 5.

Then we start a second simulation, given that $s_0 = 5$, up to coupling of queue Q_1. The trajectory of this simulation is represented in Figure 5.5.a. Note that

there is only solid line to represent the evolution of Q_0 since this queue has already coupled. The system finally couples in state $(3, 2)$ and $\tau_1^b (s_0 = 5) = 5$. Recall from Example 5.1 that τ_1^b for this sequence of events was equal to 6. So indeed, $\tau_1^b (s_0 = 5) \leqslant \tau_1^b$.

The composition of both simulations of queue Q_0 (from Figure 5.3.a) and of queue Q_1 (Figure 5.5.a) into a single simulation is shown in Figure 5.5.b. This construction yields a series of events which assures coupling of the system. $\triangle$
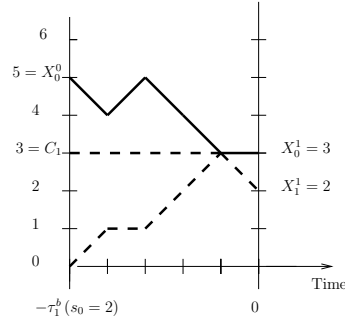
(a) *The simulation of the tandem network with up to coupling of Q_1, given that $X_0^0 = 5$.*
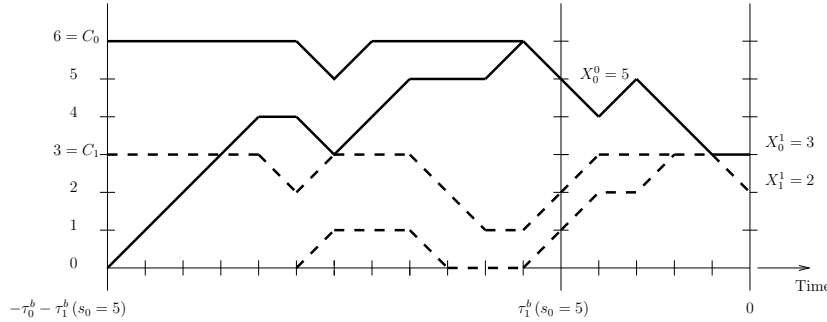
(b) *The composition of the two previous simulations assures global coupling of the network.*

Figure 5.5: *The construction of a bound on the coupling time in a tandem network with $C_0 = 6$ and $C_1 = 3$.*

Model 1: ∞-model

Let us now consider the first new model. This model has one difference from the original one: all queues are replaced by infinite queues, except for queue Q_i which stays the same. In the following, all the notations related to this new network will be expressed by appending the ∞ symbol to all variables corresponding to this new circuit.

Once this model achieves the steady state, the input stream in queue i is Poisson, by using Burke's Theorem (see Appendix A). The rate of the input stream in queue i consists of the exogenous input stream in queue i and of the proportion of the

rate at which customers are directed to queue i from others queues. Thus the rate of the input stream in queue i is given by ℓ_i, the solution of the flow equations:

$$\ell_i = \sum_{j<i} \ell_j \frac{\mu_{ji}}{\mu_j} + \gamma_i.$$

The network is said to be *stable* for queue i as soon as $\ell_i < \mu_i$. We assume stability for all i in the sequel. Remark that strictly speaking, queue i is stable, since it has a finite capacity. However, we use in the following our definition for the concept of stability.

One can construct a sequence of backward simulations for the new network in the same way as for the original network. This provides the quantities

$$^{\infty}X_j^{i-1}, \quad ^{\infty}\tau_i^b(s_0 = {}^{\infty}X_0^{i-1}, \ldots, s_{i-1} = {}^{\infty}X_{i-1}^{i-1}),$$

$$^{\infty}\tau_i^f(s_0 = {}^{\infty}X_0^{i-1}, \ldots, s_{i-1} = {}^{\infty}X_{i-1}^{i-1}),$$

and

$$^{\infty}h_{C_i \to 0}(s_0 = {}^{\infty}X_0^{i-1}, \ldots, s_{i-1} = {}^{\infty}X_{i-1}^{i-1}).$$

The monotony property given above implies that $X_i^j \leqslant_{st} {}^{\infty}X_i^j$ and

$$h_{C_i \to 0}(s_0 = X_0^{i-1}, \ldots, s_{i-1} = X_{i-1}^{i-1}) \leqslant_{st} {}^{\infty}h_{C_i \to 0}(s_0 = {}^{\infty}X_0^{i-1}, \ldots, s_{i-1} = {}^{\infty}X_{i-1}^{i-1}).$$

Model 2: Isolated queue with null events

The next step is to build yet another model. This model is made of a single $M/M/1/C_i$ queue with three types of events

type 1: Arrivals of customers with rate ℓ_i (provided that the number of customers is smaller than C_i).

type 2: Departures with rate μ_i (provided that the number of customers is positive).

type 3: Null events (with no effect on the queue) with rate $\Lambda - \ell_i - \mu_i$, where Λ is the sum of all rates of the original network, i.e $\Lambda = \sum_{i \in \mathcal{S}} \gamma_i + \sum_{i \in \mathcal{S}} \mu_i$.

For this isolated model, let us introduce the uniformizing probabilities

$$\begin{aligned} p_i &= \frac{\ell_i}{\ell_i + \mu_i}, \\ q_i &= 1 - p_i, \\ d_i &= \frac{\Lambda - \ell_i - \mu_i}{\Lambda}. \end{aligned}$$

Let F_k be the time to go from state k to state 0 in the isolated system. A one step analysis gives

$$\begin{aligned} \mathbb{E}[F_k] &= 1 + d_i \mathbb{E}[F_k] + \frac{\ell_i}{\Lambda} \mathbb{E}[F_{(k+1) \wedge C_i}] + \frac{\mu_i}{\Lambda} \mathbb{E}[F_{(k-1) \vee 0}] \\ &= \frac{1}{1 - d_i} \left(1 + \frac{\ell_i}{\Lambda} \mathbb{E}[F_{(k+1) \wedge C_i}] + \frac{\mu_i}{\Lambda} \mathbb{E}[F_{(k-1) \vee 0}] \right) \\ &= \frac{1}{1 - d_i} + p_i \mathbb{E}[F_{(k+1) \wedge C_i}] + q_i \mathbb{E}[F_{(k-1) \vee 0}]. \end{aligned}$$

We get the same equation as (4.31) except for the additional constant which is now $\frac{1}{1-d_i}$ instead of 1, so that the solution is the same as before up to a multiplicative factor of $\frac{1}{1-d_i} = \frac{\Lambda}{\ell_i+\mu_i}$. Using Equation (4.29), one gets

$$\mathbb{E}[F_{C_i}] = \frac{\Lambda}{\ell_i + \mu_i} \left(\frac{C_i}{q_i - p_i} - \frac{p_i\left(1 - \left(\frac{p_i}{q_i}\right)^{C_i}\right)}{(q_i - p_i)^2} \right). \tag{5.1}$$

Lemma 5.2. *Under the foregoing notations and assumptions,*

$$^{\infty}h_{C_i\to 0}(s_0 = {}^{\infty}X_0^{i-1}, \dots, s_{i-1} = {}^{\infty}X_{i-1}^{i-1}) = F_{C_i},$$

in distribution.

Proof. First, using Lemma 5.1 for the new network with infinite queues (except for Q_i), the state $({}^{\infty}X_0^{i-1}, \dots, {}^{\infty}X_{i-1}^{i-1})$ is steady state distributed. Using Burke's Theorem, this implies that the input stream in queue Q_i is Poisson with rate ℓ_i, when one runs a simulation starting in any state in $\mathcal{S} \cap \{s_i = C_i, s_j = {}^{\infty}X_j^{i-1}, j < i\}$.

Now, during this simulation, one can couple the addition, subtraction and null events for queue Q_i in isolation and for Q_i in the complete network of infinite queues, all of them having the same laws. This implies that the state of queue Q_i in both systems is the same under that coupling. Hence, they reach 0 at the same time: $^{\infty}h_{C_i\to 0}(s_0 = {}^{\infty}X_0^{i-1}, \dots, s_{i-1} = {}^{\infty}X_{i-1}^{i-1}) = F_{C_i}$ in distribution. $\square$

Derivation of the bound

We are ready to put everything together in expectation:

$$\mathbb{E}\tau^b \quad \leqslant_{st} \quad \sum_i \mathbb{E}[h_{C_i\to 0}(s_0 = X_0^{i-1}, \dots, s_{i-1} = X_{i-1}^{i-1})] \tag{5.2}$$

$$\leqslant \quad \sum_i \mathbb{E}[{}^{\infty}h_{C_i\to 0}(s_0 = {}^{\infty}X_0^{i-1}, \dots, s_{i-1} = {}^{\infty}X_{i-1}^{i-1})] \tag{5.3}$$

$$= \quad \sum_i \mathbb{E}[F_{C_i}]. \tag{5.4}$$

The sequence of inequalities may not hold in distribution because the variables X^i and thus $h_{C_i\to 0}(s_0 = X_0^{i-1}, \dots, s_{i-1} = X_{i-1}^{i-1})$ are not independent.

Using (5.1),

$$\mathbb{E}\tau^b \leqslant \sum_i \frac{\Lambda}{\ell_i + \mu_i} \left(\frac{C_i}{q_i - p_i} - \frac{p_i\left(1 - \left(\frac{p_i}{q_i}\right)^{C_i}\right)}{(q_i - p_i)^2} \right).$$

In subsection 4.1.2 we have seen that $\left(\frac{C_i}{q_i-p_i} - \frac{p_i\left(1-\left(\frac{p_i}{q_i}\right)^{C_i}\right)}{(q_i-p_i)^2} \right) \leqslant C_i + C_i^2$ for

$p_i \leqslant q_i$. We summarize the results of this part in the following theorem.

Theorem 5.1. *In an acyclic stable network of $K + 1$ $M/M/1/C_i$ queues with Bernoulli routing and losses in case of overflow, the coupling time from the past satisfies in expectation,*

$$\mathbb{E}[\tau^b] \leqslant \sum_{i=0}^{K} \frac{\Lambda}{\ell_i + \mu_i} \left(\frac{C_i}{q_i - p_i} - \frac{p_i(1 - \left(\frac{p_i}{q_i}\right)^{C_i})}{(q_i - p_i)^2} \right) \leqslant \sum_{i=0}^{K} \frac{\Lambda}{\ell_i + \mu_i}(C_i + C_i^2). \tag{5.5}$$

46

Remark 5.1. Observe that in case Q_i has only exogenous arrivals, the arrivals at Q_i occur according to a Poisson process. Then we have

$$\mathbb{E}\left[F_{C_i}\right] = \mathbb{E}\left[^{\infty}h_{C_i \to 0}\right] = \mathbb{E}\left[h_{C_i \to 0}\right].$$

Therefore, we can bound the coupling time on this queue in the network by considering it as a single M/M/1/C queue, except for the additionally factor of $\Lambda/(\ell_i + \mu_i)$. Thus, we have:

$$\begin{aligned}
\mathbb{E}\left[h_{C_i \to 0}\right] &\geqslant \min\left\{\mathbb{E}\left[h_{C_i \to 0}\right], \mathbb{E}\left[h_{0 \to C_i}\right], \frac{\Lambda}{\ell_i + \mu_i} \cdot \frac{C_i^2 + C_i}{2}\right\}, \\
&\geqslant \mathbb{E}\tau_i^b.
\end{aligned}$$

with

$$\mathbb{E}\left[h_{C_i \to 0}\right] = \frac{\Lambda}{\ell_i + \mu_i}\left(\frac{C_i}{q_i - p_i} - \frac{p_i\left(1 - \left(\frac{p_i}{q_i}\right)\right)^{C_i}}{(q_i - p_i)^2}\right)$$

and

$$\mathbb{E}\left[h_{0 \to C_i}\right] = \frac{\Lambda}{\ell_i + \mu_i}\left(\frac{C_i}{q_i - p_i} - \frac{p_i\left(1 - \left(\frac{p_i}{q_i}\right)\right)^{C_i}}{(q_i - p_i)^2}\right).$$

.

Remark 5.2. Suppose we have a queue Q_i with arrival rate ℓ_i and a service rate μ_i such that $\ell_i > \mu_j$, i.e the queue is instable. Let Q_j be a queue which is directly fed by the departures of Q_i, thus $\mu_{ij} > 0$. Coupling in queue Q_i occurs in state $X_i^i < \infty$. In the ∞- network, we have,

$$^{\infty}h_{C_j \to 0}\left(s_i = X_i^i\right) \leqslant {}^{\infty}h_{C_j \to 0}\left(s_i = \infty\right).$$

By supposing that Q_i has an infinite number of customers, the departure process of Q_i is a Poisson process with rate $\min\{\ell_i, \mu_i\}$. Since $\mu_i < \ell_i$, we can model the departure rate of Q_i by μ_i. Since our bound is based on the time to get from state C_j to state 0, this improves the bound for Q_j.

Chapter 6

Numerical experiments

In the previous chapter, we derived a bound on the coupling time in acyclic networks. In this chapter we will compare this bound with experimental values of the coupling time. Before presenting the experiments, we can already indicate three factors which may be responsible for the inaccuracy of the bound given by Theorem 5.1.

- The first factor is the replacement of the max by the sum. We believe that it may be a hard task to get rid of this first approximation because of the intricate dependencies between the queues. Furthermore, experiments reported below show that this may not even be possible in many cases (see Figure 6.3).

- Another factor which may increase the inaccuracy of our bounds is the fact that most events change the states of several queues at the same time, while the bound given here disregards this. This may add a factor 2 between the true coupling time and the bound given in Theorem 5.1.

- The most important factor which jeopardizes the quality of the bound is the stability issue. If one of the queues is unstable, the bound provided by Equation (5.5), also called the light traffic bound in Proposition 4.5, is very bad (as seen in Figure 4.5). The reason for this is that an unstable queue, tends to couple in state C, while the bound is based upon coupling in state 0.

  Nevertheless, under certain conditions we are able to allow instable queues and derive a bound which is not too bad. This is based on combining Remark 5.1 and 5.2. For a queue Q_0 with only exogenous arrivals, we can bound the coupling time of this queue by using the bound introduced by Remark 5.1. In case the queues which are directly nourished by this queue are stable with respect to the service rate of Q_0, we obtain a rather good bound. This will be further investigated in 6.1.2. So far we have not been able to come up with a better bound for unstable queues, unless for this particular case.

  However, when all queues are stable (and even more so when the load is smaller than 2/3), the bound tends to be more accurate. One should however note that, on a practical point of view, most actual networks which require stationary performance evaluations are indeed stable.

In the experiments, we focus on the stability issue in section 6.1 and on the dependencies between queues that block the replacement of max by sum in section 6.2. For each experiment, the number of simulation runs equals 10,000. In the construction of a bound, we use the result of Theorem 5.1 that

$$\mathbb{E}[\tau^b] \leqslant \sum_{i=0}^{K} B_i,$$

with

$$B_i = \frac{\Lambda}{\ell_i + \mu_i} \left(\frac{C_i}{q_i - p_i} - \frac{p\left(1 - \left(\frac{p_i}{q_i}\right)^{C_i}\right)}{(q_i - p_i)^2} \right).$$

6.1 Stability

For the first series of experiments, we re-use the network introduced in Chapter 3. The flow chart of this model is depicted in Figure 6.1.
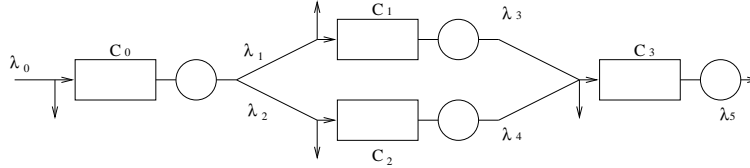

Figure 6.1: *Flow chart of our model*

6.1.1 Stability of last queue

We run three simulations, and the following parameters are fixed: The input rate is $\lambda_0 = 0.4$ and the rates of the other events are $\lambda_1 = 1.4$, $\lambda_2 = 0.6$, $\lambda_3 = 0.8$ and $\lambda_4 = 0.5$. To investigate the stability issue, we set a different value for λ_5 in each simulation. Recall that ℓ_i is the solution of

$$\ell_i = \sum_{j<i} \ell_j \frac{\mu_{ji}}{\mu_j} + \gamma_i,$$

with γ_i the exogenous arrival rate at queue i and that $p_i = \ell_i/(\ell_i + \mu_i)$. Now we can determine ℓ_i, μ_i and p_i for $i = 0, 1, 2, 3$:

i	input stream ℓ_i	service rate μ_i	probability of arrival p_i
0	0.40	2.0	1/6
1	0.28	0.8	7/27
2	0.12	0.5	6/31
3	0.4	λ_5	$0.4/(0.4+\lambda_5)$

The number of simulation runs is 10,000. The capacity C_i is the same for all the four queues and we let it vary from 1 to 20.

In the first model, we set $\lambda_5 = 0.2$ such that the last queue Q_3 is instable. For the second model, we set $\lambda_5 = 0.6$ such that Q_3 is stable and in the third model we set $\lambda_5 = 0.4$ such that the last queue is barely instable.

Model 1: Q_3 is instable

In this model, $\lambda_5 = 0.2$ so that queue Q_3 is unstable. Now $\Lambda = 3.9$, $\mu_3 = 0.2$ and $p_3 = 2/3$.

Using the expression for the B_i's, we obtain:

$$B_0 = \frac{3.9}{2.4}\left(\frac{3}{2}C - \frac{3}{8} + \frac{3}{8}\left(\frac{1}{5}\right)^C\right)$$

$$= \frac{39}{16}\left(\frac{C}{4} - 1 + \left(\frac{1}{5}\right)^C\right).$$

In a similar way we obtain B_i for the three remaining queues:

$$B_1 = \frac{15}{2}C - \frac{105}{26} + \frac{105}{26}\left(\frac{7}{20}\right)^C,$$

$$B_2 = \frac{195}{19}C - \frac{1170}{361} + \frac{1170}{361}\left(\frac{6}{25}\right)^C,$$

$$B_3 = -\frac{39}{2}C - 39 + 39 \cdot 2^C.$$

Now the bound on the backward coupling time becomes $\sum_{i=0}^{K} B_i$ with the B_i's as above. Note that the bounds B_0, B_1 and B_2 get lineair in C as C increases. However, the bound B_3 is exponentially increasing as C get large and therefore makes that our bound explodes. This is shown in Figure 6.2 which displays the bound as well as the mean coupling time computed over the 10,000 simulation runs. A ratio larger than 10 with respect to the true coupling time is reached when $C = 5$. It should also be noticed that the bound is convex in C while the coupling time is not.
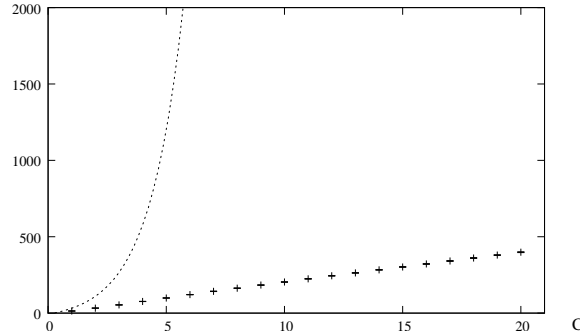

Figure 6.2: *This figure displays the coupling time (dots) with 95% confidence intervals, and the bound given by Equation (5.5) when queue Q_3 is unstable ($\lambda_5 = 2/10$), while the capacity C varies from 1 to 20.*

Model 2: Q_3 is stable

In the model, λ_5 equals 0.6, and all queues are stable with a load smaller that 2/3. Now $\Lambda = 4.3$. For queue 3 we get $\mu_3 = 0.6$ and $p_3 = 2/5$. We obtain the bound $\sum_{i=0}^{3} B_i$ with:

$$B_0 = \frac{43}{16}C - \frac{43}{64} + \frac{43}{64}\left(\frac{1}{5}\right)^C,$$

$$B_1 = \frac{215}{26}C - \frac{1505}{338} + \frac{1505}{338}\left(\frac{7}{20}\right)^C,$$

$$B_2 = \frac{215}{19}C - \frac{1290}{361} + \frac{1290}{361}\left(\frac{6}{25}\right)^C,$$

$$B_3 = \frac{43}{2}C - 43 + 43\left(\frac{2}{3}\right)^C.$$

Figure 6.3 shows this bound and the mean coupling time computed by simulation runs. Both curves appear to be almost linear in C (this is true for the bound: when q_i/p_i is small, $\mathbb{E}F_{C_i}$ is almost linear in C_i) and the ratio is smaller than 1.3.

The third curve in Figure 6.3 is $\max_{i\in\{0,\dots3\}} B_i$. Notice that

$$\max_{i\in\{0,\dots3\}} B_i < \mathbb{E}\left[\tau^b\right]$$

and thus we cannot replace the sum by the max. This is to be related with the first item in the comments above.
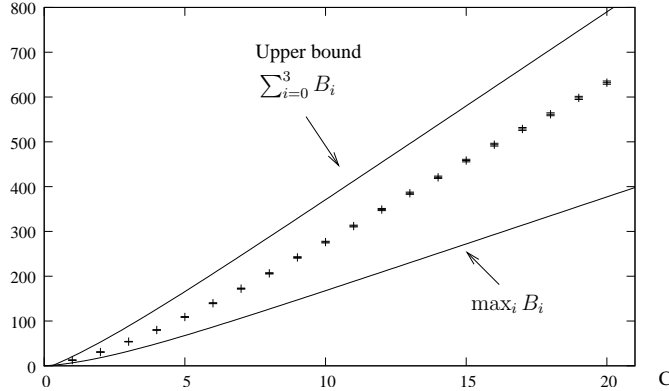

Figure 6.3: *Here are the bound given by Equation (5.5), the mean coupling time (dots) with 95% confidence intervals and the maximum over Equations (5.5) for all queues, when queue Q_3 is stable ($\lambda_5 = 6/10$), while the capacity C varies from 1 to 20.*

Model 3: Q_3 is barely instable

This time, we set $\lambda_0 = 0.4$, so that $\ell_3 = \mu_3$ and thus Q_3 is barely unstable. This would correspond to the maximal coupling time for Q_3 if it was alone. Furthermore, we have $\Lambda = 4.1$ and $p_3 = 1/2$. For the B_i's we obtain:

$$B_0 = \frac{41}{16}C - \frac{41}{64} + \frac{41}{64}\left(\frac{1}{5}\right)^C,$$

$$B_1 = \frac{205}{26}C - \frac{1435}{338} + \frac{1435}{338}\left(\frac{7}{20}\right)^C,$$

$$B_2 = \frac{205}{19}C - \frac{1230}{361} + \frac{1230}{361}\left(\frac{6}{25}\right)^C,$$

$$B_3 = \frac{43}{8}\left(C^2 + C\right).$$

51

and then $\mathbb{E}\tau^b \leqslant \sum_{i=0}^{3} B_i$.

Note that for queue Q_3, we use a bound in $C + C^2$ which is a bad approximation because of the loss of the factor 2 when compared with the bound for isolated queues.

Figure 6.4 displays the backward coupling time and the bound. Note that the total gap has a ratio which is almost 2. In that case both the coupling time and the bound exhibit a convex behaviour with respect to C. A ratio smaller than 2 is indeed interesting because efficient perfect simulation algorithm use a doubling window technique to reduce the complexity and their running time (see Equation (2.9)) so that our bound gives a good estimation of the mean running time of the algorithms.
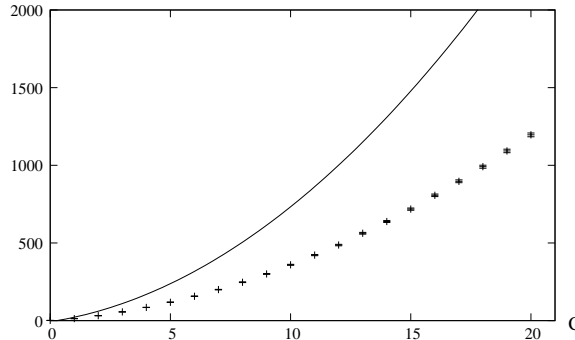

Figure 6.4: *Display of the coupling time (dots) with 95% confidence interval and the bound given by Equation* (5.5) *when queue Q_3 is barely unstable ($\lambda_5 = 4/10$) while the capacity C varies from 1 to 20.*

6.1.2 Stability of first queue

In this model, we use the same flow chart as in the three preceding examples, but with different parameters. We let $\lambda_1 = 0.5$, $\lambda_2 = 0.5$, $\lambda_3 = 1.0$, $\lambda_4 = 0.8$ and $\lambda_5 = 1.2$. For all four queues, we take again the capacity equal to 10. The exogenous input rate λ_0 is varying from 0 to 10. Again we run 10,000 simulation runs.

Observe that as soon as $\lambda_0 \geqslant 1$, the first queue becomes instable. From Remark 5.1 it follows that the instability of Q_0 is not a problem, since we can treat this queue as a isolated M/M/1/C queue, except for the additional factor of $\Lambda/(\ell_i + \mu_i)$.

However, stability is an issue for queues that are fed by departures of other queues. Since we determine the bound on these queues by the time it takes to get from C to 0, the queue need to be stable in order to obtain an acceptable bound. From Remark 5.1 it follows that we can model the departure rate of queue Q_0 with rate by $\min\{\lambda_0, \mu_0\}$. Thus as soon as the input rate exceeds 1, we take the departure process of Q_0 with rate 1. We can also apply this reasoning for the other queues. Then we find the following values for ℓ_i:

	$\lambda_0 < 1$	$\lambda_0 \geqslant 1$
ℓ_0	λ_0	λ_0
ℓ_1	$0.5 \cdot \lambda_0$	0.5
ℓ_2	$0.5 \cdot \lambda_0$	0.5
ℓ_3	$1.0 \cdot \lambda_0$	1.0

Note that the queues Q_1, Q_2 and Q_3 are stable for all input rates λ_0. Now, we will construct the bound on the coupling time. Since we can treat Q_0 as an isolated

simple queue, we calculate the three bounds form Theorem 4.5 with the additional factor.:

$$
\begin{array}{rcll}
B_0 &=& \frac{\Lambda}{1+\lambda_0}\left(\frac{10}{q_0-p_0}-\frac{p_0\left(1-\left(\frac{p_0}{q_0}\right)\right)^{10}}{(q_0-p_0)^2}\right) & \text{Light Traffic bound} \\[3ex]
B_0' &=& \frac{\Lambda}{1+\lambda_0}\cdot\frac{10^2+10}{2} & \text{Critical bound} \\[3ex]
B_0'' &=& \frac{\Lambda}{1+\lambda_0}\left(\frac{10}{q_0-p_0}-\frac{p_0\left(1-\left(\frac{p_0}{q_0}\right)\right)^{10}}{(q_0-p_0)^2}\right) & \text{High Traffic bound}
\end{array}
$$

In establishing the B_i for $i=1,2,3$ we need to distinguish between the case with $\lambda_0<1$ and with $\lambda_0\geqslant 1$.

For $\lambda_0<1$ we get:

$$
B_1 = \frac{\Lambda}{1+0.5\lambda_0}\left(\frac{10}{q_1-p_1}-\frac{p_1\left(1-\left(\frac{p_1}{q_1}\right)\right)^{10}}{(q_1-p_1)^2}\right) \quad \text{with } p_1=\frac{0.5\lambda_0}{0.5\lambda_0+1},
$$

$$
B_2 = \frac{\Lambda}{0.8+0.5\lambda_0}\left(\frac{10}{q_2-p_2}-\frac{p_2\left(1-\left(\frac{p_2}{q_2}\right)\right)^{10}}{(q_2-p_2)^2}\right) \quad \text{with } p_2=\frac{0.5\lambda_0}{0.5\lambda_0+0.8},
$$

$$
B_3 = \frac{\Lambda}{1.2+0.5\lambda_0}\left(\frac{10}{q_3-p_3}-\frac{p_3\left(1-\left(\frac{p_3}{q_3}\right)\right)^{10}}{(q_3-p_3)^2}\right) \quad \text{with } p_3=\frac{0.5\lambda_0}{0.5\lambda_0+1.2},
$$

and for $\lambda_0\geqslant 1$ we get:

$$
B_1' = \frac{\Lambda}{1+0.5}\left(\frac{10}{q_1-p_1}-\frac{p_1\left(1-\left(\frac{p_1}{q_1}\right)\right)^{10}}{(q_1-p_1)^2}\right) \quad \text{with } p_1=\frac{0.5}{0.5+1},
$$

$$
B_2' = \frac{\Lambda}{0.8+0.5}\left(\frac{10}{q_2-p_2}-\frac{p_2\left(1-\left(\frac{p_2}{q_2}\right)\right)^{10}}{(q_2-p_2)^2}\right) \quad \text{with } p_2=\frac{0.5}{0.5+0.8},
$$

$$
B_3' = \frac{\Lambda}{1.2+0.5}\left(\frac{10}{q_3-p_3}-\frac{p_3\left(1-\left(\frac{p_3}{q_3}\right)\right)^{10}}{(q_3-p_3)^2}\right) \quad \text{with } p_3=\frac{0.5}{0.5+1.2}.
$$

Now we can construct the following bounds on the mean coupling time of the network:

$$
\begin{array}{rcl}
\text{Bound 1} &=& B_0+B_1+B_2+B_3, \\
\text{Bound 2} &=& B_0'+B_1+B_2+B_3, \\
\text{Bound 3} &=& B_0'+B_1'+B_2'+B_3', \\
\text{Bound 4} &=& B_0''+B_1'+B_2'+B_3'.
\end{array}
$$

Note that Bound 1 is the bound we obtain without using Remark 5.2 and Remark 5.1. Bound 2 is obtained by using Remark 5.1 and for Bound 2 and Bound 4 both Remarks are used. These four bounds with the mean coupling time issued from the simulations is shown in Figure 6.5. Remark that modelling the departure rate of Q_0 by its service rate as soon as λ_0 depasses 1, highly improves the bound. Futhermore, remark that the bound has a similar form as the coupling time.

6.2 Dependencies between queues

In the introduction we pointed out that the replacement of the max by the sum makes our bound in some cases not appropriate. In Figure 6.3 we already saw that sometimes the max is lower than the real coupling time obtained in simulation runs. In this section, we show that the dependencies between the queues in the network can play a role in the mean coupling time. To show this, we compare two queueing
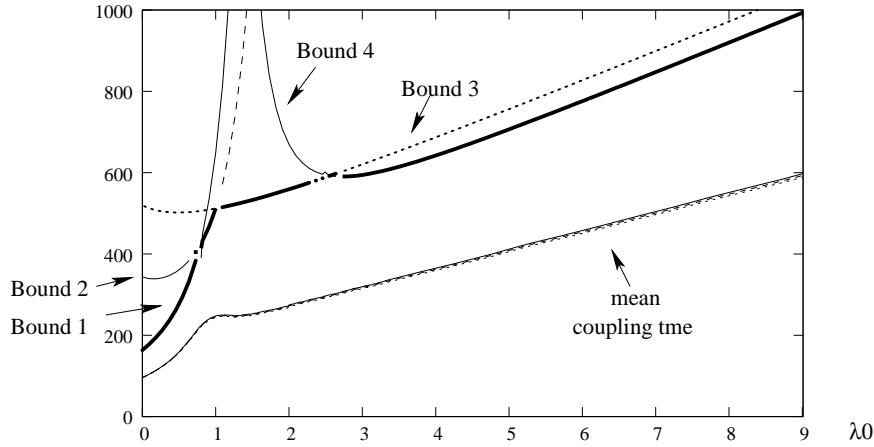
Figure 6.5: *The four bounds and the mean coupling time (dots) with 95% confidence intervals.*

systems. Both systems consist of three queues in series, each of capacity 10 (Figure 6.6). In both queues, we set the arrival rate $\lambda_0 = 1$. In the first queueing system, we set the service rates in an increasing rate: $\mu_0 = 2$, $\mu_1 = 4$ and $\mu_2 = 8$. In the second queue we set the service rates in a decreasing order: $\mu_0 = 8$, $\mu_1 = 4$ and $\mu_2 = 2$. Applying the method to construct a bound, we obtain the same bound for both systems, and one can show that

$$\mathbb{E}\tau^b \quad \leqslant \quad \sum_{i=0}^{2} \frac{\Lambda}{\ell_i + \mu_i} \left(\frac{C_i}{q_i - p_i} - \frac{p_i \left(1 - \left(\frac{p_i}{q_i} \right)^{C_i} \right)}{(q_i - p_i)^2} \right) < 205.$$
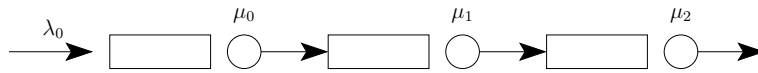

Figure 6.6: *The flow chart of the three queues in series*

By running 10,000 simulation runs for each model, we obtain a mean coupling time of 148.2 for the network with increasing service rates and a mean coupling time of 193.4 for the model with decreasing service rates (see Table 6.1). Note that in the decreasing service rate model, the mean coupling time is relatively close to our bound. What can explain the difference between these means? From Table 6.2 we see that in the increasing model, the coupling time is in more than half of the time determined by the time it takes the first queue to couple. For the decreasing model, in almost all cases, it is the last queue that couples the last and thus determine the coupling time of the system.

The explanation for this behaviour is the following. In the increasing service rate model, the ratio between the arrival rate and the service rate is equal. However, due to uniformization, most of the events affect Q_2 and the least events affect Q_0. Therefore, Q_2 tends to couple faster then Q_0.

For the decreasing service rate model, the system is not in the stationary mode from the beginning. Therefore, note that at the beginning Q_1 and Q_2 are instable. Only queue Q_0 is stable (even exteremly stable) and thus will couple very fast. Then after coupling, Q_0 is stationary and thus the arrival rate at Q_1 is 1. Hence, Q_1 will couple and Q_2 is the last queue to couple. We see from Table 6.2 that in almost all

	mean coupling time			
	global system	Q_0	Q_1	Q_2
series with increasing rates	148.2	133.2	85.4	55.1
series with decreasing rates	193.4	20.1	66.5	193.4

Table 6.1: *Mean coupling time of the queueing systems with increasing and decreasing rates, and the mean coupling time per queue.*

	Frequency		
	Q_0	Q_1	Q_2
series with increasing rates	0.5298	0.2936	0.1766
series with decreasing rates	0.0000	0.0008	0.9992

Table 6.2: *Frequency of queue with longest coupling time.*

simulation runs, indeed the last queue is responsible for the mean coupling time. Thus this system couples exactly in the order we used to construct the bound.

We see that the intricate dependencies of the queues play a strong role in the coupling time. We also showed that in some networks, the queues indeed couple almost queue by queue, as is supposed in our construction on the bound. Hence, it is hard to replace the sum. It might be subject for further research to investigate the dependencies between the queues in the network in order to find a better bound.

Chapter 7

Conclusion

The aim of this thesis was to study the coupling time of the coupling from the past algorithm in queueing networks. Now we will shortly summarise the main result, state a recommendation for effectively using the CFTP algorithm and point out topics for further research.

In the analysis of the coupling time, we first focussed on single finite capacities queues. We derived a recurrent expression for the exact coupling time and showed that the mean coupling time in a single finite capacity queue is maximal when the input rate and output rate are equal. Moreover, we derived three easy calculable upper bounds on the coupling time that are based on hitting times: a light traffic, high traffic and critical bound (page 30). Using formal series, we derived a stochastical bound on the moments of the coupling time in a single finite capacity queue.

The light traffic bound served to build a bound on the mean coupling time in queueing networks. Experiments showed that the mean coupling time shows a asymptotic lineair dependence with respect to the exogenous arrival rate. The bound we established features this linear dependence.

7.1 Recommendation for using the algorithm

In chapter 3 we explained how one can apply the CFTP algorithm after uniformization of a queueing system. We explained that if one picks an event whose enabling condition is not met, the systems stays unchanged. However, choosing such an event does indeed increase the coupling time. When one uses the CFTP algorithm in order to obtain a steady state distributed sample, one would like to avoid that one picks events that do not change the system at all. Therefore, one can determine for each state the set of admissible events and pick an event out of this set for every state. However, to do this, one needs to check in what state one is and this testing increases the complexity. Therefore, one should avoid to test on every state, but more likely to test for some specific events that might occur very often. The effect of picking events that cannot be carried out, increases when there is an event that dominates the uniformized Markov chain. Therefore, one should test only whether a dominating event is admissible. This means that one only tests on the coordinates of the queues that are involved with this particular event. For example, if the input rate at Queue Q_i is very large compared to its service rate, one only needs to check whether the number of customers in Q_i equals C_i. In testing this way, one only needs to test one coordinate of the state $s \in \mathcal{S}$.

This effect explains the linear behaviour of the coupling time with respect to the input rate at the topological first queue at the system: a very large input rate with

respect to the service rate, makes that a lot of customers arrive and immediately leave the system. But it does increase the coupling time.

7.2 Topics for further research

The topics for further research are strongly related to the three factors we indicated in Chapter 6 that may be responsible for the inaccuracy of the bound.

- One might study the intricate dependencies between the queues. Our bound does not take into account the structure between the queues. However, in section 6.2 we have shown that the structure of the chain does play a role in the mean coupling time.

- In the establishment of our bound, we studied acyclic networks with Bernoulli routings. One can extend the analysis to cyclic networks, or networks with other routing policies.

- The bound we derived is based on the time it takes each queue of the network to reach state 0 (light traffic bound of 4.5). Therefore, the bound is terribly bad in case the network contains unstable queues. We have partly overcome this, since we can obtain acceptable bounds when only queues with only exogenous arrivals are unstable. While we have only been able to show that the light traffic bound holds for each queue, we conjecture that the heavy traffic bound and the critical bound should also hold. This would yield an overall quadratic bound:

$$\mathbb{E}[\tau^b] \leqslant \sum_{i=0}^{K} \frac{\Lambda}{\ell_i + \mu_i} O(C_i^2),$$

for any monotone Markovian network of queues with a finite state space. Furthermore under light or heavy traffic in all queues, the bound should rather be linear:

$$\mathbb{E}[\tau^b] \leqslant \sum_{i=0}^{K} \frac{\Lambda}{\ell_i + \mu_i} O(C_i).$$

To illustrate this conjecture, we have run simulations for the network displayed in Figure 3.1 with the following parameters. The rates are $\lambda_0 = 0.4, \lambda_1 = 1.4$, $\lambda_2 = 0.6$, $\lambda_3 = 0.8$, $\lambda_4 = 0.5$. The capacity is fixed to 10 in all queues and we let λ_5 (the service rate in Q_3) vary from 0 to 4. As long as $\lambda_5 < 0.4$, Q_3 is unstable and our proven bound (B_1) is poor. As soon as λ_5 is large enough our bound becomes acceptable. In Figure 7.1, note that both the bound and the mean coupling time τ^b have a linear asymptotic growth in λ_5. The Figure also displays the heavy traffic bound B_2 and the critical bound B_3. Should these two bounds hold, the minimum of B_1, B_2, B_3 (in bold in the figure) would provide a remarkable bound on the coupling time, up to an additional constant, since these bounds scale very well with the number of queues. This explains why perfect simulation of monotone queueing networks is so fast, especially when dealing with large scale networks as in [12] where systems with up to 32 queues of capacity 30 (the state space is of size $31^{32} \approx 10^{47}$) are sampled over a classical desktop computer is less than 20 milli-seconds. This is good enough to estimate rare event probabilities.

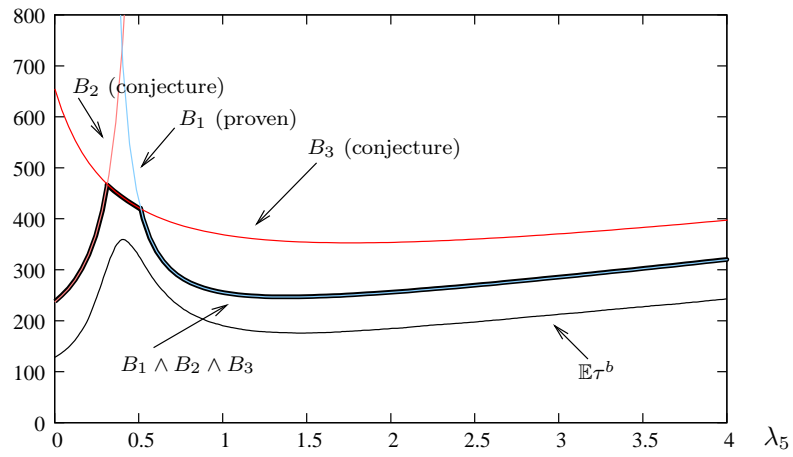**Figure 7.1:** *This figure displays the actual coupling time $\mathbb{E}\tau^b$ together with the proven light traffic bound B_1, the conjectured heavy traffic bound B_2, the conjectured critical bound B_3 and the minimum of the three bounds.*

Appendix A

Markov chain theory

This appendix summarises some results on discrete time and continuous time Markov chains. For a more detailed exposition, the reader is referred to handbooks on Markov chain theory like [1, 5, 9, 11].

Discrete time Markov chain

A Markov chain $\{X_n\}_{n \in \mathbb{N}}$ is a stochastic process in discrete time on a state space $\mathcal{S}$ with the property that

$$\mathbb{P}\big(X_{n+1} = s_{n+1} \mid X_0 = s_0, \ldots, X_n = s_n\big) = \mathbb{P}\big(X_{n+1} = s_{n+1} \mid X_n = s_n\big),$$

for every possible value of $s_0, \ldots s_{n+1} \in \mathcal{S}$. A Markov chain is called time-homogeneous if

$$p_{ij} = \mathbb{P}\big(X_{n+1} = j \mid X_n = i\big) \quad \text{for every } i, j \in \mathcal{S}. \quad \text{and for every } n \in \mathbb{N}.$$

The probabilities p_{ij} are the one-step transition probabilities. These one step probabilities satisfy

$$p_{ij} \geqslant 0 \quad \text{for } i, j \in \mathcal{S} \qquad \text{and} \qquad \sum_{j \in \mathcal{S}} p_{ij} = 1.$$

Let the matrix $P = (p_{ij})$ denote the matrix whose entries are the one step probabilities. From now one, we suppose that the Markov chains are time-homogeneous. The n-step probabilities p_{ij}^n are defined as

$$p_{ij}^n = \mathbb{P}\big(X_n = j \mid X_0 = i\big),$$

and denote the probability of going from state i to state j in exactly n steps.

A state i is accessible from state j if there exists an integer n such that $p_{ij}^n > 0$. State i and j *communicate* with each other if j is accessible from state i and i is accessible from j. Two states that communicate with each other are said to be in the same *class*. A Markov chain consisting of only one class is *irreducible*.

Let f_{ij}^n denote the first passage probability of state j, provided we start in state i. That is

$$f_{ij}^n = \mathbb{P}\left(X_n = j, \quad X_m \neq j \quad \text{for } 1 \leqslant m \leqslant n - 1 \mid X_0 = i\right).$$

Now

$$f_{ij} = \sum_{n=1}^{\infty} f_{ij}^n = \mathbb{P}\left(\text{there exists an } n \in \mathbb{N} : X_n = j \mid X_0 = i\right).$$

A state is called *recurrent* if $f_{ii} = 1$ and a state is *transient* if $f_{ii} < 1$. Thus, if a state recurrent, the Markov chain will visit state i infinitely often and when the state is transient, the state will be visited only a finite number of times. One can show that if $\sum_{n=0}^{\infty} p_{ii}^n = \infty$, then state i is recurrent and that if $\sum_{n=0}^{\infty} p_{ii}^n < \infty$, the state i is transient. Recurrence and transience are class properties.

A recurrent state i is called *positive recurrent* if

$$\sum_{n=1}^{\infty} n f_{ii}^n < \infty.$$

Thus for a recurrent state the expected time to return to state i is finite. Positive recurrence is also class property.

A state i has *period* d if $p_{ii}^n = 0$ whenever n is not divisible by d, and d is the largest integer with this property. A state with period 1 is said to be *aperiodic*. Periodicity is a class property.

In an irreducible Markov chain with finite state space $\mathcal{S}$, all states are positive recurrent.

A probability distribution $(\pi_j, j \in \mathcal{S})$ is called the *stationary distribution* of a Markov chain if

$$\pi_j = \sum_{i \in \mathcal{S}} \pi_i p_{ij}, \quad j \in \mathcal{S}.$$

This stationary distribution π_j of a state j can be interpreted as the long run proportion of time that the process is in state j. Thus $\pi_j = \lim_{n \to \infty} p_{ij}^n$.

One of the main results for finite-state Markov chains is the following:

Theorem A.1. *Let $\{X_n\}_{n \in \mathbb{N}}$ be a aperiodic and irreducible Markov chain on a finite state space. Then there exists a unique stationary distribution.*

Continuous time Markov processes

A stochastic process $\{X(t), t \geqslant 0\}$ is a continuous time Markov process if

(i) The amount of time spent in a state $i \in \mathcal{S}$ before a transition to an other state $j \in \mathcal{S}$ is exponentially distributed with parameter ν_i.

(ii) When the process leaves state i, it enters state j with some probability p_{ij}:

$$\begin{cases} p_{ii} & = & 0, & \text{for all } i \in \mathcal{S} \\ \sum_{j \neq i} p_{ij} & = & 1 & \text{for all } i \in \mathcal{S} \end{cases} \tag{A.1}$$

We suppose that the parameters ν_i are bounded and that the set $\mathcal{S}$ of all states is finite. The discrete time Markov chain $\{X_n\}_{n \in \mathbb{N}}$ with transition matrix $P = (p_{ij})$ is called the *embedded chain*. The transient probabilities

$$p_{ij}(t) = \mathbb{P}(X(t+s) = j \mid X(s) = i)$$

denote the probability that a process currently in state i will be in state j at t time units later.

Let $q_{ij} = \nu_i \, p_{ij}$ for $i \neq j$ denote the rate at which a process makes a transition from state i to state j. This rate is called the *infinitesimal transition rate*.

A probability distribution $(p_j, j \in \mathcal{S})$ is called the *stationary distribution* of a continuous time Markov chain if

$$\nu_j p_j = \sum_{k \neq j} q_{kj} p_k, \quad j \in \mathcal{S}.$$

This stationary distribution p_j of a state j can be interpreted as the long run proportion of time that the process is in state j. The main result is:

Theorem A.2. *Let $\{X(t), t \geqslant 0\}$ be a continuous time Markov process on a finite state space. Let the embedded Markov chain be irreducible and aperiodic. Then $X(t)$ has a stationary distribution p_j.*

A continuous time Markov process is *time reversible* if

$$p_i \, q_{ij} = p_j \, q_{ji} \qquad \text{for all } i, j \in \mathcal{S}.$$

This means that the rate at which the process goes directly from state i to state j is equal to the rate at which it goes directly from j to i.

Theorem A.3 (Burke's Theorem). *Consider an M/M/1 queue in steady state with arrival rate λ and service rate μ and $\lambda < \mu$. Then the departure process is a Poisson process with parameter λ.*

Proof. ([9], p. 378) In any interval of length t, the number of transitions from state i to $i+1$ must equal within one the number of transitions from $i+1$ to i. By letting $t \to \infty$, the number of transitions goes to infinity and the rate of transitions from i to $i+1$ equals the rate of transitions from $i+1$ to i. Thus the M/M/1 queue is time reversible. Let $N(t)$ count the number of customers in the M/M/1 queue. By going forward in time, the points where $N(t)$ increase by one, correspond with the arrivals of customers and thus is a Poisson process with parameter λ. Since the process is time reversible, the points at which the reversed process increases by one must also represent a Poisson process with parameter λ. These latter points are exactly the points when customers depart in the reversed process (see Figure A). Hence, the departure process is Poisson with parameter λ. $\qquad\square$
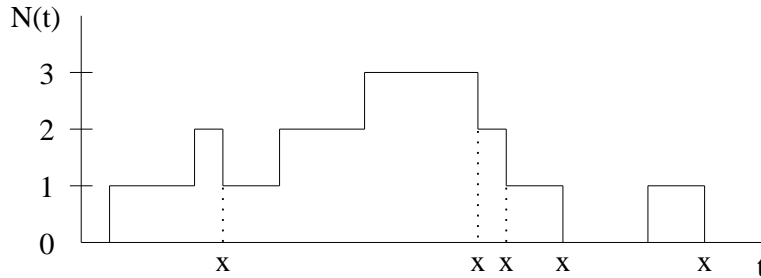

Figure A.1: *Realization of $N(t)$. The time points at which arrivals occur in the reversed process are indicated by* x.

Uniformization

The aim of uniformization is to construct for a given continuous time Markov chain a stochastic process such that the distribution of the sojourn time in a state is independent of the specific state one is in. Therefore, take a finite number ν such that $\nu \geqslant \nu_i$.

We define $\left\{\overline{X}_n\right\}_{n\in\mathbb{N}}$ a discrete time Markov chain whose transition matrix $\overline{P} = \left(\overline{p}_{ij}\right)$ is given by

$$\overline{p}_{ij} = \begin{cases} \frac{\nu_i}{\nu}\, p_{ij} & i \neq j; \\ 1 - \frac{\nu_i}{\nu} & i = j. \end{cases}$$

Note that the discrete time Markov chain $\overline{X}_n$ does allow transitions from a state to itself, whereas the the embedded chain X_n does not. Let $\{N(t), t \geqslant 0\}$ be

a Poisson process with parameter ν. Now define the continuous time stochastic process $\left\{\overline{X}\left(t\right), t \geqslant 0\right\}$ by

$$\overline{X}\left(t\right) = \overline{X}_{N(t)}, \qquad t \geqslant 0. \tag{A.2}$$

In other words, this stochastic process $\overline{X}\left(t\right)$ is driven by a Poisson process with parameter ν to determine *when* a transition takes place and a discrete time Markov chain $\overline{X}_n$ to determine *which* transition takes place. In fact, the transitions out of state i are delayed by a factor ν/ν_i while the time it takes until the next transition is condensed by a factor ν_i/ν. This explains that the continuous time Markov process $X\left(t\right)$ is probabilistically identical to the new constructed process $\overline{X}\left(t\right)$.

The uniformized model is thus driven by a single Poisson process and a discrete time Markov chain $\overline{X}_n$. The stationary distribution of the uniformized process is therefore determined by the stationary distribution of the discrete Markov chain $\overline{X}_n$. Because the model before uniformization and after uniformization are probabilistically identical, this means that the stationary distribution of the continuous time Markov process $\left\{X\left(t\right), t \geqslant 0\right\}$ is also determined by the stationary distribution of the discrete time Markov chain $\overline{X}_n$.

Appendix B

Useful theorems and results

This appendix consists of theorems and other results that are used in this thesis.

Strassen's Theorem

Let X, Y be random variables. We say that X is stochastically smaller than Y if $\mathbb{P}(X > t) \leqslant \mathbb{P}(Y > t)$ for every t. We note $X \leqslant_{st} Y$. Let $\mathcal{D}(\mathbb{R}^n)$ denote the set of probabilities on $\mathbb{R}^n$.

Theorem B.1 (Strassen's Theorem). *([2], p. 377-378) Let F and G be two cumulative distribution functions in $\mathcal{D}(\mathbb{R}^n)$. Now $F \leqslant_{st} G$ if and only if there exist two $\mathbb{R}^n$ random variables X and Y defined on the same probability space with probability distribution F and G respectively and such that $X \leqslant Y$ almost surely.*

Proof. For dimension one, the proof is obtained from the following construction: Let U be the random variable uniformly distributed on $[0, 1]$. Let $X = F^{-1}(U)$ and $Y = G^{-1}(U)$ with the inverse of F defined by $F^{-1}(u) = \inf\{x \text{ such that } F(x) > u\}$. It follows from the assumption $F \leqslant_{st} G$ that $X = F^{-1}(U) \leqslant G^{-1}(U) = Y$.

   On the other hand, if $X \leqslant Y$ almost surely, then $\mathbb{P}(X \leqslant x) = F(x) \geqslant G(x) = \mathbb{P}(Y \leqslant x)$ for all $x \geqslant 0$, and this is equivalent to $F \leqslant_{st} G$. $\qquad\square$

Random Walks

A random walk with absorbing barriers is a walk on $0, \ldots, m$ with p the probability of going up and $q = 1 - p$ the probability of going down. Let P_i denote the probability that absorption occurs in state m when starting in i and let $Q_i = 1 - P_i$ denote the probability that absorption occurs in 0.

Proposition B.1. *In a random walk on $0, \ldots m$, we have*

$$P_i = \begin{cases} \frac{1 - a^i}{1 - a^m} & \text{if } p \neq \frac{1}{2}, \\ \frac{i}{m} & \text{if } p = \frac{1}{2}, \end{cases}$$

with $a = q/p$.

Proof. ([1] p. 65-66) Conditioning delivers the following recurrent relation for P_i:

$$P_i = pP_{i+1} + qP_{i-1} \quad \text{for } i = 1, \ldots, m-1,$$

with the boundary conditions $P_0 = 0$ and $P_m = 1$. The characteristic polynomial is $px^2 - x + q = 0$.

In case $p \neq q$, the roots of the caracteristic polynomial are $r_1 = 1$ and $r_2 = q/p = a$. The general solution becomes:

$$P_i = Ar_1^i + Br_2^i = A + Ba^i, \tag{B.1}$$

with $A, B \in \mathbb{R}$ constants. These constants can be determined using the boundary conditions. We obtain that $A = -B = 1/(1-a^m)$ such that

$$P_i = \frac{1 - a^i}{1 - a^m} \quad \text{for } p \neq q.$$

In case $p = q$, there is only one root, namely $r = 1$. The general solution therefore is

$$P_i = Ar_1^i + Bir_1^i = A + Bi. \tag{B.2}$$

Again, we use the boundary conditions to determine A and B. We find that $A = 0$ and $B = 1/m$.

Thus:

$$P_i = \frac{i}{m} \quad \text{for } p = q.$$

$\square$

By setting $1 - P_i$ we obtain the next result:

Corollary B.1. *In a random walk on $0, \ldots, m$, we have*

$$Q_i = \begin{cases} \frac{a^m}{a^m - 1} \frac{1 - a^i}{} & \text{if } p \neq \frac{1}{2}, \\ \frac{m-i}{m} & \text{if } p = \frac{1}{2}, \end{cases}$$

with $a = q/p$.

Let T_i denote the absorption time of a random walk on $0, \ldots, m$ starting in i, and let α_0 denote the probability of absorption in 0.

Lemma B.1.

$$\mathbb{E}[T_i] = \begin{cases} \frac{1}{q-p} - \frac{m}{q-p} \frac{1 - a^i}{1 - a^m}, & p \neq \frac{1}{2}, \\ i(m - i), & p = \frac{1}{2}. \end{cases} \tag{B.3}$$

Proof. By a one step analysis, we get

$$\mathbb{E}[T_i] = 1 + p\mathbb{E}[T_{i+1}] + q\mathbb{E}[T_{i-1}], \quad \text{for } 1 \leqslant i \leqslant m - 1, \tag{B.4}$$

with the boundary conditions $\mathbb{E}[T_0] = \mathbb{E}[T_m] = 0$. The general solution equals (B.1) for $p \neq q$ and (B.2) for $p = q$, but this time we also need to find a particular solution since the factor $+1$ appears in the recurrence.

In case $p \neq q$, we try the particular solution $Ci + D$. Using (B.4), we find that $C = 1/(q-p)$ and $D = 0$. Thus the solution becomes:

$$A + Ba^i + \frac{i}{q - p}.$$

By sing the boundary conditions we get:

$$\mathbb{E}[T_i] = \frac{1}{q - p} - \frac{m}{q - p} \frac{1 - a^i}{1 - a^m}.$$

In case $p = q$, we try the particular solution $Ci^2 + Di + E$. Using (B.4), we find that $C = -1$ and $D = E = 0$. Then the solution becomes:

$$A + Bi - i^2.$$

By using the boundary conditions, we obtain:

$$\mathbb{E}\left[T_i\right] = i\left(m - 1\right).$$

$\square$

Catalan Numbers

The Catalan numbers are given by the recurrence

$$C_n = \sum_{k=0}^{n-1} C_k C_{n-1-k},$$

with the initial values $C_0 = C_1 = 1$.

Let $C(x)$ be the generating function of the Catalan numbers. Then

$$
\begin{aligned}
xC\left(x\right)^2 &= x\left(C_0 + C_1 x + C_2 x^2 + \dots\right)\left(C_0 + C_1 x + C_2 x^2 + \dots\right) \\
&= C_0^2 + \left(C_0 C_1 + C_1 C_0\right) x^2 + \left(C_0 C_2 + C_1 C_1 + C_2 C_0\right) x^3 + \dots \\
&= C\left(x\right) - 1.
\end{aligned}
$$

Solving for $C(x)$ yields:

$$C(x) = \frac{1 - \sqrt{1 - 4x}}{2x}.$$

Note that we choose the minus sign, since when choosing for the plus sign, then $\lim_{x \to 0} C(x) = \infty$. A direct expression for C_n is $\binom{2n}{n}\frac{1}{n+1}$. Hence,

$$C\left(x\right) = \sum_{n=0}^{\infty} \binom{2n}{n} \frac{x^n}{n+1}.$$

65

Bibliography

[1] Pierre Brémaud. *Markov chains: Gibb fields, Monte-Carlo simulation, and queues.* Springer-Verlag, 1999.

[2] François Baccelli and Pierre Brémaud. *Elements of Queueing Theory.* Springer, second edition, 2003.

[3] M. Crane and D.L. Iglehart. Simulating stable stochastic systems, iii: Regenerative processes and discrete-event simulation. *Operation Research*, 23:33–45, 1975.

[4] P. Glasserman and D.D. Yao. *Monotone Structure in Discrete-Event Systems.* Wiley Inter-Science, Series in Probability and Mathematical Statistics, 1994.

[5] O. Häggström. *Finite Markov Chains and Algorithmic Applications.* Cambridge University Press, second edition, 2002.

[6] S.G. Henderson and P.W. Glynn. Regenerative steady-state simulation of discrete-event systems. *ACM Trans. Model. Comput. Simul.*, 11(4):313–345, 2001.

[7] J. Propp and D. Wilson. Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Structures and Algorithms*, 9(1&2):223–252, 1996.

[8] Oren Patashnik Ronald L. Graham, Donald E. Knuth. *Concrete mathematics: a foundation in computer science.* Addison-Wesley, 1989.

[9] S. M. Ross. *Probability models.* Academic Press, eighth edition, 2003.

[10] W.J. Stewart. *Introduction to the Numerical Solution of Markov Chains.* Princeton, 1994.

[11] Henk C. Tijms. *A first course in stochastic models.* Wiley, 2003.

[12] J.-M. Vincent. Perfect simulation of monotone systems for rare event probability estimation. In *Winter Simulation Conference*, Orlando, dec 2005.

[13] J.-M. Vincent. Perfect simulation of queueing networks with blocking and rejection. In *Saint IEEE conference*, pages 268–271, Trento, 2005.

[14] J.-M. Vincent and C. Marchand. On the exact simulation of functionals of stationary Markov chains. *Linear Algebra and its Applications*, 386:285–310, 2004.