Richard J. Boucherie

# Markovian queueing networks

## Lecture notes LNMB course MQSN

September 5, 2020

Springer

# Contents

## Part I Solution concepts for Markovian networks of queues

# Part I
# Solution concepts for Markovian networks of queues

# Chapter 1
# Preliminaries

This chapter reviews and discusses the basic assumptions and techniques that will be used in this monograph. Proofs of results given in this chapter are omitted, but can be found in standard textbooks on Markov chains and queueing theory, e.g. [**?**, **?**, **?**, **?**, **?**, **?**, **?**]. Results from these references are used in this chapter without reference except for cases where a specific result (e.g. theorem) is inserted into the text.

## 1.1 Basic results for Markov chains

Consider a stochastic process $\{N(t),\ t \in T\}$ taking values in a countable state space $S$. Applications will usually assume that $S \subseteq \mathbb{N}_0^J$ and that $t$ represents time. Then a state $\mathbf{n} = (n_1, \ldots, n_J) \in S$ is a vector with components $n_i \in \mathbb{N}_0$, $i = 1, \ldots, J$. For a *discrete-time* stochastic process $T$ is the set of integers: $T = \mathbb{N}_0$, or $T = \mathbb{Z}$, whereas for a *continuous-time* stochastic process $T$ is the positive real line: $T = \mathbb{R}_0^+ = [0, \infty)$ or the real line $T = \mathbb{R}$. A vector $\mathbf{n} \in \mathbb{R}^J$ is called non-negative if $n_i \geq 0$, $i = 1, \ldots, J$, and positive if it is non-negative and non-null. In this monograph emphasis will be on continuous-time stochastic processes. Therefore, in the sequel all results are given for continuous-time stochastic processes only. The exposition in this section focusses on Markov chains with countable state space $S$. We will not impose further structure on the states $\mathbf{n} \in S$.

A stochastic process is a *stationary process* if $(N(t_1), N(t_2), \ldots, N(t_k))$ has the same distribution as $(N(t_1 + \tau), N(t_2 + \tau), \ldots, N(t_k + \tau))$ for all $k \in \mathbb{N}$, $t_1, t_2, \ldots, t_k \in T$, $\tau \in T$. The stochastic process $\{N(t),\ t \in T\}$ is a *Markov process* if for every $k \geq 1$, $t_1 < \cdots < t_k < t_{k+1}$, and any $\mathbf{n}_1, \ldots, \mathbf{n}_{k+1}$ in $S$, the joint distribution of $(N(t_1), \ldots, N(t_{k+1}))$ is such that

$$\begin{aligned} &\mathbb{P}\{N(t_{k+1}) = \mathbf{n}_{k+1} | N(t_1) = \mathbf{n}_1, \ldots, N(t_k) = \mathbf{n}_k\} \\ &\qquad\qquad = \mathbb{P}\{N(t_{k+1}) = \mathbf{n}_{k+1} | N(t_k) = \mathbf{n}_k\}, \end{aligned} \tag{1.1}$$

whenever the conditioning event $(N(t_1)=\mathbf{n}_1,\dots,N(t_k)=\mathbf{n}_k)$ has positive probability. In words, for a Markov process the state at a given time contains all information about the past evolution necessary to probabilistically predict the future evolution of the Markov process.

A Markov process is *time-homogeneous* if the conditional probability $\mathbb{P}\{N(s+t)=\mathbf{n}'|N(t)=\mathbf{n}\}$ is independent of $t$ for all $s>0$, $\mathbf{n},\mathbf{n}'\in S$. For a time-homogeneous Markov process the *transition probability* from state $\mathbf{n}$ to state $\mathbf{n}'$ in time $t$ is defined as

$$P(\mathbf{n},\mathbf{n}';t)=\mathbb{P}\left\{N(s+t)=\mathbf{n}'|N(s)=\mathbf{n}\right\},\quad s,t>0.$$

The *transition matrix* $P(t)=(P(\mathbf{n},\mathbf{n}';t),\ \mathbf{n},\mathbf{n}'\in S)$ has non-negative entries (1.2) and row sums equal to one (1.3). The *Markov property* (1.1) implies that the transition probabilities satisfy the *Chapman-Kolmogorov equations* (1.4). In addition, assume that the transition matrix is *standard* (1.5). For all $\mathbf{n},\mathbf{n}'\in S$, $s,t\in T$, a *standard transition matrix* satisfies:

$$P(\mathbf{n},\mathbf{n}';t)\geq 0; \tag{1.2}$$

$$\sum_{\mathbf{n}'\in S}P(\mathbf{n},\mathbf{n}';t)=1; \tag{1.3}$$

$$P(\mathbf{n},\mathbf{n}';s+t)=\sum_{\mathbf{n}''\in S}P(\mathbf{n},\mathbf{n}'';s)P(\mathbf{n}'',\mathbf{n}';t); \tag{1.4}$$

$$\lim_{t\downarrow 0}P(\mathbf{n},\mathbf{n}';t)=\delta_{\mathbf{n},\mathbf{n}'}. \tag{1.5}$$

$\delta_{\mathbf{n},\mathbf{n}'}$ is the *Kronecker-delta*, $\delta_{\mathbf{n},\mathbf{n}'}=1$ if $\mathbf{n}=\mathbf{n}'$ and $\delta_{\mathbf{n},\mathbf{n}'}=0$ if $\mathbf{n}\neq\mathbf{n}'$. For a standard transition matrix it is natural to extend the definition of $P(\mathbf{n},\mathbf{n}';\cdot)$ to $[0,\infty)$ by setting $P(\mathbf{n},\mathbf{n}';0)=\delta_{\mathbf{n},\mathbf{n}'}$. Then for all $\mathbf{n},\mathbf{n}'$ the transition probabilities are *uniformly continuous* on $[0,\infty)$. Furthermore, each $P(\mathbf{n},\mathbf{n}';t)$ is either identically zero for all $t>0$ or never zero for $t>0$ (Lévy's dichotomy [**?**, Theorem II.5.2]).

For a standard transition matrix the *transition rate* from state $\mathbf{n}$ to state $\mathbf{n}'$ can be defined as

$$q(\mathbf{n},\mathbf{n}')=\lim_{h\downarrow 0}\frac{P(\mathbf{n},\mathbf{n}';h)-\delta_{\mathbf{n},\mathbf{n}'}}{h}.$$

For all $\mathbf{n},\mathbf{n}'\in S$ this limit exists. For $\mathbf{n}\neq\mathbf{n}'$ this limit is finite (1.6), whereas for $\mathbf{n}=\mathbf{n}'$ the limit may be infinite. For practical systems the limit for $\mathbf{n}=\mathbf{n}'$ is finite too. In the sequel it is assumed that the limit exists for $\mathbf{n}=\mathbf{n}'$: (1.7). A Markov process is called a continuous-time *Markov chain* if for all $\mathbf{n},\mathbf{n}'\in S$ the limit exists and is finite (1.6), (1.7). In addition it is assumed that the rate matrix is *conservative* (1.8). Then for all $\mathbf{n},\mathbf{n}'$ the rate matrix satisfies

$$0 \le q(\mathbf{n},\mathbf{n}') < \infty, \quad \mathbf{n}' \neq \mathbf{n}; \tag{1.6}$$

$$0 \le q(\mathbf{n}) := -q(\mathbf{n},\mathbf{n}) < \infty; \tag{1.7}$$

$$\sum_{\mathbf{n}' \in S} q(\mathbf{n},\mathbf{n}') = 0. \tag{1.8}$$

For a rate matrix that satisfies (1.6), (1.7), the definition of the transition rates implies that the transition probabilities can be expressed in the transition rates. This gives, for $\mathbf{n},\mathbf{n}' \in S$,

$$P(\mathbf{n},\mathbf{n}';h) = \delta_{\mathbf{n},\mathbf{n}'} + q(\mathbf{n},\mathbf{n}')h + \mathrm{o}(h) \quad \text{for } h \downarrow 0, \tag{1.9}$$

where $\mathrm{o}(h)$ denotes a function $g(h)$ with the property that $g(h)/h \to 0$ as $h \downarrow 0$. For small positive values of $h$, for $\mathbf{n}' \neq \mathbf{n}$, the term $q(\mathbf{n},\mathbf{n}')h$ may be interpreted as the conditional probability, up to order $\mathrm{o}(h)$, that the Markov chain $\{N(t)\}$ makes a transition to state $\mathbf{n}'$ during $(t,t+h)$ given that the process is in state $\mathbf{n}$ at time $t$. From (1.7), (1.8), note that $q(\mathbf{n}) = \sum_{\mathbf{n}' \neq \mathbf{n}} q(\mathbf{n},\mathbf{n}')$. If $q(\mathbf{n})$ is finite, $q(\mathbf{n})h$ is the conditional probability that $\{N(t)\}$ leaves this state during $(t,t+h)$ given that $\{N(t)\}$ is in state $\mathbf{n}$ at time $t$. As a consequence, $q(\mathbf{n},\mathbf{n}')$ can be interpreted as the rate at which transitions occur, i.e., as transition rates. To elaborate on the transition rates and on the role of stability, consider the conditional probability that the process remains in $\mathbf{n}$ during $(s,s+h)$ if the process is in $\mathbf{n}$ at time $s$. This conditional probability is

$$\mathbb{P}\{N(\tau) = \mathbf{n},\ s < \tau < s+h | N(s) = \mathbf{n}\} = \mathrm{e}^{-q(\mathbf{n})h}, \quad h > 0.$$

The *exit-time* from state $\mathbf{n}$, $\varepsilon(\mathbf{n})$, defined as

$$\varepsilon(\mathbf{n}) = \inf\{t :\ t > 0,\ N(t+s) \neq \mathbf{n}\}$$

given that the process is in state $\mathbf{n}$ at time $s$, has a negative-exponential distribution with mean $q(\mathbf{n})^{-1}$.

For every initial state $N(0) = \mathbf{n}$, $\{N(t),\ t \in T\}$ is a *pure-jump process*, which means that the process jumps from state to state and remains in each state a *strictly positive* sojourn-time with probability 1. For the Markovian case, the process remains in state $\mathbf{n}$ for a negative-exponentially distributed sojourn-time with mean $q(\mathbf{n})^{-1}$. In addition, conditional on the process departing from state $\mathbf{n}$ it jumps to state $\mathbf{n}'$ with probability $p(\mathbf{n},\mathbf{n}') = q(\mathbf{n},\mathbf{n}')/q(\mathbf{n})$. This second interpretation is sometimes used as a definition of a continuous-time Markov chain and is used to construct such processes. The Markov chain represented via the holding times $q(\mathbf{n})$ and transition probabilities $p(\mathbf{n},\mathbf{n}')$, $\mathbf{n},\mathbf{n}' \in S$, is referred to as the *Markov jump chain* of the Markov chain $\{N(t)\}$. Note that we obtain the Markov chain with transition rates $q(\mathbf{n},\mathbf{n}')$ from the Markov jump chain with holding times with mean $q(\mathbf{n})^{-1}$ and transition probabilities $p(\mathbf{n},\mathbf{n}')$ as $q(\mathbf{n},\mathbf{n}') = q(\mathbf{n})p(\mathbf{n},\mathbf{n}')$, $\mathbf{n},\mathbf{n}' \in S$.

From the Chapman-Kolmogorov equations two systems of differential equations for the transition probabilities can be obtained. To this end, observe that for a standard transition matrix every element $P(\mathbf{n},\mathbf{n}';\cdot)$ has a continuous derivative in $(0,\infty)$, which is continuous at zero if the rate matrix satisfies (1.6), (1.7) [**?**, Theorem II.12.8]. Conditioning on the first jump of the Markov chain in $(0,t]$ yields the so-called *Kolmogorov backward equations* (1.10), whereas conditioning on the last jump in $(0,t]$ gives the *Kolmogorov forward equations* (1.11). The validity of this method is discussed below. These equations read for $\mathbf{n},\mathbf{n}' \in S$, $t \geq 0$,

$$\frac{dP(\mathbf{n},\mathbf{n}';t)}{dt} = \sum_{\mathbf{n}''\in S} q(\mathbf{n},\mathbf{n}'')P(\mathbf{n}'',\mathbf{n}';t), \tag{1.10}$$

$$\frac{dP(\mathbf{n},\mathbf{n}';t)}{dt} = \sum_{\mathbf{n}''\in S} P(\mathbf{n},\mathbf{n}'';t)q(\mathbf{n}'',\mathbf{n}'). \tag{1.11}$$

If the rate matrix satisfies (1.6), (1.7), then starting from the initial state $N(0) = \mathbf{n}$, a first jump of the Markov chain exists for $t > 0$. As a consequence conditioning on this first jump is allowed. In contrast, the last jump of the Markov chain in $(0,t]$ is not properly defined. It may be that also for a rate matrix that satisfies (1.6), (1.7) jumps will accumulate in such a way that $\{N(t)\}$ will make infinitely many jumps in finite time. In this case $\{N(t)\}$ is not properly defined from the rate matrix for all $t > 0$.

**Example 1.1.1 (Explosion in a pure birth process)** Consider the Markov chain $\{N(t),\ t \in [0,\infty)\}$, at state space $S = \mathbb{N}_0$ with transition rates

$$q(\mathbf{n},\mathbf{n}') = \begin{cases} q(\mathbf{n}), & \text{if } \mathbf{n}' = \mathbf{n}+1, \\ -q(\mathbf{n}), & \text{if } \mathbf{n}' = \mathbf{n}, \\ 0, & \text{otherwise,} \end{cases}$$

with initial distribution $\mathbb{P}(N(0) = \mathbf{n}) = \delta(\mathbf{n},0)$. Then $\{N(t)\}$ is a pure birth process that spends a negative-exponentially distributed time with rate $q(\mathbf{n})$ in state $\mathbf{n}$ and then jumps to state $\mathbf{n}+1$ with probability 1, $\mathbf{n} \in S$. Let $\xi(\mathbf{n})$ denote the time spent in state $\mathbf{n}$, and $\xi = \sum_{\mathbf{n}=0}^{\infty} \xi(\mathbf{n})$ the time spent in the states $0,1,2,\ldots$. Let $q(\mathbf{n}) = 2^{\mathbf{n}}$, then

$$\mathbb{E}\{\xi\} = \sum_{\mathbf{n}=0}^{\infty} \mathbb{E}\{\xi(\mathbf{n})\} = \sum_{\mathbf{n}=0}^{\infty} 2^{-\mathbf{n}} = 2$$

by monotone convergence. As $\mathbb{E}\{\xi\} < \infty$ it must be that $\mathbb{P}(\xi < \infty) = 1$ and therefore $\{N(t)\}$ is explosive (diverges to infinity in finite time).[1] □

An additional assumption on the rate matrix guaranteeing the existence of a last jump in $(0,t]$ is regularity. A pure-jump Markov chain is *regular* if for every initial state $N(0) = \mathbf{n}$ the number of transitions in finite time is finite with probability 1. For a regular Markov chain the last jump before $t$ is well-defined and conditioning on the last jump before $t$ is allowed. Thus if a pure-jump Markov chain satisfies (1.6), (1.7) and is regular, then for all $t > 0$ the evolution of the process is uniquely determined

[1] We may actually show the following stronger result: *A pure birth process is explosive if and only if* $\sum_{\mathbf{n}=0}^{\infty} q(\mathbf{n})^{-1} < \infty$.

by the transition rates, that is specification of the transition rates is sufficient to completely characterize the process.

Regularity is a property of the rate matrix. It can be shown [**?**] that the rate matrix is regular if and only for some $\nu > 0$ the system of equations

$$\sum_{\mathbf{n}' \in S} q(\mathbf{n}, \mathbf{n}') x(\mathbf{n}') = \nu x(\mathbf{n}), \quad \mathbf{n} \in S,$$

has no bounded solution other than $\{x(\mathbf{n}) = 0,\ \mathbf{n} \in S\}$. This characterization of regularity may be difficult to apply in practical situations. A simple sufficient condition ensuring regularity of a Markov chain is the existence of a uniform finite upper bound on $q(\mathbf{n})$. If such a bound exists, i.e., if a constant $C$ exists such that for all $\mathbf{n} \in S$

$$q(\mathbf{n}) \leq C < \infty,$$

then the Markov chain is said to be *uniformizable* and the forward and backward equations have the same solution. Uniformizability can be too strong for practical applications as it excludes, for example, the infinite-server queue (see Example 2.3.1). More general sufficient conditions can be found in, e.g., [**?**, Section 4-3]. A detailed discussion of regularity is beyond the scope of this monograph. The behaviour of irregular Markov chains is, for example, discussed in [**?**, **?**].

The following theorem summarizes the results on regularity and the forward and backward equations stated above.

**Theorem 1.1.2 ([?, Theorem II.18.3])** *For a conservative, regular, continuous-time Markov chain the forward equations (1.11) and the backward equations (1.10) have the same unique solution* $\{P(\mathbf{n}, \mathbf{n}'; t),\ \mathbf{n}, \mathbf{n}' \in S,\ t \geq 0\}$. *Moreover, this unique solution is the transition matrix of the Markov chain.*

In particular, Theorem 1.1.2 states that either the forward or the backward equations can be solved to find the transition matrix

$$P(t) = \mathrm{e}^{Qt} = \sum_{n=0}^{\infty} \frac{(Qt)^n}{n!}, \quad t \geq 0.$$

Usually the forward equations are easier to use in practical cases as they allow for an interpretation using probability fluxes (see below).

For any *initial distribution* $\{p_{(0)}(\mathbf{n}),\ \mathbf{n} \in S\}$ defined as

$$p_{(0)}(\mathbf{n}) = \mathbb{P}\{N(0) = \mathbf{n}\}, \quad \sum_{\mathbf{n} \in S} p_{(0)}(\mathbf{n}) = 1,$$

the *time-dependent distribution* $\{p(\mathbf{n}, t),\ \mathbf{n} \in S\}$ defined as

$$p(\mathbf{n}, t) = \mathbb{P}\{N(t) = \mathbf{n}\}, \quad \sum_{\mathbf{n} \in S} p(\mathbf{n}, t) = 1,$$

can be obtained from the forward equations (1.11). Pre-multiplication of the forward equations (1.11) with the initial distribution $\{p_{(0)}(\mathbf{n}),\ \mathbf{n} \in S\}$ gives for the time-dependent distribution the following version of the *Kolmogorov forward equations* for $\mathbf{n}' \in S$, $t \geq 0$,

$$\begin{cases} \dfrac{dp(\mathbf{n}',t)}{dt} = \displaystyle\sum_{\mathbf{n}\neq\mathbf{n}'} \left\{ p(\mathbf{n},t)q(\mathbf{n},\mathbf{n}') - p(\mathbf{n}',t)q(\mathbf{n}',\mathbf{n}) \right\}, \\ \\ p(\mathbf{n}',0) = p_{(0)}(\mathbf{n}'). \end{cases} \tag{1.12}$$

From the interpretation of the transition rates obtained from (1.9), for $\mathbf{n} \neq \mathbf{n}'$, the probability that the process jumps from $\mathbf{n}$ to $\mathbf{n}'$ in the interval $(t,t+h)$ is $p(\mathbf{n},t)q(\mathbf{n},\mathbf{n}')h + \mathrm{o}(h)$. Therefore, $p(\mathbf{n},t)q(\mathbf{n},\mathbf{n}')$ may be called the *probability flux* or *probability flow* from state $\mathbf{n}$ to state $\mathbf{n}'$. The forward equations now express that the rate of change of the *probability mass* of state $\mathbf{n}'$, $\frac{dp(\mathbf{n}',t)}{dt}$, equals the net probability flux from $S \setminus \{\mathbf{n}'\}$ to $\mathbf{n}'$. Thus the Kolmogorov forward equations express an intuitively obvious relation for the time-dependent probabilities. A similar straightforward interpretation of the backward equations is not available.

**Remark 1.1.3 (Uniformization)** The *embedded Markov chain* of $\{N(t),\ t \in \mathbb{R}_0^+\}$ is the discrete-time Markov chain $\{Y(t),\ t \in \mathbb{N}_0\}$ at state space $S$ with transition probabilities $p(\mathbf{n},\mathbf{n}') = q(\mathbf{n},\mathbf{n}')/q(\mathbf{n})$, $\mathbf{n},\mathbf{n}' \in S$, that follows the transitions of $\{N(t)\}$. If $q(\mathbf{n}) = q$ for all $\mathbf{n} \in S$ then $\{N(t)\}$ makes transitions at constant rate $q$ and the state after $k$ transitions is determined by the $k$-step transition probabilities of $\{Y(t)\}$.

If $\{N(t)\}$ is uniformizable with $\sup_{\mathbf{n}\in S} q(\mathbf{n}) \leq C < \infty$ we may define the discrete-time Markov chain $\{X(t),\ t \in \mathbb{N}_0\}$ at state space $S$ with transition probabilities, for $\mathbf{n},\mathbf{n}' \in S$,

$$p_u(\mathbf{n},\mathbf{n}') = \begin{cases} q(\mathbf{n},\mathbf{n}')/C, & \text{if } \mathbf{n}' \neq \mathbf{n}, \\ 1 - q(\mathbf{n})/C, & \text{if } \mathbf{n}' = \mathbf{n}. \end{cases}$$

Note that for $p_u(\mathbf{n},\mathbf{n}') = p(\mathbf{n},\mathbf{n}')q(\mathbf{n})/C$ for $\mathbf{n}' \neq \mathbf{n}$. Thus, $\{X(t)\}$ is an embedded Markov chain with transitions occurring at the event times of a Poisson process with rate $C$. In state $\mathbf{n} \in S$ with probability $1 - q(\mathbf{n})/C$ the Markov chain makes a self-transition, and with probability $q(\mathbf{n})/C$ the Markov chain makes a transition to another state, and this state is $\mathbf{n}'$ with probability $p(\mathbf{n},\mathbf{n}')$.[2] Let $P_u = (p_u(\mathbf{n},\mathbf{n}'),\ \mathbf{n},\mathbf{n}' \in S)$. Then for all $\mathbf{n},\mathbf{n}' \in S$ and $t > 0$

$$P(t) = \sum_{k=0}^{\infty} \frac{(Ct)^k}{k!} \mathrm{e}^{-Ct} (P_u)^k. \tag{1.13}$$

Uniformization transfers the continuous-time Markov chain $\{N(t)\}$ into the discrete-time Markov chain $\{X(t)\}$. Evaluation of $P(t)$ for fixed $t$ via (1.13) is efficient as $(P_u)^k$ can be computed efficiently. Observe, however, that the sum must be evaluated

[2] Observe that for $\{X(t)\}$ the exit-time from state $\mathbf{n}$ is $\varepsilon(\mathbf{n}) = \sum_{k=1}^{K} X_k$, where $K$ has a geometric distribution with succes probability $q(\mathbf{n})/C$, and the $X_k$, $k = 1,2,\ldots$, are i.i.d. negative-exponentially distributed with rate $C$. Hence $\varepsilon(\mathbf{n})$ has a negative-exponential distribution with rate $q(\mathbf{n})$.

for each $t$ separately, so that uniformization does not provide an elegant construction for $P(t)$ for all $t$. See [**?**] for a survey on uniformization. □

The remaining part of this section considers the stationary or equilibrium behaviour of Markov chains. Throughout it will be assumed that the rate matrix satisfies (1.6), (1.7), is conservative and regular. Although these assumptions are not necessary for a large part of the discussd67dion below, the discussion particularizes to conservative, regular Markov chains when the stationary distribution is related to the invariant distribution (the equilibrium solution of the Kolmogorov forward equations). When the assumptions are crucial to the theory they will be explicitly repeated.

If $P(t) = (p(\mathbf{n},\mathbf{n}';t),\ \mathbf{n},\mathbf{n}' \in S)$ is a transition matrix then the following limit exists for all $\mathbf{n},\mathbf{n}' \in S$

$$\lim_{t\to\infty} p(\mathbf{n},\mathbf{n}';t) = \upsilon(\mathbf{n},\mathbf{n}').$$

The matrix $\Upsilon = (\upsilon(\mathbf{n},\mathbf{n}'),\ \mathbf{n},\mathbf{n}' \in S)$ satisfies for all $\mathbf{n},\mathbf{n}' \in S$, $s > 0$,

$$\begin{aligned}\upsilon(\mathbf{n},\mathbf{n}') &= \sum_{\mathbf{n}''\in S} \upsilon(\mathbf{n},\mathbf{n}'')p(\mathbf{n}'',\mathbf{n}';s)\\ &= \sum_{\mathbf{n}''\in S} p(\mathbf{n},\mathbf{n}'';s)\upsilon(\mathbf{n}'',\mathbf{n}') = \sum_{\mathbf{n}''\in S} \upsilon(\mathbf{n},\mathbf{n}'')\upsilon(\mathbf{n}'',\mathbf{n}').\end{aligned}$$

Furthermore, $\upsilon(\mathbf{n},\mathbf{n}') \geq 0$ for all $\mathbf{n},\mathbf{n}' \in S$, and if $\upsilon(\mathbf{n},\mathbf{n}) \neq 0$ then $\sum_{\mathbf{n}'\in S} \upsilon(\mathbf{n},\mathbf{n}') = 1$. Therefore, $\Upsilon$ characterizes the stationary behaviour, but cannot be immediately associated with the stationary distribution. For $\Upsilon$ to be the stationary distribution additional assumptions guaranteeing that $\upsilon(\mathbf{n},\mathbf{n}) \neq 0$ must be made.

A state $\mathbf{n}$ is *absorbing* if the process cannot leave state $\mathbf{n}$, that is $p(\mathbf{n},\mathbf{n};t) = 1$ for all $t \geq 0$. For a non-absorbing state $\mathbf{n}$ the *recurrence-time* $\tau(\mathbf{n})$ is defined as

$$\tau(\mathbf{n}) = \inf\{t:\ t > \varepsilon(\mathbf{n}),\ N(t) = \mathbf{n} \text{ if } N(0) = \mathbf{n}\},$$

where $\varepsilon(\mathbf{n})$ is the exit-time from state $\mathbf{n}$. $\tau(\mathbf{n})$ is the time it takes the process to return to state $\mathbf{n}$ if it starts at $\mathbf{n}$. A state $\mathbf{n}$ is called *recurrent* if recurrence to $\mathbf{n}$ is certain, i.e., if $\mathbb{P}\{\tau(\mathbf{n}) < \infty\} = 1$. Otherwise it is *transient*. A recurrent state is *positive-recurrent* if $\mathbb{E}\{\tau(\mathbf{n})\} < \infty$, that is if the expected return-time to state $\mathbf{n}$ is finite. Otherwise it is *null-recurrent*.

State $\mathbf{n}$ is *reachable* from state $\mathbf{n}'$ if passage from $\mathbf{n}$ to $\mathbf{n}'$ is possible, that is if $P(\mathbf{n},\mathbf{n}';t) > 0$ for some positive $t$. Two states *communicate* if each one is reachable from the other. A set $V \subset S$ is *closed* if the process cannot leave $V$, so that $q(\mathbf{n},\mathbf{n}') = 0$ for $\mathbf{n} \in V$, $\mathbf{n}' \in S \setminus V$. A set $V \subset S$ is *irreducible* if it is closed and all its states communicate. Two irreducible sets are disjoint, so the state space $S$ can be decomposed into disjoint irreducible sets $V_1$, $V_2, \ldots$, and a non-irreducible set $W$. For the equilibrium behaviour of $\{N(t)\}$ the process may be analysed at each irreducible set separately. Therefore, without loss of generality, for equilibrium analysis the Markov chain may be assumed irreducible at $S$, that is $S$ is an irreducible set. In this case all states $\mathbf{n} \in S$ are of the same type (transient, null-recurrent, positive-recurrent).

A measure $m = (m(\mathbf{n}),\ \mathbf{n} \in S)$ such that $0 \leq m(\mathbf{n}) < \infty$ for all $\mathbf{n} \in S$ and $m(\mathbf{n}) > 0$ for some $\mathbf{n} \in S$ is called a *stationary measure* if for all $\mathbf{n}' \in S$, $t \geq 0$,

$$m(\mathbf{n}') = \sum_{\mathbf{n} \in S} m(\mathbf{n}) P(\mathbf{n}, \mathbf{n}'; t),$$

and is called an *invariant measure* if for all $\mathbf{n} \in S$,

$$\sum_{\mathbf{n}' \neq \mathbf{n}} \left\{ m(\mathbf{n}) q(\mathbf{n}, \mathbf{n}') - m(\mathbf{n}') q(\mathbf{n}', \mathbf{n}) \right\} = 0. \tag{1.14}$$

The relation between stationary and invariant measures is rather complicated [**?**]. Based on regularity of the rate matrix a simple relation between these measures can be obtained. If the Markov chain is irreducible and positive-recurrent at $S$ then there exists a unique (up to a multiplicative factor) stationary measure $m$ which is positive ($m(\mathbf{n}) > 0$ for all $\mathbf{n} \in S$). From this result, for a regular and irreducible pure-jump process, if a finite mass ($\sum_{\mathbf{n} \in S} m(\mathbf{n}) < \infty$) invariant measure $m$ exists then the process is positive-recurrent and $m$ is the unique stationary measure. In the literature, an irreducible positive-recurrent process with invariant measure having finite mass is called *ergodic*.

Ergodicity is an important property of a process as it guarantees the existence of a unique *stationary distribution* $\pi$, that is a stationary measure summing to unity. Furthermore, if $\{N(t)\}$ is ergodic and $\pi$ is the stationary distribution then $P(\mathbf{n}, \mathbf{n}'; t) \to \pi(\mathbf{n}')$ $(t \to \infty)$ for all $\mathbf{n}, \mathbf{n}' \in S$, or equivalently, $P(\mathbf{n}, t) \to \pi(\mathbf{n})$ $(t \to \infty)$ for all $\mathbf{n} \in S$ for any initial distribution $P_0$. As a consequence $\pi$ may be called *equilibrium distribution*. Moreover, if $\{N(t)\}$ is ergodic then for any $f : S \to [0, \infty)$ such that $\sum_{\mathbf{n} \in S} f(\mathbf{n}) \pi(\mathbf{n}) < \infty$, with probability 1

$$\lim_{T \to \infty} \frac{1}{T} \int_0^T f(N(t)) dt = \mathbb{E}_\pi \{ f(N) \} \equiv \sum_{\mathbf{n} \in S} \pi(\mathbf{n}) f(\mathbf{n}).$$

In particular, for $f(N(t)) = \mathbb{1}\{N(t) = \mathbf{n}\}$, the *indicator* of the event $\{N(t) = \mathbf{n}\}$, i.e., $\mathbb{1}\{A\} = 1$ if $A$ occurs and 0 otherwise,

$$\lim_{T \to \infty} \frac{1}{T} \int_0^T \mathbb{1}\{N(t) = \mathbf{n}\} dt = \pi(\mathbf{n}).$$

Thus $\pi(\mathbf{n})$ is the *long-run fraction of time* the process spends in state $\mathbf{n}$. The result may be extended to a function $h : S \times S \to [0, \infty)$ on the transitions of $\{N(t)\}$. If $\sum_{\mathbf{n}, \mathbf{n}' \in S} \pi(\mathbf{n}) q(\mathbf{n}, \mathbf{n}') h(\mathbf{n}, \mathbf{n}') < \infty$, then with probability 1

$$\lim_{T \to \infty} \frac{1}{T} \sum_{k=1}^{\infty} h(N(\tau_{k-1}), N(\tau_k)) \mathbb{1}(\tau_k \in (0, T]) = \sum_{\mathbf{n}, \mathbf{n}' \in S} \pi(\mathbf{n}) q(\mathbf{n}, \mathbf{n}') h(\mathbf{n}, \mathbf{n}'), \tag{1.15}$$

where $0 = \tau_0 < \tau_1 < \tau_2 < \cdots$ are the transition epochs of $\{N(t)\}$. Conditions for the process to be ergodic can be found, for example, in [**?**, **?**].

The following theorem summarizes the relation between stationary, invariant and equilibrium distributions, and is the basis for determining the stationary or equilibrium distribution.

**Theorem 1.1.4 (Equilibrium distribution)** *Let $\{N(t),\ t \geq 0\}$ be a conservative, regular, irreducible continuous-time Markov chain.*

*(i) If a positive finite mass invariant measure $m$ exists then the Markov chain is positive-recurrent (ergodic). In this case $\pi = (\pi(\mathbf{n}),\ \mathbf{n} \in S)$ defined as $\pi(\mathbf{n}) = m(\mathbf{n}) \left[\sum_{\mathbf{n}\in S} m(\mathbf{n})\right]^{-1}$, $\mathbf{n} \in S$, is the unique stationary distribution and $\pi$ is the equilibrium distribution, i.e., for all $\mathbf{n}, \mathbf{n}' \in S$,*

$$\lim_{t\to\infty} P(\mathbf{n},\mathbf{n}';t) = \pi(\mathbf{n}'),$$

*independent of the initial distribution.*

*(ii) If a positive finite mass invariant measure does not exist then for all $\mathbf{n}, \mathbf{n}' \in S$,*

$$\lim_{t\to\infty} P(\mathbf{n},\mathbf{n}';t) = 0.$$

The main result of Theorem 1.1.4 is that the stationary or equilibrium distribution can be obtained as the unique probability solution to (1.14). The equations (1.14) for $m = \pi$, the invariant distribution, can be obtained from the Kolmogorov forward equations. To this end note that the transition matrix $P(t)$ is the unique solution to (1.11). Furthermore, for a standard transition matrix $\frac{dP(\mathbf{n},\mathbf{n}';t)}{dt} \to 0$ $(t \to \infty)$ for all $\mathbf{n}, \mathbf{n}' \in S$. Thus for $t \to \infty$ (1.11) reduces to (1.14). Similar to the interpretation of (1.12), the equations (1.14) for $m = \pi$ can be interpreted as balancing the flow of probability mass on $S$. To this end $\pi(\mathbf{n})$ is interpreted as the probability mass at state $\mathbf{n}$ and $q(\mathbf{n},\mathbf{n}')$ as the conductance of the direct path from $\mathbf{n}$ to $\mathbf{n}'$. Then $\pi(\mathbf{n})q(\mathbf{n},\mathbf{n}')$ is the flux of probability mass from $\mathbf{n}$ to $\mathbf{n}'$ and (1.14) states that the flow of probability mass leaving $\mathbf{n}$ is balanced by the flow of probability mass entering $\mathbf{n}$. Therefore, (1.14) is usually referred to as *global balance equations*.

## 1.2 Three solution concepts

This section introduces three approaches to obtain the stationary or equilibrium distribution that will form the basis for the analysis in Chapters 2, 3, and 4, respectively: reversibility, partial balance, and Kelly's lemma.

**Assumption 1.2.1** *Throughout this monograph, let $\{N(t),\ t \geq 0\}$ be a conservative, ergodic, continuous-time Markov chain with initial distribution $\mathbb{P}(\mathbf{n},0) = \pi(\mathbf{n})$, $\mathbf{n} \in S$. Let $N$ be the random variable recording the state of $\{N(t),\ t \geq 0\}$ in equilibrium with distribution $\pi$.*

As is discussed in Section 1.1, under Assumption 1.2.1 the equilibrium distribution or stationary distribution, $\pi = (\pi(\mathbf{n}),\ \mathbf{n} \in S)$, can be obtained as the unique solution to the *global balance equations*

$$\sum_{\mathbf{n}'\neq\mathbf{n}} \left\{\pi(\mathbf{n})q(\mathbf{n},\mathbf{n}') - \pi(\mathbf{n}')q(\mathbf{n}',\mathbf{n})\right\} = 0, \quad \mathbf{n} \in S, \tag{1.16}$$

also called full balance equations or total balance equations as these equations express balance of the total probability flow in and out of each state $\mathbf{n}$. Solving the global balance equations is often very hard. Almost all solutions available in literature satisfy more stringent balance relations.

Note that under Assumption 1.2.1 the Markov chain $\{N(t),\ t \geq 0\}$ is stationary:

**Theorem 1.2.2** *If Markov chain* $\{N(t),\ t \geq 0\}$ *has initial distribution* $\mathbb{P}(\mathbf{n},0) = \pi(\mathbf{n})$*,* $\mathbf{n} \in S$*, then* $\{N(t),\ t \geq 0\}$ *is stationary and* $\mathbb{P}(\mathbf{n},t) = \pi(\mathbf{n})$*,* $\mathbf{n} \in S$*, for all* $t \geq 0$*.*

### 1.2.1 Reversibility

The most stringent balance relation is transition balance. A Markov chain satisfies *transition balance* if for all $\mathbf{n},\mathbf{n}' \in S$ the transition rate from $\mathbf{n}$ to $\mathbf{n}'$ equals the transition rate from $\mathbf{n}'$ to $\mathbf{n}$, that is for all $\mathbf{n},\mathbf{n}' \in S$

$$q(\mathbf{n},\mathbf{n}') = q(\mathbf{n}',\mathbf{n}).$$

If a Markov chain satisfies transition balance then $m(\mathbf{n}) = 1$ for all $\mathbf{n} \in S$ satisfies the global balance equations (1.16). The equilibrium distribution $\pi$ exists only if $S$ is finite, in which case $\pi(\mathbf{n}) = |S|^{-1}$, $\mathbf{n} \in S$, with $|S|$ the *cardinality* of $S$.

A less restrictive form of balance often encountered in physical systems is detailed balance [**?**, **?**, **?**]. A Markov chain satisfies *detailed balance* if a distribution $\pi = (\pi(\mathbf{n}),\ \mathbf{n} \in S)$ exists that satisfies the *detailed balance equations* (1.17), for all $\mathbf{n},\mathbf{n}' \in S$,

$$\pi(\mathbf{n})q(\mathbf{n},\mathbf{n}') - \pi(\mathbf{n}')q(\mathbf{n}',\mathbf{n}) = 0. \tag{1.17}$$

Detailed balance is an important equilibrium concept. Summing (1.17) over all $\mathbf{n}' \in S$ yields that a distribution $\pi$ that satisfies the detailed balance equations is the stationary distribution. The detailed balance equations state that the probability flow between each pair of states is balanced.

Detailed balance is related to reversibility. A stochastic process $\{N(t),\ -\infty < t < \infty\}$ is *reversible* if $(N(t_1), N(t_2), \ldots, N(t_n))$ has the same distribution as $(N(\tau - t_1), N(\tau - t_2), \ldots, N(\tau - t_n))$ for all $n \in \mathbb{N}$, $t_1, t_2, \ldots, t_n \in \mathbb{R}$, $\tau \in \mathbb{R}$. If a stochastic process is reversible and the direction of time is reversed, then the probabilistic behaviour of the process remains the same. The algebraic detailed balance property and the probabilistic reversibility property are the basis for the analysis in Chapter 2.

**Theorem 1.2.3 (Reversibility and detailed balance)** *Let* $\{N(t),\ t \in T\}$*,* $T = \mathbb{R}$*, be a stationary Markov chain with transition rates* $q(\mathbf{n},\mathbf{n}')$*,* $\mathbf{n},\mathbf{n}' \in S$*.* $\{N(t)\}$ *is reversible if and only if there exists a distribution* $\pi = (\pi(\mathbf{n}),\ \mathbf{n} \in S)$ *that satisfies the* *detailed balance equations. When there exists such a distribution $\pi$, then $\pi$ is the equilibrium distribution of $\{N(t)\}$.*

**Proof.** See Chapter 2. □

### 1.2.2 Partial balance

Partial balance is less restrictive than detailed balance. Define for $\mathbf{n} \in S$ a collection of mutually exclusive sets $\{A_k(\mathbf{n}),\ k \in I(\mathbf{n})\}$, $I(\mathbf{n}) \subseteq \mathbb{N}$, such that $\bigcup_{k\in I(\mathbf{n})} A_k(\mathbf{n}) = S$. A Markov chain is *partially balanced over* $\{A_k(\mathbf{n}),\ k \in I(\mathbf{n})\}$ if a distribution $\pi = (\pi(\mathbf{n}),\ \mathbf{n} \in S)$ exists such that for all $\mathbf{n} \in S$, $k \in I(\mathbf{n})$,

$$\sum_{\mathbf{n}'\in A_k(\mathbf{n})} \left\{\pi(\mathbf{n})q(\mathbf{n},\mathbf{n}') - \pi(\mathbf{n}')q(\mathbf{n}',\mathbf{n})\right\} = 0. \tag{1.18}$$

The following result follows by summation of (1.18) over $k \in I(\mathbf{n})$.

**Theorem 1.2.4 (Partial balance)** *A distribution $\pi = (\pi(\mathbf{n}),\ \mathbf{n} \in S)$ satisfying the partial balance equations (1.18) is a stationary distribution.*

Chapter 3 explores partial balance as a means to obtain the equilibrium distribution of Markov chains.

### 1.2.3 Kelly's lemma

The transition rates of the time-reversed Markov chain are given in the following theorem.

**Theorem 1.2.5** *Let $\{N(t),\ t \in T\}$, $T = \mathbb{R}$, be a stationary Markov chain with transition rates $q(\mathbf{n},\mathbf{n}')$, $\mathbf{n},\mathbf{n}' \in S$ and equilibrium distribution $\pi = (\pi(\mathbf{n}),\ \mathbf{n} \in S)$. The time-reversed process $\{N(\tau - t),\ t \in T\}$ is a conservative, regular, irreducible continuous-time stationary Markov chain with transition rates $q^r(\mathbf{n},\mathbf{n}')$, $\mathbf{n},\mathbf{n}' \in S$, given by*

$$q^r(\mathbf{n},\mathbf{n}') = \frac{\pi(\mathbf{n}')}{\pi(\mathbf{n})}q(\mathbf{n}',\mathbf{n})$$

*and the same equilibrium distribution $\pi = (\pi(\mathbf{n}),\ \mathbf{n} \in S)$.*

**Proof.** See Chapter 4. □

An important consequence of Theorem 1.2.5 is Kelly's lemma that will be the basis for the analysis in Chapter 4.

**Theorem 1.2.6 (Kelly's lemma)** *Let $\{N(t),\ t \in T\}$, $T = \mathbb{R}$, be a stationary Markov chain with transition rates $q(\mathbf{n},\mathbf{n}')$, $\mathbf{n},\mathbf{n}' \in S$. If we can find a collection of numbers $q^r(\mathbf{n},\mathbf{n}')$, $\mathbf{n},\mathbf{n}' \in S$, such that*

$$\sum_{\mathbf{n}'\neq\mathbf{n}} q(\mathbf{n},\mathbf{n}') = \sum_{\mathbf{n}'\neq\mathbf{n}} q^r(\mathbf{n},\mathbf{n}'), \quad \mathbf{n}\in S,$$

*and a distribution* $\pi = (\pi(\mathbf{n}),\ \mathbf{n}\in S)$ *such that*

$$\pi(\mathbf{n})q^r(\mathbf{n},\mathbf{n}') = \pi(\mathbf{n}')q(\mathbf{n}',\mathbf{n}), \quad \mathbf{n},\mathbf{n}'\in S,$$

*then* $q^r(\mathbf{n},\mathbf{n}')$*,* $\mathbf{n},\mathbf{n}'\in S$*, are the transition rates of the time-reversed Markov chain* $\{N(\tau-t),\ t\in T\}$ *and* $\pi = (\pi(\mathbf{n}),\ \mathbf{n}\in S)$*, is the equilibrium distribution of both Markov chains.*

**Proof.** See Chapter 4 □

# Chapter 2
# Reversibility, Poisson flows and feedforward networks

## 2.1 The birth-death process

A *birth-death process* is a Markov chain $\{N(t),\ t \in T\}$, $T = [0,\infty)$, or $T = \mathbb{R}$, at state space $S \subseteq \mathbb{N}_0$ with transition rates, for $\mathbf{n}, \mathbf{n}' \in S$,

$$q(\mathbf{n},\mathbf{n}') = \begin{cases} \lambda(\mathbf{n}) & \text{if } \mathbf{n}' = \mathbf{n}+1, \quad \text{(birth rate)}, \\ \mu(\mathbf{n})\mathbb{1}(\mathbf{n} > 0), & \text{if } \mathbf{n}' = \mathbf{n}-1, \quad \text{(death rate)}, \\ -\lambda(\mathbf{n}) - \mu(\mathbf{n}), & \text{if } \mathbf{n}' = \mathbf{n},\ \mathbf{n} > 0, \\ -\lambda(\mathbf{n}), & \text{if } \mathbf{n}' = \mathbf{n},\ \mathbf{n} = 0, \end{cases}$$

for $\lambda : S \to [0,\infty)$, $\mu : S \to (0,\infty)$. We are interested in the distribution

The Kolmogorov forward equations (1.12) read

$$\frac{dP(\mathbf{n},t)}{dt} = P(\mathbf{n}-1,t)\lambda(\mathbf{n}-1) + P(\mathbf{n}+1,t)\mu(\mathbf{n}+1) - P(\mathbf{n},t)[\lambda(\mathbf{n}) + \mu(\mathbf{n})], \qquad \mathbf{n} > 0,$$

$$\frac{dP(\mathbf{n},t)}{dt} = P(\mathbf{n}+1,t)\mu(\mathbf{n}+1) - P(\mathbf{n},t)\lambda(\mathbf{n}), \quad \mathbf{n} = 0.$$

Except for a few special cases, an (elegant) solution $P(\mathbf{n},t)$, $\mathbf{n} \in S$, is not available, see [**?**] for solutions for some special cases. The global balance equations (1.16) read

$$0 = \pi(\mathbf{n}-1)\lambda(\mathbf{n}-1) + \pi(\mathbf{n}+1)\mu(\mathbf{n}+1) - \pi(\mathbf{n})[\lambda(\mathbf{n}) + \mu(\mathbf{n})], \quad \mathbf{n} > 0,$$

$$0 = \pi(1)\mu(1) - \pi(0)\lambda(0).$$

Starting with the balance equation for $\mathbf{n} = 0$ we readily obtain that $\pi$ satisfies the *detailed balance equations*

$$\pi(\mathbf{n})\lambda(\mathbf{n}) = \pi(\mathbf{n}+1)\mu(\mathbf{n}+1), \quad \mathbf{n} \in S,$$

that may be iteratively solved to obtain the following result.

**Theorem 2.1.1** *Let $\{N(t)\}$ be a birth-death process at state space $S = \mathbb{N}_0$, with birth rates $\lambda(\mathbf{n})$ and death rates $\mu(\mathbf{n})$. If*

$$\pi(0)^{-1} := \left[\sum_{\mathbf{n}=0}^{\infty} \prod_{\mathbf{r}=0}^{\mathbf{n}-1} \frac{\lambda(\mathbf{r})}{\mu(\mathbf{r}+1)}\right] < \infty, \tag{2.1}$$

*then the equilibrium distribution is*

$$\pi(\mathbf{n}) = \pi(0) \prod_{\mathbf{r}=0}^{\mathbf{n}-1} \frac{q(\mathbf{r},\mathbf{r}+1)}{q(\mathbf{r}+1,\mathbf{r})} = \pi(0) \prod_{\mathbf{r}=0}^{\mathbf{n}-1} \frac{\lambda(\mathbf{r})}{\mu(\mathbf{r}+1)}, \quad \mathbf{n} \in S. \tag{2.2}$$

Note that we may always find an invariant measure $m$ that satisfies $m(\mathbf{n})\lambda(\mathbf{n}) = m(\mathbf{n}+1)\mu(\mathbf{n}+1)$, $\mathbf{n} \in S$, and that the normalisation condition (2.1) guarantees that the invariant measure $m$ is summable to obtain the equilibrium distribution (2.2).

**Example 2.1.2 (The $M|M|1$ queue)** Let customers arrive to a queue according to a Poisson process (the arrival process) with rate $\lambda$. Suppose there is a single server serving the customers in order of arrival and that customers' service times have a negative-exponential distribution with mean $\mu^{-1}$ and are independent of each other and of the arrival process. This queue is referred to as the *single server queue* or $M|M|1$ queue. The Markov chain $\{N(t),\ t \in T\}$, $T = [0,\infty)$, that records the number of customers in the queue is a birth-death process at state space $S = \mathbb{N}_0$ with birth and death rates

$$q(\mathbf{n},\mathbf{n}') = \begin{cases} \lambda(\mathbf{n}) = \lambda, & \text{if } \mathbf{n}' = \mathbf{n}+1, \quad \text{(birth rate)}, \\ \mu(\mathbf{n}) = \mu\mathbb{1}(\mathbf{n} > 0), & \text{if } \mathbf{n}' = \mathbf{n}-1, \quad \text{(death rate)} \end{cases}$$

and equilibrium distribution

$$\pi(\mathbf{n}) = (1-\rho)\rho^{\mathbf{n}}, \quad \mathbf{n} \in S,$$

provided that the queue is *stable*:

$$\rho := \frac{\lambda}{\mu} < 1.$$

□

**Example 2.1.3 (The $M|M|1|c$ queue)** Reconsider the $M|M|1$ queue, but now with finite waiting room that may contain at most $c-1$ customers. Let the system start in state 0 at time 0. Customers arriving to the queue containing $c$ customers (1 in service and $c-1$ waiting) are discarded. The Markov chain $\{N(t),\ t \in T\}$, $T = [0,\infty)$, that records the number of customers in the queue is a birth-death process at state space $S = \{0,1,2,\dots,c\}$ with birth and death rates

$$q(\mathbf{n},\mathbf{n}') = \begin{cases} \lambda(\mathbf{n}) = \lambda\mathbb{1}(\mathbf{n} < c), & \text{if } \mathbf{n}' = \mathbf{n}+1, \quad \text{(birth rate)}, \\ \mu(\mathbf{n}) = \mu\mathbb{1}(\mathbf{n} > 0), & \text{if } \mathbf{n}' = \mathbf{n}-1, \quad \text{(death rate)}. \end{cases}$$

For the $M|M|1|c$ queue starting in state 0 the process remains in the set $S$: the birth rate in state $c$ equals 0, so that the set of detailed balance equations is truncated at state $c$:

$$\pi(\mathbf{n})\lambda(\mathbf{n}) = \pi(\mathbf{n}+1)\mu(\mathbf{n}+1), \quad \mathbf{n} = 0,\dots,c-1.$$

The equilibrium distribution is that of the $M|M|1$ queue truncated to $S$:

$$\pi(\mathbf{n}) = \pi(0)\rho^{\mathbf{n}}, \quad \mathbf{n} \in \{0,1,\dots,c\},$$

with

$$\pi(0) = \left[\sum_{\mathbf{n}=0}^{c} \rho^{\mathbf{n}}\right]^{-1} = \frac{1-\rho}{1-\rho^{c+1}}.$$

□

**Example 2.1.4 (The $M|M|s$ queue)** Let customers arrive to a queue according to a Poisson process with rate $\lambda$. Suppose there are $s$, $s \geq 1$, servers serving the customers in parallel (each server serves one customer) in order of arrival and that customers' service times have a negative-exponential distribution with mean $\mu^{-1}$ and are independent of each other and of the arrival process. This queue is referred to as the *multi server queue* or $M|M|s$ queue. The Markov chain $\{N(t),\ t \in T\}$, $T = [0,\infty)$, that records the number of customers in the queue is a birth-death process at state space $S = \mathbb{N}_0$ with birth and death rates

$$q(\mathbf{n},\mathbf{n}') = \begin{cases} \lambda(\mathbf{n}) = \lambda, & \text{if } \mathbf{n}' = \mathbf{n}+1, \quad \text{(birth rate)}, \\ \mu(\mathbf{n}) = \mu\min(\mathbf{n},s), & \text{if } \mathbf{n}' = \mathbf{n}-1, \quad \text{(death rate)} \end{cases}$$

and equilibrium distribution

$$\pi(\mathbf{n}) = \begin{cases} \pi(0)\dfrac{\rho^{\mathbf{n}}}{\mathbf{n}!}, & \text{if } 0 \leq \mathbf{n} < s, \\[2ex] \pi(0)\dfrac{\rho^{\mathbf{n}}}{s^{\mathbf{n}-s}s!}, & \text{if } \mathbf{n} \geq s, \end{cases}$$

with normalising constant

$$\pi(0)^{-1} = \sum_{n=0}^{s-1} \frac{\rho^n}{n!} + \frac{\rho^s}{(s-\rho)(s-1)!},$$

provided that the queue is stable:

$$\rho := \frac{\lambda}{\mu} < s.$$

□

## 2.2 Detailed balance

Several properties of birth-death processes carry over to Markov chains that satisfy detailed balance.

**Definition 2.2.1 (Detailed balance)** *A Markov chain $\{N(t)\}$ at state space $S$ with transition rates $q(\mathbf{n},\mathbf{n}')$, $\mathbf{n},\mathbf{n}' \in S$, satisfies detailed balance if a distribution $\pi = (\pi(\mathbf{n}),\ \mathbf{n} \in S)$ exists that satisfies for all $\mathbf{n},\mathbf{n}' \in S$ the detailed balance equations:*

$$\pi(\mathbf{n})q(\mathbf{n},\mathbf{n}') - \pi(\mathbf{n}')q(\mathbf{n}',\mathbf{n}) = 0. \tag{2.3}$$

Summing (2.3) over all $\mathbf{n}' \in S$ gives the following result.

**Theorem 2.2.2** *If a distribution $\pi$ satisfies the detailed balance equations then $\pi$ is the equilibrium distribution.*

The detailed balance equations state that the probability flow between each pair of states is balanced.

The equilibrium distribution of the birth-death process may be iteratively obtained and is characterised via the ratio of the product of the transition rates (birth rates) on a path from state 0 to state $\mathbf{n}$ and the transition rates (death rates) on the reversed path from state $\mathbf{n}$ to state 0, see (2.2). This result may be generalized to Markov chains satisfying detailed balance. Kolmogorov's criteria provide this characterization and give a useful insight into the nature of detailed balance.

**Lemma 2.2.3 (Kolmogorov's criterion)** *A Markov chain $\{N(t)\}$ satisfies detailed balance if and only if its transition rates satisfy for all $r \in \mathbb{N}$ and any finite sequence of states $\mathbf{n}_1,\mathbf{n}_2,\dots,\mathbf{n}_r \in S$, $\mathbf{n}_r = \mathbf{n}_1$,*

$$\prod_{i=1}^{r-1} q(\mathbf{n}_i,\mathbf{n}_{i+1}) = \prod_{i=1}^{r-1} q(\mathbf{n}_{r-i+1},\mathbf{n}_{r-i}). \tag{2.4}$$

**Proof.** If $\{N(t)\}$ satisfies detailed balance, then for $i = 1,\dots,r$

$$\pi(\mathbf{n}_i)q(\mathbf{n}_i,\mathbf{n}_{i+1}) = \pi(\mathbf{n}_{i+1})q(\mathbf{n}_{i+1},\mathbf{n}_i).$$

Multiplying these equations for the finite sequence of states $\mathbf{n}_1,\mathbf{n}_2,\dots,\mathbf{n}_r \in S$, $\mathbf{n}_r = \mathbf{n}_1$, yields (2.4).

Conversely, suppose the transitions rates satisfy (2.4). Let $\mathbf{n}_0 \in S$ be an arbitrary state. Since $\{N(t)\}$ is irreducible, for all $\mathbf{n} \in S$ there exists a sequence of states $\mathbf{n} = \mathbf{n}_{r+1},\mathbf{n}_r,\dots,\mathbf{n}_1,\mathbf{n}_0$ such that $\prod_{i=0}^{r} q(\mathbf{n}_{r-i+1},\mathbf{n}_{r-i}) > 0$. Let

$$\pi(\mathbf{n}) = G\prod_{i=0}^{r} q(\mathbf{n}_i,\mathbf{n}_{i+1}) \left[\prod_{i=0}^{r} q(\mathbf{n}_{r-i+1},\mathbf{n}_{r-i})\right]^{-1}. \tag{2.5}$$

Observe that $\pi(\mathbf{n})$ does not depend on the sequence of states: if $\mathbf{n} = \mathbf{n}_{r+1} = \mathbf{n}'_{r+1}$, $\mathbf{n}'_r,\dots,\mathbf{n}'_1,\mathbf{n}_0 = \mathbf{n}'_0$ is another sequence of states such that $\prod_{i=0}^{r} q(\mathbf{n}'_{r-i+1},\mathbf{n}'_{r-i}) > 0$

then (2.4) implies that

$$\prod_{i=0}^{r} q(\mathbf{n}'_i, \mathbf{n}'_{i+1}) \left[ \prod_{i=0}^{r} q(\mathbf{n}'_{r-i+1}, \mathbf{n}'_{r-i}) \right]^{-1} = \prod_{i=0}^{r} q(\mathbf{n}_i, \mathbf{n}_{i+1}) \left[ \prod_{i=0}^{r} q(\mathbf{n}_{r-i+1}, \mathbf{n}_{r-i}) \right]^{-1}.$$

Furthermore, irreducibility and (2.4) imply that $\pi(\mathbf{n}) > 0$ for all $\mathbf{n} \in S$. It remains to show that $\pi$ satisfies detailed balance. To this end, consider $\mathbf{n}' \in S$. If $q(\mathbf{n},\mathbf{n}') = q(\mathbf{n}',\mathbf{n}) = 0$ then detailed balance is trivially satisfied. If $q(\mathbf{n}',\mathbf{n}) > 0$, then we may extend the sequence of states $\mathbf{n} = \mathbf{n}_{r+1}, \mathbf{n}_r, \dots, \mathbf{n}_1, \mathbf{n}_0$ to $\mathbf{n}', \mathbf{n} = \mathbf{n}_{r+1}, \mathbf{n}_r, \dots, \mathbf{n}_1, \mathbf{n}_0$ and (2.4) implies that we may define

$$\pi(\mathbf{n}') = G \left( \prod_{i=0}^{r} q(\mathbf{n}_i, \mathbf{n}_{i+1}) \right) q(\mathbf{n}, \mathbf{n}') \left[ \left( \prod_{i=0}^{r} q(\mathbf{n}_{r-i+1}, \mathbf{n}_{r-i}) \right) q(\mathbf{n}', \mathbf{n}) \right]^{-1}.$$

Hence, $\pi$ satisfies detailed balance $\pi(\mathbf{n})q(\mathbf{n},\mathbf{n}') = \pi(\mathbf{n}')q(\mathbf{n}',\mathbf{n})$ and therefore global balance. As $\{N(t)\}$ is assumed to be ergodic, it must be that $\pi$ is summable, so that $G$ may be chosen such that $\pi$ is a distribution and therefore $\{N(t)\}$ satisfies detailed balance. □

Equation (2.4) reflects that any finite path in the state space which returns to its initial point $\mathbf{n}_1$ has the same probability whether this path is traced in one direction or the other. This implies that a reversible Markov chain shows no net circulation in the state space. In practice, relations (2.4) usually have to be established for a small number of simple paths only, and (2.4) for general paths then follows by decomposition of these paths into simple paths.

The following result is a direct consequence of Kolmogorov's criterion (2.4), in particular of the definition of $\pi$ in (2.5). It provides a construction method for the equilibrium distribution by analogy with (2.2).

**Lemma 2.2.4 (Kolmogorov's criterion)** *For a Markov chain that satisfies detailed balance the equilibrium distribution $\pi$ is given by*

$$\pi(\mathbf{n}) = \pi(\mathbf{n}') \frac{q(\mathbf{n}_1, \mathbf{n}_2) q(\mathbf{n}_2, \mathbf{n}_3)}{q(\mathbf{n}_2, \mathbf{n}_1) q(\mathbf{n}_3, \mathbf{n}_2)} \cdots \frac{q(\mathbf{n}_{r-1}, \mathbf{n}_r)}{q(\mathbf{n}_r, \mathbf{n}_{r-1})}, \tag{2.6}$$

*for arbitrary $\mathbf{n}' \in S$ for all $r \in \mathbb{N}$ and any path $\mathbf{n}_1, \mathbf{n}_2, \dots, \mathbf{n}_r \in S$ such that $\mathbf{n}_1 = \mathbf{n}'$, $\mathbf{n}_r = \mathbf{n}$ for which the denominator is positive.*

The truncation property illustrated for the $M|M|1|s$ queue in Example 2.1.3 carries over to Markov chains satisfying detailed balance. The proof follows by insertion of the proposed distribution into the detailed balance equations (2.3).

**Theorem 2.2.5 (Truncation)** *Consider a Markov chain $\{N(t)\}$ at state space $S$ with transition rates $q(\mathbf{n},\mathbf{n}')$, $\mathbf{n},\mathbf{n}' \in S$, that satisfies detailed balance and has equilibrium distribution $\pi$. Let $V \subset S$.*

*Let $r > 0$. If the transition rates are altered by changing $q(\mathbf{n},\mathbf{n}')$ to $rq(\mathbf{n},\mathbf{n}')$ for $\mathbf{n} \in V$, $\mathbf{n}' \in S \setminus V$, then the resulting Markov chain $\{N_r(t)\}$ satisfies detailed balance* *and has equilibrium distribution*

$$\pi_r(\mathbf{n}) = \begin{cases} G\pi(\mathbf{n}), & \mathbf{n} \in V, \\ Gr\pi(\mathbf{n}), & \mathbf{n} \in S \setminus V, \end{cases}$$

*where G is the normalizing constant.*

*If $r = 0$ then the Markov chain is* truncated *to V and the resulting Markov chain sastisfies detailed balance with equilibrium distribution*

$$\pi_0(\mathbf{n}) = \pi(\mathbf{n}) \left[ \sum_{\mathbf{n} \in V} \pi(\mathbf{n}) \right]^{-1}, \quad \mathbf{n} \in V.$$

**Example 2.2.6 (Network of parallel $M|M|1$ queues and truncation)** Consider a network consisting of two $M|M|1$ queues in parallel. Queue $j$ has arrival rate $\lambda_j$ and service rate $\mu_j$, $j = 1,2$. The Markov chains $\{N_j(t)\}$ recording the number of customers in queue $j$, $j = 1,2$, are assumed independent. The Markov chain $\{N(t) = (N_1(t), N_2(t))\}$ at state space $S = \mathbb{N}_0^2$, where state $\mathbf{n} = (n_1, n_2)$ and $n_j$ records the number of customers in queue $j$, $j = 1,2$, has transition rates, for $\mathbf{n}, \mathbf{n}' \in S$, $\mathbf{n}' \neq \mathbf{n}$,

$$q(\mathbf{n}, \mathbf{n}') = \begin{cases} \lambda_j, & \text{if } \mathbf{n}' = \mathbf{n} + \mathbf{e}_j, \quad j = 1,2, \\ \mu_j, & \text{if } \mathbf{n}' = \mathbf{n} - \mathbf{e}_j, \quad j = 1,2, \end{cases} \tag{2.7}$$

where $\mathbf{e}_j$ is the $j$-th unit vector with 1 in position $j$ and 0 elsewhere. The random variables $N_j := N_j(\infty)$ recording the equilibrium number of customers in queue $j$ are independent random variables, so that the equilibrium distribution of $\{N(t)\}$ is the product of the marginal equilibrium distributions of the number of customers $\pi_j(n_j)$ in queue $j$:

$$\pi(\mathbf{n}) = \prod_{j=1}^{2} \pi_j(n_j), \quad \mathbf{n} \in S,$$

with

$$\pi_j(n_j) = (1 - \rho_j)\rho_j^{n_j}, \quad n_j \in \mathbb{N}_0,$$

provided that

$$\rho_j := \frac{\lambda_j}{\mu_j} < 1, \quad j = 1,2.$$

Now consider the network of two $M|M|1$ queues in parallel with common capacity restriction $n_1 + n_2 \leq c$. Customers arriving to the network with $c$ customers present are discarded. The Markov chain $\{N(t) = (N_1(t), N_2(t))\}$ has state space $S_c = \{(n_1, n_2) : n_j \geq 0,\ j = 1,2,\ n_1 + n_2 \leq c\}$ and transition rates (2.7) truncated to $S_c$. Invoking Theorem 2.2.5. the equilibrium distribution is

$$\pi(\mathbf{n}) = G_c \prod_{j=1}^{2} \rho_j^{n_j}, \quad \mathbf{n} \in S_c = \{(n_1, n_2) : n_j \geq 0,\ j = 1,2,\ n_1 + n_2 \leq c\},$$

with normalising constant

$$G_c = \left[ \sum_{n_1=0}^{c} \sum_{n_2=0}^{c-n_1} \prod_{i=1}^{2} \rho_i^{n_i} \right]^{-1}.$$

□

## 2.3 Erlang loss networks

**Example 2.3.1** (**$M|M|\infty$ queue**) Let customers arrive to a queue according to a Poisson process with rate $\lambda$. Suppose there is an ample supply of servers serving the customers so that each customer receives its own server. Let customers' service times have a negative-exponential distribution with mean $\mu^{-1}$, independent of each other and of the arrival process. This queue is referred to as the *infinite server queue* or $M|M|\infty$ queue. The Markov chain $\{N(t),\ t \in T\}$, $T = [0, \infty)$, that records the number of customers in the queue is a birth-death process at state space $S = \mathbb{N}_0$ with birth and death rates

$$q(\mathbf{n}, \mathbf{n}') = \begin{cases} \lambda, & \text{if } \mathbf{n}' = \mathbf{n} + 1, \quad \text{(birth rate)}, \\ \mathbf{n}\mu, & \text{if } \mathbf{n}' = \mathbf{n} - 1, \quad \text{(death rate)} \end{cases}$$

and equilibrium distribution the Poisson distribution

$$\pi(\mathbf{n}) = \mathrm{e}^{-\rho} \frac{\rho^{\mathbf{n}}}{\mathbf{n}!}, \quad \mathbf{n} \in S,$$

where

$$\rho := \frac{\lambda}{\mu}.$$

□

**Example 2.3.2** (**$M|M|s|s$ queue**) Now assume the number of servers is finite, say $s$, and let customers that find all servers occupied be rejected and discarded. Let the system start in state 0 at time 0. The Markov chain $\{N(t),\ t \in T\}$, $T = [0, \infty)$, that records the number of customers in the queue is a birth-death process at state space $S = \{0, 1, 2, \ldots, s\}$ with birth and death rates

$$q(\mathbf{n}, \mathbf{n}') = \begin{cases} \lambda \mathbb{1}(\mathbf{n} < s), & \text{if } \mathbf{n}' = \mathbf{n} + 1, \quad \text{(birth rate)}, \\ \mathbf{n}\mu, & \text{if } \mathbf{n}' = \mathbf{n} - 1, \quad \text{(death rate)}. \end{cases}$$

This queue is referred to as the $M|M|s|s$ queue or *Erlang loss queue* named after the founding father of queueing theory: A.K. Erlang. For the $M|M|s|s$ queue starting in state 0 the process remains in the set $S$. According to the Truncation Theorem 2.2.5 the equilibrium distribution is that of the $M|M|\infty$ queue truncated to $S$:

$$\pi(\mathbf{n}) = \pi(0)\frac{\rho^{\mathbf{n}}}{\mathbf{n}!}, \quad \mathbf{n} \in \{0,1,\dots,s\},$$

with

$$\pi(0)^{-1} = \sum_{\mathbf{n}=0}^{s} \frac{\rho^{\mathbf{n}}}{\mathbf{n}!}.$$

The normalising constant can be recursively evaluated as follows.

$$\begin{aligned} &T(s) = 1,\ S(s) = 1, \\ &\begin{cases} T(k) := \frac{k+1}{\rho} T(k+1), \\ S(k) := S(k+1) + T(k), \end{cases} \quad k = s-1,\dots,0. \end{aligned}$$

Then $\pi(0) = T(0)/S(0)$. The recursion avoids evaluating the factorials. Note that the recursion also yields all equilibrium probabilities:

$$\pi(\mathbf{n}) = \frac{T(\mathbf{n})}{S(1)}, \quad \mathbf{n} = 0,\dots,s.$$

□

Now consider a multidimensional network of $J$ parallel $M|M|\infty$ queues as follows. Let customers arrive to a queue $j$ according to a Poisson process with rate $\lambda_j$, $j = 1,\dots,J$. Suppose there is an ample supply of servers at each queue. Let customers' service times at queue $j$ have a negative-exponential distribution with mean $\mu_j^{-1}$, independent of each other and of the arrival processes. The Markov chain $\{N(t),\ t \in T\}$, $T = [0,\infty)$, that records the number of customers in the queues has state space $S = \mathbb{N}_0^J$ and states $\mathbf{n} = (n_1,\dots,n_J)$, with $n_j$ recording the number of customers in the queue $j$, $j = 1,\dots,J$, and transition rates

$$q(\mathbf{n},\mathbf{n}') = \begin{cases} \lambda_j, & \text{if } \mathbf{n}' = \mathbf{n} + \mathbf{e}_j, \\ n_j\mu_j, & \text{if } \mathbf{n}' = \mathbf{n} - \mathbf{e}_j. \end{cases}$$

The random variables recording the number of customers in different queues are clearly independent so that the equilibrium distribution is

$$\pi(\mathbf{n}) = \prod_{j=1}^{J} \mathrm{e}^{-\rho_j} \frac{\rho_j^{n_j}}{n_j!}, \quad \mathbf{n} \in S = \mathbb{N}_0^J,$$

where $\rho_j := \lambda_j/\mu_j$.

The *Erlang loss network* is the truncation of the network of parallel infinite server queues to the polytope

$$S = \{\mathbf{n} \in \mathbb{N}_0^J : \mathbf{n} \geq 0,\ \mathbf{n}A \leq \mathbf{c}\},$$

where $\mathbf{c} = (c_1, \ldots, c_K) \in \mathbb{R}^K$ and $A$ is a $J \times K$ matrix. The Markov chain $\{N(t),\ t \in T\}$, $T = [0, \infty)$, that records the number of customers in the queues of the Erlang loss network has state space $S = \{\mathbf{n} \in \mathbb{N}_0^J : \mathbf{n} \geq 0,\ \mathbf{n}A \leq \mathbf{c}\}$ and states $\mathbf{n} = (n_1, \ldots, n_J)$, with $n_j$ recording the number of customers in the queue $j$, $j = 1, \ldots, J$, and transition rates

$$q(\mathbf{n}, \mathbf{n}') = \begin{cases} \lambda_j \mathbb{1}(A(\mathbf{n} + \mathbf{e}_j) < \mathbf{c}), & \text{if } \mathbf{n}' = \mathbf{n} + \mathbf{e}_j, \\ n_j \mu_j, & \text{if } \mathbf{n}' = \mathbf{n} - \mathbf{e}_j. \end{cases}$$

According to the Truncation Theorem 2.2.5 the equilibrium distribution is that of the network of parallel $M|M|\infty$ queues truncated to $S$ :

$$\pi(\mathbf{n}) = \pi(0) \prod_{j=1}^{J} \frac{\rho_j^{n_j}}{n_j!}, \quad \mathbf{n} \in S,$$

where

$$\pi(0) = \left[ \sum_{\mathbf{n} \in S} \prod_{j=1}^{J} \frac{\rho_j^{n_j}}{n_j!} \right]^{-1}.$$

Observe that

$$\pi(\mathbf{n}) = \prod_{j=1}^{J} \frac{\rho_j^{n_j}}{n_j!} \left[ \sum_{\mathbf{n} \in S} \prod_{j=1}^{J} \frac{\rho_j^{n_j}}{n_j!} \right]^{-1} = \prod_{j=1}^{J} \frac{\rho_j^{n_j}}{n_j!} \mathrm{e}^{-\rho_j} \left[ \sum_{\mathbf{n} \in S} \prod_{j=1}^{J} \frac{\rho_j^{n_j}}{n_j!} \mathrm{e}^{-\rho_j} \right]^{-1}, \quad \mathbf{n} \in S,$$

is a multidimensional Poisson distribution truncated to $S$. This allows for an efficient Monte-Carlo summation method to evaluate the normalising constant $\pi(0)$, see Chapter **??** for details.

## 2.4 Reversibility

The algebraic detailed balance property is related to the probabilistic reversibility property.

**Definition 2.4.1 (Reversibility)** *A stochastic process $\{N(t),\ t \in \mathbb{R}\}$ is reversible if $(N(t_1), N(t_2), \ldots, N(t_k))$ has the same distribution as $(N(\tau - t_1), N(\tau - t_2), \ldots, N(\tau - t_k))$ for all $k \in \mathbb{N}$, $t_1, t_2, \ldots, t_k \in \mathbb{R}$, $\tau \in \mathbb{R}$.*

If a stochastic process is reversible and the direction of time is reversed, then the probabilistic behaviour of the process remains the same. We readily obtain the following result.

**Theorem 2.4.2** *If $\{N(t)\}$ is reversible then $\{N(t)\}$ is stationary.*

The algebraic detailed balance property is related to the probabilistic reversibility property.

**Theorem 2.4.3 (Reversibility and detailed balance)** *Let $\{N(t),\ t \in \mathbb{R}\}$ be a stationary Markov chain with transition rates $q(\mathbf{n},\mathbf{n}')$, $\mathbf{n},\mathbf{n}' \in S$. $\{N(t)\}$ is reversible if and only if there exists a distribution $\pi = (\pi(\mathbf{n}),\ \mathbf{n} \in S)$ that satisfies the detailed balance equations. When there exists such a distribution $\pi$, then $\pi$ is the equilibrium distribution of $\{N(t)\}$.*

**Proof.** If $\{N(t)\}$ is reversible, then for all $t,h \in \mathbb{R}$, $h > 0$, $\mathbf{n},\mathbf{n}' \in S$:

$$\mathbb{P}(N(t+h) = \mathbf{n}',\ N(t) = \mathbf{n}) = \mathbb{P}(N(t) = \mathbf{n}',\ N(t+h) = \mathbf{n}).$$

$\{N(t),\ t \in \mathbb{R}\}$ is a stationary Markov chain. Let $\pi(\mathbf{n}) = \mathbb{P}(N(t) = \mathbf{n})$, $t \in \mathbb{R}$. Then

$$\frac{\mathbb{P}(N(t+h) = \mathbf{n}'|N(t) = \mathbf{n})}{h}\pi(\mathbf{n}) = \frac{\mathbb{P}(N(t+h) = \mathbf{n}|N(t) = \mathbf{n}')}{h}\pi(\mathbf{n}').$$

Letting $h \to 0$ yields the detailed balance equations (2.3).

Now assume there exists a distribution $\pi = (\pi(\mathbf{n}),\ \mathbf{n} \in S)$ that satisfies the detailed balance equations. Recall that the Markov jump chain $\{N(t)\}$ remains in state $\mathbf{n}$ for a negative-exponentially distributed sojourn time with mean $q(\mathbf{n})^{-1}$, and has transition probabilities $p(\mathbf{n},\mathbf{n}') = q(\mathbf{n},\mathbf{n}')/q(\mathbf{n})$. Now consider $\{N(t)\}$ for $t \in [-H,H]$. Suppose in the interval $[-H,H]$ $\{N(t)\}$ moves along the sequence of states $\mathbf{n}_1,\dots,\mathbf{n}_k$ and has (remaining) sojourn time $h_i$ in each state $\mathbf{n}_i$, $i = 1,\dots,k-1$, and remains in state $\mathbf{n}_k$ for at least $h_k$ until time $H$. With probability $\pi(\mathbf{n}_1) = \mathbb{P}(N(-H) = \mathbf{n}_1)$ $\{N(t)\}$ starts in state $\mathbf{n}_1$ at time $-H$. The probability density with respect to $h_1,\dots,h_k$ for this sequence is

$$\pi(\mathbf{n}_1)q(\mathbf{n}_1)\mathrm{e}^{-q(\mathbf{n}_1)h_1}p(\mathbf{n}_1,\mathbf{n}_2)q(\mathbf{n}_2)\mathrm{e}^{-q(\mathbf{n}_2)h_2}\cdots q(\mathbf{n}_{k-1})\mathrm{e}^{-q(\mathbf{n}_{k-1})h_{k-1}}p(\mathbf{n}_{k-1},\mathbf{n}_k)\mathrm{e}^{-q(\mathbf{n}_k)h_k}, \quad (2.8)$$

where $\mathrm{e}^{-q(\mathbf{n}_k)h_k}$ is the probability that $\{N(t)\}$ resides in state $\mathbf{n}_k$ for at least $h_k$. Observe that $q(\mathbf{n}_i)p(\mathbf{n}_i,\mathbf{n}_{i+1}) = q(\mathbf{n}_i,\mathbf{n}_{i+1})$ and that Kolmogorov's criterion (Lemma 2.2.4) implies that

$$\pi(\mathbf{n}_1)q(\mathbf{n}_1,\mathbf{n}_2)q(\mathbf{n}_2,\mathbf{n}_3)\cdots q(\mathbf{n}_{k-1},\mathbf{n}_k) = \pi(\mathbf{n}_k)q(\mathbf{n}_k,\mathbf{n}_{k-1})\cdots q(\mathbf{n}_3,\mathbf{n}_2)q(\mathbf{n}_2,\mathbf{n}_1),$$

which implies that the probability density (2.8) equals the probability density for the reversed path that starts in $\mathbf{n}_k$ at time $H$ and moves along the sequence of states $\mathbf{n}_k,\dots,\mathbf{n}_1$ with (remaining) sojourn time $h_i$ in states $\mathbf{n}_i$, $i = k,\dots,2$, and remains in state $\mathbf{n}_1$ for at least $h_1$ until time $-H$. Thus, $(N(t_1),N(t_2),\dots,N(t_k))$ has the same distribution as $(N(-t_1),N(t_2),\dots,N(-t_k))$ that for all $\tau \in \mathbb{R}$ has the same distribution as $(N(\tau - t_1),N(\tau - t_2),\dots,N(\tau - t_k))$ since $\{N(t)\}$ is stationary. □

**Example 2.4.4 (Departure process from the $M|M|1$ queue)** The arrival process to the $M|M|1$ queue is a Poisson process with rate $\lambda$. If $\lambda < \mu$ the departure process from the $M|M|1$ queue has rate $\lambda$ as this rate is comprised of departure rate $\mu$ during the fraction of time the queue is busy, $1 - \pi(0) = \rho$, and departure rate 0 during the fraction of time the queue is idle, $\pi(0) = 1 - \rho$, so that the departure rate is

$\mu \cdot \rho + 0 \cdot (1-\rho) = \lambda$. Alternatively, the departure rate in an ergodic Markov chain for the $M|M|1$ queue must equal the arrival rate.[1]

Reversibility allows us to conclude that the departure *process* from the $M|M|1$ queue is a Poisson process. The Markov chain $\{N(t)\}$ recording the number of customer in the $M|M|1$ queue with arrival rate $\lambda$ and service rate $\mu$ satisfies detailed balance. Therefore, the Markov chain $\{N^r(t)\}$ in reversed time has Poisson arrivals at rate $\lambda$ and service rate $\mu$. The Markov chain $\{N^r(t)\}$ is characterised by its transition rates. Therefore $\{N^r(t)\}$ is the Markov chain of an $M|M|1$ queue with Poisson arrivals at rate $\lambda$ and negative-exponential service at rate $\mu$. As epochs of the arrival process for the reversed queue coincide with the epochs of the arrival process for the original queue, it must be that the departure process from the $M|M|1$ queue is a Poisson *process* with rate $\lambda$. □

## 2.5 Burke's theorem and feedforward networks of $M|M|1$ queues

Example 2.4.4 characterises the departure process from the $M|M|1$ queue. In a tandem network of two queues the departure process of the first $M|M|1$ queue will be the arrival process for the second queue in the tandem. This allows us to obtain the marginal distribution of the number of customers in each queue in this network of two queues: the second queue is an $M|M|1$ queue with Poisson arrival process. However, this is not sufficient to characterise the joint distribution of the number of customers in the two queues. This requires the following stronger result.

**Theorem 2.5.1 (Burke's theorem)** *Let $\{N(t)\}$ record the number of customers in the $M|M|1$ queue with arrival rate $\lambda$ and service rate $\mu$, $\lambda < \mu$. Let $\{D(t)\}$ record the customers' departure process from the queue. In equilibrium the departure process $\{D(t)\}$ is a Poisson process with rate $\lambda$, and $N(t)$ is independent of $\{D(s),\ s < t\}$.*

The independence property in Burke's theorem is surprising and counterintuitive, as departures occur at rate $\mu$ when the queue is busy and at rate 0 when the queue is idle, see Example 2.4.4. When the queue has been empty for a long time, then clearly the departure rate has been zero for a long time. However, Burke's theorem gives no information on the number of customers in the queue, and therefore also not of the busy and idle periods of the queue, only of the departure process.

Note that the arrival process $\{A(t)\}$ is a Poisson process and that $N(t)$ is independent of $\{A(s),\ s > t\}$ as Poisson arrivals in disjoint intervals are independent. Thus, Burke's theorem implies that $\{A(s),\ s > t\}$, $N(t)$, and $\{D(s),\ s < t\}$ are independent.

**Proof of Burke's theorem.** The proof uses a refinement of the arguments in Example 2.4.4. The epochs at which $\{N(t)\}$ jumps upwards (arrivals to the queue) form

[1] Note the for $\lambda > \mu$ the departure rate eventually will be $\mu$ as this $M|M|1$ queue eventually will never be idle.

a Poisson process with rate $\lambda$. The $M|M|1$ queue is reversible, so that the epochs at which $\{N(-t)\}$ jumps upwards must also form a Poisson process with rate $\lambda$. If $\{N(-t)\}$ jumps upwards at time $t^*$ then $\{N(t)\}$ must jump downwards at time $t^*$. Therefore, the departure process forms a Poisson process with rate $\lambda$. Moreover, since $\{N(t)\}$ is reversible the departure process up to time $t^*$ and $N(t^*)$ have the same distribution as the arrival process after $-t^*$ and $N(-t^*)$. As the arrival process is a Poisson process the arrival process after $-t^*$ is independent of $N(-t^*)$. Hence, the departure process up to time $t^*$ is independent of $N(t^*)$. □

**Remark 2.5.2 (Burke's theorem for reversible processes)** Note that the proof of Burke's theorem only uses that the arrival process is a Poisson process and that $\{N(t)\}$ is reversible. Therefore, the result of Burke's theorem remains valid for any birth-death process with constant birth rates $\lambda(\mathbf{n}) = \lambda$, $\mathbf{n} \in \mathbb{N}_0$. □

Consider a *tandem network* of two $M|M|1$ queues with Poisson arrival process with rate $\lambda$ to queue 1 and service rates $\mu_i$ at queue $i$, $i = 1,2$. Provided that $\rho_i = \lambda/\mu_i < 1$, $i = 1,2$, Example 2.4.4 shows that the Markov chain $\{N(t) = (N_1(t),N_2(t))\}$ at state space $S = \mathbb{N}_0^2$, where $\mathbf{n} = (n_1,n_2)$ and $n_i$ the number of customers in queue $i$, $i = 1,2$, has marginal equilibrium distributions $\pi_i(n_i) = (1-\rho_i)\rho_i^{n_i}$, $n_i \in \mathbb{N}_0$, for the number of customers at queue $i$, $i = 1,2$. Burke's theorem enables us to obtain the equilibrium distribution for the tandem of $M|M|1$ queues. To this end, let $t^*$ be fixed but arbitrary. Observe that the number of customers $N_2(t^*)$ of queue 2 at time $t^*$ is determined by the arrival process to queue 2 before $t^*$ and the service process at queue 2 before $t^*$. The arrival process to queue 2 is the departure process from queue 1. Burke's theorem states that the departure process from queue 1 before $t^*$ and $N_1(t^*)$, the number of customers at queue 1 at time $t^*$, are independent. Hence, in equilibrium, at time $t^*$ the random variables $N_1(t^*)$ and $N_2(t^*)$ are independent, so that, in equilibrium, at fixed but arbitrary time $t^*$,

$$\pi(\mathbf{n}) = \prod_{i=1}^{2} \pi_i(n_i), \quad \mathbf{n} \in S,$$

with

$$\pi_i(n_i) = (1-\rho_i)\rho_i^{n_i}, \quad n_i \in \mathbb{N}_0,$$

provided that

$$\rho_i := \frac{\lambda}{\mu_i} < 1, \quad i = 1,2.$$

This result readily extends to a tandem network of $J$ $M|M|1$ queues. Let $\{N(t) = (N_1(t),\ldots,N_J(t))\}$ at state space $S = \mathbb{N}_0^J$, where $\mathbf{n} = (n_1,\ldots,n_J)$ and $n_i$ the number of customers in queue $i$, $i = 1,\ldots,J$, record the number of customers in a tandem of $J$ $M|M|1$ queues with Poisson arrival process with rate $\lambda$ to queue 1 and service rates $\mu_i$ at queue $i$, with $\rho_i = \lambda/\mu_i < 1$, $i = 1,\ldots,J$. Consider fixed time $t^*$. Burke's theorem implies that $N_1(t^*)$ is independent of $(N_2(t^*),\ldots,N_J(t^*))$. Similarly, $N_j(t^*)$ is independent of $(N_{j+1}(t^*),\ldots,N_J(t^*))$, $j = 2,\ldots,J-1$. Thus at time $t^*$ the random variables $N_1(t^*),\ldots,N_J(t^*)$ are independent random variables, so that in equilibrium

$$\pi(\mathbf{n}) = \prod_{j=1}^{J}(1-\rho_j)\rho_j^{n_j}, \quad n_j \in \mathbb{N}_0,\ j=1,\ldots,J.$$

**Remark 2.5.3 (Independence of processes)** Clearly, the processes $\{N_j(t)\}$, $j = 1,\ldots,J$, are not independent. If queue $j$ has build up a very large queue at time $t^*$, then the arrival process to queue $j+1$ will temporarily (but for a long time) have negative-exponential interarrival times with rate $\mu_j > \lambda$ so that the queue length at queue $j+1$ subsequent to $t^*$ is likely to grow. □

**Remark 2.5.4 (Sojourn times)** Reversibility allows us to conclude an even stronger result. Let $W_j$ denote the sojourn time (including service time) of a customer at queue $j$ in a tandem of $M|M|1$ First In First Out queues (see Example 4.2.2). Then in equilibrium $W_j$, $j = 1,\ldots,J$, are independent random variables, see Chapter **??**. □

In a *feedforward network* of $J$ $M|M|1$ queues a customer leaving queue $j$ can be routed to any of the queues $j+1,\ldots,J$, or may leave the network. Let $p_{ij}$ denote the fraction of customers routing from queue $i$ to queue $j > i$, and $p_{i0}$ the fraction of customers leaving the network from queue $i$, $\sum_{j=i+1}^{J} p_{ij} + p_{i0} = 1$. Customers arrive to the network according to a Poisson process with rate $\mu_0$. A fraction $p_{0j}$ of these customers is routed to queue $j$, $j = 1,\ldots,J$, and $\sum_{j=1}^{J} p_{0j} = 1$. The service rate at queue $j$ is $\mu_j$, $j = 1,\ldots,J$. Burke's theorem implies that all flows of customers among the queues are Poisson flows. The arrival rate $\lambda_j$ of customers to queue $j$ is obtained from superposition and random splitting of Poisson processes[2]:

$$\lambda_j = \mu_0 p_{0j} + \sum_{i=1}^{j-1} \lambda_i p_{ij}, \quad j = 1,\ldots,J, \tag{2.9}$$

where $\mu_0 p_{0j}$ is the Poisson arrival rate of customers arriving to the network at queue $j$ obtained from random splitting of the Poisson arrival process with rate $\mu_0$, and $\lambda_i p_{ij}$ is the Poisson flow of customers from queue $i$ to queue $j$ obtained from random splitting of the Poisson departure process with rate $\lambda_i$ from queue $i$. The set of equations (2.9) is referred to as *traffic equations* as these equations determine the mean flow of customers in the network.

**Theorem 2.5.5 (Equilibrium distribution: feedforward network)** *Let $\{N(t) = (N_1(t),\ldots,N_J(t))\}$ at state space $S = \mathbb{N}_0^J$, where $\mathbf{n} = (n_1,\ldots,n_J)$ and $n_j$ the number of customers in queue $j$, $j = 1,\ldots,J$, record the number of customers in the feedforward network of $J$ $M|M|1$ queues described above. If $\rho_j = \lambda_j/\mu_j < 1$, with $\lambda_j$ the solution of the traffic equations (2.9), $j = 1,\ldots,J$, then the equilibrium distribution is the product of the marginal distributions of the queues:*

[2] Let $\{A_i(t)\}$ be Poisson processes with rates $\lambda_i$, $i = 1,2$. The superposition $\{A_1(t)+A_2(t)\}$ is a Poisson process with rate $\lambda_1+\lambda_2$. Random splitting with $p \in (0,1)$ of a Poisson process $\{A(t)\}$ with rate $\lambda$ yields two independent Poisson processes $\{A_1(t)\}$ and $\{A_2(t)\}$ with rates $\lambda_1 = p\lambda$, $\lambda_2 = (1-p)\lambda$.

$$\pi(\mathbf{n}) = \prod_{j=1}^{J}(1-\rho_j)\rho_j^{n_j}, \quad n_j \in \mathbb{N}_0,\ j = 1,\ldots,J. \tag{2.10}$$

The result may be further extended to feedforward networks that include queues that can be modelled as birth-death processes with constant birth rates. We will not pursue this approach since it breaks down when a customer may return to a queue it visited before: if a customer revisits a queue in the interval $(t, t+h)$ then the arrival process in the interval $(t, t+h)$ is no longer independent from the arrival process before time $t$ as the revisiting customer arrived before time $t$ in its previous visit. This is in contradiction with the independent increment property of the Poisson process, so that the arrival process cannot be a Poisson process. Chapter 3 considers networks with more general routing.

## 2.6 Literature

is due to R.R.P. Jackson [**?**] and

# Chapter 3
# Partial balance and networks with Markovian routing

## 3.1 Networks of $M|M|1$ queues

Consider a queueing network consisting of $J$ queues labelled $1,2\dots,J$. In this queueing network customers of a *a single type* route among the queues to receive a desired service. At queue $i$ a customer requires an amount of service that is negative-exponentially distributed with rate $\mu_i$, $i=1,\dots,J$, that is, if the required amount of service is worked off at rate 1 then *the service-time at queue i is negative-exponentially distributed with rate $\mu_i$, $i=1,\dots,J$.* Let $p_{ij}$ denote the fraction of customers that upon service completion route from queue $i$ to queue $j$, $j=1,\dots,J$, and $p_{i0}$ the fraction of customers leaving the network from queue $i$, $\sum_{j=0}^{J} p_{ij}=1$, $i=1,\dots,J$. Customers arrive to the network according to a Poisson process with rate $\mu_0$. A fraction $p_{0j}$ of these customers is routed to queue $j$, $j=1,\dots,J$, and $\sum_{j=1}^{J} p_{0j}=1$. Customers are served one-by-one, and arrive one-by-one so that only one customer can move between the queues of the queueing network at a time. At the queues *customer positions are not taken into account*. As a consequence, $n_j$, the number of customers at the queue $j$, $j=1,\dots,J$, give a full description of the state of the queueing network.

The evolution of the number of customers in the queues is recorded by the Markov chain $\{N(t)=(N_1(t),\dots,N_J(t)),\ t\in\mathbb{R}\}$ at state space $S\subseteq\mathbb{N}_0^J$ with states $\mathbf{n}=(n_1,\dots,n_J)$. Let $\mathbf{e}_j$ denote the $j$-th unit vector that has entry 1 in position $j$, 0 elsewhere, and $\mathbf{e}_0$ the zero-vector with all entries 0.

If $\{N(t)\}$ is in state $\mathbf{n}$ and a customer routes from queue $i$ to queue $j$ in the queueing network then the next state of $\{N(t)\}$ is $\mathbf{n}-e_i+e_j$, $i,j=0,\dots,J$. Here queue 0 is introduced to represent the outside. If a customer routes from queue $i$ to queue 0 then this customer leaves the queueing network and if a customer routes from queue 0 to queue $j$ then this customers enters the queueing network at queue $j$, $j=1,\dots,J$.

The queueing network introduced above is called *open* as arrivals to the queueing network and departures from the queueing network are possible. In this case $S=\mathbb{N}_0^J$. The number of customers in the queueing network is not constant. The transition rates of $\{N(t)\}$ for an open network are, for $\mathbf{n} \neq \mathbf{n}'$, $\mathbf{n}, \mathbf{n}' \in S$,

$$q(\mathbf{n},\mathbf{n}') = \begin{cases} \mu_i p_{ij}, & \text{if } \mathbf{n}' = \mathbf{n} - e_i + e_j,\ i,j = 0,\ldots,J, \\ 0, & \text{otherwise.} \end{cases} \tag{3.1}$$

A queueing network is called *closed* if arrivals to the queueing network and departures from the queueing network are not possible. We may obtain a closed network from the description above by setting $\mu_0 = 0$ and $p_{j0} = 0$, $j = 1,\ldots,J$. The number of customers in a closed network is constant: $S = S_M = \{\mathbf{n} : \sum_{j=1}^{J} n_j = M\}$ for some $M$, the number of customers in the network. The transition rates of $\{N(t)\}$ for a closed network are, for $\mathbf{n} \neq \mathbf{n}'$, $\mathbf{n}, \mathbf{n}' \in S$,

$$q(\mathbf{n},\mathbf{n}') = \begin{cases} \mu_i p_{ij}, & \text{if } \mathbf{n}' = \mathbf{n} - e_i + e_j,\ i,j = 1,\ldots,J, \\ 0, & \text{otherwise.} \end{cases} \tag{3.2}$$

Below, we will show that the so-called *product-form equilibrium distribution* (2.10) carries over to the Markov chain $\{N(t)\}$ with general routing probabilities $p_{ij}$, $i,j = 0,\ldots,J$. We first consider unicity of the solution of the traffic equations.

**Lemma 3.1.1 (Traffic equations: open network)** *Consider an open network. Assume that the routing matrix $P = (p_{ij},\ i,j = 0,\ldots,J)$ is irreducible.*[1] *Then the traffic equations*

$$\lambda_j = \mu_0 p_{0j} + \sum_{i=1}^{J} \lambda_i p_{ij}, \quad j = 1,\ldots,J, \tag{3.3}$$

*have a unique non-negative solution* $\{\lambda_j,\ j = 1,\ldots,J\}$.

Observe that the traffic equations (3.3) also imply a traffic equation for queue 0: summing (3.3) over $j = 1,\ldots,J$ yields

$$\sum_{j=1}^{J} \lambda_j = \sum_{j=1}^{J}\sum_{i=0}^{J} \lambda_j p_{ji} = \sum_{j=1}^{J} \lambda_j p_{j0} + \sum_{j=1}^{J}\sum_{i=1}^{J} \lambda_j p_{ji},$$
$$\sum_{j=1}^{J} \left\{ \mu_0 p_{0j} + \sum_{i=1}^{J} \lambda_i p_{ij} \right\} = \mu_0 + \sum_{j=1}^{J}\sum_{i=1}^{J} \lambda_i p_{ij},$$

and therefore

$$\mu_0 = \sum_{j=1}^{J} \lambda_j p_{j0}, \tag{3.4}$$

reflecting that the arrival rate to the network equals the departure rate from the network.

[1] A matrix $P = (p_{ij},\ i,j = 0,\ldots,J)$ with non-negative entries $p_{ij}$, $i,j = 0,\ldots,J$, is *irreducible* if for every pair $i,j$ there exists an $n \in \mathbb{N}$ such that $P^n_{ij} > 0$. Observe that for a probability matrix this is equivalent to the statement that the discrete-time Markov chain at state space $\{0,\ldots,J\}$ with transition matrix $P$ is irreducible.

**Lemma 3.1.2 (Traffic equations: closed network)** *Consider a closed network. Assume that the routing matrix $P = (p_{ij},\ i,j = 1,\dots,J)$ is irreducible. Then the traffic equations*

$$\lambda_j = \sum_{i=1}^{J} \lambda_i p_{ij}, \quad j = 1,\dots,J, \tag{3.5}$$

*have a unique non-negative solution $\{\lambda_j,\ j = 1,\dots,J\}$ such that $\sum_{j=1}^{J} \lambda_j = 1$.*

**Proof of Lemmas 3.1.1, 3.1.2.** First consider the open network. Let $\lambda_0 = \mu_0$, recall (3.4), and observe that $\sum_{i=0}^{J} p_{ji} = 1,\ j = 0,\dots,J$, then the traffic equations for the open network (3.3) read,

$$\sum_{i=0}^{J} \lambda_j p_{ji} = \sum_{i=0}^{J} \lambda_i p_{ij}, \quad j = 0,\dots,J. \tag{3.6}$$

These equations are the global balance equations for the Markov chain at state space $S_{to} = \{0,\dots,J\}$ with transition rates $q(i,j) = p_{ij},\ i,j \in S$.[2] As the routing matrix is irreducible, this Markov chain has a unique equilibrium distribution $\pi(i),\ i \in S_{to}$. The solution $\{\lambda_j,\ j = 0,\dots,J\}$ of the traffic equations must be proportional to $\pi$, in particular $\lambda_j = \mu_0 \pi(j)/\pi(0),\ j = 0,\dots,J$, so that the solution of the traffic equations is unique and non-negative.

For the closed network, the traffic equations (3.5) are the global balance equations for the Markov chain with state space $S_{tc} = \{1,\dots,J\}$. The normalising condition $\sum_{j=1}^{J} \lambda_j = 1$ guarantees unicity of the solution. □

**Remark 3.1.3 (Markovian routing)** Customers route among the queues according to the transition probabilities $p_{ij},\ i,j = 0,\dots,J$, of a discrete-time Markov chain. This is referred to as Markovian routing. □

**Theorem 3.1.4 (Equilibrium distribution: open network of $M|M|1$ queues)** *Consider the Markov chain $\{N(t)\}$ at state space $S = \mathbb{N}_0^J$ with transition rates (3.1) for the open network of $M|M|1$ queues. Assume that the routing matrix $P = (p_{ij},\ i,j = 0,\dots,J)$ is irreducible and let $\{\lambda_j,\ j = 1,\dots,J\}$ be the unique solution of the traffic equations (3.3). If $\rho_j := \lambda_j/\mu_j < 1,\ j = 1,\dots,J$, then $\{N(t)\}$ has unique equilibrium distribution*

$$\pi(\mathbf{n}) = G_o \prod_{j=1}^{J} \rho_j^{n_j}, \quad \mathbf{n} \in S, \tag{3.7}$$

*where*

$$G_o = \prod_{j=1}^{J} (1 - \rho_j).$$

*Moreover, the equilibrium distribution (3.8) satisfies* partial balance*, for all $\mathbf{n} \in S$,*

[2] This is the Markov chain for a single customer that routes among the queues $0,\dots,J$.

$$\sum_{j=0}^{J} \left\{ \pi(\mathbf{n}) q(\mathbf{n}, \mathbf{n} - \mathbf{e}_i + \mathbf{e}_j) - \pi(\mathbf{n} - \mathbf{e}_i + \mathbf{e}_j) q(\mathbf{n} - \mathbf{e}_i + \mathbf{e}_j, \mathbf{n}) \right\}, \quad i = 0, \ldots, J.$$

**Theorem 3.1.5 (Equilibrium distribution: closed network of $M|M|1$ queues)** *Consider the Markov chain $\{N(t)\}$ at state space $S = S_M = \{\mathbf{n} : \sum_{j=1}^{J} n_j = M\}$ with transition rates (3.2) for the closed network of $M|M|1$ queues containing $M$ customers. Assume that the routing matrix $P = (p_{ij},\ i, j = 1, \ldots, J)$ is irreducible and let $\{\lambda_j,\ j = 1, \ldots, J\}$ be the unique solution of the traffic equations (3.5) such that $\sum_{j=1}^{J} \lambda_j = 1$. Let $\rho_j := \lambda_j / \mu_j$, $j = 1, \ldots, J$. Then $\{N(t)\}$ has unique equilibrium distribution*

$$\pi(\mathbf{n}) = G_M \prod_{j=1}^{J} \rho_j^{n_j}, \quad \mathbf{n} \in S, \tag{3.8}$$

*where*

$$G_M = \left[ \sum_{\mathbf{n} \in S} \prod_{j=1}^{J} \rho_i^{n_j} \right]^{-1}.$$

*Moreover, the equilibrium distribution (3.8) satisfies* partial balance*, for all $\mathbf{n} \in S$,*

$$\sum_{j=1}^{J} \left\{ \pi(\mathbf{n}) q(\mathbf{n}, \mathbf{n} - \mathbf{e}_i + \mathbf{e}_j) - \pi(\mathbf{n} - \mathbf{e}_i + \mathbf{e}_j) q(\mathbf{n} - \mathbf{e}_i + \mathbf{e}_j, \mathbf{n}) \right\}, \quad i = 1, \ldots, J.$$

**Remark 3.1.6 (Product-form equilibrium distribution)** Observe that the equilibrium distribution of the open network of $M|M|1$ queues is a product of the marginal distributions $\pi_j(n_j) = (1 - \rho_j)\rho_j^{n_j}$, $n_j \in \mathbb{N}_0$, of the $M|M|1$ queues with arrival rate $\lambda_j$ and service rate $\mu_j$, $j = 1, \ldots, J$:

$$\pi(\mathbf{n}) = \prod_{j=1}^{J} \pi_j(n_j), \quad \mathbf{n} \in \mathbb{N}_0^J.$$

Thus, the random variables $N_j := N_j(\infty)$ recording the equilibrium number of customers in queue $j$, $j = 1, \ldots, J$, are independent random variables. Clearly, the processes $N_j(t)$, $j = 1, \ldots, J$, are not independent, also recall Remark 2.5.3.

The equilibrium distribution of the closed network of $M|M|1$ queues equals that of the open network except for normalisation. As a consequence, the random variables $N_j := N_j(\infty)$ for the closed network are not independent. The distribution for the open and closed network of $M|M|1$ queues are called product-form distributions. □

Theorems 3.1.4, 3.1.5 introduce the concept of *partial balance* that is an essential element in the analysis of networks of queues in this monograph. We have the following result that we include here to highlight the role of partial balance.

**Lemma 3.1.7 (Partial balance)** *Consider an open network of queues. A measure $m = (m(\mathbf{n}),\ \mathbf{n} \in S)$ that satisfies partial balance*

$$\sum_{j=0}^{J} \left\{ m(\mathbf{n})q(\mathbf{n},\mathbf{n}-\mathbf{e}_i+\mathbf{e}_j) - m(\mathbf{n}-\mathbf{e}_i+\mathbf{e}_j)q(\mathbf{n}-\mathbf{e}_i+\mathbf{e}_j,\mathbf{n}) \right\} = 0, \quad i = 0,\ldots,J, \tag{3.9}$$

*is an invariant measure.*

*Consider a closed network of queues. A measure $m = (m(\mathbf{n}),\ \mathbf{n} \in S)$ that satisfies partial balance*

$$\sum_{j=1}^{J} \left\{ m(\mathbf{n})q(\mathbf{n},\mathbf{n}-\mathbf{e}_i+\mathbf{e}_j) - m(\mathbf{n}-\mathbf{e}_i+\mathbf{e}_j)q(\mathbf{n}-\mathbf{e}_i+\mathbf{e}_j,\mathbf{n}) \right\} = 0, \quad i = 1,\ldots,J, \tag{3.10}$$

*is an invariant measure.*

**Proof.** Summation of the partial balance equations (3.9) over $i = 0,\ldots,J$ for the open network and of the partial balance equations (3.10) over $i = 1,\ldots,J$ for the closed network yields the global balance equations (1.14). □

**Proof of Theorems 3.1.4 and 3.1.5.** We will first consider the open network and show that $m(\mathbf{n}) = \prod_{k=1}^{J} \rho_k^{n_k}$, $\mathbf{n} \in S$, is an invariant measure using partial balance (3.9) and then invoke Theorem 1.1.4 to complete the proof. Inserting $m$ and the transition rates (3.1) into the partial balance equations (3.9) for the open network yields, for $i = 0,\ldots,J$, $\mathbf{n} \in \mathbb{N}_0^J$,

$$\begin{aligned}
&\sum_{j=0}^{J} \left\{ m(\mathbf{n})q(\mathbf{n},\mathbf{n}-\mathbf{e}_i+\mathbf{e}_j) - m(\mathbf{n}-\mathbf{e}_i+\mathbf{e}_j)q(\mathbf{n}-\mathbf{e}_i+\mathbf{e}_j,\mathbf{n}) \right\} \\
&\quad = \sum_{j=0}^{J} \left\{ \prod_{k=1}^{J} \rho_k^{n_k} \mu_i p_{ij} \mathbb{1}(\mathbf{n}-\mathbf{e}_i \in \mathbb{N}_0^J) - \prod_{k=1}^{J} \rho_k^{n_k-\delta_{ki}+\delta_{kj}} \mu_j p_{ji} \mathbb{1}(\mathbf{n}-\mathbf{e}_i \in \mathbb{N}_0^J) \right\},
\end{aligned}$$

where the indicator $\mathbb{1}(\mathbf{n}-\mathbf{e}_i \in \mathbb{N}_0^J)$ in the first term reflects that a customer cannot be served in queue $i$, $i = 1,\ldots,J$, when that queue is empty, and in the second term reflects that for the state $\mathbf{n}-\mathbf{e}_i+\mathbf{e}_j$ to be contained in $S$ it must be that $\mathbf{n}-\mathbf{e}_i \in \mathbb{N}_0^J$ and recall that $\mathbf{e}_0$ is the zero-vector.

First consider the partial balance equations for queue $i = 0$. Rearranging terms and inserting $\rho_i = \lambda_i/\mu_i$, $i = 1,\ldots,J$, yields

$$\begin{aligned}
&\sum_{j=0}^{J} \left\{ m(\mathbf{n})q(\mathbf{n},\mathbf{n}-\mathbf{e}_i+\mathbf{e}_j) - m(\mathbf{n}-\mathbf{e}_i+\mathbf{e}_j)q(\mathbf{n}-\mathbf{e}_i+\mathbf{e}_j,\mathbf{n}) \right\} \mathbb{1}(i=0) \\
&\quad = \left\{ \mu_0 - \sum_{j=1}^{J} \lambda_j p_{j0} \right\} \prod_{k=1}^{J} \rho_k^{n_k} \mathbb{1}(\mathbf{n} \in \mathbb{N}_0^J) = 0,
\end{aligned} \tag{3.11}$$

since $\mu_0 - \sum_{j=1}^{J} \lambda_j p_{j0} = 0$, recall (3.4).

Now consider the partial balance equations for queues $i \neq 0$. Rearranging terms yields

$$\sum_{j=0}^{J} \left\{ m(\mathbf{n})q(\mathbf{n},\mathbf{n}-\mathbf{e}_i+\mathbf{e}_j) - m(\mathbf{n}-\mathbf{e}_i+\mathbf{e}_j)q(\mathbf{n}-\mathbf{e}_i+\mathbf{e}_j,\mathbf{n}) \right\} \mathbb{1}(i \neq 0)$$

$$= \left\{ \sum_{j=0}^{J} \lambda_i p_{ij} - \mu_0 p_{0i} - \sum_{j=1}^{J} \lambda_j p_{ji} \right\} \prod_{k=1}^{J} \rho_k^{n_k - \delta_{ki}} \mathbb{1}(\mathbf{n}-\mathbf{e}_i \in \mathbb{N}_0^J)\mathbb{1}(i \neq 0) = 0, \quad (3.12)$$

since $\sum_{j=0}^{J} \lambda_i p_{ij} - \mu_0 p_{0i} - \sum_{j=1}^{J} \lambda_j p_{ji} = 0$, recall (3.3).

Irreducibility of $\{N(t)\}$ follows from irreducibility of the routing matrix. If $\rho_j < 1$, $j = 1,\dots,J$, then $m(\mathbf{n}) = \prod_{k=1}^{J} \rho_k^{n_k}$, $\mathbf{n} \in S$, has finite mass. Theorem 1.1.4 completes the proof of Theorem 3.1.4.

The proof of Theorem 3.1.5 follows from the proof above by setting $\mu_0 = 0$, $p_{j0} = 0$, $j = 1,\dots,J$. □

The normalising constant $G_o$ for the open network of $M|M|1$ queues is available in closed form. For the closed network of $M|M|1$ queues Buzen's algorithm provides an efficient recursion to evaluate the normalising constant $G_M$. The complexity of Buzen's algorithm is $O(JM)$.

**Algorithm 3.1.8 (Buzen's Algorithm)** *Define $G(m, j)$, $m = 0,\dots,M$, $j = 1,\dots,J$. Set*

$$G(0, j) = 1, \quad j = 1,\dots,J,$$
$$G(m,1) = \rho_1^m, \quad m = 0,\dots,M.$$

*For $j = 2,\dots,J$, $m = 1,\dots,M$, do*

$$G(m, j) = G(m, j-1) + \rho_j G(m-1, j). \quad (3.13)$$

*Then $G_M = G(M,J)^{-1}$.*

Buzen's algorithm yields

$$G(m, j) = \sum_{\{\mathbf{n}: n_1+\cdots+n_m=j\}} \prod_{i=1}^{m} \rho_j^{n_i},$$

the inverse of the normalising constant for the closed network of $j$ $M|M|1$ queues with $m$ customers, $m = 1,\dots,M$, $j = 1,\dots,J$. The recursion can readily be concluded observing that the first term in the right-hand side of (3.13) covers the case in which $n_m = 0$ and the second term the case $n_m > 0$.

The marginal distribution of the number of customers at queue $j$ in the network containing $M$ customers is

$$\pi_j(n_j) = G_M \rho_j^{n_j} [G_{M-n_j}^{-1} - \rho_j G_{M-n_j-1}^{-1}], \quad n_j = 0,\dots,M-1,$$
$$\pi_j(M) = G_M \rho_j^{n_j},$$

and the mean number of customers at queue $j$ is

$$\mathbb{E}[N_j] = \sum_{m=1}^{M} \rho_j^m \frac{G_M}{G_{M-m}}.$$

**Remark 3.1.9 (Jackson and Gordon-Newell networks)** Open networks of $M|M|1$ queues are often called Jackson networks, referring to J.R. Jackson who first obtained their equilibrium distribution [**?**] and closed networks of $M|M|1$ queues are called Gordon-Newell networks, referring to W.J. Gordon and G.F. Newell who obtained their equilibrium distribution [**?**]. In the exposition in this chapter we have introduced queue 0 to represent the outside of the network of queues which allows a unified analysis of open and closed networks. □

## 3.2 Kelly-Whittle networks

A closer examination of the partial balance equations (3.9), (3.10) and (3.11), (3.12) reveals that a multiplicative factor $\phi(\mathbf{n})^{-1}$ in the transition rates $q(\mathbf{n},\mathbf{n}')$ may be absorbed in the equilibrium distribution and that an additional function $\psi(\mathbf{n}-\mathbf{e}_i)$ in the transition rates $q(\mathbf{n},\mathbf{n}-\mathbf{e}_i+\mathbf{e}_j)$, $i,j=0,\dots,J$, is merely a constant in the partial balance equations for each fixed $i$, $i=0,\dots,J$. The Kelly-Whittle network provides precisely this generalisation of the Jackson network.

A *Kelly-Whittle network* is a Markov chain $\{N(t)\}$ at state space $S \subseteq \mathbb{N}_0^J$ with transition rates, for $\mathbf{n}' \neq \mathbf{n}$,

$$q(\mathbf{n},\mathbf{n}') = \begin{cases} \dfrac{\psi(\mathbf{n}-\mathbf{e}_i)}{\phi(\mathbf{n})}\mu_i p_{ij}, & \text{if } \mathbf{n}' = \mathbf{n}-\mathbf{e}_i+\mathbf{e}_j,\ i,j=0,\dots,J, \\ 0, & \text{otherwise,} \end{cases} \tag{3.14}$$

where $\psi : \mathbb{N}_0^J \to [0,\infty)$ and $\phi : \mathbb{N}_0^J \to (0,\infty)$. We will consider closed networks as special case of open networks with $\mu_0 = 0$ and $p_{i0} = 0$, $i = 1,\dots,J$. Note that we may absorb $\mu_i$, $i = 1,\dots,J$, in $\psi$ and $\phi$ via the transformation

$$\psi(\mathbf{n}) := \psi(\mathbf{n}) \prod_{j=1}^{J} \mu_i^{-n_i}, \quad \phi(\mathbf{n}) := \phi(\mathbf{n}) \prod_{j=1}^{J} \mu_i^{-n_i}, \quad \mathbf{n} \in \mathbb{N}_0^J.$$

We have the following result.

**Theorem 3.2.1 (Equilibrium distribution: Kelly-Whittle network)** *Consider the Kelly-Whittle network $\{N(t)\}$ at state space $S \subseteq \mathbb{N}_0^J$ with transition rates (3.14). Assume that the routing matrix $P = (p_{ij},\ i,j = 0,\dots,J)$ is irreducible and let $\{\lambda_j,\ j = 1,\dots,J\}$ be the solution of the traffic equations (3.3). Let $\rho_j = \lambda_j/\mu_j$, $j = 1,\dots,J$. Assume that*

$$G_{KW}^{-1} = \sum_{\mathbf{n}\in S} \phi(\mathbf{n}) \prod_{j=1}^{J} \rho_j^{n_j} < \infty,$$

*and that $\{N(t)\}$ is irreducible. Then $\{N(t)\}$ has unique equilibrium distribution*

$$\pi(\mathbf{n}) = G_{KW}\phi(\mathbf{n})\prod_{j=1}^{J}\rho_j^{n_j}, \quad \mathbf{n}\in S. \tag{3.15}$$

*Moreover, the equilibrium distribution (3.15) satisfies* partial balance*, for all* $\mathbf{n}\in S$*,*

$$\sum_{j=0}^{J}\left\{\pi(\mathbf{n})q(\mathbf{n},\mathbf{n}-\mathbf{e}_i+\mathbf{e}_j) - \pi(\mathbf{n}-\mathbf{e}_i+\mathbf{e}_j)q(\mathbf{n}-\mathbf{e}_i+\mathbf{e}_j,\mathbf{n})\right\} = 0, \quad i=0,\dots,J.$$

Observe that Theorem 3.2.1 includes the assumption that $\{N(t)\}$ is irreducible. A sufficient condition for irreducibility is that $\psi(\mathbf{n}) > 0$ for all $\mathbf{n}\in S\cap\mathbb{N}_0^J$.

**Remark 3.2.2 (Product-form equilibrium distribution)** The equilibrium distribution (3.15) is a product of a part determined by the service rates and a part determined by the routing probabilities:

$$\pi(\mathbf{n}) = G_{KW}\left(\phi(\mathbf{n})\prod_{j=1}^{J}\mu_j^{-n_j}\right)\left(\prod_{j=1}^{J}\lambda_j^{n_j}\right), \quad \mathbf{n}\in S,$$

which is often referred to as a product-form distribution. Recall that in Remark 3.1.6 the term product-form was introduced to represent a product over the queues. □

**Proof of Theorem 3.2.1.** We will first show that $m(\mathbf{n}) = \phi(\mathbf{n})\prod_{j=1}^{J}\rho_j^{n_j}$, $\mathbf{n}\in S$, is an invariant measure using partial balance (3.9). Theorem 1.1.4 then completes the proof.

Inserting $m$ and the transition rates (3.14) into the partial balance equations (3.9) for the open network yields, for $i=0,\dots,J$, $\mathbf{n}\in\mathbb{N}_0^J$, and denoting $\lambda_0=\mu_0$,

$$\begin{aligned}
&\sum_{j=0}^{J}\left\{m(\mathbf{n})q(\mathbf{n},\mathbf{n}-\mathbf{e}_i+\mathbf{e}_j) - m(\mathbf{n}-\mathbf{e}_i+\mathbf{e}_j)q(\mathbf{n}-\mathbf{e}_i+\mathbf{e}_j,\mathbf{n})\right\}\\
&= \sum_{j=0}^{J}\left\{\phi(\mathbf{n})\prod_{k=1}^{J}\rho_k^{n_k}\frac{\psi(\mathbf{n}-\mathbf{e}_i)}{\phi(\mathbf{n})}\mu_i p_{ij}\right.\\
&\qquad\left. -\phi(\mathbf{n}-\mathbf{e}_i+\mathbf{e}_j)\prod_{k=1}^{J}\rho_k^{n_k-\delta_{ki}+\delta_{kj}}\frac{\psi(\mathbf{n}-\mathbf{e}_i)}{\phi(\mathbf{n}-\mathbf{e}_i+\mathbf{e}_j)}\mu_j p_{ji}\right\},\\
&= \sum_{j=0}^{J}\left\{\lambda_i p_{ij} - \lambda_j p_{ji}\right\}\psi(\mathbf{n}-\mathbf{e}_i)\prod_{k=1}^{J}\rho_k^{n_k-\delta_{ki}} = 0,
\end{aligned} \tag{3.16}$$

where the last step follows observing that the term $\psi(\mathbf{n}-\mathbf{e}_i)\prod_{k=1}^{J}\rho_k^{n_k-\delta_{ki}}$ is a constant for the summation over $j$ and invoking the traffic equations (3.3). □

**Remark 3.2.3 (Base states)** Note that the indicator $\mathbb{1}(\mathbf{n}-\mathbf{e}_i\in\mathbb{N}_0^J)$ that appears in (3.12) to guarantee that the *base state* $\mathbf{m}=\mathbf{n}-\mathbf{e}_i$ of non-moving customers in the transition $\mathbf{n} \to \mathbf{n} - \mathbf{e}_i + \mathbf{e}_j$ is non-negative is not included in (3.16). The function $\psi : \mathbb{N}_0^J \to [0,\infty)$ takes this restriction into account. Further note that the base state $\mathbf{m} = \mathbf{n} - \mathbf{e}_i$ is the same for the transitions $\mathbf{n} \to \mathbf{n} - \mathbf{e}_i + \mathbf{e}_j$ and $\mathbf{n} - \mathbf{e}_i + \mathbf{e}_j \to \mathbf{n}$. The function $\psi$ is defined on the base states

$$S^b = \{\mathbf{m} \in \mathbb{N}_0^J : \exists i, j \in \{0,\ldots,J\},\ i \neq j \text{ s.t. } \mathbf{m}+\mathbf{e}_i, \text{ and } \mathbf{m}+\mathbf{e}_j \in S\},$$

whereas the function $\phi$ is defined on state space $S$. □

**Remark 3.2.4 (Poisson arrivals)** If the arrival process to the Kelly-Whittle network is a Poisson process it must be that the transition rates $q(\mathbf{n},\mathbf{n}+\mathbf{e}_j)$ are independent of the state $\mathbf{n}$, $\mathbf{n} \in S = \mathbb{N}_0^J$, $j = 1,\ldots,J$. Hence, it must be that $\psi(\mathbf{n}) = \phi(\mathbf{n})$, $\mathbf{n} \in \mathbb{N}_0^J$. □

**Example 3.2.5 (Independent queues)** A natural example is the network in which the service rate at queue $i$ only depends on the number of customers at queue $i$:

$$q(\mathbf{n},\mathbf{n}-\mathbf{e}_i+\mathbf{e}_j) = \begin{cases} \kappa_i(n_i)\mu_i p_{ij}, & i,j = 1,\ldots,J, \\ \kappa_i(n_i)\mu_i p_{i0}, & i = 1,\ldots,J,\ j = 0, \\ \mu_0 p_{0j}, & i = 0,\ j = 1,\ldots,J, \end{cases} \tag{3.17}$$

for $\kappa_i : \mathbb{N}_0 :\to (0,\infty)$, $i = 1,\ldots,J$. These transition rates may be written in the form (3.14) as follows. First, let $\eta_i : \mathbb{N}_0 :\to (0,\infty)$, $i = 1,\ldots,J$, be defined as

$$\eta_i(n)^{-1} = \prod_{r=1}^{n} \kappa_i(r), \quad n \in \mathbb{N}_0,\ i = 1,\ldots,J.$$

Then

$$\kappa_i(n) = \frac{\eta_i(n-1)}{\eta_i(n)}, \quad n \in \mathbb{N},\ i = 1,\ldots,J,$$

so that (3.17) has the form (3.14) with

$$\psi(\mathbf{n}) = \phi(\mathbf{n}) = \prod_{i=1}^{J} \eta_i(n_i), \quad \mathbf{n} \in N_0^J.$$

From (3.17) observe that each birth-death process with constant birth rates may be included. Typical examples are, for $n \in \mathbb{N}$, $i = 1,\ldots,J$,

$$\begin{aligned} \kappa_i(n) &= 1, && \text{single server queue,} \\ \kappa_i(n) &= \min(n,s), && s \text{ server queue,} \\ \kappa_i(n) &= n, && \text{infinite server queue.} \end{aligned}$$

□

**Example 3.2.6 (Reversible service rates; $\phi$-balance; balanced fairness)** Consider the Markov chain $\{N(t)\}$ at state space $S = \mathbb{N}_0^J$ with transition rates, for $\mathbf{n}' \neq \mathbf{n}$,

$$q(\mathbf{n},\mathbf{n}') = \begin{cases} \kappa_i(\mathbf{n})\mu_i p_{ij}, & \text{if } \mathbf{n}' = \mathbf{n}-\mathbf{e}_i+\mathbf{e}_j,\ i,j = 0,\ldots,J, \\ 0, & \text{otherwise,} \end{cases}$$

with $\kappa_i(\mathbf{n}) : S \to [0,\infty)$ such that $\kappa_i(\mathbf{n}) > 0$ if $n_i > 0$ and $\kappa_i(\mathbf{n}) = 0$ if $n_i = 0$, $i = 1,\dots,J$. The equilibrium distribution may be obtained in closed form (3.15) if the service rate function $\kappa_i(\mathbf{n})$ satisfies the *$\phi$-balance property*, that is if a function $\phi : S \to (0,\infty)$ exists such that for $\mathbf{n}$, $\mathbf{n}-\mathbf{e}_j+\mathbf{e}_k \in S$, $j,k = 0,\dots,J$,

$$\phi(\mathbf{n})\kappa_j(\mathbf{n}) = \phi(\mathbf{n}-\mathbf{e}_j+\mathbf{e}_k)\kappa_k(\mathbf{n}-\mathbf{e}_j+\mathbf{e}_k),$$

i.e., the service rate process at state space $S$ with transition rates $q(\mathbf{n},\mathbf{n}-\mathbf{e}_i+\mathbf{e}_j) = \kappa_i(\mathbf{n})$, $i,j = 0,\dots,J$, is reversible with invariant measure $\phi(\mathbf{n})$. We readily obtain that $\kappa_i(\mathbf{n})$ satisfies the *$\phi$-balance property* if and only if for some function $\psi : S^b \to (0,\infty)$, for $j = 0,\dots,J$,

$$\kappa_j(\mathbf{n}) = \frac{\psi(\mathbf{n}-\mathbf{e}_j)}{\phi(\mathbf{n})},$$

i.e., the service rate process of the Kelly-Whittle network is a reversible process, that is also referred to as *balanced fairness*.

Kolmogorov's criterion (2.4) or (2.6) implies that the service rate function $\kappa_i(\mathbf{n})$ satisfies the *$\phi$-balance property* if and only if

$$\phi(\mathbf{n}) = \phi(\mathbf{n}')\prod_{i=1}^{r}\frac{\kappa_{j_i}(\mathbf{n}_i)}{\kappa_{k_i}(\mathbf{n}_{i+1})}$$

for arbitrary $\mathbf{n}' \in S$ for all $r \in \mathbb{N}$ and any path $\mathbf{n}_1,\mathbf{n}_2,\dots,\mathbf{n}_r \in S$ such that $\mathbf{n}_1 = \mathbf{n}'$, $\mathbf{n}_r = \mathbf{n}$, and $\mathbf{n}_{i+1} = \mathbf{n}_i - \mathbf{e}_{j_i} + \mathbf{e}_{k_i}$, $i = 1,\dots,r-1$. □

**Remark 3.2.7 (Customer types)** Consider a queueing network of $J$ queues labelled $1,2\dots,J$ in which customers of types $u = 1,\dots,U$ route among the queues. The evolution of the number of customers of different types in the queues is recorded by the Markov chain $\{N(t),\ t \in \mathbb{R}\}$ at state space $S \subseteq \mathbb{N}_0^{J\times U}$ with states $\mathbf{n} = (\mathbf{n}_1,\dots,\mathbf{n}_J)$, $\mathbf{n}_j = (n_j(u),\ u = 1,\dots,U)$, with $n_j(u)$ denoting the number of customers of type $u$ in queue $j$. Let $\mathbf{e}_j(u)$ denote the $ju$-th unit vector. Let $p_{ij}(u,u')$ denote the fraction of customers of type $u$ that upon service completion in queue $i$ route to queue $j$ and turn into a type $u'$ customer, $i,j = 0,\dots,J$, $u,u' = 1,\dots,U$.

A *Kelly-Whittle network* with multiple customer types is a Markov chain $\{N(t)\}$ at state space $S \subseteq \mathbb{N}_0^{J\times U}$ with transition rates, for $\mathbf{n}' \neq \mathbf{n}$,

$$q(\mathbf{n},\mathbf{n}') = \begin{cases} \dfrac{\psi(\mathbf{n}-\mathbf{e}_i(u))}{\phi(\mathbf{n})}\mu_i(u)p_{ij}(u,u'), & \text{if } \mathbf{n}' = \mathbf{n}-\mathbf{e}_i(u)+\mathbf{e}_j(u'), \\ & i,j = 0,\dots,J,\ u,u' = 1,\dots,U, \\ 0, & \text{otherwise,} \end{cases}$$

where $\psi : \mathbb{N}_0^{J\times U} \to [0,\infty)$ and $\phi : \mathbb{N}_0^{J\times U} \to (0,\infty)$. Assume that the routing matrix $P = (p_{ij}(u,u'),\ i,j = 0,\dots,J,\ u,u' = 1,\dots,U)$ is irreducible and let $\{\lambda_j(u),\ j = 1,\dots,J,\ u = 1,\dots,U\}$ be the solution of the corresponding traffic equations (3.3). Let $\rho_i(u) = \lambda_i(u)/\mu_i(u)$, $i = 1,\dots,J$, $u = 1,\dots,U$. Assume that

$$G_{KW}^{-1} = \sum_{\mathbf{n}\in S} \phi(\mathbf{n}) \prod_{j=1}^{J} \prod_{u=1}^{U} \rho_j(u)^{n_j(u)} < \infty,$$

and that $\{N(t)\}$ is irreducible. Then $\{N(t)\}$ has unique equilibrium distribution

$$\pi(\mathbf{n}) = G_{KW}\,\phi(\mathbf{n}) \prod_{j=1}^{J} \prod_{u=1}^{U} \rho_j(u)^{n_j(u)}, \quad \mathbf{n}\in S.$$

□

## 3.3 Partial balance

Define for $\mathbf{n}\in S$ a collection of mutually exclusive sets $\{A_k(\mathbf{n}),\ k\in I(\mathbf{n})\}$, $I(\mathbf{n})\subseteq\mathbb{N}$, such that $\bigcup_{k\in I(\mathbf{n})} A_k(\mathbf{n}) = S$. A Markov chain is *partially balanced over* $\{A_k(\mathbf{n}),\ k\in I(\mathbf{n})\}$ if a distribution $\pi = (\pi(\mathbf{n}),\ \mathbf{n}\in S)$ exists such that for all $\mathbf{n}\in S$, $k\in I(\mathbf{n})$,

$$\sum_{\mathbf{n}'\in A_k(\mathbf{n})} \left\{\pi(\mathbf{n})q(\mathbf{n},\mathbf{n}') - \pi(\mathbf{n}')q(\mathbf{n}',\mathbf{n})\right\} = 0. \tag{3.18}$$

The following result follows by summation of (3.18) over $k\in I(\mathbf{n})$.

**Theorem 3.3.1 (Partial balance)** *A distribution $\pi = (\pi(\mathbf{n}),\ \mathbf{n}\in S)$ satisfying the partial balance equations (3.18) is a stationary distribution.*

For Kelly-Whittle networks we may identify the following nested set of balance equations, for $\mathbf{n}\in S$:

Transition balance:

$$q(\mathbf{n},\mathbf{n}-\mathbf{e}_i+\mathbf{e}_j) = q(\mathbf{n}-\mathbf{e}_i+\mathbf{e}_j,\mathbf{n}), \quad i,j=0,\ldots,J; \tag{3.19}$$

Detailed balance:

$$\pi(\mathbf{n})q(\mathbf{n},\mathbf{n}-\mathbf{e}_i+\mathbf{e}_j) = \pi(\mathbf{n}-\mathbf{e}_i+\mathbf{e}_j)q(\mathbf{n}-\mathbf{e}_i+\mathbf{e}_j,\mathbf{n}), \quad i,j=0,\ldots,J; \tag{3.20}$$

Partial balance:

$$\sum_{j=0}^{J} \pi(\mathbf{n})q(\mathbf{n},\mathbf{n}-\mathbf{e}_i+\mathbf{e}_j) = \sum_{j=0}^{J} \pi(\mathbf{n}-\mathbf{e}_i+\mathbf{e}_j)q(\mathbf{n}-\mathbf{e}_i+\mathbf{e}_j,\mathbf{n}), \quad i=0,\ldots,J; \tag{3.21}$$

Global balance:

$$\sum_{i,j=0}^{J} \pi(\mathbf{n})q(\mathbf{n},\mathbf{n}-\mathbf{e}_i+\mathbf{e}_j) = \sum_{i,j=0}^{J} \pi(\mathbf{n}-\mathbf{e}_i+\mathbf{e}_j)q(\mathbf{n}-\mathbf{e}_i+\mathbf{e}_j,\mathbf{n}). \tag{3.22}$$

Obviously, transition balance implies detailed balance, which implies partial balance, which in turn implies global balance. The balance equations have the following clear interpretation:

> Detailed balance states that the probability flow out of state **n** due to a customer served at queue *i* that routes queue *j* balances with the probability flow into state **n** due to a customer served at queue *j* that routes to queue *i*.
>
> Partial balance states that the probability flow out of state **n** due to a customer served at queue *i* balances with the probability flow into state **n** due to a customer routing to queue *i*.
>
> Global balance states that the probability flow out of state **n** due to a customer served at some queue routing to some other queue balances with the probability flow into state **n** due to a customer served at some queue routing to some other queue.

The sets $A_k(\mathbf{n})$, $\mathbf{n} \in S$, are for $\mathbf{n} \in S$:

Global balance (3.22):

$$\begin{aligned} A_1(\mathbf{n}) &= \bigcup_{i,j=0}^{J} \{\mathbf{n} - \mathbf{e}_i + \mathbf{e}_j\}, \\ A_2(\mathbf{n}) &= S \setminus A_1(\mathbf{n}) \text{ and observe that } q(\mathbf{n},\mathbf{n}') = 0 \text{ for } \mathbf{n}' \in A_2(\mathbf{n}), \end{aligned}$$

Partial balance (3.21):

$$\begin{aligned} A_i(\mathbf{n}) &= \bigcup_{j=0}^{J} \{\mathbf{n} - \mathbf{e}_i + \mathbf{e}_j\},\ i = 0,\dots,J, \\ A_{J+1}(\mathbf{n}) &= S \setminus \bigcup_{i=0}^{J} A_i(\mathbf{n}) \text{ and observe that } q(\mathbf{n},\mathbf{n}') = 0 \text{ for } \mathbf{n}' \in A_{J+1}(\mathbf{n}), \end{aligned}$$

Detailed balance (3.20) and transition balance (3.19):

$$\begin{aligned} A_{i,j}(\mathbf{n}) &= \{\mathbf{n} - \mathbf{e}_i + \mathbf{e}_j\},\ i,j = 0,\dots,J, \\ A_{J\cdot J+1}(\mathbf{n}) &= S \setminus \bigcup_{i,j=0}^{J} A_{i,jk}(\mathbf{n}) \text{ and observe that } q(\mathbf{n},\mathbf{n}') = 0 \text{ for } \mathbf{n}' \in A_{J\cdot J+1}(\mathbf{n}), \end{aligned}$$

where we have used $i,j$ instead of $k$ in the labelling of the sets for detailed and transition balance for ease of notation.

Observe that the traffic equations (3.3) are the key-element in the proof of Theorem 3.2.1. Conversely, if $\pi$ given in (3.15) satisfies partial balance (3.21) then the traffic equations must be satisfied, which is readily obtained observing that the terms involging $\phi$ cancel.

**Theorem 3.3.2 (Partial balance and the traffic equations)** *The traffic equations (3.3) are a necessary and sufficient condition for $\pi$ given in (3.15) to satisfiy partial balance (3.21).*

**Remark 3.3.3 (Interpretation of the traffic equations)** The solution of traffic equations (2.9) for the feedforward network of Section 2.5 determines the rates of the Poisson arrival processes to the queues. For closed networks of $M|M|1$ queues and Kelly-Whittle networks the flows of customers among the queues are not Poisson. However, also for a Kelly-Whittle network with Poisson arrivals we may interpret the solution $\lambda_j$, $j = 1,\ldots,J$, of traffic equations as the arrival rate of customers.

The number of customers that route from queue $i$ to queue $k$ in the time-interval $(0,t]$ is

$$H_{ij}(t) = \sum_{k=0}^{\infty} \mathbb{1}(N_{\tau_k} = N_{\tau_{k-1}} - \mathbf{e}_i + \mathbf{e}_j,\ \tau_k \in (0,t]),$$

where $0 = \tau_0 < \tau_1 < \tau_2 < \cdots$ are the transition epochs of $\{N(t)\}$. The average number of customers moving from queue $i$ to queue $j$ is, recall (1.15),

$$\lambda_{ij} = \lim_{T\to\infty} \frac{H_{ij}(T)}{T} = \sum_{\mathbf{n}\in S} \pi(\mathbf{n}) q(\mathbf{n}, \mathbf{n} - \mathbf{e}_i + \mathbf{e}_j), \quad i,j = 0,\ldots,J.$$

Consider the network with Poisson arrivals, so that $\psi(\mathbf{n}) = \phi(\mathbf{n})$, $\mathbf{n} \in \mathbb{N}_0^J$, recall Remark 3.2.4. The average number of customers arriving to queue $j$ from queue $i$ is, with $\lambda_0 = \mu_0$,

$$\begin{aligned}\lambda_{ij} &= \sum_{\mathbf{n}\in S} G_{KW} \phi(\mathbf{n}) \prod_{j=1}^{J} \rho_j^{n_j} \frac{\phi(\mathbf{n}-\mathbf{e}_i)}{\phi(\mathbf{n})} \mu_i p_{ij} \\ &= \lambda_i p_{ij} \sum_{\mathbf{n}\in S, n_i>0} G_{KW} \phi(\mathbf{n}-\mathbf{e}_i) \prod_{j=1}^{J} \rho_j^{n_j - \delta_{ij}} = \lambda_i p_{ij}.\end{aligned}$$

□

Partial balance plays a crucial role in the analysis of networks of queues. As an illustration, consider the open network of 2 $M|M|1$ queues, with Poisson arrival rate $\mu_0$, service rates $\mu_i$ at queue $i$, $i = 1,2$ and routing probabilities $p_{ij}$, $i,j = 0,1,2$. The transition diagram is depicted in Figure 3.1. Global balance states that for each $\mathbf{n} \in \mathbb{N}_0^2$ the total probability flow along all transitions out of state $\mathbf{n}$ equals the total probability flow along all transitions into state $\mathbf{n}$. Partial balance breaks this balance of flows into triangles: partial balance for queue 0 (in red in Figure 3.1) balances the flows from $\mathbf{n}$ to $\mathbf{n} - \mathbf{e}_0 + \mathbf{e}_1$ and $\mathbf{n} - \mathbf{e}_0 + \mathbf{e}_2$ (out of queue 0) with the flow to $\mathbf{n}$ from $\mathbf{n} + \mathbf{e}_1$ and $\mathbf{n} + \mathbf{e}_2$ (into queue 0):

$$\sum_{j\in\{1,2\}} \pi(\mathbf{n}) q(\mathbf{n}, \mathbf{n} - \mathbf{e}_0 + \mathbf{e}_j) = \sum_{j\in\{1,2\}} \pi(\mathbf{n} - \mathbf{e}_0 + \mathbf{e}_j) q(\mathbf{n} - \mathbf{e}_i + \mathbf{e}_j, \mathbf{n});$$

partial balance for queue 1 (in blue in Figure 3.1) balances the flows from $\mathbf{n}$ to $\mathbf{n} - \mathbf{e}_1 + \mathbf{e}_0$ and $\mathbf{n} - \mathbf{e}_1 + \mathbf{e}_2$ (out of queue 1) with the flow to $\mathbf{n}$ from $\mathbf{n} - \mathbf{e}_1 + \mathbf{e}_0$ and

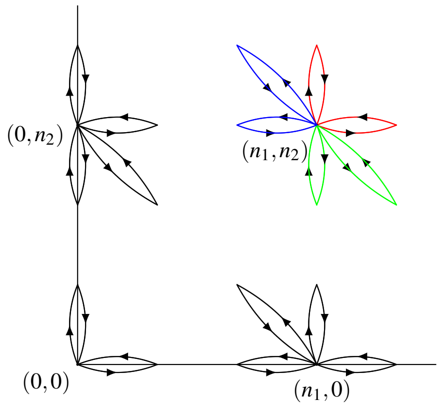

**Fig. 3.1** Open network of two $M|M|1$ queues.

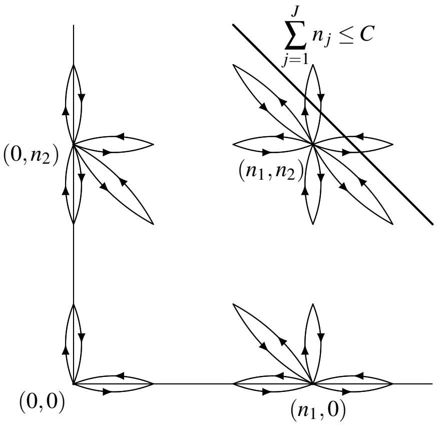

**Fig. 3.2** Open network with finite capacity.

$\mathbf{n}-\mathbf{e}_1+\mathbf{e}_2$ (into queue 1):

$$\sum_{j\in\{0,2\}} \pi(\mathbf{n})q(\mathbf{n},\mathbf{n}-\mathbf{e}_1+\mathbf{e}_j) = \sum_{j\in\{0,2\}} \pi(\mathbf{n}-\mathbf{e}_1+\mathbf{e}_j)q(\mathbf{n}-\mathbf{e}_1+\mathbf{e}_j,\mathbf{n}),$$

partial balance for queue 2 (in green in Figure 3.1) balances the flows from $\mathbf{n}$ to $\mathbf{n}-\mathbf{e}_2+\mathbf{e}_0$ and $\mathbf{n}-\mathbf{e}_2+\mathbf{e}_1$ (out of queue 2) with the flow to $\mathbf{n}$ from $\mathbf{n}-\mathbf{e}_2+\mathbf{e}_0$ and $\mathbf{n}-\mathbf{e}_2+\mathbf{e}_1$ (into queue 2):

$$\sum_{j\in\{0,1\}} \pi(\mathbf{n})q(\mathbf{n},\mathbf{n}-\mathbf{e}_2+\mathbf{e}_j) = \sum_{j\in\{0,1\}} \pi(\mathbf{n}-\mathbf{e}_2+\mathbf{e}_j)q(\mathbf{n}-\mathbf{e}_2+\mathbf{e}_j,\mathbf{n}).$$

Observe that at the boundary $\mathbf{n}=(n_1,0)$, $n_1\in\mathbb{N}$, departures from queue 2 are prohibited and transitions to $(n_1,0)$ due to an arrival of a customer to queue 2 cannot occur: there remain four transitions out of state $(n_1,0)$ and four transitions into state $(n_1,0)$. These transitions coincide with those of partial balance for queue 0 and partial balance for queue 1. The probability flow from state $(n_1,0)$ due to a customer departing from queue 2 equals 0 as and the probability flow to state $(n_1,0)$ due to a customer arriving queue 2 equals 0.[3] Global balance in state $(n_1,0)$ is satisfied as a consequence of partial balance for queues 0 and 1. At the boundary $(0,n_2)$ global balance is satisfied as a consequence of partial balance for queues 0 and 2, and at the origin $(0,0)$ global balance is satisfied as a consequence of partial balance for queue 0.

Partial balance allows us to incorporate state space restrictions. Let $\{N(t)\}$ record the number of customers in the queues of an open network of $M|M|1$ queues in which the total number of customers is restricted not to exceed $c$ and arrivals finding $c$ customers in the network are blocked and cleared. The state space is $S=S_c=\{\mathbf{n}\in$

[3] We may also state that partial balance for queue 2 in state $(n_1,0)$ reads $0=0$.

$\mathbb{N}_0^J : n_j \geq 0,\ j = 1,\dots,J,\ \sum_{j=1}^J n_j \leq c\}$ and the transition rates are, for $\mathbf{n}' \neq \mathbf{n}$,

$$q(\mathbf{n},\mathbf{n}') = \begin{cases} \mu_0 p_{0j}, & \text{if } \mathbf{n}' = \mathbf{n} + \mathbf{e}_j, \text{ and } \sum_{j=1}^J n_j < c,\ j = 0,\dots,J, \\ \mu_i p_{ij}, & \text{if } \mathbf{n}' = \mathbf{n} - \mathbf{e}_i + \mathbf{e}_j, \text{ and } n_i > 0,\ i,j = 1,\dots,J, \\ \mu_i p_{i0}, & \text{if } \mathbf{n}' = \mathbf{n} - \mathbf{e}_i, \text{ and } n_i > 0,\ i = 1,\dots,J, \\ 0, & \text{otherwise.} \end{cases}$$

Observe that at the boundary $\sum_{j=1}^J n_j = c$ partial balance for queue 0 is "removed", but that partial balance for queues $j = 1,\dots,J$ remains satisfied by the invariant measure $m(\mathbf{n}) = \prod_{j=1}^J \rho_j^{n_j}$, $\mathbf{n} \in S_c$, also see Figure 3.2 for $J = 2$. Hence, with $\{\lambda_j,\ j = 1,\dots,J\}$ the unique solution of the traffic equations (3.3), $\{N(t)\}$ has unique equilibrium distribution

$$\pi(\mathbf{n}) = G_c \prod_{j=1}^J \rho_j^{n_j}, \quad \mathbf{n} \in S_c = \{\mathbf{n} \in \mathbb{N}_0^J : n_j \geq 0,\ j = 1,\dots,J,\ \sum_{j=1}^J n_j \leq c\},$$

where

$$G_c = \left[\sum_{\mathbf{n}\in S_c} \prod_{j=1}^J \rho_i^{n_j}\right]^{-1}.$$

The truncation result is similar to that in Theorem 2.2.5 for reversible Markov chains. The result may be generalised by analogy to that of Theorem 2.2.5 for a cut between $V$ and $S \setminus V$ such that all partial balance equations involving states $\mathbf{n} \in V$ and $\mathbf{n}' \in S \setminus V$ remain satisfied.

**Theorem 3.3.4 (Truncation)** *Consider Markov chain $\{N(t)\}$ at state space $S$ with transition rates $q(\mathbf{n},\mathbf{n}')$, $\mathbf{n},\mathbf{n}' \in S$, and equilibrium distribution $\pi$. Let $V \subset S$. Let $0 \leq r < 1$ and suppose that the transition rates are altered from $q(\mathbf{n},\mathbf{n}')$ to $rq(\mathbf{n},\mathbf{n}')$ for $\mathbf{n} \in V$, $\mathbf{n}' \in S \setminus V$. The resulting Markov chain $\{N_r(t)\}$ has equilibrium distribution*

$$\pi_r(\mathbf{n}) = \begin{cases} G\pi(\mathbf{n}), & \mathbf{n} \in V, \\ Gr\pi(\mathbf{n}), & \mathbf{n} \in S \setminus V, \end{cases}$$

*where $G$ is the normalizing constant, if and only if $\pi$ satisfies*

$$\sum_{\mathbf{n}'\in S\setminus V} \pi(\mathbf{n})q(\mathbf{n},\mathbf{n}') = \sum_{\mathbf{n}'\in S\setminus V} \pi(\mathbf{n}')q(\mathbf{n}',\mathbf{n}), \quad \mathbf{n} \in V.$$

**Remark 3.3.5 (Backward partial balance; networks with vacancies)** Partial balance (3.21) seems the obvious choice for $\{N(t)\}$ recording the number of customers in the queues of a network. However, in the global balance equations (3.22) we might also consider the terms for fixed $j$. This is referred to as *backward partial balance*, for each $\mathbf{n} \in S$,

$$\sum_{i=0}^{J} \pi(\mathbf{n})q(\mathbf{n},\mathbf{n}-\mathbf{e}_i+\mathbf{e}_j) = \sum_{i=0}^{J} \pi(\mathbf{n}-\mathbf{e}_i+\mathbf{e}_j)q(\mathbf{n}-\mathbf{e}_i+\mathbf{e}_j,\mathbf{n}), \quad j=0,\dots,J,$$

and may be interpreted as follows:

> Backward partial balance states that the probability flow out of state **n** due to a customer arriving to queue $j$ balances with the probability flow into state **n** due to a customer being served at queue $j$.

Backward partial balance represents the flow of empty spaces in a queueing network containing finite capacity queues. □

## 3.4 State-dependent routing and blocking protocols

A further examination of the proof of Theorem 3.2.1reveals that we may generalise Markovian routing to state-dependent routing provided that the *state-dependent traffic equations* that will be introduced below have a non-negative solution. A *Kelly-Whittle network with state-dependent routing* is a Markov chain $\{N(t)\}$ at state space $S \subseteq \mathbb{N}_0^J$ with transition rates, for $\mathbf{n}' \neq \mathbf{n}$,

$$q(\mathbf{n},\mathbf{n}') = \begin{cases} \dfrac{\psi(\mathbf{n}-\mathbf{e}_i)\theta_i(\mathbf{n}-\mathbf{e}_i)}{\phi(\mathbf{n})}\mu_i b_{ij}(\mathbf{n}-\mathbf{e}_i), & \text{if } \mathbf{n}'=\mathbf{n}-\mathbf{e}_i+\mathbf{e}_j,\ i,j=0,\dots,J, \\ 0, & \text{otherwise,} \end{cases} \tag{3.23}$$

where $\phi : S \to (0,\infty)$ and $\psi, \theta_i, b_{ij} : S^b \to [0,\infty)$, and $S^b$ is the set of base states:

$$S^b = \{\mathbf{m} \in \mathbb{N}_0^J : \exists i,j \in \{0,\dots,J\},\ i \neq j \text{ s.t. } \mathbf{m}+\mathbf{e}_i \text{ and } \mathbf{m}+\mathbf{e}_j \in S\}.$$

We have the following result.

**Theorem 3.4.1 (Equilibrium distribution: Kelly-Whittle network with state-dependent routing)** *Consider the Kelly-Whittle network with state-dependent routing $\{N(t)\}$ at state space $S \subseteq \mathbb{N}_0^J$ with transition rates (3.23). Assume that a solution $H : S \to [0,\infty)$ exists of the state-dependent traffic equations, for $\mathbf{n} \in S$, $i = 0,\dots,J$:*

$$\sum_{j=0}^{J} H(\mathbf{n})\theta_i(\mathbf{n}-\mathbf{e}_i)b_{ij}(\mathbf{n}-\mathbf{e}_i) = H(\mathbf{n}-\mathbf{e}_i)\theta_0(\mathbf{n}-\mathbf{e}_i)\mu_0 b_{0i}(\mathbf{n}-\mathbf{e}_i) + \sum_{j=1}^{J} H(\mathbf{n}-\mathbf{e}_i+\mathbf{e}_j)\theta_j(\mathbf{n}-\mathbf{e}_i)b_{ji}(\mathbf{n}-\mathbf{e}_i). \tag{3.24}$$

*Assume that*

$$G^{-1} = \sum_{\mathbf{n}\in S} \phi(\mathbf{n}) \prod_{j=1}^{J} \left(\frac{1}{\mu_j}\right)^{n_j} H(\mathbf{n}) < \infty,$$

*and that $\{N(t)\}$ is irreducible. Then $\{N(t)\}$ has unique equilibrium distribution*

$$\pi(\mathbf{n}) = G\phi(\mathbf{n}) \prod_{j=1}^{J} \left(\frac{1}{\mu_j}\right)^{n_j} H(\mathbf{n}), \quad \mathbf{n} \in S. \tag{3.25}$$

*Moreover, the equilibrium distribution (3.25) satisfies partial balance, for all* $\mathbf{n} \in S$*,*

$$\sum_{j=0}^{J} \left\{ \pi(\mathbf{n}) q(\mathbf{n}, \mathbf{n} - \mathbf{e}_i + \mathbf{e}_j) - \pi(\mathbf{n} - \mathbf{e}_i + \mathbf{e}_j) q(\mathbf{n} - \mathbf{e}_i + \mathbf{e}_j, \mathbf{n}) \right\} = 0, \quad i = 0, \ldots, J. \tag{3.26}$$

**Remark 3.4.2 (Routing function)** The function $b_{ij} : S^b \to [0, \infty)$, $i, j = 0, \ldots, J$, represents routing of customers among the queues. Without loss of generality, we may assume

$$b_i(\mathbf{m}) := \sum_{j=0}^{J} b_{ij}(\mathbf{m}) = 1, \quad \mathbf{m} \in S^b, \ i = 0, \ldots, J,$$

as we may absorb $b_i(\mathbf{m})$ in $\theta_i(\mathbf{m})$ for all $\mathbf{m} \in S^b$. □

**Remark 3.4.3 (Product-form equilibrium distribution)** The distribution (3.25) is referred to as product-form distribution as it is a product of a part determined by the service rates, $\phi(\mathbf{n}) \prod_{j=1}^{J} \mu_j^{-n_j}$, and a part determined by the routing function, $H(\mathbf{n})$, also recall Remark 3.2.2. □

**Remark 3.4.4 (State-dependent traffic equations)** Observe that the state-dependent traffic equations (3.24) are just as difficult to solve as the partial balance equations (3.26). In applications, often the routing function contains the Markov routing probabilities $p_{ij}$, $i, j = 0, \ldots, J$, and a function of the base state:

$$b_{ij}(\mathbf{m}) = p_{ij} f(\mathbf{m}), \quad \mathbf{m} \in S^b,$$

for some $f : S^b \to [0, \infty)$. With $\lambda_j$, $j = 1, \ldots, J$, the solution of the traffic equations (3.3), often $H(\mathbf{n}) = \prod_{j=1}^{J} \lambda_j^{n_j}$, $\mathbf{n} \in S$, is a solution of the state-dependent traffic equations (3.24). □

**Proof of Theorem 3.4.1.** It is sufficient to show that $\pi$ satisfies partial balance (3.26). Theorem 1.1.4 then completes the proof.

Inserting $\pi$ and the transition rates (3.23) into the partial balance equations (3.26) yields, for $i = 0, \ldots, J$, $\mathbf{n} \in \mathbb{N}_0^J$,

$$
\begin{aligned}
&\sum_{j=0}^{J}\left\{\pi(\mathbf{n})q(\mathbf{n},\mathbf{n}-\mathbf{e}_i+\mathbf{e}_j)-\pi(\mathbf{n}-\mathbf{e}_i+\mathbf{e}_j)q(\mathbf{n}-\mathbf{e}_i+\mathbf{e}_j,\mathbf{n})\right\}\\
&=\sum_{j=0}^{J}\left\{\prod_{k=1}^{J}\left(\frac{1}{\mu_k}\right)^{n_k}H(\mathbf{n})\psi(\mathbf{n}-\mathbf{e}_i)\theta_i(\mathbf{n}-\mathbf{e}_i)\mu_i b_{ij}(\mathbf{n}-\mathbf{e}_i)\right.\\
&\qquad\left.-\prod_{k=1}^{J}\left(\frac{1}{\mu_k}\right)^{n_k-\delta_{ki}+\delta_{kj}}H(\mathbf{n}-\mathbf{e}_i+\mathbf{e}_j)\psi(\mathbf{n}-\mathbf{e}_i)\theta_j(\mathbf{n}-\mathbf{e}_i)\mu_j b_{ji}(\mathbf{n}-\mathbf{e}_i)\right\},\\
&=\left\{H(\mathbf{n})\theta_i(\mathbf{n}-\mathbf{e}_i)b_{ij}(\mathbf{n}-\mathbf{e}_i)-H(\mathbf{n}-\mathbf{e}_i)\theta_0(\mathbf{n}-\mathbf{e}_i)\mu_0 b_{0i}(\mathbf{n}-\mathbf{e}_i)\right.\\
&\qquad\left.-\sum_{j=1}^{J}H(\mathbf{n}-\mathbf{e}_i+\mathbf{e}_j)\theta_j(\mathbf{n}-\mathbf{e}_i)b_{ji}(\mathbf{n}-\mathbf{e}_i)\right\}\psi(\mathbf{n}-\mathbf{e}_i)\prod_{k=1}^{J}\left(\frac{1}{\mu_k}\right)^{n_k-\delta_{ki}}=0,
\end{aligned}
$$

where the last step follows observing that the term $\psi(\mathbf{n}-\mathbf{e}_i)\prod_{k=1}^{J}\rho_k^{n_k-\delta_{ki}}$ is a constant for the summation over $j$ and invoking the state-dependent traffic equations (3.24). □

**Example 3.4.5 (Capacity constraints: no product-form)** Consider the open tandem Jackson network of 2 queues, where queue 1 has capacity restriction $c_1$. As in the $M|M|1|c_1$ queue, if queue 1 contains $c_1$ customers a customer arriving to queue 1 is discarded. The state space is $S=\{\mathbf{n}: 0\leq n_1\leq c_1,\ 0\leq n_2\}$. In the transition diagram, in state $(c_1,n_2)$ transitions from state $(c_1,n_2)$ to state $(c_1+1,n_2)$ (arrivals are discarded) and from state $(c_1+1,n_2-1)$ to state $(c_1,n_2)$ (this state is not in $S$) are removed. As a consequence, in state $(c_1,n_2)$ partial balance is not satisfied for queue 0 and queue 2. We may verify that a product-form distribution does *not* satisfy global balance. The product-form preserving blocking protocols introduced below modify the transition rates such that partial balance remains valid and the equilibrium distribution is a product-form distribution. □

**Example 3.4.6 (Stop-protocol)** Consider the open Kelly-Whittle network of 2 queues with capacity constraint $c_1$ at queue 1. Arrivals to queue 1 are discarded when $n_1=c_1$. The transition diagram is depicted in Figure 3.3, where the transitions crossing the vertical line indicating the region $n_1\leq c$ are to be deleted. As a consequence, partial balance is not satisfied for states $(c_1,n_2)$, $n_2=0,1,\ldots$. Partial balance is restored when the other transitions of partial balance for queue 0 and queue 2 are removed in states $(c_1,n_2)$, $n_2=0,1,\ldots$, as depicted in Figure 3.4. This modification is referred to as the *stop-protocol*. For general Kelly-Whittle networks with transition rates (3.14) and finite capacity constraints $n_j\leq c_j$, $j=1,\ldots,J$, the stop-protocol is as follows:

**Stop-protocol:** If queue $i$ in a Kelly-Whittle network with finite capacity constraints becomes saturated ($n_i=c_i$) then stop service at *all* other queues $j=1,\ldots,J$, $j\neq i$, and stop the arrival process to the network.

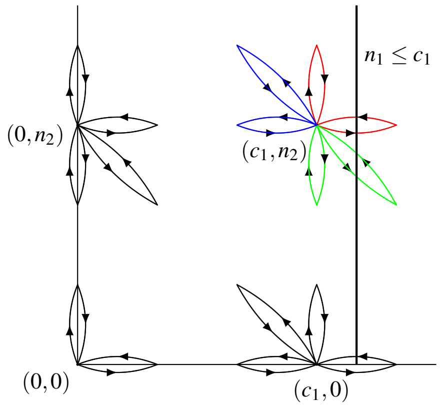

**Fig. 3.3** Two queues with finite capacity.

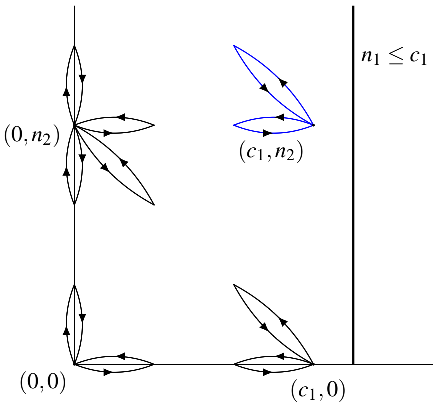

**Fig. 3.4** Two queues: stop-protocol.

A direct consequence of the stop-protocol is that two queues cannot become saturated simultaneously. For the open network the state space is

$$S_{\mathbf{c},o} = \{\mathbf{n} \in \mathbb{N}_0^J : 0 \le n_j \le c_j,\ 0 \le n_i + n_j < c_i + c_j,\ i \ne j,\ i,j = 1,\ldots,J\}$$

and for the closed network containing $M$ customers

$$S_{\mathbf{c},M} = \{\mathbf{n} \in \mathbb{N}_0^J : 0 \le n_j \le c_j,\ 0 \le n_i + n_j < c_i + c_j,\ i \ne j,\ i,j = 1,\ldots,J,\ \sum_{j=1}^{J} n_j = M\}.$$

Under the stop-protocol, the transition rates (3.14) have the form (3.23) with

$$\begin{aligned} \theta_i(\mathbf{m}) &= 1, \quad i = 0,\ldots,J,\ \mathbf{m} \in S_{\mathbf{c}}^b, \\ f(\mathbf{m}) &= \mathbb{1}(m_j < c_j,\ j = 1,\ldots,J), \quad \mathbf{m} \in S_{\mathbf{c}}^b, \\ b_{ij}(\mathbf{m}) &= p_{ij} f(\mathbf{m}), \quad i,j = 0,\ldots,J,\ \mathbf{m} \in S_{\mathbf{c}}^b \end{aligned}$$

and

$$S_{\mathbf{c}}^b = S_{\mathbf{c},o}^b = \{\mathbf{m} \in \mathbb{N}_0^J : 0 \le m_j \le c_j - 1\}$$

for open networks and

$$S_{\mathbf{c}}^b = S_{\mathbf{c},M}^b = \{\mathbf{m} \in \mathbb{N}_0^J : 0 \le m_j \le c_j - 1,\ \sum_{j=1}^{J} n_j = M - 1\}$$

for closed networks. The state-dependent traffic equations (3.24) now reduce to the traffic equations (3.3), that is,

$$H(\mathbf{n}) = \prod_{j=1}^{J} \lambda_j^{n_j}, \quad \mathbf{n} \in S_{\mathbf{c}},$$

satisfies the state-dependent traffic equations (3.24). Assume that

$$G_{\mathbf{c}}^{-1} = \sum_{\mathbf{n}\in S_{\mathbf{c}}} \phi(\mathbf{n}) \prod_{j=1}^{J} \rho_j^{n_j} < \infty,$$

and that $\{N(t)\}$ is irreducible. Then $\{N(t)\}$ has unique equilibrium distribution

$$\pi(\mathbf{n}) = G_{\mathbf{c}}\phi(\mathbf{n}) \prod_{j=1}^{J} \rho_j^{n_j}, \quad \mathbf{n} \in S_{\mathbf{c}}.$$

□

**Example 3.4.7 (Recirculate-protocol)** Under the recirculate-protocol, if a queue is saturated then at all other queues departing customers are recirculated into their originating station as newly arriving customers to undergo a new service:

**Recirculate-protocol:** If queue $i$ in a Kelly-Whittle network with finite capacity constraints becomes saturated ($n_i = c_i$) then at all *all* other queues $j = 1,\dots,J$, $j \neq i$, departing customers are recirculated into their originating station as newly arriving customers to undergo a new service, and arriving customers are discarded.[4]

The state spaces coincide with those under the stop-protocol. Under the recirculate-protocol, the transition rates (3.14) have the form (3.23) with

$$\begin{aligned}
\theta_i(\mathbf{m}) &= 1, \quad i = 0,\dots,J,\ \mathbf{m} \in S_{\mathbf{c}}^b,\\
f(\mathbf{m}) &= \mathbb{1}(m_j < c_j,\ j = 1,\dots,J), \quad \mathbf{m} \in S_{\mathbf{c}}^b,\\
b_{ij}(\mathbf{m}) &= p_{ij} f(\mathbf{m}), \quad i \neq j,\ i,j = 0,\dots,J,\ \mathbf{m} \in S_{\mathbf{c}}^b,\\
b_{ii}(\mathbf{m}) &= \sum_{j\neq i} p_{ij}(1 - f(\mathbf{m})), \quad i = 1,\dots,J,\ \mathbf{m} \in S_{\mathbf{c}}^b,
\end{aligned}$$

where $b_{ii}(\mathbf{m})$ represents a dummy-transition from state $\mathbf{m}+\mathbf{e}_i$ to state $\mathbf{m}+\mathbf{e}_i$. The equilibrium distribution coincides with the equilibrium distribution under the stop-protocol. □

**Example 3.4.8 (Jump-over-protocol)** The stop- and recirculate-protocols above change the behaviour of all stations of the network when a single station is saturated. It may be more natural to only modify the behaviour of the saturated queue. For example, consider the tandem of 2 single server queues, where queue 2 has capacity restriction $c_2$. Queue 1 has Poisson arrivals at rate $\mu_0$. Suppose queue 1 continues working no matter what the state of queue 2, but that customers arriving

[4] The arrival process has negative-exponential interarrival times, so that we may also say that arrivals are recirculated to station 0.

to queue 2 when $n_2 = c_2$ are discarded just like in the $M|M|1|c_2$ queue. Burke's theorem 2.5.1 implies that in equilibrium the number of customers $N_1, N_2$ in the queues are independent random variables and therefore that the equilibrium distribution is the product of the marginal queue length distributions. The jump-over-protocol generalises this argument to Kelly-Whittle networks.

The jump-over-protocol modifies only the behaviour of the customer that is blocked. Under the jump-over-protocol, a customer that is blocked to enter a station jumps over this station and attempts to enter the next station as if it was served at the saturated queue.

**Jump-over-blocking** If queue $i$ in a Kelly-Whittle network with finite capacity constraints becomes saturated ($n_i = c_i$) then a customer arriving to queue $i$ will immediately select a new station $j$ with probability $p_{ij}$, $j = 0,\ldots,J$, $i = 1,\ldots,J$.

We may generalise this protocol to allow customers in all states to jump over a station:

**Generalised jump-over-blocking** A customer arriving at station $i$ when $n_i$ customers are present will be accepted with probability $a_i(n_i)$, and will jump over the station with probability $1 - a_i(n_i)$. A rejected customer selects a new station $j$ with probability $p_{ij}$, $j = 0,\ldots,J$, $i = 1,\ldots,J$.

Let $c_j = \inf\{k : a_j(k) = 0,\ k = 0,1,2,\ldots\}$, $j = 1,\ldots,J$. The state space is

$$S_{jo,\mathbf{c},o} = \{\mathbf{n} \in \mathbb{N}_0^J : 0 \le n_j \le c_j,\ i = 1,\ldots,J\}$$

for the open network, and for the closed network containing $M$ customers

$$S_{jo,\mathbf{c},M} = \{\mathbf{n} \in \mathbb{N}_0^J : 0 \le n_j \le c_j,\ j = 1,\ldots,J,\ \sum_{j=1}^{J} n_j = M\}.$$

Define the matrices $P(\mathbf{m}) = (p_{ij}a_j(m_j),\ i,j = 0,\ldots,J)$, and $P_*(\mathbf{m}) = (p_{ij}(1 - a_j(m_j)),\ i,j = 0,\ldots,J)$. The transition rates of the Kelly-Whittle network under the generalised-jump-over protocol have the form (3.23), for $\mathbf{m} \in S^b_{jo,\mathbf{c}}$, with

$$\theta_i(\mathbf{m}) = 1, \quad i = 0,\ldots,J,$$
$$b_{ij}(\mathbf{m}) = p_{ij}a_j(m_j) + (P_*(\mathbf{m})P(\mathbf{m}))_{ij} + (P_*^2(\mathbf{m})P(\mathbf{m}))_{ij} + \cdots = \sum_{k=0}^{\infty} (P_*^k(\mathbf{m})P(\mathbf{m}))_{ij},$$

where

$$S^b_{jo,\mathbf{c}} = S^b_{jo,\mathbf{c},o} = S_{jo,\mathbf{c},o}$$

for the open network and

$$S^b_{jo,\mathbf{c}} = S^b_{jo,\mathbf{c},M} = S^b_{jo,\mathbf{c},M-1}$$

for the closed network. For $a_j(m_j) = \mathbb{1}(m_j \le c_j)$

$$H(\mathbf{n}) = \prod_{j=1}^{J} \lambda_j^{n_j}, \quad \mathbf{n} \in S_{jo,\mathbf{c}},$$

satisfies the state-dependent traffic equations (3.24). Assume that

$$G_{jo,\mathbf{c}}^{-1} = \sum_{\mathbf{n} \in S_{jo,\mathbf{c}}} \phi(\mathbf{n}) \prod_{j=1}^{J} \rho_j^{n_j} < \infty,$$

and that $\{N(t)\}$ is irreducible. Then $\{N(t)\}$ has unique equilibrium distribution

$$\pi(\mathbf{n}) = G_{jo,\mathbf{c}} \phi(\mathbf{n}) \prod_{j=1}^{J} \rho_j^{n_j}, \quad \mathbf{n} \in S_{jo,\mathbf{c}}.$$

□

## 3.5 Literature

Staat al in Remark 3.1.9: Open networks of $M|M|1$ queues are often called Jackson networks, referring to J.R. Jackson who first obtained their equilibrium distribution [**?**] and closed networks of $M|M|1$ queues are called Gordon-Newell networks, referring to W.J. Gordon and G.F. Newell who obtained their equilibrium distribution [**?**]. In the exposition in this chapter we have introduced queue 0 to represent the outside of the network of queues which allows a unified analysis of open and closed networks.

hier iets over HTPvD eerste met psi

HvD over blocking etc

Chandy Martin psi=phi

follow Serfozo??

Noem ook hoodstukken uit handbook

# Chapter 4
# Kelly's lemma and networks with fixed routes

## 4.1 The time-reversed process and Kelly's Lemma

For the stationary Markov chain $\{N(t)\}$ with state space $S$ and transition rates $q(\mathbf{n},\mathbf{n}')$, $\mathbf{n},\mathbf{n}' \in S$, the time-reversed process $\{N^r(t)\} = \{N(\tau - t)\}$, for some $\tau \in \mathbb{R}$, is a stationary Markov chain with state space $S$. From $\mathbb{P}(N(t+h) = \mathbf{n}', N(t) = \mathbf{n})$, the joint distribution of $N(t)$ and $N(t+h)$ for some $t \in \mathbb{R}$, $h > 0$, $\mathbf{n},\mathbf{n}' \in S$, we obtain by conditioning on $N(t+h)$ in the left-hand side and on $N(t)$ in the right-hand side, for $\mathbf{n},\mathbf{n}' \in S$, $t \in \mathbb{R}$, $h > 0$,

$$\mathbb{P}(N(t) = \mathbf{n} | N(t+h) = \mathbf{n}') = \frac{\mathbb{P}(N(t) = \mathbf{n})}{\mathbb{P}(N(t+h) = \mathbf{n}')} \mathbb{P}(N(t+h) = \mathbf{n}' | N(t) = \mathbf{n}). \quad (4.1)$$

Dividing by $h$ and taking the limit $h \downarrow 0$ we obtain the following result.

**Theorem 4.1.1** *Let $\{N(t),\ t \in T\}$, $T = \mathbb{R}$, be a stationary Markov chain with transition rates $q(\mathbf{n},\mathbf{n}')$, $\mathbf{n},\mathbf{n}' \in S$, and equilibrium distribution $\pi(\mathbf{n})$, $\mathbf{n} \in S$. The time-reversed process $\{N(\tau - t),\ t \in T\}$ is a conservative, regular, irreducible continuous-time stationary Markov chain with transition rates $q^r(\mathbf{n},\mathbf{n}')$, $\mathbf{n},\mathbf{n}' \in S$ given by*

$$q^r(\mathbf{n},\mathbf{n}') = \frac{\pi(\mathbf{n}')}{\pi(\mathbf{n})} q(\mathbf{n}',\mathbf{n}),$$

*and the same equilibrium distribution $\pi(\mathbf{n})$, $\mathbf{n} \in S$.*

**Remark 4.1.2 (Assumptions of Theorem 4.1.1)** Observe that it is essential that $\{N(t),\ t \in T\}$ is stationary. To see this, consider (4.1). If $\{N(t),\ t \in T\}$ is not stationary, then it must be that $\mathbb{P}(N(t) = \mathbf{n})$ or $\mathbb{P}(N(t+h) = \mathbf{n}')$ depend on $t$ for some $\mathbf{n},\mathbf{n}' \in S$, so that $\mathbb{P}(N(t) = \mathbf{n} | N(t+h) = \mathbf{n}')$ will depend on $t$ and therefore $\{N(\tau - t),\ t \in T\}$ will not be time-homogeneous. □

An important consequence of Theorem 4.1.1 is the following theorem that will be the basis for the analysis in this chapter.

**Theorem 4.1.3 (Kelly's lemma)** *Let $\{N(t),\, t \in T\}$, $T = \mathbb{R}$, be a stationary Markov chain with transition rates $q(\mathbf{n},\mathbf{n}')$, $\mathbf{n},\mathbf{n}' \in S$. If we can find a collection of numbers $q^r(\mathbf{n},\mathbf{n}')$, $\mathbf{n},\mathbf{n}' \in S$, such that*

$$\sum_{\mathbf{n}' \neq \mathbf{n}} q(\mathbf{n},\mathbf{n}') = \sum_{\mathbf{n}' \neq \mathbf{n}} q^r(\mathbf{n},\mathbf{n}'), \quad \mathbf{n} \in S, \tag{4.2}$$

*and a distribution $\pi = (\pi(\mathbf{n}),\, \mathbf{n} \in S)$ such that*

$$\pi(\mathbf{n}) q^r(\mathbf{n},\mathbf{n}') = \pi(\mathbf{n}') q(\mathbf{n}',\mathbf{n}), \quad \mathbf{n},\mathbf{n}' \in S, \tag{4.3}$$

*then $q^r(\mathbf{n},\mathbf{n}')$, $\mathbf{n},\mathbf{n}' \in S$, are the transition rates of the time-reversed Markov chain $\{N(\tau - t),\, t \in T\}$ and $\pi(\mathbf{n})$, $\mathbf{n} \in S$, is the equilibrium distribution of both Markov chains.*

**Proof.** From (4.3) and (4.2) it follows that

$$\begin{aligned} \sum_{\mathbf{n}' \in S} \pi(\mathbf{n}') q(\mathbf{n}',\mathbf{n}) &= \sum_{\mathbf{n}' \in S} \pi(\mathbf{n}) q^r(\mathbf{n},\mathbf{n}') \\ &= \pi(\mathbf{n}) \sum_{\mathbf{n}' \in S} q(\mathbf{n},\mathbf{n}'). \end{aligned}$$

Theorem 4.1.1 concludes the proof. □

**Example 4.1.4 (The $M|M|1$ queue)** The $M|M|1$ queue has Poisson arrival rate $\lambda$ and service rate $\mu$. For $\lambda < \mu$ the departure rate must be $\lambda$, recall Remark 2.4.4. A natural guess for the arrival rate of the time-reversed process is $q^r(\mathbf{n},\mathbf{n}+1) = \lambda$, $\mathbf{n} \in \mathbb{N}_0$. A further guess could then be $q^r(\mathbf{n},\mathbf{n}-1) = \mu$, $\mathbf{n} \in \mathbb{N}$, and an educated guess for the equilibrium distribution is $\pi(\mathbf{n}) = (1-\rho)\rho^{\mathbf{n}}$, with $\rho = \lambda/\mu$. Clearly, Kelly's lemma 4.1.3 is satisfied. Thus, the time-reversed process is an $M|M|1$ queue with Poisson arrivals at rate $\lambda$ and service rate $\mu$. This could also be concluded observing that the Markov chain $\{N(t)\}$ recording the number of customer in the $M|M|1$ queue is a reversible Markov chain. □

**Example 4.1.5 (Kelly-Whittle networks)** Consider a Kelly-Whittle network $\{N(t)\}$ at state space $S \subseteq \mathbb{N}_0^J$ with transition rates, for $\mathbf{n}' \neq \mathbf{n}$,

$$q(\mathbf{n},\mathbf{n}') = \begin{cases} \dfrac{\psi(\mathbf{n}-\mathbf{e}_i)}{\phi(\mathbf{n})} \mu_i p_{ij}, & \text{if } \mathbf{n}' = \mathbf{n} - \mathbf{e}_i + \mathbf{e}_j,\ i,j = 0,\dots,J, \\ 0, & \text{otherwise,} \end{cases} \tag{4.4}$$

where $\psi : \mathbb{N}_0^J \to [0,\infty)$ and $\phi : \mathbb{N}_0^J \to (0,\infty)$. As argued in Remark 3.2.6 the service process is a reversible process. The proof of Theorems 3.1.4 and 3.1.5 argues that the routing process is a Markov chain with equilibrium distribution $\lambda_j$, $j = 1,\dots,J$, up to normalisation. According to Kelly's lemma 4.1.3, the time-reversed routing process is the Markov chain with transition probabilities

$$p^r_{ij} := \frac{\lambda_j}{\lambda_i} p_{ji}, \quad i,j = 0,\dots,J.$$

A natural guess for the transition rates of the time-reversed Kelly-Whittle network $\{N^r(t)\}$ is, for $\mathbf{n} \neq \mathbf{n}'$,

$$q^r(\mathbf{n},\mathbf{n}') = \begin{cases} \dfrac{\psi(\mathbf{n}-\mathbf{e}_i)}{\phi(\mathbf{n})}\mu_i p^r_{ij}, \text{ if } \mathbf{n}' = \mathbf{n}-\mathbf{e}_i+\mathbf{e}_j,\ i,j=0,\dots,J, \\ 0, \qquad\qquad\qquad \text{otherwise.} \end{cases} \tag{4.5}$$

Observe that

$$\begin{aligned} \sum_{\mathbf{n}'\neq\mathbf{n}} q(\mathbf{n},\mathbf{n}') &= \sum_{i,j=0}^{J} \frac{\psi(\mathbf{n}-\mathbf{e}_i)}{\phi(\mathbf{n})}\mu_i p_{ij} = \sum_{i=0}^{J} \frac{\psi(\mathbf{n}-\mathbf{e}_i)}{\phi(\mathbf{n})}\mu_i, \\ \sum_{\mathbf{n}'\neq\mathbf{n}} q^r(\mathbf{n},\mathbf{n}') &= \sum_{i,j=0}^{J} \frac{\psi(\mathbf{n}-\mathbf{e}_i)}{\phi(\mathbf{n})}\mu_i p^r_{ij} = \sum_{i,j=0}^{J} \frac{\psi(\mathbf{n}-\mathbf{e}_i)}{\phi(\mathbf{n})}\mu_i \frac{\lambda_j}{\lambda_i} p_{ji} = \sum_{i=0}^{J} \frac{\psi(\mathbf{n}-\mathbf{e}_i)}{\phi(\mathbf{n})}\mu_i, \end{aligned}$$

where the last equality is due to the traffic equations (3.3).

An educated guess for the equilibrium distribution is $\pi(\mathbf{n}) = G_{KW}\phi(\mathbf{n})\prod_{j=1}^{J}\rho_j^{n_j}$, $\mathbf{n} \in S$, that clearly satisfies (4.3):

$$\begin{aligned} \pi(\mathbf{n})q^r(\mathbf{n},\mathbf{n}') &= G_{KW}\phi(\mathbf{n})\prod_{k=1}^{J}\rho_k^{n_k}\frac{\psi(\mathbf{n}-\mathbf{e}_i)}{\phi(\mathbf{n})}\mu_i p^r_{ij} \\ &= G_{KW}\phi(\mathbf{n})\prod_{k=1}^{J}\rho_k^{n_k}\frac{\psi(\mathbf{n}-\mathbf{e}_i)}{\phi(\mathbf{n})}\mu_i \frac{\lambda_j}{\lambda_i} p_{ji} \\ &= G_{KW}\phi(\mathbf{n}-\mathbf{e}_i+\mathbf{e}_j)\prod_{k=1}^{J}\rho_k^{n_k-\delta_{ki}+\delta_{kj}}\frac{\psi(\mathbf{n}-\mathbf{e}_i)}{\phi(\mathbf{n}-\mathbf{e}_i+\mathbf{e}_j)}\mu_j p_{ji} \\ &= \pi(\mathbf{n}-\mathbf{e}_i+\mathbf{e}_j)q(\mathbf{n}-\mathbf{e}_i+\mathbf{e}_j,\mathbf{n}). \end{aligned}$$

Kelly's lemma 4.1.3 implies that $q^r(\mathbf{n},\mathbf{n}')$ given in (4.5) are the transition rates of the time-reversed Kelly-Whittle network and that $\pi(\mathbf{n}) = G_{KW}\phi(\mathbf{n})\prod_{j=1}^{J}\rho_j^{n_j}$, $\mathbf{n} \in S$, is the equilibrium distribution of both the Kelly-Whittle network and the time-reversed Kelly-Whittle network provided $\sum_{\mathbf{n}\in S}\phi(\mathbf{n})\prod_{j=1}^{J}\rho_j^{n_j} < \infty$. □

## 4.2 Queue disciplines

A complete description of a queue in a network with customer classes $c$, $c = 1,\dots,C$, requires a description of the position of the customers in the queue as well as rules for the position of new customers and the attention of the server towards different customers.

**Definition 4.2.1 (Queue with customer types: $(\kappa, \gamma, \delta)$-protocol)** *Let customers in the queue be ordered: if the queue contains n customers then these customers are in positions* $1, \ldots, n$, $n \in \mathbb{N}$. *Assume the queue operates as follows:*

- *a customer of class c requires a negative-exponentially distributed amount of service with rate* $\mu(c)$*;*
- *if* $n > 0$ *customers are present service is provided at rate* $\kappa(n) > 0$*;*
- *a fraction* $\gamma(\ell, n)$ *of the service effort is directed to the customer in position* $\ell$, $\ell = 1, \ldots, n$*; if the customer in position* $\ell$ *completes service and leaves the queue customers in positions* $\ell+1, \ell+2, \ldots, n$ *move to positions* $\ell, \ell+1, \ldots, n-1$, *respectively;*
- *a customer that arrives moves into position* $\ell$ *with probability* $\delta(\ell, n+1)$*; customers previously in positions* $\ell, \ell+1, \ldots, n$ *move to positions* $\ell+1, \ell+2, \ldots, n+1$, *respectively,*

*where, for* $n \in \mathbb{N}$*:*

$$\sum_{\ell=1}^{n} \gamma(\ell, n) = 1, \quad \sum_{\ell=1}^{n} \delta(\ell, n) = 1.$$

**Example 4.2.2 (Queue disciplines)** The $(\kappa, \gamma, \delta)$-protocol is a flexible model to describe various queueing disciplines. A queue operates under the

**First-In-First-Out protocol (FIFO)** if customers that arrive join the tail of the queue and service is provided by a single server to the customer at the front of the queue,

**Last-In-First-Out-Preemptive-Resume protocol (LIFO-PR)** if customers that arrive join the tail of the queue and service is provided by a single server to the customer at the tail of the queue, *and* upon arrival of a new customer at the tail of the queue service of the customer in service is interrupted and the new customer at the tail of the queue is served,

**Processor-Sharing protocol (PS)** if a single server equally shares its attention to all customers present in the queue and (for symmetry) a customer that arrives is placed at a random position in the queue,

**Infinite-server protocol (INF)** if each customer receives its own server and (for symmetry) a customer that arrives is placed at a random position in the queue.

Table 4.1 provides the functions $\kappa, \gamma, \delta$ for these protocols. We have included the

**Table 4.1** The functions $\kappa, \gamma, \delta$, $\ell = 1, \ldots, n$, $n \in \mathbb{N}$:

| | $\kappa(n)$ | $\gamma(\ell,n)$ | $\delta(\ell,n)$ |
|---|---|---|---|
| FIFO | $\mathbb{1}(n>0)$ | $\mathbb{1}(\ell=1)$ | $\mathbb{1}(\ell=n)$ |
| LIFO-PR | $\mathbb{1}(n>0)$ | $\mathbb{1}(\ell=n)$ | $\mathbb{1}(\ell=n)$ |
| PS | $\mathbb{1}(n>0)$ | $1/n$ | $1/n$ |
| INF | $n$ | $1/n$ | $1/n$ |

indicator $\mathbb{1}(n > 0)$ instead of 1 for the service rate $\kappa$ to emphasise that the service rate is constant whenever customers are present. □

**Example 4.2.3 (Multi-class LIFO-PR queue: equilibrium distribution)** Consider the multi-class LIFO-PR queue as introduced above. Let customers of class $c$ arrive according to a Poisson process with rate $\lambda(c)$, $c = 1,\dots,C$, and let $\rho(c) = \lambda(c)/\mu(c)$, $c = 1,\dots,C$. Assume that $\rho := \sum_{c=1}^{C}\rho(c) < 1$, where $\rho$ the mean amount of work arriving to the queue per unit time. If $n > 0$ customers are present, let $\mathbf{c} = (c(1),\dots,c(n))$, $c(i) \in \{1,\dots,C\}$, record the class of the customers in position $i$, $i = 1,\dots,n$. Let $\{N(t)\}$ record the state of the Markov chain at state space

$$S = \{\mathbf{c} : \mathbf{c} = (c(1),\dots,c(n)),\ c(i) \in \{1,\dots,C\},\ i = 1,\dots,n,\ n \in \mathbb{N}_0\}. \tag{4.6}$$

The transition rates are, for $\mathbf{c} = (c(1),\dots,c(n))$, $\mathbf{c}' \neq \mathbf{c}$, $\mathbf{c},\mathbf{c}' \in S$,

$$q(\mathbf{c},\mathbf{c}') = \begin{cases} \lambda(c), & \text{if } \mathbf{c}' = (c(1),\dots,c(n),c),\ c \in \{1,\dots,C\}, \\ \mu(c(n)), & \text{if } \mathbf{c}' = (c(1),\dots,c(n-1)). \end{cases}$$

As customers are placed at the tail of the queue and served from the tail of the queue, a natural *guess* for the time-reversed multi-class LIFO-PR queue is the multi-class LIFO-PR queue with the same rates:

$$q^r(\mathbf{c},\mathbf{c}') = \begin{cases} \lambda(c), & \text{if } \mathbf{c}' = (c(1),\dots,c(n),c),\ c \in \{1,\dots,C\}, \\ \mu(c(n)), & \text{if } \mathbf{c}' = ((c(1),\dots,c(n-1)). \end{cases}$$

We will now use Kelly's lemma 4.1.3 to show that

$$\pi(\mathbf{c}) = (1-\rho)\prod_{i=1}^{n}\rho(c(i)), \quad \mathbf{c} = (c(1),\dots,c(n)) \in S, \tag{4.7}$$

is the equilibrium distribution. Clearly, (4.2) is satisfied:

$$\begin{aligned} \sum_{\mathbf{c}'\neq\mathbf{c}} q(\mathbf{c},\mathbf{c}') &= \sum_{c=1}^{C}\lambda(c) + \mu(c(n)), \\ \sum_{\mathbf{c}'\neq\mathbf{c}} q^r(\mathbf{c},\mathbf{c}') &= \sum_{c=1}^{C}\lambda(c) + \mu(c(n)). \end{aligned}$$

For $\mathbf{c} = (c(1),\dots,c(n))$ it is sufficient to check (4.3) for arrivals and departures. To this end, let $\mathbf{c}' = (c(1),\dots,c(n),c)$, $\mathbf{c}'' = (c(1),\dots,c(n-1))$, then

$$\begin{aligned} \pi(\mathbf{c})q^r(\mathbf{c},\mathbf{c}') = \pi(\mathbf{c}')q(\mathbf{c}',\mathbf{c}) &\Leftrightarrow \lambda(c) = \rho(c)\mu(c), \\ \pi(\mathbf{c})q^r(\mathbf{c},\mathbf{c}'') = \pi(\mathbf{c}'')q(\mathbf{c}'',\mathbf{c}) &\Leftrightarrow \rho(c(n))\mu(c(n)) = \lambda(c(n)), \end{aligned}$$

that are trivially satisfied. Kelly's lemma 4.1.3 implies that (4.7) is indeed the equilibrium distribution. □

**Example 4.2.4 (Multi-class FIFO queue: equilibrium distribution)** Now consider the multi-class FIFO queue. Let customers of class $c$, $c = 1,\dots,C$, arrive with rate $\lambda(c)$, $c = 1,\dots,C$, and let $\rho(c) = \lambda(c)/\mu(c)$, $c = 1,\dots,C$. Assume that $\rho := \sum_{c=1}^{C}\rho(c) < 1$. Let $\{N(t)\}$ record the state of the Markov chain at state space (4.6) with transition rates, for $\mathbf{c} = (c(1),\dots,c(n))$, $\mathbf{c}' \neq \mathbf{c}$, $\mathbf{c},\mathbf{c}' \in S$,

$$q(\mathbf{c},\mathbf{c}') = \begin{cases} \lambda(c), & \text{if } \mathbf{c}' = (c(1),\dots,c(n),c),\ c \in \{1,\dots,C\}, \\ \mu(c(1)), & \text{if } \mathbf{c}' = (c(2),\dots,c(n)). \end{cases}$$

Note that these rates differ from those for multi-class LIFO-PR via the departure rate: customers now depart from position 1. As customers are placed at the tail of the queue and served from the head of the queue, a natural *guess* for the time-reversed multi-class FIFO queue is the queue in which customers are placed at the head of the queue and are served from the tail of the queue:

$$q^r(\mathbf{c},\mathbf{c}') = \begin{cases} \lambda(c), & \text{if } \mathbf{c}' = (c,c(1),\dots,c(n)),\ c \in \{1,\dots,C\}, \\ \mu(c(n)), & \text{if } \mathbf{c}' = (c(1),\dots,c(n-1)). \end{cases}$$

Observe that these are the transition rates for a multi-class FIFO queue with reversed numbering of customer positions: customers are served in position $n$ and new customers are placed in postion 1. Now consider condition (4.2) in Kelly's lemma, for $\mathbf{c} = (c(1),\dots,c(n))$,

$$\sum_{\mathbf{c}'\neq\mathbf{c}} q(\mathbf{c},\mathbf{c}') = \sum_{c=1}^{C}\lambda(c) + \mu(c(1)),$$
$$\sum_{\mathbf{c}'\neq\mathbf{c}} q^r(\mathbf{c},\mathbf{c}') = \sum_{c=1}^{C}\lambda(c) + \mu(c(n)).$$

For these terms to be equal it must be that the service rates are identical for all customer classes: $\mu(c) = \mu$, $c = 1,\dots,C$. In that case, following the reasoning for the multi-class LIFO-PR queue we readily obtain that (4.7) is the equilibrium distribution of the multi-class FIFO queue. For $\mathbf{c} = (c(1),\dots,c(n))$ it is sufficient to check (4.3) for arrivals and departures. To this end, let $\mathbf{c}' = (c,c(1),\dots,c(n))$, $\mathbf{c}'' = (c(1),\dots,c(n-1))$, then

$$\pi(\mathbf{c})q^r(\mathbf{c},\mathbf{c}') = \pi(\mathbf{c}')q(\mathbf{c}',\mathbf{c}) \Leftrightarrow \lambda(c) = \rho(c)\mu(c),$$
$$\pi(\mathbf{c})q^r(\mathbf{c},\mathbf{c}'') = \pi(\mathbf{c}'')q(\mathbf{c}'',\mathbf{c}) \Leftrightarrow \rho(c(n))\mu(c(n)) = \lambda(c(n)),$$

that are trivially satisfied.[1] Kelly's lemma 4.1.3 implies that (4.7) with $\mu(c) = \mu$, $c = 1,\dots,C$, is indeed the equilibrium distribution.

Alternatively, we might *guess* that the time-reversed multi-class FIFO queue coincides with the original multi-class FIFO queue, i.e., serving customers at the head of the queue (position 1) and placing new customers at the tail of the queue (po-

[1] Note that these equations do not require that $\mu(c) = \mu$, $c = 1,\dots,C$.

sition $n+1$). However, in that case condition (4.3) of Kelly's lemma cannot be satisfied. To see this, consider $\mathbf{c} = (c(1),\ldots,c(n))$ and $\mathbf{c}' = (c(1),\ldots,c(n),c)$, then $q^r(\mathbf{c},\mathbf{c}') = \lambda(c)$, whereas $q(\mathbf{c}',\mathbf{c}) = 0$. □

By analogy with the examples above we may investigate the equilibrium distribution for a general queue operating under the $(\kappa,\gamma,\delta)$-protocol. We will consider two cases: equal service rates for all customer classes and *symmetric queues*.

**Theorem 4.2.5 (Equal service rates)** *Let $\{N(t)\}$ record the state of a queue operating under the $(\kappa,\gamma,\delta)$-protocol to which customers of class c arrive according to a Poisson process with rate $\lambda(c)$, $c = 1,\ldots,C$. Assume that $\mu(c) = \mu$, $c = 1,\ldots,C$, and that*

$$G = \left[\sum_{n=0}^{\infty}\prod_{\ell=1}^{n}\frac{\rho}{\kappa(\ell)}\right]^{-1} < \infty,$$

*where $\rho = \lambda/\mu$ with $\lambda = \sum_{c=1}^{C}\lambda(c)$ the arrival rate to the queue. Then $\{N(t)\}$ has unique equilibrium distribution*

$$\pi(\mathbf{c}) = G\prod_{\ell=1}^{n}\frac{\rho(c(\ell))}{\kappa(\ell)}, \quad \mathbf{c} = (c(1),\ldots,c(n)). \tag{4.8}$$

**Proof.** The transition rates of $\{N(t)\}$ are, for $\mathbf{c} = (c(1),\ldots,c(n))$, $\mathbf{c}' \neq \mathbf{c}$, $\mathbf{c},\mathbf{c}' \in S$, $c \in \{1,\ldots,C\}$,

$$q(\mathbf{c},\mathbf{c}') = \begin{cases} \lambda(c)\delta(\ell,n+1), & \text{if } \mathbf{c}' = (c(1),\ldots,c(\ell),c,c(\ell+1),\ldots,c(n)),\\ \mu(c(\ell))\kappa(n)\gamma(\ell,n), & \text{if } \mathbf{c}' = c(1),\ldots,c(\ell-1),c(\ell+1),\ldots,c(n)). \end{cases} \tag{4.9}$$

Following the arguments in Example 4.2.4, for $c \in \{1,\ldots,C\}$, let

$$q^r(\mathbf{c},\mathbf{c}') = \begin{cases} \lambda(c)\gamma(\ell,n+1), & \text{if } \mathbf{c}' = (c(1),\ldots,c(\ell),c,c(\ell+1),\ldots,c(n)),\\ \mu(c(\ell))\kappa(n)\delta(\ell,n), & \text{if } \mathbf{c}' = c(1),\ldots,c(\ell-1),c(\ell+1),\ldots,c(n)), \end{cases} \tag{4.10}$$

which are the transition rates of a queue under the $(\kappa,\delta,\gamma)$-protocol with the role of $\gamma$ and $\delta$ reversed. Under the assumption $\mu(c) = \mu$, $c = 1,\ldots,C$, for the distribution $\pi$ in (4.8) both conditions of Kelly's lemma 4.1.3 are satisfied. □

**Definition 4.2.6 (Symmetric queue)** *A queue operating under the $(\kappa,\gamma,\delta)$-protocol is called a symmetric queue if*

$$\gamma(\ell,n) = \delta(\ell,n), \quad \ell = 1,\ldots,n,\ n \in \mathbb{N}.$$

**Theorem 4.2.7 (Symmetric queue)** *Let $\{N(t)\}$ record the state of a symmetric queue to which customers of class c arrive according to a Poisson process with rate $\lambda(c)$, $c = 1,\ldots,C$. Assume that*

$$G = \left[\sum_{n=0}^{\infty}\prod_{\ell=1}^{n}\frac{\rho}{\kappa(\ell)}\right]^{-1} < \infty,$$

*where $\rho = \sum_{c=1}^{C} \lambda(c)/\mu(c)$ the mean amount of work arriving to the queue per unit time. Then $\{N(t)\}$ has unique equilibrium distribution*

$$\pi(\mathbf{c}) = G \prod_{\ell=1}^{n} \frac{\rho(c(\ell))}{\kappa(\ell)}, \quad \mathbf{c} = (c(1), \ldots, c(n)). \tag{4.11}$$

**Proof.** The transition rates of $\{N(t)\}$ are, for $\mathbf{c} = (c(1), \ldots, c(n))$, $\mathbf{c}' \neq \mathbf{c}$, $\mathbf{c}, \mathbf{c}' \in S$, $c \in \{1, \ldots, C\}$, given in (4.9). Following the arguments in Example 4.2.3, let $q^r$ be given in (4.10). Under the assumption $\gamma(\ell, n) = \delta(\ell, n)$, for the distribution $\pi$ in (4.11) both conditions of Kelly's lemma 4.1.3 are satisfied. □

**Remark 4.2.8 (Conditions Kelly's lemma)** Observe that the assumptions $\mu(c) = \mu$, $c = 1, \ldots, C$, in Theorem 4.2.5 and $\gamma(\ell, n) = \delta(\ell, n)$, $\ell = 1, \ldots, n$, $n \in \mathbb{N}$, in Theorem 4.2.7 are not required for condition (4.3) of Kelly's lemma, and are sufficient for condition (4.2) in Kelly's lemma, for $\mathbf{c} = (c(1), \ldots, c(n))$:

$$\begin{aligned}
\sum_{\mathbf{c}' \neq \mathbf{c}} q(\mathbf{c}, \mathbf{c}') &= \sum_{c=1}^{C} \sum_{\ell=1}^{n+1} \lambda(c) \delta(\ell, n+1) + \sum_{\ell=1}^{n} \mu(c(\ell)) \kappa(n) \gamma(\ell, n) \\
&= \sum_{c=1}^{C} \lambda(c) + \kappa(n) \sum_{\ell=1}^{n} \mu(c(\ell)) \gamma(\ell, n), \\
\sum_{\mathbf{c}' \neq \mathbf{c}} q^r(\mathbf{c}, \mathbf{c}') &= \sum_{c=1}^{C} \sum_{\ell=1}^{n+1} \lambda(c) \gamma(\ell, n+1) + \sum_{\ell=1}^{n} \mu(c(\ell)) \kappa(n) \delta(\ell, n) \\
&= \sum_{c=1}^{C} \lambda(c) + \kappa(n) \sum_{\ell=1}^{n} \mu(c(\ell)) \delta(\ell, n).
\end{aligned}$$

For condition (4.2) to be satisfied it must be that, for all $\mathbf{c} = (c(1), \ldots, c(n)) \in S$,

$$\sum_{\ell=1}^{n} \mu(c(\ell)) \gamma(\ell, n) = \sum_{\ell=1}^{n} \mu(c(\ell)) \delta(\ell, n),$$

for which $\mu(c) = \mu$, $c = 1, \ldots, C$, or $\gamma(\ell, n) = \delta(\ell, n)$, $\ell = 1, \ldots, n$, $n \in \mathbb{N}$, are sufficient conditions. □

**Remark 4.2.9 (Aggregation of customer numbers)** Consider the multi-class FIFO, LIFO-PR or PS queue. From Theorems 4.2.5, 4.2.7, the equilibrium distribution is

$$\pi(\mathbf{c}) = (1 - \rho) \rho^n \prod_{\ell=1}^{n} \frac{\rho(c(\ell))}{\rho}, \quad \mathbf{c} = (c(1), \ldots, c(n)),$$

with $\rho(c) = \lambda(c)/\mu(c)$, and $\rho = \sum_{c=1}^{C} \lambda(c)/\mu(c)$ the mean amount of work arriving to the queue per unit time. If $\mu(c) = \mu$, $c = 1, \ldots, C$, then $\rho(c)/\rho = \lambda(c)/\lambda$ is the probability that a customer in the queue is of type $c$, $c = 1, \ldots, C$.

The equilibrium distribution of the total number of customers $n(c)$ of class $c$, $c = 1, \ldots, C$, may be obtained by summation over all states $\mathbf{c}$ such that $\sum_{\ell=1}^{n} \mathbb{1}(c(\ell) =$

$c\} = n(c)$:

$$\pi(n(1),\dots,n(C)) = (1-\rho)\rho^n \frac{n!}{n(1)!\cdots n(C)!}\prod_{c=1}^{C}\left(\frac{\rho(c)}{\rho}\right)^{n(c)}.$$

The equilibrium distribution of the total number of customers is

$$\pi(n) = (1-\rho)\rho^n, \quad n \in \mathbb{N}_0.$$

□

## 4.3 Networks with customer types and fixed routes

Let customers of types $u = 1,\dots,U$ arrive to a network of $J$ queues according to a Poisson process with rate $\mu_0(u)$, $u = 1,\dots,U$. A customer's type uniquely determines his route through the network along the sequence of queues

$$r(u,1), r(u,2),\dots, r(u,L(u)),$$

i.e., the route of a customer of type $u$ has $L(u)$ stages $r(u,s)$, $s = 1,\dots,L(u)$, and starts in queue $r(u,1)$ at stage 1, passes through $L(u)$ queues and leaves the network in stage $L(u)$ from queue $L(u)$.[2][3] A customer may visit the same queue at multiple stages. The next queue on its route is then determined by the stage of its current visit to that queue. In addition, a customer may pass through a queue a fixed number of times.

Let queue $j$ operate according to the $(\kappa_j,\gamma_j,\delta_j)$-protocol, $j = 1,\dots,J$, recall Definition 4.2.1. Let $c_j(\ell) = (u_j(\ell),\ s_j(\ell))$, with $u_j(\ell)$ the type and $s_j(\ell)$ the stage of the customer in position $\ell$ in queue $j$, $j = 1,\dots,J$. The state of queue $j$ containing $n_j$ customers is $\mathbf{c}_j = (c_j(1),\dots,c_j(n_j))$, and the state of the network is $\mathbf{c} = (\mathbf{c}_1,\dots,\mathbf{c}_J)$. Let $\{N(t)\}$ record the state of the Markov chain at state space $S = \{\mathbf{c} = (\mathbf{c}_1,\dots,\mathbf{c}_J)\}$ as described above. It is convenient to introduce the following notation. For $\mathbf{c} = (\mathbf{c}_1,\dots,\mathbf{c}_J)$, let

$C^{(u,s)}_{(\ell,j),(\ell',k)}\mathbf{c}$ denote state $\mathbf{c}'$ obtained from state $\mathbf{c}$ by removing the customer of type $u$ in stage $s$ in position $\ell$ from queue $j$ and adding that customer in position $\ell'$ to queue $k$.

For $j = 0$ a customer arrives to the network (and $\ell$ is redundant and will be set to zero), and we use $(u,0)$ to indicate that a customer of type $u$ arrives to the network. For $k = 0$ a customer departs from the network (and $\ell' = 0$). Note that for

[2] We will not assume that $U$ is finite. Therefore, we assume that $\sum_{u=1}^{U}\mu_0(u) < \infty$.

[3] We may model customer types that stay in the network by setting $L(u) = 1$. This also allows modelling a closed network in the same notation as used for the open network. We will not discuss closed networks separately.

$C^{(u,s)}_{(\ell,j),(\ell',k)}\mathbf{c} \in S$ it must be that $j = r(u,s)$, $k = r(u,s+1)$ for some $s = 0,\dots,L(u)$, with convention that $s+1 := 0$ if $s = L(u)$.[4] Observe that it may be that $C^{(u,s)}_{(\overline{\ell},j),(\ell',k)}\mathbf{c} = C^{(u,s)}_{(\ell,j),(\ell',k)}\mathbf{c}$, for $\overline{\ell} \neq \ell$, for example when all customers in queue $j$ are of type $u$ and in the same stage of their route. The transition rates are, for $u = 1,\dots,U$, $\mathbf{c}' \neq \mathbf{c}$, $\mathbf{c},\mathbf{c}' \in S$,

$$q(\mathbf{c},\mathbf{c}') =$$

$$\begin{cases} \displaystyle\sum_{\left\{\overline{\ell}' \,:\, C^{(u,0)}_{(0,0),(\overline{\ell}',k)}\mathbf{c} = C^{(u,0)}_{(0,0),(\ell',k)}\mathbf{c}\right\}} \mu_0(u)\delta_k(\overline{\ell}', n_k+1), & \text{if } \mathbf{c}' = C^{(u,0)}_{(0,0),(\ell',k)}\mathbf{c}, \\ \displaystyle\sum_{\left\{\overline{\ell},\overline{\ell}' \,:\, C^{(u,s)}_{(\overline{\ell},j),(\overline{\ell}',k)}\mathbf{c} = C^{(u,s)}_{(\ell,j),(\ell',k)}\mathbf{c}\right\}} \mu_j(u)\kappa_j(n_j)\gamma_j(\overline{\ell},n_j)\delta_k(\overline{\ell}'_k, n_k+1), & \text{if } \mathbf{c}' = C^{(u,s)}_{(\ell,j),(\ell',k)}\mathbf{c}, \\ \displaystyle\sum_{\left\{\overline{\ell}, \,:\, C^{(u,L(u))}_{(\overline{\ell},j),(0,0)}\mathbf{c} = C^{(u,L(u))}_{(\ell,j),(0,0)}\mathbf{c}\right\}} \mu_j(u)\kappa_j(n_j)\gamma_j(\overline{\ell},n_j), & \text{if } \mathbf{c}' = C^{(u,L(u))}_{(\ell,j),(0,0)}\mathbf{c}. \end{cases}$$

Let $\lambda_j(u,s)$ denote the arrival rate of customers of type $u$ to queue $j = r(u,s)$, $s = 1,\dots,L(u)$, $u = 1,\dots,U$. Then it must be that

$$\lambda_j(u,s) = \begin{cases} \mu_0(u), & \text{if } j = r(u,s), \\ 0, & \text{otherwise.} \end{cases}$$

The mean amount of work arriving to queue $j$ per unit time is

$$\rho_j = \sum_{u=1}^{U}\sum_{s=1}^{L(u)} \frac{\lambda_j(u,s)}{\mu_j(u)}, \quad j = 1,\dots,J.$$

We may now obtain the equilibrium distribution for the network with fixed routes.

**Theorem 4.3.1 (Network with fixed routes)** *Let $\{N(t)\}$ record the state of a network of queues with fixed routes in which queue $j$ operates according to the $(\kappa_j,\gamma_j,\delta_j)$-protocol. Assume that, for all $\mathbf{c}_j = (c_j(1),\dots,c_j(n))$,*

$$\sum_{\ell=1}^{n} \mu_j(u_j(\ell))\gamma_j(\ell,n) = \sum_{\ell=1}^{n} \mu_j(u_j(\ell))\delta_j(\ell,n), \tag{4.12}$$

*and*

$$G_j = \left[\sum_{n=0}^{\infty}\prod_{\ell=1}^{n} \frac{\rho_j}{\kappa_j(\ell)}\right]^{-1} < \infty,$$

*Let*

[4] Note that the type of the customer in position $\ell$ in queue $j$ is unique. The index $(u,s)$ is added for notational convenience to specify the customer's service requirement.

$$\pi_j(\mathbf{c}_j) = G_j \prod_{\ell=1}^{n_j} \frac{\rho_j(c_j(\ell))}{\kappa_j(\ell)}, \quad \mathbf{c}_j = (c_j(1), \ldots, c_j(n_j)),$$

*with* $\rho_j(c_j(\ell)) = \lambda_j(u_j(\ell), s_j(\ell))/\mu_j(u_j(\ell))$*. Then* $\{N(t)\}$ *has unique equilibrium distribution*

$$\pi(\mathbf{c}) = \prod_{j=1}^{J} \pi_j(\mathbf{c}_j), \quad \mathbf{c} \in S.$$

**Proof.** A natural guess for the reversed process is that customers of type $u$ arrive according to a Poisson process with rate $\mu_0(u)$ to queue $L(u)$ and follow the reversed route $r(u, L(u)), \ldots, r(u, 1)$, and that the transition rates have the role of $\gamma$ and $\delta$ reversed: for $u = 1, \ldots, U$, $\mathbf{c}' \neq \mathbf{c}$, $\mathbf{c}, \mathbf{c}' \in S$,

$$q^r(\mathbf{c}', \mathbf{c}) =$$

$$\begin{cases} \displaystyle\sum_{\left\{\overline{\ell}' \,:\, C^{(u,0)}_{(0,0),(\overline{\ell}',k)}\mathbf{c} = C^{(u,0)}_{(0,0),(\ell',k)}\mathbf{c}\right\}} \mu_k(u)\kappa_k(n_k+1)\delta_k(\overline{\ell}', n_k+1), & \text{if } \mathbf{c}' = C^{(u,0)}_{(0,0),(\ell',k)}\mathbf{c}, \\ \displaystyle\sum_{\left\{\overline{\ell}, \overline{\ell}' \,:\, C^{(u,s)}_{(\overline{\ell},j),(\overline{\ell}',k)}\mathbf{c} = C^{(u,s)}_{(\ell,j),(\ell',k)}\mathbf{c}\right\}} \mu_k(u)\kappa_k(n_k+1)\delta_k(\overline{\ell}'_k, n_k+1)\gamma_j(\overline{\ell}, n_j), & \text{if } \mathbf{c}' = C^{(u,s)}_{(\ell,j),(\ell',k)}\mathbf{c}, \\ \displaystyle\sum_{\left\{\overline{\ell}, \,:\, C^{(u,L(u))}_{(\overline{\ell},j),(0,0)}\mathbf{c} = C^{(u,L(u))}_{(\ell,j),(0,0)}\mathbf{c}\right\}} \mu_0(u)\gamma_j(\overline{\ell}, n_j), & \text{if } \mathbf{c}' = C^{(u,L(u))}_{(\ell,j),(0,0)}\mathbf{c}. \end{cases}$$

Observe that $\pi_j$ is the equilibrium distribution of queue $j$, recall Remark 4.2.8, and that $\pi$ is a distribution. For the proposed distribution $\pi$ and transition rates $q^r$ we have, for $\mathbf{c} = (\mathbf{c}_1, \ldots, \mathbf{c}_J)$, $\mathbf{c}_j = (c_j(1), \ldots, c_j(n_j))$, $j = 1, \ldots, J$,

$$\begin{aligned} \sum_{\mathbf{c}'} q(\mathbf{c}, \mathbf{c}') &= \sum_{u=1}^{U} \mu_0(u) + \sum_{j=1}^{J} \sum_{\ell_j=1}^{n_j} \mu_j(u_j(\ell_j))\kappa_j(n_j)\gamma_j(\ell_j, n_j), \\ \sum_{\mathbf{c}'} q^r(\mathbf{c}, \mathbf{c}') &= \sum_{u=1}^{U} \mu_0(u) + \sum_{k=1}^{J} \sum_{\ell_k=1}^{n_k} \mu_k(u_k(\ell_k))\kappa_k(n_k)\delta_k(\ell_k, n_k), \end{aligned}$$

so that (4.3) is satisfied due to (4.12). For $\mathbf{c}' = C^{(u,s)}_{(\ell,j),(\ell',k)}\mathbf{c}$, with $j, k \neq 0$, we have

$$\pi(\mathbf{c})q(\mathbf{c}, \mathbf{c}') = \pi(\mathbf{c}) \sum_{\left\{\overline{\ell}, \overline{\ell}' \,:\, C^{(u,s)}_{(\overline{\ell},j),(\overline{\ell}',k)}\mathbf{c} = C^{(u,s)}_{(\ell,j),(\ell',k)}\mathbf{c}\right\}} \mu_j(u)\kappa_j(n_j)\gamma_j(\overline{\ell}, n_j)\delta_k(\overline{\ell}'_k, n_k+1),$$

$$\begin{aligned} &\pi(\mathbf{c}')q^r(\mathbf{c}', \mathbf{c}) \\ &= \pi(\mathbf{c}) \sum_{\left\{\overline{\ell}, \overline{\ell}' \,:\, C^{(u,s)}_{(\overline{\ell},j),(\overline{\ell}',k)}\mathbf{c} = C^{(u,s)}_{(\ell,j),(\ell',k)}\mathbf{c}\right\}} \frac{\rho_k(c_k(\ell'_k))}{\rho_j(c_j(\ell_j))} \frac{\kappa_j(n_j)}{\kappa_k(n_k+1)} \mu_k(u)\kappa_k(n_k+1) \\ &\qquad \times \delta_k(\overline{\ell}'_k, n_k+1)\gamma_j(\overline{\ell}, n_j). \end{aligned}$$

The arrival rate of type $u$ customers equals $\mu_0(u)$ at all stages of their route, which implies that $\rho_k(c_k(\ell'_k))\mu_k(u) = \rho_j(c_j(\ell_j))\mu_j(u)$ so that $\pi(\mathbf{c})q(\mathbf{c},\mathbf{c}') = \pi(\mathbf{c}')q^r(\mathbf{c}',\mathbf{c})$. The other cases $j = 0$ and $k = 0$ follow by analogy, so that (4.2) is satisfied. Kelly's lemma 4.1.3 completes the proof. □

**Remark 4.3.2 (Fixed routes or Markovian routing)** We may also consider the network in which queue $j$ operates according to the $(\kappa_j, \gamma_j, \delta_j)$-protocol, $j = 1,\dots,J$, and customers select the next queue using Markovian routing. The analysis follows the analysis of the network with customer types presented in Remark 3.2.7.

We may use a network with fixed routes to model a network with Markovian routing by introducing a type for each possible customer route. Conversely, a network with fixed routes may be modelled using Markovian routing and customer types as described in Remark 3.2.7 via degenerate routing probabilities combined with types recording the stages.

Fixed routes may be the natural description in some applications, whereas Markovian routing may be natural in other applications. □

**Remark 4.3.3 (BCMP networks)** Networks consisting of multi-class queues operating under the FIFO, LIFO-PR, PS, and INF protocols are commonly referred to as BCMP networks, referring to F. Baskett, K.M. Chandy, R.R. Muntz, and F.G. Palacios [**?**]. The exposition in this section follows the lines of F.P. Kelly [**?**]. □

## 4.4 Quasi-reversibility

Burke's theorem 2.5.1 implies that the output process from a reversible queue before time $t$, the input process to that queue after $t$ and the state of the queue at $t$ are independent. This independence property implies that the equilibrium distribution in a feedforward network is a product over the marginal distributions of the queues and is the key-property to obtain a product-form equilibrium distribution in feedforward networks. Quasi-reversibility formalises this independence property as a starting point to obtain a product-form distribution for networks with fairly general queues.

Let customers of classes $c = 1,\dots,C$ arrive to a queue to receive service. Customers arrive one-by-one and all customers that arrive to the queue eventually leave without changing their class. Let $\{N(t)\}$ record the state of the queue. We will assume that $\{N(t),\ t \in \mathbb{R}\}$ is a Markov chain at state space $S$ and states $\mathbf{n} \in S$ with transition rates $q(\mathbf{n},\mathbf{n}')$, $\mathbf{n},\mathbf{n}' \in S$, and equilibrium distribution $\pi(\mathbf{n})$, $\mathbf{n} \in S$. We do not impose structure on the states except for the assumption that the state changes each time a customer arrives to the queue or departs from the queue. Let $S(c,\mathbf{n}) \subset S$ denote the set of states that may be obtained from state $\mathbf{n}$ when a customer of class $c$ arrives to the queue, $c = 1,\dots,C$, $\mathbf{n} \in S$. Let $\{A_c(t),\ t \in \mathbb{R}\}$ and $\{D_c(t),\ t \in \mathbb{R}\}$ record the arrival and departure processes of customers of class $c$.

**Definition 4.4.1 (Quasi-reversibility)** *A stationary Markov chain $\{N(t)\}$ recording the evolution of a queue to which customers of classes $c = 1,\dots,C$ arrive to* *receive service and eventually leave without changing their class is quasi-reversible if for all $t \in \mathbb{R}$ the state at time $t$, $N(t)$, is independent of $\{A_c(s),\ s > t\}$, the arrival process of class $c$ customers after time $t$, and independent of $\{D_c(s),\ s < t\}$, the departure process of class $c$ customers prior to time $t$, $c = 1,\dots,C$.*

Observe that the time-reversed Markov chain $\{N(-t)\}$ is also quasi-reversible as we may identify the arrival process of the forward Markov chain with the departure process of the time-reversed Markov chain.

**Theorem 4.4.2** *If $\{N(t)\}$ is a quasi-reversible Markov chain, then*

*(i) the arrival processes $\{A_c(t),\ t \in \mathbb{R}\}$, $c = 1,\dots,C$, form independent Poisson processes;*
*(ii) the departure processes $\{D_c(t),\ t \in \mathbb{R}\}$, $c = 1,\dots,C$, form independent Poisson processes.*

**Proof.** By the definition of quasi-reversibility the arrival rate of class $c$ customers in the interval $(t, t+h)$ is independent of $N(t)$. Hence, $\sum_{\mathbf{n}' \in S(c,\mathbf{n})} q(\mathbf{n},\mathbf{n}')$, the arrival rate of class $c$ customers given that $N(t) = \mathbf{n}$ depends only on the class of the customer:

$$\lambda(c) := \sum_{\mathbf{n}' \in S(c,\mathbf{n})} q(\mathbf{n},\mathbf{n}'), \quad c = 1,\dots,C, \tag{4.13}$$

for all $\mathbf{n} \in S$ and since $\{N(t)\}$ is a Markov chain is independent of $\{N(s),\ s < t\}$. Thus, the arrival process has stationary and independent increments so that it is a Poisson process.

Consider the time-reversed process. Identification of arrivals (departures) of the forward process with departures (arrivals) of the time-reversed process combined with the observation that the time-reversed process is quasi-reversible, implies that the departure process of class $c$ customers is Poisson with rates $\lambda(c)$. □

**Remark 4.4.3 (Algebraic characterisation of quasi-reversibility)** The last argument in the proof above also shows that

$$\lambda(c) = \sum_{\mathbf{n}' \in S(c,\mathbf{n})} q^r(\mathbf{n},\mathbf{n}'), \tag{4.14}$$

with $q^r$ the transition rates of the time-reversed Markov chain. Combining (4.13), and (4.14) yields, for $\mathbf{n} \in S$,

$$\lambda(c) = \sum_{\mathbf{n}' \in S(c,\mathbf{n})} q(\mathbf{n},\mathbf{n}') = \sum_{\mathbf{n}' \in S(c,\mathbf{n})} q^r(\mathbf{n},\mathbf{n}'). \tag{4.15}$$

Inserting the time-reversed transition rates yields, for $\mathbf{n} \in S$,

$$\sum_{\mathbf{n}' \in S(c,\mathbf{n})} \pi(\mathbf{n}) q(\mathbf{n},\mathbf{n}') = \sum_{\mathbf{n}' \in S(c,\mathbf{n})} \pi(\mathbf{n}') q(\mathbf{n}',\mathbf{n}). \tag{4.16}$$

Note that the summation in (4.16) is over subset of the states: those states that may be reached due to arrival of class $c$ customers. The total rate out of state $\mathbf{n}$ may also include transitions in which no customers arrive or depart from the queue. The partial balance property (4.16) provides an *algebraic characterisation of quasi-reversibility*:

> In equilibrium the flow out of state $\mathbf{n}$ due to a customer of type $c$ arriving to the queue balances with the probability flow into state $\mathbf{n}$ due to a customer of type $c$ departing from the queue. □

**Remark 4.4.4 (Reversibility and quasi-reversibility)** Reversibility and quasi-reversibility are separate notions. The birth-death process with state-dependent birth rates is reversible but not quasi-reversible since the arrival process depends on the state of the Markov chain. If we split state 1 in an $M|M|1$ queue with arrival rate $\lambda$ and service rate $\mu$ into two states, 1a and 1b, and set the transition rate from state 0 to state 1a to $p\lambda$ and to state 1b to $(1-p)\lambda$, and the transition rates from state 2 to state 1a to $(1-p)\mu$ and to state 1b to $p\mu$, $0 < p < 1$, then the arrival and departure processes to the system are not affected so that the Markov chain is quasi-reversible, but the Markov chain is reversible only if $p = 1/2$. □

Burke's theorem 2.5.1 is based on the assumption that the arrival process to a reversible queue is a Poisson process. Properties (i) and (ii) above are *not* sufficient for the Markov chain $\{N(t)\}$ to be quasi-reversible as these properties do not mention independence of the state of the Markov chain.

In applications the arrival process is often such that $N(t)$ is independent of $\{A_c(s),\ s > t\}$ and the form of the reversed process may allow us to conclude that $N(t)$ is also independent of $\{D_c(s),\ s < t\}$ so that the Markov chain $\{N(t)\}$ is quasi-reversible. For example, the Markov chain $\{N(t)\}$ recording the number of customers in the $M|M|1$ queue with one class of customers is quasi-reversible as a consequence of Burke's theorem 2.5.1. This result carries over to reversible queues with Poisson arrivals and one class of customers. For queues with multiple customer classes, Example 4.2.3 shows that the time-reversed multi-class LIFO-PR queue with class-dependent service rates is also a multi-class LIFO-PR queue with class-dependent service rates. Identifying the arrivals to the time-reversed queue with the departures from the forward time queue shows that the queue is quasi-reversible. Similarly, Example 4.2.4 shows that for the multi-class FIFO queue to be quasi-reversible it must be that the service rate does not depend on the class of the customers. The result for the multi-class LIFO-PR queue may be generalised to symmetric queues.

**Theorem 4.4.5 (Symmetric queue is quasi-reversible)** *Let $\{N(t)\}$ record the state of a symmetric queue to which customers of class $c$ arrive according to independent Poisson processes with rate $\lambda(c)$, $c = 1,\dots,C$. Then $\{N(t)\}$ is quasi-reversible.*

**Proof.** Arrivals of class $c$ customers, $c = 1,\dots,C$, occur according to independent Poisson processes. Therefore $N(t)$ is independent of $\{A_c(s),\ s > t\}$, $c = 1,\dots,C$. The transition rates of $\{N(t)\}$ are, for $\mathbf{c} = (c(1),\dots,c(n))$, $\mathbf{c}' \neq \mathbf{c}$, $\mathbf{c},\mathbf{c}' \in S$, $c = 1,\dots,C$,

$$q(\mathbf{c},\mathbf{c}') = \begin{cases} \lambda(c)\gamma(\ell,n+1), & \text{if } \mathbf{c}' = (c(1),\dots,c(\ell),c,c(\ell+1),\dots,c(n)), \\ \mu_{c(\ell)}\kappa(n)\gamma(\ell,n), & \text{if } \mathbf{c}' = c(1),\dots,c(\ell-1),c(\ell+1),\dots,c(n)). \end{cases} \quad (4.17)$$

Following the proof of Theorem 4.2.7, these are also the transition rates of the time-reversed queue. The arrival process to the time-reversed queue is a Poisson process. As arrivals in the time-reversed process coincide with departures of $\{N(t)\}$, we find that $N(t)$ is independent of $\{D_c(s),\ s < t\}$, which completes the proof. □

**Example 4.4.6 (Order Independent (OI) queue)** Let customers of type $c$ arrive to a queue according to independent Poisson processes with rates $\lambda(c)$, $c = 1,\dots,C$. If $n > 0$ customers are present, let $\mathbf{c} = (c(1),\dots,c(n))$, $c(i) \in \{1,\dots,C\}$, record the class of the customers in position $i$, $i = 1,\dots,n$. Let $\{N(t)\}$ record the state of the Markov chain at state space

$$S = \{\mathbf{c} : \mathbf{c} = (c(1),\dots,c(n)),\ c(i) \in \{1,\dots,C\},\ i = 1,\dots,n,\ n \in \mathbb{N}_0\}. \quad (4.18)$$

The queue operates under the following modification of the $(\kappa,\gamma,\delta)$-protocol:

- a customer of class $c$ requires a negative-exponentially distributed amount of service with rate $\mu(c)$;
- if $n > 0$ customers are present, in state $\mathbf{c}$ service is provided at rate $\kappa(\mathbf{c}) > 0$;
- in state $\mathbf{c} = (c(1),\dots,c(n))$ a fraction $\gamma(\ell,\mathbf{c})$ of the service effort is directed to the customer in position $\ell$, $\ell = 1,\dots,n$; if the customer in position $\ell$ completes service and leaves the queue customers in positions $\ell+1,\ell+2,\dots,n$ move to positions $\ell,\ell+1,\dots,n-1$, respectively;
- a customer that arrives in state $\mathbf{c} = (c(1),\dots,c(n))$ moves into position $n+1$.

In contrast with the standard $(\kappa,\gamma,\delta)$-protocol, we will *not* assume that

$$\sum_{\ell=1}^{n} \gamma(\ell,\mathbf{c}) = 1,$$

so that part of the service effort might be wasted. The transition rates are, for $\mathbf{c} = (c(1),\dots,c(n))$, $\mathbf{c}' \neq \mathbf{c}$, $\mathbf{c},\mathbf{c}' \in S$,

$$q(\mathbf{c},\mathbf{c}') = \begin{cases} \lambda(c), & \text{if } \mathbf{c}' = (c(1),\dots,c(n),c),\ c \in \{1,\dots,C\}, \\ \kappa(\mathbf{c})\mu(c(\ell))\gamma(\ell,\mathbf{c}), & \text{if } \mathbf{c}' = (c(1),\dots,c(\ell-1),c(\ell+1),\dots,c(n)). \end{cases}$$

The queue is an *Order Independent queue* if, there exist $\eta(n)$, $n \in \mathbb{N}$, and $s(\ell,\mathbf{c})$, $\ell = 1,\dots,n$, $\mathbf{c} = (c(1),\dots,c(n)) \in S$, $n \in \mathbb{N}$, such that for all $\mathbf{c} = (c(1),\dots,c(n)) \in S$ and all $\ell = 1,\dots,n$, the service rates can be written as

$$\kappa(\mathbf{c})\mu(c(\ell))\gamma(\ell,\mathbf{c}) = \eta(n)s(\ell,\mathbf{c})$$

such that

(i) $s(\ell,c(1),\dots,c(n)) = s(\ell,c(1),\dots,c(\ell))$, $\ell = 1,\dots,n$,
(ii) $k(c(1),\dots,c(n)) := \sum_{\ell=1}^{n} s(\ell,c(\sigma(1)),\dots,c(\sigma(n)))$ for all permutations $(\sigma(1),\dots,\sigma(n))$ of $(1,\dots,n)$,

(iii) $\eta(n) > 0$ for $n > 0$ and $s_1(c) > 0$, $c = 1, \dots, C$.

The function $s(\ell, \mathbf{c})$ regulates the rate at which service is provided to position $\ell$. Condition (i) requires that the service rate of a customer in the queue depends only on its own class and the customer classes of customers in front of it in the queue. Condition (ii) is the distinguishing condition for the OI queue and requires that the total service rate is independent of the order of the customers in the queue. Condition (iii) is a necessary and sufficient condition for the Markov chain to be irreducible.

The OI queue is quasi-reversible and has equilibrium distribution

$$\pi(c(1), \dots, c(n)) = G_{OI} \prod_{\ell=1}^{n} \frac{\lambda(c(\ell))}{\eta(\ell) k(c(1), \dots, c(\ell))}, \quad \mathbf{c} \in S, \tag{4.19}$$

with normalising constant

$$G_{OI}^{-1} = \sum_{\mathbf{c} \in S} \prod_{\ell=1}^{n} \frac{\lambda(c(\ell))}{\eta(\ell) k(c(1), \dots, c(\ell))}$$

provided that $G_{OI} < \infty$.

The global balance equations are

$$\pi(\mathbf{0}) \sum_{c=1}^{C} \lambda(c) = \eta(1) k(c) \sum_{c=1}^{C} \pi(c) \tag{4.20}$$

and for $n \geq 1$

$$\begin{aligned}
&\pi(c(1), \dots, c(n)) \left( \sum_{c=1}^{C} \lambda(c) + \eta(n) k(c(1), \dots, c(n)) \right) \\
&= \sum_{c=1}^{C} \sum_{\ell=0}^{n} \pi((c(1), \dots, c(\ell), c, c(\ell+1), \dots, c(n)) \eta(n+1) \\
&\qquad\qquad \times s(\ell+1, c(1), \dots, c(\ell), c, c(\ell+1), \dots, c(n)) \\
&\quad + \quad \lambda(c(n)) \pi(c(1), \dots, c(n-1)).
\end{aligned} \tag{4.21}$$

Observe that the distribution (4.19) satisfies

$$\pi(c(1), \dots, c(n)) \eta(n) k(c(1), \dots, c(n)) = \lambda(c(n)) \pi(c(1), \dots, c(n-1)), \tag{4.22}$$

so that it is sufficient to show that $\pi$ satisfies the algebraic characterisation of quasi-reversibility (4.16), for $c = 1, \dots, C$:

$$\sum_{\ell=0}^{n} \frac{\pi(c(1), \dots, c(\ell), c, c(\ell+1), \dots, c(n))}{\pi(c(1), \dots, c(n))} \eta(n+1) s(\ell+1, c(1), \dots, c(\ell), c) = \lambda(c). \tag{4.23}$$

Following [**?**], this may be shown by induction in $n$. First observe that (4.23) is satisfied for $n = 0$:

$$\frac{\pi(c)}{\pi(\mathbf{0})}\eta(1)s(1,c)=\frac{\lambda(c)}{\eta(1)k(c)}\eta(1)s(1,c)=\lambda(c),$$

since $s(1,c)=k(c)$. Now assume (4.23) is satisfied for $1,2,\dots,n-1$. We have

$$\begin{aligned}
&\sum_{\ell=0}^{n}\frac{\pi(c(1),\dots,c(\ell),c,c(\ell+1),\dots,c(n))}{\pi(c(1),\dots,c(n))}\eta(n+1)s(\ell+1,c(1),\dots,c(\ell),c)\\
&=\ \sum_{\ell=0}^{n-1}\frac{\pi(c(1),\dots,c(\ell),c,c(\ell+1),\dots,c(n))}{\pi(c(1),\dots,c(n))}\eta(n+1)s(\ell+1,c(1),\dots,c(\ell),c)\\
&\qquad+\frac{\pi(c(1),\dots,c(n),c)}{\pi(c(1),\dots,c(n))}\eta(n+1)s(n+1,c(1),\dots,c(n),c)\\
&\overset{(4.22)}{=}\sum_{\ell=0}^{n-1}\frac{\pi(c(1),\dots,c(\ell),c,c(\ell+1),\dots,c(n-1))}{\pi(c(1),\dots,c(n-1))k(c(1),\dots,c(\ell),c,c(\ell+1),\dots,c(n))}\\
&\qquad\qquad\times\frac{\eta(n)k(c(1),\dots,c(n))}{\eta(n+1)}\eta(n+1)s(\ell+1,c(1),\dots,c(\ell),c)\\
&\qquad+\frac{\lambda(c)}{\eta(n+1)k(c(1),\dots,c(n),c)}\eta(n+1)s(n+1,c(1),\dots,c(n),c)\\
&\overset{\text{(ii)}}{=}\ \frac{k(c(1),\dots,c(n))}{k(c(1),\dots,c(n),c)}\sum_{\ell=0}^{n-1}\frac{\pi(c(1),\dots,c(\ell),c,c(\ell+1),\dots,c(n-1))}{\pi(c(1),\dots,c(n-1))}\\
&\qquad\qquad\times\eta(n)s(\ell+1,c(1),\dots,c(\ell),c)\\
&\qquad+\lambda(c)\frac{s(n+1,c(1),\dots,c(n),c)}{k(c(1),\dots,c(n),c)}\\
&\overset{(*)}{=}\ \lambda(c)\frac{k(c(1),\dots,c(n))+s(n+1,c(1),\dots,c(n),c)}{k(c(1),\dots,c(n),c)}\\
&\overset{(\text{ii}*)}{=}\ \lambda(c),
\end{aligned}$$

where (ii) is due to $k(\mathbf{c})$ being invariant under permutations (condition (ii)), $(*)$ due to the induction hypothesis, and (ii$*$) due to the definition of $k(\mathbf{c})$.

Several queues and queue disciplines may be modeled as OI queues, including the $M|M|1|c$ queue and the PS and INF queue disciplines. The LCFS-PR discipline violates condition (i) and therefore cannot be modeled as OI queue.

An important example of an OI queue is the *MultiServer centre with Concurrent Classes of Customers* (MSCCC queue). Customers of class $c=1,\dots,C$ arrive to an MSCCC queue according to independent Poisson processes with rates $\lambda(c)$, $c=1,\dots,C$, and have negative-exponential service times with rate $\mu$. Customers of each class are served in the order of their arrival. When a server becomes free, the queue is searched from the front looking for the first customer to admit into service subject to the following constraints: at most $K$ customers can be in service and at most $B_c\geq 1$ customers of class $c$ can be in service, $c=1,\dots,C$. Let

$$\eta(n)=1,\quad n=1,2,\dots,$$

and

$$s(\ell, c(1),\dots,c(n)) = \begin{cases} \mu, & \text{if the customer in position } \ell \text{ is served in } (c(1),\dots,c(n)), \\ 0, & \text{otherwise.} \end{cases}$$

Let $n(c) = \sum_{\ell=1}^{n} \mathbb{1}(c(\ell) = c\}$ the total number of customers of class $c$, $c = 1,\dots,C$. Then

$$k(c(1),\dots,c(n)) = \mu \min\{K, \sum_{c=1}^{C} \min\{n(c), B_c\}\}$$

is equal for all permutations of the customers, and all conditions of the OI queue are satisfied. The MSCCC queue is a quasi-reversible generalisation of the FCFS queue. □

## 4.5 Networks of quasi-reversible queues with fixed routes

Let customers of types $u = 1,\dots,U$ arrive to a network of $J$ quasi-reversible queues according to a Poisson process with rate $\mu_0(u)$, $u = 1,\dots,U$. Following the notation introduced in Section 4.3, a customer's type uniquely determines his route through the network along the sequence of queues $r(u,1), r(u,2),\dots,r(u,L(u))$.

Let $\{N_j(t)\}$ at state space $S_j$ with transition rates $q_j(\mathbf{c}_j, \mathbf{c}'_j)$, $\mathbf{c}_j, \mathbf{c}'_j \in S_j$, record the state of queue $j$ with customers of class $(u,s)$ arriving according to Poisson processes with rate $\lambda_j(u,s)$ and let $\pi_j = (\pi_j(\mathbf{c}_j),\ \mathbf{c}_j \in S_j)$ denote the equilibrium distribution of $\{N_j(t)\}$, $j = 1,\dots,J$. Then $q_j$ and $\pi_j$ satisfy (4.15), (4.16), in particular $\lambda_j(u,s) = \sum_{\mathbf{c}'_j \in S_j((u,s),\mathbf{c}_j)} q_j(\mathbf{c}_j, \mathbf{c}'_j)$.

Consider $\{N(t)\}$ at state space $S = S_1 \times \cdots \times S_J$, the Cartesian product of the state spaces of the queues, with states $\mathbf{c} = (\mathbf{c}_1,\dots,\mathbf{c}_J)$. It is convenient to introduce the following notation. For $\mathbf{c} = (\mathbf{c}_1,\dots,\mathbf{c}_J)$, and $j,k = 0,\dots,J$, let

$C_{j,k}^{(u,s)}\mathbf{c}$ denote the set of states $\mathbf{c}'$ obtained from state $\mathbf{c}$ by removing the customer of type $u$ in stage $s$ from queue $j$ and adding that customer in stage $s+1$ to queue $k$:

$$(C_{j,k}^{(u,s)}\mathbf{c})_i = \begin{cases} \{\mathbf{c}_i\}, & \text{if } i \neq j,k, \\ S_k((u,s+1),\mathbf{c}_k), & \text{if } i = k, \\ \{\mathbf{c}'_j \text{ s.t. } \mathbf{c}_j \in S_j((u,s),\mathbf{c}'_j)\}, & \text{if } i = j, \end{cases}$$

with the convention that for $j = 0$ a customer arrives to the network which also implies that $k = r(u,1)$ and we set $s = 0$, and for $k = 0$ a customer departs from the network which implies that $j = r(u,L(u))$ and we set $L(u)+1 = 0$. Note that $j,k$ are uniquely defined by $(u,s)$ and are added for notational convenience. The transition rates are, for $u = 1,\dots,U$, $\mathbf{c} \neq \mathbf{c}'$, $\mathbf{c},\mathbf{c}' \in S$,

$$q(\mathbf{c},\mathbf{c}') =$$

$$\begin{cases} q_k(\mathbf{c}_k,\mathbf{c}'_k), & \text{if } \mathbf{c}' \in C^{(u,1)}_{0,k}\mathbf{c}, & \text{(arrival)}, \\ q_j(\mathbf{c}_j,\mathbf{c}'_j)\dfrac{q_k(\mathbf{c}_k,\mathbf{c}'_k)}{\sum_{\mathbf{c}'_k\in S_k((u,s+1),\mathbf{c}_k)} q_k(\mathbf{c}_k,\mathbf{c}'_k)}, & \text{if } \mathbf{c}' \in C^{(u,s)}_{j,k}\mathbf{c}, & \text{(routing)}, \\ q_j(\mathbf{c}_j,\mathbf{c}'_j), & \text{if } \mathbf{c}' \in C^{(u,L(u))}_{j,0}\mathbf{c}, & \text{(departure)}, \\ q_j(\mathbf{c}_j,\mathbf{c}'_j), & \text{if } \mathbf{c}_j,\mathbf{c}'_j \in S_j,\ \mathbf{c}'_i = \mathbf{c}_i,\ i \neq j, & \text{(internal)}, \end{cases} \tag{4.24}$$

where an internal transition of queue $j$ is a transition without the arrival or departure of a customer from queue $j$, $j = 1,\dots,J$. Observe that quasi-reversibility implies that

$$\frac{q_k(\mathbf{c}_k,\mathbf{c}'_k)}{\sum_{\mathbf{c}'_k\in S_k((u,s+1),\mathbf{c}_k)} q_k(\mathbf{c}_k,\mathbf{c}'_k)} = \frac{q_k(\mathbf{c}_k,\mathbf{c}'_k)}{\lambda_k(u,s+1)}. \tag{4.25}$$

This is the probability that state $\mathbf{c}'_k$ is selected from the set $S_k((u,s+1),\mathbf{c}_k)$, by analogy with the probability $\delta_k(\ell,\mathbf{c}_k+1)$ that a customer is placed in position $\ell$ in the network of Section 4.3.

**Theorem 4.5.1 (Network of quasi-reversible queues with fixed routes)** *Let $\{N(t)\} = \{(N_1(t),\dots,N_J(t))\}$ record the state of a network of $J$ quasi-reversible queues to which customers of types $u = 1,\dots,U$ arrive according to independent Poisson processes with rates $\mu_0(u)$ to follow fixed route $r(u,1),r(u,2),\dots,r(u,L(u))$, $u = 1,\dots,U$. Let $S_j$, $q_j$, and $\pi_j$ denote the state space, transition rates and equilibrium distribution of queue $j$, $j = 1,\dots,J$. Then $\{N(t)\}$ has equilibrium distribution*

$$\pi(\mathbf{c}_1,\dots,\mathbf{c}_J) = \prod_{j=1}^{J} \pi_j(\mathbf{c}_j), \quad (\mathbf{c}_1,\dots,\mathbf{c}_J) \in S = S_1 \times \cdots \times S_J. \tag{4.26}$$

**Remark 4.5.2 (Feedforward network of quasi-reversible queues)** If the routes are such that the network is a feedforward network, then by analogy with the results in Section 2.5 for feedforward networks of $M|M|1$ queues based on Burke's theorem 2.5.1, for a network of quasi-reversible queues states $N_j(t)$ of the queues are independent at fixed times so that the equilibrium distribution of the network is (4.26). Moreover, the arrival process of customers of type $u$ at each of the queues is a Poisson process with rate $\mu_0(u)$, $u = 1,\dots,U$. □

**Proof of Theorem 4.5.1.** A natural guess for the time-reversed process is that customers of types $u = 1,\dots,U$ arrive according to a Poisson process with rate $\mu_0(u)$, $u = 1,\dots,U$, and route through the network along the sequence of queues in reversed order $r(u,L(u)),\dots,r(u,1)$ and that each queue operates according to its time-reversed transition rates. Recall that quasi-reversibility implies (4.25). Define, for $u = 1,\dots,U$, $\mathbf{c} \neq \mathbf{c}'$, $\mathbf{c},\mathbf{c}' \in S$,

$$q^r(\mathbf{c}',\mathbf{c}) =$$

$$\begin{cases} q^r_k(\mathbf{c}'_k,\mathbf{c}_k), & \text{if } \mathbf{c}' \in C^{(u,1)}_{0,k}\mathbf{c}, \quad \text{(departure)}, \\ q^r_k(\mathbf{c}'_k,\mathbf{c}_k)\dfrac{q^r_j(\mathbf{c}'_j,\mathbf{c}_j)}{\lambda_j(u,s)}, & \text{if } \mathbf{c}' \in C^{(u,s)}_{j,k}\mathbf{c}, \quad \text{(routing)}, \\ q^r_j(\mathbf{c}'_j,\mathbf{c}_j), & \text{if } \mathbf{c}' \in C^{(u,L(u))}_{j,0}\mathbf{c}, \quad \text{(arrival)}, \\ q^r_j(\mathbf{c}'_j,\mathbf{c}_j), & \text{if } \mathbf{c}_j,\mathbf{c}'_j \in S_j,\ \mathbf{c}'_i = \mathbf{c}_i,\ i \neq j, \quad \text{(internal)}. \end{cases} \tag{4.27}$$

For a routing transition from queue $j = r(u,s)$ to queue $k = r(u,s+1)$ it must be that $\lambda_j(u,s) = \lambda_k(u,s+1)$, which implies that

$$\pi_j(\mathbf{c}_j)\pi_k(\mathbf{c}_k)q_j(\mathbf{c}_j,\mathbf{c}'_j)\frac{q_k(\mathbf{c}_k,\mathbf{c}'_k)}{\lambda_k(u,s+1)} = \pi(\mathbf{c}'_j)\pi_k(\mathbf{c}'_k)q^r_k(\mathbf{c}'_k,\mathbf{c}_k)\frac{q^r_j(\mathbf{c}'_j,\mathbf{c}_j)}{\lambda_j(u,s)},$$

so that condition (4.3) of Kelly's lemma 4.1.3 is satisfied for routing transitions. By analogy, condition (4.3) of Kelly's lemma 4.1.3 is satisfied for arrival/departure and internal transitions. Condition (4.2) is readily verified, which shows that (4.27) are the transition rates of the time-reversed process and that (4.26) is the equilibrium distribution of $\{N(t)\}$. □

**Remark 4.5.3 (Closed networks of quasi-reversible queues)** If customers of type $u$ departing from queue $r(u,L(u))$ immediately return to queue $r(u,1)$, $u = 1,\ldots,U$, the network is a closed network of quasi-reversible queues. The Markov chain $\{N(t)\}$ at state space $S_M$ with transition rates (4.24) consisting only of the parts (routing) and (internal) records the number of customers in the queues. Except for normalisation, the equilibrium distribution remains that of the open network:

$$\pi(\mathbf{c}_1,\ldots,\mathbf{c}_J) = G_M\prod_{j=1}^{J}\pi_j(\mathbf{c}_j), \quad (\mathbf{c}_1,\ldots,\mathbf{c}_J) \in S_M \subset S_1 \times \cdots \times S_J,$$

with $G_M$ the normalising constant. □

**Remark 4.5.4 (Markovian routing)** We may also consider the network in which customers select the next queue using Markovian routing, recall Remark 4.3.2. Chapter **??** considers a network with state-dependent routing. □

## 4.6 Literature

follow Kelly