The roles of the deviation matrix and coupling in determining the value of capacity in $M/M/1/C$ queues

Peter Braunsteins, Sophie Hautphenne and Peter Taylor

March 23, 2015

Abstract

In an $M/M/1/C$ queue, customers are lost when they arrive to find C customers already present. Assuming that each arriving customer brings a certain amount of revenue, we are interested in calculating the value of an extra waiting place in terms of the expected amount of extra revenue that the queue will earn over a finite time horizon $[0, t]$.

There are different ways of approaching this problem. One involves the derivation of Markov renewal equations, conditioning on the first instance at which the state of the queue changes, a second involves expressing the value of capacity in terms of the entries of a transient analogue of the deviation matrix, discussed by Coolen-Schrijner and van Doorn in [6], and a third involves an elegant coupling argument.

In this paper, we shall compare and contrast these approaches and, in particular, use the coupling analysis to explain why the value of an extra unit of capacity remains the same when the arrival and service rates are interchanged when the queue starts at full capacity.

1 Introduction

Assume that you are the manager of an $M/M/1/C$ queue, and you have the option of acquiring or releasing waiting places, which we shall refer to as *units of capacity*, at time points $t_1, t_2, \ldots$ Customers accepted into the queue remain there until they are served and generate θ units of revenue, whereas customers that arrive when the queue is full are turned away and subsequently generate no revenue. At each time point t_j, your task is to

determine how much to pay for an additional unit of capacity or how much you would want to be reimbursed for relinquishing a unit of capacity. To determine these prices, you consider that each customer rejected from the queue is a lost opportunity that costs θ units of revenue, and calculate the expected revenue lost over the time period $[t_j, t_{j+1})$, given the arrival rate λ, service rate μ and number of customers $X(t_j)$ for various values of the capacity C.

In this paper we shall compare and contrast three different ways of approaching the above problem. Specifically we consider the problem of calculating the expected lost revenue of an $M/M/1/C$ queue over a finite time horizon $[0, t)$. First, in Section 2, we shall employ a Markov renewal analysis similar to that used in [5] for the $M/M/C/C$ queue. Then, in Section 3, we shall relate this to a transient analogue of the *deviation matrix* defined in Coolen-Schrijner and van Doorn [6], at the same time deriving a number of interesting properties of this matrix. In Section 4, we shall adopt a completely different approach, coupling the evolution of $M/M/1/C$ queues with different initial numbers of customers. In Section 5, we shall progress to a discussion of how to determine 'buying' and 'selling' prices for capacity using coupling arguments. In particular, we shall establish the counter-intuitive result that the buying and selling prices remain identical if the arrival and service rates are interchanged. Finally in Section 6, we shall make some concluding remarks.

2 Expected Lost Revenue

Our first method is based on that of [5], where the expected lost revenue was calculated for an $M/M/C/C$ queue.

Let $R_{n,C}(t)$ be the expected revenue lost in the time interval $[0, t)$ given capacity C and initial queue length $n \in \{0, \ldots, C\}$, and let $R_{n,C}(t|x)$ be the same quantity, conditional on the first change in queue length occurring at time x. Since revenue is lost at rate $\theta\lambda$ when the queue is full and not at all when less than C customers are present, we have

$$R_{n,C}(t|x) = \begin{cases} 0, & 0 \leq n < C, t < x \\ \theta\lambda t, & n = C, t < x \\ R_{1,C}(t-x), & n = 0, t \geq x \\ \frac{\mu}{\lambda+\mu} R_{n-1,C}(t-x) + \frac{\lambda}{\lambda+\mu} R_{n+1,C}(t-x) & 0 < n < C, t \geq x \\ \theta\lambda x + R_{C-1,C}(t-x) & n = C, t \geq x. \end{cases} \tag{2.1}$$

With $F_n(\cdot)$ the distribution function of the time until the first transition occurs when there are n customers in the queue at time 0, we have

$$R_{n,C}(t) = \int_0^t R_{n,C}(t|x)dF_n(x). \tag{2.2}$$

Substituting Equation (2.1) into Equation (2.2), three distinct cases arise:

Case 1: $n = 0$**.** In this case, $F_0(x) = 1 - e^{-\lambda x}$, and from (2.1), we obtain

$$\begin{aligned} R_{0,C}(t) &= \int_0^t R_{0,C}(t|x)dF_0(x) \\ &= \int_0^t R_{1,C}(t-x)\lambda e^{-\lambda x}dx. \end{aligned} \tag{2.3}$$

Case 2: $0 < n < C$**.** In this case, $F_n(x) = 1 - e^{-(\lambda+\mu)x}$, and (2.1) gives

$$\begin{aligned} R_{n,C}(t) &= \int_0^t R_n(t|x)dF_n(x) \\ &= \int_0^t \left[\mu R_{n-1}(t-x) + \lambda R_{n+1}(t-x)\right] e^{-(\lambda+\mu)x}dx. \end{aligned} \tag{2.4}$$

Case 3: $n = C$**.** In this case, $F_C(x) = 1 - e^{-\mu x}$, so by (2.1),

$$\begin{aligned} R_{C,C}(t) &= \int_0^t R_{C,C}(t|x)dF_C(x) + \int_t^\infty R_{C,C}(t|x)dF_C(x) \\ &= \int_0^t R_{C-1,C}(t-x)\mu e^{-\mu x}dx + \frac{\theta\lambda}{\mu}\left(1 - e^{-\mu t}\right). \end{aligned} \tag{2.5}$$

For complex s with $\Re(s) > 0$, by taking the Laplace transform of (2.3)-(2.5), we see that $\tilde{R}_{n,C}(s) = \int_0^\infty e^{-st}R_{n,C}(t)dt$ satisfies the system of second order difference equations

$$\tilde{R}_{0,C}(s) = \frac{\lambda}{s+\lambda}\tilde{R}_{1,C}(s) \tag{2.6}$$

$$\tilde{R}_{n,C}(s) = \frac{\lambda}{s+\lambda+\mu}\tilde{R}_{n+1,C}(s) + \frac{\mu}{s+\lambda+\mu}\tilde{R}_{n-1,C}(s), \quad 0 < n < C \tag{2.7}$$

$$\tilde{R}_{C,C}(s) = \frac{\mu}{s+\mu}\tilde{R}_{C-1,C}(s) + \frac{\theta\lambda}{s(s+\mu)}, \tag{2.8}$$

which has the explicit solution

$$\tilde{R}_{n,C}(s) = A(s)r_1^n(s) + B(s)r_2^n(s), \quad 0 \le n \le C, \tag{2.9}$$

where,

$$r_{1,2}(s) = \frac{\lambda + \mu + s \pm \sqrt{(\lambda + \mu + s)^2 - 4\mu\lambda}}{2\lambda}, \tag{2.10}$$

$$A(s) = -B(s)\left[\frac{1 - \frac{\lambda}{s+\lambda}r_2(s)}{1 - \frac{\lambda}{s+\lambda}r_1(s)}\right], \tag{2.11}$$

and

$$B(s)\left[r_2^{C-1}\left(r_2(s) - \frac{\mu}{\mu + s}\right) - r_1^{C-1}(s)\left(\frac{1 - \frac{\lambda}{s+\lambda}r_2(s)}{1 - \frac{\lambda}{s+\lambda}r_1(s)}\right)\left(r_1(s) - \frac{\mu}{s+\mu}\right)\right] = \frac{\theta\lambda}{s(\mu + s)}. \tag{2.12}$$

The Laplace transform $\tilde{R}_{n,C}(s)$ can be easily inverted numerically using, for example, the method described in [1].

Example: the $M/M/1/5$ loss system

When $C = 5$, $\theta = 1$, $\lambda = 3$ and $\mu = 5$, the values of $R_{n,C}(t)$ for $t \in [0, 10]$ are given in Figure 2.1. In Figure 2.2, $R_{n,C}(t)$ is again plotted with the same values of C and θ, but this time with $\mu = 3$ and $\lambda = 5$. In both cases $R_{0,5}(t)$ is the lowest curve and $R_{5,5}(t)$ is the highest.

In Figures 2.1 and 2.2, we observe that, when t becomes large, $R_{n,C}(t)$ is well approximated by a linear function with a common gradient for all n. This is due to the convergence of the continuous-time Markov chain $\{X(t) : t \ge 0\}$, whose state gives the number of customers, to its stationary distribution, $\boldsymbol{\pi}$, as t becomes large. In the steady state, arriving customers are rejected with probability

$$\pi_C = \frac{(1 - \rho)\rho^C}{1 - \rho^{C+1}}$$

where $\rho = \lambda/\mu$, and hence the gradient of these linear approximations is $\theta\lambda\pi_C$.

The difference in the height of the functions $R_{n,C}(t)$ reflects the difference in the expected lost revenue before the steady state is reached. The deviation and transient deviation matrices, which are the focus of the next section, turn out to be important in understanding this behavior.

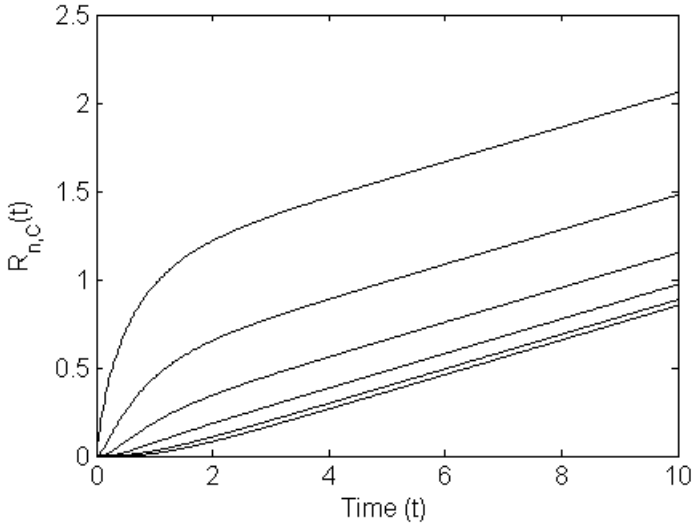

Figure 2.1: Expected lost revenue function for $n = 0, ..., 5$ when $C = 5$, $\lambda = 3$ and $\mu = 5$

3 The deviation matrix

As in Section 2, let $\{X(t) : t \geq 0\}$ denote the queue length process of an $M/M/1/C$ queue, and $p_{n,C}(t) = \mathrm{P}[X(t) = C|X(0) = n]$. Observe that the difference between the expected loss function $R_{n,C}(t)$ and the linear function $\lambda\theta\pi_C t$ can be expressed as

$$\begin{aligned} R_{n,C}(t) - \lambda\theta\pi_C t &= \lambda\theta \int_0^t p_{n,C}(u)du - \lambda\theta\pi_C t \\ &= \lambda\theta \int_0^t [p_{n,C}(u) - \pi_C]\, du. \end{aligned} \tag{3.1}$$

The last integral is a transient version of the *deviation matrix* corresponding to the Markov chain $\{X(t) : t \geq 0\}$. For an irreducible, positive-recurrent, continuous-time Markov chain on the state space $\mathcal{S}$ with generator Q, this matrix was studied by Coolen-Schrijner and van Doorn in [6]. It is the matrix whose $(i, j)th$ element is

$$D_{i,j} = \int_0^\infty [p_{i,j}(u) - \pi_j]\, du,$$

where $p_{i,j}(u) = \mathrm{P}[X(u) = j|X(0) = i] = [\exp(Qu)]_{i,j}$, and $\boldsymbol{\pi} \equiv (\pi_j)$ is the stationary distribution, which satisfies $\boldsymbol{\pi}Q = 0$ and $\boldsymbol{\pi}\mathbf{1} = 1$.

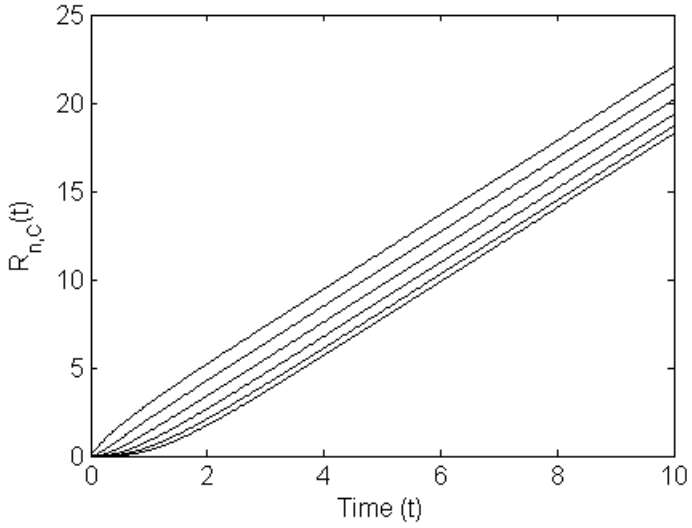

Figure 2.2: Expected lost revenue function for $n = 0, ..., 5$ when $C = 5$, $\lambda = 5$ and $\mu = 3$

For a specified column vector $\boldsymbol{g}$, the deviation matrix comes into play when solving Poisson's equation,

$$Q\boldsymbol{h} = \boldsymbol{g} - w\mathbf{1}, \tag{3.2}$$

for the vector-scalar pair $(\boldsymbol{h}, w)$. When the state space of $\{X(t)\}$ is finite, the solution to (3.2) is

$$\begin{aligned} \boldsymbol{h} &= -D\boldsymbol{g} + c\mathbf{1}, \\ w &= \boldsymbol{\pi}\boldsymbol{g}, \end{aligned}$$

where c is a constant that needs to be specified.

The entries of the deviation matrix D can be expressed in terms of expected first passage times:

$$D_{i,j} = \pi_j \left(m^e_j - m_{i,j}\right), \tag{3.3}$$

where $m_{i,j}$ is the mean first entrance time from state i to state j, and m^e_j is the mean first entrance time to state j from the stationary distribution, that is,

$$m_{i,j} = \mathrm{E}\big[\inf\{t \,:\, X(t) = j\} \mid X(0) = i\big], \qquad m^e_j = \sum_i \pi_i m_{i,j},$$

see [6]. Conversely, Equation (3.3) allows us to express the mean first passage times in terms of the entries of the deviation matrix:

$$m_{i,j} = \pi_j^{-1} (D_{j,j} - D_{i,j}).$$

When the state space is infinite, the deviation matrix does not always exist. Coolen-Schrijner and van Doorn established a necessary and sufficient condition for existence, see [6, Theorem 4.1]:

Theorem 1 *The deviation matrix D of $\{X(t)\}$ exists if and only if $m_j^e < \infty$ for some (and then for every) state $j \in \mathcal{S}$.*

It follows that D exists for any ergodic Markov chain on a finite number of states and, in particular, the deviation matrix exists for the $M/M/1/C$ queue.

Let $\Pi = \mathbf{1}\boldsymbol{\pi}$. The deviation matrix can be written explicitly in terms of the generator Q as

$$D = (\Pi - Q)^{-1} - \Pi. \tag{3.4}$$

It satisfies the properties

$$\begin{aligned} D\mathbf{1} &= \mathbf{0}, && (3.5)\\ \boldsymbol{\pi} D &= \mathbf{0}, && (3.6)\\ D(-Q) &= (-Q)D = I - \Pi, && (3.7)\\ (-Q)D(-Q) &= -Q, && (3.8)\\ D(-Q)D &= D. && (3.9) \end{aligned}$$

The last three properties imply that not only is D a generalised inverse of $-Q$, it is the *group*, or *Drazin* inverse of $-Q$.

For an M/M/1/C queue the mean first passage times $m_{i,j}$ have an explicit expression, and so have the entries of the deviation matrix, see for instance [9] and [7].

3.1 The transient deviation matrix

Observe that Equations (2.3)-(2.5) for the expected loss function can be rewritten as

$$\begin{aligned}
R_{0,C}(t) &= e^{-\lambda t}\int_0^t R_{1,C}(u)\lambda e^{\lambda u}du, \\
R_{n,C}(t) &= e^{-(\lambda+\mu)t}\int_0^t \left[\mu R_{n-1,C}(u)+\lambda R_{n+1,C}(u)\right]e^{(\lambda+\mu)u}du, \quad 1\le n\le C-1 \\
R_{C,C}(t) &= e^{-\mu t}\int_0^t R_{C-1,C}(u)\mu e^{\mu u}dx+\frac{\theta\lambda}{\mu}\left(1-e^{-\mu t}\right),
\end{aligned}$$

which can then be transformed into a time-dependent version of Poisson's equation of the form

$$\boldsymbol{R}'(t) = Q\boldsymbol{R}(t)+\boldsymbol{g},$$

where $\boldsymbol{g}^\top = (0,\ldots,0,\lambda\theta) = \lambda\theta\boldsymbol{e}_C$. In (3.1), we effectively wrote the solution,

$$R_{n,C}(t) = \lambda\theta\pi_C t + \lambda\theta D_{n,C}(t), \tag{3.10}$$

in terms of the (n,C)th entry of the matrix

$$D(t) = \int_0^t \left[P(u)-\Pi\right]du.$$

We shall call this matrix the *transient deviation matrix*. Indeed, as $t\to\infty$, $D(t)\to D$.

Lemma 1 *If the matrix D has finite entries, then the matrices D and $D(t)$ are related via the equation*

$$D(t) = \left[I-e^{Qt}\right]D. \tag{3.11}$$

Proof. By definition of D, we have

$$\begin{aligned}
\left[I-e^{Qt}\right]D &= \int_0^\infty \left[e^{Qu}-\Pi\right]du - e^{Qt}\int_0^\infty \left[e^{Qu}-\Pi\right]du. \\
&= \int_0^\infty \left[e^{Qu}-\Pi\right]du - \int_0^\infty \left[e^{Q(t+u)}-\Pi\right]du \\
&= \int_0^\infty \left[e^{Qu}-\Pi\right]du - \int_t^\infty \left[e^{Qv}-\Pi\right]dv \\
&= D(t)
\end{aligned}$$

All integrals are finite by the assumed existence of D. ■

As a consequence of Lemma 1 and (3.4), we can write $D(t)$ explicitly in terms of the generator Q via the expression

$$D(t) = \left[I - e^{Qt}\right](\Pi - Q)^{-1}. \tag{3.12}$$

We can also write $D(t)$ in terms of Q via an expression that does not involve a matrix inverse, as described in the next lemma.

Lemma 2 *The transient deviation matrix can be expressed in the form*

$$D(t) = [I_m, 0_m] \exp\left(\begin{bmatrix} Q & I_m \\ 0 & 0 \end{bmatrix} t\right) [0_m, I_m]^\top + \mathbf{1}\boldsymbol{\pi} t, \tag{3.13}$$

where the identity and zero matrices I_m and 0_m are of the same size as D, and the zero entries in the matrix exponential are scalar.

Proof. We use a particular case of Theorem 1 in [3] which states that the integral

$$\int_0^t e^{Au} B e^{C(t-u)} du, \tag{3.14}$$

where A, B and C are matrices of appropriate size, can be evaluated as the upper right corner of the matrix eponential $\exp(Mt)$ where

$$M = \begin{bmatrix} A & B \\ 0 & C \end{bmatrix}.$$

Here we have

$$D(t) = \int_0^t e^{Qu} du + \mathbf{1}\boldsymbol{\pi} t,$$

and we use the above result with $A = Q$, $B = I_m$, and $C = 0$. ■

It is easy to show from (3.11) that the transient deviation matrix satisfies the properties

$$\begin{aligned} D(t)\mathbf{1} &= \mathbf{0} \\ \boldsymbol{\pi} D(t) &= \mathbf{0} \\ D(t)(-Q) &= (-Q)D(t) = I - e^{Qt} \\ D'(t) &= e^{Qt} - \Pi \\ \int_0^t D(u)\, du &= [tI - D(t)]D. \end{aligned}$$

Furthermore, we can express the transient deviation matrix as the Drazin inverse of a particular matrix, as shown in the next lemma.

Lemma 3 *The matrices $D(t)$ and*

$$W(t) := -Q\left(I - e^{Qt} + \Pi\right)^{-1}$$

are Drazin inverses of each other.

Proof. We use the facts that the Drazin inverse of $-Q$ is D, the Drazin inverse of $(I - e^{Qt})$ is $\left(I - e^{Qt} + \Pi\right)^{-1} - \Pi$, and the Drazin inverse of the Drazin inverse of a matrix X is X itself (by the symmetry of (3.7)-(3.9) in $X = -Q$ and D). ■

An explicit expression for the entries of the transient deviation matrix can also be obtained in terms of the expected first passage times m_{ij}: using (3.11) and (3.3),

$$\begin{aligned} D_{i,j}(t) &= D_{ij} - \sum_k [\exp(Qt)]_{i,k}\, D_{k,j} \\ &= \pi_j\,(m^e_j - m_{i,j}) - \sum_k [\exp(Qt)]_{i,k}\,\pi_j\,(m^e_j - m_{k,j}) \qquad (3.15) \\ &= \pi_j\,(\sum_k [\exp(Qt)]_{i,k}\, m_{k,j} - m_{i,j}). \qquad (3.16) \end{aligned}$$

Note that since D and $\exp(Qt)$ commute in (3.11), the entries of $D(t)$ can equivalently be written as

$$D_{i,j}(t) = \pi_j\,(\sum_k m_{i,k}[\exp(Qt)]_{k,j} - m_{i,j}). \qquad (3.17)$$

It is not easy to obtain an explicit expression for the entries of $\exp(Qt)$, even in the simple case of the M/M/1/C queue. This is, however, possible in the Laplace transform domain, as we now show. For complex s with $\Re(s) > 0$, let

$$\tilde{D}(s) = \int_0^\infty D(t) e^{-st} dt$$

be the Laplace transform of the transient deviation matrix. Since the Laplace transform of $\exp(Qt)$ is given by $\Phi(s) := (sI - Q)^{-1}$, using (3.11) and some algebraic manipulations, we obtain

$$\tilde{D}_{i,j}(s) = s^{-1}[\Phi_{i,j}(s) - s^{-1}\pi_j]. \qquad (3.18)$$

Alternatively, from (3.16) and (3.17),

$$\begin{aligned} \tilde{D}_{i,j}(s) &= \pi_j\,(\sum_k \Phi_{i,k}(s)\, m_{k,j} - m_{i,j} s^{-1}) \\ &= \pi_j\,(\sum_k m_{i,k}\,\Phi_{k,j}(s) - m_{i,j} s^{-1}). \end{aligned}$$

In the M/M/1/C case, the entries of $\Phi(s)$ can be obtained explicitly using the results in Huang and McColl [8] on the analytical inversion of tridiagonal matrices:

$$\begin{aligned}
\Phi_{0,0}(s) &= (s+\lambda-\lambda\mu\, y_3/y_2)^{-1} \\
\Phi_{j,j}(s) &= \left[s+\lambda+\mu-\lambda\mu\,(z_{j-1}/z_j + y_{j+3}/y_{j+2})\right]^{-1}, \quad 1 \le j \le C-1 \\
\Phi_{C,C}(s) &= (s+\mu-\lambda\mu\, z_{C-1}/z_C)^{-1} \\
\Phi_{i,j}(s) &= (\mathbb{1}_{\{i<j\}}\lambda^{j-i} z_i/z_j + \mathbb{1}_{\{i>j\}}\mu^{i-j} y_{i+2}/y_{j+2})\Phi_{jj}(s), \quad 0 \le i \ne j \le C,
\end{aligned}$$

where

$$\begin{aligned}
y_i &= \frac{[y_-(s+\mu)-1]}{y_+^{C+1}(y_- - y_+)} y_+^i + \frac{[1-y_+(s+\mu)]}{y_-^{C+1}(y_- - y_+)} y_-^i, \quad 2 \le i \le C+2, \\
z_i &= \frac{(s+\lambda-z_-)}{(z_+ - z_-)}(z_+^i - z_-^i) + z_-^i, \quad 0 \le i \le C,
\end{aligned}$$

with

$$y_\pm = \frac{(s+\lambda+\mu) \pm \sqrt{(s+\lambda+\mu)^2 - 4\mu\lambda}}{2\lambda\mu} \quad \text{and} \quad z_\pm = \lambda\mu\, y_\pm.$$

Observe that taking the Laplace transform of (3.10) and using (2.9) provides another way to obtain an explicit expression for the entries of the last column of $\tilde{D}(s)$ in the M/M/1/C queue:

$$\begin{aligned}
\tilde{D}_{i,C}(s) &= (\lambda\theta)^{-1}\tilde{R}_{i,C}(s) - \pi_C \int_0^\infty t e^{-st} dt \\
&= (\lambda\theta)^{-1}[A(s) r_1^i(s) + B(s) r_2^i(s)] - \pi_C s^{-2},
\end{aligned}$$

where $A(s)$, $r_{1,2}(s)$ and $B(s)$ are given in (2.10)-(2.12). Finally, note that the other columns of the deviation matrix could be obtained in a similar way by generalising the argument developed in Section 2 to the case where revenue is lost when the queue length is different from C.

In the next section, we will use coupling methods and develop a completely different approach to compute explicit expressions for the elements of the last column of D and $D(t)$ in an M/M/1/C queue.

4 Coupling in the $M/M/1/C$ queue

In the rest of the paper, we denote by $Q_{n,C}(t)$ the queue length of an $M/M/1/C$ queue at time t, given that it starts with n customers at time

0, with arrivals and potential services generated by the Poisson processes $\{\mathcal{A}(t)\}$ and $\{\mathcal{S}(t)\}$ with rates λ and μ, respectively. The number of customers $Q_{n,C}(t)$ can be written explicitly in terms of these processes via the expression

$$Q_{n,C}(t) = n + (\mathcal{A}(t) - U_{n,C}(t)) - (\mathcal{S}(t) - L_{n,C}(t)), \tag{4.1}$$

where the *lower regulating process* $L_{n,C}(t)$ counts the number of potential services in $[0,t]$ which occur when the system is empty, and the *upper regulating process* $U_{n,C}(t)$ counts the number of arrivals which occur in $[0,t]$ when the system is at full capacity, given that the initial queue length is n. We see that $U_{n,C}(t)$ gives the number of customers rejected from the queue, and so

$$R_{n,C}(t) = \theta \mathbb{E}\left(U_{n,C}(t)\right). \tag{4.2}$$

This means,

$$R_{n+1,C}(t) - R_{n,C}(t) = \theta \mathbb{E}(U_{n+1,C}(t)) - \theta \mathbb{E}(U_{n,C}(t)). \tag{4.3}$$

When $U_{n+1,C}(t)$ and $U_{n,C}(t)$ are defined together on a specific probability space, Equation (4.3) can be simplified, as shown in the next two subsections. Expressions for $R_{n+1,C}(t) - R_{n,C}(t)$ and $\lim_{t\to\infty}\left[R_{n+1,C}(t) - R_{n,C}(t)\right]$ can then be used to calculate the final column of the transient deviation and deviation matrices, respectively, as we demonstrate in a third subsection.

4.1 The coupling

Suppose we have two $M/M/1/C$ queueing systems with same values of C, λ, θ and μ, but different initial queue lengths $n, n+1 \in \{0,1,...,C\}$. Considering them to be defined on different probability spaces, we denote the corresponding random variables at time t by

$$(\mathcal{A}_n(t), \mathcal{S}_n(t), Q_{n,C}(t), U_{n,C}(t), L_{n,C}(t)) \tag{4.4}$$

and

$$(\mathcal{A}_{n+1}(t), \mathcal{S}_{n+1}(t), Q_{n+1,C}(t), U_{n+1,C}(t), L_{n+1,C}(t)). \tag{4.5}$$

Note that we added the subsript n or $n+1$ to the arrival and potential service processes in order to differentiate the probability space on which they live. Since the arrival and service processes of the two systems have the same distribution, we know that

$$(\{\mathcal{A}_n(t)\}, \{\mathcal{S}_n(t)\}) =_d (\{\mathcal{A}_{n+1}(t)\}, \{\mathcal{S}_{n+1}(t)\}). \tag{4.6}$$

Now define two new queueing systems on the same probability space

$$(\hat{\mathcal{A}}(t), \hat{\mathcal{S}}(t), \hat{Q}_{n,C}(t), \hat{U}_{n,C}(t), \hat{L}_{n,C}(t), \hat{Q}_{n+1,C}(t), \hat{U}_{n+1,C}(t), \hat{L}_{n+1,C}(t)) \quad (4.7)$$

such that

$$(\{\hat{\mathcal{A}}(t)\}, \{\hat{\mathcal{S}}(t)\}) =_d (\{\mathcal{A}_n(t)\}, \{\mathcal{S}_n(t)\}) =_d (\{\mathcal{A}_{n+1}(t)\}, \{\mathcal{S}_{n+1}(t)\}). \quad (4.8)$$

Since the queue length process and the upper and lower regulating processes are functions of $\mathcal{A}_{(.)}(t)$, $\mathcal{S}_{(.)}(t)$, C and the initial queue length n or $n+1$, Equation (4.8) implies,

$$(\mathcal{A}_n(t), \mathcal{S}_n(t), Q_{n,C}(t), U_{n,C}(t), L_{n,C}(t)) =_d (\hat{\mathcal{A}}(t), \hat{\mathcal{S}}(t), \hat{Q}_{n,C}(t), \hat{U}_{n,C}(t), \hat{L}_{n,C}(t)), \quad (4.9)$$

with the same equality also holding for system $n+1$. This means,

$$R_{n+1,C}(t) - R_{n,C}(t) = \theta \mathbb{E}(\hat{U}_{n+1,C}(t) - \hat{U}_{n,C}(t)), \quad (4.10)$$

with both queueing systems generated by the same arrival and potential service processes. The purpose of defining the queueing systems in this way is to induce a *march coupling* (see [4]) on $(\hat{Q}_{n+1,C}(t), \hat{Q}_{n,C}(t)-)$. The term 'march' is made in reference to the tendency of $\hat{Q}_{n+1,C}(t)$ and $\hat{Q}_{n,C}(t)$ to move together. This can be understood by inspecting the quasi-birth-and-death process $\{(\hat{Q}_{n,C}(t), \hat{Q}_{n+1,C}(t) - \hat{Q}_{n,C}(t)) : t \geq 0\}$ on the state space $\{(k,l) : k = 0, 1, \ldots, C-1, l = 0, 1\} \cup \{(C, 0)\}$, which has initial state $(n, 1)$ and the following transition rates: for $k = 0$,

$$(0,0) \to (1,0) \quad \text{at rate } \lambda \quad (4.11)$$

$$(0,1) \to \begin{cases} (1,1), & \text{at rate } \lambda \\ (0,0), & \text{at rate } \mu; \end{cases} \quad (4.12)$$

for $0 < k < C-1$,

$$(k,0) \to \begin{cases} (k+1,0), & \text{at rate } \lambda \\ (k-1,0), & \text{at rate } \mu \end{cases} \quad (4.13)$$

$$(k,1) \to \begin{cases} (k+1,1), & \text{at rate } \lambda \\ (k-1,1), & \text{at rate } \mu; \end{cases} \quad (4.14)$$

for $k = C-1$,

$$(C-1,0) \to \begin{cases} (C,0), & \text{at rate } \lambda \\ (C-2,0), & \text{at rate } \mu \end{cases} \quad (4.15)$$

$$(C-1,1) \to \begin{cases} (C,0), & \text{at rate } \lambda \\ (C-2,1), & \text{at rate } \mu; \end{cases} \tag{4.16}$$

and for $k = C$,

$$(C,0) \to (C-1,0) \quad \text{at rate } \mu. \tag{4.17}$$

On inspection of these rates we can see that phase $l = 0$ is absorbing. This means that once equal, the queue length processes remain equal forever. There are two ways in which the queue length processes can become equal:

- First, a potential service can occur when $(\hat{Q}_{n,C}(t), \hat{Q}_{n+1,C}(t) - \hat{Q}_{n,C}(t)) = (0,1)$ resulting in a served customer in system $n+1$ and a 'wasted' service in system n, which causes $\hat{L}_{n,C}(t)$ to increase by 1 and both queueing systems to become empty.
- Alternatively, an arrival can occur when $(\hat{Q}_{n,C}(t), \hat{Q}_{n+1,C}(t) - \hat{Q}_{n,C}(t)) = (C-1,1)$. In this case the arrival is accepted to system n but is rejected from system $n+1$, which causes $\hat{U}_{n+1,C}(t)$ to increase by 1 and both queues to become full.

Both possibilities are illustrated in Figure 4.1.

Let $T = \inf\{t \geq 0 : \hat{Q}_{n+1,C}(t) - \hat{Q}_{n,C}(t) = 0\}$ be the coupling time. Note that T occurs at the time of the first 'wasted' service in system n or the first 'rejected' customer in system $n+1$. Before T, no customers are lost by either system and after T, customers are lost from both systems simultaneously. This means the only time at which $\{\hat{U}_{n+1,C}(t) - \hat{U}_{n,C}(t)\}$ can increment is T. If T occurs at the time of a 'wasted' service ($\hat{Q}_{n+1,C}(T) = 0$), neither system loses a customer at T and $\hat{U}_{n+1,C}(t) - \hat{U}_{n,C}(t) = 0$ for all $t > T$, but if T occurs at the time of a 'lost' customer ($\hat{Q}_{n,C}(T) = C$), system $n+1$ loses a customer and system n does not, which means $\hat{U}_{n+1,C}(t) - \hat{U}_{n,C}(t) = 1$ for all $t > T$. Thus,

$$\hat{U}_{n+1,C}(t) - \hat{U}_{n,C}(t) = \mathbb{1}_{\{t>T\}} \mathbb{1}_{\{\hat{Q}_{n,C}(T)=C\}}, \tag{4.18}$$

which gives,

$$R_{n+1,C}(t) - R_{n,C}(t) = \theta \mathbb{E}(\mathbb{1}_{\{t>T\}} \mathbb{1}_{\{\hat{Q}_{n,C}(T)=C\}}). \tag{4.19}$$

4.2 The effect of an additional customer at time 0

Let $\Delta_{n+1,C}(t) = R_{n+1,C}(t) - R_{n,C}(t)$, and $\Delta_{n+1,C}(t|x)$ be the same quantity conditional on the first arrival or potential service occurring at time x. Using

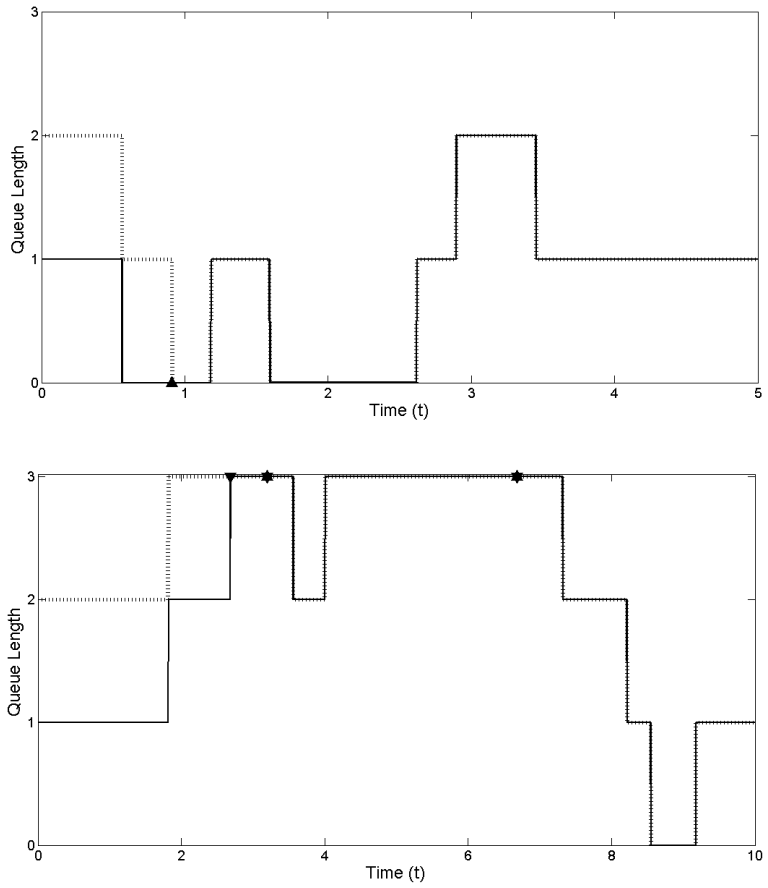

Figure 4.1: Possible realization of $\hat{Q}_{1,3}(t)$ (solid line) and $\hat{Q}_{2,3}(t)$ (bold dashed line). In the upper graph the queue length processes couple after an additional service is missed when there is one customer initially and in the lower graph the queue length processes couple after an additional customer is lost when there are two customers initially.

transition rates (4.11)-(4.17) and Equation (4.19), we see that

$$\Delta_{n+1,C}(t|x) = \begin{cases} 0, & n = 0, t \geq x \\ 0, & 0 \leq n \leq C, t < x \\ \frac{\mu}{\lambda+\mu}\Delta_{n,C}(t-x) + \frac{\lambda}{\lambda+\mu}\Delta_{n+2,C}(t-x), & 0 < n \leq C, t \geq x \\ \theta, & n = C+1. \end{cases} \tag{4.20}$$

Let $F(\cdot)$ denote the distribution of the time until the first arrival or potential service. Then,

$$\Delta_{n,C}(t) = \int_0^\infty \Delta_{n,C}(t|x)dF(x), \tag{4.21}$$

where $F(\cdot)$ is exponential with parameter $\lambda+\mu$. The Laplace transform of $\Delta_{n+1,C}(t)$, denoted by $\tilde{\Delta}_{n+1,C}(s)$, can be found using a method similar to that of Section 2. This has the form

$$\tilde{\Delta}_{n+1,C}(s) = A(s)r_1^{n+1}(s) + B(s)r_2^{n+1}(s), \tag{4.22}$$

where,

$$r_{1,2}(s) = \frac{\lambda+\mu+s \pm \sqrt{(\lambda+\mu+s)^2 - 4\mu\lambda}}{2\lambda}, \tag{4.23}$$

$$A(s) = \frac{\theta}{s(r_1^{C+1}(s) - r_2^{C+1}(s)}, \tag{4.24}$$

and $B(s) = -A(s)$.

In Figures 2.1 and 2.2 we noted that as t increases $R_{n+1,C}(t) - R_{n,C}(t)$ stabilized at a constant value. Using a similar method an analytic expression for $\lim_{t\to\infty}[R_{n+1,C}(t) - R_{n,C}(t)]$ can be derived, as we show now.

Theorem 2 *If $R_{n+1,C}(t)$ and $R_{n,C}(t)$ are calculated using common arrival and service rates, λ and μ, and common revenue lost per customer, θ, then,*

$$\lim_{t\to\infty}[R_{n+1,C}(t) - R_{n,C}(t)] = \begin{cases} \theta\dfrac{1-(\mu/\lambda)^{n+1}}{1-(\mu/\lambda)^{C+1}}, & \lambda \neq \mu \\ \theta\dfrac{n+1}{C+1}, & \lambda = \mu, \end{cases} \tag{4.25}$$

for any $C \in \mathbb{N}$ and $n \in \{0,1,...,C-1\}$.

Proof. We have

$$R_{n+1,C}(t) - R_{n,C}(t) \;=\; \theta\mathbb{E}(\mathbb{1}_{\{t>T\}}\mathbb{1}_{\{\hat{Q}_{n,C}(T)=C\}}). \tag{4.26}$$

Define the *free process* starting in state $n+1$ as $\hat{X}_{n+1}(t) = (n+1)+\hat{A}(t)-\hat{S}(t)$, and note that T is a stopping time with $\mathbb{E}(T) < \infty$. Then

$$\theta\mathbb{E}(\mathbb{1}_{\{t>T\}}\mathbb{1}_{\{\hat{Q}_{n,C}(T)=C\}}) = \theta\mathbb{P}(t \geq T)\mathbb{P}(\hat{X}_{n+1}(T) = C+1 | t \geq T) \tag{4.27}$$

$$\to \theta\mathbb{P}(\hat{X}_{n+1}(T) = C+1) \quad \text{as } t \to \infty. \tag{4.28}$$

The probability that an increment of $\{\hat{X}_{n+1}(t)\}$ corresponds to an arrival is $p = \lambda/(\lambda+\mu)$ and the probability that an increment of $\{\hat{X}_{n+1}(t)\}$ corresponds to a service is $1-p = \mu/(\lambda+\mu)$. As t becomes large, we can consider the discrete time counterpart of $\{\hat{X}_{n+1}(t)\}$ and note that $\mathbb{P}(\hat{X}_{n+1}(T) = C)$ can

be found by solving an appropriate gambler's ruin problem (see for instance [2]) which has a well known solution leading to

$$\theta\mathbb{P}(\hat{X}_{n+1}(T) = C+1) \;=\; \begin{cases} \theta\left[1 - \dfrac{\left(\frac{1-p}{p}\right)^{n+1} - \left(\frac{1-p}{p}\right)^{C+1}}{1-\left(\frac{1-p}{p}\right)^{C+1}}\right], & \lambda \neq \mu \\ \theta\dfrac{n+1}{C+1}, & \lambda = \mu. \end{cases} \tag{4.29}$$

When $\lambda \neq \mu$, Equation (4.29) can be simplified to give the desired result. ■

4.3 The transient deviation and deviation matrices

In the previous subsection, we obtained explicit expressions for $\tilde{\Delta}_{n+1,C}(s)$ and $\Delta_{n+1,C} := \lim_{t\to\infty}[R_{n+1,C}(t) - R_{n,C}(t)]$. We now show how these expressions can be used to calculate the last column of the transient deviation and the deviation matrix, that is, $D_{n,C}(t)$ and $D_{n,C}$ respectively.

Recall that when the $M/M/1/C$ queue reaches its stationary distribution revenue is lost at a constant rate given by $\theta\lambda\pi_C t$. This means that if the initial queue length follows the stationary distribution, then the lost revenue is given by $\theta\lambda\pi_C t$, that is,

$$\sum_{k=0}^{C} \pi_k R_{k,C}(t) = \lambda\theta\pi_C t. \tag{4.30}$$

When $k < n$ an alternative expression for $R_{k,C}(t)$ is,

$$R_{k,C}(t) = R_{n,C}(t) - \sum_{i=k+1}^{n} \Delta_{i,C}(t), \tag{4.31}$$

and similarly when $k > n$ we have,

$$R_{k,C}(t) = R_{n,C}(t) + \sum_{i=n+1}^{k} \Delta_{i,C}(t). \tag{4.32}$$

From Equations (4.30), (4.31) and (4.32) we get,

$$\theta\lambda D_{n,C}(t) = R_{n,C}(t) - \lambda\theta\pi_C t \tag{4.33}$$

$$= \sum_{k=0}^{n-1} \pi_k \sum_{i=k+1}^{n} \Delta_{i,C}(t) - \sum_{k=n+1}^{C} \pi_k \sum_{i=n+1}^{k} \Delta_{i,C}(t). \tag{4.34}$$

This implies,

$$\theta\lambda D_{n,C} = \lim_{t\to\infty}\left[R_{n,C}(t) - \lambda\theta\pi_C t\right] \tag{4.35}$$

$$= \lim_{t\to\infty}\left[\sum_{k=0}^{n-1}\pi_k\sum_{i=k}^{n-1}\Delta_{i,C}(t) - \sum_{k=n+1}^{C}\pi_k\sum_{i=n+1}^{k}\Delta_{i,C}(t)\right] \tag{4.36}$$

$$= \sum_{k=0}^{n-1}\pi_k\sum_{i=k+1}^{n}\Delta_{i,C} - \sum_{k=n+1}^{C}\pi_k\sum_{i=n+1}^{k}\Delta_{i,C}. \tag{4.37}$$

Equations (4.34) and (4.37) give the values of the final column of the transient deviation and deviation matrices, respectively.

5 Buying and selling prices

The use of coupling gave us another perspective on the effect of one additional initial customer on the expected loss function, and enabled us to derive an alternative method to calculate the final column of the transient deviation and deviation matrices. Insight gained from coupling can also be used to understand a number of puzzling results, as we show in this section.

5.1 Defining the buying and selling prices

We now return to the capacity planning example discussed in Section 2. Having determined the expected lost revenue in the time interval $[0, t)$ given an initial queue length n and capacity C, the manager must then be able to convert these results into the price at which extra capacity should be bought or sold. Suppose the manager has a planning horizon of a single period with length t. The manager should then purchase an additional unit of capacity only when the reduction in the expected lost revenue over $[0, t)$ is greater than the cost required to purchase the unit of capacity. We therefore let the buying price, $B_{n,C}(t)$, be

$$B_{n,C}(t) = R_{n,C}(t) - R_{n,C+1}(t) \text{ for all } n \in \{0, ..., C\}, \tag{5.1}$$

and for similar reasoning we let the selling price $S_{n,C}(t)$ be

$$S_{n,C}(t) \;=\; \begin{cases} R_{n,C-1}(t) - R_{n,C}(t), & n \in \{0, ..., C-1\} \\ R_{C-1,C-1}(t) - R_{C,C}(t) + \theta, & n = C. \end{cases} \tag{5.2}$$

Note that when $n = C$, we assume that selling a unit of capacity causes a single customer to be lost immediately. This incurs a penalty, which we have taken to be equal to θ in Equation (5.2). It is arguable that expelling a customer who is already present has a greater detrimental consequence, say in terms of customer goodwill, than refusing entry to a customer. If we wanted to model such a consideration, we could do so by incorporating a penalty greater than θ in Equation (5.2).

Example: the $M/M/1/5$ loss system

For a 'low blocking' system with $C = 5$, $\theta = 1$, $\lambda = 3$, and $\mu = 5$ the buying and selling prices, $B_{n,C}(t)$ (dotted lines) and $S_{n,C}(t)$ (continuous lines) for $n = 3, 4, 5$ are displayed in the upper graph of Figure 5.1. The same results for a 'high blocking' system with $C = 5$, $\theta = 1$, $\lambda = 5$, and $\mu = 3$ are given in the lower graph of Figure 5.1.

One feature, which is of particular interest to us, is that the selling prices in the high blocking and low blocking systems converge to the same gradient, with the same also true for the buying prices. Furthermore, when $n = C$ the selling prices for the high blocking and low blocking systems are identical for all $t \geq 0$. These are counter-intuitive results, since it is reasonable to expect that the manager of the high blocking system would place a higher value on capacity in both the short and long term.

The convergence to the same asymptotic gradient can be understood by considering the difference in the asymptotic gradient of the expected lost revenue functions. Using simple algebraic arguments it can be shown that

$$\lambda(\pi_C(\rho) - \pi_{C+1}(\rho)) \;=\; \mu(\pi_C(\rho^{-1}) - \pi_{C+1}(\rho^{-1}))$$

or equivalently,

$$\lambda\left[\frac{(1-\rho)\rho^C}{1-\rho^{C+1}} - \frac{(1-\rho)\rho^{C+1}}{1-\rho^{C+2}}\right] \;=\; \mu\left[\frac{(1-\rho^{-1})\rho^{-C}}{1-\rho^{-(C+1)}} - \frac{(1-\rho^{-1})\rho^{-(C+1)}}{1-\rho^{-(C+2)}}\right].$$

However, it is initially unclear why the selling prices for the high blocking and low blocking systems are identical when $n = C$. It turns out that this is a more general phenomenon.

Theorem 3 *Consider two $M/M/1/C$ queues with identical values of C and θ, a 'high blocking' queue with parameters λ^{HB}, μ^{HB} and a 'low blocking' queue with parameters λ^{LB}, μ^{LB}, which are such that $\lambda^{HB} > \lambda^{LB}$, $\lambda^{HB} = \mu^{LB}$*

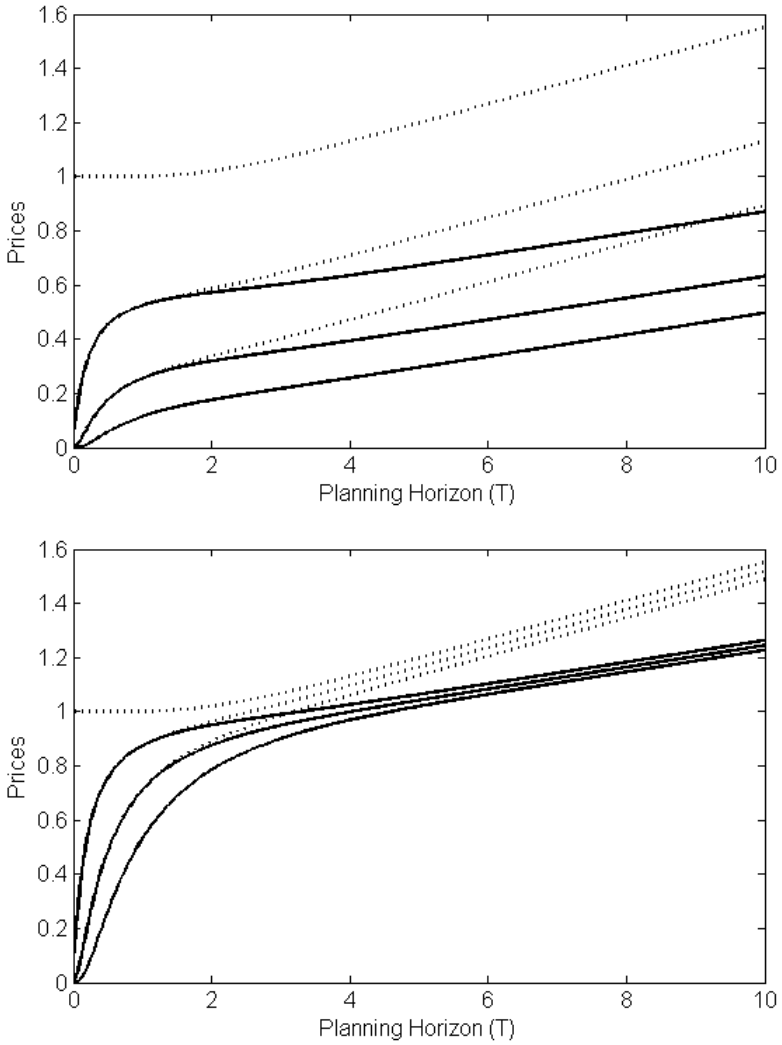

Figure 5.1: Top: Buying (solid) and selling (dotted) price functions for $n = 3, 4, 5$ when $C = 5$, $\lambda = 3$ and $\mu = 5$ (low blocking system). Bottom: Buying (solid) and selling (dotted) price functions for $n = 3, 4, 5$ when $C = 5$, $\lambda = 5$ and $\mu = 3$ (high blocking system).

and $\mu^{HB} = \lambda^{LB}$. *If* $S^{HB}_{n,C}(t)$ *is the selling price of capacity in the high blocking queue and* $S^{LB}_{n,C}(t)$ *is the selling price of capacity in the low blocking queue, then*

$$S^{HB}_{C,C}(t) = S^{LB}_{C,C}(t) \text{ for all } t \geq 0. \tag{5.3}$$

This result is proved in the next subsection in which we first consider general identities for the selling price.

5.2 The coupling

Let $U_{n,C-1}(t)$ and $U_{n,C}(t)$ count the number of customers rejected from an $M/M/1/C-1$ and an $M/M/1/C$ queue respectively, with common values of θ, n, λ and μ and no defined dependence. We use the convention that if there are initially C customers in the $M/M/1/C-1$ queue, then one customer is lost instantaneously, so that

$$U_{C,C-1}(t) = 1 + U_{C-1,C-1}(t). \tag{5.4}$$

Using the definition (5.2) of the selling price we have,

$$\begin{aligned} S_{n,C}(t) &= R_{n,C-1}(t) - R_{n,C}(t) && (5.5)\\ &= \theta\mathbb{E}(U_{n,C-1}(t)) - \theta\mathbb{E}(U_{n,C}(t)). && (5.6) \end{aligned}$$

The state of the two queueing systems $M/M/1/C-1$ and $M/M/1/C$ at time t is respectively given by

$$(\mathcal{A}_{C-1}(t), \mathcal{S}_{C-1}(t), Q_{n,C-1}(t), U_{n,C-1}(t), L_{n,C-1}(t)) \tag{5.7}$$

and

$$(\mathcal{A}_{C}(t), \mathcal{S}_{C}(t), Q_{n,C}(t), U_{n,C}(t), L_{n,C}(t)) \,. \tag{5.8}$$

As in Section 4, we again define two new queueing systems on the same probability space

$$(\hat{\mathcal{A}}(t), \hat{\mathcal{S}}(t), \hat{Q}_{n,C-1}(t), \hat{U}_{n,C-1}(t), \hat{L}_{n,C-1}(t), \hat{Q}_{n,C}(t), \hat{U}_{n,C}(t), \hat{L}_{n,C}(t)) \tag{5.9}$$

such that

$$(\{\hat{\mathcal{A}}(t)\}, \{\hat{\mathcal{S}}(t)\} =_d (\{\mathcal{A}_{C-1}(t)\}, \{\mathcal{S}_{C-1}(t)\}) =_d (\{A_C(t)\}, \{\mathcal{S}_C(t)\}) \,. \tag{5.10}$$

We refer to these queueing systems as system C and system $C-1$, respectively. Condition (5.10) ensures that

$$(\mathcal{A}_C(t), \mathcal{S}_C(t), Q_{n,C}(t), U_{n,C}(t), L_{n,C}(t)) =_d (\hat{\mathcal{A}}(t), \hat{\mathcal{S}}(t), \hat{Q}_{n,C}(t), \hat{U}_{n,C}(t), \hat{L}_{n,C}(t)), \tag{5.11}$$

with the same equality also holding for system $C-1$. This implies,

$$S_{n,C}(t) = \theta\mathbb{E}(\hat{U}_{n,C-1}(t) - \hat{U}_{n,C}(t)). \tag{5.12}$$

By allowing both queues to be generated by common arrival and service processes we again induce a march coupling on the queue length processes, $(\hat{Q}_{n,C-1}(t), \hat{Q}_{n,C}(t))$. The quasi-birth and death process

$\{(\hat{Q}_{n,C-1}(t), \hat{Q}_{n,C}(t) - \hat{Q}_{n,C-1}(t)) : t \geq 0\}$ on the state space $\{(k,l) : k = 0, 1, \ldots, C-1, l = 0, 1\}$ has the following transition rates: for $k = 0$,

$$(0,0) \to (1,0) \text{ at rate } \lambda \tag{5.13}$$

$$(0,1) \to \begin{cases} (1,1), & \text{at rate } \lambda \\ (0,0), & \text{at rate } \mu; \end{cases} \tag{5.14}$$

for $0 < k < C-1$,

$$(k,0) \to \begin{cases} (k+1,0), & \text{at rate } \lambda \\ (k-1,0), & \text{at rate } \mu \end{cases} \tag{5.15}$$

$$(k,1) \to \begin{cases} (k+1,1), & \text{at rate } \lambda \\ (k-1,1), & \text{at rate } \mu; \end{cases} \tag{5.16}$$

and for $k = C-1$,

$$(C-1,0) \to \begin{cases} (C-1,1), & \text{at rate } \lambda \\ (C-2,0), & \text{at rate } \mu \end{cases} \tag{5.17}$$

$$(C-1,1) \to (C-2,1) \text{ at rate } \mu. \tag{5.18}$$

With reference to Figure 5.2, we see that for all $t \geq 0$, either $\hat{Q}_{n,C}(t) - \hat{Q}_{n,C-1}(t) = 0$ or $\hat{Q}_{n,C}(t) - \hat{Q}_{n,C-1}(t) = 1$. When $\hat{Q}_{n,C}(t) - \hat{Q}_{n,C-1}(t) = 1$:

- both systems lose customers at precisely the same times (for example, at $t = 2.4, 18.9$ in Figure 5.2) which means during these periods $\{\hat{U}_{n,C-1}(t) - \hat{U}_{n,C}(t)\}$ remains constant, and
- only system $C-1$ can 'waste' a potential service. Once this occurs (for example, at $t = 3.9$ in Figure 5.2), the queue length processes become equal.

When $\hat{Q}_{n,C}(t) - \hat{Q}_{n,C-1}(t) = 0$:

- both systems 'waste' services at precisely the same times (for example, at $t = 4.5, 8.5, 9.5, 10, 10.7$ in Figure 5.2), but
- only system $C-1$ can lose a customer and, once this occurs (for example at $t = 1.6, 16.6$ in Figure 5.2), the queue length processes become unequal, resulting in an additional customer lost by system $C-1$. This increases $\{\hat{U}_{n,C-1}(t) - \hat{U}_{n,C}(t)\}$ by one.

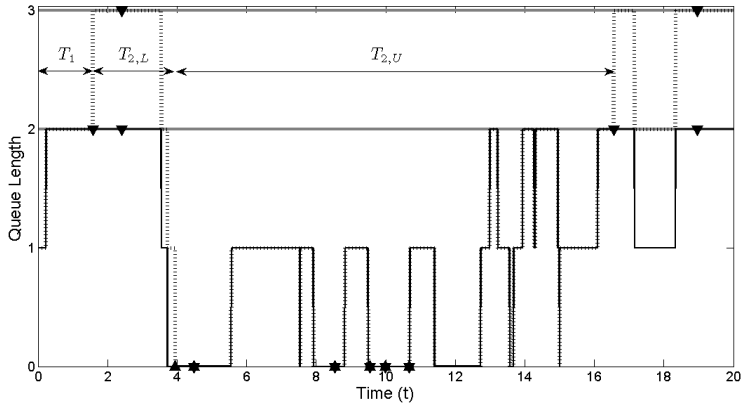

Figure 5.2: A possible realization of $\hat{Q}_{1,2}(t)$ (solid line) and $\hat{Q}_{1,3}(t)$ (dashed line). Stars represent simulatenous missed services in both systems, upward pointing triangles represent missed services in system $C-1$ and downward pointing triangles represent lost customers. Additional customers are lost at times 1.6 and 16.6 and an additional service is missed at time 3.9.

Since systems C and $C-1$ have the same initial queue length n we have, $\hat{Q}_{n,C}(0) - \hat{Q}_{n,C-1}(0) = 0$. Let T_1 be the first time that $\hat{Q}_{n,C}(0) - \hat{Q}_{n,C-1}(0) = 1$. This is also the time that system $C-1$ loses its first customer ($T_1 = 1.6$ in Figure 5.2). At this time system C does not lose a customer and therefore

$$T_1 = \inf\left\{t \geq 0 : \hat{U}_{n,C-1}(t) - \hat{U}_{n,C}(t) = 1\right\}. \tag{5.19}$$

After T_1, further losses of customers from system $C-1$ which are not matched by losses of customers from system C occur according to a renewal process where the ith renewal of $\{\hat{U}_{n,C-1}(t) - \hat{U}_{n,C}(t)\}$ can be viewed as the $(i+1)$st additional lost customer. To see this, let the ith renewal occur at time t, where $\hat{Q}_{n,C}(t) = C$ and $\hat{Q}_{n,C-1}(t) = C-1$. For another additional customer to be lost two events must take place:

- First, system $C-1$ must miss an additional service after initially being full. Let this occur at time $t+T_{i,L}$. At this time both queues are empty. In particular $T_{i,L}$ is the time taken for system C to become empty after initially being full.
- Second, system $C-1$ must lose an additional customer. Let this occur at time $t+T_{i,L}+T_{i,U}$ and at this time both queues are full. In particular, $T_{i,U}$ is the time taken for system C to become full after initially being empty.

Further renewals then continue to be generated by this sequence of events. We now have the base to prove Theorem 3.

Proof of Theorem 3. Using the definitions of T_1, $T_{i,L}$ and $T_{i,U}$ we have,

$$\hat{U}_{n,C-1}(t) - \hat{U}_{n,C}(t) = \sup\{m : \sum_{k=1}^{m} [1_{k=1} T_1 + 1_{k\geq 2} (T_{k,U} + T_{k,L})] \leq t\}. \quad (5.20)$$

Note that since $\lambda^{HB} = \mu^{LB}$ and $\mu^{HB} = \lambda^{LB}$ we have,

$$T_{i,U}^{HB} =_d T_{i,L}^{LB} \text{ for all } i \in \mathbb{N} \quad (5.21)$$

and

$$T_{i,U}^{HB} =_d T_{i,L}^{LB} \text{ for all } i \in \mathbb{N}. \quad (5.22)$$

This is because when the arrival and service rates are swapped the distribution of time taken for system C to become full after starting empty becomes the distribution of time taken from system C to become empty after starting full, and vice versa.

For a general initial queue length n, we do not have $T_1^{HB} =_d T_1^{LB}$ and thus, we do not have $\hat{U}_{n,C-1}^{HB}(t) - \hat{U}_{n,C}^{HB}(t) =_d \hat{U}_{n,C-1}^{LB}(t) - \hat{U}_{n,C}^{LB}(t)$. However, when $n = C$ and a unit of capacity is sold, a single customer is lost immediately. Hence,

$$T_1^{HB} = T_1^{LB} = 0. \quad (5.23)$$

From Equations (5.20), (5.21), (5.22) and (5.23), we have,

$$\hat{U}_{C,C-1}^{HB}(t) - \hat{U}_{C,C}^{HB}(t) =_d \hat{U}_{C,C-1}^{LB}(t) - \hat{U}_{C,C}^{LB}(t) \text{ for all } t \geq 0. \quad (5.24)$$

and hence,

$$\begin{aligned} \left|S_C^{HB}(t) - S_C^{LB}(t)\right| &= \theta|\mathbb{E}(\hat{U}_{C,C-1}^{HB}(t) - \hat{U}_{C,C}^{HB}(t)) - \mathbb{E}(\hat{U}_{C,C-1}^{LB}(t) - \hat{U}_{C,C}^{LB}(t))| \\ &= 0 \end{aligned}$$

for all $t \geq 0$. This completes the proof. ■

Note that,

$$\lim_{t\to\infty} \frac{S_{n,C}(t)}{t} = \frac{\theta}{\mathbb{E}(T_{i,L}) + \mathbb{E}(T_{i,U})}. \quad (5.25)$$

From Equations (5.25), (5.21) and (5.22) we have

$$\lim_{t\to\infty} \frac{S_{n,C}^{HB}(t)}{t} = \lim_{t\to\infty} \frac{S_{n,C}^{LB}(t)}{t}, \quad (5.26)$$

for $C \in \mathbb{N}$ and $n \in \{0, ..., C\}$. This explains, from a stochastic point of view, the common asymptotic gradients of the selling prices observed in Figure 5.1.

Observe that the $T_{i,L}$ are independent phase-type random variables with initial state $(C-1, 1)$ and absorbing state $(0, 0)$, and the $T_{i,U}$ are independent phase-type random variables with initial state $(0, 0)$ and absorbing state $(C-1, 1)$. The transition rates associated with both phase-type distributions are given in (5.13)-(5.18), and the expectations appearing in the denominator of (5.25) can be computed explicitly.

Theorem 3 also gives us properties of the last entry of the expected lost revenue vector and the lower right entry of the transient deviation matrix.

Corollary 1 *For all* $t \geq 0$,

$$R^{HB}_{C,C}(t) = R^{LB}_{C,C}(t) + \theta \left(\lambda^{HB} - \lambda^{LB}\right) t. \tag{5.27}$$

Proof: We can write

$$R_{C,C}(t) = R_{0,0}(t) + \sum_{k=1}^{C} - \left(R_{k-1,k-1}(t) - R_{k,k}(t)\right) \tag{5.28}$$

$$= \theta\lambda t + C\theta + \sum_{k=1}^{C} -S_{k,k}(t), \tag{5.29}$$

which gives

$$R^{HB}_{C,C}(t) - R^{LB}_{C,C}(t) = \theta\lambda^{HB} t + C\theta + \sum_{k=1}^{C} -S^{HB}_{k,k}(t) \tag{5.30}$$

$$- \left[\theta\lambda^{LB} t + C\theta + \sum_{k=1}^{C} -S^{LB}_{k,k}(t)\right] \tag{5.31}$$

$$= \theta \left(\lambda^{HB} - \lambda^{LB}\right) t \tag{5.32}$$

by Theorem 3. ■

Corollary 2 *For all* $t \geq 0$,

$$\lambda^{HB} D^{HB}_{C,C}(t) = \lambda^{LB} D^{LB}_{C,C}(t). \tag{5.33}$$

Proof: On the one hand, we have

$$\theta \left(\lambda^{HB} D^{HB}_{C,C}(t) - \lambda^{LB} D^{LB}_{C,C}(t)\right) = R^{HB}_{C,C}(t) - \lambda^{HB}\theta\pi^{HB}_{C} t - R^{LB}_{C,C}(t) + \lambda^{LB}\theta\pi^{LB}_{C} t \tag{5.34}$$

$$= \theta \left[\lambda^{HB} - \lambda^{HB}\pi^{HB}_{C} - \lambda^{LB} + \lambda^{LB}\pi^{LB}_{C}\right] t, \tag{5.35}$$

while, on the other hand,

$$\lambda - \lambda\pi_C = \lambda - (\mu - \lambda)\frac{(\lambda/\mu)^{C+1}}{1-(\lambda/\mu)^{C+1}} \tag{5.36}$$

$$= \frac{\lambda - \mu(\lambda/\mu)^{C+1}}{1-(\lambda/\mu)^{C+1}}. \tag{5.37}$$

Thus, the right hand side of (5.35) is equal to

$$\theta t \left(\frac{\lambda^{HB} - \mu^{HB}(\lambda^{HB}/\mu^{HB})^{C+1}}{1-(\lambda^{HB}/\mu^{HB})^{C+1}} - \frac{\lambda^{LB} - \mu^{LB}(\lambda^{LB}/\mu^{LB})^{C+1}}{1-(\lambda^{LB}/\mu^{LB})^{C+1}} \right). \tag{5.38}$$

After observing that $\lambda^{HB} = \mu^{LB}$ and $\lambda^{LB} = \mu^{HB}$, some minor algebraic manipulation gives the result that expression (5.38) is equal to zero. ■

6 Concluding remarks

In the previous sections, we have investigated three different approaches for computing the expected loss function associated with an $M/M/1/C$ queue, and the links between these approaches were highlighted. The approach presented in Section 2 relies on a conditioning argument; the approach presented in Section 3 makes use of the deviation and the transient deviation matrix; and the approach used in Sections 4 and 5 involves coupling arguments.

The results of Sections 2, 3 and 4 can be generalised to the case where the system has $1 \leq s \leq C$ servers. In particular, the same conditioning argument as in Section 2 was used in [5] to compute the Laplace transform of the loss function corresponding to an $M/M/C/C$ queue. This argument can be generalised to any value of s, however, it leads to cumbersome expressions when $1 < s < C$.

The entries of the deviation matrix of an $M/M/s/C$ queue were computed in [9]. The entries of the Laplace transform of the transient deviation matrix can be computed using (3.18), where explicit expressions for $\Phi_{i,j}$ in the $M/M/s/C$ can be obtained using the results in [8].

The coupling argument developed in Section 4 can also be generalised to the $M/M/s/C$ case. This leads to a generalisation of Theorem 2 and of the technique to compute the last column of the deviation and transient deviation matrix. Note that the results in Theorem 2 can also be obtained directly using the explicit expression for the entries of the last column of the deviation matrix since $\lim_{t\to\infty}[R_{n+1,C}(t) - R_{n,C}(t)] = D_{n+1,C} - D_{n,C}$.

Finally, the results of Section 5 are specific to the single server case and do not generalise to multiple servers.

Acknowledgements

The authors would like to acknowledge the support of the Australian Research Council (ARC) through Laureate Fellowship FL130100039 and the ARC Centre of Excellence for the Mathematical and Statistical Frontiers (ACEMS). Sophie Hautphenne would further like to thank the ARC for support through Discovery Early Career Researcher Award DE150101044.

References

[1] J. Abate and W. Whitt. Numerical inversion of laplace transforms of probability distributions. *ORSA Journal on computing*, 7(1):36–43, 1995.

[2] S. Asmussen and H. Albrecher. *Ruin probabilities*, volume 14. World Scientific, 2010.

[3] F. Carbonell, J. C. Jimenez, and L. M. Pedroso. Computing multiple integrals involving matrix exponentials. *Journal of Computational and Applied Mathematics*, 213(1):300–305, 2008.

[4] M. Chen. Optimal Markovian couplings and applications. *Acta Mathematica Sinica*, 10(3):260–275, 1994.

[5] B. A. Chiera and P. G. Taylor. What is a unit of capacity worth? *Probability in the Engineering and Informational Sciences*, 16(04):513–522, 2002.

[6] P. Coolen-Schrijner and E. A. van Doorn. The deviation matrix of a continuous-time Markov chain. *Probability in the Engineering and Informational Sciences*, 16(3):351–366, 2002.

[7] S. Hautphenne, Y. Kerner, Y. Nazarathy, and P. Taylor. The intercept term of the asymptotic variance curve for some queueing output processes. To appear in *The European Journal of Operations Research*.

[8] Y. Huang and W. McColl. Analytical inversion of general tridiagonal matrices. *Journal of Physics A: Mathematical and General*, 30(22):7919, 1997.

[9] G. M. Koole and F. M. Spieksma. On deviation matrices for birth–death processes. *Probability in the Engineering and Informational Sciences*, 15(02):239–258, 2001.