The Role of Coupling and the Deviation Matrix in Calculating the Value of Capacity for Queueing Systems

Peter Braunsteins, Sophie Hautphenne and Peter Taylor

The University of Melbourne

June 9, 2014

A Finite-Capacity Single-Server Queue

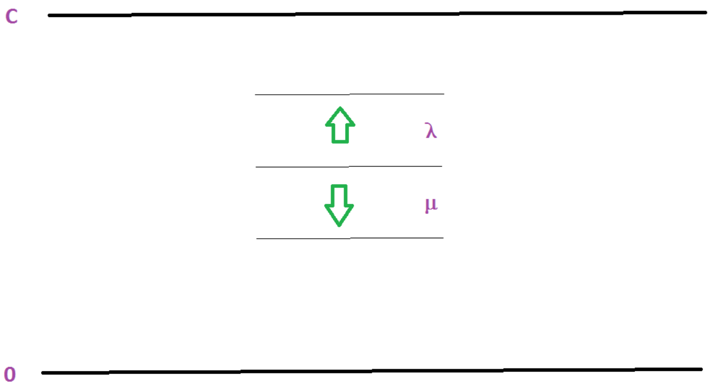
Consider a single server queue with capacity C, including the server.

- Customers arrive according to Poisson process $A(t)$ with parameter λ.
- Service times S_i are exponential with parameter μ.
- Each accepted customer generates θ dollars revenue.
- Customers that arrive when the queue is full are turned away and subsequently generate no revenue.

The queue manager can observe the state and has the option of buying or selling capacity at the start of each time period of length T. How should he/she make this decision?

A Finite-Capacity Single-Server Queue

We use a continuous-time Markov chain model.



A Finite-Capacity Single-Server Queue

Chiera and Taylor (2002) approached a similar problem by letting

$$R_n(t) = \mathbb{E}[\int_0^t \lambda\theta I(Q(u) = C)|Q(0) = n]du$$

denote the expected revenue lost in the interval $[0, t]$, given that there are n connections at time 0 and

$$R_n(t|x) = \mathbb{E}[\int_0^t \lambda\theta I(Q(u) = C)|Q(0) = n, \tau = x]du$$

be the same quantity conditional on the fact that the first time τ that the queue changes state is x.

A Finite-Capacity Single-Server Queue

By thinking about the modelling, we can derive

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

ACEMS

A Finite-Capacity Single-Server Queue

Integrating with respect to the time of the first transition, we see that

$$R_0(t) = \int_0^t R_1(t-x)\lambda e^{-\lambda x} dx,$$

$$R_n(t) = \int_0^t \left[\mu R_{n-1}(t-x) + \lambda R_{n+1}(t-x)\right] e^{-(\lambda+\mu)x} dx$$

and

$$R_C(t) = \int_0^t R_{C-1}(t-x)\mu e^{-\mu x} dx + \frac{\theta\lambda}{\mu}\left(1 - e^{-\mu t}\right).$$

A Finite-Capacity Single-Server Queue

Now taking Laplace transforms, it follows that $\widetilde{R}_n(s)$ satisfies the equations

$$\widetilde{R}_0(s) = \frac{\lambda}{s+\lambda}\widetilde{R}_1(s),$$

$$\widetilde{R}_n(s) = \frac{\lambda}{s+\lambda+\mu}\widetilde{R}_{n+1}(s) + \frac{\mu}{s+\lambda+\mu}\widetilde{R}_{n-1}(s),$$

for $0 < n < C$, and

$$\widetilde{R}_C(s) = \frac{1}{s+\mu}\left(\mu\widetilde{R}_{C-1}(s) + \frac{\theta\lambda}{s}\right).$$

A Finite-Capacity Single-Server Queue

The solution to these equations is

$$\widetilde{R}_n(s) = A(s) r_1(s)^n + B(s) r_2(s)^n,$$

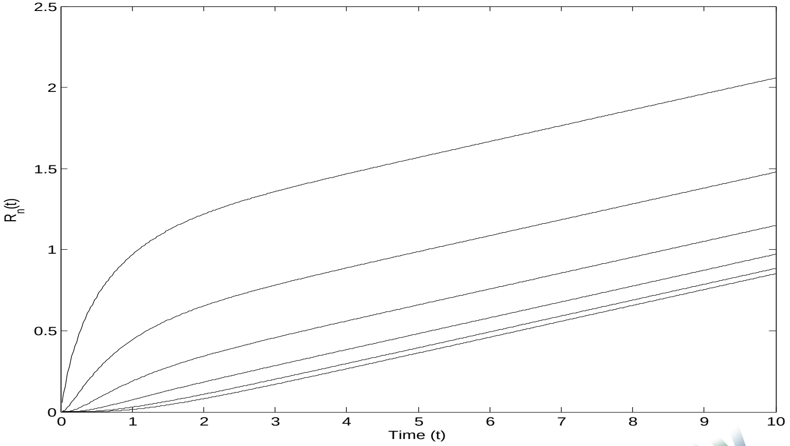
where

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

and the constants $A(s)$ and $A(s)$ can be derived from the boundary conditions.

We used the Euler method as described by Abate and Whitt (1995) to invert the transform of $\widetilde{R}_n(s)$ to yield $R_n(T)$.

The lost revenue functions for $n = 0, \ldots, 5$ when $C = 5$, $\lambda = 3$ and $\mu = 5$



Buying and selling prices

Now, indexing the lost revenue function by the capacity, we can derive “buying" and “selling" values, $B_{n,C}(T)$ and $S_{n,C}(T)$ of bandwidth when there are initially $n < C$ customers present via the expressions

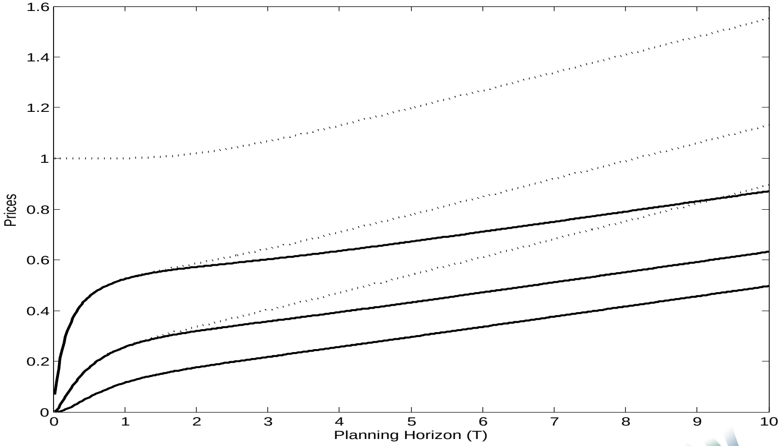
$$\begin{aligned} B_{n,C}(T) &= R_{n,C}(T) - R_{n,C+1}(T) \\ S_{n,C}(T) &= R_{n,C-1}(T) - R_{n,C}(T). \end{aligned}$$

When $n = C$, we write

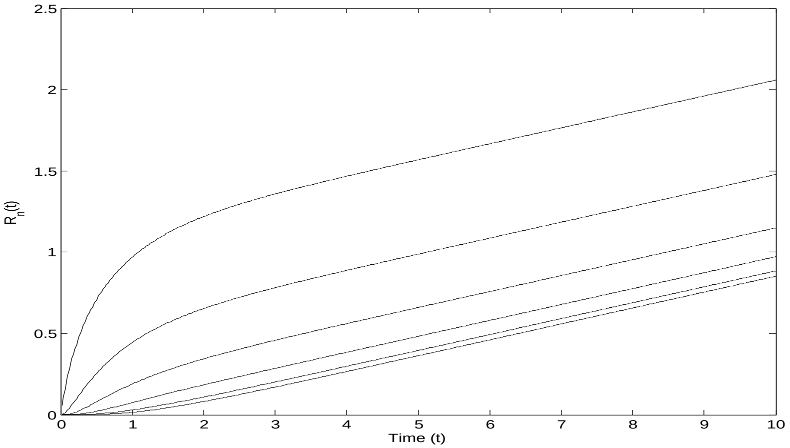
$$\begin{aligned} B_{C,C}(T) &= R_{C,C}(T) - R_{C,C+1}(T) \\ S_{C,C}(T) &= R_{C-1,C-1}(T) - R_{C,C}(T) + f(\theta). \end{aligned}$$

where $f(\theta)$ represents a penalty function for ejecting a customer.

Buying and selling prices for $n = 3, 4, 5$ when $C = 5$, $\lambda = 3$ and $\mu = 5$



The lost revenue functions again



The lost revenue functions again

All the functions have slope equal to $\lambda\theta\pi_C$, which is the stationary rate of losing revenue.

The difference between $R_n(T)$ and the line $R = \lambda\theta\pi_C T$ is

$$\begin{aligned} \Delta_n(T) &= \lambda\theta \int_0^T p_{n,C}(u)du - \lambda\theta\pi_C T \\ &= \lambda\theta \int_0^T \left[p_{n,C}(u) - \pi_C\right] du. \end{aligned}$$

This reminds us of the deviation matrix corresponding to the generator Q of the Markov chain.

The deviation matrix

For a continuous-time Markov chain with generator Q, the *deviation matrix* D was discussed by Coolen-Schrijner and van Doorn (2002). The used this terminology for the matrix whose $(i,j)th$ element is

$$D_{ij} = \int_0^\infty \left[p_{ij}(u) - \pi_j\right] du$$

where $p_{ij}(u) = P(X(t) = j | X(0) = u)$ and $\boldsymbol{\pi}^t \equiv (\pi_j)$ is the stationary distribution, which satisfies

$$\boldsymbol{\pi}^t Q = 0.$$

with

$$\boldsymbol{\pi}^t \mathbf{1} = 1.$$

The deviation matrix

With $\Pi = \mathbf{1}\pi^t$, it is relatively easy to show that

$$D(-Q) = (-Q)D = I - \Pi,$$

$$(-Q)D(-Q) = -Q$$

and

$$D(-Q)D = D$$

so, not only is D a generalised inverse of $-Q$, it is the group, or Drazin, inverse of $-Q$.

The deviation matrix

For a specified column vector $\boldsymbol{g}$, the deviation matrix is useful for solving Poisson's equation

$$-Q\boldsymbol{h} = \boldsymbol{g} - w\mathbf{1}.$$

for the vector/scalar pair $(\boldsymbol{h}, w)$.

When the state space is finite, the solution is

$$\boldsymbol{h} = -D\boldsymbol{g} + c\mathbf{1},$$

with

$$w = \pi\boldsymbol{g}$$

and c a constant that needs to be specified.

The deviation matrix

Our equations

$$R_0(t) = \int_0^t R_1(t-x)\lambda e^{-\lambda x} dx,$$

$$R_n(t) = \int_0^t \left[\mu R_{n-1}(t-x) + \lambda R_{n+1}(t-x)\right] e^{-(\lambda+\mu)x} dx$$

and

$$R_C(t) = \int_0^t R_{C-1}(t-x)\mu e^{-\mu x} dx + \frac{\theta\lambda}{\mu}\left(1 - e^{-\mu t}\right).$$

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

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

where $\boldsymbol{g}^t = (0, \ldots, 0, \lambda\theta e^{\mu t})$.

The deviation matrix

In our capacity planning example, we effectively wrote the solution in terms of

$$D(T) = \int_0^T [P(u) - \Pi]\, du$$

rather than

$$D = \int_0^\infty [P(u) - \Pi]\, du.$$

These matrices are related via the equation

$$D(T) = \left[I - e^{QT}\right] D$$

but, since e^{QT} is hard to calculate, it is not easy to see how to get $D(T)$ this way.

Coupling

Now we use a completely different approach. First consider the continuous-time Markov chain model of the number of customers to be driven by two independent Poisson processes

- the process $A(t)$ of 'potential arrivals' with rate λ, and
- the process $S(t)$ of 'potential services' with rate μ.

The 'free' process that starts with n customers

$$\tilde{X}(t) = n + A(t) - S(t)$$

takes values on all of the integers and, when $\lambda < \mu$, it will drift towards $-\infty$ with probability one.

However, on the set $\{1, \ldots, C - 1\}$, the process behaves like our single-server queue.

Coupling

To make the system behave exactly like our single server queue, we introduce two new processes $U(t)$ and $L(t)$ that count the number of arriving customers lost due to the queue being full, and the number of services wasted due to it being empty, respectively.

So we have

$$X(t) = n + [A(t) - U(t)] - [S(t) - L(t)]$$

where $U(t)$ increases only when $X(t) = C$ and an arrival occurs, and $L(t)$ increases only when $X(t) = 0$ and a potential service occurs. This is the two-sided *regulator* or *Skorokhod map*.

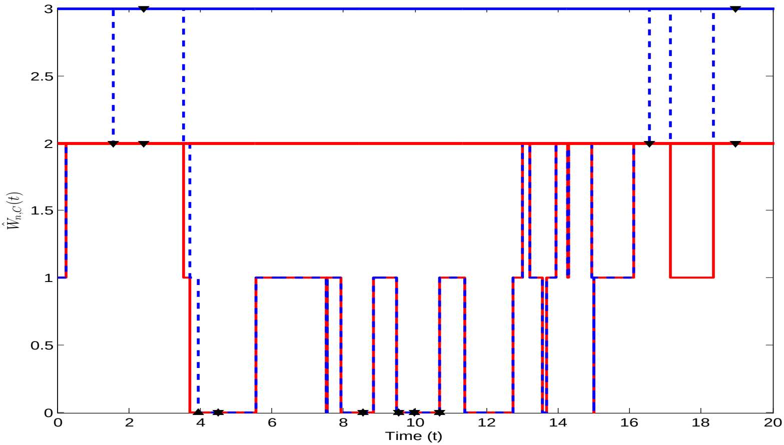
Coupling

When the capacity is C, the process $\theta U_C(t)$ gives us the amount of revenue that we have lost up to time t: similarly when the capacity is $C+1$, the process $\theta U_{C+1}(t)$ gives us the amount of revenue that we have lost up to time t. So the buying price is

$$B_c(t) = \mathbb{E}\left[\theta\left(U_C(t) - U_{C+1}(t)\right)\right]$$

Instead of analysing $\theta U_C(t)$ and $\theta U_{C+1}(t)$ driven by independent pairs of Poisson processes $(A_C(t), S_C(t))$ and $(A_{C+1}(t), S_{C+1}(t))$ respectively, the trick is to drive the capacity C and $C+1$ queues with the same pair of Poisson processes.

Coupling



Coupling

Assume both queues start with n customers.

Their sample paths remain coupled until the first time τ_1 that a customer arrives when $X_{C+1}(\tau_1) = X_C(\tau_1) = C$. This customer is accommodated in the capacity $C + 1$ queue, but lost from the capacity C queue.

After time τ_1, we have $X_{C+1}(t) = X_C(t) + 1$ until time τ_2 when $X_{C+1}(t) = 1$ and $X_C(t) = 0$ and a provisional service occurs. The queues are then coupled again, both with no customers.

Coupling

The successive couping/uncoupling intervals form an alternating renewal process, with every ‘uncoupling’ renewal corresponding to an increase by one in the difference $U_{C+1}(t) - U_C(t)$.

We can thus characterise the buying price function at time T in terms of the expected number of uncoupling renewals by time T.

We can again approach this via Laplace transforms.

Coupling

For example, asymptotically,

$$\lim_{T \to \infty} \frac{B(T)}{T} = \frac{\theta}{m_U + m_C}$$

where m_U is the mean time between a coupling time instant and an uncoupling time instant, and m_C is the mean time between an uncoupling time instant and a coupling time instant.