Optimal design of a bulk service queue

Messaoud Bounkhel †
Lotfi Tadj ‡

† King Saud University
College of Science, Department of Mathematics,
P.O. Box 2455, Riyadh 11451, Saudi Arabia
bounkhel@ksu.edu.sa

† American University in Dubai
School of Business Administration, Department of Management and e-business
P.O. Box 28282, Dubai, United Arab Emirates
ltadj@aud.edu

Abstract

This paper uses a control-theoretic approach, the so-called predictive control, for the optimal design of a single-server, finite capacity, bulk-service queueing system. The problem is presented as an optimal control problem with one state vector variable and one control variable. The optimal service rate is obtained explicitly under a general quadratic cost function. Illustrative numerical examples are presented and the effect of the initial state distribution on the shape of the optimal solution is demonstrated.

Keywords: Queue, optimal control, predictive control, optimal design

1 Introduction

Operations Research models are usually classified as being either descriptive or prescriptive models. Descriptive models are models which *describe* some current real-world situations, while prescriptive models are models which *prescribe* what the real-world situation should be, that is, the "optimal" situation at which to aim, Gross and Harris (1998).

The development of queueing theory is primarily dominated by descriptive models. Some attention has been given nonetheless to optimization. It is customary to refer to prescriptive queueing models under the title *optimal design and control of queues*. Optimal design models are also called *economic* (*cost or profit*) *models* or *static models*, while optimal control models are also called *rate-control models* or *dynamic models*, see Gross and Harris (1998). Optimal design models are generally standard queues with a superimposed cost or profit function to be optimized with respect to parameters such as

the arrival rate λ, the service rate μ, or the number of available channels c. Optimal control models try to find the optimal operating policy, that is, rules for turning the server on and off that result in the lowest long-run cost, see Cooper (1990). Most research on the optimal control of queueing models focuses on three different threshold policies that were introduced in the literature: the N-policy, introduced by Yadin and Naor (1963), the D-policy introduced by Balachandran (1973), and the T-policy introduced by Heyman (1977).

Since the main reason why one adopts a control policy for a queueing system is to control the system 'economically', optimization of costs is necessary. A cost function is designed specifically and optimal thresholds that yield minimum cost are sought. Most cost functions are built according to the following considerations

- Activating and deactivating the server result in fixed start-up and shut-down costs, respectively.
- When the server is turned off, an idle cost for power, heat, maintenance, etc. is charged, and when the server is turned on, attendant, fuel, or other costs may be added to the dormant cost to form the running cost.
- The holding cost is a penalty for delaying a customer in the system.

We want to note here that the denomination "optimal control" is somewhat misleading. The survey of Tadj and Choudhury (2005) shows that even though the appellation "optimal control" is used, it is optimization techniques rather than optimal control techniques that are used to deal with such queueing systems. Indeed, the solution of the optimization considered is a single point or set of points instead of being a whole function of time.

Optimal control theory has proven to be very efficient in obtaining optimal solutions for dynamical systems. Sethi and Thompson (2000) report applications of optimal control theory in various areas of Management Science and Operations Research such as finance (Davis and Elzinga (1972), Elton and Gruber (1975), Sethi (1978)), economics (Arrow and Kurz (1970), Feichtinger (1988), Kamien and Schwartz (1998)), marketing (Feichtinger *et al.* (1994), Sethi (1977)), maintenance (Pierskalla and Voelker (1976), Rapp (1974)), environment and transportation (Bounkhel and Tadj (2006), Khemlnitsky and Gerchak (2002)), and the consumption of natural resources (Amit (1986), Derzko and Sethi (1981)). Interestingly enough, to the best of our knowledge, the literature on the application of optimal control theory in the theory of queues seems to be non-existent.

The goal of this paper is to use the optimal control theory to determine the optimal solution in an optimal design problem. Our starting point is the paper of Selim (1997). He considers the queueing system $M/M^N/1/N$ where customers are served in batches of maximum size N. He obtains the time-dependent probability distribution for the number of customers in the system. To predict the optimal service rate, he proposes a minimization problem during a finite time horizon. Then, by taking the case of a linear objective function and by assuming piecewise constant service rate μ, he reformulates that problem into a standard *multi-variable optimization problem*. To solve the same problem, we consider a different and more general model, in the sense that there is no restriction on the service rate μ. Our approach is to propose a quadratic objective function and then to use some techniques from optimal control theory, the so-called predictive control, to obtain the optimal service rate μ not necessarily a piecewise constant function.

The rest of the paper is organized as follows: In Section 2, we recall from Selim (1997) the dynamics of the process and the time-dependent probability distribution for the number of customers in the system which will be used in the next sections to predict the optimal service rate. Section 3 is devoted to state our model and to solve it using the

predictive control method. Illustrative examples are given in Section 4 where the effect of the initial state distribution on the optimal service rate, is investigated.

2 Preliminaries

Consider the M/ M^N /1/ N queueing system and let λ and μ denote the arrival rate and a group service rate, respectively. Also, for $n = 0, \cdots, N$, denote by $p_n(t)$ the transient state probability that there are n customers in the system at time t and let $P(t) = (p_0(t), p_1(t), \cdots, p_N(t))$.

The queueing process is Markovian and the Kolmogorov system of difference-differential equations is given by

$$\dot{P}(t) = P(t)(\mu F + \lambda G), \tag{2.1}$$

where the infinitesimal operator $\mu F + \lambda G$ is such that the $(N+1)\times(N+1)$ matrices F and G are given by

$$F = \begin{bmatrix} 0 & & & & & \\ 1 & -1 & & & & \\ 1 & & -1 & & & \\ \vdots & & & \ddots & & \\ 1 & & & & -1 & \\ 1 & & & & & -1 \end{bmatrix} \text{ and } F = \begin{bmatrix} -1 & 1 & & & & \\ & -1 & 1 & & & \\ & & -1 & 1 & & \\ \vdots & & & \ddots & \ddots & \\ & & & & -1 & 1 \\ & & & & & 0 \end{bmatrix}$$

Selim (1997) has shown that the time-dependent solution of the bulk service system is given by

$$P(t) = P(0)M(t),$$

where $M(t)$ is the transition matrix whose elements $m_{ij}(t)$ are given by

$$m_{ij}(t) = \begin{cases} \left[\lambda^j \mu\left[1 - \sum_{k=0}^{j} g(k)\right] + \sigma_{ij}\lambda^{j-i}(\lambda+\mu)^{i+1} g(j-i)\right]/(\lambda+\mu)^{j+1}, & 0 \le i \le N, 0 \le j < N, \\ \left(\lambda/(\lambda+\mu)\right)^N \left[1 - \sum_{k=0}^{N-1} g(k)\right], & i = 0, j = N, \\ \left(\lambda/(\lambda+\mu)\right)^N \left[1 - \sum_{k=0}^{N-1} g(k) + \sum_{r=1}^{i} \left[(\lambda+\mu)/\lambda\right]^r g(N-r)\right], & 1 \le i \le N, j = N, \end{cases}$$

where

$$\sigma_{ij} = \begin{cases} 1, & i \le j, \\ 0, & \text{otherwise}, \end{cases}$$

and

$$g(k) = e^{-(\lambda+\mu)t}\frac{[(\lambda+\mu)t]^k}{k!}.$$

3 Optimal Design of the Bulk Service Queue

In this section we introduce the objective function J to minimize with respect to the service rate μ. Given a planning horizon of length H, given $T > 0$ with $T << H$, and an instant of time t in the finite horizon $[0, H]$, we propose the following general form of J:

$$J(P, \mu) = \int_t^{t+T} C_1(s, p_N(s))ds + \int_t^{t+T} C_2(s, P(s))ds + \int_t^{t+T} C_3(s, \mu(s))ds, \tag{3.1}$$

where $C_1:[t,t+T]\times[0,1]\to[0,\infty)$, $C_2:[t,t+T]\times[0,1]^{N+1}\to[0,\infty)$, and $C_3:[t,t+T]\times[0,\infty]\to[0,\infty)$ are differentiable functions. The first integral penalizes the system each time its full capacity is reached. The second integral penalizes each state of the system, since it contains the probabilities of all possible states from being idle to being full. The last integral can be interpreted as the cost of the provided service. Selim (1997) considers the case when all the functions $C_i,(i=1,2,3)$ are linear with respect to the second variable. In the present paper, we assume that all those functions are quadratic with respect to the second variable so that

$$\begin{aligned} C_1(t,p_N(t)) &= \frac{1}{2}\alpha p_N^2(t), \\ C_2(t,P(t)) &= \frac{1}{2}\sum_{k=0}^{N} k\beta_k p_k^2(t), \\ C_3(t,\mu(t)) &= \frac{1}{2}\gamma\mu^2(t), \end{aligned}$$

where $\alpha,\gamma,\beta_k\,(k=0,\cdots,N)$ are nonnegative constants. The quadratic form of the objective function is the most used form in optimal control theory (see for example Sethi and Thompson (2000) and the references therein). Inserting the above forms of $C_i,(i=1,2,3)$ in (3.1), the objective function can be rewritten as follows:

$$J(P,\mu)=\frac{1}{2}\int_t^{t+T}\left[\| P(s)\|_Q^2+\gamma\mu^2(s)\right]ds, \tag{3.2}$$

where Q is the diagonal matrix $Q=\mathrm{diag}\{q_0,\cdots q_N\}$ with $q_N=\alpha+N\beta_N$ and $q_k=k\beta_k, k=0,\cdots N-1$.

The optimal control problem we are considering is the following:

(P)
$$\begin{aligned} &\min_{\mu\geq 0} J(P,\mu) \\ &\text{subject to} \quad \dot{P}(t)=P(t)(\mu(t)F+\lambda G) \end{aligned}$$

The model is represented as an optimal control problem with one state variable $P(t)$ and one control variable $\mu(t)$.

Let $f(t)=\| P(t)\|_Q^2+\gamma\mu^2(t)$. The interval $[t,t+T]$ is divided into $2m$ subintervals of equal width $h=\frac{T}{2m}$. Using the composite Simpson's rule for $2m$ intervals, the objective function can be approximated as

$$J\approx\frac{h}{3}\left[f(t)+2\sum_{i=1}^{m-1}f(t+2ih)+4\sum_{i=1}^{m}f(t+(2i-1)h)+f(t+2mh)\right]. \tag{3.3}$$

By the first-order Taylor approximation, we have for $k=1,\cdots,2m$

$$P(s+kh)\approx P(s)+kh\dot{P}(s),$$

which ensures by using the state equation (2.1)

$$P(s+kh)\approx P(s)\{Id+kh[\mu(s)F+\lambda G]\},$$

where Id stands for the $(N+1)\times(N+1)$ identity matrix. Inserting this approximation of P in the definition of f we get

$$f(t+2kh)=\| P(t)\|_Q^2+4k^2h^2\,\| A(t)P(t)\|_Q^2+4khP(t)^{\mathrm{T}}QA(t)P(t)+\gamma\mu^2(t+2kh),$$

and

$$f(t+(2k-1)h)=\| P(t)\|_Q^2+(2k-1)^2h^2\,\| A(t)P(t)\|_Q^2+2(2k-1)hP(t)^{\mathrm{T}}QA(t)P(t)+\gamma\mu^2(t+(2k-1)h),$$

where $A(t)=\mu(t)F+\lambda G$. Summing these equations from $k=1$ to $m-1$ and from $k=1$ to m, respectively, yields

$$\sum_{k=1}^{m-1}F(t+2kh)= (m-1)\parallel P(t)\parallel_Q^2+\frac{2}{3}h^2m(m-1)(2m-1)\parallel A(t)P(t)\parallel_Q^2 + 2hm(m-1)P(t)^{\mathrm{T}}QA(t)P(t)+\gamma\sum_{k=1}^{m-1}\mu^2(t+2kh)$$

and

$$\sum_{k=1}^{m}F(t+(2k-1)h)= m\parallel P(t)\parallel_Q^2+\sum_{k=1}^{m}(2k-1)^2h^2\parallel A(t)P(t)\parallel_Q^2 + 2\sum_{k=1}^{m}(2k-1)hP(t)^{\mathrm{T}}QA(t)P(t)+\gamma\sum_{k=1}^{m}\mu^2(t+(2k-1)h)$$

Some computations give

$$\parallel A(t)P(t)\parallel_Q^2=M_1(t)\mu^2(t)+M_2(t)\mu(t)+M_3(t),$$

and

$$P(t)^{\mathrm{T}}QA(t)P(t)=M_4(t)\mu(t)+M_5(t),$$

where

$$M_1(t)=\parallel FP(t)\parallel_Q^2,\ M_2(t)=2\lambda P(t)^{\mathrm{T}}F^{\mathrm{T}}QGP(t),\ M_3(t)=\lambda\parallel GP(t)\parallel_Q^2,$$

$$M_4(t)=P(t)^{\mathrm{T}}QFP(t),\ \text{and}\ M_5(t)=\lambda P(t)^{\mathrm{T}}QGP(t).$$

Hence, the objective function can be further approximated as

$$J\approx\frac{h}{3}\left\{\tilde{\gamma}\mu^2(t)+2\gamma\left[\sum_{k=1}^{m-1}\mu^2(t+2kh)+2\sum_{k=1}^{m}\mu^2(t+(2k-1)h)\right]+\gamma\mu^2(t+2mh)+\hat{\gamma}\mu(t)+L(P(t),t)\right\}$$

where

$$\tilde{\gamma}(t)=\gamma+4h^2m^2(2m-1)M_1(t),$$

$$\hat{\gamma}(t)=4hm[hm(2m-1)M_2(t)+(3m-1)M_4(t)],$$

$$L(P(t),t)=2m\parallel P(t)\parallel_Q^2+\parallel P(t+2mh)\parallel_Q^2+hm(2m-1)\left[\frac{h}{3}(4m-1)M_3(t)+2M_5(t)\right],$$

and then

$$J\approx\frac{h}{3}\left[U(t)^{\mathrm{T}}\mathbf{R}(t)U(t)+H(t)^{\mathrm{T}}U(t)+L(P(t),t)\right]. \tag{3.4}$$

where

$$U(t)=\left[\mu(t),\mu(t+h),\mu(t+2mh)\right]^{\mathrm{T}},$$

$$\mathbf{R}(t)=\mathrm{diag}\{\tilde{\gamma}(t),4\gamma,2\gamma,\cdots,2\gamma,4\gamma,\gamma\}\ \text{and}\ H(t)=(\hat{\gamma}(t),0,0,\cdots,0).$$

Note that the expression $L(P(t),t)$ is independent of $U(t)$. Hence, it is easy to see that the minimum of J is reached at $U(t)$ satisfying

$$\mathbf{R}U(t)=-\frac{1}{2}H(t),$$

which gives

$$\mu(t)=-\frac{\hat{\gamma}(t)}{2\tilde{\gamma}(t)}, \tag{3.5}$$

since $\tilde{\gamma}(t)>0$.

4 Numerical Illustrations

We show in this section some of the service rate patterns that can be obtained by the proposed procedure. We consider a system with the following arrival and service rates: $\lambda = 5$ and $\mu = 2$. We take for unit penalties $\alpha = 5, \beta = 5, \gamma = 5$, and also assume that $T = 40, m = 50$, and $N = 11$. All these parameters may affect the shape of the solution. Another choice that also affects the solution is the initial state distribution $P(0)$. To show this effect we chose to consider the following different initial distributions:

	uniform	**arithmetic**	**truncated geometric**	**binomial**	**deterministic**
$p_i(0)$ $(i = 0, \cdots, N)$	$\frac{1}{N+1}$	$\frac{2i}{N(N+1)}$	$\frac{(1-r)r^i}{1-r^N}$ $(0 < r < 1)$	$\binom{N}{i} r^i (1-r)^{N-i}$ $(0 < r < 1)$	$\begin{cases} 1, & \text{if } i = j, \\ 0, & \text{if } i \neq j. \end{cases}$ $(j = 0, \cdots, N)$

Figure 1 shows the optimal service rate for the different choices of the initial distribution. We note three patterns for the optimal service rate: (i) it increases and then converges to a fixed value, $\mu^* = 2.038$ in this case (first column); (ii) it increases above that fixed service rate and then decreases to tend to μ^* (second column); (iii) it decreases to tend to μ^* (third column). Comparing these patterns with the initial distribution used, we noted that the first pattern happens when the mass of $P(0) = (p_i(0); i = 0, \cdots, N)$ is concentrated at the lower values of i, the second pattern happens when the mass of $P(0)$ is concentrated at the middle values of i, and the third pattern happens when the mass of $P(0)$ is concentrated at the higher values of i. In other words, the three patterns are due to the skewness of $P(0)$: the first pattern (first column) when $P(0)$ is skewed to the right, the second pattern (second column) when it is not skewed, and the third pattern (third column) when it is skewed to the left.

5 References

Amit, R. (1986) Petroleum reservoir exploitation: switching from primary to secondary recovery, Operations Research, Vol. 34, No. 4, pp. 534-549.

Arrow, K.J. and Kurz, M. (1970) Public Investment, the Rate of Return, and Optimal Fiscal Policy, The John Hopkins Press, Baltimore.

Balachandran K.R. (1973) Control policies for a single server system, Management Science, Vol. 19, No. 9, pp. 1013-1018.

Bounkhel, M. and Tadj, L. (2006) Minimizing energy use for a road expansion in a transportation system using optimal control theory, Applied Mathematics E-Notes (AMEN), Vol. 6, pp. 159-166.

Cooper R.B. (1990) Introduction to the Theory of Queues, 3rd ed. CEE Press Books, Washington, D.C.

Davis, B.E. and Elzinga, D.J. (1972) The solution of an optimal control problem in financial modelling, Operations Research, Vol. 19, pp. 1419-1433.

Derzko, N.A. and Sethi, S.P. (1981) Optimal exploration and consumption of a natural resource: deterministic case, Optimal Control Applications & Methods, Vol. 2, No.1, pp. 1-21.

Elton, E. and Gruber, M. (1975) Finance as a Dynamic Process, Prentice-Hall, Englewood Cliffs, NJ.

Feichtinger, G. (ed.) (1988) Optimal Control Theory and Economic Analysis, Vol. 3, North-Holland, Amsterdam.

Feichtinger, G., Hartl, R.F., and Sethi, S.P. (1994) Dynamic optimal control models in advertising: recent developments, Management Science, Vol. 40, No. 2 pp. 195-226.

Gross D. and Harris C.M. (1998) Fundamentals of Queueing Theory, 3rd ed. John Wiley and Sons, New York.

Heyman D.P. (1977) The T-policy for the M/G/1 queue, Management Science, Vol. 23, No. 7, pp. 775-778.

Kamien, M.I. and Schwartz, N.L. (1998) Dynamic Optimization: The Calculus of Variations and Optimal Control in Economics and Management, 2nd ed., Fourth Impression, North-Holland, New York.

Khemlnitsky, E. and Gerchak, Y. (2002) Optimal control approach to production systems with Inventory-level-dependent demand, IIE Transactions on Automatic Control, Vol. 47, pp. 289-292.

Pierskalla, W.P. and Voelker, J.A. (1976) Survey of maintenance models: the control and surveillance of deteriorating systems, Naval Research Logistics Quarterly, Vol. 23, pp. 353-388.

Rapp, B. (1974) Models for Optimal Investment and Maintenance Decisions, Almqvist & Wiksell, Stockholm; Wiley, New York.

Selim, S.Z. (1997) Time-dependent solution and optimal control of a bulk service queue, Journal of Applied Probability, Vol. 34, pp. 258-266.

Sethi, S.P. (1977) Dynamic optimal control models in advertising: a survey, SIAM Review, Vol. 19, No. 4, pp. 685-725.

Sethi, S.P. (1978) Optimal equity financing model of Krouse and Lee: corrections and extensions, Journal of Financial and Quantitative Analysis, Vol. 13, No. 3, pp. 487-505.

Sethi, S.P. and Thompson, G.L. (2000) Optimal Control Theory: Applications to Management Science and Economics, 2nd ed., Kluwer Academic Publishers, Dordrecht.

Tadj, L. and Choudhury, G. (2005) Optimal design and control of queues, TOP, Vol. 13, No. 1, pp. 359-414.

Yadin M. and Naor P. (1963) Queueing systems with a removable service station, Operational Research Quarterly, Vol. 14, pp. 393-405.

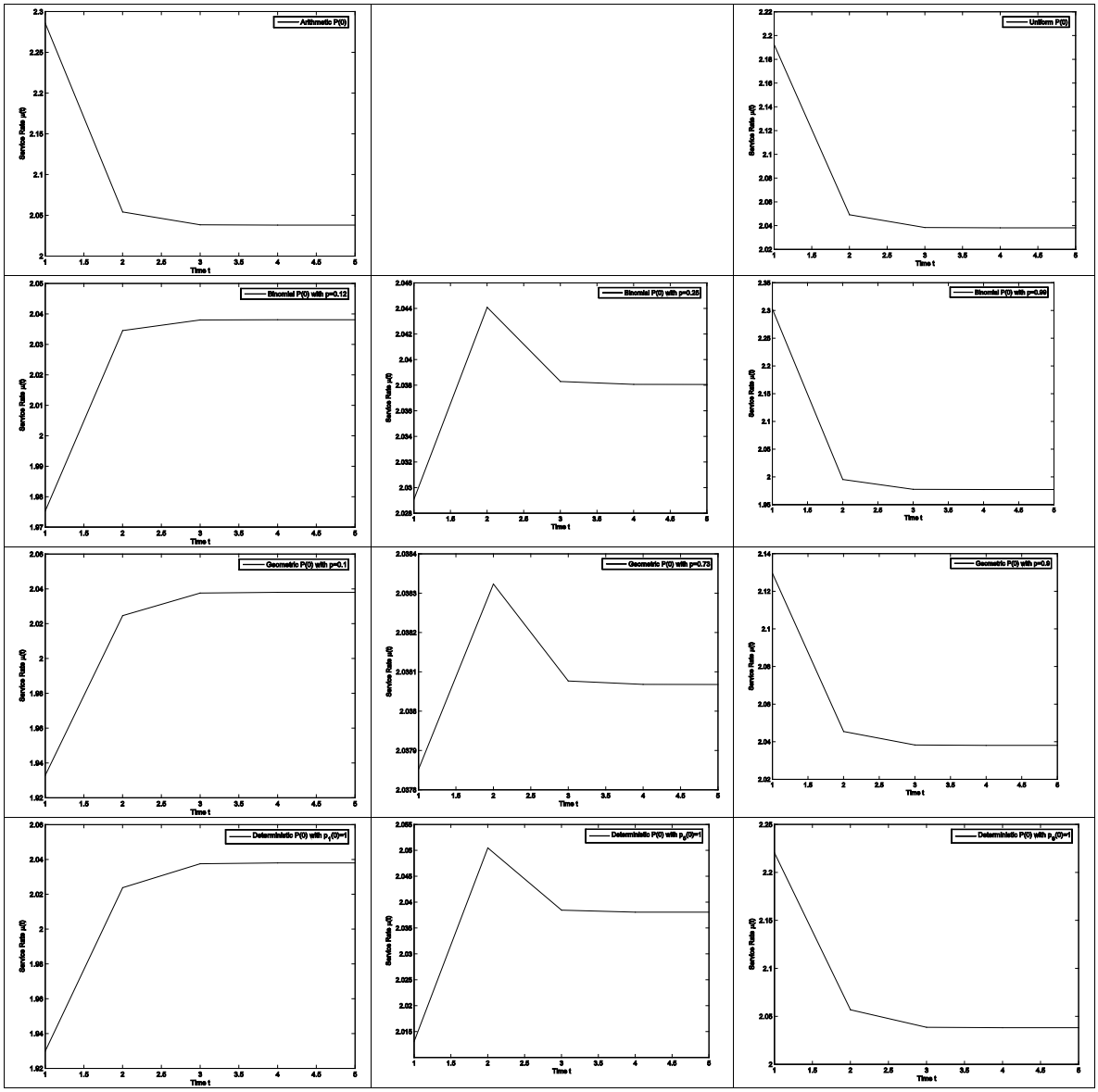

Figure 1: Optimal solution for Example 3.1.