Operations Research Letters 45 (2017) 538–542

Contents lists available at ScienceDirect

Operations Research Letters

journal homepage: www.elsevier.com/locate/orl

An approximating diffusion control problem for dynamic admission and service rate control in a $G/M/N+G$ queue

Yaşar Levent Koçağa

Information and Decision Sciences Department, Sy Syms School of Business, Yeshiva University, New York, NY 10033, United States

ARTICLE INFO

Article history:
Received 3 October 2016
Received in revised form 2 August 2017
Accepted 2 August 2017
Available online 16 August 2017

Keywords:
Admission control
Service rate control
Customer abandonment
Diffusion control problem
Halfin–Whitt (QED) heavy traffic regime
Average cost

ABSTRACT

We study a diffusion control problem that is motivated by the dynamic admission and service rate control problem for a $G/M/N+G$ queue. The objective is to minimize long run average cost. Because the original queueing control problem is not tractable, we solve the approximating diffusion control problem that arises under the QED heavy traffic regime and show that its optimal solution has two components: (1) a threshold control that regulates the diffusion and (2) a feedback-type drift rate control.

1. Introduction

We consider a service system where the system manager can exercise both admission control and service rate control to reduce congestion. Earlier literature on joint admission and service rate control includes [3,1], and [4]. Our work generalizes these papers by considering multiple servers as well as general arrival and abandonment distributions. In particular, we assume that the service system under consideration can be modeled as a $G/M/N+G$ multi-server queue operating in QED heavy traffic regime also known as the Halfin–Whitt heavy traffic regime. Customer inter-arrival times are assumed to follow a general distribution with mean $\frac{1}{\lambda}$ and variance σ_A^2. We assume service times follow an exponential distribution and that the service rate can be adjusted. Specifically, there exists a base service rate μ which can be increased to reduce congestion. Customers waiting for service abandon according to a general distribution $F(\cdot)$ which satisfies $F'(0)=\gamma$. All distributions are assumed to be independent and identically distributed. The system manager can also dynamically control arrivals by deciding whether to admit or reject arrivals. Customers waiting in queue incur a waiting cost of c_w per unit time, and each customer abandonment costs a. To rule out uninteresting cases, we assume $c_w + a\gamma > 0$. There is also an idleness cost $h_i \geq 0$ per idle server per unit time. This cost can be thought of as an *opportunity cost* and it also captures lost revenues due to idleness when calls generate income as in outbound call centers (see also section 2.4 in [7] for a detailed discussion on this cost structure). Service rate adjustments beyond the base service rate μ incur a cost $c(\cdot)$ and each customer rejected costs p. We list more specific assumptions regarding the cost structure in Section 2. Finally, the objective of the system manager is to minimize long run average cost also known as ergodic cost.

Note that the queueing control problem (QCP) described above is not tractable since the arrival and abandonment distributions are not Markovian. Thus, we follow a similar approach to [2] and [3], and we solve the associated diffusion control problem (DCP) that arises when the original queueing system operates in the QED heavy traffic regime. Our primary contribution is to solve the DCP and establish the solution has two components: (1) a threshold control that regulates the diffusion and (2) a drift rate control that increases in system state. We also show that the threshold is finite if and only if the *effective abandonment cost* is larger than the cost of rejection and establish a relationship between the optimal threshold and the optimal long run average cost. The optimality of the alluded policy assumes the existence of a solution to the Hamilton–Jacobi–Bellman (HJB) optimality equations. Furthermore, both the drift control and threshold control depend explicitly on this solution. However, finding a solution to the HJB optimality equations is not a trivial task. In particular, when the original queueing system operates under QED, the underlying diffusion can assume both positive and negative values. Hence, solving the HJB optimality equations requires solving two separate differential equations without initial conditions and smooth-pasting them at the origin. Moreover, the solutions of these equations is not trivial either and does not follow from the standard theory of differential equations. Thus, we follow a novel approach and construct a solution to the

E-mail address: kocaga@yu.edu.

http://dx.doi.org/10.1016/j.orl.2017.08.005
0167-6377/

HJB optimality equations via a nested parametrization approach as in [7]. This approach also allows us to devise an algorithm, which calculates the optimal threshold and diffusion cost and can be used to construct an optimal policy for the motivating QCP.

The rest of the paper is organized as follows. In Section 2, we describe and solve the approximating diffusion control problem and construct a solution to the HJB optimality equations. We make concluding remarks in Section 3. All proofs and a detailed numerical study are provided in the Electronic Companion (EC) (see Appendix A).

2. The diffusion control problem

In this section, we will present and solve the diffusion control problem that serves to approximate the joint admission and service rate control problem for the underlying queueing system. In Section 2.1, we present the approximating diffusion and its cost structure which leads us to formulate the diffusion control problem. In Section 2.2, we provide a solution to the approximating DCP under the assumption that a solution to the HJB optimality equations exists. In Section 2.3, we construct a solution to the HJB optimality equations.

2.1. The approximating control problem

We consider a filtered probability space $((\Omega, \mathcal{F}, P), \{\mathcal{F}_t\})$ and let $B(t)$ denote a Brownian motion with respect to the associated filtration. Then, the evolution of the state process is given by

$$X(t) = X(0) + \sigma B(t) - \int_0^t [\beta(X(s)) + u(s)]\, ds - U(t), \tag{2.1}$$

where

$$\sigma^2 = \sigma_A^2 + \mu \text{ and } \beta(x) = \begin{cases} \beta_0 + \mu x & x \le 0 \\ \beta_0 + \gamma x & x > 0, \end{cases}$$

and approximates the centered and scaled number of customers in the system. The processes u and U in (2.1) approximate the scaled service rate adjustment and scaled cumulative number of rejected customers, respectively. When there is no control, i.e., when $u(t) = U(t) = 0$ for all t, the process in (2.1) becomes an Ornstein–Uhlenbeck process with infinitesimal drift $-\beta(x)$ and infinitesimal variance σ^2. The infinitesimal drift is linear in the state variable and depends on the base service rate μ and the abandonment distribution via $\gamma = F'(0)$. The constant β_0 in the infinitesimal drift denotes the *service grade* and approximates the capacity imbalance of the original queueing system.

We let Π denote the set of admissible control policies and assume that all $\pi := (u, U) \in \Pi$ satisfy the following assumptions:

(A1) There exists a weak solution to (2.1) and $\frac{E|X(t)|}{t} \to 0$ as $t \to \infty$.

(A2) The control process u is non-negative, progressively measurable, and locally integrable.

(A3) The control process U is non-negative, non-decreasing, and RCLL.

Assumptions (A1)–(A3) allow us to restrict attention to policies under which Eq. (2.1) is well-defined and the process $X(t)$ is non-explosive. This is important because assumptions (A2) and (A3) allow for $u(t) \ge 0$ when $X(t) < 0$, and for U to increase when $X(t) < 0$. Hence, assumption (A1) limits the drift that pushes the process away from zero. Note that if we also assume an upper bound on $u(t)$ when $X(t) < 0$ (see, for example equation (3.4) and the discussion following in [7]), and also assume that U increases only when $X \ge 0$, then assumption (A1) follows directly and is not necessary.

Next, we make the following assumption on the control cost function $c(\cdot)$:

(A4) The control cost $c(\cdot)$ is non-decreasing and continuous. It also satisfies $c(0) = 0$ and

$$\inf\left\{\frac{c(x)}{x} : x \in R, x \ge y\right\} \uparrow \infty \text{ as } y \uparrow \infty. \tag{2.2}$$

Then, we can define the function $\phi : \Re \to [0, \infty)$

$$\phi(y) := \sup_x \{yx - c(x)\} \tag{2.3}$$

and let $\psi(y)$ denote the smallest maximizer in (2.3). It can be verified (see for example [2]) that both $\phi(y)$ and $\psi(y)$ are non-decreasing functions and that $\phi'(y) = \psi(y)$. These functions will play a critical role in characterizing the optimal policy.

Next, we define the *effective waiting cost* $h_w := c_w + a\gamma > 0$. Alternatively, we can define $\bar{p} := \frac{h_w}{\gamma} = a + \frac{c_w}{\gamma}$ as the *effective abandonment cost*. Finally, we define the *holding cost* as

$$h(x) := h_i x^- + h_w x^+.$$

Then, the cumulative diffusion cost up to time t, under an admissible policy $\pi = (u, U) \in \Pi$, is given by

$$\xi_\pi(t) = \int_0^t (c(u(t)) + h(X(t)))\, dt + pU(t). \tag{2.4}$$

Our objective is to minimize the long run average (ergodic) cost, within the class of admissible policies, so we solve

$$\min_{\pi \in \Pi} \liminf_{t \to \infty} \frac{E[\xi_\pi(t)]}{t}. \tag{2.5}$$

Our first result is a verification lemma, which characterizes the minimum achievable cost for any admissible policy $\pi = (u, U) \in \Pi$. Later, in Theorem 1, we will show that the optimal policy indeed achieves this lower bound.

Lemma 1 (*Minimum Achievable Cost*)**.** *Suppose there exists a twice continuously differentiable function V having bounded first derivative, and a positive constant κ that satisfies the HJB equation*

$$\min\left\{\frac{\sigma^2}{2}V''(x) - \beta(x)V'(x) - \phi(V'(x)) + h(x) - \kappa,\ p - V'(x)\right\} = 0, \tag{2.6}$$

for all $x \in \Re$. Then, for any admissible control $\pi = (u, U) \in \Pi$ having associated cumulative cost $\xi_\pi(t)$ as given in (2.4),

$$\liminf_{t \to \infty} \frac{E[\xi_\pi(t)]}{t} \ge \kappa.$$

2.2. Diffusion control problem solution

First, we show that a solution to the HJB optimality equation (2.6) exists and that its specific behavior depends on the relationship between $\bar{p}$ and p.

Proposition 1 (*Existence of a Solution to HJB Equations*)**.** *There exists a twice continuously differentiable function having bounded derivative and constant κ that satisfy (2.6) as follows:*

(i) *If $p < \bar{p}$, there exists a positive $b^\star := \inf\{x \ge 0 : V'(x) = p\}$ such that*

$$\frac{\sigma^2}{2}V''(x) - \beta(x)V'(x) - \phi(V'(x)) + h(x) = \kappa, \tag{2.7}$$

for $x \in (-\infty, b^\star]$ and

$$\frac{\sigma^2}{2}V''(x) - \beta(x)V'(x) - \phi(V'(x)) + h(x) > \kappa, \tag{2.8}$$