Constrained Optimality of an M/G/1 Queueing System

Dong Jin Kim [1)]

Abstract

This paper studies constrained optimization of an M/G/1 queue with a server that can be switched on and off. One criterion is an average number of customers in the system and another criterion is an average operating cost per unit time, where operating costs consist of switching and running costs. With the help of queueing theory, we solve the problems of optimization of one of these criteria under a constraint for another one.

1. Introduction

This paper studies constrained optimization of an M/G/1 queuing system with a removable server. One criterion is the average number of customers in the system, and the other criterion is the average operating cost per unit time. The operating costs consist of switching costs and running costs.

Queuing systems with removable servers model service and production systems without inventories, with backorders, and with stochastic demand. For the M/G/1 model considered in this paper, orders arrive in accordance with a Poisson process, they are served sequentially, and processing times are i.i.d. random variables. The question is when the system should start and finish production(or service) if set- up or shut-down costs are non zero.

Single server queues with removable servers have been studied by many researchers such as Yadin and Naor, Heyman, Bell and many others, where they deal mainly with the optimization of either the average cost per unit time or the expected total discounted cost. Yadin and Naor introduced a notion of the N-policy implying that the server is on when the queue size reaches the number N and off when the system becomes empty.

For N=1,2,... and for $p \in (0,1]$, we say that the policy is a <p, N>policy if it prescribes the following actions: (i) Switch the server off when the system becomes empty, (ii) Switch the nonoperating server on if there are more than N customers in the system, (iii) If the server is off and the number of customers in the system becomes N, switch the server on with probability p and leave it off with probability (1-p), and (iv) Do not switch the server at other epochs.

We say that the policy is an <N, p>policy if it prescribes the following actions: (i) Switch the nonoperating server on when there are N customers in the system, (ii) When the system becomes empty, switch the server off with probability p and leave it on with

1) University of Maryland University College, Department of Mathematics

probability (1-p), and (iii) Do not switch the server at other epochs. The essential differences between these two policies is that for a <p, N>policy a decision maker selects actions randomly at an arrival epoch of an Nth customer who sees the server off, whereas for an <N, p>policy a decision maker chooses actions randomly at a service completion epoch when the system becomes empty.

We show that either the <p, N> policy or <N, p> policy is optimal for the constrained problem and find the explicit formulas for the optimal N and p.

The disadvantage of the above policies to apply to the real life situations is due to the fact that they are randomized which is a typical case for constrained sequential decision problems. To get around this difficulty of randomized policies we will construct equivalent non randomized policies, which are more practical to implement in applications.

2. Problem Formulation

We consider an M/G/1 queue. The Poisson arrival process has intensity λ, and service times are i.i.d. non negative random variables with distribution function G, which has a finite and positive mean $1/\mu$ and finite variance σ^2 . The system utilization factor is $\rho = \lambda/\mu$ and is always assumed to be less than unity. Let L be the average number of customers in an M/G/1 queue which is by Pollaczek-Khintchin formula, $L = \rho + \rho^2 (\sigma^2 \mu^2 + 1)/[2(1-\rho)]$. The cost structure is the operating costs, which contain the start-up cost C_1 , the shut-down cost C_2 , and the running costs D_1 per unit time when the server is on and D_2 per unit time when the server is off. Without loss of generosity, we assume that $D_1 > D_2$ and $C_1 + C_2 > 0$.

The server may be either on or off. If the server is off, it may be switched on at arrival epochs and if the server is on, it may be switched off at service completion epochs.

For a policy π, let the corresponding average number of customers in the system be $L(\pi)$ and the average operating cost per unit time be $C(\pi)$. We solve the following optimization problems: (1) Minimization of the average operating cost per unit time under the constraint that the average number of customers in the system does not exceed a given level and (2) Minimization of the average number of customers in the system under the condition that the average operating cost per unit time does not exceed a given level.

Now, we have the following problems:

Problem 1 : Minimize $C(\pi)$ subject to $L(\pi) \leq \alpha$

Problem 2 : Minimize $L(\pi)$ subject to $C(\pi) \leq \beta$, where α and β are given numbers.

3. Optimal policies for constrained problems

If Problems described in section 2 is feasible then there exists an optimal policy that is either <p, N>, or <M, p>, or 0-policy, where 0-policy is to keep the server on all the time. Let $D(N) = C(N+1) + R_2/(N+1)$, where $R_2 = (C_1 + C_2)\lambda(1-\rho)$ and $M = \min\{i : D(i) \leq D_1 , i = 2,3,\ldots\}$.

Theorem1: Suppose that Problem 1 is given. If $\alpha < L$, then the problem 1 is infeasible. If $\alpha \geq L$, then the problem is feasible and the optimal policy has the following form:

(i) If either $D_1 \geq C(1)$, or $D_1 \leq C(1)$ and $\alpha \geq L(M)$, then the <p, N>policy is optimal with $N=[2(\alpha-L)+1]$ and $p=(N+1)\{N+2(L-\alpha)\}/\{2(N+L-\alpha)\}$, where [x] is an integer part of x.
(ii) If $D_1 \leq C(1)$ and $L < \alpha \leq L(M)$, then the <M, p>policy is optimal, where $p=2(\alpha-L)/\{(M-1)(M-2(\alpha-L))\}$.
(iii)If $D_1 \leq C(1)$ and $\alpha = L$, then the 0-policy is optimal.

Theorem2: Suppose that Problem 2 is given. If $\beta < D_2$, then the problem is infeasible. If $D_2 \leq \beta \leq (1-\rho)D_2 + \rho D_1$, then the policy $\pi(*)$ to keep the server off at all time is optimal with $L(\pi(*))=\infty$. If $\beta > (1-\rho)D_2 + \rho D_1$, then the problem is feasible and the optimal policy has the following form:
(i) If either $D_1 \geq C(1)$, or $D_1 \leq C(1)$ and $\beta \leq C(M)$, then the <p, N>policy is optimal with $N=[R_2/\beta_1]$ and $p=(N+1)-R_2/\beta_1$, where $R_2 = (C_1 + C_2)\lambda(1-\rho)$and $\beta_1 = \beta - \{(1-\rho)D_2 + \rho D_1\}$
(ii) If $D_1 \leq C(1)$ and $C(M) \leq \beta \leq D_1$, then the <M, p>policy is optimal where $p=(\beta_1 - R_1)/\{R_2 - R_1 - \beta_1(M-1)\}$, and $R_1 = (D_1 - D_2)(1-\rho)$.
(iii) If $D_1 \leq C(1)$ and $\beta \geq D_1$, then the 0-policy is optimal.

4. Equivalent non randomized optimal policies

In this section we will describe two non randomized policies and show that they are equivalent to the above optimal policies of theorems 2 and 3.

For some $t \geq 0$, [t, N] policy prescribes switching the nonoperating server on either when time t passed after the queue reaches the size N or when the queue reaches the size (N+1), whatever occurs first, and the server should be switched off when the system becomes empty. At all other epoches, the state of the server should remain the same.

An [N, t] policy prescribes switching the server off in t units of time after the system becomes empty, if there is no arrival during these t units of time, and the nonoperating server should be switched on when the number of customers in the system becomes N. If a customer arrives within t units of time after the system becomes empty, the server remains on and it serves the customers until the busy period ends.

Theorem 3: For N=1,2,... and $p \in (0,1]$, we set $t=(-1/\lambda)\ln p$. Then
(i) <p, N> policy and the [t, N] policy have the same performance
(ii) <N, p> policy and the [N, t] policy have the same performance

References

Altman, E & Nain, P(1993), Optimal control of an M/G/1 queue with vacations, IEEE Transactions on Automatic Control 38: 1766-1775

Altman, E. & Shwartz, A(1993), Timesharing policies for controlled Markov chains, Operations Research 41: 1116-1124

Bell, C. E.(1971), Characterization and computation of optimal policies for operating an M/G/1 queueing system with removable server, Operations Research 19: 208-218

Feinberg, E. A.(1994), Constrained semi-Markov decision process with average rewards, ZOR-Mathematical Methods of Operations Research 39:257-288

Feinberg, E. A & Kim, D. J.(1994), Optimal switching policies for M/G/1 queues with two performance criteria, Operations Research proceedings. Berlin:227-232

Feinberg, E. A & Kim, D. J.(1996), Bicriterion optimization of and M/G/1 queue with a removable server, Probability in the Engineering and Informational Sciences 10: 57-73

Heyman, D(1968), Optimal operating policies for M/G/1 queueing systems, Operations Research 16:362-383

Lee, H.S. & Srinivasan, M.M.(1989), Control policies for the M/G/1queueing system, Management science 35: 708-721

Yadin, M & Naor, P(1963), Queueing systems with a removable service station, Operations research Quarterly 14: 393-405