Scheduling in a queuing system with impatience and setup costs

Alain Jean-Marie, Emmanuel Hyon

Scheduling in a queuing system with impatience and setup costs

Alain Jean-Marie [*] Emmanuel Hyon [†]

May 10, 2010

Abstract

We consider a single server queue in discrete time, in which customers must be served before some limit sojourn time of geometrical distribution. A customer who is not served before this limit leaves the system: it is impatient. The fact of serving customers and the fact of losing them due to impatience induce costs. The purpose is to decide when to serve the customers so as to minimize costs. We use a Markov Decision Process with infinite horizon and discounted cost. We establish the structural properties of the stochastic dynamic programming operator, and we deduce that the optimal policy is of threshold type. In addition, we are able to compute explicitly the optimal value of this threshold according to the parameters of problem.

Keywords: Scheduling, queuing system, impatience, deadline, optimal control, Markov Decision Processes

[*]INRIA and LIRMM, CNRS/Université Montpellier 2, 161 Rue Ada, F-34392 Montpellier, ajm@lirmm.fr

[†]Université Paris Ouest Nanterre la Défense, LIP6, UPMC 4 place Jussieu, F-75252 Paris Cedex, Emmanuel.Hyon@u-paris10.fr

1 Introduction

In this paper we are interested in the optimal control of *services* in a queuing system with impatient customers (or, equivalently said, customers with deadlines). The set-up of customer services, the storage of the customers in the queue as well as their departure from the queue due to impatience (called "losses" in the remainder of this paper) induce some costs and it has to be decided when to begin the service in order to minimize these costs.

Controlled queuing models, deterministic as well as stochastic, have been largely studied in the literature since their application fields are numerous: networking (see [1] and references therein), resources allocation (see [8] and references therein) to quote just a few. Nevertheless most of these works do not consider impatient customers but rather losses due to buffer overflow. Yet, the phenomenon of impatience, associated with deadlines or "timeouts", has become non negligible in several fields: cellular communication networks [3], call center [11], revenue management or reservations problems (see [13] for discrete finite horizon problems), real-time systems etc.

The literature does features papers on the performance evaluation of queues with impatience, but relatively few on optimal control of such queues. One branch, represented by [4], is concerned with finding the optimal scheduling algorithm so as to minimize deadline misses. But, this question of the order of service is not relevant here, since we assume deadlines with memoryless distribution. Another direction is to consider the optimal routing between several queues [9], still in order to minimize average deadline overrun.

Since our focus is on the control of service in the queue, the most relevant source of ideas should be the literature on controlling batch services (and indeed, our longer-term objective is to solve the same problem but with batch service – we come back to this topic below). The problem of optimally controlling a batch server in a queue (without impatience) has been addressed in [5] and [12] (see also the references therein). Its resolution is based on the Markov Decision Process (MDP) formalism, and goes through establishing some structural properties of the value function and the dynamic programming operator. This allows to deduce some properties of the optimal policy, which in turn implies that the solution is a threshold (or *control limit*) policy.

Unfortunately, it appears that extending the techniques developed in [12] to queues with impatience is not straightforward. Indeed, it has been noted in [10] (quoted in [9]) that impatience tends to destroy the structural properties that are commonly used for proving the optimality of threshold policies. As it has been underlined in [6], these structural properties are most often a key point in the study of admission control policy and optimality proofs. In this paper, we show that structural properties exist despite the presence of losses.

We now turn to the description of the model and the results. In this paper, we consider a slotted queueing model, where the slot is the time unit. Customers are assumed to arrive at the beginning of each slot. The number of arrivals at each slot is assumed to be an i.i.d. sequence of random variables. These arrivals are stored in an infinite buffer in which they wait for to be admitted in the server to be processed. This admission decision is made by a controller. The service duration is assumed to be equal to the duration of a slot. Customers are impatient: while they are in the buffer, they can leave spontaneously the system with fixed probability independently from the past and from each other. On the other hand, customers admitted in service are not impatient anymore.

The beginning of a new service induces a setup cost, and holding customers in the buffer as well as losing customers due to impatience also induce a cost. The control objective is to minimize discounted costs accumulated over time; we choose the infinite-horizon, discounted cost criterion for this purpose.

In order to find the optimal control policy, we adopt the structural approach for Markov Decision

Processes, as described for instance in Puterman [14]. Indeed, given our assumptions, the problem can be set as a MDP with a one-dimensional, countable state space (the number of customers present) and a discrete 0/1 action space.

We prove that the Bellman and optimized Bellman operators enjoys some structural properties: propagation of monotonicity and convexity, submodularity for increasing convex functions. In turn, this yields that the value function of the decision process associated with the problem is increasing convex. Therefore the structural results of Puterman allow to conclude that the optimal policy is a threshold policy. The detailed proofs can be found in [7].

Furthermore, using a sample path comparison between policies with different values of the threshold, we explicitly compute the threshold value as a function of the parameters, and we conclude that the optimal policy is actually "*always serve*" or "*never serve*", based on a simple criterion derived from the cost parameters and the loss rate.

A so simple result would suggest that using simpler methods should be more efficient to prove this result. Nevertheless, this does not seem to be the case. We consider several alternatives, and explain why they cannot be applied so easily. In particular, sample path arguments are not easy to work with: we exhibit a counter example in which direct proof with interchange arguments fails. One the other hand, the proof which consists in verifying directly that the value of the "always serve" policy solves the Bellman equation, requires the same type of structural analysis that we perform in general.

As a conclusion, we turn to the question of generalizing the result (optimality of a threshold policy) to batch services with batch sizes B larger than 1. While some properties of the stochastic programming operator are conserved, we provide example in which the essential submodularity property vanishes (for a value of B equal to 5). It turns out that the stochastic dynamic operator which governs the evolution of the batch queue lacks elementary stochastic increasingness and convexity properties. In addition, the fact that customers are impatient prevents the use of several notions of "K-convexity" (see *e.g.* [2] and [12]) already used for the study of batch problems.

On the other hand, no experimental evidence has contradicted, so far, the possibility that the optimal control still be of threshold type (similar issues are equally addressed in [15] for the dual problem of batch admission control). The challenge of further research on the topic will therefore be to find the appropriate properties that can be propagated by the dynamic programming operator in this case.

References

[1] E. Altman. *Handbook of Markov Decision Processes Methods and Applications*, chapter Applications of Markov Decision Processes in Communication Networks : a survey. Kluwer, 2001.

[2] E. Altman and G. Koole. On submodular value functions and complex dynamic programming. *Stochastic Models*, 14:1051–1072, 1998.

[3] J. R. Artalejo and V. Pla. On the impact of customer balking, impatience and retrials in telecommunication systems. *Computers and Mathematics with Applications*, 57(2):217–229, 2009.

[4] P. P. Bhattacharya and A. Ephremides. Optimal scheduling with strict deadlines. *IEEE Trans. Automatic Control*, 34(7):721–728, July 1989.

[5] R. K. Deb and R. F. Serfozo. Optimal control of batch service queues. *Advances in Applied Probability*, 5(2):340–361, 1973.

[6] P. Glasserman and D. Yao. *Monotone Structure in Discrete-Event Systems*. Wiley, 1994.

[7] E. Hyon and A. Jean-Marie. Scheduling in a queuing system with impatience and setup costs. Technical Report RR-6881, INRIA, 2009.

[8] A. J. Kleywegt and J. D. Papastavrou. The dynamic and stochastic knapsack problem. *Operations Research*, 46:17–35, 1998.

[9] Y. L. Kocaga and A. R. Ward. Admission control for a multi-server queue with abandonment. Manuscript, July 2009.

[10] G. Koole. Monotonicity in markov reward and decision chains: Theory and applications. *Foundation and Trends in Stochastic Systems*, 1(1), 2006.

[11] G. Koole and A. Mandelbaum. Queueing models of call centers: An introduction. *Annals of Operations Research*, 113:41–59, 2002.

[12] K. P. Papadaki and W. B. Powell. Exploiting structure in adaptative dynamic programming algorithms for a stochastic batch service problem. *European Journal of Operational Research*, 142:108–127, 2002.

[13] J. D. Papastavrou, S. Rajagopalan, and A. J. Kleywegt. The dynamic and stochastic knapsack problem with deadlines. *Management Science*, 42(12):1706–1718, 1996.

[14] M. Puterman. *Markov Decision Processes Discrete Stochastic Dynamic Programming*. Wiley, 2005.

[15] E. B. Çil, E. L. Örmeci, and F. Karaesmen. Structural results on a batch acceptance problem for capacitated queues. *Mathematical Methods of Operations Research*, 66:263–274, 2007.