Journal of Southeast University (English Edition) Vol. 35, No. 2, pp. 252 - 256 June 2019 ISSN 1003—7985

Appointment scheduling with customer impatience based on operating cost model

Song Minshan[1,2] Zhang Yulin[1]

([1]School of Economics and Management, Southeast University, Nanjing 211189, China)
([2]School of Science, Jiangsu University of Science and Technology, Zhenjiang 212003, China)

Abstract: An appointment scheduling problem is studied with the consideration of customer impatience. On the assumption that both the time of leaving queue and the time of service are exponentially distributed, in order to minimize the joint cost, the optimal appointment schedule of the fixed number of customers is studied. The joint cost function is composed of customers' expected delay time and service availability time. The expected delay time of each customer in the queue is recursively computed in terms of customer interarrival time. Furthermore, the effect of impatience on the optimal schedule as well as the total operating cost is studied. The results show that as the impatience rate increases, the optimal interarrival time becomes shorter and the interarrival time of the last few customers gradually approaches that of the customers in the middle. In addition, impatient behaviors can increase the joint cost.

Key words: appointment; customer impatience; queueing theory; single server

DOI: 10.3969/j.issn.1003-7985.2019.02.016

Appointment scheduling has been extensively studied over the past sixty years[1]. Wang[2] studied the optimal arrival time intervals for the appointment in a single-server system with exponential service time. He found that the optimal interarrival time presents a dome pattern. Kuiper et al.[3] investigated the problem with general service time distribution and confirmed the dome pattern of the optimal schedule. Zhang et al.[4] studied the case of unknown random service time distribution. They pointed out that different optimization objectives lead to different time interval structures. Some appointment scheduling studies considered customer behaviors such as no-show and unpunctuality[5-6]. The results indicate that customer behaviors have certain impact on the optimal interarrival time. Those studies focused on appointment scheduling without involving impatient behavior, though impatience is common in the appointment system. Due to the uncertainty of service time, customers who arrived may find that the previous customer is still receiving service, so they have to queue; however, they will leave if they are not served within a certain time after joining the queue. Such queueing systems include appointment systems in banks, hospitals, and car maintenance etc.

Meanwhile, there is much research on the waiting time of the queueing system with impatient behavior. Movaghar[7] investigated the waiting time for the multi-server M/M/s queue with impatience. Choi et al.[8] considered the M/M/1 queue with impatience and customer priority. Daley[9] studied the waiting time of the G/G/1 queue with impatience. Choi et al.[10] considered performance measures including the waiting time of the MAP/M/c queue with impatient customers. Sakuma and Takine[11] studied the waiting time of a multi-class M/PH/1 queue. Wang and Wu[12] considered the waiting time for a M/M/1 queue with constant impatience and the last-in first-out rule.

As the service environment is not always constant, the queuing theory can bring useful insights to appointment system design. Thus, in this paper, an $S(n)/M/1$ queueing model is established for the appointment scheduling problem with customer impatience. The expected delay time (waiting time plus service time) is calculated with impatience, when the interarrival time is no longer subjected to a distribution but decision variables. Moreover, the impact of impatient behaviors on appointment scheduling is investigated.

1 Model

Assuming that there is a single server providing services in the appointment system, N is denoted as the number of customers to be scheduled. The service time for each customer is independent and identically distributed (i.i.d.) following an exponential distribution. Customers may leave the system due to impatience. That is, if an impatient customer has a deadline, he/she will keep the deadline until the beginning of service. When the waiting time of an impatient customer exceeds his/her deadline, the customer will leave the system. The

Received 2017-12-07, **Revised** 2019-04-29.

Biographies: Song Minshan(1986—), female, Ph. D. candidate; Zhang Yulin(corresponding author), male, doctor, professor, zhangyl@seu.edu.cn.

Foundation items: The National Natural Science Foundation of China (No. 71671036), the Scientific Innovation Research of Graduate Students in Jiangsu Province (No. KYLX_0211).

Citation: Song Minshan, Zhang Yulin. Appointment scheduling with customer impatience based on operating cost model [J]. Journal of Southeast University (English Edition), 2019, 35(2): 252 - 256. DOI: 10.3969/j.issn.1003-7985.2019.02.016.

customers' deadlines are assumed i. i. d. following an exponential distribution with parameter θ, and θ is the impatient rate. Usually, the impatient rate θ will not exceed the service rate μ. Thus, it is assumed that $\theta < \mu$. Customers will arrive punctually by appointment and there is no rescheduling. There is a fixed number of customers to be scheduled at the beginning. The objective is to determine a schedule for customers, thus minimizing the sum of expected delay time (waiting time plus service time) cost of customers and expected availability cost of the server. Customers' delay cost per unit time is denoted as c_w and the service cost per unit time is denoted as c_s.

The interarrival time between the i-th and $(i+1)$-th arrival is denoted as x_i. The objective function is a combination of expected delay cost and expected server availability cost. The goal is to determine the interarrival vector $\bar{\boldsymbol{x}} = \{x_1, x_2, \ldots, x_{N-1}\}^{\mathrm{T}}$ by minimizing the objective cost function. The general form of the problem is

$$\min_{x} \Phi_1(\bar{\boldsymbol{x}}, \bar{\boldsymbol{w}}) = c_w \boldsymbol{e}\bar{\boldsymbol{w}} + c_s(\boldsymbol{e}\bar{\boldsymbol{x}} + \boldsymbol{e}_N\bar{\boldsymbol{w}}) \quad \text{s. t.} \quad \bar{\boldsymbol{x}} \geqslant 0 \tag{1}$$

where w_i denotes the expected delay time of customer i; $\bar{\boldsymbol{w}} = \{w_1, w_2, \ldots, w_N\}^{\mathrm{T}}$ is the vector of expected delay time; $\boldsymbol{e}$ is a row vector of order N with all the elements equal to one; $\boldsymbol{e}_N$ is a row vector of dimension N with the N-th element equal to one and all others equal to zero. The first term in the objective function is the total customer delay cost, while the second term is the total server availability cost, which is the departure time of the last customer. Without losing generality, the objective function can be simplified by dividing $(c_w + c_s)$. Therefore, the simplified objective can be written as

$$\min_{\bar{x}} \Phi(\bar{\boldsymbol{x}}, \bar{\boldsymbol{w}}) = \alpha\boldsymbol{e}\bar{\boldsymbol{w}} + (1-\alpha)(\boldsymbol{e}\bar{\boldsymbol{x}} + \boldsymbol{e}_N\bar{\boldsymbol{w}}) \quad \text{s. t.} \quad \bar{\boldsymbol{x}} \geqslant 0 \tag{2}$$

where $\alpha = \dfrac{c_w}{c_w + c_s} (0 \leqslant \alpha \leqslant 1)$ is the relative cost factor of customer delay.

2 Expected Delay Time

The purpose of this section is to establish the structure of the expected delay time w_i, and w_i is represented as a function of interarrival time vector $\bar{\boldsymbol{x}}$. First, we need to give the departure rate for a system with impatient customers. In a stationary queuing system with m servers, the probability that a customer misses his/her deadline when the total customer number n in the system is

$$r_n = \begin{cases} 0 & \text{if } n \leqslant m \\ (n-m)\theta & \text{if } n > m \end{cases} \tag{3}$$

Assuming that there is a single server in the system with service rate μ, the departure rate including impatience is $\mu_n = \mu + r_n$, where n is the number of customers in the system and $r_n = (n-1)\theta$.

When the i-th customer arrives and finds that there are $n(0 \leqslant n \leqslant N-1)$ customers in the system, his/her delay time will be the absorption time of the departure process with $n+1$ phase. The absorption state is the state when all $n+1$ customers leave the system. In this case, the distribution of delay time for customer i is the phase-type (PH) distribution of absorption time.

Let $\bar{\boldsymbol{p}}_i = \{p_i(0), p_i(1), \ldots, p_i(N-1)\}$ be a row vector. $p_i(j)$ represents the probability that there are j customers in the system just before the i-th arrival. Note that $p_i(j) = 0$ for $j \geqslant i$.

Proposition 1 The delay time distribution is described by $(\bar{\boldsymbol{p}}_i, \boldsymbol{T})$, where

$$\boldsymbol{T} = \begin{bmatrix} -\mu & & & & \\ \mu+\theta & -(\mu+\theta) & & & \\ & \ddots & & & \\ & & -[\mu+(N-2)\theta] & & \\ & & \mu+(N-1)\theta & -[\mu+(N-1)\theta] \end{bmatrix} \tag{4}$$

According to the conclusion of Neuts[13], the n-th moment of the PH distribution $(\bar{\boldsymbol{p}}_i, \boldsymbol{T})$ is given as

$$E[x^n] = (-1)^n n! \bar{\boldsymbol{p}}_i \boldsymbol{T}^{-n} \boldsymbol{e}' \tag{5}$$

Let $n = 1$, and the expected delay time for customer i can be derived directly by Eq. (5).

$$w_i = E[W_i] = -\bar{\boldsymbol{p}}_i \boldsymbol{T}^{-1} \boldsymbol{e}' \tag{6}$$

Next, the probability row vector $\bar{\boldsymbol{p}}_i$ is derived. The recursive formula for the probability row vector $\bar{\boldsymbol{p}}_i$ is obtained by the following proposition.

Proposition 2 The probability row vector for customer i is the multiplication of the probability row vector for customer $i-1$ and departure matrix $\boldsymbol{D}(x_{i-1})$.

$$\bar{\boldsymbol{p}}_i = \bar{\boldsymbol{p}}_{i-1} \boldsymbol{D}(x_{i-1}) \qquad i > 1 \tag{7}$$

where $\bar{\boldsymbol{p}}_i$ is the probability vector and the initial vector $\bar{\boldsymbol{p}}_1 = \{1, 0, \ldots, 0\}$. The description of departure matrix is

$$\boldsymbol{D}(x) = \begin{bmatrix} s_0(x) & d_{10}(x) & 0 & 0 & \cdots & 0 \\ s_1(x) & d_{21}(x) & d_{20}(x) & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ s_{N-3}(x) & d_{N-2,N-3}(x) & d_{N-2,N-4}(x) & d_{N-2,N-5}(x) & \cdots & 0 \\ s_{N-2}(x) & d_{N-1,N-2}(x) & d_{N-1,N-3}(x) & d_{N-1,N-4}(x) & \cdots & d_{N-1,0}(x) \\ 0 & 0 & 0 & 0 & \cdots & 1 \end{bmatrix} \tag{8}$$

where $d_{i,j}(x)$ represents the probability that there are j departures in time interval x as i customers who are at the beginning of x. Furthermore, $s_i(x) = 1 - \sum_{j=0}^{i} d_{i+1,j}(x)$.

The proposition can be directly obtained by the definition of probability vector $\bar{\boldsymbol{p}}_i$ and departure matrix $\boldsymbol{D}(x)$. Note that at any time interval $[t_{i-1}, t_i)$, there is only one entry at the initial time t_{i-1}, and the transient process at time interval $[t_{i-1}, t_i)$ can be regarded as a pure death process. With such a recognition, the elements in the departure matrix $\boldsymbol{D}(x)$ can be represented as follows.

Proposition 3 The elements in the departure matrix are

$$d_{i,j}(x) = \begin{cases} \prod_{h=i-j-1}^{i} \mu_h \sum_{m=i-j}^{i} A_m \mathrm{e}^{-\mu_m x} & j > 0 \\ \mathrm{e}^{-\mu_i x} & j = 0 \end{cases} \tag{9}$$

where $A_m = \dfrac{1}{\prod_{k=i-j,\, k\neq m}^{i} (\mu_k - \mu_m)}$ and $\mu_i = \mu + (i-1)\theta$.

Proof There is no entry at any time interval x_{n-1}. Let $\pi_i(t)$ denote the probability that there are i persons at time t in time interval x_{n-1}. In this case, the transfer process can be regarded as the pure death process with the transition rate $\pi_{i,i-1}(h) = \mu_i h + o(h)$. The corresponding transition matrix is

$$\boldsymbol{Z} = \begin{bmatrix} 0 & 0 & & & & \\ \mu_1 & -\mu_1 & & & & \\ & \mu_2 & -\mu_2 & & & \\ & & & \ddots & & \\ & & & & \mu_m & -\mu_m \end{bmatrix} \tag{10}$$

It is assumed that there are m persons at the initial time, which means that $\pi_m(0) = 1$. We can write the corresponding differential equation as

$$\left.\begin{aligned} \pi'_m(t) &= -\mu_m \pi_m(t) \\ \pi'_j(t) &= -\mu_j \pi_j(t) + \mu_{j+1} \pi_{j+1}(t) \quad 1 \leqslant j < m \end{aligned}\right\} \tag{11}$$

We can first derive the boundary solution as

$$\pi_m(t) = \mathrm{e}^{-\mu_m t} \tag{12}$$

Eq. (12) corresponds to $d_{m,0}(t)$ in the departure matrix.

Using the L-S transform, Eq. (11) can be transferred as

$$\left.\begin{aligned} \pi_m(s) &= \frac{1}{s + \pi_m(s)} \\ (s+\mu_j)\pi_j(s) &= \mu_{j+1}\pi_{j+1}(s) \qquad 1 \leqslant j < m \end{aligned}\right\} \tag{13}$$

Then, do the inverse L-S transform for $\pi_j(s)$. The probability that there are j persons in the system at time t with initial m people in the system is

$$\pi_j(t) = (\mu_{j+1} \ldots \mu_m) \sum_{i=j}^{m} A_i \mathrm{e}^{-\mu_i t} \tag{14}$$

where $A_i = \dfrac{1}{(\mu_j - \mu_i) \ldots (\mu_{i-1} - \mu_i)(\mu_{i+1} - \mu_i) \ldots (\mu_m - \mu_i)}$

Eq. (14) corresponds to $d_{m,m-j}(t)$ in the departure matrix.

Finally, the expected delay for customer i can be derived by previous propositions and it is presented in Theorem 1.

Theorem 1 The expected delay for customer i is

$$w_i = E[W_i] = -\bar{\boldsymbol{p}}_1 \prod_{m=1}^{i-1} \boldsymbol{D}(x_m) \boldsymbol{T}^{-1} \boldsymbol{e}' \tag{15}$$

Theorem 1 shows that the expected delay for customer i can be represented by the time interval vector $\bar{\boldsymbol{x}}$.

In order to minimize the objective cost function, the derivative is taken with respect to the interarrival vector $\bar{\boldsymbol{x}}$ by the chain rule.

$$\frac{\mathrm{d}\Phi(\bar{\boldsymbol{x}}, \bar{\boldsymbol{w}})}{\mathrm{d}\bar{\boldsymbol{x}}} = \frac{\partial \Phi(\bar{\boldsymbol{x}}, \bar{\boldsymbol{w}})}{\partial \bar{\boldsymbol{w}}} \cdot \left[\frac{\partial \bar{\boldsymbol{w}}}{\partial \bar{\boldsymbol{x}}}\right]^{\mathrm{T}} + \frac{\partial \Phi(\bar{\boldsymbol{x}}, \bar{\boldsymbol{w}})}{\partial \bar{\boldsymbol{x}}} \tag{16}$$

The optimal interarrival vector $\bar{\boldsymbol{x}}^*$ can be obtained by setting the first-order derivative equal to zero vector.

As it is difficult to obtain the explicit solution for the optimal interarrival time vector, the numerical solution calculated by software Matlab is used as an alternative. The sequential quadratic programming method is used in order to find the optimal interarrival time vector.

3 Results

In this section, the optimal interarrival time for impatient customers and the influence of impatience on an optimal schedule are investigated. The number of customers to be scheduled, the relative cost factor for the delay time and the service rate are set to be 10, 0.5 and 1, respectively, in the numerical simulation.

3.1 Optimal interarrival time for impatient customers

In the numerical simulation, the sequential quadratic programming method is used to find the optimal schedule that minimizes the objective function (see Fig. 1). The horizontal axis is the sequence of interarrival numbers, and the vertical axis is the interarrival time. The results show that the optimal schedule with impatience still has a "dome" pattern, i.e., the first few and the last few cus-

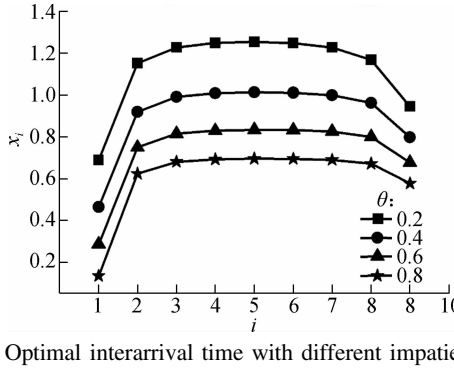

Fig. 1 Optimal interarrival time with different impatience rates ($N = 10, \alpha = 0.5, \mu = 1$)

tomers have almost zero interarrival time and the other customers have almost the same interarrival time. However, as the impatience rate increases, the shape of the "dome" becomes less conspicuous. Furthermore, the optimal interarrival time for each customer increases when the impatience rate decreases.

An apparent explanation is that when the impatience rate increases, the probability that a customer leaves the system in unit time increases; therefore, an optimal schedule should cut down the interarrival time to compensate for the time waste induced by the impatient behavior.

3.2 Impact of impatience

The optimal schedule without impatience has a longer interarrival time than the schedule with impatience (see Tab. 1). For the schedule with impatience, customers have the possibility to leave without getting service, leaving the system empty or reducing the waiting time for the later customers. The shortened interarrival time of the optimal schedule will reduce the waste time caused by impatience. As the impatient rate decreases, the optimal interarrival time with impatience is shorter than that without impatience.

Tab. 1 Comparison on optimal schedule under different θ

Interarrival time	Without impatience	$\theta=0.2$	$\theta=0.4$	$\theta=0.6$	$\theta=0.8$
x_1	1.005 990 268	0.689 103 251	0.464 669 100	0.285 494 301	0.133 782 000
x_2	1.509 248 190	1.152 831 013	0.919 941 158	0.749 957 427	0.623 465 733
x_3	1.589 487 454	1.228 351 533	0.990 729 262	0.815 130 305	0.680 378 432
x_4	1.611 149 100	1.249 757 684	1.008 812 368	0.830 056 489	0.692 506 749
x_5	1.609 995 736	1.254 446 875	1.013 317 737	0.833 746 973	0.695 449 375
x_6	1.590 836 669	1.248 796 207	1.011 130 311	0.832 737 211	0.694 897 396
x_7	1.545 477 937	1.228 163 201	0.999 771 567	0.825 650 328	0.690 006 680
x_8	1.441 725 780	1.169 019 379	0.962 336 843	0.800 174 243	0.671 507 939
x_9	1.125 311 506	0.945 559 816	0.798 307 965	0.676 727 767	0.576 514 163

Next, the influence on the cost formed by impatience is considered. The cost gaps between optimal schedules with and without impatience are shown in Fig. 2.

It is clear from Fig. 2 that ignoring impatient behavior will lead to an increasing cost for the system, and the cost increases as the impatient rate increases. An appointment system should take the impatient behavior into consideration when scheduling the customers.

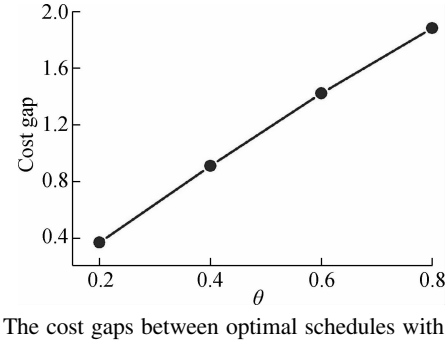

Fig. 2 The cost gaps between optimal schedules with and without impatience ($N=10, \alpha=0.5, \mu=1$)

4 Conclusion

An appointment scheduling for impatient customers is studied in this article. The optimal interarrival time to the fixed number of impatient customers will minimize the joint costs of expected delay and expected service availability. The delay time (waiting time plus service time) distribution for the $S(n)/M/1$ queue is investigated, and the recursive expression of expected delay time with impatience is given. Furthermore, the optimal schedule is calculated numerically by the sequential quadratic programming method. Numerical results of the optimal schedule are investigated. The "dome" pattern of optimal schedule gradually becomes less obvious as the impatience rate increases, and the higher the impatience rate, the shorter the interarrival time. Moreover, a comparison of the systems with and without impatience is given. The optimal schedule for systems with impatience has shorter arrival time intervals for each customer than the systems without impatience, thus avoiding the time waste induced by impatient behavior. Additionally, it is noted that a large cost will be borne if impatience is overlooked. In such a case, customer impatience should not be ignored.

References

[1] Bailey N T J. A study of queues and appointment systems in hospitalout-patient departments, with special reference to waiting-times[J]. *Journal of the Royal Statistical Society: Series B (Methodological)*, 1952, **14**(2): 185 − 199. DOI: 10.1111/j.2517-6161.1952.tb00112.x.

[2] Wang P P. Static and dynamic scheduling of customer arrivals to a single-server system[J]. *Naval Research Logistics*, 1993, **40**(3): 345 − 360. DOI: 10.1002/1520-6750(199304)40:3345::aid-nav3220400305>3.0.co;2-n.

[3] Kuiper A, Kemper B, Mandjes M. A computational approach to optimized appointment scheduling[J]. *Queueing Systems*, 2015, **79**(1): 5 − 36. DOI: 10.1007/s11134-014-9398-6.

[4] Zhang Y L, Shen S Q, Erdogan S A. Distributionally robust appointment scheduling with moment-based ambiguity set[J]. *Operations Research Letters*, 2017, **45**(2): 139 − 144. DOI: 10.1016/j.orl.2017.01.010.

[5] Liu N. Optimal choice for appointment scheduling window under patient no-show behavior[J]. *Production and*

Operations Management, 2016, **25**(1): 128 – 142. DOI: 10.1111/poms.12401.
[6] Jiang B W, Tang J F, Yan C J. A stochastic programming model for outpatient appointment scheduling considering unpunctuality[J]. *Omega*, 2019, **82**: 70 – 82. DOI: 10.1016/j.omega.2017.12.004.
[7] Movaghar A. On queueing with customer impatience until the beginning of service[J]. *Queueing Systems*, 1998, **29**(2/3/4): 337 – 350.
[8] Choi B D, Kim B, Chung J. M/M/1 queue with impatient customers of higher priority[J]. *Queueing Systems, Theory and Applications*, 2001, **38**(1): 49 – 66. DOI: 10.1023/A:1010820112080.
[9] Daley D J. General customer impatience in the queue GI/G/1[J]. *Journal of Applied Probability*, 1965, **2**(1): 186 – 205. DOI: 10.2307/3211884.
[10] Choi B D, Kim B, Zhu D B. MAP/M/c queue with constant impatient time[J]. *Mathematics of Operations Research*, 2004, **29**(2): 309 – 325. DOI: 10.1287/moor.1030.0081.
[11] Sakuma Y, Takine T. Multi-class M/PH/1 queues with deterministic impatience times[J]. *Stochastic Models*, 2017, **33**(1): 1 – 29. DOI: 10.1080/15326349.2016.1197778.
[12] Wang F, Wu X Y. On the waiting time for a M/M/1 queue with impatience [EB/OL]. (2017-04-06)[2019-05-23]. https://arxiv.org/abs/1704.01709.
[13] Neuts M F. *Matrix-geometric solutions in stochastic models: An algorithmic approach*[M]. Courier Corporation, 1994: 1 – 36.

基于运作成本模型的不耐烦顾客预约时间安排

宋皎珊[1,2] 张玉林[1]

([1] 东南大学经济管理学院, 南京 211189)
([2] 江苏科技大学理学院, 镇江 212003)

摘要: 研究了预约中考虑不耐烦行为的顾客到达时间安排问题. 假设顾客不耐烦时长及服务时长服从指数分布, 以最小化联合成本为目标,研究了固定数量顾客的最优预约到达时间安排. 联合成本函数由顾客期望延迟时间和服务时间构成. 根据顾客的预约到达时间间隔,推导出了每个顾客期望延迟时间的递推表达式. 进一步研究了不耐烦行为对于最优预约到达时间安排及总成本函数的影响. 结果表明,随着不耐烦率的增加,最优预约到达时间间隔越来越小,并且最后几名顾客的预约时间间隔逐渐接近中间的顾客, 且不耐烦行为会造成联合成本的显著增加.

关键词: 预约; 顾客不耐烦; 排队论; 单服务台
中图分类号: F224