

ON OPTIMAL AND EQUILIBRIUM RETRIAL RATES IN A QUEUEING SYSTEM

REFAEL HASSIN

*Department of Statistics and Operations Research
Tel Aviv University
69978 Tel Aviv, Israel*

MOSHE HAVIV

*Department of Econometrics
The University of Sydney
Sydney, New South Wales 2006, Australia
and
Department of Statistics
The Hebrew University of Jerusalem
91905 Jerusalem, Israel*

We discuss socially optimal and equilibrium retrial rates in a single-server queueing model. We extend known results, compare the two rates, and suggest ways to impose tolls on retrials (or rebates for waiting) in order to equate the equilibrium rate with the socially optimal one.

1. INTRODUCTION

We consider the following single-server queue model: Calls arrive according to a Poisson process with average rate λ per time unit. Service requirements are i.i.d. with mean τ and finite variance σ^2 . Let $\rho = \lambda\tau$ be the system's utilization factor and denote $\sigma^2 + \tau^2$ by S^2 . In the case that service is exponential, we denote τ by $1/\mu$, where μ is the service rate. In this case, $S^2 = 2/\mu^2$. If upon initiation of a call the server is busy, the call is repeated later. Between retrials, the call is said to be *in orbit*. The times between retrials are independent and expo-

nentially distributed with an expected value of $1/\theta$ (θ is the *retrial rate*). Each retrial costs c and the cost of waiting is w per unit of time.

There are two questions of interest here. The first is what is the social optimal retrial rate, denoted θ^* . This is the rate that minimizes the long-run total expected cost paid by a customer (equivalently, it minimizes the expected cost per customer). This question was answered by Kulkarni [3]. See Eq. 1, later. The second question is what is the Nash equilibrium retrial rate. A retrial rate defines a Nash equilibrium if given that it is used by all customers then an individual minimizes its own expected cost by using this rate itself. This question was recently answered by Elcan [1], assuming exponential service distribution.

Here we compute the Nash equilibrium retrial rate directly from results contained in Kulkarni [3]. Moreover, we remove the assumption that the service distribution is exponential. In addition, we make some observations on this rate and on ways to impose a toll on retrials (or a rebate proportional to the time in orbit) such that the resulting Nash equilibrium rate coincides with the social optimal rate. We emphasize that such a toll is justified in our model even when retrials do not lead to more consumption of resources, and the only actual cost directly involved with them is the cost inflicted on the retrying customers.

Kulkarni [2] showed that the expected time in orbit per call equals

$$W = \frac{\rho}{1 - \rho} \left(\frac{1}{\theta} + \frac{1}{2} \frac{S^2}{\tau} \right),$$

and, thus, the average cost per call is

$$(w + c\theta)W = \frac{\rho}{1 - \rho} \left(\frac{cS^2}{2\tau} \theta + \frac{w}{\theta} \right) + \frac{\rho}{1 - \rho} \left(\frac{wS^2}{2\tau} + c \right).$$

This cost is minimized at

$$\theta^* = (1/S)\sqrt{2w\tau/c}. \quad (1)$$

In particular, for the case of exponential service with $\tau = 1/\mu$ and $S^2 = 2/\mu^2$, we get that $\theta^* = \sqrt{w\mu/c}$. It is interesting to note, as observed in Kulkarni [2], that θ^* is independent of the arrival rate. We also note that when the retrial rate of θ^* is used the two terms that depend on θ in the cost function are equal. If we consider the rest of the costs as structural costs that cannot be changed by the decisionmaker, then we can conclude that, excluding that part of the costs, the waiting costs and the retrial costs coincide. This resembles the *economic order quantity* inventory control model, where the holding costs and the setup costs coincide under the optimal ordering policy.

2. THE EQUILIBRIUM RETRIAL RATE

The next theorem shows that there exists a unique equilibrium rate, θ_e , and provides an explicit formula for computing it. Note, as expected, that the social optimal and the equilibrium rates depend on the ratio w/c and not on the individual cost parameters.

THEOREM 2.1:

$$\theta_e = \frac{w\rho + \sqrt{w^2\rho^2 + 16w\tau c(1-\rho)(2-\rho)/S^2}}{4c(1-\rho)} \quad (2)$$

Moreover, this is the unique Nash equilibrium retrial rate. For exponential service,

$$\theta_e = \frac{w\rho + \sqrt{w^2\rho^2 + 8\mu wc(1-\rho)(2-\rho)}}{4c(1-\rho)}.$$

PROOF: Let $g(\gamma, \theta)$ be the expected waiting time for an individual who uses a retrial rate of γ while everybody else uses the retrial rate of θ . Equation (4.2) of Kulkarni [3] states that

$$g(\gamma, \theta) = \frac{\rho}{(1-\rho)\gamma} + \frac{\lambda S^2}{2(1-\rho)} + \frac{\lambda S^2 \rho}{2(1-\rho)} \frac{\theta - \gamma}{(1-\rho)\theta + \gamma}. \quad (3)$$

Also, let $f(\gamma, \theta)$ be the expected cost of the aforementioned individual. Then,

$$f(\gamma, \theta) = wg(\gamma, \theta) + \gamma cg(\gamma, \theta)$$

and

$$\begin{aligned} \left. \frac{d}{d\gamma} f(\gamma, \theta) \right|_{\gamma=\theta} &= c \left(\frac{\rho}{(1-\rho)\theta} + \frac{\lambda S^2}{2(1-\rho)} \right) \\ &+ (w + c\theta) \left(-\frac{\rho}{(1-\rho)\theta^2} + \frac{\lambda S^2}{2} \frac{\rho}{1-\rho} \frac{1}{\theta(\rho-2)} \right). \end{aligned}$$

A necessary condition for θ to define a Nash equilibrium is that

$$\left. \frac{d}{d\gamma} f(\gamma, \theta) \right|_{\gamma=\theta} = 0.$$

This gives

$$2\lambda c S^2 (1-\rho)\theta^2 - \lambda w S^2 \rho \theta - 2w\rho(2-\rho) = 0.$$

This quadratic equation has a unique positive root, θ_e .

Moreover, the zero derivative here corresponds to a minimum. This is the case due to the facts that $\lim_{\gamma \rightarrow 0} f(\gamma, \theta_e) = \infty$, that $\lim_{\gamma \rightarrow \infty} f(\gamma, \theta_e) = \infty$, and that for any other γ , $0 < \gamma < \infty$, $f(\gamma, \theta_e)$ is finite. Therefore, θ_e is a unique retrial rate that corresponds to a Nash equilibrium. ■

COROLLARY 2.2: For a fixed value of τ ,

- θ_e is a monotone increasing function of λ ;
- when $\rho \uparrow 1$, then $\theta_e \uparrow \infty$;
- when $\rho \downarrow 0$, then $\theta_e \downarrow \theta^*$.

Remark 2.3: From Eq. 2 it follows that the Nash equilibrium retrial rate is a monotone decreasing function of the variance of the service requirement (for

any given but fixed expected service length). Likewise, from Eq. 1, we deduce that the same is true with regard to the social optimal rate. This phenomenon can be explained as follows: The waiting time increases with S^2 . It then takes more time to clear the system after it has been found to be busy by a retrying customer when the variance is bigger; therefore, it pays to wait longer before retrying.

COROLLARY 2.4: *The equilibrium retrial rate is higher than the optimal retrial rate.*

PROOF: Because $2 - \rho > 2(1 - \rho)$,

$$\theta_e > \frac{\sqrt{32w\tau c(1-\rho)^2/S^2}}{4c(1-\rho)} = (1/S)\sqrt{2w\tau/c} = \theta^*. \quad \blacksquare$$

When a customer selects its retrial rate it ignores the way that this act affects the other customers. Thus, the social optimal retrial rate and equilibrium retrial rate differ. Such a phenomenon was noticed for the first time with respect to a queueing system by Naor [4]. In our model, when a customer in orbit repeats its call, it may deter another customer, new or in orbit, from succeeding in its own call. This means that a call, and in particular a repeated call, is associated with negative externalities. We have seen that the equilibrium retrial rate is larger than the social optimal rate. This reflects the fact that the length of the queue, t time units after the server was observed to be busy, stochastically decreases with t . Therefore, the expected negative externalities associated with a retrial decrease with the time since the previous trial. Hence, a customer who ignores these externalities, while optimizing its own welfare, retries too soon.

We now suggest two alternative ways to resolve the difference between the optimal and equilibrium retrial rates. One is by partial compensation (a *rebate*) to customers for their actual waiting time. The other is by taxing unsuccessful retrials. We note that the second option is the easier to administrate.

THEOREM 2.5: *There is a positive rebate value $r < w$ such that if paid to each customer per unit of waiting time then the resulting equilibrium retrial rate (which will be computed by replacing w with $w - r$ in Eq. 2) will coincide with θ^* . Moreover, r is uniquely determined by the equation*

$$\frac{(w-r)\rho + \sqrt{(w-r)^2\rho^2 + 16(w-r)\tau c(1-\rho)(2-\rho)/S^2}}{4c(1-\rho)} = (1/S)\sqrt{2w\tau/c}.$$

The same effect can be achieved by imposing a positive toll t per retrial such that the cost per retrial will be $c + t$ (instead of c). Evidently, $t = rc/(w - r)$.

PROOF: We already found in Corollary 2.4 that θ_e is greater than θ^* . Also, note that θ_e is continuously monotone increasing in w . Therefore, it is sufficient to show that for sufficiently small new values of w the resulting equilibrium retrial rate is smaller than the original θ^* . When $w - r$ goes down to 0, the value of θ_e

also goes down to 0 and therefore, for some intermediate value of r , equality holds. In particular, the equation stated in the theorem has a unique solution r , $0 < r < w$. Finally, the value of t is found by noticing that the social optimal and the equilibrium rates are functions of the retrial cost c and of the per unit of time waiting cost w , only through their ratio. ■

Acknowledgment

We thank Eitan Altman for fruitful discussions and, in particular, for telling us about Elcan's paper.

References

1. Elcan, A. (1994). Optimal customer return rate for an M/M/1 queueing system with retrials. *Probability in the Engineering and Informational Sciences* 8: 521-539.
2. Kulkarni, V.G. (1983). A game theoretic model for two types of customers competing for service. *Operations Research Letters* 2: 119-122.
3. Kulkarni, V.G. (1983). On queueing systems with retrials. *Journal of Applied Probability* 20: 380-389.
4. Naor, P. (1969). The regulation of queue size by levying tolls. *Econometrica* 31: 15-24.