Minimum Waiting Time in Queues with Simultaneous Arrival of Customers

in the cases of Independence and Dependence

Ramin Behzad ; M.Reza Salehirad

Allameh Tabataba'i University

The 12th edition of Young European Queueing Theorists

December 2018

The Need to more flexible queueing models

- In classical queueing systems, a customer is allowed to wait only in one queue to receive the service.
- In practice, Customers may tend to simultaneously take turn in more than one queue in order to reduce the waiting time.
- Most queueing models assume independence between the related random variables.
- In practice, they will be dependent.

Considerations

- We consider two queue rendering the same service.
- Customers may wait for the service simultaneously in the two queues.
- Customer withdraws from one of the requests made if the service is received through the other queue.
- The above condition brings some abandonment in each queue.
- How long should a given customer wait for the service under the above-mentioned circumstance ? $\rightarrow$ Minimum waiting time random variable.

Objectives

- To obtain the distribution function of the minimum waiting time (MWT).
- To compute the expected value of the MWT.
- To study the situation under the cases of Independence and Dependence for the waiting time random variables.

Model Description

We denote T_1 and T_2 as the waiting times in the two queues

$$T_M = \min(T_1, T_2) = T_1.I_{\{T_1 < T_2\}} + T_2.I_{\{T_1 > T_2\}}$$

$$E(T_M) = E\left(T_1.I_{\{T_1 \leq T_2\}}\right) + E\left(T_2.I_{\{T_1 > T_2\}}\right)$$

$$P(T_M > t) = P(\min(T_1, T_2) > t) = P(T_1 > t).\ P(T_2 > t)$$

Variables of the Model

- T_0: Virtual waiting time
- S: Service time random variable
- R: Customer's patience time, an exponential random variable
- $T_a = \min\{T_0, R\}$: actual waiting times are considered as right-censored by patience times.

Wating Time Model

For an $M/M/N$ queueing model, with λ as the average rate of customers arriving at the queue and μ as the average rate of serving customers, the appropriate staffing level, for moderate to large values of $S = \frac{\lambda}{\mu}$, is as follows

$$N = S + \beta\sqrt{S} \tag{1}$$

where β is a positive constant that depends on the desired level of service.

Wating Time Model

Adding abandonment to this queueing system and under exponentially distributed patience time random variable with parameter θ, i.e. $M/M/N+M$, Garnett et al.(2002) proposed the following survival function for the actual waiting time (T_a),

$$P(T_a > t) = \omega(-\beta, \sqrt{\frac{\mu}{\theta}}) \cdot \frac{h(\beta \cdot \sqrt{\frac{\mu}{\theta}})}{\Psi(\beta \cdot \sqrt{\frac{\mu}{\theta}}, \sqrt{N\mu\theta} \cdot t)} \cdot e^{-\theta t}, \quad t \geq 0 \quad (2)$$

where $h(x) = \dfrac{\phi(x)}{\Phi(-x)}$ (ϕ and Φ are the standard Normal density and distribution functions),

$$\omega(x, y) = \left[1 + \frac{h(-xy)}{yh(x)}\right]^{-1} \text{ and } \Psi(x, y) = \frac{\phi(x)}{1 - \Phi(x+y)}.$$

Wating Time Model

Proposition 1. The *pdf* of the actual waiting time (T_a) is given by

$$f_{T_a}(t) = \frac{\theta\left[1-\Phi\left(\beta\sqrt{\frac{\mu}{\theta}}+\sqrt{N\mu\theta}t\right)\right]+\sqrt{N\mu\theta}\cdot\phi\left(\beta\sqrt{\frac{\mu}{\theta}}+\sqrt{N\mu\theta}t\right)}{\phi\left(\beta\cdot\sqrt{\frac{\mu}{\theta}}\right)}$$

$$\cdot\,\omega(-\beta,\sqrt{\frac{\mu}{\theta}})\cdot h(\beta\cdot\sqrt{\frac{\mu}{\theta}})\cdot e^{-\theta t}, \qquad t>0$$

$$\text{and} \quad P(T_a=0)=1-\omega(-\beta,\sqrt{\frac{\mu}{\theta}}) \tag{3}$$

Wating Time Model

The following approximation for $E(T_a)$,

$$E(T_a) \approx \left[1 - \frac{h\left(\beta\sqrt{\frac{\mu}{\theta}}\right)}{h\left(\beta\sqrt{\frac{\mu}{\theta}} + \sqrt{\frac{\theta}{N\mu}}\right)}\right] \omega\left(-\beta, \sqrt{\frac{\mu}{\theta}}\right) \frac{1}{\theta}. \quad (4)$$

Minimum Wating Time

Theorem 1. For the random variables T_1 and T_2 as independent actual waiting time of the two queues, the minimum waiting time (T_M), for a customer waiting simultaneously in the two queues, has the following *cdf*

$$F_{T_M}(t) = 1 - \frac{\omega(-\beta, \sqrt{\frac{\mu}{\theta}})\omega(-\beta', \sqrt{\frac{\mu'}{\theta'}})h(\beta\sqrt{\frac{\mu}{\theta}})h(\beta'\sqrt{\frac{\mu'}{\theta'}})}{\Psi(\beta\sqrt{\frac{\mu}{\theta}}, \sqrt{\mu\theta}\cdot t)\Psi(\beta'\sqrt{\frac{\mu'}{\theta'}}, \sqrt{\mu'\theta'}\cdot t)} \cdot e^{-(\theta+\theta')t} \tag{5}$$

Minimum Wating Time

Theorem 2. Considering (3) as the *pdf* for independent actual waiting time random variables T_1 and T_2 with the respective parameters, expectation of the minimum waiting time, $E(T_M)$, for a customer waiting in the two queues is given by the following:

$$
\begin{aligned}
E(T_M) = {} & A \cdot E\left[\int_0^{T_2} t_1 \left[1 - \Phi\left(\beta\sqrt{\frac{\mu}{\theta}} + \sqrt{\mu\theta} \cdot t_1\right)\right] \cdot e^{-\theta t_1} dt_1\right] \\
& + B \cdot E\left[\int_0^{T_2} t_1 \cdot \phi\left(\beta\sqrt{\frac{\mu}{\theta}} + \sqrt{\mu\theta} \cdot t_1\right) \cdot e^{-\theta t_1} dt_1\right] \\
& + A' \cdot E\left[\int_0^{T_1} t_2 \left[1 - \Phi\left(\beta'\sqrt{\frac{\mu'}{\theta'}} + \sqrt{\mu'\theta'} \cdot t_2\right)\right] \cdot e^{-\theta' t_2} dt_2\right] \\
& + B' \cdot E\left[\int_0^{T_1} t_2 \cdot \phi\left(\beta'\sqrt{\frac{\mu'}{\theta'}} + \sqrt{\mu'\theta'} \cdot t_2\right) \cdot e^{-\theta' t_2} dt_2\right]
\end{aligned}
\tag{6}
$$

Wating Time Model

where

$$A = \frac{\theta \cdot \omega(-\beta, \sqrt{\frac{\mu}{\theta}}) \cdot h(\beta\sqrt{\frac{\mu}{\theta}})}{\phi(\beta\sqrt{\frac{\mu}{\theta}})}$$

$$B = \frac{\sqrt{\mu\theta} \cdot \omega(-\beta, \sqrt{\frac{\mu}{\theta}}) \cdot h(\beta\sqrt{\frac{\mu}{\theta}})}{\phi(\beta\sqrt{\frac{\mu}{\theta}})}$$

A' and B' are the same as A and B respectively, with the parameters (β', μ', θ').

- The four expectations do not lead to closed form expressions.
- We utilize a Monte Carlo Method, namely importance sampling, to compute the expectations.

Minimum Wating Time

To compute the first term in the right-hand side of (6), which is denoted by I_1, let $g_1(T_2)$ be the integral inside this term, then

$$I_1 = A \cdot E[g_1(T_2)] = A \cdot \int_0^\infty g_1(t_2) \cdot f(t_2) dt_2$$

$$= \frac{A \cdot \omega(-\beta', \sqrt{\frac{\mu'}{\theta'}}) \cdot h(\beta' \sqrt{\frac{\mu'}{\theta'}})}{\theta' \cdot \phi^2(\beta' \sqrt{\frac{\mu'}{\theta'}})} \cdot$$

$$E_{exp}\Bigg[g_1(T_2)\Bigg(\theta' \phi(\beta' \sqrt{\frac{\mu'}{\theta'}})[1 - \Phi(\beta' \sqrt{\frac{\mu'}{\theta'}} + \sqrt{\mu'\theta'}\, T_2)]$$

$$+ \phi(\beta' \sqrt{\frac{\mu'}{\theta'}}) \sqrt{\mu'\theta'} \phi(\beta' \sqrt{\frac{\mu'}{\theta'}} + \sqrt{\mu'\theta'}\, T_2)\Bigg)\Bigg] \qquad (7)$$

Minimum Wating Time

where $E_{exp}(\cdot)$ implies an expectation to be computed under exponential distribution with parameter θ'. Let the expression inside this expectation which has been multiplied by $g_1(T_2)$ be denoted by $R_1(T_2)$.
In addition, by considering a uniform distribution on $(0, t_2)$ for U, i.e. $U \sim U(0, t_2)$, $g_1(t_2)$ can be written as follows

$$\begin{aligned} g_1(t_2) &= \int_0^{t_2} u \cdot t_2 \left[1 - \Phi(\beta\sqrt{\frac{\mu}{\theta}} + \sqrt{\mu\theta} \cdot u)\right] \cdot e^{-\theta u} \cdot \frac{1}{t_2} du \\ &= E_U[K_1(U)] \end{aligned} \tag{8}$$

where $K_1(u) = u \cdot t_2 \left[1 - \Phi(\beta\sqrt{\frac{\mu}{\theta}} + \sqrt{\mu\theta} \cdot u)\right] \cdot e^{-\theta u}$ and $E_U(\cdot)$ denotes an expectation under uniform distribution on $(0, t_2)$.

Minimum Wating Time

By applying a two-stage importance sampling method, we use the following algorithm to approximate I_1:

i) Generate n random values from $exp(\theta')$ to be considered as values of t_2.

ii) For each value of t_2 obtained in (i), compute $R_1(t_2)$.

iii) For each value of t_2 obtained in (i), compute $g_1(t_2)$. To this end, the following steps are taken:

a) Generate n random numbers from $U(0, t_2)$ distribution to be considered as values of u.

b) For each value of u, compute $K_1(u)$.

c) Calculate $\frac{1}{n}\sum_{i=1}^{n} K_1(u_i)$ to be utilized as an approximation for $g_1(t_2)$.

Minimum Wating Time

iv) Take $\widehat{E}_1 = \frac{1}{n}\sum_{i=1}^{n} g_1(t_{2i}) \cdot R_1(t_{2i})$ as an approximation for $E_{exp}[g_1(T_2) \cdot R_1(T_2)]$.

v) Finally, take $\dfrac{A \cdot w(-\beta', \sqrt{\frac{\mu'}{\theta'}})h(\beta'\sqrt{\frac{\mu'}{\theta'}})}{\theta'\phi^2(\beta'\sqrt{\frac{\mu'}{\theta'}})} \cdot \widehat{E}_1$ as an approximation for I_1.

Minimum Wating Time

Theorem 3. Let λ_1 and λ_2 are the average rates of customers ariving at the first and second queues and μ_1 and μ_2 are the average rates of serving customers respectively. Denoting R_1 and R_2 as the exit time random variables for the first and second queues and considering an exponentially distributed service time, we have

- $R_1 \sim exp(\mu_2 - \lambda_2)$
- $R_2 \sim exp(\mu_1 - \lambda_1)$

Numerical Results

Two online free file converters:

Parameters	(λ_i)	(μ_i)	θ_i	β_i	$E(T_{ai})$
Queue I	6.8	7.3	0.3	0.0709	0.3241
Queue II	5.8	6.1	0.5	0.0504	0.2683

Numerical Results

This table shows the first (I_1), the second (J_1), the third (I'_1) and the forth (J'_1) terms of the right-hand side of (6) and the estimation for the expectation of the minimum waiting time, $\widehat{E}(T_M)$, with 10000 iterations

Quantities	Values
I_1	0.00864266
J_1	0.06323516
I'_1	0.01455535
J'_1	0.06971538
$\widehat{E}(T_M)$	0.15614855

Copula Approach for the Dependence Case

Let H be a joint distribution function with margins F and G, then by Sklar's theorem there exists a copula C such that for all x, y in $\overline{\mathbb{R}}$,

$$H(x, y) = C(F(x), G(y))$$

If F and G are continuous, then C is unique; otherwise, C is uniquely determined on Ran$F \times$Ran$G$. Conversely, if C is a copula and F and G are distribution functions, then the function H defined above is a joint distribution function with margins F and G.

Copula Approach

Proposition 2. Let T_1 and T_2 are continuous *rvs* with F and G as the *cdfs* respectively, and with copula C. Then,

$$P\left[\min\left(T_1, T_2\right) \leq t\right] = F(t) + G(t) - C(F(t), G(t))$$

- The strongest possible dependency between T_1 and T_2 is given by Comonotone copula $C_M(u, v) = \min(u, v)$. Thus, for two identically distributed *rvs* T_1 and T_2, we have

$$F_{T_M}(t) = 2F(t) - \min\left(F(t), F(t)\right) = 2F(t) - F(t) = F(t)$$

Copula Approach

- In case of the strongest possible dependency, the expectation of the *MWT* is equal to the ones for each *rv*.
- The strongest negative dependency between two *rvs* is given by Countermonotone copula $C_W(u, v) = \max(u + v - 1, 0)$.

Copula Approach

Example. For the two waiting time *rvs* T_1 and T_2, identically distributed with F as the *cdf*, with the parameters $\mu = 7.3$, $\theta = 0.3$ and $\beta = 0.0709$, the expectation of the *MWT* is as follows.

- Under the Comonotone copula: $E(T_M) = 0.3241$
- Under the Countermonotone copula: $E(T_M) = 0.0603$

FGM Copula

Farlie-Gumbel-Morgenstern (FGM) Copula is defined for $\alpha \in [-1, 1]$ as follows:

$$H(x, y) = F(x)\, G(y) \left[1 + \alpha \overline{F}(x)\, \overline{G}(y)\right]$$

which leads to the following representation.

$$C(u_1, u_2) = u_1 u_2 \left[1 + \alpha (1 - u_1)(1 - u_2)\right], \qquad u_1, u_2 \in [0, 1]$$

FGM Copula

Proposition 3. For the *rvs* T_1 and T_2, identically distributed with F as the *cdf*, the expectation of the *MWT* under *FGM* copula with α as the copula parameter is given by

$$\begin{aligned} E(T_M) = \int_0^\infty t f_{T_M}(t)dt = 2\int_0^\infty t f(t)dt - 2\int_0^\infty t f(t)F(t)dt \\ - 2\alpha \int_0^\infty t f(t)F(t)\left(1-F(t)\right)^2 dt \\ + 2\alpha \int_0^\infty t f(t)\left(1-F(t)\right)\left[F(t)\right]^2 dt \end{aligned} \tag{9}$$

FGM Copula

Expectation of the *MWT* via *FGM* copula:

α	$E(T_M)$
-0.9	0.1113962
-0.6	0.121431
0.6	0.1615698
0.9	0.1716046

Copula Approach

For (X_1, X_2) with distribution function F and f as its pdf, the df of $X_2|X_1 = x_1$ is denoted by $F_{2|1}(.|x_1)$. Then, the random variables U_1 and U_2^* defined by $U_1 = F_1(X_1)$, $U_2^* = F_{2|1}(X_2|X_1)$ are independent. Moreover, U_2^* is uniformly distributed on $(0, 1)$. Thus, the dependence is removed by this transform. The conditional distribution for the *FGM* copula C, denoted by $C_{2|1}(u_2|u_1)$ is

$$C_{2|1}(u_2|u_1) = \frac{\partial C(u_1, u_2)}{\partial u_1} = u_2 A + (1 - A) u_2^2$$

where $A = 1 + \alpha(1 - 2u_1)$.

Algorithm for Simulating from Copula

i) Simulate u_1 and v_2 independently from Uniform distribution on $(0,1)$.

ii) Return (u_1, u_2) where $u_2 = C_{2|1}^{-1}(v_2|u_1)$. The inverse $C_{2|1}^{-1}(v_2|u_1)$ is obtained by solving the following equation $(in \quad u_2)$:
$C_{2|1}(u_2|u_1) = v_2$ which for FGM copula leads to

$$(1-A)\,u_2^2 + Au_2 - v_2 = 0$$

and then we have $u_2 = \frac{2v_2}{A+\sqrt{A^2-4(A-1)v_2}}$.

iii) Finally, simulate $(x_1, x_2) = \left(F_1^{-1}(u_1), F_2^{-1}(u_2)\right)$.

FGM Copula

FGM copula parameters $\alpha = -0.9, -0.6, 0.6$ and 0.9, which are equivalent to $\tau = -0.2, -0.134, 0.134$ and 0.2 respectively, are considered.
We then simulate 1000 values from T_1 and T_2 using the *FGM* copula.

a) Generate u_1 from uniform distribution on $(0, 1)$.

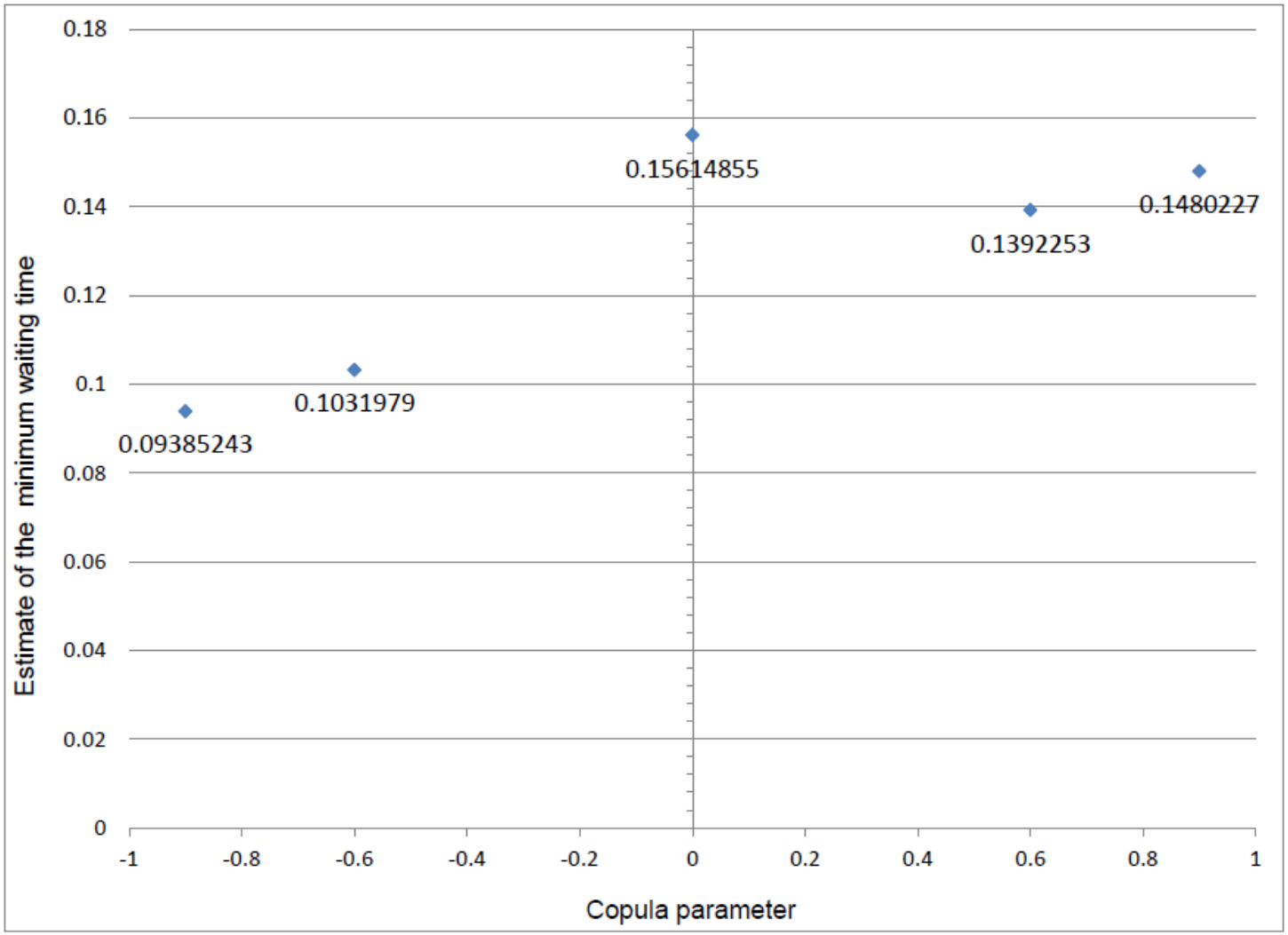
b) If $u_1 > P(T_1 = 0) = 1 - \omega\left(-\beta, \sqrt{\frac{\mu}{\theta}}\right)$ then return $T_1 = F_1^{-1}(u_1)$ through considering (2), otherwise return $T_1 = 0$.

- The steps to simulate T_2 are analogous to the above and $u_2 = C_{2|1}^{-1}(v_2|u_1)$ is considered.

Results for the Dependence Case

	Variables	Mean	Standard deviation
$\alpha = -0.9$	T_1	0.3197498	0.3415226
	T_2	0.268793	0.2993884
	$\min(T_1, T_2)$	0.09385243	0.1428322
$\alpha = -0.6$	T_1	0.3197498	0.3415226
	T_2	0.2680972	0.2988734
	$\min(T_1, T_2)$	0.1031979	0.1571282
$\alpha = 0.6$	T_1	0.3197498	0.3415226
	T_2	0.2631755	0.2965341
	$\min(T_1, T_2)$	0.1392253	0.1973006
$\alpha = 0.9$	T_1	0.3197498	0.3415226
	T_2	0.2619269	0.2955623
	$\min(T_1, T_2)$	0.1480227	0.2048147

Results for the Dependence Case



Concluding Remarks

- Sometimes customers simultaneously stand in more than one queue in order to reduce their waiting time.
- Distribution function and expectation of the minimum waiting time random variable obtained.
- Generalization to more than two queues may be considered.
- Copula approach is applied to take into account the dependency of the waiting time random variables.

Thanks for your attention.