1 GI/M/1 queue

The GI/GI/1 queue is the single-server model when the arrival process is a renewal process with iid interarrival times \( \{T_n\} \) with common df \( A(x) = P(T \leq x) \) with \( E(T) = 1/\lambda \), the service times \( \{S_n\} \) are iid with common df \( G(x) = P(S \leq x) \) with \( E(S) = 1/\mu \), and the two sequences (interarrival, service time) are independent. Two special cases for which we can successfully analyze the model in steady-state is when either one or the other of \( A, G \) is exponential; the former denoted by GI/M/1 and the latter by M/G/1.

Here we study the FIFO GI/M/1 and shall see that the steady-state distribution of the number in system, \( L \), as found by an arrival is geometric; \( P(L^a = n) = (1-\alpha)\alpha^n, \ n \geq 0 \), with the parameter \( \alpha = P(D > 0) \), the long-run proportion of arrivals who find the system busy. \( D \) here denotes steady-state delay. We will also see that while in general it is not possible to solve for \( \alpha \) in closed form, one can easily compute it numerically as the solution to a fixed point problem.

Before we analyze the GI/M/1 in full generality, let us first review the very special case of the M/M/1, that is, when arrivals are Poisson.

1.1 FIFO M/M/1

For the M/M/1 queue (arrival rate \( \lambda \), service rate \( \mu \)) letting \( L(t) \) denote the number of customers in the system at time \( t \), recall that \( \{L(t) : t \geq 0\} \) forms a birth and death process. The balance equations for the stationary probabilities \( \{P_n : n \geq 0\} \), \( \lambda P_n = \mu P_{n+1}, \ n \geq 0 \), have a probability solution if and only if \( \rho \overset{\text{def}}{=} \lambda/\mu < 1 \) in which case \( P_n = (1-\rho)\rho^n, \ n \geq 0 \), a geometric distribution (with mass at 0). Letting \( L^a \) denote a rv with the steady-state distribution of the number in system as found by an arrival, PASTA implies that \( P(L^a = n) = P_n, \ n \geq 0 \); hence \( P(L^a = n) = (1-\rho)\rho^n, \ n \geq 0 \). From here we can easily obtain the stationary distribution for sojourn time \( W \) (the total amount of time a customer spends in the system from arrival to departure), and delay \( D \) (time spent waiting in the line (queue)). Recall that \( W = D + S \), where \( S \) (a service time) is independent of \( D \).

**Proposition 1.1** For the FIFO M/M/1 queue with \( \rho < 1 \), stationary sojourn time \( W \) has an exponential distribution with rate \( \mu(1-\rho) \); \( P(W > x) = e^{-\mu(1-\rho)x}, \ x \geq 0 \).

For stationary delay \( D \), \( P(D = 0) = 1 - \rho \), \( P(D > x) = \rho e^{-\mu(1-\rho)x}, \ x \geq 0 \); \( D \sim (1-\rho)\delta_0 + \rho \exp(\mu(1-\rho)). \) (This means that \( (D \mid D > 0) \sim \exp(\mu(1-\rho)) \).

First we need

**Lemma 1.1** Suppose \( \{S_j\} \) are iid with an exponential distribution at rate \( \mu \), and independently \( M \) has a geometric distribution with success probability \( p \); \( P(M = k) = (1-\rho)^{k-1}p, \ k \geq 1 \). Then the random sum \( Y = \sum_{j=1}^M S_j \) has an exponential distribution at rate \( \mu p \).

**Proof** : We offer two proofs. The first is an immediate consequence of our partitioning theorem for Poisson processes: Consider a Poisson process with interarrival times \( \{S_j\} \); \( t_n = \sum_{j=1}^n S_j, \ n \geq 1 \). Let us partition this point process by using iid Bernoulli \( (p) \) rvs,
{B_n: n ≥ 1}: If B_n = 1, then this is called a “success” and t_n is a type 1 arrival, where as if B_n = 0 it is a type 2 arrival. We know that the sequence of type 1 arrivals, denoted by (say) \{t_n(1): n ≥ 1\} form a Poisson process at thinned rate \mu p. Thus, in particular, the first arrival time \(T_1(1)\) has an exponential distribution at rate \mu p. Now note that \(Y = t_1(1) = \sum_{j=1}^{M} S_j\), where \(M\) is interpreted as the number of Bernoulli trials until the first success. Thus indeed \(Y\) has the desired exponential distribution.

Our second method is analytical: We prove that the Laplace transform of \(Y\) is given by

\[ E(e^{-sY}) = \frac{\beta}{\beta + s}, \quad s ≥ 0, \]

where \(\beta = \mu p\). To this end, by conditioning on \(M = k\) we obtain

\[ E(e^{-sY} \mid M = k) = \left(\frac{\mu}{\mu + s}\right)^k. \]

Thus, \(E(e^{-sY}) = E(z^M)\), where \(z = \frac{\mu}{\mu + s}\), the moment generating function of \(M\) evaluated at this value of \(z\). As the reader can check (via summing a geometric series) \(E(z^M) = \sum_{k=1}^{\infty} z^k(1-p)^k p = zp/(1 - z(1 - \rho))\). Plugging in the value \(z = \frac{\mu}{\mu + s}\) then yields our result.

**Proof :** [Proposition 1.1] Note that we can represent \(W\) as the sum of all \(L_a\) service times that an arrival finds (in steady-state, e.g., an arrival who arrives way out in the infinite future), in front of him plus his own added on service time \(W = \sum_{j=1}^{M} S_j\), where \(M = L_a + 1\).

The memoryless property of the exponential distribution tells us that even the remaining service time of the customer found in service has (as if it is brand new) the same exponential distribution and is independent of the others; hence the above sum has iid \(S_j\) and they are independent of \(L_a\). Note that \(P(M = k) = P(L_a = k - 1) = (1 - \rho)^{k-1}, k ≥ 1\) a geometric distribution with success probability \(p = 1 - \rho\); thus the result follows from Lemma 1.1.

Similarly \(D = \sum_{j=1}^{L_a} S_j\), and \(P(D = 0) = P(L_a = 0) = 1 - \rho\), and \(P(D > 0) = \rho\). Note that \((D \mid D > 0) = (D \mid L_a ≥ 1)\), and in fact \((L_a \mid L_a ≥ 1)\) has the same distribution as \(L_a + 1\) (geometric with success probability \(1 - \rho\)). Thus \((D \mid D > 0)\) has the same distribution as \(\sum_{j=1}^{L_a + 1} S_j\) which is the same as \(W\), exponential at rate \(\mu(1 - \rho)\).

1.2 FIFO GI/M/1

As mentioned in the Introduction we will prove for the FIFO GI/M/1 that the steady-state distribution of the number in system, \(L_a\), as found by an arrival is geometric; \(P(L_a = n) = (1 - \alpha)\alpha^n, n ≥ 0\), with the parameter \(\alpha = P(D > 0)\), the long-run proportion of arrivals who find the system busy. As a consequence, just as for the M/M/1, steady-state sojourn time \(W\) has an exponential distribution with rate \(\mu(1 - \alpha)\), and \(D \sim (1 - \alpha)\delta_0 + \alpha \exp(\mu(1 - \alpha))\). Whereas \(\alpha = \rho\) for the M/M/1 (via PASTA), \(\alpha\) will now, more generally, be either larger or smaller than \(\rho\), depending on whether \(A\) is less or more “deterministic” than the exponential distribution. In this context we will
study batch Poisson arrival processes. We will also find the time-stationary distribution for number in system.

**Markov chain analysis**

Letting \( X_n = L(t_n -) \), the number of customers found in the system by the \( n \)th customer \( C_n \), it is immediate that \( \{X_n : n \geq 0\} \) form a discrete-time, discrete-space Markov chain on \( \{0, 1, 2, \ldots\} \): upon arrival of \( C_n \), the next interarrival time \( T_n \sim \lambda \) independent of the past, and if \( X_n > 0 \) the remaining service time of the customer in service found by \( C_n \) has an exponential distribution and is independent of the past just like for those (if any) waiting in queue. Thus the future is independent of the past given the present state \( X_n \). We now proceed to set up the transition matrix \( P = (P_{ij}) \) for the chain, observe that it is irreducible, and then solve \( \pi = \pi P \). Here \( \pi_j \) denotes the long-run proportion of arrivals who find \( j \) in the system and \( L^a \) denotes a rv with this stationary distribution: \( P(L^a = j) = \pi_j, \ j \geq 0 \).

To proceed with our objective, we need to introduce some notation. Let \( \{N(t) : t \geq 0\} \) denote a Poisson process at rate \( \mu \), and let \( Y = N(T) \), the number of Poisson events during an (independent) interarrival time. When \( C_n \) arrives finding \( X_n \), the next arrival \( C_{n+1} \) will find \( X_n + 1 \) minus the number of departures that occurred during the time interval \( (t_n, t_{n+1}) \). The crucial observation is that whenever the system is busy, the departure process is a Poisson process at rate \( \mu \), and thus \( X_n \) can be represented by the recursion

\[
X_{n+1} = (X_n + 1 - Y_n)_+, \ n \geq 0,
\]

where \( Y_n = N(t_{n+1}) - N(t_n) \sim \lambda \), and are iid since the \( T_n \) are iid and the Poisson process has stationary and independent increments. \( Y \) thus can be interpreted as the potential number of departures during an interarrival time. Given that \( T = t, Y = N(t) \) is Poisson with mean \( \mu t \). Thus we conclude that

\[
a_n \overset{\text{def}}{=} P(Y = n) = \int_0^{\infty} e^{-\lambda t} \frac{(\lambda t)^n}{n!} dA(t).
\]

We are assuming that \( 0 < E(T) = 1/\lambda < \infty \) which implies that \( a_n > 0, n \geq 0 \). We also note that \( E(Y) = \mu E(T) = \rho^{-1} \), and the moment generating function MGF of \( Y \) is

\[
M(z) = E(z^Y) = \sum_{n=0}^{\infty} z^n a_n, \ |z| \leq 1.
\]

Computing \( P_{i,j} = P(X_{n+1} = j | X_n = i) \) is easy. For example if \( X_n = 0 \), then \( X_{n+1} = 0 \) or 1 depending on whether or not \( C_n \) has completed service by the time \( C_{n+1} \) arrives \( T_n \) units of time later; e.g. whether or not \( Y_n \geq 1 \) or \( Y_n = 0 \). Thus \( P_{0,0} = P(Y \geq 1) \), and \( P_{0,1} = P(Y = 0) = a_0 \). Similarly, \( P_{1,0} = P(Y \geq 2) \), \( P_{1,1} = P(Y = 1) = a_1 \), \( P_{1,2} = P(Y = 0) = a_0 \). Continuing in this fashion yields

\[
P = \begin{pmatrix}
P(Y \geq 1) & a_0 & 0 & 0 & \cdots \\
P(Y \geq 2) & a_1 & a_0 & 0 & \cdots \\
P(Y \geq 3) & a_2 & a_1 & a_0 & \cdots \\
\vdots & \vdots & \vdots & \vdots & \ddots \\
P(Y \geq n + 1) & a_n & a_{n-1} & a_{n-2} & \cdots
\end{pmatrix}.
\]
Irreducibility is immediate since \( a_n > 0, \ n \geq 0 \). Thus the chain is positive recurrent if and only if there exists a probability solution \( \pi \) to \( \pi = \pi \mathbf{P} \).

We “guess” the solution to be geometric; \( \pi_j = (1 - \alpha) \alpha^j, \ j \geq 0 \) for a yet to be determined \( 0 < \alpha < 1 \). If our guess works, then we are done. \( \pi = \pi \mathbf{P} \) for \( \pi_1 \) (e.g., using column \( 1 = (a_0, a_1, \ldots)^T \)) is given by

\[
\pi_1 = \sum_{i=0}^{\infty} \pi_i P_{1,i} = \sum_{i=0}^{\infty} \pi_i a_i,
\]

which then yields

\[
(1 - \alpha) \alpha = \sum_{i=0}^{\infty} (1 - \alpha) \alpha^i a_i.
\]

or more simply

\[
\alpha = \sum_{j=0}^{\infty} \alpha^j a_j = M(\alpha),
\]

where we recognize from (1) that the above infinite series is the MGF of \( Y, M(z) \), evaluated at \( z = \alpha \). Therefore we are seeking a solution to the fixed point equation \( \alpha = M(\alpha) \) for \( 0 < \alpha < 1 \). (We already know that \( M(1) = 1 \) but we need a solution strictly less than 1.) In other words we must show that the function \( M(x), \ x \in [0, 1] \) intersects the line \( y = x \) for a value of \( x \in (0, 1) \). Noting that \( M(x) \) is a positive function that satisfies \( M(0) = a_0 > 0, M(1) = 1 \) and is both strictly increasing (\( M'(x) > 0 \)) and strictly convex (\( M''(x) > 0 \)), we deduce from elementary calculus that this intersection can occur if and only if it occurs only once on \( (0, 1) \) in which case it occurs because the function \( M(x) \) crosses \( y = x \) from above at our desired point \( \alpha < 1 \), then crosses again from below at \( x = 1 \). But this can happen if and only if the slope \( M'(1) > 1 \).

Since \( M'(1) = E(Y) = \rho^{-1} \) we conclude that it is both necessary and sufficient that \( \rho < 1 \). Thus the chain is positive recurrent if \( \rho < 1 \) and either null recurrent or transient otherwise.

Note that \( \alpha = P(L^a > 0) = P(D > 0) \). For the M/M/1 queue, \( \alpha = \rho \) (via PASTA), but now this is no longer so. Also note that given \( T = t, Y \sim \text{Poisson}(\mu t) \) and thus \( E(z^Y|T = t) = e^{-\mu t(1-z)}, \) or \( E(z^Y|T) = e^{-\mu T(1-z)} \). Since \( E\{E(z^Y|T)\} = E(z^Y) \), we obtain

\[
M(z) = E(e^{-\mu T(1-z)}) = L_A(\mu(1-z)),
\]

expressing the MGF of \( Y \) entirely in terms of the Laplace transform (LT) of \( A \). Summarizing, and following the same analysis as in Proposition 1.1 for the M/M/1 case:

**Proposition 1.2** For the GI/M/1 queue with interarrival time df \( A \) having Laplace transform \( L_A(s) = E(e^{-sT}), \ s \geq 0 \), stability holds iff \( \rho < 1 \) in which case

1. \( P(L^a = j) = (1 - \alpha) \alpha^j, \ j \geq 0, \ where \( \alpha = P(D > 0) \), the long-run proportion of arrivals who find the server busy, is the unique solution in \( (0, 1) \) of the fixed point equation \( \alpha = L_A(\mu(1-\alpha)) \). \( E(L^a) = \alpha/(1-\alpha) \).

2. Both \( W \) and \( (D|D > 0) \) have an exponential distribution at rate \( \mu(1-\alpha) \) and the distribution of \( D \) is the mixture \( D \sim (1 - \alpha) \delta_0 + \alpha \exp(\mu(1-\alpha)) \). In particular \( d = E(D) = \alpha/(\mu(1-\alpha)) \), and \( w = E(W) = 1/(\mu(1-\alpha)) \).
1.2.1 Computing \( \alpha \), comparisons with \( \rho \).

For the M/M/1 queue, \( \mathcal{L}_A(s) = \lambda/(\lambda + s) \) and the solution to \( \alpha = \mathcal{L}_A(\mu(1 - \alpha)) \) is immediately seen to be \( \alpha = \rho \), thus confirming what we already know.

For other \( A \), \( \alpha \) usually can not be solved for in closed form, but can easily be solved for numerically by a standard iterative scheme from numerical analysis for finding fixed points: Choose any initial value \( \alpha_0 \in (0, 1) \), then define \( \alpha_{n+1} = M(\alpha_n) \), \( n \geq 0 \). It can be shown that \( \alpha_n \to \alpha \) as \( n \to \infty \).

\textit{Application to D/M/1:} \( \alpha < \rho \)

For example, suppose interrival times are deterministic renewal, \( P(T = 1/\lambda) = 1 \). This is called the D/M/1 queue (D for deterministic). Then \( \mathcal{L}_A(s) = e^{-s/\lambda} \), and we need to solve the equation \( \alpha = e^{-\rho^{-1}(1-\alpha)} \). Consider \( \rho = 0.5 \) in which case we must solve \( \alpha = e^{-(1-\alpha)} \). The iterative scheme is easily carried out even on a simple calculator: Using \( \alpha_0 = 0.5 \), we compute \( \alpha_1 = e^{-(1-0.5)} = e^{-0.5} = 0.37 \), and \( \alpha_2 = e^{-(1-0.37)} = 0.28 \). Continuing, the sequence \( \alpha_n \) decreases down to \( \alpha = 0.20 \).

This is quite interesting for we see that in this example \( \alpha = P(D > 0) = 0.20 \) is considerably smaller than \( \rho = 0.50 \), meaning that congestion is greatly reduced when replacing a Poisson arrival process by a deterministic renewal process at the same rate. Intuitively this makes sense because when arrivals are deterministic renewal, there is less variability in the model, and hence less chance of a queue building up. After all, a queue builds up because either some service times are unusually large, or some interarrival times are unusually small; with deterministic renewal arrivals the latter is not possible.

\textit{H/M/1:} \( \alpha > \rho \)

By this same logic, one would expect that using an \( A \) that has \textit{more} variability than the exponential would lead to an \( \alpha > \rho \). This turns out to be true, and a simple example is provided by using an \( A \) of the mixture form \( A = (1-p)\delta_0 + p(exp(p\lambda)) \), containing some mass at 0. This is a special case of a hyperexponential distribution, denoted by \( H \) (which more generally is a mixture of two or more distinct exponential distributions), and the model then denoted by H/M/1. This particular \( H \) is equivalent to allowing customers to arrive in batches (busloads) which then cause a buildup in the queue. The batch sizes are geometrically distributed with mean \( 1/p \), and the batches arrive according to a Poisson process at rate \( p\lambda \). Whereas the mean of \( A \) is \( \lambda \), the second moment is \( 2/p\lambda^2 \), so that the variance is larger than for the \( exp(\lambda) \) distribution, \( 2/\lambda^2 \), and tends to \( \infty \) as \( p \to 0 \). One nice way to prove that \( \alpha > \rho \) for this model is to realize that since the buses arrive as a Poisson process, they see time averages (PASTA). So the probability (in steady-state) that a customer \( C \) finds the system busy is equal to the probability that \( C \)'s bus finds the system busy, \( \rho \), plus the probability that the bus finds the system empty \( (1-\rho) \) and there is at least one person in front of \( C \) on the bus. Thus \( \alpha \) is of the form

\[
\alpha = \rho + (1-\rho)\gamma, \tag{2}
\]

where \( \gamma \) is the probability that there is at least one person in front of \( C \) on the bus.

In fact we can compute this exactly: Let \( B \) denote a batch. Then \( P(B = n) = (1-p)^{n-1}p, \; n \geq 1, \; E(B) = 1/p \). When the batch \( B \) arrives, if \( B = n \), then we index the customers as \( C_1, \ldots, C_n \) and they join the queue in order (\( C_1 \) first, \( C_2 \) second and so on). \( C_1 \) will be delayed only if the server is busy when the batch arrived, but all other
customers within the batch (if any) are always delayed. Letting $C$ denote a randomly selected customer (from among all customers over all batches), we now compute the probability, denoted by $\gamma$, that there is at least one person in front of $C$ in $C'$s batch. It is easier to compute $1 - \gamma$, the probability that $C$ is first in the batch. Observing that exactly one customer within each batch is first, it follows (from the renewal reward theorem), that $1 - \gamma = 1/E(B) = p$. Thus $\gamma = 1 - p$, and we conclude that $\alpha = \rho + (1 - \rho)(1 - p)$. The interested reader can additionally confirm this by showing that this $\alpha$ satisfies $\alpha = \mathcal{L}_A(\mu(1 - \alpha))$.

1.2.2 Batch Poisson arrival process; the BM/M/1 queue

We shall here generalize our batch arrival process in the previous section by allowing the batch size distribution to be general, $P(B = n) = b_n$, $n \geq 1$, with finite mean $E(B)$. We assume that the batches arrive according to a Poisson process at rate $\lambda$. We call this a batch Poisson arrival process, denoted by BM; the corresponding queue is then denoted by BM/M/1. Unless the batch-size distribution is geometric, such an arrival process is not renewal (the interarrival times of customers is not iid) and thus this is not a GI/M/1 queue, and $L^a$ will not have a geometric distribution, but we can still compute $P(D > 0)$ using the same ideas used in (2).

The arrival rate of customers is $\lambda E(B)$ and so $\rho = \lambda E(B)/\mu$. Our analysis for computing $P(D > 0)$ using PASTA and the renewal reward theorem remains valid; $P(D > 0) = \rho + (1 - \rho)\gamma$, where now, more generally, $1 - \gamma = 1/E(B)$; thus $P(D > 0) = \rho + (1 - \rho)(1 - 1/E(B))$.

Position within a batch of a randomly chosen customer

Whereas we computed the probability that $C$ is first within a batch, $1/E(B)$, it is of interest to compute more generally the probability that $C$ is $j$th within a batch, $j \geq 1$. Let $J$ denote $C$'s position within the batch. Our interpretation of $P(J = j)$ is as the long-run proportion of customers who are $j$th within a batch. Observing that exactly one customer within each batch can be $j$th and only if the batch size is at least of size $j$, we can use the renewal reward theorem with reward $R = I\{B \geq j\}$ to obtain the pmf

$$P(J = j) = \frac{P(B \geq j)}{E(B)}, \quad j \geq 1.$$  

Note how this looks almost the same as the density function for the equilibrium distribution of $B$, a continuous distribution,

$$f_e(x) = \frac{P(B > x)}{E(B)}, \quad x \geq 0.$$  

This is no coincidence, for $J$ is the discrete-time analog to equilibrium batch size $B_e$, what we obtain if time $t$ is discrete as is $B$ itself, $t \in \{0, 1, 2, \ldots\}$. $J$ can be viewed as the stationary forward recurrence time of a discrete-time renewal process having iid cycle lengths distributed as $B$.

We can now compute $E(J)$ by computing $\sum jP(J = j)$, but it is also instructive to use the renewal reward theorem: The reward over a cycle is the sum of all positions over
the batch and is given by \( R = 1 + \cdots + B = B(B + 1)/2; \) thus

\[
E(J) = \frac{E(B(B + 1))}{2E(B)}.
\]

Again, notice the similarity with \( E(B_e) = E(B^2)/2E(B). \)

**Remark 1.1** The discrete-time inspection paradox: \( C \) is more likely to lie within a larger batch, since larger batches hold more customers. The pmf of the size of \( C' \)'s batch is given by (use rewards \( R = jI\{B = j\} \))

\[
\frac{jP(B = j)}{E(B)}, \ j \geq 1,
\]

and the mean is \( E(B^2)/E(B). \)

**Remark 1.2** \( J - 1 \geq 0 \) is the number of customers, in \( C' \)'s batch, who are in front of \( C \); \( E(J - 1) = E((B)(B - 1))/E(B). \)

**Remark 1.3** If \( B \) has a geometric distribution, \( P(B = j) = (1 - p)^{j-1}p, \ j \geq 1, \) then \( J \) has the same distribution as \( B \); the geometric distribution is the unique memoryless distribution in discrete time, just as the exponential distribution is the unique memoryless distribution in continuous time.

### 1.2.3 Time-stationary distribution for the FIFO GI/M/1

Let \( L \) denote time-stationary number in system; \( P(L = j) = P_j, \) the proportion of time there are \( j \) in system, \( j \geq 0. \) \( P_0 = 1 - \rho, \) but to compute the other \( P_j, \) we imagine that the queue is a series of rooms, the server is room 0, followed by the first room (first slot in the queue) then the second and so on. A customer arriving finding \( j \geq 1 \) in the system starts by waiting in room \( j, \) and then consecutively, after each service completion, passes through each of the rooms \( j - 1, \ldots, 0. \) The average sojourn time in each room is \( 1/\mu, \) and for \( j \geq 1 \) the average number of customers in the \( j^{th} \) room is \( P(L > j) = E(I\{L > j\}) \) while the rate at which customers enter the \( j^{th} \) room is \( \lambda P(L^a > j - 1). \) Applying \( l = \lambda w \) to the \( j^{th} \) room yields \( P(L > j) = \rho P(L^a > j - 1) \) and consequently

\[
P_j = P(L > j - 1) - P(L > j)
= \rho(P(L^a > j - 2) - P(L^a > j - 1))
= \rho \pi_{j-1}
= \rho(1 - \alpha)\alpha^{j-1}, \ j \geq 1.
\]

In particular \( l = E(l) = \rho/(1 - \alpha). \)