A BAYESIAN ANALYSIS OF A QUEUEING SYSTEM WITH UNLIMITED SERVICE

by

C. Armero
Universitat de València,
Spain

and

M.J. Bayarri
Universitat de València,
Spain

Technical Report #93-50

Department of Statistics
Purdue University

September 1993

*

A BAYESIAN ANALYSIS OF A QUEUEING SYSTEM WITH UNLIMITED SERVICE

by

C. Armero and M.J. Bayarri
Universitat de València, Spain
Universitat de València, Spain

Abstract

A queueing system occurs when "customers" arrive to some facility requiring a certain type of "service" provided by the "servers". Both the arrival pattern and the service requirements are usually taken to be random. If all the servers are busy when customers arrive, they usually wait in line to get served. Queues possess a number of mathematical challenges and have been mainly approached from a probability point of view, and statistical analysis are very scarce. In this paper we present a Baysesian analysis of a Markovian queue in which customers are immediately served upon arrival, and hence no waiting lines form. Emergency and self-service facilities provide many examples. Technically such services can be modelled as queues with an infinite number of servers. The mathematical simplicity of these queues allow for closed-form exploration of a number of issues that arise when statistically analyzing queues, whether or not the queue is in equilibrium.

Key words and phrases: Comparison of experiments; conjugate families; Kummer's function; Non-informative priors; Prediction; Steady-state; Transient behavior.

1. Introduction

Queues, or waiting lines, are so common in our daily lives that precise definitions seem superfluous. A queueing system occurs anytime "customers" demand "service" from some facility; usually both the arrival and service times are assumed to be random. If all the "servers" are busy when new customers arrive, these will usually wait in line for the next available server. The term "customers" is very broad and can refer to telephone calls arriving to a switchboard, machine failures, demands of CPU time, patients arriving at emergency services in a hospital, etc. (Obviously, the same applies also to the terms "service" and "servers".) The behavior of a queueing system is characterized by the arrival pattern, the number of servers and the model for service times, the queue discipline (usually first-in-first-out, or FIFO, but many others are possible), the size of the "waiting room," the total size of the population of customers, ... etc.

Queueing theory is a field of impressive activity and growth, both in theoretical research and applications. (Some key references can be found in Armero and Bayarri, 1993 b.) Nevertheless, most of this vast effort is devoted to the construction of models and the study of the mathematical properties of such models and comparatively little effort has been devoted to the statistical analysis of such systems (a good review is Bhat and Rao, 1987; see also Lehoczky, 1990). Bayesian analysis is basically absent; in fact, to the best of our knowledge, the only Bayesian references are Muddapur (1972), Reynolds (1973), Armero (1985, 1993), McGrath, Ross and Singpurwalla (1987), McGrath and Singpurwalla (1987), and Armero and Bayarri (1993a, 1993b). The statistical analysis of queueing systems possess a number of interesting and challenging questions and Bayesian methods are especially well suited to deal with them, not only because they can easily incorporate prior information (which can be substantial in queueing systems that have been running for some length of time), but also, and most importantly, because they can handle in a natural way frequently occurring issues such as restrictions in the parameter space (contrast the classical analysis in Schruben and Kulkarni, 1982, with its Bayesian counterpart in Armero and Bayarri, 1993b) and prediction problems (Armero and Bayarri, 1993a). Also, in complex queueing systems, densities are not available in closed form and analysis has to rely on simulations, which are performed for fixed values (estimates) of the unknown parameters; a natural way to incorporate the inherent uncertainty is to simulate the parameter values from their posterior distribution (an approximation to this procedure was reported by Butler and Huzurbazar, 1993). Last, there are systems, such as networks of related queues, for which the only feasible analysis is a Bayesian or an Empirical Bayesian one (Thiruvaiyaru and Basawa, 1992).

In this paper we address some of these issues in a very simple, yet frequently occurring, queueing system, namely one in which infinite servers are assumed and hence congestion does not occur. Of course, in real life there can not be infinite servers; technically, a system in which a customer receives immediate service, without delays, is usually modelled as a queueing system with infinite servers. Common examples are provided by emergency services (such as ambulances, police, firemen ... etc.) and by any self-service facility. Other examples are provided by cars traversing a bridge, say, in a non-congestion hour (the "service" time is the time needed to complete the traverse), or by "customers" turning on their TV sets in a certain time period, etc. Perhaps, the most frequent use of queues with infinite servers is to approximate the behavior of queues with a finite, but very large, number of servers in which congestion rarely occurs; this is in fact the case of the ambulance service example mentioned above, or the very important example in which the "customers" are the lines in use in a large communication network. Some queueing systems with a finite number of servers are studied in the references given above.

The paper has 7 sections, this introduction being Section 1. In Section 2 we introduce the queueing system we shall be dealing with the rest of the paper, namely the $M/M/\infty$ queue. Section 3 is devoted to discussion and comparison of several possible experiments that can be performed to obtain information about the parameters governing the queue. Section 4 introduces a new family of distributions (We call them *Kummer distributions*) that will be needed in the rest of the paper. In Section 5 we derive posterior distributions of several quantities of interest, under both non-informative and conjugate priors (conjugacy is possible through a wide range of different experiments, with non proportional likelihoods). Section 6 is devoted to the very important issue of prediction under steady-state. Transient behavior is briefly discussed in Section 7.

2. The $M/M/\infty$ Queue

It has become standard to describe simple queueing systems in terms of Kendall's notation (Kendall, 1953), which consists of a series of symbols separated by slashes, A/S/C/R/Q, where A characterizes the interarrival-time distribution, S the service distribution, C is the number of channels of service, R the restriction on the system capacity, and Q is the queue discipline. The last two are omitted when $R = \infty$ (no restrictions) and Q=FIFO (first-in-first-out). Also, exponential distributions are denoted by M. In this paper we deal with the $M/M/\infty$ queue, that is, both the inter-arrival time and the service time are assumed to have an exponential distribution, and the number of servers is ∞. (Bayesian analysis of other types of Markovian queues is presented in the references mentioned in Section 1.) In a $M/M/\infty$ queue, customers arrive to the service according to a

Poisson process with mean λ (so that inter-arrival times are exponential with mean $1/\lambda$), and service times are independent of the arrivals and follow an exponential distribution with mean $1/\mu$. The parameter λ is usually called the arrival rate, and μ the service rate.

Although the inferential aim can sometimes be to estimate the parameters characterizing the system (λ and μ in our case), it is most common in a queueing system that interest focuses on predicting observable quantities that describe the utilization, efficiency and congestion of the system. The so-called *measures of performance* include quantities such as the number of customers in the system, the time that a customer spends in the system, length of busy periods (or idle periods)... etc. (Often, the expected value of these quantities are themselves called measures of performance). Also, it is often the case that queuing systems have been working for a long period of time (under similar conditions) or they are supposed to, and interest then lies in the prediction of these quantities in a queue in equilibrium (or steady-state). Unlike $M/M/c$ queues, a $M/M/\infty$ will always reach its steady-state for every value of μ and λ. We mainly will be interested in the prediction, under steady-state, of the number of customers in the system, N, and of the time, W, that a customer spends in the system (the waiting time). Other measures of performance directly concerned with the servers, such as the length of busy or idle periods, obviously do not apply to the $M/M/\infty$ queue. The same is true of measures concerned with the queue itself, such as the time a customer has to queue before being served, or the length of the waiting line. The results that follow can be found in most standard texts in queueing theory, such as Gross and Harris (1985).

In a $M/M/\infty$ queue in equilibrium, the distribution of the (steady-state) number of customers in the system, N, is Poisson with parameter $\frac{\lambda}{\mu}$, so that

$$p(N = n|\mu, \lambda) = \frac{(\lambda/\mu)^n \; e^{-\lambda/\mu}}{n!}, \; n = 0, 1, 2, \ldots \tag{2.1}$$

In this queue, N is also the number of busy servers (another possible measure of performance). The quantity λ/μ, that is, the mean number of arrivals per unit time when the time unit is taken to be the mean service time, is called, in a general queueing system, the *offered load*; it is a dimensionless quantity whose unit is the *erlang*. We shall denote λ/μ by θ. Notice, from (2.1), that, in the $M/M/\infty$ queue, the offered load θ is also the expected value of N, so that it can be interpreted as the mean number of servers that the customer population "wants" to be able to hold simultaneously (Cooper, 1990).

Since in a $M/M/\infty$ queue a customer is served without ever queueing, the time that she/he spends in the system W obviously equals the time it takes to serve that customer, so that the waiting time distribution is given by the exponential service distribution, that is

$$p(w|\mu,\lambda) = p(w|\mu) = \mu e^{-\mu w}, \quad w > 0. \tag{2.2}$$

It is worth noting that (2.2) also holds for a $M/G/\infty$ queue, where G stands for a general distribution for the service time.

As for transient behavior, it turns out that, unlike what happens with most queueing systems, transient probabilities can be derived in a closed form for a $M/M/\infty$ queue. Indeed, the distribution of the number of customers in the system at time t, N(t), is Poisson with mean $\lambda(1-e^{-\mu t})/\mu$, so that

$$p(N(t)=n|\mu,\lambda) = \frac{1}{n!}\ [(1-e^{-\mu t})\frac{\lambda}{\mu}]^n\ exp\{-(1-e^{-\mu t})\frac{\lambda}{\mu}\}, \quad n = 0,1,2\ldots \tag{2.3}$$

It is noteworthy that (2.3) holds also for $M/G/\infty$ systems, even for a time-dependent arrival rate and time-dependent service time distribution (Newell, 1982). Also, it is immediate to check that (2.1) can be obtained from (2.3) by letting t go to ∞. (The probabilities (2.3) are derived under the usual assumption that the system is empty at time $t = 0$.)

3. Experiments and Likelihoods

A queue can be observed in a wide variety of ways and many different quantities can be recorded. The classical experiments all seem to observe the system during some interval of time (0, T] and they can get rather cumbersome to perform. Benes (1957) takes T to be fixed in advance, and classifies the customers as belonging to four different categories dependent on whether they are or are not in the system when observation begins, and whether or not they have left the system when observation ends; the observed quantities are the number of customers belonging to each category and the arrival and departure times during the period of observation. Considering a $M/M/\infty$ queue as a particular case of a birth-and-death process (the state of the process is the number of customers in the system, a "birth" occurring each time a customer arrives, and a "death" each time a customer leaves the system after being served), Wolf (1965), and Basawa and Prakasa Rao (1980), propose two other experiments, both of them also observing the system in a certain period (0,T] of time. For Wolf (1965), T is fixed and the quantities to be observed are the number of "births", the number of "deaths", and the total time the system spends in state $i, i = 0, 1, 2, \ldots$. The experiment proposed by Basawa and Prakasa Rao observes the same quantities, but now T is random and the system is observed until a fixed number of transitions (arrivals and departures) have occurred. (That is, the total number of "births" and "deaths" is fixed here.)

Even though the experiments above are very different from each other from a frequentist perspective, due to the lack of memory of the exponential distribution, all of them result in the same likelihood function for the given data. Hence, according to the likelihood principle (see, for instance, Berger and Wolpert, 1988) they do provide exactly the same information about λ and μ. We will use a much simpler experiment to perform that also results in the same likelihood function. Specifically, we assume that n_a inter-arrival times and n_s service completions are observed; the observation of the arrival and service processes do not need to be simultaneous. Also, n_a and n_s can be fixed or random (as long as their distributions do not depend on λ nor μ). Let X_i denote the service time of the $i-th$ customer, $i = 1, 2, \ldots, n_s$, and Y_j the time elapsed between the arrivals of customers j and $j-1, \;\; j = 1, \; 2, \ldots, n_a$ (as a notational device, assume that customer 0 is the first one entering the queue during the observation period). Then, according to the hypothesis of a $M/M/\infty$ queue, $\boldsymbol{Y} = (Y_1, \ldots, Y_{n_a})$ is a random sample from an exponential distribution with parameter λ, and $\mathbf{X} = (X_1, \ldots, X_{n_s})$ is an independent random sample from an exponential distribution with parameter μ. Hence, the likelihood function of λ and μ based on observations $\mathbf{x}, \mathbf{y}$ is given by:

$$\ell_1(\lambda, \mu) \propto \lambda^{n_a} \;\; e^{-\lambda t_a} \;\; \mu^{n_s} \;\; e^{-\mu t_s} \tag{3.1}$$

where $t_a = \sum y_i$ and $t_s = \sum x_i$ are the observed values of the sufficient statistics. The above experiment will be our basic experiment and we refer to it as $\mathcal{E}_1$.

If the queue is in equilibrium, observing the initial system size — customarily denoted by ν — can be relatively easy and can provide additional information about the queue.

Accordingly, we consider experiment $\mathcal{E}_2$ that observes $\mathbf{x}$ and $\mathbf{y}$ as in $\mathcal{E}_1$, and also observes ν. The probability of ν customers in the system can be derived from (2.1), and $\mathcal{E}_2$ results in the following likelihood function:

$$\ell_2(\lambda, \mu) \propto \lambda^{n_a} e^{-\lambda t_a} \quad \mu^{n_s} \quad e^{-\mu t_s} \quad \left(\frac{\lambda}{\mu}\right)^{\nu} \quad e^{-\lambda/\mu}. \tag{3.2}$$

(This is certainly the same likelihood obtained in the experiment considered by Benes (1957), which simply adds the observation of ν to the experiment he considered when the queue is not necessarily in equilibrium and that was described at the beginning of this section.)

It sometimes happens that experimenters do not have access to the service facility, or that service times are difficult or very expensive to observe. It so happens that, in an $M/M/\infty$ queue in equilibrium, we still can make inferences and predictions based on the observed initial system size, ν, and the n_a inter-arrival times, $\mathbf{x}$, even if the service facility is inaccessible. We refer to this experiment as $\mathcal{E}_3$, which results in the likelihood

$$\ell_3(\lambda, \mu) \propto \lambda^{n_a} e^{-\lambda t_a} \left(\frac{\lambda}{\mu}\right)^{\nu} e^{-\lambda/\mu}. \tag{3.3}$$

Likewise, it might be possible to observe ν and the service times (which is the case when the service facility is easy to observe, since ν is also the initial number of busy servers) but very difficult, time consuming, expensive, or even impossible, to keep a detailed record of the arrivals. Hence, we consider still another experiment, $\mathcal{E}_4$, which consists in observing ν and $\mathbf{y}$, resulting in the likelihood

$$\ell_4(\lambda, \mu) \propto \mu^{n_s} e^{-\mu t_s} \left(\frac{\lambda}{\mu}\right)^{\nu} e^{-\lambda/\mu}. \tag{3.4}$$

These four experiments do produce the basic four types of likelihood functions that we most usually encounter when observing a $M/M/\infty$ queue. Sometimes experimental conditions will determine the experiment to perform, but it may also happen that we can choose among two or more of them. A complete approach to the problem of choosing the experiment to perform should include not only the costs of the different experiments, but also the aim (estimation, prediction, ... etc.) of experimentation. (It is indeed well known that experiments that are best, according to most criteria, for one of the parameters may not be so for a different parameter.) We will only attempt here a simple, yet illuminating, exploratory comparison among the experiments, in terms of the asymptotic variance of both parameters.

It is a standard result in Bayesian statistics that (under suitable regularity conditions), for large n, the approximate posterior distribution of a parameter vector φ is normal with mean vector φ and variance-covariance matrix given by $I_n^{-1}(\hat{\varphi})$, where $\hat{\varphi}$ is the maximum likelihood estimate of φ and $I_n^{-1}(\hat{\varphi})$ is the inverse of Fisher information matrix evaluated at the MLE $\hat{\varphi}$. The asymptotic posterior variances of λ and μ, that we will denote by $\mathrm{Var}^*(\lambda)$ and $\mathrm{Var}^*(\mu)$, can then be directly obtained from $I_n^{-1}(\hat{\mu}, \hat{\lambda})$, as well as the asymptotic covariance, $\mathrm{Cov}^*(\lambda, \mu)$, of λ and μ. A lengthy but otherwise straightforward computation provides the inverses of the Fisher information matrices for the four experiments, and the resulting $\mathrm{Var}^*(\mu)$, $\mathrm{Var}^*(\lambda)$ and $\mathrm{Cov}^*(\lambda, \mu)$ are described below. We assume $n_a = n_s = n$ throughout. (For convenience, and later use, we provide the expressions for the Fisher information matrices in the Appendix.)

Experiment $\mathcal{E}_1$. (Observe inter-arrivals and service times.)

$$\mathrm{Var}_1^*(\lambda) = \frac{\hat{\lambda}^2}{n}, \quad \mathrm{Var}_1^*(\mu) = \frac{\hat{\mu}^2}{n}, \quad \mathrm{Cov}_1^*(\lambda, \mu) = 0. \tag{3.5}$$

As we could expect, this is a "middle" experiment in terms of the information provided (as measured in terms of asymptotic variances) when compared with the other three. It is also the only one for which λ and μ are independent.

Experiment $\mathcal{E}_2$. (Observe inter-arrivals and service times, and initial size.)

$$\begin{aligned}
\mathrm{Var}_2^*(\lambda) &= \frac{n\hat{\mu} + \hat{\lambda}}{n\hat{\mu} + 2\hat{\lambda}} \; \frac{\hat{\lambda}^2}{n} = \frac{n + \hat{\theta}}{n + 2\hat{\theta}} \; \mathrm{Var}_1^*(\lambda) \\
\mathrm{Var}_2^*(\mu) &= \frac{n\hat{\mu} + \hat{\lambda}}{n\hat{\mu} + 2\hat{\lambda}} \; \frac{\hat{\mu}^2}{n} = \frac{n + \hat{\theta}}{n + 2\hat{\theta}} \; \mathrm{Var}_1^*(\mu) \\
\mathrm{Cov}_2^*(\lambda, \mu) &= \frac{\hat{\mu}\hat{\lambda}^2}{n(n\hat{\mu} + 2\hat{\lambda})} = \frac{\hat{\theta}}{n + 2\hat{\theta}} \; \frac{\hat{\mu}\hat{\lambda}}{n}.
\end{aligned} \tag{3.6}$$

This is, as common sense suggests, the most informative experiment of the four considered. Both variances are smaller than the corresponding ones in $\mathcal{E}_1$ by a factor that, depending on the load, ranges between 1 and 2. Thus, observing the initial system size does not add much information (according to this criterion) when the load θ is small ($\theta \to 0$), but the asymptotic variance can be reduced by up to a half if we do observe ν in very busy systems ($\theta \to \infty$). Also, λ and μ are no longer independent.

Experiment $\mathcal{E}_3$. (Observe inter-arrival times and initial size.)

$$\mathrm{Var}_3^*(\lambda) = \frac{\hat{\lambda}^2}{n} \quad = \mathrm{Var}_1^*(\lambda) \quad . \tag{3.7}$$

Again we find that (3.7) completely agrees with intuition. From (3.7) we see that we lose nothing in estimating λ as compared to $\mathcal{E}_1$, since they are equivalent in this regard. Also, from the expression for Fisher information given in the Appendix, it can be seen that we do not have a consistent estimator of μ in this experiment (which should be expected), so asymptotic normality does not apply to the joint distribution, or to the marginal distribution of μ. It should be noted, however, that λ is asymptotically normal (Ghosh, 1993).

Experiment $\mathcal{E}_4$. (Observe service times and initial size.)

$$\mathrm{Var}_4^*(\lambda) = \frac{n\hat{\mu} + \hat{\lambda}}{\hat{\lambda}} \; \frac{\hat{\lambda}^2}{n} = \left(\frac{n}{\hat{\theta}} + 1\right) \; \mathrm{Var}_1^*(\lambda) \quad . \tag{3.8}$$

This experiment exhibits a completely symmetrical behavior to that of experiment $\mathcal{E}_3$, λ now being the parameter that we can not consistently estimate, and analogous comments apply.

4. The Kummer distribution

Before presenting the statistical analysis of the $M/M/\infty$ queue, we introduce a new family of continuous distributions that appear frequently in the derivations to come.

Definition 4.1 - A random variable X has a *Kummer distribution* with parameters α, β, γ, δ ($\alpha > 0$, $\beta > 0$, $\delta > 0$) if it has a continuous distribution whose p.d.f. is

$$Ku(x|\alpha,\beta,\gamma,\delta) = C\frac{x^{\alpha-1}e^{-\beta x}}{(1+\delta x)^{\gamma}}, \quad x > 0 \tag{4.1}$$

and $Ku(x|\alpha,\beta,\gamma,\delta) = 0$ otherwise, where the proportionality constant C is such that

$$C^{-1} = \frac{\Gamma(\alpha)}{\delta^{\alpha}} \ U(\alpha, \ \ \alpha+1-\gamma, \ \ \beta/\delta), \tag{4.2}$$

and $U(a,b,z)$ is one of Kummer's functions (a confluent hypergeometric function), with integral representation, for $a > 0, z > 0$, given by

$$\Gamma(a)U(a,b,z) = \int_0^{\infty} e^{-zt} \ \ t^{a-1} \ \ (1+t)^{b-a-1} \ dt. \tag{4.3}$$

(See, for instance, Abramowitz and Stegun, 1964.) □

Kummer's function U is a standard mathematical function and appears in many well known mathematical packages. Sometimes, $U(a,b,z)$ is also denoted by $\Psi(a;b;z)$, or by $z^{-a}{}_2F_0\ (a, \ \ a+1-b; \ ;-1/z)$. We highlight two special cases that will be needed for future reference:

$$U(a,a+1,z) = \ 1/z^{a}. \tag{4.4}$$

This follows directly from (4.3) and the definition of the gamma function. Also, even though the function U is not defined for $z=0$, we abuse notation and write

$$\Gamma(a)U(a,b,0) = \int_0^{\infty} t^{a-1} \ \ (1+t)^{b-a-1} \ dt. \tag{4.5}$$

But the integral on the RHS of (4.5) *does* converge for $b < 1$, and it is given by $\Gamma(a)\Gamma(1- \ b)/\Gamma(a+1-b)$. Hence we can define

$$U(a,b,0) = \frac{\Gamma(1-b)}{\Gamma(a+1-b)}, \quad \text{for } b < 1, \tag{4.6}$$

and extend the definition of $U(a,b,z)$ to also allow $U(a,b,0)$, as given in (4.6), when $b<1$.

We have called the distribution in Definition 4.1 the *Kummer distribution* because it is derived from Kummer's function in much the same way as the Gamma distribution is derived from the Gamma function and the beta distribution from the Beta function. It generalizes both the Gamma and the F distributions. Indeed, from (4.1), (4.2), (4.4), and (4.6), the following equivalences can be established:

$$X \sim Ku(\alpha,\beta,0,\delta) \to X \sim Ga(\alpha,\beta)$$

$$\text{for } \gamma > \alpha: \quad X \sim Ku(\alpha,0,\gamma,\delta) \to \frac{(\gamma-\alpha)\delta}{\alpha}X \sim F(2\alpha, \quad 2\gamma-2\alpha). \tag{4.7}$$

Clearly, $F(\nu_1,\nu_2) = Ku(\frac{\nu_1}{2},0,\frac{\nu_1+\nu_2}{2},1)$, but the relation in (4.7) will be more directly applicable for our purposes. Finally, from (4.1) it can be seen that $Ku(x|\alpha,\beta,\gamma,0)$ for $\alpha>0,\ \beta>0$ can be properly defined as long as the constant C is calculated directly, without resorting to the U function. Hence, we shall also use this special case and the fact that

$$X \sim Ku(\alpha,\beta,\gamma,0) \to X \sim Ga(\alpha,\beta). \tag{4.8}$$

The moments of a $Ku(\alpha,\beta,\gamma,\delta)$ distribution can easily be computed in terms of the U function and they are given by

$$E(X^k) = \frac{\Gamma(\alpha+k)}{\delta^k\Gamma(\alpha)} \ \frac{U(\alpha+k, \quad \alpha+k+1-\gamma, \quad \beta/\delta)}{U(\alpha, \quad \alpha+1-\gamma, \quad \beta/\delta)}. \tag{4.9}$$

5. Posterior distributions

When introducing and studying the properties of a $M/M/\infty$ queue, it is natural to do so in terms of λ and μ, since they are the parameters governing the arrivals to and departures from the system. Nevertheless, it can be seen from (2.1), (2.2) and (2.3) that the distributions of the relevant measures of performance of the queue depend on λ and μ through $\theta = \lambda/\mu$ and μ. (Recall also that θ is the expected steady-state number of busy servers or number of customer in the system, and that μ is the inverse of the average time that a customer spends in the system at steady-state.) Hence, for inferential and prediction purposes, it seems more natural to work in terms of μ and θ, and we will do so from now on.

It can be seen, from (3.1), (3.2), (3.3), and (3.4), that the four likelihoods for (θ, μ), corresponding to the four different experiments, can be expressed in a unified way as

$$\ell(\mu, \theta) \propto \mu^{n_t} \quad e^{-\mu(\delta_3 t_s + \delta_4 t_a \theta)} \quad \theta^m \quad e^{-\delta_1 \theta}, \tag{5.1}$$

where $n_t = \delta_4 n_a + \delta_3 n_s$, $m = \delta_4 n_a + \delta_1 \nu, t_s = \sum_1^{n_s} x_i$ (total of service times), and $t_a = \sum_1^{n_a} y_i$ (total of inter-arrival times) are observed, and $\delta_1, \delta_3, \delta_4$ are $0-1$ constants that identify the performed experiment, so that $\delta_i = 0$ in $\mathcal{E}_i$, $i = 1, 3, 4$. Notice that $\delta_i = 0$, for some i, implies $\delta_j = 1$ for $j \neq i$, $i = 1, 3, 4$. Also, $\delta_1 = \delta_3 = \delta_4 = 1$ for experiment $\mathcal{E}_2$. A conjugate prior density based on the form (5.1) is

$$p(\mu, \theta) \propto \quad \mu^{n_0 - 1} \quad e^{-\mu(b_0 + \beta_0 \ \theta)} \quad \theta^{\alpha_0 - 1} \quad e^{-k_0 \theta}, \tag{5.2}$$

and it defines a proper density for $n_0 > 0, b_0 > 0, \beta_0 > 0, \alpha_0 > 0$ and $k_0 > 0$. It is also proper for $K_0 = 0$ provided that $n_0 > \alpha_0$, for $\beta_0 = 0$ provided that $b_0 > 0$, and also for $b_0 = 0$, provided that $\alpha_0 > n_0$ and $\beta_0 > 0$. A distribution with the density $p(\mu, \theta)$ in (5.2) will be called, following the customary procedure, a *Gamma-Kummer distribution*, and denoted by $GaKu(n_0, b_0, \beta_0, \alpha_0, k_0)$. The name stems from the fact that, if $(\mu, \theta) \sim Ga(n, b, \beta, \alpha, k)$ then the conditional distribution of μ is Gamma,

$$\begin{aligned} p(\mu \mid \theta) &= Ga(\mu \mid n, b + \beta\theta) \\ &= \frac{(b + \beta\theta)^n}{\Gamma(n)} \quad \mu^{n-1} \quad e^{-\mu(b + \beta\theta)}, \quad \mu > 0, \end{aligned} \tag{5.3}$$

and, for $b > 0$, the marginal distribution of θ is Kummer,

$$p(\theta) = Ku(\theta | \alpha, k, n, \beta/b) = \frac{(\beta/b)^\alpha}{\Gamma(\alpha) U(\alpha, \alpha + 1 - n, bk/\beta)} \quad \frac{\theta^{\alpha - 1} \quad e^{-k\theta}}{[1 + \theta(\beta/b)]^n}, \quad \theta > 0. \tag{5.4}$$

If $b = 0$, and $\alpha > n$, it can be seen from (5.2) and (5.3) that the marginal distribution of θ is a Gamma

$$p(\theta) = Ga(\theta | \alpha - n, \quad k) = \frac{k^{\alpha - n}}{\Gamma(\alpha - n)} \quad \theta^{\alpha - n - 1} \quad e^{-k\theta}, \quad \theta > 0. \tag{5.5}$$

For ease of notation, whenever $b = 0$ (with $\alpha > n$) in a $Ku(\theta | \alpha, k, n, \beta/b)$, we will consider it to represent the Gamma density in (5.5).

If the prior distribution of (θ, μ) is $GaKu(n_0, b_0, \beta_0, \alpha_0, k_0)$, then it follows from (5.1) and (5.2) that the posterior density is given by

$$p(\mu, \theta|\text{data}) \quad \propto \quad \ell(\mu, \theta) \;\; p(\mu, \theta) = GaKu(\mu, \theta|n_1, b_1, \beta_1, \alpha_1, k_1), \tag{5.6}$$

where $n_1 = n_t + n_0$, $b_1 = b_0 + \delta_3 t_s$, $\beta_1 = \beta_0 + \delta_4 ta$, $\alpha_1 = \alpha_0 + m$, $k_1 = k_0 + \delta_1$. Therefore, the Gamma-Kummer family of distributions is conjugate for *all* the experiments considered. It is indeed remarkable, and quite a peculiar characteristic of this simple queueing system, that we can use the same *conjugate* prior for such different experiments. (It should be noted that the likelihood functions are *not* proportional.)

For assessment purposes, we find it easiest to think in terms of the arrival rate λ and the mean service time $\frac{1}{\mu}$. Besides, in a $M/M/\infty$ queue, it is very natural to assume that, a priori, these two quantities are independent. Accordingly, we assume

$$p(\lambda, \mu^{-1}) = Ga(\lambda|\alpha_0, \beta_0) \;\; Ga^{-1}(\mu|a_0, b_0) \;\; \text{or} \;\; p(\lambda, \mu) = Ga(\lambda|\alpha_0, \;\; \beta_0) \;\; Ga(\mu|a_0, b_0),$$

which results in

$$p(\mu, \theta) = GaKu(\mu, \theta|\alpha_0 + a_0, b_0, \beta_0, \alpha_0, 0). \tag{5.7}$$

(Notice that, since $\alpha_0 + a_0 > \alpha_0$, $k_0 = 0$ is allowed.) It should be noted that λ and μ would be a posteriori independent (that is, $K_1 = 0$), *only* if experiment $\mathcal{E}_1$ is performed. We nevertheless see no reasons whatsoever, either operational or methodological, why the extra observation of ν in $\mathcal{E}_2$, for instance, should change the prior opinions about (μ, θ). Hence, we do propose (5.7) as a prior for *all* experiments. (The corresponding posterior is obviously given by (5.6) where, in the expression of n_1, $n_0 = \alpha_0 + a_0$ and $k_1 = \delta_1$. Notice that, if experiment $\mathcal{E}_1$ is performed, the distribution of θ is simply a re-scaled F distribution, and that, as just mentioned, λ and μ are independent a posteriori.)

The same comments apply to the choice of a convenient non-informative, automatic or default prior. Most non-informative priors are model-dependent, so a different prior would be chosen depending on the experiment to be performed, even if the likelihoods are proportional. It has been argued that, since the choice of the experiment can itself reflect prior information, this phenomenom is, not only natural, but even desirable. Thus, for instance, a Bernoulli parameter expected to be very small could make a Negative Binomial experiment be selected instead of a Binomial experiment, and the default prior should reflect this extra information. None of this would seem to apply to our experiments: which one is ultimately performed would typically depend only on what is easy or cheap to observe, and usually will not carry any information about λ nor μ. Hence, our proposal is to use a unique "automatic" prior in all four experiments, and we suggest using the Jeffreys prior for the *least* informative experiments, which as we saw in section 3, are experiments $\mathcal{E}_3$ and $\mathcal{E}_4$. (The Jeffreys priors for all four experiments are derived in the Appendix.) It turns out, not surprisingly, that both these non-informative priors are equal and given by

$$p^N(\mu, \theta) \propto \quad \mu^{-1} \quad \theta^{-1/2}. \tag{5.8}$$

The non-informative (5.8) can be obtained from (5.2) by letting $n_0 = 0,\ b_0 = 0,\ \beta_0 = 0,\ \ \alpha_0 = \frac{1}{2},\ \ k_0 = 0$, and it is obviously improper. The corresponding posterior distribution is

$$p^N(\mu, \theta|data) = GaKu(\mu, \theta|\delta_4 n_a + \delta_3 n_s, \delta_3 t_s, \delta_4 t_a, \delta_4 n_a + \delta_1 \nu + \frac{1}{2}, \delta_1), \tag{5.9}$$

which can be checked to be proper for *all four* experiments. (If we had used the Jeffreys prior for experiment 1, then the resulting posteriors would be improper if experiments $\mathcal{E}_3$ or $\mathcal{E}_4$ were performed and $\nu = 0$ were observed. Similarly, use of the Jeffreys prior for experiment $\mathcal{E}_2$ can result in improper posteriors when used with data from experiments $\mathcal{E}_1, \mathcal{E}_3$ or $\mathcal{E}_4$.) From (5.9), (5.4) and (5.5), the marginal posterior distribution for θ is

$$p^N(\theta|data) = Ku(\theta|\delta_1 n_a + \delta_1 \nu + \frac{1}{2},\quad \delta_1, \delta_4 n_a + \delta_3 n_s,\quad \delta_4 ta/\delta_3 ts), \tag{5.10}$$

or, more explicitly:

$$\begin{aligned} \frac{(n_s - 0.5)/t_s}{(n_a + 0.5)/t_a}\ \theta \sim F(2n_a + 1,\quad 2n_s - 1), \quad &\text{for } \mathcal{E}_1, \\ \theta \sim Ku(n_a + \nu + \frac{1}{2},\quad 1,\quad n_a + n_s,\ t_a/t_s), \quad &\text{for } \mathcal{E}_2, \\ \theta \sim Ga(\nu + \frac{1}{2},\quad 1),\ \text{for } \mathcal{E}_3 \text{ and } \mathcal{E}_4.& \end{aligned} \tag{5.11}$$

We recall that the vast majority of experiments proposed for analyzing queues result in likelihoods, and hence posterior distributions, as in $\mathcal{E}_1$. For this important particular case, the non-informative (5.11) results in the usual estimator

$$E(\theta|data) = \frac{2n_s - 1}{2n_s - 3}\ \frac{(n_a + 0.5)/t_a}{(n_s - 0.5)/t_s}, \tag{5.12}$$

which, for moderate n_a, n_s, will be very similar to the MLE $\hat{\theta} = \hat{\lambda}/\hat{\mu} = (n_a/t_a)/(n_s/t_s)$. Another interesting feature in (5.11) is the behavior under the $\mathcal{E}_3$ and $\mathcal{E}_4$ experiments, for which observations in either the service or arrival processes are lacking. (5.11) then says that, with no prior information (or, more specifically, with the given non-informative prior that keeps θ and μ independent), observations of the other process do not help in estimating θ, and hence both experiments result in the same posterior.

Another parameter of interest in a $M/M/\infty$ queue is μ. In general, the joint posterior distribution of (μ, θ) is a $GaKu(\mu, \theta | n_1, b_1, \ \beta_1, \ \alpha_1, k_1)$, as given in (5.6). Then, directly from (5.2), it follows that

$$\begin{aligned} p(\mu|data) &\propto \mu^{n_1-1} \ e^{-\mu b_1} \int_0^\infty \theta^{\alpha_1} \ e^{-\theta(k_1+\mu\beta_1)} \ d\theta \\ &\propto \frac{\mu^{n_1-1} \ e^{-\mu b_1}}{(k_1+\mu\beta_1)^{\alpha_1}} \propto Ku(\mu|n_1, b_1, \alpha_1, \beta_1/k_1). \end{aligned} \tag{5.13}$$

With the proposed informative prior (5.7), $\mu \sim Ga(a_0, b_0)$ a priori, and if experiment $\mathcal{E}_1$ is performed, $k_1 = 0$. Hence, it follows from (5.5) that

$$p(\mu|data) = Ga(\mu|n_1 - \alpha_1, b_1) = Ga(\mu|a_0 + n_s, b_0 + t_s). \tag{5.14}$$

Also, non-informative analysis results in the joint posterior (5.9), so that (5.13) becomes

$$p^N(\mu|data) = Ku(\mu|\delta_4 n_a + \delta_3 n_s, \ \ \delta_3 t_s, \ \ \frac{1}{2} + \delta_4 n_a + \delta_1 \nu, \delta_4 t_a/\delta_1), \tag{5.15}$$

or, more explicitly:

$$\begin{aligned} \mu &\sim Ga(n_s - \frac{1}{2}, t_s), \quad \text{for } \mathcal{E}_1, \\ \mu &\sim Ku(n_a + n_s, \ \ t_s, \ \ \frac{1}{2} + n_a + \nu, \ \ t_a), \quad \text{for } \mathcal{E}_2, \\ \frac{(\nu + \frac{1}{2})}{(n_a/t_a)} \ \mu &\sim F(2n_a, 1 + 2\nu), \quad \text{for } \mathcal{E}_3, \\ \mu &\sim Ga(n_s, t_s), \quad \text{for } \mathcal{E}_4. \end{aligned} \tag{5.16}$$

Similarly to the marginal analysis for θ, it is also true here that the Bayes estimate in $\mathcal{E}_1$, $E(\mu|data) = (n_s - \frac{1}{2})/t_s$, is very close to the MLE $\hat{\mu} = n_s/t_s$. Experiments $\mathcal{E}_3$ and $\mathcal{E}_4$ also behave as expected: When we do not have observations on service times, then we use both the initial size ν and observations on the arrival process to estimate μ; on the other hand, if we do have observations on the service times (and none on the arrivals) then the added information of ν becomes irrelevant for estimating μ. Notice that the only learning process in which μ and θ are independent a posteriori takes place with $\mathcal{E}_4$ *and* a non-informative prior.

Usually, the parameter λ is not as important as θ or μ. Nevertheless, it might be of interest in some problems and we next derive its marginal posterior distribution. If, in general, the joint density of (μ, θ) is $GaKu(\theta, \mu | n_1,\ b_1,\ \beta_1,\ \alpha_1,\ k_1)$, then the density of (μ, λ) can be deduced to be

$$p(\mu, \lambda | data) = C\mu^{n_1-\alpha_1-1}\ \ e^{-b_1\mu}\ e^{-\lambda k_1/\mu}\ \ \lambda^{\alpha_1-1}\ e^{-\beta_1\lambda}, \tag{5.17}$$

where $C^{-1} = \ \Gamma(n_1)\ \ \Gamma(\alpha_1)\ U(\alpha_1, \alpha_1 + 1 - n_1, k_1b_1/\beta_1)\ \ /\ (b_1^{n_1-\alpha_1}\beta_1^{\alpha_1})$. Hence

$$p(\lambda | data) = C\lambda^{\alpha_1-1}\ \ e^{-\lambda\beta_1} \int_0^\infty \mu^{n_1-\alpha_1-1}\ \ e^{-\mu b_1}\ \ e^{-\lambda k_1/\mu}\ \ d\mu. \tag{5.18}$$

The integral in (5.18) must be evaluated numerically, in general. When experiment $\mathcal{E}_1$ is performed and the prior (5.7) or the non-informative (5.8) is used, then $k_1 = \delta_1 = 0$, so that

$$p(\lambda | data) = Ga(\lambda | \alpha_1, \beta_1), \qquad\qquad \text{for } \mathcal{E}_1. \tag{5.19}$$

In the non-informative case, $\alpha_1 = \frac{1}{2} + n_a$, $\ \beta_1 = t_a$. Also, when experiment $\mathcal{E}_3$ is performed and the non-informative (5.8) is used, then it follows from (5.9) that $b_1 = 0$ and

$$\begin{aligned} p^N(\lambda | data) &\propto \lambda^{n_a+\nu-\frac{1}{2}}\ \ e^{-\lambda ta} \int_0^\infty \ \frac{1}{\mu^{\nu+\frac{3}{2}}}\ \ e^{-\lambda/\mu}\ \ d\mu \\ &\propto Ga(\lambda | n_a, t_a), \end{aligned} \tag{5.20}$$

so that $\mathcal{E}_1$ and $\mathcal{E}_3$ result in virtually identical (for moderate n_a) posterior distributions for λ (and λ and μ are independent a posteriori.)

Even though (5.18) can not usually be expressed in closed form, the moments can easily be computed in terms of Kummer's function U. We derive here the mean for further use.

$$
\begin{aligned}
E(\lambda|data) &= \int_0^\infty \lambda \; p(\lambda|data) \;\; d\lambda \\
&= C \int_0^\infty \quad \mu^{n_1-\alpha_1-1} \;\; e^{-\mu b_1} \quad \frac{\Gamma(\alpha_1+1)}{(\beta_1+k_1/\mu)^{\alpha_1+1}} \; d\mu \\
&= \left(\frac{\alpha_1 n_1}{\beta_1}\right) \;\; \left(\frac{b_1 k_1}{\beta_1}\right)^{n_1-\alpha_1} \quad \frac{U(n_1+1, n_1-\alpha_1+1, k_1 b_1/\beta_1)}{U(\alpha_1, \alpha_1+1-n_1, k_1 b_1/\beta_1)} \\
&= \frac{\alpha_1 n_1}{\beta_1} \;\; \frac{U(\alpha_1+1, \;\; \alpha_1+1-n_1, \;\; k_1 b_1/\beta_1)}{U(\alpha_1, \;\; \alpha_1+1-n_1, \;\; k_1 b_1/\beta_1)},
\end{aligned}
\tag{5.21}
$$

where the last equality in (5.21) follows from the relation $U(a, b, z) = z^{1-b} \;\; U(1+a-b, \;\; 2-b, z)$ (see Abramowitz and Stegun, 1964). It can be checked that expectations for the distributions (5.19) and (5.20) can be obtained as particular cases.

6. Prediction under steady-state.

As mentioned in the introduction, the typical objective when analyzing a queue is to learn about measures of performance of the queue. In a $M/M/\infty$ queue in equilibrium, such measures include the number of customers in the system, N, which is equal to the number of busy servers, and the waiting time, W, which is the time a customer spends in the system, and, since the customer never waits in line, is equal to the service time. Queueing theoreticians analyzing queues from a probability point of view aim to obtain expressions for the *mean* of these quantities (for any given values of the unknown parameters) and to obtain important relationships among some of them; classical statisticians aim to obtain estimators of these mean values. Bayesians can not only compute estimators but can obtain their entire posterior distributions, as we saw in Section 5, since $E(N|\lambda, \mu) = \theta$ and $E(W|\lambda, \mu) = 1/\mu$. Bayesian methods allow one to even go one step further and determine the predictive distributions of N and W, from which *direct* probability statements can be made about these quantities.

We derive first the predictive distribution of N. It follows from (2.1) that $p(N = n|\mu, \lambda) = p(N = n|\theta) =$ Poisson $(n|\theta)$, and from (5.4) and (5.6) that the posterior distribution of θ is $Ku(\alpha_1,\ \ k_1,\ n_1,\ \ \beta_1/b_1)$. Hence, for $b_1 > 0, \beta_1 > 0$,

$$\begin{aligned}
p(n|data) &= \int_0^\infty \ \frac{\theta^n e^{-\theta}}{n!}\ \ Ku(\theta|\alpha_1, k_1, n_1, \beta_1/b_1)\ d\theta \\
&= \frac{(\beta_1/b_1)^{\alpha_1}}{n!\ \ \Gamma(\alpha_1)\ \ U(\alpha_1, \alpha_1+1-n_1, b_1k_1/\beta_1)} \int_0^\infty \frac{\theta^{n+\alpha_1-1}\ \ e^{\theta(k_1+1)}}{(1+\beta_1/b_1)^{n_1}}\ d\theta \\
&= \frac{\Gamma(\alpha_1+n)}{n!\ \ \Gamma(\alpha_1)(\beta_1/b_1)^n}\ \frac{U(\alpha_1+n, \alpha_1+n+1-n_1, (k_1+1)b_1/\beta_1)}{U(\alpha_1, \alpha_1+1-n_1, b_1k_1/\beta_1)}\ . \qquad (6.1)
\end{aligned}$$

The only substantial simplification of (6.1) occurs under non-informative analysis with experiments $\mathcal{E}_3$ (so that $b_1 = 0$) and $\mathcal{E}_4$ (so that $\beta_1 = 0$). If follows that then $p(\theta|data) = Ga(\theta|\nu + \frac{1}{2}, 1)$ and

$$p^N(n|data) = \frac{\Gamma(n+\nu+\frac{1}{2})}{n!\ \ \Gamma(\nu+\frac{1}{2})\ \ 2^{n+\nu+\frac{1}{2}}}. \qquad (6.2)$$

(Distributions, such as (6.2), appear frequently in Bayesian analysis and are sometimes referred to as *Poisson-Gamma.*) Minor simplification also occurs in the denominator of (6.1) when $k_1 = 0$ (experiment $\mathcal{E}_1$), in which case the U function in the denominator reduces to $\Gamma(n_1 - \alpha_1)/\Gamma(n_1)$ (see (4.6)).

Since $E(N|\theta) = \theta$, it follows that $E(N|data) = E(\theta|data)$ so that, from (4.9) (assume $\beta_1, b_1 > 0$),

$$E(N|data) = \frac{\alpha_1 U(\alpha_1+1, \alpha_1+2-n_1, k_1b_1/\beta_1)}{(\beta_1/b_1)\ \ U(\alpha_1, \alpha_1+1-n_1, k_1b_1/\beta_1)}. \qquad (6.3)$$

Higher order moments can similarly be computed as simple functions of $E(\theta^r|data)$, and these can immediately be derived from (4.9). Most importantly, from (6.1) we can compute probabilities of *direct* interest, such as the probability that the system is heavily utilized ($Pr(N > N_0|data)$ for some N_0), or the probability that the system is empty,

$$Pr(N = 0|data) = \frac{U(\alpha_1, \alpha_1+1-n_1, (K_1+1)b_1/\beta_1)}{U(\alpha_1, \alpha_1+1-n_1, K_1b_1/\beta_1)}. \qquad (6.4)$$

We now turn to the computation of the predictive distribution of the waiting time, W. It follows from (2.2), or simply from the definition of the $M/M/\infty$ queue, that $p(w|\mu, \lambda) = p(w|\mu) = Ex(w|\mu)$. Also, in general, the marginal distribution of μ, as given by (5.13), is $Ku(\mu|n_1, b_1, \alpha_1, \beta_1/k_1)$. Therefore, for $k_1 > 0, \beta_1 > 0$,

$$p(w|data) = \int_0^\infty \mu e^{-\mu w}\ \ Ku(\mu|n_1, b_1, \alpha_1, \beta_1/k_1)\ \ d\mu$$

$$= \frac{(\beta_1/k_1)^{n_1}}{\Gamma(n_1)\ \ U(n_1, n_1+1-\alpha_1, \beta_1/k_1)} \int_0^\infty \frac{\mu^{n_1}\ \ e^{-\mu(w+b_1)}}{(1+\mu\beta_1/k_1)^{\alpha_1}}\ d\mu$$

$$= \frac{k_1 n_1\ \ U(n_1+1, n_1+2-\alpha_1, (w+b_1)k_1/\beta_1)}{\beta_1\ \ U(n_1, n_1+1-\alpha_1, \beta_1/k_1)}. \tag{6.5}$$

For $k_1 = 0$, that is, for experiment $\mathcal{E}_1$ (with both the informative (5.7) and non-informative (5.8)), it follows from (5.5) that $p(\mu|data) = Ga(\mu|n_1 - \alpha_1, b_1)$, so that

$$p(w|data) = \frac{(n_1-\alpha_1)b_1^{(n_1-\alpha_1)}}{(w+b_1)^{n_1-\alpha_1+1}}. \tag{6.6}$$

Also, for $\beta_1 = 0$ (that is for the non-informative case with $\mathcal{E}_4$), it follows from (5.16) that $p(\mu|data) = Ga(n_s, t_s)$, so that $p(w|data)$ will be of the same form (6.6) with n_s, t_s substituted for $(n_1 - \alpha_1)$ and b_1 respectively. (Distributions, such as (6.6), are also common in Bayesian analysis and are sometimes called *Gamma-Gamma* distributions.)

Since $W \sim Ex(\mu)$, $E(W^r|\mu) = \Gamma(r+1)/\mu^r$. Also, if, in general, $\mu \sim Ku(\alpha, \beta, \gamma, \delta)$, then, for $r < \alpha$

$$E\left(\frac{1}{\mu^r}\right) = \frac{\delta^r\ \ \Gamma(\alpha - r)\ \ U(\alpha - r, \alpha - r + 1 - \gamma, \beta/\delta)}{\Gamma(\alpha)\ \ U(\alpha, \alpha+1-\gamma, \beta/\delta)}. \tag{6.7}$$

Thus, for $r < n_1$ (recall that $n_1 = n_0 + \delta_4 n_a + \delta_3 n_s$), $E(W^r|data) = \Gamma(r+1)\ \ E(1/\mu^r)$, as given in (6.7). Interestingly enough, and in contrast to the distribution of N, the predictive distribution of W has only $n_1 - 1$ moments. In particular (for $k_1 > 0$ and $\beta_1 > 0$),

$$E(W|data) = \frac{(\beta_1/k_1)\ \ U(n_1 - 1, n_1 - \alpha_1, k_1 b_1/\beta_1)}{(n_1 - 1)\ \ U(n_1, n_1 + 1 - \alpha_1, k_1 b_1/\beta_1)}. \tag{6.8}$$

An important result in queueing theory, known as *Little's formula*, applies to any system in statistical equilibrium (that meets very general regularity conditions) and establishes that the expected number of customers in the system is equal to the arrival rate times the expected time spent in the system. In our case, Little's formula establishes that

$$E(N|\lambda, \mu) = \lambda E(W|\mu). \tag{6.9}$$

A question that might arise is whether the relation also holds unconditionally, that is, whether or not

$$E(N|data) = E(\lambda|data)\ \ E(W|data)\ \ . \tag{6.10}$$

It follows from (6.9) that (6.10) holds only if the random variables λ and $E(W|\mu) = 1/\mu$ are independent a posteriori, which is only true for experiment $\mathcal{E}_1$ (with both the informative independent prior and the non-informative prior) and $\mathcal{E}_3$ with non-informative prior. This fact can also be checked directly. Indeed, it follows from the expressions for the posterior expectations of N, λ and W, as given in (6.3), (5.21) and (6.8), that (6.10) holds only if

$$U(\alpha_1, \alpha_1 + 1 - s_1, k_1 b_1/\beta_1) = n_1 \quad U(\alpha_1 + 1, \alpha_1 + 1 - n_1, k_1 b_1/\beta_1); \tag{6.11}$$

but, when applying the following general relation for the U function, (see Abramowitz and Stegun, 1964),

$$(b-a)U(a,b,z) + U(a-1,b,z) = zU(a,b+1,z),$$

we get

$$\begin{aligned} -n_1 \quad U(\alpha_1 + 1, \alpha_1 + 1 - n_1, k_1 b_1/\beta_1) \ + \quad U(\alpha_1, \alpha_1 + 1 - n_1, k_1 b_1/\beta_1) \\ = (k_1 b_1/\beta_1) \quad U(\alpha_1 + 1, \alpha_1 + 2 - n_1, k_1 b_1/\beta_1). \end{aligned} \tag{6.12}$$

The RHS of (6.12) equals 0 (which establishes (6.11)) only if $k_1 = 0$ (experiment $\mathcal{E}_1$) or $b_1 = 0$ (experiment $\mathcal{E}_3$ *and* non-informative analysis).

7. Transient behavior

An added bonus when studying $M/M/\infty$ queues is that the analysis of the transient behavior of the queue is relatively simple. Since we will not be assuming steady-state in this section, the only experiment that makes sense is experiment $\mathcal{E}_1$. Recall that the $Ga(\lambda|\alpha_0, \beta_0) \ \ Ga(\mu|a_0, b_0)$ prior for (λ, μ) is equivalent to

$$p(\theta, \mu) \propto \theta^{\alpha_0 - 1} \quad e^{-\beta_0 \theta \mu} \quad \mu^{a_0 + \alpha_0 - 1} \quad e^{-b_0 \mu}. \tag{7.1}$$

Hence, it seems more natural (and will be seen to also be more convenient computationally) to express $p(\theta, \mu)$ as

$$p(\theta, \mu) = p(\theta|\mu) \quad p(\mu) = Ga(\theta|\alpha_0, \quad \beta_0 \mu) \quad Ga(\mu|a_0, \quad b_0), \tag{7.2}$$

instead of using the $p(\theta)p(\mu|\theta)$ representation that we have been using so far. The posterior then is $Ga(\theta|\alpha_1, \beta_1\mu) \ \ Ga(\mu|a_1, b_1)$, with $\alpha_1 = \alpha_0 + n_a, \ \ \beta_1 = \beta_0 + t_a, \ \ a_1 = a_0 + n_s, \ b_1 = b_0 + t_s$.

Since, in this queue, the waiting time is simply the service time, whose distribution is time-independent, we only need to compute the predictive distribution of $N(t)$, the number of customers in the system at time t (also, the number of busy servers). We recall that $N(t)|\theta, \mu \sim Po(\theta(1-e^{-\mu t}))$, as given in (2.3). Hence:

$$\begin{aligned} p[N(t)=n|\mu, data] &= \int_0^\infty \quad Po(n|\theta(1-e^{-\mu t})) \;\; Ga(\theta|\alpha_1, \beta_1\mu) \;\; d\theta \\ &= \frac{\beta_1^{\alpha_1}\mu^{\alpha_1}}{\Gamma(\alpha_1)} \;\; \frac{\Gamma(\alpha_1+n)}{n!} \;\; \frac{(1-e^{-\mu t)n}\mu_1^{\alpha_1}}{(1-e^{-\mu t}+\beta_1\mu)^{\alpha_1+n}}, \end{aligned} \tag{7.3}$$

and finally

$$\begin{aligned} p[N(t)=n|data] &= \int_0^\infty \quad p[N(t)=n|\mu, data] \;\; Ga(\mu|a_1, b_1) \;\; d\mu \\ &= \frac{\beta_1^{\alpha_1} b_1^{a_1}}{\Gamma(\alpha_1)\;\;\Gamma(a_1)} \quad \frac{\Gamma(\alpha_1+n)}{n!} \quad \int_0^\infty \; \frac{(1-e^{-\mu t})^n \mu^{a_1+\alpha_1-1} \;\; e^{-\mu b_1}}{(1-e^{-\mu t}+\beta_1\mu)^{\alpha_1+n}} \;\; d\mu. \end{aligned} \tag{7.4}$$

The integral in (7.4) must be evaluated numerically, but it is a simple univariate integral.

The moments can be derived in closed form using the following result, whose proof is straightforward:

Result 7.1

If a random variable X has a Ga(a, b) distribution, then, for $n < a$

$$E\left(\frac{e^{-\gamma x}}{X^n}\right) \;=\frac{b^a \;\; \Gamma(a-n)}{\Gamma(a)(\gamma+b)^{a-n}}. \tag{7.5}$$

To compute $E[N(t)|data]$, note first that, since $E[N(t)|\mu, \theta] = \theta(1-e^{-\mu t})$, then

$$E[N(t)|\mu, data] = \frac{\alpha_1(1-e^{-\mu t})}{\beta_1\mu}, \tag{7.6}$$

so that, by (7.5),

$$\begin{aligned} E[N(t)|data] &= \frac{\alpha_1}{\beta_1} \;\; [E^{\mu/data}\left(\frac{1}{\mu}\right) - E^{\mu/data}\left(\frac{e^{-\mu t}}{\mu}\right)] \\ &= \frac{\alpha_1 b_1}{(a_1-1)\beta_1} \;\; \left[1+\left(\frac{b_1}{b_1+t}\right)^{a_1-1}\right] \;\; . \end{aligned} \tag{7.7}$$

In a similar way, since $E[N^2(t)|\mu, \theta] = \theta^2(1-e^{-\mu t})^2 + \theta(1-e^{-\mu t})$, it can be shown, after some algebra, that

$$\begin{aligned} E[N^2(t)|data] =& \frac{\alpha_1(\alpha_1+1)b_1^2}{(a_1-1)(a_1-2)\beta_1^2} + \frac{\alpha_1 b_1}{(a_1-1)\beta_1} + \frac{\alpha_1 b_1^{a_1}}{\beta_1(a_1-1)(b_1+t)^{a_1-1}} \\ &+ \frac{\alpha_1(\alpha_1+1)b_1^{a_1}}{\beta_1^2(a_1-1)(a_1-2)}\left[\frac{t}{(2t+b_1)^{a_1-2}} - \frac{2}{(t+b_1)^{a_1-2}}\right] \;\; . \end{aligned} \tag{7.8}$$

We finally consider an extremely interesting issue that we might want to explore when analyzing a transient $M/M/\infty$ queue, namely to investigate whether the queue is basically in equilibrium, or, more generally, to predict how long the queue will have to run so that its behavior, will, with high probability, be close enough to steady-state. The issue, in its entire complexity, would require careful determination of the analysis to be performed (for instance, predicting $N(t)$ for some t as opposed to predicting N) and determination of how "close enough" is going to be measured (maybe as some measure of distance, such as the Kullback-Leibler directed divergence between the predictives for $N(t)$ and N). In this paper, we content ourselves with a very simple, preliminary analysis that produces the desired prediction.

Our proposal is to report the value of t for which, with high posterior probability, the transient $p(N(t) = n|\theta, \mu)$ would be close enough to the steady-state $p(N = n|\theta)$, or equivalently, for which their ratio,

$$\frac{p(N(t) = n|\theta, \mu)}{p(N = n|\theta)} = (t - e^{-\mu t})^n \exp\,(\theta e^{-\mu t}), \tag{7.9}$$

is smaller than 1. Since (7.9) is a decreasing function of n, it follows that, if (7.9) is close to 1 for $n = 0$, then it will be close to 1 for all values of n. (We are here taking, as a measure of distance between the distributions of $N(t)$ and N, the maximum of the probability ratio; again, more sophisticated measures of distance between these two distributions could be used.) Hence, the desired t is such that

$$Pr\{\,\exp\,(\theta e^{-\mu t}) < 1 + \mathcal{E}|data\} > 1 - \alpha, \tag{7.10}$$

for fixed values of $\mathcal{E}$ and α. We can rewrite (7.10) as

$$Pr\{\theta e^{-\mu t} < \log(1 + \mathcal{E})|data\} > 1 - \alpha. \tag{7.11}$$

Defining $\omega_t = \theta e^{-\mu t}$, it is immediate from the joint posterior of (θ, μ) that

$$p(\omega_t, \mu|data) = Ga(\omega_t|\alpha_1, \beta_1 \mu e^{\mu t}) Ga(\mu|a_1, b_1). \tag{7.12}$$

From (7.12), the marginal posterior of ω_t can be numerically computed and the equation (7.11) numerically solved. To the best of our knowledge, this simple and useful analysis has no classical counterpart.

Appendix

We present here the Fisher information matrices for the four experiments $\mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3, \mathcal{E}_4$, as derived from the likelihoods (3.1), (3.2), (3.3), and (3.4), respectively. Under rather general regularity conditions, the Fisher information about a random vector θ from a sample $\mathbf{z} = (z_1, \ldots, z_n)$ with a joint density $f(\mathbf{z}|\theta)$ is given by the matrix $\mathbf{I}(\theta)$ with i-j element given by

$$I_{ij}(\theta) = -E_\theta\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j} \;\; f(\mathbf{z}|\theta)\right].$$

Also, the Jeffreys prior is given by $p^N(\theta) \propto det[\mathbf{I}(\theta)]^{\frac{1}{2}}$, where $\mathbf{I}(\theta)$ refers to the Fisher information in a single observation z and $det[\mathbf{A}]$ stands for the determinant of matrix $\mathbf{A}$. When we apply these results to (3.1), (3.2), (3.3), and (3.4) we obtain:

Experiment $\mathcal{E}_1$

$$\mathbf{I}_1(\lambda, \mu) = \begin{pmatrix} \frac{n_a}{\lambda^2} & 0 \\ 0 & \frac{n_s}{\mu^2} \end{pmatrix}, \quad \det[\mathbf{I}_1(\lambda, \mu] = \frac{n_a}{\lambda^2} \; \frac{n_s}{\mu^2},$$

$$p_1^N(\lambda, \mu) \propto \lambda^{-1} \; \mu^{-1}, \quad \text{and } p_1^N(\theta, \mu) \propto \theta^{-1}\mu^{-1}.$$

Experiment $\mathcal{E}_2$

$$\mathbf{I}_2(\lambda, \mu) = \begin{pmatrix} \frac{\mu n_a + \lambda}{\mu\lambda^2} & -\frac{1}{\mu^2} \\ -\frac{1}{\mu^2} & \frac{\mu n_s + \lambda}{\mu^3} \end{pmatrix}, \quad \det[\mathbf{I}_2(\lambda, \mu)] = \frac{(\mu n_a + \lambda)(\mu n_s + \lambda) - \lambda^2}{\mu^4\lambda^2} \quad ,$$

$$p_2^N(\lambda, \mu) \propto \frac{(\mu^2 + 2\mu\lambda)^{\frac{1}{2}}}{\mu^2\lambda}, \quad \text{and} \quad p_2^N(\theta, \mu) \propto \frac{(2\theta + 1)^{\frac{1}{2}}}{\theta} \; \frac{1}{\mu}.$$

Experiment $\mathcal{E}_3$

$$\mathbf{I}_3(\lambda, \mu) = \begin{pmatrix} \frac{\mu n_a + \lambda}{\mu\lambda^2} & -\frac{1}{\mu^2} \\ -\frac{1}{\mu^2} & \frac{\lambda}{\mu^3} \end{pmatrix}, \quad \det[\mathbf{I}_3(\lambda, \mu)] = \frac{n_a}{\lambda\mu^3},$$

$$p_3^N(\lambda, \mu) \propto \lambda^{-\frac{1}{2}}\mu^{-\frac{3}{2}}, \quad \text{and} \quad p_3^N(\theta, \mu) \propto \theta^{-\frac{1}{2}}\mu^{-1}.$$

Experiment $\mathcal{E}_4$

$$\mathbf{I}_4(\lambda,\mu)=\begin{pmatrix}\frac{1}{\mu\lambda} & -\frac{1}{\mu^2}\\ -\frac{1}{\mu^2} & \frac{n_s\mu+\lambda}{\mu^3}\end{pmatrix},\quad \det[\mathbf{I}_4(\lambda,\mu)]=\frac{n_s}{\lambda\mu^3},$$

$$p_4^N(\lambda,\mu)\propto\lambda^{-\frac{1}{2}}\mu^{-\frac{3}{2}},\quad\text{and}\quad p_4^N(\theta,\mu)\propto\theta^{-\frac{1}{2}}\mu^{-1}\quad .$$

References

Abramowitz, M., and Stegun, I.A. (1964). *Handbook of Mathematical Functions.* New York: Dover.

Armero, C. (1985). Bayesian analysis of $M/M/\infty/FIFO$ queues. In *Bayesian Statistics 2* (J.M. Bernardo, M.H. DeGroot, D.V. Lindley, and A.F.M. Smith, eds.), 613-618. Amsterdam: North-Holland.

Armero, C. (1993). Bayesian inference in Markovian queues. *Queueing Systems* (in press).

Armero, C., and Bayarri, M.J. (1993a). Bayesian prediction in $M/M/1$ queues. *Queueing Systems* (in press).

Armero, C., and Bayarri, M.J. (1993b). Prior assessments for prediction in queues. *The Statistician* (in press).

Basawa, J.V., and Prakasa Rao B.L.S. (1980). *Statistical Inference for Stochastic Processes.* New York: Academic Press.

Benes, V.E. (1957). A sufficient set of statistics for a simple telephone exchange model. *Bell Systems Technical Journal* **36**, 939-964.

Berger, J.O., and Wolpert, R.L. (1988). *The Likelihood Principle, Second Edition.* Hayward, CA: Institute of Mathematical Statistics Monograph Series.

Bhat, U.N., and Rao, S.S. (1987). Statistical analysis of queueing systems. *Queueing Systems* **1**, 217-247.

Butler, R.W., and Huzurbazar, A. (1993). Computation of Bayesian predictive distributions. Presented at the 1993 Joint Statistical Meetings of the ASA, Biometric Society (ENAR and WNAR), and IMS.

Cooper, R.B. (1990). Queueing theory. In *Handbooks in OR & MS, Vol. 2* (Heyman, D.P., and Sobel, M.J., eds.), chapter 10. Amsterdam: Elsevier Science Publishers B.V. (North-Holland).

Gosh, J.K. (1993). Personal communication.

Gross, D., and Harris, C.M. (1985). *Fundamentals of Queueing Theory, Second Edition.*

New York: Wiley.

Kendall, D.G. (1953). Stochastic processes occurring in the theory of queues and their analysis by the method of imbedded Markov chains. *Annals of Mathematical Statistics* **24**, 338-354.

Lehoczky, J. (1990). Statistical Methods. In *Handbooks in OR & MS, Vol. 2* (Heyman, D.P., and Sobel, M.J., eds.), Chapter 6. Amsterdam: Elsevier Science Publishers B.V. (North-Holland).

McGrath, M.F., Gross, D., and Singpurwalla, N.D. (1987). A Subjective Bayesian approach to the theory of queues I-modeling. *Queueing Systems* **1**, 317-333.

McGrath, M.F., and Singpurwalla, N.D. (1987). A Subjective Bayesian approach to the theory of queues II-inference and information in $M/M/1$ queues. *Queueing Systems* **1**, 335-353.

Muddapur, M.V. (1972). Bayesian estimates of parameters in some queueing models. *Annals of the Institute of Mathematics* **24**, 327-331.

Newell, G.F. (1982). *Applications of Queueing Theory, Second Edition.* London: Chapman and Hall.

Reynolds, J.F. (1973). On estimating the parameters in some queueing models. *Australian Journal of Statistics* **15 1**, 35-43.

Shruben, L., and Kulkarni, R. (1982). Some consequences of estimating parameters for the $M/M/1$ queue. *Operations Research Letters* **1**, 75-78.

Thiruvaiyaru, D. and Basawa, I.V. (1992). Empirical Bayes estimation for queueing systems and networks. *Queueing Systems* **11**, 179-202.

Wolf, R.W. (1957). Problems of statistical inference for birth and death queueing models. *Operation Research* **13**, 343-357.