Fast and Scalable Nonparametric Bayesian Prediction for the $M/G/1$ Queue

Steffen Ventz[1] and Pietro Muliere[2]

[1]Harvard School of Public Health and Dana-Farber Cancer Institute

[2]Bocconi University

October 25, 2014

Abstract

In this article we develop a nonparametric Bayesian approach to prediction for the $M/G/1$ queue, focusing on the imbedded semi-Markov process of the queue at the departure times. Our approach is motivated by queues with a large number of data points and high-frequency systems, where times consuming MCMC/ABC algorithms might be infeasible and a nonparametric approach is desirable to avoid parametric assumptions. We define a reinforced stochastic model for the analysis of the $M/G/1$ queue through a system of predictive distributions. Using the theory of partial exchangeable processes, we prove that the reinforced stochastic process is the predictive probability model of a Bayesian semi-Markov mixture model. This result enabled fast and scalable Bayesian nonparametric prediction based on the reinforced stochastic process.

1

1 INTRODUCTION

In this article we develop a nonparametric Bayesian approach to fast and scalable prediction for a single counter queueing system with a Poisson arrival process and a general service-time distribution G, called the $M/G/1$ queue.

Queueing systems are a useful class of models for a variety of applied problems, including internet traffic, parcel services, telecommunication, cloud computing and operations research. Most inference and prediction methods for queueing processes utilize the frequentist paradigm (Jain and Templeton, 1988; Pitts, 1994; Teugels and Vanmarcke, 1994). The Bayesian approach to the analysis of queueing systems has recently gained popularity and modern Markov-Chain Monte-Carlo (MCMC) methods enable simple, but computational demanding implementation. Bayesian analysis of queueing systems includes (Armero, 1985; McGrath and Singpurwalla, 1987; Armero and Bayarri, 1994b,a, 1996; Armero and Conesa, 2000, 2006) for Markovian queues and (Wiper, 1998; Butler and Huzurbazar, 2000; Insua et al., 1998; Wiper et al., 2001; Ausin et al., 2004) for the analysis of non-Markovian queueing systems, where either the inter-arrival time or the service-time distribution is non-exponential. Moreover, (Ausin et al., 2008; Sutton and Jordan, 2011) discussed Bayesian methods for the general $G/G/1$ queues and queueing networks.

In the current work we introduce a Bayesian nonparametric predictive approach to the analysis of the $M/G/1$ queue. The queueing model has been treated previously in a Bayesian semi-parametric approach in (Ausin et al.,

2004) and in a nonparametric Bayesian fashion in (Conti, 1999, 2004). In (Conti, 1999, 2004) the author discussed a queue with discrete arrival and service-times, whereas we focus on a nonparametric analysis for continuous arrival time and an arbitrary service-time distribution. We focus on the imbedded semi-Markov process of the $M/G/1$ queue at the departure times. The imbedded process has the same limiting behavior as the underlying $M/G/1$ queue and provides an alternative experimental design for inference of the queue in transient and steady state. An interesting MCMC and approximate Bayesian computation algorithm (ABC) for the imbedded departure process for a queueing model with uniform $U(\theta_1, \theta_2)$ service-time distribution has been studied recently in (Shestopaloff and Neal, 2014; Blum and François, 2010). Our predictive approach is motivated by queues with a large number of data points and high-frequency systems, where times consuming MCMC/ABC algorithms might be infeasible and a nonparametric approach is desirable to avoid parametric assumptions. We define a reinforced stochastic model for the analysis of the $M/G/1$ queue through a system of predictive distributions. Using the theory of partial exchangeable processes developed in (Diaconis and Freedman, 1980; Epifani et al., 2002; Muliere et al., 2003), we prove that the reinforced stochastic process is in fact the predictive probability model of a Bayesian semi-Markov mixture model. It is shown that the nonparametric prior for the random semi-Markov kernel is a beta-Stacy Neutral-to-the-Right process (Walker and Muliere, 1997; Muliere et al., 2003). This result enabled Bayesian nonparametric prediction for the $M/G/1$ queue based on the reinforced stochastic process. The proposed approach bypasses time-consuming MCMC/ABC algorithms, which

makes the method scalable to large datasets. We provide numerical illustrations of the proposed method in an extensive simulation study and an open-source R package which implements the model.

The paper is structured as follows. Section 2 gives a brief summary of the structure of an $M/G/1$ queue, which will be required in Section 3, where we introduce a reinforced probability model for the analysis of the $M/G/1$ queue. In section 4 we show that the reinforced process is a semi-Markov mixture model and can therefore be utilized for Bayesian predictive inference for the $M/G/1$ queue as described in Section 5. Section 6 gives a numerical illustration of the proposed approach in a systematic simulation study for several $M/G/1$ queues. We then conclude the article in Section 7.

2   The M/G/1 queue

We consider a one-counter queueing system where items arrive according to a homogeneous Poisson process with arrival rate λ, where items are processed according to a "first-in-first-out" rule according to some probability G on $[0, \infty)$. The random function $X(t), t \geq 0$ denotes the number of items in the system at time $t \geq 0$ and can be sufficiently summarized through the jump times and jump chain, i.e. the time points at which an item enters or departures from the queue and the number of items in the system right after this jump time.

Except for the special case where $G = \text{Exp}(\mu)$ is the exponential distibution, the jump chain and holding times are neither Markov nor semi-Markov. Hence for a general service time distribution G the $M/G/1$ queue is non-

trivial to study and one usually defines an augmented or imbedded process which is simpler to study (Medhi, 2003). We will utilize the imbedded process $Y = (Y(t), t \geq 0)$ which is equal to X at the departure times of the original queueing system $Y(t) = X(S_n+)$ if $S_n \leq t < S_{n+1}$, where $S_n > 0$ denotes the departure time of the n-th item. By definition, Y satisfies the relation

$$Y(S_n) = \min\{Y(S_{n-1}) - 1, 0\} + A_n \ \ \text{for } n \geq 1, \tag{1}$$

where A_n denotes the number of arriving items during the n-th service time (Ross, 1992, p. 89). Since the arrival process of an $M/G/1$ queue is Markov, $\{A_n\}_n$ is independent and identicaly distributed with discrete probability $Q = (Q_k)_{k \geq 0}$ where $Q_k = \int_0^\infty (\lambda s)^k e^{-\lambda s} G(ds)/k!$. Hence the process Y is semi-Markov with embedded Markov jump chain $\{Y(S_n)\}_{n \geq 0}$, with transition matrix Π equal to $\Pi_{k,j} = Q_{j-k-1+\delta_0(k)} I(j \geq k)$, and holding time distribution $P(d(S_{n+1} - S_n))$ equal to G if $Y_n > 0$ and $G \star Exp(\lambda)$ if $Y_n = 0$.

3 The row-wise exchangeable predictive model

In this section we introduce a reinforced probability model for the imbedded semi-Markov process of the $M/G/1$ queue. Specifically we will define a probability model through a system of predictive distributions in such a way that the predicted probability of the next event will be a function of the number of times such an event occurred during the compete history

of the process. This enables sequential reinforced learning on the future behavior of the $M/G/1$ process. In section 4, it will be shown that the probability model corresponds to a Bayesian semi-Markov mixture process with a nonparametric prior, and hence the system of predictive distributions can be interpreted as a nonparametric Bayesian prediction model for the imbedded departure process of the $M/G/1$ queue.

We start by introducing three sequences of random variables which will be used for this purpose. The first sequence will be used to model the discrete increments of the queue at departure times. Let A_1 be a discrete random variable on $\mathbb{N}_0$ with probability mass function

$$\mathbb{P}(A_1 = k) = \frac{s_k}{m_k} \prod_{j=0}^{k-1} \left(1 - \frac{s_j}{m_j} \right) \quad \text{for } k \geq 0,$$

where s and m are non-negative sequences such that (i) $s_j < m_j$, (ii) $0 < \inf_j\{s_j\}$ and (iii) $\sup_j\{m_j\} < +\infty$. Moreover, for $n \geq 1$ and given $\mathcal{A}_n = \sigma(\{A_j\}_{1 \leq j \leq n})$, let A_{n+1} be distributed as

$$\mathbb{P}(A_{n+1} = k | \mathcal{A}_n) = \frac{s_k + v_n(k)}{m_k + w_n(k)} \prod_{j=0}^{k-1} \left(1 - \frac{s_j + v_n(j)}{m_j + w_n(j)} \right),$$

where $v_n(k) = \sum_{i=1}^n I(A_i = k)$ and $w_n(k) = \sum_{i=1}^n I(A_i \geq k)$ denote the counting and risk processes for the first n elements of $(A_j)_{j \geq 1}$, respecttively.

Secondly, we define two sequences of non-negative real variables, which will be used to model the departure times of the queue. For this propose let $\alpha^{(i)}$ with $i = 1, 2$ be two measures on the positive real line $(\mathbb{R}_+, \mathbb{B}(\mathbb{R}_+))$, and let $\beta^{(i)} : [0, \infty) \to (0, \infty)$ be a measurable step function. For ease

of exposure we decompose the measure $\alpha^{(i)}$ into a discrete part $\alpha_d^{(i)}$ with discontinuity set $\mathcal{J}^{(i)} = \{t_j\}_{j\geq 1}$ and an absolutely continuous part $\alpha_c^{(i)}$, i.e. $\alpha^{(i)}[0,t] = \alpha_c^{(i)}(0,t] + \sum_{j:t_j\in\mathcal{J}^{(i)}, t_i\leq t} \alpha_d^{(i)}\{t_i\}$. We assume that $0 \notin \mathcal{J}^{(i)}$ and $(\alpha^{(i)}, \beta^{(i)})$ satisfy

$$\prod_{t_j\in\mathcal{J}^{(i)}} \left(1 - \frac{\alpha_d^{(i)}\{t_j\}}{\alpha_d^{(i)}\{t_j\} + \beta^{(i)}(t_j)}\right) \exp\left\{-\int_{[0,\infty)} \frac{\alpha_c^{(i)}(ds)}{\beta^{(i)}(s)}\right\} = 0. \qquad (2)$$

Now, for $i = 1, 2$ let $S_1^{(i)}$ be a non-negative random variable with complementary distribution function (ccdf)

$$\mathbb{P}\left(S_1^{(i)} > t\right) = \prod_{j:t_j\in\mathcal{J}^{(i)}, t_j\leq t} \left(1 - \frac{\alpha_d^{(i)}\{t_j\}}{\beta^{(i)}(t_j) + \alpha_d^{(i)}\{t_j\}}\right) \exp\left\{-\int_{(0,t]} \frac{\alpha_c^{(i)}(ds)}{\beta^{(i)}(s)}\right\}.$$

Moreover, for $n > 1$ and given the sigma-field $\mathcal{S}_n^{(i)} = \sigma\left(\{S_j^{(i)}\}_{1\leq j\leq n}\right)$, let $S_{n+1}^{(i)}$ be distributed as

$$\mathbb{P}\left(S_{n+1}^{(i)} > t | \mathcal{S}_n^{(i)}\right) = \prod_{j:t_j\in\mathcal{J}_n^{(i)}, t_j\leq t} \left(1 - \frac{\alpha_{n,d}^{(i)}\{t_j\}}{\beta_n^{(i)}(t_j) + \alpha_{d,n}^{(i)}\{t_j\}}\right) \exp\left\{-\int_{(0,t]} \frac{\alpha_c^{(i)}(ds)}{\beta_n^{(i)}(s)}\right\},$$

where the discrete measure $\alpha_{n,d}\{t\} = \alpha_d^{(i)}\{t\} + \sum_{i\leq n} I(S_j^{(i)} = t)$ and the step function $\beta_n^{(i)}(t) = \beta^{(i)}(t) + \sum_{i\leq n} I(S_j^{(i)} > t)$ are updated by the counting and risk processes for the first n random variables. Assumption (2) and $0 \notin \mathcal{J}^{(i)}$ imply that the system of predictive distributions is proper without positive mass at zero or infinity, which will ensure that the queueing system has a.s. non-zero and finite holding-times.

Finally, similarly to the departure process of the $M/G/1$ queue, we define

the jump process $Y(t) = \sum_{n \geq 0} Y_n I(S_n \leq t \leq S_{n+1})$ for $t \geq 0$ as

$$(Y_{n+1}, S_{n+1}) = \Big(\max(Y_n - 1, 0) + A_{n+1},$$
$$S_n + I(Y_n = 0) S^{(1)}_{h(n)} + I(Y_n > 0) S^{(2)}_{n+1-h(n)} \Big) \text{ for } n \geq 1, \quad (3)$$

where $h(n) = \sum_{j=0}^n I(Y_j = 0)$ denotes the number of times the system was idle until time S_n and we set $(S_0, Y_0) = (0, a_0)$ for $a_0 \in \mathbb{N}_0$. The interpretation of (3) is similar to the imbedded semi-Markov process of a $M/G/1$ queue. Differentiating between non-idle and idle holding times is not necessary but has some advantages for statistical prediction as demonstrated below.

4 A Bayesian mixture representation

In this section we show that the process defined in (3) can be represented as a Bayesian mixture model. We will first establish an invariance property for the jump chain $\{Y_n\}$ and the sequence of states and jump times $\{(Y_n, S_n)\}$, which will be required for the mixture representation. Since the jump chain has a restricted set of transitions, we call a sequence $(j_k)_{k=0}^n$ of non-negative integers admissible if $j_k \geq j_{k-1} - I(j_{k-1} > 0)$ for all $k \leq n$. Moreover, following (Diaconis and Freedman, 1980), we call two integer sequences $\mathbf{i} = (i_k)_0^n$ and $\mathbf{j} = (j_k)_0^m$ equivalent if $i_0 = j_0$ and for every pair of non-negative integers s, v the number of transitions from s to v among states in $\mathbf{i}$ equals the same number of transitions in $\mathbf{j}$. The following lemma establishes the finite dimensional law of the jump chain and an invariance property.

Lemma 1. *Lemma 1 (i) For every $n \geq 1$ and every sequence of states*
$\boldsymbol{j} = (j_k)_{k=0}^n$

$$\mathbb{P}(\cap_{k=0}^n \{Y_k = j_k\}) \prod_{i \geq 0} \frac{s_i^{[v_n(i)]} (m_i - s_i)^{[u_n(i)]}}{m_l^{[w_n(i)]}}$$

if $\boldsymbol{j}$ is admissible and 0 otherwise, where $\prod_{i=0}^{-1} := 1$, $a^{[n]} = a(a+1)\cdots(a+n-1)$ and $u_n(i) = w_n(i) - v_n(i)$.

(ii) The jump chain $\{Y_n\}$ is partial exchangeable according to (Diaconis and Freedman, 1980), i.e. for every two equivalent integer sequences $\boldsymbol{i}$ and $\boldsymbol{j}$ of length $n+1 > 0$

$$\mathbb{P}(\cap_{k=0}^n \{Y_k = i_k\}) = \mathbb{P}(\cap_{k=0}^n \{Y_k = j_k\}).$$

(iii) For every fixed $i \in \mathbb{N}_0$, $\mathbb{P}(\liminf_n Y_n = i | Y_0 = i) = 1$.

The proof of the lemma and all remaining results are given in the appendix.

The partial exchangeability property (ii) can be established from the finite dimensional probability law as stated in (i). Moreover, part (ii) and (iii) of the preceding lemma facilitates a direct application of Diaconis and Freedman representation theorem (Diaconis and Freedman, 1980), which states that every recurrent partial exchangeable process is a mixture of Markov chains. It follows that the sequence $\{Y_n\}$ is a Markov mixture model given some random transition matrix. The exact prior measure for the transition matrix will be determined below.

We now turn to the joint model of jump chain and holding times (3) and

establish a joint mixture representation. For this purpose we define for each state $i \in \mathbb{N}_0$ the sequence of hitting times to i by $\tau_i(1) = \inf\{m \geq 0 : Y_m = i\}$ and $\tau_i(n+1) = \inf\{m > \tau_i(n) : Y_m = i\}$ for $n > 1$. Furthermore, following (Epifani et al., 2002), we introduce the sequence of successor states of state i as $V_i(n) = Y_{\tau_i(n)+1}$ if $\tau_i(n) < \infty$ and ∞ otherwise. Similarly we define the sequence of holding times to state i as $T_i(n) = S_{\tau_i(n)+1} - S_{\tau_i(n)}$ if $\tau_{i(n)} < \infty$ and $+\infty$ otherwise. The following lemma extends Lemma1 (ii) and shows that the joint process $\{(Y_n, S_n)\}_n$ is row-wise partial exchangeable.

Lemma 2. *Lemma 2 Define the matrix (V, T) of successor states and holding times as $(V, T) = \{(V_i(n), T_i(n))\}_{i \geq 0, n \geq 1}$. Then (V, T) is row-wise partial exchangeable according to (Epifani et al., 2002), i.e. for all $N \geq 0$ and $n > 1$*

$$
\mathbb{P}\left[\bigcap_{i=0}^{N}\bigcap_{k=1}^{n}\left\{V_i(k) = j_{i,k}, T_i(k) \leq t_{i,k}\right\}\right] = \mathbb{P}\left[\bigcap_{i=0}^{N}\bigcap_{k=1}^{n}\left\{V_i(k) = j_{i,\sigma_i(k)}, T_i(k) \leq t_{i,\sigma_i(k)}\right\}\right]
$$

for any permutation σ_i of $\{1, \cdots, n\}$, for $1 \leq i \leq N$.

Using lemma 1 and 2, we can now use the theory of row-wise partial exchangeable processes (Epifani et al., 2002; Muliere et al., 2003) and show that the process defined in (3) is a Bayesian mixture model. Recall that a continuous time jump process is semi-Markov if the jump chain and holding time are jointly Markov on $\mathcal{S} = \mathbb{N}_0 \times [0, \infty)$. Also recall that a semi-Markov kernel $\mathbf{W} = (W_i, i \in \mathbb{N}_0)$ is a sequence of probability measures on $\mathcal{S}$ which is an elements of $\mathbb{W}$, where $\mathbb{W}$ denotes the space of all sequences of probability measures on $\mathcal{S}$.

Corollary 1. *Corollary 1(i) There exists a $\mathbb{W}$-valued random element $\boldsymbol{W}$ such that for every admissible sequence $(i_k)_{0 \leq k \leq n}$ of states and positive real numbers $(t_k)_{1 \leq k \leq n}$, $n \geq 1$*

$$\mathbb{P}\left[\bigcap_{k=0}^{n}\left\{Y_k = i_k, S_k - S_{k-1} \leq t_k\right\}\Big| X_0 = i_0\right] = \mathbb{E}\left[\prod_{1 \leq k \leq n} W_{i_{k-1}}(i_k, [0, t_k])\right].$$

(ii) The random semi-Markov kernel $\boldsymbol{W} = \{W_i(\cdot, \cdot)\}_{i \geq 0}$ has the form

$$W_i(j, [0, t]) = \begin{cases} \Pi_{0,j} G^{(1)}(t) & \text{if } i = 0, t \geq 0 \\ \\ \Pi_{i,j} G^{(2)}(t) & \text{if } i \geq 1, t \geq 0; \end{cases} \tag{4}$$

with random transition matrix $\Pi_{i,j} = Q_{j-i+1-\delta_0(i)} I(i \geq 0, j \geq i)$ where $Q_j = \theta_j \prod_{k<j}(1 - \theta_k)$ and $\theta_j \overset{ind.}{\sim} Beta(s_j, m_k - s_j)$. Moreover, the random distribution function $G^{(i)}$, $i = 1, 2$, is a beta-Stacy Neutral-to-the-Right process on $\mathbb{R}_+$ with parameters $(\beta^{(i)}, \alpha^{(i)})$ and with Levy measure

$$v(ds, dt) = \frac{\exp\{-s\beta^{(i)}(t)\}}{1 - \exp\{-s\}} ds \alpha_c^{(i)}(dt). \tag{5}$$

The corollary states that, given the random semi-Markov kernel $\mathbf{W}$, the reinforced process Y defined in (3) is semi-Markov and behaves like the $M/G/1$ queue at the departure epochs. The nonparametric prior for $\mathbf{W}$ is a product of a random matrix and two Neutral-to-the-Right processes (Doksum, 1974; Walker and Muliere, 1997) where $G^{(2)}$ is the nonparametric prior for the service time distribution G.

Remark 1. *Remark 1 It is not difficult to see that, given an observed sample*

of N jump times and states $\{(Y_n, S_n)\}_{n \leq N}$, the process $\{(Y_n, S_n)\}_{n \geq N}$ is again a mixture of semi-Markov chains with the same nonparametric prior, where the prior parameters $v, w, \alpha^{(i)}, \beta^{(i)}, i = 1, 2$ are replaced by $s_N = s + v_N, m_M = m + w_N, \alpha_N^{(i)}, \beta_N^{(i)}$, with $\alpha_{N,c}^{(i)} = \alpha_c^{(i)}$, $\alpha_{N,d}^{(i)} = \alpha_d^{(i)} + N^{(i)}(\cdot)$ and $\beta_N^{(i)} = \beta^{(i)} + R^{(i)}(\cdot)$; where the counting and risk processes equal to

$$N^{(1)}(x) = \sum_{1 \leq n \leq N} I(S_n - S_{n-1} = x, Y_{n-1} = 0),$$

$$R^{(1)}(x) = \sum_{1 \leq n \leq N} I(S_n - S_{n-1} > x, Y_{n-1} = 0),$$

$$N^{(2)}(x) = \sum_{1 \leq n \leq N} I(S_n - S_{n-1} = x, Y_{n-1} > 0),$$

$$R^{(2)}(x) = \sum_{1 \leq n \leq N} I(S_n - S_{n-1} > x, Y_{n-1} > 0) \quad \text{for } x \geq 0.$$

In the next section we use the reinforced stochastic process (3) for fast and scalable nonparametric Bayesian prediction of main quantities of the $M/G/1$ queue.

5   Bayesian nonparametric prediction

In this section we use the semi-Markov mixture process (3) for nonparametric Bayesian prediction for the $M/G/1$ queue. The R package, NPBMG1, available in the supplementary material, implements the proposed method. Assume data for an $M/G/1$ process are observed until the N-th departure time and summarized as $(Y, S)_N := \{(Y_n, S_n)\}_{0 \leq n \leq N}$. This data reduction is always feasible even if the original data are not collected according to this format. Based on the current knowledge one wants to predict the main

characteristics of the system in transient and steady state without knowing the exact service time distribution and the arrival rate. Furthermore, as new observations arrive we want to update our estimates in real time. We model the queue through the process (3) and its predictive distribution as described below. We first consider the system in transient state and turn thereafter to the steady state equilibrium.

Number of items in the queue at departure epochs

Predicting the number of items at the next departure epoch can be done directly using the predictive distribution of the increments A. In particular, the transition probability from $i \geq 0$ items to $j \geq i - 1 + \delta_0(i)$ items, given the current information $\mathcal{F}_N = \sigma((Y, S)_N)$, is given by

$$\mathbb{E}[\Pi_{i,j}|\mathcal{F}_N] = \mathbb{P}(Y_{N+1} = j|Y_N = i, \mathcal{F}_n)$$
$$= \frac{s_{j-i+1-\delta_0(i)} + v_N(j - i + 1 - \delta_0(i))}{m_{j-i+1-\delta_0(i)} + w_N(j - i + 1 - \delta_0(i))} \prod_{l=0}^{j-i-\delta_0(i)} \left(1 - \frac{s_l + v_N(l)}{m_l + w_N(l)}\right)$$

if $j \geq i - 1 + \delta_0(i) \geq 0$ and 0 otherwise. Similarly the tail probability of more than $j \geq i - 1 + \delta_0(i)$ items and the expected number of items at the next departure time is given by

$$\mathbb{P}(Y_{N+1} > j|Y_N = i, \mathcal{F}_n) = \prod_{l=0}^{j-i+1-\delta_0(i)} \left(1 - \frac{s_l + v_N(l)}{m_l + w_N(l)}\right),$$
$$\mathbb{E}[Y_{N+1}|\mathcal{F}_N] = Y_N + 1 - \delta_0(Y_N) + \sum_{j \geq 0} \prod_{l=0}^{j} \left(1 - \frac{s_l + v_N(l)}{m_l + w_N(l)}\right).$$

13

Mean service time

The distribution function of the service-times given the current information can be estimated from the predictive distribution of the holding times

$$
\begin{aligned}
\mathbb{E}[G(t)|\mathcal{F}_N] &= \mathbb{P}\Big[S^{(2)}_{N+1-h(N)} \le t\big|\mathcal{F}_N\Big] \\
&= 1 - \prod_{j:t\ge t_j\in\mathcal{J}^{(2)}} \left(1 - \frac{\alpha^{(2)}_{N,d}\{t_j\}}{\beta^{(2)}_N(t_j)+\alpha^{(2)}_{N,d}\{t_j\}}\right) \exp\left\{-\int_{(0,t]}\frac{\alpha^{(2)}_c(dx)}{\beta^{(2)}_N(x)}\right\}.
\end{aligned}
$$

The evaluation of the integral in the last expression across t is relatively time consuming. In the actual numerical implementation, described in the next section, we sample repeatedly from the posterior $\mathbb{P}(dG^{(2)}|\mathcal{F}_N)$. In this way we can provide highest-posterior-density (HPD) credibility sets for the service-time cdf $G^{(2)}$ and the mean service time $\mathbb{E}[S^{(2)}|G^{(2)}]$. We developed a simple and fast algorithm to sample from the posterior $\mathbb{P}(dG^{(2)}|\mathcal{F}_N)$. The detailed steps of the algorithm are described in the supplementary material.

The Traffic intensity and the steady state equilibrium

The traffic intensity, defined as $\rho = \mathbb{E}[A_n; Q]$ if probability mass function Q of A_n is known, is the single most important index of the queue and determines whether the system converges to a steady state equilibrium. The queue will converge to a steady state limit if and only if $\rho < 1$ (Medhi, 2003). The Bayesian point predictor for ρ equals

$$
E[\rho|\mathcal{F}_N] = E[A_N|\mathcal{F}_N] = \sum_{j\ge 0}\prod_{l=0}^{j}\left(1 - \frac{s_l + v_N(l)}{m_l + w_N(l)}\right),
$$

14

Moreover, we introduce a Bayesian hypothesis test for convergence to a steady-state equilibrium by specifying a weighted 0-1 loss function with cost w_1 and w_0 for choosing falsely the model $M_1 = \{\rho < 1\}$ over the model $M_0 = \{\rho \geq 1\}$ and vice versa. The optimal Bayesian decision then selects model M_1 over M_0 if $\mathbb{P}[\rho < 1|\mathcal{F}_N] > w_1/(w_1+w_0)$. To compute the posterior probability for model M_1, we apply the mixture representation in Corollary 1 and resort to Monte-Carlo based inference as described in Algorithm 1.

Algorithm 1 Monte-Carlo based steady state inference

1: **procedure** INFERENCE FOR THE TRAFFIC INTENSITY ρ
2:     **for** c=1:C **do**
3:         sample $\theta_j^{(c)} \sim \text{Beta}(s_j + v_N(j), m_j + w_N(j) - v_N(j))$ for $j \geq 0$
4:         set $\rho_c := \sum_{j \geq 0} \prod_{l=1}^{j}(1 - \theta_l)$ and $I_c = I(\rho_c < 1)$
5:     **end for**
6:     compute $J_C = \sum_{c=1}^{C} I_c$ and set $\mathbb{P}[\rho < 1|\mathcal{F}_N] \approx J_C/C$
7:     set $(L_\alpha, U_\alpha) = \arg\min_{k=1,...,C-\lfloor(1-\alpha)C\rfloor}[\rho_{(k)}, \rho_{(k+\lfloor(1-\alpha)C\rfloor)}]$
8: **end procedure**

Steady state prediction

If, given the current information $\mathcal{F}_N$, the Bayesian decision is to choose the ergodic model M_1, we want to predict the queue at the steady state equilibrium. For this purpose, we have to restrict the posterior and the predictive distribution to the model M_1. We use the fact that the invariant distribution $\psi = (\psi_j)_{j \geq 0}$ of the $M/G/1$ system, i.e. the distribution of the number of items in the system at steady state, is identical to the invariant distribution of the imbedded jump chain at the departure times $\{Y_n\}$ (Medhi, 2003), where ψ solves $\psi\Pi = \psi$. As before $\Pi_{i,j} = Q_{j-i+1-\delta_0(j)}I(j \geq i, i > 0)$, where Q under the posterior probability equals $Q_k = \theta_k \prod_{l<k}(1 - \theta_l)$ and θ_l

has a beta distribution as described in Algorithm 1. After some algebraic manipulations we can express the invariant distribution as

$$\psi_0 = 1 - \rho \tag{6}$$

$$\psi_k = \frac{\psi_0 S_{k-1} + \sum_{i=1}^{k-1} \psi_i S(k-i)}{Q_0} \quad \text{for} \quad k \geq 1, \tag{7}$$

where $S(k) = \sum_{j \geq k+1} Q_j = \prod_{j=1}^{k}(1 - \theta_j)$. We utilize the recursive relation (6) and (7), and build upon Algorithm 1, to predict the invariant distribution and the expected number of items in the system at equilibrium. The main computational steps are summarized in Algorithm 2.

Lastly, for an $M/G/1$ queue such that $\rho < 1$, the mean length of the busy period in equilibrium, is given by $\mu(Q, G) = \mathbb{E}[S_n^{(2)}|G]/(1 - \mathbb{E}[A_n|Q])$. The predicted busy period given the current data $\mathcal{F}_N$ is obtained by using the Monte-Carlo computations of Algorithm 2.

6   Numerical Illustration

In this section we give a numerical example of the procedure outlined in the previous section. In particular we consider four different queueing models summarized in Table 1.

For each system we simulated 1,000 queueing processes of length 10^i, $i = 3, 4, 5$ time units. The procedure was implemented in R without using parallel computing, which would be straight forward as all Monte-Carlo computing steps are non-iterative. Table 2 summarizes the average predictive performance of the proposed method over the 1,000 queueing processes

Algorithm 2 Monte-Carlo based steady-state inference (continued)

1: **procedure** INFERENCE FOR THE INVARIANT DISTRIBUTION ψ
2:     **for** c=1:C such that $I_c = 1$ **do**
3:         compute $S_k^{(c)} = \prod_{j \leq k}(1 - \theta_k^{(c)})$ for $k \geq 0$
4:         set $\psi_0^{(c)} = 1 - \rho_c$ and
5:         compute recursively $\psi_k^{(c)}$ for $k > 0$ according to (7)
6:     **end for**
7:     compute $\widehat{\psi}_k = \sum_{c:I_c=1} \psi_k^{(c)}/J_C$ for $k \geq 0$
8:     set $\lim_{t \to +\infty} \mathbb{E}[X(t)|\mathcal{F}_N, M_1] \approx \sum_{k \geq} k\widehat{\psi}_k$
9: **end procedure**

1: **procedure** INFERENCE FOR THE MEAN LENGTH OF THE BUSY PERIOD
2:     **for** c=1:C such that $I_c = 1$ **do**
3:         generate $\overline{G}_c(\cdot) = 1 - G_c(\cdot)$ from $\mathbb{P}[G^{(2)}|\mathcal{F}_N]$
4:         compute the mean service time $\mu_c = \int d\overline{G}_c$
5:     **end for**
6:     Compute $\mathbb{E}[\mu(Q,G)|\mathcal{F}_N] \approx J_C^{-1} \sum_{c:I_c=1} \mu_c/(1 - \rho_c)$
7: **end procedure**

---

for each of the four queueing models and the 3 observational periods. For all quantities considered posterior estimates quickly concentrate around the true values. Except for the mean number of items in equilibrium the 90 percent highest density credibility intervals have frequentist coverage close to the nominal values. The coverage of the credibility interval for the mean number of items in equilibrium is higher than the nominal value making prediction more conservative.

7 Conclusions

Queueing models are a class of stochastic processes used in many applied problems. In this paper we introduced a fast and scalable approach to the Bayesian analysis of the $M/G/1$ queuing systems. The proposed method

17

Model	Arrival rate	service-time distribution	traffic intensity
1	.3	$\text{Exp}(.4)$	.75
2	.3	$\text{Uni}(1,4)$	.75
3	.3	$\text{Gam}(5,2)$	.75
4	.3	$.3\text{Gam}(.2,1)+.2\text{Gam}(.4,.1)+.4\text{Gam}(.4,3)+.1\text{Gam}(5,.25)$	.75

Table 1: True queueing models used in the simulation study

uses a predictive nonparametric approach utilizing the theory of reinforced stochastic process and row-wise partial exchangeable process developed in (De Finetti, 1980; Epifani et al., 2002; Muliere et al., 2003). As a consequence inference can be done using either closed form expressions or simple Monte-Carlo schemes and avoids therefore time consuming MCMC methods or posterior approximations like ABC methods. Future work will be directed towards the $G/G/1$ queue using an augmentation approach where the queueing system is embedded into an augmentation semi-Markov process.

References

D. Aldous. *Exchangeability and Related Topics.* Springer, 1985.

C. Armero. *Bayesian Statistics 2*, chapter Bayesian Analysis of $M/M/1/\infty/FIFO$ queues, pages 613–617. Elsevier Science Publishing Company, 1985.

C. Armero and M. Bayarri. Prior assessments for prediction in queues. *The Statistician*, 43:139–153, 1994a.

C. Armero and M. Bayarri. *Bayesian Statistics 5*, chapter Bayesian Questions and Answers in Queues, pages 3–23. Oxford University Press, 1996.

C. Armero and M. J. Bayarri. Bayesian prediction in M/M/1 queues. *Queueing Systems*, 15:401–417, 1994b.

C. Armero and D. Conesa. Prediction in markovian bulk arrival queues. *Queueing Systems*, 34:327 – 350, 2000.

C. Armero and D. Conesa. Bayesian hierarchical models in manufacturing bulk service queues. *Journal of Statistical Planning and Inference*, 136:335 – 354, 2006.

M. Ausin, M. Wiper, and R. Lillo. Bayesian estimation for the $M/G/1$ queue using a phase-type approximation. *Journal of Statistical Planning and Inference*, 118: 83–101, 2004.

M. Ausin, M. Wiper, and R. Lillo. Bayesian prediction of the transient behaviour and busy period in short- and long-tailed $GI/G/1$ queueing systems. *Computational Statistics & Data Analysis*, 52:1615–1635, 2008.

Michael GB Blum and Olivier François. Non-linear regression models for approximate bayesian computation. *Statistics and Computing*, 20(1):63–73, 2010.

R. Butler and A. Huzurbazar. Bayesian prediction for waiting times in stochastic models. *Canadian Journal of Statistics*, 28:311–325, 2000.

Pier Luigi Conti. Large sample bayesian analysis for Geo/G/1 discrete-time queueing models. *Annals of Statistics*, 27:1785–1807, 1999.

Pier Luigi Conti. Bootstrap approximations for bayesian analysis of $Geo/G/1$ discrete-time queueing models. *Journal of Statistical Planning and Inference*, 120:65–84, 2004.

B. De Finetti. *Studies in Inductive Logic and Probability*, chapter On the Condition of Partial Exchangeability, pages 193–205. University of California Press, 1980.

Peris Diaconis and David A. Freedman. De Finetti's theorem for Markov chains. *The Annals of Probability*, 8:115–130, 1980.

Kjell Doksum. Tailfree and neutral random probabilities and their posterior distributions. *The Annals of Probability*, 2:183–201, 1974.

I. Epifani, S. Fortini, and L. Ladellic. A characterization for mixtures of semi-markov processes. *Statistics and Probability Letters*, 60:445–457, 2002.

T.S. Ferguson and M.J. Klass. A representation of independent increment processes without gaussian components. *Annals of Mathematical Statistics*, 42:1634 – 1643, 1972.

S. Fortin, L. Ladelli, and E. Regazzini. Exchangeability, predictive distributions and parametric models. *Sankhya*, 62:86 – 109, 2000.

S. Fortini, L. Ladelli, G. Petris, and E. Regazzini. On mixtures of distributions of markov chains. *Stochastic Processes and their Applications*, 100:147–165, 2002.

N.L. Hjort. Nonparametric bayes estimators based on beta processes in models for life history data. *The Annals of Statistics*, 18:1259 – 1294, 1990.

D.R. Insua, M. Wiper, and F. Ruggeri. Bayesian analysis of $M/E_r/1$ and $M/H_k/1$ queues. *Queueing Systems*, 30:289–308, 1998.

Sudha Jain and J. G. C Templeton. Statistical inference for $G/M/1$ queueing system. *Oper. Res. Lett.*, 7(6):309–313, December 1988. ISSN 0167-6377.

O. Kallenberg. Characterizations and embedding properties in exchangeability. *Probability theory and Related Fields*, 60:249–281, 1982.

J. Lee. Sampling methods of neutral to the right processes. *Journal of Computational and Graphical Statistics*, 16:656–671, 2007.

G. Link. *Studies in Inductive Logic and Probability*, chapter Representation Theorems of the de Finetti Type for Partial Symmetric Probability Measures, pages 207–231. University of California Press, 1980.

Michael F McGrath and Nozer D Singpurwalla. A subjective bayesian approach to the theory of queues II : Inference and information in $M/M/1$ queues. *Queueing Systems*, 1(4):335–353, 1987.

J. Medhi. *Stochastic Models in Queuing Theory*. Academic Press, 2003.

P Muliere, P Secchi, and SG Walker. Urn schemes and reinforced random walks. *Stochastic processes and their applications*, 88(1):59–78, 2000.

P. Muliere, P. Secchi, and S.G. Walker. Reinforced random processes in continuous time. *Stochastic Processes and their Applications*, 104:117–130, 2003.

Susan M Pitts. Nonparametric estimation of the stationary waiting time distribution function for the $gi/g/1$ queue. *The Annals of Statistics*, 22:1428–1446, 1994.

Sheldon M. Ross. *Applied Probability Models With Optimization Applications*. Dover Publications, 1992.

Alexander Y Shestopaloff and Radford M Neal. On bayesian inference for the $M/G/1$ queue with efficient mcmc sampling. *arXiv preprint arXiv:1401.5548*, 2014.

C. Sutton and M.I. Jordan. Bayesian inference for queueing networks and modeling of internet services. *The Annals of Applied Statistics*, 254 – 282:5, 2011.

JL Teugels and A Vanmarcke. Estimating the stationary distribution in a $GI/M/1$ queue. *Stochastic processes and their applications*, 54(1):113–119, 1994.

S. Walker and P. Damien. Representations of lévy processes without gaussian components. *Biometrika*, 87:477 – 483, 2000.

Stephen Walker and Pietro Muliere. Beta-stacy processes and a generalization of the polya-urn scheme. *Annals of Statistics*, 25:1762 – 1780, 1997.

D. Williams. *Probability with Martingales*. Cambridge University Press, 1991.

M. Wiper, D. R. Insua, and F. Ruggeri. Mixture of gamma distributions with application. *Journal of Computational and Graphical Statistics*, 10:440–454, 2001.

M. P. Wiper. Bayesian analysis of $Er/M/1$ and $Er/M/c$ queues. *Journal of Statistical Planning and Inference*, 69(1):65–79, June 1998.

A Simulation of Beta-Stacy processes

The Beta-Stacy Neutral-to-the-Right random distribution function $G \sim BS(\alpha, \beta)$ with base measure $\alpha = \alpha_d + \alpha_c$ and precision function β is a stochastic process define as $1 - G(\cdot) \stackrel{d}{=} \exp\{-Z_c(\cdot) - Z_d(\cdot)\}$ where $Z_c(\cdot) = \sum_j Z_{c,j} \delta_{T_j}$ is an independent increment process with random jump locations on $[0, \infty)$ and Levy measure

$$\nu(ds, dt) = k(s|t)ds\alpha_c(dt) = \frac{\exp\{-s\beta(t)\}ds}{1 - \exp\{-s\}}\alpha_c(dt), \tag{8}$$

and $Z_d = \sum_{t_j \in \mathcal{I}} Z_{d,j} \delta_{t_j}$ is an independent increment process jumps at deterministic locations $\mathcal{I} = \{t \geq 0 : \alpha_d\{t\} > 0\}$ of size $W_j = 1 - e^{-Z_{d,j}} \sim$

Beta($\alpha_d\{t_j\}, \beta(t_j)$) (Ferguson and Klass, 1972; Walker and Muliere, 1997; Walker and Damien., 2000; Lee, 2007). Moreover, from (Doksum, 1974; Walker and Muliere, 1997), if $\mathbf{X} = \{X_i\}_{1 \leq i \leq n}$ is a random sample from G, then the posterior distribution of G, given $\mathbf{X}$ is again a Beta-Stacy process with parameter (α_n, β_n) where $\alpha_n = \alpha_c + [\alpha_d + N]$ and $\beta_n = \beta + Y$, for $N([0,t]) = \sum_{i \leq n} I(X_i \leq t)$ and $Y(t) = \sum_{i \leq n} I(X_i > t)$ for $t \geq 0$.

To simulate from $BS(\alpha_n, \beta_n)$ we use the fact that if we define $H = \sum_j (1 - e^{-Z_{c,j}})\delta_{T_j}$ then H is an independent increment process with Levy measure

$$v_H(ds, dt) = \frac{k(-\log(1-s)|t)}{1-s}\alpha_c(dt) = \beta(t)s^{-1}(1-s)^{\beta(t)-1}\frac{\alpha_c(dt)}{\beta(t)}, \quad (9)$$

which is the Levy measure of a Beta process (Hjort, 1990) with precision parameter β and base measure α_c/β, say $H \sim \mathrm{BP}(\beta, \alpha_c/\beta)$. We can, for example use the ϵ-truncations method of (Lee, 2007) to simulate a Beta process $H = \sum_j H_j \delta_{T_j}$ first and then used the transformation $Z_c = \sum_j -\log(1 - H_j)\delta_{T_j}$ to obtain a realization from a Beta-Stacy process. The detailed steps are summarized in algorithm 3.

Algorithm 3 ϵ-truncations method for a Beta-Stacy process

1: generate $M \sim \mathrm{Poisson}\big(\alpha_c(0,T]/\epsilon\big)$

2: generate $T_j \overset{iid}{\sim} I(T_i \in (0,T])\alpha_c(\cdot)/\alpha_c(0,T]$ for $j = 1, \cdots, M$

3: generate $H_j|T_{(j)} \sim \mathrm{Beta}(\epsilon, \beta(T_{(j)}))$ for $j = 1, \cdots, M$

4: generate $W_j \sim \mathrm{Beta}(\alpha_d\{t_j\}, \beta(t_j))$ for $t_j \in \mathcal{J}$

5: set $F_\epsilon(t) = 1 - \prod_{j:t_j<t}(1 - W_j) \times \prod_{k:T_k \leq t}(1 - H_k)$ for $t \in [0,T]$

B Proofs

Proof. Proof of lemma 1 (i) Let $(j_k)_{k=0}^n$ be an admissible integer sequence and define $s_n(j) = s_j + v_n(j)$ and $m_n(j) = m_j + w_n(j)$. From the definition of the conditional distribution of A_{i+1} given $\mathcal{A}_i$ and the fact that $A_{i+1} = Y_{i+1} - \max(Y_i - 1, 0)$ we can express the joined distribution as

$$
\begin{aligned}
&P(\cap_{k=0}^n \{Y_k = j_k\}) \\
&= \prod_{k=1}^n P(A_k = j_k - j_{k-1} + 1 - \delta_0(j_{k-1})|\mathcal{A}_{k-1}) \\
&= \prod_{k=1}^n \frac{s_{k-1}(j_k - (j_{k-1} - 1)_+)}{m_{k-1}(j_k - (j_{k-1} - 1)_+)} \prod_{i=0}^{j_k - (j_{k-1}-1)_+ - 1} \left(1 - \frac{s_{k-1}(i)}{m_{k-1}(i)}\right) \\
&= \frac{\prod_{k=1}^n s_{k-1}(j_k - (j_{k-1}-1)_+) \prod_{k=1}^n \prod_{i=0}^{j_k-(j_{k-1}-1)_+ - 1}[m_i(k-1) - s_{k-1}(i)]}{\prod_{k=1}^n \prod_{i=0}^{j_k-(j_{k-1}-1)_+ - 1} m_{k-1}(i)} \\
&= \prod_{j \geq 0} \frac{s_j^{[v_n(j)]}(m_j - s_j)^{[u_n(j)]}}{m_j^{[w_n(j)]}}
\end{aligned}
$$

(ii) Let **i** and **j** be to equivalent integer sequences and consider first the case where one sequence, say **i**, is inadmissible. Hence, there exists an index $1 \leq k \leq n$ and two integer r, l such that $r = i_k < i_{k-1} - 1 = l$ and therefore the number of transitions from l to r is positive in **i**. By equivalence the number of transitions from l to r in **j** is positive two and hence both **i** and **j** are null events. Now, consider the case where both sequences are admissible. Since $u_n(\cdot), w_n(\cdot)$ are functions of $v_n(\cdot)$, from part (i) the finite dimensional law of Y depends only on $v_n(\cdot)$. Therefore, it suffices to show that the counting process $v_n(\cdot)$ for $(X_0, \cdots, X_n) = \mathbf{i}$ and $(X_0, \cdots, X_n) = \mathbf{j}$ is identical. But since **i** and **j** are equivalent and admissible, for any integer

r, l

$$\sum_{k=0}^{n} I(i_k = r) = I(i_0 = r) + \sum_{l \leq r+1-\delta_0(r)} \sum_{k=1}^{n} I(i_{k-1} = l, i_k = r)$$

$$= I(j_0 = r) + \sum_{l \leq r+1-\delta_0(r)} \sum_{k=1}^{n} I(j_{k-1} = l, j_k = r) = \sum_{k=0}^{n} I(j_k = r)$$

(iii) To prove the last part of the lemma we need the following lemma. $\square$

Lemma 3. *Lemma A.1 For $m \in \mathbb{N}_0$ fixed define $S_n(m) = \prod_{i=0}^{m} \left(1 - \frac{s_i + v_n(i)}{m_i + w_n(i)} \right)$ for $n \geq 0$.*

(i) Then, $S_n(m) \xrightarrow{a.s.} S(m)$ as $n \to +\infty$.

(ii) Moreover, the limit $S(m) > 0$ with probability one.

Proof. Proof of lemma A.1 We show that $\{S_n(m)\}$ is a martingale with respect to the filtration $(\mathcal{F}_n)$. Since $S_n(m) \in [0, 1]$, (i) follows from the Martingale-convergence theorem. From (Walker and Muliere, 1997; Muliere et al., 2000) $A = (A_k)_{k \geq 1}$ is exchangeable and therefore an iid sequence given the tail sigma field of A, say $\mathcal{A}$. Fix $n \geq 1$, then

$$\mathbb{E}[S_{n+1}(m)|\mathcal{F}_n] = \mathbb{E}[\mathbb{P}[A_{n+2} > m)|\mathcal{F}_{n+1}]|\mathcal{F}_n]$$

$$= \mathbb{E}[\mathbb{E}[I(A_{n+2} > m)|\mathcal{F}_{n+1}]|\mathcal{F}_n]$$

$$= \mathbb{E}[\mathbb{E}[\mathbb{E}[I(A_{n+2} > m)|\mathcal{A}]|\mathcal{F}_{n+1}]|\mathcal{F}_n]$$

$$= \mathbb{E}[\mathbb{E}[I(A_{n+2} > m)|\mathcal{A}]|\mathcal{F}_n] \text{ (by the tower property)}$$

$$= \mathbb{E}[\mathbb{E}[I(A_{n+1} > m)|\mathcal{A}]|\mathcal{F}_n] \text{ (by conditional iid)}$$

$$= \mathbb{E}[I(A_{n+1} > m)|\mathcal{F}_n] = S_n(m).$$

(ii) Moreover, Proposition 5.2 of (Fortin et al., 2000) implies that $\left(S_n^{(m)}\right)_m$ convergence in distribution to a random complementary distribution function $S = (S(m), m \in \mathbb{N}_0)$ where form (Walker and Muliere, 1997; Muliere et al., 2000) the limit equals $S(m) = \prod_{k=1}^{m}(1 - \theta_k)$ with independent $\theta_k \sim$ Beta$(s_k, m_k - s_k)$. Therefore

$$\mathbb{P}(S(m) > 0) = \mathbb{P}\left[\prod_{k=1}^{m}(1 - \theta_k) > 0\right] = \prod_{k=0}^{m}\mathbb{P}(\theta_k < 1) = 1.$$

The last equality holds since by assumption $m_k - s_k > 0$ for $k = 0, \cdots, m$. Since limits are unique a.s. it follows that $S(m) > 0$ a.s.. $\qquad \square$

Proof. Proof of lemma 1(iii) From the predictive probability law of $\{A_n\}$ and since $Y_{n+1} = A_{n+1} + \max(Y_n - 1, 0)$, for a fix $x_0 \in \mathbb{N}_0$

$P(Y_{n+1} = x_0 | \mathcal{F}_n, Y_0 = x_0)$

$= \frac{s_{x_0 - X_n + 1 - \delta_0(X_n)} + v_n(x_0 - X_n + 1 - \delta_0(X_n))}{m_{x_0 - X_n + I(X_n \geq 1)} + w_n(x_0 - X_n + 1 - \delta_0(X_n))} \prod_{j=0}^{x_0 - X_n - \delta_0(X_n)} \left(1 - \frac{s_j + v_n(j)}{m_j + w_n(j)}\right)$

$= \frac{s_{x_0 - X_n + 1 - \delta_0(X_n)} + v_n(x_0 - X_n + 1 - \delta_0(X_n))}{m_{x_0 - X_n + I(X_n \geq 1)} + w_n(x_0 - X_n + 1 - \delta_0(X_n))} S_n(x_0 - X_n - \delta_0(X_n))$

$\geq \frac{\inf_i s_i}{\sup_i m_i + n} S_n(x_0) \ a.s. \qquad (10)$

where the last inequality follows since $S_m(\cdot)$ is non-increasing and $X_n \geq 0$ a.s. Hence we have that

$$\sum_{n=1}^{\infty} P(Y_n = x_0 | \mathcal{F}_n) \geq \sum_{n=1}^{\infty} \frac{\inf_i s_i}{\sup_i m_i + n} S_n(x_0) \ a.s. \qquad (11)$$

If the sum on the right hand side of (11) divergence with probability one,

then from Levy's extension of the Borel-Cantelli Lemma (Williams, 1991) it follows that $\mathbb{P}(\limsup_n Y_n = x_0 | Y_0 = x_0) = 1$. Since the first term in the sum on the right hand side of (11) is of order $O(1/n)$ it suffices to show that $S_n(x_n) \overset{a.s.}{\to} S(x_0)$ and $S(x_0) > 0$ a.s.. But both facts follow from the previous lemma. $\qquad\square$

Proof. Proof of lemma 2 By Lemma 1, Y is recurrent and partial exchangeable, which by Theorem 2 in (Fortini et al., 2002) is equivalent to $V = \{V_j(k)\}_{k \geq 1, j \geq 0}$ being row-wise exchangeability, i.e. $\{V_i(n); i \geq 0, n \geq 1\} \overset{d}{=} \{V_i(\sigma)i(n)); i \geq 0, n \geq 1\}$ for any finite permutation σ_k of $\mathbb{N}$ for $k \geq 0$. Moreover, from (Walker and Muliere, 1997; Muliere et al., 2003) $\left\{S_n^{(i)}\right\}_{n \geq 1}$ are exchangeable as well for $i = 1, 2$. The successor states and howling times $V = \{V_j(k)\}_{k \geq 1, j \geq 0}$ and $T = \{T_i(k)\}_{k \geq 1, j \geq 0}$ are tied together only through the hitting times $(\tau_i(n); n \geq 1, i \in \mathbb{N}_0)$. Let $\{j_{i,k}\}_{i \leq N, k \leq n}$ be a sequence of successor states. Without lost of generality assume $j_{0,k} \geq 0$ and $j_{i,k} \geq i - 1$ for all $1 \leq k \leq N$ and $1 \leq i \leq n$, since otherwise the lemma follows trivially since the sequence $\{j_{i,k}\}_{i \leq N, k \leq n}$ and any row-wise permutations of it is a

27

null event. Then, a.s.

$$\mathbb{P}\left[\bigcap_{i=0}^{N}\bigcap_{k=1}^{n}\{T_i(k)\leq t_{i,k}\}\bigg|\bigcap_{i=0}^{N}\bigcap_{k=1}^{n}\{V_i(k)=j_{i,k},\tau_i(n)\}\right]$$

$$=\mathbb{P}\left[\bigcap_{k=1}^{n}\{S_k^{(1)}\leq t_{0,k}\}\bigcap_{1\leq i\leq N}\bigcap_{k=1}^{n}\{S_{\tau_i(k)+1-h(\tau_i(n))}^{(2)}\leq t_{i,k}\}\bigg|\bigcap_{i=0}^{N}\bigcap_{k=1}^{n}\{\tau_i(n)\}\right]$$

$$=\mathbb{P}\left[\bigcap_{k=1}^{n}\{S_k^{(1)}\leq t_{0,\sigma_0(k)}\}\bigcap_{1\leq i\leq N}\bigcap_{k=1}^{n}\{S_{(i-1)n+k}^{(2)}\leq t_{i,\sigma_i(k)}\}\bigg|\bigcap_{i=0}^{N}\bigcap_{k=1}^{n}\{\tau_i(n)\}\right]$$

$$=\mathbb{P}\left[\bigcap_{i=0}^{N}\bigcap_{k=1}^{n}\{T_i(k)\leq t_{i,\sigma_i(k)}\}\bigg|\bigcap_{i=0}^{N}\bigcap_{k=1}^{n}\{V_i(k)=j_{i,\sigma_i(k)},\tau_i(n)\}\right]. \tag{12}$$

where the second equality follow be exchangeability of $\{S_k^{(i)}\}, i=1,2$. Now,

Lemma 2 follows from

$$\mathbb{P}\left[\bigcap_{i=0}^{N}\bigcap_{k=1}^{n}\{V_i(k)=j_{i,k},T_i(k)\leq t_{i,k}\}\right]$$

$$=\mathbb{P}\left[\bigcap_{i=0}^{N}\bigcap_{k=1}^{n}\{T_i(k)\leq t_{i,k}\}\bigg|\bigcap_{i=0}^{N}\bigcap_{k=1}^{n}\{V_i(k)=j_{i,k}\}\right]\mathbb{P}\left[\bigcap_{i=0}^{N}\bigcap_{k=1}^{n}\{V_i(k)=j_{i,k}\}\right]$$

$$=\mathbb{E}\left[\mathbb{P}\left[\bigcap_{i=0}^{N}\bigcap_{k=1}^{n}\{T_i(k)\leq t_{i,k}\}\bigg|\bigcap_{i=0}^{N}\bigcap_{k=1}^{n}\{V_i(k)=j_{i,k},\tau_i(n)\}\right]\bigg|\bigcap_{i=0}^{N}\bigcap_{k=1}^{n}\{V_i(k)=j_{i,k}\}\right]$$

$$\times\mathbb{P}\left[\bigcap_{i=0}^{N}\bigcap_{k=1}^{n}\{V_i(k)=j_{i,k}\}\right]$$

$$=\mathbb{E}\left[\mathbb{P}\left[\bigcap_{i=0}^{N}\bigcap_{k=1}^{n}\{T_i(k)\leq t_{i,\sigma_i(k)}\}\bigg|\bigcap_{i=0}^{N}\bigcap_{k=1}^{n}\{V_i(k)=j_{i,\sigma_i(k)},\tau_i(n)\}\right]\bigg|\bigcap_{i=0}^{N}\bigcap_{k=1}^{n}\{V_i(k)=j_{i,\sigma_i(k)}\}\right]$$

$$\times\mathbb{P}\left[\bigcap_{i=0}^{N}\bigcap_{k=1}^{n}\{V_i(k)=j_{i,\sigma_i(k)}\}\right]$$

$$=\mathbb{P}\left[\bigcap_{i=0}^{N}\bigcap_{k=1}^{n}\{V_i(k)=j_{i,\sigma_i(k)},T_i(k)\leq t_{i,\sigma_i(k)}\}\right]$$

where the third equality follows from (12) and the fact that V is row-wise exchangeable. $\qquad \square$

Proof. Proof of Corollary 1 (i) Since (V, T) is row-wise partial exchangeable de Finetti's theorem for row-wise partial exchangeable arrays (Link, 1980; De Finetti, 1980; Epifani et al., 2002) gives the existence of a $\mathbb{W}$-valued random element $W = (W_i)_{i \in \mathbb{N}_0}$ such that $(V_i(n), T_i(n))_n \stackrel{iid}{\sim} W_i(\cdot, \cdot)$ for any $i \in \mathbb{N}_0$. Since $\{V_i(n), T_i(n) | \tau_i(n) = k\} = \{Y_{k+1}, S_{k+1} - S_k | Y_k = i, \tau_i(n) = k\}$ this gives (i).

(ii) Moreover, from (Epifani et al., 2002), for $i \in \mathbb{N}_0$ and $t \geq 0$ the empirical transition kernel $W_i^{(n)}(j, [0, t]) = n^{-1} \sum_{k=1}^{n} I(V_i(k) = j, T_i(k) \leq t)$ converges weakly to $W_i(j, [0, t]) \stackrel{d}{=} \Pi_{i,j} F_{i,j}(t)$; and both $\Pi_{i,j}^{(n)} = \sum_{k=1}^{n} I(V_i(k) = j)$ and $F_{i,j}^{(n)}(t) = \sum_{k=1}^{n} I(T_i(k) \leq t)$ converge weakly to $\Pi_{i,j}$ and $F_{i,j}(t)$.

Form (Fortin et al., 2000; Walker and Muliere, 1997; Muliere et al., 2000), $\lim_n n^{-1} \sum_{1 \leq k \leq n} I(A_k = j) \stackrel{d}{=} \lim_n P(A_{n+1} = j | \mathcal{F}_n) \stackrel{d}{=} Q_j$ a.s., where $Q = (Q_j, j \geq 0)$ is a random probability with distribution as stated in Corollary 1 (ii). The distributional form of Π follows now from $\mathbb{P}[V_i(n) = j] = \mathbb{P}[A_{\tau_i(n)+1} = j - i + I(i > 0)] = \mathbb{P}[A_n = j - i + I(i > 0)]$, where the second equality follows from a stopping-time result for exchangeable random variables in (Kallenberg, 1982) and (Aldous, 1985) Theorem 6.1 page p.42.

Similarly, since $S^{(2)}$ is exchangeable and independent of $\{Y_n\}$, by the same stopping-time property for exchangeable random variables, for $i > 0, j \geq i+1-\delta_0(i), t \geq 0$ and $\tau_{i,j}(n+1) = \inf\{m > \tau_{i,j}(n) : Y_m = i, Y_{m+1} = j\}$

$$F_{i,j} \stackrel{d}{=} \lim_n n^{-1} \sum_{k=1}^{n} I(S^{(2)}_{\tau_{i,j}(k)+1-h(\tau_{i,j}(k))} \leq t) \stackrel{d}{=} \lim_n n^{-1} \sum_{k=1}^{n} I(S^{(2)}_{k+1-h(k)} \leq t) =: G^{(2)}(t).$$

Which shows that the limit, if it exists, does not dependent on $i > 0$, $j \geq 0$ and Π. Moreover, the exchangeable sequence $\left\{S_n^{(2)}\right\}_n$ has the same system of predictive distributions as an exchangeable sequence of random variables with beta-Stacy mixing measure and parameters $(\alpha^{(2)}, \beta^{(2)})$, $G \sim BS(\alpha^{(2)}, \beta^{(2)})$ (Walker and Muliere, 1997; Muliere et al., 2000, 2003). Since the predictive distribution uniquely determined the system of finite dimensional probability law of $\left\{S_n^{(2)}\right\}_n$ this implies $G^{(2)} \sim BS(\alpha^{(2)}, \beta^{(2)})$. The case $F_{0,i} \overset{d}{=} G^{(1)} \sim BS(\alpha^{(1)}, \beta^{(1)})$ can be shown similarly.

$\square$

Model	Queue length	Point estimator		90%-CI	
		Bias	MSE	Width	Coverage
		Traffic Intensity ρ			
1	1e+03	-9.2e-03	4.2e-03	2.1e-01	8.9e-01
1	1e+04	-6.2e-04	4.1e-04	6.8e-02	9.0e-01
1	1e+05	-9.0e-05	4.2e-05	2.1e-02	9.2e-01
2	1e+03	-1.7e-03	2.5e-03	1.7e-01	9.0e-01
2	1e+04	-2.5e-04	2.7e-04	5.4e-02	9.0e-01
2	1e+05	-3.2e-05	2.6e-05	1.7e-02	9.1e-01
3	1e+03	-3.2e-03	2.8e-03	1.7e-01	9.0e-01
3	1e+04	-1.0e-03	2.9e-04	5.5e-02	9.0e-01
3	1e+05	-3.6e-04	2.9e-05	1.7e-02	9.1e-01
4	1e+03	-1.6e-02	6.4e-03	2.5e-01	8.7e-01
4	1e+04	-8.6e-04	6.6e-04	8.3e-02	8.8e-01
4	1e+05	2.9e-05	6.9e-05	2.6e-02	8.8e-01
		Service time mean			
1	1e+03	5.9e-02	3.3e-02	5.9e-01	9.0e-01
1	1e+04	8.9e-03	2.8e-03	1.7e-01	8.9e-01
1	1e+05	1.8e-03	2.7e-04	5.4e-02	9.0e-01
2	1e+03	2.8e-03	3.4e-03	1.9e-01	8.8e-01
2	1e+04	1.6e-04	3.4e-04	5.9e-02	8.9e-01
2	1e+05	-1.5e-04	3.4e-05	1.9e-02	9.0e-01
3	1e+03	1.8e-02	6.1e-03	2.5e-01	9.0e-01
3	1e+04	1.8e-03	5.7e-04	7.7e-02	8.9e-01
3	1e+05	1.7e-04	6.2e-05	2.4e-02	8.8e-01
4	1e+03	1.3e-01	1.1e-01	9.5e-01	8.8e-01
4	1e+04	2.6e-02	7.3e-03	2.7e-01	8.8e-01
4	1e+05	3.9e-03	6.1e-04	8.1e-02	8.9e-01
		Mean Number of items in equilibrium			
1	1e+03	-1.7e-02	1.3e+0	5.8e+0	9.2e-01
1	1e+04	8.2e-03	1.2e-01	1.4e+0	9.5e-01
1	1e+05	3.1e-04	1.2e-02	4.3e-01	9.5e-01
2	1e+03	5.3e-02	3.3e-01	2.6e+0	9.5e-01
2	1e+04	2.8e-03	2.7e-02	6.3e-01	9.3e-01
2	1e+05	1.9e-04	2.5e-03	2.0e-01	9.5e-01
3	1e+03	4.8e-02	3.9e-01	2.8e+0	9.3e-01
3	1e+04	-4.8e-03	3.2e-02	6.9e-01	9.4e-01
3	1e+05	-3.8e-03	3.2e-03	2.2e-01	9.4e-01
4	1e+03	-4.9e-01	3.1e+0	9.7e+0	8.6e-01
4	1e+04	-3.5e-02	5.3e-01	2.9e+0	9.4e-01
4	1e+05	-1.0e-03	5.6e-02	9.0e-01	9.4e-01

Table 2: Predictive performance for all queuing models and different length of observation