Using Infinite Server Queues Theory In Stress Testing

by

Guichang Zhang

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Mathematics
in
Quantitative Finance

Waterloo, Ontario, Canada, 2016

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

In this thesis, we propose and study a framework to model stress testing, using an infinite server queues theory, such that this framework is aligned with and integrates several existing frameworks currently used in the industry. On the other hand, memoryless property plays a critical rule in the framework of industry models. Our proposed framework will provide a mathematical analysis for the memoryless property as well.

Acknowledgements

I would like to thank my supervisor Professor Tony Wirjanto and the readers for their valuable comments and suggestions.

Dedication

This thesis is dedicated to my wife, Liping Wang, who always supports me in chasing success, and also to my son, Yuchen Zhang.

Table of Contents

1 Introduction **1**

1.1 Stress testing and capital planning background 1

1.2 Infinite-server queueing theory background 5

1.3 Proposed queueing theory framework 6

2 A Non-homogeneous Poisson process **8**

3 Infinite-server queues **16**

3.1 A Kolmogorov-type equation for the $M(t)/M(t)/\infty$ queue 16

3.2 Generalized equation for the $M(t)/M(t)/\infty$ queue 20

3.3 Main theorems for the $M(t)/M(t)/\infty$ queue 21

3.4 The $M(t)/G(t)/\infty$ queue 25

3.5 The mean and variance of the queue-length process 28

4 Non-homogeneous Poisson process in infinite-server queues **31**

5 Implementation of the infinite-server queue framework and Conclusion **34**

References **36**

Chapter 1

Introduction

In this chapter, we first introduce research topic of this thesis - stress testing practice in financial industry. The requirements and backgrounds of the stress testing are then elaborated and the existing approaches are compared by reviewing the corresponding literature.

Then the backgrounds and main results in finite server queues are presented in the next section. The reviewed literature is selected to be relevant to the infinite-server queues which depend on time dependent parameters.

1.1 Stress testing and capital planning background

The Comprehensive Capital Analysis and Review (CCAR) is an annual exercise by the Federal Reserve to assess whether the largest Bank Holding Companies (BHCs) operating in the United States have sufficient capital to continue operations through times of economic and financial stress and whether they have robust, forward-looking capital-planning processes that account for their unique risks.

As part of this exercise, the Federal Reserve evaluates institutions' capital adequacy, internal capital adequacy assessment processes, and their individual plans to make capital distributions, such as dividend payments or stock repurchases. Dodd-Frank Act stress testing (DFAST)- a complementary exercise to CCAR - is a forward-looking component conducted by the Federal Reserve and financial companies supervised by the Federal Reserve to help assess whether institutions have sufficient capital to absorb losses and support operations during adverse economic conditions.

While DFAST is complementary to CCAR, both efforts are distinct testing exercises that rely on similar processes, data, supervisory exercises, and requirements. The Federal Reserve coordinates these processes to reduce duplicative requirements and to minimize regulatory burden.

In the CCAR or DFAST exercise, the core component is to calculate the retail credit loss projections (for example, Net Charge Off (NCO) or Specific Provision for Credit Loss (SPCL)), under different stressed scenarios. Usually the calculation of credit loss projections comprises three sections: probability of default (PD), loss given default (LGD) and exposure at default (EAD). The current practice usually sets EAD constant without any models, and employs different approaches to model PD and LGD. There are a lot of formations to express the credit loss projections. As an example, the definition of SPCL is given below.

$$SPCL_{forecast} = PD_{forecast} \times LGD_{forecast} \times EAD$$

where the Specific Provision for Credit Loss - SPCL is an estimate of losses that will be incurred on exposures that have been identified as impaired.

In this thesis, we mainly focus on the modelling approaches for the PD. Our framework might be applicable to the LGD as well. In the following, we will briefly introduce the existing modelling approaches for PD estimation currently used in the industry.

- Segment-level econometric models

  First, several segments within the retail portfolio are generated by assuming that the loan applicants in each segment will respond to the changing economics conditions in a homogenous way. In this approach, each segment is modelled separately, based on independently selected macroeconomic drivers.

- Transition matrix approach

  This approach captures the transition probability from one state of the loan to another one (for example from CURRENT to 30 days-past-due (DPD)). At the segment level described in the first approach, those transition probabilities are regressed against macroeconomic variables as well as loan-level factors.

- Loan-level hazard rate models

  The hazard rate model predicts a probability of default or a payoff event over time given that the loan has survived.

In the following, we will elaborate the transition matrix and hazard rate modelling approaches. Our new framework originates from these two approaches, but provides a higher level of integration and a convenient insight into the models' properties (for example memoryless property).

With the transition matrix approach, each loan at a time point includes several different states: current, 1-29 DPD, 30-59 DPD, 60-89 DPD, 90-119 DPD, DEFAULT, and Prepayment, which can transit to others at the next time point. The advantage of this approach is that it not only models the likelihood of the terminal events (prepayment and default), but also tracks credit path and balance migration of individual loans. For any given point in time, the loan state movement starts in one state and moves successively to another state in a monthly basis until termination. The transition matrix framework provides monthly predictions of delinquency status transition probabilities among active delinquency statuses and into terminal prepayment and default outcomes. Transition probabilities are then iterated each period to generate one-month ahead and cumulative event probabilities of default/prepayment and the distribution of delinquency stocks at each period in the forecasting horizon.

The first application was by Cyert et al (1962), who developed a Markov chain model of customer's repayment behavior. Subsequently more complex models have been developed by Ho (2001), Thomas, Ho, and Scherer (2001) and Trench et al. (2003). Schniederjans and Loch (1994) used Markov chain models to model the marketing aspects of a customer relationship in the banking environment.

Behavioral score based Markov chain models are sometimes used in the industry (see Scallan, 1998), but mainly as ways of assessing provisioning estimates, and they do not include economic drivers. Most recently, Malik and Thomas (2012) applied the transition matrix approach to the the consumer credit rating.

The transition matrix model relies on a finite state space non-stationary first order Markov chain assumption. In other words, a one-month-ahead transition probability only depends on a loan's current state information/status and does not exhibit any path-dependence pattern. This is the memoryless property of Markov chain. Under our framework, we study this property closely. The memoryless property is a limitation here to fail to characterize the path information. In the non-Markovian framework, one may use Backward Stochastic Differential Equation (BSDE) techniques.

In the recent years survival analysis together with hazard rate modelling has been introduced into credit scoring. Survival analysis is the area of statistics that deals with the analysis of lifetime data. The variable of interest is the time to an event.

Survival analysis in credit scoring was introduced by Narain (1992), see e.g., Stepanova

and Thomas (2002). It was further developed by Thomas et al. (1999). The event of interest is default. Narain (1992) applied an accelerated life exponential model to personal loan data. He found that this model estimated the number of failures well at each time interval. Next, he showed that credit granting decisions could be improved by using the methods of survival analysis as compared to a multiple regression framework. Finally, the author argued that survival analysis can be used in all credit operations in which there are predictor variables and the time to an event is of interest. Thomas et al. (1999) made a comparison of the performance of exponential, Weibull, and Cox models with a logistic regression and found that the survival-analysis methods are competitive with, and sometimes superior to, the traditional logistic-regression approach. This indicates that survival analysis may be useful for accurate PD estimation for a fixed 12 months horizon for various types of loans, which, in turn, is useful for PD estimation within the Basel II Accord (Tong et al., 2012).

Suppose that T is the length of time before a facility defaults. The randomness of T can be described in the following three standard ways.

The distribution function describes the probability that the time to event (T) is less than or equal to a fixed time (t) and is given by:

$$F(t) = P(T \leq t)$$

From this, the survival function, the probability that the time to event (T) is larger than a fixed time (t), can be derived as:

$$S(t) = 1 - F(t)$$

The second way works through a probability density function $f(t)$, which is given by:

$$f(t) = \lim_{\Delta t \downarrow 0} \frac{P(t \leq T \leq t + \Delta t)}{\Delta t}$$

The last description is given by the hazard function $h(t)$, which is given by:

$$h(t) = \lim_{\Delta t \downarrow 0} \frac{P(t < T \leq t + \Delta t | T > t)}{\Delta t} = \frac{f(t)}{1 - F(t)} = \frac{f(t)}{S(t)}$$

The hazard function $h(t)$, also called an incidence rate, instantaneous risk or force of mortality, is the event rate at t among those at risk at time t. It is actually the same as the upcoming intensity function $\lambda(t)$ to be introduced in the next Section.

1.2 Infinite-server queueing theory background

In this subsection, we focus on infinite-server queues with time dependent parameters. In particular, we study $M(t)/M(t)/\infty$ queue and its generalization $M(t)/G(t)/\infty$ queue. Eick, Massey and Whitt (1993a and 1993b) studied the $M(t)/G/\infty$ queue from the perspective of a point-wise stationary approximation for multi server queues. Mandelbaum and Massey (1995) employed the concepts of time-dependent fluid and diffusion approximations to explore the asymptotic behavior of time-dependent queues. See also Mandelbaum, Massey and Reiman (1998) and Mandelbaum et al. (1999) for similar approaches. More recently, Pang and Whitt (2010) used heavy-traffic limits to approximate infinite-server queues under a more general setting. See also Massey (2002) and Fralix and Adan (2008) for related work. Zhang and Srinivasan (2013) applied a martingale approach to obtain explicit solutions to the infinite-server queues, which are the main foundation of our framework. The approach we use is similar to that of Abramov (2006 and 2007), however substantial modifications are introduced to make the derivation transparent and easy to adopt. Ellis (2010) derived similar expressions for the mean and variance of the $M(t)/G(t)/\infty$ queue by using the techniques developed in El-Sherbiny (2010).

Below, we present the $M(t)/M(t)/\infty$ and $M(t)/G(t)/\infty$ queues framework.

The arrival process $A(t)$ is assumed to be a nonhomogeneous Poisson process with a deterministic arrival rate function $\lambda(t), \quad 0 \leq t < \infty$. There is an infinite number of servers and the service times associated with each server for each customer are independent and assumed to be exponentially distributed with a deterministic rate function $\mu(t), \quad 0 \leq t < \infty$, i.e., we assume that there is a potential nonhomogeneous Poisson process $\pi(t)$ with rate $\mu(t)$ for each server. In this thesis, we assume that λ and μ as functions of t are nonnegative, measurable and integrable over any bounded interval. For the $M/M/\infty$ queue, $\lambda(t) \equiv \lambda, \mu(t) \equiv \mu$. When $A(t)$ is a nonhomogeneous Poisson process with rate $\lambda(t)$, we know that (see Brémaud, 1981, p.25) it has the following Doob-Meyer decomposition

$$A(t) = \int_0^t \lambda(s)ds + M_A(t),$$

where $M_A(t)$ is a martingale. When $A(t)$ is a homogeneous Poisson process with constant rate λ, this reduces to

$$A(t) = \lambda t + M_A(t).$$

For the $M(t)/G(t)/\infty$ queue, the arrival process is the same as the $M(t)/M(t)/\infty$, while the departure process is different. Under the $M(t)/G(t)/\infty$ queue framework, the

service times associated with each server for each customer are independent and assumed to follow a general distribution function $G(s, \cdot)$ if the service started at time s, which is time dependent.

1.3 Proposed queueing theory framework

The objective of this thesis is to estimate the probability of default (PD) of the loans within a specific segment of a retail portfolio. The loans in each segment of the portfolio are assumed to be homogeneous in response to macroeconomic condition changes and share the same PD for the purpose of calculating the SPCL.

For a segment of the retail portfolio, there are a lot of loans (The number of the loans denoted by N is usually more than a thousand even in a small portfolio). Under the queueing theory framework, when a loan is booked with a bank, we treat this loan entering a server (arrival); when a loan is either defaulted or mature, we treat it as if it leaves the server (departure). Since N is usually very large ($N > 1000$), we will use the infinite-server queue ($M(t)/M(t)/\infty$ and $M(t)/G(t)/\infty$) to model this procedure.

Note that for the $M(t)/M(t)/\infty$ queue, we have two kinds of departures: either defaulted or mature. In this thesis, we assume that maturity departure is negligible (For the mortgage portfolio, the loan amortization is usually longer than ten years). Future research can take into consideration of maturity departures.

We conclude that our proposed new queueing theory framework incorporates the aforementioned transition matrix and hazard rate approaches for the credit scoring purpose, according to the analyses in the following chapters. For example, the memoryless property, which is critical to the transition matrix and hazard rate approaches, are examined in the context of non-homogeneous Poisson Processes in the next chapter. For the main results regarding infinite-sever queues, we refer to our previous research by Zhang and Srinivasan (2013).

The main contributions of this thesis are as follows.

- We provide a new framework for the estimation of PD in the stress testing environment, which incorporates the transition matrix and hazard rate approaches;
- This new framework provides a more convenient way to re-examine the memoryless property, which plays an important role in all the approaches. For the first time, we define the quasi-memoryless property;

- Within the new framework, we investigate the quasi-memoryless property with the infinite-server queue theory.

Chapter 2

A Non-homogeneous Poisson process

It is well known that a Poisson Process and a Non-homogeneous Poisson Process (NHPP for short) are widely used models in queueing theory. There are a lot of publications on Poisson Processes. The NHPP is probably the best known generalization of the Poisson Process (see for example Ross (2003)). In this thesis, we focus on the distribution of inter-arrival times and memoryless property. At first, we give the following definition.

Definition 2.1 *The counting process* $N = \{N(t) : t \geq 0\}$ *is said to be a non-homogeneous Poisson Process with a time-varying intensity function* $\lambda(t), t \geq 0$*, if:*
1. $N(0) = 0$*;*
2. N *has independent increments;*
3. $P\{N(t+h) - N(t) = 1\} = \lambda(t)h + o(h)$*;*
4. $P\{N(t+h) - N(t) \geq 2\} = o(h)$

Notice that from 3 and 4 in the definition, we obtain

$$P\{N(t+h) - N(t) = 0\} = 1 - \lambda(t)h + o(h)$$

Below we give some properties about NHPP (Ross 2003).

Theorem 2.2 *If the NHPP* N *is defined in the Definition 2.1, then we have*

$$\lambda(t) = \lim_{h \to 0} \frac{P\{N(t+h) = i+1 | N(t) = i\}}{h}, \quad \forall t \geq 0, \quad \forall i = 0, 1, 2, \cdots .$$

Notice that $\lambda(t)$ is independent of state i.

Define

$$\mathbf{Q}(t) = (q_{ij}(t), i, j = 0, 1, 2, \cdots), \quad \forall t \geq 0$$
$$q_{ij}(t) = \lim_{h \to 0} \frac{P\{N(t+h) = j | N(t) = i\}}{h}, \forall i \neq j$$
$$q_{ii}(t) = -\sum_{j \neq i} q_{ij}(t)$$

then we can get the Q-matrix as follows

$$\mathbf{Q}(t) = \begin{pmatrix} -\lambda(t) & \lambda(t) & & \\ & -\lambda(t) & \lambda(t) & \\ & & \ddots & \ddots \end{pmatrix}$$

Obviously the NHPP N is a pure birth process with the starting state 0 and Q-matrix $\mathbf{Q}(t)$.

Define the integrated intensity function $\Lambda(t) = \int_0^t \lambda(s)ds$.

Theorem 2.3 *If the NHPP N is defined in the Definition 2.1, then we obtain*

$$P\{N(t) - N(s) = k\} = \frac{(\Lambda(t) - \Lambda(s))^k}{k!} e^{-[\Lambda(t) - \Lambda(s)]}, \quad k = 0, 1, 2, \cdots, \quad 0 \leq s \leq t \quad (2.1)$$

In the following, we consider the inter-arrival time distribution. Denote by $X_i, i = 1, 2, \cdots$, the i-th inter-arrival time random variable.

Theorem 2.4 *If the NHPP N is defined in the Definition 2.1, then we have the following inter-arrival time distribution*

$$G_{X_1}(t) = 1 - e^{-\Lambda(t)}, \quad \forall t \geq 0;$$
$$G_{X_i,s}(t) = 1 - e^{-[\Lambda(t+s) - \Lambda(s)]}, \quad \forall t \geq 0, \quad i = 2, 3, \cdots$$

where

$$G_{X_1}(t) = P(X_1 \leq t)$$
$$G_{X_i,s}(t) = P(X_i \leq t | \sum_{k=1}^{i-1} X_k = s) \quad (2.2)$$

Notice that in Theorem 2.4, for $i, j = 2, 3, \cdots$, we have

$$G_{X_i,s}(t) = G_{X_j,s}(t)$$

provided that it is conditional on the same jump point s (no matter how many jumps before s). In fact, if we define $G_{X_1,s}(t)$ in a similar way as (2.2) properly, we also have

$$G_{X_i,s}(t) = G_{X_1,s}(t), \quad i = 2, 3, \cdots \tag{2.3}$$

We will make this point clear after the next theorem. Henceforth we call this property as a conditional identical distribution for the inter-arrival time distribution of NHPP.

Corollary 2.5 *Denote by $X_i, i = 1, 2, \cdots$ the i-th inter-arrival time random variable in homogeneous Poisson Process N with the arriving rate λ. Then $X_i, i = 1, 2, \cdots$ are i.i.d and exponentially distributed with parameter λ.*

$$G_{X_i}(t) = P(X_i \leq t) = 1 - e^{-\lambda t}, \quad \forall t \geq 0$$

Define

$$S_i = \sum_{j=1}^{i} X_j, \quad \forall i = 1, 2, \cdots; \quad S_0 = 0$$

For each $i = 1, 2, \cdots$, S_i is the i-th jump point of NHPP N, which is a random variable and has the following properties,

$$N(S_i-) = i - 1; \quad N(S_i) = i$$

$$0 \leq S_1 \leq S_2 \leq S_3 \leq \cdots$$

$$X_i = S_i - S_{i-1}$$

Now let us consider the independent property of inter-arrival times X_i of NHPP. The following result shows that such X_i have a conditional independent property.

Theorem 2.6 *NHPP N is defined in the Definition 2.1. $X_i, i = 1, 2, \cdots$, is its i-th inter-arrival time random variable. $S_n = \sum_{i=1}^{n} X_i, n = 1, 2, \cdots$, is its n-th jumping point. Then we have the following conditional independent property for all $X_i's$.*

$$\begin{aligned} & P(X_{n+1} > t | S_n = s, X_n = t_n, \cdots, X_1 = t_1) \\ = \; & P(X_{n+1} > t | S_n = s) \\ = \; & e^{-[\Lambda(t+s)-\Lambda(s)]} \end{aligned}$$

where $t \geq 0, 0 < t_i < s, i = 1, 2, \cdots, n, \sum_{i=1}^{n} t_i = s$.

Proof:

$$
\begin{aligned}
& P(X_{n+1} > t | S_n = s, X_n = t_n, \cdots, X_1 = t_1) \\
= \; & P(N(t+s) - N(s) = 0 | N(s) - N(\tau) = 1, \forall \tau \in [s - t_n, s); N(s - t_n) - N(\tau) = 1, \\
& \forall \tau \in [s - t_n - t_{n-1}, s - t_n); \cdots; N(t_1) - N(\tau) = 1, \forall \tau \in [0, t_1)) \\
= \; & P(N(t+s) - N(s) = 0) \\
= \; & e^{-[\Lambda(t+s) - \Lambda(s)]} \\
= \; & P(X_{n+1} > t | S_n = s)
\end{aligned}
$$

In the second equality of the above equation, we have used the independent increments property of NHPP N. The above result shows that given the current jumping point, the following inter-arrival time is independent of the previous inter-arrival times.

□

Theorem 2.7 *Assume that*
1. NHPP N is defined in the Definition 2.1;
2. s_1 *is a jump point of N (*s_1 *is deterministic),* Y_1 *is the inter-arrival time of N between* s_1 *and the next jump point* s_2 *(*s_2 *is random). i.e.* $Y_1 = s_2 - s_1$*;*
3. The current time is $s(s_1 < s)$*, there is no jump between* s_1 *and* s *for N;*
4. Y_2 *is the time random variable between the current time* s *and* s_2*. i.e.* $Y_2 = s_2 - s$*, then for all* $t \geq s$ *we have:*

$$
\begin{aligned}
P(Y_2 > t - s | Y_1 > s - s_1) &= P(Y_1 > t - s_1 | Y_1 > s - s_1) \\
&= e^{-[\Lambda(t) - \Lambda(s)]} \\
&= 1 - G_{Y_2, s}(t - s) \\
&= P(Y_2 > t - s | s \textit{ is a jump point})
\end{aligned}
$$

Proof:

$$
\begin{aligned}
P(Y_2 > t - s | Y_1 > s - s_1) &= P(Y_1 > t - s_1 | Y_1 > s - s_1) \\
&= \frac{P(Y_1 > t - s_1 | s_1 \text{ is a jump point})}{P(Y_1 > s - s_1 | s_1 \text{ is a jump point})} \\
&= \frac{e^{-[\Lambda(t) - \Lambda(s_1)]}}{e^{-[\Lambda(s) - \Lambda(s_1)]}} \\
&= e^{-[\Lambda(t) - \Lambda(s)]} \\
&= 1 - G_{Y_2, s}(t - s) \\
&= P(Y_2 > t - s | s \text{ is a jump point})
\end{aligned}
$$

□

What information does Theorem 2.7 give us? Conditional on the inter-arrival time Y_1 being larger than $s - s_1$, where s_1 is a jump point, the probability of $Y_2 > t - s$, that is the probability of $Y_1 > t - s_1$, equals to the probability of $Y_2 > t - s$ conditional on s is a jump point. It tells us that when we consider the distribution of Y_2 at the current time, we can treat Y_2 as the inter-arrival time of N and treat s as the jump point from when Y_2 starts, even if s is not necessarily the real jump point for N. This property is also a kind of memoryless property. Here we name it as quasi-memoryless property. Due to this property, we can consider $M(t)/M(t)/\infty$ queue at any beginning time instead of 0. We will make it clear in section 3.

Corollary 2.8 *Denote by $X_i, i = 1, 2, \cdots$, the i-th inter-arrival time random variable in a homogeneous Poisson Process N with the arrival rate λ. Then $X_i, i = 1, 2, \cdots$ have the memoryless property. That is*

$$P(X_i > t + s | X_i > s) = P(X_i > t)$$

Now let us revisit Theorem 2.4. Applying Theorem 2.7 to X_1 in Theorem 2.4 and changing X_1 as a time interval length from current time s to the first jump point, we obtain

$$\begin{aligned} G_{X_1,s}(t) &= P(X_1 \leq t | \text{ there is no jump before the current time s}) \\ &= 1 - e^{-[\Lambda(t+s)-\Lambda(s)]} \end{aligned}$$

which coincides with the equation in (2.3).

Considering Theorem 2.4 together with Theorem 2.7, we can get a stronger result as follows.

Theorem 2.9 *Assume that the NHPP N is defined in the Definition 2.1. Denote by X the time interval length random variable from the current time s, which is deterministic, to the next jump point. Then we have*

$$G_{X,s}(t) = P(X \leq t | \text{ s is the current time}) = 1 - e^{-[\Lambda(t+s)-\Lambda(s)]}, \quad \forall t \geq 0$$

Notice that in Theorem 2.9, the result reduces to Theorem 2.4 if s is a jump point. Otherwise, we will use Theorem 2.7 instead.

In Theorem 2.4, we have a distribution for the first inter-arrival time given by

$$G_{X_1}(t) = 1 - e^{-\Lambda(t)}, \quad \forall t \geq 0;$$

What are the distributions for the other inter-arrival times? At first, in order to make $G_{X_1}(t)$ properly normalized in the time interval $[0, \infty)$, we should add some limit on the intensity function $\lambda(t)$ as follows

$$\Lambda(\infty) = +\infty$$

Theorem 2.10 *Assume that the conditions in Theorem 2.4 hold, then we have*

$$G_{S_1}(t) = P(S_1 \leq t) = G_{X_1}(t) = 1 - e^{-\Lambda(t)}, \quad \forall t \geq 0;$$

$$G_{S_i}(t) = P(S_i \leq t) = 1 - \sum_{j=0}^{i-1} \frac{\Lambda^j(t)}{j!} e^{-\Lambda(t)}, \quad \forall t \geq 0; \quad i = 2, 3, \cdots;$$

$$G_{X_2}(t) = P(X_2 \leq t) = 1 - \int_0^\infty \lambda(s) e^{-\Lambda(t+s)} ds, \quad \forall t \geq 0$$

$$G_{X_i}(t) = P(X_i \leq t) = 1 - \int_0^\infty \lambda(s) e^{-\Lambda(t+s)} \frac{\Lambda^{i-2}(s)}{(i-2)!} ds, \quad \forall t \geq 0; \quad i = 2, 3, \cdots$$

Proof: For $i = 2, 3, \cdots$ and $t \geq 0$, we have

$$\begin{aligned} 1 - G_{S_i}(t) &= P(S_i > t) \\ &= P(N(t) = 0 \text{ or } N(t) = 1 \text{ or } \cdots \text{ or } N(t) = i-1) \\ &= \textstyle\sum_{j=0}^{i-1} P(N(t) = j) \\ &= \textstyle\sum_{j=0}^{i-1} \frac{\Lambda^j(t)}{j!} e^{-\Lambda(t)} \end{aligned}$$

Therefore,

$$G_{S_i}(t) = 1 - \sum_{j=0}^{i-1} \frac{\Lambda^j(t)}{j!} e^{-\Lambda(t)}$$

$$\begin{aligned} G_{X_2}(t) &= P(X_2 \leq t) \\ &= \textstyle\int_0^\infty P(X_2 \leq t | X_1 = s) dG_{X_1}(s) \\ &= \textstyle\int_0^\infty (1 - e^{-[\Lambda(t+s) - \Lambda(s)]}) \lambda(s) e^{-\Lambda(s)} ds \\ &= 1 - \textstyle\int_0^\infty \lambda(s) e^{-\Lambda(t+s)} ds \end{aligned}$$

For $i = 2, 3, \cdots$ and $t \geq 0$, we have

$$\begin{aligned} G_{X_i}(t) &= P(X_i \leq t) \\ &= \int_0^\infty P(X_i \leq t | S_{i-1} = s) dG_{S_{i-1}}(s) \\ &= \int_0^\infty (1 - e^{-[\Lambda(t+s)-\Lambda(s)]}) \lambda(s) e^{-\Lambda(s)} \frac{\Lambda^{i-2}(s)}{(i-2)!} ds \\ &= 1 - \int_0^\infty \lambda(s) e^{-\Lambda(t+s)} \frac{\Lambda^{i-2}(s)}{(i-2)!} ds \end{aligned}$$

□

Considering the Theorems 2.4-2.9 together, we can easily derive the following strong conditional independent property for NHPP.

Theorem 2.11 *Assume that NHPP N, X_i and S_n are given in Theorem 2.6. Denote by $R_s^{n+1} = S_{n+1} - s$ the residual life random variable from the current time s inside the $(n+1)$-th inter-arrival time. Denote by $C_s^{n+1} = s - S_n$ the current life random variable from the n-th jump point to the current time s. Then,*

$$\begin{aligned} & P(R_s^{n+1} > t | S_{n+1} > s, C_s^{n+1} = \tau, X_n = t_n, \cdots, X_1 = t_1) \\ = \; & P(R_s^{n+1} > t | S_{n+1} > s) \\ = \; & e^{-[\Lambda(t+s)-\Lambda(s)]} \end{aligned}$$

where $t \geq 0, 0 < \tau < s, 0 < t_i < s, i = 1, 2, \cdots, n, \sum_{i=1}^n t_i + \tau = s$.

According to Brémaud (1981, P25, T5), we have the following theorem.

Theorem 2.12 *Given the probability space $(\Omega, \mathcal{F}, P : F = \{\mathcal{F}_t, t \geq 0\})$ where $\{\mathcal{F}_t, t \geq 0\}$ is a natural filtration generated by the stochastic process $N = \{N_t : t \geq 0\}$. Assume that N is the NHPP with rate $\lambda(t)$, then it has the following Doob-Meyer decomposition*

$$N(t) = \int_0^t \lambda(s) ds + M_N(t) = \Lambda(t) + M_N(t)$$

where $M_N(t)$ is an F-martingale. In addition, $N(t) - N(s)$ is independent of $\mathcal{F}_s$ provided $0 \leq s \leq t$.

Now let us apply Theorem 2.12 to the NHPP N. For all $0 \leq s \leq t$, we have

$$\begin{aligned} 0 &= E[M_N(t) - M_N(s)|\mathcal{F}_s] \\ &= E[[N(t) - N(s)] - [\Lambda(t) - \Lambda(s)]|\mathcal{F}_s] \\ &= E[N(t) - N(s)] - [\Lambda(t) - \Lambda(s)] \end{aligned} \tag{2.4}$$

From (2.4), we get

$$E[N(t) - N(s)] = \Lambda(t) - \Lambda(s)$$

which coincides with the equation in (2.1).

Chapter 3

Infinite-server queues

3.1 A Kolmogorov-type equation for the $M(t)/M(t)/\infty$ queue

Customers in the $M(t)/M(t)/\infty$ queue arrive according to a nonhomogeneous Poisson process with time-varying rate $\lambda(t)$. There are infinite servers with the service time following the exponential distribution with time-varying rate $\mu(t)$. In this subsection, we derive a Kolmogorov forward equation (KFE) using martingale methods (Abramov, 2006; 2007; Brémaud, 1981).

Let $\{Q(t) : t \geq 0\}$ denote a queue-length process of the $M(t)/M(t)/\infty$ queue. For all $t \geq s \geq 0$, the flow equation for $Q(t)$ can be written as,

$$Q(t) = Q(s) + A(t) - A(s) - \int_s^t \sum_{i=1}^{\infty} I\{Q(\tau-) \geq i\} d\pi_i(\tau), \quad \forall t \geq s \geq 0, \tag{3.1}$$

or equivalently,

$$dQ(t) = dA(t) - \sum_{i=1}^{\infty} I\{Q(t-) \geq i\} d\pi_i(t), \tag{3.2}$$

where A and $\pi_i, i = 1, 2, \cdots$, are independent Poisson processes with rate $\lambda(t)$ and $\mu(t)$, respectively, which we assume to have no common jumps. For all $t \geq s \geq 0$, let

$$I_{ij}(s,t) = I\{Q(t) = j | Q(s) = i\}, \quad i, j = 0, 1, \cdots$$

be the indicator function of the queue length process with the convention that

$$I_{ij}(s,t) \equiv 0, \quad \text{if } i = -1 \text{ or } j = -1.$$

Denote the transition probability matrix of the queue length process $p_{ij}(s,t)$ by

$$\mathbf{P}(s,t) = (p_{ij}(s,t))_{i,j=0,1,\cdots},$$

where

$$p_{ij}(s,t) = P\{Q(t) = j | Q(s) = i\}, \qquad t \geq s \geq 0,$$

with the convention that

$$p_{ij}(s,t) \equiv 0, \quad \text{if } i = -1 \text{ or } j = -1.$$

Working with the indicator function $I_{ij}(s,t)$, we obtain a difference differential equation as follows:

$$\begin{aligned}
\triangle I_{ij}(s,t) = & \; I_{ij}(s,t) - I_{ij}(s,t-) \\
= & \; I\{Q(t) = j | Q(s) = i\} - I\{Q(t-) = j | Q(s) = i\} \\
= & \; I\{Q(t-) = j-1 | Q(s) = i\} I\{Q(t) = j | Q(t-) = j-1, Q(s) = i\} \\
& + I\{Q(t-) = j+1 | Q(s) = i\} I\{Q(t) = j | Q(t-) = j+1, Q(s) = i\} \\
& + I\{Q(t-) = j | Q(s) = i\} I\{Q(t) = j | Q(t-) = j, Q(s) = i\} \\
& - I\{Q(t-) = j | Q(s) = i\} \\
\doteq & \; I\{Q(t-) = j-1 | Q(s) = i\} I\{Q(t) = j | Q(t-) = j-1\} \\
& + I\{Q(t-) = j+1 | Q(s) = i\} I\{Q(t) = j | Q(t-) = j+1\} \\
& + I\{Q(t-) = j | Q(s) = i\} I\{Q(t) = j | Q(t-) = j\} \\
& - I\{Q(t-) = j | Q(s) = i\} \\
= & \; I\{Q(t-) = j-1 | Q(s) = i\} \triangle A(t) \\
& + I\{Q(t-) = j+1 | Q(s) = i\} \triangle \Pi_{j+1}(t) \\
& + I\{Q(t-) = j | Q(s) = i\} [1 - \triangle A(t) - \triangle \Pi_j(t)] \\
& - I\{Q(t-) = j | Q(s) = i\} \\
= & \; I_{ij-1}(s,t-) \triangle A(t) + I_{ij+1}(s,t-) \triangle \Pi_{j+1}(t) \\
& - I_{ij}(s,t-) \triangle A(t) - I_{ij}(s,t-) \triangle \Pi_j(t),
\end{aligned} \tag{3.3}$$

where $\doteq$ means equal in distribution and $\Pi_j(t) = \sum_{i=1}^{j} \pi_i(t)$. Integrating both sides of (3.3) from s to t and using the Doob-Meyer decomposition we have

$$\begin{aligned}
I_{ij}(s,t) \doteq & \; I_{ij}(s,s) + \int_s^t I_{ij-1}(s,\tau-) dA(\tau) + \int_s^t I_{ij+1}(s,\tau-) d\Pi_{j+1}(\tau) \\
& - \int_s^t I_{ij}(s,\tau-) dA(\tau) - \int_s^t I_{ij}(s,\tau-) d\Pi_j(\tau) \\
= & \; I_{ij}(s,s) + \int_s^t I_{ij-1}(s,\tau) \lambda(\tau) d\tau + (j+1) \int_s^t I_{ij+1}(s,\tau) \mu(\tau) d\tau \\
& - \int_s^t I_{ij}(s,\tau) \lambda(\tau) d\tau - j \int_s^t I_{ij}(s,\tau) \mu(\tau) d\tau + M_{ij}(s,t),
\end{aligned} \tag{3.4}$$

where $\{M_{ij}(s,t) : t \geq s \geq 0\}$ is martingale for each $i, j \geq 0$ with

$$\begin{aligned} M_{ij}(s,t) = \; & \int_s^t [I_{ij-1}(s,\tau-) - I_{ij}(s,\tau-)][dA(\tau) - \lambda(\tau)d\tau] \\ & + \int_s^t I_{ij+1}(s,\tau-)[d\Pi_{j+1}(\tau) - (j+1)\mu(\tau)d\tau] \\ & - \int_s^t I_{ij}(s,\tau-)[d\Pi_j(\tau) - j\mu(\tau)d\tau]. \end{aligned}$$

Noting that

$$E[M_{ij}(s,t)] = 0, \quad \forall t \geq s \geq 0, \quad \forall i, j \geq 0,$$

and taking expectation on both sides of (3.4) and using the martingale property of M_{ij}, we obtain

$$\begin{aligned} p_{ij}(s,t) = \; & p_{ij}(s,s) + \int_s^t p_{ij-1}(s,\tau)\lambda(\tau)d\tau + (j+1)\int_s^t p_{ij+1}(s,\tau)\mu(\tau)d\tau \\ & - \int_s^t p_{ij}(s,\tau)\lambda(\tau)d\tau - j\int_s^t p_{ij}(s,\tau)\mu(\tau)d\tau, \end{aligned}$$

and hence

$$\frac{\partial p_{ij}(s,t)}{\partial t} = \lambda(t)p_{ij-1}(s,t) + (j+1)\mu(t)p_{ij+1}(s,t) - \lambda(t)p_{ij}(s,t) - j\mu(t)p_{ij}(s,t). \tag{3.5}$$

From (3.5), using matrix notation, we obtain KFE as follows,

$$\frac{\partial \mathbf{P}(s,t)}{\partial t} = \mathbf{P}(s,t)\mathbf{Q}(t), \quad \mathbf{P}(s,s) = \mathbf{I}, \tag{3.6}$$

where $\mathbf{I}$ is the identity matrix and the time dependent generator of the queue length process $Q(t)$ is

$$\mathbf{Q}(t) = \begin{pmatrix} -\lambda(t) & \lambda(t) & & & & & & \\ \mu(t) & -\lambda(t)-\mu(t) & \lambda(t) & & & & & \\ & 2\mu(t) & -\lambda(t)-2\mu(t) & \lambda(t) & & & & \\ & & \ddots & \ddots & \ddots & & & \\ & & & n\mu(t) & -\lambda(t)-n\mu(t) & \lambda(t) & & \\ & & & & \ddots & \ddots & \ddots & \\ & & & & & \ddots & \ddots & \end{pmatrix}.$$

For the $M/M/\infty$ queue, the KFE reduces to

$$\frac{\partial \mathbf{P}(s,t)}{\partial t} = \mathbf{P}(s,t)\mathbf{Q}, \quad \mathbf{P}(s,s) = \mathbf{I}, \tag{3.7}$$

where

$$\mathbf{Q} = \begin{pmatrix} -\lambda & \lambda & & & & & & \\ \mu & -\lambda-\mu & \lambda & & & & & \\ & 2\mu & -\lambda-2\mu & \lambda & & & & \\ & & \ddots & \ddots & \ddots & & & \\ & & & n\mu & -\lambda-n\mu & \lambda & & \\ & & & & \ddots & \ddots & \ddots & \\ & & & & & \ddots & \ddots & \end{pmatrix}.$$

As is well known in the literature, we can also obtain a KBE for the $M/M/\infty$ queue exploiting the Markov and homogenous properties. This can be accomplished with the following modifications to the equation (3.3).

$$\begin{aligned} \triangle I_{ij}(s,t) = & \; I_{ij}(s,t) - I_{ij}(s,t-) \\ = & \; I\{Q(t)=j|Q(s)=i\} - I\{Q(t-)=j|Q(s)=i\} \\ \doteq & \; \triangle A(s+t-(t-))I\{Q(t)=j|Q(s+t-(t-))=i+1\} \\ & +\triangle \Pi_i(s+t-(t-))I\{Q(t)=j|Q(s+t-(t-))=i-1\} \\ & +[1-\triangle A(s+t-(t-))-\triangle \Pi_i(s+t-(t-))] \\ & I\{Q(t)=j|Q(s+t-(t-))=i\} \\ & -I\{Q(t-)=j|Q(s)=i\} \\ \doteq & \; \triangle A(t)I\{Q(t-)=j|Q(s)=i+1\} \\ & +\triangle \Pi_i(t)I\{Q(t-)=j|Q(s)=i-1\} \\ & +[1-\triangle A(t)-\triangle \Pi_i(t)]I\{Q(t-)=j|Q(s)=i\} \\ & -I\{Q(t-)=j|Q(s)=i\} \\ = & \; \triangle A(t)I\{Q(t-)=j|Q(s)=i+1\} \\ & +\triangle \Pi_i(t)I\{Q(t-)=j|Q(s)=i-1\} \\ & -\triangle A(t)I\{Q(t-)=j|Q(s)=i\} \\ & -\triangle \Pi_i(t)I\{Q(t-)=j|Q(s)=i\} \\ = & \; \triangle A(t)I_{i+1j}(s,t-) + \triangle \Pi_i(t)I_{i-1j}(s,t-) \\ & -\triangle A(t)I_{ij}(s,t-) - \triangle \Pi_i(t)I_{ij}(s,t-). \end{aligned} \tag{3.8}$$

Notice that the Markov property and the homogeneous property of Q and Poisson processes have been explicitly applied in (3.8). Continuing with the general procedure outlined above, we obtain the KBE for the $M/M/\infty$ queue as

$$\frac{\partial \mathbf{P}(s,t)}{\partial t} = \mathbf{Q}\mathbf{P}(s,t), \quad \mathbf{P}(s,s) = \mathbf{I}. \tag{3.9}$$

We also note that the KBE does not exist for the $M(t)/M(t)/\infty$ queue, due to the loss of homogeneous property. We will develop other versions of Kolmogorov equations for this queue in the next subsection below.

3.2 Generalized equation for the $M(t)/M(t)/\infty$ queue

In this subsection, we consider system dynamics in terms of time $s < t$ instead of time t. Following the procedure established in the previous subsection, we develop a variation of the Kolmogorov equation called a Generalized Backward Equation (GBE). Using an indicator function $I_{ij}(s,t)$ where $t \geq s \geq 0, i, j = 0, 1, \cdots$, we obtain

$$\begin{aligned}
\triangle I_{ij}(s,t) = & \; I_{ij}(s,t) - I_{ij}(s-,t) \\
= & \; I\{Q(t) = j | Q(s) = i\} - I\{Q(t) = j | Q(s-) = i\} \\
= & \; I_{ij}(s,t) - [I\{Q(t) = j, Q(s) = i+1 | Q(s-) = i\} \\
& + I\{Q(t) = j, Q(s) = i | Q(s-) = i\} \\
& + I\{Q(t) = j, Q(s) = i-1 | Q(s-) = i\}] \\
\doteq & \; I_{ij}(s,t) - [I\{Q(t) = j | Q(s) = i+1\} I\{Q(s) = i+1 | Q(s-) = i\} \\
& + I\{Q(t) = j | Q(s) = i\} I\{Q(s) = i | Q(s-) = i\} \\
& + I\{Q(t) = j | Q(s) = i-1\} I\{Q(s) = i-1 | Q(s-) = i\}] \\
= & \; I_{ij}(s,t) - [I_{i+1j}(s,t) \triangle A(s) + I_{i-1j}(s,t) \triangle \Pi_i(s) \\
& + I_{ij}(s,t)(1 - \triangle A(s) - \triangle \Pi_i(s))] \\
= & \; I_{ij}(s,t) \triangle A(s) - I_{i+1j}(s,t) \triangle A(s) \\
& + I_{ij}(s,t) \triangle \Pi_i(s) - I_{i-1j}(s,t) \triangle \Pi_i(s).
\end{aligned}$$

A major difference between this equation and the equation in (3.3) used to derive the KFE is the way in which we decompose events and the events we condition on. For example, in (3.3), $I_{ij}(s,t)$ is decomposed into three events conditioned on the queue length process at time $t-$, i.e., $Q(t-)$ which results in differentials in terms of t (from $t-$ to t). While in the current equation, $I_{ij}(s-,t)$ is decomposed into three events conditioned on the queue length process at time s, i.e., $Q(s)$ which results in differentials in terms of s (from $s-$ to s). Note that there is a minus sign on the right hand side of the GBE given in (3.10).

Now continuing with the general procedure, the GBE for the $M(t)/M(t)/\infty$ queue can be written as

$$\frac{\partial \mathbf{P}(s,t)}{\partial s} = -\mathbf{Q}(t)\mathbf{P}(s,t), \quad \mathbf{P}(t,t) = \mathbf{I}, \tag{3.10}$$

where $\mathbf{Q}(t)$ is given in (3.6).

Remark 3.1 *See Feller (1940) where the proof of the existence of solutions to the GBE on a general state space was provided.*

For the $M/M/\infty$ queue, the GBE reduces to

$$\frac{\partial \mathbf{P}(s,t)}{\partial s} = -\mathbf{Q}\mathbf{P}(s,t), \quad \mathbf{P}(t,t) = \mathbf{I},$$

where $\mathbf{Q}$ is given in (3.7). A similar argument yields the following Generalized Forward Equation (GFE) for the $M/M/\infty$ queue, but such an equation does not exist for the $M(t)/M(t)/\infty$ queue.

$$\frac{\partial \mathbf{P}(s,t)}{\partial s} = -\mathbf{P}(s,t)\mathbf{Q}, \quad \mathbf{P}(t,t) = \mathbf{I}.$$

3.3 Main theorems for the $M(t)/M(t)/\infty$ queue

Now we summarize the results for the $M(t)/M(t)/\infty$ queue given in Theorem 3.2.

Theorem 3.2 *For the $M(t)/M(t)/\infty$ queue with the arrival rate $\lambda(t)$ and service rate $\mu(t)$, the KFE and GBE are as follows*

$$\textbf{(KFE)} \qquad \frac{\partial \mathbf{P}(s,t)}{\partial t} = \mathbf{P}(s,t)\mathbf{Q}(t), \quad \mathbf{P}(s,s) = \mathbf{I}; \tag{3.11}$$

$$\textbf{(GBE)} \qquad \frac{\partial \mathbf{P}(s,t)}{\partial s} = -\mathbf{Q}(s)\mathbf{P}(s,t), \quad \mathbf{P}(t,t) = \mathbf{I}, \tag{3.12}$$

which have the same solution

$$p_{ij}(s,t) = \sum_{k=0}^{j} \left(\begin{array}{c} i \\ j-k \end{array} \right) \frac{e^{-r_1 p_1} p_1^j q_1^{i-j+k} r_1^k}{k!}, \tag{3.13}$$

where $p_1 = e^{-\int_s^t \mu(\tau)d\tau}$, $q_1 = 1 - p_1$, $r_1 = \int_s^t \lambda(\tau) e^{\int_s^\tau \mu(l)dl} d\tau$ and $\mathbf{Q}(t)$ is given in (3.6).

Proof: We only need to prove the second part of this theorem. For the KFE, we adopt a generating function method. Define the generating function as

$$G(z,s,t) = \sum_{j=0}^{\infty} z^j p_{ij}(s,t).$$

Using this, the equation in (3.11) reduces to the following partial differential equation (PDE)

$$\frac{\partial G}{\partial t} - \mu(t)(1-z)\frac{\partial G}{\partial z} = \lambda(t)(z-1)G.$$

Using the boundary condition $G(z,s,s) = z^i$, the solution of the PDE can be written as

$$G(z,s,t) = e^{-r_1 p_1 (1-z)}(q_1 + p_1 z)^i,$$

where $p_1 = e^{-\int_s^t \mu(\tau)d\tau}$, $q_1 = 1 - p_1$ and $r_1 = \int_s^t \lambda(\tau) e^{\int_s^\tau \mu(l)dl} d\tau$.

The transient transition probabilities of $Q(t)$ can be obtained by inversion, and are given by

$$p_{ij}(s,t) = \sum_{k=0}^{j} \left(\begin{array}{c} i \\ j-k \end{array} \right) \frac{e^{-r_1 p_1} p_1^j q_1^{i-j+k} r_1^k}{k!}.$$

The proof for the GBE is similar to this one.

Corollary 3.3 *For the $M/M/\infty$ queue with a constant arrival rate λ and service rate μ, the following four kinds of Kolmogorov type equations summarize the dynamics of the system in terms of its transition probability functions.*

$$\textbf{(KFE)} \qquad \frac{\partial \mathbf{P}(s,t)}{\partial t} = \mathbf{P}(s,t)\mathbf{Q}, \quad \mathbf{P}(s,s) = \mathbf{I}; \tag{3.14}$$

$$\textbf{(KBE)} \qquad \frac{\partial \mathbf{P}(s,t)}{\partial t} = \mathbf{Q}\mathbf{P}(s,t), \quad \mathbf{P}(s,s) = \mathbf{I}; \tag{3.15}$$

$$\textbf{(GFE)} \qquad \frac{\partial \mathbf{P}(s,t)}{\partial s} = -\mathbf{P}(s,t)\mathbf{Q}, \quad \mathbf{P}(t,t) = \mathbf{I}; \tag{3.16}$$

$$\textbf{(GBE)} \qquad \frac{\partial \mathbf{P}(s,t)}{\partial s} = -\mathbf{Q}\mathbf{P}(s,t), \quad \mathbf{P}(t,t) = \mathbf{I}, \tag{3.17}$$

all of which have the same solution as given by

$$p_{ij}(s,t) = \sum_{k=0}^{j} \left(\begin{array}{c} i \\ j-k \end{array} \right) \frac{e^{-\lambda q_2/\mu} p_2^{j-k} q_2^{i-j+2k}}{k!} \left(\frac{\lambda}{\mu}\right)^k, \tag{3.18}$$

where $p_2 = e^{-\mu(t-s)}$ and $q_2 = 1 - p_2$.

As we mentioned earlier, for the $M(t)/M(t)/\infty$ queue the corresponding KBE and GFE do not exist. However, assuming that such equations existed, and we call them as quasi-KBE (QKBE) and quasi-GFE (QGFE), we can derive explicit solutions to these equations and summarize them in the next theorem.

Theorem 3.4 *Assuming that the following equations existed for the $M(t)/M(t)/\infty$ queue,*

$$\textbf{(QKBE)} \qquad \frac{\partial \mathbf{P}(s,t)}{\partial t} = \mathbf{Q}(t)\mathbf{P}(s,t), \quad \mathbf{P}(s,s) = \mathbf{I}; \tag{3.19}$$

$$\textbf{(QGFE)} \qquad \frac{\partial \mathbf{P}(s,t)}{\partial s} = -\mathbf{P}(s,t)\mathbf{Q}(s), \quad \mathbf{P}(t,t) = \mathbf{I}. \tag{3.20}$$

Both equations admit the following solution

$$p_{ij}(s,t) = \sum_{k=0}^{j} \left(\begin{array}{c} i \\ j-k \end{array} \right) \frac{e^{-r_2 p_1} p_1^j q_1^{i-j+k} r_2^k}{k!}, \tag{3.21}$$

where p_1, q_1 are given in (3.13) and $r_2 = \int_s^t \lambda(\tau) e^{\int_\tau^t \mu(l)dl} d\tau$.

Note that the only difference between solutions (3.13) and (3.21) is the parameter r (r_1 in (3.13) and r_2 in (3.21)). Next, we will explain why KFE and GFE (KBE and GBE, respectively) have the same solution for the $M/M/\infty$ queue in two different ways.

Remark 3.5 *Firstly, utilizing the homogeneous property for the $M/M/\infty$ queue, we note that*

$$\begin{aligned} p_{ij}(s,t) &= P(Q(t) = j | Q(s) = i) \\ &= P(Q(T+t) = j | Q(T+s) = i) \\ &= P(Q(T+t-s) = j | Q(T) = i) \\ &= P(Q(T-s) = j | Q(T-t) = i) \\ &= p_{ij}(T-t, T-s), \end{aligned} \tag{3.22}$$

which implies that

$$\mathbf{P}(s,t) = \mathbf{P}(T-t, T-s),$$

where $0 \le s \le t \le T$ for fixed T.

In KFE (3.14), using the above homogeneous property, we can show that

$$\mathbf{P}(s,t)\mathbf{Q} = \mathbf{P}(T-t, T-s)\mathbf{Q} = \frac{\partial \mathbf{P}(T-t, T-s)}{\partial (T-s)} = -\frac{\partial \mathbf{P}(s,t)}{\partial s},$$

which is the GFE (equation (3.16)). Similarly we can obtain the KFE (equation (3.14)) from the GFE (equation(3.16)).

Remark 3.6 *Secondly, the solution in (3.18) implies that*

$$\frac{\partial p_2}{\partial t} = -\frac{\partial p_2}{\partial s}; \quad \frac{\partial q_2}{\partial t} = -\frac{\partial q_2}{\partial s}.$$

Consequently, we have

$$\frac{\partial p_{ij}(s,t)}{\partial t} = -\frac{\partial p_{ij}(s,t)}{\partial s},$$
$$\frac{\partial \mathbf{P}(s,t)}{\partial t} = -\frac{\partial \mathbf{P}(s,t)}{\partial s}.$$

This argument provides an explanation of why KFE and GFE have the same solution for the $M/M/\infty$ queue.

Remark 3.7 *For the $M(t)/M(t)/\infty$ queue, the result in equation (3.22) does not hold due to the lack of a homogeneous property. In addition, notice that for the solutions in (3.13) and (3.21), we obtain*

$$\frac{\partial p_1}{\partial t} = -\mu(t)p_1; \quad \frac{\partial p_1}{\partial s} = \mu(s)p_1;$$
$$\frac{\partial q_1}{\partial t} = \mu(t)p_1; \quad \frac{\partial q_1}{\partial s} = -\mu(s)p_1;$$
$$\frac{\partial r_1}{\partial t} = \lambda(t)p_1^{-1}; \quad \frac{\partial r_1}{\partial s} = -\lambda(s) - \mu(s)r_1;$$
$$\frac{\partial r_2}{\partial t} = \lambda(t) + \mu(t)r_2; \quad \frac{\partial r_2}{\partial s} = -\lambda(s)p_1^{-1},$$

which imply that the equations in (3.19) and (3.20) have the solution given in (3.21) instead of the one given in (3.13). However, for the special case of the $M/M/\infty$ queue,

$$\lambda(t) \equiv \lambda; \quad \mu(t) \equiv \mu,$$

$$r_1 = \int_s^t \lambda e^{\int_s^\tau \mu dl} d\tau = \frac{\lambda}{\mu}(e^{\mu(t-s)} - 1)$$

and

$$r_2 = \int_s^t \lambda e^{\int_\tau^t \mu dl} d\tau = \frac{\lambda}{\mu}(e^{\mu(t-s)} - 1),$$

reduce to the same common value, i.e. $r_1 = r_2$. It is clear that in this case when the arrival and service rates are constant why all of the four types of equations (KFE, KBE, GFE and GBE) exist and all of them have the same solution. When the rates are time dependent this equality does not hold and only the KFE and GBE have the same solution.

3.4 The $M(t)/G(t)/\infty$ queue

In this subsection, we provide an interpretation of the solution to the $M(t)/M(t)/\infty$ queue given in (3.13). With the help of this argument, we can directly write out an explicit transient transition probability for the $M(t)/G(t)/\infty$ queue. But we should notice that, due to the loss of the memoryless property for the $M(t)/G(t)/\infty$ queue, we can only consider the system dynamics starting at time 0 instead of starting at an arbitrary time $s > 0$. Therefore the ensuing analysis in this subsection will always assume $s = 0$.

When $s = 0$, the solution given in (3.13) for the $M(t)/M(t)/\infty$ queue reduces to

$$p_{ij}(t) = \sum_{k=0}^{j} \left(\begin{array}{c} i \\ j-k \end{array} \right) \frac{e^{-r_0} p_0^{j-k} q_0^{i-j+k} r_0^k}{k!}, \tag{3.23}$$

where $p_0 = e^{-\int_0^t \mu(\tau)d\tau}$, $q_0 = 1 - p_0$, $r_0 = \int_0^t \lambda(\tau) e^{-\int_\tau^t \mu(l)} dl d\tau$.

The solution in (3.23) can be expressed as convolution of a Poisson random variable with parameter r_0 and a Binomial random variable with parameters (i, p_0). In the following, we provide an interpretation of this solution. Let S_τ be a random variable representing the service time starting from the time $\tau \geq 0$. Let

$$Q(t) = Q_1(t) + Q_2(t),$$

where $Q_1(t)$ represents the number of remaining customers out of the initial i customers who are still in service at time t and $Q_2(t)$ represents the number of remaining customers out of the new arrivals who are still in service at time t. Given that $Q(0) = i$ which implies that $Q_1(0) = i$ and $Q_2(0) = 0$, in order to have $Q(t) = j$, we should have the following possibilities:

$$Q_1(t) = j - k; \quad Q_2(t) = k, \quad 0 \leq k \leq j.$$

Noting that Q_1 is independent of Q_2, we can write

$$\begin{aligned} p_{ij}(t) &= P(Q(t) = j | Q(0) = i) \\ &= \textstyle\sum_{k=0}^{j} P(Q_1(t) = j-k, Q_2(t) = k | Q_1(0) = i, Q_2(0) = 0) \\ &= \textstyle\sum_{k=0}^{j} P(Q_1(t) = j-k | Q_1(0) = i) P(Q_2(t) = k | Q_2(0) = 0) \\ &= \textstyle\sum_{k=0}^{j} \left(\begin{array}{c} i \\ j-k \end{array} \right) p_0^{j-k} q_0^{i-j+k} \frac{e^{-r_0} r_0^k}{k!}, \end{aligned}$$

where $p_0 = e^{-\int_0^t \mu(\tau)d\tau} = P(S_0 > t)$ is the probability that an initial customer has not completed the service by time t and $q_0 = 1 - p_0 = P(S_0 \leq t)$ is the probability that an

initial customer has completed the service by time t and

$$r_0 = \int_0^t \lambda(\tau) e^{-\int_\tau^t \mu(l)} dl d\tau = \int_0^t \lambda(\tau) P(S_\tau > t - \tau) d\tau.$$

Now $Q_1(t) = j - k$ and $Q_1(0) = i$ imply that there will be $j - k$ customers in the queue and $i - j + k$ customers have completed the service by time t. Therefore the transition probability has the binomial distribution. As for Q_2, we can treat it as an adjusted birth process with a time-dependent rate $\lambda(\tau)P(S_\tau > t - \tau)$ which means that the arriving rate $\lambda(\tau)$ at time τ will be adjusted by the probability of a customer still being in the queue after time $t - \tau$. Therefore, the transition probability of Q_2 at time t follows the Poisson distribution with parameter r_0. Solution (3.18) can be explained in a similar way.

From the above discussion, we can directly write out a transient solution for the $M(t)/G(t)/\infty$ queue since the solution only involves $P(S_\tau > t - \tau)$ and $P(S_\tau \leq t - \tau)$. For the $M(t)/G(t)/\infty$ queue, the arrival rate is $\lambda(t)$, and a service time started at time τ has distribution function $G(\tau, \cdot)$. Then

$$p_{ij}(t) = \sum_{k=0}^{j} \left(\begin{array}{c} i \\ j-k \end{array} \right) p_3^{j-k} q_3^{i-j+k} \frac{e^{-r_3} r_3^k}{k!},$$

where

$$p_3 = P(S_0 > t) = 1 - G(0,t), \quad q_3 = 1 - p_3 = P(S_0 \leq t) = G(0,t),$$
$$r_3 = \int_0^t \lambda(\tau) P(S_\tau > t - \tau) d\tau = \int_0^t \lambda(\tau)(1 - G(\tau, t - \tau)) d\tau.$$

This result is summarized in Theorem 3.8.

Theorem 3.8 *For the $M(t)/G(t)/\infty$ queue with the arrival rate $\lambda(t)$, and the time dependent service time started at time τ having the distribution function $G(\tau, \cdot)$, we have*

$$p_{ij}(t) = \sum_{k=0}^{j} \left(\begin{array}{c} i \\ j-k \end{array} \right) \frac{e^{-r_3} p_3^{j-k} q_3^{i-j+k} r_3^k}{k!},$$

where

$$p_3 = 1 - G(0,t), \quad q_3 = 1 - p_3, \quad r_3 = \int_0^t \lambda(\tau)(1 - G(\tau, t - \tau)) d\tau.$$

Corollary 3.9 *In the special case when the service times are generally distributed (do not depend on time and have a general distribution function), for the $M/G/\infty$ queue with the arrival rate λ and the service time distribution function $G(t)$, we have*

$$p_{ij}(t) = \sum_{k=0}^{j} \left(\begin{array}{c} i \\ j-k \end{array} \right) \frac{e^{-r_4} p_4^{j-k} q_4^{i-j+k} r_4^k}{k!}, \tag{3.24}$$

where

$$p_4 = 1 - G(t), \quad q_4 = 1 - p_4, \quad r_4 = \int_0^t \lambda(1 - G(t-\tau))d\tau.$$

Corollary 3.10 *For the $M(t)/G/\infty$ queue with the time varying arrival rate $\lambda(t)$ and the service time distribution function $G(t)$, we have*

$$p_{ij}(t) = \sum_{k=0}^{j} \left(\begin{array}{c} i \\ j-k \end{array} \right) \frac{e^{-r_5} p_5^{j-k} q_5^{i-j+k} r_5^k}{k!},$$

where

$$p_5 = 1 - G(t), \quad q_5 = 1 - p_5, \quad r_5 = \int_0^t \lambda(\tau)(1 - G(t-\tau))d\tau.$$

Observe that in (3.24), when time t goes into infinity, $p_4 = 0$, $q_4 = 1$ and $r_4 = \lambda E[S]$. Hence,

$$P_j = \lim_{t\to\infty} p_{ij}(t) = \frac{(\lambda E[S])^j}{j!} e^{-\lambda E[S]},$$

which implies that in equilibrium the $M/G/\infty$ queue has the Poisson distribution with the mean $\lambda E[S]$. The Poisson distribution has nothing to do with the starting state i and is insensitive to the distribution of S for fixed mean $E[S]$.

Remark 3.11 *In this subsection, we always assume that all of the i customers start service simultaneously at $t = 0$. Under this condition, the interpretation given here is very intuitive.*

Remark 3.12 *In fact, the result of Corollary 3.9 was proved by Takcs (1962, P.160-161). The method used in that book involved a non-homogeneous Bernoulli splitting of the Poisson arrival process. Similar arguments appear in several standard texts, for example see Gross and Harris (1985) and Kulkarni (1995). Tijms (1986) established a backward differential*

equation to get such result. Eick, Massey and Whitt (1993b) derived a similar result. In fact, assuming that our $M(t)/G/\infty$ system started empty in the distant past, i.e., at $t = -\infty$, we can obtain their result via the following equation

$$r_5 = \int_{-\infty}^{t} \lambda(\tau)(1 - G(t-\tau))d\tau = E[\lambda(t - S_e)]E[S] = m(t),$$

where $m(t)$ is the mean of the queue-length process at time t.

3.5 The mean and variance of the queue-length process

In this subsection, as before we will assume that the system starts at time $s = 0$.

For the $M(t)/M(t)/\infty$ queue, taking expectation on both sides of equation (3.1) when $s = 0$, we have

$$E[Q(t)] = E[Q(0)] + \int_0^t \lambda(s)ds - \int_0^t \mu(s)E[Q(s)]ds, \quad \forall t \geq 0, \tag{3.25}$$

which has the following solution

$$E[Q(t)] = e^{-\int_0^t \mu(s)ds} \left(E[Q(0)] + \int_0^t \lambda(s) e^{\int_0^s \mu(\tau)d\tau} ds \right).$$

Remark 3.13 *The expression given in (3.25) can also be interpreted as a Kolmogorov forward equation. Please see Robert (2003, P.361) where we let $f(x) = x$.*

It is not easy to calculate the variance of the queue length process using the flow equation in (3.1). In addition, for the $M(t)/G(t)/\infty$ queue, we do not have the corresponding flow equation (similar to (3.1)) for the queue length process. Here we use the decomposition of the queue length process employed earlier to obtain expressions for the variance. Recall that

$$Q(t) = Q_1(t) + Q_2(t), \tag{3.26}$$

where Q_1 is independent of Q_2. It is immediate that

$$E[Q(t)] = E[Q_1(t)] + E[Q_2(t)] \tag{3.27}$$

and

$$Var[Q(t)] = Var[Q_1(t)] + Var[Q_2(t)]. \tag{3.28}$$

As we have noted earlier, Q_1 is binomially distributed and Q_2 is distributed as a Poisson random variable. It is well known that if a random variable X follows a binomial distribution $B(i,p)$ then

$$E[X] = ip, Var[X] = ip(1-p), \tag{3.29}$$

and for random variable Y with a Poisson distribution $Pois(\lambda)$, we have

$$E[Y] = Var[Y] = \lambda. \tag{3.30}$$

Note that

$$\begin{aligned} E[Q_1(t)] &= \sum_{j=0}^{\infty} jP(Q_1(t)=j) \\ &= \sum_{j=0}^{\infty} j \sum_{i=0}^{\infty} P(Q_1(t)=j|Q_1(0)=i)P(Q_1(0)=i) \\ &= \sum_{i=0}^{\infty} P(Q_1(0)=i) \sum_{j=0}^{\infty} jP(Q_1(t)=j|Q_1(0)=i) \\ &= \sum_{i=0}^{\infty} P(Q_1(0)=i)ip \\ &= pE[Q_1(0)], \end{aligned} \tag{3.31}$$

$$\begin{aligned} E[Q_1^2(t)] &= \sum_{j=0}^{\infty} j^2P(Q_1(t)=j) \\ &= \sum_{j=0}^{\infty} j^2 \sum_{i=0}^{\infty} P(Q_1(t)=j|Q_1(0)=i)P(Q_1(0)=i) \\ &= \sum_{i=0}^{\infty} P(Q_1(0)=i) \sum_{j=0}^{\infty} j^2P(Q_1(t)=j|Q_1(0)=i) \\ &= \sum_{i=0}^{\infty} P(Q_1(0)=i)[ipq + (ip)^2] \\ &= pqE[Q_1(0)] + p^2E[Q_1^2(0)] \end{aligned} \tag{3.32}$$

and

$$\begin{aligned} Var[Q_1(t)] &= E[Q_1^2(t)] - (E[Q_1(t)])^2 \\ &= pqE[Q_1(0)] + p^2Var[Q_1(0)]. \end{aligned} \tag{3.33}$$

Combining (3.26-3.33) with earlier results, we have the following expressions for the variance.

For the $M(t)/M(t)/\infty$ queue,

$$Var[Q(t)] = (1-e^{-\int_0^t \mu(s)ds})e^{-\int_0^t \mu(s)ds}E[Q(0)] + e^{-2\int_0^t \mu(s)ds}Var[Q(0)] + \int_0^t \lambda(s)e^{-\int_s^t \mu(\tau)d\tau}ds.$$

For the $M(t)/G(t)/\infty$ queue,

$$E[Q(t)] = (1-G(0,t))E[Q(0)] + \int_0^t \lambda(s)(1-G(s,t-s))ds$$

and

$$Var[Q(t)] = (1-G(0,t))G(0,t)E[Q(0)]+(1-G(0,t))^2Var[Q(0)]+\int_0^t \lambda(s)(1-G(s,t-s))ds.$$

Remark 3.14 *For the $M/M/\infty$ queue, the mean of the queue length process reduces to*

$$E[Q(t)] = (E[Q(0)] - \lambda/\mu)e^{-\mu t} + \lambda/\mu.$$

We always assume that $0 \leq E[Q(0)] < \infty$. Then

$$E[Q(\infty)] = \lambda/\mu.$$

When $E[Q(0)] = \lambda/\mu$,

$$E[Q(t)] \equiv \lambda/\mu,$$

which implies that the expectation of a queue-length process is a constant as time varies. Its variance is

$$Var[Q(t)] = (1 - e^{-\mu t})e^{-\mu t}E[Q(0)] + e^{-2\mu t}Var[Q(0)] + \frac{\lambda}{\mu}(1 - e^{-\mu t}).$$

Remark 3.15 *For the $M/G/\infty$ queue,*

$$E[Q(t)] = (1 - G(t))E[Q(0)] + \int_0^t \lambda(1 - G(t-s))ds$$

and

$$Var[Q(t)] = (1 - G(t))G(t)E[Q(0)] + (1 - G(t))^2Var[Q(0)] + \int_0^t \lambda(1 - G(t-s))ds.$$

Observe that $E[Q(\infty)] = Var[Q(\infty)] = \lambda E[S]$.

Remark 3.16 *For the $M(t)/G/\infty$ queue,*

$$E[Q(t)] = (1 - G(t))E[Q(0)] + \int_0^t \lambda(s)(1 - G(t-s))ds$$

and

$$Var[Q(t)] = (1 - G(t))G(t)E[Q(0)] + (1 - G(t))^2Var[Q(0)] + \int_0^t \lambda(s)(1 - G(t-s))ds.$$

Chapter 4

Non-homogeneous Poisson process in infinite-server queues

In the previous sections, we have modelled the infinite-server queues with NHPPs for the arriving process and the potential service process. First we focus on the main result Theorem 3.2 for the $M(t)/M(t)/\infty$ queue.

Notice that the solution given in (3.13) is expressed as the convolution of Poisson distribution with parameter $r_1 p_1$ and a Binomial distribution with parameter (i, p_1). According to Zhang and Srinivasan (2013), we can explain its meaning in the following way. Let S_τ denote the random variable of service time after time τ. Denote

$$Q(t) = Q_1(t) + Q_2(t)$$

where Q_1 is the service process with rate $\mu(t)$ which is a pure death process and Q_2 is the same $M(t)/M(t)/\infty$ queue such that

$$Q_1(s) = i; \quad Q_1(t) = j - k; \quad Q_2(s) = 0; \quad Q_2(t) = k; \quad 0 \le k \le j$$

Obviously, Q_1 is independent of Q_2.

$$\begin{aligned} p_{ij}(s,t) &= P(Q(t) = j | Q(s) = i) \\ &= \textstyle\sum_{k=0}^{j} P(Q_1(t) = j-k, Q_2(t) = k | Q_1(s) = i, Q_2(s) = 0) \\ &= \textstyle\sum_{k=0}^{j} P(Q_1(t) = j-k | Q_1(s) = i) P(Q_2(t) = k | Q_2(s) = 0) \\ &= \textstyle\sum_{k=0}^{j} \begin{pmatrix} i \\ j-k \end{pmatrix} p_1^{j-k} q_1^{i-j+k} \cdot \frac{e^{-r_1 p_1} (r_1 p_1)^k}{k!} \\ &= \textstyle\sum_{k=0}^{j} \begin{pmatrix} i \\ j-k \end{pmatrix} \frac{e^{-r_1 p_1} p_1^j q_1^{i-j+k} r_1^k}{k!} \end{aligned}$$

where

$$p_1 = e^{-\int_s^t \mu(\tau)d\tau} = P(S_s > t-s| \text{ s is the current time}) \tag{4.1}$$

which is the probability of a customer still being kept inside the queue after time t-s when the current time is s.

$$q_1 = 1 - p_1 = P(S_s \leq t-s| \text{ s is the current time}) \tag{4.2}$$

which is the probability of a customer leaving the queue after time t-s when the current time is s.

$$r_1 p_1 = \int_s^t \lambda(\tau) e^{-\int_\tau^t \mu(l)dl} d\tau = \int_s^t \lambda(\tau) P(S_\tau > t-\tau | \tau \text{ is the current time}) d\tau \tag{4.3}$$

$Q_1(t) = j-k, Q_1(s) = i$ means that there will be $j-k$ customers being kept inside the queue and $i-j+k$ customers leaving the queue after time $t-s$. Therefore, the transition probability has a binomial distribution. As for Q_2, we can treat it as an adjusted birth process (also NHPP) with a time-dependent rate $\lambda(\tau)P(S_\tau > t-\tau|\tau$ is the current time) which means that the arriving rate $\lambda(\tau)$ at time τ will be adjusted by the probability of a customer being kept inside the queue after time $t-\tau$. Therefore, the transition probability of Q_2 follows a Poisson distribution with parameter $r_1 p_1$ according to Theorem 2.3.

Notice that in (4.1)-(4.3), we have applied Theorem 2.9. For the NHPP Q_1, $Q_1(s) = i$ means that there are i customers being serviced at the current time s. Obviously, there is a different elapsed time for each customer. But the expression given in (4.1) tells us that all of the i customers seem to begin to be serviced at the same time s. This is known as the quasi-memoryless property.

In the following, let us consider the $M/G/\infty$ queue and $M(t)/G/\infty$ queue. The results are presented in Corollaries 3.9 and 3.10, respectively.

Can we suppose to have such results as Theorem (3.2) at any starting time s? The answer is no. Even for $M/G/\infty$ queue, generally, we can not get the following result

$$p_{ij}(s,t) = \sum_{k=0}^{j} \begin{pmatrix} i \\ j-k \end{pmatrix} \frac{e^{-r_4} p_4^{j-k} q_4^{i-j+k} r_4^k}{k!} \tag{4.4}$$

where

$$p_4 = 1 - G(t-s); \quad q_4 = 1 - p_4; \quad r_4 = \int_s^t \lambda(1 - G(t-\tau))d\tau$$

The reason that (4.4) does not hold is that there is a different elapsed time for each customer in the queue at current time s and the service time random variable is not quasi-memoryless generally. In fact, the non-homogeneous Poisson process is the only process which has the quasi-memoryless property.

Chapter 5

Implementation of the infinite-server queue framework and Conclusion

The CCAR stress testing practice requires that we forecast the future PD usually quarterly. Accounts with 90 plus days delinquency or with a bankruptcy indicator are considered as default. The quarterly average default rate is calculated as the ratio of the number of defaults throughout the course of the quarter (3-month default observation window), divided by the total number of performing accounts at the start of the quarter. As such, we denote the time points $t_0, t_1, \ldots$ as the current quarter, the next quarter and so on. As stated before, $\lambda(t), \mu(t)$ denote the arrival and departure intensities respectively. Based on the above definition of quarterly average default rate, the arrivals within the current quarter will not be counted. Then departures within the current quarter can be calculated through two infinite-server queues: $Q(t)$ with $\lambda(t), \mu(t)$ and $Q'(t)$ with $\lambda'(t), \mu(t)$ where $t \in [t_i, t_{i+1}], i = 0, 1, 2, \ldots, Q'(t_i) = Q_(t_i)$ and $\lambda'(t) = 0, t \in [t_i, t_{i+1}]$.

Then the quarterly average default rate at time t_1 can be expressed as

$$PD_{t_{i+1}} = \frac{E[Q(t_i)] - E[Q'(t_{i+1})]}{E[Q(t_i)]}, i = 0, 1, 2, \ldots$$

It is obvious that $PD_{t_{i+1}} \in [0, 1], i = 0, 1, 2, \ldots$ since $0 \leq E[Q'(t_{i+1})] \leq E[Q(t_i)]$.

Here the parameters $\lambda(t), \mu(t)$ can be modelled in terms of the macroeconomic variables and loan level variables, as has been done in the hazard rate models. However this topic is beyond the scope of analysis in this thesis.

This thesis aims at proposing a new theoretical framework for the CCAR stress testing practice, under which the inherent nature of the memoryless property can be investigated

more systematically. In addition, the new framework was shown to incorporate the existing modelling approaches for CCAR practice including the transition matrix method and hazard rate method.

Future research can look into the prepayment and maturity departure events. For example, denoting by $\mu_2(t), \mu_3(t)$ the departure intensities for them respectively, two more departure streams can be modelled by the infinite-server queue theory.

References

[1] Abramov, V.M. (2006). Analysis of multiserver retrial queueing system: A martingale approach and an algorithm of solution, Ann. Oper. Res. 141, 19-50.

[2] Abramov, V.M. (2007). Multiserver queueing systems with retrial and losses, Anziam J. 48, 297-314.

[3] Brémaud, P. (1981). Point Processes and Queues: Martingale Dynamics, Springer.

[4] Cyert, R.M., Davidson, H.J. and Thompson, G.L. (1962). Estimation of allowance for doubtful accounts by Markov chains. Management Science 8, 287-303.

[5] Eick, S., Massey, W.A. and Whitt,W. (1993a). $M_t/G/\infty$ queues with sinusoidal arrival rates, Management Science 39, 241-252.

[6] Eick, S., Massey, W.A. and Whitt,W. (1993b). The physics of the $M_t/G/\infty$ queue, Operations Research 41, 731-742.

[7] Feller, W. (1940). On the Integro-Differential Equations of Purely Discontinuous Markoff Processes. Transactions of the American Mathematical Society 48(3), 488-515.

[8] Gross, D. and Harris, C.M. (1985). Fundamentals of Queueing Theory (2nd ed.). John Wiley & Sons, Inc. New York, USA.

[9] Ho, J. (2001). Modelling bank customers behaviour using data warehouses and incorporating economic indicators. Ph.D. thesis, University of Edinburgh, Edinburgh.

[10] Kulkarni, V.G. (1995). Modelling and Analysis of Stochastic Systems. Texts in Statistical Science Series. Chapman and Hall, Ltd., London.

[11] Malik, M. and Thomas, L.C. (2012). Transition matrix models of consumer credit ratings. International Journal of Forecasting 28, 261-272.

[12] Mandelbaum, A. and Massey, W.A. (1995). Strong approximations for time dependent queues. MOR 20(1), 33-64.

[13] Mandelbaum, A., Massey, W.A. and Reiman, M.I. (1998). Strong approximations for Markovian service networks. Queueing Systems 30, 149-201.

[14] Mandelbaum, A., Massey, W.A., Reiman, M.I. and Rider, B. (1999). Time varying multiserver queues with abandonment and retrials, in: Teletraffic Engineering in a Competitive World, Proc. of the 16th Internat. Teletraffic Congress, 355-364.

[15] Massey, W.A. (2002). The analysis of queues with time-varying rates for telecommunication models. Telecommunication Systems 21(2-4), 173-204.

[16] Narain, B. (1992). Survival analysis and the credit granting decision, In: Thomas L.C., Crook J.N., Edelman D.B. (eds), Credit Scoring and Credit Control, Oxford University Press: Oxford, UK, 109-122.

[17] Philippe, R. (2003). Stochastic Networks and Queues, Springer.

[18] Ross, S.M. (2003). Probability Models, Academic Press, Eighth Edition.

[19] Scallan, G. (1998). Bad debt projection models, an overview of modelling approaches. http://www.scoreplus.com/docs/BadDebt.pdf.

[20] Schniederjans, M.J. and Loch, K.D. (1994). An aid for strategic marketing in the banking industry: a Markov analysis. Computers and Operations Research 21, 281-287.

[21] Stepanova, M. and Thomas, L. (2002). Survival analysis methods for personal loan data. Operations Research 50(2) , 277-289.

[22] Takcs, L. (1962). Introduction to the Theory of Queues. Oxford University Press, New York.

[23] Thomas, L., Banasik, J., and Crook, J. (1999). Not if but when loans default. J. Oper. Res. Soc. 50 , 1185-1190.

[24] Thomas, L.C., Ho, J. and Scherer, W.T. (2001). Time will tell: Behavioural scoring and the dynamics of consumer risk assessment. IMA Journal of Management Mathematics 12, 89-103.

[25] Tijms, H.C. (1986). Stochastic modelling and analysis: A computational approach. Wiley Series in Probability and Mathematical Statistics. Wiley, New York.

[26] Tong, E.N., Mues, C. and Tomas, L.C. (2012). Mixture cure models in credit scoring: If and when borrowers default. European Journal of Operational Research 218, 132-139.

[27] Trench, M.S., Pederson, S.P., Lau, E.T., Lizhi, M., Hui, W., and Nair, S.K. (2003). Managing credit lines for Bank One credit cards. Interfaces, 33(5), 4-22.

[28] Zhang, G. and Srinivasan, R. (2013). Infinite-server queues with time-varying rates. International Journal of Mathematics in Operational Research 5(1): 91-109.