Martingale decomposition for large queue asymptotics

Masakiyo Miyazawa
Department of Information Sciences
Tokyo University of Science

Abstract Asymptotic analyses are one of major topics in the recent queueing theory. We focus on them for large queues, which may be classified into two types. One is large deviations, typically for a fixed model. Another is to study a limit of a sequence of queueing models under appropriate scaling of time, space and/or modeling primitives. Diffusion approximation is such an example. We revisit the unified approach of [8] for both asymptotic problems. Expanding its framework, we discuss more about the martingale which plays a key role in it. Some new examples and conjectures are also presented.

1. Introduction

Asymptotic analyses have been actively studied in the recent queueing theory. This is because queueing models, particularly, queueing networks, become very complicated and their exact analyses are getting harder. We focus on asymptotic analyses for large queues, and aim to understand their limiting behaviors through their modeling primitives.

There are two different types of asymptotic analyses for large queues. One is large deviations, which is typically studied for a fixed model. Among them, we are particularly interested in the tail asymptotic behaviors for the stationary distribution of a queueing process. Another is to study a limit of a sequence of queueing models under appropriate scaling of time, space and/or modeling primitives. This would give a theoretical support for approximation using the limiting model. We call it a weak limit approximation. A diffusion approximation is such an example, which means to approximate the queueing model by a reflecting diffusion process. If the drift and variance of this diffusion process do not depend on their current position when it is apart from the boundary of its state space, then it is called a reflecting Brownian motion.

Those two asymptotic problems have been studied separately in the literature. Recently, the author [8] studied a unified approach which is applicable to both asymptotic analyses. Martingales play a key role there. We revisit this unified approach focusing on those martingales. It is hardly said how they are important, but it at least widens our view for studying the queueing theory.

The unified approach of [8] uses a piecewise deterministic Markov process, PDMP for short, for describing queueing models. Its sample path is composed of two parts, deterministic continuous part and discontinuous part, called jumps, by which randomness is created. This process is widely applicable, but known to be hard to analysis. Because of those facts, PDMP is often used for describing models, but rarely analyzed. So, other methods have been employed for analysis. Discretizing the state space and using Markov chains either in continuous time or in discrete time may be most popular. For the diffusion approximation, it is standard to use a continuous mapping of the input data, which is often obtained as the solution of sample path equations, called Skorohod problem (e.g., see [10]). However, there are some gaps from direct analysis of the PDMP.

Thus, it is interesting to directly study the PDMP. Its time evolution is easily presented by a stochastic integral equation using a test function, which maps the states of the PDMP to real values (see (3.2)). This stochastic equation has no martingale term because of deterministic sample paths between jumps. However, it includes random jump terms. This causes difficulty for analysis. Davis [2] who introduced PDMP replaces them by a martingale, imposing the so called boundary condition on the test function. However, it is not easy to find a class of the test functions which characterize a distribution on the state space. This is important in [2] because it aims to get the stationary distribution of the PDMP. Thus, Davis's approach had not been well developed.

Those situations may be overcome by choosing a smaller class of the test functions which may not characterize a distribution on the state space. This is a basic idea in [8]. One may be afraid that such an approach would discard important information in the stochastic integral equation. This is true for full analysis, but we are interested in large queues under the stationary distribution, and lost information might be not so important for them. As we will see, this is indeed the case for both asymptotic analyses, large deviations and weak limit approximations.

Another difference in the approach of [8] from that in [2] is a role of a martingale. Davis [2] uses it for obtaining an extended generator, which enables to use standard results on Markov processes (see, e.g., [3]). On the other hand, the martingale itself plays important role in [8]. For this, it is crucial to find good test functions and techniques to apply them for asymptotic analyses. Those have been done for a single queue with multiple heterogeneous servers in [8]. A similar idea has been used in [1] for the diffusion approximation of a generalized jackson network, GJN for short.

In this paper, we first revisit the approach of [8], extending it for a more general PDMP. We simplify some arguments by assuming stronger conditions because of space of this paper, but detail more about martingales as far as possible. We also consider some new examples for this unified approach. One is the workload process of the $G/G/1$ queue as a toy model for the main component of the PDMP to be continuously changed. This type of PDMP has not been considered in [1, 8]. Another is the tail asymptotics of the stationary distribution of the joint queue length in the GJN. For this, we discuss how we can use a martingale for them, and present some conjectures.

2. Piecewise Deterministic Markov process, PDMP

We formally introduce a piecewise deterministic Markov process. Let $X(t)$ be the state of this process at time t, then it has two components $Z(t)$ and $C(t)$, namely,

$$X(t) = (Z(t), C(t)), \qquad t \geq 0,$$

where $Z(t)$ is a state to be interested, and $C(t)$ is a time counter for the next jump. As always, $X(t)$ is assumed to be right-continuous and has the left-hand limit at each time t. Its state space has the following structure.

(a) $Z(t)$ takes values in a complete and separable topological space S_1.

(b) For each $Z(t)$, there is a finite set $\mathcal{K}(Z(t))$, and $C(t)$ takes values in $\mathbb{R}_+^{\mathcal{K}(Z(t))}$ which is the set of all vectors $\boldsymbol{y}$ whose entry y_i takes values in $\mathbb{R}_+$ for $i \in \mathcal{K}(Z(t))$, where $\mathbb{R}_+$ is the set of all nonnegative real numbers. The entry of $C(t)$ with index i is denoted by $C_i(t)$.

Typical examples of S_1 are a countable set with discrete topology, the d-dimensional real vector space $\mathbb{R}^d$ with Euclid metric and the set of some measures on $\mathbb{R}$ with certain metric, where $\mathbb{R}$ is the set of all real numbers.

Let $S = \{(z, \boldsymbol{y}); z \in S_1, \boldsymbol{y} \in \mathbb{R}_+^{\mathcal{K}(z)}\}$, which is the state space of $X(t)$. We assume the following dynamics.

(c) $X(t)$ is a continuous deterministic function of t except for jump instants, which are denoted by an increasing sequence $\{t_n; n = 1, 2, \ldots\}$. For convenience, we let $t_0 = 0$.

(d) There is a set $\mathcal{M}(S)$ of continuous functions from S to $\mathbb{R}$ such that

(d1) The distribution of $X(t)$ is determined by $\mathbb{E}(f(X(t))) < \infty$ for $f \in \mathcal{M}(S)$.

(d2) For $t \in (t_{n-1}, t_n)$, $f(X(t))$ has a continuous derivative $\mathcal{A}f(X(t))$ for $f \in \mathcal{M}(S)$, where $\mathcal{A}$ is an operator on $\mathcal{M}(S)$, that is, $\mathcal{A}f \in \mathcal{M}(S)$ for $f \in \mathcal{M}(S)$.

(d3) For $t \in (t_{n-1}, t_n)$, $\mathcal{K}(Z(t))$ is unchanged, and $C_i(t)$ is non-increasing in t for each $i \in \mathcal{K}(Z(t))$.

(e) For $s > 0$ and $n \geq 1$, $t_n > t_{n-1} + s$ if and only if $C_i((t_{n-1} + u)-) > 0$ for all $u \in [0, s]$ and all $i \in \mathcal{K}(Z(t_{n-1}))$, where $C_i(t-) = \lim_{\epsilon \downarrow 0} C_i(t - \epsilon)$.

(f) The conditional distribution of $X(t_n)$ given $\{X(u); u < t_n\}$ is a function of $X(t_n-)$ for $n \geq 1$, which is characterized by the transition kernel Q given below.

$$Qf\big(X(t-)\big) = \mathbb{E}\big(f(X(t))|X(t-)\big), \qquad X(t-) \in \Gamma, f \in \mathcal{M}(S), \tag{2.1}$$

where Γ is the subset of S such that the some entry of $\boldsymbol{y}$ vanishes for $(z, \boldsymbol{y}) \in S$. This Γ is referred to as a terminal set, while Q is referred to as a jump kernel.

Obviously, $\{X(t); t \geq 0\}$ satisfying the conditions (a)–(f) is a Markov process, whose dynamics is specified by $\mathcal{A}$ and Q. It extends the framework of the PDMP in [8]. This process is essentially the same as a piecewise deterministic Markov process, PDMP for short, introduced by Davis [2]. We refer to it as the same name. It is noticed that we exclude jumps generated by the main component $Z(t)$, but they may be included in $C(t)$.

The PDMP covers a wide range of queues and their networks. Here are some examples.

Example 2.1 ($GI/G/1$ queue)**.** *Consider a FCFS single server queue in which customers arrive subject to a renewal process and their service times are independent and identically distributed, where FCFS means first-come first served. This queueing model is named as the $GI/G/1$ queue in the literature. Let $L(t)$ be the number of customers in system at time t. Let $R_e(t)$ and $R_s(t)$ be the residual arrival and service times at time t, where we let $R_s(t) = 0$ for $L(t) = 0$. Then,*

$$X(t) = (L(t), R_e(t), R_s(t)), \qquad t \geq 0,$$

is a PDMP, for which $Z(t) = L(t)$ and $C(t) = (R_e(t), R_s(t))$, and therefore $S_1 = \mathbb{Z}_+$ and $\mathcal{K}(z) = \{e, s\}$ for $z > 0$ and $\mathcal{K}(z) = \{e\}$ for $z = 0$, where $\mathbb{Z}_+$ is the set of all nonnegative integers. By the assumptions, for $f \in \mathcal{M}(S)$, we have

$$\mathcal{A}f(z, \boldsymbol{y}) = -\frac{\partial}{\partial y_e} f(z, \boldsymbol{y}) - \frac{\partial}{\partial y_s} f(z, \boldsymbol{y}) 1(z > 0), \tag{2.2}$$

$$Qf(z, \boldsymbol{y}) = \begin{cases} \mathbb{E}(f(z+1, (T_e, y_s)), & y_e = 0, y_s > 0 \\ \mathbb{E}(f(z-1, (y_e, T_s)), & z \geq 1, y_e > 0, y_s = 0, \\ \mathbb{E}(f(z, (T_e, T_s)), & z \geq 1, y_e = 0, y_s = 0, \\ 0, & \textit{otherwise}, \end{cases} \tag{2.3}$$

where T_e and T_s are random variables subject to the inter-arrival and service times, respectively, which are independent of everything else. Thus, $X(t)$ is a PDMP.

Similarly, let $V(t)$ be the total unfinished work at time t, then it is easy to see that $(V(t), R_e(t))$ is a PDMP. In this case, $S = \mathbb{R}_+^2$, and, for $(z, y) \in S$,

$$\mathcal{A}f(z, y) = -\frac{\partial}{\partial v} f(z, y) 1(z > 0) - \frac{\partial}{\partial y} f(z, y), \tag{2.4}$$

$$Qf(z, y) = \begin{cases} \mathbb{E}(f(z + T_s, T_e)), & y = 0, \\ 0, & \textit{otherwise}. \end{cases} \tag{2.5}$$

■

Example 2.2 (Generalized Jackson network)**.** *Consider a d-node queueing network with single servers at nodes in which exogenous customers arrive at each node subject to a renewal process and service times at each node are independent and identically distributed which are independent of everything else. Each node has an infinite buffer, customers are served in the FCFS manner, and they are routed to the next nodes or leave the network according to a given probability which only depends on the current node when their service completed.*

Let $\mathcal{J} = \{1, 2, \ldots, d\}$, and let $\mathcal{E}$ be the set of nodes which have exogenous arrivals. For time t and node $i \in \mathcal{J}$, let $L_i(t)$ be the number of customers, and let $R_{s,i}(t)$ be the residual service times, respectively, where we set $R_{s,i}(t) = 0$ when $L_i(t) = 0$. For $i \in \mathcal{E}$, let $R_{e,i}(t)$ be the residual time for an exogenous arrival. Let p_{ij} be the probability that a customer completing service at node i is routed to node j for $i, j \in \mathcal{J}$, where those customer leave the outside of the network with probability:

$$p_{i0} \equiv 1 - \sum_{i \in \mathcal{J}} p_{ij}.$$

For each node i, let $F_{e,i}$ be the interarrival time distribution of exogenous customers, and let $F_{s,i}$ be the service time distribution. Denote the vectors whose i-th entries are $L_i(t), R_{e,i}(t)$ for $i \in \mathcal{E}$, $R_{s,i}(t)$ by $\boldsymbol{L}(t), \boldsymbol{R}_e(t), \boldsymbol{R}_s(t)$, respectively, and define $X(t)$ as

$$X(t) = (\boldsymbol{L}(t), \boldsymbol{R}_e(t), \boldsymbol{R}_s(t)), \qquad t \geq 0.$$

Then, $\{p_{ij}; i,j \in \mathcal{J}\}$, $\{F_{e,i}; i \in \mathcal{E}\}$ and $\{F_{s,i}; i \in \mathcal{J}\}$ are the modeling primitives, and it is not hard to see that $X(t)$ is a PDMP. ▮

For some models, we need $C(t)$ to be measure valued. For example, it is required for the network in which customers are classified into multiple classes, each class has their own routing probabilities, and they are served in the FCFS manner irrespective to their classes.

3. Martingale decomposition for PDMP

Let $X(\cdot)$ be a PDMP satisfying the conditions (a)–(f). We consider its evolution in time by a stochastic integral equation. Let $\mathcal{F}_t = \sigma(X(u); u \leq t)$, where $\sigma(\cdot)$ stands for the minimal σ-field. $\{\mathcal{F}_t; t \geq 0\}$ is called a filtration. Then, $X(\cdot)$ is a strong Markov process with respect to $\{\mathcal{F}_t; t \geq 0\}$. Define the counting process $N(\cdot) \equiv \{N(t); t \geq 0\}$ for the jump instants of this PDMP as

$$N(t) = \sum_{n=1}^{\infty} 1(t_i \leq t), \qquad t \geq 0. \tag{3.1}$$

Since $X(t)$ is absolutely continuous in $t \in \mathbb{R}_+$ with respect to the sum of the Lebesque measure and the counting process $N(\cdot)$, we obviously have a stochastic integral equation.

$$f(X(t)) = f(X(0)) + \int_0^t \mathcal{A}f(X(u))du + \int_0^t \Delta f(X(u))dN(u), \quad f \in \mathcal{M}(S), \tag{3.2}$$

where $\Delta f(X(u)) = f(X(u)) - f(X(u-))$. Note that $\Delta N(u) > 0$ if and only if $X(u-) \in \Gamma$, which causes a jump, that is, $X(u-) \neq X(u)$.

Our arguments will be based on the following fact due to Davis [2].

Lemma 3.1 (A version of Proposition 4.3 of Davis [2])**.** *If a Borel-measurable function f from S to $\mathbb{R}$ satisfies that*

$$\mathbb{E}\Big(\int_0^t \big|f(X(u)) - Qf(X(u-))\big|dN(u)\Big) < \infty, \qquad \textit{for each } t > 0, \tag{3.3}$$

then

$$M_Q(\cdot) \equiv \Big\{ \int_0^t \big(f(X(u)) - Qf(X(u-))\big)dN(u); t \geq 0\Big\} \tag{3.4}$$

is an $\mathcal{F}_t$-martingale, that is, $\mathbb{E}(|M_Q(t)|) < \infty$ and $\mathbb{E}(M_Q(t)|\mathcal{F}_s) = M_Q(s)$ for $0 \leq s < t$.

A simple proof of this lemma can be found in [8]. The following fact is immediate from (3.2) and Lemma 3.1.

Lemma 3.2 (A special case of Theorem 5.5 of [2])**.** *For $f \in \mathcal{M}(S)$, if (3.3) holds and if*

$$M(t) \equiv f(X(t)) - f(X(0)) - \Big(\int_0^t \mathcal{A}f(X(u))du + \int_0^t (Qf(X(u-)) - f(X(u-)))dN(u)\Big) \tag{3.5}$$

satisfies that $\mathbb{E}(|M(t)|) < \infty$, then $M(\cdot) \equiv \{M(t); t \geq 0\}$ is an $\mathcal{F}_t$-martingale. In particular, if f satisfies that

$$Qf(\boldsymbol{x}) = f(\boldsymbol{x}), \qquad \forall \boldsymbol{x} \in \Gamma, \tag{3.6}$$

then the $\mathcal{F}_t$-martingale $M(t)$ is simplified to

$$M(t) = f(X(t)) - f(X(0)) - \int_0^t \mathcal{A}f(X(u))du, \qquad t \geq 0. \tag{3.7}$$

Although Davis [2] refers to (3.6) as a boundary condition, we refer to (3.6) as a terminal condition following the terminology of [8]. Note that (3.7) can be written as

$$f(X(t)) = f(X(0)) + \int_0^t \mathcal{A}f(X(u))du + M(t), \qquad t \geq 0. \tag{3.8}$$

Apart from the terminal condition (3.6), this representation is standard for a Markov process which has $\mathcal{A}$ as an extended generator (e.g., see [3]). In the present case, we have a simpler $\mathcal{A}$, but the terminal condition causes difficulty as we discussed in Section 1. So far, we attack this problem by finding a class of test functions satisfying (3.6) which may not characterize the distribution of $X(t)$ but is still useful for asymptotic analysis. We call (3.8) a martingale decomposition.

One might wonder why this martingale decomposition is useful for studying large queues. There are two reasons for this.

1) The martingale term disappears after taking the expectation. This is convenient to derive a stationary equation from (3.8).

2) It enables us to change measures. This is useful to study the asymptotic problems.

These two facts are demonstrated for the queue length of a many server queue in [8]. Although no martingale is used, the fact 1) is essentially used in [1] for the queue length vector of the generalized Jackson network of Example 2.2. In this paper, we consider workload in the $GI/G/1$ queue of Example 2.1, while we derive the martingale of Lemma 3.3 for the generalized Jackson network for discussing the tail asymptotics of the stationary joint queue length distribution.

3.1. Martingales for the workload

Let us consider the workload in the $GI/G/1$ queue. As in Example 2.1, we set $X(t) = (V(t), R_e(t))$, for the workload $V(t)$ and residual arrival time $R_e(t)$ at time t. Thus, the state space S of $X(t)$ is $\mathbb{R}_+^2$. For simplicity, we assume that the distributions of T_e and T_s have light tails. If this is not the case, we need truncation arguments, which works well as shown in [8]. Let $\lambda = 1/\mathbb{E}(T_e)$ and $\mu = 1/\mathbb{E}(T_s)$, and assume that they are positive. In this subsection, we do not need the stability condition that $\rho \equiv \lambda/\mu < 1$, so $X(t)$ could be unstable.

Using parameters $\theta, \eta \in \mathbb{R}$ and $v \in \mathbb{R}_+$, we choose the following test function,

$$f(z, y) = e^{\theta z + \eta y}, \qquad (z, y) \in S. \tag{3.9}$$

Let us consider η for this function f to satisfy (3.6). From (2.5), we choose η as the solution of the following equation:

$$\widehat{F}_s(\theta)\widehat{F}_e(\eta) = 1, \tag{3.10}$$

where $\widehat{F}_e(\eta) = \mathbb{E}(e^{\eta T_e})$ and $\widehat{F}_s(\theta) = \mathbb{E}(e^{\theta T_s})$, then the terminal condition (3.6) is satisfied . By the light tail assumptions, the moment generating functions $\widehat{F}_e(\eta)$ and $\widehat{F}_s(\theta)$ are finite in neighborhoods of the origin for η and θ. $\widehat{F}_e(\eta)$ is increasing in η as long as it is finite. For simplicity, we assume its range is $(0, \infty)$. Then, letting $\theta_* = \sup\{\theta \in \mathbb{R}; \widehat{F}_s(\theta) < \infty\}$. we have

$$\eta = \widehat{F}_e^{-1}\big((\widehat{F}_s(\theta))^{-1}\big), \qquad \theta \in (-\infty, \theta_*).$$

We denote this η by $\eta(\theta)$, and denote the f of (3.9) with this $\eta(\theta)$ for η by f_θ, respectively.

As remarked in [8], the inverse function $\widehat{F}_e^{-1}$ has nice properties. For example, by Theorem 1 of Glynn and Whitt [4] (see also (2.20) of [8]), we have

$$\widehat{F}_e^{-1}(e^{-\theta}) = -\lim_{t\to\infty} \frac{1}{t} \log \mathbb{E}\big(e^{\theta N_e(t)}\big),$$

and therefore, using the notation $U(t) \equiv \sum_{n=1}^{N_e(t)} T_s(n)$, which is an accumulated work brought by customer arriving in the time interval $[0, t]$,

$$\eta(\theta) = -\lim_{t\to\infty} \frac{1}{t} \log \mathbb{E}\big(e^{\theta U(t)}\big), \qquad \theta \in (-\infty, \theta_*). \tag{3.11}$$

Thus, it is not hard to see the following facts similar to Lemmas 2.4 and 2.5 of [8] and using Taylor expansion.

Lemma 3.3. *The terminal condition (3.6) holds for $f = f_\theta$ and $\theta < \theta_*$.*

Lemma 3.4. *(a) $\eta(\theta)$ is decreasing and concave for $\theta < \theta_*$. (b) For each $\delta > 0$, we have*

$$|\eta(\theta)| \leq \max\Big(\rho, \frac{1}{\delta}|\eta(\delta)|\Big)|\theta|, \qquad \forall \theta \in [-\delta, \delta], \tag{3.12}$$

$$\eta(\theta) = -\rho\theta - \frac{1}{2}\lambda(\sigma_s^2 + \rho^2\sigma_e^2)\theta^2 + o(\theta^2), \qquad \textit{as } \theta \to 0. \tag{3.13}$$

A figure for $\eta(\theta)$ is given below.

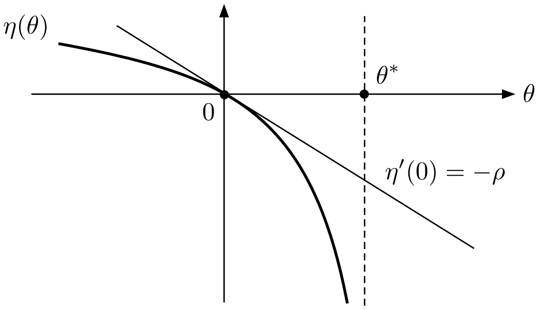

Figure 1: A typical shape for $\eta(\theta)$

Applying the test function f_θ to (3.7), we have the martingale $M_\theta(t)$:

$$\begin{aligned} M_\theta(t) =& f_\theta(V(t), R_e(t)) - f_\theta(V(0), R_e(0)) \\ &+ \int_0^t (\theta + \eta(\theta)) f_\theta(V(u), R_e(u)) du - \int_0^t \theta f_\theta(V(u), R_e(u)) 1(V(u) = 0) du. \end{aligned} \tag{3.14}$$

This martingale will be useful to study a diffusion approximation for the workload process $V(t)$. Let $Y(t)$ be a left-continuous process, which is called predictable because $Y(t-)$ is $\mathcal{F}_{t-}$-measurable. Assume that $Y(t)$ has bounded in each finite interval, then the integral of $Y(t)$ with respect to $M_\theta(t)$, which is denoted by

$$Y \cdot M_\theta(t) \equiv 1 + \int_0^t Y(s) M_\theta(ds)$$

can be well defined as a natural extension of a Riemann-Stieltjes integral for taking step functions for $Y(t)$, and it is a $\mathcal{F}_t$-martingale again (see Section 4d of Chapter I of [5]). Choose $Y(t)$ as

$$Y(t) = \frac{1}{f_\theta(X(0))} \exp\Big(-\int_0^t \frac{\mathcal{A}f_\theta(X(u))}{f_\theta(X(u))} du\Big),$$

which is obviously continuous in t, so $Y \cdot M_\theta(t)$ is martingale. We denote it by E^{f_θ}. Thus, we have

$$\begin{aligned} E^{f_\theta}(t) &= 1 + \int_0^t \frac{1}{f_\theta(X(0))} \exp\Big(-\int_0^s \frac{\mathcal{A}f_\theta(X(u))}{f_\theta(X(u))} du\Big)\Big(df_\theta(X(s)) - \mathcal{A}f_\theta(X(s))ds\Big) \\ &= \frac{f_\theta(X(t))}{f_\theta(X(0))} \exp\Big(-\int_0^t \frac{\mathcal{A}f_\theta(X(u))}{f_\theta(X(u))} du\Big) \\ &= \exp\Big(\theta(V(t) - V(0)) + \eta(\theta)(R_e(t) - R_e(0)) \\ &\qquad + (\theta + \eta(\theta))t - \theta \int_0^t 1(V(u) = 0) du\Big). \end{aligned} \tag{3.15}$$

Let ν be the distribution of $X(0)$, and denote the probability measure under this initial distribution by $\mathbb{P}_\nu$. Since the martingale $E^{f_\theta}(t)$ is positive, we can be used to change $\mathbb{P}_\nu$ to a new probability measure $\widetilde{\mathbb{P}}_\nu^{(\theta)}$ by

$$\frac{d\widetilde{\mathbb{P}}_\nu^{(\theta)}}{d\mathbb{P}_\nu}\Big|_{\mathcal{F}_t} = E^{f_\theta}(t), \qquad t \ge 0.$$

This is called exponential change of measures (see [9]). This implies that

$$d\mathbb{P}_\nu = (E^{f_\theta}(t))^{-1} d\widetilde{\mathbb{P}}_\nu^{(\theta)} \quad \text{on } \mathcal{F}_t, \tag{3.16}$$

which is useful for studying the tail asymptotic of the stationary distribution of $V(t)$.

3.2. Martingales for the generalized Jackson network

We consider the GJN of Example 2.2. In this case, the state space S is given by

$$S = \{(\boldsymbol{z}, \boldsymbol{y}_e, \boldsymbol{y}_s); z \in \mathbb{Z}_+^d, \boldsymbol{y}_e \in \mathbb{R}_+^{\mathcal{E}}, \boldsymbol{y}_s \in \mathbb{R}_+^{\mathcal{K}(z)}\},$$

where $\mathcal{K}(z) = \{i \in \mathcal{J}; z_i \ge 1\}$. Similar to the $GI/G/1$ case, we assume that $F_{e,i}$ and $F_{s,i}$ have light tails for simplicity. Let

$$t_i(\boldsymbol{\theta}) = e^{-\theta_i}\Big(\sum_{j \in \mathcal{J}} p_{ij} e^{\theta_j} + p_{i0}\Big), \qquad \boldsymbol{\theta} \in \mathbb{R}^d, i \in \mathcal{J}.$$

Using parameters $\boldsymbol{\theta}, \boldsymbol{\zeta} \in \mathbb{R}^d$ and $\boldsymbol{\eta} \in \mathbb{R}^{\mathcal{E}}$, we choose the following test function,

$$f_{\boldsymbol{\theta}}(\boldsymbol{z}, \boldsymbol{y}_e, \boldsymbol{y}_s) = w(\boldsymbol{z}, \boldsymbol{\theta}) e^{\langle \boldsymbol{\theta}, \boldsymbol{z}\rangle + \langle \boldsymbol{\eta}(\boldsymbol{\theta}), \boldsymbol{y}_e\rangle + \langle \boldsymbol{\zeta}(\boldsymbol{\theta}), \boldsymbol{y}_s\rangle},$$

where w is given by

$$w(\boldsymbol{z}, \boldsymbol{\theta}) = e^{\sum_{i \in \mathcal{J}} \theta_i 1(z_i = 0)},$$

and $\boldsymbol{\eta}(\boldsymbol{\theta})$ and $\boldsymbol{\zeta}(\boldsymbol{\theta})$ are the solutions of $\boldsymbol{\eta} \equiv \{\eta_i; i \in \mathcal{E}\}$ and $\boldsymbol{\zeta} \equiv \{\zeta_i; i \in \mathcal{J}\}$, respectively, of the following equations.

$$e^{\theta_i} \widehat{F}_{e,i}(\eta_i) = 1, \quad i \in \mathcal{E}, \qquad t_i(\boldsymbol{\theta}) \widehat{F}_{s,i}(\zeta_i) = 1, \quad i \in \mathcal{J}.$$

Because the light tail assumption, $\boldsymbol{\eta}(\boldsymbol{\theta})$ and $\boldsymbol{\zeta}(\boldsymbol{\theta})$ are well defined at least in some neighborhood of the origin of their vector spaces. Then, the terminal condition (3.6) is satisfied, and similar results in Lemma 3.4 hold (see Section 4 of [1] for details). For convenience, we define $\gamma(\boldsymbol{\theta})$ as

$$\gamma(\boldsymbol{\theta}) = -\sum_{i \in \mathcal{E}} \eta_i(\theta_i) - \sum_{i \in \mathcal{J}} \zeta_i(\boldsymbol{\theta}), \qquad \boldsymbol{\theta} \in \mathbb{R}^d.$$

Thus, similar to the $GI/G/1$ case in Section 3.1 work, we have the following martingales,

$$\begin{aligned} M_{\boldsymbol{\theta}}(t) \equiv{}& f_{\boldsymbol{\theta}}(\boldsymbol{L}(t), \boldsymbol{R}_e(t), \boldsymbol{R}_s(t)) - f_{\boldsymbol{\theta}}(\boldsymbol{L}(0), \boldsymbol{R}_e(0), \boldsymbol{R}_s(0)) \\ &- \int_0^t \Big(\gamma(\boldsymbol{\theta}) + \sum_{i \in \mathcal{J}} \zeta_i(\boldsymbol{\theta}) 1(L_i(u) = 0)\Big) f_{\boldsymbol{\theta}}(\boldsymbol{L}(u), \boldsymbol{R}_e(u), \boldsymbol{R}_s(u)) du, \end{aligned} \tag{3.17}$$

$$\begin{aligned} E^{f_{\boldsymbol{\theta}}}(t) \equiv{}& f_{\boldsymbol{\theta}}(X(0))^{-1} \exp\Big[\langle \boldsymbol{\theta}, \boldsymbol{L}(t)\rangle + \langle \boldsymbol{\eta}(\boldsymbol{\theta}), \boldsymbol{R}_e(t)\rangle + \langle \boldsymbol{\zeta}(\boldsymbol{\theta}), \boldsymbol{R}_s(t)\rangle \\ &- \gamma(\boldsymbol{\theta}) t - \int_0^t \sum_{i \in \mathcal{J}} \zeta_i(\boldsymbol{\theta}) 1(L_i(u) = 0) du\Big]. \end{aligned} \tag{3.18}$$

4. Applications

In this section, we apply martingales obtained in Section 3 to the two asymptotic problems. We first consider two weak limit approximations for the stationary workload distribution of the the $GI/G/1$ queue. Arguments are quite similar to those in [8], but we require a Taylor expansion of $\eta(\theta)$, which is different from the one considered in [8]. We next consider the generalized Jackson network. The heavy traffic approximation has been fully studied for the stationary joint queue length distribution in [1], so we here only consider its tail asymptotics.

4.1. Weak limit approximations for the $GI/G/1$ queue

Consider the $GI/G/1$ of Example 2.1, and assume now the stability condition that $\rho < 1$. Then, $X(t) \equiv (V(t), R_e(t))$ has the stationary distribution, denoted by π, as is well known. Let $X(0)$ have this π, then $\{X(t)\}$ is a stationary process. For this stationary process, we take the expectation of (3.14) for $\theta \le 0$, then we have

$$(\theta + \eta(\theta))\mathbb{E}(e^{\theta V(0)+\eta(\theta)R_e(0)}) - \theta\mathbb{E}(e^{\eta(\theta)R_e(0)}1(V(0) = 0)) = 0. \tag{4.1}$$

This can be used to derive a heavy traffic approximation for the stationary workload distribution.

For this, we consider a sequence of the $GI/G/1$ queues, indexed by $n = 1, 2, \ldots$. Let $T_e^{(n)}, T_s^{(n)}$ be random variables subjects to the inter-arrival and service times of the n-th system respectively. Their distributions are denoted by $F_e^{(n)}, F_s^{(n)}$, respectively. Similarly, the function $\eta(\theta)$ of the n-th system is denoted by $\eta^{(n)}(\theta)$. Let

$$\begin{aligned}
&\lambda^{(n)} = 1/\mathbb{E}(T_e^{(n)}), \qquad \mu^{(n)} = 1/\mathbb{E}(T_s^{(n)}),\\
&(\sigma_e^{(n)})^2 = \mathbb{E}((T_e^{(n)} - \mathbb{E}(T_e^{(n)}))^2), \qquad (\sigma_s^{(n)})^2 = \mathbb{E}((T_e^{(s)} - \mathbb{E}(T_s^{(n)}))^2),
\end{aligned}$$

We assume that $\lambda^{(n)}, \mu^{(n)}, \sigma_e^{(n)}, \sigma_s^{(n)}$ converge to constants $\lambda, \mu, \sigma_e, \sigma_s$ as $n \to \infty$, and the heavy traffic condition that, for a sequence $r_n > 0$ vanishing as $n \to \infty$,

$$1 - \rho^{(n)} = r_n, \text{ where } \rho^{(n)} = \lambda^{(n)}/\mu^{(n)}. \tag{4.2}$$

Obviously, we must have $\lambda = \mu$, and the $V^{(n)}$ has the stationary distribution for each $n \ge 1$.

Let $V^{(n)}, R_e^{(n)}$ be random variables subject to the stationary distributions of the workload and residual arrival time, respectively, of the n-th system. Then, substituting $r_n\theta$ into the θ's of (4.1) and using (4.2), we have

$$(r_n\theta + \eta^{(n)}(r_n\theta))\mathbb{E}(e^{\theta r_n V^{(n)}+\eta^{(n)}(r_n\theta)R_e^{(n)}}) - r_n^2\theta\mathbb{E}(e^{\eta^{(n)}(r_n\theta)R_e^{(n)}}|V^{(n)} = 0)) = 0. \tag{4.3}$$

Then, letting $n \to \infty$ and applying the n-system version of Lemma 3.4 and using the fact that $\rho^{(n)} = 1 - r_n$, we have

$$\lim_{n\to\infty} \mathbb{E}(e^{\theta(1-\rho^{(n)})V^{(n)}}) = \frac{2}{2 - \lambda(\sigma_s^2 + \sigma_e^2)\theta}. \tag{4.4}$$

Thus, we have proved that the distribution of $(1 - \rho^{(n)})V^{(n)}$ weakly converges to the exponential distribution with mean $\lambda(\sigma_s^2 + \sigma_e^2)/2$, which is the well known heavy traffic approximation.

What is interesting in this approach, we can consider another approximation increasing the variances in such a way that

$$\lim_{n\to\infty} s_n(\sigma_e^{(n)})^2 = b_e^2, \qquad \lim_{n\to\infty} s_n(\sigma_s^{(n)})^2 = b_s^2, \tag{4.5}$$

for a sequence s_n vanishing as $n \to \infty$ and constants $b_e, b_s \ge 0$ satisfying $b_e + b_s > 0$. However, we still require the heavy traffic condition (4.2). Then, similar to the heavy traffic case, (4.1) yields

$$\lim_{n\to\infty} \mathbb{E}(e^{\theta(1-\rho^{(n)})s_n V^{(n)}}) = \frac{2}{2 - \lambda(b_s^2 + \rho^2 b_e^2)\theta}. \tag{4.6}$$

Thus, the limiting distribution is still exponential.

4.2. Tail asymptotics for the generalized Jackson network

Consider the GJN of Example 2.2. Denote the means of $T_{e,i}, T_{s,i}$ by $m_{e,i}$ and $m_{s,i}$, respectively. Let $\lambda_{e,i} = 1/m_{e,i}$ for $i \in \mathcal{E}$. Let α_i for $i \in \mathcal{J}$ be the solutions of the following traffic equation.

$$\alpha_i = \lambda_i 1(i \in \mathcal{E}) + \sum_{j \in \mathcal{J}} \alpha_j p_{ji}, \qquad i \in \mathcal{J}.$$

It is easy to see that the solutions uniquely exist if the $d \times d$ matrix $P \equiv \{p_{ij}; i,j \in \mathcal{J}\}$ is strictly substochastic and if $\overline{P} \equiv \{p_{ij}; i,j \in \{0\} \cup \mathcal{J}\}$ is irreducible, where $p_{00} = 0$, and $p_{0i} = \lambda_i 1(i \in \mathcal{E})/\sum_{j \in \mathcal{E}} \lambda_j$ for $i \in \mathcal{J}$. We assume these conditions without loss of generality. Let $\rho_i = \alpha_i m_i$, and assume the stability condition that $\rho_i < 1$ for all $i \in \mathcal{J}$.

Braverman et al. [1] study a similar approach to this GJN, not using martingales. They derive the limiting distribution of the stationary joint queue length distribution under the stability condition in addition to the standard heavy traffic conditions, which corresponds to the diffusion approximation for the joint queue length process (see, e.g., [10]).

We here consider the light tail asymptotic of that stationary distribution, using martingale. For this, we assume that $F_{e,i}$ and $F_{s,i}$ have light tails for all possible i. This tail asymptotic problem is known to be hard. Except for a tandem or out-tree network, which is essentially reduced to a single queue, the problem is solved only for the two node GJN under the phase-type setting in [7].

We change measures from $\mathbb{P}_\nu$ to $\widetilde{\mathbb{P}}^{(\boldsymbol{\theta})}_\nu$ similarly to (3.16) by using $E^{f_{\boldsymbol{\theta}}}$ of (3.18). For this, let τ^+_A, τ^-_A be the first exit and return times of $L(t)$ to $A \subset S_1$ such that $\tau^+_A < \tau^-_A$. Let ν^+_A the distribution of $X(t)$ at time τ^+_A given that $X(0)$ is subject to the normalized stationary distribution ν_A limited on $\{(z, \boldsymbol{y}) \in S; z \in A\}$. Denote a random vector subject to the stationary distribution of $X(t)$ by $X \equiv (\boldsymbol{L}, \boldsymbol{R}_e, \boldsymbol{R}_s)$. Then, the cycle formula yields, for measurable $B \subset S_1 \setminus A$,

$$\mathbb{P}(\boldsymbol{L} \in B) = c(A)\mathbb{E}_{\nu^+_A}\Big(\int_0^{\tau^-_A} 1(\boldsymbol{L}(u) \in B)du\Big), \tag{4.7}$$

where $c(A) = \mathbb{P}(\boldsymbol{L} \in A)/\mathbb{E}_{\nu^+_A}(\tau^-_A - \tau^+_A)$. Let

$$\begin{aligned}
&B_K(\boldsymbol{c}, n) = \{z \in S_1; \langle \boldsymbol{c}, \boldsymbol{z}\rangle \ge n, z_i = 0, \forall i \in K, n_j \ge 1, \forall j \in \mathcal{J} \setminus K\}, \quad \boldsymbol{c} \in S_1, n \ge 1,\\
&\tau^+_{\boldsymbol{c},n} = \inf\{t > 0; \boldsymbol{L}(t) \in B_K(\boldsymbol{c}, n)\},\\
&Y_{\tau^+_{\boldsymbol{c},n}}(B) = \mathbb{E}_{\nu^+_A}\Big(\int_{\tau^+_{\boldsymbol{c},n}}^{\tau^-_A} 1(\boldsymbol{L}(u) \in B_K(\boldsymbol{c}, n))du \big| \mathcal{F}_{\tau^+_{\boldsymbol{c},n}-}\Big) 1(\tau^+_{\boldsymbol{c},n} < \tau^-_A),
\end{aligned}$$

and applying (3.16) with $E^{f_{\boldsymbol{\theta}}}$ of (3.18) with $A \equiv \{z \in S; z_j = 0, j \in \mathcal{J} \setminus K\}$, we have

$$\begin{aligned}
\mathbb{P}(\boldsymbol{L} \in B) = c(A)\widetilde{\mathbb{E}}^{(\boldsymbol{\theta})}_{\nu^+_A}\Big(f_{\boldsymbol{\theta}}(X(0))Y_{\tau^+_{\boldsymbol{c},n}-}(B)\exp\Big[-\langle \boldsymbol{\theta}, \boldsymbol{L}(\tau^+_{\boldsymbol{c},n}-)\rangle - \langle \boldsymbol{\eta}(\boldsymbol{\theta}), \boldsymbol{R}_e(\tau^+_{\boldsymbol{c},n}-)\rangle\\
- \langle \boldsymbol{\zeta}(\boldsymbol{\theta}), \boldsymbol{R}_s(\tau^-_{\boldsymbol{c},n}-)\rangle + \gamma(\boldsymbol{\theta})\tau^+_{\boldsymbol{c},n} + \sum_{i \in K} \zeta_i(\boldsymbol{\theta}) \int_0^{\tau^+_{\boldsymbol{c},n}} 1(L_i(u) = 0)du\Big]\Big).
\end{aligned} \tag{4.8}$$

This formula suggests the following conjectures. For $K \subset \mathcal{J}$, let

$$\begin{aligned}
&\Gamma^+_K = \{\boldsymbol{\theta} \in \mathbb{R}^d; \gamma(\boldsymbol{\theta}) \le 0, \zeta_j(\boldsymbol{\theta}) \le 0, \forall j \in K\},\\
&\Gamma^-_K = \{\boldsymbol{\theta} \in \mathbb{R}^d; \gamma(\boldsymbol{\theta}) \ge 0, \zeta_j(\boldsymbol{\theta}) \ge 0, \forall j \in K\},\\
&\varphi_K(\boldsymbol{\theta}) = \mathbb{E}(e^{\sum_{i \in \mathcal{J} \setminus K} \theta_i L_i} 1(L_j = 0, \forall j \in K)), \qquad \text{(moment generating function)},\\
&\nabla\gamma(\boldsymbol{\theta}) = \Big(\frac{\partial}{\partial\theta_1}\gamma(\boldsymbol{\theta}), \frac{\partial}{\partial\theta_2}\gamma(\boldsymbol{\theta}), \ldots, \frac{\partial}{\partial\theta_d}\gamma(\boldsymbol{\theta})\Big), \qquad \text{(gradient vector)},
\end{aligned}$$

then, choosing $A = \{z \in S_1; \exists i \in \mathcal{J}, z_i = 0\}$ and $B = B_K(\boldsymbol{c}_{\mathcal{J}\setminus K}, n)$ for (4.8), where $\boldsymbol{x}_D$ is $|D|$-dimensional vector whose i-entry is x_i for $i \in D \subset \mathcal{J}$, we conjecture that, letting $K^c = \mathcal{J} \setminus K$,

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}(B_K(\boldsymbol{c}_{K^c}, n)) \le - \sup\{\langle \boldsymbol{c}_{K^c}, \boldsymbol{\theta}\rangle; \boldsymbol{\theta} \in \Gamma^+_K, \varphi_{K^c}(\boldsymbol{\theta}) < \infty\}, \tag{4.9}$$

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}(B_K(\boldsymbol{c}_{K^c}, n)) \ge - \inf\{\langle \boldsymbol{c}_{K^c}, \boldsymbol{\theta}\rangle; \boldsymbol{\theta} \in \Gamma^-_K, \nabla\gamma(\boldsymbol{\theta})_K \le \boldsymbol{0}, \nabla\gamma(\boldsymbol{\theta})_{K^c} \propto \boldsymbol{c}_{K^c}\}, \tag{4.10}$$

where $\boldsymbol{x} \propto \boldsymbol{y}$ stands for the vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ to be parallel.

We now explain the reason why (4.9) and (4.10) are expected. We first note that

$$|\langle \boldsymbol{\theta}, \boldsymbol{L}(\tau^+_{\boldsymbol{c},n}-)\rangle - \langle \boldsymbol{\theta}, \boldsymbol{c}\rangle n| \le |\boldsymbol{\theta}| \equiv \sum_{i\in\mathcal{J}} |\theta_i|, \quad Y_{\tau^+_{\boldsymbol{c},n}}(B_K(\boldsymbol{c}, n)) \text{ is uniformly bounded in } n.$$

Hence, from (4.8), $e^{\langle \boldsymbol{\theta}, \boldsymbol{c}\rangle n}\mathbb{P}(\boldsymbol{L} \in B)$ should be finite for $\boldsymbol{\theta} \in \Gamma^+_K$ and if $\varphi_{K^c}(\boldsymbol{\theta}) < \infty$, where the latter is needed for $\widetilde{\mathbb{E}}^{\boldsymbol{\theta}}_{\nu_A}(f_{\boldsymbol{\theta}}(X(0))) < \infty$ for the martingale $M_{\boldsymbol{\theta}}$ to be well defined. Thus, the upper bound would not be hard to prove. A difficult part is the lower bound (4.10). The extra condition in (4.9) is not needed because we can start $X(0)$ on a compact set A. However, we require that, for some $\delta > 0$, $\widetilde{\mathbb{P}}^{(\boldsymbol{\theta})}_{\nu^+_A}(\tau^+_{\boldsymbol{c},n} < \tau^+_A) > \delta$ for all $n \ge 1$. This must be related to the drift condition of $\boldsymbol{L}(t)$ inside of $\mathbb{R}^d_+$, which is the reason why the gradient conditions are added in (4.10).

Note that the upper bound (4.9) is similar to those for the d-dimensional reflecting random walk (see Theorem 6.1 and (6.9) of [6]), while the lower bound (4.10) for $K = \emptyset$ correspond to (6.6) in Lemma 6.2 of the same paper. Once (4.9) and (4.10) are obtained, the algorithm in Theorem 6.1 of [6] would give the decay rate in an arbitrary direction. However, at the present, we have not yet proved it, so this is also a conjecture.

References

[1] Braverman, A., Dai, J. and Miyazawa, M. (2015). Heavy traffic approximation for the stationary distribution of a generalized jackson network: the BAR approach. Submitted for publication.

[2] Davis, M. H. A. (1984). Piecewise deterministic Markov processes: a general class of non-diffusion stochastic models. *Journal of Royal Statist. Soc. series B*, **46** 353–388.

[3] Ethier, S. N. and Kurtz, T. G. (1986). *Markov Processes: Characterization and Convergence.* Wiley, New York.

[4] Glynn, P. and Whitt, W. (1994). Logarithmic asymptotics for steady-state tail probabilities in a single-server queue. *Journal of Applied Probability*, **31** 131–156.

[5] Jacod, J. and Shiryaev, A. N. (2003). *Limit Theorems for stochastic processes.* 2nd ed. Springer, Berlin.

[6] Miyazawa, M. (2011). Light tail asymptotics in multidimensional reflecting processes for queueing networks. *TOP*, **19** 233–299.

[7] Miyazawa, M. (2015). A superharmonic vector for a nonnegative matrix with QBD block structure and its application to a Markov modulated two dimensional reflecting process. *Queueing Systems*, **81** 1–48.

[8] Miyazawa, M. (2016). A unified approach for large queue asymptotics in a heterogeneous multiserver queue. To appear in Advances in Applied Probability.

[9] Palmowski, Z. and Rolski, T. (2002). A technique of the exponential change of measure for Markov processes. *Bernoulli*, **8** 767–785.

[10] Reiman, M. I. (1984). Open queueing networks in heavy traffic. *Mathematics of Operations Research*, **9** 441–458.