DISSERTATION

SADDLEPOINT APPROXIMATION TO FUNCTIONAL EQUATIONS IN QUEUEING THEORY AND INSURANCE MATHEMATICS

Submitted by

Sunghoon Chung

Department of Statistics

In partial fulllment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Fall 2010

Doctoral Committee:

Department Chair: F. Jay Breidt

Advisor: Ronald W. Butler

Louis L. Scharf
Phillip L. Chapman
Jennifer A. Hoeting

ABSTRACT

SADDLEPOINT APPROXIMATION TO FUNCTIONAL EQUATIONS IN QUEUEING THEORY AND INSURANCE MATHEMATICS

We study the application of saddlepoint approximations to statistical inference when the moment generating function (MGF) of the distribution of interest is an explicit or an implicit function of the MGF of another random variable which is assumed to be observed. In other words, let $\mathcal{W}(s)$ be the MGF of the random variable W of interest. We study the case when $\mathcal{W}(s) = h\{\mathcal{G}(s);\lambda\}$, where $\mathcal{G}(s)$ is an MGF of G for which a random sample can be obtained, and h is a smooth function. If $\hat{\mathcal{G}}(s)$ estimates $\mathcal{G}(s)$, then $\hat{\mathcal{W}}(s) = h\{\hat{\mathcal{G}}(s);\hat{\lambda}\}$ estimates $\mathcal{W}(s)$. Generally, it can be shown that $\hat{\mathcal{W}}(s)$ converges to $\mathcal{W}(s)$ by the strong law of large numbers, which implies that $\hat{F}(t)$, the cumulative distribution function (CDF) corresponding to $\hat{\mathcal{W}}(s)$, converges to $F(t)$, the CDF of W, almost surely. If we set $\hat{\mathcal{W}}^*(s) = h\{\hat{\mathcal{G}}^*(s);\hat{\lambda}^*\}$, where $\hat{\mathcal{G}}^*(s)$ and $\hat{\lambda}^*$ are the empirical MGF and the estimator of λ from bootstrapping, the corresponding CDF $\hat{F}^*(t)$ can be used to construct the confidence band of $F(t)$.

In this dissertation, we show that the saddlepoint inversion of $\hat{\mathcal{W}}(s)$ is not only fast, reliable, stable, and accurate enough for a general statistical inference, but also easy to use without deep knowledge of the probability theory regarding the stochastic process of interest.

For the first part, we consider nonparametric estimation of the density and the CDF of the stationary waiting times W and W_q of an M/G/1 queue. These estimates are computed using saddlepoint inversion of $\hat{\mathcal{W}}(s)$ determined from the Pollaczek-Khinchin formula. Our saddlepoint estimation is compared with estimators based on other approximations, including the Cramér-Lundberg approximation.

For the second part, we consider the saddlepoint approximation for the busy period distribution $F_B(t)$ in a M/G/1 queue. The busy period B is the first passage time for the queueing system to pass from an initial arrival (1 in the system) to 0 in the system. If $\mathcal{B}(s)$ is the MGF of B, then $\mathcal{B}(s)$ is an implicitly defined function of $\mathcal{G}(s)$ and λ, the inter-arrival rate, through the well-known Kendall-Takács functional equation. As in the first part, we show that the saddlepoint approximation can be used to obtain $\hat{F}_B(t)$, the CDF corresponding to $\hat{\mathcal{B}}(s)$ and simulation results show that confidence bands of $F_B(t)$ based on bootstrapping perform well.

ACKNOWLEDGMENTS

I am grateful for my committee members, Dr. Scharf, Dr. Chapman, and Dr. Hoeting, for their suggestions concerning my dissertation. Dr. Scharf and Dr. Hoeting have pointed out numerous typos and grammatical errors in this work. I would, also, like to thank my advisor, Dr. Butler, for his kindness to agree, first and foremost, to be my advisor, for of all the discussions to keep me curious about our topic and to make me realize that there were yet more questions to answer, and for the financial support that allowed me to concentrate on the research during many summers in Fort Collins. Without his suggestions, proofreading, and countless corrections, my report would not be shaped into this PhD dissertation.

I also wanted to thank each of the professors in this department from whom I have learned statistics. Before arriving at the Statistics building, I had never even heard the term "hypothesis test," among others, and hopefully this work shows that their time spent on teaching me was not wasted.

Finally, I cherish my friendships with Austin, Vicky, Arkom, Courtney, Mark, Rongning, Martha, Derek, Lidong, Wenying, Ming-ying, Aggie and Amber, which made my life at CSU less suffering and kept me sane over the years.

DEDICATION

아버지 어머니께 감사드리며 제 박사 학위 논문을 헌정합니다.

I dedicate my dissertation to my father and mother for their endless support throughout this rather long and winding journey.

TABLE OF CONTENTS

Summary of Dissertation

We study the application of saddlepoint approximations to statistical inference when the moment generating function (MGF) of the distribution of interest is an explicit or an implicit function of the MGF of another random variable which is assumed to be observed. In other words, let $\mathcal{W}(s)$ be the MGF of the random variable W of interest. We study the case when $\mathcal{W}(s) = h\{\mathcal{G}(s); \lambda\}$, where $\mathcal{G}(s)$ is a MGF of G for which a random sample can be obtained, and h is a smooth function. If $\hat{\mathcal{G}}(s)$ estimates $\mathcal{G}(s)$, then $\hat{\mathcal{W}}(s) = h\{\hat{\mathcal{G}}(s); \hat{\lambda}\}$ estimates $\mathcal{W}(s)$. Generally, it can be shown that $\hat{\mathcal{W}}(s)$ converges to $\mathcal{W}(s)$ by the strong law of large numbers, which implies that $\hat{F}(t)$, the cumulative distribution function (CDF) corresponding to $\hat{\mathcal{W}}(s)$, converges to $F(t)$, the CDF of W, almost surely. If we set $\hat{\mathcal{W}}^*(s) = h\{\hat{\mathcal{G}}^*(s); \hat{\lambda}^*\}$, where $\hat{\mathcal{G}}^*(s)$ and $\hat{\lambda}^*$ are the empirical MGF and the estimator of λ from bootstrapping, the corresponding CDF $\hat{F}^*(t)$ can be used to construct the confidence band of $F(t)$.

With bootstrapping in mind, it is clear that the success of this approach depends on the existence of a fast, reliable, and stable method of inverting $\hat{\mathcal{W}}(s)$ to obtain $\hat{F}(t)$. Also, one might consider deriving an asymptotic formula based on $\mathcal{W}(s)$ and attempt to use the empirical version of that asymptotic formula with $\hat{\mathcal{W}}(s)$. However, deriving asymptotic formulas requires deep knowledge of the stochastic processes of interest and probability theory and this may not be possible.

In this dissertation, we show that the saddlepoint inversion of $\hat{\mathcal{W}}(s)$ is not only fast, reliable, stable, and accurate enough for a general statistical inference, but also easy to use without requiring deep knowledge of the probability theory regarding the stochastic process

of interest. In the preface of *An Introduction to the Bootstrap*, it was remarked, "Statistics is a subject of amazingly many uses and surprisingly few effective practitioners. The traditional road to statistical knowledge is blocked, for most, by a formidable wall of mathematics. Our approach here breeches that wall. The bootstrap is a computer-based method of statistical inference that can answer many statistical questions without formulas. Our goal in this book is to arm scientists and engineers, as well as statisticians, with computational techniques that they can use to analyze and understand complicated data sets." Though the author dares not compare his work to the "primary reference" of the bootstrap method, he believes the approach in this dissertation is very similar to the remark in spirit - if the MGF is available, it is possible to do meaningful statistical inference without seemingly formidable knowledge of mathematics and probability, through the saddlepoint approximation.

The most famous such formula in queueing theory is the Pollaczek-Khinchin formula for the classical M/G/1 queue. This queue has Poisson process input, general service time, and a single server that uses FCFS (first-come, first-served) principle. The Pollaczek-Khinchin formula specifies $\mathcal{W}(s)$, the MGF of the stationary waiting time distribution, in terms of $\mathcal{G}(s)$, the MGF of the general service time distribution. Chapter 1 considers nonparametric estimation of the density and the CDF of the stationary waiting time W of an M/G/1 queue. These estimates are computed using saddlepoint inversion of $\hat{\mathcal{W}}(s)$ determined from the Pollaczek-Khinchin formula. The bootstrap is also used to construct a confidence band for the CDF. Implementation of the bootstrap becomes computationally feasible primarily with the use of saddlepoint approximation.

The stationary waiting time in system W consists of the stationary waiting time in queue W_q plus the service time G. Chapter 2 considers estimation of the CDF for W_q, or $F_q(t)$ using saddlepoint methods similar to those in Chapter 1. This particular distribution has received considerable attention in the area of insurance mathematics. Distribution $F_q(t)$, in the insurance context gives the probability of eventual ruin (bankruptcy) for the company when it starts with initial cash reserve t and is subject to a compound distribution of claims. The prominence of such ruin computations has lead to various procedures in the insurance/applied probability literature for approximating $F_q(t)$. Among the procedures is the well-known Cramér-Lundberg approximation. Our saddlepoint estimation is compared with estimators based on these other approximations including the Cramér-Lundberg approximation. The comparison shows that saddlepoint approximations are more accurate than all other approximations when cash reserves are small to moderately large but not excessively large. For very large cash reserves all approximations appear to work equally well although saddlepoint approximation also shows wider applicability.

Chapter 3 considers saddlepoint approximation for the busy period distribution $F_B(t)$ in a M/G/1 queue. The busy period B is the first passage time for the queueing system to pass from an initial arrival (one in the system) to 0 in the system. If $\mathcal{B}(s)$ is the MGF of B, then $\mathcal{B}(s)$ is an implicitly defined function of $\mathcal{G}(s)$ and λ, the inter-arrival rate, through the well-known Kendall-Takács functional equation. As in Chapter 1, we show that the saddlepoint approximation can be used to obtain $\hat{F}_B(t)$, the CDF corresponding to $\hat{\mathcal{B}}(s)$ and simulation results show that confidence bands of $F_B(t)$ based on bootstrapping perform well.

In Chapter 4, we re-direct our attention from the saddlepoint approximation to moment estimators based on $\hat{\mathcal{W}}(s)$ and $\hat{\mathcal{B}}(s)$. We investigate the bootstrap confidence interval (CI) of EW, $\operatorname{Var} W$, EB, and $\operatorname{Var} B$. We show that the CI based on the percentile is the only viable method of the methods we consider and suggest a modified bootstrap percentile CI as a better method for these cases, based on simulation results.

Chapter 1

Saddlepoint Approximation to Pollaczek-Khinchin Formula

1.1 Chapter overview

When the service time distribution and the arrival rate are given, the stationary waiting time distribution can be recognized up to its moment generating function (or Laplace transform) only by Pollaczek-Khinchin formula and because of this restriction, the cumulative distribution function (CDF) or density function of the stationary waiting time distribution can only be obtained by direct inversion if possible (see [47, 48], and [9] for the matrix-geometric solution for phase type service distribution, §1.4 for Erlang service time distribution, [15] and [7] for certain heavy tailed service time distributions, and [52] for Pareto service time distribution), asymptotic approximation (see §2.1 and the reference therein), or through numerical inversion methods ([6],[4],[62]). In fact, at least in queueing theory it has been a prominent example to illustrate the numerical Laplace transform inversion method. For a general overview of the numerical Laplace transform inversion method, see [20].

Let $F(t)$ and $f(t) = F'(t)$ be the unknown CDF and the density functions for the stationary waiting time in a M/G/1 queueing system. This chapter shows how saddlepoint approximation can be used to formulate nonparametric estimators $\hat{F}_0(t)$ and $\hat{f}_0(t)$ for these functions based on the data obtained from observation of the queue.

To obtain these saddlepoint estimators, we must first consider the MGFs involved, in particular the MGF of $f(t)$ denoted as $\mathcal{W}(s)$, the moment generating function (MGF) of stationary waiting time in system (so waiting time $W = W_q + G$, where W_q is stationary waiting time or delay and G is the subsequent service time.) Let $\mathcal{G}(s)$ be the moment generating function of the service time distribution. Then, the Pollaczek-Khinchin formula for M/G/1 is

$$\mathcal{W}(s) = \frac{(1-\rho)\, s\mathcal{G}(s)}{s + \lambda - \lambda\mathcal{G}(s)}, \tag{1.1.1}$$

where λ is the arrival rate, $\mathcal{G}'(0) = 1/\mu$ is the average service time, and $\rho = \lambda/\mu < 1$ is the stability parameter indicating stationarity when less than 1.

The saddlepoint inversion of $\mathcal{W}(s)$ could lead to approximations $F_0(t)$ and $f_0(t)$ however λ and $\mathcal{G}(s)$ in (1.1.1) are unknown. We assume it is possible to obtain a random sample of service times, $\{G_j : j = 1 \ldots, n\}$ and inter-arrival times $\{I_j : j = 1, \ldots, n\} \overset{\text{iid}}{\sim} \text{Exp}(\lambda)$. For simplicity, we assume that the random sample sizes of $\{I_j\}$ and $\{G_j\}$ are both n though this assumption can be relaxed. Let

$$\hat{\rho} = \frac{\hat{\lambda}}{n} \sum_{j=1}^{n} G_j = \frac{\sum G_j}{\sum I_j} \tag{1.1.2}$$

where $\hat{\lambda} = 1/\overline{I}_n$ is the maximum likelihood estimator of the rate of inter-arrival times. Then, we obtain

$$\hat{\mathcal{W}}(s) = \frac{(1-\hat{\rho})s\hat{\mathcal{G}}(s)}{s + \hat{\lambda} - \hat{\lambda}\hat{\mathcal{G}}(s)}, \tag{1.1.3}$$

which is exactly the plug-in estimator of the MGF of W, i.e., $\mathcal{G}(s)$ is replaced by its empirical MGF

$$\hat{\mathcal{G}}(s) = \frac{1}{n} \sum_{i=1}^{n} e^{sG_i}. \tag{1.1.4}$$

$\hat{\mathcal{W}}(s)$ is inverted using saddlepoint methods to determine $\hat{F}_0(t)$, the saddlepoint CDF approximation, and $\hat{f}_0(t)$, the saddlepoint density approximation, as approximations to $F_0(t)$ and $f_0(t)$ which serve as approximations for $F(t)$ and $f(t)$.

The remaining part of this chapter is organized as follows: In the rest of this section, we introduce the saddlepoint approximation. In Section 2, we study the properties of the plug-in estimators, $\hat{\mathcal{W}}(s)$ and $\hat{\mathcal{K}}(s)$. They are locally uniformly a.s. convergent to $\mathcal{W}(s)$ and $\mathcal{K}(s)$, respectively. We also investigate the limit behavior of $n^{-\delta}\{\hat{\mathcal{W}}(s)-\mathcal{W}(s)\}$ and derive the result regarding the convergence rate of $\hat{F}_0(t)-F_0(t)$.

In Section 3, we show how the saddlepoint approximation can be used to obtain the bootstrap confidence interval for the CDF of W. In Section 4, we perform the parametric estimation of the CDF and PDF of M/E_k/1 queue through analytic inverse Laplace transform and compare with the saddlepoint approximations. In the Appendix we give the proof of the theorems which are not directly related to the theme of this chapter but which we use in Section 2.

1.1.1 Saddlepoint density and CDF approximation methods

Let $\mathcal{K}(s)$ be the cumulant generating function (CGF) of the stationary waiting time distribution W such that

$$\mathcal{K}(s)=\log\mathcal{W}(s),$$

and let Φ and ϕ be the CDF and PDF of the standard normal distribution respectively. Then the Lugannani and Rice saddlepoint approximation for the CDF $F(t)$ of W is defined in [43] by

$$F_0(t)=\begin{cases}\Phi(r_t)+\phi(r_t)(1/r_t-1/u_t) & \text{if } t\neq\mathcal{K}'(0)\\ \frac{1}{2}+\frac{\mathcal{K}'''(0)}{6\sqrt{2\pi}\mathcal{K}''(0)^{3/2}} & \text{if } t=\mathcal{K}'(0),\end{cases}$$

where r_t and u_t are defined by

$$\begin{aligned}r_t&=\operatorname{sgn}(s_t)\sqrt{2\{s_t t-\mathcal{K}(s_t)\}}\\ u_t&=s_t\sqrt{\mathcal{K}''(s_t)},\end{aligned}$$

and s_t denotes the unique solution to the saddlepoint equation,

$$\mathcal{K}'\left(s_t\right) = t \tag{1.1.5}$$

over the range $s_t \in (-\infty, c)$, where $(-\infty, c)$ is the largest open neighborhood of 0 on which $\mathcal{K}\left(s\right)$ is convergent. Using the same notation, the saddlepoint density estimation is defined in [24] by

$$f\left(t\right) = \frac{\phi\left(r_t\right)}{\sqrt{\mathcal{K}''\left(s_t\right)}}.$$

Since $\mathcal{K}''\left(s\right) > 0$ over $(-\infty, c)$, the solution is well defined. However, it is worth noting that the solution of the saddlepoint equation is meaningful only when $s_t \in (-\infty, c)$, which should be considered when numerical implementations is involved. See the remarks of §1.1.4 of [17] for the related explanation.

Saddlepoint approximations require higher-order derivatives of $\mathcal{K}\left(s\right)$, $\mathcal{K}'\left(s\right)$ and $\mathcal{K}''\left(s\right)$ in particular. Using our plug-in estimator (1.1.3), we have

$$\hat{\mathcal{K}}\left(s\right) = \log \hat{\mathcal{W}}\left(s\right) = \log \frac{\left(1-\hat{\rho}\right) s\hat{\mathcal{G}}\left(s\right)}{s + \hat{\lambda} - \hat{\lambda}\hat{\mathcal{G}}\left(s\right)}.$$

Taking the first derivative gives the estimators of $\mathcal{K}'$ as

$$\hat{\mathcal{K}}'\left(s\right) = \frac{d}{ds}\{\log \hat{\mathcal{W}}\left(s\right)\} = \frac{\hat{\lambda}\{1-\hat{\mathcal{G}}\left(s\right)\}\hat{\mathcal{G}}\left(s\right) + s(s+\hat{\lambda})\hat{\mathcal{G}}'\left(s\right)}{s\hat{\mathcal{G}}\left(s\right)\left\{s + \hat{\lambda} - \hat{\lambda}\hat{\mathcal{G}}\left(s\right)\right\}}, \tag{1.1.6}$$

where

$$\hat{\mathcal{G}}'\left(s\right) = \frac{d}{ds}\hat{\mathcal{G}}\left(s\right) = \frac{1}{n}\sum_{i=1}^{n} G_i e^{sG_i}.$$

Further differentiation gives the expression for second derivative of $\hat{\mathcal{K}}\left(s\right)$ needed in saddlepoint approximations. We also note that the empirical saddlepoint approximation is defined by

$$\hat{F}_0\left(t\right) = \begin{cases} \Phi\left(\hat{r}_t\right) + \phi\left(\hat{r}_t\right)\left(1/\hat{r}_t - 1/\hat{u}_t\right) & \text{if } t \neq \hat{\mathcal{K}}'\left(0\right) \\ \frac{1}{2} + \frac{\hat{\mathcal{K}}'''(0)}{6\sqrt{2\pi}\hat{\mathcal{K}}''(0)^{3/2}} & \text{if } t = \hat{\mathcal{K}}'\left(0\right), \end{cases}$$

where $\hat{\mathcal{K}}^{(j)}(0)$, $j = 1, 2, 3, 4$ can be calculated with Lemma 23.

It is known that $D^k\hat{\mathcal{G}}(s) \xrightarrow{\text{a.s.}} D^k\mathcal{G}(s)$ for any fixed s and $k \in \mathbb{N}_0$, where D is the differentiation operator (D^0 is the identity operator) and $\mathbb{N}_0$ is the set of all natural numbers and 0. Moreover, it has been proved in [31] that for any finite interval of the region on which $\mathcal{G}$ is convergent, $D^k\hat{\mathcal{G}}(s) \xrightarrow{\text{a.s.}} D^k\mathcal{G}(s)$ and $D^k \log\hat{\mathcal{G}}(s) \xrightarrow{\text{a.s.}} D^k \log\mathcal{G}(s)$ uniformly, which is sometimes called locally uniform a.s. convergence in this paper.

From now on, we use $F_0(t)$ for the Lugannani and Rice saddlepoint CDF approximation using the true CGF, and $\hat{F}_0(t)$ for the Lugannani and Rice saddlepoint CDF approximation using the empirical CGF. $f_0(t)$ and $\hat{f}_0(t)$ are defined similarly.

The idea of using the empirical MGF for saddlepoint approximations was proposed in [27] to apply to bootstrap simulation approximation. In [31], the properties of saddlepoint approximation using the empirical CGF were investigated and considered useful when the CGF of the interested distribution is intractable and only a sample from the distribution is available. As in [31], we can decompose the error of using the empirical saddlepoint approximation,

$$\hat{F}_0(t) - F(t) = \hat{F}_0(t) - F_0(t) + F_0(t) - F(t)$$

and it can be shown that $\hat{F}_0(t) \xrightarrow{\text{a.s.}} F_0(t)$, and $\hat{F}_0(t)$ is asymptotically biased unless $F_0(t) = F(t)$. However, as shown in [17], saddlepoint estimation is highly accurate and in most instances this asymptotic bias is so small that it can be ignored for practical purposes.

1.2 Properties of empirical cgf and the plug-in estimator

We investigate properties of our plug-in estimator (1.1.3) and the resulting saddlepoint estimate of the PDF and CDF of the stationary waiting time distribution of M/G/1 queues. In our discussion, we assume the inter-arrival time distribution (I_i) and the service time distribution (G_i) have finite second moments and are independent of each other.

For the results regarding the asymptotic normality of $\hat{\mathcal{W}}(s)$ and the asymptotic result regarding $\hat{F}_0(t)$, we mostly follow the approach of [31] though most of the proofs are the author's own.

1.2.1 Solvability of the saddlepoint equation

We now consider solvability of the saddlepoint equation (1.1.5) and its empirical counter part $t = \hat{\mathcal{K}}'(\hat{s}_t)$ with the assumption that either $\mathcal{G}(s)$ is known or the empirical MGF $\hat{\mathcal{G}}(s)$ is obtained.

Note that $\mathcal{K}'(s)$ (and its empirical counter part $\hat{\mathcal{K}}'(s)$) is a strictly increasing continuous function. Let $(-\infty, c)$ be the largest open connected set including 0 on which $\mathcal{K}'(s)$ is convergent. Because $\mathcal{K}'(s)$ is continuous, $\mathcal{K}'\{(-\infty, c)\}$, its image under $\mathcal{K}'$, is also connected (Theorem 23.5 of [46]).

Thus, for any $t \in \mathcal{K}'\{(-\infty, c)\}$, the saddlepoint solution s_t exists by the intermediate value theorem (Theorem 24.3 of [46]) and the fact $\mathcal{K}'$ is a one to one function on $(-\infty, c)$.

Daniels ([24], §6) showed that $F(t) = 0$ for $t < a$ if and only if $\mathcal{K}'(s_t) = t$ has no real root for $t < a$, which implies, with the fact mentioned above, that

$$\lim_{s \to -\infty} \mathcal{K}'(s) = \inf\{t : F(t) > 0\}, \tag{1.2.1}$$

the infimum of the support of W, which will be denoted by w_0. We investigate how $w_0 = \lim_{s \to -\infty} \mathcal{K}'(s)$ is related to $\mathcal{G}(s)$ first.

Lemma 1. *The infimum of the support of W is the same as the infimum of the support of G. Thus, $w_0 = g_0$, where g_0 is the infimum of the support of G.*

Proof. We have

$$\begin{aligned}
\lim_{s \to -\infty} \frac{d}{ds} \log \mathcal{W}(s) &= \lim_{s \to -\infty} \frac{\lambda\{1 - \mathcal{G}(s)\}\mathcal{G}(s) + s(s+\lambda)\mathcal{G}'(s)}{s\mathcal{G}(s)\{s + \lambda - \lambda\mathcal{G}(s)\}} \\
&= \lim_{s \to -\infty} \frac{\lambda\{1 - \mathcal{G}(s)\}}{s\{s + \lambda - \lambda\mathcal{G}(s)\}} + \lim_{s \to -\infty} \left\{ \frac{s + \lambda}{s + \lambda - \lambda\mathcal{G}(s)} \cdot \frac{\mathcal{G}'(s)}{\mathcal{G}(s)} \right\} \\
&= \lim_{s \to -\infty} \frac{\mathcal{G}'(s)}{\mathcal{G}(s)} = \lim_{s \to -\infty} \frac{d}{ds} \log \mathcal{G}(s).
\end{aligned}$$

□

Thus, for our plug-in estimator $\hat{\mathcal{K}}(s)$, using the same argument as in the proof of the above lemma with the corresponding estimator instead, we obtain that $G_{(1)}$, the first order statistic, is the infimum of the support of $\hat{\mathcal{K}}(s)$ by Lemma 22.

If $\lim_{s \nearrow c} \mathcal{K}'(s) = \infty$, then the saddlepoint equation has a unique solution for any $t \in (g_0, \infty)$, which is known as steepness (p. 86 and p.117 of [10]). We now show what property of $\mathcal{G}(s)$ makes $\mathcal{W}(s)$ steep. To understand this better, we need to consider the behavior of a MGF at the boundary point of the domain of the MGF. For any MGF (or CGF), the largest connected set containing 0 will be called a convergence strip of that MGF (or the CGF). There are two different types of convergence strip, namely open interval, $(-\infty, b)$, and half open (closed) interval, $(-\infty, b]$ $(b < \infty)$. We note that $\lim_{s \nearrow b} \mathcal{G}(s) = \infty$ if the convergence strip is $(-\infty, b)$ and $\mathcal{G}(b) < \infty$ if the convergence strip is $(-\infty, b]$.

For the steepness property of $\mathcal{K}(s)$, we first note that by the definition of $\mathcal{W}(s)$, if the service time distribution has MGF $\mathcal{G}(s)$ which is convergent on $(-\infty, b)$ (or $(-\infty, b]$), then the convergence strip of $\mathcal{W}(s)$ (and $\mathcal{K}(s)$ and $\mathcal{K}'(s)$) must be included in $(-\infty, b)$ (or $(-\infty, b]$), i.e., the convergence strip of $\mathcal{W}(s)$ must be a subset of the convergence strip of $\mathcal{G}(s)$.

If c is the smallest positive root of $\mathcal{D}(s) = s + \lambda - \lambda\mathcal{G}(s)$, the denominator of $\mathcal{W}(s)$, and it exists in $(0, b]$, then

$$\lim_{s \nearrow c} \mathcal{K}'(s) = \infty,$$

or $\mathcal{W}(s)$ is steep.

Note that 0 is a removable singularity of $\mathcal{W}(s)$ (see Lemma 23) and we actually understand $\mathcal{W}(s)$ as

$$\mathcal{W}(s) = \begin{cases} \frac{(1-\rho)s\mathcal{G}(s)}{s+\lambda-\lambda\mathcal{G}(s)} & \text{if } s \neq 0 \\ 1 & \text{if } s = 0 \end{cases},$$

which makes $\mathcal{W}(s)$ continuous on the convergence strip of $\mathcal{W}(s)$. The same interpretations are used for $\mathcal{K}(s) = \log \mathcal{W}(s)$, $\mathcal{K}'(s) = \{\log \mathcal{W}(s)\}'$, etc and their corresponding estimator $\hat{\mathcal{W}}(s)$, and so on.

First, we check when c exists.

Lemma 2. *Let $\mathcal{G}(s)$ have the open convergence strip $(-\infty, b)$ or $(-\infty, b]$ and define $\mathcal{D}(s) = s + \lambda - \lambda\mathcal{G}(s)$. Then, the unique positive root of $\mathcal{D}(s)$, $c(\leq b)$ exists if and only if $b > 0$ and*

$$\lim_{s \nearrow b} \mathcal{D}(s) \leq 0.$$

Proof. Observing $\mathcal{D}(0) = 0$, and $\mathcal{D}'(0) = 1 - \lambda\mathcal{G}'(0) = 1 - \rho > 0$ (from the stability condition), and $\mathcal{G}'(s)$ is a strictly increasing function ($\mathcal{G}''(s) > 0$), we can conclude that $\mathcal{D}'(s) = 1 - \lambda\mathcal{G}'(s)$ has only one (positive) root d if it exists in $(0, b)$ and $\mathcal{D}(d) > 0$ is the maximum by the first derivative test.

Therefore, by the intermediate value theorem, the unique positive root c of $\mathcal{D}(s)$ satisfying $d < c \leq b$ exists if and only if $\lim_{s \nearrow b} \mathcal{D}(s) \leq 0$. □

Thus, the convergence strip of $\mathcal{W}(s)$ is $(-\infty, c)$ if c exists or is the same as the convergence strip of $\mathcal{G}(s)$ (either $(-\infty, b)$ or $(-\infty, b]$) if c does not exist. For $\hat{\mathcal{W}}(s)$, $\hat{c}$, the positive root of $\hat{\mathcal{D}}(s) = s + \hat{\lambda} - \hat{\lambda}\hat{\mathcal{G}}(s)$, always exists because $\lim_{s \to \infty} \hat{\mathcal{G}}(s) = \infty$. The existence of c also depends on λ:

Example 3. Let G follow the inverse Gaussian distribution with $EG = \operatorname{Var} G = 1$. Then its CGF is

$$1 - \sqrt{1 - 2s} \qquad \text{for } s \leq \frac{1}{2},$$

which has first derivative, $1/\sqrt{1-2s}$ for $s \leq 1/2$. Thus, $\mathcal{G}(s)$ is steep. Solving the equation,

$$1/2 + \lambda - \lambda\mathcal{G}(1/2) = 1/2 + \lambda - \lambda e = 0,$$

with respect to λ, we have $\lambda = 1/\{2(e-1)\} \approx 0.2909884$. Because $d\mathcal{D}/d\lambda = d\{s + \lambda - \lambda\mathcal{G}(s)\}/d\lambda = 1 - \mathcal{G}(s) < 0$ for $s > 0$, we have

$$\mathcal{D}(1/2) \begin{cases} < 0 & \text{if } \lambda > 1/\{2(e-1)\}, \\ > 0 & \text{if } \lambda < 1/\{2(e-1)\}. \end{cases}$$

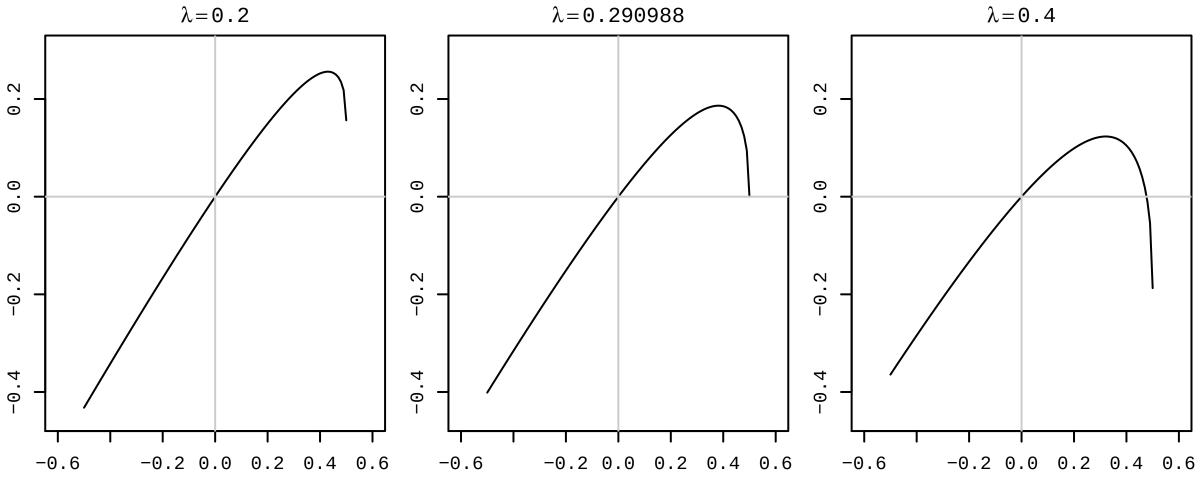

Figure 1.2.1: The graphs of $\mathcal{D}(s) = s + \lambda - \lambda\mathcal{G}(s)$, where $\mathcal{G}(s) = \exp(1 - \sqrt{1-2s})$ with $\lambda = .2$ (left), $1/\{2(e-1)\} \approx 0.2909884$ (middle), and .4 (right).

Thus, if $\lambda \geq 1/\{2(e-1)\}$, then c exits and if $\lambda < 1/\{2(e-1)\}$, c does not exist. See Figure 1.2.1 for the case of $\lambda = 0.2$, 0.290988, and 0.4 respectively.

Note that if the convergence strip of $\mathcal{G}(s)$ is $(-\infty, b)$, $\mathcal{W}(s)$ is steep since $\lim_{s\nearrow b}\mathcal{G}(s) = \infty$ implies $\lim_{s\nearrow b}\mathcal{D}(s) = -\infty$. For the case of $(-\infty, b]$, if $b > 0$, we have

$$\lim_{s\nearrow b}\mathcal{K}'(s) = \frac{\lambda\{1-\mathcal{G}(b)\}\mathcal{G}(b) + b(b+\lambda)\lim_{s\nearrow b}\mathcal{G}'(s)}{b\mathcal{G}(b)\{b+\lambda-\lambda\mathcal{G}(b)\}},$$

so that $\mathcal{W}(s)$ is steep if $\mathcal{G}(s)$ is steep at b or $c\,(\leq b)$ exists. For the case of $(-\infty, b]$ and $b = 0$, we have $\lim_{s\nearrow 0}\mathcal{K}'(s) = EW = EG + \lambda EG^2/\{2(1-\rho)\}$ (see Lemma 23 or use $W = \sum_{j=1}^{N} V_j + G_1$), so that $\mathcal{W}(s)$ is steep if $EG^2 = \infty$. These can be summarized as follows.

Lemma 4. *$\mathcal{W}(s)$ is steep if and only if one of the following holds:*

1. *c exists,*

2. *$\mathcal{G}(s)$ is steep at $b > 0$, or*

3. *$b = 0$ and $EG^2 = \infty$.*

In summary, we have the following proposition.

Proposition 5. *If $\mathcal{W}(s)$ is steep, then the saddlepoint equation (1.1.5) can be solved for any $t \in (g_0, \infty)$ where $g_0 = \inf\{t : G(t) > 0\}$. If not, then the saddlepoint equation (1.1.5) can be solved for any $t \in (g_0, \mathcal{K}'(b)]$. For the saddlepoint equation of the plug-in estimator (1.1.6), it can be solved for any $t \in (G_{(1)}, \infty)$.*

Proof. We only need to show the part for the plug-in estimator (1.1.6), which can be derived from the fact that $\hat{\mathcal{G}}(s)$ is convergent on $\mathbb{R}$ and $\lim_{s\to\infty} \hat{\mathcal{G}}(s) = \infty$. □

Example 6. Suppose G follows Pareto(.8,5) distribution which has the density,

$$5\left(\frac{4}{5}\right)^5 \frac{1}{x^6}\mathbf{1}_{[4/5,\infty)}(x).$$

Then, $\mathcal{G}(s)$ is defined only on $(-\infty, 0]$ and $EG^2 = 16/15$. Thus, $\mathcal{W}(s)$ is not steep and the saddlepoint equation can be solved for $t \in (.8, 1]$ only. Because the nth raw moment of the Pareto(α, β), $\beta\alpha^n/(\alpha - n)$ exists for $\alpha > n$, $\mathcal{W}(s)$ is steep at 0 if $\alpha \leq 2$. Note that $\mathcal{G}(s)$ itself is not steep if $\alpha > 1$ since $\lim_{s\nearrow 0}\{\log \mathcal{G}(s)\}' = EG = \beta\alpha/(\alpha - 1) < \infty$. Thus, Pareto$(1.5, \beta)$ is an example for which $\mathcal{G}(s)$ is not steep but it makes $\mathcal{W}(s)$ steep.

1.2.2 Properties of the plug-in estimators

We note that $E\hat{\mathcal{G}}(s) = \frac{1}{n}\sum_{i=1}^n Ee^{sG_i} = \mathcal{G}(s)$ and similarly, for any $k = 0, 1, \ldots,$

$$ED^k\hat{\mathcal{G}}(s) = D^k\mathcal{G}(s).$$

It is proved in [31] that in the space of continuous functions with the supremum norm, $\sqrt{n}\{D^k\hat{\mathcal{G}}(s) - D^k\mathcal{G}(s)\}$ converges weakly to a Gaussian process of zero mean and the covariance structure of

$$n\,\mathrm{Cov}\{D^k\mathcal{G}_n(s), D^k\mathcal{G}_n(r)\} = D^{2k}\mathcal{G}(s + r) - D^k\mathcal{G}(s)\,D^k\mathcal{G}(r)$$

on any compact subset of $(-\infty, b/2)$. Note that only on $(-\infty, b/2]$ the existence of $E\{\hat{\mathcal{G}}(s)^2\} = \mathcal{G}(2s)$ is guaranteed and because of it, the central limit theorem (CLT) cannot be applied outside of $(-\infty, b/2)$.

If $b < \infty$, and s is on the outside of $(-\infty, b/2)$ (i.e., $s \in (b/2, b)$), then by Marcinkiewicz-Zygmund strong law (see Theorem 3.2.3 of [65] or §6.7 of [35]),

$$\hat{\mathcal{G}}(s) - \mathcal{G}(s) = o(n^{-\delta}), \tag{1.2.2}$$

for any δ satisfying $0 \leq \delta < 1 - s/b < 1/2$. We note that this also holds for any $s \in (-\infty, b)$ even if $b = \infty$ with the condition $0 \leq \delta < 1/2$.

As mentioned in [31], for $s > b/2 (< \infty)$,, $n^{s/b-1}\left\{\hat{\mathcal{G}}(s) - \mathcal{G}(s)\right\}$ has a nondegenarate limit distribution if and only if e^{tG} is in some domain of attraction (and the limit distribution is stable. See §5 of Chapter 17 of [30] or §2 of Chapter 15 of [61]). Because we do not assume e^{tG} to follow a stable law, for $s \in (b/2, b)$, we will use (1.2.2) only to obtain the convergence rates of $D^k\hat{\mathcal{W}}(s)$ and $D^k\hat{\mathcal{K}}(s)$.

We now prove similar results for $\hat{\mathcal{W}}(s)$ and $\hat{\mathcal{K}}(s)$. Since we are interested in applying the saddlepoint approximation method, we first show that the the convergence strip of $\hat{\mathcal{W}}(s)$ converges to the convergence strip of $\mathcal{W}(s)$.

We first show that if c exists, then $\hat{c} \to c$ a.s. For this, we first need the following lemma.

Lemma 7. *Suppose the unique (and positive) root of the equation $\mathcal{D}'(s) = 0$ (, or $\mathcal{G}'(s) = \lambda^{-1}$), d exists on $(-\infty, b)$ and let $\hat{d}$ be the unique (and positive) root of equation $\hat{\mathcal{D}}'(s) = 0$ (, or $\hat{\mathcal{G}}'(s) = \hat{\lambda}^{-1}$). Then $\hat{d} \to d$ a.s.*

Proof. Because $\mathcal{G}'(s)$ is a strictly increasing function, for any $\varepsilon > 0$ satisfying $\varepsilon < d$ and $d + \varepsilon < b$, we have $\mathcal{G}'(d-\varepsilon) - \lambda^{-1} < 0 < \mathcal{G}'(d+\varepsilon) - \lambda^{-1}$. Since $\hat{\mathcal{G}}'(d \pm \varepsilon) \to \hat{\mathcal{G}}(d \pm \varepsilon)$ a.s. and $\hat{\lambda}^{-1} \to \lambda^{-1}$ a.s., we have

$$\hat{\mathcal{G}}'(d-\varepsilon) - \hat{\lambda}^{-1} < 0 < \hat{\mathcal{G}}'(d+\varepsilon) - \hat{\lambda}^{-1}$$

for all but finite n.

Therefore, $\hat{d} \in (d - \varepsilon, d + \varepsilon)$ for all but finite n and since $\varepsilon > 0$ can be arbitrary small, $\hat{d} \to d$ a.s. □

As we show in the proof of Lemma 2, $\mathcal{D}(s)$ has the maximum at d and $\mathcal{D}(s)$ is strictly decreasing for $s > d$ and c must be greater than d.

Lemma 8. *Suppose c, the positive root of $\mathcal{D}(s) = 0$, exists ($c \leq b$) and let $\hat{c}$ be the positive root of $\hat{\mathcal{D}}(s) = s + \hat{\lambda} - \hat{\lambda}\hat{\mathcal{G}}(s)$. Then $\hat{c} \xrightarrow{\text{a.s.}} c$.*

Proof. Suppose first that $c < b$. and we follow the proof of Theorem 3.10.1 (page 95) of [8]. For any ε satisfying $\varepsilon < b - c$ and $\varepsilon < c$, we have

$$\mathcal{D}(c - \varepsilon) > 0 > \mathcal{D}(c + \varepsilon)$$

and since $\hat{\mathcal{D}}(c \pm \varepsilon)$ converges to $\mathcal{D}(c \pm \varepsilon)$ a.s., we have

$$\hat{\mathcal{D}}(c - \varepsilon) > 0 > \hat{\mathcal{D}}(c + \varepsilon)$$

for any sufficiently large n in which, $\hat{c} \in (c - \varepsilon, c + \varepsilon)$. Because ε was arbitrary, $\hat{c}$ converges to c a.s.

For the case of $c = b$, or $b + \lambda - \lambda \lim_{s \nearrow b} \mathcal{G}(b) = 0$, we have $\lim_{s \nearrow b} \mathcal{G}(b) = b/\lambda + 1$, so that $\mathcal{G}(b)$ is convergent at b with the value $b/\lambda + 1$. Thus, $\hat{\mathcal{D}}(b) \to \mathcal{D}(b)$ a.s. too.

The above argument is valid for showing $\hat{c} > c - \varepsilon$ but cannot be used to show $\hat{c} < c + \varepsilon$ since $\mathcal{D}(c + \varepsilon)$ is not defined. For this, we use d which is defined in Lemma 7.

As remarked after Lemma 7, $d < c$ and since $\hat{d} \to d$ a.s., we have $\hat{d} < c$ for all but finite n. Therefore, $\hat{\mathcal{D}}(s)$ is decreasing on $[c, \infty)$ and $\hat{\mathcal{D}}(c + \varepsilon) < \hat{\mathcal{D}}(c)$ for all but finite n.

Because $\hat{\mathcal{D}}(c) \to \mathcal{D}(c) = 0$ a.s., we conclude that $\hat{\mathcal{D}}(c + \varepsilon) < 0$ for all but finite n, which implies $\hat{c} < c + \varepsilon$ for all but finite n. □

Theorem 9. *With the same notation as in Lemma 8, $\hat{c} - c = o(n^{-\delta})$ a.s. for any $0 \leq \delta < \min\{1/2, 1 - c/b\}$. If $c < b/2$, or $c = b/2$ and $\mathcal{G}(b) < \infty$, then*

$$\sqrt{n}(\hat{c} - c) \Rightarrow N\left[0, \lambda^2 \{1 - 2\mathcal{G}(c) + \mathcal{G}(2c)\} / \{1 - \lambda\mathcal{G}'(c)\}^2\right].$$

Proof. Suppose first $c \leq b/2$. Let $\{I_j\}$ be iid observations of the inter-arrival time with $\mathrm{Exp}\,(\lambda)$ distribution. By the multivariate CLT, if we set $\overline{I} = n^{-1}\sum_{j=1}^{n} I_j$ and $\overline{G} = n^{-1}\sum_{j=1}^{n} G_j$, we have

$$\sqrt{n}\left\{\begin{pmatrix} \overline{I} \\ \hat{\mathcal{G}}\,(s) \end{pmatrix} - \begin{pmatrix} \lambda^{-1} \\ \mathcal{G}\,(s) \end{pmatrix}\right\} \Rightarrow N\,(0, \Sigma)\,,$$

where

$$\Sigma = \begin{pmatrix} \lambda^{-2} & 0 \\ 0 & \mathcal{G}\,(2s) - \mathcal{G}\,(s)^2 \end{pmatrix}.$$

Define $h\,(x, y) = s + x^{-1} - x^{-1}y$ and note $h\{\lambda^{-1}, \mathcal{G}\,(s)\} = s + \lambda - \lambda\mathcal{G}\,(s) = \mathcal{D}\,(s)$ and $h\{\overline{I}, \hat{\mathcal{G}}\,(s)\} = s + \hat{\lambda} - \hat{\lambda}\hat{\mathcal{G}}\,(s) = \hat{\mathcal{D}}\,(s)$. By the multivariate delta method (Theorem 8.22 on page 61 of [42]), we have

$$\sqrt{n}\left\{\hat{\mathcal{D}}\,(s) - \mathcal{D}\,(s)\right\} = \sqrt{n}\left[h\{\overline{I}, \hat{\mathcal{G}}\,(s)\} - h\{\lambda^{-1}, \mathcal{G}\,(s)\}\right] \Rightarrow N\left[0, \lambda^2\left\{1 - 2\mathcal{G}\,(s) + \mathcal{G}\,(2s)\right\}\right].$$

By the mean value theorem, there exist c' in between $\hat{c}$ and c such that

$$\hat{c} - c = \frac{\hat{\mathcal{D}}\,(\hat{c}) - \hat{\mathcal{D}}\,(c)}{\hat{\mathcal{D}}'\,(c')} = \frac{0 - \hat{\mathcal{D}}\,(c)}{\hat{\mathcal{D}}'\,(c')} = \frac{\mathcal{D}\,(c) - \hat{\mathcal{D}}\,(c)}{\hat{\mathcal{D}}'\,(c')}. \tag{1.2.3}$$

Since $\hat{\mathcal{G}}'\,(s) \to \mathcal{G}'\,(s)$ locally uniformly a.s., it can be shown that $\hat{\mathcal{D}}'\,(s) = 1 - \hat{\lambda}\hat{\mathcal{G}}'\,(s)$ converges uniformly to $\mathcal{D}'\,(s) = 1 - \lambda\mathcal{G}'\,(s)$ on any compact subset around c. Since $c' \to c$ a.s., $\hat{\mathcal{D}}'\,(c') \to \mathcal{D}'\,(c)$ (see Theorem 7.3.5 of [66] or §0.1 of [53]). Thus, by Slutsky's lemma, we obtain

$$\sqrt{n}\,(\hat{c} - c) = \frac{-1}{\hat{\mathcal{D}}'\,(c')}\sqrt{n}\left\{\mathcal{D}\,(c) - \hat{\mathcal{D}}\,(c)\right\} \Rightarrow N\left[0, \lambda^2\frac{1 - 2\mathcal{G}\,(c) + \mathcal{G}\,(2c)}{\{1 - \lambda\mathcal{G}'\,(c)\}^2}\right]$$

Now, let $0 \leq \delta < \max\{1/2, 1 - c/b\}$. Since we assume $EI_i^2 < \infty$, for $0 \leq \delta < 1/2$, we obtain $\hat{\lambda} - \lambda = o(n^{-\delta})$ by Marcinkiewicz-Zygmund strong law and Proposition 35. As we noted earlier, $\hat{\mathcal{G}}(c) - \mathcal{G}(c) = o(n^{-\delta})$ for $0 \le \delta < \min\{1/2, 1 - c/b\}$. Because

$$\begin{aligned} \left|\hat{\mathcal{D}}(s) - \mathcal{D}(s)\right| &\le \left|\hat{\lambda} - \lambda\right| + \left|\hat{\lambda}\hat{\mathcal{G}}(s) - \hat{\lambda}\mathcal{G}(s)\right| + \left|\hat{\lambda}\mathcal{G}(s) - \lambda\mathcal{G}(s)\right| \\ &= \left|\hat{\lambda} - \lambda\right| + (|\lambda| + 1)\left|\hat{\mathcal{G}}(s) - \mathcal{G}(s)\right| + |\mathcal{G}(s)|\left|\hat{\lambda} - \lambda\right| \\ &= (|\mathcal{G}(s)| + 1)\left|\hat{\lambda} - \lambda\right| + (|\lambda| + 1)\left|\hat{\mathcal{G}}(s) - \mathcal{G}(s)\right| \qquad (1.2.4) \end{aligned}$$

we conclude $\hat{\mathcal{D}}(c) - \mathcal{D}(c) = o(n^{-\delta})$. Using (1.2.3) with the same argument about $\hat{\mathcal{D}}'(c')$, we obtain the desired result. □

Example 10. For M/M/1 queue with $G \sim$Exp(μ), which has the MGF, $\mathcal{G}(s) = (1 - s/\mu)^{-1}$. Therefore $b = \mu$ and solving $\mathcal{D}(s) = 0$ with respect s, we have $c = \mu - \lambda$. By Theorem 9, $\hat{c}$ is asymptotically normal if $\mu - \lambda < \mu/2$, or $\mu < 2\lambda$.

We later learned [1] that our c coincides with the adjustment coefficient (or the Lundberg exponent) in ruin probability theory and some similar results of ours were established already. See [33], [32], [55], or [8] for examples. Note that in the cited references, the arrival rate λ was assumed to be fixed and in [32] and [55], the mode of convergence of $\hat{c}$ is the convergence in probability. [33] was acknowledged in [8] as the original source of the theorem.

We now consider the behavior of $\hat{c}$ when c does not exist.

Proposition 11. [2]*If c does not exist, then $\hat{c} \to b$ a.s.*

Proof. As we showed in the proof of Lemma 4, the nonexistence of c implies either $b = 0$ or $\mathcal{D}(b) > 0$ (if $b > 0$). Suppose $b = 0$. Since $\mathcal{G}(\varepsilon) = \infty$ for any $\varepsilon > 0$, we have $\hat{\mathcal{G}}(\varepsilon) \nearrow \infty$ a.s. as $n \to \infty$. Thus, for all but finite n, $0 < \hat{c} < \varepsilon$. Because ε is arbitrary, we have $\hat{c} \to 0$ a.s.

Now suppose $b > 0$. Because $\hat{\mathcal{D}}(b) \to \mathcal{D}(b) > 0$ a.s., $\hat{c} > b$ all but finite n. Again, for any $\varepsilon > 0$, we have $\hat{\mathcal{G}}(b + \varepsilon) \to \mathcal{G}(b + \varepsilon) = \infty$, which implies $\hat{\mathcal{D}}(b + \varepsilon) \to \mathcal{D}(b + \varepsilon) = -\infty$. Thus, $b < \hat{c} < b + \varepsilon$ for all but finite n, which implies $\hat{c} \to b$ a.s. □

As we mentioned, for $\hat{\mathcal{W}}(s)$, the convergence strip is $(-\infty, \hat{c})$ since $\hat{c}$ always exists. Thus,

[1] Theorem 9 was originally suggested by Dr. Butler for me to prove.

[2] This is suggested by Dr. Butler and gave the idea of the proof.

the previous proposition with Lemma 8 shows that the convergence strip of $\hat{\mathcal{W}}(s)$ converges to the convergence strip of $\mathcal{W}(s)$.

Figure 1.2.2 shows boxplots and the histograms of 10^4 $\hat{c}$ of different sample sizes $n = 50$, 200, 10^3, 10^4, 10^5 with Pareto(0.8,5) service distribution and $\hat{\lambda} = .5$ (i.e., $\{I_j\}$ were not used and only $\{G_j\}_{j=1}^n$ were used). One can see that though the maximum and the the interquartile range of $\{\hat{c}\}$ are decreasing as n increases, the convergence of $\hat{c}$ to $0(= b)$ would be very slowly. Note that since $\mathcal{G}(s)$ is convergent on $(-\infty, 0]$ only, we do not have any stochastic order of the convergence for $\hat{c}$.

We now show that $\hat{\mathcal{W}}(s)$, $\hat{\mathcal{K}}(s)$, $\hat{\mathcal{K}}'(s)$, and $\hat{\mathcal{K}}''(s)$ are reasonable estimators. First, we have:

Lemma 12. *For each $s \in (-\infty, c)$ and $k \in \mathbb{N}_0$,*

$$D^k\hat{\mathcal{W}}(s) \xrightarrow{\text{a.s.}} D^k\mathcal{W}(s)$$
$$D^k\hat{\mathcal{K}}(s) \xrightarrow{\text{a.s.}} D^k\mathcal{K}(s),$$

as $n \to \infty$.

Proof. Let $\{I_j\}$ be iid of the inter-arrival time with $\operatorname{Exp}(\lambda)$ distribution and $\{G_j\}$ be iid of the stationary service time distribution with mean μ^{-1} and the variance σ_G^2. By the strong law of large number (SLLN), $\overline{I} \xrightarrow{\text{a.s.}} \lambda^{-1}$, $\overline{G} \xrightarrow{\text{a.s.}} \mu^{-1}$, and $D^k\hat{\mathcal{G}}(s) \xrightarrow{\text{a.s.}} D^k\mathcal{G}(s)$ for each $k \in \mathbb{N}_0$. For fixed $s \in (-\infty, c)$, define $h : \mathbb{R}^3 \to \mathbb{R}$ by $h(x, y, z) = \left(1 - yx^{-1}\right) sz / \left(s + x^{-1} - x^{-1}z\right)$, which is continuous.

Thus, by the continuous mapping theorem for the almost sure convergence,

$$\hat{\mathcal{W}}(s) = h\{\overline{I}, \overline{G}, \hat{\mathcal{G}}(s)\} \xrightarrow{\text{a.s.}} h\left\{\lambda^{-1}, \mu^{-1}, \mathcal{G}(s)\right\} = \mathcal{W}(s).$$

The other cases can be shown in similar way. □

Theorem 13. *For $k \in \mathbb{N}_0$, $D^k\hat{\mathcal{W}}$ and $D^k\hat{\mathcal{K}}$ converges to $D^k\mathcal{W}$ and $D^k\hat{\mathcal{K}}_n$ respectively locally*

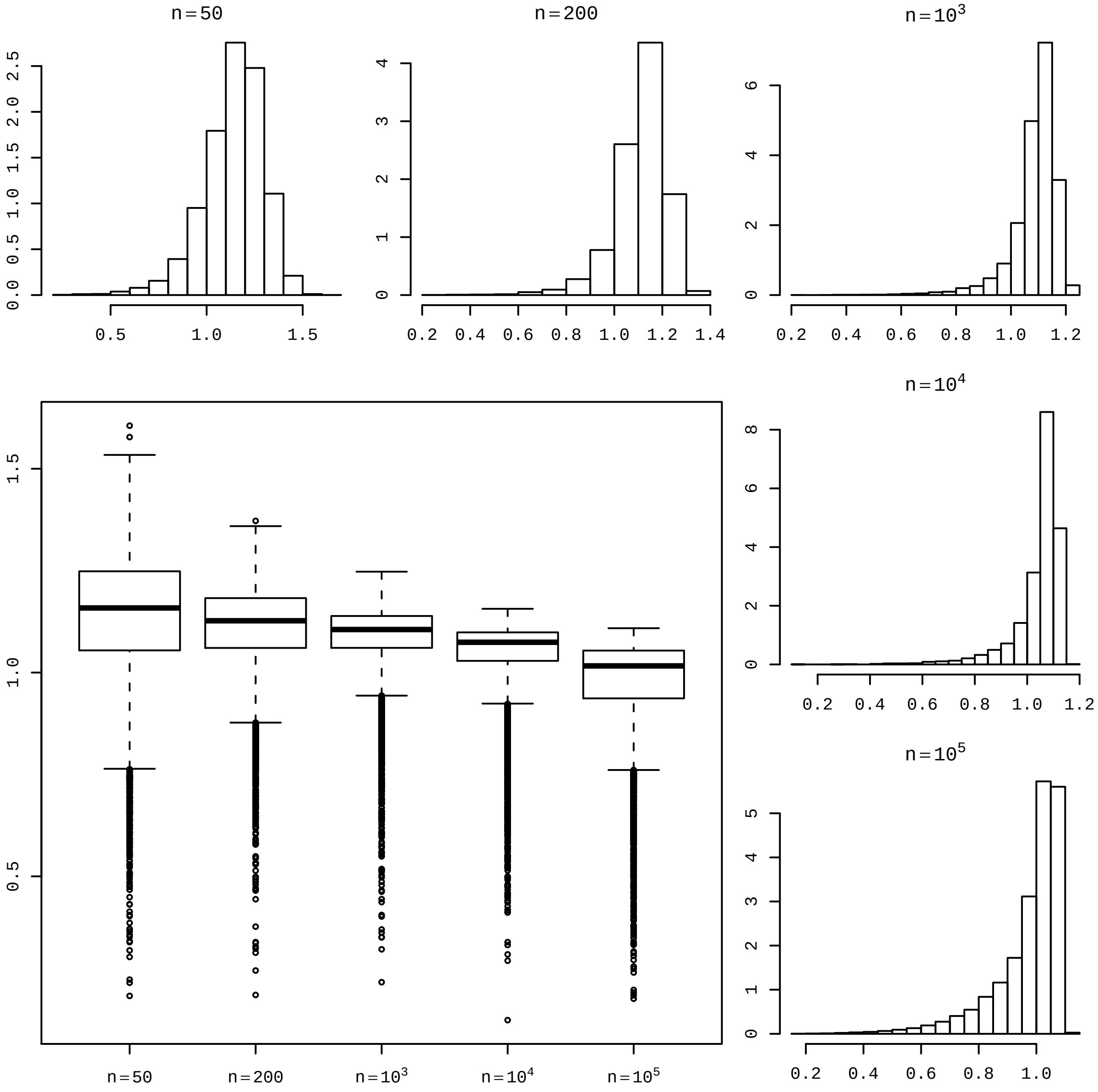

Figure 1.2.2: The boxplots and the histograms of 10^4 $\hat{c}$ from $\{G_j\}_{j=1}^{n}$ with the different sample sizes $n = 50$, 200, 10^3, 10^4, 10^5 with Pareto(0.8,5) service distribution. $\hat{\lambda}$ is fixed to be 0.5

uniformly a.s. on $(-\infty, c)$ *(i.e., for any* $[s_1, s_2] \subset (-\infty, c)$*, and* $k \in \mathbb{N}_0$*,*

$$\sup_{s_1 \le s \le s_2} \left| D^k \hat{\mathcal{W}}(s) - D^k \mathcal{W}(s) \right| \xrightarrow{a.s.} 0$$
$$\sup_{s_1 \le s \le s_2} \left| D^k \hat{\mathcal{K}}(s) - D^k \mathcal{K}(s) \right| \xrightarrow{a.s.} 0).$$

Proof. We use the same notation as in the proof of Lemma 12. Let $[s_1, s_2]$ be fixed. As a MGF of the stationary waiting time distribution when the service time distribution follows the empirical CDF $n^{-1} \sum_{i=1}^{n} \mathbf{1}_{[0,G_i]}(t)$, $\hat{\mathcal{W}}(s)$ is C^∞, positive, and non-decreasing on $(-\infty, \hat{c})$[3].

Since $\hat{c} \xrightarrow{\text{a.s.}} c > s_2$, by discarding first finite $\hat{\mathcal{W}}$, we can assume that $\hat{c} > s_2$ a.s.

We first show that $\hat{\mathcal{W}}(s)$ are (uniformly) equicontinuous on $[s_1, s_2]$. Fix n_0 such that $|\hat{\mathcal{W}}'(s_2) - \mathcal{W}'(s_2)| < 1$ for any $n \ge n_0$. Then, by the mean value theorem and the non-decreasing property of $\hat{\mathcal{W}}'(s)$ ($\hat{\mathcal{W}}''(s) > 0$), we have that for any $n \ge n_0$ and $r, s \in [s_1, s_2]$,

$$\left| \hat{\mathcal{W}}(r) - \hat{\mathcal{W}}(s) \right| \le \left| \hat{\mathcal{W}}'(s_2) \right| |r - s| \le \left(\mathcal{W}'(s_2) + 1 \right) |r - s|,$$

which give us the (uniform) equicontinuity.

Now, we claim that on $[s_1, s_2]$, $\hat{\mathcal{W}}(s) \to \mathcal{W}(s)$ a.s. Let $\mathbb{Q}$ be the set of all rational numbers. By Lemma 12, and the countable sub-additivity of probability measure, for any $s \in \mathbb{Q} \cap (-\infty, c)$, $\hat{\mathcal{W}}(s) \to \mathcal{W}(s)$ a.s. Let $s \in [s_1, s_2]$ and pick $s_n \in \mathbb{Q} \cap (-\infty, c)$ such that $s_n \to s$. Then,

$$\left| \hat{\mathcal{W}}(s) - \mathcal{W}(s) \right| \le \left| \hat{\mathcal{W}}(s) - \hat{\mathcal{W}}(s_n) \right| + \left| \hat{\mathcal{W}}(s_n) - \mathcal{W}(s_n) \right| + \left| \mathcal{W}(s_n) - \mathcal{W}(s) \right|.$$

The first term on the right hand side of the above inequality can be majorized since $\hat{\mathcal{W}}(s)$ are (uniformly) equicontinuous. The second term can be majorized since $\hat{\mathcal{W}}(s_n) \xrightarrow{\text{a.s.}} \mathcal{W}(s_n)$ and the last term can be majorized since $\mathcal{W}(s)$ is continuous.

To be precise, let $\varepsilon > 0$ be given. Pick m_0 such that $|\mathcal{W}(s_m) - \mathcal{W}(s)| < \varepsilon/3$ for any

[3] In [31], the author use problem 13-A-7 of [54], whose proof was not provided in [54]. Even if the same proof can be used here, we use a different way.

$m \geq m_0$. Also, pick m_1 such that $m \geq m_1$ implies $|\hat{\mathcal{W}}(s) - \hat{\mathcal{W}}(s_m)| < \varepsilon/3$ for any n. Then, finally, set $m = \max\{m_0, m_1\}$ and pick n_0 such that $|\hat{\mathcal{W}}_n(s_m) - \mathcal{W}(s_m)| < \varepsilon/3$ for any $n \geq n_0$. Thus, $|\hat{\mathcal{W}}(s) - \mathcal{W}(s)| < \varepsilon$ for any $n \geq n_0$. Since $\varepsilon > 0$ is arbitrary, we proved the claim.

From here, we can reach the desired result in several ways. We may use $\hat{\mathcal{W}}(s)$ is convex since the point-wise convergence of convex functions implies locally uniform convergence (Theorem E of [54] or theorem 3.1.4 of [36]), which cannot be used generally in other cases.

We may use that pointwise convergence of equicontinuous function implies the locally uniform convergence (Exercise 7.16 of [58]) or, we may use that the pointwise convergence of non-decreasing functions implies the locally uniform convergence (Theorem 37), both of which can be used for $D^k\hat{\mathcal{W}}$ cases.

Since $\log x$ is a smooth function, locally uniform a.s. convergence of $\hat{\mathcal{W}}(s)$ transferred to $\hat{\mathcal{K}}(s) = \log \hat{\mathcal{W}}(s)$ too. For $k \geq 1$, we note that

$$D^k\hat{\mathcal{K}}(s) = h\{\hat{\mathcal{W}}(s), \hat{\mathcal{W}}'(s), \cdots, D^k\hat{\mathcal{W}}(s)\}/\{\hat{\mathcal{W}}(s)\}^{2k},$$

where $h(x_1, \cdots, x_k)$ is a polynomial of $x_1, \cdots, x_k$. Thus, locally uniform a.s. convergences of $D^k\hat{\mathcal{W}}(s)$ transferred to $D^k\hat{\mathcal{K}}(s)$ for any k by the argument like the delta method. □

Lemma 14. *The finite dimensional distributions of $\sqrt{n}\{\hat{\mathcal{W}}(s) - \mathcal{W}(s)\}$ converges weakly to multivariate normal distributions on $(-\infty, \min\{c, b/2\})$. If $s \in (b/2, c)$, then $\{\hat{\mathcal{W}}(s) - \mathcal{W}(s)\} = o\left(n^{-\delta}\right)$ a.s. for any $0 \leq \delta < \min\{1/2, 1 - s/b\}$.*

Proof. We use the multivariate CLT and the multivariate delta method to establish the weak convergence of finite dimensional distributions. For brevity, we only show 2-dimensional convergence. The general cases can be shown in a similar way.

By the multivariate CLT, we have

$$\sqrt{n}\left\{\begin{pmatrix}\overline{I}\\ \overline{G}\\ \hat{\mathcal{G}}(s)\\ \hat{\mathcal{G}}(r)\end{pmatrix}-\begin{pmatrix}\lambda^{-1}\\ \mu^{-1}\\ \mathcal{G}(s)\\ \mathcal{G}(r)\end{pmatrix}\right\}\Rightarrow N(0,\Sigma),$$

where

$$\Sigma=\begin{pmatrix}\lambda^{-2} & 0 & 0 & 0\\ 0 & \sigma_G^2 & \mathcal{G}'(s)-\mu^{-1}\mathcal{G}(s) & \mathcal{G}'(r)-\mu^{-1}\mathcal{G}(r)\\ 0 & \mathcal{G}'(s)-\mu^{-1}\mathcal{G}(s) & \mathcal{G}(2s)-\mathcal{G}(s)^2 & \mathcal{G}(s+r)-\mathcal{G}(s)\mathcal{G}(r)\\ 0 & \mathcal{G}'(r)-\mu^{-1}\mathcal{G}(r) & \mathcal{G}(s+r)-\mathcal{G}(s)\mathcal{G}(r) & \mathcal{G}(2r)-\mathcal{G}(r)^2\end{pmatrix}.$$

For fixed s and r, define functions $h_i(x,y,z_1,z_2)$ $(i=1,2)$ by

$$h_1(x,y,z_1,z_2)=\frac{\left(1-yx^{-1}\right)sz_1}{s+x^{-1}-x^{-1}z_1}$$
$$h_2(x,y,z_1,z_2)=\frac{\left(1-yx^{-1}\right)rz_2}{r+x^{-1}-x^{-1}z_2}.$$

Observe that $\hat{\mathcal{W}}(s)=h_1\{\overline{I},\overline{G},\hat{\mathcal{G}}(s),\hat{\mathcal{G}}(r)\}$ and $\hat{\mathcal{W}}(r)=h_2\{\overline{I},\overline{G},\hat{\mathcal{G}}(s),\hat{\mathcal{G}}(r)\}$. Also, note that $\mathcal{W}(s)=h_1\{\lambda^{-1},\mu^{-1},\mathcal{G}(s),\mathcal{G}(r)\}$ and $\mathcal{W}(r)=h_2\{\lambda^{-1},\mu^{-1},\mathcal{G}(s),\mathcal{G}(r)\}$. By the multivariate delta method, $\sqrt{n}[\{\hat{\mathcal{W}}(s),\hat{\mathcal{W}}(r)\}^\intercal-\{\mathcal{W}(s),\mathcal{W}(r)\}^\intercal]$ converges weakly to $N(0,\Sigma')$, where

$$\Sigma'=\mathbf{J_h}\left\{\lambda^{-1},\mu^{-1},\mathcal{G}(s),\mathcal{G}(r)\right\}\Sigma\mathbf{J_h}\left\{\lambda^{-1},\mu^{-1},\mathcal{G}(s),\mathcal{G}(r)\right\}^\intercal.$$

Here, $\mathbf{J_h}\{\lambda^{-1},\mu^{-1},\mathcal{G}(s),\mathcal{G}(r)\}$ denotes the 2×4 Jacobin matrix of $\mathbf{h}=(h_1,h_2)$ evaluated at $\{\lambda^{-1},\mu^{-1},\mathcal{G}(s),\mathcal{G}(r)\}$.

Now, let $0\le\delta<\min\{1/2,1-c/b\}$. By Marcinkiewicz-Zygmund strong law, we have

$$\{\overline{I},\overline{G},\hat{\mathcal{G}}(s),\hat{\mathcal{G}}(t)\}-\left\{\lambda^{-1},\mu^{-1},\mathcal{G}(s),\mathcal{G}(t)\right\}=o\left(n^{-\delta}\right)\quad\text{a.s.}$$

Apply Proposition 35 in the appendix to h_1, we obtain the desired result. □

Since we do not need multi-dimensional asymptotic normality of our estimator for our applications of saddlepoint approximation, we now concentrate on 1-dimensional asymptotic normality of our estimators.

However, we note that it can be shown that $\sqrt{n}\{D^k\hat{\mathcal{W}}(s)-D^k\mathcal{W}(s)\}$ and $\sqrt{n}\{D^k\hat{\mathcal{K}}(s)-D^k\mathcal{K}(s)\}$ converge weakly to Gaussian processes on $C_{[s_1,s_2]}$ for any $[s_1,s_2]\subset(-\infty,\min\{c,b/2\})$ (see Theorem 24 and the following remark), where $C_{[s_1,s_2]}$ is the metric space of continuous functions on $[s_1,s_2]$ with the supremum (uniform) norm.

Since $\log x$ is a smooth function, the asymptotic normality of $\sqrt{n}\{\hat{\mathcal{W}}(s)-\mathcal{W}(s)\}$ can be transferred to $\hat{\mathcal{K}}(s)=\log\hat{\mathcal{W}}(s)$ by the 1-dimensional delta method and we may use Proposition 35 for a.s. convergence of the order, $o\left(n^{-\delta}\right)$:

Corollary 15. *$\sqrt{n}\{\hat{\mathcal{K}}(s)-\mathcal{K}(s)\}$ converges weakly to $N\{0,\sigma^2_{\mathcal{K}}(s)\}$ on $(-\infty,\min\{c,b/2\})$, where*

$$\sigma^2_{\mathcal{K}}(s)=\sigma^2_{\mathcal{W}}(s)\{\mathcal{W}(s)\}^{-2}.$$

If $s\in(b/2,c)$, then $\{\hat{\mathcal{K}}(s)-\mathcal{K}(s)\}=o\left(n^{-\delta}\right)$ a.s. for any $0\le\delta<\min\{1/2,1-s/b\}$.

For simplicity, unlike Lemma (14), we concentrate on the 1-dimensional asymptotic normality of $\hat{\mathcal{K}}'(s)$ and $\hat{\mathcal{K}}''(s)$, respectively.

Theorem 16. *Let $\hat{\mathcal{K}}'(s)$ be as in (1.1.6) and $\hat{\mathcal{K}}''(s)=\partial\hat{\mathcal{K}}'(s)/\partial s$. Then $\sqrt{n}\{\hat{\mathcal{K}}'(s)-\mathcal{K}'(s)\}\Rightarrow N\{0,\sigma^2_{\mathcal{K}'}(s)\}$ and $\sqrt{n}\{\hat{\mathcal{K}}''(s)-\mathcal{K}''(s)\}\Rightarrow N\{0,\sigma^2_{\mathcal{K}''}(s)\}$ on any $s\in(-\infty,\min\{c,b/2\})$, where $\sigma^2_{\mathcal{K}'}$ and $\sigma^2_{\mathcal{K}''}$ are as in the proof. If $s\in(b/2,c)$, $\{\hat{\mathcal{K}}'(s)-\mathcal{K}'(s)\}=o\left(n^{-\delta}\right)$ a.s. and $\{\hat{\mathcal{K}}''(s)-\mathcal{K}''(s)\}=o\left(n^{-\delta}\right)$ a.s. for any $0\le\delta<\min\{1/2,1-s/b\}$.*

Proof. For fixed s, $(e^{sG_i},G_ie^{sG_i},G_i^2e^{sG_i})$ are iid with the mean $(\mathcal{G}(s),\mathcal{G}'(s),\mathcal{G}''(s))$. The covariance can be calculated as

$$\operatorname{Cov}\left(e^{sG_i},G_ie^{sG_i}\right)=E\left(e^{sG_i}G_ie^{sG_i}\right)-Ee^{sG_i}E\left(G_ie^{sG_i}\right)=\mathcal{G}'(2s)-\mathcal{G}(s)\mathcal{G}'(s),$$

so that we have the covariance matrix $\Sigma = (\sigma_{ij})$, where

$$\sigma_{ij} = \mathcal{G}^{(i+j-2)}(2s) - \mathcal{G}^{(i-1)}(s)\,\mathcal{G}^{(j-1)}(s)\,.$$

By multivariate CLT, we have

$$\sqrt{n}\left\{\begin{pmatrix} \hat{\mathcal{G}}(s) \\ \hat{\mathcal{G}}'(s) \\ \hat{\mathcal{G}}''(s) \end{pmatrix} - \begin{pmatrix} \mathcal{G}(s) \\ \mathcal{G}'(s) \\ \mathcal{G}''(s) \end{pmatrix}\right\} \Rightarrow N(0,\Sigma)$$

and by CLT and the delta method,

$$\sqrt{n}\left(\hat{\lambda} - \lambda\right) \Rightarrow N\left(0,\lambda^2\right)$$

and note that $\hat{\lambda}$ is independent of $\hat{\mathcal{G}}^{(k)}(s)$ for any $k = 0, 1, \ldots$. Now, define $h_1(x_1, x_2, x_3)$ and $h_2(x_1, x_2, x_3, x_4)$ by

$$\begin{aligned} h_1(x_1,x_2,x_3) =& \frac{x_1(1-x_2)x_2 + s(s+x_1)x_3}{sx_2(s+x_1-x_1x_3)} \\ h_2(x_1,x_2,x_3,x_4) =& \frac{1}{\{sx_2(s+x_1-x_1x_2)\}^2}\Big(2x_1(s+x_1)x_2^3 \\ & - x_1^2x_2^4 - \{s(s+x_1)x_3\}^2 + s^2(s+x_1)x_2\left\{2x_1x_3^2 + (s+\lambda)x_4\right\} \\ & - x_1x_2^2\left[2s+\lambda+s^2\left\{2x_3+(s+x_1)x_4\right\}\right]\Big) \end{aligned}$$

and note that $h_1\{\lambda, \mathcal{G}(s), \mathcal{G}'(s)\} = \{\log \mathcal{W}(s)\}'$ and $h_2\{\lambda, \mathcal{G}(s), \mathcal{G}'(s), \mathcal{G}''(s)\} = \{\log \mathcal{W}(s)\}''$. By the multivariate CLT, $\sqrt{n}\{\hat{\mathcal{K}}'(s) - \mathcal{K}'(s)\} \Rightarrow N\{0, \sigma_{\mathcal{K}'}(s)\}$, where using

$$Q = s + \lambda - \lambda\mathcal{G}(s)\,,$$

$\sigma^2_{\mathcal{K}'}(s)$ can be written as

$$\begin{aligned}
\sigma^2_{\mathcal{K}'}(s) =& \lambda^2 \left\{ \frac{\partial h_1}{\partial x_1} \left(\lambda, \mathcal{G}(s), \mathcal{G}'(s)\right) \right\}^2 \\
&+ \sum_{i=1}^{2} \sum_{j=1}^{2} \sigma_{ij} \frac{\partial h_1}{\partial x_{i+1}} \left\{\lambda, \mathcal{G}(s), \mathcal{G}'(s)\right\} \cdot \frac{\partial h_1}{\partial x_{j+1}} \left\{\lambda, \mathcal{G}(s), \mathcal{G}'(s)\right\} \\
=& \frac{1}{Q^4} \Bigg(\frac{(s+\lambda)^4 \mathcal{G}(2s)\, \mathcal{G}'(s)^2}{\mathcal{G}(s)^4} + 2\lambda^2 \mathcal{G}(s) \left\{-1 + \lambda \mathcal{G}'(s)\right\} \\
&+ \frac{2(s+\lambda)^3 \mathcal{G}'(s) \left\{-2\lambda \mathcal{G}(2s)\, \mathcal{G}'(s) - (s+\lambda) \mathcal{G}'(2s)\right\}}{\mathcal{G}(s)^3} \\
&+ \frac{2\lambda(s+\lambda)^2 \mathcal{G}'(s) \left[\mathcal{G}(2s) \left\{1 + 2\lambda \mathcal{G}'(s)\right\} + 3(s+\lambda) \mathcal{G}'(2s)\right] + (s+\lambda)^4 \mathcal{G}''(2s)}{\mathcal{G}(s)^2} \\
&+ \frac{2\lambda(s+\lambda) \left(-2\lambda \mathcal{G}(2s)\, \mathcal{G}'(s) + (s+\lambda) \left[\left\{-1 - 2\lambda \mathcal{G}'(s)\right\} \mathcal{G}'(2s) - (s+\lambda) \mathcal{G}''(2s)\right]\right)}{\mathcal{G}(s)} \\
&+ \lambda^2 [1 + \mathcal{G}(2s) + 2s \mathcal{G}'(s) - \lambda(2s+\lambda) \mathcal{G}'(s)^2 + (s+\lambda)\{2\mathcal{G}'(2s) + (s+\lambda)\mathcal{G}''(2s)\}] \Bigg).
\end{aligned}$$

By the similar way, $\sqrt{n}\{\hat{\mathcal{K}}''(s) - \mathcal{K}''(s)\} \Rightarrow N\{0, \sigma^2_{\mathcal{K}''}(s)\}$, where using

$$\begin{aligned}
A =& 2[\lambda \mathcal{G}(s)^2 + (s+\lambda) \left\{s + \lambda - 2\lambda \mathcal{G}(s)\right\} \mathcal{G}'(s)] \\
B =& 2\lambda \mathcal{G}(s)^3 - 4\lambda^2 \mathcal{G}(s)^3 \mathcal{G}'(s) \\
&+ 2(s+\lambda)\{(s+\lambda)^2 - 3\lambda(s+\lambda)\mathcal{G}(s) + 3\lambda^2 \mathcal{G}(s)^2\}\mathcal{G}'(s)^2 \\
&- (s+\lambda)\mathcal{G}(s) \left\{s + \lambda - 2\lambda \mathcal{G}(s)\right\} \left\{s + \lambda - \lambda \mathcal{G}(s)\right\} \mathcal{G}''(s),
\end{aligned}$$

$\sigma^2_{\mathcal{K}''}(s)$ can be written as

$$\begin{aligned}\sigma^2_{\mathcal{K}''}(s) =& \lambda^2 \left[\frac{\partial h_2}{\partial x_{i+1}} \left\{\lambda, \mathcal{G}(s), \mathcal{G}'(s), \mathcal{G}''(s)\right\}\right]^2 \\ & + \sum_{i=1}^{3}\sum_{j=1}^{3} \sigma_{ij} \frac{\partial h_2}{\partial x_{i+1}} \left\{\lambda, \mathcal{G}(s), \mathcal{G}'(s), \mathcal{G}''(s)\right\} \cdot \frac{\partial h_2}{\partial x_{j+1}} \left\{\lambda, \mathcal{G}(s), \mathcal{G}'(s)' \mathcal{G}''(s)\right\} \\ =& \frac{1}{Q^6}\Bigg(\lambda^2 \left[-2\left\{-1+\mathcal{G}(s) - s\mathcal{G}'(s)\right\}\left\{-1+\lambda\mathcal{G}'(s)\right\} + Qs\mathcal{G}''(s)\right]^2 \\ & \quad + \frac{B(B\sigma_{11} - QA\mathcal{G}(s)\sigma_{12} + Q^2(s+\lambda)\mathcal{G}(s)^2\sigma_{13})}{\mathcal{G}(s)^6} \\ & \quad - \frac{QA(B\sigma_{12} - QA\mathcal{G}(s)\sigma_{22} + Q^2(s+\lambda)\mathcal{G}(s)^2\sigma_{23})}{\mathcal{G}(s)^5} \\ & \quad + \frac{Q^2(s+\lambda)(B\sigma_{13} - QA\mathcal{G}(s)\sigma_{23} + Q^2(s+\lambda)\mathcal{G}(s)^2\sigma_{33})}{\mathcal{G}(s)^4}\Bigg).\end{aligned}$$

Now, let $0 \le \delta < \min\{1/2, 1 - c/b\}$. Apply Proposition 35 to g and h, we obtain the desired result. □

1.2.3 Asymptotic results of the saddlepoint approximation

Lemma 17. *Let t be in the range of $\mathcal{K}'(s)$ determined in Proposition 5, where the saddlepoint equation for $\mathcal{K}'(s_t) = t$ can be solved and $s_t \in (-\infty, c)$ is the unique solution. Then, for large enough $n \in \mathbb{N}$, there exists a sequence $\hat{s}_t$ such that $\hat{\mathcal{K}}'(\hat{s}_t) = t$ and $\hat{s}_t \xrightarrow{a.s.} s_t$.*

Proof. Fix t in the range determined in Proposition 5, and recall that saddlepoint equation for $\hat{\mathcal{K}}(s_t) = t$ can be solved if $t > G_{(1)}$. Since the empirical CDF of $\{G_i\}$ at t is 0 if and only if $t < G_{(1)}$, the Glivenko-Cantelli theorem implies $G_{(1)} \xrightarrow{\text{a.s.}} g_0 = \inf\{t : G(t) > 0\}$. Thus, $\hat{s}_t$ is well-defined for large enough $n \in \mathbb{N}$.

Let ε be any real number satisfying $0 < \varepsilon < \min\{c, b\} - s_t$, so that $\mathcal{K}'(s_t + \varepsilon)$ is convergent. Recall that $\mathcal{K}'(s)$ (and $\hat{\mathcal{K}}'(s)$) is strictly increasing, we have

$$\mathcal{K}'(s_t - \varepsilon) < \mathcal{K}'(s_t) = t < \mathcal{K}'(s_t + \varepsilon).$$

Since $\hat{\mathcal{K}}'(s_t \pm \varepsilon) \to \mathcal{K}'(s_t \pm \varepsilon)$ a.s., for large enough n, we have

$$\hat{\mathcal{K}}'\left(s_t - \varepsilon\right) < t = \hat{\mathcal{K}}'\left(\hat{s}_t\right) < \hat{\mathcal{K}}'\left(s_t + \varepsilon\right),$$

which implies

$$s_t - \varepsilon < \hat{s}_t < s_t + \varepsilon.$$

Because ε can be arbitrarily small, we conclude $\hat{s}_t \to s_t$ a.s. □

Corollary 18. *Let t be in the range determined in Proposition 5, where the saddlepoint equation for $\mathcal{K}'\left(s_t\right) = \log \mathcal{W}\left(s_t\right) = t$ can be solved, then*

1. $\hat{f}_0\left(t\right) \xrightarrow{a.s.} f_0\left(t\right)$.

2. $\hat{F}_0\left(t\right) \xrightarrow{a.s.} F_0\left(t\right)$ *for* $t \neq EW$. *For* $t = EW$, *we have*

$$\hat{F}_0\{\hat{\mathcal{K}}_n'\left(0\right)\} := \frac{1}{2} + \frac{\hat{\mathcal{K}}'''\left(0\right)}{6\sqrt{2\pi}\{\hat{\mathcal{K}}''\left(0\right)\}^{3/2}} \xrightarrow{a.s.} F_0\left(EW\right).$$

Proof. By Lemma 17, $\hat{f}_0\left(t\right)$ and $\hat{F}\left(t\right)$ are well-defined for large enough n. For given t, since $\hat{\mathcal{K}}''\left(s\right) \to \mathcal{K}''\left(s\right)$ and $\hat{\mathcal{K}}\left(s\right) \to \mathcal{K}\left(s\right)$ locally uniformly and $\hat{s}_t \xrightarrow{\text{a.s.}} s_t$, we have $\hat{\mathcal{K}}''(\hat{s}_t) \xrightarrow{\text{a.s.}} \mathcal{K}''(s_t)$ and $\hat{\mathcal{K}}(\hat{s}_t) \xrightarrow{\text{a.s.}} \mathcal{K}(s_t)$. Thus, by the continuous mapping theorem for a.s. convergence, we have

$$\hat{f}\left(t\right) = \frac{1}{\sqrt{2\pi\hat{\mathcal{K}}''\left(\hat{s}_t\right)}} \exp\{\hat{\mathcal{K}}\left(\hat{s}_t\right) - \hat{s}_t t\} \xrightarrow{\text{a.s.}} \frac{1}{\sqrt{2\pi\mathcal{K}''\left(s_t\right)}} \exp\left\{\mathcal{K}\left(s_t\right) - s_t t\right\} = f_0\left(t\right).$$

Now, for $\hat{F}_0\left(t\right)$, the same argument holds if $t \neq EW$, which implies $s_t \neq 0$. The a.s. convergence of $\hat{F}_0\{\hat{\mathcal{K}}'\left(0\right)\}$ to $F_0\left\{\mathcal{K}'\left(0\right)\right\}$ comes from SLLN and the continuous mapping theorem for a.s. convergence as in the remark after Lemma 23. □

Theorem 19. *Assuming the condition of Lemma 17, for $s_t \in (-\infty, \min\{c, b/2\})$, we have $\sqrt{n}\{\hat{s}_t - s_t\} \Rightarrow N(0, \sigma_{s_t}^2)$, where $\sigma_{s_t}^2$ is shown in the proof. If $s_t \in (b/2, c)$, then $\hat{s}_t - s_t = o\left(n^{-\delta}\right)$ a.s. for any $0 \leq \delta < \min\{1/2, 1 - s_t/b\}$.*

Proof. From Theorem 16, $\sqrt{n}\{\hat{\mathcal{K}}'(s)-\mathcal{K}'(s)\} \Rightarrow N\{0,\sigma^2_{\mathcal{K}'}(s)\}$ on any $s \in (-\infty, \min\{b/2, c\})$. By the mean value theorem, there exist $\hat{r}$ in between $\hat{s}_t$ and s_t such that

$$\hat{s}_t - s_t = \frac{\hat{\mathcal{K}}'(\hat{s}_t) - \hat{\mathcal{K}}'(s_t)}{\hat{\mathcal{K}}''(\hat{r})} = \frac{t - \hat{\mathcal{K}}'(s_t)}{\hat{\mathcal{K}}''(\hat{r})} = \frac{\mathcal{K}'(s_t) - \hat{\mathcal{K}}'(s_t)}{\hat{\mathcal{K}}''(\hat{r})}. \tag{1.2.5}$$

Since $\hat{s}_t \xrightarrow{\text{a.s.}} s_t$, we have $\hat{r} \xrightarrow{\text{a.s.}} s_t$. Thus a.s. locally uniform convergence of $\hat{\mathcal{K}}''(s) \to \mathcal{K}''(s)$ implies $\hat{\mathcal{K}}''(\hat{r}) \to \mathcal{K}''(s_t)$. Thus, by Slutsky's lemma, we obtain

$$\sqrt{n}(\hat{s}_t - s_t) = \frac{-1}{\hat{\mathcal{K}}''(\hat{r})}\sqrt{n}\{\hat{\mathcal{K}}'(s_t) - \mathcal{K}'(s_t)\} \Rightarrow N\left[0, \frac{\sigma^2_{\mathcal{K}'}(s_t)}{\{\mathcal{K}''(s_t)\}^2}\right].$$

Now, suppose $0 \le \delta < \min\{1/2, 1 - c/b\}$. Applying Theorem 16 to (1.2.5), we obtain the desired result. □

Theorem 20. *If we assume the condition of Lemma 17. Then*

1. $\hat{f}_0(t) - f_0(t) = O_p(n^{-1/2})$ *for* $s_t \in (-\infty, \min\{c, b/2\})$. *If* $s_t \in (b/2, c)$, *then* $\hat{f}_0(t) - f_0(t) = o(n^{-\delta})$ *for* $0 \le \delta < \min\{1/2, 1 - s_t/b\}$.

2. *If* $t \neq EW$, *then* $\hat{F}_0(t) - F_0(t) = O_p(n^{-1/2})$ *for* $s_t \in (-\infty, \min\{c, b/2\})$ *and* $\hat{F}_0(t) - F_0(t) = o(n^{-\delta})$ *for* $s_t \in (b/2, c)$ *and* $0 \le \delta < \min\{1/2, 1 - s_t/b\}$. *If* $t = EW$, *then* $\hat{F}_0\{\hat{\mathcal{K}}'(0)\} - F_0(t) = O_p(n^{-1/2})$.

Proof. 1. We first show $O_p(n^{-1/2})$ part. Since $\exp x$ is a C^∞ function on $(0,\infty)$, by Proposition 36 , it suffice to show $\log\hat{f}_0(t) - \log f_0(x) = O_p(n^{-1/2})$. We have

$$\begin{aligned}\log\hat{f}_0(t) - \log f_0(t) = &-\frac{1}{2}\{\log\hat{\mathcal{K}}''(\hat{s}_t) - \log\mathcal{K}''(s_t)\} \\ &+ \{\hat{\mathcal{K}}(\hat{s}_t) - \mathcal{K}(s_t)\} - t(\hat{s}_t - s_t).\end{aligned}$$

We note that

$$\begin{aligned}\hat{\mathcal{K}}(\hat{s}_t) - \mathcal{K}(s_t) &= \{\hat{\mathcal{K}}(\hat{s}_t) - \hat{\mathcal{K}}(s_t)\} + \{\hat{\mathcal{K}}(s_t) - \mathcal{K}(s_t)\} \\ &= \hat{\mathcal{K}}'(\hat{r})(\hat{s}_t - s_t) + O_p(n^{-1/2}),\end{aligned}$$

where $\hat{r}$ is in between $\hat{s}_t$ and s_t.

Since $\hat{\mathcal{K}}'(s)$ converges locally uniformly to $\mathcal{K}'(s)$ a.s. and $\hat{r} \to s_t$ a.s., for large enough n, $|\hat{\mathcal{K}}'_n(\hat{r})| \leq |\mathcal{K}'(s_t)| + 1$, which implies $\hat{\mathcal{K}}(\hat{s}_t) - \mathcal{K}(s_t) = O_p(n^{-1/2})$. A similar argument can be used to show $\hat{\mathcal{K}}''(\hat{s}_t) - \mathcal{K}''(s_t) = O_p(n^{-1/2})$, which again implies $\log\hat{\mathcal{K}}''(\hat{s}_t) - \log\mathcal{K}''(s_t) = O_p(n^{-1/2})$. Thus, we obtain the desired result.

For $0 \leq \delta < \min\{1/2, 1 - c/b\}$, the similar argument to the above can be applied with Proposition 35 to obtain $o_p(n^{-\delta})$ part.

2. Again, we show $O_p(n^{-1/2})$ part first. Let $t \neq EW$. Thus, $s_t \neq 0$ and r_t, $u_t \neq 0$. In the above proof, we showed $\{s_t t - \mathcal{K}(s_t)\} - \{\hat{s}_t w - \hat{\mathcal{K}}_n(\hat{s}_t)\} = O_p(n^{-1/2})$. Since $s_t \neq 0$ and $\hat{s}_t \xrightarrow{\text{a.s.}} s_t$, for large enough n, $\operatorname{sgn}\hat{s}_t = \operatorname{sgn} s_t$ a.s. Since the function $\sqrt{2x}$ is a C^1 function on $(0,\infty)$, we conclude $\hat{r}_t - r_t = O_p(n^{-1/2})$.

   To show $\hat{u}_t - u_t = O_p(n^{-1/2})$, we consider

$$\log\hat{u}_t - \log u_t = \log\hat{s}_t - \log s_t + \frac{1}{2}\{\log\hat{\mathcal{K}}''(\hat{s}_t) - \log\mathcal{K}''(s_t)\}.$$

   As shown in part (1), $\log\hat{\mathcal{K}}''(\hat{s}_t) - \log K''(s_t) = O_p(n^{-1/2})$. Thus, we conclude $\hat{u}_t - u_t = O_p(n^{-1/2})$.

   Finally, observing $\hat{F}_0(w) = h(\hat{r}_t, \hat{u}_t)$, where $h(x,y) = \Phi(x) + \phi(x)(1/x - 1/y)$ and h is C^1 function away from 0, Proposition 36 give us $\hat{F}_0(t) - F_0(t) = O_p\left(n^{-1/2}\right)$.

   For $t = EW$, as in Lemma 23, for $j = 1, 2, 3$, $\mathcal{K}^{(j)}(0)$ can be written as $h_j(\lambda, \mu'_1, \ldots, \mu'_j)$, where h_j is a continuous function and μ'_i is ith (non-central) moment of the service time distribution.

   By CLT, $\mu'_i - \hat{\mathcal{G}}^{(i)}(0) = O_p(n^{-1/2})$ for each i. Now, observe

$$\hat{F}_0\{\hat{\mathcal{K}}'(0)\} - F_0(EW) = \frac{\hat{\mathcal{K}}'''(0)}{6\sqrt{2\pi}\hat{\mathcal{K}}''(0)^{3/2}} - \frac{\mathcal{K}'''(0)}{6\sqrt{2\pi}\left(\mathcal{K}''(0)\right)^{3/2}}$$

and

$$\log \frac{\hat{\mathcal{K}}'''(0)}{6\sqrt{2\pi}\hat{\mathcal{K}}''(0)^{3/2}} - \log \frac{\mathcal{K}'''(0)}{6\sqrt{2\pi}\left(\mathcal{K}''(0)\right)^{3/2}} = \log \hat{\mathcal{K}}'''(0) - \log \mathcal{K}'''(0) - \frac{3}{2}\{\log \hat{\mathcal{K}}''(0) - \log \mathcal{K}''(0)\}$$

and applying Proposition 36 repeatedly, we obtain the desired result.

For $o(n^{1/r-1})$ part where $0 \leq \delta < \min\{1/2, 1-c/b\}$, a similar argument to the above can be applied with Proposition 35.

□

Remark 21. If $\mathcal{W}(s)$ is not steep, the limit of $\hat{F}_0(t)$, $F_0(t)$ may not exist. As we see in Example 6, if we set $G\sim$Pareto(.8,5), then $F_0(t)$ exists only for $t \in (.8, 1]$. Thus, Corollary 18 and Theorem 20 are meaningful only when $\mathcal{W}(s)$ is steep.

1.2.4 Miscellaneous results

Lemma 22. *If we use the empirical MGF $n^{-1}\sum_{i=1}^{n} e^{sx_i}$ in the saddlepoint approximation for the distribution of X, when x_i's are random sample of X, the range of x of the saddle point equation (1.1.5) is $(\min_i\{x_i\}, \max_i\{x_i\})$.*

Proof. Since $\hat{\mathcal{K}}_X(s) = \log n^{-1}\sum_{i=1}^{n} e^{sx_i}$, we have

$$\begin{aligned}\lim_{s\to\infty} \hat{\mathcal{K}}'_X(s) &= \lim_{s\to\infty} \frac{\sum_{i=1}^{n} x_i e^{sx_i}}{\sum_{i=1}^{n} e^{sx_i}} = \lim_{s\to\infty} \frac{\sum_{i=1}^{n} x_i e^{s(x_i - \max\{x_i\})}}{\sum_{i=1}^{n} e^{s(x_i - \max\{x_i\})}} \\ &= \lim_{s\to\infty} \frac{\max\{x_i\} e^{s}}{e^{s}} = \max_i\{x_i\}\end{aligned}$$

and the limit for the case of $s \to -\infty$ can be handled similarly. □

The following is needed to calculate $\hat{F}_0\{\hat{\mathcal{K}}'(0)\}$, $F_0(EW)$, the second order Taylor polynomial approximations of $\mathcal{K}'$ and $\mathcal{K}''$ around 0, and the first order Taylor polynomial approximation of $1/r_t - 1/u_t$ around 0 for a numerical implementation.

Lemma 23. *Let $\mathcal{K}(s) = \log \mathcal{W}(s)$, where $\mathcal{W}(s)$ is defined as in (1.1.1) and define $\mu_k' = \mathcal{G}^{(k)}(0)$ (i.e., $\mu_1 = 1/\mu$ and $\lambda\mu_1 = \rho$. Note that μ_k' is the kth (non-central) moment of the service time distribution). Then,*

$$\begin{aligned}
\lim_{s\nearrow 0} \mathcal{K}'(s) =& \mu_1' + \frac{\lambda\mu_2'}{2(1-\rho)},\\
\lim_{s\nearrow 0} \mathcal{K}''(s) =& \frac{\lambda^2 {\mu_2'}^2}{4(1-\rho)^2} + \frac{\lambda\mu_3'}{3(1-\rho)} + \left(\mu_2' - {\mu_1'}^2\right),\\
\lim_{s\nearrow 0} \mathcal{K}'''(s) =& 2{\mu_1'}^3 - 3\mu_1'\mu_2' + \mu_3' - \frac{{\mu_1'}^2\mu_4' - 2\mu_1'\mu_2'\mu_3' + {\mu_2'}^3}{4{\mu_1'}^3}\\
&+ \frac{3{\mu_2'}^3 - 4\mu_1'\mu_2'\mu_3' + {\mu_1'}^2\mu_4'}{4{\mu_1'}^3(1-\rho)} + \frac{2\mu_1'\mu_2'\mu_3' - 3{\mu_2'}^3}{4{\mu_1'}^3(1-\rho)^2} + \frac{{\mu_2'}^3}{4{\mu_1'}^3(1-\rho)^3}\\
\lim_{s\nearrow 0} \mathcal{K}^{(4)}(s) =& \left(-6{\mu_1'}^4 + 12{\mu_1'}^2\mu_2' - 4\mu_1'\mu_3' + \mu_4'\right) + \frac{3{\mu_2'}^4\lambda^4}{8(1-\rho)^4}\\
&+ {\mu_2'}^2\left\{\frac{\mu_3'\lambda^3}{(1-\rho)^3} - 3\right\} + \frac{\mu_5'\lambda}{5(1-\rho)} + \frac{\lambda^2\left(2{\mu_3'}^2 + 3\mu_2'\mu_4'\right)}{6(1-\rho)^2}.
\end{aligned}$$

Proof. As $s \nearrow 0$, $\mathcal{K}'(s)$ becomes a $0/0$ form. Applying L'Hospital's rule for the left-hand limit (Theorem 30.2 of [57]) twice and using the fact, $\mathcal{G}(0) = 1$, we obtain $\mathcal{K}'(0)$. For $\mathcal{K}''(0)$, $\mathcal{K}'''(0)$, and $\mathcal{K}^{(4)}(0)$, we need to apply L'Hospital's rule 4 times, 6 times, and 8 times, respectively. □

By the method of moments, we obtain the estimators of $\mathcal{K}^{(j)}(0)$ for $j = 1, 2, 3, 4$. For example

$$\hat{\mathcal{K}}'(0) = \hat{\mathcal{G}}'(0) + \frac{\hat{\lambda}\hat{\mathcal{G}}'(0)}{2(1-\hat{\rho})} \xrightarrow{\text{a.s.}} \mu_1' + \frac{\lambda\mu_2'}{2(1-\rho)} = \mathcal{K}'(0),$$

where the a.s. convergence comes from the SLLN and the continuous mapping theorem for a.s. convergence. The a.s. convergence also holds for $\hat{\mathcal{K}}^{(j)}$, $j = 2, 3, 4$.

Theorem 24. *$\sqrt{n}\{\hat{\mathcal{W}}(s) - \mathcal{W}(s)\}$ converges weakly to a Gaussian process with zero mean in the metric space $C_{[s_1,s_2]}$ for any $[s_1, s_2] \subset (-\infty, \min\{c, b/2\})$.*

Proof. We fix $[s_1, s_2] \subset (-\infty, \min\{c, b/2\})$. Referring to Theorem 7.1 of [11] and previous Lemma 14, we only need to show tightness, for which it is enough to show (see Theorem 7.5

of [11]) that for any $\varepsilon > 0$,

$$\lim_{|s-r|\to 0} \limsup_n P\left(\left|\sqrt{n}\left(\hat{\mathcal{W}}(s) - \mathcal{W}(s)\right) - \sqrt{n}\left(\hat{\mathcal{W}}(r) - \mathcal{W}(r)\right)\right| \geq \varepsilon\right) = 0. \tag{1.2.6}$$

To this end, we split the term inside of P as following: Define function $h(s,x,y,z)$ by

$$h(s,x,y,z) = \frac{\left(1 - yx^{-1}\right) sz}{s + x^{-1} - x^{-1}z}$$

and observe that $\hat{\mathcal{W}}(s) = h\{s, \overline{I}, \overline{G}, \hat{\mathcal{G}}(s)\}$. We set

$$q_n(s) = h\{s, \lambda^{-1}, \mu^{-1}, \hat{\mathcal{G}}(s)\} \quad \text{and} \quad q_n(r) = h\{r, \lambda^{-1}, \mu^{-1}\hat{\mathcal{G}}(r)\}$$

which give us

$$\begin{aligned} &\left|\sqrt{n}\{\hat{\mathcal{W}}(s) - \mathcal{W}(s)\} - \sqrt{n}\{\hat{\mathcal{W}}(r) - \mathcal{W}(r)\}\right| \\ &\qquad \leq \left|\sqrt{n}\{\hat{\mathcal{W}}(s) - q_n(s)\} - \sqrt{n}\{\hat{\mathcal{W}}(r) - q_n(r)\}\right| \\ &\qquad\qquad + \left|\sqrt{n}\{q_n(s) - \mathcal{W}(s)\} - \sqrt{n}\{q_n(r) - \mathcal{W}(r)\}\right|. \end{aligned}$$

Referring to above inequality, we have

$$\begin{aligned} P(\text{LHS} \geq \varepsilon) \leq & P(\text{RHS} \geq \varepsilon) \\ \leq & P\left(\left|\sqrt{n}\{\hat{\mathcal{W}}_n(s) - q_n(s)\} - \sqrt{n}\{\hat{\mathcal{W}}_n(r) - q_n(r)\}\right| \geq \varepsilon/2\right) \\ & + P\left(\left|\sqrt{n}\{q_n(s) - \mathcal{W}(s)\} - \sqrt{n}\{q_n(r) - \mathcal{W}(r)\}\right| \geq \varepsilon/2\right) \end{aligned}$$

and (1.2.6) can be obtained by showing the the last two probabilities of the above inequalities satisfies (1.2.6) by themselves.

As we mentioned before, it is proved that $\sqrt{n}\{\hat{\mathcal{G}}(s) - \mathcal{G}(s)\}$ converges to a Gaussian process on $C_{[s_1,s_2]}$ in [23] and [31]. We first show that $\sqrt{n}\{q_n(s) - \mathcal{W}(s)\}$ converges weakly to a Gaussian process on $C_{[s_1,s_2]}$.

We use the functional delta method (Theorem 20.8 of [69]). Define $\phi : C_{[s_1,s_2]} \to C_{[s_1,s_2]}$ by

$$\phi\left(\theta\right)\left(s\right) = h\{s, \lambda^{-1}, \mu^{-1}, \theta\left(s\right)\} = \frac{\left(1-\rho\right)s\theta\left(s\right)}{s+\lambda-\lambda\theta\left(s\right)}.$$

First, we have

$$\frac{h\left\{s, \lambda^{-1}, \mu^{-1}, \mathcal{G}\left(s\right)+t\xi_t\left(s\right)\right\} - h\left\{s, \lambda^{-1}, \mu^{-1}, \mathcal{G}\left(s\right)\right\}}{t}, \\ = \frac{\xi_t\left(s\right)s\left(1-\rho\right)\left(s+\lambda\right)}{\left\{s+\lambda-\lambda\mathcal{G}\left(s\right)\right\}\left[s+\lambda-\lambda\left\{\mathcal{G}\left(s\right)+t\xi_t\left(s\right)\right\}\right]}$$

which give us that the Hadamard derivative of ϕ at $\mathcal{G}$ would be

$$\phi'_{\mathcal{G}}\left(\xi\right)\left(s\right) = \frac{\xi\left(s\right)\left(1-\rho\right)s\left(s+\lambda\right)}{\left\{s+\lambda-\lambda\mathcal{G}\left(s\right)\right\}^2}.$$

It is clear that $\phi'_{\mathcal{G}}$ is a continuous linear map. We have

$$\frac{\phi\left(\mathcal{G}+t\xi_t\right)\left(s\right)-\phi\left(\mathcal{G}\right)\left(s\right)}{t} - \phi'_{\mathcal{G}}\left(\xi\right)\left(s\right) \\ = \frac{s\left(s+\lambda\right)\left(1-\rho\right)\left[\left\{s+\lambda-\lambda\mathcal{G}\left(s\right)+t\lambda\xi\left(s\right)\right\}\xi_t\left(s\right)-\left\{s+\lambda-\lambda\mathcal{G}\left(s\right)\right\}\xi\left(s\right)\right]}{\left\{s+\lambda-\lambda\mathcal{G}\left(s\right)\right\}^2\left[s+\lambda-\lambda\left\{\mathcal{G}\left(s\right)+t\xi_t\left(s\right)\right\}\right]},$$

which implies

$$\left\|\frac{\phi\left(\theta+t\xi_t\right)\left(s\right)-\phi\left(\theta\right)\left(s\right)}{t} - \phi'_{\theta}\left(\xi\right)\left(s\right)\right\| \to 0$$

as $t \to 0$ for every $\xi_t \to \xi$, where $\|\cdot\|$ is the supremum on $C_{[s_1,s_2]}$. Thus ϕ is Hadamard differentiable and

$$\sqrt{n}\left\{q_n\left(s\right)-\mathcal{W}\left(s\right)\right\} = \sqrt{n}\left\{\phi\left(\hat{\mathcal{G}}\right)\left(s\right)-\phi\left(\mathcal{G}\right)\left(s\right)\right\} \Rightarrow \frac{Y_s\left(1-\rho\right)s\left(s+\lambda\right)}{\left\{s+\lambda-\lambda\mathcal{G}\left(s\right)\right\}^2},$$

where Y_s is the limit Gaussian process of $\sqrt{n}\{\hat{\mathcal{G}}\left(s\right)-\mathcal{G}\left(s\right)\}$. Because $\sqrt{n}\{q_n\left(s\right)-\mathcal{W}\left(s\right)\}$ has the weak convergence limit, they are relatively compact. Note that $C_{[s_1,s_2]}$ is separable and complete (Example 1.3 of [11]) on which the relative compactness implies tightness (Theorem 5.2 of [11]). Therefore, $\sqrt{n}\{q_n\left(s\right)-\mathcal{W}\left(s\right)\}$ is tight and satisfies (1.2.6).

Now, using the multivariate mean value theorem again, we have

$$\begin{aligned}
&\sqrt{n}\left\{\hat{\mathcal{W}}(s) - q_n(s)\right\} - \sqrt{n}\left\{\hat{\mathcal{W}}(r) - q_n(r)\right\} \\
&= \frac{\partial h}{\partial x}(\mathbf{b}_n)\sqrt{n}\left(\overline{I} - \lambda^{-1}\right) + \frac{\partial h}{\partial y}(\mathbf{b}_n)\sqrt{n}\left(\overline{G} - \lambda^{-1}\right) \\
&\qquad\qquad - \left\{\frac{\partial h}{\partial x}(\mathbf{b}'_n)\sqrt{n}\left(\overline{I} - \lambda^{-1}\right) + \frac{\partial h}{\partial y}(\mathbf{b}'_n)\sqrt{n}\left(\overline{G} - \mu^{-1}\right)\right\} \\
&= \left\{\frac{\partial h}{\partial x}(\mathbf{b}_n) - \frac{\partial h}{\partial x}(\mathbf{b}'_n)\right\}\sqrt{n}\left(\overline{I}_n - \lambda^{-1}\right) + \left\{\frac{\partial h}{\partial y}(\mathbf{b}_n) - \frac{\partial h}{\partial y}(\mathbf{b}'_n)\right\}\sqrt{n}\left(\overline{G} - \mu^{-1}\right) \\
&\qquad \Rightarrow \left\{h_x\left(s, \mu^{-1}, \lambda^{-1}, \mathcal{G}(r)\right) - h_x\left(t, \mu^{-1}, \lambda^{-1}, \mathcal{G}(s)\right)\right\} N\left(0, \sigma_G^2\right) \\
&\qquad\qquad + \left\{h_y\left(s, \mu^{-1}, \lambda^{-1}, \mathcal{G}(r)\right) - h_y\left(t, \mu^{-1}, \lambda^{-1}, \mathcal{G}(s)\right)\right\} N\left(0, \sigma_I^2\right),
\end{aligned}$$

where $\mathbf{b}_n$ is a vector in between $\{s, \overline{I}, \overline{G}, \hat{\mathcal{G}}(s)\}$ and $\{s, \mu^{-1}, \lambda^{-1}, \hat{\mathcal{G}}(s)\}$, and $\mathbf{b}'_n$ is a vector in between $\{r, \overline{I}, \overline{G}, \hat{\mathcal{G}}(r)\}$ and $\{r, \mu^{-1}, \lambda^{-1}, \hat{\mathcal{G}}(r)\}$ and $N(0, \sigma_G^2) \perp\!\!\!\perp N(0, \sigma_I^2)$.

We note that as a function of s, $h_x\{s, \mu^{-1}, \lambda^{-1}, \mathcal{G}(s)\}$ and $h_y\{s, \mu^{-1}, \lambda^{-1}, \mathcal{G}(s)\}$ satisfy Lipschitz condition on $[s_1, s_2]$ since h_x and h_y are differentiable with respect s on the compact domain $[s_1, s_2]$.

Therefore, we have

$$\lim_{|s-r|\to 0} \lim_{n\to\infty} P\left(\left|\sqrt{n}\{\hat{\mathcal{W}}(s) - q_n(s)\} - \sqrt{n}\{\hat{\mathcal{W}}(r) - q_n(r)\}\right| \geq \varepsilon/2\right) = 0$$

and conclude (1.2.6). □

We note that the ways to prove Lemma 14 and Theorem 24 can be used to show that $\sqrt{n}\{D^k\hat{\mathcal{W}}(s) - D^k\mathcal{W}(s)\}$ and $\sqrt{n}\{D^k\hat{\mathcal{K}}(s) - D^k\mathcal{K}(s)\}$, $k \in \mathbb{N}$ converge weakly to Gaussian processes on $C_{[s_1,s_2]}$ for any $[s_1, s_2] \subset (-\infty, \min\{c, b/2\})$.

1.3 Estimation of the CDF and the PDF of stationary waiting times

In this section, we show how the saddlepoint approximation can be used to estimate the PDF and CDF of the stationary waiting time distribution of M/G/1 queues. The assumptions remain the same as in the previous section. We observe the service times and inter-arrival times so that the random sample of these can be used to get information about the waiting time. To this end, we use the "empirical" moment generating function of (1.1.3).

We will show in §1.3.1 that we can obtain a simulated value whose MGF is $\hat{\mathcal{W}}(s)$ and we will denote this by $\hat{W}$, which also can be used for the estimation of CDF and PDF of W by the standard methods to estimate CDF and PDF, namely the empirical CDF and the kernel density estimation. However, obtaining the random sample of $\hat{W}$ is time consuming and we propose that the saddlepoint inversion of $\hat{\mathcal{W}}(s)$ should be used instead.

We fix notations first. The plug in estimator, $\hat{\mathcal{W}}(s)$ of $\mathcal{W}(s)$, the MGF of stationary waiting time distribution, is defined by (1.1.3). Let $\hat{F}$ be the corresponding (true) CDF for $\hat{\mathcal{W}}(s)$ such that $\hat{W} \sim \hat{F}$ and let $\hat{F}^{\dagger}(t)$ be the empirical CDF of $\hat{W}_1, \cdots, \hat{W}_m \overset{\text{iid}}{\sim} \hat{F}(t)$. Also, we recall that $\hat{F}_0(x)$ is the saddlepoint CDF approximation calculated from $\hat{\mathcal{W}}(s)$ and $F_0(t)$ is the saddlepoint CDF approximation calculated from $\mathcal{W}(s)$.

Let $A \rightsquigarrow B$ denote A approximates B. As we showed in the introduction section, the following diagram holds;

$$\begin{tikzcd}
 & & \left\{\hat{F}_0(t)\right. \arrow[r, "\text{a.s.}", "n\to\infty"'] \arrow[d, rightsquigarrow] & \left.F_0(t)\right\} \arrow[d, rightsquigarrow] \\
\left\{\hat{\mathcal{W}}(s)\right. \arrow[r, "\text{a.s.}", "n\to\infty"'] & \left.\mathcal{W}(s)\right\} \arrow[ur, Rightarrow] \arrow[r, Leftrightarrow] & \left\{\hat{F}(t)\right. \arrow[r, "\text{a.s.}", "n\to\infty"'] & \left.F(t)\right\} \\
 & & \hat{F}^{\dagger}(t) \arrow[u, "\text{a.s.}", "m\to\infty"'] & F^{\dagger}(t) \arrow[u, "\text{a.s.}", "m\to\infty"']
\end{tikzcd}$$

Let $\{G_j^*\}_{j=1}^n$ be a (non-parametric) bootstrap sample from $\{G_j\}_{j=1}^n$ (i.e., each G_i^* is independently sampled from $\{G_j\}_{j=1}^n$) and $\hat{\lambda}^* = n / \sum_{j=1}^n I_j^*$, where $\{I_j^*\}_{j=1}^n$ is a parametric bootstrap sample from $\text{Exp}(\hat{\lambda})$ (i.e., $I_j^* \overset{\text{iid}}{\sim} \text{Exp}(\hat{\lambda})$). Then, we can construct $\hat{\mathcal{W}}^*(s)$, a bootstrap replication of $\hat{\mathcal{W}}(s)$ as in

$$\hat{\mathcal{W}}^*(s) = \frac{(1-\hat{\rho}^*)\, s\hat{\mathcal{G}}^*(s)}{s + \hat{\lambda}^* - \hat{\lambda}^* \hat{\mathcal{G}}^*(s)},$$

where $\hat{\rho}^* = \sum_{j=1}^n G_j^* / \sum_{j=1}^n I_j^*$ and $\hat{\mathcal{G}}^*(s) = n^{-1} \sum_{j=1}^n \exp\{sG_j^*\}$. Let $\hat{F}^*(t)$ be the corresponding CDF of $\hat{\mathcal{W}}^*(s)$. Then, we expect that the bootstrap sampling of $\{\hat{F}^*(t)\}$ approximates the sampling distribution of $\{\hat{F}(t)\}$ and since $\hat{F}(t)$ converges to $F(t)$, we may use $\{\hat{F}^*(t)\}$ to construct bootstrap confidence intervals of $F(t)$.

Because $\hat{F}(t)$ and $\hat{F}^*(t)$ are generally not known, we need to use an estimator of them and using the empirical CDFs $\hat{F}^\dagger(t)$ and $\hat{F}^{*\dagger}(t)$ directly would be time consuming. We propose to use instead the saddlepoint CDF estimator of $\hat{F}(t)$ and $\hat{F}^*(t)$ denoted as $\hat{F}_0(t)$ and $\hat{F}_0^*(t)$ respectively. Rigorously, since $\hat{F}_0(t)$ converges to $F_0(t)$, a confidence interval based on the bootstrapped saddlepoint CDF approximations, $\{\hat{F}_n^*(t)\}$ is a CI of $F_0(t)$ not of $F(t)$.

For this, we check whether $\hat{F}_0(t)$ approximates $\hat{F}(t)$ well or not, which can be estimated by $\hat{F}^\dagger(t)$ with a sufficiently large sample size m of the simulated $\hat{W}^*$. If it is, then we may assume $\{\hat{F}_0^*(t)\}$ approximates $\{\hat{F}^*(t)\}$ well and the usage of $\{\hat{F}_0^*(t)\}$ to construct CI for $F(t)$ is legit.

Note that our saddlepoint estimator $\hat{F}_0(w)$ is in fact an estimator of $\hat{F}(t)$, which we believe is the best estimator of $F(t)$[4], or

$$\hat{F}_0(t) \rightsquigarrow \hat{F}(t) \rightsquigarrow F(t).$$

This section is organized as follows. In Subsection 1, we show how we can obtain the random sample of $\hat{W}$. In Subsection 2, we check the convergence of $\hat{\mathcal{W}}(x)$ and $\hat{c}$ for a M/M/1 queue case. In Subsection 3, we check how $F_0(t)$ approximates $F(t)$ well for service time

[4] Currently, this is just a claim by the author who does not know how to support this claim yet.

distributions of Exp(2) and Gamma$(3,3)$, for which the stationary waiting time distributions can be obtained analytically. In Subsection 4, we show how $\hat{F}_0(t)$ approximates $\hat{F}(t)$ by comparing to $\hat{F}^{\dagger}(t)$. In Subsection 5, we show how the bootstrap CI's based on $\{\hat{F}_0^*(t)\}$ of $F(t)$ work by simulation studies. In Subsection 6, we show a case of why we need to check whether $\hat{F}(t) \rightsquigarrow F(t)$ or not and why the blind usage of the saddlepoint CDF approximation to $F(t)$ may be dangerous.

Our choices of service time distributions in this section are in Table 1.3.1(see Table 4.1.2 for the PDF and CDF) but for Exp(1) service time, it is known that $W \sim \text{Exp}(.5)$, which has a relatively large range (99.5th percentile is about 10.597) comparing to other cases and we will use Exp(2) (with $\lambda = 1$) as the service time distribution for the M/M/1 queue case.

For the convenience of the reader, the following table summarize the symbols used in this section.

$$\begin{tikzcd}
 & & & F_0(t) & \text{saddlepoint CDF} \\
\text{simulated value} & W & \mathcal{W}(s) \arrow[l] \arrow[r] \arrow[ur] \arrow[dr] & F(t) & \text{true CDF} \\
 & & & F^{\dagger}(t) & \text{empirical CDF}
\end{tikzcd}$$

$$\begin{tikzcd}
 & & & \hat{F}_0(t) & \text{saddlepoint CDF} \\
\text{simulated value} & \hat{W} & \hat{\mathcal{W}}(s) \arrow[l] \arrow[r] \arrow[ur] \arrow[dr] & \hat{F}(t) & \text{true CDF} \\
 & & & \hat{F}^{\dagger}(t) & \text{empirical CDF}
\end{tikzcd}$$

$$\begin{tikzcd}
 & \hat{F}_0^*(t) & \text{saddlepoint CDF} \\
\hat{\mathcal{W}}^*(s) \arrow[r] \arrow[ur] & \hat{F}^*(t) & \text{true CDF}
\end{tikzcd}$$

After the majority of our work had been done, we later found that similar studies of nonparametric estimations using empirical service time to obtain the CI of the CDF of W_q, the stationary waiting time in queue, were done in [33], [32] and [51].

In [33], bounds of the CDF of W_q were obtained from the asymptotic normality (Theorem 9) of $\hat{c}$, which is also an empirical estimation of the adjustment coefficient or the Lundberg exponent in ruin probability. See §2.1.3 for the related matter.

In [32], the Cramér-Lundberg approximation with the empirical MGF of the service time is used with the jackknife method to obtain the CI of CDF of W_q. We added a comparison of their estimators in chapter 2.

In [51], the properties of $\hat{F}_q(t)$, as a nonparametric estimate of the CDF, of the stationary waiting time in queue for a GI/G/1 queue were studied based on the empirical process theory. The author showed that $\hat{F}_q(t)$ converges to $F_q(t)$ in the supremum norm and $\sqrt{n}\{\hat{F}_q(t) - F_q(t)\}$ converges weakly to a Gaussian process in the supremum norm. A numerical example of the bootstrap confidence band of $\hat{F}_q(t)$ in the GI/M/1 queue with a Uniform(0,2) arrival time distribution was also given, where $\hat{F}_q(t)$ was approximated by a numerical inversion based on the fast Fourier transform.

1.3.1 A solution by simulation

When we have a random sample of service times and an estimate of λ, we can, in principle, obtain a random sample of the stationary waiting time distribution approximation by running a queueing simulation with the empirical service time distribution instead of unknown true service time distribution. Of course, as in most Monte Carlo methods, this simulation needs burn-in to simulate stationary distributions and the outputs of this simulation are not iid, so this is not an efficient estimation and is rather time consuming. In [34], it is reported that even the estimated mean waiting time in queue by simulation method is quite inaccurate.

However, in our case of M/G/1, it is known that there is a better way to simulate. Write the Pollaczek-Khinchin formula (1.1.1) as

$$\frac{(1-\rho)\,s\mathcal{G}(s)}{s+\lambda-\lambda\mathcal{G}(s)}=\frac{(1-\rho)}{1-\rho\mu\left\{\frac{1-\mathcal{G}(s)}{-s}\right\}}\mathcal{G}(s)\,.$$

Then, using the fact that $\mu\left\{1-\mathcal{G}(s)\right\}/(-s)$ is the MGF of the equilibrium distribution of $\mathcal{G}(x)$ (i.e., the distribution whose density is $h(t)=\mu\left\{1-G(t)\right\}$) and $(1-\rho)/(1-\rho s)$ is the generating function of a geometric distribution with mass function $\rho^k(1-\rho)$, we recognize that

$$W\sim V_1+\cdots+V_N+G_1, \tag{1.3.1}$$

where $P(N=k)=\rho^k(1-\rho)$, $k=0,1,\cdots$, and $\{V_j\}$ are iid with the density $\mu\left\{1-G(t)\right\}$, and V, N, and G are independent.

Therefore, if we have random sample of $\{G_i\}_{i=1}^n$ at hand and plug-in its empirical CDF, $\hat{G}(t)$, then the density of V is estimated by $\hat{\mu}\left\{1-\hat{G}(t)\right\}$ and the estimated CDF of V is

$$\hat{H}(v)=\hat{\mu}\int_0^v\left(1-\hat{G}(t)\right)dt=\begin{cases}\hat{\mu}v & \text{if } 0\le v<G_{(1)}\\ \hat{\mu}\left\{\sum_{i=1}^k G_{(i)}+(n-k)\,v\right\}/n & \text{if } G_{(k)}\le v<G_{(k+1)}\\ 1 & \text{if } v\ge G_{(n)}\end{cases}, \tag{1.3.2}$$

where $G_{(k)}$ means kth order statistic. A random sample $\hat{V}_1,\cdots,\hat{V}_{\hat{N}}$ of size $\hat{N}$ is obtained from CDF $\hat{H}(v)$ where $\hat{N}$ has the probability mass function (PMF) $\hat{\rho}^k(1-\hat{\rho})$, where $\hat{\rho}$ given in (1.1.2). Each $\hat{V}_i$ is computed by taking the probability integral transform $\hat{H}^{-1}(\cdot)$ of a uniformly distributed variable. If $\hat{G}$ is randomly generated from $\{G_j\}_{j=1}^n$, we obtain a simulated value from $\hat{F}(t)$ as $\hat{W}=\hat{V}_1+\cdots+\hat{V}_{\hat{N}}+\hat{G}$. Clearly, if we know $G(t)$, we can obtain the random sample of W in similar way. See Table 1.3.1 and related remarks.

It can be shown that $\hat{W}\Rightarrow W$. We showed that $\hat{\mathcal{W}}(s)\to\mathcal{W}(s)$ almost surely (a.s.) in Lemma 12. Because $W\ge 0$, the Laplace transform of W, $\mathcal{W}(-s)$ exists for any $s\ge 0$. Thus, for any $s\ge 0$

$$E\left(e^{-s\hat{W}}|\mathcal{F}_n\right)=\hat{\mathcal{W}}(-s)\to\mathcal{W}(-s)\quad\text{a.s.},$$

where $\mathcal{F}_n = \sigma\left\{I_1, \cdots, I_n, G_1, \cdots, G_n\right\}$, the σ-field generated by $\{I_1, \cdots, I_n, G_1, \cdots, G_n\}$. Note that because $\hat{\mathcal{W}}(s)$ is strictly increasing, for $s \geq 0$ and for any $n \in \mathbb{N}$,

$$0 < \hat{\mathcal{W}}(-s) \leq \hat{\mathcal{W}}(0) = 1,$$

so that $\hat{\mathcal{W}}(-s)$ is uniformly bounded by the integrable function 1. Then, by the dominated convergence theorem, we have that for any $s \geq 0$,

$$\begin{aligned} \lim_{n\to\infty} E\left(e^{-s\hat{W}}\right) &= \lim_{n\to\infty} E\left\{E\left(e^{-s\hat{W}}|\mathcal{F}_n\right)\right\} = E\left\{\lim_{n\to\infty} E\left(e^{-s\hat{W}}|\mathcal{F}_n\right)\right\} \\ &= E\left\{\mathcal{W}(-s)\right\} = \mathcal{W}(-s), \end{aligned}$$

which implies $\hat{W} \Rightarrow W$ (i.e, $\hat{W}$ converges to W weakly) by the continuity theorem of the Laplace transform (Theorem 2 (p. 431) of [30] or Example 5.5 of [11]).

Estimators for the CDF or PDF of W can be computed from the suggested simulation, however not without a considerable amount of simulation. We show that alternative estimators that use saddlepoint approximations require far less computing time and are much simpler to compute without any loss in accuracy. Thus saddlepoint methods can replace the simulation.

Example 25. For Pareto(x_m,α), the mean is $\alpha x_m/(\alpha-1)$ and the PDF and the CDF are

$$g(x) = \frac{\alpha x_m^\alpha}{x^{\alpha+1}} \mathbf{1}_{(x_m,\infty)} \quad \text{and} \quad G(x) = \left\{1 - \left(\frac{x_m}{x}\right)^\alpha\right\} \mathbf{1}_{(x_m,\infty)},$$

so that

$$g_e(x) = \mu\left\{1 - G(x)\right\} = \frac{\alpha-1}{\alpha x_m}\left\{\mathbf{1}_{[0,x_m]} + \left(\frac{x_m}{x}\right)^\alpha \mathbf{1}_{(x_m,\infty)}\right\}$$

and

$$\begin{aligned} G_e(x) &= \frac{\alpha-1}{\alpha x_m}\left\{x\,\mathbf{1}_{[0,x_m]} + \left(x_m + \frac{x_m}{\alpha-1} - \frac{x_m^\alpha}{\alpha-1}\frac{1}{x^{\alpha-1}}\right)\mathbf{1}_{(x_m,\infty)}\right\} \\ &= \frac{\alpha-1}{\alpha x_m} x\,\mathbf{1}_{[0,x_m]} + \left(1 - \frac{x_m^{\alpha-1}}{\alpha}\frac{1}{x^{\alpha-1}}\right)\mathbf{1}_{(x_m,\infty)}. \end{aligned}$$

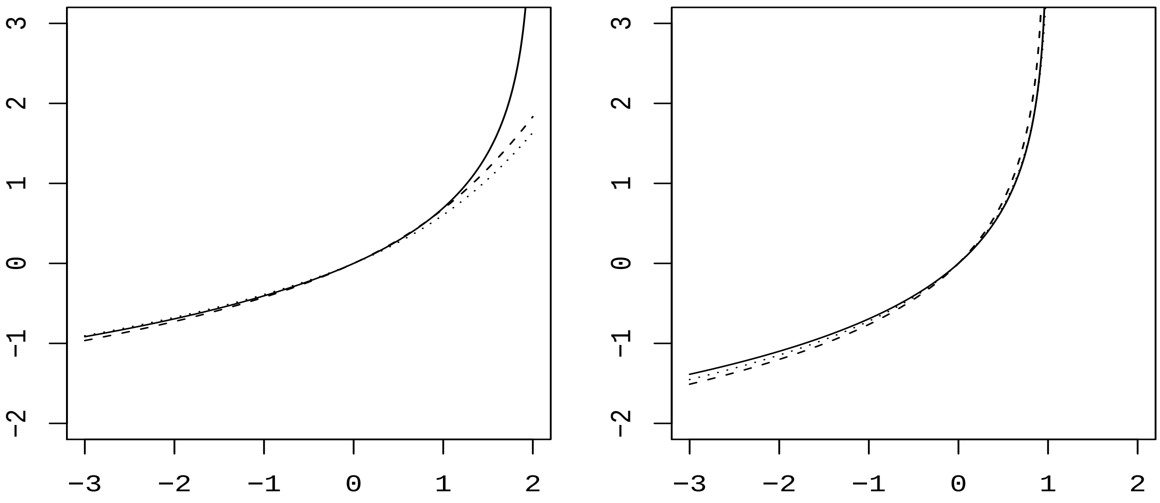

Figure 1.3.1: The true CGF and the estimated CGF of the the service time (left) and the stationary waiting time (right) of M/M/1 queue. Left: The solid curve is the true CGF of the service time, $\log(1-s/2)^{-1}$ and the dashed (dotted) curve is the estimated CGF of the random sample size 200 (50) . Right: The solid curve is the true CGF of the stationary waiting times, $\log(1-s)^{-1}$ and the dashed (dotted) curve is of its estimated CGF, $\hat{\mathcal{W}}_n(s)$ of a random sample size $n = 200$ (50) .

Thus, by solving $G_e(x) = t$ to obtain

$$G_e^{-1}(t) = \begin{cases} t/\mu & \text{if } t \leq 1-\alpha^{-1} \\ x_m/\left(\alpha(1-t)\right)^{1/(\alpha-1)} & \text{if } t > 1-\alpha^{-1} \end{cases}$$

For example, if $G \sim \text{Pareto}(\frac{2}{3}, 3)$, then $G_e^{-1}(t) = t\,\mathbf{1}_{[0,2/3]}(t) + 2/\left(3\sqrt{3(1-t)}\right)\mathbf{1}_{(2/3,\infty)}$.

1.3.2 Convergence of $\hat{\mathcal{W}}$ and $\hat{c}$

In the previous section, we study the convergence behavior of $\hat{\mathcal{K}}(s)$ and its isolated singularity (pole) $\hat{c}$. We first examine these feature with a simulation study of M/M/1 queues with the arrival rate, $\lambda = 1$ and the average service time, $\mu^{-1} = 1/2$ (so, $\mu = 2$ is the service rate of the queue), which will give insight about how the convergence of $\hat{\mathcal{K}}(s)$ works.

In this case, the stationary waiting time follows $\text{Exp}(\mu-\lambda) \sim \text{Exp}(1)$ distribution. In our simulation study, n, the sample size of a random sample are set to either 50 or 200.

The estimated $\log \hat{\mathcal{G}}(s)$ is drawn in Figure 1.3.1 with the true CGF, $\log(1 - s/2)^{-1}$. It is clear from the graph that the estimated CGF, $\log \hat{\mathcal{G}}(s)$, does not have the pole at 2, which can be explained with the fact that it is an analytic function by the definition in (1.1.4).

See the right graph of Figure 1.3.1 for the estimated CGF and the true CGF of the stationary waiting time. Recall that

$$\lim_{s \nearrow c} \mathcal{K}(s) = \infty.$$

Clearly, the quality of our estimator $\hat{\mathcal{K}}(s)$ is governed by how well the pole of $\hat{\mathcal{W}}(s)$, $\hat{c}$ estimates the pole of the true $\mathcal{W}(s)$, c, which Theorem 9 is about.

From Example 10, $c = \mu - \lambda$ and $\hat{c}$ is asymptotically normal if $\mu < 2\lambda$. To check the convergence of $\hat{c}$ to c, we obtain random samples of $\hat{c}_n$ of size 10^5 for $n = 50$ and $n = 200$, respectively but with setting $\lambda = 3/2$ and $\mu = 2$ to make sure $\mu < 2\lambda$ (or $c = .5 < \mu/2 = 1$). See the top 4 graphs of Figure 1.3.2 for the histograms of $\sqrt{n}(\hat{c} - .5)$ and the density curves of the limiting distribution $N(0, 27/4)$.

To see what happens when $b/2 < c < b$, we also obtain 10^5 $\hat{c}$ for $n = 50$ and $n = 200$, respectively with setting $\lambda = 1/2$ and $\mu = 2$, which results in $\mu > 2\lambda$ (or $1 = \mu/2 < c = 3/2 < 2 = \mu$. The graphs are the bottom 4 graphs of Figure 1.3.2, which show that though $\hat{c}$ are very close to following a normal distribution, the convergence of $\hat{c}$ to 3/2 is rather slower than $c = .5$ case (For $c = .5$ case, $\bar{\hat{c}} = .650$ (.516) for $n = 50$ (200) and for $c = 1.5$ case, $\bar{\hat{c}} = 1.913$ (1.701) for $n = 50$ (200)).

For case $n = 200$, one may inquire $\hat{c}$ converges to a normal distribution even if $c > b/2$. We simulate to obtain 10^5 $\hat{c}$ with $n = 10^3$, 10^4, and 10^5 to see if the convergence is true. Figure 1.3.3 shows the histograms of $\sqrt{n}(\hat{c} - 3/2)$, QQ-plots of $\hat{c}$ assuming the normal distribution, and boxplots of them. From the boxplots, it is clear that $\hat{c}$ converges to 1.5. However, the histograms show that the convergence is slower than the order of $O(\sqrt{n})$ and QQ-plots show that $\hat{c}$ does not converge to a normal distribution.

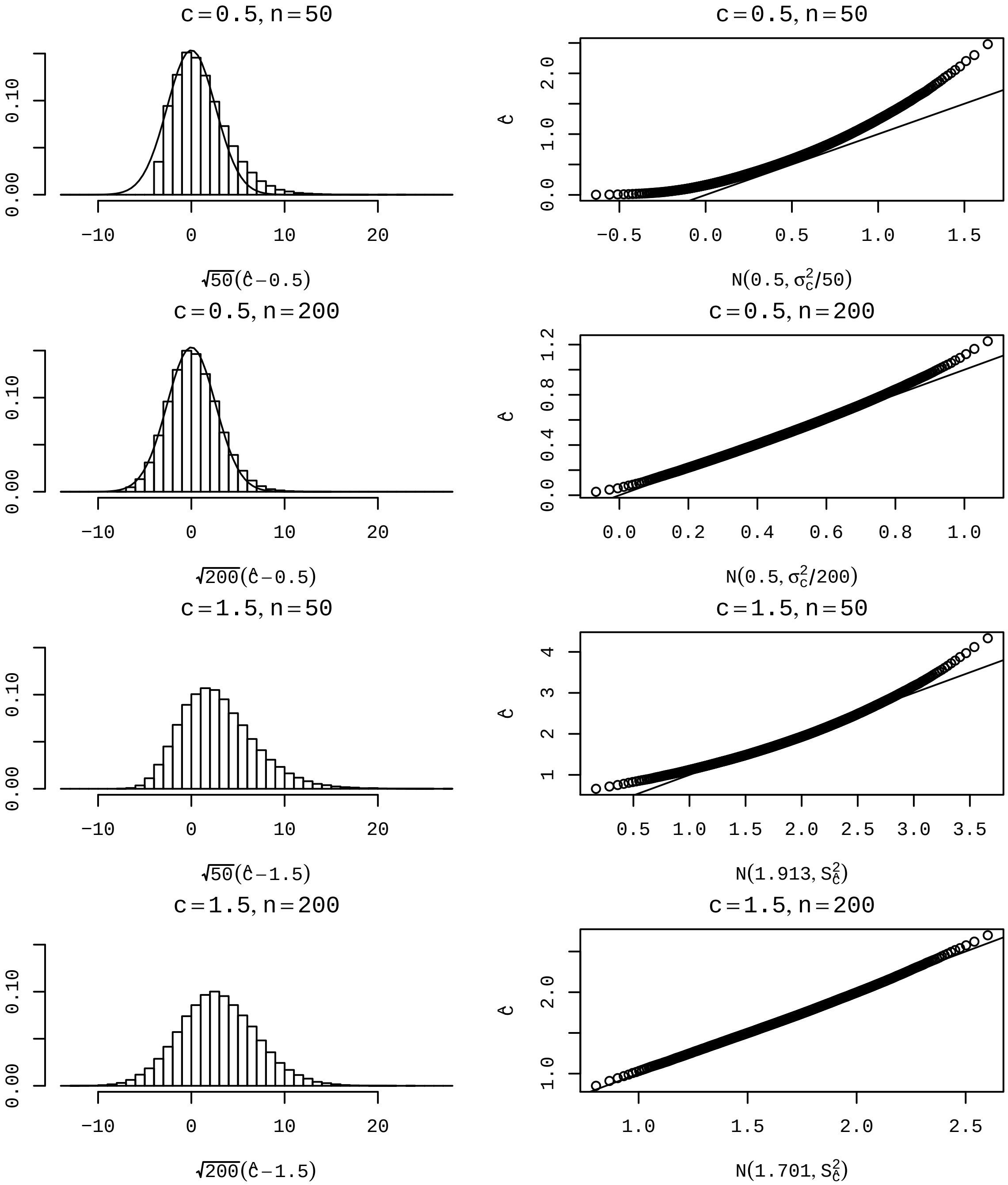

Figure 1.3.2: Histograms and QQ-plots of simulated 10^5 $\hat{c}$. The top 4 graphs: Histograms of 10^5 $\sqrt{n}\,(\hat{c}-c)$ and QQ-plots of $\hat{c}$ against $N\left(.5, \sigma_c^2/n\right)$, where $\hat{c}$ is the root of $\hat{\mathcal{D}}\,(s) = 0$, for M/M/1 queue model. Here, $\lambda = 3/2$, $\mu = 2$, $c = 1/2$ and the overlapped curve of the histogram is the density of the limiting distribution of $\sqrt{n}\,(\hat{c}-c)$, $N\,(0, 27/4)$. The bottom 4 graphs: Histograms of 10^5 $\sqrt{n}\,(\hat{c}-c)$, where$\lambda = 1/2$, $\mu = 2$, $c = 3/2$. For the QQ-plots of $\hat{c}$, $N\left(\bar{\hat{c}}, S_{\hat{c}}^2\right)$ ($\bar{\hat{c}} = 1.913$ (1.701) for $n = 50$ (200)) is used.

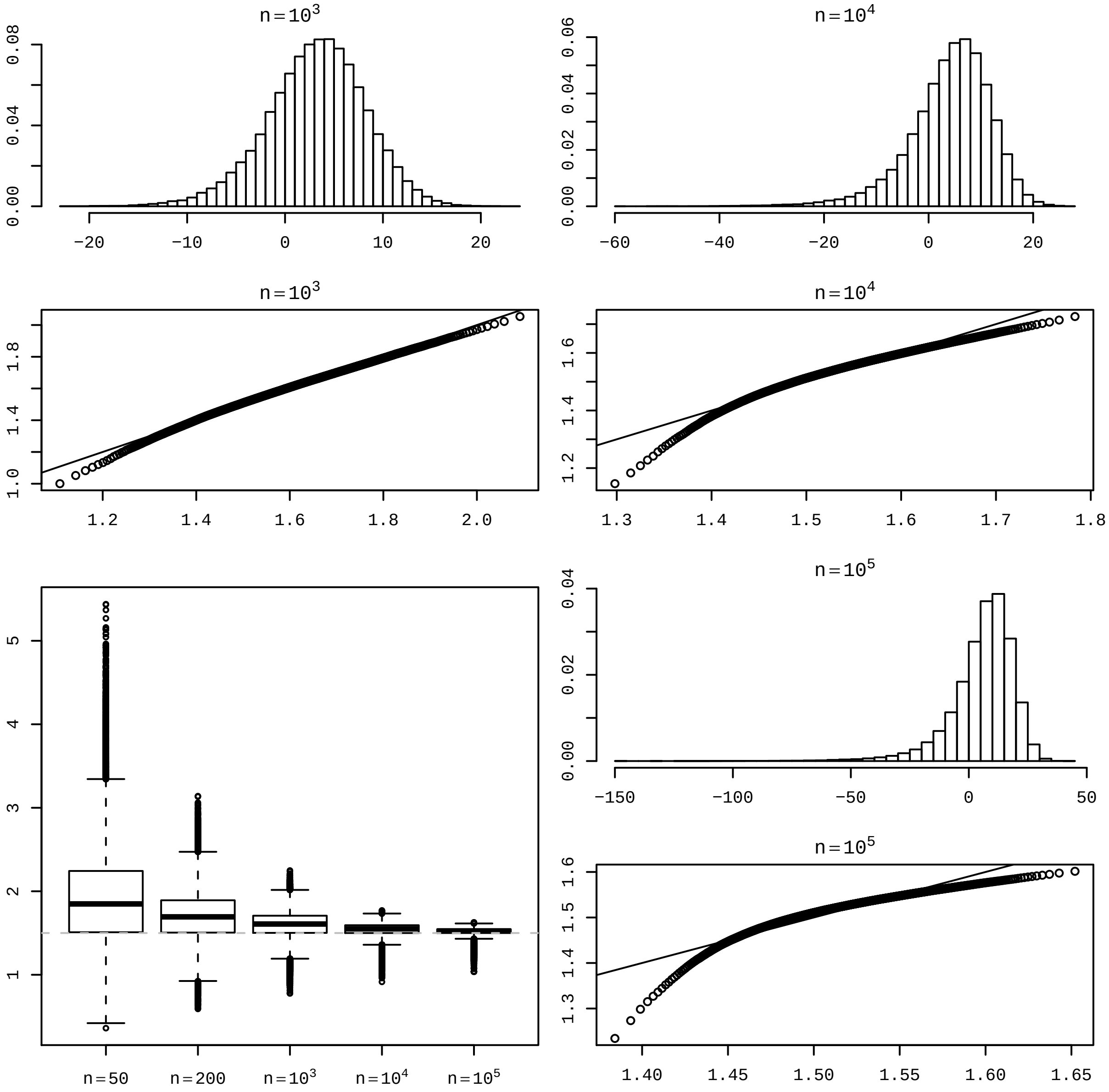

Figure 1.3.3: The histograms of $\sqrt{n}\,(\hat{c} - 3/2)$, QQ-plots of $\hat{c}$ assuming a normal distribution, and the boxplots of 10^5 $\hat{c}$ with $n = 10^3$, 10^4, and 10^5 where where$\lambda = 1/2$, $\mu = 2$, $c = 3/2$.

1.3.3 Saddlepoint approximations for W in M/E$_k$/1 queue

Exponential and gamma distributions have a closed form CGF, so that the corresponding $\mathcal{W}(s)$ also has a closed form and saddlepoint approximations can be calculated directly from $\mathcal{W}(s)$. See SFigure 1.3.4 for the saddlepoint PDF and CDF estimation of the stationary waiting times with their the % relative errors when the service time distributions are Exp(2) and Gamma$(3,3)$. The arrival rate λ is set to 1 and 2 respectively.

Note that for the % relative error for the PDF estimation with Exp(2) service time, the normalized saddlepoint PDF approximation can be used with

$$\begin{aligned} f_0(0) &= \lim_{s\to-\infty} \frac{1}{\sqrt{2\pi\mathcal{K}''(s)}} \exp\left(\mathcal{K}(s) - s\mathcal{K}'(s)\right) \\ &= \frac{1}{\sqrt{2\pi}} \lim_{s\to-\infty} \exp\left(\frac{-s}{1-s}\right) = \frac{\exp(1)}{\sqrt{2\pi}} \approx 1.084438. \end{aligned}$$

since $\lim_{s\to-\infty}\mathcal{K}'(s) = 0$. If $f_0(0)$ is not analytically tractable, it can be estimated through spline extrapolation estimation. All % relative errors are within first digit, which is a well known characteristic for the saddlepoint approximation.

Gamma(3,3) service distribution is a case that the Laplace transform can be inverted analytically. See Theorem 26. When the shape parameter α of Gamma(α, β) distribution is a natural number k, it is known as a Erlang distribution in queueing theory and usually denoted by E_k.

The true PDF $f(t)$ and the CDF $F(t)$ are compared with the saddlepoint approximations $f_0(t)$ and $F_0(t)$ in Figure 1.3.4, which shows that % relative errors are first digit in either of these cases too.

Note that for the % relative error of the CDF, the following formula is used;

$$\frac{100\left(F_0(t) - F(t)\right)}{\min\left\{F(t), 1-F(t)\right\}}.$$

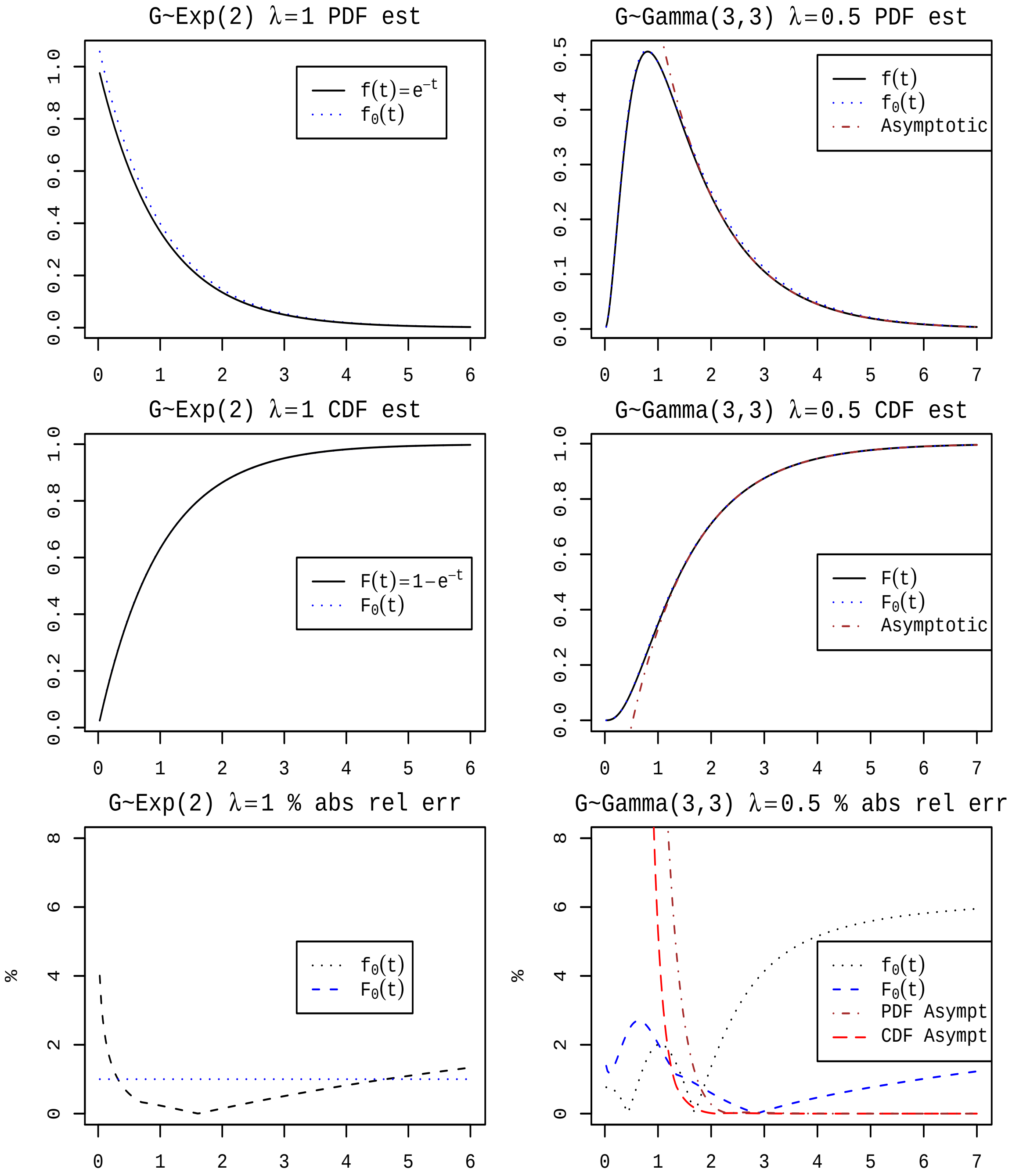

Figure 1.3.4: Top: Saddlepoint (unnormalized) PDF approximations, $f_0(t)$ of the waiting time distribution of M/M/1 queue (left), where $\mu = 2$ and $\lambda = 1$ and M/G/1 queue with Gamma$(3,3)$ service time distribution with $\lambda = .5$ (right). The solid curves are the true densities. Middle: Saddlepoint CDF approximations, $F_0(t)$. The solid curve is the true CDF and the dotted curve is the approximation from true CGF, $F_0(t)$ and the dot-dashed curve of Gamma(3,3) case is the asymptotic from Example 28. Bottom: The percentage absolute relative errors of normalized PDF (dotted) and CDF (dashed) estimations from the top and the middle graphs. The dot-dashed and long dash curves are of the asymptotic from Example 28.

1.3.4 Comparison to the simulation method of (1.3.1)

In this section, we check the performance of the saddlepoint approximation as estimators of the CDF $\hat{F}(t)$ and the PDF of $\hat{W}$.

Let m be the number of generated $\hat{W}$ to obtain $\hat{F}^{\dagger}(t)$, the empirical CDF of $\hat{W}$. For our simulation study, we generally set $m = 10^7$.

See Figure 1.3.5 and 1.3.6 for the comparison of the saddlepoint CDF and PDF approximation with the empirical CDF, $\hat{F}^{\dagger}(t)$, and the histogram of 10^7 $\hat{W}$ for the sample sizes $n = 50$ and $n = 200$, respectively. The graphs show the characteristic of saddlepoint approximations; Approximating closely the PDF and CDF as smooth functions.

Let

$$\hat{W}_q \sim \sum_{j=1}^{\hat{N}} \hat{V}_j, \tag{1.3.3}$$

where $\hat{V} \overset{\text{iid}}{\sim} \hat{H}(v)$ of (1.3.2) and $\hat{N}$ has the PMF of $\hat{\rho}^k(1-\hat{\rho})$. Since $\hat{V}$ has the PDF $\hat{\mu}\{1-\hat{G}(t)\}$, which is discontinuous only at finite points, $\hat{V}$ is a continuous r.v. and as a random sum of $\hat{V}$, we may see that the CDF of $\hat{W}_q$ is (absolutely) continuous on $(0,\infty)$ (See [63] and the reference therein for a formal proof).

Because $\hat{G}_j \sim \hat{G}(t)$, the empirical CDF of $\{G_j\}_{j=1}^n$, $\hat{G}_j$ acts as a discrete random variable. Note that the CDF of $\hat{W}_q$ can be decomposed as

$$(1-\hat{\rho})\,\mathbf{1}_{\{0\}}(t) + \hat{\rho}\hat{F}_q^{+}(t),$$

where $\hat{F}_q^{+}(t)$ is the CDF of $\sum_{j=1}^{\hat{N}^{+}} \hat{V}_j$ and $\hat{N}^{+}$ has the PMF of $P(\hat{N}^{+} = k)$=$\hat{\rho}^{k-1}(1-\hat{\rho})$ for $k = 1, 2, \cdots$. Then, $\hat{F}_q^{+}(t)$ is continuous but due to the point mass of $(1-\hat{\rho})$ at 0, the CDF of $\hat{W} = \hat{W}_q + \hat{G}$ still has the discrete part. In Figure 1.3.5 and 1.3.6, the effect of the discrete random variable part can be seen from the rougher lower quantile parts of the empirical CDF's of $\hat{W}$.

We check the performance of the saddlepoint approximation $\hat{F}_0(t)$ as an estimator of $\hat{F}(t)$ by checking the average of the relative errors against $\hat{F}^{\dagger}(t)$. Obtaining each $\hat{F}_k^{\dagger}(t)$

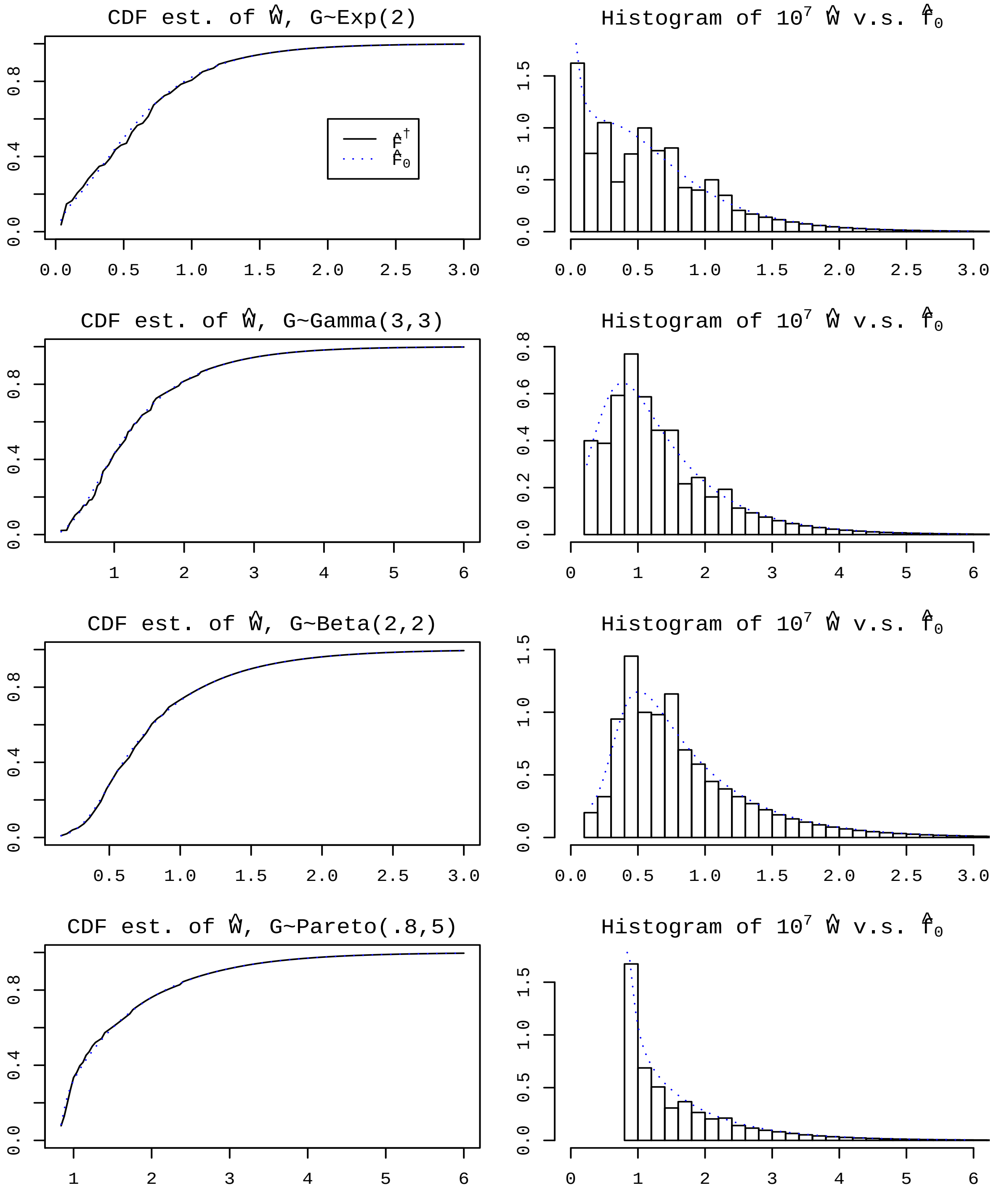

Figure 1.3.5: Saddlepoint approximation for $\hat{F}(t)$ and $\hat{f}(t)$. Left: Empirical CDF's, $\hat{F}^{\dagger}(t)$, of 10^7 $\hat{W}$ using (1.3.1) and (1.3.2) with the saddlepoint CDF approximations $\hat{F}(t)$ from $\hat{\mathcal{W}}(s)$. The x-axes are cut to include at least 99.5% of $\hat{W}$ of size $m = 10^7$. From the top, the $G_j \overset{\text{iid}}{\sim} \text{Exp}(2)$, Gamma$(3, 3)$, Beta$(2, 2)$, and Pareto$(4/5, 5)$. Right: Histogram of the 10^7 $\hat{W}$ used in the left graphs with the saddlepoint PDF approximation from $\hat{\mathcal{W}}(s)$.

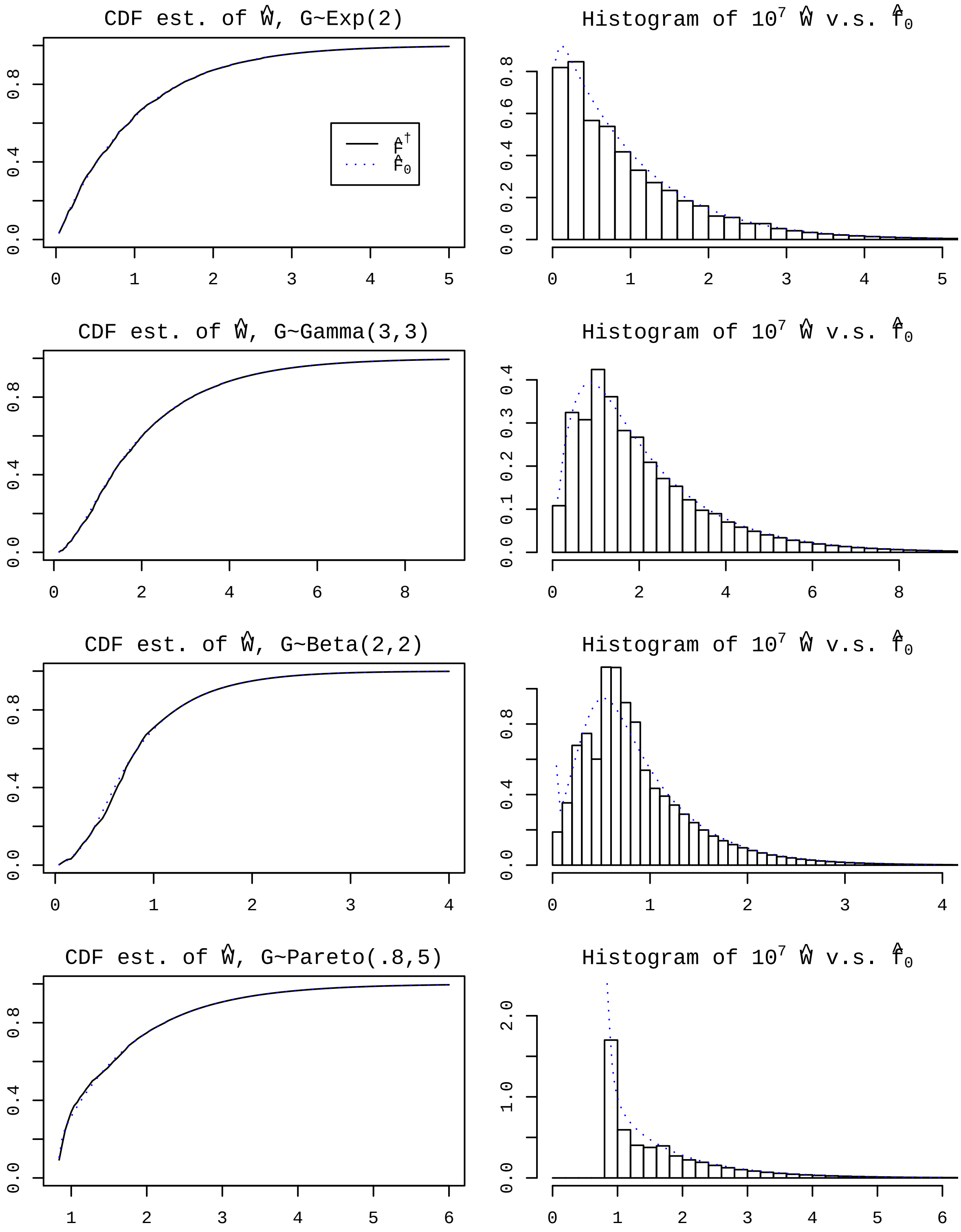

Figure 1.3.6: Same as Figure 1.3.5 but with the sample size $n = 200$.

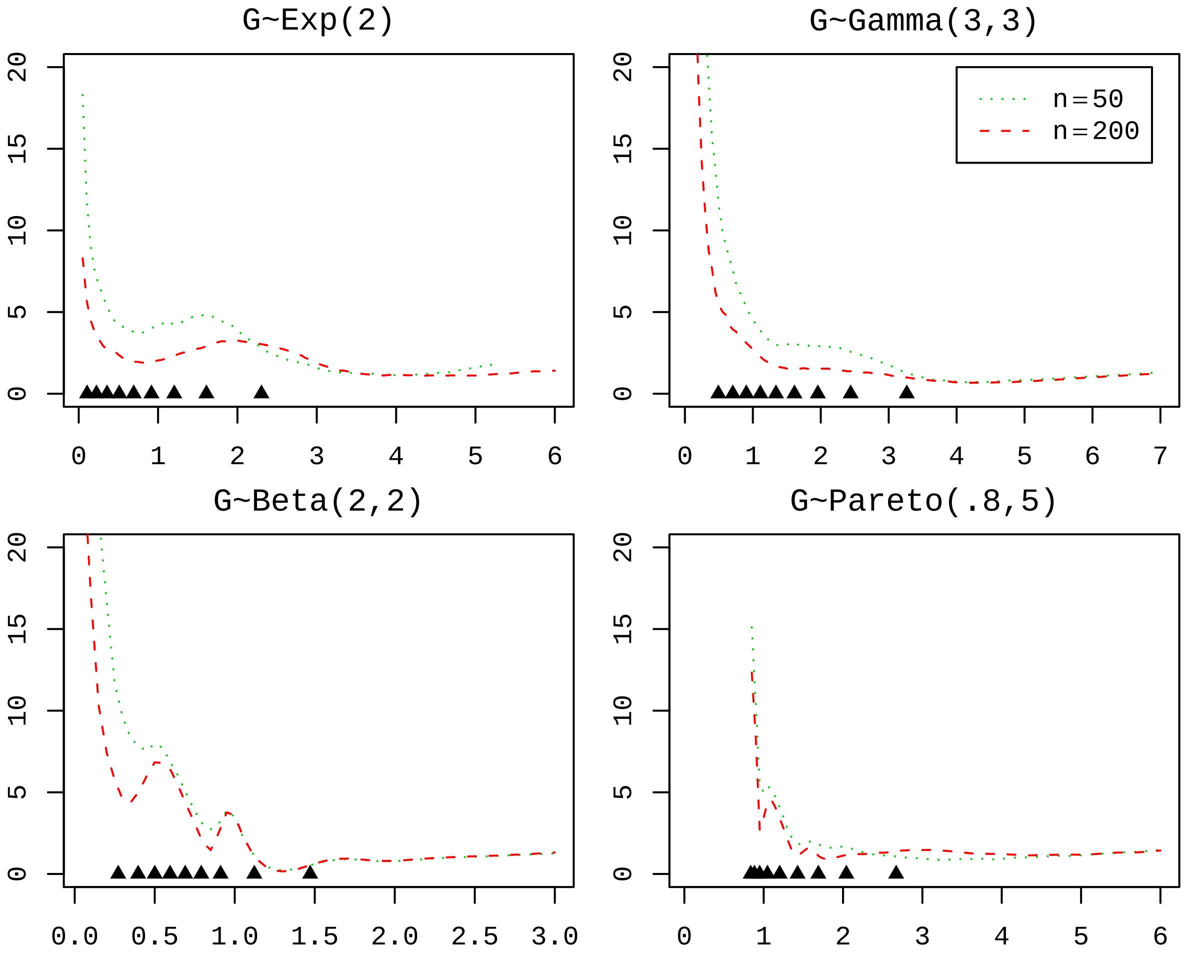

Figure 1.3.7: The average percentage relative absolute error of $\hat{F}_0(t)(n = 50$, dotted) and $\hat{F}_0(t)$ $(n = 200$, dashed) against $\hat{F}^{\dagger}(t)$ from $l = 10^3$ random samples of $\{\{G_j\}_{j=1}^n, \{I_j\}_{j=1}^n\}_{k=1}^l$ for different service times. In each graph, ▲ denotes a decile (from 10% to 90%) of the true distribution of W.

and $\hat{F}_{0k}(t)$ from $l = 10^3$ random samples of $\{\{G_j\}_{j=1}^n, \{I_j\}_{j=1}^n\}_{k=1}^l$ as we did in Figure 1.3.5 and 1.3.6, we calculate the average of $l = 10^3$ percentage relative absolute errors of $\hat{F}_{0k}(t)$'s, which is defined by

$$\frac{1}{l}\sum_{k=1}^{l}\frac{100\left|\hat{F}_{0k}(t)-\hat{F}_k^{\dagger}(t)\right|}{\min\left\{\hat{F}_k^{\dagger}(t),1-\hat{F}_k^{\dagger}(t)\right\}}$$

and estimates the mean percentage relative absolute error of $\hat{F}_0(t)$ as an estimator of $F_0(t)$. Figure 1.3.7 shows the result. All of them show a first digit percentage relative error for the tail area on average.

1.3.5 CI of CDF by the saddlepoint approximation with bootstrapping

In this subsection, we compare the performances of different bootstrap confidence intervals by simulation studies.

The number of samples for the simulation, l is set to 2500. Thus, with 90% confidence, the true coverage probability is within the observed coverage probability $\pm .01$ by either Wilson's confidence interval (CI) or Agresti-Coull's CI for the binomial proportion p.

Because of the restricted range of $0 \le \hat{F}_0^* (t) \le 1$, the bootstrap sample size $B = 10^4$ should be enough for estimating the percentile of $\{\hat{F}_0^* (t)\}$.

As we noted before, for Pareto service time distributions and beta service time distributions, analytically closed forms of $F(t)$ are not known (However, for a non-closed form of $W_q(t)$ with Pareto service time distribution, see [52]). Therefore, for the service time distributions of Beta$(2, 2)$ and Pareto$(.8, 5)$, the true CDF of W are estimated by simulating $m = 3 \cdot 10^7$ values of W. By Dvoretzky-Kiefer-Wolfowitz inequality (see [45] and the reference therein), with $\varepsilon = .0005$ and $m = 3 \cdot 10^7$,

$$P\left(\sup_t \left|F(t) - F^\dagger(t)\right| > \varepsilon\right) \le 2e^{-2m\varepsilon^2} \approx 6.118046 \times 10^{-7}, \tag{1.3.4}$$

so that theoretically, we can assure that $|F(t) - F^\dagger(t)| < .0005$ for any t with probability of 0.9999994. If we fix the probability p_0, then we have

$$2e^{-2m\varepsilon^2} = p_0 \Longrightarrow \varepsilon = \sqrt{\frac{\log 2/p_0}{2m}},$$

so that ε has the order of $m^{-1/2}$.

We check (1.3.4) for Exp(2) and Gamma(3,3) service time distributions by comparing $F^\dagger(t)$ with $F(t)$, where the latter is obtained from Theorem 26. See Figure 1.3.8 for $\log_{10}$-scaled absolute errors of $f^\dagger(t)$ (the kernel density estimations) and $F^\dagger(t)$ (the empirical CDF) from the random sample of $3 \cdot 10^7 W$. We also add the absolute errors of the saddlepoint approximations $F_0(t)$ and $f_0(t)$ for a comparison.

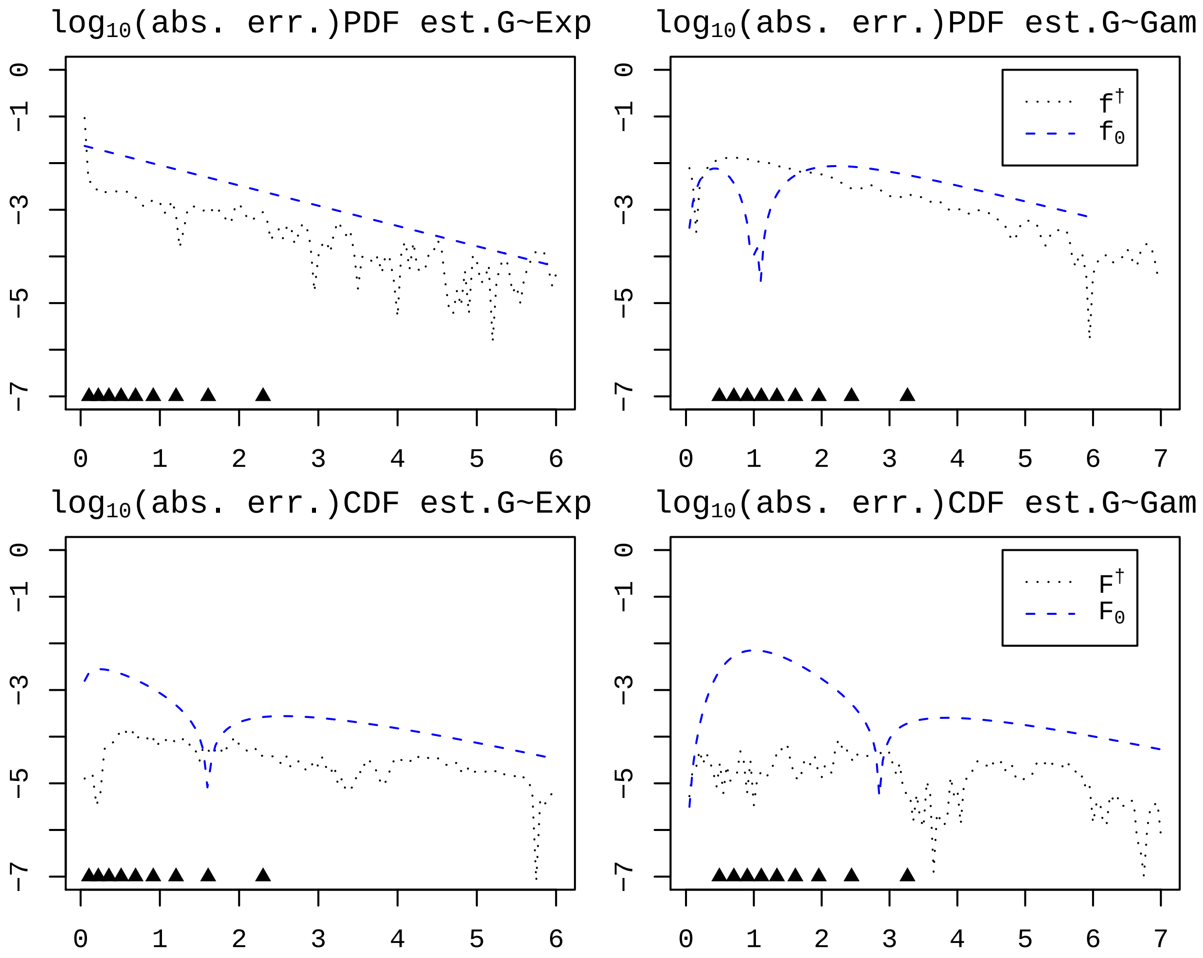

Figure 1.3.8: $\log_{10}$(absolute error) of PDF (top) and CDF (bottom) estimations of the stationary waiting time W of M/G/1 queue with Exp(2) (left) and Gamma(3,3) (right) service time distributions and $\rho = .5$. The dotted curves are of $f^{\dagger}(t)$ (the kernel density estimation) and $F^{\dagger}(t)$ (the empirical CDF) from the random samples of the size $3 \cdot 10^{7}$ and the dashed curves are of normalized $f_0(t)$ and $F_0(t)$. In each graph, ▲ denotes a decile (from 10% to 90%) of the distribution of W, the stationary waiting distribution.

Table 1.3.1: $h\left(t\right)$ and $H\left(t\right)$ of (1.3.1) for corresponding service time distributions.

G	$h\left(t\right)=\mu\left\{1-G\left(t\right)\right\}$	$H\left(t\right)=\int_0^t h\left(x\right)dx$
$\text{Exp}\left(2\right)$	$2e^{-2t}$	$1-e^{-2t}$
$\text{Gamma}\left(3,3\right)$	$\frac{1}{3}3e^{-3t}+\frac{1}{3}9te^{-3t}+\frac{1}{3}\frac{27}{2}t^2e^{-3t}$	$1-e^{-3t}-2te^{-3t}-\frac{3}{2}t^2e^{-3t}$
$\text{Beta}\left(2,2\right)$	$2\left(1-3t^2+2t^3\right)\mathbf{1}_{[0,1]}\left(t\right)$	$t\left(2-2t^2+t^3\right)\mathbf{1}_{[0,1]}\left(t\right)+\mathbf{1}_{(1,\infty)}(t)$
$\text{Pareto}\left(\frac{4}{5},5\right)$	$\mathbf{1}_{[0,4/5]}\left(t\right)+\left(\frac{4}{5}\right)^5t^{-5}\mathbf{1}_{(4/5,\infty)}\left(t\right)$	$t\,\mathbf{1}_{[0,4/5]}\left(t\right)+\left(1-\frac{256}{3125t^4}\right)\mathbf{1}_{(4/5,\infty)}\left(t\right)$

See Table (1.3.1) for the $H(t)$'s used for the random sample generation of V and W. For Beta$(2,2)$ and Pareto$(.8,5)$ service time distribution, $H\left(t\right)$ is a rational function and $H^{-1}\left(\cdot\right)$ can be calculated analytically to be used for the inverse transform algorithms.

Note that for Gamma$(3,3)$ service time distribution, $H^{-1}\left(\cdot\right)$ does not have a closed form but, since the distribution of V is a mixture of three Gamma distributions with the probability of $1/3$ each, we may use

$$V\sim\frac{1}{3}\text{Exp}\left(3\right)+\frac{1}{3}\text{Gamma}\left(2,3\right)+\frac{1}{3}\text{Gamma}\left(3,3\right)$$

to obtain the random sample of V.

The bootstrap CI's we are testing are the bootstrap standard percentile CI (BP), BCa, and the HDR method which are discussed in Chapter 4. Since the histogram for $\{\hat{F}^*_{0j}\left(t\right):1\leq j\leq 10^4\}$ is skewed to the left when t is large, the bootstrap standard CI is not considered. Also due to the fact that $\{\hat{F}^*_0\left(t\right)\}$ are skewed, we found that the bootstrap basic percentile confidence intervals perform poorly in the tail area of W, which is excluded in our simulation study.

The four example in Table 1.3.1(see table 4.1.2 for the PDF and CDF) are considered in Figure 1.3.9, 1.3.10, 1.3.11, and 1.3.12, respectively. Each figure shows one of the calculated CI's from the 2500 random samples $\{G_j,I_j\}_{j=1}^n$ used, the average coverage probabilities, and the average interval lengths.

Note that for the average interval lengths, we plot the average of $U\left(t\right)-L\left(t\right)$, $U\left(t\right)-F\left(t\right)$, and $L\left(t\right)-F\left(t\right)$, respectively, where $L\left(t\right)$ and $U\left(t\right)$ are the lower limit and upper limit of the confidence intervals. Thus, the top curves in the middle graphs are the average interval lengths, and the middle curves are the average distances between the upper limit of CI's, $U(t)$ and $F(t)$, and the bottom curves are the average distances between the lower limit $L(t)$ and $F(t)$. As mentioned, for the tail area of W, $\{\hat{F}_0^*(t)\}$ is skewed to the left, which leads to $F(t) - L(t)$ as the dominant part in the interval length for the tail area.

Two things need to be commented: First, the bigger sample size of n=200 gives us shorter interval lengths compared to those of the sample size n=50 cases and also the better coverage probabilities. Secondly, though HDR method has the shortest interval length among the three different CI's, its average coverage probabilities is still comparable to the bootstrap standard percentile method.

There is no clear winner among the three different CI's but if we need to choose one method considering the coverage probabilities only, BCa method would be recommended since it works consistently well over all the different cases.

1.3.6 When saddlepoint approximations need a caution

From our simulation study and experience, the saddlepoint approximation works well for either estimating $F(t)$ (when the true MGF of G is available and is a closed form) or $\hat{F}(t)$. However, we recommend checking whether the saddlepoint approximation $\hat{F}_0(t)$ is close to $\hat{F}^\dagger(t)$ or not before constructing the bootstrap CI because the approximation does not always give the percentage relative errors to first digits.

We recall that $\hat{F}_0(t)$ converges to $F_0(t)$ but not to $F(t)$. Since $\hat{F}_0^*(t)$ converges to $\hat{F}_0(t)$, the bootstrap CI will be "biased" as much as the difference between $F_0(t)$ and $F(t)$.

There are two possible causes for $\hat{F}_0(t)$ not approximating $\hat{F}^\dagger(t)$ (or $\hat{F}(t)$) well. One is that the sample is just a "bad" sample and increasing the sample size will be the remedy for this case. This is because as long as $F_0(t)$, the limit of $\hat{F}_0(t)$, is close enough to $F(t)$, the limit of $\hat{F}^\dagger(t)$ (as $n, m \to \infty$), then $\hat{F}_0(x)$ will approximate $\hat{F}^\dagger(t)$ (and $\hat{F}(t)$ as a result) when the sample size is large enough.

Another possible cause is that $F_0(t)$ itself is not a good approximation of $F(t)$ in which case, just raising the sample size will not cure the problem since $\hat{F}_0(t)$ and $\hat{F}^\dagger(t)$ converge

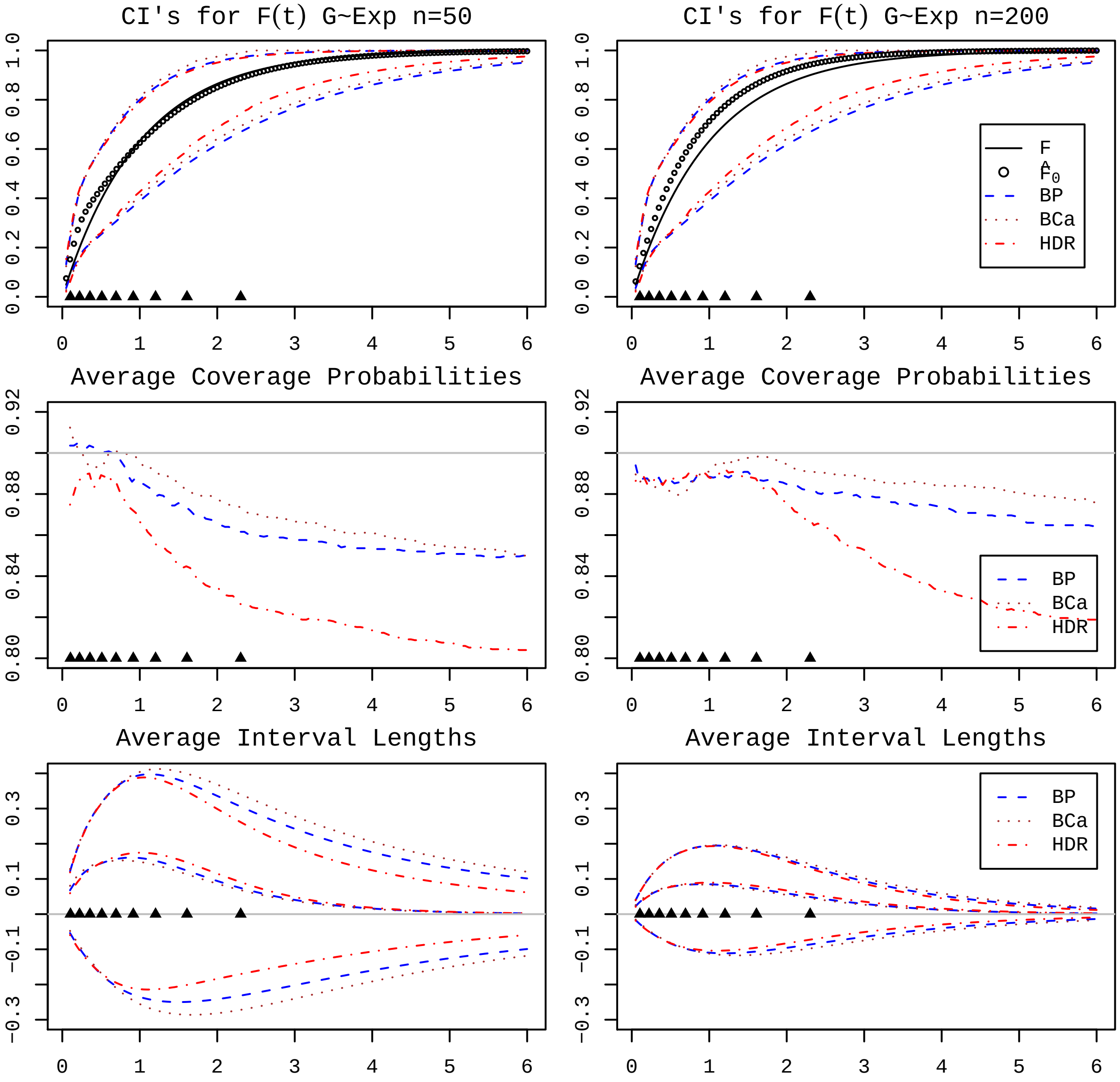

Figure 1.3.9: CI's for $F(t)$ where $G\sim\text{Exp}(2)$ and $\lambda = 1$. Top: Calculated CI's for a random sample of $\{G_j, I_j\}_{j=1}^n$ for n=50 (left) and n=200 (right). Middle: the average coverage probabilities from l=2500 random samples of $\{G_j, I_j\}_{j=1}^n$ for n=50 (left) and n=200 (right). Bottom: Average interval lengths of CI's. The top curves are the interval lengths, $U(t) - L(t)$ and the middle curves are of $U(t) - F(t)$ and bottom curves are of $L(t) - F(t)$. In each graph, ▲ denotes a decile (from 10% to 90%) of W, the stationary waiting distribution.

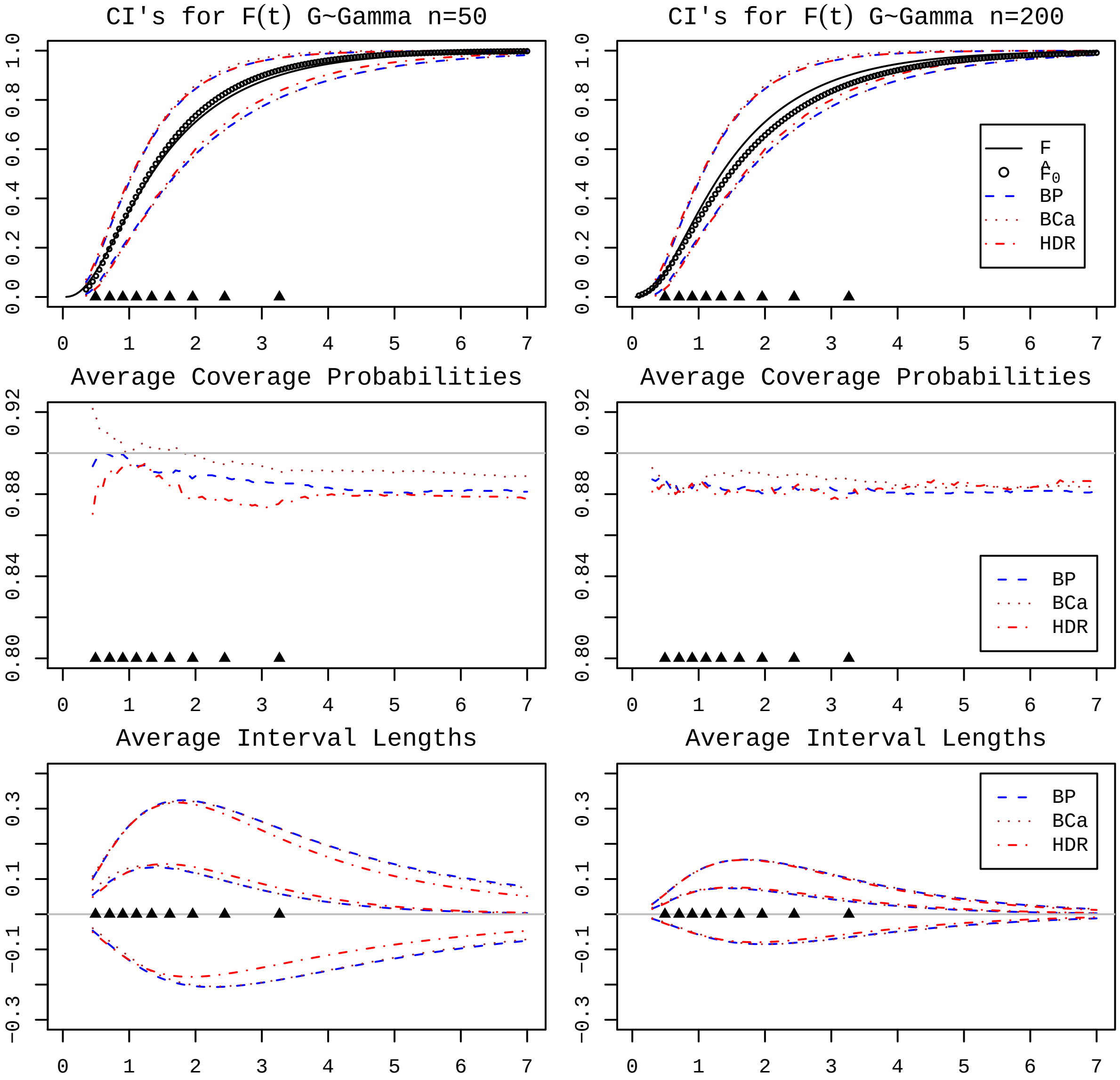

Figure 1.3.10: Same as Figure (1.3.9) for G~Gamma(3, 3).

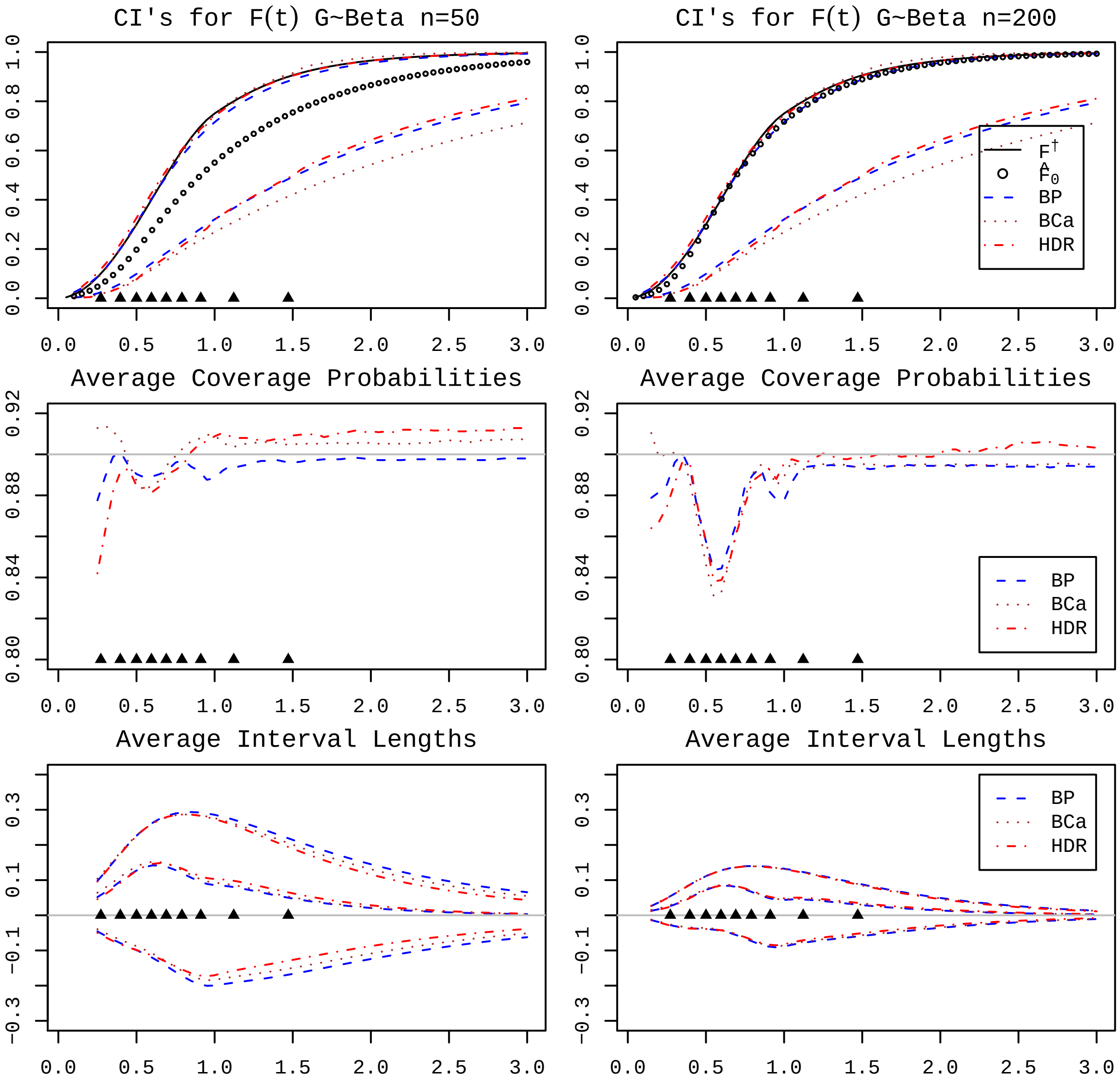

Figure 1.3.11: Same as Figure (1.3.9) for $G\sim$Beta$(2,2)$.

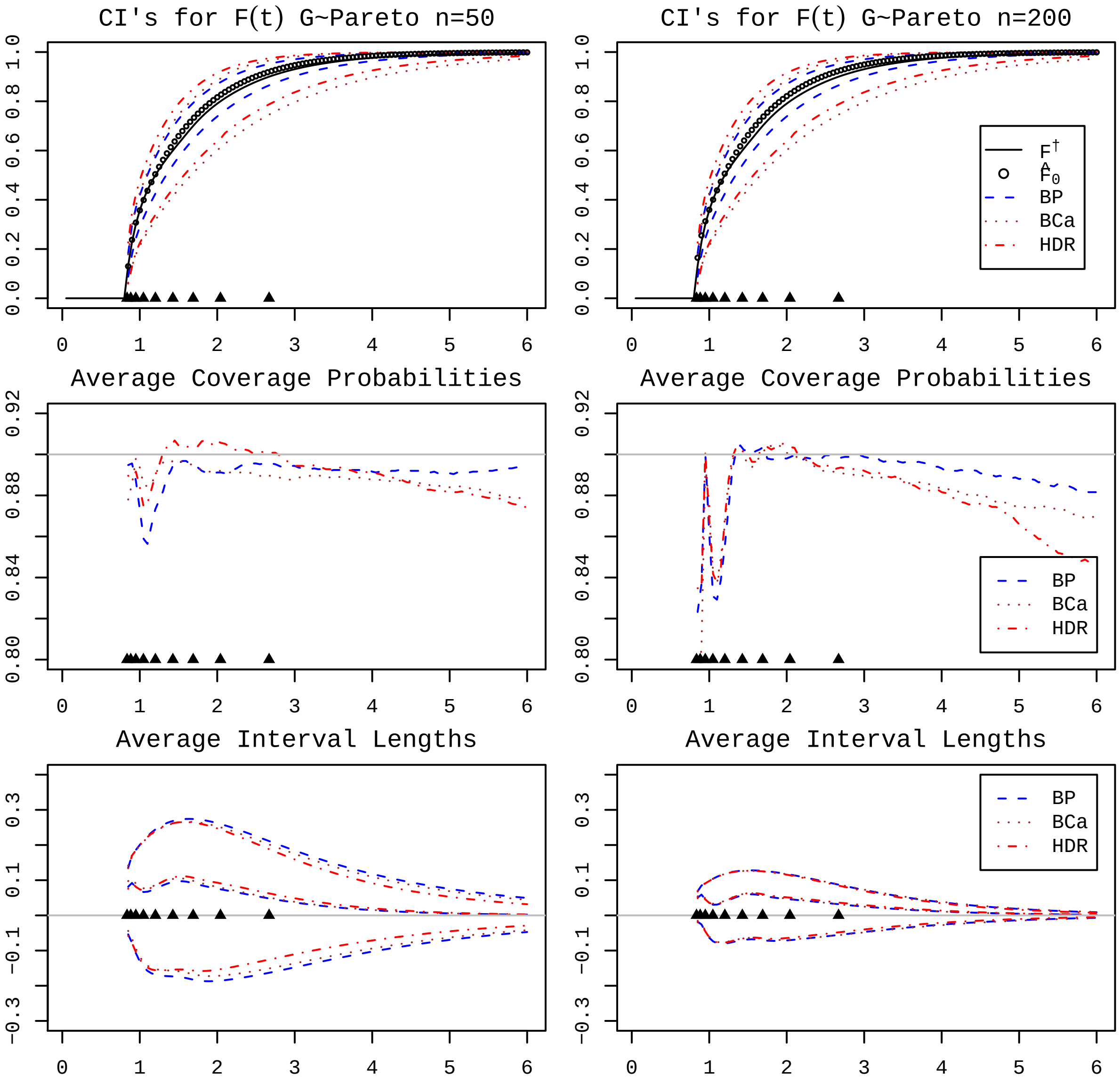

Figure 1.3.12: Same as Figure (1.3.9) for G~Pareto$(.8, 5)$.

to $F(t)$ while $\hat{F}_0(t)$ converges to $F_0(t)$.

Here are two examples for the two cases we mentioned in the above. In the first example, G~LogNormal$(-0.5, 1)$ and in the second example, G~Uniform$(0, 2)$. Note that in either case, μ=1 and we set λ=.5 to get ρ=.5.

See Figure 1.3.13 for the graphs of $\hat{F}_0(t)$ and its percentage relative error for the both cases. Here, the sample size of $\hat{W}$ to obtain $\hat{F}^{\dagger}(t)$ as an estimator of $\hat{F}(t)$ is set to $m = 10^7$.

We already mentioned that $\hat{F}_0(t)$ has higher relative errors on lower percentile area of $F_0(t)$ than the tail area due to the fact that $\hat{W} = \hat{W}_q + \hat{G}$ and $\hat{G}$ is discrete. However, these high relative errors were reduced after around the 40th percentile in Figure 1.3.7 unlike the G~LogNormal$(-.5, 1)$ case in Figure 1.3.13, in which the percentage relative errors are still close to 10% in 90th percentile.

Similarly, for the G~Uniform$(0, 2)$ case in Figure 1.3.13, the relative errors are persistently high till 70th percentile. Note that though the y-axis is cut to include -40% and 40% only to show the relative errors more clearly, the maximum was 193% which occurs at $t = .1$. In both examples, the errors are rather larger than the cases we covered in Figure 1.3.7 however in both examples, relative errors are acceptable above the 50th percentile.

For the case of G~LogNormal$(-5, 1)$, we also suspect that the persistence of the relative error to around 90th percentile area is due to the fact that the range of observed $\{G_j\}_{j=1}^{n}$ is rather large compared to the sample size of n=50. Note that LogNormal distribution is one of heavy-tail distributions (the MGF exists on the negative half axis only) and the range of $\{G_j\}$ used in Figure 1.3.13 was (.802, 6.79) with the interquartile range of (.358, .958). Thus, the discreteness of $\hat{G}$ is rather severe in this case.

In other words, the sample $\{G_j\}$ does not spread the range densely enough. The remedy of this case was a bigger size for the random sample. When we increase the sample size to 200 including the original sample of size 50, the percentage relative error between 30th percentile and 90th percentile were reduced about half from the case of n=50.

The case of G~Uniform$(0, 2)$ is different. Note that the domain of the distribution is the bound set of $(0, 2)$ so that the sample spreads densely on the interval enough unlike the first case of G~LogNormal. Also note that in Figure 1.3.7, though Beta$(2, 2)$ has a similar characteristic of the restricted domain of $(0, 1)$, the relative error was not severe as G~Uniform$(0, 2)$ of Figure 1.3.13.

Figure 1.3.14 shows the relative errors of $F_0\left(t\right)$ and $F_0\left(t\right) = (1-\rho) + \rho F_0^+\left(t\right)$ of W and W_q against their empirical CDF estimation $F^\dagger\left(t\right)$'s where the simulation size is $m = 3 \cdot 10^7$. Plots of other approximation methods were also drawn, which will be dealt in Chapter 2. It is known that the uniform distribution is a distribution for which the saddlepoint approximation dose not work exceptionally well, which can be seen by comparing the shape of $F^\dagger\left(t\right)$ and $F_0\left(t\right)$ on $(0, 2)$ of the top right graph in Figure 1.3.14. Also compare this characteristic with the case of W_q in the the top left graph of the same figure.

Because this high relative error is inherited from the poor performance of $F_0\left(t\right)$, even for the sample size of n=200, $\hat{F}_0\left(t\right)$ consistently overestimates from around 30th percentile till around 70th percentile as $\hat{F}_0\left(t\right)$ of $n = 50$ case does though the percentage relative errors are smaller and within 10%.

Thus, if raising the sample size does not improve the error, we may suspect that (unknown) $F_0\left(t\right)$ does not approximate (unknown) $F\left(t\right)$ well.

1.4 Estimation of W and W_q in M/E$_k$/1 queues

In this section, we concentrate on the parametric estimations of W in M/E_k/1 queue and compare the parametric bootstrap confidence band using the parametric estimation to the confidence band of the previous section, the confidence band based on (nonparametric) bootstrap with the saddlepoint approximation.

1.4.1 Inverse Laplace transform of $\mathcal{W}\left(s\right)$ and $\mathcal{W}_q\left(s\right)$

It is known that for M/G/1 queues, if the service time distribution is a phase-type distribution, than the stationary waiting time distribution is also a phase type distribution. See [9, 47, 48] for a proof.

Since $\mathcal{W}\left(s\right)$ of M/E_k/1 queue is a rational function, built-in routines of inverse Laplace

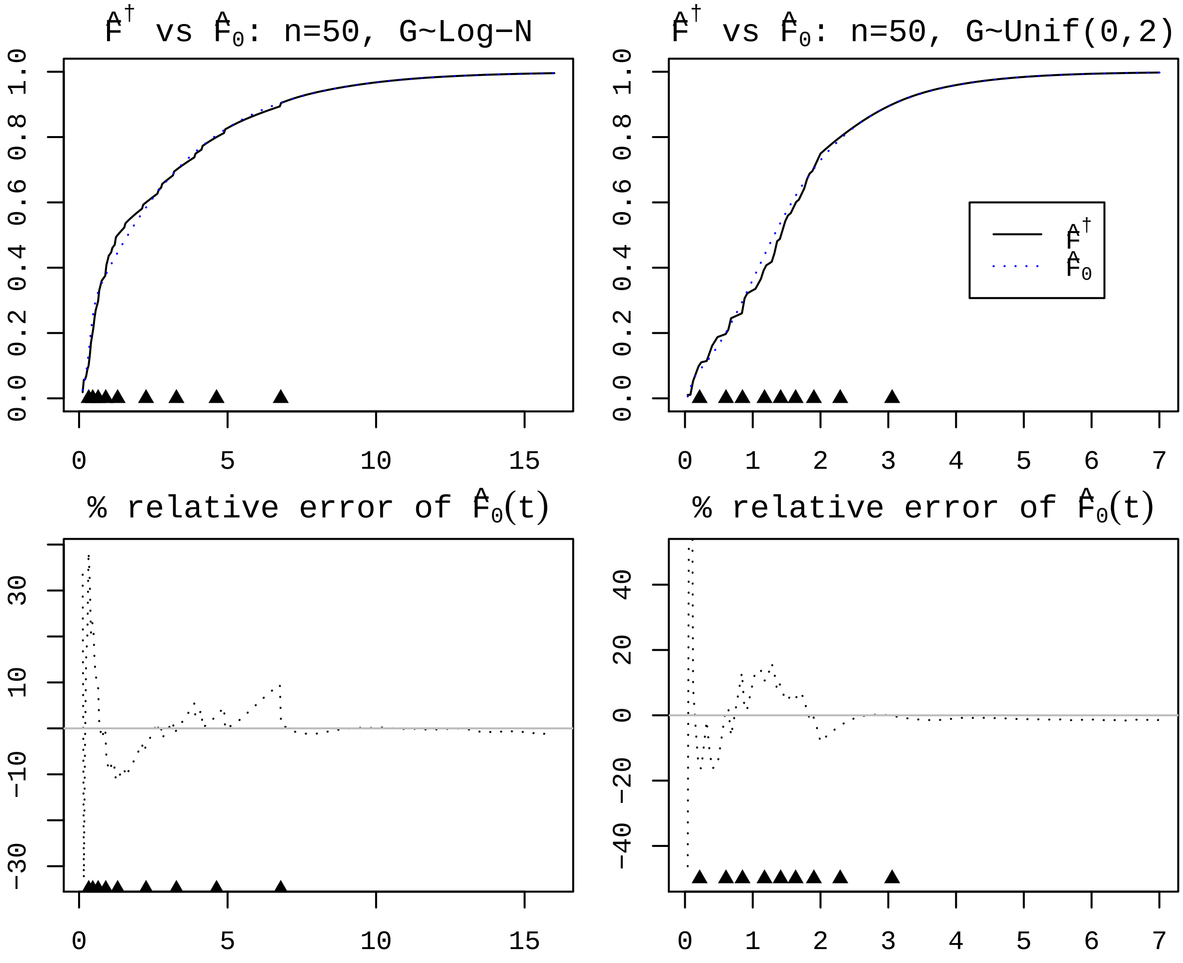

Figure 1.3.13: The cases of the saddlepoint approximation does not work so well. Left: $\hat{F}_0\,(t)$ and $\hat{F}^{\dagger}\,(t)$ for $G{\sim}\text{LogNormal}(-.5, 1)$ and its % relative error with $n = 50$. Right: $\hat{F}_0\,(t)$ and $\hat{F}^{\dagger}\,(t)$ for $G\ {\sim}\text{Uniform}(0, 2)$ and its % relative error with $n = 50$. In each graph, ▲ denotes a decile (from 10% to 90%) of $\hat{W}^*$.

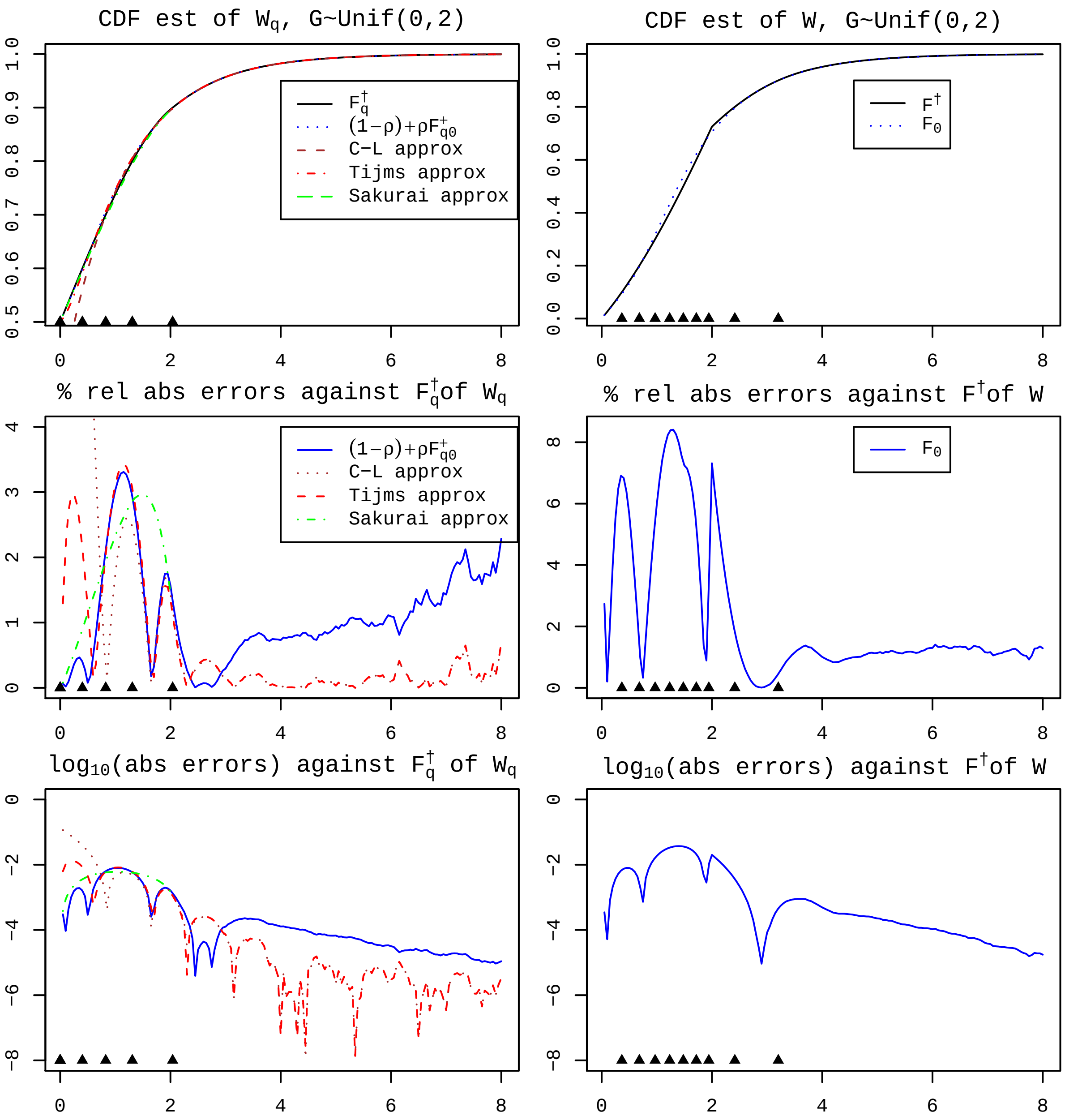

Figure 1.3.14: CDF estimations of W_q(left) and W (right) for G ~Uniform(0,2) with λ=.5. In each graph, ▲ denotes a decile (from 10% to 90%) of W_q and W, the estimated stationary waiting distribution.

transform of well-known symbolic mathematics programs like Mathematica can be used to find the PDF and CDF of the stationary waiting time distribution.

In fact, we can calculate the PDF and the CDF by hand for Gamma(3,3) service time distribution with $\lambda = .5$. We have

$$\begin{aligned}\mathcal{L}\left\{f\left(t\right)\right\}\left(s\right) = \mathcal{W}\left(-s\right) &= \frac{13.5}{13.5 + s\left(22.5 + s\left(8.5 + s\right)\right)} \\ &= \frac{13.5}{\left(s + .840474\right)\left(s^2 + 7.65943s + 16.0624\right)},\end{aligned}$$

which can be decomposed into partial fractions.

Then, using the Laplace transform of $\left(s + a\right)^{-1}$ is e^{-at} and the (inverse) Laplace transform is linear, we can obtain the PDF. See the appendix of [41] for related explanation and examples.

This result is known as the expansion theorem (see [20] or [25]): Let a Laplace transform $\mathcal{L}\{f\left(t\right)\}$ be a rational function having a form of

$$\frac{B\left(s\right)}{A\left(s\right)},$$

where degree of $A\left(s\right)$ is greater than $B\left(s\right)$. Suppose

$$A\left(s\right) = \left(s - a_1\right)\left(s - a_2\right)\cdots\left(s - a_k\right),$$

where a_j are all distinct. Then, the partial fraction expansion of $\mathcal{L}\{f\left(t\right)\}$ is

$$\mathcal{L}\left\{f\left(t\right)\right\}\left(s\right) = \sum_{j=1}^{k} \frac{B\left(a_j\right)}{\left(s - a_j\right)A'\left(a_j\right)},$$

which implies

$$f\left(t\right) = \sum_{j=1}^{k} \frac{B\left(a_j\right)}{A'\left(a_j\right)} e^{a_j t}.$$

Note that

$$\mathcal{L}\left\{F\left(x\right)\right\}\left(s\right)=\frac{\mathcal{L}\left\{f\left(x\right)\right\}}{s},$$

so that the CDF of W can be obtained similarly.

Generally, we can invert the Laplace transform of $\mathcal{W}\left(-s\right)$ of M/E_k/1 queue analytically to get the CDF and the PDF as a closed form using the expansion theorem.

Theorem 26. [5]*For M/E_k/1 queues with the service time distribution of Gamma(k,β), the CDF and the PDF of the waiting time distribution are*

$$\begin{aligned}f\left(t\right)&=\left(1-\rho\right)\beta^{k}\sum_{j=1}^{k}\frac{\left(a_j-\beta\right)^2\exp\left(a_j-\beta\right)t}{ka_j^{k-1}\left(a_j-\beta\right)\left(a_j-\beta-\lambda\right)+\lambda\left(a_j^k-\beta^k\right)}\\ F\left(t\right)&=1-\left(1-\rho\right)\beta^{k}\sum_{j=1}^{k}\frac{\exp\left(a_j-\beta\right)t}{a_j^{k-1}\left\{k\left(\lambda+\beta\right)-\left(k+1\right)a_j\right\}},\end{aligned}$$

where a_j's are the (distinct) roots of polynomial of r,

$$r^{k+1}-\left(\lambda+\beta\right)r^k+\lambda\beta^k, \tag{1.4.1}$$

which is not β (β is a root of $r^{k+1}-\left(\lambda+\beta\right)r^k+\lambda\beta^k$).

For W_q, the stationary waiting time in the queue, we will use $F_q\left(t\right)$ for the CDF and $f_q\left(t\right)$ for the PDF. Then we have

$$\begin{aligned}f_q\left(t\right)&=\left(1-\rho\right)\mathbf{1}_{\{0\}}\left(x\right)+\lambda\left(1-\rho\right)\sum_{j=1}^{k}\frac{\left(a_j^k-\beta^k\right)\left(a_j-\beta\right)\exp\left(a_j-\beta\right)t}{ka_j^{k-1}\left(a_j-\beta\right)\left(a_j-\beta-\lambda\right)+\lambda\left(a_j^k-\beta^k\right)}.\\ F_q\left(t\right)&=1-\left(1-\rho\right)\sum_{j=1}^{k}\frac{a_j\exp\left(a_j-\beta\right)t}{k\left(\lambda+\beta\right)-\left(k+1\right)a_j}.\end{aligned}$$

[5] Originally, I was sure that this theorem should have been obtained already but was not able to find such result in any queuing theory textbook - Because M/E_k/1 queue is one of the standard example used in texts, I assumed that if the theorem was known before, it should appear in some of the textbooks. Since I could not find any reference other than [47, 48], and [9] of the matrix-geometric solution for phase type service distributions, which is clearly an overkill, I decided to prove it myself. Of course, later I found a reference which is dated back to 1953 (see Remark 27).

Proof. The MGF of Gamma(k, β) is

$$\left(\frac{1}{1 - s/\beta}\right)^k,$$

so that the Laplace transform of the PDF $f(t)$ is

$$\begin{aligned}\mathcal{L}\{f(t)\}(s) = \mathcal{W}(-s) &= \frac{(1 - \lambda k/\beta)\, s\mathcal{G}(-s)}{s - \lambda + \lambda\mathcal{G}(-s)} = \cdots \\ &= \frac{(\beta - k\lambda)\,\beta^{k-1}s}{s(s+\beta)^k - \lambda(s+\beta)^k + \lambda\beta^k} = \frac{(1-\rho)\,\beta^k s}{s(s+\beta)^k - \lambda(s+\beta)^k + \lambda\beta^k}.\end{aligned}$$

Let

$$Q(s) := s(s+\beta)^k - \lambda(s+\beta)^k + \lambda\beta^k \tag{1.4.2}$$

and because $Q(0) = 0$, we know $Q(s)$ has the s term, which can be seen also by

$$\begin{aligned}Q(s) &= s\sum_{j=0}^{k}\binom{k}{j}s^{k-j}\beta^j - \lambda\left\{\sum_{j=0}^{k-1}\binom{k}{j}s^{k-j}\beta^j + \beta^k\right\} + \lambda\beta^k \\ &= s\sum_{j=0}^{k}\binom{k}{j}s^{k-j}\beta^j - \lambda s\left\{\sum_{j=0}^{k-1}\binom{k}{j}s^{k-j-1}\beta^j\right\} \\ &= s\left\{\sum_{j=0}^{k}\binom{k}{j}s^{k-j}\beta^j - \lambda\sum_{j=0}^{k-1}\binom{k}{j}s^{k-j-1}\beta^j\right\}\end{aligned}$$

and we have

$$\mathcal{L}\{f(t)\}(s) = (\beta - k\lambda)\,\beta^{k-1}\left\{\sum_{j=0}^{k}\binom{k}{j}s^{k-j}\beta^j - \lambda\sum_{j=0}^{k-1}\binom{k}{j}s^{k-j-1}\beta^j\right\}^{-1}. \tag{$*$}$$

By the fundamental theorem of algebra, there are $k+1$ roots of $Q(s)$ including 0. We now show that $Q(s)$ has simple roots only so that the roots are all distinct. To simplify calculations, let $r = s + \beta$. Then

$$Q(s) = (r - \beta)\,r^k - \lambda r^k + \lambda\beta^k = r^{k+1} - (\lambda + \beta)\,r^k + \lambda\beta^k =: Q_1(r) \tag{1.4.3}$$

and if a is a root of $Q_1(r)$ then $a-\beta$ is a root of $Q(s)$ and vice versa. Note that a and $(a-\beta)$ has the same multiplicity as a root of $Q_1(r)$ and $Q(s)$, respectively.

Suppose that there is a root a_j whose multiplicity is greater than 1, which implies that

$$Q_1(a_j) = Q_1'(a_j) = 0.$$

Because we have

$$Q_1'(r) = (k+1)r^k - k(\lambda+\beta)r^{k-1} = r^{k-1}\{(k+1)r - k(\lambda+\beta)\},$$

the roots of $Q_1'(r)$ are either 0 of multiplicity $k-1$ or

$$r = \frac{k}{k+1}(\lambda+\beta).$$

Since $Q_1(0) = \lambda\beta^k \neq 0$, a_j must be $(\lambda+\beta)k/(k+1)$.

We now show $(\lambda+\beta)k/(k+1)$ cannot be a root of $Q_1(r)$ when the stationary condition $\rho < 1$ is satisfied, which gives us that there is no root whose multiplicity is greater than 1. To this end, we define $Q_2(\lambda)$ by

$$Q_2(\lambda) := Q_1\left\{\frac{k}{k+1}(\lambda+\beta)\right\} = \lambda\beta^k - \frac{1}{k}\left(\frac{k}{k+1}\right)^{k+1}(\lambda+\beta)^{k+1}.$$

Then, we find a real root of $Q_2'(\lambda)$ by calculating

$$\begin{aligned}\frac{dQ_2(\lambda)}{d\lambda} = \beta^k - \left(\frac{k}{k+1}\right)^k(\lambda+\beta)^k = 0 \Longleftrightarrow \beta^k &= \left(\frac{k}{k+1}\right)^k(\lambda+\beta)^k \\ \Longrightarrow \beta &= \left(\frac{k}{k+1}\right)(\lambda+\beta) \Longrightarrow \lambda = \frac{\beta}{k}\end{aligned}$$

so that the (only) real root of $Q_2'(\lambda)$ is β/k because $\lambda, \beta, k > 0$. Since $Q_2''(\beta/k) \neq 0$ and $Q_2(\beta/k) = 0$, we can see that β/k is the only real root of $Q_2(\lambda)$ of the multiplicity 2. Since $Q_2''(\lambda) < 0$, $Q_2'(\lambda)$ is non-increasing.

Thus, with the fact that β/k is the only real root of $Q_2'(\lambda)$, we conclude that

$$Q_2'(\lambda)\begin{cases} >0 & \text{if } \lambda<\beta/k \\ <0 & \text{if } \lambda>\beta/k \end{cases},$$

so that $Q_2(\lambda)$ has the global maximum 0 at β/k. Because of the stationary condition

$$\rho=\lambda\frac{k}{\beta}<1 \Longleftrightarrow \lambda<\frac{\beta}{k},$$

$Q_2(\lambda)$ will be always less than 0 so that $(\lambda+\beta)k/(k+1)$ cannot be a root of $Q_1(r)$.

Therefore, every root of $Q_1(s)$ is distinct and we can apply the expansion theorem. From $(*)$, instead of using

$$d\left\{\sum_{j=0}^{k}\binom{k}{j}s^{k-j}\beta^j-\lambda\sum_{j=0}^{k-1}\binom{k}{j}s^{k-j-1}\beta^j\right\}/ds$$

directly, we use

$$\begin{aligned}\frac{d\{Q(r)/s\}}{ds} &= \frac{ks(s+\beta)^{k-1}(s-\lambda)+\lambda\left(-\beta^k-(s+\beta)^k\right)}{s^2}\\ &= \frac{kr^{k-1}(r-\beta)(r-\beta-\lambda)+\lambda\left(r^k-\beta^k\right)}{(r-\beta)^2}\end{aligned}$$

for simplicity of our formula and we obtain the desired result.

For $F(t)$, using

$$\mathcal{L}\{F(t)\}(s)=\frac{\mathcal{L}\{f(t)\}(s)}{s}=\frac{(1-\rho)\beta^k}{Q(s)}$$

and

$$Q'(0)=\beta^k-k\lambda\beta^{k-1}=(1-\rho)\beta^k,$$

we have

$$F(t)=(1-\rho)\beta^k\left(\frac{e^{0\cdot x}}{Q'(0)}+\sum_{j=1}^{k}\frac{e^{(a_j-\beta)w}}{Q'(a_j-\beta)}\right),$$

which results in the desired result.

For W_q, the stationary waiting time in the queue, it is known that

$$\mathcal{W}_q(s) = \frac{(1-\rho)s}{s+\lambda-\lambda\mathcal{G}(s)},$$

so that we have

$$\begin{aligned}\mathcal{L}\{f_q(t)\}(s) = \mathcal{W}_q(-s) &= \frac{(1-k\lambda/\beta)s}{s-\lambda+\lambda\mathcal{G}(-s)}\\ &= \frac{(1-\rho)s(s+\beta)^k}{s(s+\beta)^n-\lambda(s+\beta)^k+\lambda\beta^k} = (1-\rho)\left[1+\frac{\lambda\left\{(s+\beta)^k-\beta^k\right\}}{Q_1(r)}\right],\end{aligned}$$

Using the previous calculation and the fact that $\mathcal{L}\{\mathbf{1}_{\{0\}}(x)\}=1$, we have

$$f_q(t) = (1-\rho)\mathbf{1}_{\{0\}}(x) + \sum_{j=1}^{k} \frac{\lambda(1-\rho)\left(a_j^k-\beta^k\right)(a_j-\beta)e^{(a_j-\beta)w_q}}{ka_j^{k-1}(a_j-\beta)(a_j-\beta-\lambda)+\lambda\left(a_j^k-\beta^k\right)}.$$

Again, since $\mathcal{L}\{F_q(t)\}(s) = \mathcal{L}\{f_q(t)\}/s$, we have

$$\mathcal{L}\{F_q(t)\}(s) = \frac{(1-\rho)r^k}{Q_1(r)}$$

and in a similar way, we obtain

$$\begin{aligned}F_q(t) &= (1-\rho)\left\{\frac{\beta^k}{(1-\rho)\beta^k} + \sum_{j=1}^{k}\frac{a_j^k e^{(a_j-\beta)t}}{a_j^{k-1}\{(k+1)a_j-k(\lambda+\beta)\}}\right\}\\ &= 1-(1-\rho)\sum_{j=1}^{k}\frac{a_j e^{(a_j-\beta)w_q}}{k(\lambda+\beta)-(k+1)a_j}.\end{aligned}$$

□

Remark 27. We later found that in [64], the formula of the PDF of the waiting time distribution for G/G/1 queue were obtained when both MGF of the service time distribution and the inter-arrival time are reciprocals of polynomial functions, which includes M/E_k/1 .

See theorem 4 and the example in §5.3 therein. However, the roots of $Q_1(s)$ being distinct were supposed, which we proved in Theorem 26.

Example 28. For a M/M/1 queue, $k = 1$ and $\beta = \mu$ and

$$r^2 - (\lambda + \mu) r + \lambda\mu = 0$$

implies $a_1 = \lambda$ and

$$f(t) = \left(1 - \frac{\lambda}{\mu}\right) \mu \frac{(\lambda - \mu)^2 e^{(\lambda-\mu)t}}{(\lambda - \mu)(-\mu) + \lambda(\lambda - \mu)} = (\mu - \lambda) e^{-(\mu-\lambda)t}$$

and

$$F(t) = 1 - \left(1 - \frac{\lambda}{\mu}\right) \mu \frac{e^{(\lambda-\mu)t}}{(\lambda + \mu) - 2\lambda} = 1 - e^{-(\mu-\lambda)t},$$

which coincides the known result of $W \sim \text{Exp}(\mu - \lambda)$. For W_q, we have

$$f_q(t) = (1 - \rho)\mathbf{1}_{\{0\}}(x) + \lambda(1 - \rho) e^{-(\mu-\lambda)t}.$$
$$F_q(t) = 1 - \left(1 - \frac{\lambda}{\mu}\right) \frac{\lambda e^{-(\mu-\lambda)x}}{\mu - \lambda} = 1 - \rho e^{-(\mu-\lambda)t}.$$

For $G \sim$Gamma(3,3) with $\lambda = .5$, we have $a_1 - \beta = a_1 - 3 = -0.8404738$ and $a_2 - 3$ and $a_3 - 3$ are $-3.8297631 \pm 1.1812212i$. Recall that for $x, y \in \mathbb{R}$,

$$e^{x+yi} = e^x (\cos y + i \sin y),$$

and if $b < a < 0$, then

$$c_1 e^{ax} + c_2 e^{bx} = c_1 e^{ax} \left\{1 + \frac{c_2}{c_1} e^{(b-a)x}\right\} = c_1 e^{ax} \left\{1 + O\left(e^{(b-a)x}\right)\right\} = c_1 e^{ax} \left\{1 + o\left(e^{-\varepsilon x}\right)\right\}$$

for any $\varepsilon < |b-a|$. Thus, we have

$$F(t) = 1 - 1.5547537e^{-0.8404738t}\left\{1 + o\left(e^{-\varepsilon t}\right)\right\}$$
$$F_q(t) = 1 - 0.57992693e^{-0.8404738t}\left\{1 + o\left(e^{-\varepsilon t}\right)\right\}$$

for any $\varepsilon < 2.9892893$.

We can have the similar results for the PDF's, which are the derivatives of the above asymptotic approximation.

For the graph of the asymptotic of $F(t)$ and $f(t)$, see Figure 1.3.4.

Remark 29. As we saw in the previous example, a_j can be a complex number. However, $F(t)$ and the others given in Theorem 26 are all real valued; For $a \in \mathbb{C}$, let $\overline{a}$ be conjugate of a (i.e., $\overline{x+yi} = x - yi$ for $x, y \in \mathbb{R}$).

It is known that if a_j is a root of real polynomial function, then $\overline{a_j}$ is also a root of the polynomial ($0 = \overline{Q_1(a_j)} = Q_1(\overline{a_j})$ in our case). Thus,

$$\begin{aligned}\overline{F(t)} &= \overline{\left(1 - (1-\rho)\beta^k \sum_{j=1}^{k} \frac{e^{(a_j-\beta)t}}{a_j^{k-1}\left\{k(\lambda+\beta) - (k+1)a_j\right\}}\right)} \\ &= 1 - (1-\rho)\beta^k \sum_{j=1}^{k} \frac{e^{(\overline{a_j}-\beta)t}}{(\overline{a_j})^{k-1}\left\{k(\lambda+\beta) - (k+1)\overline{a_j}\right\}} = F(t),\end{aligned}$$

so that $F(t)$ is a real number.

The above example of the asymptotic result can be generalized:

Corollary 30. *Among the roots $\{a_j\}_{j=1}^k$ of (1.4.1) other than β, there exists only one positive real roots and it has the maximum absolute value and the maximum real part among all the roots. Thus, if we call the real root a_1, then*

$$1 - F(t) = (1-\rho)\beta^k \frac{e^{(a_1-\beta)t}}{a_1^{k-1}\left\{k(\lambda+\beta) - (k+1)a_1\right\}}\left\{1 + o\left(e^{-\varepsilon t}\right)\right\},$$
$$1 - F_q(t) = (1-\rho)\frac{a_1 e^{(a_1-\beta)x}}{k(\lambda+\beta) - (k+1)a_1}\left\{1 + o\left(e^{-\varepsilon t}\right)\right\}$$

for any $\varepsilon > \max_{2\leq j\leq k}\{a_1 - \Re a_j\}$.

Proof. First, we show that there exists only one real root of $Q_1(r)$ other than β. Define $\mathcal{D}(s)$ as in the proof of Lemma 8 by

$$\mathcal{D}(s) := s + \lambda - \lambda \mathcal{G}(s) = s + \lambda - \lambda \left(\frac{\beta}{\beta - s}\right)^k .$$

Then for $s \neq \beta$, $\mathcal{D}(s) = 0$ implies

$$0 = s(\beta - s)^k + \lambda(\beta - s) - \lambda\beta^k = -Q(-s),$$

where $Q(s)$ is defined in (1.4.2). Thus, if $a_i - \beta$ is a root of $Q(s)$ (, or a_i is a root of $Q_1(r)$), then $\beta - a_i$ is also a root of $\mathcal{D}(s)$. In Lemma 2, we showed that there exists a unique positive real root of $\mathcal{D}(s)$, c which is smaller than $b = \beta$. Thus, $a_1 - \beta$ must be $-c$.

Let $\Re a$ be the real part of a (i.e., $\Re(x + yi) = x$ for $x, y \in \mathbb{R}$). We note that since $F(t)$ and $F_q(t)$ are CDF's, $\Re(a_j - \beta) < 0$, or

$$\Re a_j < \beta \tag{$*$}$$

(if not, $e^{(a_j - \beta)t} \nearrow \infty$ as $t \to \infty$).

For each root a_j, $j > 1$, we have

$$Q_1(a_j) = 0 \Longleftrightarrow a_j^k \{a_j - (\lambda + \beta)\} = -\lambda\beta^k,$$

which implies

$$|a_j|^k |a_j - (\lambda + \beta)| = \lambda\beta^k.$$

By $(*)$, we have $\Re a_j - (\lambda + \beta) < -\lambda$ so that

$$|a_j - (\lambda + \beta)| > \lambda,$$

which implies $|a_j|^k < \beta^k$, or

$$|a_j| < \beta.$$

Since $(\lambda + \beta) > 0$, we have

$$|a_j - (\lambda + \beta)| \geq ||a_j| - (\lambda + \beta)|\,,$$

which implies

$$|a_j|^k\,||a_j| - (\lambda + \beta)| \leq \lambda\beta^k,$$

so that we have

$$Q_1\left(|a_j|\right) = |a_j|^k\left\{|a_j| - (\lambda + \beta)\right\} + \lambda\beta^k = -\,|a_j|^k\,||a_j| - (\lambda + \beta)| + \lambda\beta^k > 0.$$

From the result of proof of Lemma 2, we have the following table of $Q_1\left(r\right)$:

r	0	...	a_1	$\frac{k(\lambda+\beta)}{k+1}$	β	...	$\lambda+\beta$
$Q_1'\left(r\right)$	0	-	-	0	+	+	+
$Q_1\left(r\right)$	$\lambda\beta^k$	+	0	-	0	+	$\lambda\beta^k$

Because $0 < |a_j| < \beta$ and $Q_1(|a_j|) > 0$, we conclude that

$$|a_j| < a_1 \quad \text{and} \quad \Re a_j < a_1$$

and the argument of Example 28 can be applied to obtain the desired results. □

Example 31. The asymptotic result for $F_q\left(t\right)$ in Corollary 30 coincides with the Cramér-Lundberg approximation. Recall that $\gamma = \left(1 - \rho\right)/\left\{\lambda\mathcal{G}'\left(c\right) - 1\right\}$ and $\mathcal{G}\left(s\right) = \left\{\beta/\left(\beta - s\right)\right\}^k$. We have

$$\mathcal{G}'\left(s\right) = \frac{k}{\beta}\left(\frac{\beta}{\beta - s}\right)^{k+1} = \frac{k}{\beta - s}\mathcal{G}\left(s\right).$$

Then, using $c = \beta - a_1$ and $\lambda - \lambda\mathcal{G}(c) + c = 0$, we obtain

$$\lambda\mathcal{G}'(c) - 1 = \frac{k}{a_1}\lambda\mathcal{G}(c) - 1 = \frac{k(\lambda + \beta - a_1)}{a_1} - 1 = \frac{k(\lambda + \beta) - (k+1)a_1}{a_1},$$

which implies

$$\gamma = (1 - \rho)\frac{a_1}{k(\lambda + \beta) - (k+1)a_1}.$$

1.4.2 Parametric estimation of the CDF and the PDF of W and W_q

Theorem 26 gives us estimators of CDF and PDF of W and W_q when we know that the queues in question is M/E$_k$/1 and the random sample of inter-arrival $\{I_j\}$ and the service time $\{G_j\}$ are available.

From the sample, we obtain $\hat{\lambda}$ and $\hat{\beta}$, which give us

$$\hat{Q}_1(r) = r^{k+1} - \left(\hat{\lambda} + \hat{\beta}\right) r^k + \hat{\lambda}\hat{\beta}^k.$$

Solving $\hat{Q}_1(r) = 0$, we obtain $\hat{a}_j$ and by replacing a_j, λ, and β with $\hat{a}_j$, $\hat{\lambda}$, and $\hat{\beta}$ in the formula in Theorem 26, we have the estimator of the PDF's and CDF's. For example,

$$\hat{F}(t) = 1 - (1 - \hat{\rho})\hat{\beta}^k \sum_{j=1}^{k} \frac{\left(\hat{a}_j - \hat{\beta}\right)^2 e^{\left(\hat{a}_j - \hat{\beta}\right)t}}{k\hat{a}_j^{k-1}\left(\hat{a}_j - \hat{\beta}\right)\left(\hat{a}_j - \hat{\beta} - \hat{\lambda}\right) + \hat{\lambda}\left(\hat{a}_j^k - \hat{\beta}^k\right)} \tag{1.4.4}$$

The validity of our estimators comes from following:

Theorem 32. *Suppose* $\{G_j\}_{j=1}^n \overset{\text{iid}}{\sim}$ *Gamma*(k, β) *with* k *known and* $\{I_j\}_{j=1}^n \overset{\text{iid}}{\sim}$ *Exp*(λ)*. If we define* $\hat{\lambda} = n/\sum I_j$ *and* $\hat{\beta} = nk/\sum G_j$*, then we have the following:*

1. *As a complex function,* $\hat{Q}_1(z) = z^{k+1} - (\hat{\lambda} + \hat{\beta})z^k + \hat{\lambda}\hat{\beta}^k$ *converges locally uniformly to* $Q_1(z) = z^{k+1} - (\lambda + \beta)z^k + \lambda\beta^k$.

2. *For any* $z \in \mathbb{C}$, $|\hat{Q}_1(z) - Q(z)| = O_p\left(n^{-1/2}\right)$.

3. *Moreover, for each* $r \in \mathbb{R}$,

$$\sqrt{n}\left\{\hat{Q}_1(r) - Q_1(r)\right\} \Rightarrow N\left\{0, \lambda^2\left(t^k - \beta^k\right)^2 + k\beta^2(t^k - k\lambda\beta^{k-1})^2\right\}$$

4. *The roots of the equation* $\hat{Q}_1(r) = 0$ *converges to the roots of* $Q_1(r) = 0$ *a.s.*

5. *Let* $\hat{a}_1$ *be the unique positive real root of* $\hat{Q}_1(r) = 0$ *which converges to* a_1 *which is the unique positive real root of* $Q_1(r) = 0$. *Then,*

$$\sqrt{n}\left(\hat{a}_1 - a_1\right) \Rightarrow N\left(0, \frac{\lambda^2\left(a_1^k - \beta^k\right)^2 + k\beta^2(a_1^k - k\lambda\beta^{k-1})^2}{\left[a_1^{k-1}\left\{(k+1)a_1 - k(\lambda+\beta)\right\}\right]^2}\right).$$

6. *Let* $\hat{a}_j$ *be a root of* $\hat{Q}(r) = 0$, $j > 1$, *which converges to* a_j. *Then,* $\hat{a}_j - a_j = O_p\left(n^{-1/2}\right)$.

Proof. We show (1) first. Observe that if D is a compact set in $\mathbb{C}$, then for any $z \in D$,

$$\begin{aligned}\left|\hat{Q}_1(z) - Q_1(z)\right| &= \left|z^k\left(\hat{\lambda} + \hat{\beta} - \lambda - \beta\right) + \left(\hat{\lambda}\hat{\beta}^k - \lambda\beta^k\right)\right| \\ &\le \left|z^k\right|\left(\left|\hat{\lambda} - \lambda\right| + \left|\hat{\beta} - \beta\right|\right) + \hat{\lambda}\left|\hat{\beta}^k - \beta^k\right| + \beta^k\left|\hat{\lambda} - \lambda\right| \\ &\le \sup_{z\in D}|z| \cdot \left(\left|\hat{\lambda} - \lambda\right| + \left|\hat{\beta} - \beta\right|\right) + (\lambda+1)\left|\hat{\beta}^k - \beta^k\right| + \beta^k\left|\hat{\lambda} - \lambda\right| \qquad (*)\end{aligned}$$

for large enough n. Because $\hat{\lambda} \to \lambda$ and $\hat{\beta} \to \beta$ a.s., the last line in the inequality converges to 0 a.s. and we conclude (1).

For (2), note that $\hat{\lambda} - \lambda = O_p\left(n^{-1/2}\right)$ and $\hat{\beta} - \beta = O_p\left(n^{-1/2}\right)$, where the latter implies $\hat{\beta}^k - \beta^k = O_p\left(n^{-1/2}\right)$ by Proposition 36. Because $O_p\left(n^{-1/2}\right) + O_p\left(n^{-1/2}\right) = O_p\left(n^{-1/2}\right)$, $(*)$ implies the desired result.

For (3), use multivariate CLT as in the proof of Theorem 9; We have

$$\sqrt{n}\left\{\begin{pmatrix}\overline{I} \\ \overline{G}\end{pmatrix} - \begin{pmatrix}\lambda^{-1} \\ k\beta^{-1}\end{pmatrix}\right\} \Rightarrow N\left\{0, \begin{pmatrix}\lambda^{-2} & 0 \\ 0 & k/\beta^2\end{pmatrix}\right\},$$

and define

$$h\left(x,y\right) := r^{k+1} - \left(x^{-1} + ky^{-1}\right) r^{k} + x^{-1}\left(ky^{-1}\right)^{k}.$$

Then,

$$\left(\frac{\partial h}{\partial x}, \frac{\partial h}{\partial y}\right) = \left[x^{-2}\left\{r^{k} - \left(k/y\right)^{k}\right\}, kr^{k}y^{-2} - \left(k/y\right)^{k+1}x^{-1}\right]$$

and we have

$$\sqrt{n}\left\{h\left(\overline{I},\overline{G}\right) - h\left(\lambda^{-1}, k\beta^{-1}\right)\right\} \Rightarrow N\left\{0, \lambda^{2}\left(t^{k} - \beta^{k}\right)^{2} + k\beta^{2}\left(t^{k}/k - \lambda\beta^{k-1}\right)^{2}\right\}.$$

For (4), let $C(a_{j,}\varepsilon)$ be a circle of radius of ε at a_j (i.e., $C(a_{j,}\varepsilon) := \{z \in \mathbb{C} : |a_j - a| = \varepsilon\}$). Then, since $\hat{Q}_1\left(z\right)$ converges to $Q_1\left(z\right)$ locally uniformly a.s., by the residue theorem, if

$$\varepsilon < \min_{k \neq j} \left|a_j - a_k\right|,$$

then we have

$$\int_{C(a_j,\varepsilon)} \frac{1}{\hat{Q}_1\left(z\right)} dz \to \int_{C(a_j,\varepsilon)} \frac{1}{Q_1\left(z\right)} dz = \lim_{z \to a_j} \frac{\left(z - a_j\right)}{Q_1\left(z\right)} \neq 0 \quad \text{a.s.}$$

Recall that $\int_{C(a_j,\varepsilon)} \frac{1}{\hat{Q}_1(z)} dz = 0$ unless a root of $\hat{Q}_1\left(z\right) = 0$ is inside of $C(a_j, \varepsilon)$ since $1/\hat{Q}_j\left(z\right)$ is analytic on $\mathbb{C} \backslash \{\hat{a}_j\}_{j=1}^{k+1}$ by Cauchy's theorem. Since ε can be arbitrary, there is a sequence of roots of $\hat{Q}_1\left(z\right) = 0$ which converges to a_j a.s.

For (5), an almost same argument we used in the proof of Theorem 9 can be used here and we obtain

$$\sqrt{n}\left(\hat{a}_1 - a_1\right) = \frac{-1}{\hat{Q}_1'\left(\hat{a}_1'\right)} \sqrt{n}\left\{\hat{Q}_1\left(a_1\right) - Q_1\left(a_1\right)\right\},$$

which converges to the normal distribution of the desired result.

For (6), similarly to the proof of Theorem 19, we write

$$\left|\hat{a}_j - a\right| = \left|\frac{0 - \hat{Q}_1\left(a\right)}{\hat{Q}_1'\left(z'\right)}\right| = \left|\frac{1}{\hat{Q}_1'\left(z'\right)}\right| \left|Q_1\left(a\right) - \hat{Q}_1\left(a\right)\right|,$$

where z' is a complex number between $\hat{a}_j$ and a. Thus, $z' \to a$ a.s. and a similar argument to the proof of (1) can be used to show that $\hat{Q}_1'(z)$ converges locally uniformly a.s. to $Q_1'(z)$, which implies $\hat{Q}_1'(z') \to Q_1'(a) \neq 0$ a.s. Thus, we obtain the desired result. □

Remark 33. 1. Note that $\hat{\lambda} = n/\sum X_j$ and $\hat{\beta} = nk/\sum G_j$ are the MLE, so that $\hat{Q}_1(t)$ is the MLE too. Moreover, because a_j's only depends on k, λ, and β, we may write

$$a_j = a_j(k, \beta, \lambda).$$

Then, we have

$$\hat{a}_j = a_j\left(k, \hat{\beta}, \hat{\lambda}\right),$$

so that $\hat{a}_j$ is also the MLE of a_j, which implies $\hat{F}(t)$, $\hat{f}(t)$, $\hat{F}_q(t)$, and $\hat{f}_q(t)$ are all the MLE's of $F(t)$, $f(t)$, $F_q(t)$, and $f_q(t)$.

2. Though we assume that k is known, if $\hat{k}$ is a natural number valued estimator of k and converges to k a.s., then Theorem 32 is still valid with k being replaced by $\hat{k}$ because the a.s. convergence of $\hat{k}$ to k as a sequence of natural numbers implies $\hat{k}_n = k$ all but finite n.

Corollary 34. *Assuming the condition of Theorem 32, let $\hat{F}(t)$ be the estimators of $F(t)$ as in (1.4.4), and let $\hat{f}(t)$, $\hat{F}_q(t)$, and $\hat{f}_q(t)$ be the similarly defined estimators of $F(t)$, $f(t)$, $F_q(t)$, and $f_q(t)$, respectively. Then the following holds:*

1. *The convergence is locally uniformly a.s.*

2. *$\hat{F}(t) - F(t) = O_p\left(n^{-1/2}\right)$, and the same relation also hold for the estimators, $\hat{f}(t)$, $\hat{F}_q(t)$, and $\hat{f}_q(t)$.*

Proof. The arguments we used in the proof of Theorem 13 and Theorem 20 can be applied here. □

1.4.3 Parametric bootstrap CI and comparisons to the saddlepoint approximations

The bootstrap method of constructing CI we used in the previous subsection can be used for the parametric estimators of $F(t)$ defined in (1.4.4).

However, because G is known to follow a gamma distribution, we use parametric bootstrapping instead to compute resampled estimators $\hat{\rho}^*$, $\{\hat{a}_j^*\}$, $\hat{\beta}^*$, and $\hat{\lambda}^*$ used in (1.4.4) to determine $\hat{F}^*(t)$. The same methods (BP, BCa, and HDR) are used to construct CI but using the parametric bootstrap.

Using the same random sample we used for Figure 1.3.9 and Figure 1.3.10, we obtain three parametric bootstrapped CI's for Exp(2) and Gamma(3,3) service time as shown in Figure 1.4.1 and 1.4.2. As we did in Figure 1.3.9 and Figure 1.3.10, the average coverage probabilities and the average interval lengths are computed.

The parametric coverage probabilities in these figures are slightly better than those in Figure 1.4.1 and 1.4.2 using the nonparametric bootstrap except for gamma service time with $n = 200$.

However, the average lengths of CI of parametric estimators are very similar to those of the nonparametric saddlepoint approximation estimators. Figure 1.4.3 and 1.4.4 shows the comparisons of the average and the sample standard deviation of the lengths of the methods for all cases and one can see that the non-parametric CI lengths are very compatible to the lengths of parametric CI's in terms of the mean and standard deviation.

Figure 1.4.5 shows the estimated absolute biases and MSE of the parametric estimates and the saddlepoint approximation of $F(t)$ for G~Exp(2) (with $\lambda = 1$) and G~Gamma(3,3) (with $\lambda =.5$) on $\log_{10}$-scale from $l = 10^4$ random samples. Again, as a estimator of $F(t)$, the saddlepoint approximation is on a par with the parametric estimation considering that k is assumed to be known in the parametric estimation method.

Perhaps what is most impressive about the nonparametric bootstrap CI's is that they match the performance of their parametric counterparts well into the tail of $F(t)$.

Figure 1.4.6 shows the estimated % relative absolute bias of Figure 1.4.5. Each graphs

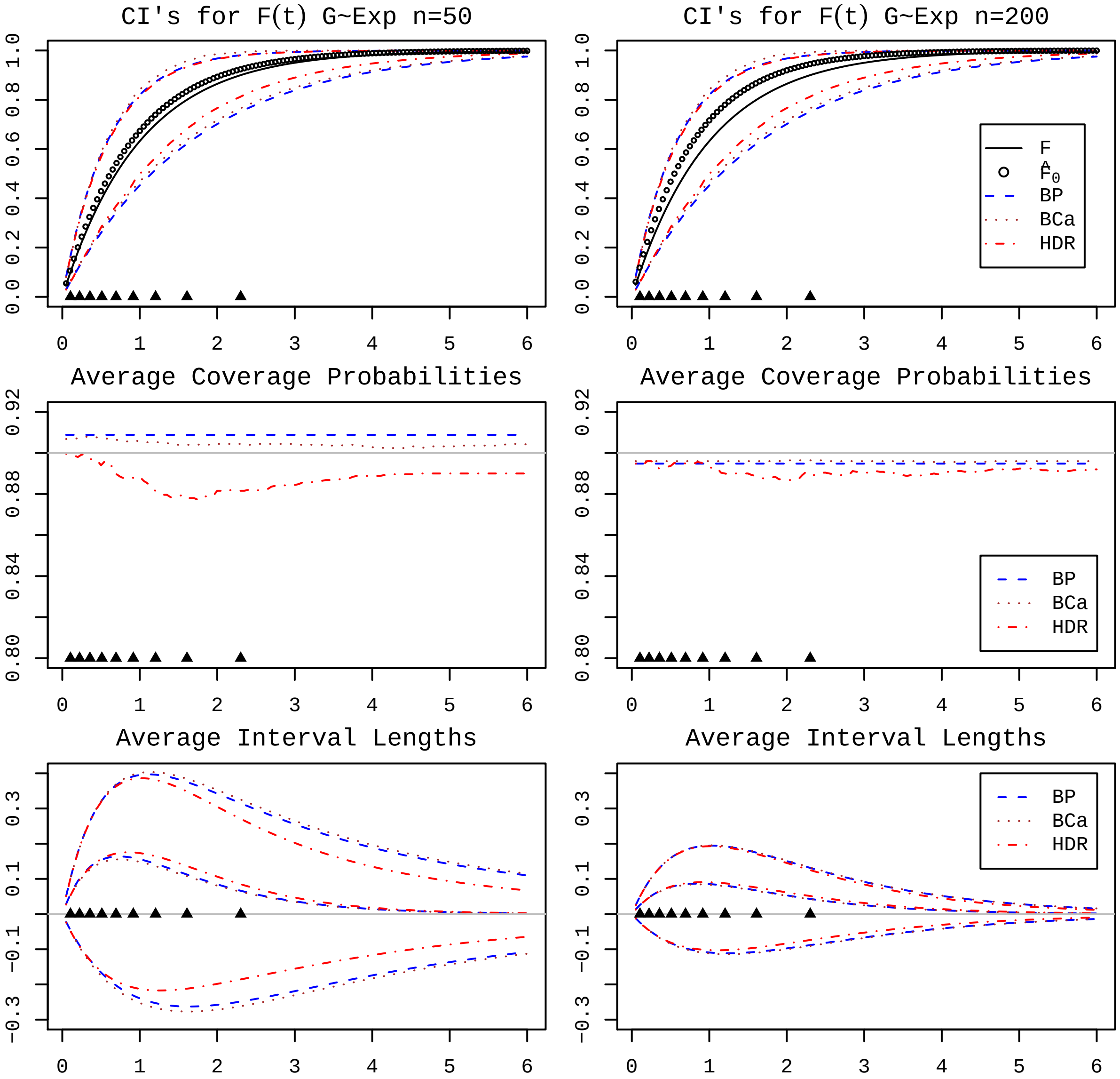

Figure 1.4.1: Parametric bootstrapped CI's for $F(t)$ where G~Exp(2) and $\lambda = 1$. Top: Calculated CI's for the same random sample of $\{G_j, I_j\}_{j=1}^n$ for n=50 (left) and n=200 (right) of the top graphs of Figure 1.3.9. Middle: the average coverage probabilities from the same l=2500 random samples of $\{G_j, I_j\}_{j=1}^n$ of Figure 1.3.9 for n=50 (left) and n=200 (right). Bottom: Average interval lengths of each CI's. The top curves are the interval lengths, $U(t) - L(t)$ and the middle curves are of $U(t) - F(t)$ and bottom curves are of $L(t) - F(t)$. In each graph, ▲ denotes a decile (from 10% to 90%) of W, the stationary waiting distribution.

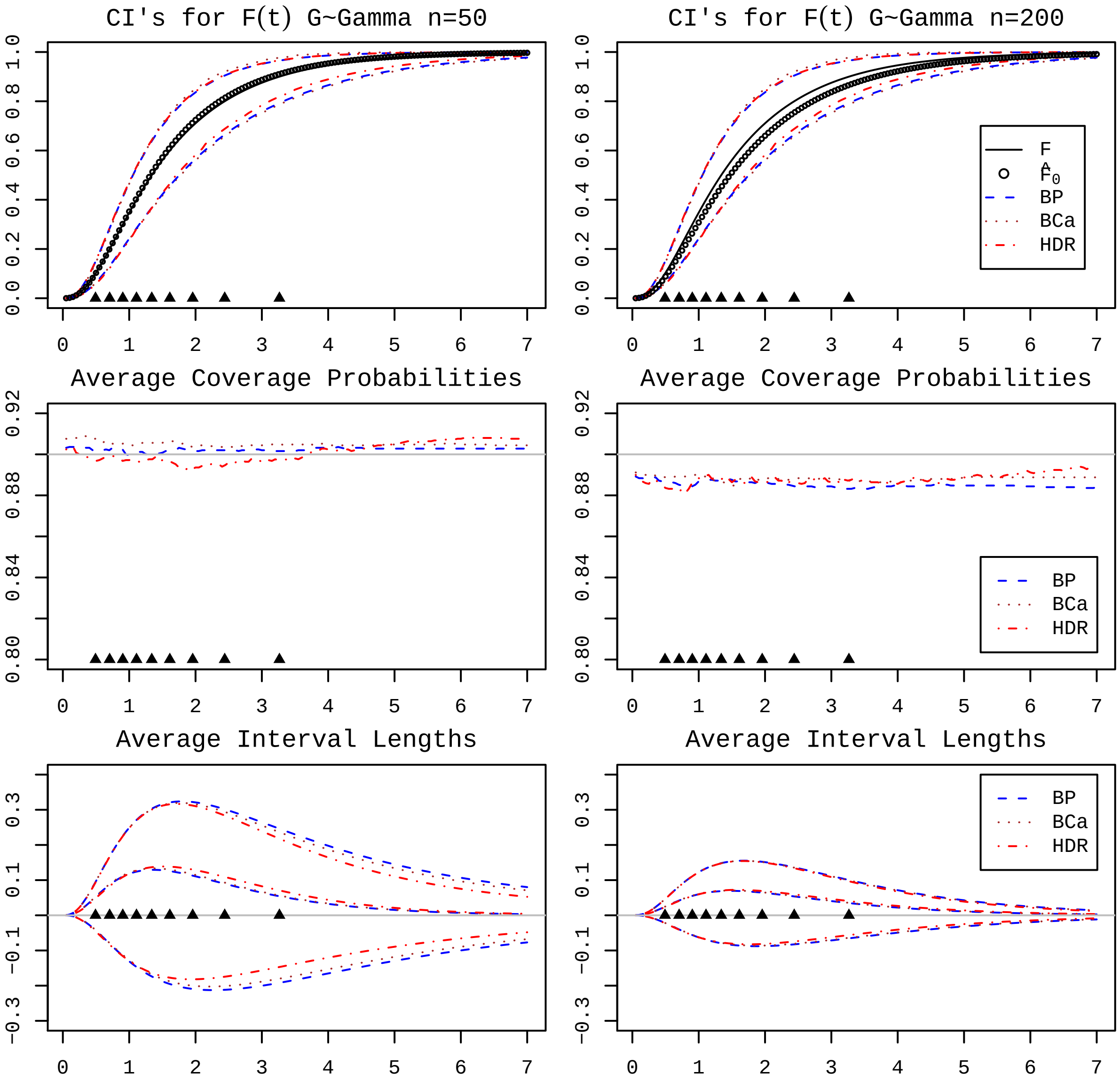

Figure 1.4.2: Same as Figure 1.4.1 for $G\sim$Gamma$(3, 3)$.

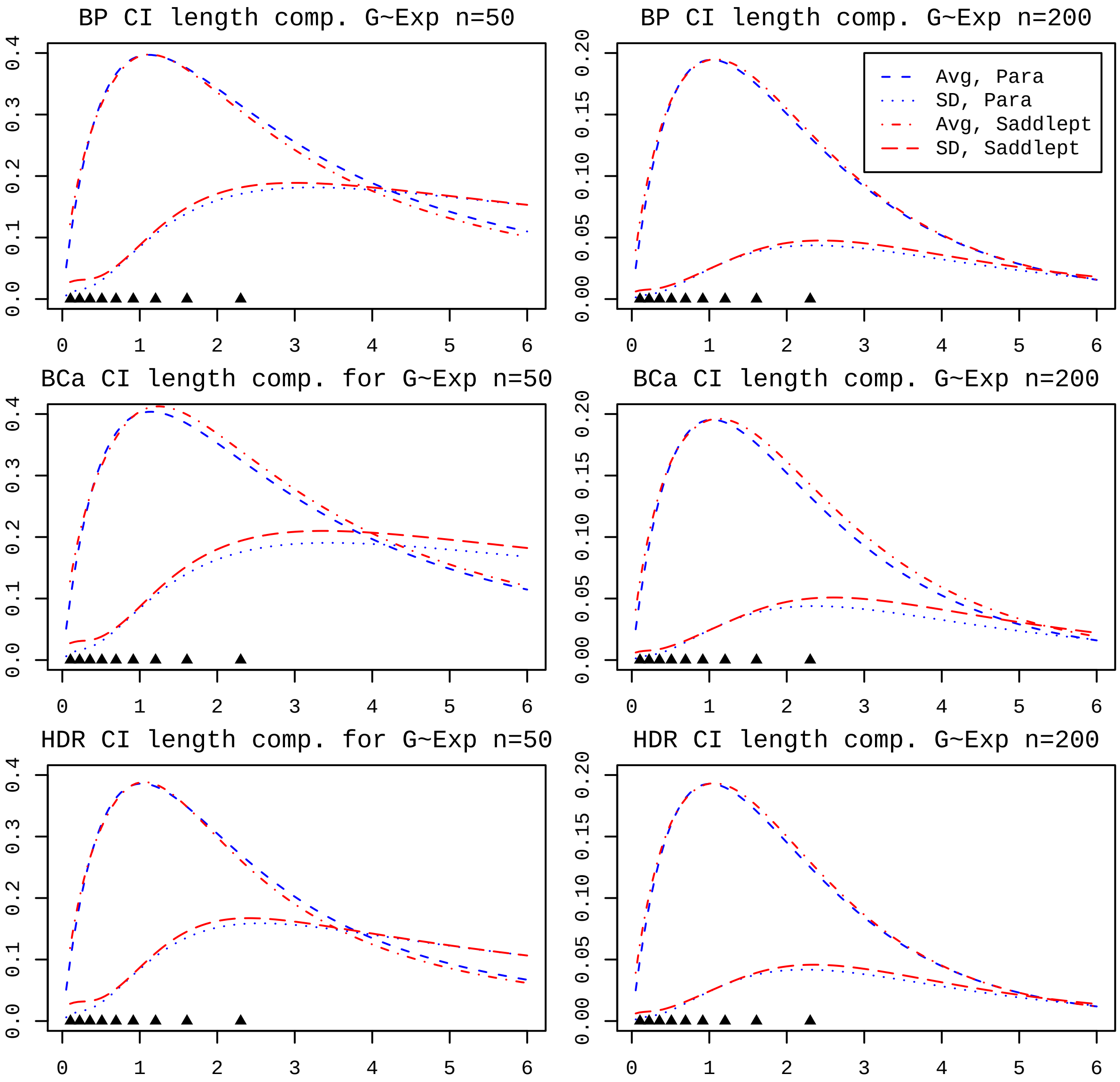

Figure 1.4.3: The mean and the standard deviation of the lengths of the parametric (Figure 1.4.1) and the non-parametric (Figure 1.3.9) bootstrapping CI's for $F(t)$ where $G{\sim}\text{Exp}(2)$ and $\lambda = 1$. Top: Bootstrapped percentile (BP) CI's. Middle: BCa CI's. Bottom: HDR CI's.

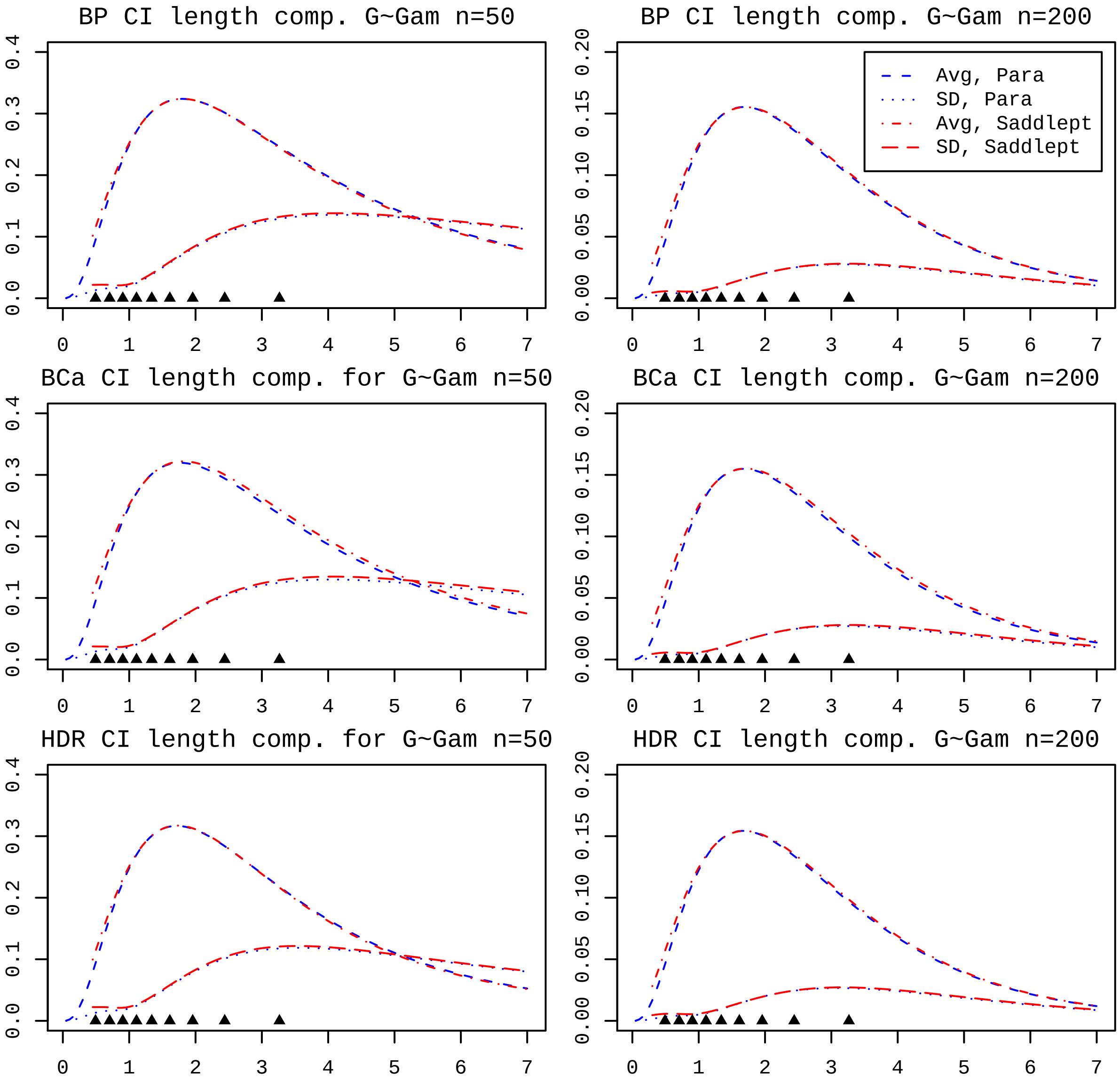

Figure 1.4.4: Same as Figure 1.4.3 for G~Gamma(3, 3).

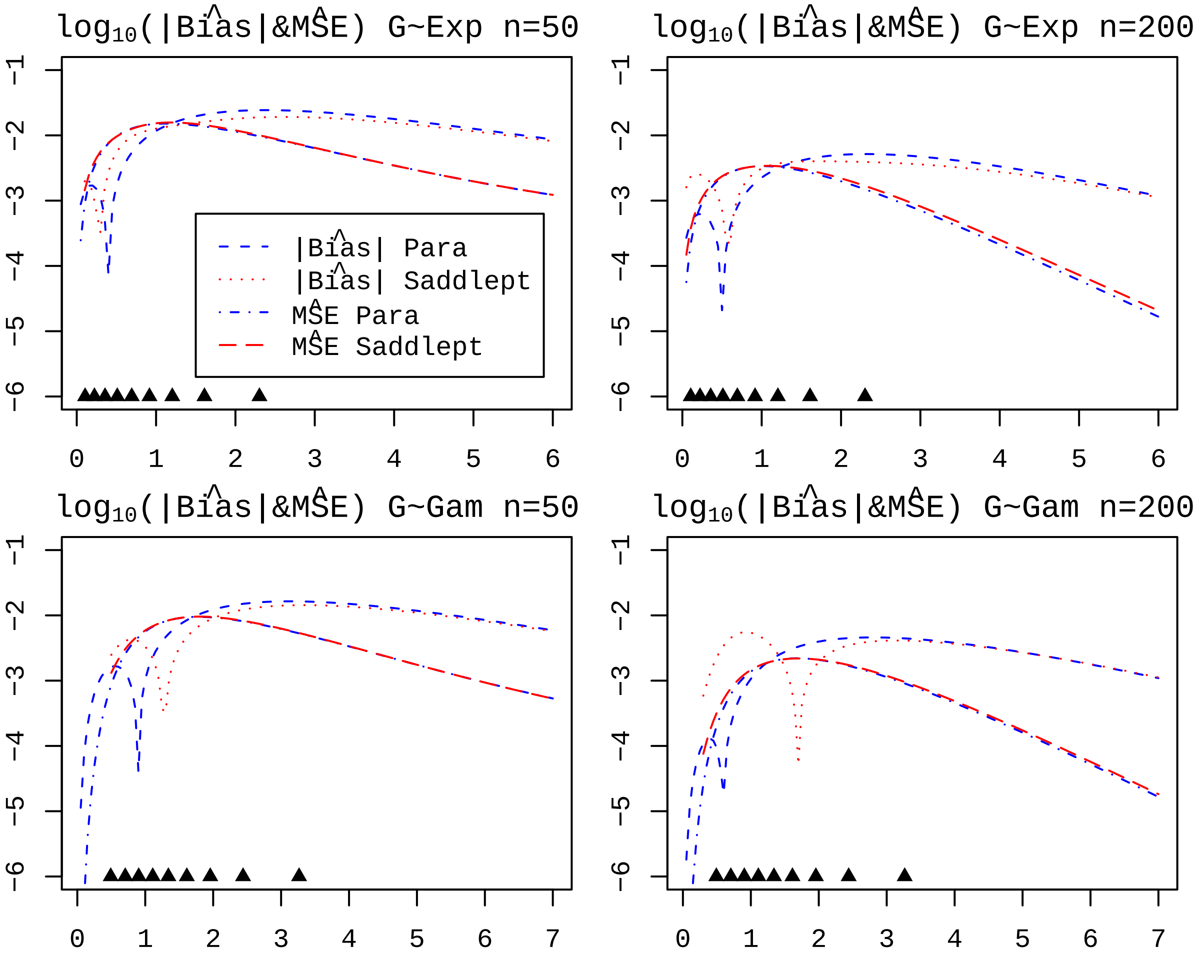

Figure 1.4.5: Estimated absolute biases and MSE's of the parametric estimations and the saddlepoint approximation of $F(t)$ for M/M/1 queue and M/E_k/1 queues in $\log_{10}$-scale from $l = 10^4$ random samples. Top: G~Exp(2) with $\lambda = 1$. Bottom: G~Gamma(3,3) with $\lambda = .5$.

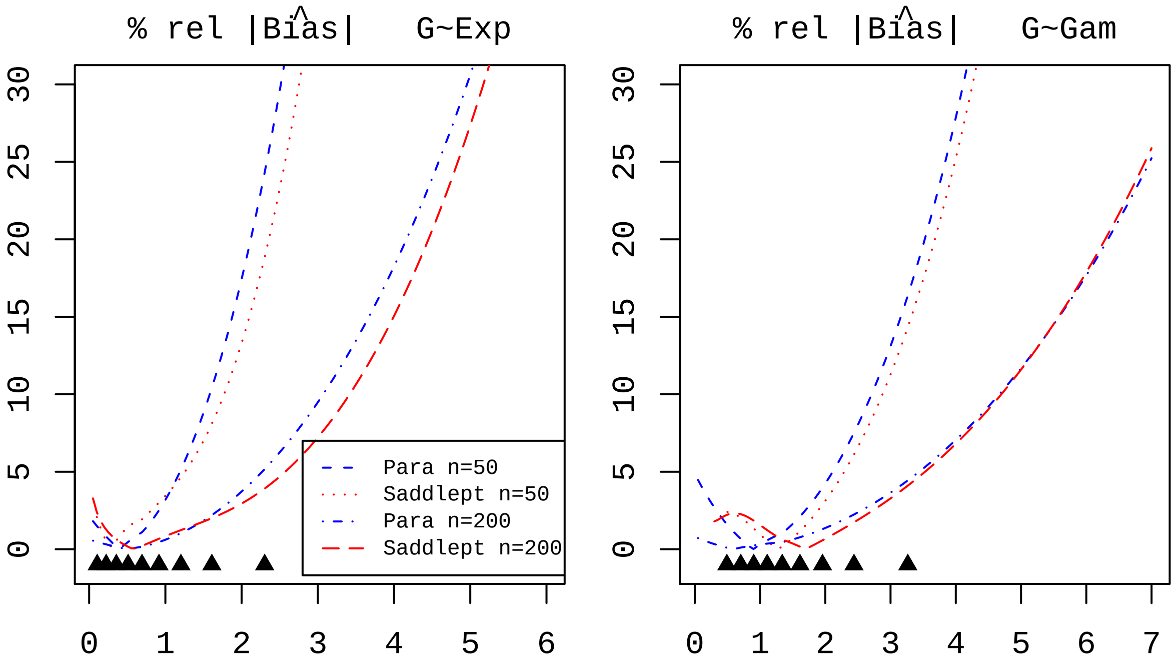

Figure 1.4.6: The estimated % relative bias from figure 1.4.5.

shows that the relative bias is lower than 5% till 90th percentile but grows exponentially for the tail area. This is not unexpected. Note that from Corollary 30, the asymptotic functions in each case have the form $1 - ae^{-cx}$ so that after 50th percentile, the relative bias is

$$\frac{100\left|\text{Bias}F\left(t\right)\right|}{\min\left\{F\left(t\right),1-F\left(t\right)\right\}} \sim 100\left|\text{Bias}F\left(t\right)\right|\left(a^{-1}e^{ct}\right).$$

However, rather surprisingly, the performance of the saddlepoint approximation is slightly better than the parametric estimation for the lower quantile areas for our cases in the view of the relative bias.

Because the estimated Biases and MSE's are very close to each other, we may suspect that both $\hat{F}_0\left(t\right)$ and the parametric estimation of $\hat{F}\left(t\right)$ are, in fact, very close to each other, which we find true. See Figure 1.4.7 for the sample mean and the sample standard deviation of the absolute difference of two estimators from Figure 1.4.5 on $\log_{10}$-scale. Both cases show that the differences become smaller as the sample sizes become bigger.

In summary, our simulation study for the $\text{M}/\text{E}_k/1$ case shows that the saddlepoint ap-

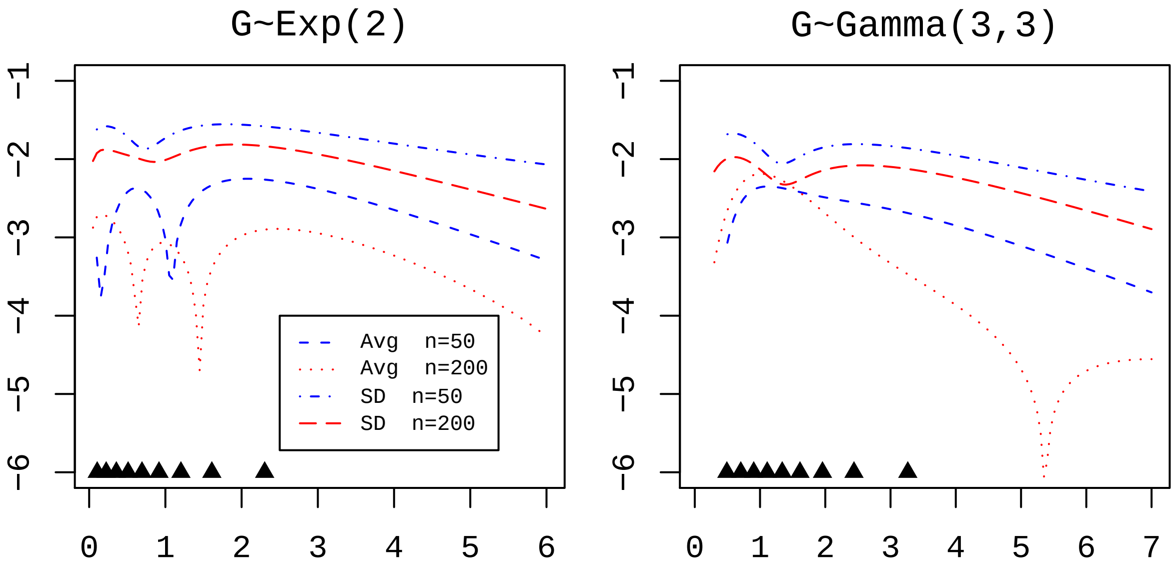

Figure 1.4.7: The sample mean (Avg) and the sample standard deviation (SD) of the absolute difference of the parametric estimators and the saddlepoint approximations from Figure 1.4.5 on $\log_{10}$-scale. The left is of G ~Exp(2) case and the right is of G ~Gamma(3,3) case.

proximation as a non-parametric estimator of $F(t)$ performs equally well with the parametric estimator of (1.4.4) and it bolsters the validity of the saddlepoint approximation with the bootstrapping as a good statistical inference method.

1.5 Conclusion

We showed that the saddlepoint approximation with the empirical MGF can be used as a reliable approximation method.

It was remarked that as a smooth approximation method, the performance of the estimation will be hindered if the estimated CDF is not smooth. However, even in that case, the saddlepoint approximation can be regarded as a good smooth approximation as we showed in estimating $\hat{F}(t)$, the distribution of $\hat{W}$.

Though a simulation approach may be possible in certain cases, it is not a time-efficient method. In our case, using R, to draw one of 8 graphs in Figure 2.4.1, took about 130-170 hours to obtain $l = 10^3$ independent simulations of $\hat{F}^{\dagger}(t)$ using the sample size $m = 10^7$

on our computer.[6] However, it took only about 68 minutes to obtain the corresponding saddlepoint approximations to draw the graph for $G\sim$Gamma(3,3) and $n = 200$, which needs to solve the saddlepoint equation 140 times for 140 grid points of $(1/20, 2/20, \cdots, 7)$ in Figure 2.4.1. This was the maximum time taken among the 8 graphs.

The saddlepoint approximation can be used as a good, general purpose, and stable approximation without a deep knowledge of probability and statistics theory. With the bootstrapping method, many statistical inferences can be made without referring to asymptotic normality results, which are hard to derive without deep insight into how the stochastic process works.

With the advance of empirical process theory, which will provide the validity of the using the empirical MGF in more general setting, we believe that the saddlepoint approximation will become an important tool to use and appealing to general practitioners of statistical inference.

1.6 Appendix

Here we collect technical results we used in the proofs of Section 2. We believe these must be known results but we could not find references during our search and decide to provide the proofs for completeness.

Proposition 35. *Let* $\mathbf{X}_n = \left(X_n^1, \ldots, X_n^m\right)$ *be* m*-dimensional random vector and* $\mathbf{a} \in \mathbb{R}^m$ *and assume* $\mathbf{X}_n - \mathbf{a} = o\left(r_n\right)$ *a.s. in the sense of*

$$\sup_{1\le i\le m} \frac{X_n^i - a^i}{r_n} \xrightarrow{a.s.} 0 \quad \text{as } n \to \infty,$$

where $r_n \to 0$ *as* $n \to \infty$*. If* $g : \mathbb{R}^m \to \mathbb{R}$ *is* C^1 *at* $\mathbf{a}$ *and* $g'\left(\mathbf{a}\right) \neq \mathbf{0}$*. Then*

$$g\left(\mathbf{X}_n\right) - g\left(\mathbf{a}\right) = o\left(r_n\right) \quad a.s.$$

[6] The computer is equipped with AMD Athlon® 64 X2 4000+. Because the R we used is a single thread program, only one core was used for the calculations and the simulations.

Proof. By the multivariate mean value theorem,

$$g\left(\mathbf{X}_n\right)-g\left(\mathbf{a}\right)=g'\left(\mathbf{Y}_n\right)\left(\mathbf{X}_n-\mathbf{a}\right)=\sum_{i=1}^{m}\frac{\partial g}{\partial x^i}\left(Y_n^i\right)\left(X_n^i-a^i\right),$$

where $\mathbf{Y}_n$ is m-dimensional random vector such that Y_n^i's are all in between X_n^i and a^i. Since $X_n^i-a^i=o\left(r_n\right)$ a.s., $X_n^i\xrightarrow{\text{a.s.}}a^i$, which implies $Y_n^i\xrightarrow{\text{a.s.}}a^i$, or $\partial g\left(Y_n^i\right)/\partial x^i\xrightarrow{\text{a.s.}}\partial g\left(a^i\right)/\partial x^i$. Then, we have

$$\lim_{n\to\infty}\frac{g\left(\mathbf{X}_n\right)-g\left(\mathbf{a}\right)}{r_n}=\sum_{i=1}^{m}\lim_{n\to\infty}\frac{\partial g}{\partial x^i}\left(Y_n^i\right)\frac{X_n^i-a^i}{r_n}=0,$$

which is the desired result. □

Proposition 36. *Let $\mathbf{X}_n=\left(X_n^1,\ldots,X_n^m\right)$ be m-dimensional random vector and $\mathbf{a}\in\mathbb{R}^m$ and assume and $\mathbf{X}_n-\mathbf{a}=O_p\left(r_n\right)$ in the sense of*

$$X_n^i-a^i=O_p\left(r_n\right)\quad \text{as } n\to\infty \text{ for any } 1\leq i\leq m,$$

where $r_n\to 0$ as $n\to\infty$. If $g:\mathbb{R}^m\to\mathbb{R}$ is C^1 at a and $g'\left(\mathbf{a}\right)\neq\mathbf{0}$. Then

$$g\left(\mathbf{X}_n\right)-g\left(\mathbf{a}\right)=O_p\left(r_n\right).$$

Proof. We show that $X_n^i\xrightarrow{p}a^i$ for each i as $n\to\infty$ by contradiction. Suppose X_n^i does not converges to a^i in probability, which implies that there exists $\varepsilon>0$ such that

$$\limsup_n P\left(\left|X_n^i-a^i\right|>\varepsilon\right)\to c>0. \tag{$*$}$$

Since $X_n^i-a^i=O_p\left(r_n\right)$, we can pick $M\in\mathbb{N}$ such that

$$P\left(\frac{\left|X_n^i-a^i\right|}{r_n}>M\right)<c/2\quad\text{for any } n\geq n_0.$$

Since $r_n\to 0$, we can pick n_1 such that $r_nM\leq\varepsilon$ for any $n\geq n_1$. Thus, for any $n\geq$

$\max\{n_0, n_1\}$, we have

$$P\left(\left|X_n^i - a^i\right| > \varepsilon\right) \leq P\left(\left|X_n^i - a^i\right| > r_n M\right) < c/2,$$

which contradicts to $(*)$. Thus, we conclude $X_n^i \xrightarrow{p} a^i$.

Now, by the multivariate mean value theorem,

$$g\left(\mathbf{X}_n\right) - g\left(\mathbf{a}\right) = g'\left(\mathbf{Y}_n\right)\left(\mathbf{X}_n - \mathbf{a}\right) = \sum_{i=1}^{m} \frac{\partial g}{\partial x^i}\left(Y_n^i\right)\left(X_n^i - a^i\right),$$

where $\mathbf{Y}_n$ is m-dimensional random vector such that Y_n^i's are all in between X_n^i and a^i. Thus, $Y_n^i \xrightarrow{p} a^i$, which implies $\partial g\left(Y_n^i\right)/\partial x^i \xrightarrow{p} \partial g\left(a^i\right)/\partial x^i$, or

$$\frac{\partial g}{\partial x^i}\left(Y_n^i\right) = \frac{\partial g}{\partial x^i}\left(a^i\right) + o_p(1).$$

Thus, we have

$$\begin{aligned}\sum_{i=1}^{m} \frac{\partial g}{\partial x^i}\left(Y_n^i\right)\left(X_n^i - a^i\right) &= \sum_{i=1}^{m} \frac{\partial g}{\partial x^i}\left(a^i\right)\left(X_n^i - a^i\right) + \sum_{i=1}^{m} o_p(1)\left(X_n^i - a^i\right) \\ &= \sum_{i=1}^{m} \frac{\partial g}{\partial x^i}\left(a^i\right)\left(X_n^i - a^i\right) + o_p\left(r_n\right),\end{aligned}$$

[7]which implies the desired result because $O_p(r_n) + O_p(r_n) = O_p(r_n)$, $O_p(r_n)\,o_p(1) = O_p(r_n)$, and $O_p(r_n) + o_p(r_n) = O_p(r_n)$.

The following result can be used for the theorem known as Pólya's lemma (Exercise 7.2 of [61]. See also Exercise 7.13 of [58]), which is that F_n uniformly converges to F if $F_n \Rightarrow F$ and F is continuous. □

Theorem 37. *Let $f_n : \mathbb{R} \to \mathbb{R}$, $n \in \mathbb{N}$ be a sequence of non-decreasing functions. If f is continuous and $f_n(x) \to f(x)$ pointwise for a countable dense set in $\mathbb{R}$, then*

[7] After completing the proof, I found Proposition 6.1.5 of Time Series: Theory and Methods, which has a more generalized result. Note that $g'(x)$ does not need to be continuous as in the Delta method (Theorem 1.8.12 of [42]). However, a function whose derivative is not continuous is barely occurred in statistics so that the applicability of the proposition will not be weakened generally.

1. *$f_n(x) \to f(x)$ for any $x \in \mathbb{R}$,*

2. *f is non-decreasing, and*

3. *$f_n \to f$ locally uniformly.*

Proof. 1. Let S be the countable dense set on which f_n converges to f pointwise. Let x be any real number. For any $\varepsilon > 0$, there $\delta > 0$ such that $|x - y| < \delta$ implies $|f(x) - f(y)| < \varepsilon$ due to the continuity of f. Since S is dense in $\mathbb{R}$, there exists $y_1, y_2 \in S$ such that $|y_i - x| < \delta$ for $i = 1, 2$ and $y_1 \le x \le y_2$. Because f_n is non-decreasing, we have

$$f_n(y_1) \le f_n(x) \le f_n(y_2),$$

which implies

$$\limsup_{n\to\infty} f_n(x) \le \limsup_{n\to\infty} f_n(y_2) = \lim_{n\to\infty} f_n(y_2) = f(y_2) < f(x) + \varepsilon$$

and similarly,

$$f(x) - \varepsilon < \liminf_{n\to\infty} f_n(x).$$

Because ε is arbitrary, $\lim_{n\to\infty} f_n(x) = f(x)$.

2. We now show f is non-decreasing. Suppose there exists $x, y \in \mathbb{R}$ such that $x < y$ but $f(x) > f(y)$. Let $\varepsilon = \{f(x) - f(y)\}/2$ and pick $n \in \mathbb{N}$ such that $|f_n(x) - f(x)| < \varepsilon$ and $|f_n(y) - f(y)| < \varepsilon$. Then by our choice of ε, $f_n(y) < f(y) + \varepsilon < f(x) - \varepsilon < f_n(x)$, which contradicts to the fact f_n is non-decreasing.

3. For the locally uniform convergence of f_n, let $[a, b]$ be given. It suffice to show f_n converges to f continuously on $[a, b]$ (see theorem 7.3.5 of [66] or §0.0 of [53]), which can be shown similarly to the proof (1).

□

Chapter 2

Saddlepoint Approximation to Pollaczek-Khinchin Formula of W_q

In the previous chapter, we showed that the saddlepoint approximation with (non-parametric) bootstrapping is an efficient tool for statistical inference. In this chapter, we compare the saddlepoint approximation with other known methods of estimating the CDF of W_q, the waiting time in queue. We show that the saddlepoint CDF approximation $\hat{F}_{q0}(t)$ is a better approximation to $F_q(t)$ and $\hat{F}_q(t)$, the CDF's of W_q and $\hat{W}_q$, respectively, than the Cramér-Lundberg approximation and other known asymptotic approximations.

2.1 Waiting time in queue and its CDF estimation

As mentioned in the previous chapter, the smoothness of saddlepoint approximation hinders the performance of approximations for the lower quantile area where the effect of the discrete part, $\hat{G}_j$ is dominant. Let W_q be the stationary waiting time in the queue and $F_q(t)$ the CDF of W_q.

In insurance mathematics, it is known that there is a connection between ruin probability and the stationary waiting time of M/G/1 queue. Following [68], let $S(\tau)$ be a compound

Poisson process with rate λ for the total claims in the time interval $[0,\tau]$ defined by

$$S(\tau) = \sum_{j=1}^{N(\tau)} C_j,$$

where C_j represents successive claims with $C_j \overset{\text{iid}}{\sim} F_C(t)$ and $EC_j = \mu^{-1}$. We assume that the insurance company's reserve increases at constant rate $\sigma > 0$ and $\sigma > \lambda/\mu$.

If the company's initial reserve is $t > 0$, then the surplus of the company at time τ is $t + \sigma\tau - S(\tau)$ and the probability that the the company will be ruined eventually is

$$\psi(t) = P\{S(\tau) > t + \sigma\tau \text{ for some } \tau \geq 0\}.$$

It is known that if we set $G_j = C_j$, then

$$1 - F_q(t) = \psi(\sigma t).$$

In this chapter, we compare the saddlepoint approximation to other known approximations. We note that §4 of Chapter 1 in [8] gives a birds-eye view on the related result and method of finding the CDF. For a review and a comparison for the performance of several known numerical inversion methods of $\hat{\mathcal{W}}_q(s)$, see [63].

2.1.1 Saddlepoint approximation

As we did for $\mathcal{W}(s)$, because

$$\mathcal{W}_q(s) = \frac{(1-\rho)s}{s + \lambda - \lambda\mathcal{G}(s)},$$

we have

$$\hat{\mathcal{W}}_q(s) = \frac{(1-\hat{\rho})s}{s + \hat{\lambda} - \hat{\lambda}\hat{\mathcal{G}}(s)},$$

where $\hat{\rho}$, $\hat{\lambda}$, and $\hat{\mathcal{G}}(s)$ are defined as in $\hat{\mathcal{W}}$ with the sample size n suppressed.

The saddlepoint approximation can be calculated as usual and are denoted as $\hat{F}_{q0}(t)$ and

$\hat{f}_{q0}(t)$. See Figure 2.1.1 and 2.1.2 for these saddlepoint CDF and PDF approximations of $\hat{W}_q$, which accurately approximate the tail areas of the empirical CDFs, and the histograms of 10^7 $\hat{W}_q$ values. These approximations are not so accurate around 0 since there is a point mass (atom) at 0 as

$$P(W_q = 0) = P(N = 0) = (1 - \rho).$$

As a smooth function, the saddlepoint approximation tries to connect the point mass smoothly, which is evident by comparing the histograms and the PDF estimations.

A better way to approach is to get rid of the point mass at 0. Following [6], let $F_q^+(t)$ be defined by

$$F_q^+(t) = \frac{F_q(t) - (1 - \rho)}{\rho} \tag{2.1.1}$$

(i.e, $F_q(t) = (1 - \rho) + \rho F_q^+(t)$). Then the MGF of $F_q^+(t)$, $\mathcal{W}_q^+(s)$ is

$$\mathcal{W}_q^+(s) = \frac{\mathcal{W}(s) - (1 - \rho)}{\rho} = \frac{\mu(1 - \rho)\{\mathcal{G}(s) - 1\}}{s + \lambda - \lambda\mathcal{G}(s)}.$$

Denote the plug-in estimator as $\hat{\mathcal{W}}^+(s)$. Since

$$\lim_{s \to -\infty} \left\{\log \mathcal{W}_q^+(s)\right\}' = \lim_{s \to -\infty} \{\log \hat{\mathcal{W}}_q^+(s)\}' = 0,$$

the saddlepoint equation can be solved for any $t > 0$ which is unlike the case of $\mathcal{W}(s)$ (see Corollary 1).

Let $\hat{F}_{q0}^+(t)$ be the saddlepoint CDF approximation and $\hat{f}_{q0}^+(t)$ be the saddlepoint PDF approximation using $\hat{\mathcal{W}}^+(s)$. Then, $(1 - \hat{\rho}) + \hat{\rho}\hat{F}_{q0}^+(t)$ and $\hat{\rho}\hat{f}_{q0}^+(t)$ are the the better approximations of $\hat{F}_q(t)$ and $\hat{f}_{q0}(t)\mathbf{1}_{(0,\infty)}(t)$ respectively, which can be seen in Figure 2.1.1 and 2.1.2.

2.1.2 Efficient simulation method for W_q

We observed that $(1 - \rho) + \rho F_{q0}^+$ and $(1 - \hat{\rho}) + \hat{\rho}\hat{F}_{q0}^+$ can be used in lieu of a simulation approach based on sampling values of W_q and $\hat{W}_q$. If $m = 10^7$ values of W_q and $\hat{W}_q$ are

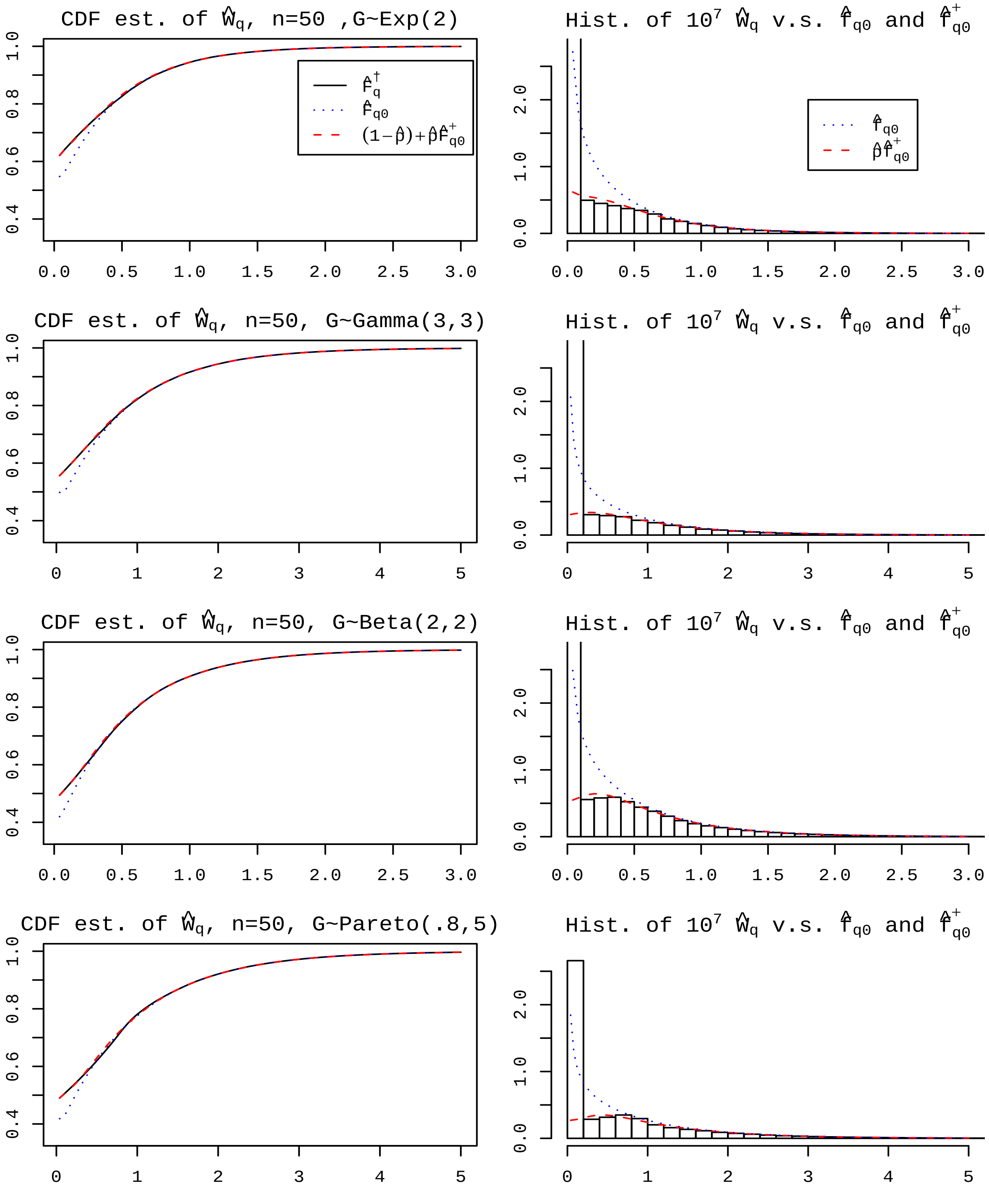

Figure 2.1.1: Left: The empirical CDFs, $\hat{F}_q^{\dagger}(t)$, of 10^7 $\hat{W}_q$ values as in (1.3.3) with the saddlepoint CDF approximations $\hat{F}_{q0}(t)$ from $\hat{\mathcal{W}}_q(s)$ and $(1-\hat{\rho})+\hat{\rho}\hat{F}_{q0}^{\dagger}(t)$ in (2.1.1). From the top, the $G_j \overset{\text{iid}}{\sim} \text{Exp}(2)$, Gamma(3,3), Beta(2,2), and Pareto(4/5,5). Right: Histogram of 10^7 $\hat{W}_q$ values as compares with the saddlepoint PDF approximation from $\hat{\mathcal{W}}_q(s)$ and $\hat{\rho}\hat{f}_{q0}^{+}(t)$. The t-axes are cut to include at least 99.5% of W_q^*.

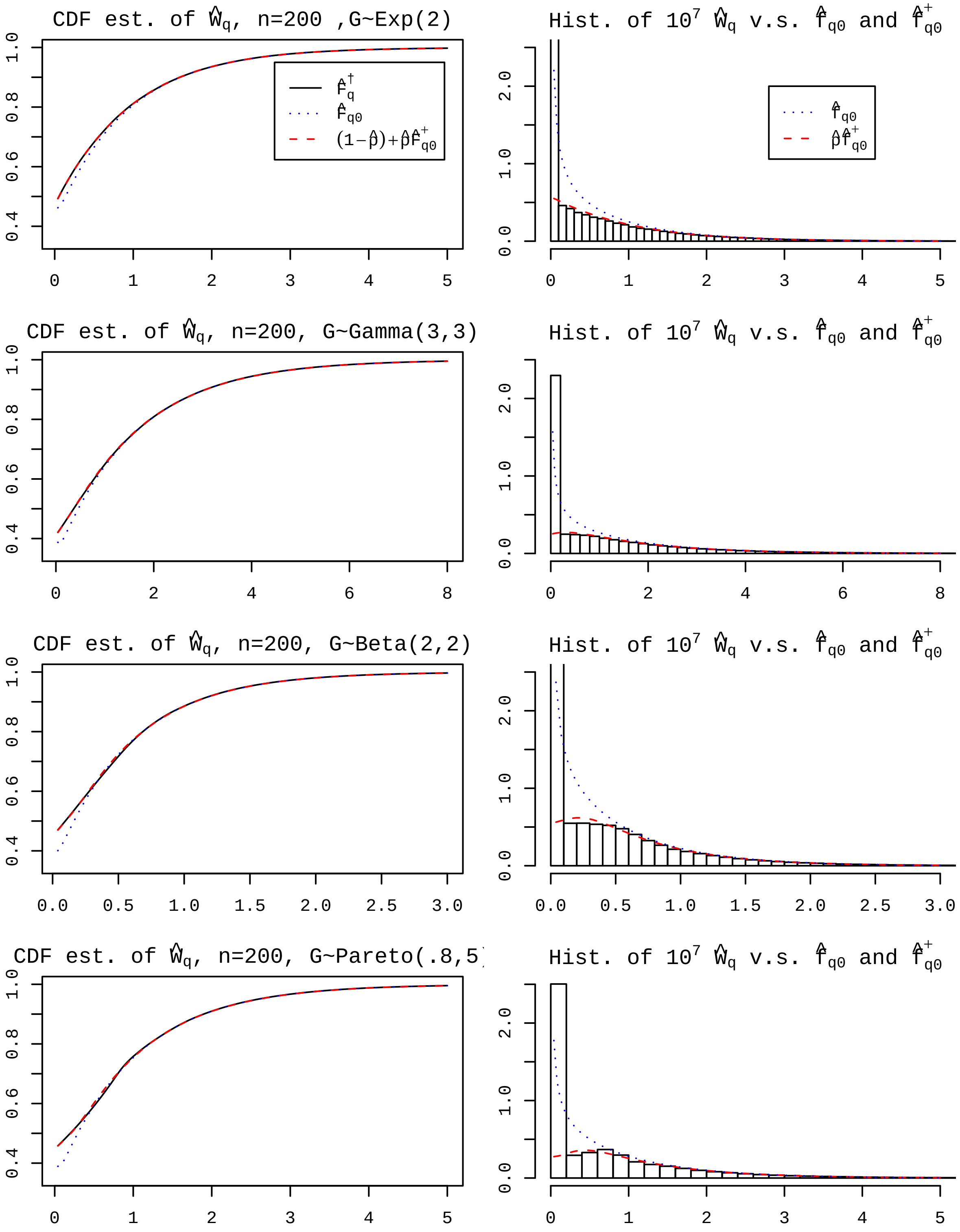

Figure 2.1.2: Same as Figure 2.1.1 but with the sample size $n = 200$.

simulated to obtain the random sample following (1.3.2), about $m(1-\rho)$ of the simulated values will be 0. Thus, for the estimator of F_q or $\hat{F}_q$, it is more efficient to use

$$(1-\rho)+\rho F_q^{+\dagger} \quad \text{or} \;\; (1-\hat{\rho})+\hat{\rho}\hat{F}_q^{+\dagger},$$

where $F_q^{+\dagger}$ is the empirical CDF of

$$W_q^+ = V_1 + \cdots V_{N^+}$$

and the PMF of N^+ is defined by

$$P\left(N^+ = k\right) = \rho^{k-1}(1-\rho), \qquad k = 1, 2, \cdots,$$

or

$$N^+ \sim N+1.$$

In this way, all of the simulated values are non-zero and can be used to estimate $F_q^{+\dagger}$ or $\hat{F}_q^{+\dagger}$. Note that in this dissertation, the method of simulating W_q^+ was not used to make our simulation method consistent through different chapters.

2.1.3 Cramér-Lundberg approximation

In risk theory or ruin probability literature, there is a well-known approximation for $\psi(t)$, the Cramér-Lundberg approximation (see [30], [68], [55], or [8]): Let

$$\mathcal{G}_e(s) = \mu \int_0^\infty e^{st}\{1-G(t)\}\,dt$$

be the MGF of V (this is usually denoted by G_e), the equilibrium distribution of G. Suppose $\mathcal{G}_e(s)$ is finite for some $s>0$ and if

$$\lim_{s \to s'} \mathcal{G}_e(s) = \infty,$$

where

$$s' := \sup\left\{s : \mathcal{G}_e(s) < \infty\right\},$$

then with the assumption of $\sigma = 1$,

$$\psi(x) = 1 - F_q(t) \sim \gamma e^{-\delta t}$$

where δ is the unique solution of equation

$$\mathcal{G}_e(\delta) = \rho^{-1} \tag{2.1.2}$$

on $(0, s')$ and

$$\gamma = \frac{1-\rho}{\rho\delta\mu\int_0^\infty te^{\delta x}\left\{1 - G(t)\right\}dt}.$$

Here "$\sim$" means that $\lim_{x\to\infty}\{1 - F(t)\}/\gamma e^{-\delta t} = 1$.

Note that the above equations are normally how δ and γ are presented in the ruin probability literature. The following proposition is useful to calculate δ and γ, and we later found that a similar representation appeared in [8]. It is well known that

$$\mathcal{G}_e(s) = \frac{\mu\left\{1 - \mathcal{G}(s)\right\}}{-s}. \tag{2.1.3}$$

Note that $s = 0$ is a removable singularity, so that $\mathcal{G}(s) < \infty$ if and only if $\mathcal{G}_e(s) < \infty$. Thus, the convergence strip of $\mathcal{G}_e(s)$ is the same of $\mathcal{G}(s)$ and $s' = b$.

Proposition 38. *Suppose the condition of Cramér-Lundberg approximation is satisfied. Then, the root δ of (2.1.2) is c of Lemma 2. Moreover, if there is an integrable function $l(t)$ such that*

$$te^{s_1 t}\left\{1 - G(t)\right\} \le l(t)$$

for some c satisfying $c < s_1 < b$ then

$$\mu\int_0^\infty te^{ct}\{1 - G(t)\}dt = \mathcal{G}_e'(c), \tag{2.1.4}$$

and

$$\gamma = \frac{1-\rho}{\lambda \mathcal{G}'(c) - 1}. \tag{2.1.5}$$

Proof. Equation (2.1.2) can be written as

$$\mathcal{G}_e(\delta) = \frac{\mu\{1-\mathcal{G}(\delta)\}}{-\delta} = \rho^{-1} = \frac{\mu}{\lambda},$$

which implies

$$\lambda - \lambda\mathcal{G}(\delta) + \delta = 0, \tag{$*$}$$

Therefore, δ must be c because the root is unique by Lemma 2.

For the second part, note that the function

$$\varphi(s,t) := e^{st}\{1-G(t)\}$$

is integrable on $(0, b-\varepsilon)$ for some $\varepsilon > 0$ by the assumption and differentiable with respect to s. Because $d\varphi(s,t)/ds$ is dominated by the integrable function $l(t)$, we conclude (2.1.4) (see p.154 of [38]).

Thus, we have

$$\begin{aligned}\mu\int_0^\infty te^{ct}\{1-G(t)\}\,dt = \mathcal{G}_e'(c) &= \frac{d}{ds}\left[\frac{\mu\{1-\mathcal{G}(s)\}}{-s}\right]\Big|_{s=c} \\ &= \mu\left\{\frac{\mathcal{G}'(s)\,s + 1 - \mathcal{G}(s)}{s^2}\right\}\Big|_{s=c} = \frac{\mu\{\mathcal{G}'(c)\,c + 1 - \mathcal{G}(c)\}}{c^2},\end{aligned}$$

and

$$\begin{aligned}\gamma &= \frac{1-\rho}{\rho\delta\mu\int_0^\infty te^{\delta t}\{1-G(t)\}dt} = \frac{1-\rho}{\rho c\left[\frac{\mu\{\mathcal{G}'(c)c+1-\mathcal{G}(c)\}}{c^2}\right]} \\ &= \frac{(1-\rho)\,c}{\lambda\{1-\mathcal{G}(c)+c\mathcal{G}'(c)\}} = \frac{(1-\rho)\,c}{\lambda\{1-\mathcal{G}(c)\}+\lambda c\mathcal{G}'(c)} \\ &\overset{(*)}{=} \frac{(1-\rho)\,c}{-c+\lambda c\mathcal{G}'(c)} = \frac{1-\rho}{\lambda\mathcal{G}'(c)-1},\end{aligned}$$

which is the desired result. □

Note that the assumption of the Cramér-Lundberg approximation and proposition 38 are satisfied for the empirical MGF of $\hat{\mathcal{G}}(s)$. For more detailed information and generalizations of this approximation, see [1], [2], [8], [55], and [72].

If we substitute ρ, λ, c, $\mathcal{G}(s)$ with estimators, $\hat{c}$, $\hat{\lambda}$, $\hat{p}$, $\hat{\mathcal{G}}(s)$ as we did for the saddlepoint approximations $\hat{\mathcal{W}}(s)$, the Cramér-Lundberg approximation can be used to estimate $F_q(t)$ and $\hat{F}_q(t)$. For example,

$$\hat{\gamma} := \frac{1-\hat{\rho}}{\hat{\lambda}\hat{\mathcal{G}}'(\hat{c})-1}. \tag{2.1.6}$$

We note that our estimator $\hat{\gamma}$ of γ is different from the estimator in [32], in which λ is assumed to be known and set to 1. Their estimate is

$$\hat{\gamma}_{GJ} := \frac{\hat{\mathcal{G}}''(0)/2 + \hat{R}_1/\hat{c}}{\hat{\mathcal{G}}''(0)/2 + \hat{\mathcal{G}}'''(0)\,\hat{c}/2 + \hat{R}_2}, \tag{2.1.7}$$

where

$$\hat{R}_1 := \frac{1}{n}\sum_{j=1}^{n}\{e^{\hat{c}G_j} - 1 - \hat{c}G_j - (\hat{c}G_j)^2/2\},$$
$$\hat{R}_2 := -\hat{R}_1 + \frac{\hat{c}}{n}\sum_{j=1}^{n} G_j\{e^{\hat{c}G_j} - 1 - \hat{c}G_j - (\hat{c}G_j)^2/2\}.$$

Our nonparametric estimation of the Cramér-Lundberg approximation can be regarded as the Cramér-Lundberg approximation of $\hat{F}_q(t)$, which is not the case for $1 - \hat{\gamma}_{GJ}e^{-\hat{c}t}$. Moreover, for $\lambda = 1$, we have

$$\gamma = \frac{1-EG}{\mathcal{G}'(p)-1} \Longrightarrow \hat{\gamma} = \frac{1-\overline{G}}{n^{-1}\sum G_j e^{G_j\hat{c}} - 1},$$

which is much simpler than $\hat{\gamma}_{GJ}$ and intuitively, $\hat{\gamma}$ will have a smaller MSE.

Figure 2.1.3 shows density estimates of the sampling distributions for $(\hat{\gamma}-\gamma)$ and $(\hat{\gamma}_{GJ}-\gamma)$ by simulating 10^4 values with $\lambda = 1$. In other words, from the sample service time $\{G_j\}_{j=1}^{n}$ only (i.e., $\hat{\lambda}$ is fixed as 1), values for $\hat{\gamma}$ and $\hat{\gamma}_{GJ}$ are calculated 10^4 times to give the density estimates shown. They clearly show that the mean squared error (MSE) of $\hat{\gamma}$ is smaller and from now on, we only consider $\hat{\gamma}$ when we deal with the Cramér-Lundberg approximation.

In [33], a approximate CI for γe^{-ct} was given as

$$\left\{\hat{\gamma}\exp\left(-\hat{c}-\frac{z_{\alpha/2}\hat{\sigma}_c}{\sqrt{n}}\right)t, \hat{\gamma}\exp\left(-\hat{c}+\frac{z_{\alpha/2}\hat{\sigma}_c}{\sqrt{n}}\right)t\right\}$$

by observing

$$\frac{\hat{\gamma}\exp\left(-\hat{c}t\right)}{\gamma\exp\left(-ct\right)}=\frac{\hat{\gamma}}{\gamma}\exp\left(c-\hat{c}\right)\sqrt{n}\frac{t}{\sqrt{n}}\Rightarrow e^{-N\left(0,\sigma_c^2\right)}$$

if $t/\sqrt{n}$ converges to a finite number.

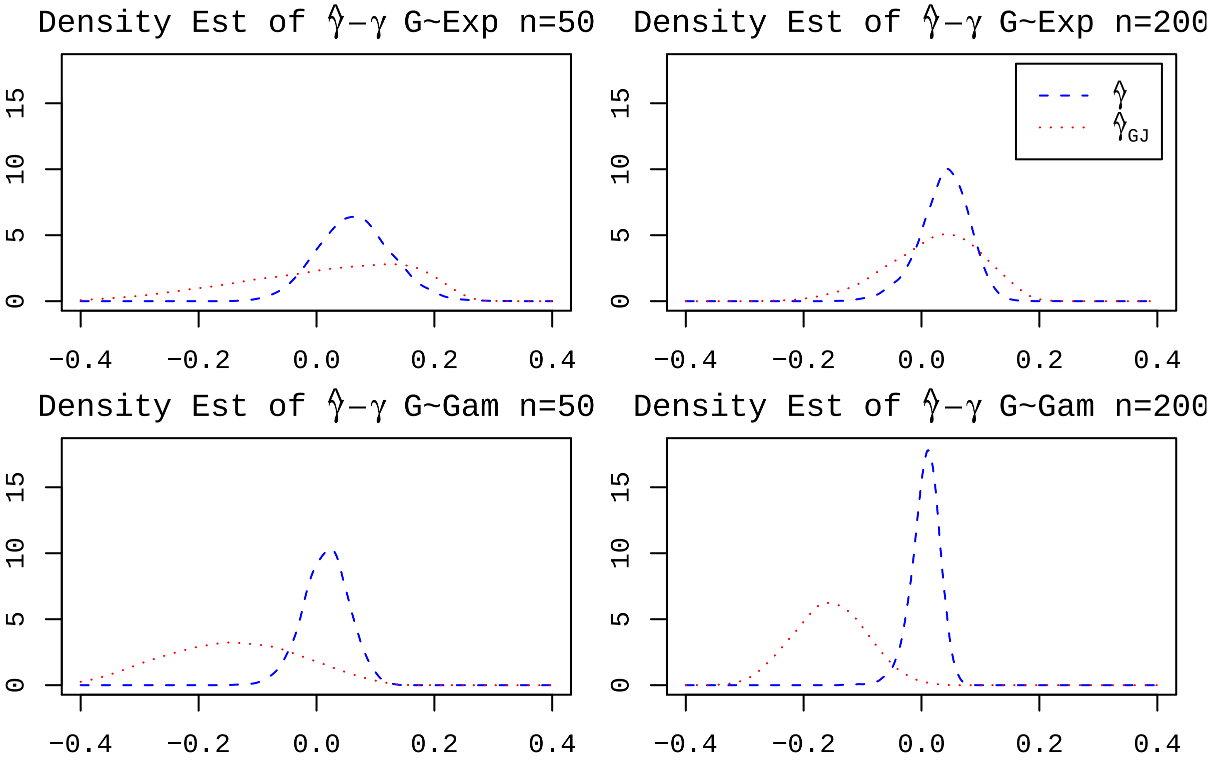

Figure 2.1.3: The density estimation of the sampling distribution for $(\hat{\gamma}-\gamma)$ and $(\hat{\gamma}_{GJ}-\gamma)$ by simulating 10^4 values with $\lambda=1$ from the same service time sample of $\{G_j\}_{j=1}^n$. Top: G~Exp(2). The left graph is for $n=50$ and the right graph is for $n=200$. Bottom: G~Gamma(3,6).

2.1.4 Tijms approximation

Tijms ([68]) suggested adding another exponential term to improve Cramér-Lundberg approximation. Tijms' approximation for $1 - F(t)$ is

$$\gamma e^{-ct} + (\rho - \gamma) e^{-\beta t}, \tag{2.1.8}$$

where

$$\beta = \frac{\rho - \gamma}{EW_q - \gamma/c}. \tag{2.1.9}$$

Note that the first term of Tijms approximation (2.1.8) is the exponential term of Cramér-Lundberg approximation and it was mentioned that the Tijms approximation can be applied only if

$$\beta > c \tag{2.1.10}$$

since then for large t, the approximation would agree to the Cramér-Lundberg approximation, which is an asymptotic expansion of $1 - F_q(t)$.

If $t = 0$, the value of Tijms approximation is ρ, which is the true value of $1 - F_q(0)$ and β is chosen to match the first moment of W_q, i.e. β is the solution of the equation,

$$\int_0^\infty \left\{\gamma e^{-ct} + (\rho - \gamma) e^{-\beta t}\right\} dt = \frac{\gamma}{c} + \frac{\rho - \gamma}{\beta} = EW_q.$$

Example 39. For M/M/1 queue, $c = \mu - \lambda$ and $\gamma = \rho$. By Example 28, the Cramér-Lundberg approximation is exact. For the Tijms approximation, the second terms in the Tijms approximation has weight $\rho - \gamma = 0$.

For Erlang service time distribution (i.e., $G \sim \text{Gamma}(k, \beta)$), see Example 31.

Again, if we use $\hat{\beta} = (\hat{\rho} - \hat{\gamma})/(\widehat{EW_q} - \hat{\gamma}/\hat{c})$, where $\widehat{EW_q}$ is the moment estimator defined by

$$\widehat{EW_q} := \hat{\mathcal{W}}_q'(0) = \frac{\overline{G^2}}{2(\overline{I} - \overline{G})} = \frac{\hat{\mathcal{G}}''(0)}{2\{\overline{I} - \hat{\mathcal{G}}'(0)\}},$$

then the empirical version of the Tijms approximation can be used as the approximation of $\hat{F}_q(t)$. Figures 2.1.4 and 2.1.5 show the Cramér-Lundberg approximation and the Tijms approximation of $\hat{F}_q(t)$ of Figure 2.1.1 and 2.1.2 on the left graphs and their % absolute relative errors against $\hat{F}_q^{\dagger}(t)$.

2.1.5 Other approximations - Willmot's and Sakurai's

Based on the idea of the Tijms approximation, Willmot ([71]) proposed that

$$P(W_q \geq t) \sim \gamma e^{-ct} + (\rho - \gamma)\overline{C}(t),$$

where $1 - \overline{C}(x)$ is a distribution function satisfying

$$(\rho - \gamma)\int_0^\infty \overline{C}(t)\,dt = EW_q - \frac{\gamma}{c},$$

so that the first moment is matched. Thus, his approximation also preserves $P(W_q = 0)$ and EW_q as with Tijms' approximation.

The suggested choice of distribution of C is Gamma(k, θ), where k and θ are selected to match the first two moments of W_q. Let μ_j' be the jth non-central moment of G. Solving the following equations simultaneously,

$$\begin{cases} (\rho - \gamma)\dfrac{k}{\theta} = \left(EW_q - \dfrac{\gamma}{c}\right) \\ (\rho - \gamma)\left(\dfrac{k + k^2}{\theta^2}\right) + \gamma\left(\dfrac{2}{c^2}\right) = EW_q^2, \end{cases}$$

where

$$EW_q^2 = \frac{\lambda\left\{3\mu_2'\lambda + 2\mu_3'(1 - \rho)\right\}}{6(1 - \rho)^2},$$

we have the solutions,

$$k = \left\{\frac{(\rho - \gamma)\left(EW_q^2 - \frac{2\gamma}{c^2}\right)}{\left(EW_q - \frac{\gamma}{c}\right)^2} - 1\right\}^{-1}$$

$$\theta = k\frac{(\rho - \gamma)}{\left(EW_q - \frac{\gamma}{c}\right)}.$$

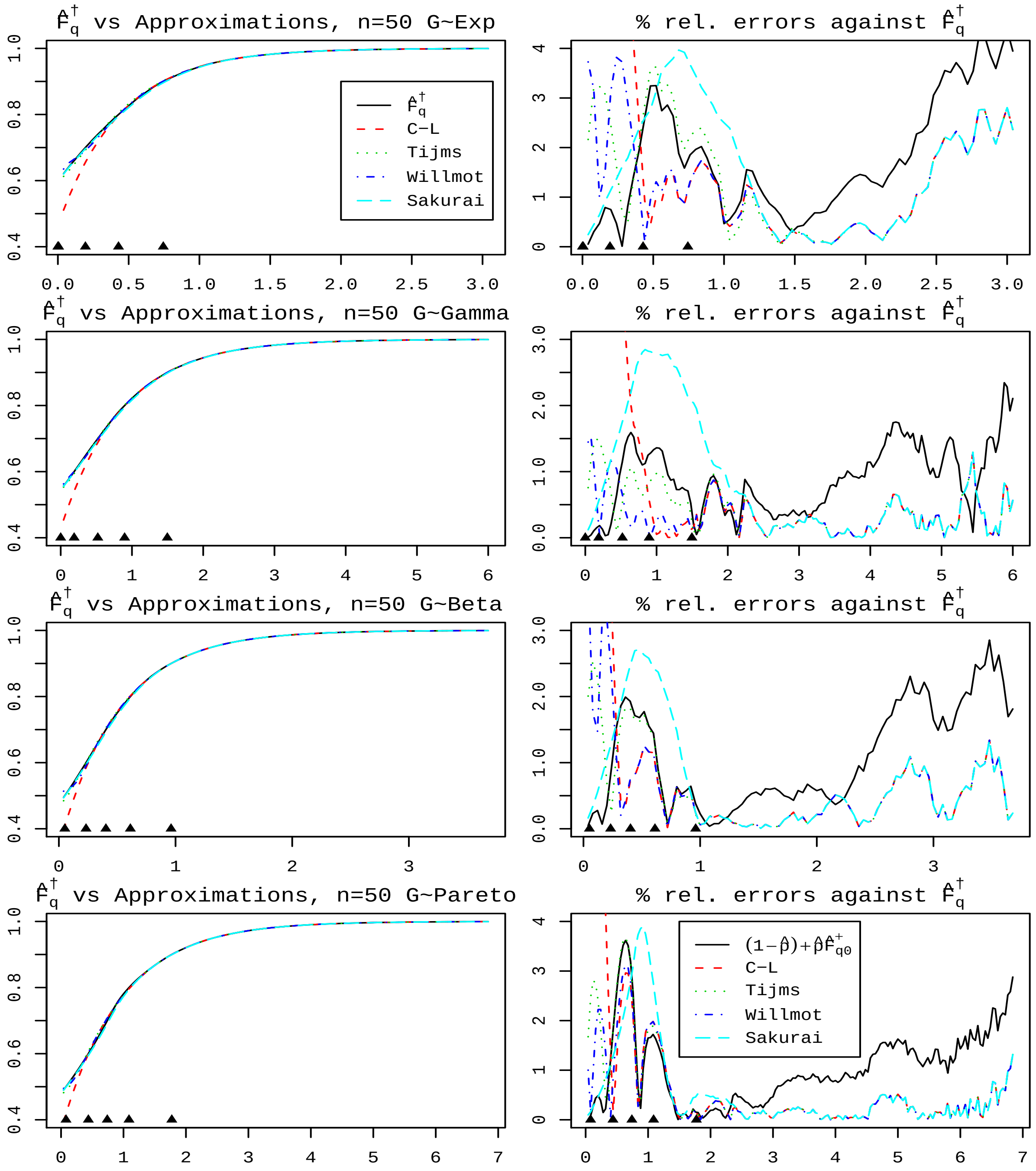

Figure 2.1.4: Left: The empirical CDF's, $\hat{F}_q^\dagger(t)$, of Figure (2.1.1) with the Cramér-Lundberg approximations, the Tijms approximation, the Willmot approximation, and the Sakurai approximation. Right: % relative Absolute errors of each approximation against $\hat{F}_q^\dagger(t)$. In each graph, ▲ denotes a decile (from 10% to 90%) of the sampling distribution of $\hat{W}_q$.

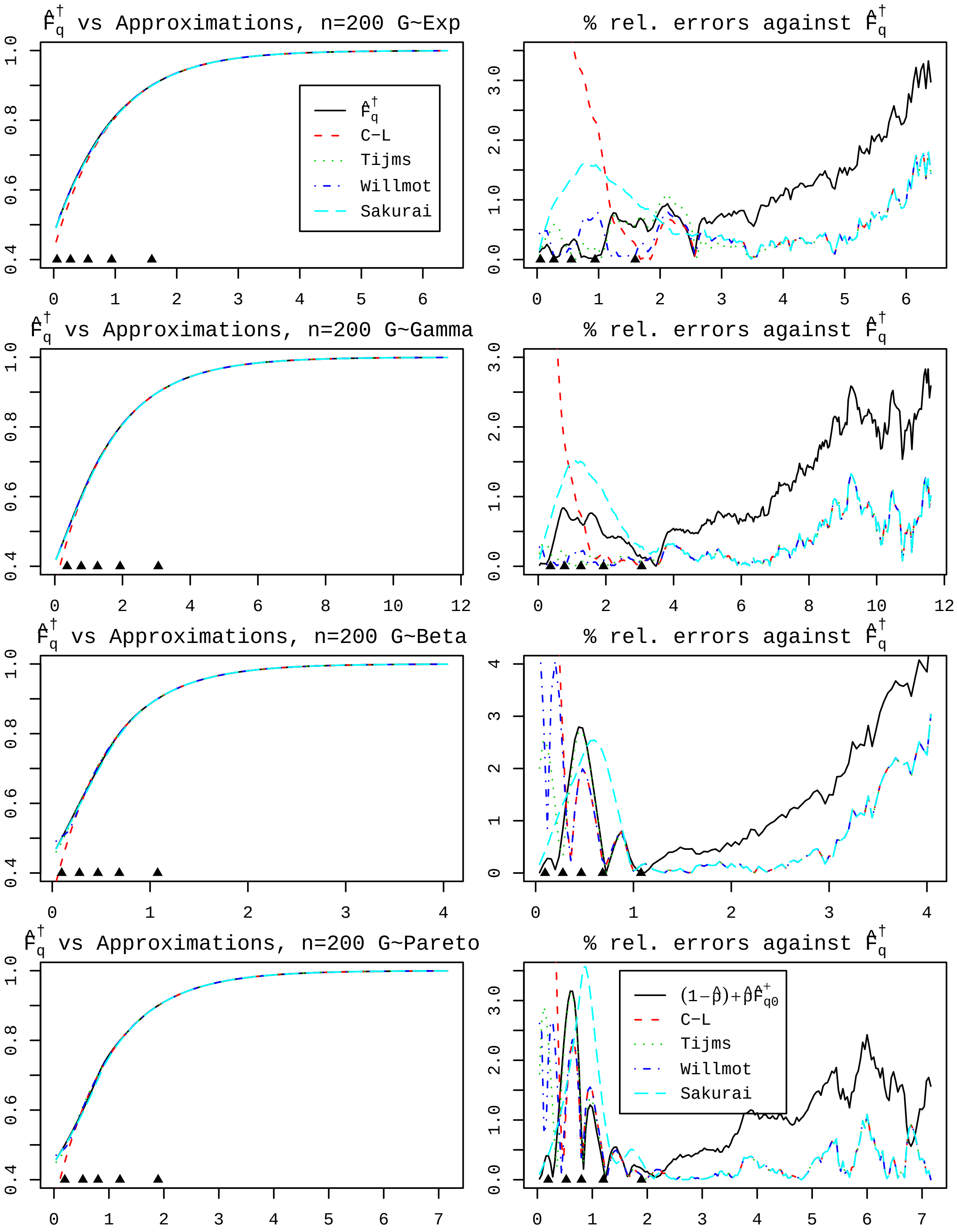

Figure 2.1.5: Same as Figure 2.1.4 with $n = 200$.

or

$$\begin{aligned} k &= \left\{ \frac{\beta^2}{\rho - \gamma} \left(EW_q^2 - \frac{2\gamma}{c^2} \right) - 1 \right\}^{-1} \\ \theta &= k\beta, \end{aligned}$$

where β is defined in (2.1.9).

In [71], it was noted that his approximation is asymptotically valid if

$$k > 1 - \frac{\rho - cEW_q}{(\rho - \gamma)} \tag{2.1.11}$$

and if the calculated k is negative or not satisfying the above condition, the smallest positive integer satisfying the above condition is used as k.

Note that the right term of the inequality in the above condition (2.1.11) can be simplified as

$$1 - \frac{\rho - cEW_q}{(\rho - \gamma)} = \frac{\rho - \gamma - \rho + cEW_q}{\rho - \gamma} = \left(\frac{EW_q - \gamma/c}{\rho - \gamma} \right) c = \frac{c}{\beta},$$

so that the Willmot's condition (2.1.11) can be written as

$$k > \frac{c}{\beta}.$$

Comparing (2.1.11), Tijms' condition (2.1.10) can be regarded as Willmot's condition with k fixed as 1 and the Willmot's approximation is more flexible than Tijms' condition. We also note that if k is chosen to be 1, then $\theta = \beta$, so that the Willmot's approximation coincides with the Tijms' approximation.

Sakurai's approximation ([59]) is based on truncating the argument of G_e ([13] and [18]), the equilibrium distribution of G.

Let $G_e(t) = \mu \int_0^t \{1 - G(x)\} \, dx$. Sakurai proposed an approximation,

$$P(W_q > t) \approx \gamma_t e^{-\delta_t t} + \frac{\rho}{1 - \rho G_e(t)} \{1 - G_e(t)\},$$

where γ_t and δ_t are similarly defined for truncated G_e. In other words, δ_t is the solution of

$$\int_0^t e^{\delta_t x} dG_e\left(x\right) = \rho^{-1} \tag{2.1.12}$$

and

$$\gamma_t := \frac{1-\rho}{\rho\delta_t \int_0^t x e^{\delta_t x} dG_e\left(x\right)}.$$

Note that since $G_e\left(0\right) = 0$, the Sakurai's approximation also preserves $P\left(W_q = 0\right) = 1-\rho$ assuming $\gamma_0 = 0$. Though this approximation was developed for asymptotic approximation with heavy-tailed service time distributions where Cramér-Lundberg approximation cannot be used, it was also claimed to be useful for light tailed service distributions.

Since $\int_0^t e^{\delta_t x} dG_e\left(x\right)$ and $\int_0^t x e^{\delta_t x} dG_e\left(x\right)$ do not always have closed forms, it was suggested to use numerical integrations or moments based estimators in [59]. However, for the case of the empirical MGF, δ_t and γ_t can be calculated by the following proposition.

Proposition 40. *Let $G_{(j)}$ be the jth order statistic of $\{G_j\}_{j=1}^n$. Then the following hold:*

1. *$\hat{\delta}_t$ is the root of the equation $\hat{\lambda} - \hat{\lambda}\hat{\mathcal{G}}_t(\hat{\delta}_t) + \hat{\delta}_t = 0$ and $\hat{\gamma}_t$ is given by*

$$\hat{\gamma}_t = \frac{1-\hat{\rho}}{\hat{\lambda}\hat{\mathcal{G}}_t'(\hat{\delta}_t) - 1},$$

where

$$\hat{\mathcal{G}}_t\left(s\right) = \frac{1}{n}\left(\sum_{j=1}^{n'} \exp G_{(j)}s + \left(n-n'\right)\exp(ts)\right)$$

is the empirical MGF of $\left\{G_{(1)}, G_{(2)}, \cdots, G_{(n')}, \underbrace{t, \cdots, t}_{n-n'}\right\}$ *with*

$$n' := \max\left\{j : G_{(j)} < t\right\}.$$

2. *For $t \geq \inf\left\{t : G\left(t\right) = 1\right\}$ ($t > G_{(n)} := \max G_i$), the Sakurai approximation coincides with the Cramér-Lundberg approximation. Thus, for the empirical approximation, if*

$t \geq G_{(n)}$, the empirical Sakurai approximation coincides with the empirical Cramér-Lundberg approximation.

Proof. Let $n' := \max\left\{j : G_{(j)} < t\right\}$. Then, since $\hat{G}(t) = n'/n$, we have

$$\begin{aligned}\int_0^t e^{sx} d\hat{G}_e(x) =& \hat{\mu}\left(\int_{G_{(n')}}^t + \int_{G_{(n'-1)}}^{G_{(n')}} + \cdots + \int_0^{G_{(1)}}\right) e^{sx}\left\{1 - \hat{G}(x)\right\} dx \\ =& \frac{\hat{\mu}}{s}\bigg\{\frac{n-n'}{n}\left(\exp st - \exp sG_{(n')}\right) + \frac{n-n'+1}{n}\left(\exp sG_{(n')} - \exp sG_{(n'-1)}\right) \\ & + \cdots + \exp sG_{(1)} - 1\bigg\} \\ =& \frac{\hat{\mu}}{s}\left\{\frac{n-n'}{n}\exp st + \frac{1}{n}\sum_{j=1}^{n'}\exp sG_{(j)} - 1\right\}.\end{aligned}$$

and solving (2.1.12) implies

$$\begin{aligned}& \frac{\hat{\mu}}{\hat{\delta}_t}\left\{\frac{n-n'}{n}\left(\exp\hat{\delta}_t t\right) + \frac{1}{n}\sum_{j=1}^{n'}\exp\hat{\delta}_t G_{(j)} - 1\right\} = \frac{\hat{\mu}}{\hat{\lambda}} \\ \Longrightarrow & \frac{1}{\hat{\delta}_t}\left\{\frac{n-n'}{n}\left(\exp\hat{\delta}_t t\right) + \frac{1}{n}\sum_{j=1}^{n'}\exp\hat{\delta}_t G_{(j)} - 1\right\} = \frac{1}{\hat{\lambda}} \\ \Longrightarrow & \hat{\lambda}\left\{\frac{n-n'}{n}\left(\exp\hat{\delta}_t t\right) + \frac{1}{n}\sum_{j=1}^{n'}\exp\hat{\delta}_t G_{(j)} - 1\right\} - \hat{\delta}_t = 0.\end{aligned}$$

Observing that $\frac{n-n'}{n}\exp st + \frac{1}{n}\sum_{j=1}^{n'}\exp sG_{(j)}$ is the empirical MGF of

$$\left\{G_{(1)}, G_{(2)}, \cdots, G_{(n')}, \underbrace{t, \cdots, t}_{n-n'}\right\},$$

then $\hat{\delta}_t$ solves $\hat{\lambda}\{\hat{\mathcal{G}}_t(\hat{\delta}_t) - 1\} - \hat{\delta}_t = 0$.

For $\hat{\gamma}_t$, we just point out that the conditions of Proposition 38 are satisfied in this case.

For (2), we note that if $t \geq \inf\{t : G(t) = 1\}$, then

$$\int_0^t e^{\delta_t x} dG_e(x) = \mathcal{G}_e(\delta_t),$$

which implies $\delta_t = c$ and $\hat{\gamma}_t = \gamma$. Also, we note $G_e(t) = 0$ because $\mu\{1 - G(t)\} = 0$. Thus, the Sakurai approximation coincides with the Cramér-Lundberg approximation. □

2.2 Comparisons of the approximation of $F_q(t)$ with gamma and uniform service distributions

For the comparison of the approximations, we first use Gamma (3,3) service time distribution with $\lambda = .5$, which is the only common example of the primer references of each approximation ([68], [71], [59]). See Table 2.2.1 for the values of each approximation of $P(W_q > x)$ calculated from the true MGF and Figure 2.2.1 for their the percentage absolute errors and absolute errors.

We also compare those approximations for Uniform(0,2) service time distribution with λ=.5 in Figure 1.3.14.

Each figure shows that good performance of the saddlepoint approximation method over all the range of W_q.

Note that since G has the bounded range of $(0, 2)$, the Sakurai approximation coincides with the Cramér-Lundberg approximation for $t > 2$ in the Uniform(0,2) service time distribution case.

Table 2.2.1: Comparison of various approximations of $P(W_q > x)$, where W_q is the stationary waiting time in queue with Gamma(3,3) service time and λ=.5. C-L denotes the Cramér-Lundberg approximation.

t	$P(W_q > t)$	Saddlepoint	C-L	Tijms	Willmot	Sakurai
.005	.4987	.4987	.5775	.4989	.4982	.4988
.1	.4744	.4743	.5332	.4764	.4728	.4752
.25	.4342	.4338	.4700	.4361	.4344	.4367
.5	.3664	.3644	.3809	.3665	.3672	.3711
.75	.3033	.3006	.3088	.3026	.3037	.3089
1	.2484	0.2460	.2502	.2476	.2484	.2536
2	.107988	.107620	.107981	.107896	.107950	.109327
3	.046595	.046682	.046594	.046592	.046594	.046751
4	.02010579	.02021472	.02010577	.020105684	.02010577	.02011895
6	.003743644	.003783579	.003743644	.003743644	.003743644	.003743713
8	.000697057	.000707296	.000697057	.000697057	.000697057	.000697057

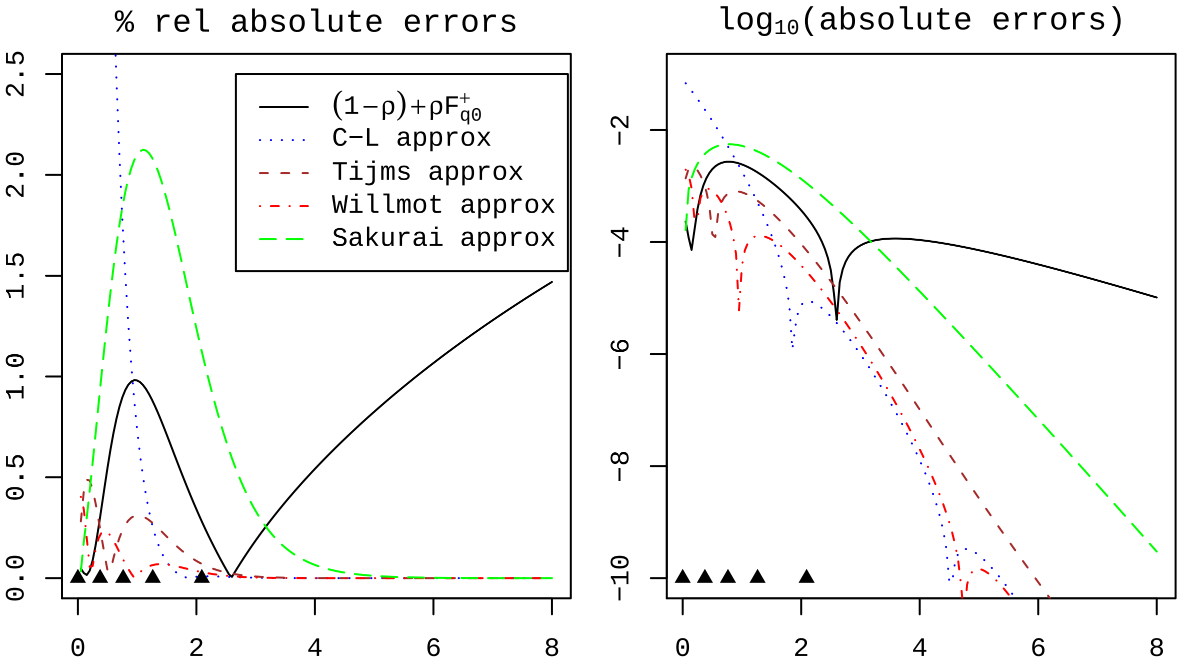

Figure 2.2.1: The percentage relative absolute errors and $\log_{10}$(absolute errors) of different approximation of the CDF of the stationary waiting time in queue W_q with Gamma(3,3) service time distribution and λ=.5. In each graph, ▲ denotes a decile (from 50% to 90%) of the distribution of W_q .

2.3 Comparisons of the empirical approximations as estimators of $F_q\left(t\right)$

We now look into the performance of empirical versions of the approximations as estimators of $F_q\left(t\right)$. Figure 2.3.1 and 2.3.2 shows the estimated absolute bias and the MSE on $\log_{10}$-scales of each approximation of the $l = 10^4$ random samples.

Note that all estimators of constants from the 4 different approximations mentioned above (γ and c of the Cramér-Lundberg approximation, β of the Tijms approximation, k and θ of the Willmot approximation, and δ_x and γ_x of the Sakurai approximation) based on the empirical MGF, converge to their true values a.s. so that their resulting empirical approximations all converge a.s. to their true approximations. This can be shown in a similar way to the proof of the a.s. convergence of $\hat{F}_{q0}\left(t\right)$ to $F_{q0}\left(t\right)$ (Corollary 18).

Thus, as with the saddlepoint approximation, empirical versions of these approximations can be regarded as the estimators of $F_q\left(t\right)$ and also of $\hat{F}_q\left(t\right)$.

We first note that the Tijms condition is not satisfied all the time and the Willmot approximation has a similar defect too, which will be explained. Therefore, if the Tijms' condition is not satisfied for a sample $\{\{G_j\}_{j=1}^n, \{I_j\}_{j=1}^n\}$, the Cramér-Lundberg approximation for that sample is counted as the Tijms approximation for that sample and the same goes for the Willmot approximation.

However, the Willmot approximation still has the several unusual cases which make the average bloating in the tail area as one can see in the graphs of G~Exp(2), and G~Pareto(.8,5) with $n = 200$ case of Figure 2.4.1. From the result of [71], the tail of the Willmot approximation should be very close to the Cramér-Lundberg approximation when the Willmot condition is satisfied, which can always be done by choosing a accordingly unless $\beta \leq 0$.

Because the purpose of the chapter is to evaluate the performance of the saddlepoint approximation and compare it to four other approximation methods, we did not investigate why this behavior of the Willmot approximation happens.

Table 2.3.1: The estimated probability of $\hat{\beta} \leq \hat{c}$ (the case not satisfying Tijms' condition) and $\hat{\beta} \leq 0$ (the case that the Willmot approximation cannot be used) from $l = 10^4$ random samples of $\{\{G_j\}_{j=1}^n, \{I_j\}_{j=1}^n\}_{k=1}^l$ for different service times.

	G ∼Exp		G ∼Gamma		G ∼Beta		G ∼Pareto	
n	50	200	50	200	50	200	50	200
$\hat{P}\left(\hat{\beta} \leq \hat{c}\right)$	.1164	.1987	.0058	.0078	.0000	.0000	.0197	.0879
$\hat{P}\left(\hat{\beta} \leq 0\right)$	.0424	.0994	.0027	.0062	.0000	.0000	.0180	.0736

The defect of the Willmot approximation we mentioned is that there is a chance that $\hat{\beta} \leq 0$, in which case, the Tijms condition cannot be met, but Willmot's condition is still met with $\hat{k} = 1$. However, $\hat{\theta}$ becomes negative so that the Willmot approximation cannot be used.

Note that $\{\hat{\beta} \leq 0\} \subset \{\hat{\beta} \leq \hat{c}\}$. Thus, in summary, if the Willmot approximation cannot be used, then the Tijms approximation cannot be used but not vice versa and there is a chance that the Willmot approximation cannot be used for a random sample of $\{\{G_j\}, \{I_j\}\}$. Table 2.3.1 shows the estimated probability of $\hat{\beta} \leq \hat{c}$ and $\hat{\beta} \leq 0$ from $l = 10^4$ random samples of $\{\{G_j\}_{j=1}^n, \{I_j\}_{j=1}^n\}_{k=1}^l$ for different service times and different sample sizes n=50, 200.

The true CDFs of the exponential and the gamma service time distributions were obtained from Theorem (26) but for the beta and the Pareto service time distribution, the empirical CDF $F_q^{\dagger}(t)$ was used from a sample of simulated W_q of the size $m = 3 \cdot 10^7$ in place of $F_q(t)$.

Again, if the Tijms' condition is not satisfied for a sample $\{\{G_j\}_{j=1}^n, \{I_j\}_{j=1}^n\}$, the Cramér-Lundberg approximation for that sample is counted as the Tijms approximation for that sample and the same goes for the Willmot approximation.

A more clear way to see the estimated bias would be comparing the bias to the true value of $F_q(t)$. Figure 2.3.3 shows the % relative absolute bias of different approximations. For example, the % relative absolute bias of the saddlepoint CDF approximation is defined by

$$\frac{100\left|E\{\hat{F}_{q0}(x)\} - F_q(t)\right|}{\min\{F_q(t), 1 - F_q(t)\}}, \tag{2.3.1}$$

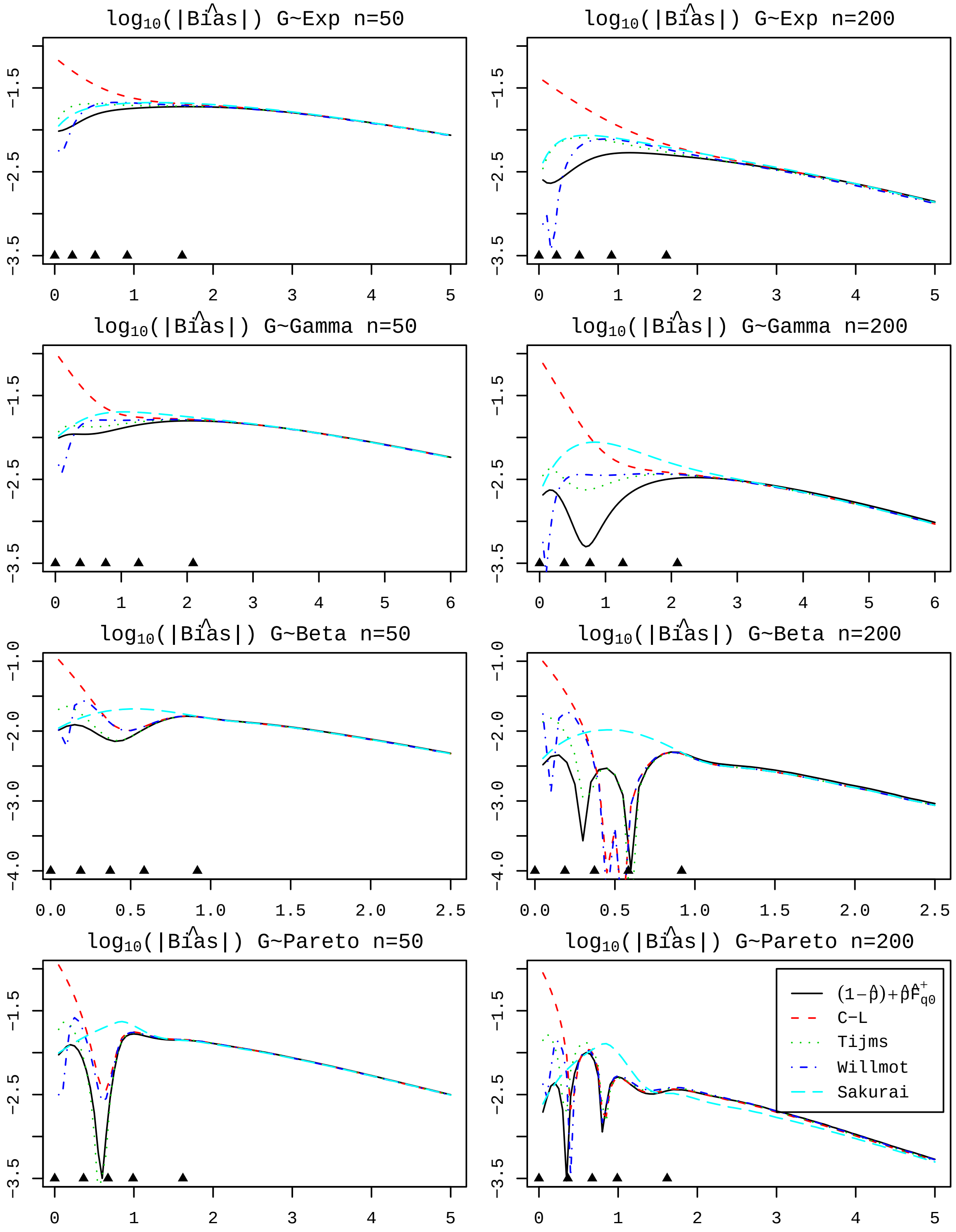

Figure 2.3.1: The estimated absolute bias of the different empirical approximations in $\log_{10}$-scale for W_q from the $l = 10^4$ random samples of $\{\{G_j\}_{j=1}^n, \{I_j\}_{j=1}^n\}_{k=1}^l$ used in Table 2.3.1. The x-axis is cut off to include at least 99.5% of W_q.

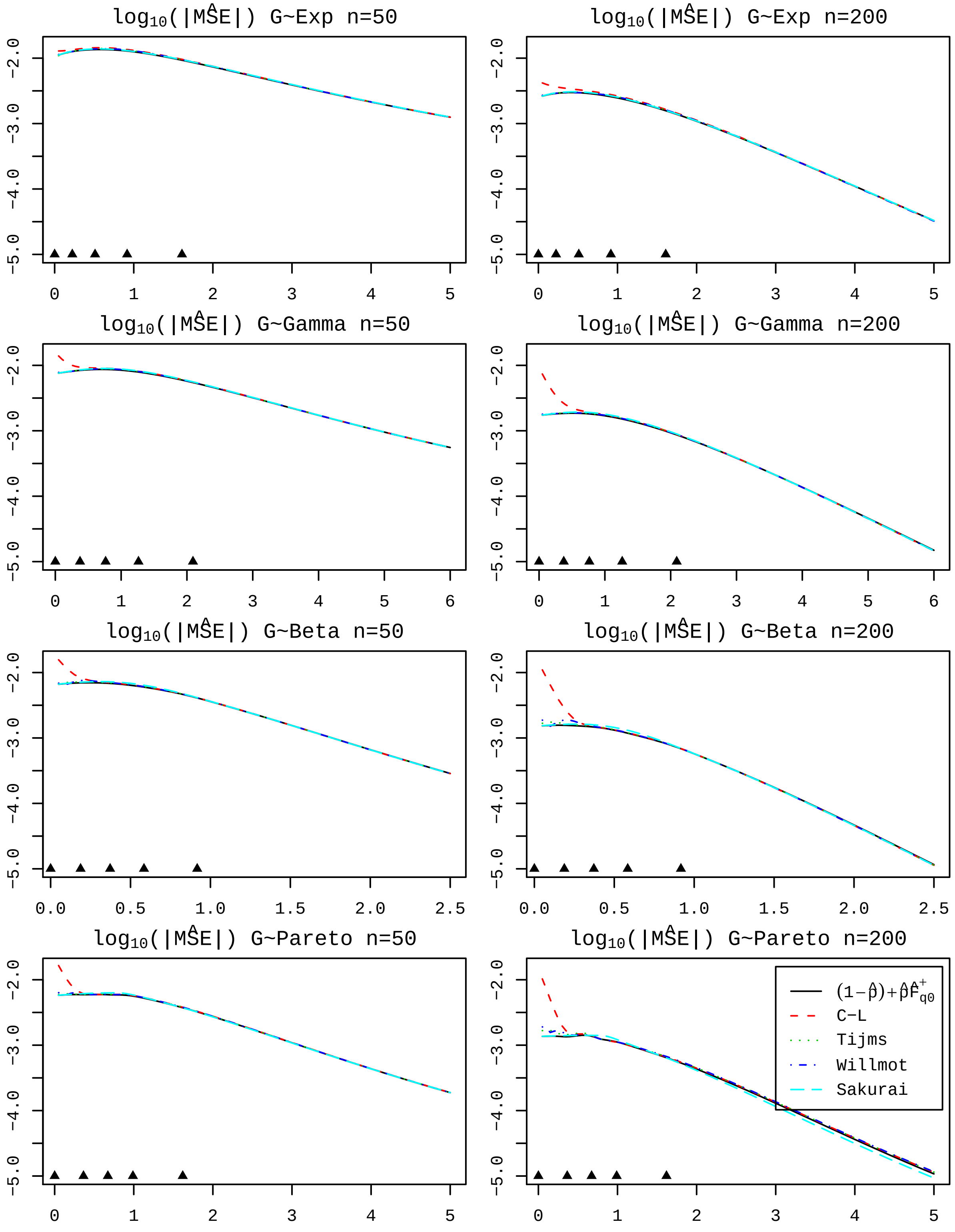

Figure 2.3.2: The estimated MSE from Figure 2.3.1.

and is estimated by

$$\begin{aligned}\frac{100}{\min\{F_q\left(t\right),1-F_q\left(t\right)\}}&\left|\frac{1}{l}\sum_{j=1}^{l}\hat{F}_{q0,j}\left(t\right)-F_q\left(t\right)\right|\\&=\frac{100}{\min\{F_q\left(t\right),1-F_q\left(t\right)\}}\left|\hat{\text{Bias}}\right|,\end{aligned}$$

where $\hat{F}_{q0,j}\left(t\right)$ is the saddlepoint CDF approximation from jth set of the random sample. One can see the saddlepoint approximation is competitive well with the asymptotic approximation in the tail areas.

Graphs for the exponential, gamma, and beta service times with $n = 200$ show that the % relative absolute bias is smaller than 10% till 90th percentile but it grow exponentially for the tail area. This is not unexpected. Note that for the light tail service time distribution, by Cramér-Lundberg approximation, $1 - F\left(t\right) \sim \gamma \exp -ct$. Thus, after 50th percentile, the relative bias is

$$\frac{100\left|\text{Bias}F\left(t\right)\right|}{\min\left\{F\left(t\right),1-F\left(t\right)\right\}} \sim 100\left|\text{Bias}F\left(t\right)\right|\left(\gamma^{-1}e^{ct}\right).$$

Unless $\left|\text{Bias}F\left(t\right)\right|$ decreases exponentially faster than e^{-ct}, the % relative absolute bias will increase exponentially. A similar argument is possible for Pareto service time distribution case. By the result of [49], we have

$$1-F\left(t\right)\sim\frac{\rho}{1-\rho}\{1-G_e\left(t\right)\}=\frac{256}{3125t^4}\mathbf{1}_{\left(4/5,\infty\right)}\left(t\right),$$

so that the % relative absolute bias will be $\left|\text{Bias}F\left(t\right)\right|O\left(t^4\right)$.

From each figures we conclude that the saddlepoint approximation is the best approximation among 5 different approximations in the cases we compared.

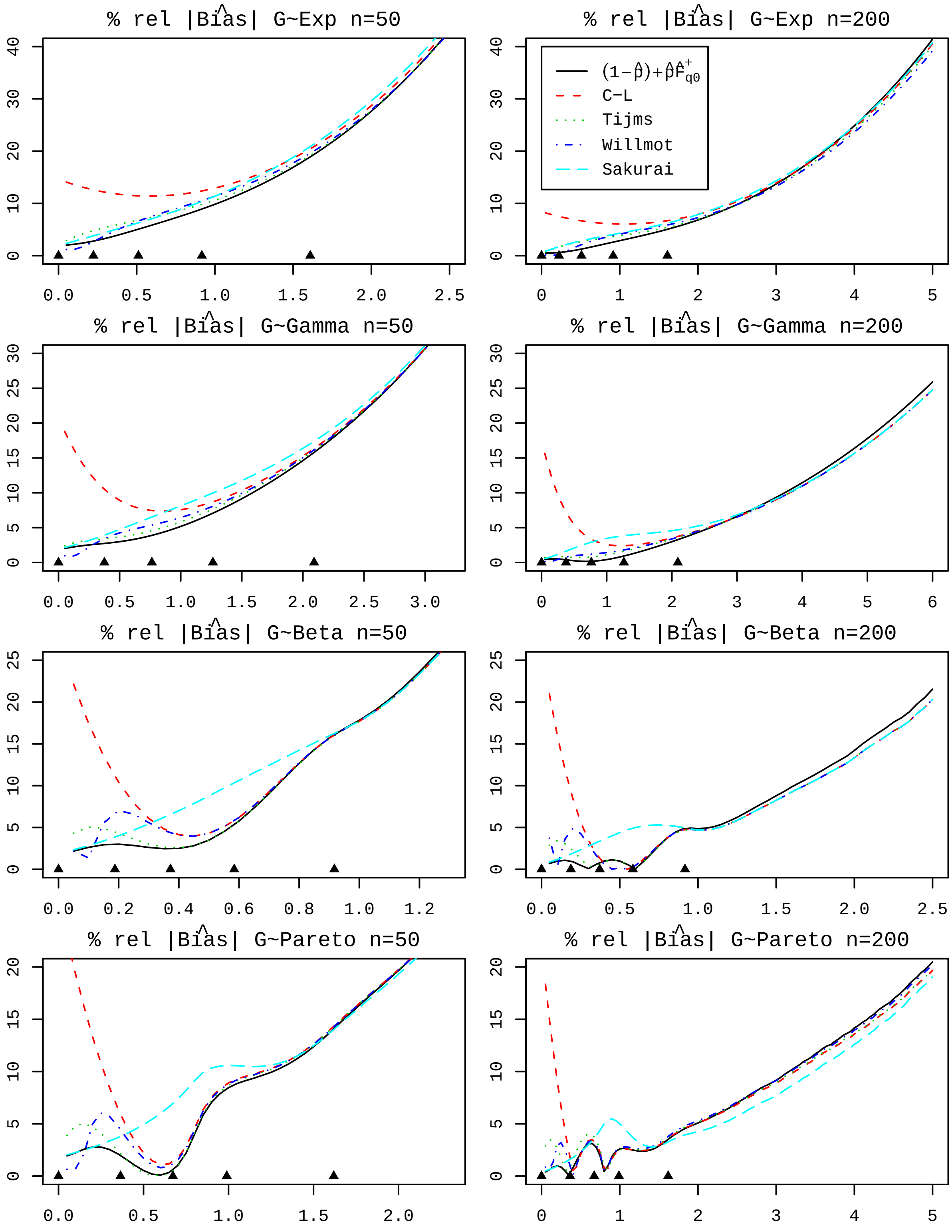

Figure 2.3.3: The estimated % relative absolute bias of the empirical approximations of Figure 2.3.1.

2.4 Comparisons of the empirical approximations as estimators of $\hat{F}_q(t)$

We now look into to the performance of the empirical versions of the 5 different approximation methods as estimators of $\hat{F}_q(t)$. Figures 2.1.4 and 2.1.5 show the Willmot approximation and the Sakurai approximation of $\hat{F}_q(t)$ on the left graphs and their % absolute relative errors against $\hat{F}_q^{\dagger}(t)$ on the right. All of the graphs show that the saddlepoint approximation performs better than the other methods if x is smaller than 90th percentile of $\hat{W}_q$ and still perform comparatively in the tail areas. Note each end point of the x-axes is greater than 99.8th percentile for each $\hat{W}_q$.

As we did in Figure 1.3.7, we obtained each empirical approximation from the random samples we used in Figure 1.3.7 and compute the the average of the percentage absolute relative errors against $\hat{F}_q^{\dagger}(t)$, which is obtained from the random sample of $\hat{W}_q$ of the size, $m = 10^7$. See Figure 2.4.1 for the graphs.

We remark one thing regarding the graphs. For $G \sim \text{Exp}(2)$ with $n = 50$ case, there is one sample whose estimated CDF is 1 at $t = 4.8$ and at $t = 6$, 6 out of 1000 have the estimated CDF value 1. In this case, the estimated relative error becomes ∞ so that the average was not drawn after $t = 4.8$.

2.5 Conclusion

We have shown that the saddlepoint approximation with the empirical MGF can be used as a reliable approximation method to approximate the CDF (and the PDF) of W_q as we did in Chapter 1. Comparing the known asymptotic approximations specifically for $F_q(t)$ and $\hat{F}_q(t)$, the saddlepoint approximation performs on par with them even in the tail area, where those approximation methods have been targeted to work. Note that obtaining asymptotic approximations for a specific stochastic process requires a deep knowledge of probability theory and real (and complex) analysis and is not always possible. Saddlepoint

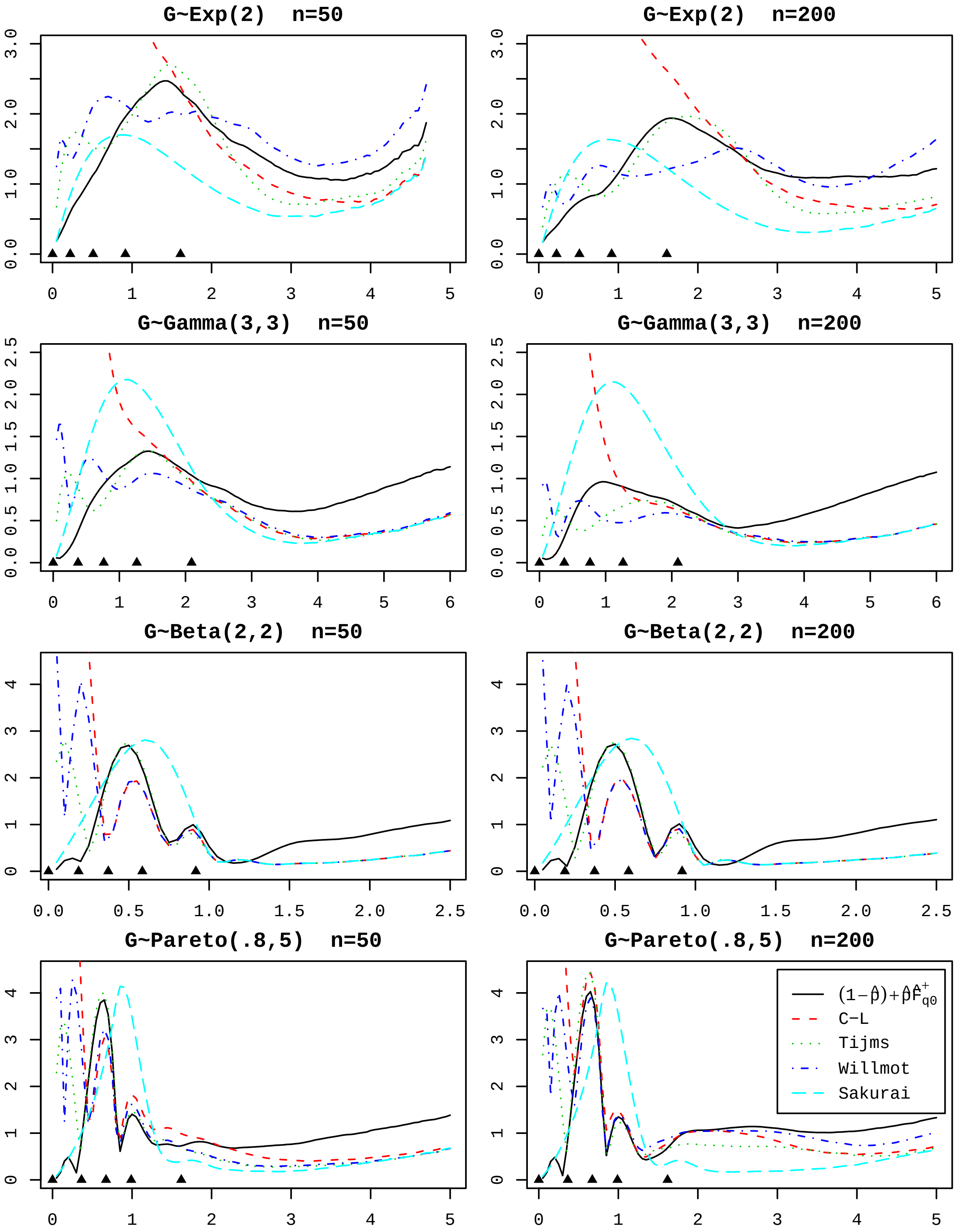

Figure 2.4.1: The average percentage relative absolute error of the different approximation methods against $\hat{F}_q^\dagger(t)$ from the $l = 10^3$ random samples used in Figure 1.3.7. In each graph, ▲ denotes a decile (from 50% to 90%) of the distribution of W_q .

approximation does not require such knowledge to apply but shows the superior performance.

The summary of our findings is as follows. When $\mathcal{W}_q(s)$ is steep, then $F_{q0}(t)$, as a probability approximation to $F_q(t)$, works better than its four competitors when t is smaller than the 90th percentile. Above the 90th percentile, the Cramér-Lundberg approximation and its variations dominate. As a statistical estimator, $\hat{F}_{q0}(t)$ is more accurate than all other plug-in estimators for all $t \geq 0$. From the examples, this appears to hold for both light-tailed and heavy-tailed distributions. It even appears to hold for cases in which $\mathcal{W}_q(s)$ is not steep (so $F_{q0}(t)$ cannot be computed) although $\hat{F}_{q0}(t)$ can always be computed since $\hat{\mathcal{W}}_q(s)$ is always steep.

We consider the two cases $\mathcal{W}_q(s)$ steep and $\mathcal{W}_q(s)$ not steep separately.

- $\mathcal{W}_q(s)$ steep: If $0 < c < b/2$, then $\hat{F}_{q0}(t)$ is a $O_p\left(n^{-1/2}\right)$ consistent estimator of $F_{q0}(t)$ which, by regularity of the saddlepoint theory involved, should be very close to $F_q(t)$ for all t. If $0 < b/2 < c < b$, then $\hat{F}_{q0}(t)$ is a $O_p\left(n^{-\delta}\right)$ $(0 \leq \delta < \min\{1/2, 1 - c/b\})$ consistent estimator and still $F_{q0}(t) \approx F_q(t)$. So, $\hat{F}_{q0}$ is justified. If $b = 0$, then $\hat{\mathcal{W}}_q(-s) \to \mathcal{W}(-s)$ as $n \to \infty$. So, from a less rigorous perspective, we expect $\hat{F}_q(t) \approx F_q(t)$. However, $\hat{\mathcal{W}}(s)$ converges on $\Re s < \hat{c}(>0)$ so approximation $\hat{F}_{q0}(t) \approx \hat{F}_q(t)$ is still a regular case for saddlepoint approximation for which great accuracy can still expected. Thus, $\hat{F}_{q0}(t) \approx \hat{F}_q(t) \approx F_q(t)$ and we anticipate $\hat{F}_{q0}(t)$ still performs well as an estimate of $F_q(t)$, despite the fact that $F_q(t)$ may have a heavy tail.

  From another perspective, steepness assumes that $F_{q0}(t)$ exists and if $F_{q0}(t) \approx F_q(t)$, then albeit slowly, $\hat{F}_{q0}(t) \to F_{q0}(t)$ a.s. (follows from $\hat{\mathcal{W}}(s) \to \mathcal{W}(s)$ a.s. for $\Re s \leq 0$). So again, $\hat{F}_{q0}(t) \approx F_{q0}(t) \approx F_q(t)$.

- $\mathcal{W}_q(s)$ not steep: Numerical computations and simulations suggest $\hat{F}_{q0}(t)$ remains a good estimator of $F_q(t)$ even when $F_{q0}(t)$ does not exist and $\mathcal{W}_q(s)$ is not steep. When claim size follows a Pareto(0.8, 5) distribution, then $\mathcal{W}_q(s)$ is not steep and $b = 0$ but $\hat{F}_{q0}(t)$ still demonstrates excellent performance under simulations. Nonrigorous justification follows the case in which $b = 0$ and $\mathcal{W}(s)$ is steep. Essentially, $\hat{F}_q(t) \approx F_q(t)$ by $\hat{W} \Rightarrow W$ and $\hat{F}_{q0}(t) \approx \hat{F}_q(t)$ is still a regularly setting for saddlepoint approximation.

In conclusion, there are good reasons to expect that $\hat{F}_{q0}(t)$ will provide a good estimator of $F_q(t)$ even for settings in which W_q is heavy-tailed and perhaps lacking steepness.

2.6 Appendix

Here, we summarize some results we find after the author's PhD defense was done. By Theorem 20, if $c \leq b/2$, then $\hat{F}_0(t) - F_0(t) = O_p\left(n^{-1/2}\right)$ for any $t \in \mathbb{R}$ as in our simulation examples of G~Exp(2) ($c = 1 = b/2$), G~Gamma(3,3) ($c = .840474 < 1.5 = b/2$), and G~Beta(2,2) ($b = \infty$) cases. When G~Pareto(0.8,5), we cannot have any rate of the convergence result for $t > EW_q$. However, from Figure 2.3.1 and 2.3.2 we do not observe any slow convergence rate of $\hat{F}_0(t)$ on $t > 8/15 = EW_q$ for G~Pareto(0.8,5) case comparing to the other three service time distribution cases. In fact, the convergence rates look all the same for those 4 cases.

This phenomenon can be explained by considering the behavior of $\hat{F}_q(t)$ instead of $\hat{F}_{q0}(t)$. First, we show the uniform convergence of $\hat{F}_q(t)$.

Theorem 41. *$\hat{F}_q(t)$ converges to $F_q(t)$ uniformly on $[0, \infty)$.*

Proof. In a similar way shown in §1.3.1, we can show $\hat{W}_q \Rightarrow W_q$. Note that $F_q(t)$ is continuous. Thus, by the theorem known as Pólya's lemma (Exercise 7.2 of [61]. See also Exercise 7.13 of [58], which can be proved by Theorem 37), $\hat{F}_q(t)$ converges to $F_q(t)$ uniformly. □

In [51], $\hat{F}_q$ was regarded as an image of $\{\hat{G}(t), \hat{I}(t)\}$ under an operator map Φ, where $\hat{G}(t)$ and $\hat{I}(t)$ are the empirical CDF estimates of the service time distribution and the inter-arrival distribution, respectively. Theorem 41 can be proved using Theorem 3.1 and Lemma 3.2 of [51] though it requires $EG^{\gamma} < \infty$ for some $\gamma > 2$.

Let D_∞ be the space of all cádlág (right-continuous with the existence of left limits) functions on $[0, \infty]$ with the supremum norm and the open ball σ-field. For $f \in D_\infty$, $f(\infty) := \lim_{t \to \infty} f(t)$. The following can be obtained from Theorem 4.3 of [51].

Theorem 42. *If $EG^{\gamma} < \infty$ for some $\gamma > 4$, then*

$$\sqrt{n}\left(\hat{F}_q - F\right) \Rightarrow Z \quad \textit{as } n \to \infty \quad \textit{in } D_\infty$$

where Z is a Gaussian process.

Proof. By the proof of Theorem 4.3 of [51] with Theorem 4.1 and Lemma 4.2 of [51], we only need to show that for any $\beta > 2$, $\sqrt{n}\{\hat{I}(t) - I(t)\}$ converges weakly to a continuous stochastic process in D_β, where D_β is a subspace of D_∞ contains all the cádlág real-valued function f on $[0, \infty]$ such that $(1+x)^\beta f(x) \in D_\infty$ with the metric $\|f\|_\beta = \|(1+x)^\beta f(x)\|_\infty$ and $\hat{I}(t) = 1 - e^{-\hat{\lambda}t}$ (i.e., instead of using the empirical CDF of I_j's, we use the parametric estimate).

We use the functional delta method (Theorem 20.8 of [69]). Define $\phi : (0, \infty) \to D_\beta$ by $\phi(a) = e^{-at}$. For a fixed t, regarding $e^{-\lambda t}$ as a function of λ, we use Tayler's theorem (Theorem 5.15 of [58]) to get

$$e^{-(\lambda+h)t} = e^{-\lambda t} - e^{-\lambda t}h + \frac{e^{-\lambda_t' t}}{2}h^2,$$

where $|\lambda_t' - \lambda| \le h$. This suggests that the Fréchet derivative of ϕ at λ, $\phi_\lambda' : (0, \infty) \to D_\beta$, will be $\phi_\lambda'(h) = -e^{-\lambda t}h$ and we show that it actually is.

Clearly, ϕ_λ' is linear and continuous. Moreover, we have that for any $t \ge 0$,

$$\left|\phi(\lambda+h) - \phi(\lambda) - \phi_\lambda'(h)\right| = \left|\frac{e^{-\lambda_t' t}}{2}h^2\right| \le h^2 \left|\frac{e^{-(\lambda-|h|)t}}{2}\right|,$$

which implies

$$\left\|\phi(\lambda+h) - \phi(\lambda) - \phi_\lambda'(h)\right\|_\beta \le \frac{h^2}{2} \cdot \sup_{x\in[0,\infty)} \left|(1+t)^\beta e^{-(\lambda-|h|)t}\right| = O\left(h^2\right) \quad \text{as } |h| \searrow 0.$$

Thus, ϕ_λ' is the Fréchet derivative of ϕ at λ and by the functional delta method, we have

$$-\sqrt{n}\left\{\hat{I}(t) - I(t)\right\} = \sqrt{n}\left\{\phi\left(\hat{\lambda}\right) - \phi(\lambda)\right\} \Rightarrow \phi_\lambda'(Y) = -te^{-\lambda t}Y,$$

where $Y \sim N\left(0, \lambda^2\right)$ is the limit distribution of $\sqrt{n}(\hat{\lambda} - \lambda)$.

Therefore $\sqrt{n}\{\hat{I}\left(t\right) - I\left(t\right)\} \Rightarrow Yte^{-\lambda t}$ in D_β and we complete the proof. □

The previous theorem give us

$$\sup_{t \in [0,\infty]} \left|\hat{F}_q\left(t\right) - F_q\left(t\right)\right| = O_p\left(n^{-1/2}\right),$$

so that

$$\sup_{t \in [0,\infty]} \left|\hat{F}_{q0}\left(t\right) - F_q\left(t\right)\right| \leq \sup_{t \in [0,\infty]} \left|\hat{F}_{q0}\left(t\right) - \hat{F}_q\left(t\right)\right| + O_p\left(n^{-1/2}\right).$$

In a regular circumstance, as n increases, $\hat{F}_q\left(t\right)$ is getting smoother and $\hat{F}_{q0}\left(t\right)$, as a smooth estimate of $\hat{F}_q\left(t\right)$, will approximate $\hat{F}_q\left(t\right)$ better. Thus, even in the case of G~Pareto(0.8,5), $\hat{F}_{q0}\left(t\right)$ still behaves similarly to the other three cases and acts as a good estimate of $F_q\left(t\right)$.

Chapter 3

Saddlepoint Approximation to Kendall's Functional Equation

3.1 Introduction

In this chapter, we consider the application of the saddlepoint approximation to nonparametric estimation of M/G/1 queue busy time period distributions.

A saddlepoint approximation is used to obtain a nonparametric CDF estimator of this distribution using the empirical moment generating function of the service time distribution. Also, using the bootstrap method, we calculate a confidence interval (CI) for the saddlepoint estimation that occurs at each point of the CDF. We note that the CDF we estimate is not directly computable but its MGF can be derived by Kendall's functional equation as an implicit function of the MGF of the service time distribution which we assume can be estimated from data.

Kendall's functional equation for M/G/1 and an empirical MGF of stationary service times

Let $\mathcal{B}(s)$ be the moment generating function of the duration of the busy period B in the M/G/1 queue and let $\mathcal{G}(s)$ be the moment generating function of service time distribution.

Then Kendall-Takács functional equation for M/G/1 is

$$\mathcal{B}(s) = \mathcal{G}\{s - \lambda + \lambda\mathcal{B}(s)\}, \tag{3.1.1}$$

where λ is the arrival rate. It can be shown (Example (a), XIII.4 of ([30]) and p. 232 of [21]) that for $s \in (-\infty, 0]$, the equation (3.1.1) has a unique root $\mathcal{B}(s)$ and the distribution of busy time period is proper iff $\rho = \lambda/\mu < 1$, where $\mu = 1/\mathcal{G}'(0)$ is the service rate of the queue. In [3], it was shown that an iteration method of solution for $\mathcal{B}(s)$ can be used even for complex s if $\Re s < 0$.

Example 43. If $\mathcal{G}(s)$ has a simple form as in the M/M/1, $\mathcal{B}(s)$ can be calculated in the following way. The MGF of the exponential(μ) distribution is $(1 - s/\mu)^{-1}$. By setting $x = \mathcal{B}(s)$, then (3.1.1) requires the solution of

$$x = \left(1 - \frac{s - \lambda + \lambda x}{\mu}\right),$$

a quadratic equation with respect to x. Solutions are

$$x = \frac{(\lambda + \mu - s) \pm \sqrt{(\lambda + \mu - s)^2 - 4\lambda\mu}}{2\lambda}.$$

Since $\mathcal{B}(0)$ must be 1, the solution must be

$$\mathcal{B}(s) = \left\{(\lambda + \mu - s) - \sqrt{(\lambda + \mu - s)^2 - 4\lambda\mu}\right\}/2\lambda.$$

Clearly, $(\lambda + \mu - s)^2 - 4\lambda\mu$ must be non-negative, which gives us the convergence strip for $\mathcal{B}(s)$,

$$(-\infty, \mu + \lambda - 2\sqrt{\lambda\mu}] = (-\infty, (\sqrt{\mu} - \sqrt{\lambda})^2]. \tag{3.1.2}$$

Unless $\mathcal{G}(s)$ has a simple closed form as in M/M/1 queues, obtaining $\mathcal{B}(s)$ analytically is seldom possible.

Because we need to obtain $\mathcal{B}(s)$ for $s \geq 0$ to use saddlepoint approximations, a stable numerical method other than the iteration method mentioned above is needed. Moreover, Example 43 shows that the convergence strip of $\mathcal{B}\left(s\right)$ should be considered in the numerical method. These will be dealt with later.

Note that there is a series representation of the CDF and PDF of the busy period B. Let $F_B\left(t\right)$ be the CDF and $f_B\left(t\right)$ be the PDF of T. Then, it is known that

$$f_B\left(t\right)=\sum_{j=0}^{\infty}e^{-\lambda t}\frac{\left(\lambda t\right)^j}{\left(j+1\right)!}\frac{dG^{(j+1)*}}{dt}\left(t\right), \tag{3.1.3}$$

where $dG^{(n+1)*}\left(t\right)/dt$ is the PDF of $\sum_{i=1}^{n+1}G_i$ with iid service times G_i. Equation (3.1.3) is called the Takács series representation for the M/G/1 busy period density though it was independently obtained in [22] and [67].

If we replace $\mathcal{G}\left(s\right)$ by its empirical moment generating function

$$\hat{\mathcal{G}}\left(s\right)=\frac{1}{n}\sum_{i=1}^{n}e^{sG_i},$$

and use the numerical solution of Kendall-Takács functional equation (3.1.1) with $\mathcal{G}\left(s\right)$ replaced by $\hat{\mathcal{G}}\left(s\right)$, we obtain the derivative estimators of $\hat{\mathcal{B}}^{(k)}\left(s\right)$, $k=0,1,2$,3, which are needed for saddlepoint approximation. For example, $\hat{\mathcal{B}}\left(s\right)$ is defined by the solution of

$$\hat{\mathcal{B}}\left(s\right)=\hat{\mathcal{G}}\{s-\hat{\lambda}+\hat{\lambda}\hat{\mathcal{B}}\left(s\right)\}, \tag{3.1.4}$$

where $\hat{\lambda}=1/\overline{I}$ is the maximum likelihood estimator of λ using the inter-arrival times, $I_j\overset{\text{iid}}{\sim}\text{Exp}\left(\lambda\right)$. Note that we assume that the sample size of $\{I_i\}$ and $\{G_i\}$ are the same n to simplify the notation.

3.2 Solving the Kendall-Takács functional equation for saddlepoint approximations

To use the saddlepoint approximation, we need to find $\mathcal{B}^{(k)}(s)$, $k = 0, 1, 2$, which require solving the Kendall-Takács functional equation (3.1.1). In this section, we discuss how it can be solved generally and exploit the properties of $\mathcal{B}(s)$ and $\hat{\mathcal{B}}(s)$. We note that in our discussion, we assume that either λ and $\mathcal{G}(s)$ are known or $\hat{\lambda}$ and $\hat{\mathcal{G}}(s)$ are obtained from data. Also we assume that the inter-arrival time distribution (I_i) and the service time distribution (G_i) have finite second moments, are independent of each other, and $\rho = \lambda/\mu = EG_i/EI_i < 1$.

3.2.1 Solvability of the Kendall-Takács functional equation

Differentiate both sides of the Kendall-Takács functional equation (3.1.1) using the chain rule and solve for $\mathcal{B}'(s)$, to obtain

$$\mathcal{B}'(s) = \frac{\mathcal{G}'\{s - \lambda + \lambda\mathcal{B}(s)\}}{1 - \lambda\mathcal{G}'\{s - \lambda + \lambda\mathcal{B}(s)\}}. \tag{3.2.1}$$

Higher-order derivatives $\mathcal{B}^{(k)}(s)$ for $k = 2, \cdots$ can be obtained in a similar way. See Proposition 49. Thus, the empirical estimator of $\mathcal{B}'(s)$ is

$$\hat{\mathcal{B}}'(s) = \frac{\hat{\mathcal{G}}'\{s - \hat{\lambda} + \hat{\lambda}\hat{\mathcal{B}}(s)\}}{1 - \hat{\lambda}\hat{\mathcal{G}}'\{s - \hat{\lambda} + \hat{\lambda}\hat{\mathcal{B}}(s)\}}.$$

Suppose d, the (unique) positive root of $\mathcal{G}'(s) = \lambda^{-1}$, exists on $(-\infty, b)$, where b is the supremum of the convergence strip of $\mathcal{G}(s)$ (i.e., $d \leq b$). Then d determines the convergence strip of $\mathcal{B}(s)$.

Lemma 44. *Suppose $\rho = \lambda/\mu < 1$ and $(-\infty, b)$ is the convergence strip of $\mathcal{G}(s)$. If d, the (unique) positive root of the equation $\mathcal{G}'(s) = \lambda^{-1}$, exists on $(0, b)$, then, the Kendall-Takács functional equation (3.1.1) has a solution satisfying $\mathcal{B}(0) = 1$ on the domain of*

$(-\infty, d+\lambda-\lambda\mathcal{G}(d)]$ *only (or, the convergence strip of* $\mathcal{B}(s)$ *is* $(-\infty, d+\lambda-\lambda\mathcal{G}(d)]$*) and*

$$\mathcal{B}\{d+\lambda-\lambda\mathcal{G}(d)\} = \mathcal{G}(d). \tag{3.2.2}$$

Proof. The first step is to show that the Kendall-Takács functional equation (3.1.1) admits a solution $\mathcal{B}_I(s)$ iff $s \in (-\infty, c_1]$, with $c_1 = d+\lambda-\lambda\mathcal{G}(d)$, such that $\mathcal{B}_I(0) = 1$. The second step requires showing that $\mathcal{B}_I(s) = \mathcal{B}(s)$ for $s \in (-\infty, c_1]$.

For the first step, suppose there exists an $s > 0$ for which $\mathcal{B}_I(s) > 0$ is defined. Set $r = s - \lambda + \lambda\mathcal{B}_I(s)$ and rewrite the Kendall-Takács functional equation (3.1.1) as

$$\mathcal{G}(r) = \mathcal{B}_I(s) = \frac{r}{\lambda} + \left(1 - \frac{s}{\lambda}\right). \tag{3.2.3}$$

Since $\mathcal{G}(r)$ is convex, the line $r/\lambda + \{\mathcal{G}(r) - d/\lambda\}$ in r is tangent to $\mathcal{G}(s)$ at $s = d$ and parallel to the line in (3.2.3). Therefore, (3.2.3) admits one or two solutions iff $(1 - s/\lambda) \geq \mathcal{G}(d) - d/\lambda$, or $s \leq c_1$.

A value of $s > c_1$ contradicts to the assumption that $\mathcal{B}_I(s)$ is well-defined and is therefore outside of the convergence strip of $\mathcal{B}_I(\cdot)$. If $s < c$, then there are 2 solutions to (3.2.3) in r and $\mathcal{B}_I(s)$ is taken as the intersection with $r < d$. Note that if $s = 0$, then $r = 0$ and $\mathcal{B}_I(0) = 0/\lambda + (1 - 0/\lambda) = 1$ so the lower solution is the correct one to give a MGF. Thus, $\mathcal{B}_I(s)$ is well-defined for $s < c_1$. If s is in the boundary, (3.2.3) gives the relationship $\mathcal{B}_I(c_1) = \mathcal{G}(r)$. The implicit function theorem assures that $\mathcal{B}_I(s)$ is analytic for $\Re s < c_1$.

For the second step, we use an analytic continuation argument from complex variables. Since $\mathcal{B}(s) = \mathcal{B}_I(s)$ on $s \in (-\infty, 0]$ and both functions are analytic, the analytic continuation of $\mathcal{B}(s)$ to $\Re s < c_1$ must agree with $\mathcal{B}_I(s)$. Since $\mathcal{B}_I(s)$ is left continuous at $s = c$, then so also is $\mathcal{B}(s)$. Hence $(-\infty, c]$ must be the convergence strip of $\mathcal{B}(\cdot)$. □

Example 45. For M/M/1 queue, solving the equation,

$$\mathcal{G}'(s) = \frac{\mu}{(s-\mu)^2} = \frac{1}{\lambda},$$

we have

$$d=\sqrt{\mu}\left(\sqrt{\mu}-\sqrt{\lambda}\right),$$

(the other root of the equation, $\sqrt{\mu}(\sqrt{\mu}+\sqrt{\lambda})>\mu$ is outside of the domain of the saddlepoint equation for $\mathcal{G}$). Since

$$\mathcal{G}\left(d\right)=\mathcal{G}\{\sqrt{\mu}(\sqrt{\mu}-\sqrt{\lambda})\}=\sqrt{\frac{\mu}{\lambda}},$$

the domain of the saddlepoint equation for $\mathcal{B}\left(s\right)$ is

$$(-\infty,\sqrt{\mu}(\sqrt{\mu}-\sqrt{\lambda})+\lambda-\lambda\sqrt{\frac{\mu}{\lambda}}]=(-\infty,(\sqrt{\mu}-\sqrt{\lambda})^2],$$

which coincides with the previous result (3.1.2).

Remark 46. From now on, we will use $\mathcal{D}\left(s\right)=s+\lambda-\lambda\mathcal{G}\left(s\right)$ as in Lemma 2. If d exists on $(0,b)$, then the supremum of the convergence strip of $\mathcal{B}\left(s\right)$ is $\mathcal{D}\left(d\right)$. From Lemma 44,

$$\lim_{s\nearrow\mathcal{D}(d)}\mathcal{B}\left(s\right)=\mathcal{G}\left(d\right)<\infty,$$

which indicates that the busy period B has a heavy tail.

Note that $\mathcal{D}'\left(s\right)=1-\lambda\mathcal{G}'\left(s\right)$ is a strictly decreasing function with $\mathcal{D}'\left(0\right)=1-\lambda/\mu=1-\rho>0$. As with $\mathcal{W}\left(s\right)$, there is a easy way to check the existence of d. By the intermediate value theorem, $\mathcal{D}'\left(s\right)$ has a zero at d if and only if $\lim_{s\nearrow b}\mathcal{D}'\left(s\right)\leq 0$ and $b>0$.

Lemma 47. *The positive root d of $\mathcal{G}'\left(s\right)=\lambda^{-1}$ (or $\mathcal{D}'\left(s\right)=0$) exists if and only if $b>0$ and $\lim_{s\nearrow b}\mathcal{G}'\left(s\right)-\lambda^{-1}\geq 0$.*

It is clear that if d does not exist, then we need to consider how b will be related to the convergence strip of $\mathcal{B}\left(s\right)$. For this and simpler formulas regarding $\mathcal{B}^{(k)}\left(s\right)$, $k=1,2,3$, we will use the following function.

Lemma 48. *If we define a function $\mathcal{R}$ of s by*

$$\mathcal{R}\left(s\right)=s-\lambda+\lambda\mathcal{B}\left(s\right),$$

then $\mathcal{R}(s)$ is a strictly increasing continuous function.

Proof. We have

$$\mathcal{R}'(s) = 1 + \lambda\mathcal{B}'(s) > 1$$

since $\mathcal{B}'(s) \geq 0$. $\square$

Proposition 49. *Let $\mathcal{R}(s)$ be the function defined in Lemma 48 and let $\mathcal{K}(s) = \log\mathcal{B}(s)$. For any s, let $r = \mathcal{R}(s)$. Then, we have*

$$\begin{aligned}
\mathcal{B}'(s) &= \frac{\mathcal{G}'(r)}{1-\lambda\mathcal{G}'(r)} \\
\mathcal{B}''(s) &= \frac{\mathcal{G}''(r)}{\{1-\lambda\mathcal{G}'(r)\}^2} \\
\mathcal{B}'''(s) &= \frac{3\lambda\mathcal{G}''(r)^2 + \{1-\lambda\mathcal{G}'(r)\}\mathcal{G}'''(r)}{\{1-\lambda\mathcal{G}'(r)\}^3}
\end{aligned}$$

and

$$\begin{aligned}
\mathcal{K}'(s) =& \frac{\mathcal{G}'(r)}{\mathcal{G}(r)\{1-\lambda\mathcal{G}'(r)\}} \\
\mathcal{K}''(s) =& \frac{\mathcal{G}(r)\mathcal{G}''(r) - \mathcal{G}'(r)^2\{1-\lambda\mathcal{G}'(r)\}}{\mathcal{G}(r)^2\{1-\lambda\mathcal{G}'(r)\}^3} \\
\mathcal{K}'''(s) =& \frac{1}{\mathcal{G}(r)^3\{1-\lambda\mathcal{G}'(r)\}^5}[2\lambda^2\mathcal{G}'(r)^5 - 4\lambda\mathcal{G}'(r)^4 + 2\mathcal{G}'(r)^3 + 3\lambda\mathcal{G}(r)\mathcal{G}'(r)^2\mathcal{G}''(r) \\
& + \mathcal{G}(r)^2\{3\lambda\mathcal{G}''(r)^2 + \mathcal{G}'''(r)\} - \mathcal{G}(r)\mathcal{G}'(r)\{3\mathcal{G}''(r) + \lambda\mathcal{G}(r)\mathcal{G}'''(r)\}]
\end{aligned}$$

Proof. To get $\mathcal{B}(s)$, differentiate both sides of Kendall-Takács functional equation (3.1.1)

$$\mathcal{B}'(s) = \mathcal{G}'\{s-\lambda+\lambda\mathcal{B}(s)\}\{1+\lambda\mathcal{B}'(s)\}$$

and solve for $\mathcal{B}'(s)$ to get

$$\mathcal{B}'(s) = \frac{\mathcal{G}'\{s-\lambda+\lambda\mathcal{B}(s)\}}{1-\lambda\mathcal{G}'\{s-\lambda+\lambda\mathcal{B}(s)\}}. \tag{$*$}$$

Replacing $s-\lambda+\lambda\mathcal{B}(s)$ with r, we obtain the desired formula.

$\mathcal{B}''(s)$ can be obtained by differentiating both sides of $(*)$, and replacing $\mathcal{B}'(s)$ with the formula we obtained and using

$$\mathcal{B}(s) = \mathcal{G}(r)$$

we get the result for $\mathcal{B}''(s)$. $\mathcal{B}^{(3)}$ can be obtained similarly.

Using

$$\mathcal{K}'(s) = \frac{\mathcal{B}'(s)}{\mathcal{B}(s)} \qquad \mathcal{K}''(s) = \frac{\mathcal{B}(s)\mathcal{B}''(s) - \mathcal{B}'(s)^2}{\mathcal{B}(s)^2},$$

and so on, we obtain the results for $\mathcal{K}^{(k)}$, $k = 1, 2, 3$. □

Corollary 50. *Let $\mathcal{K}(s) = \log \mathcal{B}(s)$ and define $\mu'_k = \mathcal{G}^{(k)}(0)$ (i.e., $\mu'_1 = 1/\mu$ and $\lambda\mu_1 = \rho$. Note that μ'_k is the kth (non-central) moment of the service time distribution). Then,*

$$\begin{aligned}
\mathcal{K}'(0) =& \frac{\mu'_1}{(1-\lambda\mu'_1)} = \frac{\mu'_1}{(1-\rho)}, \\
\mathcal{K}''(0) =& \frac{\mu'_2 - {\mu'_1}^2(1-\lambda\mu'_1)}{(1-\lambda\mu'_1)^3} = \frac{\mu'_2 - {\mu'_1}^2(1-\rho)}{(1-\rho)^3}, \\
\mathcal{K}'''(0) =& \frac{2\lambda^2{\mu'_1}^5 - 4\lambda{\mu'_1}^4 + 2{\mu'_1}^3 + 3\lambda{\mu'_1}^2\mu'_2 - \mu'_1(3\mu'_2 + \lambda\mu'_3) + 3\lambda{\mu'_2}^2 + \mu'_3}{(1-\rho)^5}.
\end{aligned}$$

Now, we consider what happens if d does not exist. Using the above function, the Kendall-Takács functional equation can be written as

$$\mathcal{B}(s) = \mathcal{G}\left\{\mathcal{R}(s)\right\}. \tag{3.2.4}$$

If the convergence strip of $\mathcal{G}(s)$ is $(-\infty, b)$ for $b > 0$ then $\lim_{s\nearrow b}\mathcal{G}(s) = \infty$, which implies

$$\begin{aligned}
\lim_{s\nearrow b}\mathcal{G}'(s) = \lim_{s\nearrow b}\int_0^\infty xe^{xs}dG(x) \geq \lim_{s\nearrow b}\int_\varepsilon^\infty xe^{xs}dG(x) \\
\geq \lim_{s\nearrow b}\varepsilon\int_\varepsilon^\infty e^{xs}dG(x) = \infty.
\end{aligned}$$

Thus $d \in (0, b)$ exists as the root $\mathcal{G}'(d) = 1/\lambda > 1/\mu$. Therefore only for setting in which the convergence strip of $\mathcal{G}(s)$ is half open, or $(-\infty, b]$, can d potentially not exists.

Suppose such a setting in which $\mathcal{G}(s)$ converges on $(-\infty, b]$ for $b \geq 0$. Clearly, by (3.2.4),

$\mathcal{R}(s)$ cannot be greater than b. Because $\mathcal{G} \circ \mathcal{R}$ is is a strictly increasing function, the mapping $s \mapsto \mathcal{B}(s)$ in (3.2.4) is a bijection of

$$(-\infty, \mathcal{R}^{-1}(b)]$$

onto the range of $\mathcal{B}(\cdot)$ with boundary value

$$\mathcal{B}\{\mathcal{R}^{-1}(b)\} = \mathcal{G}(b) \tag{3.2.5}$$

in this case.

We now show the relationship between $\mathcal{R}(s)$ and $\mathcal{D}(s)$. Fix s in the convergence strip of $\mathcal{B}(s)$. We have

$$\begin{aligned} r = \mathcal{R}(s) = s - \lambda + \lambda \mathcal{B}(s) &\Longrightarrow r = s - \lambda + \lambda \mathcal{G}(r) \\ &\Longrightarrow r + \lambda - \lambda \mathcal{G}(r) = s \\ &\Longrightarrow \mathcal{D}(r) = s, \end{aligned} \tag{3.2.6}$$

which implies

$$\mathcal{D}\{\mathcal{R}(s)\} = s$$

or $\mathcal{D}$ is a left inverse of $\mathcal{R}$;

$$r = \mathcal{R}(s) \Longrightarrow \mathcal{D}(r) = s.$$

Generally, $\mathcal{D}(s)$ is not the inverse of $\mathcal{R}(s)$ because $\mathcal{D}(s)$ is decreasing on (d, b). Note that as a function of $\mathcal{G}(s)$, the domain of $\mathcal{D}(s)$ is the same as the convergence strip of $\mathcal{G}(s)$. By the same argument, the domain of $\mathcal{R}(s)$ is the convergence strip of $\mathcal{B}(s)$. To simplify the notation, we define

$$\langle d \rangle_b = \begin{cases} d & \text{if } d \text{ exists in } (0, b) \text{ for } b > 0, \\ b & \text{otherwise.} \end{cases} \tag{3.2.7}$$

Because $\mathcal{D}(s)$ has its maximum at d, $\mathcal{D}(s)$ is a 1-1, strictly increasing function on $(-\infty, \langle d \rangle_b]$.

Moreover, if d exists, $\mathcal{D}(d) = d + \lambda - \lambda\mathcal{G}(d)$, is the supremum of the convergence strip of $\mathcal{B}(s)$. Also, if d does not exist, $\mathcal{D}$ is a 1-1, strictly increasing function on $(-\infty, b]$. From (3.2.5) and (3.1.1), we have $b = \mathcal{R}^{-1}(b) - \lambda + \lambda\mathcal{G}(b)$, which implies

$$\mathcal{R}^{-1}(b) = b + \lambda - \lambda\mathcal{G}(b) = \mathcal{D}(b).$$

Thus, the domain of $\mathcal{R}$ is the same as the range of $\mathcal{D}$ and the restriction of $\mathcal{D}(s)$ on $(-\infty, \langle d\rangle_b]$ is the inverse function of $\mathcal{R}(s)$. If we define $\langle\mathcal{D}(d)\rangle_{\mathcal{D}(b)}$ in a similar way to (3.2.7), then the convergence strip of $\mathcal{B}(s)$(and the domain of $\mathcal{R}(s)$) is $(-\infty, \langle\mathcal{D}(d)\rangle_{\mathcal{D}(b)}]$.

Theorem 51. *On the convergence strip of $\mathcal{B}(s)$, $(-\infty, \langle\mathcal{D}(d)\rangle_{\mathcal{D}(b)}]$, we have $\mathcal{B}(s) = \mathcal{G}(r)$, where r is the (unique) root on $(-\infty, \langle d\rangle_b]$ of*

$$s = \mathcal{D}(r) = r + \lambda - \lambda\mathcal{G}(r),$$

or

$$\mathcal{B}(s) = \mathcal{G}\{\mathcal{D}|^{-1}_{(-\infty,\langle d\rangle_b]}(s)\} = \frac{\mathcal{D}|^{-1}_{(-\infty,\langle d\rangle_b]}(s) - s}{\lambda} + 1, \tag{3.2.8}$$

where $\mathcal{D}|_{(-\infty,\langle d\rangle_b]}(s)$ is the restriction of $\mathcal{D}$ on $(-\infty, \langle d\rangle_b]$.

Proof. We only need to show the last equality. From $\mathcal{D}(s) = s + \lambda - \lambda\mathcal{G}(s)$, we have

$$\mathcal{G}(s) = \frac{s - \mathcal{D}(s)}{\lambda} + 1,$$

so that

$$\mathcal{G}\{\mathcal{D}|^{-1}_{(-\infty,\langle d\rangle_b]}(s)\} = \frac{\mathcal{D}|^{-1}_{(-\infty,\langle d\rangle_b]}(s) - \mathcal{D}\{\mathcal{D}|^{-1}_{(-\infty,\langle d\rangle_b]}(s)\}}{\lambda} + 1$$

and we obtain the last equality. □

Remark 52. The author found later that the first equality of (3.2.8) was shown using martingale method in [56], as "a new explicit formula" without any consideration for the convergence strip of $\mathcal{B}(s)$. It was mentioned that $\mathcal{D}^{-1}(s)$ exists around 0 because $\mathcal{D}'(0) > 0$. The relationship between $\mathcal{R}(s)$ and $\mathcal{D}(s)$ (3.2.6) was also found in [12] (equation (32)).

3.2.2 Solvability of the saddlepoint equation

We are now ready to check the solvability of the saddlepoint equation. Following the discussion for the solvability of the saddlepoint equation for the Pollaczek-Khinchin formula, we check the infimum of the support of $\mathcal{B}$.

Lemma 53. *The infimum of the support for* $\mathcal{K}(s) = \log \mathcal{B}(s)$ *is the same as the infimum of support for* $\log \mathcal{G}(s)$, *or* $\inf\{t : F_B(t) = 0\} = g_0 := \inf\{t : G(t) > 0\}$

Proof. As a MGF, $\lim_{s \to -\infty} \mathcal{B}(s) = 0$, so that we have

$$\lim_{s \to -\infty} \{s - \lambda + \lambda \mathcal{B}(s)\} = -\infty.$$

Since $r : s \mapsto s - \lambda + \lambda \mathcal{B}(s)$ is a continuous function of s on $(-\infty, 0)$,

$$\begin{aligned} \lim_{s \to -\infty} \mathcal{K}(s)' = \lim_{s \to -\infty} \frac{\mathcal{B}'(s)}{\mathcal{B}(s)} &= \lim_{s \to -\infty} \frac{\mathcal{G}'\{s - \lambda + \lambda \mathcal{B}(s)\}}{1 - \lambda \mathcal{G}'\{s - \lambda + \lambda \mathcal{B}(s)\}} \left[\frac{1}{\mathcal{G}\{s - \lambda + \lambda \mathcal{B}(s)\}} \right] \\ &= \lim_{r \to -\infty} \left[\frac{\mathcal{G}'(r)}{\mathcal{G}(r)\{1 - \lambda \mathcal{G}'(r)\}} \right] = \lim_{r \to -\infty} \left\{ \frac{\mathcal{G}'(r)}{\mathcal{G}(r)} \right\} = \lim_{s \to -\infty} \{\log \mathcal{G}(s)\}', \end{aligned}$$

in which $\lim_{s \to -\infty} \mathcal{G}'(s) = 0$ is used and we obtain the desired result. □

Thus, as in the case of Pollaczek-Khinchin formula case, the infimum of s, where the empirical saddlepoint equation can be solved is $\min_i G_i$. We now investigate the steepness property of $\mathcal{B}(s)$.

Proposition 54. *Let b be the supremum of the convergence strip of $\mathcal{G}(s)$. Then, $\mathcal{B}(s)$ is steep if and only if one of the following holds:*

1. *d exists.*

2. *$b = 0$ and $EG = \infty$.*

Proof. It is clear that the existence of d implies the steepness of $\mathcal{B}(s)$. Suppose $b > 0$ and d does not exist, or $\lim_{s \nearrow b}\{1 - \lambda \mathcal{G}'(s)\} > 0$, which must be finite. Thus, $\mathcal{G}'(b)$ is well defined and finite and $\mathcal{G}(b) - 1 = \int_0^b \mathcal{G}'(s)\, ds$ is finite too. Thus,

$$\lim_{s \nearrow \mathcal{D}(b)} \mathcal{K}'(s) = \lim_{s \nearrow \mathcal{D}(b)} \frac{\mathcal{G}'(\mathcal{R}(s))}{\mathcal{G}(\mathcal{R}(s))\{1 - \lambda \mathcal{G}'(\mathcal{R}(s))\}} = \lim_{s \nearrow b} \frac{\mathcal{G}'(s)}{\mathcal{G}(s)(1 - \lambda \mathcal{G}'(s))} < \infty.$$

Thus, for $b > 0$, $\mathcal{K}(s)$ cannot be steep unless d exists.

When $b = 0$, the convergence strip of $\mathcal{B}(s)$ is $(-\infty, 0]$ because $\mathcal{D}(0) = 0$ and $\lim_{s \nearrow 0} \mathcal{K}'(s) = EG/(1-\rho)$ implies that $\mathcal{B}(s)$ is steep at 0 only if $EG = \infty$. □

Proposition 55. *If $\mathcal{B}(s)$ is steep, then the saddlepoint equation (3.2.9) can be solved for any $t \in (g_0, \infty)$ where $g_0 = \inf\{t : G(t) > 0\}$. If not, then the saddlepoint equation (3.2.9) can be solved for any $t \in (g_0, \mathcal{G}'(b)/[\mathcal{G}(b)\{1 - \lambda \mathcal{G}'(b)\}])$. The saddlepoint equation of the plug-in estimator (3.1.4), can be solved for any $t \in (\min G_i, \infty)$.*

To use the saddlepoint approximation, it is necessary to calculate $\mathcal{B}^{(k)}(s)$ $k = 0, 1, 2$. It may seem daunting at first since to solve the saddlepoint equation

$$\mathcal{K}'(s_t) = \frac{\mathcal{B}'(s_t)}{\mathcal{B}(s_t)} = t \tag{3.2.9}$$

numerically, any root finding algorithm must use a convergent sequence $\{s_j\} \to s_t$ and for each iteration, $\mathcal{B}(s_j)$ and $\mathcal{B}'(s_j)$ need to be calculated. This requires solving the Kendall-Takács functional equation, which requires another root finding algorithm. For example, suppose one needs on average n iterations (i.e., $s_1, s_2, \cdots, s_n$) for solving the saddlepoint equation and in each iteration, one needs to solve $s_j = \mathcal{D}(r)$ to calculate $\mathcal{B}(s_j)$ and $\mathcal{B}'(s_j)$. Assuming we need n iteration on average for the inside iterations (i.e., $r_{j1}, r_{j2}, \cdots, r_{jn}$), then a total of n^2 iterations are needed to solve the saddlepoint equation.

From Proposition 49, we can see that if we know r, then $\mathcal{B}(s)$ is not needed for $\mathcal{K}^{(k)}$, $k = 1, 2, 3$, which are the quantities needed for saddlepoint approximation. Of course, we need to find s_t for given t in the convergence strip of $\mathcal{B}(s)$ not just r. However, by the following theorem, we actually do not need at all to solve the Kendall-Takács functional equation (3.1.1) (,or equivalently $s_j = \mathcal{D}(r_j)$) to use the saddlepoint approximation.

Theorem 56. *For any t in the range of $\mathcal{K}'(s)$ described in Proposition 55, the solution of the saddlepoint equation s_t is given by*

$$s_t = r_t + \lambda - \lambda\mathcal{G}(r_t) = \mathcal{D}(r_t), \tag{3.2.10}$$

where r_t is the (unique) solution of the equation

$$t = \frac{\mathcal{G}'(r_t)}{\mathcal{G}(r_t)\{1 - \lambda\mathcal{G}'(r_t)\}}. \tag{3.2.11}$$

Thus, $\mathcal{B}(s_t) = \mathcal{G}(r_t)$ and $\mathcal{B}^{(k)}(s_t)$ and $\mathcal{K}^{(k)}(s_t)$ k=1,2,3 can be obtained by plugging r_t into argument r in the formulas of Proposition 49.

Proof. We can prove this in two ways. Let t be as in the assumption and s_t the unique solution of the saddlepoint equation $\mathcal{K}'(s_t) = t$. Define r_t by $r_t := \mathcal{R}(s_t)$, which is defined in Lemma 48. Because $\mathcal{R}$ is strictly increasing, there is no other s satisfying $\mathcal{R}(s_t) = r_t$, so that r_t is 1-1 with s_t. Using $\mathcal{B}(s_t) = \mathcal{G}(r_t)$, we have

$$r_t = \mathcal{R}(s_t) = s_t - \lambda + \lambda\mathcal{B}(s_t) = s_t - \lambda + \lambda\mathcal{G}(r_t),$$

which gives us equation (3.2.10).

Also, we can prove it in a more direct manner. Define a real valued function h by

$$h(x) = \frac{\mathcal{G}'(x)}{\mathcal{G}(x)\{1 - \lambda\mathcal{G}'(x)\}} \tag{3.2.12}$$

on the domain $(-\infty, d)$ if d exists and $(-\infty, b)$ if d does not exist. Then,

$$\mathcal{K}'(s) = h \circ \mathcal{R}(s).$$

Because $\mathcal{K}'$ and $\mathcal{R}$ are strictly increasing function, h is strictly increasing function too (Or, we may just use

$$h'(x) = \frac{\mathcal{G}'(x)^2\{1 - \lambda\mathcal{G}'(x)\} + \mathcal{G}(x)\mathcal{G}''(x)}{\mathcal{G}(x)^2\{1 - \lambda\mathcal{G}'(x)\}^2} > 0).$$

Since t is in the range of $\mathcal{K}'(s)$, it is also in the range of h, so that there exists unique r_t satisfying equation (3.2.11) and $\mathcal{R}^{-1}(r_t) = s_t$ satisfies the saddlepoint equation. □

Remark 57. If we use this $\mathcal{R}$ function to solve $h(r) = t$, the domain of h given by $(-\infty, \langle d \rangle_b]$ should be used. Note that using Theorem 56, we can draw the graph of the saddlepoint CDF and PDF approximation. For example, the saddlepoint CDF approximation can be written as a function of s_t, $\mathcal{K}^{(k)}(s_t)$, $k = 1, 2, 3$, which are all functions of r by Theorem 56 and Proposition 49. Thus, the graph can be drawn by the parametric curve $r \mapsto \{\mathcal{K}'(r), F_{B0}(r)\}$ for $r \in (-\infty, \langle d \rangle_b]$ without ever solving a saddlepoint equation.

3.2.3 Non-normal based saddlepoint density and CDF approximations

Theorem (44) implies that the CDF of the busy period will have a relatively heavy right tail. Experience shows us that the "usual" saddlepoint approximation performs poorly for the relatively heavy tailed distribution. In [73] and [14], it was shown that for the first passage distribution for a random walk, the inverse Gaussian (IG) based saddlepoint approximation works well. This is understandable because the IG is the distribution of the first passage time of Brownian motion with positive drift. A heuristic derivation of this fact can be seen in [70]. It is also shown in [5] that IG distribution approximates the CDF of the busy period of M/M/1 queue for large t.

For the derivation and other non-normal based saddlepoint approximations, see [73] and [17]. Following their approach, we set $\lambda = 1$, one of two parameters of IG, to make IG distribution considered a one parameter distribution. Then, the PDF and the CDF become

$$f_{\text{IG}}(x; \alpha) = \frac{1}{\sqrt{2\pi x^3}} \exp\left\{ \frac{-(x-\alpha)^2}{2\alpha^2 x} \right\}$$
$$F_{\text{IG}}(x; \alpha) = \Phi\left(\frac{\sqrt{x}}{\alpha} - \frac{1}{\sqrt{x}} \right) + \exp\left(\frac{2}{\alpha} \right) \Phi\left(-\frac{\sqrt{x}}{\alpha} - \frac{1}{\sqrt{x}} \right)$$

and its CGF becomes

$$L(\mathfrak{s}) = \frac{1}{\alpha} - \sqrt{\frac{1}{\alpha^2} - 2\mathfrak{s}} \quad \text{for } \mathfrak{s} \le \frac{1}{2\alpha^2}.$$

Let $F_{\text{IG}}(x;\alpha)$ be the CDF of IG. The IG based saddlepoint approximation of the cumulative distribution function for random variable B with its CGF $\mathcal{K}(s)$ is

$$F_{B0}(t) = \begin{cases} F_{\text{IG}}(z_t;\alpha_t) + f_{\text{IG}}(z_t;\alpha_t)\left(\frac{1}{\mathfrak{s}_t} - \frac{z_t^{3/2}}{u_t}\right) & \text{if } t \neq EB \\ F_{\text{IG}}\left\{\frac{\mathcal{K}'''(0)^2}{9\mathcal{K}''(0)^3}; \frac{\mathcal{K}'''(0)^2}{9\mathcal{K}''(0)^3}\right\} & \text{if } t = EB, \end{cases}$$

where

$$\begin{aligned} u_t &= s_t\sqrt{\mathcal{K}''(s_t)} \\ w_t &= \operatorname{sgn}(s_t)\sqrt{2\{s_t t - \mathcal{K}(s_t)\}} \\ \alpha_t &= \frac{\mathcal{K}'''(s_t)^2}{3\mathcal{K}''(s_t)^3}\left(3 + w_t\sqrt{\frac{\mathcal{K}'''(s_t)^2}{\mathcal{K}''(s_t)^3}}\right)^{-1} \qquad \text{for } \mathcal{K}'''(s_t) > 0 \\ z_t &= \alpha_t + \frac{\alpha_t^2}{2}\left(w_t^2 + w_t\sqrt{w_t^2 + \frac{4}{\alpha_t}}\right) \\ \mathfrak{s}_t &= \frac{1}{2}\left(\alpha_t^{-2} - z_t^{-2}\right) \end{aligned}$$

and s_t is the root of the usual saddlepoint equation

$$\mathcal{K}'(s_t) = t.$$

Note that the parameter of the base distribution α_t is not fixed and varies over t, which is different from the usual normal-based saddlepoint approximation case. Also, note that it is assumed that $\mathcal{K}'''(s_t) > 0$, which holds if B is a non-negative r.v. Finally, we note that non-normal base saddlepoint CDF approximation is invariant under affine transformation as the normal base saddlepoint CDF approximation is.

Note that for the saddlepoint density approximation, the inverse Gaussian (IG) based saddlepoint is given by

$$f_{B0}(t) = f_{\text{IG}}(z_t)\sqrt{\frac{L''(\mathfrak{s}_t)}{\mathcal{K}''(s_t)}} = f_{\text{IG}}(z_t)\sqrt{\frac{z_t^3}{\mathcal{K}''(s_t)}},$$

which is, in fact, exactly the same as the normal-based saddlepoint PDF approximation. In other words, it can be shown that

$$f_{\mathrm{IG}}\left(z_t\right)\sqrt{\frac{z_t^3}{\mathcal{K}''\left(s_t\right)}}=\frac{1}{\sqrt{2\pi\mathcal{K}''\left(s_t\right)}}\exp\{\mathcal{K}\left(s_t\right)-s_t t\},$$

which is the normal based saddlepoint PDF approximation. Therefore, for the PDF approximations, we will just use the normal- based saddlepoint approximation.

3.3 Convergence of the Plug-in estimators

Here we study the convergence properties of the plug-in estimators $D^k\hat{\mathcal{B}}\left(s\right)$ and $D^k\hat{\mathcal{K}}\left(s\right)$ for $k=0,1,2,3$. We consider the convergence properties of d first. It was shown in Lemma 7 that $\hat{d}\to d$ a.s. As for the case of $\hat{c}$, $\hat{d}$ is asymptotically normal.

Proposition 58. *$\hat{d}-d=o(n^{-\delta})$ a.s. for any $0<\delta<\min\{1/2,1-d/b\}$. If $d<b/2$, or $d=b/2$ and $\mathcal{G}\left(b\right)<\infty$, then*

$$\sqrt{n}\left(\hat{d}-d\right)=N\left[0,\frac{\mathcal{G}''\left(2s\right)-\mathcal{G}'\left(s\right)}{\left\{\mathcal{G}''\left(d\right)\right\}^2}\right].$$

Proof. As we mentioned in Chapter 1, $\sqrt{n}\{\hat{\mathcal{G}}'\left(s\right)-\mathcal{G}'\left(s\right)\}$ is asymptotically normal on $(-\infty,b/2)$ (and also on $b/2$ if $\mathcal{G}\left(b\right)<\infty$). For $d<b/2$ (or $d=b/2$ and $\mathcal{G}\left(b\right)<\infty$), by the mean value theorem,

$$\hat{d}-d=\frac{\hat{\mathcal{G}}'\left(\hat{d}\right)-\hat{\mathcal{G}}'\left(d\right)}{\hat{\mathcal{G}}''\left(d'\right)}=\frac{\lambda^{-1}-\hat{\mathcal{G}}'\left(d\right)}{\hat{\mathcal{G}}''\left(d'\right)}=\frac{\mathcal{G}'\left(d\right)-\hat{\mathcal{G}}'\left(d\right)}{\hat{\mathcal{G}}''\left(d'\right)}.$$

Because $\hat{\mathcal{G}}''\left(s\right)$ converges to $\mathcal{G}''\left(s\right)$ locally uniformly and $d'\to d$ a.s., $\hat{\mathcal{G}}''\left(d'\right)\to\mathcal{G}''\left(d\right)$. By Slutsky's lemma, we have

$$\sqrt{n}\left(\hat{d}-d\right)=\frac{-1}{\hat{\mathcal{G}}''\left(d'\right)}\sqrt{n}\left\{\mathcal{G}'\left(d\right)-\hat{\mathcal{G}}'\left(d\right)\right\}\Rightarrow N\left[0,\frac{\mathcal{G}''\left(2s\right)-\mathcal{G}'\left(s\right)}{\left\{\mathcal{G}''\left(d\right)\right\}^2}\right].$$

For $d>b/2$ and $0<\delta<\delta<\min\{1/2,1-d/b\}$, $\hat{d}-d=o\left(n^{-\delta}\right)$ is implied by the fact that

$\mathcal{G}'(d) - \hat{\mathcal{G}}'(d) = o\left(n^{-\delta}\right)$ and $\hat{\mathcal{G}}''(d') \to \mathcal{G}''(d) < 0$ a.s. □

Proposition 59. *If d does not exist, then $\hat{d} \to b$ a.s.*

Proof. As we showed in the proof of Lemma 47, $b > 0$ and $\lim_{s \nearrow b} \mathcal{G}'(s) - \lambda^{-1} \geq 0$. the nonexistence of c implies either $b = 0$ or $\mathcal{G}'(b) < \lambda^{-1}$ (if $b > 0$). Note that if $b = 0$, then $\mathcal{G}'(b) = \mathcal{G}'(0) = \mu^{-1} < \lambda^{-1}$ by the stability condition $\rho = \lambda/\mu < 1$. Suppose $b = 0$. Since $\mathcal{G}'(\varepsilon) = \infty$ for any $\varepsilon > 0$, we have $\hat{\mathcal{G}}'(\varepsilon) \nearrow \infty$ a.s. as $n \to \infty$. Thus, for all but finite n, $0 < \hat{d} < \varepsilon$. Because ε is arbitrary, we have $\hat{d} \to 0$ a.s.

Now suppose $b > 0$. Because $\hat{\mathcal{G}}'(b) \to \mathcal{G}'(b) < \lambda^{-1}$ a.s., $\hat{d} > b$ all but finite n. For any $\varepsilon > 0$, we have $\hat{\mathcal{G}}'(b + \varepsilon) \to \mathcal{G}'(b + \varepsilon) = \infty$, which implies $\hat{c} < b + \varepsilon$ eventually. Altogether, we conclude $b < \hat{d} < b + \varepsilon$ for all but finite n, which implies $\hat{d} \to b$ a.s. □

We now consider the convergence properties of $D^k \hat{\mathcal{B}}(s) = \hat{\mathcal{B}}^{(k)}(s)$ and $D^k \hat{\mathcal{K}}(s) = \hat{\mathcal{K}}^{(k)}(s)$. Note that we define $\hat{\mathcal{B}}(s) = \hat{\mathcal{G}}(\hat{r})$, where $\hat{r}$ is the (unique) root on $(-\infty, \langle \hat{d} \rangle_{\hat{b}}]$ of

$$s = \hat{\mathcal{D}}(\hat{r}) = \hat{r} + \hat{\lambda} - \hat{\lambda} \hat{\mathcal{G}}(\hat{r}).$$

$\hat{\mathcal{B}}^{(k)}(s)$ and $\hat{\mathcal{K}}^{(k)}(s)$ are defined accordingly too. Thus, we need to check the convergence properties of

$$\mathcal{D}|^{-1}_{(-\infty, \langle d \rangle_b]}(s) = \mathcal{R}|_{(-\infty, \langle \mathcal{D}(d) \rangle_{\mathcal{D}(b)}]}(s)$$

first. We showed that $\mathcal{D}(s)$ is strictly increasing on $(-\infty, \langle d \rangle_b]$. Because $\hat{\lambda} \to \lambda$ a.s. and $\hat{\mathcal{G}}(s) \to \mathcal{G}(s)$ locally uniformly a.s., we have that $\hat{\mathcal{D}}(s)$ converges to $\hat{\mathcal{D}}(s)$ locally uniformly a.s. In the proof of Theorem 9, we showed that $\hat{\mathcal{D}}(s) - \mathcal{D}(s) = o(n^{-\delta})$ a.s. for any $0 < \delta < \min\{1/2, 1 - s/b\}$ and if $s < b/2$, or $s = b/2$ and $\mathcal{G}(b) < \infty$, then

$$\sqrt{n}\left\{\hat{\mathcal{D}}(s) - \mathcal{D}(s)\right\} \Rightarrow N[0, \lambda^2 \{1 - 2\mathcal{G}(s) + \mathcal{G}(2s)\}].$$

In fact, for any $s_1 < s_2 < b/2$, $\sqrt{n}\{\hat{\mathcal{D}}(s) - \mathcal{D}(s)\}$ weakly converges to a Gaussian process on $C_{[s_1, s_2]}$ because $\sqrt{n}\{\hat{\mathcal{G}}(s) - \mathcal{G}(s)\}$ converges weakly to a Gaussian process on $C_{[s_2, s_2]}$ and $\hat{\lambda} \to \lambda$ a.s. by Slutsy's lemma.

Because $\hat{\mathcal{D}}(s) \to \mathcal{D}(s)$ pointwise, we can show $\hat{\mathcal{D}}^{-1}(s) \to \mathcal{D}^{-1}(s)$ pointwise on $(-\infty, \langle d \rangle_b]$. One way to show this is using a similar argument to the proof of Lemma 17 though other arguments without using the differentiability of $\hat{\mathcal{D}}$ is possible. See the proof of Theorem 63. With Theorem 37, we have $\hat{\mathcal{D}}^{-1}(s) \to \mathcal{D}^{-1}(s)$ locally uniformly on $(-\infty, \langle d \rangle_b]$. We state here formally as a proposition.

Proposition 60. *For any $s \in (-\infty, \langle \mathcal{D}(d) \rangle_{\mathcal{D}(b)}]$, $\hat{\mathcal{D}}^{-1}(s) \to \mathcal{D}^{-1}(s)$* locally uniformly a.s.

With Proposition 60, we have the following result.

Theorem 61. *For any $s \in (-\infty, \langle \mathcal{D}(d) \rangle_{\mathcal{D}(b)}]$ and $k \in \mathbb{N}_0$, $D^k\hat{\mathcal{B}}(s)$ and $D^k\hat{\mathcal{K}}(s)$ converges to $D^k\mathcal{B}(s)$ and $D^k\mathcal{K}(s)$ locally uniformly a.s., respectively.*

Proof. We start with the case of $D^k\hat{\mathcal{B}}(s)$ first. Invoking Theorem 37, we only need to show that for any s, $D^k\hat{\mathcal{B}}(s)$ converges to $D^k\mathcal{B}(s)$ a.s. because $D^k\hat{\mathcal{B}}(s)$ are all increasing function for any k. By Proposition 60, we have $\hat{\mathcal{D}}^{-1}(s) \to \mathcal{D}^{-1}(s)$. Because $\hat{\mathcal{G}}(s) \to \mathcal{G}(s)$ locally uniformly, $\hat{\mathcal{B}}(s) = \hat{\mathcal{G}}\{\hat{\mathcal{D}}^{-1}(s)\} \to \mathcal{G}\{\mathcal{D}^{-1}(s)\} = \mathcal{B}(s)$.

Similarly, we have $\hat{\mathcal{G}}^{(k)}\{\hat{\mathcal{D}}^{-1}(s)\} \to \mathcal{G}^{(k)}\{\mathcal{D}^{-1}(s)\}$ because $\hat{\mathcal{G}}^{(k)}(s) \to \mathcal{G}(s)$ locally uniformly a.s. From Proposition 49, we have

$$D^k\hat{\mathcal{B}}(s) = h_{1,k}\left\{\hat{\lambda}, \hat{\mathcal{G}}(r), \hat{\mathcal{G}}'(r), \cdots, \hat{\mathcal{G}}^{(k)}(r)\right\} / \{1 - \lambda\mathcal{G}'(r)\}^k\Big|_{r=\hat{\mathcal{D}}^{-1}(s)},$$

where $h_{1,k}(x_1, \cdots, x_{k+1})$ is a polynomial of $x_1, \cdots, x_{k+1}$. By the continuous mapping theorem, $D^k\hat{\mathcal{B}}(s) \to D^k\mathcal{B}(s)$, which complete the case of $D^k\hat{\mathcal{B}}(s)$.

For $D^k\hat{\mathcal{K}}(s)$, it is not obvious that $D^k\hat{\mathcal{K}}(s)$ is an increasing function for $k \geq 3$. Thus, instead of using Theorem 37, we observe that for each k,

$$D^k\hat{\mathcal{K}}(s) = h_{2,k}\left\{\hat{\lambda}, \hat{\mathcal{G}}(r), \hat{\mathcal{G}}'(r), \cdots, \hat{\mathcal{G}}^{(k)}(r)\right\} / \mathcal{G}(r)^k \left\{1 - \lambda\mathcal{G}'(r)\right\}^{2k+1}\Big|_{r=\hat{\mathcal{D}}^{-1}(s)},$$

where $h_{2,k}(x_1, \cdots, x_{k+1})$ is a polynomial of $x_1, \cdots, x_{k+1}$. Thus, $D^k\hat{\mathcal{K}}(s) \to D^k\mathcal{K}(s)$ pointwise a.s. on $(-\infty, \langle \mathcal{D}(d) \rangle_{\mathcal{D}(b)}]$. For the locally uniform convergence, it suffice to show $\hat{\mathcal{K}}(s)$ converges to $\mathcal{K}(s)$ continuously on $[s_1, s_2]$ for any $s_1 < s_2 \leq, \langle \mathcal{D}(d) \rangle_{\mathcal{D}(b)}$ (see theorem 7.3.5

of [66] or §0.0 of [53]). If $s_n \to s$, then $\hat{\mathcal{G}}_n^{(k)}\{\hat{\mathcal{D}}_n^{-1}(s_n)\} \to \mathcal{G}^{(k)}\{\mathcal{D}^{-1}(s)\}$ as before and by the continuous mapping theorem, $D^k\hat{\mathcal{K}}_n(s_n) \to D^k\mathcal{K}(s)$ under the function $h_{2,k}$ and we complete the proof. □

Thus, as we shown in chapter 2, the previous result give us $\hat{B} \Rightarrow B$ or

Proposition 62. *$\hat{F}_B(s) \to F_B(s)$. Moreover, if the service time distribution G is continuous, then $\hat{F}_B$ converges uniformly.*

Proof. If G is continuous, so is F_B by (3.1.3). The theorem known as Pólya's lemma (Exercise 7.2 of [61]. See also Exercise 7.13 of [58], which can be proved by Theorem 37) gives

$$\sup_{x\geq 0}\left|\hat{F}_B(x) - F_B(x)\right| \to 0 \quad \text{as } n \to \infty.$$

□

We now have the following theorem regarding the convergence of $\hat{F}_{B0}(t)$.

Theorem 63. *$\hat{F}_{B0}(t)$ converges to $F_{B0}(t)$ locally uniformly a.s.*

Proof. By the definition of $\hat{F}_{B0}(t)$, we may write

$$\hat{F}_{B0}(t) = H\left\{\hat{\mathcal{K}}(r_t), t, \hat{\mathcal{K}}'(r_t), \hat{\mathcal{K}}''(r_t), \hat{\mathcal{K}}'''(r_t)\right\},$$

where $H : \mathbb{R}^5 \to \mathbb{R}$ is a continuous function.

If we define h as in (3.2.12) and define

$$\hat{h}(x) = \frac{\hat{\mathcal{G}}'(x)}{\hat{\mathcal{G}}(x)\{1 - \lambda\hat{\mathcal{G}}'(x)\}},$$

then we have

$$\hat{r}_t = \hat{h}^{-1}(t).$$

One can show that for each t, $\hat{h}^{-1}(t) \to h^{-1}(t)$ as in proof of Lemma 17. Another way to show without using the differentiability of h but using the monotonicity of h is as follows.

We claim that $\hat{h}_n(x_n) \to h(y)$ implies $x_n \to y$. Then setting $x_n = \hat{h}_n^{-1}(x)$ and $y = h^{-1}(x)$ we get the desired result. Suppose that $\hat{h}_n(x_n) \to h(y)$ but $x_n \not\to y$. The latter is equivalent to that there exist $\varepsilon > 0$ and an infinite sequence n' such that $|x_{n'} - y| > \varepsilon$. So, we have $\hat{h}_{n'}(x_{n'}) \to h(y)$ but $|x_{n'} - y| > \varepsilon$. Note that the latter implies

$$\begin{aligned}\left|\hat{h}_{n'}(x_{n'}) - \hat{h}_{n'}(y)\right| &\geq \min\left\{\left|\hat{h}_{n'}(y-\varepsilon) - \hat{h}_{n'}(y)\right|, \left|\hat{h}_{n'}(y+\varepsilon) - \hat{h}_{n'}(y)\right|\right\} \\ &\to \min\left\{\left|h(y-\varepsilon) - h(y)\right|, \left|h(y+\varepsilon) - h(y)\right|\right\} > 0\end{aligned}$$

though $|\hat{h}_{n'}(x_{n'}) - h_{n'}(y)| \to 0$. Thus, we get the contradiction.

By Theorem 37, we have $\hat{h}^{-1}(t) \to h^{-1}(t)$ locally uniformly on $(-\infty, \langle d\rangle_b]$ and pointwise on $(-\infty, \langle\mathcal{D}(d)\rangle_{\mathcal{D}(b)}]$. Thus, $\hat{F}_{B0}(t) \to F_{B0}(t)$ pointwise on $(-\infty, \langle\mathcal{D}(d)\rangle_{\mathcal{D}(b)}]$ by the continuous mapping theorem. For the locally uniform convergence, if $t_n \to t$, then we have

$$\hat{h}_n^{-1}(t_n) \to h^{-1}(t)$$

and

$$\hat{\mathcal{K}}_n^{(j)}\left\{\hat{h}_n^{-1}(t_n)\right\} \to \mathcal{K}^{(j)}\left\{h^{-1}(t)\right\} \quad \text{for } j = 1, 2, 3.$$

By the continuous mapping theorem, we have

$$\hat{F}_{B0}(t_n) \to F_{B0}(t),$$

which completes the proof. □

3.4 Estimation of the CDF and the PDF of busy periods

In this section, we show how the saddlepoint approximation can be used to estimate the PDF and CDF of the busy time periods of M/G/1 queues as we did in the previous chapter. Our choice of the service time distributions for the study is the same as before. Because the MGF's of exponential distribution and gamma distribution have closed form, we are able to calculate $F_{B0}(t)$ and $f_{B0}(t)$ explicitly.

3.4.1 True CDF and PDF for M/M/1 queue and M/E_k/1 queue

Note that for M/M/1 queues, the PDF of the busy time period has a explicit form: Let modified Bessel functions (or the hyperbolic Bessel functions) $I_\alpha(x)$ be defined by

$$I_\alpha(x) = \sum_{k=0}^{\infty} \frac{1}{k!\Gamma(k+\alpha+1)} \left(\frac{x}{2}\right)^{2k+\alpha},$$

which is $i^{-\alpha} J_\alpha(ix)$, where

$$J_\alpha(x) = \sum_{k=0}^{\infty} \frac{(-1)^k}{k!\Gamma(k+\alpha+1)} \left(\frac{x}{2}\right)^{2k+\alpha}$$

is "ordinary" Bessel function of the first kind. It is known (p. 474 and p. 483 of [30]) that for M/M/1 case, the density of the busy period is

$$\sqrt{\frac{\mu}{\lambda}} \frac{\exp\{-(\mu+\lambda)t\}}{t} I_1\left(2\sqrt{\lambda\mu}t\right). \tag{3.4.1}$$

One can obtain the CDF of the busy period through the numerical integration of (3.4.1) or the following method:

A gamma distribution service time is one of the case that (3.1.3) can be used to obtain an approximation of the CDF; Using $\sum_{i=1}^{j} G_i \sim \text{Gamma}(\alpha, jk)$ if $G_i \sim \text{Gamma}(\alpha, k)$,

$$\begin{aligned} f_B(t) &= \sum_{j=0}^{\infty} e^{-\lambda t} \frac{(\lambda t)^j}{(j+1)!} dG^{(j+1)*}(t) = \sum_{j=0}^{\infty} e^{-\lambda t} \frac{(\lambda t)^j}{(j+1)!} \left[\frac{\alpha^{k(j+1)} t^{k(j+1)-1}}{\Gamma\{k(j+1)\}} e^{-\alpha t} \right] \\ &= \sum_{j=0}^{\infty} e^{-(\lambda+\alpha)t} \frac{\lambda^j \alpha^{k(j+1)} t^{kj+k+j-1}}{(j+1)!\Gamma\{k(j+1)\}}. \end{aligned} \tag{3.4.2}$$

Let $\gamma(s,t) := \int_0^t x^{s-1} e^{-x} dx$, the lower incomplete Gamma function. Using

$$\int_0^t x^{s-1} e^{-ax} dx = \int_0^{at} \frac{1}{a} \left(\frac{x}{a}\right)^{s-1} e^{-x} dx = a^{-s} \gamma(s, at),$$

we have

$$F_B\left(t\right) = \int_0^t f_B\left(x\right) dx = \sum_{j=0}^{\infty} \frac{\lambda^j \alpha^{k(j+1)}}{(j+1)!\Gamma\{k(j+1)\}} \int_0^t e^{-(\lambda+\alpha)x} x^{kj+k+j-1} dx$$

$$= \sum_{j=0}^{\infty} \frac{\lambda^j \alpha^{k(j+1)}}{(j+1)!\Gamma\{k(j+1)\}\left(\lambda+\alpha\right)^{kj+k+j}} \gamma\{kj+k+j, (\lambda+\alpha)t\}. \qquad (3.4.3)$$

Thus, if G~Gamma(3,3) and I ~Exp(1/2), then we have

$$F_B\left(t\right) = \sum_{j=0}^{\infty} \frac{216^{j+1}}{(j+1)!\Gamma\{3(j+1)\}7^{4j+3}} \gamma(4j+3, 7t/2),$$

and using the finite summation of j up to n, we obtain the approximation. The case of G~Exp(2) can be done in similar way.

For Gamma(3,3) and Exp(2) service time distribution, the approximation of $F_B\left(t\right)$ was obtained by setting $n = 42$ using R because the largest integer allowed for the argument in the Gamma function in R was $172 (= 4 \cdot 42 + 3)$.

3.4.2 Simulation method

Let I_j ~Exp(λ) be inter-arrival times and G_j be the service time, $j \in \mathbb{N}$. Suppose at time $t = 0$, there are no customers in the queue. Then, the server becomes busy when the first customer arrives, which is at time I_1. When the server finishes serving the first customer, (i.e., at time $I_1 + G_1$) if there is no customer in the queue (or the second customer did not arrive in between $I_1 + G_1$ and I_1, which means $I_1 + I_2 > I_1 + G_1$), then the server is off and the busy time period is G_1 . If not and there is no customer in the queue at the time $I_1 + G_1 + G_2$ (,or the third customer did not arrive in between $I_1 + I_2$ and $I_1 + G_2 + G_2$,), then the busy time period is $G_1 + G_2$ and so on. Thus,

$$B \sim \sum_{j=1}^{N} G_j, \text{ where } N = \min\left\{m \in \mathbb{N} : \sum_{1}^{m+1} I_j > I_1 + \sum_{1}^{m} G_j\right\}.$$

Note that I_j are identically distributed so that

$$N = \min\left\{ m \in \mathbb{N} : \sum_2^{m+1} I_j > \sum_1^m G_j \right\} \sim \min\left\{ m \in \mathbb{N} : \sum_1^m I_j > \sum_1^m G_j \right\}.$$
$$= \min\left\{ m \in \mathbb{N} : \sum_1^m (I_j - G_j) > 0 \right\}$$

Thus, the following algorithm gives us a random sample $\{B_j\}$:

1. Generate random vectors $(I_1, I_2, \cdots, I_l)$ of the inter-arrival times and $(G_1, G_2, \cdots, G_l)$ of the service times, where l is a fixed number.

2. Calculate the cumulative sums

$$\left(I_1, \sum_1^2 I_j, \sum_1^3 I_j, \cdots, \sum_1^l I_j \right) \quad \text{and} \quad \left(G_1, \sum_1^2 G_j, \sum_1^3 G_j, \cdots, \sum_1^l G_j \right).$$

3. Find the minimum N such that

$$\sum_j^N G_j < \sum_{j=1}^N I_j$$

and set

$$B_1 = \sum_{j=1}^N G_j.$$

If $\sum_j^n G_j > \sum_{j=1}^n I_j$ for all $n = 1, 2, \cdots, l$, generate more random vectors of I_j and G_j to get $(I_1, \sum_1^2 I_j, \sum_1^3 I_j, \cdots, \sum_1^{2l} I_j)$ and $(G_1, \sum_1^2 G_j, \sum_1^3 G_j, \cdots, \sum_1^{2l} G_l)$ until obtaining N.

4. Repeat the step 1 through 3 to obtain $B_2, \cdots, B_m$.

3.4.3 The saddlepoint approximations from the true MGF and the approximated MGF

See Figure 3.4.1 for the saddlepoint PDF and CDF estimation of the busy time periods with their the % relative errors when the service time distributions are Exp(2) and Gamma(3, 3).

The arrival rate λ is set to 1 and 2 respectively. For M/M/1 case, the (true) CDF is calculated by numerical integration of the true PDF (3.4.1) and F_{Φ} denotes the normal based CDF saddlepoint approximation. For Gamma(3,3) service time distribution, the approximated PDF and CDF were obtained using (3.4.2) and (3.4.3). Also the empirical CDF, $F_B^{\dagger}$ and the kernel density estimation $f_B^{\dagger}$ were obtained from the random sample B of size $3 \cdot 10^7$ to check their percentage relative errors.

For Beta(2,2) and Pareto(.8,5) service distribution cases, the true CDFs were approximated from the average of 50 $F_B^{\dagger}(t)$, each of which is calculated from a random sample of B size $m = 3 \cdot 10^7$. For beta and Pareto distribution, the empirical MGF of the service time distribution from $\left(10^5 - 1\right)$ quantile points of

$$\left\{F^{-1}\left(i * 10^{-5}\right) : 1 \leq i \leq 10^{-5} - 1\right\}$$

are used to obtain $\tilde{F}_{B0}$ and $\tilde{f}_{B0}$, which are approximations of F_0 and f_0, respectively. This approximation was used in [62] to compute the waiting time distribution by numerical inversion of the Laplace transform. It is clear that $\tilde{F}_{B0}(t)$ works in the tail area. The t-axes are cutoff to include 99.5 percentile of B and we checked the relative errors of $\tilde{F}_{B0}(t)$ up to 99.95 percentile of B, which are still close to 0. Thus, at least for our cases, when the MGF of G is not steep (Pareto distribution), one may obtain the asymptotic result by using the approximated MGF of G with the saddlepoint approximation.

We note that the irregularity of the tail area of % relative error for the $\hat{f}_0(x)$ of Exp(2) and Gamma(3,3) service times indicates that the kernel density estimation is not smooth for that tail area and suggests to use different bandwidths for that area or "transform."

3.4.4 Confidence Band of the CDF of B

As we did in Chapter 1, we can build the confidence band using $\{\hat{F}_{B0}^{*}(t)\}$. We do simulation studies using the same service time distributions as in Chapter 1. Because the numerical routine to calculate the saddlepoint approximation $\hat{F}_{B0}$ is substantially slower than the calculation of $\hat{F}_0$, the number of sample for the simulation, l is set to 1000. Also, the

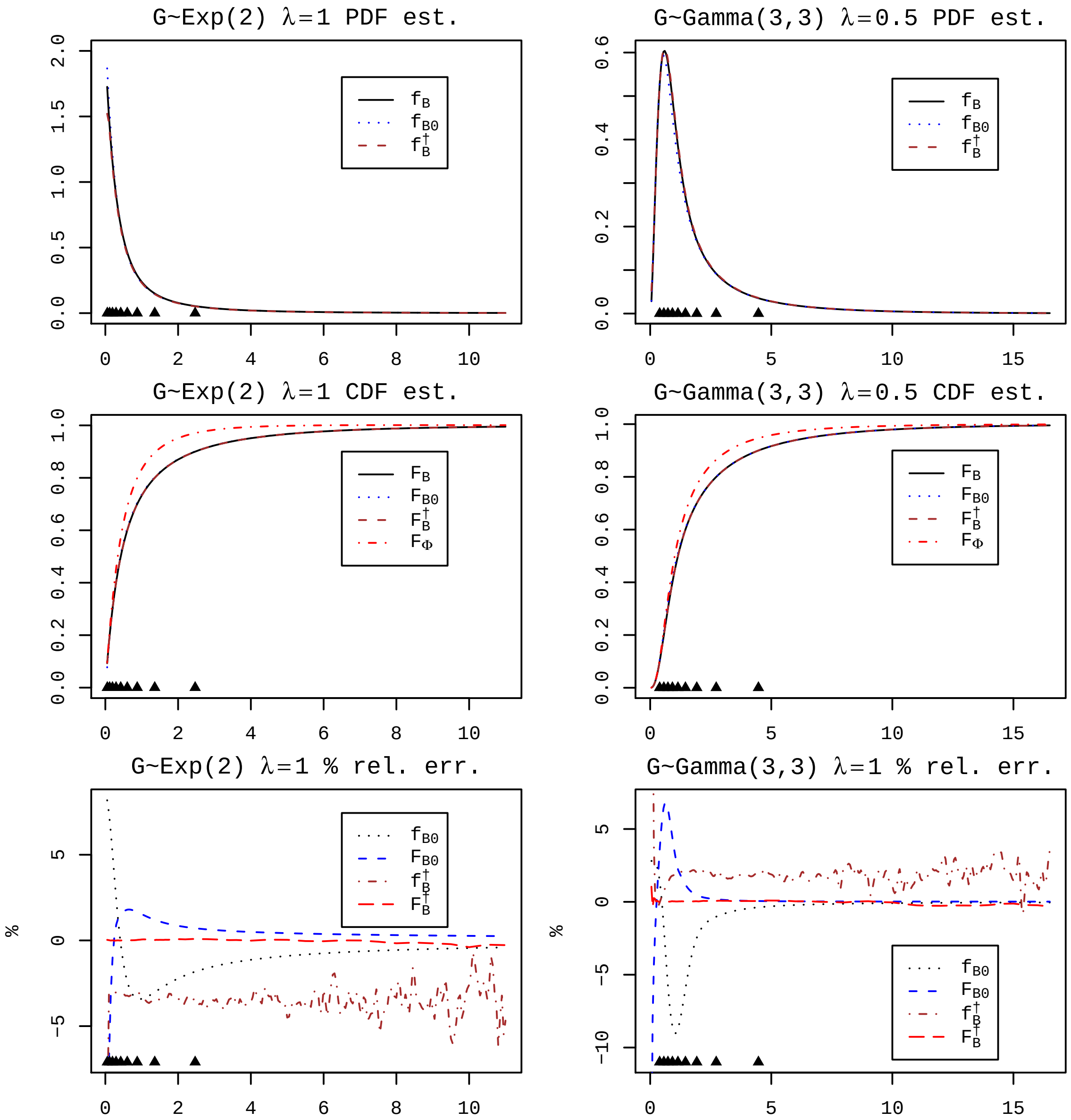

Figure 3.4.1: The percentage relative errors of F_{B0} and f_{B0} against true CDF and PDF (left: G~Exp(2)) and empirical CDF, $F_B^\dagger$ and the kernel density $f_B^\dagger(t)$ (right: G~Gamma(3,3)). In each graph, ▲ denotes a decile (form 10% to 90%) of the distribution of the busy periods. Top: Saddlepoint (unnormalized) PDF approximations with the histogram of the busy period random sample of the size 3×10^7 for G ~Gamma(3,3). Middle: The inverse Gaussian distribution based saddlepoint CDF approximations F_0 and the normal based saddlepoint CDF approximation F_Φ. Bottom: The percentage relative errors of PDF (dotted) and CDF (dashed) estimations from the top and the middle graphs. t-axes are cutoff to include up to 99.5 percentile of B.

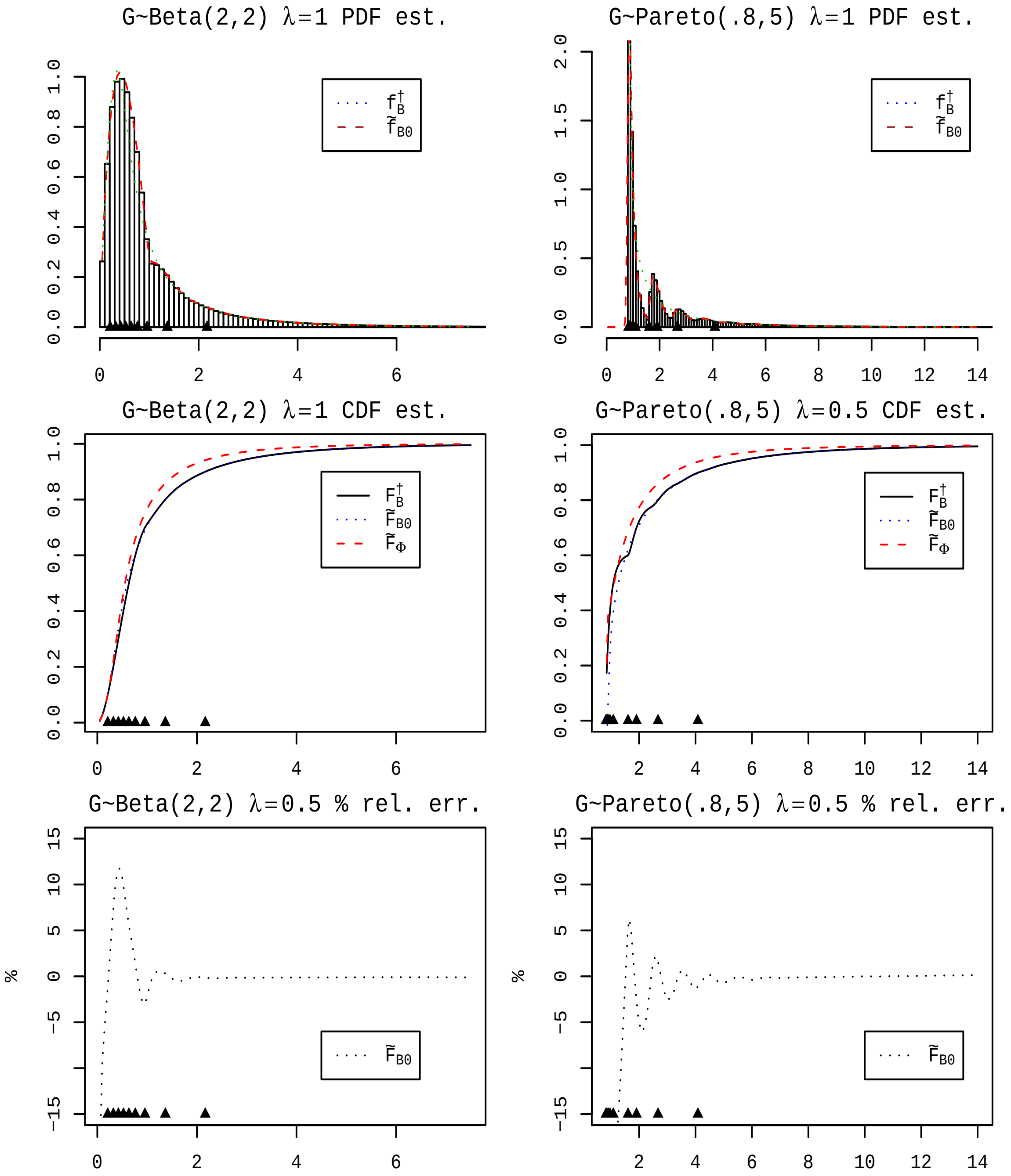

Figure 3.4.2: Similar to Figure 3.4.1 with beta and Pareto service time distribution.

bootstrap sample size $B\#$ is set to $5 \cdot 10^3$ for $n = 50$. See Figure 3.4.3 and 3.4.4 for an example, the average coverage probabilities, and the average interval lengths of each service time distributions. As in the case of the confidence band of $F(t)$, there is no clear winner in our simulation result. For example, the HDR method does not work well for $G\sim$Exp(2) and Gamma(3,3) cases but works well for beta and Pareto service time distributions. The estimated coverage probabilities and the average lengths of BP and BCa method are very close to each other.

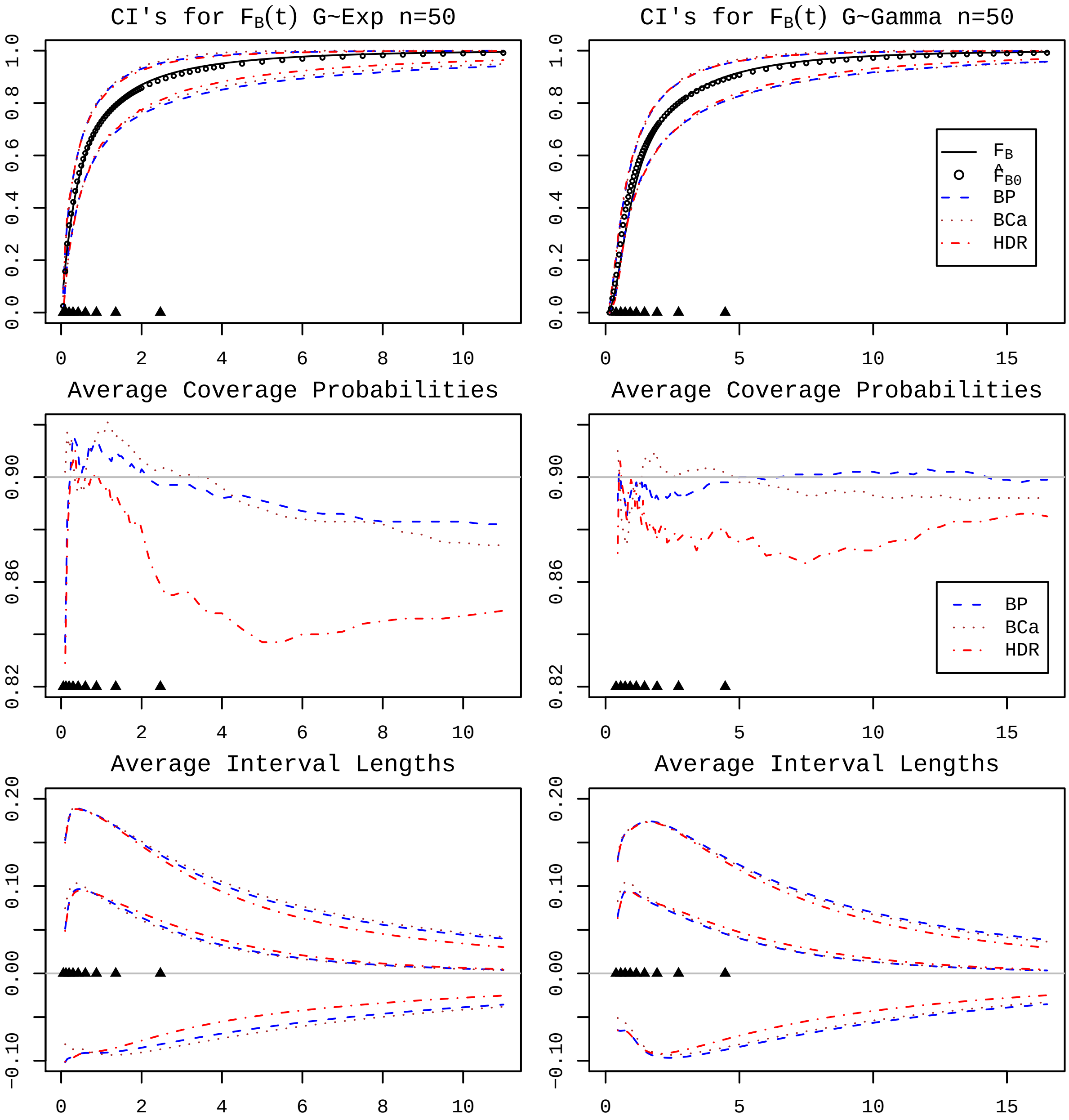

Figure 3.4.3: CI's for $F_B(t)$ where G~Exp(2) ($\lambda = 1$) and G~Gamma(3,3) ($\lambda = .5$) Top: Calculated CI's for a random sample of $\{G_j, I_j\}_{j=1}^{n}$ for n=50. Middle: the average coverage probabilities from l=1000 random samples of $\{G_j, I_j\}_{j=1}^{n}$ for n=50. Bottom: Average interval lengths of each CI's. The top curves are the interval lengths, $U(t) - L(t)$ and the middle curves are of $U(t) - F_B(t)$ and bottom curves are of $L(t) - F_B(t)$. In each graph, ▲ denotes the deciles (from 10% to 90%) of B, the busy period. t-axes are cutoff to include up to 99.5 percentile of B.

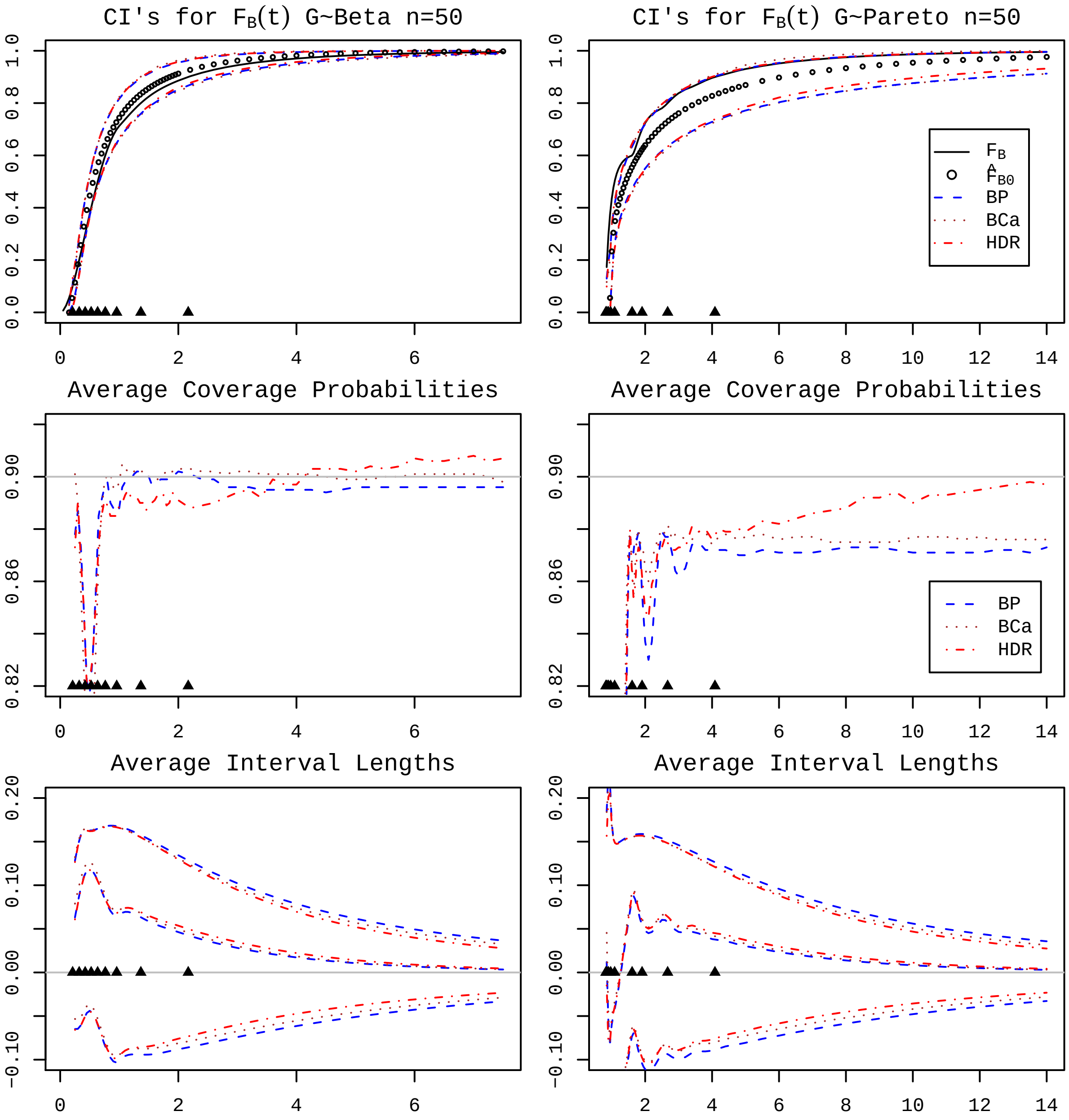

Figure 3.4.4: Same as Figure 1.3.9 for $G\sim$Beta$(2,2)$ and $G\sim$Pareto$(.8,5)$.

Chapter 4

Estimation of EW, $\operatorname{Var} W$, EB, and $\operatorname{Var} B$

4.1 Estimates of the mean and the variance of stationary waiting times

In this chapter, we compare the performance of the different bootstrap confidence intervals for EW, $\operatorname{Var} W$, EB, and $\operatorname{Var} W$ and propose a new bootstrap modified percentile CI. We show that the proposed method yields better coverage probabilities than standard bootstrap CI's.

4.1.1 Moment estimators and its features

Let $\mu_k' := EG^k$ (i.e., $\mu_1' = 1/\mu$ and $\lambda\mu_1' = \rho$. Note that μ_k' is the kth (non-central) moment of the service time distribution). Using $EW^k = \lim_{s\to 0} \mathcal{W}^{(k)}(s)$, where $\mathcal{W}^{(k)}(s)$ is the kth derivative of $\mathcal{W}(s)$ with Lemma 23, we obtain

$$\begin{aligned} EW =& \mu_1' + \frac{\lambda \mu_2'}{2(1-\rho)} = \mu_1' + \frac{\mu_2'}{2(\lambda^{-1} - \mu_1')}. \\ \operatorname{Var}(W) =& \left\{ \frac{\lambda \mu_2'}{2(1-\rho)} \right\}^2 + \frac{\lambda \mu_3'}{3(1-\rho)} + \operatorname{Var} G \\ =& \left\{ \frac{\mu_2'}{2(\lambda^{-1} - \mu_1')} \right\}^2 + \frac{\mu_3'}{3(\lambda^{-1} - \mu_1')} + \{\mu_2' - {\mu_1'}^2\}. \end{aligned}$$

By replacing μ_k', the kth moment of the service time distribution, with its kth sample moment $\hat{\mathcal{G}}^{(k)}(0) := n^{-1} \sum_{j=1}^{n} G_j^k$ and λ^{-1} with $\overline{I}$, where the inter-arrival time, $I_j \overset{\text{iid}}{\sim} \operatorname{Exp}(\lambda)$ (note that $\overline{I}$ is also the maximum likelihood estimator of λ), we obtain the method of moments (MOM) estimators for EW as

$$\begin{aligned} \widehat{EW} =& \hat{\mathcal{G}}'(0) + \frac{\hat{\mathcal{G}}''(0)}{2\{\overline{I} - \hat{\mathcal{G}}'(0)\}}, \\ \widehat{\operatorname{Var}(W)} =& \left\{ \frac{\hat{\mathcal{G}}''(0)}{2\left(\overline{I} - \hat{\mathcal{G}}'(0)\right)} \right\}^2 + \frac{\hat{\mathcal{G}}'''(0)}{3\{\overline{I} - \hat{\mathcal{G}}'(0)\}} + \{\hat{\mathcal{G}}''(0) - \hat{\mathcal{G}}'(0)^2\}. \end{aligned}$$

By Lemma 12, we know

$$\begin{aligned} \widehat{EW} &\xrightarrow{\text{a.s.}} EW \\ \widehat{\operatorname{Var}(W)} &\xrightarrow{\text{a.s.}} \operatorname{Var}(W), \end{aligned}$$

so that the MOM estimators are strongly consistent. Theorem 16 implies the following corollary:

Corollary 64. *With the same assumption as in Theorem 16, we obtain*

$$\begin{aligned} \sqrt{n}(\widehat{EW} - EW) &\Rightarrow N\left(0, \sigma_{EW}^2\right), \\ \sqrt{n}(\widehat{\operatorname{Var} W} - \operatorname{Var} W) &\Rightarrow N\left(0, \sigma_{\operatorname{Var} W}^2\right). \end{aligned}$$

σ^2_{EW} and $\sigma^2_{\mathrm{Var}\,W}$ can be calculated using Theorem 16. For example,

$$\sigma^2_{EW} = \left\{ \frac{\mu_2'\lambda}{2\,(1-\rho)^2} \right\}^2 + B\{CA + (\mu_2' - {\mu_1'}^2)B\} + A\{(\mu_4' - {\mu_2'}^2)A - CB\}$$

with

$$A = \frac{\lambda}{2\,(1-\rho)}, \qquad B = 1 + \frac{\mu_2'\lambda^2}{2\,(1-\rho)^2}, \qquad C = \mu_3' - \mu_1'\mu_2'.$$

The above corollary gives us approximate $(1-\alpha)$ confidence intervals,

$$\widehat{EW} \pm z_{\alpha/2}\frac{\sigma_{EW}}{\sqrt{n}}, \quad \text{and} \quad \widehat{\mathrm{Var}\,(W)} \pm z_{\alpha/2}\frac{\sigma_{\mathrm{Var}\,W}}{\sqrt{n}}. \tag{4.1.1}$$

However, as ρ approaches to 1,

$$\sigma_{EW} = O((1-\rho)^{-2}), \quad \text{and,} \quad \sigma_{\mathrm{Var}\,W} = O\left((1-\rho)^{-3}\right),$$

which will result in intervals that are too wide to be of practical sue when ρ is very close to 1. This has been confirmed by our simulation study.

Figure 4.1.1 shows the histograms of 10^5 random sample of $\sqrt{n}(\widehat{EW} - EW)$ for $M/M/1$ queue with $\mu = 1$ and $\lambda = .1$, .5, and .9 from the top to the bottom (thus, the corresponding ρ's are .1, .5, and .9 respectively) and $n = 25$ and 100 from the left to the right. Note that if the calculated $\hat{\rho} = \overline{G}/\overline{X} > 1$, the sample $\{X_1, \cdots, X_n, G_1, \cdots, G_n\}$ is discarded and the random samples we use always satisfy the condition $\hat{\rho} < 1$.

The overlapping smooth curves are the density curves of the limiting distributions $N(0, \sigma^2_{EW})$. Also note that for $\rho = .5$ only up to 93.61% ($n = 25$) and 82.58% ($n = 100$) of the samples are shown in the histograms due to the relatively large ranges. For $\rho = .9$, only 95.86% ($n = 25$) and 87.23% ($n = 100$) of the random sample of size 10^5 are shown. See Figure 4.1.2 for their boxplots using log transform.

For $\rho = .5$ and $\rho = .9$, the feature of a heavy right-hand tail is even more clear from Figure 4.1.3 giving QQ-plots of $\widehat{EW}$ against its limiting distribution $N(EW, \sigma^2_{EW}/n)$. The plots show that $\widehat{EW}$ is biased. Also, the observed heavy outliers in Figure 4.1.2 have a

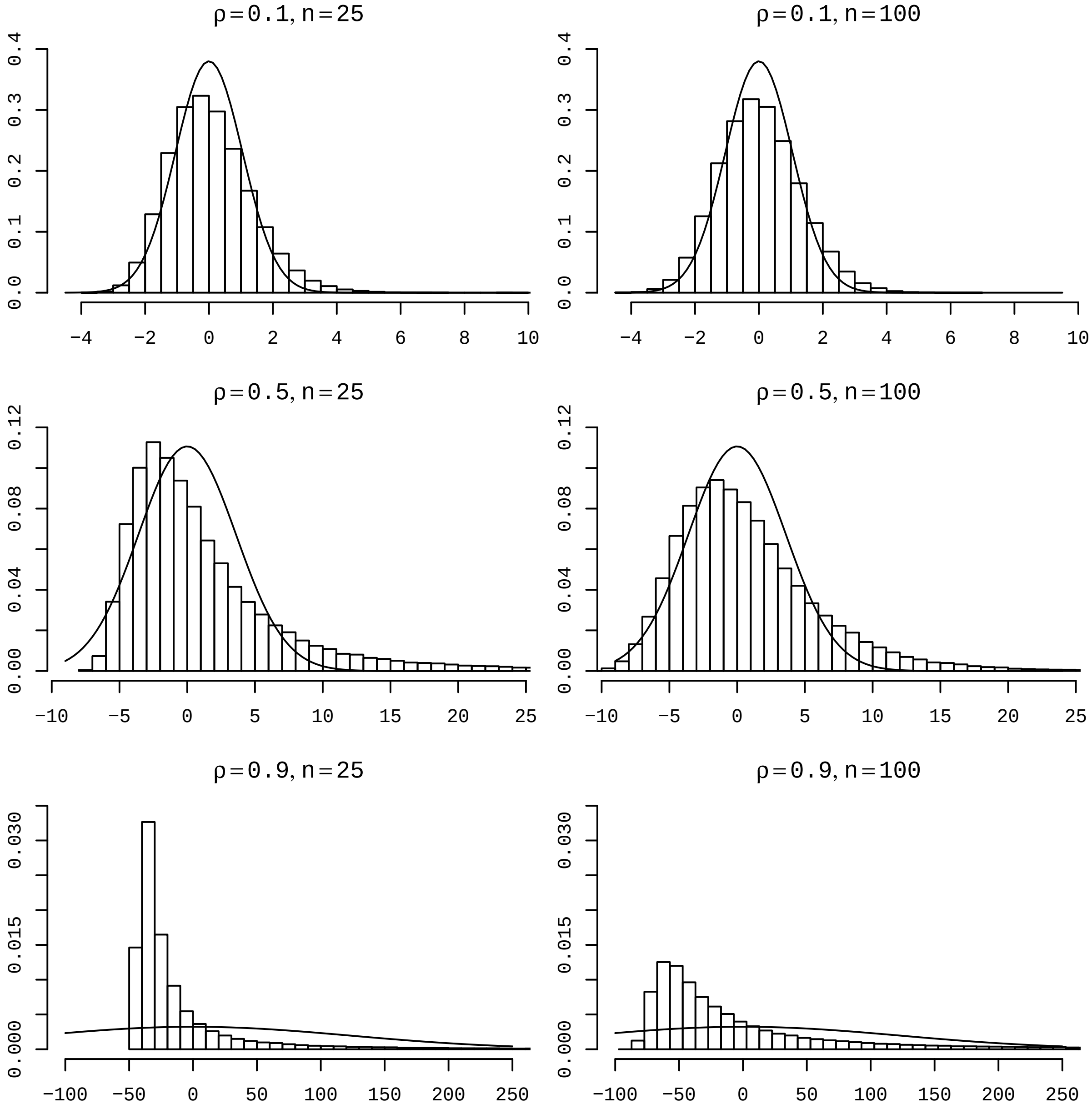

Figure 4.1.1: Histograms and density curve of the limiting distribution of $\sqrt{n}(\widehat{EW} - EW)$ for M/M/1 queue where $\mu = 1$ and $\lambda = .1, .5, .9$ from the top to bottom.

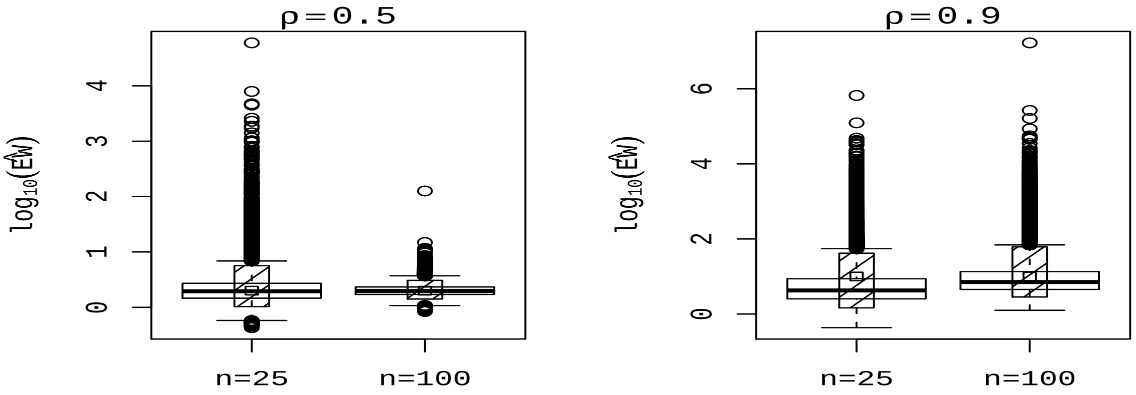

Figure 4.1.2: Boxplots of $10^5 \log_{10} \widehat{EW}$ for $M/M/1$ queues with $\mu = 1$ and $\lambda = .5, .9$. The shaded boxes are the the intervals between the 5% quantile and the 95% quantile for each case and the square points represent the locations of the true EW's.

strong effect on estimation of the sample mean and the sample SD of $\widehat{EW}$, which can be seen in Table 4.1.1. Moreover, the magnitude of the sample standard deviations for $\rho = .9$ in Table 4.1.1 suggests that $\widehat{EW}$ may not have second moments. In fact, generally, even the first moments of $\widehat{EW}$ do not exist.

Theorem 65. *If the distribution of the service time G is not one point distributed at 0, (i.e., $P(G=0) < 1$) and $P(\overline{I} > \overline{G}) > 0$, then conditional the expectation of $\widehat{EW}$ given event $\{\overline{I} > \overline{G}\}$ does not exist.*

Proof. Applying Jensen's inequality, we obtain $(\sum G_j/n)^2 \leq \sum G_j^2/n$, we have

$$E(\widehat{EW} \,|\, \overline{I} > \overline{G}) \geq \frac{1}{2} E(\overline{G}^2/(\overline{I} - \overline{G}) \,|\, \overline{I} > \overline{G}).$$

Let $E(X;\, A)$ denote $E(X\, \mathbf{1}_A)$. Pick any $a > 0$ such that $P\left(\overline{G} > a\right) > 0$. Then we have

$$E\left(\frac{\overline{G}^2}{\overline{I} - \overline{G}} \,|\, \overline{I} > \overline{G}\right) \geq E\left(\frac{a^2}{\overline{I} - \overline{G}}; \overline{G} > a \,|\, \overline{I} > \overline{G}\right) = \frac{a^2}{P(\overline{I} > \overline{G})} \int_a^\infty \int_y^\infty \frac{1}{x - y} f_{\overline{I}}(x)\, f_{\overline{G}}(y)\, dx dy$$
$$= \frac{a^2}{P(\overline{I} > \overline{G})} \int_a^\infty f_{\overline{G}}(y) \left(\int_y^\infty \frac{f_{\overline{I}}(x)}{x - y} dx\right) dy$$

Because $n\overline{I} \sim \text{Gamma}\,(n, \lambda)$, $f_{\overline{I}}(x) > 0$ on $[a, \infty)$ so that the inside integral of the last line diverges for any $y > a$, which completes the proof. □

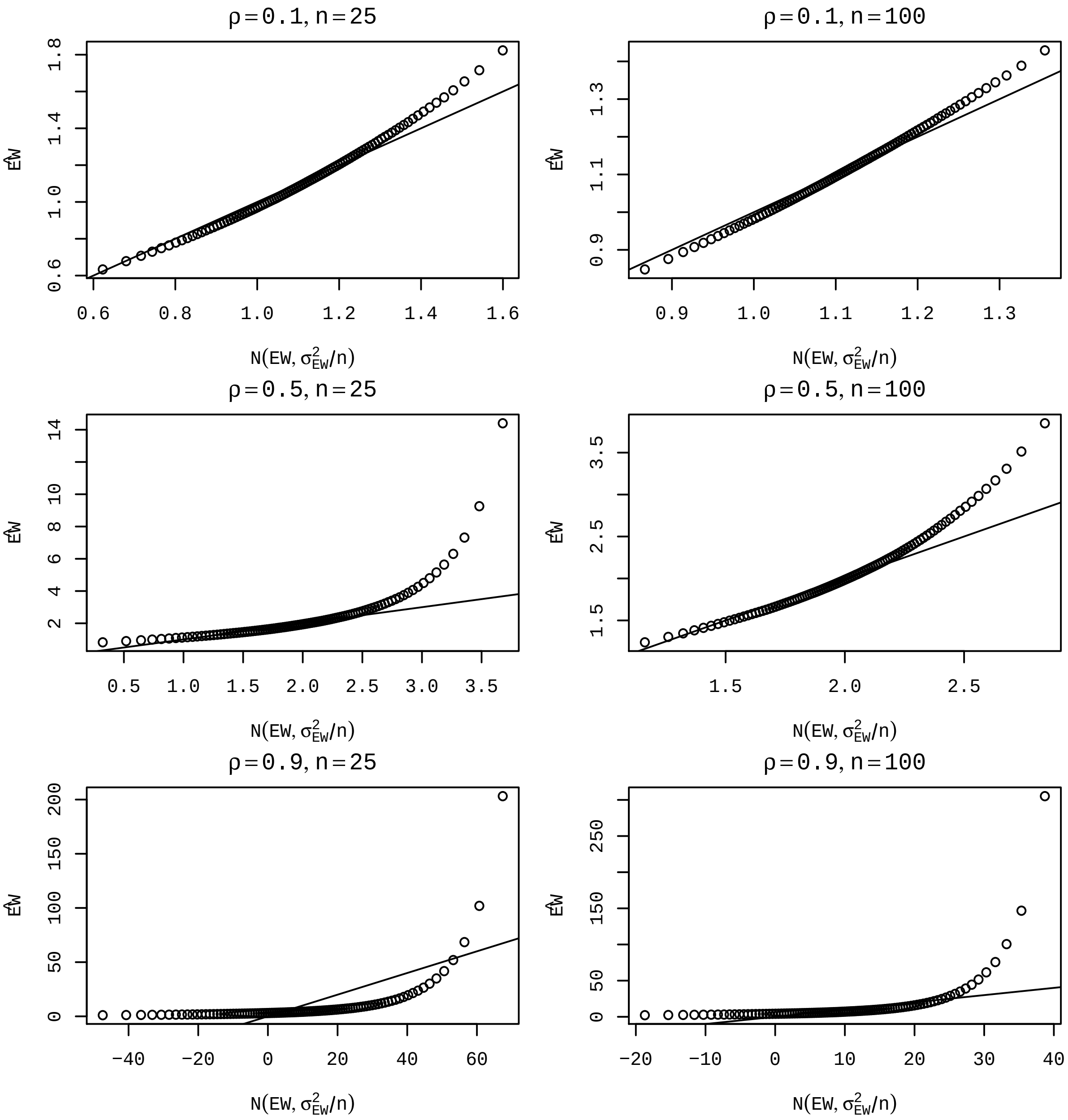

Figure 4.1.3: QQ-plots of $\widehat{EW}$ for for $M/M/1$ queue where $\mu = 1$ and $\lambda = .1, .5, .9$ from the top to bottom with the limiting distribution, $N(EW, \sigma^2_{\widehat{EW}})$. The solid line is the graph of the line $y = x$.

Table 4.1.1: Table for the sample mean, sample median, and sample standard deviation (SD) of the random sample of $\widehat{EW}$ of size 10^5, which is the same random sample used in Figure 4.1.1, 4.1.2, and 4.1.3. E, G, B, and P mean Exp(1), Gamma$(3, 1/3)$, Beta$(2, 2)$, and Pareto$(.8, 5)$, respectively.

			Mean		Med		EW	SD		$\frac{\sigma_{\widehat{EW}}}{\sqrt{n}}$
G	ρ	n	$\widehat{EW}$	$\widetilde{EW}$	$\widehat{EW}$	$\widetilde{EW}$		$\widehat{EW}$	$\widetilde{EW}$	
E	.1	25	1.12	1.12	1.10	1.12	1.11	0.25	0.24	0.21
		100	1.11	1.11	1.11	1.11		0.13	0.13	0.10
	.5	25	3.77	4.60	1.94	2.03	2.00	192.86	468.91	0.72
		100	2.08	2.08	1.98	2.01		0.67	0.54	0.36
	.9	25	31.05	28.92	4.24	4.49	10.00	2186.73	1398.40	24.68
		100	204.01	41.01	7.11	7.21		53032.81	3314.66	12.34
G	.1	25	1.08	1.08	1.07	1.08	1.07	0.14	0.14	0.12
		100	1.07	1.07	1.07	1.07		0.07	0.07	0.06
	.5	25	2.03	2.12	1.68	1.69	1.67	13.67	29.65	0.37
		100	1.70	1.71	1.67	1.67		0.25	0.25	0.18
	.9	25	27.73	19.74	3.92	3.92	7.00	2562.92	678.04	13.66
		100	25.61	26.44	5.74	5.74		1021.21	1083.82	6.83
B	.1	25	0.54	0.54	0.53	0.53	0.53	0.05	0.05	0.05
		100	0.53	0.53	0.53	0.53		0.03	0.03	0.02
	.5	25	0.93	1.01	0.81	0.81	0.80	3.24	20.93	0.15
		100	0.81	0.81	0.80	0.80		0.09	0.09	0.07
	.9	25	10.55	7.91	1.91	1.91	3.20	502.01	133.34	5.85
		100	10.16	15.09	2.71	2.70		303.56	665.62	2.93
P	.1	25	1.06	1.06	1.05	1.06	1.06	0.06	0.06	0.05
		100	1.06	1.06	1.06	1.06		0.03	0.03	0.03
	.5	25	1.81	1.72	1.54	1.55	1.53	17.41	2.04	0.25
		100	1.56	1.56	1.53	1.54		0.15	0.15	0.12
	.9	25	39.56	31.92	3.64	3.66	5.80	6768.93	3043.94	9.90
		100	20.57	28.47	5.02	5.02		530.67	2376.56	4.95

See Figure 4.1.4 for the boxplots of 10^3 $\overline{\widehat{EW}}$ and $\overline{\widehat{EW}^2}$. For $\rho \geq .5$, the sample mean of 10^4 iid sample of $\widehat{EW}$ and $\widehat{EW}^2$ are varying too widely except for Beta$(2,2)$ service distribution of $\rho = .5$, which supports Theorem 65.

It is worth checking whether this feature of heavy right-hand tail is just for exponential service distributions or common for any other service distributions; if the latter is true, our approach to the inference regarding EW should be based on this feature. We suspect the latter is true; Since $\widehat{EW}$ is a function of $\left(\overline{I}, \hat{\mathcal{G}}'(0), \hat{\mathcal{G}}''(0)\right)$, it depends more on the distribution of the means of $\{I_j, G_j, G_j^2\}_{j=1}^n$ than the distribution of I_j and G_j themselves. By CLT, we have

$$\sqrt{n}\left\{\begin{pmatrix} \overline{G} \\ \overline{G^2} \end{pmatrix} - \begin{pmatrix} \mu_1' \\ \mu_2' \end{pmatrix}\right\} \Rightarrow N\left\{0, \begin{pmatrix} \mu_2' - (\mu_1')^2 & \mu_3' - \mu_1'\mu_2' \\ \mu_3' - \mu_1'\mu_2' & \mu_4' - (\mu_2')^2 \end{pmatrix}\right\}$$

and it is known that

$$n\overline{I} \sim \text{Gamma}\,(n, \lambda)\,.$$

Thus, if we define $\widetilde{EW}$ as

$$\widetilde{EW} = Y_1 + \frac{Y_2}{2\,(X'/n - Y_1)},$$

where

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim N\left\{\begin{pmatrix} \mu_1' \\ \mu_2' \end{pmatrix}, \frac{1}{n}\begin{pmatrix} \mu_2' - {\mu_1'}^2 & \mu_3' - \mu_1'\mu_2' \\ \mu_3' - \mu_1'\mu_2' & \mu_4' - {\mu_2'}^2 \end{pmatrix}\right\}$$

and

$$X' \sim \text{Gamma}\,(n, \lambda)\,,$$

then we may expect that the conditional distribution

$$\widetilde{EW} \,|\, Y_1, Y_2, \left(X'/n - Y_1\right) > 0$$

will approximate the conditional distribution of $\widehat{EW} \mid \left(\overline{I} - \overline{G}\right) > 0$.

Figure 4.1.5 shows that the distribution of $\widetilde{EW}$ approximates the distribution of $\widehat{EW}$

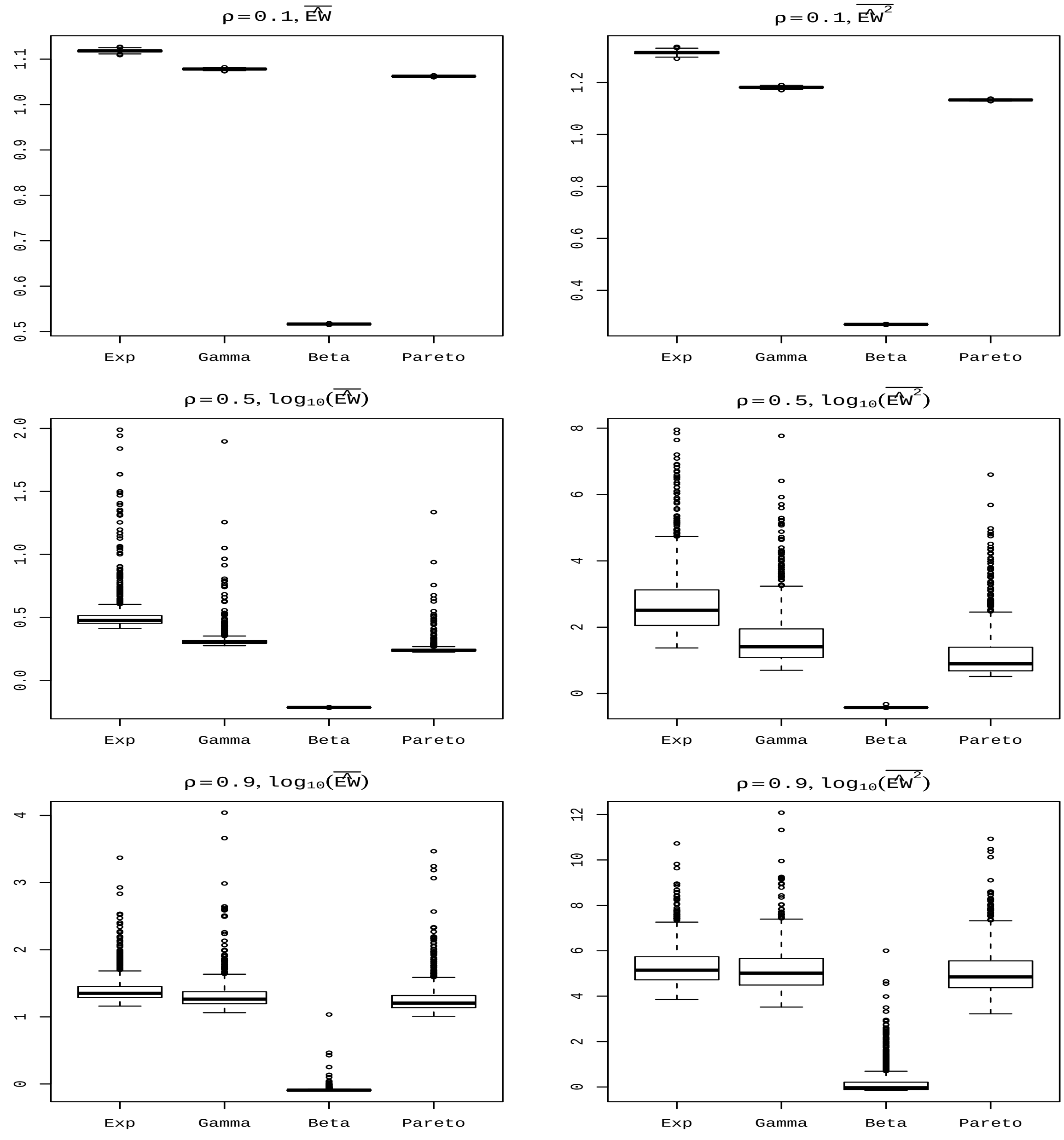

Figure 4.1.4: Boxplots of $10^3\overline{\widehat{EW}}$ and $\overline{\widehat{EW}^2}$ for $\rho = .1$, and $\log_{10}(\overline{\widehat{EW}})$ and $\log_{10}(\overline{\widehat{EW}^2})$ for $\rho = .5$ and .9. Here, $\overline{\widehat{EW}} = \sum_{j=1}^{10^4} \widehat{EW}_j/10^4$ and $\overline{\widehat{EW}^2} = \sum_{j=1}^{10^4} \widehat{EW}_j^2/10^4$. The service time distributions used are Exp(1), Gamma(3, 3), Beta(2, 2) and Pareto(4/5, 5) respectively.

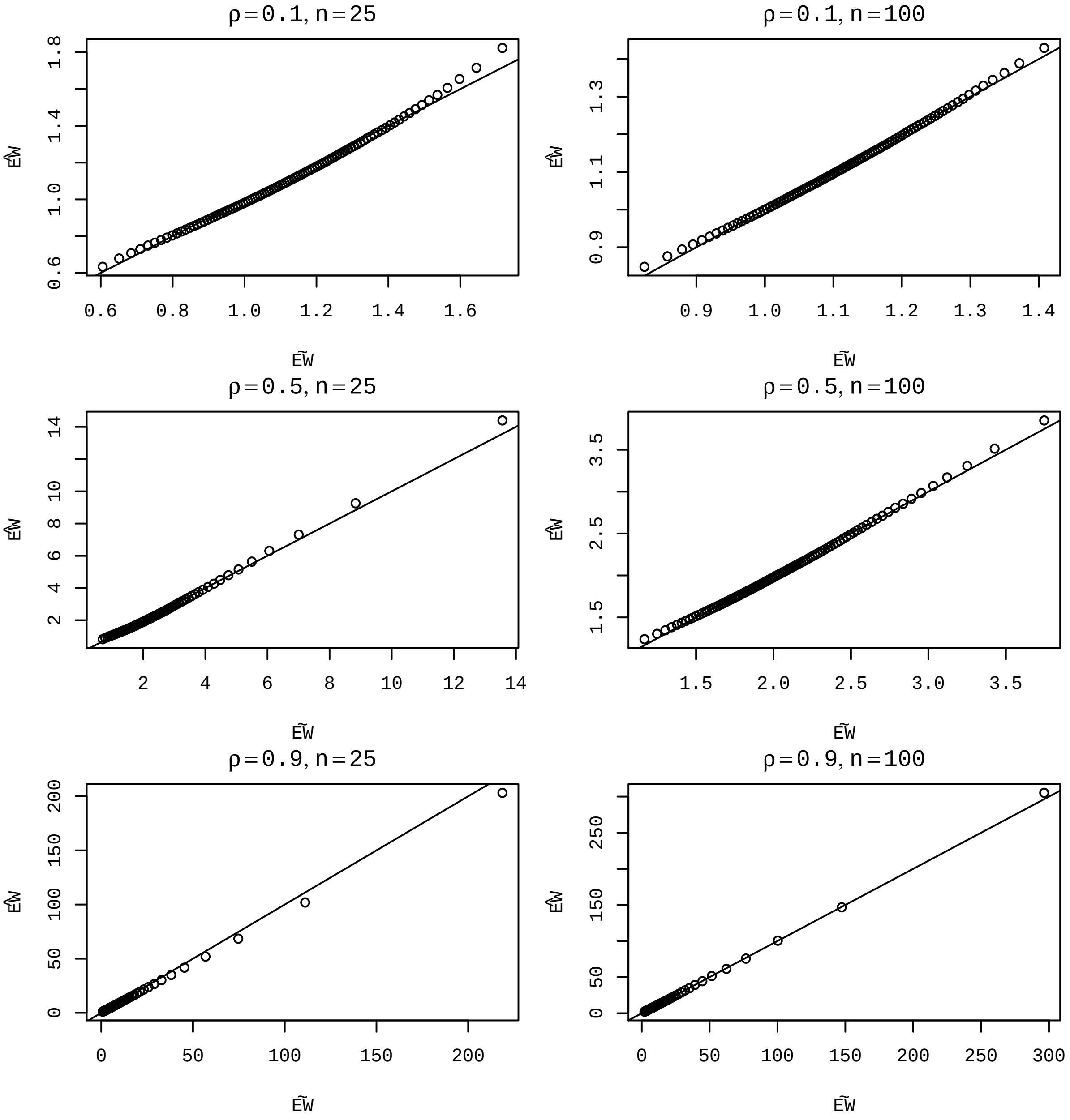

Figure 4.1.5: QQ-plots of 10^5 random sample of $\widehat{EW}$ against 10^5 random sample of $\widetilde{EW}$ for $M/M/1$ queue with $\mu = 1$ and $\lambda = .1$, .5, and .9 from the top to the bottom. The solid line is the graph of the line $y = x$.

well for $M/M/1$ case. The other service distributions we also consider are Beta$(2,2)$ which has the bounded range of $(0,1)$ and Pareto$(4/5,5)$ which is known to have heavy right-hand tail (the nth raw moment of Pareto(α,β) exists only for $n<\beta$). See Figure 4.1.6, which also confirms our finding that even with different service time distributions, $\widetilde{EW}$ approximates $\widehat{EW}$ well. Also, since we only use the first 4 moments for $\widetilde{EW}$, the behavior of $\widehat{EW}$ on the right hand tail is more or less independent of service time distribution but depends on ρ.

In summary, we find that the distribution of $\widehat{EW}$ has a heavier right hand tail as ρ approaches 1, $E(\widehat{EW})$ does not exist, and these facts hold regardless of the distribution of G, the service times.

4.1.2 Confidence intervals of EW

Here, we introduce several ways of constructing $100\,(1-\alpha)\,\%$ confidence intervals of $\widehat{EW}$ based on bootstrap sampling and compare their performances. For properties and applications of bootstrapping method in general, we refer to [29, 26] and [60].

Bootstrap sampling of $\widehat{EW}$

We explain how we obtain the bootstrap sample $\widehat{EW}^*$ from the sample of the inter-arrival times of $\{I_j\}_{j=1}^n$ and the service times of $\{G_j\}_{j=1}^n$ with the assumptions of M/G/1 queue model. We also add two more requirements to the regular assumption of M/G/1 model.

First, we assume that $\rho<1$, so that Pollaczek-Khinchin formula (1.1.1) is valid to use and $EW<\infty$. We also assume that the sample we obtained satisfies the requirement of $\hat{\rho}=\sum I_j/\sum G_j<1$.

Secondly, though the sample size of I_j and G_j does not need to be the same in the our bootstrap sampling process, we assume they are for the brevity of the notation.

Also note that in this dissertation, we follow the convention that the PDF of Exp(λ) is $\lambda\exp(-\lambda x)$ and similarly for Gamma(n,λ). Thus, if $I_j\overset{\text{iid}}{\sim}$ Exp(λ), then from our convention, $EI_j=\lambda^{-1}$ and $\sum_{j=1}^n I_j\sim$ Gamma(n,λ). We obtain the bootstrap sampling of $\widehat{EW}$ as follows:

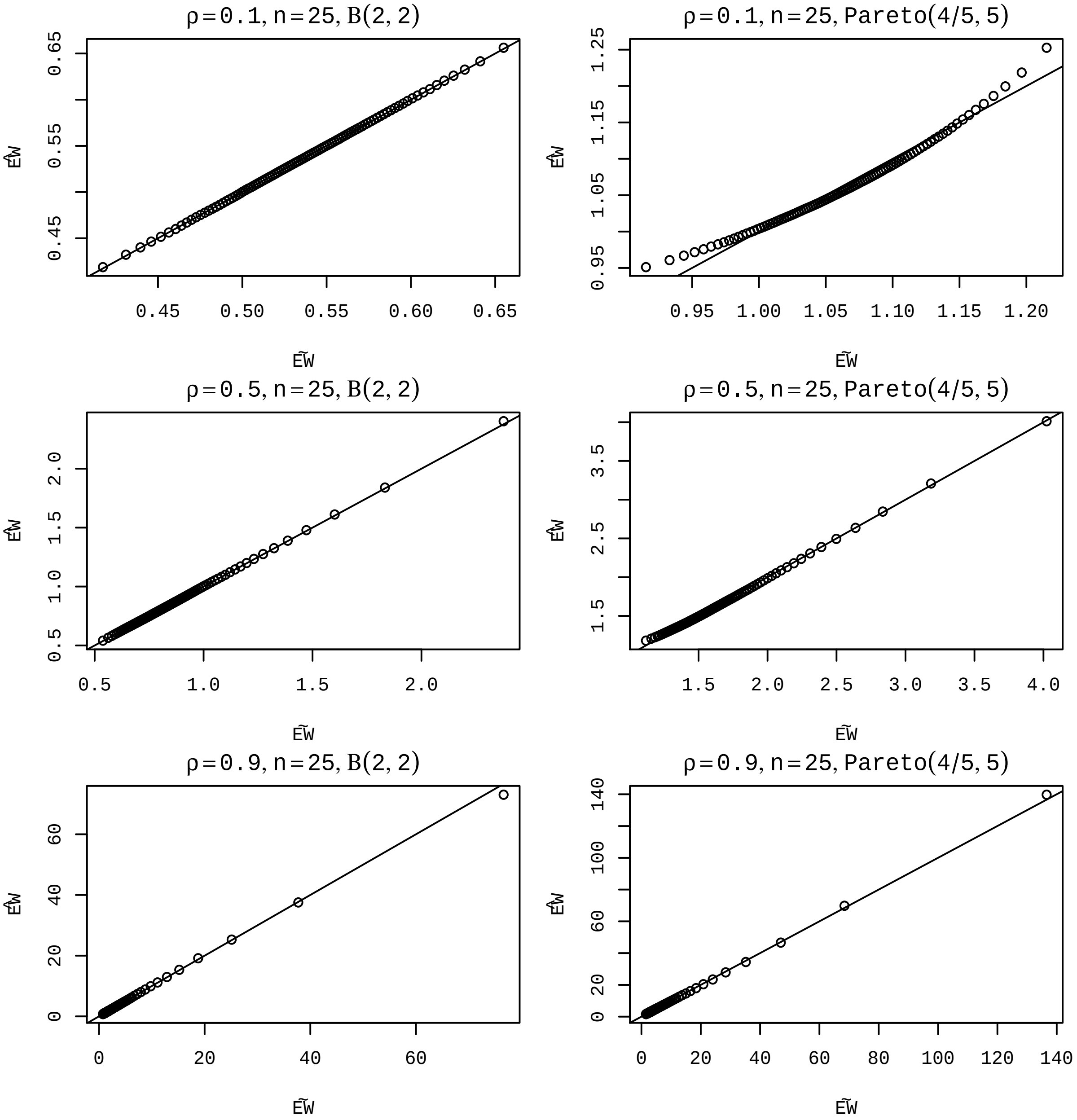

Figure 4.1.6: QQ-plots of 10^5 random sample of $\widehat{EW}$ against 10^5 random sample of $\widetilde{EW}$ for $M/G/1$ queue with $\rho = .1$, .5, and .9 from the top to the bottom for $n = 25$. The left graphs are of $G \sim$ Beta$\,(2, 2)$ ($\mu = 1/2$ and $\lambda = .2$, 1, 1.8) and the right graphs are of $G \sim$ Pareto$\,(.8, 5)$ ($\mu = 1$ and $\lambda = .1$, .5, .9). The solid line is the graph of the line $y = x$.

1. Find the sample mean, $\overline{I}$ of $\{I_j\}$ and set $\hat{\lambda} = 1/\overline{I}$, which is the maximum likelihood estimator of λ.

2. Pick a random sample of size n, $\{I_j^*\}_{j=1}^n$ from $\text{Exp}(\hat{\lambda})$ (or, one random sample of $\text{Gamma}(\hat{\lambda}, n)$ and assign it as $\sum_{j=1}^n I_j^*$) and pick a random sample of size n, $\{G_j^*\}_{j=1}^n$ from $\{G_j\}_{j=1}^n$ with replacement.

3. If
$$\left(\sum_{j=1}^n I_j^* - \sum_{j=1}^n G_j^* \right) \leq 0$$
(or $\hat{\rho}^* = \sum_{j=1}^n G_j^* / \sum_{j=1}^n I_j^* \geq 1$), we discard the bootstrap random samples and repeat 2. If not, calculate $\widehat{EW}^*$ by
$$\widehat{EW}^* = \overline{G^*} + \frac{\sum_{j=1}^n \left(G_j^* \right)^2}{2 \left(\sum_{j=1}^n I_j^* - \sum_{j=1}^n G_j^* \right)}.$$

4. Repeat the steps 2 and 3 till we obtain $\{\widehat{EW}^*\}$ of the size of $B\#$.

This is called parametric bootstrapping, when we use the knowledge of $I_j \overset{\text{iid}}{\sim} \text{Exp}(\lambda)$. The bootstrap sampling of G^* is called a nonparametric bootstrapping. Because of the assumption $\rho < 1$, we discard the bootstrap random sample $\{I_j^*, G_j^*\}_{j=1}^*$ if $(\sum_{j=1}^n I_j^* - \sum_{j=1}^n G_j^*) < 0$ in the step 3, and $\widehat{EW}^*$ is always finite.

Traditional bootstrap confidential intervals

The most recognized bootstrap confidence intervals (CI) are of standard bootstrap, percentile bootstrap, bootstrap-t, and BCa methods. We briefly explain their constructions and short comings for our case. For general introduction of bootstrap confidence intervals, we refer to [29, 26, 60] and [44].

- Standard bootstrap CI: The limits of CI are

$$\widehat{EW} \pm z_{\alpha/2} \cdot \text{SD}(\widehat{EW}^*),$$

where $\mathrm{SD}(\widehat{EW}^*)$ is the sample standard deviation of $\widehat{EW}^*$, or

$$\mathrm{SD}(\widehat{EW}^*) = \left\{ \sum_{j=1}^{n} (\widehat{EW}_j^* - \overline{\widehat{EW}^*})^2 / (B-1) \right\}^{1/2}.$$

As noted in [44], this method requires that $\widehat{EW}$ follows approximately normal distribution and $\mathrm{SD}(\widehat{EW}^*)$ is a a good approximation of σ_{EW} (Bias adjustment can be used if $\widehat{EW} - \overline{\widehat{EW}^*}$ is not closed to 0 as noted in [26]).

- Percentile intervals: There are two types of this method; Let $\widehat{EW}^{*(\alpha)}$ be α-quantile of $\left\{ \widehat{EW}_j^* \right\}_{j=1}^{B\#}$. Then the $(1-\alpha)$ percentile intervals is

$$\left(\widehat{EW}^{*(\alpha/2)}, \widehat{EW}^{*(1-\alpha/2)} \right). \tag{4.1.2}$$

Suppose that there exists an increasing function $g(x)$ such that $g(\widehat{EW}^*)$ is symmetric around $g(EW)$, so that the exact confidence interval, (L, U) can be obtained. It can be shown that $g^{-1}(L) = \widehat{EW}^{*(\alpha/2)}$ and $g^{-1}(U) = \widehat{EW}^{*(1-\alpha/2)}$ (see [60]) because the percentile interval is transformation-respecting ([29]). The other interval is based on the observation that if the distribution of $\widehat{EW} - EW$ can be approximated by the distribution of $\widehat{EW}^* - \widehat{EW}$, then the resulting approximate interval would be

$$\left(2\widehat{EW} - \widehat{EW}^{*(1-\alpha/2)}, 2\widehat{EW} - \widehat{EW}^{*(\alpha/2)} \right), \tag{4.1.3}$$

which is called the basic confidence interval in [26]. Since $\widehat{EW}^*$ is heavily skewed to the right, it is even possible to have negative lower end limit for the latter interval.

- Bootstrap-t CI: For each of $\widehat{EW}_1^*, \cdots, \widehat{EW}_{B\#}^*$, define T_j^* by

$$T_j^* := \frac{\widehat{EW}_j^* - \widehat{EW}}{\mathrm{SD}(\widehat{EW}_j^*)}$$

and the bootstrap-t CI is defined by

$$\left(\widehat{EW} - T^{*(\alpha/2)}\mathrm{SD}\left(\widehat{EW}^*\right), \widehat{EW} - T^{*(1-\alpha/2)}\mathrm{SD}\left(\widehat{EW}^*\right)\right),$$

where $T^{*(\alpha/2)}$ and $T^{*(1-\alpha/2)}$ are $(\alpha/2)$ and $(1-\alpha/2)$ quantile of T_j^*. Unless there is a known simple formula for SD $\left(\widehat{EW}_j^*\right)$, it needs to be calculated from nested bootstrap sampling.

- BCa (bootstrap accelerated bias-corrected percentile) CI: We first introduce bias-corrected percentile CI. Suppose there is an increasing function g such that

$$g(\widehat{EW}^*) - g\left(EW\right) + z_0 \sim N\left(0, 1\right),$$

which is more general assumption than we had for the bootstrap percentile method. Let $\Phi\left(x\right)$ be the CDF of the standard normal distribution and $\hat{F}_{\widehat{EW}^*}$ be the empirical CDF of $\widehat{EW}^*$. Then we have

$$P(\widehat{EW}^* \leq EW) = P\{g(\widehat{EW}^*) - g\left(EW\right) + z_0 \leq z_0\} = \Phi\left(z_0\right),$$

which implies

$$z_0 = \Phi^{-1}\left(\hat{F}_{\widehat{EW}^*}(\widehat{EW})\right)$$

and similar calculation give us the CI,

$$\left(\hat{F}^{-1}_{\widehat{EW}^*}\{\Phi(2z_0 - z_{\alpha/2})\}, \hat{F}^{-1}_{\widehat{EW}^*}\{\Phi(2z_0 + z_{\alpha/2})\}\right), \tag{4.1.4}$$

which is a precursor ([28]) of BCa CI, which assumes an increasing function $g\left(x\right)$ and constant a such that

$$\frac{g(\widehat{EW}^*) - g\left(EW\right)}{1 + ag\left(EW\right)} + z_0 \sim N\left(0, 1\right).$$

Then, we can find the limits of CI in a similar way to the above to obtain

$$\left(\hat{F}^{-1}_{\widehat{EW}^*}\left\{\Phi\left(z_0+\frac{z_0-z_{\alpha/2}}{1-a\left(z_0-z_{\alpha/2}\right)}\right)\right\},\hat{F}^{-1}_{\widehat{EW}^*}\left\{\Phi\left(z_0+\frac{z_0+z_{\alpha/2}}{1-a\left(z_0+z_{\alpha/2}\right)}\right)\right\}\right)$$

and a is called the acceleration constant. There are several estimators of it and one used in [29] is

$$\frac{\sum_{j=1}^{n}\left(\overline{\widehat{EW}_{-}}-\widehat{EW}_{-j}\right)^3}{6\left\{\sum_{j=1}^{n}\left(\overline{\widehat{EW}_{-}}-\widehat{EW}_{-j}\right)^2\right\}^{3/2}},$$

where $\widehat{EW}_{-j}$ is the estimator of EW with jth observed random sample (I_j, G_j) deleted and $\overline{\widehat{EW}_{-}}{=}\sum_{j=1}^{n}\widehat{EW}_{-j}/n$.

A somewhat naive[1] simulation study of these 4 different bootstrap CI's of the means of the stationary waiting time distributions with phase-type service distributions was done in [19]. They reported that though the standard bootstrap CI performed best at coverage probability, only the percentile bootstrap CI is practical to use since for $\rho \geq .5$, the average width of the other three CI methods was too wide. In one of their simulation results, they report that for $M/H_4/1$ (H_4 means a mixture of 4 exponential distribution, which is a hyper-exponential distribution) queue with $\rho = .9$ and $n = 20$, the average width of BCa CI was 53,393.57 while the average width of percentile bootstrap CI was 27.97. However, they did not explain why this occurs.

By Theorem 65, we cannot expect that the sample moment of bootstrapped sample $\widehat{EW}^*$ is converging to any number and the estimated moments are unreliable to use, which is why the bootstrap CI methods other than percentile method give too wide widths. As we saw in the previous section any method which needs moment estimates will result in too

[1] Although their result is also confirmed by our simulation study that the percentile method will be the recommended choice among 4 bootstrap CI's, we believe that their simulation study has three drawbacks; Even though $M/G/1$ queue is assumed (so that the inter-arrival time follows exponential distribution) they used non-parametric bootstrap for I_j^*'s. Secondly, it seems that they follows [29] word for word literally so that their B, the number of bootstrap sample is only 1000 (and for the estimation of standard error of bootstrap-t interval, $B = 25$), which are not big enough considering heavy tail of $\widehat{EW}^*$. Thirdly, they did not investigate why the other three (standard, bootstrap-t, and BCa) have much wider average lengths but their coverage probabilities still perform poorly comparing the percentile method. They only reported the result of their simulation study.

wide width because the estimator does not have the moments of any orders.

For an example, we observed in Figure 4.1.1, 4.1.3, and 4.1.6 that the distribution of $\widehat{EW}$ is heavily skewed as ρ gets close to 1 regardless of the service time distributions. See Table (4.1.1) for the sample mean, sample median, and sample standard deviation (SD) of $\widehat{EW}$ of M/G/1 queue with $\mu = 1$ and $\lambda = .1, .5, .9$ for 4 different service time distributions. The heavy right hand tail has a strong effect on the estimation of sample mean and the sample SD of $\widehat{EW}$ and because of it, any CI based on the estimation of the standard deviation of $\widehat{EW}$ will be too wide.

The percentile method is robust against skewness (in Table 4.1.1, the sample median is much closer to EW than the sample mean for $\rho = .5$ and $\rho = .9$) and does not require any moment estimation, which is why it gives relatively practical widths and performs better than the other three method for 90% CI in [19].

Note that the choice of lower and upper limit of percentile method (4.1.2) is not the only way to pick the limits. If $f(x)$ be the density function of a random variable X, $100\,(1-\alpha)\,\%$ highest density region (HDR) is the subset $R(f_\alpha)$ satisfying

$$R\left(f_\alpha\right) = \left\{x : f\left(x\right) \geq f_\alpha\right\},$$

where f_α is the largest constant such that

$$P\left\{X \in R\left(f_\alpha\right)\right\} \geq 1-\alpha.$$

By the definition, the HDR is allowed to be the union of disjoint intervals and has the smallest possible volume in the sample space of X (see [37] for more discussion and its usage to data representations). In [40], HDR with bias correction methods was used to obtain bootstrap confidence intervals, where the bootstrap sampling distribution has the similar characteristic of $\widehat{EW}$, heavily skewed right tail.

Because the sampling distribution of $\widehat{EW}$ is unimodal and heavily skewed to the right, a direct application of the HDR method will result in an interval close to

$$\left(\widehat{EW}^{*0}, \widehat{EW}^{*(1-\alpha)}\right).$$

To reduce this effect, we may use the HDR method after a transform like the Box-Cox power transform. It seems log transform works well, which is also heuristically appealing; by the mean value theorem, if we set $\widehat{EW}^*_{(k)}$ to be the kth order statistic of $\widehat{EW}^*$, then

$$\log \widehat{EW}^*_{(k+1)} - \log \widehat{EW}^*_{(k)} = \left(\frac{1}{a}\right)\left(\widehat{EW}^*_{(k+1)} - \widehat{EW}^*_{(k)}\right),$$

where a is a constant satisfying

$$\widehat{EW}^*_{(k)} \leq a \leq \widehat{EW}^*_{(k+1)}.$$

Since the magnitude of $\widehat{EW}^*$ is mostly affected by the term $(\overline{I}^* - \overline{G}^*)^{-1}$, the reciprocal factor $(1/a)$ reduce the skewness and $(1/a)$ gets smaller as k gets bigger.

If (L, U) is a $100\,(1-\alpha)\,\%$ HDR of $\log \widehat{EW}^*$, then the $100\,(1-\alpha)\,\%$ confidence interval of EW is (e^L, e^U). Note that unlike the bootstrap percentile interval, HDR method is not transform-respecting and different transforms will result in different upper limits and lower limits for the confidence intervals.

Though generally finding HDR interval needs kernel density estimation (, which requires a transform because of skewness of $\widehat{EW}^*$ and we find log works well for this too), we may use the discrete version of HDR because of the unimodality of $\widehat{EW}^*$. For $(1-\alpha)\,100\%$ HDR interval, let a_0 be defined by

$$a_0 := \underset{0 \leq a \leq \alpha}{\text{argmin}} \left\{\widehat{EW}^{*(1-a)} - \widehat{EW}^{*(a)}\right\},$$

then

$$\left(\widehat{EW}^{*(a_0)}, \widehat{EW}^{*(1-a_0)}\right)$$

Table 4.1.2: Service time distribution considered

G	PDF	$\mu'_1, \mu'_2, \mu'_3, \mu'_4$
$\mathrm{Exp}\,(1)$	e^{-x}	1, 2, 6, 24
$\mathrm{Gamma}\,(3,3)$	$3^3x^2e^{-3x}/\Gamma\,(3)$	$1,\ 1\frac{1}{3},\ 2\frac{2}{9},\ 4\frac{4}{9}$
$\mathrm{Beta}\,(2,2)$	$6x\,(1-x)$	$\frac{1}{2},\ \frac{3}{10},\ \frac{1}{5},\ \frac{1}{7}$
$\mathrm{Pareto}\left(\frac{4}{5},5\right)$	$\left\{5\left(\frac{4}{5}\right)^5/x^6\right\}\mathbf{1}_{[4/5,\infty)}\,(x)$	$1,\ \frac{16}{15},\ \frac{32}{25},\ \frac{256}{125}$

will be the HDR interval.

4.1.3 Simulation study of the first round

Here, we show our own simulation study for the bootstrap percentile confidence intervals, asymptotic confidence intervals, and HDR after log transform for the mean of the stationary waiting time distribution with a confidence level $\alpha = .1$ (i.e., 90% confidence intervals). The service time distributions we consider are $\mathrm{Exp}\,(1)$, $\mathrm{Gamma}\,(3,3)$, $\mathrm{Beta}\,(2,2)$, and $\mathrm{Pareto}\,(4/5,5)$. See the table for the PDF's and the moments μ'_j, $j = 1, \cdots, 4$. We compare the estimated coverage probabilities and the estimated expected lengths from $l = 2500$ random samples of size $n = 25$ and 100 respectively for different $\rho = .1$, .5, .9, and different service time distributions of Table 4.1.2. Among the confidence intervals for a binomial proportion recommended in ([16]), the Wilson interval and the Agresti-Coull interval have relatively easy formula to decide the sample size needed to obtain a preassigned interval width (for the sample size for other CI intervals reviewed in [16], see [50]). We set the required interval length $w = .02$ with the confidence level $\alpha = .1$ and assume the true coverage probability $p = .9$ to decide the sample size $l = 2500$, which is enough for either of Wilson's and Agresti-Coull's intervals. Thus, with 90% confidence, the true coverage probability is within the observed coverage probability $\pm .01$. The bootstrap sample size $B\# = 10^5$ for each case. Note that because the sampling distribution of $\widehat{EW}^*$ has the heavy right tail as we saw in the previous subsection, we suggest the bootstrap sample size $B\#$ to be at least greater than 10^5.

From Table 4.1.3, one would pick the standard bootstrap percentile confidence interval as the best method; its observed coverage probabilities are close to the nominal level of

Table 4.1.3: The observed coverage probabilities of the bootstrap percentile confidence intervals (BP_1 of (4.1.2, the standard) and BP_2 of (4.1.3, the basic), HDR after log transform, and the bias-corrected percentile CI (BC of (4.1.4)), and the asymptotic CI (Asym of (4.1.1) of 90% confidence level. av.c denotes the average coverage probabilities and av.l denotes the average length of the confidence intervals.

			$\text{BP}_{1,2}$			HDR		BC		Asym	
G	ρ	n	av.c_1	av.c_2	av.l	av.c	av.l	av.c	av.l	av.c	av.l
E	.1	25	.861	.858	0.79	.868	0.80	.866	0.80	.804	0.66
		100	.880	.878	0.41	.881	0.41	.878	0.41	.815	0.34
	.5	25	.884	.772	8.32	.890	5.68	.833	9720.56	.781	124.75
		100	.879	.853	2.02	.897	1.85	.878	2.09	.839	1.35
	.9	25	.820	.312	25.69	.692	15.49	.754	120076.71	.538	6995.21
		100	.908	.510	45.75	.844	26.81	.791	161050.93	.684	159452.15
G	.1	25	.866	.866	0.44	.866	0.44	.868	0.44	.814	0.38
		100	.887	.886	0.22	.886	0.22	.886	0.22	.835	0.19
	.5	25	.886	.828	4.17	.915	2.86	.852	957.85	.838	80.20
		100	.894	.881	0.87	.915	0.83	.897	0.87	.865	0.66
	.9	25	.885	.389	21.84	.772	12.40	.772	62361.73	.622	104061.46
		100	.942	.594	35.63	.887	20.21	.805	73351.31	.760	16514.90
B	.1	25	.884	.893	0.17	.879	0.17	.893	0.17	.849	0.15
		100	.894	.892	0.08	.891	0.08	.894	0.08	.854	0.08
	.5	25	.897	.879	1.52	.943	1.06	.880	163.34	.888	1.86
		100	.891	.890	0.33	.904	0.32	.888	0.33	.871	0.27
	.9	25	.914	.412	10.26	.809	5.75	.799	21308.35	.646	18959.37
		100	.944	.593	16.19	.888	9.11	.802	38522.58	.757	1163.32
P	.1	25	.852	.826	0.19	.844	0.18	.848	0.19	.800	0.16
		100	.873	.867	0.10	.874	0.10	.870	0.10	.825	0.09
	.5	25	.902	.812	2.39	.922	1.60	.884	799.77	.867	3.44
		100	.891	.864	0.53	.906	0.50	.887	0.53	.880	0.44
	.9	25	.928	.408	19.37	.816	10.43	.796	74585.97	.664	21958.32
		100	.944	.606	29.40	.892	16.37	.807	63381.32	.768	3426020.91

.9 over different values of ρ and the mean length of intervals are within practical range of usage (HDR after log transform has the smallest mean lengths but the observed coverage probabilities for $\rho = .9$ case are not close to the nominal level). However, as we will see in Table 4.1.5, when $\rho = .95$, the performance deteriorates to an unacceptable level for a reasonable 90% CI. In the next subsection, we examine why this occurs and suggest better CI methods.

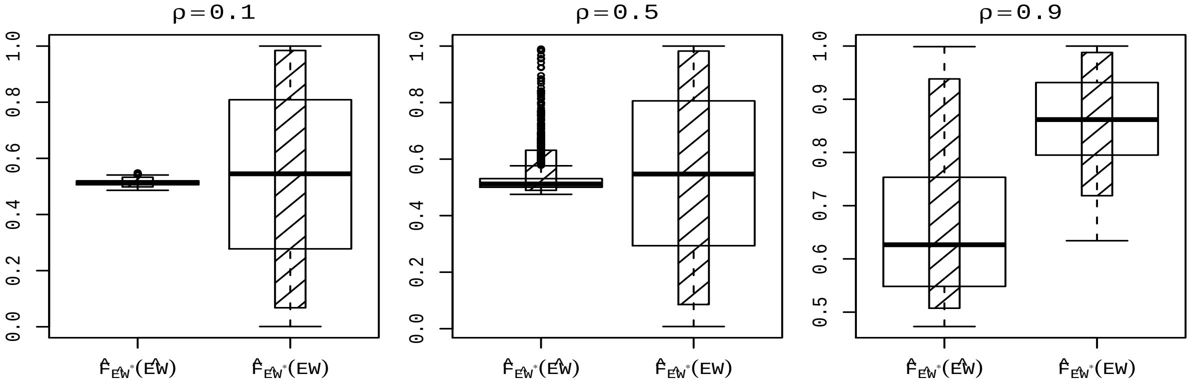

Figure 4.1.7: Boxplots of 2500 $\hat{F}_{\widehat{EW}^*}(\widehat{EW})$ and $\hat{F}_{\widehat{EW}^*}(EW)$ from the simulation of the Table 4.1.3 with the service time distribution of Exp(1). The shaded boxes are the the intervals between the 5% quantile and the 95% quantile for each case. For $\hat{F}_{\widehat{EW}^*}(EW)$, they are (0.0677, 0.9842), (0.0856, 0.9824), and (0.7188, 0.9878), respectively from the left to the right ($\rho = .1$, .5, .9).

4.1.4 Varied percentile limit CI

Close look at percentile methods

Let $\hat{F}_{\widehat{EW}^*}$ be the sampling distribution of $\widehat{EW}^*$ (i.e, $\hat{F}_{\widehat{EW}^*}$ in the previous subsection, the empirical CDF of $\widehat{EW}^*$). Figure 4.1.7 shows boxplots of 2500 $\hat{F}_{\widehat{EW}^*}(\widehat{EW})$ and $\hat{F}_{\widehat{EW}^*}(EW)$ for different ρ (=.1, .5, and .9) of M/M/1 queue case, which are obtained from the simulation of the previous subsection and Table 4.1.3.

Since $E(\widehat{EM}) = \infty$, the mean bias is meaningless in estimating EW and we rather consider the median bias instead;

$$\text{Median Bias of } \hat{\theta} = \text{Median of } \hat{\theta} - \theta.$$

Because

$$\text{Med}\,\widehat{EW}^* - \widehat{EW} < (>)\, 0 \Longleftrightarrow \hat{F}_{\widehat{EW}^*}(EW) > (<)\, .5,$$

Figure 4.1.7 confirms that when $\rho = .9$, $\widehat{EW}^*$ is median biased in most of the 2500 random samples and we also see this phenomenon in Table 4.1.1 , to which one may attribute the lower average coverage probability of the bootstrap percentile CI. This suspicion may be supported by the observation that the equal tail 90% percentile of 2500 $\hat{F}_{\widehat{EW}^*}(EW)$ for

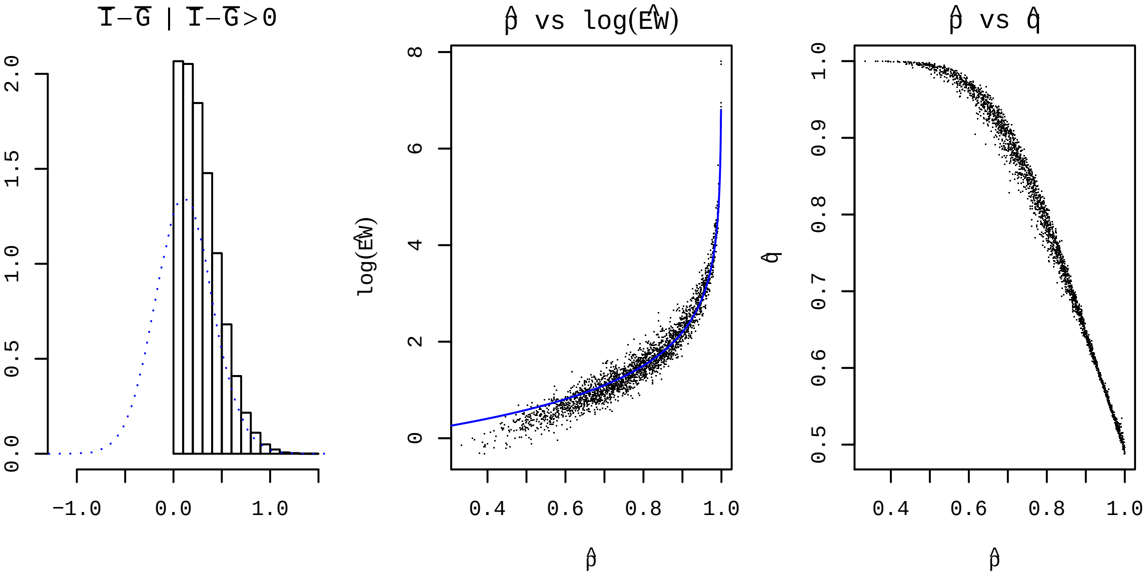

Figure 4.1.8: Left: Histogram of $10^5\left(\overline{I}-\overline{G}\right)$ conditional on $\left(\overline{I}-\overline{G}\right)>0$ for M/M/1 queue with $\rho=.9$ and $n=25$ ($G_j\sim\text{Exp}\,(1)$ and $X_j\sim\text{Exp}\,(.9)$). The dotted curve is the kernel density estimation of unconditional $10^5\ \left(\overline{I}-\overline{G}\right)$. Middle: $\log(\widehat{EW})$ against $\hat{\rho}$ from the simulation of Table 4.1.3 for M/M/1 with $\rho=.9$. The solid curve is of $\log\{.9\,(1-x)^{-1}\}$. Right: $\hat{q}$ of (4.1.6) against $\hat{\rho}$ from the simulation of Table 4.1.3 for M/M/1 with $\rho=.9$.

$\rho=.9$ is (0.7188, 0.9878), which means that if we set the bootstrap percentile CI to be $(\widehat{EW}^{*(.7188)},\widehat{EW}^{*(.9878)})$, then the estimated coverage probability will be .9 for the 2500 random samples we used for Table 4.1.3 (Note that the 90% quantile of $\hat{F}_{\widehat{EW}^*}(EW)$ for $\rho=.9$ is 0.9714 so that the upper limit of the 90% bootstrap percentile CI should be greater than .9714 to have the average coverage probability not smaller than .9).

The reason why there is the median bias for ρ being close to 1 is because we impose the requirement of $(\overline{I}^*-\overline{G}^*)>0$ for the bootstrap sampling. Thus, the resulting $\widehat{EW}^*$ is in fact conditional on $(\overline{I}^*-\overline{G}^*)>0$. See Figure 4.1.8 for the histogram of $10^5\ (\overline{I}_{25}-\overline{G}_{25})$ conditional on $\overline{I}_{25}-\overline{G}_{25}>0$ with the assumption of M/M/1 queues with $\rho=.9$. By comparing the overlapped dotted curve, which is the kernel density estimation of 10^5 values of $\left(\overline{I}-\overline{G}\right)$ unconditionally, to the histogram, one can see that $\left(\overline{I}-\overline{G}\right)\mid\left(\overline{I}-\overline{G}\right)>0$ will be mean (and median) biased against the mean (and the median) of the unconditional $\left(\overline{I}-\overline{G}\right)$.

We note that the condition was also imposed for the 2500 random samples of $\{I_j\}$ and $\{G_j\}$ of the simulation study of Table 4.1.5. See Table 4.1.4 for the sample mean and the median of $\hat{\rho}$ of Table 4.1.5.

Table 4.1.4: The sample mean and the sample median of $\hat{\rho}$ of the simulation of Table (4.1.5)

	Exp(1)		Gamma(3, 3)		Beta(2, 2)		Pareto(.8, 5)	
ρ	$\bar{\hat{\rho}}$	Med $\hat{\rho}$	$\bar{\hat{\rho}}$	Med $\hat{\rho}$	$\bar{\hat{\rho}}$	Med $\hat{\rho}$	$\bar{\hat{\rho}}$	Med $\hat{\rho}$
.1	0.105	0.100	0.105	0.102	0.104	0.101	0.104	0.101
.5	0.517	0.498	0.523	0.508	0.520	0.506	0.518	0.504
.9	0.778	0.791	0.810	0.824	0.816	0.827	0.826	0.834

From the definition of $\widehat{EW}$, when ρ is close to 1, $\widehat{EW} = O((1-\hat{\rho})^{-1})$, or equivalently, $\left(\overline{I} - \overline{G}\right)$ will be dominant. See the graph in the middle in 4.1.8, which shows $\log(\widehat{EW})$ against $\hat{\rho}$ of M/M/1 queue with $\rho = .9$ from the simulation of Table 4.1.3. Note that

$$\widehat{EW} - \mu_1' = \frac{\lambda \mu_2'}{2\,(1-\rho)} = \frac{.9}{1-\rho}$$

for $G \sim \text{Exp}\,(1)$ and $\lambda = .9$ and the overlapped curve is of $\log\{.9/\,(1-x)\}$, which confirms that $(1-\hat{\rho})$ is dominant in the estimator $\widehat{EW}$ when ρ is close to 1 and we can see that the median (or mean) bias of $\left(\overline{I} - \overline{G}\right) \mid \left(\overline{X} - \overline{G}\right) > 0$ will result in the median bias of $\widehat{EW}$ for ρ being close to 1.

As we mentioned, one can try to apply a bias correction method within bootstrap sampling estimation of $\widehat{EW}^*$ but the adjustment may push the sample to the unstationary condition $\hat{\rho}^* \geq 1$. For this, one possible remedy is using Kilian's method ([39]) but we find that the bias correction is unnecessary.

Varied percentile limit confidence interval

We propose one variation of bootstrap percentile CI, which we call varied percentile limit (VPL) CI. There are two requirements we consider to develop this method. First, it is based on the percentiles of $\widehat{EW}^*$ and for a $(1-\alpha)100\%$ confidence interval, our proposed interval should contain $(1-\alpha)100\%$ of $\widehat{EW}^*$. Secondly, the calculations of the lower limit and upper limit of the CI should be simple; we choose a CI with a simple formula with a reasonable coverage probability rather than a CI of more complicated formula with a possibly better performance.

Since we are considering CI of $(1-\alpha)100\%$, the limits of the CI will have the form of

$$\left(\widehat{EW}^{*(a)}, \widehat{EW}^{*(1-\alpha+a)}\right) \quad a \in (0, \alpha). \tag{4.1.5}$$

See Figure 4.1.9 for the graphs of observed average coverage probabilities of CI having the form of 4.1.5 with the confidence level α=.1 for different a from the simulation of Table 4.1.3. It is clear that a percentile CI with a fixed a will not have uniformly good coverage probabilities over different service time distributions and different ρ's. In other words, we need to consider a CI whose limits are adjustable.

Let q be the probability of $\left(\overline{I}-\overline{G}\right)>0$, which can be estimated by

$$\hat{q} = \frac{B\#}{\text{total }\#\text{of}\left(\overline{I}^{*}-\overline{G}^{*}\right)}, \tag{4.1.6}$$

where B is the bootstrap sample size of $\widehat{EW}^{*}$ satisfying $(\overline{I}^{*}-\overline{G}^{*})>0$. Thus, $\hat{q}$ is the rate of taking $\{I_j^*, G_j^*\}$ in bootstrap sampling step 3. Clearly, $\hat{q}$ will be smaller as $\hat{\rho}$ is getting close to 1 as we can see the right graph in Figure 4.1.8, which shows $\hat{\rho}$ against $\hat{q}$ from the simulation of Table 4.1.3 for M/M/1 with $\rho=.9$. The varied percentile limit CI is the CI having the form of 4.1.5, where the constant a is defined by

$$a = .9\alpha\hat{q}. \tag{4.1.7}$$

Note that the constant of .9 is multiplied to avoid the case of $a=\alpha$, which will result in the upper limit of $\widehat{EW}^{*(1)}$, the maximum of $\widehat{EW}^{*}$.

We now explain why our choice of a works (and how we derived formula (4.1.7)). See Figure 4.1.10 for the graphs of $\hat{F}_{\widehat{EW}^*}(EW)$ against $\hat{q}$ from the simulation of Table 4.1.3 for M/M/1 with different ρ=.1 , .5, and .9. We notice that for a lower $\hat{q}$, $\hat{F}_{\widehat{EW}^*}(EW)$ is low too. Note that the low value of $\hat{q}$ means that $\widehat{EW}^{*}$ is more (median) biased against EW and $\widehat{EW}$. Then intuitively one may think that a higher upper limit of the CI will be needed to capture the true value of EW. Thus that observation contradicts our intuition.

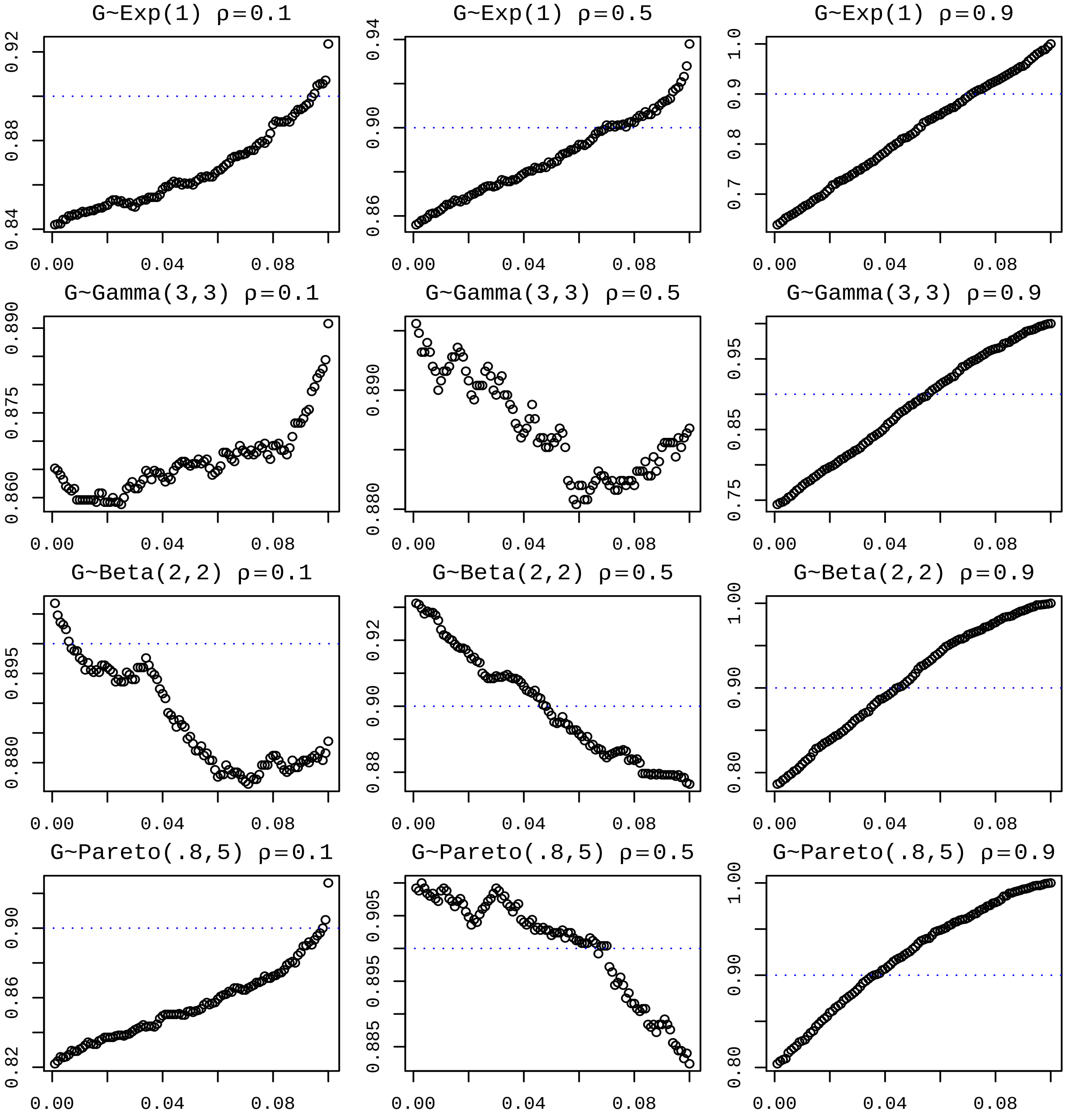

Figure 4.1.9: Observed average coverage probabilities of $(\widehat{EW}^{*(a)}, \widehat{EW}^{*(.9+a)})$ for different a from the simulation of Table 4.1.3.

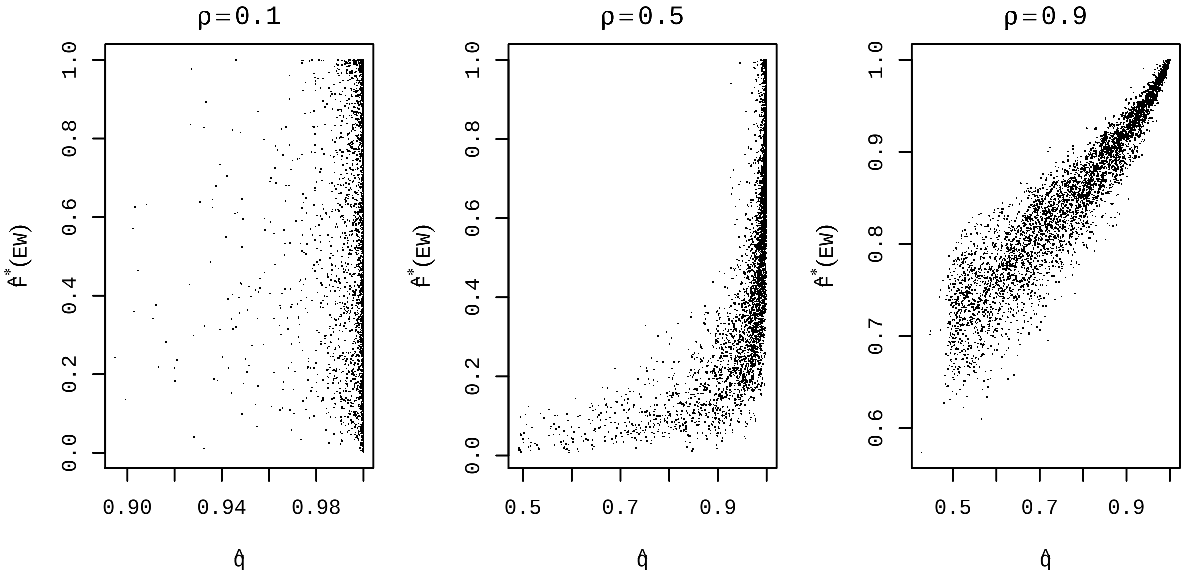

Figure 4.1.10: $\hat{F}_{\widehat{EW}^*}(EW)$ against $\hat{q}$ from the simulation of Table 4.1.3 for M/M/1 with different ρ=.1 , .5, and .9.

Our explanation is that if the true value of ρ is as high as .9, a "good" random sample of $\{I_j\}$ and $\{G_j\}$ will result in higher $\hat{\rho}$ (i.e., $\hat{\rho}$ is close to ρ) and lower value of $\hat{q}$. Thus, even if $\widehat{EW}^*$ will be biased, the true value of EW will not be far from $\widehat{EW}$ and within 90% of $\widehat{EW}^*$. As one can see, a high value of $\hat{F}_{\widehat{EW}^*}(EW)$ only happens for a high value of $\hat{q}$ when $\rho = .9$ so that to capture EW in CI, a higher upper limit percentile is needed and this also holds for ρ=.5 case. Clearly, there is not much of pattern between $\hat{q}$ and $\hat{F}_{\widehat{EW}^*}(EW)$ when ρ=.1 but our strategy of making the limit adjust according to the value of $\hat{q}$ inversely still works.

See Table 4.1.5 for the average coverage probabilities, average lengths, and the sample standard deviations of lengths of the varied percentile limit CI for $n = 25$ and $B\# = 10^6$. Note that the random samples used are the same as in Table 4.1.3 and we add ρ=.3, .5, and .95 cases to see how the performance of CI's are varies over different ρ more thoroughly.

For $n = 100$ case, see Table 4.1.6. In either cases, we can see that the average coverage of varied percentile limit CI are always higher than those of the standard percentile CI.

Table 4.1.5: Average coverage probabilities, the average length, and the sample standard deviation of lengths of standard percentile (BP_1), varied percentile limit, and varied percentile limit (VPL) with lower limit adjustment of EW for different ρ=.1, .3, .5, .7, .9, .95 with different service time distribution of Table 4.1.3 with $n = 25$ and $B\# = 10^6$.

			BP_1			Varied percentile limit			VPL with Lo limit adj		
G	ρ	EW	av.c	av.l	SD	av.c	av.l	SD	av.c	av.l	SD
E	.1	1.11	0.86	0.79	0.28	0.89	0.95	0.38	0.93	1.00	0.39
	.3	1.43	0.86	1.97	2.54	0.89	3.73	6.53	0.92	3.81	6.55
	.5	2.00	0.88	8.32	11.14	0.93	18.41	18.66	0.93	18.50	18.65
	.7	3.33	0.91	18.64	16.24	0.97	33.04	18.57	0.97	33.09	18.51
	.9	10.00	0.82	25.69	16.04	0.95	38.26	14.03	0.95	38.24	13.95
	.95	20.00	0.64	26.63	15.59	0.91	38.58	13.11	0.91	38.55	13.03
G	.1	1.07	0.87	0.44	0.10	0.87	0.51	0.13	0.92	0.54	0.13
	.3	1.29	0.87	0.89	0.67	0.87	1.43	1.93	0.92	1.48	1.94
	.5	1.67	0.89	4.17	5.94	0.90	9.75	10.76	0.93	9.82	10.76
	.7	2.56	0.95	13.10	11.58	0.99	24.31	13.31	0.98	24.35	13.27
	.9	7.00	0.89	21.83	13.02	0.98	31.82	9.94	0.98	31.80	9.86
	.95	13.67	0.75	23.13	12.71	0.96	32.49	9.00	0.96	32.45	8.92
B	.1	0.53	0.88	0.17	0.02	0.88	0.19	0.02	0.92	0.20	0.02
	.3	0.63	0.89	0.32	0.20	0.87	0.48	0.55	0.92	0.50	0.56
	.5	0.80	0.90	1.52	2.09	0.88	3.67	4.15	0.92	3.70	4.15
	.7	1.20	0.96	5.85	5.26	0.99	10.96	5.88	0.98	10.98	5.86
	.9	3.20	0.91	10.26	5.84	0.99	15.16	4.15	0.99	15.15	4.12
	.95	6.20	0.80	11.23	5.80	0.98	15.74	3.71	0.98	15.72	3.67
P	.1	1.06	0.85	0.19	0.11	0.89	0.22	0.14	0.92	0.24	0.14
	.3	1.23	0.89	0.45	0.43	0.89	0.71	1.10	0.93	0.74	1.11
	.5	1.53	0.90	2.39	3.71	0.89	5.91	7.65	0.93	5.94	7.65
	.7	2.24	0.97	9.99	9.61	1.00	19.17	11.30	0.98	19.20	11.27
	.9	5.80	0.93	19.36	11.37	0.99	28.28	8.01	0.99	28.26	7.95
	.95	11.13	0.82	21.06	11.05	0.98	29.42	6.99	0.98	29.38	6.93

Lower limit adjustment

From Table 4.1.5 and 4.1.6, one can see that the nominal coverage probabilities will not be $(1 - \alpha)$ for either of the standard percentile method or varied percentile limit method (we remind you that with 90% confidence, the true coverage probability is within the observed coverage probability ±.01). Also, we would like to remind you that even though the varied percentile limit CI has the higher estimated coverage percentiles than the standard bootstrap percentile CI for all the cases in our simulation study, it was devised to perform well for higher ρ.

Table 4.1.6: Same as Table 4.1.5 with $n = 100$.

			BP$_1$			Varied percentile limit			VPL with Lo limit adj		
G	ρ	EW	av.c	av.l	SD	av.c	av.l	SD	av.c	av.l	SD
E	.1	1.11	0.88	0.40	0.07	0.89	0.47	0.09	0.93	0.50	0.10
	.3	1.43	0.88	0.74	0.23	0.89	0.92	0.32	0.93	0.97	0.33
	.5	2.00	0.88	2.02	1.67	0.90	3.46	4.81	0.93	3.55	4.83
	.7	3.33	0.90	14.37	18.87	0.93	33.19	32.69	0.94	33.33	32.68
	.9	10.00	0.91	45.76	31.24	0.98	74.34	28.39	0.98	74.35	28.25
	.95	20.00	0.83	51.98	30.46	0.97	78.73	24.27	0.97	78.69	24.13
G	.1	1.07	0.89	0.22	0.03	0.90	0.25	0.03	0.94	0.27	0.03
	.3	1.29	0.89	0.36	0.07	0.90	0.43	0.09	0.94	0.46	0.09
	.5	1.67	0.89	0.87	0.41	0.89	1.26	0.94	0.94	1.31	0.95
	.7	2.56	0.90	6.24	8.70	0.90	15.44	18.45	0.93	15.54	18.46
	.9	7.00	0.94	35.62	24.57	0.99	58.63	22.07	0.99	58.64	21.96
	.95	13.67	0.89	43.22	24.77	0.98	64.54	18.34	0.98	64.50	18.20
B	.1	0.53	0.89	0.08	0.00	0.89	0.09	0.01	0.94	0.10	0.01
	.3	0.63	0.89	0.13	0.02	0.89	0.16	0.02	0.93	0.17	0.02
	.5	0.80	0.89	0.33	0.13	0.89	0.47	0.25	0.93	0.49	0.26
	.7	1.20	0.89	2.58	4.08	0.89	6.19	7.91	0.93	6.23	7.91
	.9	3.20	0.94	16.18	11.53	0.99	26.88	10.47	0.99	26.88	10.42
	.95	6.20	0.90	20.19	11.89	0.99	30.17	8.71	0.99	30.15	8.65
P	.1	1.06	0.87	0.10	0.03	0.90	0.12	0.04	0.93	0.12	0.05
	.3	1.23	0.89	0.19	0.08	0.90	0.23	0.11	0.94	0.24	0.11
	.5	1.53	0.89	0.53	0.35	0.88	0.76	0.86	0.93	0.79	0.87
	.7	2.24	0.89	4.06	6.54	0.89	9.87	13.67	0.93	9.94	13.68
	.9	5.80	0.94	29.38	21.45	0.99	49.48	20.03	0.99	49.49	19.94
	.95	11.13	0.90	37.37	22.05	0.99	56.34	16.48	0.99	56.30	16.37

To obtain the nominal level of α, the restriction of (4.1.5), including only $100(1-\alpha)\%$ CI's of $\widehat{EW}^*$, needs to be loosened. We suggest to use the following adjusted lower limit:

$$\widehat{EW}^{*(\alpha/(2\hat{q}))}, \tag{4.1.8}$$

which would be the $\alpha/2$ quantile of unconditioned $\widehat{EW}^*$ since in the bootstrap sampling processes $B\#/\hat{q}$ is the total number of bootstrap random samples of $(\overline{I}^* - \overline{G}^*)$ needed to obtain $B\#$ for which $(\overline{I}^* - \overline{G}^*) > 0$. Because $\hat{q} \leq 1$, the adjusted lower limit (4.1.8) will be always greater than or equal to the lower limit of the standard bootstrap percentile CI.

Note that

$$.9\alpha\hat{q} \le \frac{\alpha}{2\hat{q}} \Longleftrightarrow \hat{q} \le \frac{\sqrt{5}}{3} \approx .745,$$

so that the adjusted lower limit (4.1.8) is higher than the lower limit of plain VPL CI if $\hat{q} < .745$. See Table 4.1.5 and 4.1.6 for the simulation study result of the varied percentile limit (VPL) method with lower limit adjustment. The estimated coverage probabilities are the best among the three methods with comparable average lengths and the standard deviation of the lengths to VPL method and for high ρ, the average lengths of VPL with the lower limit adjustment is slightly smaller than those of VPL CI.

4.1.5 CI of $\operatorname{Var}\mathbf{W}$

We compare the performances of the three methods, the standard bootstrap percentile, the varied percentile limit, and the varied percentile limit with the lower limit adjustment of (4.1.8). See Table 4.1.7 and 4.1.8. Note that the bootstrap random sample size, $B\#$ is 10^6 again. Though VPL with lower limit adjustment still perform the best among the three.

4.2 Estimation of the mean and the variance of busy time periods

From Corollary 50, we obtain the moment estimator of ET and $\operatorname{Var}T$ as

$$\widehat{EB} = \frac{\hat{\mu}_1}{(1-\hat{\rho})} = \frac{\overline{G}}{\left(1-\overline{G}/\overline{I}\right)},$$

$$\widehat{\operatorname{Var}B} = \frac{\hat{\mu}_2 - \hat{\mu}_1^2\,(1-\hat{\rho})}{(1-\hat{\rho})^3} = \frac{\overline{G^2}}{\left(1-\overline{G}/\overline{I}\right)^3} - \frac{\overline{G}^2}{\left(1-\overline{G}/\overline{I}\right)^2}.$$

The a.s. convergence comes from SLLN. As with the moment estimators of EW and $\operatorname{Var}W$, $\widehat{EB}$ and $\widehat{\operatorname{Var}B}$ also has $(1-\hat{\rho})$ as the denominator term. Thus, we can expect that these estimators will have similar behaviors to $\widehat{EW}$ and $\widehat{\operatorname{Var}W}$. In similar way, one can show that $E(\widehat{EB}) = \infty$ and the same goes for $\widehat{\operatorname{Var}B}$.

We compare the performances of the three methods, the standard bootstrap percentile,

Table 4.1.7: Average coverage probabilities, the average length, and the sample standard deviation of lengths of standard percentile (BP_1), varied percentile limit, and varied percentile limit (VPL) with lower limit adjustment of $\operatorname{Var} W$ for different ρ=.1, .3, .5, .7, .9, .95 with different service time distribution of 4.1.3 with $n = 25$ and $B = 10^6$.

			BP_1			Varied percentile limit			VPL with Lo limit adj		
G	ρ	$\operatorname{Var} W$	av.c	av.l	SD	av.c	av.l	SD	av.c	av.l	SD
E	.1	1.23	0.68	1.78	1.83	0.76	2.23	2.62	0.77	2.31	2.64
	.3	2.04	0.77	13.87	88.47	0.84	64.16	401.11	0.87	64.31	401.14
	.5	4.00	0.86	214.11	610.50	0.93	734.33	1231.4	0.93	734.51	1231.4
	.7	11.11	0.90	660.30	1007.1	0.97	1523.4	1352.1	0.97	1523.4	1351.9
	.9	100.00	0.81	983.28	1029.2	0.95	1758.3	1071.4	0.95	1758.2	1071.1
	.95	400.00	0.63	1019.7	1007.2	0.91	1757.8	1004.1	0.91	1757.7	1003.8
G	.1	0.42	0.75	0.45	0.31	0.81	0.55	0.39	0.83	0.58	0.40
	.3	0.73	0.84	2.11	6.55	0.88	7.42	45.24	0.91	7.48	45.25
	.5	1.52	0.88	60.40	225.11	0.90	229.91	448.94	0.93	230.01	448.93
	.7	4.48	0.94	335.39	533.42	0.99	823.04	686.08	0.98	823.09	685.96
	.9	43.00	0.88	699.31	687.31	0.98	1188.0	598.88	0.98	1187.88	598.64
	.95	174.85	0.75	752.51	675.22	0.96	1214.7	552.97	0.96	1214.58	552.73
B	.1	0.07	0.87	0.04	0.01	0.91	0.05	0.01	0.93	0.05	0.01
	.3	0.12	0.90	0.24	0.88	0.90	0.74	5.60	0.93	0.75	5.61
	.5	0.27	0.91	7.90	30.91	0.90	33.97	70.40	0.93	33.98	70.40
	.7	0.85	0.97	68.33	109.16	1.00	166.63	136.97	0.98	166.64	136.94
	.9	8.54	0.91	151.58	143.74	0.99	264.94	115.99	0.99	264.91	115.94
	.95	35.07	0.80	173.27	147.19	0.98	280.32	107.71	0.98	280.28	107.65
P	.1	0.12	0.65	0.18	0.63	0.77	0.23	0.85	0.78	0.23	0.85
	.3	0.30	0.86	0.88	3.96	0.93	2.59	22.50	0.95	2.61	22.51
	.5	0.78	0.90	23.65	105.09	0.90	104.16	291.99	0.94	104.21	291.99
	.7	2.61	0.96	214.28	374.65	1.00	537.03	559.36	0.98	537.07	559.29
	.9	26.95	0.93	551.52	556.78	0.99	932.14	459.53	0.99	932.03	459.35
	.95	110.86	0.82	618.13	556.98	0.98	986.42	410.16	0.98	986.29	409.97

the varied percentile limit, and the varied percentile limit with the lower limit adjustment of (4.1.8). See Table 4.2.1, 4.2.2, 4.2.3, and 4.2.4. Again $B\#$ is set to 10^6.

4.3 Conclusion

The difficulty in the constructing CI for EW, $\operatorname{Var} W$, EB and $\operatorname{Var} B$ comes from two facts. First, having a reciprocal form, the expected values for their sample moments are infinite. Any CI using the empirical moments cannot be used and a CI based on the bootstrap percentile method is the only viable method. Secondly, by imposing the stable condition

Table 4.1.8: Same as table 4.1.5 with $n = 100$.

			BP_1			Varied percentile limit			VPL with Lo limit adj		
G	ρ	Var W	av.c	av.l	SD	av.c	av.l	SD	av.c	av.l	SD
E	.1	1.23	0.80	1.09	0.64	0.86	1.32	0.83	0.88	1.39	0.85
	.3	2.04	0.82	2.51	1.78	0.87	3.38	2.69	0.90	3.51	2.73
	.5	4.00	0.86	13.61	36.51	0.89	47.89	280.00	0.92	48.18	280.09
	.7	11.11	0.90	638.07	1787.2	0.94	2348.7	3673.0	0.95	2349.2	3672.9
	.9	100.00	0.91	3337.3	3751.1	0.98	6764.1	3879.6	0.98	6763.9	3878.6
	.95	400.00	0.84	3933.2	3727.7	0.97	7244.2	3431.3	0.97	7243.7	3430.3
G	.1	0.42	0.84	0.26	0.10	0.88	0.31	0.13	0.91	0.32	0.13
	.3	0.73	0.88	0.58	0.23	0.89	0.75	0.32	0.93	0.78	0.33
	.5	1.52	0.89	2.57	2.95	0.90	5.00	17.00	0.94	5.09	17.02
	.7	4.48	0.90	137.02	511.16	0.90	637.82	1340.6	0.94	638.13	1340.6
	.9	43.00	0.94	2047.1	2325.8	0.99	4208.5	2362.7	0.99	4208.4	2362.0
	.95	174.85	0.89	2699.0	2512.8	0.98	4822.6	2088.9	0.98	4822.2	2088.1
B	.1	0.07	0.89	0.02	0.00	0.91	0.02	0.00	0.94	0.03	0.00
	.3	0.12	0.90	0.07	0.02	0.90	0.09	0.03	0.94	0.09	0.03
	.5	0.27	0.89	0.38	0.31	0.89	0.66	0.84	0.93	0.68	0.85
	.7	0.85	0.90	27.67	121.24	0.89	111.78	263.08	0.93	111.83	263.08
	.9	8.54	0.94	432.59	505.39	0.99	894.00	518.50	0.99	893.98	518.35
	.95	35.07	0.90	597.83	567.45	0.99	1058.27	463.65	0.99	1058.17	463.45
P	.1	0.12	0.71	0.12	0.23	0.80	0.15	0.31	0.81	0.15	0.31
	.3	0.30	0.84	0.29	0.58	0.90	0.39	0.90	0.93	0.40	0.91
	.5	0.78	0.89	1.35	4.66	0.89	2.79	27.83	0.94	2.84	27.84
	.7	2.61	0.89	72.16	371.92	0.89	317.21	867.05	0.93	317.39	867.08
	.9	26.95	0.94	1454.5	1765.0	0.99	3066.5	1841.3	0.99	3066.4	1840.8
	.95	110.86	0.90	2055.6	1965.1	0.99	3703.2	1645.1	0.99	3702.9	1644.5

$\hat{\rho} < 1$, the resulting bootstrap resamples of $\{I_j^*, G_j^*\}_{j=1}^n$ were biased. Thus, the performance of the bootstrap standard percentile method will be hindered by its equal tailed construction.

We showed that by considering variable percentile limit based on the estimation of $q = P(\overline{I} - \overline{G}) > 0$, our percentile methods work better than the bootstrap standard percentile method.

However, we need to mention two observations regarding our suggested CI. When ρ is close to 1, the CI based on a larger sample size works poorly. For $n = 100$, their average length and standard deviation are bigger than those of $n = 25$ case without much improvement on the coverage probability due to the fact that the estimator has the reciprocal form of $1/(1 - \hat{\rho})$, which dominates the estimates as we showed in Figure 4.1.8 (Because a larger

Table 4.2.1: Average coverage probabilities, the average length, and the sample standard deviation of lengths of standard percentile (BP_1), varied percentile limit, and varied percentile limit (VPL) with lower limit adjustment of EB for different ρ=.1, .3, .5, .7, .9, .95 with different service time distribution of Table 4.1.3 with $n = 25$ and $B = 10^6$.

			BP_1			Varied percentile limit			VPL with Lo limit adj		
G	ρ	EB	av.c	av.l	SD	av.c	av.l	SD	av.c	av.l	SD
E	.1	1.11	0.87	0.81	0.26	0.90	0.99	0.35	0.93	1.04	0.37
	.3	1.43	0.88	2.10	2.57	0.89	4.05	6.63	0.93	4.13	6.65
	.5	2.00	0.90	9.05	12.05	0.92	20.19	19.83	0.93	20.28	19.82
	.7	3.33	0.93	20.17	17.26	0.98	36.05	19.28	0.98	36.09	19.21
	.9	10.00	0.85	27.88	17.17	0.96	41.79	14.40	0.96	41.77	14.32
	.95	20.00	0.67	28.80	16.47	0.92	42.05	13.31	0.92	42.02	13.23
G	.1	1.11	0.87	0.48	0.11	0.87	0.57	0.14	0.92	0.60	0.15
	.3	1.43	0.87	1.20	0.99	0.87	2.03	2.88	0.92	2.09	2.89
	.5	2.00	0.89	6.28	9.05	0.89	14.91	16.52	0.92	15.00	16.52
	.7	3.33	0.95	20.07	17.82	0.99	37.34	20.31	0.98	37.40	20.25
	.9	10.00	0.89	33.54	20.05	0.99	48.97	15.07	0.99	48.92	14.95
	.95	20.00	0.76	35.51	19.53	0.96	49.96	13.61	0.96	49.90	13.49
B	.1	0.56	0.89	0.19	0.03	0.88	0.22	0.03	0.92	0.23	0.03
	.3	0.71	0.89	0.48	0.34	0.87	0.77	0.96	0.92	0.79	0.96
	.5	1.00	0.89	2.55	3.57	0.88	6.30	7.18	0.92	6.34	7.18
	.7	1.67	0.97	9.97	9.10	0.99	18.71	10.17	0.97	18.74	10.14
	.9	5.00	0.91	17.31	9.96	0.99	25.61	7.20	0.99	25.59	7.14
	.95	10.00	0.80	18.90	9.87	0.98	26.52	6.46	0.98	26.48	6.40
P	.1	1.11	0.87	0.22	0.10	0.89	0.27	0.13	0.93	0.28	0.14
	.3	1.43	0.89	0.72	0.66	0.88	1.19	1.78	0.93	1.22	1.79
	.5	2.00	0.90	4.31	6.57	0.89	10.74	13.24	0.93	10.81	13.25
	.7	3.33	0.97	18.61	17.64	1.00	35.83	20.32	0.97	35.88	20.27
	.9	10.00	0.93	36.49	20.97	0.99	53.43	14.24	0.99	53.38	14.12
	.95	20.00	0.82	39.87	20.51	0.98	55.78	12.46	0.98	55.70	12.33

sample size n results in a smaller the variance of $\hat{\rho}$, we would have a shorter 90% CI for $1 - \hat{\rho}$, which would be more close to 0 comparing to the $n = 25$ case).

Also note that for $\operatorname{Var} W$, we found that $B\#$ should be greater that 10^6 at least in our case. For example, with $B\# = 10^5$, the observed coverage probabilities were .019, .075, and .075 for the standard percentile CI, the VPL, and the VPL with lower limit adjustment, respectively when G~Beta(2,2) and $\rho = .3$ with $n = 100$. When we set $B\# = 10^6$, we observed the acceptable coverage probabilities for all three different CI's. For EW, this phenomenon was not observed. The observed result were comparable for both $B\# = 10^5$ and $B\# = 10^6$, which is explained by $\widehat{EW} = O((1-\hat{\rho})^{-1})$ but $\widehat{\operatorname{Var} W} = O((1-\hat{\rho})^{-2})$.

Table 4.2.2: Same as Table 4.2.1 with $n = 100$.

			BP_1			Varied percentile limit			VPL with Lo limit adj		
G	ρ	EB	av.c	av.l	SD	av.c	av.l	SD	av.c	av.l	SD
E	.1	1.11	0.88	0.41	0.07	0.89	0.47	0.08	0.93	0.51	0.08
	.3	1.43	0.88	0.74	0.19	0.88	0.93	0.27	0.93	0.98	0.28
	.5	2.00	0.89	2.04	1.64	0.89	3.51	4.79	0.93	3.60	4.81
	.7	3.33	0.91	14.67	19.28	0.93	33.96	33.17	0.94	34.11	33.17
	.9	10.00	0.92	46.81	31.89	0.99	76.11	28.58	0.99	76.12	28.43
	.95	20.00	0.84	53.12	31.06	0.97	80.60	24.23	0.97	80.56	24.07
G	.1	1.11	0.89	0.24	0.03	0.90	0.27	0.03	0.94	0.29	0.03
	.3	1.43	0.89	0.46	0.09	0.89	0.56	0.11	0.94	0.59	0.12
	.5	2.00	0.90	1.24	0.61	0.89	1.81	1.41	0.94	1.88	1.43
	.7	3.33	0.91	9.34	13.10	0.90	23.24	27.81	0.93	23.38	27.83
	.9	10.00	0.94	53.75	37.11	0.99	88.51	33.25	0.99	88.52	33.09
	.95	20.00	0.89	65.20	37.33	0.98	97.41	27.55	0.98	97.35	27.36
B	.1	0.56	0.89	0.09	0.01	0.89	0.11	0.01	0.94	0.11	0.01
	.3	0.71	0.89	0.19	0.03	0.89	0.23	0.04	0.93	0.24	0.04
	.5	1.00	0.89	0.54	0.23	0.89	0.77	0.44	0.93	0.80	0.45
	.7	1.67	0.89	4.36	6.91	0.89	10.49	13.43	0.93	10.56	13.44
	.9	5.00	0.94	27.17	19.40	0.99	45.13	17.68	0.99	45.14	17.60
	.95	10.00	0.90	33.80	19.97	0.99	50.51	14.71	0.99	50.48	14.60
P	.1	1.11	0.88	0.11	0.03	0.90	0.13	0.04	0.94	0.14	0.04
	.3	1.43	0.89	0.28	0.07	0.88	0.35	0.09	0.93	0.37	0.10
	.5	2.00	0.89	0.91	0.46	0.88	1.31	1.07	0.92	1.36	1.08
	.7	3.33	0.89	7.40	11.64	0.89	18.05	24.64	0.93	18.18	24.67
	.9	10.00	0.94	55.03	40.02	0.99	92.73	37.24	0.99	92.76	37.08
	.95	20.00	0.90	70.27	41.42	0.99	105.94	30.72	0.99	105.88	30.50

Table 4.2.3: Average coverage probabilities, the average length, and the sample standard deviation of lengths of standard percentile (BP_1), varied percentile limit, and varied percentile limit (VPL) with lower limit adjustment of $\mathrm{Var}\, B$ for different ρ=.1, .3, .5, .7, .9, .95 with different service time distribution of Table 4.1.3 with $n = 25$ and $B = 10^6$.

			BP_1			Varied percentile limit			VPL with Lo limit adj		
G	ρ	$\mathrm{Var}\, B$	av.c	av.l	SD	av.c	av.l	SD	av.c	av.l	SD
E	.1	1.51	0.75	2.73	2.94	0.82	3.89	5.28	0.84	4.00	5.32
	.3	3.79	0.84	284.91	4760.8	0.89	2792.0	25565	0.92	2792.3	25565
	.5	12.00	0.89	14063	62521	0.93	57565	113893	0.94	57566	113893
	.7	62.96	0.92	52656	103650	0.98	139543	135392	0.98	139543	135390
	.9	1900	0.85	88127	116888	0.96	174655	111789	0.96	174654	111787
	.95	15600	0.66	91303	112208	0.92	175786	103849	0.92	175785	103848
G	.1	0.59	0.81	0.78	0.51	0.86	1.05	0.75	0.89	1.09	0.76
	.3	1.85	0.87	25.32	269.99	0.89	340.17	4564.9	0.92	340.33	4565.0
	.5	6.67	0.89	5968.4	37357	0.89	27932	71078	0.93	27933	71077
	.7	38.27	0.95	45388	100043	0.99	126104	125458	0.99	126104	125457
	.9	1233	0.89	111207	141292	0.99	202090	119257	0.99	202088	119254
	.95	10267	0.76	121785	139573	0.96	208998	111170	0.96	208996	111167
B	.1	0.10	0.89	0.09	0.03	0.90	0.12	0.05	0.93	0.13	0.06
	.3	0.36	0.90	3.17	50.56	0.89	34.40	751.16	0.93	34.43	751.17
	.5	1.40	0.90	798.82	5693.4	0.89	4543.6	12885	0.93	4543.7	12885
	.7	8.33	0.97	10975	24352	1.00	29872	30780	0.98	29872	30779
	.9	275.0	0.92	27508	34815	0.99	51902	28692	0.99	51902	28692
	.95	2300	0.80	32319	36149	0.98	55709	27201	0.98	55709	27201
P	.1	0.23	0.83	0.34	0.80	0.90	0.48	1.22	0.93	0.49	1.22
	.3	1.07	0.90	11.08	207.07	0.90	126.88	3222.5	0.94	126.96	3222.6
	.5	4.53	0.91	2552.1	21885	0.89	14088	45189	0.93	14088	45189
	.7	28.40	0.97	37802	87392	1.00	106030	114361	0.98	106030	114359
	.9	966.7	0.93	118856	148916	0.99	213749	113517	0.99	213747	113514
	.95	8133	0.82	136977	152398	0.98	231727	106130	0.98	231724	106127

Table 4.2.4: Same as Table 4.2.3 with $n = 100$.

				BP_1			Varied percentile limit			VPL with Lo limit adj	
G	ρ	$\mathrm{Var}\,T$	av.c	av.l	SD	av.c	av.l	SD	av.c	av.l	SD
E	.1	1.51	0.83	1.41	0.72	0.87	1.75	0.95	0.90	1.84	0.98
	.3	3.79	0.87	6.37	4.46	0.89	9.63	7.89	0.93	9.89	8.01
	.5	12.00	0.88	157.88	1410.7	0.90	2097.7	29765	0.94	2098.8	29766
	.7	62.96	0.91	73789	321887	0.93	341606	659446	0.94	341609	659444
	.9	1900	0.92	537288	782133	0.98	1225335	821770	0.98	1225331	821758
	.95	15600	0.84	655950	806161	0.97	1345360	737670	0.97	1345353	737657
G	.1	0.59	0.87	0.38	0.13	0.89	0.46	0.16	0.93	0.49	0.17
	.3	1.85	0.89	2.10	0.96	0.90	2.95	1.48	0.94	3.05	1.52
	.5	6.67	0.90	25.26	84.94	0.89	103.77	1598.6	0.93	104.29	1598.7
	.7	38.27	0.91	21801	160336	0.90	145351	404630	0.93	145354	404630
	.9	1233	0.94	589325	883035	0.99	1361443	926579	0.99	1361438	926565
	.95	10267	0.89	823988	999743	0.98	1615795	850111	0.98	1615785	850094
B	.1	0.10	0.89	0.04	0.01	0.90	0.05	0.01	0.93	0.05	0.01
	.3	0.36	0.90	0.34	0.14	0.89	0.48	0.21	0.93	0.49	0.22
	.5	1.40	0.89	4.16	5.84	0.89	9.87	26.11	0.93	9.97	26.16
	.7	8.33	0.89	6055.4	44202	0.89	30476	97892	0.93	30477	97891
	.9	275.0	0.94	146314	226194	0.99	337517	240478	0.99	337516	240475
	.95	2300	0.90	215933	267135	0.99	413778	225346	0.99	413776	225341
P	.1	0.23	0.82	0.18	0.24	0.88	0.23	0.34	0.91	0.24	0.34
	.3	1.07	0.89	1.15	1.06	0.89	1.61	1.88	0.93	1.67	1.90
	.5	4.53	0.89	14.33	72.21	0.88	63.53	1398.4	0.93	63.86	1398.5
	.7	28.40	0.89	14802	132149	0.89	91525	328265	0.93	91528	328265
	.9	966.7	0.94	573151	924861	0.99	1351858	977923	0.99	1351854	977908
	.95	8133	0.90	872311	1093736	0.99	1713439	926400	0.99	1713428	926380

Bibliography

[1] J. Abate, G. L. Choudhury, and W. Whitt. Waiting-time tail probabilities in queues with long-tail service-time distributions. *Queueing Systems*, 16(3-4):311–338, 1994.

[2] J. Abate, G. L. Choudhury, and W. Whitt. Exponential approximations for tail probabilities in queues, .1. waiting-times. *Operations Research*, 43(5):885–901, September 1995.

[3] J Abate and W Whitt. Solving probability transform functional-equations for numerical inversion. *Operations Research Letters*, 12(5):275–281, NOV 1992.

[4] Joseph Abate, Gagan L. Choudhur, and Ward Whitt. *Computational Probability*, chapter 8 An Introduction to Numerical Transform Inversion and its Application to Probability Models, pages 257–323. Kluwer, Boston, 1999. 0-7923-8617-5.

[5] Joseph Abate and Ward Whitt. Approximations for the M/M/1 busy-period distribution. In *Queueing Theory and its Applications*, volume 7 of *CWI Monogr.*, pages 149–191. North-Holland, Amsterdam, 1988.

[6] Joseph Abate and Ward Whitt. The Fourier-series method for inverting transforms of probability distributions. *Queueing Systems*, 10(1):5–87, January 1992.

[7] Joseph Abate and Ward Whitt. Explicit M/G/1 waiting-time distributions for a class of long-tail service-time distributions. *Operations Research Letters*, 25(1):25 – 31, 1999.

[8] Søren Asmussen. *Ruin Probabilities*, volume 2 of *Advanced Series on Statistical Science & Applied Probability*. World Scientific Publishing Co. Inc., River Edge, NJ, 2000.

[9] Søren Asmussen and Tomasz Rolski. Computational methods in risk theory: A matrix-algorithmic approach. *Insurance: Mathematics and Economics*, 10(4):259 – 274, 1992.

[10] Ole Barndorff-Nielsen. *Information and Exponential Families in Statistical Theory.* John Wiley & Sons Ltd., Chichester, 1978. Wiley Series in Probability and Mathematical Statistics.

[11] Patrick Billingsley. *Convergence of Probability Measures.* Wiley Series in Probability and Statistics: Probability and Statistics. John Wiley & Sons Inc., New York, second edition, 1999. A Wiley-Interscience Publication.

[12] J. P. C. Blanc. On the numerical inversion of busy-period related transforms. *Operations Research Letters*, 30(1):33–42, February 2002.

[13] Jean-Louis Bon and Vladimir Kalashnikov. Some estimates of geometric sums. *Statistics & Probability Letters*, 55(1):89 – 97, 2001.

[14] James G. Booth and Andrew T. A. Wood. An example in which the Lugannani-Rice saddlepoint formula fails. *Statistics & Probability Letters*, 23(1):53–61, April 1995.

[15] OJ Boxma and JW Cohen. The M/G/1 queue with heavy-tailed service time distribution. *ieee Journal on Selected Areas in Communications*, 16(5):749–763, JUN 1998.

[16] Lawrence D. Brown, T. Tony Cai, and Anirban DasGupta. Interval estimation for a binomial proportion. *Statist. Sci.*, 16(2):101–133, 2001. With comments and a rejoinder by the authors.

[17] Ronald W. Butler. *Saddlepoint Approximations with Applications (Cambridge Series in Statistical and Probabilistic Mathematics).* Cambridge University Press, 2007.

[18] J. Cai and J. Garrido. A unified approach to the study of tail probabilities of compound distributions. *Journal of Applied Probability*, 36(4):1058–1073, December 1999.

[19] Y. K. Chu and J. C. Ke. Confidence intervals of mean response time for an M/G/1 queueing system: Bootstrap simulation. *Applied Mathematics and Computation*, 180(1):255–263, September 2006.

[20] Alan M. Cohen. *Numerical Methods for Laplace Transform Inversion*, volume 5 of *Numerical Methods and Algorithms*. Springer, New York, 2007.

[21] Robert B. Cooper. *Introduction to Queueing Theory (Second Edition)*. Elsevier/North-Holland [Elsevier Science Publishing Co., New York; North-Holland Publishing Co., Amsterdam], 1981.

[22] D. R. Cox and Walter L. Smith. *Queues*. Methuen's Monographs on Statistical Subjects. Methuen & Co. Ltd., London, 1961.

[23] S. Csörgő. The empirical moment generating function. In *Nonparametric statistical inference, Vol. I, II (Budapest, 1980)*, volume 32 of *Colloq. Math. Soc. János Bolyai*, pages 139–150. North-Holland, Amsterdam, 1982.

[24] H. E. Daniels. Saddlepoint approximations in statistics. *Ann. Math. Statist.*, 25:631–650, 1954.

[25] Brian Davies. *Integral Transforms and their Applications*, volume 41 of *Texts in Applied Mathematics*. Springer-Verlag, New York, third edition, 2002.

[26] A. C. Davison and D. V. Hinkley. *Bootstrap methods and their application*, volume 1 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 1997.

[27] Anthony C. Davison and David V. Hinkley. Saddlepoint approximations in resampling methods. *Biometrika*, 75(3):417–431, sep 1988.

[28] Bradley Efron. Nonparametric standard errors and confidence intervals. *Canad. J. Statist.*, 9(2):139–172, 1981. With discussion and a reply by the author.

[29] Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*, volume 57 of *Monographs on Statistics and Applied Probability.* Chapman and Hall, New York, 1993.

[30] William Feller. *An Introduction to Probability Theory and its Applications. Vol. II.* Second edition. John Wiley & Sons Inc., New York, 1971.

[31] Andrey Feuerverger. On the empirical saddlepoint approximation. *Biometrika*, 76(3):457–464, Sept 1989.

[32] Gaver, D. P. and Jacobs, P. A. Nonparametric estimation of the probability of a long delay in the M/G/1 queue. *Journal of the Royal Statistical Society. Series B (Methodological)*, 50(3):392–402, 1988.

[33] Jan Grandell. Empirical bounds for ruin probabilities. *Stochastic Processes and their Applications*, 8(3):243 – 255, 1979.

[34] D. Gross, J.F. Shortle, M.J. Fischer, and D.M.B. Masi. Difficulties in simulating queues with pareto service. In *Simulation Conference, 2002. Proceedings of the Winter*, volume 1, pages 407–415 vol.1, Dec. 2002.

[35] Allan Gut. *Probability: a Graduate Course.* Springer Texts in Statistics. Springer, New York, 2005.

[36] Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Fundamentals Of Convex Analysis.* Grundlehren Text Editions. Springer-Verlag, Berlin, 2001. Abridged version of *Convex Analysis and Minimization Algorithms. I* [Springer, Berlin, 1993; MR1261420 (95m:90001)] and *II* [ibid.; MR1295240 (95m:90002)].

[37] R. J. Hyndman. Computing and graphing highest density regions. *American Statistician*, 50(2):120–126, May 1996.

[38] Frank Jones. *Lebesgue Integration on Euclidean Space.* Jones and Bartlett Publishers, Boston, MA, 1993.

[39] L. Kilian. Small-sample confidence intervals for impulse response functions. *Review Of Economics And Statistics*, 80(2):218–230, May 1998.

[40] J. H. Kim, P. Silvapulle, and R. J. Hyndman. Half-life estimation based on the bias-corrected bootstrap: A highest density region approach. *Computational Statistics & Data Analysis*, 51(7):3418–3432, April 2007.

[41] Leonard Kleinrock. *Queueing Systems. Volume 1: Theory.* Wiley-Interscience, 1 edition, 1 1975.

[42] E. L. Lehmann and George Casella. *Theory of Point Estimation.* Springer Texts in Statistics. Springer-Verlag, New York, second edition, 1998.

[43] Robert Lugannani and Stephen Rice. Saddle point approximation for the distribution of the sum of independent random variables. *Advances in Applied Probability*, 12(2):475–490, Jun 1980.

[44] Bryan F. J. Manly. *Randomization, Bootstrap and Monte Carlo Methods in Biology.* Chapman & Hall/CRC Texts in Statistical Science Series. Chapman & Hall/CRC, Boca Raton, FL, third edition, 2007.

[45] P. Massart. The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *The Annals of Probability*, 18(3):1269–1283, jul 1990.

[46] James Munkres. *Topology (2nd Edition).* Prentice Hall, 1999.

[47] Marcel F. Neuts. *Matrix-geometric Solutions in Stochastic Models, An algorithmic approach*, volume 2 of *Johns Hopkins Series in the Mathematical Sciences.* Johns Hopkins University Press, Baltimore, Md., 1981.

[48] Marcel F. Neuts. *Structured Stochastic Matrices of $M/G/1$ Type and Their Applications*, volume 5 of *Probability: Pure and Applied.* Marcel Dekker Inc., New York, 1989.

[49] A. G. Pakes. Tails of waiting-time distributions. *Journal of Applied Probability*, 12(3):555–564, 1975.

[50] Walter W. Piegorsch. Sample sizes for improved binomial confidence intervals. *Comput. Statist. Data Anal.*, 46(2):309–316, 2004.

[51] Susan M. Pitts. Nonparametric estimation of the stationary waiting time distribution function for the GI/G/1 queue. *The Annals of Statistics*, 22(3):1428–1446, sep 1994.

[52] Colin M. Ramsay. Exact waiting time and queue size distributions for equilibrium M/G/1 queues with Pareto service. *Queueing Systems*, 57(4):147–155, DEC 2007.

[53] Sidney I. Resnick. *Extreme Values, Regular Variation, and Point Processes (Applied Probability)*. Springer, 1987.

[54] A. Wayne Roberts and Dale E. Varberg. *Convex Functions*. Academic Press [A subsidiary of Harcourt Brace Jovanovich, Publishers], New York-London, 1973. Pure and Applied Mathematics, Vol. 57.

[55] Tomasz Rolski, Hanspeter Schmidli, Volker Schmidt, and Jozef Teugels. *Stochastic Processes for Insurance and Finance*. Wiley Series in Probability and Statistics. John Wiley & Sons Ltd., Chichester, 1999.

[56] W. A. Rosenkrantz. Calculation of the Laplace transform of the length of the busy period for the M/G/1 queue via martingales. *Annals of Probability*, 11(3):817–818, 1983.

[57] Kenneth A. Ross. *Elementary Analysis: the Theory of Calculus*. Springer-Verlag, New York, 1980. Undergraduate Texts in Mathematics.

[58] Walter Rudin. *Principles of Mathematical Analysis*. McGraw-Hill Book Co., New York, third edition, 1976. International Series in Pure and Applied Mathematics.

[59] T. Sakurai. Approximating M/G/1 waiting time tail probabilities. *Stochastic Models*, 20(2):173–191, 2004.

[60] Jun Shao and Dong Sheng Tu. *The Jackknife and Bootstrap.* Springer Series in Statistics. Springer-Verlag, New York, 1995.

[61] Galen R. Shorack. *Probability for statisticians.* Springer Texts in Statistics. Springer-Verlag, New York, 2000.

[62] J. F. Shortle, P. H. Brill, M. J. Fischer, D. Gross, and D. M. B. Masi. An algorithm to compute the waiting time distribution for the M/G/1 queue. *Informs Journal on Computing*, 16(2):152–161, 2004.

[63] John F. Shortle, Martin J. Fischer, and Percy H. Brill. Waiting-time distribution of M/$D-n$/1 queues through numerical Laplace inversion. *Informs Journal on Computing*, 19(1):112–120, WIN 2007.

[64] Walter L. Smith. On the distribution of queueing times. *Mathematical Proceedings of the Cambridge Philosophical Society*, 49(03):449–461, 1953.

[65] William F. Stout. *Almost Sure Convergence.* Academic Press [A subsidiary of Harcourt Brace Jovanovich, Publishers], New York-London, 1974. Probability and Mathematical Statistics, Vol. 24.

[66] Robert S. Strichartz. *The Way of Analysis.* Jones & Bartlett Pub, 1995.

[67] Lajos Takács. *Introduction to the theory of queues.* University Texts in the Mathematical Sciences. Oxford University Press, New York, 1962.

[68] Henk C. Tijms. *A First Course in Stochastic Models.* John Wiley & Sons Ltd., Chichester, 2003.

[69] A. W. van der Vaart. *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics.* Cambridge University Press, Cambridge, 1998.

[70] Whitmore, G. A. and Seshadri, V. A heuristic derivation of the inverse gaussian distribution. *The American Statistician*, 41(4):280–281, nov 1987.

[71] G. E. Willmot. On a class of approximations for ruin and waiting time probabilities. *Operations Research Letters*, 22(1):27–32, February 1998.

[72] Gordon E. Willmot and X. Sheldon Lin. *Lundberg Approximations for Compound Distributions with Insurance Applications*, volume 156 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 2001.

[73] A. T. A. Wood, J. G. Booth, and R. W. Butler. Saddlepoint approximations to the cdf of some statistics with nonnormal limit distributions. *Journal Of The American Statistical Association*, 88(422):680–686, June 1993.