BATCHES, BURSTS, AND SERVICE SYSTEMS

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Andrew Monroe Daw

August 2020

BATCHES, BURSTS, AND SERVICE SYSTEMS

Andrew Monroe Daw, Ph.D.

Cornell University 2020

In a plethora of natural phenomena, events occur in flurries, clusters, or bunches. Modern service systems are no exception to this. This can be by design, such as in batches of jobs being sent to a data center for processing, or simply by circumstance, such as in bursts of newly infected flu patients arriving to a health clinic or in the virality of new interactions with a popular social media post. This thesis is concerned with the modeling, exploration, and analysis of these batch and burst arrival processes through the lens of applied probability. Often, this builds on the idea of self-exciting Hawkes process, in which each arrival increases the likelihood of another arrival occurring soon after, forming quick bursts of several successive arrivals. By comparison, batches are taken to be truly simultaneous, with multiple entities entering the system at precisely the same epoch. In the course of this dissertation, batches are both compared to bursts and used as tools to develop deeper understanding of bursts. These objects are also both applied in a variety of settings, most notably in the problem of staffing teleoperation support systems for autonomous vehicles. This analysis reveals that batches and bursts have a pronounced effect on service systems, and thus must be addressed.

BIOGRAPHICAL SKETCH

Andrew Daw was born and raised in Jacksonville, Florida, the largest city in the conterminous United States in terms of area, or the largest city in the entire United States in terms of area not located in Alaska. He received his undergraduate education from the University of Florida, majoring in Industrial and Systems Engineering. Upon completion of his bachelors degree in 2015, he began his doctoral studies in the School of Operations Research & Information Engineering at Cornell University, where he has been advised by Professor Jamol Pender.

ACKNOWLEDGEMENTS

These acknowledgements have been written during social distancing in Spring 2020 and, given those circumstances and the length of this dissertation overall, it doesn't seem like they should stand apart from the rest of the text for their brevity. I certainly have many people to thank, and I will inevitably forget to thank many more. Before naming specific people, I want to express my gratitude to the National Science Foundation for their support through the Graduate Research Fellowships Program under grant DGE-1650441.

It has been the privilege of my academic career to have been a part of Cornell ORIE and be around so many remarkable people. First and foremost, I am incredibly fortunate to have been advised by Jamol Pender, who has offered me much more than I ever expected to get from a PhD. His patient guidance has surely played a crucial role in my growth as a researcher, teacher, and person, and I am very grateful for the many opportunities that he has provided to me. I likely cannot express the full extent of my admiration and respect, so I hope to at least convey my fullest gratitude. In addition to my advisor, I am also grateful to have had Jim Dai and Sid Resnick as my two minor committee members. Both have been quick to offer advice, encouragement, and opportunity, and I feel very privileged to have had access to their expertise and attention. It speaks to ORIE's strengths as a community that I feel I have reason to thank virtually every faculty member; I am grateful to have been part of such an open and energized environment. I would like to specifically acknowledge Kathryn Caggiano, Brenda Dietrich, Dave Goldberg, Itai Gurvich, Mark Lewis, David Shmoys, and David Williamson for their routine advice and insights, and I am also particularly grateful to Shane Henderson, who has been exceptionally supportive across my Cornell career.

Another fantastic aspect of the ORIE community has been the wonderful staff members, and I would like to acknowledge two in particular. First, I am so grateful to have met Dionysios Panagitsas. It's hard to imagine my time at Cornell without Dennis's frequent advice, fantastic stories, or constant encouragement, and I feel very lucky to have been *just* Greek enough to have been the recipient of this. Secondly, I am also very grateful to have been part of this department when Tara Woodard was, and I'm sure that I'm not the only PhD student to say that Tara was who first made me feel at home in ORIE. To that end, I would also like to acknowledge Dean Jan Allen, who was similarly welcoming for me at Cornell overall. Outside of Cornell, I would like to thank the researchers with whom I have been fortunate to collaborate; in particular I am grateful to Robert Hampshire at the University of Michigan, Brian Fralix at Clemson University, and Antonio Castellanos and Galit Yom-Tov at the Technion. I am also appreciative of the many people at the University of Florida who helped set me on this path, with particular acknowledgement to Cole Smith for his continuing advice and guidance.

Perhaps the most difficult part of writing these acknowledgments has been deciding how to properly express my gratitude for my fellow PhD students. I've benefited immensely from having all of you as peers; it's been humbling and inspiring to be around each of you. I was fortunate to join ORIE alongside a fantastic 2015 cohort. From talking about everything from dogs and travel to diversity and inclusion with my long-running officemate Pamela Badian-Pessot, to the most reliably friendly check-ins with Tom Fei; from picnic planning and joyful conference reunions with Mika Sumida, to learning from my ever impressive Pender-lab-mate Shuang Tao; and from Alberto Vera's casual brilliance to Matthew Zalesak's genuine kindness and unending supply of day-brightening

tidbits, there are so many reasons that I am grateful to have gone through my PhD alongside this group. This is especially true for Sam Gutekunst, who I am undoubtedly lucky to have had as a roommate and as a model for success in all facets of a PhD, and who is remarkably willing to serve as a group's method of social cohesion, even if that is at his own expense.

The 2016 cohort was also particularly impactful for me. I'm glad to have had them in ORIE for nearly my entire time there, and it's easy to find reasons why: Woo-Hyung Cho's exemplar leadership and consistent encouragement, Lijun Ding's coordination of the mentoring program, Ben Grimmer's routine reminders to think like a computer scientist, Chamsi Hssaine's friendship, thoughtfulness, and excellent recommendations, Angela Zhou's inspirational thoughts on virtually any subject, Amy Zhang's kindness. I'm also grateful to have gotten to know Yilun Chen, who has been an inspiration even before I arrived at ORIE. There are also many people worth celebrating in the 2017 cohort, but I'd like to give a special shout-out to Vasilis Charisopoulos, with whom I am honored and grateful to have co-taught. Similarly for the 2018's, I want to thank Sean Sinclair for being a better mentor than me. For the 2019's, I wish I had gotten the chance to know you more this final semester, but I look forward to seeing the great things that you do. There are several students outside the department that I'm grateful to have met as well, including Daniel Freund, Sergio Palomo, Sophia Novitzky, and Faisal Alkaabneh.

There were also many fantastic people in ORIE when I joined. I relished bouncing ideas with excellent probability researchers and teachers like David Eckman, Emily Fischer, and Tiandong Wang, celebrating Fridays through barn trips with friends like James Dong, Steve Pallone, Pat Steele, Julian Sun, and Calvin Wylie, exchanging music recommendations and frisbee passes with Cory

Girard, and trying (but failing) to catch them all with Venus Lo. I am additionally very appreciative for the excellent standards set by these senior students and the corresponding mentorship that they passed down, such as from Anton Braverman, Michael Choi, Daniel Fleischman, and Alice Paul. Finally, I was also especially fortunate that Dave Lingenbrink was one of the very first people I met at Cornell, and even more so that that friendship has spanned bowling alleys, people movers, and last chance power drives.

These acknowledgements would undoubtedly be incomplete without addressing the wonderful and unwavering motivation I've had from my family. I am deeply grateful to my parents, Keith and Ashley Daw, for their love and encouragement, and for the priority that they have always placed on my education. I am also very appreciative for the constant uplifting from my two younger sisters, Elizabeth and Caroline, whose brother I am proud to be. If it takes a village to do a PhD, I've certainly had mine. I would be remiss if I did not also acknowledge the decade-plus of support I've received from my Stamatogiannakis family: Ellie, Emmanuel, Anna, Alixe, Nick, and Anna Stark.

Finally, and most of all, thank you to Maria. Thank you for taking my dreams in among your own and chasing them alongside me. You have always given me a smile whenever I've needed it most, and you have set the bar for purpose and mission in the ways that you serve your classrooms and communities. I am so grateful to have you as a partner in life, and I can't wait to see where our next chapter goes.

"We are caught in an inescapable network of mutuality, tied in a single garment of destiny. Whatever affects one directly, affects all indirectly."

King (1963)

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

At the heart of this dissertation is the idea of bursts, meaning temporally clustered flurries of arrivals to a system. By comparison to periodicity or seasonality, bursts need not follow a regular schedule. The times that these clusters occur are unpredictable, and many new entities could arrive to the system within a short period of time without advance warning. Empirical evidence shows that this phenomenon is ubiquitous; data-driven examples include call centers (e.g. Glynn et al. (2019); Ibrahim et al. (2016); L'Ecuyer et al. (2018)), financial activity (e.g. Aït-Sahalia et al. (2015); Azizpour et al. (2016)), and communication and social media (e.g. Farajtabar et al. (2017); Malmgren et al. (2008); Rizoiu et al. (2017)). Motivated by such observations, this thesis is concerned with understanding the structure of bursts of arrivals, providing managerial insights on how to handle and prepare for their occurrence, and developing new techniques for analyzing the phenomenon. Primarily, the bursts discussed herein stem from two types of sources: exogenously driven stimuli and endogenously created self-excitation. For the former, this means that external events create a rush of new arrivals to a system, like when many people call to report a single emergency incident, as studied in L'Ecuyer et al. (2018). For the latter, this means that the occurrence of new activity increases the likelihood of additional activity soon afterwards, like how one firm defaulting increases the risk that other firms will default as well, as is studied in Azizpour et al. (2016). This idea of *self-exciting point processes* originated in Hawkes (1971), which proposed an arrival process model such that "the current intensity of events is determined by events in the past."

Drawing upon motivations from service systems, the following chapters both employ and extend these arrival processes. This involves using batches of arrivals as both a subject of research and a tool for analysis. For example, in Chapter 6 we study the problem of staffing a remote support center for the teleoperation assistance of driverless cars. In this problem, autonomous vehicles disengage their self-driving capabilities and receive human help, provided there are enough operators available. Due to the rapid bursts of requests and the simulation-based crowd-sourcing style of support, we identify batches as a salient feature of the associated queueing model. We show that increasing the batch size raises the necessary staffing level linearly, whereas increasing the arrival rate is known to only have a sub-linear effect on the staffing level, yielding a fundamental contribution for queueing theory. As a practical problem contribution, we use national driving data to show that a remote operation center is more efficient than having in-car safety drivers for every vehicle, producing a key insight for managing fleets of driverless cars or trucks at scale. These findings are powered by our methodological contributions, as we have proved a novel *batch scaling* limit theorem connecting the multi-server queueing models to storage processes.

This multi-server batch scaling methodology is a generalization of infinite server techniques developed in Chapter 3 and 4. In Chapter 4, this yields an important connection between the self-exciting Hawkes process and a linear birth-death-immigration process, which can be thought of as an infinite server queue with state-dependent arrival rates. This model offers many of the same properties as the Hawkes process but with greater analytical tractability. Moreover, this parsimonious pre-limit object offers fundamental understanding into the concept of self-excitement since it provides a formal link to conceptually

similar processes, such as epidemic models. This work can be seen as an extension of the Hawkes process driven infinite server queues studied in Chapter 2, as the model in Chapter 4 is both excited by arrivals and inhibited by departures. Thus, this process pioneers the idea of *ephemeral self-excitement*, in which an arriving entity only increases the rate of new arrivals so long as it remains in the system.

This dissertation has also led to additional insights into broader classes of stochastic processes beyond the motivating burst arrival models. For example, by recognizing an underlying nesting structure to differential equations for the moments of the Hawkes process intensity, we have identified a computationally efficient, matrix-based calculation in Chapter 5 that is able to yield higher order closed form moments than what was previously known in the literature. This technique can be used to calculate the moments of many other Markov processes, and we provide example applications for popular models such as Itô diffusions, growth-collapse processes, and shot noise processes. Before proceeding with the contents of this dissertation, let us first review the self-exciting Hawkes process, as this stochastic model serves as the subject of much of the following research.

1.1 The Hawkes Arrival Process

Introduced and pioneered through the series of papers Hawkes (1971); Hawkes (1971); Hawkes and Oakes (1974), the Hawkes process is a stochastic intensity point process in which the current rate of arrivals is dependent on the history of arrival process itself. Formally, this is defined as follows: let (λ_t, N_t) be an

intensity and counting process pair such that

$$\mathrm{P}\left(N_{t+\Delta} - N_t = 1 \mid \mathcal{F}_t^N\right) = \lambda_t \Delta + o(\Delta),$$
$$\mathrm{P}\left(N_{t+\Delta} - N_t > 1 \mid \mathcal{F}_t^N\right) = o(\Delta),$$
$$\mathrm{P}\left(N_{t+\Delta} - N_t = 0 \mid \mathcal{F}_t^N\right) = 1 - \lambda_t \Delta + o(\Delta),$$

where $\mathcal{F}_t^N$ is the filtration of N_t up to time t and λ_t is given by

$$\lambda_t = \lambda^* + \int_{-\infty}^{t} g(t-u)\mathrm{d}N_u,$$

where $\lambda^* > 0$ and $g : \mathbb{R}^+ \to \mathbb{R}^+$ is such that $\int_0^\infty g(x)\mathrm{d}x < 1$. Through this definition, the intensity λ_t captures the history of the arrival process up to time t. Thus, λ_t encapsulates the sequence of past events and uses it to determine the rate of future occurrences. We refer to λ^* as the baseline intensity and $g(\cdot)$ as the excitation kernel. The baseline intensity represents an underlying stationary arrival rate and the excitation kernel governs the effect that the history of the process has on the current intensity. A common modeling choice is to set $g(x) = \alpha e^{-\beta x}$, where $\beta > \alpha > 0$. This is often referred to as the "exponential" kernel and it is perhaps the most widely used form of the Hawkes process. In this case, (λ_t, N_t) is a Markov process obeying the stochastic differential equation

$$\mathrm{d}\lambda_t = \beta(\lambda^* - \lambda_t)\mathrm{d}t + \alpha \mathrm{d}N_t. \tag{1.1}$$

That is, at arrival epochs λ_t jumps upward by amount α and the $N_{t,\lambda}$ increases by 1; between arrivals λ_t decays exponentially at rate β towards the baseline intensity λ^*. Thus, each arrival increases the likelihood of additional arrivals occurring soon afterwards – hence, it self-excites. This form of the Hawkes process is also often alternatively stated with an initial value for λ_t, say $\lambda_0 \geq \lambda^*$. In this case, if one applies Ito's lemma to the kernel function $e^{-\beta t}\lambda_t$, then one can show that

$$\lambda_t = \lambda^* + e^{-\beta t}(\lambda_0 - \lambda^*) + \alpha \int_0^t e^{-\beta(t-s)} dN_s, \tag{1.2}$$

as in Da Fonseca and Zaatour (2014), which also discusses the impact of the initial value of the intensity λ_0. This process is known to be stable for $\alpha < \beta$, see Laub et al.. Additionally, it is Markovian when conditioned on the present value of both the counting process and the intensity, which is also given in Laub et al.. For the rest of this study we will restrict our setting to this exponential kernel assumption. When we use the term "Hawkes process" we assume that it has such a kernel. Before proceeding with a review of relevant Hawkes process results from the literature, we motivate the use of this process by showing both its similarities and its differences with the Poisson process.

1.1.1 Comparison to the Poisson Process

In Equation 1.2, note that if $\alpha = 0$ and $\lambda_0 = \lambda^*$ then $\lambda_t = \lambda^*$ for all t. In this case, the Hawkes process is equivalent to a stationary Poisson process with rate λ^*. However, if $\alpha = 0$ but $\lambda_0 \neq \lambda^*$ it is equivalent to a non-stationary Poisson process. So, conceptually, a Poisson process is a Hawkes process without excitement. Furthermore, a Hawkes process with $\lambda_0 = \lambda^*$ is in essence a stationary Poisson process until the first arrival occurs. However, once an arrival occurs the intensity process jumps by an amount α from the initial value and then begins to decay towards the baseline rate according to the exponential decay rate β. This is demonstrated in the example in Figure 1.1 below. This simulation, in addition to all the others throughout this work, is constructed by use of the algorithm described in Ogata (1981).

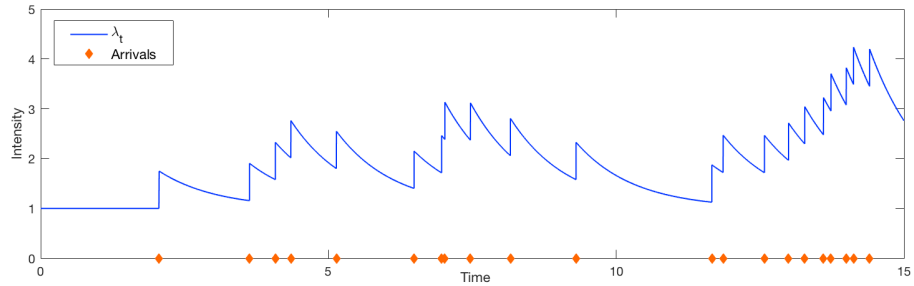


Figure 1.1: Simulated λ_t, where $\alpha = 3/4, \beta = 1$, and $\lambda^* = 1$.

This example also shows another key difference between the Hawkes and Poisson processes. Because the self-excitation increases the likelihood of another arrival occurring soon after, the Hawkes process tends to cluster arrivals together across time. This means that the variance of a Hawkes process will be larger than that of a Poisson process, which is known to be equal to its mean. Below we demonstrate this through simulated limit distributions of the Hawkes process compared with the known Poisson probability mass function (PMF), each with the same mean.

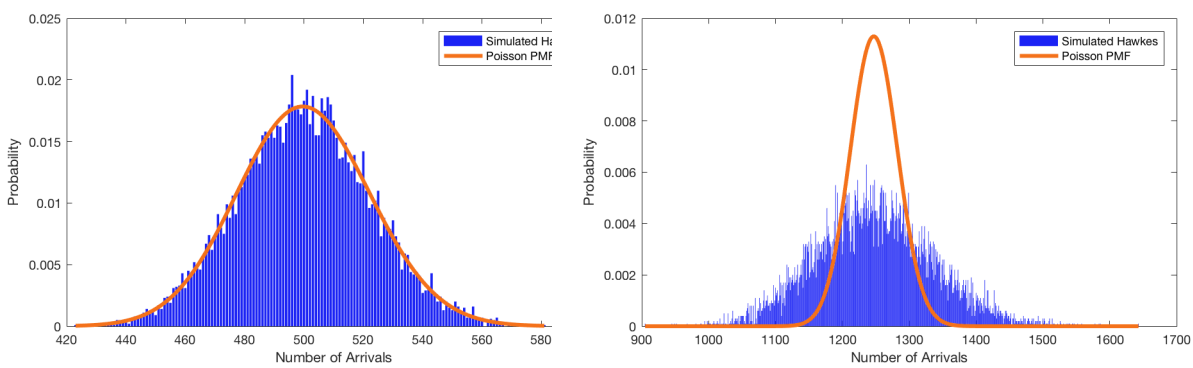


Figure 1.2: Limit Distributions for $\lambda^* = \beta = 1$ and $\alpha = 0$ (left) and 0.6 (right).

The simulated results are based on 10,000 replications, each with an end time of 500. As described previously, the two processes are equivalent for $\alpha = 0$. However, as α increases, the similarity between the Hawkes process and the

Poisson process starts to disappear. Through these examples, we observe that the Hawkes process behaves quite differently from the Poisson process since it has heavier tails and therefore, is more variable. Thus, this provides theoretical motivation for the following investigations.

1.1.2 Review of Relevant Hawkes Process Literature

We now review a brief selection of Hawkes process results that support our following analysis of Hawkes processes in queueing systems. These results can be found in greater detail in Dassios and Zhao (2011); Da Fonseca and Zaatour (2014, 2015), as discussed specifically after each result statement. This review is primarily focused on the transient and stationary moments of the Hawkes process, and is included both for the sake of completeness and understanding of the problem, but also so that it may be incorporated later in this work. In the first statement, Proposition 1.1.1, differential equations for the moments of the Hawkes process are provided.

Proposition 1.1.1. *Given a Hawkes process $X_t = (\lambda_t, N_t)$ with dynamics given by Equation 1.1, then we have the following differential equations for the moments of N_t and λ_t,*

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[N_t^m\right] = \sum_{j=0}^{m-1}\binom{m}{j}\mathrm{E}\left[\lambda_t N_t^j\right] \tag{1.3}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[\lambda_t^m\right] = m\beta\lambda^*\mathrm{E}\left[\lambda_t^{m-1}\right] - m\beta\mathrm{E}\left[\lambda_t^m\right] + \sum_{j=0}^{m-1}\binom{m}{j}\alpha^{m-j}\mathrm{E}\left[\lambda_t^{j+1}\right] \tag{1.4}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[\lambda_t^m N_t^l\right] = m\beta\lambda^*\mathrm{E}\left[\lambda_t^{m-1}N_t^l\right] - m\beta\mathrm{E}\left[\lambda_t^m N_t^l\right] + \sum_{(j,k)\in S}\binom{m}{j}\binom{l}{k}\alpha^{m-j}\mathrm{E}\left[\lambda_t^{j+1}N_t^k\right] \tag{1.5}$$

where $S = (\{0,\ldots,m\}\times\{0,\ldots,l\}) \setminus \{(m,l)\}$.

Proof. This follows directly from the approach involving the infinitesimal gen-

erator described in Sections 2.1 and 2.2 of Da Fonseca and Zaatour (2014), followed by simplification using the binomial theorem. For the first and second moments of N_t and λ_t and the first product moment, these equations are stated exactly in that work. □

As has been observed in the literature, the differential equations for the moments form a system of linear ordinary differential equations that have explicit solutions. We now provide the exact dynamics of the first two moments of the Hawkes process since this is of particular relevance to our later analysis. We also define notation that will be used throughout the remainder of this work.

Proposition 1.1.2. *Given a Hawkes process $X_t = (\lambda_t, N_t)$ with dynamics given by Equation 1.1 with $\alpha < \beta$, then the mean, variance, and covariance of N_t and λ_t are provided by the following equations for all $t \geq 0$,*

$$\mathrm{E}[\lambda_t] = \lambda_\infty + (\lambda_0 - \lambda_\infty)\, e^{-(\beta-\alpha)t} \tag{1.6}$$

$$\mathrm{E}[N_t] = \lambda_\infty t + \frac{\lambda_0 - \lambda_\infty}{\beta - \alpha}\left(1 - e^{-(\beta-\alpha)t}\right) \tag{1.7}$$

$$\mathrm{Var}(\lambda_t) = \frac{\alpha^2\lambda_\infty}{2(\beta-\alpha)} + \frac{\alpha^2(\lambda_0 - \lambda_\infty)}{\beta-\alpha}e^{-(\beta-\alpha)t} - \frac{\alpha^2(2\lambda_0 - \lambda_\infty)}{2(\beta-\alpha)}e^{-2(\beta-\alpha)t} \tag{1.8}$$

$$\begin{aligned}\mathrm{Var}(N_t) = {} & \frac{\beta^2\lambda_\infty}{(\beta-\alpha)^2}t + \frac{\alpha^2(2\lambda_0 - \lambda_\infty)}{2(\beta-\alpha)^3}\left(1 - e^{-2(\beta-\alpha)t}\right) - \frac{2\alpha\beta(\lambda_0 - \lambda_\infty)}{(\beta-\alpha)^2}te^{-(\beta-\alpha)t} \\ & + \left(\frac{\beta+\alpha}{(\beta-\alpha)^2}(\lambda_0 - \lambda_\infty) - \frac{2\alpha\beta}{(\beta-\alpha)^3}\lambda_\infty\right)(1 - e^{-(\beta-\alpha)t})\end{aligned} \tag{1.9}$$

$$\begin{aligned}\mathrm{Cov}[\lambda_t, N_t] = {} & \left(\frac{\alpha\lambda_\infty}{\beta-\alpha} + \frac{\alpha^2\lambda_\infty}{2(\beta-\alpha)^2}\right)\left(1 - e^{-(\beta-\alpha)t}\right) + \frac{\alpha^2(2\lambda_0 - \lambda_\infty)}{2(\beta-\alpha)^2}\left(e^{-2(\beta-\alpha)t} - e^{-(\beta-\alpha)t}\right) \\ & + \frac{\alpha\beta(\lambda_0 - \lambda_\infty)}{\beta-\alpha}te^{-(\beta-\alpha)t}\end{aligned} \tag{1.10}$$

where

$$\lambda_\infty = \frac{\beta\lambda^*}{\beta-\alpha}.$$

Proof. The proof of this result can be found in Section 3.4 of Dassios and Zhao

(2011) (as a particular case where $\rho = 0$) and in Section 3.2 of Da Fonseca and Zaatour (2015), or by solving the corresponding ODE system stated above Proposition 1.1.1. □

By further observation of Proposition 1.1.2 or simply by further review of the references in this section, the steady-state behavior of various Hawkes process statistics is also available. These expressions are stated in the following corollary.

Corollary 1.1.3. *Given a Hawkes process $X_t = (\lambda_t, N_t)$ with dynamics given by Equation 1.1 with $\alpha < \beta$, then the steady state values of the mean and variance of the intensity and of the covariance between the intensity and the counting process are as follows:*

$$\lim_{t\to\infty} \mathrm{E}\left[\lambda_t\right] = \frac{\beta\lambda^*}{\beta-\alpha} = \lambda_\infty, \tag{1.11}$$

$$\lim_{t\to\infty} \mathrm{Var}\left(\lambda_t\right) = \frac{\alpha^2\lambda_\infty}{2(\beta-\alpha)}, \tag{1.12}$$

$$\lim_{t\to\infty} \mathrm{Cov}\left[\lambda_t, N_t\right] = \frac{\alpha\lambda_\infty}{\beta-\alpha} + \frac{\alpha^2\lambda_\infty}{2(\beta-\alpha)^2}. \tag{1.13}$$

In Proposition 1.1.2 and Corollary 1.1.3, we assume that $\alpha < \beta$, which is a known stability condition in the literature, as detailed in Laub et al.. However, we can also consider the case where $\alpha \geq \beta$ and investigate the behavior of the system through its transient mean values. This is performed in the following corollary.

Corollary 1.1.4. *Given a Hawkes process $X_t = (\lambda_t, N_t)$ with dynamics given by Equation 1.1 with $\alpha \geq \beta$, the transient mean intensity and transient mean of the counting*

process for $t \geq 0$ are

$$\mathrm{E}[\lambda_t] = \frac{\beta\lambda^*}{\alpha-\beta}\left(e^{(\alpha-\beta)t} - 1\right) + \lambda_0 e^{(\alpha-\beta)t} \tag{1.14}$$

$$\mathrm{E}[N_t] = \left(\frac{\beta\lambda^*}{(\alpha-\beta)^2} + \frac{\lambda_0}{\alpha-\beta}\right)\left(e^{(\alpha-\beta)t} - 1\right) - \frac{\beta\lambda^*}{\alpha-\beta}t \tag{1.15}$$

when $\alpha > \beta$, and

$$\mathrm{E}[\lambda_t] = \beta\lambda^* t + \lambda_0 \tag{1.16}$$

$$\mathrm{E}[N_t] = \frac{\beta\lambda^*}{2}t^2 + \lambda_0 t \tag{1.17}$$

when $\alpha = \beta$.

As is stated in the stability condition, we see that the limits of these functions as t goes to infinity diverge for $\alpha \geq \beta$. The effect of the relationship of α and β on the system can be observed in the following graph.

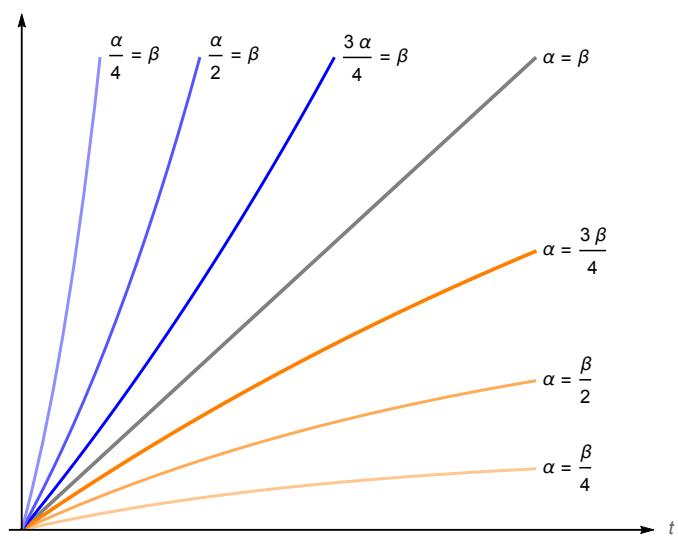


Figure 1.3: Transient Mean Intensity for $\alpha < \beta$, $\alpha = \beta$, and $\alpha > \beta$.

For the majority of this work we will consider settings in which the arrival process is stable and so we will assume $\alpha < \beta$. However, there are settings in

which the transient behavior of the unstable arrival process is of interest, and so in our analysis of the queueing system we will also explore the mean behavior of queues under such arrival conditions.

CHAPTER 2

QUEUES DRIVEN BY HAWKES PROCESSES

2.1 Introduction

Historically, the Hawkes process has been studied predominantly in financial settings. However, it has only recently received a significant amount of attention in more general contexts. In this chapter, we are particularly interested in socially informed queueing systems, and we use these systems as a motivation for both studying the Hawkes process and applying it to queueing models. For example, in situations in which a person does not know the value of competing offers or services, she may decide to pursue the option that has the most other people already waiting for it. When one can't be sure of what is earned by waiting, the willingness of others to wait can often be the best indicator.

As a quick example for the sake of building intuition, consider walking past a street performer. If there is only a handful of other people watching, one may not feel a desire to stop and see the performance. However, if there is a large crowd already watching it is more enticing to join the group and see what is happening. This is the basic motivation of self-exciting and clustering arrival processes. Although this example is simple, the concept itself has powerful implications for service systems. Several naturally occurring examples of these systems were detailed in a recent Chicago Booth Review article (Mordfin, 2015).

Contents of this chapter have been published in Daw and Pender (2018).

These examples include cellular companies paying employees to join the lines outside stores during product launches and pastry enthusiasts waiting hours in queue to buy baked goods from the famed Dominique Ansel Bakery in New York. (The article even includes a story of a German man joining a long queue in 1947 without any knowledge of what awaited him, only to find it was for visas to the United States!)

In this chapter, we will discuss Hawkes driven queues in two main applications: the viral nature of modern web traffic and the appeal associated with the lengths of queues for nightclubs. In socially informed internet traffic, webpages experience arrivals of users in clusters due to the contagion-like spread of information. If one user shares a webpage, others become more likely to view and share it as well. We demonstrate this through an example from Twitter data and explore the impact of a click. The night club example can be seen as an effect of having to pay a cover fee up front to enter the venue. Because club-goers must pay before ever seeing inside, the number of others already in queue to enter the club gives a sense of the attraction they are awaiting. In this setting we consider the managerial control problem of how quickly to admit customers to maximize earnings. Again, in these examples the occurrence of an event or arrival of a customer increases the likelihood that another will happen soon after.

We model these sort of settings through queueing systems in which the arrivals occur according to a Hawkes process and in which service times follow phase-type distributions. This general type of service allows for more detailed modeling while preserving key characteristics for queues, such as the Markov property. Mathematically, this work is most similar to recent work by Gao and Zhu (2018a) and Koops et al. (2018). Moreover, transient moments for infinite

server queues with Markovian arrivals are also among the findings in Koops et al. (2018), an independent and concurrent work. However the moments in Koops et al. (2018) are only derived for exponential service distributions, whereas we give expressions for any phase-type service distribution. Additionally, we analyze the Hawkes/D/∞ queue and give an explicit analysis for its first two moments. Conceptually, our motivation is most similar to Debo et al. (2012). While the model in Debo et al. (2012) is similar to this one in concept, it is quite different in its probabilistic structure. Rather than using a Hawkes process for the arrivals, the authors model the scenario through a Poisson process with a probability of arrivals joining or balking that increases with the length of the queue. This describes the setting well, but there are a few limitations and room for additional considerations. For example, recency plays no role in the influence of the next arrival. For queues of identical length, that model considers the most recent arrival occurring a minute ago to be equivalent to it occurring an hour ago. Additionally, because events arrive according to a time-homogeneous Poisson process and then either join or balk, the rate at which arrivals join the queue is bounded by the overall arrival rate, a constant. This prevents any kind of "viral" behavior for the events, so a large influx of arrivals over a short time is unlikely to occur. By contrast, these behaviors are inherent to our model. We will explore these ideas and others after the following descriptions of this chapter's composition.

2.1.1 Main Contributions of Chapter

In this chapter, we provide exact expressions for the mean, variance, and covariance of the Hawkes process driven queue for all time, in both transient and

steady state. These moments are derived for general phase-type service; we also provide examples for hyper-exponential and Erlang service. These results are derived by exploiting linear ordinary differential equations. We also derive expressions for all moments of the queue. We verify these functions via comparisons to simulations. We also derive a partial differential equation for the moment generating function and the cumulant moment generating function for the Hawkes/PH/∞ queue. We are able to show that the solution of the potentially high dimensional PDE for the MGF can be reduced to solving one differential equation, which does not have a closed form expression except in some special cases. Moreover, we analyze the Hawkes/D/∞ queue where the service times are deterministic. We derive exact expressions for the mean, variance, and auto-covariance of the queue length process. Throughout this work we show the relevance of the Hawkes process by direct comparison to the Poisson process and through novel applications. In our applications, we investigate the long run effects of the self-excitement structure, design an optimal control problem, and describe how to solve it numerically.

2.1.2 Organization of Chapter

The remainder of this chapter is organized into three main sections. In Section 1.1, we give an overview of results and properties in the Hawkes process literature that are relevant to this work and we then investigate the infinite server Hawkes process driven queue with deterministic service. In Section 2.3, we perform the main analysis of this work, which is the investigation of infinite server queues with Hawkes process arrivals and phase-type distributed service. In doing so, we first provide model definitions and technical lemmas, then derive

expressions for the moments of the queue, followed by the auto-covariance and moment and cumulant generating functions. In Section 2.4, we apply this work to two novel settings, trending web traffic and night clubs.

2.2 $Hawkes/D/\infty$ Queue

Before moving on to the phase-type distributed service systems, we will first investigate the deterministic service setting. Since we have a good understanding about the Hawkes process itself, we can leverage our knowledge to analyze the $Hawkes/D/\infty$ queue where D is deterministic and is equal to the exact amount of time each customer spends in service. We exploit the fact that the $Hawkes/D/\infty$ queue can be written as the difference between the Hawkes process evaluated at time t and the Hawkes process evaluated at time $t - D$ i.e.

$$Q_t = N_t - N_{t-D}. \tag{2.1}$$

This representation of the $Hawkes/D/\infty$ queue leads us to a theorem that provides explicit expressions for the mean, variance, and auto-covariance of the $Hawkes/D/\infty$ queueing process. However, before we state the result, we need a lemma that describes the transient auto-covariance of the Hawkes process. This lemma will be extremely useful for our future calculations of other quantities of interest for the $Hawkes/D/\infty$ queue.

Lemma 2.2.1. *Let N_t be a Hawkes process with dynamics given by Equation 1.1 with $\alpha < \beta$ and suppose N_t is initialized at zero. If we define $C(t, \tau)$ as*

$$C(t, \tau) \equiv \mathrm{Cov}[N_t, N_{t-\tau}], \tag{2.2}$$

then

$$C(t,\tau) = \frac{\alpha\left(1-e^{-(\beta-\alpha)\tau}\right)}{2(\beta-\alpha)^3}\left((2\beta-\alpha)\lambda_\infty - 2e^{-(\beta-\alpha)(t-\tau)}\left(\alpha\lambda_0 + \beta(\lambda_\infty-\lambda_0)(\beta-\alpha)(t-\tau) + (\beta-\alpha)\lambda_\infty\right)\right)$$
$$+\left(\lambda_\infty + \frac{2\alpha\lambda_\infty}{\beta-\alpha} + \frac{\alpha^2\lambda_\infty}{(\beta-\alpha)^2}\right)(t-\tau) + \frac{\alpha^2(2\lambda_0-\lambda_\infty)}{2(\beta-\alpha)^3}\left(1-e^{-(\beta-\alpha)(2t-\tau)}\right) - \frac{2\alpha\beta(\lambda_0-\lambda_\infty)}{(\beta-\alpha)^2}$$
$$\cdot (t-\tau)e^{-(\beta-\alpha)(t-\tau)} + \left(\frac{\beta+\alpha}{(\beta-\alpha)^2}(\lambda_0-\lambda_\infty) - \frac{2\alpha\beta}{(\beta-\alpha)^3}\lambda_\infty\right)(1-e^{-(\beta-\alpha)(t-\tau)}) \tag{2.3}$$

for all $t \geq \tau \geq 0$; otherwise $C(t,\tau) = 0$.

Proof. To see this, we manipulate the definition of the auto-covariance to find an expression in terms of other known functions. Starting from the definition of covariance, we have

$$\mathrm{Cov}\left[N_t, N_{t-\tau}\right] = \mathrm{E}\left[N_t N_{t-\tau}\right] - \mathrm{E}\left[N_t\right]\mathrm{E}\left[N_{t-\tau}\right]$$

and by Proposition 1.1.2 we have expressions for $\mathrm{E}[N_t]$ and $\mathrm{E}[N_{t-\tau}]$. Thus, we focus on $\mathrm{E}[N_t N_{t-\tau}]$. However, for brevity's sake we do not yet substitute these known expressions into the equation. By the tower property, we have that

$$C(t,\tau) = \mathrm{E}\left[\mathrm{E}\left[N_t N_{t-\tau} \mid \mathcal{F}_{t-\tau}\right]\right] - \mathrm{E}\left[N_t\right]\mathrm{E}\left[N_{t-\tau}\right]$$

where $\mathcal{F}_{t-\tau}$ is the filtration of the Hawkes process up to time $t-\tau$. Through this conditioning, $N_{t-\tau}$ is known in the inner expectation, and so we can replace $\mathrm{E}[\mathrm{E}[N_t N_{t-\tau} \mid \mathcal{F}_{t-\tau}]]$ with $\mathrm{E}[\mathrm{E}[N_t \mid \mathcal{F}_{t-\tau}]N_{t-\tau}]$. Then, again by Proposition 1.1.2 we have that $\mathrm{E}[N_t \mid \mathcal{F}_{t-\tau}] = \lambda_\infty\tau + \frac{\lambda_{t-\tau}-\lambda_\infty}{\beta-\alpha}\left(1-e^{-(\beta-\alpha)\tau}\right) + N_{t-\tau}$. Making use of this, we now have that

$$C(t,\tau) = \lambda_\infty\tau\mathrm{E}\left[N_{t-\tau}\right] + \mathrm{E}\left[\lambda_{t-\tau}N_{t-\tau}\right]\frac{1-e^{-(\beta-\alpha)\tau}}{\beta-\alpha} - \frac{\lambda_\infty}{\beta-\alpha}\mathrm{E}\left[N_{t-\tau}\right]$$
$$\cdot\left(1-e^{-(\beta-\alpha)\tau}\right) + \mathrm{E}\left[N_{t-\tau}^2\right] - \mathrm{E}\left[N_t\right]\mathrm{E}\left[N_{t-\tau}\right],$$

and by the definitions of covariance and variance this is equivalent to

$$C(t,\tau) = \lambda_\infty \tau \mathrm{E}\left[N_{t-\tau}\right] + \frac{\mathrm{Cov}\left[\lambda_{t-\tau}, N_{t-\tau}\right] + \mathrm{E}\left[\lambda_{t-\tau}\right]\mathrm{E}\left[N_{t-\tau}\right]}{\beta-\alpha}\left(1 - e^{-(\beta-\alpha)\tau}\right)$$
$$- \frac{\lambda_\infty}{\beta-\alpha}\mathrm{E}\left[N_{t-\tau}\right]\left(1 - e^{-(\beta-\alpha)\tau}\right) - \mathrm{E}\left[N_t\right]\mathrm{E}\left[N_{t-\tau}\right] + \mathrm{Var}\left(N_{t-\tau}\right) + \mathrm{E}\left[N_{t-\tau}\right]^2.$$

Here we can recognize that each term in this expression has a known form from Proposition 1.1.2. Hence, by substituting these expressions and simplifying, we achieve the stated result. □

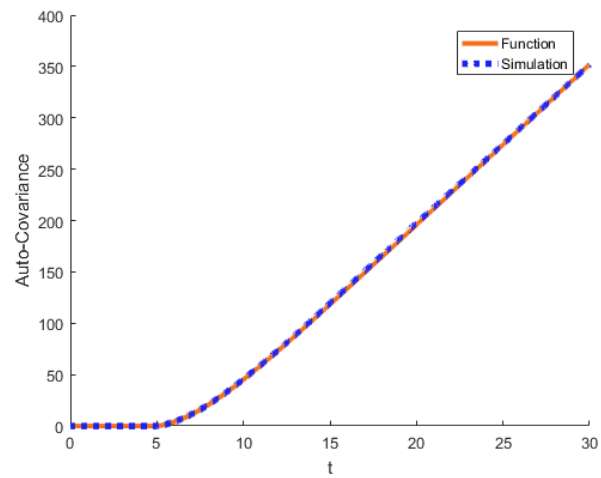


Figure 2.1: Auto-covariance of the Hawkes Process with $D = 5$, $\lambda^* = 1$, $\alpha = \frac{3}{4}$, and $\beta = \frac{5}{4}$.

With the expression for the transient auto-covariance of the Hawkes process in hand, we can now give explicit forms of the mean, variance, and auto-covariance of the $Hawkes/D/\infty$ queue.

Theorem 2.2.2. *The transient mean of the $Hawkes/D/\infty$ when $\alpha < \beta$ is given by the following expression*

$$\mathbb{E}[Q_t] = \begin{cases} \lambda_\infty t + \frac{\lambda_0 - \lambda_\infty}{\beta-\alpha}\left(1 - e^{-(\beta-\alpha)t}\right) & \text{if } t \le D, \\ \lambda_\infty D + \frac{\lambda_0 - \lambda_\infty}{\beta-\alpha}\left(e^{-(\beta-\alpha)(t-D)} - e^{-(\beta-\alpha)t}\right) & \text{if } t > D. \end{cases} \tag{2.4}$$

Thus, the steady state mean queue length is

$$\mathbb{E}[Q_\infty] = \lambda_\infty D. \tag{2.5}$$

Moreover, the transient variance of the Hawkes/D/∞ queue is given by the following expression

$$\mathrm{Var}[Q_t] = \begin{cases} C(t,0) & \text{if } t \le D, \\ C(t,0) + C(t-D,0) - 2\,C(t,D) & \text{if } t > D. \end{cases} \tag{2.6}$$

Lastly, the transient auto-covariance of the Hawkes/D/∞ queue is given by the following expression when $\tau \ge D$,

$$\mathrm{Cov}[Q_t, Q_{t-\tau}] = \begin{cases} 0 & \text{if } t \le \tau, \\ C(t,\tau) - C(t-D,\tau-D) & \text{if } \tau < t \le \tau + D \\ C(t,\tau) + C(t-D,\tau) - C(t,\tau+D) - C(t-D,\tau-D) & \text{if } \tau + D < t \end{cases} \tag{2.7}$$

and when $\tau < D$, then

$$\mathrm{Cov}[Q_t, Q_{t-\tau}] = \begin{cases} 0 & \text{if } t \le \tau, \\ C(t,\tau) & \text{if } \tau < t \le D, \\ C(t,\tau) - C(t-\tau,D-\tau) & \text{if } D < t \le \tau + D \\ C(t,\tau) + C(t-D,\tau) - C(t,\tau+D) - C(t-\tau,D-\tau) & \text{if } \tau + D < t. \end{cases} \tag{2.8}$$

Proof. Throughout this proof we make use of the form of the auto-covariance of N_t given in Lemma 2.2.1. The transient mean is straightforward since it follows from the linearity property of expectation and just taking the difference of the

two means. Moreover, for the variance we have

$$\begin{aligned}
\mathrm{Var}[Q_t] &= \mathrm{Var}[N_t - N_{t-D}] \\
&= \mathrm{Var}[N_t] + \mathrm{Var}[N_{t-D}] - 2\mathrm{Cov}[N_t, N_{t-D}] \\
&= \mathrm{Var}[N_t] + \mathrm{Var}[N_{t-D}] - 2\,C(t, D) \\
&= C(t, 0) + C(t - D, 0) - 2\,C(t, D).
\end{aligned}$$

Finally for the auto-covariance, if $\tau \geq D$ we have that

$$\mathrm{Cov}[Q_t, Q_{t-\tau}] = \begin{cases} 0 & \text{if } t \leq \tau, \\ \mathrm{Cov}\left[N_t - N_{t-D}, N_{t-\tau}\right] & \text{if } \tau < t \leq \tau + D \\ \mathrm{Cov}\left[N_t - N_{t-D}, N_{t-\tau} - N_{t-\tau-D}\right] & \text{if } \tau + D < t \end{cases}$$

by the definition of the $Hawkes/D/\infty$ queue and from the linearity of covariance. Now, for $\tau < D$, we have that

$$\mathrm{Cov}[Q_t, Q_{t-\tau}] = \begin{cases} 0 & \text{if } t \leq \tau, \\ \mathrm{Cov}\left[N_t, N_{t-\tau}\right] & \text{if } \tau < t \leq D, \\ \mathrm{Cov}\left[N_t - N_{t-D}, N_{t-\tau}\right] & \text{if } D < t \leq \tau + D \\ \mathrm{Cov}\left[N_t - N_{t-D}, N_{t-\tau} - N_{t-\tau-D}\right] & \text{if } \tau + D < t. \end{cases}$$

Again by the definition of the deterministic, Hawkes-driven, infinite server queue and the linearity of covariance, we achieve the stated result. □

2.3 $Hawkes/PH/\infty$ Queue

In this section, we will explore queueing systems in which arrivals occur according to a Hawkes process. This section is organized in the following manner. In

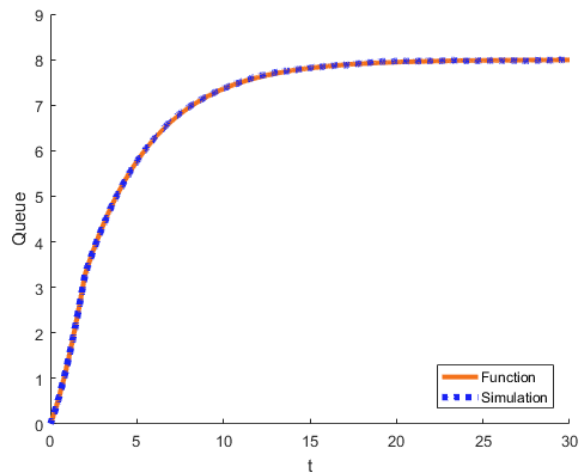


Figure 2.2: Mean of the $Hawkes/D/\infty$ Queue with $D = 5$, $\lambda^* = 1$, $\alpha = \frac{3}{4}$, and $\beta = \frac{5}{4}$.

Subsection 2.3.1, we provide key model definitions such as the phase-type distribution and we detail technical lemmas that support our analysis. Next, in Subsection 2.3.2, we derive differential equations for all moments of the queueing system and solve for exact expressions for the first and second moments. In Subsection 2.3.3, we consider the stationary limits of queues with stable arrival processes and investigate the transient behavior of those with unstable arrivals. Afterwards, we consider the auto-covariance of the queue in Subsection 2.3.4. Finally, in Subsection 2.3.5 we derive partial differential equations for the moment generating function and the cumulant moment generating function for this system.

2.3.1 Model Definitions and Technical Lemmas

To begin, we define the phase-type distribution. This form of service, formally defined below, can be thought of as a sequence of sub-services that have independent and exponentially distributed durations. We use this primarily for

two factors. The first is that this is more general than just exponential service, and it can be shown that phase-type distributions can weakly approximate any non-negative continuous distribution, see Cox (1955). Secondly, because the phase-type distribution is comprised of independent exponential service times, a queueing system with such service distributions is Markovian. Thus, these two properties together give us a system that is both flexible in application and practical in terms of analysis. A phase-type distribution with n phases represents the time taken from an initial state to an absorbing state of a continuous time Markov chain (CTMC) with the following infinitesimal generator matrix,

$$\Gamma = \begin{bmatrix} 0 & \mathbf{0} \\ \mathbf{s} & \mathbf{S} \end{bmatrix}.$$

Here $\mathbf{0}$ is a $1 \times n$ zero vector, $\mathbf{s}$ is an $n \times 1$ vector, and $\mathbf{S}$ is an $n \times n$ matrix. Note $\mathbf{s} = -\mathbf{S}\mathbf{v}$ where $\mathbf{v}$ is an $n \times 1$ vector of ones. The matrix $\mathbf{S}$ and the initial distribution θ, which is a $1 \times n$ vector, identify the phase-type distributions. The number of phases in $\mathbf{S}$ is n. The matrix $\mathbf{S}$ and vector $\mathbf{s}$ can be expressed as:

$$\mathbf{S} = \begin{bmatrix} -\mu_1 & \cdots & \mu_{1,n} \\ \vdots & \ddots & \vdots \\ \mu_{n,1} & \cdots & -\mu_n \end{bmatrix}, \quad \mathbf{s} = (\mu_{1,0}, \ldots, \mu_{n,0})^{\mathrm{T}}, \tag{2.9}$$

where the μ_{ij}'s agree with the definition of the infinitesimal generator matrix $\mathbf{\Gamma}$. For notational consistency, we use a term *phase* to indicate the state of CTMC of the phase-type distributions throughout this chapter. Additionally, we now note that in all following use of the matrix S we will not use a bold notation as in those settings additional emphasis that it is a matrix is not necessary.

With the phase-type distributions as described above, we build a Markovian queueing model referred to as the $Hawkes/PH/\infty$ queue. We assume that the system starts with no customers and that there are infinitely many servers.

Further, we suppose that there are n phases of service and the transition rate between two distinct phases i and j is μ_{ij}. Let $\theta \in [0,1]^n$ be a distribution over the phases such that the probability that an arriving entity joins the ith phase is θ_i, with $\sum_{i=1}^n \theta_i = 1$. An entity departs the system at rate μ_{i0}, where i is the entity's phase of service before leaving. For brevity of notation, define $\mu_i \equiv \mu_{i0} + \mu_{i1} + \cdots + \mu_{i,i-1} + \mu_{i,i+1} + \mu_{i,n}$. Let $Q_t \in \mathbb{N}^n$ represent the number of entities in the queueing system, with $Q_{t,i}$ representing the number in phase i of service i.e.

$$Q_t = \sum_{i=1}^{n} Q_{t,i}\mathbf{v}_i \tag{2.10}$$

where $\mathbf{v}_i$ is the unit column vector in the ith coordinate. We let (λ_t, N_t) represent a Hawkes process as described in Equation 1.1. We will now find the infinitesimal generator for real valued functions of the state space, $f : \mathbb{R}^+ \times \mathbb{N} \times \mathbb{N}^n \to \mathbb{R}$. For simplicity of notation, when describing the difference in values of f for changed arguments we will only list the variables that change, rather than listing all n queueing phase variables. This generator is shown below.

$$\mathcal{L}f(x) = \underbrace{\beta(\lambda^* - \lambda_t)\frac{\partial f(x)}{\partial \lambda_t}}_{\text{Excitation Decay}} + \underbrace{\sum_{i=1}^{n} \lambda_t \theta_i \left(f(\lambda_t + \alpha, N_t + 1, Q_{t,i} + 1) - f(x)\right)}_{\text{Arrivals}} \tag{2.11}$$

$$+ \underbrace{\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n} \mu_{ij} Q_{t,i}\Big(f(\lambda_t, N_t, Q_{t,i} - 1, Q_{t,j} + 1) - f(x)\Big)}_{\text{Transfers}} + \underbrace{\sum_{i=1}^{n} \mu_{i0} Q_{t,i}\left(f(\lambda_t, N_t, Q_{t,i} - 1) - f(x)\right)}_{\text{Departures}}$$

Here, x is an element of the state space $(\mathbb{R}^+ \times \mathbb{N} \times \mathbb{N}^n)$. We can use this to obtain Dynkin's formula for the full $Hawkes/PH/\infty$ queueing system. We have that

$$\mathrm{E}_t\left[f(X_s)\right] = f(X_t) + E_t\left[\int_t^s \mathcal{L}f(X_u)du\right], \tag{2.12}$$

where $X_t = (\lambda_t, N_t, Q_t)$. This gives rise to the following lemma.

Lemma 2.3.1. *Let f be a function such that Equation 2.12 holds. Then,*

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[f(X_t)\right] = \mathrm{E}\left[\mathcal{L}f(X_t)\right]$$

for all $t \geq 0$.

Proof. This is achieved through use of Fubini's theorem and the fundamental theorem of calculus. Using Equation 2.12 we have that

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[f(X_t)\right] &= \frac{\mathrm{d}}{\mathrm{d}t}\left(f(X_0) + \mathrm{E}\left[\int_0^t \mathcal{L}f(X_u)\mathrm{d}u\right]\right) \\ &= \frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[\int_0^t \mathcal{L}f(X_u)\mathrm{d}u\right] = \frac{\mathrm{d}}{\mathrm{d}t}\int_0^t \mathrm{E}\left[\mathcal{L}f(X_u)\right]\mathrm{d}u = \mathrm{E}\left[\mathcal{L}f(X_t)\right]\end{aligned}$$

and this completes this proof. □

Remark. It is important for the reader to recognize that this is equivalent to Dynkin's theorem. In most textbooks, Dynkin's theorem is proved for sufficiently differentiable and more importantly bounded functions. However, this assumption of boundedness can often be relaxed. In fact this relaxation of the boundedness is very common for extending results like Ito's lemma and the Feynman Kac formula for unbounded, but polynomial bounded functions. This is often extended by stopping the process when it hits a certain level by using stopping times. Then one applies the previous results for bounded functions and takes limits as the bound tends to infinity. For the interested reader, see Lemma 2 of Oelschlager (1984) for a proof.

Now, before using these differential equations to find explicit functions as we did previously, we will first introduce a series of technical lemmas to aid our analysis. These lemmas are presented without proof as they follow from standard approaches for matrix exponentials and integration. First, we give a form for the indefinite integral of the exponential of a non-singular matrix.

Lemma 2.3.2. *Let* $L \in \mathbb{R}^{n\times n}$ *be invertible. Then, if the integral of* e^{Lt} *exists it can be*

expressed

$$\int e^{Lt}\,\mathrm{d}t = L^{-1}e^{Lt} + c$$

where c is some constant of integration.

Proof. The proof follows from standard approaches. □

The second lemma now provides explicit forms for the definite integral from 0 to t of the product of an exponential of an invertible matrix, a vector, a scalar power of the variable of integration, and a scalar exponential function of the variable of integration.

Lemma 2.3.3. *Let $L \in \mathbb{R}^{n\times n}$ be invertible, let $v \in \mathbb{R}^n$, let $\eta \in \mathbb{N}$, and let $\gamma \in \mathbb{R}$. Then, if $L + \gamma I$ is invertible,*

$$\int_0^t e^{Ls} v s^{\eta} e^{\gamma s}\,\mathrm{d}s = \sum_{k=0}^{\eta} \frac{\eta!}{(\eta-k)!}(-1)^k\,(L+\gamma I)^{-(k+1)}\left(e^{Lt} v t^{\eta-k} e^{\gamma t}\right) - \eta!(-1)^{\eta}\,(L+\gamma I)^{-(\eta+1)}\,v$$

for $t > 0$.

Proof. The proof follows from the preceding lemma, induction, and integration by parts. □

The next lemma is a quick demonstration of commutativity of the inverse of a matrix exponential and an inverse of the same matrix shifted in the direction of the identity.

Lemma 2.3.4. *Let $A \in \mathbb{R}^{n\times n}$ be invertible and let $b, c \in \mathbb{R}$ be such that $cA + bI$ is also invertible. Then,*

$$e^{-A}\,(cA+bI)^{-1} = (cA+bI)^{-1}\,e^{-A}.$$

Proof. The proof follows from the definition of the matrix exponential. □

These lemmas now come together to give the general solution to differential equations of a certain form.

Lemma 2.3.5. *Let $g(t) \in \mathbb{R}^n$ be a function described by the first-derivative dynamics*

$$\dot{g}(t) = -Lg(t) + \sum_{i \in \mathcal{S}} v_i t^{\eta_i} e^{\gamma_i t}$$

with an initial condition of $g(0) = g_0$, where $L \in \mathbb{R}^{n \times n}$ is invertible and $\mathcal{S}$ is a finite index set such that $v_i \in \mathbb{R}^n$, $\eta_i \in \mathbb{N}$, and $\gamma_i \in \mathbb{R}$ for each $i \in \mathcal{S}$. Then, if $L + \gamma_i I$ is invertible for all $i \in \mathcal{S}$ the explicit function for $g(t)$ is given by

$$g(t) = \sum_{i \in \mathcal{S}} \sum_{k=0}^{\eta_i} \frac{\eta_i!(-1)^k}{(\eta_i - k)!} (L + \gamma_i I)^{-(k+1)} \left(v_i t^{\eta_i - k} e^{\gamma_i t} \right) - \eta_i!(-1)^{\eta_i} (L + \gamma_i I)^{-(\eta_i + 1)} e^{-Lt} v_i + e^{-Lt} g_0$$

for all $t \geq 0$.

Proof. The proof follows from standard differential equation techniques and the three preceding lemmas. □

Now, before introducing one final lemma we first define a useful matrix. For $\gamma, c \in \mathbb{R}$, $v \in \mathbb{R}^n$, and $L \in \mathbb{R}^{n \times n}$, let $M_{\gamma,v,L}(t) \in \mathbb{R}^{n \times n}$ be such that

$$M_{\gamma,v,L}(t) = \int_0^t e^{(\gamma I - L^{\mathrm{T}})s} v v^{\mathrm{T}} e^{-Ls} \, \mathrm{d}s \tag{2.13}$$

for all $t \geq 0$. Element-wise, we can express this matrix after integration as

$$\left(M_{\gamma,v,L}(t) \right)_{i,j} = \begin{cases} \sum_{k=1}^n \sum_{l=1}^n v_k v_l \sum_{r=0}^\infty \sum_{w=0}^\infty \frac{(L^r)_{k,i}(L^w)_{l,j}}{\gamma^{r+w+1}} \binom{r+w}{r} \left(e^{\gamma t} \sum_{z=0}^{r+w} \frac{(-\gamma t)^z}{z!} - 1 \right) & \text{if } \gamma \neq 0, \\ \sum_{k=1}^n \sum_{l=1}^n v_k v_l \sum_{r=0}^\infty \sum_{w=0}^\infty \frac{(L^r)_{k,i}(L^w)_{l,j} t^{r+w+1}}{r!w!(r+w+1)} & \text{if } \gamma = 0. \end{cases}$$

This function provides shorthand when integrating a particular function that otherwise does not produce a nice linear algebraic form. The difficulty of expressing this integral in matrix form stems from the fact that L and vv^{T} need not commute. With defining $M_{\gamma,v,L}(t)$ we circumvent this issue by integrating on the element-level, but if L and vv^{T} were to commute we could avoid this function entirely, as we will later see. For now, this definition leads us to our next lemma.

Lemma 2.3.6. *Let $\eta, \gamma, c \in \mathbb{R}$, $\nu \in \mathbb{R}^n$, $L \in \mathbb{R}^{n\times n}$ be such that L, $\gamma I + L$, and $(\eta+1)\gamma I - L$ are each invertible. Then,*

$$\begin{aligned}
&\int_0^t \Big(\left((\eta+1)\gamma I - L^{\mathrm{T}}\right)^{-1} \left(e^{(\eta\gamma I - L^{\mathrm{T}})s} - e^{-\gamma I s}\right) \nu\nu^{\mathrm{T}} c e^{-Ls} + e^{-L^{\mathrm{T}}s} \nu\nu^{\mathrm{T}} c \left(e^{(\eta\gamma I - L)s} - e^{-\gamma I s}\right) \\
&\quad \cdot ((\eta+1)\gamma I - L)^{-1} \Big) \mathrm{d}s \\
=&c \left((\eta+1)\gamma I - L^{\mathrm{T}}\right)^{-1} \Big((\eta+2)\gamma M_{\eta\gamma,\nu,L}(t) + e^{(\eta\gamma I - L^{\mathrm{T}})t} \nu\nu^{\mathrm{T}} e^{-Lt} - \nu\nu^{\mathrm{T}} + \nu\nu^{\mathrm{T}} \left(e^{-(\gamma I + L)t} - I\right) (\gamma I + L)^{-1} \\
&\quad \cdot ((\eta+1)\gamma I - L) + \left((\eta+1)\gamma I - L^{\mathrm{T}}\right) (\gamma I + L^{\mathrm{T}})^{-1} \left(e^{-(\gamma I + L^{\mathrm{T}})t} - I\right) \nu\nu^{\mathrm{T}} \Big) ((\eta+1)\gamma I - L)^{-1}
\end{aligned}$$

for all $t \geq 0$.

Proof. The proof follows from the given definition of $M_{\gamma,\nu,L}(t)$, the product rule, and the preceding lemma. □

With these lemmas and definitions now in hand we can proceed to our analysis of the $Hawkes/PH/\infty$ queueing system. These results, stated in the following theorem, make use of the form of the infinitesimal generator in Lemma A.1.1, with simplification through linearity of expectation and the binomial theorem.

2.3.2 Mean Dynamics of the $Hawkes/PH/\infty$ Queue

To begin investigation of the $Hawkes/PH/\infty$ queueing system, we first derive differential equations for the moments of the number in each phase of service and the intensity.

Theorem 2.3.7. *Consider a queueing system with arrivals occurring in accordance to a Hawkes process (λ_t, N_t) with dynamics given in Equation 1.1 and phase-type distributed*

service. Then we have differential equations for the moments of $Q_{t,i}$ given by

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[Q_{t,i}^m\right] = \theta_i \sum_{g=0}^{m-1}\binom{m}{g}\mathrm{E}\left[\lambda_t Q_{t,i}^g\right] + \sum_{g=0}^{m-1}\sum_{\substack{j=1\\ j\neq i}}^{n}\binom{m}{g}\mu_{ji}\mathrm{E}\left[Q_{t,j}Q_{t,i}^g\right] \tag{2.14}$$
$$+ \sum_{g=1}^{m}\binom{m}{g-1}\mu_i(-1)^{m-g+1}\mathrm{E}\left[Q_{t,i}^g\right],$$

for the products of $Q_{t,i}$ and $Q_{t,j}$ where $i \neq j$ given by

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[Q_{t,i}^m Q_{t,j}^l\right] = \theta_i \sum_{g=0}^{m-1}\binom{m}{g}\mathrm{E}\left[\lambda_t Q_{t,j}^l Q_{t,i}^g\right] + \theta_j \sum_{h=0}^{l-1}\binom{l}{h}\mathrm{E}\left[\lambda_t Q_{t,i}^m Q_{t,j}^h\right] \tag{2.15}$$
$$+ \sum_{\substack{k=1\\ i\neq k\neq j}}^{n}\sum_{g=0}^{m-1}\binom{m}{g}\mu_{ki}\mathrm{E}\left[Q_{t,k}Q_{t,i}^g Q_{t,j}^l\right] + \sum_{\substack{k=1\\ j\neq k\neq i}}^{n}\sum_{h=0}^{l-1}\binom{l}{h}\mu_{kj}\mathrm{E}\left[Q_{t,k}Q_{t,i}^m Q_{t,j}^h\right]$$
$$+ \mu_i \sum_{g=0}^{m-1}\binom{m}{g}(-1)^{m-g}\mathrm{E}\left[Q_{t,j}^l Q_{t,i}^{g+1}\right] + \mu_{ij}\sum_{g=0}^{m}\sum_{h=0}^{l-1}\binom{m}{g}\binom{l}{h}(-1)^{m-g}\mathrm{E}\left[Q_{t,i}^{g+1}Q_{t,j}^h\right]$$
$$+ \mu_j \sum_{h=0}^{l-1}\binom{l}{h}(-1)^{l-h}\mathrm{E}\left[Q_{t,i}^m Q_{t,j}^{h+1}\right] + \mu_{ji}\sum_{h=0}^{l}\sum_{g=0}^{m-1}\binom{l}{h}\binom{m}{g}(-1)^{l-h}\mathrm{E}\left[Q_{t,j}^{h+1}Q_{t,i}^g\right],$$

and for the products of λ_t and $Q_{t,i}$ given by

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[\lambda_t^m Q_{t,i}^l\right] = \beta\lambda^* m\mathrm{E}\left[\lambda_t^{m-1}Q_{t,i}^l\right] - \beta m\mathrm{E}\left[\lambda_t^m Q_{t,i}^l\right] + \theta_i\sum_{g=0}^{m}\sum_{h=0}^{l-1}\binom{m}{g}\binom{l}{h} \tag{2.16}$$
$$\cdot\, \alpha^{m-g}\mathrm{E}\left[\lambda_t^{g+1}Q_{t,i}^h\right] + \sum_{g=0}^{m-1}\binom{m}{g}\alpha^{m-g}\mathrm{E}\left[\lambda_t^{g+1}Q_{t,i}^l\right] + \mu_i\sum_{h=0}^{l-1}\binom{l}{h}$$
$$\cdot\, (-1)^{l-h}\mathrm{E}\left[\lambda_t^m Q_{t,i}^{h+1}\right] + \sum_{\substack{j=1\\ j\neq i}}^{n}\sum_{h=0}^{l-1}\binom{l}{h}\mu_{ji}\mathrm{E}\left[\lambda_t^m Q_{t,j}Q_{t,i}^h\right],$$

where $t \geq 0$.

Proof. We can first observe that each of these moments can be generalized to

$\mathrm{E}\left[\lambda_t^m Q_{t,i}^l Q_{t,j}^k\right]$. From Lemma A.1.1 we see that

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[\lambda_t^m Q_{t,i}^l Q_{t,j}^k\right] = \mathrm{E}\Big[&\beta(\lambda^* - \lambda_t)m\lambda_t^{m-1} Q_{t,i}^l Q_{t,j}^k + \lambda_t\theta_i\left((\lambda_t+\alpha)^m (Q_{t,i}+1)^l Q_{t,j}^k - \lambda_t^m Q_{t,i}^l Q_{t,j}^k\right) \\
&+ \lambda_t\theta_j\left((\lambda_t+\alpha)^m Q_{t,i}^l (Q_{t,j}+1)^k - \lambda_t^m Q_{t,i}^l Q_{t,j}^k\right) + \sum_{\substack{x=1\\ j\neq x\neq i}}^{n} \lambda_t\theta_x Q_{t,i}^l Q_{t,j}^k\left((\lambda_t+\alpha)^m - \lambda_t^m\right) \\
&+ \sum_{\substack{x=1\\ i\neq x\neq j}}^{n} \mu_{xi} Q_{t,x}\lambda_t^m Q_{t,j}^k\left((Q_{t,i}+1)^l - Q_{t,i}^l\right) + \sum_{\substack{x=1\\ j\neq k\neq i}}^{n} \mu_{xj} Q_{t,x}\lambda_t^m Q_{t,i}^l\left((Q_{t,j}+1)^k - Q_{t,j}^k\right) \\
&+ \sum_{\substack{x=0\\ i\neq x\neq j}}^{n} \mu_{ix} Q_{t,i}\lambda_t^m Q_{t,j}^k\left((Q_{t,i}-1)^l - Q_{t,i}^l\right) + \sum_{\substack{x=0\\ j\neq x\neq i}}^{n} \mu_{jx} Q_{t,j}\lambda_t^m Q_{t,i}^l\left((Q_{t,j}-1)^k - Q_{t,j}^k\right) \\
&+ \mu_{ij} Q_{t,i}\lambda_t^m\left((Q_{t,i}-1)^l (Q_{t,j}+1)^k - Q_{t,i}^l Q_{t,j}^k\right) + \mu_{ji} Q_{t,j}\lambda_t^m\left((Q_{t,j}-1)^k (Q_{t,i}+1)^l - Q_{t,i}^l Q_{t,j}^k\right)\Big]
\end{aligned}$$

where we have combined the transfers from one phase to another and departures from that phase into the same summation by starting the index at 0. Using the binomial theorem and linearity of expectation, we have the following:

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[\lambda_t^m Q_{t,i}^l Q_{t,j}^k\right] &= \beta\lambda^* m\mathrm{E}\left[\lambda_t^{m-1} Q_{t,i}^l Q_{t,j}^k\right] - \beta m\mathrm{E}\left[\lambda_t^m Q_{t,i}^l Q_{t,j}^k\right] + \sum_{\substack{x=1\\ j\neq x\neq i}}^{n}\sum_{y=0}^{m-1}\binom{m}{y}\theta_x\alpha^{m-y} \\
&\cdot \mathrm{E}\left[\lambda_t^{y+1} Q_{t,i}^l Q_{t,j}^k\right] + \theta_i\left(\sum_{x=0}^{m}\sum_{y=0}^{l}\binom{m}{x}\binom{l}{y}\alpha^{m-x}\mathrm{E}\left[\lambda_t^{x+1} Q_{t,i}^y Q_{t,j}^k\right] - \mathrm{E}\left[\lambda_t^{m+1} Q_{t,i}^l Q_{t,j}^k\right]\right) \\
&+ \theta_j\left(\sum_{x=0}^{m}\sum_{y=0}^{k}\binom{m}{x}\binom{k}{y}\alpha^{m-x}\mathrm{E}\left[\lambda_t^{x+1} Q_{t,i}^l Q_{t,j}^y\right] - \mathrm{E}\left[\lambda_t^{m+1} Q_{t,i}^l Q_{t,j}^k\right]\right) + \sum_{\substack{x=1\\ i\neq x\neq j}}^{n}\sum_{y=0}^{l-1}\binom{l}{y}\mu_{xi} \\
&\cdot \mathrm{E}\left[\lambda_t^m Q_{t,x} Q_{t,i}^y Q_{t,j}^k\right] + \sum_{\substack{x=1\\ i\neq x\neq j}}^{n}\sum_{y=0}^{k-1}\binom{k}{y}\mu_{xj}\mathrm{E}\left[\lambda_t^m Q_{t,x} Q_{t,i}^l Q_{t,j}^y\right] + \sum_{\substack{x=0\\ i\neq x\neq j}}^{n}\sum_{y=0}^{l-1}\binom{l}{y}(-1)^{l-y}\mu_{ix}\mathrm{E}\left[\lambda_t^m Q_{t,i}^{y+1} Q_{t,j}^k\right] \\
&+ \sum_{\substack{x=0\\ i\neq x\neq j}}^{n}\sum_{y=0}^{k-1}\binom{k}{y}(-1)^{k-y}\mu_{jx}\mathrm{E}\left[\lambda_t^m Q_{t,i}^l Q_{t,j}^{y+1}\right] + \mu_{ij}\left(\sum_{x=0}^{l}\sum_{y=0}^{k}\binom{l}{x}\binom{k}{y}(-1)^{l-x}\mathrm{E}\left[\lambda_t^m Q_{t,i}^{x+1} Q_{t,j}^y\right]\right. \\
&\left.- \mathrm{E}\left[\lambda_t^m Q_{t,i}^{l+1} Q_{t,j}^k\right]\right) + \mu_{ji}\left(\sum_{x=0}^{l}\sum_{y=0}^{k}\binom{l}{x}\binom{k}{y}(-1)^{k-y}\mathrm{E}\left[\lambda_t^m Q_{t,i}^x Q_{t,j}^{y+1}\right] - \mathrm{E}\left[\lambda_t^m Q_{t,i}^l Q_{t,j}^{k+1}\right]\right).
\end{aligned}$$

Now we simplify by recognizing that $\sum_{x\neq j}\mu_{ix} = \mu_i - \mu_{ij}$ and $\sum_{i\neq x\neq j}\theta_x = 1-\theta_i-\theta_j$. This leaves us with

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[\lambda_t^m Q_{t,i}^l Q_{t,j}^k\right] &= \beta\lambda^* m\mathrm{E}\left[\lambda_t^{m-1} Q_{t,i}^l Q_{t,j}^k\right] - \beta m\mathrm{E}\left[\lambda_t^m Q_{t,i}^l Q_{t,j}^k\right] + \sum_{y=0}^{m-1}\binom{m}{y}\alpha^{m-y}\mathrm{E}\left[\lambda_t^{y+1} Q_{t,i}^l Q_{t,j}^k\right] \\
&+ \theta_i\sum_{x=0}^{m}\sum_{y=0}^{l-1}\binom{m}{x}\binom{l}{y}\alpha^{m-x}\mathrm{E}\left[\lambda_t^{x+1} Q_{t,i}^y Q_{t,j}^k\right] + \theta_j\sum_{x=0}^{m}\sum_{y=0}^{k-1}\binom{m}{x}\binom{k}{y}\alpha^{m-x}\mathrm{E}\left[\lambda_t^{x+1} Q_{t,i}^l Q_{t,j}^y\right] \\
&+ \sum_{\substack{x=1\\ i\neq x\neq j}}^{n}\sum_{y=0}^{l-1}\binom{l}{y}\mu_{xi}\mathrm{E}\left[\lambda_t^m Q_{t,x} Q_{t,i}^y Q_{t,j}^k\right] + \sum_{\substack{x=1\\ i\neq x\neq j}}^{n}\sum_{y=0}^{k-1}\binom{k}{y}\mu_{xj}\mathrm{E}\left[\lambda_t^m Q_{t,x} Q_{t,i}^l Q_{t,j}^y\right] \\
&+ \mu_i\sum_{y=0}^{l-1}\binom{l}{y}(-1)^{l-y}\mathrm{E}\left[\lambda_t^m Q_{t,i}^{y+1} Q_{t,j}^k\right] + \mu_{ij}\sum_{x=0}^{l}\sum_{y=0}^{k-1}\binom{l}{x}\binom{k}{y}(-1)^{l-x}\mathrm{E}\left[\lambda_t^m Q_{t,i}^{x+1} Q_{t,j}^y\right] \\
&+ \mu_j\sum_{y=0}^{k-1}\binom{k}{y}(-1)^{k-y}\mathrm{E}\left[\lambda_t^m Q_{t,i}^l Q_{t,j}^{y+1}\right] + \mu_{ji}\sum_{y=0}^{k}\sum_{x=0}^{l-1}\binom{l}{x}\binom{k}{y}(-1)^{k-y}\mathrm{E}\left[\lambda_t^m Q_{t,i}^x Q_{t,j}^{y+1}\right]
\end{aligned}$$

which is equivalent to each stated result when $m = k = 0$, $k = 0$, and $m = 0$, respectively. □

We can now observe that we can form closed systems of linear ordinary differential equations from these equations. To do so, we restrict our focus to the equations for moments of combined power at most $m \in \mathbb{Z}^+$. Of course, the collection of equations that is of most practical interest is found by setting $m = 2$, as this yields a system for the means and variances. This now gives rise to Corollary 2.3.8, which states the differential equations for the mean, variance, and covariances of queues driven by Hawkes processes.

Corollary 2.3.8. *Consider a queueing system with arrivals occurring in accordance to a Hawkes process (λ_t, N_t) with dynamics given in Equation 1.1 and phase-type distributed service. Then, we have the following differential equations for the mean, variance, and*

covariances of the number of entities in each phase and in the system as a whole:

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[Q_{t,i}\right] = \theta_i \mathrm{E}\left[\lambda_t\right] + \sum_{\substack{j=1\\ j\neq i}}^{n} \mu_{ji}\mathrm{E}\left[Q_{t,j}\right] - \mu_i \mathrm{E}\left[Q_{t,i}\right] \tag{2.17}$$

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{Var}\left(Q_{t,i}\right) = \theta_i \mathrm{E}\left[\lambda_t\right] + 2\theta_i \mathrm{Cov}\left[\lambda_t, Q_{t,i}\right] + 2\sum_{\substack{j=1\\ j\neq i}}^{n} \mu_{ji}\mathrm{Cov}\left[Q_{t,i}, Q_{t,j}\right] + \mu_i \mathrm{E}\left[Q_{t,i}\right] \\ + \sum_{\substack{j=1\\ j\neq i}}^{n} \mu_{ji}\mathrm{E}\left[Q_{t,j}\right] - 2\mu_i \mathrm{Var}\left(Q_{t,i}\right)\end{aligned} \tag{2.18}$$

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{Cov}\left[\lambda_t, Q_{t,i}\right] = (\alpha - \beta - \mu_i)\mathrm{Cov}\left[\lambda_t, Q_{t,i}\right] + \alpha\theta_i \mathrm{E}\left[\lambda_t\right] + \sum_{\substack{j=1\\ j\neq i}}^{n} \mu_{ji}\mathrm{Cov}\left[\lambda_t, Q_{t,j}\right] \\ + \theta_i \mathrm{Var}\left(\lambda_t\right)\end{aligned} \tag{2.19}$$

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{Cov}\left[Q_{t,i}, Q_{t,j}\right] = \theta_i \mathrm{Cov}\left[\lambda_t, Q_{t,j}\right] + \theta_j \mathrm{Cov}\left[\lambda_t, Q_{t,i}\right] - (\mu_i + \mu_j)\mathrm{Cov}\left[Q_{t,i}, Q_{t,j}\right] \\ + \sum_{\substack{k=1\\ k\neq i}}^{n} \mu_{ki}\mathrm{Cov}\left[Q_{t,k}, Q_{t,j}\right] + \sum_{\substack{k=1\\ k\neq i}}^{n} \mu_{kj}\mathrm{Cov}\left[Q_{t,k}, Q_{t,i}\right] - \mu_{ij}\mathrm{E}\left[Q_{t,i}\right] - \mu_{ji}\mathrm{E}\left[Q_{t,j}\right].\end{aligned} \tag{2.20}$$

We will find that it is quite useful to also be able to state the equations in Corollary 2.3.8 in linear algebraic form. Recall that the vector of the number in each phase of service is $Q_t \in \mathbb{N}^n$, the distribution of arrivals into phases is $\theta \in [0,1]^n$, and the sub-generator-matrix for the n phases of service is $S \in \mathbb{R}^{n\times n}$ so that $S_{i,i} = -\mu_i$ for each $i \in \{1,\ldots,n\}$ and $S_{i,j} = \mu_{i,j}$ for all $j \neq i$. We now also incorporate the notation $\mathrm{diag}\left(x\right) \in \mathbb{R}^{n\times n}$ for $x \in \mathbb{R}^n$ as $\mathrm{diag}\left(x\right) \equiv \sum_{i=1}^{n} \mathbf{V}_i x \mathbf{v}_i^{\mathrm{T}}$, where $\mathbf{v}_i \in \mathbb{R}^n$ is the unit column vector in the direction of the ith coordinate and $\mathbf{V}_i = \mathbf{v}_i\mathbf{v}_i^{\mathrm{T}}$, meaning that the ith diagonal element is 1 and the rest are 0. Together, we have that the vector form of Equation 2.17 is

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[Q_t\right] = \theta \mathrm{E}\left[\lambda_t\right] + S^{\mathrm{T}}\mathrm{E}\left[Q_t\right],$$

the vector form of Equation 2.19 is

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{Cov}\left[\lambda_t, Q_t\right] = \left(S^{\mathrm{T}} - (\beta-\alpha)I\right)\mathrm{Cov}\left[\lambda_t, Q_t\right] + \alpha\theta\mathrm{E}\left[\lambda_t\right] + \theta\mathrm{Var}\left(\lambda_t\right),$$

and the matrix form of Equations 2.18 and 2.20 is

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{Cov}\left[Q_t, Q_t\right] &= S^{\mathrm{T}}\mathrm{Cov}\left[Q_t, Q_t\right] + \mathrm{Cov}\left[Q_t, Q_t\right]S + \theta\mathrm{Cov}\left[\lambda_t, Q_t\right]^{\mathrm{T}} + \mathrm{Cov}\left[\lambda_t, Q_t\right]\theta^{\mathrm{T}} \\ &\quad + \mathrm{diag}\left(\theta\mathrm{E}\left[\lambda_t\right] + S^{\mathrm{T}}\mathrm{E}\left[Q_t\right]\right) - S^{\mathrm{T}}\mathrm{diag}\left(\mathrm{E}\left[Q_t\right]\right) - \mathrm{diag}\left(\mathrm{E}\left[Q_t\right]\right)S\end{aligned}$$

where the diagonal elements of the matrix $\mathrm{Cov}\left[Q_t, Q_t\right]$ correspond to the variance of the number in each phase of service and the off-diagonal elements represent the covariance between two phases of service. We can now use the technical lemmas in Subsection 2.3.1 to find explicit linear algebraic solutions to the closed system of differential equations in Corollary 2.3.8.

Theorem 2.3.9. *Consider a queueing system with arrivals occurring in accordance to a Hawkes process (λ_t, N_t) with dynamics given in Equation 1.1 with $\alpha < \beta$ and phase-type distributed service. Let $S \in \mathbb{R}^{n\times n}$ be the sub-generator matrix for the transient states in the phase-distribution CTMC and let $\theta \in [0,1]^n$ be the initial distribution for arrivals to these states. If $S + (\beta-\alpha)I$ is invertible, then the vector of the mean number in service in each phase of service is*

$$\mathrm{E}\left[Q_t\right] = \lambda_\infty\left(-S^{\mathrm{T}}\right)^{-1}(I - e^{S^{\mathrm{T}}t})\theta - (\lambda_0 - \lambda_\infty)\left(S^{\mathrm{T}} + (\beta-\alpha)I\right)^{-1}(e^{-(\beta-\alpha)t}I - e^{S^{\mathrm{T}}t})\theta \quad (2.21)$$

where $\lambda_\infty = \frac{\beta\lambda^*}{\beta-\alpha}$. Further, the vector of covariances between the intensity and each phase of service is

$$\begin{aligned}\mathrm{Cov}\left[\lambda_t, Q_t\right] &= \frac{\alpha(2\beta-\alpha)\lambda_\infty}{2(\beta-\alpha)}\left((\beta-\alpha)I - S^{\mathrm{T}}\right)^{-1}\left(I - e^{(S^{\mathrm{T}}-(\beta-\alpha)I)t}\right)\theta - \frac{\alpha\beta(\lambda_0-\lambda_\infty)}{\beta-\alpha} \\ &\quad \cdot\left(S^{\mathrm{T}}\right)^{-1}\left(e^{-(\beta-\alpha)t}I - e^{(S^{\mathrm{T}}-(\beta-\alpha)I)t}\right)\theta + \frac{\alpha^2(2\lambda_0-\lambda_\infty)}{2(\beta-\alpha)}\left(S^{\mathrm{T}} + (\beta-\alpha)I\right)^{-1} \\ &\quad \cdot\left(e^{-2(\beta-\alpha)t}I - e^{(S^{\mathrm{T}}-(\beta-\alpha)I)t}\right)\theta\,. \end{aligned} \quad (2.22)$$

Finally, the matrix of covariances between phases of service is given by

$$\begin{aligned}
\mathrm{Cov}\left[Q_t, Q_t\right] = {} & \frac{\alpha(2\beta-\alpha)\lambda_\infty}{2(\beta-\alpha)}\left((\beta-\alpha)I - S^{\mathrm{T}}\right)^{-1}\Bigg(2(\beta-\alpha)e^{S^{\mathrm{T}}t}M_{0,\theta,S}(t)e^{St} + \theta\theta^{\mathrm{T}} - e^{S^{\mathrm{T}}t}\theta\theta^{\mathrm{T}}e^{St} \\
& + e^{S^{\mathrm{T}}t}\theta\theta^{\mathrm{T}}\left(e^{-(\beta-\alpha)t}I - e^{St}\right)((\beta-\alpha)I + S)^{-1}((\beta-\alpha)I - S) + \left((\beta-\alpha)I - S^{\mathrm{T}}\right)\left((\beta-\alpha)I + S^{\mathrm{T}}\right)^{-1} \\
& \cdot\left(e^{-(\beta-\alpha)t}I - e^{S^{\mathrm{T}}t}\right)\theta\theta^{\mathrm{T}}e^{St}\Bigg)((\beta-\alpha)I - S)^{-1} + \frac{\alpha\beta(\lambda_0-\lambda_\infty)}{\beta-\alpha}\left(S^{\mathrm{T}}\right)^{-1}\Bigg((\beta-\alpha)e^{S^{\mathrm{T}}t}M_{-(\beta-\alpha),\theta,S}(t)e^{St} \\
& + e^{-(\beta-\alpha)t}\theta\theta^{\mathrm{T}} - e^{S^{\mathrm{T}}t}\theta\theta^{\mathrm{T}}e^{St} - e^{S^{\mathrm{T}}t}\theta\theta^{\mathrm{T}}\left(e^{-(\beta-\alpha)t}I - e^{St}\right)((\beta-\alpha)I + S)^{-1}S - S^{\mathrm{T}}\left((\beta-\alpha)I + S^{\mathrm{T}}\right)^{-1} \\
& \cdot\left(e^{-(\beta-\alpha)t}I - e^{S^{\mathrm{T}}t}\right)\theta\theta^{\mathrm{T}}e^{St}\Bigg)S^{-1} - \frac{\alpha^2(2\lambda_0-\lambda_\infty)}{2(\beta-\alpha)}((\beta-\alpha)I + S^{\mathrm{T}})^{-1}\Bigg(e^{-2(\beta-\alpha)t}\theta\theta^{\mathrm{T}} - e^{S^{\mathrm{T}}t}\theta\theta^{\mathrm{T}}e^{St} \\
& - e^{S^{\mathrm{T}}t}\theta\theta^{\mathrm{T}}\left(e^{-(\beta-\alpha)t}I - e^{St}\right) - \left(e^{-(\beta-\alpha)t}I - e^{S^{\mathrm{T}}t}\right)\theta\theta^{\mathrm{T}}e^{St}\Bigg)((\beta-\alpha)I + S)^{-1} - \lambda_\infty\mathrm{diag}\Big(\left(S^{\mathrm{T}}\right)^{-1} \\
& \cdot\left(I - e^{S^{\mathrm{T}}t}\right)\theta\Big) - (\lambda_0-\lambda_\infty)\mathrm{diag}\left(\left(S^{\mathrm{T}} + (\beta-\alpha)I\right)^{-1}\left(e^{-(\beta-\alpha)t}I - e^{S^{\mathrm{T}}t}\right)\theta\right) \qquad (2.23)
\end{aligned}$$

where all $t \geq 0$.

Proof. Throughout this proof we use the fact that a matrix being invertible implies that its transpose is invertible as well. To begin, we can see from Corollary 2.3.8 that

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[Q_t\right] = S^{\mathrm{T}}\mathrm{E}\left[Q_t\right] + \theta\mathrm{E}\left[\lambda_t\right] = S^{\mathrm{T}}\mathrm{E}\left[Q_t\right] + \theta\left(\lambda_\infty + (\lambda_0-\lambda_\infty)\,e^{-(\beta-\alpha)t}\right)$$

and so we apply Lemma 2.3.5. Let $\nu_1 = \theta\lambda_\infty$ and $\eta_1 = \gamma_1 = 0$, and let $\nu_2 = \theta(\lambda_0 - \lambda_\infty)$, $\eta_2 = 0$, and $\gamma_2 = -(\beta-\alpha)$. We assume that the queue starts empty. Then, we have

$$\begin{aligned}
\mathrm{E}\left[Q_t\right] = {} & -\left(S^{\mathrm{T}}\right)^{-1}\theta\lambda_\infty + \left(S^{\mathrm{T}}\right)^{-1}e^{-S^{\mathrm{T}}t}\theta\lambda_\infty - \left(S^{\mathrm{T}} + (\beta-\alpha)I\right)^{-1}\theta(\lambda_0-\lambda_\infty)e^{-(\beta-\alpha)t} \\
& + \left(S^{\mathrm{T}} + (\beta-\alpha)I\right)^{-1}e^{-S^{\mathrm{T}}t}\theta(\lambda_0-\lambda_\infty)
\end{aligned}$$

which now simplifies to the stated result. Note that S is invertible because it is diagonally dominant by definition and we have assumed the invertibility of $S + (\beta-\alpha)I$, which implies non-singularity of the respective transposes. We

find the stated result for $\mathrm{Cov}\,[\lambda_t, Q_t]$ through repeating the same technique to the corresponding differential equation systems, where again we make use of the linear algebraic representation. Thus, we are left to solve for the covariance matrix. Note that from Corollary 2.3.8, the variance of each phase and the covariance between phases can form one linear algebraic form as the covariance matrix, as shown below.

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{Cov}\,[Q_t, Q_t] &= S^{\mathrm{T}}\mathrm{Cov}\,[Q_t, Q_t] + \mathrm{Cov}\,[Q_t, Q_t]S + \theta\mathrm{Cov}\,[\lambda_t, Q_t]^{\mathrm{T}} + \mathrm{Cov}\,[\lambda_t, Q_t]\theta^{\mathrm{T}} \\ &\quad + \mathrm{diag}\left(\theta\mathrm{E}\,[\lambda_t] + S^{\mathrm{T}}\mathrm{E}\,[Q_t]\right) - S^{\mathrm{T}}\mathrm{diag}\,(\mathrm{E}\,[Q_t]) - \mathrm{diag}\,(\mathrm{E}\,[Q_t])S\end{aligned}$$

Using the product rule and multiplying through by matrix exponentials on the right and left, we can also express this as below:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\left(e^{-S^{\mathrm{T}}t}\mathrm{Cov}\,[Q_t, Q_t]e^{-St}\right) &= e^{-S^{\mathrm{T}}t}\theta\mathrm{Cov}\,[\lambda_t, Q_t]^{\mathrm{T}}e^{-St} + e^{-S^{\mathrm{T}}t}\mathrm{Cov}\,[\lambda_t, Q_t]\theta^{\mathrm{T}}e^{-St} \\ &\quad + e^{-S^{\mathrm{T}}t}\mathrm{diag}\left(\theta\mathrm{E}\,[\lambda_t] + S^{\mathrm{T}}\mathrm{E}\,[Q_t]\right)e^{-St} - e^{-S^{\mathrm{T}}t}S^{\mathrm{T}}\mathrm{diag}\,(\mathrm{E}\,[Q_t])e^{-St} \\ &\quad - e^{-S^{\mathrm{T}}t}\mathrm{diag}\,(\mathrm{E}\,[Q_t])S\,e^{-St}.\end{aligned}$$

For the pair of $\mathrm{Cov}\,[\lambda_t, Q_t]$ terms, we use Lemma 2.3.6 in conjunction with the explicit function for $\mathrm{Cov}\,[\lambda_t, Q_t]$ to find

$$\begin{aligned}&\int_0^t \left(e^{-S^{\mathrm{T}}s}\theta\mathrm{Cov}\,[\lambda_s, Q_s]^{\mathrm{T}}e^{-Ss} + e^{-S^{\mathrm{T}}s}\mathrm{Cov}\,[\lambda_s, Q_s]\theta^{\mathrm{T}}e^{-Ss}\right)\mathrm{d}s \\ &= \frac{\alpha(2\beta-\alpha)\lambda_\infty}{2(\beta-\alpha)}\left((\beta-\alpha)I - S^{\mathrm{T}}\right)^{-1}\Big(2(\beta-\alpha)M_{0,\theta,S}(t) + e^{-S^{\mathrm{T}}t}\theta\theta^{\mathrm{T}}e^{-St} - \theta\theta^{\mathrm{T}} + \theta\theta^{\mathrm{T}}\left(e^{-((\beta-\alpha)I+S)t} - I\right) \\ &\cdot ((\beta-\alpha)I + S)^{-1}((\beta-\alpha)I - S) + \left((\beta-\alpha)I - S^{\mathrm{T}}\right)\left((\beta-\alpha)I + S^{\mathrm{T}}\right)^{-1}\left(e^{-((\beta-\alpha)I+S^{\mathrm{T}})t} - I\right)\theta\theta^{\mathrm{T}}\Big) \\ &\cdot ((\beta-\alpha)I - S)^{-1} + \frac{\alpha\beta(\lambda_0-\lambda_\infty)}{\beta-\alpha}\left(S^{\mathrm{T}}\right)^{-1}\Big((\beta-\alpha)M_{-(\beta-\alpha),\theta,S}(t) + e^{-((\beta-\alpha)I+S^{\mathrm{T}})t}\theta\theta^{\mathrm{T}}e^{-St} - \theta\theta^{\mathrm{T}} \\ &- \theta\theta^{\mathrm{T}}\left(e^{-((\beta-\alpha)I+S)t} - I\right)((\beta-\alpha)I + S)^{-1}S - S^{\mathrm{T}}\left((\beta-\alpha)I + S^{\mathrm{T}}\right)^{-1}\left(e^{-((\beta-\alpha)I+S^{\mathrm{T}})t} - I\right)\theta\theta^{\mathrm{T}}\Big)S^{-1} \\ &- \frac{\alpha^2(2\lambda_0-\lambda_\infty)}{2(\beta-\alpha)}((\beta-\alpha)I + S^{\mathrm{T}})^{-1}\Big(e^{-(2(\beta-\alpha)I+S^{\mathrm{T}})t}\theta\theta^{\mathrm{T}}e^{-St} - \theta\theta^{\mathrm{T}} - \theta\theta^{\mathrm{T}}\left(e^{-((\beta-\alpha)I+S)t} - I\right) \\ &- \left(e^{-((\beta-\alpha)I+S^{\mathrm{T}})t} - I\right)\theta\theta^{\mathrm{T}}\Big)((\beta-\alpha)I + S)^{-1}\end{aligned}$$

and so we now integrate the remaining terms in the covariance matrix differential equations. Note that the product rule for three terms is $(fgh)' = f'gh + fg'h + fgh'$. We have already used this in concatenating the covariance matrix terms in the differential equation, and we can now make use of it again. Recall that $\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}[Q_t] = S^{\mathrm{T}}\mathrm{E}[Q_t] + \theta\mathrm{E}[\lambda_t]$. Using this realization, the integral of the remaining three terms is

$$\begin{aligned}
&\int_0^t \left(e^{-S^{\mathrm{T}}s}\mathrm{diag}\left(\theta\mathrm{E}[\lambda_s] + S^{\mathrm{T}}\mathrm{E}[Q_s]\right)e^{-Ss} - e^{-S^{\mathrm{T}}s}S^{\mathrm{T}}\mathrm{diag}\left(\mathrm{E}[Q_s]\right)e^{-Ss} - e^{-S^{\mathrm{T}}s}\mathrm{diag}\left(\mathrm{E}[Q_s]\right)S\,e^{-Ss}\right)\mathrm{d}s \\
&= e^{-S^{\mathrm{T}}t}\mathrm{diag}\left(\mathrm{E}[Q_t]\right)e^{-St} \\
&= -e^{-S^{\mathrm{T}}t}\mathrm{diag}\left(\left(S^{\mathrm{T}}\right)^{-1}\left(I - e^{S^{\mathrm{T}}t}\right)\theta\right)e^{-St}\lambda_\infty - e^{-S^{\mathrm{T}}t}\mathrm{diag}\left(\left(S^{\mathrm{T}} + (\beta-\alpha)I\right)^{-1}\left(e^{-(\beta-\alpha)t}I - e^{S^{\mathrm{T}}t}\right)\theta\right) \\
&\qquad \cdot e^{-St}(\lambda_0 - \lambda_\infty)
\end{aligned}$$

where we are justified in moving the differentiation through the diagonalization and distributing it across sums via the definition of diagonalization as a linear combination. Combining this with the integral for the covariance between the queue and intensity and multiplying each side by the corresponding exponentials, we achieve the stated result. □

As a brief example, consider a Hawkes process driven queueing system with infinite servers and suppose that the service is phase-type distributed with initial distribution $\theta = \mathbf{v}_1$ and the following sub-generator matrix:

$$S_{\mathrm{Cox}} = \begin{bmatrix} -4 & 3 & 0 & 0 & 0 \\ 0 & -2 & 1 & 0 & 0 \\ 0 & 0 & -3 & 2 & 0 \\ 0 & 0 & 0 & -5 & 4 \\ 0 & 0 & 0 & 0 & -1 \end{bmatrix}. \tag{2.24}$$

This is referred to as a Coxian distribution. It is characterized by each phase of service having an associated probability of either system departure or advance-

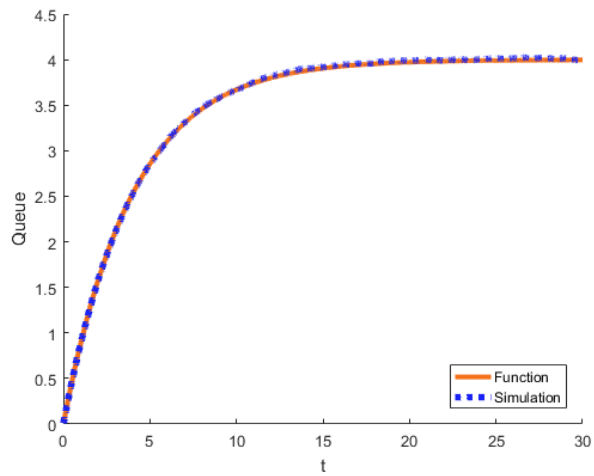


Figure 2.3: Example Mean of the $Hawkes/PH/\infty$ Queue with Sub-Generator Matrix S_{Cox} as in Equation 2.24.

ment to the next phase upon service completion. In this example, $\lambda^* = 1$, $\alpha = \frac{3}{4}$, and $\beta = 1$. The simulation is based on 100,000 replications.

Remark. We now note that the assumed nonsingularity of $S + (\beta - \alpha)I$ is necessary to implement the technical lemmas, but need not hold in order for a closed form solution to exist. If these conditions do not hold, one can instead make use of the structure of invertibility that is implied by a specific phase-type distribution. In Corollaries 2.3.10 and 2.3.11, we demonstrate this for Erlang and hyper-exponential service, respectively. Like we have seen in Theorem 2.3.9, these expressions can be found through solving systems of differential equations provided by Corollary 2.3.8.

We start with the case of service times following a Erlang distribution. In this case, we define $N \in \mathbb{R}^{n \times n}$ as the matrix of all ones on the first lower diagonal and zeros otherwise. Then, $S^{\mathrm{T}} = n\mu(N - I)$ for this phase-type distribution. Observe that N is a nilpotent matrix of a particular structure: for $k \in \mathbb{N}$, N^k is the matrix

of all ones on the kth lower diagonal if $k \leq n-1$ and is the zero matrix otherwise. Additionally, in this case $\theta = \mathbf{v}_1$ as all arrivals occur in the first phase. With this in hand, we see that

$$\begin{aligned}\left(M_{\gamma,\mathbf{v}_1,n\mu(I-N^{\mathrm{T}})}(t)\right)_{i,j} &= \left(M_{\gamma+2n\mu,\mathbf{v}_1,n\mu N^{\mathrm{T}}}(t)\right)_{i,j} \\ &= \begin{cases} \binom{i+j-2}{i-1}(n\mu)^{i+j-2}\frac{e^{(\gamma+2n\mu)t}\sum_{k=0}^{i+j-2}\frac{(-(\gamma+2n\mu)t)^k}{k!}-1}{(\gamma+2n\mu)^{i+j-1}} & \text{if } \gamma+2n\mu \neq 0 \\ \frac{(tn\mu)^{i+j-1}}{n\mu(i-1)!(j-1)!(i+j-1)} & \text{if } \gamma+2n\mu = 0 \end{cases}\end{aligned}$$

and we make use of this in the following corollary.

Corollary 2.3.10. *Consider a queueing system with arrivals occurring in accordance to a Hawkes process (λ_t, N_t) with dynamics given in Equation 1.1 with $\alpha < \beta$ and Erlang distributed service with n phases and mean $\frac{1}{\mu}$. Then, when $n\mu \neq \beta - \alpha$, the vector of mean number in each phase of service is given by*

$$\mathrm{E}\left[Q_t\right] = \frac{\lambda_\infty}{n\mu}(I - e^{n\mu(N-I)t})\mathbf{v} - (\lambda_0 - \lambda_\infty)\left(n\mu N - (n\mu - \beta + \alpha)I\right)^{-1}\left(e^{-(\beta-\alpha)t}I - e^{n\mu(N-I)t}\right)\mathbf{v}_1, \tag{2.25}$$

and when $n\mu = \beta - \alpha$, this vector is

$$\mathrm{E}\left[Q_t\right] = \frac{\lambda_\infty}{n\mu}(I - e^{n\mu(N-I)t})\mathbf{v} + (\lambda_0 - \lambda_\infty)\, e^{n\mu(N-I)t}x(t), \tag{2.26}$$

where $\lambda_\infty = \frac{\beta\lambda^*}{\beta-\alpha}$ and $x : \mathbb{R}^+ \to \mathbb{R}^n$ is such that $x_i(t) = \frac{(-n\mu)^{i-1}t^i}{i!}$. Further, when $n\mu \neq \beta - \alpha$ the vector of covariances between the number in each phase of service and the intensity is

$$\begin{aligned}\mathrm{Cov}\left[\lambda_t, Q_t\right] &= \lambda_\infty\left(\alpha + \frac{\alpha^2}{2(\beta-\alpha)}\right)\left((n\mu+\beta-\alpha)I - n\mu N\right)^{-1}\left(I - e^{(n\mu N-(n\mu+\beta-\alpha)I)t}\right)\mathbf{v}_1 \\ &+ \frac{\alpha\beta(\lambda_0-\lambda_\infty)}{n\mu(\beta-\alpha)}\left(e^{-(\beta-\alpha)t}I - e^{(n\mu N-(n\mu+\beta-\alpha)I)t}\right)\mathbf{v} + \frac{\alpha^2(2\lambda_0-\lambda_\infty)}{2(\beta-\alpha)}\left(n\mu N - (n\mu-\beta+\alpha)I\right)^{-1} \\ &\cdot\left(e^{-2(\beta-\alpha)t}I - e^{(n\mu N-(n\mu+\beta-\alpha)I)t}\right)\mathbf{v}_1,\end{aligned} \tag{2.27}$$

and when $n\mu = \beta - \alpha$, this is

$$\mathrm{Cov}\left[\lambda_t, Q_t\right] = \lambda_\infty \left(\frac{\alpha}{n\mu} + \frac{\alpha^2}{2(n\mu)^2}\right)(2I - N)^{-1}\left(I - e^{n\mu(N-2I)t}\right)\mathbf{v}_1 + (\lambda_0 - \lambda_\infty)\left(\frac{\alpha}{n\mu} + \frac{\alpha^2}{(n\mu)^2}\right)$$
$$\cdot \left(e^{-n\mu t} I - e^{n\mu(N-2I)t}\right)\mathbf{v} - \frac{\alpha^2(2\lambda_0 - \lambda_\infty)}{2n\mu} e^{n\mu(N-2I)t} x(t). \tag{2.28}$$

Finally, when $n\mu \neq \beta - \alpha$, the matrix of the covariance between the number in the phases of service is given by

$$\mathrm{Cov}\left[Q_t, Q_t\right] = \frac{\alpha(2\beta - \alpha)\lambda_\infty}{2(\beta - \alpha)}\left((n\mu + \beta - \alpha)I - n\mu N\right)^{-1}\Bigg(2(\beta - \alpha)e^{n\mu(N-I)t} M_{2n\mu,\mathbf{v}_1,n\mu N^{\mathrm{T}}}(t) e^{n\mu(N^{\mathrm{T}}-I)t}$$
$$+ \mathbf{v}_1\mathbf{v}_1{}^{\mathrm{T}} - e^{n\mu(N-I)t}\mathbf{v}_1\mathbf{v}_1{}^{\mathrm{T}} e^{n\mu(N^{\mathrm{T}}-I)t} + e^{n\mu(N-I)t}\mathbf{v}_1\mathbf{v}_1{}^{\mathrm{T}}\left(e^{-(\beta-\alpha)t} I - e^{n\mu(N^{\mathrm{T}}-I)t}\right)(n\mu N^{\mathrm{T}} - (n\mu - \beta + \alpha)I)^{-1}$$
$$\cdot \left((n\mu + \beta - \alpha)I - n\mu N^{\mathrm{T}}\right) + \left((n\mu + \beta - \alpha)I - n\mu N\right)(n\mu N - (n\mu - \beta + \alpha)I)^{-1}\left(e^{-(\beta-\alpha)t} I - e^{n\mu(N-I)t}\right)$$
$$\cdot \mathbf{v}_1\mathbf{v}_1{}^{\mathrm{T}} e^{n\mu(N^{\mathrm{T}}-I)t}\Bigg)\left((n\mu + \beta - \alpha)I - n\mu N^{\mathrm{T}}\right)^{-1} + \frac{\alpha\beta(\lambda_0 - \lambda_\infty)}{(n\mu)^2(\beta - \alpha)}(N - I)^{-1}\Bigg((\beta - \alpha)e^{n\mu(N-I)t}$$
$$\cdot M_{2n\mu-\beta+\alpha,\mathbf{v}_1,n\mu N^{\mathrm{T}}}(t) e^{n\mu(N^{\mathrm{T}}-I)t} + e^{-(\beta-\alpha)t}\mathbf{v}_1\mathbf{v}_1{}^{\mathrm{T}} - e^{n\mu(N-I)t}\mathbf{v}_1\mathbf{v}_1{}^{\mathrm{T}} e^{n\mu(N^{\mathrm{T}}-I)t} - n\mu e^{n\mu(N-I)t}\mathbf{v}_1\mathbf{v}_1{}^{\mathrm{T}}$$
$$\cdot \left(e^{-(\beta-\alpha)t} I - e^{n\mu(N^{\mathrm{T}}-I)t}\right)(n\mu N^{\mathrm{T}} - (n\mu - \beta + \alpha)I)^{-1}(N^{\mathrm{T}} - I) - n\mu(N - I)(n\mu N - (n\mu - \beta + \alpha)I)^{-1}$$
$$\cdot \left(e^{-(\beta-\alpha)t} I - e^{n\mu(N-I)t}\right)\mathbf{v}_1\mathbf{v}_1{}^{\mathrm{T}} e^{n\mu(N^{\mathrm{T}}-I)t}\Bigg)(N^{\mathrm{T}} - I)^{-1} - \frac{\alpha^2(2\lambda_0 - \lambda_\infty)}{2(\beta - \alpha)}(n\mu N - (n\mu - \beta + \alpha)I)^{-1}$$
$$\cdot \Bigg(e^{-2(\beta-\alpha)t}\mathbf{v}_1\mathbf{v}_1{}^{\mathrm{T}} - e^{n\mu(N-I)t}\mathbf{v}_1\mathbf{v}_1{}^{\mathrm{T}} e^{n\mu(N^{\mathrm{T}}-I)t} - e^{n\mu(N-I)t}\mathbf{v}_1\mathbf{v}_1{}^{\mathrm{T}}\left(e^{-(\beta-\alpha)t} I - e^{n\mu(N^{\mathrm{T}}-I)t}\right)$$
$$- \left(e^{-(\beta-\alpha)t} I - e^{n\mu(N-I)t}\right)\mathbf{v}_1\mathbf{v}_1{}^{\mathrm{T}} e^{n\mu(N^{\mathrm{T}}-I)t}\Bigg)(n\mu N^{\mathrm{T}} - (n\mu - \beta + \alpha)I)^{-1} + \frac{\lambda_\infty}{n\mu}\mathrm{diag}\left(\left(I - e^{n\mu(N-I)t}\right)\mathbf{v}\right)$$
$$- (\lambda_0 - \lambda_\infty)\mathrm{diag}\left((n\mu N - (n\mu - \beta + \alpha)I)^{-1}\left(e^{-(\beta-\alpha)t} I - e^{n\mu(N-I)t}\right)\mathbf{v}_1\right), \tag{2.29}$$

whereas when $n\mu = \beta - \alpha$, this matrix is

$$\mathrm{Cov}\left[Q_t, Q_t\right] = \mathrm{diag}\left(\frac{\lambda_\infty}{n\mu}(I - e^{n\mu(N-I)t})\mathbf{v} + (\lambda_0 - \lambda_\infty)\, e^{n\mu(N-I)t} x(t)\right) + e^{n\mu(N-I)t}\Bigg(\lambda_\infty\left(\frac{\alpha}{n\mu} + \frac{\alpha^2}{2(n\mu)^2}\right)$$
$$\cdot \left(\left(M_{2n\mu,\mathbf{v}_1,n\mu N^{\mathrm{T}}}(t) - x(t)\mathbf{v}_1{}^{\mathrm{T}}\right)\left(2I - N^{\mathrm{T}}\right)^{-1} + (2I - N)^{-1}\left(M_{2n\mu,\mathbf{v}_1,n\mu N^{\mathrm{T}}}(t) - \mathbf{v}_1 x^{\mathrm{T}}(t)\right)\right) + (\lambda_0 - \lambda_\infty)$$
$$\cdot \left(\frac{\alpha}{n\mu} + \frac{\alpha^2}{(n\mu)^2}\right)\left(M_{2n\mu,\mathbf{v}_1,n\mu N^{\mathrm{T}}}(t)\left(I - N^{\mathrm{T}}\right)^{-1} + (I - N)^{-1} M_{2n\mu,\mathbf{v}_1,n\mu N^{\mathrm{T}}}(t) - x(t)\mathbf{v}^{\mathrm{T}} - \mathbf{v}x^{\mathrm{T}}(t)\right)$$
$$- \frac{\alpha^2(2\lambda_0 - \lambda_\infty)}{2n\mu}\left(X(t) + X^{\mathrm{T}}(t)\right)\Bigg)e^{n\mu(N^{\mathrm{T}}-I)t}, \tag{2.30}$$

where all $t \geq 0$ and $X : \mathbb{R}^+ \to \mathbb{R}^{n\times n}$ is such that $X_{i,j}(t) = \frac{(-n\mu)^{i+j-2} t^{i+j-1}}{(i-1)!\, j!\,(i+j)}$.

As with the Erlang, we also provide explicit formulas for the hyper-exponential distribution. In this case we have that $S = -D$ where D is a diagonal matrix of the rates of service in each phase. This allows it to commute with the symmetric $\theta\theta^{\mathrm{T}}$, giving us

$$M_{\gamma,\theta,-D}(t) = \int_0^t e^{(\gamma I+D)s}\theta\theta^{\mathrm{T}}e^{Ds}\,\mathrm{d}s = \int_0^t e^{(\gamma I+2D)s}\,\mathrm{d}s\theta\theta^{\mathrm{T}} = (\gamma I + 2D)^{-1}\left(e^{(\gamma I+2D)t} - I\right)\theta\theta^{\mathrm{T}}$$

as long as $\gamma I + 2D$ is invertible. However, we also seek to address the case where $(\beta - \alpha)I + S = (\beta - \alpha)I - D$ is not invertible. In the hyper-exponential service setting, $(\beta - \alpha)I - D$ being singular implies that some $\mu_i = \beta - \alpha$, but it is not clear which or for how many μ_i this is the case. So, we instead use the element-level equations in Corollary 2.3.8 to solve for the explicit expressions. This method is preferable to the linear algebra approach for hyper-exponential service since in this setting $\mu_{ij} = 0$ for every i and j.

Corollary 2.3.11. *Consider a queueing system with arrivals occurring in accordance to a Hawkes process (λ_t, N_t) with dynamics given in Equation 1.1 with $\alpha < \beta$ and hyper-exponential distributed service with n phases and distinct service rates $\mu_1, \ldots, \mu_n$. Then, the mean number in phase $i \in \{1, \ldots, n\}$ of service is*

$$\mathrm{E}\left[Q_{t,i}\right] = \begin{cases} \frac{\lambda_\infty}{\mu_i}\left(1 - e^{-\mu_i t}\right)\theta_i + \frac{\lambda_0 - \lambda_\infty}{\mu_i - \beta + \alpha}\left(e^{-(\beta-\alpha)t} - e^{-\mu_i t}\right)\theta_i & \text{if } \mu_i \neq \beta - \alpha, \\ \frac{\lambda_\infty}{\mu_i}\left(1 - e^{-\mu_i t}\right)\theta_i + (\lambda_0 - \lambda_\infty)\,\theta_i t e^{-\mu_i t} & \text{if } \mu_i = \beta - \alpha, \end{cases} \tag{2.31}$$

where $\lambda_\infty = \frac{\beta\lambda^*}{\beta-\alpha}$. Furthermore the covariance between the number in phase i of service

and the intensity is

$$\mathrm{Cov}\left[\lambda_t, Q_{t,i}\right] = \begin{cases} \frac{\alpha\theta_i(2\beta-\alpha)\lambda_\infty}{2(\beta-\alpha)(\mu_i+\beta-\alpha)}\left(1-e^{-(\mu_i+\beta-\alpha)t}\right) + \frac{\alpha\beta\theta_i(\lambda_0-\lambda_\infty)}{\mu_i(\beta-\alpha)}\big(e^{-(\beta-\alpha)t} \\ \quad -e^{-(\mu_i+\beta-\alpha)t}\big) - \frac{\alpha^2\theta_i(2\lambda_0-\lambda_\infty)}{2(\beta-\alpha)(\mu_i-\beta+\alpha)}\left(e^{-2(\beta-\alpha)t} - e^{-(\mu_i+\beta-\alpha)t}\right) & \text{if } \mu_i \neq \beta-\alpha, \\ \frac{\alpha\theta_i(2\mu_i+\alpha)\lambda_\infty}{4\mu_i^2}\left(1-e^{-2\mu_i t}\right) + \frac{\alpha\beta\theta_i(\lambda_0-\lambda_\infty)}{\mu_i^2}\big(e^{-\mu_i t} - e^{-2\mu_i t}\big) \\ \quad -\frac{\alpha^2\theta_i(2\lambda_0-\lambda_\infty)}{2\mu_i} t e^{-2\mu_i t} & \text{if } \mu_i = \beta-\alpha. \end{cases} \tag{2.32}$$

Then, the covariance between the number in phase i of service and the number in phase j of service where $i, j \in \{1, \ldots, \}$ and $i \neq j$ is

$$\mathrm{Cov}\left[Q_{t,i}, Q_{t,j}\right] = \begin{cases} \frac{\alpha\theta_i\theta_j(2\beta-\alpha)\lambda_\infty}{2(\beta-\alpha)(\mu_j+\beta-\alpha)}\left(\frac{1-e^{-(\mu_i+\mu_j)t}}{\mu_i+\mu_j} - \frac{e^{-(\mu_j+\beta-\alpha)t}-e^{-(\mu_i+\mu_j)t}}{\mu_i-\beta+\alpha}\right) \\ \quad +\frac{\alpha\beta\theta_i\theta_j(\lambda_0-\lambda_\infty)}{\mu_j(\beta-\alpha)}\left(\frac{e^{-(\beta-\alpha)t}-e^{-(\mu_i+\mu_j)t}}{\mu_i+\mu_j-\beta+\alpha} - \frac{e^{-(\mu_j+\beta-\alpha)t}-e^{-(\mu_i+\mu_j)t}}{\mu_i-\beta+\alpha}\right) \\ \quad -\frac{\alpha^2\theta_i\theta_j(2\lambda_0-\lambda_\infty)}{2(\beta-\alpha)(\mu_j-\beta+\alpha)}\left(\frac{e^{2(\beta-\alpha)t}-e^{-(\mu_i+\mu_j)t}}{\mu_i+\mu_j-2\beta+2\alpha} - \frac{e^{-(\mu_j+\beta-\alpha)t}-e^{-(\mu_i+\mu_j)t}}{\mu_i-\beta+\alpha}\right) \\ \quad +\frac{\alpha\theta_i\theta_j(2\beta-\alpha)\lambda_\infty}{2(\beta-\alpha)(\mu_i+\beta-\alpha)}\left(\frac{1-e^{-(\mu_i+\mu_j)t}}{\mu_i+\mu_j} - \frac{e^{-(\mu_i+\beta-\alpha)t}-e^{-(\mu_i+\mu_j)t}}{\mu_j-\beta+\alpha}\right) \\ \quad +\frac{\alpha\beta\theta_i\theta_j(\lambda_0-\lambda_\infty)}{\mu_i(\beta-\alpha)}\left(\frac{e^{-(\beta-\alpha)t}-e^{-(\mu_i+\mu_j)t}}{\mu_i+\mu_j-\beta+\alpha} - \frac{e^{-(\mu_i+\beta-\alpha)t}-e^{-(\mu_i+\mu_j)t}}{\mu_j-\beta+\alpha}\right) \\ \quad -\frac{\alpha^2\theta_i\theta_j(2\lambda_0-\lambda_\infty)}{2(\beta-\alpha)(\mu_i-\beta+\alpha)}\left(\frac{e^{2(\beta-\alpha)t}-e^{-(\mu_i+\mu_j)t}}{\mu_i+\mu_j-2\beta+2\alpha} - \frac{e^{-(\mu_i+\beta-\alpha)t}-e^{-(\mu_i+\mu_j)t}}{\mu_j-\beta+\alpha}\right) & \text{if } \mu_i \neq \beta-\alpha \neq \mu_j, \\ \frac{\alpha\theta_i\theta_j(2\beta-\alpha)\lambda_\infty}{4\mu_j^2}\left(\frac{1-e^{-(\mu_i+\mu_j)t}}{\mu_i+\mu_j} - \frac{e^{-2\mu_j t}-e^{-(\mu_i+\mu_j)t}}{\mu_i-\mu_j}\right) + \frac{\alpha\beta\theta_i\theta_j(\lambda_0-\lambda_\infty)}{\mu_j^2} \\ \quad \cdot\left(\frac{e^{-\mu_j t}-e^{-(\mu_i+\mu_j)t}}{\mu_i} - \frac{e^{-2\mu_j t}-e^{-(\mu_i+\mu_j)t}}{\mu_i-\mu_j}\right) - \frac{\alpha^2\theta_i\theta_j(2\lambda_0-\lambda_\infty)}{2\mu_j} \\ \quad \cdot\left(\frac{te^{-2\mu_i t}}{\mu_j-\mu_i} + \frac{e^{-(\mu_i+\mu_j)t}-e^{-2\mu_i t}}{(\mu_j-\mu_i)^2}\right) + \frac{\alpha\theta_i\theta_j(2\beta-\alpha)\lambda_\infty}{2\mu_j(\mu_i+\mu_j)}\left(\frac{1-e^{-(\mu_i+\mu_j)t}}{\mu_i+\mu_j}\right. \\ \quad \left.-te^{-(\mu_i+\mu_j)t}\right) + \frac{\alpha\beta\theta_i\theta_j(\lambda_0-\lambda_\infty)}{\mu_i\mu_j}\left(\frac{e^{-\mu_j t}-e^{-(\mu_i+\mu_j)t}}{\mu_i} - te^{-(\mu_i+\mu_j)t}\right) \\ \quad -\frac{\alpha^2\theta_i\theta_j(2\lambda_0-\lambda_\infty)}{2\mu_j(\mu_i-\mu_j)}\left(\frac{e^{2\mu_j t}-e^{-(\mu_i+\mu_j)t}}{\mu_i-\mu_j} - te^{-(\mu_i+\mu_j)t}\right) & \text{if } \mu_i \neq \beta-\alpha = \mu_j, \end{cases} \tag{2.33}$$

Finally, the variance of the number in phase $i \in \{1, \ldots, n\}$ of service is given by

$$
\mathrm{Var}\left(Q_{t,i}\right) = \begin{cases}
\frac{\lambda_\infty \theta_i}{\mu_i}\left(1 - e^{-\mu_i t}\right) + \frac{\alpha\theta_i^2(2\beta-\alpha)\lambda_\infty}{2\mu_i(\beta-\alpha)(\mu_i+\beta-\alpha)}\left(1 - e^{-2\mu_i t}\right) - \Big(\frac{\alpha\theta_i^2(2\beta-\alpha)\lambda_\infty}{(\beta-\alpha)(\mu_i+\beta-\alpha)} \\
\quad + \frac{2\alpha\beta\theta_i^2(\lambda_0-\lambda_\infty)}{\mu_i(\beta-\alpha)} - \frac{\alpha^2\theta_i^2(2\lambda_0-\lambda_\infty)}{(\beta-\alpha)(\mu_i-\beta+\alpha)}\Big)\frac{e^{-(\mu_i+\beta-\alpha)t}-e^{-2\mu_i t}}{\mu_i-\beta+\alpha} + \Big((\lambda_0 - \lambda_\infty)\theta_i \\
\quad + \frac{\mu_i(\lambda_0-\lambda_\infty)\theta_i}{\mu_i-\beta+\alpha} + \frac{2\alpha\beta\theta_i^2(\lambda_0-\lambda_\infty)}{\mu_i(\beta-\alpha)}\Big)\frac{e^{-(\beta-\alpha)t}-e^{-2\mu_i t}}{2\mu_i-\beta+\alpha} - \frac{\alpha^2\theta_i^2(2\lambda_0-\lambda_\infty)}{2(\beta-\alpha)(\mu_i-\beta+\alpha)^2} \\
\quad \cdot\left(e^{-2(\beta-\alpha)t} - e^{-2\mu_i t}\right) - \frac{(\lambda_0-\lambda_\infty)\theta_i}{\mu_i-\beta+\alpha}\left(e^{-\mu_i t} - e^{-2\mu_i t}\right) & \text{if } \mu_i \neq \beta-\alpha \neq 2\mu_i, \\
\frac{\lambda_\infty \theta_i}{\mu_i}\left(1 - e^{-\mu_i t}\right) + \frac{\alpha\theta_i^2(2\beta-\alpha)\lambda_\infty}{2\mu_i(\beta-\alpha)(\mu_i+\beta-\alpha)}\left(1 - e^{-2\mu_i t}\right) - \Big(\frac{\alpha\theta_i^2(2\beta-\alpha)\lambda_\infty}{(\beta-\alpha)(\mu_i+\beta-\alpha)} \\
\quad + \frac{2\alpha\beta\theta_i^2(\lambda_0-\lambda_\infty)}{\mu_i(\beta-\alpha)} - \frac{\alpha^2\theta_i^2(2\lambda_0-\lambda_\infty)}{(\beta-\alpha)(\mu_i-\beta+\alpha)}\Big)\frac{e^{-(\mu_i+\beta-\alpha)t}-e^{-2\mu_i t}}{\mu_i-\beta+\alpha} + \Big((\lambda_0 - \lambda_\infty)\theta_i \\
\quad + \frac{\mu_i(\lambda_0-\lambda_\infty)\theta_i}{\mu_i-\beta+\alpha} + \frac{2\alpha\beta\theta_i^2(\lambda_0-\lambda_\infty)}{\mu_i(\beta-\alpha)}\Big)te^{-2\mu_i t} - \frac{\alpha^2\theta_i^2(2\lambda_0-\lambda_\infty)}{2(\beta-\alpha)(\mu_i-\beta+\alpha)^2}\big(e^{-2(\beta-\alpha)t} \\
\quad -e^{-2\mu_i t}\big) - \frac{(\lambda_0-\lambda_\infty)\theta_i}{\mu_i-\beta+\alpha}\left(e^{-\mu_i t} - e^{-2\mu_i t}\right) & \text{if } 2\mu_i = \beta-\alpha, \\
\frac{\lambda_\infty \theta_i}{\mu_i}\left(1 - e^{-\mu_i t}\right) + \frac{\alpha\theta_i^2(2\beta-\alpha)\lambda_\infty}{4\mu_i^3}\left(1 - e^{-2\mu_i t}\right) - \Big(\frac{\alpha\theta_i^2(2\beta-\alpha)\lambda_\infty}{2\mu_i^2} \\
\quad + \frac{2\alpha\beta\theta_i^2(\lambda_0-\lambda_\infty)}{\mu_i^2}\Big)te^{-2\mu_i t} + \Big((\lambda_0 - \lambda_\infty)\theta_i + \frac{2\alpha\beta\theta_i^2(\lambda_0-\lambda_\infty)}{\mu_i^2}\Big) \\
\quad \cdot\frac{e^{-\mu_i t}-e^{-2\mu_i t}}{\mu_i} - \frac{\alpha^2\theta_i^2(2\lambda_0-\lambda_\infty)}{2\mu_i}t^2e^{-2\mu_i t} + (\lambda_0 - \lambda_\infty)\theta_i\Big(\frac{te^{-\mu_i t}}{\mu_i} \\
\quad + \frac{e^{-2\mu_i t}-e^{-\mu_i t}}{\mu_i^2}\Big) & \text{if } \mu_i = \beta-\alpha,
\end{cases} \tag{2.34}
$$

where all $t \geq 0$.

We now note that in both Corollary 2.3.10 and Corollary 2.3.11, taking $n = 1$ reduces the setting to exponential service. We demonstrate the simplification and use of the singe-phase expressions in finding the auto-covariance of the $Hawkes/M/\infty$ queue, shown in Proposition 2.3.14. We also note that these findings compare quite nicely to simulations in numerical demonstrations. In Subsection 2.3.6, we provide several example figures of these equations and their simulated counterparts.

Now that we have investigated the transient behavior of the $Hawkes/PH/\infty$ queue for a variety of settings it is natural to consider the behavior of the system in steady-state. This, along with the behavior of the system with an unstable arrival process, is the focus of the next subsection.

2.3.3 Limiting Behavior of the $Hawkes/PH/\infty$ Queue

In many situations, the steady-state behavior of a queueing system may be of particular interest. With that in mind, we now investigate the mean and variance of the $Hawkes/PH/\infty$ queue as time goes to infinity.

Corollary 2.3.12. *Consider a queueing system with arrivals occurring in accordance to a Hawkes process (λ_t, N_t) with dynamics given in Equation 1.1 and phase-type distributed service. Let $S \in \mathbb{R}^{n\times n}$ be the sub-generator matrix for the transient states in the phase-distribution CTMC and let $\theta \in [0,1]^n$ be the initial distribution for arrivals to these states. Then, the steady-state mean number in each phase of service is given by the vector*

$$\mathcal{Q}_\infty \equiv \lim_{t\to\infty} \mathrm{E}\left[Q_t\right] = \lambda_\infty\left(-S^{\mathrm{T}}\right)^{-1}\theta \tag{2.35}$$

where $\lambda_\infty = \frac{\beta\lambda^*}{\beta-\alpha}$. Further, the vector of steady-state covariances between the number in each phase of service and the intensity is

$$\mathcal{C}_\infty \equiv \lim_{t\to\infty} \mathrm{Cov}\left[\lambda_t, Q_t\right] = \lambda_\infty \frac{\alpha(2\beta-\alpha)}{2(\beta-\alpha)}\left((\beta-\alpha)I - S^{\mathrm{T}}\right)^{-1}\theta\,. \tag{2.36}$$

Finally, the matrix of steady-state covariances between each phase of service $\lim_{t\to\infty}\mathrm{Cov}\left[Q_t, Q_t\right]$, denoted $\mathcal{V}_\infty$, is given by the solution to the Lyapunov equation

$$S^{\mathrm{T}}\mathcal{V}_\infty + \mathcal{V}_\infty S + \mathcal{M} = 0 \tag{2.37}$$

where $\mathcal{M} = \theta C_\infty^{\mathrm{T}} + C_\infty \theta^{\mathrm{T}} - S^{\mathrm{T}} \mathrm{diag}\,(Q_\infty) - \mathrm{diag}\,(Q_\infty) S$. *If* S *is symmetric, then* $\mathcal{V}_\infty = -\frac{1}{2} S^{-1} \mathcal{M}$.

Proof. The proof follows by either taking the limit of the equations in Theorem 2.3.9 or setting the corresponding differential equations to 0 and finding the equilibrium solution. □

Remark. We note that in steady-state the invertibility conditions from Theorem 2.3.9 are no longer necessary. We can further observe that these equations reveal an interesting relationship among these steady-state values for the case of single phase service. For μ as the rate of exponential service, Corollary 2.3.12 yields

$$\mathcal{V}_\infty = Q_\infty + \frac{1}{\mu} C_\infty = \frac{\lambda_\infty}{\mu} \left(1 + \frac{\alpha(2\beta - \alpha)}{2(\beta - \alpha)(\mu + \beta - \alpha)} \right). \tag{2.38}$$

Thus, we have that the steady-state variance of the number in system for the $Hawkes/M/\infty$ queue is equal to the mean number in system plus the expected service duration times the steady-state covariance between the number in system and the intensity. Thus this provides an explicit contrast with Poisson-driven queues, as the steady-state distribution of a $M/M/\infty$ system is known to be Poisson distributed with rate equal to the steady-state mean number in system. This implies that the steady-state variance for such a queue is equal to its steady-state mean, unlike the relationship we observe for the $Hawkes/M/\infty$ system in Equation 2.38.

However, as we have noted, if $\alpha \geq \beta$ the Hawkes process is unstable and so steady-state analysis of the queue will not apply. Thus, in this scenario we instead investigate the transient behavior of the mean of the queue under the unstable arrival process.

Corollary 2.3.13. *Consider a queueing system with arrivals occurring in accordance to a Hawkes process (λ_t, N_t) with dynamics given in Equation 1.1 with $\alpha \geq \beta$ and phase-type distributed service. Let $S \in \mathbb{R}^{n\times n}$ be the sub-generator matrix for the transient states in the phase-distribution CTMC and let $\theta \in [0,1]^n$ be the initial distribution for arrivals to these states. Then the vector of mean number in service in each phase of service is given by*

$$\begin{aligned}\mathrm{E}\left[Q_t\right] = &\left((\alpha-\beta)I - S^{\mathrm{T}}\right)^{-1}\left(e^{(\alpha-\beta)t}I - e^{S^{\mathrm{T}}t}\right)\theta\left(\frac{\beta\lambda^*}{\alpha-\beta} + \lambda_0\right)\\ &+ (S^{\mathrm{T}})^{-1}\left(I - e^{S^{\mathrm{T}}t}\right)\theta\frac{\beta\lambda^*}{\alpha-\beta}\end{aligned} \tag{2.39}$$

when $\alpha > \beta$ and

$$\mathrm{E}\left[Q_t\right] = -(S^{\mathrm{T}})^{-1}\left(I - e^{S^{\mathrm{T}}t}\right)\theta(\lambda_0 - \beta\lambda^*) - (S^{\mathrm{T}})^{-1}\theta\beta\lambda^* t \tag{2.40}$$

when $\alpha = \beta$.

2.3.4 Auto-covariance of the $Hawkes/PH/\infty$ Queue

We now consider the auto-covariance of the number in this queueing system, $Q_t \in \mathbb{R}^n$. Analogous to the auto-covariance for the number of arrivals from the Hawkes process discussed in Subsection 2.2, this matrix quantity is defined as

$$\mathrm{Cov}\left[Q_t, Q_{t-\tau}\right] = \mathrm{E}\left[Q_t Q_{t-\tau}^{\mathrm{T}}\right] - \mathrm{E}\left[Q_t\right]\mathrm{E}\left[Q_{t-\tau}\right]^{\mathrm{T}}$$

where $t \geq \tau \geq 0$ and otherwise the covariance is equal to 0. For an infinite server queue with Hawkes process arrivals and phase-type distributed service, the findings in Subsection 2.3.2 give us expressions for $\mathrm{E}\left[Q_t\right]$ and $\mathrm{E}\left[Q_{t-\tau}\right]$. Let $\mathcal{F}_s$ be the filtration of the queueing system, the Hawkes process, and the intensity

at time $s \geq 0$. Then, assuming $S + (\beta - \alpha)I$ is invertible, conditional expectation yields

$$\begin{aligned}
\mathrm{E}\left[Q_t Q_{t-\tau}^{\mathrm{T}}\right] &= \mathrm{E}\left[\mathrm{E}\left[Q_t \mid F_{t-\tau}\right] Q_{t-\tau}^{\mathrm{T}}\right] \\
&= \mathrm{E}\Big[\Big(\lambda_\infty \left(-S^{\mathrm{T}}\right)^{-1} (I - e^{S^{\mathrm{T}}\tau})\theta - (\lambda_{t-\tau} - \lambda_\infty)\left(S^{\mathrm{T}} + (\beta - \alpha)I\right)^{-1} (e^{-(\beta-\alpha)\tau} I \\
&\quad - e^{S^{\mathrm{T}}\tau})\theta + e^{S^{\mathrm{T}}\tau} Q_{t-\tau}\Big) Q_{t-\tau}^{\mathrm{T}}\Big] \\
&= \lambda_\infty \left(-S^{\mathrm{T}}\right)^{-1} (I - e^{S^{\mathrm{T}}\tau})\theta \mathrm{E}\left[Q_{t-\tau}\right]^{\mathrm{T}} - \left(S^{\mathrm{T}} + (\beta - \alpha)I\right)^{-1} (e^{-(\beta-\alpha)\tau} I - e^{S^{\mathrm{T}}\tau})\theta \\
&\quad \cdot \left(\mathrm{E}\left[\lambda_{t-\tau} Q_{t-\tau}^{\mathrm{T}}\right] - \lambda_\infty \mathrm{E}\left[Q_{t-\tau}\right]^{\mathrm{T}}\right) + e^{S^{\mathrm{T}}\tau} \mathrm{E}\left[Q_{t-\tau} Q_{t-\tau}^{\mathrm{T}}\right]
\end{aligned}$$

by application of the expression for the vector of the mean number in each phase given in Theorem 2.3.9, modified to start at time $t - \tau$. Upon recognizing that $\mathrm{E}\left[\lambda_{t-\tau} Q_{t-\tau}^{\mathrm{T}}\right] = \mathrm{Cov}\left[\lambda_{t-\tau}, Q_{t-\tau}\right] + \mathrm{E}\left[\lambda_{t-\tau}\right]\mathrm{E}\left[Q_{t-\tau}\right]^{\mathrm{T}}$ and $\mathrm{E}\left[Q_{t-\tau} Q_{t-\tau}^{\mathrm{T}}\right] = \mathrm{Cov}\left[Q_{t-\tau}, Q_{t-\tau}^{\mathrm{T}}\right] + \mathrm{E}\left[Q_{t-\tau}\right]\mathrm{E}\left[Q_{t-\tau}\right]^{\mathrm{T}}$, we have that

$$\begin{aligned}
\mathrm{Cov}\left[Q_t, Q_{t-\tau}\right] &= \lambda_\infty \left(-S^{\mathrm{T}}\right)^{-1} \left(I - e^{S^{\mathrm{T}}\tau}\right)\theta \mathrm{E}\left[Q_{t-\tau}\right]^{\mathrm{T}} - \left(S^{\mathrm{T}} + (\beta - \alpha)I\right)^{-1} \left(e^{-(\beta-\alpha)\tau} I - e^{S^{\mathrm{T}}\tau}\right) \\
&\quad \cdot \theta \left(\mathrm{Cov}\left[\lambda_{t-\tau}, Q_{t-\tau}\right]^{\mathrm{T}} + \mathrm{E}\left[\lambda_{t-\tau}\right]\mathrm{E}\left[Q_{t-\tau}\right]^{\mathrm{T}} - \lambda_\infty \mathrm{E}\left[Q_{t-\tau}\right]^{\mathrm{T}}\right) + e^{S^{\mathrm{T}}\tau} \mathrm{Cov}\left[Q_{t-\tau}, Q_{t-\tau}\right] \\
&\quad + \left(e^{S^{\mathrm{T}}\tau} \mathrm{E}\left[Q_{t-\tau}\right] - \mathrm{E}\left[Q_t\right]\right) \mathrm{E}\left[Q_{t-\tau}\right]^{\mathrm{T}}
\end{aligned} \tag{2.41}$$

and that each term in this expression can be calculated by applying Theorem 2.3.9. In this section we give an explicit expression for the auto-covariance of the $Hawkes/M/\infty$ queue. In this setting with service rate μ, the same approach as above yields

$$\begin{aligned}
\mathrm{Cov}\left[Q_t, Q_{t-\tau}\right] &= \frac{\lambda_\infty}{\mu}\left(1 - e^{-\mu\tau}\right) \mathrm{E}\left[Q_{t-\tau}\right] + e^{-\mu\tau} \mathrm{Var}\left(Q_{t-\tau}\right) + \mathrm{Cov}\left[\lambda_{t-\tau}, Q_{t-\tau}\right] \frac{e^{-(\beta-\alpha)\tau} - e^{-\mu\tau}}{\mu - \beta + \alpha} \\
&\quad + \left(\mathrm{E}\left[\lambda_{t-\tau}\right] - \lambda_\infty\right) \mathrm{E}\left[Q_{t-\tau}\right] \frac{e^{-(\beta-\alpha)\tau} - e^{-\mu\tau}}{\mu - \beta + \alpha} + e^{-\mu\tau} \mathrm{E}\left[Q_{t-\tau}\right]^2 - \mathrm{E}\left[Q_t\right]\mathrm{E}\left[Q_{t-\tau}\right]
\end{aligned} \tag{2.42}$$

when $\mu \neq \beta - \alpha$ and

$$\begin{aligned}\mathrm{Cov}\left[Q_t, Q_{t-\tau}\right] = \frac{\lambda_\infty}{\mu}\left(1 - e^{-\mu\tau}\right)\mathrm{E}\left[Q_{t-\tau}\right] + e^{-\mu\tau}\mathrm{Var}\left(Q_{t-\tau}\right) + \mathrm{Cov}\left[\lambda_{t-\tau}, Q_{t-\tau}\right]\tau e^{-\mu\tau} \\ + \left(\mathrm{E}\left[\lambda_{t-\tau}\right] - \lambda_\infty\right)\mathrm{E}\left[Q_{t-\tau}\right]\tau e^{-\mu\tau} + e^{-\mu\tau}\mathrm{E}\left[Q_{t-\tau}\right]^2 - \mathrm{E}\left[Q_t\right]\mathrm{E}\left[Q_{t-\tau}\right]\end{aligned} \tag{2.43}$$

when $\mu = \beta - \alpha$, where each of these makes use of Corollary 2.3.11 with $n = 1$, $\theta_1 = 1$, and $\mu_i = \mu$. These expressions are made explicit in the following proposition.

Proposition 2.3.14. *Consider a queueing system with arrivals occurring in accordance to a Hawkes process (λ_t, N_t) with dynamics given in Equation 1.1 with $\alpha < \beta$ and exponentially distributed service with rate μ. Then, for $t \geq \tau \geq 0$ the auto-covariance of the number in system is*

$$\begin{aligned}\mathrm{Cov}\left[Q_t, Q_{t-\tau}\right] &= \frac{\lambda_\infty}{\mu}\left(1 - e^{-\mu\tau}\right)\left(\frac{\lambda_\infty}{\mu}\left(1 - e^{-\mu(t-\tau)}\right) + \frac{\lambda_0 - \lambda_\infty}{\mu - \beta + \alpha}\left(e^{-(\beta-\alpha)(t-\tau)} - e^{-\mu(t-\tau)}\right)\right) + \frac{\lambda_\infty}{\mu} \\ &\cdot \left(e^{-\mu\tau} - e^{-\mu t}\right) + \frac{\alpha(2\beta - \alpha)\lambda_\infty}{2\mu(\beta - \alpha)(\mu + \beta - \alpha)}\left(e^{-\mu\tau} - e^{-\mu(2t-\tau)}\right) - \left(\frac{\alpha(2\beta - \alpha)\lambda_\infty}{(\beta - \alpha)(\mu + \beta - \alpha)} + \frac{2\alpha\beta(\lambda_0 - \lambda_\infty)}{\mu(\beta - \alpha)}\right. \\ &- \left.\frac{\alpha^2(2\lambda_0 - \lambda_\infty)}{(\beta - \alpha)(\mu - \beta + \alpha)}\right)\frac{e^{-(\mu+\beta-\alpha)t+(\beta-\alpha)\tau} - e^{-\mu(2t-\tau)}}{\mu - \beta + \alpha} + \left(\lambda_0 - \lambda_\infty + \frac{\mu(\lambda_0 - \lambda_\infty)}{\mu - \beta + \alpha} + \frac{2\alpha\beta(\lambda_0 - \lambda_\infty)}{\mu(\beta - \alpha)}\right) \\ &\cdot h(t - \tau)e^{-\mu\tau} - \frac{\alpha^2(2\lambda_0 - \lambda_\infty)}{2(\beta - \alpha)(\mu - \beta + \alpha)^2}\left(e^{-2(\beta-\alpha)t-(\mu-2\beta+2\alpha)\tau} - e^{-\mu(2t-\tau)}\right) - \frac{\lambda_0 - \lambda_\infty}{\mu - \beta + \alpha}\left(e^{-\mu t}\right. \\ &- \left.e^{-\mu(2t-\tau)}\right) + e^{-\mu\tau}\left(\frac{\lambda_\infty}{\mu}\left(1 - e^{-\mu(t-\tau)}\right) + \frac{\lambda_0 - \lambda_\infty}{\mu - \beta + \alpha}\left(e^{-(\beta-\alpha)(t-\tau)} - e^{-\mu(t-\tau)}\right)\right)^2 + \frac{e^{-(\beta-\alpha)\tau} - e^{-\mu\tau}}{\mu - \beta + \alpha} \\ &\cdot \left(\frac{\alpha(2\beta - \alpha)\lambda_\infty}{2(\beta - \alpha)(\mu + \beta - \alpha)}\left(1 - e^{-(\mu+\beta-\alpha)(t-\tau)}\right) + \frac{\alpha\beta(\lambda_0 - \lambda_\infty)}{\mu(\beta - \alpha)}\left(e^{-(\beta-\alpha)(t-\tau)} - e^{-(\mu+\beta-\alpha)(t-\tau)}\right)\right. \\ &- \left.\frac{\alpha^2(2\lambda_0 - \lambda_\infty)}{2(\beta - \alpha)(\mu - \beta + \alpha)}\left(e^{-2(\beta-\alpha)(t-\tau)} - e^{-(\mu+\beta-\alpha)(t-\tau)}\right)\right) + (\lambda_0 - \lambda_\infty)\frac{e^{-(\beta-\alpha)\tau} - e^{-\mu\tau}}{\mu - \beta + \alpha}\left(\frac{\lambda_0 - \lambda_\infty}{\mu - \beta + \alpha}\right. \\ &\cdot \left.\left(e^{-2(\beta-\alpha)(t-\tau)} - e^{-(\mu+\beta-\alpha)(t-\tau)}\right) + \frac{\lambda_\infty}{\mu}\left(e^{-(\beta-\alpha)(t-\tau)} - e^{-(\mu+\beta-\alpha)(t-\tau)}\right)\right) - \left(\frac{\lambda_\infty}{\mu}\left(1 - e^{-\mu t}\right) + \frac{\lambda_0 - \lambda_\infty}{\mu - \beta + \alpha}\right. \\ &\cdot \left.\left(e^{-(\beta-\alpha)t} - e^{-\mu t}\right)\right)\left(\frac{\lambda_\infty}{\mu}\left(1 - e^{-\mu(t-\tau)}\right) + \frac{\lambda_0 - \lambda_\infty}{\mu - \beta + \alpha}\left(e^{-(\beta-\alpha)(t-\tau)} - e^{-\mu(t-\tau)}\right)\right)\end{aligned} \tag{2.44}$$

when $\mu \neq \beta - \alpha$ and

$$\begin{aligned}
\mathrm{Cov}\left[Q_t, Q_{t-\tau}\right] = {} & \frac{\lambda_\infty}{\mu}\left(1 - e^{-\mu\tau}\right)\left(\frac{\lambda_\infty}{\mu}\left(1 - e^{-\mu(t-\tau)}\right) + \left(\lambda_0 - \lambda_\infty\right)(t-\tau)e^{-\mu(t-\tau)}\right) + \frac{\lambda_\infty}{\mu}\left(e^{-\mu\tau} - e^{-\mu t}\right) \\
& + \frac{\alpha(2\beta-\alpha)\lambda_\infty}{4\mu^3}\left(e^{-\mu\tau} - e^{-\mu(2t-\tau)}\right) - \left(\frac{\alpha(2\beta-\alpha)\lambda_\infty}{2\mu^2} + \frac{2\alpha\beta(\lambda_0-\lambda_\infty)}{\mu^2}\right)(t-\tau)e^{-\mu(2t-\tau)} + \Big(\lambda_0 - \lambda_\infty \\
& + \frac{2\alpha\beta(\lambda_0-\lambda_\infty)}{\mu^2}\Big)\frac{e^{-(\beta-\alpha)t-(\mu-\beta+\alpha)\tau} - e^{-\mu(2t-\tau)}}{\mu} - \frac{\alpha^2(2\lambda_0-\lambda_\infty)}{2\mu}(t-\tau)^2 e^{-\mu(2t-\tau)} + (\lambda_0-\lambda_\infty) \\
& \cdot \left(\frac{(t-\tau)e^{-\mu t}}{\mu} + \frac{e^{-\mu(2t-\tau)} - e^{-\mu t}}{\mu^2}\right) + e^{-\mu\tau}\left(\frac{\lambda_\infty}{\mu}\left(1 - e^{-\mu(t-\tau)}\right) + (\lambda_0-\lambda_\infty)(t-\tau)e^{-\mu(t-\tau)}\right)^2 \\
& + \left(\frac{\alpha(2\mu+\alpha)\lambda_\infty}{4\mu^2}\left(1 - e^{-2\mu(t-\tau)}\right) + \frac{\alpha\beta(\lambda_0-\lambda_\infty)}{\mu^2}\left(e^{-\mu(t-\tau)} - e^{-2\mu(t-\tau)}\right) - \frac{\alpha^2(2\lambda_0-\lambda_\infty)}{2\mu}(t-\tau)e^{-2\mu(t-\tau)}\right) \\
& \cdot \tau e^{-\mu\tau} + \tau(\lambda_0-\lambda_\infty)e^{-\mu t}\left(\frac{\lambda_\infty}{\mu}\left(1 - e^{-\mu(t-\tau)}\right) + (\lambda_0-\lambda_\infty)(t-\tau)e^{-\mu(t-\tau)}\right) - \left(\frac{\lambda_\infty}{\mu}\left(1 - e^{-\mu t}\right)\right. \\
& \left. + (\lambda_0-\lambda_\infty)\,te^{-\mu t}\right)\left(\frac{\lambda_\infty}{\mu}\left(1 - e^{-\mu(t-\tau)}\right) + (\lambda_0-\lambda_\infty)(t-\tau)e^{-\mu(t-\tau)}\right) \qquad (2.45)
\end{aligned}$$

when $\mu = \beta - \alpha$, where $h(s) = se^{-2\mu s}$ if $2\mu = \beta - \alpha$ and $h(s) = \frac{e^{-(\beta-\alpha)s} - e^{-2\mu s}}{2\mu - \beta + \alpha}$ if $2\mu \neq \beta - \alpha$ for all $s \geq 0$.

Proof. The stated forms follow by simplification of the expressions in Corollary 2.3.11, yielding

$$\mathrm{E}\left[Q_t\right] = \frac{\lambda_\infty}{\mu}\left(1 - e^{-\mu t}\right) + \frac{\lambda_0 - \lambda_\infty}{\mu - \beta + \alpha}\left(e^{-(\beta-\alpha)t} - e^{-\mu t}\right)$$

for the mean of the $Hawkes/M/\infty$ queue,

$$\begin{aligned}
\mathrm{Cov}\left[\lambda_t, Q_t\right] = {} & \frac{\alpha(2\beta-\alpha)\lambda_\infty}{2(\beta-\alpha)(\mu+\beta-\alpha)}\left(1 - e^{-(\mu+\beta-\alpha)t}\right) + \frac{\alpha\beta(\lambda_0-\lambda_\infty)}{\mu(\beta-\alpha)}\left(e^{-(\beta-\alpha)t} - e^{-(\mu+\beta-\alpha)t}\right) \\
& - \frac{\alpha^2(2\lambda_0-\lambda_\infty)}{2(\beta-\alpha)(\mu-\beta+\alpha)}\left(e^{-2(\beta-\alpha)t} - e^{-(\mu+\beta-\alpha)t}\right)
\end{aligned}$$

for the covariance between the queue and the intensity, and

$$\begin{aligned}
\mathrm{Var}\left(Q_t\right) = {} & \frac{\lambda_\infty}{\mu}\left(1 - e^{-\mu t}\right) + \frac{\alpha(2\beta-\alpha)\lambda_\infty}{2\mu(\beta-\alpha)(\mu+\beta-\alpha)}\left(1 - e^{-2\mu t}\right) - \left(\frac{\alpha(2\beta-\alpha)\lambda_\infty}{(\beta-\alpha)(\mu+\beta-\alpha)} + \frac{2\alpha\beta(\lambda_0-\lambda_\infty)}{\mu(\beta-\alpha)}\right. \\
& \left. - \frac{\alpha^2(2\lambda_0-\lambda_\infty)}{(\beta-\alpha)(\mu-\beta+\alpha)}\right)\frac{e^{-(\mu+\beta-\alpha)t} - e^{-2\mu t}}{\mu-\beta+\alpha} + \left(\lambda_0 - \lambda_\infty + \frac{\mu(\lambda_0-\lambda_\infty)}{\mu-\beta+\alpha} + \frac{2\alpha\beta(\lambda_0-\lambda_\infty)}{\mu(\beta-\alpha)}\right) \\
& \cdot \frac{e^{-(\beta-\alpha)t} - e^{-2\mu t}}{2\mu-\beta+\alpha} - \frac{\alpha^2(2\lambda_0-\lambda_\infty)}{2(\beta-\alpha)(\mu-\beta+\alpha)^2}\left(e^{-2(\beta-\alpha)t} - e^{-2\mu t}\right) - \frac{\lambda_0-\lambda_\infty}{\mu-\beta+\alpha}\left(e^{-\mu t} - e^{-2\mu t}\right)
\end{aligned}$$

for the variance of the queue, all in the case where $\mu \neq \beta - \alpha$. The remaining derivation follows directly from substitution of these functions and the corresponding expressions for remaining cases and epochs into Equations 2.42 and 2.43. □

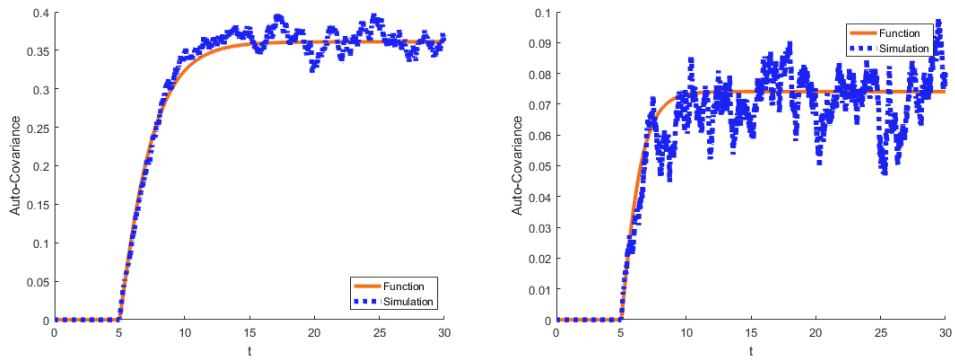


Figure 2.4: Auto-covariance of the $Hawkes/M/\infty$ Queue for $\tau = 5$, where $\alpha = \frac{3}{4}$, $\beta = \frac{5}{4}$, $\lambda^* = \mu = 1$ (left) and $\alpha = 1, \beta = 2$, $\lambda^* = \mu = 1$ (right).

In Figure 2.4 the expressions in Proposition 2.3.14 are compared to simulations, based on 100,000 replications.

2.3.5 Generating Functions for the $Hawkes/PH/\infty$ Queue

To complement these findings, we also derive a form for the moment generating function for a general queueing system driven by a Hawkes process.

Theorem 2.3.15. *Consider a queueing system with arrivals occurring in accordance to a Hawkes process (λ_t, N_t) with dynamics given in Equation 1.1 with $\alpha < \beta$ and phase-type distributed service. Let $\delta \in \mathbb{R}_+^{n+1}$ and let $M(\delta, t) = M(\delta_0, \ldots, \delta_n, t) = \mathrm{E}\left[e^{\delta_0 \lambda_t + \sum_{i=1}^n \delta_i Q_{t,i}}\right]$. Then, the moment generating function for the queueing system $M(\delta, t)$*

is given by the solution to the following partial differential equation,

$$\frac{\partial M(\delta,t)}{\partial t} = \delta_0\beta\lambda^* M(\delta,t) + \left(\sum_{i=1}^{n}\theta_i(e^{\delta_0\alpha+\delta_i}-1) - \delta_0\beta\right)\frac{\partial M(\delta,t)}{\partial \delta_0} + \sum_{i=1}^{n}\left(\mu_{i0}(e^{-\delta_i}-1) + \sum_{k\neq i}\mu_{ik}(e^{\delta_k-\delta_i}-1)\right)\frac{\partial M(\delta,t)}{\partial \delta_i}. \tag{2.46}$$

Proof. This proof makes use of techniques similar to the prior theorems, and so we omit the preceding infinitesimal generator steps. Note that $\frac{\partial M(\delta,t)}{\partial t} = \frac{\partial}{\partial t}\mathrm{E}\left[e^{\delta_0\lambda_t+\sum_{i=1}^{n}\delta_i Q_{t,i}}\right]$. From this, we start with the following.

$$\begin{aligned}\frac{\partial M(\delta,t)}{\partial t} = \mathrm{E}\Bigg[&\delta_0\beta(\lambda^*-\lambda_t)e^{\delta_0\lambda_t+\sum_{i=1}^{n}\delta_i Q_{t,i}} + \sum_{j=1}^{n}\lambda_t\theta_j\left(e^{\delta_0(\lambda_t+\alpha)+\sum_{k\neq j}\delta_k Q_{t,k}+\delta_j(Q_{t,j}+1)} - e^{\delta\lambda_t+\sum_{i=1}^{n}\delta_i Q_{t,i}}\right)\\
&+\sum_{k=1}^{n}\sum_{j\neq k}\mu_{jk}Q_{t,j}\left(e^{\delta_0\lambda_t+\sum_{l\neq j\wedge l\neq k}\delta_l Q_{t,l}+\delta_j(Q_{t,j}-1)+\delta_k(Q_{t,k}+1)} - e^{\delta_0\lambda_t+\sum_{i=1}^{n}\delta_i Q_{t,i}}\right)\\
&+\sum_{j=1}^{n}\mu_{j0}Q_{t,j}\left(e^{\delta_0\lambda_t+\sum_{k\neq j}\delta_k Q_{t,k}+\delta_j(Q_{t,k}-1)} - e^{\delta_0\lambda_t+\sum_{i=1}^{n}\delta_i Q_{t,i}}\right)\Bigg]\end{aligned}$$

Now, we distribute terms and notice that the difference of exponentials here can be expressed as the following products.

$$\begin{aligned}\frac{\partial M(\delta,t)}{\partial t} = \mathrm{E}\Bigg[&\delta_0\beta\lambda^* e^{\delta_0\lambda_t+\sum_{i=1}^{n}\delta_i Q_{t,i}} - \delta_0\beta\lambda_t e^{\delta_0\lambda_t+\sum_{i=1}^{n}\delta_i Q_{t,i}} + \sum_{j=1}^{n}\lambda_t\theta_j e^{\delta_0\lambda_t+\sum_{i=1}^{n}\delta_i Q_{t,i}}\left(e^{\delta_0\alpha+\delta_j}-1\right)\\
&+\sum_{k=1}^{n}\sum_{j\neq k}\mu_{jk}Q_{t,j}e^{\delta_0\lambda_t+\sum_{i=1}^{n}\delta_i Q_{t,i}}\left(e^{\delta_k-\delta_j}-1\right) + \sum_{j=1}^{n}\mu_{j0}Q_{t,j}e^{\delta_0\lambda_t+\sum_{i=1}^{n}\delta_i Q_{t,i}}\left(e^{-\delta_j}-1\right)\Bigg]\end{aligned}$$

Here, we can now use linearity of expectation and group like terms.

$$\begin{aligned}\frac{\partial M(\delta,t)}{\partial t} = &\delta_0\beta\lambda^*\mathrm{E}\left[e^{\delta_0\lambda_t+\sum_{i=1}^{n}\delta_i Q_{t,i}}\right] + \left(\sum_{j=1}^{n}\theta_j(e^{\delta_0\alpha+\delta_j}-1) - \delta_0\beta\right)\mathrm{E}\left[\lambda_t e^{\delta_0\lambda_t+\sum_{i=1}^{n}\delta_i Q_{t,i}}\right]\\
&+\sum_{j=1}^{n}\left(\mu_{j0}(e^{-\delta_j}-1) + \sum_{k\neq j}\mu_{jk}(e^{\delta_k-\delta_j}-1)\right)\mathrm{E}\left[Q_{t,j}e^{\delta_0\lambda_t+\sum_{i=1}^{n}\delta_i Q_{t,i}}\right]\end{aligned}$$

Finally, here we recognize the form of partial derivatives of $M(\delta,t)$ in each expectation, and so we simplify to the desired result. □

We can use this to also find a partial differential equation for the natural logarithm of the moment generating function. This is called the cumulant moment generating function, as the derivative of this function yields the cumulant moments.

Corollary 2.3.16. *Consider a queueing system with arrivals occurring in accordance to a Hawkes process (λ_t, N_t) with dynamics given in Equation 1.1 and phase-type distributed service. Let $\delta \in \mathbb{R}_+^{n+1}$ and let $G(\delta, t) = G(\delta_0, \ldots, \delta_n, t) = \log\left(\mathrm{E}\left[e^{\delta_0 \lambda_t + \sum_{i=1}^n \delta_i Q_{t,i}}\right]\right)$. Then, the cumulant moment generating function for the queueing system $G(\delta, t)$ is given by the solution to the following partial differential equation,*

$$\begin{aligned}\frac{\partial G(\delta,t)}{\partial t} &= \delta_0 \beta \lambda^* + \left(\sum_{i=1}^n \theta_i (e^{\delta_0 \alpha + \delta_i} - 1) - \delta_0 \beta\right) \frac{\partial G(\delta,t)}{\partial \delta_0} \\ &\quad + \sum_{i=1}^n \left(\mu_{i0}(e^{-\delta_i} - 1) + \sum_{k \neq i} \mu_{ik}(e^{\delta_k - \delta_i} - 1)\right) \frac{\partial G(\delta,t)}{\partial \delta_i}.\end{aligned} \tag{2.47}$$

Proof. To begin, we see from the derivative of the logarithm and the chain rule that

$$\frac{\partial G(\delta,t)}{\partial t} = \frac{\partial}{\partial t} \log\left(\mathrm{E}\left[e^{\delta_0 \lambda_t + \sum_{i=1}^n \delta_i Q_{t,i}}\right]\right) = \frac{\frac{\partial}{\partial t} \mathrm{E}\left[e^{\delta_0 \lambda_t + \sum_{i=1}^n \delta_i Q_{t,i}}\right]}{\mathrm{E}\left[e^{\delta_0 \lambda_t + \sum_{i=1}^n \delta_i Q_{t,i}}\right]}$$

and here we can recognize that these expectations are the moment generating function. Using Theorem 2.3.15, we have

$$\begin{aligned}\frac{\partial G(\delta,t)}{\partial t} &= \delta_0 \beta \lambda^* + \left(\sum_{i=1}^n \theta_i (e^{\delta_0 \alpha + \delta_i} - 1) - \delta_0 \beta\right) \frac{\frac{\partial}{\partial \delta_0} \mathrm{E}\left[e^{\delta_0 \lambda_t + \sum_{i=1}^n \delta_i Q_{t,i}}\right]}{\mathrm{E}\left[e^{\delta_0 \lambda_t + \sum_{i=1}^n \delta_i Q_{t,i}}\right]} \\ &\quad + \sum_{i=1}^n \left(\mu_{i0}(e^{-\delta_i} - 1) + \sum_{k \neq i} \mu_{ik}(e^{\delta_k - \delta_i} - 1)\right) \frac{\frac{\partial}{\partial \delta_i} \mathrm{E}\left[e^{\delta_0 \lambda_t + \sum_{i=1}^n \delta_i Q_{t,i}}\right]}{\mathrm{E}\left[e^{\delta_0 \lambda_t + \sum_{i=1}^n \delta_i Q_{t,i}}\right]}.\end{aligned}$$

Now we recognize that $\frac{\frac{\partial}{\partial \delta_i} \mathrm{E}\left[e^{\delta_0 \lambda_t + \sum_{i=1}^n \delta_i Q_{t,i}}\right]}{\mathrm{E}\left[e^{\delta_0 \lambda_t + \sum_{i=1}^n \delta_i Q_{t,i}}\right]} = \frac{\partial G(\delta,t)}{\partial \delta_i}$, and so we have the stated result.

□

Comparing these two partial differential equations, we see that the expression for the cumulant moment generating function only depends on the partial derivatives, not on the function itself. In some cases the cumulant moment generating function is better since it directly will compute the variance, skewness, and higher order cumulants directly without having to know the relationships between cumulants and moments. Moreover, the cumulant moments have shift and scale invariance properties, which are often desired. The PDE in Corollary 2.3.16 produces a form that provides insight to the solution through use of the method of characteristics, which we now show in the following theorem.

Theorem 2.3.17. *Consider a queueing system with arrivals occurring in accordance to a Hawkes process (λ_t, N_t) with dynamics given in Equation 1.1 and phase-type distributed service with transient state sub-generator matrix $S \in \mathbb{R}^{n\times n}$. Let $\delta \in \mathbb{R}_+^{n+1}$ and let $G(\delta, t) = G(\delta_0, \ldots, \delta_n, t) = \log\left(\mathrm{E}\left[e^{\delta_0\lambda_t + \sum_{i=1}^n \delta_i Q_{t,i}}\right]\right)$. Then, the cumulant moment generating function for the queueing system $G(\delta, t)$ is given by*

$$G(\delta, t) = \beta\lambda^* \int_0^t h(z)\mathrm{d}z + h(0)\lambda_0 \tag{2.48}$$

where $h(z)$ is the solution to the ordinary differential equation

$$\dot{h}(z) = 1 - e^{\alpha h(z)}\theta^{\mathrm{T}}\left(\mathbf{v} + e^{-S(z-t)}\left(e^{\mathrm{diag}(\delta)} - I\right)\mathbf{v}\right) + \beta h(z)$$

with initial value $h(t) = \delta_0$.

Proof. We proceed by the method of characteristics for the PDE given in Corollary 2.3.16. To do so, let z be a parametrization variable and let $\Delta_0, \Delta_1, \ldots, \Delta_n$ be characteristics variables. From recognizing the linearity of the PDE, we see that we can implement the method of characteristics by setting $\dot{\Delta}_i(z) := \frac{\mathrm{d}\Delta_i(z)}{\mathrm{d}z}$ equal to the function serving as coefficient of $\frac{\partial G(\delta,t)}{\partial \delta_i}$ in the PDE for each $i \in \{0, \ldots, n\}$,

each with initial condition that $\Delta_i(t) = \delta_i$. This yields the following system of characteristic ODE's:

$$\dot{\Delta}_0(z) = 1 - e^{\Delta_0 \alpha} \sum_{j \neq i} \theta_j e^{\Delta_j} + \Delta_0 \beta,$$

$$\dot{\Delta}_i(z) = \mu_i - \mu_{i0} e^{-\Delta_i} - \sum_{j \neq i} \mu_{ij} e^{\Delta_j - \Delta_i} \qquad \forall i \in \{1, \ldots, n\}.$$

We now let $x \in \mathbb{R}^n$ be such that $x_i = e^{\Delta_i}$. Note that this substitution can also be expressed $x = e^{\mathrm{diag}(\Delta)}\mathbf{v}$, as this will be of use in solving the system. Then, we have that $\dot{x}_i(z) = x_i(z)\dot{\Delta}_i(z)$. In this form, the last n characteristic ODE's can be expressed as

$$\dot{x}(z) = -S\,x(z) + S\,\mathbf{v}$$

which means that

$$x(z) = \mathbf{v} + e^{-S(z-t)}\left(e^{\mathrm{diag}(\delta)} - I\right)\mathbf{v}$$

where we have used the initial condition $x(t) = e^{\mathrm{diag}(\Delta(t))} = e^{\mathrm{diag}(\delta)}$. We now note that to follow the method of characteristics fully and receive a closed form solution to the PDE we would want to solve the remaining characteristic ODE

$$\dot{\Delta}_0(z) = 1 - e^{\Delta_0 \alpha} \sum_{j \neq i} \theta_j e^{\Delta_j} + \Delta_0 \beta = 1 - e^{\Delta_0 \alpha} \theta^{\mathrm{T}} x + \Delta_i \beta$$

which has initial condition that $\Delta_0(t) = \delta_0$. Because this form of ODE is not known to have a closed form solution in terms of standard math functions, we let $h(z)$ be defined as the solution to this initial value problem. Then, we now complete the method of characteristics by solving

$$\dot{g}(z) = \beta\lambda^* \Delta_0(z) = \beta\lambda^* h(z)$$

with the initial condition that $g(0) = G(\Delta(0), 0) = \Delta_0(0)\lambda_0 = h(0)\lambda_0$. Since this ODE is already separated, we have

$$g(z) - h(0)\lambda_0 = g(z) - g(0) = \int_0^z \dot{g}(\xi)\mathrm{d}\xi = \beta\lambda^* \int_0^z h(\xi)\mathrm{d}\xi.$$

Thus, we now have

$$G(\delta, t) = g(t) = \beta\lambda^* \int_0^t h(\xi)\mathrm{d}\xi + h(0)\lambda_0$$

and this is the stated result. □

While the ODE in this statement may not be able to be solved for a closed form expression outside of special cases, this reduction of the PDE to an ODE simplifies numerical implementations. We now note that this of course extends to the moment generating function as well by simply taking the exponential of the cumulant generating function.

2.3.6 Simulation Study

To conclude Section 2.3 we provide a collection of simulation examples that verify the accuracy of our expressions for the moments in a variety of settings. In each example we derive the simulated functions via 100,000 replications of the procedure described in Ogata (1981). We start with the mean and variance of a single phase system, as shown in the pair of plots below in Figure 2.5.

As a second example, we also consider a three-phase Erlang distributed service. We use two different parameter settings, one in which the mean service duration is 1 and another in which the mean service length is 6. In the first case, $\alpha = \frac{1}{2}$, $\beta = \frac{3}{4}$, and $\lambda^* = 1$. In the latter, $\alpha = \frac{3}{4}$, $\beta = \frac{5}{4}$, and $\lambda^* = 1$. The mean is shown in Figure 2.6, the variance in Figure 2.7, the covariance of the queue and the intensity in Figure 2.8, and the covariance of the phases of the queue in Figure 2.9.

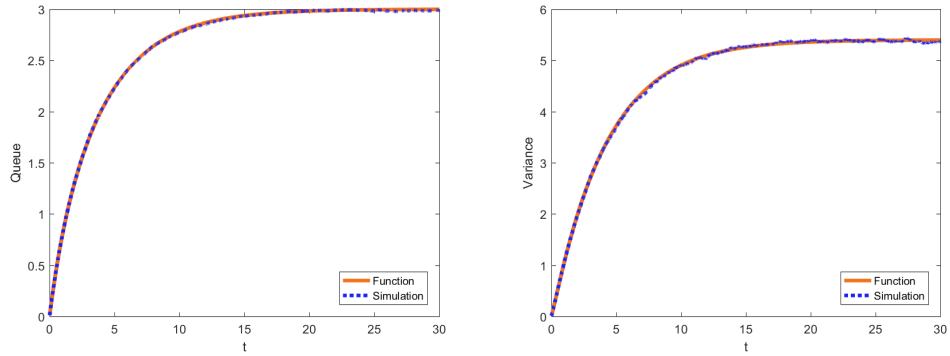


Figure 2.5: Mean (left) and Variance (right) of Q_t in $Hawkes/M/\infty$, $\alpha = \frac{1}{2}$, $\beta = \frac{3}{4}$, $\lambda^* = \mu = 1$.

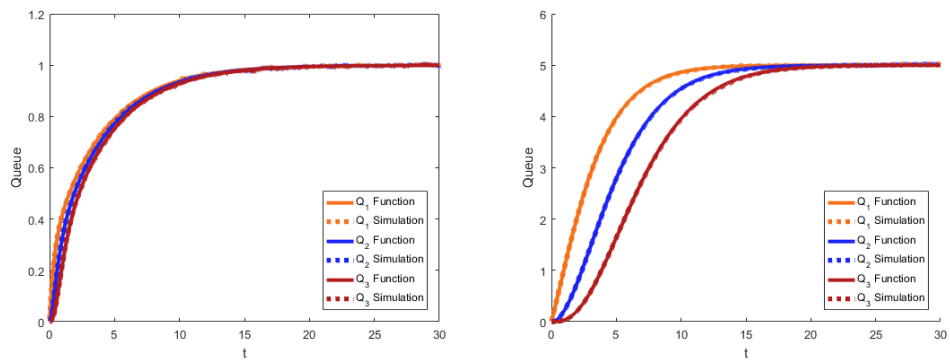


Figure 2.6: Mean of the $Hawkes/E_3/\infty$ Queue, where $\alpha = \frac{1}{2}$, $\beta = \frac{3}{4}$, $\lambda^* = 1$, $\frac{1}{\mu} = 1$ (left) and $\alpha = \frac{3}{4}, \beta = \frac{5}{4}$, $\lambda^* = 1$, $\frac{1}{\mu} = 6$ (right).

In addition to the Erlang setting, we also verify the performance of the hyper-exponential service equations. We again consider a three phase distributed service and display a pair of scenarios. In both parameter groups $\theta = [.15, .4, .45]^{\mathrm{T}}$ and $\mu = [1, 4, 6]^{\mathrm{T}}$. In the first setting we consider $\alpha = \frac{1}{2}, \beta = 1$, and $\lambda^* = 2$, whereas in the second setting $\alpha = 1, \beta = 2$, and $\lambda^* = 2$. These are displayed in the same order as the Erlang examples are: mean in Figure 2.10, variance in Figure 2.11, covariance with the intensity in Figure 2.12, and covariance of the queues in Figure 2.13.

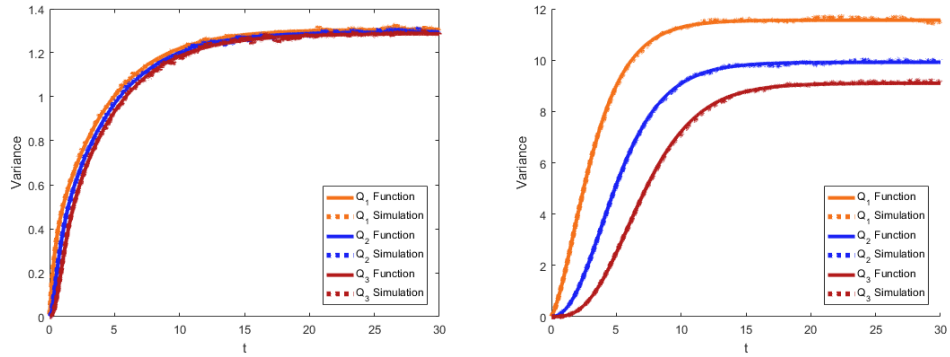


Figure 2.7: Variance of the $Hawkes/E_3/\infty$ Queue, where $\alpha = \frac{1}{2}, \beta = \frac{3}{4}, \lambda^* = 1, \frac{1}{\mu} = 1$ (left) and $\alpha = \frac{3}{4}, \beta = \frac{5}{4}, \lambda^* = 1, \frac{1}{\mu} = 6$ (right).

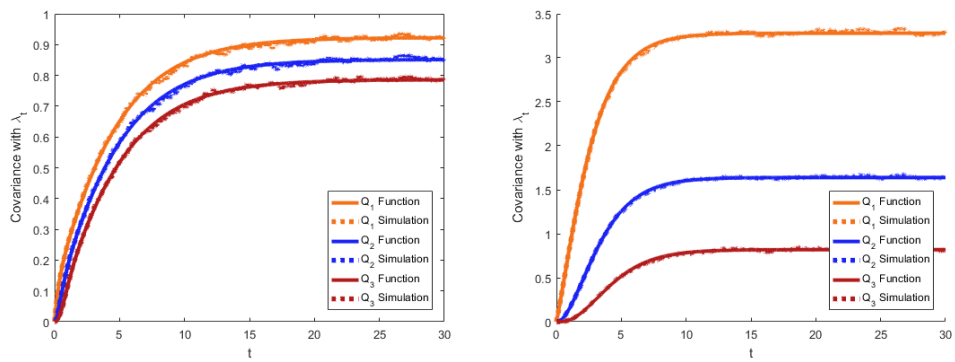


Figure 2.8: Covariance of $Hawkes/E_3/\infty$ Queue, where $\alpha = \frac{1}{2}, \beta = \frac{3}{4}, \lambda^* = 1, \frac{1}{\mu} = 1$ (left) and $\alpha = \frac{3}{4}, \beta = \frac{5}{4}, \lambda^* = 1, \frac{1}{\mu} = 6$ (right).

In conducting these simulation experiments we have made an interesting observation. Consider the following example: let $\lambda^* = 1$, $\alpha = 1$, and $\beta = 2$. Then, let $D = 1$ be the fixed service length in a $Hawkes/D/\infty$ system and let $\mu = 1$ be the parameter of the exponential distribution in a $Hawkes/M/\infty$ system. We plot the simulated variances of these two systems in Figure 2.14 based on 10,000 replications, in which we find that the variance is larger in the deterministic service setting.

While this relationship may seem unexpected, there is an intuitive expla-

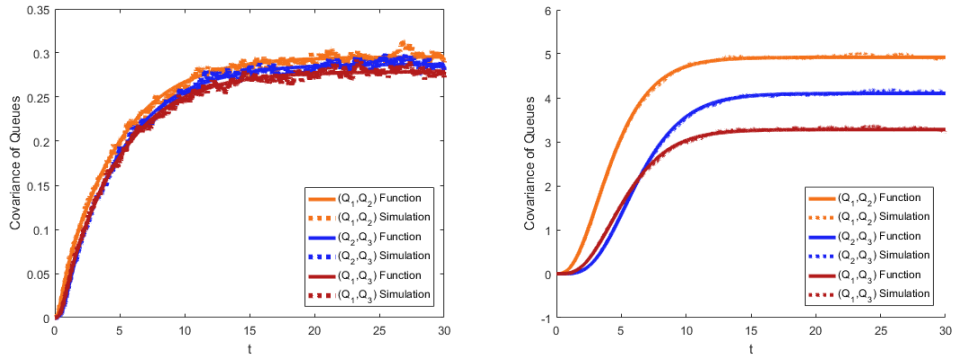


Figure 2.9: Covariance between Phases in the $Hawkes/E_3/\infty$ Queue, where $\alpha = \frac{1}{2}, \beta = \frac{3}{4}, \lambda^* = 1, \frac{1}{\mu} = 1$ (left) and $\alpha = \frac{3}{4}, \beta = \frac{5}{4}, \lambda^* = 1, \frac{1}{\mu} = 6$ (right).

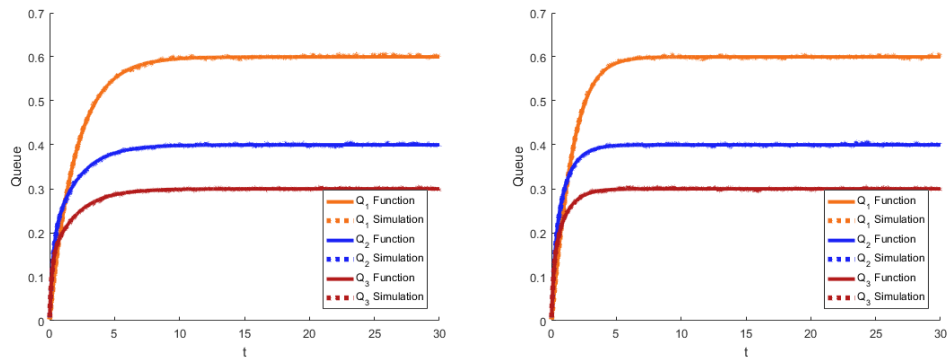


Figure 2.10: Mean of the $Hawkes/H_3/\infty$ Queue, where $\alpha = \frac{1}{2}$, $\beta = 1$, $\lambda^* = 2$, $\theta = [.15, .4, .45]^{\mathrm{T}}$, $\mu = [1, 4, 6]^{\mathrm{T}}$ (left) and $\alpha = 1$, $\beta = 2$, $\lambda^* = 2$, $\theta = [.15, .4, .45]^{\mathrm{T}}$, $\mu = [1, 4, 6]^{\mathrm{T}}$ (right).

nation for it. Because the Hawkes process exhibits clustering behavior in the arrival times, a service system with fixed service length will also experience clusters of departures times. By comparison, a system with random service durations has the opportunity to counteract the clustering behavior and disperse the departure times. In Proposition 2.3.18 we show that the steady-state variance in the deterministic service setting is greater than that of the exponential service setting.

Proposition 2.3.18. *For equal Hawkes process parameters λ^*, α, and β and equivalent*

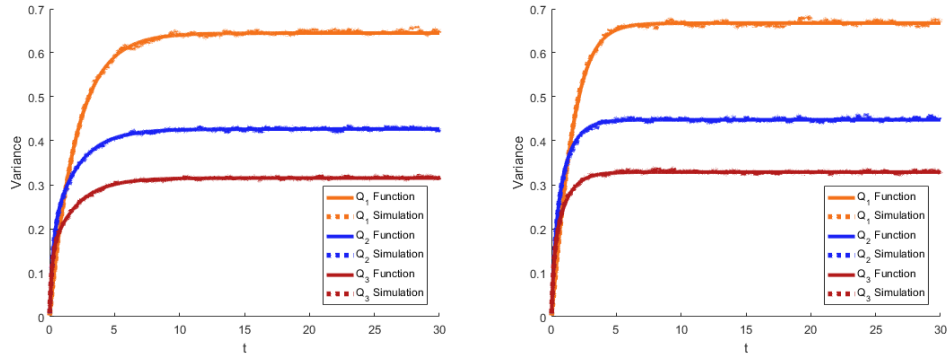


Figure 2.11: Variance of the $Hawkes/H_3/\infty$ Queue, where $\alpha = \frac{1}{2}$, $\beta = 1$, $\lambda^* = 2$, $\theta = [.15, .4, .45]^{\mathrm{T}}$, $\mu = [1, 4, 6]^{\mathrm{T}}$ (left) and $\alpha = 1$, $\beta = 2$, $\lambda^* = 2$, $\theta = [.15, .4, .45]^{\mathrm{T}}$, $\mu = [1, 4, 6]^{\mathrm{T}}$ (right).

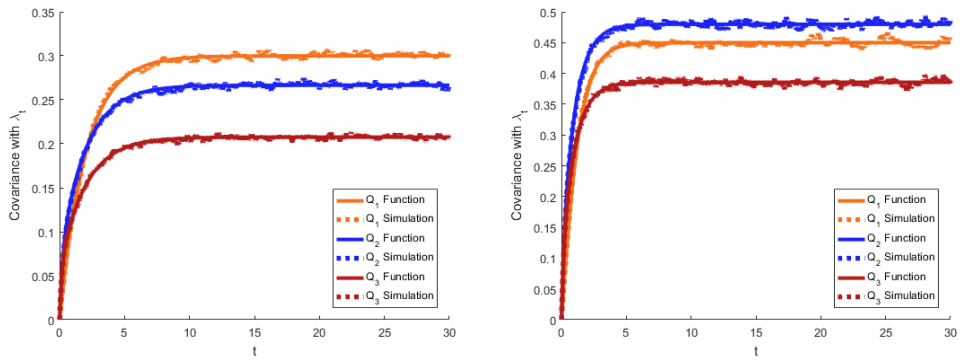


Figure 2.12: Covariance of λ_t and the $Hawkes/H_3/\infty$ Queue, where $\alpha = \frac{1}{2}$, $\beta = 1$, $\lambda^* = 2$, $\theta = [.15, .4, .45]^{\mathrm{T}}$, $\mu = [1, 4, 6]^{\mathrm{T}}$ (left) and $\alpha = 1$, $\beta = 2$, $\lambda^* = 2$, $\theta = [.15, .4, .45]^{\mathrm{T}}, \mu = [1, 4, 6]^{\mathrm{T}}$ (right).

service parameters $D = \frac{1}{\mu} > 0$, the steady-state variance of the $Hawkes/D/\infty$ queue is greater than the steady-state variance of the $Hawkes/M/\infty$ queue.

Proof. Let $\beta > \alpha > 0$ and let $\lambda^* > 0$. Further, let $D = \frac{1}{\mu} > 0$. By Theorem 2.2.2, the steady-state variance of the $Hawkes/D/\infty$ queue is

$$\mathcal{V}_{\mathrm{D}} \equiv \lambda_\infty D\left(1 + \frac{2\alpha\beta - \alpha^2}{(\beta - \alpha)^2}\right) - \lambda_\infty(1 - e^{-(\beta-\alpha)D})\frac{2\alpha\beta - \alpha^2}{(\beta - \alpha)^3}.$$

Likewise, Corollary 2.3.12 gives the steady-state variance in the exponential ser-

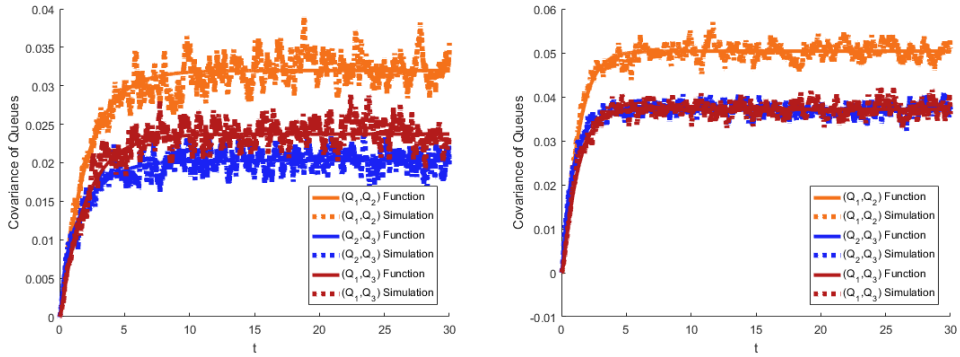


Figure 2.13: Covariance between Phases in the $Hawkes/H_3/\infty$ Queue, where $\alpha = \frac{1}{2}$, $\beta = 1$, $\lambda^* = 2$, $\theta = [.15, .4, .45]^{\mathrm{T}}$, $\mu = [1, 4, 6]^{\mathrm{T}}$ (left) and $\alpha = 1$, $\beta = 2$, $\lambda^* = 2$, $\theta = [.15, .4, .45]^{\mathrm{T}}$, $\mu = [1, 4, 6]^{\mathrm{T}}$ (right).

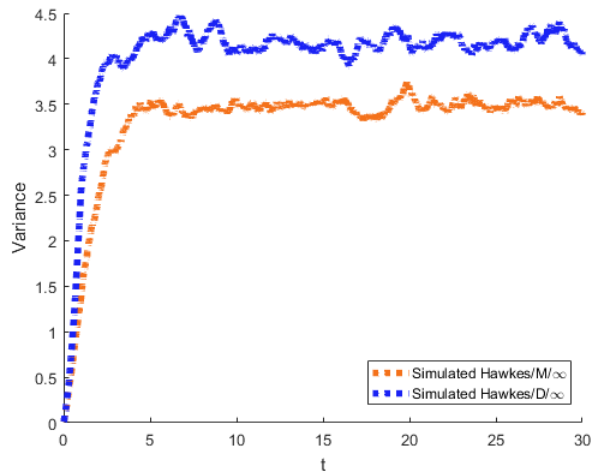


Figure 2.14: Comparison of Variances in $Hawkes/M/\infty$ and $Hawkes/D/\infty$ Queues when $\frac{1}{\mu} = D = 1$, with $\lambda^* = 1$, $\alpha = 1$, and $\beta = 2$.

vice case as

$$\mathcal{V}_{\mathrm{M}} \equiv \frac{\lambda_\infty}{\mu}\left(1 + \frac{2\alpha\beta - \alpha^2}{2(\beta - \alpha)(\mu + \beta - \alpha)}\right),$$

as noted in Remark 2.3.3. Then, the difference between these terms is

$$\mathcal{V}_{\mathrm{D}} - \mathcal{V}_{\mathrm{M}} = \frac{\lambda_\infty}{\mu}\left(\frac{2\alpha\beta - \alpha^2}{(\beta - \alpha)^2}\right)\left(1 - \frac{\beta - \alpha}{2(\mu + \beta - \alpha)} - \frac{\mu - \mu e^{-(\beta-\alpha)\frac{1}{\mu}}}{\beta - \alpha}\right),$$

where we have substituted $\frac{1}{\mu}$ for D. Because of the assumed relationships among the parameters, $\mathcal{V}_{\mathrm{D}} - \mathcal{V}_{\mathrm{M}}$ is positive if and only if the expression inside the

lattermost parenthesis is. Multiplying this expression by $\frac{2}{\mu^2}(\mu+\beta-\alpha)(\beta-\alpha) > 0$ and simplifying yields

$$\Upsilon\left(\frac{\beta-\alpha}{\mu}\right) \equiv \left(\frac{\beta-\alpha}{\mu}\right)^2 - 2\left(1 - e^{-\frac{\beta-\alpha}{\mu}}\right) + 2\left(\frac{\beta-\alpha}{\mu}\right)e^{-\frac{\beta-\alpha}{\mu}}.$$

We can re-parameterize this expression as $\Upsilon(x)$ for $x \equiv \frac{\beta-\alpha}{\mu}$. By checking the first derivative of $\Upsilon(x)$, we see that it is strictly increasing for $x \geq 0$. Since $\Upsilon(0) = 0$ and $\frac{\beta-\alpha}{\mu} > 0$ for any valid α, β, and μ, we have that $\mathcal{V}_{\mathrm{D}} - \mathcal{V}_{\mathrm{M}} > 0$. □

In Figure 2.15 we observe that this behavior can also occur in non-Markovian service settings, shown here for lognormal distributions based on 10,000 simulation replications. In this experiment each lognormal distribution has a mean of 1 and the variances increase from 0 to 5 with a step size of 0.5. Note that all the mean queue lengths appear to be converging to 1 in steady-state. Further, we see that the means of systems with higher variance in the lognormal service distribution are converging more slowly than those of lower lognormal variance. However, the opposite relationship appears to hold in terms of the variances of the queues: the higher the variance of the lognormal, the lower the variance of the queue.

2.4 Applications

To motivate this study and demonstrate its findings, we now briefly discuss two applications of this work, one concerned with viral internet traffic and one covering night clubs. Each is inspired by the self-excitement behavior of the Hawkes process, and in these settings we consider the impact and influence

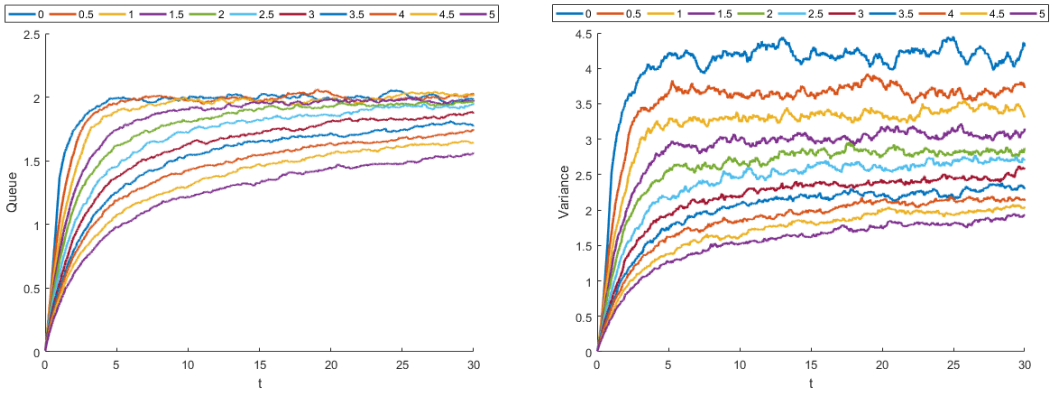


Figure 2.15: Mean (left) and Variance (right) of the $Hawkes/Lognormal/\infty$ with $\lambda^* = 1$, $\alpha = 1$, and $\beta = 2$ where Mean Service Durations is 1 and Service Variance Increases from 0 to 5.

one arrival can have on a system and how managers of such systems might try to harness that influence for some kind of benefit.

2.4.1 Trending Web Traffic

In May 2017 website rankings for the United States, Youtube, Facebook, and Reddit each ranked among the top 5 most visited websites, with Twitter in the top 10 and LinkedIn and Instagram both in the top 15, per Alexa Alexa the Web Information Company (2017). For Facebook, Reddit, and Twitter in particular, users' interactions with the sites frequently involve viewing links to external media like videos, articles, and shopping sales. A user's exposure to a webpage and her likelihood to share it herself is directly influenced by whether she sees the link from other users. As users choose to visit and potentially re-share links posted by other users, the link may start trending or become "viral." This means that it is receiving high levels of traffic and arrivals to the site, and this may lead to even more arrivals while the users continue to share it on various social

platforms. For a business or organization, going viral can lead to significant jumps in exposure, interest, and revenue.

As a basic example, we analyzed publicly available Twitter data McKelvey and Menczer (2013). This data set covers all tweets featuring both a URL and a hashtag from November 2012 and includes the tweet timestamp, the hashtags used, and the URL's linked, as well as an anonymous user ID. Perhaps the most notable event captured among the reactions in this data set is the 2012 U.S. Presidential election, which was held on November 6. Among the bountiful election-related tweets are 106 posts of the music video for Young Jeezy's 2008 song *My President* from the start of November 5 to midday on November 7. A plot of the timestamps of these tweets along with the total number of tweets occurring by that time is below. Note the flurry of posts once the election results were announced; 60 of the data's 106 postings of the video occur within an hour's time. A quick numerical investigation suggests that this type of extreme viral reaction may be more likely in certain parameter settings. In 100,000 simulation replications of a system with $\lambda^* = 0.5$, $\alpha = 19.5$, and $\beta = 20$, 82.4% of the trials had a majority of arrivals occur within one time quartile. By comparison, in the same number of replications for a system with $\lambda^* = 1$, $\alpha = 0.5$, and $\beta = 1$, this only occurred for 18.0% of the experiments. However, even outside of the main spike in this data, users seem to be posting the video in clustered time segments, approximately at the 6, 20, 45, 48, and 52 hour marks. These clusters suggest that these arrivals could be appropriately modeled by a Hawkes process, particularly when compared to a Poisson process.

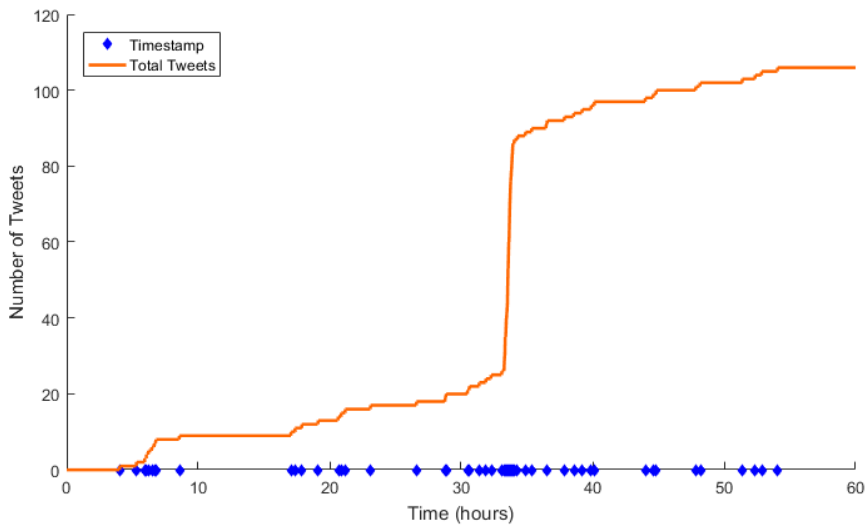


Figure 2.16: Tweets of Young Jeezy - *My President* music video from November 5 - 7, 2012.

Using what we have observed from this data as inspiration, we now model users arriving to a webpage as a Hawkes process. Because of the viral behavior we have seen in this type of arrivals, we will investigate the impact of a click. Consider a Hawkes Process N_t with baseline intensity λ^*, initial intensity λ_0, jump size α, and decay parameter β. Now, let $\hat{N}_t$ represent an independent Hawkes process that is identical to N_t in terms of parameters with the exception that it experienced an arrival at time 0, whereas N_t starts empty. This means that the baseline intensity, jump size, and decay parameter are the same for $\hat{N}_t$ as they were for N_t, but the initial intensity is $\lambda_0 + \alpha$ and $\hat{N}_0 = 1$. Then, by Proposition 1.1.1,

$$\begin{aligned}
\mathrm{E}\left[\hat{N}_t\right] - \mathrm{E}\left[N_t\right] &= \lambda_\infty t + \frac{\lambda_0 + \alpha - \lambda_\infty}{\beta - \alpha}\left(1 - e^{-(\beta-\alpha)t}\right) + 1 - \lambda_\infty t - \frac{\lambda_0 - \lambda_\infty}{\beta - \alpha}\left(1 - e^{-(\beta-\alpha)t}\right) \\
&= \frac{\beta}{\beta - \alpha} - \frac{\alpha}{\beta - \alpha} e^{-(\beta-\alpha)t} \longrightarrow \frac{\beta}{\beta - \alpha} \text{ as } t \to \infty
\end{aligned}$$

which shows that the gap between the two expectations is positive and grows throughout time. However, this is simply tracking the number of visitors; it

does not account for the time the users spend on the site. To capture this, we can extend this arrival model to a queueing model in which the service represents the time the user spends on the webpage. Provided the website is well hosted, this can be modeled as an infinite server queue as any user can visit the webpage that chooses to do so. If the time each user spends on the page is independently and exponentially distributed with rate μ, we see that the expected number of users on the page at time t is $\mathrm{E}[Q_t]$. Then, from time 0 to time T the expected total time spent on the page across all users $\sigma(T)$ is

$$\begin{aligned}\sigma(T) = \int_0^T \mathrm{E}[Q_t]\,\mathrm{d}t &= \int_0^T \left(\frac{\lambda_\infty}{\mu}\left(1-e^{-\mu t}\right) + \frac{\lambda_0-\lambda_\infty}{\mu-\beta+\alpha}\left(e^{-(\beta-\alpha)t}-e^{-\mu t}\right)\right)\mathrm{d}t \\ &= \frac{\lambda_\infty}{\mu}\left(T-\frac{1-e^{-\mu T}}{\mu}\right) + \frac{\lambda_0-\lambda_\infty}{\mu-\beta+\alpha}\left(\frac{1-e^{-(\beta-\alpha)T}}{\beta-\alpha}-\frac{1-e^{-\mu T}}{\mu}\right)\end{aligned}$$

where we have applied the results of Corollary 2.3.11 for hyper-exponential service with $n = 1$ and $\mu \neq \beta - \alpha$, thus yielding exponential service. Now, suppose that a website earns m dollars per unit of time in advertising revenue for each user on the site. Then, the expected earnings by time T is $A(T) = m\sigma(T)$. We can now repeat the value of a click experiment when also considering service. Let Q_t be a queueing system with exponential service at rate μ, infinite servers, and Hawkes process arrivals with parameters λ^*, α, and β and assume the queue starts empty. Then, let $\hat{Q}_t$ be the analogous adaptation of Q_t that $\hat{N}_t$ is to N_t. Let $A(T)$ and $\hat{A}(T)$ be the corresponding expected dwell time revenues, each with earning rate m. Note that the expected time the initial customer has spent in the system by time T is $\min\{S, T\}$ where S is the duration of her service. Hence the revenue associated with her visit to the page by time T is $m\frac{1-e^{-\mu T}}{\mu}$. Then,

$$\begin{aligned}\hat{A}(T) - A(T) &= m\frac{1-e^{-\mu T}}{\mu} + m\frac{\alpha}{\mu-\beta+\alpha}\left(\frac{1-e^{-(\beta-\alpha)T}}{\beta-\alpha}-\frac{1-e^{-\mu T}}{\mu}\right) \\ &= \frac{m}{\mu}\left(1+\frac{1}{\beta-\alpha}\right) - m\frac{\alpha e^{-(\beta-\alpha)T}}{(\beta-\alpha)(\mu-\beta+\alpha)} - m\frac{(\mu-\beta)e^{-\mu T}}{\mu(\mu-\beta+\alpha)},\end{aligned}$$

which can be shown to also always grow with T via its first derivative. We can also further observe that as $\alpha \to \beta$ each of these gaps grows towards infinity, and thus so grows the impact of a click in viral settings.

Note that this model can also be used for internet-inspired applications other than users arriving to internet pages. For example, as mobile carriers continue to add cloud storage based services and allow customers to upload pictures from their smart phones as soon as they are taken, the $Hawkes/M/\infty$ queue can be used to describe the number of pictures being uploaded at once. For further reading on the Hawkes process and its use in internet traffic applications see Rizoiu et al. (2017), in which the authors develop a novel Hawkes-process-based model for the popularity of online content in great detail.

2.4.2 Club Queue

From our Hawkes driven infinite server queue with phase-type service distributions, we can construct what we refer to as the *Club Queue*. This stems from an application perhaps uncommon to queueing systems, a nightclub. This setting features a key characteristic: the best club has the most people waiting for it. Because of this, the Hawkes process naturally represents the excitation exhibited by club-goers joining a queue as many club-goers might call their friends to join them. With this application in mind, it is important to understand the characteristics of nightclubs. Many nightclubs have waiting spaces for potential customers outside the club. Moreover, inside the club is where much of the activity happens. Thus, using phase-type distributions we can model the inside and outside of the club as two phases of services or a two dimensional

phase-type queue. The first phase of service can be considered "admittance" to the service with the second step being the service itself. Because the clubs' bouncers have the ability to admit customers into the venue from any position in the external queue and because each customer determines how long she stays in the club, we model this scenario as an infinite server queue. This process is visualized below, where μ_O and μ_I are the rates of each step of service.

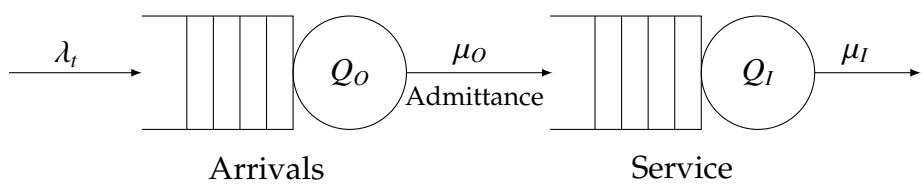


Figure 2.17: Club Queue Process Diagram.

We can represent the Club Queue using the two dimensional vector of queue lengths $Q(t)$ for $t \geq 0$, with coordinates $Q_I(t)$ and $Q_O(t)$ representing the service systems inside and outside the club, respectively. A fundamental managerial task is to figure out at what rate to admit club-goers into the club to maximize profitability while making the club attractive from the outside. This is non-trivial as a short line outside the club might signal to others that the club is not interesting and make them choose to not go inside the club. However, if the line is too long, there are many customers not actively generating revenue for the club and becoming frustrated with the wait outside. With this in mind, we construct the following objective function that maximizes the rate at which the bouncer of the club should let club-goers inside the club over the finite time horizon $[0, T]$, where $T > 0$.

$$\zeta(\mu_O(t)) = r_O\mu_O \mathrm{E}\left[Q_O(t)\right] + r_I \mathrm{E}\left[Q_I(t)\right] - c(\mu_O \mathrm{E}\left[Q_O(t)\right] - k)^2 - w\mu_O^2 \tag{2.49}$$

Here $r_O \geq 0$ and $r_I \geq 0$ are revenues generated from the cover outside and inside the club respectively. We also have that c is a penalty for having the overall admittance rate be too slow or too fast and finally, w is a penalty for admitting each individual customer too quickly. A complete formulation of this optimal control problem is presented next.

Problem 2.4.1 (Unconstrained Club Profit Model)**.**

$$\max_{\{\mu_O \geq 0\}} \quad \int_0^T \left[r_O \mu_O(t) \mathrm{E}\left[Q_O(t)\right] + r_I \mathrm{E}\left[Q_I(t)\right] - c(\mu_O(t)\mathrm{E}\left[Q_O(t)\right] - k)^2 - w\mu_O(t)^2 \right] \mathrm{d}t$$

subject to

$$\dot{\mathrm{E}}[\lambda(t)] = \beta \cdot (\lambda^* - \mathrm{E}[\lambda(t)]) + \alpha \cdot \mathrm{E}[\lambda(t)]$$

$$\dot{\mathrm{E}}[Q_O(t)] = \mathrm{E}[\lambda(t)] - \mu_O(t) \cdot \mathrm{E}[Q_O(t)]$$

$$\dot{\mathrm{E}}[Q_I(t)] = \mu_O(t) \cdot \mathrm{E}[Q_O(t)] - \mu_I \cdot \mathrm{E}[Q_I(t)]$$

The solution to this problem gives the optimal rate to admit club-goers across time in order to maximize the difference between club revenue and the queue length and admittance rate penalties. This is characterized by the following theorem.

Theorem 2.4.1. *The optimal solution to Problem 2.4.1 is given by $\mu_O^*(t)$, where*

$$\mu_O^*(t) = \frac{(r_O + 2ck - \gamma_1 + \gamma_2)\mathrm{E}\left[Q_O(t)\right]}{2w + 2c\mathrm{E}\left[Q_O(t)\right]^2} \tag{2.50}$$

for all $t \in [0, T]$.

Proof. We start by transforming the optimization model into a single Hamiltonian equation, which can be thought of as an unconstrained version of the Lagrangian. For this problem, we have the Hamiltonian $\mathcal{H}$ as

$$\mathcal{H}(t, \gamma) = \zeta(\mu_O(t)) - \gamma_1 \left(\dot{\mathrm{E}}[Q_O(t)] - \mathrm{E}[\lambda(t)] + \mu_O \mathrm{E}[Q_O(t)] \right) - \gamma_2 \left(\dot{\mathrm{E}}[Q_I(t)] - \mu_O \mathrm{E}[Q_O(t)] + \mu_I \mathrm{E}[Q_I(t)] \right) - \gamma_3 \left(\dot{\mathrm{E}}[\lambda(t)] - \beta \cdot (\lambda^*(t) - \mathrm{E}[\lambda(t)]) - \alpha \cdot \mathrm{E}[\lambda(t)] \right)$$

where each $\gamma_i \in \mathbb{R}$ for $i \in \{1, 2, 3\}$. To achieve optimality in the control problem, the method ensures that $\mu_O(t)$ is such that $\frac{\mathrm{d}\mathcal{H}}{\mathrm{d}\mu_O(t)} = 0$ for all $t \in [0, T]$. We see that the derivative of the Hamiltonian with respect to $\mu_O(t)$ is

$$\frac{\mathrm{d}\mathcal{H}}{\mathrm{d}\mu_O(t)} = r_O \mathrm{E}\left[Q_O(t)\right] - 2c\mu_O(t)\mathrm{E}\left[Q_O(t)\right]^2 + 2ck\mathrm{E}\left[Q_O(t)\right] - 2w\mu_O(t) - \gamma_1 \mathrm{E}\left[Q_O(t)\right] + \gamma_2 \mathrm{E}\left[Q_O(t)\right].$$

Thus, the optimal $\mu_O^*(t)$ is found by solving

$$0 = \frac{\mathrm{d}\mathcal{H}}{\mathrm{d}\mu_O(t)} = (r_O + 2ck - \gamma_1 + \gamma_2)\mathrm{E}\left[Q_O(t)\right] - (2c\mathrm{E}\left[Q_O(t)\right]^2 + 2w)\mu_O^*(t)$$

for $\mu_O^*(t)$, which yields the expression in Equation 2.50. Because the objective function is concave in $\mu_O(t)$ at every t, we have that this solution corresponds to a maximum. □

Using the differential equations shown in Section 2.3, this optimization problem can be solved numerically by the Forward Backward sweep method as in Niyirora and Pender (2016); Qin and Pender (2017); Lenhart and Workman (2007). We now give two example outputs of this method below.

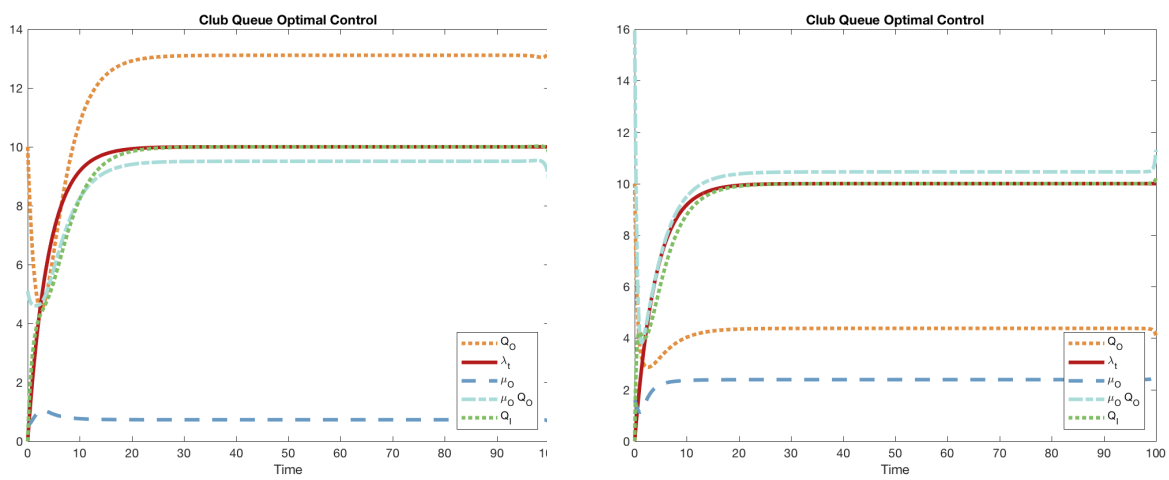


Figure 2.18: Example Forward Backward Sweep Implementation.

In the scenario on the left, the parameters are as follows: r_O, the external

entrance revenue rate, is equal to 100 units of currency per units of time. The revenue per person inside, r_I, is equal to 100 units of currency per person. The cost of deviating from the desired admittance rate k, c, is also 100, whereas $k = 8$. Finally, the penalty for admitting individuals too quickly, $w = 150$. On the right, w is instead 100 and $k = 12$. These changes have significant impacts on the resulting solution. On the left the outside queue is allowed to grow roughly three times as large whereas on the right μ_O is approximately twice the size of that on the left.

2.5 Conclusion and Final Remarks

In this chapter, we have analyzed a new infinite server stochastic queueing model that is driven by a Hawkes arrival process and phase-type distributed service. We are able to derive the exact moments and moment generating function for the Hawkes driven queue as well as the Hawkes process itself.

Although we have analyzed this queueing model in great detail, there are many extensions that are worthy of future study. One extension that we intend to explore is the impact of a non-stationary baseline intensity in the spirit of Massey and Pender (2013); Pender (2014a); Engblom and Pender (2014); Pender (2016a, 2015a,b, 2016b). In one simple example, we could set the baseline be $\lambda^*(t) = \lambda^* + \rho \cdot \sin(t)$. This analysis of a non-stationary baseline intensity is important not only because arrival rates of customers are not constant over time, but also because it is important to know how to distinguish and separate the impact of the time varying arrival rate from the impact of the stochastic dynamics of the self-excitation. The extension of one periodic function such as $\sin(t)$

seems analytically tractable, however, additional functions may require Fourier analysis.

Other extensions include the modeling of different types of queueing models other than the infinite server model. For example, it would be interesting to apply our analysis to the Erlang-A queueing model with abandonments. With regard to obtaining analytical expressions for the Erlang-A model, this is a non-trivial problem because even the Erlang-A queueing model with a Poisson arrival process is analytically somewhat intractable. This presents new challenges for deriving analytical formulas and approximations for the moment behavior of this type of queueing model. Work by Massey and Pender (2011); Pender (2014c,b, 2015a, 2016c); Daw and Pender (2019a) shows that simple closure approximations or spectral expansions can be effective at approximating the dynamics of the Erlang-A model and variants. Thus, a natural extension is to apply these techniques to the Erlang-A setting when it is driven by a Hawkes process. Not only do these approximations have the potential to describe the moment dynamics, but they can be used to stabilize performance measures like in Pender and Massey (2017). A detailed analysis of these extensions will provide a better understanding how the information that operations managers provide to their customers will affect the dynamics of these real world systems like in Pender et al. (2017a, 2018, 2017b). We plan to explore these extensions in subsequent work.

CHAPTER 3

ON THE DISTRIBUTIONS OF INFINITE SERVER QUEUES WITH BATCH ARRIVALS

3.1 Introduction

Queueing systems with batch arrivals have enjoyed a long and rich history of study, at least on the time scale of queueing theory. Researchers have been exploring models of this sort for no less than six decades, based on the April 1958 submission date of Miller Jr (1959). Given this stretch of time, a wide variety of systems and settings have been considered under the banner of batch arrivals. Much of the earliest work focuses on single server models, including Miller Jr (1959); Lucantoni (1991); Masuyama and Takine (2002); Liu and Templeton (1993) and Foster (1964), although infinite server models followed soon after, such as work by Shanbhag (1966) and Brown and Ross (1969). Later work has expanded the concept into a variety of related models, such as for priority queues Takagi and Takahashi (1991) and for handling server vacations Lee et al. (1995). Additionally, there is some work that proves heavy traffic limit theorems for queues with batch arrivals. Examples of this include Chiamsiri and Leonard (1981); Pang and Whitt (2012); Pender (2013). These papers show that one can approximate the queue length process with Brownian motion and Ornstein-Uhlenbeck processes and also show that one can exploit the approximations even in multi-server and non-Markovian settings.

Contents of this chapter have been published in Daw and Pender (2019b).

In this chapter we consider queues with arrivals occurring at times following a Poisson process, with consideration given to both non-stationary and stationary rates. We analyze both general and exponential service as conducted by infinitely many servers. Additionally, this work addresses both fixed and random batch sizes. Our analysis starts with the fixed batch size case. We begin by analyzing the transient behavior of the queue with Markovian service and time-varying arrival rates, providing explicit forms for the moment generating function, mean, and variance. Then, we show that if the arrival rate is stationary the resulting steady-state distribution can be written as a sum of independent, non-identical, scaled Poisson random variables. This leads us to uncover connections to the harmonic numbers and generalizations of the Hermite distribution. By viewing the batch arrival queue as a collection of infinite server sub-queues that receive solitary arrivals simultaneously, we are able to extend this Poisson sum construction to general service distributions. This perspective also provides an avenue for us to extend to random batch sizes. We also give fluid and diffusion scalings of the queue in the case of random batch sizes, as well as extending many of the results we found for fixed batch sizes.

One can note that the batch arrival queue may not always be given the name "batch," as many authors choose to use the term "bulk" instead. Predominantly, this reflects two leading strands of applications, where "bulk" often gives a connotation of transportation settings whereas "batch" frequently implies applications in communications. Just as practical by any other name, this family of models has also been studied in a wide variety of applications beyond these two. Perhaps one most distinct from other types of queueing models is particle splitting in DNA caused by radiation, as discussed in Sachs et al. (1992). In this application, primary particles arrive at a cell nucleus and cause DNA

double-strand breaks. These double-strand breaks occur in near simultaneity and are thus modeled as arriving in batches of random size, as it is possible that any number double-strand breaks will be induced. After they are induced, the double-strand breaks are then processed by cellular enzymes, corresponding to service in the queueing model. Another interesting and modern application of these models is in cloud-based data processing. In this case, the batches arriving to the system are collections of jobs submitted simultaneously. These jobs are then served by each being processed individually and returned. For more discussion, detailed models, and specific analysis for this setting, see works such as Lu et al. (2011); Pender and Phung-Duc (2016); Xie et al. (2017); Yekkehkhany et al. (2018) and references therein.

3.1.1 Main Contributions of Chapter

Our contributions in this chapter can be summarized as follows:

i) We show that an infinite server queue with batch arrivals at Poisson process epochs is equivalent in steady state distribution to a sum of scaled independent Poisson random variables, including for generally distributed service and randomly distributed batch sizes. For exponential service, this reveals a connection to the harmonic numbers and generalized Hermite distributions.

ii) We derive a limit of the queue length process in which the batch size grows infinitely large and the number of entities in the system is scaled inverse proportionally, yielding a novel distribution characterized by the exponential integral functions. For distributions that meet a divisibility condition,

we find that this also holds for random batch sizes.

iii) In the case of time-varying arrival rates we give a transient moment generating function for fixed batch sizes as well as means and variance for both fixed and randomly sized batches.

iv) We give fluid and diffusion limits of the queue for stationary arrival rates for batches of random size.

3.1.2 Organization of Chapter

The body of the remainder of this chapter is organized in two main sections: Sections 3.2 and 3.3. In Section 3.2 we consider systems in which the size of the batches is fixed. Similarly, we devote Section 3.3 to the case of randomly distributed batch sizes. At the beginning of each section we give a detailed overview of the contents within and provide context for the analysis in term of this project's scope. After these sections we conclude in Section 3.4.

3.2 Batches of Deterministic Size

In this section we will consider infinite server queues with arrivals occurring in batches of a fixed size. We will assume that the arrival epochs occur according to a Poisson process, including both stationary and non-stationary models. We also will investigate both exponentially and generally distributed service.

This section starts with studying the case of Markovian arrivals and service in transient state in Subsection 3.2.1. For a time-varying arrival rate, we give

the mean, variance, and moment generating function. We then use this in Subsection 3.2.2 to find the steady-state distribution of the queue. Upon observing that this can be represented as a sum of scaled Poisson random variables, we establish connections to generalized Hermite distributions and to the harmonic numbers. Taking motivation from this, we derive the distribution of the limit of the scaled system as the batch size grows infinitely large. Finally, in Subsection 3.2.3, we examine the batch queue as a collection of infinite server subqueues that simultaneously receive solitary arrivals. In doing so we extend our understanding of the steady-state distribution to the case of general service.

3.2.1 Transient Analysis of the Markovian Setting

We begin our analysis with the case of non-stationary Poisson arrival epochs and Markovian service. In Kendall notation, this is the $M_t^n/M/\infty$ queue. We let Q_t represent the number of entities present in the queueing system at time $t \geq 0$, which we often refer to as the "number in system." We will use this notation throughout the remainder of this work, where the precise setting of the queue will be implied by context. In this fully Markovian setting, we can use Dynkin's infinitesimal generator theorem to support our analysis. Specifically, we can note that for a sufficiently regular function $f : \mathbb{N} \to \mathbb{R}$, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[f(Q_t)\right] = \mathrm{E}\left[\lambda(t)\left(f(Q_t+n) - f(Q_t)\right) + \mu Q_t\left(f(Q_t-1) - f(Q_t)\right)\right], \tag{3.1}$$

for a batch arrival queue with arrival intensity $\lambda(t) > 0$. We will see in this subsection that this infinitesimal generator approach gives us a potent toolkit for exploring this model. Moreover, the insights we find in Markovian settings now and in Subsection 3.2.2 will provide intuition that will guide our investigation of this system when the Markov property does not hold.

To begin, we now derive the moment generating function of the number in system. We do so for a system with a non-stationary arrival rate given by a Fourier series, allowing these results to hold for all periodic arrival patterns.

Proposition 3.2.1. *For $\theta \in \mathbb{R}$, let $\mathcal{M}(\theta, t) = \mathrm{E}\left[e^{\theta Q_t}\right]$ be the moment generating function of the number in system of an infinite server queue with periodic arrival rate $\lambda + \sum_{k=1}^{\infty} a_k \cos(kt) + b_k \sin(kt) > 0$, arrival batch size $n \in \mathbb{Z}^+$, and exponential service rate $\mu > 0$. Then, $\mathcal{M}(\theta, t)$ is given by*

$$\begin{aligned}\mathcal{M}(\theta, t) = \left(e^{-\mu t}(e^{\theta} - 1) + 1\right)^{Q_0} e^{\sum_{j=1}^{n} \binom{n}{j}(e^{\theta}-1)^j \left(\frac{\lambda}{j\mu}\left(1-e^{-j\mu t}\right)+\sum_{k=1}^{\infty} \frac{(a_k j\mu - b_k k)}{k^2+j^2\mu^2}\left(\cos(kt)-e^{-j\mu t}\right)\right)} \\ \cdot e^{\sum_{j=1}^{n} \binom{n}{j}(e^{\theta}-1)^j \sum_{k=1}^{\infty} \frac{(a_k k + b_k j\mu)\sin(kt)}{k^2+j^2\mu^2}}\end{aligned} \tag{3.2}$$

for all time $t \geq 0$, where Q_0 is the initial number in system.

Proof. From Equation 3.1, the MGF is given by the solution to the partial differential equation

$$\frac{\partial}{\partial t}\mathcal{M}(\theta, t) = \left(\lambda + \sum_{k=1}^{\infty} a_k \cos(kt) + b_k \sin(kt)\right)\left(e^{n\theta} - 1\right)\mathcal{M}(\theta, t) + \mu\left(e^{-\theta} - 1\right)\frac{\partial}{\partial \theta}\mathcal{M}(\theta, t)$$

with initial solution $\mathcal{M}(\theta, 0) = e^{\theta Q_0}$. Because $\frac{\mathrm{d}\log(f(x))}{\mathrm{d}x} = \frac{1}{f(x)}\frac{\mathrm{d}f(x)}{\mathrm{d}x}$, we can observe that the partial differential equation for the cumulant generating function $G(\theta, t) = \log\left(\mathrm{E}\left[e^{\theta Q_t}\right]\right)$ is

$$\mu(1 - e^{-\theta})\frac{\partial G(\theta, t)}{\partial \theta} + \frac{\partial G(\theta, t)}{\partial t} = \left(\lambda + \sum_{k=1}^{\infty} a_k \cos(kt) + b_k \sin(kt)\right)(e^{n\theta} - 1),$$

with the initial condition $G(\theta, 0) = \log\left(\mathrm{E}\left[e^{\theta Q_0}\right]\right) = \theta Q_0$. We will now solve this PDE by the method of characteristics. We begin by establishing the characteris-

tic ODE's and corresponding initial solutions as follows:

$$\begin{aligned}
\frac{\mathrm{d}\theta}{\mathrm{d}s}(r,s) &= \mu(1-e^{-\theta}), && \theta(r,0) = r,\\
\frac{\mathrm{d}t}{\mathrm{d}s}(r,s) &= 1, && t(r,0) = 0,\\
\frac{\mathrm{d}g}{\mathrm{d}s}(r,s) &= \left(\lambda + \sum_{k=1}^{\infty} a_k\cos(kt) + b_k\sin(kt)\right)(e^{n\theta}-1), && g(r,0) = rQ_0.
\end{aligned}$$

The first two of these initial value problems yield the following solutions.

$$\begin{aligned}
\theta(r,s) &= \log(e^{c_1(r)+\mu s}+1) && \rightarrow \quad \theta(r,s) = \log\left((e^r-1)e^{\mu s}+1\right)\\
t(r,s) &= s + c_2(r) && \rightarrow \quad t(r,s) = s
\end{aligned}$$

Therefore we can simplify the remaining characteristic ODE to

$$\begin{aligned}
\frac{\mathrm{d}g}{\mathrm{d}s}(r,s) &= \left(\lambda + \sum_{k=1}^{\infty} a_k\cos(ks) + b_k\sin(ks)\right)\left(((e^r-1)e^{\mu s}+1)^n - 1\right)\\
&= \left(\lambda + \sum_{k=1}^{\infty} a_k\cos(ks) + b_k\sin(ks)\right)\sum_{j=1}^{n}\binom{n}{j}(e^r-1)^j e^{j\mu s},
\end{aligned}$$

and this produces the general solution of

$$g(r,s) = c_3(r) + \sum_{j=1}^{n}\binom{n}{j}(e^r-1)^j\left(\frac{\lambda}{j\mu} + \sum_{k=1}^{\infty}\frac{(a_k j\mu - b_k k)\cos(ks)}{k^2+j^2\mu^2} + \frac{(a_k k + b_k j\mu)\sin(ks)}{k^2+j^2\mu^2}\right)e^{j\mu s}.$$

This now equates to

$$\begin{aligned}
g(r,s) = rQ_0 &+ \sum_{j=1}^{n}\binom{n}{j}(e^r-1)^j\left(\frac{\lambda}{j\mu}\left(e^{j\mu s}-1\right) + \sum_{k=1}^{\infty}\frac{(a_k j\mu - b_k k)}{k^2+j^2\mu^2}\left(\cos(ks)e^{j\mu s}-1\right)\right.\\
&\left. + \sum_{k=1}^{\infty}\frac{(a_k k + b_k j\mu)\sin(ks)}{k^2+j^2\mu^2}e^{j\mu s}\right)
\end{aligned}$$

as the solution to the initial value problem. We now find the solution to the original PDE by solving for each characteristic variable in terms of t and θ and then substituting these expression into $g(r,s)$. That is, for $s = t$ and

$r = \log\left(e^{-\mu t}(e^{\theta} - 1) + 1\right)$, we have that

$$\begin{aligned} G(\theta, t) &= g\left(\log\left(e^{-\mu t}(e^{\theta} - 1) + 1\right), t\right) \\ &= \log\left(e^{-\mu t}(e^{\theta} - 1) + 1\right) Q_0 + \sum_{j=1}^{n} \binom{n}{j} (e^{\theta} - 1)^j \Bigg(\frac{\lambda}{j\mu}\left(1 - e^{-j\mu t}\right) + \sum_{k=1}^{\infty} \frac{(a_k j\mu - b_k k)}{k^2 + j^2\mu^2} \\ &\quad \cdot \left(\cos(kt) - e^{-j\mu t}\right) + \sum_{k=1}^{\infty} \frac{(a_k k + b_k j\mu)}{k^2 + j^2\mu^2} \sin(kt) \Bigg). \end{aligned}$$

To conclude the proof, we note that $\mathcal{M}(\theta, t) = e^{G(\theta,t)}$. □

We now extend this analysis through two following corollaries. First, for systems with a stationary arrival rate, say $\lambda > 0$, we can further specify the moment generating function explicitly in Corollary 3.2.2. This will be of use when we explore the distribution of the queue in steady-state, which we begin in Subsection 3.2.2. As with Proposition 3.2.1, the uniqueness of moment generating functions will aid us in later exploration of the distributions within this model and within generalizations of it.

Corollary 3.2.2. *For $\theta \in \mathbb{R}$, let $\mathcal{M}(\theta, t) = \mathrm{E}\left[e^{\theta Q_t}\right]$ be the moment generating function of the number in system of an infinite server queue with stationary arrival rate $\lambda > 0$, arrival batch size $n \in \mathbb{Z}^+$, and exponential service rate $\mu > 0$. Then, $\mathcal{M}(\theta, t)$ is given by*

$$\mathcal{M}(\theta, t) = \left(e^{-\mu t}(e^{\theta} - 1) + 1\right)^{Q_0} e^{\lambda \sum_{j=1}^{n} \binom{n}{j} \frac{(e^{\theta}-1)^j}{j\mu}\left(1 - e^{-j\mu t}\right)} \tag{3.3}$$

for all time $t \geq 0$, where Q_0 is the initial number in system.

For the second direct result of Proposition 3.2.1, we can also give explicit expressions for the transient mean and variance of the queue. We derive these equations from the first and second derivatives, respectively, of the cumulant generating function $\log(\mathrm{E}\left[e^{Q_t}\right])$.

Corollary 3.2.3. *Let Q_t be an infinite server queue with periodic arrival rate $\lambda + \sum_{k=1}^{\infty} a_k \cos(kt) + b_k \sin(kt) > 0$, arrival batch size $n \in \mathbb{Z}^+$, and exponential service rate $\mu > 0$. Then, the mean and variance of the queue are given by*

$$\begin{aligned}
\mathrm{E}\left[Q_t\right] &= Q_0 e^{-\mu t} + \frac{n\lambda}{\mu}\left(1 - e^{-\mu t}\right) + \sum_{k=1}^{\infty} \frac{n(a_k\mu - b_k k)}{k^2 + \mu^2}\left(\cos(kt) - e^{-\mu t}\right) \\
&\quad + \sum_{k=1}^{\infty} \frac{n(a_k k + b_k \mu)}{k^2 + \mu^2} \sin(kt) \qquad (3.4) \\
\mathrm{Var}\left(Q_t\right) &= Q_0\left(e^{-\mu t} - e^{-2\mu t}\right) + \frac{n\lambda}{\mu}\left(1 - e^{-\mu t}\right) + \sum_{k=1}^{\infty} \frac{n(a_k\mu - b_k k)}{k^2 + \mu^2}\left(\cos(kt) - e^{-\mu t}\right) \\
&\quad + \sum_{k=1}^{\infty} \frac{n(a_k k + b_k \mu)}{k^2 + \mu^2} \sin(kt) + \frac{n(n-1)\lambda}{2\mu}\left(1 - e^{-2\mu t}\right) + \sum_{k=1}^{\infty} \frac{n(n-1)(2a_k\mu - b_k k)}{k^2 + 4\mu^2} \\
&\quad \cdot \left(\cos(kt) - e^{-2\mu t}\right) + \sum_{k=1}^{\infty} \frac{n(n-1)(a_k k + 2b_k \mu)}{k^2 + 4\mu^2} \sin(kt) \qquad (3.5)
\end{aligned}$$

for all time $t \geq 0$, where Q_0 is the initial number in system.

In the remainder of this work we will explore various modifications of this model, including general service and randomized batch sizes. The results of this subsection will serve as cornerstone throughout much of this upcoming analysis, both supporting the underlying derivation techniques and providing the intuition for new perspectives.

3.2.2 The Markovian System with Stationary Arrival Rates

Our first departure from our initial model will be modest: instead of studying the fully Markovian, non-stationary, fixed batch size system in transient time we will now move to addressing the stationary case, with much of our analysis focused on the system in steady-state. This simplified setting will allow us to extract greater intuition from our prior findings, which in turn will support

generalization of the service distribution and randomization of the batch sizes. To begin, we find a representation of the steady-state distribution of the queue length in terms of a sum of independent, scaled Poisson random variables.

Proposition 3.2.4. *In steady-state the distribution of the number in system of an infinite server queue with stationary arrival rate $\lambda > 0$, arrival batch size $n \in \mathbb{Z}^+$, and exponential service rate $\mu > 0$ is*

$$Q_\infty(n) \stackrel{D}{=} \sum_{j=1}^{n} jY_j \tag{3.6}$$

where $Y_j \sim \text{Pois}\left(\frac{\lambda}{j\mu}\right)$ are independent.

Proof. From Proposition 3.2.1, we have that the steady-state moment generating function of the queue is given by

$$\lim_{t\to\infty} \mathcal{M}(\theta, t) = e^{\lambda \sum_{k=1}^{n} \binom{n}{k} \frac{\left(e^\theta - 1\right)^k}{k\mu}}.$$

To satisfy our stated Poisson form, we are now left to show that $\sum_{k=1}^{n} \binom{n}{k} \frac{(e^\theta-1)^k}{k} = \sum_{k=1}^{n} \frac{e^{k\theta}-1}{k}$ for all $n \in \mathbb{Z}^+$. We proceed by induction. In the base case of $n = 1$ we have $e^\theta - 1 = e^\theta - 1$ and so we are left to show the inductive step. We now assume $\sum_{k=1}^{n} \binom{n}{k} \frac{(e^\theta-1)^k}{k} = \sum_{k=1}^{n} \frac{e^{k\theta}-1}{k}$ holds at n. Then, by the Pascal triangle identity $\binom{n}{k} = \binom{n+1}{k} - \binom{n}{k-1}$ and our inductive hypothesis we can observe

$$\sum_{k=1}^{n} \frac{e^{k\theta} - 1}{k} = \sum_{k=1}^{n} \binom{n}{k} \frac{(e^\theta - 1)^k}{k} = \sum_{k=1}^{n} \left(\binom{n+1}{k} - \binom{n}{k-1} \right) \frac{(e^\theta - 1)^k}{k}.$$

Now, by applying the identity $\binom{n}{k-1} = \frac{k}{n+1}\binom{n+1}{k}$ and distributing the summation we can further note that

$$\begin{aligned}
\sum_{k=1}^{n} \left(\binom{n+1}{k} - \binom{n}{k-1} \right) \frac{(e^\theta - 1)^k}{k} &= \sum_{k=1}^{n} \left(\binom{n+1}{k} - \frac{k}{n+1}\binom{n+1}{k} \right) \frac{(e^\theta - 1)^k}{k} \\
&= \sum_{k=1}^{n} \binom{n+1}{k} \frac{(e^\theta - 1)^k}{k} - \frac{\sum_{k=1}^{n} \binom{n+1}{k} (e^\theta - 1)^k}{n+1}.
\end{aligned}$$

Now, we can use the binomial theorem to see that

$$\sum_{k=1}^{n}\binom{n+1}{k}(e^{\theta}-1)^{k} = (e^{\theta}-1+1)^{n+1}-1-(e^{\theta}-1)^{n+1} = e^{(n+1)\theta}-1-(e^{\theta}-1)^{n+1},$$

and so we can now simplify and find

$$\sum_{k=1}^{n}\binom{n+1}{k}\frac{(e^{\theta}-1)^{k}}{k} - \frac{\sum_{k=1}^{n}\binom{n+1}{k}(e^{\theta}-1)^{k}}{n+1} = \sum_{k=1}^{n}\binom{n+1}{k}\frac{(e^{\theta}-1)^{k}}{k} + \frac{(e^{\theta}-1)^{n+1}}{n+1} - \frac{e^{(n+1)\theta}-1}{n+1}.$$

Hence, in conjunction with our initial equation, we have that

$$\sum_{k=1}^{n}\frac{e^{k\theta}-1}{k} = \sum_{k=1}^{n}\binom{n+1}{k}\frac{(e^{\theta}-1)^{k}}{k} + \frac{(e^{\theta}-1)^{n+1}}{n+1} - \frac{e^{(n+1)\theta}-1}{n+1},$$

and by rearranging terms we now complete the inductive approach:

$$\sum_{k=1}^{n+1}\frac{e^{k\theta}-1}{k} = \sum_{k=1}^{n+1}\binom{n+1}{k}\frac{(e^{\theta}-1)^{k}}{k}.$$

We can now observe that we have a moment generating function that is a product of moment generating functions of scaled Poisson random variables, which yields the stated result. □

While we will continue to explore the stationary arrival rate setting throughout this subsection, we note that this Poisson sum representation will be a leading inspiration in the sequel. Specifically, in Subsection 3.2.3 we will find intuition for this result by viewing the batch arrival queue as a collection of subsystems.

Remark. In addition to this Poisson sum representation, we can also express the steady-state MGF in terms of the truncated polylogarithm function and harmonic numbers. From the MGF of the queue length in steady state for $\theta < 0$, we can observe that

$$\lim_{t\to\infty}\mathcal{M}(\theta,t) = e^{\frac{\lambda}{\mu}\sum_{k=1}^{n}\frac{e^{k\theta}-1}{k}} = e^{\frac{\lambda}{\mu}(\mathrm{Li}(e^{\theta},n,1)-H_n)}$$

where we have H_n as the nth harmonic number, given by $\sum_{k=1}^{n} \frac{1}{k}$, and where the truncated polylogarithm function $\mathrm{Li}(z, n, s)$ is defined as

$$\mathrm{Li}(z, n, s) = \sum_{k=1}^{n} \frac{z^k}{k^s}.$$

This decomposition into Poisson random variables can be quite useful from a computational standpoint. It allows us to simulate the steady state quite easily since we only need to simulate n Poisson random variables instead of simulating an actual queue, which could be quite expensive. We can now observe that this construction also yields an interesting connection to both the harmonic number and Hermite distributions, as suggested in the remark above. To motivate our following analysis, suppose that $n = 2$. Then, steady-state queue length has steady-state moment generating function given by

$$\mathcal{M}_n(\theta, \infty) = e^{\frac{\lambda}{\mu}(e^{\theta}-1)+\frac{\lambda}{2\mu}(e^{2\theta}-1)}.$$

We can now observe that this MGF corresponds to a Hermite distribution with parameters $\frac{\lambda}{\mu}$ and $\frac{\lambda}{2\mu}$. This implies that the steady-state CDF of the queue at $n = 2$ is

$$P(Q_\infty(2) \leq k) = e^{-\frac{3\lambda}{2\mu}} \sum_{i=0}^{\lfloor k \rfloor} \sum_{j=0}^{\lfloor i/2 \rfloor} \frac{\left(\frac{\lambda}{\mu}\right)^{i-2j} \left(\frac{\lambda}{2\mu}\right)^{j}}{(i-2j)!\,j!} = e^{-\frac{3\lambda}{2\mu}} \sum_{i=0}^{\lfloor k \rfloor} \sum_{j=0}^{\lfloor i/2 \rfloor} \frac{\left(\frac{\lambda}{\mu}\right)^{i-j} 2^{-j}}{(i-2j)!\,j!}.$$

Furthermore, the steady-state PMF of the queue length is given by

$$P(Q_\infty(2) = i) = e^{-\frac{3\lambda}{2\mu}} \sum_{j=0}^{\lfloor i/2 \rfloor} \frac{\left(\frac{\lambda}{\mu}\right)^{i-j} 2^{-j}}{(i-2j)!\,j!}.$$

This observation prompts us to ponder generalizations for $n \geq 3$. The term "generalized Hermite distribution" has taken on slightly varying (yet always interesting) definitions for different authors. For readers interested in the Hermite distribution and popular generalizations of it, we suggest Kemp and Kemp

(1965); Gupta and Jain (1974), and Milne and Westcott (1993). In our setting we note that the coefficients of $\frac{\lambda}{\mu}$ in the MGF for batch size n will be $1, \frac{1}{2}, \frac{1}{3}, \ldots, \frac{1}{n}$. For this reason, we think of this particular generalization of Hermite distributions to be the *harmonic Hermite distribution*. We can now note that because of this harmonic structure we can instead fully characterize the distribution simply by n and $\frac{\lambda}{\mu}$. In the following proposition we find a useful recursion for the probability mass function of this distribution at all $n \in \mathbb{Z}^+$.

Proposition 3.2.5. *Let $Q_t(n)$ be an infinite server batch arrivals queue with arrival rate $\lambda > 0$, batch size $n \in \mathbb{Z}^+$, and service rate $\mu > 0$. Then, the steady-state distribution of the queue is given by the recursion*

$$\mathbb{P}(Q_\infty(n) = j) = p_j = \sum_{i=1}^{n} i p_{j-i} \frac{\lambda}{ij\mu} = \sum_{i=1}^{n} p_{j-i} \frac{\lambda}{j\mu}, \tag{3.7}$$

where $p_0 = e^{-\frac{\lambda}{\mu} H_n}$ for H_n as the nth harmonic number and $p_k = 0$ for all $k < 0$. Thus, we say that $Q_\infty(n)$ follows the "harmonic Hermite distribution" with parameter n.

Proof. We know from our Poisson representation of the steady state queue length that the steady-state moment generating function is

$$M(\theta) = \sum_{j=0}^{\infty} \mathbb{P}(Q_\infty(n) = j)\theta^j = \sum_{j=0}^{\infty} p_j \theta^j = \exp\left(\sum_{i=1}^{n} \frac{\lambda}{i\mu} \left(\theta^i - 1\right) \right).$$

If we take the logarithm of both sides we see that we have

$$\log\left(\sum_{j=0}^{\infty} p_j \theta^j \right) = \sum_{i=1}^{n} \frac{\lambda}{i\mu} \left(\theta^i - 1\right).$$

Now we take the derivative of both sides with respect to the parameter θ and this yields the following expression

$$\frac{\sum_{j=1}^{\infty} j p_j \theta^{j-1}}{\sum_{j=0}^{\infty} p_j \theta^j} = \sum_{i=1}^{n} \frac{\lambda}{\mu} \theta^{i-1}.$$

By moving the denominator to the righthand side, we have that

$$\sum_{j=1}^{\infty} j p_j \theta^{j-1} = \left(\sum_{j=0}^{\infty} p_j \theta^j\right)\left(\sum_{i=1}^{n} \frac{\lambda}{\mu}\theta^{i-1}\right).$$

Finally, by matching similar powers of θ on the left and right sides, we complete the proof. □

From the above result, we see that for the steady state queue length $Q_\infty(n)$ we can derive the specific probabilities,

$$\begin{aligned} p_0 &= e^{-\frac{\lambda}{\mu}H_n}, \\ p_1 &= \frac{\lambda}{\mu}p_0 = \frac{\lambda}{\mu}e^{-\frac{\lambda}{\mu}H_n}, \\ p_2 &= \frac{\lambda}{2\mu}(p_0 + p_1) = \frac{\lambda}{2\mu}e^{-\frac{\lambda}{\mu}H_n} + \frac{\lambda^2}{2\mu^2}e^{-\frac{\lambda}{\mu}H_n}. \end{aligned}$$

We can repeat this process as needed for any desired probability. From Proposition 3.2.4, we can observe that the mean number in system grows linearly with the batch size, meaning that the mean of the nth harmonic Hermite distribution is

$$\mathrm{E}\left[Q_\infty(n)\right] = \sum_{j=1}^{n} j\mathrm{E}\left[Y_j\right] = \frac{n\lambda}{\mu}. \tag{3.8}$$

We can observe further that the second moment and variance are quadratic functions of n:

$$\mathrm{E}\left[Q_\infty(n)^2\right] = \mathrm{E}\left[\left(\sum_{j=1}^{n} jY_j\right)^2\right] = \frac{n(n+1)\lambda}{2\mu} + n^2\frac{\lambda^2}{\mu^2},$$

$$\mathrm{Var}[Q_\infty(n)] = \mathrm{E}\left[Q_\infty(n)^2\right] - \mathrm{E}\left[Q_\infty(n)\right]^2 = \frac{n(n+1)\lambda}{2\mu}.$$

We note that from Proposition 3.2.4 and the following remark, the moment generating function of this distribution is given by

$$\lim_{t\to\infty} \mathcal{M}(\theta, t) = e^{\frac{\lambda}{\mu}\sum_{k=1}^{n}\frac{e^{k\theta}-1}{k}} = e^{\frac{\lambda}{\mu}(\mathrm{Li}(e^\theta, n, 1) - H_n)}. \tag{3.9}$$

If one is to consider this system as the batch size grows infinitely large we can see from Equations 3.8 and 3.9 that the number in system will grow proportionally, tending to infinity as n does. This leads us to ponder the limiting object of the scaled number in system $\frac{Q_t(n)}{n}$ as the batch size grows.

We begin by using Equation 3.9 with θ replaced by $\frac{\theta}{n}$ to see that the steady-state moment generating function of this scaled queue length is

$$\lim_{t\to\infty} \mathcal{M}(\theta, t) = e^{\frac{\lambda}{\mu}\sum_{k=1}^{n} \frac{e^{\frac{k}{n}\theta}-1}{k}}. \tag{3.10}$$

Furthermore, by replacing θ with $\frac{\theta}{n}$ and $Q_0(n)$ with $\frac{Q_0(n)}{n}$ in Proposition 3.2.1, we can note that the transient moment generating function for this scaled system with constant arrival rate is given by

$$\mathrm{E}\left[e^{\theta\cdot\frac{Q_t(n)}{n}}\right] \equiv \mathcal{M}_n(\theta, t) = \left(e^{-\mu t}(e^{\frac{\theta}{n}} - 1) + 1\right)^{\frac{Q_0}{n}} e^{\lambda \sum_{k=1}^{n} \binom{n}{k} \frac{\left(e^{\theta/n}-1\right)^k}{k\mu}\left(1-e^{-k\mu t}\right)}.$$

Additionally, we can also observe that the steady-state distribution of the scaled queue can also be interpreted as a sum of Poisson random variables through direction application of Proposition 3.2.4 or by inspection of Equation 3.10. This representation is

$$\frac{Q_\infty(n)}{n} \stackrel{D}{=} \sum_{j=1}^{n} \frac{j}{n} Y_j, \tag{3.11}$$

where again $Y_j \sim \mathrm{Pois}\left(\frac{\lambda}{j\mu}\right)$.

We now consider the limit as $n \to \infty$, in which we are both sending the size of batches of arrivals to infinity while also scaling the size of the queue inversely. We can use this construction to move beyond just the mean and variance and instead explicitly state every cumulant of the scaled queue. In Proposition 3.2.6 we give exact expressions of all steady-state cumulants of the scaled queue as functions of the Bernoulli numbers. Further, we find a convenient form of every cumulant of the scaled queue as the batch size grows to infinity.

Proposition 3.2.6. *Let $\lambda > 0$ be the arrival rate of batches of size $n \in \mathbb{Z}^+$ to an infinite server queue with exponential service rate $\mu > 0$. Then, the kth steady-state cumulant of the scaled queue $C^k\left[\frac{Q_\infty(n)}{n}\right]$ is given by*

$$C^k\left[\frac{Q_\infty(n)}{n}\right] = \frac{\frac{n^k}{k} + \frac{1}{2}n^{k-1} + \sum_{j=2}^{k-1}\frac{\mathrm{B}_j}{j!}(k-1)_{j-1}n^{k-j}}{n^k}. \tag{3.12}$$

where $(n)_i = \frac{n!}{(n-i)!}$ is the ith falling factorial of n and B_i is the ith Bernoulli number, which is defined as

$$B_i = \sum_{k=0}^{i}\sum_{j=0}^{k}(-1)^j\binom{k}{j}\frac{(j+1)^i}{k+1}.$$

Moreover, we have that $\lim_{n\to\infty} C^k\left[\frac{Q_\infty(n)}{n}\right] = \frac{\lambda}{k\mu}$.

Proof. From our prior observation that $\frac{Q_\infty(n)}{n} \overset{D}{=} \sum_{j=1}^{n}\frac{j}{n}Y_j$ where $Y_j \sim \mathrm{Pois}\left(\frac{\lambda}{j\mu}\right)$, we have that

$$C^k\left[\frac{Q_\infty(n)}{n}\right] = C^k\left[\sum_{j=1}^{n}\frac{j}{n}Y_j\right] = \sum_{j=1}^{n}C^k\left[\frac{j}{n}Y_j\right] = \sum_{j=1}^{n}\frac{j^k}{n^k}C^k\left[Y_j\right] = \frac{\lambda}{\mu n^k}\sum_{j=1}^{n}j^{k-1},$$

from the independence of these Poisson distributions. Now, by using Faulhaber's formula as given in Knuth (1993), we achieve the stated result. □

Just as we built from inherited expressions for the mean and variance to specify every cumulant in Proposition 3.2.6, we can also find the limit of the transient-state moment generating function for the scaled queue given in Equation 3.9.

Proposition 3.2.7. *Let Q_t be an infinite server queue with arrival rate $\lambda > 0$, arrival batch size $n \in \mathbb{Z}^+$, and exponential service rate $\mu > 0$. For $\theta \in \mathbb{R}$, let*

$$\mathcal{M}_\infty(\theta, t) = \lim_{n\to\infty}\mathrm{E}\left[e^{\frac{\theta Q_t(n)}{n}}\right].$$

Then, $\mathcal{M}_\infty(\theta, t)$ is given by

$$\mathcal{M}_\infty(\theta, t) = \begin{cases} e^{\frac{\lambda}{\mu}\left(\mathrm{Ei}(\theta) - \mathrm{Ei}(\theta e^{-\mu t}) - \mu t\right)} & \text{if } \theta > 0, \\ e^{\frac{\lambda}{\mu}\left(E_1(-\theta e^{-\mu t}) - E_1(-\theta) - \mu t\right)} & \text{if } \theta < 0, \\ 1 & \text{if } \theta = 0, \end{cases} \tag{3.13}$$

for all time $t \geq 0$, where the exponential integral functions $\mathrm{Ei}(x)$ *and* $E_1(x)$ *are defined*

$$\mathrm{Ei}(x) = -\int_{-x}^{\infty} \frac{e^{-s}}{s} \mathrm{d}s, \quad E_1(x) = \int_{x}^{\infty} \frac{e^{-s}}{s} \mathrm{d}s,$$

and are real-valued for $x > 0$.

Proof. While conventions may vary by application area, in this work we use the definition of exponential integral function given by

$$\mathrm{Ei}(x) = -\int_{-x}^{\infty} \frac{e^{-s}}{s} \mathrm{d}s.$$

By taking the limit of the MGF of the scaled queue, we have that

$$\frac{\partial}{\partial t}\mathcal{M}_\infty(\theta, t) = \lambda\left(e^{\theta} - 1\right)\mathcal{M}_\infty(\theta, t) - \mu\theta\frac{\partial}{\partial \theta}\mathcal{M}_\infty(\theta, t)$$

with initial solution $\mathcal{M}_\infty(\theta, 0) = \lim_{n\to\infty} e^{\frac{\theta Q_0}{n}} = 1$. In the same manner as the proof of Theorem 3.2.1, we solve the PDE of the cumulant generating function through use of the method of characteristics. We start by establishing the characteristic ODE's:

$$\begin{aligned} &\frac{\mathrm{d}\theta}{\mathrm{d}s}(r, s) = \mu\theta, && \theta(r, 0) = r, \\ &\frac{\mathrm{d}t}{\mathrm{d}s}(r, s) = 1, && t(r, 0) = 0, \\ &\frac{\mathrm{d}g}{\mathrm{d}s}(r, s) = \lambda(e^{\theta} - 1), && g(r, 0) = 0. \end{aligned}$$

We now solve the first two initial value problems and find

$$\begin{aligned} \theta(r, s) &= c_1(r)e^{\mu s} && \rightarrow \quad \theta(r, s) = re^{\mu s}, \\ t(r, s) &= s + c_2(r) && \rightarrow \quad t(r, s) = s. \end{aligned}$$

This allows us to simplify the third characteristic equation to

$$\frac{\mathrm{d}g}{\mathrm{d}s}(r,s) = \lambda(e^{re^{\mu s}} - 1).$$

Because $\theta = re^{\mu s}$, we can note that r and θ will match in sign: $r > 0$ if and only if $\theta > 0$. If $\theta > 0$, the general solution to this ODE is

$$g(r,s) = c_3(r) + \frac{\lambda}{\mu}\left(\mathrm{Ei}(re^{\mu s}) - \mu s\right),$$

whereas if $\theta < 0$, the solution is instead

$$g(r,s) = c_3(r) - \frac{\lambda}{\mu}\left(E_1(-re^{\mu s}) + \mu s\right).$$

This follows from the fact that for $x > 0$ the exponential integral functions are such that $\mathrm{Ei}(x) = -E_1(-x) - i\pi$; that is, the real parts of $E_1(-x)$ and $-\mathrm{Ei}(x)$ are the same. Moreover, for $x > 0$ one can consider $\mathrm{Ei}(x)$ as the real part of $-E_1(-x)$. Additionally, $E_1(x)$ is real for all $x > 0$. Hence, we use each definition of the exponential integral function when appropriate. As an alternative, we could replace each of these functions with `real`$(-E_1(-x))$ to have a single expression for both positive and negative x. For a collection of facts regarding the exponential integral functions, see Pages 228-237 of Abramowitz and Stegun (1965).

Now, using this we have that the corresponding solutions to the initial value problems will be

$$g(r,s) = \begin{cases} \frac{\lambda}{\mu}\left(\mathrm{Ei}(re^{\mu s}) - \mathrm{Ei}(r) - \mu s\right) & \text{if } r > 0, \\ \frac{\lambda}{\mu}\left(E_1(-r) - E_1(-re^{\mu s}) - \mu s\right) & \text{if } r < 0. \end{cases}$$

Hence, for $s = t$ and $r = \theta e^{-\mu t}$, this yields

$$G(\theta,t) = g\left(\theta e^{-\mu t}, t\right) = \begin{cases} \frac{\lambda}{\mu}\left(\mathrm{Ei}(\theta) - \mathrm{Ei}(\theta e^{-\mu t}) - \mu t\right) & \text{if } \theta > 0, \\ \frac{\lambda}{\mu}\left(E_1(-\theta e^{-\mu t}) - E_1(-\theta) - \mu t\right) & \text{if } \theta < 0. \end{cases}$$

By $\mathcal{M}_\infty(\theta,t) = e^{G_\infty(\theta,t)}$, we complete the proof. □

By consequence, we can also give the moment generating function in steady-state.

Corollary 3.2.8. *The moment generating function of the scaled number in system in steady-state as $n \to \infty$ is given by*

$$\mathcal{M}_\infty(\theta) = \begin{cases} \theta^{-\frac{\lambda}{\mu}} e^{\frac{\lambda}{\mu}(\mathrm{Ei}(\theta)-\gamma)} & \text{if } \theta > 0, \\ (-\theta)^{-\frac{\lambda}{\mu}} e^{-\frac{\lambda}{\mu}(E_1(-\theta)+\gamma)} & \text{if } \theta < 0, \\ 1 & \text{if } \theta = 0, \end{cases} \tag{3.14}$$

where γ is the Euler-Mascheroni constant.

Proof. From Abramowitz and Stegun (1965), for $x > 0$ we can expand the exponential integral functions as

$$\mathrm{Ei}(x) = \gamma + \log(x) + \sum_{k=1}^{\infty} \frac{x^k}{kk!}, \quad E_1(x) = -\gamma - \log(x) - \sum_{k=1}^{\infty} \frac{(-x)^k}{kk!}, \tag{3.15}$$

where γ is the Euler-Mascheroni constant. By expanding $\mathrm{Ei}(\theta e^{-\mu t})$ and $E_1(-\theta e^{-\mu t})$ in the respective cases of positive and negative θ and taking the limit as $t \to \infty$, we achieve the stated result. □

As a demonstration of the convergence of the steady-state moment generating functions of the batch scaled queues to the expression given in Corollary 3.2.8, we plot the first four cases in comparison to the limiting scenario in Figure 3.1.

While it can be argued that even in steady-state the form of this moment generating function is unfamiliar, we can still observe interesting characteristics of it. In particular, for $\theta < 0$ we can uncover a connection back to the harmonic numbers. We now discuss this in the following remark.

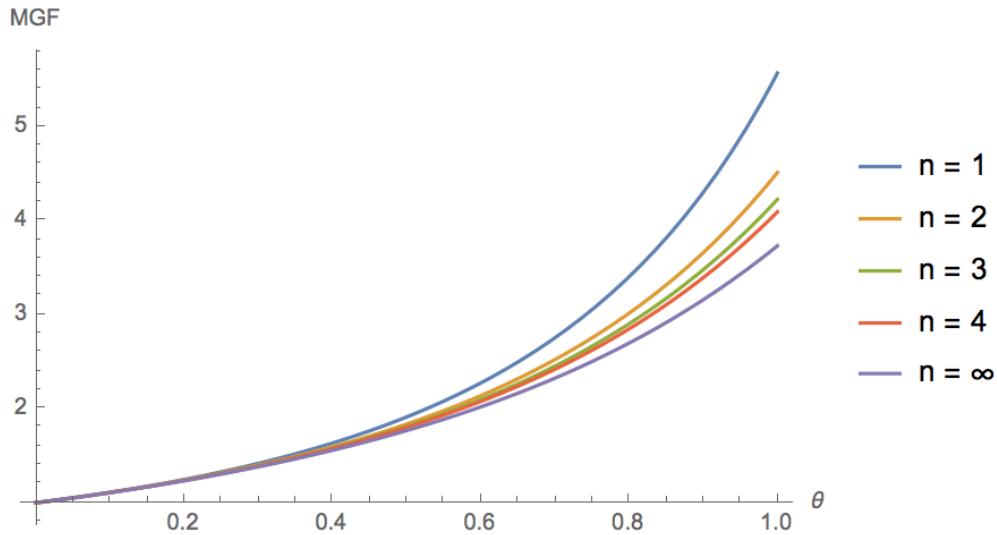


Figure 3.1: Steady-state MGF of the scaled queue for increasing batch size where $\frac{\lambda}{\mu} = 1$.

Remark. Using Equation 3.15, we can note that for $\theta < 0$ the steady-state moment generating function of limit of the scaled queue can be expressed

$$M(\theta) = (-\theta)^{-\frac{\lambda}{\mu}} e^{-\frac{\lambda}{\mu}(E_1(-\theta)+\gamma)} = e^{-\frac{\lambda}{\mu}\left(E_1(-\theta)+\gamma+\log(-\theta)\right)} = e^{-\frac{\lambda}{\mu}\left(-\sum_{k=1}^{\infty}\frac{\theta^k}{kk!}\right)}.$$

From Dattoli and Srivastava (2008), we have that $-e^x \sum_{k=1}^{\infty} \frac{(-x)^k}{kk!}$ is an exponential generating function for the harmonic numbers. That is,

$$-e^x \sum_{k=1}^{\infty} \frac{(-x)^k}{kk!} = \sum_{n=1}^{\infty} \frac{x^n}{n!} H_n$$

where H_n is the nth harmonic number. Thus, for $\theta < 0$ the steady-state moment generating function of this limiting object can be further simplified to

$$M(\theta) = e^{-\frac{\lambda}{\mu}\left(-\sum_{k=1}^{\infty}\frac{\theta^k}{kk!}\right)} = e^{-\frac{\lambda}{\mu}\sum_{n=1}^{\infty} H_n e^{\theta}\frac{(-\theta)^n}{n!}} = e^{-\frac{\lambda}{\mu}\mathrm{E}[H_N]},$$

where $N \sim \text{Pois}(-\theta)$.

In addition to this remark's connection of the moment generating function and the harmonic numbers, we can also gain insight into this limiting object

through Monte Carlo methods. Using Equation 3.11, we have a simple and efficient approximate simulation method for this process through summing scaled Poisson random numbers. Furthermore, this approximation of course becomes increasingly precise as n grows. As an example of this, we give the simulated steady-state densities across different relationships of λ and μ in Figures 3.2. In addition to the interesting shapes of the densities across the different settings, one can see the limiting form of the relationships given by the recursion in Proposition 3.2.5 in these plots. We can note that one could also calculate these through a numerical inverse Laplace transform of the steady-state moment generating function in Corollary 3.2.8, although this may likely incur significantly more computational costs than the simulation procedure.

So far we have only considered exponentially distributed service. In the next subsection we will address this and extend this Poisson sum representation of the steady-state distribution to hold for general service. We do this through viewing the n-batch-size system as being composed of n sub-systems that experience single arrivals simultaneously.

3.2.3 Generalizing through Sub-System Perspectives

Because of the infinite server construction of this model, we can also interpret this system as being a network of sub-systems that also feature infinitely many servers. However, this network's mutuality is not in its services but rather in its arrivals. Specifically, in this subsection we will think of infinite server queues with batch arrivals of size n as being n infinite server queues that all receive individual arrivals simultaneously. From this perspective, one can quickly observe

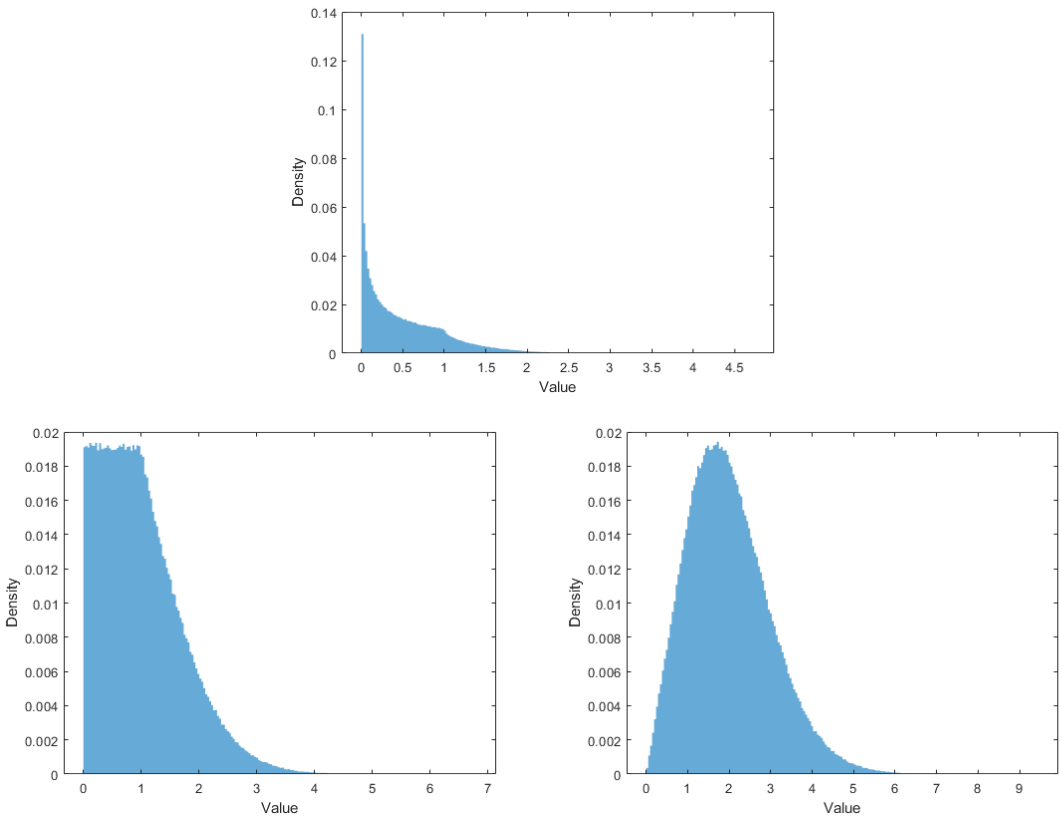


Figure 3.2: Approximate steady-state density of the scaled queue limit for size where $\frac{\lambda}{\mu} = \frac{1}{2}$ (top), $\frac{\lambda}{\mu} = 1$ (left), and $\frac{\lambda}{\mu} = 2$ (right), using 1,000,000 simulation replications and $n = 2,000$.

that marginally each subsystem will be distributed as a standard infinite server queue.

For example, if the batch system is the $M_t^n/M/\infty$ that we first considered in Subsection 3.2.1, then each of these sub-queues are $M_t/M/\infty$ systems. These sub-systems are coupled through the coincidence of their arrival times but otherwise operate independently from one another. To quantify the relationship between these systems, in Proposition 3.2.9 we derive the transient covariance between two sub-systems for a general time-varying arrival rate.

Proposition 3.2.9. *Let the batch arrival queue Q_t with batch size $n \in \mathbb{Z}^+$ be represented*

as a superposition of n infinite server single arrival queues $\{Q_{t,i} \mid 1 \leq i \leq n\}$ *that all receive arrivals simultaneously and each have independent exponentially distributed service, as described above. Let* $\lambda(t) > 0$ *be the non-stationary rate of simultaneous arrivals and let* $\mu > 0$ *be the rate of service. Then, for distinct* $i, j \in \{1, \ldots, n\}$*, the covariance between* $Q_{t,i}$ *and* $Q_{t,j}$ *is given by*

$$\mathrm{Cov}\left[Q_{t,i}, Q_{t,j}\right] = e^{-2\mu t} \int_0^t \lambda(s) e^{2\mu s} \mathrm{d}s \tag{3.16}$$

for all $t \geq 0$.

Proof. From Equation 3.1, we can solve for the product moment of the two subsystems through the ODE

$$\frac{\mathrm{d}}{\mathrm{d}t} \mathrm{E}\left[Q_{t,i} Q_{t,j}\right] = \lambda(t)\left(\mathrm{E}\left[Q_{t,i}\right] + \mathrm{E}\left[Q_{t,j}\right] + 1\right) - 2\mu \mathrm{E}\left[Q_{t,i} Q_{t,j}\right].$$

The solution to this differential equation is given by

$$\mathrm{E}\left[Q_{t,i} Q_{t,j}\right] = Q_{0,i} Q_{0,j} e^{-2\mu t} + e^{-2\mu t} \int_0^t \lambda(s)\left(\mathrm{E}\left[Q_{s,i}\right] e^{2\mu s} + \mathrm{E}\left[Q_{s,j}\right] e^{2\mu s} + e^{2\mu s}\right) \mathrm{d}s.$$

By substituting the corresponding forms of $\mathrm{E}\left[Q_{s,k}\right] = Q_{0,k} e^{-\mu s} + e^{-\mu s} \int_0^s \lambda(u) e^{\mu u} \mathrm{d}u$ in for each of the two means, we have

$$\begin{aligned}\mathrm{E}\left[Q_{t,i} Q_{t,j}\right] = Q_{0,i} Q_{0,j} e^{-2\mu t} + e^{-2\mu t} \int_0^t \lambda(s) \Bigg(& e^{2\mu s} + \left(Q_{0,i} + \int_0^s \lambda(u) e^{\mu u} \mathrm{d}u\right) e^{\mu s} \\ & + \left(Q_{0,j} + \int_0^s \lambda(u) e^{\mu u} \mathrm{d}u\right) e^{\mu s} \Bigg) \mathrm{d}s,\end{aligned}$$

and this simplifies to the following

$$\begin{aligned}\mathrm{E}\left[Q_{t,i} Q_{t,j}\right] = {} & Q_{0,i} Q_{0,j} e^{-2\mu t} + e^{-2\mu t} \int_0^t \lambda(s) e^{2\mu s} \mathrm{d}s + \left(Q_{0,i} + Q_{0,j}\right) e^{-2\mu t} \int_0^t \lambda(s) e^{\mu s} \mathrm{d}s \\ & + 2 e^{-2\mu t} \int_0^t \lambda(s) e^{\mu s} \int_0^s \lambda(u) e^{\mu u} \mathrm{d}u \mathrm{d}s.\end{aligned}$$

We can now use the fact that for a function $F : \mathbb{R}^+ \to \mathbb{R}$ defined such that $F(t) = \int_0^t f(s)\mathrm{d}s$ for a given $f(\cdot)$, integration by parts implies

$$\int_0^t f(s)F(s)\mathrm{d}s = F(t)^2 - \int_0^t F(s)f(s)\mathrm{d}s,$$

and so $\int_0^t f(s)F(s)\mathrm{d}s = \frac{F(t)^2}{2}$. This allows us to simplify to

$$\begin{aligned}\mathrm{E}\left[Q_{t,i}Q_{t,j}\right] = Q_{0,i}Q_{0,j}e^{-2\mu t} + e^{-2\mu t}\int_0^t \lambda(s)e^{2\mu s}\mathrm{d}s + \left(Q_{0,i} + Q_{0,j}\right)e^{-2\mu t}\int_0^t \lambda(s)e^{\mu s}\mathrm{d}s \\ + e^{-2\mu t}\left(\int_0^t \lambda(s)e^{\mu s}\mathrm{d}s\right)^2,\end{aligned}$$

and now we turn our focus to the product of the means. Here we distribute the multiplication to find that

$$\begin{aligned}\mathrm{E}\left[Q_{t,i}\right]\mathrm{E}\left[Q_{t,j}\right] &= \left(Q_{0,i}e^{-\mu t} + e^{-\mu t}\int_0^t \lambda(s)e^{\mu s}\mathrm{d}s\right)\left(Q_{0,j}e^{-\mu t} + e^{-\mu t}\int_0^t \lambda(s)e^{\mu s}\mathrm{d}s\right) \\ &= Q_{0,i}Q_{0,j}e^{-2\mu t} + (Q_{0,i} + Q_{0,j})e^{-2\mu t}\int_0^t \lambda(s)e^{\mu s}\mathrm{d}s + e^{-2\mu t}\left(\int_0^t \lambda(s)e^{\mu s}\mathrm{d}s\right)^2\end{aligned}$$

and by subtracting this expression from that of the product moment, we complete the proof. □

As a consequence of this, we can specify the covariance between sub-systems in the non-stationary and stationary arrival settings we have considered thus far in this report. Further, for stationary arrival rates we capitalize on simplified expressions to also give an explicit expression for the correlation coefficient between two sub-systems.

Corollary 3.2.10. *Let Q_t be an infinite server queue with arrival batch size $n \in \mathbb{Z}^+$ and exponential service rate $\mu > 0$. Further, let $Q_{t,k}$ for $k \in \{1, \ldots, n\}$ be infinite server queues with solitary arrivals and exponential service rate $\mu > 0$, so that $\sum_{k=1}^n Q_{t,k} = Q_t$ for all $t \geq 0$. Let $i, j \in \{1, \ldots, n\}$ be distinct. Then, if the arrival rate is given by*

$\lambda + \sum_{k=1}^{\infty} a_k \cos(kt) + b_k \sin(kt) > 0$, *the covariance between* $Q_{t,i}$ *and* $Q_{t,j}$ *is*

$$\mathrm{Cov}\left[Q_{t,i}, Q_{t,j}\right] = \frac{\lambda}{2\mu}\left(1 - e^{-2\mu t}\right) + \sum_{k=1}^{\infty} \frac{a_k}{k^2 + 4\mu^2}\left(2\mu\cos(kt) + k\sin(kt) - 2\mu e^{-2\mu t}\right)$$
$$+ \sum_{k=1}^{\infty} \frac{b_k}{k^2 + 4\mu^2}\left(2\mu\sin(kt) - k\cos(kt) + k e^{-2\mu t}\right), \tag{3.17}$$

and if the arrival rate is given by $\lambda > 0$, *the covariance between* $Q_{t,i}$ *and* $Q_{t,j}$ *is*

$$\mathrm{Cov}\left[Q_{t,i}, Q_{t,j}\right] = \frac{\lambda}{2\mu}\left(1 - e^{-2\mu t}\right), \tag{3.18}$$

where all $t \geq 0$. *Finally, the correlation between two sub-systems in the stationary setting can be calculated as*

$$\mathrm{Corr}[Q_{t,i}, Q_{t,j}] = \frac{\frac{\lambda}{2\mu}\left(1 - e^{-2\mu t}\right)}{\sqrt{\left(Q_{0,i}\left(e^{-\mu t} - e^{-2\mu t}\right) + \frac{\lambda}{\mu}\left(1 - e^{-\mu t}\right)\right)\left(Q_{0,j}\left(e^{-\mu t} - e^{-2\mu t}\right) + \frac{\lambda}{\mu}\left(1 - e^{-\mu t}\right)\right)}},$$

hence for stationary arrival rates, $\mathrm{Corr}[Q_{t,i}, Q_{t,j}] \to \frac{1}{2}$ *as* $t \to \infty$.

Thus, we find that for a fully Markovian batch arrival queue with stationary arrival rate the correlation among any two sub-systems in steady-state is $\frac{1}{2}$, regardless of the arrival or service parameters. In some sense this seems to capture a balance between the effect of arrivals and of services on an infinite server system, with the latter being independent between these systems and the former being perfectly correlated.

Now, we can pause to note that we have actually made an implicit modeling choice by separating the batch into n identical sub-systems. In this set-up we have decided to route all customers within one batch equivalently, but we are free to make other routing decisions and still maintain the n sub-systems construction. With that in mind, it seems natural to wonder if we can uncover distributional structure of the full system if we choose our routing procedure

carefully. We will now find that not only is this true, but we in fact already have already seen a suggestion on what type of routing to consider.

From Proposition 3.2.4, we have seen that the steady-state distribution of the $M^n/M/\infty$ system is equivalent to that of $\sum_{j=1}^{n} jY_j$ where $Y_j \sim \text{Pois}(\frac{\lambda}{j\mu})$ are independent. We can also note that just as the minimum of the independent sample $S_1, \ldots, S_n \sim \text{Exp}(\mu)$ will be exponentially distributed with rate $n\mu$, for $S_{(i)}$ as the ith ordered statistic of the n-sample we have that $S_{(i)} - S_{(i-1)} \sim \text{Exp}((n - i + 1)\mu)$. Of course, the sum of these differences will telescope so that $\sum_{j=1}^{i} S_{(j)} - S_{(j-1)} = S_{(i)}$.

Taking this as inspiration, we will now assume that upon the arrival of a batch we can now know the duration of each customer's service. We then take the sub-queues to be such that the first sub-system always receives the service with the shortest duration, the second sub-system receives the second shortest service, and so on. Thus, we will route each batch of customers according to the order statistics within each batch. For reference, we visualize this sub-system construction in Figure 3.3.

We can note that while the covariance structure we explored in Proposition 3.2.9 and Corollary 3.2.10 do not apply for this new routing, the sub-systems are certainly still correlated. Due to the order-statistics structuring of the service in each queue, we can note that now both the arrival processes and the service distributions will be dependent. However, we can in fact use our understanding of this dependence to not only understand how these systems relate to one another, but also to interpret how they form the structure of the full batch system as a whole. In this way, we will now consider a $M^n/G/\infty$ system. As follows in Theorem 3.2.11, we will find that the order-statistics-routing inspiration we

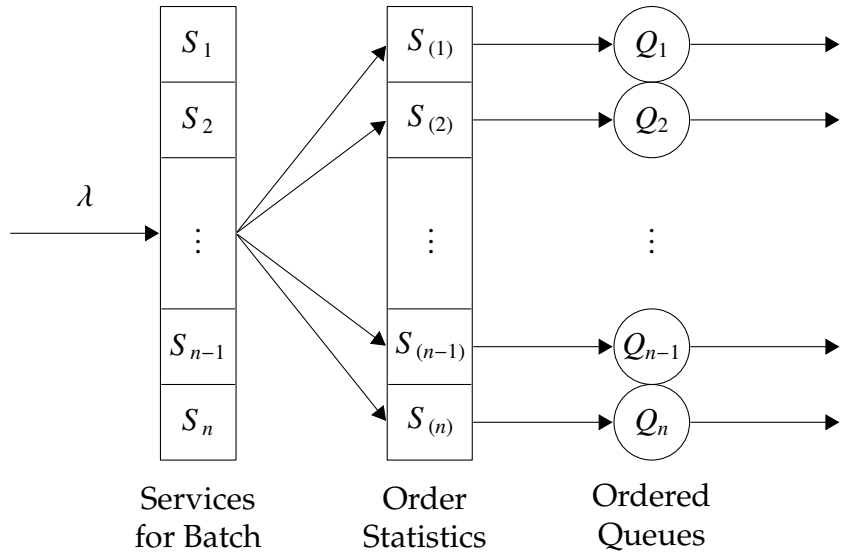


Figure 3.3: Queueing diagram for the batch arrival queue with infinite servers, in which the arriving entities are routed according to the ordering of their service durations.

have used from Proposition 3.2.4 leads us to a generalized Poisson sum result for general service distributions.

Theorem 3.2.11. *Let $Q_t(n)$ be an $M^n/G/\infty$ queue. That is, let $Q_t(n)$ be an infinite server queue with stationary arrival rate $\lambda > 0$, arrival batch size $n \in \mathbb{Z}^+$, and general service distribution G. Then, the steady-state distribution of the number in system $Q_\infty(n)$ is*

$$Q_\infty(n) \stackrel{D}{=} \sum_{j=1}^{n} (n - j + 1) Y_j \tag{3.19}$$

where $Y_j \sim \text{Pois}\left(\lambda \mathrm{E}\left[S_{(j)} - S_{(j-1)}\right]\right)$ are independent, with $S_{(1)} \leq \cdots \leq S_{(n)}$ as order statistics of the distribution G and with $S_{(0)} = 0$.

Proof. As we have discussed in the paragraphs preceding this statement, we will consider the full queueing system as being composed of n infinite server sub-systems to which we route the arriving customers in each batch. That is, let $Q_1, \ldots, Q_n$ be infinite server queues of which we will consider the steady-state

behavior. Upon the arrival of a batch, we order the customers according to the duration of their service. Then, we send the customer with the earliest service completion to Q_1, the customer with the second earliest to Q_2, and so on.

When viewing each sub-system on its own, we see that Q_j is an infinite server queue with single arrivals according to a Poisson process with rate λ and service distribution matching that of $S_{(j)}$, the jth order statistics of G. Thus, we can see that in steady-state $Q_j \sim \text{Pois}\left(\lambda \text{E}\left[S_{(j)}\right]\right)$ through the literature for $M/G/\infty$ queues, such as in Eick et al. (1993). While we can further observe that $Q_\infty(n) = \sum_{j=1}^{n} Q_j$, we must take care in re-assembling the sub-queues. In particular, we can note that $S_{(j)}$ shares a similar structure with $S_{(j-1)}$. Each order statistic can be viewed as a construction of the gaps between the lower ordered quantities:

$$S_{(j)} = \sum_{k=1}^{j} S_{(k)} - S_{(k-1)}.$$

Thus, from the thinning property of the Poisson distribution and the linearity of expectation, we can write the distribution of Q_j as a sum of independent Poisson RV's, as given by

$$Q_j \sim \sum_{k=1}^{j} \text{Pois}\left(\lambda \text{E}\left[S_{(k)} - S_{(k-1)}\right]\right).$$

We can note further that $j-1$ of the Poisson components of Q_j are the exact components of Q_{j-1}, with $j-2$ of these components also shared with Q_{j-2}, $j-3$ with Q_{j-3}, and so on. Then, we see that the Poisson component $\text{Pois}\left(\lambda \text{E}\left[S_{(j)} - S_{(j-1)}\right]\right)$ is repeated $n - j + 1$ times across this sub-system construction of $Q_\infty(n)$, as it appears in each of the Poisson sum expressions of Q_j, Q_{j+1}, $\ldots$, Q_{n-1}, and Q_n. Assembling $Q_\infty(n)$ in this way, we complete the proof. □

One can also note that this order statistic sub-system structure also provides some motivation for the occurrence of the harmonic numbers that we observed

in Subsection 3.2.2 when viewing the largest order statistic, which we discuss now in the following remark.

Remark. For $S_i \sim \text{Exp}(\mu)$, one can see through the telescoping construction of the order statistics that

$$\mathrm{E}\left[S_{(n)}\right] = \sum_{i=1}^{n} \mathrm{E}\left[S_{(i)} - S_{(i-1)}\right] = \sum_{i=1}^{n} \frac{1}{(n-i+1)\mu} = \frac{1}{\mu} H_n.$$

Now, throughout this section we have operated on the assumption that the batch size is a known, fixed constant. While this may be applicable in some settings there are certainly many settings where the batch size is unknown and varies between arrivals. Thus, we address this in Section 3.3 and find that many of the results we have shown thus far can be replicated for models with random batch size.

3.3 Random Batch Sizes

We will now consider systems in which the size of an arriving batch is drawn from an independent and identically distributed sequence of random variables. We will treat the distribution of the batch size as general throughout this work. As in Section 3.2, we assume that the times of arrivals are given by a Poisson process, with consideration given to both stationary and non-stationary rates, and we will again analyze both exponential and general service distributions.

We start by giving the mean and variance of the system for time-varying arrival rates with exponential service in Subsection 3.3.1. Then, in Subsection 3.3.2 we give three limiting results for the stationary arrivals model: a batch scaling, a fluid limit, and a diffusion limit. Finally in Subsection 3.3.3 we extend the Pois-

son sum construction of the steady-state distribution to hold for random batch sizes.

One can note that many of these results are generalizations or extensions of findings from Section 3.2, thus implying them as a special case and perhaps even building a case for them to be omitted. Rather, these findings are critical to the narrative of this report. As we will see, the results for fixed batch size provide the analytic foundations and conceptual inspirations from which we derive much of the analysis in this section.

3.3.1 Mean and Variance for Time-Varying, Markovian Case

To begin our exploration into random batch size systems, we'll start simple: we'll look at a fully Markovian (albeit time-varying) system and find the mean and variance, using conditional probability and our results from Section 3.2. Specifically, in this subsection we will consider the $M_t^N/M/\infty$ queue. That is, take an infinite server queue with a general non-stationary arrival rate. We suppose that arrivals occur in batches of random size from a sequence of independent and identically distributed random variables. Furthermore, we suppose that service is exponentially distributed. We now give the mean and variance of this system in Proposition 3.3.1.

Proposition 3.3.1. *Let Q_t be an infinite server queue with finite, time-varying arrival rate $\lambda(t) > 0$, exponential service rate $\mu > 0$, and random batch size with finite mean, $\mathrm{E}[N]$. Then, the mean number in system is given by*

$$\mathrm{E}[Q_t] = Q_0 e^{-\mu t} + e^{-\mu t}\mathrm{E}[N] \int_0^t \lambda(s) e^{\mu s} \mathrm{d}s, \tag{3.20}$$

for all $t \geq 0$. Then, if the batch size distribution has finite second moment $\mathrm{E}\left[N^2\right]$, the variance of the number in system is given by

$$\begin{aligned}\mathrm{Var}\left(Q_t\right) = Q_0\left(e^{-\mu t} - e^{-2\mu t}\right) + e^{-2\mu t}\left(\mathrm{E}\left[N^2\right] - \mathrm{E}\left[N\right]\right)\int_0^t \lambda(s)e^{2\mu s}\mathrm{d}s \\ + e^{-\mu t}\mathrm{E}\left[N\right]\int_0^t \lambda(s)e^{\mu s}\mathrm{d}s,\end{aligned} \tag{3.21}$$

again for all $t \geq 0$.

Proof. Using the infinitesimal generator method, we have that the first and second moments of this system are given by the solutions to

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[Q_t\right] &= \lambda(t)\mathrm{E}\left[N_1\right] - \mu\mathrm{E}\left[Q_t\right], \\ \frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[Q_t^2\right] &= \lambda(t)\left(2\mathrm{E}\left[Q_t\right]\mathrm{E}\left[N_1\right] + \mathrm{E}\left[N_1^2\right]\right) - 2\mu\mathrm{E}\left[Q_t^2\right] + \mu\mathrm{E}\left[Q_t\right],\end{aligned}$$

where $\{N_i \mid i \in \mathbb{Z}^+\}$ are the i.i.d. batch sizes that are also independent of the queue. Through noting that

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{Var}\left(Q_t\right) = \frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[Q_t^2\right] - 2\mathrm{E}\left[Q_t\right]\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[Q_t\right] = \lambda(t)\mathrm{E}\left[N_1^2\right] + \mu\mathrm{E}\left[Q_t\right] - 2\mu\mathrm{Var}\left(Q_t\right),$$

we can solve for the stated results. □

In addition to providing a direct comparison to the fixed batch size case in conjunction with Corollary 3.2.3, Proposition 3.3.1 also provides a building block for the remainder of this section. In particular, in the following subsection we will develop a series of limiting results for this queueing system, including fluid and diffusion limits. In those cases, we will use this result for added interpretation. To expedite comparison in cases of stationary arrival rates, we now give the mean and variance for such systems in Corollary 3.3.2. Additionally, to also facilitate comparison to Corollary 3.2.3, we provide expressions for periodic arrival rates in Corollary 3.3.3.

Corollary 3.3.2. *Let Q_t be an infinite server queue with stationary arrival rate $\lambda > 0$, exponential service rate $\mu > 0$, and random batch size with mean $\mathrm{E}[N]$. Then, the mean number in system is given by*

$$\mathrm{E}[Q_t] = Q_0 e^{-\mu t} + \frac{\lambda \mathrm{E}[N]}{\mu}\left(1 - e^{-\mu t}\right), \tag{3.22}$$

for all $t \geq 0$. Then, if the batch size distribution has finite second moment $\mathrm{E}\left[N^2\right]$, the variance of the number in system is given by

$$\mathrm{Var}(Q_t) = Q_0\left(e^{-\mu t} - e^{-2\mu t}\right) + \frac{\lambda \mathrm{E}[N]}{\mu}\left(1 - e^{-\mu t}\right) + \frac{\lambda}{2\mu}\left(\mathrm{E}\left[N^2\right] - \mathrm{E}[N]\right)\left(1 - e^{-2\mu t}\right), \tag{3.23}$$

again for all $t \geq 0$.

Corollary 3.3.3. *Let Q_t be an infinite server queue with periodic arrival rate $\lambda + \sum_{k=1}^{\infty} a_k \cos(kt) + b_k \sin(kt) > 0$, exponential service rate $\mu > 0$, and random batch size with finite mean, $\mathrm{E}[N]$. Then, the mean number in system is given by*

$$\begin{aligned}\mathrm{E}[Q_t] = Q_0 e^{-\mu t} &+ \frac{\lambda \mathrm{E}[N]}{\mu}\left(1 - e^{-\mu t}\right) + \sum_{k=1}^{\infty} \frac{\mathrm{E}[N](a_k\mu - b_k k)}{k^2 + \mu^2}\left(\cos(kt) - e^{-\mu t}\right) \\ &+ \sum_{k=1}^{\infty} \frac{\mathrm{E}[N](a_k k + b_k\mu)}{k^2 + \mu^2}\sin(kt),\end{aligned} \tag{3.24}$$

for all $t \geq 0$. Then, if the batch size distribution has finite second moment $\mathrm{E}\left[N^2\right]$, the variance of the number in system is given by

$$\begin{aligned}\mathrm{Var}(Q_t) = Q_0\left(e^{-\mu t} - e^{-2\mu t}\right) &+ \frac{\lambda \mathrm{E}[N]}{\mu}\left(1 - e^{-\mu t}\right) + \sum_{k=1}^{\infty} \frac{EN(a_k\mu - b_k k)}{k^2 + \mu^2}\left(\cos(kt) - e^{-\mu t}\right) \\ &+ \sum_{k=1}^{\infty} \frac{\mathrm{E}[N](a_k k + b_k\mu)}{k^2 + \mu^2}\sin(kt) + \frac{\lambda}{2\mu}\left(\mathrm{E}\left[N^2\right] - \mathrm{E}[N]\right)\left(1 - e^{-2\mu t}\right) \\ &+ \left(\mathrm{E}\left[N^2\right] - \mathrm{E}[N]\right)\left(\sum_{k=1}^{\infty} \frac{2a_k\mu - b_k k}{k^2 + 4\mu^2}\left(\cos(kt) - e^{-2\mu t}\right) + \sum_{k=1}^{\infty} \frac{a_k k + 2b_k\mu}{k^2 + 4\mu^2}\sin(kt)\right),\end{aligned} \tag{3.25}$$

again for all $t \geq 0$.

3.3.2 Limiting Results for Stationary Arrival Rates

We will now focus on systems with stationary arrival rates throughout the analysis in this subsection. In doing so, we derive limit theorems for various scalings of this process. To begin, we show a brief technical lemma for the limit of non-negative random variables that can be represented as sums of independent and identically distributed random variables.

Lemma 3.3.4. *Let $X(n)$ be any random variable that $X(n) = \sum_{k=1}^{n} Y_k$ where Y_k are i.i.d. non-negative, discrete random variables. Then, the moment generating function of $X(n)$ is such that*

$$\mathrm{E}\left[e^{\frac{\theta X(n)}{n}}\right] \to e^{\mathrm{E}[Y_1]\theta}$$

as $n \to \infty$.

Proof. By the strong law of large numbers, we have that

$$\lim_{n\to\infty} \frac{X(n)}{n} = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} Y_k \overset{\text{a.s.}}{=} \mathrm{E}\left[Y_1\right],$$

and this implies convergence in distribution, which is equivalent to convergence of moment generating functions. □

We can note that this condition is a weaker form of infinite divisibility. Thus, in addition to holding for any infinitely divisible and non-negative random variables such as the Poisson, and negative binomial distributions, Lemma 3.3.4 also holds for some distributions that are not infinitely divisible, such as the binomial. Using this lemma we can now find our first limit theorem for random batch sizes, a batch scaling result akin to Proposition 3.2.7.

Theorem 3.3.5. *For $n \in \mathbb{Z}^+$, let $Q_t(n)$ be an infinite server queue with batch arrivals where the batch size is drawn from the i.i.d. sequence $\{N_i(n) \mid i \in \mathbb{Z}^+\}$. Let $\lambda > 0$ be the arrival rate and let $\mu > 0$ be the rate of exponentially distributed service. Then, suppose that for any i and n there is a sequence of i.i.d. non-negative, discrete random variables $\{B_k \mid k \in \mathbb{Z}^+\}$ such that $N_i(n) = \sum_{k=1}^{n} B_k$. Then, the limiting moment generating function of the batch scaled object*

$$\lim_{n\to\infty} \mathrm{E}\left[e^{\frac{\theta}{n}Q_t(n)}\right] = \begin{cases} e^{\frac{\lambda}{\mu}\left(\mathrm{Ei}(\theta\mathrm{E}[B_1])-\mathrm{Ei}\left(\theta\mathrm{E}[B_1]e^{-\mu t}\right)-\mu t\right)} & \text{if } \theta > 0, \\ e^{\frac{\lambda}{\mu}\left(E_1\left(-\theta\mathrm{E}[B_1]e^{-\mu t}\right)-E_1(-\theta\mathrm{E}[B_1])-\mu t\right)} & \text{if } \theta < 0, \\ 1 & \text{if } \theta = 0, \end{cases} \tag{3.26}$$

for all $t \geq 0$.

Proof. Because this system is Markovian, we can calculate the time derivative of the moment generating function for a given n as

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[e^{\frac{\theta}{n}Q_t(n)}\right] &= \mathrm{E}\left[\lambda\left(e^{\frac{\theta}{n}N_1(n)} - 1\right)e^{\frac{\theta}{n}Q_t(n)} + \mu Q_t(n)\left(e^{-\frac{\theta}{n}} - 1\right)e^{\frac{\theta}{n}Q_t(n)}\right] \\ &= \lambda\left(\mathrm{E}\left[e^{\frac{\theta}{n}N_1(n)}\right] - 1\right)\mathrm{E}\left[e^{\frac{\theta}{n}Q_t(n)}\right] + n\mu\left(e^{-\frac{\theta}{n}} - 1\right)\mathrm{E}\left[\frac{Q_t(n)}{n}e^{\frac{\theta}{n}Q_t(n)}\right]. \end{aligned}$$

This can then be re-expressed in partial differential equation form as

$$\frac{\partial \mathcal{M}^n(\theta,t)}{\partial t} = \lambda\left(\mathrm{E}\left[e^{\frac{\theta}{n}N_1(n)}\right] - 1\right)\mathcal{M}^n(\theta,t) + n\mu\left(e^{-\frac{\theta}{n}} - 1\right)\frac{\partial \mathcal{M}^n(\theta,t)}{\partial \theta},$$

where $\mathcal{M}^n(\theta,t) = \mathrm{E}\left[e^{\frac{\theta}{n}Q_t(n)}\right]$. Now, through Lemma 3.3.4, we see that the limit of this partial differential equation is given by

$$\frac{\partial \mathcal{M}^\infty(\theta,t)}{\partial t} = \lambda\left(e^{\theta\mathrm{E}[B_1]} - 1\right)\mathcal{M}^\infty(\theta,t) - \mu\theta\frac{\partial \mathcal{M}^\infty(\theta,t)}{\partial \theta}.$$

We achieve the stated result through a straightforward update of the method of characteristics approach in Proposition 3.2.7. □

We can note that a similar batch scaling of infinite server queues is discussed in de Graaf et al. (2017), in which the authors show that the limiting process can be interpreted as a shot noise process. However, that work considers a different class of batch size distributions, as the authors define their batch size distribution in terms of the distribution of the marks through use of a ceiling rounding function. In this way, that paper is more oriented around the distribution of the marks in the shot noise process rather than the size of the batches.

From this result, we can identify a relationship between the moment generating functions of the deterministic and random batch size queues under batch scalings. Let $\mathcal{M}_n^\infty(\theta, t)$ be the limiting moment generating function of the fixed batch size queue as given in Proposition 3.2.7 and let $\mathcal{M}_N^\infty(\theta, t)$ be the same for the random batch size queue as we have now seen in Theorem 3.3.5. Then, we can observe that

$$\mathcal{M}_N^\infty(\theta, t) = \mathcal{M}_n^\infty(\theta \mathrm{E}\left[B_1\right], t),$$

whenever the distribution of the random batch sizes meets the "finite divisibility" condition as described in Lemma 3.3.4. The relationship between these limiting objects provides a direct comparison between the two different batch types.

As two additional limiting results, we now provide fluid and diffusion limits for scaling the arrival rate in Theorems 3.3.6 and 3.3.7, respectively. We did not give fluid or diffusion limits for the deterministic batch cases in Section 3.2, so these two limits are built from scratch within this section. Although we did not develop such limits explicitly for the $M^n/M/\infty$ system, we will find that these limits can still be used to draw comparisons between this system and the $M^N/M/\infty$ queue simply by treating the random batch size as deterministically

distributed. We now begin with the fluid limit.

Theorem 3.3.6. *For $n \in \mathbb{Z}^+$, let $Q_t(n)$ be an infinite server queue with batch arrivals where the batch size is drawn from the i.i.d. sequence $\{N_i \mid i \in \mathbb{Z}^+\}$. Let $n\lambda > 0$ be the arrival rate and let $\mu > 0$ be the rate of exponentially distributed service. Then, the limiting moment generating function of the fluid scaling is given by*

$$\lim_{n\to\infty} \mathrm{E}\left[e^{\frac{\theta}{n}Q_t(n)}\right] = e^{\frac{\lambda \mathrm{E}[N_1]\theta}{\mu}\left(1-e^{-\mu t}\right)+Q_0\theta e^{-\mu t}}, \tag{3.27}$$

for all $t \geq 0$.

Proof. We begin with the infinitesimal generator equation for the time derivative of the moment generating function at a given n. This is

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[e^{\frac{\theta}{n}Q_t(n)}\right] &= \mathrm{E}\left[n\lambda\left(e^{\frac{\theta N_1}{n}}-1\right)e^{\frac{\theta}{n}Q_t(n)} + \mu Q_t(n)\left(e^{-\frac{\theta}{n}}-1\right)e^{\frac{\theta}{n}Q_t(n)}\right] \\ &= n\lambda\left(\mathrm{E}\left[e^{\frac{\theta N_1}{n}}\right]-1\right)\mathrm{E}\left[e^{\frac{\theta}{n}Q_t(n)}\right] + \mu n\left(e^{-\frac{\theta}{n}}-1\right)\mathrm{E}\left[\frac{Q_t(n)}{n}e^{\frac{\theta}{n}Q_t(n)}\right],\end{aligned}$$

which can also be expressed in partial differential equation form as

$$\frac{\partial \mathcal{M}^n(\theta,t)}{\partial t} = n\lambda\left(\mathrm{E}\left[e^{\frac{\theta N_1}{n}}\right]-1\right)\mathcal{M}^n(\theta,t) + \mu n\left(e^{-\frac{\theta}{n}}-1\right)\frac{\partial \mathcal{M}^n(\theta,t)}{\partial \theta},$$

where $M^n(\theta,t) = \mathrm{E}\left[e^{\frac{\theta}{n}Q_t(n)}\right]$. By a Taylor expansion of the function $e^{\frac{\theta N_1}{n}}$ and by taking the limit as $n \to \infty$, we can see that this yields

$$\frac{\partial \mathcal{M}^\infty(\theta,t)}{\partial t} = \lambda\theta \mathrm{E}\left[N_1\right]\mathcal{M}^\infty(\theta,t) - \mu\theta\frac{\partial \mathcal{M}^\infty(\theta,t)}{\partial \theta}.$$

Using the initial condition $\mathcal{M}^\infty(\theta,0) = e^{Q_0\theta}$, we can see that the solution to this partial differential equation will be

$$\mathcal{M}^\infty(\theta,t) = e^{\frac{\lambda \mathrm{E}[N_1]\theta}{\mu}\left(1-e^{-\mu t}\right)+Q_0\theta e^{-\mu t}},$$

and this completes the proof. □

From Corollary 3.3.2, we see that the mean number in system for the $M^N/M/\infty$ queue is $\frac{\lambda \mathrm{E}[N_1]}{\mu}(1-e^{-\mu t}) + Q_0 e^{-\mu t}$. Thus, this fluid limit moment generating function is equivalent to $e^{\theta \mathrm{E}[Q_t]}$ for all $t \geq 0$ and all θ, showing that the fluid limit converges to the mean. We now find a connection to both the mean and the variance through a diffusion limit in Theorem 3.3.7.

Theorem 3.3.7. *For $n \in \mathbb{Z}^+$, let $Q_t(n)$ be an infinite server queue with batch arrivals where the batch size is drawn from the i.i.d. sequence $\{N_i \mid i \in \mathbb{Z}^+\}$. Let $n\lambda > 0$ be the arrival rate and let $\mu > 0$ be the rate of exponentially distributed service. Then, the limiting moment generating function of the diffusion scaling is given by*

$$\lim_{n\to\infty} \mathrm{E}\left[e^{\frac{\theta}{\sqrt{n}}\left(Q_t(n) - \frac{n\lambda \mathrm{E}[N_1]}{\mu}\right)}\right] = e^{\frac{\lambda\theta^2}{4\mu}\left(\mathrm{E}[N_1]+\mathrm{E}\left[N_1^2\right]\right)\left(1-e^{-\mu t}\right)+\theta Q_0 e^{-\mu t}} \tag{3.28}$$

which gives a steady-state approximation of $X \sim \mathrm{Norm}\left(\frac{\lambda \mathrm{E}[N_1]}{\mu}, \frac{\lambda}{2\mu}\left(\mathrm{E}\left[N_1\right] + \mathrm{E}\left[N_1^2\right]\right)\right)$.

Proof. Through use of the infinitesimal generator, we have that the time derivative of the moment generating function for a given n can be expressed

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[e^{\frac{\theta}{\sqrt{n}}\left(Q_t(n) - \frac{n\lambda \mathrm{E}[N_1]}{\mu}\right)}\right] \\
&= \mathrm{E}\left[n\lambda\left(e^{\frac{\theta N_1}{\sqrt{n}}} - 1\right)e^{\frac{\theta}{\sqrt{n}}\left(Q_t(n) - \frac{n\lambda \mathrm{E}[N_1]}{\mu}\right)} + \mu Q_t(n)\left(e^{-\frac{\theta}{\sqrt{n}}} - 1\right)e^{\frac{\theta}{\sqrt{n}}\left(Q_t(n) - \frac{n\lambda \mathrm{E}[N_1]}{\mu}\right)}\right] \\
&= \mathrm{E}\left[\sqrt{n}\lambda\left(\theta N_1 + \frac{\theta^2 N_1^2}{2\sqrt{n}} + \mathrm{O}\left(\frac{\theta^3 N_1^3}{6n}\right)\right)e^{\frac{\theta}{\sqrt{n}}\left(Q_t(n) - \frac{n\lambda \mathrm{E}[N_1]}{\mu}\right)}\right] \\
&\quad + \mathrm{E}\left[\mu\sqrt{n}\left(\frac{Q_t(n)}{\sqrt{n}} - \frac{n\lambda \mathrm{E}\left[N_1\right]}{\sqrt{n}\mu} + \frac{n\lambda \mathrm{E}\left[N_1\right]}{\sqrt{n}\mu}\right)\left(e^{-\frac{\theta}{\sqrt{n}}} - 1\right)e^{\frac{\theta}{\sqrt{n}}\left(Q_t(n) - \frac{n\lambda \mathrm{E}[N_1]}{\mu}\right)}\right],
\end{aligned}$$

where here we have used a Taylor expansion of the function $e^{\frac{\theta N_1}{\sqrt{n}}}$. Now, for $\mathcal{M}^n(\theta, t) = \mathrm{E}\left[e^{\frac{\theta}{\sqrt{n}}\left(Q_t(n) - \frac{n\lambda \mathrm{E}[N_1]}{\mu}\right)}\right]$, this equation can be written as a partial differential equation as follows:

$$\begin{aligned}
\frac{\partial \mathcal{M}^n(\theta, t)}{\partial t} &= \lambda\theta\sqrt{n}\mathrm{E}\left[N_1\right]\mathcal{M}^n(\theta, t) + \frac{\lambda\theta^2}{2}\mathrm{E}\left[N_1^2\right]\mathcal{M}^n(\theta, t) + \sqrt{n}\lambda\mathrm{E}\left[\mathrm{O}\left(\frac{\theta^3 N_1^3}{6n}\right)e^{\frac{\theta}{\sqrt{n}}\left(Q_t(n) - \frac{n\lambda \mathrm{E}[N_1]}{\mu}\right)}\right] \\
&\quad + \sqrt{n}\mu\left(e^{-\frac{\theta}{\sqrt{n}}} - 1\right)\frac{\partial \mathcal{M}^n(\theta, t)}{\partial \theta} + n\lambda\mathrm{E}\left[N_1\right]\left(e^{-\frac{\theta}{\sqrt{n}}} - 1\right)\mathcal{M}^n(\theta, t).
\end{aligned}$$

As we take $n \to \infty$ this PDE becomes

$$\frac{\partial \mathcal{M}^{\infty}(\theta,t)}{\partial t} = \frac{\lambda\theta^2}{2}\mathrm{E}\left[N_1\right]\mathcal{M}^{\infty}(\theta,t) + \frac{\lambda\theta^2}{2}\mathrm{E}\left[N_1^2\right]\mathcal{M}^{\infty}(\theta,t) - \mu\theta\frac{\partial \mathcal{M}^{\infty}(\theta,t)}{\partial \theta},$$

and this yields a solution of

$$\mathcal{M}^{\infty}(\theta,t) = e^{\frac{\lambda\theta^2}{4\mu}\left(\mathrm{E}[N_1]+\mathrm{E}\left[N_1^2\right]\right)\left(1-e^{-\mu t}\right)+\theta Q_0 e^{-\mu t}}.$$

To observe the steady-state distribution, we take the limit as $t \to \infty$ and observe that this produces the moment generating function for a Gaussian. □

By comparison to the limits of the expresions in Corollary 3.3.2 as $t \to \infty$, we can now observe that this steady-state approximation is equal in mean and variance to the steady-state queue.

3.3.3 Extending the Order Statistics Sub-Systems

In Subsection 3.2.3 we found that the steady-state distribution of infinite server queues with fixed batch size and general service can be written as a sum of scaled Poisson random variables, providing a succinct interpretation of the process and an efficient simulation procedure for approximate calculations. The underlying observation that supported this approach was that we can think of an infinite server queue with batch arrivals as a collection of infinite server queues with solitary arrivals that occur simultaneously. Using the thinning property of Poisson processes, we now extend this result to queues with random batch sizes and general service.

Theorem 3.3.8. *Let Q_t be a $M^N/G/\infty$ queue. That is, let Q_t an infinite server queue with stationary arrival rate $\lambda > 0$, arrival batch of random size according to the i.i.d. sequence of non-negative integer valued random variables $\{N_i \mid i \in \mathbb{Z}^+\}$, and general*

service distribution G. Then, the steady-state distribution of the number in system Q_∞ is

$$Q_\infty \stackrel{D}{=} \sum_{n=1}^{\infty} \sum_{j=1}^{n} (n-j+1) Y_{j,n} \tag{3.29}$$

where $Y_{j,n} \sim \text{Pois}\left(\lambda p_n \text{E}\left[S_{(j,n)} - S_{(j-1,n)}\right]\right)$ *are independent, with* $S_{(1,n)} \leq \cdots \leq S_{(n,n)}$ *as order statistics of the distribution G when* $N_i = n$*, where* $S_{(0,n)} = 0$ *for all n and* $p_n = \text{P}(N_1 = n)$*.*

Proof. To begin, we suppose that there is some $m \in \mathbb{Z}^+$ such that $\text{P}(N_i \in \{0, \ldots, m\}) = 1$. Then, using the thinning property of Poisson processes, we separate the arrival process into m arrival streams where the nth arrival rate is λp_n. Then, by Theorem 3.2.11 the steady-state distribution of the number in system from the nth stream is

$$\sum_{j=1}^{n} (n-j+1) \text{Pois}\left(\lambda p_n \text{E}\left[S_{(j,n)} - S_{(j-1,n)}\right]\right).$$

Then, since the m thinned Poisson streams are independent, we have that the full combined system will be distributed as

$$\sum_{n=1}^{m} \sum_{j=1}^{n} (n-j+1) \text{Pois}\left(\lambda p_n \text{E}\left[S_{(j,n)} - S_{(j-1,n)}\right]\right).$$

Through taking the limit as $m \to \infty$, we achieve the stated result. □

We can note that Theorem 3.3.8 also provides a method for approximate empirical calculation through simulation. This representation can also be simplified if more information is known about the distribution of the batch size or of the service, or both. As an example, we give the distribution for the fully Markovian system in the following corollary.

Corollary 3.3.9. *Let Q_t be a $M^N/M/\infty$ queue. That is, let Q_t an infinite server queue with stationary arrival rate $\lambda > 0$, arrival batch of random size according to the i.i.d. sequence of non-negative integer valued random variables $\{N_i \mid i \in \mathbb{Z}^+\}$, and exponentially distributed service at rate $\mu > 0$. Then, the steady-state distribution of the number in system Q_∞ is*

$$Q_\infty \overset{D}{=} \sum_{j=1}^{\infty} jY_j \tag{3.30}$$

where $Y_j \sim \text{Pois}\left(\frac{\lambda}{j\mu}\bar{F}_N(j)\right)$ are independent, where $\bar{F}_N(j) = \mathrm{P}(N_1 \geq j)$.

One can note that the moment generating function for this system in steady-state is

$$\mathrm{E}\left[e^{\theta Q_\infty}\right] = e^{\sum_{j=1}^{\infty} \frac{\lambda}{j\mu}\bar{F}_N(j)\left(e^{j\theta}-1\right)},$$

and that this also admits a connection to the generalized Hermite distributions we discussed in Subsection 3.2.2. In particular, this generalized Hermite distribution can be characterized by $\frac{\lambda}{\mu}$, which is again the mean of the distribution, and the complementary cumulative distribution function of the batch size distribution, which dictates the coefficients at each j. For this reason, it may be possible that the steady-state distribution of the queue may be simplified even further for particular batch size distributions.

Because Theorem 3.3.8 is again built upon an order statistics sub-queue perspective, it is natural to wonder how the distribution of the batch size would affect those sub-systems. In particular, we now consider the following scenario: suppose that the batch size is bounded by some constant, say k, and that we have k sub-systems. For each arriving batch, the customer with the shortest service duration will go to the first sub-system, the second shortest to the second sub-system, and so on, but only up to the number that have just arrived: if this

batch is of size $k-1$, the kth sub-queue will not receive an arrival. In this way, the ith sub-queue represents the number in system that were the ith smallest in their batch. In the following proposition we find the conditions on the batch size distribution under which the distributions of the sub-queues will be equivalent.

Proposition 3.3.10. *Consider a $M^B/G/\infty$ queueing system in which the distribution of B has support on $\{1,\ldots,k\}$. Let $\phi \in [0,1]^{k-1}$ be such that $\phi_i = \mathrm{P}(B=i)$, yielding $\mathrm{P}(B=k) = 1-\sum_{i=1}^{k-1}\phi_i$. Let $S_{(i,j)}$ be the ith order statistics in a sample of size j from the service distribution. Furthermore, let Q_i be steady-state number in system of an infinite server sub-queue to which the customer with the ith smallest service duration in an arriving batch will be routed whenever there are at least i customers in the batch. Let $M \in \mathbb{R}^{k-1\times k-1}$ be an upper triangular matrix such that*

$$M_{i,j} = \frac{\mathrm{E}\left[S_{(i,j)}\right]}{\mathrm{E}\left[S_{(k,k)}\right]-\mathrm{E}\left[S_{(i,k)}\right]},$$

for $i \le j$ and $M_{i,j}=0$ otherwise. For $\mathbf{v} \in \mathbb{R}^{k-1}$ as the all-ones column vector, if ϕ is such that

$$\mathbf{v} = \left(M + \mathbf{v}\mathbf{v}^{\mathrm{T}}\right)\phi,$$

then $Q_i \overset{D}{=} Q_j$ for all sub-queues i and j. Moreover, if $1+\mathbf{v}^{\mathrm{T}}M^{-1}\mathbf{v} \neq 0$, then the distributions of the sub-queues are equivalent if and only if $\phi = (M+\mathbf{v}\mathbf{v}^{\mathrm{T}})^{-1}\mathbf{v}$.

Proof. We start by considering the mean of each queue and solving for ϕ such that all the means are equal. Let λ be the batch arrival rate. Then, the mean of Q_i is

$$\mathrm{E}\left[Q_i\right] = \sum_{j=i}^{k-1}\lambda\phi_j\mathrm{E}\left[S_{(i,j)}\right] + \lambda\left(1-\sum_{j=1}^{k-1}\phi_j\right)\mathrm{E}\left[S_{(i,k)}\right],$$

as entities only arrive to Q_i when $B \ge i$. We can note that for Q_k this is

$$\mathrm{E}\left[Q_k\right] = \lambda\left(1-\sum_{j=1}^{k-1}\phi_j\right)\mathrm{E}\left[S_{(k,k)}\right].$$

Then, we can see that all the queue means will be equal if $\mathrm{E}[Q_i] = \mathrm{E}[Q_k]$ for all i. Thus, we want to solve for ϕ such that

$$0 = \sum_{j=i}^{k-1} \lambda\phi_j \mathrm{E}\left[S_{(i,j)}\right] + \lambda\left(1 - \sum_{j=1}^{k-1}\phi_j\right)\mathrm{E}[S_{(i,k)}] - \lambda\left(1 - \sum_{j=1}^{k-1}\phi_j\right)\mathrm{E}[S_{(k,k)}],$$

for all i. Rearranging this equation and dividing by $\lambda(\mathrm{E}[S_{(k,k)}] - \mathrm{E}[S_{(i,k)}])$, we receive

$$\sum_{j=i}^{k-1} \frac{\mathrm{E}\left[S_{(i,j)}\right]}{\mathrm{E}[S_{(k,k)}] - \mathrm{E}[S_{(i,k)}]}\phi_j + \sum_{j=1}^{k-1}\phi_j = 1.$$

We can now observe that this forms the linear system $(M + \mathbf{v}\mathbf{v}^{\mathrm{T}})\phi = \mathbf{v}$, and so we have shown that if ϕ satisfies this system then the means of the sub-queues will be equal. We can note moreover that $M + \mathbf{v}\mathbf{v}^{\mathrm{T}}$ is a rank one update of the matrix M. Thus, it is known that $M + \mathbf{v}\mathbf{v}^{\mathrm{T}}$ will be invertible if $1 + \mathbf{v}^{\mathrm{T}}M^{-1}\mathbf{v} \neq 0$; see Lemma 1.1 of Ding and Zhou (2007). In that case, we know that the unique solution to this system is $\phi = (M + \mathbf{v}\mathbf{v}^{\mathrm{T}})^{-1}\mathbf{v}$.

As we noted in the proof of Theorem 3.3.8, the steady-state distribution of an $M/G/\infty$ queue is $\mathrm{Pois}(\lambda\mathrm{E}[S])$ when the arrival rate is λ and service distribution is equivalent to the random variables S. We can now note further that $\lambda\mathrm{E}[S]$ is the steady-state mean of such a queueing system. The distribution of Q_i is then given by $\mathrm{Pois}(\mathrm{E}[Q_i])$ for each $i \in \{1, \ldots, k\}$, and thus is equivalent across all sub-queues. □

For added motivation, we now consider the two dimensional case in the following remark.

Remark. If $k = 2$, M and ϕ are scalars, given by

$$M = \frac{\mathrm{E}[S]}{\mathrm{E}[S_{2,2}] - \mathrm{E}[S_{1,2}]}, \quad \phi = \frac{\mathrm{E}[S_{2,2}] - \mathrm{E}[S_{1,2}]}{\mathrm{E}[S] + \mathrm{E}[S_{2,2}] - \mathrm{E}[S_{1,2}]}.$$

In this case, we can note that if $\mathrm{P}(B = 1) = \phi$, then in steady-state the distribution of the workload in the system from the easier jobs from all batches will

be equivalent to that of the harder jobs. If $\mathrm{P}(B = 1) > \phi$ the number of harder jobs will stochastically dominate the number of easier jobs, and vice-versa is $\mathrm{P}(B = 1) < \phi$.

This result implies if we have the ability to choose the probability of batch sizes, we can construct each of the sub-systems which are organized by the order statitics to have the same queue length distribution. Thus, providing equal work to all of the queues.

3.4 Conclusion and Final Remarks

In this chapter, we have found parallels between infinite server queues with batch arrivals, sums of scaled Poisson random variables, and Hermite distributions. Moreover, we also connect the stochastic objects to analytic quantities and functions of external interest, such as the harmonic numbers, the exponential integral function, the Euler-Mascheroni constant, and the polylogarithm function. In addition to being interesting in their own right, these connections have helped us to specify exact forms of valuable quantities related to this queueing system, including generating functions for the queue and for the limit of the queue scaled by the batch size. Thus, we have gained both insight into the queue itself and perspective on the model's place in operations research and applied mathematics more broadly.

For this reason, we believe continued work on these fronts is merited. For example, while we have some intuition for the harmonic Hermite distribution discussed in Subsection 3.2.2, we have less of an understanding of the limiting distribution of the scaled queue in that subsection and extended for random

batch sizes in Subsection 3.3.2. Having more knowledge of what distribution might produce a moment generating function comprised of exponential integral function. Finding such a distribution could not only teach us about this queueing system, it would also likely be worth studying entirely on its own. Additionally, providing further connections of this distribution back to the harmonic numbers and the associated Hermite distribution would also be of interest, such as in the connection of the limiting moment generating function to the expected value of a harmonic number evaluated at a Poisson random variable that we remarked in Subsection 3.2.2. One could also consider control problems for the routing of arrivals to sub-systems, like what we discuss for the case of random batch sizes in Subsection 3.3.3.

For future expansions of this work into other areas of queueing, we can group the main themes of potential further investigations in three categories. First, the extension of our batch model beyond infinite server queues to multi-server queues, queues with abandonment, and networks of infinite server queues, a la Mandelbaum and Zeltyn (2007); Massey and Pender (2013); Engblom and Pender (2014); Gurvich et al. (2013); Pender (2014a); Daw and Pender (2019a). It would be interesting to explore our limit theorems in these cases to understand the impact of having a finite number of servers. Second, it would also be interesting to explore the impact of the batch arrivals in the context of queues with delayed information as in Pender et al. (2017a,b, 2018). It would be of interest to know whether or not the batch arrivals would influence the Hopf bifurcations or oscillations that occur in the delayed information queues. Additionally, one could explore findings of this work, like the steady-state distribution representation or the batch scaling, in contexts where there is dependence among the service durations within each batch of arrivals, such as those studied

in Pang and Whitt (2012); Falin (1994). Finally, we are particularly interested in studying the impact of batch arrivals in the context of self-exciting arrival processes such as Hawkes processes like in the work of Gao and Zhu (2018a); Koops et al. (2018) and in Chapter 2. We intend to pursue the ideas described here as well as other related concepts in our future work.

CHAPTER 4

AN EPHEMERALLY SELF-EXCITING POINT PROCESS

4.1 Introduction

What's past is prologue – unavoidably, the present is shaped by what has already occurred. The current state of the world is indebted to our history. Our actions, behaviors, and decisions are both precursory and prescriptive to those that follow, and this can be observed across a variety of different scenarios. For example, the spread of an infectious disease is accelerated as more people become sick and dampened as they recover. In finance, a flurry of recent transactions can prompt new buyers or sellers to enter a market. On social media platforms, as more and more users interact with a post it can become trending or viral and thus be broadcast to an even larger audience.

Self-exciting processes are an intriguing family of stochastic models in which the history of events influences the future. Hawkes (1971) introduced the concept of self-excitement – defining what is now known as the Hawkes process, a model in which "the current intensity of events is determined by events in the past." That is, the Hawkes process is a stochastic intensity point process that depends on the history of the point process itself. The rate of new event occurrences increases as each event occurs. As time passes between occurrences, the

Contents of this chapter are, at the time of this dissertation's writing, under review for publication and are publicly available as a preprint (Daw and Pender, 2020a). Previous publicly available drafts of this work used the moniker "Queue-Hawkes process," but this has been rebranded to avoid confusion and ambiguity. Alas, "the process formerly known as the Queue-Hawkes" is likely not an improvement in this regard.

intensity is governed by a deterministic excitement kernel. Most often, this kernel is specified so that the intensity jumps upward at event epochs and strictly decreases in the interim. In this way, occurrences beget occurrences; hence the term "self-exciting." Unlike the Poisson process, disjoint increments are not independent in sample paths of Hawkes process. Instead, they are positively correlated and, by definition, the events of the former influence the events of the latter. Furthermore, the Hawkes process is known to be over-dispersed – meaning that its variance is larger than its mean – which is commonly found in real world data, whereas the Poisson process has equal mean and variance.

Because of the practical relevance of these model features, self-exciting processes have been used in a wide variety of applications, many of which are quite recent additions to the literature. Seismology was among the first domains to incorporate these models, such as in Ogata (1988), as the occurrence of an earthquake increases the risk of subsequent seismic activity in the form of aftershocks. Finance has since followed as a popular application and is now perhaps the most prolific area of work. In these studies, self-excitement is used to capture the often contagious nature of financial activity, see e.g. Errais et al. (2010); Bacry et al. (2013); Bacry and Muzy (2014); Da Fonseca and Zaatour (2014); Aït-Sahalia et al. (2015); Azizpour et al. (2016); Rambaldi et al. (2017); Gao et al. (2018); Wu et al. (2019). Similarly, there have been many recent internet and social media scenarios that have been modeled using self-exciting processes, drawing upon the virality of modern web traffic. For example, see Farajtabar et al. (2017); Rizoiu et al. (2017, 2018). Notably, this also includes use of Hawkes processes for constructing data-driven methods in the artificial intelligence and machine learning literatures, such as Du et al. (2015); Mei and Eisner (2017); Xu et al. (2017). In an intriguing area of work, self-excitement has also been used

to model inter-personal communication; for example in application to conversation audio recordings in Masuda et al. (2013) or in studying email correspondence in Malmgren et al. (2008); Halpin and De Boeck (2013). Hawkes processes have also recently been used to represent arrivals to service systems in queueing models, e.g. in Gao and Zhu (2018a,b); Koops et al. (2018) as well as in Chapter 2 of this thesis. This is of course not an exhaustive list of works in these areas, nor is it a complete account of all the modern applications of self-excitement. Examples of other notable uses include neuroscience Truccolo et al. (2005); Krumin et al. (2010), environmental management Gupta et al. (2018), public health Zino et al. (2018), movie trailer generation Xu et al. (2015), energy conservation Li and Zha (2018), and industrial preventative maintenance Yan et al. (2013).

As the variety of uses for self-excitement has continued to grow, the number of Hawkes process generalizations has kept pace. By modifying the definition of the Hawkes process in some way, the works in this generalized self-exciting process literature provide new perspectives on these concepts while also empowering and enriching applications. For example, Brémaud and Massoulié (1996) introduce a non-linear Hawkes process that adapts the definition of the process intensity to feature a general, non-negative function of the integration over the process history, as opposed to the linear form given originally. Similarly, the quadratic Hawkes process model given by Blanc et al. (2017) allows for excitation kernels that have quadratic dependence on the process history, rather than simply linear. This is also an example of a generalization motivated by application, as the authors seek to capture time reversal asymmetry observed in financial data. As another finance-motivated generalization, Dassios and Zhao (2011) propose the dynamic contagion process. This model can be thought of as a hybrid between a Hawkes process and a shot-noise process, as the stochastic

intensity of the model features both self-excited and externally excited jumps. The authors take motivation from an application in credit risk, in which the dynamics are shaped by both the process history and by exogenous shocks. The affine point processes studied in e.g. Errais et al. (2010); Zhang et al. (2009, 2015) are also motivated by credit risk applications. The models in these works combine the self-exciting dynamics of Hawkes process with those of an affine jump-diffusion process, imbedding modeling concepts of feedback and dependency into the process intensity. An exact simulation procedure for the Hawkes process with CIR intensity, a generalization of the Hawkes process that is a special case of the affine point process, is shown in Dassios and Zhao (2017). In that case, the authors discuss an application to portfolio loss processes.

There have also been several Hawkes process generalizations proposed in social media and data analytics contexts. For example, Rizoiu et al. (2018) introduces a finite population Hawkes process that couples self-excitement dynamics with those of the Susceptible-Infected-Recovered (SIR) process. Drawing upon the use of the SIR process for the spread of both disease and ideas, the authors propose this SIR-Hawkes process as a method of studying information cascades. Similarly, Mei and Eisner (2017) introduce the neural Hawkes process as a new point process model in the machine learning literature. As the name suggests, this model combines self-excitement with concepts from neural networks. Specifically, a recurrent neural network effectively replaces the excitation kernel, governing the effect of the past events on the rate of future occurrences. In the literature for Bayesian nonparametric models, Du et al. (2015) present the Dirichlet-Hawkes process for topic clustering in document streams. In this case, the authors combine a Hawkes process and a Dirichlet process, so that the intensity of the stream of new documents is self-exciting while the type

of each new document is determined by the Dirichlet process, leading to a preferential attachment structure among the document types.

In this chapter, we propose the *ephemerally self-exciting process*, a novel generalization of the Hawkes process. Rather than regulating the excitement through the gradual, deterministic decay provided by an excitation kernel function, we instead incorporate randomly timed down-jumps. We will refer to this random length of time as the arrival's activity duration. The down-jumps are equal in size to the up-jumps, and between events the arrival rate does not change. Thus, this process increases in arrival rate upon each occurrence, and these increases are then mirrored some time later by decreases in the arrival rate once the activity duration expires. In this way, the self-excitement is ephemeral: it is only in effect as long as the excitement is active. Much of the body of this work will discuss how this ephemeral, piece-wise constant self-excitement compares to the eternal but ever-decaying notion from Hawkes's original definition. As we will see in our analysis, this new process is both a promising model of self-excitement and an explanation of its origins in natural phenomena.

4.1.1 Practical Relevance

While this chapter will not be focused on any one application, in this subsection we summarize several domain areas in which the models in this work can be applied. A natural example is in public health and the management of epidemics. For example, consider influenza. When a person becomes sick with the flu, she increases the rate of spread of the virus through her contact with others. This creates a self-exciting dynamic of the spread of the virus. How-

ever, a person only spreads a disease as long as she is contagious; once she has recovered she no longer has a direct effect on the rate of new infections. From a system-level perspective, the ephemerally self-exciting process can thus be thought of as modeling the arrivals of new infections, capturing the self-exciting and ephemeral nature of sick patients. This motivates the use of this model as an arrival process to queueing models for healthcare, as the rate of *arrivals to* clinics serving patients with infectious diseases should depend on the number of people currently infected. The health-care service can also be separately modeled, as an infinite server queue may be a fitting representation for the number of infected individuals but the clinic itself likely has limited capacity. This concept of course extends to the modeling and management of any other viral disease, including the novel coronavirus that has caused the COVID-19 pandemic.

Of course, epidemic models need not be exclusively applied to disease spread. These same ideas can be used for information spread and product adoption, such as in the aforementioned Hawkes-infused models in Rizoiu et al. (2018) and Zino et al. (2018). In these contexts, one can think of the duration in system as being the time a person actively promotes a concept or product. A single person only affects the self-excitement of the idea or product spread as long as she is in the system, which distinguishes this model from those in the aforementioned works. Epidemic models have also been used to study social issues, such as the contagious nature of imprisonment demonstrated by Lum et al. (2014). We discuss the relevance of ephemeral self-excitement for epidemics in detail in Subsection 4.3.3 by relating this model to the Susceptible-Infected-Susceptible (SIS) process through a convergence in distribution. In fact, throughout Section 4.3 we establish connections from this process to other relevant stochastic models. This includes classical processes such as branching

processes and random walks, as well as models popular both in Bayesian nonparametrics and in preferential attachment settings, such as the Dirichlet process and the Chinese restaurant process.

In the context of service systems, self-excitement can be motivated by the same rationale that inspires restaurants to seat customers at the tables by the windows. Potential new customers could choose to dine at the establishment because they can see others already eating there, taking an implicit recommendation from those already being served. This same example also motivates the ephemerality. After a customer seated by the window finishes her dinner and departs, any passing potential patron only sees an empty table; the implicit recommendation vanishes with the departing customer. A similar dynamic can be observed in online streaming platforms. For example on popular music streaming services like Spotify and Apple Music, users can see what songs and albums have been recently played by their friends. If a user sees that many of her friends have listened to the same album recently, she may be more inclined to listen to it as well. However, this applies only as long as the word "recently" does. If her friends don't play the album within a certain amount of time, the platform will no longer promote the album to her in that fashion. Again, this displays the ephemerality of the underlying self-excitement: the album grows more attractive as more users listen to it, but only as long as those listens are "recent" enough.

In finance, limit order books (LOB's) are among the many concepts that have been modeled using Hawkes process, such as in Rambaldi et al. (2017); Bacry et al. (2016). LOB's have also been studied through queueing models, where one can model the state of the LOB (or, more specifically, the number of un-

resolved bids and asks) as the length of a queueing process. Moreover, there has been recent work that models this process as not just a queue, but a queue with Hawkes process arrivals; for example see Guo et al. (2015); Gao and Zhu (2018a). Conceptually, the self-excitement may arise from traders reacting to the activity of other traders, creating runs of transactions. However, the desire to not act on stale information may mean that this excitement only lasts as long as trades are actively being conducted. In fact, the idea of the self-excitement in LOB models being "queue-reactive" has just very recently been considered by Wu et al. (2019), a related work to this one.

One can also consider failures in a mechanical system as an application of this model. For example, consider a network of water pipes. When one pipe breaks or bursts, it can place stress on the pipes connected to it. This stress may then cause further failures within the pipe network. However once the pipe is properly repaired it should no longer place strain on its surrounding components. Thus, the increase in pipe failure rate caused by a failure is only in effect until the repair occurs, inducing ephemeral self-excitement. The self-excitement (albeit without the ephemerality) arising in this scenario was modeled using Hawkes processes in Yan et al. (2013), which includes an empirical study. A similar problem for electrical systems is considered in Ertekin et al. (2015). The reactive point process considered in that work is perhaps the model most similar to the ones studied herein, as the rate of new power failures both increases at the prior failure times and decreases upon inspection or repair. However, a key difference is that in Ertekin et al. (2015), the authors treat the inspection times as controlled by management, whereas in this chapter the model is fully stochastic and thus the repair durations are random. Regardless, that work is an excellent example of how generalized self-exciting processes can be used to shape prac-

tical policy. Because power outages have significant and wide-reaching consequences, it is critical to understand the inter-dependency between these events and to study the resulting ephemerally self-exciting process that arises in these electrical grid failures.

4.1.2 Organization and Contributions of Chapter

Let us now detail the remainder of this chapter's organization, as well as the contributions therein.

- In Section 4.2, we define the ephemerally self-exciting process (ESEP), a Markovian model for self-excitement that lasts for only a finite amount of time. After defining the model, we develop fundamental distributional quantities and compare the ESEP to the Hawkes process.
- In Section 4.3, we relate the ESEP to many other important and well-known stochastic processes. This includes branching processes, which gives us further comparisons between the Hawkes process and the ESEP, models for preferential attachment and Bayesian statistics, and epidemic models. The lattermost of these motivates the ESEP as a representation for the times of infection within an epidemic, and this also provides a formal link between the conceptually similar concepts of epidemics and self-excitement.
- In Section 4.4, we broaden our exploration of ephemeral self-excitement to non-Markovian models with general activity durations and batches of arrivals. In this general setting, we establish a limit theorem providing an alternate construction of general Hawkes processes. This batch scaling

limit thus yields intuition for the observed occurrence of self-excitement in natural phenomena and stands as a fundamental cornerstone for studying such processes.

In addition to these main avenues of study, we also have extensive auxiliary analysis housed in this chapter's appendix. Appendix A.1 contains lemmas and side results that support our analysis but are outside the main narrative. In Appendix A.2, we explore a model that is a hybrid between the ESEP and the Hawkes process, in that it regulates the excitement with both down-jumps and decay. Appendix A.3 is devoted to a finite capacity version of the ESEP, in which arrivals that would put the active number in system above the capacity are blocked from occurring. Finally, Appendix A.4 contains an algebraically cumbersome proof of a result from Section 4.2.

4.2 Modeling Ephemeral Self-Excitement

We begin this chapter by defining our ephemerally self-exciting model and conducting an initial analysis of some fundamental quantities. These quantities include the transient moment generating function and the steady-state distribution. Before doing so though, let us first review the Hawkes process, which is the original self-exciting probability model.

4.2.1 Defining the Ephemerally Self-Exciting Process

As we have discussed in the introduction, a plethora of natural phenomena exhibit self-exciting features but only for a finite amount of time. This prompts the notion of ephemeral self-excitement. By comparison to the traditional Hawkes process we have reviewed in Subsection 5.3.2, we seek a model in which a new occurrence increases the arrival rate only so long as the newly entered entity remains active in the system. Thus, we now define the *ephemerally self-exciting process* (ESEP), which trades the Hawkes process's eternal decay for randomly drawn expiration times. Moreover, in the following Markovian model, exponential decay is replaced with exponentially distributed durations. In Section 4.4, we extend these concepts to generally distributed service. As another generalization, in Appendix A.2 we consider a Markovian model with both decay and down-jumps. For now, we explore the effects of ephemerality through the ESEP model in Definition 4.2.1.

Definition 4.2.1 (Ephemerally self-exciting process)**.** For times $t \geq 0$, a baseline intensity $\eta^* > 0$, intensity jump size $\alpha > 0$, and expiration rate $\beta > 0$, let N_t be a counting process with stochastic intensity η_t such that

$$\eta_t = \eta^* + \alpha Q_t, \tag{4.1}$$

where Q_t is incremented with N_t and then is depleted at unit down-jumps according to the rate βQ_t. Then, we say that (η_t, N_t) is an *ephemerally self-exciting process* (ESEP).

We will assume that η_0 and Q_0 are known initial values such that $\eta_0 = \eta^* + \alpha Q_0$. In addition to this definition, one could also describe the ESEP through

its dynamics. In particular, the behavior of this process can be summarily cast through the life cycle of its arrivals:

i) At each arrival, the arrival rate η_t increases by α.

ii) Each arrival remains active for an activity duration drawn from an i.i.d. sequence of exponential random variables with rate β.

iii) At the expiration of a given activity duration, η_t decreases by α.

The ephemerality of the ESEP is embodied by this cycle. Because arrivals only contribute to the intensity for the length of their activity duration, their effect on the process's excitation vanishes when this clock expires. Furthermore, there is an affine relationship between the number of active "exciters" – meaning unexpired arrivals still causing excitation – and the intensity, i.e. $\eta_t = \eta^* + \alpha Q_t$. Thus, we could also track the arrival rate through Q_t in place of η_t and still have full understanding of this process. This also means that results are readily transferrable between these two processes; we will often make use of this fact. x Because the ESEP is quite parsimonious, there are many alternative perspectives we could take to gain additional understanding of it. For example, one could consider Q_t a Markovian queueing system with infinitely many servers and a state dependent arrival rate. Equivalently, one could also describe the ESEP as a Markov chain on the non-negative integers where transitions at state i are to $i + 1$ at rate $\eta^* + \alpha i$ and to $i - 1$ at rate μi, with the counting process then defined as the epochs of the upward jumps in this chain. A visualization of this linear birth-death-immigration process is given in Figure 4.1. Stability for this chain occurs when $\beta > \alpha$; we will assume this hereforward although it of course is not necessary for transient results.

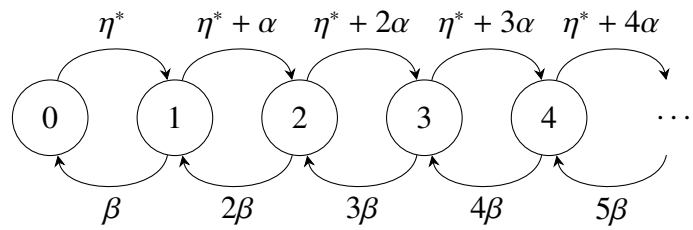


Figure 4.1: The transition diagram of the Markov chain for Q_t.

In the remainder of this subsection, let us now develop a few fundamental quantities for this stochastic process, particularly its intensity and active number in system as these capture the self-exciting behavior of the process. First, in Proposition 4.2.1 we compute the transient moment generating function for the intensity η_t. As we have noted, this can also be used to immediately derive the same transform for Q_t, and the proof makes use of this fact.

Proposition 4.2.1. *Let $\eta_t = \eta^* + \alpha Q_t$ be the intensity of an ESEP with baseline intensity $\eta^* > 0$, intensity jump $\alpha > 0$, and expiration rate $\beta > \alpha$. Then, the moment generating function for η_t is given by*

$$\mathrm{E}\left[e^{\theta\eta_t}\right] = \left(\frac{\beta - \alpha e^{\alpha\theta} - \beta(1-e^{\alpha\theta})e^{-(\beta-\alpha)t}}{\beta - \alpha e^{\alpha\theta} - \alpha(1-e^{\alpha\theta})e^{-(\beta-\alpha)t}}\right)^{\frac{\eta_0-\eta^*}{\alpha}} \left(\frac{\beta e^{\alpha\theta}}{\beta - \alpha e^{\alpha\theta}} - \frac{\alpha e^{\alpha\theta}}{\beta - \alpha e^{\alpha\theta}}\left(\frac{\beta - \alpha e^{\alpha\theta} - \beta(1-e^{\alpha\theta})e^{-(\beta-\alpha)t}}{\beta - \alpha e^{\alpha\theta} - \alpha(1-e^{\alpha\theta})e^{-(\beta-\alpha)t}}\right)\right)^{\frac{\eta^*}{\alpha}},$$

for all $t \geq 0$ and $\theta < \frac{1}{\alpha}\log\left(\frac{\beta}{\alpha}\right)$.

Proof. We will approach this through the perspective of the active number in system, Q_t. Using Lemma A.1.1, we have that the probability generating function for Q_t, say $\mathcal{P}(z,t) = \mathrm{E}\left[z^{Q_t}\right]$, is given by the solution to the following partial differential equation:

$$\frac{\partial}{\partial t}\mathrm{E}\left[z^{Q_t}\right] = \mathrm{E}\left[(\eta^* + \alpha Q_t)\left(z^2 - z\right)z^{Q_t-1} + \beta Q_t\,(1-z)\,z^{Q_t-1}\right],$$

which is equivalently expressed

$$\frac{\partial}{\partial t}\mathcal{P}(z,t) = \eta^*\,(z-1)\,\mathcal{P}(z,t) + \left(\alpha\left(z^2 - z\right) + \beta\,(1-z)\right)\frac{\partial}{\partial z}\mathcal{P}(z,t),$$

with initial condition $\mathcal{P}(z,0) = z^{Q_0}$. The solution to this initial value problem is given by

$$\mathcal{P}(z,t) = \left(\frac{\beta - \alpha z - \beta(1-z)e^{-(\beta-\alpha)t}}{\beta - \alpha z - \alpha(1-z)e^{-(\beta-\alpha)t}}\right)^{Q_0} \left(\frac{\beta}{\beta - \alpha z} - \frac{\alpha}{\beta - \alpha z}\left(\frac{\beta - \alpha z - \beta(1-z)e^{-(\beta-\alpha)t}}{\beta - \alpha z - \alpha(1-z)e^{-(\beta-\alpha)t}}\right)\right)^{\frac{\eta^*}{\alpha}},$$

yielding the probability generating function for Q_t. By setting $z = e^{\theta}$ we receive the moment generating function. Finally, using the affine relationship $\eta_t = \eta^* + \alpha Q_t$, we have that

$$\mathrm{E}\left[e^{\theta\eta_t}\right] = \mathrm{E}\left[e^{\theta(\eta^* + \alpha Q_t)}\right] = e^{\theta\eta^*}\mathrm{E}\left[e^{\alpha\theta Q_t}\right],$$

with $\eta_0 = \eta^* + \alpha Q_0$. □

As we have mentioned, this Markov chain can be shown to be stable for $\beta > \alpha$ through standard techniques. Thus, using the moment generating function from Proposition 4.2.1, we can find the steady-state distributions of the intensity and the active number in system by taking the limit of t. We can quickly observe that this leads to a negative binomial distribution, as we state in Theorem 4.2.2. Because of the varying definitions of the negative binomial distribution, we state the probability mass function explicitly.

Theorem 4.2.2. *Let $\eta_t = \eta^* + \alpha Q_t$ be an ESEP with baseline intensity $\eta^* > 0$, intensity jump $\alpha > 0$, and expiration rate $\beta > \alpha$. Then, the activer number in system in steady-state follows a negative binomial distribution with probability of success $\frac{\alpha}{\beta}$ and number of failures $\frac{\eta^*}{\alpha}$, which is to say that the steady-state probability mass function is*

$$\mathrm{P}\left(Q_\infty = k\right) = \frac{\Gamma\left(k + \frac{\eta^*}{\alpha}\right)}{\Gamma\left(\frac{\eta^*}{\alpha}\right)k!}\left(\frac{\beta - \alpha}{\beta}\right)^{\frac{\eta^*}{\alpha}}\left(\frac{\alpha}{\beta}\right)^k. \tag{4.2}$$

Consequently, the steady-state distribution of the intensity is given by a shifted and scaled negative binomial with probability of success $\frac{\alpha}{\beta}$ and number of failures $\frac{\eta^*}{\alpha}$, shifted by η^* and scaled by α.

Proof. Using Proposition 4.2.1, we can see that the steady-state moment generating function of Q_t is given by

$$\lim_{t\to\infty} \mathrm{E}\left[e^{\theta Q_t}\right] = \left(\frac{\beta-\alpha}{\beta-\alpha e^{\theta}}\right)^{\frac{\eta^*}{\alpha}}.$$

We can observe that this steady-state moment generating function is equivalent to that of a negative binomial. By the affine transformation $\eta_t = \eta^* + \alpha Q_t$, we find the steady-state distribution for the intensity. □

Let us pause to note that this explicit characterization of the steady-state intensity is already an advantage of the ESEP over the traditional Markovian Hawkes process, for which there is not a closed form intensity stationary distribution available. As a consequence of Theorem 4.2.2, we can observe that the steady-state mean of the intensity is $\eta_\infty := \frac{\beta\eta^*}{\beta-\alpha}$. Interestingly, this would also be the steady-state mean of the Hawkes process when given the same baseline intensity, the same intensity jump size, and an exponential decay rate equal to the rate of expiration. This leads us to ponder how the processes would otherwise compare when given equivalent parameters. In Proposition 4.2.3 we find that although this equivalence of means extends to transient settings, for all higher moments the ESEP dominates the Hawkes process in terms of both the intensity and the counting process.

Proposition 4.2.3. *Let $(\eta_t, N_{t,\eta})$ be an ESEP intensity and counting process pair with jump size $\alpha > 0$, expiration rate $\beta > \alpha$, and baseline intensity $\eta^* > 0$. Similarly, let $(\lambda_t, N_{t,\lambda})$ be a Hawkes process intensity and counting process pair with jump size $\alpha > 0$, decay rate $\beta > 0$, and baseline intensity $\eta^* > 0$. Then, if the two processes have equal initial values, their means will satisfy*

$$\mathrm{E}\left[\lambda_t\right] = \mathrm{E}\left[\eta_t\right], \qquad \mathrm{E}\left[N_{t,\lambda}\right] = \mathrm{E}\left[N_{t,\eta}\right], \tag{4.3}$$

and for $m \geq 2$ their m^{th} moments are ordered such that

$$\mathrm{E}\left[\lambda_t^m\right] \leq \mathrm{E}\left[\eta_t^m\right], \qquad \mathrm{E}\left[N_{t,\lambda}^m\right] \leq \mathrm{E}\left[N_{t,\eta}^m\right], \tag{4.4}$$

for all time $t \geq 0$.

Proof. Let us start with the means. For the intensities, we can note that these are given by the solutions to

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[\lambda_t\right] = \alpha\mathrm{E}\left[\lambda_t\right] - \beta(\mathrm{E}\left[\lambda_t\right] - \eta^*) \qquad \text{and} \qquad \frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[\eta_t\right] = \alpha\mathrm{E}\left[\eta_t\right] - \beta\alpha\left(\frac{\mathrm{E}\left[\eta_t\right] - \eta^*}{\alpha}\right),$$

and through simplification one can quickly observe that these two ODE's are equivalent. Thus, because we have assumed that the processes have the same initial values, we find that $\mathrm{E}\left[\lambda_t\right] = \mathrm{E}\left[\eta_t\right]$. Since

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[N_{t,\lambda}\right] = \mathrm{E}\left[\lambda_t\right] \qquad \text{and} \qquad \frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[N_{t,\eta}\right] = \mathrm{E}\left[\eta_t\right],$$

this equality immediately extends to the means of the counting processes as well. We will now use these equations for the means as the base cases for inductive arguments, beginning again with the intensities. For the inductive step, we will assume that the intensity moment ordering holds for moments 1 to $m-1$. The mth moment of the Hawkes process intensity is thus given by the solution to

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[\lambda_t^m\right] = \sum_{k=0}^{m-1}\binom{m}{k}\mathrm{E}\left[\lambda_t^{k+1}\right]\alpha^{m-k} - m\beta\mathrm{E}\left[\lambda_t^m\right] + m\beta\eta^*\mathrm{E}\left[\lambda_t^{m-1}\right] := f_\lambda(t, \mathrm{E}\left[\lambda_t^m\right]),$$

where $f_\lambda(t, \mathrm{E}\left[\lambda_t^m\right])$ is meant to capture that this ODE depends on the value of the mth moment and of the lower moments, which by the inductive hypothesis we take as known functions of the time t. Then, the mth moment of the ESEP intensity will satisfy

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[\eta_t^m\right] = \sum_{k=0}^{m-1}\binom{m}{k}\mathrm{E}\left[\eta_t^{k+1}\right]\alpha^{m-k} + \frac{\beta}{\alpha}\sum_{k=0}^{m-1}\binom{m}{k}\mathrm{E}\left[\eta_t^{k+1}\right](-\alpha)^{m-k} - \frac{\beta\eta^*}{\alpha}\sum_{k=0}^{m-1}\binom{m}{k}\mathrm{E}\left[\eta_t^{k}\right](-\alpha)^{m-k}.$$

By pulling the $k = m - 1$ terms off the top of each summation, we can re-express this ODE as

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[\eta_t^m\right] = \sum_{k=0}^{m-1}\binom{m}{k}\mathrm{E}\left[\eta_t^{k+1}\right]\alpha^{m-k} - m\beta\mathrm{E}\left[\eta_t^m\right] + m\beta\eta^*\mathrm{E}\left[\eta_t^{m-1}\right]$$
$$+ \frac{\beta}{\alpha}\sum_{k=0}^{m-2}\binom{m}{k}(-\alpha)^{m-k}\left(\mathrm{E}\left[\eta_t^{k+1}\right] - \eta^*\mathrm{E}\left[\eta_t^k\right]\right),$$

and through the definition of $f_\lambda(\cdot)$ and the inductive hypothesis, we can find the following lower bound:

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[\eta_t^m\right] \geq f_\lambda(t, \mathrm{E}\left[\eta_t^m\right]) + \frac{\beta}{\alpha}\sum_{k=0}^{m-2}\binom{m}{k}(-\alpha)^{m-k}\left(\mathrm{E}\left[\eta_t^{k+1}\right] - \eta^*\mathrm{E}\left[\eta_t^k\right]\right).$$

This right-most term can then be expressed

$$\frac{\beta}{\alpha}\sum_{k=0}^{m-2}\binom{m}{k}(-\alpha)^{m-k}\left(\mathrm{E}\left[\eta_t^{k+1}\right] - \eta^*\mathrm{E}\left[\eta_t^k\right]\right) = \frac{\beta}{\alpha}\mathrm{E}\left[(\eta_t - \eta^*)\left((\eta_t - \alpha)^m - \eta_t^m + m\alpha\eta_t^{m-1}\right)\right],$$

and we can now reason about the quantity inside the expectation. By definition, we have that $\eta_t \geq \eta^*$ surely, and furthermore we can observe that if $\eta_t - \eta^* > 0$, then $\eta_t \geq \eta^* + \alpha > \alpha$. Thus, let us consider $(\eta_t - \alpha)^m - \eta_t^m + m\alpha\eta_t^{m-1}$ assuming $\eta_t > \alpha$. Dividing through by η_t^m, we have the expression

$$\left(1 - \frac{\alpha}{\eta_t}\right)^m - 1 + \frac{m\alpha}{\eta_t}. \tag{4.5}$$

Since $(1 - x)^m - 1 + mx$ is equal to 0 at $x = 0$ and is non-decreasing on $x \in [0, 1)$ via a first-derivative check, we can note that (4.5) is non-negative for all $\eta_t > \eta^*$. Thus, we have that

$$\mathrm{E}\left[(\eta_t - \eta^*)\left((\eta_t - \alpha)^m - \eta_t^m + m\alpha\eta_t^{m-1}\right)\right] \geq 0,$$

and by consequence,

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[\eta_t^m\right] \geq f_\lambda(t, \mathrm{E}\left[\eta_t^m\right]),$$

completing the proof of the intensity moment ordering via Lemma A.1.2. For the counting processes, let us again assume as an inductive hypothesis that the

moment ordering holds for moments 1 through $m-1$, with the mean equality serving as the base case. Then, the mth moment of the ESEP counting process will satisfy

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[N_{t,\eta}^{m}\right] = \sum_{k=0}^{m-1}\binom{m}{k}\mathrm{E}\left[\eta_t N_{t,\eta}^{k}\right],$$

and the ODE for the mth Hawkes counting process moment is analogous. By the FKG inequality, we can observe that

$$\sum_{k=0}^{m-1}\binom{m}{k}\mathrm{E}\left[\eta_t N_{t,\eta}^{k}\right] \geq \sum_{k=0}^{m-1}\binom{m}{k}\mathrm{E}\left[\eta_t\right]\mathrm{E}\left[N_{t,\eta}^{k}\right].$$

By the inductive hypothesis, we have that $\mathrm{E}\left[N_{t,\eta}^{k}\right] \geq \mathrm{E}\left[N_{t,\lambda}^{k}\right]$ for each $k \leq m-1$, and thus we can observe that

$$\sum_{k=0}^{m-1}\binom{m}{k}\mathrm{E}\left[\eta_t\right]\mathrm{E}\left[N_{t,\eta}^{k}\right] \geq \sum_{k=0}^{m-1}\binom{m}{k}\mathrm{E}\left[\lambda_t\right]\mathrm{E}\left[N_{t,\lambda}^{k}\right].$$

Finally, by another application of Lemma A.1.2, we have $\mathrm{E}\left[N_{t,\lambda}^{m}\right] \leq \mathrm{E}\left[N_{t,\eta}^{m}\right]$. □

The fact that the ESEP variance dominates the Hawkes variance should not be surprising, since the presence of both up and down jumps means that the ESEP sample paths should be subject to more abrupt changes. Nevertheless, this also shows that the ESEP is more over-dispersed than the Hawkes process is. This may be an attractive feature for data modeling. It is worth noting that matrix computations are available for all moments of these intensities via Chapter 5, through which one could use the method of moments to fit the processes to data.

4.2.2 The Ephemerally Self-Exciting Counting Process

Thus far we have studied the intensity of the ESEP, as this process is by definition tracking the self-excitement. However, this excitation is manifested in the actual arrivals from the process, which are counted in N_t. We now turn our attention to developing fundamental quantities for this counting process. To begin, we give the probability generating function of the counting process in closed form below in Proposition 4.2.4. One can note that by comparison, the generating functions of the Hawkes process are instead only expressible as functions of ordinary differential equations with no known closed form solutions, see for example Chapter 2.

Proposition 4.2.4. *Let N_t be the number of arrivals by time $t \geq 0$ in an ESEP with baseline intensity $\eta^* > 0$, intensity jump $\alpha > 0$, and expiration rate $\beta > \alpha$. Then, the probability generating function of N_t is given by*

$$\mathrm{E}\left[z^{N_t}\right] = e^{\frac{\eta^*(\beta-\alpha)}{2\alpha}t}\left(\frac{2e^{\frac{t}{2}\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}{1-\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}+\left(1+\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}\right)e^{t\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}\right)^{\frac{\eta^*}{\alpha}}$$
$$\cdot\left(\frac{\beta+\alpha}{2\alpha}+\frac{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}{2\alpha}\left(\frac{1-\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}-\left(1+\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}\right)e^{t\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}{1-\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}+\left(1+\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}\right)e^{t\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}\right)\right)^{Q_0} \tag{4.6}$$

where Q_0 is the active number in system at time 0.

Proof. Due to the cumbersome length of some equations, please see Appendix A.4 for the proof. □

In addition to calculating the probability generating function, we can also

find a matrix calculation for the transient probability mass function of the counting process. To do so, we recognize that the time until the next arrival occurs can be treated as the time to absorption in a continuous time Markov chain. By building from this idea to construct a transition matrix for several successive arrivals, we find the form for the distribution given in Proposition 4.2.5.

Proposition 4.2.5. *Let N_t be the number of arrivals by time t in an ESEP with baseline intensity $\eta^* > 0$, intensity jump $\alpha > 0$, and expiration rate $\beta > \alpha$. Further, let $Q_0 = k$ be the initial active number in system. Then for $i \in \mathbb{N}$, define the matrices $\mathbf{D}_i \in \mathbb{R}^{k+i+1 \times k+i+1}$ and $\mathbf{S}_i \in \mathbb{R}^{k+i+1 \times k+i+2}$ as*

$$\mathbf{D}_i = \begin{bmatrix} -(\eta^* + (k+i)(\alpha+\beta)) & (k+i)\beta & & & \\ & -(\eta^* + (k+i-1)(\alpha+\beta)) & & & \\ & & \ddots & & \\ & & & -(\eta^* + \alpha + \beta) & \beta \\ & & & & -\eta^* \end{bmatrix},$$

and

$$\mathbf{S}_i = \begin{bmatrix} \eta^* + \alpha(k+i) & & & & 0 \\ & \eta^* + \alpha(k+i-1) & & & 0 \\ & & \ddots & & \vdots \\ & & & \eta^* + \alpha & 0 \\ & & & \eta^* & 0 \end{bmatrix}.$$

Further, let $\mathbf{Z}_n \in \mathbb{R}^{\hat{d}_n \times \hat{d}_n}$ *for* $\hat{d}_n = \frac{n(n+1)}{2} + (n+1)(k+1)$ *be a matrix such that*

$$\mathbf{Z}_n = \begin{bmatrix} \mathbf{D}_0 & \mathbf{S}_0 & & & & \\ & \mathbf{D}_1 & \mathbf{S}_1 & & & \\ & & \ddots & \ddots & & \\ & & & \ddots & \mathbf{S}_{n-2} & \\ & & & & \mathbf{D}_{n-1} & \mathbf{S}_{n-1} \\ & & & & & \mathbf{D}_n \end{bmatrix}.$$

Then, the probability that $N_t = n$ *is given by*

$$\mathrm{P}(N_t = n) = \mathbf{v}_1^{\mathrm{T}} e^{\mathbf{Z}_n t} \mathbf{v}_{:} \tag{4.7}$$

where $\mathbf{v}_j \in \mathbb{R}^{\hat{d}_n}$ *is the unit column vector for the* j^{th} *coordinate and* $\mathbf{v}_{:} = \sum_{j=0}^{k+n} \mathbf{v}_{\hat{d}_n - j}$.

Proof. This follows directly from viewing $\mathbf{Z}_n$ as a sub-matrix of the generator matrix of a CTMC, much like one can do to calculate probabilities of phase-type distributions. Specifically, the sub-generator matrix is defined on the state space $\mathcal{S} = \bigcup_{i=0}^{n} \{(0,i),(1,i),\ldots,(k+i-1,i),(k+i,i)\}$. In this scenario, the state (s_1, s_2) represents having s_1 entities in system and having seen s_2 arrivals since time 0. Then, $\mathbf{D}_i$ is the sub-generator matrix for transitions among the sub-state space $\{(k+i,i),(k+i-1,i),\ldots,(1,i),(0,i)\}$ to itself (where the states are ordered in that fashion). Similarly, $\mathbf{S}_i$ is for transitions from states in $\{(k+i,i),(k+i-1,i),\ldots,(1,i),(0,i)\}$ to states in $\{(k+i+1,i+1),(k+i,i+1),\ldots,(1,i+1),(0,i+1)\}$. Then, one can consider this from an absorbing CTMC perspective since if $n+1$ arrivals occur it is not possible to transition back to any state in which n arrivals had occurred. Hence, we only need to use the matrix $\mathbf{Z}_n$ to consider up to n arrivals. Then, $e^{\mathbf{Z}_n t}$ is the sub-matrix for probabilities of transitions among states in $\mathcal{S}$, where the rows will sum to less than 1 as it is possible that the chain has experienced more than n arrivals by time t. Finally, because $Q_0 = k$ we know that

the chain states in state $(k, 0)$; further, because we are seeking the probability that there have been exactly n arrivals by time t we want the probability of transitions from $(k, 0)$ to any of the states in $\{(k+n, n), (k+n-1, n), \ldots, (1, n), (0, n)\}$. □

With these fundamental quantities in hand, let us now turn to explore more nuanced connections between the ESEP and other stochastic processes in the following section. Doing so will provide further comparison between the Hawkes process and the ESEP, and moreover will formally connect the notion of self-excitement to similar concepts such as contagion, virality, and rich-get-richer effects.

4.3 Relating Ephemeral Self-Excitement to Branching Processes, Random Walks, and Epidemics

Aside from the original definition, the most frequently utilized result for Hawkes processes is perhaps the immigration-birth representation first shown in Hawkes and Oakes (1974). By viewing a portion of arrivals as immigrants – externally driven and stemming from a homogenous Poisson process – and then viewing the remaining portion as offspring – excitation-driven descendants of the immigrants and the prior offspring – one can take new perspectives on self-exciting processes. From this position, if an arrival is a descendant then it has a unique parent, the excitement of which spurred this arrival into existence. Every entity has the potential to generate offspring. This viewpoint takes on added meaning in the context of ephemeral self-excitement, as an entity only has the opportunity to generate descendants so long as it remains in the system. In this section, we will use this idea to connect self-exciting processes to well-known

stochastic models that have applications ranging from public health to Bayesian statistics. Furthermore, these connections will also help us form comparisons between the Hawkes process and the model we have introduced, the ESEP. The different dynamics are at the forefront of this process comparison, as the branching structure is dictated by the self-excitation caused by a single arrival. For the Hawkes process, this increase in the arrival rate is eternal but ever-diminishing, whereas in the ESEP the jump is ephemeral but constant when it does exist.

4.3.1 Discrete Time Perspectives through Branching Processes

Let us first view these processes through a discrete time lens as branching processes. In this subsection we will interpret classical branching processes results in application to these self-exciting processes. Taking the immigration-birth representation as inspiration, we start by considering the distribution of the total number of offspring of a single arrival. That is, we want to calculate the probability mass function for the number of arrivals that are generated directly from the excitement caused by the initial arrival. To constitute the *total* number of offspring, we will consider all the children of this initial entity across all time. For the ESEP, this equates to the number of arrivals generated by the entity throughout its duration in the system; in the Hawkes process this counts the number of arrivals spurred by the entity as time goes to infinity. Given that the stability conditions are satisfied throughout, in Proposition 4.3.1 we calculate these distributions by way of inhomogeneous Poisson processes, yielding a Poisson mixture form for each.

Proposition 4.3.1. *Let $\beta > \alpha > 0$. Let X^η be the number of new arrivals generated by the excitement caused by an arbitrary initial arrival throughout its duration in the*

system in an ESEP with jump size α and expiration rate β. Then, this offspring distribution is geometrically distributed with probability mass function

$$\mathrm{P}\left(X^{\eta} = k\right) = \left(\frac{\beta}{\alpha+\beta}\right)\left(\frac{\alpha}{\alpha+\beta}\right)^{k}. \tag{4.8}$$

Similarly, let X^{λ} be the number of new arrivals generated by the excitement caused by an arbitrary initial arrival in a Hawkes process with jump size α and decay rate β. This offspring distribution is then Poisson distributed with probability mass function

$$\mathrm{P}\left(X^{\lambda} = k\right) = \frac{e^{-\frac{\alpha}{\beta}}}{k!}\left(\frac{\alpha}{\beta}\right)^{k}. \tag{4.9}$$

where all $k \in \mathbb{N}$.

Proof. Without loss of generality, we assume that the initial arrival in each process occurred at time 0. Then, at time $t \geq 0$ the excitement generated by these initial arrivals has intensities given by $\alpha e^{-\beta t}$ and $\alpha \mathbf{1}\{t < S\}$ for the Hawkes and ESEP processes, respectively, where $S \sim \mathrm{Exp}(\beta)$. Using Daley and Vere-Jones (2007), one can note that the offspring distributions across all time can then be expressed as

$$X^{\lambda} \sim \mathrm{Pois}\left(\alpha \int_0^{\infty} e^{-\beta t}\mathrm{d}t\right) \qquad \text{and} \qquad X^{\eta} \sim \mathrm{Pois}\left(\alpha \int_0^{\infty} \mathbf{1}\{t < S\}\mathrm{d}t\right),$$

which are equivalently stated $X^{\lambda} \sim \mathrm{Pois}\left(\alpha/\beta\right)$ and $X^{\eta} \sim \mathrm{Pois}\left(\alpha S\right)$. This now immediately yields the stated distributions for X^{λ} and X^{η}, as the Poisson-Exponential mixture is known to yield a geometric distribution, see for example the overview of Poisson mixtures in Karlis and Xekalaki (2005). □

We now move towards considering the total progeny of an initial arrival, meaning the total number of arrivals generated by the excitement of an initial arrival *and* the excitement of its offspring, and of their offspring, and so on across

all time. It is important to note that by comparison to the number of offspring, the progeny includes the initial arrival itself. As we will see, the stability of the self-exciting processes implies that this total number of descendants is almost surely finite. This demonstrates the necessity of immigration for these processes to survive. From the offspring distributions in Proposition 4.3.1, the Hawkes descendant process is a Poisson branching process and, similarly, the ESEP is a geometric branching process. These are well-studied models in branching processes, so we have many results available to us. In fact, we now use a result for random walks with potentially multiple simultaneous steps forward to derive the progeny distributions for these two processes. This is through the well-known hitting time theorem, stated below in Lemma 4.3.2.

Lemma 4.3.2 (Hitting Time Theorem)**.** *The total progeny Z of a branching process with descendant distribution equivalent to X_1 is*

$$\mathrm{P}(Z = k) = \frac{1}{k}\mathrm{P}(X_1 + X_2 + \cdots + X_k = k - 1),$$

where $X_1, \ldots, X_k$ are i.i.d. for all $k \in \mathbb{Z}^+$.

Proof. See Otter (1949) for the original statement and proof in terms of random walks; a review and elementary proof are given in the brief note Van der Hofstad and Keane (2008). □

We now use the hitting time theorem to give the total descendants distributions for the Hawkes process and the ESEP in Proposition 4.3.3. This is a common technique for branching processes, and it now yields valuable insight into these two self-exciting models.

Proposition 4.3.3. *Let $\beta > \alpha > 0$. Let Z^η be a random variable for the total progeny of an arbitrary arrival in an ESEP with intensity jump α and expiration rate β. Likewise,*

let Z^λ be a random variable for the total progeny of an arbitrary arrival in a Hawkes process with intensity jump α and decay rate β. Then, the probability mass functions for Z^η and Z^λ are given by

$$\mathrm{P}\left(Z^\eta = k\right) = \frac{1}{k}\binom{2k-2}{k-1}\left(\frac{\beta}{\beta+\alpha}\right)^k\left(\frac{\alpha}{\beta+\alpha}\right)^{k-1} \textit{ and } \quad \mathrm{P}\left(Z^\lambda = k\right) = \frac{e^{-\frac{\alpha}{\beta}k}}{k!}\left(\frac{\alpha k}{\beta}\right)^{k-1}, \quad (4.10)$$

where $k \in \mathbb{Z}^+$.

Proof. This follows by applying Lemma 4.3.2 to Proposition 4.3.1. Because the sum of independent Poisson random variables is Poisson distributed with the sum of the rates, we have that

$$\frac{1}{k}\mathrm{P}\left(X_1^\lambda + X_2^\lambda + \ldots X_k^\lambda = k-1\right) = \frac{1}{k}\mathrm{P}\left(K_1 = k-1\right),$$

where $K_1 \sim \mathrm{Pois}\left(\frac{\alpha k}{\beta}\right)$. This now yields the expression for the probability mass function for Z^λ. Similarly for Z_η we note that the sum of independent geometric random variables has a negative binomial distribution, which implies that

$$\frac{1}{k}\mathrm{P}\left(X_1^\eta + X_2^\eta + \ldots X_k^\eta = k-1\right) = \frac{1}{k}\mathrm{P}\left(K_2 = k-1\right),$$

where $K_2 \sim \mathrm{NegBin}\left(k, \frac{\alpha}{\beta+\alpha}\right)$, and this completes the proof. □

For a visual comparison of the descendants in the ESEP and the Hawkes process, we plot these two progeny distributions for equivalent parameters in Figure 4.2. As suggested by the variance ordering in Proposition 4.2.3, the tail of the ESEP progeny distribution is heavier than that of the Hawkes process.

We can note that while one can calculate the mean of each progeny via the probability mass functions in Proposition 4.3.3, they can also easily be found using Wald's identity. Through standard infinitesimal generator approaches, one

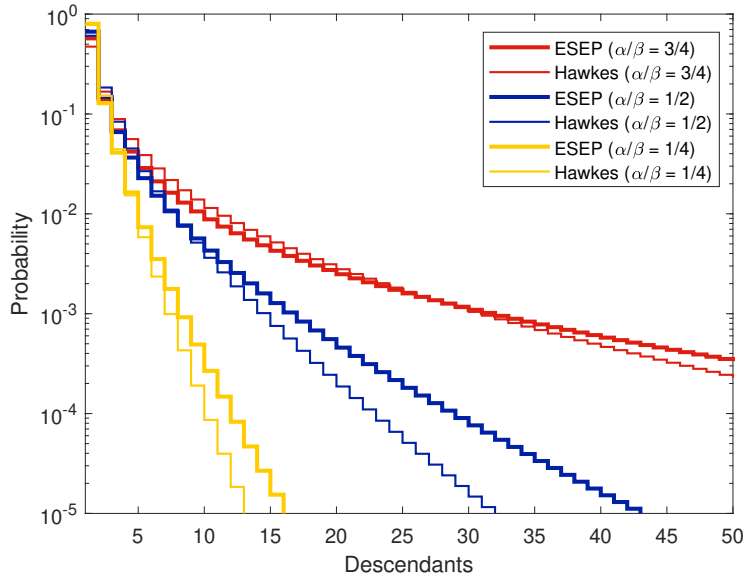


Figure 4.2: Progeny distributions for the ESEP and the Hawkes process with matching parameters.

can calculate that the expected number of arrivals (including by immigration) in the ESEP is

$$\mathrm{E}\left[N_t\right] = \frac{\beta\eta^* t}{\beta-\alpha} + \frac{\eta_0 - \eta_\infty}{\beta-\alpha}(1 - e^{-(\beta-\alpha)t}).$$

However, using these branching process representations, we can also express this as

$$\mathrm{E}\left[N_t\right] = \mathrm{E}\left[\sum_{i=1}^{M_t} Z_i(t)\right],$$

where M_t is a Poisson process with rate η^* and $Z_i(t)$ are the total progeny up to time $t \geq 0$ that descend from the ith immigrant arrival. Now, by applying Wald's identity to the limit of $\frac{1}{t}\mathrm{E}\left[N_t\right]$ as $t \to \infty$, we see that

$$\frac{\beta\eta^*}{\beta-\alpha} = \lim_{t\to\infty} \frac{\mathrm{E}\left[N_t\right]}{t} = \lim_{t\to\infty} \frac{1}{t}\mathrm{E}\left[\sum_{i=1}^{M_t} Z_i(t)\right] = \eta^*\mathrm{E}\left[Z^\eta\right],$$

and so $\mathrm{E}\left[Z^\eta\right] = \frac{\beta}{\beta-\alpha}$. By analogous arguments for the Hawkes process, we see that $\mathrm{E}\left[Z^\lambda\right] = \frac{\beta}{\beta-\alpha}$.

As a final branching process comparison between these two processes, we calculate the distribution of the total number of generations of descendants of an initial arrival in the ESEP and the Hawkes process. That is, let the first entity be the first generation, its offspring be the second generation, their offspring the third, and so on. In Proposition 4.3.4 we find the probability mass function for the ESEP in closed form and a recurrence relation for the cumulative distribution function for the Hawkes process.

Proposition 4.3.4. *Let $\beta > \alpha > 0$. Let $\mathcal{G}^\eta$ be the number of distinct arrival generations across in full the progeny of an initial arrival in an ESEP with intensity jump α and service rate β. Then, the probability mass function for $\mathcal{G}^\eta$ is given by*

$$\mathrm{P}\left(\mathcal{G}^\eta = k\right) = \frac{\alpha^{k-1}(\beta-\alpha)}{\beta^k - \alpha^k} - \frac{\alpha^k(\beta-\alpha)}{\beta^{k+1} - \alpha^{k+1}}. \tag{4.11}$$

Likewise, let $\mathcal{G}^\lambda$ be the number of distinct arrival generations in the full progeny of an initial arrival for a Hawkes process with intensity jump $\alpha > 0$ and decay rate β. Then, $\mathcal{G}^\lambda$ has cumulative distribution function $F_{\mathcal{G}^\lambda}(k) = \mathrm{P}\left(\mathcal{G}^\lambda \leq k\right)$ satisfying the recursion

$$F_{\mathcal{G}^\lambda}(k) = e^{-\frac{\alpha}{\beta}\left(1 - F_{\mathcal{G}^\lambda}(k-1)\right)}, \tag{4.12}$$

where $F_{\mathcal{G}^\lambda}(0) = 0$ and all $k \in \mathbb{Z}^+$.

Proof. Let Y_k^λ and Y_k^η be Galton-Watson branching processes defined as

$$Y_k^\lambda = \sum_{i=1}^{Y_{k-1}^\lambda} X_{\lambda,i}^{(k)}, \qquad\qquad Y_k^\eta = \sum_{i=1}^{Y_{k-1}^\eta} X_{\eta,i}^{(k)}, \tag{4.13}$$

with $X_{\lambda,i}^{(k)} \overset{i.i.d.}{\sim} \mathrm{Pois}\left(\frac{\alpha}{\beta}\right)$, $X_{\eta,i}^{(k)} \overset{i.i.d.}{\sim} \mathrm{Geo}\left(\frac{\alpha}{\alpha+\beta}\right)$, and $Y_0^\lambda = Y_0^\eta = 1$. These processes then have probability generating functions

$$\mathcal{P}_k^\lambda(z) = \sum_{j=0}^{\infty} z^j \mathrm{P}\left(Y_k^\lambda = j\right) \quad \text{and} \quad \mathcal{P}_k^\eta(z) = \sum_{j=0}^{\infty} z^j \mathrm{P}\left(Y_k^\eta = j\right),$$

that are given by the recursions $\mathcal{P}^{\lambda}_{k+1}(z) = \mathcal{P}_{X^{\lambda}}\left(\mathcal{P}^{\lambda}_{k}(z)\right)$ and $\mathcal{P}^{\lambda}_{k+1}(z) = \mathcal{P}_{X^{\eta}}\left(\mathcal{P}^{\eta}_{k}(z)\right)$ with $\mathcal{P}^{\lambda}_{1}(z) = \mathcal{P}_{X^{\lambda}}(z)$ and $\mathcal{P}^{\eta}_{1}(z) = \mathcal{P}_{X^{\eta}}(z)$, where $\mathcal{P}_{X^{\lambda}}(z)$ and $\mathcal{P}_{X^{\eta}}(z)$ are the probability generating functions of $X^{(1)}_{\lambda,1}$ and $X^{(1)}_{\eta,1}$, respectively; see e.g. Section XII.5 of Feller (1957). One can then use induction to observe that

$$\mathcal{P}^{\eta}_{k}(z) = 1 - \frac{\alpha^{k}(1-z)}{\beta^{k} + \sum_{j=1}^{k} \alpha^{j}\beta^{k-j}(1-z)},$$

whereas $\mathcal{P}^{\lambda}_{k}(z) = e^{-\frac{\alpha}{\beta}\left(1-\mathcal{P}^{\lambda}_{k-1}(z)\right)}$, with $\mathcal{P}^{\lambda}_{1}(z) = e^{-\frac{\alpha}{\beta}(1-z)}$. Because of their shared offspring distribution constructions, the number of the progeny in the kth arrival generations of the Hawkes process and the ESEP are equivalent in distribution to Y^{λ}_{k} and Y^{η}_{k}, respectively. In this way, we can express $\mathcal{G}^{\lambda}$ and $\mathcal{G}^{\eta}$ as

$$\mathcal{G}^{\lambda} = \inf\{k \in \mathbb{Z}^{+} \mid Y^{\lambda}_{k} = 0\} \quad \text{and} \quad \mathcal{G}^{\eta} = \inf\{k \in \mathbb{Z}^{+} \mid Y^{\eta}_{k} = 0\}.$$

This leads us to observe that the events $\{\mathcal{G}^{\lambda} = j\}$ and $\{Y^{\lambda}_{j} = 0, Y^{\lambda}_{j-1} > 0\}$ are equivalent, as are $\{\mathcal{G}^{\eta} = j\}$ and $\{Y^{\eta}_{j} = 0, Y^{\eta}_{j-1} > 0\}$. Focusing for now on $\mathcal{G}^{\lambda}$, we have that

$$\mathrm{P}\left(Y^{\lambda}_{j} = 0, Y^{\lambda}_{j-1} > 0\right) = \sum_{i=1}^{\infty} \mathrm{P}\left(X^{(1)}_{\lambda,1} = 0\right)^{i} \mathrm{P}\left(Y^{\lambda}_{j-1} = i\right) = \mathcal{P}^{\lambda}_{j-1}\left(\mathrm{P}\left(X^{(1)}_{\lambda,1} = 0\right)\right) - \mathrm{P}\left(Y^{\lambda}_{j-1} = 0\right),$$

and since $\mathrm{P}(K = 0) = \mathcal{P}(0)$ for any non-negative discrete random variable K with probability generating function $\mathcal{P}(z)$, this yields

$$\mathrm{P}\left(\mathcal{G}^{\lambda} = j\right) = \mathcal{P}^{\lambda}_{j}(0) - \mathcal{P}^{\lambda}_{j-1}(0).$$

Using $\mathcal{P}^{\lambda}_{0}(0) = 0$, this telescoping sum now produces the stated form of the cumulative distribution function for $\mathcal{G}^{\lambda}$. By analogous arguments for $\mathcal{G}^{\eta}$, we complete the proof. □

In the following subsection we focus on the ESEP, using the insight we have now gained from branching processes to connect this process to stochastic models for preferential attachment that are popular in the Bayesian nonparametric and machine learning literatures.

4.3.2 Similarities with Preferential Attachment and Bayesian Statistics Models

In the branching process perspective of the ESEP, consider the total number of active families at one point in time. That is, across all the entities present in the system at a given time, we are interested in the number of distinct progeny to which these entities belong. As each arrival occurs, the new entity either belongs to one of the existing families, meaning that the entity is a descendant, or it forms a new family, which is to say that it is an immigrant. If the entity is joining an existing family, it is more likely to join families that have more presently active family members.

We can note that these dynamics are quite similar to the definition of the Chinese Restaurant Process (CRP), see Chapter 11 in Aldous (1985). The CRP models the successive arrival of customers to the restaurant that has infinitely many tables that each have infinitely many seats. Each arriving customer chooses which table to join based on the decisions of those before. Specifically, the nth customer to arrive joins table i with probability $\frac{c_i}{n-1+\lambda}$ or otherwise starts a new table with probability $\frac{\lambda}{n-1+\lambda}$, where c_i is the number at table i and $\lambda > 0$. As the number seated at table i grows larger, it is increasingly likely that the next customer will choose to sit at table i. In the ESEP, a new arrival at time $t \geq 0$ was generated as part of active excitement family i with probability $\frac{\alpha Q_{t,i}}{\alpha Q_t + \eta^*}$ and otherwise was an externally generated arrival with probability $\frac{\eta^*}{\alpha Q_t + \eta^*}$, where $Q_{t,i}$ is the number of active exciters in the system at time t in the ith excitement family with $Q_t = \sum_i Q_{t,i}$. By normalizing the numerator and denominator of these probabilities by $\frac{1}{\alpha}$, we see that these dynamics match the CRP almost exactly. The difference is hardly a novel idea for restaurants – in the ESEP diners even-

tually leave. This departure then decreases the number of customers at the table, making it less attractive to the next person to arrive.

In addition to being an intriguing stochastic model, the CRP is also of interest for Bayesian statistics and machine learning through its connection to Bayesian nonparametric mixture models, specifically Dirichlet process mixtures. By consequence, the CRP then also has commonality with urn models and models for preferential attachment, see e.g. Blackwell et al. (1973). The CRP is also established enough to have its own generalizations, such as the distance dependent CRP in Blei and Frazier (2011), in which the probability a customer joins a table is dependent on a distance metric, and the recurrent CRP in Ahmed and Xing (2008), in which the restaurant closes at the end of each day forcing all of that day's customers to simultaneously depart. Drawing inspiration from the CRP and from the branching process perspectives of the ESEP, we investigate the distribution of the number of active families in the ESEP. Equivalently stated, this is the number of active tables in a continuous time CRP in which customers leave after their exponentially distributed meal durations. To begin, we first find the expected amount of time until a newly formed table becomes empty.

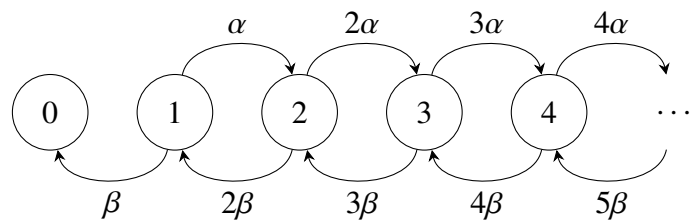
Proposition 4.3.5. *Suppose that an ESEP receives an initial arrival at time 0. Let X_t be the number of entities in the system at time $t \geq 0$ that are progeny of the initial arrival and let τ be a stopping time such that $\tau = \inf\{t \geq 0 \mid X_t = 0\}$. Then, the expected value of τ is*

$$\mathrm{E}[\tau] = \frac{1}{\alpha} \log\left(\frac{\beta}{\beta - \alpha}\right), \tag{4.14}$$

where $\alpha > 0$ is the intensity jump size and $\beta > \alpha$ is the expiration rate.

Proof. To observe this, we note that X_t can be viewed as the state of an absorbing continuous time Markov chain on the non-negative integers. State 0 is the single

absorbing state and in any other state j the two possible transitions are to $j+1$ at rate αj and to $j-1$ at rate μj, as visualized below.



Then, τ is the time of absorption into state 0 when starting in state 1 and so $\mathrm{E}[\tau]$ can be calculated by standard first step analysis approaches, yielding

$$\mathrm{E}[\tau] = \sum_{i=1}^{\infty} \frac{1}{\alpha i} \prod_{j=1}^{i} \frac{\alpha j}{\beta j} = \frac{1}{\alpha} \sum_{i=1}^{\infty} \frac{1}{i} \left(\frac{\alpha}{\beta}\right)^{i} = \frac{1}{\alpha} \log\left(\frac{1}{1-\frac{\alpha}{\beta}}\right),$$

and this simplifies to the stated result. □

Proposition 4.3.5 gives the expectation of the total time of an excitement family is active in the system. Using this, in Proposition 4.3.6 we now employ a classical queueing theory result to find the exact distribution of the number of active families simultaneously in the system in steady-state.

Proposition 4.3.6. *Let B be the number of distinct excitement families that have progeny active in the system in steady-state of an ESEP with baseline intensity $\eta^* > 0$, intensity jump $\alpha > 0$, and expiration rate $\beta > \alpha$. Then, $B \sim \mathrm{Pois}\left(\frac{\eta^*}{\alpha} \log\left(\frac{\beta}{\beta-\alpha}\right)\right)$.*

Proof. We first note that new excitement families are started when a baseline-generated arrival occurs, which follows a Poisson process with rate ν^*. The duration excitement family's time in system then has mean given by Proposition 4.3.5. Because there is no limitation on the number of possible families in the system at once, this is equivalent to an infinite server queue with Poisson

process arrivals and generally distributed service, an $M/G/\infty$ queue in Kendall notation. This process is known to have Poisson distributed steady-state distribution, see e.g. Eick et al. (1993), with mean given by the product of the arrival rate and the mean service duration, which yields the stated form for B. □

An interesting consequence of the number of active families being Poisson distributed and the total number in system being negative binomially distributed is that it suggests that the number of simultaneously active family members is logarithmically distributed. We observe this via the known compound Poisson representation of the negative binomial distribution Willmot (1986). For $B \sim \text{Pois}\left(\frac{\eta^*}{\alpha}\log\left(\frac{\beta}{\beta-\alpha}\right)\right)$, $Q \sim \text{NegBin}\left(\frac{\alpha}{\beta}, \frac{\eta^*}{\alpha}\right)$, and $L_i \overset{\text{iid}}{\sim} \text{Log}\left(\frac{\alpha}{\beta}\right)$, then one can observe that

$$Q \overset{D}{=} \sum_{i=1}^{B} L_i,$$

where $\mathrm{P}(L_1 = k) = \left(\frac{\alpha}{\beta}\right)^k \left(k\log\left(\frac{\beta}{\beta-\alpha}\right)\right)^{-1}$ for all $k \in \mathbb{Z}^+$. Thus, the idea that the number of active members of each family is logarithmically distributed follows from the fact that this is a sum of positive integer valued random variables, of which there are as many as there are active families, and this sum is equal to the total number in system.

4.3.3 Connections to Epidemic Models

As a final observation regarding the ESEP and its connections to other stochastic models, consider disease spread. As we discussed in the introduction to this chapter, when a person becomes sick with a contagious disease she increases the rate of new infection through her contact with others. Furthermore when a person recovers from a disease such as the flu, she is no longer contagious and

thus she no longer contributes to the rate of disease spread. While we have discussed in the introduction that this scenario has the hallmarks of self-excitement qualitatively, a classic model for studying this phenomenon is the Susceptible-Infected-Susceptible (SIS) process.

In the SIS model there is a finite population of $N \in \mathbb{Z}^+$ individuals. Each individual takes on one of two states, either infected or susceptible. Let I_t be the number infected at time $t \geq 0$ and S_t be the number susceptible. In the continuous time stochastic SIS model, each infected individual recovers after an exponentially distributed duration of the illness. Once a person recovers from the disease, she becomes susceptible again. Because there is a finite population, the rate of new infection depends on both the number infected and the number susceptible; a new person falls ill at a rate proportional to $I_t \cdot \frac{S_t}{N}$. Because this CTMC would be absorbed into state $I_t = 0$, it is common to include an exogenous infection rate proportional to just $\frac{S_t}{N}$. We will refer to this model as the stochastic SIS with exogenous infections, and Figure 4.3 shows rate diagram for the transitions from infected to susceptible and from susceptible to infected. For the sake of comparison, we set the exogenous infection rate as η^*, the epidemic infection rate as α, and the recovery rate as β.

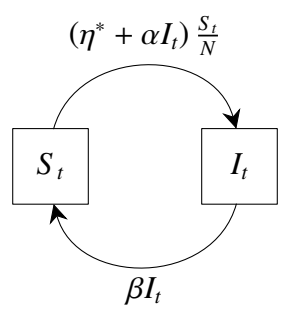


Figure 4.3: Stochastic SIS model with exogenous infections

One can note that there are immediate similarities between this process and

the ESEP. That is, new infections increase the infection rate while recoveries decrease it, and infections can be the result of either external or internal stimuli. However, the primary difference between these two models is that the SIS process has a finite population, whereas the ESEP does not. In Proposition 4.3.7 we find that as this population size grows large the difference between these models fades, yielding that the distribution of the number infected in the exogenously driven SIS model converges to the distribution of the queue length in the ESEP.

Proposition 4.3.7. *Let I_t be the number of infected individuals at time $t \geq 0$ in an exogenously driven stochastic SIS model with population size $N \in \mathbb{Z}^+$, exogenous infection rate $\eta^* > 0$, epidemic infection rate $\alpha > 0$, and recovery rate $\beta > 0$. Then, as $N \to \infty$*

$$I_t \overset{D}{\Longrightarrow} Q_t,$$

where Q_t is the active number in system at time t for an ESEP with baseline intensity η^*, intensity jump α, and expiration rate β.

Proof. Because the SIS model is a Markov process, one can use the infinitesimal generator approach to find a time derivative for the moment generating function of the number of infected individuals at time $t \geq 0$. Thus, by noting that $S_t = N - I_t$ we have that

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[e^{\theta I_t}\right] &= \mathrm{E}\left[\frac{\alpha I_t S_t}{N}\left(e^{\theta}-1\right)e^{\theta I_t} + \beta I\left(e^{-\theta}-1\right)e^{\theta I_t} + \frac{\eta^* S_t}{N}\left(e^{\theta}-1\right)e^{\theta I_t}\right] \\ &= \mathrm{E}\left[\frac{\alpha I_t(N-I_t)}{N}\left(e^{\theta}-1\right)e^{\theta I_t}\right] + \mathrm{E}\left[\beta I_t\left(e^{-\theta}-1\right)e^{\theta I_t}\right] + \mathrm{E}\left[\frac{\eta^*(N-I_t)}{N}\left(e^{\theta}-1\right)e^{\theta I_t}\right],\end{aligned}$$

which we can re-express in partial differential equation form as

$$\begin{aligned}\frac{\partial \mathrm{E}\left[e^{\theta I_t}\right]}{\partial t} &= \left(\alpha\left(e^{\theta}-1\right) + \beta\left(e^{-\theta}-1\right) - \frac{\eta^*}{N}\left(e^{\theta}-1\right)\right)\frac{\partial \mathrm{E}\left[e^{\theta I_t}\right]}{\partial \theta} - \frac{\alpha}{N}\left(e^{\theta}-1\right)\frac{\partial^2 \mathrm{E}\left[e^{\theta I_t}\right]}{\partial \theta^2} \\ &\quad + \eta^*\left(e^{\theta}-1\right)\mathrm{E}\left[e^{\theta I_t}\right].\end{aligned}$$

Now as the population size $N \to \infty$, this converges to

$$\frac{\partial \mathrm{E}\left[e^{\theta I_t}\right]}{\partial t} = \left(\alpha\left(e^{\theta}-1\right)+\beta\left(e^{-\theta}-1\right)\right)\frac{\partial \mathrm{E}\left[e^{\theta I_t}\right]}{\partial \theta} + \eta^*\left(e^{\theta}-1\right)\mathrm{E}\left[e^{\theta I_t}\right],$$

which we can recognize as the partial differential equation for the moment generating function of the ESEP through its own infinitesimal generator. □

As a demonstration of this convergence, we plot the empirical steady-state distribution of the SIS process for increasing population size below in Figure 4.4. Note that in this example the distributions appear fairly close for populations of size $N = 1,000$. On the scale of the populations of cities or even some larger high schools, this is quite small. At a medium university size of $N = 10,000$, the distributions are essentially indistinguishable.

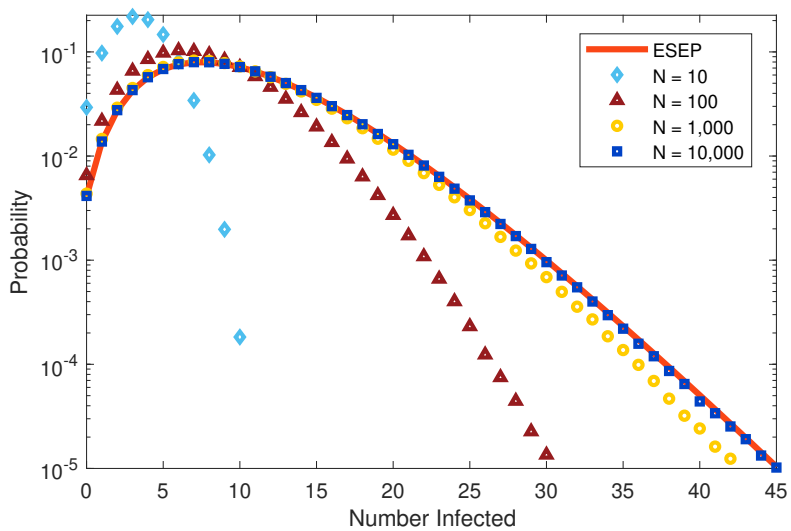


Figure 4.4: Steady-state distribution of the number infected in the exogenously driven SIS model for increasing population size N, where $\eta^* = 10$, $\alpha = 2$, and $\beta = 3$.

We would be remiss if we did not note that connections from epidemic models to birth-death processes are not new. For example, Ball (1983) demonstrated

that epidemic models converge to birth-death processes, and Singh and Myers (2014) even noted that the exogenously driven Susceptible-Infected-Recovered (SIR) model – that is, people cannot become re-infected – converges to a linear birth-death-immigration process; however, these works did not outright form connections to self-exciting processes. In Rizoiu et al. (2018), the similarities between the Hawkes process and the SIR process are shown and formal connections are made, although this is through a generalization of the Hawkes process defined on a finite population rather than through increasing the epidemic model population size. Regardless, the topics considered in these prior works serve to expand the practical relevance of the ESEP, as they note that these epidemic models are also of use outside of public health. For example, the contagious nature of these models has also been used to study topics like product adoption, idea spread, and social influence. These all also naturally relate to the concept of self-excitement, and in Proposition 4.3.7 we observe that this connection can be formalized.

These connections also take on an added importance in the contemporaneous context of the COVID-19 public health crisis. It is worth noting that the convergence of distributions in Proposition 4.3.7 does not require our commonly assumed stability condition $\beta > \alpha$. Thus, this convergence also covers potentially pandemic scenarios in which $\alpha \geq \beta$. In such settings, one can quickly observe that the mean number infected by time t is such that $\mathrm{E}[N_t] \in O(e^{(\alpha-\beta)t})$, reproducing the exponential growth exhibited by this novel coronavirus in these early stages. The ESEP arrival process then captures the times of new infections, some portion of which may then be used as the arrival process to a queueing model for the cases that require hospitalization.

4.4 Constructing Eternal Self-Excitement from Ephemeral Self-Excitement

Thus far we have exclusively considered a Markovian model for ephemeral self-excitement. However, just as the Hawkes process need not be Markovian, we do not have to restrict ourselves to settings in which the Markov property exists. Recall from Subsection 5.3.2 that the general definition from Hawkes (1971) described the intensity as

$$\lambda_t = \lambda^* + \int_{-\infty}^{t} g(t-u)\mathrm{d}N_{u,\lambda} = \lambda^* + \sum_{i=1}^{N_t} g(t-A_i),$$

where $g : \mathbb{R}^+ \to \mathbb{R}^+$ and $\int_0^\infty g(x)\mathrm{d}x < 1$ for stability. One could also consider a marked Hawkes process that draws jump sizes from a sequence of i.i.d. positive random variables $\{M_i \mid i \in \mathbb{Z}^+\}$, in which case the summation form of intensity would instead be expressed

$$\lambda_t = \lambda^* + \sum_{i=1}^{N_t} M_i g(t-A_i).$$

The ESEP model has provided us a natural comparison for the popular Markovian case where $g(x) = \alpha e^{-\beta x}$ (with no marks), but let us now consider more general excitation kernels and jump sizes. To do so, we will make two main generalizations while preserving the other key elements of the ESEP, such as the affine relationship between the intensity and the number of active exciters. First, we will change the activity duration to be a general distribution, mimicking the general excitation kernel. Secondly, we also change from solitary arrivals to batch arrivals, meaning groups of events that occur simultaneously at each arrival epoch. The size of these batches may be drawn from an i.i.d. sequence of positive integer random variables, so we will use the parameter n to represent

the relative size of this batch through the mean of the batch size distribution. This leads us to the nth *general ephemerally self-exciting process* (n-GESEP), defined now in Definition 4.4.1.

Definition 4.4.1 (n^{th} general ephemerally self-exciting process)**.** For times $t \geq 0$, a baseline intensity $\eta^* > 0$, cumulative distribution function $G : \mathbb{R}^+ \to [0, 1]$, i.i.d. sequence of positive random variables $\{B_i \mid i \in \mathbb{Z}^+\}$, and $n \in \mathbb{Z}^+$ such that $\mathrm{E}[B_i] \in O(n)$, let $N_t(n)$ be a counting process for the arrival epochs occurring according to the stochastic intensity $\eta_t(n)$, defined such that

$$\eta_t(n) = \eta^* + \frac{\alpha}{n} Q_t(n), \tag{4.15}$$

where $Q_t(n)$ is incremented by $B_i(n)$ at the ith arrival epoch in $N_t(n)$ and then is depleted at unit down-jumps at the expiration of each individual arrival's activity duration, which are i.i.d. draws from the distribution $G(\cdot)$ across all batches and all epochs. Then, we say that $(\eta_t(n), N_t(n))$ is the nth *general ephemerally self-exciting process* (n-GESEP).

It is important to note that in this definition, a batch of size n occurring at the current time would increase the present arrival rate by α. However, each of these n activity durations are mutually independent. Thus, despite the common arrival epoch, each expiration should cause an instantaneous decrease of just α/n if the activity duration distribution is continuous. It is also worth noting that this n-GESEP model encapsulates the simple generalization of the ESEP with general service durations, as this is given by $(\eta_t(1), N_t(1))$ with $\mathrm{P}(B_1(1) = 1)$.

Just as it can often be beneficial to think of the length of an infinite server queue as a sum over all customers that remain in service, it will be quite useful for us to think about the number active within the n-GESEP as the sum of all

arrivals that have not yet expired. For A_i as the ith arrival epoch and $S_{i,j}$ as the jth activity duration within the ith batch, this can be expressed

$$Q_t(n) = \sum_{i=1}^{N_t(n)} \sum_{j=1}^{B_i} \mathbf{1}\{t < A_i + S_{i,j}\},$$

or equivalently for the intensity,

$$\eta_t = \eta^* + \frac{\alpha}{n} \sum_{i=1}^{N_t(n)} \sum_{j=1}^{B_i} \{t < A_i + S_{i,j}\}.$$

A nice consequence of this representation is that the ephemerality is at the forefront. Because the event within the indicator is that the current time t is less than the sum of a given arrival time and activity duration, this indicator counts whether that particular excitement is currently active. Thus, this indicator will switch from 1 to 0 when the present time passes a given expiration time, causing the intensity to drop by α/n. From this point forward, the excitement brought by this particular arrival no longer has a direct effect on the system.

We will now use the n-GESEP to establish a main result of this work, in which we connect this general form of ephemeral self-excitement to general forms of the traditional, eternal notion of self-excitement. In Theorem 4.4.1, we prove a scaling limit that incorporates random batch distributions and general service to construct marked, general decay Hawkes processes. We refer to this limit as a "batch scaling," as we are letting the relative batch size n grow large, which simultaneously shrinks the size of the excitement generated by each individual entity within a batch of arrivals. Thus, while the effect of one individual exciter shrinks, the collective effect of a batch arrival remains fixed. In the limit, this provides an alternate construction of the Hawkes process.

Theorem 4.4.1. *For $t \geq 0$ and $n \in \mathbb{Z}^+$, let $(\eta_t(n), N_t(n))$ be a n-GESEP with baseline intensity $\eta^* > 0$, intensity jump size $\alpha > 0$, activity duration CDF $G(\cdot)$, and i.i.d. batch*

size sequence of non-negative, discrete random variables $\{B_i \mid i \in \mathbb{Z}^+\}$. *Additionally, let* $\eta_0(n) = \eta^*$, *i.e.* $Q_0(n) = 0$. *Then, for all* $t \geq 0$, *this n-GESEP process is such that*

$$\eta_t(n) \overset{D}{\Longrightarrow} \lambda_t \text{ and } N_t(n) \overset{D}{\Longrightarrow} N_{t,\lambda} \tag{4.16}$$

as $n \to \infty$, *where* $(\lambda_t, N_{t,\lambda})$ *is the general Hawkes process intensity and counting process pair such that*

$$\lambda_t = \eta^* + \sum_{i=1}^{N_{t,\lambda}} M_i \bar{G}(t - A_i), \tag{4.17}$$

where $\{A_i \mid i \in \mathbb{Z}^+\}$ *are the Hawkes process arrival epochs,* $\bar{G}(x) = 1 - G(x)$ *for all* $x \geq 0$, *and* $\{M_i \mid i \in \mathbb{Z}^+\}$ *is an i.i.d. sequence of positive real random variables such that* $\alpha B_1/n \overset{D}{\Longrightarrow} M_1$ *and* $B_1/n^2 \xrightarrow{p} 0$.

Proof. We will organize the proof into two parts. Each part is oriented around the process arrival times, as these fully determine the sample path of the Hawkes process. We will first show through induction that the distributions of inter-arrival times converge. Then, we will demonstrate that given the same arrival times, the dynamics of the processes converge.

To begin, let A_i^η for $i \in \mathbb{Z}^+$ be the time of the ith arrival in the n-GESEP and similarly let A_i^λ be the ith arrival time for the Hawkes process. We start with the base case: for the time of the first arrival, we can note that for all n-GESEP models,

$$\mathrm{P}\left(A_1^\eta > x\right) = e^{-\eta^* x},$$

as $Q_0(n) = 0$ and thus the first arrival is driven by the external baseline rate. Likewise for the Hawkes process, since Equation 4.17 implies that $\lambda_t = \nu^*$ for $0 \leq t < A_1^\lambda$, we can see that

$$\mathrm{P}\left(A_1^\lambda > x\right) = e^{-\eta^* x},$$

and thus $\mathrm{P}\left(A_1^\lambda > x\right) = \mathrm{P}\left(A_1^\eta > x\right)$. As an inductive hypothesis, we now assume that $\{A_1^\eta, \ldots, A_k^\eta\}$ converge in joint and marginal distributions to $\{A_1^\lambda, \ldots, A_k^\lambda\}$ where $k \in \mathbb{Z}^+$. Now, for the Hawkes process we can observe that

$$\mathrm{P}_k\left(A_{k+1}^\lambda - A_k^\lambda > x\right) := \mathrm{P}\left(A_{k+1}^\lambda - A_k^\lambda > x \mid \{A_1^\lambda, \ldots, A_k^\lambda\}\right) = \mathrm{E}_k\left[e^{-\int_0^x \lambda_{A_k^\lambda + t}\mathrm{d}t}\right],$$

because when conditioned on the arrival history, the Hawkes process behaves like an imhogoneous Poisson process until the next arrival occurs. Using Equation 4.17, we can express this as

$$\mathrm{E}_k\left[e^{-\int_0^x \lambda_{A_k^\lambda + t}\mathrm{d}t}\right] = e^{-\eta^* x}\mathrm{E}_k\left[e^{-\int_0^x \sum_{i=1}^k M_i \bar{G}\left(A_k^\lambda - A_i^\lambda + t\right)\mathrm{d}t}\right] = e^{-\eta^* x}\prod_{i=1}^k \mathrm{E}_k\left[e^{-M_i \int_0^x \bar{G}\left(A_k^\lambda - A_i^\lambda + t\right)\mathrm{d}t}\right].$$

Turning to the ESEP epochs, we define $N_{i,j}^\eta((t, t+s])$ as the number of arrivals on the time interval $(t, t+s]$ that are generated by the excitement caused by the jth entity within the ith batch. Furthermore, let $N_*^\eta((t, t+s])$ be the number of arrivals on $(t, t+s]$ that are generated by the external, baseline rate η^*. Then, using this notation we have that

$$\begin{aligned}
\mathrm{P}_k\left(A_{k+1}^\eta - A_k^\eta > x\right) &:= \mathrm{P}\left(A_{k+1}^\eta - A_k^\eta > x \mid \{A_1^\eta, \ldots, A_k^\eta\}\right) \\
&= \mathrm{P}_k\left(\bigcap_{i=1}^k \bigcap_{j=1}^{B_i} \left\{N_{i,j}^\eta\left((A_k^\eta, A_k^\eta + x]\right) = 0\right\} \cap \left\{N_*^\eta\left((A_k^\eta, A_k^\eta + x]\right) = 0\right\}\right) \\
&= \mathrm{E}_k\left[\prod_{i=1}^k \prod_{j=1}^{B_i} \mathbf{1}\left\{N_{i,j}^\eta\left((A_k^\eta, A_k^\eta + x]\right) = 0\right\} \mathbf{1}\left\{N_*^\eta\left((A_k^\eta, A_k^\eta + x]\right) = 0\right\}\right].
\end{aligned}$$

From the independence of each of these arrival processes, we can move the probability for no arrivals in the external arrival process and the product over i outside of the expectation to receive

$$\begin{aligned}
&\mathrm{E}_k\left[\prod_{i=1}^k \prod_{j=1}^{B_i} \mathbf{1}\left\{N_{i,j}^\eta\left((A_k^\eta, A_k^\eta + x]\right) = 0\right\} \mathbf{1}\left\{N_*^\eta\left((A_k^\eta, A_k^\eta + x]\right) = 0\right\}\right] \\
&= e^{-\eta^* x}\prod_{i=1}^k \mathrm{E}_k\left[\prod_{j=1}^{B_i} \mathbf{1}\left\{N_{i,j}^\eta\left((A_k^\eta, A_k^\eta + x]\right) = 0\right\}\right].
\end{aligned}$$

Consider an arbitrary entity, say the jth entity in the ith batch. Let $S_{i,j}$ be its service duration. If this entity has departed from the queue before A_k^η, then it cannot generate further arrivals and thus

$$\mathrm{P}_k\left(N_{i,j}^\eta\left((A_k^\eta, A_k^\eta + x]\right) = 0 \mid S_{i,j} \leq A_k^\eta - A_i^\eta\right) = 1.$$

Likewise, if it does not depart until after $A_k^\eta + x$, then the probability that it generates an arrival on $(A_k^\eta, A_k^\eta + x]$ is

$$\mathrm{P}_k\left(N_{i,j}^\eta\left((A_k^\eta, A_k^\eta + x]\right) = 0 \mid S_{i,j} \geq A_k^\eta - A_i^\eta + x\right) = e^{-\frac{\alpha}{n}x}.$$

Finally, if the entity departs in the interval $(A_k^\eta, A_k^\eta + x]$, the probability it generates an arrival before departing is

$$\mathrm{P}_k\left(N_{i,j}^\eta\left((A_k^\eta, A_k^\eta + x]\right) = 0 \mid S_{i,j} = A_k^\eta - A_i^\eta + z\right) = e^{-\frac{\alpha}{n}z},$$

where $0 < z < x$. Therefore through conditioning on each entity's service duration, we have that

$$\begin{aligned}
&e^{-\eta^* x}\prod_{i=1}^{k}\mathrm{E}_k\left[\prod_{j=1}^{B_i}\mathbf{1}\left\{N_{i,j}^\eta\left((A_k^\eta, A_k^\eta + x]\right) = 0\right\}\right] \\
&\quad = e^{-\eta^* x}\prod_{i=1}^{k}\mathrm{E}_k\left[\prod_{j=1}^{B_i}\left(G(A_k^\eta - A_i^\eta) + e^{-\frac{\alpha}{n}x}\bar{G}(A_k^\eta - A_i^\eta + x) + \int_0^x e^{-\frac{\alpha}{n}z}g(A_k^\eta - A_i^\eta + z)\mathrm{d}z\right)\right],
\end{aligned}$$

where $g(\cdot)$ is the density corresponding to $G(\cdot)$. Since the term inside the inner product does not depend on the specific entity within a batch but rather just the batch itself, we can evaluate this inside the expectation as

$$\begin{aligned}
&e^{-\eta^* x}\prod_{i=1}^{k}\mathrm{E}_k\left[\prod_{j=1}^{B_i}\left(G(A_k^\eta - A_i^\eta) + e^{-\frac{\alpha}{n}x}\bar{G}(A_k^\eta - A_i^\eta + x) + \int_0^x e^{-\frac{\alpha}{n}z}g(A_k^\eta - A_i^\eta + z)\mathrm{d}z\right)\right] \\
&\quad = e^{-\eta^* x}\prod_{i=1}^{k}\mathrm{E}_k\left[\left(G(A_k^\eta - A_i^\eta) + e^{-\frac{\alpha}{n}x}\bar{G}(A_k^\eta - A_i^\eta + x) + \int_0^x e^{-\frac{\alpha}{n}z}g(A_k^\eta - A_i^\eta + z)\mathrm{d}z\right)^{B_i}\right].
\end{aligned}$$

Since the base term of this exponent is deterministic, we will simplify it as follows. Using integration by parts on $\int_0^x e^{-\frac{\alpha}{n}z}g(A_k^\eta - A_i^\eta + z)\mathrm{d}z$ and expanding

$\bar{G}(x) = 1 - G(x)$, this simplifies to

$$G(A_k^\eta - A_i^\eta) + e^{-\frac{\alpha}{n}x}\bar{G}(A_k^\eta - A_i^\eta + x) + \int_0^x e^{-\frac{\alpha}{n}z} g(A_k^\eta - A_i^\eta + z)\mathrm{d}z = e^{-\frac{\alpha}{n}x} + \frac{\alpha}{n}\int_0^x e^{-\frac{\alpha}{n}z} G(A_k^\eta - A_i^\eta + z)\mathrm{d}z.$$

If we express $e^{-\frac{\alpha}{n}x}$ in integral form via $e^{-\frac{\alpha}{n}x} = 1 - \frac{\alpha}{n}\int_0^x e^{-\frac{\alpha}{n}z}\mathrm{d}z$, we can further simplify this expression of the base to

$$e^{-\frac{\alpha}{n}x} + \frac{\alpha}{n}\int_0^x e^{-\frac{\alpha}{n}z} G(A_k^\eta - A_i^\eta + z)\mathrm{d}z = 1 - \frac{\alpha}{n}\int_0^x e^{-\frac{\alpha}{n}z}\bar{G}(A_k^\eta - A_i^\eta + z)\mathrm{d}z.$$

This form makes it quick to observe that this base term is at most 1. Thus we are justified in taking the expectation of this term raised to B_i, since that is equivalent to the probability generating function of the batch size and this exists for all discrete random variables when evaluated on values less than or equal to 1 in absolute value. Returning to this expectation, we first note that for all x, rearranging the Taylor expansion of e^x produces

$$1 + x = e^x - \sum_{j=2}^{\infty}\frac{x^j}{j!} = e^x\left(1 - e^{-x}\sum_{j=2}^{\infty}\frac{x^j}{j!}\right) = e^{x + \log\left(1 - e^{-x}\sum_{j=2}^{\infty}\frac{x^j}{j!}\right)}.$$

Thus we re-express the expectation in exponential function form as

$$\begin{aligned} &e^{-\eta^* x}\prod_{i=1}^{k}\mathrm{E}_k\left[\left(G(A_k^\eta - A_i^\eta) + e^{-\frac{\alpha}{n}x}\bar{G}(A_k^\eta - A_i^\eta) + \int_0^x e^{-\frac{\alpha}{n}z} g(A_k^\eta - A_i^\eta + z)\mathrm{d}z\right)^{B_i}\right] \\ &= e^{-\eta^* x}\prod_{i=1}^{k}\mathrm{E}_k\left[e^{-\frac{\alpha}{n}B_i\int_0^x e^{-\frac{\alpha}{n}z}\bar{G}(A_k^\eta - A_i^\eta + z)\mathrm{d}z + O\left(\frac{B_i}{n^2}\right)}\right]. \end{aligned}$$

Through use of a Taylor expansion on $e^{-\frac{\alpha}{n}z}$ and absorbing higher terms into the $O\left(\frac{B_i}{n^2}\right)$ notation, we can further simplify to

$$e^{-\eta^* x}\prod_{i=1}^{k}\mathrm{E}_k\left[e^{-\frac{\alpha}{n}B_i\int_0^x e^{-\frac{\alpha}{n}z}\bar{G}(A_k^\eta - A_i^\eta + z)\mathrm{d}z + O\left(\frac{B_i}{n^2}\right)}\right] = e^{-\eta^* x}\prod_{i=1}^{k}\mathrm{E}_k\left[e^{-\frac{\alpha}{n}B_i\int_0^x \bar{G}(A_k^\eta - A_i^\eta + z)\mathrm{d}z + O\left(\frac{B_i}{n^2}\right)}\right].$$

We can now take the limit as $n \to \infty$ and observe that

$$e^{-\eta^* x}\prod_{i=1}^{k}\mathrm{E}_k\left[e^{-\frac{\alpha}{n}B_i\int_0^x \bar{G}(A_k^\eta - A_i^\eta + z)\mathrm{d}z + O\left(\frac{B_i}{n^2}\right)}\right] \longrightarrow e^{-\eta^* x}\prod_{i=1}^{k}\mathrm{E}_k\left[e^{-M_i\int_0^x \bar{G}(A_k^\eta - A_i^\eta + z)\mathrm{d}z}\right],$$

as we have that $\frac{\alpha}{n}B_1 \overset{D}{\Longrightarrow} M_1$ and $\frac{B_i}{n^2} \overset{D}{\Longrightarrow} 0$. This is now equal to the Hawkes process inter-arrival probability $\mathrm{P}_k\left(A^\lambda_{k+1} - A^\lambda_k > x\right)$. Hence by induction and total probability the arrival times converge, completing the first part of the proof.

For the second part of the proof, we now show that the dynamics of the processes converge when we condition on having the same fixed arrival times, which we now denote $\{A_i \mid i \in \mathbb{Z}^+\}$ for both processes. Since $N_t(n)$ is defined as the counting process of arrival epcochs rather than total number of arrivals, $N_t(n) = N_{t,\lambda}$ for all n and all t. We now treat the intensity in two cases, the jump at arrivals and the dynamics between these times. For the first case, we take $k \in \mathbb{Z}^+$ and let $\lambda_{A_{k^-}} = \inf_{A_{k-1} \le t < A_k} \lambda_t$ and $\eta_{A_{k^-}}(n) = \inf_{A_{k-1} \le t < A_k} \eta_t(n)$ for all n, where $A_0 = 0$. Then, the jump in the n-GESEP intensity at the kth jump is such that

$$\eta_{A_k}(n) - \eta_{A_{k^-}}(n) = \frac{\alpha}{n}B_k \overset{D}{\Longrightarrow} M_k = \lambda_{A_k} - \lambda_{A_{k^-}},$$

as $n \to \infty$. For the behavior between arrival times we first note that for S_j independent and distributed with CDF $G(\cdot)$ for all $j \in \mathbb{Z}^+$, the probability generating function of $\frac{1}{n}\sum_{j=1}^{B_1} \mathbf{1}\{y < S_j\}$ is

$$\mathrm{E}\left[z^{\frac{1}{n}\sum_{j=1}^{B_1} \mathbf{1}\{y<S_j\}}\right] = \mathrm{E}\left[\left(G(y) + \bar{G}(y)z^{\frac{1}{n}}\right)^{B_1}\right] = \mathrm{E}\left[\left(1 - \bar{G}(y)\left(1 - e^{\frac{1}{n}\log z}\right)\right)^{B_1}\right],$$

and by a Taylor expansion approach similar to what we used in the proof's first part, we can see that

$$\mathrm{E}\left[\left(1 - \bar{G}(y)\left(1 - e^{\frac{1}{n}\log z}\right)\right)^{B_1}\right] = \mathrm{E}\left[e^{-B_1\bar{G}(y)\left(1-e^{\frac{1}{n}\log z}\right)+O\left(\frac{B_1}{n^2}\right)}\right] = \mathrm{E}\left[e^{\frac{1}{n}B_1\bar{G}(y)\log(z)+O\left(\frac{B_1}{n^2}\right)}\right].$$

Taking the limit as $n \to \infty$, this yields $\mathrm{E}\left[z^{\frac{1}{n}\sum_{j=1}^{B_1} \mathbf{1}\{y<S_j\}}\right] \longrightarrow \mathrm{E}\left[z^{\bar{G}(y)M_1}\right]$, which is to say that

$$\frac{1}{n}\sum_{j=1}^{B_1} \mathbf{1}\{y < S_j\} \overset{D}{\Longrightarrow} \bar{G}(y)M_1.$$

Using this, we can now see that for $k \in \mathbb{Z}^+$ and $0 \leq x < A_{k+1} - A_k$, the intensity of the n-GESEP satisfies

$$\eta_{A_k+x}(n) = \eta^* + \frac{\alpha}{n} \sum_{i=1}^{k} \sum_{j=1}^{B_i} \mathbf{1}\{A_k + x < A_i + S_{i,j}\} \overset{D}{\Longrightarrow} \eta^* + \sum_{i=1}^{k} M_i \bar{G}(A_k - A_i + x) = \lambda_{A_k+x},$$

as $n \to \infty$. Thus, both the jump sizes of $\eta_t(n)$ and the behavior of $\eta_t(n)$ between jumps converge to that of λ_t, completing the proof. □

For an empirical demonstrate of this convergence, in Figure 4.5 we plot cumulative distribution functions for the intensity of the Markovian n-GESEP across multiple batch sizes and compare them to the empirical distribution of the Markovian Hawkes process. As one can see, in each of the two parameter settings with $n = 8$, the distribution of the n-GESEP intensity is quite close to that of the Hawkes intensity. Loosely speaking, one can also note that the convergence appears to be faster in the case displayed on the right hand side, which has larger parameter values. In future work, it will be of interest to consider the rate of convergence for this batch-scaling and how those depend on the process parameters.

As a reference, we list the components of the ephemerally self-exciting models and their corresponding limiting quantities in the general Hawkes process below in Table 4.1. We can note that because the limiting excitation kernel given in Theorem 4.4.1 is a complementary cumulative distribution function it is exclusively non-increasing, meaning that the excitement after each arrival immediately decays. It can be observed that this includes the two most popular excitation kernels, the exponential and power-law forms that we detailed in Subsection 5.3.2. However, it does not include kernels that have a "hump remote from the origin" that Hawkes mentions briefly in the original paper Hawkes (1971). If desired, this can be remedied through extension to multi-phase service in the

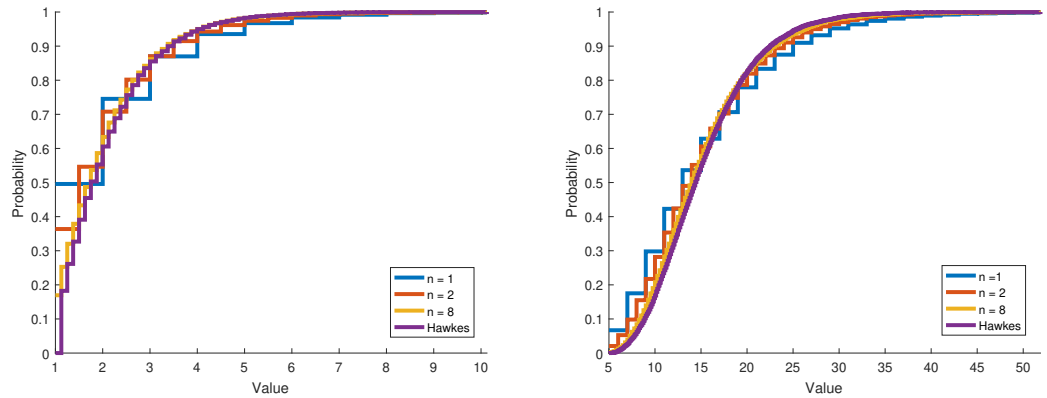


Figure 4.5: Empirical steady-state CDF of the n-GESEP intensity where $\nu^* = \alpha = 1$ and $\mu = 2$ (left); and where $\nu^* = 5$, $\alpha = 2$ and $\mu = 3$ (right), based on 10,000 replications.

n-GESEP, with the intensity defined as an affine relationship with one of the later phases.

$n \longrightarrow \infty$		
Batch	$\Longrightarrow$	Mark
Expire	$\Longrightarrow$	Decay
ESEP	$\Longrightarrow$	Hawkes

Table 4.1: Overview of convergence details in the batch-scaling of the ESEP.

Before concluding, let us remark that in addition to providing conceptual understanding into the Hawkes process itself, the alternate construction through the batch scaling in Theorem 4.4.1 is also of practical relevance in explaining the use of the Hawkes process in many application settings. For example, in biological applications such as the environmental management problem considered in Gupta et al. (2018), one of the invasive species studied may produce multiple offspring simultaneously but only for the duration of its life cycle. That is, many species give birth in litters, creating batch arrivals, but of course only reproduce during their lifetime, yielding ephemerality. Furthermore, the numerical experiments in Figure 4.5 suggest that n need not be overly large for the n-GESEP and

the Hawkes process to be comparable in distribution.

As another example, consider the spread of information on communication and social media platforms. This setting has recently been a popular application of Hawkes processes, see e.g. Du et al. (2015); Farajtabar et al. (2017); Halpin and De Boeck (2013); Rizoiu et al. (2017, 2018). When a user shares a post on these platforms, it is immediately and simultaneously dispatched to the real-time feeds of many other users, creating a batch increase of the response rate from the other users. The post then will typically only be seen on news feeds for a short period of time, as new content comes in to replace it. On top of this, social media administrators have been adopting a trend of intentionally introducing ephemerality into their platforms. For example, the expiration of posts and messages has been a defining feature of Snapchat since its inception. Facebook and Instagram have recently adopted the same behavior with "stories," and Twitter has responded in kind with the appropriately named "fleets." Just as in the case of biological offspring processes, Theorem 4.4.1 offers an explanation of why the Hawkes process has become a popular and successful model in this space. Moreover, the insights from these connections can be deepened through the additional model relationships that we have discussed in Section 4.3.

4.5 Conclusion

Time is fleeting; excitement is ephemeral. In this chapter we have introduced the *ephemerally self-exciting process* (ESEP), a point process driven by the pieces of its history that remain presently active. That is, each arrival excites the arrival rate until the expiration of its randomly drawn activity duration, at which point

its individual influence vanishes. Throughout this work we have compared ephemeral self-excitement to eternal self-excitement through contrast with the well-known Hawkes process. These comparisons include an ordering of the moments of the two processes in Proposition 4.2.3 and through study of each process's branching structure in Subsection 4.3.1. We have also used the ESEP to relate ephemeral self-excitement to other well known stochastic models, including preferential attachment, random walks, and epidemics. Finally, we have also considered a generalized model with batch arrivals and general activity duration distributions, which we refer to as the nth *general ephemerally self-exciting process* (n-GESEP). This n-GESEP model provides an alternate construction of general marked Hawkes processes through a batch scaling limit. In Theorem 4.4.1 we prove that the n-GESEP model converges to a Hawkes process as its batch arrival size grows large, in which the limiting Hawkes process has an excitation kernel matching the tail CDF of the activity duration distribution and has marks given by the scaled limit of the batch sizes. As we have discussed, this limit both provides intuition for the occurrence of self-excitement in natural phenomena and relates the Hawkes process to the other stochastic models we connected to ephemeral self-excitement.

This presents many different directions for future research. First, we have frequently campaigned in this work that our results motivate the ESEP (and by extension, the n-GESEP) as a promising model for self-excitement in its own right. This follows from its tractability for analysis and its amenability to connection with other models. Because of this promise, we believe that further exploration and application of the ESEP holds great potential. Another natural and relevant avenue would be to continue to explore the connection between ephemeral self-excitement and the other stochastic models we have discussed.

For example, one could study the connection between ESEP and epidemic models on more complex networks or with more complex dynamics. Doing so would give a point process representation for the times of infection in a more realistic epidemic model, which could be quite useful in practice for resource allocation and policy design. Similar deepened connections could also be pursued for other models such as preferential attachment. In general, we believe the concept of ephemeral self-excitement merits further theoretical exploration and detailed empirical application, both of which we look forward to pursuing.

CHAPTER 5

MATRIX CALCULATIONS FOR MOMENTS OF MARKOV PROCESSES

5.1 Introduction

In recently studying the intensity of Markovian Hawkes process, originally defined in Hawkes (1971), we have been interested in computing all the moments of this process. In surveying the literature for this process, there does not seem to be any closed form transient solutions at the fourth order or higher (see Proposition 5 of Gao and Zhu (2018c) for moments one through three), and both steady-state solutions and ordinary differential equations have only been available up to the fourth moment, see Da Fonseca and Zaatour (2014); Errais et al. (2010). Similarly, Aït-Sahalia et al. (2015) give expressions for the fourth transient moment of a self-exciting jump-diffusion model up to squared error in the length of time, and one could simplify these expressions to represent the Markovian Hawkes intensity with the same error. The standard methodology for finding moments is to differentiate the moment generating function to obtain the moments, however, this is intractable for practical reasons, see for example Errais et al. (2010). The problem of finding the moments of the Hawkes process is also the subject of the recent interesting research in Cui et al. (2019); Cui and Wu (2019), works that are concurrent and independent from this one. In Cui et al. (2019), the authors propose a new approach for calculating moments that they construct from elementary probability arguments and also relate to the in-

Contents of this chapter are, at the time of this dissertation's writing, under review for publication and are publicly available as a preprint (Daw and Pender, 2020b).

finitesimal generator. Like the infinitesimal generator, this new methodology produces differential equations that can be solved algebraically or numerically to yield the process moments, and the authors provide closed form transient expressions up to the second moment. Cui and Wu (2019) extends this methodology to cases of Gamma decay kernels. In other recent previous works, including Koops et al. (2018) and Chapter 2, the authors have identified the differential equation for an arbitrary moment of the Hawkes process, although the closed form solutions for these equations have remained elusive and prompted closer investigation. Upon inspecting the differential equation for a given moment of the Hawkes process intensity, one can notice that this expression depends on the moments of lower order. Thus, to compute a given moment one must solve a system of differential equations with size equal to the order of the moment, meaning one must at least implicitly solve for all the lower order moments first. This same pattern occurs in Cui et al. (2019). Noticing this nesting pattern leads one to wonder: what other processes have moments that follow this structure?

In this chapter, we explore this question by identifying what exactly this nesting structure is. In the sequel, we will define a novel sequence of matrices that captures this pattern. Just as Matryoshka dolls – the traditional Russian nesting figurines – stack inside of one another, these matrices are characterized by their encapsulation of their predecessors in the sequence. Hence, we refer to this sequence as **Matryoshkan matrices**. As we will show, these matrices can be used to describe the linear system of differential equations that arise in solving for the moments of the Hawkes process, as well as the moments of a large class of other Markov processes. In fact, the only assumption we make on these processes is that their moments satisfy differential equations that do not depend on any higher order moments. As we will demonstrate through detailed examples,

this includes a wide variety of popular stochastic processes, such as Itô diffusions and shot noise processes. By utilizing this nesting structure we are able to solve for the moments of these processes in closed form. By comparison to traditional methods of solving these systems of differential equations numerically, the advantage of the approach introduced herein is the fact that the moments can be computed at a specific point in time rather than on a path through time. This yields a methodology that is both efficient and precise.

This methodology also has the potential to be quite relevant in practice. Of course, these techniques can be used to efficiently calculate the commonly used first four moments, thus obtaining the mean, variance, skewness, and kurtosis. Moreover though, let us note that the higher moment calculations are also of practical use. For example, these higher moments can be used in Markov-style concentration inequalities, as the higher order should improve the accuracy of the tail bounds. To that end, one can also use the vector of moments to approximate generating functions such as the moment generating function of Laplace-Stieltjes transforms. This could then be used to characterize the stationary distribution of the process, for example, or to provide approximate calculations of quantities such as the cumulative distribution function through transform methods. The calculation of moments can also be highly relevant for many applications in mathematical finance. For example, these techniques may hold great potential for polynomial processes, see for example Filipović and Larsson (2019). One could also expect this efficiently calculated vector of moments to be of use in estimation through method of moment techniques. Again in this case, access to higher order moments should improve fit.

The remainder of this chapter is organized as follows. In Section 5.2, we

introduce Matryoshkan matrix sequences and identify some of their key properties. In Section 5.3 we use these matrices to find the moments of a large class of general Markov processes. We also give specific examples. In Section 5.4, we demonstrate the numerical performance of this method in comparison to traditional differential equation techniques. In Section 5.5, we conclude. Throughout the course of this study, we make the following contributions:

i) We define a novel class of matrix sequences that we call Matryoshkan matrix sequences for their nesting structure. We identify key properties of these matrices such as their inverse and matrix exponentials.

ii) Through these Matryoshkan matrices, we solve for closed form expressions for the moments of a large class of Markov processes. Furthermore, we demonstrate the general applicability of this technique through application to notable stochastic processes including Hawkes processes, shot noise process, Itô diffusions, growth-collapse processes, and linear birth-death-immigration processes. In the case of the Hawkes process and growth-collapse processes this resolves an open problem, as closed form expressions of these general transient moments were not previously known in the literature.

iii) We compare the precision and computation time of our methodology to numerically solving the underlying differential equations. In observing empirical superiority of the Matryoshkan matrix approach, we demonstrate the efficiency of calculating the moments at a given point, rather than on a path through time.

5.2 Matryoshkan Matrix Sequences

For the sake of clarity, let us begin this section by introducing general notation patterns we will use throughout this chapter. Because of the heavy use of matrices in this work, we reserve boldface upper case variables for these objects, such as $\mathbf{I}$ for the identity matrix. Similarly we let boldface lower case variables be vectors, such as $\mathbf{v}$ for the vector of all ones or $\mathbf{v}_i$ being the unit vector in the ith direction. One can assume that all vectors are column vectors unless otherwise noted. Scalar terms will not be bolded. A special matrix that we will use throughout this work is the diagonal matrix, which we denote $\mathbf{diag}(\mathbf{a})$, which is a square matrix with the values of the vector $\mathbf{a}$ along its diagonal and zeros otherwise. We will also make use of a generalization of this, denoted $\mathbf{diag}(\mathbf{a}, k)$, which instead contains the values of $\mathbf{a}$ on the kth off-diagonal, with negative k being below the diagonal and positive k being above.

Let us now introduce a sequence of matrices that will be at the heart of this work. We begin as follows: consider a sequence of lower triangular matrices $\{\mathbf{M}_n, n \in \mathbb{Z}^+\}$ such that

$$\mathbf{M}_n = \begin{bmatrix} \mathbf{M}_{n-1} & \mathbf{0}_{n-1\times 1} \\ \mathbf{m}_n & m_{n,n} \end{bmatrix}, \tag{5.1}$$

where $\mathbf{m}_n \in \mathbb{R}^{n-1}$ is a row vector, $m_{n,n} \in \mathbb{R}$, and $\mathbf{M}_1 = m_{1,1}$, an initial value. Taking inspiration from Matryoshka dolls, the traditional Russian nesting dolls, we will refer to these objects as *Matryoshkan* matrices. Using their nested and triangular structures, we can make four quick observations of note regarding Matryoshkan matrices.

Proposition 5.2.1. *Each of the following statements is a consequence of the definition of Matryoshkan matrices given by Equation 5.1:*

i) *If* $\mathbf{X}_n \in \mathbb{R}^{n\times n}$ *and* $\mathbf{Y}_n \in \mathbb{R}^{n\times n}$ *are both Matryoshkan matrix sequences, then so are* $\mathbf{X}_n + \mathbf{Y}_n$ *and* $\mathbf{X}_n\mathbf{Y}_n$.

ii) *If* $m_{i,i} \neq 0$ *for all* $i \in \{1, \ldots, n\}$ *then the Matryoshkan matrix* $\mathbf{M}_n \in \mathbb{R}^{n\times n}$ *is nonsingular. Moreover, the inverse of* $\mathbf{M}_n$ *is given by the recursion*

$$\mathbf{M}_n^{-1} = \begin{bmatrix} \mathbf{M}_{n-1}^{-1} & \mathbf{0}_{n-1\times 1} \\ -\frac{1}{m_{n,n}}\mathbf{m}_n\mathbf{M}_{n-1}^{-1} & \frac{1}{m_{n,n}} \end{bmatrix}. \tag{5.2}$$

iii) *If* $m_{i,i} \neq 0$ *for all* $i \in \{1, \ldots, n\}$ *and are all distinct then the matrix exponential of the Matryoshkan matrix* $\mathbf{M}_n \in \mathbb{R}^{n\times n}$ *multiplied by* $t \in \mathbb{R}$ *follows the recursion*

$$e^{\mathbf{M}_n t} = \begin{bmatrix} e^{\mathbf{M}_{n-1}t} & \mathbf{0}_{n-1\times 1} \\ \mathbf{m}_n\left(\mathbf{M}_{n-1} - m_{n,n}\mathbf{I}\right)^{-1}\left(e^{\mathbf{M}_{n-1}t} - e^{m_{n,n}t}\mathbf{I}\right) & e^{m_{n,n}t} \end{bmatrix}. \tag{5.3}$$

iv) *If* $m_{i,i} \neq 0$ *for all* $i \in \{1, \ldots, n\}$ *and are all distinct then the matrices* $\mathbf{U}_n \in \mathbb{R}^{n\times n}$ *and* $\mathbf{D}_n \in \mathbb{R}^{n\times n}$ *are such that*

$$\mathbf{M}_n\mathbf{U}_n = \mathbf{U}_n\mathbf{D}_n$$

for the Matryoshkan matrix $\mathbf{M}_n \in \mathbb{R}^{n\times n}$ *when defined recursively as*

$$\mathbf{U}_n = \begin{bmatrix} \mathbf{U}_{n-1} & \mathbf{0}_{n-1\times 1} \\ \mathbf{m}_n\left(\mathbf{D}_{n-1} - m_{n,n}\mathbf{I}\right)^{-1}\mathbf{U}_{n-1} & 1 \end{bmatrix}, \ \mathbf{D}_n = \begin{bmatrix} \mathbf{D}_{n-1} & \mathbf{0}_{n-1\times 1} \\ \mathbf{0}_{1\times n-1} & m_{n,n} \end{bmatrix}. \tag{5.4}$$

One can pause to note that in some sense any lower triangular matrix could be considered Matryoshkan, or at least be able to satisfy these properties. However, we note that some of the most significant insights we can gain from the Matryoshkan structure are the recursive implications available for sequences of matrices. Moreover, it is the combination of the nested relationship of consecutive matrices and the lower triangular structure that enables us to find these patterns. We will now see how this notion of Matryoshkan matrix sequences and the associated properties above can be used to specify element-wise solutions to a sequence of differential equations.

Lemma 5.2.2. *Let* $\mathbf{M}_n \in \mathbb{R}^{n\times n}$, $\mathbf{c}_n \in \mathbb{R}^n$, *and* $\mathbf{s}_n(t) : \mathbb{R}^+ \to \mathbb{R}^n$ *be such that*

$$\mathbf{M}_n = \begin{bmatrix} \mathbf{M}_{n-1} & \mathbf{0}_{n-1\times 1} \\ \mathbf{m}_n & m_{n,n} \end{bmatrix}, \quad \mathbf{c}_n = \begin{bmatrix} \mathbf{c}_{n-1} \\ c_n \end{bmatrix}, \quad \text{and} \quad \mathbf{s}_n(t) = \begin{bmatrix} \mathbf{s}_{n-1}(t) \\ s_n(t) \end{bmatrix}$$

where $\mathbf{m}_n \in \mathbb{R}^{n-1}$ *is a row vector,* $\mathbf{c}_{n-1} \in \mathbb{R}^{n-1}$, $s_n(t) \in \mathbb{R}$, *and* $\mathbf{M}_1 = m_{1,1}$. *Further, suppose that*

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{s}_n(t) = \mathbf{M}_n\mathbf{s}_n(t) + \mathbf{c}_n.$$

Then, if $m_{k,k} \neq 0$ *for all* $k \in \{1, \ldots, n\}$, *the vector function* $\mathbf{s}_n(t)$ *is given by*

$$\mathbf{s}_n(t) = e^{\mathbf{M}_n t}\mathbf{s}_n(0) - \mathbf{M}_n^{-1}\left(\mathbf{I} - e^{\mathbf{M}_n t}\right)\mathbf{c}_n, \tag{5.5}$$

and if all $m_{k,k} \neq 0$ *for all* $k \in \{1, \ldots, n\}$ *are distinct, the* n^{th} *scalar function* $s_n(t)$ *is given by*

$$\begin{aligned} s_n(t) = {} & \mathbf{m}_n\left(\mathbf{M}_{n-1} - m_{n,n}\mathbf{I}\right)^{-1}\left(e^{\mathbf{M}_{n-1}t} - e^{m_{n,n}t}\mathbf{I}\right)\left(\mathbf{s}_{n-1}(0) + \frac{\mathbf{c}_{n-1}}{m_{n,n}}\right) + e^{m_{n,n}t}s_n(0) \\ & - \frac{c_n}{m_{n,n}}\left(1 - e^{m_{n,n}t}\right) + \frac{\mathbf{m}_n}{m_{n,n}}\mathbf{M}_{n-1}^{-1}\left(\mathbf{I} - e^{\mathbf{M}_{n-1}t}\right)\mathbf{c}_{n-1}, \end{aligned} \tag{5.6}$$

where $t \geq 0$.

With this lemma in hand, we can now move to using these matrix sequences for calculating Markov process moments. To do so, we will use the infinitesimal generator, a key tool for Markov processes, to find the derivatives of the moments through time. By identifying a Matryoshkan matrix structure in these differential equations, we are able to apply Lemma 5.2.2 to find closed form expressions for the moments.

5.3 Calculating Moments through Matryoshkan Matrix Sequences

In this section we connect Matryoshkan matrix sequences with the moments of Markov processes. In Subsection 5.3.1, we prove the main result, which is the computation of the moment of a general Markov process through Matryoshkan matrices. To demonstrate the applicability of this result, we now apply it to a series of examples. First in Subsection 5.3.2 we obtain the moments of the self-exciting Hawkes process, for which finding moments in closed form has been an open problem. Then in Subsection 5.3.3 we study the Markovian shot noise process, a stochastic intensity process that trades self-excitement for external shocks. Next in Subsection 5.3.4 we showcase the use of these techniques for diffusive dynamics through application to Itô diffusions. Finally, in Subsection 5.3.6 we apply this technique to a process with jumps both upwards and downwards, a linear birth-death-immigration process we have studied previously called the Affine Queue-Hawkes process. In each scenario, we describe the process of interest, define the infinitesimal generator, and identify the matrix structure. Through this, we solve for the process moments.

5.3.1 The Moments of General Markov Processes

The connection between Matryoshkan matrices and Markov processes is built upon a key tool for Markov processes, the infinitesimal generator. For a detailed introduction to infinitesimal generators and their use in studying Markov processes, see e.g. Ethier and Kurtz (2009). For a Markov process X_t on state space

$\mathbb{S}$, the infinitesimal generator on a function $f : \mathbb{S} \to \mathbb{R}$ is defined

$$\mathcal{L}f(x) = \lim_{\tau \to 0} \frac{\mathrm{E}\left[f(X_\tau) \mid X_0 = x\right] - f(x)}{\tau}.$$

In our context and in many others, the power of the infinitesimal generator comes through use of Dynkin's formula, which gives us that

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[f(X_t)\right] = \mathrm{E}\left[\mathcal{L}f(X_t)\right].$$

To study the moments of a Markov process, we are interested in functions f that are polynomials. Let's suppose now that $\mathcal{L}x^n$ for any $n \in \mathbb{Z}^+$ is polynomial in the lower powers of x for a given Markov process X_t. Then, we can then write

$$\mathcal{L}X_t^n = \theta_{0,n} + \sum_{i=1}^{n} \theta_{i,n} X_t^i,$$

which implies that the differential equation for the nth moment of this process is

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[X_t^n\right] = \theta_{0,n} + \sum_{i=1}^{n} \theta_{i,n} \mathrm{E}\left[X_t^i\right],$$

for some collection of constants $\theta_{0,n}, \theta_{1,n}, \ldots, \theta_{n,n}$. Thus, to solve for the nth moment of X_t we must first solve for all the moments of lower order. We can also observe that to solve for the $(n-1)^{\text{th}}$ moment we must have all the moments below it. In comparing these systems of differential equations, we can see that all of the equations in the system for the $(n-1)^{\text{th}}$ moment are also in the system for the nth moment. No coefficients are changed in any of these lower moment equations, the only difference between the two systems is the inclusion of the differential equation for the nth moment in its own system. Hence, the nesting Matryoshkan structure arises. By expressing each system of linear ordinary differential equations in terms of a vector of moments, a matrix of coefficients, and a vector of shift terms, we can use these matrix sequences to capture how one moment's system encapsulates all the systems below it. This observation then

allows us to calculate all the moments of the process in closed form, as we now show in Theorem 5.3.1.

Theorem 5.3.1. *Let X_t be a Markov process such that the time derivative of its nth moment can be written*

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[X_t^n\right] = \theta_{0,n} + \sum_{i=1}^{n} \theta_{i,n}\mathrm{E}\left[X_t^i\right], \tag{5.7}$$

for any $n \in \mathbb{Z}^+$, where $t \geq 0$ and $\theta_{i,n} \in \mathbb{R}$ for all $i \leq n$. Let $\boldsymbol{\Theta}_n \in \mathbb{R}^{n\times n}$ be defined recursively by

$$\boldsymbol{\Theta}_n = \begin{bmatrix} \boldsymbol{\Theta}_{n-1} & \mathbf{0}_{n-1\times 1} \\ \boldsymbol{\theta}_n & \theta_{n,n} \end{bmatrix}, \tag{5.8}$$

where $\boldsymbol{\theta}_n = [\theta_{1,n}, \ldots, \theta_{n-1,n}]$ and $\boldsymbol{\Theta}_1 = \theta_{1,1}$. Furthermore, let $\boldsymbol{\theta}_{0,n} = [\theta_{0,1}, \ldots, \theta_{0,n}]^{\mathrm{T}}$. Then, if $\theta_{k,k} \neq 0$ for all $k \in \{1, \ldots, n\}$ are distinct, the nth moment of X_t can be expressed

$$\begin{aligned}\mathrm{E}\left[X_t^n\right] = \boldsymbol{\theta}_n \left(\boldsymbol{\Theta}_{n-1} - \theta_{n,n}\mathbf{I}\right)^{-1} \left(e^{\boldsymbol{\Theta}_{n-1}t} - e^{\theta_{n,n}t}\mathbf{I}\right)\left(\mathbf{x}_{n-1}(x_0) + \frac{\boldsymbol{\theta}_{0,n-1}}{\theta_{n,n}}\right) + x_0^n e^{\theta_{n,n}t} - \frac{\theta_{0,n}(1 - e^{\theta_{n,n}t})}{\theta_{n,n}} \\ + \frac{\boldsymbol{\theta}_n}{\theta_{n,n}}\boldsymbol{\Theta}_{n-1}^{-1}\left(\mathbf{I} - e^{\boldsymbol{\Theta}_{n-1}t}\right)\boldsymbol{\theta}_{0,n-1},\end{aligned} \tag{5.9}$$

where x_0 is the initial value of X_t and where $\mathbf{x}_n(a) \in \mathbb{R}^n$ is such that $(\mathbf{x}_n(a))_i = a^i$. If X_t has a stationary distribution, then the nth steady-state moment $\mathrm{E}\left[X_\infty^n\right]$ is given by

$$\mathrm{E}\left[X_\infty^n\right] = \frac{1}{\theta_{n,n}}\left(\boldsymbol{\theta}_n\boldsymbol{\Theta}_{n-1}^{-1}\boldsymbol{\theta}_{0,n-1} - \theta_{0,n}\right), \tag{5.10}$$

and these steady-state moments satisfy the recursive relationship

$$\mathrm{E}\left[X_\infty^{n+1}\right] = -\frac{1}{\theta_{n+1,n+1}}\left(\boldsymbol{\theta}_{n+1}\mathbf{s}_n^X(\infty) + \theta_{0,n+1}\right), \tag{5.11}$$

where $\mathbf{s}_n^X(\infty) \in \mathbb{R}^n$ is the vector of steady-state moments such that $\left(\mathbf{s}_n^X(\infty)\right)_i = \mathrm{E}\left[X_\infty^i\right]$.

Proof. Using the definition of $\boldsymbol{\Theta}_n$ in Equation 5.8, Equation 5.7 gives rise to the system of ordinary differential equations given by

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{s}_n^X(t) = \boldsymbol{\Theta}_n\mathbf{s}_n^X(t) + \boldsymbol{\theta}_{0,n}, \tag{5.12}$$

where $\mathbf{s}_n^X(t) \in \mathbb{R}^n$ is the vector of transient moments at time $t \geq 0$ such that $\left(\mathbf{s}_n^X(t)\right)_i = \mathrm{E}\left[X_t^i\right]$ for all $1 \leq i \leq n$. We can observe that by definition the matrices $\mathbf{\Theta}_n$ form a Matryoshkan sequence, and thus by Lemma 5.2.2, we achieve the stated transient solution. To prove the steady-state solution, we can first note that if the process has a steady-state distribution then the vector $\mathbf{s}_n^X(\infty) \in \mathbb{R}^n$ defined $\left(\mathbf{s}_n^X(\infty)\right)_i = \mathrm{E}\left[X_\infty^i\right]$ will satisfy

$$0 = \mathbf{\Theta}_n \mathbf{s}_n^X(\infty) + \boldsymbol{\theta}_{0,n}, \tag{5.13}$$

as this is the equilibrium solution to the differential equation corresponding to each of the moments. This system has a unique solution since $\mathbf{\Theta}_n$ is nonsingular due to the assumption that the diagonal values are unique and non-zero. Using Proposition 5.2.1, we find the nth moment by

$$\mathrm{E}\left[X_\infty^n\right] = -\mathbf{v}_n^{\mathrm{T}} \mathbf{\Theta}_n^{-1} \boldsymbol{\theta}_{0,n} = \begin{bmatrix} \frac{1}{\theta_{n,n}} \boldsymbol{\theta}_n \mathbf{\Theta}_{n-1}^{-1} & -\frac{1}{\theta_{n,n}} \end{bmatrix} \begin{bmatrix} \boldsymbol{\theta}_{0,n-1} \\ \theta_{0,n} \end{bmatrix} = \frac{1}{\theta_{n,n}} \left(\boldsymbol{\theta}_n \mathbf{\Theta}_{n-1}^{-1} \boldsymbol{\theta}_{0,n-1} - \theta_{0,n} \right),$$

which completes the proof of Equation 5.27. To conclude, one can note that each line of the linear system in Equation 5.13 implies the stated recursion. □

5.3.2 Application to Hawkes Process Intensities

For our first example of this method let us turn to our motivating application, the Markovian Hawkes process intensity. Via Hawkes (1971), this process is defined as follows. Let λ_t be stochastic arrival process intensity such that

$$\lambda_t = \lambda^* + (\lambda_0 - \lambda^*) e^{-\beta t} + \int_0^t \alpha e^{-\beta(t-s)} \mathrm{d}N_s = \lambda^* + (\lambda_0 - \lambda^*) e^{-\beta t} + \sum_{i=1}^{N_t} \alpha e^{-\beta(t-A_i)},$$

where $\{A_i \mid i \in \mathbb{Z}^+\}$ is the sequence of arrival epochs in the point process N_t, with

$$\mathrm{P}\left(N_{t+s} - N_t = 0 \mid \mathcal{F}_t\right) = \mathrm{P}\left(N_{t+s} - N_t = 0 \mid \lambda_t\right) = e^{-\int_0^s \lambda_{t+u} \mathrm{d}u},$$

where $\mathcal{F}_t$ is the filtration generated by the history of λ_t up to time t. We will assume that $\beta > \alpha > 0$ so that the process has a stationary distribution, and we will also let $\lambda^* > 0$ and $\lambda_0 > 0$. Note that the process behaves as follows: at arrivals λ_t increases by α and in the interims it decays exponentially at rate β towards the baseline level λ^*. In this way, (λ_t, N_t) is referred to as a *self-exciting* point process, as the occurrence of an arrival increases the intensity and thus increases the likelihood that another arrival will occur soon afterwards. Because the intensity λ_t forms a Markov process, we can write its infinitesimal generator for a (sufficiently regular) function $f : \mathbb{R}^+ \to \mathbb{R}$ as follows:

$$\mathcal{L}f(\lambda_t) = \lambda_t \left(f(\lambda_t + \alpha) - f(\lambda_t)\right) - \beta \left(\lambda_t - \lambda^*\right) \frac{\mathrm{d}f(\lambda_t)}{\mathrm{d}\lambda_t}.$$

Note that this expression showcases the process dynamics that we have described, as the first term on the right-hand side corresponds to the product of the arrival rate and the change in the process when an arrival occurs while the second term captures the decay.

To obtain the nth moment we must consider $f(\cdot)$ of the form $f(x) = x^n$. In the simplest case when $n = 1$ this formula yields an ordinary differential equation for the mean, which can be written

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[\lambda_t\right] = \alpha \mathrm{E}\left[\lambda_t\right] - \beta \left(\mathrm{E}\left[\lambda_t\right] - \lambda^*\right) = \beta\lambda^* - (\beta - \alpha)\mathrm{E}\left[\lambda_t\right].$$

By comparison for the second moment at $n = 2$ we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[\lambda_t^2\right] = \mathrm{E}\left[\lambda_t\left((\lambda_t + \alpha)^2 - \lambda_t^2\right) - 2\beta\lambda_t(\lambda_t - \lambda^*)\right] = (2\beta\lambda^* + \alpha^2)\mathrm{E}\left[\lambda_t\right] - 2(\beta - \alpha)\mathrm{E}\left[\lambda_t^2\right],$$

and we can note that while the ODE for the mean is autonomous, the second moment equation depends on both the mean and the second moment. Thus, to solve for the second moment we must also solve for the mean, leading us to the

following system of differential equations:

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix}\mathrm{E}\left[\lambda_t\right]\\ \mathrm{E}\left[\lambda_t^2\right]\end{bmatrix}=\begin{bmatrix}-(\beta-\alpha) & 0\\ 2\beta\lambda^*+\alpha^2 & -2(\beta-\alpha)\end{bmatrix}\begin{bmatrix}\mathrm{E}\left[\lambda_t\right]\\ \mathrm{E}\left[\lambda_t^2\right]\end{bmatrix}+\begin{bmatrix}\beta\lambda^*\\ 0\end{bmatrix}.$$

Moving on to the third moment, the infinitesimal generator formula yields

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[\lambda_t^3\right]=\mathrm{E}\left[\lambda_t\left((\lambda_t+\alpha)^3-\lambda_t^3\right)-3\beta\lambda_t^2\left(\lambda_t-\lambda^*\right)\right]=\alpha^3\mathrm{E}\left[\lambda_t\right]+3(\beta\lambda^*+\alpha^2)\mathrm{E}\left[\lambda_t^2\right]-3(\beta-\alpha)\mathrm{E}\left[\lambda_t^3\right],$$

and hence we see that this ODE now depends on all of the first three moments. Thus, to solve for $\mathrm{E}\left[\lambda_t^3\right]$ we need to solve the system of ordinary differential equations

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix}\mathrm{E}\left[\lambda_t\right]\\ \mathrm{E}\left[\lambda_t^2\right]\\ \mathrm{E}\left[\lambda_t^3\right]\end{bmatrix}=\begin{bmatrix}-(\beta-\alpha) & 0 & 0\\ 2\beta\lambda^*+\alpha^2 & -2(\beta-\alpha) & 0\\ \alpha^3 & 3(\beta\lambda^*+\alpha^2) & -3(\beta-\alpha)\end{bmatrix}\begin{bmatrix}\mathrm{E}\left[\lambda_t\right]\\ \mathrm{E}\left[\lambda_t^2\right]\\ \mathrm{E}\left[\lambda_t^3\right]\end{bmatrix}+\begin{bmatrix}\beta\lambda^*\\ 0\\ 0\end{bmatrix},$$

and this now suggests the Matryoshkan structure of these process moments: we can note that the system for the second moment is nested within the system for the third moment. That is, the matrix for the three dimensional system contains the two dimensional system in its upper left-hand block, just as the vector of the first three moments has the first two moments in its first two coordinates. In general, we can see that the nth moment will satisfy the ODE given by

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[\lambda_t^n\right]&=\mathrm{E}\left[\lambda_t\left((\lambda_t+\alpha)^n-\lambda_t^n\right)-n\beta\lambda_t^{n-1}\left(\lambda_t-\lambda^*\right)\right]\\ &=\sum_{k=1}^{n}\binom{n}{k-1}\alpha^{n-k+1}\mathrm{E}\left[\lambda_t^k\right]-n\beta\mathrm{E}\left[\lambda_t^n\right]+n\beta\lambda^*\mathrm{E}\left[\lambda_t^{n-1}\right],\end{aligned}$$

where we have simplified by use of the binomial theorem. Thus, the system of differential equations needed to solve for the nth moment uses the matrix from the $(n-1)^{\text{th}}$ system augmented below by the row

$$\begin{bmatrix}\alpha^n & n\alpha^{n-1} & \binom{n}{2}\alpha^{n-2} & \dots & \binom{n}{n-3}\alpha^3 & \binom{n}{n-2}\alpha^2+n\beta\lambda^* & -n(\beta-\alpha)\end{bmatrix}, \tag{5.14}$$

and buffered on the right by a column of zeros. To collect these coefficients into a coherent structure, let us define the matrix $\mathcal{P}_n(a) \in \mathbb{R}^{n\times n}$ for $a \in \mathbb{R}$ such that

$$(\mathcal{P}_n(a))_{i,j} = \begin{cases} \binom{i}{j-1} a^{i-j+1} & i \geq j, \\ 0 & i < j. \end{cases} \tag{5.15}$$

If we momentarily disregard the terms with β in the general augment row in Equation 5.14, one can observe that the remaining terms in this vector are given by the bottom row of the matrix $\mathcal{P}_n(\alpha)$. Furthermore, by definition $\{\mathcal{P}_n(a) \mid n \in \mathbb{Z}^+\}$ forms a Matryoshkan matrix sequence. We can also note that $\mathcal{P}_n(a)$ can be equivalently defined as

$$\mathcal{P}_n(a) = \sum_{k=1}^{n} a \begin{bmatrix} \mathbf{0}_{n-k\times n-k} & \mathbf{0}_{n-k\times k} \\ \mathbf{0}_{k\times n-k} & \mathbf{L}_k(a) \end{bmatrix},$$

where $\mathbf{L}_k(a) = e^{a\mathbf{diag}(1:k-1,-1)}$ is the kth lower triangular Pascal matrix, i.e. the nonzero terms in $\mathbf{L}_k(1)$ yield the first k rows of Pascal's triangle. Alternatively, $\mathcal{P}_n(a)$ can be found by creating a lower triangular matrix from the strictly lower triangular values in $\mathbf{L}_{n+1}(a)$. For these reasons, we refer to the sequence of $\mathcal{P}_n(a)$ as *Matryoshkan Pascal matrices*. For brief overviews and beautiful properties of Pascal matrices, see Brawer and Pirovino (1992); Call and Velleman (1993); Zhang (1997); Edelman and Strang (2004). As we have seen in the preceding derivation, Matryoshkan Pascal matrices arise naturally in using the infinitesimal generator for calculating moments of Markov processes. This follows from the application of the binomial theorem to jump terms. Now, in the case of the Markovian Hawkes process intensity we find closed form expressions for all transient moments in Corollary 5.3.2.

Corollary 5.3.2. *Let λ_t be the intensity of a Hawkes process with baseline intensity $\lambda^* > 0$, intensity jump $\alpha > 0$, and decay rate $\beta > \alpha$. Then, the nth moment of λ_t is given*

by

$$\mathrm{E}\left[\lambda_t^n\right] = \mathbf{m}_n^\lambda \left(\mathbf{M}_{n-1}^\lambda + n(\beta-\alpha)\mathbf{I}\right)^{-1} \left(e^{\mathbf{M}_{n-1}^\lambda t} - e^{-n(\beta-\alpha)t}\mathbf{I}\right)\left(\mathbf{x}_{n-1}(\lambda_0) - \frac{\beta\lambda^*\mathbf{v}_1}{n(\beta-\alpha)}\right) + \lambda_0^n e^{-n(\beta-\alpha)t}$$
$$+ \mathbb{I}_{\{n=1\}}\frac{\beta\lambda^*}{\beta-\alpha}\left(1-e^{-(\beta-\alpha)t}\right) - \frac{\beta\lambda^*}{n(\beta-\alpha)}\mathbf{m}_n^\lambda\left(\mathbf{M}_{n-1}^\lambda\right)^{-1}\left(\mathbf{I}-e^{\mathbf{M}_{n-1}^\lambda t}\right)\mathbf{v}_1, \tag{5.16}$$

for all $t \geq 0$ *and* $n \in \mathbb{Z}^+$*, where* $\mathbf{v}_1 \in \mathbb{R}^n$ *is the unit vector in the first coordinate,* $\mathbf{M}_n^\lambda = \beta\lambda^*\mathbf{diag}(2:n,-1) - \beta\mathbf{diag}(1:n) + \mathcal{P}_n(\alpha)$*,* $\mathbf{m}_n^\lambda = \left[\left(\mathbf{M}_n^\lambda\right)_{n,1}, \ldots, \left(\mathbf{M}_n^\lambda\right)_{n,n-1}\right]$ *is given by*

$$\left(\mathbf{m}_n^\lambda\right)_j = \begin{cases} \binom{n}{j-1}\alpha^{n-j+1} & \text{if } j < n-1, \\ \binom{n}{n-2}\alpha^2 + n\beta\lambda^* & \text{if } j = n-1, \end{cases}$$

and $\mathbf{x}_n(a) \in \mathbb{R}^n$ *is such that* $(\mathbf{x}_n(a))_i = a^i$*. In steady-state, the* n^{th} *moment of* λ_t *is given by*

$$\lim_{t\to\infty}\mathrm{E}\left[\lambda_t^n\right] = -\frac{\beta\lambda}{n(\beta-\alpha)}\mathbf{m}_n^\lambda\left(\mathbf{M}_{n-1}^\lambda\right)^{-1}\mathbf{v}_1, \tag{5.17}$$

for $n \geq 2$ *with* $\lim_{t\to\infty}\mathrm{E}\left[\lambda_t\right] = \frac{\beta\lambda^*}{\beta-\alpha}$*. Moreover, the* $(n+1)^{th}$ *steady-state moment of the Hawkes process intensity is given by the recursion*

$$\lim_{t\to\infty}\mathrm{E}\left[\lambda_t^{n+1}\right] = \frac{1}{(n+1)(\beta-\alpha)}\mathbf{m}_{n+1}^\lambda\mathbf{s}_n^\lambda, \tag{5.18}$$

for all $n \in \mathbb{Z}^+$*, where* $\mathbf{s}_n^\lambda \in \mathbb{R}^n$ *is the vector of steady-state moments defined such that* $\left(\mathbf{s}_n^\lambda\right)_i = \lim_{t\to\infty}\mathrm{E}\left[\lambda_t^i\right]$ *for* $1 \leq i \leq n$*.*

5.3.3 Application to Shot Noise Processes

As a second example of calculating moments through Matryoshkan matrices, consider a Markovian shot noise process; see e.g. Daley and Vere-Jones (2003) for an introduction. That is, let ψ_t be defined such that

$$\psi_t = \sum_{i=1}^{N_t} J_i e^{-\beta(t-A_i)},$$

where $\beta > 0$, $\{J_i \mid i \in \mathbb{Z}^+\}$ is a sequence of i.i.d. positive random variables, N_t is a Poisson process at rate $\lambda > 0$, and $\{A_i \mid i \in \mathbb{Z}^+\}$ is the sequence of arrival times in the Poisson process. These dynamics yield the following infinitesimal generator:

$$\mathcal{L}f(\psi_t) = \lambda\left(f(\psi_t + J_i) - f(\psi_t)\right) - \beta\psi_t \frac{\mathrm{d}f(\psi_t)}{\mathrm{d}\psi_t}.$$

We can note that this is similar to the Hawkes process discussed in Subsection 5.3.2, as the right-hand side contains a term for jumps and a term for exponential decay. However, this infinitesimal generator formula also shows key differences between the two processes, as the jumps in the shot noise process are of random size and they occur at the fixed, exogenous rate $\lambda > 0$. Supposing the mean jump size is finite, this now yields that the mean satisfies the ordinary differential equation

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[\psi_t\right] = \lambda\mathrm{E}\left[J_1\right] - \beta\mathrm{E}\left[\psi_t\right],$$

whereas if $\mathrm{E}\left[J_1^2\right] < \infty$, the second moment of the shot noise process is given by the solution to

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[\psi_t^2\right] = \mathrm{E}\left[\lambda\left((\psi_t + J_1)^2 - \psi_t^2\right) - 2\beta\psi_t^2\right] = \lambda\mathrm{E}\left[J_1^2\right] + 2\lambda\mathrm{E}\left[J_1\right]\mathrm{E}\left[\psi_t\right] - 2\beta\mathrm{E}\left[\psi_t^2\right],$$

which depends on both the second moment and the mean. This gives rise to the linear system of differential equations

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix}\mathrm{E}\left[\psi_t\right] \\ \mathrm{E}\left[\psi_t^2\right]\end{bmatrix} = \begin{bmatrix}-\beta & 0 \\ 2\lambda\mathrm{E}\left[J_1\right] & -2\beta\end{bmatrix}\begin{bmatrix}\mathrm{E}\left[\psi_t\right] \\ \mathrm{E}\left[\psi_t^2\right]\end{bmatrix} + \begin{bmatrix}\lambda\mathrm{E}\left[J_1\right] \\ \lambda\mathrm{E}\left[J_1^2\right]\end{bmatrix},$$

and by observing that the differential equation for the third moment depends on the first three moments if the third moment of the jump size is finite,

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[\psi_t^3\right] = \mathrm{E}\left[\lambda\left((\psi_t + J_1)^3 - \psi_t^3\right) - 3\beta\psi_t^3\right] = \lambda\mathrm{E}\left[J_1^3\right] + 3\lambda\mathrm{E}\left[J_1^2\right]\mathrm{E}\left[\psi_t\right] + 3\lambda\mathrm{E}\left[J_1\right]\mathrm{E}\left[\psi_t^2\right] - 3\beta\mathrm{E}\left[\psi_t^3\right],$$

we can see that the system for the first two moments again are contained in the system for the first three moments:

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix}\mathrm{E}\left[\psi_t\right]\\ \mathrm{E}\left[\psi_t^2\right]\\ \mathrm{E}\left[\psi_t^3\right]\end{bmatrix}=\begin{bmatrix}-\beta & 0 & 0\\ 2\lambda\mathrm{E}\left[J_1\right] & -2\beta & 0\\ 3\lambda\mathrm{E}\left[J_1^2\right] & 3\lambda\mathrm{E}\left[J_1\right] & -3\beta\end{bmatrix}\begin{bmatrix}\mathrm{E}\left[\psi_t\right]\\ \mathrm{E}\left[\psi_t^2\right]\\ \mathrm{E}\left[\psi_t^3\right]\end{bmatrix}+\begin{bmatrix}\lambda\mathrm{E}\left[J_1\right]\\ \lambda\mathrm{E}\left[J_1^2\right]\\ \lambda\mathrm{E}\left[J_1^3\right]\end{bmatrix}.$$

By use of the binomial theorem, we can observe that if $\mathrm{E}\left[J_1^n\right] < \infty$ then the nth moment of the shot noise process satisfies

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[\psi_t^n\right]=\mathrm{E}\left[\lambda\left((\psi_t+J_1)^n-\psi_t^n\right)-n\beta\psi_t^n\right]=\sum_{k=0}^{n-1}\binom{n}{k}\mathrm{E}\left[J_1^{n-k}\right]\mathrm{E}\left[\psi_t^k\right]-n\beta\mathrm{E}\left[\psi_t^n\right],$$

which means that the nth dimensional system is equal to the preceding one augmented below by the row vector

$$\begin{bmatrix}n\lambda\mathrm{E}\left[J_1^{n-1}\right] & \binom{n}{2}\lambda\mathrm{E}\left[J_1^{n-2}\right] & \binom{n}{3}\lambda\mathrm{E}\left[J_1^{n-3}\right] & \dots & \binom{n}{n-2}\lambda\mathrm{E}\left[J_1^2\right] & n\lambda\mathrm{E}\left[J_1\right] & -n\beta\end{bmatrix},$$

and to the right by zeros. Bringing this together, this now leads us to Corollary 5.3.3.

Corollary 5.3.3. *Let ψ_t be the intensity of a shot noise process with epochs given by a Poisson process with rate $\lambda > 0$, jump sizes drawn from the i.i.d. sequence of random variables $\{J_i \mid i \in \mathbb{Z}^+\}$, and exponential decay at rate $\beta > 0$. If $\mathrm{E}\left[J_1^n\right] < \infty$, the nth moment of ψ_t is given by*

$$\begin{aligned}\mathrm{E}\left[\psi_t^n\right]=&\,\mathbf{m}_n^\psi\left(\mathbf{M}_{n-1}^\psi+n\beta\mathbf{I}\right)^{-1}\left(e^{\mathbf{M}_{n-1}^\psi t}-e^{-n\beta t}\mathbf{I}\right)\left(\mathbf{x}_{n-1}\left(\psi_0\right)-\frac{\lambda\mathbf{j}_{n-1}}{n\beta}\right)+\psi_0^n e^{-n\beta t}\\ &+\frac{\lambda\mathrm{E}\left[J_1^n\right]}{n\beta}\left(1-e^{-n\beta t}\right)-\frac{\lambda\mathbf{m}_n^\psi}{n\beta}\left(\mathbf{M}_{n-1}^\psi\right)^{-1}\left(\mathbf{I}-e^{\mathbf{M}_{n-1}^\psi t}\right)\mathbf{j}_{n-1},\end{aligned}\tag{5.19}$$

for all $t \geq 0$ and $n \in \mathbb{Z}^+$, where $\mathbf{j}_n \in \mathbb{R}^n$ is such that $(\mathbf{j}_n)_i = \mathrm{E}\left[J_1^i\right]$, $\mathbf{M}_n^\psi \in \mathbb{R}^{n\times n}$ is recursively defined

$$\mathbf{M}_n^\psi=\begin{bmatrix}\mathbf{M}_{n-1}^\psi & \mathbf{0}_{n-1\times 1}\\ \mathbf{m}_n^\psi & -n\beta\end{bmatrix},$$

with the row vector $\mathbf{m}_n^\psi \in \mathbb{R}^{n-1}$ *defined such that* $\left(\mathbf{m}_n^\psi\right)_i = \binom{n}{i}\lambda \mathrm{E}\left[J_1^{n-i}\right]$ *and with* $\mathbf{M}_1^\psi = -\beta$*, and where* $\mathbf{x}_n(a) \in \mathbb{R}^n$ *is such that* $(\mathbf{x}_n(a))_i = a^i$*. In steady-state, the* $(n+1)^{th}$ *moment of the shot noise process is given by*

$$\lim_{t\to\infty} \mathrm{E}\left[\psi_t^n\right] = \frac{\lambda}{n\beta}\left(\mathrm{E}\left[J_1^n\right] - \mathbf{m}_n^\psi \left(\mathbf{M}_{n-1}^\psi\right)^{-1} \mathbf{j}_{n-1}\right), \tag{5.20}$$

for $n \geq 2$ *where* $\lim_{t\to\infty} \mathrm{E}\left[\psi_t\right] = \frac{\lambda}{\beta}\mathrm{E}\left[J_1\right]$*. Moreover, if* $\mathrm{E}\left[J_1^{n+1}\right] < \infty$ *the* $(n+1)^{th}$ *moment of the shot noise process is given by the recursion*

$$\lim_{t\to\infty} \mathrm{E}\left[\psi_t^{n+1}\right] = \frac{1}{(n+1)\beta}\left(\mathbf{m}_{n+1}^\psi \mathbf{s}_n^\psi + \mathrm{E}\left[J_1^{n+1}\right]\right), \tag{5.21}$$

for all $n \in \mathbb{Z}^+$*, where* $\mathbf{s}_n^\psi \in \mathbb{R}^n$ *is the vector of steady-state moments defined such that* $(\mathbf{s}_n^\psi)_i = \lim_{t\to\infty} \mathrm{E}\left[\psi_t^i\right]$ *for* $1 \leq i \leq n$*.*

5.3.4 Application to Itô Diffusions

For our third example, we consider an Itô diffusion; see e.g. Oksendal (2013) for an overview. Let S_t be given by the stochastic differential equation

$$\mathrm{d}S_t = g(S_t)\mathrm{d}t + h(S_t)\mathrm{d}B_t,$$

where B_t is a Brownian motion and $g(\cdot)$ and $h(\cdot)$ are real-valued functions. Then, the infinitesimal generator for this process is given by

$$\mathcal{L}f(S_t) = g(S_t)\frac{\mathrm{d}f(S_t)}{\mathrm{d}S_t} + \frac{h(S_t)^2}{2}\frac{\mathrm{d}^2 f(S_t)}{\mathrm{d}S_t^2},$$

where $f : \mathbb{R} \to \mathbb{R}$. Because we will be considering functions of the form $f(x) = x^n$ for $n \in \mathbb{Z}^+$, we will now specify the forms of $g(\cdot)$ and $h(\cdot)$ to be $g(x) = \mu + \theta x$ for some $\mu \in \mathbb{R}$ and $\theta \in \mathbb{R}$ and $h(x) = \sigma x^{\gamma/2}$ for some $\sigma \in \mathbb{R}$ and $\gamma \in \{0, 1, 2\}$. One can note that this encapsulates a myriad of relevant stochastic processes including many that are popular in the financial models literature, such as

Ornstein-Uhlenbeck (OU) processes, geometric Brownian motion (GBM), and Cox-Ingersoll-Ross (CIR) processes. In this case, the infinitesimal generator becomes

$$\mathcal{L}f(S_t) = (\mu + \theta S_t)\frac{\mathrm{d}f(S_t)}{\mathrm{d}S_t} + \frac{\sigma^2 S_t^\gamma}{2}\frac{\mathrm{d}^2 f(S_t)}{\mathrm{d}S_t^2},$$

meaning that we can express the ordinary differential equation for the mean as

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}[S_t] = \mu + \theta\mathrm{E}[S_t],$$

and similarly the second moment will be given by the solution to

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[S_t^2\right] = \mathrm{E}\left[2(\mu + \theta S_t)S_t + \sigma^2 S_t^\gamma\right] = 2\mu\mathrm{E}[S_t] + 2\theta\mathrm{E}\left[S_t^2\right] + \sigma^2\mathrm{E}[S_t^\gamma].$$

For the sake of example, we now let $\gamma = 1$ as is the case in the CIR process. Then, the first two transient moments of S_t will be given by the solution to the system

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix}\mathrm{E}[S_t] \\ \mathrm{E}\left[S_t^2\right]\end{bmatrix} = \begin{bmatrix}\theta & 0 \\ 2\mu + \sigma^2 & 2\theta\end{bmatrix}\begin{bmatrix}\mathrm{E}[S_t] \\ \mathrm{E}\left[S_t^2\right]\end{bmatrix} + \begin{bmatrix}\mu \\ 0\end{bmatrix}.$$

By observing that the third moment differential equation is

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[S_t^3\right] = \mathrm{E}\left[3(\mu + \theta S_t)S_t^2 + 3\sigma^2 S_t^{\gamma+1}\right] = 3\mu\mathrm{E}\left[S_t^2\right] + 3\theta\mathrm{E}\left[S_t^3\right] + 3\sigma^2\mathrm{E}\left[S_t^{\gamma+1}\right],$$

we can note that the third moment system for $\gamma = 1$ is

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix}\mathrm{E}[S_t] \\ \mathrm{E}\left[S_t^2\right] \\ \mathrm{E}\left[S_t^3\right]\end{bmatrix} = \begin{bmatrix}\theta & 0 & 0 \\ 2\mu + \sigma^2 & 2\theta & 0 \\ 0 & 3\mu + 3\sigma^2 & 3\theta\end{bmatrix}\begin{bmatrix}\mathrm{E}[S_t] \\ \mathrm{E}\left[S_t^2\right] \\ \mathrm{E}\left[S_t^3\right]\end{bmatrix} + \begin{bmatrix}\mu \\ 0 \\ 0\end{bmatrix},$$

and this showcases the Matryoshkan nesting structure, as the second moment system is contained within the third. Because the general n^{th} moment for $n \geq 2$ has differential equation given by

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}[S_t^n] = \mathrm{E}\left[n(\mu + \theta S_t)S_t^{n-1} + \frac{n(n-1)\sigma^2}{2}S_t^{n+\gamma-2}\right] = n\mu\mathrm{E}\left[S_t^{n-1}\right] + n\theta\mathrm{E}[S_t^n] + \frac{n(n-1)\sigma^2}{2}\mathrm{E}\left[S_t^{n+\gamma-2}\right],$$

we can see that the $(n-1)^{\text{th}}$ system can be augmented below by the row vector $\gamma = 1$

$$\begin{bmatrix} 0 & 0 & \dots & 0 & n\mu + \frac{n(n-1)\sigma^2}{2} & n\theta \end{bmatrix},$$

and to the right by zeros. Through this observation, we can now give the moments of Itô diffusions in Corollary 5.3.4.

Corollary 5.3.4. *Let S_t be an Itô diffusion that satisfies the stochastic differential equation*

$$\mathrm{d}S_t = (\mu + \theta S_t)\mathrm{d}t + \sigma S_t^{\gamma/2}\mathrm{d}B_t, \tag{5.22}$$

where B_t is a Brownian motion and with $\mu, \theta, \sigma \in \mathbb{R}$ and $\gamma \in \{0, 1, 2\}$. Then, the nth moment of S_t is given by

$$\begin{aligned} \mathrm{E}\left[S_t^n\right] = {} & \mathbf{m}_n^S \left(\mathbf{M}_{n-1}^S - \chi_n \mathbf{I}\right)^{-1} \left(e^{\mathbf{M}_{n-1}^S t} - e^{\chi_n t}\mathbf{I}\right)\left(\mathbf{x}_{n-1}(S_0) + \frac{\mu \mathbf{v}_1 + \sigma^2 \mathbb{I}_{\{\gamma=0\}}\mathbf{v}_2}{\chi_n}\right) + S_0^n e^{\chi_n t} \\ & - \left(\mu \mathbb{I}_{\{n=1\}} + \sigma^2 \mathbb{I}_{\{\gamma=0, n=2\}}\right)\frac{1 - e^{\chi_n t}}{\chi_n} + \frac{\mathbf{m}_n^S}{\chi_n}\left(\mathbf{M}_{n-1}^S\right)^{-1}\left(\mathbf{I} - e^{\mathbf{M}_{n-1}^S t}\right)\left(\mu \mathbf{v}_1 + \sigma^2 \mathbb{I}_{\{\gamma=0\}}\mathbf{v}_2\right), \end{aligned} \tag{5.23}$$

for all $t \geq 0$ and $n \in \mathbb{Z}^+$, where $\chi_n = n\theta + \frac{n}{2}(n-1)\sigma^2 \mathbb{I}_{\{\gamma=2\}}$, $\mathbf{M}_n^S = \theta \mathbf{diag}(1:n) + \mu \mathbf{diag}(2:n, -1) + \frac{\sigma^2}{2}\mathbf{diag}(\mathbf{d}_{n+\gamma-2}^{2-\gamma}, \gamma - 2)$ for $\mathbf{d}_k^j \in \mathbb{R}^k$ such that $(\mathbf{d}_k^j)_i = (j+i)(j+i-1)$, and $\mathbf{m}_n^S = \left[\left(\mathbf{M}_n^S\right)_{n,1}, \dots, \left(\mathbf{M}_n^S\right)_{n,n-1}\right]$ is such that

$$\left(\mathbf{m}_n^S\right)_j = \begin{cases} n\mu + \frac{n(n-1)\sigma^2}{2}\mathbb{I}_{\{\gamma=1\}} & j = n-1, \\ \frac{n(n-1)\sigma^2}{2}\mathbb{I}_{\{\gamma=0\}} & j = n-2, \\ 0 & 1 \leq j < n-2, \end{cases}$$

and $\mathbf{x}_n(a) \in \mathbb{R}^n$ is such that $(\mathbf{x}_n(a))_i = a^i$. If $\theta < 0$ and $\gamma \in \{0, 1\}$, then the nth steady-state moment of S_t is given by

$$\lim_{t\to\infty} \mathrm{E}\left[S_t^n\right] = \frac{\mu}{\chi_n}\mathbf{m}_n^S \left(\mathbf{M}_{n-1}^S\right)^{-1}\mathbf{v}_1, \tag{5.24}$$

for $n \geq 2$ *with* $\lim_{t\to\infty} \mathrm{E}[S_t] = -\frac{\mu}{\theta}$. *Moreover, the* $(n+1)^{th}$ *steady-state moment of* S_t *is given by the recursion*

$$\lim_{t\to\infty} \mathrm{E}\left[S_t^{n+1}\right] = -\frac{1}{\chi_n}\mathbf{m}_{n+1}^S \mathbf{s}_n^S, \tag{5.25}$$

for all $n \in \mathbb{Z}^+$, *where* $\mathbf{s}_n^S \in \mathbb{R}^n$ *is the vector of steady-state moments defined such that* $\left(\mathbf{s}_n^S\right)_i = \lim_{t\to\infty} \mathrm{E}\left[S_t^i\right]$ *for* $1 \leq i \leq n$.

As a consequence of these expressions we can also gain insight for the moments of an Itô diffusion in the case of non-integer $\gamma \in [0, 2]$, as is used in volatility models such as the CEV model and the SABR model, see e.g. Henry-Labordere (2008). This can be achieve through bounding the differential equations, as the nth moment of such a diffusion is again given by

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}[S_t^n] = n\mu\mathrm{E}\left[S_t^{n-1}\right] + n\theta\mathrm{E}[S_t^n] + \frac{n(n-1)\sigma^2}{2}\mathrm{E}\left[S_t^{n+\gamma-2}\right],$$

and the right-most term in this expression can be bounded above and below

$$\mathrm{E}\left[S_t^{n+\lfloor\gamma\rfloor-2}\right] \leq \mathrm{E}\left[S_t^{n+\gamma-2}\right] \leq \mathrm{E}\left[S_t^{n+\lceil\gamma\rceil-2}\right],$$

and the differential equations given by substituting these bounded terms form a closed system solvable by Corollary 5.3.4. Assuming the true differential equation and the upper and lower bounds all share an initial value, the solution to the bounded equations bounds the solution to the true moment equation, see Hale and Lunel (2013).

5.3.5 Application to Growth-Collapse Processes

For a fourth example, we consider growth-collapse processes with Poisson driven shocks. These processes have been studied in variety of contexts, see e.g.

Boxma et al. (2006); Kella (2009); Kella and Löpker (2010); Boxma et al. (2011). More recently, these processes and their related extensions have seen renewed interest in the study of the crypto-currency Bitcoin, see for example Frolkova and Mandjes (2019); Koops (2018); Javier and Fralix (2019); Fralix (2019). While growth-collapse processes can be defined in many different ways, for this example we use a definition in the style of Section 4 from Boxma et al. (2006). We let Y_t be the state of the growth collapse model and let $\{U_i \mid i \in \mathbb{Z}^+\}$ be a sequence of independent Uni(0, 1) random variables that are also independent from the state and history of the growth-collapse process. Then, the infinitesimal generator of Y_t is given by

$$\mathcal{L}f(Y_t) = \lambda \frac{\mathrm{d}f(Y_t)}{\mathrm{d}Y_t} + \mu \left(f(U_i Y_t) - f(Y_t) \right).$$

Thus, Y_t experiences linear growth at rate $\lambda > 0$ throughout time but it also collapses at epochs given by a Poisson process with rate $\mu > 0$. At the ith collapse epoch the process falls to a fraction of its current level, specifically it jumps down to $U_i Y_t$. Using the infinitesimal generator, we can see that the mean of this growth-collapse process satisfies

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[Y_t\right] = \lambda + \mu \left(\mathrm{E}\left[U_1 Y_t\right] - \mathrm{E}\left[Y_t\right] \right) = \lambda - \frac{\mu}{2}\mathrm{E}\left[Y_t\right],$$

and its second moment will satisfy

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[Y_t^2\right] = 2\lambda \mathrm{E}\left[Y_t\right] + \mu \left(\mathrm{E}\left[U_1^2 Y_t^2\right] - \mathrm{E}\left[Y_t^2\right] \right) = 2\lambda \mathrm{E}\left[Y_t\right] - \frac{2\mu}{3}\mathrm{E}\left[Y_t^2\right].$$

Therefore, we can write the linear system of differential equations for the second moment of this growth-collapse process as

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix} \mathrm{E}\left[Y_t\right] \\ \mathrm{E}\left[Y_t^2\right] \end{bmatrix} = \begin{bmatrix} -\frac{\mu}{2} & 0 \\ 2\lambda & -\frac{2\mu}{3} \end{bmatrix} \begin{bmatrix} \mathrm{E}\left[Y_t\right] \\ \mathrm{E}\left[Y_t^2\right] \end{bmatrix} + \begin{bmatrix} \lambda \\ 0 \end{bmatrix}.$$

Moving to the third moment, via the infinitesimal generator we write its differential equation as

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[Y_t^3\right] = 3\lambda\mathrm{E}\left[Y_t^2\right] + \mu\left(\mathrm{E}\left[U_1^3 Y_t^3\right] - \mathrm{E}\left[Y_t^3\right]\right) = 3\lambda\mathrm{E}\left[Y_t^2\right] - \frac{3\mu}{4}\mathrm{E}\left[Y_t^3\right],$$

which thus shows that the system of differential equations for the third moment is

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix}\mathrm{E}\left[Y_t\right]\\ \mathrm{E}\left[Y_t^2\right]\\ \mathrm{E}\left[Y_t^3\right]\end{bmatrix} = \begin{bmatrix}-\frac{\mu}{2} & 0 & 0\\ 2\lambda & -\frac{2\mu}{3} & 0\\ 0 & 3\lambda & -\frac{3\mu}{4}\end{bmatrix}\begin{bmatrix}\mathrm{E}\left[Y_t\right]\\ \mathrm{E}\left[Y_t^2\right]\\ \mathrm{E}\left[Y_t^3\right]\end{bmatrix} + \begin{bmatrix}\lambda\\ 0\\ 0\end{bmatrix},$$

which evidently encapsulates the system for the first two moments. We can note that the general n^{th} moment will satisfy

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[Y_t^n\right] = n\lambda\mathrm{E}\left[Y_t^{n-1}\right] + \mu\left(\mathrm{E}\left[U_1^n Y_t^n\right] - \mathrm{E}\left[Y_t^n\right]\right) = n\lambda\mathrm{E}\left[Y_t^{n-1}\right] - \frac{n\mu}{n+1}\mathrm{E}\left[Y_t^n\right],$$

and thus the system for the nth moment is given by appending the row vector

$$\begin{bmatrix}0 & 0 & \dots & 0 & n\lambda & -\frac{n\mu}{n+1}\end{bmatrix}$$

below the matrix from the $(n-1)^{\text{th}}$ system augmented by zeros on the right. Following this derivation, we reach the following general expressions for the moments in Corollary 5.3.5. Furthermore, we can note that because of the relative simplicity of this particular structure, we are able to solve the recursion for the steady-state moments and give these terms explicitly.

Corollary 5.3.5. *Let Y_t be a growth-collapse process with growth rate $\lambda > 0$ and uniformly sized collapses occurring according to a Poisson process with rate $\mu > 0$. Then, the nth moment of Y_t is given by*

$$\begin{aligned}\mathrm{E}\left[Y_t^n\right] = {} & n\lambda\mathbf{v}_{n-1}^{\mathrm{T}}\left(\mathbf{M}_{n-1}^Y + \frac{n\mu}{n+1}\mathbf{I}\right)^{-1}\left(e^{\mathbf{M}_{n-1}^Y t} - e^{-\frac{n\mu t}{n+1}}\mathbf{I}\right)\left(\mathbf{x}_{n-1}(y_0) - \frac{(n+1)\lambda\mathbf{v}_1}{n\mu}\right) + y_0^n e^{-\frac{n\mu t}{n+1}}\\ & + \frac{(n+1)\lambda\mathbb{I}_{\{n=1\}}}{n\mu}\left(1 - e^{-\frac{n\mu t}{n+1}}\right) - \frac{(n+1)\lambda^2}{\mu}\mathbf{v}_{n-1}^{\mathrm{T}}\left(\mathbf{M}_{n-1}^Y\right)^{-1}\left(\mathbf{I} - e^{\mathbf{M}_{n-1}^Y t}\right)\mathbf{v}_1, \qquad (5.26)\end{aligned}$$

where y_0 *is the initial value of* Y_t, $\mathbf{x}_n(a) \in \mathbb{R}^n$ *is such that* $(\mathbf{x}_n(a))_i = a^i$, $\mathbf{M}_n^Y = \lambda \mathbf{diag}(2:n,-1) - \mu \mathbf{diag}\left(\frac{1}{2} : \frac{n}{n+1}\right)$, *and* $\mathbf{m}_n^Y = [\left(\mathbf{M}_n^Y\right)_{n,1}, \ldots, \left(\mathbf{M}_n^Y\right)_{n,n-1}]$ *is such that* $\mathbf{m}_n^Y = n\lambda \mathbf{v}_{n-1}^{\mathrm{T}}$. *Moreover the* n^{th} *steady-state moment of* Y_t *is given by*

$$\lim_{t\to\infty} \mathrm{E}\left[Y_t^n\right] = 2n!\left(\frac{\lambda}{\mu}\right)^n, \tag{5.27}$$

for $n \in \mathbb{Z}^+$.

5.3.6 Application to Ephemerally Self-Exciting Processes

As a final detailed example of the applicability of Matryoshkan matrices, we now consider a stochastic process we have analyzed in Chapter 4. This process is a linear birth-death-immigration process in which the occurrence of an arrival increases the arrival rate by an amount $\alpha > 0$, like in the Hawkes process, and this increase expires after an exponentially distributed duration with some rate $\mu > \alpha$. In this way, this process is an *ephemerally self-exciting* process. Given a baseline intensity $\nu^* > 0$, let Q_t be such that new arrivals occur at rate $\nu^* + \alpha Q_t$ and then the overall rate until the next excitement expiration is μQ_t. One can then think of Q_t as the number of entities still causing active excitement at time $t \geq 0$. We will refer to Q_t as the number in system for this ephemerally self-exciting process. The infinitesimal generator for a function $f : \mathbb{N} \to \mathbb{R}$ is thus

$$\mathcal{L}f(Q_t) = (\nu^* + \alpha Q_t)\left(f(Q_t + 1) - f(Q_t)\right) + \mu Q_t\left(f(Q_t - 1) - f(Q_t)\right),$$

which again captures the dynamics of the process, as the first term on the right hand-side is the product of the up-jump rate and the change in function value upon an increase in the process while the second term is the product of the down-jump rate and the corresponding process decrease. This now yields an

ordinary differential equation for the mean given by

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[Q_t\right] = \nu^* + \alpha\mathrm{E}\left[Q_t\right] - \mu\mathrm{E}\left[Q_t\right] = \nu^* - (\mu - \alpha)\mathrm{E}\left[Q_t\right],$$

while the second moment will satisfy

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[Q_t^2\right] &= \mathrm{E}\left[(\nu^* + \alpha Q_t)\left((Q_t + 1)^2 - Q_t^2\right) + \mu Q_t\left((Q_t - 1)^2 - Q_t^2\right)\right] \\ &= (2\nu^* + \mu + \alpha)\,\mathrm{E}\left[Q_t\right] + \nu^* - 2(\mu - \alpha)\mathrm{E}\left[Q_t^2\right].\end{aligned}$$

Thus, the first two moments are given by the solution to the linear system

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix}\mathrm{E}\left[Q_t\right] \\ \mathrm{E}\left[Q_t^2\right]\end{bmatrix} = \begin{bmatrix}-(\mu - \alpha) & 0 \\ 2\nu^* + \mu + \alpha & -2(\mu - \alpha)\end{bmatrix}\begin{bmatrix}\mathrm{E}\left[Q_t\right] \\ \mathrm{E}\left[Q_t^2\right]\end{bmatrix} + \begin{bmatrix}\nu^* \\ \nu^*\end{bmatrix},$$

and by observing that the third moment differential equation is

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[Q_t^3\right] &= \mathrm{E}\left[(\nu^* + \alpha Q_t)\left((Q_t + 1)^3 - Q_t^3\right) + \mu Q_t\left((Q_t - 1)^3 - Q_t^3\right)\right] \\ &= (3\nu^* + 3\alpha + 3\mu)\,\mathrm{E}\left[Q_t^2\right] + (3\nu^* + \alpha - \mu)\,\mathrm{E}\left[Q_t\right] + \nu^* - 3(\mu - \alpha)\mathrm{E}\left[Q_t^3\right],\end{aligned}$$

we can observe that the third moment system does indeed encapsulate that of the second moment:

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix}\mathrm{E}\left[Q_t\right] \\ \mathrm{E}\left[Q_t^2\right] \\ \mathrm{E}\left[Q_t^3\right]\end{bmatrix} = \begin{bmatrix}-(\mu - \alpha) & 0 & 0 \\ 2\nu^* + \mu + \alpha & -2(\mu - \alpha) & 0 \\ 3\nu^* + \alpha - \mu & 3\nu^* + 3\alpha + 3\mu & -3(\mu - \alpha)\end{bmatrix}\begin{bmatrix}\mathrm{E}\left[Q_t\right] \\ \mathrm{E}\left[Q_t^2\right] \\ \mathrm{E}\left[Q_t^3\right]\end{bmatrix} + \begin{bmatrix}\nu^* \\ \nu^* \\ \nu^*\end{bmatrix}.$$

In general, the nth moment is given by the solution to

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[Q_t^n\right] &= \mathrm{E}\left[(\nu^* + \alpha Q_t)\left((Q_t + 1)^n - Q_t^n\right) + \mu Q_t\left((Q_t - 1)^n - Q_t^n\right)\right] \\ &= \nu^* + \nu^*\sum_{k=1}^{n-1}\binom{n}{k}\mathrm{E}\left[Q_t^k\right] + \alpha\sum_{k=1}^{n}\binom{n}{k-1}\mathrm{E}\left[Q_t^k\right] + \mu\sum_{k=1}^{n}\binom{n}{k-1}\mathrm{E}\left[Q_t^k\right](-1)^{n-k-1},\end{aligned}$$

which means that the nth system is given by augmenting the previous system below by

$$\begin{bmatrix}n\nu^* + \alpha + \mu(-1)^n & \binom{n}{2}\nu^* + n\alpha + n\mu(-1)^{n-1} & \dots & n\nu^* + \binom{n}{2}\alpha + \binom{n}{2}\mu & -n(\mu - \alpha)\end{bmatrix},$$

and to the right by zeros. By comparing this row vector to the definition of the Martyoshkan Pascal matrices in Equation 5.15, we arrive at explicit forms for the moments of this process shown now in Corollary 5.3.6.

Corollary 5.3.6. *Let Q_t be the number in system for an ephemerally self-exciting process with baseline intensity $\nu^* > 0$, intensity jump $\alpha > 0$, and duration rate $\mu > \alpha$. Then, the nth moment of Q_t is given by*

$$\mathrm{E}\left[Q_t^n\right] = \mathbf{m}_n^Q \left(\mathbf{M}_{n-1}^Q + n(\mu-\alpha)\mathbf{I}\right)^{-1} \left(e^{\mathbf{M}_{n-1}^Q t} - e^{-n(\mu-\alpha)t}\mathbf{I}\right)\left(\mathbf{x}_{n-1}(Q_0) - \frac{\nu^*\mathbf{v}}{n(\mu-\alpha)}\right) + Q_0^n e^{-n(\mu-\alpha)t}$$
$$+ \frac{\nu^*}{n(\mu-\alpha)}\left(1 - e^{-n(\mu-\alpha)t}\right) - \frac{\nu^*\mathbf{m}_n^Q}{n(\mu-\alpha)}\left(\mathbf{M}_{n-1}^Q\right)^{-1}\left(\mathbf{I} - e^{\mathbf{M}_{n-1}^Q t}\right)\mathbf{v}, \tag{5.28}$$

for all $t \geq 0$ and $n \in \mathbb{Z}^+$, where $\mathbf{M}_n^Q = \nu^*\boldsymbol{\mathcal{P}}_n(1)\mathbf{diag}(\mathbf{v}, -1) + \alpha\boldsymbol{\mathcal{P}}_n(1) + \mu\boldsymbol{\mathcal{P}}_n(-1)$, and $\mathbf{m}_n^Q = \left[\left(\mathbf{M}_n^Q\right)_{n,1}, \ldots, \left(\mathbf{M}_n^Q\right)_{n,n-1}\right]$ is such that $\left(\mathbf{m}_n^Q\right)_j = \binom{n}{j}\nu^* + \binom{n}{j-1}\alpha + \binom{n}{j-1}\mu(-1)^{n-j-1}$ and $\mathbf{x}_n(a) \in \mathbb{R}^n$ is such that $(\mathbf{x}_n(a))_i = a^i$. In steady-state, the nth moment of Q_t is given by

$$\lim_{t\to\infty} \mathrm{E}\left[Q_t^n\right] = \frac{\nu^*}{\mu-\alpha}\left(1 - \mathbf{m}_n^Q\left(\mathbf{M}_{n-1}^Q\right)^{-1}\mathbf{v}\right), \tag{5.29}$$

for $n \geq 2$ with $\lim_{t\to\infty}\mathrm{E}\left[Q_t\right] = \frac{\nu^*}{\mu-\alpha}$. Moreover the $(n+1)^{th}$ steady-state moment of the ephemerally self-exciting process is given by the recursion

$$\lim_{t\to\infty} \mathrm{E}\left[Q_t^{n+1}\right] = \frac{1}{(n+1)(\mu-\alpha)}\left(\mathbf{m}_{n+1}^Q\mathbf{s}_n^Q + \nu^*\right), \tag{5.30}$$

for all $n \in \mathbb{Z}^+$, where $\mathbf{s}_n^Q \in \mathbb{R}^n$ is the vector of steady-state moments defined such that $\left(\mathbf{s}_n^Q\right)_i = \lim_{t\to\infty}\mathrm{E}\left[Q_t^i\right]$ for $1 \leq i \leq n$.

5.3.7 Additional Applications by Combination and Permutation

While the preceding examples are the the only detailed examples we include in this chapter, we can note that these Matryoshkan matrix methods can be

applied to many other settings. In fact, one can observe that these example derivations can be applied directly to processes that feature a combination of their structures, such as the dynamic contagion process introduced in Dassios and Zhao (2011). The dynamic contagion process is a point process that is both self-excited and externally excited, meaning that its intensity experiences jumps driven by its own activity and by the activity of an exogenous Poisson process. In this way, the process combines the behavior of the Hawkes and shot noise processes. Hence, its infinitesimal generator can be written using a combination of expressions used in Subsections 5.3.2 and 5.3.3, implying that all moments of the process can be calculated through this methodology. Similarly, these methods can also be readily applied to processes that combine dynamics from Hawkes processes and from Itô diffusions, such as affine point processes. These processes, studied in e.g. Errais et al. (2010); Zhang et al. (2015); Gao and Zhu (2019), feature both self-excitement and diffusion behavior and thus have an infinitesimal generator that can be expressed using terms from the generators for Hawkes and Itô processes. Similarly, one could study the combination of externally driven jumps and diffusive behavior such as in affine jump diffusions, see e.g. Duffie et al. (2000). Of course, one can also consider permutations of the model features seen in our examples, such as trading fixed size jumps for random ones to form marked Hawkes processes or changing to randomly sized batches of arrivals in the ephemerally self-exciting process. In general, the key requirement from the assumptions in Theorem 5.3.1 is the closure of the system of moment differential equations specified in Equation 5.7. This is equivalent to having the infinitesimal generator of any polynomial being a polynomial of

order no more than the original. That is, infinitesimal generators of the form

$$\mathcal{L}f(X_t) = \underbrace{(\alpha_0 + \alpha_1 X_t)\,(f(X_t + A_i) - f(X_t))}_{\text{Up-jumps}} + \underbrace{(\alpha_2 + \alpha_3 X_t)\,(f(X_t - B_i) - f(X_t))}_{\text{Down-jumps}}$$
$$+ \underbrace{(\alpha_4 + \alpha_5 X_t)\frac{\mathrm{d}f(X_t)}{\mathrm{d}X_t}}_{\text{Drift, decay, or growth}} + \underbrace{(\alpha_6 + \alpha_7 X_t + \alpha_8 X_t^2)\frac{\mathrm{d}^2 f(X_t)}{\mathrm{d}X_t^2}}_{\text{Diffusion}} + \underbrace{\alpha_9\,(f(C_i X_t) - f(X_t))}_{\text{Expansion or collapse}},$$

can be handled by this methodology, where $\alpha_j \in \mathbb{R}$ for all j and where the sequences $\{A_i\}$, $\{B_i\}$ and $\{C_i\}$ are of mutually independent random variables. Finally we note that this example generator need not be exclusive, as it is possible that other dynamics may also meet the closure requirements in Equation 5.7.

5.4 Complexity Analysis and Numerical Experiments

In this section we address the calculations within this method through comparison with parsimonious differential equation techniques. Specifically, we compare both the result and the calculation time of these two methods in computing the transient moments of these stochastic processes. For the Matryoshkan-based approach, we define the calculation time as the time needed to complete the matrix computations of the moments at the specified point in time. In the differential equation approach, we take the calculation time as the time needed to reach the specified time through applying Euler's iterative method to the linear system of ODE's in Equation 5.12, starting at time 0. We choose to compare to Euler's method because it is the fastest technique available in terms of total run time. Of course in practice in solving differential equations one may use more sophisticated approaches such as higher order Runge-Kutta methods, but such techniques are inherently at least as computationally demanding as Euler's method. As we will see however, our direct Matryoshkan matrix ap-

proach is more efficient than Euler's method outside of very short time intervals, while also delivering much more accurate answers. In Subsection 5.4.1 we compare these two in a formal complexity analysis, and in Subsection 5.4.2 we compare empirically through numerical experiments. Because the same matrices are used in both approaches, the time to form the matrices is omitted from each empirical calculation time reported in Subsection 5.4.2, although we do address the complexity of this pre-computation step in Subsection 5.4.1.

5.4.1 Complexity Characterization

In computing the moments through either the differential equations approach or the Matryoshkan matrix approach, one must first form the matrix that describes the system of differential equations. Thus, before comparing the two methods let us first quickly discuss this common pre-computation step. To form the matrix needed to solve for the nth moment, there will be $O(n^2)$ operations. Specifically, there are $\frac{1}{2}n(n+1)$ elements to write in this lower triangular matrix, which can be naturally conducted by writing vectors of size increasing from 1 to n through the nested Matryoshkan structure of the matrix.

Turning now to Euler's method, we will let $\Delta > 0$ be the time-step size parameter. For simplicity we will assume that the desired time point t is a multiple of Δ, meaning that there will be t/Δ iterations within the method to calculate the moment across time from the known initial value. Since the matrix multiplications will take $O(n^2)$ time at each step, the total complexity of Euler's method will thus be $O(n^2 t/\Delta)$, with the global error known to be $O(\Delta)$ (Butcher, 2016). By comparison in the direct Matryoshkan calculations, the matrix exponential cal-

culations imply that this method is $O(n^3)$. There is of course no dependence on Δ, and there exist methods in which the coefficients on n^3 do not depend on t, but rather only at lower powers of n (Moler and Van Loan, 2003). It is also possible this calculation could be expedited, at least in terms of the hidden coefficients, through leveraging Proposition 5.2.1 and the triangularity of the matrices.

In many applications, we would expect the number of Eulerian time steps should be much larger than the order of the largest moment, i.e. $t/\Delta \gg n$. For example, even at a very high order moment like $n = 100$, taking a modest time step of $\Delta = .01$ would then put times $t \geq 1$ as at least as expensive for Euler's method as for the Matryoshkan approach. Of course, as the time step becomes more refined or as longer time horizons are considered, the Matryoshkan matrix calculations should become even more competitive by comparison. As we have discussed, this superiority in complexity should also immediately extend to comparisons to other numerical differential equation techniques that are themselves more complex than Euler's method. Furthermore, differential equation techniques should incur an time-step dependent error in their solution, and the closed form solutions of the Matryoshkan will not be subject to this.

5.4.2 Empirical Comparisons and Speed Tests

To demonstrate the computational efficiency and numerical precision of this method in practice, we now examine the five detailed examples covered in Subsections 5.3.2, 5.3.3, 5.3.4, 5.3.5, and 5.3.6, apply our Matryoshkan matrix methodology, and compare its performance to solving the differential equations numerically through Euler's method. To compare the results produced by the

two methods, we give the absolute and relative error of Euler's method to the Matryoshkan method. That is, for m_{D} as the Eulerian moment differential equation solution and m_{M} as the Matryoshkan calculated moment, we define these errors as

$$\text{Absolute Error} = |m_{\mathsf{D}} - m_{\mathsf{M}}| \quad \text{and} \quad \text{Relative Error} = \frac{|m_{\mathsf{D}} - m_{\mathsf{M}}|}{m_{\mathsf{M}}}.$$

All calculations are performed using simple MATLAB code on a 64-bit Windows machine with 16 GB of memory. In all four examples, we evaluate four different step sizes for Euler's method: 0.01, 0.001, 0.0001, and 0.00001. All time and error results presented in the following tables are found through averaging the results of 20 trial calculations per scenario.

We begin with the moments of the Hawkes process intensity, the first example we have discussed and the original motivation for this work. It is worth noting that to the best of our knowledge even the recognition of the matrix structure of the moments ODE system is a new contribution, as the highest order moment with explicitly stated differential equation is the fourth moment, presented without proof or solution in **?**. Similarly, the highest order moment with closed form solution (either transient or stationary) previously given in the literature appears to be the second. In Table 5.1, we give the errors and the time incurred for computing the first 4, 10, 20, and 100 moments. For this example we take a baseline intensity of $\lambda^* = 1$, an intensity jump size of $\alpha = 1$, a decay rate of $\beta = 2$, and we compute the moments for time $t = 10$. Additionally, we assume the initial value of the intensity is equal to the baseline. As can be quickly observed, the calculation time in the Matryoshkan computation outpaces Euler's method in all scenarios regardless of the Eulerian step size. Furthermore, when Euler's method is performed with a smaller, more precise step size, we find that its solution is increasingly close to the solution given by the Matryoshkan

method as both the relative error and the absolute error decrease with the step size. In the most precise setting for Euler's method, the Matryoshkan method's run time is 4 orders of magnitude smaller for the 4, 10, and 20 moment settings and 3 orders of magnitude smaller for the 100th moment.

$n = 4$	Run Time	Absolute Error	Relative Error
Matryoshkan	2.1×10^{-4} sec	·	·
Euler $\Delta = 10^{-2}$	1.4×10^{-3} sec	3.0×10^{-4}	5.0×10^{-6}
Euler $\Delta = 10^{-3}$	1.3×10^{-2} sec	3.1×10^{-5}	5.1×10^{-7}
Euler $\Delta = 10^{-4}$	1.3×10^{-1} sec	3.1×10^{-6}	5.2×10^{-8}
Euler $\Delta = 10^{-5}$	1.3×10^{0} sec	3.1×10^{-7}	5.2×10^{-9}
$n = 10$	Run Time	Absolute Error	Relative Error
Matryoshkan	2.3×10^{-4} sec	·	·
Euler $\Delta = 10^{-2}$	1.5×10^{-3} sec	4.4×10^{1}	1.0×10^{-5}
Euler $\Delta = 10^{-3}$	1.4×10^{-2} sec	4.5×10^{0}	1.1×10^{-6}
Euler $\Delta = 10^{-4}$	1.4×10^{-1} sec	4.5×10^{-1}	1.1×10^{-7}
Euler $\Delta = 10^{-5}$	1.4×10^{0} sec	4.5×10^{-2}	1.1×10^{-8}
$n = 20$	Run Time	Absolute Error	Relative Error
Matryoshkan	2.5×10^{-4} sec	·	·
Euler $\Delta = 10^{-2}$	1.5×10^{-3} sec	7.8×10^{12}	1.7×10^{-5}
Euler $\Delta = 10^{-3}$	1.5×10^{-2} sec	8.0×10^{11}	1.8×10^{-6}
Euler $\Delta = 10^{-4}$	1.4×10^{-1} sec	8.0×10^{10}	1.8×10^{-7}
Euler $\Delta = 10^{-5}$	1.4×10^{0} sec	8.0×10^{9}	1.8×10^{-8}
$n = 100$	Run Time	Absolute Error	Relative Error
Matryoshkan	4.3×10^{-3} sec	·	·
Euler $\Delta = 10^{-2}$	5.0×10^{-3} sec	3.0×10^{145}	5.1×10^{-5}
Euler $\Delta = 10^{-3}$	4.9×10^{-2} sec	3.1×10^{144}	5.2×10^{-6}
Euler $\Delta = 10^{-4}$	4.6×10^{-1} sec	3.1×10^{143}	5.2×10^{-7}
Euler $\Delta = 10^{-5}$	4.4×10^{0} sec	3.1×10^{142}	5.3×10^{-8}

Table 5.1: Comparison of run time and errors for Hawkes process moment calculation via Matryoshkan matrix method and via Euler differential equation methods as the moment size increases.

We find similarly effective performance for the shot noise process. In the example shown in Table 5.2, we suppose that the Poisson process arrival rate is $\lambda = 1$ and that the distribution of the shot noise is LogNormal with $\mu = 0$ and $\sigma = 1$. Furthermore, we assume that the exponential decay rate is $\beta = 4$ and we evaluate the moments at time $t = 5$. Because of the scale of these moments, we

now perform these computations for $n = 5$, 10, 15, and 20. Again we see that as the step size in Euler's method decreases the differences between the pair of results shrinks, although in this case the run times of the Matryoshkan method and Euler's method with step size 0.01 are of the same magnitude. Nevertheless, as Euler's method grows increasingly precise the Matryoshkan approach becomes more favorable; its run time is 3 orders of magnitude smaller than the most precise Euler's computational duration.

$n = 5$	Run Time	Absolute Error	Relative Error
Matryoshkan	2.1×10^{-4} sec	·	·
Euler $\Delta = 10^{-2}$	1.9×10^{-4} sec	1.3×10^{-9}	3.2×10^{-10}
Euler $\Delta = 10^{-3}$	2.0×10^{-3} sec	1.4×10^{-10}	3.5×10^{-11}
Euler $\Delta = 10^{-4}$	1.9×10^{-2} sec	1.4×10^{-11}	3.5×10^{-12}
Euler $\Delta = 10^{-5}$	1.9×10^{-1} sec	1.6×10^{-12}	3.9×10^{-13}
$n = 10$	Run Time	Absolute Error	Relative Error
Matryoshkan	2.2×10^{-4} sec	·	·
Euler $\Delta = 10^{-2}$	1.9×10^{-4} sec	2.6×10^{2}	5.6×10^{-10}
Euler $\Delta = 10^{-3}$	1.9×10^{-3} sec	2.9×10^{1}	6.0×10^{-11}
Euler $\Delta = 10^{-4}$	1.9×10^{-2} sec	2.9×10^{0}	6.1×10^{-12}
Euler $\Delta = 10^{-5}$	1.9×10^{-1} sec	3.2×10^{-1}	6.7×10^{-13}
$n = 15$	Run Time	Absolute Error	Relative Error
Matryoshkan	2.6×10^{-4} sec	·	·
Euler $\Delta = 10^{-2}$	2.2×10^{-4} sec	1.6×10^{39}	8.4×10^{-10}
Euler $\Delta = 10^{-3}$	2.0×10^{-3} sec	1.8×10^{38}	9.1×10^{-11}
Euler $\Delta = 10^{-4}$	2.0×10^{-2} sec	1.8×10^{37}	9.2×10^{-12}
Euler $\Delta = 10^{-5}$	2.0×10^{-1} sec	2.0×10^{36}	1.0×10^{-12}
$n = 20$	Run Time	Absolute Error	Relative Error
Matryoshkan	3.4×10^{-4} sec	·	·
Euler $\Delta = 10^{-2}$	2.1×10^{-4} sec	2.1×10^{115}	1.1×10^{-9}
Euler $\Delta = 10^{-3}$	2.1×10^{-3} sec	2.3×10^{114}	1.2×10^{-10}
Euler $\Delta = 10^{-4}$	2.1×10^{-2} sec	2.3×10^{113}	1.2×10^{-11}
Euler $\Delta = 10^{-5}$	2.1×10^{-1} sec	2.5×10^{112}	1.3×10^{-12}

Table 5.2: Comparison of run time and errors for shot noise process moment calculation via Matryoshkan matrix method and via Euler differential equation methods as the moment size increases.

In Table 5.3 we perform these computational experiments for the moments of a Cox-Ingersoll-Ross (CIR) process, which is an Itô diffusion with parameter

$\gamma = 1$. Moreover we assume that $\mu = 1$, $\theta = 1$, and $\sigma = 1$, and we compute the moments at time $t = 5$ for $n = 4$, 10, 20, and 100. Like in the Hawkes process example, the Matryoshkan approach outperforms Euler's method in terms of calculation time in this example in all moment scenarios and step sizes. Moreover we again see that as the step size decreases the error between the two methods decreases while the Eulerian computation time becomes much larger than the Matryoshkan run time. Specifically for the first 4, 10, and 20 moment calculation experiments the Matryoshkan is faster by 4 orders of magnitude and in the $n = 100$ setting is is 3 orders of magnitude faster.

$n = 4$	Run Time	Absolute Error	Relative Error
Matryoshkan	2.3×10^{-4} sec	·	·
Euler $\Delta = 10^{-2}$	1.4×10^{-3} sec	6.8×10^{-3}	9.3×10^{-4}
Euler $\Delta = 10^{-3}$	1.3×10^{-2} sec	4.8×10^{-4}	6.6×10^{-5}
Euler $\Delta = 10^{-4}$	1.3×10^{-1} sec	4.9×10^{-5}	6.6×10^{-6}
Euler $\Delta = 10^{-5}$	1.3×10^{0} sec	6.8×10^{-6}	9.4×10^{-7}
$n = 10$	Run Time	Absolute Error	Relative Error
Matryoshkan	2.4×10^{-4} sec	·	·
Euler $\Delta = 10^{-2}$	1.4×10^{-3} sec	8.2×10^{2}	2.3×10^{-3}
Euler $\Delta = 10^{-3}$	1.3×10^{-2} sec	5.8×10^{1}	1.6×10^{-4}
Euler $\Delta = 10^{-4}$	1.3×10^{-1} sec	5.8×10^{0}	1.6×10^{-5}
Euler $\Delta = 10^{-5}$	1.3×10^{0} sec	8.3×10^{-1}	2.3×10^{-6}
$n = 20$	Run Time	Absolute Error	Relative Error
Matryoshkan	2.7×10^{-4} sec	·	·
Euler $\Delta = 10^{-2}$	1.4×10^{-3} sec	1.8×10^{11}	4.3×10^{-3}
Euler $\Delta = 10^{-3}$	1.3×10^{-2} sec	1.3×10^{10}	2.9×10^{-4}
Euler $\Delta = 10^{-4}$	1.3×10^{-1} sec	1.3×10^{9}	2.9×10^{-5}
Euler $\Delta = 10^{-5}$	1.3×10^{0} sec	1.8×10^{8}	4.3×10^{-6}
$n = 100$	Run Time	Absolute Error	Relative Error
Matryoshkan	2.0×10^{-3} sec	·	·
Euler $\Delta = 10^{-2}$	2.8×10^{-3} sec	4.8×10^{127}	1.3×10^{-2}
Euler $\Delta = 10^{-3}$	2.7×10^{-2} sec	2.1×10^{126}	5.6×10^{-4}
Euler $\Delta = 10^{-4}$	2.8×10^{-1} sec	2.1×10^{125}	5.6×10^{-5}
Euler $\Delta = 10^{-5}$	2.7×10^{0} sec	4.7×10^{124}	1.2×10^{-5}

Table 5.3: Comparison of run time and errors for CIR process moment calculation via Matryoshkan matrix method and via Euler differential equation methods as the moment size increases.

We evaluate the Matryoshkan matrix method for the growth-collapse process in Table 5.4. Like in Section 4 of Boxma et al. (2006), we take $\lambda = 1$ and we also set $\mu = \frac{1}{2}$. We evaluate the moments at time $t = 8$ and the moments $n = 5$, 10, 15, and 20. In addition to observing that the Matryoshkan approach is an order of magnitude faster than any of the differential equation methods, we can also note that the relative errors are the largest we have seen in these experiments across all the Eulerian step sizes. At best, the relative error is of order 10^{-6}, and in this case the Matryoshkan method run time is four orders of magnitude faster. As was the case for each of the other processes, we can observe that as the step size is decreased, the increased precision in Euler's method yields results closer and closer to the moments calculated by the Matryoshkan approach.

As a final table of computation comparisons, we now evaluate the moments of the Affine Queue-Hawkes process with baseline intensity $\nu^* = 1$, intensity jump size $\alpha = 2$, and duration rate $\mu = 3$. Table 5.5 contains the calculation times and errors for computing the first 4, 10, 20, and 100 moments of this process at time $t = 5$. Again a familiar pattern emerges, as the Matryoshkan method performance is comparable or better relative to Euler's method across all experiment scenarios. At the largest Eulerian step size the run times are of the same order but for each order of magnitude decrease in step size the method's step size becomes approximately a factor of 10 times slower than the Matryoshkan calculation. As this step size decreases the error between the two computations again decreases, implying that the Matryoshkan method also outperforms the differential equation approach in accuracy.

$n = 5$	Run Time	Absolute Error	Relative Error
Matryoshkan	2.2×10^{-4} sec	·	·
Euler $\Delta = 10^{-2}$	1.9×10^{-3} sec	4.9×10^{0}	1.4×10^{-3}
Euler $\Delta = 10^{-3}$	1.1×10^{-2} sec	7.3×10^{-1}	2.1×10^{-4}
Euler $\Delta = 10^{-4}$	1.1×10^{-1} sec	4.9×10^{-2}	1.4×10^{-5}
Euler $\Delta = 10^{-5}$	1.1×10^{0} sec	4.9×10^{-3}	1.4×10^{-6}
$n = 10$	Run Time	Absolute Error	Relative Error
Matryoshkan	2.2×10^{-4} sec	·	·
Euler $\Delta = 10^{-2}$	1.2×10^{-3} sec	1.1×10^{6}	1.7×10^{-2}
Euler $\Delta = 10^{-3}$	1.2×10^{-2} sec	1.6×10^{5}	2.6×10^{-3}
Euler $\Delta = 10^{-4}$	1.2×10^{-1} sec	1.1×10^{4}	1.7×10^{-4}
Euler $\Delta = 10^{-5}$	1.2×10^{0} sec	1.1×10^{3}	1.7×10^{-5}
$n = 15$	Run Time	Absolute Error	Relative Error
Matryoshkan	2.4×10^{-4} sec	·	·
Euler $\Delta = 10^{-2}$	1.2×10^{-3} sec	9.7×10^{10}	6.3×10^{-2}
Euler $\Delta = 10^{-3}$	1.2×10^{-2} sec	1.2×10^{10}	7.9×10^{-3}
Euler $\Delta = 10^{-4}$	1.2×10^{-1} sec	9.9×10^{8}	6.4×10^{-4}
Euler $\Delta = 10^{-5}$	1.2×10^{0} sec	9.9×10^{7}	6.4×10^{-5}
$n = 20$	Run Time	Absolute Error	Relative Error
Matryoshkan	2.6×10^{-4} sec	·	·
Euler $\Delta = 10^{-2}$	1.2×10^{-3} sec	5.7×10^{15}	1.3×10^{-1}
Euler $\Delta = 10^{-3}$	1.2×10^{-2} sec	6.9×10^{14}	1.6×10^{-2}
Euler $\Delta = 10^{-4}$	1.2×10^{-1} sec	6.1×10^{13}	1.4×10^{-3}
Euler $\Delta = 10^{-5}$	1.2×10^{0} sec	6.1×10^{12}	1.4×10^{-4}

Table 5.4: Comparison of run time and errors for growth-collapse process moment calculation via Matryoshkan matrix method and via Euler differential equation methods as the moment size increases.

5.5 Conclusion

In this work, we have defined a novel sequence of matrices called Matryoshkan matrices that stack like their Russian nesting doll namesakes. In doing so, we have found a computationally efficient manner of calculating the moments of a large class of Markov processes that satisfy a closure condition for the time derivatives of their transient moments. Furthermore, this has yielded closed form expressions for the transient and steady state moments of these process. Notably, this includes the intensity of the Hawkes process, for which finding

$n = 4$	Run Time	Absolute Error	Relative Error
Matryoshkan	2.0×10^{-4} sec	·	·
Euler $\Delta = 10^{-2}$	3.3×10^{-4} sec	1.7×10^{-1}	7.8×10^{-4}
Euler $\Delta = 10^{-3}$	3.2×10^{-3} sec	1.2×10^{-2}	5.6×10^{-5}
Euler $\Delta = 10^{-4}$	3.2×10^{-2} sec	1.2×10^{-3}	5.6×10^{-6}
Euler $\Delta = 10^{-5}$	3.1×10^{-1} sec	1.7×10^{-4}	7.9×10^{-7}
$n = 10$	Run Time	Absolute Error	Relative Error
Matryoshkan	2.3×10^{-4} sec	·	·
Euler $\Delta = 10^{-2}$	3.5×10^{-4} sec	8.5×10^{6}	1.9×10^{-3}
Euler $\Delta = 10^{-3}$	3.3×10^{-3} sec	6.1×10^{5}	1.3×10^{-4}
Euler $\Delta = 10^{-4}$	3.3×10^{-2} sec	6.1×10^{4}	1.3×10^{-5}
Euler $\Delta = 10^{-5}$	3.3×10^{-1} sec	8.6×10^{3}	1.9×10^{-6}
$n = 20$	Run Time	Absolute Error	Relative Error
Matryoshkan	2.6×10^{-4} sec	·	·
Euler $\Delta = 10^{-2}$	3.6×10^{-4} sec	6.2×10^{22}	3.6×10^{-3}
Euler $\Delta = 10^{-3}$	3.5×10^{-3} sec	4.3×10^{21}	2.5×10^{-4}
Euler $\Delta = 10^{-4}$	3.5×10^{-2} sec	4.3×10^{20}	2.5×10^{-5}
Euler $\Delta = 10^{-5}$	3.5×10^{-1} sec	6.2×10^{19}	3.6×10^{-6}
$n = 100$	Run Time	Absolute Error	Relative Error
Matryoshkan	4.3×10^{-3} sec	·	·
Euler $\Delta = 10^{-2}$	1.9×10^{-3} sec	5.3×10^{193}	1.2×10^{-2}
Euler $\Delta = 10^{-3}$	1.8×10^{-2} sec	2.7×10^{192}	6.3×10^{-4}
Euler $\Delta = 10^{-4}$	1.8×10^{-1} sec	2.7×10^{191}	6.2×10^{-5}
Euler $\Delta = 10^{-5}$	1.7×10^{0} sec	5.1×10^{190}	1.2×10^{-5}

Table 5.5: Comparison of run time and errors for Affine Queue-Hawkes process moment calculation via Matryoshkan matrix method and via Euler differential equation methods as the moment size increases.

an expression for the nth moment had been an open problem. Other examples we have discussed include Itô diffusions from the mathematical finance literature and shot noise processes from the physics literature, which showcases the breadth of this methodology. Furthermore, our computational experiments have demonstrated the efficiency of computing at a point in time rather than through time, which is a key benefit of this method over traditional approaches for solving differential equations numerically.

We can note that there are many applications of this methodology that we

have not explored in this chapter and are thus opportunities for future work. For example, the vector form of the moments arising from this matrix based method naturally lends itself to use in the method of moments. Thus, Matryoshkan matrices have the potential to greatly simplify estimation for the myriad of Markov processes to which they apply. Additionally, this vector of solutions may also be of use in providing computationally tractable approximations of moment generating functions. That is, by a Taylor expansion one can approximate a moment generating function by a weighted sum of its moments. Because this chapter's Matryoshkan matrix methods enable efficient calculation of higher order moments, this enables higher order approximations of the moment generating function.

As another important direction of future work, we are also interested in extending these techniques to multivariate Markov processes. This is of practical relevance in many of the settings we have described, such as point processes driven by the Hawkes or shot noise process intensities. The challenge in this case arises in the fact that a moment's differential equation now depends on the lower product moments rather than just the lower moments, so the nesting structure is not as neatly organized. Nevertheless, addressing this generalization is an extension worth pursuing, as this would render these techniques even more applicable.

CHAPTER 6

STAFFING A TELEOPERATIONS SYSTEM FOR AUTONOMOUS VEHICLES

6.1 Introduction

With apologies to the likes of Bruce Springsteen and Frank Ocean, it seems that the car may not be a symbol of individual experiences for much longer. A recent market report from Intel and Strategy Analytics (Lanctot (2017)) sees a "generational sea change" in which consumers and businesses alike shift from vehicle ownership towards mobility-as-a-service. This study projects that in 2050 this *passenger economy* will generate a global revenue of $7 trillion, defined and driven by the proliferation of shared autonomous vehicles. Such potent predictions for the value of autonomous vehicles are not uncommon. For example, a recent report from Allied Market Research (Jadhav (2018)) projects that the market value of autonomous vehicles will grow from $54.23 billion in 2019 to $556.67 billion in 2026. This valuation includes the sale and development of driverless car technology, as well as revenue from mobility-as-a-service applications such as taxis/ride share, freight, and public transit. Even taking economic values aside, autonomous vehicles are also poised to bring broad societal benefits, offering safer, smarter, and more sustainable transit, as detailed in Burns (2013). These landmark changes expected from autonomous vehicles are predi-

Contents of this chapter are, at the time of this dissertation's writing, under review for publication and are publicly available as a preprint (Daw et al., 2019). Robert C. Hampshire is also a co-author of that paper, and his contributions, guidance, and insights on this chapter are greatly appreciated.

cated on a projected progression in their capability. The Society of Automotive Engineers (SAE) has created an industry standard for classifying driving automation, distilled into six levels in SAE On-Road Automated Vehicle Standards Committee (2018). These classifications are based around the relative responsibilities of humans and automation, ranging from no automation (level 0) to full automation (level 5). Level 4, the most advanced classification that has been achieved, is characterized by the vehicle handling all driving functionality so long as the current conditions satisfy some constraints, such as the route being within a specific well-mapped, geo-fenced area. Whenever these constraints are not satisfied, a human driver must assume the vehicle's operation or otherwise the car cannot function. Level 4 automation is exemplified by the localized driverless shuttle or taxi services that have been offered in test by companies such as Uber and Waymo (Google's self driving effort). To achieve the full market earnings projected in Lanctot (2017), driverless car technology must advance from level 4 to level 5; all constraints must be removed.

However, there have been recent concerns that this might be an impossible goal. In November 2018, Waymo CEO John Krafcik said publicly that autonomous vehicles won't ever be able to drive in all conditions, meaning that level 5 automation will never be achieved Tibken (2018). Recently, Kalra and Paddock (2016) have also found that it may take up to hundreds of billions of miles to be driven by autonomous vehicles before we can confidently conclude that driverless cars are as reliable as human drivers. The underlying source of these concerns is that there are too many exceptional circumstances vehicles can encounter and too many different situations that can occur while driving. Further complicating this predicament is that there are unknown unknowns, or as stated by Krafcik, "you don't know what you don't know" Tibken (2018). In a

sense, this means that the constraints in the vehicles' level 4 capability may not be known, or that knowing the constraints may require particularly restrictive conditions. Hence, the presence of these edge cases has shifted the expectation for autonomous vehicles; they are now projected to remain at level 4 autonomy for the foreseeable future. While the constraints may become less strict as the technology improves, the prevailing thought is that these limitations will continue to exist, and this persistence means that autonomous vehicles must continue to rely on human expertise to be able to function.

In level 4 automation, the times at which an autonomous vehicle needs human help are called **disengagements**, meaning that the vehicle disengages its autonomous operation. By definition, disengagements occur when the car is faced with an unacceptable level of uncertainty, effectively exceeding the constraints of its abilities. Currently, the standard practice of handling disengagements is to have an **in-car safety driver**, meaning a person in the drivers seat who takes control of the autonomous vehicle when needed. This is of course inefficient at scale, as it is a one-to-one pairing of people to cars at all times. What else can be done to safeguard driverless cars? This question has been of interest to both government and industry; legislation and LLC's alike are being created in response to it. At their core, the common idea behind these new ventures is essentially a call center for driverless cars. Referred to as autonomous vehicle **teleoperations systems**, these remote support centers move the human operators from behind the wheel to a centralized location, so that one person has the potential to help many different vehicles across the course of the day.

As an example of the development of teleoperations centers, the state of California has recently introduced regulations requiring a **remote operator** –

Figure 6.1: An example remote operation setup used by the startup Designated Driver (2019).

meaning a person not in the vehicle who can monitor and control the car when needed – for testing autonomous vehicles that are truly driverless and do not have a safety driver onboard (California Department of Motor Vehicles (2018)). Other states are beginning to follow California's lead. As of July 2019, five more states require teleoperation systems in the absence of drivers and another five allow teleoperation but do not mandate it, per Rosenszweig (2019). Outside of the U.S., five other countries have legislation allowing or mandating teleoperation: Canada (Ontario specifically), Japan, Finland, the Netherlands, and the United Kingdom Rosenszweig (2019). As described in Davies (2019), this is also seen in the private sector, where both startups and major car companies are engineering new technology to ensure the safety, reliability, and efficiency of these remote teleoperation systems for autonomous vehicles. By introducing the opportunity to have human driver input in uncertain scenarios, these remote operator centers enable autonomous vehicles to function in environments they otherwise could not. Based on the mission statements of recent startups, this modern service system can cater to its driverless car customers in differ-

ent ways. One type of teleoperation explored by these companies is **real-time remote operation**, such as what is offered by start-ups like Designated Driver and Phantom Auto. This involves a remote operator taking over the driving of the car for a period of time via a teleoperations system. An example of this technology as used by Designated Driver is given in Figure 6.1.

A recent line of promising and continuing development in teleoperations technology incorporates **human-in-the-loop** approaches in which the remote operators provide input without directly driving the vehicle. For example, the startup Ottopia offers a service which they refer to as "advanced teleoperation," as described in Sawers (2018). Rather than having teleoperators assume full control of the driving, Ottopia's advanced teleoperation utilizes "path choice" or "path drawing," in which remote operators either select or draw a path for the vehicle to take. In this way, this service combines human expertise and machine execution rather than switching between the two. An artistic rendering of Ottopia's path choice procedure is shown in Figure 6.2. Major car companies are also pursuing combinations of this type. For example, Nissan has recently unveiled a teleoperation service they refer to as "Seamless Autonomous Mobility," which was inspired by NASA's interplanetary robot supervision software, Visual Environment for Remote Virtual Exploration. Like Ottopia's path drawing, Nissan's Seamless Autonomous Mobility also uses a line drawing interface for the teleoperations, which they refer to as the operator painting a path for the vehicle to take. This interface is visualized in Figure 6.3, in which the teleoperator's painted path can be seen in the center of the image.

Because this type of passive input pairs machine execution and human expertise without requiring direct driving control by the remote operator, it creates

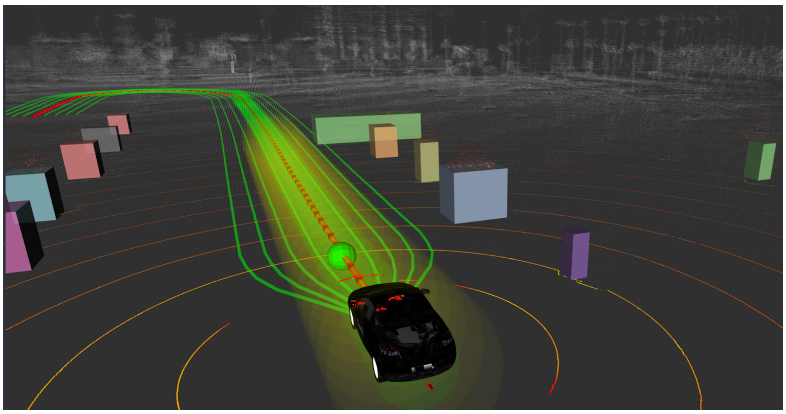


Figure 6.2: A visualization of "path choice" within Ottopia's advanced teleoperation Ottopia Team (2019).

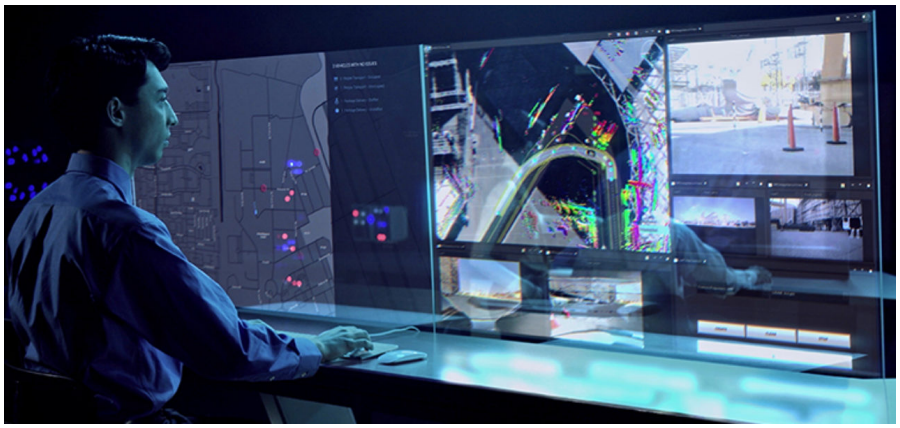


Figure 6.3: The line-drawing interface used in Nissan's Seamless Autonomous Mobility Designated Driver (2019).

an opportunity for assistance that is faster than the time to drive. Through the recent work in Lundgard et al. (2018); Chung et al. (2019), this type of teleoperation service can even be *instantaneously crowdsourced* by pre-fetching input on possible future scenarios simulated from the current state of a vehicle. The idea of this technique for driverless car assistance is visualized in Figure 6.4. When the autonomous vehicle is approaching uncertainty ahead, it sends its current state information to the teleoperations center and requests support. Using this information, simulations of the car's future environment are generated and passed to a pool of remote operators. These operators then supply quick

and specific driving instructions, such as path choice or path drawing, for each of the simulated scenarios, and this input is then passed back to the vehicle before it encounters the time of the simulated scenarios. At that point in time, the autonomous vehicle can now reduce its uncertainty and determine an action via its library of human assistance, which spans a variety of possible environments. Using language from Lundgard et al. (2018), we refer to this human-supported AI service as **look-ahead assistance**. By this construction, one can note that look-ahead assistance offers a just-in-time policy training and generation procedure within a reinforcement learning framework.

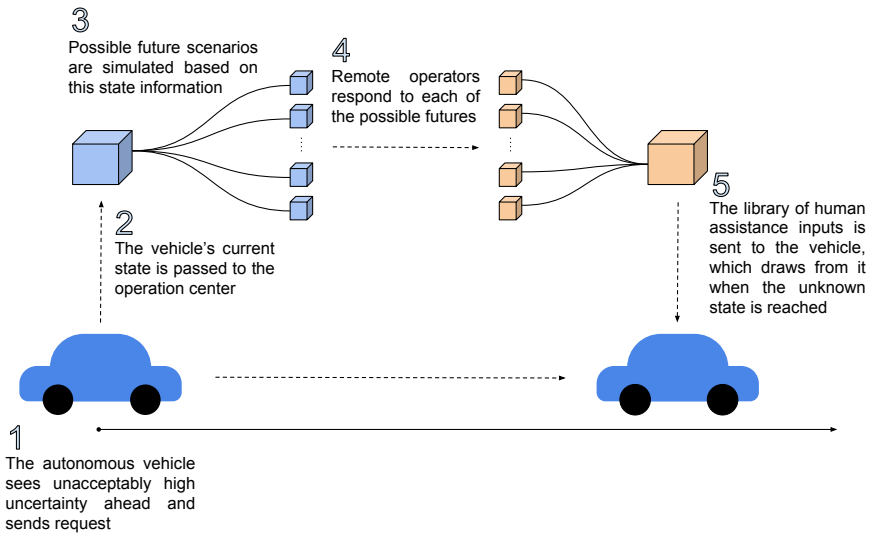


Figure 6.4: A visual guide to the human-in-the-loop AI look-ahead assistance, as based on the description in Lundgard et al. (2018).

At each disengagement, look-ahead assistance creates batches of tasks to be handled by the teleoperators. Because the simulation step generates a collection of possible future scenarios to be evaluated, each time a vehicle requests human support it will actually engage many people at once. From the core components of look-ahead assistance, we can actually reason that the support system becomes safer and safer if the batches in this procedure are to become larger

and larger. For example, from the fact that the simulation is generating different future states the vehicle could encounter, increasing the batch size from this perspective would mean that additional possible scenarios are being explored and evaluated. Even if these tail events are highly unlikely to occur, this still provides added human assurance for the vehicle and can even be used for additional training of the vehicle. As another source of larger batch sizes, it has been seen that adopting intentional redundancy and assigning the same job to multiple workers can increase the accuracy of the crowd-sourced response, as noted in Lundgard et al. (2018). Hence, by duplicating scenarios within the batch of simulated future environments, look-ahead assistance becomes more robust.

This dynamic underscores the tradeoff between in-car safety drivers and remote teleoperations. An in-car safety driver is a one-to-one pairing of operators and vehicles at all times. By comparison, look-ahead assistance is a many-to-one relationship of operators and vehicles, but only when the vehicles need help. Naturally, this raises the question: which pairing is more efficient? Furthermore, this leads to the important long run concern: can teleoperations systems help lift level 4 automation to achieve the vision of the potentially unattainable level 5? To answer these questions, we must first solve the problem of staffing a teleoperations system. As motivated by the preceding discussion, the structure of look-ahead assistance leads us to model the teleoperations service system as a queue with batch arrivals, particularly one in which the batches are large in size. We will study the $G_t^{B(n)}/GI/cn$ model, i.e. a queue receiving batch arrivals of size drawn from an i.i.d. sequence at epochs generated by some general, possibly correlated and time-varying point process, in which each job in the batch receives service of general i.i.d. duration from one of a finitely many servers. In

spirit, our approach is effectively a batch analog of well-known multi-server heavy traffic limits for queueing models, such as in the seminal work from Halfin and Whitt (1981). By comparison to those limits in which the arrival rate and the number of servers both grow linearly, in the **large batch analysis** we conduct, we will study how the queue length process changes as both the batch size and the number of servers grow linearly.

In this limiting analysis, we prove new connections between queues and storage processes, and these connections enable us to develop staffing methodologies for queues with large batch arrivals. In the context of the teleoperations system, we will use these results to determine the number of remote operators needed to reliably support the center's autonomous vehicles. In doing so, we find that self-driving technology is already at a level of performance that suggests that teleoperations systems are substantially more efficient than in-car safety drivers. Moreover, in a long-term view this motivates teleoperations systems as the key to helping the tractability of level 4 automation achieve the high value impact of the ideal but potentially unachievable level 5. Human expertise offers the assurance to assist driverless vehicles when unexpected, uncertain situations occur, and we find that teleoperations systems offer ways to deliver this assistance efficiently. While the teleoperations context will be the focus of this chapter, our large batch analysis of course need not only apply in this setting. Large batches of jobs can occur in data processing centers, mass-transit systems, and disasters or mass casualty emergencies.

6.1.1 Review of Relevant Literature

The large batch setting in this chapter separates this work from previous queueing theoretic studies on batch arrival multi-server queues, which either assume a less general arrival process than we consider here or, perhaps more critically, assume bounded batch sizes. For example, Neuts (1978) and Baily and Neuts (1981) each use matrix-geometric approaches to study the $GI^B/M/c$ queue under the assumption that the batch size distribution B is bounded, with Neuts (1978) considering the stationary setting and Baily and Neuts (1981) the transient. In each setting, the bounded-ness of the batch size is essential, as this bound dictates the size of the underlying matrices. This bounded batch size assumption is also used to study the $GI^B/M/c$ model in Zhao (1994) and Chaudhry and Kim (2016), with the former giving explicit expressions for the generating function and an equation satisfied by the steady-state probabilities and the latter providing efficient computational methods while also simplifying the approach of the former. Again the bound is essential, as these approaches are built upon root-finding methods where the number of roots is equal to the batch size. By comparison, the general unbounded batch size setting has often called for approximate approaches, such as the bounds on the $GI^B/G/c$ system that were constructed in Yao et al. (1984) through comparison to single arrival queues. Yao (1985) then gives tighter bounds for the $M^B/M/c$ queue using the $M^B/G/1$ system and demonstrates that these bounds can be used to approximate the $GI^B/G/c$. Computational methods have also been provided in Cromie et al. (1979) for the fully Markovian setting, the $M^B/M/c$ queue, although these were only done for three specific batch distributions: constant size, geometric, and Poisson.

In studying this large batch setting we will prove batch scaling limits of the

queue, in which the batch size and the number of servers grow large and the queue length is scaled inversely. Through the batch scaling limits, we connect the general batch arrival queueing models to storage processes, another class of stochastic processes. Similar albeit less general scalings have been explored recently in de Graaf et al. (2017) and in Chapter 3 of this dissertation, although the limiting process was not characterized in Chapter 3 . Specifically, the limits we prove in this work for the $G^{B(n)}/GI/cn$ queue generalize the batch scaling results of $M^{B(n)}/M/\infty$ queueing systems shown in de Graaf et al. (2017) and in Chapter 3, which converge to shot noise processes with exponential decay. The limiting relationship between infinite server queues and shot noise processes was also discussed as motivation in Kella and Whitt (1999), although this relationship was presented without proof. This connection allows us to make use of a broad literature on storage processes, which can be seen as a generalization of shot noise processes.

Storage processes, which can also be referred to as dams, content processes, or even fluid queues, are positive valued, continuous time stochastic processes in which the process level will jump upwards by some amount at epochs given by a point process. Between jumps the process will decrease according to some function of its state. That is, there is a function, often denoted $r : \mathbb{R}^+ \to \mathbb{R}^+$, such that the rate of the process's decline when in state x is $r(x)$. For example, $r(x) \propto x$ would recover the exponential decay of a shot noise process. In generality, the release dynamics may also be a function of the history of the process rather than just the current state, such a setting will be necessary to study the multi-server queue's limiting form in the case of non-Markovian service.

Because storage processes have a long history of study, we are able to draw

upon a rich literature of interesting ideas. Many of the results that will be most relevant to us are focused on the stationary distributions of storage processes. Even on its own the study of stationary distributions of storage processes has a rich history, with early work including expressions of stationary distributions for shot noise processes given in Gilbert and Pollak (1960). Later work found similar results for more general settings, including Cinlar and Pinsky (1972); Yeo (1974, 1976); Rubinovitch and Cohen (1980); Kaspi (1984). A line of study that will be particularly useful for us can be found in Brockwell (1977); Brockwell et al. (1982), as these works find integral equations for the stationary distributions of storage processes with a general release rule $r(\cdot)$. These forms will be of great use to us in our staffing analysis. For precursors to this work in a different but no less interesting setting, see Harrison and Resnick (1976, 1978). Another elegant area of study is the duality of the storage processes, for example see Kaspi and Perry (1989); Perry and Stadje (2003). Connections between queues and storage processes are not new in general, as the single server queue has been known to be directly related to storage processes. For an overview of these connections and the related ideas, see Prabhu (2012).

Our analysis of the multi-server queueing model is predicated on an understanding of the infinite server model, and thus we also prove a generalization of the limits in de Graaf et al. (2017) and in Chapter 3 for the $G_t^{B(n)}/GI/\infty$ model. To prove this generalization beyond the Markovian setting, we develop an approach that is entirely agnostic to the arrival epoch process, which is what enables our results to be immediately applicable to queues with time-varying and/or correlated inter-arrival times. This approach of leveraging the infinite server queue to understand the multi-server system is similar to the techniques used by Reed (2009) to extend the Halfin-Whitt heavy traffic lim-

its to non-Markovian service durations. In particular, Reed (2009) serves as a key predecessor and inspiration for our proof methodology. The infinite server model offers a natural counterpart for studying multi-server queues because it features the same arrival process with every job beginning its service immediately, meaning each job's time in system has no dependency on the current level of the process. Hence, comparisons to infinite server queues are often powerful tools for multi-server queueing analysis, for example in other works such as Eick et al. (1993); Massey and Whitt (1994); Jennings et al. (1996).

A key tool in our calculations will be the convergence of a sum of exponential functions to the indicator function $\mathbf{1}\{x \leq c\}$, as shown in Sullivan et al. (1980). This is of particular use to us in calculating the probability of a storage process exceeding a threshold, through which we drive our staffing analysis. One can note that an alternative approach to this would be to leverage asymptotic normality results, such as those in Lane (1984); Rice (1977) for shot noise processes. While this may work well in some settings, we can note that for systems that require a very rapid service rate, such as in the look-ahead service, the rate of arrivals may not be fast enough to justify a Gaussian approximation. This is of particular concern for approximating the tails of the distribution, which is at the heart of this problem. However, such an approximation could be promising in areas large enough to have a significant number of miles driven, and thus we discuss a normal-based approximation in our numerical studies in Section 6.4. Moreover, this simple approximation helps us build intuition for our general results on how batches impact service systems.

Because batches can be thought of as particularly rapid bursts of arrivals, our work aligns with recent studies of queueing models with bursty arrival

processes, such as Gao and Zhu (2018a); Koops et al. (2017, 2018); L'Ecuyer et al. (2018); Boxma et al. (2018) and as in Chapter 2. In these works, there are two main model characteristics that produce temporal clusters of arrival epochs: self-excitement and external stimuli. The classic examples of these processes are the Hawkes and Cox processes, respectively. Originally defined in Hawkes (1971), the Hawkes process (in its simplest form) has an arrival intensity that jumps upward by a fixed amount when each arrival occurs and decays exponentially towards a baseline rate between epochs. Thus, this process is said to be self-exciting as the occurrence of an event increases the likelihood that another will occur soon after. Self-excitement has often been thought of as a contagion or viral process, and this has recently been formalized in Rizoiu et al. (2018) and in Chapter 4. Similarly, the analogous Cox process also has an arrival intensity with upward jumps and exponential decay, see e.g. Daley and Vere-Jones (2003). However, the times of these jumps are not the same as the arrival epochs; they are instead given by an external Poisson process. For this reason, the Cox process can be thought of a non-stationary Poisson process with stochastic intensity driven by another, exogenous Poisson process, and thus is often referred to as a doubly-stochastic Poisson process. We can note that self-excitement and external stimuli need not be mutually exclusive, as discussed briefly in Hawkes (1971) and explored in depth in the "dynamic contagion process" introduced in Dassios and Zhao (2011). While our attention in this work will focus on batches of arrivals rather than bursts of arrivals, our results apply naturally to these settings either through the batch arrivals or the general arrival process we consider.

For an interesting concurrent work that also uses queueing theory to address problems in the management of autonomous vehicles, we refer the reader

to Mirzaeian et al. (2018). In that work, the authors compare policies for regulating the mix of autonomous vehicles and human-driven vehicles on highways, analyzing designated lane and integrated traffic policies across various levels of vehicle arrival rates to the highway and proportions of autonomous vehicles in the market. As the authors note, autonomous vehicles have the ability to form *platoons* on highways, meaning batches of cars that travel together with small gaps between them. Using this idea, the authors find that autonomous vehicles can make highways significantly more efficient. This potential for societal improvement aligns with the empirical findings in Zhang et al. (2019), which observes a decrease in traffic accidents following the introduction of autonomous vehicles.

6.1.2 Contributions and Organization

The remainder of this chapter is organized as follows. In Section 6.2, we develop general queueing models for autonomous vehicle teleoperations systems. With batch arrivals as the salient feature of these service systems, we also then establish connections from these queueing systems to storage processes, another type of stochastic process, as part of our analysis of the large batch regime. Then in Section 6.3, we leverage this connection to storage processes and develop methodology for staffing these batch arrival queueing models of the teleoperation systems. Finally in Section 6.4, we calculate necessary staffing levels using a variety of public data on driving and on autonomous vehicles, including the 2018 California disengagement reports (GM Cruise LLC (2019); Waymo LLC (2019)), the 2017 National Household Travel Survey Federal Highway Administration (2017), the 2014 and 2018 New York City Taxi Factbooks New York City

Taxi & Limousine Commission (2014, 2018), and a 2019 taxi study from the Los Angeles Department of Transportation Sam Schwartz Engineering (2019). Additionally, we also numerically investigate the effect of dependence within batches of jobs and observe how this can affect the large batch setting. Broadly speaking, the proofs contained in the main body of this chapter are in the context of using an infinite server queue to understand a multi-server queue, a recurring theme to this work. We note that we also consider a blocking model in which jobs that would have to wait are lost, and this analogous but adjacent analysis is conducted in the appendix. The appendix also contains technical lemmas and proofs for our Legendre computational techniques, as well as an exploration of a specialized batch distribution setting.

This work leads us to the following contributions:

i) Our analysis finds that even at the current level of driverless vehicle technology, teleoperations are already substantially more staffing-efficient than in car safety drivers and thus offer a desirable autonomous vehicle safety system. Moreover, the efficiency of this way humans can support level 4 driverless vehicles promotes teleoperations systems as the potential solution for helping these constrained vehicles achieve the long term goals of the potentially infeasible level 5 automation.

ii) This analysis is conducted through the study of queueing systems that receive large batches of arrivals. As a general queueing theoretic takeaway, we find that batches have a pronounced effect on the performance of the service system, and thus are important for any service manager to consider.

iii) Methodologically speaking, our large batch analysis is based around what is, to the best of our knowledge, the first batch scaling limit of a multi-

server queueing system, specifically the $G_t^{B(n)}/GI/cn$ queue. We also provide a batch scaling limit for the $G_t^{B(n)}/GI/\infty$ queue, which constitutes a significant generalization over previous batch scaling results, which only existed for the $M^{B(n)}/M/\infty$ system.

6.2 Modeling the Remote Support Center Using Queueing Theory

In the introduction, we have described possible arrangements of autonomous vehicle teleoperations systems, including the pre-fecthing look-ahead assistance method introduced in Lundgard et al. (2018). This simulation-based methodology will be the motivating scenario for the queueing model we define and analyze in this section, as its instantaneous crowd-sourcing structure offers the potential to engage many human driving experts for each autonomous vehicle. Additionally, this structure leads to a service system that differs from many models commonly studied in the literature. Because the simulation yields multiple future scenarios at each request for support, jobs arrive to the system in batches rather than in single-file fashion. Furthermore, this arrangement even differs from other batch arrival service systems in that there is a benefit to receiving batches that are large in size. As we have discussed, the crowdsourcing and simulation components within look-ahead assistance imply that large batches of jobs will produce a safer and more robust teleoperations system. From the crowdsourcing perspective, this is because taking on intentional redundancy and assigning the same task to multiple people is known to yield more accurate collective answers. From the simulation perspective, this is because simulating

more scenarios means that the vehicle is prepared for additional possible future states, further protecting the vehicle against the tail of what can occur. On top of this, simulating additional unlikely scenarios can also provide valuable driving input for training the autonomous vehicle, as the challenges observed on real roads are the edge cases that have prevented autonomous vehicle technology from progressing from level 4 to level 5. In this way, look-ahead assistance resembles a rare event simulation of sorts for autonomous vehicles, as it is generates possible future states once the vehicle has encountered the periphery of its understanding. It is also worth noting that although we focus on the simulation structure as the source of batches, batch arrivals may occur for other reasons as well, such as the bursty arrival processes or the dense platoons of autonomous vehicles that we have discussed in the introduction. In conjunction with the simulation, these sources would create *batches of batches*, again implying large batches of arrivals.

As a service system, the look-ahead teleoperations model seeks to provide timely human input for each of these future scenarios so that the vehicle is well-equipped to navigate the uncertainty. Hence, the objective of our study is to staff this system so that there are sufficiently many human experts available to inform the machine's driving execution. In our analysis this amounts to staffing a multi-server queueing system receiving large batches of arrivals. Hence, in Subsection 6.2.1 we precisely define a general delay queueing model with batch arrivals. In addition to this, we will also define a general infinite server queueing model. While this system constitutes an idealized variant of the multi-sever model in which there is always sufficiently many teleoperators, it is not necessarily meant to represent the teleoperations scenario. Rather, we will make use of the infinite server model as a tractable first step towards understanding the

finite server model. A recurring theme throughout this work, we use this comparison as tool for analyzing the multi-server queue with large batch arrivals in Subsection 6.2.2, in which we connect these queueing systems to storage processes.

6.2.1 Defining the Queueing Model

For $n \in \mathbb{Z}^+$, let $Q_t^C(n)$ be the number of jobs in the system at time $t \geq 0$ in a $G_t^{B(n)}/GI/cn$ queueing model. That is, suppose that arrivals to the queue occur according to some general point process $\{N_t \mid t \geq 0\}$ with arrival epochs $\{A_i \mid i \in \mathbb{Z}^+\}$, where the distributions of the inter-arrival times may be correlated and time-varying. A batch of jobs enters the system at each arrival epoch, with the size of the ith batch being the ith element of the sequence of positive integer random variables $\{B_i(n) \mid i \in \mathbb{Z}^+\}$, which is i.i.d. across i and independent across n. Moreover, suppose that that the mean batch size grows linearly with n, i.e. $\mathrm{E}[B_1(n)] \in O(n)$. Then, suppose that there are cn servers for some constant $c > 0$, and that any job that finds all servers busy waits, meaning that $Q_t^C(n)$ is a delay model. Because n indexes the size of the batches, we can note the constant c corresponds to ratio between the number of agents and this relative batch size. Let $\{S_{i,j} \mid (i,j) \in \mathbb{Z}^+ \times \mathbb{Z}^+\}$ be a sequence of i.i.d. positive real random variables such that $S_{i,j}$ is the service duration of the jth job within the ith batch. We let $G(\cdot)$ be the CDF of these random variables with $\bar{G}(\cdot)$ as the complementary CDF: $\bar{G}(x) = 1 - G(x)$ for all $x > 0$. We will refer to this system as the delay model, as jobs will be delayed before beginning service if there are no immediately available servers. Similarly, let us define $Q_t(n)$ as the queue length process for the infinite server analog of $Q_t^C(n)$, which is the $G_t^{B(n)}/GI/\infty$ system.

We again take N_t as the arrival process, $\{B_i(n) \mid i \in \mathbb{Z}^+\}$ as the batch sizes, and $\{S_{i,j} \mid (i,j) \in \mathbb{Z}^+ \times \mathbb{Z}^+\}$ as the service durations, where by comparison to the delay model this system has infinitely many servers. Because no jobs wait to begin service, this means that this infinite server model will give us a tractable first step towards understanding the delay model under large batch arrivals.

We can interpret these models in terms of the teleoperations systems as follows. Each arrival epoch represents a disengagement, i.e. A_i is the time of the ith disengagement. At the ith disengagement, $B_i(n)$ jobs enter the system simultaneously to be handled by the teleoperators. In the delay model there are cn teleoperators in total, so jobs may have to wait before receiving assistance from a teleoperator. By comparison, the infinite server model represents the unachievable ideal of having as many teleoperators as could possibly be needed, and no job will have to wait to be handled. This strikes to the heart of the problem we consider in this work, as there is an inherent time-sensitivity within look-ahead assistance. Hence, we seek to staff the teleoperations center so that the probability that the system exceeds its capacity achieves some low target. One can consider the probability that the number of jobs in system is above the number of teleoperators, i.e. $\mathrm{P}\left(Q_t^C(n) \geq cn\right)$, or the probability that an entering batch would cause the system to exceed its capacity, $\mathrm{P}\left(Q_t^C(n) + B_1(n) > cn\right)$. Or, as a generalization of this, one could also consider the probability

$$\mathrm{P}\left(Q_t^C(n) + pB_1(n) > cn\right),$$

for some $p \in (0, 1]$. It is worth noting that although these events are the same in the single arrivals case, in the case of batch arrivals they are not. In our analysis, we will focus on the extreme cases, using either $\mathrm{P}\left(Q_t^C(n) \geq cn\right)$ or $\mathrm{P}\left(Q_t^C(n) + B_1(n) > cn\right)$ within our staffing problem objective. The complements of each of these events can be seen to plainly represent their performance goals.

For the former this is that at least some part of the batch is able to begin service immediately, and for the latter this is that all jobs in the batch begins service immediately. Thus, for a target probability $\epsilon > 0$, our goal is to find a number of agents such that these exceedance probabilities are no more than ϵ, which is to say that the corresponding complementary events occur with probability at least $1 - \epsilon$. This amounts to finding a constant c such that the given exceedance probability is sufficiently small. Because this quantity converts the relative batch size n to the number of teleoperators cn, we will refer to c as the operator to batch size ratio. To find such a staffing level in the presence of large batch arrivals, we first need to develop theory on how these queues behave in such conditions. To do so, we now prove connections between the queueing models $Q_t^C(n)$ and $Q_t(n)$ and storage processes in Subsection 6.2.2.

6.2.2 From Queues to Storage Processes

To motivate the concept of what we will refer to as a "batch scaling" of a queueing system, let us make an informal comparison to a queue's fluid limit. Like in a fluid limit, imagine shrinking the size of each arriving entity in a queueing model. However, rather than increasing the rate that entities arrive, like in the fluid limit, suppose that instead we increase the number of entities that enter the system at each arrival epoch. In this way, we isolate the arrival epoch process. The distribution of the inter-arrival times is the same for all n, yet the distribution queue's departure process changes with n. In the limit, we find that these batch scaled queueing processes converge to storage processes, which have also been referred to as dams or even fluid queues. Informally, the idea of these continuous time processes is as follows. Much like how in this work we think of a

queue by its queue length, i.e. the number of entities present in the system at the given time, a storage process is concerned with the total "content" currently in system. By comparison to the queue this content is a non-negative real number, rather than a non-negative integer. Like a queue, the content in the storage process jumps upward at times given by some point process, but unlike the queue, the content in the storage process simply drains or releases deterministically between the jump epochs. We will refer to the manner in which the content drains as the "release rule" of the storage process, which may depend on the current content level or even on the history of the process. In this way one can see how these processes have been a natural fit in the literature for modeling dams. The jumps can represent an amount of water added to the reservoir in sudden large amounts, such as from rainfall, whereas the release from draining or evaporation is gradual and continuous.

As a preliminary, let us now introduce the terms and assumptions that we will use throughout our batch scaling analysis. At the risk of overloading notation, we will let N_t be a point process that is equivalent in distribution to the process for the queue arrival epochs (thus we do not use distinguishing notation) and we let A_i be the corresponding i^{th} arrival epoch. We suppose that there is an i.i.d. sequence of positive random variables $\{M_i \mid i \in \mathbb{Z}^+\}$ such that $\frac{B_1(n)}{n} \overset{D}{\Longrightarrow} M_1$ as $n \to \infty$, with $\frac{B_1(n)}{n^2} \xrightarrow{p} 0$. In this assumption, the batch scaling of the queue converts discrete batches of entities to continuous jumps in content, or "marks." Furthermore, we suppose that the known initial value of the infinite server queue converges to an analogous initial value in the limit, i.e. $\frac{Q_0(n)}{n} \to \psi_0$.

Using these definitions, we now prove our batch scaling results. To first gain understanding in a tractable setting, we start with the infinite server model.

That is, for $Q_t(n)$ we will now show the convergence of $G_t^{B_1(n)}/GI/\infty$ queues to general shot noise processes, which can be viewed as infinite capacity storage, or dam, processes. Let the shot noise process ψ_t at time $t \geq 0$ be defined as

$$\psi_t = \psi_0 \bar{G}_0(t) + \sum_{i=1}^{N_t} M_i \bar{G}(t - A_i), \tag{6.1}$$

which will jump upward according to the sequence $\{M_i \mid i \in \mathbb{Z}^+\}$ and then decay downward according to the complementary CDF $\bar{G}(\cdot) = 1 - G(\cdot)$. This can be thought of as an infinite capacity storage process, as there is no bound on the amount of content that can enter the system at once. Furthermore, as in an infinite server queue, the manner in which the content brought by one arrival departs has no dependance on any of the arrivals before it or on the amount of the content currently in the system. Following this intuition, we now formalize the connections between the shot noise process and the general infinite server queue in Theorem 6.2.1.

Theorem 6.2.1. *As $n \to \infty$, the batch scaling of the $G_t^{B(n)}/GI/\infty$ queue $Q_t(n)$ yields*

$$\frac{Q_t(n)}{n} \overset{D}{\Longrightarrow} \psi_t, \tag{6.2}$$

pointwise in $t \geq 0$, where ψ_t is a shot noise process as defined in Equation 6.1, i.e. an infinite capacity storage process. If N_t is a stationary Poisson process with rate $\lambda > 0$, this implies that the moment generating function of $\frac{Q_t(n)}{n}$ converges to

$$\mathrm{E}\left[e^{\frac{\theta}{n} Q_t(n)}\right] \longrightarrow e^{\theta \psi_0 \bar{G}_0(t) + \lambda \int_0^t \left(\mathrm{E}\left[e^{\theta M_1 \bar{G}(x)}\right] - 1\right) \mathrm{d}x}, \tag{6.3}$$

as $n \to \infty$.

Proof. We will show the convergence of the batch scaling of the queue through analyzing its moment generating function. To begin, we note that the infinite

server queue length can be expressed in terms of indicator functions as

$$Q_t(n) = \sum_{j=1}^{Q_0(n)} \mathbf{1}\{t < S_{0,j}\} + \sum_{i=1}^{N_t} \sum_{j=1}^{B_i(n)} \mathbf{1}\{t < A_i + S_{i,j}\},$$

where $S_{i,j}$ is the service duration of the jth customer within the ith batch and $S_{0,j}$ is the remaining service time of the jth job that was in service at time 0. In this way, the first term on the right hand side represents the number of jobs in the system at time 0 that remain in the system at time t, whereas the double summation counts the number of jobs from each batch that remain in service at time t. Because there are infinitely many servers, we can note that the number jobs remaining since time 0 is independent from the number of jobs in the system that entered after time 0. Hence, we will consider these groups separately. Starting with those jobs initially present, we can note that since $\{S_{0,j} \mid 1 \leq j \leq Q_0(n)\}$ are the only stochastic terms, the law of large numbers yields that

$$\frac{1}{n} \sum_{j=1}^{Q_0(n)} \mathbf{1}\{t < S_{0,j}\} = \frac{Q_0(n)}{n} \frac{1}{Q_0(n)} \sum_{j=1}^{Q_0(n)} \mathbf{1}\{t < S_{0,j}\} \xrightarrow{a.s.} \psi_0 \bar{G}_0(t)$$

Thus, without loss of generality, we will hereforward assume that the queue starts empty. We then write the moment generating function of $Q_t(n)$ at $\frac{\theta}{n}$ as

$$\mathrm{E}\left[e^{\frac{\theta Q_t(n)}{n}}\right] = \mathrm{E}\left[\exp\left(\frac{\theta}{n} \sum_{i=1}^{N_t} \sum_{j=1}^{B_i(n)} \mathbf{1}\{t < A_i + S_{i,j}\}\right)\right].$$

By conditioning on the filtration of the counting process $\mathcal{F}_t^N$, total expectation yields that

$$\mathrm{E}\left[\exp\left(\frac{\theta}{n} \sum_{i=1}^{N_t} \sum_{j=1}^{B_i(n)} \mathbf{1}\{t < A_i + S_{i,j}\}\right)\right] = \mathrm{E}\left[\prod_{i=1}^{N_t} \mathrm{E}\left[\exp\left(\frac{\theta}{n} \sum_{j=1}^{B_i(n)} \mathbf{1}\{t < A_i + S_{i,j}\}\right) \middle| \mathcal{F}_t^N\right]\right].$$

Focusing on the inner expectation, we again use the tower property. We now condition on the batch size $B_i(n)$, which leaves the service duration as the only uncertain quantity. The indicator is thus a Bernoulli random variable with success probability $\bar{G}(t - A_i)$, and since these are i.i.d. within the batch we have

that

$$\begin{aligned}
\mathrm{E}\left[\exp\left(\frac{\theta}{n}\sum_{j=1}^{B_i(n)}\mathbf{1}\{t < A_i + S_{i,j}\}\right)\middle|\mathcal{F}_t^N\right] &= \mathrm{E}\left[\prod_{j=1}^{B_i(n)}\mathrm{E}\left[e^{\frac{\theta}{n}\mathbf{1}\{t<A_i+S_{i,j}\}}\middle|\mathcal{F}_t^N, B_i(n)\right]\middle|\mathcal{F}_t^N\right] \\
&= \mathrm{E}\left[\left(G(t-A_i) + \bar{G}(t-A_i)e^{\frac{\theta}{n}}\right)^{B_i(n)}\middle|\mathcal{F}_t^N\right] \\
&= \mathrm{E}\left[\left(1 + \bar{G}(t-A_i)(e^{\frac{\theta}{n}} - 1)\right)^{B_i(n)}\middle|\mathcal{F}_t^N\right].
\end{aligned}$$

By now using the identity $x = e^{\log(x)}$, we can transform this to

$$\begin{aligned}
\mathrm{E}\left[\left(1 + \bar{G}(t-A_i)(e^{\frac{\theta}{n}} - 1)\right)^{B_i(n)}\middle|\mathcal{F}_t^N\right] &= \mathrm{E}\left[\exp\left(\log\left(\left(1 + \bar{G}(t-A_i)(e^{\frac{\theta}{n}} - 1)\right)^{B_i(n)}\right)\right)\middle|\mathcal{F}_t^N\right] \\
&= \mathrm{E}\left[e^{B_i(n)\log\left(1+\bar{G}(t-A_i)(e^{\frac{\theta}{n}}-1)\right)}\middle|\mathcal{F}_t^N\right],
\end{aligned}$$

which we can now re-express further through two series expansions. Specifically, using a Taylor and a Mercator series expansion on $e^{\frac{\theta}{n}} - 1$ and $\log\left(1 + \bar{G}(t-A_i)(e^{\frac{\theta}{n}} - 1)\right)$, respectively, we simplify to

$$\mathrm{E}\left[e^{B_i(n)\log\left(1+\bar{G}(t-A_i)(e^{\frac{\theta}{n}}-1)\right)}\middle|\mathcal{F}_t^N\right] = \mathrm{E}\left[e^{\frac{\theta B_i(n)\bar{G}(t-A_i)}{n} + O\left(\frac{B_i(n)}{n^2}\right)}\middle|\mathcal{F}_t^N\right].$$

Returning to the original expectation, we now have that

$$\mathrm{E}\left[\prod_{i=1}^{N_t}\mathrm{E}\left[\exp\left(\frac{\theta}{n}\sum_{j=1}^{B_i(n)}\mathbf{1}\{t < A_i + S_{i,j}\}\right)\middle|\mathcal{F}_t^N\right]\right] = \mathrm{E}\left[e^{\sum_{i=1}^{N_t}\frac{\theta B_i(n)\bar{G}(t-A_i)}{n} + O\left(\frac{B_i(n)}{n^2}\right)}\right],$$

and as $n \to \infty$, this converges to

$$\mathrm{E}\left[e^{\sum_{i=1}^{N_t}\frac{\theta B_i(n)\bar{G}(t-A_i)}{n} + O\left(\frac{B_i(n)}{n^2}\right)}\right] \longrightarrow \mathrm{E}\left[e^{\theta\sum_{i=1}^{N_t} M_i\bar{G}(t-A_i)}\right],$$

which yields the stated result for the queue. To now yield the specific form of the generating function when N_t is a Poisson process, we note that when conditioned on the quantity N_t we have

$$\mathrm{E}\left[e^{\sum_{i=1}^{N_t}\theta M_i\bar{G}(t-A_i)}\right] = \mathrm{E}\left[\mathrm{E}\left[e^{\sum_{i=1}^{N_t}\theta M_i\bar{G}(t-A_i)} \mid N_t\right]\right] = \mathrm{E}\left[\mathrm{E}\left[e^{\theta M_1\bar{G}(U_1(0,t))}\right]^{N_t}\right],$$

where $U_i(0,t) \sim \text{Uni}(0,t)$ are i.i.d and independent of M_i. Then, conditioning on M_1, this inner expectation can be expressed

$$\mathrm{E}\left[e^{\theta M_1 \bar{G}(U_1(0,t))}\right] = \mathrm{E}\left[\mathrm{E}\left[e^{\theta M_1 \bar{G}(U_1(0,t))} \mid M_1\right]\right] = \mathrm{E}\left[\frac{1}{t}\int_0^t e^{\theta M_1 \bar{G}(x)}\mathrm{d}x\right].$$

By exchanging the order of integration and expectation via Fubini's theorem and substituting into the moment generating function for the Poisson process, we achieve the corresponding stated form. □

For a visual example of this convergence, in Figure 6.5 we plot the empirical distributions of four infinite server queues with different batch sizes and compare them to the simulated distribution of the limiting shot noise process. In this scenario the batches are Poisson distributed with rate n. Through the scaling, this produces deterministic jumps of size 1 in the storage process. As one can observe, as the batch size increase the queue's cumulative distribution function becomes increasingly similar to the cumulative distribution function for the shot noise process.

Following the convergence shown in Theorem 6.2.1, we have noted that the infinite server queue and the limiting shot noise process share a key similarity. In both models, the manner that new arrivals leave the system has no dependence on the rest of the process. In the infinite server queue, each job immediately enters service without waiting regardless of the number of jobs already in the system. In the shot noise process, a new jump immediately begins to drain, and the manner in which this content is released is determined only by the time that has elapsed since its arrival. Of course, this is not true for the delay queueing model $Q_t^C(n)$. Some of the jobs within an arriving batch may have to wait if there are not sufficiently many available servers. Moreover, the time that a job has to wait is dependent on the status of the jobs that have entered service

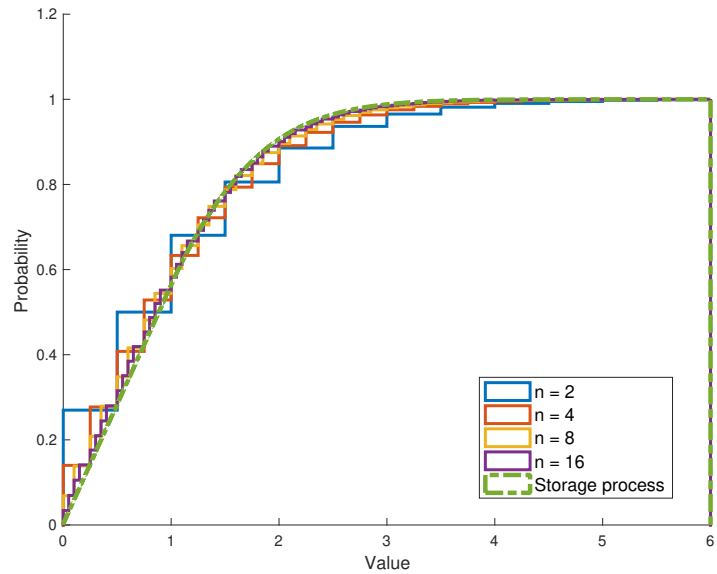


Figure 6.5: Simulated demonstration of convergence in distribution of an $M^{B(n)}/M/\infty$ queue to a shot noise process, based on 100,000 replications with $t = 10$, $\lambda = 1$, $\mu = 1$, and $B_1(n) \sim \text{Pois}(n)$.

ahead of it. If a job does not have to wait, its movement within the system is the same as if there were infinitely many servers, but if it does have to wait this will not be the case. Thus, we can categorize jobs as either having entered service immediately or having needed to wait. This observation will guide the proof of our multi-server batch scaling limit in Theorem 6.2.2, and it will also now serve as inspiration for our definition of the limiting storage process.

Using the same arrival process definitions used in the shot noise process, let us now define the generalized c-threshold storage process ψ_t^C at time $t \geq 0$ as

$$\psi_t^C = \left(\psi_0^C \wedge c\right)\bar{G}_0(t) + \left(\psi_0^C - c\right)^+ \bar{G}(t) + \sum_{i=1}^{N_t} M_i \bar{G}(t - A_i) + \int_0^t \left(\psi_{t-s}^C - c\right)^+ \mathrm{d}G(s), \quad (6.4)$$

where we let ψ_0^C be the limit of the normalized initial value of the delay model queue length, meaning $\frac{Q_0^C(n)}{n} \to \psi_0^C$. In comparing the definitions of this process and the shot noise process given in Equation 6.1, we can note that there

are similar terms between the two. In particular, each process has terms for the decay of the initial value and of the marks according to the tail CDF of the service distribution. However, Equation 6.4 also includes terms that adjust these values, notably the integral over the history of the process's exceedance of the threshold level c, from which this process gets its name. As we now prove in Theorem 6.2.2, this correction integral arises in the batch scaling of the delay model as the limiting adjustment for the content that has to wait to drain. Just as this construction is cast in comparison to the shot noise process, this proof is centered around the relationship between the infinite server queue and the multi-server delay model.

Theorem 6.2.2. *As $n \to \infty$, the batch scaling of the $G_t^{B(n)}/GI/cn$ queue $Q_t^C(n)$ yields*

$$\frac{Q_t^C(n)}{n} \overset{D}{\Longrightarrow} \psi_t^C, \tag{6.5}$$

pointwise in $t \geq 0$, where ψ_t^C is a generalized threshold storage process as defined in Equation 6.4.

Proof. In a manner similar to the proof of the infinite server to shot noise convergence in Theorem 6.2.1, we begin by decomposing the queue length process into a sum of indicators. By comparison to the infinite server decomposition however, these indicators depend not only on the batch arrival epochs and the individual service durations, but also on the lengths of time that jobs wait to begin service while the servers were occupied. Defining $W_{i,j}$ as the total time the jth job within the ith batch spends waiting, we can express the queue length

in the delay model queue at time t as

$$Q_t^C(n) = \sum_{i=1}^{N_t} \sum_{j=1}^{B_i(n)} \mathbf{1}\{t < A_i + S_{i,j}\} + \sum_{i=1}^{N_t} \sum_{j=1}^{B_i(n)} \mathbf{1}\{A_i + S_{i,j} \leq t < A_i + S_{i,j} + W_{i,j}\}$$
$$+ \sum_{j=1}^{(Q_0^C(n) \wedge cn)} \mathbf{1}\{t < S_{0,j}\} + \sum_{j=1}^{(Q_0^C(n) - cn)^+} \mathbf{1}\{t < W_{\cdot,j} + S_{\cdot,j}\}. \tag{6.6}$$

One can interpret this decompositions as follows. The first double summation across arrival epochs and batch sizes gives an idealized infinite server representation that would be accurate if no jobs had to wait to begin service. The second double summation then corrects that under-counting for any jobs that had to wait and have not yet completed service at time t. The third and fourth terms then capture the initial state of the system, with the third term counting which jobs have remained in service from time 0 to time t and with the fourth term counting the number of jobs that were waiting at time 0 and have not completed service by t. Here we use $S_{0,j}$ to represent the remaining service times of the jobs that are in service at time 0 and we use $W_{\cdot,j}$ and $S_{\cdot,j}$ to represent the waiting and service times for the jobs that are present in the system at time 0 but were not in service. In this notation, the residual service time $S_{0,j}$ need not be equivalent in distribution to $S_{i,j}$ for $i \in \mathbb{Z}^+$, whereas $S_{\cdot,j}$ is equivalent to $S_{i,j}$.

To begin moving towards the storage process limit, we first show a batch-arrival-queue analog of Proposition 2.1 from Reed (2009). That is, we seek to justify

$$\int_0^t \left(Q_{t-s}^C - cn\right)^+ \mathrm{d}G(s) = \sum_{i=1}^{N_t} \sum_{j=1}^{B_i(n)} \left(\bar{G}(t - A_i - W_{i,j}) - \bar{G}(t - A_i)\right)$$
$$+ \sum_{j=1}^{(Q_0^C(n) - cn)^+} \left(\bar{G}(t - W_{\cdot,j}) - \bar{G}(t)\right), \tag{6.7}$$

and this follows from a generalization of the arguments from Reed (2009). Starting with the summations over the tail CDF terms, one can re-express these in

terms of integrals over the service distribution measure, and these integrals can then be adjusted to a standard interval of $[0, t]$ through the introduction of indicator functions:

$$\begin{aligned}
&\sum_{i=1}^{N_t}\sum_{j=1}^{B_i(n)}\left(\bar{G}(t-A_i-W_{i,j})-\bar{G}(t-A_i)\right)+\sum_{j=1}^{(Q_0^C(n)-cn)^+}\left(\bar{G}(t-W_{\cdot,j})-\bar{G}(t)\right)\\
&=\sum_{i=1}^{N_t}\sum_{j=1}^{B_i(n)}\int_{(t-A_i-W_{i,j})^+}^{t-A_i}\mathrm{d}G(s)+\sum_{j=1}^{(Q_0^C(n)-cn)^+}\int_{(t-W_{\cdot,j})^+}^{t}\mathrm{d}G(s)\\
&=\sum_{i=1}^{N_t}\sum_{j=1}^{B_i(n)}\int_0^t\mathbf{1}\{A_i\le t-s<A_i+W_{i,j}\}\mathrm{d}G(s)+\sum_{j=1}^{(Q_0^C(n)-cn)^+}\int_0^t\mathbf{1}\{t-s<W_{i,j}\}\mathrm{d}G(s).
\end{aligned}$$

Then, one can recognize that the number of jobs waiting at an arbitrary time $u \geq 0$ can be written

$$\left(Q_u^C(n)-cn\right)^+=\sum_{j=1}^{N_u}\sum_{j=1}^{B_i(n)}\mathbf{1}\{A_i\le u<A_i+W_{i,j}\}+\sum_{j=1}^{(Q_0^C(n)-cn)^+}\mathbf{1}\{u<W_{\cdot,j}\},$$

as the first term on the right-hand side captures the number jobs still waiting across each batch of arrivals and the second term captures the number of jobs that have been waiting since time 0. Thus, by exchanging the order of summation and integration, we can now observe that

$$\begin{aligned}
&\sum_{i=1}^{N_t}\sum_{j=1}^{B_i(n)}\int_0^t\mathbf{1}\{A_i\le t-s<A_i+W_{i,j}\}\mathrm{d}G(s)+\sum_{j=1}^{(Q_0^C(n)-cn)^+}\int_0^t\mathbf{1}\{t-s<W_{i,j}\}\mathrm{d}G(s)\\
&=\int_0^t\left(\sum_{i=1}^{N_t}\sum_{j=1}^{B_i(n)}\mathbf{1}\{A_i\le t-s<A_i+W_{i,j}\}+\sum_{j=1}^{(Q_0^C(n)-cn)^+}\mathbf{1}\{t-s<W_{i,j}\}\right)\mathrm{d}G(s)\\
&=\int_0^t\left(Q_{t-s}^C(n)-cn\right)^+\mathrm{d}G(s),
\end{aligned}$$

and thus we achieve Equation 6.7.

Returning now to the decomposition of the queue length in Equation 6.6, we

can use the equality from Equation 6.7 to re-express the queue length as

$$
\begin{aligned}
Q_t^C(n) &= \sum_{i=1}^{N_t}\sum_{j=1}^{B_i(n)} \mathbf{1}\{t < A_i + S_{i,j}\} + \sum_{i=1}^{N_t}\sum_{j=1}^{B_i(n)} \mathbf{1}\{A_i + S_{i,j} \le t < A_i + S_{i,j} + W_{i,j}\} \\
&\quad - \sum_{i=1}^{N_t}\sum_{j=1}^{B_i(n)} \left(\bar{G}(t - A_i - W_{i,j}) - \bar{G}(t - A_i)\right) - \sum_{j=1}^{(Q_0^C(n)-cn)^+} \left(\bar{G}(t - W_{\cdot,j}) - \bar{G}(t)\right) \\
&\quad + \int_0^t \left(Q_{t-s}^C(n) - cn\right)^+ \mathrm{d}G(s) + \sum_{j=1}^{(Q_0^C(n)\wedge cn)} \mathbf{1}\{t < S_{0,j}\} + \sum_{j=1}^{(Q_0^C(n)-cn)^+} \mathbf{1}\{t < W_{\cdot,j} + S_{\cdot,j}\} \\
&= \sum_{i=1}^{N_t}\sum_{j=1}^{B_i(n)} \mathbf{1}\{t < A_i + S_{i,j}\} + \sum_{j=1}^{(Q_0^C(n)\wedge cn)} \mathbf{1}\{t < S_{0,j}\} + \int_0^t \left(Q_{t-s}^C(n) - cn\right)^+ \mathrm{d}G(s) \\
&\quad + \left(Q_0^C(n) - cn\right)^+ \bar{G}(t) + \sum_{i=1}^{N_t}\sum_{j=1}^{B_i(n)} \left(\mathbf{1}\{t - A_i - W_{i,j} < S_{i,j}\} - \bar{G}(t - A_i - W_{i,j})\right) \\
&\quad - \sum_{i=1}^{N_t}\sum_{j=1}^{B_i(n)} \left(\mathbf{1}\{t - A_i < S_{i,j}\} - \bar{G}(t - A_i)\right) + \sum_{j=1}^{(Q_0^C(n)-cn)^+} \left(\mathbf{1}\{t - W_{\cdot,j} < S_{\cdot,j}\} - \bar{G}(t - W_{\cdot,j})\right).
\end{aligned}
$$

Through this decomposition, we will now prove that the batch scaling of the queue length converges to the generalized storage process. We proceed through induction on the arrival times. For the base case, let $0 \le t < A_1$. Then, the normalized queue length at time t can be written

$$
\begin{aligned}
\frac{Q_t^C(n)}{n} &= \frac{1}{n}\sum_{j=1}^{(Q_0^C(n)\wedge cn)} \mathbf{1}\{t < S_{0,j}\} + \frac{1}{n}\int_0^t \left(Q_{t-s}^C(n) - cn\right)^+ \mathrm{d}G(s) + \frac{1}{n}\left(Q_0^C(n) - cn\right)^+ \bar{G}(t) \\
&\quad + \frac{1}{n}\sum_{j=1}^{(Q_0^C(n)-cn)^+} \left(\mathbf{1}\{t - W_{\cdot,j} < S_{\cdot,j}\} - \bar{G}(t - W_{\cdot,j})\right),
\end{aligned}
$$

which we now analyze piece by piece. By the law of large numbers, the assumption on the initial values, and the continuous mapping theorem, we have that as $n \to \infty$

$$
\frac{1}{n}\sum_{j=1}^{(Q_0^C(n)\wedge cn)} \mathbf{1}\{t < S_{0,j}\} \xRightarrow{D} \left(\psi_0^C \wedge c\right)\bar{G}_0(t),
$$

where $\bar{G}_0(\cdot)$ is the complementary CDF of the residual service durations of the

initial jobs in service at time 0. We can also similarly observe that

$$\frac{1}{n}\left(Q_0^C(n) - cn\right)^+ \bar{G}(t) \stackrel{D}{\Longrightarrow} \left(\psi_0^C - c\right)^+ \bar{G}(t),$$

as $n \to \infty$. Now, for the summation over jobs that were waiting to begin service at time 0, we can employ a martingale argument such as that used in e.g. Andrews (1988). Let $\mathcal{S}_j$ for $0 \leq j \leq \left(Q_0^C(n) - cn\right)^+$ be the filtration generated by the collection of service times of the jobs initially in service at time 0 and of the first j jobs to enter service after time 0, i.e. $\mathcal{S}_j = \sigma\left(\{S_{0,1}, \ldots, S_{0,cn}, S_{\cdot,1}, \ldots, S_{\cdot,j}\}\right)$. Then, one can note that for $j < \left(Q_0^C(n) - cn\right)^+$, $W_{\cdot,j+1}$ is $\mathcal{S}_j$ measurable, as the previous service durations dictate the time that this job waits. Thus, we can recognize that $\mathrm{E}\left[\mathbf{1}\{t - W_{\cdot,j} < S_{\cdot,j}\} \mid \mathcal{S}_j\right] = \bar{G}(t - W_{\cdot,j})$. This implies that the summation is a martingale difference sequence, and thus we have that

$$\frac{1}{n}\sum_{j=1}^{(Q_0^C(n)-cn)^+} \left(\mathbf{1}\{t - W_{\cdot,j} < S_{\cdot,j}\} - \bar{G}(t - W_{\cdot,j})\right) \stackrel{D}{\Longrightarrow} 0,$$

as $n \to \infty$. Thus, as $n \to \infty$ the queue length on $0 \leq t < A_1$ converges to a process $z(\cdot)$ satisfying

$$z(t) = \left(\psi_0^C \wedge c\right)\bar{G}_0(t) + \left(\psi_0^C - c\right)^+ \bar{G}(t) + \int_0^t (z(t-s) - c)^+ \, \mathrm{d}G(s).$$

We can observe that on this time interval each of these terms are deterministic, and thus Proposition 3.1 of Reed (2009) yields that the function $z(\cdot)$ that solves this equation is unique. Since this matches the expression for ψ_t^C on $0 \leq t < A_i$ as given by Equation 6.4, we have that $\frac{Q_t^C(n)}{n} \stackrel{D}{\Longrightarrow} \psi_t^C$ as $n \to \infty$ for $0 \leq t < A_1$. At the precise epoch of the first arrival, we can note that we furthermore have the convergence of the process immediately after the batch of jobs arrives, which is a direct consequence of the preceding arguments and assumption that $\frac{B_1(n)}{n} \stackrel{D}{\Longrightarrow} M_1$ as $n \to \infty$. Thus, $\frac{Q_t^C(n)}{n} \stackrel{D}{\Longrightarrow} \psi_t^C$ as $n \to \infty$ for $0 \leq t \leq A_1$, satisfying the base case of our inductive argument.

For the inductive step, we now assume that $\frac{Q_s^C(n)}{n} \overset{D}{\Longrightarrow} \psi_s^C$ as $n \to \infty$ for s such that $0 \leq s \leq A_i$ and some $i \in \mathbb{Z}^+$. Let us now take t such that $A_i \leq t < A_{i+1}$. We have established that we can decompose the normalized queue length as

$$\begin{aligned}\frac{Q_t^C(n)}{n} &= \frac{1}{n}\sum_{i=1}^{N_t}\sum_{j=1}^{B_i(n)} \mathbf{1}\{t < A_i + S_{i,j}\} + \frac{1}{n}\sum_{j=1}^{(Q_0^C(n)\wedge cn)} \mathbf{1}\{t < S_{0,j}\} + \frac{1}{n}\int_0^t \left(Q_{t-s}^C(n) - cn\right)^+ \mathrm{d}G(s) \\ &\quad + \frac{1}{n}\left(Q_0^C(n) - cn\right)^+ \bar{G}(t) + \frac{1}{n}\sum_{i=1}^{N_t}\sum_{j=1}^{B_i(n)} \left(\mathbf{1}\{t - A_i - W_{i,j} < S_{i,j}\} - \bar{G}(t - A_i - W_{i,j})\right) \\ &\quad - \frac{1}{n}\sum_{i=1}^{N_t}\sum_{j=1}^{B_i(n)} \left(\mathbf{1}\{t - A_i < S_{i,j}\} - \bar{G}(t - A_i)\right) + \frac{1}{n}\sum_{j=1}^{(Q_0^C(n)-cn)^+} \left(\mathbf{1}\{t - W_{\cdot,j} < S_{\cdot,j}\} - \bar{G}(t - W_{\cdot,j})\right),\end{aligned}$$

and we can again analyze this piece-by-piece. By the batch scaling convergence of infinite server queues to shot noise processes in Theorem 6.2.1, we can observe that as $n \to \infty$

$$\frac{1}{n}\sum_{i=1}^{N_t}\sum_{j=1}^{B_i(n)} \mathbf{1}\{t < A_i + S_{i,j}\} \overset{D}{\Longrightarrow} \sum_{i=1}^{N_t} M_i \bar{G}(t - A_i),$$

and

$$\frac{1}{n}\sum_{i=1}^{N_t}\sum_{j=1}^{B_i(n)} \left(\mathbf{1}\{t - A_i < S_{i,j}\} - \bar{G}(t - A_i)\right) \overset{D}{\Longrightarrow} 0.$$

Similarly, analogous arguments to the base case show that the initial condition terms are such that

$$\frac{1}{n}\sum_{j=1}^{(Q_0^C(n)\wedge cn)} \mathbf{1}\{t < S_{0,j}\} \overset{D}{\Longrightarrow} \left(\psi_0^C \wedge c\right)\bar{G}_0(t),$$

$$\frac{1}{n}\left(Q_0^C(n) - cn\right)^+ \bar{G}(t) \overset{D}{\Longrightarrow} \left(\psi_0^C - c\right)^+ \bar{G}(t),$$

and

$$\frac{1}{n}\sum_{j=1}^{(Q_0^C(n)-cn)^+} \left(\mathbf{1}\{t - W_{\cdot,j} < S_{\cdot,j}\} - \bar{G}(t - W_{\cdot,j})\right) \overset{D}{\Longrightarrow} 0.$$

For the remaining double summation over arrival epochs and batch sizes, we can again make use of a martingale structure. For $i \in \mathbb{Z}^+$ and $j \in \mathbb{Z}^+$, let us define

the sigma algebra generated by the arrival times and service times of all jobs up to and including to the jth job within the ith batch, which is

$$\mathcal{S}_{i,j} = \sigma\Bigg(\{S_{0,1}, \ldots, S_{0,cn}, S_{\cdot,1}, \ldots, S_{\cdot,(cn-Q_0^C(n))^+}\} \cup \{A_1, \ldots, A_i\} \\ \cup \left(\bigcup_{k=1}^{i-1} \{S_{k,1}, \ldots, S_{k,B_k(n)}\}\right) \cup \{S_{i,1}, \cdots, S_{i,j}\}\Bigg).$$

Then, we have that $W_{i,j+1}$ is $\mathcal{S}_{i,j}$ measurable since the queue is operating under first-come-first-serve, meaning that only the previous jobs determine how long the jth job in the ith batch waits. Thus, $\mathrm{E}\left[\mathbf{1}\{t - A_i - W_{i,j+1} < S_{i,j+1}\} \mid \mathcal{S}_{i,j}\right] = \bar{G}(t - A_i - W_{i,j})$. Therefore through martingale differences we have that

$$\frac{1}{n}\sum_{i=1}^{N_t}\sum_{j=1}^{B_i(n)}\left(\mathbf{1}\{t - A_i - W_{i,j} < S_{i,j}\} - \bar{G}(t - A_i - W_{i,j})\right) \overset{D}{\Longrightarrow} 0,$$

as $n \to \infty$. Bringing these pieces together we now have that for $t \in [A_i, A_{i+1})$ the queue length process converges to a process $z_i(\cdot)$ satisfying

$$z_i(t) = \left(\psi_0^C \wedge c\right)\bar{G}_0(t) + \left(\psi_0^C - c\right)^+ \bar{G}(t) + \sum_{i=1}^{N_t} M_i \bar{G}(t - A_i) + \int_0^t (z_i(t-s) - c)^+ \,\mathrm{d}G(s).$$

From the inductive hypothesis and the uniqueness given by Proposition 3.1 of Reed (2009), we have that $z_i(s) = \psi_s^C$ must hold for all $s \leq A_i$. One can then observe that $z_i(t)$ is deterministic for $A_i < t < A_{i+1}$, meaning that Proposition 3.1 of Reed (2009) further implies that $z_i(t) = \psi_t^C$ on this interval as well. To complete the inductive step, we can note that by the given convergence of the batch sizes to the jump sizes, we also have that $Q^n_{A_{i+1}}(n)/n \overset{D}{\Longrightarrow} \psi^C_{A_{i+1}}$ as $n \to \infty$, and this completes the proof. □

To provide some intuition about this stochastic process, let us consider the case of exponentially distributed service. For the sake of this example, suppose that $\bar{G}(x) = e^{-\mu x}$ for some $\mu > 0$. Then, Equation 6.4 yields that

$$\psi_t^C = \psi_0^C e^{-\mu t} + \sum_{i=1}^{N_t} M_i e^{-\mu(t-A_i)} + \int_0^t \left(\psi_{t-s}^C - c\right)^+ \mu e^{-\mu s} \mathrm{d}s.$$

Multiplying and dividing by $e^{-\mu t}$ inside the integral, we can re-express this as

$$\psi_t^C = \psi_0^C e^{-\mu t} + \sum_{i=1}^{N_t} M_i e^{-\mu(t-A_i)} + e^{-\mu t} \int_0^t \left(\psi_{t-s}^C - c\right)^+ \mu e^{\mu(t-s)} \mathrm{d}s,$$

and by changing the variable of integration to be s instead of $t - s$, we furthermore have

$$\psi_t^C = \psi_0^C e^{-\mu t} + \sum_{i=1}^{N_t} M_i e^{-\mu(t-A_i)} + e^{-\mu t} \int_0^t \left(\psi_s^C - c\right)^+ \mu e^{\mu s} \mathrm{d}s.$$

Since we know that the process jumps by M_i at the ith arrival, let us take $t \in (A_i, A_{i+1})$ and focus on the behavior between jumps. Because storage processes are deterministic on inter-jump intervals, we can take the derivative with respect to time and observe that for $t \in (A_i, A_{i+1})$,

$$\begin{aligned}
\frac{\mathrm{d}\psi_t^C}{\mathrm{d}t} &= -\mu\psi_0^C e^{-\mu t} - \mu \sum_{i=1}^{N_t} M_i e^{-\mu(t-A_i)} - \mu e^{-\mu t} \int_0^t \left(\psi_s^C - c\right)^+ \mu e^{\mu s} \mathrm{d}s + \mu \left(\psi_t^C - c\right)^+ \\
&= -\mu\psi_t^C + \mu \left(\psi_t^C - c\right)^+ \\
&= -\mu \left(\psi_t^C \wedge c\right).
\end{aligned}$$

Hence, in the case of exponential service the inter-jump dynamics of this process can be easily summarized. If ψ_t^C is above the threshold c it drains linearly, if it is below c it decays exponentially. This corresponds to a storage process with threshold release rule $r(x) = \mu(x \wedge c)$, which further motivates why we refer to the general limiting object as the generalized c-threshold storage process.

Just as we gave a visual example of the convergence of infinite server queues to shot noise processes in Figure 6.5, we now plot a series of simulated delay model distributions in Figure 6.6 and compare them to a c-threshold storage process. For this example we suppose that the batch sizes are geometrically distributed with probability of success $\frac{1}{n}$, and this yields jumps that are exponentially distributed with unit rate in the batch scaling limit. As an additional

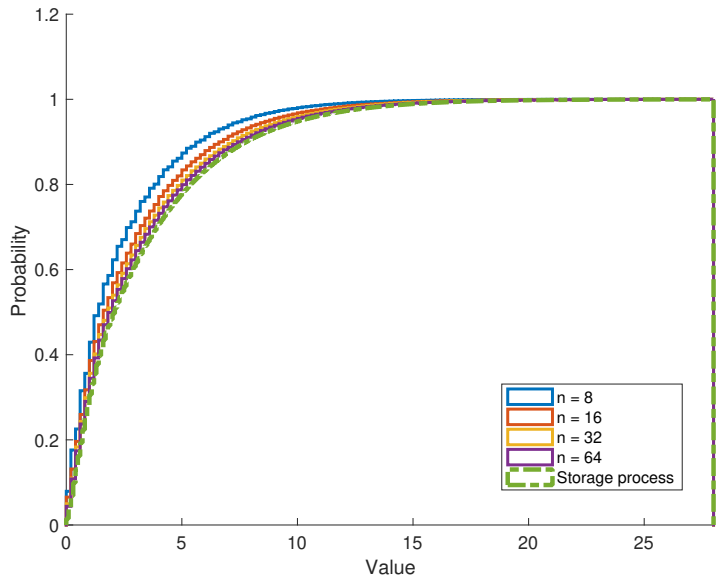


Figure 6.6: Simulated demonstration of convergence in distribution of an $M^{B(n)}/M/cn$ queue to a c-threshold storage process, based on 100,000 replications with $t = 10$, $\lambda = 3$, $\mu = 2$, $c = 2$, and $B_1(n) \sim \text{Geo}\left(\frac{1}{n}\right)$.

example of the limiting threshold dynamics, in Figure 6.7 we plot a simulated scaled queue length sample path along with the calculated storage process values when given the same arrival epochs. One can observe the change in release behavior as the process crosses the capacity level c. Above the level c, the content drains linearly and below c it decays exponentially towards zero.

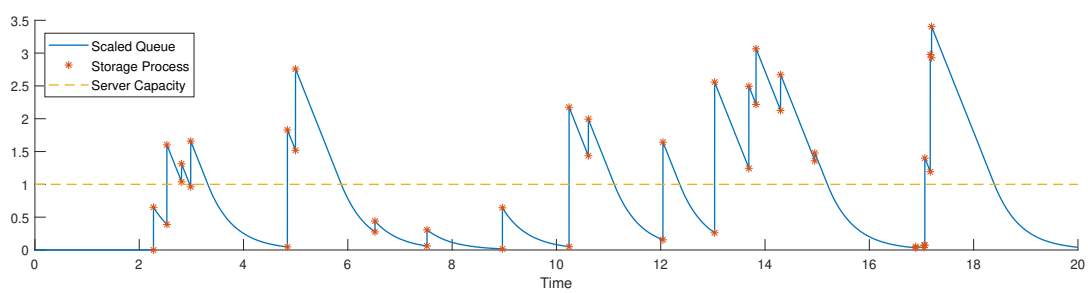


Figure 6.7: A comparison of the simulated scaled queue length process and the calculated storage process sample paths defined on the same arrival process.

6.3 Staffing the Teleoperations System

Now that we have developed an understanding of queues with large batch arrivals through connections to storage processes, in this section we will leverage this insight and use the storage process models to staff the teleoperations system. In Section 6.2, we defined the staffing problem as finding a operator to batch size ratio c such that the desired exceedance probability, $\mathrm{P}\left(Q_t^C(n) \geq cn\right)$ or $\mathrm{P}\left(Q_t^C(n) + B_i(n) > cn\right)$, is smaller than some target $\epsilon > 0$. By normalizing these events by n, we can see that the batch scaling limit in Theorem 6.2.2 yields that

$$\mathrm{P}\left(Q_t^C(n) \geq cn\right) \to \mathrm{P}\left(\psi_t^C > c\right) \quad \text{and} \quad \mathrm{P}\left(Q_t^C(n) + B_i(n) \geq cn\right) \to \mathrm{P}\left(\psi_t^C + M_i > c\right),$$

as $n \to \infty$, allowing us to "staff" the storage process instead. One general approach to this problem would be to take a simulation-based approach, such as the well-known iterative staffing algorithm introduced in Feldman et al. (2008). It is thus worth noting that the results of Theorems 6.2.1 and 6.2.2 have an immediate consequence of greatly simplifying the simulation of batch arrival queueing systems. For large batch sizes, one can simply simulate a storage process instead. This only requires generating random variables for the arrival epochs and jump sizes; one need not simulate service durations. In the large batch setting, this can deliver substantial savings in computation complexity, as large batches mean that a large number of service durations must be generated.

To draw upon results from the storage process literature and calculate explicit staffing levels, we will now assume that we are in the Markovian setting with N_t as a Poisson process with rate $\lambda > 0$ and with exponential service at rate $\mu > 0$. In the c-threshold storage process this corresponds to a release rule of $r(x) = \mu(x \wedge c)$. Similarly in the shot noise process the release rule would be $r(x) = \mu x$. Following standard stability assumptions for multi-server queueing

models we will also suppose $\lambda \mathrm{E}[B_1(n)] < cn\mu$ for all $n \in \mathbb{Z}^+$ and we suppose that in the limit we have $\lambda \mathrm{E}[M_1] < c\mu$ as well. Thus, the objects we use to determine the staffing levels will be the storage and shot noise processes in steady-state. We denote these as ψ_∞^C and ψ_∞, respectively. We now cite a result from the storage process literature providing integral equations for the steady-state densities of ψ_∞ and ψ_∞^C in Lemma 6.3.1.

Lemma 6.3.1. *The steady-state density of the shot noise process $f_\infty(\cdot)$ exists and is given by the unique solution to the integral equation*

$$f_\infty(x) = \frac{\lambda}{\mu x} \int_0^x \mathrm{P}(M_1 > x - y) f_\infty(y) \mathrm{d}y, \tag{6.8}$$

for all $x > 0$. Furthermore, the steady-state density of the c-threshold storage process $f_C(\cdot)$ exists and is given by the unique solution to the integral equation

$$f_C(x) = \frac{\lambda}{\mu(x \wedge c)} \int_0^x \mathrm{P}(M_1 > x - y) f_C(y) \mathrm{d}y, \tag{6.9}$$

for all $x > 0$.

Proof. This follows directly from Theorem 5 of Brockwell et al. (1982). □

As an alternate representation of the integrals in Lemma 6.3.1, we can observe that in the case of the threshold storage process, for example, we have

$$\int_0^x \mathrm{P}(M_1 > x - y) f_C(y) \mathrm{d}y = \mathrm{P}\left(M_1 + \psi_\infty^C > x\right) - \mathrm{P}\left(\psi_\infty^C > x\right), \tag{6.10}$$

since $\int_0^\infty \mathrm{P}(M_1 > x - y) f_C(y) \mathrm{d}y = \mathrm{P}\left(M_1 + \psi_\infty^C > x\right)$ and $\mathrm{P}(M_1 > x - y) = 1$ for all $y > x$. This expression will be of use to us in relating the two processes, further enabling us to use the shot noise process to understand the c-threshold storage process, just as we have used the infinite server queue to understand the multi-server queue. To begin, in Theorem 6.3.2 we will now use this alternate

expression to justify our study of the stationary setting through a validation of the interchange of the limits of time and of the batch scaling.

Theorem 6.3.2. *In the stationary Markovian infinite server and delay queueing models, the interchange of limits of time and batch scaling is justified. That is,*

$$\lim_{n\to\infty}\lim_{t\to\infty}\mathrm{P}\left(\frac{Q_t(n)}{n}\le x\right)=\lim_{t\to\infty}\lim_{n\to\infty}\mathrm{P}\left(\frac{Q_t(n)}{n}\le x\right), \tag{6.11}$$

and

$$\lim_{n\to\infty}\lim_{t\to\infty}\mathrm{P}\left(\frac{Q_t^C(n)}{n}\le x\right)=\lim_{t\to\infty}\lim_{n\to\infty}\mathrm{P}\left(\frac{Q_t^C(n)}{n}\le x\right), \tag{6.12}$$

for all $x>0$.

Proof. For the infinite server queueing model, this interchange can be quickly observed through differential equations for the moment generating functions of $Q_t(n)$ and ψ_t. Let $\mathcal{M}^n(\theta,t)$ be the moment generating function of the scaled Markovian infinite server queue, i.e. $\mathcal{M}^n(\theta,t)=\mathrm{E}\left[e^{\frac{\theta}{n}Q_t(n)}\right]$. Then, $\mathcal{M}^n(\theta,t)$ satisfies

$$\frac{\partial\mathcal{M}^n(\theta,t)}{\partial t}=\lambda\left(\mathrm{E}\left[e^{\frac{\theta}{n}B_1(n)}\right]-1\right)\mathcal{M}^n(\theta,t)+n\mu\left(e^{-\frac{\theta}{n}}-1\right)\frac{\partial\mathcal{M}^n(\theta,t)}{\partial\theta},$$

since $\frac{\partial\mathcal{M}^n(\theta,t)}{\partial\theta}=\mathrm{E}\left[\frac{Q_t(n)}{n}e^{\frac{\theta}{n}Q_t(n)}\right]$. Then, we have that for any $n\in\mathbb{Z}^+$ the moment generating function of the steady-state queue, say $\mathcal{M}^n(\theta,\infty)$, will be given by the solution to the time-equilibrium ordinary differential equation

$$0=\lambda\left(\mathrm{E}\left[e^{\frac{\theta}{n}B_1(n)}\right]-1\right)\mathcal{M}^n(\theta,\infty)+n\mu\left(e^{-\frac{\theta}{n}}-1\right)\frac{\mathrm{d}\mathcal{M}^n(\theta,\infty)}{\mathrm{d}\theta}.$$

As $n\to\infty$, the limiting steady-state object will then satisfy

$$0=\lambda\left(\mathrm{E}\left[e^{\theta M_1}\right]-1\right)\mathcal{M}^\infty(\theta,\infty)-\mu\theta\frac{\mathrm{d}\mathcal{M}^\infty(\theta,\infty)}{\mathrm{d}\theta}.$$

By comparison, the moment generating function of the shot noise process that yielded in the Markovian case of the batch scaling in Theorem 6.2.1, say $\mathcal{M}^\psi(\theta,t)$,

will satisfy

$$\frac{\partial \mathcal{M}^{\psi}(\theta,t)}{\partial t} = \lambda\left(\mathrm{E}\left[e^{\theta M_1}\right]-1\right)\mathcal{M}^{\psi}(\theta,t) - \mu\theta\frac{\partial \mathcal{M}^{\psi}(\theta,t)}{\partial \theta},$$

which implies that in steady-state the shot noise process moment generating function, say $\mathcal{M}^{\psi}(\theta,\infty)$, is given by the solution to

$$0 = \lambda\left(\mathrm{E}\left[e^{\theta M_1}\right]-1\right)\mathcal{M}^{\psi}(\theta,\infty) - \mu\theta\frac{\partial \mathcal{M}^{\psi}(\theta,\infty)}{\partial \theta}.$$

Hence, $\mathcal{M}^{\psi}(\theta,\infty) = \mathcal{M}^{\infty}(\theta,\infty)$, justifying Equation 6.11. To now prove Equation 6.12, we start with describing the balance equations for the queue. Letting $\pi_i^n = \lim_{t\to\infty} \mathrm{P}\left(Q_t(n) = i\right)$ for every $i \in \mathbb{N}$, we have that these steady-state probabilities satisfy

$$(\lambda + \mu(i \wedge cn))\,\pi_i^n = \lambda \sum_{j=1}^{i} \mathrm{P}\left(B_1(n) = j\right)\pi_{i-j}^n + \mu(i+1 \wedge cn)\pi_{i+1}^n,$$

for any $n \in \mathbb{Z}^+$. By induction, we can observe that this implies that the probabilities satisfy the recurrence relation

$$\pi_i^n = \frac{\lambda}{\mu(i \wedge cn)} \sum_{j=1}^{i} \mathrm{P}\left(B_1(n) \geq j\right)\pi_{i-j}^n,$$

for all $i \in \mathbb{Z}^+$. At $i = 1$ this follows immediately from the global balance equation for π_0^n, so we proceed to the inductive step and assume that the hypothesis holds on $i \in \{1,\ldots,k\}$ for some $k \in \mathbb{Z}^+$. Then, through this assumption and the balance equation for π_k^n, we can observe that

$$\lambda\pi_k^n + \lambda \sum_{j=1}^{k} \mathrm{P}\left(B_1(n) \geq j\right)\pi_{k-j}^n = \lambda \sum_{j=1}^{k} \mathrm{P}\left(B_1(n) = j\right)\pi_{k-j}^n + \mu(k+1 \wedge cn)\pi_{k+1}^n,$$

and since $\mathrm{P}\left(B_1(n) \geq 1\right)$ this simplifies to

$$\mu(k+1 \wedge cn)\pi_{k+1}^n = \lambda\pi_k^n + \lambda \sum_{j=1}^{k} \mathrm{P}\left(B_1(n) \geq j+1\right)\pi_{k-j}^n = \lambda \sum_{j=1}^{k+1} \mathrm{P}\left(B_1(n) \geq j\right)\pi_{k+1-j}^n,$$

which completes the induction. With this confirmation of the recursion, let us now observe an alternate representation of the summation within it. That is, for $Q^C_\infty(n)$ as the delay model in steady-state, one can note through the law of total probability that

$$\begin{aligned}\mathrm{P}\left(B_1(n) + Q^C_\infty(n) \geq i\right) &= \sum_{j=0}^{\infty} \mathrm{P}\left(B_1(n) \geq i - j\right) \pi^n_j \\ &= \sum_{j=0}^{i-1} \mathrm{P}\left(B_1(n) \geq i - j\right) \pi^n_j + \mathrm{P}\left(Q^C_\infty(n) \geq i\right),\end{aligned}$$

since $\mathrm{P}\left(B_1(n) \geq i - j\right) = 1$ for all $j \geq i$. This then implies that one can re-express the recurrence relation as

$$\pi^n_i = \frac{\lambda}{\mu(i \wedge cn)} \left(\mathrm{P}\left(B_1(n) + Q^C_\infty(n) \geq i\right) - \mathrm{P}\left(Q^C_\infty(n) \geq i\right)\right),$$

and we can now use this to give a representation for $F^n(x) \equiv \mathrm{P}\left(Q^C_\infty(n) \leq xn\right)$. Since $F^n(x) = \sum_{i=0}^{\lfloor xn \rfloor} \pi^n_i$, we have that

$$F^n(x) = \pi^n_0 + \sum_{i=1}^{\lfloor xn \rfloor} \frac{\lambda}{\mu(i \wedge cn)} \left(\mathrm{P}\left(B_1(n) + Q^C_\infty(n) \geq i\right) - \mathrm{P}\left(Q^C_\infty(n) \geq i\right)\right).$$

By changing the step size of the summation to being in increments of $\frac{1}{n}$, this sum becomes

$$F^n(x) = \pi^n_0 + \sum_{\substack{i=\frac{1}{n}, \\ \Delta=\frac{1}{n}}}^{\lfloor xn \rfloor / n} \frac{\lambda}{n\mu(i \wedge c)} \left(\mathrm{P}\left(\frac{B_1(n)}{n} + \frac{Q^C_\infty(n)}{n} \geq i\right) - \mathrm{P}\left(\frac{Q^C_\infty(n)}{n} \geq i\right)\right).$$

Letting X_∞ be equivalent in distribution to the limiting object of $\frac{Q^C_\infty(n)}{n}$ as $n \to \infty$, we have that $F^\infty(x) = \mathrm{P}\left(X_\infty \leq x\right)$ is given by

$$F^\infty(x) = \int_0^x \frac{\lambda}{\mu(z \wedge c)} \left(\mathrm{P}\left(M_1 + X_\infty \geq z\right) - \mathrm{P}\left(X_\infty \geq z\right)\right) \mathrm{d}z, \tag{6.13}$$

for all $x > 0$, since $\pi^n_0 \to 0$ and $\frac{B_1(n)}{n} \overset{D}{\Longrightarrow} M_1$ as $n \to \infty$. Using Lemma 6.3.1 and the alternate representation of the integral in Equation 6.10, one can note that

$F^C(x) = \mathrm{P}\left(\psi_\infty^C \leq x\right)$ will be given by

$$F^C(x) = \int_0^x \frac{\lambda}{\mu(z \wedge c)} \left(\mathrm{P}\left(M_1 + \psi_\infty^C \geq z\right) - \mathrm{P}\left(\psi_\infty^C \geq z\right)\right) \mathrm{d}z.$$

From Lemma 6.3.1 we have that $F^C(x)$ is the unique distribution satisfying this equation and thus $X_\infty \stackrel{D}{=} \psi_\infty^C$, completing the proof. □

Having now justified the interchange of limits, we will now develop an asymptotic approach to calculate the exceedance probabilities for general batch sizes in Subsection 6.3.1 through use of the results from the storage process literature. It is worth noting that in specific settings the integral equations in Lemma 6.3.1 can yield results directly. An example of this is in the case of exponential distributed marks, which arise as the limit of geometrically distributed batches. The resulting equations in this setting are straightforward to solve, and we demonstrate this in Section C.3 of the Appendix.

6.3.1 Asymptotic Analysis for General Batch Sizes

To calculate the exceedance probabilities for ψ_∞^C, we will again draw upon its relationship with the tractable shot noise process, ψ_∞. Furthermore, we will also make use of a transform method for computing the cumulative distribution function and truncated expectation of a random variable through use of orthogonal Legendre polynomials. This approach is based on an generalization of Sullivan et al. (1980), in which the authors provide a representation for the indicator function through a sum of exponential functions. In Section C.1 of the Appendix, we extend this result for use in studying continuous random variables. Through use of the resulting Lemma C.1.1, we derive the following expressions for the exceedance probabilities in Theorem 6.3.3.

Theorem 6.3.3. *In the Markovian case, the threshold exceedance probabilities for ψ_∞^C are given by*

$$\mathrm{P}\left(\psi_\infty^C > c\right) = \lim_{m\to\infty} \frac{\frac{\lambda}{\mu}\mathrm{E}\left[M_1\right] - \sigma_{m,c}^{(C1)}}{c - \sigma_{m,c}^{(C1)}}, \tag{6.14}$$

and

$$\mathrm{P}\left(\psi_\infty^C + M_1 > c\right) = \lim_{m\to\infty} \frac{\frac{\lambda}{\mu}\mathrm{E}\left[M_1\right] - \sigma_{m,c}^{(C2)}}{c - \sigma_{m,c}^{(C2)}} + \frac{\sigma_{m,c}^{(C1)}\left(\frac{c\mu}{\lambda} - \mathrm{E}\left[M_1\right]\right)}{\left(c - \sigma_{m,c}^{(C1)}\right)\left(c - \sigma_{m,c}^{(C2)}\right)}, \tag{6.15}$$

where for $m \in \mathbb{Z}^+$ and c as the capacity threshold, $\sigma_{m,c}^{(C1)}$ is given by

$$\sigma_{m,c}^{(C1)} = \sum_{k=1}^{m} \frac{c\lambda}{\mu k}\left(1 - \mathrm{E}\left[e^{-\frac{k}{c}M_1}\right]\right) \frac{a_k^m e^{-\lambda\int_0^\infty\left(1-\mathrm{E}\left[e^{-\frac{k}{c}M_1 e^{-\mu x}}\right]\right)\mathrm{d}x}}{\sum_{i=1}^m a_i^m e^{-\lambda\int_0^\infty\left(1-\mathrm{E}\left[e^{-\frac{i}{c}M_1 e^{-\mu x}}\right]\right)\mathrm{d}x}}, \tag{6.16}$$

and $\sigma_{m,c}^{(C2)}$ is given by

$$\sigma_{m,c}^{(C2)} = \sum_{k=1}^{m} \frac{\mathrm{E}\left[M_1 e^{-\frac{k}{c}M_1}\right] + \mathrm{E}\left[e^{-\frac{k}{c}M_1}\right]\frac{c\lambda}{\mu k}\left(1 - \mathrm{E}\left[e^{-\frac{k}{c}M_1}\right]\right)}{\sum_{i=1}^m a_i^m \mathrm{E}\left[e^{-\frac{i}{c}M_1}\right] e^{-\lambda\int_0^\infty\left(1-\mathrm{E}\left[e^{-\frac{i}{c}M_1 e^{-\mu x}}\right]\right)\mathrm{d}x}} a_k^m e^{-\lambda\int_0^\infty\left(1-\mathrm{E}\left[e^{-\frac{k}{c}M_1 e^{-\mu x}}\right]\right)\mathrm{d}x}, \tag{6.17}$$

with a_k^m as defined in Equation C.1.

Proof. To begin, we first recall that Equation 6.10 gives us that

$$(x \wedge c) f_C(x) = \frac{\lambda}{\mu}\left(\mathrm{P}\left(M_1 + \psi_\infty^C > x\right) - \mathrm{P}\left(\psi_\infty^C > x\right)\right),$$

and by integrating each side across all x this further implies that

$$\mathrm{E}\left[\psi_\infty^C \wedge c\right] = \frac{\lambda}{\mu}\left(\mathrm{E}\left[M_1 + \psi_\infty^C\right] - \mathrm{E}\left[\psi_\infty^C\right]\right),$$

yielding that $\mathrm{E}\left[\psi_\infty^C \wedge c\right] = \frac{\lambda}{\mu}\mathrm{E}\left[M_1\right]$. This same expectation can also be expressed through conditioning as

$$\mathrm{E}\left[\psi_\infty^C \wedge c\right] = c\mathrm{P}\left(\psi_\infty^C > c\right) + \mathrm{E}\left[\psi_\infty^C \mid \psi_\infty^C \le c\right]\left(1 - \mathrm{P}\left(\psi_\infty^C > c\right)\right),$$

and thus by setting these two expressions equal to one another we find that

$$\mathrm{P}\left(\psi_\infty^C > c\right) = \frac{\frac{\lambda}{\mu}\mathrm{E}\left[M_1\right] - \mathrm{E}\left[\psi_\infty^C \mid \psi_\infty^C \le c\right]}{c - \mathrm{E}\left[\psi_\infty^C \mid \psi_\infty^C \le c\right]}. \tag{6.18}$$

Although we do not know this truncated mean of ψ_∞^C in closed form, we can observe that

$$\mathrm{E}\left[\psi_\infty^C \mid \psi_\infty^C \le c\right] = \mathrm{E}\left[\psi_\infty \mid \psi_\infty \le c\right],$$

because the integral equations of the these truncated densities are equivalent for all $x \in (0, c]$, as can be observed through Lemma 6.3.1. Now, by total probability we can recognize that

$$\mathrm{E}\left[\psi_\infty \mid \psi_\infty \le c\right] = \frac{\mathrm{E}\left[\psi_\infty \mathbf{1}\{\psi_\infty \le c\}\right]}{\mathrm{P}\left(\psi_\infty \le c\right)}.$$

For $m \in \mathbb{Z}^+$, we now define the quantities $\sigma_{m,c}^{(1)}$ and $\sigma_{m,c}^{(2)}$ as

$$\sigma_{m,c}^{(1)} = \sum_{k=1}^{m} a_k^m e^{-\lambda \int_0^\infty \left(1 - \mathrm{E}\left[e^{-\frac{k}{c}M_1 e^{-\mu x}}\right]\right)\mathrm{d}x},$$

and

$$\sigma_{m,c}^{(2)} = \sum_{k=1}^{m} \frac{c\lambda a_k^m}{\mu k}\left(1 - \mathrm{E}\left[e^{-\frac{k}{c}M_1}\right]\right) e^{-\lambda \int_0^\infty \left(1 - \mathrm{E}\left[e^{-\frac{k}{c}M_1 e^{-\mu x}}\right]\right)\mathrm{d}x}.$$

Using Theorem 6.2.1 and Lemma C.1.1, we have that $\sigma_{m,c}^{(1)} \to \mathrm{P}\left(\psi_\infty \le c\right)$ and $\sigma_{m,c}^{(2)} \to \mathrm{E}\left[\psi_\infty \mathbf{1}\{\psi_\infty \le c\}\right]$ as $m \to \infty$. Thus, by substituting $\sigma_{m,c}^{(C1)} = \sigma_{m,c}^{(2)}/\sigma_{m,c}^{(1)}$ into Equation 6.18 and simplifying, we achieve the stated form in Equation 6.14.

To now prove Equation 6.15, we start by finding an identity for $\mathrm{E}\left[\psi_\infty^C + M_1 \wedge c\right]$. Because Lemma 6.3.1 implies that the threshold storage process density $f_C(x)$ satisfies

$$(x \wedge c) f_C(x) = \frac{\lambda}{\mu}\left(\mathrm{P}\left(\psi_\infty^C + M_1 > x\right) - \mathrm{P}\left(\psi_\infty^C > x\right)\right),$$

we are able to observe that

$$\begin{aligned}
\mathrm{E}\left[\psi_\infty^C \mathbf{1}\{\psi_\infty^C < c\}\right] &= \int_0^c (x \wedge c) f_C(x)\mathrm{d}x \\
&= \frac{\lambda}{\mu}\int_0^c \mathrm{P}\left(\psi_\infty^C + M_1 > x\right)\mathrm{d}x - \frac{\lambda}{\mu}\int_0^c \mathrm{P}\left(\psi_\infty^C > x\right)\mathrm{d}x \\
&= \frac{\lambda}{\mu}\mathrm{E}\left[\psi_\infty^C + M_1 \wedge c\right] - \frac{\lambda}{\mu}\mathrm{E}\left[\psi_\infty^C \wedge c\right].
\end{aligned}$$

Because we know that $\mathrm{E}\left[\psi_\infty^C \wedge c\right] = \frac{\lambda}{\mu}\mathrm{E}\left[M_1\right]$ and $\mathrm{E}\left[\psi_\infty^C \mid \psi_\infty^C \leq c\right] = \mathrm{E}\left[\psi_\infty \mid \psi_\infty \leq c\right]$, we can note that this now implies that expectation of the minimum of the threshold and the storage process plus a jump is equal to

$$\mathrm{E}\left[\psi_\infty^C + M_1 \wedge c\right] = \frac{\mu}{\lambda}\mathrm{E}\left[\psi_\infty \mid \psi_\infty \leq c\right]\mathrm{P}\left(\psi_\infty^C \leq c\right) + \frac{\lambda}{\mu}\mathrm{E}\left[M_1\right],$$

all of which on the right-hand side we now know how to calculate. Then, by mimicking the conditioning decomposition we used previously on $\mathrm{E}\left[\psi_\infty^C \wedge c\right]$, we can note that $\mathrm{E}\left[\psi_\infty^C + M_1 \wedge c\right]$ is also equal to

$$\mathrm{E}\left[\psi_\infty^C + M_1 \wedge c\right] = c\mathrm{P}\left(\psi_\infty^C + M_1 > c\right) + \mathrm{E}\left[\psi_\infty^C + M_1 \mid \psi_\infty^C + M_1 \leq c\right]\left(1 - \mathrm{P}\left(\psi_\infty^C + M_1 > c\right)\right).$$

By setting these two expressions for $\mathrm{E}\left[\psi_\infty^C + M_1 \wedge c\right]$ equal to one another and solving for $\mathrm{P}\left(\psi_\infty^C + M_1 > c\right)$, we have that

$$\mathrm{P}\left(\psi_\infty^C + M_1 > c\right) = \frac{\frac{\mu}{\lambda}\mathrm{E}\left[\psi_\infty \mid \psi_\infty \leq c\right]\mathrm{P}\left(\psi_\infty^C \leq c\right) + \frac{\lambda}{\mu}\mathrm{E}\left[M_1\right] - \mathrm{E}\left[\psi_\infty^C + M_1 \mid \psi_\infty^C + M_1 \leq c\right]}{c - \mathrm{E}\left[\psi_\infty^C + M_1 \mid \psi_\infty^C + M_1 \leq c\right]} \tag{6.19}$$

Again through the integral equations, we can recognize that $\mathrm{E}\left[\psi_\infty^C + M_1 \mid \psi_\infty^C + M_1 \leq c\right] = \mathrm{E}\left[\psi_\infty + M_1 \mid \psi_\infty + M_1 \leq c\right]$. Because M_1 is independent from the state of the shot noise process ψ_∞, we have that

$$\mathrm{E}\left[e^{\theta(\psi_\infty + M_1)}\right] = \mathrm{E}\left[e^{\theta\psi_\infty}\right]\mathrm{E}\left[e^{\theta M_1}\right] = \mathrm{E}\left[e^{\theta M_1}\right]e^{-\lambda\int_0^\infty\left(1-\mathrm{E}\left[e^{\theta M_1 e^{-\mu x}}\right]\right)\mathrm{d}x},$$

by use of Theorem 6.2.1. Then, for $m \in \mathbb{Z}^+$ let us additionally define $\sigma_{m,c}^{(3)}$ and $\sigma_{m,c}^{(4)}$ such that

$$\sigma_{m,c}^{(3)} = \sum_{k=1}^{m} a_k^m \mathrm{E}\left[e^{-\frac{k}{c}M_1}\right]e^{-\lambda\int_0^\infty\left(1-\mathrm{E}\left[e^{-\frac{k}{c}M_1 e^{-\mu x}}\right]\right)\mathrm{d}x},$$

and

$$\sigma_{m,c}^{(4)} = \sum_{k=1}^{m} a_k^m \left(\mathrm{E}\left[M_1 e^{-\frac{k}{c}M_1} \right] + \mathrm{E}\left[e^{-\frac{k}{c}M_1} \right] \frac{c\lambda}{\mu k} \left(1 - \mathrm{E}\left[e^{-\frac{k}{c}M_1} \right] \right) \right) e^{-\lambda \int_0^\infty \left(1 - \mathrm{E}\left[e^{-\frac{k}{c}M_1 e^{-\mu x}} \right] \right) \mathrm{d}x}.$$

Through these definitions, Lemma C.1.1 yields that $\sigma_{m,c}^{(3)} \to \mathrm{P}\left(\psi_\infty^C + M_1 \leq c\right)$ and $\sigma_{m,c}^{(4)} \to \mathrm{E}\left[(\psi_\infty + M_1)\mathbf{1}\{\psi_\infty + M_1 \leq c\}\right]$ as $m \to \infty$. Thus we have that $\sigma_{m,c}^{(C2)} = \sigma_{m,c}^{(4)} / \sigma_{m,c}^{(3)} \to \mathrm{E}\left[\psi_\infty + M_1 \mid \psi_\infty + M_1 \leq c\right]$, and this completes the proof. □

The derivations behind this computational methodology also enable us to provide a closed form expression for the utilization of the teleoperators, meaning the ratio between the mean number of busy servers and the total number employed. Because the number of busy servers is the minimum of the number of jobs in the system and the staffing level, we can make use of the expected value of the minimum of the storage process and the threshold c. In Corollary 6.3.4, we observe that with this expectation in hand, the utilization is easy to compute.

Corollary 6.3.4. *In the Markovian case, the steady-state utilization of the teleoperators in the large batch setting is given by* $\frac{1}{c}\mathrm{E}\left[\psi_\infty^C \wedge c\right] = \frac{\lambda}{c\mu}\mathrm{E}\left[M_1\right]$.

Proof. In the proof of Theorem 6.3.3, we have seen that $\mathrm{E}\left[\psi_\infty^C \wedge c\right] = \frac{\lambda}{\mu}\mathrm{E}\left[M_1\right]$. Through the batch scaling in Theorem 6.2.2, the teleoperation utilization converges to

$$\mathrm{E}\left[\frac{1}{cn}\left(Q_\infty^C(n) \wedge cn\right)\right] \longrightarrow \frac{1}{c}\mathrm{E}\left[\psi_\infty^C \wedge c\right] = \frac{\lambda}{c\mu}\mathrm{E}\left[M_1\right],$$

as $n \to \infty$. □

As a side consequence of the proof of Theorem 6.3.3, we can also identify a practical, closed-form upper bound on $\mathrm{P}\left(\psi_\infty^C > c\right)$. To do so we bound first find

a lower bound for the truncated mean $\mathrm{E}\left[\psi_\infty^C \mid \psi_\infty^C \leq c\right] = \mathrm{E}\left[\psi_\infty \mid \psi_\infty \leq c\right]$. Letting $\bar{f}(x)$ be the truncated density on $(0, c]$, through Lemma 6.3.1 we then have that

$$\mathrm{E}\left[\psi_\infty \mid \psi_\infty \leq c\right] = \int_0^c x\bar{f}(x)\mathrm{d}x = \int_0^c \frac{\lambda}{\mu}\left(\mathrm{P}\left(M_1 + \psi_\infty > x \mid \psi_\infty \leq c\right) - \mathrm{P}\left(\psi_\infty > x \mid \psi_\infty \leq c\right)\right)\mathrm{d}x.$$

Because $\int_0^c \mathrm{P}\left(\psi_\infty > x \mid \psi_\infty \leq c\right)\mathrm{d}x = \mathrm{E}\left[\psi_\infty \mid \psi_\infty \leq c\right]$, we have

$$\mathrm{E}\left[\psi_\infty \mid \psi_\infty \leq c\right] = \frac{\lambda}{\lambda+\mu}\int_0^c \mathrm{P}\left(M_1 + \psi_\infty > x \mid \psi_\infty \leq c\right)\mathrm{d}x.$$

Then, by observing that $\mathrm{P}\left(M_1 + \psi_\infty > x \mid \psi_\infty \leq c\right) \geq \mathrm{P}\left(M_1 > x\right)$ through the independence of the two quantities and the fact that each is positive, we furthermore have

$$\mathrm{E}\left[\psi_\infty \mid \psi_\infty \leq c\right] \geq \frac{\lambda}{\lambda+\mu}\int_0^c \mathrm{P}\left(M_1 > x\right)\mathrm{d}x = \frac{\lambda}{\lambda+\mu}\mathrm{E}\left[M_1 \wedge c\right].$$

Using the decomposition in Equation 6.18, this now yields the upper bound

$$\mathrm{P}\left(\psi_\infty^C > c\right) \leq \frac{\frac{\lambda}{\mu}\mathrm{E}\left[M_1\right] - \frac{\lambda}{\lambda+\mu}\mathrm{E}\left[M_1 \wedge c\right]}{c - \frac{\lambda}{\lambda+\mu}\mathrm{E}\left[M_1 \wedge c\right]}.$$

This bound is most helpful in cases of small λ, as in that case ψ_∞ is likely to be small. This small arrival rate setting is practical for our motivating teleoperations scenario. As we have discussed, jobs arrive to the teleoperations system at disengagements. As self-driving technology improves, disengagements should become increasingly rare. In fact, based on the public statistics of industry leading organizations, disengagements may already be quite rare. In the following section, we will explore this data in greater detail. Using the theoretical results we have obtained for these system, in Section 6.4 we will calculate the necessary staffing levels for teleoperations systems in a variety of real world settings through a collection of public data sets on autonomous vehicles and driving behavior.

6.4 Numerical Results and Experiments

Before exploring the practical implications of our theoretical results, let us first review the data we use in the subsequent calculations. These data sets are of two main types: data on autonomous vehicle performance and data on driving behavior in the United States. The former comes from the California disengagement reports, which are disclosures of all autonomous vehicle disengagements required from every company testing driverless cars on the state's public roads. We will focus our analysis on the reports from two industry leaders, GM Cruise LLC (2019) and Waymo LLC (2019). These reports allow us to calculate a *disengagement rate*, meaning the average number of miles driven between autonomous vehicle disengagements, which enables us to convert from number of miles driven per hour to number of disengagements per hour, which is the arrival rate to the teleoperations queueing system. Using the most recently available data, Waymo leads the industry with a rate of 11,154.3 miles per disengagement and GM Cruise is in second with 5,204.9 miles per disengagement. It is worth noting though that GM Cruise tests exclusively in downtown San Francisco, and we will use this rate to address urban driving situations.

For driving behavior data, we will consider two different types of drivers: individuals and taxis. As we have discussed in the introduction, taxis are likely to be the first widespread public deployment of autonomous vehicles, since the technology involves expensive hardware that may be prohibitively expensive for individual consumers and is more cost-effective in a highly utilized vehicle like a taxi. For taxi data, we turn to the 2014 and 2018 New York City Taxicab Factbooks and their accompanying datasets New York City Taxi & Limousine Commission (2014, 2018), which document the driving activity in NYC in ex-

tensive detail, and a 2019 taxi study from the Los Angeles Department of Transportation with a focus on key arterial roadways Sam Schwartz Engineering (2019). For individual driving behavior, we use the most recent National Household Travel Survey (NHTS) from the Federal Highway Administration, which contains summary statistics on driving in major census designated geographic areas in the United States Federal Highway Administration (2017). From this data, we know that the average American drives approximately 13,476 miles each year, which puts the performance of Waymo's and GM Cruise's disengagement rates in an impressive context.

6.4.1 Staffing the Teleoperations System from Data

As an opening numerical experiment, let us consider the early phases of autonomous vehicle public deployment. In Figure 6.8, we plot the number of remote operators needed to support an urban fleet of driverless taxis as the number of vehicles increases. In terms of the teleoperations system, as the number of vehicles grows so will the rate of arrivals for service grow. Based on the 2014 NYC Taxi Factbook New York City Taxi & Limousine Commission (2014), we assume that each vehicle drives 70,000 miles per year. Furthermore, we are considering the PM shift, in which NYC taxis average 63.0% of their daily driven miles per the 2018 NYC Taxi Factbook New York City Taxi & Limousine Commission (2018). To capture disengagements in a dense urban road network, we use GM Cruise's 2018 San Francisco disengagement rate. Then, we conservatively assume an average service time of 1 minute, and we suppose that the scaled limit of batch distribution leads to a jump size distribution that is log normally distributed with mean 1 and variance 0.5. Although the moment gen-

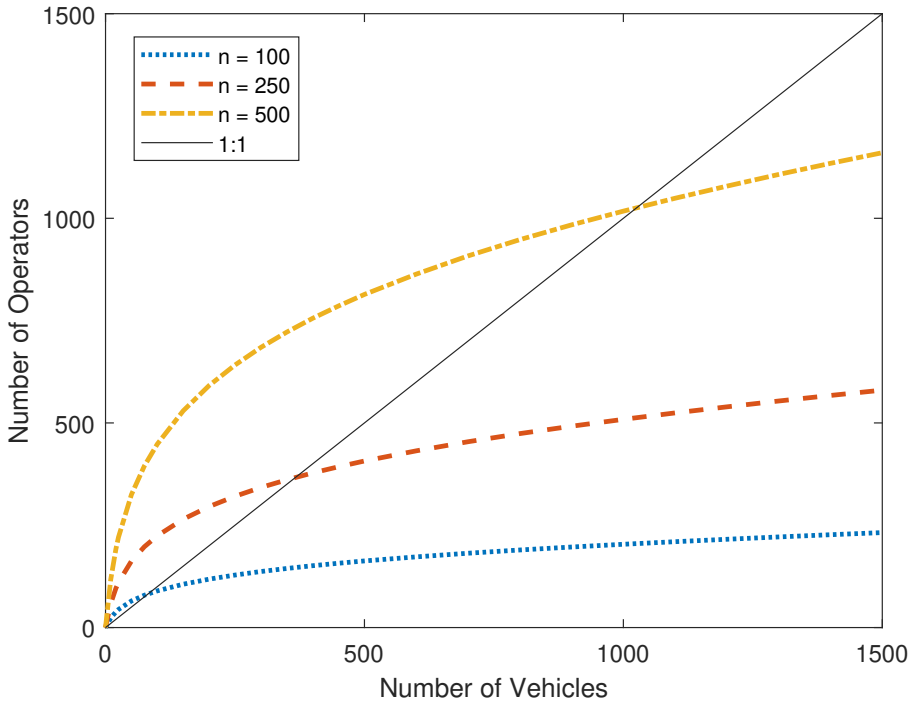


Figure 6.8: Number of teleoperators needed to support an autonomous taxi fleet as the fleet size grows, based on data from the 2014 and 2018 NYC Taxi Factbooks New York City Taxi & Limousine Commission (2014, 2018) and GM Cruise's 2018 CA disengagement reports GM Cruise LLC (2019).

erating function of the log normal is not known in closed form, we are able to use the approximation provided in Asmussen et al. (2016), which performs quite well in simulation comparisons. We calculate the operator to batch size ratio c needed to achieve an $\epsilon = 0.001$ exceedance probability for $\mathrm{P}\left(\psi_\infty^C > c\right)$ through use of Theorem 6.3.3. Then, the three dashed and dotted curves in Figure 6.8 are the corresponding number of operators for relative batch sizes of $n \in \{100, 250, 500\}$. For comparison, we also plot a 45° line that represents a one-to-one pairing of operators and vehicles, capturing the requisite staffing of either standard taxis or the in-car safety driver autonomous vehicle support method. As can be observed, by fleets of size 1,000, even the most extreme batch size setting ($n = 500$) has at least as efficient staffing as the number of vehicles, if not substantially

more efficient. For reference, in the lowest hour of activity there was still an average of over 30,000 taxis actively providing service in NYC New York City Taxi & Limousine Commission (2018). This suggests that teleoperations systems have significant potential for efficiency, and may be the key to unlocking the benefits of the ideal but potentially unreachable level 5 autonomous vehicles at level 4.

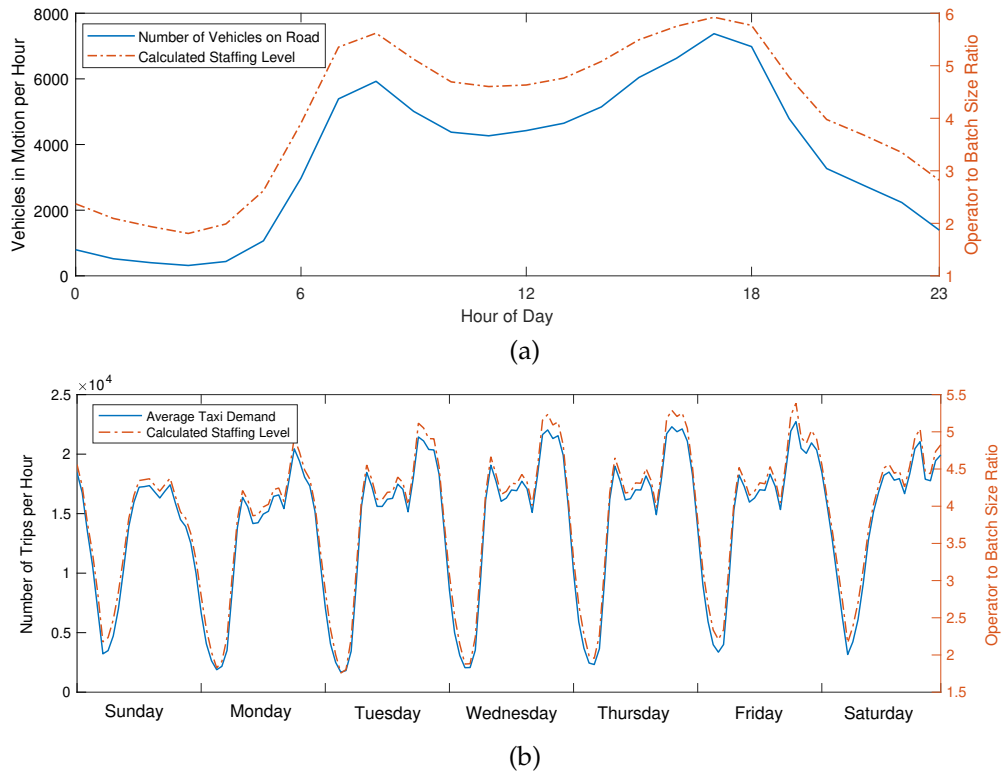


Figure 6.9: Necessary staffing-to-batch ratio across time to support (a) typical traffic on major arterial roadways in Los Angeles using LADOT data Sam Schwartz Engineering (2019) and (b) medallion taxi demand in New York City, based on data from the 2018 NYC Taxi Factbook New York City Taxi & Limousine Commission (2018).

This leads us to consider the necessary staffing ratios for supporting a much larger network of taxis and in time varying settings. Thus, in Figure 6.9 we compute the operator to batch size ratios needed to support the demand in each hour of the day for all vehicles on key arterial roadways in Los Angeles (Fig-

ure 6.9a) and for all medallion taxis in New York City (Figure 6.9b), using the same disengagement rate, service time, and jump distributions as in Figure 6.8. In the Los Angeles setting, the staffing ratio is compared to the total number of vehicles (taxis or otherwise) in motion per hour in the 14 arterial corridors as provided on page 73 of Sam Schwartz Engineering (2019). In the New York setting, this is compared to the total number of trips served by medallion taxis (i.e. yellow cabs) in each hour of each day of the week using data from New York City Taxi & Limousine Commission (2018).

Table 6.1: Staffing ratios and utilizations as calculated for total peak hour traffic in the ten largest U.S. metropolitan areas, based on data from the 2017 National Household Travel Survey Federal Highway Administration (2017).

	Metropolitan area	Annual miles driven (M)	Number of vehicles (M)	Operator to batch size ratio ($p = 0$)	Operator to batch size ratio ($p = 1$)	Teleoperator utilization ($p = 0$)	Teleoperator utilization ($p = 1$)
1	New York, NY	93,512	9.33	27.1	28.5	87.3%	83.0%
2	Los Angeles, CA	71,791	7.80	21.5	22.9	84.4%	79.2%
3	Dallas, TX	50,231	4.81	15.9	17.3	79.9%	73.3%
4	Chicago, IL	49,348	5.75	15.7	17.1	79.7%	73.0%
5	Atlanta, GA	42,547	3.86	13.9	15.3	77.6%	70.3%
6	Houston, TX	42,431	3.99	13.8	15.3	77.6%	70.3%
7	Washington, DC	41,199	4.57	13.5	14.9	77.2%	69.8%
8	Minneapolis, MN	34,540	3.14	11.7	13.2	74.6%	66.5%
9	Philadelphia, PA	32,781	3.60	11.2	12.7	73.8%	65.5%
10	Phoenix, AZ	31,408	2.97	10.9	12.3	73.1%	64.6%

This now motivates us to project the total staffing level needed to support all vehicle traffic at peak hour in the ten largest US metropolitan areas in terms of annual miles driven, as per the 2017 NHTS Federal Highway Administration (2017). Since this is an idealized long term scenario, we will now use Waymo LLC (2019)'s industry leading rate of 11,154.3 miles per disengagement and to compute both $\mathrm{P}\left(\psi_\infty^C > c\right)$ and $\mathrm{P}\left(\psi_\infty^C + M_1 > c\right)$, abbreviated as cases $p = 0$ and $p = 1$ based on our discussion in Subsection 6.2.1, we will now assume that

the scaled limit of the batch size distribution produces a deterministic jump size. Note that this occurs for any batch size distribution that at n the batch size can be equivalently decomposed into n i.i.d. random variables, and this family of "finitely divisible" random variables includes commonly used distributions such as the Poisson and the binomial, see Chapter 3. We will again take a mean service time of 1 minute and a target exceedance probability of $\epsilon = 0.001$. We can observe from New York City Taxi & Limousine Commission (2018) that approximately 6.1% of the miles driven occur in the most active hour, and thus we will use this as the base of our peak hour analysis. In Table 6.1 we give the resulting operator to batch size ratios for these two performance measures in each metropolitan area, and we compare it to the known number of miles driven annually and the total number of vehicles owned in the area Federal Highway Administration (2017). Through use of Corollary 6.3.4, we also compute the utilization of the teleoperators in each city. One can note that in this high volume setting, the utilization is also fairly high. Because Figure 6.9 and Table 6.1 list the operator to batch size ratio, it is important that note that this ratio c does not outright depend on the batch size itself and only requires that the batch size is large enough. Thus, the ratios in Table 6.1 hold for any relative batch size n that is sufficiently large, and these staffing levels can be found by simply multiplying c by n.

By inspecting Figure 6.8 or Table 6.1, we can observe an important fact about not only this teleoperations system, but also batch arrival service systems in general. As one can observe in Table 6.1, the operator to batch size ratio does not increase linearly with the annual miles driven (which is proportional to the rate of disengagements). For example, the number of miles driven each year in the Minneapolis metropolitan area is less than half of how many miles are driven

in the L.A. area each year (34.5B to 71.8B), yet the resulting ratios of operators to batch size do not exhibit the same relationship (11.7 to 21.5 for $p = 0$, and 13.2 to 22.9 for $p = 1$). This effect is perhaps even more noticeable in Figure 6.8, as each curve in the graph is eventually dominated by the 45° line. It is worth noting that this sub-linear effect of the arrival rate on the staffing level is common in queueing theory, as can be seen in the literature on the pooling principle or square root staffing rule, for example. This same idea is what tips the scale towards teleoperations in the trade-off between its many-to-one when needed staffing level and the one-to-one always pairing when using drivers or in-car safety operators, and, moreover, in comparison with Theorems 6.2.1 and 6.2.2 this shows us the strong impact that batches can have on a service system. Unlike the arrival rate's sublinear effect on the staffing level, these results imply that batches have a precisely linear effect. Since the staffing level grows linearly with n, doubling the batch size will exactly double the necessary staffing level. This can be seen in Figure 6.8, as the curve for $n = 500$ is precisely double the curve for $n = 250$. This emphasizes one of the key take-aways our work: batches of arrivals have a pronounced impact on service systems and thus must be addressed with care.

6.4.2 Observations from Normal Approximations

This comparison with the square root staffing rule also motivates an approximation for the staffing level. Just as Gaussian approximations of infinite server queues can be used to study multi-server queues à la works like Halfin and Whitt (1981); Jennings et al. (1996); Garnett et al. (2002), we can combine central limit theorem results for the infinite capacity shot noise process and the batch

scaling results of Theorems 6.2.1 and 6.2.2 to propose a normal-based approximation for multi-server queues with large batch arrivals. Using Theorem 6.2.1, we can note that the steady-state mean and variance of a Poisson-driven shot noise process are

$$\rho_\psi \equiv \mathrm{E}\left[\psi_t\right] = \lambda \mathrm{E}\left[M_1\right] \int_0^\infty \bar{G}(x)\mathrm{d}x \quad \text{and} \quad \sigma_\psi^2 \equiv \mathrm{Var}\left(\psi_t\right) = \lambda \mathrm{E}\left[M_1^2\right] \int_0^\infty \bar{G}^2(x)\mathrm{d}x.$$

Then, central limit results in Rice (1977); Lane (1984) yield that

$$\frac{\psi_t - \rho_\psi}{\sigma_\psi} \stackrel{D}{\Longrightarrow} N(0,1),$$

as $\lambda \to \infty$. Then, the normal approximation yields a storage process analog of the square root staffing rule such that

$$\tilde{c}_\epsilon = \rho_\psi + z_\epsilon \sigma_\psi$$

is an approximate threshold level for the c-threshold storage process ψ_t^C so that

$$\mathrm{P}\left(\psi_t^C > \tilde{c}_\epsilon\right) \approx \mathrm{P}\left(\psi_t > \tilde{c}_\epsilon\right) \approx 1 - \Phi(z_\epsilon) = \epsilon,$$

where $\Phi(\cdot)$ is the CDF of a standard normal distribution. Then, to convert from storage processes to large batch arrival queues, Theorem 6.2.2 suggests multiplying $\tilde{c}_\epsilon$ by the relative batch size n to receive an approximate queueing staffing level of

$$n\tilde{c}_\epsilon = n\left(\rho_\psi + z_\epsilon \sigma_\psi\right)$$

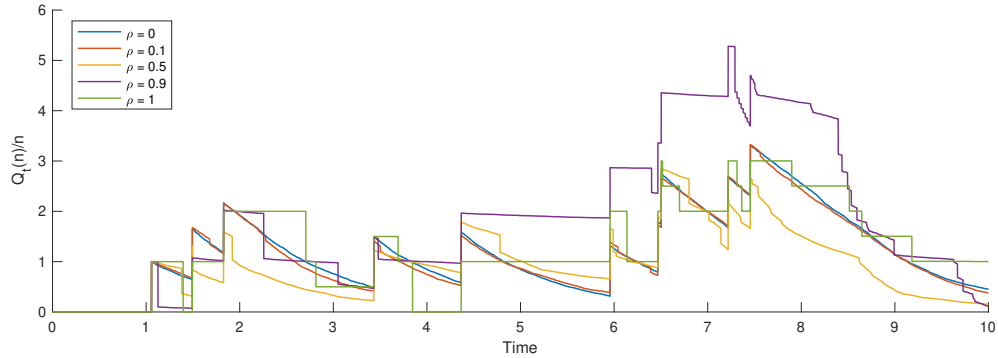
for the target exceedance probability ϵ. While this is only an approximation, it does capture the sublinear growth of the staffing level with increase in λ as well as the precisely linear growth of the staffing with n. This is an important observation, as a seemingly plausible approach to staffing the multi-server queue with batches of size n could be to use well-known staffing levels for the single arrival multi-server queue with arrival rate λn. While the infinite server analogs

of these systems will be equal in mean, we can see through the preceding analysis that addressing batch arrivals as if they merely constituted an increase in the arrival rate will always lead to inadequate staffing.

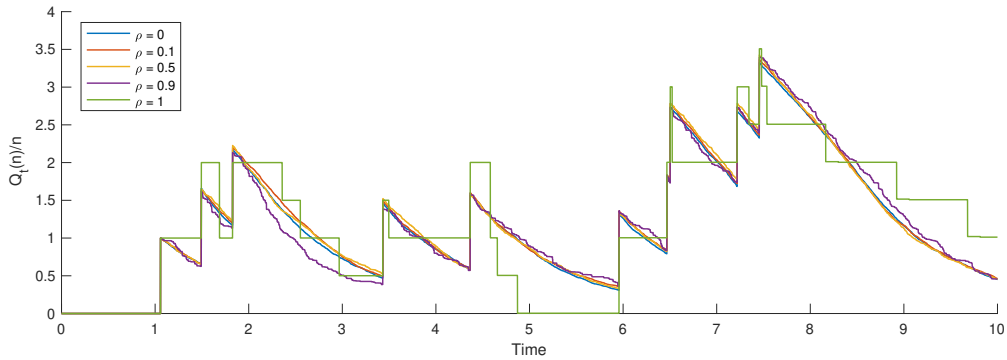
It is also worth noting that we can verify this linear-in-n and sublinear-in-λ behavior from known expressions for the infinite server batch arrival queue, e.g. in Chapter 3. For example, for the simple case of the $M^n/M/\infty$ model with arrival rate λ and service rate μ, the mean and variance are known to be $\frac{n\lambda}{\mu}$ and $\frac{n(n+1)\lambda}{2\mu}$. A corresponding normal approximation would thus be

$$\frac{n\lambda}{\mu} + z_\epsilon \sqrt{\frac{n(n+1)\lambda}{2\mu}} = n\rho_\psi + nz_\epsilon \sqrt{\frac{n+1}{n}}\sigma_\psi > n\tilde{c}_n.$$

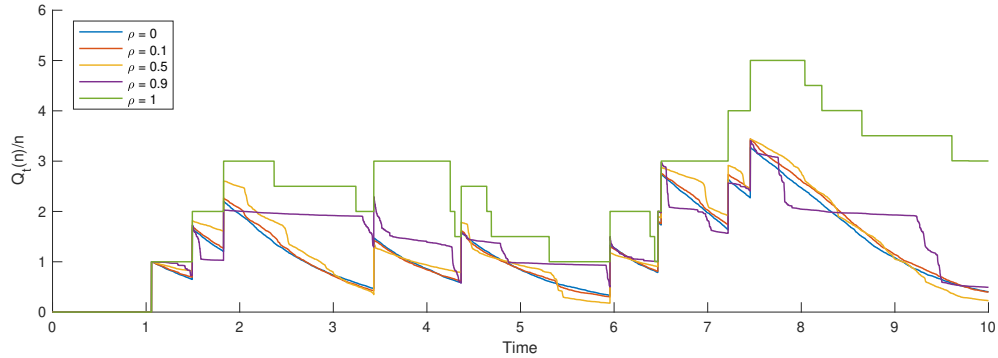
Therefore through these approximations, the standard deviation multiplier $\sqrt{\frac{n+1}{n}}$ gives us insight into how the staffing level of the queue converges to threshold level of the storage process and demonstrates that this normal approximation of the shot noise process should be expected to be smaller than the true batch arrival queue. However, in considering either of these derivations of normal-based approximations, it is important to recall our previous discussion that the rate that jobs arrive to the teleoperations system may not be significantly larger than the rate of service, thus implying that the needed central limit theorem conditions may not apply. In fact, one should expect for these conditions to be decreasingly applicable as driverless technology continues to improve and the rate of disengagements decreases. Given this disclaimer, we emphasize that we have included these Gaussian approximations as a guide for intuition, and they need not hold in practice.



(a)



(b)



(c)

Figure 6.10: A comparison of queue length sample paths with dependent service durations within each arriving batch for various dependence structures. In all experiments, the service distribution is unit rate exponential service and the batch sizes are deterministic, with $n = 1000$ and $c = 1.5$.

6.4.3 Exploration of Dependence within Batches

For a final numerical experiment, let us now explore dependence within batches of jobs. Because the simulated future scenarios are generated from a common

initial state in the look-ahead teleoperation, it is possible that the operators' response times could be correlated. As an empirical exploration of this, in Figure 6.10 we plot normalized queue length processes under three different dependency structures. In each setting, there is a probability $\rho \in [0, 1]$ that each successive service time will be dependent. In the first case, (a), each service time in a batch has probability ρ of being equal to the first duration within that batch and otherwise will be drawn independently, i.e. for an arbitrary batch i and $j \geq 2$,

$$S_{i,j} = \begin{cases} S_{i,1} & \text{with probability } \rho, \\ \tilde{S}_{i,j} & \text{otherwise,} \end{cases}$$

where $\tilde{S}_{i,j}$ is an independent draw from the service distribution. In case (b) this is instead equal to the previous time with probability ρ and independent otherwise, meaning

$$S_{i,j} = \begin{cases} S_{i,j-1} & \text{with probability } \rho, \\ \tilde{S}_{i,j} & \text{otherwise,} \end{cases}$$

and in (c) each time is an average over all previous service times within the batch with probability ρ and again otherwise independently drawn:

$$S_{i,j} = \begin{cases} \frac{1}{j-1} \sum_{k=1}^{j-1} S_{i,k} & \text{with probability } \rho, \\ \tilde{S}_{i,j} & \text{otherwise.} \end{cases}$$

In each of these settings, we plot simulated sample paths for $\rho \in \{0, 0.1, 0.5, 0.9, 1\}$ and we hold the arrival epochs fixed across all the experiments. When $\rho = 0$ all durations are independent regardless of the dependency setting, and thus these processes are effectively identical on this sample path due to the results in Theorem 6.2.2. Similarly, if $\rho = 1$ the service times within each batch

are perfectly correlated. Moreover, these processes are equivalently distributed across the dependency settings. In the case of infinitely many servers, the normalized queue length process can be trivially identified as a piecewise constant jump process (and in the case of deterministic batch sizes, this is an infinite server queue). However in the multi-server case the batch arrival queue is not as easily understood, and even more insight is lost in the intermediate settings of $\rho \in \{0.1, 0.5, 0.9\}$. The previous time dependency in case (b) shows a subdued system level dependency, as the difference in sample paths between $\rho = 0$ and $\rho \in \{0.1, 0.5, 0.9\}$ is not as pronounced as in cases (a) and (c). This illustrates that batch scaling limits subject to dependency within batches may merit its own future study. It is worth noting though that in the infinite server setting there are immediately available extensions of Theorem 6.2.1. For example, for each arriving batch in case (a), there is a binomially distributed number of jobs that are identical in duration, with the remaining jobs independently drawn. Under the batch scaling limit, this means that a ρ fraction of each jump will contribute to a piecewise constant jump process while the remaining $1 - \rho$ will function as part of a shot noice process. Moreover, because the limits in Theorem 6.2.1 make use of the law of large numbers, one could recover the infinite server batch scaling if the service durations are weakly dependent.

It is also worth noting that even though the jobs within look-ahead assistance are simulated from the same initial states, evidence from the pioneering implementations of this methodology suggests that job durations can vary significantly even on the same simulated future scenario. In fact, Lundgard et al. (2018) finds that the best of the instantaneously crowdsourced decisions often come from the operators who take the most time to respond. With this variance within batch in mind, perhaps it is more appropriate to consider a generalized

model that addresses the effect of the underlying initial state through service duration distributions that change across batches. In the infinite server model this can still be understood in the large batch setting, as this would correspond to a service distribution that changes with each arrival epoch, i.e. $\bar{G}_i(\cdot)$ for the ith arrival. In this way, Theorem 6.2.1 can be naturally extended. This generalization is more challenging to address in the multi-server setting, however, and thus this non-identically distributed model makes for an interesting and important direction of future study.

6.5 Conclusion, Discussion, and Future Work

In this chapter we have studied the staffing performance of a teleoperations system for autonomous vehicles. We model this modern service system as a queue with batch arrivals, as each disengagement of a vehicle's autonomous operation creates a batch of jobs through the pre-fetching look-ahead assistance method. Because this safety support methodology is based on instantaneously crowd-sourced input on simulated future scenarios, we have noted that larger batches actually imply a safer teleoperations system. Thus, we study this queueing system in the presence of large batch arrivals. Through our large batch analysis, we connect the queueing models to storage processes through novel scaling results. Via these storage processes, we are able to calculate the probability of the system exceeding capacity, which drives our staffing methodology. We are able to compute these quantities by leveraging the storage processes literature and introducing a technical lemma that connects sums of evaluations of the moment generating function to quantities such as the cumulative distribution function and the expectation of a random variable multiplied by an indicator function.

Through our numerical experiments, we have both validated these analytic results and uncovered interesting relationships. In particular, based on industry and government data we have found that teleoperations centers offer substantial staffing efficiencies, making them the potential key for achieving the dreams of level 5 automation at the more realistic level 4. More generally, we also have verified the intuition that the size of the batches of arrivals has more impact on a system's performance than the arrival rate does, a key takeaway for batch arrival service systems broadly.

We believe that there are a variety of opportunities for directly related future work. For example, we are also interested in studying networks of teleoperations centers that support different coverage regions, in which cars may frequently cross into different areas of support. Additionally, we would like to investigate how job and server types affect this teleoperations system, as in practice there may be both different classes of jobs and different skill sets of remote operators. In this case, we can draw upon queueing results such as in Gurvich and Whitt (2009, 2010); Adan et al. (2010); Adan and Weiss (2012). As another potential direction of work, we could seek to extend the storage processes literature that we have used here. For example, we are interested in using something such as a lack of bias assumption to extend some Poisson based results to non-Poisson settings Melamed and Whitt (1990). We also believe that there is opportunity for clever numerical implementations of the sums in Lemma C.1.1, as we discuss briefly in the appendix. We also intend to investigate different batch scalings of these models and we remain interested in extending these scalings to other related systems. Our insight on the strong effect of batches may also extend to systems with bursts of arrivals that occur in quick succession but are not quite simultaneous, as sufficiently rapid bursts may closely resemble batches.

This effect amounts to an important managerial insight, particularly when the arrival data is discretely observed and it is unclear whether bursts or batches occur. This is of great interest to us, and this relationship is another direction we intend to pursue.

As a closing comment, we note that this work is only among the first steps in planning and managing a driverless vehicle teleoperation system. Many important questions remain across a variety of different disciplines. For example, the look-ahead service requires sophisticated pairings of artificial intelligence and human expertise and this will necessitate careful study and attention to detail to be implemented at scale. Furthermore, this service system asks a great deal of its remote operators, and the profession of repeatedly performing high stakes driving tasks is certainly strenuous enough to prompt study of how to manage this cognitive load. Additionally, there are of course challenges in the design of the communication system supporting this center and there are questions on how to structure the market of the teleoperations services and regulate their operation. Moreover, wide-spread automation is poised to redefine many aspects of society, and it is important to ensure that people aren't displaced by these changes. Because of the variety of the relevant issues, these teleoperations centers pose questions that are not just important and intriguing, but also ones that are well suited to the breadth of the operations research community and we hope that this marks the beginning of wide study of teleoperations systems for autonomous vehicles in OR.

APPENDIX A

ADDENDUM FOR CHAPTER 4

A.1 Lemmas and Auxiliaries

In this section of the appendix we give technical lemmas to support our analysis and brief auxiliary results that are of interest but not within the narrative of the body of this report. We begin by giving the infinitesimal generator form for time derivatives of the expectations of functions of our process. This is a valuable tool available to us because the ESEP is Markov, and it supports much of our analysis throughout this work.

Lemma A.1.1. *For a sufficiently regular function $f : (\mathbb{R}^+ \times \mathbb{N}) \to \mathbb{R}$, the generator of the ESEP is given by*

$$\mathcal{L}f(\eta_t, N_t) = \underbrace{\sum_{i=1}^{n} \eta_t \left(f(\eta_t + \alpha, N_t + 1) - f(\eta_t, N_t)\right)}_{\text{Arrivals}} + \underbrace{\beta\left(\frac{\eta_t - \eta^*}{\alpha}\right)\left(f\left(\eta_t - \alpha, N_t\right) - f(\eta_t, N_t)\right)}_{\text{Expirations}}. \tag{A.1}$$

Then, the time derivative of the expectation of $f(\eta_t, N_t)$ is given by

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[f(\eta_t, N_t)\right] = \mathrm{E}\left[\mathcal{L}f(\eta_t, N_t)\right] \tag{A.2}$$

for all $t \geq 0$.

Proof. This is a direct result of the ESEP belonging to the family of piece-wise deterministic Markov processes, as defined in Davis (1984). Moreover, the specific regularity conditions are given in Theorem 5.5 of that work. □

Note that this is also immediately applicable to the active number in system process perspective of the ESEP, as $Q_t = (\eta_t - \eta^*)/\alpha$. Thus, we will leverage this

infinitesimal generator for studying each of η_t, Q_t, and N_t throughout both the main body of the text and these appendices.

As another supporting lemma, let us summarize a result that can be used with the infinitesimal generator to relate two different Markov processes. Throughout this work we make comparisons between different processes, in particular between the ESEP and the Hawkes process. One way that we do this is to investigate the differential equations found with use of Lemma A.1.1. In Lemma A.1.2 we provide the method by which we make such comparisons.

Lemma A.1.2 (A Comparison Lemma)**.** *Let $f : \mathbb{R}^2 \to \mathbb{R}$ be a continuous function in both variables. If we assume that initial value problem*

$$\frac{\mathrm{d}x(t)}{\mathrm{d}t} = f(t, x(t)), \; x(0) = x_0 \tag{A.3}$$

has a unique solution for the time interval $[0, T]$ and

$$\frac{\mathrm{d}y(t)}{\mathrm{d}t} \leq f(t, y(t)) \quad \text{for } t \in [0, T] \text{ and } y(0) \leq x_0 \tag{A.4}$$

then $x(t) \geq y(t)$ for all $t \in [0, T]$.

Proof. The the proof of this result is given in Hale and Lunel (2013). □

For a result that is both auxillary on the surface and beneficial in proofs, in Proposition A.1.3 we give the probability generating function for the number in system and the number of departures, or expirations, in the ESEP. The departure process is largely outside of the scope of this work, but this result is instrumental in the proof of the probability generating function for the counting process in Proposition 4.2.4, which is given in Appendix A.4.

Proposition A.1.3. *Let Q_t be the active number in system at time $t \geq 0$ of an ESEP with baseline intensity $\eta^* > 0$, intensity jump size $\alpha > 0$, and expiration rate $\beta > \alpha$.*

Then, let D_t be the number of arrivals by time t that are no longer active. Then, the joint probability generating function of Q_t and D_t, denoted $G(z_1, z_2, t) \equiv \mathrm{E}\left[z_1^{Q_t} z_2^{D_t}\right]$, is given by

$$G(z_1, z_2, t) = z_2^{D_0} e^{\frac{\eta^*(\beta-\alpha)}{2\alpha}t} \left(1 - \left(\tanh\left(\frac{t}{2}\sqrt{(\beta+\alpha)^2 - 4\alpha\beta z_2} + \tanh^{-1}\left(\frac{\beta+\alpha-2\alpha z_1}{\sqrt{(\beta+\alpha)^2 - 4\alpha\beta z_2}}\right)\right)\right)^2\right)^{\frac{\eta^*}{2\alpha}}$$
$$\cdot \left(\frac{\beta+\alpha}{2\alpha} - \frac{\sqrt{(\beta+\alpha)^2 - 4\alpha\beta z_2}}{2\alpha}\tanh\left(\frac{t}{2}\sqrt{(\beta+\alpha)^2 - 4\alpha\beta z_2} + \tanh^{-1}\left(\frac{\beta+\alpha-2\alpha z_1}{\sqrt{(\beta+\alpha)^2 - 4\alpha\beta z_2}}\right)\right)\right)^{Q_0}$$
$$\cdot \left(\cosh\left(\tanh^{-1}\left(\frac{2\alpha z_1 - \beta - \alpha}{\sqrt{(\beta+\alpha)^2 - 4\alpha\beta z_2}}\right)\right)\right)^{\frac{\eta^*}{\alpha}}, \tag{A.5}$$

where Q_0 and D_0 are the active number in the system and the count of departures at time 0, respectively.

Proof. We will show this through the method of characteristics. We can first observe through Lemma A.1.1 that

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[z_1^{Q_t} z_2^{D_t}\right] = \mathrm{E}\left[(\eta^* + \alpha Q_t)(z_1 - 1) z_1^{Q_t} z_2^{D_t} + \beta Q_t \left(\frac{z_2}{z_1} - 1\right) z_1^{Q_t} z_2^{D_t}\right],$$

and so $G(z_1, z_2, t)$ is given by the following partial differential equation:

$$\frac{\partial}{\partial t}G(z_1, z_2, t) + \left(\alpha(z_1 - z_1^2) + \beta(z_1 - z_2)\right)\frac{\partial}{\partial z_1}G(z_1, z_2, t) = \eta^*(z_1 - 1)G(z_1, z_2, t).$$

To simplify our analysis, we will instead solve for $\log(G(z_1, z_2, t))$, which through the chain rule will by given by the solution to the partial differential equation expressed

$$\frac{\partial}{\partial t}\log(G(z_1, z_2, t)) + \left(\alpha(z_1 - z_1^2) + \beta(z_1 - z_2)\right)\frac{\partial}{\partial z_1}\log(G(z_1, z_2, t)) = \eta^*(z_1 - 1),$$

with initial condition $\log(G(z_1, z_2, 0)) = \log(z_1^{Q_0} z_2^{D_0})$. This now gives us the charac-

teristic equations as follows:

$$\begin{aligned}
\frac{\mathrm{d}z_1}{\mathrm{d}s}(r,s) &= \alpha(z_1 - z_1^2) + \beta(z_1 - z_2), && z_1(r,0) = r \\
\frac{\mathrm{d}t}{\mathrm{d}s}(r,s) &= 1, && t(r,0) = 0 \\
\frac{\mathrm{d}g}{\mathrm{d}s}(r,s) &= \eta^*(z_1 - 1), && g(r,0) = \log(r^{Q_0} z_2^{D_0}).
\end{aligned}$$

Solving the first two equations we see that

$$z_1(r,s) = \frac{\beta+\alpha}{2\alpha} + \frac{\sqrt{(\beta+\alpha)^2 - 4\alpha\beta z_2}}{2\alpha} \tanh\left(\frac{s}{2}\sqrt{(\beta+\alpha)^2 - 4\alpha\beta z_2} - \tanh^{-1}\left(\frac{\beta+\alpha-2\alpha r}{\sqrt{(\beta+\alpha)^2 - 4\alpha\beta z_2}}\right)\right)$$

$$t(r,s) = s,$$

which allows us to now solve for $g(r,s)$. Using the solution to $z_1(r,s)$, the ordinary differential equation for $g(r,s)$ is given by

$$\begin{aligned}
\frac{\mathrm{d}g}{\mathrm{d}s}(r,s) = {} & \frac{\eta^*\sqrt{(\beta+\alpha)^2 - 4\alpha\beta z_2}}{2\alpha} \tanh\left(\frac{s}{2}\sqrt{(\beta+\alpha)^2 - 4\alpha\beta z_2} - \tanh^{-1}\left(\frac{\beta+\alpha-2\alpha r}{\sqrt{(\beta+\alpha)^2 - 4\alpha\beta z_2}}\right)\right) \\
& + \frac{\eta^*(\beta-\alpha)}{2\alpha},
\end{aligned}$$

which yields a solution of

$$\begin{aligned}
g(r,s) = {} & \log(r^{Q_0} z_2^{D_0}) + \frac{\eta^*(\beta-\alpha)}{2\alpha} s + \frac{\eta^*}{2\alpha}\log\left(1 - \frac{(\beta+\alpha-2\alpha r)^2}{(\beta+\alpha)^2 - 4\alpha\beta z_2}\right) \\
& + \frac{\eta^*}{\alpha}\log\left(\cosh\left(\frac{s}{2}\sqrt{(\beta+\alpha)^2 - 4\alpha\beta z_2} - \tanh^{-1}\left(\frac{\beta+\alpha-2\alpha r}{\sqrt{(\beta+\alpha)^2 - 4\alpha\beta z_2}}\right)\right)\right).
\end{aligned}$$

Now, from these equations we can express the characteristics variables in terms of the original arguments as $s = t$ and

$$r = \frac{\beta+\alpha}{2\alpha} - \frac{\sqrt{(\beta+\alpha)^2 - 4\alpha\beta z_2}}{2\alpha} \tanh\left(\frac{t}{2}\sqrt{(\beta+\alpha)^2 - 4\alpha\beta z_2} - \tanh^{-1}\left(\frac{2\alpha z_1 - \beta - \alpha}{\sqrt{(\beta+\alpha)^2 - 4\alpha\beta z_2}}\right)\right).$$

Then, by performing the substitution $G(z_1, z_2, t) = e^{g(r(z_1,z_2,t), s(z_1,z_2,t))}$ and simplifying, we achieve the stated result. □

As another auxiliary result, in Proposition A.1.4 we give the steady-state moment generating function for the 2-GESEP with exponentially distribution activity durations and deterministic batch sizes, meaning pairs of arrivals.

Proposition A.1.4. *Consider the following 2-GESEP: arrivals occur at rate $\eta_t(2) = \eta^* + \frac{\alpha}{2} Q_t(2)$, where $Q_t(2)$ receives arrivals batches of size 2. Each activity duration is independent and exponentially distributed with rate $\beta > \alpha > 0$. Then, the steady-state moment generating function of $Q_t(2)$ is given by*

$$\begin{aligned}
&\mathrm{E}\left[e^{\theta Q_\infty(2)}\right] \equiv \lim_{t\to\infty} \mathrm{E}\left[e^{\theta Q_t(2)}\right] \\
&= \exp\left(\frac{2\eta^*}{\sqrt{\alpha(\alpha+8\beta)}}\left(\tanh^{-1}\left(\left(2e^\theta+1\right)\sqrt{\frac{\alpha}{\alpha+8\beta}}\right) - \tanh^{-1}\left(3\sqrt{\frac{\alpha}{\alpha+8\beta}}\right)\right)\right)\left(\frac{2\beta-2\alpha}{2\beta-\alpha(e^\theta+e^{2\theta})}\right)^{\frac{\eta^*}{\alpha}}
\end{aligned} \tag{A.6}$$

Proof. Using Lemma A.1.1, we see that the moment generating function will be given by the solution to

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[e^{\theta Q_t(2)}\right] = \mathrm{E}\left[\left(\eta^* + \frac{\alpha Q_t(2)}{2}\right)\left(e^{\theta(Q_t(2)+2)} - e^{\theta Q_t(2)}\right) + \beta Q_t(2)\left(e^{\theta(Q_t(2)-1)} - e^{\theta Q_t(2)}\right)\right],$$

which can be equivalently expressed in PDE form as

$$\frac{\partial}{\partial t}\mathcal{M}_2(\theta,t) = \eta^*\left(e^{2\theta}-1\right)\mathcal{M}_2(\theta,t) + \left(\frac{\alpha}{2}\left(e^{2\theta}-1\right) + \beta\left(e^{-\theta}-1\right)\right)\frac{\partial}{\partial\theta}\mathcal{M}_2(\theta,t),$$

where $\mathcal{M}_2(\theta,t) = \mathrm{E}\left[e^{\theta Q_t(2)}\right]$. To solve for the steady-state moment generating function we consider the ODE given by

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\mathcal{M}_2(\theta,\infty) = \frac{\eta^*\left(1-e^{2\theta}\right)\mathcal{M}_2(\theta,\infty)}{\frac{\alpha}{2}\left(e^{2\theta}-1\right)+\beta\left(e^{-\theta}-1\right)},$$

with the initial condition that $\mathcal{M}_2(0,\infty) = 1$. Through taking the derivative of the expression in Equation A.6, we verify the result. □

A.2 Exploring a Hybrid Self-Exciting Model

In the main body of the text we have defined the ESEP, a model that features arrivals that self-excite but only for a finite period of time. By comparison to

the traditional Hawkes process models for self-excitement, the effect from one arrival does not decay through time but is fixed at a constant value for as long as it remains active. In this way, the ESEP features ephemeral but piecewise constant self-excitement whereas the Hawkes process has eternal but ever decreasing self-excitement. One can note though that ephemeral self-excitement need not be piecewise constant. A model could feature both decay and down-jumps as manners of regulating its self-excitement. In this section of the appendix, we will consider such a model, specifically a Markovian one. To begin, let us now define the *hybrid ephemerally self-exciting process* (HESEP) in Defintion A.2.1.

Definition A.2.1 (Hybrid ephemerally self-exciting process)**.** Let $t \geq 0$ and suppose that $\nu^* > 0$, $\alpha > 0$, $\beta \geq 0$, and $\mu \geq 0$. Then, define ν_t, $N_{t,\nu}$, and $Q_{t,\nu}$ such that:

i) $N_{t,\nu}$ is an arrival process driven by the intensity ν_t,

ii) $Q_{t,\nu}$ is the number of arrivals from $N_{t,\nu}$ that have not yet expired according to their i.i.d. Exp(μ) activity durations,

iii) ν_t is governed by

$$\mathrm{d}\nu_t = \beta(\nu^* - \nu_t)\mathrm{d}t + \alpha \mathrm{d}N_{t,\nu} - \frac{\nu_t - \nu^*}{Q_{t,\nu}}\mathrm{d}D_{t,\nu}$$

where $D_{t,\nu} = N_{t,\nu} - Q_{t,\nu}$.

Then, we say that the intensity-queue-counting process triplet $(\nu_t, Q_{t,\nu}, N_{t,\nu})$ is a *hybrid ephemerally self-exciting process* (HESEP) with baseline intensity ν^*, intensity jump size α, decay rate β, expiration rate μ, and initial values $(\nu_0, Q_0^\nu, N_0^\nu)$.

By definition, one can view the HESEP as a hybrid between the ESEP and Hawkes process models. If $\beta = 0$ then we recover the ESEP; if $\mu = 0$ then we recover the Hawkes process. In this way, much of the dynamics are quite familiar:

up-jumps of size α at teach arrival, exponential decay between events at rate β, and down-jumps upon each activity duration expiration. Perhaps the least intuitive part of this definition is the size of the down-jump, as this depends on the current levels of the intensity and the number of active exciters in the system. This draws inspiration from Markovian infinite server queues. In an $M/M/\infty$, all jobs currently in the system are equally likely to be the next to leave, regardless of the order they entered the system. Similarly, in the HESEP, each exciter in the system is equally likely to be the next to leave. Moreover, the down-jump size is the same regardless of which exciter is next to leave. When an expiration of an activity durations means that there are no longer any presently active exciters, by definition the intensity will return to the baseline value. One can note that this down-jump size is actually always bounded below by 0 since the intensity decays down towards the baseline ν^* and bounded above by α since $\nu_t - \nu^* \leq \alpha Q_{t,\nu}$ due to the fact that each arrival increases ν_t by α before it decays. As a quick interesting fact regarding this process, in Proposition A.2.1 we show that the size of a downjump, $(\nu_t - \nu^*)/Q_{t,\nu}$, does not have down-jumps itself.

Proposition A.2.1. *Let $\phi_t = \frac{\nu_t - \nu^*}{Q_{t,\nu}}$ be the size of a down-jump occurring at time $t \geq 0$. Suppose that $b \geq a \geq 0$ is such that $Q_{t,\nu}$ is positive for all $t \in [a, b]$. Then, the ϕ_t has no downward jumps on $[a, b]$.*

Proof. Suppose that $[a, b]$ is such as interval, and then for $t \in [a, b]$ we note that

$$\frac{\nu_t - \frac{\nu_t - \nu^*}{Q_{t,\nu}} - \nu^*}{Q_{t,\nu} - 1} = \frac{Q_t \nu_t - \nu_t + \nu^* - Q_{t,\nu}\nu^*}{Q_{t,\nu}(Q_{t,\nu} - 1)} = \frac{\nu_t - \nu^*}{Q_{t,\nu}},$$

and this is equal to ϕ_t. □

A.2.1 Sandwiching the Hybrid Model

As one might expect for a so-called "hybrid" model, the HESEP can be connected to the ESEP and the Hawkes process in many different ways. In Proposition A.2.2, we show that one can actual sandwich this model between its two extremes. That is, we show that the means of the processes are equal when given the same parameters, whereas the variances are ordered with Hawkes as the smallest and ESEP as the largest.

Proposition A.2.2. *Let $\alpha > 0$, $\beta > 0$, and $\mu > 0$ be such that $\mu + \beta > \alpha > 0$. Additionally, let $\nu^* > 0$. Let ν_t be an HESEP with baseline intensity ν^*, intensity jump size α, decay rate β, and service rate μ. Similarly, let λ_t be the intensity of a Hawkes process with baseline intensity ν^*, intensity jump α, and decay rate $\mu + \beta$. Finally, let η_t be the intensity of an ESEP with baseline intensity ν^*, intensity jump α, and service rate $\mu + \beta$, then the means of these process intensities are all equal:*

$$\mathrm{E}\left[\lambda_t\right] = \mathrm{E}\left[\nu_t\right] = \mathrm{E}\left[\eta_t\right]. \tag{A.7}$$

Furthermore, the process variances are ordered such that

$$\mathrm{Var}\left(\lambda_t\right) \leq \mathrm{Var}\left(\nu_t\right) \leq \mathrm{Var}\left(\eta_t\right). \tag{A.8}$$

Additionally, let $N_{t,\nu}$, $N_{t,\lambda}$, and $N_{t,\eta}$ be the counting processes of the HESEP, Hawkes process, and ESEP, respectively. Then, the means of these counting process are equal

$$\mathrm{E}\left[N_{t,\lambda}\right] = \mathrm{E}\left[N_{t,\lambda}\right] = \mathrm{E}\left[N_{t,\eta}\right], \tag{A.9}$$

and the variances of these counting processes are again ordered such that

$$\mathrm{Var}\left(N_{t,\lambda}\right) \leq \mathrm{Var}\left(N_{t,\nu}\right) \leq \mathrm{Var}\left(N_{t,\eta}\right). \tag{A.10}$$

Finally, the covariances among each intensity and counting process pair are likewise ordered such that

$$\mathrm{Cov}\left[\lambda_t, N_{t,\lambda}\right] \leq \mathrm{Cov}\left[\nu_t, N_{t,\nu}\right] \leq \mathrm{Cov}\left[\eta_t, N_{t,\eta}\right], \tag{A.11}$$

where $t \geq 0$ and where all intensities have the same initial value.

Proof. By a quick check of the differential equations for each mean, we can directly observe that $\mathrm{E}\left[\nu_t\right] = \mathrm{E}\left[\eta_t\right] = \mathrm{E}\left[\lambda_t\right]$. To show the variance ordering we begin by considering the ODE for the second moment of ν_t:

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[\nu_t^2\right] = 2\beta\left(\nu^*\mathrm{E}\left[\nu_t\right] - \mathrm{E}\left[\nu_t^2\right]\right) + \alpha^2\mathrm{E}\left[\nu_t\right] + 2\alpha\mathrm{E}\left[\nu_t^2\right] + \mu\mathrm{E}\left[\left(\left(\nu_t - \frac{\nu_t - \nu^*}{Q_{t,\nu}}\right)^2 - \nu_t^2\right)Q_{t,\nu}\right].$$

Now, let's observe that

$$\mathrm{E}\left[\left(\left(\nu_t - \frac{\nu_t - \nu^*}{Q_{t,\nu}}\right)^2 - \nu_t^2\right)Q_{t,\nu}\right] = 2\left(\nu^*\mathrm{E}\left[\nu_t\right] - \mathrm{E}\left[\nu_t^2\right]\right) + \mathrm{E}\left[\frac{(\nu_t - \nu^*)^2}{Q_{t,\nu}}\right],$$

which follows by expanding the squared term. Because $\frac{(\nu_t-\nu^*)^2}{Q_{t,\nu}} \geq 0$, this gives us that

$$\mathrm{E}\left[\left(\left(\nu_t - \frac{\nu_t - \nu^*}{Q_{t,\nu}}\right)^2 - \nu_t^2\right)Q_{t,\nu}\right] \geq 2\left(\nu^*\mathrm{E}\left[\nu_t\right] - \mathrm{E}\left[\nu_t^2\right]\right). \tag{A.12}$$

This inequality now allows us to directly compare $\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[\nu_t^2\right]$ to $\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[\lambda_t^2\right]$ and $\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[\eta_t^2\right]$. First, we can use Equation A.12 to see that

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[\nu_t^2\right] &= 2\beta\left(\nu^*\mathrm{E}\left[\nu_t\right] - \mathrm{E}\left[\nu_t^2\right]\right) + \alpha^2\mathrm{E}\left[\nu_t\right] + 2\alpha\mathrm{E}\left[\nu_t^2\right] + \mu\mathrm{E}\left[\left(\left(\nu_t - \frac{\nu_t - \nu^*}{Q_{t,\nu}}\right)^2 - \nu_t^2\right)Q_{t,\nu}\right] \\ &\geq 2(\mu+\beta)\left(\nu^*\mathrm{E}\left[\nu_t\right] - \mathrm{E}\left[\nu_t^2\right]\right) + \alpha^2\mathrm{E}\left[\nu_t\right] + 2\alpha\mathrm{E}\left[\nu_t^2\right].\end{aligned}$$

Because we have already shown that $\mathrm{E}\left[\lambda_t\right] = \mathrm{E}\left[\nu_t\right]$, we see that $\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[\lambda_t^2\right] \leq \frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[\nu_t^2\right]$ when evaluated at the same point and thus by Lemma A.1.2, $\mathrm{Var}\left(\lambda_t\right) \leq \mathrm{Var}\left(\nu_t\right)$. By analogous arguments for η_t, we achieve the stated result. For the counting

process means, we can now observe that all the differential equations are such that

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[N_{t,\lambda}\right] = \mathrm{E}\left[\lambda_t\right] = \frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[N_{t,\nu}\right] = \mathrm{E}\left[\nu_t\right] = \frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[N_{t,\eta}\right] = \mathrm{E}\left[\eta_t\right].$$

We assume that all counting processes start at 0 and thus we have that the counting process means are equal throughout time. This also implies that the products of means, $\mathrm{E}\left[\lambda_t\right]\mathrm{E}\left[N_{t,\lambda}\right]$, $\mathrm{E}\left[\nu_t\right]\mathrm{E}\left[N_{t,\nu}\right]$, and $\mathrm{E}\left[\eta_t\right]\mathrm{E}\left[N_{t,\eta}\right]$, are equal. Hence to show the ordering of the covariances we will focus solely on the expectations of the products. This differential equation is given by

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[\nu_t N_{t,\nu}\right] = -(\mu+\beta-\alpha)\mathrm{E}\left[\nu_t N_{t,\nu}\right] + (\mu+\beta)\nu^*\mathrm{E}\left[N_{t,\nu}\right] + \alpha\mathrm{E}\left[\nu_t\right] + \mathrm{E}\left[\nu_t^2\right],$$

and we can note that the coefficients are the same for each of the processes. Not including the function for which we want to solve, $\mathrm{E}\left[\nu_t N_{t,\nu}\right]$, we can also observe that every function is equivalent across the processes other than the second moments of the intensities. We have shown that these second moments are in fact ordered and therefore we receive the stated ordering of the covariances. Finally, we observe that the differential equation for the second moment of each counting process is of the form

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[(N_{t,\nu})^2\right] = \mathrm{E}\left[\nu_t\right] + 2\mathrm{E}\left[\nu_t N_{t,\nu}\right].$$

From the ordering of the covariances and the equivalences of the means, we can conclude the proof. □

As a simple consequence of Propositon A.2.2, we can note that because the Hawkes process is over-dispersed, i.e. its variance is larger than its mean, so too are the ESEP and HESEP. One can note that these bounds on the variance of the HESEP are not only useful for comparison but also practical for the study of the HESEP itself, as the differential equations for the variance via the infinitesimal generator is not closed.

A.2.2 Strong Convergence of the HESEP Counting Process

In the final three subsections of analysis of the HESEP, we obtain a trio of limiting results. We begin with the almost sure convergence of the ratio of the HESEP counting process and time, which is an elementary renewal result in the style of Blackwell (1948) or Lindvall (1977), for example. However, by comparison to the context of such works, we can bound the mean and variance of the HESEP via the ESEP and we are instead solely interested in establishing the convergence, as we will obtain additional results by consequence. Using these expressions for the first two moments, the proof of Theorem A.2.3 follows standard approaches using the Borel-Cantelli lemma. In Corollary A.2.4 we use this renewal result to find a strong law of large numbers for the dependent and non-identically distributed inter-arrival times of the HESEP by way of the continuous mapping theorem, which is another standard technique.

Theorem A.2.3. *Let $(\nu_t, Q_{t,\nu}, N_{t,\nu})$ be a HESEP with baseline intensity ν^*, intensity jump $\alpha > 0$, intensity decay rate $\beta \geq 0$, and rate of exponentially distributed service $\mu \geq 0$, where $\mu + \beta > \alpha$. Then,*

$$\frac{N_{t,\nu}}{t} \xrightarrow{\text{a.s.}} \nu_\infty \tag{A.13}$$

as $t \to \infty$, where $\nu_\infty = \frac{(\mu+\beta)\nu^*}{\mu+\beta-\alpha}$.

Proof. We will show this through use of the Borel-Cantelli Lemma. Let $\epsilon > 0$ be arbitrary and define the event E_s for $s \in \mathbb{N}$ as

$$E_s = \left\{ \sup_{t \in (s^2,(s+1)^2]} \frac{|N_{t,\nu} - \mathrm{E}\left[N_{t,\nu}\right]|}{t} > \epsilon \right\}.$$

We now note that $N_{t,\nu} - \mathrm{E}\left[N_{t,\nu}\right]$ is a martingale by definition, and so $|N_{t,\nu} - \mathrm{E}\left[N_{t,\nu}\right]|$

is a sub-martingale. Additionally, we can observe that

$$\mathrm{P}(E_s) \leq \mathrm{P}\left(\sup_{t\in(s^2,(s+1)^2]} |N_{t,\nu} - \mathrm{E}[N_{t,\nu}]| > s^2\epsilon\right)$$

because $s^2 \leq t$ for any t. By Doob's martingale inequality, we have

$$\mathrm{P}\left(\sup_{t\in(s^2,(s+1)^2]} |N_{t,\nu} - \mathrm{E}[N_{t,\nu}]| > s^2\epsilon\right) \leq \frac{\mathrm{E}\left[|N_{(s+1)^2,\nu} - \mathrm{E}\left[N_{(s+1)^2,\nu}\right]|^2\right]}{s^4\epsilon^2} = \frac{\mathrm{Var}\left(N_{(s+1)^2,\nu}\right)}{s^4\epsilon^2}.$$

From Proposition A.2.2, we note that the variance of an HESEP counting process with baseline intensity ν^*, intensity jump size α, decay rate β, and service rate μ is upper-bounded by the variance of an ESEP counting process with baseline ν^*, jump size α, and service rate $\mu+\beta$. Using the explicit form of the ESEP counting process variance as computed through Lemma A.1.1, we have the bound

$$\begin{aligned}\mathrm{Var}\left(N_{(s+1)^2,\nu}\right) \leq \mathrm{Var}\left(N_{(s+1)^2,\eta}\right) &= \frac{((\mu+\beta)^2+\alpha^2)\nu_\infty}{(\mu+\beta-\alpha)^2}(s+1)^2 - \frac{2\alpha\mu(\nu_0-\nu_\infty)}{(\mu+\beta-\alpha)^3}\Big(e^{-(\mu+\beta-\alpha)(s+1)^2} \\ &+ (\mu+\beta-\alpha)(s+1)^2 e^{-(\mu+\beta-\alpha)(s+1)^2}\Big) + \left(\frac{\nu_0-\nu_\infty}{\mu+\beta-\alpha} - \frac{\alpha\mu\nu_\infty}{(\mu+\beta-\alpha)^3} - \frac{(\alpha^2+\alpha(\mu+\beta))\nu_0}{(\mu+\beta-\alpha)^3}\right) \\ &\cdot\left(1-e^{-(\mu+\beta-\alpha)(s+1)^2}\right) + \left(\frac{(\alpha^2+\alpha(\mu+\beta))\nu_0}{2(\mu+\beta-\alpha)^3} - \frac{\alpha(\mu+\beta)\nu_\infty}{2(\mu+\beta-\alpha)^3}\right)\left(1-e^{-2(\mu+\beta-\alpha)(s+1)^2}\right).\end{aligned}$$

Together, this implies that $\mathrm{P}(E_s) \in O\left(\frac{1}{s^2}\right)$. Therefore $\sum_{s=0}^{\infty}\mathrm{P}(E_s) < \infty$, and so by the Borel-Cantelli Lemma, $\frac{|N_{t,\nu}-\mathrm{E}[N_{t,\nu}]|}{t} \xrightarrow{\text{a.s.}} 0$. Since $\lim_{t\to\infty}\frac{\mathrm{E}[N_{t,\nu}]}{t} = \nu_\infty$, we complete the proof. □

As an immediate consequence of this, we achieve a law of large numbers for the dependent inter-arrival times.

Corollary A.2.4. *Let $(\nu_t, Q_{t,\nu}, N_{t,\nu})$ be an HESEP counting process with baseline intensity $\nu^* > 0$, intensity jump $\alpha > 0$, intensity decay rate $\beta \geq 0$, and rate of exponentially distributed service $\mu \geq 0$, where $\mu+\beta > \alpha$. Further, let S_k^ν denote the kth inter-arrival time for $k \in \mathbb{Z}^+$. Then,*

$$\frac{1}{n}\sum_{k=1}^{n} S_k^\nu \xrightarrow{\text{a.s.}} \frac{1}{\nu_\infty} \tag{A.14}$$

as $n \to \infty$, where $\nu_\infty = \frac{(\mu+\beta)\nu^*}{\mu+\beta-\alpha}$.

Proof. Let A_n^ν denote the time of the nth arrival for each $n \in \mathbb{Z}^+$, which is to say that $A_n^\nu = \sum_{k=1}^n S_k^\nu$. Now, observe that the time of the most recent arrival up to time t, $A_{N_{t,\nu}}^\nu$, can be bounded as

$$t - S_{N_{t,\nu}+1}^\nu \leq A_{N_{t,\nu}}^\nu \leq t,$$

since if $t - S_{N_{t,\nu}+1}^\nu > A_{N_{t,\nu}}^\nu$, then arrival $N_{t,\nu} + 1$ would have occurred before time t. Now, we also note that because $\nu^* > 0$ then $N_{t,\nu} \to \infty$ as $t \to \infty$ and this implies that

$$\frac{S_{N_{t,\nu}+1}}{N_{t,\nu}} \xrightarrow{\text{a.s.}} 0$$

as $t \to \infty$. From Proposition A.2.3 and the continuous mapping theorem, we know that $\frac{t}{N_{t,\nu}} \to \frac{1}{\nu_\infty}$ and $\frac{t - S_{N_{t,\nu}+1}}{N_{t,\nu}} \to \frac{1}{\nu_\infty}$ almost surely. By the sandwiching $A_{N_{t,\nu}}$, this yields the stated result. □

Because the Hawkes process and the ESEP are special cases of this hybrid model, we can note that both the renewal result and the law of large numbers apply directly to each.

Corollary A.2.5. *Let $(\lambda_t, N_{t,\lambda})$ be the intensity and count of a Hawkes process with baseline intensity $\lambda^* > 0$, intensity jump $\alpha > 0$, and decay rate $\beta > \alpha$. Similarly, let $(\eta_t, N_{t,\eta})$ be the intensity and counting process pair for an ESEP with baseline intensity $\nu^* > 0$, intensity jump $\alpha > 0$, and rate of exponentially distributed service $\mu > \alpha$. Then, for S_k^λ and S_k^η as the kth inter-arrival times for the Hawkes process and the ESEP process, respectively, we have that*

$$\frac{N_{t,\lambda}}{t} \xrightarrow{\text{a.s.}} \lambda_\infty, \qquad \frac{N_{t,\eta}}{t} \xrightarrow{\text{a.s.}} \eta_\infty, \tag{A.15}$$

and

$$\frac{1}{n}\sum_{k=1}^n S_k^\lambda \xrightarrow{\text{a.s.}} \frac{1}{\lambda_\infty}, \qquad \frac{1}{n}\sum_{k=1}^n S_k^\eta \xrightarrow{\text{a.s.}} \frac{1}{\eta_\infty}, \tag{A.16}$$

where $\lambda_\infty = \frac{\beta\lambda^*}{\beta-\alpha}$ and $\eta_\infty = \frac{\mu\nu^*}{\mu-\alpha}$.

A.2.3 Baseline Fluid Limit of the HESEP

In this subsection and in the sequel, we consider a baseline scaling of the HESEP. That is, we investigate limiting properties of the process as the baseline intensity grows large and the intensity and queue length are normalized in some fashion. To begin, we take the normalization as directly proportional to the baseline scaling, which is the fluid limit. The derivation of this is empowered by the following lemma, which allows us to make use of Taylor expansions.

Lemma A.2.6. *Suppose that for some $b > 0$, $-b \leq z_n(t) \leq 0$ for all values of n. Then there exists constants C_1 and C_2 where $C_1 \leq C_2$, which imply the following bounds for sufficiently large values of n*

$$z_n(t) + \frac{C_1}{n} \leq n \cdot \left(e^{\frac{z_n(t)}{n}} - 1\right) \leq z_n(t) + \frac{C_2}{n}. \tag{A.17}$$

Proof. The proof follows by performing a second order Taylor expansion for the exponential function and observing that since $z_n(t)$ lies in a compact interval, we can construct uniform lower and upper bounds for the exponential function. □

With this lemma in hand, we now proceed to finding the fluid limit in Theorem A.2.7. In this case, we scale the baseline intensity by n, whereas we scale the intensity and the queue length by $\frac{1}{n}$. As one would expect to see, we find that the fluid limit converges to the means of the intensity and queue.

Theorem A.2.7. *For $n \in \mathbb{Z}$, let the nth fluid-scaled HESEP $(\nu_t(n), Q_{t,\nu}(n))$ be defined such that the baseline intensity is $n\nu^*$, the intensity jump size is $\alpha > 0$, the intensity decay rate is $\beta \geq 0$, and the rate of exponentially distributed service is $\mu > 0$, where $\mu + \beta > \alpha$. Then, for the scaled quantities $(\frac{\nu_{t,\nu}(n)}{n}, \frac{Q_{t,\nu}(n)}{n})$, the limit of the moment generating*

function

$$\tilde{\mathcal{M}}^{\infty}(t,\theta_\nu,\theta_Q) \equiv \lim_{n\to\infty} \mathrm{E}\left[e^{\frac{\theta_\nu}{n}\nu_t(n)+\frac{\theta_Q}{n}Q_{t,\nu}(n)}\right], \tag{A.18}$$

is given by

$$\tilde{\mathcal{M}}^{\infty}(t,\theta_\nu,\theta_Q) = e^{\theta_\nu \mathrm{E}[\nu_t]+\theta_Q \mathrm{E}\left[Q_{t,\nu}\right]}, \tag{A.19}$$

for all $t \geq 0$.

Proof. The proof will follow in two steps. The first step is to show that the limiting moment generating function converges to a PDE given by $\tilde{\mathcal{M}}^{\infty}$ using properties of the exponential function and Lemma A.2.6. The second step is to solve this PDE using the method of characteristics. Finally, by the uniqueness of moment generating functions, we can assert that the random variables to which our limit converges are deterministic functions of time, which are also known as the fluid limit. We begin with the infinitesimal generator form which simplifies through the linearity of expectation as

$$\begin{aligned}
\frac{\partial}{\partial t}\tilde{\mathcal{M}}^{n}(t,\theta_\nu,\theta_Q) &\equiv \frac{\partial}{\partial t}\mathrm{E}\left[e^{\frac{\theta_\nu}{n}\nu_t(n)+\frac{\theta_Q}{n}Q_{t,\nu}(n)}\right] \\
&= \mathrm{E}\left[\beta(\nu^* n-\nu_t(n))\frac{\theta_\nu}{n}e^{\frac{\theta_\nu}{n}\nu_t(n)+\frac{\theta_Q}{n}Q_{t,\nu}(n)}\right] + \mathrm{E}\left[\nu_t(n)\left(e^{\frac{\alpha\theta_\nu}{n}+\frac{\theta_Q}{n}}-1\right)e^{\frac{\theta_\nu}{n}\nu_t(n)+\frac{\theta_Q}{n}Q_{t,\nu}(n)}\right] \\
&\quad + \mathrm{E}\left[\mu Q_{t,\nu}(n)\left(e^{-\frac{\theta_\nu(\nu_t(n)-\nu^* n)}{nQ_{t,\nu}(n)}-\frac{\theta_Q}{n}}-1\right)e^{\frac{\theta_\nu}{n}\nu_t(n)+\frac{\theta_Q}{n}Q_{t,\nu}(n)}\right] \\
&= \beta\nu^*\theta_\nu \mathrm{E}\left[e^{\frac{\theta_\nu}{n}\nu_t(n)+\frac{\theta_Q}{n}Q_t(n)}\right] - \beta\theta_\nu \mathrm{E}\left[\frac{\nu_t(n)}{n}e^{\frac{\theta_\nu}{n}\nu_t(n)+\frac{\theta_Q}{n}Q_{t,\nu}(n)}\right] \\
&\quad + n\left(e^{\frac{\alpha\theta_\nu}{n}+\frac{\theta_Q}{n}}-1\right)\mathrm{E}\left[\frac{\nu_t(n)}{n}e^{\frac{\theta_\nu}{n}\nu_t(n)+\frac{\theta_Q}{n}Q_t(n)}\right] \\
&\quad + \frac{\mu}{n}\mathrm{E}\left[Q_{t,\nu}(n)n\left(e^{-\frac{\theta_\nu(\nu_t(n)-\nu^* n)}{nQ_{t,\nu}(n)}-\frac{\theta_Q}{n}}-1\right)e^{\frac{\theta_\nu}{n}\nu_t(n)+\frac{\theta_Q}{n}Q_{t,\nu}(n)}\right] \\
&= \beta\nu^*\theta_\nu\tilde{\mathcal{M}}(t,\theta_\nu,\theta_Q) + \left(n\left(e^{\frac{\alpha\theta_\nu}{n}+\frac{\theta_Q}{n}}-1\right)-\beta\theta_\nu\right)\frac{\partial}{\partial\theta_\nu}\tilde{\mathcal{M}}(t,\theta_\nu,\theta_Q) \\
&\quad + \frac{\mu}{n}\mathrm{E}\left[Q_{t,\nu}(n)\left(-\frac{\theta_\nu(\nu_t(n)-\nu^* n)}{Q_{t,\nu}(n)}-\theta_Q+\frac{\epsilon_n}{n}\right)e^{\frac{\theta_\nu}{n}\nu_t(n)+\frac{\theta_Q}{n}Q_{t,\nu}(n)}\right],
\end{aligned}$$

where the last equality holds for sufficiently large n, where ϵ_n is in some bounded interval as according to Lemma A.2.6. Then, by rearranging further we can see that in limit this becomes

$$
\begin{aligned}
&\frac{\partial}{\partial t}\tilde{\mathcal{M}}^n(t,\theta_\nu,\theta_Q)\\
&= \beta\nu^*\theta_\nu M^n(t,\theta_\nu,\theta_Q) + \left(n\left(e^{\frac{\alpha\theta_\nu}{n}+\frac{\theta_Q}{n}}-1\right)-\beta\theta_\nu\right)\frac{\partial}{\partial\theta_\nu}\tilde{\mathcal{M}}^n(t,\theta_\nu,\theta_Q) - \mu\theta_\nu \mathrm{E}\left[\frac{\nu_t(n)}{n}e^{\frac{\theta_\nu}{n}\nu_t(n)+\frac{\theta_Q}{n}Q_{t,\nu}(n)}\right]\\
&\quad + \mu\theta_\nu\nu^*\mathrm{E}\left[e^{\frac{\theta_\nu}{n}\nu_t(n)+\frac{\theta_Q}{n}Q_{t,\nu}(n)}\right] - \mu\theta_Q\mathrm{E}\left[\frac{Q_{t,\nu}(n)}{n}e^{\frac{\theta_\nu}{n}\nu_t(n)+\frac{\theta_Q}{n}Q_{t,\nu}(n)}\right] + \frac{\mu\epsilon_n}{n}\mathrm{E}\left[Q_{t,\nu}(n)e^{\frac{\theta_\nu}{n}\nu_t(n)+\frac{\theta_Q}{n}Q_{t,\nu}(n)}\right]\\
&= (\mu+\beta)\nu^*\theta_\nu\tilde{\mathcal{M}}^n(t,\theta_\nu,\theta_Q) + \left(n\left(e^{\frac{\alpha\theta_\nu}{n}+\frac{\theta_Q}{n}}-1\right)-(\mu+\beta)\theta_\nu\right)\frac{\partial}{\partial\theta_\nu}\tilde{\mathcal{M}}^n(t,\theta_\nu,\theta_Q)\\
&\quad - \mu\theta_Q\frac{\partial}{\partial\theta_Q}\tilde{\mathcal{M}}^n(t,\theta_\nu,\theta_Q) + \frac{\mu\epsilon_n}{n^2}\mathrm{E}\left[Q_{t,\nu}(n)e^{\frac{\theta_\nu}{n}\nu_t(n)+\frac{\theta_Q}{n}Q_{t,\nu}(n)}\right]\\
&\xrightarrow{n\to\infty} (\mu+\beta)\nu^*\theta_\nu\tilde{\mathcal{M}}^\infty(t,\theta_\nu,\theta_Q) + (\theta_Q-(\mu+\beta-\alpha)\theta_\nu)\frac{\partial}{\partial\theta_\nu}\tilde{\mathcal{M}}^\infty(t,\theta_\nu,\theta_Q) - \mu\theta_Q\frac{\partial}{\partial\theta_Q}\tilde{\mathcal{M}}^\infty(t,\theta_\nu,\theta_Q).
\end{aligned}
$$

We now solve this partial differential equation for $\tilde{\mathcal{M}}^\infty(t,\theta_\nu,\theta_Q)$ through the method of characteristics. For simplicity's sake we will instead use this procedure to solve for $G(t,\theta_\nu,\theta_Q) = \log\left(\tilde{\mathcal{M}}^\infty(t,\theta_\nu,\theta_Q)\right)$. This PDE is given by

$$
(\mu+\beta)\nu^*\theta_\nu = \frac{\partial}{\partial t}G(t,\theta_\nu,\theta_Q) + \mu\theta_Q\frac{\partial}{\partial\theta_Q}G(t,\theta_\nu,\theta_Q) + ((\mu+\beta-\alpha)\theta_\nu-\theta_Q)\frac{\partial}{\partial\theta_\nu}G(t,\theta_\nu,\theta_Q),
$$

with boundary condition $G(0,\theta_\nu,\theta_Q) = \theta_Q Q_0 + \theta_\nu\nu_0$. This corresponds to the following system of characteristics equations:

$$
\begin{aligned}
&\frac{\mathrm{d}\theta_Q}{\mathrm{d}z}(x,y,z) = \mu\theta_Q, && \theta_Q(x,y,0) = x\\
&\frac{\mathrm{d}\theta_\nu}{\mathrm{d}z}(x,y,z) = (\mu+\beta-\alpha)\theta_\nu-\theta_Q, && \theta_\nu(x,y,0) = y\\
&\frac{\mathrm{d}t}{\mathrm{d}z}(x,y,z) = 1, && t(x,y,0) = 0\\
&\frac{\mathrm{d}g}{\mathrm{d}z}(x,y,z) = (\mu+\beta)\nu^*\theta_\nu = (\mu+\beta-\alpha)\nu_\infty\theta_\nu, && g(x,y,0) = xQ_0 + y\nu_0.
\end{aligned}
$$

If $\beta \neq \alpha$, the solutions to these initial value problems are given by:

$$\begin{aligned}
\theta_Q(x,y,z) &= xe^{\mu z}, \\
\theta_\nu(x,y,z) &= ye^{(\mu+\beta-\alpha)z} + \frac{x}{\beta-\alpha}\left(e^{\mu z} - e^{(\mu+\beta-\alpha)z}\right), \\
&= \left(y - \frac{x}{\beta-\alpha}\right)e^{(\mu+\beta-\alpha)z} + \frac{xe^{\mu z}}{\beta-\alpha}, \\
t(x,y,z) &= z, \\
g(x,y,z) &= xQ_0 + y\nu_0 + \nu_\infty\left(y - \frac{x}{\beta-\alpha}\right)\left(e^{(\mu+\beta-\alpha)z} - 1\right) + \frac{x\nu_\infty(\mu+\beta-\alpha)(e^{\mu z}-1)}{\mu(\beta-\alpha)}.
\end{aligned}$$

Now, we can solve for the characteristic variables in terms of the original variables and find $x = \theta_Q e^{-\mu t}$, $y = \theta_\nu e^{-(\mu+\beta-\alpha)t} + \frac{\theta_Q}{\beta-\alpha}\left(e^{-\mu t} - e^{-(\mu+\beta-\alpha)t}\right)$, and $z = t$, so this gives a PDE solution of

$$\begin{aligned}
G(t,\theta_Q,\theta_\nu) &= g\left(\theta_Q e^{-\mu t}, \theta_\nu e^{-(\mu+\beta-\alpha)t} + \frac{\theta_Q}{\beta-\alpha}\left(e^{-\mu t} - e^{-(\mu+\beta-\alpha)t}\right), t\right) \\
&= Q_0\theta_Q e^{-\mu t} + \nu_0\left(\theta_\nu e^{-(\mu+\beta-\alpha)t} + \frac{\theta_Q}{\beta-\alpha}\left(e^{-\mu t} - e^{-(\mu+\beta-\alpha)t}\right)\right) \\
&\quad + \nu_\infty\left(\theta_\nu - \frac{\theta_Q}{\beta-\alpha}\right)\left(1 - e^{-(\mu+\beta-\alpha)t}\right) + \frac{\theta_Q\nu_\infty(\mu+\beta-\alpha)(1-e^{-\mu t})}{\mu(\beta-\alpha)}.
\end{aligned}$$

If instead $\beta = \alpha$, the solutions to the characteristic ODE's are as follows:

$$\begin{aligned}
\theta_Q(x,y,z) &= xe^{\mu z}, \\
\theta_\nu(x,y,z) &= e^{\mu z}\left(y - xz\right), \\
t(x,y,z) &= z, \\
g(x,y,z) &= xQ_0 + y\nu_0 + \nu_\infty y\left(e^{\mu z} - 1\right) - \frac{x\nu_\infty}{\mu}\left(e^{\mu z}(\mu z - 1) + 1\right).
\end{aligned}$$

This makes our expressions for the characteristic variables $x = \theta_Q e^{-\mu t}$, $y = \theta_\nu e^{-\mu t} + \theta_Q t e^{-\mu t}$, and $z = t$. This now makes the PDE solution

$$\begin{aligned}
G(t,\theta_Q,\theta_\nu) &= g\left(\theta_Q e^{-\mu t}, \theta_\nu e^{-\mu t} + \theta_Q t e^{-\mu t}, t\right) \\
&= Q_0\theta_Q e^{-\mu t} + \nu_0\theta_\nu e^{-\mu t} + \nu_0\theta_Q t e^{-\mu t} + \nu_\infty\left(\theta_\nu + \theta_Q t\right)\left(1 - e^{-\mu t}\right) - \frac{\nu_\infty\theta_Q}{\mu}\left(\mu t - 1 + e^{-\mu t}\right),
\end{aligned}$$

and we can observe that each of these cases simplify to the corresponding means of the queue and the intensity, which yields the stated result. □

A.2.4 Baseline Diffusion Limit of the HESEP

To now consider a diffusion limit we will still scale the baseline intensity by n, but we now instead scale the process intensity and the queue length by $\frac{1}{\sqrt{n}}$. More specifically, we scale the centered version of the processes by $\frac{1}{\sqrt{n}}$. While we can make use of some of the techniques used for the fluid limit in Theorem A.2.7, the diffusion scaling also involves second order terms. It is challenging to calculate such quantities for the HESEP. Thus, we will use the same idea bound the quantities above and below via

$$0 \leq \frac{(\nu_t - \nu^*)^2}{Q_{t,\nu}} \leq \alpha(\nu_t - \nu^*), \tag{A.20}$$

which follows from our previously discussed bounds on the down-jump size. By doing so, we create upper and lower bounds for the true diffusion limit of the HESEP. To facilitate a variety of approximations that fit within these bounds, we introduce the parameter $\gamma \in [0, 1]$, with $\gamma = 0$ corresponding to the lower bound and $\gamma = 1$ as the upper.

Theorem A.2.8. *For $n \in \mathbb{Z}$, let the nth diffusion-scaled HESEP $(\nu_t(n), Q_{t,\nu}(n))$ be defined such that the baseline intensity is $n\nu^*$, the intensity jump size is $\alpha > 0$, the intensity decay rate is $\beta \geq 0$, and the rate of exponentially distributed service is $\mu > 0$, where $\mu + \beta > \alpha$. For the scaled quantities $(\frac{\nu_t(n)}{\sqrt{n}}, \frac{Q_{t,\nu}(n)}{\sqrt{n}})$, let $\hat{\mathcal{M}}^\infty(t, \theta_\nu, \theta_Q)$ be defined*

$$\hat{\mathcal{M}}^\infty(t, \theta_\nu, \theta_Q) \equiv \lim_{n\to\infty} \mathrm{E}\left[e^{\frac{\theta_\nu}{\sqrt{n}}(\nu_t(n) - n\nu_\infty) + \frac{\theta_Q}{\sqrt{n}}\left(Q_{t,\nu}(n) - \frac{n\nu_\infty}{\mu}\right)}\right], \tag{A.21}$$

if the limit converges. Then for $\beta \neq \alpha$, this is bounded above and below by $\mathcal{B}_0(t, \theta_\nu, \theta_Q) \leq \hat{\mathcal{M}}^\infty(t, \theta_\nu, \theta_Q) \leq \mathcal{B}_1(t, \theta_\nu, \theta_Q)$, where $\mathcal{B}_\gamma(t, \theta_\nu, \theta_Q)$ is given by

$$\begin{aligned}\mathcal{B}_\gamma(t, \theta_\nu, \theta_Q) = {} & e^{\nu_0\theta_\nu e^{-(\mu+\beta-\alpha)t} + \frac{\nu_0\theta_Q}{\beta-\alpha}\left(e^{-\mu t} - e^{-(\mu+\beta-\alpha)t}\right) + Q_0\theta_Q e^{-\mu t} + \left(\theta_\nu - \frac{\theta_Q}{\beta-\alpha}\right)^2\left(\frac{\gamma\alpha\mu(\nu_\infty - \nu^*)}{2} + \frac{\alpha^2\nu_\infty}{2}\right)\frac{1-e^{-2(\mu+\beta-\alpha)t}}{2(\mu+\beta-\alpha)}} \\ & \cdot e^{\left(\theta_\nu\theta_Q - \frac{\theta_Q^2}{\beta-\alpha}\right)\left(\left(\frac{\gamma\alpha\mu}{\beta-\alpha} + \mu\right)(\nu_\infty - \nu^*) + \frac{\alpha\beta\nu_\infty}{\beta-\alpha}\right)\frac{1-e^{-(2\mu+\beta-\alpha)t}}{2\mu+\beta-\alpha}} \\ & \cdot e^{\theta_Q^2\left(\frac{\gamma\alpha\mu(\nu_\infty - \nu^*)}{2(\beta-\alpha)^2} + \frac{\mu(\nu_\infty - \nu^*)}{\beta-\alpha} + \frac{\nu_\infty}{2} + \frac{\nu_\infty\beta^2}{2(\beta-\alpha)^2}\right)\frac{1-e^{-2\mu t}}{2\mu}},\end{aligned} \tag{A.22}$$

whereas if $\beta = \alpha$, it is instead

$$\mathcal{B}_\gamma(t,\theta_\nu,\theta_Q) = e^{\nu_0\theta_\nu e^{-\mu t}+\nu_0\theta_Q t e^{-\mu t}+Q_0\theta_Q e^{-\mu t}+\left(\left(\frac{\gamma\alpha(\theta_\nu+\theta_Q t)^2}{2}+\theta_\nu\theta_Q+\theta_Q^2 t\right)\frac{1-e^{-2\mu t}}{2}-\left(\gamma\alpha(\theta_\nu\theta_Q+\theta_Q^2 t)+\theta_Q^2\right)\frac{2\mu t-1+e^{-2\mu t}}{4\mu}\right.}$$
$$\cdot\, e^{\frac{\gamma\alpha\theta_Q^2}{2}\left(\frac{2\mu t(\mu t-1)+1-e^{-2\mu t}}{4\mu^2}\right)\Big)(\nu_\infty-\nu^*)+\frac{\nu_\infty}{2}\left(\left(\theta_Q^2+(\theta_Q+\alpha\theta_\nu)^2+2\left(\alpha^2\theta_\nu\theta_Q+\alpha\theta_Q^2\right)t+\alpha^2\theta_Q^2 t^2\right)\frac{1-e^{-2\mu t}}{2\mu}\right.}$$
$$\cdot\, e^{\left.-2\left(\alpha^2\theta_\nu\theta_Q+\alpha\theta_Q^2+\alpha^2\theta_Q^2 t\right)\left(\frac{2\mu t-1+e^{-2\mu t}}{4\mu^2}\right)+\alpha^2\theta_Q^2\left(\frac{2\mu t(\mu t-1)+1-e^{-2\mu t}}{4\mu^3}\right)\right)}, \tag{A.23}$$

for $\gamma \in [0,1]$ with $t \geq 0$ and $\nu_\infty = \frac{(\mu+\beta)\nu^*}{\mu+\beta-\alpha}$.

Proof. We begin by bounding the quantity $Q_{t,\nu}(n)\left(\frac{\nu_t(n)-n\nu^*}{Q_{t,\nu}(n)}\right)^2$ above and below by observing

$$0 \leq Q_{t,\nu}(n)\left(\frac{\nu_t(n)-n\nu^*}{Q_{t,\nu}(n)}\right)^2 = (\nu_t(n)-n\nu^*)\left(\frac{\nu_t(n)-n\nu^*}{Q_{t,\nu}(n)}\right) \leq \alpha(\nu_t(n)-n\nu^*).$$

To consolidate the development of the two bounds into one approach, we introduce the extra parameter $\gamma \in \{0,1\}$ and replace $Q_{t,\nu}(n)\left(\frac{\nu_t(n)-n\nu^*}{Q_{t,\nu}(n)}\right)^2$ by $\gamma\alpha(\nu_t(n)-n\nu^*)$ in the following diffusion limit derivation. In this notation, $\gamma = 0$ yields the lower bound and $\gamma = 1$ the upper. These two cases share the same start – identifying the moment generating function form of the pre-limit object. By Lemma A.1.1, this is

$$\begin{aligned}
\frac{\partial}{\partial t}\hat{\mathcal{M}}^n(\theta_\nu,\theta_Q,t) &= \frac{\partial}{\partial t}\mathrm{E}\left[e^{\theta_\nu\left(\frac{\nu_t(n)-n\nu_\infty}{\sqrt{n}}\right)+\theta_Q\left(\frac{Q_{t,\nu}(n)-\frac{n\nu_\infty}{\mu}}{\sqrt{n}}\right)}\right] \\
&= \mathrm{E}\left[\nu_t(n)\left(e^{\frac{\alpha\theta_\nu+\theta_Q}{\sqrt{n}}}-1\right)e^{\theta_\nu\left(\frac{\nu_t(n)-n\nu_\infty}{\sqrt{n}}\right)+\theta_Q\left(\frac{Q_{t,\nu}(n)-\frac{n\nu_\infty}{\mu}}{\sqrt{n}}\right)}\right] \\
&\quad + \mathrm{E}\left[\mu Q_{t,\nu}(n)\left(e^{-\frac{\theta_\nu}{\sqrt{n}}\left(\frac{\nu_t(n)-n\nu^*}{Q_{t,\nu}(n)}\right)-\frac{\theta_Q}{\sqrt{n}}}-1\right)e^{\theta_\nu\left(\frac{\nu_t(n)-n\nu_\infty}{\sqrt{n}}\right)+\theta_Q\left(\frac{Q_{t,\nu}(n)-\frac{n\nu_\infty}{\mu}}{\sqrt{n}}\right)}\right] \\
&\quad + \mathrm{E}\left[\frac{\beta\theta_\nu}{\sqrt{n}}\left(n\nu^*-\nu_t(n)\right)e^{\theta_\nu\left(\frac{\nu_t(n)-n\nu_\infty}{\sqrt{n}}\right)+\theta_Q\left(\frac{Q_{t,\nu}(n)-\frac{n\nu_\infty}{\mu}}{\sqrt{n}}\right)}\right].
\end{aligned}$$

As a first step, we simplify this expression through the linearity of expectation.

Moving deterministic terms outside of the expectation and re-scaling, we have

$$\begin{aligned}\frac{\partial}{\partial t}\hat{\mathcal{M}}^n(\theta_\nu,\theta_Q,t) = &\ \sqrt{n}\left(e^{\frac{\alpha\theta_\nu+\theta_Q}{\sqrt{n}}}-1\right)\mathrm{E}\left[\frac{\nu_t(n)}{\sqrt{n}}e^{\theta_\nu\left(\frac{\nu_t(n)-n\nu_\infty}{\sqrt{n}}\right)+\theta_Q\left(\frac{Q_{t,\nu}(n)-\frac{n\nu_\infty}{\mu}}{\sqrt{n}}\right)}\right]\\ &+\mathrm{E}\left[\mu Q_{t,\nu}(n)\left(e^{-\frac{\theta_\nu}{\sqrt{n}}\left(\frac{\nu_t(n)-n\nu^*}{Q_{t,\nu}(n)}\right)-\frac{\theta_Q}{\sqrt{n}}}-1\right)e^{\theta_\nu\left(\frac{\nu_t(n)-n\nu_\infty}{\sqrt{n}}\right)+\theta_Q\left(\frac{Q_{t,\nu}(n)-\frac{n\nu_\infty}{\mu}}{\sqrt{n}}\right)}\right]\\ &+\beta\theta_\nu\nu^*\sqrt{n}\mathrm{E}\left[e^{\theta_\nu\left(\frac{\nu_t(n)-n\nu_\infty}{\sqrt{n}}\right)+\theta_Q\left(\frac{Q_{t,\nu}(n)-\frac{n\nu_\infty}{\mu}}{\sqrt{n}}\right)}\right]-\beta\theta_\nu\mathrm{E}\left[\frac{\nu_t(n)}{\sqrt{n}}e^{\theta_\nu\left(\frac{\nu_t(n)-n\nu_\infty}{\sqrt{n}}\right)+\theta_Q\left(\frac{Q_{t,\nu}(n)-\frac{n\nu_\infty}{\mu}}{\sqrt{n}}\right)}\right].\end{aligned}$$

For the terms on the first and third lines in the right-hand side of the above equation, we are able to re-express the expectation in terms of the moment generating function or its derivatives. For the first and second lines, we perform Taylor expansions and truncate terms from the third order and above. This now yields

$$\begin{aligned}\frac{\partial}{\partial t}\hat{\mathcal{M}}^n(\theta_\nu,\theta_Q,t) = &\left(\alpha\theta_\nu+\theta_Q+\frac{(\alpha\theta_\nu+\theta_Q)^2}{2\sqrt{n}}+O\left(\frac{1}{n}\right)\right)\left(\frac{\partial}{\partial\theta_\nu}\hat{\mathcal{M}}^n(\theta_\nu,\theta_Q,t)+\nu_\infty\sqrt{n}\hat{\mathcal{M}}^n(\theta_\nu,\theta_Q,t)\right)\\ &+\mathrm{E}\Bigg[\mu Q_{t,\nu}(n)\left(-\frac{\theta_\nu}{\sqrt{n}}\left(\frac{\nu_t(n)-n\nu^*}{Q_{t,\nu}(n)}\right)-\frac{\theta_Q}{\sqrt{n}}+\frac{1}{2n}\left(\theta_\nu\left(\frac{\nu_t(n)-n\nu^*}{Q_{t,\nu}(n)}\right)+\theta_Q\right)^2+O\left(n^{-\frac{3}{2}}\right)\right)\\ &\cdot e^{\theta_\nu\left(\frac{\nu_t(n)-n\nu_\infty}{\sqrt{n}}\right)+\theta_Q\left(\frac{Q_{t,\nu}(n)-\frac{n\nu_\infty}{\mu}}{\sqrt{n}}\right)}\Bigg]+\beta\theta_\nu\nu^*\sqrt{n}\hat{\mathcal{M}}^n(\theta_\nu,\theta_Q,t)-\beta\theta_\nu\left(\frac{\partial}{\partial\theta_\nu}\hat{\mathcal{M}}^n(\theta_\nu,\theta_Q,t)+\nu_\infty\sqrt{n}\hat{\mathcal{M}}^n(\theta_\nu,\theta_Q,t)\right).\end{aligned}$$

We now begin distributing and combining like terms through linearity of expectation. Moreover, we distribute within the expectation on the second line and

cancel $Q_{t,\nu}(n)$ across the numerator and denominator where possible.

$$
\begin{aligned}
\frac{\partial}{\partial t}\hat{\mathcal{M}}^n(\theta_\nu,\theta_Q,t) = {} & \left(\alpha\theta_\nu + \theta_Q + \frac{(\alpha\theta_\nu+\theta_Q)^2}{2\sqrt{n}} - \beta\theta_\nu + O\left(\frac{1}{n}\right)\right)\left(\frac{\partial}{\partial\theta_\nu}\hat{\mathcal{M}}^n(\theta_\nu,\theta_Q,t) + \nu_\infty\sqrt{n}\hat{\mathcal{M}}^n(\theta_\nu,\theta_Q,t)\right) \\
& - \mu\theta_\nu \mathrm{E}\left[\frac{\nu_t(n)}{\sqrt{n}} e^{\theta_\nu\left(\frac{\nu_t(n)-n\nu_\infty}{\sqrt{n}}\right)+\theta_Q\left(\frac{Q_{t,\nu}(n)-\frac{n\nu_\infty}{\mu}}{\sqrt{n}}\right)}\right] + \mu\theta_\nu\nu^*\sqrt{n}\mathrm{E}\left[e^{\theta_\nu\left(\frac{\nu_t(n)-n\nu_\infty}{\sqrt{n}}\right)+\theta_Q\left(\frac{Q_{t,\nu}(n)-\frac{n\nu_\infty}{\mu}}{\sqrt{n}}\right)}\right] \\
& - \mu\theta_Q \mathrm{E}\left[\frac{Q_{t,\nu}(n)}{\sqrt{n}} e^{\theta_\nu\left(\frac{\nu_t(n)-n\nu_\infty}{\sqrt{n}}\right)+\theta_Q\left(\frac{Q_{t,\nu}(n)-\frac{n\nu_\infty}{\mu}}{\sqrt{n}}\right)}\right] \\
& + \frac{\mu\theta_\nu^2}{2n}\mathrm{E}\left[Q_{t,\nu}(n)\left(\frac{\nu_t(n)-n\nu^*}{Q_{t,\nu}(n)}\right)^2 e^{\theta_\nu\left(\frac{\nu_t(n)-n\nu_\infty}{\sqrt{n}}\right)+\theta_Q\left(\frac{Q_{t,\nu}(n)-\frac{n\nu_\infty}{\mu}}{\sqrt{n}}\right)}\right] \\
& + \frac{\mu\theta_\nu\theta_Q}{\sqrt{n}}\mathrm{E}\left[\frac{\nu_t(n)}{\sqrt{n}} e^{\theta_\nu\left(\frac{\nu_t(n)-n\nu_\infty}{\sqrt{n}}\right)+\theta_Q\left(\frac{Q_{t,\nu}(n)-\frac{n\nu_\infty}{\mu}}{\sqrt{n}}\right)}\right] - \mu\theta_\nu\theta_Q\nu^*\mathrm{E}\left[e^{\theta_\nu\left(\frac{\nu_t(n)-n\nu_\infty}{\sqrt{n}}\right)+\theta_Q\left(\frac{Q_{t,\nu}(n)-\frac{n\nu_\infty}{\mu}}{\sqrt{n}}\right)}\right] \\
& + \frac{\mu\theta_Q^2}{2\sqrt{n}}\mathrm{E}\left[\frac{Q_{t,\nu}(n)}{\sqrt{n}} e^{\theta_\nu\left(\frac{\nu_t(n)-n\nu_\infty}{\sqrt{n}}\right)+\theta_Q\left(\frac{Q_{t,\nu}(n)-\frac{n\nu_\infty}{\mu}}{\sqrt{n}}\right)}\right] + O\left(\frac{1}{n}\right)\mathrm{E}\left[\frac{Q_{t,\nu}(n)}{\sqrt{n}} e^{\theta_\nu\left(\frac{\nu_t(n)-n\nu_\infty}{\sqrt{n}}\right)+\theta_Q\left(\frac{Q_{t,\nu}(n)-\frac{n\nu_\infty}{\mu}}{\sqrt{n}}\right)}\right] \\
& + \beta\theta_\nu\nu^*\sqrt{n}\hat{\mathcal{M}}^n(\theta_\nu,\theta_Q,t).
\end{aligned}
$$

For all remaining components of this equation that are still expressed in terms of the expectation, we substitute equivalent forms in terms of the moment generating function or its partial derivatives. Furthermore, we will now replace $Q_{t,\nu}(n)\left(\frac{\nu_t(n)-n\nu^*}{Q_{t,\nu}(n)}\right)^2$ by $\gamma\alpha(\nu_t(n)-n\nu^*)$ inside the expectation and re-express the expectation in terms of the moment generating function accordingly. To denote that we have now made this replacement and changed the function, we add γ

as a subscript to the moment generating function, i.e. $\hat{\mathcal{M}}^n_\gamma(\theta_\nu, \theta_Q, t)$.

$$
\begin{aligned}
\frac{\partial}{\partial t}\hat{\mathcal{M}}^n_\gamma(\theta_\nu, \theta_Q, t) &= \left(\alpha\theta_\nu + \theta_Q + \frac{(\alpha\theta_\nu + \theta_Q)^2}{2\sqrt{n}} - \beta\theta_\nu + O\left(\frac{1}{n}\right)\right)\left(\frac{\partial}{\partial\theta_\nu}\hat{\mathcal{M}}^n_\gamma(\theta_\nu, \theta_Q, t) + \nu_\infty\sqrt{n}\hat{\mathcal{M}}^n_\gamma(\theta_\nu, \theta_Q, t)\right) \\
&\quad - \mu\theta_\nu\left(\frac{\partial}{\partial\theta_\nu}\hat{\mathcal{M}}^n_\gamma(\theta_\nu, \theta_Q, t) + \nu_\infty\sqrt{n}\hat{\mathcal{M}}^n_\gamma(\theta_\nu, \theta_Q, t)\right) + \mu\theta_\nu\nu^*\sqrt{n}\hat{\mathcal{M}}^n_\gamma(\theta_\nu, \theta_Q, t) \\
&\quad - \mu\theta_Q\left(\frac{\partial}{\partial\theta_Q}\hat{\mathcal{M}}^n_\gamma(\theta_\nu, \theta_Q, t) + \frac{\nu_\infty\sqrt{n}}{\mu}\hat{\mathcal{M}}^n_\gamma(\theta_\nu, \theta_Q, t)\right) \\
&\quad + \frac{\gamma\alpha\mu\theta_\nu^2}{2\sqrt{n}}\left(\frac{\partial}{\partial\theta_\nu}\hat{\mathcal{M}}^n_\gamma(\theta_\nu, \theta_Q, t) + \nu_\infty\sqrt{n}\hat{\mathcal{M}}^n_\gamma(\theta_\nu, \theta_Q, t)\right) - \frac{\gamma\alpha\mu\nu^*\theta_\nu^2}{2}\hat{\mathcal{M}}^n_\gamma(\theta_\nu, \theta_Q, t) \\
&\quad + \frac{\mu\theta_\nu\theta_Q}{\sqrt{n}}\left(\frac{\partial}{\partial\theta_\nu}\hat{\mathcal{M}}^n_\gamma(\theta_\nu, \theta_Q, t) + \nu_\infty\sqrt{n}\hat{\mathcal{M}}^n_\gamma(\theta_\nu, \theta_Q, t)\right) - \mu\theta_\nu\theta_Q\nu^*\hat{\mathcal{M}}^n_\gamma(\theta_\nu, \theta_Q, t) \\
&\quad + \frac{\mu\theta_Q^2}{2\sqrt{n}}\left(\frac{\partial}{\partial\theta_Q}\hat{\mathcal{M}}^n_\gamma(\theta_\nu, \theta_Q, t) + \frac{\nu_\infty\sqrt{n}}{\mu}\hat{\mathcal{M}}^n_\gamma(\theta_\nu, \theta_Q, t)\right) + O\left(\frac{1}{n}\right)\left(\frac{\partial}{\partial\theta_Q}\hat{\mathcal{M}}^n_\gamma(\theta_\nu, \theta_Q, t) + \frac{\nu_\infty\sqrt{n}}{\mu}\hat{\mathcal{M}}^n_\gamma(\theta_\nu, \theta_Q, t)\right) \\
&\quad + \beta\theta_\nu\nu^*\sqrt{n}\hat{\mathcal{M}}^n_\gamma(\theta_\nu, \theta_Q, t).
\end{aligned}
$$

Before we find the limiting object, we first combine like terms of the moment generating function, consolidating coefficients and absorbing into $O(\cdot)$ notation where possible.

$$
\begin{aligned}
\frac{\partial}{\partial t}\hat{\mathcal{M}}^n_\gamma(\theta_\nu, \theta_Q, t) &= \left(\theta_Q - (\mu + \beta - \alpha)\theta_\nu + O\left(\frac{1}{\sqrt{n}}\right)\right)\frac{\partial}{\partial\theta_\nu}\hat{\mathcal{M}}^n_\gamma(\theta_\nu, \theta_Q, t) - \left(\mu\theta_Q - O\left(\frac{1}{\sqrt{n}}\right)\right)\frac{\partial}{\partial\theta_Q}\hat{\mathcal{M}}^n_\gamma(\theta_\nu, \theta_Q, t) \\
&\quad \left(\frac{\gamma\alpha\mu(\nu_\infty - \nu^*)\theta_\nu^2}{2} + \mu\theta_\nu\theta_Q(\nu_\infty - \nu^*) + \frac{\theta_Q^2\nu_\infty}{2} + \frac{(\alpha\theta_\nu + \theta_Q)^2\nu_\infty}{2} + O\left(\frac{1}{\sqrt{n}}\right)\right)\hat{\mathcal{M}}^n_\gamma(\theta_\nu, \theta_Q, t).
\end{aligned}
$$

Taking the limit as $n \to \infty$, we receive

$$
\begin{aligned}
\frac{\partial}{\partial t}\hat{\mathcal{M}}^\infty_\gamma(\theta_\nu, \theta_Q, t) &= (\theta_Q - (\mu + \beta - \alpha)\theta_\nu)\frac{\partial}{\partial\theta_\nu}\hat{\mathcal{M}}^\infty_\gamma(\theta_\nu, \theta_Q, t) - \mu\theta_Q\frac{\partial}{\partial\theta_Q}\hat{\mathcal{M}}^\infty_\gamma(\theta_\nu, \theta_Q, t) \\
&\quad + \left(\frac{\gamma\alpha\mu(\nu_\infty - \nu^*)\theta_\nu^2}{2} + \mu\theta_\nu\theta_Q(\nu_\infty - \nu^*) + \frac{\theta_Q^2\nu_\infty}{2} + \frac{(\alpha\theta_\nu + \theta_Q)^2\nu_\infty}{2}\right)\hat{\mathcal{M}}^\infty_\gamma(\theta_\nu, \theta_Q, t).
\end{aligned}
$$

We will now solve this limiting partial differential equation through the method of characteristics. To simplify this approach, we let $G_\gamma(\theta_\nu, \theta_Q, t) = \log\left(\hat{\mathcal{M}}^\infty_\gamma(\theta_\nu, \theta_Q, t)\right)$, which is the cumulant generating function. The resulting PDE

for the cumulant generating function is then

$$\frac{\partial}{\partial t}G_\gamma(\theta_\nu,\theta_Q,t) + ((\mu+\beta-\alpha)\theta_\nu - \theta_Q)\frac{\partial}{\partial\theta_\nu}G_\gamma(\theta_\nu,\theta_Q,t) + \mu\theta_Q\frac{\partial}{\partial\theta_Q}G_\gamma(\theta_\nu,\theta_Q,t)$$
$$= \frac{\gamma\alpha\mu(\nu_\infty-\nu^*)\theta_\nu^2}{2} + \mu\theta_\nu\theta_Q(\nu_\infty-\nu^*) + \frac{\theta_Q^2\nu_\infty}{2} + \frac{(\alpha\theta_\nu+\theta_Q)^2\nu_\infty}{2},$$

with initial condition $G_\gamma(\theta_\nu,\theta_Q,0) = \theta_\nu\nu_0 + \theta_Q Q_0$. Thus, the resulting system of characteristic equations is

$$\begin{aligned}
&\frac{\mathrm{d}t}{\mathrm{d}z}(x,y,z) = 1, && t(x,y,0) = 0,\\
&\frac{\mathrm{d}\theta_\nu}{\mathrm{d}z}(x,y,z) = (\mu+\beta-\alpha)\theta_\nu - \theta_Q, && \theta_\nu(x,y,0) = x,\\
&\frac{\mathrm{d}\theta_Q}{\mathrm{d}z}(x,y,z) = \mu\theta_Q, && \theta_Q(x,y,0) = y,\\
&\frac{\mathrm{d}g}{\mathrm{d}z}(x,y,z) = \left(\frac{\gamma\alpha\mu\theta_\nu^2}{2} + \mu\theta_\nu\theta_Q\right)(\nu_\infty-\nu^*) + \left(\theta_Q^2 + (\alpha\theta_\nu+\theta_Q)^2\right)\frac{\nu_\infty}{2}, && g(x,y,0) = x\nu_0 + yQ_0.
\end{aligned}$$

Assuming $\beta \neq \alpha$, we can solve these first three ordinary differential equations to find that

$$t = z, \qquad \theta_Q = ye^{\mu z}, \qquad \theta_\nu = \left(x - \frac{y}{\beta-\alpha}\right)e^{(\mu+\beta-\alpha)z} + \frac{y}{\beta-\alpha}e^{\mu z},$$

which we now use to solve the remaining equation. Re-writing the characteristic equation for g, we have

$$\begin{aligned}
\frac{\mathrm{d}g}{\mathrm{d}z}(x,y,z) = {} & \frac{\gamma\alpha\mu(\nu_\infty-\nu^*)}{2}\left(\left(x-\frac{y}{\beta-\alpha}\right)^2 e^{2(\mu+\beta-\alpha)z} + \frac{2}{\beta-\alpha}\left(xy - \frac{y^2}{\beta-\alpha}\right)e^{(2\mu+\beta-\alpha)z} + \frac{y^2}{(\beta-\alpha)^2}e^{2\mu z}\right)\\
& + \mu(\nu_\infty-\nu^*)\left(\left(xy-\frac{y^2}{\beta-\alpha}\right)e^{(2\mu+\beta-\alpha)z} + \frac{y^2}{\beta-\alpha}e^{2\mu z}\right) + \frac{\nu_\infty}{2}\left(\alpha^2\left(x-\frac{y}{\beta-\alpha}\right)^2 e^{2(\mu+\beta-\alpha)z}\right.\\
& \left. + \frac{2\alpha\beta}{\beta-\alpha}\left(xy-\frac{y^2}{\beta-\alpha}\right)e^{(2\mu+\beta-\alpha)z} + \left(1+\frac{\beta^2}{(\beta-\alpha)^2}\right)y^2e^{2\mu z}\right),
\end{aligned}$$

and so by grouping coefficients of like exponential functions and then integrat-

ing with respect to z, this solves to

$$\begin{aligned}
g(x,y,z) = x\nu_0 + yQ_0 &+ \left(x - \frac{y}{\beta-\alpha}\right)^2 \left(\frac{\gamma\alpha\mu(\nu_\infty - \nu^*)}{2} + \frac{\alpha^2\nu_\infty}{2}\right)\frac{e^{2(\mu+\beta-\alpha)z}-1}{2(\mu+\beta-\alpha)} \\
&+ \left(xy - \frac{y^2}{\beta-\alpha}\right)\left(\left(\frac{\gamma\alpha\mu}{\beta-\alpha}+\mu\right)(\nu_\infty-\nu^*) + \frac{\alpha\beta\nu_\infty}{\beta-\alpha}\right)\frac{e^{(2\mu+\beta-\alpha)z}-1}{2\mu+\beta-\alpha} \\
&+ y^2\left(\frac{\gamma\alpha\mu(\nu_\infty-\nu^*)}{2(\beta-\alpha)^2} + \frac{\mu(\nu_\infty-\nu^*)}{\beta-\alpha} + \frac{\nu_\infty}{2} + \frac{\nu_\infty\beta^2}{2(\beta-\alpha)^2}\right)\frac{e^{2\mu z}-1}{2\mu}.
\end{aligned}$$

From the solutions to the characteristic equations, we can express each of x, y, and z in terms of the three cumulant generating function parameters:

$$z = t, \qquad y = \theta_Q e^{-\mu t}, \qquad x = \theta_\nu e^{-(\mu+\beta-\alpha)t} + \frac{\theta_Q}{\beta-\alpha}\left(e^{-\mu t} - e^{-(\mu+\beta-\alpha)t}\right).$$

Thus, we can then solve for $G_\gamma(\theta_\nu, \theta_Q, t)$ via

$$\begin{aligned}
G_\gamma(\theta_\nu,\theta_Q,t) &= g\left(\theta_\nu e^{-(\mu+\beta-\alpha)t} + \frac{\theta_Q}{\beta-\alpha}\left(e^{-\mu t} - e^{-(\mu+\beta-\alpha)t}\right), \theta_Q e^{-\mu t}, t\right) \\
&= \nu_0\theta_\nu e^{-(\mu+\beta-\alpha)t} + \frac{\nu_0\theta_Q}{\beta-\alpha}\left(e^{-\mu t} - e^{-(\mu+\beta-\alpha)t}\right) + Q_0\theta_Q e^{-\mu t} \\
&\quad + \left(\theta_\nu - \frac{\theta_Q}{\beta-\alpha}\right)^2\left(\frac{\gamma\alpha\mu(\nu_\infty-\nu^*)}{2} + \frac{\alpha^2\nu_\infty}{2}\right)\frac{1-e^{-2(\mu+\beta-\alpha)t}}{2(\mu+\beta-\alpha)} \\
&\quad + \left(\theta_\nu\theta_Q - \frac{\theta_Q^2}{\beta-\alpha}\right)\left(\left(\frac{\gamma\alpha\mu}{\beta-\alpha}+\mu\right)(\nu_\infty-\nu^*) + \frac{\alpha\beta\nu_\infty}{\beta-\alpha}\right)\frac{1-e^{-(2\mu+\beta-\alpha)t}}{2\mu+\beta-\alpha} \\
&\quad + \theta_Q^2\left(\frac{\gamma\alpha\mu(\nu_\infty-\nu^*)}{2(\beta-\alpha)^2} + \frac{\mu(\nu_\infty-\nu^*)}{\beta-\alpha} + \frac{\nu_\infty}{2} + \frac{\nu_\infty\beta^2}{2(\beta-\alpha)^2}\right)\frac{1-e^{-2\mu t}}{2\mu}.
\end{aligned}$$

By Lemma A.1.2, we have that $\hat{\mathcal{M}}_0^\infty(\theta_\nu,\theta_Q,t) \le \hat{\mathcal{M}}^\infty(\theta_\nu,\theta_Q,t) \le \hat{\mathcal{M}}_1^\infty(\theta_\nu,\theta_Q,t)$ and since $\hat{\mathcal{M}}_\gamma^\infty(\theta_\nu,\theta_Q,t) = e^{G_\gamma(\theta_\nu,\theta_Q,t)}$, we have completed the proof of the joint moment generating function bounds when $\beta \neq \alpha$. We now apply this to the two marginal generating functions by setting the opposite space parameter to 0. That is, for the intensity we let $\theta_Q = 0$, yielding

$$\hat{\mathcal{M}}_\gamma^\infty(\theta_\nu,0,t) = \exp\left(\nu_0\theta_\nu e^{-(\mu+\beta-\alpha)t} + \frac{\theta_\nu^2}{2}\left(\gamma\alpha\mu(\nu_\infty-\nu^*) + \alpha^2\nu_\infty\right)\frac{1-e^{-2(\mu+\beta-\alpha)t}}{2(\mu+\beta-\alpha)}\right),$$

whereas for the queue we take $\theta_\nu = 0$ and receive

$$\begin{aligned}\hat{\mathcal{M}}_\gamma^\infty(0,\theta_Q,t) = \exp\Bigg(&\frac{\nu_0\theta_Q}{\beta-\alpha}\left(e^{-\mu t} - e^{-(\mu+\beta-\alpha)t}\right) + \frac{\theta_Q^2}{(\beta-\alpha)^2}\left(\frac{\gamma\alpha\mu(\nu_\infty-\nu^*)}{2} + \frac{\alpha^2\nu_\infty}{2}\right)\frac{1-e^{-2(\mu+\beta-\alpha)t}}{2(\mu+\beta-\alpha)}\\ &+ Q_0\theta_Q e^{-\mu t} - \frac{\theta_Q^2}{\beta-\alpha}\left(\left(\frac{\gamma\alpha\mu}{\beta-\alpha}+\mu\right)(\nu_\infty-\nu^*) + \frac{\alpha\beta\nu_\infty}{\beta-\alpha}\right)\frac{1-e^{-(2\mu+\beta-\alpha)t}}{2\mu+\beta-\alpha}\\ &+ \theta_Q^2\left(\frac{\gamma\alpha\mu(\nu_\infty-\nu^*)}{2(\beta-\alpha)^2} + \frac{\mu(\nu_\infty-\nu^*)}{\beta-\alpha} + \frac{\nu_\infty}{2} + \frac{\nu_\infty\beta^2}{2(\beta-\alpha)^2}\right)\frac{1-e^{-2\mu t}}{2\mu}\Bigg).\end{aligned}$$

Now if $\beta = \alpha$, the solution to the characteristic ODE for θ_ν is instead

$$\theta_\nu = xe^{\mu z} - yze^{\mu z},$$

whereas the solutions for θ_Q and t are unchanged: $\theta_Q = ye^{\mu z}$ and $t = z$. This then implies that ODE for g is given by

$$\begin{aligned}\frac{\mathrm{d}g}{\mathrm{d}z}(x,y,z) = &\left(\frac{\gamma\alpha\mu}{2}\left(x^2e^{2\mu z} - 2xyze^{2\mu z} + y^2z^2e^{2\mu z}\right) + \mu\left(xye^{2\mu z} - y^2ze^{2\mu z}\right)\right)(\nu_\infty-\nu^*)\\ &+ \frac{\nu_\infty}{2}\left((2y^2+\alpha^2x^2+2\alpha xy)e^{2\mu z} - 2(\alpha^2xy+\alpha y^2)ze^{2\mu z} + \alpha^2y^2z^2e^{2\mu z}\right),\end{aligned}$$

which yields a solution of

$$\begin{aligned}g(x,y,z) = x\nu_0 + yQ_0 + \Bigg(&\left(\frac{\gamma\alpha x^2}{2} + xy\right)\frac{e^{2\mu z}-1}{2} - \left(\gamma\alpha xy + y^2\right)\frac{e^{2\mu z}(2\mu z-1)+1}{4\mu}\\ &+ \frac{\gamma\alpha y^2}{2}\left(\frac{e^{2\mu z}\left(2\mu z(\mu z-1)+1\right)-1}{4\mu^2}\right)\Bigg)(\nu_\infty-\nu^*) + \frac{\nu_\infty}{2}\Bigg(\left(2y^2+\alpha^2x^2+2\alpha xy\right)\frac{e^{2\mu z}-1}{2\mu}\\ &- 2(\alpha^2xy+\alpha y^2)\left(\frac{e^{2\mu z}(2\mu z-1)+1}{4\mu^2}\right) + \alpha^2y^2\left(\frac{e^{2\mu z}\left(2\mu z(\mu z-1)+1\right)-1}{4\mu^3}\right)\Bigg).\end{aligned}$$

In this case the inverse solutions are

$$z = t, \qquad y = \theta_Q e^{-\mu t}, \qquad x = \theta_\nu e^{-\mu t} + \theta_Q t e^{-\mu t},$$

and so $G_\gamma(\theta_\nu, \theta_Q, t)$ is given by

$$
\begin{aligned}
G_\gamma(\theta_\nu, \theta_Q, t) &= g(\theta_\nu e^{-\mu t} + \theta_Q t e^{-\mu t}, \theta_Q e^{-\mu z}, t) \\
&= \nu_0 \theta_\nu e^{-\mu t} + \nu_0 \theta_Q t e^{-\mu t} + Q_0 \theta_Q e^{-\mu t} + \left(\left(\frac{\gamma\alpha(\theta_\nu + \theta_Q t)^2}{2} + \theta_\nu \theta_Q + \theta_Q^2 t \right) \frac{1 - e^{-2\mu t}}{2} \right. \\
&\quad - \left(\gamma\alpha(\theta_\nu \theta_Q + \theta_Q^2 t) + \theta_Q^2 \right) \frac{2\mu t - 1 + e^{-2\mu t}}{4\mu} + \frac{\gamma\alpha\theta_Q^2}{2} \left(\frac{2\mu t(\mu t - 1) + 1 - e^{-2\mu t}}{4\mu^2} \right) \Bigg) (\nu_\infty - \nu^*) \\
&\quad + \frac{\nu_\infty}{2} \Bigg(\left(2\theta_Q^2 + \alpha^2 \theta_\nu^2 + 2\alpha^2 \theta_\nu \theta_Q t + \alpha^2 \theta_Q^2 t^2 + 2\alpha\theta_\nu \theta_\nu + 2\alpha\theta_Q^2 t \right) \frac{1 - e^{-2\mu t}}{2\mu} \\
&\quad - 2 \left(\alpha^2 \theta_\nu \theta_Q + \alpha^2 \theta_Q^2 t + \alpha\theta_Q^2 \right) \left(\frac{2\mu t - 1 + e^{-2\mu t}}{4\mu^2} \right) + \alpha^2 \theta_Q^2 \left(\frac{2\mu t(\mu t - 1) + 1 - e^{-2\mu t}}{4\mu^3} \right) \Bigg).
\end{aligned}
$$

By taking $\hat{\mathcal{M}}_\gamma^\infty(\theta_\nu, \theta_Q, t) = e^{G_\gamma(\theta_\nu, \theta_Q, t)}$, we complete the proof. □

As a consequence of these diffusion approximations, we can give normally distributed approximations for the steady-state distributions of the HESEP intensity and queue length. These are stated below in Corollary A.2.9 again in terms of γ. One can note that the approximate intensity variance in Equation A.24 can be used to provide upper and lower bounds on the HESEP variance that may be tighter than the bounds from the ESEP and the Hawkes process in Proposition A.2.2

Corollary A.2.9. *Let $(\nu_t, Q_{t,\nu})$ be an HESEP with baseline intensity $\nu^* > 0$, intensity jump $\alpha > 0$, decay rate $\beta > 0$, and rate of exponential service $\mu > 0$, with $\mu + \beta > \alpha$. Then, the steady-state distributions of processes ν_t and $Q_{t,\nu}$ are approximated by the random variables $X_\nu(\gamma) \sim N(\nu_\infty, \sigma_\nu^2(\gamma))$ and $X_Q(\gamma) \sim N(\frac{\nu_\infty}{\mu}, \sigma_Q^2(\gamma))$, respectively, where*

$$
\sigma_\nu^2(\gamma) = \frac{\gamma\alpha\mu(\nu_\infty - \nu^*) + \alpha^2 \nu_\infty}{2(\mu + \beta - \alpha)}, \tag{A.24}
$$

and if $\beta \neq \alpha$ then

$$
\begin{aligned}
\sigma_Q^2(\gamma) = &\frac{\gamma\alpha\mu(\nu_\infty - \nu^*) + \alpha^2 \nu_\infty}{2(\beta - \alpha)^2(\mu + \beta - \alpha)} - \frac{(2\gamma\alpha\mu + 2\mu(\beta - \alpha))(\nu_\infty - \nu^*) + 2\alpha\beta\nu_\infty}{(\beta - \alpha)^2(2\mu + \beta - \alpha)} \\
&+ \frac{\gamma\alpha\mu(\nu_\infty - \nu^*) + \nu_\infty \beta^2}{2\mu(\beta - \alpha)^2} + \frac{\nu_\infty - \nu^*}{\beta - \alpha} + \frac{\nu_\infty}{2\mu},
\end{aligned} \tag{A.25}
$$

whereas if $\beta = \alpha$ *then*

$$\sigma_Q^2(\gamma) = \left(\frac{1}{2\mu} + \frac{\gamma\alpha}{4\mu^2}\right)(\nu_\infty - \nu^*) + \left(\frac{1}{\mu} + \frac{\alpha}{2\mu^2} + \frac{\alpha^2}{4\mu^3}\right)\nu_\infty, \tag{A.26}$$

with $\nu_\infty = \frac{(\mu+\beta)\nu^*}{\mu+\beta-\alpha}$ *and* $\gamma \in [0, 1]$.

In Figures A.1 and A.2 we plot the simulated steady-state distributions of an HESEP with large baseline intensities, as calculated from 100,000 replications. We then also plot the densities corresponding to the upper and lower approximate diffusion distributions as well as an additional candidate approximation with $\gamma = \frac{\mu}{\mu+\beta}$. We motivate this choice by a ratio of mean approximations of the terms in Equation A.20:

$$\frac{\frac{(\nu_\infty - \nu^*)^2}{\frac{\nu_\infty}{\mu}}}{\alpha(\nu_\infty - \nu^*)} = \frac{\mu(\nu_\infty - \nu^*)}{\alpha\nu_\infty} = \frac{\mu}{\mu + \beta}.$$

In Figure A.1 the baseline intensity is equal to 100, whereas in Figure A.2 it is 1,000. While there are known limitations of Gaussian approximations for queueing processes such as is discussed in Massey and Pender (2013), we see that these approximations appear to be quite close, particularly so for the $\nu^* = 1,000$ case. The upper and lower bounds predictably over- and under-approximate the tails, while the case of $\gamma = \frac{\mu}{\mu+\beta}$ closely mimics the true distribution.

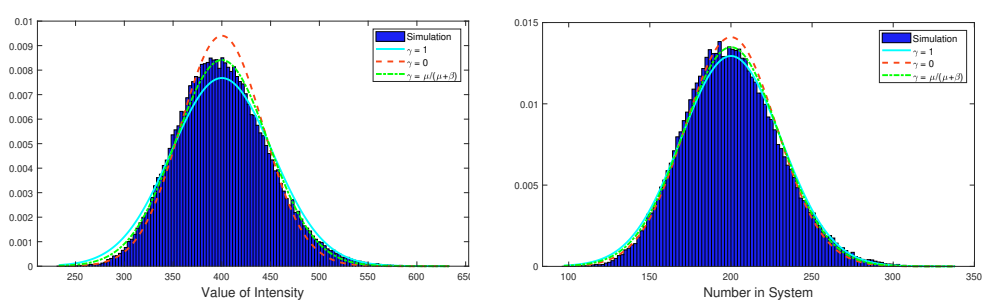


Figure A.1: Histogram comparing the simulated steady-state HESEP intensity (left) and queue (right) to their diffusion approximations evaluated at multiple values of γ, where $\nu^* = 100$, $\alpha = 3$, $\beta = 2$, and $\mu = 2$.

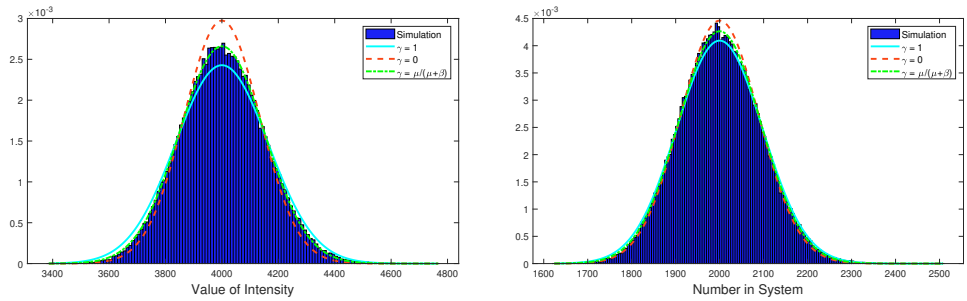


Figure A.2: Histogram comparing the simulated steady-state HESEP intensity (left) and queue (right) to their diffusion approximations evaluated at multiple values of γ, where $\nu^* = 1,000$, $\alpha = 3$, $\beta = 2$, and $\mu = 2$.

A.3 An Ephemeral Self-Exciting Process with Finite Capacity and Blocking

Drawing inspiration from the works that originated queueing theory, we will now consider the change in the ESEP if there is an upper bound on the total number of active exciters. That is, we suppose that there is a finite capacity and no excess buffer beyond them, so that any entities that arrive and find the system full are blocked from entry, thus not registering an arrival nor causing any excitement. As an employee of the Copenhagen Telephone company, A.K. Erlang developed these pioneering queueing models to determine the probability that a call would be blocked based on the capacity of the telephone network trunk line. Often referred to as the Erlang-B model, this queueing system remains relevant not just modern telecommunication systems, but broadly across industries as varied as healthcare operations and transportation. For English translations of the seminal Erlang papers and a biography of the author, see Brockmeyer et al. (1948). In those original works, Erlang supposed that calls arrive perfectly independently, that they have no influence or relationship with one another. In the remainder of this subsection we investigate the scenario

where these calls instead exhibit self-excitement, which is a potential explanation for the over-dispersion that has been seen in industrial call center data, as detailed in e.g. Ibrahim et al. (2016). Another potential application for this model is a website that may receive viral traffic but is also liable to crash if there are too many simultaneous visitors. Additionally, this finite capacity model could be used to represent a restaurant that becomes more enticing the more patrons it has in its limited seating area, like we have discussed in the introduction. To begin, we find the steady-state distribution of this process in Proposition A.3.1. Drawing further inspiration from Erlang's work, we will refer to this finite capacity ESEP model as the *blocking ephemerally self-exciting process* (ESEP-B).

Proposition A.3.1. *Let $\eta_t^{\mathrm{B}} = \eta^* + \alpha Q_t^{\mathrm{B}}$ be a ESEP-B, with baseline intensity $\eta^* > 0$, intensity jump $\alpha > 0$, expiration rate $\beta > \alpha$, and capacity $c \in \mathbb{Z}^+$. That is, if $Q_t^{\mathrm{B}} = c$ any arrivals that occur will be blocked and not recorded. Then, the steady-state distribution of the active number in system is given by*

$$\mathrm{P}\left(Q_\infty^{\mathrm{B}} = n\right) = \frac{\mathrm{P}\left(Q_\infty^{\eta} = n\right)}{1 - I_{\frac{\alpha}{\beta}}\left(c+1, \frac{\eta^*}{\alpha}\right)} = \frac{\Gamma\left(n + \frac{\eta^*}{\alpha}\right)\left(\frac{\beta-\alpha}{\beta}\right)^{\frac{\eta^*}{\alpha}}\left(\frac{\alpha}{\beta}\right)^n}{\Gamma\left(\frac{\eta^*}{\alpha}\right) n! \left(1 - I_{\frac{\alpha}{\beta}}\left(c+1, \frac{\eta^*}{\alpha}\right)\right)}, \tag{A.27}$$

for $0 \le n \le c$ and 0 otherwise, where $\mathrm{P}\left(Q_\infty^{\eta} = n\right)$ is as stated in Theorem 4.2.2. Furthermore, the mean and variance of the number in system are given by

$$\mathrm{E}\left[Q_\infty^{\mathrm{B}}\right] = \frac{\eta_\infty}{\beta}\left(\frac{1 - I_{\frac{\alpha}{\beta}}\left(c, \frac{\eta^*+\alpha}{\alpha}\right)}{1 - I_{\frac{\alpha}{\beta}}\left(c+1, \frac{\eta^*}{\alpha}\right)}\right), \tag{A.28}$$

$$\begin{aligned}\mathrm{Var}\left(Q_\infty^{\mathrm{B}}\right) = {} & \frac{\eta_\infty}{\beta}\left(\frac{\eta_\infty}{\beta} + \frac{\alpha}{\beta-\alpha}\right)\left(\frac{1 - I_{\frac{\alpha}{\beta}}\left(c-1, \frac{\eta^*+2\alpha}{\alpha}\right)}{1 - I_{\frac{\alpha}{\beta}}\left(c+1, \frac{\eta^*}{\alpha}\right)}\right) - \frac{\eta_\infty^2}{\beta^2}\left(\frac{1 - I_{\frac{\alpha}{\beta}}\left(c, \frac{\eta^*+\alpha}{\alpha}\right)}{1 - I_{\frac{\alpha}{\beta}}\left(c+1, \frac{\eta^*}{\alpha}\right)}\right)^2 \\ & + \frac{\eta_\infty}{\beta}\left(\frac{1 - I_{\frac{\alpha}{\beta}}\left(c, \frac{\eta^*+\alpha}{\alpha}\right)}{1 - I_{\frac{\alpha}{\beta}}\left(c+1, \frac{\eta^*}{\alpha}\right)}\right),\end{aligned} \tag{A.29}$$

where $\eta_\infty = \frac{\beta\eta^*}{\beta-\alpha}$ *and* $I_z(a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\int_0^z x^{a-1}(1-x)^{b-1}\mathrm{d}x$ *for* $z \in [0,1]$, $a > 0$ *and* $b > 0$ *is the regularized incomplete beta function.*

Proof. To show each of these, we first note that for $k \in \mathbb{Z}^+$, $x > 0$, and $p \in (0,1)$,

$$\sum_{n=0}^{k} \frac{\Gamma(n+x)}{\Gamma(x)\,n!}(1-p)^x p^n = 1 - I_p(k+1, x). \tag{A.30}$$

Hence, we can use Equation A.30 to see that

$$\sum_{n=0}^{c} \mathrm{P}(Q_\infty^\eta = n) = \sum_{n=0}^{c} \frac{\Gamma\left(n+\frac{\eta^*}{\alpha}\right)}{\Gamma\left(\frac{\eta^*}{\alpha}\right)n!}\left(\frac{\beta-\alpha}{\beta}\right)^{\frac{\eta^*}{\alpha}}\left(\frac{\alpha}{\beta}\right)^n = 1 - I_{\frac{\alpha}{\beta}}\left(c+1, \frac{\eta^*}{\alpha}\right).$$

Because the ESEP is a birth-death process it is reversible. Thus, by truncation we achieve the steady-state distribution, see e.g. Corollary 1.10 in Kelly (2011). Then, the steady-state mean of the number in system is given by

$$\begin{aligned}
\mathrm{E}\left[Q_\infty^B\right] &= \sum_{n=1}^{c} \frac{n\Gamma\left(n+\frac{\eta^*}{\alpha}\right)\left(\frac{\beta-\alpha}{\beta}\right)^{\frac{\eta^*}{\alpha}}\left(\frac{\alpha}{\beta}\right)^n}{\Gamma\left(\frac{\eta^*}{\alpha}\right)n!\left(1-I_{\frac{\alpha}{\beta}}\left(c+1,\frac{\eta^*}{\alpha}\right)\right)} \\
&= \frac{\frac{\eta^*}{\beta-\alpha}}{1-I_{\frac{\alpha}{\beta}}\left(c+1,\frac{\eta^*}{\alpha}\right)} \sum_{n=1}^{c} \frac{\Gamma\left(n-1+\frac{\eta^*+\alpha}{\alpha}\right)\left(\frac{\beta-\alpha}{\beta}\right)^{\frac{\eta^*+\alpha}{\alpha}}\left(\frac{\alpha}{\beta}\right)^{n-1}}{\Gamma\left(\frac{\eta^*+\alpha}{\alpha}\right)(n-1)!} \\
&= \frac{\eta_\infty}{\beta}\left(\frac{1-I_{\frac{\alpha}{\beta}}\left(c,\frac{\eta^*+\alpha}{\alpha}\right)}{1-I_{\frac{\alpha}{\beta}}\left(c+1,\frac{\eta^*}{\alpha}\right)}\right),
\end{aligned}$$

where we have again used Equation A.30 to simplify the summation. Likewise, the second moment in steady-state can be written

$$\begin{aligned}
\mathrm{E}\left[\left(Q_\infty^B\right)^2\right] &= \sum_{n=1}^{c} \frac{n^2\Gamma\left(n+\frac{\eta^*}{\alpha}\right)\left(\frac{\beta-\alpha}{\beta}\right)^{\frac{\eta^*}{\alpha}}\left(\frac{\alpha}{\beta}\right)^n}{\Gamma\left(\frac{\eta^*}{\alpha}\right)n!\left(1-I_{\frac{\alpha}{\beta}}\left(c+1,\frac{\eta^*}{\alpha}\right)\right)} \\
&= \sum_{n=2}^{c} \frac{\Gamma\left(n+\frac{\eta^*}{\alpha}\right)\left(\frac{\beta-\alpha}{\beta}\right)^{\frac{\eta^*}{\alpha}}\left(\frac{\alpha}{\beta}\right)^n}{\Gamma\left(\frac{\eta^*}{\alpha}\right)(n-2)!\left(1-I_{\frac{\alpha}{\beta}}\left(c+1,\frac{\eta^*}{\alpha}\right)\right)} + \sum_{n=1}^{c} \frac{\Gamma\left(n+\frac{\eta^*}{\alpha}\right)\left(\frac{\beta-\alpha}{\beta}\right)^{\frac{\eta^*}{\alpha}}\left(\frac{\alpha}{\beta}\right)^n}{\Gamma\left(\frac{\eta^*}{\alpha}\right)(n-1)!\left(1-I_{\frac{\alpha}{\beta}}\left(c+1,\frac{\eta^*}{\alpha}\right)\right)} \\
&= \frac{\eta^*(\eta^*+\alpha)}{(\beta-\alpha)^2} \sum_{n=2}^{c} \frac{\Gamma\left(n-2+\frac{\eta^*+2\alpha}{\alpha}\right)\left(\frac{\beta-\alpha}{\beta}\right)^{\frac{\eta^*+2\alpha}{\alpha}}\left(\frac{\alpha}{\beta}\right)^{n-2}}{\Gamma\left(\frac{\eta^*+2\alpha}{\alpha}\right)(n-2)!\left(1-I_{\frac{\alpha}{\beta}}\left(c+1,\frac{\eta^*}{\alpha}\right)\right)} + \mathrm{E}\left[Q_\infty^{\mathrm{B}}\right] \\
&= \frac{\eta_\infty}{\beta}\left(\frac{\eta_\infty}{\beta} + \frac{\alpha}{\beta-\alpha}\right)\frac{1-I_{\frac{\alpha}{\beta}}\left(c-1,\frac{\eta^*+2\alpha}{\alpha}\right)}{1-I_{\frac{\alpha}{\beta}}\left(c+1,\frac{\eta^*}{\alpha}\right)} + \mathrm{E}\left[Q_\infty^{\mathrm{B}}\right]
\end{aligned}$$

where once more these sums have been simplified through Equation A.30. □

As a demonstration of these findings, we now plot both the steady-state distribution and the mean and variance of this blocking system in Figure A.3. As can be observed in the figure, this system remains over-dispersed even when truncated. We can observe further that this holds in generality as follows. To observe this, we state two known properties of the regularized incomplete beta function:

$$I_z(a,b) = I_z(a+1,b) + \frac{z^a(1-z)^b}{aB(a,b)}, \qquad I_z(a,b+1) = I_z(a,b) + \frac{z^a(1-z)^b}{bB(a,b)}, \tag{A.31}$$

where $B(a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}$ is the beta function. Using these together, we can observe that

$$I_z(a,b) > I_z(a+1,b-1).$$

Thus, we can see that $I_{\frac{\alpha}{\beta}}\left(c+1, \frac{\eta^*}{\alpha}\right) < I_{\frac{\alpha}{\beta}}\left(c, \frac{\eta^*+\alpha}{\alpha}\right) < I_{\frac{\alpha}{\beta}}\left(c-1, \frac{\eta^*+2\alpha}{\alpha}\right) < 1$, and this implies

$$1 > \frac{1 - I_{\frac{\alpha}{\beta}}\left(c, \frac{\eta^*+\alpha}{\alpha}\right)}{1 - I_{\frac{\alpha}{\beta}}\left(c+1, \frac{\eta^*}{\alpha}\right)} > \frac{1 - I_{\frac{\alpha}{\beta}}\left(c-1, \frac{\eta^*+2\alpha}{\alpha}\right)}{1 - I_{\frac{\alpha}{\beta}}\left(c+1, \frac{\eta^*}{\alpha}\right)}.$$

We now note that the variance is written as the sum of the mean and a positive term and is thus over-dispersed.

We can also note that in the classical Erlang-B model, the famous "Poisson arrivals see time averages" (PASTA) result implies that the steady-state fraction of arrivals that are blocked is equal to the probability that the queue is at capacity in steady-state, see Wolff (1982). This is not so for the ESEP-B, as the arrival rate is state-dependent and, more specifically, increases with the queue length. However, in Proposition A.3.2 we find that an equivalent result holds asymptotically as the baseline intensity and the capacity grow large simultaneously. We note that large baseline intensity and capacity are realistic scenarios for many

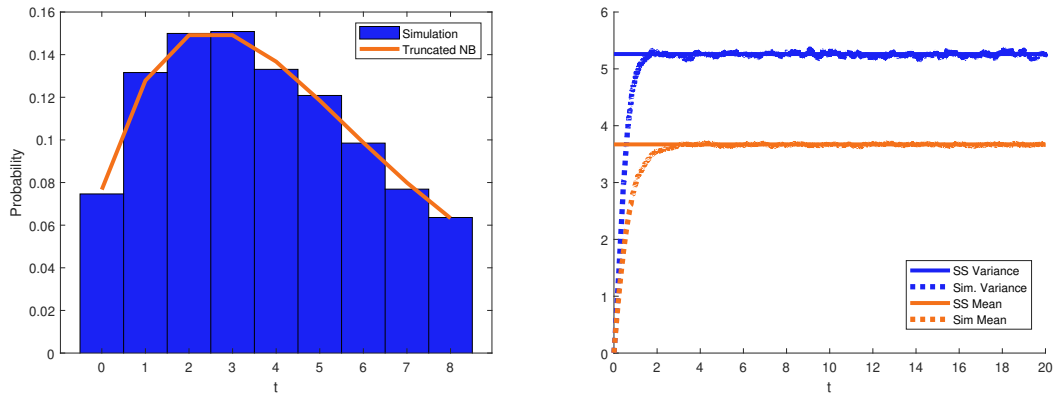


Figure A.3: Steady-state distribution (left) and mean and variance (right) of the ESEP-B with $\eta^* = 5$, $\alpha = 2$, $\beta = 3$, and $c = 8$ (Right), based on 10,000 replications.

practically relevant applications, including the aforementioned website crashing scenario.

Proposition A.3.2. *Let $\eta_t^{\mathrm{B}} = \eta^* + \alpha Q_t^{\mathrm{B}}$ be an ESEP-B, with baseline intensity $\eta^* > 0$, intensity jump $\alpha > 0$, exponential service rate $\beta > \alpha$, and capacity $c \in \mathbb{Z}^+$. Then, the fraction of arrivals in steady-state that are blocked π_{B} is given by*

$$\pi_{\mathrm{B}} = \frac{(\eta^* + \alpha c)\mathrm{P}\left(Q_\infty^\eta = c\right)}{\sum_{k=0}^{c}(\eta^* + \alpha k)\mathrm{P}\left(Q_\infty^\eta = k\right)} = \frac{(\eta^* + \alpha c)\mathrm{P}\left(Q_\infty^{\mathrm{B}} = c\right)}{\eta^* + \alpha \mathrm{E}\left[Q_\infty^{\mathrm{B}}\right]}, \tag{A.32}$$

where $\mathrm{P}\left(Q_\infty^\eta = k\right)$ is as given in Theorem 4.2.2 and $\mathrm{P}\left(Q_\infty^{\mathrm{B}} = c\right)$ and $\mathrm{E}\left[Q_\infty^{\mathrm{B}}\right]$ are as given in Proposition A.3.1. Moreover, if the baseline intensity and the capacity are redefined to be $\eta^* n$ and cn for $n \in \mathbb{Z}^+$, then

$$\frac{\pi_{\mathrm{B}}}{\mathrm{P}\left(Q_\infty^{\mathrm{B}} = c\right)} \longrightarrow 1, \tag{A.33}$$

as $n \to \infty$.

Proof. The expression for steady-state fraction of arrivals blocked π_{B} in Equation A.32 follows as a direct consequence from observing that the $\eta^* + \alpha k$ is the arrival rate when the queue is in state k. We are thus left to show Equation A.33.

By use of Equation A.32, we have that the ratio of π_B and $\mathrm{P}\left(Q_\infty^{\mathrm{B}} = c\right)$ is

$$\frac{\pi_{\mathrm{B}}}{\mathrm{P}\left(Q_\infty^{\mathrm{B}} = c\right)} = \frac{\eta^* + \alpha c}{\eta^* + \alpha \mathrm{E}\left[Q_\infty^{\mathrm{B}}\right]} = \frac{\eta^* + \alpha c}{\eta^* + \frac{\alpha\eta^*}{\beta-\alpha}\left(\frac{1-I_{\frac{\alpha}{\beta}}\left(c, \frac{\eta^*}{\alpha}+1\right)}{1-I_{\frac{\alpha}{\beta}}\left(c+1, \frac{\eta^*}{\alpha}\right)}\right)},$$

by use of Proposition A.3.1. Substituting in the scaled forms of the baseline intensity and capacity $\eta^* n$ and cn and then dividing the numerator and denominator by cn, this is

$$\frac{\eta^* n + \alpha c n}{\eta^* n + \frac{\alpha\eta^* n}{\beta-\alpha}\left(\frac{1-I_{\frac{\alpha}{\beta}}\left(cn, \frac{\eta^* n}{\alpha}+1\right)}{1-I_{\frac{\alpha}{\beta}}\left(cn+1, \frac{\eta^* n}{\alpha}\right)}\right)} = \frac{\frac{\eta^*}{c} + \alpha}{\frac{\eta^*}{c} + \frac{\eta^*}{c}\left(\frac{\alpha}{\beta-\alpha}\right)\left(\frac{1-I_{\frac{\alpha}{\beta}}\left(cn, \frac{\eta^* n}{\alpha}+1\right)}{1-I_{\frac{\alpha}{\beta}}\left(cn+1, \frac{\eta^* n}{\alpha}\right)}\right)}.$$

From the definition and symmetry of the regularized incomplete beta function, we can note that the ratio of these functions is such that

$$\frac{1 - I_{\frac{\alpha}{\beta}}\left(cn, \frac{\eta^* n}{\alpha} + 1\right)}{1 - I_{\frac{\alpha}{\beta}}\left(cn + 1, \frac{\eta^* n}{\alpha}\right)} = \frac{I_{1-\frac{\alpha}{\beta}}\left(\frac{\eta^* n}{\alpha} + 1, cn\right)}{I_{1-\frac{\alpha}{\beta}}\left(\frac{\eta^* n}{\alpha}, cn + 1\right)} = \frac{\alpha c}{\eta^*}\left(\frac{\int_0^{1-\frac{\alpha}{\beta}} x^{\frac{n\eta^*}{\alpha}} (1-x)^{cn-1}\,\mathrm{d}x}{\int_0^{1-\frac{\alpha}{\beta}} x^{\frac{n\eta^*}{\alpha}-1} (1-x)^{cn}\,\mathrm{d}x}\right).$$

We can now recognize an identity for the hypergeometric function ${}_2F_1(a, b; c; z)$, and thus re-express this ratio as

$$\begin{aligned}
\frac{\alpha c}{\eta^*}\left(\frac{\int_0^{1-\frac{\alpha}{\beta}} x^{\frac{n\eta^*}{\alpha}} (1-x)^{cn-1}\,\mathrm{d}x}{\int_0^{1-\frac{\alpha}{\beta}} x^{\frac{n\eta^*}{\alpha}-1} (1-x)^{cn}\,\mathrm{d}x}\right) &= \frac{\alpha c}{\eta^*}\left(\frac{\frac{1}{\frac{\eta^* n}{\alpha}+1}\left(1-\frac{\alpha}{\beta}\right)^{\frac{\eta^* n}{\alpha}+1}\left(\frac{\alpha}{\beta}\right)^{cn} {}_2F_1\left(c + \frac{\eta^* n}{\alpha} + 1, 1; cn + 2; 1 - \frac{\alpha}{\beta}\right)}{\frac{1}{\frac{\eta^* n}{\alpha}}\left(1-\frac{\alpha}{\beta}\right)^{\frac{\eta^* n}{\alpha}}\left(\frac{\alpha}{\beta}\right)^{cn+1} {}_2F_1\left(c + \frac{\eta^* n}{\alpha} + 1, 1; cn + 1; 1 - \frac{\alpha}{\beta}\right)}\right) \\
&= \frac{\alpha c}{\eta^*}\left(\frac{\eta^* n}{\eta^* n + \alpha}\right)\left(\frac{\beta-\alpha}{\alpha}\right)\frac{{}_2F_1\left(c + \frac{\eta^* n}{\alpha} + 1, 1; cn + 2; 1 - \frac{\alpha}{\beta}\right)}{{}_2F_1\left(c + \frac{\eta^* n}{\alpha} + 1, 1; cn + 1; 1 - \frac{\alpha}{\beta}\right)}.
\end{aligned}$$

As $n \to \infty$, this yields

$$\frac{\alpha c}{\eta^*}\left(\frac{\eta^* n}{\eta^* n + \alpha}\right)\left(\frac{\beta-\alpha}{\beta}\right)\frac{{}_2F_1\left(c + \frac{\eta^* n}{\alpha} + 1, 1; cn + 2; 1 - \frac{\alpha}{\beta}\right)}{{}_2F_1\left(c + \frac{\eta^* n}{\alpha} + 1, 1; cn + 1; 1 - \frac{\alpha}{\beta}\right)} \longrightarrow \frac{\alpha c}{\eta^*}\left(\frac{\beta-\alpha}{\alpha}\right),$$

which thus implies that

$$\frac{\frac{\eta^*}{c} + \alpha}{\frac{\eta^*}{c} + \frac{\eta^*}{c}\left(\frac{\alpha}{\beta-\alpha}\right)\left(\frac{1-I_{\frac{\alpha}{\beta}}\left(cn, \frac{\eta^* n}{\alpha}+1\right)}{1-I_{\frac{\alpha}{\beta}}\left(cn+1, \frac{\eta^* n}{\alpha}\right)}\right)} \longrightarrow \frac{\frac{\eta^*}{c} + \alpha}{\frac{\eta^*}{c} + \frac{\eta^*}{c}\left(\frac{\alpha}{\beta-\alpha}\right)\frac{\alpha c}{\eta^*}\left(\frac{\beta-\alpha}{\alpha}\right)} = 1,$$

and this completes the proof. □

As an example of the convergence stated in Proposition A.3.2, we compare the probability of the system being at capacity and the fraction of blocked arrivals below in Figure A.4. In this figure, η^* and c are increased simultaneously according to a fixed ratio. Although at the initial values it is clear that a PASTA-esque result does not hold, as the baseline intensity and capacity both increase one can see that the two curves tend toward one another in each of the different parameter settings.

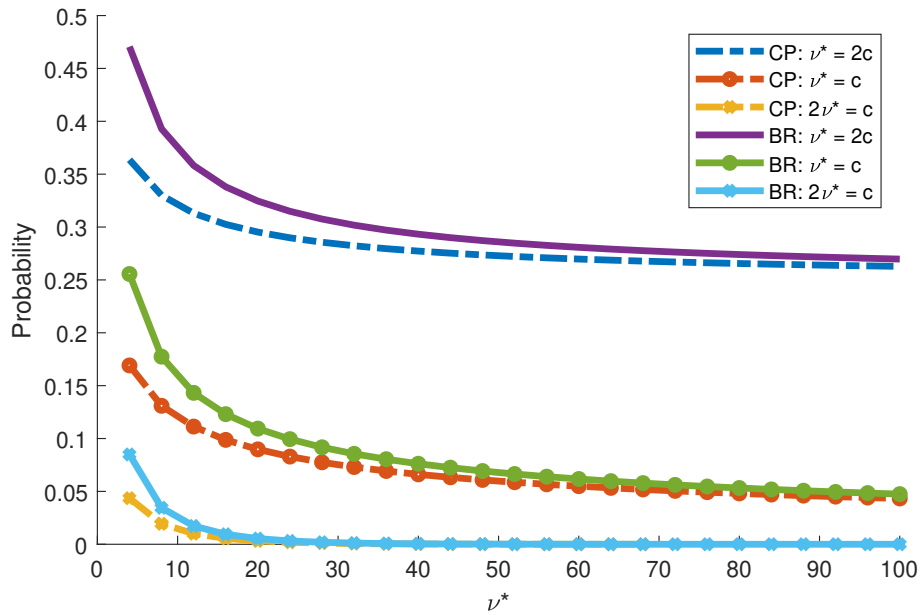


Figure A.4: Comparison of the ratio of blocked arrivals (BR) and the probability of system being at capacity (CP) when increasing η^* and c simultaneously, where $\alpha = 2$ and $\beta = 3$.

A.4 Proof of Proposition 4.2.4

Proof. Using Proposition A.1.3, we proceed through use of exponential identities for the hyperbolic functions. Specifically, we will make use of the following:

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}, \tag{A.34}$$

$$\cosh(x) = \frac{e^x + e^{-x}}{2}, \tag{A.35}$$

and

$$\tanh^{-1}(x) = \frac{1}{2}\log\left(\frac{1+x}{1-x}\right). \tag{A.36}$$

Using these identities we can further observe that

$$\cosh\left(\tanh^{-1}(x)\right) = \frac{e^{\tanh^{-1}(x)} + e^{-\tanh^{-1}(x)}}{2} = \frac{\left(\frac{1+x}{1-x}\right)^{\frac{1}{2}} + \left(\frac{1-x}{1+x}\right)^{\frac{1}{2}}}{2}.$$

Now, for any time $t \geq 0$ we can note that $N_t = Q_t + D_t$. Thus, we have that

$$\mathrm{E}\left[z^{N_t}\right] = \mathrm{E}\left[z^{Q_t} z^{D_t}\right] = G(z, z, t),$$

where $G(z_1, z_2, t)$ is as given in Proposition A.1.3. Setting $z_1 = z_2 = z$ and $D_0 = N_0 - Q_0$, this is

$$\begin{aligned} G(z,z,t) = {} & z^{N_0 - Q_0} e^{\frac{\eta^*(\beta-\alpha)}{2\alpha}t} \left(1 - \left(\tanh\left(\frac{t}{2}\sqrt{(\beta+\alpha)^2 - 4\alpha\beta z} + \tanh^{-1}\left(\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2 - 4\alpha\beta z}}\right)\right)\right)^2\right)^{\frac{\eta^*}{2\alpha}} \\ & \cdot \left(\frac{\beta+\alpha}{2\alpha} - \frac{\sqrt{(\beta+\alpha)^2 - 4\alpha\beta z}}{2\alpha}\tanh\left(\frac{t}{2}\sqrt{(\beta+\alpha)^2 - 4\alpha\beta z} + \tanh^{-1}\left(\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2 - 4\alpha\beta z}}\right)\right)\right)^{Q_0} \\ & \cdot \left(\cosh\left(\tanh^{-1}\left(\frac{2\alpha z - \beta - \alpha}{\sqrt{(\beta+\alpha)^2 - 4\alpha\beta z}}\right)\right)\right)^{\frac{\eta^*}{\alpha}}. \end{aligned} \tag{A.37}$$

Using the hyperbolic identities and simplifying, this is

$$G(z,z,t) = z^{N_0-Q_0} e^{\frac{\eta^*\eta^*(\beta-\alpha)}{2\alpha}t} \left(\frac{2e^{\frac{t}{2}\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}{1 - \frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}} + \left(1 + \frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}\right) e^{t\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}} \right)^{\frac{\eta^*}{\alpha}}$$

$$\cdot \left(\frac{\beta+\alpha}{2\alpha} + \frac{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}{2\alpha} \left(\frac{1 - \frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}} - \left(1 + \frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}\right) e^{t\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}{1 - \frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}} + \left(1 + \frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}\right) e^{t\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}} \right) \right)^{Q_0},$$

which is the stated result. However, the simplifications used to reach this form require multiple parts and several steps and so we can these individually now. We start with the hyperbolic tangent function that appears on the first and second lines of Equation A.37. Using Equations A.34 and A.36, this is

$$\begin{aligned}
&- \tanh\left(\frac{t}{2}\sqrt{(\beta+\alpha)^2-4\alpha\beta z} + \tanh^{-1}\left(\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}} \right) \right) \\
&= -\frac{e^{\left(\frac{t}{2}\sqrt{(\beta+\alpha)^2-4\alpha\beta z}+\frac{1}{2}\log\left(\frac{1+\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}{1-\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}\right)\right)} - e^{-\left(\frac{t}{2}\sqrt{(\beta+\alpha)^2-4\alpha\beta z}+\frac{1}{2}\log\left(\frac{1+\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}{1-\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}\right)\right)}}{e^{\left(\frac{t}{2}\sqrt{(\beta+\alpha)^2-4\alpha\beta z}+\frac{1}{2}\log\left(\frac{1+\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}{1-\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}\right)\right)} + e^{-\left(\frac{t}{2}\sqrt{(\beta+\alpha)^2-4\alpha\beta z}+\frac{1}{2}\log\left(\frac{1+\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}{1-\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}\right)\right)}} \\
&= -\frac{e^{\left(t\sqrt{(\beta+\alpha)^2-4\alpha\beta z}+\log\left(\frac{1+\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}{1-\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}\right)\right)} - 1}{e^{\left(t\sqrt{(\beta+\alpha)^2-4\alpha\beta z}+\log\left(\frac{1+\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}{1-\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}\right)\right)} + 1} \\
&= \frac{1 - \left(\frac{1+\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}{1-\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}\right) e^{t\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}{1 + \left(\frac{1+\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}{1-\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}\right) e^{t\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}} \\
&= \frac{1 - \frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}} - \left(1 + \frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}\right) e^{t\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}{1 - \frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}} + \left(1 + \frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}\right) e^{t\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}.
\end{aligned}$$

Thus, the second line of Equation A.37 simplifies as

$$\left(\frac{\beta+\alpha}{2\alpha} - \frac{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}{2\alpha}\tanh\left(\frac{t}{2}\sqrt{(\beta+\alpha)^2-4\alpha\beta z} + \tanh^{-1}\left(\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}\right)\right)\right)^{Q_0}$$

$$= \left(\frac{\beta+\alpha}{2\alpha} + \frac{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}{2\alpha}\left(\frac{1-\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}} - \left(1+\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}\right)e^{t\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}{1-\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}} + \left(1+\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}\right)e^{t\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}\right)\right)^{Q_0}$$

$$= \left(\frac{\beta+\alpha}{2\alpha} + \frac{\frac{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}{2\alpha} - \frac{\beta+\alpha-2\alpha z}{2\alpha} - \left(\frac{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}{2\alpha} + \frac{\beta+\alpha-2\alpha z}{2\alpha}\right)e^{t\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}{1-\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}} + \left(1+\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}\right)e^{t\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}\right)^{Q_0}$$

$$= \left(\frac{\frac{(\beta+\alpha)^2-2\beta\alpha z-2\alpha^2 z}{2\alpha\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}\left(e^{t\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}-1\right) + \frac{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}{2\alpha} + z - \left(\frac{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}{2\alpha} - z\right)e^{t\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}{1-\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}} + \left(1+\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}\right)e^{t\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}\right)^{Q_0}$$

$$= \left(\frac{\left(\frac{(\beta-\alpha)z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}\right)\left(e^{t\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}-1\right) + z + ze^{t\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}{1-\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}} + \left(1+\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}\right)e^{t\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}\right)^{Q_0}.$$

Following the same approach, the first line of Equation A.37 rearranges to

$$\left(1-\left(\frac{e^{\left(\frac{t}{2}\sqrt{(\beta+\alpha)^2-4\alpha\beta z}+\frac{1}{2}\log\left(\frac{1+\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}{1-\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}\right)\right)}-e^{-\left(\frac{t}{2}\sqrt{(\beta+\alpha)^2-4\alpha\beta z}+\frac{1}{2}\log\left(\frac{1+\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}{1-\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}\right)\right)}}{e^{\left(\frac{t}{2}\sqrt{(\beta+\alpha)^2-4\alpha\beta z}+\frac{1}{2}\log\left(\frac{1+\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}{1-\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}\right)\right)}+e^{-\left(\frac{t}{2}\sqrt{(\beta+\alpha)^2-4\alpha\beta z}+\frac{1}{2}\log\left(\frac{1+\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}{1-\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}\right)\right)}}\right)^2\right)^{\frac{\eta^*}{2\alpha}}$$

$$=\left(1-\left(\frac{1-\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}-\left(1+\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}\right)e^t\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}{1-\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}+\left(1+\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}\right)e^t\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}\right)^2\right)^{\frac{\eta^*}{2\alpha}}$$

$$=\left(\frac{4\left(1-\frac{(\beta+\alpha-2\alpha z)^2}{(\beta+\alpha)^2-4\alpha\beta z}\right)e^t\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}{\left(1-\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}+\left(1+\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}\right)e^t\sqrt{(\beta+\alpha)^2-4\alpha\beta z}\right)^2}\right)^{\frac{\eta^*}{2\alpha}}$$

$$=\left(\frac{\frac{4\alpha\sqrt{z-z^2}}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}e^{\frac{t}{2}\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}{1-\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}+\left(1+\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}\right)e^t\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}\right)^{\frac{\eta^*}{\alpha}}.$$

Finally, the third line of Equation A.37 is simplified through use of Equations A.35 and A.36. This expression is then given by

$$\left(\frac{\left(\frac{1+\frac{2\alpha z-\beta-\alpha}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}{1-\frac{2\alpha z-\beta-\alpha}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}\right)^{\frac{1}{2}}+\left(\frac{1-\frac{2\alpha z-\beta-\alpha}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}{1+\frac{2\alpha z-\beta-\alpha}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}\right)^{\frac{1}{2}}}{2}\right)^{\frac{\eta^*}{\alpha}}=\left(\left(\frac{\left(\frac{1-\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}{1+\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}\right)^{\frac{1}{2}}+\left(\frac{1+\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}{1-\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}\right)^{\frac{1}{2}}}{2}\right)^2\right)^{\frac{\eta^*}{2\alpha}}$$

$$=\left(\frac{\frac{1-\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}{1+\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}+2+\frac{1+\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}{1-\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}}{4}\right)^{\frac{\eta^*}{2\alpha}}$$

$$=\left(\frac{\frac{1-\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}{1+\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}+\frac{1+\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}{1+\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}+\frac{1-\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}{1-\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}+\frac{1+\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}{1-\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}}{4}\right)^{\frac{\eta^*}{2\alpha}}$$

$$=\left(\frac{1-\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}+1+\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}{2\left(1-\frac{(\beta+\alpha-2\alpha z)^2}{(\beta+\alpha)^2-4\alpha\beta z}\right)}\right)^{\frac{\eta^*}{2\alpha}}$$

$$=\left(\frac{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}{2\alpha\sqrt{z-z^2}}\right)^{\frac{\eta^*}{\alpha}}.$$

Together these forms give the stated result. □

APPENDIX B

ADDENDUM FOR CHAPTER 5

B.1 Proof of Proposition 5.2.1

Proof. For clarity's sake and ease of reference, we will also enumerate the proofs of each statement.

i) Suppose $\mathbf{X}_n$ and $\mathbf{Y}_n$ are each Matryoshkan matrices. Then, by Equation 5.1, we have that

$$\mathbf{X}_n + \mathbf{Y}_n = \begin{bmatrix} \mathbf{X}_{n-1} & \mathbf{0}_{n-1\times 1} \\ \mathbf{x}_n & x_{n,n} \end{bmatrix} + \begin{bmatrix} \mathbf{Y}_{n-1} & \mathbf{0}_{n-1\times 1} \\ \mathbf{y}_n & y_{n,n} \end{bmatrix} = \begin{bmatrix} \mathbf{X}_{n-1} + \mathbf{Y}_{n-1} & \mathbf{0}_{n-1\times 1} \\ \mathbf{x}_n + \mathbf{y}_n & x_{n,n} + y_{n,n} \end{bmatrix},$$

and

$$\mathbf{X}_n\mathbf{Y}_n = \begin{bmatrix} \mathbf{X}_{n-1} & \mathbf{0}_{n-1\times 1} \\ \mathbf{x}_n & x_{n,n} \end{bmatrix} \begin{bmatrix} \mathbf{Y}_{n-1} & \mathbf{0}_{n-1\times 1} \\ \mathbf{y}_n & y_{n,n} \end{bmatrix} = \begin{bmatrix} \mathbf{X}_{n-1}\mathbf{Y}_{n-1} & \mathbf{0}_{n-1\times 1} \\ \mathbf{x}_n\mathbf{Y}_{n-1} + x_{n,n}\mathbf{y}_n & x_{n,n}y_{n,n} \end{bmatrix}.$$

We can now again invoke Equation 5.1 to observe that these forms satisfy this definition and thus are also Matryoshkan matrices.

ii) Let $\mathbf{M}_n \in \mathbb{R}^{n\times n}$ be a Matryoshkan matrix with all non-zero diagonal elements $m_{i,i}$ for $i \in \{1, \ldots, n\}$. By definition $\mathbf{M}_n$ is lower triangular and hence its eigenvalues are on its diagonal. Since all the eigenvalues are non-zero by assumption, $\mathbf{M}_n$ is invertible. Moreover, it is known that the inverse of a lower triangular matrix is lower triangular as well. Thus, we will now solve for lower triangular matrix $\mathbf{W}_n \in \mathbb{R}^{n\times n}$ such that $\mathbf{I}_n = \mathbf{M}_n\mathbf{W}_n$ where $\mathbf{I}_n \in \mathbb{R}^{n\times n}$ is the identity. This can be written

$$\begin{bmatrix} \mathbf{I}_{n-1} & \mathbf{0}_{n-1\times 1} \\ \mathbf{0}_{1\times n-1} & 1 \end{bmatrix} = \mathbf{I}_n = \mathbf{M}_n\mathbf{W}_n = \begin{bmatrix} \mathbf{M}_{n-1} & \mathbf{0}_{n-1\times 1} \\ \mathbf{m}_n & m_{n,n} \end{bmatrix} \begin{bmatrix} \mathbf{A} & \mathbf{0}_{n-1\times 1} \\ \mathbf{b} & c \end{bmatrix},$$

where $\mathbf{A} \in \mathbb{R}^{n-1\times n-1}$, $b \in \mathbb{R}^{1\times n-1}$, and $c \in \mathbb{R}$. Because $m_{i,i} \neq 0$ for all $i \in \{1, \ldots, n-1\}$, we also know that $\mathbf{M}_{n-1}$ is non-singular. Thus, we can see that $\mathbf{A} = \mathbf{M}_{n-1}^{-1}$ from $\mathbf{M}_{n-1}\mathbf{A} = \mathbf{I}_{n-1}$. Likewise, $cm_{n,n} = 1$ implies $c = \frac{1}{m_{n,n}}$. Then, we have that

$$\mathbf{0}_{1\times n-1} = \mathbf{m}_n\mathbf{A} + m_{n,n}\mathbf{b} = \mathbf{m}_n\mathbf{M}_{n-1}^{-1} + m_{n,n}\mathbf{b},$$

and so $\mathbf{b} = -\frac{1}{m_{n,n}}\mathbf{m}_n\mathbf{M}_{n-1}^{-1}$. This completes the solution for $\mathbf{W}_n$, and hence provides the inverse of $\mathbf{M}_n$.

iii) To begin, we will prove that

$$\mathbf{M}_n^k = \begin{bmatrix} \mathbf{M}_{n-1}^k & \mathbf{0}_{n-1\times 1} \\ \mathbf{m}_n \sum_{j=0}^{k-1} \mathbf{M}_{n-1}^j m_{n,n}^{k-1-j} & m_{n,n}^k \end{bmatrix}$$

for $k \in \mathbb{Z}^+$. We proceed by induction. The base case, $k = 1$, holds by definition. Therefore we suppose that the hypothesis holds at k. Then, at $k+1$ we can observe that

$$\begin{aligned}
\mathbf{M}_n^{k+1} &= \mathbf{M}_n\mathbf{M}_n^k \\
&= \begin{bmatrix} \mathbf{M}_{n-1} & \mathbf{0}_{n-1\times 1} \\ \mathbf{m}_n & m_{n,n} \end{bmatrix} \begin{bmatrix} \mathbf{M}_{n-1}^k & \mathbf{0}_{n-1\times 1} \\ \mathbf{m}_n \sum_{j=0}^{k-1} \mathbf{M}_{n-1}^j m_{n,n}^{k-1-j} & m_{n,n}^k \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{M}_{n-1}^{k+1} & \mathbf{0}_{n-1\times 1} \\ \mathbf{m}_n\mathbf{M}_{n-1}^k + \mathbf{m}_n \sum_{j=0}^{k-1} \mathbf{M}_{n-1}^j m_{n,n}^{k-j} & m_{n,n}^{k+1} \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{M}_{n-1}^{k+1} & \mathbf{0}_{n-1\times 1} \\ \mathbf{m}_n \sum_{j=0}^{k} \mathbf{M}_{n-1}^j m_{n,n}^{k-j} & m_{n,n}^{k+1} \end{bmatrix},
\end{aligned}$$

which completes the induction. We now observe further that for matrices $\mathbf{A} \in \mathbb{R}^{n\times n}$ and $\mathbf{B} \in \mathbb{R}^{n\times n}$ such that $\mathbf{AB} = \mathbf{BA}$ and $\mathbf{A} - \mathbf{B}$ is non-singular,

$$\sum_{j=0}^{k-1} \mathbf{A}^j\mathbf{B}^{k-1-j} = (\mathbf{A} - \mathbf{B})^{-1}\left(\mathbf{A}^k - \mathbf{B}^k\right).$$

This relationship can verified by multiplying the left-hand side by $\mathbf{A} - \mathbf{B}$:

$$(\mathbf{A} - \mathbf{B}) \sum_{j=0}^{k-1} \mathbf{A}^j \mathbf{B}^{k-1-j} = \sum_{j=0}^{k-1} \mathbf{A}^{j+1} \mathbf{B}^{k-1-j} - \sum_{j=0}^{k-1} \mathbf{A}^j \mathbf{B}^{k-j} = \mathbf{A}^k - \mathbf{B}^k.$$

This allows us to observe that

$$\mathbf{M}_n^k = \begin{bmatrix} \mathbf{M}_{n-1}^k & \mathbf{0}_{n-1\times 1} \\ \mathbf{m}_n(\mathbf{M}_{n-1} - m_{n,n}\mathbf{I})^{-1}\left(\mathbf{M}_{n-1}^k - m_{n,n}^k \mathbf{I}\right) & m_{n,n}^k \end{bmatrix},$$

and thus

$$\begin{aligned} e^{\mathbf{M}_n t} &= \sum_{k=0}^{\infty} \frac{t^k \mathbf{M}_n^k}{k!} = \sum_{k=0}^{\infty} \frac{t^k}{k!} \begin{bmatrix} \mathbf{M}_{n-1}^k & \mathbf{0}_{n-1\times 1} \\ \mathbf{m}_n(\mathbf{M}_{n-1} - m_{n,n}\mathbf{I})^{-1}\left(\mathbf{M}_{n-1}^k - m_{n,n}^k \mathbf{I}\right) & m_{n,n}^k \end{bmatrix} \\ &= \begin{bmatrix} e^{\mathbf{M}_{n-1} t} & \mathbf{0}_{n-1\times 1} \\ \mathbf{m}_n(\mathbf{M}_{n-1} - m_{n,n}\mathbf{I})^{-1}\left(e^{\mathbf{M}_{n-1} t} - e^{m_{n,n} t}\mathbf{I}\right) & e^{m_{n,n} t} \end{bmatrix}, \end{aligned}$$

which completes the proof. Note that because $\mathbf{M}_{n-1}$ is triangular and because we have assumed $m_{1,1}, \ldots, m_{n,n}$ are distinct, we know that $\mathbf{M}_{n-1} - m_{n,n}\mathbf{I}$ is invertible.

iv) From the statement, we seek a matrix $\mathbf{A} \in \mathbb{R}^{n-1\times n-1}$, a row vector $\mathbf{b} \in \mathbb{R}^{1\times n-1}$, and scalar $c \in \mathbb{R}$ such that

$$\begin{bmatrix} \mathbf{M}_{n-1} & \mathbf{0}_{n-1\times 1} \\ \mathbf{m}_n & m_{n,n} \end{bmatrix} \begin{bmatrix} \mathbf{A} & \mathbf{0}_{n-1\times 1} \\ \mathbf{b} & c \end{bmatrix} = \begin{bmatrix} \mathbf{A} & \mathbf{0}_{n-1\times 1} \\ \mathbf{b} & c \end{bmatrix} \begin{bmatrix} \mathbf{D}_{n-1} & \mathbf{0}_{n-1\times 1} \\ \mathbf{0}_{1\times n-1} & m_{n,n} \end{bmatrix}$$

where $\mathbf{D}_{n-1} \in \mathbb{R}^{n-1\times n-1}$ is a diagonal matrix with values $m_{1,1}, \ldots, m_{n-1,n-1}$. From the triangular structure of $\mathbf{M}_n$, we know that $\mathbf{D}_n$ contains all the eigenvalues of $\mathbf{M}_n$. We will now solve the resulting sub-systems. From $\mathbf{M}_{n-1}\mathbf{A} = \mathbf{A}\mathbf{D}_{n-1}$, we take $\mathbf{A} = \mathbf{U}_{n-1}$. Substituting this forward, we see that

$$\mathbf{m}_n \mathbf{U}_{n-1} + m_{n,n}\mathbf{b} = \mathbf{m}_n \mathbf{A} + m_{n,n}\mathbf{b} = \mathbf{b}\mathbf{D}_{n-1}$$

and so $b = \mathbf{m}_n \mathbf{U}_{n-1}(\mathbf{D}_{n-1} - m_{n,n}\mathbf{I})^{-1}$, where as in step (iii) we are justified in inverting $\mathbf{D}_{n-1} - m_{n,n}\mathbf{I}$ due to the fact that $m_{1,1}, \ldots, m_{n,n}$ are distinct. Finally, we take $c = 1$, as any value will satisfy $cm_{n,n} = cm_{n,n}$.

□

B.2 Proof of Lemma 5.2.2

Proof. The vector solution in Equation 5.5 is known and is thus displayed for reference. Expanding this expression in bracket-notation form, by use of Proposition 5.2.1 this is

$$\begin{bmatrix} \mathbf{s}_{n-1}(t) \\ s_n(t) \end{bmatrix} = \begin{bmatrix} e^{\mathbf{M}_{n-1}t} & \mathbf{0}_{n-1\times 1} \\ \mathbf{m}_n \left(\mathbf{M}_{n-1} - m_{n,n}\mathbf{I}\right)^{-1}\left(e^{\mathbf{M}_{n-1}t} - e^{m_{n,n}t}\mathbf{I}\right) & e^{m_{n,n}t} \end{bmatrix} \begin{bmatrix} \mathbf{s}_{n-1}(0) \\ s_n(0) \end{bmatrix}$$
$$- \begin{bmatrix} \mathbf{M}_{n-1}^{-1} & \mathbf{0}_{n-1\times 1} \\ -\frac{1}{m_{n,n}}\mathbf{m}_n\mathbf{M}_{n-1}^{-1} & \frac{1}{m_{n,n}} \end{bmatrix} \begin{bmatrix} \mathbf{I} - e^{\mathbf{M}_{n-1}t} & \mathbf{0}_{n-1\times 1} \\ -\mathbf{m}_n \left(\mathbf{M}_{n-1} - m_{n,n}\mathbf{I}\right)^{-1}\left(e^{\mathbf{M}_{n-1}t} - e^{m_{n,n}t}\mathbf{I}\right) & 1 - e^{m_{n,n}t} \end{bmatrix} \begin{bmatrix} \mathbf{c}_{n-1} \\ c_n \end{bmatrix}.$$

Thus, we can find $s_n(t)$ by multiplying each left side of the equality by a unit row vector in the direction of the nth coordinate, which we denote $\mathbf{v}_n^{\mathrm{T}}$. This yields

$$\begin{aligned} s_n(t) &= \mathbf{v}_n^{\mathrm{T}} \begin{bmatrix} \mathbf{s}_{n-1}(t) \\ s_n(t) \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{m}_n \left(\mathbf{M}_{n-1} - m_{n,n}\mathbf{I}\right)^{-1}\left(e^{\mathbf{M}_{n-1}t} - e^{m_{n,n}t}\mathbf{I}\right) & e^{m_{n,n}t} \end{bmatrix} \begin{bmatrix} \mathbf{s}_{n-1}(0) \\ s_n(0) \end{bmatrix} \\ &\quad - \begin{bmatrix} -\frac{1}{m_{n,n}}\mathbf{m}_n\mathbf{M}_{n-1}^{-1} & \frac{1}{m_{n,n}} \end{bmatrix} \begin{bmatrix} \mathbf{I} - e^{\mathbf{M}_{n-1}t} & \mathbf{0}_{n-1\times 1} \\ -\mathbf{m}_n \left(\mathbf{M}_{n-1} - m_{n,n}\mathbf{I}\right)^{-1}\left(e^{\mathbf{M}_{n-1}t} - e^{m_{n,n}t}\mathbf{I}\right) & 1 - e^{m_{n,n}t} \end{bmatrix} \begin{bmatrix} \mathbf{c}_{n-1} \\ c_n \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{m}_n \left(\mathbf{M}_{n-1} - m_{n,n}\mathbf{I}\right)^{-1}\left(e^{\mathbf{M}_{n-1}t} - e^{m_{n,n}t}\mathbf{I}\right) & e^{m_{n,n}t} \end{bmatrix} \begin{bmatrix} \mathbf{s}_{n-1}(0) \\ s_n(0) \end{bmatrix} \\ &\quad - \begin{bmatrix} -\frac{1}{m_{n,n}}\mathbf{m}_n\mathbf{M}_{n-1}^{-1} & \frac{1}{m_{n,n}} \end{bmatrix} \begin{bmatrix} \left(\mathbf{I} - e^{\mathbf{M}_{n-1}t}\right)\mathbf{c}_{n-1} \\ -\mathbf{m}_n \left(\mathbf{M}_{n-1} - m_{n,n}\mathbf{I}\right)^{-1}\left(e^{\mathbf{M}_{n-1}t} - e^{m_{n,n}t}\mathbf{I}\right)\mathbf{c}_{n-1} + c_n(1 - e^{m_{n,n}t}) \end{bmatrix}. \end{aligned}$$

Then by taking these inner products, we receive

$$s_n(t) = \mathbf{m}_n \left(\mathbf{M}_{n-1} - m_{n,n}\mathbf{I}\right)^{-1} \left(e^{\mathbf{M}_{n-1}t} - e^{m_{n,n}t}\mathbf{I}\right) \mathbf{s}_{n-1}(0) + s_n(0)e^{m_{n,n}t} + \mathbf{m}_n \mathbf{M}_{n-1}^{-1} \left(\mathbf{I} - e^{\mathbf{M}_{n-1}t}\right) \frac{\mathbf{c}_{n-1}}{m_{n,n}}$$
$$+ \mathbf{m}_n \left(\mathbf{M}_{n-1} - m_{n,n}\mathbf{I}\right)^{-1} \left(e^{\mathbf{M}_{n-1}t} - e^{m_{n,n}t}\mathbf{I}\right) \frac{\mathbf{c}_{n-1}}{m_{n,n}} - \frac{c_n}{m_{n,n}} \left(1 - e^{m_{n,n}t}\right),$$

and this simplifies to the stated solution. □

APPENDIX C

ADDENDUM TO CHAPTER 6

C.1 Technical Lemmas and Proofs

To support our general batch analysis we will now introduce a technical lemma, which extends Sullivan et al. (1980) to a probabilistic context. In Sullivan et al. (1980), the authors use shifted, asymmetric Legendre polynomials to produce a sum of exponential functions of $x \geq 0$ that converges to the indicator function $\mathbf{1}\{x \leq c\}$ for any constant $c > 0$. These approximations make use of the generalized hypergeometric function ${}_3F_2(\cdot)$, which is defined

$$ {}_3F_2(a_1, a_2, a_3, b_1, b_2, x) = \sum_{i=0}^{\infty} \frac{(a_1)_i (a_2)_i (a_3)_i}{(b_1)_i (b_2)_i} \frac{x^i}{i!}, $$

where $(c)_i = \prod_{j=0}^{i-1}(c+j)$ is a rising factorial. By use of the dominated convergence theorem, in Lemma C.1.1 we generalize this result using a sum of moment generating functions of a continuous non-negative random variable. We find convergence to the cumulative distribution function of the random variable, as well as to the expectation of the product between the random variable and an indicator function. Therefore, this lemma provides a method to find this cumulative probability and expectation when one only has access to the moment generating function of the random variable. This is paramount to our staffing analysis, and because of its generality we believe it may also be of use in other applications. For clarity's sake, we note that the moment generating functions used in this technique are for strictly negative space parameters and thus will exist for all distributions. These functions can thus be viewed as Laplace transforms of the density with real, negative arguments. It is worth noting that the batch scal-

ings enable us to use this lemma, as the storage processes satisfy the required condition of continuous support but the queueing models do not.

Lemma C.1.1. *Let X be a non-negative continuous random variable and let $\mathcal{M}(\cdot)$ be its moment generating function and let $\mathcal{M}'(\cdot)$ be its first derivative, i.e. $\mathcal{M}(z) = \mathrm{E}\left[e^{zX}\right]$ and $\mathcal{M}'(z) = \frac{\mathrm{d}}{\mathrm{d}\theta}\mathrm{E}\left[e^{\theta X}\right]|_{\theta=z}$. Then, for the sequence $\{a_k^m \mid m, k \in \mathbb{Z}^+\}$ given by*

$$a_k^m = (-1)^{k+1}\binom{m}{k}\binom{m+k}{k}{}_3F_2\left(k, -m, m+1; 1, k+1; \frac{1}{e}\right), \tag{C.1}$$

the summation over the products between a_k^m and $\mathcal{M}\left(-\frac{k}{c}\right)$ is such that

$$\lim_{m\to\infty}\sum_{k=1}^{m} a_k^m\,\mathcal{M}\left(-\frac{k}{c}\right) = \mathrm{P}\left(X \le c\right), \tag{C.2}$$

whereas the summation over the products between a_k^m and $\mathcal{M}'\left(-\frac{k}{c}\right)$ is such that

$$\lim_{m\to\infty}\sum_{k=1}^{m} a_k^m\,\mathcal{M}'\left(-\frac{k}{c}\right) = \mathrm{E}\left[X\mathbf{1}\{X \le c\}\right], \tag{C.3}$$

for all $c > 0$.

Proof. For $x \ge 0$ and $m \in \mathbb{Z}^+$, let the function $\mathcal{L}_m(x)$ be defined as

$$\mathcal{L}_m(x) = \sum_{k=1}^{m} a_k^m e^{-\frac{kx}{c}}, \tag{C.4}$$

where each a_k^m is as given in Equation C.1. By Sullivan et al. (1980), we have that

$$\int_0^\infty \left(\mathcal{L}_m(x) - \mathbf{1}\{x \le c\}\right)^2 \mathrm{d}x \longrightarrow 0,$$

as $m \to \infty$, which implies that $\mathcal{L}_m(x) \longrightarrow \mathbf{1}\{x \le c\}$ pointwise for $x \in [0, c)$ and $x \in (c, \infty)$ as $m \to \infty$. Furthermore, from Sullivan et al. (1980) we also have that $\mathcal{L}_m(x)$ can be equivalently expressed

$$\mathcal{L}_m(x) = -\int_0^1 \tilde{P}_m\left(\frac{w}{e}\right)\frac{\mathrm{d}}{\mathrm{d}w}\tilde{P}_m\left(we^{-\frac{x}{c}}\right)\mathrm{d}w, \tag{C.5}$$

where $\tilde{P}_m(\cdot)$ is a shifted, asymmetric Legendre polynomial defined by

$$\tilde{P}_m(w) = \sum_{k=0}^{m} \binom{m}{k}\binom{m+k}{k}(-w)^k,$$

for $w \in [0, 1]$. For reference, this can be connected to a standard Legendre polynomial $P_m(\cdot)$ via the transformation $\tilde{P}_m(w) = P_m(1 - 2w)$. To employ the dominated convergence theorem, we now bound $|\mathcal{L}_m(x)|$ as follows. Via the integral definition in Equation C.5, we can observe that the values of this function at the origin are $\mathcal{L}_m(0) = 1 + (-1)^{m+1}\tilde{P}_m(1/e)$, meaning that $\mathcal{L}_m(0) \in (0, 2)$ for all m. Hence, we now focus on the quantity when x is positive. In this case, we can see that

$$\sup_{x>0} \left| \int_0^1 \tilde{P}_m\left(\frac{w}{e}\right) \frac{\mathrm{d}}{\mathrm{d}w} \tilde{P}_m\left(we^{-\frac{x}{c}}\right) \mathrm{d}w \right| \leq \sup_{x>0} \left| \int_0^1 \frac{\mathrm{d}}{\mathrm{d}w} \tilde{P}_m\left(we^{-\frac{x}{c}}\right) \mathrm{d}w \right|,$$

which can be explained as follows. Note x dictates how much or how little to integrate along $\frac{\mathrm{d}}{\mathrm{d}w}\tilde{P}_m(we^{-\frac{x}{c}})$. That is, at $x = 0$, the integral evaluates $\frac{\mathrm{d}}{\mathrm{d}w}\tilde{P}_m(w)$ at every point in its domain $[0, 1]$ but for positive x the derivative is only evaluated from 0 to $e^{-\frac{x}{c}}$. Because we know that the shifted Legendre polynomial is bounded on $-1 \leq \tilde{P}_m(\cdot) \leq 1$, the integral on the left hand side is subject to negative values in both $\tilde{P}_m(w/e)$ and $\frac{\mathrm{d}}{\mathrm{d}w}\tilde{P}_m(we^{-\frac{x}{c}})$, whereas the right hand side only has $\frac{\mathrm{d}}{\mathrm{d}w}\tilde{P}_m(we^{-\frac{x}{c}})$. Note furthermore that $\tilde{P}_m(w/e)$ and $\frac{\mathrm{d}}{\mathrm{d}w}\tilde{P}_m(we^{-\frac{x}{c}})$ cannot match in sign at every $w \in [0, 1]$, as $\tilde{P}_m(w/e)$ is a polynomial of degree m while $\frac{\mathrm{d}}{\mathrm{d}w}\tilde{P}_m(we^{-\frac{x}{c}})$ is a polynomial of degree $m - 1$. Thus, any interval that the integral on the left hand side evaluates over can be improved upon in the right hand side by evaluating only on a subinterval in which the derivative is positive, and it does so with a larger value as $\tilde{P}_m(w/e) \leq 1$. Integrating on the right hand side now leads us to the simpler form

$$\sup_{x>0} \left| \int_0^1 \frac{\mathrm{d}}{\mathrm{d}w} \tilde{P}_m\left(we^{-\frac{x}{c}}\right) \mathrm{d}w \right| = \sup_{x>0} \left| 1 - \tilde{P}_m\left(e^{-\frac{x}{c}}\right) \right| \leq 2,$$

where the final bound again follows through the observation that $-1 \leq \tilde{P}_m(\cdot) \leq 1$. With this bound in hand, to use the dominated convergence theorem we now

review the specific convergence from Sullivan et al. (1980). From Sullivan et al. (1980), we have that $\mathcal{L}_m(x) \to \mathbf{1}\{x \le c\}$ pointwise for $x \in [0, c)$ and $x \in (c, \infty)$. At the point of discontinuity in the indicator function at $x = c$, it can be observed that $\mathcal{L}_m(c) \to \frac{1}{2}$ as $m \to \infty$. Because the random variable X is assumed to be continuous, the singleton $\{c\}$ is of measure 0 and thus $\mathcal{L}_m(x) \to \mathbf{1}\{x \le c\}$ almost everywhere, justifying use of the dominated convergence theorem. Using this, we now have that

$$\mathrm{E}\left[\mathcal{L}_m(X)\right] \longrightarrow \mathrm{E}\left[\mathbf{1}\{X \le c\}\right] = \mathrm{P}\left(X \le c\right) \quad \text{and} \quad \mathrm{E}\left[X\mathcal{L}_m(X)\right] \longrightarrow \mathrm{E}\left[X\mathbf{1}\{X \le c\}\right],$$

as $m \to \infty$. Using the definition of $\mathcal{L}_m(x)$ in Equation C.4 and linearity of expectation, one can write

$$\mathrm{E}\left[\mathcal{L}_m(X)\right] = \sum_{k=1}^{m} a_k^m \mathrm{E}\left[e^{-\frac{kX}{c}}\right] \quad \text{and} \quad \mathrm{E}\left[X\mathcal{L}_m(X)\right] = \sum_{k=1}^{m} a_k^m \mathrm{E}\left[Xe^{-\frac{kX}{c}}\right],$$

and by observing that $\mathcal{M}'\left(-\frac{k}{c}\right) = \mathrm{E}\left[Xe^{-\frac{kX}{c}}\right]$, we complete the proof. □

As a related numerical discussion, let us demonstrate how we perform approximate implementations of the expressions in Theorems 6.3.3 and C.2.4 as based on the Legendre exponential forms given in Lemma C.1.1. As an initial observation, we can note that as m grows large, calculations of the coefficients given in Equation C.1 become subject to numerical inaccuracies, such as overflow, due to the large binomial coefficients. While this could potentially be assuaged by use of Stirling's approximation or something similar, in our numerical experiments we have seen that such techniques may not be necessary for strong performance. However, we can note that the convergences in these results need not be monotone, hence we will not simply take the expression for the largest m before numerical instability is observed. To explain through example, we will calculate the empirical exceedance probability in the delay queueing

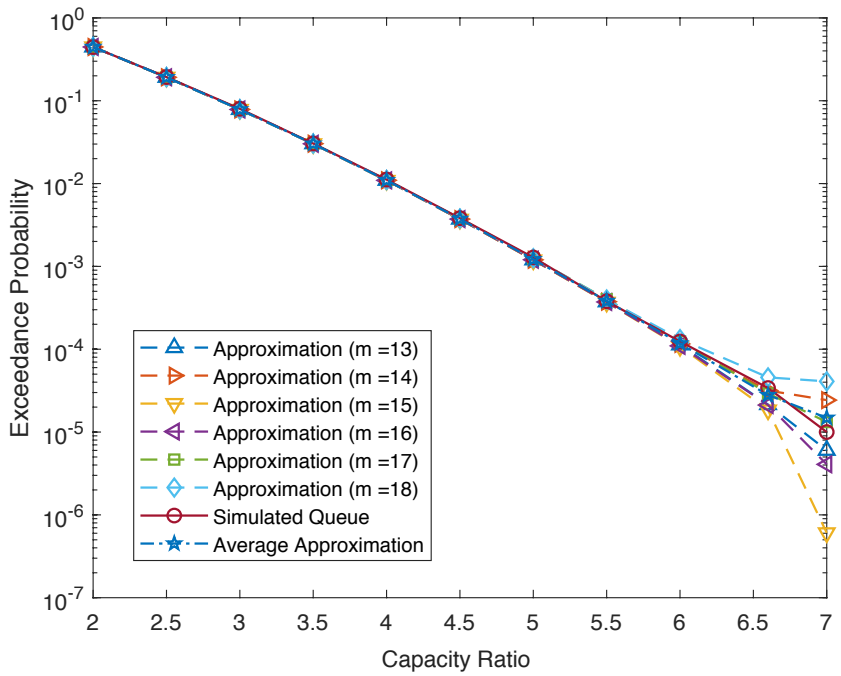


Figure C.1: Comparison of Legendre approximations and the empirical exceedance probability in a simulated queue with fixed size batches of size $n = 100$, $\lambda = 3$, and $\mu = 2$.

model via simulation and compare it to various approximate Legendre sums. Based on Theorem 6.3.3, we have that

$$\mathrm{P}\left(Q_\infty^C(n) > cn\right) \approx \mathrm{P}\left(\psi_\infty^C > c\right) \approx \frac{\frac{\lambda}{\mu}\mathrm{E}\left[M_1\right] - \sigma_{m,c}^{(C)}}{c - \sigma_{m,c}^{(C)}},$$

and so we will consider candidate m values, which we plot in Figure C.1.

As one can see, for relatively small values of m the approximation performs quite well, as the simulated values and the approximation are virtually indistinguishable before the true probability is approximately of order 10^{-5}. However, if desired we can improve this further by taking the average among the candidate approximations. We can see that this does well in this example, and we can quickly show it will do no worse than the worst individual approximation. For p as the true probability and p_{σ_m} as the approximation at m, by the triangle

inequality we have that

$$\left|\sum_{k=m_0}^{m_1} \frac{p_{\sigma_k}}{m_1 - m_0 + 1} - p\right| = \left|\sum_{k=m_0}^{m_1} \frac{p_{\sigma_k} - p}{m_1 - m_0 + 1}\right| \leq \sum_{k=m_0}^{m_1} \frac{\left|p_{\sigma_k} - p\right|}{m_1 - m_0 + 1} \leq \max_{m_0 \leq k \leq m_1} \left|p_{\sigma_k} - p\right|.$$

Thus, a loose description of an approximation heuristic based on these Legendre limits is as follows: compute multiple candidate approximations, remove clear errors caused by numerical instabilities and pre-convergence gaps, and take the average of the remaining candidates. While our experiments suggest that this simple approach does well, we can note that it could be possible to develop more sophisticated numerical approximations based on these limits and we find this to be an interesting direction of future research.

C.2 Analysis of Blocking Model Queue

As we referred to in the introduction, the analysis that we have performed in the main body of this work can easily be extended to a blocking batch arrival queueing model. In this section of the appendix, we will briefly reproduce key results for this system. In a similar definition to the delay multi-server model, let $Q_t^B(n)$ be the queue length process for a $G_t^{B(n)}/GI/cn/cn$ queueing system. That is, we will take the same general assumptions on the arrival epochs, batch distribution, and service distribution as in $Q_t^C(n)$ and $Q_t(n)$, but we will now assume that there are cn servers with no space for waiting. Thus, any arriving batch that contains more jobs than available servers will experience partial blocking, meaning that the excess jobs that do not find an available server will be blocked and will not enter the system. It is worth noting that a blocking system could be an acceptable model for the look-ahead teleoperations system, as the pre-fetching context could imply that there is no time for a job to wait. We will also

define a finite capacity storage process ψ_t^B for $t \geq 0$ such that

$$\psi_t^B = \psi_0^B \bar{G}_0(t) + \sum_{i=1}^{N_t} \left(M_i \wedge \psi_{A_i^-}^B \right) \bar{G}\left(t - A_i\right), \tag{C.6}$$

where $\psi_{A_i^-}^B = \lim_{t \uparrow A_i} \psi_t^B$ and $\frac{Q_0^B(n)}{n} \to \psi_0^B$ as $n \to \infty$. Then, in Theorem C.2.1 we prove the analogous batch scaling limit for the blocking queueing model, which converges to this finite capacity shot noise process.

Theorem C.2.1. *As $n \to \infty$, the batch scaling of the $G^{B(n)}/GI/cn/cn$ queue $Q_t^B(n)$ yields*

$$\frac{Q_t^B(n)}{n} \overset{D}{\Longrightarrow} \psi_t^B, \tag{C.7}$$

pointwise in $t \geq 0$, where ψ_t^B is a finite storage process as defined in Equation C.6.

Proof. We will again take an inductive approach and show convergence on the inter-arrival times. As a base case, let $t \in [0, A_1)$. Then, we can observe through the law of large numbers and the assumed initial conditions that

$$\frac{Q_t^B(n)}{n} = \frac{1}{n} \sum_{j=1}^{Q_0^B(n)} \mathbf{1}\{t < S_{0,j}\} = \frac{Q_0^B(n)}{n} \frac{1}{Q_0^B(n)} \sum_{j=1}^{Q_0^B(n)} \mathbf{1}\{t < S_{0,j}\} \xrightarrow{a.s.} \psi_0^B \bar{G}_0(t),$$

as $n \to \infty$. Then, at the time of first arriving batch, we can note that the new queue length will be $Q_{A_1}^B(n) = \left(Q_{A_1^-}^B(n) + B_1(n) \wedge cn \right)$. Taking this as a difference from the moment just before the batch arrives, we can observe that as $n \to \infty$

$$\frac{Q_{A_1}^B(n)}{n} = \frac{Q_{A_1^-}^B(n)}{n} + \left(\frac{B_1(n)}{n} \wedge c - \frac{Q_{A_1^-}^B(n)}{n} \right) \overset{D}{\Longrightarrow} \psi_{A_1^-}^B + \left(M_1 \wedge c - \psi_{A_1^-}^B \right),$$

by the previous observation and the assumed convergence of the batch size distribution. Hence, we proceed to the inductive step. We now take $t \in [A_k, A_{k+1})$ and assume convergence for all time less than or equal to A_k as an inductive hypothesis. Then, on this time interval we can decompose the queue length as

$$Q_t^B(n) = \sum_{j=1}^{Q_0^B(n)} \mathbf{1}\{t < S_{0,j}\} + \sum_{i=1}^{N_t} \sum_{j=1}^{\tilde{B}_i(n)} \mathbf{1}\{t < S_{i,j}\},$$

where

$$\tilde{B}_i(n) = Q^B_{A_i}(n) - Q^B_{A_i^-}(n) = \left(B_i(n) \wedge cn - Q^B_{A_i^-}(n)\right).$$

Scaling inversely by n, we find that

$$\begin{aligned}
\frac{Q^B_t(n)}{n} &= \frac{1}{n}\sum_{j=1}^{Q^B_0(n)} \mathbf{1}\{t < S_{0,j}\} + \frac{1}{n}\sum_{i=1}^{N_t}\sum_{j=1}^{\tilde{B}_i(n)} \mathbf{1}\{t < S_{i,j}\} \\
&= \frac{Q^B_0(n)}{n}\frac{1}{Q^B_0(n)}\sum_{j=1}^{Q^B_0(n)} \mathbf{1}\{t < S_{0,j}\} + \sum_{i=1}^{N_t}\frac{\tilde{B}_i(n)}{n}\frac{1}{\tilde{B}_i(n)}\sum_{j=1}^{\tilde{B}_i(n)} \mathbf{1}\{t < S_{i,j}\} \\
&\overset{D}{\Longrightarrow} \psi^B_0 \bar{G}_0(t) + \sum_{i=1}^{N_t}\left(M_i \wedge c - \psi^B_{A_i^-}\right)\bar{G}(t - A_i),
\end{aligned}$$

as $n \to \infty$, since

$$\frac{\tilde{B}_i(n)}{n} = \left(\frac{B_i(n)}{n} \wedge c - \frac{Q^B_{A_i^-}(n)}{n}\right) \overset{D}{\Longrightarrow} \left(M_i \wedge c - \psi^B_{A_i^-}\right)$$

by the inductive hypothesis and the continuous mapping theorem. Then, at the arrival epoch A_{k+1} the admitted batch size converges to

$$\frac{1}{n}\left(Q^B_{A_{k+1}}(n) - Q^B_{A^-_{k+1}}(n)\right) = \frac{1}{n}\left(B_{k+1}(n) \wedge cn - Q^B_{A^-_{k+1}}(n)\right) \overset{D}{\Longrightarrow} \left(M_{k+1} \wedge c - \psi^B_{A^-_{k+1}}\right),$$

which completes the proof. □

Just as we have visualized the convergence of the infinite server and delay model queues in Figures 6.5 and 6.6, respectively, we plot the analogous demonstration for the blocking model queues and finite capacity storage processes in Figure C.2. For this example the batches are binomially distributed with number of trials n and probability of success $\frac{1}{2}$, and this yields deterministic jumps of size $\frac{1}{2}$ in the finite capacity storage process storage process. One can observe that the scaled queue lengths and the finite capacity storage process all lie on the interval $[0, 2]$, and that as the batch size grows large the distributions of the queue appear to approach that of the storage process.

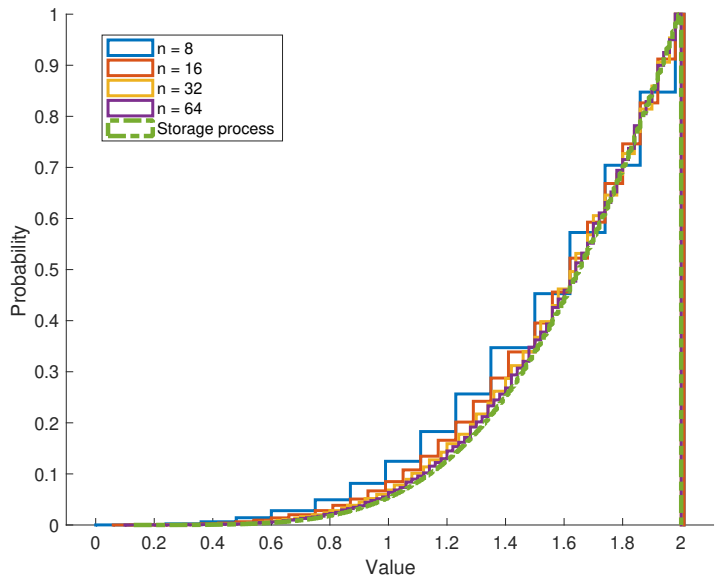


Figure C.2: Simulated demonstration of convergence in distribution of an $M^{B(n)}/M/cn/cn$ queue to a finite capacity storage process, based on 100,000 replications with $t = 10$, $\lambda = 5$, $\mu = 1$, $c = 2$, and $B_1(n) \sim \text{Bin}\left(n, \frac{1}{2}\right)$.

Likewise in the blocking model, we want to compute the probability that some portion of an arriving batch is blocked, which means that the sum of the pre-arrival queue length and the incoming batch size exceeds the number of servers. Again we find a connection to the corresponding storage process, as this converges to the probability that the finite storage processes ψ_t^B is above its capacity as the relative batch size grows large:

$$\mathrm{P}\left(Q_t^B(n) + B(n) > cn\right) = \mathrm{P}\left(\frac{Q_t^B(n)}{n} + \frac{B(n)}{n} > c\right) \longrightarrow \mathrm{P}\left(\psi_t^B + M > c\right). \tag{C.8}$$

To compute this in the Markovian setting, we make use of a second lemma from the storage process literature. We cite a truncation result for the steady-state density of ψ_∞^B, which we can view as analogous to several known queueing results. It is known that the reversibility of the $M/M/\infty$ queue implies that truncation yields the stationary distribution of the $M/M/c/c$ Erlang-B model. This can even be observed in non-reversible models, as the steady-state distribution of

a batch arrival blocking model can be obtained via truncating the steady-state distribution of an infinite server queue with batch arrivals. This can be seen as a result of Chapter 3. Now in Lemma C.2.2 we see that this is also known in the storage process literature, as the density of a finite dam with Poisson process epochs can be found via truncation of a shot noise process.

Lemma C.2.2. *In the Markovian setting, the steady-state density of the finite capacity storage process $f_B(\cdot)$ exists and is given by*

$$f_B(x) = \frac{f_\infty(x)}{\int_0^c f_\infty(y)\mathrm{d}y}, \tag{C.9}$$

for all $0 < x \leq c$, where $f_\infty(\cdot)$ is the steady-state density of the shot noise process.

Proof. See Section 8 of Brockwell (1977). □

This truncation also immediately yields validity to an interchange of limits, which we can now quickly state in the following proposition.

Proposition C.2.3. *In the stationary Markovian blocking queueing model, the interchange of limits of time and batch scaling is justified. That is,*

$$\lim_{n\to\infty}\lim_{t\to\infty} \mathrm{P}\left(\frac{Q_t^B(n)}{n} \leq x\right) = \lim_{t\to\infty}\lim_{n\to\infty} \mathrm{P}\left(\frac{Q_t^B(n)}{n} \leq x\right), \tag{C.10}$$

for all $x \in [0, c]$.

Proof. Because the distribution of $Q_t^B(n)$ is a truncation of the distribution of $Q_t(n)$ to $[0, cn]$, we can write the cumulative distribution function as

$$\mathrm{P}\left(\frac{Q_t^B(n)}{n} \leq x\right) = \frac{\mathrm{P}\left(\frac{Q_t(n)}{n} \leq x\right)}{\mathrm{P}\left(\frac{Q_t(n)}{n} \leq c\right)},$$

implying that this interchange is an immediate consequence of Theorem 6.3.2.

□

Using this relationship between the finite capacity storage process and the shot noise process, we can now find a Legendre-based computation for the exceedance probability that can be expressed in terms of quantities that we have identified already in the main text.

Theorem C.2.4. *In the stationary Markovian setting, the exceedance probability for ψ_∞^C is given by*

$$\mathrm{P}\left(\psi_\infty^B + M_1 > c\right) = \lim_{m\to\infty} \frac{\frac{\lambda+\mu}{\lambda}\sigma_{m,c}^{(C1)} - \sigma_{m,c}^{(C2)}}{c - \sigma_{m,c}^{(C2)}}, \tag{C.11}$$

where $\sigma_{m,c}^{(C1)}$ and $\sigma_{m,c}^{(C2)}$ are as given in Theorem 6.3.3 and with a_k^m as defined in Equation C.1.

Proof. Again by conditioning, we can observe that the mean of the minimum between the threshold c and the jump-added steady-state content is

$$\mathrm{E}\left[\psi_\infty^B + M_1 \wedge c\right] = c\mathrm{P}\left(\psi_\infty^B + M_1 > c\right) + \mathrm{E}\left[\psi_\infty^B + M_1 \mid \psi_\infty^B + M_1 \le c\right]\left(1 - \mathrm{P}\left(\psi_\infty^B + M_1 > c\right)\right).$$

Then, by use of the integral equations for the density, we can see that the mean of the finite capacity storage process satisfies

$$\mathrm{E}\left[\psi_\infty^B\right] = \int_0^c x f_B(x)\mathrm{d}x = \int_0^c \frac{\lambda}{\mu}\left(\mathrm{P}\left(\psi_\infty^B + M_1 > x\right) - \mathrm{P}\left(\psi_\infty^B > x\right)\right)\mathrm{d}x,$$

which implies that

$$\frac{\lambda+\mu}{\lambda}\mathrm{E}\left[\psi_\infty^B\right] = \mathrm{E}\left[\psi_\infty^B + M_1 \wedge c\right].$$

Hence, we can then express the exceedance probability as

$$\mathrm{P}\left(\psi_\infty^B + M_1 > c\right) = \frac{\frac{\lambda+\mu}{\lambda}\mathrm{E}\left[\psi_\infty^B\right] - \mathrm{E}\left[\psi_\infty^B + M_1 \mid \psi_\infty^B + M_1 \le c\right]}{c - \mathrm{E}\left[\psi_\infty^B + M_1 \mid \psi_\infty^B + M_1 \le c\right]}.$$

Both the finite storage process mean and the truncated mean of the jump-added storage process can be calculated using Lemma C.1.1. By the distributional equivalence of the finite storage process and the truncated shot

noise as stated in Lemma C.2.2, we can note that $\mathrm{E}\left[\psi_\infty^B\right] = \mathrm{E}\left[\psi_\infty \mid \psi_\infty \le c\right]$ and $\mathrm{E}\left[\psi_\infty^B + M_1 \mid \psi_\infty^B + M_1 \le c\right] = \mathrm{E}\left[\psi_\infty + M_1 \mid \psi_\infty + M_1 \le c\right]$, and thus through the results of Theorem 6.2.2 we complete the proof. □

C.3 Exact Analysis for Geometrically Distributed Batches

As an example of an approach to this staffing problem under more specific assumptions, in this subsection we will suppose that the batch size distribution in the queueing models is geometric, i.e. we let $B_1(n) \sim \mathrm{Geo}\left(\frac{\alpha}{n}\right)$ for some $\alpha > 0$. We can observe that having geometrically distributed batch sizes implies that the jumps in the storage process will be exponentially distributed, as

$$\mathrm{E}\left[e^{\frac{\theta}{n}B_1(n)}\right] = \frac{\frac{\alpha}{n}}{1-\left(1-\frac{\alpha}{n}\right)e^{\frac{\theta}{n}}} = \frac{\alpha}{\alpha e^{\frac{\theta}{n}} - n\left(e^{\frac{\theta}{n}}-1\right)} \longrightarrow \frac{\alpha}{\alpha-\theta} = \mathrm{E}\left[e^{\theta M_1}\right],$$

with $M_1 \sim \mathrm{Exp}(\alpha)$. In this situation, the assumed stability condition simplifies to $\lambda < \alpha c \mu$. As is often the case for exponential random variables, we will find that this leads to significant tractability. Thus, throughout this section we will use the assumption that $M_1 \sim \mathrm{Exp}(\alpha)$ together with Lemmas 6.3.1 and C.2.2 to find the steady-state densities of ψ_∞, ψ_∞^C, and ψ_∞^B in closed form. The assumption of exponential marks will be explicitly stated at the beginning of each statement for clarity's sake. Because the shot noise process will again be the cornerstone for the c-threshold and finite storage processes, we begin by showing that its steady-state value is gamma distributed.

Proposition C.3.1. *Suppose that $M_1 \sim \mathrm{Exp}(\alpha)$. Then, $\psi_\infty \sim \mathrm{Gamma}\left(\frac{\lambda}{\mu}, \alpha\right)$ in the Markovian setting.*

Proof. By Lemma 6.3.1, we know that for all $x > 0$ the density $f_\infty(x)$ will satisfy

the integral equation

$$xf_\infty(x) = \frac{\lambda}{\mu} e^{-\alpha x} \int_0^x e^{\alpha y} f_\infty(y) \mathrm{d}y.$$

By taking the derivative of each side with respect to x and simplifying, we find the ordinary differential equation

$$f'_\infty(x) = \frac{1}{x}\left(\frac{\lambda}{\mu} - 1\right) f_\infty(x) - \alpha f_\infty(x),$$

which yields a solution of $f_\infty(x) = k_1 e^{-\alpha x} x^{\frac{\lambda}{\mu}-1}$ for some constant k_1. By requiring that $\int_0^\infty f_\infty(x)\mathrm{d}x = 1$ and solving for the normalizing constant k_1, we find the density of a Gamma$\left(\frac{\lambda}{\mu}, \alpha\right)$ random variable. □

Using this same technique, we can also derive the steady-state density of the c-threshold storage process ψ_∞^C. In this case, we find in Proposition C.3.2 that the threshold release rule manifests itself as a piecewise stationary density. In particular, the shape of the distribution below the threshold is proportional to a gamma distribution like what was shown in Proposition C.3.1 and above the threshold the density is proportional to an exponential distribution. We can note that this then resembles the conditional distribution of the conditional waiting time in an $M/M/c$ queue and the conditional distribution of the workload process in an $M/M/1$ queue, which is one of the classic connections between queues and storage (or dam) processes with linear drain; see for example Prabhu (2012). Thus, just as a multiserver queue can be seen as a hybrid between an infinite server queue and a single server queue, the c-threshold storage process can be connected to the storage processes corresponding to the infinite and single server queues, and the structure of each can be plainly observed in the steady-state density of ψ_∞^C under these assumptions.

Proposition C.3.2. *Suppose that $M_1 \sim \mathrm{Exp}(\alpha)$. Then, in the Markovian setting ψ_∞^C has probability density function given by*

$$f_C(x) = \begin{cases} \frac{\alpha^{\lambda/\mu}(\alpha c\mu-\lambda)e^{-\alpha x}x^{\frac{\lambda}{\mu}-1}}{(\alpha c\mu-\lambda)\Gamma\left(\frac{\lambda}{\mu}\right)-\alpha c\mu\Gamma\left(\frac{\lambda}{\mu},\alpha c\right)+\mu\Gamma\left(\frac{\lambda}{\mu}+1,\alpha c\right)} & 0 \le x \le c, \\ \frac{\left(\alpha-\frac{\lambda}{c\mu}\right)\left(\mu\Gamma\left(\frac{\lambda}{\mu}+1,\alpha c\right)-\lambda\Gamma\left(\frac{\lambda}{\mu},\alpha c\right)\right)e^{-\left(\alpha-\frac{\lambda}{c\mu}\right)(x-c)}}{(\alpha c\mu-\lambda)\Gamma\left(\frac{\lambda}{\mu}\right)-\alpha c\mu\Gamma\left(\frac{\lambda}{\mu},\alpha c\right)+\mu\Gamma\left(\frac{\lambda}{\mu}+1,\alpha c\right)} & x > c. \end{cases} \tag{C.12}$$

Proof. From the integral equation given in Lemma 6.3.1, we can note that the density $f_C(x)$ satisfies

$$xf_C(x)e^{\alpha x} = \frac{\lambda}{\mu}\int_0^x e^{\alpha y}f_C(y)\mathrm{d}y,$$

for $x \le c$, which we have seen yields $f_C(x) = k_1 e^{-\alpha x}x^{\frac{\lambda}{\mu}-1}$ for some constant k_1 through the proof of Proposition C.3.1. Similarly for $x > c$, Lemma 6.3.1 also implies that

$$f_C(x)e^{\alpha x} = \frac{\lambda}{c\mu}\int_0^x e^{\alpha y}f_C(y)\mathrm{d}y,$$

and thus this first derivative satisfies the equation

$$f_C'(x) = -\left(\alpha - \frac{\lambda}{c\mu}\right)f_C(x).$$

By consequence, $f_C(x) = k_2 e^{-\left(\alpha-\frac{\lambda}{c\mu}\right)x}$ for $x > c$ and some constant k_2. To solve for k_1 and k_2, we can use the fact that the density must integrate to 1 to observe

$$1 = \int_0^\infty f_C(x)\mathrm{d}x = k_1\alpha^{-\frac{\lambda}{\mu}}\left(\Gamma\left(\frac{\lambda}{\mu}\right) - \Gamma\left(\frac{\lambda}{\mu},\alpha c\right)\right) + \frac{k_2}{\alpha - \frac{\lambda}{c\mu}}e^{-\left(\alpha-\frac{\lambda}{c\mu}\right)c}.$$

Similarly, because $\mathrm{E}\left[\phi_\infty^C \wedge c\right] = \frac{\lambda}{\mu}\mathrm{E}\left[M_1\right] = \frac{\lambda}{\alpha\mu}$ as seen in the proof of Theorem 6.3.3, we can also note that

$$\frac{\lambda}{\alpha\mu} = \int_0^\infty (x\wedge c)f_C(x)\mathrm{d}x = k_1\alpha^{-\frac{\lambda}{\mu}-1}\left(\Gamma\left(\frac{\lambda}{\mu}+1\right) - \Gamma\left(\frac{\lambda}{\mu}+1,\alpha c\right)\right) + \frac{ck_2}{\alpha - \frac{\lambda}{c\mu}}e^{-\left(\alpha-\frac{\lambda}{c\mu}\right)c}.$$

This now gives us a system of linear equations of k_1 and k_2. By solving and simplifying, we achieve the stated form. □

For the finite storage process ψ_∞^B, we will now derive its steady-state density by use of the density for ψ_∞ in Proposition C.3.1 and the truncation equation given in Lemma C.2.2. Thus, as we have found that the shot noise process steady-state is equivalent to a gamma random variable, we find in Proposition C.3.3 that we can view the resulting density for ψ_∞^B as a truncated gamma distribution.

Proposition C.3.3. *Suppose that* $M_1 \sim \mathrm{Exp}(\alpha)$. *Then,* ψ_∞^B *has probability density function given by*

$$f_B(x) = \frac{\alpha^{\frac{\lambda}{\mu}} x^{\frac{\lambda}{\mu}-1} e^{-\alpha x}}{\Gamma\left(\frac{\lambda}{\mu}\right) - \Gamma\left(\frac{\lambda}{\mu}, \alpha\right)}, \tag{C.13}$$

for all $0 < x \leq c$.

Proof. By Proposition C.3.1, we know that the shot noise process is gamma distributed in steady-state, i.e. it has a density that is proportional to $e^{-\alpha x} x^{\frac{\lambda}{\mu}-1}$. By Proposition C.2.2, we can normalize this expression so that $\int_0^c f_B(x) = 1$, and this yields the stated form. □

BIBLIOGRAPHY

Abramowitz, Milton and Irene A Stegun (1965), *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*, volume 55. Courier Corporation.

Adan, Ivo, Cor Hurkens, and Gideon Weiss (2010), "A reversible Erlang loss system with multitype customers and multitype servers." *Probability in the Engineering and Informational Sciences*, 24, 535–548.

Adan, Ivo and Gideon Weiss (2012), "A loss system with skill-based servers under assign to longest idle server policy." *Probability in the Engineering and Informational Sciences*, 26, 307–321.

Ahmed, Amr and Eric Xing (2008), "Dynamic non-parametric mixture models and the recurrent chinese restaurant process: with applications to evolutionary clustering." In *Proceedings of the 2008 SIAM International Conference on Data Mining*, 219–230, SIAM.

Aït-Sahalia, Yacine, Julio Cacho-Diaz, and Roger JA Laeven (2015), "Modeling financial contagion using mutually exciting jump processes." *Journal of Financial Economics*, 117, 585–606.

Aldous, David J (1985), "Exchangeability and related topics." In *École d'Été de Probabilités de Saint-Flour XIII—1983*, 1–198, Springer.

Alexa the Web Information Company (2017), "Top sites in United States."

Andrews, Donald WK (1988), "Laws of large numbers for dependent non-identically distributed random variables." *Econometric Theory*, 4, 458–467.

Asmussen, Søren, Jens Ledet Jensen, and Leonardo Rojas-Nandayapa (2016), "On the Laplace transform of the lognormal distribution." *Methodology and Computing in Applied Probability*, 18, 441–458.

Azizpour, Shahriar, Kay Giesecke, and Gustavo Schwenkler (2016), "Exploring the sources of default clustering." *Journal of Financial Economics*.

Bacry, Emmanuel, Sylvain Delattre, Marc Hoffmann, and Jean-Francois Muzy (2013), "Some limit theorems for Hawkes processes and application to financial statistics." *Stochastic Processes and their Applications*, 123, 2475–2499.

Bacry, Emmanuel, Thibault Jaisson, and Jean-François Muzy (2016), "Estimation of slowly decreasing Hawkes kernels: application to high-frequency order book dynamics." *Quantitative Finance*, 16, 1179–1201.

Bacry, Emmanuel and Jean-François Muzy (2014), "Hawkes model for price and trades high-frequency dynamics." *Quantitative Finance*, 14, 1147–1166.

Baily, David E and Marcel F Neuts (1981), "Algorithmic methods for multi-server queues with group arrivals and exponential services." *European Journal of Operational Research*, 8, 184–196.

Ball, Frank (1983), "The threshold behaviour of epidemic models." *Journal of Applied Probability*, 20, 227–241.

Blackwell, David (1948), "A renewal theorem." *Duke Mathematical Journal*, 15, 145–150.

Blackwell, David, James B MacQueen, et al. (1973), "Ferguson distributions via Pólya urn schemes." *The Annals of Statistics*, 1, 353–355.

Blanc, Pierre, Jonathan Donier, and J-P Bouchaud (2017), "Quadratic Hawkes processes for financial prices." *Quantitative Finance*, 17, 171–188.

Blei, David M and Peter I Frazier (2011), "Distance dependent chinese restaurant processes." *Journal of Machine Learning Research*, 12, 2461–2488.

Boxma, Onno, Offer Kella, and Michel Mandjes (2018), "Infinite-server systems with Coxian arrivals." *Working paper*.

Boxma, Onno, Offer Kella, and David Perry (2011), "On some tractable growth-collapse processes with renewal collapse epochs." *Journal of Applied Probability*, 48, 217–234.

Boxma, Onno, David Perry, Wolfgang Stadje, and Shelemyahu Zacks (2006), "A Markovian growth-collapse model." *Advances in Applied Probability*, 38, 221–243.

Brawer, Robert and Magnus Pirovino (1992), "The linear algebra of the Pascal matrix." *Linear Algebra and Its Applications*, 174, 13–23.

Brémaud, Pierre and Laurent Massoulié (1996), "Stability of nonlinear Hawkes processes." *The Annals of Probability*, 1563–1588.

Brockmeyer, E, HL Halstrom, and Arne Jensen (1948), *The life and works of AK Erlang*. Copenhagen: Copenhagen Telephone Co.

Brockwell, Peter J, Sidney I Resnick, and Richard L Tweedie (1982), "Storage processes with general release rule and additive inputs." *Advances in Applied Probability*, 14, 392–433.

Brockwell, PJ (1977), "Stationary distributions for dams with additive input and content-dependent release rate." *Advances in Applied Probability*, 9, 645–663.

Brown, Mark and Sheldon M Ross (1969), "Some results for infinite server Poisson queues." *Journal of Applied Probability*, 6, 604–611.

Burns, Lawrence D (2013), "Sustainable mobility: a vision of our transport future." *Nature*, 497, 181.

Butcher, John Charles (2016), *Numerical methods for ordinary differential equations*. John Wiley & Sons.

California Department of Motor Vehicles (2018), "Article 3.7 – Testing of Autonomous Vehicles." *Title 13, Division 1, Chapter 1*.

Call, Gregory S and Daniel J Velleman (1993), "Pascal's matrices." *The American Mathematical Monthly*, 100, 372–376.

Chaudhry, Mohan L and James J Kim (2016), "Analytically elegant and computationally efficient results in terms of roots for the $GI^X/M/c$ queueing system." *Queueing Systems*, 82, 237–257.

Chiamsiri, Singha and Michael S Leonard (1981), "A diffusion approximation for bulk queues." *Management Science*, 27, 1188–1199.

Chung, John Joon Young, Fuhu Xiao, Nicholas Recker, Kammeran Barnes, Nikola Banovic, and Walter S Lasecki (2019), "Accident prevention with predictive instantaneous crowdsourcing." In *CHI'19 Workshop on "Looking into the Future: Weaving the Threads of Vehicle Automation"*, ACM.

Cinlar, E and M Pinsky (1972), "On dams with additive inputs and a general release rule." *Journal of Applied Probability*, 9, 422–429.

Cox, David R (1955), "A use of complex probabilities in the theory of stochastic

processes." In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 51, 313–319, Cambridge University Press.

Cromie, MV, ML Chaudhry, and WK Grassmann (1979), "Further results for the queueing system $M^X/M/c$." *Journal of the Operational Research Society*, 30, 755–763.

Cui, Lirong, Alan Hawkes, and He Yi (2019), "An elementary derivation of moments of Hawkes processes." *Journal of Applied Probability*, (To Appear).

Cui, Lirong and Bei Wu (2019), "Moments for Hawkes processes with gamma decay kernel functions." *Working paper by personal communication*.

Da Fonseca, José and Riadh Zaatour (2014), "Hawkes process: Fast calibration, application to trade clustering, and diffusive limit." *Journal of Futures Markets*, 34, 548–579.

Da Fonseca, José and Riadh Zaatour (2015), "Clustering and mean reversion in a Hawkes microstructure model." *Journal of Futures Markets*, 35, 813–838.

Daley, Daryl J and David Vere-Jones (2003), *An introduction to the theory of point processes: Volume I: Elementary Theory and Methods*. Springer Science & Business Media.

Daley, Daryl J and David Vere-Jones (2007), *An introduction to the theory of point processes: volume II: general theory and structure*. Springer Science & Business Media.

Dassios, Angelos and Hongbiao Zhao (2011), "A dynamic contagion process." *Advances in Applied Probability*, 43, 814–846.

Dassios, Angelos and Hongbiao Zhao (2017), "Efficient simulation of clustering jumps with CIR intensity." *Operations Research*, 65, 1494–1515.

Dattoli, Giuseppe and HM Srivastava (2008), "A note on harmonic numbers, umbral calculus and generating functions." *Applied Mathematics Letters*, 21, 686–693.

Davies, Alex (2019), "The war to remotely control self-driving cars heats up." *Wired*.

Davis, Mark HA (1984), "Piecewise-deterministic markov processes: A general class of non-diffusion stochastic models." *Journal of the Royal Statistical Society. Series B (Methodological)*, 353–388.

Daw, Andrew, Robert C Hampshire, and Jamol Pender (2019), "Beyond safety drivers: Staffing a teleoperations system for autonomous vehicles." *arXiv preprint arXiv:1907.12650*.

Daw, Andrew and Jamol Pender (2018), "Queues driven by Hawkes processes." *Stochastic Systems*, 8, 192–229.

Daw, Andrew and Jamol Pender (2019a), "New perspectives on the Erlang-A queue." *Advances in Applied Probability*, 51.

Daw, Andrew and Jamol Pender (2019b), "On the distributions of infinite server queues with batch arrivals." *Queueing Systems*, 91, 367–401.

Daw, Andrew and Jamol Pender (2020a), "An ephemerally self-exciting point process." *arXiv preprint arXiv:1811.04282*.

Daw, Andrew and Jamol Pender (2020b), "Matrix calculations for moments of markov processes." *arXiv preprint arXiv:1909.03320*.

de Graaf, WF, Willem RW Scheinhardt, and RJ Boucherie (2017), "Shot-noise fluid queues and infinite-server systems with batch arrivals." *Performance evaluation*, 116, 143–155.

Debo, Laurens G, Christine Parlour, and Uday Rajan (2012), "Signaling quality via queues." *Management Science*, 58, 876–891.

Designated Driver (2019), "Remote operator setup."

Ding, Jiu and Aihui Zhou (2007), "Eigenvalues of rank-one updated matrices with some applications." *Applied Mathematics Letters*, 20, 1223–1226.

Du, Nan, Mehrdad Farajtabar, Amr Ahmed, Alexander J Smola, and Le Song (2015), "Dirichlet-Hawkes processes with applications to clustering continuous-time document streams." In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 219–228, ACM.

Duffie, Darrell, Jun Pan, and Kenneth Singleton (2000), "Transform analysis and asset pricing for affine jump-diffusions." *Econometrica*, 68, 1343–1376.

Edelman, Alan and Gilbert Strang (2004), "Pascal matrices." *The American Mathematical Monthly*, 111, 189–197.

Eick, Stephen G, William A Massey, and Ward Whitt (1993), "The physics of the $M_t/G/\infty$ queue." *Operations Research*, 41, 731–742.

Engblom, Stefan and Jamol Pender (2014), "Approximations for the moments of nonstationary and state dependent birth-death queues." *arXiv preprint arXiv:1406.6164*.

Errais, Eymen, Kay Giesecke, and Lisa R Goldberg (2010), "Affine point processes and portfolio credit risk." *SIAM Journal on Financial Mathematics*, 1, 642–665.

Ertekin, Şeyda, Cynthia Rudin, Tyler H McCormick, et al. (2015), "Reactive point processes: A new approach to predicting power failures in underground electrical systems." *The Annals of Applied Statistics*, 9, 122–144.

Ethier, Stewart N and Thomas G Kurtz (2009), *Markov processes: characterization and convergence*, volume 282. John Wiley & Sons.

Falin, Gennadi (1994), "The $M^k/G/\infty$ batch arrival queue by heterogeneous dependent demands." *Journal of Applied Probability*, 31, 841–846.

Farajtabar, Mehrdad, Jiachen Yang, Xiaojing Ye, Huan Xu, Rakshit Trivedi, Elias Khalil, Shuang Li, Le Song, and Hongyuan Zha (2017), "Fake news mitigation via point process based intervention." In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 1097–1106, JMLR. org.

Federal Highway Administration (2017), "2017 national household travel survey." *U.S. Department of Transportation, Washington, DC*.

Feldman, Zohar, Avishai Mandelbaum, William A Massey, and Ward Whitt (2008), "Staffing of time-varying queues to achieve time-stable performance." *Management Science*, 54, 324–338.

Feller, Willliam (1957), *An introduction to probability theory and its applications*, 2 edition, volume 1. John Wiley & Sons.

Filipović, Damir and Martin Larsson (2019), "Polynomial jump-diffusion models." *Swiss Finance Institute Research Paper*.

Foster, FG (1964), "Batched queuing processes." *Operations Research*, 12, 441–449.

Fralix, Brian (2019), "On classes of bitcoin-inspired infinite-server queueing systems."

Frolkova, Maria and Michel Mandjes (2019), "A bitcoin-inspired infinite-server model with a random fluid limit." *Stochastic Models*, 35, 1–32.

Gao, Xuefeng, Xiang Zhou, and Lingjiong Zhu (2018), "Transform analysis for Hawkes processes with applications in dark pool trading." *Quantitative Finance*, 18, 265–282.

Gao, Xuefeng and Lingjiong Zhu (2018a), "Functional central limit theorems for stationary Hawkes processes and application to infinite-server queues." *Queueing Systems*, 1–46.

Gao, Xuefeng and Lingjiong Zhu (2018b), "Large deviations and applications for Markovian Hawkes processes with a large initial intensity." *Bernoulli*, 24, 2875–2905.

Gao, Xuefeng and Lingjiong Zhu (2018c), "Limit theorems for Markovian Hawkes processes with a large initial intensity." *Stochastic Processes and their Applications*, 128, 3807–3839.

Gao, Xuefeng and Lingjiong Zhu (2019), "Affine point processes: Refinements to large-time asymptotics." *arXiv preprint arXiv:1903.06371*.

Garnett, Ofer, Avishai Mandelbaum, and Martin Reiman (2002), "Designing a call center with impatient customers." *Manufacturing & Service Operations Management*, 4, 208–227.

Gilbert, EN and HO Pollak (1960), "Amplitude distribution of shot noise." *The Bell System Technical Journal*, 39, 333–350.

Glynn, Peter W., L. Jeff Hong, and Xiaowei Zhang (2019), "Modeling call center arrivals: A tale of three timescales." *Working paper*.

GM Cruise LLC (2019), "Autonomous vehicle disengagement reports." *California Department of Motor Vehicles*.

Guo, Xin, Zhao Ruan, and Lingjiong Zhu (2015), "Dynamics of order positions and related queues in a limit order book." *arXiv preprint arXiv:1505.04810*.

Gupta, Amrita, Mehrdad Farajtabar, Bistra Dilkina, and Hongyuan Zha (2018), "Discrete interventions in Hawkes processes with applications in invasive species management." In *IJCAI*, 3385–3392.

Gupta, RP and GC Jain (1974), "A generalized Hermite distribution and its properties." *SIAM Journal on Applied Mathematics*, 27, 359–363.

Gurvich, Itai, Junfei Huang, and Avishai Mandelbaum (2013), "Excursion-based universal approximations for the Erlang-A queue in steady-state." *Mathematics of Operations Research*, 39, 325–373.

Gurvich, Itai and Ward Whitt (2009), "Queue-and-idleness-ratio controls in many-server service systems." *Mathematics of Operations Research*, 34, 363–396.

Gurvich, Itai and Ward Whitt (2010), "Service-level differentiation in many-server service systems via queue-ratio routing." *Operations Research*, 58, 316–328.

Hale, Jack K and Sjoerd M Verduyn Lunel (2013), *Introduction to functional differential equations*, volume 99. Springer Science & Business Media.

Halfin, Shlomo and Ward Whitt (1981), "Heavy-traffic limits for queues with many exponential servers." *Operations Research*, 29, 567–588.

Halpin, Peter F and Paul De Boeck (2013), "Modelling dyadic interaction with hawkes processes." *Psychometrika*, 78, 793–814.

Harrison, J Michael and Sidney I Resnick (1976), "The stationary distribution and first exit probabilities of a storage process with general release rule." *Mathematics of Operations Research*, 1, 347–358.

Harrison, J Michael and Sidney I Resnick (1978), "The recurrence classification of risk and storage processes." *Mathematics of Operations Research*, 3, 57–66.

Hawkes, Alan G (1971), "Point spectra of some mutually exciting point processes." *Journal of the Royal Statistical Society. Series B (Methodological)*, 438–443.

Hawkes, Alan G (1971), "Spectra of some self-exciting and mutually exciting point processes." *Biometrika*, 58, 83–90.

Hawkes, Alan G. and David Oakes (1974), "A cluster process representation of a self-exciting process." *Journal of Applied Probability*, 11, 493–503.

Henry-Labordere, Pierre (2008), *Analysis, geometry, and modeling in finance: Advanced methods in option pricing*. Chapman and Hall/CRC.

Ibrahim, Rouba, Han Ye, Pierre L'Ecuyer, and Haipeng Shen (2016), "Modeling and forecasting call center arrivals: A literature survey and a case study." *International Journal of Forecasting*, 32, 865–874.

Jadhav, Akshay (2018), "Autonomous vehicle market by level of automation (level 3, level 4, and level 5) and component (hardware, software, and service) and application (civil, robo taxi, self-driving bus, ride share, self-driving

truck, and ride hail) - global opportunity analysis and industry forecast, 2019-2026." *Allied Market Research*.

Javier, Kayla and Brian Fralix (2019), "An exact analysis of a class of Markovian bitcoin models."

Jennings, Otis B, Avishai Mandelbaum, William A Massey, and Ward Whitt (1996), "Server staffing to meet time-varying demand." *Management Science*, 42, 1383–1394.

Kalra, Nidhi and Susan M Paddock (2016), "Driving to safety: How many miles of driving would it take to demonstrate autonomous vehicle reliability?" *Transportation Research Part A: Policy and Practice*, 94, 182–193.

Karlis, Dimitris and Evdokia Xekalaki (2005), "Mixed poisson distributions." *International Statistical Review*, 73, 35–58.

Kaspi, Haya (1984), "Storage processes with Markov additive input and output." *Mathematics of Operations Research*, 9, 424–440.

Kaspi, Haya and David Perry (1989), "On a duality between a non-Markovian storage/production process and a Markovian dam process with state-dependent input and output." *Journal of Applied Probability*, 26, 835–844.

Kella, Offer (2009), "On growth-collapse processes with stationary structure and their shot-noise counterparts." *Journal of Applied Probability*, 46, 363–371.

Kella, Offer and Andreas Löpker (2010), "A Markov-modulated growth collapse model." *Probability in the Engineering and Informational Sciences*, 24, 99–107.

Kella, Offer and Ward Whitt (1999), "Linear stochastic fluid networks." *Journal of Applied Probability*, 36, 244–260.

Kelly, Frank P (2011), *Reversibility and stochastic networks*. Cambridge University Press.

Kemp, CD and Adrienne W Kemp (1965), "Some properties of the 'Hermite' distribution." *Biometrika*, 52, 381–394.

King, Martin Luther, Jr (1963), "Letter from Birmingham jail."

Knuth, Donald E (1993), "Johann Faulhaber and sums of powers." *Mathematics of Computation*, 61, 277–294.

Koops, David (2018), "Predicting the confirmation time of bitcoin transactions." *arXiv preprint arXiv:1809.10596*.

Koops, David T, Onno J Boxma, and MRH Mandjes (2017), "Networks of $\cdot/G/\infty$ queues with shot-noise-driven arrival intensities." *Queueing Systems*, 86, 301–325.

Koops, David T, Mayank Saxena, Onno J Boxma, and Michel Mandjes (2018), "Infinite-server queues with Hawkes input." *Journal of Applied Probability*, 55, 920–943.

Krumin, Michael, Inna Reutsky, and Shy Shoham (2010), "Correlation-based analysis and generation of multiple spike trains using Hawkes models with an exogenous input." *Frontiers in Computational Neuroscience*, 4, 147.

Lanctot, Roger (2017), "Accelerating the future: The economic impact of the emerging passenger economy." *Strategy Analytics*.

Lane, John A (1984), "The central limit theorem for the Poisson shot-noise process." *Journal of Applied Probability*, 21, 287–301.

Laub, Patrick J., Thomas Taimre, and Philip K. Pollett (????), "Hawkes processes." *arXiv preprint arXiv:1507.02822*.

L'Ecuyer, Pierre, Klas Gustavsson, and Leif Olsson (2018), "Modeling bursts in the arrival process to an emergency call center." In *2018 Winter Simulation Conference (WSC)*, 525–536, IEEE.

Lee, Soon Seok, Ho Woo Lee, Seung Hyun Yoon, and Kyung C Chae (1995), "Batch arrival queue with N-policy and single vacation." *Computers & Operations Research*, 22, 173–189.

Lenhart, Suzanne and John T Workman (2007), *Optimal control applied to biological models*. CRC Press.

Li, Liangda and Hongyuan Zha (2018), "Energy usage behavior modeling in energy disaggregation via Hawkes processes." *ACM Transactions on Intelligent Systems and Technology (TIST)*, 9, 36.

Lindvall, Torgny (1977), "A probabilistic proof of blackwell's renewal theorem." *The Annals of Probability*, 5, 482–485.

Liu, Liming and James GC Templeton (1993), "Autocorrelations in infinite server batch arrival queues." *Queueing Systems*, 14, 313–337.

Lu, Yi, Qiaomin Xie, Gabriel Kliot, Alan Geller, James R Larus, and Albert Greenberg (2011), "Join-idle-queue: A novel load balancing algorithm for dynamically scalable web services." *Performance Evaluation*, 68, 1056–1071.

Lucantoni, David M (1991), "New results on the single server queue with a batch Markovian arrival process." *Communications in Statistics. Stochastic Models*, 7, 1–46.

Lum, Kristian, Samarth Swarup, Stephen Eubank, and James Hawdon (2014), "The contagious nature of imprisonment: an agent-based model to explain racial disparities in incarceration rates." *Journal of the Royal Society Interface*, 11, 20140409.

Lundgard, Alan, Yiwei Yang, Maya L Foster, and Walter S Lasecki (2018), "Bolt: Instantaneous crowdsourcing via just-in-time training." In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 467, ACM.

Malmgren, R Dean, Daniel B Stouffer, Adilson E Motter, and Luís AN Amaral (2008), "A poissonian explanation for heavy tails in e-mail communication." *Proceedings of the National Academy of Sciences*, 105, 18153–18158.

Mandelbaum, Avishai and Sergey Zeltyn (2007), "Service engineering in action: the Palm/Erlang-A queue, with applications to call centers." In *Advances in services innovations*, 17–45, Springer.

Massey, William A and Jamol Pender (2011), "Poster: skewness variance approximation for dynamic rate multiserver queues with abandonment." *ACM SIGMETRICS Performance Evaluation Review*, 39, 74–74.

Massey, William A and Jamol Pender (2013), "Gaussian skewness approximation for dynamic rate multi-server queues with abandonment." *Queueing Systems*, 75, 243–277.

Massey, William A and Ward Whitt (1994), "An analysis of the modified offered-load approximation for the nonstationary Erlang loss model." *The Annals of Applied Probability*, 1145–1160.

Masuda, Naoki, Taro Takaguchi, Nobuo Sato, and Kazuo Yano (2013), "Self-

exciting point process modeling of conversation event sequences." In *Temporal Networks*, 245–264, Springer.

Masuyama, Hiroyuki and Tetsuya Takine (2002), "Analysis of an infinite-server queue with batch Markovian arrival streams." *Queueing Systems*, 42, 269–296.

McKelvey, Karissa Rae and Filippo Menczer (2013), "Truthy: Enabling the study of online social networks." In *Proceedings of the 2013 conference on Computer supported cooperative work companion*, 23–26, ACM.

Mei, Hongyuan and Jason M Eisner (2017), "The neural hawkes process: A neurally self-modulating multivariate point process." In *Advances in Neural Information Processing Systems*, 6754–6764.

Melamed, Benjamin and Ward Whitt (1990), "On arrivals that see time averages." *Operations Research*, 38, 156–172.

Miller Jr, Rupert G (1959), "A contribution to the theory of bulk queues." *Journal of the Royal Statistical Society. Series B (Methodological)*, 320–337.

Milne, Robin Kingsley and Mark Westcott (1993), "Generalized multivariate Hermite distributions and related point processes." *Annals of the Institute of Statistical Mathematics*, 45, 367–381.

Mirzaeian, Neda, Soo-Haeng Cho, and Alan Andrew Scheller-Wolf (2018), "A queueing model and analysis for autonomous vehicles on highways." *Available at SSRN 3278330*.

Moler, Cleve and Charles Van Loan (2003), "Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later." *SIAM Review*, 45, 3–49.

Mordfin, Robin (2015), "Why long lines can be good for shoppers, and business."

Neuts, Marcel F (1978), "An algorithmic solution to the $GI/M/C$ queue with group arrivals." Technical report, Delaware Univ. Newark Dept. of Statistics and Computer Science.

New York City Taxi & Limousine Commission (2014), "2014 taxicab fact book." *City of New York*.

New York City Taxi & Limousine Commission (2018), "2018 TLC factbook." *City of New York*.

Niyirora, Jerome and Jamol Pender (2016), "Optimal staffing in nonstationary service centers with constraints." *Naval Research Logistics (NRL)*, 63, 615–630.

Oelschlager, Karl (1984), "A martingale approach to the law of large numbers for weakly interacting stochastic processes." *The Annals of Probability*, 458–479.

Ogata, Yosihiko (1981), "On Lewis' simulation method for point processes." *IEEE Transactions on Information Theory*, 27, 23–31.

Ogata, Yosihiko (1988), "Statistical models for earthquake occurrences and residual analysis for point processes." *Journal of the American Statistical association*, 83, 9–27.

Oksendal, Bernt (2013), *Stochastic differential equations: an introduction with applications*. Springer Science & Business Media.

Otter, Richard (1949), "The multiplicative process." *The Annals of Mathematical Statistics*, 206–224.

Ottopia Team (2019), "Advanced teleoperation: Unlocking the full potential of level 4 fleets today." *Medium*.

Pang, Guodong and Ward Whitt (2012), "Infinite-server queues with batch arrivals and dependent service times." *Probability in the Engineering and Informational Sciences*, 26, 197–220.

Pender, Jamol (2013), "Poisson and Gaussian approximations for multi-server queues with batch arrivals and batch abandonment."

Pender, Jamol (2014a), "Gram Charlier expansion for time varying multiserver queues with abandonment." *SIAM Journal on Applied Mathematics*, 74, 1238–1265.

Pender, Jamol (2014b), "Laguerre polynomial expansions for time varying multiserver queues with abandonment."

Pender, Jamol (2014c), "A Poisson–Charlier approximation for nonstationary queues." *Operations Research Letters*, 42, 293–298.

Pender, Jamol (2015a), "Nonstationary loss queues via cumulant moment approximations." *Probability in the Engineering and Informational Sciences*, 29, 27–49.

Pender, Jamol (2015b), "The truncated normal distribution: Applications to queues with impatient customers." *Operations Research Letters*, 43, 40–45.

Pender, Jamol (2016a), "An analysis of nonstationary coupled queues." *Telecommunication Systems*, 61, 823–838.

Pender, Jamol (2016b), "Risk measures and their application to staffing nonsta-

tionary service systems." *European Journal of Operational Research*, 254, 113–126.

Pender, Jamol (2016c), "Sampling the functional Kolmogorov forward equations for nonstationary queueing networks." *INFORMS Journal on Computing*, 29, 1–17.

Pender, Jamol and William A Massey (2017), "Approximating and stabilizing dynamic rate Jackson networks with abandonment." *Probability in the Engineering and Informational Sciences*, 31, 1–42.

Pender, Jamol and Tuan Phung-Duc (2016), "A law of large numbers for M/M/c/delayoff-setup queues with nonstationary arrivals." In *International Conference on Analytical and Stochastic Modeling Techniques and Applications*, 253–268, Springer.

Pender, Jamol, Richard H Rand, and Elizabeth Wesson (2017a), "Queues with choice via delay differential equations." *International Journal of Bifurcation and Chaos*, 27, 1730016.

Pender, Jamol, Richard H Rand, and Elizabeth Wesson (2017b), "Strong approximations for queues with customer choice and constant delays."

Pender, Jamol, Richard H Rand, and Elizabeth Wesson (2018), "An analysis of queues with delayed information and time-varying arrival rates." *Nonlinear Dynamics*, 91, 2411–2427.

Perry, David and Wolfgang Stadje (2003), "Duality of dams via mountain processes." *Operations Research Letters*, 31, 451–458.

Prabhu, Narahari Umanath (2012), *Stochastic storage processes: queues, insurance*

risk, dams, and data communication, volume 15. Springer Science & Business Media.

Qin, Ziyuan and Jamol Pender (2017), "Dynamic control for nonstationary queueing networks."

Rambaldi, Marcello, Emmanuel Bacry, and Fabrizio Lillo (2017), "The role of volume in order book dynamics: a multivariate Hawkes process analysis." *Quantitative Finance*, 17, 999–1020.

Reed, Josh (2009), "The $G/GI/N$ queue in the Halfin–Whitt regime." *The Annals of Applied Probability*, 19, 2211–2269.

Rice, John (1977), "On generalized shot noise." *Advances in Applied Probability*, 9, 553–565.

Rizoiu, Marian-Andrei, Swapnil Mishra, Quyu Kong, Mark Carman, and Lexing Xie (2018), "SIR-Hawkes: Linking epidemic models and Hawkes processes to model diffusions in finite populations." In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*, 419–428, International World Wide Web Conferences Steering Committee.

Rizoiu, Marian-Andrei, Lexing Xie, Scott Sanner, Manuel Cebrian, Honglin Yu, and Pascal Van Hentenryck (2017), "Expecting to be HIP: Hawkes intensity processes for social media popularity." In *Proceedings of the 26th International Conference on World Wide Web*, 735–744, International World Wide Web Conferences Steering Committee.

Rosenszweig, Amit (2019), "Av industry needs teleoperation laws." *Automotive News*.

Rubinovitch, Michael and JW Cohen (1980), "Level crossings and stationary distributions for general dams." *Journal of Applied Probability*, 17, 218–226.

Sachs, Rainer K, Pei-Li Chen, Philip J Hahnfeldt, and Lynn R Hlatky (1992), "DNA damage caused by ionizing radiation." *Mathematical Biosciences*, 112, 271–303.

SAE On-Road Automated Vehicle Standards Committee (2018), "Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles." *SAE International: Warrendale, PA, USA*.

Sam Schwartz Engineering (2019), "LADOT taxi and for-hire vehicle study." *Los Angeles Department of Transportation*.

Sawers, Paul (2018), "Ottopia's remote assistance platform for autonomous cars combines humans with ai." *VentureBeat*.

Shanbhag, DN (1966), "On infinite server queues with batch arrivals." *Journal of Applied Probability*, 3, 274–279.

Singh, Sarabjeet and Christopher R Myers (2014), "Outbreak statistics and scaling laws for externally driven epidemics." *Physical Review E*, 89, 042108.

Sullivan, J, L Crone, and J Jalickee (1980), "Approximation of the unit step function by a linear combination of exponential functions." *Journal of Approximation Theory*, 28, 299–308.

Takagi, Hideaki and Yoshitaka Takahashi (1991), "Priority queues with batch Poisson arrivals." *Operations Research Letters*, 10, 225–232.

Tibken, Shara (2018), "Waymo CEO: Autonomous cars won't ever be able to drive in all conditions." *CNET*.

Truccolo, Wilson, Uri T Eden, Matthew R Fellows, John P Donoghue, and Emery N Brown (2005), "A point process framework for relating neural spiking activity to spiking history, neural ensemble, and extrinsic covariate effects." *Journal of Neurophysiology*, 93, 1074–1089.

Van der Hofstad, Remco and Michael Keane (2008), "An elementary proof of the hitting time theorem." *The American Mathematical Monthly*, 115, 753–756.

Waymo LLC (2019), "Autonomous vehicle disengagement reports." *California Department of Motor Vehicles*.

Willmot, Gord (1986), "Mixed compound poisson distributions." *ASTIN Bulletin: The Journal of the IAA*, 16, S59–S79.

Wolff, Ronald W (1982), "Poisson arrivals see time averages." *Operations Research*, 30, 223–231.

Wu, Peng, Marcello Rambaldi, Jean-François Muzy, and Emmanuel Bacry (2019), "Queue-reactive Hawkes models for the order flow." *arXiv preprint arXiv:1901.08938*.

Xie, Qiaomin, Mayank Pundir, Yi Lu, Cristina L Abad, and Roy H Campbell (2017), "Pandas: Robust locality-aware scheduling with stochastic delay optimality." *IEEE/ACM Transactions on Networking (TON)*, 25, 662–675.

Xu, Hongteng, Dixin Luo, and Hongyuan Zha (2017), "Learning Hawkes processes from short doubly-censored event sequences." In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 3831–3840, JMLR. org.

Xu, Hongteng, Yi Zhen, and Hongyuan Zha (2015), "Trailer generation via a point process-based visual attractiveness model." In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

Yan, Junchi, Yu Wang, Ke Zhou, Jin Huang, Chunhua Tian, Hongyuan Zha, and Weishan Dong (2013), "Towards effective prioritizing water pipe replacement and rehabilitation." In *Twenty-Third International Joint Conference on Artificial Intelligence*.

Yao, David D (1985), "Some results for the queues $M^X/M/c$ and $GI^X/G/c$." *Operations Research Letters*, 4, 79–83.

Yao, David DW, ML Chaudhry, and JGC Templeton (1984), "On bounds for bulk arrival queues." *European Journal of Operational Research*, 15, 237–243.

Yekkehkhany, Ali, Avesta Hojjati, and Mohammad H Hajiesmaili (2018), "GB-PANDAS:: Throughput and heavy-traffic optimality analysis for affinity scheduling." *ACM SIGMETRICS Performance Evaluation Review*, 45, 2–14.

Yeo, Geoffrey (1976), "A dam with general release rule." *The ANZIAM Journal*, 19, 469–477.

Yeo, GF (1974), "A finite dam with exponential release." *Journal of Applied Probability*, 11, 122–133.

Zhang, Xiao-Wei, Peter W Glynn, Kay Giesecke, and Jose Blanchet (2009), "Rare event simulation for a generalized Hawkes process." In *Proceedings of the 2009 Winter Simulation Conference (WSC)*, 1291–1298, IEEE.

Zhang, Xiaowei, Jose Blanchet, Kay Giesecke, and Peter W Glynn (2015), "Affine point processes: Approximation and efficient simulation." *Mathematics of Operations Research*, 40, 797–819.

Zhang, Yingjie, Jinyang Zheng, and Yong Tan (2019), "From automobile to autonomous: Does self-driving improve traffic conditions?" *Available at SSRN 3374040*.

Zhang, Zhizheng (1997), "The linear algebra of the generalized Pascal matrix." *Linear Algebra and Its Applications*, 250, 51–60.

Zhao, Yiqiang (1994), "Analysis of the $GI^X/M/c$ model." *Queueing Systems*, 15, 347–364.

Zino, Lorenzo, Alessandro Rizzo, and Maurizio Porfiri (2018), "Modeling memory effects in activity-driven networks." *SIAM Journal on Applied Dynamical Systems*, 17, 2830–2854.