This article was downloaded by: [71.190.135.153] On: 29 June 2020, At: 05:50
Publisher: Institute for Operations Research and the Management Sciences (INFORMS)
INFORMS is located in Maryland, USA

Operations Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Beyond Heavy-Traffic Regimes: Universal Bounds and Controls for the Single-Server Queue

Junfei Huang, Itai Gurvich

To cite this article:
Junfei Huang, Itai Gurvich (2018) Beyond Heavy-Traffic Regimes: Universal Bounds and Controls for the Single-Server Queue. Operations Research 66(4):1168-1188. https://doi.org/10.1287/opre.2017.1715

Full terms and conditions of use: https://pubsonline.informs.org/Publications/Librarians-Portal/PubsOnLine-Terms-and-Conditions

This article may be used only for the purposes of research, teaching, and/or private study. Commercial use or systematic downloading (by robots or other automatic processes) is prohibited without explicit Publisher approval, unless otherwise noted. For more information, contact permissions@informs.org.

The Publisher does not warrant or guarantee the article's accuracy, completeness, merchantability, fitness for a particular purpose, or non-infringement. Descriptions of, or references to, products or publications, or inclusion of an advertisement in this article, neither constitutes nor implies a guarantee, endorsement, or support of claims made of that product, publication, or service.

Copyright © 2018, INFORMS

Please scroll down for article—it is on subsequent pages

With 12,500 members from nearly 90 countries, INFORMS is the largest international association of operations research (O.R.) and analytics professionals and students. INFORMS provides unique networking and learning opportunities for individual professionals, and organizations of all types and sizes, to better understand and use O.R. and analytics tools and methods to transform strategic visions and achieve better outcomes.
For more information on INFORMS, its publications, membership, or meetings visit http://www.informs.org

http://pubsonline.informs.org/journal/opre/

OPERATIONS RESEARCH
Vol. 66, No. 4, July–August 2018, pp. 1168–1188
ISSN 0030-364X (print), ISSN 1526-5463 (online)

Beyond Heavy-Traffic Regimes: Universal Bounds and Controls for the Single-Server Queue

Junfei Huang,[a] **Itai Gurvich**[b]

[a] Department of Decision Sciences and Managerial Economics, CUHK Business School, Chinese University of Hong Kong, Shatin, Hong Kong; [b] School of Operations Research and Information Engineering, Cornell Tech, New York, New York 10011

Contact: junfeih@cuhk.edu.hk, http://orcid.org/0000-0002-3764-354X (JH); gurvich@cornell.edu, http://orcid.org/0000-0001-9746-7755 (IG)

Received: May 9, 2016
Revised: April 3, 2017; September 24, 2017
Accepted: November 15, 2017
Published Online in Articles in Advance: July 26, 2018

Subject Classifications: queues: approximations, optimization; probability: diffusion
Area of Review: Stochastic Models

https://doi.org/10.1287/opre.2017.1715

Copyright: © 2018 INFORMS

Abstract. Central-limit (Brownian) approximations are widely used for the performance analysis and optimization of queueing networks because of their tractability relative to the original queueing models. The stationary distributions of the approximations are used as proxies for those of the queues. The convergence of suitably scaled and centered processes provides mathematical support for the use of these Brownian models. As with the central limit theorem, to establish convergence, one must impose assumptions directly on the primitives or indirectly on the parameters of a related optimization problem. These assumptions reflect an interpretation of the underlying parameters—a classification into so-called heavy-traffic regimes that specify a scaling relationship between the utilization and the arrival rate. Here, it matters whether a utilization of 90% in a queue with an arrival rate of $\lambda = 100$ is read as $\rho(\lambda) = 0.9 = 1 - 1/\sqrt{\lambda}$ or as $\rho(\lambda) \equiv 0.9$, because different interpretations lead to different limits and, in turn, to different approximations. However, from a heuristic point of view, there is an immediate Brownian (i.e., normal) analogue of the queueing model that is derived directly from the primitives and requires no scaling interpretation of the parameters. In this model, the drift is that of the original queue, and the noise term is replaced by a Brownian motion with the same variance. This is intuitive and appealing as a tool, but it lacks mathematical justification. In this paper, we prove that for the fundamental $M/GI/1 + GI$ queue, this direct intuitive approach works: the Brownian model is accurate uniformly over a family of patience distributions and universally in the heavy-traffic regime. The validity of this approach extends to dynamic control in that the solution of the directly derived diffusion control problem is universally accurate. To build mathematical support for the accuracy of this model, we introduce a framework built around "queue families" that allows us to treat various patience distributions simultaneously, and it uncovers the role of a concentration property of the queue.

Funding: This research was supported in part by the National Science Foundation [NSF Grant CMMI-1662294] and the Hong Kong Research Grants Council [Projects 24500314 and 14502815].
Supplemental Material: The e-companion is available at https://doi.org/10.1287/opre.2017.1715.

Keywords: *M/GI/1 + GI* • universal approximation • stationary distribution • Stein's method

1. Introduction

***The Basic Building Block* (*the* $M/M/1$ *Queue*).** The fundamental building block of queueing theory is the $M/M/1$ queue in which Poisson arrivals with Exponential service requirements are processed by a single server. The queue is stable when the arrival rate λ is strictly smaller than the service rate μ ($\rho := \lambda/\mu < 1$), in which case the stationary waiting time, W, has the distribution $\mathbb{P}\{W > x\} = \rho e^{-\mu(1-\rho)x}$, with moments

$$\mathbb{E}[W^k] = \frac{\rho k!}{(\mu(1-\rho))^k}.$$

Viewed as a process, the waiting time in the $M/M/1$ queue (which equates with the workload) satisfies the evolution

$$W(t) = W(0) + \sum_{i=1}^{A(t)} s_i - (t - I(t))$$
$$= W(0) + \rho t - (t - I(t)) + \left(\sum_{i=1}^{A(t)} s_i - \rho t\right),$$

where $\{s_i, i \geqslant 1\}$ represent the customer service requirements and are Exponential random variables with a mean of $1/\mu$, $A(t)$ is the number of arrivals by time t, and $I(t)$ is the cumulative idle time of the server by time t. The compound Poisson input satisfies, at each t, the central-limit-theorem approximation $\sum_{i=1}^{A(t)} s_i \approx \rho t + Z(t)$, where $Z(t)$ is a zero mean normal random variable with variance $\lambda \mathbb{E}[s_1^2]t = 2\lambda t/\mu^2$. It is thus heuristically natural to replace the input process

by $\rho t + (\sqrt{2\lambda}/\mu)B(t)$, where $B = (B(t), t \geqslant 0)$ is a standard Brownian motion, and propose, as an approximation, the *Brownian queue*:

$$\hat{W}(t) = W(0) + \rho t - (t - \hat{I}(t)) + \frac{\sqrt{2\lambda}}{\mu} B(t),$$

where the Brownian idleness $\hat{I}(t)$ is nonnegative and increases only when $\hat{W}(t) = 0$, keeping the latter positive ($\hat{W}$ is a so-called *reflected Brownian motion*). Direct derivations of a Brownian analogue of a queueing network have a long history that precedes the rigorization through limit theorems; see Harrison and Nguyen (1993) and Harrison and Williams (1987) for an exposition and discussion of this approach.

If $\rho < 1$, the Brownian queue's stationary distribution is Exponential with a mean of $2\lambda/(2\mu^2(1-\rho)) = \rho/(\mu(1-\rho))$ so that $\mathbb{E}[\hat{W}^k] = k!\rho^k/(\mu(1-\rho))^k$. The approximation gap for the kth moment is

$$\begin{aligned}
&|\mathbb{E}[W^k] - \mathbb{E}[\hat{W}^k]| \\
&\quad = \frac{\rho k!}{(\mu(1-\rho))^k}(1-\rho^{k-1}) = \frac{1-\rho^{k-1}}{\rho^{k-1}}\mathbb{E}[\hat{W}^k] \\
&\quad = \frac{k(1-\rho^{k-1})}{\mu\rho^{k-2}(1-\rho)}\mathbb{E}[\hat{W}^{k-1}] \leqslant \frac{k(k-1)}{\rho^{k-2}}\mathbb{E}[s_1]\mathbb{E}[\hat{W}^{k-1}],
\end{aligned}$$

where the inequality follows from $\rho < 1$ (required for stability) and $(1-\rho^{k-1})/(1-\rho) = 1 + \rho + \cdots + \rho^{k-2} \leqslant k-1$.

The gap is 0 for the first moment ($k = 1$). For the second moment ($k = 2$),

$$|\mathbb{E}[\hat{W}^2] - \mathbb{E}[W^2]| = 2\mathbb{E}[s_1]\mathbb{E}[\hat{W}],$$

and for $k > 2$,

$$|\mathbb{E}[W^k] - \mathbb{E}[\hat{W}^k]| \approx k(k-1)\mathbb{E}[s_1]\mathbb{E}[\hat{W}^{k-1}], \tag{1}$$

as ρ approaches 1.

In contrast with the $M/M/1$ queue, the $M/GI/1+GI$ queue (a single-server queue with general service time and patience distributions) is analytically intractable, *even* in its Markovian instance, the $M/M/1+M$ queue.[1] A notable exception is the $M/G/1$ (infinite patience) queue, where the Pollaczek–Khinchine formula captures the *first* moment of both the queue and its Brownian counterpart. It is this intractability that renders Brownian approximations valuable.

Here, we prove that the approximation quality in (1) persists in the generality of the $M/GI/1+GI$ queue:

$$|\mathbb{E}[W^k] - \mathbb{E}[\hat{W}^k]| \leqslant C\mathbb{E}[s_1]\mathbb{E}[\hat{W}^{k-1}], \tag{2}$$

for a constant C that does not depend on λ or μ and depends only in a limited way on the patience distribution. Here, $\hat{W}$ is derived from the intuitive Brownian counterpart of the $M/GI/1+GI$ queue whose heuristic derivation, as in the case of the $M/M/1$ queue, does not require familiarity with limit theory.

With finite patience, the waiting time of a customer arriving at time t is the minimum of his willingness to wait (his patience) and the offered waiting time $V(t)$. The latter is the sum of the residual service time of the customer in service and the service requirements of the customers in the queue who will not abandon before being served. Because $V(t)$ captures the time the arriving customer will have to wait to enter service, it is often referred to as the *virtual* waiting time at t; it materializes as a customer's real waiting time only if the customer's patience exceeds it. With infinite patience, a customer's waiting time equals his virtual waiting time.

The process $(V(t), t \geqslant 0)$ is the key mathematical object. Denoting by F_a the patience distribution and by $\bar{F}_a := 1 - F_a$ its complement, $\lambda\bar{F}_a(V(t))$ is the rate of "effective" arrivals at time t (i.e., those that increase the virtual wait). Each of these patient customers brings, in expectation, $1/\mu$ work. The instantaneous drift is then $\lambda\bar{F}_a(V(t))/\mu - 1 = \rho\bar{F}_a(V(t)) - 1$.

In the $M/M/1$ case, the variance of the compound Poisson input is $\lambda\mathbb{E}[s_1^2]$. With abandonment, it seems appropriate to replace λ with the throughput rate $\lambda \wedge \mu$. We arrive at the following (reflected) diffusion:

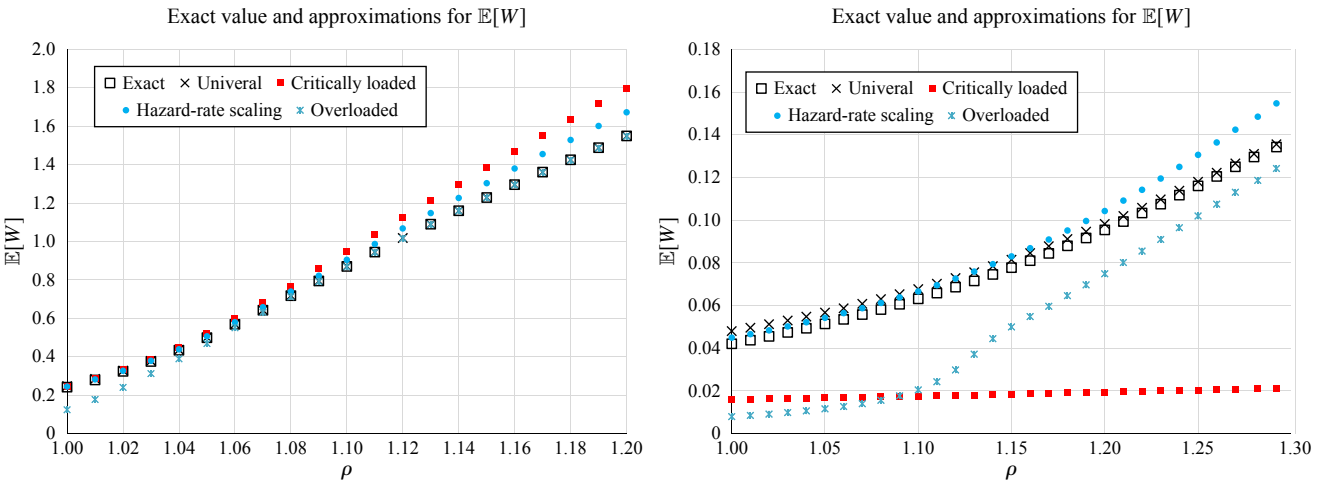
$$\begin{aligned}
\hat{V}(t) = V(0) &+ \int_0^t \rho\bar{F}_a(\hat{V}(s))\,ds - t \\
&+ \sqrt{(\lambda \wedge \mu)\mathbb{E}[s_1^2]}B(t) + \hat{I}(t).
\end{aligned}$$

With infinite patience ($\bar{F}_a(x) \equiv 1$) and $\rho < 1$, the drift term reduces to $-(1-\rho)t$ and the diffusion coefficient to $\lambda\mathbb{E}[s_1^2]$, and we recover the Brownian counterpart of the $M/GI/1$ queue. The stationary distribution of $\hat{V}(t)$ can be expressed in a closed form (see Equation (4)) that can be used for performance analysis and optimization.

That the diffusion model is tractable is not, however, enough. We simply constructed this Brownian queue by keeping the (state-dependent) drift and replacing the centered input process by a Brownian motion with the same variance. A question remains as to whether this model, derived heuristically, can be universally used as a valid approximation for the original queue. Our answer is a strong affirmative.

Universality in Regimes. Universality is best understood in contrast to the heavy-traffic limit theory. The implicit idea of the heavy-traffic theory is to embed a given queue as an element in a convergent sequence of queues. The limit of this sequence is subsequently used as an approximation for the original queue. In this embedding, interpretation is unavoidable because we can embed an $M/GI/1+GI$ queue with $\rho = 1.1$ and $\lambda = 110$ in at least two distinct ways: we can either treat the utilization 1.1 as $\rho(\lambda) \approx 1 + 1/\sqrt{\lambda}$, in which case

Figure 1. (Color online) Comparison of the First-Moment Approximations for the $M/M/1+GI$ Queue



Notes. Left: $\mu=100$, patience $F_a(x)=1-e^{-0.2x}-0.1xe^{-0.2x}$ (Erlang), and varying ρ. The actual performance is plotted as a square and the universal as an ×. That the × is inside the square captures the universal (in ρ) precision of the approximation. This is the only approximation that shows consistently good performance. Right: $\mu=100$ and $F_a(x)=0.9(1-e^{-x})+0.1(1-e^{-200x})$ (Hyperexponential). The critical loading and overloaded approximation are inaccurate. Implicit in these approximations is that the patience parameters are small or at least moderate relative to λ,μ. This is violated in this example. The universal and hazard-rate approximations both perform well. The hazard-rate approximation performs slightly better for lower values of ρ, but its performance deteriorates somewhat for larger values.

the appropriate analysis is the heavy-traffic analysis of a sequence of *critically loaded* queues with $\sqrt{\lambda}(1-\rho)\approx -1$, or we can treat the utilization 1.1 as a constant that does not scale with λ, $\rho(\lambda)\equiv 1.1$, in which case the appropriate limit approximation is obtained by studying a sequence of *overloaded* queues. These two embeddings lead to two different limits with different stationary distributions and hence to different approximations; see Ward and Glynn (2005) and Jennings and Reed (2012). Figure 1 compares the expected average delay against four approximations: our proposed universal approximation, a critical loading approximation (based on modeling the patience only through its density at 0), an overloaded-queue approximation, and a hazard-rate approximation developed (for critically loaded queues) in Reed and Ward (2008). A detailed discussion of these approximations and further numerical examples are provided in the online appendix to this paper.

An informative discussion of the heavy-traffic embedding step appears in Ward and Glynn (2003), who offer, as a remedy, a common (universal) *process* approximation for the $M/M/1+M$ queue. They prove that in a process convergence sense, the gap between the suitably scaled queueing process and their universal Brownian queue is small across multiple heavy-traffic regimes. Ward (2012) advances a similar idea in which a universal diffusion process is proposed for the $GI/M/N+GI$ queue, with universality relative to the number of servers, N (single server or many servers), but restricted to critical loading.

Universality in Concentration. With Exponential patience

$$V\approx \bar{w}+\frac{1}{\sqrt{\lambda}}\mathcal{N},$$

where $\bar{w}$ is the first-order approximation[2] and $\mathcal{N}$ is a random variable whose parameters (mean and standard deviation) do not depend on λ. In particular, the *concentration* of the stationary distribution around $\bar{w}$ is of the order of $\lambda^{-1/2}$; to obtain meaningful limits, we must consider the scaling $\sqrt{\lambda}(V-\bar{w})$ as is common in the heavy-traffic literature. But this scaling is restrictive. Fixing the regime, the concentration can vary with the patience distribution, and fixing the patience distribution, the concentration can vary across regimes. If the patience is a shape-2 Erlang distribution, for example, an overloaded queue has a $1/\sqrt{\lambda}$ concentration but the critically loaded one has $V\approx\bar{w}+(1/\lambda^{1/3})\mathcal{N}$ (with $\bar{w}=0$). For meaningful limits one must use the scaling $\lambda^{1/3}(V-\bar{w})$. Our framework, based on queue families, allows us to group together patience distributions that differ in their natural concentration and obtain a unified result that is "blind" to these differences.

Our notion of queue families requires that we formalize this notion of concentration and ground it in the queue's primitives. Loosely speaking, the concentration is the ratio of the variation and the drift. In the stable $M/GI/1$ queue, the drift is $-(1-\rho)$ and the variance is $\sigma^2=\lambda\mathbb{E}[s_1^2]$. The ratio between the variance and the absolute value of the drift is $\lambda\mathbb{E}[s_1^2]/(1-\rho)$, which, by the Pollaczek–Khinchine formula, is twice the expected waiting time. If $1-\rho\approx 1/\sqrt{\lambda}$, the concentration is $1/\sqrt{\lambda}$.

In the $M/GI/1+GI$ queue, the drift is state dependent and the concentration is a fixed point—namely, the value of c at which the ratio of variation to drift equals c:

$$\frac{\lambda\mathbb{E}[s_1^2]((1/\rho)\wedge 1)}{|\rho\bar{F}_a(\bar{w}+c)-1|}=c. \tag{3}$$

We prove that $c\approx\mathbb{E}[|V-\bar{w}|]$—namely, that it indeed captures the *concentration* of V around its first-order approximation.

With Exponential patience, the concentration is of the order of $1/\sqrt{\lambda}$ (see Example 1) so that it suffices to restrict our attention to the "standard" scaling; see Gurvich et al. (2014). The more elaborate structure brought about by the general patience distribution motivates (indeed, necessitates) an infrastructure that can accommodate multiple modes of scaling simultaneously.

The concentration may depend on all the primitives of the queue—the arrival rate and service and patience distributions—and it can vary within a queue family. Common to all members in the queue family is a parameter H that bounds the behavior *relative* to the concentration. The patience distribution and other primitives of the queue vary within the queue family, but the bounds apply universally to any primitives within it.

Analysis. Having set up the framework of queue families, our analysis follows the line of work by Gurvich et al. (2014), Gurvich (2014), and Braverman and Dai (2017). Braverman and Dai (2017) provide a road map based on the following three ingredients of (1) generator coupling, (2) gradient bounds for the Poisson equation, and (3) a priori moment bounds. This generator-comparison methodology, inspired by Stein's method, is conceptually related to the closure approximations proposed in Pender and Engblom (2014), which rely on the forward equations and Poisson–Charlier polynomials to establish asymptotics-free bounds for birth and death processes. Stein's method has also been used to approximate the invariant measures of diffusions using Malliavin calculus; see, for example, Kusuoka and Tudor (2012).

Although we do not seek to contribute here to Stein's method, our paper expands its scope. Our analysis of the $M/GI/1+GI$ queue is the first application of this method to nonexponential patience, which in turn, requires venturing beyond the standard $\sqrt{\lambda}$ diffusion scaling. Our paper also provides a first extension of these ideas to optimal dynamic control. Conceptually, Stein's method extends to control in an (almost) natural way in that the Poisson equation used for the performance analysis is replaced with the Hamilton-Jacobi-Bellman (HJB) equation. Some care is needed because (1) a priori, the optimal control might be history dependent, which means that martingale arguments have to be used instead of the direct generator coupling; and (2) in the absence of a given control, the a priori moment bounds in the performance analysis are replaced by bounds under "good controls."

Our main contribution, then, is not in the mechanics of establishing the error bounds. Rather, our purpose is to visit one of the most fundamental queues and build an infrastructure that allows us to establish, in an accessible way, the universality and accuracy of a (indeed, *the*) simple Brownian queue, thus circumventing the assumptions about heavy-traffic regimes.

We hope that our paper will have not only mathematical value but also modeling value in that we formally and rigorously expand the toolbox of the modeler who can now follow the natural heuristic without attributing regime interpretation to the parameters or being concerned with limits.

Heavy-traffic regimes are important. They provide an elegant and insightful way to map the underlying economic parameters to the capacity decisions. A flexible universal approximation does not obviate those insights but allows for direct (maybe simpler) analysis and optimization of queues.

Notation. We use the convention that $\mathbb{N}=\{0,1,2,\dots\}$. Following standard terminology, we denote by $|x|$ the absolute value of a real number x. For an l-times differentiable function $f:\mathbb{R}\to\mathbb{R}$, we write $f^{(l)}(\cdot)$ for its lth derivative. For three positive numbers a,b,c, we write $a=b\pm c$ to mean $a\in[b-c,b+c]$. For a Markov process $X(\cdot)=(X(t),t\geqslant 0)$ that has a stationary distribution, we denote by X (i.e., without a time index) a random variable with this distribution.

2. The Virtual Wait Dynamics

The $M/GI/1+GI$ queue has Poisson arrivals, general independent service times, and general independent patience thresholds.

The arrival process is denoted by $A(\cdot)=(A(t),t\geqslant 0)$. The random variable s_i stands for the service time of the ith customer and is drawn from the distribution F_s. The service rate is $\mu=1/\mathbb{E}[s_1]$ and $\rho=\lambda/\mu$ is the *traffic intensity*—that is, the amount of work that arrives per unit of time.

The queue follows the work-conserving first-come-first-served (FCFS) policy. An arriving customer's service commences immediately if the server is available. Otherwise, the customer is queued. Customer i's patience threshold is v_i. The customer abandons the queue if his service has not commenced by the time his patience expired. The patience values $\{v_i,i=1,\dots\}$ form a sequence of independent and identically distributed random variables drawn from the distribution F_a. We denote by f_a the density of this distribution and by $h_a=f_a/\bar{F}_a$ its hazard rate. If patience is infinite, we have $\bar{F}_a\equiv 1$. When the patience distribution has a

finite mean, we denote the mean by $\mathbb{E}[v_1]$. For stability, we assume that

$$\lambda \bar{F}_a(\infty) < \mu.$$

This is satisfied if $\rho = \lambda/\mu < 1$ and patience is infinite and for any λ, μ provided that all customers have finite patience (i.e., if $\bar{F}_a(\infty) = 0$). The arrival process, service times, and patience thresholds are mutually independent. We refer to $p = (\lambda, F_s, F_a)$ as *the primitives* of the $M/GI/1+GI$ queue.

System Dynamics. We study the *virtual waiting time* process $V(\cdot) = (V(t), t \geqslant 0)$. The quantity $V(t)$ is the effective workload at time t that includes only the work of customers who will not abandon the queue before being served. Thus, $V(t)$ signifies the amount of time that a virtual customer arriving at t would have to wait before entering service; that is, customer i, arriving at time τ_i, is "offered" the waiting time

$$\omega_i = V(\tau_i-).$$

Upon arrival of the ith customer, the process $V(\cdot)$ increases by this customer's service time s_i if $v_i > \omega_i = V(\tau_i-)$ (the customer is sufficiently patient). The process $V(\cdot)$ decreases at a rate of 1 whenever the server is working; $I(t)$ is the cumulative idleness of the server by time t so that $t - I(t)$ is the cumulative processing of the server by time t. The virtual wait/effective workload (including all work that has arrived and will stay for service minus the amount of work processed) at time t is then given by

$$V(t) = V(0) + \sum_{i=1}^{A(t)} s_i 1_{\{v_i > \omega_i\}} - (t - I(t)).$$

As a process, $V(t)$ satisfies the obvious properties,

$$V(t) \geqslant 0, \quad \forall\, t \geqslant 0, \quad I(\cdot) \text{ is nondecreasing with } I(0) = 0,$$
$$\text{and} \quad \int_0^\infty 1_{\{V(s)>0\}}\, dI(s) = 0.$$

The last of the above is the work conservation requirement: the idleness does not increase when there are customers in the system and hence there is a strictly positive virtual wait.

It should be intuitively clear that with FCFS service, $V(\cdot)$ is a Markov process. With $\lambda \bar{F}_a(\infty) < \mu$, the Markov process has a unique stationary distribution (see Section 3.1) and *we denote by V a random variable having the stationary distribution of $V(\cdot)$.*

The First-Order (a.k.a. Fluid) Stationary Approximation. The maximum long-run throughput rate is bounded by $\lambda \wedge \mu$ (the number of customers served cannot exceed the arrival rate or the service rate). The number of customers who get served (i.e., do not abandon) per unit of time when the waiting time is w is $\lambda \bar{F}_a(w)$, and the amount of work the customers bring is $\lambda \mathbb{E}[s_1] \bar{F}_a(w) = \rho \bar{F}_a(w)$. Heuristically, then, given the primitives p, the virtual waiting time should center at a point $\bar{w}_p$ where $\lambda \bar{F}_a(\bar{w}_p) = \mu \wedge \lambda$, or, equivalently, where

$$\rho \bar{F}_a(\bar{w}_p) = 1 \wedge \rho.$$

The Brownian Queue. With abandonments, not all customers are counted in the virtual wait. Only customers whose patience, v_i, is greater than the virtual wait at their moment of arrival are counted, and their total service requirement is $\sum_{i=1}^{A(t)} s_i 1_{\{v_i > V(\tau_i-)\}}$, where τ_i is the time of the ith arrival and $V(\tau_i-)$ is the wait "offered" to that customer. Because a customer's virtual waiting time is independent of his service time and patience threshold, we expect that

$$\mathbb{E}\left[\sum_{i=1}^{A(t)} s_i 1_{\{v_i > V(\tau_i-)\}}\right] = \lambda \mathbb{E}[s_i] \int_0^t \mathbb{P}\{v_i > V(s)\}\, ds$$
$$= \rho \int_0^t \bar{F}_a(V(s))\, ds,$$

where F_a is the patience distribution.

Given a standard Brownian motion $B(\cdot)$ and primitives p, let $(\hat{V}(\cdot), \hat{I}(\cdot))$ be the unique solution to the following stochastic differential equation (SDE):

$$\hat{V}(t) = \hat{V}(0) + \int_0^t \rho \bar{F}_a(\hat{V}(s))\, ds - t + \sigma B(t) + \hat{I}(t),$$
$$\hat{V}(\cdot) \geqslant 0,$$
$$\hat{I}(\cdot) \text{ is nondecreasing and starts at 0,}$$
$$\int_0^\infty 1_{\{\hat{V}(s)>0\}}\, d\hat{I}(s) = 0,$$

where $\sigma = \sqrt{\lambda \mathbb{E}[s_1^2] \bar{F}_a(\bar{w}_p)} = \sqrt{\lambda \mathbb{E}[s_1^2]((1/\rho) \wedge 1)}$. Because $\bar{F}_a \leqslant 1$, the existence and uniqueness of a strong solution $(\hat{V}(\cdot), \hat{I}(\cdot))$ follows from theorem 3.1 of Zhang (1994). The appeal of the diffusion model is the simplicity of its stationary distribution: $\hat{V}(\cdot)$ has a unique stationary distribution (which is also a steady-state distribution) if

$$G = \left(\int_0^\infty \exp\left(2 \int_0^x \frac{\rho \bar{F}_a(u) - 1}{\sigma^2}\, du\right) dx\right)^{-1} < \infty,$$

(which holds, in particular, if $\lambda \bar{F}_a(\infty) < \mu$), in which case its density is given by

$$\hat{\pi}(dx) = G \exp\left(2 \int_0^x \frac{\rho \bar{F}_a(u) - 1}{\sigma^2}\, du\right) dx, \quad x \in [0, \infty). \tag{4}$$

We denote by $\hat{V}$ a random variable following this distribution. We will prove that $\hat{V}$ provides a universally accurate approximation to V.

3. Performance Analysis

We first introduce a notion of universality that accommodates variation in the patience distribution and the traffic intensity ρ.

Definition 1 (Queue Families)**.** Fix $H > 1$. Denote by $\mathcal{Q}(H)$ the family of primitives $p = (\lambda, F_s, F_a)$ such that

(i) *exponential service-time moments*: $\mathbb{E}[\exp(\delta_H(s_1/\mathbb{E}[s_1]))] \leqslant H$, for some $\delta_H > 0$,

and there exists a constant $c_p \geqslant \mathbb{E}[s_1]/H$ such that, as a pair, (p, c_p) satisfy

(ii) *finite load*: $\rho \in [H^{-1}, H]$, $\rho \geqslant 1 - H/(\lambda c_p)$;

(iii) *subpolynomial patience density*:[3] F_a is differentiable with density f_a that satisfies

$$f_a(y) \leqslant \frac{H}{\lambda c_p^2}\left(1 + \left|\frac{y - \bar{w}_p}{c_p}\right|^H\right);$$

(iv) *inward drift:*

$$\rho \bar{F}_a(y) - 1 \leqslant -H^{-1}\frac{1}{\lambda c_p}, \quad \text{for all } y \in [\bar{w}_p + c_p H, \infty),$$

and

$$\rho \bar{F}_a(y) - 1 \geqslant H^{-1}\frac{1}{\lambda c_p}, \quad \text{for all } y \in [0, \bar{w}_p - c_p H],$$

where the second part is satisfied trivially if $\bar{w}_p < c_p H$.

We refer to the constant c_p as the *concentration* under the primitives p. Given $H > 1$, the set $\mathcal{Q}(H)$ includes the $M/M/1$ queue with $\rho \in [H^{-1}, 1)$ ($\bar{w}_p = 0$ and $c_p = 1/(\lambda(1-\rho))$) and is, hence, nonempty. The constants in Theorem 1 are uniform over $p \in \mathcal{Q}(H)$ and depend only on H and not on c_p, which is allowed to vary with p within $\mathcal{Q}(H)$. Moreover, while ρ is restricted to $[H^{-1}, H]$, the *arrival rate is not by itself restricted* and can grow without bounds *within* $\mathcal{Q}(H)$ as long as μ grows with it.

Example 1. Consider the $M/GI/1+M$ queue with patience rate $\theta = 1$. Letting $c_p = 1/\sqrt{\lambda}$, we have for $\eta \in \{-1, 1\}$,

$$\rho \bar{F}_a(\bar{w}_p + \eta/\sqrt{\lambda}) - 1 = \rho e^{-(\bar{w}_p + \eta/\sqrt{\lambda})} - 1.$$

From the definition of $\bar{w}_p$, we have $\rho \bar{F}_a(\bar{w}_p) = \rho e^{-\bar{w}_p} = \rho \wedge 1$ so that using $e^{-\eta/\sqrt{\lambda}} \approx 1 - \eta/\sqrt{\lambda}$,

$$\rho \bar{F}_a(\bar{w}_p + \eta/\sqrt{\lambda}) - 1 \approx (\rho \wedge 1) - 1 - \eta\frac{\rho \wedge 1}{\sqrt{\lambda}}.$$

Condition (iv) is then satisfied for any $H > 1$. If $\rho \leqslant 1$, $\bar{w}_p = 0$, and we only need the first part of (iv) to hold. Furthermore, for any $H > 1$, $H \geqslant f_a(0) = 1$, so condition (iii) is satisfied. □

Table 1 lists patience distributions together with the concentration c_p and the value of H for which $p \in \mathcal{Q}(H)$. Reading from H backwards, the table defines, given H, which primitives are included in $\mathcal{Q}(H)$. These do not apply to cases in which the mean patience is short relative to the mean service time; see Remark 3. Evidently, given $H > 1$, there are multiple instances of each of these distributions (and of the service time distribution and arrival rates λ) that fit within the family $\mathcal{Q}(H)$. When restricting attention to the Exponential distribution, this table shows that, for example, taking $H = 2$, any $M/G/1+M$ queue with a light-tailed service time distribution and $\rho \in [1/2, 2]$ is a member of $\mathcal{Q}(2)$. Thus, $\mathcal{Q}(2)$ covers simultaneously underloaded, critically loaded, and overloaded queues.

We prove that $\hat{V}$ provides an accurate approximation for V across multiple performance metrics and universally (i.e., for all queues in $\mathcal{Q}(H)$). Because a queue family covers a range of values for ρ, this strong notion of universality implies, in particular, universality in heavy-traffic regimes.

Theorem 1 (Virtual Waiting Time). *Given $H > 0$ and $k \in \mathbb{N}$, there exists a constant $C^1_{H,k} > 0$ such that*

$$\begin{aligned}&\mathbb{E}[(V - \bar{w}_p)^k] - \mathbb{E}[(\hat{V} - \bar{w}_p)^k]\\ &\quad = \pm C^1_{H,k}\mathbb{E}[s_1]\mathbb{E}[|\hat{V} - \bar{w}_p|^{k-1}]\\ &\quad = \pm C^1_{H,k}\frac{\rho}{\lambda}\mathbb{E}[|\hat{V} - \bar{w}_p|^{k-1}], \quad p \in \mathcal{Q}(H).\end{aligned}$$

Remark 1 (Time Units). The constant H is independent of the time unit (seconds, minutes, or hours) and, consequently, so are the constants $C^1_{H,k}$ in Theorem 1. Given $p \in \mathcal{Q}(H)$, changing the time units (using, say, minutes instead of seconds) leaves $p \in \mathcal{Q}(H)$ *with the same* H. Table 1 illustrates this insensitivity. The concentration c_p and the constant $\bar{w}_p$ do have to be changed (dividing by 60, for example, if we move from seconds to minutes).

The virtual waiting time does obviously depend on the time unit, but the theorem's statement can be made unit free by multiplying both sides by λ^k to get

$$\begin{aligned}&\mathbb{E}[\lambda^k(V - \bar{w}_p)^k] - \mathbb{E}[\lambda^k(\hat{V} - \bar{w}_p)^k]\\ &\quad = \pm C^1_{H,k}\rho\mathbb{E}[\lambda^{k-1}|\hat{V} - \bar{w}_p|^{k-1}], \quad p \in \mathcal{Q}(H). \quad \square\end{aligned}$$

Remark 2 (When Is the Concentration $1/\sqrt{\lambda}$?). It is standard in deriving heavy-traffic limits to consider the scaling $\sqrt{\lambda}(V - \bar{w}_p)$; see Ward and Glynn (2005), Jennings and Reed (2012). This scaling is not always suitable: Table 1 shows, for example, that with $\rho = 1$ and $F_a(x) = x^m$ for $x \in [0, 1]$ and $m > 1$, the concentration is $c_p = \lambda^{-1/(m+1)}$ so that $\sqrt{\lambda}(V - \bar{w}_p)$ would "explode" with λ. To place sequences of queues within our queue-families framework, Lemma EC.2 in the e-companion specifies sufficient conditions for $1/\sqrt{\lambda}$ concentration.

Table 1. Parameters c_p and H for a Family of Patience Distributions

F_a	c_p		H
Infinite $(M/GI/1)$	$c_p^\infty := 1/(\lambda(1-\rho))$, $\rho<1$		$1/\rho$
	$\rho \leqslant 1$: $c_p^\infty \wedge c_p^c$, $\rho>1$: $c_p^c \wedge c_p^o$		
	c_p^c	c_p^o	
$\exp(\theta)$	$\sqrt{\frac{\mathbb{E}[v_1]}{\lambda}}$	$\sqrt{\frac{\mathbb{E}[v_1]}{\lambda}}$	$\max(2,\rho,1/\rho)$
HyperExp$(\boldsymbol{\theta},\boldsymbol{\varphi})$[a]	$\sqrt{\frac{\mathbb{E}[v_1]}{\lambda}}$	$\sqrt{\frac{\mathbb{E}[v_1]}{\lambda}}$	$\max(\bar{\theta},2/\underline{\theta},\rho,1/\rho)$
Gamma(k,θ)[b]	$\left(\frac{\mathbb{E}[v_1]^k}{\lambda}\right)^{1/(k+1)}$	$\sqrt{\frac{\mathbb{E}[v_1]}{\lambda(\rho-1)^{(k-1)/k}}}$	$\max\left(\frac{2^{k+1}(\rho\vee k)k^k H_0^E\Gamma(k)}{\rho\wedge 1}, U\right)$
Uniform$[0,\alpha]$	$\sqrt{\frac{\alpha}{\lambda}}$	$\sqrt{\frac{\alpha}{\lambda}}$	$\max(\rho,1/\rho)$
$F(x)=\frac{(x\wedge\alpha)^k}{\alpha^k}$	$\left(\frac{\alpha^k}{\lambda}\right)^{1/(k+1)}$	$\sqrt{\frac{\alpha}{\lambda(\rho-1)^{(k-1)/k}}}$	$\frac{k+1}{k(1\wedge\rho)}\vee(2^k(k\vee\rho))$
Beta(α,β)[c]	$\frac{1}{\lambda^{1/(\alpha+1)}}$	$\frac{1}{\sqrt{\lambda(\rho-1)^{(\alpha-1)/\alpha}}}$	$\max\left(\frac{2^{\alpha+\beta}(\rho\vee\alpha)\Gamma(\alpha+\beta)}{(\rho\wedge 1)\Gamma(\alpha)\Gamma(\beta)\min(L,1)}, U\right)$

[a]Here, $\mathbb{E}[v_1]=\sum_{k=1}^K \varphi_k/\theta_k$ with $\sum_{k=1}^K \varphi_k = 1$. In the table, $\bar{\theta} := \max_k \theta_k\mathbb{E}[v_1]$ and $\underline{\theta} := \min_k \theta_k\mathbb{E}[v_1]$. These two constants do not depend on the time units.

[b]Here, $k>1$ is the shape parameter and θ is the rate parameter. We use the superscript 1 for the special case $\theta=k$ (i.e., with the density $f_a^1(x) := (k^k/\Gamma(k))x^{k-1}e^{-kx}$). Letting F_a^1 and h_a^1 be the corresponding distribution and hazard-rate functions, we define $U := \sup_{x\geqslant 0} f_a^1(x)(1/\bar{F}_a^1(1)-1)^{-(k-1)/k}$ and $H_0^E = \max(e^k, 1/L)$, where $L = \inf_{x\geqslant 1} h_a^1(x) > 0$.

[c]Here, $\alpha,\beta>1$ are the shape parameters; $f_a(x) = (\Gamma(\alpha+\beta)/(\Gamma(\alpha)\Gamma(\beta)))x^{\alpha-1}(1-x)^{\beta-1}$. Letting h_a be the hazard-rate function, we define $U := \sup_{x\geqslant 0} f_a(x)(1/\bar{F}_a(1/2)-1)^{-(\alpha-1)/\alpha} < \infty$, $L = \inf_{x\geqslant 1/2} h_a(x) > 0$.

Exponential, Uniform, and Hyperexponential patience distributions satisfy these conditions in alignment with Table 1.

The connection to heavy-traffic sequences is explored formally in Section EC.3 of the e-companion. □

Let W follow the stationary distribution of the waiting time, which is the minimum of the stationary virtual waiting time, V, and the patience threshold v. The diffusion analogue $\hat{W}$ is similarly represented as the minimum of $\hat{V}$ and v:

$$W = v \wedge V, \quad \hat{W} = v \wedge \hat{V},$$

where v is drawn from the patience distribution F_a and is independent of V (respectively of $\hat{V}$). The following proposition generalizes our introductory observation (2) about the $M/M/1$ queue.

Proposition 1 (Waiting Time). *Given H and $k\in\mathbb{N}$, there exists a constant $C_{H,k}^2>0$ such that*

$$\mathbb{E}[W^k] - \mathbb{E}[\hat{W}^k] = \pm C_{H,k}^2\mathbb{E}[s_1]\mathbb{E}[\hat{W}^{k-1}] = \pm\rho\frac{C_{H,k}^2}{\lambda}\mathbb{E}[\hat{W}^{k-1}], \quad p\in\mathcal{Q}(H).$$

For $k=1$, $\mathbb{E}[\hat{W}^{k-1}]=1$ so the gap is bounded by $C_{H,k}^2\mathbb{E}[s_1]$ and is of the order of a *single service time*. Next, let Q be a random variable following the stationary distribution of the queue length process. Little's law and Proposition 1 yield the following corollary.

Corollary 1 (Queue Length). *Given H, there exists a constant $C_{H,1}^2>0$ such that*

$$\mathbb{E}[Q] = \lambda\mathbb{E}[W] = \lambda\mathbb{E}[\hat{W}] \pm \rho C_{H,1}^2, \quad p\in\mathcal{Q}(H).$$

Remark 3 (The Case of Light Traffic or Short Patience). For light-traffic queues or those with very impatient customers the result embedded in (2) is noninformative.

Consider the case of $k=1$ (first moment). If $\rho<1$, the stationary virtual waiting time in the $M/GI/1+GI$ queue is bounded from above by the stationary workload in the corresponding (infinite patience) $M/GI/1$ queue. For the latter, by the Pollaczek–Khinchine formula,

$$\mathbb{E}[W] = \mathbb{E}[s_1]\frac{\rho}{1-\rho}\frac{1+\mathrm{CoV}^2(s_1)}{2},$$

where $\mathrm{CoV}^2(s_1)$ is the service time distribution's squared coefficient of variation. If ρ is small, the waiting time is of the order of $\mathbb{E}[s_1]$. Thus, both the expected waiting time and the approximation error in Proposition 1 are of the order of $\rho/\lambda = \mathbb{E}[s_1]$.

A similar conclusion applies to the case in which customers' patience is of the order of (or smaller than) the service time. Because a customer's waiting time is shorter than his patience $\mathbb{E}[W]\leqslant\mathbb{E}[v_1]$, if $\mathbb{E}[v_1]$ is of the order of the mean service time, then so is the expected waiting time. Again, both the expected waiting time

and the approximation error are of the order of $\mathbb{E}[s_1]$; see numerical examples in Section 3.3. □

Finally, let Ab be the long-run (and hence stationary) fraction of customers who abandon the queue.

Proposition 2 (Abandonment). *Given H, there exists a constant $C_H > 0$ such that*

$$Ab = \mathbb{E}[F_a(\hat{V})] \pm \frac{C_H}{\lambda^2 \mathbb{E}[|\hat{V} - \bar{w}_p|^2]} = \left(1 - \frac{1}{\rho}\right)^+ + \mathbb{E}[F_a(\hat{V}) - F_a(\bar{w}_p)] \pm \frac{C_H}{\lambda^2 \mathbb{E}[|\hat{V} - \bar{w}_p|^2]}, \quad p \in \mathcal{Q}(H).$$

The approximation has two terms: $((\lambda - \mu)/\lambda)^+ = (1 - (1/\rho))^+$ is the first-order (fluid) proxy for the fraction of abandoning customers and the second term is the second-order (Brownian) correction. In the case where $\mathbb{E}[|\hat{V} - \bar{w}_p|^2]$ is proportional to $1/\lambda$, the error reduces to C_H/λ for a redefined constant C_H. Focusing on a single regime rather than seeking universal bounds may lead to tighter bounds. Bassamboo and Randhawa (2010) show, for example, that for overloaded queues the fluid approximation already achieves an accuracy of $1/\lambda$.

Thus far, our approximation errors were stated in terms of *moments of the diffusion* rather than in terms of the concentration c_p. The next result shows that *they are one and the same*: namely, that c_p, satisfying conditions (ii)–(iv) in the definition of queue families, is (up to a multiplicative constant) the diffusion's concentration $\mathbb{E}[|\hat{V} - \bar{w}_p|]$.

Lemma 1 (Concentration Bounds). *Given $H > 0$ and $k \in \mathbb{N}$, there exist constants $C^V_{H,k}$, $c^V_{H,k}$, $C^W_{H,k}$, and $c^W_{H,k} > 0$ (depending only on H and k) such that*

(i) *For the $M/GI/1 + GI$ queue,*

$$\mathbb{E}[|V - \bar{w}_p|^k] \leqslant C^V_{H,k} c_p^k,$$
$$\mathbb{E}[W^k] \leqslant C^W_{H,k}(\bar{w}_p^k + c_p^k), \quad p \in \mathcal{Q}(H).$$

(ii) *For the Brownian queue,*

$$\mathbb{E}[|\hat{V} - \bar{w}_p|^k] \in [c^V_{H,k} c_p^k, C^V_{H,k} c_p^k], \quad p \in \mathcal{Q}(H),$$

and

$$\mathbb{E}[\hat{W}^k] \in [c^W_{H,k}(\bar{w}_p^k + c_p^k), C^W_{H,k}(\bar{w}_p^k + c_p^k)], \quad p \in \mathcal{Q}(H).$$

A lower bound on $\mathbb{E}[|V - \bar{w}|^k]$ can also be established; see Remark 4. The following is now a corollary of Theorem 1, Proposition 1, Corollary 1, and Proposition 2.

Corollary 2. *Given $H > 0$ and $k \in \mathbb{N}$, there exist $C_{H,k} > 0$ (in particular, $C_{H,1} > 0$) and $C^{Ab}_H > 0$ such that, for all $p \in \mathcal{Q}(H)$,*

$$\mathbb{E}[(V - \bar{w}_p)^k] - \mathbb{E}[(\hat{V} - \bar{w}_p)^k] = \pm \frac{C_{H,k}}{\lambda} c_p^{k-1},$$
$$\mathbb{E}[W^k] - \mathbb{E}[\hat{W}^k] = \pm \frac{C_{H,k}}{\lambda}(\bar{w}_p^{k-1} + c_p^{k-1}),$$
$$\mathbb{E}[Q] = \lambda \mathbb{E}[\hat{W}] \pm C_{H,1}, \quad Ab = \mathbb{E}[F_a(\hat{V})] \pm \frac{C^{Ab}_H}{\lambda^2 c_p^2}.$$

Example 2 (The $M/GI/1$ Queue Revisited). The $M/GI/1$ queue (no abandonments) is stable if and only if $\rho < 1$, in which case $\bar{w}_p = 0$. Condition (ii) reduces to $\rho - 1 \geqslant -H/(\lambda c_p)$, and condition (iv) to $\rho - 1 \leqslant -H^{-1}/(\lambda c_p)$. Thus, with $H \geqslant 1/\rho$, $p = (\lambda, F_s, F_a) \in \mathcal{Q}(H)$ with $c_p = 1/(\lambda(1-\rho))$. By Theorem 1 and Corollary 2, there exist constants $C_{k,1}, C_{k,2}, C_{k,3}$ such that

$$\mathbb{E}[W^k] - \mathbb{E}[\hat{W}^k] = \pm \frac{C_{k,1}}{\lambda} \mathbb{E}[\hat{W}^{k-1}] = \pm \frac{C_{k,2}}{\lambda^k (1-\rho)^{k-1}} = \pm C_{k,3}(1-\rho)\mathbb{E}[\hat{W}^k]. \quad \square$$

3.1. What Makes This Work

Lemma 2. *The process $V(\cdot)$ is an ergodic strong Markov process. For a differentiable function f such that $\mathbb{E}[|f(x + s_1) - f(x)|] < \infty$ for all $x \geqslant 0$, let*

$$\mathcal{A}f(x) = -f^{(1)}(x) + \lambda \bar{F}_a(x) \mathbb{E}[f(x + s_1) - f(x)]$$

($\mathcal{A}$ is the generator of $V(\cdot)$). Let Ψ be such a continuously differentiable function with $\Psi^{(1)}(0) = 0$: (i) If $\Psi \geqslant 0$ and $\sup_{x \geqslant 0} \mathcal{A}\Psi(x) < \infty$, then $\mathbb{E}[\mathcal{A}\Psi(V)] \geqslant 0$, where V has the stationary distribution of $V(\cdot)$. (ii) If, instead, $\mathbb{E}[|\Psi(V)|] < \infty$ and $\mathbb{E}[|\Psi(V + s_1) - \Psi(V)|^2] < \infty$ (where s_1 is $F_s(\cdot)$-distributed and is independent of V), then $\mathbb{E}[\mathcal{A}\Psi(V)] = 0$.

Taking a differentiable function Ψ and using Taylor's expansion heuristically to replace

$$\mathbb{E}[\Psi(x + s_1)] \approx \Psi(x) + \Psi^{(1)}(x)\mathbb{E}[s_1] + \tfrac{1}{2}\Psi^{(2)}(x)\mathbb{E}[s_1^2],$$

we have, after some manipulations (recall that $\mu = 1/\mathbb{E}[s_1]$), that

$$\begin{aligned}\mathcal{A}\Psi(x) &\approx (\rho \bar{F}_a(x) - 1)\Psi^{(1)}(x) + \tfrac{1}{2}\lambda \mathbb{E}[s_1^2]\Psi^{(2)}(x)\bar{F}_a(x) \\ &= (\rho \bar{F}_a(x) - 1)\Psi^{(1)}(x) + \tfrac{1}{2}\lambda \mathbb{E}[s_1^2]\Psi^{(2)}(x)\bar{F}_a(\bar{w}_p) \\ &\quad + \tfrac{1}{2}\lambda \mathbb{E}[s_1^2]\Psi^{(2)}(x)(\bar{F}_a(x) - \bar{F}_a(\bar{w}_p)) \\ &= \hat{\mathcal{A}}\Psi(x) + \tfrac{1}{2}\lambda \mathbb{E}[s_1^2]\Psi^{(2)}(x)(\bar{F}_a(x) - \bar{F}_a(\bar{w}_p)), \quad (5)\end{aligned}$$

where $\hat{\mathcal{A}}$ is the generator of the diffusion; that is, $\hat{\mathcal{A}}\Psi(x) = (\rho \bar{F}_a(x) - 1)\Psi^{(1)}(x) + ((\lambda \mathbb{E}[s_1^2]\bar{F}_a(\bar{w}_p))/2)\Psi^{(2)}(x)$. Take

$$f_k(x) = \bar{f}_k\left(\frac{x - \bar{w}_p}{\mathbb{E}[|\hat{V} - \bar{w}_p|]}\right) := \frac{(x - \bar{w}_p)^k - \mathbb{E}[(\hat{V} - \bar{w}_p)^k]}{(\mathbb{E}[|\hat{V} - \bar{w}_p|])^k}, \quad (6)$$

(notice that $\mathbb{E}[f_k(\hat{V})]=0$) and suppose that Ψ is a solution to $\hat{\mathcal{A}}\Psi=-f_k$. This is the so-called Poisson equation for the diffusion. It is this "self-scaling" definition of the performance function f_k—normalized by the expectation $\mathbb{E}[|\hat{V}-\bar{w}_p|]$—that allows for the *universality in concentration*.

Taking expectations on both sides of (5) with respect to the stationary distribution of V, we have

$$\begin{aligned}&\mathbb{E}[\mathcal{A}\Psi(V)]\\&\quad\approx\mathbb{E}[\hat{\mathcal{A}}\Psi(V)]+\tfrac{1}{2}\lambda\mathbb{E}[s_1^2]\mathbb{E}[\Psi^{(2)}(V)(\bar{F}_a(V)-\bar{F}_a(\bar{w}_p))]\\&\quad=-\mathbb{E}[f_k(V)]+\tfrac{1}{2}\lambda\mathbb{E}[s_1^2]\mathbb{E}[\Psi^{(2)}(V)(\bar{F}_a(V)-\bar{F}_a(\bar{w}_p))],\end{aligned}$$

which, if $\mathbb{E}[\mathcal{A}\Psi(V)]=0$, gives

$$\mathbb{E}[f_k(V)]\approx\tfrac{1}{2}\lambda\mathbb{E}[s_1^2]\mathbb{E}[\Psi^{(2)}(V)(\bar{F}_a(V)-\bar{F}_a(\bar{w}_p))].$$

Proving that the term on the right is small would imply that $\mathbb{E}[f_k(V)]\approx 0$ and, because $f_k(x)=((x-\bar{w}_p)^k-\mathbb{E}[(\hat{V}-\bar{w}_p)^k])/(\mathbb{E}[|\hat{V}-\bar{w}_p|])^k$, that $\hat{V}$ approximates V in the sense of

$$\mathbb{E}[(V-\bar{w}_p)^k]\approx\mathbb{E}[(\hat{V}-\bar{w}_p)^k]+o((\mathbb{E}[|\hat{V}-\bar{w}_p|])^k).$$

It is clear that this argument, building on a second-order Taylor expansion, requires suitable bounds for the second and third derivatives of Ψ. The proof of Theorem 1 that follows formalizes this heuristic argument.

3.2. Proof of Theorem 1 and Propositions 1 and 2

Proof of Theorem 1. Fix $k\in\mathbb{N}$ and let f_k and $\bar{f}_k$ be as in (6). Note that with $C_0:=\max(\max(C^V_{H,k},1)/(c^V_{H,1})^k,k)$, for all x,

$$\begin{aligned}|\bar{f}_k(x)|&\leqslant C_0(1+|x|^k)\quad\text{and}\\|\bar{f}_k^{(1)}(x)|&\leqslant C_0(1+|x|^{2H+k+2}).\end{aligned}\tag{7}$$

Lemma 3. *There is a unique solution (up to an additive constant) to the initial value problem*:

$$\hat{\mathcal{A}}\Psi(x)=-f_k(x),\quad \Psi^{(1)}(0)=0,\quad x\geqslant 0.$$

Denote the unique solution by Ψ_k. For any $H>0$, there is a positive constant C_Ψ (which also depends on H and k), such that for $p\in\mathcal{Q}(H)$,

$$|\Psi_k^{(2)}(x)|\leqslant C_0C_\Psi\times\lambda\times\left(1+\left|\frac{x-\bar{w}_p}{\mathbb{E}[|\hat{V}-\bar{w}_p|]}\right|^{H+k+1}\right),$$

$$|\Psi_k^{(3)}(x)|\leqslant C_0C_\Psi\times\frac{\lambda}{\mathbb{E}[|\hat{V}-\bar{w}_p|]}\times\left(1+\left|\frac{x-\bar{w}_p}{\mathbb{E}[|\hat{V}-\bar{w}_p|]}\right|^{2H+k+2}\right).$$

The proof of Lemma 3 is presented in the e-companion. For notational simplicity, we omit the subscript k in Ψ_k and denote by Ψ a solution to $\hat{\mathcal{A}}\Psi=-f_k$ with $\Psi^{(1)}(0)=0$. Using Taylor's expansion,

$$\begin{aligned}\Psi(x+s_1)=\Psi(x)+\Psi^{(1)}(x)s_1+\tfrac{1}{2}\Psi^{(2)}(x)s_1^2\\+\tfrac{1}{6}\Psi^{(3)}(x+\Delta_x(s_1))s_1^3,\end{aligned}$$

where $\Delta_x(s_1)$ is some number between 0 and s_1 that can depend on x. Simple manipulations give

$$\begin{aligned}\mathcal{A}\Psi(x)&=(\rho\bar{F}_a(x)-1)\Psi^{(1)}(x)+\frac{\lambda\mathbb{E}[s_1^2]}{2}\Psi^{(2)}(x)\bar{F}_a(x)\\&\quad+\frac{1}{6}\lambda\mathbb{E}[s_1^3\Psi^{(3)}(x+\Delta_x(s_1))]\bar{F}_a(x)\\&=(\rho\bar{F}_a(x)-1)\Psi^{(1)}(x)+\frac{\lambda\mathbb{E}[s_1^2]}{2}\Psi^{(2)}(x)\bar{F}_a(\bar{w}_p)\\&\quad+\frac{\lambda\mathbb{E}[s_1^2]}{2}\Psi^{(2)}(x)(\bar{F}_a(x)-\bar{F}_a(\bar{w}_p))\\&\quad+\frac{1}{6}\lambda\mathbb{E}[s_1^3\Psi^{(3)}(x+\Delta_x(s_1))]\bar{F}_a(x)\\&=\hat{\mathcal{A}}\Psi(x)+\frac{\lambda\mathbb{E}[s_1^2]}{2}\Psi^{(2)}(x)(\bar{F}_a(x)-\bar{F}_a(\bar{w}_p))\\&\quad+\frac{1}{6}\lambda\mathbb{E}[s_1^3\Psi^{(3)}(x+\Delta_x(s_1))]\bar{F}_a(x).\end{aligned}$$

In turn,

$$\begin{aligned}|\mathcal{A}\Psi(x)-\hat{\mathcal{A}}\Psi(x)|&\leqslant\frac{\lambda\mathbb{E}[s_1^2]}{2}\times|\Psi^{(2)}(x)|\times|\bar{F}_a(x)-\bar{F}_a(\bar{w}_p)|\\&\quad+\frac{\lambda}{6}\mathbb{E}[s_1^3|\Psi^{(3)}(x+\Delta_x(s_1))|]\bar{F}_a(x).\end{aligned}\tag{8}$$

Next, we plug in the derivative bounds from Lemma 3 and the properties of the queue family $\mathcal{Q}(H)$. Notice that

$$\begin{aligned}|\Psi^{(3)}(x+z)|&\leqslant C_0C_\Psi\times\frac{\lambda}{\mathbb{E}[|\hat{V}-\bar{w}_p|]}\\&\quad\times\left(1+\left|\frac{x+z-\bar{w}_p}{\mathbb{E}[|\hat{V}-\bar{w}_p|]}\right|^{2H+k+2}\right)\leqslant 2^{2H+k+2}\times C_0C_\Psi\\&\quad\times\frac{\lambda}{\mathbb{E}[|\hat{V}-\bar{w}_p|]}\times\left(1+\left|\frac{x-\bar{w}_p}{\mathbb{E}[|\hat{V}-\bar{w}_p|]}\right|^{2H+k+2}\right.\\&\qquad\left.+\left|\frac{z}{\mathbb{E}[|\hat{V}-\bar{w}_p|]}\right|^{2H+k+2}\right).\end{aligned}$$

The last inequality follows from the fact $|x+y|^n\leqslant(|x|+|y|)^n\leqslant 2^n(|x|^n+|y|^n)$ for all $x,y\in\mathbb{R}$ and $n=1,2,\ldots$.

Furthermore, $\bar{F}_a(\cdot)\leqslant 1$, and per the definition of $\mathcal{Q}(H)$ (and Lemma 1),

$$\begin{aligned}|\bar{F}_a(x)-\bar{F}_a(\bar{w}_p)|&\leqslant\frac{\bar{C}}{\lambda\mathbb{E}[|\hat{V}-\bar{w}_p|]}\\&\quad\times\left(\left|\frac{x-\bar{w}_p}{\mathbb{E}[|\hat{V}-\bar{w}_p|]}\right|+\left|\frac{x-\bar{w}_p}{\mathbb{E}[|\hat{V}-\bar{w}_p|]}\right|^{H+1}\right)\end{aligned}$$

for some constant $\bar{C}$. Thus,

$$
\begin{aligned}
&|\mathcal{A}\Psi(x) - \hat{\mathcal{A}}\Psi(x)| \\
&\leqslant \frac{\rho^2\mathbb{E}[(\mu s_1)^2]}{2} \times \frac{C_0 \times C_\Psi \times \bar{C}}{\lambda\mathbb{E}[|\hat{V} - \bar{w}_p|]}\left(1 + \left|\frac{x - \bar{w}_p}{\mathbb{E}[|\hat{V} - \bar{w}_p|]}\right|^{H+k+1}\right) \\
&\quad \times \left(\left|\frac{x - \bar{w}_p}{\mathbb{E}[|\hat{V} - \bar{w}_p|]}\right| + \left|\frac{x - \bar{w}_p}{\mathbb{E}[|\hat{V} - \bar{w}_p|]}\right|^{H+1}\right) + \frac{\rho^3 2^{2H+k+2} C_0 C_\Psi}{6\lambda\mathbb{E}[|\hat{V} - \bar{w}_p|]} \\
&\quad \times \left(\mathbb{E}[(\mu s_1)^3]\left(1 + \left|\frac{x - \bar{w}_p}{\mathbb{E}[|\hat{V} - \bar{w}_p|]}\right|^{2H+k+2}\right)\right. \\
&\qquad \left. + \frac{\mathbb{E}[(\mu s_1)^{2H+k+5}]}{(\mu\mathbb{E}[|\hat{V} - \bar{w}_p|])^{2H+k+2}}\right). \qquad (9)
\end{aligned}
$$

Recall that for $p \in \mathcal{Q}(H)$, $\rho \leqslant H$, $1/(\mu c_p) \leqslant H$ and that, by the exponential moment requirement, $\mathbb{E}[(\mu s_1)^l] \leqslant B_{H,l}$ for some constant $B_{H,l}$. Because Ψ satisfies the inequalities in Lemma 3, it is bounded by a polynomial function. By Lemma 1, the assumptions of Lemma 2(ii) then hold, and hence $\mathbb{E}[\mathcal{A}\Psi(V)] = 0$. This gives us

$$
\begin{aligned}
|\mathbb{E}[f_k(V)]| &= |\mathbb{E}[\hat{\mathcal{A}}\Psi(V)]| = |\mathbb{E}[\hat{\mathcal{A}}\Psi(V)] - \mathbb{E}[\mathcal{A}\Psi(V)]| \\
&\leqslant \mathbb{E}[|\hat{\mathcal{A}}\Psi(V) - \mathcal{A}\Psi(V)|] = \frac{1}{\lambda\mathbb{E}[|\hat{V} - \bar{w}_p|]} \\
&\quad \times C_0\bar{C}_{\Psi,k} \times \mathbb{E}\left[\left(1 + \left|\frac{V - \bar{w}_p}{\mathbb{E}[|\hat{V} - \bar{w}_p|]}\right|^{2H+k+2}\right)\right] \\
&\leqslant \frac{1}{\lambda\mathbb{E}[|\hat{V} - \bar{w}_p|]}\bar{C}_{H,k} \qquad (10)
\end{aligned}
$$

for some constants $\bar{C}_{\Psi,k}$ and $\bar{C}_{H,k}$, where the last inequality follows from Lemma 1, which provides a lower bound on $\mathbb{E}[|\hat{V} - \bar{w}_p|]$ and an upper bound on $\mathbb{E}[|V - \bar{w}_p|^{2H+k+2}]$ as a function of c_p. Plugging this bound back in the definition of f_k, we have

$$
\begin{aligned}
&|\mathbb{E}[(V - \bar{w}_p)^k] - \mathbb{E}[(\hat{V} - \bar{w}_p)^k]| \\
&\qquad \leqslant \bar{C}_{H,k}\frac{(\mathbb{E}[|\hat{V} - \bar{w}_p|])^{k-1}}{\lambda} \leqslant C^1_{H,k}\frac{\mathbb{E}[|\hat{V} - \bar{w}_p|^{k-1}]}{\lambda},
\end{aligned}
$$

for an appropriate constant $C^1_{H,k}$, where the last inequality follows from the upper bound on $\mathbb{E}[|\hat{V} - \bar{w}_p|]$ and the lower bound on $\mathbb{E}[|\hat{V} - \bar{w}_p|^{k-1}]$ in Lemma 1. This concludes the proof. □

Before proceeding, note that the only properties of f_k that we use in the proof of Equation (10) are that $\mathbb{E}[f_k(\hat{V})] = 0$, that f_k is differentiable, and that it is sub-polynomial in the sense of (7). The same bounds thus apply to any function g that has these properties.

Remark 4. For any $k \geqslant 2$, the function

$$
f_k(x) = \frac{|x - \bar{w}_p|^k - \mathbb{E}[|\hat{V} - \bar{w}_p|^k]}{(\mathbb{E}[|\hat{V} - \bar{w}_p|])^k}
$$

satisfies the required conditions of f_k. As a consequence, Theorem 1 can be expanded to include the absolute-value statement

$$
\begin{aligned}
&\mathbb{E}[|V - \bar{w}_p|^k] - \mathbb{E}[|\hat{V} - \bar{w}_p|^k] \\
&\quad = \pm C^1_{H,k}\frac{\rho}{\lambda}\mathbb{E}[|\hat{V} - \bar{w}_p|^{k-1}], \quad k \geqslant 2,\ p \in \mathcal{Q}(H).
\end{aligned}
$$

By Lemma 1 $\mathbb{E}[|\hat{V} - \bar{w}_p|^k] \in [c^V_{H,k}c^k_p, C^V_{H,k}c^k_p]$, and we then have

$$
\mathbb{E}[|V - \bar{w}_p|^k] \geqslant c^V_{H,k}c^k_p - \frac{C_{H,k}}{\lambda}c^{k-1}_p,
$$

for a suitable constant $C_{H,k}$. Considering $p \in \mathcal{Q}(H)$ with $\lambda c_p > (1+\epsilon)C_{H,k}/c^V_{H,k}$ for some $\epsilon > 0$, it follows that $\mathbb{E}[|V - \bar{w}_p|^k] \geqslant (\epsilon/(1+\epsilon))c^V_{H,k}c^k_p$. The existence of such ϵ is guaranteed if λc_p grows with λ, as when $c_p = \lambda^{-\delta}$ for any $\delta \in (0,1)$.

A lower bound for $k = 1$ can also be derived but requires special treatment to circumvent the nondifferentiability of the absolute value function. □

Proof of Proposition 1. Define

$$
\begin{aligned}
g_k(x) &:= \frac{1}{\mathbb{E}[|\hat{V} - \bar{w}_p|] \times \mathbb{E}[\hat{W}^{k-1}]}(\mathbb{E}[(v \wedge x)^k] - \mathbb{E}[\hat{W}^k]) \\
&= \frac{1}{\mathbb{E}[|\hat{V} - \bar{w}_p|] \times \mathbb{E}[\hat{W}^{k-1}]}\left(\int_0^x k u^{k-1}\bar{F}_a(u)\,du - \mathbb{E}[\hat{W}^k]\right),
\end{aligned}
$$

and $\bar{g}_k(x) = g_k(\bar{w}_p + x\mathbb{E}[|\hat{V} - \bar{w}_p|])$. Because $\hat{W} = v \wedge \hat{V}$, we have $\mathbb{E}[g_k(\hat{V})] = 0$. Let $\ell_k(x) = \bar{g}_k(x) - \bar{g}_k(0)$. Then $\bar{g}_k(x) = \ell_k(x) - \mathbb{E}[\ell_k((\hat{V} - \bar{w}_p)/\mathbb{E}[|\hat{V} - \bar{w}_p|])]$. Moreover, $\bar{g}_k$ and ℓ_k are differentiable with

$$
\begin{aligned}
|\bar{g}^{(1)}_k(x)| = |\ell^{(1)}_k(x)| &= \frac{k\mathbb{E}[|\hat{V} - \bar{w}_p|]}{\mathbb{E}[|\hat{V} - \bar{w}_p|] \times \mathbb{E}[\hat{W}^{k-1}]} \\
&\quad \times \bar{F}_a(\bar{w}_p + x\mathbb{E}[|\hat{V} - \bar{w}_p|])|(\bar{w}_p + x\mathbb{E}[|\hat{V} - \bar{w}_p|])^{k-1}| \\
&\leqslant C_k\frac{k\mathbb{E}[|\hat{V} - \bar{w}_p|]}{\mathbb{E}[|\hat{V} - \bar{w}_p|] \times (\bar{w}_p^{k-1} + (\mathbb{E}[|\hat{V} - \bar{w}_p|])^{k-1})} \\
&\quad \times (\bar{w}_p + |x|\mathbb{E}[|\hat{V} - \bar{w}_p|])^{k-1} \\
&\leqslant kC_k(1 + |x|)^{k-1} \leqslant 2^{k-1}kC_k(1 + |x|^{k-1}),
\end{aligned}
$$

where C_k is chosen based on the inequalities in Lemma 1 and depends only on k and H.

From $|\ell_k(x)| \leqslant 2^{k-1}kC_k(1 + |x|^{k-1})|x| \leqslant 2^{k-1}kC_k \cdot (|x| + |x|^k)$ and Lemma 1, we get a constant C_0 such that

$$
|\bar{g}_k(x)| \leqslant C_0(1 + |x|^k) \quad \text{and} \quad |\bar{g}^{(1)}_k(x)| \leqslant C_0(1 + |x|^{2H+k+2}).
$$

The function g_k thus satisfies the conditions of Lemma 3, and following the steps leading to (10), we get $|\mathbb{E}[g_k(V)]| \leqslant C^2_{H,k}/(\lambda\mathbb{E}[|\hat{V} - \bar{w}_p|])$ for an appropriately chosen constant $C^2_{H,k}$. We thus have

$$
\begin{aligned}
\mathbb{E}[W^k] - \mathbb{E}[\hat{W}^k] &= \mathbb{E}[g_k(V)]\mathbb{E}[|\hat{V} - \bar{w}_p|]\mathbb{E}[\hat{W}^{k-1}] \\
&= \pm\frac{C^2_{H,k}}{\lambda}\mathbb{E}[\hat{W}^{k-1}]. \quad \square
\end{aligned}
$$

Proof of Proposition 2. Recall that the process $(V(t), t \geqslant 0)$ is an ergodic strong Markov process. We denote by π_V its stationary distribution and write $\mathbb{E}_{\pi_V}[\cdot]$ for the expectation operation when $V(0)$ is drawn from π_V. Note that

$$\begin{aligned} Ab &= \lim_{t\to\infty} \frac{\mathbb{E}_{\pi_V}[\sum_{i=1}^{A(t)} 1_{\{v_i \leqslant \omega_i\}}]}{\mathbb{E}_{\pi_V}[A(t)]} = \lim_{t\to\infty} \frac{\lambda t \times \mathbb{E}[F_a(V)]}{\lambda t} \\ &= \mathbb{E}[F_a(V)] \\ &= F_a(\bar{w}_p) + \mathbb{E}[F_a(V) - F_a(\bar{w}_p)] \qquad (11) \\ &= \left(1 - \frac{1}{\rho}\right)^+ + \mathbb{E}[F_a(V) - F_a(\bar{w}_p)] \\ &= \left(1 - \frac{1}{\rho}\right)^+ + \mathbb{E}[F_a(\hat{V}) - F_a(\bar{w}_p)] \pm C_H(\lambda\mathbb{E}[|\hat{V} - \bar{w}_p|])^{-2}. \end{aligned}$$

The second equality uses the following lemma.

Lemma 4. *Suppose that $V(0)$ is drawn from π_V. Then we have the identity*:

$$\mathbb{E}_{\pi_V}\left[\sum_{i=1}^{A(t)} 1_{\{v_i \leqslant \omega_i\}}\right] = \lambda t \times \mathbb{E}[F_a(V)].$$

The last equality in (11) uses $\lambda F_a(\bar{w}_p) = \lambda - \lambda \bar{F}_a(\bar{w}_p) = \lambda - \lambda \wedge \mu$ and the following argument. Let $g(x) = \bar{g}((x - \bar{w}_p)/\mathbb{E}[|\hat{V} - \bar{w}_p|])$, where $\bar{g}(x) := \ell(x) - \mathbb{E}[\ell((\hat{V} - \bar{w}_p)/\mathbb{E}[|\hat{V} - \bar{w}_p|])]$ with $\ell(x) = \lambda\mathbb{E}[|\hat{V} - \bar{w}_p|](F_a(\bar{w}_p + x\mathbb{E}[|\hat{V} - \bar{w}_p|]) - F_a(\bar{w}_p))$. Then, by the definition of $\mathcal{C}(H)$ and Lemma 1, there exists a constant $\bar{C}_0$,

$$\begin{aligned} |\bar{g}^{(1)}(x)| = |\ell^{(1)}(x)| &= |\lambda(\mathbb{E}[|\hat{V} - \bar{w}_p|])^2 f_a(\bar{w}_p + x\mathbb{E}[|\hat{V} - \bar{w}_p|])| \\ &\leqslant \bar{C}_0(1 + |x|^H), \end{aligned}$$

and from $|\ell(x)| \leqslant \bar{C}_0(1 + |x|^H)|x| = \bar{C}_0(|x| + |x|^{H+1})$ and Lemma 1, we get a constant C_0 such that

$$|\bar{g}(x)| \leqslant C_0(1 + |x|^{H+1}), \quad \text{and} \quad |\bar{g}^{(1)}(x)| \leqslant C_0(1 + |x|^{3H+3}).$$

Hence, g satisfies the condition of Lemma 3 with $k = H + 1$, so that, following the steps leading to (10), we have $\mathbb{E}[g(V)] = \pm C_H/(\lambda\mathbb{E}[|\hat{V} - \bar{w}_p|])$, and we conclude that

$$\begin{aligned} \mathbb{E}[F_a(V) - F_a(\hat{V})] &= \frac{\mathbb{E}[g(V)]}{\lambda\mathbb{E}[|\hat{V} - \bar{w}_p|]} \\ &= \pm\frac{C_H}{(\lambda\mathbb{E}[|\hat{V} - \bar{w}_p|])^2}. \quad \square \end{aligned}$$

3.3. Numerical Examples

The $M/M/1 + GI$ queue (i.e., with exponential service time) is useful for numerical comparisons. Closed-form expressions for the stationary virtual waiting are presented in Zeltyn and Mandelbaum (2005), allowing us to circumvent the complexities of steady-state simulation.

The first set of figures comprises numerical manifestations of the predictions in Theorem 1 and Proposition 1. Figure 2 reports the results for two power-law patience distributions. We plot the moments of the stationary virtual waiting time and the gap scaled by $1/\lambda$. The distribution $F_a(x) = x$ has a positive density at 0 while $F_a(x) = x^2$ does not. In the context of asymptotic convergence, the process limits require different treatment. The case $F_a(x) = x^2$ requires hazard-rate scaling; otherwise, the patience distribution disappears in the limit; see further discussion in Section EC.3. As predicted in Theorem 1, the scaled gap is bounded by a constant.

In addition to further showcasing the precision of the approximation, Figure 3 highlights a point that the mathematical results do not capture. On the left-hand side of the figure, we plot the second moment of the waiting time $\mathbb{E}[W^2]$, its approximation $\mathbb{E}[\hat{W}^2]$, and the approximation gap relative to the first moment approximation $\mathbb{E}[\hat{W}]$ divided by λ. Proposition 1 states

Figure 2. $M/M/1 + GI$ with $\rho = 1$ and $F_a(x) = x$ on $[0,1]$ (Left) and $F_a(x) = x^2$ on $[0,1]$ (Right)

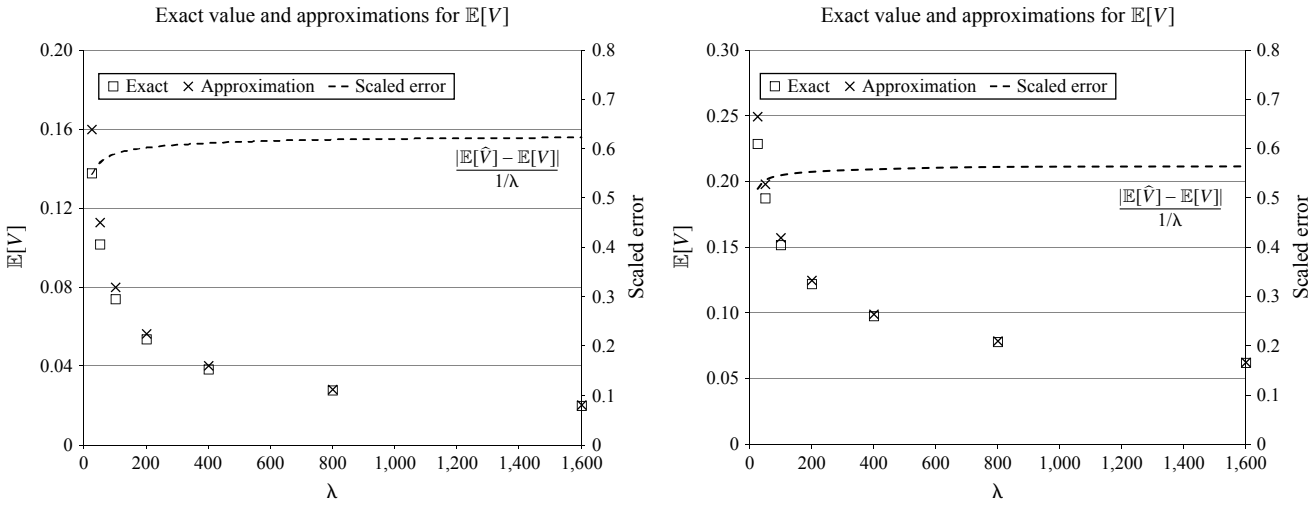
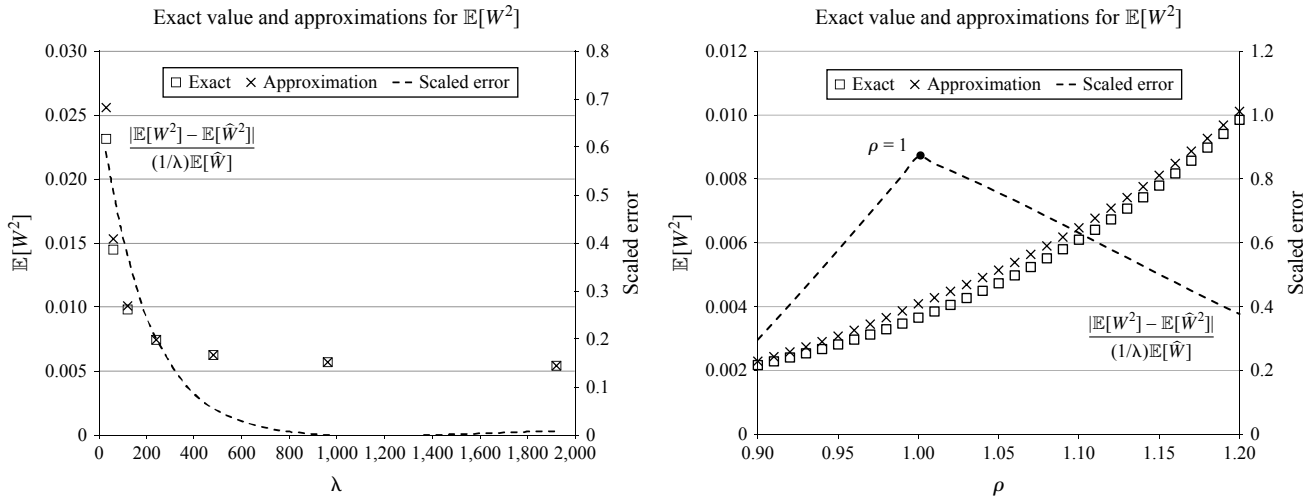


Figure 3. $M/M/1+GI$ with Patience $F_a(x) = \frac{4}{7}(1-e^{-4x}) + \frac{3}{7}(1-e^{-x/2})$ (Hyperexponential): $\rho = 1.2$ and λ (Hence Also μ) Varied (Left) and $\mu = 100$ Fixed and ρ (Hence Also λ) Varied (Right)



that this scaled error is bounded. In these numerical instances, it not only is bounded but decreases as λ grows large. The right-hand side shows that the scaled error is the greatest in the critically loaded case of $\rho \approx 1$ and decreases as the queue becomes overloaded. Thus, when focusing only on overloaded queues, it may be possible to obtain tighter bounds (as in Bassamboo and Randhawa 2010).

We repeat this exercise for a different patience distribution in Figure 4. This is an instance where the concentration c_p changes with ρ. It is of the order of $\lambda^{-1/3}$ for $\rho = 1$ but of the order of $\lambda^{-1/2}$ for $\rho = 1.2$ so that, in particular, process limit theorems for $\rho = 1$ would require different scaling than those for $\rho > 1$. The universal approximation is indifferent to this fact, as are the bounds.

Figures 5 and 6 consider settings that violate our assumptions to explore the necessity (or lack thereof) of our sufficient conditions. In Figure 5, we consider the patience distribution $F_a = \text{Gamma}(0.5, 2)$, which violates our requirement (iii) in the definition of queue families. The precision in Figure 5 is nevertheless impressive. The performance of the $M/D/1+GI$ queue is computed via simulation (the 95% confidence intervals are smaller than 0.0008).

In Figure 6, we revisit our requirement that the service times have light tails (the patience distribution is the one used in Figure 3). Our proofs are based on a second-order Taylor expansion and the bounds, consequently, depend on the third moment of the service time being finite. In fact, we also use higher moments. An exponentially decaying tail is not necessary for

Figure 4. $M/M/1+GI$ with Patience $F_a(x) = 1 - e^{-2x} - 2xe^{-2x}$ (Erlang): $\rho = 1.2$ and λ (Hence Also μ) Varied (Left) and $\mu = 100$ Fixed and ρ (Hence Also λ) Varied (Right)

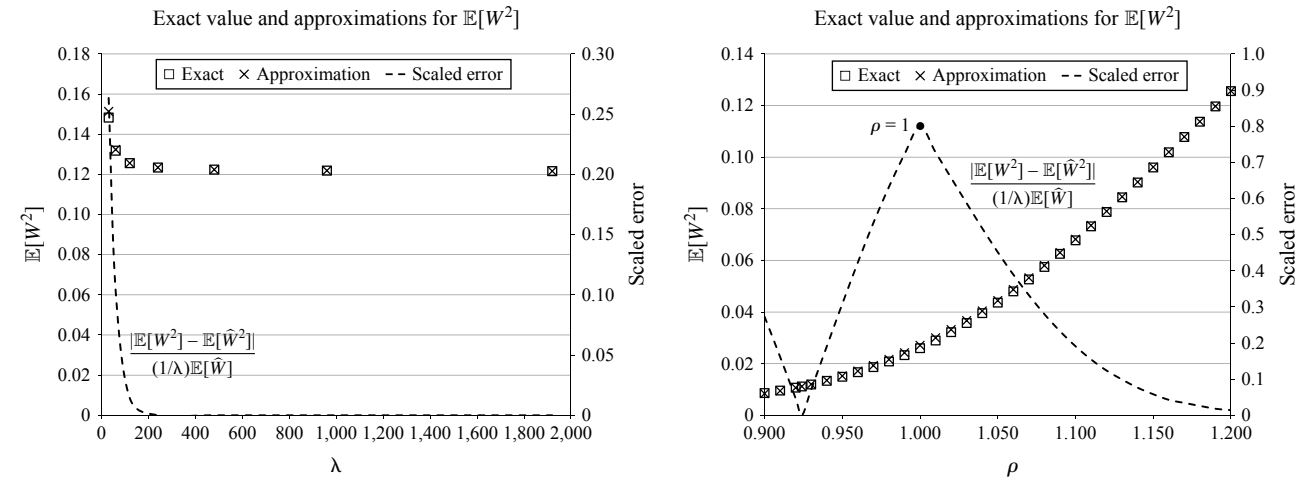
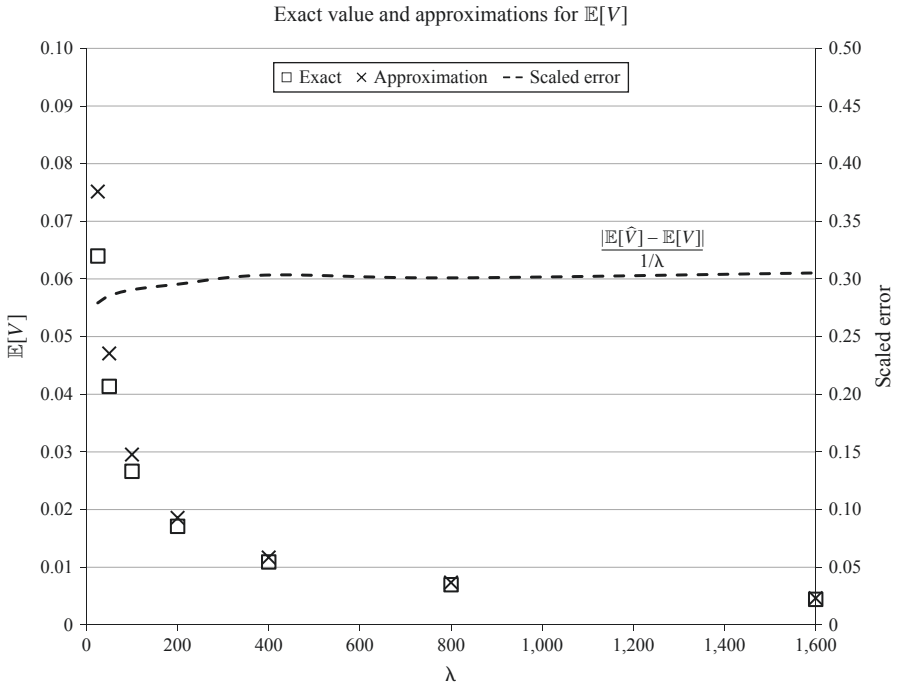


Figure 5. $M/D/1+GI$ with $\rho = 1$ ($\bar{w}_p = 0$) and $F_a = \text{Gamma}(0.5, 2)$



these moments to be finite, but it does guarantee that $\mathbb{E}[s_1^k]$ is not too large relative to $(\mathbb{E}[s_1])^k$. With subexponential distribution, the third moment can be very large, making our bounds loose. Figure 6 reports the simulation results for the $M/GI/1+GI$ queue with (i) a Pareto service time (where $\mathbb{E}[s_1^k] = (\alpha-1)^k/(\alpha^{k-1}(\alpha-k)) \cdot (\mathbb{E}[s_1])^k$, where α is the shape parameter). With $k=3$ and $\alpha = 5$, $\mathbb{E}[s_1^3] = 1.28(\mathbb{E}[s_1])^3$, the third moment is not too large and the performance is rather good. With $\alpha = 3$, the third moment is *infinite*. While the approximation error is not too large for the range of λ values we tried, it does grow slowly with λ; and (ii) a Log-Normal service time where $\mathbb{E}[s_1^k] = (1+\text{CoV}^2)^{(kn(k-1))/2} \cdot (\mathbb{E}[s_1])^k$, where CoV = 2 is the coefficient of variation of the distribution. With $k=3$, $\mathbb{E}[s_1^3] = 125(\mathbb{E}[s_1])^3$. For the Log-Normal distribution, $\mathbb{E}[s_1^k]/(\mathbb{E}[s_1])^k$ grows exponentially with k, and the approximation is relatively inaccurate.

Underloaded Queues. In Remark 3 we pointed out that the mathematical bounds are of the order of the quantity we are trying to approximate. It is a priori plausible that the bounds are not tight and that the actual error is small. Figure 7 rules this out: strictly underloaded queues (as in $\rho = 0.5$) fall outside of the scope of the Brownian models.

4. Static Optimization

We expect the universal accuracy of the Brownian approximation to translate to nearly optimal prescriptions. Randhawa (2016) shows in a many-server context how an $\mathcal{O}(1)$ performance-approximation gap translates to smaller $o(1)$ errors in staffing prescriptions.

We revisit a standard capacity optimization problem for the $M/GI/1+GI$ queue. We fix a patience distribution F_a with a bounded density $f_a(\cdot) \leqslant U$ (in particular, $\bar{F}_a(\infty) = 0$) and let the service time be given by $F_s^\mu(x) = F_s(\mu x)$, where $F_s(\cdot)$ is fixed. With these restrictions, variation within a queue family $\mathcal{Q}(H)$ reduces to variation of λ and μ, so we can write $(\lambda, \mu) \in \mathcal{Q}(H)$ instead of $p \in \mathcal{Q}(H)$ and $c_p = c_{\lambda,\mu}$.

Given strictly positive constants C_r, C_{ab}, C_w, we consider the total cost

$$\mathcal{C}(\mu) := C_r\mu + C_{ab}\lambda Ab_\mu + C_w\lambda\mathbb{E}[W_\mu],$$

and seek to solve

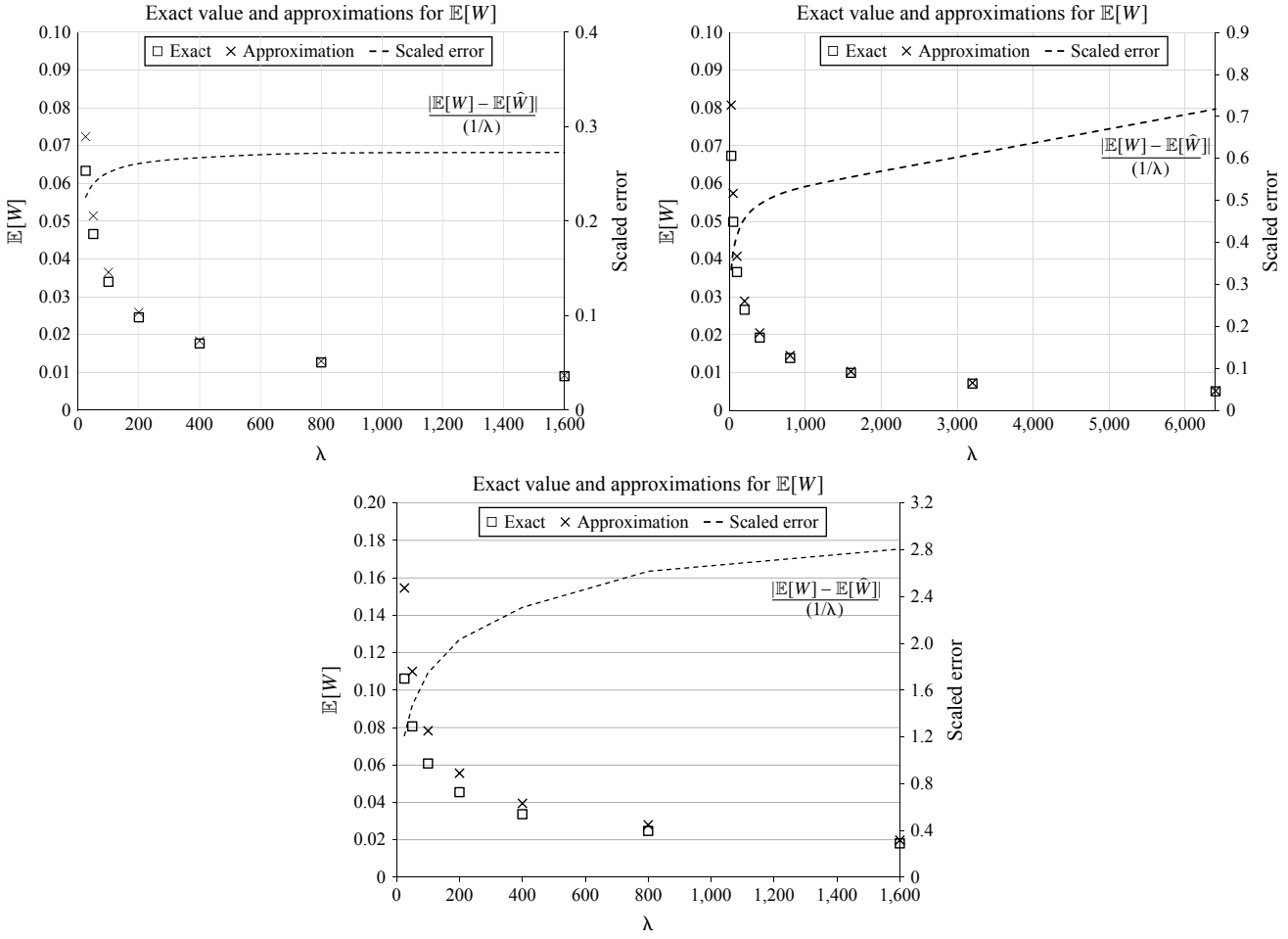
$$\mu_* = \arg\min_{\mu \geqslant \lambda\xi} \mathcal{C}(\mu). \tag{12}$$

Here, Ab_μ is the stationary fraction of customers who abandon the queue before being served, and W_μ follows the stationary distribution of the waiting time process, given that the service rate is μ. $\xi \in (0,1)$ is a prespecified lower bound on the fraction of customers who must be served.[4] The optimizer μ_* balances the cost of capacity and the combined costs of abandonment and delay.

Let V_μ be a random variable following the stationary distribution of the virtual waiting time process when the service rate is μ. Then, as in (11), $Ab_\mu = \mathbb{E}[F_a(V_\mu)]$ and

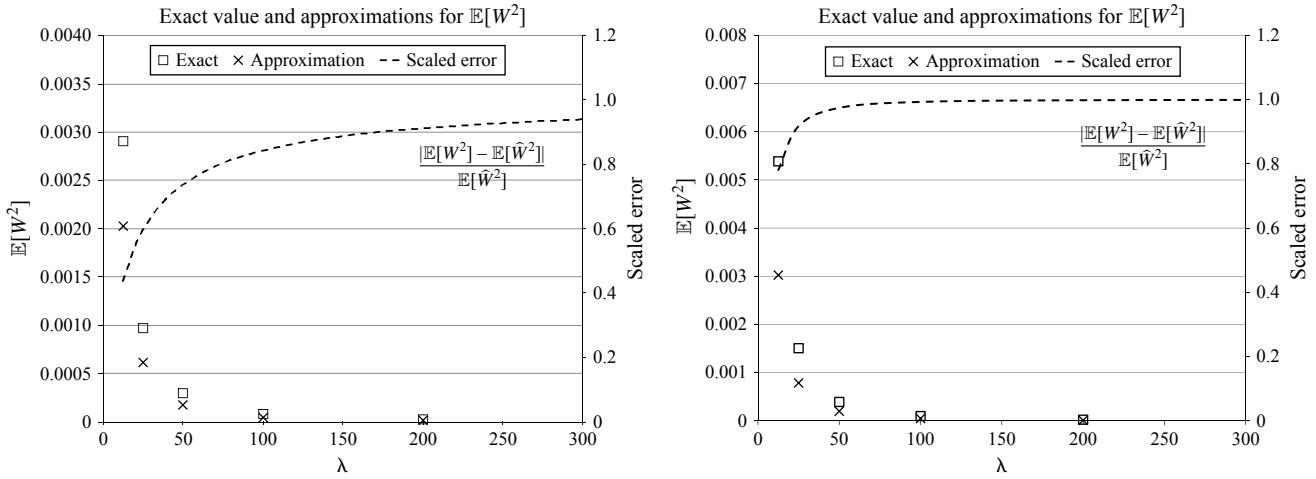
$$\mathcal{C}(\mu) = C_r\mu + C_{ab}\lambda\mathbb{E}[F_a(V_\mu)] + C_w\lambda\mathbb{E}[W_\mu].$$

Figure 6. Subexponential Service Times: Pareto Service Time Distribution with Shape Parameter $\alpha = 5$ (Top Left) and $\alpha = 3$ (Infinite Third Moment) (Top Right); Log-Normal Service Time (Bottom)



Notes. The figure displays the expected waiting time, approximation, and error as a function of the arrival rate λ. The arrival rate λ is varied, and the mean service time is set to $1/\lambda$ so that the utilization is kept at 1. Each replication runs for 100,000 time units. Per arrival rate λ in the tested range, we choose the number of replications so that, in expectation, 1.6E+9 arrivals are generated. In the collection of statistics, the first 10% of the arrivals are left out as a warm-up period.

Figure 7. An Underloaded Queue with $\rho = 0.5$ and $F_a(x) = 1 - e^{-2x} - 2xe^{-2x}$ (Erlang) (Left) and $F_a(x) = \frac{4}{7}(1 - e^{-4x}) + \frac{3}{7}(1 - e^{-x/2})$ (Hyperexponential) (Right)



The intuitive Brownian approximation for the total cost is

$$\hat{\mathscr{C}}(\mu) := C_r\mu + C_{ab}\lambda\mathbb{E}[F_a(\hat{V}_\mu)] + C_w\lambda\mathbb{E}[\hat{W}_\mu],$$

where $\hat{V}_\mu$ follows the stationary distribution of the Brownian queue with the service rate μ as in (4), and $\hat{W}_\mu = v \wedge \hat{V}_\mu$ is the Brownian approximation for the waiting time. The Brownian analogue of the optimization problem (12) is then

$$\hat{\mu}_* = \arg\min_{\mu \geqslant \lambda\xi}\{\hat{\mathscr{C}}(\mu)\}. \tag{13}$$

For $H > 1$, we define $\mathscr{Q}(H)$ as an *M-stable queue family* if, for λ such that $(\lambda, \lambda) \in \mathscr{Q}(H)$,

$$\{(\lambda, \mu)\colon \mu \in [\lambda\xi, \lambda(1 + Mc_{\lambda,\lambda})]\} \subseteq \mathscr{Q}(H).$$

Stability here is relative to changes in the service rate. The queue family is stable if it contains a sufficiently wide range of service-rate values.

Proposition 3. *Assume $\mathscr{Q}(H)$ is an M-stable family for $M > H_0 := (C_{ab} \cdot C^V_{H,1}U + C_w \cdot C^W_{H,1})/C_r$ (with $C^V_{H,1}, C^W_{H,1}$ as in Lemma 1). Then, with $C^{Ab}_H, C_{H,1}$ as in Corollary 2,*

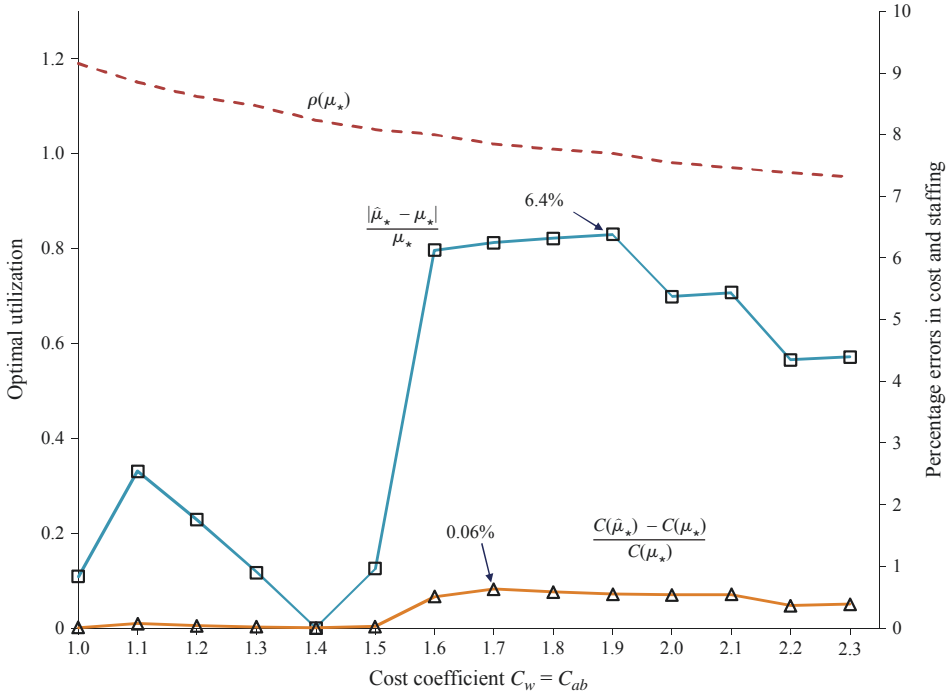
$$0 \leqslant \mathscr{C}(\hat{\mu}_*) - \mathscr{C}(\mu_*) \leqslant \left(\frac{1}{\lambda c^2_{\lambda,\mu_*}} + \frac{1}{\lambda c^2_{\lambda,\hat{\mu}_*}}\right) C_{ab}C^{Ab}_H + 2C_wC_{H,1},$$
$$\text{for all } \lambda \text{ s.t. } (\lambda,\lambda) \in \mathscr{Q}(H).$$

Consequently, if $c_{\lambda,\hat{\mu}_*}, c_{\lambda,\mu_*} \geqslant 1/\sqrt{\lambda}$, the error is bounded by the constant $2(C_{ab}C^{Ab}_H + C_wC_{H,1})$.

The stability requirement guarantees that both the optimal service rate μ_* and the approximate rate $\hat{\mu}_*$ are within the queue family so that we can build on our performance analysis bounds in Theorem 1 and its corollaries. The Uniform, Exponential, and Hyperexponential distributions with $\mathbb{E}[v_1] = 1$ are instances of patience distributions for which the queue family $\mathscr{Q}(H)$ is M-stable for any M smaller than H. These three distributions have, in fact, $c_p \equiv c_\lambda$ (the concentration depends only on the arrival rate) for $\mu \leqslant \lambda(1 + Mc_{\lambda,\lambda})$. For example, the Exponential case (λ, μ) with $\mu \leqslant \lambda + M\sqrt{\lambda}$ is in the queue family with $c_p \equiv c_\lambda = 1/\sqrt{\lambda}$. In the case of the Gamma distribution, by contrast, the concentration does depend on the service rate (through its dependence on ρ; see Table 1). For all of these distributions $c_{\lambda,\mu} \geqslant 1/\sqrt{\lambda}$, as required in the second part of the proposition.

Figure 8 is a numerical illustration for Hyperexponential patience, Exponential service time, and $\lambda = 100$. To bring out multiple utilization levels, we consider a range of values for $C_{ab} = C_w$ (normalizing the cost so that $C_r = 1$). We compare the cost at the optimal solution $\mathscr{C}(\mu_*)$ (found through search) against the cost under the approximate solution $\mathscr{C}(\hat{\mu}_*)$. At the lowest end of the cost spectrum ($C_{ab} = C_w = 1$), the queue operates optimally with a utilization of about 1.2 and can be interpreted as overloaded,[5] whereas when the coefficient increases to 2.3, the optimal utilization is below 1, which we interpret as critical loading. Notably, the

Figure 8. (Color online) Service-Rate Optimization: $\lambda = 100$, Exponential Service Time, and Hyperexponential Patience: $F_a(x) = \frac{4}{7}(1 - e^{-4x}) + \frac{3}{7}(1 - e^{-x/2})$



Note. For different values of $C_{ab} = C_w$ ($C_r \equiv 1$), the graph displays the cost optimality gap (□) and the capacity-prescription gap (△).

percentage error in the resulting cost, $(\mathscr{C}(\hat{\mu}_*)-\mathscr{C}(\mu_*))/\mathscr{C}(\mu_*)$, is smaller than the percentage error in staffing $|\hat{\mu}_*-\mu_*|/\mu_*$. This is because the objective function is relatively flat around the optimal solution.

Proof. We claim that if λ is such that $(\lambda,\lambda)\in\mathcal{Q}(H)$, then both

$$\mu_* \leqslant \lambda(1+Mc_{\lambda,\lambda}) \quad \text{and} \quad \hat{\mu}_* \leqslant \lambda(1+Mc_{\lambda,\lambda}), \tag{14}$$

where M is as in the M-stability requirement. This guarantees that both (λ,μ_*) and $(\lambda,\hat{\mu}_*)$ are in $\mathcal{Q}(H)$. Propositions 1 and 2 then guarantee that for either $\mu\in\{\mu_*,\hat{\mu}_*\}$,

$$\begin{aligned}
&|\mathscr{C}(\mu)-\hat{\mathscr{C}}(\mu)|\\
&\quad\leqslant C_{ab}\lambda\mathbb{E}[|F_a(V_\mu)-F_a(\hat{V}_\mu)|]+C_w\lambda|\mathbb{E}[W_\mu]-\mathbb{E}[\hat{W}_\mu]|\\
&\quad\leqslant C_{ab}C_H^{Ab}\frac{1}{\lambda c_{\lambda,\mu}^2}+C_wC_{H,1}.
\end{aligned}$$

By definition, $\mathscr{C}(\hat{\mu}_*)-\mathscr{C}(\mu_*)\geqslant 0$ and $\hat{\mathscr{C}}(\hat{\mu}_*)-\hat{\mathscr{C}}(\mu_*)\leqslant 0$. In sum,

$$\begin{aligned}
0&\leqslant \mathscr{C}(\hat{\mu}_*)-\mathscr{C}(\mu_*)\\
&=\mathscr{C}(\hat{\mu}_*)-\hat{\mathscr{C}}(\hat{\mu}_*)+\hat{\mathscr{C}}(\hat{\mu}_*)-\hat{\mathscr{C}}(\mu_*)+\hat{\mathscr{C}}(\mu_*)-\mathscr{C}(\mu_*)\\
&\leqslant \mathscr{C}(\hat{\mu}_*)-\hat{\mathscr{C}}(\hat{\mu}_*)+\hat{\mathscr{C}}(\mu_*)-\mathscr{C}(\mu_*)\\
&\leqslant \left(\frac{1}{\lambda c_{\lambda,\mu_*}^2}+\frac{1}{\lambda c_{\lambda,\hat{\mu}_*}^2}\right)C_{ab}C_H^{Ab}+2C_wC_{H,1}\\
&\leqslant 2(C_{ab}C_H^{Ab}+C_wC_{H,1}),
\end{aligned}$$

where the last inequality follows from our assumption that $c_{\lambda,\mu}\geqslant 1/\sqrt{\lambda}$.

To prove (14) we use the two assumptions $(\lambda,\lambda)\in\mathcal{Q}(H)$ (and has $\bar{w}_p=0$) and $F_a(x)=\int_0^x f_a(y)\,dy\leqslant Ux$ and apply Lemma 1 to get

$$\begin{aligned}
\hat{\mathscr{C}}(\lambda)&=C_r\lambda+C_{ab}\lambda\mathbb{E}[F_a(\hat{V}_\mu)]+C_w\lambda\mathbb{E}[\hat{W}_\mu]\\
&\leqslant C_r\lambda+\lambda C_{ab}C_{H,1}^VUc_{\lambda,\lambda}+\lambda C_wC_{H,1}^Wc_{\lambda,\lambda}\\
&=C_r\lambda(1+H_0c_{\lambda,\lambda}),
\end{aligned}$$

where H_0 is as in the statement of the theorem. For any service rate $\mu>\lambda+\lambda Mc_{\lambda,\lambda}$,

$$\hat{\mathscr{C}}(\mu)\geqslant C_r\mu>C_r\lambda(1+H_0c_{\lambda,\lambda})\geqslant\hat{\mathscr{C}}(\lambda),$$

so it must be the case that $\hat{\mu}_*\leqslant\lambda(1+Mc_{\lambda,\lambda})$. The proof for μ_* is identical given the corresponding bounds in Lemma 1. □

5. Dynamic Optimization

In this section, we turn to ergodic control. We consider a specific and well-studied control problem, that of service speed/rate control for the $M/G/1$ queue, similar to that considered in Doshi (1978) and Mitchell (1973).

Our objective is to show how the same approach, via an intuitive Brownian control problem, yields controls that are universally nearly optimal. As a by-product, we illustrate how the generator view (à la Stein's method) is extended to dynamic control.

In this section, the service requirement is drawn from a (fixed) distribution $F_s(\cdot)$ and the service rate is controllable. The distribution $F_s(\cdot)$ has a mean of 1, standard deviation of σ, and finite exponential moments.

This is different from our model in the previous sections, where the server works at a rate of 1 and arrival i brings an amount of work s_i whose mean is a parameter of the model (and can vary within the queue family). Here, following the model of Doshi (1978), the rate itself is the choice.

The base service rate equals λ and can be sped up using a multiplier $1+\theta$, where $\theta\geqslant 0$. The actual service rate is then

$$\mu(\theta)=\lambda(1+\theta).$$

The speed-up is $\lambda\theta$, and the larger it is, the costlier it is but the smaller the workload.

We consider the sum of a long-run-average polynomial holding cost and a quadratic control cost:

$$\mathcal{J}_{p,m}^V(\theta):=\limsup_{t\to\infty}\frac{1}{t}\mathbb{E}_x\left[\int_0^t[h(V(\theta,s))^m+(\lambda\theta(s))^2]\,ds\right],$$

where $m\geqslant 2$ and $V(\theta,t)$ is the workload at time t under the control θ. The *workload control problem* is given by

$$\mathcal{J}_{p,m}^{V,*}=\inf_{\theta\in\Theta_V}\mathcal{J}_{p,m}^V(\theta), \tag{15}$$

where Θ_V is the family of nonanticipative controls in the standard sense—that is, with respect to the history

$$\mathcal{F}_t=\sigma_f\left\{V(\theta,s),\int_0^s\theta(u)\,du;s\leqslant t\right\}.$$

Because the service time distribution is fixed and arrivals are Poisson, the pair $p=(\lambda,h)$ captures the moving pieces in the model, and given H, we let $\mathcal{Q}(H):=\{(\lambda,h)\colon\lambda\geqslant H^{-1},0<h\leqslant H\}$.

Within an asymptotic framework one can "push" the queue into different asymptotic regimes by specifying how h scales with λ. Consider the case $m=2$. In optimality, the stationary workload is proportional to $h^{-1/4}\sqrt{\lambda}$ (see Lemma 6). If $h_\lambda\to\bar{h}>0$ as $\lambda\to\infty$, the optimally controlled stationary workload is of the order of $\sqrt{\lambda}$ as in the so-called conventional heavy-traffic regime. If, instead, $h_\lambda\to 0$ as $\lambda>0$, the optimal workload is orders of magnitude larger. With any bounded sequence $h^\lambda\leqslant H$, the utilization approaches 100% as λ grows. The way in which the utilization approaches 100% is the regime.

Our result is universal and obviates the need to interpret whether $\lambda=100$ and $h=0.1$ should be read as

$h^\lambda = 1/\sqrt{\lambda}$ (and, hence, converging to 0) or as a constant 0.1. Our recommended control and the performance guarantee are given in terms of h itself rather than an interpretation thereof.

When positive, the drift of V under a speed-up control θ is $\lambda - \lambda(1+\theta) = -\lambda\theta$. Given an admissible control $(\theta(t), t \geqslant 0)$,

$$V(\theta,t) = V(0) - \lambda \int_0^t \theta(s)\,ds + \int_0^t \lambda(1+\theta(s))\mathbb{1}\{V(\theta,s)=0\}\,ds + M(t),$$

where $M(t)$ is a zero mean martingale that does not depend on the control with predictable quadratic variation

$$\langle M(t)\rangle = \lambda(1+\sigma^2)t.$$

It is heuristically intuitive to consider the reflected diffusion,

$$\hat{Y}(\theta,t) = \hat{Y}(0) - \lambda \int_0^t \theta(s)\,ds + \sqrt{\lambda(1+\sigma^2)}B(t) + \lambda \int_0^t (1+\theta(s))\,dL(s),$$
$$L(t) = \int_0^t \mathbb{1}\{\hat{Y}(\theta,s)=0\}\,ds,$$

as a proxy for V. The diffusion-optimal control is determined from the Brownian counterpart of the workload control problem (15),

$$\mathcal{J}^{Y,*}_{p,m} = \inf_{\theta\in\Theta_Y} \limsup_{t\to\infty} \frac{1}{t}\mathbb{E}_x\left[\int_0^t [h(\hat{Y}(\theta,s))^m + (\lambda\theta(s))^2]\,ds\right], \tag{16}$$

where Θ_Y is the family of processes $(\theta(s), s \geqslant 0)$ that are nonnegative and progressively measurable with respect to the self-filtration of the Brownian motion, and θ takes values in $[0,\infty)$.[6] A more complicated version of this diffusion control problem, where a finite buffer is also optimized, appears in Ghosh and Weerasinghe (2007).

As is typical for diffusion control problems, we establish a verification lemma that stipulates the optimality of a stationary control derived from the following HJB equation:

$$\min_{z\geqslant 0}\{\hat{\mathcal{A}}^z_\lambda \Psi(x) + (\lambda z)^2 + hx^m\} = \gamma_{p,m},$$
$$\Psi(0) = \Psi^{(1)}(0) = 0 \quad \text{and} \quad \Psi^{(1)}(x) \geqslant 0, \quad \text{for all } x \geqslant 0, \tag{17}$$

where, given a constant $z \geqslant 0$, $\hat{\mathcal{A}}^z_\lambda$ is the operator

$$\hat{\mathcal{A}}^z_\lambda = -\lambda z\frac{\partial}{\partial x} + \frac{1}{2}\lambda(1+\sigma^2)\frac{\partial^2}{\partial x^2}. \tag{18}$$

Given $x \geqslant 0$, $p \in \mathcal{Q}(H)$, and a pair $(\Psi_{p,m}(\cdot), \gamma_{p,m})$ solving (17), the optimal z is trivially given by

$$\mathcal{S}^*_{p,m}(x) = \frac{\Psi^{(1)}_{p,m}(x)}{2\lambda}, \tag{19}$$

so that the HJB equation translates to

$$\tfrac{1}{2}\lambda(1+\sigma^2)\Psi^{(2)}_{p,m}(x) - \tfrac{1}{4}[\Psi^{(1)}_{p,m}(x)]^2 + hx^m = \gamma_{p,m},$$
$$\Psi_{p,m}(0) = \Psi^{(1)}_{p,m}(0) = 0 \quad \text{and} \quad \Psi^{(1)}_{p,m}(x) \geqslant 0, \quad \text{for all } x \geqslant 0. \tag{20}$$

Using the speed-up $\mathcal{S}^*_{p,m}(x)$, derived from the diffusion control problem, results in the workload dynamics

$$V^*(t) = V(0) - \lambda \int_0^t \mathcal{S}^*_{p,m}(V^*(s))\,ds + \int_0^t \lambda(1+\mathcal{S}^*_{p,m}(V^*(s)))\mathbb{1}\{V^*(s)=0\}\,ds + M(t),$$

and the control follows the trajectory

$$\hat{\theta}^*_{p,m}(t) = \mathcal{S}^*_{p,m}(V^*(t)). \tag{21}$$

Theorem 2 (Universality of the Diffusion Solution for the Workload Problem). *Fix $m \geqslant 2$. The HJB equation has a unique solution. The stationary control $\hat{\theta}^*_{p,m}$ and the corresponding workload process $V(\hat{\theta}^*_{p,m}, \cdot) = V^*(\cdot)$ yield a cost that is nearly optimal for the workload control problem: there exists a constant $C_{H,m}$ such that, for all $p \in \mathcal{Q}(H)$,*

$$\mathcal{J}^V_{p,m}(\hat{\theta}^*_{p,m}) = \mathcal{J}^{V,*}_{p,m} \pm C_{H,m}B_m(\lambda,h)\mathcal{J}^{Y,*}_{p,m-1} = \mathcal{J}^{V,*}_{p,m} \pm C_{H,m}B_m(\lambda,h)\mathcal{J}^{V,*}_{p,m-1},$$

where $B_m(\lambda,h) := (h^{m-1}\lambda^{-2(m-1)})^{1/((m+1)(m+2))}$.

In the special case where $m = 2$, $\mathcal{J}^V_{p,2}(\hat{\theta}^*_{p,2}) = \mathcal{J}^{V,*}_{p,2}$: the diffusion-based stationary control is optimal (not just nearly optimal) for the workload control problem.

Notably, the optimality gap for the problem with holding cost $hV(\theta,t)^m$ is given in terms of the optimal cost in the problem with the lower-order holding cost $hV(\theta,t)^{m-1}$. This parallels Theorem 1 where the approximation gap for the kth moment of the virtual waiting time is given in terms of the $(k-1)$st moment. For $m > 2$, because

$$B_m(\lambda,h) \leqslant (H \vee 1)\lambda^{-2(m-1)/((m+1)(m+2))},$$

the optimality gap is negligible relative to the optimal cost of the lower-order problem. For $m = 2$, the optimality gap is 0.

Optimality Gaps and Generator Comparisons. The HJB equation is a Ricatti-type equation. Ricatti equations are first-order, nonlinear ODEs and are relatively well studied in the literature. Here, the added element is the requirement on the positivity of the solution $\Psi^{(1)}(x)$. The challenge for us is that we are not interested merely in solvability (existence and uniqueness) but, rather, in the derivative bounds. Having set up the ingredients, the following informal discussion parallels the one in Section 3.1, with the main exception being the replacement of the Poisson equation with the HJB equation. We spell these steps out to make these connections clear.

The workload control problem for the Markov process (not the diffusion) is studied by Doshi (1978). Restricting attention to the Markov controls, Doshi proves that *if* there exists a nonnegative service rate $(\xi(x), x \geqslant 0)$, a function Ψ, and a nonnegative constant γ^D, such that

(i) for all admissible (according to Doshi's definition) $(\zeta(t,x); t, x \geqslant 0)$

$$\lim_{t\to\infty} \frac{\mathbb{E}_x[\Psi(V(\zeta,t))]}{t} = 0, \quad x \geqslant 0, \quad \text{and}$$

(ii) the constant γ^D together with the function $\Psi(\cdot)$ satisfies the Bellman equation

$$\gamma^D = \min_{z\geqslant 0}\{(\lambda z)^2 + hx^m - \lambda(1+z)\Psi^{(1)}(x) + \lambda\mathbb{E}[\Psi(x+s_1)-\Psi(x)]\}, \quad x \geqslant 0;$$

and $\xi(x)$ is the minimizer:

$$\gamma^D = (\lambda\xi(x))^2 + hx^m - \lambda(1+\xi(x))\Psi^{(1)}(x) + \lambda\mathbb{E}[\Psi(x+s_1)-\Psi(x)], \quad x \geqslant 0,$$

then $\xi(x)$ is the workload optimal control (and $\mu(x) = \lambda(1+\xi(x))$ is the optimal service rate); see theorem 4 in Doshi (1978). In operator notation, the Bellman equation translates to

$$\gamma^D = \inf_{z\geqslant 0}\{(\lambda z)^2 + hx^m + \mathcal{A}^z_\lambda\Psi(x)\},$$

where $\mathcal{A}^z_\lambda$ is the operator

$$\mathcal{A}^z_\lambda\Psi(x) = -\lambda(1+z)\Psi^{(1)}(x) + \lambda\mathbb{E}[\Psi(x+s_1)-\Psi(x)]. \quad (22)$$

By Taylor's expansion,

$$\Psi(x+s_1)-\Psi(x) = \Psi^{(1)}(x)s_1 + \tfrac{1}{2}\Psi^{(2)}(x)s_1^2 + \mathcal{O}(|\Psi^{(3)}|^*_{x,s_1}s_1^3),$$

where $|f|^*_{x,y} = \sup_{z\in[x,x+y]}|f(z)|$. As $\mathbb{E}[s_1] = 1$ and $\mathbb{E}[s_1^2] = 1+\sigma^2$,

$$\lambda\mathbb{E}[\Psi(x+s_1)-\Psi(x)] = \lambda\Psi^{(1)}(x) + \tfrac{1}{2}\Psi^{(2)}(x)\lambda(1+\sigma^2) + \lambda\mathcal{O}(\mathbb{E}[|\Psi^{(3)}|^*_{x,s_1}s_1^3]).$$

Defining $e(x) = \mathbb{E}[|\Psi^{(3)}|^*_{x,s_1}s_1^3]$, we have

$$\mathcal{A}^z_\lambda\Psi(x) = -\lambda(1+z)\Psi^{(1)}(x) + \lambda\mathbb{E}[\Psi(x+s_1)-\Psi(x)] = -\lambda z\Psi^{(1)}(x) + \tfrac{1}{2}\lambda(1+\sigma^2)\Psi^{(2)}(x) + \lambda\mathcal{O}(e(x)).$$

We conclude that the Bellman equation should satisfy

$$\gamma^D \approx \inf_{z\geqslant 0}\{(\lambda z)^2 + hx^m + \hat{\mathcal{A}}^z_\lambda\Psi(x) + \lambda\mathcal{O}(e(x))\}, \quad x \geqslant 0,$$

where $\hat{\mathcal{A}}^z_\lambda$ is defined as for the diffusion in (18). In turn, the Bellman equation for the jump process is "almost" the HJB equation for a diffusion control problem. If $\lambda\mathcal{O}(e(x))$ is suitably bounded, then the solutions to the jump-process's Bellman equation and the HJB equation should be suitably close. This connection is at the core of the argument.

We do not rely directly on Doshi's analysis of the Markov policies in our proofs. We allow for a larger family of policies, and within this larger family, we show the universal near optimality of the stationary policy that arises from the *diffusion*'s HJB equation.

As our outline above suggests, the first ingredient in the proof of Theorem 2 is an analysis of the HJB equation and the third derivative of its solution.

For a family of positive pairs $\{(a_{p,m}, b_{p,m}), p \in \mathcal{Q}(H)\}$, we write $a_{p,m} \sim b_{p,m}$ if there exists a constant $C > 1$ such that $C^{-1} \leqslant a_{p,m}/b_{p,m} \leqslant C$ for all $p \in \mathcal{Q}(H)$.

Lemma 5 (HJB and the Diffusion Control Properties). *Fix $m \geqslant 2$. A unique solution $(\Psi_{p,m}, \gamma_{p,m})$ to the HJB equation exists for each $p \in \mathcal{Q}(H)$.*

Properties:

(i) *The third derivative satisfies, for all $p \in \mathcal{Q}(H)$,*

$$-C^0_{H,m}B_m(\lambda,h)\left(\frac{xh^{1/(m+2)}}{\lambda^{2/(m+2)}}\right)^{m-1}\gamma_{p,m-1} \leqslant \lambda\Psi^{(3)}_{p,m}(x) \leqslant C^0_{H,m}B_m(\lambda,h)\left(1+\left(\frac{xh^{1/(m+2)}}{\lambda^{2/(m+2)}}\right)^{3m/2}\right)\gamma_{p,m-1},$$

where $C^0_{H,m}$ depends only on H and m.

(ii) *The constant $\gamma_{p,m}$ satisfies $\gamma_{p,m} \sim \lambda^{2m/(m+2)}h^{2/(m+2)}$.*

(iii) *For $m = 2$, $(\Psi_{p,2}(x), \gamma_{p,2}) = (\sqrt{h}x^2, \lambda(1+\sigma^2)\sqrt{h})$, so that $\Psi^{(3)}_{p,2} \equiv 0$.*

Verification: $\gamma_{p,m}$ is the optimal long-run average cost in the diffusion control problem (i.e., $\mathcal{J}^{Y,*}_{p,m} = \gamma_{p,m}$), and it is optimal to use the stationary control (19).

Lemma 6 below is the control analogue of the concentration bounds in Lemma 1. It provides order-of-magnitude estimates that are subsequently useful for the optimality-gap bounds. Equation (24) captures how the optimally controlled workload scales with h and λ.

Lemma 6 (A Priori Bounds). *Fix $m \geqslant 2$. Then*

$$\mathcal{J}^{V,*}_{p,m} \sim \mathcal{J}^{Y,*}_{p,m}. \quad (23)$$

In particular, $\mathcal{J}^{V,*}_{p,m} \sim \gamma_{p,m} \sim \lambda^{2m/(m+2)} h^{2/(m+2)}$. *Moreover,* $V(\hat{\theta}^*_{p,m}, t)$ *is positive recurrent, and for any* $k \geqslant 2$,

$$\mathbb{E}[V(\hat{\theta}^*_{p,m}, \infty)^k] = \lim_{t\to\infty} \frac{1}{t} \mathbb{E}_x \left[\int_0^t (V(\hat{\theta}^*_{p,m}, s))^k \, ds \right] \sim \lambda^{2k/(m+2)} h^{-k/(m+2)}, \quad x \geqslant 0. \tag{24}$$

Recall the operators $\mathcal{A}^z_\lambda$ and $\hat{\mathcal{A}}^z_\lambda$ as defined in (22) and (18).

Lemma 7. *Fix* $p = (\lambda, h)$ *and* $m \geqslant 2$, *and let* $(\Psi_{p,m}, \gamma_{p,m})$ *be the solution to the HJB equation. For any admissible control* θ,

$$\mathbb{E}_x \left[\int_0^t [h(V(\theta, s))^m + (\lambda\theta(s))^2] \, ds \right] \geqslant \Psi_{p,m}(x) - \mathbb{E}_x[\Psi_{p,m}(V(\theta, t))] + \gamma_{p,m} t + \mathbb{A}^x_{p,m}(\theta, t), \quad x, t \geqslant 0,$$

where

$$\mathbb{A}^x_{p,m}(\theta, t) = \mathbb{E}_x \left[\int_0^t (\mathcal{A}^{\theta(s)}_\lambda \Psi_{p,m}(V(\theta, s)) - \hat{\mathcal{A}}^{\theta(s)}_\lambda \Psi_{p,m}(V(\theta, s))) \, ds \right], \quad x, t \geqslant 0.$$

If $\theta = \hat{\theta}^*_{p,m}$, *then the inequality is replaced with an equality. In the special case where* $m = 2$, $\mathbb{A}^x_{p,m}(\theta, t) \equiv 0$ *for any control* θ.

In proving the near optimality of the diffusion-based control $\hat{\theta}^*_{p,m}$ we must show that no other control can do much better. To that end, we require performance bounds for all "reasonable" control policies and not only for the optimal control (which we do not explicitly identify or assume to exist) or the diffusion-optimal control.

A family of admissible policies $(\theta_{p,m}, p \in \mathcal{Q}(H))$ is said to be *order optimal* if

$$\mathcal{J}^V_{p,m}(\theta_{p,m}) \sim \mathcal{J}^{V,*}_{p,m}.$$

Lemma 8. *Fix* m *and let* $(\Psi_{p,m}, \gamma_{p,m})$ *be the* (*family of*) *solutions to the HJB equation. Then, there exist constants* $C^1_{H,m}, C^2_{H,m}$ *such that, for any order optimal family of policies* $\{\theta_{p,m}, p \in \mathcal{Q}(H)\}$,

$$\liminf_{t\to\infty} \frac{1}{t} \mathbb{A}^x_{p,m}(\theta_{p,m}, t) \geqslant -C^1_{H,m} B_m(\lambda, h) \mathcal{J}^{Y,*}_{p,m-1}, \quad x \geqslant 0, \ p \in \mathcal{Q}(H),$$

and under the stationary policy $\hat{\theta}^*_{p,m}$,

$$\limsup_{t\to\infty} \frac{1}{t} \mathbb{A}^x_{p,m}(\hat{\theta}^*_{p,m}, t) \leqslant C^2_{H,m} B_m(\lambda, h) \mathcal{J}^{Y,*}_{p,m-1}, \quad x \geqslant 0, \ p \in \mathcal{Q}(H).$$

If an optimal control $\theta^*_{p,m}$ exists for each $p \in \mathcal{Q}(H)$, then the family $(\theta^*_{p,m}, \ p \in \mathcal{Q}(H))$ is order optimal, in which case Lemma 8 immediately implies bounds for the optimal controls. If optimal controls do not exist for some p, then we must work, instead, with the infimum over the admissible controls.

Proof of Theorem 2. From Lemma 7 it follows that, under any admissible control θ,

$$\mathbb{E}_x \left[\int_0^t [h(V(\theta, s))^m + (\lambda\theta(s))^2] \, ds \right] \geqslant \Psi_{p,m}(x) - \mathbb{E}_x[\Psi_{p,m}(V(\theta, t))] + \gamma_{p,m} t + \mathbb{A}^x_{p,m}(\theta, t), \quad x, t \geqslant 0.$$

Dividing by t and using Lemma 8, we have for any order optimal family of policies

$$\limsup_{t\to\infty} \frac{1}{t} \mathbb{E}_x \left[\int_0^t [h(V(\theta, s))^m + (\lambda\theta(s))^2] \, ds \right] \geqslant \gamma_{p,m} - C^1_{H,m} B_m(\lambda, h) \mathcal{J}^{Y,*}_{p,m-1}.$$

Recall that $\mathcal{J}^{V,*}_{p,m} = \inf_{\theta\in\Theta_V} \mathcal{J}^V_{p,m}(\theta) < \infty$, where finiteness follows from Lemma 6. For each p, let $\tilde{\theta}_p \in \Theta_V$ be such that

$$\mathcal{J}^V_{p,m}(\tilde{\theta}_p) \leqslant \mathcal{J}^{V,*}_{p,m} + \tfrac{1}{2} C^1_{H,m} B_m(\lambda, h) \mathcal{J}^{Y,*}_{p,m-1}.$$

By Lemma 6, $\mathcal{J}^{V,*}_{p,m} \sim \mathcal{J}^{Y,*}_{p,m} \sim \gamma_{p,m}$. Because $\gamma_{p,m-1} \leqslant (h/\lambda^2)^{2/((m+1)(m+2))} \gamma_{p,m} \leqslant H^{6/((m+1)(m+2))} \gamma_{p,m}$ and $B_m(\lambda, h) \leqslant H^{3(m-1)/((m+1)(m+2))}$, we have that $(\theta_p, \ p \in \mathcal{Q}(H))$ is an order optimal family and, in turn, that

$$\mathcal{J}^{V,*}_{p,m} \geqslant \mathcal{J}^V_{p,m}(\tilde{\theta}_p) - \tfrac{1}{2} C^1_{H,m} B_m(\lambda, h) \mathcal{J}^{Y,*}_{p,m-1} \geqslant \gamma_{p,m} - \tfrac{3}{2} C^1_{H,m} B_m(\lambda, h) \mathcal{J}^{Y,*}_{p,m-1}.$$

Using the second parts of Lemmas 7 and 8 for $\theta = \hat{\theta}^*_{p,m}$, we have

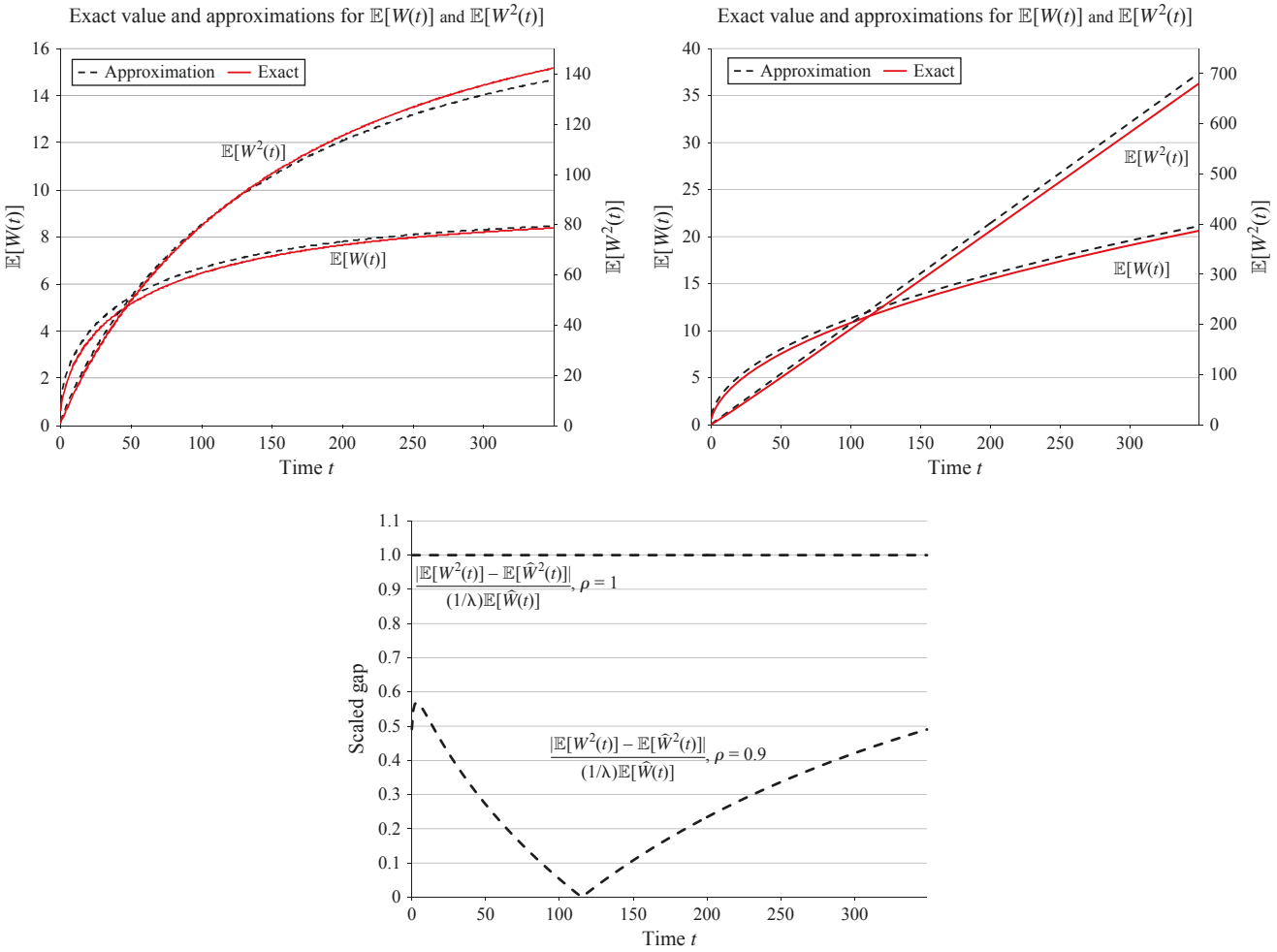
$$\mathcal{J}^{V,*}_{p,m} \leqslant \mathcal{J}^V_{p,m}(\hat{\theta}^*_{p,m}) = \limsup_{t\to\infty} \frac{1}{t} \mathbb{E}_x \left[\int_0^t [h(V(\hat{\theta}^*_{p,m}, s))^m + (\lambda \hat{\theta}^*_{p,m}(V(s)))^2] \, ds \right] \leqslant \gamma_{p,m} + C^2_{H,m} B_m(\lambda, h) \mathcal{J}^{Y,*}_{p,m-1}.$$

Thus, the error bounds hold with $C_{H,m} = \frac{3}{2} C^1_{H,m} + C^2_{H,m}$. Finally, it follows from Lemma 6 that $\mathcal{J}^{Y,*}_{p,m-1} \sim \mathcal{J}^{V,*}_{p,m-1}$. □

6. Concluding Remarks

Brownian approximations, like the central limit theorem that they generalize to queueing processes, provide significant tractability. They are typically supported by limit theorems, and these, in turn, are based on assumptions reflecting an implicit interpretation of a concrete system at hand.

Figure 9. (Color online) Time-Dependent Performance for $M/M/1$ with $\mu = 1$: $\rho = 0.9$ (Top Left) and $\rho = 1$ (Top Right), Scaled Absolute Gaps (Bottom) (for $\rho = 0.9$, the Sign of the Gap Changes at About $t = 110$.)



Notes. In the left and right panels, the black dashed lines correspond to the approximations, and the red solid lines show the exact values. As t grows, the scaled gap will eventually reach that of the stationary distributions.

In this paper, we establish the universality of the most intuitive Brownian approximation of the fundamental $M/GI/1 + GI$ queue. We cover performance analysis as well as static and dynamic optimization.

From a toolbox perspective, this paper supports the often applied heuristic of using Brownian approximations for modeling. From a technical perspective, our analysis presents a generalizable framework based on queue families that can be used, we hope, to study universality for extensions to our base model (to multiple servers, multiple customer classes, etc.).

Our analysis of *stationary* performance leaves open the question of whether Brownian approximations are universally accurate for *transient* (time-dependent) performance. A simple experiment suggests that the answer may be positive. For the $M/M/1$ queue, we computed the time-dependent first and second moments (starting empty) using known expressions for the time-dependent distribution (Asmussen 2003, theorem III.8.5). We also computed these moments for the corresponding Brownian queue (Harrison 1985, section 3.4). The results are plotted in Figure 9.

Evidently, the $M/M/1$ and its natural Brownian queue are very close for both values of ρ. The bottom graph suggests that Proposition 1 extends to time-dependent performance—that is, that $\mathbb{E}[W^2(t)] - \mathbb{E}[\hat{W}^2(t)] = \pm(C_{H,2}/\lambda)\mathbb{E}[\hat{W}(t)]$ for suitable constant $C_{H,2}$. That the gap is exactly $\mathbb{E}[\hat{W}(t)]$ for $\rho = 1$ is expected because, for a null recurrent or transient $M/M/1$, the boundary is rarely hit and the reflection has little effect. The second moments of the free Brownian motion and the free Poisson process are the same. The scaled gap varies with time for $\rho = 0.9$ but remains bounded.

The motivation for using Brownian approximations is their relative tractability. Time-dependent expectations for the $M/GI/1 + GI$ are difficult to compute, but for the Brownian queue, these expectations can be computed by solving suitable PDEs. To bound the approximation error, the theory of *strong approximations* provides a framework for sample path comparisons, but it

may be too heavy a hammer. It seems that, to compare marginal moments (rather than sample path gaps), tighter bounds can be obtained by using an approach that builds on the underlying Markovian structure of the queue. Our small experiment above suggests that it may be possible to expand universality and the bounds that accompany it, which express the error in the kth-moment approximation in terms of the $(k-1)$st moment, from $t=\infty$ to $t\in[0,\infty)$.

Acknowledgments

The authors are grateful to the area editor, associate editor, and reviewers for their proposed improvements to the original manuscript. The authors also thank Jim Dai, Rami Atar, and Amy Ward for their helpful comments.

Endnotes

[1] The exception is the very special case in which μ equals the patience rate θ.

[2] The virtual waiting time, in the first order, stabilizes at the point at which input = output: $\lambda\bar{F}_a(\bar{w})=\mu\wedge\lambda$.

[3] This condition can be relaxed to $\lambda c_p|F_a(\bar{w}_p+yc_p)-F_a(\bar{w}_p+xc_p)|\leqslant H(1+(|x|\vee|y|)^H)|y-x|$ without affecting the results that follow.

[4] Although this constraint is not necessary, it simplifies the exposition of what follows. The service rate of 0 can be ruled out by imposing, instead, a condition on the cost parameters.

[5] Since the Hyperexponential distribution has a decreasing hazard rate, it indeed follows from Bassamboo and Randhawa (2010) that for sufficiently low abandonment costs, the optimal choice is to overload the queue.

[6] To be precise, we should write $\mathcal{J}^{Y,*}_{p,m}(x)$ to capture the possible dependence on the initial condition. The independence of x does follow, as is standard, from the verification arguments in Lemma 5.

References

Asmussen S (2003) *Applied Probability and Queues* (Springer-Verlag, New York).

Bassamboo A, Randhawa RS (2010) On the accuracy of fluid models for capacity sizing in queueing systems with impatient customers. *Oper. Res.* 58(5):1398–1413.

Braverman A, Dai JG (2017) Stein's method for steady-state diffusion approximations of $M/Ph/n+M$ systems. *Ann. Appl. Probab.* 27(1):550–581.

Doshi BT (1978) Optimal control of the service rate in an $M/G/1$ queueing system. *Adv. Appl. Probab.* 10(3):682–701.

Ghosh AP, Weerasinghe AP (2007) Optimal buffer size for a stochastic processing network in heavy traffic. *Queueing Systems* 55(3): 147–159.

Gurvich I (2014) Diffusion models and steady-state approximations for exponentially ergodic Markovian queues. *Ann. Appl. Probab.* 24(6):2527–2559.

Gurvich I, Huang J, Mandelbaum A (2014) Excursion-based universal approximations for the Erlang-A queue in steady-state. *Math. Oper. Res.* 39(2):325–373.

Harrison JM (1985) *Brownian Motion and Stochastic Flow Systems* (John Wiley & Sons, New York).

Harrison JM, Nguyen V (1993) Brownian models of multiclass queueing networks: Current status and open problems. *Queueing Systems* 13(1–3):5–40.

Harrison JM, Williams RJ (1987) Brownian models of open queueing networks with homogeneous customer populations. *Stochastics: Internat. J. Probab. Stochastic Processes* 22(2):77–115.

Jennings OB, Reed JE (2012) An overloaded multiclass FIFO queue with abandonments. *Oper. Res.* 60(5):1282–1295.

Kusuoka S, Tudor CA (2012) Stein's method for invariant measures of diffusions via Malliavin calculus. *Stochastic Processes Their Appl.* 122(4):1627–1651.

Mitchell B (1973) Optimal service-rate selection in an $M/G/1$ queue. *SIAM J. Appl. Math.* 24(1):19–35.

Pender J, Engblom S (2014) Approximations for the moments of nonstationary and state dependent birth-death queues. Working paper, Cornell University, Ithaca, NY.

Randhawa RS (2016) Optimality gap of asymptotically derived prescriptions in queueing systems. *Queueing Systems* 83(1):131–155.

Reed JE, Ward AR (2008) Approximating the $GI/GI/1+GI$ queue with a nonlinear drift diffusion: Hazard rate scaling in heavy traffic. *Math. Oper. Res.* 33(3):606–644.

Ward AR (2012) Asymptotic analysis of queueing systems with reneging: A survey of results for FIFO, single class models. *Surveys Oper. Res. Management Sci.* 17(1):1–14.

Ward AR, Glynn PW (2003) A diffusion approximation for a Markovian queue with reneging. *Queueing Systems* 43(1–2):103–128.

Ward AR, Glynn PW (2005) A diffusion approximation for a $GI/GI/1$ queue with balking or reneging. *Queueing Systems* 50(4):371–400.

Zeltyn S, Mandelbaum A (2005) Call centers with impatient customers: Many-server asymptotics of the $M/M/n+G$ queue. *Queueing Systems* 51(3–4):361–402.

Zhang T (1994) On the strong solutions of one-dimensional stochastic differential equations with reflecting boundary. *Stochastic Processes Their Appl.* 50(1):135–147.

Junfei Huang is an assistant professor in the Department of Decision Sciences and Managerial Economics at the Chinese University of Hong Kong. His research interests are in asymptotic analysis and optimal control of queueing systems and their applications in manufacturing and services.

Itai Gurvich is an associate professor at Cornell's School of Operations Research and Information Engineering and at Cornell Tech. His research focuses on the performance analysis and optimization of processing networks.