

Diffusion approximation error for queueing systems with general primitives

Anton Braverman

Kellogg School of Management, Northwestern University, Evanston, IL 60208,
anton.braverman@kellogg.northwestern.edu

Ziv Scully*

School of Operations Research and Information Engineering, Cornell University, Ithaca, NY 14853,
zivscully@cornell.edu

We investigate the steady-state diffusion-approximation error for continuous-time queueing systems with generally distributed primitives. Across four canonical systems—the $G/G/1$ and $G/M/\infty$ queues, the join-the-shortest-queue system, and a two-station tandem queue—a common picture emerges: the error decomposes into interior and boundary terms. The former are simpler to handle and can be bounded using only low-order moments of the system’s primitives—when the approximation error is measured using the Wasserstein distance, three moments suffice. The boundary terms are inherently more delicate: while crude bounds are easy to obtain, sharper (e.g., order optimal) bounds require deeper, model specific, insights.

Methodologically, we extend the generator comparison approach of Stein’s method to piecewise-deterministic Markov processes (PDMPs). The discontinuous nature of the PDMP at jump times necessitates using the basic adjoint relationship (BAR), instead of the infinitesimal generator, to characterize the stationary distribution. A second-order Taylor expansion of the BAR’s jump terms, coupled with a Palm-inversion step that converts event-averaged quantities into time averages, yields the candidate diffusion generator and a transparent interior/boundary error decomposition. In parallel, we show how the prelimit generator approach—working with the Poisson equation of the queueing system instead of the diffusion process—offers a promising avenue for bounding the challenging boundary terms.

Key words: Stein’s method; piece-wise deterministic diffusion approximation; steady-state; convergence rate

*Supported by NSF grant no. CMMI-2307008

1. Introduction

Diffusion approximations play a central role in the analysis of queueing systems. They depend on the first two moments of the system's primitives and can yield valuable insights into the performance of models that are otherwise intractable. These approximations are often justified through heavy-traffic limit theorems, which ensure that the original system converges to a diffusion limit as its parameters are taken to an appropriate asymptotic regime. Yet a key limitation of such limit theorems is that they provide no information about the accuracy of the approximation when the system operates outside this asymptotic regime. Over the past decade, this gap has been addressed by the generator approach of Stein's method ([Barbour 1988, 1990](#), [Götze 1991](#)), which offers a systematic way to bound approximation errors even in non-asymptotic settings.

A notable contribution of the generator approach has been in establishing the universality of diffusion approximations—that is, demonstrating the accuracy of the diffusion model simultaneously across multiple heavy-traffic regimes, such as quality-driven, quality- and efficiency-driven, and efficiency-driven regimes (see, e.g., [Ward and Glynn \(2003\)](#), [Gurvich et al. \(2014\)](#)). This is due to the diffusion model's dependence on the finite-parameter specification of the queueing system (in contrast to the diffusion limit, which depends on the limit of the parameters). In this paper, we explore a different dimension of universality or: for continuous-time queueing models with generally distributed primitives (e.g., general interarrival and service times), we investigate how the diffusion approximation error behaves across primitive distributions.

The assumption of exponentially distributed primitives is often adopted for analytical tractability; in such models, the queue-length process naturally admits a continuous-time Markov chain (CTMC) representation. Models with general primitives are important not only because they better reflect real-world variability, but also because a system's behavior can change qualitatively when exponential assumptions are relaxed. For example, generalized Jackson networks no longer have a product-form stationary distribution when interarrivals and service times are not exponentially distributed; for further examples see [Whitt \(1986\)](#), [Dai et al. \(2004\)](#), [Bassamboo and Randhawa \(2010\)](#). In such cases, maintaining a Markovian representation requires augmenting the state space with residual inter-event

times; for instance, the $G/G/1$ queue-length process includes both the residual interarrival time and the residual service time. Phase-type primitives, which can approximate any positive-valued distribution arbitrarily well, also admit a CTMC representation, but the enlarged state space significantly complicates the error analysis (Gurvich 2014, Braverman and Dai 2017).

In what follows, we introduce the terminology of a *general clock* to denote a residual inter-event time when the corresponding distribution is general (non-exponential). For instance, in the Markovian representation of the $G/G/1$ queue length, both the residual interarrival and service times are general clocks, whereas in the $G/M/\infty$ system where service times are exponentially distributed, only the residual interarrival time is a general clock.

Our main goal is to establish bounds on the distance between the steady-state distribution of the queueing system and its diffusion approximation; in this paper, we focus on the Wasserstein distance. We seek error bounds that (a) hold across a broad class of primitive distributions, (b) depend on those distributions only through low-order statistics (e.g., moments of the distribution), and (c) converge to zero at the optimal rate. The reassurance brought by such distributional universality is particularly valuable when calibrating model primitives from data: while accurately estimating the exact distribution is often infeasible, low-order statistics—such as the first three moments—can be estimated with much greater precision.

We analyze four canonical queueing systems: the $G/G/1$ queue, the $G/M/\infty$ queue, the join-the-shortest-queue system, and the tandem queue. Across these models, a common structure emerges: the diffusion-approximation error naturally decomposes into *interior* and *boundary* terms. Establishing properties (a)–(c) for the interior terms is relatively straightforward—we can bound them using only the first three moments of the primitive distributions, and we expect this to hold more broadly beyond the cases studied here. The boundary terms, by contrast, present a greater challenge. A useful analogy is a random walk reflected at the origin to remain on $[0, \infty)$: away from the boundary the dynamics match those of the unconstrained walk, while at the reflection point the behavior becomes markedly more intricate.

To highlight this added difficulty, we analyze the boundary terms in a representative case: the workload process of the $G/G/1$ queue. A simple bound can be obtained using Hölder's inequality, which ensures properties (a) and (b) but yields a suboptimal convergence rate; this technique extends to other models as well. Achieving a bound that also satisfies (c) is substantially harder. Despite considerable effort, we were only able to establish such a bound for a restricted class of distributions with nonincreasing hazard rates (see Theorem 2). Although these partial results do not rule out the existence of simple, universally applicable bounds, they underscore the significant challenge of attaining such generality.

Achieving our stated objective requires the development of new methodological tools, which form the second main contribution of this work. We extend the generator approach of Stein's method, which, to date, has only been applied to CTMCs and discrete-time systems, to models with generally distributed primitives. Such models naturally admit a piecewise-deterministic Markov process (PDMP) representation (Davis 1984). In contrast to Davis (1984), which formulates PDMPs via hazard rates, we adopt a residual-time representation that tracks the remaining time to the next event, or jump, of the process. Extending Stein's method to such PDMPs requires new tools. In particular, we replace the infinitesimal generator with the basic adjoint relationship (BAR)—the stationary balance equation for the process—because the PDMP does not have a well-defined infinitesimal generator at the boundaries where jumps occur.

Starting from the BAR, we use a Taylor expansion to bridge the stationary equation of the queueing model to the generator of its diffusion approximation, mirroring the approach in the CTMC setting. The novelty here lies in the additional jump terms in the BAR due to the general renewal processes—these terms can be naturally interpreted as Palm expectations. To extract the diffusion generator from these jump terms, we rely on the Palm inversion formula, which links Palm expectations (event-averaged quantities) to time-averaged expectations, because the event-based nature of these terms prevents a direct Taylor expansion. We emphasize that, although our analysis has a close connection to Palm calculus, all results are derived directly from first principles, without invoking the broader machinery of that theory.

Returning to the boundary-term difficulty noted earlier, one promising way to circumvent it is the prelimit generator approach (Braverman 2022). This method works with the Poisson equation of the queueing system itself, rather than that of the diffusion approximation. In effect, it reshuffles the problem’s difficulty: instead of bounding complex boundary terms, the challenge is transferred to obtaining Stein factor bounds—bounds on the second- and third-order derivatives of the Poisson equation’s solution.

For the $G/G/1$ workload process, this reshuffling works in our favor and directly addresses one of the most stubborn boundary terms: the expected equilibrium idle-period length. This term has been extensively studied (Siegmund 1979, Li and Ou 1995, Wolff and Wang 2003), and although Blanchet and Glynn (2006) provide an asymptotic expansion for it, the only “clean” upper bound we know of is due to Köllerström (1976). In contrast, while the Stein factor bounds remain technically involved, they lead to novel and explicit bounds on this quantity.

Finally, we note that extending the prelimit approach to the PDMP setting requires tools distinct from those used for our extension of the classical generator approach. We are able to carry out this extension in settings with a single general clock, but extending it to models with multiple clocks remains an open problem. Taken together, our results broaden the scope of both the generator and prelimit approaches beyond CTMCs and discrete-time models, and yield new insights into the quality of diffusion approximations of queueing models.

1.1. Literature review

The generator comparison approach of Stein’s method provides a principled framework for comparing stationary distributions of Markov processes. While Stein’s method originates with Stein (1972), the link to Markov processes was developed by Barbour (1988, 1990) and Götze (1991). Queueing theory has been a particularly fertile domain: early applications to birth–death processes appear in Brown and Xia (2001). Subsequent work analyzed more complex systems—see, for example, Gurvich et al. (2014), Stolyar (2015), Gurvich (2014)—which, although not framed explicitly in Stein’s terminology, employ the core generator-comparison ideas. The approach was later systematized and popularized

in queueing settings by Braverman et al. (2016), Braverman and Dai (2017) for diffusion approximations, and by Ying (2016, 2017), Gast (2017) for mean-field models. Since then, the generator approach within Stein's method has remained an active line of research in the queueing community.

A related line of work analyzes approximation errors for *discrete-time* queueing systems, both via the drift method (Eryilmaz and Srikant 2012, Maguluri and Srikant 2016, Maguluri et al. 2018, Hurtado Lange and Maguluri 2022) and via Stein's method (Feng and Shi 2018, Zhou and Shroff 2020, Hurtado-Lange and Maguluri 2022). Discrete time affords simplifications not available in our continuous-time setting; nevertheless, the approximation error in these works likewise decomposes into interior and boundary components. For example, one common boundary term is related to the unused service within a slot, and is commonly controlled by imposing a bound on the number of service completions per slot.

Another relevant line of work is the BAR approach (Miyazawa 2015, 2017, Braverman et al. 2017, 2024, Dai and Xu 2024, Dai et al. 2025), which develops heavy-traffic steady-state limit theorems for queueing systems. Its core device is an exponential test function, engineered so that all but an asymptotically negligible portion of the BAR's jump terms vanish, which facilitates establishing heavy-traffic limit theorems via the convergence of moment generating functions. In contrast, our test functions come from the Poisson equation, which enables us to establish upper bounds on the Wasserstein distance.

Our $G/G/1$ example has a long history. The seminal work of Kingman (1961, 1962) initiated an extensive literature on upper and lower bounds for the expected waiting time; see Daley et al. (1992), Wolff and Wang (2003) for surveys. Because the expected waiting time and expected workload are tightly linked (Asmussen 2003, Corollary X.3.5), bounds for one translate directly to bounds for the other.

While many classical bounds are asymptotically tight as $\rho \uparrow 1$, they often do not quantify the nonasymptotic gap to the true expectation. Notable exceptions include (Boon et al. 2023, Eq. (7)), who use transform methods to obtain moment formulas and an approximation to $\mathbb{E}W$ with an $O(1/\log(1-\rho))$ error bound, albeit with an implicit constant. By contrast, our result provides an $O(1)$ error bound—uniform in ρ as $\rho \rightarrow 1$ —with an explicit constant. Another approximation that is $o(1)$ accurate (error vanishing as $\rho \rightarrow 1$) is stated

following (Bertsimas and Gamarnik 2022, Theorem 3.3). A related line of research studies *extremal* queues (Chen and Whitt 2020, 2021, 2022a,b), identifying interarrival and service-time distributions for which waiting-time bounds are attained with equality.

Beyond the many bounds for $\mathbb{E}W$, there is a growing body of work applying Stein's method to single-server queues. Using equilibrium couplings, Gaunt and Walton (2020) obtain exponential approximation bounds for the $G/G/1$ waiting time—building on the approaches of Peköz and Röllin (2011), Ross (2011) and exploiting the representation of W as a geometric sum of i.i.d. increments. They also derive error bounds for the $M/G/1$ system via the generator approach. In line with our Theorem 1, their $G/G/1$ bound depends on the expected equilibrium idle period.

In Huang and Gurvich (2018), the generator comparison approach is applied to the workload of the $M/G/1+GI$ system (Poisson arrivals with a general patience-time distribution). The focus there is on diffusion-approximation error bounds that are universal across patience-time distributions and load regimes, together with bounds for arbitrary moments of the stationary workload.

Finally, Braverman (2022) gives a simple prelimit treatment of the $M/M/1$ queue length. At the process level, Besançon et al. (2020) establish rates of convergence to diffusion limits for the $M/M/1$ and $M/M/\infty$ systems. More recently, Barbour et al. (2023) develop Gaussian-process approximations and apply them to the $G/G/\infty$ system.

1.2. Notation

We write $f^{(k)}(\xi)$ to denote a generic k th-order Taylor expansion remainder term that may change from expression to expression. For a function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ and integer k , we let $\partial_{x_i}^k f(x) = \partial^k f(x) / \partial x_i^k$. We let $\text{Lip}(1)$ denote the set of Lipschitz-1 functions. For a counting process $\{N(t) : t \geq 0\}$ with event times $\{\tau_m\}_{m=1}^\infty$, we define

$$\int_0^t f(Z(t)) dN(t) = \sum_{m=1}^\infty f(Z(\tau_m)) 1(\tau_m \leq t),$$

and we let $\Delta f(Z(t-)) = f(Z(t)) - f(Z(t-))$. Given a random variable X we say that the expectation of $f(X)$ is well defined if $\mathbb{E}|f(X)| < \infty$. Similarly, we say that the expectation of $\int_0^t f(Z(t)) dN(t)$ is well defined if

$$\int_0^t \mathbb{E}|f(Z(t))| dN(t) < \infty.$$

2. The G/G/1 workload

The workload process of the single-server queue has only one general clock tracking arrivals to the system, making it the perfect introductory example. After defining the model we derive the BAR and, with its help, some useful identities about the system. We then extract a diffusion approximation generator from the BAR using Taylor expansion and the Palm inversion formula. We conclude by using Stein's method to bound the approximation error.

2.1. The model

Consider a single-server queueing system operating under any non-idling policy, where arrivals are driven by a renewal process. Let U and S be random variables having the interarrival and service time distributions, respectively, and assume that

$$\mathbb{E}U^3 < \infty \quad \text{and} \quad \mathbb{E}S^3 < \infty. \quad (1)$$

For simplicity, we assume that simultaneous events (arrivals/departures) do not occur with probability one. Set

$$\lambda = 1/\mathbb{E}U, \quad \mu = 1/\mathbb{E}S, \quad \rho = \lambda/\mu, \quad c_U^2 = \lambda^2 \text{Var}(U), \quad c_S^2 = \mu^2 \text{Var}(S).$$

Let $V(t)$ and $R_a(t)$ be the remaining workload in the system and residual interarrival time (time until the next arrival), respectively, at time $t \geq 0$.

We are interested in the heavy-traffic regime as $\rho \rightarrow 1$. To this end we define the scaling factor δ and the scaled workload $X(t)$ by

$$\delta = (1 - \rho) \quad \text{and} \quad X(t) = \delta V(t).$$

We also define the (scaled) workload process

$$\{Z(t) = (X(t), R_a(t)) : t \geq 0\}.$$

The workload is a right-continuous with left limits (RCLL) PDMP taking values on the state space

$$\mathbb{S} = \{(x, r) \in \mathbb{R}_+^2 : x \geq 0, r > 0\}.$$

Note that $(x, 0) \notin \mathbb{S}$ for $x \geq 0$: since the process is RCLL, whenever the residual interarrival time hits zero an arrival occurs instantaneously, and the process immediately jumps to a state with strictly positive residual time.

Let $A(t)$ denote the number of arrivals on $[0, t]$. If an arrival occurs at time t , define

$$\begin{aligned} U(t) &= \text{the next interarrival time,} \\ S(t) &= \text{the workload brought by the arriving customer.} \end{aligned} \tag{2}$$

If no arrival occurs at t , set $U(t) = S(t) = 0$. Conditional on an arrival at t , $U(t) \stackrel{d}{=} U$ and $S(t) \stackrel{d}{=} S$, and both are independent of the process history over $[0, t)$.

The workload process is a regenerative process, with regeneration happening at those instances when a customer arrives to an empty system. Going forward we assume that

$$\rho < 1 \text{ and } U \text{ is nonlattice,} \tag{3}$$

which ensures the existence of a limiting steady-state distribution ([Asmussen 2003](#), Corollary X.3.3). We let $Z = (X, R_a) \in \mathbb{S}$ have this distribution.

2.2. The BAR and some basic identities

Given $Z(0)$ and a function $f : \mathbb{S} \rightarrow \mathbb{R}$ such that $df(Z(t))/dt$ exists for almost all t , the fundamental theorem of calculus yields

$$f(Z(t)) - f(Z(0)) = \int_0^t (-\delta 1(X(s) > 0) \partial_x f(Z(s)) - \partial_r f(Z(s))) ds + \int_0^t \Delta f(Z(s)) dA(s).$$

The first term accounts for the deterministic decay of the workload and residual interarrival time, and the second term accounts for the changes to the workload process at arrival instances.

If $f : \mathbb{S} \rightarrow \mathbb{R}$ satisfies certain regularity conditions (defined precisely in the following lemma), dividing both sides by t , taking expectations in steady state, and using Fubini/Tonelli to interchange expectation and integration, we obtain the following basic adjoint relationship (BAR).

LEMMA 1. Initialize $Z(0) \sim Z$. If $\mathbb{E}|f(Z)| < \infty$ and, $Z(0)$ -almost surely, $df(Z(t))/dt$ exists Lebesgue-almost everywhere except at jump times, then, provided that all expectations are well defined,

$$0 = -\delta\mathbb{E}(1(X > 0)\partial_x f(Z)) - \mathbb{E}(\partial_{r_a} f(Z)) + \frac{1}{t}\mathbb{E} \int_0^t \Delta f(Z(t))dA(t). \quad (4)$$

Going forward we fix $t = 1$ (for an example of when we might use a different t , see Remark 5). The BAR grants access to the properties of Z . The following identities will be useful.

LEMMA 2. In steady state, the following identities hold:

$$\begin{aligned} \mathbb{E}A(1) &= \lambda, \\ \mathbb{P}(X > 0) &= \rho, \\ \mathbb{E}R_a^{m-1} &= \frac{\lambda}{m}\mathbb{E}U^m, \quad m \geq 1. \end{aligned}$$

Proof of Lemma 2 Let $f(z) = r_a \wedge M$ and note that $\partial_{r_a} f(z) = 1(r_a < M)$ is defined everywhere except when $r_a = M$. To ensure that $\mathbb{E}\partial_{r_a} f(Z)$ is well defined, we choose $M > 0$ such that R_a does not have a point mass at M (any random variable can have at most countably many point masses). Using the BAR (4) yields

$$\begin{aligned} \mathbb{P}(R_a < M) &= \mathbb{E} \int_0^1 ((R_a(t) \wedge M) - (R_a(t-) \wedge M))dA(t) \\ &= \mathbb{E} \int_0^1 (U(t) \wedge M)dA(t) = \mathbb{E}(U \wedge M)\mathbb{E}A(1), \end{aligned}$$

Taking $M \rightarrow \infty$ yields $\mathbb{E}A(1) = \lambda$. Omitting the algebra, the test function $f(z) = x \wedge M$ eventually yields $\mathbb{P}(X > 0) = \rho$; for an alternative proof see (Asmussen 2003, Section X.3). The remaining identity is obtained using $f(z) = r_a^m \wedge M$. $\square$

2.3. Deriving the diffusion generator

When $f(Z) = f(X, R_a)$ ignores the second coordinate and is a function of X alone, the BAR (4) reduces to

$$0 = -\delta\mathbb{E}(1(X > 0)f'(X)) + \mathbb{E} \int_0^1 (f(X(t-) + \delta S(t)) - f(X(t-)))dA(t). \quad (5)$$

Deriving the diffusion generator requires expanding the jump term on the right-hand side. In the special case of Poisson arrivals, the “PASTA” property (Theorem 3.3.1 of [Baccelli and Brémaud \(2003\)](#); see also their equation (1.8.8)) implies

$$\begin{aligned} & -\delta\mathbb{E}(1(X > 0)f'(X)) + \mathbb{E} \int_0^1 (f(X(t-) + \delta S(t)) - f(X(t-)))dA(t) \\ & = -\delta\mathbb{E}(1(X > 0)f'(X)) + \lambda\mathbb{E}(f(X + \delta S) - f(X)). \end{aligned}$$

The right-hand side coincides with the generator of the $M/G/1$ workload process; see [Huang and Gurvich \(2018\)](#) and [Gaunt and Walton \(2020\)](#) for analyses of the $M/G/1$ workload using the generator approach.

For general arrival processes, we consider the *compensated* workload

$$\tilde{X}(t) = X(t) - \delta\rho R_a(t) \quad \text{and} \quad \tilde{X} = X - \delta\rho R_a. \quad (6)$$

The defining property of the compensated workload is that at all arrival instances t ,

$$\mathbb{E}\Delta\tilde{X}(t-) = \delta(\mathbb{E}S(t) - \rho\mathbb{E}U(t)) = \delta(\mathbb{E}S - \rho\mathbb{E}U) = 0.$$

Specializing the BAR (4) to functions of the form $f(z) = f(x - \delta\rho r_a)$ yields

$$0 = \delta\mathbb{E}((\rho - 1(X > 0))f'(\tilde{X})) + \mathbb{E} \int_0^1 \Delta f(\tilde{X}(t-))dA(t). \quad (7)$$

With the compensated process, the first-order term in the jump-term expansion of (7) equals zero. We elaborate on why this matters at the end of Section 2.3.1 after proving the following key proposition, which expands (7) into a second-order differential operator plus an error term. Recall that $f^{(k)}(\xi)$ denotes a generic k th-order Taylor expansion remainder term.

PROPOSITION 1. *If $f \in C^2(\mathbb{R})$ with $\|f''\| < \infty$ and $f''(x)$ absolutely continuous, then, provided that all expectations are well defined,*

$$\delta\mathbb{E}(\rho - 1(X > 0))f'(\tilde{X}) = -\delta^2\mathbb{E}f'(X) + \delta^2f'(0) + \epsilon_0(f), \quad (8)$$

$$\mathbb{E} \int_0^1 \Delta f(\tilde{X}(t-))dA(t) = \frac{1}{2}\delta^2\lambda\mathbb{E}(S - \rho U)^2\mathbb{E}f''(X) + \epsilon_A(f) \quad (9)$$

where

$$\begin{aligned}\epsilon_0(f) &= \delta^3 \rho \mathbb{E}(R_a f''(\xi)) - \delta^2 \rho \mathbb{E}(1(X=0) R_a f''(\xi)), \\ \epsilon_A(f) &= \frac{1}{6} \delta^3 \mathbb{E} \int_0^1 (S(t) - \rho U(t))^3 f'''(\xi(t)) dA(t) \\ &\quad - \frac{1}{2} \delta^2 \lambda \mathbb{E}(S - \rho U)^2 \mathbb{E} \int_0^1 \int_0^{U(t)} (X(t+u) - X(t-)) f'''(\xi(t+u)) du dA(t).\end{aligned}$$

We remark that although the left-hand sides of (8) and (9) are in terms of $\tilde{X}(t)$, the right-hand sides (excluding the error terms) are in terms of the uncompensated X .

The terms $-\delta^2 \mathbb{E} f'(X)$ and $\delta^2 f'(0)$ in (8) capture the drift of the diffusion and its reflection at the boundary, while the term involving $\mathbb{E} f''(X)$ in (9) captures its variability. Combining these terms and noting that

$$\lambda \mathbb{E}(S - \rho U)^2 = \lambda \text{Var}(S - \rho U) = \lambda \text{Var}(S) + \lambda \rho^2 \text{Var}(U) = \rho \mathbb{E} S(c_U^2 + c_S^2),$$

we arrive at the diffusion generator

$$G_Y f(x) = -\delta^2 f'(x) + \frac{1}{2} \delta^2 \rho \mathbb{E} S(c_U^2 + c_S^2) f''(x) + \delta^2 f'(0), \quad x \geq 0.$$

This generator corresponds to the one-dimensional reflected Brownian motion (RBM) on the nonnegative real numbers (Harrison and Reiman 1981), whose stationary distribution is exponentially distributed with mean $\rho \mathbb{E} S(c_U^2 + c_S^2)/2$. We let Y be the random variable having this distribution.

We prove Proposition 1 in the following section. Readers wishing to skip the technical details can proceed directly to Section 2.4 for the bound on $d_W(X, Y)$.

2.3.1. Palm inversion and the proof of Proposition 1 A critical element in the proof of Proposition 1 is the following relationship between event-average expectations and time-average expectations. It is a special case of the Palm inversion formula (Baccelli and Brémaud 2003, Equation (1.2.25)). Note that the conditions of Lemma 3 are trivially satisfied when $f(x)$ is bounded.

LEMMA 3. Initialize $Z(0) \sim Z$ and fix $f: \mathbb{R} \rightarrow \mathbb{R}$ with

$$\mathbb{E} |f(X)| < \infty \quad \text{and} \quad \mathbb{E} \left| \int_0^{R_a(0)} f(X(t)) dt \right| < \infty.$$

Then

$$\mathbb{E}f(X) = \mathbb{E} \int_0^1 \int_0^{U(t)} f(X(t+u)) du dA(t). \quad (10)$$

Proof of Lemma 3 Let τ_m be the time of the m th arrival and observe that

$$\int_0^1 f(X(t)) dt = \int_0^{\tau_1} f(X(t)) dt + \sum_{m=1}^{A(1)} \int_{\tau_m}^{\tau_{m+1}} f(X(t)) dt - \int_1^{\tau_{A(1)+1}} f(X(t)) dt.$$

Since $\tau_{m+1} - \tau_m = U(\tau_m)$, it follows that

$$\begin{aligned} \sum_{m=1}^{A(1)} \int_{\tau_m}^{\tau_{m+1}} f(X(t)) dt &= \sum_{m=1}^{A(1)} \int_0^{U(\tau_m)} f(X(\tau_m + u)) du \\ &= \int_0^1 \int_0^{U(t)} f(X(t+u)) du dA(t). \end{aligned}$$

Furthermore, since $\tau_1 = R_a(0)$ and $\tau_{A(1)+1} = 1 + R_a(1)$,

$$\int_0^{\tau_1} f(X(t)) dt - \int_1^{\tau_{A(1)+1}} f(X(t)) dt = \int_0^{R_a(0)} f(X(t)) dt - \int_0^{R_a(1)} f(X(1+t)) dt.$$

Therefore,

$$\begin{aligned} \mathbb{E}f(X) &= \mathbb{E} \int_0^1 f(X(t)) dt \\ &= \mathbb{E} \int_0^1 \int_0^{U(t)} f(X(t+u)) du dA(t) + \mathbb{E} \left(\int_0^{R_a(0)} f(X(t)) dt - \int_0^{R_a(1)} f(X(1+t)) dt \right) \end{aligned}$$

and the result follows by noting that since (X, R_a) is stationary, the joint distribution of $(X(t), R_a(t))$ is the same as $(X(1+t), R_a(1+t))$, so the two integrals cancel in expectation.

□

We now use Lemma 3 to expand the BAR for the compensated workload.

Proof of Proposition 1 We first prove (8), which follows from some basic algebra:

$$\begin{aligned} &\delta \mathbb{E}(\rho - 1(X > 0)) f'(X - \delta \rho R_a) \\ &= -\delta^2 \mathbb{E} f'(X - \delta \rho R_a) + \delta \mathbb{E}(1(X = 0) f'(-\delta \rho R_a)) \\ &= -\delta^2 \mathbb{E}(f'(X) + \delta(-\rho R_a) f''(\xi)) + \delta \mathbb{E}(1(X = 0)(f'(0) - \delta(\rho R_a) f''(\xi))) \\ &= -\delta^2 \mathbb{E} f'(X) + \delta^2 f'(0) + \delta^3 \rho \mathbb{E}(R_a f''(\xi)) - \delta^2 \rho \mathbb{E}(1(X = 0) R_a f''(\xi)), \end{aligned}$$

where in the last equality we used $\mathbb{P}(X=0)=1-\rho$ (Lemma 2). Next we prove (9). Since $\mathbb{E}\Delta\tilde{X}(t-)=0$ at all arrival instants t , it follows that

$$\begin{aligned}\mathbb{E}\int_0^1 \Delta f(\tilde{X}(t-))dA(t) &= \frac{1}{2}\delta^2\mathbb{E}(S-\rho U)^2\mathbb{E}\int_0^1 f''(\tilde{X}(t-))dA(t) \\ &\quad + \frac{1}{6}\delta^3\mathbb{E}\int_0^1 (S(t)-\rho U(t))^3 f'''(\xi(t))dA(t).\end{aligned}\quad (11)$$

Noting that $\tilde{X}(t-)=X(t-)$ at jump times t , to conclude (9) it suffices to show that

$$\begin{aligned}\mathbb{E}\int_0^1 f''(X(t-))dA(t) \\ = \lambda\mathbb{E}f''(X) - \lambda\mathbb{E}\int_0^1 \int_0^{U(t)} (X(t+u)-X(t-))f'''(\xi(t+u))dudA(t),\end{aligned}$$

which follows from expanding the right-hand side of (10) in Lemma 3 as follows:

$$\begin{aligned}\mathbb{E}f''(X) &= \mathbb{E}\int_0^1 U(t)f''(X(t-))dA(t) \\ &\quad + \mathbb{E}\int_0^1 \int_0^{U(t)} (X(t+u)-X(t-))f'''(\xi(t+u))dudA(t).\end{aligned}$$

□

REMARK 1. The first-order term in the expansion of $\mathbb{E}\int_0^1 \Delta f(\tilde{X}(t-))dA(t)$ in (11) is zero because the compensated workload satisfies $\mathbb{E}\Delta\tilde{X}(t-)=0$ at jump times t . In contrast, the jump-term expansion in the BAR for the uncompensated workload (5) contains a first-order term that equals $\delta\mathbb{E}S\mathbb{E}\int_0^1 f'(X(t-))dA(t)$. Similar to the final step in the proof of Proposition 1, we can relate this term to $\delta\mathbb{E}S\mathbb{E}f'(X)$ by expanding the right-hand side of (10) in Lemma 3. Namely,

$$\begin{aligned}\delta\mathbb{E}S\mathbb{E}\int_0^1 f'(X(t-))dA(t) &= \lambda\delta\mathbb{E}S\mathbb{E}f'(X) \\ &\quad - \lambda\delta\mathbb{E}S\mathbb{E}\int_0^1 f''(X(t-))\int_0^{U(t)} (X(t+u)-X(t-))dudA(t) \\ &\quad - \frac{1}{2}\lambda\delta\mathbb{E}S\mathbb{E}\int_0^1 \int_0^{U(t)} (X(t+u)-X(t-))^2 f'''(\xi(t+u))dudA(t).\end{aligned}$$

Let us informally discuss the last two terms on the right-hand side. In the following sections we will work with $f(x)$ such that $\|f''\|, \|f'''\| = O(\delta^{-2})$. Note that

$$X(t+u)-X(t-)= (X(t-)+\delta S(t)-\delta u)^+-X(t-)$$

is of order δ , so that the third term is an $O(\delta)$ error term. The second term however is $O(1)$ and cannot be treated as error. We must use the Palm inversion formula a second time to separate

$$\mathbb{E} \int_0^1 f''(X(t-)) \int_0^{U(t)} (X(t+u) - X(t-)) du dA(t)$$

into a term involving $\mathbb{E}f''(X)$ (that we include in the diffusion generator), plus a third-order error term. However, this task requires computing $\mathbb{E} \int_0^{U(t)} (X(t+u) - X(t-)) du$. This is relatively simple (although messy) for the workload process, because no other jumps happen at times $u \in [t, t+U(t))$. However, if we were working with the $G/G/1$ queue-length process, computing $\mathbb{E} \int_0^{U(t)} (X(t+u) - X(t-)) du$ would require knowing the expected number of departures during $[t, t+U(t))$, but this quantity is not straightforward.

Thus, working with the compensated process accomplishes two things. First, it reduces the volume of calculations required because we only need to use the Palm inversion formula once. Second, it eliminates the need to compute $\mathbb{E} \int_0^{U(t)} (X(t+u) - X(t-)) du$, which is generally intractable.

2.4. The Poisson equation

In order to compare the workload to its diffusion approximation, we now introduce the Poisson equation. Given $\theta, \sigma^2 > 0$ and $h: \mathbb{R} \rightarrow \mathbb{R}$ with $E|h(Y)| < \infty$, consider the Poisson equation for the exponential distribution

$$\begin{aligned} -\theta f'_h(x) + \frac{1}{2}\sigma^2 f''_h(x) &= \mathbb{E}h(Y) - h(x), \quad x \geq 0, \\ f'_h(0) &= 0. \end{aligned} \tag{12}$$

The condition $f'_h(0) = 0$ is automatically satisfied (see [Ross \(2011\)](#)), but we include it here for emphasis. Setting $x = X$ and taking expectations allows us to compare $\mathbb{E}h(Y)$ to $\mathbb{E}h(X)$ by comparing the left-hand side to the expanded BAR in [Proposition 1](#). To do so however, we require bounds on the derivatives of $f_h(x)$. The following results are standard in the literature on Stein's method for the exponential distribution ([Peköz and Röllin 2011](#), [Ross 2011](#)).

One may verify that the solution satisfies

$$f'_h(x) = -e^{\frac{2\theta}{\sigma^2}x} \int_x^\infty \frac{2}{\sigma^2} (\mathbb{E}h(Y) - h(y)) e^{-\frac{2\theta}{\sigma^2}y} dy, \quad x \geq 0.$$

Since the integral representation is well defined for all $x \in \mathbb{R}$, we may extend $f'_h(x)$, $f''_h(x)$, and the Poisson equation (12) to the entire real line.

The solution to (12) is unique up to an additive constant, so without loss of generality we normalize by setting $f_h(0) = 0$. The following lemma is a rescaled version of (Peköz and Röllin 2011, Lemma 4.1).

LEMMA 4. *Let $f_h(x)$ be the solution to (12) with $f_h(0) = 0$. Then $f'_h(0) = 0$. Moreover, if $h(x)$ is Lipschitz, then $f'''_h(x)$ is absolutely continuous and*

$$\|f''_h\| \leq 1/\theta \quad \text{and} \quad \|f'''_h\| \leq 4/\sigma^2.$$

As a consequence, for all $x \in \mathbb{R}$, both $|f'_h(x)| \leq |x|/\theta$ and $|f_h(x)| \leq x^2/(2\theta)$ (obtained by integrating the bound on $f'''_h(x)$).

2.5. Bounding the error

In this section we prove the following error bound on $d_W(X, Y)$.

THEOREM 1. *Assume that (1) and (3) hold. Then*

$$d_W(X, Y) \leq |\epsilon_0(f_h)| + |\epsilon_A(f_h)|, \tag{13}$$

where $\epsilon_0(f_h)$ and $\epsilon_A(f_h)$ are as in Proposition 1. Furthermore,

$$\begin{aligned} |\epsilon_0(f_h)| &\leq \delta \rho \lambda \mathbb{E}U^2/2 + \rho \mathbb{E}(1(X=0)R_a), \\ |\epsilon_A(f_h)| &\leq \frac{2\delta \mathbb{E}|S - \rho U|^3}{3(\mathbb{E}S)^2(c_U^2 + c_S^2)} + 2\delta(\mathbb{E}S + \lambda \mathbb{E}U^2). \end{aligned}$$

The workload X is measured in units of time, and the bounds on $|\epsilon_0(f_h)|$ and $|\epsilon_A(f_h)|$ carry the same units. For example, $\delta \rho \lambda \mathbb{E}U^2/2$ combines the scale-free factors δ and ρ with λ (time⁻¹) and $\mathbb{E}U^2$ (time²), yielding time units.

With the exception of the boundary term $\mathbb{E}(1(X=0)R_a)$, the rest are of order $\delta = (1 - \rho)$ and depend only on the first three moments of U and S . Let us say a few words about the boundary term.

A crude bound. For a crude but simple bound on the boundary term we can use the fact that $\mathbb{P}(X=0) = (1-\rho)$ together with Hölder's inequality to get

$$\mathbb{E}(R_a 1(X=0)) \leq (\mathbb{P}(X=0))^{1/p} (\mathbb{E}(R_a^q))^{1/q} = (1-\rho)^{1/p} (\lambda \mathbb{E}U^{q+1}/q)^{1/q}, \quad (14)$$

for any $p, q \geq 1$ with $1/p + 1/q = 1$. Setting $p = q = 2$, we recover a bound depending only on λ and the third moment of U , but the rate of convergence is only $\sqrt{1-\rho}$. We can get a rate of convergence of $(1-\rho)^{1-\epsilon}$ for any $\epsilon > 0$ by taking q large enough, but this comes at the expense of assuming that higher moments of U exist.

A connection to the idle period. Note that $\mathbb{E}(R_a 1(X=0)) = (1-\rho)\mathbb{E}(R_a | X=0)$. Since the workload process is a regenerative process, with regeneration happening at those instances when a customer arrives to an empty system, the workload cycles between busy and idle periods; let $\bar{I}$ be the length of such an idle period. Thus, conditioned on $X=0$, the distribution of R_a is the same as the distribution of the remaining idle time (which equals the equilibrium distribution of $\bar{I}$), so that

$$\mathbb{E}(R_a 1(X=0)) = (1-\rho)\mathbb{E}(R_a | X=0) = (1-\rho) \frac{\mathbb{E}\bar{I}^2}{2\mathbb{E}\bar{I}}.$$

Significant effort has gone into understanding the quantity on the right-hand side, which is related to the first descending ladder height (Asmussen 2003, Proposition X.1.5). Li and Ou (1995) characterize the distribution function of $\bar{I}$ as a solution to a certain (complicated) non-linear integral equation. Wolff and Wang (2003) characterize the asymptotic behavior of the equilibrium distribution of $\bar{I}$ in a $G/M/1$ system as $\rho \rightarrow 1$. Blanchet and Glynn (2006) derive a full asymptotic expansion of $\mathbb{E}\bar{I}^2/2\mathbb{E}\bar{I}$ as $\rho \rightarrow 1$ (see also Siegmund (1979)). However, none of these papers offer an upper bound on $\mathbb{E}\bar{I}^2/2\mathbb{E}\bar{I}$ in terms of simple model primitives.

The only bounds that we are aware of are found in equations (16) and (17) of Köllerström (1976). Letting $r = (S - U)^-$, these are

$$\frac{\mathbb{E}\bar{I}^2}{2\mathbb{E}\bar{I}} \leq \frac{\mathbb{E}r^3/3 + a\mathbb{E}r^2}{2\mathbb{E}r\mathbb{P}(r \geq a)a} \quad \text{and} \quad \frac{\mathbb{E}\bar{I}^2}{2\mathbb{E}\bar{I}} \leq \frac{\mathbb{E}r^3/3 + a\mathbb{E}r^2}{2\mathbb{E}(r-a)^+a},$$

for any $a > 0$ such that $\mathbb{P}(r \geq a) > 0$ and $\mathbb{E}(r-a)^+ > 0$, respectively. These bounds are an improvement over the crude bound in (14), as they imply that the boundary term

$\mathbb{E}(R_a 1(X=0))$ decays at a rate of (at most) $1 - \rho$, while only requiring three finite moments of U and S . One downside of these bounds is the need to know how the distribution function of r behaves (to control the term in the denominator).

Idle period bounds via the prelimit approach. A key contribution of this paper is to demonstrate how the prelimit generator approach can be used to bound the distance between X and Y without explicitly handling complicated boundary terms such as $\mathbb{E}\bar{I}^2/(2\mathbb{E}\bar{I})$. Rather than analyzing these boundary terms directly, the difficulty is transferred to bounding Stein factors.

The next theorem illustrates the type of bounds obtained under this approach. The main technical challenge lies in establishing universal (across distributions) bounds for the third derivative of the Poisson equation solution. Although an explicit expression is available, obtaining the bounds is delicate and requires additional structure on the interarrival distribution. One tractable family consists of interarrival distributions with a nonincreasing hazard rate, as assumed in our theorem; these cover certain heavy-tailed interarrival distributions. Another admissible family includes distributions whose hazard rate is uniformly bounded away from zero. In that case, however, $\mathbb{E}\bar{I}^2/(2\mathbb{E}\bar{I})$ can be bounded directly using the hazard-rate lower bound, and the prelimit-based bounds are strictly weaker. For this reason, we do not include that case in the theorem, although we show how to handle it in the proof.

In summary, while our machinery extends beyond these two specific distributional families, it relies essentially on exploiting structural properties of U in order to control the third-order Stein factors. After developing the prelimit approach for models with a single general clock in Section 3, we prove the following theorem in Section 4.

THEOREM 2. *Assume that (1) and (3) hold, and let Y_2 be exponentially distributed with mean $\lambda\mathbb{E}(S-U)^2/2$. Further assume that U has a nonincreasing hazard rate that is bounded from above by $\bar{\eta} < \infty$. Then for $h(x) = x$,*

$$\begin{aligned} |\mathbb{E}h(Y_2) - \mathbb{E}h(X)| &\leq \delta(K_1 + K_2 + K_3 + K_4 + K_5), \\ K_1 &= (\lambda\mathbb{E}U^2/2 + \mathbb{E}S), \\ K_2 &= \rho\mathbb{E}U^2/2 + \delta(1/\mathbb{E}Y_2)\lambda(\mathbb{E}U^2/2)(\mathbb{E}S^2 + (\mathbb{E}S)^2)(\delta + 2\bar{\eta}\delta\mathbb{E}\bar{B}), \end{aligned}$$

$$\begin{aligned} K_3 &= ((1/\mathbb{E}Y_2)(\rho\mathbb{E}U^2 + 5\lambda\mathbb{E}U^3/3) + \lambda\mathbb{E}U^3/3)(\delta + (\lambda + \bar{\eta})\delta\mathbb{E}\bar{B}) \\ K_4 &= \lambda\mathbb{E}S^3(3 + (\mathbb{E}U + \mathbb{E}\bar{I})\lambda(1 + \rho + \lambda^2\mathbb{E}U^2))(1/\mathbb{E}Y_2)\bar{\eta}\delta\mathbb{E}\bar{B}, \\ K_5 &= \mathbb{E}(S^2 - \rho U^2 + \lambda\mathbb{E}U^3/3)(3 + (\mathbb{E}U + \mathbb{E}\bar{I})\lambda(1 + \rho + \lambda^2\mathbb{E}U^2))(1/\mathbb{E}Y_2)\bar{\eta}\delta\mathbb{E}\bar{B}. \end{aligned}$$

Furthermore, $\mathbb{E}\bar{I} = \delta\mathbb{E}\bar{B}/\rho$ and

$$\mathbb{E}\bar{B} \leq \min \left\{ \frac{\rho \text{Var}(S - U)}{2(1 - \rho)\mathbb{E}(S - U)^+}, \right. \\ \left. 0.9\rho \frac{\sqrt{\text{Var}(U - S)}}{1 - \rho} \exp \left(5.4 \frac{\mathbb{E}|U - S|^3}{(\text{Var}(U - S))^{3/2}} + 0.8 \frac{\mathbb{E}(U - S)}{\sqrt{\text{Var}(U - S)}} \right) \right\}.$$

Theorem 2 can be strengthened to hold for Lipschitz $h(x)$ with Lipschitz first and second derivatives, but we restrict ourselves to $h(x) = x$ for two reasons. First, it allows us to focus on cleanly showcasing the machinery of the prelimit approach without the algebraic clutter involved with more general $h(x)$.

Second, and this is special to the $G/G/1$ system, the bound in Theorem 2 allows us to exploit a classic identity involving $\mathbb{E}\bar{I}^2/2\mathbb{E}\bar{I}$ and $\mathbb{E}Y_2 - \mathbb{E}X$. Namely, letting W denote the steady-state customer waiting time, it follows that

$$\begin{aligned} \frac{\mathbb{E}\bar{I}^2}{2\mathbb{E}\bar{I}} &= \frac{\lambda\mathbb{E}(S - U)^2}{2(1 - \rho)} - \mathbb{E}W = \frac{\lambda\mathbb{E}(S - U)^2}{2(1 - \rho)} - \frac{1}{\rho}\mathbb{E}V + \frac{\mathbb{E}S^2}{2\mathbb{E}S} \\ &\leq \frac{1}{1 - \rho}(\mathbb{E}Y_2 - \mathbb{E}X) + \frac{\mathbb{E}S^2}{2\mathbb{E}S}. \end{aligned} \quad (15)$$

The first two equalities are well-known identities ((Asmussen 2003, Equation X.2.5) and (Asmussen 2003, Corollary X.3.5)), whereas the inequality uses $\rho < 1$ and $X = (1 - \rho)V$. As a result, Theorem 2 translates into a bound on $\mathbb{E}\bar{I}^2/2\mathbb{E}\bar{I}$, and can then be applied to the bound in Theorem 1.

Proof of Theorem 1 Given $h \in \text{Lip}(1)$, consider the Poisson equation (12) with $\theta = \delta^2$ and $\sigma^2 = \delta^2\rho\mathbb{E}S(c_U^2 + c_S^2)$. Assuming that the BAR (7) holds for $f_h(\tilde{X})$, we set $x = X$ and take expected values to arrive at

$$\begin{aligned} \mathbb{E}h(Y) - \mathbb{E}h(X) &= \mathbb{E}G_Y f_h(X) \\ &= \mathbb{E}G_Y f_h(X) - \delta\mathbb{E}((\rho - 1(X > 0))f'_h(\tilde{X})) - \mathbb{E} \int_0^1 \Delta f_h(\tilde{X}(t-))dA(t). \end{aligned}$$

Taking the supremum over all $h \in \text{Lip}(1)$ and using Proposition 1 yields

$$d_W(X, Y) \leq |\epsilon_0(f_h)| + |\epsilon_A(f_h)|.$$

We now bound both $|\epsilon_0(f_h)|$ and $|\epsilon_A(f_h)|$, and then show that the BAR (7) holds for $f_h(\tilde{X})$.

Recalling that

$$\epsilon_0(f_h) = \delta^3 \rho \mathbb{E}(R_a f_h''(\xi)) - \delta^2 \rho \mathbb{E}(1(X=0) R_a f_h''(\xi)),$$

the bound on $|\epsilon_0(f_h)|$ follows from $\|f_h''\| \leq \delta^{-2}$ (Lemma 4) and $\mathbb{E}R_a = \lambda \mathbb{E}U^2/2$ (Lemma 2).

Similarly, recall that

$$\begin{aligned} \epsilon_A(f_h) &= \frac{1}{6} \delta^3 \mathbb{E} \int_0^1 (S(t) - \rho U(t))^3 f_h'''(\xi(t)) dA(t) \\ &\quad - \frac{1}{2} \delta^2 \lambda \mathbb{E}(S - \rho U)^2 \mathbb{E} \int_0^1 \int_0^{U(t)} (X(t+u) - X(t-)) f_h'''(\xi(t+u)) du dA(t). \end{aligned}$$

The first term is handled using the Stein factor bound $\|f_h'''\| \leq 4/\sigma^2$, while the second term requires controlling increments of X over an interarrival time. Namely,

$$\frac{1}{6} \delta^3 \mathbb{E} \left| \int_0^1 (S(t) - \rho U(t))^3 f_h'''(\xi(t)) dA(t) \right| \leq \frac{4\delta^3 \mathbb{E}|S - \rho U|^3}{6\delta^2 \rho \mathbb{E}S(c_U^2 + c_S^2)} \mathbb{E}A(1) = \frac{2\delta \mathbb{E}|S - \rho U|^3}{3(\mathbb{E}S)^2(c_U^2 + c_S^2)},$$

where the equality is true because $\mathbb{E}A(1) = \lambda$, and

$$\begin{aligned} &\frac{1}{2} \delta^2 \lambda \mathbb{E}(S - \rho U)^2 \mathbb{E} \left| \int_0^1 \int_0^{U(t)} (X(t+u) - X(t-)) f_h'''(\xi(t+u)) du dA(t) \right| \\ &\leq 2\mathbb{E} \int_0^1 \int_0^{U(t)} |X(t+u) - X(t-)| du dA(t) \\ &= 2\mathbb{E} \int_0^1 \int_0^{U(t)} |(X(t-) + \delta S(t) - \delta u)^+ - X(t-)| du dA(t) \\ &\leq 2\mathbb{E} \int_0^1 \int_0^{U(t)} \delta(S(t) + U(t)) du dA(t) \\ &= 2\delta \mathbb{E}(US + U^2) \mathbb{E}A(1) = 2\delta(\mathbb{E}S + \lambda \mathbb{E}U^2), \end{aligned}$$

where the last inequality follows from $|(x+y)^+ - x^+| \leq y^+$.

It remains to show that (7) holds with $f(\tilde{X}) = f_h(\tilde{X})$. By assumption (1), U and S have finite third moments. Since $\mathbb{E}R_a = \lambda \mathbb{E}U^2/2 < \infty$ (Lemma 2) and $\mathbb{E}X^2 < \infty$ (Asmussen 2003,

Theorems X.3.4 and X.2.1), the fact that $|f_h(x)| \leq x^2/(2\theta)$ (Lemma 4) yields $\mathbb{E}|f_h(\tilde{X})| < \infty$; the fact that $\mathbb{E}|f'_h(\tilde{X})| < \infty$ is argued similarly. To show that the jump term in (7) is well defined, note that

$$\begin{aligned} \mathbb{E} \left| \int_0^1 \Delta f(\tilde{X}(t-)) dA(t) \right| &= \mathbb{E} \left| \int_0^1 \int_0^{\delta S(t)} f'_h(X(t-) + u) du dA(t) \right| \\ &\leq \mathbb{E} \left| \int_0^1 \int_0^{\delta S(t)} \delta^{-2} |X(t-) + u| du dA(t) \right| \\ &\leq \mathbb{E} \int_0^1 \delta^{-1} S(t) (X(t-) + S(t)) dA(t) \\ &= \delta^{-1} (\mathbb{E}S) \mathbb{E} \int_0^1 X(t-) dA(t) + \delta^{-1} \mathbb{E}S^2 \mathbb{E}A(1) \end{aligned}$$

The second term on the right-hand side is finite because $\mathbb{E}A(1) = \lambda$ (Lemma 2). Letting τ_m be the time of the m th jump, the first term is bounded by

$$\mathbb{E} \int_0^1 X(t-) dA(t) \leq \mathbb{E}(A(1) \sup_{0 \leq t \leq 1} X(t)) \leq \sqrt{\mathbb{E}(A(1))^2} \sqrt{\mathbb{E} \left(X(0) + \delta \sum_{m=1}^{A(1)} S(\tau_m) \right)^2}.$$

The right-hand side is finite because $\mathbb{E}X^2 < \infty$, $\mathbb{E}A^2(1) < \infty$ (since $\mathbb{E}U^2 < \infty$), and that $\mathbb{E}(\sum_{m=1}^{A(1)} S(\tau_m))^2 < \infty$ (since $A(1)$ is independent of $S(\tau_m)$). $\square$

3. The G/G/1 workload: the prelimit approach

This section develops the prelimit approach for the $G/G/1$ workload and uses it to prove Theorem 2. The approach generalizes naturally to other models with a single general clock, but extending it to models with multiple general clocks remains an open problem. We highlight the added difficulty at the end of this section, and leave it as an open problem for future work.

In addition to the notation introduced in Section 2.1, let $G(x) = \mathbb{P}(U \leq x)$ be the inter-arrival distribution function and assume for simplicity that $\mathbb{P}(U > 0) = 1$. Let

$$\begin{aligned} \bar{B} &= \text{the duration of a busy period initialized by an arrival to an empty system, and} \\ \bar{I} &= \text{the duration of the subsequent idle period.} \end{aligned} \tag{16}$$

We let B_0 be the length of the initial busy period, with the convention that $B_0 = 0$ if $X(0) = 0$, and let B_n , $n \geq 1$, be the lengths of the subsequent busy periods, which are i.i.d. $\bar{B}$. We also let I_n , $n \geq 0$, be the duration of the idle period following B_n , and note that I_n , $n \geq 1$, are i.i.d. $\bar{I}$.

When $\rho < 1$, (Asmussen 2003, Propositions X.1.3 and X.3.1) say that

$$\mathbb{E}\bar{I} = \frac{1-\rho}{\rho} \mathbb{E}\bar{B} < \infty. \quad (17)$$

Furthermore, note that

$$\mathbb{E}\bar{B} = \lim_{\epsilon \downarrow 0} \mathbb{E}_{0,\epsilon} B_0 = \mathbb{E}(\mathbb{E}_{\delta S, U} B_0),$$

where $\mathbb{E}_{(x,r_a)}(\cdot) = \mathbb{E}(\cdot | Z(0) = (x, r_a))$ and the outer expectation on the right-hand side is with respect to the distributions of U and S .

Recall that R_a is the steady-state residual interarrival time. It is a standard result in renewal theory that

$$\mathbb{P}(R_a \leq x) = \lambda \int_0^x \mathbb{P}(U > t) dt, \quad x \geq 0. \quad (18)$$

At times, we will need to consider expected values with respect to some variables but not others. For example, given random variables T_n , $n \geq 1$, and some subset of them $T_{n_1}, \dots, T_{n_k}$, we write

$$\mathbb{E}^{T_{n_1}, \dots, T_{n_k}}(f(T_1, \dots, T_n))$$

to denote an expectation with respect $T_{n_1}, \dots, T_{n_k}$ only. We still write $\mathbb{E}(\cdot)$ to denote the expectation over all random quantities inside the parentheses.

Lastly, we argue that the expected time to regeneration is finite under (3) given any initial condition $(x, r_a) \in \mathbb{S}$. Namely, we show that

$$\mathbb{E}_{x,r_a} B_0 \leq \mathbb{E}(\mathbb{E}_{x+\delta S, U} B_0) < \infty. \quad (19)$$

The first inequality is true because the busy period is made longer if an arrival happens immediately. For the second inequality, let us treat the (unscaled) initial work x/δ as low-priority work and all other work as high-priority work. The low-priority work is cleared

when there is no high priority work; i.e., during the idle periods of a system seeing only the high-priority work. Thus, the initial low-priority workload is cleared when the cumulative time spent without high-priority work exceeds x/δ . If N_x is the number of idle periods required for this, then $\mathbb{E}N_x < \infty$ (Gut 1974), and by Wald's identity, the expected duration until the N_x th idle period ends equals $\mathbb{E}N_x\mathbb{E}(\bar{B} + \bar{I})$, which is finite, implying (19).

3.1. A roadmap

We begin with a roadmap outlining the technical differences of the prelimit approach. As the name suggests, the prelimit approach starts from the Poisson equation of the workload instead of the limiting exponential distribution. To start, let

$$G_Z f(z) = \lim_{t \rightarrow 0} \frac{\mathbb{E}_z f(Z(t)) - f(z)}{t}, \quad z \in \mathbb{S},$$

be the workload generator. Since $Z(t) = (X(0) - \delta t, R_a(0) - t)$ for $t < R_a(0)$,

$$G_Z f(z) = -\delta 1(x > 0) \partial_x f(z) - \partial_{r_a} f(z),$$

provided that the derivatives on the right-hand side exist. Next, given a test function $h: \mathbb{R} \rightarrow \mathbb{R}$, we define

$$F_h(z) = \int_0^\infty (\mathbb{E}_z h(X(t)) - \mathbb{E}h(X)) dt, \quad z \in \mathbb{S}.$$

Provided that $F_h(z)$ is well defined, i.e.,

$$\int_0^\infty |\mathbb{E}_z h(X(t)) - \mathbb{E}h(X)| dt < \infty, \quad z \in \mathbb{S}, \quad (20)$$

and that its first-order partial derivatives exist, it is possible to show that $F_h(z)$ solves the Poisson equation

$$G_Z F_h(z) = -\delta 1(x > 0) \partial_x F_h(z) - \partial_{r_a} F_h(z) = \mathbb{E}h(X) - h(x), \quad z = (x, r_a) \in \mathbb{S}.$$

Since the left-hand side depends on both x and r_a , it is not clear how to extract the diffusion generator (which should approximate only X) via standard Taylor expansion. Furthermore, since the right-hand side does not depend on r_a , it appears that we have a degree of freedom in the choice of r_a .

While other approaches may exist, we find it fruitful to set $r_a = R_a$ and take expectations, which, as we later show, results in

$$-\delta \partial_x \mathbb{E} F_h(x, R_a) + \lambda \mathbb{E}(F_h(x + \delta S, U) - F_h(x, U)) = \mathbb{E} h(X) - h(x), \quad x \geq 0. \quad (21)$$

Before expanding $F_h(x + \delta S, U) - F_h(x, U)$ to derive the diffusion generator, there remains the step of relating $\mathbb{E} F_h(\cdot, R_a)$ and $\mathbb{E} F_h(\cdot, U)$, which is not needed when all clocks are exponentially distributed because in that case $U \stackrel{d}{=} R_a$. This additional step becomes the key technical obstacle when attempting to apply the prelimit approach to models with two or more general clocks. We now formalize our roadmap.

3.2. The workload Poisson equation.

In this section we prove that (21) holds for all $h \in \text{Lip}(1)$, which paves the way for the generator extraction and comparison step performed in Section 3.3. We begin with a discussion of how to verify (20).

As one option, Dai and Meyn (1995) show that when $\mathbb{E} U^{p+1} < \infty$ and $\mathbb{E} S^{p+1} < \infty$, then $|\mathbb{E}_z Q(t) - \mathbb{E} Q|$ decays at a rate of $1/t^{p-1}$, where $Q(t)$ and Q is the customer count at time t and in steady state, respectively. Their result holds for queueing-network models beyond the $G/G/1$ system, but we wish to avoid using their complex machinery (in particular we would need to convert their bound in $|\mathbb{E}_z Q(t) - \mathbb{E} Q|$ into a bound on $|\mathbb{E}_z X(t) - \mathbb{E} X|$).

A second option is to verify (20) is by noticing that

$$\int_0^\infty |\mathbb{E}_z h(X(t)) - \mathbb{E} h(X)| dt \leq \int_0^\infty \mathbb{E} (|\mathbb{E}_z h(X(t)) - \mathbb{E}_Z h(X(t))|) dt,$$

where the outer expectation is taken with respect to the stationary distribution of Z . To bound the right-hand side we would need to couple the workload process starting from z with one starting from Z , and bound the expected coupling time (in terms of z and Z). Constructing such a coupling is complicated by the fact that z and Z may differ both in the initial workload and residual interarrival time.

In this paper we illustrate a third option that bypasses the need to verify (20) directly. It involves using the truncated

$$F_h^M(z) = \int_0^M (\mathbb{E}_z h(X(t)) - \mathbb{E} h(X)) dt, \quad z \in \mathbb{S},$$

and may be useful in other models where verifying (20) directly is challenging. Our starting point is the following proposition, which is proved in Appendix A.

PROPOSITION 2. *For any $h \in \text{Lip}(1)$ and almost all $M > 0$,*

$$-\delta \mathbb{E} \partial_x F_h^M(x, R_a) + \lambda \mathbb{E}(F_h^M(x + \delta S, U) - F_h^M(x, U)) = \mathbb{E}(\mathbb{E}_{x, R_a} h(X(M)) - h(x)). \quad (22)$$

We wish to take $M \rightarrow \infty$ in (22) and recover (21), but since we do not assume that $F_h(z)$ is well defined, we first need to specify what we mean by both $\partial_x \mathbb{E} F_h(x, R_a)$ and $\mathbb{E}(F_h(x + \delta S, U) - F_h(x, U))$.

We first define

$$F_h(x + \epsilon, r_a) - F_h(x, r_a) = \int_0^\infty (\mathbb{E}_{x+\epsilon, r_a} h(X(t)) - \mathbb{E}_{x, r_a} h(X(t))) dt, \quad x \geq 0, \epsilon, r_a > 0. \quad (23)$$

To argue that the right-hand side is well defined, we now introduce a simple synchronous coupling of the workload process that differs only in the initial workload and not in the initial residual time. This coupling also plays a central role in Section 4.

Given $\epsilon > 0$, let $\{Z^{(\epsilon)}(t) = (X^{(\epsilon)}(t), R_a(t)) : t \geq 0\}$ be a coupling of $\{Z(t) = (X(t), R_a(t)) : t \geq 0\}$ with initial condition $X^{(\epsilon)}(0) = X(0) + \epsilon$. Both systems share the same arrival process, and the service time of each arriving customer is identical in both systems. Similar to B_0 , we define $B_0^{(\epsilon)} = \inf\{t \geq 0 : X^{(\epsilon)}(t) = 0\}$. It follows that for every sample path,

$$\begin{aligned} \frac{\partial}{\partial t} (X^{(\epsilon)}(t) - X(t)) &= -\delta 1(X(t) = 0, t \in [0, B_0^{(\epsilon)}]), \\ Z^{(\epsilon)}(t) &= Z(t) \text{ for } t > B_0^{(\epsilon)}. \end{aligned} \quad (24)$$

We adopt the convention that $\mathbb{E}_{x, r_a}(\cdot)$ is the expected value conditional on $Z(0) = (x, r_a)$, even if the quantity inside the parentheses is a function of $Z^{(\epsilon)}(t)$.

Returning to (23), our synchronous coupling yields

$$\int_0^\infty (\mathbb{E}_{x+\epsilon, r_a} h(X(t)) - \mathbb{E}_{x, r_a} h(X(t))) dt = \mathbb{E}_{x, r_a} \int_0^{B_0^{(\epsilon)}} (h(X^{(\epsilon)}(t)) - h(X(t))) dt. \quad (25)$$

The right-hand side is well defined because $|h(X^{(\epsilon)}(t)) - h(X(t))| \leq \epsilon \|h'\|$ and $\mathbb{E}_{x, r_a} B_0^{(\epsilon)} = \mathbb{E}_{x+\epsilon, r_a} B_0 < \infty$ by (19). A similar line of reasoning yields the following two lemmas. The detailed proofs are found in Appendix B.

LEMMA 5. Let $T \geq 0$ be any random variable and define $\partial_x \mathbb{E}F_h(x, T)$ by

$$\partial_x \mathbb{E}F_h(x, T) = \lim_{\epsilon \rightarrow 0} \frac{1}{\epsilon} \mathbb{E}(F_h(x + \epsilon, T) - F_h(x, T)), \quad x \geq 0, \quad (26)$$

with the convention that $\partial_x \mathbb{E}F_h(x, T) = \partial_x F_h(x, r_a)$ when $T = r_a$ is deterministic. Then for any $h \in \text{Lip}(1)$,

$$\partial_x \mathbb{E}F_h(x, T) = \mathbb{E} \left(\mathbb{E}_{x, T} \int_0^{B_0} h'(X(t)) dt \right), \quad x \geq 0, \quad (27)$$

where the outer expectation is with respect to T . In the special case that $T = r_a$ is deterministic, (27) yields an expression for $\partial_x F_h(x, r_a)$. In particular, this implies that for any random $T \geq 0$,

$$\partial_x \mathbb{E}F_h(x, T) = \mathbb{E} \partial_x F_h(x, T), \quad x \geq 0. \quad (28)$$

LEMMA 6. For any $h \in \text{Lip}(1)$ and $x \geq 0$,

$$\lim_{M \rightarrow \infty} \mathbb{E}(\mathbb{E}_{x, R_a} h(X(M))) = \mathbb{E}h(X), \quad (29)$$

$$\lim_{M \rightarrow \infty} \mathbb{E} \partial_x F_h^M(x, R_a) = \mathbb{E} \partial_x F_h(x, R_a), \quad (30)$$

$$\lim_{M \rightarrow \infty} \mathbb{E}(F_h^M(x + \delta S, U) - F_h^M(x, U)) = \mathbb{E}(F_h(x + \delta S, U) - F_h(x, U)). \quad (31)$$

Starting from (22) of Proposition 2, we apply Lemma 6 to take $M \rightarrow \infty$ there. Since $\partial_x \mathbb{E}F_h(x, R_a) = \mathbb{E} \partial_x F_h(x, R_a)$ and $\partial_x \mathbb{E}F_h(0, R_a) = 0$ by Lemma 5 and (27), respectively, we arrive at

$$-\delta \partial_x \mathbb{E}F_h(x, R_a) + \lambda \mathbb{E}(F_h(x + \delta S, U) - F_h(x, U)) = \mathbb{E}h(X) - h(x), \quad x \geq 0. \quad (32)$$

In the following section, we replace $\mathbb{E}(F_h(x + \delta S, U) - F_h(x, U))$ by a term where the expectation is taken over R_a instead of U . We then perform a Taylor expansion to compare the left-hand side with the diffusion generator.

3.3. Generator expansion and comparison

Let S' be an independent copy of S and introduce the random variable

$$J(x, r_a) = -(x \wedge \delta r_a) + \delta S', \quad (x, r_a) \in \mathbb{S}.$$

Conditioned on $Z(0) = (x, r_a)$, J represents the jump in workload between time $t = 0$ and $t = r_a$ (the first arrival time). We present the following lemma, which is proved in Appendix C.

LEMMA 7. *For any $(x, r_a) \in \mathbb{S}$, $s \geq 0$, and $h \in \text{Lip}(1)$,*

$$F_h(x + \delta s, r_a) - F_h(x, r_a) = \mathbb{E}(F_h(x + \delta s + J(x, r_a), U) - F_h(x + J(x, r_a), U)) + \epsilon(x, r, s).$$

The U on the right-hand side is independent of S' (and thus $J(x, r_a)$) and

$$\begin{aligned} \epsilon(x, r_a, s) = & \mathbb{E}^{S'} \left(\int_{-x \wedge (\delta r_a)}^{-(x + \delta s) \wedge (\delta r_a)} \partial_x \mathbb{E}^U F_h(x + \delta s + v + \delta S', U) dv \right) \\ & + \int_0^{r_a} \left(h((x + \delta s - \delta t)^+) - h((x - \delta t)^+) \right) dt, \end{aligned}$$

where we recall that $\mathbb{E}^U(\cdot)$ and $\mathbb{E}^{S'}(\cdot)$ denote expectations with respect to U only and S' only, respectively.

To simplify notation, we define

$$\bar{F}'_h(x) = \partial_x \mathbb{E} F_h(x, R_a). \quad (33)$$

Replacing x by $x + J(x, R_a)$ in the Poisson equation (32) and taking expectations yields

$$\begin{aligned} & \mathbb{E}h(X) - \mathbb{E}h(x + J(x, R_a)) \\ = & -\delta \mathbb{E} \bar{F}'_h(x + J(x, R_a)) + \lambda \mathbb{E}(F_h(x + \delta S, R_a) - F_h(x, R_a)) - \mathbb{E}(\epsilon(x, R_a, S)). \end{aligned} \quad (34)$$

We emphasize that $\mathbb{E} \bar{F}'_h(x + J(x, R_a))$ is actually $\mathbb{E}^{R_a} \partial_x \mathbb{E}^{R'_a} F_h(x + J(x, R_a), R'_a)$, where R_a and R'_a are independent copies.

The following lemma is the result of expanding the right-hand side of (34) and taking expected values with respect to the steady-state of the approximating diffusion. It is proved in Appendix C.1.

LEMMA 8. Let Y_2 be exponentially distributed with mean $\lambda\mathbb{E}(U - S)^2/2$. Fix $h \in \text{Lip}(1)$ and assume that $\bar{F}_h''(x)$ and $\bar{F}_h'''(x)$ exist for all $x \geq 0$, and that $\mathbb{E}|F_h'(Y_2)|, \mathbb{E}|F_h''(Y_2)| < \infty$. Then

$$\begin{aligned} \mathbb{E}h(Y_2) - \mathbb{E}h(X) &= \mathbb{E}(h(Y_2) - h(Y_2 + J(Y_2, R_a))) + \mathbb{E}(\epsilon(Y_2, R_a, S)) \\ &\quad + \delta\mathbb{E}\left(1(\delta R_a \geq Y_2)(\bar{F}_h'(\delta S) - \bar{F}_h'(Y_2) - \delta(S - R_a)\bar{F}_h''(Y_2))\right) \\ &\quad + \delta\mathbb{E}\left(1(\delta R_a < Y_2)\int_0^{\delta(S - R_a)}\int_0^v\bar{F}_h'''(Y_2 + u)dudv\right) \\ &\quad - \lambda\mathbb{E}\int_0^{\delta S}(\delta S - v)\int_0^v\bar{F}_h'''(Y_2 + u)dudv. \end{aligned}$$

Lemma 8 expresses $\mathbb{E}h(Y_2) - \mathbb{E}h(X)$ in terms of the difference of the diffusion and prelimit generators, and is the prelimit-approach analog of Proposition 1.

3.4. The prelimit approach with multiple general clocks

Before departing this section, we comment on the difficulty of extending the prelimit approach to models with multiple general clocks. As an illustration, consider the $G/G/1$ workload process in which arrivals are the superposition of two independent renewal streams with interarrival distributions U_1 and U_2 and rates λ_1 and λ_2 , respectively; the workload brought by a customer (from either stream) is i.i.d. S . The state is

$$Z(t) = (V(t), R_{a,1}(t), R_{a,2}(t)),$$

where $R_{a,i}(t)$ is the residual interarrival time of stream $i \in \{1, 2\}$. For notational simplicity we do not apply heavy-traffic scaling here (i.e., set $\delta = 1$).

Analogous to the single-clock case, one can derive the Poisson equation

$$\begin{aligned} &-\partial_v\mathbb{E}F_h(v, R_{a,1}, R_{a,2}) \\ &+ \lambda_1\mathbb{E}(F_h(v + S, U_1, R_{a,2}) - F_h(v, U_1, R_{a,2})) \\ &+ \lambda_2\mathbb{E}(F_h(v + S, R_{a,1}, U_2) - F_h(v, R_{a,1}, U_2)) = \mathbb{E}h(V) - h(v), \quad v \geq 0. \end{aligned}$$

However, relating the mixed expectations

$$\mathbb{E}(F_h(v + S, U_1, R_{a,2}) - F_h(v, U_1, R_{a,2})) \quad \text{and} \quad F_h(v + S, R_{a,1}, U_2) - F_h(v, R_{a,1}, U_2),$$

to the “all-residuals” quantity

$$\mathbb{E}(F_h(v + S, R_{a,1}, R_{a,2}) - F_h(v, R_{a,1}, R_{a,2}))$$

is substantially more delicate in the presence of multiple general clocks. The single-clock argument underlying Lemma 7—which conditions on a fixed residual time r_a and “waits” until the next arrival—no longer cleanly applies: the path segment from time 0 to the next type-1 arrival has a complicated dependence on the entire trajectory of the type-2 clock. In particular, type-2 arrivals may occur before $r_{a,1}$, repeatedly resetting $R_{a,2}(t)$ and altering the workload trajectory, which introduces nontrivial cross-terms that we do not know how to handle.

We expect the technique underlying Lemma 7 to extend only to settings with a single general clock. In models with multiple general clocks, the cross-dependencies preclude a straightforward extension and appear to require genuinely new ideas. We therefore leave this extension as an open problem.

4. Stein factor bounds for the G/G/1 workload process

In this section we derive second- and third-order Stein factor bounds and then combine them with Lemma 8 to prove Theorem 2. Although we focus on the case $h(x) = x$, the same approach extends to Lipschitz test functions with Lipschitz first and second derivatives.

Throughout, we assume that U is absolutely continuous with density $G'(x)$ and hazard rate

$$\eta(x) = \frac{G'(x)}{1 - G(x)}, \quad x \geq 0,$$

and write $\bar{\eta} = \sup_{x \geq 0} \eta(x)$. Extending the analysis to interarrival distributions with point masses would require additional work; see the discussion following Lemma 13 in Section 4.1.

Our proof of Theorem 2 proceeds in stages. We first establish auxiliary results containing Stein factor bounds for the workload process. These appear in two parts: Lemma 9, proved in Section 4.1, and Lemma 10, proved in Section 4.2. Only the latter requires $\eta(x)$ to be nonincreasing; the proof can be adapted to more general hazard rate structures, but at the expense of added complexity; see Remark 4. Throughout, recall that $\bar{B}$ denotes the busy-period duration as defined in (16).

LEMMA 9. Suppose that $\bar{\eta} < \infty$ and fix $h(x) = x$. Then for any $x \geq 0$,

$$0 \leq \delta \bar{F}'_h(x) \leq x(1 + (\lambda + \bar{\eta})\mathbb{E}\bar{B}), \quad 0 \leq \delta \partial_x \mathbb{E}F_h(x, U) \leq x(1 + 2\bar{\eta}\mathbb{E}\bar{B}), \quad (35)$$

$$0 \leq \delta \bar{F}''_h(x) \leq 1 + (\lambda + \bar{\eta})\mathbb{E}\bar{B} \quad \text{and} \quad 0 \leq \delta \partial_x^2 \mathbb{E}F_h(x, U) \leq 1 + 2\bar{\eta}\mathbb{E}\bar{B}. \quad (36)$$

LEMMA 10. Suppose that $\bar{\eta} < \infty$, that $\eta(x)$ is nonincreasing, and fix $h(x) = x$. Then for any $x \geq 0$,

$$|\delta^2 \bar{F}'''_h(x)| \leq \mathbb{P}(\delta U > x) 3\lambda \bar{\eta} \mathbb{E}\bar{B} + \mathbb{P}(\delta U < x < \delta \bar{I} + \delta U) \lambda \bar{\eta} (1 + \rho + \lambda^2 \mathbb{E}U^2) \mathbb{E}\bar{B},$$

where U and $\bar{I}$ are independent.

The Stein factor bounds in Lemmas 9 and 10 both involve the busy-period duration $\mathbb{E}\bar{B}$. Thus, in order to make the bounds explicit we require an upper bound on $\mathbb{E}\bar{B}$ in terms of the interarrival and service distributions. A classical result due to [Loulou \(1978\)](#) provides such a bound:

$$\mathbb{E}\bar{B} \leq 0.9\rho \frac{\sqrt{\text{Var}(U - S)}}{1 - \rho} \exp \left(5.4 \frac{\mathbb{E}|U - S|^3}{(\text{Var}(U - S))^{3/2}} + 0.8 \frac{\mathbb{E}(U - S)}{\sqrt{\text{Var}(U - S)}} \right). \quad (37)$$

Notably, the right-hand side depends only on the first three moments of S and U . We next present an alternative bound, derived in Section 4.3.

LEMMA 11. Recall that $I_1, I_2, \dots$ are i.i.d. $\bar{I}$. For any $\rho < 1$,

$$\mathbb{E}V = \lambda \frac{\mathbb{E}S^2}{2} + \lambda \mathbb{E}\bar{B} \left[\mathbb{E}(S - U)^+ + \sum_{k=2}^{\infty} \mathbb{E} \left(S - U - \sum_{i=1}^{k-1} I_i \right)^+ \right]. \quad (38)$$

Denoting the steady-state customer waiting time distribution by W , a consequence of (38) is that

$$\mathbb{E}\bar{B} = \frac{\mathbb{E}W\mathbb{E}S}{\mathbb{E}(S - U)^+ + \sum_{k=2}^{\infty} \mathbb{E} \left(S - U - \sum_{i=1}^{k-1} I_i \right)^+} \leq \frac{\rho \text{Var}(S - U)}{2(1 - \rho)\mathbb{E}(S - U)^+}. \quad (39)$$

REMARK 2. The bound in (39) is a result of the upper bound on $\mathbb{E}W$ due to [Kingman \(1962\)](#). Tighter bounds on $\mathbb{E}W$ have been established since then ([Daley et al. 1992](#)), and any one of them could be used instead.

We require one final auxiliary lemma, proved in Appendix D.3, which uses the Stein factor bounds from Lemmas 9 and 10 to bound the right-hand side of the Taylor expansion in Lemma 8. We recall from (18) that $\mathbb{E}R_a^k = \lambda \mathbb{E}U^{k+1}/(k+1)$.

LEMMA 12. *Let $h(x) = x$. Then*

$$\mathbb{E} |h(Y_2 + J(Y_2, R)) - h(Y_2)| \leq \delta(\lambda \mathbb{E}U^2/2 + \mathbb{E}S), \quad (40)$$

$$\mathbb{E} |\epsilon(Y_2, R_a, S)| \leq \delta(\rho \mathbb{E}U^2/2 + \delta(1/\mathbb{E}Y_2)\lambda(\mathbb{E}U^2/2)(\mathbb{E}S^2 + (\mathbb{E}S)^2)(\delta + 2\bar{\eta}\delta\mathbb{E}\bar{B})), \quad (41)$$

$$\begin{aligned} \delta \mathbb{E} \left| 1(\delta R \geq Y_2) (\bar{F}'_h(\delta S) - \bar{F}'_h(Y_2) - \delta(S - R)\bar{F}''_h(Y_2)) \right| \\ \leq \delta((1/\mathbb{E}Y_2)(\rho \mathbb{E}U^2 + 5\lambda \mathbb{E}U^3/3) + \lambda \mathbb{E}U^3/3)(\delta + (\lambda + \bar{\eta})\delta\mathbb{E}\bar{B}), \end{aligned} \quad (42)$$

$$\lambda \left| \mathbb{E} \int_0^{\delta S} (\delta S - v) \int_0^v \bar{F}'''_h(Y_2 + u) du dv \right| \leq \delta \lambda \mathbb{E}S^3 C, \quad (43)$$

$$\delta \left| \mathbb{E} 1(\delta R < Y_2) \int_0^{\delta(S-R)} \int_0^v \bar{F}'''_h(Y_2 + u) du dv \right| \leq \delta \mathbb{E}(S^2 - \rho U^2 + \lambda \mathbb{E}U^3/3) C, \quad (44)$$

where $C = (3 + (\mathbb{E}U + \mathbb{E}\bar{I})\lambda(1 + \rho + \lambda^2 \mathbb{E}U^2))(1/\mathbb{E}Y_2)\bar{\eta}\delta\mathbb{E}\bar{B}$.

Proof of Theorem 2 We apply the bounds in Lemma 12 to the expression for $\mathbb{E}h(X) - \mathbb{E}h(Y_2)$ in Lemma 8. The constants K_1 through K_5 correspond exactly to the upper bounds in (40)–(44), respectively. The relationship between $\mathbb{E}\bar{I}$ and $\mathbb{E}\bar{B}$ was previously mentioned in (17) and the bound on $\mathbb{E}\bar{B}$ follows from using the lesser of the two bounds in (37) and Lemma 11. $\square$

4.1. Second-derivative bound

We now establish Lemma 9. Let $\alpha(t)$ denote the age of the interarrival process (the backward recurrence time) at time t . Our first step is to identify an expression for the second derivative of $\mathbb{E}F_h(x, T)$ with respect to x , obtained by differentiating $\partial_x F_h(x, r_a)$. We prove it in Appendix D.1.

LEMMA 13. *Fix $h(x) = x$ and suppose that $\bar{\eta} < \infty$. For any absolutely continuous random variable $T \geq 0$ with bounded density $\theta(x)$, and any $x \geq 0$,*

$$\partial_x^2 \mathbb{E}F_h(x, T) = \frac{1}{\delta} + \frac{1}{\delta} (\theta(x/\delta) + \mathbb{E}(1(T < x/\delta) \mathbb{E}_{x,T} \eta(\alpha(B_0)))) \mathbb{E}\bar{B}.$$

REMARK 3. In the proof of Lemma 13, the quantity

$$\lim_{\epsilon \rightarrow 0} \frac{1}{\epsilon} \mathbb{P}_{x, r_a}(R_a(B_0) < \epsilon/\delta)$$

arises. Since $R_a(B_0) = I_0$, this is the density at zero of the idle period I_0 . When U has a density, the latter equals $\mathbb{E}_{x, r_a} \eta(\alpha(B_0))$, which will play an important role in the third-derivative bounds. If U has point masses one cannot use the hazard-rate connection, which makes the analysis more challenging.

Proof of Lemma 9 Since the densities of R_a and U are bounded by λ and $\bar{\eta}$, respectively, Lemma 13 yields

$$\delta \partial_x^2 \mathbb{E} F_h(x, R_a) \leq 1 + (\lambda + \bar{\eta}) \mathbb{E} \bar{B} \quad \text{and} \quad \delta \partial_x^2 \mathbb{E} F_h(x, U) \leq 1 + 2\bar{\eta} \mathbb{E} \bar{B}, \quad x \geq 0,$$

which proves (36). The first-derivative bounds in (35) follow from the fact that

$$\partial_x \mathbb{E} F_h(x, T) = \partial_x \mathbb{E} F_h(0, T) + \int_0^x \partial_x^2 \mathbb{E} F_h(u, T) dt = \int_0^x \partial_x^2 \mathbb{E} F_h(u, T) dt, \quad x \geq 0$$

where the second equality is due to Lemma 5.

□

4.2. Third-derivative bound

In this section we prove Lemma 10. Differentiating both sides of the Poisson equation (32) twice yields

$$\delta \bar{F}_h'''(x) = \lambda \partial_x^2 \mathbb{E}(F_h(x + \delta S, U) - F_h(x, U)) + h''(x). \quad (45)$$

The following lemma is proved in Appendix D.2.1. It follows directly from Lemma 13 after verifying that $\partial_x^2 \mathbb{E}(F_h(x + \delta S, U) - F_h(x, U)) = \mathbb{E}^S(\partial_x^2 \mathbb{E}^U F_h(x + \delta S, U) - \partial_x^2 \mathbb{E}^U F_h(x, U))$.

LEMMA 14. Fix $h(x) = x$ and suppose that $\bar{\eta} < \infty$. Then for any $x \geq 0$,

$$\begin{aligned} \delta^2 \bar{F}_h'''(x) &= \lambda \mathbb{E}(G'(x/\delta + S) - G'(x/\delta)) \mathbb{E} \bar{B} \\ &\quad + \lambda \mathbb{E}\left(1(x/\delta < U < x/\delta + S) \mathbb{E}_{x+\delta S, U} \eta(\alpha(B_0))\right) \mathbb{E} \bar{B} \\ &\quad + \lambda \mathbb{E}\left(1(U < x/\delta) \left(\mathbb{E}_{x+\delta S, U} \eta(\alpha(B_0)) - \mathbb{E}_{x, U} \eta(\alpha(B_0))\right)\right) \mathbb{E} \bar{B}. \end{aligned}$$

The first two terms are straightforward to bound, whereas the third term

$$\mathbb{E}\left(1(U < x/\delta)\left(\mathbb{E}_{x+\delta S, U}\eta(\alpha(B_0)) - \mathbb{E}_{x, U}\eta(\alpha(B_0))\right)\right)\mathbb{E}\bar{B}$$

requires substantially more effort. Naively bounding it by $\bar{\eta}\mathbb{E}\bar{B}$ is insufficient, since we need an additional factor of δ in the bound. The following lemma is proved in Section 4.2.1, where we also describe how to establish an analogous result when the hazard rate is uniformly bounded away from zero.

LEMMA 15. *Assume that $\bar{\eta} < \infty$ and that $\eta(x)$ is nonincreasing. For any $x, s \geq 0$, and $r_a < x/\delta$,*

$$\left|\mathbb{E}_{x+\delta s, r_a}\eta(\alpha(B_0)) - \mathbb{E}_{x, r_a}\eta(\alpha(B_0))\right| \leq \bar{\eta}(1 + \lambda s + \lambda^2 \mathbb{E}U^2)\mathbb{P}(\bar{I} > x/\delta - r_a).$$

Proof of Lemma 10 We bound, one by one, each term in the expression of $\delta^2 \bar{F}_h'''(x)$ in Lemma 14. Observe that

$$\begin{aligned} \lambda \left|\mathbb{E}(G'(x/\delta + S) - G'(x/\delta))\right| \mathbb{E}\bar{B} &\leq 2\lambda\bar{\eta}\mathbb{P}(U > x/\delta)\mathbb{E}\bar{B}, \\ \lambda \mathbb{E}\left(1(x/\delta < U < x/\delta + S)\mathbb{E}_{x+\delta S, U}\eta(\alpha(B_0))\right) \mathbb{E}\bar{B} &\leq \lambda\bar{\eta}\mathbb{P}(U > x/\delta)\mathbb{E}\bar{B}, \end{aligned}$$

where in the first inequality we used $G'(x) = \eta(x)\mathbb{P}(U > x) \leq \bar{\eta}\mathbb{P}(U > x)$. Lastly we invoke Lemma 15, which implies that

$$\begin{aligned} &\lambda \mathbb{E}\left(1(U < x/\delta) \left|\mathbb{E}_{x+\delta S, U}\eta(\alpha(B_0)) - \mathbb{E}_{x, U}\eta(\alpha(B_0))\right|\right) \mathbb{E}\bar{B} \\ &\leq \lambda\bar{\eta}(1 + \lambda \mathbb{E}S + \lambda^2 \mathbb{E}U^2)\mathbb{E}^U(1(U < x/\delta)\mathbb{E}^{\bar{I}}(1(\bar{I} > x/\delta - U)))\mathbb{E}\bar{B} \\ &= \lambda\bar{\eta}(1 + \rho + \lambda^2 \mathbb{E}U^2)\mathbb{P}(U < x/\delta < \bar{I} + U)\mathbb{E}\bar{B}, \end{aligned}$$

where U and $\bar{I}$ are independent. □

4.2.1. The renewal process driven by idle times In this section we prove Lemma 15. Recall that $\alpha(t)$ is the age of the interarrival process at time $t \geq 0$. Further recall from the discussion following (16) that B_n and I_n , $n \geq 0$, are the lengths of the n th busy and idle periods, respectively. While B_0 and I_0 are the initial busy and idle period lengths, which depend on the initial condition $Z(0)$, the pairs (B_n, I_n) , $n \geq 1$, are i.i.d. with

distributions $\bar{B}$ and $\bar{I}$, respectively. Given $Z(0-) = (0, 0)$ (an arrival occurs to an empty system at time $t = 0$), consider $\alpha(B_0)$, the age at the end of the first busy period, and let $\bar{\alpha}$ be the random variable having its distribution.

Central to our analysis is the following continuous-time Markov process that behaves like the interarrival age during idle periods of the workload process. Formally, we let $\{\Phi(t) : t \geq 0\}$ be defined by the infinitesimal generator

$$G_{\Phi}f(\gamma) = \lim_{t \rightarrow 0} \frac{\mathbb{E}_{\gamma}f(\Phi(t)) - f(\gamma)}{t} = f'(\gamma) + \eta(\gamma)(\mathbb{E}f(\bar{\alpha}) - f(\gamma)), \quad \gamma \geq 0. \quad (46)$$

This process grows at unit rate and, at its jump times, $\Phi(t) \stackrel{d}{=} \bar{\alpha}$. Jumps occur with rate $\eta(\Phi(t))$.

Let $\{\Phi(t) : t \geq 0\}$ and $\{\tilde{\Phi}(t) : t \geq 0\}$ be two copies (possibly coupled) of the Markov process defined by (46). The following technical lemma is proved at the end of this section.

LEMMA 16. *Let $v = v(x) = x/\delta$. For any $(x, r_a) \in \mathbb{S}$ with $r_a < v$ and any $s \geq 0$,*

$$\begin{aligned} & \mathbb{E}_{x+\delta s, r_a} \eta(\alpha(B_0)) - \mathbb{E}_{x, r_a} \eta(\alpha(B_0)) \\ &= \mathbb{E}(\eta(\Phi(v - r_a + s)) - \eta(\tilde{\Phi}(v - r_a + s)) | \Phi(0) = \tilde{\Phi}(s), \Phi(0) \stackrel{d}{=} \bar{\alpha}). \end{aligned} \quad (47)$$

Furthermore if $\Phi(0) \stackrel{d}{=} \bar{\alpha}$, then the inter-jump times of $\{\Phi(t) : t \geq 0\}$ are i.i.d. $\bar{I}$.

The final ingredient for the proof of Lemma 15 is a coupling $\{(\tilde{\Phi}(t), \Phi(t)) : t \geq 0\}$, which allows us to bound the expression in (47). Defining

$$\eta_m(x, y) = \min\{\eta(x), \eta(y)\}, \quad \eta_{\Delta}(x, y) = \max\{\eta(x), \eta(y)\} - \min\{\eta(x), \eta(y)\},$$

we let $\{(\tilde{\Phi}(t), \Phi(t)), t \geq 0\}$ have the same distribution as the Markov process defined by the generator

$$\begin{aligned} G_J f(\tilde{\gamma}, \gamma) &= \partial_{\tilde{\gamma}} f(\tilde{\gamma}, \gamma) + \partial_{\gamma} f(\tilde{\gamma}, \gamma) \\ &\quad + \eta_m(\tilde{\gamma}, \gamma)(\mathbb{E}f(\bar{\alpha}, \bar{\alpha}) - f(\tilde{\gamma}, \gamma)) \\ &\quad + \eta_{\Delta}(\tilde{\gamma}, \gamma)1(\eta(\gamma) < \eta(\tilde{\gamma}))(\mathbb{E}f(\bar{\alpha}, \gamma) - f(\tilde{\gamma}, \gamma)) \\ &\quad + \eta_{\Delta}(\tilde{\gamma}, \gamma)1(\eta(\gamma) > \eta(\tilde{\gamma}))(\mathbb{E}f(\tilde{\gamma}, \bar{\alpha}) - f(\tilde{\gamma}, \gamma)). \end{aligned} \quad (48)$$

Note that the marginal law of either component of this process is equivalent to the Markov process defined in (46). Furthermore, when this process jumps it either couples (with rate $\eta_m(\tilde{\gamma}, \gamma)$), or only one of the components jumps.

Proof of Lemma 15 Fix $v = x/\delta \geq 0$ and $r_a < v$, and let $\tau_C = \inf\{t \geq s : \tilde{\Phi}(t) = \Phi(t)\}$, it follows that

$$\begin{aligned} & |\mathbb{E}_{x+\delta s, r_a} \eta(\alpha(B_0)) - \mathbb{E}_{x, r_a} \eta(\alpha(B_0))| \\ &= \left| \mathbb{E}(\eta(\Phi(v - r_a + s)) - \eta(\tilde{\Phi}(v - r_a + s)) | \Phi(0) = \tilde{\Phi}(s), \Phi(0) \stackrel{d}{=} \bar{\alpha}) \right| \\ &\leq \bar{\eta} \mathbb{P}(\tau_C > v - r_a + s | \Phi(0) = \tilde{\Phi}(s), \Phi(0) \stackrel{d}{=} \bar{\alpha}). \end{aligned} \quad (49)$$

If $\tilde{\Phi}(s) < \Phi(s)$, then $\eta_m(\Phi(t), \tilde{\Phi}(t)) = \eta(\tilde{\Phi}(t))$ for all $t \geq s$ until the first jump of $\{\tilde{\Phi}(t) : t \geq 0\}$, at which point coupling occurs. Since $\tilde{\Phi}(s) \stackrel{d}{=} \bar{\alpha}$, we know by Lemma 16 that the first jump after s happens after $\bar{I}$ amount of time, implying that the probability of no jump on $(s, v - r_a + s]$ is at most

$$\mathbb{P}(\bar{I} > v - r_a). \quad (50)$$

Alternatively, if $\tilde{\Phi}(s) > \Phi(s)$, then τ_C corresponds to the first jump time after s of $\{\Phi(t) : t \geq 0\}$. We will shortly prove that the probability that $\{\Phi(t) : t \geq 0\}$ does not jump on $(s, v - r_a + s]$ is at most

$$\mathbb{P}(\bar{I} > v - r_a) \mathbb{E}(N_{\Phi}(s) | \Phi(0) \stackrel{d}{=} \bar{\alpha}), \quad (51)$$

where $N_{\Phi}(t)$ is the number of jumps made by $\{\Phi(t) : t \geq 0\}$ on $[0, t]$. Adding (50) and (51) yields

$$\bar{\eta} \mathbb{P}(\tau_C > v - r_a + s | \Phi(0) = \tilde{\Phi}(s) \stackrel{d}{=} \bar{\alpha}) \leq \bar{\eta}(1 + \mathbb{E}(N_{\Phi}(s) | \Phi(0) \stackrel{d}{=} \bar{\alpha})) \mathbb{P}(\bar{I} > v - r_a).$$

Let $\{N_U(t) : t \geq 0\}$ be the counting process of a zero-delayed renewal process whose inter-event times have the interarrival distribution U . It follows that

$$\mathbb{E}(N_{\Phi}(s) | \Phi(0) \stackrel{d}{=} \bar{\alpha}) \leq \mathbb{E}N_U(s) \leq \lambda s + \lambda^2 \mathbb{E}U^2, \quad s \geq 0,$$

where the first inequality is due to the hazard rate $\eta(x)$ being nonincreasing, and the second inequality follows from Lorden's inequality (Lorden 1970).

It remains to verify (51). Let $T_0 = 0$ and T_n , $n \geq 1$, be the n th jump time of $\{\Phi(t) : t \geq 0\}$. Also let $J_n = T_n - T_{n-1}$, $n \geq 1$, be the duration between the n th and $(n-1)$ st jump. We know by Lemma 16 that $J_n \stackrel{d}{=} \bar{I}$ whenever $\Phi(0) \stackrel{d}{=} \bar{\alpha}$. It follows that (to simplify notation, all probabilities and expectations that follow are conditional on $\Phi(0) \stackrel{d}{=} \bar{\alpha}$)

$$\begin{aligned} \mathbb{P}(N_\Phi(v - r_a + s) - N_\Phi(s) = 0) &= \mathbb{P}(J_{N_\Phi(s)+1} > v - r_a + s - T_{N_\Phi(s)}) \\ &= \sum_{n=0}^{\infty} \mathbb{P}(J_{n+1} > v - r_a + s - T_n, N_\Phi(s) = n). \end{aligned}$$

Since $\{N_\Phi(s) = n\} = \{T_n < s, J_{n+1} > s - T_n\}$, the right-hand side equals

$$\begin{aligned} \sum_{n=0}^{\infty} \mathbb{P}(J_{n+1} > v - r_a + s - T_n, T_n < s) &\leq \sum_{n=0}^{\infty} \mathbb{P}(J_{n+1} > v - r_a, T_n < s) \\ &= \mathbb{P}(\bar{I} > v - r_a) \sum_{n=0}^{\infty} \mathbb{P}(T_n < s) \\ &= \mathbb{P}(\bar{I} > v - r_a) \sum_{n=0}^{\infty} \mathbb{P}(N_\Phi(s) \geq n) \\ &= \mathbb{P}(\bar{I} > v - r_a) \mathbb{E}N_\Phi(s), \end{aligned}$$

where the first equality is true because $J_{n+1} \stackrel{d}{=} \bar{I}$ and J_{n+1} is independent of T_n . $\square$

REMARK 4. We proved Lemma 15 under the assumption that $\eta(x)$ is nonincreasing. The key step was to obtain an upper bound on the coupling probability in (49), which in turn yields a bound on $\bar{F}_h'''(x)$. The latter bound is then used in Lemma 12 to establish (43) and (44), and must be sharp enough to ensure that the right-hand sides of those inequalities are of order δ . Another family of hazard rates for which the coupling probability can be controlled consists of those uniformly bounded away from zero: $\eta(x) \geq \underline{\eta} > 0$. In this case, $\eta_m(\tilde{\gamma}, \gamma) \geq \underline{\eta}$, so coupling occurs after an exponentially distributed time with rate $\underline{\eta}$, leading to the bound

$$\mathbb{P}(\tau_C > v - r_a + s \mid \Phi(0) = \tilde{\Phi}(s), \Phi(0) \stackrel{d}{=} \bar{\alpha}) \leq \bar{\eta} e^{-\underline{\eta}(v-r)}.$$

Proof of Lemma 16 Define ℓ_n , $n \geq 0$, to be the end of the n th busy period. Namely, $\ell_0 = B_0$ and

$$\ell_n = \ell_{n-1} + I_{n-1} + B_n, \quad n \geq 1.$$

Letting $T_0 = 0$ and $T_{n+1} = T_n + I_n$, $n \geq 0$, so that the n th idle period occurs on the interval $[\ell_n, \ell_n + T_{n+1} - T_n)$, we define

$$\Gamma(t) = \alpha(\ell_n + t - T_n), \quad t \in [T_n, T_{n+1}), \quad n \geq 0, \quad (52)$$

to be the age of the interarrival process at the instant when the server has idled for exactly t time units.

Recalling our synchronous coupling $\{Z^{(x)}(t) : t \geq 0\}$ introduced in (24) and the corresponding initial busy period $B_0^{(x)}$, it follows that

$$\alpha(B_0^{(x)}) = \Gamma(v), \quad x \geq 0,$$

because $B_0^{(x)}$ is precisely the instant when $\{Z(t) : t \geq 0\}$ idles for v time units. Therefore, for any $(x, r_a) \in \mathbb{S}$ with $r_a < v$ and any $s \geq 0$,

$$\begin{aligned} \mathbb{E}_{x+\delta s, r_a} \eta(\alpha(B_0)) - \mathbb{E}_{x, r_a} \eta(\alpha(B_0)) &= \mathbb{E}_{0, r_a} \eta(\alpha(B_0^{(x+\delta s)})) - \mathbb{E}_{0, r_a} \eta(\alpha(B_0^{(x)})) \\ &= \mathbb{E}_{0, r_a} (\eta(\Gamma(v+s)) - \eta(\Gamma(v))), \end{aligned} \quad (53)$$

where we recall that $\mathbb{E}_{x, r_a}(\cdot)$ is the expectation conditioned on $Z(0) = (x, r_a)$. To proceed, we now show that although $\{\Gamma(t) : t \geq 0\}$ is defined via (52), it is also a Markov process with generator (46).

We know from (52) that $\{\Gamma(t) : t \geq 0\}$ increases at a unit rate. It jumps at times T_{n+1} , $n \geq 0$, and its distribution at jump times satisfies $\Gamma(T_{n+1}) = \alpha(\ell_{n+1}) \stackrel{d}{=} \bar{\alpha}$ for all $n \geq 0$, because ℓ_{n+1} marks the end of a busy period initiated by an arrival to an empty system. Lastly, conditioned on $\Gamma(t)$, the probability that a jump occurs on the interval $(t, t+dt)$ equals

$$\frac{\mathbb{P}(\Gamma(t) < U < \Gamma(t) + dt)}{\mathbb{P}(U > \Gamma(t))} = \eta(\Gamma(t))dt + o(dt),$$

where $o(dt) \rightarrow 0$ as $dt \rightarrow 0$; the Markov property is simple to verify. Thus $\{\Gamma(t) : t \geq 0\}$ is a Markov process described by the generator (46) and, given the same initial condition, both $\{\Phi(t) : t \geq 0\}$ and $\{\Gamma(t) : t \geq 0\}$ are equal in distribution. We now argue that given $\Phi(0) \sim \bar{\alpha}$, the inter-jump times of $\{\Phi(t) : t \geq 0\}$ are i.i.d. $\bar{I}$.

By construction, the value of $\Gamma(\cdot)$ at jump times T_{n+1} satisfies $\Gamma(T_{n+1}) \stackrel{d}{=} \bar{\alpha}$, and the inter-jump durations $T_{n+2} - T_{n+1} \stackrel{d}{=} \bar{I}$ for $n \geq 0$. Since $\{\Gamma(t) : t \geq 0\}$ is a Markov process, it follows that conditioned on $\Gamma(0) \stackrel{d}{=} \bar{\alpha}$, the time until the first jump has the same distribution as $\bar{I}$.

Coming back to right-hand side of (53), suppose that $Z(0) = (0, r_a)$. Then $\Gamma(r_a) = \alpha(\ell_1) \stackrel{d}{=} \bar{\alpha}$, because $B_0 = 0$ and $I_0 = r_a$, which in turn yields $r_a \in [T_1, T_2]$ (since $T_1 = r_a$). Furthermore, conditioned on both $Z(0) = (0, r_a)$ and $\Gamma(r_a)$, the value of $\Gamma(r_a + t)$, $t \geq 0$, is independent of $Z(0)$. It follows that for any v and $r_a < v$,

$$\begin{aligned} \mathbb{E}_{0, r_a}(\eta(\Gamma(v+s)) - \eta(\Gamma(v))) &= \mathbb{E}(\eta(\Gamma(v+s)) - \eta(\Gamma(v)) | \Gamma(r_a) \stackrel{d}{=} \bar{\alpha}) \\ &= \mathbb{E}(\eta(\Phi(v+s)) - \eta(\Phi(v)) | \Phi(r_a) \stackrel{d}{=} \bar{\alpha}) \\ &= \mathbb{E}(\eta(\Phi(v-r_a+s)) - \eta(\Phi(v-r_a)) | \Phi(0) \stackrel{d}{=} \bar{\alpha}), \end{aligned}$$

where the third equality comes from the Markov property. Finally, if $\{\tilde{\Phi}(t) : t \geq 0\}$ is a copy of the Markov process defined by (46), then

$$\begin{aligned} &\mathbb{E}(\eta(\Phi(v-r_a+s)) - \eta(\Phi(v-r_a)) | \Phi(0) \stackrel{d}{=} \bar{\alpha}) \\ &= \mathbb{E}(\eta(\Phi(v-r_a+s)) - \eta(\tilde{\Phi}(v-r_a)) | \Phi(0) = \tilde{\Phi}(0), \Phi(0) \stackrel{d}{=} \bar{\alpha}) \\ &= \mathbb{E}(\eta(\Phi(v-r_a+s)) - \eta(\tilde{\Phi}(v-r_a+s)) | \Phi(0) = \tilde{\Phi}(s), \Phi(0) \stackrel{d}{=} \bar{\alpha}). \end{aligned}$$

□

4.3. An expression for $\mathbb{E}\bar{B}$

Proof of Lemma 11 We first prove (38). Let $\hat{h}(x) = x$, recall that $\bar{F}_{\hat{h}}(0) = 0$ due to (27), and consider the Poisson equation (32) with $h(x) = \hat{h}(x)$ evaluated at $x = 0$, which results in

$$\lambda \mathbb{E}(F_{\hat{h}}(\delta S, U) - F_{\hat{h}}(0, U)) = \delta \mathbb{E}V,$$

Using the synchronous coupling $\{Z^{(\epsilon)}(t)\}$ introduced in Section 3.2, it follows that

$$\delta \mathbb{E}V = \lambda \mathbb{E}\left(\int_0^{B^{(\delta S)}} \mathbb{E}_{0, U}(X^{(\delta S)}(t) - X(t)) dt\right). \quad (54)$$

From (24) we know that the difference $X^{(\delta S)}(t) - X(t)$ decays at rate δ only during the idle periods of $\{X(t) : t \geq 0\}$. Recall that the idle and busy period durations are $I_0, I_1, \dots$, and $B_0, B_1, B_2, \dots$, respectively, and that $B_0 = 0$ because $X(0) = 0$. It follows that for all times t corresponding to the busy period B_k , $k \geq 1$,

$$X^{(\delta S)}(t) - X(t) = \delta \left(S - \sum_{i=0}^{k-1} I_i \right)^+.$$

Letting $\mathcal{I} = \{t \in [0, B^{(\delta S)}] : X(t) = 0\}$, it follows that

$$\begin{aligned} \int_0^{B^{(\delta S)}} (X^{(\delta S)}(t) - X(t)) dt &= \int_{\mathcal{I}} (X^{(\delta S)}(t) - X(t)) dt + \int_{[0, B^{(\delta S)}] \setminus \mathcal{I}} (X^{(\delta S)}(t) - X(t)) dt \\ &= \delta S^2/2 + \sum_{k=1}^{\infty} B_k \delta \left(S - \sum_{i=0}^{k-1} I_i \right)^+. \end{aligned}$$

We conclude (38) by combining this equation with (54), noting that B_k is independent of $(S - I_0 - I_1 \dots - I_{k-1})^+$ and $B_k \stackrel{d}{=} \bar{B}$, and that $I_0 \stackrel{d}{=} U$ since $Z(0) = (0, U)$. The equality in (39) is true because $\mathbb{E}V = \lambda \mathbb{E}S^2/2 + \rho \mathbb{E}W$ due to Corollary X.3.5 of Asmussen (2003). The inequality follows from the well-known bound in (7') of Kingman (1962), which says that

$$\mathbb{E}W \leq \frac{\text{Var}(S - U)}{2\mathbb{E}(S - U)} = \frac{\rho \text{Var}(S - U)}{2(1 - \rho)\mathbb{E}S}.$$

□

5. The $G/M/\infty$ queue length

The $G/M/\infty$ system provides another instructive example. It features both general and exponential clocks, and in contrast to the $G/G/1$ setting, the diffusion error for the infinite-server queue is composed solely of interior terms and no boundary terms. As a result, obtaining a three-moment error bound is considerably simpler.

The $G/M/\infty$ system is an infinite-server queue where arrivals form a renewal process with generic interarrival time distribution U and exponentially distributed service times with rate μ . Let $R = \lambda/\mu$ be the offered load. We assume that $\mathbb{E}U^3 < \infty$ and set

$$\lambda = 1/\mathbb{E}U, \quad c_U^2 = \lambda^2 \text{Var}(U).$$

Let $Q(t)$ be the customer count, let $R_a(t)$ be the residual interarrival time at $t \geq 0$, and assume that

$$\{Z(t) = (Q(t), R_a(t))\}$$

is positive Harris recurrent. Let $Z = (Q, R_a)$ denote the vector having the stationary distribution. Let $\delta = 1/\sqrt{R}$ and define the normalized customer count

$$X(t) = \delta(Q(t) - R) \quad \text{and} \quad X = \delta(Q - R).$$

The following is the main result of this section.

THEOREM 3. *Let $Y \sim N(0, \sigma^2)$ with $\sigma^2 = (1 + c_U^2)$. Then*

$$\begin{aligned} d_W(X, Y) &\leq \delta(K_A + K_D), \\ K_A &= \frac{1}{3\sqrt{(1 + c_U^2)}} \mathbb{E}|1 - \lambda U|^3 + \left(1 + \frac{1}{2}\lambda^2 \mathbb{E}U^2 (\delta^4 \lambda^3 \mathbb{E}U^3/3 + \delta^2 \lambda^2 \mathbb{E}U^2 + 1)\right) \\ K_D &= \sqrt{\frac{2/\pi}{(1 + c_U^2)}} \sqrt{\mathbb{E}X^2} \sqrt{\lambda^3 \mathbb{E}U^3/3} + \frac{1}{(1 + c_U^2)} \left(\delta^2 \lambda^3 \mathbb{E}U^3/3 + \lambda^2 \mathbb{E}U^2 + \frac{1}{3}\right) \end{aligned}$$

The proof of Theorem 3 follows the same structure as the proof of Theorem 1. Namely, we use the BAR for the compensated customer count in the infinite-server queue to extract a diffusion generator and corresponding error terms in Proposition 3. Bounding the error terms requires both Stein factor bounds and a bound on the second moment of the centered queue length (unlike our $G/G/1$ example, which did not require moment bounds). The former are well known and the latter is obtained from the BAR by using a quadratic test function. After presenting said Stein factor and moment bounds, we combine everything to prove Theorem 3 at the end of the section.

For the BAR, we let $A(t)$ and $U(t)$ be defined as in Section 2.1, and let $D(t)$ be the number of departures on $[0, t]$.

LEMMA 17. *Initialize $Z(0) \sim Z$. If $\mathbb{E}|f(Z)| < \infty$ and, $Z(0)$ -almost surely, $df(Z(t))/dt$ exists for almost all t , then, provided that all expectations are well defined, for all $t \geq 0$,*

$$\begin{aligned} 0 &= -t\mathbb{E}(\partial_{r_a} f(Z)) + \mathbb{E} \int_0^t \Delta f(Z(s-)) dD(s) + \mathbb{E} \int_0^t \Delta f(Z(s-)) dA(s) \\ &= t\mathbb{E}(\partial_{r_a} f(Z)) + t\mu \mathbb{E}Q(f(Q-1, R_a) - f(Q, R_a)) + \mathbb{E} \int_0^t \Delta f(Z(s-)) dA(s). \end{aligned} \quad (55)$$

Proof of Lemma 17 The first equality follows as in Lemma 1 and the second is due to Theorem 3.3.1 of Baccelli and Brémaud (2003). $\square$

REMARK 5. For convenience, we will work with the right-hand side of (55) with $t = 1$ there. For an approach that avoids using the machinery of Baccelli and Brémaud (2003) to establish the second equality in (55), consider the first equality in (55), divide both sides by t , and let $t \rightarrow 0$ to get

$$\begin{aligned} 0 &= -\mathbb{E}(\partial_{r_a} f(Z)) + \lim_{t \rightarrow 0} \frac{1}{t} \mathbb{E} \int_0^t \Delta f(Z(s-)) dD(s) + \lim_{t \rightarrow 0} \frac{1}{t} \mathbb{E} \int_0^t \Delta f(Z(s-)) dA(s) \\ &= -\mathbb{E}(\partial_{r_a} f(Z)) + \mu \mathbb{E} Q(f(Q-1, R_a) - f(Q, R_a)) + \lim_{t \rightarrow 0} \frac{1}{t} \mathbb{E} \int_0^t \Delta f(Z(s-)) dA(s). \end{aligned}$$

The second equality is argued using the fact that $\mathbb{P}(D(t) = 1) = t\mu Q(0) + o(t)$ and $\mathbb{P}(D(t) > 1) = o(t)$ for t . Since $\mathbb{E}A(t) = \lambda t$, the presence of the limit as $t \rightarrow 0$ in the third term does not add further complexity (compared to the $t = 1$ case) to our diffusion generator extraction procedure (Proposition 3).

Defining the compensated customer count

$$\tilde{X}(t) = X(t) - \delta \lambda R_a(t) \quad \text{and} \quad \tilde{X} = X - \delta \lambda R_a, \quad (56)$$

and specializing the BAR (55) to $\tilde{X}$ yields

$$\delta \lambda \mathbb{E} f'(\tilde{X}) + \mu \mathbb{E}(Q(f(\tilde{X} - \delta) - f(\tilde{X}))) + \mathbb{E} \int_0^1 \Delta f(\tilde{X}(t-)) dA(t) = 0. \quad (57)$$

The following expansion of (57) is analogous to Proposition 1 and is proved in Appendix E.

PROPOSITION 3. *If $f \in C^2(\mathbb{R})$ with $f''(x)$ absolutely continuous and $\|f''\|, \|f'''\| < \infty$, then, provided that all expectations are well defined,*

$$\mu \mathbb{E}(Q(f(\tilde{X} - \delta) - f(\tilde{X}))) = -\delta R f'(\tilde{X}) - \mu \mathbb{E} X f'(X) + \frac{1}{2} \mu \mathbb{E} f''(X) + \epsilon_D(f), \quad (58)$$

$$\mathbb{E} \int_0^1 \Delta f(\tilde{X}(t-)) dA(t) = \frac{1}{2} \mu c_U^2 \mathbb{E} f''(X) + \epsilon_A(f) \quad (59)$$

where

$$\begin{aligned} |\epsilon_A(f)| &\leq \frac{1}{6} \delta \mu \|f'''\| \mathbb{E} |1 - \lambda U|^3 + \frac{1}{2} \delta \mu c_U^2 \|f'''\| \left(1 + \frac{1}{2} \lambda^2 \mathbb{E} U^2 (\mathbb{E}(Q - R)^2 / R^2 + 1)\right), \\ |\epsilon_D(f)| &\leq \mu \delta \|f''\| \sqrt{\mathbb{E} X^2} \sqrt{\lambda^3 \mathbb{E} U^3 / 3} + \frac{1}{2} \mu \delta \|f'''\| \delta^2 \lambda \mathbb{E}(Q R_a) + \frac{1}{6} \mu \delta \|f'''\|. \end{aligned}$$

The upper bounds on $|\epsilon_A(f_h)|$ and $|\epsilon_D(f_h)|$ contain terms involving $\mathbb{E}(Q - R)^2$ and $\mathbb{E}(QR_a)$. The following lemma presents some useful identities, including a bound on these quantities. It is proved in Appendix [E.1](#).

LEMMA 18. *Recall that $R = \lambda/\mu$. For any $m > 1$,*

$$\mathbb{E}A(1) = \lambda, \quad \mathbb{E}R_a^{m-1} = \lambda \mathbb{E}U^m / m, \quad \mathbb{E}Q = R, \quad (60)$$

$$\mathbb{E}(Q - R)^2 = \lambda \mathbb{E}QR_a \leq \lambda^3 \mathbb{E}U^3 / 3 + R\lambda^2 \mathbb{E}U^2, \quad (61)$$

$$\mathbb{E} \int_0^1 Q(t-) dA(t) = \mu \mathbb{E}(Q - R)^2 + \lambda R - \lambda. \quad (62)$$

Applying the expansions in (58)–(59) to the BAR (57) suggests the diffusion approximation with generator

$$G_Y f(x) = -\mu x f'(x) + \frac{1}{2} \mu (1 + c_U^2) f''(x), \quad x \in \mathbb{R},$$

which corresponds to the $N(0, \sigma^2)$ distribution with $\sigma^2 = (1 + c_U^2)$. The following lemma presents the corresponding Stein factor bounds; it is a rescaling of ([Chen et al. 2011](#), Lemma 2.4).

LEMMA 19. *Let $Y \sim N(0, \sigma^2)$ and let $f_{h,\sigma}(x)$ be the solution to*

$$-x f'_{h,\sigma}(x) + \frac{1}{2} \sigma^2 f''_{h,\sigma}(x) = \frac{1}{\mu} (\mathbb{E}h(Y) - h(x)), \quad x \in \mathbb{R}, \quad (63)$$

with $f_{h,\sigma}(0) = 0$. Then

$$\|f'_{h,\sigma}\| \leq \frac{2}{\mu}, \quad \|f''_{h,\sigma}\| \leq \frac{\sqrt{2/\pi}}{\mu\sigma}, \quad \text{and} \quad \|f'''_{h,\sigma}\| \leq \frac{2}{\mu\sigma^2}.$$

As a consequence, $|f_{h,\sigma}(x)| \leq 2|x|/\mu$ for all $x \in \mathbb{R}$.

Proof of Lemma 19 Given $h \in \text{Lip}(1)$, define $\bar{h}(x) = \sigma h(\sigma x)$ and $\bar{f}(x) = f_{\bar{h},1}(x)$. Let us verify that $f_{h,\sigma}(x) = \sigma \bar{f}(x/\sigma)$ solves the Poisson equation. Since

$$f'_{h,\sigma}(x) = \bar{f}'(x/\sigma) \quad \text{and} \quad f''_{h,\sigma}(x) = \bar{f}''(x/\sigma)/\sigma,$$

it follows that

$$\begin{aligned} -xf'_{h,\sigma}(x) + \frac{1}{2}\sigma^2 f''_{h,\sigma}(x) &= -x\bar{f}'(x/\sigma) + \frac{1}{2}\sigma\bar{f}''(x/\sigma) \\ &= \frac{1}{\mu}\frac{1}{\sigma}(\mathbb{E}\bar{h}(Y/\sigma) - \bar{h}(x/\sigma)) = \frac{1}{\mu}(\mathbb{E}h(Y) - h(x)), \end{aligned}$$

and our Stein factor bounds follow from the standard result (Chen et al. 2011, Lemma 2.4) that:

$$\|\bar{f}'\| \leq \frac{2}{\mu}, \quad \|\bar{f}''\| \leq \frac{\sqrt{2/\pi}}{\mu}, \quad \text{and} \quad \|\bar{f}'''\| \leq \frac{2}{\mu}.$$

□

We are ready to prove Theorem 3.

Proof of Theorem 3 Since $\|f'_h\| < \infty$ and $|f_h(x)| \leq 2|x|/\mu$ for any $h \in \text{Lip}(1)$, one readily verifies that the BAR (57) holds with $f_h(X - \delta\lambda R_a)$. Applying the Stein factor bounds of Lemma 19 and the moment bound in (61) of Lemma 18 to the upper bounds on $|\epsilon_A(f_h)|$ and $|\epsilon_D(f_h)|$ in Proposition 3 yields

$$|\epsilon_A(f_h)| \leq K_A \quad \text{and} \quad |\epsilon_D(f_h)| \leq K_D.$$

□

6. The JSQ total customer count

The JSQ model's multiple general clocks make it a more advanced example. Consider a parallel-server system with n identical servers, each with its own buffer, operating under a first-come-first-served policy. Customer arrivals form a renewal process with generic interarrival time distribution U and customer service times are i.i.d. S . Upon arrival, customers are routed to the server with the shortest queue. We assume for simplicity that simultaneous arrivals/departures do not occur with probability one, and that ties are broken uniformly at random (other tie-breaking rules can be handled as well). We set

$$\begin{aligned} \lambda &= 1/\mathbb{E}U, \quad \mu = 1/\mathbb{E}S, \quad \rho = \lambda/n\mu, \\ c_U^2 &= \lambda^2 \text{Var}(U), \quad c_S^2 = \mu^2 \text{Var}(S), \end{aligned}$$

and assume that

$$\rho < 1, \quad \mathbb{E}U^3 < \infty, \quad \text{and} \quad \mathbb{E}S^3 < \infty.$$

Let $Q_i(t)$ be the number of customers (both in service and waiting) at server i at time $t \geq 0$. Also let $R_a(t)$ be the time until the next arrival and let $R_{s,i}(t)$ be the remaining service time of the customer being served by server i ; when server i is idle, $R_{s,i}(t)$ denotes the service time of the next customer to arrive to server i . Define $Q(t) = (Q_1(t), \dots, Q_n(t))$ and $R_s(t) = (R_{s,1}(t), \dots, R_{s,n}(t))$, and let

$$Z(t) = (Q(t), R_a(t), R_s(t)).$$

We assume that $\{Z(t) : t \geq 0\}$ is positive Harris recurrent (see [Bramson \(2011\)](#) for sufficient conditions) and let $Z = (Q, R_a, R_s)$ be the vector having its stationary distribution. Define the (scaled) total customer count

$$X(t) = \delta \sum_{i=1}^n Q_i(t) \quad \text{and} \quad X = \delta \sum_{i=1}^n Q_i,$$

where $\delta = (1 - \rho)$.

In addition to positive Harris recurrence, we assume that $\mathbb{E}X^2 < \infty$. While one can establish this via a combination of the Palm inversion formula and the BAR—first by applying the test function $f(z) = \sum_{i=1}^n (q_i \wedge M + \mu r_{s,i})^2$ and letting $M \rightarrow \infty$ to obtain $\mathbb{E}X < \infty$, and then combining this with the test function $f(z) = \sum_{i=1}^n (q_i \wedge M + \mu r_{s,i})^3$ to deduce $\mathbb{E}X^2 < \infty$ —the calculations are lengthy and would distract from the paper's main message, so we take $\mathbb{E}X^2 < \infty$ as an assumption. The following is our main result.

THEOREM 4. *Assume that $\mathbb{E}X^2 < \infty$ and let Y be exponentially distributed with mean $(\rho c_U^2 + c_S^2)/2$. Then*

$$d_W(X, Y) \leq \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left(1(Q_i = 0) \left(\sum_{j=1}^n Q_j + \lambda R_a + \sum_{j=1}^n \mu R_{s,j} \right) \right) + \delta(K_0 + K_A + K_D),$$

where

$$\begin{aligned}
 K_0 &= \frac{1}{2}(\lambda^2 \mathbb{E}U^2/2 + (\lambda\mu \mathbb{E}S^2/2 + \mu\delta \mathbb{E}S)), \\
 K_A &= \frac{4}{\lambda c_U^2 + n\mu c_S^2} \left(\frac{1}{3} \lambda \mathbb{E}|1 - \lambda U|^3 + c_U^2 n \mu \lambda (\rho(\lambda/n) \mathbb{E}S^2/2 + \delta \mathbb{E}S + \lambda \mathbb{E}U^2/2) \right. \\
 &\quad \left. + \lambda c_U^2 (1 + n + \lambda \mu \mathbb{E}U^2 + \mu^2 \mathbb{E}S^2) \right), \\
 K_D &= \frac{4n}{\lambda c_U^2 + n\mu c_S^2} \left(\frac{1}{3} \frac{\lambda}{n} \mathbb{E}|1 - \mu S|^3 \right. \\
 &\quad + c_S^2 ((\lambda^2 \mu \rho + 2(n-1)\mu^2 \rho \lambda) \mathbb{E}S^2/2n + \delta(\lambda\mu + 2(n-1)\mu^2) \mathbb{E}S + \frac{1}{2} \lambda^2 \mu \mathbb{E}U^2) \\
 &\quad + \mu c_S^2 n \rho + \mu c_S^2 (\lambda/n) (\mathbb{E}S + \lambda \mathbb{E}S^2 + \lambda^2 \mathbb{E}U^2 \mathbb{E}S) \\
 &\quad \left. + c_S^2 \lambda (1 + (\lambda \mu \mathbb{E}U^2/2 + 1/\rho) + 2\mu^2 \mathbb{E}S^2) \right) + \frac{1}{2} c_S^2 \lambda/n.
 \end{aligned}$$

We prove this theorem at the end of this section, after presenting the BAR and using it to derive the diffusion generator. The only term we left unbounded in Theorem 4 is the boundary term

$$\frac{1}{n} \sum_{i=1}^n \mathbb{E} \left(1(Q_i = 0) \left(\sum_{j=1}^n Q_j + \lambda R_a + \sum_{j=1}^n \mu R_{s,j} \right) \right).$$

Based on the difficulty of bounding $\mathbb{E}R_a 1(X = 0)$ in the $G/G/1$ workload model, we anticipate the bounding $\mathbb{E}1(Q_i = 0)R_a$ and $\mathbb{E}1(Q_i = 0)R_{s,j}$ (for $j \neq i$) is also challenging, though we can get crude (sub- $(1 - \rho)$) bounds using Hölder's inequality as in (14). The

$$\mathbb{E}1(Q_i = 0) \left(\sum_{j=1}^n Q_j \right)$$

term can be interpreted as a state-space collapse (SSC) term that we expect to be of order $(1 - \rho)$; see (Hurtado-Lange and Maguluri 2022, pg. 375) for a treatment of this term in the discrete-time model. Since the JSQ routing scheme seeks to equalize queue lengths, we would expect that when $Q_i = 0$ (by Lemma 29 we know that this happens with probability $(1 - \rho)$), all other queue lengths are similar to Q_i . For recent work on the JSQ model with general interarrival and service times, see Dai et al. (2024).

Let us now present the BAR for the JSQ model. Let $A(t)$ and $U(t)$ be defined as in Section 2.1. Similarly, let $D_i(t)$ denote the number of departures from server i on $[0, t]$ and

if a departure occurs from server i at time t , we let $S_i(t)$ be the service time of the next customer to be served by that server. We also let $\Lambda_i(t)$ denote the (random) remaining time until a customer gets routed to server i . The following BAR for Z is established identically to Lemma 1.

LEMMA 20. *Initialize $Z(0) \sim Z$. If $\mathbb{E}|f(Z)| < \infty$ and, $Z(0)$ -almost surely, $df(Z(t))/dt$ exists for almost all t , then, provided that all expectations are well defined,*

$$\begin{aligned} & -\mathbb{E}(\partial_{r_a} f(Z)) - \sum_{i=1}^n \mathbb{E}(1(Q_i > 0) \partial_{r_{s,i}} f(Z)) \\ & + \mathbb{E} \int_0^1 \Delta f(Z(t-)) dA(t) + \sum_{i=1}^n \mathbb{E} \int_0^1 \Delta f(Z(t-)) dD_i(t) = 0. \end{aligned} \quad (64)$$

Define the compensated total customer count

$$\tilde{X}(t) = X(t) - \delta \lambda R_a(t) + \sum_{i=1}^n \delta \mu R_{s,i}(t) \quad \text{and} \quad \tilde{X} = X - \delta \lambda R_a + \sum_{i=1}^n \delta \mu R_{s,i}.$$

Specialized to $\tilde{X}$, the BAR (64) becomes

$$\begin{aligned} & \delta \mathbb{E} \left(\left(\lambda - \sum_{i=1}^n 1(Q_i > 0) \mu \right) f'(\tilde{X}) \right) + \mathbb{E} \int_0^1 \Delta f(\tilde{X}(t-)) dA(t) \\ & + \sum_{i=1}^n \mathbb{E} \int_0^1 \Delta f(\tilde{X}(t-)) dD_i(t) = 0. \end{aligned} \quad (65)$$

The following Proposition expands each of the terms in the BAR (65) and bounds all of the non-boundary error terms in terms of the first three moments of U and S . It is proved in Appendix F.

PROPOSITION 4. *If $f \in C^2(\mathbb{R})$ with $f''(x)$ absolutely continuous and $\|f''\|, \|f'''\| < \infty$, then, provided that all expectations are well defined,*

$$\delta \mathbb{E} \left(\left(\lambda - \sum_{i=1}^n 1(Q_i > 0) \mu \right) f'(\tilde{X}) \right) = -n\mu\delta^2 \mathbb{E}f'(X) + n\mu\delta^2 f'(0) + \epsilon_0(f), \quad (66)$$

$$\mathbb{E} \int_0^1 \Delta f(\tilde{X}(t-)) dA(t) = \frac{1}{2} \delta^2 \lambda c_U^2 \mathbb{E}f''(X) + \epsilon_A(f), \quad (67)$$

$$\mathbb{E} \int_0^1 \Delta f(\tilde{X}(t-)) dD_i(t) = \frac{1}{2} \delta^2 \mu c_S^2 \mathbb{E}f''(X) + \epsilon_{D,i}(f), \quad (68)$$

where

$$\begin{aligned}
 |\epsilon_0(f)| &\leq \frac{1}{2}n\mu\delta^3\|f''\|(\lambda^2\mathbb{E}U^2/2 + (\lambda\mu\mathbb{E}S^2/2 + \mu\delta\mathbb{E}S)) \\
 &\quad + \delta^2\mu\|f''\|\sum_{i=1}^n\mathbb{E}\left(1(Q_i=0)\left(\sum_{j=1}^nQ_j + \lambda R_a + \sum_{j=1}^n\mu R_{s,j}\right)\right), \\
 |\epsilon_A(f)| &\leq \frac{1}{2}\delta^3\|f'''\|\left(\frac{1}{3}\lambda\mathbb{E}|1-\lambda U|^3 + c_U^2n\mu\lambda(\rho(\lambda/n)\mathbb{E}S^2/2 + \delta\mathbb{E}S + \lambda\mathbb{E}U^2/2)\right. \\
 &\quad \left.+ \lambda c_U^2(1+n+\lambda\mu\mathbb{E}U^2 + \mu^2\mathbb{E}S^2)\right), \\
 |\epsilon_{D,i}(f)| &\leq \frac{1}{2}\delta^3\|f'''\|\left(\frac{1}{3}\frac{\lambda}{n}\mathbb{E}|1-\mu S|^3\right. \\
 &\quad + c_S^2((\lambda^2\mu\rho + 2(n-1)\mu^2\rho\lambda)\mathbb{E}S^2/2n + \delta(\lambda\mu + 2(n-1)\mu^2)\mathbb{E}S + \frac{1}{2}\lambda^2\mu\mathbb{E}U^2) \\
 &\quad + \mu c_S^2n\rho + \mu c_S^2(\lambda/n)(\mathbb{E}S + \lambda\mathbb{E}S^2 + \lambda^2\mathbb{E}U^2\mathbb{E}S) \\
 &\quad \left.+ c_S^2\lambda(1 + (\lambda\mu\mathbb{E}U^2/2 + 1/\rho) + 2\mu^2\mathbb{E}S^2)\right) + \frac{1}{2}\delta^3\|f''\|c_S^2\mu\lambda/n.
 \end{aligned}$$

The expansions in (66)–(68) suggest the diffusion generator

$$G_Y f(x) = -n\mu\delta^2 f'(x) + \frac{1}{2}\delta^2(\lambda c_U^2 + n\mu c_S^2)f''(x) + n\mu\delta^2 f'(0), \quad x \geq 0,$$

which, like in Section 2, corresponds to the exponential distribution with mean $(\rho c_U^2 + c_S^2)/2$. We now prove the main result of this section.

Proof of Theorem 4 Fix $h \in \text{Lip}(1)$. Since $\mathbb{E}X^2 < \infty$, we can verify that $f_h(\tilde{X})$ satisfies the conditions of Lemma 20 in the same way as we did in the proof of Theorem 1. For the Stein factor bounds, we use Lemma 4 with $\theta = n\mu\delta^2$ and $\sigma^2 = \delta^2(\lambda c_U^2 + n\mu c_S^2)/2$ there to get

$$\|f_h''\| \leq \frac{1}{n\mu\delta^2} \quad \text{and} \quad \|f_h'''\| \leq \frac{4}{\delta^2(\lambda c_U^2 + n\mu c_S^2)}.$$

Combining these with the bounds on $|\epsilon_0(f_h)|$, $|\epsilon_A(f_h)|$, and $|\epsilon_{D,i}(f_h)|$ in Proposition 4 concludes the proof. $\square$

7. The tandem queue

We now illustrate our methodology on a multidimensional system—the tandem queue. Our approach also extends naturally to generalized Jackson networks. In higher dimensions, the

mechanics are unchanged: we derive the diffusion generator from the BAR and decompose the approximation error into interior and boundary contributions. The added difficulty lies in establishing Stein factor bounds for the multidimensional RBMs; bounds are not known even for the two-dimensional RBM approximating the tandem queue.

Consider two first-come-first-served single-server stations in tandem. Customers arrive to station one according to a renewal process with generic interarrival time U ; after service at station one they move to station two and depart after completing service at station two. Service times at station i are i.i.d. S_i , independent of the arrival process and of service times at the other station. Assume $\mathbb{E}U^3 < \infty$, $\mathbb{E}S_1^3, \mathbb{E}S_2^3 < \infty$, and no simultaneous events almost surely. Set

$$\begin{aligned}\lambda &= 1/\mathbb{E}U, \quad \mu_i = 1/\mathbb{E}S_i, \quad \rho_i = \lambda/\mu_i, \\ c_U^2 &= \lambda^2 \text{Var}(U), \quad c_{S,i}^2 = \mu_i^2 \text{Var}(S_i).\end{aligned}$$

For $t \geq 0$, let $Q_i(t)$ be the queue length at station i , $R_a(t)$ the residual interarrival time, and $R_{s,i}(t)$ the remaining service at station i (if $Q_i(t) = 0$, the service time of the next customer). Define $X(t) = (\delta_1 Q_1(t), \delta_2 Q_2(t))$ with $\delta_i = 1 - \rho_i$, $R_s(t) = (R_{s,1}(t), R_{s,2}(t))$, and $Z(t) = (X(t), R_a(t), R_s(t))$. Assume $\{Z(t)\}$ is positive Harris recurrent (cf. Dai and Meyn (1995)) and let $Z = (X, R_a, R_s)$ have the stationary distribution. Let $A(t)$ and $D_i(t)$ denote arrivals and station- i departures on $[0, t]$.

The procedure of applying the generator approach to the tandem queue is similar to the three examples already considered. We therefore highlight only the key steps, leaving the formal derivations to the interested reader. For sufficiently regular functions, the BAR for Z is

$$\begin{aligned}& -\mathbb{E}(\partial_{r_a} f(Z)) - \sum_{i=1}^2 \mathbb{E}(1(X_i > 0) \partial_{r_{s,i}} f(Z)) \\ & + \mathbb{E} \int_0^1 \Delta f(Z(t-)) dA(t) + \sum_{i=1}^2 \mathbb{E} \int_0^1 \Delta f(Z(t-)) dD_i(t) = 0.\end{aligned}\tag{69}$$

Let $\tilde{X}$ have the stationary distribution of the compensated queue-length vector $\tilde{X}(t) = (\tilde{X}_1(t), \tilde{X}_2(t))$, where

$$\tilde{X}_1(t) = X_1(t) - \delta_1 \lambda R_a(t) + \delta_1 \mu_1 R_{s,1}(t), \quad \tilde{X}_2(t) = X_2(t) - \delta_2 \mu_1 R_{s,1}(t) + \delta_2 \mu_2 R_{s,2}(t).$$

The BAR for $\tilde{X}$ is

$$\begin{aligned} & \delta_1 \mathbb{E}((\lambda - \mu_1 1(Q_1 > 0)) \partial_{x_1} f(\tilde{X})) + \delta_2 \mathbb{E}((\mu_1 1(Q_1 > 0) - \mu_2 1(Q_2 > 0)) \partial_{x_2} f(\tilde{X})) \\ & + \mathbb{E} \int_0^1 \Delta f(\tilde{X}(t-)) dA(t) + \sum_{i=1}^2 \mathbb{E} \int_0^1 \Delta f(\tilde{X}(t-)) dD_i(t) = 0, \end{aligned}$$

Omitting the higher-order error terms, it follows that

$$\begin{aligned} \mathbb{E} \int_0^1 \Delta f(\tilde{X}(t-)) dA(t) & \approx \frac{1}{2} \delta_1^2 \lambda c_U^2 \mathbb{E} \partial_{x_1}^2 f(X), \\ \mathbb{E} \int_0^1 \Delta f(\tilde{X}(t-)) dD_1(t) & \approx \frac{1}{2} \mu_1 c_{S,1}^2 \mathbb{E} (\delta_1^2 \partial_{x_1}^2 f(X) - 2\delta_1 \delta_2 \partial_{x_1} \partial_{x_2} f(X) + \delta_2^2 \partial_{x_2}^2 f(X)), \\ \mathbb{E} \int_0^1 \Delta f(\tilde{X}(t-)) dD_2(t) & \approx \frac{1}{2} \delta_2^2 \mu_2 c_{S,2}^2 \mathbb{E} \partial_{x_2}^2 f(X). \end{aligned}$$

As in our prior examples, the omitted error terms can be bounded using the first three moments of the primitives as well as the second- and third-order derivatives of $f(x)$. Furthermore, omitting the boundary error terms, we have

$$\begin{aligned} & \delta_1 \mathbb{E}((\lambda - \mu_1 1(Q_1 > 0)) \partial_{x_1} f(\tilde{X})) + \delta_2 \mathbb{E}((\mu_1 1(Q_1 > 0) - \mu_2 1(Q_2 > 0)) \partial_{x_2} f(\tilde{X})) \\ & \approx -\mu_1 \delta_1^2 \mathbb{E} \partial_{x_1} f(X) + \delta_2 (\mu_1 \delta_1 - \mu_2 \delta_2) \mathbb{E} \partial_{x_2} f(X) \\ & + \mu_1 (\delta_1 1(Q_1 = 0) \partial_{x_1} f(X) - \delta_2 1(Q_1 = 0) \partial_{x_2} f(X)) + \mu_2 \delta_2 1(Q_2 = 0) \partial_{x_2} f(X). \end{aligned}$$

The second line on the right-hand side is tied to the reflection structure of the approximating RBM, which we introduce shortly. It is similar to the $f'(0)$ found in (8) of Proposition 1 and is not simply an error term.

Our expansions suggest the diffusion generator

$$\begin{aligned} G_Y f(x) = & -\mu_1 \delta_1^2 \partial_{x_1} f(x) + \delta_2 (\mu_1 \delta_1 - \mu_2 \delta_2) \partial_{x_2} f(x) + \frac{1}{2} \delta_1^2 (\lambda c_U^2 + \mu_1 c_{S,1}^2) \partial_{x_1}^2 f(x) \\ & - \delta_1 \delta_2 \mu_1 c_{S,1}^2 \partial_{x_1} \partial_{x_2} f(x) + \frac{1}{2} \delta_2^2 (\mu_1 c_{S,1}^2 + \mu_2 c_{S,2}^2) \partial_{x_2}^2 f(x), \\ & + \mu_1 (1(x_1 = 0) (\delta_1 \partial_{x_1} f(x) - \delta_2 \partial_{x_2} f(x)) + \mu_2 \delta_2 1(x_2 = 0) \partial_{x_2} f(x), \quad x \in \mathbb{R}_+^2, \end{aligned}$$

which corresponds to a two-dimensional RBM $\{Y(t) \in \mathbb{R}_+^2 : t \geq 0\}$ on the nonnegative orthant defined as follows. Set

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \lambda c_U^2 + \mu_1 c_{S,1}^2 & -\mu_1 c_{S,1}^2 \\ -\mu_1 c_{S,1}^2 & \mu_1 c_{S,1}^2 + \mu_2 c_{S,2}^2 \end{pmatrix}, \quad R = \begin{pmatrix} 1 & 0 \\ -1 & 1 \end{pmatrix},$$

let $\{\xi(t) : t \geq 0\}$ be a two-dimensional Brownian motion with drift $b = -R\mu$ and covariance matrix Σ , and let

$$Y(t) = \begin{pmatrix} \delta_1 & 0 \\ 0 & \delta_2 \end{pmatrix} \tilde{Y}(t),$$

where

$$\tilde{Y}(t) = \xi(t) + RI(t), \quad t \geq 0,$$

and $I : R \rightarrow \mathbb{R}^2$ is the unique nondecreasing process with $I(0) = 0$ and

$$\int_0^\infty \tilde{Y}_i(t) dI_i(t) = 0, \quad i = 1, 2.$$

For the existence and uniqueness of $\{\tilde{Y}(t) : t \geq 0\}$ see [Harrison and Reiman \(1981\)](#).

The final ingredient is the Poisson equation. Fix $h : \mathbb{R}^2 \rightarrow \mathbb{R}$ with $\mathbb{E}|h(Y)| < \infty$, where Y has the stationary distribution of $\{Y(t) : t \geq 0\}$, and consider

$$\begin{aligned} G_Y f_h(x) &= \mathbb{E}h(Y) - h(x), & x \in \mathbb{R}_+^2, \\ \delta_1 \partial_{x_1} f(x) - \delta_2 \partial_{x_2} f(x) &= 0, & x = (0, x_2) \in \mathbb{R}_+^2, \end{aligned} \tag{70}$$

$$\partial_{x_2} f(x) = 0, \quad x = (x_1, 0) \in \mathbb{R}_+^2. \tag{71}$$

This is known as an oblique derivative problem ([Lieberman 2013](#)) with (70) and (71) arising from the reflection structure of the RBM, which is driven by the matrix R . To bound the diffusion approximation error for the tandem queue, we require bounds on the partial derivatives, up to the third order, of $f_h(x)$. These Stein factor bounds remain an open problem.

Appendix A: Proving Proposition 2

We require several auxiliary lemmas.

LEMMA 21. *For any $h \in \text{Lip}(1)$ and $M > 0$,*

$$\partial_x F_h^M(z) = \mathbb{E}_z \int_0^{B_0 \wedge M} h'(X(t)) dt, \quad z = (x, r_a) \in \mathbb{S}.$$

Proof of Lemma 21 The proof is identical to that of Lemma 5; see Appendix B. □

The next three lemmas are proved in Appendix A.1.

LEMMA 22. For any $h \in \text{Lip}(1)$ and almost all $M > 0$,

$$-\delta \partial_x F_h^M(z) - \partial_{r_a} F_h^M(z) = \mathbb{E}_z h(X(M)) - h(x), \quad z = (x, r_a) \in \mathbb{S}.$$

In particular, $\partial_{r_a} F_h^M(z)$ is well defined for $z \in \mathbb{S}$.

LEMMA 23. For any differentiable $f: \mathbb{R}_+ \rightarrow \mathbb{R}$ with $\mathbb{E}|f(U)|, \mathbb{E}|f(R_a)|$, and $\mathbb{E}|f'(R_a)| < \infty$,

$$\mathbb{E}f'(R_a) = \frac{1}{\mathbb{E}U} (\mathbb{E}f(U) - \lim_{\epsilon \rightarrow 0} f(\epsilon)).$$

LEMMA 24. For any $h \in \text{Lip}(1)$ and $M > 0$,

$$\lim_{\epsilon \rightarrow 0} F_h^M(x, \epsilon) = \mathbb{E}F_h^M(x + \delta S, U).$$

Proof of Proposition 2 For almost all $M > 0$, Lemma 22 says that

$$-\delta \partial_x F_h^M(z) - \partial_{r_a} F_h^M(z) = \mathbb{E}_z h(X(M)) - h(x), \quad z \in \mathbb{S}. \quad (72)$$

Observe that $f(r_a) = F_h^M(x, r_a)$ satisfies the conditions of Lemma 23. Indeed, $|\mathbb{E}f(U)|, |\mathbb{E}f(R_a)| < \infty$ because $M < \infty$. Furthermore, $|\mathbb{E}f'(R_a)| < \infty$ follows from the expression for $\partial_{r_a} F_h^M(z)$ in Lemma 22, together with the observation that $|\partial_x F_h^M(z)| \leq M$, which follows from Lemma 21. Setting $r_a = R_a$ in (72) and taking expected values yields

$$\begin{aligned} \mathbb{E}(\mathbb{E}_{z, R_a} h(X(M)) - h(x)) &= -\delta \mathbb{E} \partial_x F_h^M(x, R_a) - \mathbb{E} \partial_{r_a} F_h^M(x, R_a) \\ &= -\delta \mathbb{E} \partial_x F_h^M(x, R_a) - \lambda \left(\mathbb{E} F_h^M(x, U) - \lim_{\epsilon \rightarrow 0} F_h^M(x, \epsilon) \right) \\ &= -\delta \mathbb{E} \partial_x F_h^M(x, R_a) - \lambda \left(\mathbb{E} F_h^M(x, U) - \mathbb{E} F_h^M(x + \delta S, U) \right), \end{aligned}$$

where the second and third equalities follow from Lemmas 23 and 24, respectively. $\square$

A.1. Auxiliary Lemma Proofs

Proof of Lemma 22 We define $\tilde{h}(x) = h(x) - \mathbb{E}h(X)$ for convenience, in which case

$$F_h^M(z) = \int_0^M \mathbb{E}_z \tilde{h}(X(t)) dt, \quad z \in \mathbb{S}.$$

Our goal is to prove that

$$-\delta \partial_x F_h^M(z) - \partial_{r_a} F_h^M(z) = \mathbb{E}_z h(X(M)) - h(x), \quad z = (x, r_a) \in \mathbb{S}. \quad (73)$$

Fix $z = (x, r_a) \in \mathbb{S}$ and suppose first that $x = 0$. On one hand,

$$F_h^{M+\epsilon}(0, r_a + \epsilon) = F_h^M(0, r_a + \epsilon) + \int_M^{M+\epsilon} \mathbb{E}_{0, r_a + \epsilon} \tilde{h}(X(t)) dt,$$

and on the other,

$$F_h^{M+\epsilon}(0, r_a + \epsilon) = \int_0^\epsilon \mathbb{E}_{0, r_a + \epsilon} \tilde{h}(X(t)) dt + \int_0^M \mathbb{E}_{0, r_a} \tilde{h}(X(t)) dt = \epsilon \tilde{h}(0) + F_h^M(0, r_a),$$

Equating the two expressions and dividing both sides by ϵ yields

$$\lim_{\epsilon \rightarrow 0} \frac{1}{\epsilon} (F_h^M(0, r_a + \epsilon) - F_h^M(0, r_a)) = \tilde{h}(0) - \lim_{\epsilon \rightarrow 0} \frac{1}{\epsilon} \int_M^{M+\epsilon} \mathbb{E}_{0, r_a + \epsilon} \tilde{h}(X(t)) dt.$$

The left-hand side equals $\delta \partial_x F_h^M(z) + \partial_{r_a} F_h^M(z) = \partial_{r_a} F_h^M(z)$, since $\partial_x F_h^M(0, r_a) = 0$ by Lemma 21.

Thus, to prove (73) when $x = 0$, it suffices to show that

$$\begin{aligned} & \frac{1}{\epsilon} \int_M^{M+\epsilon} (\mathbb{E}_{0, r_a + \epsilon} \tilde{h}(X(t)) - \mathbb{E}_{0, r_a} \tilde{h}(X(t))) dt \\ &= \frac{1}{\epsilon} \int_M^{M+\epsilon} \mathbb{E}_{0, r_a} (\tilde{h}(X(t-\epsilon)) - \tilde{h}(X(t))) dt \rightarrow 0 \end{aligned} \quad (74)$$

as $\epsilon \rightarrow 0$, which implies that

$$\lim_{\epsilon \rightarrow 0} \frac{1}{\epsilon} \int_M^{M+\epsilon} \mathbb{E}_{0, r_a + \epsilon} \tilde{h}(X(t)) dt = \lim_{\epsilon \rightarrow 0} \frac{1}{\epsilon} \int_M^{M+\epsilon} \mathbb{E}_{0, r_a} \tilde{h}(X(t)) dt = \mathbb{E}_{0, r_a} \tilde{h}(X(M)).$$

Observe that $|X(t-\epsilon) - X(t)|$ is bounded by the workload processed during $[t-\epsilon, t]$, which is at most $\delta\epsilon$, plus any new work that arrives during $[t-\epsilon, t]$. Letting $A([t_1, t_2])$ denote the number of customers arriving during $[t_1, t_2]$, Wald's identity says that the expected workload to arrive during $[t_1, t_2]$ equals $\mathbb{E}SE A([t_1, t_2])$. Thus, to prove (74), we observe that for any $h \in \text{Lip}(1)$ and for all $t \in [M, M+\epsilon]$,

$$\begin{aligned} \mathbb{E}_{0, r_a} |\tilde{h}(X(t-\epsilon)) - \tilde{h}(X(t))| &\leq \mathbb{E}_{0, r_a} |X(t-\epsilon) - X(t)| \\ &\leq \delta\epsilon + \mathbb{E}_{0, r_a} (\delta \mathbb{E}SE(A([t-\epsilon, t]))) \\ &\leq \delta\epsilon + \delta \mathbb{E}SE_{0, r_a}(A([M-\epsilon, M+\epsilon])). \end{aligned}$$

It suffices to argue that the right-hand side goes to zero as $\epsilon \rightarrow 0$. By the dominated convergence theorem,

$$\lim_{\epsilon \rightarrow 0} \mathbb{E}_{0, r_a} (A([M-\epsilon, M+\epsilon])) = \mathbb{E}_{0, r_a} (A([M, M])),$$

which equals the expected number of arrivals at time M . The right-hand side may be non-zero if the distribution of U has point masses. However, since the number of point masses is at most countable, then $\mathbb{E}_{0, r_a} (A([M, M])) = 0$ for all but at most countably many M . This proves (73) when $x = 0$.

The case when $x > 0$ follows similarly. We repeat the arguments, highlighting the differences. Given $z = (x, r_a)$, fix $\epsilon < x/\delta$. Then

$$F_h^{M+\epsilon}(x, r_a + \epsilon) = F_h^M(x, r_a + \epsilon) + \int_M^{M+\epsilon} \mathbb{E}_{x, r_a + \epsilon} \tilde{h}(X(t)) dt$$

and

$$F_h^{M+\epsilon}(x, r_a + \epsilon) = \int_0^\epsilon \mathbb{E}_{x, r_a + \epsilon} \tilde{h}(X(t)) dt + F_h^M(x - \delta\epsilon, r_a).$$

Equating both expressions, subtracting $F_h^M(x, r_a)$ from each side, and dividing by ϵ yields

$$\begin{aligned} & \frac{1}{\epsilon} (F_h^M(x, r_a + \epsilon) - F_h^M(x, r_a)) \\ &= \frac{1}{\epsilon} (F_h^M(x - \delta\epsilon, r_a) - F_h^M(x, r_a)) + \frac{1}{\epsilon} \int_0^\epsilon \mathbb{E}_{x, r_a + \epsilon} \tilde{h}(X(t)) dt - \frac{1}{\epsilon} \int_M^{M+\epsilon} \mathbb{E}_{x, r_a + \epsilon} \tilde{h}(X(t)) dt. \end{aligned}$$

We now argue that each of the terms on the right-hand side has a well-defined limit as $\epsilon \rightarrow 0$, implying that the left-hand side converges to $\partial_x F_h^M(z)$, which is itself well defined. The first term on the right-hand side converges to $-\partial_x F_h^M(z)$, which we know exists for all $z \in \mathbb{S}$ by Lemma 21. Furthermore,

$$\lim_{\epsilon \rightarrow 0} \frac{1}{\epsilon} \int_0^\epsilon \mathbb{E}_{x, r_a + \epsilon} \tilde{h}(X(t)) dt = \tilde{h}(x) \quad \text{and} \quad \lim_{\epsilon \rightarrow 0} \frac{1}{\epsilon} \int_M^{M+\epsilon} \mathbb{E}_{x, r_a + \epsilon} \tilde{h}(X(t)) dt = \mathbb{E}_{x, r_a} \tilde{h}(X(M)).$$

The first equality is straightforward because no arrival occurs during $[0, \epsilon]$, while the second equality is proved the same way as (74). $\square$

Proof of Lemma 23 For simplicity, we first assume that $\mathbb{P}(U > 0) = 1$. Initialize $R_a(0) \sim R_a$, let $\tau_0 = 0$ and τ_m be the time of the m th arrival, and let U_m , $m \geq 1$ be the interarrival times with $U_1 = R_a(0)$ and U_m i.i.d. U for $m \geq 2$. Then $\tau_{m+1} = \tau_m + U_{m+1}$ for $m \geq 0$, and by isolating times when jumps occur, one can verify that for any $t > 0$,

$$\begin{aligned} 0 &= \frac{1}{t} \mathbb{E}(f(R_a(t)) - f(R_a(0))) \\ &= \frac{1}{t} \mathbb{E} \int_0^t -f'(R_a(s)) ds + \frac{1}{t} \mathbb{E} \sum_{m=1}^{\infty} 1(\tau_m \leq t) (f(R_a(\tau_m)) - f(0)). \end{aligned}$$

Initialize $R_a(0) \sim R_a$, where R_a is defined by (18). Since $\mathbb{E}|f'(R_a)| < \infty$, the Fubini-Tonelli theorem says that

$$\frac{1}{t} \mathbb{E} \int_0^t -f'(R_a(s)) ds = -\frac{1}{t} \int_0^t \mathbb{E} f'(R_a(0)) ds = -\mathbb{E} f'(R_a).$$

To conclude, we argue that

$$\lim_{t \rightarrow \infty} \frac{1}{t} \mathbb{E} \sum_{m=1}^{\infty} 1(\tau_m \leq t) (f(R_a(\tau_m)) - f(0)) = \frac{1}{\mathbb{E}U} (\mathbb{E}f(U) - f(0)).$$

Since $\tau_m = \sum_{i=1}^m U_i$ and $R_a(\tau_m) = U_{m+1}$, and it follows that

$$\begin{aligned} & \frac{1}{t} \mathbb{E} \sum_{m=1}^{\infty} 1(\tau_m \leq t) (f(R_a(\tau_m)) - f(0)) \\ &= \frac{1}{t} \mathbb{E} \left(1(U_1 \leq t) (f(U_2) - f(0)) \right) + \frac{1}{t} \mathbb{E} \sum_{m=2}^{\infty} 1(U_1 + \dots + U_m \leq t) (f(U_{m+1}) - f(0)) \\ &= \frac{1}{t} \mathbb{P}(U_1 \leq t) \mathbb{E}(f(U_2) - f(0)) + \frac{1}{t} \mathbb{E} \sum_{m=2}^{\infty} 1(U_1 + \dots + U_m \leq t) (f(U_{m+1}) - f(0)). \end{aligned} \quad (75)$$

Since $U_1 \sim R_a$, we know by (18) that the first term converges to $(1/\mathbb{E}U)\mathbb{E}(f(U) - f(0))$ as $t \rightarrow 0$, and it remains to show that the second term converges to zero as $t \rightarrow 0$. Note that

$$\begin{aligned} & \frac{1}{t} \mathbb{E} \sum_{m=2}^{\infty} 1(U_1 + \dots + U_m \leq t) (f(U_{m+1}) - f(0)) \\ &= \frac{1}{t} \sum_{m=2}^{\infty} \mathbb{P}(U_1 \leq t, U_1 + U_2 \leq t, \dots, U_1 + \dots + U_m \leq t) \mathbb{E}(f(U_{m+1}) - f(0)) \\ &\leq \mathbb{E}(f(U) - f(0)) \frac{1}{t} \sum_{m=2}^{\infty} \mathbb{P}(U_1 \leq t, \dots, U_m \leq t) \\ &= \mathbb{E}(f(U) - f(0)) \frac{1}{t} \sum_{m=2}^{\infty} \mathbb{P}(U_1 \leq t) \mathbb{P}(U \leq t)^{m-1} \\ &= \mathbb{E}(f(U) - f(0)) \frac{1}{t} \mathbb{P}(U_1 \leq t) \frac{\mathbb{P}(U \leq t)}{1 - \mathbb{P}(U \leq t)} \rightarrow 0 \quad \text{as } t \rightarrow 0, \end{aligned}$$

where the first equality is by the independence of the U_m , $m \geq 1$, and the convergence to zero follows from $\mathbb{P}(U = 0) = 0$. The case $0 < \mathbb{P}(U > 0) < 1$ follows similarly. Since $f(U_{m+1}) - f(0) = 0$ in (75) if $U_{m+1} = 0$, we only need to consider those jump times τ_m when $U_{m+1} \neq 0$. In the last display, we would then replace $\mathbb{P}(U \leq t)$ by $\mathbb{P}(U \leq t, U > 0)$. $\square$

Proof of Lemma 24 Define $\tilde{h}(x) = h(x) - \mathbb{E}h(X)$ and consider first the case when $x = 0$. Then

$$\begin{aligned} F_h^M(0, \epsilon) &= \int_0^\epsilon \mathbb{E}_{0, \epsilon} \tilde{h}(X(t)) dt + \int_\epsilon^M \mathbb{E}_{0, \epsilon} \tilde{h}(X(t)) dt \\ &= \int_0^\epsilon \mathbb{E}_{0, \epsilon} \tilde{h}(X(t)) dt + \mathbb{E} \int_0^{M-\epsilon} \mathbb{E}_{\delta S, U} \tilde{h}(X(t)) dt, \end{aligned}$$

where the outer expectation is with respect to U and S . Taking $\epsilon \rightarrow 0$, the left-hand side converges to $\lim_{\epsilon \rightarrow 0} F_h^M(0, \epsilon)$ while the right-hand side converges to

$$\lim_{\epsilon \rightarrow 0} \mathbb{E} \int_0^{M-\epsilon} \mathbb{E}_{\delta S, U} \tilde{h}(X(t)) dt = \mathbb{E} \int_0^M \mathbb{E}_{\delta S, U} \tilde{h}(X(t)) dt - \lim_{\epsilon \rightarrow 0} \mathbb{E} \int_{M-\epsilon}^M \mathbb{E}_{\delta S, U} \tilde{h}(X(t)) dt.$$

The first term equals $\mathbb{E}F_h^M(\delta S, U)$ while the second term is zero because $h \in \text{Lip}(1)$. Now suppose that $x > 0$ and take $\epsilon < x/\delta$. Arguing as before,

$$F_h^M(x, \epsilon) = \int_0^\epsilon \mathbb{E}_{x, \epsilon} \tilde{h}(X(t)) dt + \mathbb{E}F_h^{M-\epsilon}(x - \delta\epsilon + \delta S, U).$$

To conclude, we use the fundamental theorem of calculus to write

$$\mathbb{E}F_h^{M-\epsilon}(x - \delta\epsilon + \delta S, U) = \mathbb{E}F_h^{M-\epsilon}(x + \delta S, U) + \mathbb{E} \int_0^{-\delta\epsilon} \partial_x F_h^{M-\epsilon}(x + v + \delta S, U) dv.$$

The second term on the right-hand side converges to zero because $|\partial_x F_h^{M-\epsilon}(z)| \leq M$ due to Lemma 21. The first term converges to $\mathbb{E}F_h^M(x + \delta S, U)$ because

$$\lim_{\epsilon \rightarrow 0} \mathbb{E} \int_0^{M-\epsilon} \mathbb{E}_{x+\delta S, U} \tilde{h}(X(t)) dt = \mathbb{E} \int_0^M \mathbb{E}_{x+\delta S, U} \tilde{h}(X(t)) dt - \lim_{\epsilon \rightarrow 0} \mathbb{E} \int_{M-\epsilon}^M \mathbb{E}_{x+\delta S, U} \tilde{h}(X(t)) dt,$$

and the second term equals zero since $h \in \text{Lip}(1)$. $\square$

Appendix B: Proofs of Lemmas 5 and 6

We recall the synchronous coupling $\{Z^{(\epsilon)}(t) : t \geq 0\}$ defined in Section 3.2.

Proof of Lemma 5 First, observe that

$$\begin{aligned} & \frac{1}{\epsilon} \mathbb{E} \left(\int_0^\infty (\mathbb{E}_{x+\epsilon, T} h(X(t)) - \mathbb{E}_{x, T} h(X(t))) dt \right) \\ &= \frac{1}{\epsilon} \mathbb{E} \left(\mathbb{E}_{x, T} \int_0^{B_0} (h(X^{(\epsilon)}(t)) - h(X(t))) dt \right) + \frac{1}{\epsilon} \mathbb{E} \left(\mathbb{E}_{x, T} \int_{B_0}^{B_0^{(\epsilon)}} (h(X^{(\epsilon)}(t)) - h(X(t))) dt \right). \end{aligned}$$

Note that $|h(X^{(\epsilon)}(t)) - h(X(t))|/\epsilon \leq \|h'\| \leq 1$ and $B_0^{(\epsilon)} \rightarrow B_0$ as $\epsilon \rightarrow 0$. Also note that for all $\epsilon < 1$,

$$\mathbb{E}_{x, T} B_0^{(\epsilon)} = \mathbb{E}_{x+\epsilon, T} B_0 \leq \mathbb{E}_{x+1, T} B_0 \leq \mathbb{E}(\mathbb{E}_{x+1+\delta S, U} B_0) < \infty,$$

where the second-last inequality follows from the fact that the busy period starting at state $(x+1, T)$ is made longer if the next arrival happens immediately. The DCT then implies that

$$\begin{aligned} & \frac{1}{\epsilon} \mathbb{E} \left(\mathbb{E}_{x, T} \int_0^{B_0} (h(X^{(\epsilon)}(t)) - h(X(t))) dt \right) \rightarrow \mathbb{E} \left(\mathbb{E}_{x, T} \int_0^{B_0} h'(X(t)) dt \right), \\ & \frac{1}{\epsilon} \mathbb{E} \left(\mathbb{E}_{x, T} \int_{B_0}^{B_0^{(\epsilon)}} (h(X^{(\epsilon)}(t)) - h(X(t))) dt \right) \rightarrow 0. \end{aligned}$$

$\square$

Proof of Lemma 6 Fix $h \in \text{Lip}(1)$. To avoid confusion, we write E^X , E^{R_a} , and E^{X, R_a} to denote expectations with respect to X , R_a , and (X, R_a) , respectively. We first prove (29). Using the tower property of conditional expectation,

$$\begin{aligned} \mathbb{E}(\mathbb{E}_{x, R_a} h(X(M))) - \mathbb{E}h(X) &= \mathbb{E}^{R_a}(\mathbb{E}_{x, R_a} h(X(M))) - \mathbb{E}^{R_a} \left(\mathbb{E}^X [\mathbb{E}_{x, R_a} h(X(M)) | R_a] \right) \\ &= \mathbb{E}^{R_a} \left(\mathbb{E}^X [\mathbb{E}_{x, R_a} h(X(M)) - \mathbb{E}_{x, R_a} h(X(M)) | R_a] \right) \\ &= \mathbb{E}^{X, R_a} \left(\mathbb{E}_{x, R_a} h(X(M)) - \mathbb{E}_{x, R_a} h(X(M)) \right). \end{aligned}$$

Using our synchronous coupling defined in (24) and the fact that $h \in \text{Lip}(1)$, it follows that

$$|\mathbb{E}_{x,R_a} h(X(M)) - \mathbb{E}_{X,R_a} h(X(M))| \leq |x - X| \mathbb{P}_{x \vee X, R_a}(B_0 > M),$$

where the probability on the right-hand side corresponds to the probability that coupling does not occur by time M . Thus,

$$\lim_{M \rightarrow \infty} |\mathbb{E}(\mathbb{E}_{x,R_a} h(X(M))) - \mathbb{E}h(X)| \leq \lim_{M \rightarrow \infty} \mathbb{E}(|x - X| \mathbb{P}_{x \vee X, R_a}(B_0 > M)) = 0.$$

The last equality follows from the DCT because $\mathbb{E}X < \infty$, and because $\lim_{M \rightarrow \infty} \mathbb{P}_{x \vee x', r_a}(B_0 > M) = 0$ for any $x, x', r_a > 0$ by (19). To prove (30), one can reuse the arguments used to prove Lemma 5 to show that

$$\partial_x F_h^M(x, r_a) = \mathbb{E}_{x, r_a} \int_0^{B_0 \wedge M} h'(X(t)) dt \rightarrow \partial_x F_h(x, r_a) \quad \text{as } M \rightarrow \infty \text{ for all } (x, r_a) \in \mathbb{S},$$

and also that $\lim_{M \rightarrow \infty} \mathbb{E} \partial_x F_h^M(x, R_a) = \mathbb{E} \lim_{M \rightarrow \infty} \partial_x F_h^M(x, R_a)$. Lastly, we prove (31). Similar to the way we argued (25),

$$\begin{aligned} F_h^M(x + \epsilon, r_a) - F_h^M(x, r_a) &= \mathbb{E}_{x, r_a} \int_0^{B_0^{(\epsilon)} \wedge M} (h(X^{(\epsilon)}(t)) - h(X(t))) dt \\ &\rightarrow F_h(x + \epsilon, r_a) - F_h(x, r_a), \quad \text{as } M \rightarrow \infty \text{ for all } (x, r_a) \in \mathbb{S}. \end{aligned}$$

Let $\hat{h}(x) = x$ and observe that since $h \in \text{Lip}(1)$ and $X^{(\epsilon)}(t) \geq X(t)$, then

$$|F_h^M(x + \epsilon, r_a) - F_h^M(x, r_a)| \leq \mathbb{E}_{x, r_a} \int_0^{B_0^{(\epsilon)} \wedge M} (X^{(\epsilon)}(t) - X(t)) dt \leq F_{\hat{h}}(x + \epsilon, r_a) - F_{\hat{h}}(x, r_a).$$

It remains to show that $\mathbb{E}(F_{\hat{h}}(x + \delta S, U) - F_{\hat{h}}(x, U)) < \infty$, because then we can use the DCT to conclude (31). The finiteness of this expectation follows from

$$\begin{aligned} \mathbb{E}(F_{\hat{h}}(x + \delta S, U) - F_{\hat{h}}(x, U)) &= \lim_{M \rightarrow \infty} \mathbb{E}(F_{\hat{h}}^M(x + \delta S, U) - F_{\hat{h}}^M(x, U)) \\ &= \lim_{M \rightarrow \infty} \mathbb{E}(\mathbb{E}_{x, R_a} X(M) - x) + \lim_{M \rightarrow \infty} \delta \mathbb{E} \partial_x F_{\hat{h}}^M(x, R_a), \end{aligned}$$

where the first equality is due to the monotone convergence theorem, since $F_{\hat{h}}^M(x + \epsilon, r_a) - F_{\hat{h}}^M(x, r_a)$ is increasing in M and is nonnegative for any $(x, r_a) \in \mathbb{S}$, and the second equality is due to (22).

The right-hand side is finite by (29) and (30). $\square$

Appendix C: Section 3.3 Proofs

Proof of Lemma 7 Recall that $J(x, r_a) = -(x \wedge \delta r_a) + \delta S'$. It follows that

$$\begin{aligned} F_h(x + \delta s, r_a) - F_h(x, r_a) &= \int_0^\infty (\mathbb{E}_{x+\delta s, r_a} h(X(t)) - \mathbb{E}_{x, r_a} h(X(t))) dt \\ &= \int_0^{r_a} (h((x + \delta s - \delta t)^+) - h((x - \delta t)^+)) dt \\ &\quad + \mathbb{E} \left(F_h(x + \delta s + J(x + \delta s, r_a), U) - F_h(x + J(x, r_a), U) \right). \end{aligned}$$

To conclude, note that

$$\begin{aligned} &\mathbb{E} \left(F_h(x + \delta s + J(x + \delta s, r_a), U) - F_h(x + J(x, r_a), U) \right) \\ &= \mathbb{E} \left(F_h(x + \delta s + J(x, r_a), U) - F_h(x + J(x, r_a), U) \right) \\ &\quad + \mathbb{E} \left(F_h(x + \delta s + J(x + \delta s, r_a), U) - F_h(x + \delta s + J(x, r_a), U) \right). \end{aligned}$$

Using the fundamental theorem of calculus, together with Lemma 5, which shows that $\partial_x \mathbb{E} F_h(x + \delta S, U) = \mathbb{E} \partial_x F_h(x + \delta S, U)$, we arrive at

$$\begin{aligned} &\mathbb{E} \left(F_h(x + \delta s + J(x + \delta s, r_a), U) - F_h(x + \delta s + J(x, r_a), U) \right) \\ &= \mathbb{E} \left(\int_{-x \wedge (\delta r_a)}^{-(x + \delta s) \wedge (\delta r_a)} \partial_x F_h(x + \delta s + v + \delta S', U) dv \right) \\ &= \mathbb{E}^{S'} \left(\int_{-x \wedge (\delta r_a)}^{-(x + \delta s) \wedge (\delta r_a)} \mathbb{E}^U \partial_x F_h(x + \delta s + v + \delta S', U) dv \right) \\ &= \mathbb{E}^{S'} \left(\int_{-x \wedge (\delta r_a)}^{-(x + \delta s) \wedge (\delta r_a)} \partial_x \mathbb{E}^U F_h(x + \delta s + v + \delta S', U) dv \right). \end{aligned}$$

Interchanging $\mathbb{E}^U$ with the integral in the second equality is justified by the Fubini-Tonelli theorem because $\mathbb{E}^{S'} \mathbb{E}^U |\partial_x F_h(x + \delta S', U)| \leq \mathbb{E}^{S'} \mathbb{E}^U \mathbb{E}_{x+\delta S', U} B_0 < \infty$ for all $x \geq 0$ by Lemma 5 and (19). $\square$

C.1. Proving Lemma 8

We recall that $\bar{F}'_h(x) = \partial_x \mathbb{E} F_h(x, R_a)$ and that $\bar{F}''_h(x)$ and $\bar{F}'''_h(x)$ are assumed to exist. We recall (34), or

$$\begin{aligned} &\mathbb{E} h(X) - \mathbb{E} h(x + J(x, R'_a)) \\ &= -\delta \mathbb{E} \bar{F}'_h(x + J(x, R'_a)) + \lambda \mathbb{E} (F_h(x + \delta S, R'_a) - F_h(x, R'_a)) - \mathbb{E}(\epsilon(x, R'_a, S)), \end{aligned} \quad (76)$$

where $J(x, r_a) = -(x \wedge \delta r_a) + \delta S'$. The following lemma expands the first two terms on the right-hand side of (76). We prove it after proving Lemma 8.

LEMMA 25. For any $x \geq 0$,

$$\begin{aligned} \bar{F}'_h(x + J(x, R'_a)) &= \bar{F}'_h(x) + \delta(S' - R'_a)\bar{F}''_h(x) + 1(\delta R'_a < x) \int_0^{\delta(S' - R'_a)} \int_0^v \bar{F}'''_h(x + u) du dv \\ &\quad + 1(\delta R'_a \geq x) (\bar{F}'_h(\delta S') - \bar{F}'_h(x) - \delta(S' - R'_a)\bar{F}''_h(x)) \\ \mathbb{E}(F_h(x + \delta S, R'_a) - F_h(x, R'_a)) &= \delta \mathbb{E} S \bar{F}'_h(x) + \frac{1}{2} \delta^2 \mathbb{E} S^2 \bar{F}''_h(x) + \mathbb{E} \int_0^{\delta S} (\delta S - v) \int_0^v \bar{F}'''_h(x + u) du dv, \end{aligned}$$

Proof of Lemma 8 Recall that $\lambda \mathbb{E} S = \rho$. Combining Lemma 25 with (76) yields

$$\begin{aligned} &\mathbb{E} h(X) - \mathbb{E} h(x + J(x, R'_a)) \\ &= -\delta(\bar{F}'_h(x) + \delta \mathbb{E}(S' - R'_a) \bar{F}''_h(x)) + \lambda(\delta \mathbb{E} S \bar{F}'_h(x) + \frac{1}{2} \delta^2 \mathbb{E} S^2 \bar{F}''_h(x)) \\ &\quad - \delta \mathbb{E} \left(1(\delta R'_a \geq x) (\bar{F}'_h(\delta S') - \bar{F}'_h(x) - \delta(S' - R'_a) \bar{F}''_h(x)) \right) \\ &\quad - \delta \mathbb{E} \left(1(\delta R'_a < x) \int_0^{\delta(S' - R'_a)} \int_0^v \bar{F}'''_h(x + u) du dv \right) \\ &\quad + \lambda \mathbb{E} \int_0^{\delta S} (\delta S - v) \int_0^v \bar{F}'''_h(x + u) du dv. \end{aligned}$$

Using the facts that $\lambda \mathbb{E} S = \rho$, $\lambda \mathbb{E} U = 1$, and that $\mathbb{E} R'_a = \lambda \mathbb{E} U^2 / 2$, we see that the first line on the right-hand side equals

$$-\delta(1 - \rho) \bar{F}'_h(x) + \frac{1}{2} \delta^2 (\lambda \mathbb{E} S^2 - 2\lambda \mathbb{E} U \mathbb{E} S' + \lambda \mathbb{E} U^2) \bar{F}''_h(x).$$

Let us call this term $G_{Y_2} \bar{F}_h(x)$. Since $F'_h(0) = 0$ due to Lemma 5 (because $B_0 = 0$ if the initial workload $X(0) = 0$), our assumptions that $\mathbb{E} |F'_h(Y_2)|, \mathbb{E} |F''_h(Y_2)| < \infty$ and integration by parts yield $\mathbb{E} G_{Y_2} \bar{F}_h(Y_2) = 0$. $\square$

Proof of Lemma 25 The expression for $\bar{F}'_h(x + J(x, R'_a))$ follows from the facts that

$$\bar{F}'_h(x + J(x, R'_a)) = 1(\delta R'_a \geq x) \bar{F}'_h(\delta S') + 1(\delta R'_a < x) \bar{F}'_h(x - \delta R'_a + \delta S')$$

and, for all $x > \delta R'_a$,

$$\bar{F}'_h(x - \delta R'_a + \delta S') = \bar{F}'_h(x) + \delta(S' - R'_a) \bar{F}''_h(x) + \int_0^{\delta(S' - R'_a)} \int_0^v \bar{F}'''_h(x + u) du dv.$$

Next, we argue that

$$\mathbb{E}(F_h(x + \delta S, R'_a) - F_h(x, R'_a)) = \mathbb{E} \int_0^{\delta S} \bar{F}'_h(x + v) dv, \quad (77)$$

so that the expression for $\mathbb{E}(F_h(x + \delta S, R'_a) - F_h(x, R'_a))$ also follows from Taylor expansion of the integrand around x . To prove (77), note that

$$\mathbb{E}(F_h(x + \delta S, R'_a) - F_h(x, R'_a)) = \mathbb{E} \int_0^{\delta S} \partial_x F_h(x + v, R'_a) dv$$

$$\begin{aligned}
 &= \mathbb{E}^S \int_0^{\delta S} \mathbb{E}^{R'_a} \partial_x F_h(x+v, R'_a) dv \\
 &= \mathbb{E}^S \int_0^{\delta S} \partial_x \mathbb{E} F_h(x+v, R'_a) dv \\
 &= \mathbb{E} \int_0^{\delta S} \bar{F}'_h(x+v) dv.
 \end{aligned}$$

The first and second-last equalities follows from Lemma 5. Once we justify the interchange of the integral and expectation in the second equality using the Fubini-Tonelli theorem, (77) will follow. Let $\hat{h}(x) = x$. Using the form of $\partial_x F_h(x, r_a)$ from Lemma 5, it follows that for any $h \in \text{Lip}(1)$,

$$|\partial_x F_h(x, r_a)| \leq \mathbb{E}_{x, r_a} B_0 = \mathbb{E}_{x, r_a} \int_0^{B_0} \hat{h}'(X(t)) dt = \partial_x F_{\hat{h}}(x, r_a).$$

Thus,

$$\mathbb{E} \int_0^{\delta S} |\partial_x F_h(x+v, R'_a)| dv \leq \mathbb{E} \int_0^{\delta S} \partial_x F_{\hat{h}}(x+v, R'_a) dv = \mathbb{E}(F_{\hat{h}}(x+\delta S, R'_a) - F_{\hat{h}}(x, R'_a)),$$

and the right-hand side is finite because the right-hand side of (31) in Lemma 6 is finite. $\square$

Appendix D: Workload Stein factor bound proofs

For the entirety of this appendix we fix $h(x) = x$ and assume that $\bar{\eta} = \sup\{\eta(x) : x \geq 0\} < \infty$.

D.1. Second-order bounds (Proof of Lemma 13)

Recall that $G(x) = \mathbb{P}(U \leq x)$. The following auxiliary lemma is needed to prove Lemma 13.

LEMMA 26. For any $\epsilon > 0$ and $(x, r_a) \in \mathbb{S}$ with $r_a < x/\delta$,

$$\frac{1}{\epsilon} \mathbb{P}_{x, r_a}(R_a(B_0) < \epsilon/\delta) \leq \bar{\eta}/\delta, \quad (78)$$

$$\lim_{\epsilon \rightarrow 0} \frac{1}{\epsilon} \mathbb{P}_{x, r_a}(R_a(B_0) < \epsilon/\delta) = \frac{1}{\delta} \mathbb{E}_{x, r_a} \eta(\alpha(B_0)). \quad (79)$$

Proof of Lemma 26 Let U_n denote the interarrival time of the n th customer, let $W_0 = V(0)$, and let $W_n = V(U_1 + \dots + U_n)$ be the workload in the system right after the n th customer arrives, which includes the workload brought by the n th customer. Let

$$\sigma = \min\{n \geq 1 : U_n > W_{n-1}\}$$

be the number of customers served in the first busy period $[0, B_0]$. Now assuming that $Z(0) = (x, r_a) \in \mathbb{S}$ with $r_a < x/\delta$, it must be that $\sigma > 1$, because $W_0 = x/\delta$ and $U_1 = r_a$. Since $\{R_a(B_0) \leq \epsilon/\delta\} = \{U_\sigma \leq W_{\sigma-1} + \epsilon/\delta\}$, it follows that

$$\begin{aligned}
 \frac{1}{\epsilon} \mathbb{P}_{x, r_a}(R_a(B_0) \leq \epsilon/\delta) &= \frac{1}{\epsilon} \sum_{n=2}^{\infty} \mathbb{P}_{x, r_a}(U_n \leq W_{n-1} + \epsilon/\delta | \sigma = n) \mathbb{P}_{x, r_a}(\sigma = n) \\
 &= \frac{1}{\epsilon} \sum_{n=2}^{\infty} \mathbb{E}_{x, r_a} \left[\mathbb{P}_{x, r_a}(U_n \leq W_{n-1} + \epsilon/\delta | \sigma = n, W_{n-1}) \middle| \sigma = n \right] \mathbb{P}_{x, r_a}(\sigma = n).
 \end{aligned}$$

To proceed, note that $\{\sigma = n\} = \{U_1 \leq W_0, \dots, U_{n-1} \leq W_{n-2}, U_n > W_{n-1}\}$ for any $n \geq 1$, implying that for any $n \geq 2$,

$$\begin{aligned}
& \mathbb{P}_{x,r_a}(U_n \leq W_{n-1} + \epsilon/\delta \mid \sigma = n, W_{n-1}) \\
&= \mathbb{P}(U_n \leq W_{n-1} + \epsilon/\delta \mid W_0 = x/\delta, U_1 = r_a, \sigma = n, W_{n-1}) \\
&= \mathbb{P}(U_n \leq W_{n-1} + \epsilon/\delta \mid W_0 = x/\delta, U_1 = r_a, U_1 \leq W_0, \dots, U_{n-1} \leq W_{n-2}, U_n > W_{n-1}, W_{n-1}) \\
&= \mathbb{P}(U_n \leq W_{n-1} + \epsilon/\delta \mid U_n > W_{n-1}, W_{n-1}) \\
&= \mathbb{P}(U \leq W_{n-1} + \epsilon/\delta \mid U > W_{n-1}, W_{n-1}) \\
&= \frac{G(W_{n-1} + \epsilon/\delta) - G(W_{n-1})}{1 - G(W_{n-1})},
\end{aligned}$$

and therefore

$$\begin{aligned}
\frac{1}{\epsilon} \mathbb{P}_{x,r_a}(R_a(B_0) \leq \epsilon/\delta) &= \frac{1}{\epsilon} \sum_{n=1}^{\infty} \mathbb{E}_{x,r_a} \left[\frac{G(W_{n-1} + \epsilon/\delta) - G(W_{n-1})}{1 - G(W_{n-1})} \mid \sigma = n \right] \mathbb{P}_{x,r_a}(\sigma = n) \\
&= \frac{1}{\epsilon} \mathbb{E}_{x,r_a} \left[\frac{G(W_{\sigma-1} + \epsilon/\delta) - G(W_{\sigma-1})}{1 - G(W_{\sigma-1})} \right]. \tag{80}
\end{aligned}$$

To prove (78), observe that the right-hand side of (80) is bounded by $\bar{\eta}/\delta$ because by the mean value theorem,

$$\frac{G(w + \epsilon/\delta) - G(w)}{1 - G(w)} = \frac{\epsilon}{\delta} \frac{G'(\xi)}{1 - G(w)} = \frac{\epsilon}{\delta} \eta(\xi) \frac{1 - G(\xi)}{1 - G(w)} \leq \frac{\epsilon}{\delta} \bar{\eta}$$

for some $\xi \in [w, w + \epsilon/\delta]$, where the last inequality follows from $\xi \geq w$ and $\eta(x) \leq \bar{\eta}$. Once we observe that $W_{\sigma-1} = \alpha(B_0)$, then (79) follows from taking $\epsilon \rightarrow 0$ in (80) and applying the dominated convergence theorem. $\square$

Proof of Lemma 13 Fix $h(x) = x$, $x \geq 0$, and $\epsilon > 0$,

$$\begin{aligned}
& \frac{1}{\epsilon} (\partial_x \mathbb{E} F_h(x + \epsilon, T) - \partial_x \mathbb{E} F_h(x, T)) \\
&= \frac{1}{\epsilon} \mathbb{E} \left(\mathbb{E}_{x,T} \int_0^{B_0} (h'(X^{(\epsilon)}(t)) - h'(X(t))) dt \right) + \frac{1}{\epsilon} \mathbb{E} \left(\mathbb{E}_{x,T} \int_{B_0}^{B_0^{(\epsilon)}} h'(X^{(\epsilon)}(t)) dt \right) \\
&= \frac{1}{\epsilon} \mathbb{E} \left(\mathbb{E}_{x,T} (B_0^{(\epsilon)} - B_0) \right),
\end{aligned}$$

where the first equality uses (27) of Lemma 5 and the second equality follows from $h'(x) = 1$.

Recall that $R_a(B_0)$ is the residual interarrival time at the end of the initial busy period ($R_a(B_0)$ also equals the length of the first idle period I_0). If $R_a(B_0) \geq \epsilon/\delta$, then there is no arrival during the interval $[B_0, B_0^{(\epsilon)})$ and thus $B_0^{(\epsilon)} = B_0 + \epsilon/\delta$, so that

$$\begin{aligned}
\frac{1}{\epsilon} \mathbb{E} \left(\mathbb{E}_{x,T} (1(R_a(B_0) \geq \epsilon/\delta) (B_0^{(\epsilon)} - B_0)) \right) &= \frac{1}{\delta} \mathbb{E} \left(\mathbb{P}_{x,T}(R_a(B_0) \geq \epsilon/\delta) \right) \\
&\xrightarrow{\epsilon \rightarrow 0} \frac{1}{\delta} \mathbb{E} \left(\mathbb{P}_{x,T}(R_a(B_0) > 0) \right) = \frac{1}{\delta}.
\end{aligned}$$

To justify the last equality, we observe that $R_a(B_0) = 0$ would imply that an arrival occurs precisely at the instant that the workload hits zero. Since the workload process is right-continuous, this would imply that $X(B_0) > 0$, which contradicts the definition of B_0 .

It remains to show that

$$\begin{aligned} & \frac{1}{\epsilon} \mathbb{E} \left(\mathbb{E}_{x,T} \left(1(R_a(B_0) < \epsilon/\delta) (B_0^{(\epsilon)} - B_0) \right) \right) \\ & \xrightarrow{\epsilon \rightarrow 0} \frac{1}{\delta} \left(\theta(x/\delta) + \mathbb{E} (1(T < x/\delta) \mathbb{E}_{x,T} \eta(\alpha(B_0))) \right) \mathbb{E} \bar{B}. \end{aligned}$$

Since $R_a(B_0) < \epsilon/\delta$ implies that an arrival occurs in $[B_0, B_0^{(\epsilon)})$, then

$$\begin{aligned} & \frac{1}{\epsilon} \mathbb{E} \left(\mathbb{E}_{x,T} \left(1(R_a(B_0) < \epsilon/\delta) (B_0^{(\epsilon)} - B_0) \right) \right) \\ &= \frac{1}{\epsilon} \mathbb{E} \left(\mathbb{E}_{x,T} \left(1(R_a(B_0) < \epsilon/\delta) R_a(B_0) \right) \right) \\ & \quad + \frac{1}{\epsilon} \mathbb{E} \left(\mathbb{E}_{x,T} \left(1(R_a(B_0) < \epsilon/\delta) \mathbb{E}^{S,U} (\partial_x F_h(\epsilon - \delta R_a(B_0) + \delta S, U)) \right) \right), \end{aligned}$$

where we remind the reader that $\mathbb{E}^{S,U}(\cdot)$ is an expectation over S and U but not $R_a(B_0)$. The first term on the right-hand side converges to zero as $\epsilon \rightarrow 0$. To analyze the second term, note that $R_a(B_0) < \epsilon/\delta$ implies that $T < x/\delta + \epsilon/\delta$. Furthermore, $x/\delta \leq T < x/\delta + \epsilon/\delta$ implies that $R_a(B_0) = T - x/\delta$ and therefore, the second term equals

$$\begin{aligned} & \frac{1}{\epsilon} \mathbb{E} \left(1(T < x/\delta) \mathbb{E}_{x,T} \left(1(R_a(B_0) < \epsilon/\delta) \mathbb{E}^{S,U} (\partial_x F_h(\epsilon - \delta R_a(B_0) + \delta S, U)) \right) \right) \\ & \quad + \frac{1}{\epsilon} \mathbb{E} \left(1(x/\delta \leq T < x/\delta + \epsilon/\delta) \mathbb{E}_{x,T} \left(\mathbb{E}^{S,U} (\partial_x F_h(\epsilon - (\delta T - x) + \delta S, U)) \right) \right). \end{aligned}$$

To characterize the limit of the first term as $\epsilon \rightarrow 0$, one can check that

$$\sup_{0 \leq x' \leq \epsilon} \left| \mathbb{E} (\partial_x F_h(x' + \delta S, U)) - \mathbb{E} (\partial_x F_h(\delta S, U)) \right| \leq \mathbb{E} (\mathbb{E}_{\delta S, U} (B_0^{(\epsilon)} - B_0)) \xrightarrow{\epsilon \rightarrow 0} 0,$$

which, together with $\mathbb{E} (\partial_x F_h(\delta S, U)) = \mathbb{E} \bar{B}$, Lemma 26, and the DCT yields

$$\begin{aligned} & \frac{1}{\epsilon} \mathbb{E} \left(1(T < x/\delta) \mathbb{E}_{x,T} \left(1(R_a(B_0) < \epsilon/\delta) \mathbb{E} (\partial_x F_h(\epsilon - \delta R_a(B_0) + \delta S, U)) \right) \right) \\ & \xrightarrow{\epsilon \rightarrow 0} \frac{1}{\delta} \mathbb{E} (1(T < x/\delta) \mathbb{E}_{x,T} \eta(\alpha(B_0))) \mathbb{E} \bar{B}. \end{aligned}$$

Similarly, using the fact that $\theta(x)$ is bounded,

$$\frac{1}{\epsilon} \mathbb{E} \left(1(x/\delta \leq T < x/\delta + \epsilon/\delta) \mathbb{E}_{x,T} \left(\mathbb{E} (\partial_x F_h(\epsilon - (\delta T - x) + \delta S, U)) \right) \right) \xrightarrow{\epsilon \rightarrow 0} \frac{1}{\delta} \theta(x/\delta) \mathbb{E} \bar{B}.$$

□

D.2. Third-order bounds

D.2.1. Proof of Lemma 14

LEMMA 27. Suppose that U has a bounded density. Then for any $x \geq 0$,

$$\partial_x^2 \mathbb{E}(F_h(x + \delta S, U) - F_h(x, U)) = \mathbb{E}^S(\partial_x^2 \mathbb{E}^U F_h(x + \delta S, U) - \partial_x^2 \mathbb{E}^U F_h(x, U)).$$

Proof of Lemma 14 Observe that

$$\begin{aligned} \delta \bar{F}_h'''(x) &= \lambda \partial_x^2 \mathbb{E}(F_h(x + \delta S, U) - F_h(x, U)) \\ &= \lambda \mathbb{E}^S(\partial_x^2 \mathbb{E}^U F_h(x + \delta S, U) - \partial_x^2 \mathbb{E}^U F_h(x, U)), \end{aligned}$$

where the first equality is due to (45) and the second is due to Lemma 27. Applying the expression for $\partial_x^2 \mathbb{E}^U F_h(\cdot, U)$ from Lemma 13 to the right-hand side yields the result. $\square$

Proof of Lemma 27 Though we do not assume $F_h(z)$ to be well defined, note that

$$\begin{aligned} &\partial_x \mathbb{E}(F_h(x + \delta S, U) - F_h(x, U)) \\ &= \lim_{\epsilon \rightarrow 0} \left(\frac{1}{\epsilon} \mathbb{E}(F_h(x + \epsilon + \delta S, U) - F_h(x + \delta S, U)) - \frac{1}{\epsilon} \mathbb{E}(F_h(x + \epsilon, U) - F_h(x, U)) \right) \\ &= \partial_x \mathbb{E} F_h(x + \delta S, U) - \partial_x \mathbb{E} F_h(x, U), \end{aligned}$$

so that by differentiating both sides with respect to x , we arrive at

$$\partial_x^2 \mathbb{E}(F_h(x + \delta S, U) - F_h(x, U)) = \partial_x^2 \mathbb{E} F_h(x + \delta S, U) - \partial_x^2 \mathbb{E} F_h(x, U).$$

By repeating the arguments used to prove Lemma 5, one can check that

$$\partial_x^2 \mathbb{E} F_h(x + \delta S, U) = \partial_x \mathbb{E}^S \partial_x \mathbb{E}^U F_h(x + \delta S, U).$$

Similarly,

$$\partial_x \mathbb{E}^S \partial_x \mathbb{E}^U F_h(x + \delta S, U) = \mathbb{E}^S \partial_x^2 \mathbb{E}^U F_h(x + \delta S, U)$$

follows from repeating the proof of Lemma 13. $\square$

D.3. Proof of Lemma 12

For convenience, let $\nu = 1/\mathbb{E}Y_2$. We use the following inequalities throughout the proof, which follow from the fact that Y_2 has density $\nu e^{-\nu y}$ and is independent of R_a .

$$\mathbb{P}(\delta R_a \geq Y_2) = \mathbb{E}(1 - e^{-\nu \delta R_a}) \leq \nu \delta \mathbb{E} R_a, \quad (81)$$

$$\mathbb{E} Y_2 1(\delta R_a \geq Y_2) = \frac{1}{\nu} \mathbb{E}((1 - e^{-\nu \delta R_a}) - \nu \delta R_a e^{-\nu \delta R_a}) \leq 2\nu \delta^2 \mathbb{E} R_a^2, \quad (82)$$

$$\mathbb{E} Y_2^2 1(\delta R_a \geq Y_2) = \frac{1}{\nu^2} \mathbb{E}(2(1 - e^{-\nu \delta R_a}) - 2\nu \delta R_a e^{-\nu \delta R_a} - (\nu \delta R_a)^2 e^{-\nu \delta R_a}) \leq 3\nu \delta^3 \mathbb{E} R_a^3. \quad (83)$$

We also use the facts that $U, S, S', R_a, \bar{I}$, and Y_2 are independent, and that S and S' have the same distribution. We first prove (40)–(42) and then (43)–(44) when $\eta(x)$ is nonincreasing, following by the case when $\underline{\eta} > 0$.

Proof of (40)–(42). Inequality (40) follows $J(x, r) = -(x \wedge \delta r_a) + \delta S'$. Next, we prove (41), we first note that the definition of $\epsilon(x, r_a, s)$ in Lemma 7 and the fact that $h(x) = x$ both imply that

$$\int_0^{r_a} |h((x + \delta s - \delta t)^+) - h((x - \delta t)^+)| dt \leq r_a \delta s.$$

Furthermore,

$$\begin{aligned} & 1(\delta r_a > x) \mathbb{E}^{S'} \left(\int_{-x}^{-(x+\delta s)} \partial_x \mathbb{E}^U F_h(x + \delta s + v + \delta S', U) dv \right) \\ & \leq 1(\delta r_a > x) \delta s \mathbb{E}^{S'} \left(\sup_{0 \leq w \leq \delta s} \partial_x \mathbb{E}^U F_h(w + \delta S', U) \right) \\ & \leq 1(\delta r_a > x) s(\delta s + \delta \mathbb{E} S') (1 + 2\bar{\eta} \mathbb{E} \bar{B}), \end{aligned}$$

where in the second inequality we used (35) of Lemma 9. Combining the bounds and using (81) yields

$$\begin{aligned} \mathbb{E} |\epsilon(Y_2, R_a, S)| & \leq \delta \mathbb{E} R_a \mathbb{E} S + \mathbb{P}(\delta R_a \geq Y_2) \delta (\mathbb{E} S^2 + (\mathbb{E} S)^2) (1 + 2\bar{\eta} \mathbb{E} \bar{B}) \\ & \leq \delta \mathbb{E} R_a \mathbb{E} S + \delta^2 \nu \mathbb{E} R_a (\mathbb{E} S^2 + (\mathbb{E} S)^2) (1 + 2\bar{\eta} \mathbb{E} \bar{B}). \end{aligned}$$

Next we prove (42). Recall from Lemma 9 that

$$\begin{aligned} |\delta \bar{F}'_h(x)| & \leq x(1 + (\lambda + \bar{\eta}) \mathbb{E} \bar{B}), \\ |\delta \bar{F}''_h(x)| & \leq (1 + x)(1 + (\lambda + \bar{\eta}) \mathbb{E} \bar{B}), \quad x \geq 0. \end{aligned}$$

Therefore,

$$\begin{aligned} & \mathbb{E} |1(\delta R_a \geq Y_2) (\delta \bar{F}'_h(\delta S) - \delta \bar{F}'_h(Y_2) - \delta(S - R_a) \delta \bar{F}''_h(Y_2))| \\ & \leq \left(\mathbb{E} (1(\delta R_a \geq Y_2) (\delta S + Y_2)) + \mathbb{E} (1(\delta R_a \geq Y_2) \delta(S - R_a) (1 + Y_2)) \right) (1 + (\lambda + \bar{\eta}) \mathbb{E} \bar{B}) \\ & \leq \delta^2 (\nu (2\mathbb{E} S \mathbb{E} R_a + 5\mathbb{E} R_a^2) + \mathbb{E} R_a^2) (1 + (\lambda + \bar{\eta}) \mathbb{E} \bar{B}). \end{aligned}$$

The last inequality follows from using (81) and (82) to show that

$$\begin{aligned} \mathbb{E} (1(\delta R_a \geq Y_2) (\delta S + Y_2)) & \leq \delta^2 \nu \mathbb{E} S \mathbb{E} R_a + 2\nu \delta^2 \mathbb{E} R_a^2, \\ \delta \mathbb{E} S \mathbb{E} (1(\delta R_a \geq Y_2) (1 + Y_2)) & \leq \delta \mathbb{E} S (\nu \delta \mathbb{E} R_a + 2\nu \delta^2 \mathbb{E} R_a^2), \\ \delta \mathbb{E} (1(\delta R_a \geq Y_2) R_a (1 + Y_2)) & \leq \delta (\nu \delta \mathbb{E} R_a^2 + \delta \mathbb{E} R_a^2), \end{aligned}$$

where in the final inequality we used the fact that $\mathbb{E} (1(\delta R_a \geq Y_2) R_a Y_2) \leq \delta \mathbb{E} R_a^2$ instead of (82).

Using the latter would have resulted in a term involving $\mathbb{E} R_a^3$.

Proof of (43)–(44). Recall that

$$\bar{C} = (3 + (\mathbb{E}U + \mathbb{E}\bar{I})\lambda(1 + \rho + \lambda^2\mathbb{E}U^2))\nu\bar{\eta}\delta\mathbb{E}\bar{B}.$$

We claim that

$$\mathbb{E}|\delta^2\bar{F}_h'''(Y_2 + u)| \leq \bar{C}, \quad u > 0, \quad (84)$$

$$\mathbb{E}^{Y_2}\left(1(\delta R_a \leq Y_2)|\delta^2\bar{F}_h'''(Y_2 + u)|\right) \leq \bar{C}, \quad u \in [-\delta R_a, 0]. \quad (85)$$

Then (43) follows by applying the bound in (84) to

$$\lambda\mathbb{E}\left(\int_0^{\delta S}(\delta S - v)\int_0^v|\bar{F}_h'''(Y_2 + u)|dudv\right) = \lambda\mathbb{E}^S\left(\int_0^{\delta S}(\delta S - v)\int_0^v\mathbb{E}^{Y_2}|\bar{F}_h'''(Y_2 + u)|dudv\right).$$

To prove (44), observe that both (84) and (85) imply that

$$\begin{aligned} & \delta\frac{1}{\delta^2}\mathbb{E}\left(\int_0^{\delta(S-R_a)}\int_0^v|1(\delta R_a < Y_2)\delta^2\bar{F}_h'''(Y_2 + u)|dudv\right) \\ &= \delta\frac{1}{\delta^2}\mathbb{E}^{S,R_a}\left(\int_0^{\delta(S-R_a)}\int_0^v\mathbb{E}^{Y_2}|1(\delta R_a < Y_2)\delta^2\bar{F}_h'''(Y_2 + u)|dudv\right) \\ &\leq \delta\mathbb{E}(S - R_a)^2\bar{C}. \end{aligned}$$

It remains to prove (84) and (85). Recall from Lemma 10 that for any x, u such that $x + u \geq 0$,

$$|\delta^2\bar{F}_h'''(x + u)| \leq (\mathbb{P}(\delta U > x + u)3\lambda + \mathbb{P}(\delta U < x + u < \delta\bar{I} + \delta U)\lambda(1 + \rho + \lambda^2\mathbb{E}U^2))\bar{\eta}\mathbb{E}\bar{B}. \quad (86)$$

Then (84) follows once we observe that for any $u > 0$,

$$\begin{aligned} \mathbb{P}(\delta U > Y_2 + u) &\leq \mathbb{P}(\delta U > Y_2) = \mathbb{E}(1 - e^{-\nu\delta U}) \leq \nu\delta\mathbb{E}U, \\ \mathbb{P}(\delta U < Y_2 + u < \delta\bar{I} + \delta U) &\leq \nu\delta(\mathbb{E}U + \mathbb{E}\bar{I}), \end{aligned} \quad (87)$$

where the last inequality is true because

$$\begin{aligned} & \mathbb{P}(\delta U < Y_2 + u < \delta\bar{I} + \delta U) \\ &= \mathbb{P}(Y_2 \leq \delta U < Y_2 + u < \delta\bar{I} + \delta U) + \mathbb{P}(\delta U < Y_2, Y_2 + u < \delta\bar{I} + \delta U) \\ &\leq \mathbb{P}(\delta U \geq Y_2) + \mathbb{P}(\delta U < Y_2 < \delta U + \delta\bar{I}) \\ &= \mathbb{E}(1 - e^{-\nu\delta U}) + \mathbb{E}e^{-\nu\delta U}(1 - e^{-\nu\delta\bar{I}}) \leq \nu\delta(\mathbb{E}U + \mathbb{E}\bar{I}). \end{aligned}$$

Similarly, (85) follows from the fact that for any $u \in [-\delta R_a, 0]$,

$$\begin{aligned} \mathbb{E}^{Y_2,U}(1(\delta R_a < Y_2)1(\delta U > Y_2 + u)) &\leq \nu\delta\mathbb{E}U, \\ \mathbb{E}^{Y_2,U,\bar{I}}(1(\delta U < Y_2 + u < \delta\bar{I} + \delta U)) &\leq \nu\delta(\mathbb{E}U + \mathbb{E}\bar{I})\mathbb{E}\bar{B}. \end{aligned} \quad (88)$$

The first inequality in (88) follows from

$$\begin{aligned} & \mathbb{E}^{Y_2, U} (1(\delta R_a < Y_2) 1(\delta U > Y_2 + u)) \\ & \leq \mathbb{E}^{Y_2, U} (1(\delta R_a < Y_2 < \delta U + \delta R_a)) = e^{-\nu \delta R_a} \mathbb{E}(1 - e^{-\nu \delta U}) \leq \nu \delta \mathbb{E} U \end{aligned}$$

and the second from

$$\begin{aligned} E^{Y_2, U, \bar{I}} (1(\delta U < Y_2 + u < \delta \bar{I} + \delta U)) & \leq E^{Y_2, U, \bar{I}} (1(\delta U - u < Y_2 < \delta \bar{I} + \delta U - u)) \\ & = \mathbb{E}(e^{-\nu(\delta U - u)} (1 - e^{-\nu \delta \bar{I}})) \leq \nu \delta \mathbb{E} \bar{I}. \end{aligned}$$

□

Appendix E: Proofs for the $G/M/\infty$ system

To prove Proposition 3 we require the following inversion formula, which is proved exactly like Lemma 3.

LEMMA 28. Initialize $Z(0) \sim Z$ and fix $f: \mathbb{R} \rightarrow \mathbb{R}$ with

$$\mathbb{E}|f(X)| < \infty \quad \text{and} \quad \mathbb{E} \left| \int_0^{R(0)} f(X(t)) dt \right| < \infty,$$

which holds, in particular, when $f(x)$ is bounded. Then

$$\mathbb{E} f(X) = \mathbb{E} \int_0^1 \int_0^{U(t)} f(X(t+u)) du dA(t) \quad (89)$$

We now have what we need to prove Proposition 3.

Proof of Proposition 3 Repeating the arguments used in the proof of Proposition 1 yields

$$\mathbb{E} \int_0^1 \Delta f(\tilde{X}(t-)) dA(t) = \frac{1}{2} \delta^2 \lambda c_U^2 \mathbb{E} f''(X) + \epsilon_A(f) = \frac{1}{2} \mu c_U^2 \mathbb{E} f''(X) + \epsilon_A(f),$$

where

$$\begin{aligned} \epsilon_A(f) &= \frac{1}{6} \delta^3 \mathbb{E} \int_0^1 (1 - \lambda U(t))^3 f'''(\xi(t)) dA(t) \\ &\quad - \frac{1}{2} \delta^2 \lambda c_U^2 \mathbb{E} \int_0^1 \int_0^{U(t)} (X(t+u) - X(t-)) f'''(\xi(t+u)) du dA(t). \end{aligned} \quad (90)$$

To bound the first term on the right-hand side, we use the fact that $\mathbb{E} A(1) = \lambda$ and $\delta^2 \lambda = \mu$ to get

$$\frac{1}{6} \delta^3 \mathbb{E} \int_0^1 |(1 - \lambda U(t))^3 f'''(\xi(t))| dA(t) \leq \frac{1}{6} \delta \mu \|f'''\| \mathbb{E} |1 - \lambda U|^3$$

To bound the second term in (90), let $D[t, t+u]$ be the number of departures on $[t, t+u]$ and recall that $X(t) = \delta(Q(t) - R)$. It follows that

$$\begin{aligned}
& \mathbb{E} \int_0^1 \int_0^{U(t)} |(X(t+u) - X(t-)) f'''(\xi(t+u))| dudA(t) \\
& \leq \|f'''\| \delta \mathbb{E} \int_0^1 \int_0^{U(t)} |Q(t+u) - Q(t-)| dudA(t) \\
& \leq \|f'''\| \delta \mathbb{E} \int_0^1 \int_0^{U(t)} (1 + D[t, t+u]) dudA(t) \\
& \leq \|f'''\| \delta \mathbb{E} \int_0^1 \int_0^{U(t)} (1 + \mu Q(t)u) dudA(t) \\
& = \|f'''\| \delta \mathbb{E} \int_0^1 (U(t) + \mu Q(t)U^2(t)/2) dA(t) \\
& = \|f'''\| \delta \left(1 + \frac{1}{2} \mathbb{E} U^2 \mu \mathbb{E} \int_0^1 Q(t) dA(t)\right) \\
& = \|f'''\| \delta \left(1 + \frac{1}{2} \mathbb{E} U^2 \mu \mathbb{E} \int_0^1 (1 + Q(t-)) dA(t)\right).
\end{aligned}$$

The third inequality is true because the number of departures on $[t, t+u]$ is bounded by a Poisson process with rate $\mu Q(t)$. Using $\mathbb{E}A(1) = \lambda$ and (62) of Lemma 18, the right-hand side equals

$$\begin{aligned}
& \|f'''\| \delta \left(1 + \frac{1}{2} \mathbb{E} U^2 \mu (\lambda + \mu \mathbb{E}(Q - R)^2 + \lambda R - \lambda)\right) \\
& = \|f'''\| \delta \left(1 + \frac{1}{2} \lambda^2 \mathbb{E} U^2 (\mathbb{E}(Q - R)^2 / R^2 + 1)\right).
\end{aligned}$$

We arrive at the stated bound on $|\epsilon_A(f)|$.

We now prove (58). Taylor expansion yields

$$\mu Q(f(\tilde{X} - \delta) - f(\tilde{X})) = -\mu \delta Q f'(\tilde{X}) + \frac{1}{2} \mu \delta^2 Q f''(\tilde{X}) + \frac{1}{6} \mu \delta^3 Q f'''(\xi).$$

Since $\mu R = \lambda$, $X = \delta(Q - R)$, and $\tilde{X} - X = -\delta \lambda R_a$, the first two terms on the right-hand side satisfy

$$\begin{aligned}
-\mu \delta Q f'(\tilde{X}) &= -\mu \delta R f'(\tilde{X}) - \mu \delta (Q - R) f'(\tilde{X}) \\
&= -\mu \delta R f'(\tilde{X}) - \mu \delta (Q - R) f'(X) - \mu \delta (Q - R) (f'(\tilde{X}) - f'(X)) \\
&= -\delta \lambda f'(\tilde{X}) - \mu X f'(X) + \mu X \delta \lambda R_a f''(\xi),
\end{aligned}$$

and

$$\begin{aligned}
\frac{1}{2} \mu \delta^2 Q f''(\tilde{X}) &= \frac{1}{2} \mu \delta^2 R f''(X) + \frac{1}{2} \mu \delta^2 (Q - R) f''(X) + \frac{1}{2} \mu \delta^2 Q (f''(\tilde{X}) - f''(X)) \\
&= \frac{1}{2} \mu f''(\tilde{X}) + \frac{1}{2} \mu \delta X f''(X) - \frac{1}{2} \mu \delta^3 Q \lambda R_a f'''(\xi),
\end{aligned}$$

implying that (91) holds with

$$\epsilon_D(f) = \mu\delta\mathbb{E}(X\lambda R_a f''(\xi)) - \frac{1}{2}\mu\delta^3\mathbb{E}(Q\lambda R_a f'''(\xi)) + \frac{1}{6}\mu\delta^3\mathbb{E}(Qf'''(\xi)). \quad (91)$$

To conclude, we recall that $\mathbb{E}Q = R$ and bound each of the three terms on the right-hand side as follows:

$$\begin{aligned} \mu\delta\mathbb{E}|X\lambda R_a f''(\xi)| &\leq \mu\delta\|f''\|\sqrt{\mathbb{E}X^2}\sqrt{\lambda^2\mathbb{E}R_a^2} = \mu\delta\|f''\|\sqrt{\mathbb{E}X^2}\sqrt{\lambda^3\mathbb{E}U^3/3}, \\ \frac{1}{2}\mu\delta^3\mathbb{E}(Q\lambda R_a f'''(\xi)) &\leq \frac{1}{2}\mu\|f'''\|\delta^3\mathbb{E}(Q\lambda R_a) \\ \frac{1}{6}\mu\delta^3\mathbb{E}(Qf'''(\xi)) &\leq \frac{1}{6}\mu\|f'''\|\delta^3\mathbb{E}Q = \frac{1}{6}\mu\|f'''\|\delta. \end{aligned}$$

□

E.1. Proof of Lemma 18

The first two equalities in (60) are argued as in Lemma 2 and the third is obtained from the BAR (55) with $f(z) = q \wedge M$. For the relationship in (61), we invoke the BAR with $f(z) = ((q \wedge M) - \lambda r_a)^2$ to get

$$\begin{aligned} 0 &= 2\lambda\mathbb{E}(Q \wedge M - \lambda R_a) + \mu\mathbb{E}(Q1(Q \leq M)((Q - 1 - \lambda R_a)^2 - (Q - \lambda R_a)^2)) \\ &\quad + \mathbb{E}\int_0^1 1(Q(t-) \leq M - 1)((Q(t-) + 1 - \lambda U(t))^2 - (Q(t-))^2) dA(t) \\ &= 2\lambda\mathbb{E}(Q \wedge M - \lambda R_a) + \mu\mathbb{E}(Q1(Q \leq M)(-2(Q - \lambda R_a) + 1)) \\ &\quad + \mathbb{E}\int_0^1 1(Q(t-) \leq M - 1)(2Q(t-)(1 - \lambda U(t)) + (1 - \lambda U(t))^2) dA(t) \end{aligned}$$

Taking $M \rightarrow \infty$ yields

$$0 = 2\lambda\mathbb{E}Q - 2\lambda^2\mathbb{E}R_a - 2\mu\mathbb{E}Q^2 + 2\mu\lambda\mathbb{E}QR_a + \mu\mathbb{E}Q + \mathbb{E}(1 - \lambda U)^2\mathbb{E}A(1).$$

Divide by 2μ , move $\mathbb{E}Q^2$ to the left-hand side, and use $\mathbb{E}Q = R$ to get

$$\begin{aligned} \mathbb{E}Q^2 &= \mathbb{E}(Q - R)^2 + R^2 \\ &= R^2 - R\lambda^2\mathbb{E}U^2/2 + \lambda\mathbb{E}QR_a + R/2 + R\mathbb{E}(1 - \lambda U)^2/2, \end{aligned}$$

implying that

$$\begin{aligned} \mathbb{E}(Q - R)^2 &= -R\lambda^2\mathbb{E}U^2/2 + \lambda\mathbb{E}QR_a + R/2 + R(1 - 2 + \lambda^2\mathbb{E}U^2)/2 \\ &= \lambda\mathbb{E}QR_a \\ &= \lambda\mathbb{E}(Q - R)R_a + R\lambda^2\mathbb{E}U^2/2 \\ &\leq \lambda\sqrt{\mathbb{E}(Q - R)^2}\sqrt{\mathbb{E}R_a^2} + R\lambda^2\mathbb{E}U^2/2. \end{aligned}$$

Since $x^2 \leq bx + c$ implies that $x^2 \leq (b^2 + x^2)/2 + c$ and, therefore, $x^2 \leq b^2 + 2c$, and since $\mathbb{E}R_a^2 = \lambda \mathbb{E}U^3/3$, we obtain (61).

To prove (62) we use $f(z) = (q \wedge m)^2$ in the BAR to get

$$\begin{aligned} 0 &= \mu \mathbb{E}(Q1(Q \leq M)(-2Q + 1)) + \mathbb{E} \int_0^1 1(Q(t-) \leq M - 1)(2Q(t-) + 1)dA(t) \\ &= -2\mu \mathbb{E}(Q^2 1(Q \leq M)) + \mu \mathbb{E}(Q1(Q \leq M)) \\ &\quad + 2\mathbb{E} \int_0^1 1(Q(t-) \leq M - 1)Q(t-)dA(t) + \mathbb{E} \int_0^1 1(Q(t-) \leq M - 1)dA(t). \end{aligned}$$

Taking $M \rightarrow \infty$ and using $\mathbb{E}Q = R$ we arrive at

$$\begin{aligned} 2\mathbb{E}Q^2 &= \mathbb{E}Q + \frac{2}{\mu} \mathbb{E} \int_0^1 Q(t-)dA(t) + \frac{1}{\mu} \mathbb{E}A(1) \\ &= R + \frac{2}{\mu} \mathbb{E} \int_0^1 Q(t-)dA(t) + R. \end{aligned}$$

Using $\mathbb{E}A(1) = \lambda$ and rearranging terms, we get

$$\frac{1}{\mu} \mathbb{E} \int_0^1 Q(t-)dA(t) = \mathbb{E}Q^2 - R = \mathbb{E}(Q - R)^2 + R^2 - R.$$

□

Appendix F: Proofs for the JSQ system

To prove Proposition 4 we require two auxiliary lemmas. The first lemma contains some useful relationships derived using the BAR.

LEMMA 29. For any $1 \leq i \neq j \leq n$ and $m > 1$,

$$\mathbb{E}A(1) = n\mathbb{E}D_i(1) = \lambda, \quad \mathbb{P}(Q_i > 0) = \rho, \quad (92)$$

$$\mathbb{E}R_a^{m-1} = \lambda \mathbb{E}U^m/m, \quad \mathbb{E}(R_{s,i}^{m-1} 1(Q_i > 0)) = (\lambda/n) \mathbb{E}S^m/m, \quad \mathbb{E}(R_{s,i}^m | Q_i = 0) = \mathbb{E}S^m, \quad (93)$$

$$\mathbb{E} \int_0^1 R_a(t) dD_i(t) \leq \mu(\mathbb{E}R_{s,i} + \mathbb{E}R_a), \quad \mathbb{E} \int_0^1 R_{s,i}(t) dA(t) \leq \lambda(\mathbb{E}R_{s,i} + \mathbb{E}R_a), \quad (94)$$

$$\mathbb{E} \int_0^1 R_{s,i}(t) dD_j(t) \leq \mu(\mathbb{E}R_{s,i} + \mathbb{E}R_{s,j}). \quad (95)$$

Proof of Lemma 29 As in the proof of Lemma 2, we prove each of (92)–(95) by plugging in truncated test functions into the BAR (65). Using $f(z) = r_a \wedge M$, $f(z) = (q_1 + \dots + q_n) \wedge M$ (together with the symmetry of the servers), and $f(z) = r_{s,i} \wedge M$ yields $\mathbb{E}A(1) = \lambda$, $n\mathbb{E}D_i(1) = \mathbb{E}A(1)$, and $\mathbb{E}S\mathbb{E}D_i(1) = \mathbb{P}(Q_i > 0)$, respectively, which proves (92). Using $f(z) = r_a^m \wedge M$ yields the first equality in (93) and using $f(z) = r_{s,i}^m \wedge M$ yields $m\mathbb{E}(R_{s,i}^{m-1} 1(Q_i > 0)) = \mathbb{E}S^m \mathbb{E}D(1) = (\lambda/n) \mathbb{E}S^m$. The third equality in (93) follows once we note that $\mathbb{E}(R_{s,i}^m 1(Q_i = 0)) = \mathbb{E}S^m \mathbb{P}(Q_i = 0)$ because when $Q_i(t) = 0$,

then $R_{s,i}(t)$ equals the service time of the next customer to arrive. Using $f(z) = (r_a \wedge M)(r_{s,i} \wedge M)$ yields

$$\begin{aligned} & \mathbb{E} \int_0^1 R_a(t) (1(Q_i(t)=0)\Lambda_i(t) + S_i(t)) dD_i(t) + \mathbb{E} \int_0^1 R_{s,i}(t) U(t) dA(t) \\ &= \mathbb{E} R_{s,i} + \mathbb{E} 1(Q_i > 0) R_a, \end{aligned}$$

from which (94) follows. The inequality (95) is proved similarly using $f(z) = (r_{s,i} \wedge M)(r_{s,j} \wedge M)$. $\square$

The second lemma is the Palm inversion formula, analogous to Lemma 3. For simplicity, we only prove it for bounded functions $f(x)$, though this assumption can be relaxed as in Lemma 3.

LEMMA 30. *For any bounded $f: \mathbb{R} \rightarrow \mathbb{R}$,*

$$\mathbb{E} f(X) = \mathbb{E} \left(\int_0^1 \int_0^{U(t)} f(X(t+u)) du dA(t) \right) \quad \text{and} \quad (96)$$

$$\mathbb{E} f(X) = \mathbb{E} \left(\int_0^1 \int_0^{1(Q_i(t)=0)\Lambda_i(t) + S_i(t)} f(X(t+u)) du dD_i(t) \right). \quad (97)$$

Proof of Lemma 30 Since $\mathbb{E} R_a < \infty$ (Lemma 29), the proof of (96) is identical to that of (10) in Lemma 3. Similarly, to prove (97), let $R_{d,i}(t) = 1(Q_i(t)=0)\Lambda_i(t) + R_{s,i}(t)$ be the remaining time until the next departure from server i and let $R_{d,i}$ denote the steady-state version. We only need to argue that $\mathbb{E} R_{d,i} < \infty$. This is implied by the facts that 1) the expected time until the next customer arrives to server i is finite because $\mathbb{E} R_a < \infty$ and because of our uniform tie-breaking rule (meaning that an arrival will be allocated an idling server with probability at least $1/n$) and 2) $\mathbb{E} R_{s,i} < \infty$ (Lemma 29). $\square$

Proof of Proposition 4 The proof has two parts. In part one we assume that (66)–(68) hold with

$$\begin{aligned} \epsilon_0(f) = & -n\mu\delta^3 \mathbb{E} \left(\left(-\lambda R_a + \sum_{i=1}^n \mu R_{s,i} \right) f''(\xi) \right) \\ & + \delta^2 \mu \sum_{i=1}^n \mathbb{E} \left(1(Q_i=0) \left(\sum_{j=1}^n Q_j - \lambda R_a + \sum_{j=1}^n \mu R_{s,j} \right) f''(\xi) \right), \end{aligned} \quad (98)$$

$$\begin{aligned} \epsilon_A(f) = & \frac{1}{6} \delta^3 \mathbb{E} \int_0^1 (1 - \lambda U(t))^3 f'''(\xi(t)) dA(t) \\ & + \frac{1}{2} \delta^3 c_U^2 \mathbb{E} \int_0^1 \left(\sum_{i=1}^n \mu R_{s,i}(t) \right) f'''(\xi(t-)) dA(t) \\ & - \frac{1}{2} \delta^2 \lambda c_U^2 \mathbb{E} \int_0^1 \int_0^{U(t)} (X(t+u) - X(t-)) f'''(\xi(t+u)) du dA(t), \end{aligned} \quad (99)$$

and

$$\begin{aligned}
\epsilon_{D,i}(f) = & \frac{1}{6}\delta^3\mathbb{E}\int_0^1(1-\mu S_i(t))^3f'''(\xi(t))dD_i(t) \\
& + \frac{1}{2}\delta^3c_S^2\mathbb{E}\int_0^1\left(-\lambda R_a(t) + \sum_{j \neq i}\mu R_{s,j}(t)\right)f'''(\xi(t-))dD_i(t) \\
& - \frac{1}{2}\delta^2\mu c_S^2\mathbb{E}\int_0^1\int_0^{1(Q_i(t)=0)\Lambda_i(t)}f''(X(t+u))dudD_i(t) \\
& - \frac{1}{2}\delta^2\mu c_S^2\mathbb{E}\int_0^1\int_{1(Q_i(t)=0)\Lambda_i(t)}^{1(Q_i(t)=0)\Lambda_i(t)+S_i(t)}(X(t+u)-X(t-))f'''(\xi(t+u))dudD_i(t), \quad (100)
\end{aligned}$$

and we bound the terms on the right-hand side, with the exception of the boundary term in $\epsilon_0(f)$. In part two, we expand all terms in the BAR (65) to verify that (66)–(68) does indeed hold with $\epsilon_0(f)$, $\epsilon_A(f)$, and $\epsilon_{D,i}(f)$ as in (98)–(100).

Part one. To prove the bound on $\epsilon_0(f)$, we bound the non-boundary term by noting that

$$n\mu\delta^3\mathbb{E}\left(\left(\lambda R_a + \sum_{i=1}^n\mu R_{s,i}\right)\|f''\|\right) \leq n\mu\delta^3\|f''\|(\lambda^2\mathbb{E}U^2/2 + (\lambda\mu\mathbb{E}S^2/2 + \mu\delta\mathbb{E}S)),$$

where the inequality follows from Lemma 29. Next we bound the three terms composing $\epsilon_A(f)$. For the first term, noting that $\mathbb{E}A(1) = \lambda$, by Lemma 29, yields

$$\frac{1}{6}\delta^3\mathbb{E}\int_0^1(1-\lambda U(t))^3f'''(\xi(t))dA(t) \leq \frac{1}{6}\delta^3\|f'''\|\lambda\mathbb{E}|1-\lambda U|^3.$$

For the second term, we use Lemma 29 to get

$$\begin{aligned}
& \frac{1}{2}\delta^3c_U^2\mathbb{E}\int_0^1\left(\sum_{i=1}^n\mu R_{s,i}(t)\right)f'''(\xi(t-))dA(t) \\
& \leq \frac{1}{2}\delta^3c_U^2\|f'''\|n\mu\lambda(\mathbb{E}R_{s,i} + \mathbb{E}R_a) \\
& = \frac{1}{2}\delta^3c_U^2\|f'''\|n\mu\lambda(\rho(\lambda/n)\mathbb{E}S^2/2 + \delta\mathbb{E}S + \lambda\mathbb{E}U^2/2).
\end{aligned}$$

For the third term, note that

$$\begin{aligned}
& \frac{1}{2}\delta^2\lambda c_U^2\mathbb{E}\int_0^1\int_0^{U(t)}(X(t+u)-X(t-))f'''(\xi(t+u))dudA(t) \\
& \leq \frac{1}{2}\delta^2\lambda c_U^2\|f'''\|\mathbb{E}\int_0^1U(t)\sup_{0 \leq u \leq U(t)}|X(t+u)-X(t-)|dA(t) \\
& \leq \frac{1}{2}\delta^2\lambda c_U^2\|f'''\|(1+n+\lambda\mu\mathbb{E}U^2+\mu^2\mathbb{E}S^2).
\end{aligned}$$

To justify the second inequality, let $D_i[t, t+u]$ be the number of service completions by server i on $[t, t+u]$ and note that

$$\sup_{0 \leq u \leq U(t)}|X(t+u)-X(t-)| \leq \delta + \delta \sum_{i=1}^n D_i[t, t+U(t)].$$

By forcing server i to continue working even if it has no customers in its buffer, we can construct a zero-delayed counting process $\{\bar{D}_i(t), t \geq 0\}$ keeping track of *potential* service completions, which would be independent of the arrival process, and satisfy

$$D_i[t, t+u] \leq 1 + \bar{D}_i(u).$$

Since $\mathbb{E}\bar{D}_i(u) \leq \mu u + \mu^2 \mathbb{E}S^2$ by Lorden's inequality (Lorden (1970)), it follows that

$$\begin{aligned} \mathbb{E} \int_0^1 U(t) \sup_{0 \leq u \leq U(t)} |X(t+u) - X(t-)| dA(t) &\leq \delta \mathbb{E} \left(U(1+n + \sum_{i=1}^n \bar{D}_i(U)) \right) \mathbb{E}A(1) \\ &\leq \delta \mathbb{E} \left(U(1+n + \mu U + \mu^2 \mathbb{E}S^2) \right) \mathbb{E}A(1) \\ &= \delta(1+n + \lambda \mu \mathbb{E}U^2 + \mu^2 \mathbb{E}S^2) \end{aligned} \quad (101)$$

where in the last equality we used $\mathbb{E}U\mathbb{E}A(1) = 1$. Lastly, we bound the four terms composing $\epsilon_{D,i}(f)$.

First, since $\mathbb{E}D_i(1) = \lambda/n$ by Lemma 29, we have

$$\frac{1}{6} \delta^3 \mathbb{E} \int_0^1 (1 - \mu S_i(t))^3 f'''(\xi(t)) dD_i(t) \leq \frac{1}{6} \delta^3 \|f'''\| \frac{\lambda}{n} \mathbb{E}|1 - \mu S|^3.$$

Second, using Lemma 29, we have

$$\begin{aligned} &\frac{1}{2} \delta^3 c_S^2 \mathbb{E} \int_0^1 \left(\lambda R_a(t) + \sum_{j \neq i} \mu R_{s,j}(t) \right) f'''(\xi(t-)) dD_i(t) \\ &\leq \frac{1}{2} \delta^3 c_S^2 \|f'''\| \left(\lambda \mu (\mathbb{E}R_{s,i} + \mathbb{E}R_a) + \sum_{j \neq i} \mu^2 (\mathbb{E}R_{s,i} + \mathbb{E}R_{s,j}) \right) \\ &\leq \frac{1}{2} \delta^3 c_S^2 \|f'''\| \left(\lambda \mu (\rho(\lambda/n) \mathbb{E}S^2/2 + \delta \mathbb{E}S + \lambda \mathbb{E}U^2/2) + 2(n-1) \mu^2 (\rho(\lambda/n) \mathbb{E}S^2/2 + \delta \mathbb{E}S) \right) \\ &= \frac{1}{2} \delta^3 c_S^2 \|f'''\| \left(\frac{\lambda^2 \mu \rho + 2(n-1) \mu^2 \rho \lambda}{2n} \mathbb{E}S^2 + \delta(\lambda \mu + 2(n-1) \mu^2) \mathbb{E}S + \frac{1}{2} \lambda^2 \mu \mathbb{E}U^2 \right). \end{aligned}$$

Third, the Palm inversion formula (Lemma 30) applied to $f(z) = 1(q_i = 0)$, together with Lemma 29 yield

$$\mathbb{E} \int_0^1 1(Q_i(t) = 0) \Lambda_i(t) dD_i(t) = \mathbb{P}(Q_i = 0) = \delta,$$

so that

$$\frac{1}{2} \delta^2 \mu c_S^2 \mathbb{E} \int_0^1 \int_0^{1(Q_i(t)=0)\Lambda_i(t)} f''(X(t+u)) du dD_i(t) \leq \frac{1}{2} \delta^3 \|f''\| c_S^2 \mu \lambda / n.$$

Fourth,

$$\begin{aligned} &\frac{1}{2} \delta^2 \mu c_S^2 \|f'''\| \mathbb{E} \int_0^1 \int_{1(Q_i(t)=0)\Lambda_i(t)}^{1(Q_i(t)=0)\Lambda_i(t)+S_i(t)} |X(t+u) - X(t-)| du dD_i(t) \\ &\leq \frac{1}{2} \delta^2 \mu c_S^2 \|f'''\| \mathbb{E} \int_0^1 S_i(t) \sup_{0 \leq u \leq S_i(t)} |X(t+1(Q_i(t)=0)\Lambda_i(t)+u) - X(t-)| dD_i(t) \end{aligned}$$

Observe that

$$\begin{aligned} \sup_{0 \leq u \leq S_i(t)} |X(t+u) - X(t-)| &\leq \delta A[t, t+1(Q_i(t)=0)\Lambda_i(t)] \\ &\quad + \delta A(t+1(Q_i(t)=0)\Lambda_i(t), t+1(Q_i(t)=0)\Lambda_i(t) + S_i(t)) \\ &\quad + \delta \sum_{j \neq i} D_j[t, t+1(Q_i(t)=0)\Lambda_i(t) + S_i(t)], \end{aligned}$$

where $A[a, b]$ and $D_j[a, b]$ are the number of arrivals and server j departures, respectively, on $[a, b]$.

To complete the bound on the fourth term in $\epsilon_{D,i}(f)$, we now bound

$$\frac{1}{2} \delta^2 \mu c_S^2 \|f'''\| \mathbb{E} \int_0^1 S_i(t) \delta A[t, t+1(Q_i(t)=0)\Lambda_i(t)] dD_i(t) \quad (102)$$

$$+ \frac{1}{2} \delta^2 \mu c_S^2 \|f'''\| \mathbb{E} \int_0^1 S_i(t) \delta A(t+1(Q_i(t)=0)\Lambda_i(t), t+1(Q_i(t)=0)\Lambda_i(t) + S_i(t)) dD_i(t) \quad (103)$$

$$+ \frac{1}{2} \delta^2 \mu c_S^2 \|f'''\| \sum_{j \neq i} \mathbb{E} \int_0^1 S_i(t) \delta D_j[t, t+1(Q_i(t)=0)\Lambda_i(t) + S_i(t)] dD_i(t). \quad (104)$$

First, we bound $A[t, t+1(Q_i(t)=0)\Lambda_i(t)]$, the number of arrivals before server i gets a customer. Recall that our tie-breaking routing rule is to select among the least loaded servers uniformly at random. If server i becomes idle at time t , then an arrival will be routed to this server with probability at least $1/n$, and it is therefore possible to construct a geometrically distributed random variable $\Gamma \geq 1$ with mean n (success probability $1/n$) that a) upper bounds the number arrivals needed to route a customer to server i and b) is independent of $S_i(t)$ and $\{D_j(t), t \geq 0\}_{j=1}^n$. It follows that $A[t, t+1(Q_i(t)=0)\Lambda_i(t)] \leq \Gamma$. Furthermore, we define the time until the Γ th arrival as

$$\bar{\Lambda}_i(t) = R_a(t) + \sum_{m=2}^{\Gamma} U(\tau_m),$$

where τ_m is the time of the m arrival after time t , and note that

$$1(Q_i(t)=0)\Lambda_i(t) \leq \bar{\Lambda}_i(t).$$

We now have the tools needed to bound (102)–(104). First, we observe that

$$\begin{aligned} &\frac{1}{2} \delta^2 \mu c_S^2 \|f'''\| \mathbb{E} \int_0^1 S_i(t) \delta A[t, t+1(Q_i(t)=0)\Lambda_i(t)] dD_i(t) \\ &\leq \frac{1}{2} \delta^3 \mu c_S^2 \|f'''\| \mathbb{E} \int_0^1 S_i(t) \Gamma dD_i(t) = \frac{1}{2} \delta^3 \mu c_S^2 \|f'''\| \mathbb{E} S \mathbb{E} \Gamma \mathbb{E} D_i(1) = \frac{1}{2} \delta^3 \mu c_S^2 \|f'''\| n \rho, \end{aligned}$$

where in the last equality we used $\mathbb{E} \Gamma = n$, $\mathbb{E} S = 1/\mu$, and $\mathbb{E} D_i(1) = \lambda/n$. Second, we argue using Lorden's inequality, as we did in (101), that

$$\frac{1}{2} \delta^2 \mu c_S^2 \|f'''\| \mathbb{E} \int_0^1 S_i(t) \delta A(t+1(Q_i(t)=0)\Lambda_i(t), t+1(Q_i(t)=0)\Lambda_i(t) + S_i(t)) dD_i(t)$$

$$\begin{aligned}
 &\leq \frac{1}{2} \delta^3 \mu c_S^2 \|f'''\| \mathbb{E} \int_0^1 S(1 + \lambda S + \lambda^2 \mathbb{E} U^2) dD_i(t) \\
 &= \frac{1}{2} \delta^3 \mu c_S^2 \|f'''\| \mathbb{E}(S(1 + \lambda S + \lambda^2 \mathbb{E} U^2)) \mathbb{E} D_i(1) \\
 &= \frac{1}{2} \delta^3 \mu c_S^2 \|f'''\| (\lambda/n) (\mathbb{E} S + \lambda \mathbb{E} S^2 + \lambda^2 \mathbb{E} U^2 \mathbb{E} S).
 \end{aligned}$$

Third, we bound (104)

$$\begin{aligned}
 &\frac{1}{2} \delta^2 \mu c_S^2 \|f'''\| \sum_{j \neq i} \mathbb{E} \int_0^1 S_i(t) \delta D_j[t, t+1(Q_i(t)=0)\Lambda_i(t) + S_i(t)] dD_i(t) \\
 &\leq \frac{1}{2} \delta^3 \mu c_S^2 \|f'''\| \sum_{j \neq i} \mathbb{E} \int_0^1 S D_j[t, t + \bar{\Lambda}_i + S] dD_i(t) \\
 &\leq \frac{1}{2} \delta^3 \mu c_S^2 \|f'''\| (n-1) \mathbb{E}(S(1 + \mu(\bar{\Lambda}_i + S) + \mu^2 \mathbb{E} S^2)) \mathbb{E} D_i(1) \\
 &= \frac{1}{2} \delta^3 \mu c_S^2 \|f'''\| (n-1) (\mathbb{E} S + \mathbb{E} \bar{\Lambda}_i + 2\mu \mathbb{E} S^2) \lambda/n \\
 &\leq \frac{1}{2} \delta^3 \mu c_S^2 \|f'''\| \lambda (\mathbb{E} S + (\mathbb{E} R_a + \mathbb{E} U \mathbb{E} \Gamma) + 2\mu \mathbb{E} S^2) \\
 &= \frac{1}{2} \delta^3 c_S^2 \|f'''\| \lambda (1 + (\lambda \mu \mathbb{E} U^2 / 2 + 1/\rho) + 2\mu^2 \mathbb{E} S^2).
 \end{aligned}$$

Part two. We first prove (66):

$$\delta \mathbb{E} \left(\left(\lambda - \sum_{i=1}^n 1(Q_i > 0) \mu \right) f'(\tilde{X}) \right) = -n\mu \delta^2 \mathbb{E} f'(\tilde{X}) + \delta \mu \sum_{i=1}^n \mathbb{E} (1(Q_i = 0) f'(\tilde{X})).$$

Recalling that $\tilde{X} = \delta \sum_{i=1}^n Q_i - \delta \lambda R_a + \delta \sum_{i=1}^n \mu R_{s,i}$, we have

$$-n\mu \delta^2 \mathbb{E} f'(\tilde{X}) = -n\mu \delta^2 \mathbb{E} f'(X) - n\mu \delta^3 \mathbb{E} \left(\left(-\lambda R_a + \sum_{i=1}^n \mu R_{s,i} \right) f''(\xi) \right),$$

and

$$\begin{aligned}
 &\delta \mu \sum_{i=1}^n \mathbb{E} (1(Q_i = 0) f'(\tilde{X})) \\
 &= \delta \mu \sum_{i=1}^n \mathbb{E} \left(1(Q_i = 0) (f'(0) + \delta \left(\sum_{i=1}^n Q_i - \lambda R_a + \sum_{i=1}^n \mu R_{s,i} \right) f''(\xi)) \right),
 \end{aligned}$$

which proves (98). The proof of (99) is very similar to the proof of (9) in Proposition 1: if an arrival happens at time t , then $\Delta \tilde{X}(t-) = \delta(1 - \lambda U(t))$ and, since $\mathbb{E}(1 - \lambda U) = 0$,

$$\begin{aligned}
 \mathbb{E} \int_0^1 \Delta f(\tilde{X}(t-)) dA(t) &= \frac{1}{2} \delta^2 \mathbb{E} (1 - \lambda U)^2 \mathbb{E} \int_0^1 f''(\tilde{X}(t-)) dA(t) \\
 &\quad + \frac{1}{6} \delta^3 \mathbb{E} \int_0^1 (1 - \lambda U(t))^3 f'''(\xi(t)) dA(t),
 \end{aligned} \tag{105}$$

and the first term expands into

$$\begin{aligned} & \frac{1}{2}\delta^2\mathbb{E}(1-\lambda U)^2\mathbb{E}\int_0^1 f''(X(t-))dA(t) \\ & + \frac{1}{2}\delta^3\mathbb{E}(1-\lambda U)^2\mathbb{E}\int_0^1 (\tilde{X}(t-)-X(t-))f'''(\xi(t-))dA(t). \end{aligned}$$

Noting that $\mathbb{E}(1-\lambda U)^2 = c_U^2$ and $\tilde{X}(t-)-X(t-) = \sum_{i=1}^n \mu R_{s,i}(t)$, to conclude (99) we use Lemma 30 and repeat the argument following (11) to get

$$\begin{aligned} & \mathbb{E}\int_0^1 f''(X(t-))dA(t) \\ & = \lambda\mathbb{E}f''(X) - \lambda\mathbb{E}\int_0^1 \int_0^{U(t)} (X(t+u)-X(t-))f'''(\xi(t+u))dudA(t). \end{aligned}$$

Lastly we prove (100). Similar to (105) and the display that follows, we expand the departure jump term to get

$$\begin{aligned} \mathbb{E}\int_0^1 \Delta f(\tilde{X}(t-))dD_i(t) & = \frac{1}{2}\delta^2\mathbb{E}(1-\mu S)^2\mathbb{E}\int_0^1 f''(X(t-))dD_i(t) \\ & + \frac{1}{2}\delta^3\mathbb{E}(1-\mu S)^2\mathbb{E}\int_0^1 (\tilde{X}(t-)-X(t-))f'''(\xi(t-))dD_i(t) \\ & + \frac{1}{6}\delta^3\mathbb{E}\int_0^1 (1-\mu S_i(t))^3 f'''(\xi(t))dD_i(t), \end{aligned}$$

where $\tilde{X}(t-)-X(t-) = -\delta\lambda R_a(t) + \sum_{j \neq i} \delta\mu R_{s,j}(t)$. We conclude the proof by using Lemma 30 to get

$$\begin{aligned} \mathbb{E}f''(X) & = \mathbb{E}\int_0^1 \int_0^{1(Q_i(t)=0)\Lambda_i(t)} f''(X(t+u))dudD_i(t) + \mathbb{E}S\int_0^1 f''(X(t-))dD_i(t) \\ & + \mathbb{E}\int_0^1 \int_{1(Q_i(t)=0)\Lambda_i(t)}^{1(Q_i(t)=0)\Lambda_i(t)+S_i(t)} (X(t+u)-X(t-))f'''(\xi(t+u))dudD_i(t). \end{aligned}$$

□

References

- Asmussen S (2003) *Applied probability and queues*, volume 51 of *Applications of Mathematics (New York)* (New York: Springer-Verlag), second edition, ISBN 0-387-00211-1, Stochastic Modelling and Applied Probability.
- Baccelli F, Brémaud P (2003) *Elements of queueing theory: Palm martingale calculus and stochastic recurrences*, volume 26 of *Applications of Mathematics* (Berlin: Springer), 2nd edition.
- Barbour A (1990) Stein's method for diffusion approximations. *Probab. Theory and Related Fields* 84(3):297–322, ISSN 0178-8051, URL <http://dx.doi.org/10.1007/BF01197887>.

- Barbour A, Ross N, Zheng G (2023) Stein's method, gaussian processes and palm measures, with applications to queueing. *The Annals of Applied Probability* 33(5):3835–3871.
- Barbour AD (1988) Stein's method and Poisson process convergence. *Journal of Appl. Probab.* 25:175–184, ISSN 00219002, URL <http://www.jstor.org/stable/3214155>.
- Bassamboo A, Randhawa RS (2010) On the accuracy of fluid models for capacity sizing in queueing systems with impatient customers. *Operations Research* 58(5):1398–1413, URL <http://dx.doi.org/10.1287/opre.1100.0815>.
- Bertsimas D, Gamarnik D (2022) *Queueing theory: classical and modern methods* (Belmont, MA: Dynamic Ideas LLC), 1st edition.
- Besançon E, Decreusefond L, Moyal P (2020) Stein's method for diffusive limits of queueing processes. *Queueing Systems* 95:173–201.
- Blanchet J, Glynn P (2006) Complete corrected diffusion approximations for the maximum of a random walk. *Annals of Applied Probability* 16(2):951–983.
- Boon M, Janssen A, van Leeuwen J (2023) Heavy-traffic single-server queues and the transform method. *Indagationes Mathematicae* 34(5):1014–1037.
- Bramson M (2011) Stability of join the shortest queue networks. *Ann. Appl. Probab.* 21(4):1568–1625, URL <http://dx.doi.org/10.1214/10-AAP726>.
- Braverman A (2022) The prelimit generator comparison approach of stein's method. *Stochastic Systems* 12(2):181–204, URL <http://dx.doi.org/10.1287/stsy.2021.0085>.
- Braverman A, Dai J, Miyazawa M (2017) Heavy traffic approximation for the stationary distribution of a generalized Jackson network: the BAR approach. *Stochastic Systems* 7(1):143–196, URL <http://projecteuclid.org/euclid.ssy/1495785619>.
- Braverman A, Dai J, Miyazawa M (2024) The BAR approach for multiclass queueing networks with SBP service policies. *Stochastic Systems* .
- Braverman A, Dai JG (2017) Stein's method for steady-state diffusion approximations of $M/Ph/n + M$ systems. *Ann. of Appl. Probab.* 27(1):550–581, ISSN 1050-5164, URL <http://dx.doi.org/10.1214/16-AAP1211>.
- Braverman A, Dai JG, Feng J (2016) Stein's method for steady-state diffusion approximations: An introduction through the Erlang-A and Erlang-C models. *Stoch. Syst.* 6:301–366, URL <http://www.i-journals.org/ssy/viewarticle.php?id=212&layout=abstract>.

- Brown TC, Xia A (2001) Stein's method and birth-death processes. *Ann. Probab.* 29(3):1373–1403, URL <http://dx.doi.org/10.1214/aop/1015345606>.
- Chen LHY, Goldstein L, Shao QM (2011) *Normal approximation by Stein's method*. Probability and its Applications (New York) (Springer, Heidelberg), ISBN 978-3-642-15006-7, URL <http://dx.doi.org/10.1007/978-3-642-15007-4>.
- Chen Y, Whitt W (2020) Extremal models for the $g_i/g_i/k$ waiting-time tail-probability decay rate. *Operations Research Letters* 48(6):770–776.
- Chen Y, Whitt W (2021) Extremal $g_i/g_i/1$ queues given two moments: exploiting tchebycheff systems. *Queueing Systems* 97(1):101–124.
- Chen Y, Whitt W (2022a) Correction to: Extremal $g_i/g_i/1$ queues given two moments: exploiting tchebycheff systems. *Queueing Systems* 102(3):553–556.
- Chen Y, Whitt W (2022b) Set-valued performance approximations for the queue given partial information. *Probability in the Engineering and Informational Sciences* 36(2):378–400.
- Dai JG, Glynn P, Xu Y (2025) Asymptotic product-form steady-state for generalized jackson networks in multi-scale heavy traffic. URL <https://arxiv.org/abs/2304.01499>.
- Dai JG, Guang J, Xu Y (2024) Steady-state convergence of the continuous-time jsq system with general distributions in heavy traffic. *SIGMETRICS Perform. Eval. Rev.* 52(2):39–41, ISSN 0163-5999, URL <http://dx.doi.org/10.1145/3695411.3695426>.
- Dai JG, Hasenbein JJ, Vande Vate JH (2004) Stability and instability of a two-station queueing network. *Annals of Applied Probability* 14:326–377, ISSN 1050-5164.
- Dai JG, Meyn SP (1995) Stability and convergence of moments for multiclass queueing networks via fluid limit models. *IEEE Transactions on Automatic Control* 40:1889–1904.
- Dai JG, Xu Y (2024) Explicit steady-state approximations for parallel server systems with heterogeneous servers. URL <https://arxiv.org/abs/2406.04203>.
- Daley D, Kreinin AY, Trengove C (1992) Inequalities concerning the waiting time in single-server queues: A survey. *Oxford Statistical Science Series* 1(9):177–177.
- Davis MHA (1984) Piecewise deterministic Markov processes: a general class of non-diffusion stochastic models. *Journal of Royal Statist. Soc. series B* 46:353–388.

- Eryilmaz A, Srikant R (2012) Asymptotically tight steady-state queue length bounds implied by drift conditions. *Queueing Systems* 72(3-4):311–359, ISSN 0257-0130, URL <http://dx.doi.org/10.1007/s11134-012-9305-y>.
- Feng J, Shi P (2018) Steady-state diffusion approximations for discrete-time queue in hospital inpatient flow management. *Naval Research Logistics (NRL)* 65(1):26–65, URL <http://dx.doi.org/10.1002/nav.21787>.
- Gast N (2017) Expected values estimated via mean-field approximation are $1/n$ -accurate. *Proceedings of the ACM on Measurement and Analysis of Computing Systems* 1(1):1–26.
- Gaunt RE, Walton N (2020) Stein's method for the single server queue in heavy traffic. *Statistics & Probability Letters* 156:108566, ISSN 0167-7152, URL <http://dx.doi.org/https://doi.org/10.1016/j.spl.2019.108566>.
- Götze F (1991) On the rate of convergence in the multivariate CLT. *Ann. Probab.* 19(2):724–739, URL <http://dx.doi.org/10.1214/aop/1176990448>.
- Gurvich I (2014) Diffusion models and steady-state approximations for exponentially ergodic Markovian queues. *Ann. Appl. Probab.* 24(6):2527–2559, URL <http://dx.doi.org/10.1214/13-AAP984>.
- 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, URL <http://dx.doi.org/10.1287/moor.2013.0606>.
- Gut A (1974) On the moments and limit distributions of some first passage times. *The Annals of Probability* 277–308.
- Harrison JM, Reiman MI (1981) Reflected Brownian motion on an orthant. *Ann. Probab.* 9(2):302–308, ISSN 0091-1798, URL [http://links.jstor.org/sici?sici=0091-1798\(198104\)9:2<302:RBMOAO>2.0.CO;2-P&origin=MSN](http://links.jstor.org/sici?sici=0091-1798(198104)9:2<302:RBMOAO>2.0.CO;2-P&origin=MSN).
- Huang J, Gurvich I (2018) Beyond heavy-traffic regimes: Universal bounds and controls for the single-server queue. *Oper. Res.* 66(4):1168–1188, URL <http://dx.doi.org/10.1287/opre.2017.1715>.
- Hurtado-Lange D, Maguluri ST (2022) A load balancing system in the many-server heavy-traffic asymptotics. *Queueing Systems* 101(3):353–391.

- Hurtado Lange DA, Maguluri ST (2022) Heavy-traffic analysis of queueing systems with no complete resource pooling. *Mathematics of Operations Research* 47(4):3129–3155, URL <http://dx.doi.org/10.1287/moor.2021.1248>.
- Kingman JFC (1961) The single server queue in heavy traffic. *Mathematical Proceedings of the Cambridge Philosophical Society* 57:902–904, URL <http://dx.doi.org/10.1017/S0305004100036094>.
- Kingman JFC (1962) On queues in heavy traffic. *J. Roy. Statist. Soc. Ser. B* 24:383–392, ISSN 0035-9246, URL [http://links.jstor.org/sici?sici=0035-9246\(1962\)24:2<383:0QIHT>2.0.CO;2-A&origin=MSN](http://links.jstor.org/sici?sici=0035-9246(1962)24:2<383:0QIHT>2.0.CO;2-A&origin=MSN).
- Köllerström J (1976) Stochastic bounds for the single-server queue. *Mathematical Proceedings of the Cambridge Philosophical Society* 80(3):521–525, URL <http://dx.doi.org/10.1017/S0305004100053135>.
- Li J, Ou J (1995) Characterizing the idle-period distribution of gi/g/1 queues. *Journal of applied probability* 32(1):247–255.
- Lieberman GM (2013) *Oblique Derivative Problems for Elliptic Equations* (WORLD SCIENTIFIC), URL <http://dx.doi.org/10.1142/8679>.
- Lorden G (1970) On excess over the boundary. *The Annals of Mathematical Statistics* 41(2):520 – 527, URL <http://dx.doi.org/10.1214/aoms/1177697092>.
- Loulou R (1978) An explicit upper bound for the mean busy period in a $GI/G/1$ queue. *Journal of Applied Probability* 15(2):452–455, URL <http://dx.doi.org/10.2307/3213419>.
- Maguluri ST, Burle SK, Srikant R (2018) Optimal heavy-traffic queue length scaling in an incompletely saturated switch. *Queueing Systems* 88(3):279–309.
- Maguluri ST, Srikant R (2016) Heavy traffic queue length behavior in a switch under the maxweight algorithm. *Stochastic Systems* 6(1):211–250, URL <http://dx.doi.org/10.1214/15-SSY193>.
- Miyazawa M (2015) Diffusion approximation for stationary analysis of queues and their networks: a review. *J. Oper. Res. Soc. Japan* 58(1):104–148, ISSN 0453-4514, URL <http://dx.doi.org/10.15807/jorsj.58.104>.
- Miyazawa M (2017) A unified approach for large queue asymptotics in a heterogeneous multiserver queue. *Adv. in Appl. Probab.* 49(1):182–220, ISSN 0001-8678, URL <http://dx.doi.org/10.1017/apr.2016.84>.

- Peköz EA, Röllin A (2011) New rates for exponential approximation and the theorems of Rényi and Yaglom. *The Annals of Probability* 39(2):587 – 608, URL <http://dx.doi.org/10.1214/10-AOP559>.
- Ross N (2011) Fundamentals of Stein's method. *Probab. Surv.* 8:210–293, ISSN 1549-5787, URL <http://dx.doi.org/10.1214/11-PS182>.
- Siegmund D (1979) Corrected diffusion approximations in certain random walk problems. *Advances in Applied Probability* 11(4):701–719, ISSN 00018678, URL <http://www.jstor.org/stable/1426855>.
- Stein C (1972) A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Probability Theory*, 583–602 (Berkeley, Calif.: University of California Press), URL <http://projecteuclid.org/euclid.bsmmsp/1200514239>.
- Stolyar AL (2015) Tightness of stationary distributions of a flexible-server system in the Halfin-Whitt asymptotic regime. *Stoch. Syst.* 5(2):239–267, URL <http://dx.doi.org/10.1214/14-SSY139>.
- Ward A, Glynn P (2003) A diffusion approximation for a markovian queue with reneging. *Queueing Systems* 43(1-2):103–128, ISSN 0257-0130, URL <http://dx.doi.org/10.1023/A:1021804515162>.
- Whitt W (1986) Deciding which queue to join: Some counterexamples. *Oper. Res.* 34(1):55–62, ISSN 0030-364X.
- Wolff RW, Wang CL (2003) Idle period approximations and bounds for the gi/g/1 queue. *Advances in Applied Probability* 35(3):773–792.
- Ying L (2016) On the approximation error of mean-field models. *Proceedings of the 2016 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Science*, 285–297 (Antibes Juan-les-Pins, France: ACM), URL <http://dx.doi.org/10.1145/2964791.2901463>.
- Ying L (2017) Stein's method for mean field approximations in light and heavy traffic regimes. *Proc. ACM Meas. Anal. Comput. Syst.* 1(1):12:1–12:27, ISSN 2476-1249, URL <http://dx.doi.org/10.1145/3084449>.
- Zhou X, Shroff N (2020) A note on stein's method for heavy-traffic analysis. *arXiv preprint arXiv:2003.06454* .