Robust Queueing for Single-Server Queues with Abandonment

Wei You

Department of Industrial Engineering and Decision Analytics, The Hong Kong University of Science and Technology, Hong Kong SAR, China, weiyou@ust.hk

Abstract

Single-server queues with customer abandonment arise in call centers and many service systems, but steady-state performance measures remain analytically intractable beyond Markovian assumptions. This paper develops Robust Queueing (RQ) approximations for the mean steady-state *virtual waiting time* (offered waiting time) in the $GI/GI/1{+}GI$ model. The approach starts from a reverse-time supremum representation of the virtual waiting time as the reflection of an *effective* net-input process that accounts for abandonments. We approximate effective net-input increments by their mean plus a robustness parameter times their standard deviation. For the drift, we introduce a Poisson-surrogate compensator and show that the associated correction term is asymptotically negligible in the long-patience regime. For variability, we propose two implementable surrogates: (i) a deterministic time-change approximation that yields a first RQ algorithm, and (ii) a refined algorithm based on a heavy-traffic limit that produces a scale-dependent variance function capturing the variance-reduction effect of abandonment. The resulting steady-state approximation reduces to a one-dimensional fixed point solvable by bisection and takes as input the arrival index of dispersion for counts (IDC), the service-time squared coefficient of variation, and the patience-time distribution. We further show how to extend the method to queues in series by feeding an approximation of the upstream departure IDC into the downstream RQ algorithm. Extensive numerical experiments demonstrate that the refined RQ approximation is accurate across underload, critical loading, and overload, and remains robust relative to existing heavy-traffic and hazard-rate-scaling benchmarks.

Keywords: robust queueing, customer abandonment, virtual waiting time, indices of dispersion, heavy-traffic limits

1 Introduction

Customer impatience and abandonment (reneging) are defining features of many modern service systems, including call centers, healthcare delivery, and online service platforms. In these settings, customers may leave without receiving service when delays are perceived as too long, altering both operational efficiency and quality of service. Much of the call-center literature therefore models systems with abandonment; see, e.g., the many-server asymptotic analysis in [30], the Erlang-A

call-center model in [8], and the survey in [6]. While these applications often involve many servers, single-server models remain important as building blocks for more complex service networks (e.g., sequential service stages) and as primitives in decomposition approximations.

In this paper we study the classical first-come-first-served $GI/GI/1{+}GI$ queue with customer abandonment. Arrivals follow a renewal process, service requirements are i.i.d. with a general distribution, and patience times are i.i.d. with a general distribution; a customer abandons if service has not begun by the time her patience expires. We focus on the *virtual waiting time* (also called the *offered waiting time*), denoted by $Z(t)$, and in particular on the stationary mean $\mathbb{E}[Z(\infty)]$. The virtual waiting time is a fundamental performance metric because it directly summarizes the system's congestion, underlies delay announcements, and can be used to approximate related quantities such as the probability of abandonment and mean queue length (see, e.g., [14, 15, 21]).

Despite its apparent simplicity, the $GI/GI/1{+}GI$ model is analytically challenging. Abandonment creates a nonlinear feedback loop: the waiting time affects which customers remain in queue, which in turn changes the future workload seen by subsequent arrivals. Exact steady-state descriptions are available only in special cases; early work establishing structural relations between actual and virtual waiting times includes [1, 20]. For general primitives, one typically relies on asymptotic approximations, numerical schemes, or simulation.

We develop new *robust queueing* (RQ) approximations for $\mathbb{E}[Z(\infty)]$ that are fast, require only low-dimensional traffic descriptors, and remain accurate away from classical heavy-traffic regimes. Our approach builds on the stochastic RQ methodology that approximates single-server performance using reverse-time supremum representations and variability summaries in the form of *indices of dispersion* (e.g., [7, 26, 27]). These ideas have been extended to open networks via IDC-based flow propagation, yielding an RQ network analyzer analogous in spirit to the classical queueing network analyzer (QNA) [24]; see [29] and references therein.

1.1 Literature Review

Exact analysis and structural properties. In single-server queues with deadlines or patience times, early work analyzed reneging and established fundamental stability and distributional relations; see, e.g., [1, 20]. Even for Poisson arrivals, general patience times lead to integral-equation characterizations rather than closed forms, and tractable steady-state formulas are typically restricted to Markovian special cases.

Heavy-traffic diffusion approximations. A major line of work develops diffusion approximations for $GI/GI/1{+}GI$ queues using the offered waiting-time process. Ward and Glynn [23] show that, under conventional heavy-traffic scaling, the offered waiting time can be approximated by a regulated Ornstein–Uhlenbeck diffusion. Reed and Ward [19] introduce *hazard-rate scaling* so that the heavy-traffic diffusion limit incorporates the full patience-time distribution through a nonlinear drift term. Lee and Weerasinghe [13] further establish heavy-traffic convergence for general patience-time distributions (allowing, e.g., state-dependent arrival intensities) and derive related

limits for queue length. A broader perspective on asymptotic regimes for queues with reneging, including conventional heavy traffic, the Halfin–Whitt regime [9], and overload, is provided in the survey [21].

Using the stationary distribution of a diffusion limit as a proxy for the steady-state queue requires an *interchange of limits* justification. For the $GI/GI/1{+}GI$ model, [14] establishes convergence of the scaled stationary offered-waiting-time distribution (and moments) to the stationary distribution of the limiting diffusion, resolving a question left open in [23]. This result is extended under more general patience-time scaling (including hazard-rate scaling) in [15].

Beyond diffusion limits derived from specific heavy-traffic parameter scalings, [10] develops *universal* performance bounds and diffusion-based approximations for the $M/GI/1{+}GI$ queue that are valid uniformly over families of patience distributions and across heavy-traffic regimes. These results provide complementary support for diffusion proxies, but they still rely on Markovian arrivals and do not directly address nonrenewal inputs arising endogenously in networks.

Robust queueing and indices of dispersion. Indices of dispersion for counts/work were introduced as variability summaries of offered traffic and were used to predict mean workload in single-server queues in [7]. The robust queueing approach in [26] shows how to convert IDC/IDW information into accurate approximations for mean steady-state workload in the general $G/G/1$ queue, with asymptotic correctness in light and heavy traffic and the ability to capture temporal dependence. Subsequent work emphasizes the value of IDC-based descriptions [27] and extends the methodology to open networks through IDC propagation [29]. Robust queueing has also been pursued from a robust-optimization perspective (e.g., [2]), but our focus is on developing stochastic-model performance approximations in the IDC-based RQ framework.

1.2 Contributions and Organization

Our contributions are as follows.

- We derive an RQ approximation for the mean steady-state virtual waiting time in the $GI/GI/1{+}GI$ queue by expressing $Z(\infty)$ as the reflection of an *effective* net-input process that counts only work that will eventually be served.
- We propose an implementable *drift approximation* for this effective net input via a Poisson-surrogate compensator. The correction term is exact for Poisson arrivals and is asymptotically negligible for renewal arrivals in a long-patience regime.
- We develop two implementable *variance surrogates* for the effective net input. The first is a deterministic time-change approximation; the second is a refined approximation justified by the heavy-traffic diffusion limit and incorporates a scale-dependent variance function capturing the variance-reduction effect of abandonment.
- We show how the resulting refined RQ approximation can be extended to queues in series: the downstream queue can be analyzed using its arrival IDC, and we approximate that

IDC by propagating variability through the upstream queue via existing departure-IDC approximations.

The remainder of the paper is organized as follows. Section 2 reviews robust queueing for the $G/GI/1$ model and develops a reverse-time representation for the $GI/GI/1{+}GI$ virtual waiting time. Section 3 develops drift approximations for the effective net input. Sections 4 and 5 present the crude and refined RQ approximations, respectively, and discuss calibration of the RQ parameter. Section 6 reports numerical experiments for single queues and tandem configurations, comparing against classical diffusion and bounds-based approximations.

2 Preliminary

2.1 Review of Robust Queueing for the $G/GI/1$ Model

Consider a single-server $G/GI/1$ queue with infinite waiting room and a first-in-first-out (FIFO) service discipline. Let $A(t)$ denote a stationary and ergodic arrival counting process with rate λ, and assume $\mathrm{Var}(A(t)) < \infty$ for all $t > 0$. Let $\{V_i\}_{i\geq 1}$ be an i.i.d. sequence of service times, independent of $A(\cdot)$, with mean $1/\mu$ and finite variance. Define the traffic intensity $\rho \triangleq \lambda/\mu$.

A convenient workload representation is based on the cumulative work (total input) process $\tilde{Y}(t) \triangleq \sum_{k=1}^{A(t)} V_k$, and the associated net-input process (under unit service capacity) $\tilde{N}(t) \triangleq \tilde{Y}(t) - t$. For a system that starts empty at time 0, the workload process $\tilde{Z}(t)$ is given by the Skorokhod mapping applied to $\tilde{N}(\cdot)$:

$$\tilde{Z}(t) = \tilde{N}(t) - \inf_{0\leq s\leq t} \tilde{N}(s) = \sup_{0\leq s\leq t} \{\tilde{N}(t) - \tilde{N}(t-s)\}. \tag{1}$$

In words, $\tilde{Z}(t)$ is the reflected net-input process, and (1) expresses the workload as the running supremum of *reverse-time* net-input increments.

Exact analysis of (1) is typically intractable in non-Markovian settings, which motivates approximations such as Robust Queueing (RQ) and Brownian queues. The RQ approximation proceeds by replacing the stochastic process inside the supremum in (1) with a deterministic surrogate, which turns the stochastic optimization into a deterministic one that is tractable numerically. Concretely, one approximates the reverse-time increment $\tilde{N}(t) - \tilde{N}(t-s)$ by its (stationary) mean plus a multiple b of its standard deviation:

$$\mathbb{E}\left[\tilde{N}(t) - \tilde{N}(t-s)\right] + b \cdot \mathrm{SD}\left(\tilde{N}(t) - \tilde{N}(t-s)\right) = -(1-\rho)s + b\sqrt{\rho s \tilde{I}_w(s)/\mu},$$

where $\mathrm{SD}(\cdot)$ denotes standard deviation, and

$$\tilde{I}_w(t) \triangleq \frac{\mathrm{Var}(\tilde{Y}(t))}{\mathbb{E}[\tilde{Y}(t)]\mathbb{E}[V_1]} = \frac{\mathrm{Var}(\tilde{Y}(t))}{\rho t/\mu}$$

is the *index of dispersion for work* (IDW) associated with the stationary version of $\tilde{Y}(t)$. The reverse-time formulation (1) is essential here: RQ approximates the *increment process* $\tilde{N}(t) - \tilde{N}(t-s)$ rather than $\tilde{N}(t)$ itself.

Remark 1 (Comparison to Brownian queues)**.** *Another classical approximation is the Brownian queue, which replaces the reverse-time net-input increment by a Brownian motion with negative drift,*

$$-(1-\rho)s + \sqrt{\rho \tilde{I}_w(\infty)/\mu} B(s),$$

where $B(\cdot)$ is a standard Brownian motion and $\tilde{I}_w(\infty) \triangleq \lim_{t\to\infty} \tilde{I}_w(t)$ (when the limit exists). In a Brownian queue, the drift is linear and the diffusion coefficient is constant; consequently, approximating $\tilde{N}(t)$ and approximating increments $\tilde{N}(t) - \tilde{N}(t-s)$ are effectively equivalent (due to stationary, independent increments). This equivalence does not *carry over to RQ: because the RQ surrogate depends on the scale-dependent variability encoded by $\tilde{I}_w(s)$, it is critical to apply the reverse-time representation first and then approximate the increment process inside the supremum.*

2.2 The Dynamics of the $G/GI/1{+}GI$ Model

We now turn to the $G/GI/1{+}GI$ queue with abandonment. Customers arrive according to a general right-continuous counting process $A(t)$. Let U denote a generic interarrival time. Each customer has an i.i.d. service time V and an i.i.d. patience time D. A customer abandons if it has not entered service before its patience time expires; the notation $+GI$ indicates that the patience-time distribution is general.

The dynamics are most conveniently described in terms of the *virtual waiting time process* $Z(t)$, defined as the waiting time at time t of a hypothetical customer arriving at t with infinite patience (also called the *offered waiting time*). Let $T_i \triangleq \inf\{t > 0 : A(t) = i\}$ be the arrival time of the ith customer. Given $Z(\cdot)$, customer i is offered waiting time $W_i \triangleq Z(T_i-)$, and is eventually served if and only if $D_i > W_i$.

In contrast to the $G/GI/1$ model without abandonment, the evolution of $Z(t)$ is driven by the *effective* workload that will eventually be processed by the server. Define the *effective arrival process*

$$A_0(t) \triangleq \sum_{k=1}^{A(t)} \mathbb{1}(D_k > W_k), \tag{2}$$

the number of arrivals by time t who will eventually enter service. Closely related is the effective total-input process

$$Y(t) \triangleq \sum_{k=1}^{A(t)} V_k \mathbb{1}(D_k > W_k),$$

which counts the total amount of work brought by customers arriving by time t who do not abandon. The corresponding effective net-input process is

$$N(t) \triangleq Y(t) - t. \tag{3}$$

Following the same reflection argument that yields (1), assuming the system starts empty at time 0, the virtual waiting time admits the reverse-time (supremum) representation

$$Z(t) = \sup_{0\le s\le t} \{N(t) - N(t-s)\}, \tag{4}$$

where $N(\cdot)$ is defined in (3).

We impose the following assumptions.

Assumption 1. *1. Arrivals occur one at a time. The arrival process $A(\cdot)$ is a stationary renewal process with rate λ and index of dispersion for counts*

$$I_a(t) \triangleq \frac{\mathrm{Var}(A(t))}{\lambda t}.$$

We assume the long-run limit exists and is finite: $I_a(\infty) = c_a^2 < \infty$.

2. The service times are i.i.d. with mean $1/\mu$, finite variance, and squared coefficient of variation (SCV) c_s^2.

3. The patience times are i.i.d. and admit the scaling representation $D = \alpha^{-1}\tilde{D}$, where $\tilde{D}$ is a nonnegative random variable with $\mathbb{E}[\tilde{D}] = 1$. Under this scaling convention, the CDF of D is $F_\alpha(t) = F(\alpha t)$. Let F denote the cumulative distribution function of $\tilde{D}$. We assume F is twice continuously differentiable on $[0, \infty)$, has support $[0, \infty)$, and its density f satisfies $f(0) < \infty$.

4. The arrival process, service times, and patience times are mutually independent.

The reverse-time representation (4) naturally motivates a RQ approximation. Specifically, we define the RQ approximation as

$$Z_{\mathrm{RQ}}(t) = \sup_{0 \le s \le t} \Big\{ \mathbb{E}[N(t) - N(t-s)] + b \cdot \mathrm{SD}\big(N(t) - N(t-s)\big) \Big\}, \tag{5}$$

where $b \ge 0$ is the robustness parameter.

Comparing (1) and (4), the essential distinction lies in the underlying net-input process. For the $GI/GI/1{+}GI$ model, the object of interest is the *effective* net-input process (3), which depends on the abandonment indicators $\mathbb{1}(D_k > W_k)$ and hence on the offered waiting times themselves. To operationalize (5), it therefore remains to characterize the mean (drift) and variability of the increment process $N(t) - N(t-s)$. Section 3 develops an approximation for the drift term $\mathbb{E}[N(t) - N(t-s)]$, and Section 5 focuses on the corresponding variance function.

3 The Drift of the Effective Net-Input Process

To evaluate the mean of the effective net-input process, it is convenient to exploit martingale representations of counting processes. Let $A(\cdot)$ be an integrable counting process with rate λ, adapted to a filtration $\mathbb{F} = \{\mathcal{F}_t\}_{t \ge 0}$. By the Doob–Meyer decomposition [12, Theorem 10.5], $M_A(t) \triangleq A(t) - \Lambda(t)$ is an $\mathbb{F}$-martingale, where $\Lambda(\cdot)$ is the (predictable) compensator of $A(\cdot)$. This representation is particularly tractable when $A(\cdot)$ is a homogeneous Poisson process, in which case $\Lambda(t) = \lambda t$. We begin with this Poisson setting.

3.1 Queues with Poisson arrivals

Let $\bar{F}_\alpha(t) \triangleq 1-F_\alpha(t)$ denote the complementary CDF of the patience time. Recall the effective arrival process $A_0(\cdot)$ defined in (2), and the effective net-input process is $N(t)$ defined in (3). Throughout this subsection, let $\mathbb{F} = \{\mathcal{F}_t\}_{t\geq 0}$ denote the natural filtration generated by the primitives up to time t (arrivals and the associated marks). In particular, $Z(t-)$ is $\mathbb{F}$-predictable.

Under Poisson arrivals, consider an arrival occurring at time u. Conditional on the pre-arrival history $\mathcal{F}_{u-}$, the offered waiting time $Z(u-)$ is known, while the patience time D is independent of $\mathcal{F}_{u-}$. Therefore,

$$\mathbb{P}\left(D > Z(u-) \mid \mathcal{F}_{u-}\right) = \bar{F}_\alpha(Z(u-)).$$

In other words, relative to the filtration $\mathbb{F}$, the process $A_0(\cdot)$ is obtained from the Poisson arrivals via *predictable thinning* with retention probability $\bar{F}_\alpha(Z(u-))$. Consequently, $A_0(\cdot)$ has $\mathbb{F}$-intensity $\lambda\bar{F}_\alpha(Z(u-))$, and its compensator is

$$\Lambda_0(t) \triangleq \lambda \int_0^t \bar{F}_\alpha(Z(u-))du.$$

Equivalently, $M_0(t) \triangleq A_0(t) - \Lambda_0(t)$ is an $\mathbb{F}$-martingale.

For the effective work process, conditional on $\mathcal{F}_{u-}$, the service time mark V is independent of $\mathcal{F}_{u-}$ and independent of D, so

$$\mathbb{E}\left[V\mathbb{1}\{D > Z(u-)\} \mid \mathcal{F}_{u-}\right] = \mathbb{E}[V]\bar{F}_\alpha(Z(u-)) = \frac{1}{\mu}\bar{F}_\alpha(Z(u-)).$$

Thus the compensator of $Y(\cdot)$ is

$$\Lambda_Y(t) \triangleq \frac{\lambda}{\mu} \int_0^t \bar{F}_\alpha(Z(u-))du.$$

Lemma 1. *Under Assumption 1, suppose that $A(\cdot)$ is a homogeneous Poisson process with rate λ. Then, for any $0 \leq s \leq t$,*

$$\mathbb{E}\left[A_0(t) - A_0(t-s)\right] = \lambda\mathbb{E}\left[\int_{t-s}^t \bar{F}_\alpha(Z(u-))du\right],$$
$$\mathbb{E}\left[N(t) - N(t-s)\right] = \frac{\lambda}{\mu}\mathbb{E}\left[\int_{t-s}^t \bar{F}_\alpha(Z(u-))du\right] - s.$$

Moreover, if $\{Z(u)\}_{u\in\mathbb{R}}$ is strictly stationary, then for all $0 \leq s \leq t$,

$$\mathbb{E}\left[A_0(t) - A_0(t-s)\right] = \lambda s\mathbb{E}\left[\bar{F}_\alpha(Z(0-))\right],$$
$$\mathbb{E}\left[N(t) - N(t-s)\right] = \left(\frac{\lambda}{\mu}\mathbb{E}\left[\bar{F}_\alpha(Z(0-))\right] - 1\right)s.$$

3.2 Queues with a General Renewal Arrival Process

Motivated by Lemma 1, we introduce the Poisson surrogate

$$\Lambda_t(s) \triangleq \lambda \int_{t-s}^t \bar{F}_\alpha(Z(u-))du, \tag{6}$$

which coincides with the compensator increment of the effective arrival process in the Poisson case. For a general renewal arrival process, we can write the exact decomposition

$$\int_{t-s}^{t} \bar{F}_\alpha(Z(u-))dA(u) = \Lambda_t(s) + \delta_t(s), \quad \text{where} \quad \delta_t(s) \triangleq \int_{t-s}^{t} \bar{F}_\alpha(Z(u-))d(A(u) - \lambda u).$$

When $A(\cdot)$ is Poisson, $A(u) - \lambda u$ is a martingale and the predictability of $Z(u-)$ implies $\mathbb{E}[\delta_t(s)] = 0$, recovering Lemma 1. For a general renewal process, however, $A(u) - \lambda u$ is not a martingale, and in general $\mathbb{E}[\delta_t(s)] \neq 0$. We therefore seek an approximation to $\mathbb{E}[\delta_t(s)]$. In particular, we show that the *expected* correction term vanishes in the long-patience/time-stationary regime as the abandonment rate $\alpha \to 0$ and $t \to \infty$.

Lemma 2. *Suppose that $A(\cdot)$ is a renewal counting process with rate λ. Then, for any fixed $s \geq 0$,*

$$\lim_{\alpha \downarrow 0} \lim_{t \to \infty} \mathbb{E}[\delta_t(s)] = 0.$$

Combining Lemma 1 and Lemma 2, we approximate the drift of the effective total-input increment and the effective net-input increment under renewal arrivals by neglecting the mean correction term:

$$\begin{aligned} \mathbb{E}\left[Y(t) - Y(t-s)\right] &= \frac{1}{\mu}\mathbb{E}\left[\int_{t-s}^{t} \bar{F}_\alpha(Z(u-))dA(u)\right] \approx \frac{1}{\mu}\mathbb{E}[\Lambda_t(s)], \\ \mathbb{E}\left[N(t) - N(t-s)\right] &= \mathbb{E}\left[Y(t) - Y(t-s)\right] - s \approx \frac{1}{\mu}\mathbb{E}[\Lambda_t(s)] - s. \end{aligned} \tag{7}$$

For stationary versions, we let $t \to \infty$. In steady state,

$$\mathbb{E}[\Lambda_t(s)] = \lambda \int_{t-s}^{t} \mathbb{E}\left[\bar{F}_\alpha(Z(u-))\right] du = \lambda s \mathbb{E}\left[\bar{F}_\alpha(Z(0))\right],$$

and we define the corresponding stationary drift surrogate

$$\Lambda^*(s) \triangleq \lambda s \mathbb{E}\left[\bar{F}_\alpha(Z(0))\right]. \tag{8}$$

Here $Z(\cdot)$ denotes the stationary virtual waiting time process.

Lemma 3. *Let $\xi \triangleq F^{-1}((\rho-1)/\rho)\mathbb{1}\{\rho \geq 1\}$, where $F^{-1}(y) \triangleq \inf\{x \geq 0 : F(x) \geq y\}$ is the generalized inverse. Then $\lim_{\alpha\downarrow 0} \lim_{t\to\infty} \alpha Z(t) = \xi$.*

4 A First Robust Queueing Algorithm

We now propose our first approximation for the variance of the effective net-input increment $N(t) - N(t-s)$ based on a (deterministic) time-change of the renewal arrival process. Combining this variance approximation with the drift approximation from Section 3 yields our first Robust Queueing (RQ) algorithm for the virtual waiting time.

Recall that the effective net-input process is a thinned renewal reward process: an arrival contributes service work if and only if its patience time exceeds the offered waiting time at arrival. Although patience times are independent of the system primitives, the thinning decision $\mathbb{1}\{D_k > W_k\}$ is correlated with the arrival process through the offered waiting time $W_k = Z(T_k-)$. As an approximation, we treat this correlation as negligible and model the effective input over $(t-s, t]$ via a stationary renewal reward process evaluated at a deterministically rescaled time.

4.1 The First RQ Algorithm

Let $\tilde{A}(\cdot)$ denote a *rate-one* version of the renewal arrival process, i.e., if $A(\cdot)$ has i.i.d. interarrival times U with $\mathbb{E}[U] = 1/\lambda$, then $\tilde{A}(\cdot)$ is the renewal counting process with interarrival times $\tilde{U} \triangleq \lambda U$ so that $\mathbb{E}[\tilde{U}] = 1$ and $\mathbb{E}[\tilde{A}(t)] = t$. Motivated by the Poisson case and the drift surrogate $\Lambda_t(s)$ in (6), we approximate the effective net-input increment by

$$N(t) - N(t-s) = \sum_{k=A(t-s)+1}^{A(t)} V_k \mathbb{1}\{D_k > W_k\} - s \approx \sum_{k=1}^{\tilde{A}(\Lambda_t(s))} V_k - s,$$

where $\Lambda_t(s) = \lambda \int_{t-s}^{t} \bar{F}_\alpha(Z(u-))du$ is defined in (6). Under this approximation,

$$\mathrm{Var}(N(t) - N(t-s)) \approx \mathrm{Var}\left(\sum_{k=1}^{\tilde{A}(\Lambda_t(s))} V_k\right) = \frac{\Lambda_t(s)}{\mu^2} I_w(\Lambda_t(s)), \tag{9}$$

where $I_w(\cdot)$ is the IDW associated with $\tilde{A}$ and the service times $\{V_k\}$:

$$I_w(t) \triangleq \frac{\mathrm{Var}\left(\sum_{k=1}^{\tilde{A}(t)} V_k\right)}{\mathbb{E}\left[\sum_{k=1}^{\tilde{A}(t)} V_k\right] \mathbb{E}[V_1]} = \frac{\mathrm{Var}\left(\sum_{k=1}^{\tilde{A}(t)} V_k\right)}{t/\mu^2}. \tag{10}$$

We assume that the IDW I_w is well defined with $c_x^2 \triangleq I_w(\infty) = \lim_{t\to\infty} I_w(t) < \infty$.

Note that $\Lambda_t(s)$ depends on the state process $Z(\cdot)$. For the RQ approximation, we replace the stochastic process $Z(\cdot)$ with its deterministic RQ counterpart $Z_{\mathrm{RQ}_1}(\cdot)$ and define

$$\Lambda_t^{\mathrm{RQ}_1}(s) \triangleq \lambda \int_{t-s}^{t} \bar{F}_\alpha(Z_{\mathrm{RQ}_1}(u))du. \tag{11}$$

Combining the drift approximation (7) with the variance surrogate (9) yields the RQ surrogate for the increment $N(t) - N(t-s)$:

$$N(t) - N(t-s) \approx \frac{\Lambda_t^{\mathrm{RQ}_1}(s)}{\mu} - s + b\sqrt{\frac{\Lambda_t^{\mathrm{RQ}_1}(s)}{\mu^2} I_w(\Lambda_t^{\mathrm{RQ}_1}(s))}.$$

Substituting this into the RQ supremum representation (5) gives the transient RQ approximation:

$$Z_{\mathrm{RQ}_1}(t) = \sup_{0\le s\le t} \left\{ \frac{\Lambda_t^{\mathrm{RQ}_1}(s)}{\mu} - s + b\sqrt{\frac{\Lambda_t^{\mathrm{RQ}_1}(s)}{\mu^2} I_w(\Lambda_t^{\mathrm{RQ}_1}(s))} \right\}. \tag{12}$$

We next consider the steady-state RQ approximation obtained by letting $t \to \infty$. Assume $\lim_{t\to\infty} Z_{\mathrm{RQ}_1}(t) = Z_{\mathrm{RQ}_1}$ exists (and is deterministic). Then, for each fixed $s \ge 0$,

$$\Lambda^{\mathrm{RQ}_1}(s) \triangleq \lim_{t\to\infty} \Lambda_t^{\mathrm{RQ}_1}(s) = \lambda \bar{F}_\alpha(Z_{\mathrm{RQ}_1}) s. \tag{13}$$

Substituting (13) into (12) and writing $\rho = \lambda/\mu$ yields the steady-state fixed-point equation

$$\begin{aligned} Z_{\mathrm{RQ}_1} &= \sup_{s\geq 0}\left\{\rho\bar{F}_\alpha(Z_{\mathrm{RQ}_1})s - s + b\sqrt{\frac{\lambda\bar{F}_\alpha(Z_{\mathrm{RQ}_1})s}{\mu^2}I_w(\lambda\bar{F}_\alpha(Z_{\mathrm{RQ}_1})s)}\right\} \\ &= \sup_{u\geq 0}\left\{\rho u - \frac{u}{\bar{F}_\alpha(Z_{\mathrm{RQ}_1})} + b\sqrt{\frac{\rho u}{\mu}I_w(\lambda u)}\right\}, \end{aligned} \tag{14}$$

where we apply the change of variables $u = \bar{F}_\alpha(Z_{\mathrm{RQ}_1})s$ so that $\lambda\bar{F}_\alpha(Z_{\mathrm{RQ}_1})s = \lambda u$.

Define the mapping

$$\Psi(z) \triangleq \sup_{u\geq 0}\left\{\rho u - \frac{u}{\bar{F}_\alpha(z)} + b\sqrt{\frac{\rho u}{\mu}I_w(\lambda u)}\right\}.$$

Since $z \mapsto \bar{F}_\alpha(z)$ is nonincreasing, $z \mapsto \Psi(z)$ is nonincreasing, and hence $z \mapsto \Psi(z) - z$ is strictly decreasing. Therefore, (14) admits at most one solution. In practice, the solution $Z_{\mathrm{RQ}_1} = Z_{\mathrm{RQ}_1}(b;\lambda,\mu,F_\alpha,I_w)$ can be computed efficiently by bisection. We discuss calibration of b in Section 4.3.

4.2 Heavy-Traffic Limits

The choice of the robustness parameter b is central to the accuracy of the RQ approximation. To motivate our calibration of b, we establish heavy-traffic limits for the steady-state RQ fixed point in (14) and compare these limits with the corresponding heavy-traffic asymptotics for the canonical $M/M/1{+}GI$ model. A key theme is how the scaling of the RQ solution (and, by comparison, the mean offered waiting time) depends on the local behavior of the patience-time distribution near the origin. Recall from Assumption 1 that the patience-time scaling is $F_\alpha(t) = F(\alpha t)$ and $\bar{F}_\alpha(t) = 1 - F_\alpha(t) = \bar{F}(\alpha t)$, where F is the CDF of $\tilde{D}$ with mean 1.

To formalize the relevant local behavior of F at 0, we impose the following regularity assumption.

Assumption 2. *There exists an integer $k = k(F) \geq 1$ such that F is k times continuously differentiable on $[0,\infty)$ and*

$$F^{(j)}(0) = 0 \text{ for } j = 0,1,\ldots,k-1, \qquad F^{(k)}(0) \neq 0,$$

where $F^{(j)}$ denotes the jth derivative. Equivalently,

$$F(x) = \frac{F^{(k)}(0)}{k!}x^k + o(x^k), \qquad x \downarrow 0.$$

We consider the coupled long-patience–heavy-traffic limit indexed by the the abandonment rate $\alpha \downarrow 0$. Specifically, we fix the service rate μ and let the arrival rate $\lambda = \rho(\alpha)\mu$ for some traffic load $\rho = \rho(\alpha)$ satisfying $\alpha^{-\gamma}(\rho(\alpha) - 1) \to c$ for some constants $\gamma > 0$ and $c \in \mathbb{R}$, as $\alpha \downarrow 0$. Thus, the system is in heavy-traffic since $\rho(\alpha) \to 1$ as $\alpha \downarrow 0$. Define the threshold

$$h = h(F) \triangleq \frac{k}{k+1}. \tag{15}$$

In this regime, the RQ solution exhibits three distinct scalings, which depends on the following threshold h determined by the threshold h:

Underloaded: if $c < 0$ and $\gamma < h$, abandonment becomes asymptotically negligible and Z_{RQ} scales as $(1-\rho)^{-1}$, consistent with the expected steady-state workload of a $GI/GI/1$ queue without abandonment.

Critically loaded: if $\gamma \geq h$, then abandonment enters the solution and Z_{RQ} scales as α^{-h}. This corresponds to the case where refined diffusion models are required, e.g. reflected Ornstein–Uhlenbeck (ROU) process [23] when $F'(0) \neq 0$ and hazard rate scaling [19] when $F'(0) = 0$.

Overloaded (approaching from above): if $c > 0$ and $\gamma < h$, then Z_{RQ} grows faster, on the scale $\alpha^{-(1-\gamma/k)}$. The case with $\gamma = 0$ corresponds to the overloaded queue studied in [11].

These scalings are summarized below. Proofs are given in Section C.4.

Theorem 1 (Heavy-traffic limit for RQ)**.** *Consider the $GI/GI/1$+GI model under Assumption 1 and Assumption 2. Fix $\mu > 0$ and let $\lambda = \rho(\alpha)\mu$ with $\alpha^{-\gamma}(\rho(\alpha) - 1) \to c$ for some $\gamma > 0$ and $c \in \mathbb{R}$. Let $c_x^2 \triangleq I_w(\infty) < \infty$, and let $Z_{\mathrm{RQ}_1,b}$ denote the (deterministic) steady-state RQ solution of* (14)*.*

1. ***(Underloaded)*** *If $c < 0$ and $\gamma < h$, then*

$$\lim_{\alpha\downarrow 0}(1-\rho(\alpha))Z_{\mathrm{RQ}_1,b} = \frac{1}{\mu}\cdot\frac{b^2}{2}\cdot\frac{c_x^2}{2}, \qquad \textit{equivalently} \qquad \lim_{\alpha\downarrow 0}(-c)\mu\alpha^{\gamma}Z_{\mathrm{RQ}_1,b} = \frac{b^2}{2}\cdot\frac{c_x^2}{2}.$$

   Moreover,

$$\lim_{\alpha\downarrow 0}\frac{Z_{\mathrm{RQ}_1,b}}{\mathbb{E}[Z_{M/M/1}]} = \frac{b^2}{2}\cdot\frac{c_x^2}{2}, \qquad \mathbb{E}[Z_{M/M/1}] \triangleq \frac{\rho(\alpha)}{\mu(1-\rho(\alpha))},$$

   where $\mathbb{E}[Z_{M/M/1}]$ is the mean steady-state workload of an $M/M/1$ queue with arrival rate λ and service rate μ (without abandonment).

2. ***(Critically loaded)*** *If $\gamma \geq h$, then there exists a finite constant $\hat{Z}_{\mathrm{RQ}_1,b} > 0$ such that*

$$\lim_{\alpha\downarrow 0}\alpha^{h}Z_{\mathrm{RQ}_1,b} = \hat{Z}_{\mathrm{RQ}_1,b}.$$

   Furthermore, $\hat{Z}_{\mathrm{RQ}_1,b}$ is the unique positive root of

$$-\mathbb{1}\{\gamma = h\}c\hat{Z}_{\mathrm{RQ}_1,b} + \frac{F^{(k)}(0)}{k!}\hat{Z}^{k+1}_{\mathrm{RQ}_1,b} = \frac{c_x^2 b^2}{4\mu}. \tag{16}$$

3. ***(Overloaded)*** *If $c > 0$ and $\gamma < h$, then*

$$\lim_{\alpha\downarrow 0}\alpha^{1-\gamma/k}Z_{\mathrm{RQ}_1,b} = \left(\frac{ck!}{F^{(k)}(0)}\right)^{1/k}.$$

   In particular, the leading-order limit is independent of b and of the work-variability parameter c_x^2, and depends on the patience-time distribution only through $F^{(k)}(0)$. Equivalently,

$$\lim_{\alpha\downarrow 0}\frac{\rho(\alpha)}{\rho(\alpha)-1}F\big(\alpha Z_{\mathrm{RQ}_1,b}\big) = 1.$$

Part (1) of Theorem 1 identifies the parameter range in which patience times are asymptotically long relative to the load gap, so that abandonment becomes negligible and the system behaves as a single-server queue without abandonment. The following corollary is a direct consequence of Whitt and You [26, Corollary 3 and Theorem 5].

Corollary 1 (Calibration of RQ for the underloaded regime)**.** *Consider the $GI/GI/1{+}GI$ model under Assumption 1 and Assumption 2. Suppose $c < 0$ and $\gamma < h$.*

1. *For Poisson arrivals ($M/GI/1{+}GI$), the RQ algorithm in* (14) *with $b = \sqrt{2}$ yields an asymptotically correct approximation of the steady-state mean virtual waiting time as $\alpha \downarrow 0$.*
2. *If, in addition, $\rho(\alpha) \downarrow 0$ (light traffic) or $\rho(\alpha) \uparrow 1$ (heavy traffic), then the same conclusion holds for general renewal arrivals ($GI/GI/1{+}GI$).*

Part (2) of Theorem 1 characterizes the regime in which the patience-time distribution influences the heavy-traffic scaling. In the canonical Markovian case $M/M/1{+}M$, one has $k = 1$ (so $h = 1/2$) and $F'(0) \neq 0$. Diffusion limits for queues with abandonment were established for the Markovian model in Ward and Glynn [22] and generalized to $GI/GI/1{+}GI$ in Ward and Glynn [23]. We restate the relevant result below.

Let $Z^\alpha(\cdot)$ denote the (steady-state) virtual waiting time process under abandonment scaling parameter α, and define the diffusion-scaled process

$$\tilde{Z}^\alpha(t) \triangleq \alpha^{1/2} Z^\alpha(\alpha^{-1} t).$$

Proposition 1 (Theorem 1, Ward and Glynn 23)**.** *Suppose $\alpha^{-1/2}(\rho(\alpha) - 1) \to c$ for some finite constant c, and assume $\tilde{Z}^\alpha(0) \Rightarrow \tilde{Z}(0)$ as $\alpha \downarrow 0$. Then $\tilde{Z}^\alpha \Rightarrow \tilde{Z}$ as $\alpha \downarrow 0$, where $\tilde{Z}$ is a reflected Ornstein–Uhlenbeck (ROU) process with drift $c - F'(0)z$ and infinitesimal variance c_x^2/μ with $c_x^2 = c_a^2 + c_s^2$. If $F'(0) > 0$, the ROU process has a unique stationary distribution, which is the law of a normal random variable truncated to $[0, \infty)$*

$$\tilde{Z}(\infty) \overset{d}{=} N\left(\frac{c}{F'(0)}, \frac{c_x^2}{2\mu F'(0)}\right) \;\Big|\; \left\{N\left(\frac{c}{F'(0)}, \frac{c_x^2}{2\mu F'(0)}\right) \geq 0\right\},$$

where $N(\mu, \sigma^2)$ is a normal random variable.

When $F'(0) > 0$, the stationary distribution above is a truncated normal with mean

$$\mathbb{E}[\tilde{Z}(\infty)] = \frac{c}{F'(0)} + \frac{\phi\left(-\frac{c}{F'(0)\sigma}\right)}{1 - \Phi\left(-\frac{c}{F'(0)\sigma}\right)}\sigma, \qquad \sigma^2 \triangleq \frac{c_x^2}{2\mu F'(0)}, \tag{17}$$

where ϕ and Φ are the standard normal density and distribution functions, respectively. If $F'(0) = 0$, the diffusion approximation in Proposition 1 degenerates to a reflected Brownian motion, where the patience-time distribution vanishes from the limit. This, however, fails to reveal the subtle scaling and heavy-traffic limit when the system load is heavier than that in the canonical $\alpha^{1/2}$ scaling; see Theorem 2 below.

Exact formulas for $\mathbb{E}[Z_\alpha]$ in the $M/M/1{+}GI$ model are available in Zeltyn and Mandelbaum [30]. The next theorem shows that the exact mean offered waiting time exhibits the same three scaling regimes as the RQ solution.

Theorem 2 (Heavy-traffic limit for $M/M/1{+}GI$)**.** *Let $\mathbb{E}[Z_\alpha]$ be the mean steady-state virtual waiting time in the $M/M/1{+}GI$ model with service rate μ and arrival rate $\lambda = \rho(\alpha)\mu$ such that $\alpha^{-\gamma}(\rho(\alpha) - 1) \to c$ for some $\gamma > 0$ and $c \in \mathbb{R}$. Let h be defined by* (15)*.*

1. ***(Underloaded)*** *If $c < 0$ and $\gamma < h$, then*

$$\lim_{\alpha \downarrow 0} (-c)\mu\alpha^{\gamma}\mathbb{E}[Z_\alpha] = \lim_{\alpha \downarrow 0} \frac{\mathbb{E}[Z_\alpha]}{\mathbb{E}[Z_{M/M/1}]} = 1,$$

   where $\mathbb{E}[Z_{M/M/1}] = \rho(\alpha)/(\mu(1-\rho(\alpha)))$ is the mean workload of an $M/M/1$ queue with arrival rate λ and service rate μ.

2. ***(Critically loaded)*** *If $\gamma \geq h$, then*

$$\lim_{\alpha \downarrow 0} \alpha^{h}\mathbb{E}[Z_\alpha] = \frac{\int_0^\infty x \exp\left\{c\mu x \mathbb{1}\{\gamma = h\} - \frac{\mu F^{(k)}(0)}{(k+1)!}x^{k+1}\right\} dx}{\int_0^\infty \exp\left\{c\mu x \mathbb{1}\{\gamma = h\} - \frac{\mu F^{(k)}(0)}{(k+1)!}x^{k+1}\right\} dx} \triangleq \psi. \tag{18}$$

   In particular, ψ depends on F only through the first nonzero derivative $F^{(k)}(0)$.

3. ***(Overloaded)*** *If $c > 0$ and $\gamma < h$, then*

$$\lim_{\alpha \downarrow 0} \alpha^{1-\gamma/k}\mathbb{E}[Z_\alpha] = \left(\frac{ck!}{F^{(k)}(0)}\right)^{1/k}.$$

4.3 Calibration of the parameter b

We now discuss calibration of the robustness parameter b. We follow a procedure similar to Whitt and You [28]: we select b by matching the heavy-traffic limits of the RQ approximation in Theorem 1 with the corresponding heavy-traffic limits of the original system in Theorem 2, specialized to the $M/M/1{+}GI$ model. In view of Corollary 1, the underloaded regime is already correctly captured by the RQ solution with $b = \sqrt{2}$. Moreover, Part (3) of Theorem 1 shows that the steady-state RQ solution matches the exact leading-order constant in the overloaded heavy-traffic regime, irrespective of the choice of b. Consequently, we work in the critically-loaded scaling $\gamma = h$ with $c \triangleq \alpha^{-h}(\rho - 1)$.

Throughout this subsection we set $\mu = 1$. For an $M/M/1$ input, the long-run index of dispersion for work satisfies $c_x^2 = I_w(\infty) = 2$. Matching the limiting constants in (16) and (18) by setting $\hat{Z}_{\mathrm{RQ}_1,b} = \psi$ therefore yields

$$b(c) \triangleq \sqrt{2\left|-c\psi + \frac{F^{(k)}(0)}{k!}\psi^{k+1}\right|}, \tag{19}$$

where $\psi = \psi(c; k, F^{(k)}(0))$ is the constant defined in (18) with $\gamma = h$ and $\mu = 1$.

The next lemma shows that this calibration automatically recovers the classical underloaded calibration $b = \sqrt{2}$ as a limiting case.

Lemma 4. *For any $k \geq 1$ and $F^{(k)}(0) > 0$, we have $\lim_{c\to-\infty} b(c) = \sqrt{2}$.*

Proof. As $c \to -\infty$, the integrals defining ψ in (18) are dominated by a neighborhood of 0, and one obtains $\psi \sim -1/c$, so that $-c\psi \to 1$ and $\psi^{k+1} = o(1)$. Substituting into (19) yields $b(c) \to \sqrt{2}$. □

Remark 2 (Universal calibration across all heavy-traffic regimes)**.** *Recall that $c \to -\infty$ corresponds to the underloaded long-patience regime (part (1) of Theorem 1 and Theorem 2), because if $\gamma < h$ then $\alpha^{-h}(\rho - 1) \to -\infty$ as $\alpha \downarrow 0$. Lemma 4 therefore implies that the critically-loaded calibration* (19) *subsumes the underloaded case as a special limit. Moreover, in the overloaded regime the leading-order scaling is asymptotically insensitive to b (Theorem 1(3)). In summary,* (19) *provides a single calibration rule that is consistent across all three heavy-traffic regimes.*

Remark 3 (Closed form for the $M/M/1{+}M$ model)**.** *For the canonical $M/M/1{+}M$ model with $\mu = 1$, one has $k = 1$, $h = 1/2$, and $F'(0) = 1$. Writing $c = \alpha^{-1/2}(\rho - 1)$, the constant ψ in* (18) *equals the mean of a truncated normal distribution (see Proposition 1), namely*

$$\psi = c + \frac{\phi(-c)}{1 - \Phi(-c)}.$$

Substituting this ψ into (19) *yields the explicit calibration*

$$b(c) = \sqrt{2\left(c + \frac{\phi(-c)}{1 - \Phi(-c)}\right)\frac{\phi(-c)}{1 - \Phi(-c)}}.$$

4.4 Approximations for Other Performance Measures

Additional steady-state performance measures can be approximated by combining the RQ approximation for the mean virtual waiting time with standard identities for $GI/GI/1{+}GI$ queues. Let Z_{RQ_1} denote the steady-state RQ approximation of the mean virtual waiting time, i.e., the solution of (14) with $b = b(c)$.

Abandonment probability. Let π denote the steady-state probability that an arriving customer abandons. In steady state, $\pi = \mathbb{E}\left[F_\alpha(W)\right]$ and $W = Z(T_i-)$, where W is the offered waiting time seen by an arrival. As a first-order mean-field approximation, we replace W by its RQ mean and set

$$\pi \approx F_\alpha(Z_{\mathrm{RQ}_1}). \tag{20}$$

Mean waiting time of served customers. For the $GI/GI/1$ model without abandonment, it is well-known that the mean steady-state workload (virtual waiting time) and the mean steady-state waiting time is connected by Pollaczek-Khintchine formula

$$\mathbb{E}[Z_{GI/GI/1}] = \rho\left(\mathbb{E}[W_{GI/GI/1}] + \frac{c_s^2 + 1}{2\mu}\right).$$

For the $GI/GI/1+GI$ model, [1] derives the corresponding extension for the waiting time of a customer *conditional on being served* in the $GI/GI/1{+}GI$ model:

$$\mathbb{E}[Z] = (1-\pi)\rho\left(\mathbb{E}[W] + \frac{c_s^2+1}{2\mu}\right), \tag{21}$$

where π is the steady-state probability of abandonment. Combining (20) and (21), we approximate

$$\mathbb{E}[W] \approx \max\left\{0, \frac{Z_{\mathrm{RQ}_1}}{\rho(1-F_\alpha(Z_{\mathrm{RQ}_1}))} - \frac{c_s^2+1}{2\mu}\right\}. \tag{22}$$

Effective queue length. Let Q_0 denote the steady-state number of customers *waiting* who will eventually enter service (i.e., the queue length associated with the *effective* arrival stream). Little's law applied to the effective stream gives $\mathbb{E}[Q_0] = \lambda(1-\pi)\mathbb{E}[W]$. Using (20) and (22) yields

$$\mathbb{E}[Q_0] \approx \max\left\{0, \mu Z_{\mathrm{RQ}_1} - \rho(1-F_\alpha(Z_{\mathrm{RQ}_1}))\frac{c_s^2+1}{2}\right\}.$$

5 A Refined Robust Queueing Algorithm

The first RQ algorithm in Section 4 uses the variance surrogate (9), which is motivated by the heuristic that the dependence between the thinning rule $\mathbb{1}\{D_i > W_i\}$ and the arrival process is negligible. In this section, we develop a refined approximation for the variance function of the effective net-input process based on a heavy-traffic limit. The resulting limit suggests a more accurate structure for the variability term in the RQ formulation, and it serves as the basis for a refined RQ algorithm.

5.1 Heavy-Traffic Limit for the Effective Net-Input Process

Diffusion limits for queues with abandonment were established for the Markovian model in Ward and Glynn [22] and extended to the $GI/GI/1{+}GI$ setting in Ward and Glynn [23]. Under the canonical diffusion scaling, the limiting process is a reflected Ornstein–Uhlenbeck (ROU) diffusion whose drift depends on the patience-time distribution only through the value of its density at the origin, $f(0)$. When $f(0) = 0$, this diffusion limit degenerates to a reflected Brownian motion and the patience-time distribution disappears from the limit.

Here we consider an alternative heavy-traffic regime under which the limiting diffusion has a *nonlinear* drift that depends on the patience-time distribution through its first nonzero derivative at 0 (cf. Assumption 2). This regime is distinct from the hazard-rate scaling proposed in Reed and Ward [19], where the diffusion limit retains information from the entire patience-time distribution.

System sequence and scaling. Consider a sequence of $GI/GI/1{+}GI$ queues indexed by the patience scaling parameter $\alpha \downarrow 0$ (equivalently, mean patience grows as α^{-1}). In the αth system, the patience-time CDF is $F_\alpha(x) = F(\alpha x)$, where F is a base CDF with $F(0) = 0$ and finite mean,

satisfying Assumption 2. Let λ^α and μ^α denote the arrival and service rates, and write $\rho^\alpha \triangleq \lambda^\alpha/\mu^\alpha$. Let $h = k/(k+1)$ be defined as in (15). We assume

$$\mu^\alpha \to \mu, \qquad \alpha^{-h}(\rho^\alpha - 1) \to c, \qquad \text{as } \alpha \downarrow 0.$$

This is precisely the critically-loaded regime with $\gamma = h$ in Theorem 1.

Let $A^\alpha(\cdot)$ be the arrival counting process, let $S^\alpha(t)$ denote the renewal counting process associated with the i.i.d. service-time sequence $\{V_i^\alpha\}_{i\geq 1}$, and let $Z^\alpha(\cdot)$ be the virtual waiting time (offered waiting time) process. Define the effective arrival and effective work-input processes as

$$A_0^\alpha(t) \triangleq \sum_{i=1}^{A^\alpha(t)} \mathbb{1}\{D_i^\alpha > W_i^\alpha\}, \qquad Y^\alpha(t) \triangleq \sum_{i=1}^{A^\alpha(t)} V_i^\alpha \mathbb{1}\{D_i^\alpha > W_i^\alpha\},$$

where $W_i^\alpha = Z^\alpha(T_i^\alpha -)$ is the offered waiting time seen by customer i and T_i^α is the ith arrival epoch. We use the time scaling $t \mapsto \alpha^{-2h}t$ and space scaling $x \mapsto \alpha^h x$. Define the fluid-scaled arrival process

$$\bar{A}^\alpha(t) \triangleq \alpha^{2h} A^\alpha(\alpha^{-2h}t),$$

and the diffusion-scaled processes

$$\begin{aligned}
\tilde{A}^\alpha(t) &\triangleq \alpha^h \left[A^\alpha(\alpha^{-2h}t) - \alpha^{-2h}\lambda^\alpha t\right], \\
\tilde{S}^\alpha(t) &= \alpha^h \left[S^\alpha(\alpha^{-2h}t) - \alpha^{-2h}\mu^\alpha t\right], \\
\tilde{A}_0^\alpha(t) &\triangleq \alpha^h \left[A_0^\alpha(\alpha^{-2h}t) - \alpha^{-2h}\lambda^\alpha t\right], \\
\tilde{Y}^\alpha(t) &\triangleq \alpha^h \left[Y^\alpha(\alpha^{-2h}t) - \alpha^{-2h}\rho^\alpha t\right], && \text{(23a)} \\
\tilde{Z}^\alpha(t) &\triangleq \alpha^h Z^\alpha(\alpha^{-2h}t), && \text{(23b)} \\
\tilde{L}^\alpha(t) &\triangleq \alpha^h L^\alpha(\alpha^{-2h}t), && \text{(23c)}
\end{aligned}$$

where $L^\alpha(\cdot)$ is the cumulative idle time in the identity

$$Z^\alpha(t) = Z^\alpha(0) + Y^\alpha(t) - t + L^\alpha(t), \qquad t \geq 0.$$

Heavy-traffic limit. Let B_a and B_s be independent standard Brownian motions and write $e(t) = t$ for the identity map. Recall that c_a^2 and c_s^2 are the asymptotic variability parameters of the arrival and service primitives, and $c_x^2 = c_a^2 + c_s^2$.

Theorem 3. *Assume the functional CLT*

$$(\tilde{A}^\alpha, \tilde{S}^\alpha, \tilde{Z}^\alpha(0)) \Rightarrow (c_a B_a \circ (\mu e), c_s B_s \circ (\mu e), Z^*(0)), \qquad \alpha \downarrow 0. \tag{24}$$

Then the joint heavy-traffic limit of the diffusion-scaled processes is

$$(\tilde{Z}^\alpha, \tilde{L}^\alpha, \tilde{Y}^\alpha, \tilde{A}_0^\alpha) \Rightarrow (Z^*, L^*, Y^*, A_0^*), \qquad \alpha \downarrow 0,$$

where (Z^*, L^*) is the unique solution to the reflected integral equation

$$Z^*(t) = Z^*(0) + \mu^{-1}c_a B_a(\mu t) + \mu^{-1}c_s B_s(\mu t) - \frac{F^{(k)}(0)}{k!}\int_0^t \big(Z^*(s)\big)^k ds + ct + L^*(t), \tag{25}$$

with $Z^*(t) \geq 0$, L^* nondecreasing, $L^*(0) = 0$, and $\int_0^\infty \mathbb{1}\{Z^*(t) > 0\}dL^*(t) = 0$. Moreover,

$$A_0^*(t) = c_a B_a(\mu t) - \mu\frac{F^{(k)}(0)}{k!}\int_0^t \big(Z^*(s)\big)^k ds,$$
$$Y^*(t) = \frac{1}{\mu}A_0^*(t) + \frac{1}{\mu}c_s B_s(\mu t) = \mu^{-1}c_a B_a(\mu t) + \mu^{-1}c_s B_s(\mu t) - \frac{F^{(k)}(0)}{k!}\int_0^t \big(Z^*(s)\big)^k ds.$$

For Robust Queueing, we will ultimately use a stationary approximation for increments of the effective net-input process (see, e.g., Whitt and You [26, Section 5.2]). Accordingly, we assume henceforth that $Z^*(0)$ is distributed according to the unique stationary distribution of Z^*.

Remark 4. *When $k = 1$, Theorem 3 reduces to the ROU limit in Proposition 1. For general $k \geq 1$, the limiting diffusion* (25) *has polynomial drift. Its stationary density is given by*

$$\pi_k(x) = \frac{1}{G_k}\exp\left\{\frac{2\mu}{c_x^2}\left(cx - \frac{F^{(k)}(0)}{(k+1)!}x^{k+1}\right)\right\}\mathbb{1}\{x \geq 0\}, \tag{26}$$

where G_k is the normalizing constant; see, e.g., Browne and Whitt [5, Section 3].

5.2 The Variance Function of the Stationary Heavy-Traffic Limit

To set the stage for our heavy-traffic approximation of the variance function, we first present a detailed characterization of the variance function of the stationary heavy-traffic limit.

Recall that the heavy-traffic limit in Theorem 3 is characterized by the parameter tuple

$$\Xi \triangleq (c, k, \mu, c_a^2, c_s^2, F^{(k)}(0)).$$

Let (Z^*, L^*) be the stationary reflected diffusion in Theorem 3, and note that the limit total-input process is given by

$$Y^*(t) = Z^*(t) - Z^*(0) - ct - L^*(t).$$

We define the variance function of the stationary heavy-traffic limit as

$$v(t; \Xi) \triangleq \mathrm{Var}\big(Y^*(t)\big) = \mathrm{Var}\big(Z^*(t) - Z^*(0) - L^*(t)\big),$$

where $Z^*(0)$ follows the stationary distribution in (26).

5.2.1 A Scaling Representation and the Variance-Reduction Function

Define the normalized polynomial drift coefficient

$$\beta \triangleq \frac{F^{(k)}(0)}{k!} > 0.$$

To obtain a convenient scaling representation, we introduce the *base* parameter-tuple

$$\Xi_0 \triangleq (c, k, 1, 1, 1, k!),$$

which corresponds to the normalization $\mu = 1$, $c_a^2 = c_s^2 = 1$ (hence $c_x^2 = 2$), and $\beta = 1$. Define the associated *base variance function*

$$v_{c,k}(t) \triangleq v(t; \Xi_0), \qquad t \geq 0. \tag{27}$$

For the corresponding $M/M/1$ benchmark *without* abandonment, the heavy-traffic limit of the total-input process is a Brownian motion with variance $2t$. We therefore define the *variance-reduction function* relative to this Brownian benchmark by

$$w_{c,k}(t) \triangleq \frac{v_{c,k}(t)}{2t}, \qquad t > 0. \tag{28}$$

The part (1) of following proposition show that the variance-reduction function is bounded from above by 1. Thus, $w_{c,k}(t)$ measures the *reduction* of variance induced by abandonment in the heavy-traffic diffusion limit, relative to the $M/M/1$ benchmark of $2t$.

Proposition 2. *The variance-reduction function $w_{c,k}(\cdot)$ satisfies*

1. $0 \leq w_{c,k}(t) \leq 1$ *for all* $t \geq 0$.
2. $\lim_{t \downarrow 0} w_{c,k}(t) = 1$, *so we set* $w_{c,k}(0) \triangleq 1$ *without loss of generality.*
3. $w_{c,k}(t)$ *is strictly decreasing in* t *for all* $t \geq 0$.
4. *The mapping* $c \mapsto w_{c,k}(\infty)$ *is strictly decreasing on* $\mathbb{R}$. *Moreover,*

$$\lim_{c \to -\infty} w_{c,k}(\infty) = 1, \qquad \lim_{c \to \infty} w_{c,k}(\infty) = 0.$$

Proposition 2 formalizes two queueing intuitions. Abandonment provides state-dependent negative feedback: $w_{c,k}(t) \to 1$ as $t \downarrow 0$, while $w_{c,k}(t)$ decreases over longer horizons as the feedback suppresses cumulative variability. The strength of this suppression increases with the scaled load c: heavier loading shifts the stationary workload upward and strengthens mean reversion, so $w_{c,k}(\infty) \downarrow 0$ as $c \to \infty$, whereas in the underloaded limit $c \to -\infty$ the workload stays near the boundary and $w_{c,k}(\infty) \uparrow 1$.

More broadly, $w_{c,k}$ quantifies the joint effect of the scaled load c, the local patience-time index k, and the time horizon t on effective-input variability. This scale dependence matters outside asymptotic heavy traffic, since reverse-time increment approximations are typically dominated by horizons comparable to the mean virtual waiting time.

The following lemma show how the variance function under any parameter-tuple can be conveniently expressed in terms of the variance-reduction function $w_{c,k}$. Let $\beta \triangleq F^{(k)}(0)/k! > 0$.

Lemma 5. *Define*

$$\tau \triangleq \left(\frac{c_x^2}{2\mu}\right)^{\frac{k-1}{k+1}} \beta^{\frac{2}{k+1}}, \qquad \tilde{c} \triangleq c\left(\frac{c_x^2}{2\mu}\right)^{-\frac{k}{k+1}} \beta^{-\frac{1}{k+1}}.$$

Then, for all $t \geq 0$,

$$v(t; c, k, \mu, c_a^2, c_s^2, F^{(k)}(0)) = \frac{c_x^2}{\mu} t w_{\tilde{c},k}(\tau t).$$

5.2.2 Computing the Variance-Reduction Function $w_{c,k}$

By Lemma 5, evaluating the heavy-traffic variance function $v(t;\Xi)$ reduces to computing the base variance-reduction function $w_{c,k}$. In general, $w_{c,k}(t)$ depends on the transient law of a reflected nonlinear diffusion and does not admit a closed-form expression. We therefore develop a tractable representation of $w_{c,k}$ based on Malliavin calculus and associated one-dimensional parabolic PDEs. These PDEs can be solved numerically and yield $w_{c,k}(t)$ for all time horizons t.

Fix $c \in \mathbb{R}$ and an integer $k \geq 1$. By the definition of $w_{c,k}(t)$, we need only consider the base parameter-tuple $(c, k, \mu, c_a^2, c_s^2, F^{(k)}(0)) = (c, k, 1, 1, 1, k!)$, so that $c_x^2 = c_a^2 + c_s^2 = 2$ and the polynomial drift coefficient $\beta = F^{(k)}(0)/k! = 1$. Specifically, in the base model the stationary heavy-traffic diffusion (25) can be written as the reflected SDE

$$Z^{c,k}(t) = Z^{c,k}(0) + \sqrt{2}B(t) + \int_0^t (c - (Z^{c,k}(s))^k) ds + L^{c,k}(t), \qquad t \geq 0, \tag{29}$$

where B is a standard Brownian motion, $Z^{c,k}(t) \geq 0$, $L^{c,k}$ is nondecreasing with $L^{c,k}(0) = 0$, and $\int_0^\infty \mathbb{1}\{Z^{c,k}(t) > 0\} dL^{c,k}(t) = 0$. We assume $Z^{c,k}(0)$ has the unique stationary distribution; then the stationary density of $Z^{c,k}$ is given in (26), which specializes to

$$\pi_{c,k}(x) = \frac{1}{G_{c,k}} \exp\left\{cx - \frac{1}{k+1}x^{k+1}\right\} \mathbb{1}\{x \geq 0\}, \qquad G_{c,k} \triangleq \int_0^\infty \exp\left\{cx - \frac{1}{k+1}x^{k+1}\right\} dx. \tag{30}$$

The associated base effective-input functional is

$$Y^{c,k}(t) \triangleq Z^{c,k}(t) - Z^{c,k}(0) - ct - L^{c,k}(t) = \sqrt{2}B(t) - \int_0^t (Z^{c,k}(s))^k ds, \qquad t \geq 0. \tag{31}$$

By definition, $v_{c,k}(t) = \mathrm{Var}(Y^{c,k}(t))$ and $w_{c,k}(t) = v_{c,k}(t)/(2t)$ for $t > 0$.

For notational convenience, define the nonnegative function $q_k(x) \triangleq kx^{k-1}$ for $x \geq 0$, and, for $z \geq 0$, let $\mathbb{P}_z$ denote the law of the reflected diffusion $Z^{c,k}$ in (29) with initial state $Z^{c,k}(0) = z$. Let $\tau_0 \triangleq \inf\{t \geq 0 : Z^{c,k}(t) = 0\}$ denote the first hitting time of the boundary. Furthermore, for $t \geq 0$ and $z \geq 0$, define

$$\psi_{c,k}(t, z) \triangleq \mathbb{E}_z\left[\exp\left\{-\int_0^{t\wedge\tau_0} q_k(Z^{c,k}(s)) ds\right\}\right]. \tag{32}$$

The function $\psi_{c,k}$ governs the conditional Malliavin derivative of the effective input. Because we initialize the diffusion in stationarity, $Z^{c,k}(0) \sim \pi_{c,k}$ is random, and thus, by the law of total variance, $\mathrm{Var}(Y^{c,k}(t))$ decomposes into a *conditional* variance term (generated by the Brownian noise on $(0, t]$)

and a second term due to randomness in the initial state. To characterize this second term, define, for $t \geq 0$ and $z \geq 0$,

$$h_{c,k}(t,z) \triangleq \mathbb{E}_z \left[\int_0^t (Z^{c,k}(s))^k ds \right]. \tag{33}$$

By (31), $\mathbb{E}_z[Y^{c,k}(t)] = -h_{c,k}(t,z)$.

Lemma 6 (Clark–Ocone representation of $w_{c,k}$)**.** *For each $t > 0$,*

$$w_{c,k}(t) = \frac{1}{t} \int_0^t \mathbb{E}_{\pi_{c,k}} \left[\psi_{c,k}(u,Z)^2 \right] du + \frac{1}{2t} \mathrm{Var}_{\pi_{c,k}} \left(h_{c,k}(t,Z) \right) \tag{34}$$

where $Z \sim \pi_{c,k}$ and $\mathrm{Var}_{\pi_{c,k}}(\cdot)$ denotes variance with respect to the stationary density (30)*.*

The representation (34) reduces the computation of $w_{c,k}(t)$ to evaluating $\psi_{c,k}(u,\cdot)$ for $u \in [0,t]$ and $h_{c,k}(t,\cdot)$. The function $\psi_{c,k}$ can be computed by solving a one-dimensional parabolic PDE, and $h_{c,k}$ can be computed by solving a second one-dimensional parabolic PDE (see Remark 5).

Proposition 3 (PDE characterization of $\psi_{c,k}$ and $h_{c,k}$)**.** *The function $\psi_{c,k}$ defined in* (32) *is the unique bounded classical solution on $[0,\infty) \times [0,\infty)$ to*

$$\begin{cases} \partial_t \psi(t,x) = \partial_{xx}\psi(t,x) + (c - x^k)\partial_x \psi(t,x) - q_k(x)\psi(t,x), & t > 0,\ x > 0, \\ \psi(0,x) = 1, & x \geq 0, \\ \psi(t,0) = 1, & t \geq 0, \\ \sup_{t\in[0,T]} \sup_{x \geq 0} |\psi(t,x)| < \infty, & \forall T < \infty. \end{cases} \tag{35}$$

Moreover, the function $h_{c,k}$ defined in (33) *is the unique classical solution on $[0,\infty) \times [0,\infty)$ to*

$$\begin{cases} \partial_t h(t,x) = \partial_{xx} h(t,x) + (c - x^k)\partial_x h(t,x) + x^k, & t > 0,\ x > 0, \\ h(0,x) = 0, & x \geq 0, \\ \partial_x h(t,0) = 0, & t \geq 0, \\ \sup_{t\in[0,T]} \sup_{x \geq 0} \frac{|h(t,x)|}{1+x^k} < \infty, & \forall T < \infty. \end{cases} \tag{36}$$

Remark 5 (Numerical evaluation of $w_{c,k}$)**.** *Given (c,k) and a time horizon $t > 0$, one can compute $w_{c,k}(t)$ via* (34) *as follows: (i) solve the PDE* (35) *and* (36) *for $\psi_{c,k}(u,x)$ and $h_{c,k}(u,x)$ over $(u,x) \in [0,t] \times [0,\infty)$, (ii) compute $\int_0^\infty \psi_{c,k}(u,x)^2 \pi_{c,k}(x)dx$ for $u \in [0,t]$ and the moments $\int_0^\infty h_{c,k}(t,x)\pi_{c,k}(x)dx$ and $\int_0^\infty h_{c,k}(t,x)^2\pi_{c,k}(x)dx$ using numerical integration, and (iii) combine the results according to* (34)*, using $\mathrm{Var}_{\pi_{c,k}}(h_{c,k}(t,Z)) = \mathbb{E}_{\pi_{c,k}}[h_{c,k}(t,Z)^2] - \mathbb{E}_{\pi_{c,k}}[h_{c,k}(t,Z)]^2$. The stationary density $\pi_{c,k}$ is explicit in* (30)*, and the computational effort is dominated by solving the one-dimensional PDEs* (35) *and* (36)*. As before, such a numerical procedure need only be performed once offline for each pair (c,k), and the resulting function $t \mapsto w_{c,k}(t)$ can be stored and used for approximating the variance function under any parameter-tuple Ξ via Lemma 5.*

Remark 6 (An explicit Poisson-equation formula for $w_{c,k}(\infty)$)**.** *The PDE approach above is tailored to computing the full function $t \mapsto w_{c,k}(t)$. Since the second term in* (34) *is $O(1/t)$, it vanishes as*

$t \to \infty$. *If one only needs the* long-run *variance-reduction constant* $w_{c,k}(\infty) \triangleq \lim_{t\to\infty} w_{c,k}(t)$, *then a more explicit one-dimensional characterization is available via a Poisson equation. Let* $\pi_{c,k}$ *be the stationary density* (30) *and define* $m_k \triangleq \mathbb{E}_{\pi_{c,k}}[Z^k]$. *Let* u *be a* C^2 *solution to the Poisson equation*

$$u''(x) + (c - x^k)u'(x) = x^k - m_k, \qquad x > 0, \qquad u'(0) = 0,$$

for the reflected generator $\mathcal{L}$. *Then one has the asymptotic-variance identity*

$$w_{c,k}(\infty) = \mathbb{E}_{\pi_{c,k}}\left[(1 + u'(Z))^2\right] = \int_0^\infty (1 + u'(x))^2 \pi_{c,k}(x)dx,$$

and the derivative u' *admits the explicit integrating-factor representation*

$$u'(x) = \frac{1}{\pi_{c,k}(x)} \int_0^x \pi_{c,k}(y)(y^k - m_k)dy.$$

This eliminates the need to solve the parabolic PDE and reduces the evaluation of $w_{c,k}(\infty)$ *to one-dimensional numerical integration.*

5.3 Heavy-Traffic Approximation of the Variance Function

In this section, we propose a heavy-traffic approximation for the time-stationary variance function of the effective net-input process $N(t)$, i.e., we assume the pre-limit system is in equilibrium at time 0. Since $N(t) = Y(t) - t$ and t is deterministic, $\mathrm{Var}(N(t)) = \mathrm{Var}(Y(t))$; hence it suffices to study the variance of the effective total-input process $Y(t)$. Throughout this subsection, $\mathbb{E}_e[\cdot]$ and $\mathrm{Var}_e(\cdot)$ denote expectation and variance under the equilibrium (stationary) distribution of the αth system.

Recall the diffusion-scaled effective total-input process in (23a):

$$\tilde{Y}^\alpha(t) \triangleq \alpha^h\Big(Y^\alpha(\alpha^{-2h}t) - \alpha^{-2h}\rho^\alpha t\Big), \qquad t \geq 0,$$

where $\rho^\alpha = \lambda^\alpha/\mu^\alpha$ and $h = k/(k+1)$. Define the heavy-traffic scaled variance function

$$\tilde{V}^\alpha(t) \triangleq \mathrm{Var}_e(\tilde{Y}^\alpha(t)) = \mathrm{Var}_e(\alpha^h Y^\alpha(\alpha^{-2h}t)) = \alpha^{2h}\mathrm{Var}_e(Y^\alpha(\alpha^{-2h}t)),$$

where the centering term $\alpha^h \cdot \alpha^{-2h}\rho^\alpha t$ is deterministic and dropped.

Assuming the usual interchange-of-limits conditions hold, Theorem 3 yields the following corollary, which characterizes the pre-limit variance function on the heavy-traffic time scale.

Corollary 2 (Heavy-traffic limit of the variance function)**.** *Assume the conditions of Theorem 3. Assume further that the* α*th system is in equilibrium at time* 0 *and that for each fixed* $t \geq 0$ *the family* $\{|\tilde{Y}^\alpha(t)|^2 : \alpha > 0\}$ *is uniformly integrable (e.g.,* $\sup_\alpha \mathbb{E}_e[|\tilde{Y}^\alpha(t)|^{2+\delta}] < \infty$ *for some* $\delta > 0$*). Then for each fixed* $t \geq 0$*,*

$$\tilde{V}^\alpha(t) = \mathrm{Var}_e(\tilde{Y}^\alpha(t)) \longrightarrow v(t;\Xi) = \mathrm{Var}\big(Y^*(t)\big), \qquad \alpha \downarrow 0,$$

where Y^* *is the limiting process in Theorem 3 and* $\Xi = (c, k, \mu, c_a^2, c_s^2, F^{(k)}(0))$ *is the parameter tuple. Moreover, with* $\tilde{c}$ *and* τ *defined in Lemma 5,*

$$v(t;\Xi) = \frac{c_x^2}{\mu} t w_{\tilde{c},k}(\tau t), \qquad t \geq 0.$$

Corollary 2 implies that for physical time horizons of order $t = O(\alpha^{-2h})$,

$$\mathrm{Var}_e(Y^\alpha(t)) \approx \frac{c_x^2}{\mu} t w_{\tilde{c},k}(\alpha^{2h}\tau t), \qquad t = O(\alpha^{-2h}).$$

Equivalently, since in heavy traffic the mean effective work input rate is asymptotically 1, $\mathbb{E}_e[Y^\alpha(t)] \approx t$ on this time scale, we may rewrite

$$\mathrm{Var}_e(Y^\alpha(t)) \approx \frac{c_x^2}{\mu} w_{\tilde{c},k}(\alpha^{2h}\tau t)\mathbb{E}_e[Y^\alpha(t)], \qquad t = O(\alpha^{-2h}). \tag{37}$$

This normalization is chosen to mirror the IDW function in the $GI/GI/1$ model without abandonment. The form of (37) is meticulously chosen to match that of the IDW in the $GI/GI/1$ model without abandonment. Specifically, define the effective IDW by

$$I_w^{\mathrm{ab}}(t) \triangleq \frac{\mathrm{Var}_e(Y^\alpha(t))}{\mathbb{E}[V_1]\mathbb{E}_e[Y^\alpha(t)]} = \mu \frac{\mathrm{Var}_e(Y^\alpha(t))}{\mathbb{E}_e[Y^\alpha(t)]}, \tag{38}$$

where $\mathbb{E}[V_1] = 1/\mu$. Then (37) implies $I_w^{\mathrm{ab}}(t) \approx c_x^2 w_{\tilde{c},k}(\alpha^{2h}\tau t)$ on the $O(\alpha^{-2h})$ time scale. The advantage of working with the IDW is that it is dimensionless, and thus cleanly separates intrinsic variability from the overall scale of the input process; see Whitt and You [26].

Remark 7. *For the $GI/GI/1$ model without abandonment, the (work) IDW satisfies*

$$\lim_{t\to\infty} \frac{\mathrm{Var}_e(\tilde{Y}(t))}{\mathbb{E}[V_1]\mathbb{E}_e[\tilde{Y}(t)]} = c_x^2, \qquad \tilde{Y}(t) = \sum_{i=1}^{A(t)} V_i.$$

In contrast, for the abandonment model, Corollary 2 implies that on the heavy-traffic time scale,

$$\lim_{t\to\infty}\lim_{\alpha\downarrow 0} \frac{\mathrm{Var}_e(Y^\alpha(\alpha^{-2h}t))}{\mathbb{E}[V_1]\mathbb{E}_e[Y^\alpha(\alpha^{-2h}t)]} = c_x^2 w_{\tilde{c},k}(\infty) \le c_x^2,$$

where $w_{\tilde{c},k}(\infty)$ quantifies the long-run variance reduction induced by abandonment; see Remark 6.

The heavy-traffic approximation in Corollary 2 is most informative on the heavy-traffic time scale $t = O(\alpha^{-2h})$. For practical applications, we also consider the long-patience limit $\alpha \downarrow 0$ over a fixed physical time horizon t.

We begin with heuristic intuition. When the traffic intensity $\rho < 1$ is fixed, sufficiently long patience implies that almost all arrivals enter service, so the effective arrival process should be asymptotically equivalent to the original arrival process as $\alpha \to 0$. This corresponds to the underloaded regime in Theorem 1. The behavior is qualitatively different when $\rho > 1$ is fixed. In that case, limited service capacity forces a persistent thinning of the arrival stream, regardless of how long patience is. Nevertheless, when the abandonment rate is small, the state-dependent selection is expected to simplify. Heuristically, only a fraction $1/\rho$ of arrivals can be served in steady state, so the effective arrival process should be well approximated by independent thinning with retention probability $1/\rho$.

The following lemma formalizes the heuristics above. Define the IDC of the (stationary renewal) arrival process over horizon t by

$$I_a(t) \triangleq \frac{\mathrm{Var}(A(t))}{\lambda t} = \frac{\mathrm{Var}(A(t))}{\mathbb{E}[A(t)]}.$$

Lemma 7. *Fix $t \geq 0$ and consider a* stationary *$GI/GI/1{+}GI$ queue with arrival rate λ, service rate μ, and patience-time scaling $F_\alpha(x) = F(\alpha x)$. Let $\rho = \lambda/\mu$. Then, as $\alpha \downarrow 0$,*

$$\mathbb{E}_e[Y^\alpha(t)] \to (\rho \wedge 1)t,$$

and

$$\frac{\mathrm{Var}_e(Y^\alpha(t))}{\mathbb{E}_e[Y^\alpha(t)]} \to \frac{1}{\mu}\left(\frac{I_a(t)}{\rho \vee 1} + \left(1 - \frac{1}{\rho \vee 1}\right) + c_s^2\right).$$

Finally, combining Corollary 2, Lemma 7, and the drift approximation $\mathbb{E}_e[Y^\alpha(t)] \approx \Lambda^*(t)/\mu$ for $\Lambda^*(\cdot)$ defined in (8), we propose the variance approximation

$$V(t) \equiv \hat{I}_w(t) w_{\tilde{c},k}(\alpha^{2h}\tau t)\frac{\Lambda^*(t)}{\mu^2}, \tag{39}$$

where

$$\hat{I}_w(t) \triangleq \frac{I_a(t)}{\rho \vee 1} + \left(1 - \frac{1}{\rho \vee 1}\right) + c_s^2. \tag{40}$$

Equivalently, the effective IDW admits the approximation

$$\hat{I}_w^{\mathrm{ab}}(t) \approx \hat{I}_w(t) w_{\tilde{c},k}(\alpha^{2h}\tau t). \tag{41}$$

Note that $\lim_{\rho\to 1} \hat{I}_w(\infty) = I_a(\infty) + c_s^2 = c_x^2$, so (39) is consistent with the heavy-traffic scaling.

Remark 8. *Our approximation of the effective IDW admits a insightful factorization. The term $\hat{I}_w(t)$ captures the intrinsic variability of the effective arrival process, while $w_{\tilde{c},k}(\alpha^{2h}\tau t)$ captures the additional variance reduction induced by abandonment through state-dependent feedback. Moreover, the quantity*

$$\frac{I_a(t)}{\rho \vee 1} + \left(1 - \frac{1}{\rho \vee 1}\right)$$

appearing in $\hat{I}_w(t)$ can be interpreted as the IDC of the effective arrival process. It is a convex combination of the original IDC $I_a(t)$ and 1 (the IDC of a Poisson process), with weights determined by the traffic intensity ρ. This decomposition highlights the regulating effect of abandonment and its dependence on system load.

Example 1. *To illustrate the accuracy of the approximation, we consider the $H_2(4)/M/1{+}M$ and $H_2(4)/M/1{+}E_2$ model. Here $H_2(4)$ denotes a hyperexponential distribution with balanced mean and squared coefficient of variation 4, and E_2 denotes an Erlang distribution with shape parameter 2. For E_2, the patience-time CDF satisfies Assumption 2 with $k = 2$, and hence $h = 2/3$. We set $\mu = 1$ and $\lambda = \rho = 1 + c\alpha^{2/3}$ with $c = 2$ and $\alpha = 2^{-i}$ for $i \in \{0, 3, 6, 9, 12\}$. Figure 1 compares simulation estimates of the effective IDW defined in* (38) *(solid curves) with the approximation in* (41) *(dashed curves). The variance-reduction function $w_{\tilde{c},2}(\cdot)$ is evaluated using the procedure described in Remark 5. Even though the approximation is derived under the long-patience limit, it performs remarkably well across a range of time horizons and patience levels.*

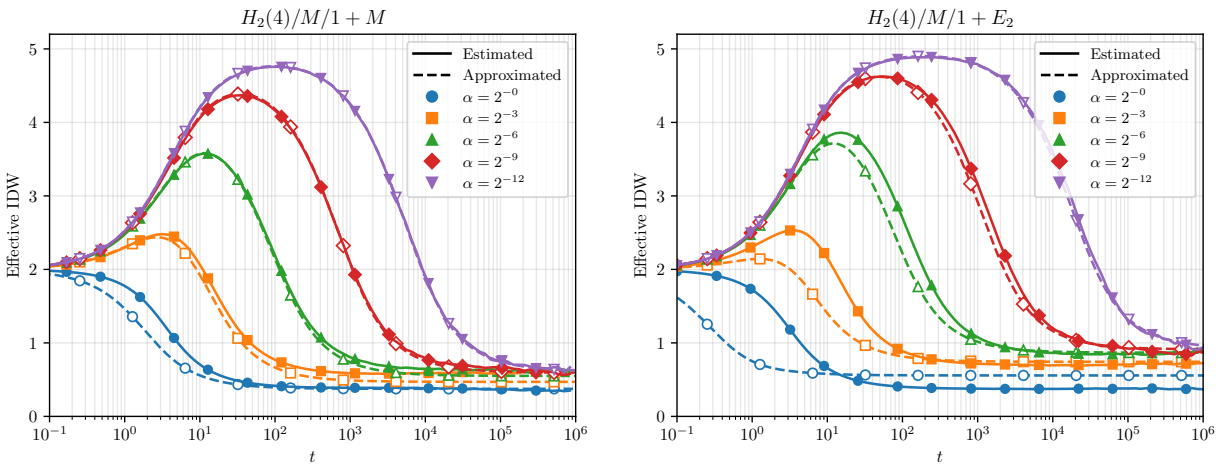

Figure 1: Simulation estimates (solid) and approximations (39) (dashed) of the effective IDW in the $H_2(4)/M/1+M$ model (left) the $H_2(4)/M/1+E_2$ model (right) with $\mu = 1$, $c = 2$, and $\alpha = 2^{-i}$ for $i \in \{0, 3, 6, 9, 12\}$.

5.4 Robust Queueing Algorithm

With the drift approximation $\Lambda_t(s)$ in (6) and the variance approximation $V(t)$ in (39), we are ready to propose a refined robust queueing formulation for the virtual waiting time. The definitions of $\Lambda_t(s)$ and $V(t)$ depend on the state process $Z(\cdot)$, so we replace $Z(\cdot)$ by its deterministic RQ counterpart $Z_{\mathrm{RQ}}(\cdot)$. Analogously to (11), for $0 \le s \le t$ we define the deterministic surrogates $\Lambda_t^{\mathrm{RQ}}(s)$ and $V_t^{\mathrm{RQ}}(s)$ as follows:

$$\Lambda_t^{\mathrm{RQ}}(s) \triangleq \lambda \int_{t-s}^{t} \bar{F}_\alpha(Z_{\mathrm{RQ}}(u))du, \qquad V_t^{\mathrm{RQ}}(s) \triangleq \hat{I}_w(s) w_{\tilde{c},k}(\alpha^{2h}\tau s) \frac{\Lambda_t^{\mathrm{RQ}}(s)}{\mu^2}.$$

The resulting RQ surrogate for the effective net-input increment is

$$N(t) - N(t-s) \approx \frac{\Lambda_t^{\mathrm{RQ}}(s)}{\mu} - s + b\sqrt{V_t^{\mathrm{RQ}}(s)}. \tag{42}$$

Substituting (42) into (5) yields the refined transient RQ approximation

$$Z_{\mathrm{RQ}}(t) = \sup_{0\le s\le t} \left\{ \frac{\Lambda_t^{\mathrm{RQ}}(s)}{\mu} - s + b\sqrt{V_t^{\mathrm{RQ}}(s)} \right\}. \tag{43}$$

We next consider the steady-state approximation obtained by letting $t \to \infty$ in (43). Assume the limit $Z_{\mathrm{RQ}} \triangleq \lim_{t\to\infty} Z_{\mathrm{RQ}}(t)$ exists and is deterministic. Then, for each fixed $s \ge 0$, the limiting drift stationary variance surrogates are given by

$$\Lambda^{\mathrm{RQ}}(s) \triangleq \lim_{t\to\infty} \Lambda_t^{\mathrm{RQ}}(s) = \lambda \bar{F}_\alpha(Z_{\mathrm{RQ}})s, \qquad V^{\mathrm{RQ}}(s) \triangleq \lim_{t\to\infty} V_t^{\mathrm{RQ}}(s) = \hat{I}_w(s) w_{\tilde{c},k}(\alpha^{2h}\tau s) \frac{\lambda \bar{F}_\alpha(Z_{\mathrm{RQ}})s}{\mu^2}.$$

Taking $t \to \infty$ in (43) yields the steady-state fixed-point equation

$$Z_{\mathrm{RQ}} = \sup_{s\ge 0} \left\{ \frac{\Lambda^{\mathrm{RQ}}(s)}{\mu} - s + b\sqrt{V^{\mathrm{RQ}}(s)} \right\}. \tag{44}$$

Note that the right-hand side of (44) with Z_{RQ} replaced by z is nonincreasing in z. Consequently, (44) admits a unique solution, which we denote by $Z_{\mathrm{RQ},b}$. In practice, $Z_{\mathrm{RQ},b}$ can be computed efficiently by bisection. We discuss calibration of the robustness parameter b in Section 5.6.

5.5 Heavy-Traffic Limit for Robust Queueing

Recall the threshold h defined in (15). We study the scaling of the refined RQ fixed point in the long-patience heavy-traffic regime $\alpha \downarrow 0$ with $\rho(\alpha) \to 1$. We assume $\rho(\alpha) = 1 + c\alpha^{\gamma}$ for some constants $\gamma > 0$ and $c \in \mathbb{R}$.

Theorem 4 (Heavy-traffic limit for refined RQ)**.** *Consider the $GI/GI/1{+}GI$ model under Assumption 1 and Assumption 2. Fix $\mu > 0$ and let $\lambda = \rho(\alpha)\mu$ with $\rho(\alpha) = 1 + c\alpha^{\gamma}$ for some $\gamma > 0$ and $c \in \mathbb{R}$. Let $c_a^2 \triangleq I_a(\infty)$ and $c_x^2 \triangleq c_a^2 + c_s^2$. For each $\alpha > 0$, let $Z_{\mathrm{RQ},b}^{\alpha}$ denote the unique solution to the refined steady-state RQ equation* (44)*.*

1. ***Underloaded.*** *If $0 < \gamma < h$ and $c < 0$, then*

$$\lim_{\alpha \downarrow 0} (-c)\mu\alpha^{\gamma} Z_{\mathrm{RQ},b}^{\alpha} = \lim_{\alpha \downarrow 0} \frac{Z_{\mathrm{RQ},b}^{\alpha}}{\mathbb{E}[Z_{M/M/1}]} = \frac{b^2}{2}\frac{c_x^2}{2},$$

   where $\mathbb{E}[Z_{M/M/1}] \triangleq \rho(\alpha)\big(\mu(1-\rho(\alpha))\big)^{-1}$ is the mean workload of an $M/M/1$ queue.

2. ***Critically loaded.*** *If $\gamma \geq h$, then there exists a finite constant $\hat{Z}_{\mathrm{RQ},b} > 0$ such that*

$$\lim_{\alpha \downarrow 0} \alpha^{h} Z_{\mathrm{RQ},b}^{\alpha} = \hat{Z}_{\mathrm{RQ},b}.$$

   Moreover, $\hat{Z}_{\mathrm{RQ},b}$ is the unique positive solution to

$$\hat{Z}_{\mathrm{RQ},b} = \sup_{u \geq 0} \left\{ \left(\mathbb{1}\{\gamma = h\}c - \beta \hat{Z}_{\mathrm{RQ},b}^{k} \right) u + b\sqrt{\frac{c_x^2}{\mu} w_{\tilde{c},k}(\tau u) u} \right\}. \tag{45}$$

3. ***Overloaded.*** *If $0 < \gamma < h$ and $c > 0$, then*

$$\lim_{\alpha \downarrow 0} \alpha^{1-\gamma/k} Z_{\mathrm{RQ},b}^{\alpha} = \left(\frac{c}{\beta}\right)^{1/k} = \left(\frac{ck!}{F^{(k)}(0)}\right)^{1/k}.$$

   In particular, the leading-order limit is independent of b and of the arrival and service variability parameters, and it depends on the patience-time distribution only through $F^{(k)}(0)$. Equivalently,

$$\lim_{\alpha \downarrow 0} \frac{\rho(\alpha)}{\rho(\alpha) - 1} F(\alpha Z_{\mathrm{RQ},b}^{\alpha}) = 1.$$

The underloaded and overloaded limits in Theorem 4 coincide with those of the first RQ algorithm in Theorem 1. The key difference lies in the critically-loaded regime. Both algorithms yield the same scaling $Z_{\mathrm{RQ}}^{\alpha} = O(\alpha^{-h})$, but the refined limit depends nontrivially on the variance-reduction function $w_{\tilde{c},k}$.

5.6 Calibration of the Parameter b

In the underloaded regime of Theorem 3, setting $b = \sqrt{2}$ recovers the original RQ algorithm. Moreover, by Corollary 1, the refined RQ algorithm is exact in the underloaded case. In the overloaded regime of Theorem 3, the value of b is immaterial, and Lemma 3 implies that the refined RQ algorithm is exact in the overloaded case as well.

The remaining challenge is the critically loaded case. Due to the additional term $w_{\tilde{c},k}(\tau u)$, the equation in (45) cannot be solved explicitly without a closed-form expression for $w_{c,k}$. We therefore calibrate b numerically by matching the exact heavy-traffic limits in (18) to the RQ heavy-traffic limit $\hat{Z}_{\mathrm{RQ},b}$. Specifically, we consider the $M/M/1+M$ model when $k = 1$ and the $M/M/1+E_k$ model when $k > 1$, where E_k denotes the Erlang distribution with shape parameter k and mean 1.

6 Numerical Experiments

We assess the accuracy of the proposed RQ approximation through extensive numerical experiments. In all experiments below, we report the *refined* RQ approximation computed from (44) with $b = \sqrt{2}$ and V defined in (39).

For comparison, we also compute the Ward–Glynn approximation [22, 23] (denoted "WG"), i.e., $\alpha^{-1/2}\mathbb{E}[\tilde{Z}(\infty)]$ with the expectation defined in (17) when $f(0) > 0$; the hazard-rate scaling approximation of [19] (denoted "Hazard rate") when $f(0) = 0$; and the Huang–Gurvich approximation of [10] (denoted "HG"). These benchmark methods are briefly reviewed in Appendix A.

The approximation of [10] is generally accurate for $M/GI/1{+}GI$ models when the system is not too lightly loaded and patience times are not too short, but it does not directly cover general arrival processes. A minor modification of the procedure in [10] allows us to apply it to models with general arrivals; see Appendix A. This modification, however, does not inherit the performance guarantee established in [10]. By contrast, the approximations of [19, 22, 23] are most accurate for critically loaded $GI/GI/1{+}GI$ models when the traffic intensity is close to 1 and patience times are long.

Evaluation of the performance. Throughout this section, we normalize the mean service time to 1. We vary the arrival rate over $\lambda \in \{1 - 2^{-k} : k = 1, 2, \ldots, 10\} \cup \{1 + 2^{-k} : k = -2, -1, \ldots, 10\}$, and the abandonment-rate parameter over $\alpha \in \{2^{-k} : k = 0, 1, \ldots, 13\}$, equivalently, the mean patience time $1/\alpha \in \{1, 2, 4, \ldots, 8192\}$. This yields $23 \times 14 = 322$ parameter combinations spanning underloaded, critically loaded, and overloaded regimes. For each parameter pair (λ, α) and each approximation, we report the signed relative error (approx $-$ exact)/exact as a heat map: blue shades indicate overestimation and red shades indicate underestimation. The color scale becomes darker as the magnitude of the error increases, and is clipped at $\pm 30\%$ relative error.

6.1 The $M/M/1{+}GI$ Models

The $M/M/1{+}GI$ models are tractable, and exact expressions are available in [30]. Our primary interest is in approximations for more general $G/GI/1{+}GI$ models; nevertheless, the $M/M/1{+}GI$

setting provides a useful baseline. A minimal requirement for any proposed approximation is that it performs well for this basic class.

Figure 2 reports results for the $M/M/1{+}M$ model (top row), the $M/M/1{+}E_2$ model (mid row), and the $M/M/1{+}H_2(4)$ model (bottom row).[1] These correspond to the cases $k = 1$ and $k = 2$ in Theorem 4, respectively. For the Erlang-2 case, the density is zero at the origin: $f(0) = 0$, while $f'(0) = 4$ under our normalization. Consequently, the WG approximation is not applicable, and we use the hazard-rate scaling approximation of [19] instead.

Figures 2 show that the refined RQ approximation is accurate over a wide range of traffic intensities and abandonment rates, and remains stable across the different patience-time distributions. The strong performance of the refined RQ approximation is achieved by carefully calibrating the parameter b in heavy traffic (Section 5.6). We emphasize, however, that this calibration is derived only from the baseline $M/M/1{+}M$ and $M/M/1{+}E_2$ models *in the heavy-traffic limit* as the mean patience time tends to infinity. The numerical results indicate that the resulting calibration is nevertheless robust: it remains effective for other patience-time distributions (e.g., for $H_2(4)$ patience times we reuse the $M/M/1{+}M$ calibration) and continues to perform well even when mean patience times are relatively short.

Unsurprisingly, all approximations work best in the critically loaded heavy-traffic regime, i.e., when $\alpha^{k/(k+1)} = c|1 - \rho|$ for small α and moderate c. This regime appears in the heat maps as a light band around $\rho \approx 1$ (small error), while departures from this scaling lead to visible degradation for the heavy-traffic-based benchmarks. In particular, for the $M/M/1{+}E_2$ model, the refined RQ heat map exhibits a thin transition stripe when λ is only slightly above 1 and the mean patience time is large; this is consistent with the $\alpha^{k/(k+1)}$ scaling boundary separating the critically loaded and overloaded regimes predicted by Theorems 2 and 3.

Away from critical loading, the benchmark methods behave quite differently. The WG and hazard-rate scaling approximations are highly accurate in their intended regime (near $\rho = 1$ with sufficiently long patience), but they can deteriorate sharply outside it, especially in overload. For example, in the $M/M/1{+}M$ model the WG approximation overestimates substantially in heavy overload and long patience; a similar (though slightly less extreme) behavior occurs for $M/M/1{+}H_2(4)$. For the $M/M/1{+}E_2$ model, the hazard-rate scaling approximation can likewise overestimate dramatically in both extreme underload and extreme overload. The HG approximation displays the opposite bias pattern: it is very accurate in overloaded regimes (even when patience is relatively short), but it significantly overestimates when the system is underloaded or near-critical and patience is short.

The most challenging regime for all methods is when the mean patience time is comparable to the mean service time (high abandonment), where the system increasingly resembles a loss model. In this regime the refined RQ approximation exhibits a consistent negative bias. As the mean patience time increases, the error decreases rapidly; already for mean patience in the range 8–32 the refined RQ error is typically in the single-digit percentage range across the grid, and for longer patience times it becomes uniformly small.

[1] E_2 is the Erlang distribution with shape parameter 2 with SCV 0.5. $H_2(4)$ is the hyperexponential distribution with balanced mean and SCV 4.

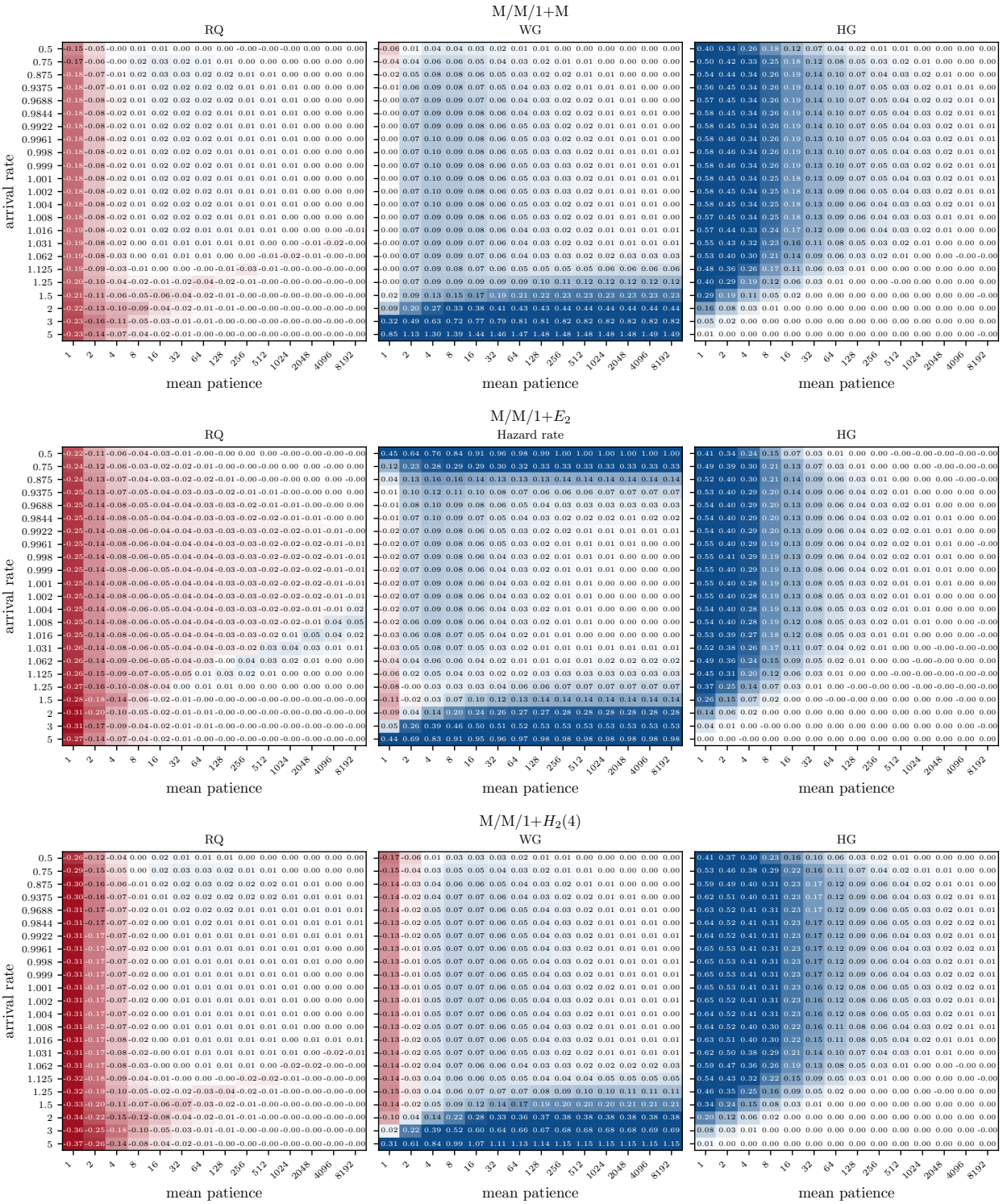

Figure 2: Signed relative error heat maps for the refined RQ approximation (left column), the Ward–Glynn approximation [22, 23] (center column; replaced by the hazard-rate scaling approximation [19] when $f(0) = 0$), and the Huang–Gurvich approximation [10] (right column), for the $M/M/1{+}M$ model (top row), the $M/M/1{+}E_2$ model (mid row), and the $M/M/1{+}H_2(4)$ model (bottom row).

6.2 The $GI/GI/1{+}GI$ Models

When the arrival process is Poisson, steady-state performance is often relatively insensitive to higher-order features of the service-time distribution beyond its mean and variance (provided the third moment is not excessively large). Consequently, approximation accuracy for $M/GI/1{+}GI$ models typically does not degrade substantially relative to the $M/M/1{+}GI$ baseline. Figure 3 illustrates this robustness for lognormal service times: the refined RQ approximation remains accurate for both the $M/LN(1,4)/1{+}H_2(4)$ and $M/LN(1,4)/1{+}E_2$ models over most of the parameter grid. Here $LN(1,4)$ denotes the lognormal distribution with mean 1 and variance 4.

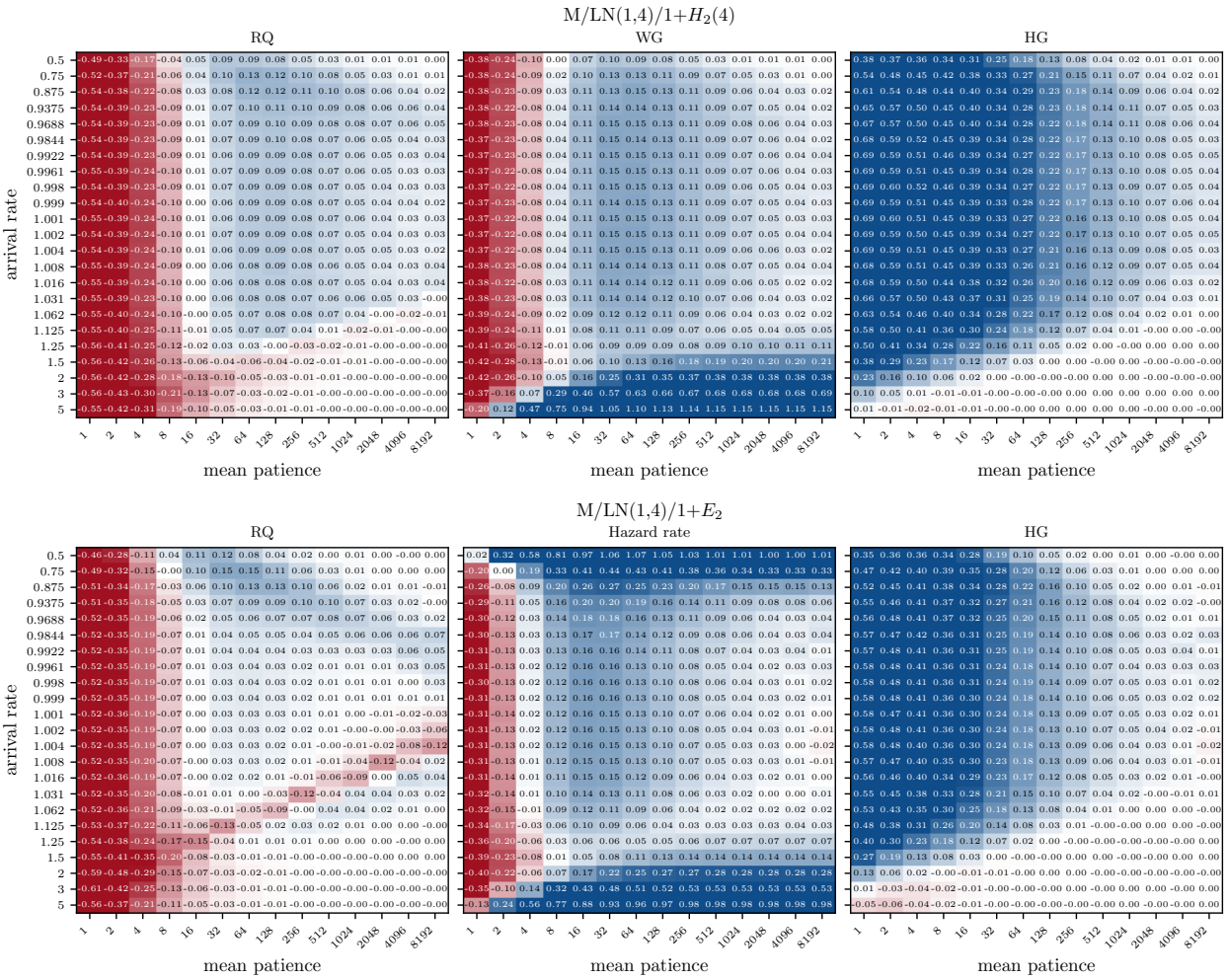

Figure 3: Signed relative error heat maps for the refined RQ approximation (left column), the Ward–Glynn approximation [22, 23] (center column; replaced by the hazard-rate scaling approximation [19] when $f(0) = 0$), and the Huang–Gurvich approximation [10] (right column), for the $M/LN(1,4)/1{+}H_2(4)$ model (top row) and the $M/LN(1,4)/1{+}E_2$ model (bottom row).

Models with non-Poisson renewal arrival processes are usually more challenging, even when the service times are exponential. Classical approaches often describe renewal input via two moments (rate and SCV c_a^2), which is justified by heavy-traffic limits where c_a^2 appears explicitly. Away from

heavy traffic, however, the *effective* arrival variability reflected in steady-state performance can lie between 1 and c_a^2, so a two-moment characterization can be too crude.

Figure 4 reports results for the $E_2/LN(1,2)/1{+}E_2$ model (top), the $H_2(4)/LN(1,2)/1{+}H_2(4)$ model (mid), and the $H_2(4)/LN(1,2)/1{+}E_2$ model (bottom). Here $LN(1,2)$ denotes the lognormal distribution with mean 1 and SCV 2. The benchmark methods behave as expected: the Ward–Glynn and hazard-rate scaling approximations are most accurate near critical loading (where they are theoretically justified), while the Huang–Gurvich approximation can deteriorate when patience times are short. By contrast, the refined RQ approximation remains stable across regimes.

6.3 Non-Renewal Arrival Processes

We next demonstrate that the RQ approximations extend naturally to *tandem* (queues-in-series) models. We consider a two-node system in which Queue 1 is a stable single-server queue without abandonment, and customers departing Queue 1 immediately join Queue 2, where abandonment is allowed. Our goal is to approximate the mean steady-state virtual waiting time at Queue 2.

Unless Queue 1 is an $M/M/1$ queue, the departure process from Queue 1 (and hence the arrival process to Queue 2) is generally *not* a renewal process. This does not pose a fundamental difficulty for our approach: in both RQ formulations, the arrival process enters through the IDC function $I_a(\cdot)$. While heavy-traffic limits for $G/GI/1{+}GI$ queues depend on a general arrival process only through the rate λ and the asymptotic variability parameter c_a^2, performance at typical traffic intensities can depend on more detailed temporal dependence. As emphasized in [27], the IDC function captures substantially more information than the standard two-moment descriptors (rate and SCV), and can therefore support more accurate approximations.

To implement the RQ approximations for Queue 2, we replace the input IDC $I_a(t)$ in (10) and (40) by the IDC of the departure process from Queue 1. We approximate this departure IDC using the IDC-propagation method developed for queueing networks in [29]; see Appendix B for the explicit formula.

Figure 5 reports signed relative errors for two tandem models: $H_2(4)/E_2/1 \to \cdot/M/1{+}H_2(4)$ (top row) and $E_2/H_2(4)/1 \to \cdot/M/1{+}E_2$ (bottom row). Overall, the refined RQ approximation remains accurate despite the non-renewal input to Queue 2, with absolute relative error below 20% in nearly all parameter instances with $\alpha \le 2^{-1}$.

7 Conclusion

In this paper we developed Robust Queueing (RQ) approximations for the $GI/GI/1{+}GI$ model, with the goal of accurately and efficiently approximating the mean steady-state virtual waiting time under general primitives. The key modeling step is to work with the reverse-time supremum representation of the offered waiting time, which reduces steady-state performance estimation to characterizing the drift and the scale-dependent variability of *effective* net-input increments. For the drift, a Poisson-surrogate compensator yields a simple approximation that is exact under

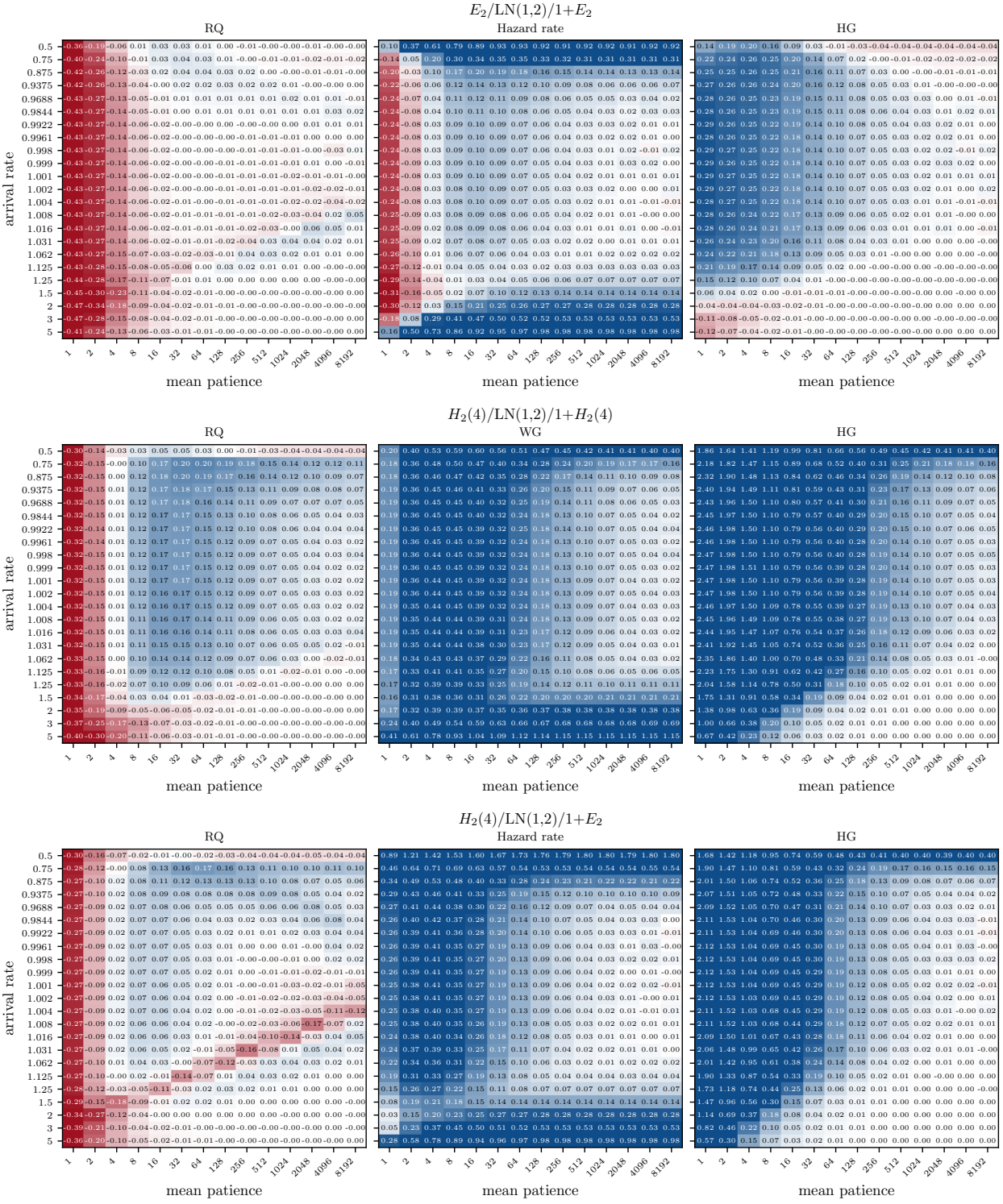

Figure 4: Signed relative error heat maps for the refined RQ approximation (left column), the Ward–Glynn approximation [22, 23] (center column, when $f(0) > 0$; replaced by the hazard-rate scaling approximation [19] when $f(0) = 0$), and the Huang–Gurvich approximation [10] (right column), for the $E_2/LN(1,2)/1{+}E_2$ (top row), $H_2(4)/LN(1,2)/1{+}H_2(4)$ (mid row), and $H_2(4)/LN(1,2)/1{+}E_2$ (bottom row) models.

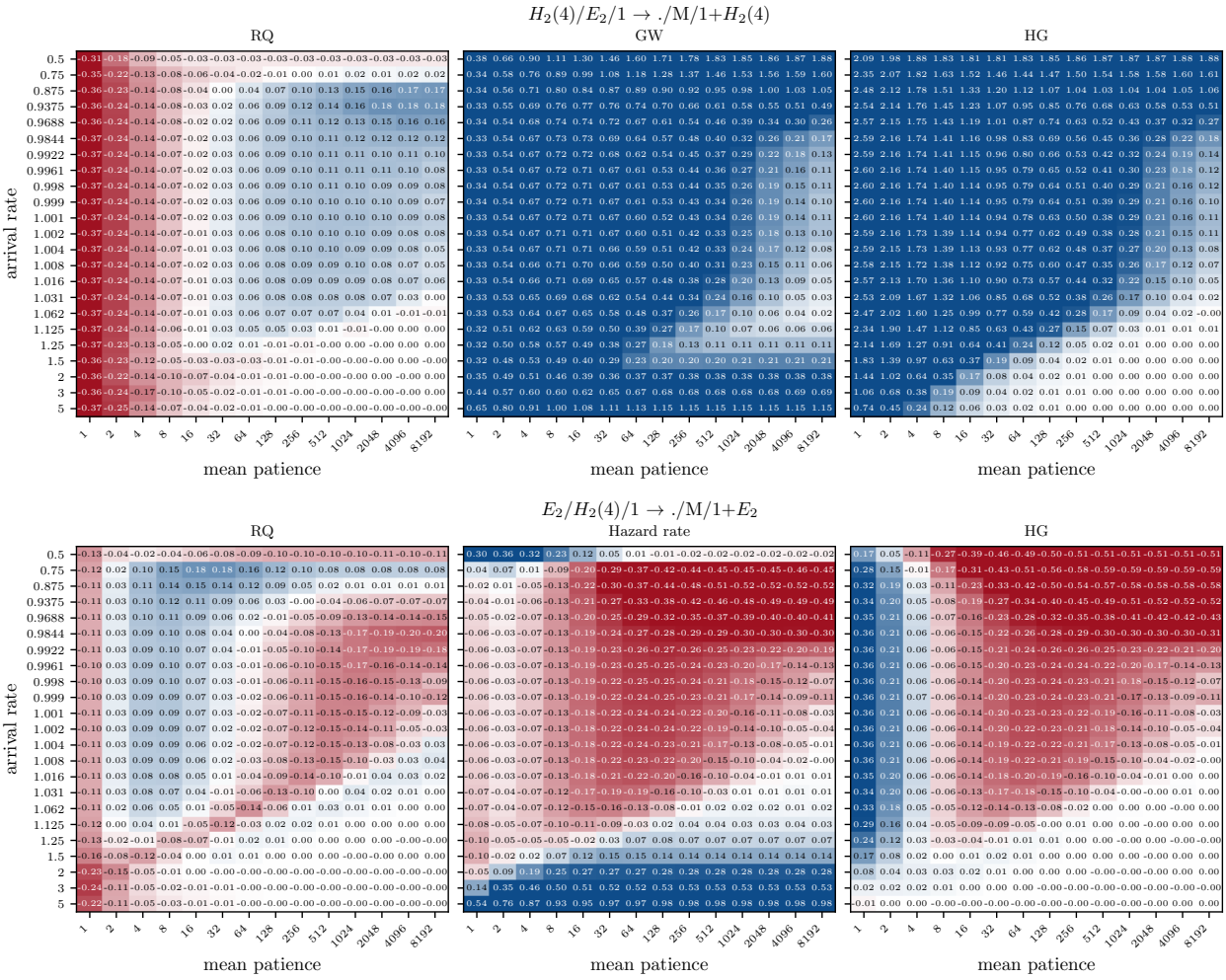

Figure 5: Signed relative error heat maps for tandem systems with non-renewal arrivals at Queue 2. Top row: $H_2(4)/E_2/1 \to \cdot/M/1{+}H_2(4)$. Bottom row: $E_2/H_2(4)/1 \to \cdot/M/1{+}E_2$. In each row, the left panel corresponds to the (refined) RQ approximation, the middle panel corresponds to the Ward–Glynn approximation when applicable (top row) or its hazard-rate-scaling variant when $f(0) = 0$ (bottom row), and the right panel corresponds to the Huang–Gurvich approximation.

Poisson arrivals and asymptotically justified under renewal arrivals in the long-patience regime. For variability, we proposed two RQ algorithms: a first algorithm based on a deterministic time-change of the renewal input, and a refined algorithm whose variance surrogate is informed by a heavy-traffic limit and explicitly incorporates the variance-reduction effect induced by abandonment. Both algorithms lead to a tractable one-dimensional fixed-point equation that can be solved rapidly by bisection, and the resulting approximation can be combined with standard identities to approximate other steady-state measures such as the abandonment probability and the mean waiting time of served customers.

Our numerical study indicates that the refined RQ approximation is accurate and stable over a wide range of traffic intensities and abandonment rates, including parameter regimes that are practically relevant but lie far outside the critically loaded scaling for which classical heavy-traffic approximations are designed.

Finally, we showed that the RQ framework naturally extends beyond renewal input: when arrivals are described through their IDC functions, the same fixed-point formulation applies to non-renewal arrival streams. This makes it possible to treat queues in series by approximating the IDC of the departure process of an upstream $GI/GI/1$ queue and feeding it into the RQ algorithm for the downstream abandonment queue. Promising directions for future work include extending these ideas to larger queueing networks with abandonment, developing data-driven procedures to estimate IDC/IDW inputs and to calibrate the robustness parameter in a model-adaptive manner, and providing sharper theoretical guarantees for the non-renewal and network settings.

Acknowledgments

W. You's research is generously supported by the Hong Kong Research Grants Council [Grant GRF 16212823] and [Theme-based Research Project T32-615/24-R].

References

[1] F Baccelli, P Boyer, and G Hebuterne. Single-server queues with impatient customers. *Advances in Applied Probability*, 16(4):887–905, 1984.

[2] Chaithanya Bandi, Dimitris Bertsimas, and Nataly Youssef. Robust queueing theory. *Operations Research*, 63(3):676–700, 2015.

[3] Patrick Billingsley. *Convergence of probability measures*. John Wiley & Sons, 2013.

[4] Mireille Bossy, Mamadou Cissé, and Denis Talay. Stochastic representations of derivatives of solutions of one-dimensional parabolic variational inequalities with neumann boundary conditions. In *Annales de l'IHP Probabilités et statistiques*, volume 47(2), pages 395–424, 2011.

[5] Sid Browne and Ward Whitt. Piecewise-linear diffusion processes. In *Advances in Queueing Theory, Methods, and Open Problems*, pages 463–480. CRC Press, 1995.

[6] JG Dai and Shuangchi He. Many-server queues with customer abandonment: A survey of diffusion and fluid approximations. *Journal of Systems Science and Systems Engineering*, 21(1):1–36, 2012.

[7] Kerry W Fendick and Ward Whitt. Measurements and approximations to describe the offered traffic and predict the average workload in a single-server queue. *Proceedings of the IEEE*, 77(1):171–194, 2002.

[8] Ofer Garnett, Avishai Mandelbaum, and Martin Reiman. Designing a call center with impatient customers. *Manufacturing & Service Operations Management*, 4(3):208–227, 2002.

[9] Shlomo Halfin and Ward Whitt. Heavy-traffic limits for queues with many exponential servers. *Operations research*, 29(3):567–588, 1981.

[10] Junfei Huang and Itai Gurvich. Beyond heavy-traffic regimes: Universal bounds and controls for the single-server queue. *Operations Research*, 66(4):1168–1188, 2018.

[11] Otis B Jennings and Josh E Reed. An overloaded multiclass fifo queue with abandonments. *Operations research*, 60(5):1282–1295, 2012.

[12] Olav Kallenberg. *Foundations of modern probability*. Springer, 3rd edition, 2021.

[13] Chihoon Lee and Ananda Weerasinghe. Convergence of a queueing system in heavy traffic with general patience-time distributions. *Stochastic Processes and their Applications*, 121(11):2507–2552, 2011.

[14] Chihoon Lee, Amy R Ward, and Heng-Qing Ye. Stationary distribution convergence of the offered waiting processes for GI/GI/1+ GI queues in heavy traffic. *Queueing Systems*, 94(1):147–173, 2020.

[15] Chihoon Lee, Amy R Ward, and Heng-Qing Ye. Stationary distribution convergence of the offered waiting processes in heavy traffic under general patience time scaling. *Queueing Systems*, 99(3):283–303, 2021.

[16] Dominique Lépingle, David Nualart, and Marta Sanz. Dérivation stochastique de diffusions réfléchies. In *Annales de l'IHP Probabilités et statistiques*, volume 25(3), pages 283–305, 1989.

[17] Ivan Nourdin and Giovanni Peccati. *Normal approximations with Malliavin calculus: from Stein's method to universality*, volume 192. Cambridge University Press, 2012.

[18] David Nualart. *The Malliavin calculus and related topics*. Springer, 2006.

[19] Josh E Reed and Amy R Ward. Approximating the GI/GI/1+GI queue with a nonlinear drift diffusion: Hazard rate scaling in heavy traffic. *Mathematics of Operations Research*, 33(3):606–644, 2008.

[20] Robert E Stanford. Reneging phenomena in single channel queues. *Mathematics of Operations Research*, 4(2):162–178, 1979.

[21] Amy R Ward. Asymptotic analysis of queueing systems with reneging: A survey of results for FIFO, single class models. *Surveys in Operations Research and Management Science*, 17(1):1–14, 2012.

[22] Amy R Ward and Peter W Glynn. A diffusion approximation for a markovian queue with reneging. *Queueing systems*, 43(1):103–128, 2003.

[23] Amy R Ward and Peter W Glynn. A diffusion approximation for a GI/GI/1 queue with balking or reneging. *Queueing Systems*, 50(4):371–400, 2005.

[24] Ward Whitt. The queueing network analyzer. *The bell system technical journal*, 62(9):2779–2815, 1983.

[25] Ward Whitt and Wei You. Heavy-traffic limit of the GI/GI/1 stationary departure process and its variance function. *Stochastic Systems*, 8(2):143–165, 2018.

[26] Ward Whitt and Wei You. Using robust queueing to expose the impact of dependence in single-server queues. *Operations Research*, 66(1):184–199, 2018.

[27] Ward Whitt and Wei You. The advantage of indices of dispersion in queueing approximations. *Operations Research Letters*, 47(2):99–104, 2019.

[28] Ward Whitt and Wei You. Time-varying robust queueing. *Operations Research*, 67(6):1766–1782, 2019.

[29] Ward Whitt and Wei You. A robust queueing network analyzer based on indices of dispersion. *Naval Research Logistics (NRL)*, 69(1):36–56, 2022.

[30] Sergey Zeltyn and Avishai Mandelbaum. Call centers with impatient customers: Many-server asymptotics of the M/M/n+G queue. *Queueing Systems*, 51(3):361–402, 2005.

Appendix

Contents

A Review of Existing Methods

Exact formula for $M/M/1+GI$ models. Let $H(x)=\int_0^x \bar{F}_\alpha(u)du$, then from [30]

$$\begin{aligned}\mathbb{E}[Z] &= \frac{\lambda\int_0^\infty x\exp(\lambda H(x)-\mu x)dx}{1+\lambda\int_0^\infty \exp(\lambda H(x)-\mu x)dx}\\ &= \frac{\int_0^\infty x\exp(\mu\int_0^x(\rho\bar{F}_\alpha(u)-1)du)dx}{1/\lambda+\int_0^\infty \exp(\mu\int_0^x(\rho\bar{F}_\alpha(u)-1)du)dx}.\end{aligned} \tag{46}$$

Approximation for critically-loaded $GI/GI/1+GI$ models based on the derivative at 0. Assuming the $F'(0)>0$, From Section 5 of [23]

$$\mathbb{E}[Z]\approx\alpha^{-1/2}\left[\frac{c}{F'(0)}+\frac{\phi\left(-\sqrt{2\mu}c/\sqrt{F'(0)\tilde{c}_x^2}\right)}{1-\Phi\left(-\sqrt{2\mu}c/\sqrt{F'(0)\tilde{c}_x^2}\right)}\sqrt{\frac{\tilde{c}_x^2}{2\mu F'(0)}}\right], \tag{47}$$

where $\tilde{c}_x^2=\rho c_a^2+(\rho\wedge 1)c_s^2$.

Approximation for critically-loaded $GI/GI/1+GI$ models based on hazard rate scaling [19].

$$\mathbb{E}[Z] \approx \frac{\int_0^\infty x \exp\left\{\frac{2\mu}{c_x^2}\int_0^x \left[\log(\bar{F}_\alpha(u)) + (\rho-1)\right] du\right\} dx}{\int_0^\infty \exp\left\{\frac{2\mu}{c_x^2}\int_0^x \left[\log(\bar{F}_\alpha(u)) + (\rho-1)\right] du\right\} dx},$$

where $c_x^2 = c_a^2 + c_s^2$.

Universal approximation for $M/GI/1+GI$ models in [10].

$$\mathbb{E}[Z] \approx \frac{\int_0^\infty x \exp\left\{\frac{2\mu}{(1+c_s^2)(\rho\wedge 1)}\int_0^x (\rho\bar{F}_\alpha(u) - 1) du\right\} dx}{\int_0^\infty \exp\left\{\frac{2\mu}{(1+c_s^2)(\rho\wedge 1)}\int_0^x (\rho\bar{F}_\alpha(u) - 1) du\right\} dx}.$$

Comparing with (46), for this approximation to be exact for the $M/M/1+GI$ model, one should add an additional constant $1/\lambda$ in the denominator and remove the modifier $\rho \wedge 1$ for the variability parameter $1 + c_s^2$.

Modification of [10] for $GI/GI/1+GI$ models. Formula from [10] can be modified to obtain a naive approximation for $GI/GI/1+GI$ models with non-Poisson renewal arrival processes. In particular, this is done by observing that exponential interarrival times have a SCV of $c_a^2 = 1$ and plugging in the corresponding SCV c_a^2 of the renewal arrival process.

$$\mathbb{E}[Z] \approx \frac{\int_0^\infty x \exp\left\{\frac{2\mu}{c_x^2(\rho\wedge 1)}\int_0^x (\rho\bar{F}_\alpha(u) - 1) du\right\} dx}{\int_0^\infty \exp\left\{\frac{2\mu}{c_x^2(\rho\wedge 1)}\int_0^x (\rho\bar{F}_\alpha(u) - 1) du\right\} dx},$$

where $c_x^2 = c_a^2 + c_s^2$.

B Tandem-Queue Approximation

This section summarizes how we extend the RQ approximations to a two-node tandem model, in which the downstream queue has abandonment and its input process is the *departure* process from an upstream $GI/GI/1$ queue. The key point is that, although the downstream arrivals are generally *non-renewal*, the refined RQ approximation only requires the arrival process through its *IDC* function; we therefore approximate the downstream arrival IDC using an IDC-based departure approximation from [25, 29].

Model. Consider two single-server FCFS queues in series. Queue 1 is a stable $GI/GI/1$ queue without abandonment. Its stationary departure process $D_1(\cdot)$ is routed to Queue 2, which is a $GI/GI/1{+}GI$ queue with patience-time distribution F_α. The external arrival rate to Queue 1 is λ, and Queue 1 has traffic intensity $\rho_1 < 1$; hence the throughput of Queue 1 is λ and the arrival rate to Queue 2 is also λ. For a stationary counting process $N(\cdot)$ with rate λ, recall the *index of dispersion for counts (IDC)*

$$I_N(t) \triangleq \frac{\mathrm{Var}\big(N(t) - N(0)\big)}{\mathbb{E}\big(N(t) - N(0)\big)} = \frac{\mathrm{Var}\big(N(t) - N(0)\big)}{\lambda t}, \qquad t > 0.$$

For renewal processes, $I_N(t)$ can be computed numerically from renewal-function or Laplace-transform representations; see, e.g., [29] and references therein. For general *non-renewal* processes (such as departures from a $GI/GI/1$ queue), we work directly with $I_N(\cdot)$.

Step 1: Approximate the departure IDC from Queue 1. Let $I_{a,1}(\cdot)$ be the IDC of the external arrival process to Queue 1 (a renewal process in our experiments), and let $I_{s,1}(\cdot)$ be the IDC of the *equilibrium service renewal process rescaled to rate* λ (i.e., the service-time renewal process with interrenewal times $\rho_1 S_1$, so that $\mathbb{E}[\rho_1 S_1] = 1/\lambda$). Let $c_{a,1}^2 \triangleq I_{a,1}(\infty)$ and $c_{s,1}^2 \triangleq I_{s,1}(\infty)$, and define $c_{x,1}^2 \triangleq c_{a,1}^2 + c_{s,1}^2$.

Following [29], we approximate the stationary departure IDC from Queue 1 by the convex combination

$$I_{d,1}(t) \approx w_{\rho_1}(t) I_{a,1}(t) + \big(1 - w_{\rho_1}(t)\big) I_{s,1}(t), \qquad t \geq 0, \tag{48}$$

where the weight function is

$$w_{\rho_1}(t) \equiv w^* \left(\frac{(1-\rho_1)^2 \lambda t}{\rho_1 c_{x,1}^2} \right),$$

and $w^*(\cdot)$ is the heavy-traffic limiting weight derived from the canonical RBM correlation structure; an explicit closed form is available (see [29]): for $u > 0$,

$$w^*(u) = \frac{1}{2u} \left((u^2 + 2u - 1)(2\Phi(\sqrt{u}) - 1) + 2\phi(\sqrt{u})\sqrt{u}(1+u) - u^2 \right), \tag{49}$$

where ϕ and Φ are the standard normal density and distribution functions, respectively. The function $w^*(u)$ is increasing and satisfies $0 \leq w^*(u) \leq 1$ [29], so (48) interpolates smoothly between service-scale variability (small t) and arrival-scale variability (large t).

Step 2: Use the departure IDC as the downstream arrival IDC. Because Queue 2 receives the departures from Queue 1, the arrival IDC to Queue 2 is exactly the departure IDC from Queue 1, so we set

$$I_{a,2}(t) \approx I_{d,1}(t), \qquad t \geq 0. \tag{50}$$

Step 3: The IDW input to the refined RQ approximation at Queue 2. In the refined RQ approximation for the $GI/GI/1{+}GI$ model, the arrival process enters through the *effective IDW* (see (40)–(41)). For the tandem system, we use (50) and set

$$\hat{I}_{w,2}(t) \triangleq \frac{I_{a,2}(t)}{\rho_2 \vee 1} + \left(1 - \frac{1}{\rho_2 \vee 1}\right) + c_{s,2}^2, \qquad t \geq 0, \tag{51}$$

where $\rho_2 = \lambda/\mu_2$ is the nominal traffic intensity of Queue 2 (ignoring abandonment), and $c_{s,2}^2$ is the service-time SCV at Queue 2 (e.g., $c_{s,2}^2 = 1$ for exponential service). The abandonment-modulated IDW is then obtained exactly as in (41) by multiplying $\hat{I}_{w,2}(t)$ with the abandonment factor $w_{\tilde{c},k}(\alpha^{2h}\tau t)$ from the refined RQ algorithm.

To approximate the mean steady-state virtual waiting time in Queue 2 for a tandem system: (i) compute/approximate $I_{a,1}(\cdot)$ and $I_{s,1}(\cdot)$, (ii) approximate $I_{d,1}(\cdot)$ via (48)–(49), (iii) set $I_{a,2}(\cdot) \approx$

$I_{d,1}(\cdot)$, and (iv) run the refined RQ procedure for Queue 2 using (51) in place of (40). This is precisely the IDC-based propagation mechanism advocated in [29].

C Proofs

C.1 Proof of Lemma 1

Proof. Taking expectations of the increment $A_0(t)-A_0(t-s) = M_0(t)-M_0(t-s)+\Lambda_0(t)-\Lambda_0(t-s)$ and $Y(t)-Y(t-s)$ gives

$$\mathbb{E}[Y(t)-Y(t-s)] = \frac{\lambda}{\mu}\mathbb{E}\left[\int_{t-s}^{t} \bar{F}_\alpha(Z(u-))du\right].$$

Since $N(t)-N(t-s) = (Y(t)-Y(t-s))-s$, we obtain the desired result.

Assume now that $Z(\cdot)$ is strictly stationary. Define $g(x) = \bar{F}_\alpha(x)$. Because $0 \le g(Z(u-)) \le 1$, Tonelli's theorem yields

$$\mathbb{E}\left[\int_{t-s}^{t} g(Z(u-))du\right] = \int_{t-s}^{t} \mathbb{E}[g(Z(u-))]du.$$

By stationarity, $Z(u-) \overset{d}{=} Z(0-)$ for all u, so $\mathbb{E}[g(Z(u-))] = \mathbb{E}[g(Z(0-))]$ and hence

$$\mathbb{E}\left[\int_{t-s}^{t} \bar{F}_\alpha(Z(u-))du\right] = s\mathbb{E}[\bar{F}_\alpha(Z(0-))].$$

This completes the proof. □

C.2 Proof of Lemma 2

Proof. Fix $s \ge 0$. Write $g_\alpha(u) \triangleq \bar{F}_\alpha(Z(u-)) \in [0,1]$. By definition,

$$\delta_t(s) = \int_{t-s}^{t} g_\alpha(u)d(A(u)-\lambda u) = \int_{t-s}^{t} g_\alpha(u)dA(u) - \lambda\int_{t-s}^{t} g_\alpha(u)du.$$

Therefore,

$$|\delta_t(s)| \le \int_{t-s}^{t} g_\alpha(u)dA(u) + \lambda\int_{t-s}^{t} g_\alpha(u)du \le A(t)-A(t-s)+\lambda s. \tag{52}$$

Let $A_t(s) \triangleq A(t)-A(t-s)$. For a renewal process, $A_t(s)$ converges in distribution, as $t\to\infty$, to the stationary renewal count over an interval of length s, and Blackwell's renewal theorem implies $\mathbb{E}[A_t(s)] \to \lambda s$ as $t\to\infty$. Since $A_t(s) \ge 0$, convergence in distribution together with convergence of means yields uniform integrability of $\{A_t(s) : t \ge 0\}$ (e.g., 3, Theorem 3.6). By (52), $\{\delta_t(s) : t \ge 0\}$ is uniformly integrable as well.

Let ξ be defined as in Lemma 3 and set $c_\rho \triangleq \bar{F}(\xi)$ (note that $\bar{F}_\alpha(x) = \bar{F}(\alpha x)$). Then we decompose

$$\delta_t(s) = c_\rho(A(t)-A(t-s)-\lambda s) + \int_{t-s}^{t}(g_\alpha(u)-c_\rho)d(A(u)-\lambda u).$$

Taking expectations gives

$$\mathbb{E}[\delta_t(s)] = c_\rho(\mathbb{E}[A(t) - A(t-s)] - \lambda s) + \mathbb{E}\left[\int_{t-s}^{t} (g_\alpha(u) - c_\rho) d(A(u) - \lambda u)\right]. \tag{53}$$

We control the second term in (53) using total variation. Let

$$\eta_{t,\alpha}(s) \triangleq \sup_{t-s\le u\le t} |g_\alpha(u) - c_\rho| \in [0,1].$$

Since $A(u) - \lambda u$ has total variation $A(t) - A(t-s) + \lambda s$ on $(t-s, t]$,

$$\left|\int_{t-s}^{t} (g_\alpha(u) - c_\rho) d(A(u) - \lambda u)\right| \le \eta_{t,\alpha}(s)(A(t) - A(t-s) + \lambda s).$$

By the dynamics of the virtual waiting time, over an interval of length s we have the crude bound

$$\sup_{t-s\le u\le t} |Z(u-) - Z(t)| \le s + \sum_{k=A(t-s)+1}^{A(t)} V_k,$$

and hence

$$\sup_{t-s\le u\le t} |\alpha Z(u-) - \alpha Z(t)| \le \alpha s + \alpha \sum_{k=A(t-s)+1}^{A(t)} V_k.$$

Because $\mathbb{E}[V_1] = 1/\mu$ and $\mathbb{E}[A(t) - A(t-s)]$ is $O(s)$, the right-hand side converges to 0 in L^1 (hence in probability) as $\alpha \downarrow 0$. Together with Lemma 3 and continuity of $\bar{F}$, it follows that

$$\lim_{\alpha\downarrow 0} \limsup_{t\to\infty} \mathbb{P}(\eta_{t,\alpha}(s) > \varepsilon) = 0, \qquad \forall \varepsilon > 0.$$

Since $\eta_{t,\alpha}(s) \le 1$ and $\{A(t) - A(t-s) + \lambda s : t \ge 0\}$ is uniformly integrable, we conclude that

$$\lim_{\alpha\downarrow 0} \lim_{t\to\infty} \mathbb{E}\left[\eta_{t,\alpha}(s)(A(t) - A(t-s) + \lambda s)\right] = 0,$$

and therefore the second term in (53) vanishes in the same iterated limit.

Finally, Blackwell's theorem gives $\mathbb{E}[A(t) - A(t-s)] - \lambda s \to 0$ as $t \to \infty$, so the first term in (53) also vanishes. This proves Lemma 2. □

C.3 Proof of Lemma 3

Proof. If $\rho < 1$, then the virtual waiting time in the system with abandonment is stochastically dominated by the workload in the corresponding $GI/GI/1$ queue without abandonment, which has a proper stationary distribution. Hence $Z(t) = O_p(1)$ in steady state, and therefore $\alpha Z(t) \Rightarrow 0$ as $\alpha \downarrow 0$.

If $\rho = 1$, Theorem 3 implies that $\alpha^h Z(t)$ converges weakly to a nondegenerate reflected diffusion limit in steady state, with $h = k/(k+1) < 1$. In particular, $Z(t) = O_p(\alpha^{-h})$, and thus $\alpha Z(t) = \alpha^{1-h}(\alpha^h Z(t)) \Rightarrow 0$.

If $\rho > 1$, we invoke Jennings and Reed [11, Theorem 1], which yields that the steady-state scaled virtual waiting time $\alpha Z(t)$ converges, as $\alpha \downarrow 0$, to the unique solution of the associated fluid equilibrium equation (see, e.g., equation (19) therein). One can verify from that characterization that the limit ξ satisfies $\rho\bar{F}(\xi) = 1$, equivalently $F(\xi) = (\rho-1)/\rho$, which gives $\xi = F^{-1}((\rho-1)/\rho)$. □

C.4 Proof of Theorem 1

For notational simplicity, assume $\rho(\alpha) = 1 + c\alpha^\gamma$ for all $\alpha > 0$. Let $Z_\alpha^{\mathrm{RQ}} \equiv Z_{\mathrm{RQ}_1,b}$ denote the steady-state RQ solution at parameter α. Equation (14) can be written as

$$Z_\alpha^{\mathrm{RQ}} = \sup_{s\geq 0}\left\{\rho(\alpha)s - \frac{s}{\bar{F}(\alpha Z_\alpha^{\mathrm{RQ}})} + b\sqrt{\frac{\rho(\alpha)sI_w(\lambda s)}{\mu}}\right\}, \qquad \lambda = \rho(\alpha)\mu. \tag{54}$$

Critically loaded ($\gamma \geq h$). Let $\hat{Z}_\alpha \triangleq \alpha^h Z_\alpha^{\mathrm{RQ}}$. Multiply (54) by α^h and set $u = \alpha^{2h}s$:

$$\hat{Z}_\alpha = \sup_{u\geq 0}\left\{\alpha^{-h}\left(\rho(\alpha) - \frac{1}{\bar{F}(\alpha^{1-h}\hat{Z}_\alpha)}\right)u + b\sqrt{\frac{\rho(\alpha)uI_w(\lambda\alpha^{-2h}u)}{\mu}}\right\}.$$

By Assumption 2,

$$F(\alpha^{1-h}x) = \frac{F^{(k)}(0)}{k!}\alpha^{k(1-h)}x^k + o(\alpha^{k(1-h)}) = \frac{F^{(k)}(0)}{k!}\alpha^h x^k + o(\alpha^h),$$

uniformly on compact x-sets. Hence

$$\frac{1}{\bar{F}(\alpha^{1-h}x)} = 1 + \frac{F^{(k)}(0)}{k!}\alpha^h x^k + o(\alpha^h), \quad \text{u.o.c. in } x,$$

and therefore

$$\alpha^{-h}\left(\rho(\alpha) - \frac{1}{\bar{F}(\alpha^{1-h}\hat{Z}_\alpha)}\right) = \mathbb{1}\{\gamma = h\}c - \frac{F^{(k)}(0)}{k!}\hat{Z}_\alpha^k + o(1).$$

Moreover, $I_w(\lambda\alpha^{-2h}u) \to c_x^2$ for each fixed $u > 0$. Passing to the limit yields $\hat{Z}_\alpha \to \hat{Z}$ satisfying

$$\hat{Z} = \sup_{u\geq 0}\left\{-\left(-\mathbb{1}\{\gamma = h\}c + \frac{F^{(k)}(0)}{k!}\hat{Z}^k\right)u + b\sqrt{\frac{c_x^2}{\mu}u}\right\}.$$

Evaluating the supremum gives

$$\hat{Z} = \frac{b^2c_x^2/\mu}{4\left(-\mathbb{1}\{\gamma = h\}c + \frac{F^{(k)}(0)}{k!}\hat{Z}^k\right)},$$

which is equivalent to (16).

Underloaded ($c < 0$ and $\gamma < h$). Let $\hat{Z}_\alpha \triangleq (1-\rho(\alpha))Z_\alpha^{\mathrm{RQ}} = (-c)\alpha^\gamma Z_\alpha^{\mathrm{RQ}}$. Multiply (54) by $(1-\rho(\alpha))$ and set $u = \alpha^{2\gamma}c^2 s$:

$$\hat{Z}_\alpha = \sup_{u\geq 0}\left\{-u + \alpha^{-\gamma}\left(1 - \frac{1}{\bar{F}(\alpha^{1-\gamma}\hat{Z}_\alpha)}\right)\frac{u}{c} + b\sqrt{\frac{\rho(\alpha)uI_w(\lambda\alpha^{-2\gamma}u/c^2)}{\mu}}\right\}.$$

Since $\gamma < h$, we have $k - \gamma(k+1) > 0$ and thus $\alpha^{-\gamma}F(\alpha^{1-\gamma}x) \to 0$ uniformly on compact x-sets, so the middle term vanishes u.o.c. Also, $I_w(\lambda\alpha^{-2\gamma}u/c^2) \to c_x^2$ for fixed $u > 0$. Therefore,

$$\hat{Z} = \sup_{u\geq 0}\left\{-u + b\sqrt{\frac{c_x^2}{\mu}u}\right\} = \frac{b^2c_x^2}{4\mu},$$

which implies the result.

Overloaded ($c > 0$ and $\gamma < h$). Let $\hat{Z}_\alpha \triangleq \alpha^{1-\gamma/k} Z_\alpha^{\mathrm{RQ}}$. Multiply (54) by $\alpha^{1-\gamma/k}$ and set $u = \alpha^{1+(k-1)\gamma/k} s$:

$$\hat{Z}_\alpha = \sup_{u \geq 0} \left\{ cu + \alpha^{-\gamma} \left(1 - \frac{1}{\bar{F}(\alpha^{\gamma/k} \hat{Z}_\alpha)} \right) u + b \sqrt{\alpha^{1-(k+1)\gamma/k} \frac{\rho(\alpha) u I_w(\lambda \alpha^{-1-(k-1)\gamma/k} u)}{\mu}} \right\}.$$

Since $\gamma < h$, the factor $\alpha^{1-(k+1)\gamma/k}$ tends to 0, so the square-root term vanishes uniformly on compact u-sets.

For the supremum to be finite, the linear coefficient of u must be nonpositive, which implies

$$c + \alpha^{-\gamma} \left(1 - \frac{1}{\bar{F}(\alpha^{\gamma/k} \hat{Z}_\alpha)} \right) \leq 0.$$

Using $1 - 1/\bar{F} = -F/\bar{F}$ and Taylor expansion,

$$\alpha^{-\gamma} \frac{F(\alpha^{\gamma/k} x)}{\bar{F}(\alpha^{\gamma/k} x)} \to \frac{F^{(k)}(0)}{k!} x^k, \quad \text{u.o.c. in } x,$$

so the previous inequality yields (for sufficiently small α)

$$c \leq \frac{F^{(k)}(0)}{k!} \hat{Z}_\alpha^k + o(1).$$

To obtain the reverse bound, fix $\varepsilon > 0$ and suppose that for some sequence $\alpha_n \downarrow 0$,

$$c + \alpha_n^{-\gamma} \left(1 - \frac{1}{\bar{F}(\alpha_n^{\gamma/k} \hat{Z}_{\alpha_n})} \right) \leq -\varepsilon.$$

Then, bounding $I_w(\cdot) \leq M$ for some finite M (using continuity of I_w and finiteness of $I_w(\infty)$), we obtain

$$\hat{Z}_{\alpha_n} \leq \sup_{u \geq 0} \left\{ -\varepsilon u + b \sqrt{\alpha_n^{1-(k+1)\gamma/k} \frac{\rho(\alpha_n) M u}{\mu}} \right\} = \frac{b^2 \rho(\alpha_n) M}{4\mu\varepsilon} \alpha_n^{1-(k+1)\gamma/k},$$

which implies $\hat{Z}_{\alpha_n} \to 0$ and thus contradicts the upper bound $c \leq (F^{(k)}(0)/k!) \hat{Z}_{\alpha_n}^k + o(1)$ with $c > 0$. Therefore, for all sufficiently small α,

$$c + \alpha^{-\gamma} \left(1 - \frac{1}{\bar{F}(\alpha^{\gamma/k} \hat{Z}_\alpha)} \right) \geq -\varepsilon,$$

and letting $\alpha \downarrow 0$ and then $\varepsilon \downarrow 0$ yields

$$c \geq \frac{F^{(k)}(0)}{k!} \hat{Z}^k, \qquad \hat{Z} = \lim_{\alpha \downarrow 0} \hat{Z}_\alpha.$$

Combining both bounds gives $\hat{Z}^k = ck!/F^{(k)}(0)$, proving the result.

C.5 Proof of Theorem 2

We repeatedly use the following uniform small-argument expansion to control $F(\alpha x)$, $\bar{F}(\alpha x)$, and $1/\bar{F}(\alpha x)$ as $\alpha \downarrow 0$.

Lemma 8 (Uniform small-argument expansion)**.** *Suppose Assumption 2 holds with index $k \geq 1$. Then for every $R > 0$,*

$$\sup_{0 \leq x \leq R} \left| \frac{F(\alpha x)}{\alpha^k} - \frac{F^{(k)}(0)}{k!} x^k \right| \longrightarrow 0 \qquad \text{as } \alpha \downarrow 0. \tag{55}$$

Consequently, for every $R > 0$,

$$\bar{F}(\alpha x) = 1 - \frac{F^{(k)}(0)}{k!} \alpha^k x^k + o(\alpha^k), \qquad \text{uniformly for } x \in [0, R], \tag{56}$$

$$\frac{1}{\bar{F}(\alpha x)} = 1 + \frac{F^{(k)}(0)}{k!} \alpha^k x^k + o(\alpha^k), \qquad \text{uniformly for } x \in [0, R]. \tag{57}$$

Proof. Fix $R > 0$. By Taylor's theorem with the mean-value remainder, for each $y \in [0, \alpha R]$ there exists $\theta = \theta(y) \in (0, 1)$ such that

$$F(y) = \sum_{j=0}^{k-1} \frac{F^{(j)}(0)}{j!} y^j + \frac{F^{(k)}(\theta y)}{k!} y^k.$$

Assumption 2 implies $F^{(j)}(0) = 0$ for $j < k$, hence $F(y) = F^{(k)}(\theta y) y^k / k!$. Taking $y = \alpha x$ with $x \in [0, R]$ gives

$$\left| \frac{F(\alpha x)}{\alpha^k} - \frac{F^{(k)}(0)}{k!} x^k \right| = \frac{x^k}{k!} \left| F^{(k)}(\theta \alpha x) - F^{(k)}(0) \right|.$$

Therefore,

$$\sup_{0 \leq x \leq R} \left| \frac{F(\alpha x)}{\alpha^k} - \frac{F^{(k)}(0)}{k!} x^k \right| \leq \frac{R^k}{k!} \sup_{0 \leq z \leq \alpha R} \left| F^{(k)}(z) - F^{(k)}(0) \right| \longrightarrow 0,$$

since $F^{(k)}$ is continuous at 0. This proves (55). Equation (56) follows from $\bar{F} = 1 - F$.

To obtain (57), note that (56) implies $\inf_{0 \leq x \leq R} \bar{F}(\alpha x) \to 1$, so for sufficiently small α we have $\inf_{0 \leq x \leq R} \bar{F}(\alpha x) \geq 1/2$. Then, uniformly over $x \in [0, R]$,

$$\frac{1}{\bar{F}(\alpha x)} = \frac{1}{1 - F(\alpha x)} = 1 + F(\alpha x) + O(F(\alpha x)^2) = 1 + \frac{F^{(k)}(0)}{k!} \alpha^k x^k + o(\alpha^k),$$

using (55) and the fact that $F(\alpha x) = O(\alpha^k)$ uniformly on $[0, R]$. □

Proof of Theorem 2. By Zeltyn and Mandelbaum [30, (9.9)], the mean steady-state virtual waiting time in the $M/M/1{+}GI$ model satisfies

$$\mathbb{E}[Z_\alpha] = \frac{\lambda J_1}{1 + \lambda J}, \tag{58}$$

where

$$J = \int_0^\infty \exp\left\{\lambda \int_0^x \bar{F}(\alpha u)du - \mu x\right\} dx,$$
$$J_1 = \int_0^\infty x \exp\left\{\lambda \int_0^x \bar{F}(\alpha u)du - \mu x\right\} dx.$$

Recall $\lambda = \rho(\alpha)\mu$ and $\rho(\alpha) = 1 + c\alpha^\gamma$.

In all three regimes considered below, the change of variables implies $\alpha x \to 0$ on the x-scale that contributes to J and J_1. Thus we may apply Lemma 8 on $[0, \alpha x]$ to obtain

$$\int_0^x \bar{F}(\alpha u)du = x - \frac{F^{(k)}(0)}{(k+1)!}\alpha^k x^{k+1} + o(\alpha^k x^{k+1}), \qquad \text{whenever } \alpha x \to 0. \tag{59}$$

Substituting (59) into the exponent yields

$$\lambda \int_0^x \bar{F}(\alpha u)du - \mu x = -\mu\big(1 - \rho(\alpha)\big)x - \lambda\frac{F^{(k)}(0)}{(k+1)!}\alpha^k x^{k+1} + o(\alpha^k x^{k+1}), \tag{60}$$

whenever $\alpha x \to 0$.

Underloaded ($c < 0$ and $\gamma < h$). Set $s = \mu(1 - \rho(\alpha))x$, so $x = s/(\mu(1 - \rho(\alpha)))$. Since $1 - \rho(\alpha) = -c\alpha^\gamma$ with $c < 0$ and $\gamma < h < 1$, we have $\alpha x = \alpha^{1-\gamma} s/(\mu(-c)) \to 0$ for each fixed s, so (60) applies. Moreover,

$$\alpha^k x^{k+1} = \frac{\alpha^k}{\mu^{k+1}(1-\rho(\alpha))^{k+1}} s^{k+1} = \frac{(-1/c)^{k+1}}{\mu^{k+1}} \alpha^{k-\gamma(k+1)} s^{k+1},$$

and $\gamma < h$ is equivalent to $k - \gamma(k+1) > 0$, so this term vanishes. Therefore, by dominated convergence,

$$\lim_{\alpha \downarrow 0} \mu(1-\rho(\alpha))J = \int_0^\infty e^{-s} ds = 1, \qquad \lim_{\alpha \downarrow 0} \mu^2(1-\rho(\alpha))^2 J_1 = \int_0^\infty s e^{-s} ds = 1.$$

Substituting into (58) yields $\mathbb{E}[Z_\alpha] \sim 1/(\mu(1-\rho(\alpha)))$ and hence $\mathbb{E}[Z_\alpha]/\mathbb{E}[Z_{M/M/1}] \to 1$.

Critically loaded ($\gamma \geq h$). Set $s = \alpha^h x$ (so $x = s/\alpha^h$). Then $\alpha x = \alpha^{1-h} s \to 0$ for fixed s, so (60) applies. Since $h = k/(k+1)$, we have $\alpha^k x^{k+1} = s^{k+1}$, and

$$-\mu(1-\rho(\alpha))x = \mu(\rho(\alpha) - 1)x = \mu c \alpha^{\gamma - h} s.$$

Using $\lambda \to \mu$ and dominated convergence, we obtain

$$\lim_{\alpha \downarrow 0} \alpha^h J = \int_0^\infty \exp\left\{c\mu s \mathbb{1}\{\gamma = h\} - \frac{\mu F^{(k)}(0)}{(k+1)!} s^{k+1}\right\} ds,$$
$$\lim_{\alpha \downarrow 0} \alpha^{2h} J_1 = \int_0^\infty s \exp\left\{c\mu s \mathbb{1}\{\gamma = h\} - \frac{\mu F^{(k)}(0)}{(k+1)!} s^{k+1}\right\} ds.$$

Substituting these limits into (58) gives (18).

Overloaded ($c > 0$ and $\gamma < h$). The overloaded case is handled via a Laplace-principle argument. Define the large parameter

$$r_\alpha \triangleq \alpha^{\gamma/h-1} = \alpha^{-\left(1-\gamma(k+1)/k\right)} \longrightarrow \infty \qquad (\alpha \downarrow 0),$$

and introduce the scaling $x = \alpha^{-(1-\gamma/k)}s$. With this scaling, the exponent can be written as

$$\lambda \int_0^x \bar{F}(\alpha u)du - \mu x = r_\alpha G_\alpha(s), \qquad G_\alpha(s) \triangleq \mu cs - \lambda \int_0^s \alpha^{-\gamma} F(\alpha^{\gamma/k} v)dv,$$

and hence

$$J = \alpha^{-1+\gamma/k} \int_0^\infty \exp\{r_\alpha G_\alpha(s)\}ds,$$
$$J_1 = \alpha^{-2+2\gamma/k} \int_0^\infty s \exp\{r_\alpha G_\alpha(s)\}ds.$$

Lemma 8 implies that $\alpha^{-\gamma}F(\alpha^{\gamma/k}v) \to (F^{(k)}(0)/k!)v^k$ uniformly on compact v-sets, and therefore $G_\alpha \to G$ uniformly on compact s-sets, where

$$G(s) \triangleq \mu cs - \mu \frac{F^{(k)}(0)}{(k+1)!} s^{k+1}.$$

The function G is strictly concave on $(0, \infty)$ and attains its unique maximum at

$$s^* = \left(\frac{ck!}{F^{(k)}(0)}\right)^{1/k}.$$

A standard Laplace-principle argument then yields that the probability measures proportional to $\exp\{r_\alpha G_\alpha(s)\}ds$ concentrate at s^* as $\alpha \downarrow 0$, and hence

$$\lim_{\alpha \downarrow 0} \alpha^{1-\gamma/k} \frac{J_1}{J} = s^*.$$

Finally, $G(s^*) > 0$ implies $J \to \infty$ and thus $\lambda J \to \infty$, so $\mathbb{E}[Z_\alpha] \sim J_1/J$. Therefore,

$$\lim_{\alpha \downarrow 0} \alpha^{1-\gamma/k} \mathbb{E}[Z_\alpha] = s^* = \left(\frac{ck!}{F^{(k)}(0)}\right)^{1/k},$$

which proves the result. □

C.6 Proof of Theorem 3

Fix $T > 0$. All processes below are viewed on $[0, T]$ and as elements of the space of càdlàg functions on $[0, T]$, $\mathbb{D}([0, T], \mathbb{R}^d)$, endowed with the J_1 topology. Write $e(t) = t$ and set

$$\beta \triangleq \frac{F^{(k)}(0)}{k!} > 0.$$

Recall $h = k/(k+1) \in (0, 1)$ and the heavy-traffic regime: $\mu^\alpha \to \mu$ and $c^\alpha \triangleq \alpha^{-h}(\rho^\alpha - 1) \to c$. Throughout, T_i^α denotes the ith arrival epoch and $W_i^\alpha = Z^\alpha(T_i^\alpha -)$ is the offered waiting time seen by customer i.

Skorohod reflection map. For a càdlàg function $y \in \mathbb{D}([0,T],\mathbb{R})$ define the classical one-dimensional Skorohod reflection map

$$\Gamma(y)(t) \triangleq y(t) - \inf_{0\le s\le t}(y(s)\wedge 0), \qquad \Lambda(y)(t) \triangleq -\inf_{0\le s\le t}(y(s)\wedge 0), \qquad t\in[0,T].$$

Then $x=\Gamma(y)$ and $\ell=\Lambda(y)$ are càdlàg, $x\ge 0$, ℓ is nondecreasing with $\ell(0)=0$, and $x=y+\ell$ together with the complementarity condition $\int_0^T \mathbb{1}\{x(t)>0\}d\ell(t)=0$. Moreover, for the sup-norm $\|f\|_T \triangleq \sup_{0\le t\le T}|f(t)|$,

$$\|\Gamma(y_1)-\Gamma(y_2)\|_T \le 2\|y_1-y_2\|_T, \qquad \|\Lambda(y_1)-\Lambda(y_2)\|_T \le 2\|y_1-y_2\|_T, \tag{61}$$

and if y is continuous then $(\Gamma(y),\Lambda(y))$ is continuous.

We will use the following fact, proved here for completeness.

Lemma 9 (Reflected integral equation with polynomial drift)**.** *Fix $T>0$ and define $g(x)\triangleq -\beta x^k$ on $\mathbb{R}_+$. For any $y\in\mathbb{D}([0,T],\mathbb{R})$ there exists a unique pair $(x,\ell)\in\mathbb{D}([0,T],\mathbb{R})^2$ such that*

$$x(t) = y(t) + \int_0^t g(x(s))ds + \ell(t), \qquad t\in[0,T], \tag{62}$$

with $x(t)\ge 0$, ℓ nondecreasing, $\ell(0)=0$, and $\int_0^T \mathbb{1}\{x(t)>0\}d\ell(t)=0$. In addition, if $y_n\to y$ uniformly on $[0,T]$, then the corresponding solutions satisfy $(x_n,\ell_n)\to(x,\ell)$ uniformly on $[0,T]$. If y is continuous, then (x,ℓ) is continuous.

Proof. Equation (62) is equivalent to the fixed point representation

$$x = \Gamma\left(y+\int_0^{\cdot} g(x(s))ds\right), \qquad \ell = \Lambda\left(y+\int_0^{\cdot} g(x(s))ds\right). \tag{63}$$

We first prove uniqueness. Suppose (x_i,ℓ_i) solve (62) for the same y, $i=1,2$. Let $u_i(t)\triangleq y(t)+\int_0^t g(x_i(s))ds$ so that $x_i=\Gamma(u_i)$. Since g is locally Lipschitz on $\mathbb{R}_+$, for any $R>0$ it is Lipschitz on $[0,R]$ with constant $L_R=\beta k R^{k-1}$. Because $g\le 0$, (63) implies $u_i\le y$ pointwise and hence $x_i=\Gamma(u_i)\le\Gamma(y)$ pointwise; in particular, $\|x_i\|_T\le\|\Gamma(y)\|_T<\infty$. Set $R\triangleq\|\Gamma(y)\|_T$. Then for all $t\le T$,

$$\|u_1-u_2\|_t \le \int_0^t |g(x_1(s))-g(x_2(s))|ds \le L_R\int_0^t \|x_1-x_2\|_s ds.$$

Using (61), $\|x_1-x_2\|_t\le 2\|u_1-u_2\|_t$, hence

$$\|x_1-x_2\|_t \le 2L_R\int_0^t \|x_1-x_2\|_s ds.$$

By Grönwall's inequality, $\|x_1-x_2\|_T=0$ and thus $x_1\equiv x_2$ and $\ell_1\equiv\ell_2$. This gives uniqueness.

For existence, define an iteration $x^{(0)}\triangleq\Gamma(y)$ and for $n\ge 0$

$$x^{(n+1)} \triangleq \Gamma\left(y+\int_0^{\cdot} g(x^{(n)}(s))ds\right).$$

As above, $0 \le x^{(n)} \le \Gamma(y)$ for all n, so all iterates are bounded by $R = \|\Gamma(y)\|_T$. Hence g is Lipschitz on the range of all iterates with constant L_R. Then

$$\|x^{(n+1)} - x^{(n)}\|_T \le 2 \left\| \int_0^{\cdot} (g(x^{(n)}(s)) - g(x^{(n-1)}(s)))ds \right\|_T \le 2L_R T \|x^{(n)} - x^{(n-1)}\|_T.$$

If $2L_R T < 1$, this is a contraction and yields convergence. For general T, split $[0,T]$ into finitely many subintervals on which $2L_R\Delta < 1$ and construct the solution by concatenation. This yields existence of a unique (x,ℓ) satisfying (62).

Continuity of the solution map under uniform convergence follows by repeating the uniqueness estimate with two different inputs y_1, y_2: if $x_i = \Gamma(y_i + \int g(x_i))$, then

$$\|x_1 - x_2\|_T \le 2\|y_1 - y_2\|_T + 2L_R \int_0^T \|x_1 - x_2\|_s ds \le 2e^{2L_R T}\|y_1 - y_2\|_T,$$

and similarly for $\ell_1 - \ell_2$ using (61). Finally, if y is continuous, then $u(t) = y(t) + \int_0^t g(x(s))ds$ is continuous and so $(x,\ell) = (\Gamma(u), \Lambda(u))$ is continuous. □

A tightness bound for $\tilde{Z}^\alpha$. Define the *unthinned* total-input process

$$Y^\alpha_{\rm tot}(t) \triangleq \sum_{i=1}^{A^\alpha(t)} V^\alpha_i, \qquad N^\alpha_{\rm tot}(t) \triangleq Z^\alpha(0) + Y^\alpha_{\rm tot}(t) - t.$$

Let $Z^\alpha_{\rm tot}(t) \triangleq \Gamma(N^\alpha_{\rm tot})(t)$ be the workload process of the corresponding $GI/GI/1$ queue *without* abandonment driven by the same primitives. Since $Y^\alpha(t) \le Y^\alpha_{\rm tot}(t)$ for all t, the monotonicity of Γ implies

$$Z^\alpha(t) \le Z^\alpha_{\rm tot}(t) \qquad \text{for all } t \ge 0, \tag{64}$$

and therefore $\tilde{Z}^\alpha(t) \le \tilde{Z}^\alpha_{\rm tot}(t)$ for the corresponding diffusion scalings.

A standard FCLT implies that the diffusion-scaled net input of the unthinned system converges to a Brownian motion with drift, and by the continuity of Γ at continuous limits, $\{\tilde{Z}^\alpha_{\rm tot}\}$ is tight in $\mathbb{D}([0,T],\mathbb{R})$. Hence by (64), $\{\tilde{Z}^\alpha\}$ is tight as well. In particular, for each T,

$$\lim_{R\to\infty} \limsup_{\alpha\downarrow 0} \ \mathbb{P}(\|\tilde{Z}^\alpha\|_T > R) = 0. \tag{65}$$

Abandonment count and the state-dependent drift. Define the *(eventual) abandonment count* among arrivals up to time t:

$$G^\alpha(t) \triangleq A^\alpha(t) - A^\alpha_0(t) = \sum_{i=1}^{A^\alpha(t)} \mathbb{1}\{D^\alpha_i \le W^\alpha_i\}.$$

Let $\tilde{G}^\alpha(t) \triangleq \alpha^h G^\alpha(\alpha^{-2h}t)$.

Let $\mathbb{F}^\alpha = \{\mathcal{F}^\alpha_t\}$ be the natural filtration generated by the arrivals and their marks up to time t. At the ith arrival time T^α_i, $W^\alpha_i = Z^\alpha(T^\alpha_i-)$ is $\mathcal{F}^\alpha_{T^\alpha_i-}$-measurable, while D^α_i is independent of $\mathcal{F}^\alpha_{T^\alpha_i-}$ with CDF $F_\alpha(x) = F(\alpha x)$. Hence

$$\mathbb{E}[\mathbb{1}\{D^\alpha_i \le W^\alpha_i\} \mid \mathcal{F}^\alpha_{T^\alpha_i-}] = F_\alpha(W^\alpha_i).$$

Therefore the process

$$M_G^\alpha(t) \triangleq \sum_{i=1}^{A^\alpha(t)} \Big(\mathbb{1}\{D_i^\alpha \le W_i^\alpha\} - F_\alpha(W_i^\alpha) \Big)$$

is an $\mathbb{F}^\alpha$-martingale. Using $G^\alpha(t) = \sum \mathbb{1}\{D_i^\alpha \le W_i^\alpha\}$, we have the decomposition

$$G^\alpha(t) = M_G^\alpha(t) + \sum_{i=1}^{A^\alpha(t)} F_\alpha(W_i^\alpha). \tag{66}$$

Define the scaled martingale $\tilde{M}_G^\alpha(t) \triangleq \alpha^h M_G^\alpha(\alpha^{-2h}t)$.

We now show that the martingale term is negligible. Note that its predictable quadratic variation satisfies

$$\langle M_G^\alpha \rangle(t) = \sum_{i=1}^{A^\alpha(t)} F_\alpha(W_i^\alpha)(1 - F_\alpha(W_i^\alpha)) \le \sum_{i=1}^{A^\alpha(t)} F_\alpha(W_i^\alpha).$$

Using (66), this implies $\langle M_G^\alpha \rangle(t) \le G^\alpha(t) + |M_G^\alpha(t)|$. Since $|M_G^\alpha|$ is controlled by $\langle M_G^\alpha \rangle$ in L^2, it suffices to note that $G^\alpha(\alpha^{-2h}T)$ is of order α^{-h} in expectation because each summand has conditional mean $F(\alpha W_i^\alpha)$, which is $O(\alpha^h)$ when $\tilde{Z}^\alpha$ is $O(1)$. A direct bound using (65) and the small-argument expansion yields $\sup_\alpha \mathbb{E}[\alpha^h G^\alpha(\alpha^{-2h}T)] < \infty$. Consequently,

$$\mathbb{E}\big[\langle \tilde{M}_G^\alpha \rangle(T)\big] = \alpha^{2h}\mathbb{E}\big[\langle M_G^\alpha \rangle(\alpha^{-2h}T)\big] = O(\alpha^h) \to 0,$$

and Doob's L^2 inequality gives

$$\sup_{0 \le t \le T} |\tilde{M}_G^\alpha(t)| \Rightarrow 0.$$

Write (66) at time $\alpha^{-2h}t$ and multiply by α^h:

$$\tilde{G}^\alpha(t) = \tilde{M}_G^\alpha(t) + \alpha^h \sum_{i=1}^{A^\alpha(\alpha^{-2h}t)} F_\alpha(W_i^\alpha). \tag{67}$$

Since $W_i^\alpha = Z^\alpha(T_i^\alpha -)$ and $\tilde{Z}^\alpha(s) = \alpha^h Z^\alpha(\alpha^{-2h}s)$, we have $\alpha W_i^\alpha = \alpha^{1-h}\tilde{Z}^\alpha(\alpha^{2h}T_i^\alpha -)$. Define the rescaled function

$$\widehat{F}_\alpha(x) \triangleq \alpha^{-h} F(\alpha^{1-h}x), \qquad x \ge 0.$$

Then

$$\alpha^h F_\alpha(W_i^\alpha) = \alpha^h F(\alpha W_i^\alpha) = \alpha^{2h}\widehat{F}_\alpha(\tilde{Z}^\alpha(\alpha^{2h}T_i^\alpha -)).$$

Also recall the fluid-scaled arrival process $\bar{A}^\alpha(t) = \alpha^{2h}A^\alpha(\alpha^{-2h}t)$, so each arrival contributes a jump of size α^{2h} to $\bar{A}^\alpha$. Hence the sum in (67) is the Lebesgue–Stieltjes integral

$$\alpha^h \sum_{i=1}^{A^\alpha(\alpha^{-2h}t)} F_\alpha(W_i^\alpha) = \int_0^t \widehat{F}_\alpha(\tilde{Z}^\alpha(s-))d\bar{A}^\alpha(s). \tag{68}$$

By Assumption 2, $F^{(j)}(0) = 0$ for $j < k$ and $F^{(k)}$ is continuous at 0, so Taylor's theorem yields: for each $R > 0$,

$$\sup_{0 \le x \le R} \Big| \widehat{F}_\alpha(x) - \beta x^k \Big| \to 0, \qquad \alpha \downarrow 0. \tag{69}$$

Moreover,

$$\bar{A}^\alpha(t) = \alpha^{2h} A^\alpha(\alpha^{-2h}t) = \lambda^\alpha t + \alpha^h \tilde{A}^\alpha(t),$$

so by (24) and $\alpha^h \to 0$,

$$\sup_{0\le t\le T}\left|\bar{A}^\alpha(t) - \lambda^\alpha t\right| \Rightarrow 0, \qquad \alpha \downarrow 0. \tag{70}$$

Combining (68), (69), (70), and the tightness (65), we obtain

$$\sup_{0\le t\le T}\left|\tilde{G}^\alpha(t) - \lambda^\alpha \beta \int_0^t \big(\tilde{Z}^\alpha(s)\big)^k ds\right| \Rightarrow 0. \tag{71}$$

Since $\lambda^\alpha = \rho^\alpha \mu^\alpha \to \mu$, (71) implies

$$\sup_{0\le t\le T}\left|\tilde{G}^\alpha(t) - \mu \beta \int_0^t \big(\tilde{Z}^\alpha(s)\big)^k ds\right| \Rightarrow 0. \tag{72}$$

Service-time fluctuation under thinning. Decompose the effective work-input process:

$$Y^\alpha(t) = \sum_{i=1}^{A^\alpha(t)} V_i^\alpha \mathbb{1}\{D_i^\alpha > W_i^\alpha\} = \frac{1}{\mu^\alpha} A_0^\alpha(t) + \sum_{i=1}^{A^\alpha(t)} \Big(V_i^\alpha - \frac{1}{\mu^\alpha}\Big)\mathbb{1}\{D_i^\alpha > W_i^\alpha\}.$$

Define the centered service-time fluctuation process

$$S_{\mathrm{fl}}^\alpha(t) \triangleq \sum_{i=1}^{A^\alpha(t)} \Big(V_i^\alpha - \frac{1}{\mu^\alpha}\Big)\mathbb{1}\{D_i^\alpha > W_i^\alpha\}, \qquad \tilde{S}_{\mathrm{fl}}^\alpha(t) \triangleq \alpha^h S_{\mathrm{fl}}^\alpha(\alpha^{-2h}t).$$

Then, using $\rho^\alpha = \lambda^\alpha/\mu^\alpha$, the definition (23a) yields the exact identity

$$\tilde{Y}^\alpha(t) = \frac{1}{\mu^\alpha}\tilde{A}_0^\alpha(t) + \tilde{S}_{\mathrm{fl}}^\alpha(t). \tag{73}$$

We next show $\tilde{S}_{\mathrm{fl}}^\alpha \Rightarrow \mu^{-1} c_s B_s(\mu e)$. Let the unthinned centered sum be

$$S_{\mathrm{all}}^\alpha(t) \triangleq \sum_{i=1}^{A^\alpha(t)} \Big(V_i^\alpha - \frac{1}{\mu^\alpha}\Big), \qquad \tilde{S}_{\mathrm{all}}^\alpha(t) \triangleq \alpha^h S_{\mathrm{all}}^\alpha(\alpha^{-2h}t).$$

By Donsker's invariance principle applied to the i.i.d. sequence $\{V_i^\alpha\}$ and the random time-change theorem (using $\bar{A}^\alpha \Rightarrow \mu e$), we have

$$\tilde{S}_{\mathrm{all}}^\alpha \Rightarrow \mu^{-1} c_s B_s \circ (\mu e) \qquad \text{in } \mathbb{D}([0,T],\mathbb{R}), \tag{74}$$

where B_s is a standard Brownian motion independent of B_a.

Now note that

$$\tilde{S}_{\mathrm{fl}}^\alpha(t) = \tilde{S}_{\mathrm{all}}^\alpha(t) - \alpha^h \sum_{i=1}^{A^\alpha(\alpha^{-2h}t)} \Big(V_i^\alpha - \frac{1}{\mu^\alpha}\Big)\mathbb{1}\{D_i^\alpha \le W_i^\alpha\}.$$

Denote the last term by $\Delta^\alpha(t)$. Conditioning on the history up to each arrival epoch, the indicator $\mathbb{1}\{D_i^\alpha \le W_i^\alpha\}$ is independent of V_i^α and $\mathbb{E}[V_i^\alpha - 1/\mu^\alpha] = 0$, so the unscaled version of the sum is an $\mathbb{F}^\alpha$-martingale. Its predictable quadratic variation satisfies

$$\langle \Delta^\alpha/\alpha^h\rangle(\alpha^{-2h}t) = \mathrm{Var}(V^\alpha) \sum_{i=1}^{A^\alpha(\alpha^{-2h}t)} \mathbb{1}\{D_i^\alpha \le W_i^\alpha\} = \mathrm{Var}(V^\alpha) G^\alpha(\alpha^{-2h}t).$$

Therefore,

$$\langle \Delta^\alpha\rangle(t) = \alpha^{2h}\mathrm{Var}(V^\alpha)G^\alpha(\alpha^{-2h}t) = \mathrm{Var}(V^\alpha)\alpha^h \tilde{G}^\alpha(t).$$

By (72), $\tilde{G}^\alpha$ is tight, and since $\alpha^h \to 0$, $\sup_{t\le T}\langle\Delta^\alpha\rangle(t) \Rightarrow 0$. Doob's inequality yields

$$\sup_{0\le t\le T} |\Delta^\alpha(t)| \Rightarrow 0. \tag{75}$$

Combining (74) and (75) gives

$$\tilde{S}_{\mathrm{fl}}^\alpha \Rightarrow \mu^{-1} c_s B_s \circ (\mu e). \tag{76}$$

Prelimit equation in reflected integral form. From the identity $Z^\alpha(t) = Z^\alpha(0) + Y^\alpha(t) - t + L^\alpha(t)$ and the scalings (23b)–(23c), we obtain for $t \in [0,T]$:

$$\tilde{Z}^\alpha(t) = \tilde{Z}^\alpha(0) + \tilde{Y}^\alpha(t) + c^\alpha t + \tilde{L}^\alpha(t), \qquad c^\alpha = \alpha^{-h}(\rho^\alpha - 1). \tag{77}$$

Using $\tilde{A}_0^\alpha = \tilde{A}^\alpha - \tilde{G}^\alpha$ and (73), we can rewrite (77) as

$$\tilde{Z}^\alpha(t) = \tilde{Z}^\alpha(0) + \frac{1}{\mu^\alpha}\tilde{A}^\alpha(t) + \tilde{S}_{\mathrm{fl}}^\alpha(t) + c^\alpha t - \frac{1}{\mu^\alpha}\tilde{G}^\alpha(t) + \tilde{L}^\alpha(t). \tag{78}$$

Define the centered abandonment error

$$\tilde{G}_c^\alpha(t) \triangleq \frac{1}{\mu^\alpha}\tilde{G}^\alpha(t) - \beta\int_0^t (\tilde{Z}^\alpha(s))^k ds. \tag{79}$$

By (71) and $\lambda^\alpha/\mu^\alpha = \rho^\alpha \to 1$,

$$\sup_{0\le t\le T} |\tilde{G}_c^\alpha(t)| \Rightarrow 0. \tag{80}$$

Substituting (79) into (78) yields

$$\tilde{Z}^\alpha(t) = \underbrace{\left(\tilde{Z}^\alpha(0) + \mu^{-\alpha}\tilde{A}^\alpha(t) + \tilde{S}_{\mathrm{fl}}^\alpha(t) + c^\alpha t - \tilde{G}_c^\alpha(t)\right)}_{=:\ \tilde{y}^\alpha(t)} + \int_0^t (-\beta(\tilde{Z}^\alpha(s))^k) ds + \tilde{L}^\alpha(t).$$

Thus $(\tilde{Z}^\alpha, \tilde{L}^\alpha)$ solves (62) with input $\tilde{y}^\alpha$ and drift $g(x) = -\beta x^k$.

Convergence of the driving term. By (24), (76), $c^\alpha \to c$, $\mu^\alpha \to \mu$, and (80), we have

$$\tilde{y}^\alpha \Rightarrow y \quad \text{in } \mathbb{D}([0,T],\mathbb{R}), \qquad y(t) \triangleq Z^*(0) + \mu^{-1}c_a B_a(\mu t) + \mu^{-1}c_s B_s(\mu t) + ct.$$

The limit y has continuous sample paths. Therefore, $\tilde{y}^\alpha \Rightarrow y$ in J_1 implies uniform convergence on $[0,T]$. Lemma 9 then implies

$$(\tilde{Z}^\alpha, \tilde{L}^\alpha) \Rightarrow (Z^*, L^*) \qquad \text{in } \mathbb{D}([0,T], \mathbb{R}^2),$$

where (Z^*, L^*) is the unique solution on $[0,T]$ to

$$Z^*(t) = y(t) - \beta \int_0^t (Z^*(s))^k ds + L^*(t), \quad Z^*(t) \geq 0, L^* \text{ nondecreasing}, \int_0^T \mathbb{1}\{Z^*(t) > 0\} dL^*(t) = 0.$$

Expanding y gives (25) on $[0,T]$. Since $T > 0$ is arbitrary, this establishes $(\tilde{Z}^\alpha, \tilde{L}^\alpha) \Rightarrow (Z^*, L^*)$ in $\mathbb{D}(\mathbb{R}_+, \mathbb{R}^2)$.

Limits for $\tilde{A}_0^\alpha$ and $\tilde{Y}^\alpha$. From $A_0^\alpha = A^\alpha - G^\alpha$ we have

$$\tilde{A}_0^\alpha(t) = \tilde{A}^\alpha(t) - \tilde{G}^\alpha(t).$$

By (24) and (72) with $\tilde{Z}^\alpha \Rightarrow Z^*$,

$$\tilde{A}_0^\alpha \Rightarrow A_0^*(t) \triangleq c_a B_a(\mu t) - \mu\beta \int_0^t (Z^*(s))^k ds.$$

Finally, by (73), $\mu^\alpha \to \mu$, the convergence of $\tilde{A}_0^\alpha$, and (76),

$$\tilde{Y}^\alpha \Rightarrow Y^*(t) \triangleq \frac{1}{\mu} A_0^*(t) + \frac{1}{\mu} c_s B_s(\mu t),$$

and expanding A_0^* gives the result.

The joint convergence of $(\tilde{Z}^\alpha, \tilde{L}^\alpha, \tilde{Y}^\alpha, \tilde{A}_0^\alpha)$ follows because each component is a continuous functional on $[0,T]$ of the jointly convergent primitives and the solution mapping in Lemma 9.

C.7 Proof of Proposition 2

C.7.1 A Technical Lemma on the Malliavin Derivative of the Reflected Diffusion

Let D denote the Malliavin derivative with respect to B, and $\mathbb{D}^{1,2}$ the Sobolev–Watanabe space on Wiener space.

Lemma 10 (Malliavin derivative of the reflected heavy-traffic diffusion)**.** *Fix $t > 0$ and let B be a one-dimensional standard Brownian motion. Consider the one-dimensional reflected diffusion Z^* on $\mathbb{R}_+$ defined by the Skorokhod equation*

$$Z^*(u) = Z^*(0) + \sqrt{2}B(u) + \int_0^u (c - g(Z^*(r)))dr + L^*(u), \qquad 0 \leq u \leq t, \tag{81}$$

where L^* is nondecreasing, $L^*(0) = 0$, $Z^*(u) \geq 0$ and $\int_0^t \mathbb{1}\{Z^*(u) > 0\} dL^*(u) = 0$. Assume $g(x) = x^k/k!$, so that $g'(x) = x^{k-1}/(k-1)! \geq 0$. Then, for every $u \in [0,t]$, one has $Z^*(u) \in \mathbb{D}^{1,2}$ and, for all $0 \leq s \leq u \leq t$,

$$D_s Z^*(u) = \sqrt{2} \exp\left(-\int_s^u g'(Z^*(r))dr\right) \mathbb{1}\left\{\inf_{r \in [s,u]} Z^*(r) > 0\right\}. \tag{82}$$

In particular,

$$0 \le D_s Z^*(u) \le \sqrt{2} \exp\left(- \int_s^u g'(Z^*(r)) dr \right), \qquad 0 \le s \le u \le t. \tag{83}$$

Moreover, the effective input functional

$$Y^*(t) \equiv Z^*(t) - Z^*(0) - ct - L^*(t) = \sqrt{2} B(t) - \int_0^t g(Z^*(r)) dr \tag{84}$$

belongs to $\mathbb{D}^{1,2}$*, with Malliavin derivative given by*

$$D_s Y^*(t) = \sqrt{2} \mathbb{1}\{s \le t\} - \int_s^t g'(Z^*(r)) D_s Z^*(r) dr, \qquad 0 \le s \le t, \tag{85}$$

and therefore $|D_s Y^*(t)| \le \sqrt{2}$ *almost surely for every* $s \in [0, t]$*.*

Proof. Lépingle et al. [16, Proposition 2.7] prove Malliavin differentiability for one-dimensional reflected diffusions and an explicit formula for the Malliavin derivative in terms of a multiplicative functional that vanishes when the path touches the boundary. Bossy et al. [4, Lemma 3.7] restate the [16] representation in a form convenient for calculations. In particular, for $s < u$,

$$D_s Z^*(u) = \sigma(Z^*(s)) \frac{J_u}{J_s} \mathbb{1}_{E_{s,u}}, \tag{86}$$

where $\sigma \equiv \sqrt{2}$ in (81), J is the multiplicative functional associated with the linearization of the flow, and $E_{s,u}$ is the event that the path does not hit the boundary on $[s, u]$, which, for reflection at 0, can be taken as $\{\inf_{r \in [s,u]} Z^*(r) > 0\}$.

By Bossy et al. [4, Proposition 2.8], one can choose J so that

$$\frac{J_u}{J_s} = \exp\left(\int_s^u \sigma'(Z^*(r)) dB(r) + \int_s^u \Big(b'(Z^*(r)) - \tfrac{1}{2} (\sigma'(Z^*(r)))^2 \Big) dr \right),$$

where $b(x) = c - g(x)$ is the drift. Here $\sigma'(x) = 0$ (since $\sigma \equiv \sqrt{2}$), hence

$$\frac{J_u}{J_s} = \exp\left(\int_s^u b'(Z^*(r)) dr \right) = \exp\left(- \int_s^u g'(Z^*(r)) dr \right).$$

Plugging this into (86) gives (82), and the bound (83) follows immediately by dropping the indicator.

By definition of $\mathbb{D}^{1,2}$ and the chain/closure properties of D (e.g., 17, Section 2.3), the identity (84) implies that $Y^*(t) \in \mathbb{D}^{1,2}$ as soon as $\int_0^t g(Z^*(r)) dr$ is in $\mathbb{D}^{1,2}$, with derivative obtained by differentiating under the integral. Using the chain rule,

$$D_s \left(\int_0^t g(Z^*(r)) dr \right) = \int_s^t g'(Z^*(r)) D_s Z^*(r) dr,$$

so (85) holds. Finally, using (83) and $g' \ge 0$,

$$\begin{aligned} 0 \le \int_s^t g'(Z^*(r)) D_s Z^*(r) dr &\le \sqrt{2} \int_s^t g'(Z^*(r)) \exp\left(- \int_s^r g'(Z^*(\ell)) d\ell \right) dr \\ &= \sqrt{2} \Big(1 - e^{- \int_s^t g'(Z^*(\ell)) d\ell} \Big) \le \sqrt{2}, \end{aligned}$$

which yields $|D_s Y^*(t)| \le \sqrt{2}$. This uniform bound implies $\mathbb{E} \int_0^t |D_s Y^*(t)|^2 ds \le 2t < \infty$. □

C.7.2 Proof of Proposition 2

Proof of the bounds. Fix $t > 0$ and recall from (28) that

$$w_{c,k}(t) = \frac{v_{c,k}(t)}{2t}, \qquad \text{where} \qquad v_{c,k}(t) \equiv \mathrm{Var}\big(Y^*(t)\big).$$

By Lemma 10, $Y^*(t) \in \mathbb{D}^{1,2}$ and its Malliavin derivative satisfies

$$|D_s Y^*(t)| \le \sqrt{2} \qquad \text{for all } s \in [0,t] \quad \text{a.s.}$$

In particular, $(D_s Y^*(t))^2 \le 2$ a.s. for every $s \in [0,t]$. Applying the Wiener-space Poincaré inequality (see, e.g., 17, Exercise 2.11.1) gives

$$\mathrm{Var}\big(Y^*(t)\big) \le \mathbb{E}\left[\int_0^t (D_s Y^*(t))^2 ds\right] \le \int_0^t 2ds = 2t.$$

Therefore, for every $t > 0$,

$$0 \le w_{c,k}(t) = \frac{\mathrm{Var}(Y^*(t))}{2t} \le 1,$$

where nonnegativity is immediate since $v_{c,k}(t)$ is a variance. □

Proof of the small-time limit. We now prove the small-time limit $\lim_{t\downarrow 0} w_{c,k}(t) = 1$. Define the process

$$\Gamma^*(t) \triangleq \int_0^t g(Z^*(u))du,$$

so that $Y^*(t) = \sqrt{2}B(t) - \Gamma^*(t)$. Under the stationary initialization, $Z^*(0)$ has exponentially decaying tails as in (26), so $\mathbb{E}[g(Z^*(0))^2] < \infty$. By Cauchy–Schwarz,

$$\big(\Gamma^*(t)\big)^2 = \left(\int_0^t g(Z^*(u))du\right)^2 \le t\int_0^t g(Z^*(u))^2 du,$$

and taking expectations plus stationarity yields

$$\mathbb{E}\left[\big(\Gamma^*(t)\big)^2\right] \le t\int_0^t \mathbb{E}\left[g(Z^*(u))^2\right] du = t^2\mathbb{E}\left[g(Z^*(0))^2\right] = O(t^2) \qquad \text{as } t \downarrow 0.$$

Hence $\mathrm{Var}(\Gamma^*(t)) = O(t^2)$. Moreover, by Cauchy–Schwarz again,

$$|\mathrm{Cov}(\sqrt{2}B(t), \Gamma^*(t))| \le \sqrt{\mathrm{Var}(\sqrt{2}B(t))\mathrm{Var}(\Gamma^*(t))} = \sqrt{(2t)O(t^2)} = O(t^{3/2}).$$

Therefore,

$$\begin{aligned}\mathrm{Var}(Y^*(t)) &= \mathrm{Var}(\sqrt{2}B(t)) + \mathrm{Var}(\Gamma^*(t)) - 2\mathrm{Cov}(\sqrt{2}B(t), \Gamma^*(t)) \\ &= 2t + O(t^2) + O(t^{3/2}) = 2t + o(t), \qquad t \downarrow 0,\end{aligned}$$

and consequently

$$\lim_{t\downarrow 0} w_{c,k}(t) = \lim_{t\downarrow 0} \frac{\mathrm{Var}(Y^*(t))}{2t} = 1.$$

We may thus define $w_{c,k}(0) \triangleq 1$. □

Proof of strict monotonicity. Let $\mathbb{F} = \{\mathcal{F}_t\}$ be the filtration of B. By Lemma 10, $Y^*(t) \in \mathbb{D}^{1,2}$ and

$$D_s Y^*(t) = \sqrt{2} - \int_s^t g'(Z^*(r)) D_s Z^*(r) dr, \qquad 0 \le s \le t,$$

where $D_s Z^*(r) \ge 0$ and $g'(Z^*(r)) \ge 0$ a.s. Hence for each fixed s, $t \mapsto D_s Y^*(t)$ is a.s. nonincreasing on $[s, \infty)$.

By the Clark–Ocone formula [18, Proposition 1.3.14],

$$Y^*(t) - \mathbb{E}[Y^*(t)] = \int_0^t \phi_s^{(t)} dB(s), \qquad \phi_s^{(t)} \triangleq \mathbb{E}[D_s Y^*(t) \mid \mathcal{F}_s],$$

and therefore by Itô isometry,

$$v_{c,k}(t) = \mathbb{E}\left[\int_0^t (\phi_s^{(t)})^2 ds\right].$$

Time-homogeneity of Z^* implies there exists a deterministic measurable function $\psi : [0, \infty) \times \mathbb{R}_+ \to \mathbb{R}_+$ such that

$$\phi_s^{(t)} = \psi(t - s, Z^*(s)).$$

Under stationary initialization, $Z^*(s) \overset{d}{=} Z^*(0)$, hence

$$v_{c,k}(t) = \int_0^t m(u) du, \qquad m(u) \triangleq \mathbb{E}\left[\psi(u, Z^*(0))^2\right].$$

Consequently,

$$w_{c,k}(t) = \frac{1}{2t} \int_0^t m(u) du.$$

To show that $w_{c,k}$ is strictly decreasing, it suffices to show that m is strictly decreasing on $(0, \infty)$.

Fix $z > 0$ and $0 < u_1 < u_2$. Let $\tau_0 \triangleq \inf\{t > 0 : Z^*(t) = 0\}$ be the first boundary hitting time. From the representation in Lemma 10, on the event $\{\tau_0 > u_1\}$ one has $D_0 Y^*(u_2) < D_0 Y^*(u_1)$ a.s. because the integrand $g'(Z^*(r)) D_0 Z^*(r)$ is strictly positive for $r \in (0, u_1]$ on that event (using $g'(x) > 0$ for $x > 0$ and $Z^*(r) > 0$ for $r \le u_1$). Moreover, for each $z > 0$ and each finite $u_1 > 0$,

$$\mathbb{P}_z(\tau_0 > u_1) > 0,$$

since the diffusion has continuous paths and nondegenerate noise. Therefore,

$$\mathbb{E}_z[D_0 Y^*(u_2)] < \mathbb{E}_z[D_0 Y^*(u_1)].$$

Since $D_0 Y^*(u) \in (0, \sqrt{2}]$, the same strict inequality holds after squaring:

$$\psi(u_2, z)^2 < \psi(u_1, z)^2.$$

Finally, the stationary density (26) has no atom at 0, so $\mathbb{P}(Z^*(0) > 0) = 1$ and taking expectation over $Z^*(0)$ yields $m(u_2) < m(u_1)$. Thus m is strictly decreasing, completing the proof. □

Proof of the monotonicity and extreme limits of $w_{c,k}(\infty)$. **A representation of $w_{c,k}(\infty)$.** Recall the reflected base diffusion (29) and define $q_k(x) = kx^{k-1}$. For $z \geq 0$ let $\mathbb{P}_{c,z}$ denote the law of $Z^{c,k}$ started from $Z^{c,k}(0) = z$, and let $\tau_0 \triangleq \inf\{t \geq 0 : Z^{c,k}(t) = 0\}$. Define the *infinite-horizon* kernel

$$\psi^{\infty}_{c,k}(z) \triangleq \lim_{t\to\infty} \psi_{c,k}(t,z) = \mathbb{E}_{c,z}\left[\exp\left\{-\int_0^{\tau_0} q_k(Z^{c,k}(s))ds\right\}\right],$$

where the equality holds by monotone convergence since $t \mapsto \int_0^{t\wedge\tau_0} q_k(Z^{c,k}(s))ds$ is nondecreasing and $0 \leq e^{-x} \leq 1$.

Let $\pi_{c,k}$ be the stationary density (30). By Lemma 6, for $t > 0$,

$$w_{c,k}(t) = \frac{1}{t}\int_0^t \mathbb{E}_{\pi_{c,k}}\left[\psi_{c,k}(u,Z)^2\right]du + \frac{1}{2t}\mathrm{Var}_{\pi_{c,k}}\left(h_{c,k}(t,Z)\right), \qquad Z \sim \pi_{c,k}.$$

By Remark 6, the second term above is $O(1/t)$ and therefore vanishes as $t \to \infty$. For each fixed z, the map $u \mapsto \psi_{c,k}(u,z)$ is nonincreasing, hence so is $u \mapsto \psi_{c,k}(u,z)^2$; therefore the map $u \mapsto \mathbb{E}_{\pi_{c,k}}[\psi_{c,k}(u,Z)^2]$ is nonincreasing. Consequently, the Cesàro limit equals the pointwise limit, and dominated convergence yields

$$\begin{aligned} w_{c,k}(\infty) = \lim_{t\to\infty} w_{c,k}(t) &= \lim_{t\to\infty}\frac{1}{t}\int_0^t \mathbb{E}_{\pi_{c,k}}\left[\psi_{c,k}(u,Z)^2\right]du \\ &= \lim_{u\to\infty}\mathbb{E}_{\pi_{c,k}}\left[\psi_{c,k}(u,Z)^2\right] = \mathbb{E}_{\pi_{c,k}}\left[\left(\psi^{\infty}_{c,k}(Z)\right)^2\right]. \end{aligned} \tag{87}$$

Monotonicity of $\psi^{\infty}_{c,k}$ in the initial state and in c. Fix $z_1 \leq z_2$ and $c_1 \leq c_2$. Because the reflection term $L^{c,k}$ is identically 0 before the first hit of 0, the stopped process $Z^{c,k}(t\wedge\tau_0)$ coincides with the solution to the *unreflected* SDE on $(0,\infty)$ stopped upon hitting 0:

$$X^c(t) = z + \sqrt{2}B(t) + \int_0^t (c - (X^c(s))^k)ds, \qquad t < \tau_0^c, \qquad \tau_0^c \triangleq \inf\{t \geq 0 : X^c(t) = 0\}.$$

Couple X^{c_1} and X^{c_2} on the same Brownian path B, with initial conditions $X^{c_i}(0) = z_i$. For $t < \tau_0^{c_1} \wedge \tau_0^{c_2}$,

$$X^{c_1}(t) - X^{c_2}(t) = (z_1 - z_2) + \int_0^t \Big((c_1 - c_2) - \left((X^{c_1}(s))^k - (X^{c_2}(s))^k\right)\Big)ds.$$

Define $\Delta(t) \triangleq X^{c_1}(t) - X^{c_2}(t)$. If $\Delta(s) > 0$, then $X^{c_1}(s) > X^{c_2}(s)$ and hence $(X^{c_1}(s))^k - (X^{c_2}(s))^k > 0$. Therefore, on $\{\Delta(s) > 0\}$,

$$(c_1 - c_2) - \left((X^{c_1}(s))^k - (X^{c_2}(s))^k\right) \leq c_1 - c_2 \leq 0.$$

Since Δ is absolutely continuous, it follows that $\Delta^+(t) \triangleq \max\{\Delta(t), 0\}$ is nonincreasing on $[0, \tau_0^{c_1} \wedge \tau_0^{c_2})$. Because $z_1 \leq z_2$, we have $\Delta^+(0) = 0$, hence $\Delta^+(t) = 0$ and

$$X^{c_1}(t) \leq X^{c_2}(t), \qquad 0 \leq t < \tau_0^{c_1} \wedge \tau_0^{c_2}.$$

In particular, by continuity of sample paths, $\tau_0^{c_1} \leq \tau_0^{c_2}$ almost surely.

Since q_k is nondecreasing on $\mathbb{R}_+$ and $X^{c_2}(t) \geq X^{c_1}(t)$ for $t \leq \tau_0^{c_1}$,

$$\int_0^{\tau_0^{c_2}} q_k(X^{c_2}(s))ds \geq \int_0^{\tau_0^{c_1}} q_k(X^{c_2}(s))ds \geq \int_0^{\tau_0^{c_1}} q_k(X^{c_1}(s))ds.$$

Exponentiating and taking expectations gives

$$\psi_{c_2,k}^{\infty}(z_2) \leq \psi_{c_1,k}^{\infty}(z_1). \tag{88}$$

Thus, for each fixed c, the function $z \mapsto \psi_{c,k}^{\infty}(z)$ is nonincreasing, and for each fixed z, the function $c \mapsto \psi_{c,k}^{\infty}(z)$ is nonincreasing.

Stochastic monotonicity of the stationary distribution in c. Let Z_c denote a random variable with density $\pi_{c,k}$ in (30). For $c_2 > c_1$, the likelihood ratio satisfies

$$\frac{\pi_{c_2,k}(x)}{\pi_{c_1,k}(x)} = \frac{G_{c_1,k}}{G_{c_2,k}} \exp\{(c_2 - c_1)x\}, \qquad x \geq 0,$$

which is strictly increasing in x. Therefore $\pi_{c_2,k}$ dominates $\pi_{c_1,k}$ in the monotone likelihood-ratio order, and in particular in the usual stochastic order:

$$Z_{c_1} \ \leq_{\mathrm{st}} \ Z_{c_2}. \tag{89}$$

$c \mapsto w_{c,k}(\infty)$ is strictly decreasing. Fix $c_2 > c_1$ and define $f_c(z) \triangleq (\psi_{c,k}^{\infty}(z))^2$. By (88), for each z we have $f_{c_2}(z) \leq f_{c_1}(z)$, and for each c the map $z \mapsto f_c(z)$ is nonincreasing. Using (87),

$$w_{c_2,k}(\infty) = \mathbb{E}_{\pi_{c_2,k}}[f_{c_2}(Z)] \leq \mathbb{E}_{\pi_{c_2,k}}[f_{c_1}(Z)] \leq \mathbb{E}_{\pi_{c_1,k}}[f_{c_1}(Z)] = w_{c_1,k}(\infty),$$

where the second inequality uses the stochastic dominance (89) and the fact that f_{c_1} is nonincreasing. This proves that $c \mapsto w_{c,k}(\infty)$ is nonincreasing. Strict monotonicity follows because (i) $\pi_{c,k}$ has a continuous density supported on $(0, \infty)$, and (ii) $z \mapsto \psi_{c,k}^{\infty}(z)$ is nonconstant and nonincreasing, so the above inequalities are strict when $c_2 > c_1$.

The limit $\lim_{c \to -\infty} w_{c,k}(\infty) = 1$. Fix $\delta > 0$ and consider $Z_c \sim \pi_{c,k}$. We first show $Z_c \Rightarrow 0$ as $c \to -\infty$. Write

$$\mathbb{P}(Z_c > \delta) = \frac{\int_\delta^\infty \exp\{cx - \frac{1}{k+1}x^{k+1}\}dx}{\int_0^\infty \exp\{cx - \frac{1}{k+1}x^{k+1}\}dx}.$$

For $c < 0$ and $x \geq \delta$, we have $cx \leq c\delta$, hence

$$\int_\delta^\infty \exp\left\{cx - \frac{1}{k+1}x^{k+1}\right\}dx \leq e^{c\delta} \int_\delta^\infty \exp\left\{-\frac{1}{k+1}x^{k+1}\right\}dx \triangleq C_\delta e^{c\delta},$$

where $C_\delta < \infty$. On the other hand,

$$\int_0^\infty \exp\left\{cx - \frac{1}{k+1}x^{k+1}\right\}dx \geq \int_0^\delta \exp\left\{cx - \frac{1}{k+1}\delta^{k+1}\right\}dx = e^{-\delta^{k+1}/(k+1)} \frac{1 - e^{c\delta}}{-c}.$$

Combining these bounds yields

$$\mathbb{P}(Z_c > \delta) \leq C_\delta e^{\delta^{k+1}/(k+1)}(-c)e^{c\delta} \xrightarrow[c \to -\infty]{} 0.$$

Thus $Z_c \Rightarrow 0$.

Next, we show that $\psi^{\infty}_{c,k}(z) \to 1$ uniformly for $z \in [0,\delta]$ as $c \to -\infty$. Fix $z \in [0,\delta]$ and consider the (unreflected) SDE X^c started from z and stopped at τ_0^c. Let $t_c \triangleq |c|^{-1/2}$ and define the event

$$E_c \triangleq \left\{ \sup_{0\le s\le t_c} |B(s)| \le \frac{\delta}{\sqrt{2}} \right\}.$$

Since $t_c \downarrow 0$, standard Brownian maximal inequalities imply $\mathbb{P}(E_c) \to 1$ as $c \to -\infty$.

On E_c, for all $0 \le t \le t_c$,

$$X^c(t) = z + \sqrt{2}B(t) + \int_0^t \big(c - (X^c(s))^k\big)ds \le z + \sqrt{2}|B(t)| + ct \le \delta + \delta + 0 = 2\delta,$$

since $c < 0$ and $ct \le 0$. Moreover, at time t_c we have

$$X^c(t_c) \le z + \sqrt{2}|B(t_c)| + ct_c \le 2\delta - |c|^{1/2},$$

which is strictly negative for all sufficiently large $|c|$. By continuity, X^c must hit 0 before time t_c, so $\tau_0^c \le t_c$ on E_c for all sufficiently negative c.

Therefore, on E_c and for large negative c,

$$\int_0^{\tau_0^c} q_k(X^c(s))ds \le q_k(2\delta)\tau_0^c \le q_k(2\delta)t_c,$$

and hence

$$\exp\left\{ -\int_0^{\tau_0^c} q_k(X^c(s))ds \right\} \ge \exp\{-q_k(2\delta)t_c\}.$$

Taking expectations yields the uniform bound

$$\inf_{0\le z\le\delta} \psi^{\infty}_{c,k}(z) \ge \exp\{-q_k(2\delta)t_c\}\mathbb{P}(E_c) \xrightarrow[c\to-\infty]{} 1,$$

since $t_c \to 0$ and $\mathbb{P}(E_c) \to 1$.

Finally, combine this with (87) and the fact that $0 \le \psi^{\infty}_{c,k} \le 1$:

$$w_{c,k}(\infty) = \mathbb{E}_{\pi_{c,k}}\left[\left(\psi^{\infty}_{c,k}(Z)\right)^2 \right] \ge \left(\inf_{0\le z\le\delta} \psi^{\infty}_{c,k}(z) \right)^2 \mathbb{P}(Z_c \le \delta) \xrightarrow[c\to-\infty]{} 1.$$

Since always $w_{c,k}(\infty) \le 1$, we conclude $\lim_{c\to-\infty} w_{c,k}(\infty) = 1$.

The limit $\lim_{c\to\infty} w_{c,k}(\infty) = 0$**.** We first obtain a uniform upper bound on $\psi^{\infty}_{c,k}$ for large c via a supermartingale argument. For $c \ge k+1$, define $f(x) \triangleq (1+x)^{-k}$, $x \ge 0$. A direct calculation shows that for $x \ge 0$,

$$(\mathcal{L}_c f)(x) - q_k(x)f(x) = f''(x) + (c - x^k)f'(x) - kx^{k-1}f(x) \tag{90}$$

$$= k(1+x)^{-k-2}\Big((k+1) - (1+x)(c + x^{k-1})\Big) \le 0, \tag{91}$$

where $\mathcal{L}_c f = f'' + (c - x^k)f'$ is the interior generator. Let X^c be the (unreflected) SDE started from z and stopped at τ_0^c as above. Itô's formula applied to $e^{-\int_0^{t\wedge\tau_0^c} q_k(X^c(s))ds} f(X^c(t\wedge\tau_0^c))$ together with (90) implies that

$$\mathbb{E}_{c,z}\left[e^{-\int_0^{t\wedge\tau_0^c} q_k(X^c(s))ds} f(X^c(t\wedge\tau_0^c)) \right] \le f(z), \qquad t \ge 0.$$

Letting $t \to \infty$ and using $X^c(\tau_0^c) = 0$ and $f(0) = 1$ yields

$$\psi_{c,k}^{\infty}(z) = \mathbb{E}_{c,z}\left[e^{-\int_0^{\tau_0^c} q_k(X^c(s))ds}\right] \le f(z) = (1+z)^{-k}, \qquad c \ge k+1.$$

Therefore, for $c \ge k+1$,

$$w_{c,k}(\infty) = \mathbb{E}_{\pi_{c,k}}\left[\left(\psi_{c,k}^{\infty}(Z)\right)^2\right] \le \mathbb{E}_{\pi_{c,k}}\left[(1+Z)^{-2k}\right]. \tag{92}$$

It remains to show that $Z_c \sim \pi_{c,k}$ diverges to $+\infty$ in probability as $c \to \infty$. Fix $M > 0$. Then

$$\mathbb{P}(Z_c \le M) = \frac{\int_0^M \exp\{cx - \frac{1}{k+1}x^{k+1}\}dx}{\int_0^{\infty} \exp\{cx - \frac{1}{k+1}x^{k+1}\}dx}.$$

For the numerator, it is bounded from above by Me^{cM}. For the denominator, restrict to $[M, M+1]$ and use the bound $-(k+1)^{-1}x^{k+1} \ge -(k+1)^{-1}(M+1)^{k+1}$:

$$\int_0^{\infty} \exp\left\{cx - \frac{1}{k+1}x^{k+1}\right\}dx \ge e^{-(M+1)^{k+1}/(k+1)} \int_M^{M+1} e^{cx}dx = e^{-(M+1)^{k+1}/(k+1)} \frac{e^{c(M+1)} - e^{cM}}{c}.$$

Hence

$$\mathbb{P}(Z_c \le M) \le Me^{cM} \cdot \frac{ce^{(M+1)^{k+1}/(k+1)}}{e^{c(M+1)} - e^{cM}} = Mce^{(M+1)^{k+1}/(k+1)} \cdot \frac{1}{e^c - 1} \xrightarrow[c\to\infty]{} 0.$$

Thus $Z_c \to \infty$ in probability as $c \to \infty$, and therefore $(1+Z_c)^{-2k} \to 0$ in probability while being bounded by 1. Consequently,

$$\mathbb{E}_{\pi_{c,k}}\left[(1+Z)^{-2k}\right] \xrightarrow[c\to\infty]{} 0.$$

Combining with (92) yields $\lim_{c\to\infty} w_{c,k}(\infty) = 0$. □

C.8 Proof of Lemma 5

Throughout the proof, set $\sigma^2 \triangleq \frac{c_x^2}{\mu}$. From Theorem 3, the stationary reflected diffusion Z^* satisfies

$$Z^*(t) = Z^*(0) + \mu^{-1}c_a B_a(\mu t) + \mu^{-1}c_s B_s(\mu t) - \beta\int_0^t (Z^*(s))^k ds + ct + L^*(t), \qquad t \ge 0. \tag{93}$$

Define rescaled Brownian motions

$$\widehat{B}_a(t) \triangleq \mu^{-1/2}B_a(\mu t), \qquad \widehat{B}_s(t) \triangleq \mu^{-1/2}B_s(\mu t),$$

which are independent standard Brownian motions. Let

$$B(t) \triangleq \frac{c_a}{c_x}\widehat{B}_a(t) + \frac{c_s}{c_x}\widehat{B}_s(t),$$

so B is a standard Brownian motion and

$$\mu^{-1}c_a B_a(\mu t) + \mu^{-1}c_s B_s(\mu t) = \frac{c_a}{\sqrt{\mu}}\widehat{B}_a(t) + \frac{c_s}{\sqrt{\mu}}\widehat{B}_s(t) = \sqrt{\frac{c_x^2}{\mu}}B(t) = \sigma B(t).$$

Hence (93) is equivalently written as

$$Z^*(t) = Z^*(0) + \sigma B(t) - \beta \int_0^t (Z^*(s))^k ds + ct + L^*(t), \qquad t \geq 0. \tag{94}$$

Recall also that

$$Y^*(t) = Z^*(t) - Z^*(0) - ct - L^*(t), \qquad t \geq 0, \tag{95}$$

so $v(t;\cdot) = \mathrm{Var}(Y^*(t))$ by definition.

Define the positive scaling constants

$$\theta \triangleq \left(\frac{\sigma^2}{2\beta}\right)^{\frac{1}{k+1}}, \qquad \tau \triangleq \beta\theta^{k-1}. \tag{96}$$

Thus,

$$\tilde{c} = \frac{c}{\beta\theta^k}. \tag{97}$$

For $u \geq 0$, define the scaled processes

$$\bar{Z}(u) \triangleq \frac{1}{\theta} Z^*\left(\frac{u}{\tau}\right), \qquad \bar{L}(u) \triangleq \frac{1}{\theta} L^*\left(\frac{u}{\tau}\right), \qquad \bar{B}(u) \triangleq \sqrt{\tau} B\left(\frac{u}{\tau}\right). \tag{98}$$

By Brownian scaling, $\bar{B}$ is a standard Brownian motion. Substitute $t = u/\tau$ and $Z^*(u/\tau) = \theta\bar{Z}(u)$ into (94):

$$\theta\bar{Z}(u) = \theta\bar{Z}(0) + \sigma B\left(\frac{u}{\tau}\right) - \beta \int_0^{u/\tau} \left(Z^*(s)\right)^k ds + c\frac{u}{\tau} + \theta\bar{L}(u).$$

For the integral term, change variables $r = \tau s$ to obtain

$$\int_0^{u/\tau} \left(Z^*(s)\right)^k ds = \int_0^{u/\tau} \left(\theta\bar{Z}(\tau s)\right)^k ds = \theta^k \tau^{-1} \int_0^u \left(\bar{Z}(r)\right)^k dr.$$

Also, $B(u/\tau) = \tau^{-1/2}\bar{B}(u)$. Dividing by θ yields

$$\bar{Z}(u) = \bar{Z}(0) + \frac{\sigma}{\theta\sqrt{\tau}}\bar{B}(u) - \frac{\beta\theta^{k-1}}{\tau} \int_0^u \left(\bar{Z}(r)\right)^k dr + \frac{c}{\theta\tau}u + \bar{L}(u). \tag{99}$$

By the choice $\tau = \beta\theta^{k-1}$ in (96), the integral coefficient satisfies

$$\frac{\beta\theta^{k-1}}{\tau} = 1.$$

Next, using $\tau = \beta\theta^{k-1}$ again,

$$\frac{\sigma}{\theta\sqrt{\tau}} = \frac{\sigma}{\theta\sqrt{\beta\theta^{k-1}}} = \frac{\sigma}{\sqrt{\beta\theta^{k+1}}}.$$

By the definition of θ in (96), we have $\theta^{k+1} = \sigma^2/(2\beta)$, so $\beta\theta^{k+1} = \sigma^2/2$ and hence

$$\frac{\sigma}{\theta\sqrt{\tau}} = \sqrt{2}.$$

Finally, by (97), (99) simplifies to

$$\bar{Z}(u) = \bar{Z}(0) + \sqrt{2}\bar{B}(u) - \int_0^u \big(\bar{Z}(r)\big)^k dr + \tilde{c}u + \bar{L}(u), \qquad u \geq 0, \tag{100}$$

together with the reflection conditions $\bar{Z}(u) \geq 0$, $\bar{L}$ nondecreasing, $\bar{L}(0) = 0$, and $\int_0^\infty \mathbb{1}\{\bar{Z}(u) > 0\} d\bar{L}(u) = 0$. Thus $(\bar{Z}, \bar{L})$ has the same law as the stationary reflected diffusion corresponding to the *base* parameter-tuple $(\tilde{c}, k, 1, 1, 1, k!)$.

Define the corresponding base effective-input process

$$\bar{Y}(u) \triangleq \bar{Z}(u) - \bar{Z}(0) - \tilde{c}u - \bar{L}(u), \qquad u \geq 0.$$

By (100),

$$\bar{Y}(u) = \sqrt{2}\bar{B}(u) - \int_0^u (\bar{Z}(r))^k dr,$$

so by definition of $v_{\tilde{c},k}$ in (27),

$$\mathrm{Var}(\bar{Y}(u)) = v_{\tilde{c},k}(u), \qquad u \geq 0,$$

and hence $\mathrm{Var}(\bar{Y}(u)) = 2uw_{\tilde{c},k}(u)$ by (28).

Next, using (98) and the identity (95), for each $t \geq 0$,

$$\begin{aligned} Y^*(t) &= Z^*(t) - Z^*(0) - ct - L^*(t) \\ &= \theta\bar{Z}(\tau t) - \theta\bar{Z}(0) - ct - \theta\bar{L}(\tau t) \\ &= \theta\Big(\bar{Z}(\tau t) - \bar{Z}(0) - \tilde{c}(\tau t) - \bar{L}(\tau t)\Big), \end{aligned}$$

where we used $\tilde{c}\tau = (c/(\beta\theta^k))(\beta\theta^{k-1}) = c/\theta$. Therefore,

$$Y^*(t) = \theta\bar{Y}(\tau t), \qquad t \geq 0.$$

Taking variances and using $\mathrm{Var}(\bar{Y}(u)) = 2uw_{\tilde{c},k}(u)$ gives

$$v(t; c, k, \mu, c_a^2, c_s^2, F^{(k)}(0)) = \mathrm{Var}(Y^*(t)) = \theta^2 \mathrm{Var}(\bar{Y}(\tau t)) = \theta^2 v_{\tilde{c},k}(\tau t) = 2\theta^2 \tau t w_{\tilde{c},k}(\tau t).$$

The result follows by substituting the definitions in (96).

C.9 Proof of Lemma 6

Let $\mathbb{F} = \{\mathcal{F}_t\}_{t\geq 0}$ be the natural filtration of B. By Lemma 10 applied to the polynomial drift $g(x) = x^k$, the functional $Y^{c,k}(t)$ in (31) belongs to $\mathbb{D}^{1,2}$, and for $0 \leq s \leq t$ its Malliavin derivative satisfies

$$D_s Y^{c,k}(t) = \sqrt{2} - \int_s^t q_k(Z^{c,k}(r)) D_s Z^{c,k}(r) dr. \tag{101}$$

Moreover, the explicit representation of $D_s Z^{c,k}$ in Lemma 10 implies $D_s Z^{c,k}(r) = 0$ for $r \geq \tau_s$, where $\tau_s \triangleq \inf\{u \geq s : Z^{c,k}(u) = 0\}$. Therefore the integral in (101) reduces to $[s, t \wedge \tau_s]$, and a differentiation of $r \mapsto \exp\{-\int_s^r q_k(Z^{c,k}(\ell))d\ell\}$ yields

$$D_s Y^{c,k}(t) = \sqrt{2} \exp\left\{-\int_s^{t\wedge\tau_s} q_k\big(Z^{c,k}(r)\big) dr\right\}, \qquad 0 \leq s \leq t. \tag{102}$$

Since $Z^{c,k}(0)$ is initialized from the stationary distribution $\pi_{c,k}$, we decompose the variance by the law of total variance:

$$\mathrm{Var}\big(Y^{c,k}(t)\big) = \mathbb{E}_{\pi_{c,k}}\Big[\mathrm{Var}\big(Y^{c,k}(t) \mid Z^{c,k}(0)\big)\Big] + \mathrm{Var}_{\pi_{c,k}}\Big(\mathbb{E}\big[Y^{c,k}(t) \mid Z^{c,k}(0)\big]\Big). \tag{103}$$

We first characterize the conditional-variance term. Fix $z \geq 0$ and work under $\mathbb{P}_z$ (so that $Z^{c,k}(0) = z$ is deterministic and $Z^{c,k}$ is adapted to $\mathbb{F}$). Then $Y^{c,k}(t) \in \mathbb{D}^{1,2}$ and the Clark–Ocone formula [18, Proposition 1.3.14] gives

$$Y^{c,k}(t) - \mathbb{E}_z\Big[Y^{c,k}(t)\Big] = \int_0^t \mathbb{E}_z\Big[D_s Y^{c,k}(t) \mid \mathcal{F}_s\Big]\, dB(s),$$

and hence, by Itô isometry,

$$\mathrm{Var}_z\big(Y^{c,k}(t)\big) = \mathbb{E}_z\left[\int_0^t \Big(\mathbb{E}_z[D_s Y^{c,k}(t) \mid \mathcal{F}_s]\Big)^2 ds\right]. \tag{104}$$

By (102) and the strong Markov property of $Z^{c,k}$, conditioning on $\mathcal{F}_s$ is equivalent to conditioning on $Z^{c,k}(s)$ under $\mathbb{P}_z$. More precisely, for each $0 \leq s \leq t$,

$$\mathbb{E}_z\Big[D_s Y^{c,k}(t) \mid \mathcal{F}_s\Big] = \sqrt{2}\mathbb{E}\left[\exp\left\{-\int_s^{t\wedge\tau_s} q_k\big(Z^{c,k}(r)\big)dr\right\} \,\Big|\, \mathcal{F}_s\right] = \sqrt{2}\psi_{c,k}\big(t-s, Z^{c,k}(s)\big),$$

where $\psi_{c,k}$ is defined in (32). Substituting into (104) yields

$$\mathrm{Var}_z\big(Y^{c,k}(t)\big) = 2\mathbb{E}_z\left[\int_0^t \psi_{c,k}\big(t-s, Z^{c,k}(s)\big)^2 ds\right].$$

Integrating over $z \sim \pi_{c,k}$ and using stationarity of $Z^{c,k}$ under $\pi_{c,k}$ gives ($u = t - s$)

$$\mathbb{E}_{\pi_{c,k}}\Big[\mathrm{Var}\big(Y^{c,k}(t) \mid Z^{c,k}(0)\big)\Big] = 2\int_0^t \mathbb{E}_{\pi_{c,k}}\Big[\psi_{c,k}\big(t-s, Z\big)^2\Big]\, ds = 2\int_0^t \mathbb{E}_{\pi_{c,k}}\Big[\psi_{c,k}\big(u, Z\big)^2\Big]\, du,$$

For the second term in (103), note from (31) that, for each $z \geq 0$,

$$\mathbb{E}\big[Y^{c,k}(t) \mid Z^{c,k}(0) = z\big] = -\mathbb{E}_z\left[\int_0^t \big(Z^{c,k}(s)\big)^k ds\right] = -h_{c,k}(t,z),$$

where $h_{c,k}$ is defined in (33). Therefore,

$$\mathrm{Var}_{\pi_{c,k}}\Big(\mathbb{E}\big[Y^{c,k}(t) \mid Z^{c,k}(0)\big]\Big) = \mathrm{Var}_{\pi_{c,k}}\left(h_{c,k}(t,Z)\right).$$

Substituting the two terms into (103) yields

$$\mathrm{Var}\big(Y^{c,k}(t)\big) = 2\int_0^t \mathbb{E}_{\pi_{c,k}}\Big[\psi_{c,k}(u,Z)^2\Big]\, du + \mathrm{Var}_{\pi_{c,k}}\left(h_{c,k}(t,Z)\right).$$

Dividing by $2t$ gives (34).

C.10 Proof of Lemma 7

Proof. Let T_i be the ith arrival epoch, $W_i^\alpha = Z^\alpha(T_i-)$ the offered waiting time, and define the service indicator $I_i^\alpha \triangleq \mathbf{1}\{D_i^\alpha > W_i^\alpha\}$, where $D_i^\alpha \sim F_\alpha$. Then the effective arrival count and effective work input over $[0,t]$ are $A_0^\alpha(t) = \sum_{i=1}^{A(t)} I_i^\alpha$ and $Y^\alpha(t) = \sum_{i=1}^{A(t)} V_i I_i^\alpha$. Condition on $\sigma(A(t), \{I_i^\alpha\}_{i\le A(t)})$. Since $\{V_i\}$ are i.i.d., independent of the arrival process and patience times, we have

$$\mathbb{E}\left[Y^\alpha(t) \mid A(t), \{I_i^\alpha\}\right] = \mathbb{E}[V_1] \sum_{i=1}^{A(t)} I_i^\alpha = \frac{1}{\mu} A_0^\alpha(t),$$

and

$$\mathrm{Var}\left(Y^\alpha(t) \mid A(t), \{I_i^\alpha\}\right) = \mathrm{Var}(V_1) \sum_{i=1}^{A(t)} I_i^\alpha = \frac{c_s^2}{\mu^2} A_0^\alpha(t),$$

because $\mathrm{Var}(V_1) = c_s^2/\mu^2$. Taking expectations and applying the law of total variance yields

$$\mathbb{E}_e[Y^\alpha(t)] = \frac{1}{\mu}\mathbb{E}_e[A_0^\alpha(t)], \qquad \mathrm{Var}_e(Y^\alpha(t)) = \frac{c_s^2}{\mu^2}\mathbb{E}_e[A_0^\alpha(t)] + \frac{1}{\mu^2}\mathrm{Var}_e(A_0^\alpha(t)). \tag{105}$$

Hence

$$\frac{\mathrm{Var}_e(Y^\alpha(t))}{\mathbb{E}_e[Y^\alpha(t)]} = \frac{1}{\mu}\left(c_s^2 + \frac{\mathrm{Var}_e(A_0^\alpha(t))}{\mathbb{E}_e[A_0^\alpha(t)]}\right). \tag{106}$$

It remains to identify the limit of the mean and variance of $A_0^\alpha(t)$.

Let $\mathcal{G}^\alpha \triangleq \sigma\big(A(t), \{W_i^\alpha\}_{i\le A(t)}\big)$. Conditional on $\mathcal{G}^\alpha$, the patience times $\{D_i^\alpha\}$ are independent, so $\{I_i^\alpha\}_{i\le A(t)}$ are independent Bernoulli random variables with success probabilities

$$p_i^\alpha \triangleq \mathbb{P}(D_i^\alpha > W_i^\alpha \mid \mathcal{G}^\alpha) = \bar{F}(\alpha W_i^\alpha).$$

Therefore,

$$\mathbb{E}[A_0^\alpha(t) \mid \mathcal{G}^\alpha] = \sum_{i=1}^{A(t)} p_i^\alpha, \qquad \mathrm{Var}(A_0^\alpha(t) \mid \mathcal{G}^\alpha) = \sum_{i=1}^{A(t)} p_i^\alpha(1-p_i^\alpha).$$

Unconditioning and using the law of total variance gives

$$\mathbb{E}_e[A_0^\alpha(t)] = \mathbb{E}_e\left[\sum_{i=1}^{A(t)} p_i^\alpha\right], \qquad \mathrm{Var}_e(A_0^\alpha(t)) = \mathbb{E}_e\left[\sum_{i=1}^{A(t)} p_i^\alpha(1-p_i^\alpha)\right] + \mathrm{Var}_e\left(\sum_{i=1}^{A(t)} p_i^\alpha\right). \tag{107}$$

We claim that

$$\sup_{0\le u\le t}\left|\alpha Z^\alpha(u) - \alpha Z^\alpha(0)\right| \Rightarrow 0, \qquad \alpha \downarrow 0. \tag{108}$$

Indeed, $Z^\alpha(\cdot)$ decreases at unit rate between arrivals and increases by at most V_i at arrival i (since only served customers contribute positive jumps, and $I_i^\alpha \le 1$). Thus, pathwise,

$$\sup_{0\le u\le t}|Z^\alpha(u) - Z^\alpha(0)| \le t + \sum_{i=1}^{A(t)} V_i.$$

Multiplying by α and taking expectations,

$$\mathbb{E}\left[\sup_{0\le u\le t}|\alpha Z^\alpha(u) - \alpha Z^\alpha(0)|\right] \le \alpha t + \alpha\mathbb{E}\left[\sum_{i=1}^{A(t)} V_i\right] = \alpha t + \alpha\mathbb{E}[A(t)]\mathbb{E}[V_1] = \alpha t + \alpha\lambda t\cdot\frac{1}{\mu} \to 0,$$

so (108) holds.

Since $\bar{F}$ is continuous and bounded, (108) implies

$$\sup_{1\le i\le A(t)} \left|p_i^\alpha - \bar{F}(\alpha Z^\alpha(0))\right| \Rightarrow 0, \qquad \alpha \downarrow 0. \tag{109}$$

By stationarity, Lemma 3 and continuity of $\bar{F}$, we also have $\bar{F}(\alpha Z^\alpha(0)) \Rightarrow \bar{F}(\xi)$. Define $p \triangleq \bar{F}(\xi)$. If $\rho \le 1$, then $\xi = 0$ and hence $p = \bar{F}(0) = 1$; if $\rho > 1$, then by definition of ξ, $F(\xi) = (\rho-1)/\rho$ and thus $p = 1 - F(\xi) = 1/\rho$. Equivalently,

$$p = \frac{1}{\rho \vee 1}. \tag{110}$$

Combining (109) with $\bar{F}(\alpha Z^\alpha(0)) \Rightarrow p$ yields

$$\sup_{1\le i\le A(t)} |p_i^\alpha - p| \Rightarrow 0. \tag{111}$$

Using (107) and the bound

$$\left|\sum_{i=1}^{A(t)} p_i^\alpha - pA(t)\right| \le A(t) \sup_{1\le i\le A(t)} |p_i^\alpha - p|,$$

together with $\mathbb{E}[A(t)] < \infty$ and (111), we obtain

$$\mathbb{E}_e[A_0^\alpha(t)] = \mathbb{E}_e\left[\sum_{i=1}^{A(t)} p_i^\alpha\right] \to \mathbb{E}[pA(t)] = p\mathbb{E}[A(t)] = p\lambda t.$$

Similarly, since $\mathbb{E}[A(t)^2] < \infty$ (because $\mathrm{Var}(A(t)) < \infty$), the same bound implies $\sum_{i=1}^{A(t)} p_i^\alpha \to pA(t)$ in L^2, and hence

$$\mathrm{Var}_e\left(\sum_{i=1}^{A(t)} p_i^\alpha\right) \to \mathrm{Var}(pA(t)) = p^2\mathrm{Var}(A(t)).$$

For the conditional-variance term, note that $x \mapsto x(1-x)$ is Lipschitz on $[0,1]$, so

$$\left|\sum_{i=1}^{A(t)} p_i^\alpha(1-p_i^\alpha) - p(1-p)A(t)\right| \le CA(t) \sup_{1\le i\le A(t)} |p_i^\alpha - p|$$

for a universal constant C, which again implies

$$\mathbb{E}_e\left[\sum_{i=1}^{A(t)} p_i^\alpha(1-p_i^\alpha)\right] \to p(1-p)\mathbb{E}[A(t)] = p(1-p)\lambda t.$$

Plugging into (107) yields

$$\mathrm{Var}_e(A_0^\alpha(t)) \to p(1-p)\lambda t + p^2\mathrm{Var}(A(t)).$$

Therefore,

$$\frac{\mathrm{Var}_e(A_0^\alpha(t))}{\mathbb{E}_e[A_0^\alpha(t)]} \to \frac{p(1-p)\lambda t + p^2\mathrm{Var}(A(t))}{p\lambda t} = (1-p) + p\frac{\mathrm{Var}(A(t))}{\lambda t} = (1-p) + pI_a(t).$$

Using (110), this equals

$$(1-p)+pI_a(t)=\left(1-\frac{1}{\rho\vee 1}\right)+\frac{I_a(t)}{\rho\vee 1}.$$

By (105),

$$\mathbb{E}_e[Y^\alpha(t)]\to\frac{1}{\mu}p\lambda t=\frac{\rho}{\rho\vee 1}t=(\rho\wedge 1)t,$$

and by (106),

$$\frac{\mathrm{Var}_e(Y^\alpha(t))}{\mathbb{E}_e[Y^\alpha(t)]}\to\frac{1}{\mu}\left(c_s^2+(1-p)+pI_a(t)\right)=\frac{1}{\mu}\left(\frac{I_a(t)}{\rho\vee 1}+\left(1-\frac{1}{\rho\vee 1}\right)+c_s^2\right).$$

This proves the lemma. □

C.11 Proof of Proposition 3

Fix $z\geq 0$ and consider the reflected diffusion $Z^{c,k}$ started from $Z^{c,k}(0)=z$. Let $\tau_0\triangleq\inf\{t\geq 0: Z^{c,k}(t)=0\}$ and let $\mathcal{L}$ denote the interior generator on $(0,\infty)$,

$$(\mathcal{L}f)(x)=f''(x)+(c-x^k)f'(x),\qquad x>0.$$

PDE for $\psi_{c,k}$. Since the exponential functional in (32) is stopped at τ_0, only the pre-hitting dynamics on $(0,\infty)$ matter (in particular, the reflection term is inactive on $[0,\tau_0)$). Define

$$\psi(t,z)=\mathbb{E}_z\left[\exp\left\{-\int_0^{t\wedge\tau_0}q_k(Z^{c,k}(s))ds\right\}\right].$$

A standard Feynman–Kac argument for diffusions stopped upon hitting the boundary[2] shows that ψ satisfies the Kolmogorov backward equation $\partial_t\psi=\mathcal{L}\psi-q_k\psi$ on $(0,\infty)$, with initial condition $\psi(0,\cdot)\equiv 1$. Moreover, since $t\wedge\tau_0=0$ when $z=0$, the boundary condition is $\psi(t,0)=1$ for all $t\geq 0$. Finally, boundedness follows from $0\leq\psi\leq 1$. Uniqueness within the class of bounded classical solutions follows from the maximum principle for linear parabolic equations on $(0,\infty)$ with Dirichlet boundary data at 0. This proves (35).

PDE for $h_{c,k}$. Recall $h_{c,k}(t,z)=\mathbb{E}_z\left[\int_0^t(Z^{c,k}(s))^k ds\right]$ from (33). Let h be any $C^{1,2}$ solution to (36) satisfying the stated growth bound. Fix $t>0$ and apply Itô's formula for reflected diffusions to the process $h(t-s,Z^{c,k}(s))$ for $0\leq s\leq t$. Using (29) and the definition of $\mathcal{L}$, we obtain

$$\begin{aligned}dh(t-s,Z^{c,k}(s))=&\Big(-\partial_t h+\mathcal{L}h\Big)(t-s,Z^{c,k}(s))ds+\sqrt{2}\partial_x h(t-s,Z^{c,k}(s))dB(s)\\&+\partial_x h(t-s,Z^{c,k}(s))dL^{c,k}(s).\end{aligned}$$

Since h satisfies (36), we have $(-\partial_t h+\mathcal{L}h)(t-s,x)=-x^k$ for $x>0$, and hence

$$d\left(h(t-s,Z^{c,k}(s))+\int_0^s(Z^{c,k}(r))^k dr\right)=\sqrt{2}\partial_x h(t-s,Z^{c,k}(s))dB(s)+\partial_x h(t-s,Z^{c,k}(s))dL^{c,k}(s).$$

[2]Equivalently, apply Itô's formula to $\exp\left\{-\int_0^{s\wedge\tau_0}q_k(Z^{c,k}(r))dr\right\}\psi(t-s,Z^{c,k}(s\wedge\tau_0)), 0\leq s\leq t$, and use optional stopping at $t\wedge\tau_0$.

Because $L^{c,k}$ increases only when $Z^{c,k}(s) = 0$, the Neumann boundary condition $\partial_x h(\cdot, 0) = 0$ implies $\partial_x h(t-s, Z^{c,k}(s))dL^{c,k}(s) = 0$. Consequently, the process

$$M(s) \triangleq h(t-s, Z^{c,k}(s)) + \int_0^s (Z^{c,k}(r))^k dr, \qquad 0 \le s \le t,$$

is a local martingale. The stated growth bound ensures that M is integrable and hence a true martingale. Taking expectations and using $h(0, \cdot) \equiv 0$ yields

$$\begin{aligned} h(t,z) = \mathbb{E}_z[M(0)] = \mathbb{E}_z[M(t)] &= \mathbb{E}_z\left[h(0, Z^{c,k}(t)) + \int_0^t (Z^{c,k}(r))^k dr\right] \\ &= \mathbb{E}_z\left[\int_0^t (Z^{c,k}(r))^k dr\right] = h_{c,k}(t,z). \end{aligned}$$

Thus any classical solution of (36) in the stated growth class coincides with $h_{c,k}$, proving uniqueness. Existence (and classical regularity) for (36) with Neumann boundary data follows from standard parabolic theory for linear equations on the half-line with smooth coefficients, and the martingale argument above identifies the solution with the stochastic representation (33).

C.12 Proof of Theorem 4

The proof follows the steps of the proof of Theorem 1, and we only record the modifications induced by the refined variance surrogate. For notational simplicity, assume $\rho(\alpha) = 1 + c\alpha^\gamma$ for all $\alpha > 0$. We assume $\gamma > 0$, so that $\rho(\alpha) \to 1$ as $\alpha \downarrow 0$. For each α, let Z_α denote the unique solution of (44). Recall $h = k/(k+1)$ and $\beta \triangleq F^{(k)}(0)/k!$. In the critically-loaded regime, the diffusion drift parameter is $\mathbb{1}\{\gamma = h\}c$, and $w_{\tilde{c},k}$ is parameterized by $\tilde{c}$ constructed as in Lemma 5.

Define, for $s \ge 0$, the auxiliary function

$$g_\alpha(s) \triangleq \hat{I}_w\left(\frac{s}{\bar{F}_\alpha(Z_\alpha) + \alpha\zeta/\lambda}\right) w_{\tilde{c},k}\left(\frac{\alpha^{2h}\tau s}{\bar{F}_\alpha(Z_\alpha) + \alpha\zeta/\lambda}\right).$$

Perform the change of variables $u = (\bar{F}_\alpha(Z_\alpha) + \alpha\zeta/\lambda)s$ in (44) and then rename u as s. This yields the equivalent fixed-point representation

$$Z_\alpha = \sup_{s \ge 0}\left\{\rho(\alpha)s - \frac{s}{\bar{F}_\alpha(Z_\alpha) + \alpha\zeta/\lambda} + b\sqrt{\frac{\rho(\alpha)s}{\mu} g_\alpha(s)}\right\}. \tag{112}$$

We analyze (112) in the three regimes.

Critically loaded. Assume $\gamma \ge h$ and define $\hat{Z}_\alpha \triangleq \alpha^h Z_\alpha$. Multiplying (112) by α^h and setting $u = \alpha^{2h}s$ yields

$$\begin{aligned} \hat{Z}_\alpha &= \sup_{u \ge 0}\left\{\alpha^{-h}\left(\rho(\alpha) - \frac{1}{\bar{F}(\alpha^{1-h}\hat{Z}_\alpha) + \alpha\zeta/\lambda}\right)u + b\sqrt{\frac{\rho(\alpha)u}{\mu} g_\alpha(\alpha^{-2h}u)}\right\} \\ &= \sup_{u \ge 0}\left\{\alpha^{\gamma-h}cu + \frac{-\alpha^{-h}F(\alpha^{1-h}\hat{Z}_\alpha) + \alpha^{1-h}\zeta/\lambda}{\bar{F}(\alpha^{1-h}\hat{Z}_\alpha) + \alpha\zeta/\lambda}u + b\sqrt{\frac{\rho(\alpha)u}{\mu} g_\alpha(\alpha^{-2h}u)}\right\}. \end{aligned}$$

Assumption 2 implies $\alpha^{-h}F(\alpha^{1-h}x) \to \beta x^k$ uniformly on compact x sets. Since $\alpha^{1-h} \to 0$, the term $\alpha^{1-h}\zeta/\lambda$ vanishes and the denominator converges to 1. Moreover, $\rho(\alpha) \to 1$, $\bar{F}(\alpha^{1-h}\hat{Z}_\alpha) \to 1$, and

$$g_\alpha(\alpha^{-2h}u) = \hat{I}_w\left(\frac{\alpha^{-2h}u}{\bar{F}(\alpha^{1-h}\hat{Z}_\alpha)+\alpha\zeta/\lambda}\right) w_{\tilde{c},k}\left(\frac{\tau u}{\bar{F}(\alpha^{1-h}\hat{Z}_\alpha)+\alpha\zeta/\lambda}\right).$$

Since $\alpha^{-2h}u \to \infty$ for each fixed $u > 0$ and $\hat{I}_w(t) \to c_x^2$ as $t \to \infty$, we have $g_\alpha(\alpha^{-2h}u) \to c_x^2 w_{\tilde{c},k}(\tau u)$ for each fixed $u \geq 0$. Let $\hat{Z}$ be any subsequential limit of $\hat{Z}_\alpha$. Passing to the limit in the previous display yields

$$\hat{Z} = \sup_{u\geq 0}\left\{\left(\mathbb{1}\{\gamma = h\}c - \beta\hat{Z}^k\right)u + b\sqrt{\frac{c_x^2}{\mu}w_{\tilde{c},k}(\tau u)u}\right\}.$$

The right-hand side is strictly decreasing in $\hat{Z}$, so the fixed point has a unique positive solution. Therefore, $\hat{Z}_\alpha \to \hat{Z}$ as $\alpha \downarrow 0$, which proves part (2).

Underloaded. Assume $c < 0$ and $0 < \gamma < h$, and define $\hat{Z}_\alpha \triangleq (1-\rho(\alpha))Z_\alpha = -c\alpha^\gamma Z_\alpha$. Multiplying (112) by $-c\alpha^\gamma$ and setting $u = \alpha^{2\gamma}c^2 s$ yields

$$\hat{Z}_\alpha = \sup_{u\geq 0}\left\{-u - \frac{-\alpha^{-\gamma}F(\alpha^{1-\gamma}\hat{Z}_\alpha)+\alpha^{1-\gamma}\zeta/\lambda}{\bar{F}(\alpha^{1-\gamma}\hat{Z}_\alpha)+\alpha\zeta/\lambda}\frac{u}{c} + b\sqrt{\frac{\rho(\alpha)u}{\mu}g_\alpha(\alpha^{-2\gamma}u/c^2)}\right\}.$$

Since $\gamma < h$ we have $k - \gamma(k+1) > 0$, and Assumption 2 implies $\alpha^{-\gamma}F(\alpha^{1-\gamma}x) \to 0$ uniformly on compact x sets. The term $\alpha^{1-\gamma}\zeta/\lambda$ also vanishes, so the middle fraction converges to 0. Moreover, $\alpha^{-2\gamma}u/c^2 \to \infty$ for each fixed $u > 0$ and $\alpha^{2h}\tau\alpha^{-2\gamma}u/c^2 \to 0$, so $g_\alpha(\alpha^{-2\gamma}u/c^2) \to c_x^2 w_{\tilde{c},k}(0)$. By Proposition 2, $w_{\tilde{c},k}(0) = 1$, and letting $\alpha \downarrow 0$ gives

$$\hat{Z} = \sup_{u\geq 0}\left\{-u + b\sqrt{\frac{c_x^2}{\mu}u}\right\} = \frac{b^2c_x^2}{4\mu}.$$

This proves part (1).

Overloaded. Assume $c > 0$ and $0 < \gamma < h$, and define $\hat{Z}_\alpha \triangleq \alpha^{1-\gamma/k}Z_\alpha$. Multiplying (112) by $\alpha^{1-\gamma/k}$ and setting $u = \alpha^{1+(k-1)\gamma/k}s$ yields

$$\hat{Z}_\alpha = \sup_{u\geq 0}\left\{cu + \alpha^{-\gamma}\left(1 - \frac{1}{\bar{F}(\alpha^{\gamma/k}\hat{Z}_\alpha)+\alpha\zeta/\lambda}\right)u + b\sqrt{\alpha^{1-(k+1)\gamma/k}\frac{\rho(\alpha)u}{\mu}g_\alpha(\alpha^{-1-(k-1)\gamma/k}u)}\right\}.$$

Since $\gamma < h$, we have $1 - (k+1)\gamma/k > 0$, so the square root term vanishes uniformly on compact u sets. For the supremum to be finite, the linear coefficient of u must be nonpositive, which implies

$$c + \alpha^{-\gamma}\left(1 - \frac{1}{\bar{F}(\alpha^{\gamma/k}\hat{Z}_\alpha)+\alpha\zeta/\lambda}\right) \leq 0.$$

Using $1 - 1/x = -(x-1)/x$ and the expansion $F(\alpha^{\gamma/k}x) = \beta\alpha^\gamma x^k + o(\alpha^\gamma)$ yields

$$c \leq \beta\hat{Z}_\alpha^k + o(1). \tag{113}$$

To obtain the reverse bound, fix $\varepsilon > 0$ and suppose that along some sequence $\alpha_n \downarrow 0$,

$$c + \alpha_n^{-\gamma}\left(1 - \frac{1}{\bar{F}(\alpha_n^{\gamma/k}\hat{Z}_{\alpha_n}) + \alpha_n\zeta/\lambda}\right) \le -\varepsilon.$$

Since $w_{\tilde{c},k}$ is bounded by 1 and $\hat{I}_w$ is bounded on $[0,\infty)$, there exists a finite constant M such that $g_\alpha(s) \le M$ for all α and s. The display above then implies

$$\hat{Z}_{\alpha_n} \le \sup_{u\ge 0}\left\{-\varepsilon u + b\sqrt{\alpha_n^{1-(k+1)\gamma/k}\frac{\rho(\alpha_n)Mu}{\mu}}\right\} = \frac{b^2\rho(\alpha_n)M}{4\mu\varepsilon}\alpha_n^{1-(k+1)\gamma/k},$$

so $\hat{Z}_{\alpha_n} \to 0$. This contradicts (113) with $c > 0$. Therefore, for all sufficiently small α,

$$c + \alpha^{-\gamma}\left(1 - \frac{1}{\bar{F}(\alpha^{\gamma/k}\hat{Z}_{\alpha}) + \alpha\zeta/\lambda}\right) \ge -\varepsilon.$$

Letting $\alpha \downarrow 0$ and then $\varepsilon \downarrow 0$ yields $c \ge \beta\hat{Z}^k$ for any subsequential limit $\hat{Z}$ of $\hat{Z}_\alpha$. Combining this with (113) yields $\hat{Z}^k = c/\beta$. Therefore, $\hat{Z}_\alpha \to (c/\beta)^{1/k} = (ck!/F^{(k)}(0))^{1/k}$ as $\alpha \downarrow 0$, which proves part (3).