Mathematics of Operations Research

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

Excursion-Based Universal Approximations for the Erlang-A Queue in Steady-State

Itai Gurvich, Junfei Huang, Avishai Mandelbaum

To cite this article:
Itai Gurvich, Junfei Huang, Avishai Mandelbaum (2014) Excursion-Based Universal Approximations for the Erlang-A Queue in Steady-State. Mathematics of Operations Research 39(2):325-373. http://dx.doi.org/10.1287/moor.2013.0606

MATHEMATICS OF OPERATIONS RESEARCH
Vol. 39, No. 2, May 2014, pp. 325–373
ISSN 0364-765X (print) | ISSN 1526-5471 (online)

http://dx.doi.org/10.1287/moor.2013.0606

Excursion-Based Universal Approximations for the Erlang-A Queue in Steady-State

Itai Gurvich
Kellogg School of Management, Northwestern University, Evanston, Illinois 60208,
i-gurvich@kellogg.northwestern.edu

Junfei Huang
Department of Decision Sciences and Managerial Economics, Faculty of Business Administration,
Chinese University of Hong Kong, Shatin, N.T., Hong Kong, junfeih@gmail.com

Avishai Mandelbaum
William Davidson Faculty of Industrial Engineering and Management, Technion—Israel Institute of Technology,
Technion City, Haifa 32000, Israel, avim@ie.technion.ac.il

We revisit many-server approximations for the well-studied Erlang-A queue. This is a system with a single pool of i.i.d. servers that serve one class of impatient i.i.d. customers. Arrivals follow a Poisson process and service times are exponentially distributed as are the customers' patience times. We propose a diffusion approximation that applies simultaneously to all existing many-server heavy-traffic regimes: quality and efficiency driven, efficiency driven, quality driven, and nondegenerate slowdown. We prove that the approximation provides accurate estimates for a broad family of steady-state metrics. Our approach is "metric-free" in that we do not use the specific formulas for the steady-state distribution of the Erlang-A queue. Rather, we study excursions of the underlying birth-and-death process and couple these to properly defined excursions of the corresponding diffusion process. Regenerative process and martingale arguments, together with derivative bounds for solutions to certain ordinary differential equations, allow us to control the accuracy of the approximation. We demonstrate the appeal of universal approximation by studying two staffing optimization problems of practical interest.

Keywords: many-server queues; diffusion approximation; steady-state; universal; Erlang-A; excursions; regenerative processes; staffing; call centers
MSC2000 subject classification: Primary: 60F17; secondary: 60G10
OR/MS subject classification: Primary: queues/approximations; secondary: probability/regenerative processes, probability/diffusion
History: Received April 7, 2012; revised November 24, 2012. Published online in *Articles in Advance* June 27, 2013.

1. Introduction. Heavy-traffic limits provide tractable means to approximate and optimize the performance of various queueing systems. Often, the limit is characterized by a diffusion process. When the diffusion process admits a steady-state distribution, that distribution can serve (under appropriate conditions) as an approximation for the steady-state distribution of the pre-limit queueing system.

The diffusion-limit approach to the study of queueing systems has been successfully applied to study large-scale service systems as part of what came to be known as "many-server heavy-traffic approximations." Our focus here is on approximations to the $M/M/n+M$ (also known as the Erlang-A) queue—this is a queue with Poisson arrivals, i.i.d. exponential service times, n servers, and i.i.d. exponential patience times. The Erlang-A queue is a central building block in the study of service systems, most notably call centers, where abandonment plays a nonnegligible role; see Mandelbaum and Zeltyn [27, §2].

In the many-server-approximations framework, one considers a sequence of queues with individual service rate μ and abandonment rate θ, indexed by the arrival rate λ. Letting n^λ be the number of servers in the λth queue, define

$$\rho^\lambda = \frac{\lambda}{\mu n^\lambda}$$

to be the *offered utilization*. Let $X^\lambda(t)$ be the number of customers in the system (in service or in queue) at time t. The process $X^\lambda = (X^\lambda(t), t \geq 0)$ is then a birth-and-death (B&D) process on the nonnegative integers, with birth rate $\lambda(x) \equiv \lambda$ and death rate $\mu(x) = \mu(x \wedge n^\lambda) + \theta(x - n^\lambda)^+$ in state x.

The specific value of $\rho = \lim_{\lambda\to\infty} \rho^\lambda$ (assuming it exists) induces a useful categorization into operational regimes, which relates ρ to the fundamental metric of the probability of delay; see Garnett et al. [15]. If $\rho < 1$, we say that the system operates in the quality-driven (QD) regime. Here, capacity is significantly greater than the load and the fraction of customers experiencing any delay before entering service converges to 0 as $\lambda \to \infty$. Further, the number of abandoning customers decreases to 0 exponentially fast as λ grows indefinitely; see Iglehart [18] and Whitt [39]. This is in contrast to the efficiency-driven (ED) regime in which $\rho > 1$, where essentially all customers are delayed before being served and a nonnegligible fraction of customers abandons;

see, e.g., Whitt [40]. Finally, the case in which $\rho = 1$ and $\sqrt{\lambda}(1-\rho^\lambda) \to \beta \in (-\infty, \infty)$ is referred to as the quality-and-efficiency-driven (QED) regime because it offers a combination of high efficiency and quality of service. The deepest characteristic of the QED regime, introduced by Halfin and Whitt [17] for Erlang-C, is in terms of the limiting probability of delay, which is to be strictly between 0 and 1. For Erlang-A, an additional characterization is in terms of the fraction of abandoning customers, which approaches 0 at a rate of $1/\sqrt{\lambda}$; see, e.g., Garnett et al. [15]. A QED refinement of the ED regime (ED + QED) was introduced in Mandelbaum and Zeltyn [28], in order to generate staffing that accommodates constraints on the probability that waiting time exceeds a fixed target T.

More recently, an additional many-server regime was studied by Atar [5], who entitled it the nondegenerate-slowdown (NDS) regime. As in the QED regime, one sets $\sqrt{\lambda}(1-\rho^\lambda) \to \beta \in (-\infty, \infty)$ but, in contrast to the QED regime, the individual service rate scales here with λ proportionally to $\sqrt{\lambda}$. In this regime, in particular, n^λ is proportional to $\sqrt{\lambda}$. The NDS regime offers a hybrid of the QED and ED regimes—as in the former, the fraction of abandoning customers approaches 0 at the rate of $1/\sqrt{\lambda}$ whereas in the latter, the probability of delay approaches 1 as λ approaches ∞. We consider the NDS regime in §C.

We refer the reader to Garnett et al. [15], Mandelbaum and Zeltyn [28], Zeltyn and Mandelbaum [41], and Atar [5] for more detailed discussions of operational regimes. Toward constructing a universal approximation, it is useful to identify

$$\Delta^\lambda = \frac{\lambda}{\mu} - \left(\frac{\lambda}{\mu} - n^\lambda\right)^+ \left(1 - \frac{\mu}{\theta}\right) \tag{1}$$

as the "balancing" point in the state-space of X^λ at which the inflow rate equals the outflow rate; i.e., $\lambda = \mu(n^\lambda \wedge \Delta^\lambda) + \theta(\Delta^\lambda - n^\lambda)^+$. This state serves as a first-order proxy for the number of customers in steady-state. When $n^\lambda < \lambda/\mu$, we have that $\Delta^\lambda = n^\lambda + (\lambda - n^\lambda\mu)/\theta$ so that the balancing point is where the queue is strictly positive. If $n^\lambda > \lambda/\mu$, then $\Delta^\lambda = \lambda/\mu$ so that the queue is, in first order, empty. Using the known diffusion-limit results (see Ward [38]), one can verify that under any of the multi-server regimes ED, QD, QED, or NDS, the process convergence

$$\frac{X^\lambda - \Delta^\lambda}{\sqrt{\lambda}} \implies \hat{X} \tag{2}$$

holds, where $\hat{X}$ is an Ornstein-Uhlenbeck (OU) type process whose specific structure depends on the specific regime.

Given this process convergence, one further expects the steady-state of the diffusion process $\hat{X}$ to provide an approximation for the steady-state of the pre-limit queues; that is,

$$\hat{X}^\lambda(\infty) := \frac{X^\lambda(\infty) - \Delta^\lambda}{\sqrt{\lambda}} \implies \hat{X}(\infty),$$

where $\hat{X}(\infty)$ has the steady-state distribution of the corresponding OU type process. Making such approximation rigorous requires a limit interchange result; see the discussion in Ward [38, p. 6]. This has been proved for the QED regime in Garnett et al. [15], for the ED regime in Whitt [40], and for the QD regime in Whitt [39] (whose arguments for Erlang-C apply also to the Erlang-A queue). It has not been proved yet for the NDS regime. A byproduct of our analysis is that the limit interchange holds universally; see Remark 4.1.

That the process limit is regime dependent motivates the universal approximation for the Erlang-A queue that is proposed in Ward [38]. The author introduces a Brownian approximation, $\check{Y}^\lambda$ for each λ, that covers the QED, NDS, and conventional (or single server) heavy-traffic regimes. The proposed approximation is universal in the sense of *process* convergence in the QED and NDS regimes: if one assumes QED scaling, then $(\check{Y}^\lambda - n^\lambda)/\sqrt{\lambda}$ converges weakly to the OU process characteristic of the QED regime. If, in contrast, one assumes the NDS regime (or the single-server conventional heavy-traffic one), $(\check{Y}^\lambda - n^\lambda)/\sqrt{\lambda}$ converges weakly to the reflected OU process, which is characteristic of this regime; see Ward [38, Theorem 4.1].

Such process convergence is *not* the subject of this paper. Our approach to universality is different. We are primarily interested in pre-limit approximations (rather than limits) for *steady-state* metrics and their associated error bounds. We do not use weak convergence or diffusion limits per se. Instead, for each λ, we offer a diffusion process, Y^λ, where the parameters λ, μ, θ and n^λ appear explicitly in its characterization (see (3) below). We prove that regardless of the underlying regime, $X^\lambda(\infty)$ and $Y^\lambda(\infty)$ are "close" to each other in terms of their expected performance metrics; see §1.2. Accordingly, we refer to our proposed process Y^λ as a *universal diffusion*.

The universality of the approximation and, more specifically, the performance bounds that we provide build on a novel analysis approach. To elaborate, a possible approach toward steady-state approximations is to use

the explicit expressions for the distribution of $X^\lambda(\infty)$. One computes, for each integer k, the corresponding steady-state probability $\mathbb{P}\{X^\lambda(\infty)=k\}$ and uses it to obtain various performance metrics; see, e.g., Mandelbaum and Zeltyn [27, Appendix A]. To compare the B&D process to the diffusion process, one can analytically bound, for example, the gap between $\mathbb{P}\{X^\lambda(\infty)\geq k\}$ and $\mathbb{P}\{Y^\lambda(\infty)\geq k\}$. This is the nature of the approach in Zhang et al. [42] and Bassamboo and Randhawa [6].

In contrast, we do not use the specific expressions for the steady-state distribution of $X^\lambda(\infty)$. Rather, we introduce an excursion-based approach that circumvents the exact expressions. Our contribution has, then, four interrelated elements: (a) *universal approximation:* We have a family of diffusion processes such that, for each λ, the diffusion process explicitly depends on the system parameters and applies to all regimes. (b) *Refined bounds:* We provide order-of-magnitude bounds for the accuracy of the proposed approximation for a large family of performance metrics. (c) *Universal optimization:* We demonstrate this via two (asymptotically) optimal staffing problems. (d) *Excursion-based analysis:* Our analysis relies on the regenerative and martingale structure of both the diffusion and the B&D processes and on properties of smooth solutions to certain ordinary differential equations.

We next expand on each of the above.

1.1. A "universal" approximation. For each λ, we propose Y^λ to be the diffusion process given by the unique solution to the stochastic differential equation (SDE)

$$Y^\lambda(t)=Y^\lambda(0)+\lambda t-\mu\int_0^t(Y^\lambda(s)\wedge n^\lambda)\,ds-\theta\int_0^t(Y^\lambda(s)-n^\lambda)^+\,ds+\sqrt{2\lambda}B(t). \tag{3}$$

There is an intimate relation between the diffusion process Y^λ and the limit process that arises in the QED regime. Assuming that $\beta^\lambda:=(n^\lambda\mu-\lambda)/\sqrt{\lambda}\equiv\beta$, the process $\hat{Y}^\lambda=(Y^\lambda-n^\lambda)/\sqrt{\lambda}$ would satisfy the SDE

$$\hat{Y}^\lambda(t)=\hat{Y}^\lambda(0)-\beta t+\mu\int_0^t(\hat{Y}^\lambda(s))^-\,ds-\int_0^t\theta(\hat{Y}^\lambda(s))^+\,ds+\sqrt{2}B(t),$$

which is the OU type process obtained as a limit in the QED regime; see Ward [38, Theorem 2.2]. In a sense, then, we "universalize" the QED diffusion by allowing its drift and diffusion coefficient to depend explicitly on the parameters, μ, λ, and n^λ; see further discussion in our Remark 2.1. The process Y^λ could also play the role of a "strong approximation" for X^λ (see Remark 2.2). This implies that for any of the multi-server regimes (QED, ED, QD, NDS),

$$\frac{Y^\lambda-\Delta^\lambda}{\sqrt{\lambda}}\implies\hat{X};$$

here, $\hat{X}$ is any of the four OU type processes, obtained from the scaled and centered queueing processes in (2), each corresponding to an underlying regime. Although establishing process limits is not the subject of this paper, the fact that $(Y^\lambda-\Delta^\lambda)/\sqrt{\lambda}$ has the "correct" limits serves as a strong indication of it being a natural choice for a universal approximation.

We prove that our universal approximation provides accurate steady-state metrics regardless of the underlying regime. Such universality is useful for purposes of performance analysis, data inference, and optimization. The value for performance analysis is clear, as demonstrated in §1.2. Indeed, considering a fixed queueing system, it is useful to have performance metrics that are relatively precise yet offer the tractability of diffusion approximations. Such approximations of queues have been recently used also for the purpose of structural inference (see, e.g., Allon et al. [2]). In this context, a universal approximation allows one to avoid a priori assumptions about the operational regime that underlies the data. Finally, in §1.3, we describe the application of our approximation to universal optimization.

1.2. Error bounds: Performance analysis. Our main result, Theorem 1, states that $Y^\lambda(\infty)$ provides an accurate universal approximation for the original B&D process. By this we mean that for each nonnegative integer m, *universally*

$$\mathbb{E}[(X^\lambda(\infty)-\Delta^\lambda)^m]-\mathbb{E}[(Y^\lambda(\infty)-\Delta^\lambda)^m]=\mathcal{O}(\sqrt{\lambda}^{m-1}) \tag{4}$$

as well as

$$\sup_{x\geq0}\left|\mathbb{P}\{X^\lambda(\infty)\geq x\}-\mathbb{P}\{Y^\lambda(\infty)\geq x\}\right|=\mathcal{O}(\sqrt{\lambda}^{-1});$$

here we use the convention that for two sequences $\{x^\lambda\}$ and $\{y^\lambda\}$, $x^\lambda = \mathcal{O}(y^\lambda)$ if $\limsup_{\lambda\to\infty}(|x^\lambda|/|y^\lambda|) < \infty$. Letting $\tilde{Q}^\lambda(\infty) := (Y^\lambda(\infty) - n^\lambda)^+$, we obtain as a consequence of (4) that

$$\mathbb{E}[Q^\lambda(\infty)] - \mathbb{E}[\tilde{Q}^\lambda(\infty)] = \mathcal{O}(1) \tag{5}$$

or, in other words, that the queue length is approximated, up to a constant, by the "queue" of the Brownian approximation. We cover a rather broad family of functions, of which the power functions in (4) are special cases.

The universality of the approximation comes at some cost. If, for example, one restricts attention to the QED regime, the errors in (5) exceed those of Zhang et al. [42]: the guaranteed precision is $o(1)$; namely, the error vanishes in absolute terms as λ grows. There is also a "complexity cost" when specializing to the ED regime. In Bassamboo and Randhawa [6] it is shown that in the ED regime and for the special metric of the expected queue length (see (5)), the simple fluid model is as precise as our more complicated universal approximation. The returns for these "costs" are the universality of our proposed approximation, the generality of our performance metrics, and the expression-free nature of our proofs.

1.3. Universal optimization. Typical optimization problems seek to minimize capacity costs subject to service level constraints (see Mandelbaum and Zeltyn [28]) or, alternatively, minimize a weighted cost of capacity and service level (e.g., Bassamboo et al. [7], Bassamboo and Randhawa [6], and the references therein). In this context, a caveat with heavy-traffic limits is that these require imposing assumptions on the scaling of the constraints or of the cost coefficients.

1.3.1. Constraint satisfaction. As a case in point, consider the problem of minimizing the number of servers while maintaining a pre-specified bound, α, on the fraction of abandonments. Limit-based solutions depend on the way in which α scales with λ. If it is not scaled, as in $\alpha(\lambda) \equiv \alpha$, then the system operates optimally in the ED regime and it is asymptotically optimal to use $n^\lambda = (\lambda/\mu)(1-\alpha) + o(\lambda)$ servers; see Mandelbaum and Zeltyn [28, §4.3]. If, on the other hand, $\alpha(\lambda) = c/\sqrt{\lambda}$ for some $c > 0$, a rather different solution emerges. Here, the system operates optimally in the QED regime and the recommended staffing has the so-called square-root staffing solution $n^\lambda = \lambda/\mu + \beta\sqrt{\lambda/\mu} + o(\sqrt{\lambda})$, where β is a function of c, μ, and θ; see Mandelbaum and Zeltyn [28, §4.3]. From a practical point of view, then, using heavy-traffic *limits* requires an interpretation step. If, for example, $\lambda = 100$, $\mu = 1$, and $\theta = 3$, a 5% abandonment target may be interpreted as corresponding to $\alpha(\lambda) \equiv \alpha = 0.05$ or alternatively as $\alpha(100) = 0.5/\sqrt{100} = 0.05$. The real optimal solution obtained by using an Erlang calculator (4 call centers [1]) is 101 servers. Our universal approximation provides the same solution; see §5. If one assumes that α does not scale with λ, the ED-based recommendation is $(\lambda/\mu)(1-\alpha) = 95$ servers. When applied to the queueing system, this results in an 8.1% abandonment rate instead of the targeted 5%. If, on the other hand, one interprets the constraint as $\alpha(\lambda) = 0.5/\sqrt{\lambda}$, the QED-based solution is 101 servers, which recovers the precise solution in this case. This in particular supports the robustness of the QED regime (which is mathematically supported by our results and by the connection, discussed above, between the QED diffusion and our universal approximation). The ED staffing level does produce reasonably good solutions when λ is larger. With $\lambda = 1{,}000$, for example, the ED staffing level amounts to using 950 agents (the precise optimal solution is 954, as also identified by the universal approximation). Using 950 servers will result in 5.3% abandonment, which is only a minor violation of the target.

1.3.2. Cost minimization. The need for an interpretation step arises also in the context of cost minimization, where one seeks to minimize weighted costs of staffing, waiting, and abandonment. Such optimization problems were studied for the Erlang-C queue (i.e., with no abandonment) in Borst et al. [8] via limit arguments, and we revisit this problem for the Erlang-A queue in §5. Specifically, assume that μ and θ are fixed and let $\mathbb{E}[Q(\lambda, n)]$ be the expected queue length when the arrival rate is λ and there are n servers. Similarly, let $\mathrm{Ab}(\lambda, n)$ be the fraction of abandoning customers. Let C_s^λ, C_q^λ, and C_{ab}^λ be, respectively, the cost per server per unit of time, the cost incurred by a customer waiting one unit of time, and the cost per customer abandonment. Consider the optimization problem

$$\min_{n\in\mathbb{N}} \left\{ C_s^\lambda n + C_q^\lambda \mathbb{E}[Q(\lambda, n)] + C_{ab}^\lambda \lambda \,\mathrm{Ab}(\lambda, n) \right\}.$$

A single example suffices as a case in point for the introduction (additional numerical experiments appear in §5). For the case $\mu = \theta = 1$, $\lambda = 100$, $C_s^\lambda \equiv 2$, and $C_q^\lambda = C_{ab}^\lambda \equiv 10$, the optimal solution (identified through direct enumeration and an Erlang calculator) is 113 servers. This is also the solution recommended by our universal approximation. If one interprets $C_s^\lambda, C_q^\lambda, C_{ab}^\lambda$ as being constants (that do not scale with λ), the system operates optimally in the QED regime and an asymptotically optimal solution is given by a square-root staffing rule;

see Bassamboo et al. [7, Proposition 1]. Asymptotic optimality in the context of cost minimization has not been yet studied at the generality of the Erlang-C queue (Borst et al. [8]). For example, it is not known what asymptotically optimal recommendation emerges should one interpret the cost coefficients as corresponding to $C_s^\lambda \equiv 2$ but $C_q^\lambda = C_{ab}^\lambda = \sqrt{\lambda}$. For our purposes, the important fact is that the universal approximation, being explicitly dependent on the parameters, can be directly applied without the need to interpret the parameters and results, in this case, in an accurate recommendation. We return to both the constraint satisfaction and cost minimization problems in §5.

1.4. The excursion-based argument. For stable B&D processes, steady-state metrics are given by averages over finite (albeit random) horizons. Specifically, the positive recurrence of $\tilde{X}^\lambda = X^\lambda - \Delta^\lambda$ guarantees that for every function f that is integrable with respect to its steady-state distribution,

$$\mathbb{E}[f(\tilde{X}^\lambda(\infty))] = \frac{\mathbb{E}_1[\int_0^{\tau^\lambda} f(\tilde{X}^\lambda(s))\,ds]}{\mathbb{E}_1[\tau^\lambda]},$$

where τ^λ is the first hitting time of $\tilde{X}^\lambda$ at 1 after hitting 0, $\tilde{X}^\lambda(\infty)$ is a random variable having the steady-state distribution of $\tilde{X}^\lambda$, and $\mathbb{E}_y$ is the expectation conditional on $\tilde{X}^\lambda(0) = y$. (There are other ways to choose the regenerative cycle but this specific choice will be useful in what follows.) For the diffusion process $\tilde{Y}^\lambda = Y^\lambda - \Delta^\lambda$, it similarly holds that

$$\mathbb{E}[f(\tilde{Y}^\lambda(\infty))] = \frac{\mathbb{E}_1[\int_0^{\tilde{\tau}^\lambda} f(\tilde{Y}^\lambda(s))\,ds]}{\mathbb{E}_1[\tilde{\tau}^\lambda]},$$

for appropriate functions f, where $\tilde{\tau}^\lambda$ is the first hitting time of $\tilde{Y}^\lambda$ at 1 after hitting 0 and, with abuse of notation, $\mathbb{E}_y$ also denotes expectation conditional on $\tilde{Y}^\lambda(0) = y$. Thus, toward obtaining a universal Brownian approximation, it suffices to approximate $\tilde{X}^\lambda$ on the (random) finite horizon $[0, \tau^\lambda)$ by $\tilde{Y}^\lambda$ on the time interval $[0, \tilde{\tau}^\lambda)$; and because the duration of the excursion becomes small ($\mathcal{O}(1/\sqrt{\lambda})$), it guarantees that $\tilde{Y}^\lambda$ and $\tilde{X}^\lambda$ do not "drift apart" and enables an accurate approximation.

Brownian approximations over finite horizons (rather than limits for scaled processes) are well studied through strong approximations. These can be used for various queueing systems; see, e.g., Mandelbaum et al. [29] (which covers, in particular, the Erlang-A queue) as well as Chen [10], Chen and Shen [11], and the references therein. We observe that the process Y^λ in (3) is simpler than a direct strong approximation of the Erlang-A queue. Indeed, a strong approximation of X^λ would be given by a standard Brownian motion B and the unique strong solution $\breve{Y}^\lambda$ of

$$\breve{Y}^\lambda(t) = X^\lambda(0) + \lambda t - \mu \int_0^t (\breve{Y}^\lambda(s) \wedge n^\lambda)\,ds - \theta \int_0^t (\breve{Y}^\lambda(s) - n^\lambda)^+\,ds + \int_0^t \sigma^\lambda(\breve{Y}^\lambda(s))\,dB(s), \tag{6}$$

where, for each $x \geq 0$,

$$(\sigma^\lambda(x))^2 = \lambda + \mu(x \wedge n^\lambda) + \theta(x - n^\lambda)^+. \tag{7}$$

The simplicity of Y^λ, relative to $\breve{Y}^\lambda$, is facilitated by the relationship between the steady-state metrics and excursions of (short) random length; see Remark 2.2. Although strong approximations turn out to be inappropriate for the purpose of getting the error bounds that we seek to prove, the idea of treating the approximation of steady-state metrics as that of performance-comparison over finite horizons, albeit random, is valid and lies at the core of our analysis, as we explain next.

Let $\tilde{\tau}_u^\lambda$ be the first time that the diffusion process $\tilde{Y}^\lambda = Y^\lambda - \Delta^\lambda$ hits 0. Let $\mathcal{A}^\lambda$ be the generator of $\tilde{Y}^\lambda$. Then it is a matter of standard arguments that

$$\mathcal{V}^\lambda(y) = \mathbb{E}_y\left[\int_0^{\tilde{\tau}_u^\lambda} f(\tilde{Y}^\lambda(s))\,ds\right]$$

solves the ordinary differential equation (ODE)

$$\mathcal{A}^\lambda \mathcal{V}^\lambda = -f, \qquad \mathcal{V}^\lambda(0) = 0;$$

see Equation (32). Similarly, let τ_u^λ be the first hitting time of $\tilde{X}^\lambda = X^\lambda - \Delta^\lambda$ to the state 0. Let $\mathcal{B}^\lambda$ be the generator of the B&D process $\tilde{X}^\lambda$. Applying Dynkin's formula (heuristically at this stage) one obtains that for each $y > 0$,

$$\mathbb{E}_y[\mathcal{V}^\lambda(\tilde{X}^\lambda(\tau_u^\lambda))] = \mathcal{V}^\lambda(y) + \mathbb{E}_y\left[\int_0^{\tau_u^\lambda} \mathcal{B}^\lambda \mathcal{V}^\lambda(\tilde{X}^\lambda(s))\,ds\right];$$

see Equation (40). Recalling that $\tilde{X}^\lambda(\tau_u^\lambda)=0$, $\mathscr{V}^\lambda(0)=0$, and $\mathscr{A}^\lambda\mathscr{V}^\lambda=-f$, we then have that

$$\mathscr{V}^\lambda(y)-\mathbb{E}_y\left[\int_0^{\tau_u^\lambda} f(\tilde{X}^\lambda(s))\,ds\right]=\mathbb{E}_y\left[\int_0^{\tau_u^\lambda}(\mathscr{A}^\lambda\mathscr{V}^\lambda(\tilde{X}^\lambda(s))-\mathscr{B}^\lambda\mathscr{V}^\lambda(\tilde{X}^\lambda(s)))\,ds\right]. \tag{8}$$

In particular, to bound the gap

$$\mathbb{E}_y\left[\int_0^{\tau_u^\lambda} f(\tilde{X}^\lambda(s))\,ds\right]-\mathbb{E}_y\left[\int_0^{\tilde{\tau}_u^\lambda} f(\tilde{Y}^\lambda(s))\,ds\right],$$

it is enough to bound the right-hand side of (8). It is here where much of the challenge lies. We use preliminary order bounds on the hitting times, gradient bounds for $\mathscr{V}^\lambda$ (see Lemma 4.7), and martingale arguments to bound this error term. We can then approximate the integrals over excursions of the B&D process by those of the diffusion process. Finally, the cycle $[0,\tau^\lambda)$—starting at 1 until returning to 1 after hitting 0—can be decomposed into two parts—an upper excursion (starting at 1 until hitting 0) and a lower excursion (starting at 0 until hitting 1). The above arguments are applied separately to each of these excursions and then combined to bound the gap between $\mathbb{E}[f(\tilde{X}^\lambda(\infty))]$ and $\mathbb{E}[f(\tilde{Y}^\lambda(\infty))]$.

The idea of considering a sequence of Brownian queues and using gradient bounds, together with a martingale argument, to show that a Brownian approximation is "close" to the real queue is adopted from Ata and Gurvich [4]. There, it is used toward the study of an optimal control problem in a multi-class queue. Specifically, our function $\mathscr{V}^\lambda$ serves as the analogue of the value function of the diffusion control problem in Ata and Gurvich [4]. To the best of our knowledge, we are the first to use such process-based analysis to obtain error bounds on the steady-state distributions.

To summarize, the three key elements in our analysis are (i) regenerative structure of the queueing and diffusion process, (ii) derivative bounds for the "value" function of the diffusion process, and (iii) martingale properties of the queueing and diffusion processes. In §6 we discuss the potential application of these ideas to other queueing systems.

Notation. Our main results concern bounds that are uniform in the arrival rate λ. Following standard terminology, we write $a^\lambda=\mathcal{O}(b^\lambda)$ for two sequences $\{a^\lambda\}$ and $\{b^\lambda\}$ such that $\limsup_{\lambda\to\infty}(|a^\lambda|/|b^\lambda|)<\infty$. The queueing processes that we consider are assumed to be right-continuous with left limits and we let $\mathscr{D}[0,\infty)$ be the space of such functions on $[0,\infty)$. For $x\in\mathscr{D}$ we denote $\Delta x(t)=x(t)-x(t-)$. The B&D process X^λ and the diffusion process Y^λ that we will construct are real-valued Markov processes. For a Markov process $\mathscr{Z}$ on a complete and separable metric space, we write $\mathbb{P}_x\{\mathscr{Z}(t)\in\cdot\}$ for the conditional probability $\mathbb{P}\{\mathscr{Z}(t)\in\cdot\,|\mathscr{Z}(0)=x\}$. The operator $\mathbb{E}_x[\,\cdot\,]$ is then the expectation with respect to the probability distribution $\mathbb{P}_x\{\,\cdot\,\}$. In the analysis below, the probability and the corresponding expectation are applied interchangeably to the B&D process and the diffusion process; the correct interpretation will be clear from the context. A distribution π is said to be a stationary distribution if for any bounded continuous functions f, $\mathbb{E}_\pi[f(\mathscr{Z}(0))]=\mathbb{E}_\pi[f(\mathscr{Z}(t))]$ for all $t\geq 0$. It is said to be the steady-state distribution if for every such function and all $x\in\mathscr{X}$, $\mathbb{E}_x[f(\mathscr{Z}(t))]\to\mathbb{E}_\pi[f(\mathscr{Z}(0))]$ as $t\to\infty$.

When considering a Markov process, $\mathscr{Z}$, that admits a unique steady-state distribution, we use $\mathscr{Z}(\infty)$ to denote a random variable with this steady-state distribution. We use the conventions $\mathbb{R}_+=[0,\infty)$ and $\mathbb{N}=\{0,1,2,\dots\}$. For an l-times differentiable function $f\colon\mathbb{R}\to\mathbb{R}$, we write $f^{(l)}(\,\cdot\,)$ for its lth derivative. Finally, we use the term *absolute constant* when referring to a strictly positive constant that does not depend on λ.

2. Martingale representation and the universal diffusion. We consider a family of $M/M/n+M$ queues indexed by the arrival rate, $\lambda\in\mathbb{R}_+$. The service rate $\mu>0$ and the patience parameter $\theta>0$ are fixed throughout the sequence. The number of servers in the λth system is n^λ.

Let $Z^\lambda(t)$ be the number of busy servers in the λth system at time t and $Q^\lambda(t)$ the queue length at that time. The process $X^\lambda(t)=Z^\lambda(t)+Q^\lambda(t)$ captures the headcount—the total number of customers in the system at time t—is then a B&D process on the nonnegative integers. With $\theta>0$, it is known that X^λ always admits a steady-state distribution. We denote by $X^\lambda(\infty)$ a random variable that has this distribution.

It is standard to construct the sample paths of X^λ, through time changes of unit-rate Poisson processes, in the following way:

$$X^\lambda(t)=X^\lambda(0)+E(\lambda t)-S\left(\mu\int_0^t Z^\lambda(s)\,ds\right)-N\left(\theta\int_0^t Q^\lambda(s)\,ds\right),\quad t\geq 0, \tag{9}$$

where $E(\,\cdot\,)$, $S(\,\cdot\,)$, $N(\,\cdot\,)$ are independent unit-rate Poisson processes. Since there can be no idle servers simultaneously with a positive queue, we have

$$Q^\lambda(t)=(X^\lambda(t)-n^\lambda)^+\qquad\text{and}\qquad Z^\lambda(t)=X^\lambda(t)\wedge n^\lambda. \tag{10}$$

As a result, (9) is equivalently written as

$$X^\lambda(t) = X^\lambda(0) + E(\lambda t) - S\left(\mu \int_0^t (X^\lambda(s) \wedge n^\lambda)\, ds\right) - N\left(\theta \int_0^t (X^\lambda(s) - n^\lambda)^+\, ds\right).$$

Let

$$M_a^\lambda(t) = E(\lambda t) - \lambda t,$$
$$M_s^\lambda(t) = S\left(\mu \int_0^t (X^\lambda(s) \wedge n^\lambda)\, ds\right) - \mu \int_0^t (X^\lambda(s) \wedge n^\lambda)\, ds,$$
$$M_r^\lambda(t) = N\left(\theta \int_0^t (X^\lambda(s) - n^\lambda)^+\, ds\right) - \theta \int_0^t (X^\lambda(s) - n^\lambda)^+\, ds.$$

Each of these processes is a square-integrable martingale with respect to the filtration $\mathbb{F}^\lambda = (\mathcal{F}_t^\lambda, t \geq 0)$, given by

$$\mathcal{F}_t^\lambda = \sigma\left\{E(\lambda s), S\left(\mu \int_0^s (X^\lambda(u) \wedge n^\lambda)\, du\right), N\left(\theta \int_0^s (X^\lambda(u) - n^\lambda)^+\, du\right); s \leq t\right\};$$

see Pang et al. [31, §2]. In turn,

$$M^\lambda(t) = M_a^\lambda(t) - M_s^\lambda(t) - M_r^\lambda(t) \tag{11}$$

is itself a square-integrable martingale with respect to $\mathbb{F}^\lambda$. We write

$$X^\lambda(t) = X^\lambda(0) + \lambda t - \mu \int_0^t (X^\lambda(s) \wedge n^\lambda)\, ds - \theta \int_0^t (X^\lambda(s) - n^\lambda)^+\, ds + M^\lambda(t). \tag{12}$$

Letting

$$b^\lambda(x) = \lambda - \mu(x \wedge n^\lambda) - \theta(x - n^\lambda)^+, \tag{13}$$

we arrive at the representation

$$X^\lambda(t) = X^\lambda(0) + \int_0^t b^\lambda(X^\lambda(s))\, ds + M^\lambda(t), \quad t \geq 0.$$

A sequence of "Brownian queues." For each λ, introduce a standard Brownian motion $B = (B(t), t \geq 0)$ and, given an initial condition $Y^\lambda(0)$, consider the diffusion process Y^λ defined through the following stochastic differential equation (SDE):

$$Y^\lambda(t) = Y^\lambda(0) + \int_0^t b^\lambda(Y^\lambda(s))\, ds + \sqrt{2\lambda} B(t). \tag{14}$$

The Lipschitz continuity of the drift guarantees that (given B and $Y^\lambda(0)$) there is a unique solution Y^λ to (14). Furthermore, the process Y^λ is a semi-martingale with respect to the self-filtration of the Brownian motion B.

Remark 2.1 (On the Universality of the Diffusion Coefficient). The diffusion process Y^λ and the B&D process X^λ share the drift function $b^\lambda(\cdot)$. The predictable quadratic variation of the martingale M^λ is given by

$$\langle M^\lambda \rangle(t) = \lambda t + \mu \int_0^t Z^\lambda(s)\, ds + \theta \int_0^t Q^\lambda(s)\, ds.$$

Note that, in steady-state, one has $\lambda = \mu\, \mathbb{E}[Z^\lambda(t)] + \theta\, \mathbb{E}[Q^\lambda(t)]$ for all $0 \leq t \leq \infty$; thus, $\mathbb{E}[\langle M^\lambda \rangle(t)] = 2\lambda t$, and it is intuitively reasonable to construct our universal approximation Y^λ with the diffusion coefficient $\sqrt{2\lambda}$.

Diffusion coefficients that do not depend on the state are prevalent when considering diffusion limits of queueing systems. Indeed, the state independence of the diffusion coefficient extends beyond Markovian queues; see, e.g., the recent work of Kaspi and Ramanan [21] and, specifically, Corollary 5.13 there. Interestingly, a key outcome of our results is that even without scaling, one can ignore the state dependence of the diffusion coefficients for approximations of steady-state metrics. This is further discussed in the next remark.

Remark 2.2 (On the Connection to Strong Approximations). A strong approximation to the B&D process X^λ is the diffusion process $\check{Y}^\lambda$ defined in (6). Formally, from strong approximation theorems Mandelbaum et al. [29], one can choose the Brownian motion (and in turn $\check{Y}^\lambda$) such that a.s. for each $t \geq 0$,

$$\sup_{s \leq t} \left| X^\lambda(s) - \check{Y}^\lambda(s) \right| = \mathcal{O}(\ln \lambda).$$

(The $O(\cdot)$ does depend on t.) Given that $\check{Y}^\lambda$ preserves the state dependence in its diffusion coefficient (see (7)), one expects that replacing our universal diffusion process (14) with $\check{Y}^\lambda$ results in better sample-path bounds. Indeed, an analysis similar to the one in Chen and Yao [12] yields, a.s. for each $t \geq 0$,

$$\sup_{s\leq t} |X^\lambda(s) - Y^\lambda(s)| = \mathcal{O}((\lambda \ln\ln\lambda)^{1/4}(\ln\lambda)^{1/2}).$$

It follows that $\check{Y}^\lambda$ provides more accurate sample path approximations than does Y^λ. However, in steady state this is not the case: Y^λ is as accurate as $\check{Y}^\lambda$, which is appealing from a practical point of view, given the former's relative simplicity.

The steady state of the "Brownian queue." The diffusion process Y^λ has a piecewise-linear drift $b^\lambda(\cdot)$ (13), which "pushes" Y^λ towards the "center" Δ^λ (see (1) and the discussion below it). From this and Browne and Whitt [9], it follows that

Lemma 2.1. *For each $\lambda \in \mathbb{R}_+$, the diffusion process Y^λ has a unique stationary distribution that is also its steady-state distribution. Moreover, $\mathbb{E}[(Y^\lambda(\infty))^m] < \infty$ for each $m \in \mathbb{N}$.*

Henceforth we denote by $Y^\lambda(\infty)$ a random variable having the steady-state distribution of Y^λ. Letting $\beta^\lambda = (n^\lambda\mu - \lambda)/\sqrt{\lambda}$, the density of $Y^\lambda(\infty) - n^\lambda$ is given by (see, e.g., Browne and Whitt [9])

$$\pi^\lambda(x) = \begin{cases} \dfrac{\sqrt{\mu}}{\sqrt{\lambda}} \dfrac{\phi(\sqrt{\mu}(x/\sqrt{\lambda} + \beta^\lambda/\mu))}{\Phi(\beta^\lambda/\sqrt{\mu})} p(\beta^\lambda, \mu, \theta), & \text{if } x \leq 0, \\ \dfrac{\sqrt{\theta}}{\sqrt{\lambda}} \dfrac{\phi(\sqrt{\theta}(x/\sqrt{\lambda} + \beta^\lambda/\theta))}{1 - \Phi(\beta^\lambda/\sqrt{\theta})} (1 - p(\beta^\lambda, \mu, \theta)), & \text{if } x > 0, \end{cases} \tag{15}$$

where

$$p(\beta^\lambda, \mu, \theta) = \left[1 + \sqrt{\frac{\mu}{\theta}} \frac{\phi(\beta^\lambda/\sqrt{\mu})}{\Phi(\beta^\lambda/\sqrt{\mu})} \frac{1 - \Phi(\beta^\lambda/\sqrt{\theta})}{\phi(\beta^\lambda/\sqrt{\theta})}\right]^{-1},$$

and ϕ and Φ are, respectively, the standard normal density and cumulative distribution functions.

Significantly, the specific expression of π^λ is not needed for the theory that we are developing (our excursion-based framework). It plays a role only in concrete calculations, for example when solving optimization problems associated with Erlang-A; see §5. Indeed, in such calculations, one takes advantage of the form of π^λ, which is more amenable to analysis than the steady-state of X^λ.

The remainder of the paper. We state the main result and important corollaries in §3. Section 4 is dedicated to the proof of the main result. Section 5 then studies implications of the universal approximation to two well-studied optimization problems. Finally, §6 provides concluding remarks and discusses possible extensions of our framework. Throughout the remainder of the paper, we state and prove the key results while relegating proofs of auxiliary lemmas to the appendix.

3. Main results.

Recall (1) where

$$\Delta^\lambda = \frac{\lambda}{\mu} - \left(\frac{\lambda}{\mu} - n^\lambda\right)^+ \left(1 - \frac{\mu}{\theta}\right).$$

To simplify notation, we assume without loss of generality that $\Delta^\lambda \in \mathbb{N}$; see Remark 4.2. Define

$$\tilde{X}^\lambda(t) = X^\lambda(t) - \Delta^\lambda \qquad \text{and} \qquad \tilde{Y}^\lambda(t) = Y^\lambda(t) - \Delta^\lambda.$$

Definition 3.1 (Subpolynomial Functions). A sequence of differentiable functions $f^\lambda\colon \mathbb{R} \to \mathbb{R}$ is said to be uniformly subpolynomial of order $m \in \{1, 2, \dots\}$ if there exist absolute constants a_1, a_2 such that, for all λ,

$$|f^\lambda(x)| \leq a_1\sqrt{\lambda}^m + a_2|x|^m \qquad \text{and} \qquad |(f^\lambda)^{(1)}(x)| \leq a_1\sqrt{\lambda}^{m-1} + a_2|x|^{m-1}.$$

It is said to be uniformly subpolynomial of order $m = 0$ if there exists an absolute constant a_3 and a sequence $\{a^\lambda\}$ such that, for all λ,

$$|f^\lambda(x)| \vee |(f^\lambda)^{(1)}(x)| \leq a_3 \quad \forall\, x \in \mathbb{R} \qquad \text{and} \qquad (f^\lambda)^{(1)}(x) = 0 \quad \forall\, x \notin (a^\lambda, a^\lambda + 1).$$

We let $\mathcal{S}_m$ denote the family of uniformly subpolynomial function-sequences of order m.

The following is the main result of our paper. It is proved in §4.

THEOREM 1. *Fix $m \in \mathbb{N}$ and $\{f^\lambda\} \in \mathcal{S}_m$. Then,*

$$\mathbb{E}[f^\lambda(\tilde{X}^\lambda(\infty))] - \mathbb{E}[f^\lambda(\tilde{Y}^\lambda(\infty))] = \mathcal{O}(\sqrt{\lambda}^{m-1}).$$

We next state three corollaries that draw implications of practical interest from Theorem 1. All three corollaries are proved in the appendix. The performance metrics that we consider—queue length, probability of delay, and variance—are themselves not subpolynomial functions of $\tilde{X}^\lambda$ but implications of Theorem 1 can be drawn for these via relatively straightforward manipulations.

The first corollary is concerned with the expected steady-state queue length and is instrumental in our exploration of optimization problems in §5. Recall (10) that $Q^\lambda(\infty) = (X^\lambda(\infty) - n^\lambda)^+$ and, for the proposed approximation, define $\tilde{Q}^\lambda(\infty) = (Y^\lambda(\infty) - n^\lambda)^+$, which, using (15), satisfies

$$\mathbb{E}[\tilde{Q}^\lambda(\infty)] = \frac{\sqrt{\lambda}}{\sqrt{\theta}}[1 - p(\beta^\lambda, \mu, \theta)]\big[h(\beta^\lambda/\sqrt{\theta}) - \beta^\lambda/\sqrt{\theta}\big], \tag{16}$$

with h being the hazard rate of the standard normal distribution; i.e., $h(x) = \phi(x)/(1 - \Phi(x))$.

COROLLARY 1.

$$\mathbb{E}[Q^\lambda(\infty)] - \mathbb{E}[\tilde{Q}^\lambda(\infty)] = \mathcal{O}(1).$$

EXAMPLE 3.1 (QUEUE LENGTH). (i) *Fixed λ, varying n*: Consider the Erlang-A queue with $\lambda = 500$, $\mu = 1$, $\theta = 0.5$. The top graph in Figure 1 displays $\mathbb{E}[Q^\lambda(\infty)]$ versus the universal approximation $\mathbb{E}[\tilde{Q}^\lambda(\infty)]$ as a function of the number of servers n. The bottom graph displays the errors $\mathbb{E}[\tilde{Q}^\lambda(\infty)] - \mathbb{E}[Q^\lambda(\infty)]$, again as a function of n.

(ii) *Fixed ρ, varying λ*: Here we consider various values of the offered utilization $\rho^\lambda = \lambda/(n^\lambda\mu)$. For each fixed value of the utilization, we vary λ between 20 and 2,000 and increase n^λ as needed to keep ρ^λ fixed. We then plot the absolute errors $|\mathbb{E}[\tilde{Q}^\lambda(\infty)] - \mathbb{E}[Q^\lambda(\infty)]|$ as well as the function $1/\sqrt{\lambda}$. The result is displayed in Figure 2 where the solid line corresponds to the absolute error and the dashed line to $1/\sqrt{\lambda}$. This numerical experiment suggests that the error may be, in fact, $\mathcal{O}(1/\sqrt{\lambda})$ and, in particular, smaller than the $\mathcal{O}(1)$ predicted by Corollary 1. The next case, shows, however, that the bound $\mathcal{O}(1)$ is tight.

(iii) *Varying ρ^λ with λ (QED)*: We vary λ between 20 and 2,000 and set the capacity to a square-root staffing $n^\lambda = \lceil \lambda/\mu + \beta\sqrt{\lambda/\mu} \rceil$ with $\beta = 1$. The result is displayed in Figure 3 where the upper graph displays the queues

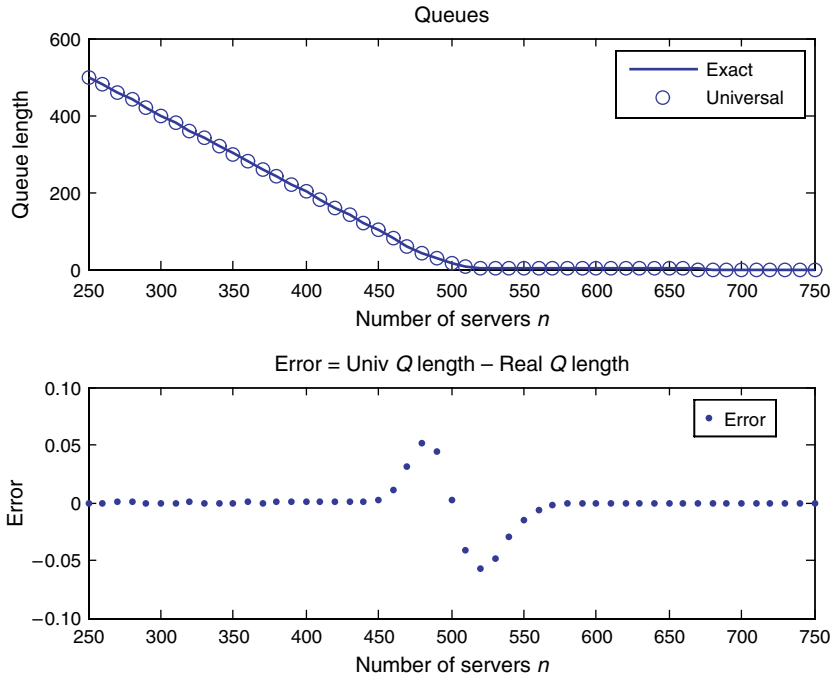

FIGURE 1. Expected queue approximation: fixed λ, varying n ($250 \le n \le 750$).

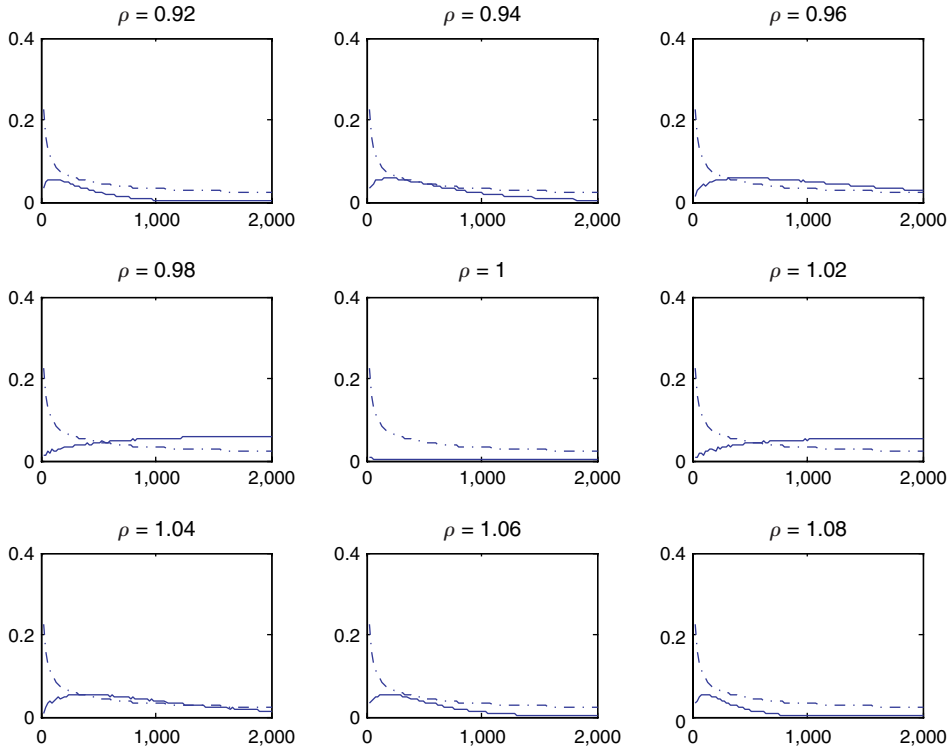

FIGURE 2. Expected queue approximation: fixed ρ, varying λ ($20 \le \lambda \le 2{,}000$).

in the Erlang-A queue and in the universal approximation and the bottom graph displays the absolute error. Here, the error approaches a constant as λ grows large.

REMARK 3.1 (ON PARAMETERS IN NUMERICAL EXPERIMENTS). In the above and subsequent examples, we are using $\mu > \theta$: in words, average patience exceeds average service times. Based on our practical experience (SEELab [34]), such a relation is prevalent, with $\mu/\theta = 2$ being not uncommon. We should add that, based on

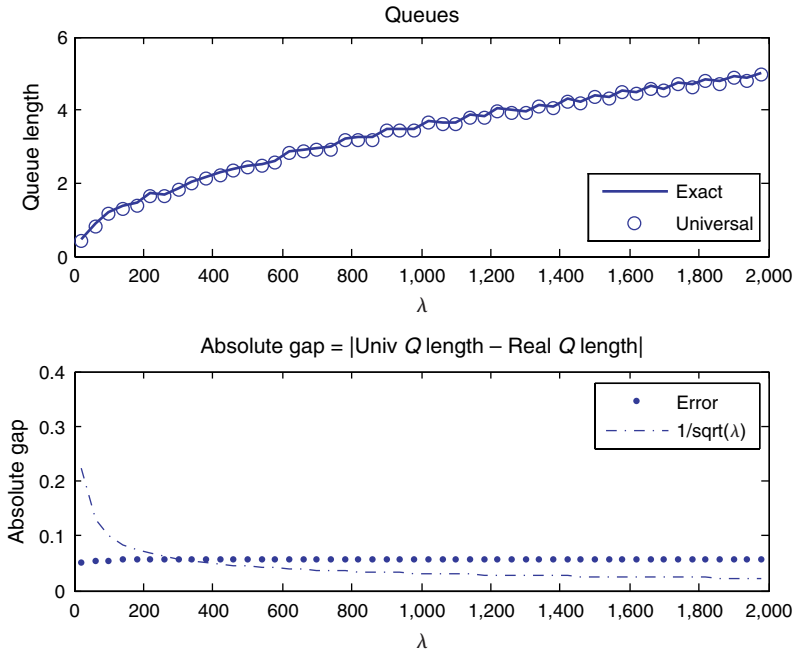

FIGURE 3. Expected queue approximation: varying ρ^λ with λ (QED) ($20 \le \lambda \le 2{,}000$).

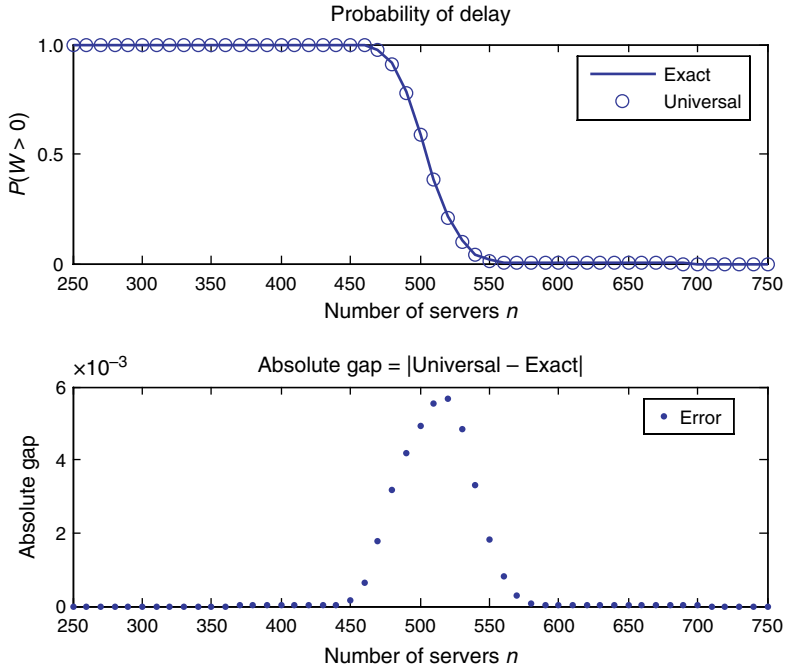

FIGURE 4. Probability of delay: fixed λ, varying n ($250 \le n \le 750$).

extensive numerical experiments that we performed, the numerical outcomes reported here are representative of the full range of θ values above and below μ (with the exception of an overloaded system with $\theta << \mu$, which approximates an unstable Erlang-C queue).

COROLLARY 2.

$$\sup_{x\ge 0}\left|\mathbb{P}\{X^\lambda(\infty)\ge x\}-\mathbb{P}\{Y^\lambda(\infty)\ge x\}\right| = \mathcal{O}(\sqrt{\lambda}^{-1}), \tag{17}$$

or equivalently, for any sequence $\{a^\lambda\}$,

$$\mathbb{P}\{X^\lambda(\infty)\ge a^\lambda\}-\mathbb{P}\{Y^\lambda(\infty)\ge a^\lambda\} = \mathcal{O}(\sqrt{\lambda}^{-1}). \tag{18}$$

EXAMPLE 3.2 (PROBABILITY OF DELAY). An important application of Corollary 2 is the probability of delay, corresponding to $a^\lambda = n^\lambda$ in (18) for each λ. In this example we compare $\mathbb{P}\{X^\lambda(\infty)\ge n^\lambda\}$ to $\mathbb{P}\{Y^\lambda(\infty)\ge n^\lambda\}$.

(i) *Fixed* λ, *varying* n: We fix the parameters as in Example 3.1(i). The top graph in Figure 4 displays $\mathbb{P}\{X^\lambda(\infty)\ge n^\lambda\}$ versus its universal approximation $\mathbb{P}\{Y^\lambda(\infty)\ge n^\lambda\}$. The bottom graph displays the absolute error $|\mathbb{P}\{X^\lambda(\infty)\ge n^\lambda\}-\mathbb{P}\{Y^\lambda(\infty)\ge n^\lambda\}|$.

(ii) *Fixed* ρ, *varying* λ: We fix the parameters as in Example 3.1(ii) but replace the queue length with the delay probability. The result is displayed in Figure 5.

(iii) *Varying* ρ^λ *with* λ (*QED*): We repeat the setting of Example 3.1(iii). In Figure 6 it is seen that the bound $\mathcal{O}(1/\sqrt{\lambda})$ is tight.

Our last corollary compares the variance of $Q^\lambda(\infty)$ to that of $\tilde{Q}^\lambda(\infty)$.

COROLLARY 3.

$$\mathrm{Var}(Q^\lambda(\infty)) - \mathrm{Var}(\tilde{Q}^\lambda(\infty)) = \mathcal{O}(\sqrt{\lambda}).$$

EXAMPLE 3.3 (VARIANCE OF QUEUE LENGTH). In this example we compare $\mathrm{Var}(Q^\lambda(\infty))$ to $\mathrm{Var}(\tilde{Q}^\lambda(\infty))$.

(i) *Fixed* λ, *varying* n: We fix the parameters to be as in Example 3.1(i). The top graph in Figure 7 displays $\mathrm{Var}(Q^\lambda(\infty))$ versus its universal approximation $\mathrm{Var}(\tilde{Q}^\lambda(\infty))$. The bottom graph displays the error $\mathrm{Var}(\tilde{Q}^\lambda(\infty)) - \mathrm{Var}(Q^\lambda(\infty))$.

(ii) *Fixed* ρ, *varying* λ: We fix the parameters as in Example 3.1(ii) but replace the expectation of the queue length with its variance. The result is displayed in Figure 8.

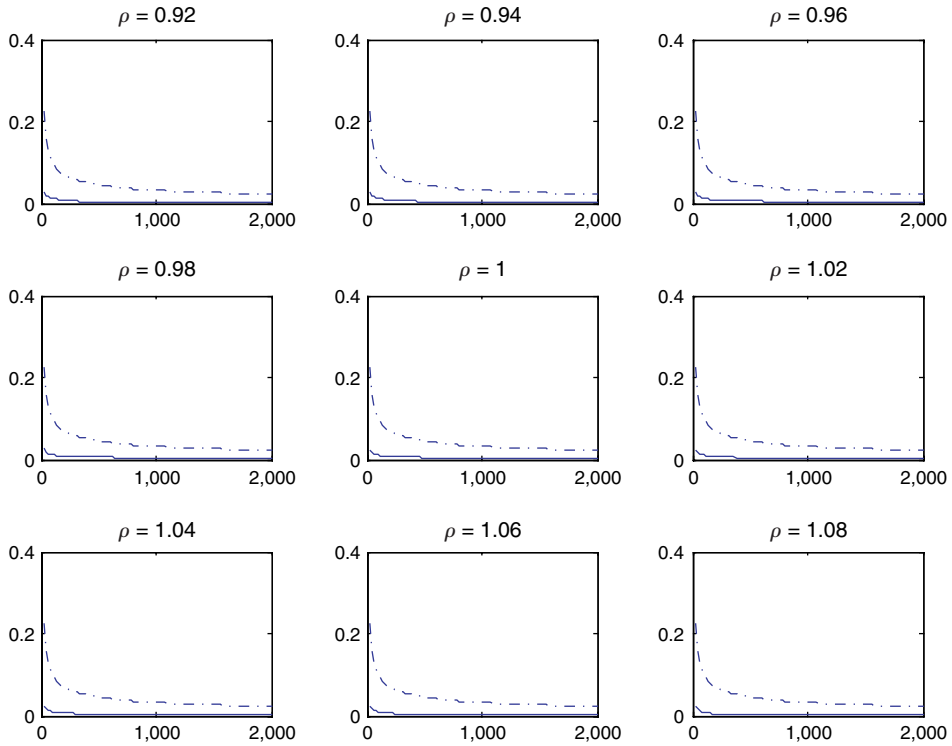

FIGURE 5. Probability of delay: fixed ρ, varying λ ($20 \le \lambda \le 2{,}000$).

(iii) *Varying ρ^λ with λ (QED)*: We fix the parameters as in Example 3.1(iii). As before, it is seen in Figure 9 that the bound $\mathcal{O}(\sqrt{\lambda})$ is tight.

4. Proof of the main result. This section contains the proof of Theorem 1. It is divided into three subsections. In §4.1 we define regeneration times for $\tilde{X}^\lambda$ and $\tilde{Y}^\lambda$—in both cases, these are based on return times to Δ^λ. We then distinguish between cycles that are above Δ^λ (upper excursions) and those that are below (lower excursions). Sections 4.2 and 4.3 are then devoted, respectively, to the study of the upper and lower excursions.

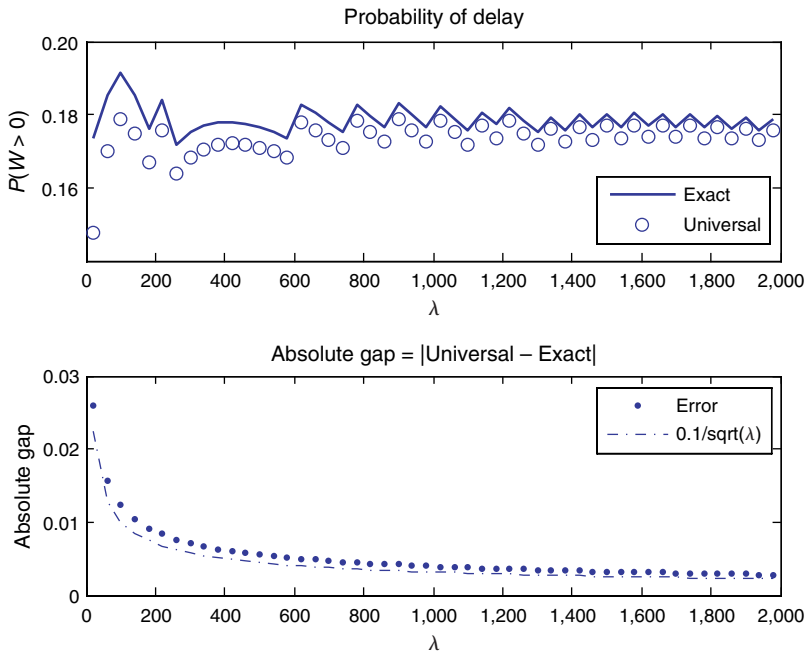

FIGURE 6. Probability of delay: varying ρ^λ with λ (QED) ($20 \le \lambda \le 2{,}000$).

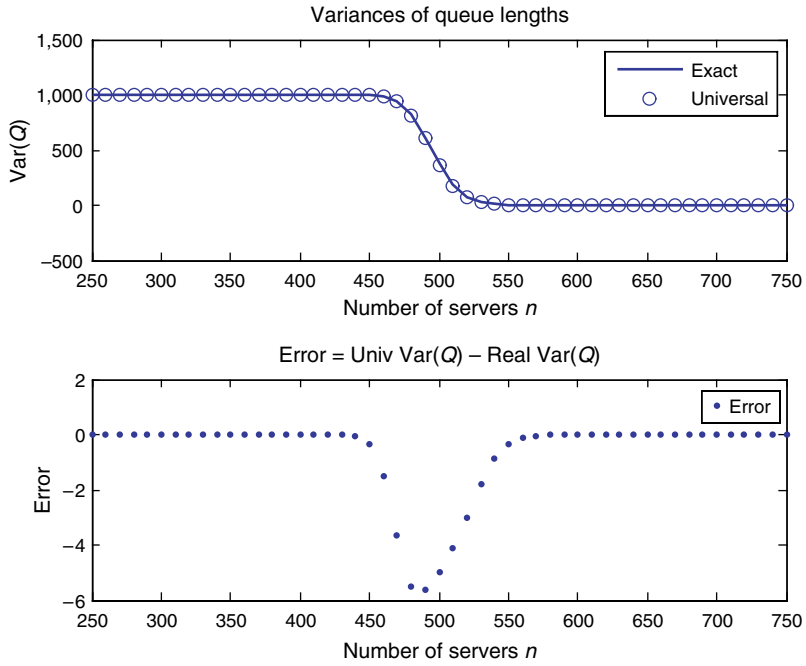

FIGURE 7. Variance of queue length: fixed λ, varying n ($250 \le n \le 750$).

4.1. The regenerative structure. A starting point for our analysis is the intimate relationship between regenerative structure and steady-state distributions. Recall that $\tilde{X}^\lambda = X^\lambda - \Delta^\lambda$, $\tilde{Y}^\lambda = Y^\lambda - \Delta^\lambda$, and Δ^λ is assumed to be integer. This assumption is made without loss of generality; see Remark 4.2.

Whereas consecutive visits to a point $y \in \mathbb{N}$ constitute a renewal process for the process $\tilde{X}^\lambda$ on the basis of which a regenerative process can be constructed, this is not so for the diffusion process $\tilde{Y}^\lambda$; see, e.g., Asmussen [3, p. 174]. Because we wish to compare the B&D process and the diffusion process, we use a

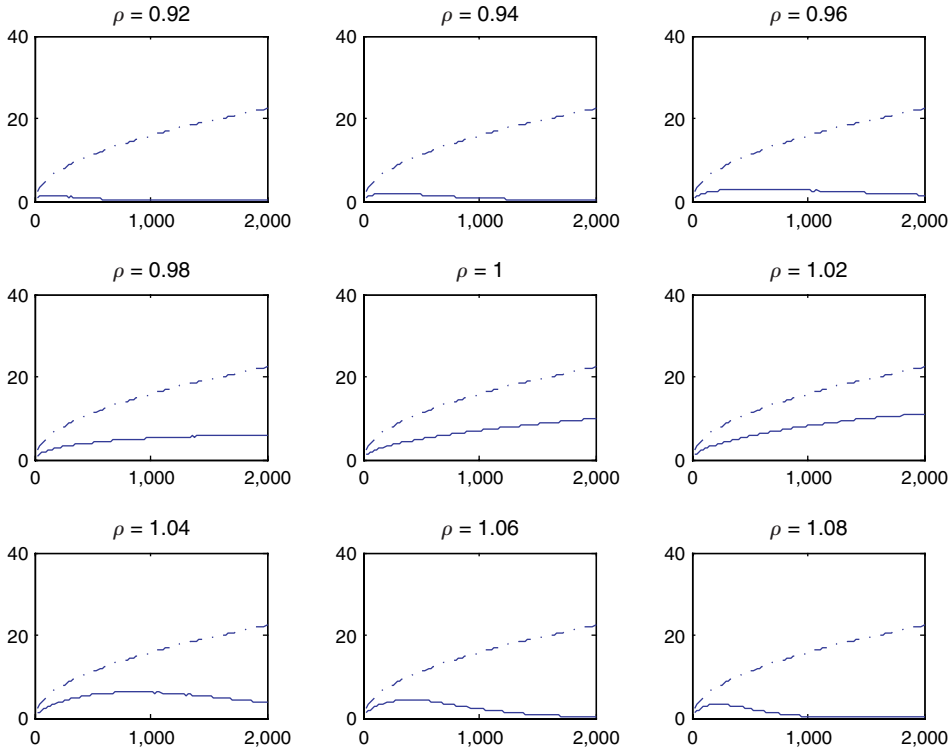

FIGURE 8. Variance of queue length: fixed ρ, varying λ ($20 \le \lambda \le 2{,}000$).

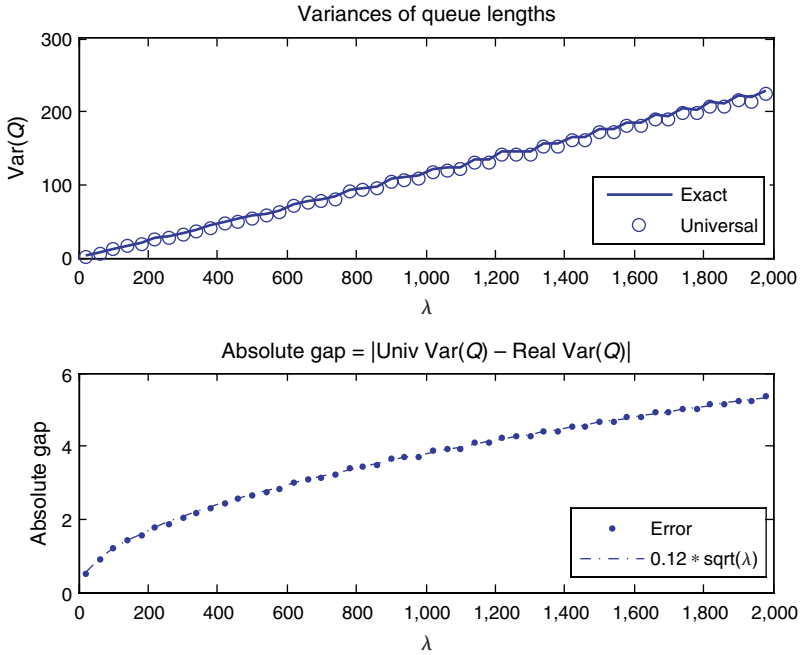

FIGURE 9. Variance of queue length: varying ρ^λ with λ (QED) ($20 \le \lambda \le 2{,}000$).

common definition for the underlying renewal process. We define (for both) a regeneration as the first visit to state 1 after visiting state 0. Formally, let

$$\tau_u^\lambda(s) := \inf\{t \ge s\colon \tilde{X}^\lambda(t) = 0\}, \qquad \tau_l^\lambda(s) := \inf\{t \ge s\colon \tilde{X}^\lambda(t) = 1\},$$

and

$$\tau^\lambda(s) = \tau_l^\lambda(\tau_u^\lambda(s)).$$

Define similarly

$$\tilde{\tau}_u^\lambda(s) := \inf\{t \ge s\colon \tilde{Y}^\lambda(t) = 0\}, \qquad \tilde{\tau}_l^\lambda(s) = \inf\{t \ge s\colon \tilde{Y}^\lambda(t) = 1\},$$

and

$$\tilde{\tau}^\lambda = \tilde{\tau}_l^\lambda(\tilde{\tau}_u^\lambda(s)).$$

For obvious reasons we refer henceforth to the interval $[0, \tau_u^\lambda)$ (respectively, $[0, \tilde{\tau}_u^\lambda)$) as the *upper excursion* for $\tilde{X}^\lambda$ (respectively, $\tilde{Y}^\lambda$) and to the interval $[\tau_u^\lambda, \tau^\lambda)$ (respectively, $[\tilde{\tau}_u^\lambda, \tilde{\tau}^\lambda)$) as the *lower excursion* for $\tilde{X}^\lambda$ (respectively, $\tilde{Y}^\lambda$).

The composition of stopping times is well defined. Both $\tilde{Y}^\lambda$ and $\tilde{X}^\lambda$ are strong Markov processes (e.g., Asmussen [3, Theorem 1.1], Karatzas and Shreve [19, Theorem 4.20]) and have a regenerative structure with $\tilde{\tau}^\lambda$ and τ^λ being the regeneration times for $\tilde{Y}^\lambda$ and $\tilde{X}^\lambda$, respectively. Define $T_0^\lambda = 0$ and recursively define $T_{i+1}^\lambda = \tau^\lambda(T_i^\lambda)$ (if $T_i^\lambda = \infty$, one sets $T_{i+1}^\lambda = \infty$). With $\tilde{X}^\lambda(0) = 1$, the sequence $T_0^\lambda < T_1^\lambda < \cdots$ constitutes an (undelayed) renewal process. A renewal process for $\tilde{Y}^\lambda$ is constructed similarly.

To simplify notation, let

$$\tau_u^\lambda = \tau_u^\lambda(0),\ \tau_l^\lambda = \tau_l^\lambda(0),\ \tau^\lambda = \tau^\lambda(0) \qquad \text{and} \qquad \tilde{\tau}_u^\lambda = \tilde{\tau}_u^\lambda(0),\ \tilde{\tau}_l^\lambda = \tilde{\tau}_l^\lambda(0),\ \tilde{\tau}^\lambda = \tilde{\tau}^\lambda(0).$$

Both $\tilde{X}^\lambda$ and $\tilde{Y}^\lambda$ have a positive drift "pushing" them up when sufficiently smaller than 0, and since $\theta > 0$, they have a negative drift pushing them down when sufficiently greater than 0, so that one expects τ^λ and $\tilde{\tau}^\lambda$ to be "well-behaved." This is formally justified by the following lemma.

LEMMA 4.1. *Fix $\lambda \in \mathbb{R}_+$. Then, there exists a constant $\vartheta_0 > 0$ (possibly depending on λ) such that $\mathbb{E}_y[e^{\vartheta_0 \tau^\lambda}] < \infty$ for all $y \in \mathbb{N}$ and $\mathbb{E}_y[e^{\vartheta_0 \tilde{\tau}^\lambda}] < \infty$ for all $y \in \mathbb{R}$. In turn, $\mathbb{E}_1[(\tau^\lambda)^m] < \infty$ and $\mathbb{E}_1[(\tilde{\tau}^\lambda)^m] < \infty$, for each $m \in \mathbb{N}$.*

Provided that f is integrable under the steady-state distributions of $\tilde{X}^\lambda$ and $\tilde{Y}^\lambda$, we have

$$\mathbb{E}[f(\tilde{X}^\lambda(\infty))] = \frac{\mathbb{E}_1[\int_0^{\tau^\lambda} f(\tilde{X}^\lambda(s))\,ds]}{\mathbb{E}_1[\tau^\lambda]} \qquad \text{and} \qquad \mathbb{E}[f(\tilde{Y}^\lambda(\infty))] = \frac{\mathbb{E}_1[\int_0^{\tilde{\tau}^\lambda} f(\tilde{Y}^\lambda(s))\,ds]}{\mathbb{E}_1[\tilde{\tau}^\lambda]}.$$

From the strong Markov property it follows that for any such function f,

$$\mathbb{E}_1\left[\int_{\tau_u^\lambda}^{\tau^\lambda} f(\tilde{X}^\lambda(s))\,ds\right] = \mathbb{E}_0\left[\int_0^{\tau_l^\lambda} f(\tilde{X}^\lambda(s))\,ds\right]$$

and, particularly, $\mathbb{E}_1[\tau^\lambda - \tau_u^\lambda] = \mathbb{E}_0[\tau_l^\lambda]$. Similar observations apply to $\tilde{Y}^\lambda$ with τ^λ, τ_u^λ, τ_l^λ replaced by $\tilde{\tau}^\lambda$, $\tilde{\tau}_u^\lambda$, and $\tilde{\tau}_l^\lambda$. For $y \geq 0$, define

$$V_u^\lambda(f,y) = \mathbb{E}_y\left[\int_0^{\tau_u^\lambda} f(\tilde{X}^\lambda(s))\,ds\right] \qquad \text{and} \qquad \mathscr{V}_u^\lambda(f,y) = \mathbb{E}_y\left[\int_0^{\tilde{\tau}_u^\lambda} f(\tilde{Y}^\lambda(s))\,ds\right],$$

and for $y \leq 0$,

$$V_l^\lambda(f,y) = \mathbb{E}_y\left[\int_0^{\tau_l^\lambda} f(\tilde{X}^\lambda(s))\,ds\right] \qquad \text{and} \qquad \mathscr{V}_l^\lambda(f,y) = \mathbb{E}_y\left[\int_0^{\tilde{\tau}_l^\lambda} f(\tilde{Y}^\lambda(s))\,ds\right].$$

Thus,

$$\begin{aligned}\mathbb{E}[f(\tilde{X}^\lambda(\infty))] &= \frac{\mathbb{E}_1[\int_0^{\tau_u^\lambda} f(\tilde{X}^\lambda(s))\,ds + \int_{\tau_u^\lambda}^{\tau^\lambda} f(\tilde{X}^\lambda(s))\,ds]}{\mathbb{E}_1[\tau^\lambda]} \\ &= \frac{V_u^\lambda(f,1) + V_l^\lambda(f,0)}{\mathbb{E}_1[\tau^\lambda]},\end{aligned}$$

and

$$\mathbb{E}[f(\tilde{Y}^\lambda(\infty))] = \frac{\mathscr{V}_u^\lambda(f,1) + \mathscr{V}_l^\lambda(f,0)}{\mathbb{E}_1[\tilde{\tau}^\lambda]}.$$

Consequently

$$\begin{aligned}\mathbb{E}[f(\tilde{X}^\lambda(\infty))] - \mathbb{E}[f(\tilde{Y}^\lambda(\infty))] = {} & \frac{V_u^\lambda(f,1) - \mathscr{V}_u^\lambda(f,1)}{\mathbb{E}_1[\tilde{\tau}^\lambda]} + \frac{V_l^\lambda(f,0) - \mathscr{V}_l^\lambda(f,0)}{\mathbb{E}_1[\tilde{\tau}^\lambda]} \\ & + \frac{V_u^\lambda(f,1) + V_l^\lambda(f,0)}{\mathbb{E}_1[\tau^\lambda]} - \frac{V_u^\lambda(f,1) + V_l^\lambda(f,0)}{\mathbb{E}_1[\tilde{\tau}^\lambda]},\end{aligned}$$

and

$$\begin{aligned}|\,\mathbb{E}[f(\tilde{X}^\lambda(\infty))] - \mathbb{E}[f(\tilde{Y}^\lambda(\infty))]| \leq {} & \frac{|\mathscr{V}_u^\lambda(f,1) - V_u^\lambda(f,1)|}{\mathbb{E}_1[\tilde{\tau}^\lambda]} + \frac{|\mathscr{V}_l^\lambda(f,0) - V_l^\lambda(f,0)|}{\mathbb{E}_1[\tilde{\tau}^\lambda]} \\ & + \frac{|V_u^\lambda(f,1) + V_l^\lambda(f,0)|}{\mathbb{E}_1[\tau^\lambda]\mathbb{E}_1[\tilde{\tau}^\lambda]}\,|\,\mathbb{E}_1[\tau^\lambda] - \mathbb{E}_1[\tilde{\tau}^\lambda]|. \qquad (19)\end{aligned}$$

Theorem 1 is a direct corollary of the above together with the bounds provided in Theorem 2 below.

THEOREM 2. *Fix $m \in \mathbb{N}$ and $\{f^\lambda\} \in \mathscr{S}_m$. Then,*

$$V_u^\lambda(f^\lambda,1) = \mathscr{O}(\sqrt{\lambda}^{m-1}), \qquad (\mathbb{E}_1[\tau_u^\lambda])^{-1} = \mathscr{O}(\sqrt{\lambda}), \qquad (20)$$

$$V_u^\lambda(f^\lambda,1) - \mathscr{V}_u(f^\lambda,1) = \mathscr{O}(\sqrt{\lambda}^{m-2}), \qquad (21)$$

$$V_l^\lambda(f^\lambda,0) = \mathscr{O}(\sqrt{\lambda}^{m-1}), \qquad (\mathbb{E}_0[\tau_l^\lambda])^{-1} = \mathscr{O}(\sqrt{\lambda}), \qquad (22)$$

$$V_l^\lambda(f^\lambda,0) - \mathscr{V}_l^\lambda(f^\lambda,0) = \mathscr{O}(\sqrt{\lambda}^{m-2}). \qquad (23)$$

Setting $f^\lambda \equiv \mathbf{1}$ in (20) and (22) gives

$$\mathbb{E}_1[\tau_u^\lambda] = \mathcal{O}(\sqrt{\lambda}^{-1}) \qquad \text{and} \qquad \mathbb{E}_0[\tau_l^\lambda] = \mathcal{O}(\sqrt{\lambda}^{-1}).$$

A further immediate corollary of Theorem 2 is that

$$\mathcal{V}_u(f^\lambda, 1) = \mathcal{O}(\sqrt{\lambda}^{m-1}), \qquad \mathcal{V}_l(f^\lambda, 0) = \mathcal{O}(\sqrt{\lambda}^{m-1}),$$

so that setting again $f^\lambda \equiv \mathbf{1}$,

$$\mathbb{E}_1[\tilde{\tau}_u^\lambda] = \mathcal{O}(\sqrt{\lambda}^{-1}) \qquad \text{and} \qquad \mathbb{E}_0[\tilde{\tau}_l^\lambda] = \mathcal{O}(\sqrt{\lambda}^{-1}).$$

Finally, noting that $\mathbb{E}_1[\tau^\lambda] \geq \mathbb{E}_1[\tau_u^\lambda]$ and $\mathbb{E}_1[\tilde{\tau}^\lambda] \geq \mathbb{E}_1[\tilde{\tau}_u^\lambda]$, (20) guarantees that

$$(\mathbb{E}_1[\tau^\lambda])^{-1} = \mathcal{O}(\sqrt{\lambda}) \qquad \text{and} \qquad (\mathbb{E}_1[\tilde{\tau}^\lambda])^{-1} = \mathcal{O}(\sqrt{\lambda}).$$

The decomposition in (19) allows us to conduct a separate analysis for the *upper excursion* and the *lower excursion*. Section 4.2 is dedicated to the former, and §4.3 is dedicated to the latter. We conclude this subsection with a remark about limit interchange.

Remark 4.1 (Implications to Limit Interchange). Given the process-convergence $\hat{X}^\lambda := \tilde{X}^\lambda/\sqrt{\lambda} \Rightarrow \hat{X}$ (see §1), one expects that $\hat{X}^\lambda(\infty) := \tilde{X}^\lambda(\infty)/\sqrt{\lambda} \Rightarrow \hat{X}(\infty)$. This conclusion is proved via an interchange-of-limits argument that, as pointed out in Ward [38], has been proved in the QED and ED regimes but not yet in the NDS regime. The key step in establishing limit interchange is proving that the family of random variables $\{\hat{X}^\lambda(\infty), \lambda \geq 0\}$ is tight as a sequence of random variables in $\mathbb{R}$.

Such tightness is a byproduct of our results. Indeed, by Theorem 2, there exists a constant c such that for all λ, $\mathbb{E}[(\tilde{X}^\lambda(\infty))^2] \leq c\lambda$. In particular, $\limsup_{\lambda\to\infty} \mathbb{E}[(\hat{X}^\lambda(\infty))^2] < \infty$, implying that the scaled sequence $\{\hat{X}^\lambda(\infty), \lambda \geq 0\}$ is not only tight but in fact uniformly integrable. In both the NDS and QED regimes $\Delta^\lambda = n^\lambda + \mathcal{O}(\sqrt{\lambda})$ so that the uniform integrability of $\hat{X}^\lambda(\infty)$ implies that of $\{(X^\lambda(\infty) - n^\lambda)/\sqrt{\lambda}, \lambda \geq 0\}$, which is the centering used in the literature; see Ward [38]. For the ED regime, we center around $\Delta^\lambda = n^\lambda + (\lambda - n^\lambda\mu)/\theta$, which is $\mathcal{O}(\sqrt{\lambda})$ away from that of Whitt [40].

4.2. Upper excursion. Propositions 3 and 4 prove, respectively, Equations (20) and (21) in Theorem 2. They are proved in §§4.2.1 and 4.2.2, respectively.

Proposition 3 (Order Bounds). *Fix $m \in \mathbb{N}$ and $\{f^\lambda\} \in \mathcal{S}_m$. Then,*

$$(\mathbb{E}_1[\tau_u^\lambda])^{-1} = \mathcal{O}(\sqrt{\lambda}) \qquad \textit{and} \qquad V_u^\lambda(f^\lambda, 1) = \mathcal{O}(\sqrt{\lambda}^{m-1}).$$

Proposition 4 (Gap Bounds). *Fix $m \in \mathbb{N}$ and $\{f^\lambda\} \in \mathcal{S}_m$. Then,*

$$V_u^\lambda(f^\lambda, 1) - \mathcal{V}_u^\lambda(f^\lambda, 1) = \mathcal{O}(\sqrt{\lambda}^{m-2}).$$

Let

$$\ell^\lambda(x) = -\lambda + \mu((x + \Delta^\lambda) \wedge n^\lambda) + \theta(x + \Delta^\lambda - n^\lambda)^+. \tag{24}$$

Starting at $x \geq 0$ and using (12) we have, for $t \leq \tau_u^\lambda$, that

$$\tilde{X}^\lambda(t) = \tilde{X}^\lambda(0) - \int_0^t \ell^\lambda(\tilde{X}^\lambda(s))\, ds + M^\lambda(t), \tag{25}$$

where M^λ is as in (11). For $\tilde{Y}^\lambda$ and $t \leq \tilde{\tau}_u^\lambda$, we have

$$\tilde{Y}^\lambda(t) = \tilde{Y}^\lambda(0) - \int_0^t \ell^\lambda(\tilde{Y}^\lambda(s))\, ds + \sqrt{2\lambda}B(t).$$

Recalling that $\ell^\lambda(0) = 0$, it follows that there exist absolute constants $\vartheta_i > 0$, $i = 1, 2, 3$ such that $\vartheta_1 x \leq \ell^\lambda(x) \leq \vartheta_2 x$. In fact, it will suffice for our proofs that

$$-\vartheta_3\sqrt{\lambda} + \vartheta_1 x \leq \ell^\lambda(x) \leq \vartheta_3\sqrt{\lambda} + \vartheta_2 x, \quad x \geq 0. \tag{26}$$

Having the proofs rely only on this weaker bound will facilitate the extension of our proofs to the NDS regime.

The following two simple lemmas will be useful in the proofs of Propositions 3 and 4. Here, recall that $E(\lambda t)$ is the number of arrivals by time t in the λth system.

LEMMA 4.2. *Fix $\lambda, t \in \mathbb{R}_+$, $y \in \mathbb{N}$, and a nonnegative nondecreasing function $g(\cdot)$ such that $\mathbb{E}[g(y+E(\lambda t))] < \infty$. Then,*

$$\mathbb{E}_y\left[\int_0^{t\wedge\tau_u^\lambda} g(\tilde{X}^\lambda(s))\,ds\right] < \infty.$$

LEMMA 4.3. *Fix λ, $t \in \mathbb{R}_+$, $y \in \mathbb{N}$, and a function $g(\cdot)$. If*

$$\mathbb{E}_y\left[\int_0^{t\wedge\tau_u^\lambda} (2\lambda + \ell^\lambda(\tilde{X}^\lambda(s)))(g^2(\tilde{X}^\lambda(s)))^2\,ds\right] < \infty, \tag{27}$$

then

$$\mathbb{E}_y\left[\int_0^{t\wedge\tau_u^\lambda} g(\tilde{X}^\lambda(s-))\,dM^\lambda(s)\right] = 0,$$

and

$$\mathbb{E}_y\left[\sum_{s\le t\wedge\tau_u^\lambda:\,|\Delta\tilde{X}^\lambda(s)|>0} g(\tilde{X}^\lambda(s-))(\Delta\tilde{X}^\lambda(s-))^2\right] = \mathbb{E}_y\left[\int_0^{t\wedge\tau_u^\lambda} (2\lambda + \ell^\lambda(\tilde{X}^\lambda(s)))g(\tilde{X}^\lambda(s))\,ds\right]. \tag{28}$$

Condition (27) holds, in particular, if $g(\cdot)$ is nonnegative, nondecreasing, and such that $\mathbb{E}[(g(y+E(\lambda t)))^2 \cdot (y+E(\lambda t))] < \infty$.

Note that since $E(\lambda t)$ has finite moments of all orders, both Lemmas 4.2 and 4.3 hold with g polynomial. Moreover, the function g can be replaced by any (not necessarily nondecreasing) function f such that $|f(x)| \le h(x)$, for all $x \ge 0$, where h satisfies the conditions of Lemmas 4.2 and 4.3.

4.2.1. Proof of Proposition 3. Starting at $x \ge 0$, $\tilde{X}^\lambda$ has, on $[0, \tau_u^\lambda)$, the law of a B&D process on the positive integers, with birth rate λ in all states and death rate $\lambda + \ell^\lambda(x)$ when in state $x > 0$. Let $U^\lambda = (U^\lambda(t), t \ge 0)$ be a B&D process with these birth and death rates and observe that $\lambda < \lambda + \ell^\lambda(x)$ for all $x > 0$ so that U^λ admits a steady-state distribution. We have the following simple lemmas:

LEMMA 4.4. *For any nondecreasing function $f\colon \mathbb{R} \to \mathbb{R}$,*

$$\mathbb{E}[f(U^\lambda(\infty))] \le \frac{\mathbb{E}_1[\int_0^{\tau_u^\lambda} f(\tilde{X}^\lambda(s))\,ds]}{\mathbb{E}_1[\tau_u^\lambda]} \le \mathbb{E}[f(U^\lambda(\infty)+1)].$$

LEMMA 4.5. *Fix $\lambda \in \mathbb{R}_+$. Then,*

$$\mathbb{E}_1\left[\int_0^{\tau_u^\lambda} \ell^\lambda(\tilde{X}^\lambda(s))\,ds\right] = 1.$$

Recalling that ℓ^λ is nondecreasing, taking $f = \ell^\lambda$ in Lemma 4.4, and using Lemma 4.5, we conclude that

$$\mathbb{E}[\ell^\lambda(U^\lambda(\infty))] \le \frac{1}{\mathbb{E}_1[\tau_u^\lambda]} \le \mathbb{E}[\ell^\lambda(U^\lambda(\infty)+1)],$$

which, in turn, proves that

$$\big(\mathbb{E}[\ell^\lambda(U^\lambda(\infty)+1)]\big)^{-1} \le \mathbb{E}_1[\tau_u^\lambda] \le \big(\mathbb{E}[\ell^\lambda(U^\lambda(\infty))]\big)^{-1}. \tag{29}$$

The following then provides bounds for the left- and right-hand sides.

LEMMA 4.6. *There exist absolute constants ϑ_l, ϑ_u such that for all $\lambda \in \mathbb{R}_+$,*

$$\vartheta_l\sqrt{\lambda} \le \mathbb{E}[\ell^\lambda(U^\lambda(\infty))] \le \mathbb{E}[\ell^\lambda(U^\lambda(\infty)+1)] \le \vartheta_u(1+\sqrt{\lambda}). \tag{30}$$

Also, there exist absolute constants $\{\vartheta_{u,m}, m \in \mathbb{N}\}$ such that for all $\lambda \in \mathbb{R}_+$,

$$\mathbb{E}[(U^\lambda(\infty)+1)^m] \le \vartheta_{u,m}(1+\sqrt{\lambda})^m. \tag{31}$$

Combining Lemma 4.6 and (29) it follows that there exist absolute constants ϑ_l and ϑ_u such that

$$\vartheta_u^{-1}(1+\sqrt{\lambda})^{-1} \le \mathbb{E}_1[\tau_u^\lambda] \le \vartheta_l^{-1}(\sqrt{\lambda})^{-1}.$$

This proves that $(\mathbb{E}_1[\tau_u^\lambda])^{-1} = \mathcal{O}(\sqrt{\lambda})$. Finally, from Lemmas 4.4 and 4.6 and using $f(x) = x^m$ there, we have

$$\mathbb{E}_1\left[\int_0^{\tau_u^\lambda} (\tilde{X}^\lambda(s))^m\,ds\right] \le \vartheta_{2,m}(1+\sqrt{\lambda})^{m-1}, \quad m \in \mathbb{N}.$$

The statement of the proposition now follows recalling that $\{f^\lambda\} \in \mathcal{S}_m$.

4.2.2. Proof of Proposition 4. Given $m \in \mathbb{N}$ and $\{f^\lambda\} \in \mathcal{S}_m$ consider, for each λ, the ODE on $[0, \infty)$:

$$\begin{aligned} -\ell^\lambda(x)(u)^{(1)}(x) + \lambda(u)^{(2)}(x) &= -f^\lambda(x), \\ u(0) &= 0. \end{aligned} \tag{32}$$

We first identify some properties of solutions to this equation.

Lemma 4.7. *Fix $m \in \mathbb{N}$ and $\{f^\lambda\} \in \mathcal{S}_m$. Then there exist absolute constants $A_{i,m}$, $i = 1, 2, 3$ such that for each λ, there is an infinitely differentiable solution $u^\lambda_{f^\lambda}$ to (32) such that, for all $x \geq 0$,*

$$|(u^\lambda_{f^\lambda})^{(1)}(x)| \leq A_{1,m}\big(x^{m-1} + (\sqrt{\lambda})^{m-1}\big), \tag{33}$$

$$|(u^\lambda_{f^\lambda})^{(2)}(x)| \leq A_{2,m}\left(\frac{x^m}{\lambda} + \sqrt{\lambda}^{m-2}\right), \tag{34}$$

$$|(u^\lambda_{f^\lambda})^{(3)}(x)| \leq A_{3,m}\left(\frac{x^{m+1}}{\lambda^2} + (\sqrt{\lambda})^{m-3}\right) \tag{35}$$

if $m \geq 1$ and

$$|(u^\lambda_{f^\lambda})^{(1)}(x)| \leq A_{1,m}\frac{1}{\sqrt{\lambda}}, \tag{36}$$

$$|(u^\lambda_{f^\lambda})^{(2)}(x)| \leq A_{2,m}\frac{1}{\lambda}, \tag{37}$$

$$|(u^\lambda_{f^\lambda})^{(3)}(x)| \leq A_{3,m}\left(\frac{x}{\lambda^2} + \frac{1}{\sqrt{\lambda^3}} + \frac{|(f^\lambda)^{(1)}(x)|}{\lambda}\right) \tag{38}$$

if $m = 0$. This solution satisfies the identity

$$u^\lambda_{f^\lambda}(y) = \mathcal{V}^\lambda_u(f^\lambda, y)\left(=: \mathbb{E}_y\left[\int_0^{\tilde{\tau}^\lambda_u} f^\lambda(\tilde{Y}^\lambda(s))\,ds\right]\right). \tag{39}$$

Lemma 4.7 guarantees that for each λ, $\mathcal{V}^\lambda_u(f^\lambda, y)$ is the unique solution to (32) satisfying (33)–(35) (for $m \geq 1$) or (36)–(38) (for $m = 0$). For the remainder of this proof, we use the simplified notation $\mathcal{V}^\lambda(y) = \mathcal{V}^\lambda_u(f^\lambda, y)$. The function sequence $\{f^\lambda\} \in \mathcal{S}_m$ will be fixed.

Fix $y \in \mathbb{N}$. By Ito's lemma

$$\begin{aligned} \mathcal{V}^\lambda(\tilde{X}^\lambda(t \wedge \tau^\lambda_u)) = \mathcal{V}^\lambda(y) &+ \sum_{s \leq t \wedge \tau^\lambda_u : |\Delta \tilde{X}^\lambda(s)| > 0} (\Delta \mathcal{V}^\lambda(\tilde{X}^\lambda(s)) - (\mathcal{V}^\lambda)^{(1)}(\tilde{X}^\lambda(s-))\Delta \tilde{X}^\lambda(s)) \\ &- \int_0^{t \wedge \tau^\lambda_u} (\mathcal{V}^\lambda)^{(1)}(\tilde{X}^\lambda(s-))\ell^\lambda(\tilde{X}^\lambda(s))\,ds - \int_0^{t \wedge \tau^\lambda_u} (\mathcal{V}^\lambda)^{(1)}(\tilde{X}^\lambda(s-))\,dM^\lambda(s). \end{aligned} \tag{40}$$

By Lemma 4.2 with $g(x) = (\mathcal{V}^\lambda)^{(1)}(x)\ell^\lambda(x)$, the first integral has a finite expectation. Subsequently, Lemma 4.3 with $g(x) = (2\lambda + \ell^\lambda(x))((\mathcal{V}^\lambda)^{(1)}(x))^2$ guarantees that the stochastic integral has expectation 0. Finally, since $\tilde{X}^\lambda$ is nonexplosive, it has a finite number of jumps on each finite interval so that $\tilde{X}^\lambda(s-)$ can be replaced with $\tilde{X}^\lambda(s)$ in all integrals that follow. Taking expectations on both sides of (40) we then obtain,

$$\begin{aligned} \mathbb{E}_y[\mathcal{V}^\lambda(\tilde{X}^\lambda(t \wedge \tau^\lambda_u))] = \mathcal{V}^\lambda(y) &+ \mathbb{E}_y\left[\sum_{s \leq t \wedge \tau^\lambda_u : |\Delta \tilde{X}^\lambda(s)| > 0} (\Delta \mathcal{V}^\lambda(\tilde{X}^\lambda(s)) - (\mathcal{V}^\lambda)^{(1)}(\tilde{X}^\lambda(s-))\Delta \tilde{X}^\lambda(s))\right] \\ &- \mathbb{E}_y\left[\int_0^{t \wedge \tau^\lambda_u} (\mathcal{V}^\lambda)^{(1)}(\tilde{X}^\lambda(s))\ell^\lambda(\tilde{X}^\lambda(s))\,ds\right]. \end{aligned}$$

Subtracting and adding terms, and recalling that $\mathcal{V}^\lambda$ solves (32), yields

$$\begin{aligned} &\mathbb{E}_y[\mathcal{V}^\lambda(\tilde{X}^\lambda(t \wedge \tau^\lambda_u))] \\ &= \mathcal{V}^\lambda(y) + \mathbb{E}_y\left[\sum_{s \leq t \wedge \tau^\lambda_u : |\Delta \tilde{X}^\lambda(s)| > 0} \big(\Delta \mathcal{V}^\lambda(\tilde{X}^\lambda(s)) - (\mathcal{V}^\lambda)^{(1)}(\tilde{X}^\lambda(s-))\Delta \tilde{X}^\lambda(s) - \tfrac{1}{2}(\mathcal{V}^\lambda)^{(2)}(\tilde{X}^\lambda(s-))(\Delta \tilde{X}^\lambda(s))^2\big)\right] \end{aligned}$$

$$+\tfrac{1}{2}\left(\mathbb{E}_y\left[\sum_{s\le t\wedge\tau_u^\lambda:\,|\Delta\tilde X^\lambda(s)|>0}(\mathscr{V}^\lambda)^{(2)}(\tilde X^\lambda(s-))(\Delta\tilde X^\lambda(s))^2\right]-\mathbb{E}_y\left[\int_0^{t\wedge\tau_u^\lambda}(2\lambda+\ell^\lambda(\tilde X^\lambda(s)))(\mathscr{V}^\lambda)^{(2)}(\tilde X^\lambda(s))\,ds\right]\right)$$

$$+\tfrac{1}{2}\,\mathbb{E}_y\left[\int_0^{t\wedge\tau_u^\lambda}\ell^\lambda(\tilde X^\lambda(s))(\mathscr{V}^\lambda)^{(2)}(\tilde X^\lambda(s))\,ds\right]-\mathbb{E}_y\left[\int_0^{t\wedge\tau_u^\lambda}f^\lambda(\tilde X^\lambda(s))\,ds\right].$$

Thus,

$$\begin{aligned}&\left|\mathscr{V}^\lambda(y)-\mathbb{E}_y\left[\int_0^{t\wedge\tau_u^\lambda}f^\lambda(\tilde X^\lambda(s))\,ds\right]\right|\\&\quad\le\left|\mathbb{E}_y[\mathscr{V}^\lambda(\tilde X^\lambda(t\wedge\tau_u^\lambda))]\right|\\&\qquad+\tfrac{1}{2}\left|\mathbb{E}_y\left[\sum_{s\le t\wedge\tau_u^\lambda:\,|\Delta\tilde X^\lambda(s)|>0}(\mathscr{V}^\lambda)^{(2)}(\tilde X^\lambda(s-))(\Delta\tilde X^\lambda(s))^2\right]-\mathbb{E}_y\left[\int_0^{t\wedge\tau_u^\lambda}(2\lambda+\ell^\lambda(\tilde X^\lambda(s)))(\mathscr{V}^\lambda)^{(2)}(\tilde X^\lambda(s))\,ds\right]\right|\\&\qquad+\left|\mathbb{E}_y\left[\sum_{s\le t\wedge\tau_u^\lambda:\,|\Delta\tilde X^\lambda(s)|>0}\left(\Delta\mathscr{V}^\lambda(\tilde X^\lambda(s))-(\mathscr{V}^\lambda)^{(1)}(\tilde X^\lambda(s-))\Delta\tilde X^\lambda(s)-\tfrac{1}{2}(\mathscr{V}^\lambda)^{(2)}(\tilde X^\lambda(s-))(\Delta\tilde X^\lambda(s))^2\right)\right]\right|\\&\qquad+\tfrac{1}{2}\left|\mathbb{E}_y\left[\int_0^{t\wedge\tau_u^\lambda}\ell^\lambda(\tilde X^\lambda(s))(\mathscr{V}^\lambda)^{(2)}(\tilde X^\lambda(s))\,ds\right]\right|.\end{aligned}$$

Using Lemma 4.3 with $g=(\mathscr{V}^\lambda)^{(2)}$ there we have

$$\mathbb{E}_y\left[\sum_{s\le t\wedge\tau_u^\lambda:\,|\Delta\tilde X^\lambda(s)|>0}(\mathscr{V}^\lambda)^{(2)}(\tilde X^\lambda(s-))(\Delta\tilde X^\lambda(s))^2\right]=\mathbb{E}_y\left[\int_0^{t\wedge\tau_u^\lambda}(2\lambda+\ell^\lambda(\tilde X^\lambda(s)))(\mathscr{V}^\lambda)^{(2)}(\tilde X^\lambda(s))\,ds\right],$$

so that the terms in the third line cancel each other. Also, since the jump size of $\tilde X^\lambda$ is ±1 we have, by Taylor's expansion, that

$$\begin{aligned}&\mathbb{E}_y\left[\sum_{s\le t\wedge\tau_u^\lambda:\,|\Delta\tilde X^\lambda(s)|>0}\left(\Delta\mathscr{V}^\lambda(\tilde X^\lambda(s))-(\mathscr{V}^\lambda)^{(1)}(\tilde X^\lambda(s-))\Delta\tilde X^\lambda(s)-\tfrac{1}{2}(\mathscr{V}^\lambda)^{(2)}(\tilde X^\lambda(s-))(\Delta\tilde X^\lambda(s))^2\right)\right]\\&\qquad\le\mathbb{E}_y\left[\sum_{s\le t\wedge\tau_u^\lambda:\,|\Delta\tilde X^\lambda(s)|>0}\tfrac{1}{2}\left|(\mathscr{V}^\lambda)^{(3)}(\tilde X^\lambda(s-)+\eta^\lambda_{\tilde X^\lambda(s-)})\right|\right],\end{aligned}$$

for some $\eta^\lambda_{\tilde X^\lambda(s-)}\in(-1,1)$. Thus,

$$\begin{aligned}&\left|\mathscr{V}^\lambda(y)-\mathbb{E}_y\left[\int_0^{t\wedge\tau_u^\lambda}f^\lambda(\tilde X^\lambda(s))\,ds\right]\right|\\&\quad\le\left|\mathbb{E}_y[\mathscr{V}^\lambda(\tilde X^\lambda(t\wedge\tau_u^\lambda))]\right|\\&\qquad+\left|\mathbb{E}_y\left[\sum_{s\le t\wedge\tau_u^\lambda:\,|\Delta\tilde X^\lambda(s)|>0}\left(\Delta\mathscr{V}^\lambda(\tilde X^\lambda(s))-(\mathscr{V}^\lambda)^{(1)}(\tilde X^\lambda(s-))\Delta\tilde X^\lambda(s)-\tfrac{1}{2}(\mathscr{V}^\lambda)^{(2)}(\tilde X^\lambda(s-))(\Delta\tilde X^\lambda(s))^2\right)\right]\right|\\&\qquad+\tfrac{1}{2}\left|\mathbb{E}_y\left[\int_0^{t\wedge\tau_u^\lambda}\ell^\lambda(\tilde X^\lambda(s))(\mathscr{V}^\lambda)^{(2)}(\tilde X^\lambda(s))\,ds\right]\right|\\&\quad\le\left|\mathbb{E}_y[\mathscr{V}^\lambda(\tilde X^\lambda(t\wedge\tau_u^\lambda))]\right|+\mathbb{E}_y\left[\sum_{s\le t\wedge\tau_u^\lambda:\,|\Delta\tilde X^\lambda(s)|>0}\tfrac{1}{2}\left|(\mathscr{V}^\lambda)^{(3)}(\tilde X^\lambda(s-)+\eta^\lambda_{\tilde X^\lambda(s-)})\right|\right]\\&\qquad+\tfrac{1}{2}\,\mathbb{E}_y\left[\int_0^{t\wedge\tau_u^\lambda}\ell^\lambda(\tilde X^\lambda(s))|(\mathscr{V}^\lambda)^{(2)}(\tilde X^\lambda(s))|\,ds\right]\end{aligned}\tag{41}$$

Note that $\tilde{X}^\lambda \geq 1$ for all $t < \tau_u^\lambda$ so that, in particular, $(\tilde{X}^\lambda(s-) + \eta^\lambda_{\tilde{X}^\lambda(s-)})^m \leq (\tilde{X}^\lambda(s-)+1)^m$. Recalling that $\ell^\lambda(x) \leq \vartheta(\sqrt{\lambda}+x)$ (with $\vartheta = \vartheta_2 \vee \vartheta_3$ in (26)), for $m \geq 1$, the bounds (34)–(35) then yield

$$\begin{aligned}
&\left|\mathscr{V}^\lambda(y) - \mathbb{E}_y\left[\int_0^{t\wedge\tau_u^\lambda} f^\lambda(\tilde{X}^\lambda(s))\,ds\right]\right| \\
&\quad\leq \left|\mathbb{E}_y[\mathscr{V}^\lambda(\tilde{X}^\lambda(t\wedge\tau_u^\lambda))]\right| + \mathbb{E}_y\left[\sum_{s\leq t\wedge\tau_u^\lambda:\,|\Delta\tilde{X}^\lambda(s)|>0} \frac{A_{3,m}}{2}\left(\frac{(\tilde{X}^\lambda(s-))^{m+1}}{\lambda^2} + (\sqrt{\lambda})^{m-3}\right)\right] \\
&\qquad + \tfrac{1}{2}\,\mathbb{E}_y\left[\int_0^{t\wedge\tau_u^\lambda} A_{2,m}\vartheta\big(\sqrt{\lambda}+\tilde{X}^\lambda(s)\big)\left(\frac{(\tilde{X}^\lambda(s))^m}{\lambda} + \sqrt{\lambda}^{m-2}\right)ds\right]. \qquad (42)
\end{aligned}$$

We will here need the following lemma.

LEMMA 4.8. (i) *Fix $m \in \mathbb{N}$, $\{f^\lambda\} \in \mathscr{S}_m$ and $\lambda \in \mathbb{R}_+$. Let $\mathscr{V}^\lambda$ be the solution to (32) as in Lemma 4.7. Then, for any $y \in \mathbb{N}$,*

$$\lim_{t\to\infty} \mathbb{E}_y[\mathscr{V}^\lambda\big(\tilde{X}^\lambda(t\wedge\tau_u^\lambda)\big)] = 0; \qquad (43)$$

(ii) *For any $l \geq 0$ and $y \in \mathbb{N}$,*

$$\lim_{t\to\infty} \mathbb{E}_y\left[\sum_{s\leq t\wedge\tau_u^\lambda:\,|\Delta\tilde{X}^\lambda(s)|>0} (\tilde{X}^\lambda(s-))^l\right] = \mathbb{E}_y\left[\int_0^{\tau_u^\lambda} (2\lambda + \ell^\lambda(\tilde{X}^\lambda(s)))(\tilde{X}^\lambda(s))^l\,ds\right]. \qquad (44)$$

Letting $t \to \infty$ in (42) we have by Lemma 4.8 that

$$\begin{aligned}
&\left|\mathscr{V}^\lambda(y) - \mathbb{E}_y\left[\int_0^{\tau_u^\lambda} f^\lambda(\tilde{X}^\lambda(s))\,ds\right]\right| \\
&\quad\leq \frac{A_{3,m}}{2}\sqrt{\lambda}^{m-3}\,\mathbb{E}_y\left[\int_0^{\tau_u^\lambda}(2\lambda+\ell^\lambda(\tilde{X}^\lambda(s)))\,ds\right] + \frac{A_{3,m}}{2\lambda^2}\,\mathbb{E}_y\left[\int_0^{\tau_u^\lambda}(2\lambda+\ell^\lambda(\tilde{X}^\lambda(s)))(\tilde{X}^\lambda(s))^{m+1}\,ds\right] \\
&\qquad + \frac{A_{2,m}\vartheta}{2\lambda}\,\mathbb{E}_y\left[\int_0^{\tau_u^\lambda}(\tilde{X}^\lambda(s))^{m+1}\,ds + \sqrt{\lambda}\int_0^{\tau_u^\lambda}(\tilde{X}^\lambda(s))^m\,ds + (\sqrt{\lambda})^m\int_0^{\tau_u^\lambda}\tilde{X}^\lambda(s)\,ds + (\sqrt{\lambda})^{m+1}\tau_u^\lambda\right].
\end{aligned}$$

Using Proposition 3 and the bound (26) we have (recall $m \in \mathbb{N}$)

$$\mathbb{E}_y\left[\int_0^{\tau_u^\lambda}(2\lambda+\ell^\lambda(\tilde{X}^\lambda(s)))\,ds\right] = \mathscr{O}(\sqrt{\lambda}),$$

$$\mathbb{E}_y\left[\int_0^{\tau_u^\lambda}(2\lambda+\ell^\lambda(\tilde{X}^\lambda(s)))(\tilde{X}^\lambda(s))^{m+1}\,ds\right] = \mathscr{O}(\sqrt{\lambda}^{m+2}),$$

and

$$\mathbb{E}_y\left[\int_0^{\tau_u^\lambda}(\tilde{X}^\lambda(s))^l\,ds\right] = \mathscr{O}(\sqrt{\lambda}^{l-1}), \quad \text{for } l = 0,\,1,\,m,\,m+1.$$

In turn, there exist absolute constants $A_{4,m}$, $A_{5,m}$, and $A_{6,m}$ for which

$$\left|\mathscr{V}^\lambda(y) - \mathbb{E}_y\left[\int_0^{\tau_u^\lambda} f^\lambda(\tilde{X}^\lambda(s))\,ds\right]\right| \leq A_{4,m}\sqrt{\lambda}^{m-2} + \frac{A_{5,m}}{\lambda^2}(\sqrt{\lambda})^{m+2} + \frac{A_{6,m}}{\lambda}(\sqrt{\lambda})^m = \mathscr{O}(\sqrt{\lambda}^{m-2}).$$

Using the definition of $V_u^\lambda(f^\lambda, y)$, we conclude that

$$V_u^\lambda(f^\lambda, y) - \mathscr{V}^\lambda(y) = \mathscr{O}(\sqrt{\lambda}^{m-2}).$$

For $m = 0$, we must take care of the extra term on the right-hand side of (38) (compared to (35)). In particular, in the transition from (41) to (42) we have the extra term:

$$\mathbb{E}_y\left[\sum_{s\leq t\wedge\tau_u^\lambda:\,|\Delta\tilde{X}^\lambda(s)|>0} \left|(f^\lambda)^{(1)}(\tilde{X}^\lambda(s-)+\eta^\lambda_{\tilde{X}^\lambda(s-)})\right|\right] \leq a_3\,\mathbb{E}_y\left[\sum_{s\leq t\wedge\tau_u^\lambda:\,|\Delta\tilde{X}^\lambda(s)|>0} \mathbb{I}_{\{X^\lambda(s-)\in(a^\lambda-1,\,a^\lambda+2)\}}\right].$$

By Lemma 4.3 we have

$$\begin{aligned}
&\frac{1}{\lambda}\mathbb{E}_y\Bigg[\sum_{s\le t\wedge\tau_u^\lambda:\,|\Delta\tilde{X}^\lambda(s)|>0}\mathbb{1}_{\{\tilde{X}^\lambda(s-)\in(a^\lambda-1,a^\lambda+2)\}}\Bigg]\\
&\quad=\frac{1}{\lambda}\mathbb{E}_y\Bigg[\int_0^{t\wedge\tau_u^\lambda}(2\lambda+\ell^\lambda(\tilde{X}^\lambda(s)))\mathbb{1}_{\{\tilde{X}^\lambda(s)\in(a^\lambda-1,\,a^\lambda+2)\}}\,ds\Bigg]\\
&\quad\le 2\,\mathbb{E}_y\Bigg[\int_0^{t\wedge\tau_u^\lambda}\mathbb{1}_{\{\tilde{X}^\lambda(s)\in(a^\lambda-1,\,a^\lambda+2)\}}\,ds\Bigg]+\frac{1}{\lambda}\,\mathbb{E}_y\Bigg[\int_0^{t\wedge\tau_u^\lambda}\ell^\lambda(\tilde{X}^\lambda(s)))\,ds\Bigg].
\end{aligned}\tag{45}$$

For the first element on the right-hand side (for $y=1$),

$$\begin{aligned}
\mathbb{E}_1\Bigg[\int_0^{t\wedge\tau_u^\lambda}\mathbb{1}_{\{\tilde{X}^\lambda(s)\in(a^\lambda-1,a^\lambda+2)\}}\,ds\Bigg]&\le\mathbb{E}_1\Bigg[\int_0^{\tau^\lambda}\mathbb{1}_{\{\tilde{X}^\lambda(s)\in(a^\lambda-1,a^\lambda+2)\}}\,ds\Bigg]\\
&=\mathbb{E}_1[\tau^\lambda]\mathbb{P}\big\{\tilde{X}^\lambda(\infty)\in(a^\lambda-1,a^\lambda+2)\big\}.
\end{aligned}$$

In Proposition 5 below, we will show that $\mathbb{E}_1[\tau_l^\lambda]=\mathcal{O}(1/\sqrt{\lambda})$, which, together with Proposition 3, shows that $\mathbb{E}_1[\tau^\lambda]=\mathcal{O}(1/\sqrt{\lambda})$. By Lemma A.1, $\mathbb{P}\{\tilde{X}^\lambda(\infty)\in(a^\lambda-1,a^\lambda+2)\}=\mathcal{O}(1/\sqrt{\lambda})$ so that $\mathbb{E}_1[\tau^\lambda]\mathbb{P}\{\tilde{X}^\lambda(\infty)\in(a^\lambda-1,a^\lambda+2)\}=\mathcal{O}(1/\lambda)$. Using $|\ell^\lambda(x)|\le\vartheta(\sqrt{\lambda}+x)$ (see (26)) and Proposition 3, the second element on the right-hand side of (45) is itself $\mathcal{O}(1/\lambda)$, which concludes the proof of the proposition.

4.3. Lower excursion. In this section we consider the lower excursion—these are the time intervals $[\tau_u^\lambda,\tau^\lambda)$ and $[\tilde{\tau}_u^\lambda,\tilde{\tau}^\lambda)$ for $\tilde{X}^\lambda$ and $\tilde{Y}^\lambda$, respectively. Define

$$\check{X}^\lambda(t)=-(X^\lambda(t)-\Delta^\lambda)=-\tilde{X}^\lambda(t)\qquad\text{and}\qquad\check{Y}^\lambda(t)=-(Y^\lambda(t)-\Delta^\lambda)=-\tilde{Y}^\lambda(t).$$

Then $\check{X}^\lambda\ge 0$ on $[\tau_u^\lambda,\tau_l^\lambda)$ and $\check{Y}^\lambda>-1$ on $[\tilde{\tau}_u^\lambda,\tilde{\tau}_l^\lambda)$, respectively. At time τ_u^λ, $\check{X}^\lambda=0$ and τ^λ is its hitting time of -1 (similarly for $\check{Y}^\lambda$, $\tilde{\tau}_u^\lambda$, and $\tilde{\tau}^\lambda$).

Define

$$\check{\ell}^\lambda(x)=\lambda-\mu((-x+\Delta^\lambda)\wedge n^\lambda)-\theta(-x+\Delta^\lambda-n^\lambda)^+.\tag{46}$$

With $\check{X}^\lambda(0)=y\ge 0$, the process $\check{X}^\lambda$ satisfies the following on $[0,\tau_l^\lambda)$

$$\check{X}^\lambda(t)=\check{X}^\lambda(0)-\int_0^t\check{\ell}^\lambda(\check{X}^\lambda(s))\,ds-M^\lambda(t).$$

Similarly, $\check{Y}^\lambda$ satisfies on $[0,\tilde{\tau}_l^\lambda)$

$$\check{Y}^\lambda(t)=\check{Y}^\lambda(0)-\int_0^t\check{\ell}^\lambda(\check{Y}^\lambda(s))\,ds-\sqrt{2\lambda}B^\lambda(t).$$

The function $\check{\ell}^\lambda(x)$ is nondecreasing with $\check{\ell}^\lambda(0)=0$. There is a clear symmetry between the upper excursion and the lower excursion and the following are exact analogues of Propositions 3 and 4.

Proposition 5 (Order Bounds). *Fix $m\in\mathbb{N}$ and $\{f^\lambda\}\in\mathcal{S}_m$. Then,*

$$(\mathbb{E}_0[\tau_l^\lambda])^{-1}=\mathcal{O}(\sqrt{\lambda})\qquad\text{and}\qquad V_l^\lambda(f^\lambda,0)=\mathcal{O}(\sqrt{\lambda}^{m-1}).$$

Proposition 6 (Gap Bounds). *Fix $m\in\mathbb{N}$ and $\{f^\lambda\}\in\mathcal{S}_m$. Then,*

$$V_l^\lambda(f^\lambda,0)-\mathcal{V}_l^\lambda(f^\lambda,0)=\mathcal{O}(\sqrt{\lambda}^{m-2}).$$

The proof of Proposition 5 is very similar to that of Proposition 3 for the upper excursion and is based on analogues of Lemmas 4.4 and 4.6 that are proved identically. The proof of Proposition 6 is, as well, similar to that of Proposition 4 for the upper excursion. The key step is writing the appropriate ODE and identifying the gradient bounds. Specifically, fixing $\lambda\in\mathbb{R}_+$, $m\in\mathbb{N}$, define $\check{f}^\lambda$ by $\check{f}^\lambda(x)=f^\lambda(-x)$ and consider the following ODE on $[0,\infty)$:

$$\begin{aligned}
-\check{\ell}^\lambda(x)(\check{u})^{(1)}(x)+\lambda(\check{u})^{(2)}(x)&=-\check{f}^\lambda(x),\\
\check{u}(-1)&=0.
\end{aligned}\tag{47}$$

LEMMA 4.9. *Fix $m \in \mathbb{N}$ and $\{\check{f}^\lambda\} \in \mathcal{S}_m$. Then, for each λ, there exists an infinitely differentiable solution $u^\lambda_{\check{f}^\lambda}$ for (32) that satisfies the derivative bounds in Lemma 4.7. Furthermore,*

$$\check{u}^\lambda_{\check{f}^\lambda}(y) = \mathscr{V}^\lambda_l(\check{f}^\lambda, y)\left(=: \mathbb{E}_y\left[\int_0^{\check{\tau}^\lambda_l} \check{f}^\lambda(\check{Y}^\lambda(s))\,ds\right]\right).$$

Lemma 4.9 is proved identically to Lemma 4.7. From here, the lower excursion has exact analogues of Lemmas 4.2, 4.3, and 4.8 that are proved identically (in fact, the boundedness of the state space of $\check{X}^\lambda$ further simplifies the proofs). The proof of Proposition 6 is, in turn, identical to that of Proposition 4. We omit the detailed argument here but point the reader to Appendix C, where we provide a complete proof for the lower excursion in the case of the NDS regime.

We conclude this section with a remark about the case $\Delta^\lambda \notin \mathbb{N}$.

REMARK 4.2 (NONINTEGER Δ^λ). Thus far we have assumed that $\Delta^\lambda \in \mathbb{N}$. To explain why that assumption is made without loss of generality, assume that $\Delta^\lambda \notin \mathbb{N}$. We then use a slightly different centering for X^λ. Specifically, define $\tilde{Y}^\lambda$ together with its underlying regenerative structure as before. We redefine $\tilde{X}^\lambda = X^\lambda - \lceil \Delta^\lambda \rceil$ and redefine its regenerative structure with respect to $\lceil \Delta^\lambda \rceil$ in an obvious way.

All the order-bound arguments in §4.2.1 remain unchanged, and only a minor change is required in the proof of the gap bounds in Proposition 3. First, in the dynamics of $\tilde{X}^\lambda$ (see (25)), we must replace $\ell^\lambda(x)$ with

$$\begin{aligned}\bar{\ell}^\lambda(x) &= -\lambda + \mu((x + \lceil \Delta^\lambda \rceil) \wedge n^\lambda) + \theta(x + \lceil \Delta^\lambda \rceil - n^\lambda)^+ \\ &= \ell^\lambda(x) + \big(\mu((x + \lceil \Delta^\lambda \rceil) \wedge n^\lambda) - \mu((x + \Delta^\lambda) \wedge n^\lambda)\big) + \big(\theta(x + \lceil \Delta^\lambda \rceil - n^\lambda)^+ - \theta(x + \Delta^\lambda - n^\lambda)^+\big).\end{aligned}$$

Note that

$$|\bar{\ell}^\lambda(x) - \ell^\lambda(x)| \le \mu + \theta. \tag{48}$$

Then, in the proof of Proposition 4 (specifically in Equation (40)), there will be an extra term $\int_0^{t\wedge\tau^\lambda_u} (\mathscr{V}^\lambda)^{(1)}(\tilde{X}^\lambda(s))(\ell^\lambda(\tilde{X}^\lambda(s)) - \bar{\ell}^\lambda(\tilde{X}^\lambda(s)))\,ds$. Given (48), we then have that

$$\left|\mathbb{E}_y\left[\int_0^{t\wedge\tau^\lambda_u} (\mathscr{V}^\lambda)^{(1)}(\tilde{X}^\lambda(s))(\ell^\lambda(\tilde{X}^\lambda(s)) - \bar{\ell}^\lambda(\tilde{X}^\lambda(s)))\,ds\right]\right| \le (\mu+\theta)\,\mathbb{E}_y\left[\int_0^{t\wedge\tau^\lambda_u} |(\mathscr{V}^\lambda)^{(1)}(\tilde{X}^\lambda(s))|\,ds\right].$$

Proceeding as in the text after Equation (42), we conclude that this term is $\mathcal{O}(\sqrt{\lambda}^{m-2})$ so that the gap bound is not compromised. Exactly the same change would apply to the proofs of the lower excursion.

For the analysis in the next section (specifically, toward Lemma 5.2), it is useful to note that the same argument applies, in fact, to any perturbation of the drift of the diffusion by a constant. Specifically, assume that we replace Y^λ with $\bar{Y}^\lambda$ that is defined by

$$\bar{Y}^\lambda(t) = \bar{Y}^\lambda(0) + \int_0^t \tilde{b}^\lambda(\bar{Y}^\lambda(s))\,ds + \sqrt{2\lambda}B(t), \quad t \ge 0,$$

where $\tilde{b}^\lambda$ differs from b^λ only with respect to the constant n^λ; i.e., $\tilde{b}^\lambda(x) = \lambda - \mu(x \wedge \tilde{n}^\lambda) - \theta(x - \tilde{n}^\lambda)^+$. Let $\tilde{\Delta}^\lambda$ be defined from Δ^λ by replacing n^λ with $\tilde{n}^\lambda$. Then, $\Delta^\lambda - \tilde{\Delta}^\lambda = \mathcal{O}(1)$ provided that $n^\lambda - \tilde{n}^\lambda = \mathcal{O}(1)$. Redefine $\tilde{Y}^\lambda = Y^\lambda - \tilde{\Delta}^\lambda$. Let

$$\tilde{\ell}^\lambda(x) = -\lambda + \mu((x + \tilde{\Delta}^\lambda) \wedge \tilde{n}^\lambda) + \theta(x + \tilde{\Delta}^\lambda - \tilde{n}^\lambda)^+.$$

Then, with $n^\lambda - \tilde{n}^\lambda = \mathcal{O}(1)$, we have that $|\tilde{\ell}^\lambda(x) - \ell^\lambda(x)| \le \vartheta$ for an absolute constant ϑ whose value depends on $n^\lambda - \tilde{n}^\lambda = \mathcal{O}(1)$. The arguments above apply here as well; in particular, the gap bounds persist after an $\mathcal{O}(1)$ perturbation of n^λ.

5. Universal optimization of the Erlang-A queue. We revisit two staffing problems that have been analyzed in the literature using asymptotic analysis and limits; recall the discussion in §1. In §5.1 we consider the problem of minimizing the number of servers subject to a constraint on the fraction of abandoning customers. In §5.2 we consider a cost minimization problem where one seeks to minimize a combined cost of staffing, abandonment, and holding.

To make explicit the dependence of the steady-state distribution on the number of servers, we let X^λ_n be the headcount process in the Erlang-A queue with n servers and similarly define Y^λ_n for the universal diffusion. The service rate μ and the patience rate θ are fixed and do not appear in the notation. We define

$$Q^\lambda_n(\infty) = (X^\lambda_n(\infty) - n)^+ \qquad \text{and} \qquad \tilde{Q}^\lambda_n(\infty) = (Y^\lambda_n(\infty) - n)^+$$

to be the steady-state queue and its proposed universal approximation.

5.1. Constraint satisfaction. Denote by $\mathrm{Ab}(n,\lambda)$ the fraction of abandonments when the arrival rate is λ and the number of servers is n. Consider the constraint satisfaction problem

$$N^*(\lambda)=\min\{n\in\mathbb{N}\colon\ \mathrm{Ab}(n,\lambda)\le\alpha(\lambda)\}. \tag{49}$$

That is, $N^*(\lambda)$ is the least number of servers required to meet a target abandonment fraction $\alpha(\lambda)$ when the arrival rate is λ. The instances $\alpha(\lambda)\equiv\alpha$ and $\alpha(\lambda)=\alpha/\sqrt{\lambda}$, discussed in the introduction, are covered here as special cases.

It is known that

$$\lambda\,\mathrm{Ab}(n,\lambda)=\theta\,\mathbb{E}[Q_n^\lambda(\infty)]; \tag{50}$$

see, e.g., Mandelbaum and Zeltyn [27, §4.4]. As a result, (49) is equivalently written as

$$N^*(\lambda)=\min\left\{n\in\mathbb{N}\colon\ \mathbb{E}[Q_n^\lambda(\infty)]\le\frac{\lambda}{\theta}\alpha(\lambda)\right\}.$$

As an approximation to $N^*(\lambda)$, we propose to solve the problem

$$\tilde N^*(\lambda)=\inf\left\{n\in\mathbb{N}\colon\ \mathbb{E}[\tilde Q_n^\lambda(\infty)]\le\frac{\lambda}{\theta}\alpha(\lambda)\right\}. \tag{51}$$

The following provides a characterization of $\tilde N^*(\lambda)$.

Lemma 5.1. *Suppose that* $\limsup_{\lambda\to\infty}\alpha(\lambda)<1$; *then, for all sufficiently large* λ, *there exists a unique solution* $n^*(\lambda)\in\mathbb{R}$ *to the equation*

$$\mathbb{E}[\tilde Q_n^\lambda(\infty)]=\frac{\lambda\alpha(\lambda)}{\theta}. \tag{52}$$

By the monotonicity of $\mathbb{E}[\tilde Q_n^\lambda(\infty)]$ in n (see Mandelbaum and Zeltyn [28]) we have that $\tilde N^*(\lambda)=\lceil n^*(\lambda)\rceil$ with the latter as in Lemma 5.1. Using the explicit expressions for $\mathbb{E}[\tilde Q_n^\lambda(\infty)]$ (see (16)) and (52), we have that

$$n^*(\lambda)=\frac{\lambda}{\mu}(1-\alpha(\lambda))+\frac{\sqrt{\lambda}}{\mu}\big(\beta(\lambda,\alpha(\lambda))+\sqrt{\lambda}\alpha(\lambda)\big), \tag{53}$$

where $\beta(\lambda,\alpha(\lambda))$ is the unique solution β to

$$[1-p(\beta,\mu,\theta)]h(\beta/\sqrt{\theta})+p(\beta,\mu,\theta)\frac{\beta}{\sqrt{\theta}}=\frac{1}{\sqrt{\theta}}\big(\beta+\sqrt{\lambda}\alpha(\lambda)\big).$$

The following is the main result of this section.

Theorem 7. *The staffing* $\tilde N^*(\lambda)$ *is asymptotically feasible for* (49); *namely,*

$$\mathrm{Ab}(\tilde N^*(\lambda),\lambda)-\alpha(\lambda)=\mathcal{O}(\lambda^{-1}). \tag{54}$$

If, in addition, $\alpha(\lambda)\ge\vartheta\sqrt{\lambda}^{-1}$ *for some absolute constant* ϑ, *then* $\tilde N^*(\lambda)$ *is asymptotically optimal for* (49); *namely,*

$$N^*(\lambda)-\tilde N^*(\lambda)=\mathcal{O}(1). \tag{55}$$

By definition $\theta\,\mathbb{E}[\tilde Q_{n^*(\lambda)}^\lambda]=\lambda\alpha(\lambda)$. Using (50) we have

$$\big|\mathrm{Ab}(\tilde N^*(\lambda),\lambda)-\alpha(\lambda)\big|\le\frac{\theta}{\lambda}\big|\mathbb{E}[Q_{\tilde N^*(\lambda)}^\lambda(\infty)]-\mathbb{E}[\tilde Q_{\tilde N^*(\lambda)}^\lambda(\infty)]\big|+\frac{\theta}{\lambda}\big|\mathbb{E}[\tilde Q_{\tilde N^*(\lambda)}^\lambda(\infty)]-\mathbb{E}[\tilde Q_{n^*(\lambda)}^\lambda(\infty)]\big|.$$

The first term on the right-hand side is $\mathcal{O}(1/\lambda)$ by Corollary 1. For the second term we have the following lemma.

Lemma 5.2. *Fix two sequences* $\{n_1^\lambda\}$ *and* $\{n_2^\lambda\}$ *of nonnegative numbers such that* $n_2^\lambda-n_1^\lambda=\mathcal{O}(1)$. *Then* $\mathbb{E}[\tilde Q_{n_1^\lambda}^\lambda(\infty)]-\mathbb{E}[\tilde Q_{n_2^\lambda}^\lambda(\infty)]=\mathcal{O}(1)$.

Let $n_2^\lambda = \tilde{N}^*(\lambda)$ and $n_1^\lambda = n^*(\lambda)$. Then, $|n_2^\lambda - n_1^\lambda| \le 1$ by construction and recalling (50) and (52), we can apply Lemma 5.2 to conclude that $|\mathrm{Ab}(\tilde{N}^*(\lambda), \lambda) - \alpha(\lambda)| = \mathcal{O}(\lambda^{-1})$ as required.

For (55), we make the simple observation that if $\liminf_{\lambda\to\infty} \sqrt{\lambda}\alpha(\lambda) > 0$, then $\limsup_{\lambda\to\infty}(\beta(\lambda, \alpha(\lambda)) + \sqrt{\lambda}\alpha(\lambda)) < \infty$, which then implies (see (53)) that

$$\limsup_{\lambda\to\infty}(n^*(\lambda)\mu - \lambda)/\sqrt{\lambda} < \infty. \tag{56}$$

Assume, toward contradiction, that there is a sequence $\lambda \to \infty$ such that

$$|N^*(\lambda) - \tilde{N}^*(\lambda)| \to \infty.$$

Then either $N^*(\lambda) - \tilde{N}^*(\lambda) \to \infty$ or $N^*(\lambda) - \tilde{N}^*(\lambda) \to -\infty$. Recall that $\tilde{N}^*(\lambda) = \lceil n^*(\lambda) \rceil$ with the latter being the solution to (52). Thus, it holds in particular that either $N^*(\lambda) - n^*(\lambda) \to \infty$, or $N^*(\lambda) - n^*(\lambda) \to -\infty$.

The following lemma will be useful in what follows.

Lemma 5.3. *Fix two sequences* $\{n_1^\lambda\}$ *and* $\{n_2^\lambda\}$ *of nonnegative numbers such that* $\limsup_{\lambda\to\infty}(\mu n_2^\lambda - \lambda)/\sqrt{\lambda} < \infty$. *Then,* $\mathbb{E}[\tilde{Q}^\lambda_{n_1^\lambda}(\infty)] - \mathbb{E}[\tilde{Q}^\lambda_{n_2^\lambda}(\infty)] \to -\infty$ *if* $n_2^\lambda - n_1^\lambda \to -\infty$, *and* $\mathbb{E}[\tilde{Q}^\lambda_{n_1^\lambda}(\infty)] - \mathbb{E}[\tilde{Q}^\lambda_{n_2^\lambda}(\infty)] \to \infty$ *if* $n_2^\lambda - n_1^\lambda \to \infty$.

Assume first that $N^*(\lambda) - n^*(\lambda) \to \infty$. Fix a positive sequence $K^\lambda \to \infty$ such that $n^*(\lambda) + K^\lambda < N^*(\lambda)$ for all λ and let $\bar{n}^\lambda = n^*(\lambda) + K^\lambda$. Let $n_1^\lambda = \bar{n}^\lambda$ and $n_2^\lambda = n^*(\lambda)$. Then since $n^*(\lambda)$ (and, in turn, n_2^λ) satisfies (56), we can apply Lemma 5.3 to have $\mathbb{E}[\tilde{Q}^\lambda_{n^*(\lambda)}(\infty)] - \mathbb{E}[\tilde{Q}^\lambda_{\bar{n}^\lambda}(\infty)] \to \infty$. In particular, since (see (52)) $\mathbb{E}[\tilde{Q}^\lambda_{n^*(\lambda)}(\infty)] = \lambda\alpha(\lambda)/\theta$, we have

$$\lim_{\lambda\to\infty}\left(\mathbb{E}[\tilde{Q}^\lambda_{\bar{n}^\lambda}(\infty)] - \frac{\lambda\alpha(\lambda)}{\theta}\right) = -\infty. \tag{57}$$

By Corollary 1, $\mathbb{E}[\tilde{Q}^\lambda_{\bar{n}^\lambda}(\infty)] - \mathbb{E}[Q^\lambda_{\bar{n}^\lambda}(\infty)] = \mathcal{O}(1)$ so that (57) implies

$$\limsup_{\lambda\to\infty}\left(\mathbb{E}[Q^\lambda_{\bar{n}^\lambda}(\infty)] - \frac{\lambda\alpha(\lambda)}{\theta}\right) = -\infty,$$

and, in turn, that $\bar{n}^\lambda$ is feasible for (49) for all sufficiently large λ. Since $N^*(\lambda) > \bar{n}^\lambda$ by construction, this is a contradiction to the optimality of $N^*(\lambda)$ and we may conclude that $\limsup_{\lambda\to\infty} N^*(\lambda) - n^*(\lambda) < \infty$. The proof that $\liminf_{\lambda\to\infty} N^*(\lambda) - n^*(\lambda) > -\infty$ is similar and uses the second part of Lemma 5.3. The detailed argument is omitted. □

Example 5.1 (Constrained Staffing). (i) *Unscaled targets*: We fix $\mu = 1$ and $\theta = 1/3$ and consider the case $\alpha(\lambda) \equiv \alpha \in \{0.05, 0.2\}$; i.e., the target fraction of abandonment does not scale with λ. The figure pairs (10, 11) and (12, 13) correspond to $\alpha = 0.05$ and $\alpha = 0.2$, respectively. For each value of λ (in jumps of 20), we solve (49) and (51) to obtain $N^*(\lambda)$ and $\tilde{N}^*(\lambda)$. The plots on the left-hand side of Figures 10 and 12 support (54) in showing that $\mathrm{Ab}(\tilde{N}^*(\lambda), \lambda)$ never violates the target $\alpha(\lambda)$. The plots on the right-hand side compare $\tilde{N}^*(\lambda)$ to the true optimal solution to (49) and support that the constraint is satisfied with little or no compromise to staffing costs. The plots in Figures 11 and 13 display the respective error ratios $(\alpha(\lambda) - \mathrm{Ab}(\tilde{N}^*(\lambda), \lambda))/\alpha(\lambda)$ and $(\tilde{N}^*(\lambda) - N^*(\lambda))/N^*(\lambda)$. It is notable that the staffing error is 0 (that is, $\tilde{N}^*(\lambda) = N^*(\lambda)$) for almost all values of λ except for a small set of values (see Figure 13) where the staffing error is a single server.

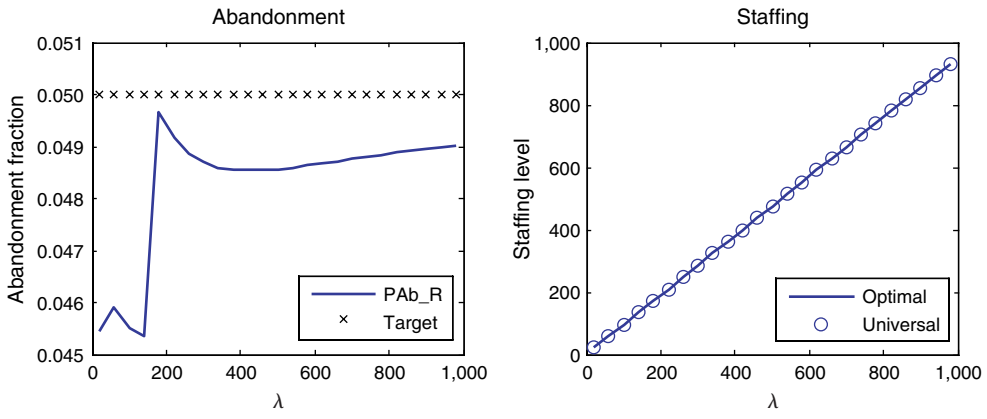

Figure 10. Constrained staffing, unscaled targets $\alpha^\lambda = \alpha = 0.05$.

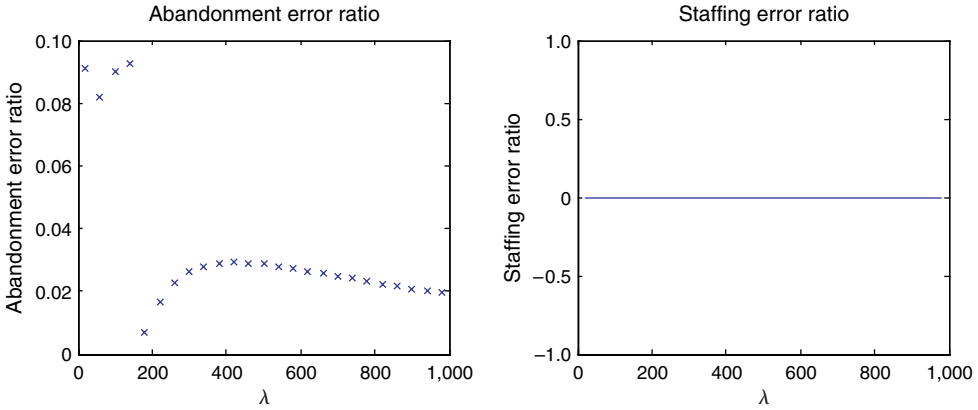

FIGURE 11. Constrained staffing, unscaled targets $\alpha^\lambda = \alpha = 0.05$ (error ratio).

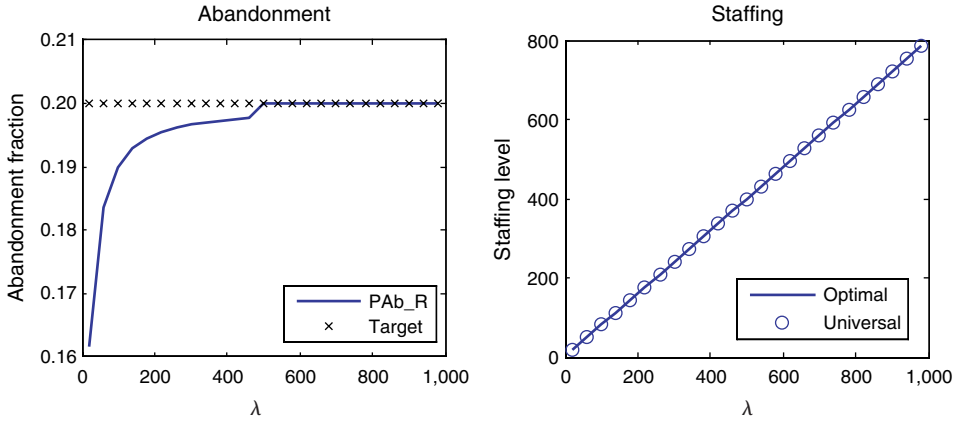

FIGURE 12. Constrained staffing $\alpha^\lambda = \alpha = 0.2$.

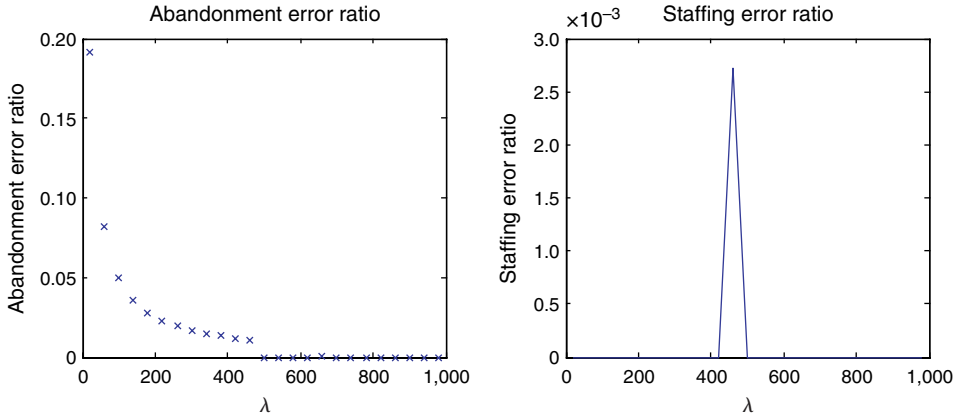

FIGURE 13. Constrained staffing $\alpha^\lambda = \alpha = 0.2$ (error ratio).

(ii) *Scaled targets*: We set $\mu = 1$ and $\theta = 1/3$ and repeat the experiment above but, this time, with scaled targets of the form $\alpha(\lambda) = \alpha/\sqrt{\lambda}$ with $\alpha \in \{0.5, 2\}$. Figures 14 and 15 correspond to $0.5/\sqrt{\lambda}$. Figures 16 and 17 correspond to the case $\alpha(\lambda) = 2/\sqrt{\lambda}$.

5.2. Cost minimization. Given cost parameters C_s^λ, C_{ab}^λ, and C_q^λ, consider the cost

$$C_X(\lambda, n) = C_s^\lambda n + C_{ab}^\lambda \lambda \,\mathrm{Ab}(n, \lambda) + C_q^\lambda \,\mathbb{E}[Q_n^\lambda(\infty)].$$

Recalling (50), $C_X(\lambda, n)$ is equivalently written by

$$C_X(\lambda, n) = C_s^\lambda n + (C_{ab}^\lambda \theta + C_q^\lambda)\,\mathbb{E}[Q_n^\lambda(\infty)].$$

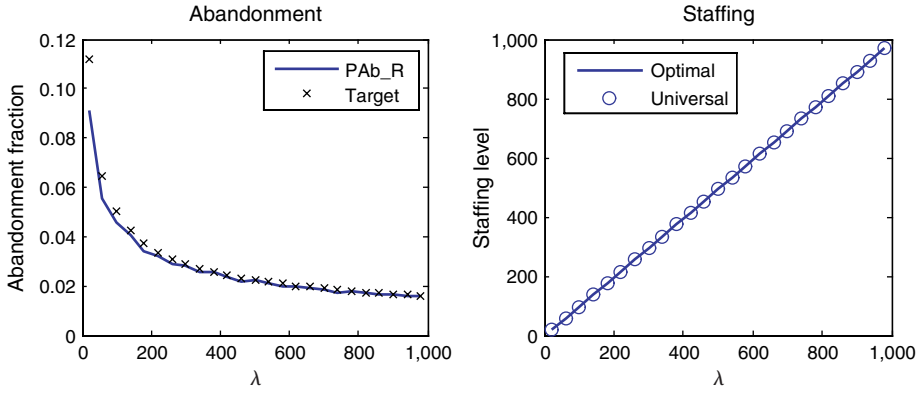

FIGURE 14. Constrained staffing $\alpha^\lambda = 0.5/\sqrt{\lambda}$.

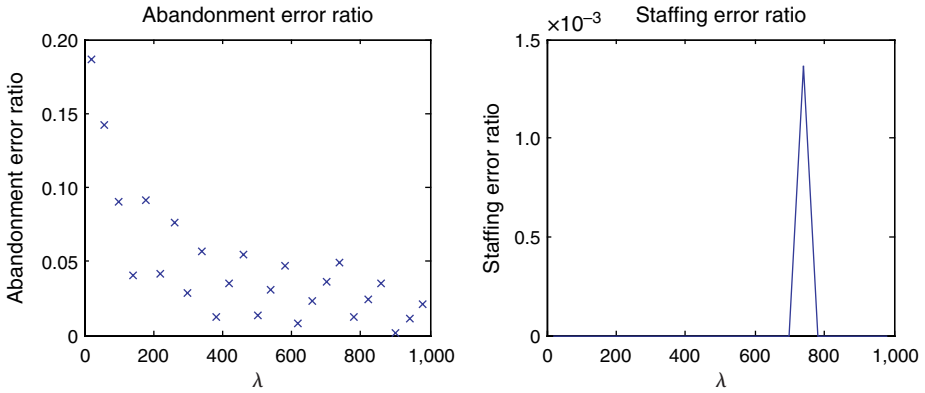

FIGURE 15. Constrained staffing $\alpha^\lambda = 0.5/\sqrt{\lambda}$ (error ratio).

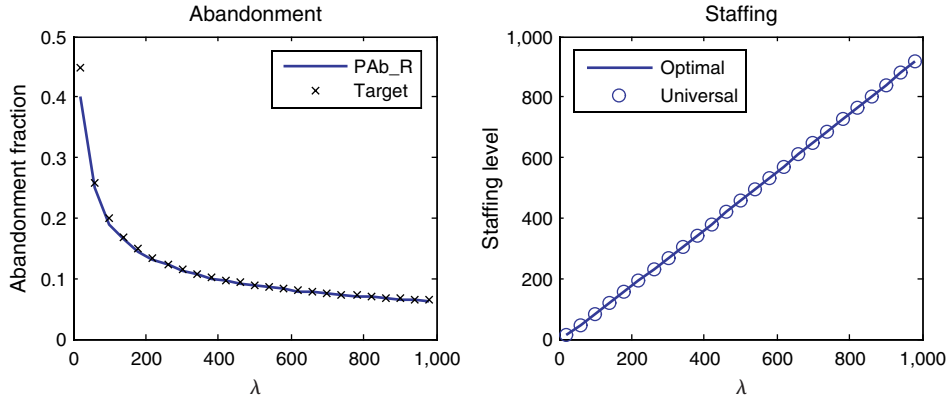

FIGURE 16. Constrained staffing $\alpha^\lambda = \alpha = 2/\sqrt{\lambda}$.

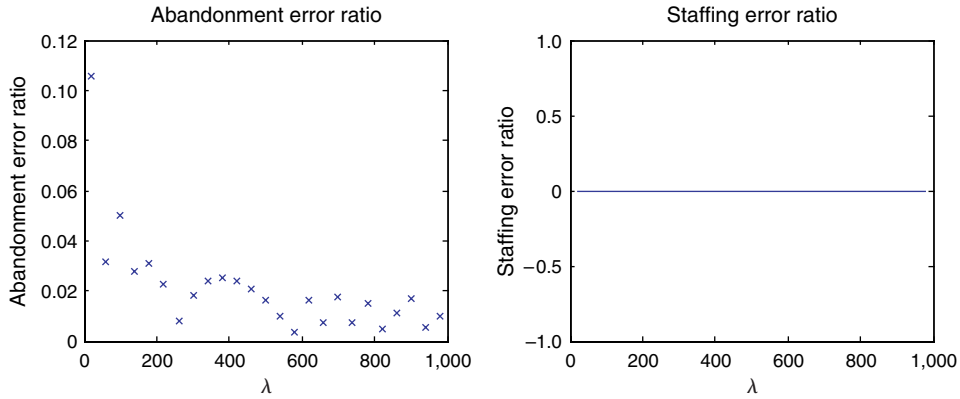

FIGURE 17. Constrained staffing $\alpha^\lambda = \alpha = 2/\sqrt{\lambda}$ (error ratio).

Define

$$N^*(\lambda) = \arg\min_{n\in\mathbb{N}} C_X(\lambda, n). \tag{58}$$

Similarly recall that $\tilde{Q}_n^\lambda = (Y^\lambda(\infty) - n)^+$, define

$$C_Y(\lambda, n) = C_s^\lambda n + (C_{ab}^\lambda \theta + C_q^\lambda)\,\mathbb{E}[\tilde{Q}_n^\lambda(\infty)],$$

and redefine

$$\tilde{N}^*(\lambda) = \arg\min_{n\in\mathbb{N}} C_Y(\lambda, n). \tag{59}$$

In both (58) and (59) if there are multiple minimizers we choose the minimal among them. Recall that $\mathbb{E}[\tilde{Q}_n^\lambda(\infty)]$ is given by (16), with $\beta^\lambda = \beta_n^\lambda = (n\mu - \lambda)/\sqrt{\lambda}$.

THEOREM 8. *The staffing level $\tilde{N}^*(\lambda)$ is asymptotically optimal for (58) in the sense of*

$$C_X(\lambda, \tilde{N}^*(\lambda)) - C_X(\lambda, N^*(\lambda)) = \mathcal{O}(\max(C_{ab}^\lambda, C_q^\lambda)).$$

REMARK 5.1. Note that if $C_{ab}^\lambda \equiv C_{ab}$ and $C_q^\lambda \equiv C_q$, then Theorem 8 states that the cost gap is $\mathcal{O}(1)$. Also, whereas Theorem 8 imposes no restriction on the cost parameters, the interesting cases (and the ones we consider in our numerical experiments below) are those where $\tilde{N}^*(\lambda) > 0$ which, in turn, holds only if $C_s^\lambda/\mu < C_{ab}^\lambda + C_q^\lambda/\theta$. This inequality is assumed, for example, in Bassamboo and Randhawa [6].

Recent work (Randhawa [32]) suggests that it may be possible to improve on our approximation gap. Interpreted to our context, this work suggests that since $\mathbb{E}[Q_n^\lambda(\infty)] - \mathbb{E}[\tilde{Q}_n^\lambda(\infty)] = \mathcal{O}(1)$, the optimality gap is $o(\max(C_{ab}^\lambda, C_q^\lambda))$. However, it is not clear that the conditions required in Randhawa [32, §1] are satisfied for the universal diffusion and the Erlang-A queue.

REMARK 5.2. The total cost is a natural criterion of optimality in this context of the cost minimization problem. Nevertheless, one may be interested also in how "close" the recommended staffing is to the optimal staffing. Because the subject of this paper is the universal approximation rather than the specific optimization problem, we do not pursue the proof of this result here. Interestingly, in our numerical examples below not only is the cost under $\tilde{N}^*(\lambda)$, $C_X(\lambda, \tilde{N}^*(\lambda))$, very close to the true optimal cost but also $\tilde{N}^*(\lambda)$ is identical to $N^*(\lambda)$ for most values of λ.

PROOF. By the definition of $N^*(\lambda)$ and $\tilde{N}^*(\lambda)$,

$$C_X(\lambda, N^*(\lambda)) \le C_X(\lambda, \tilde{N}^*(\lambda)) \qquad \text{and} \qquad C_Y(\lambda, \tilde{N}^*(\lambda)) \le C_Y(\lambda, N^*(\lambda)).$$

Hence

$$\begin{aligned} 0 &\le C_X(\lambda, \tilde{N}^*(\lambda)) - C_X(\lambda, N^*(\lambda)) \\ &= C_X(\lambda, \tilde{N}^*(\lambda)) - C_Y(\lambda, \tilde{N}^*(\lambda)) + C_Y(\lambda, \tilde{N}^*(\lambda)) - C_Y(\lambda, N^*(\lambda)) + C_Y(\lambda, N^*(\lambda)) - C_X(\lambda, N^*(\lambda)). \end{aligned}$$

Since $C_Y(\lambda, \tilde{N}^*(\lambda)) - C_Y(\lambda, N^*(\lambda)) \le 0$ we have that

$$\begin{aligned} 0 &\le C_X(\lambda, \tilde{N}^*(\lambda)) - C_X(\lambda, N^*(\lambda)) \\ &\le (C_{ab}^\lambda \theta + C_q^\lambda)\,|\,\mathbb{E}[\tilde{Q}_{N^*(\lambda)}(\infty)] - \mathbb{E}[Q_{N^*(\lambda)}^\lambda(\infty)]| + (C_{ab}^\lambda \theta + C_q^\lambda)\,|\,\mathbb{E}[\tilde{Q}_{\tilde{N}^*(\lambda)}^\lambda(\infty)] - \mathbb{E}[Q_{\tilde{N}^*(\lambda)}^\lambda(\infty)]|. \end{aligned}$$

The result now follows from Corollary 1. □

EXAMPLE 5.2 (COST MINIMIZATION). (i) *Unscaled parameters*: Let $\mu = 1$, $\theta = 1/2$, $C_{ab}^\lambda = C_q^\lambda = 2$, $C_s^\lambda = 1$. For values of λ from 20 to 2,000 (in jumps of 20), we solve for $N^*(\lambda)$ in (58) and $\tilde{N}^*(\lambda)$ in (59). The graph on the left-hand side of Figure 18 displays $C_X(\tilde{N}^*(\lambda), \lambda)$ and $C_X(N^*(\lambda), \lambda)$ as a function of λ supports Theorem 8. The graph on the right-hand side displays $N^*(\lambda)$ and $\tilde{N}^*(\lambda)$, suggesting that the corresponding staffing levels are also close. The plots in Figure 19 display the error ratios $|C_X(\tilde{N}^*(\lambda), \lambda) - C_X(N^*(\lambda), \lambda)|/C_X(N^*(\lambda), \lambda))$ and $|\tilde{N}^*(\lambda) - N^*(\lambda)|/N^*(\lambda)$, respectively. The staffing error is, in absolute values, 0 servers except for a single point around $\lambda = 1{,}600$ in which the error is a single server.

(ii) *Scaled parameters*: We reconsider the setting above (in particular, $\mu = 1$ and $\theta = 1/2$) but now with scaled cost parameters. Specifically, we set $C_s^\lambda = 1$ but $C_{ab}^\lambda = C_q^\lambda = 2\sqrt{\lambda}$. Figure 20 displays the costs and staffing levels and Figure 21 displays the error ratios.

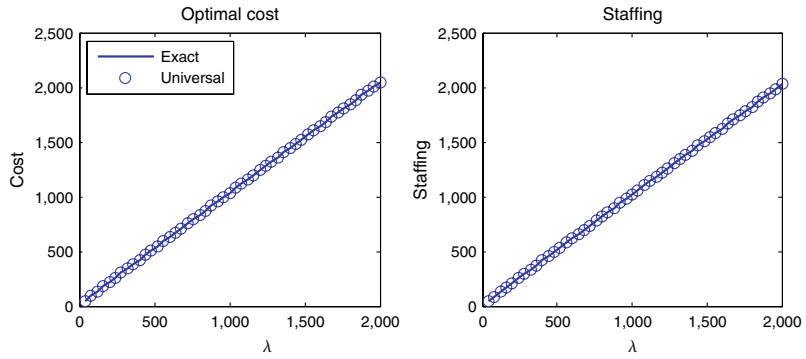

FIGURE 18. Cost minimization: $\mu = 1$, $\theta = 1/2$, $C_{ab}^{\lambda} = C_q^{\lambda} = 2$, and $C_s^{\lambda} = 1$.

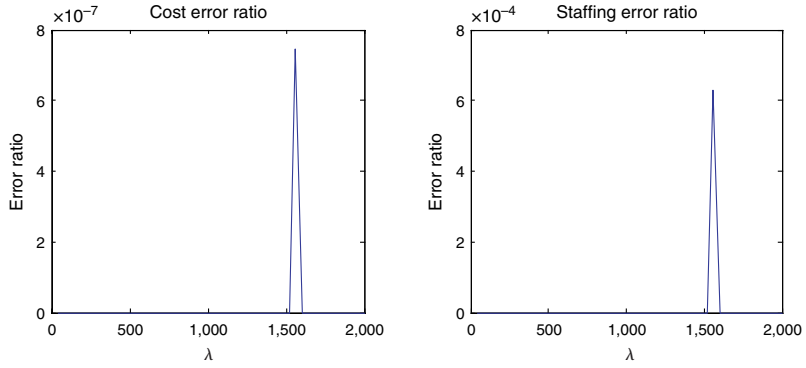

FIGURE 19. Cost minimization: $\mu = 1$, $\theta = 1/2$, $C_{ab}^{\lambda} = C_q^{\lambda} = 2$, and $C_s^{\lambda} = 1$ (error ratio).

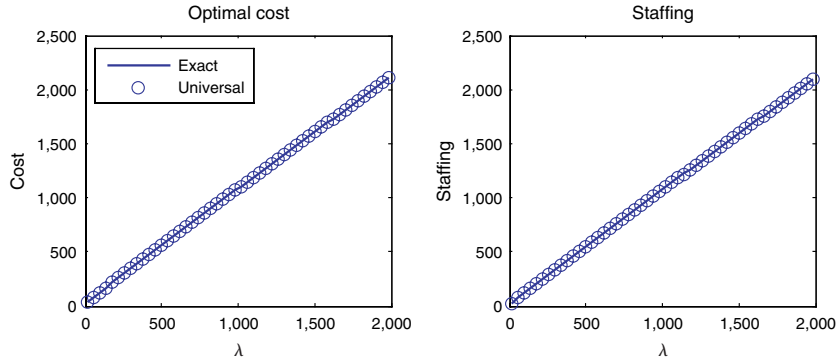

FIGURE 20. Cost minimization, scaled parameters: $\mu = 1$, $\theta = 1/2$, $C_{ab}^{\lambda} = C_q^{\lambda} = 2\sqrt{\lambda}$, and $C_s^{\lambda} = 1$.

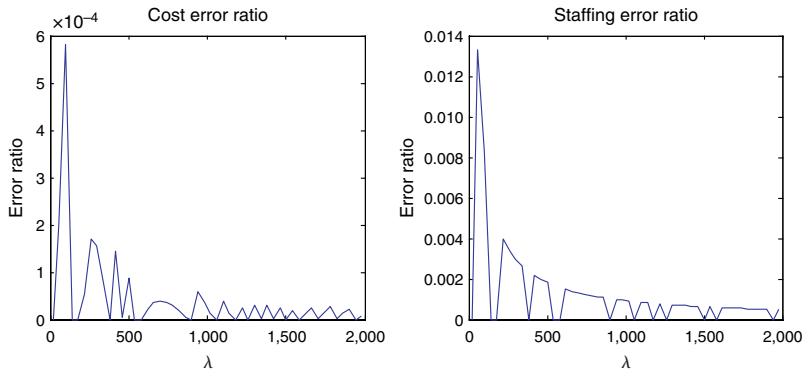

FIGURE 21. Cost minimization, scaled parameters: $\mu = 1$, $\theta = 1/2$, $C_{ab}^{\lambda} = C_q^{\lambda} = 2\sqrt{\lambda}$, and $C_s^{\lambda} = 1$ (error ratio).

6. Concluding remarks.

6.1. Virtual waiting-time distribution. Our analysis relies on the Markovian structure of the headcount process X^λ and, in turn, covers only performance metrics that can be represented as functionals of this process. The virtual waiting time at time t, $V^\lambda(t)$, is defined as the time-to-service of a customer equipped with infinite patience who arrives at time t. Mathematically, $V^\lambda(t)$ is a first passage time that depends, in particular, on the dynamics of X^λ after time t (e.g., service completions and abandonment); see Talreja and Whitt [35, Equation (1.1)]. Thus, a "universal" approximation for the virtual waiting time does not follow directly from our results.

Existing results on heavy-traffic limits for V^λ, particularly Whitt [40] and Talreja and Whitt [35], suggest that in great generality,

$$\lambda(V^\lambda(\infty) - w^\lambda) \overset{d}{\approx} Q^\lambda(\infty) - q^\lambda,$$

where

$$w^\lambda = \frac{1}{\theta}\ln\left(\frac{\lambda}{n^\lambda\mu} \vee 1\right) \qquad \text{and} \qquad q^\lambda = \frac{(\lambda - n^\lambda\mu)^+}{\theta}.$$

Heuristically, given the analysis in this paper, the following sequence of approximations should hold for any sequence t^λ:

$$\mathbb{P}\{V^\lambda(\infty) - w^\lambda > t^\lambda\} \approx \mathbb{P}\{Q^\lambda(\infty) > q^\lambda + \lambda t^\lambda\} \approx \mathbb{P}\{\tilde{Q}^\lambda(\infty) > q^\lambda + \lambda t^\lambda\}.$$

Our Corollary 2 guarantees that $\mathbb{P}\{Q^\lambda(\infty) > q^\lambda + \lambda t^\lambda\} - \mathbb{P}\{\tilde{Q}^\lambda(\infty) > q^\lambda + \lambda t^\lambda\} = \mathcal{O}(1/\sqrt{\lambda})$. Hence, to show that

$$\mathbb{P}\{V^\lambda(\infty) - w^\lambda > t^\lambda\} - \mathbb{P}\{\tilde{Q}^\lambda(\infty) > q^\lambda + \lambda t^\lambda\} = \mathcal{O}(1/\sqrt{\lambda}),$$

it suffices to prove that

$$\mathbb{P}\{V^\lambda(\infty) - w^\lambda > t^\lambda\} - \mathbb{P}\{Q^\lambda(\infty) > q^\lambda + \lambda t^\lambda\} = \mathcal{O}(1/\sqrt{\lambda}). \tag{60}$$

We conjecture that the heuristic above is, in fact, valid and that the excursion-based analysis can help in establishing (60). We leave this as an important problem for future research and conclude this discussion with a numerical experiment that supports our conjecture: set $\mu = 1$ and $\theta = 0.5$ and fix $t^\lambda = -2/\lambda$. The performance metric in question is then $\mathbb{P}\{V^\lambda(\infty) \geq w^\lambda - 2/\lambda\}$. We then repeat Example 3.2 replacing the probability of delay with the metric $\mathbb{P}\{V^\lambda(\infty) \geq w^\lambda - 2/\lambda\}$. The results are displayed in Figures 22–24 and suggest that indeed $\mathbb{P}\{\tilde{Q}^\lambda(\infty) > q^\lambda + \lambda t^\lambda\}$ provides an accurate approximation for $\mathbb{P}\{V^\lambda(\infty) \geq w^\lambda - 2/\lambda\}$.

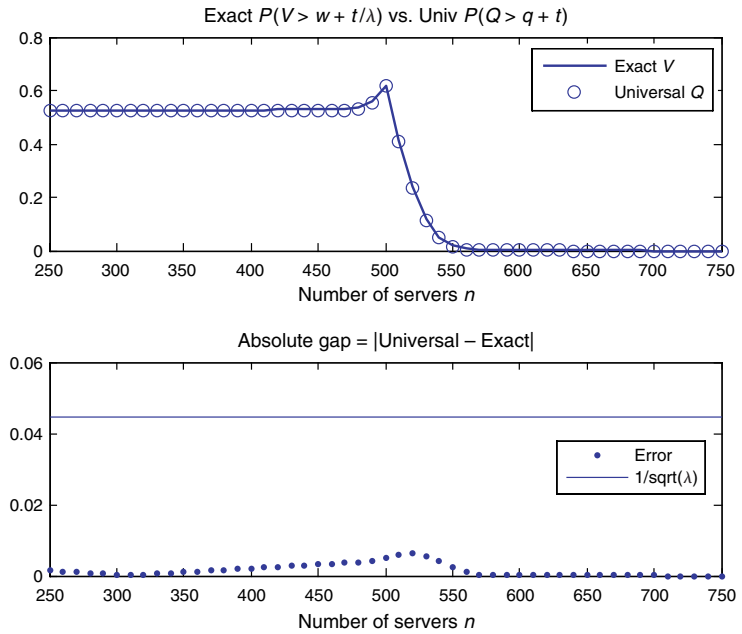

FIGURE 22. Approximating $\mathbb{P}\{V^\lambda(\infty) \geq w^\lambda - 2/\lambda\}$: fixed λ, varying n $(250 \leq n \leq 750)$.

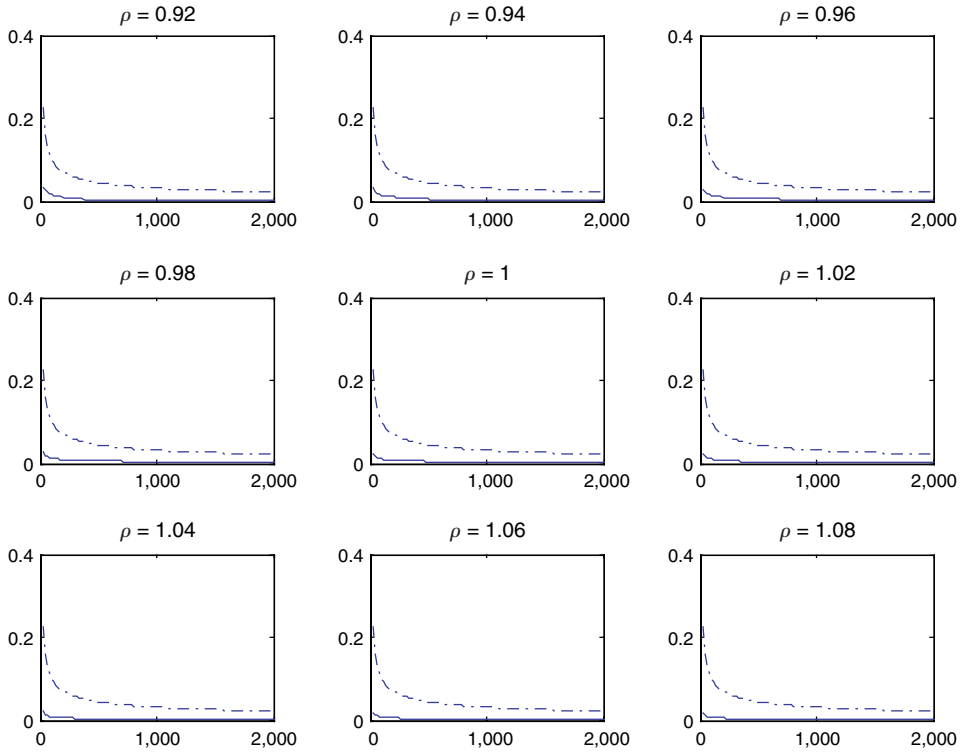

FIGURE 23. Approximating $\mathbb{P}\{V^\lambda(\infty) \geq w^\lambda - 2/\lambda\}$ fixed ρ, varying λ $(20 \leq \lambda \leq 2{,}000)$.

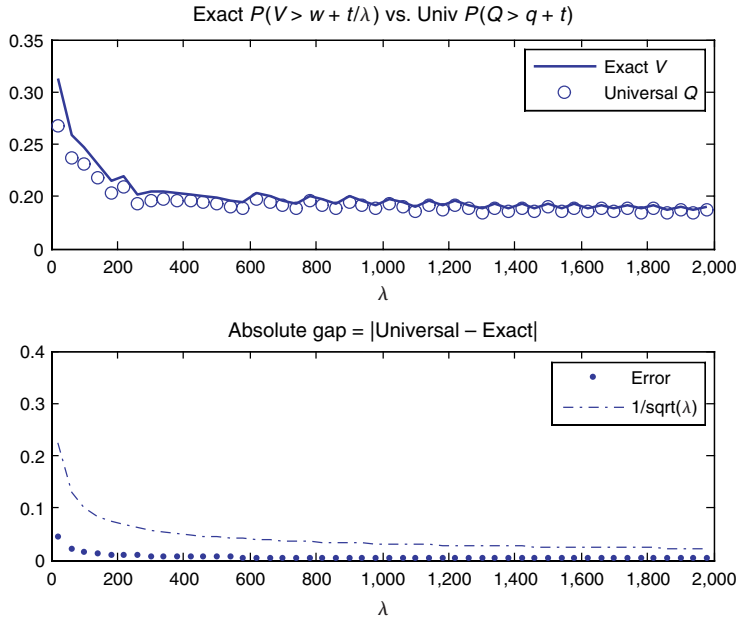

FIGURE 24. Approximating $\mathbb{P}\{V^\lambda(\infty) \geq w^\lambda - 2/\lambda\}$: varying ρ^λ with λ $(20 \leq \lambda \leq 2{,}000)$.

6.2. Toward a framework. It may be possible to extend our excursion-based approach to other queueing systems such as *networks* of queues or queues with *nonexponential* distributions. We could perhaps also cover *time-varying* models, after stabilizing their performance via appropriate staffing (Feldman et al. [13], Liu and Whitt [25]). Regardless of the setting, the following elements seem to be prerequisites for our framework to apply.

Markov structure and regeneration. We require models with dynamics that can be characterized by a Markov chain that has an appropriate regeneration point. Birth-and-death processes, of which the Erlang-A queue is

a special case, are to some extent the simplest cases that adhere to this structure. More generally, it is often possible to define a sufficiently rich state descriptor that renders the dynamics Markovian. A regenerative set then often replaces the regeneration point; see, e.g., Kaspi and Mandelbaum [20]. Because the intimate connection between stationary measures and cycle averages are known to hold also for Markov processes with more general regenerative sets (see, e.g., Asmussen [3, Chapter VII.3]), it may be possible to extend our analysis to these more general settings. This seems a challenging direction.

Martingale properties. To be able to apply Ito's lemma, as in our proof of Proposition 4, the dynamics must be represented as a semi-martingale. This, however, would not be enough. In order to obtain refined bounds, these martingales must be relatively "tractable." To illustrate the underlying complexity, we consider, for example, a general renewal arrival process. Provided that the interarrival times have finite second moments, it is then known (see Konstantopoulos et al. [24]) that $A(t) - \int_0^t h(a(s))\,ds$ is a martingale with respect to a properly defined filtration, where $h(\cdot)$ is the hazard rate of the interarrival time and $a(\cdot)$ is the age process. This martingale has the predictable quadratic variation process $\sigma^2\mu^2\int_0^t h(a(s))\,ds$, where σ^2 is the standard deviation of the interarrival time and $1/\mu$ is the mean interarrival time. That $\int_0^t h(a(s))\,ds \approx \mu t$, as $t \to \infty$, guarantees that the quadratic variation of the above martingale approaches $\sigma^2\mu^3 t$, and it is used in Konstantopoulos et al. [24] to establish a functional central limit theorem. Whereas such convergence suffices for purposes of weak convergence, we expect that some estimates on the "distance" $\int_0^t h(a(s))\,ds - \mu t$ are needed for refined bounds. The results in Konstantopoulos and Last [23] may be helpful in that regard.

Order bounds. Preliminary order bounds on steady-state metrics and on the expectation of underlying hitting times played a crucial role in our analysis. For the special case of the Erlang-A queue, we establish such bounds directly using Lyapunov function arguments; see Lemma 4.6. In exploring extensions to other queueing systems, it is useful that existing research already provides such bounds. For the case of generalized Jackson networks, as an example, order bounds for the steady-state queue length are given by Gamarnik and Zeevi [14].

Gradient bounds. This, in a sense, is the simplest of the required preliminaries. Given a differential equation that characterizes excursion-performance of the approximating diffusion process, one must establish gradient bounds for its solutions. For the diffusion in the current paper, we have explicit solutions for the corresponding ODE that allow one to directly derive the gradient bounds. In more general settings, the ODE may be replaced by a more complex Partial Differential Equation (PDE) for which closed-form solutions are not available. Yet it is plausible that, in those cases, gradient bounds can be established indirectly by relying on the rich theory of gradient bounds for solutions to PDEs.

Acknowledgments. The authors dedicate this paper to the late Uri Rothblum, their friend and teacher. Uri was the editor of MOR at his untimely passing. They are grateful to the editor, associate editor, and two anonymous referees for their careful constructive comments that guided our revision. The joint research of J. Huang and A. Mandelbaum was partially supported by the National University of Singapore (NUS); the funds for the promotion of research and sponsored research at the Technion, Haifa, Israel; the Technion SEELab; the Statistics and Applied Mathematical Sciences Institute (SAMSI) of the NSF, North Carolina, USA; the Department of Statistics and Operations Research (STOR), the University of North Carolina at Chapel Hill; the Department of Information, Operations and Management Sciences (IOMS), Leonard N. Stern School of Business, New York University; and the Department of Statistics, The Wharton School, University of Pennsylvania—the hospitality of all these institutions is gratefully acknowledged and truly appreciated. The work of A. Mandelbaum was partially supported by BSF [Grants 2005175 and 2008480], as well as ISF [Grant 1357/08]. Work on this paper was conducted also during visits of J. Huang to the Kellogg School of Management at Northwestern University. I. Gurvich and J. Huang would like to thank Kellogg and the Center for Mathematical Studies in Economics and Management Science (CMS-EMS) for their support of these visits.

Appendix A. Proofs of corollaries.

Proof of Corollary 1. We can, without loss of generality, assume that either $n^\lambda \geq \Delta^\lambda$ for all λ or that $n^\lambda < \Delta^\lambda$ for all λ. Otherwise, the argument below applies to each subsequence. We first consider the case $n^\lambda \geq \Delta^\lambda$. For each λ, let $f^\lambda(x) = (x + \Delta^\lambda - n^\lambda)^+$. Then,

$$\mathbb{E}[Q^\lambda(\infty)] = \mathbb{E}[f^\lambda(\tilde{X}^\lambda(\infty))] \qquad \text{and} \qquad \mathbb{E}[\tilde{Q}^\lambda(\infty)] = \mathbb{E}[f^\lambda(\tilde{Y}^\lambda(\infty))].$$

Fixing $0 < \epsilon < 1$, define $g_\epsilon^\lambda : \mathbb{R} \to \mathbb{R}$ as follows:

$$g_\epsilon^\lambda(x) := \begin{cases} 0, & x \leq -[\Delta^\lambda - n^\lambda] - \epsilon, \\ \dfrac{1}{4\epsilon}(x + \Delta^\lambda - n^\lambda + \epsilon)^2, & -[\Delta^\lambda - n^\lambda] - \epsilon \leq x < -[\Delta^\lambda - n^\lambda] + \epsilon, \\ x + \Delta^\lambda - n^\lambda, & x \geq -[\Delta^\lambda - n^\lambda] + \epsilon. \end{cases}$$

Then, $|g_\epsilon^\lambda(x)| \le 1 + |x|$ and $(g_\epsilon^\lambda)^{(1)}(x) \le 1$, for all x. The sequence $\{g_\epsilon^\lambda\}$ is, in turn, subpolynomial of order 1. From Theorem 1 it then follows that

$$\mathbb{E}[g_\epsilon^\lambda(\tilde{X}^\lambda(\infty))] - \mathbb{E}[g_\epsilon^\lambda(\tilde{Y}^\lambda(\infty))] = \mathcal{O}(1).$$

By construction, $|g_\epsilon^\lambda(x) - f^\lambda(x)| \le \epsilon/4$, for all $x \in \mathbb{R}$, which proves the result for the case $n^\lambda \ge \Delta^\lambda$. For the case $n^\lambda < \Delta^\lambda$, it suffices to prove

$$\mathbb{E}[(X^\lambda(\infty) - n^\lambda)^-] - \mathbb{E}[(Y^\lambda(\infty) - n^\lambda)^-] = \mathbb{E}[(n^\lambda - X^\lambda(\infty))^+] - \mathbb{E}[(n^\lambda - Y^\lambda(\infty))^+] = \mathcal{O}(1). \tag{A1}$$

Indeed,

$$\begin{aligned}\mathbb{E}[Q^\lambda(\infty)] - \mathbb{E}[\tilde{Q}^\lambda(\infty)] &= \mathbb{E}[X^\lambda(\infty)] - \mathbb{E}[Y^\lambda(\infty)] \\ &\quad + \mathbb{E}[(X^\lambda(\infty) - n^\lambda)^-] - \mathbb{E}[(Y^\lambda(\infty) - n^\lambda)^-],\end{aligned}$$

and by Theorem 1 we have that $\mathbb{E}[X^\lambda(\infty)] - \mathbb{E}[Y^\lambda(\infty)] = \mathbb{E}[\tilde{X}^\lambda(\infty)] - \mathbb{E}[\tilde{Y}^\lambda(\infty)] = \mathcal{O}(1)$. To prove (A1), let $\tilde{f}^\lambda(x) = (n^\lambda - \Delta^\lambda - x)^+$. Then,

$$\mathbb{E}[(X^\lambda(\infty) - n^\lambda)^-] = \mathbb{E}[\tilde{f}^\lambda(\tilde{X}^\lambda(\infty))]$$

and similarly for Y^λ. Fixing $0 < \epsilon < 1$, define $\tilde{g}_\epsilon^\lambda : \mathbb{R} \to \mathbb{R}$ as follows:

$$\tilde{g}_\epsilon^\lambda(x) := \begin{cases} n^\lambda - \Delta^\lambda - x, & x \le -[\Delta^\lambda - n^\lambda] - \epsilon, \\ \dfrac{1}{4\epsilon}(x + \Delta^\lambda - n^\lambda - \epsilon)^2, & -[\Delta^\lambda - n^\lambda] - \epsilon \le x < -[\Delta^\lambda - n^\lambda] + \epsilon, \\ 0, & x \ge -[\Delta^\lambda - n^\lambda] + \epsilon. \end{cases}$$

Then, $|\tilde{g}_\epsilon^\lambda(x)| \le 1 + |x|$ and $(\tilde{g}_\epsilon^\lambda)^{(1)}(x) \le 1$, for all x. The sequence $\{\tilde{g}_\epsilon^\lambda\}$ is, in turn, subpolynomial of order 1, and it follows from Theorem 1 that

$$\mathbb{E}[\tilde{g}_\epsilon^\lambda(\tilde{X}^\lambda(\infty))] - \mathbb{E}[\tilde{g}_\epsilon^\lambda(\tilde{Y}^\lambda(\infty))] = \mathcal{O}(1).$$

By construction, $|\tilde{g}_\epsilon^\lambda(x) - \tilde{f}^\lambda(x)| \le \epsilon/4$, for all $x \in \mathbb{R}$, and the result of the corollary follows. □

For Corollary 2 we require a lemma that guarantees that $X^\lambda(\infty)$ has no significant mass concentrated on any fixed point. (A similar result holds also for the density of Y^λ, but it is not needed for any of our derivations.)

Lemma A.1. *There exists an absolute constant ϑ such that for any $k \in \mathbb{N}$,*

$$\mathbb{P}\{X^\lambda(\infty) = k\} \le \vartheta \sqrt{\lambda}^{-1}.$$

Proof. The result can be equivalently proved for $\tilde{X}^\lambda(\infty) = X^\lambda(\infty) - \Delta^\lambda$. For $x \in \{-\Delta^\lambda, \dots, 0, 1, 2, \dots\}$, let $\nu^\lambda(x) = \mathbb{P}\{\tilde{X}^\lambda(\infty) = x\}$. We claim that 0 is a maximizer of ν^λ. Indeed, using the balance equation $\lambda\nu^\lambda(x) = (\mu((x + \Delta^\lambda) \wedge n^\lambda) + \theta(x + \Delta^\lambda - n^\lambda)^+)\nu^\lambda(x + 1)$, it is evident that $\nu^\lambda(\cdot)$ is nondecreasing for $x \le 0$ and nonincreasing for $x \ge 0$. In turn, it suffices to prove that $\nu^\lambda(0) = \mathcal{O}(\sqrt{\lambda}^{-1})$. Let τ_0^λ be the hitting time of $\tilde{X}^\lambda$ to 0. Since any point $x \in \mathbb{N}$ (in particular 0) is a regeneration point for the B&D process $\tilde{X}^\lambda$, we have

$$\nu^\lambda(0) = \frac{\mathbb{E}_0[\int_0^{\tau_0^\lambda} \mathbb{1}\{\tilde{X}^\lambda(s) = 0\}\, ds]}{\mathbb{E}_0[\tau_0^\lambda]}.$$

During such a regenerative cycle, the process $\tilde{X}^\lambda$ visits 0 only once so that $\mathbb{E}_0[\int_0^{\tau_0^\lambda} \mathbb{1}\{\tilde{X}^\lambda(s) = 0\}\, ds] = 1/(\lambda + \mu(\Delta^\lambda \wedge n^\lambda) + \theta(\Delta^\lambda - n^\lambda)^+) = 1/(2\lambda)$ and, in particular,

$$\nu^\lambda(0) = \frac{1}{2\lambda\, \mathbb{E}_0[\tau_0^\lambda]}.$$

Next note that $\mathbb{E}_0[\tau_0^\lambda] \ge p_1^\lambda\, \mathbb{E}_1[\tau_u^\lambda]$, where τ_u^λ is as defined in §4.1 and $p_1^\lambda = \lambda/(\lambda + \mu(\Delta^\lambda \wedge n^\lambda) + \theta(\Delta^\lambda - n^\lambda)^+) = 1/2$ is the transition probability from 0 to 1. By Proposition 3, $(\mathbb{E}_1[\tau_u^\lambda])^{-1} = \mathcal{O}(\sqrt{\lambda})$, and, in turn,

$$\nu^\lambda(0) \le \frac{1}{\lambda\, \mathbb{E}_1[\tau_u^\lambda]} = \mathcal{O}(\sqrt{\lambda}^{-1}).$$

This completes the proof. □

Proof of Corollary 2. Equation (17) directly implies (18). The converse holds noting that if (17) is not true, there must exist a sequence $\{a^\lambda\}$ such that (18) does not hold. We focus on proving (18). Let $\tilde{a}^\lambda = a^\lambda - \Delta^\lambda$. Using the centered process $\tilde{X}^\lambda$ and $\tilde{Y}^\lambda$, it is equivalent to prove

$$\mathbb{P}\{\tilde{X}^\lambda(\infty) \geq \tilde{a}^\lambda\} - \mathbb{P}\{\tilde{Y}^\lambda(\infty) \geq \tilde{a}^\lambda\} = \mathcal{O}(\sqrt{\lambda}^{-1}).$$

We can construct two sequences of increasing continuously differentiable functions $\{f^\lambda\} \in \mathcal{S}_0$ and $\{g^\lambda\} \in \mathcal{S}_0$ such that for all x and λ, the following properties hold:

$$f^\lambda(x) = \mathbb{1}_{\{x \geq \tilde{a}^\lambda\}}, \quad x \in (-\infty, \tilde{a}^\lambda - 1] \cup [\tilde{a}^\lambda, \infty)$$

and

$$|f^\lambda(x) - \mathbb{1}_{\{x \geq \tilde{a}^\lambda\}}| \leq g^\lambda(x) \leq \mathbb{1}_{\{x \in [\tilde{a}^\lambda - 2,\, \tilde{a}^\lambda + 1]\}}.$$

Then,

$$\begin{aligned}
&\mathbb{P}\{\tilde{X}^\lambda(\infty) \geq \tilde{a}^\lambda\} - \mathbb{P}\{\tilde{Y}^\lambda(\infty) \geq \tilde{a}^\lambda\} \\
&\quad = \mathbb{E}[f^\lambda(\tilde{X}^\lambda(\infty))] - \mathbb{E}[f^\lambda(\tilde{Y}^\lambda(\infty))] + \big(\mathbb{E}[\mathbb{1}_{\{\tilde{X}^\lambda(\infty) \geq \tilde{a}^\lambda\}}] - \mathbb{E}[f^\lambda(\tilde{X}^\lambda(\infty))]\big) + \big(\mathbb{E}[\mathbb{1}_{\{\tilde{Y}^\lambda(\infty) \geq \tilde{a}^\lambda\}}] - \mathbb{E}[f^\lambda(\tilde{Y}^\lambda(\infty))]\big),
\end{aligned}$$

and we have

$$\begin{aligned}
&\big|\mathbb{P}\{\tilde{X}^\lambda(\infty) \geq \tilde{a}^\lambda\} - \mathbb{P}\{\tilde{Y}^\lambda(\infty) \geq \tilde{a}^\lambda\}\big| \\
&\quad \leq \big|\mathbb{E}[f^\lambda(\tilde{X}^\lambda(\infty))] - \mathbb{E}[f^\lambda(\tilde{Y}^\lambda(\infty))]\big| + \big|\mathbb{E}[g^\lambda(\tilde{X}^\lambda(\infty))]\big| + \big|\mathbb{E}[g^\lambda(\tilde{Y}^\lambda(\infty))]\big| \\
&\quad \leq \big|\mathbb{E}[f^\lambda(\tilde{X}^\lambda(\infty))] - \mathbb{E}[f^\lambda(\tilde{Y}^\lambda(\infty))]\big| + \big|\mathbb{E}[g^\lambda(\tilde{X}^\lambda(\infty))] - \mathbb{E}[g^\lambda(\tilde{Y}^\lambda(\infty))]\big| + 2\,\mathbb{E}[g^\lambda(\tilde{X}^\lambda(\infty))].
\end{aligned}$$

The first two terms in the last line above are bounded by Theorem 1. Finally, by Lemma A.1,

$$\mathbb{E}[g^\lambda(\tilde{X}^\lambda(\infty))] = \mathbb{P}\{\tilde{X}^\lambda(\infty) \in (\tilde{a}^\lambda - 2, \tilde{a}^\lambda + 1)\} = \mathcal{O}(\sqrt{\lambda}^{-1}),$$

which concludes the proof of the corollary. □

Proof of Corollary 3. Define $q^\lambda = (\Delta^\lambda - n^\lambda)^+$. Then

$$\begin{aligned}
&\mathbb{E}[(Q^\lambda(\infty) - \mathbb{E}[Q^\lambda(\infty)])^2] - \mathbb{E}[(\tilde{Q}^\lambda(\infty) - \mathbb{E}[\tilde{Q}^\lambda(\infty)])^2] \\
&\quad = \mathbb{E}[(Q^\lambda(\infty) - q^\lambda + q^\lambda - \mathbb{E}[Q^\lambda(\infty)])^2] - \mathbb{E}[(\tilde{Q}^\lambda(\infty) - q^\lambda + q^\lambda - \mathbb{E}[\tilde{Q}^\lambda(\infty)])^2] \\
&\quad = \mathbb{E}[(Q^\lambda(\infty) - q^\lambda)^2] - \mathbb{E}[(\tilde{Q}^\lambda(\infty) - q^\lambda)^2] - ((\mathbb{E}[Q^\lambda(\infty)] - q^\lambda)^2 - (\mathbb{E}[\tilde{Q}^\lambda(\infty)] - q^\lambda)^2). \qquad \text{(A2)}
\end{aligned}$$

For the last term in the above,

$$\begin{aligned}
&(\mathbb{E}[Q^\lambda(\infty)] - q^\lambda)^2 - (\mathbb{E}[\tilde{Q}^\lambda(\infty)] - q^\lambda)^2 \\
&\quad = 2(\mathbb{E}[Q^\lambda(\infty)] - q^\lambda)(\mathbb{E}[Q^\lambda(\infty)] - \mathbb{E}[\tilde{Q}^\lambda(\infty)]) - (\mathbb{E}[Q^\lambda(\infty)] - \mathbb{E}[\tilde{Q}^\lambda(\infty)])^2.
\end{aligned}$$

For all $x, y \in \mathbb{R}$, it holds that $-(x-y)^- \leq (x)^+ - (y)^+ \leq (x-y)^+$ so that

$$-(x - \Delta^\lambda)^- \leq (x - n^\lambda)^+ - (\Delta^\lambda - n^\lambda)^+ \leq (x - \Delta^\lambda)^+.$$

Recalling that $q^\lambda = (\Delta^\lambda - n^\lambda)^+$ and that $Q^\lambda(\infty) = (X^\lambda(\infty) - n^\lambda)^+$, we have that

$$|\mathbb{E}[Q^\lambda(\infty)] - q^\lambda| = |\mathbb{E}[(X^\lambda(\infty) - n^\lambda)^+] - (\Delta^\lambda - n^\lambda)^+| \leq \mathbb{E}[|X^\lambda(\infty) - \Delta^\lambda|] \leq \sqrt{\mathbb{E}[(X^\lambda(\infty) - \Delta^\lambda)^2]} = \mathcal{O}(\sqrt{\lambda}),$$

where the last inequality follows from Jensen's inequality and the last equality follows from Theorem 2; see Remark 4.1. By Corollary 1, we have that $\mathbb{E}[Q^\lambda(\infty)] - \mathbb{E}[\tilde{Q}^\lambda(\infty)] = 1$, and we conclude that

$$(\mathbb{E}[Q^\lambda(\infty)] - q^\lambda)^2 - (\mathbb{E}[\tilde{Q}^\lambda(\infty)] - q^\lambda)^2 = \mathcal{O}(\sqrt{\lambda}). \qquad \text{(A3)}$$

Revisiting (A2), it is clear that, to complete the proof, it remains to prove that

$$\mathbb{E}[(Q^\lambda(\infty) - q^\lambda)^2] - \mathbb{E}[(\tilde{Q}^\lambda(\infty) - q^\lambda)^2] = \mathcal{O}(\sqrt{\lambda}).$$

Defining the function $f^\lambda(x) = [(x + \Delta^\lambda - n^\lambda)^+ - (\Delta^\lambda - n^\lambda)^+]^2$, we rewrite

$$\mathbb{E}[(Q^\lambda(\infty) - q^\lambda)^2] - \mathbb{E}[(\tilde{Q}^\lambda(\infty) - q^\lambda)^2] = \mathbb{E}[f^\lambda(\tilde{X}^\lambda(\infty))] - \mathbb{E}[f^\lambda(\tilde{Y}^\lambda(\infty))].$$

The sequence $\{f^\lambda\}$ is not subpolynomial since f^λ is not differentiable at $-(\Delta^\lambda - n^\lambda)$ when $\Delta^\lambda > n^\lambda$. Instead, given $\epsilon^\lambda \leq q^\lambda/2 \wedge 1$, define g^λ_ϵ as follows:

(i) If $\Delta^\lambda \leq n^\lambda$, $g^\lambda_\epsilon(x) = f^\lambda(x)$ for all $x \in \mathbb{R}$;

(ii) If $\Delta^\lambda > n^\lambda$, $g_\epsilon^\lambda(x) = f^\lambda(x)$ for $x \geq -\Delta^\lambda + n^\lambda$, and for $x < -\Delta^\lambda + n^\lambda$,

$$g_\epsilon^\lambda(x) = (\Delta^\lambda - n^\lambda)^2 + \epsilon^\lambda - \frac{(\Delta^\lambda - n^\lambda)^2}{\epsilon^\lambda}\left(\left(x + \frac{\epsilon^\lambda}{\Delta^\lambda - n^\lambda} + \Delta^\lambda - n^\lambda\right)^+\right)^2.$$

Then $\{g_\epsilon^\lambda\} \in \mathscr{S}_2$ and $\sup_{x\in\mathbb{R}} |f^\lambda(x) - g^\lambda(x)| \leq (\epsilon^\lambda)^2$. As in the proof of Corollary 1, we then have that

$$\begin{aligned}
&\mathbb{E}[(Q^\lambda(\infty) - q^\lambda)^2] - \mathbb{E}[(\tilde{Q}^\lambda(\infty) - q^\lambda)^2] \\
&\quad = \mathbb{E}[f^\lambda(\tilde{X}^\lambda(\infty))] - \mathbb{E}[f^\lambda(\tilde{Y}^\lambda(\infty))] \\
&\quad = \mathbb{E}[f^\lambda(\tilde{X}^\lambda(\infty))] - \mathbb{E}[g_\epsilon^\lambda(\tilde{X}^\lambda(\infty))] + \mathbb{E}[g_\epsilon^\lambda(\tilde{X}^\lambda(\infty))] \\
&\qquad - \mathbb{E}[g_\epsilon^\lambda(\tilde{Y}^\lambda(\infty))] + \mathbb{E}[g_\epsilon^\lambda(\tilde{Y}^\lambda(\infty))] - \mathbb{E}[f^\lambda(\tilde{Y}^\lambda(\infty))] \\
&\quad = \mathcal{O}(\sqrt{\lambda}).
\end{aligned} \tag{A4}$$

The corollary now follows by plugging (A3) and (A4) into the last line in (A2). □

Appendix B. Proofs of auxiliary lemmas.

Proof of Lemma 4.1. We start with the B&D process $\tilde{X}^\lambda$. Let $\mathscr{A}^\lambda$ be its generator and let $g(x) = e^{\delta x}$. Then, $g(x) \geq 1$ for all $x \geq 0$ and

$$\begin{aligned}
\mathscr{A}^\lambda g(x) &= \lambda(e^{\delta(x+1)} - e^{\delta x}) + (\lambda + \ell^\lambda(x))(e^{\delta(x-1)} - e^{\delta x}) \\
&= e^{\delta x}(\lambda(e^\delta - 1) + (\lambda + \ell^\lambda(x))(e^{-\delta} - 1)),
\end{aligned}$$

where $\ell^\lambda(x)$ is as in (24). Since $e^{-\delta} = 1 - \delta + o(\delta)$ and $e^\delta = 1 + \delta + o(\delta)$ we have

$$\mathscr{A}^\lambda g(x) = e^{\delta x}(\lambda\delta - \delta(\lambda + \ell^\lambda(x))) + (\lambda + \ell^\lambda(x))o(\delta)e^{\delta x} = -e^{\delta x}\delta\ell^\lambda(x) + (\lambda + \ell^\lambda(x))o(\delta)e^{\delta x}.$$

Since $\ell^\lambda(x)$ is strictly increasing, we can choose K_1 (which may depend on λ) sufficiently large so that $\ell^\lambda(x) \geq \lambda$ for all $x \geq K_1$. We can subsequently choose δ sufficiently small so that $o(\delta) \leq \delta/4$ and find c_2 and c_3 (which may depend on λ) such that

$$\mathscr{A}^\lambda g(x) \leq -c_3 g(x) + c_2 \mathbb{I}\{x \leq K_2\},$$

for all $x \in \mathbb{N}$. Since $g(x) \geq 1$, we can conclude the existence of ϑ_0 such that for all $y > K_1$, $\mathbb{E}_y[e^{\vartheta_0 \tau_{K_1}^\lambda}] \leq e^{\delta y} < \infty$, where $\tau_{K_1}^\lambda$ is the hitting time of K_1 (see, e.g., Roberts and Rosenthal [33, Corollary 2]). A similar argument is applied to $\tau_{-K_2}^\lambda$ (for some $-K_2 \geq -\Delta^\lambda$) to show the existence of ϑ_0 (possibly re-chosen) such that for all $y < -K_2$, $\mathbb{E}_y[e^{\vartheta_0 \tau_{-K_2}^\lambda}] < \infty$. Letting $\mathscr{K} = \{-K_2, \ldots, 0, 1, \ldots K_1\}$, we have established that for all $y \notin \mathscr{K}$, $\mathbb{E}_y[e^{\vartheta_0 \tau_{\mathscr{K}}^\lambda}] < \infty$ where $\tau_{\mathscr{K}}^\lambda$ is the hitting time of the set $\mathscr{K}$.

Finally, the existence of an exponential moment for the return time of a continuous-time Markov chain to a finite set implies that $\mathbb{E}_y[e^{\vartheta_0 \tau_y^\lambda}] < \infty$ for any y (where τ_y^λ is the hitting time of y); see Meyn and Tweedie [30, Chapter 15]. This concludes the proof for $\tilde{X}^\lambda$.

For the diffusion process $\tilde{Y}^\lambda$ it follows from Loukianova et al. [26, Theorem 1.1] that there exists ϑ_0 (possibly depending on λ and different from the above ϑ_0) such that both $\mathbb{E}_0[e^{\vartheta_0 \tilde{\tau}_u^\lambda}] < \infty$ and $\mathbb{E}_0[e^{\vartheta_0 \tilde{\tau}_l^\lambda}] < \infty$. Indeed, to apply the result in Loukianova et al. [26], one must verify certain conditions on the speed density and scale density (see Browne and Whitt [9, p. 471]) of the diffusion and these conditions can be verified directly. Finally, by the strong Markov property, $\mathbb{E}_y[e^{\vartheta_0 \tilde{\tau}^\lambda}] = \mathbb{E}_y[e^{\vartheta_0 \tilde{\tau}_u^\lambda}]\mathbb{E}_0[e^{\vartheta_0 \tilde{\tau}_l^\lambda}] < \infty$ for some constant ϑ_0. □

Proof of Lemma 4.2. Since $g(\cdot)$ is nondecreasing and $\tilde{X}^\lambda(t) \leq \tilde{X}^\lambda(0) + E(\lambda t)$ for all $t \geq 0$, we have for all such t that

$$\mathbb{E}_y\left[\int_0^{t\wedge\tau_u^\lambda} |g(\tilde{X}^\lambda(s))|\, ds\right] \leq t\,\mathbb{E}_y[g(y + E(\lambda t))] < \infty. \quad \square$$

Proof of Lemma 4.3. Recall that the predictable quadratic variation of the martingale M^λ is given by

$$\langle M^\lambda \rangle(t) = \lambda t + \mu \int_0^t Z^\lambda(s)\, ds + \theta \int_0^t Q^\lambda(s)\, ds,$$

which satisfies $\langle M^\lambda\rangle(s) = \int_0^s (2\lambda + \ell^\lambda(\tilde{X}^\lambda(u)))\,du$ for each t and all $s \le t \wedge \tau_u^\lambda$. By the optional sampling theorem, the stopped martingale $M^\lambda(\cdot \wedge \tau_u^\lambda)$ is itself a martingale and, furthermore, the stochastic integral

$$\int_0^{t\wedge\tau_u^\lambda} g(\tilde{X}^\lambda(s-))\,dM^\lambda(s)$$

is then itself a zero-mean martingale provided that (27) holds; see, e.g., Van der Vaart [37, Theorem 5.25].

We turn to the second part of the lemma. Since the jumps of $\tilde{X}^\lambda$ are of size 1, we have for $t \le \tau_u^\lambda$ that

$$\sum_{s\le t}(\Delta\tilde{X}^\lambda(s))^2 = \sum_{s\le t}|\Delta\tilde{X}^\lambda(s)| = E(\lambda t) + S\left(\mu\int_0^t Z^\lambda(s)\,ds\right) + N\left(\theta\int_0^t Q^\lambda(s)\,ds\right).$$

Recalling the square-integrable martingales M_a^λ, M_s^λ, and M_r^λ defined in §2 and the definition of $\ell^\lambda(\cdot)$ in (24) we have

$$\mathcal{M}^\lambda(t) := \sum_{s\le t}(\Delta\tilde{X}^\lambda(s))^2 - \int_0^t (2\lambda + \ell^\lambda(\tilde{X}^\lambda(s)))\,ds$$

is a square-integrable martingale. The stochastic integral

$$\mathcal{N}^\lambda(t) := \int_0^t g(\tilde{X}^\lambda(s-))\,d\mathcal{M}^\lambda(s)$$

is itself a square-integrable martingale provided that (27) holds; see, e.g., Van der Vaart [37, Theorem 5.25]. Note that (28) is equivalently written as $\mathbb{E}_y[\mathcal{N}^\lambda(t\wedge\tau_u^\lambda)] = 0$, which now holds by optional stopping. This concludes the proof. □

Proof of Lemma 4.4. Let $\tilde{X}_J^\lambda$ (J here stands for "jump") be a process that has the transition law of $\tilde{X}^\lambda$ on the states $\{1, 2, \dots\}$ but jumps instantaneously back to 1 when hitting 0. By Lemma 4.1, the process $\tilde{X}_J^\lambda$ is a positive recurrent Markov process, and consecutive visits to 0 are regeneration points. Thus,

$$\frac{\mathbb{E}_1[\int_0^{\tau_u^\lambda} f(\tilde{X}^\lambda(s))\,ds]}{\mathbb{E}_1[\tau_u^\lambda]} = \frac{\mathbb{E}_1[\int_0^{\hat{\tau}_J^\lambda} f(\tilde{X}_J^\lambda(s))\,ds]}{\mathbb{E}_1[\hat{\tau}_J^\lambda]}, \tag{B1}$$

where $\hat{\tau}_J^\lambda$ is the first hitting time of $\tilde{X}_J^\lambda$ to 0. We also have (see, e.g., Asmussen [3, Theorem 3.1])

$$\mathbb{E}[f(\tilde{X}_J^\lambda(\infty))] = \frac{\mathbb{E}_1[\int_0^{\hat{\tau}_J^\lambda} f(\tilde{X}_J^\lambda(s))\,ds]}{\mathbb{E}_1[\hat{\tau}_J^\lambda]} = \lim_{t\to\infty}\frac{1}{t}\int_0^t f(\tilde{X}_J^\lambda(s))\,ds.$$

The processes U^λ and $\tilde{X}_J^\lambda$ share the same transition law for all states except for 0 (in which $\tilde{X}_J^\lambda$ jumps instantaneously to 1). Initializing both U^λ and $\tilde{X}_J^\lambda$ at time $t=0$ in state 1 and using that $\ell^\lambda(x)$ is nondecreasing, it is straightforward to construct U^λ and $\tilde{X}_J^\lambda$ on a common sample space so that on each sample path,

$$U^\lambda(t) \le \tilde{X}_J^\lambda(t) \le U^\lambda(t) + 1.$$

In particular, since f is nondecreasing,

$$\frac{1}{t}\int_0^t f(U^\lambda(s))\,ds \le \frac{1}{t}\int_0^t f(\tilde{X}_J^\lambda(s))\,ds \le \frac{1}{t}\int_0^t f(U^\lambda(s)+1)\,ds$$

for each $t > 0$. Taking the limit $t\to\infty$, we then have that $\mathbb{E}[f(U^\lambda(\infty))] \le \mathbb{E}[f(\tilde{X}_J^\lambda(\infty))] \le \mathbb{E}[f(U^\lambda(\infty)+1)]$, which, by (B1), implies the result of the lemma. □

Proof of Lemma 4.5. Taking expectations in (25) and using the optional stopping theorem, we have that

$$\mathbb{E}_1[\tilde{X}^\lambda(t\wedge\tau_u^\lambda)] = 1 - \mathbb{E}_1\left[\int_0^{t\wedge\tau_u^\lambda}\ell^\lambda(\tilde{X}^\lambda(s))\,ds\right]. \tag{B2}$$

Lemma 4.2 with $g = \ell^\lambda$ there guarantees that the expectation of the integral is finite. The process $(E(\lambda t) - \lambda t,\ t \ge 0)$ is a martingale with respect to $\mathbb{F}^\lambda$. By the optional stopping theorem,

$$\mathbb{E}_1[E(\lambda(t\wedge\tau_u^\lambda))] = \lambda\,\mathbb{E}_1[t\wedge\tau_u^\lambda] \le \lambda\,\mathbb{E}_1[\tau_u^\lambda],$$

for all $t \ge 0$. By the monotone convergence theorem and Lemma 4.1, we have that $\mathbb{E}_1[E(\lambda(\tau_u^\lambda))] < \infty$. Because

$$0 \le \tilde{X}^\lambda(t\wedge\tau_u^\lambda) \le 1 + E(\lambda(t\wedge\tau_u^\lambda)) \le 1 + E(\lambda\tau_u^\lambda)$$

for all $t \ge 0$, thus $\tilde{X}^\lambda(t\wedge\tau_u^\lambda)$ is uniformly integrable in t, and taking $t\to\infty$ in (B2) we obtain the result of the lemma. □

Proof of Lemma 4.6. We first prove the upper bounds in (30) and (31). Specifically, we prove that there exist absolute constants $\{\vartheta_{u,m}, m \in \mathbb{N}\}$ such that, for any $m > 1$, $\mathbb{E}[(U^\lambda(\infty))^m] \le \vartheta_{u,m}\sqrt{\lambda}^m$. Equation (31) then follows trivially for $m > 1$ and it follows for $m = 1$ by the fact that $\mathbb{E}[U^\lambda(\infty)] \le \sqrt{\mathbb{E}[(U^\lambda(\infty))^2]}$. Finally, the upper bound in (30) is because $|\ell^\lambda(x)| \le (\vartheta_2 + \vartheta_3)(\sqrt{\lambda} + x)$ for all $x \ge 0$; see (26).

Let $\mathcal{U}^\lambda$ be the generator of the B&D process $(U^\lambda(t), t \ge 0)$. Let $g(x) = x^m$. Then, for all $x \in \mathbb{N}$,

$$\mathcal{U}^\lambda g(x) = \lambda(g(x+1) - g(x)) + (\lambda + \ell^\lambda(x))(g(x-1) - g(x)).$$

Since $g(x+1) - g(x) = \sum_{k \le m, k \ne 0} \binom{m}{k} x^{m-k}$ and $g(x-1) - g(x) = \sum_{k \le m, k \ne 0} \binom{m}{k} x^{m-k}(-1)^k$,

$$\begin{aligned}\mathcal{U}^\lambda g(x) &= \sum_{k \le m, k \ne 0} \ell^\lambda(x)\binom{m}{k}x^{m-k}(-1)^k + \sum_{k \text{ even}, k \ne 0} (2\lambda)\binom{m}{k}x^{m-k} \\ &= \sum_{2 \le k \le m} \ell^\lambda(x)\binom{m}{k}x^{m-k}(-1)^k + \sum_{k \text{ even}, k \ne 0} (2\lambda)\binom{m}{k}x^{m-k} - mx^{m-1}\ell^\lambda(x).\end{aligned}$$

By the structure of $\ell^\lambda(x)$, we can choose absolute constants c_1, c_2 such that $\mathcal{U}^\lambda g(x) \le -c_2 x^m$ for $x \ge c_1\sqrt{\lambda}$. There then exists another absolute constant c_3 such that $\mathcal{U}^\lambda g(x) \le c_3\sqrt{\lambda}^m$, for $x \le c_1\sqrt{\lambda}$. Overall, we can find absolute constants c_4, c_5 (possibly depending on m) such that

$$\mathcal{U}^\lambda g(x) \le -c_4 x^m + c_5\sqrt{\lambda}^m,$$

for all $x \ge 0$. Applying expectations we conclude (see, e.g., Glynn and Zeevi [16, Corollary 1]) that

$$\mathbb{E}[(U^\lambda(\infty))^m] \le \frac{c_5}{c_4}\sqrt{\lambda}^m.$$

Letting $\vartheta_{u,m} = c_5/c_4$ concludes the argument.

We use a Lyapunov function argument also to prove that $\mathbb{E}[\ell^\lambda(U^\lambda(\infty))] \ge \vartheta_l\sqrt{\lambda}$ for some absolute constant ϑ_l. Define a (sequence of) functions g^λ as follows (the absolute constant c_6 will be determined below):

$$g^\lambda(x) = \begin{cases} x^2; & x \le \lceil\sqrt{c_6\lambda}\rceil; \\ (2\lceil\sqrt{c_6\lambda}\rceil + 1)x - (\lceil\sqrt{c_6\lambda}\rceil + 1)\lceil\sqrt{c_6\lambda}\rceil; & x > \lceil\sqrt{c_6\lambda}\rceil. \end{cases}$$

Then

$$\mathcal{U}^\lambda g^\lambda(x) = \begin{cases} \lambda; & x = 0; \\ 2\lambda - (2x-1)\ell^\lambda(x); & 1 \le x \le \lceil\sqrt{c_6\lambda}\rceil; \\ -(2\lceil\sqrt{c_6\lambda}\rceil + 1)\ell^\lambda(x); & x > \lceil\sqrt{c_6\lambda}\rceil. \end{cases}$$

We claim that we can choose c_6 in the definition of g^λ together with absolute constants c_7, c_8 such that

$$\mathcal{U}^\lambda g^\lambda(x) \ge c_7\lambda - c_8\sqrt{\lambda}\ell^\lambda(x),$$

in which case, since $\mathbb{E}[\mathcal{U}^\lambda g^\lambda(U^\lambda(\infty))] = 0$, we conclude that $\mathbb{E}[\ell^\lambda(U^\lambda(\infty))] \ge c_7\sqrt{\lambda}/c_8$ as required. To show the existence of such constants, we pick $c_7 \le 1$ and choose c_6, c_8 such that

$$\lambda - (2\lceil\sqrt{c_6\lambda}\rceil - 1)\ell^\lambda(\lceil\sqrt{c_6\lambda}\rceil) \ge 0, \tag{B3}$$

$$(c_8\sqrt{\lambda} - (2\lceil\sqrt{c_6\lambda}\rceil + 1))\ell^\lambda(\lceil\sqrt{c_6\lambda}\rceil) \ge c_7\lambda. \tag{B4}$$

Specifically, since ℓ^λ satisfies $\ell^\lambda(x) \ge (\mu \wedge \theta)x$, we can choose c_6 sufficiently small in (B3) and subsequently choose c_8 sufficiently large in (B4) so that both hold. □

Proof of Lemma 4.7. Fix $d^\lambda = \int_0^\infty (f^\lambda(s)/\lambda)e^{-\int_0^s \ell^\lambda(u)/\lambda\, du}\, ds$ and consider the first-order ODE

$$\begin{aligned} -\ell^\lambda(x)\nu^\lambda(x) + \lambda(\nu^\lambda)^{(1)}(x) &= -f^\lambda(x), \\ \nu^\lambda(0) &= d^\lambda. \end{aligned} \tag{B5}$$

From the Lipschitz continuity of the coefficients and known results in ODE theory (see, e.g., Teschl [36, Theorem 2.2 and Lemma 2.3]) it follows that the ODE (B5) has a unique solution that is infinitely differentiable. In turn, $u^\lambda_{f^\lambda}(x) = \int_0^x \nu^\lambda(u)\, du$ is a solution for (32).

By direct differentiation it is verified that

$$(u^\lambda_{f^\lambda})^{(1)}(x) = \nu^\lambda(x) = \int_x^\infty \frac{f^\lambda(s)}{\lambda} e^{-\int_x^s (\ell^\lambda(x)/\lambda)\,du}\,ds$$

is the corresponding unique solution to (B5). Recall (see (24)) that

$$\begin{aligned}\ell^\lambda(x) &= -\lambda + \mu((x+\Delta^\lambda)\wedge n^\lambda) + \theta(x+\Delta^\lambda - n^\lambda)^+\\ &= -\lambda + \mu(n^\lambda\wedge\Delta^\lambda) + \theta(\Delta^\lambda - n^\lambda)^+\\ &\quad + \mu((x+\Delta^\lambda)\wedge n^\lambda - \Delta^\lambda\wedge n^\lambda) + \theta((x+\Delta^\lambda - n^\lambda)^+ - (\Delta^\lambda - n^\lambda)^+)\\ &\geq (\mu\wedge\theta)x - \kappa\sqrt{\lambda},\end{aligned}$$

where by the definition of Δ^λ, $-\lambda + \mu(n^\lambda\wedge\Delta^\lambda) + \theta(\Delta^\lambda - n^\lambda)^+ \geq -\kappa\sqrt{\lambda}$ for some $\kappa > 0$ and $x \geq 0$. Thus, in (26) one can take $\vartheta_1 = \mu\wedge\theta$ and $\vartheta_3 = \kappa$. (In fact, the above holds with $\kappa = 0$. We allow for $\kappa > 0$ so that this proof can be reused without change for the NDS regime in §C.)

Subsequently, we can choose $C_{0,m}$, $C_{1,m}$, and $C_{2,m}$ such that

$$\begin{aligned}|(u^\lambda_{f^\lambda})^{(1)}(x)| &\leq \int_x^\infty \frac{a_1\sqrt{\lambda}^m + a_2 s^m}{\lambda} e^{-\int_x^s ((\mu\wedge\theta)u - \kappa\sqrt{\lambda})/\lambda\,du}\,ds\\ &= \frac{a_1\sqrt{\lambda}^{m-1}}{\mu\wedge\theta} e^{((\mu\wedge\theta)x - \kappa\sqrt{\lambda})^2/(2(\mu\wedge\theta)\lambda)} \int_{((\mu\wedge\theta)x/\sqrt{\lambda})-\kappa}^\infty e^{-(s^2/(2(\mu\wedge\theta)))}\,ds\\ &\quad + \frac{a_2\sqrt{\lambda}^{m-1}}{(\mu\wedge\theta)^{m+1}} e^{((\mu\wedge\theta)x - \kappa\sqrt{\lambda})^2/(2(\mu\wedge\theta)\lambda)} \int_{((\mu\wedge\theta)x/\sqrt{\lambda})-\kappa}^\infty (s+\kappa)^m e^{-(s^2/(2(\mu\wedge\theta)))}\,ds\\ &\leq \frac{a_1\sqrt{\lambda}^{m-1}}{\mu\wedge\theta} C_{0,m} + \frac{a_2\sqrt{\lambda}^{m-1}}{(\mu\wedge\theta)^{m+1}}\left(C_{1,m} + C_{2,m}\left(\frac{\mu\wedge\theta}{\sqrt{\lambda}}x\right)^{m-1}\right)\\ &= \left(\frac{a_1 C_{0,m}}{\mu\wedge\theta} + \frac{a_2 C_{1,m}}{(\mu\wedge\theta)^{m+1}}\right)\sqrt{\lambda}^{m-1} + C_{2,m}\frac{a_2}{(\mu\wedge\theta)^2}x^{m-1}.\end{aligned}$$

For the second inequality, we use the fact that $\lim_{z\to\infty} e^{z^2/2}\int_z^\infty e^{-s^2/2}\,ds = 0$ and $\lim_{x\to\infty}(e^{z^2/2}/z^{m-1})\int_z^\infty (s+a)^m e^{-s^2/2}\,ds = 1$. Thus, in particular, there exists K such that when $z \geq K$, $e^{z^2/2}\int_z^\infty (s+a)^m e^{-s^2/2}\,ds \leq 2z^{m-1}$. For values $z \leq K$, the function $g(z) = e^{z^2/2}\int_z^\infty (s+a)^m e^{-s^2/2}\,ds$ (being continuous) is bounded by a constant. Finally, set $A_{1,m} = ((a_1 C_{0,m})/\mu\wedge\theta + (a_2 C_{1,m})/(\mu\wedge\theta)^{m+1}) \vee (C_{2,m}(a_2/(\mu\wedge\theta)^2))$ to obtain (33).

To prove (34) and (35) we plug the bound on $(u^\lambda_{f^\lambda})^{(1)}(x)$ back into the ODE (32). Using the subpolynomiality of f^λ, we can choose an absolute constant $A_{2,m}$ such that

$$\begin{aligned}|(u^\lambda_{f^\lambda})^{(2)}(x)| &\leq \frac{1}{\lambda}\left(a_1\sqrt{\lambda}^m + a_2 x^m + (1+\ell^\lambda(x))|(u^\lambda_{f^\lambda})^{(1)}(x)|\right)\\ &\leq \frac{1}{\lambda}\left(a_1\sqrt{\lambda}^m + a_2 x^m + A_{1,m}(1+\ell^\lambda(x))(x^{m-1} + (\sqrt{\lambda})^{m-1})\right)\\ &\leq \frac{1}{\lambda}A_{2,m}\left(x^m + (\sqrt{\lambda})^m\right).\end{aligned}$$

Taking derivatives on both sides of (32), we get

$$|(u^\lambda_{f^\lambda})^{(3)}(x)| \leq \frac{1}{\lambda}\left(|(f^\lambda)^{(1)}(x)| + \theta\vee\mu|(u^\lambda_{f^\lambda})^{(1)}(x)| + |\ell^\lambda(x)(u^\lambda_{f^\lambda})^{(2)}(x)|\right).$$

Plugging in the bounds for the first and second derivatives and using (26) concludes the proof, recalling (see Definition 3.1) that for $m \geq 1$, $|(f^\lambda)^{(1)}(x)| \leq a_1\sqrt{\lambda}^{m-1} + a_2|x|^{m-1}$.

We next prove (39). We fix λ for the remainder of the proof so that the constants in the various bounds may depend on λ and are not necessarily absolute constants. The following is a standard argument in relating SDEs to PDEs/ODEs. We provide the detailed argument for completeness.

The process $\tilde{Y}^\lambda$ satisfies trivially a piecewise-linear growth condition on the drift and the (constant) diffusion coefficient. By the assumptions of the lemma $u^\lambda_{f^\lambda}$ has first and second derivatives that grow at sub-exponential rate so that $u^\lambda_{f^\lambda}$ is in the domain of the generator of $\tilde{Y}^\lambda$ (see, e.g., Klebaner [22, Theorem 6.11]). In turn,

$$\mathbb{E}_y[u^\lambda_{f^\lambda}(\tilde{Y}^\lambda(t\wedge\tilde{\tau}^\lambda))] = u^\lambda_{f^\lambda}(y) - \mathbb{E}_y\left[\int_0^{t\wedge\tilde{\tau}^\lambda_u} \ell^\lambda(\tilde{Y}^\lambda(s))(u^\lambda_{f^\lambda})^{(1)}(\tilde{Y}^\lambda(s))\,ds\right] + \mathbb{E}_y\left[\int_0^{t\wedge\tilde{\tau}^\lambda_u} \lambda(u^\lambda_{f^\lambda})^{(2)}(\tilde{Y}^\lambda(s))\,ds\right].$$

Since $u^\lambda_{f^\lambda}$ solves (32) we have, for each $t\ge 0$, that

$$\mathbb{E}_y[u^\lambda_{f^\lambda}(\tilde{Y}^\lambda(t\wedge\tilde{\tau}^\lambda_u))] = u^\lambda_{f^\lambda}(y) - \mathbb{E}_y\left[\int_0^{t\wedge\tilde{\tau}^\lambda_u} f^\lambda(\tilde{Y}^\lambda(s))\,ds\right]. \tag{B6}$$

Lemma 4.1 guarantees that $\tilde{\tau}^\lambda_u$ is almost surely finite so that

$$\lim_{t\to\infty}\tilde{Y}^\lambda(t\wedge\tilde{\tau}^\lambda_u)=0 \quad\text{and}\quad \lim_{t\to\infty}u^\lambda_{f^\lambda}(\tilde{Y}^\lambda(t\wedge\tilde{\tau}^\lambda_u))=0$$

almost surely. To conclude that

$$\mathbb{E}_y[u^\lambda_{f^\lambda}(\tilde{Y}^\lambda(t\wedge\tilde{\tau}^\lambda_u))]\to 0 \quad \text{as } t\to\infty, \tag{B7}$$

it remains to show that $u^\lambda_{f^\lambda}(\tilde{Y}^\lambda(t\wedge\tilde{\tau}^\lambda_u))$ is uniformly integrable in t. To that end, by (33) we have

$$u^\lambda_{f^\lambda}(\tilde{Y}^\lambda(t\wedge\tilde{\tau}^\lambda_u)) \le A_{1,m}\left(\frac{1}{m}(\tilde{Y}^\lambda(t\wedge\tilde{\tau}^\lambda_u))^m + (\sqrt{\lambda})^{m-1}\tilde{Y}^\lambda(t\wedge\tilde{\tau}^\lambda_u)\right),$$

so it remains only to prove the uniform integrability of $\{(\tilde{Y}^\lambda(t\wedge\tilde{\tau}^\lambda_u))^m, t\ge 0\}$. In fact, it suffices to prove the uniform integrability of $\{(B(t\wedge\tilde{\tau}^\lambda_u))^m, t\ge 0\}$. Indeed, since $\tilde{Y}^\lambda\ge 0$ on $[0,\tilde{\tau}^\lambda_u)$ and using (26), we have

$$\tilde{Y}^\lambda(t) = y - \int_0^t \ell^\lambda(\tilde{Y}^\lambda(s))\,ds + \sqrt{2\lambda}B(t) \le y + c_1\sqrt{\lambda}t + \sqrt{2\lambda}B(t) \tag{B8}$$

so that since $\mathbb{E}_1[(\tilde{\tau}^\lambda_l)^k]<\infty$ for any $k\in\mathbb{N}$ (see Lemma 4.1), the uniform integrability of $\{(\tilde{Y}^\lambda(t\wedge\tilde{\tau}^\lambda_u))^m, t\ge 0\}$ follows from that of $\{(B(t\wedge\tilde{\tau}^\lambda_u))^m, t\ge 0\}$, which we prove next.

The process $\mathscr{B}(t)=\exp(\eta B(t)-(\eta^2/2)t)$ is a martingale for any $\eta\in(-\infty,\infty)$ and so is the stopped martingale $\mathscr{B}(t\wedge\tilde{\tau}^\lambda_u)$; see, e.g., Klebaner [22, Theorems 3.7 and 7.14]. In turn, $\mathbb{E}_y[\mathscr{B}(t\wedge\tilde{\tau}^\lambda_u)]=1$ for all $y, t\ge 0$. (Recall that here $\mathbb{E}_y[\,\cdot\,]$ is the expectation conditional on $\tilde{Y}^\lambda(0)=y$ and not on $\mathscr{B}(0)=y$.) By Holder's inequality

$$\mathbb{E}_y\left[\exp\left(\frac{\eta}{2}B(t\wedge\tilde{\tau}^\lambda_u)\right)\right] = \mathbb{E}_y\left[\sqrt{\mathscr{B}(t\wedge\tilde{\tau}^\lambda_u)\exp\left(\frac{\eta^2}{2}(t\wedge\tilde{\tau}^\lambda_u)\right)}\right] \le \sqrt{\mathbb{E}_y[\mathscr{B}(t\wedge\tilde{\tau}^\lambda_u)]}\sqrt{\mathbb{E}_y\left[\exp\left(\frac{\eta^2}{2}(t\wedge\tilde{\tau}^\lambda_u)\right)\right]} = \sqrt{\mathbb{E}_y\left[\exp\left(\frac{\eta^2}{2}(t\wedge\tilde{\tau}^\lambda_u)\right)\right]}.$$

A similar argument is applied to the martingale $\tilde{\mathscr{B}}(t)=\exp(-\eta B(t)-(\eta^2/2)t)$. By Lemma 4.1 $\mathbb{E}_y[\exp((\eta^2/2)\tilde{\tau}^\lambda_u)]<\infty$ for all sufficiently small η. By the dominated convergence theorem it then holds that $\lim_{\eta\to 0}\mathbb{E}_y[\exp((\eta^2/2)\tilde{\tau}^\lambda_u)]=1$, and we can choose η small such that $\sqrt{\mathbb{E}_y[\exp((\eta^2/2)(t\wedge\tilde{\tau}^\lambda_u))]}\le 2$. Fixing such η, we have

$$\mathbb{E}_y\left[\exp\left(\frac{\eta}{2}|B(t\wedge\tilde{\tau}^\lambda_u)|\right)\right] \le \mathbb{E}_y\left[\exp\left(\frac{\eta}{2}B(t\wedge\tilde{\tau}^\lambda_u)\right)\right] + \mathbb{E}_y\left[\exp\left(-\frac{\eta}{2}B(t\wedge\tilde{\tau}^\lambda_u)\right)\right] \le 4.$$

In particular, $\mathbb{E}_y[\exp(\eta/2|B(t\wedge\tilde{\tau}^\lambda_u)|)]$ is uniformly bounded in t. We conclude that the sequence $\{(B(t\wedge\tilde{\tau}^\lambda_u))^m, t\ge 0\}$ is uniformly integrable and so is, by (B8), the family $\{(\tilde{Y}^\lambda(t\wedge\tilde{\tau}^\lambda_u))^m, t\ge 0\}$, which proves (B7). It remains to show that as $t\to\infty$,

$$\mathbb{E}_y\left[\int_0^{t\wedge\tilde{\tau}^\lambda_u} f^\lambda(\tilde{Y}^\lambda(s))\,ds\right] \to \mathbb{E}_y\left[\int_0^{\tilde{\tau}^\lambda_u} f^\lambda(\tilde{Y}^\lambda(s))\,ds\right]. \tag{B9}$$

By the almost sure finiteness of $\tilde{\tau}_u^\lambda$ (and the finiteness of $\tilde{Y}^\lambda$ on finite intervals), we have that

$$\int_0^{t\wedge\tilde{\tau}_u^\lambda} f^\lambda(\tilde{Y}^\lambda(s))\,ds \to \int_0^{\tilde{\tau}_u^\lambda} f^\lambda(\tilde{Y}^\lambda(s))\,ds$$

almost surely. Since $f^\lambda \in \mathcal{S}_m$, it suffices to prove that

$$\mathbb{E}_y\left[\int_0^{\tilde{\tau}_u^\lambda} (a_1\sqrt{\lambda}^m + a_2(\tilde{Y}^\lambda(s))^m)\,ds\right] < \infty, \tag{B10}$$

which will allow us to apply the dominated convergence theorem to obtain (B9). Equation (B10) follows from (B8) by the uniform integrability of $\{|B(t\wedge\tilde{\tau}_u^\lambda)|, t\ge 0\}$, which, through Doob's inequality, implies also that of $\{\sup_{0\le s\le t}|B(s\wedge\tilde{\tau}_u^\lambda)|,\ t\ge 0\}$. We conclude that (B9) holds. Plugging (B9) and (B10) into (B6) completes the proof of the lemma. □

Proof of Lemma 4.8. We first prove (43). Recall that given that $\tilde{X}^\lambda(0) = y \ge 0$, $\tilde{X}^\lambda(t\wedge\tau_u^\lambda) \le y + E(\lambda(t\wedge\tau_u^\lambda))$. Using Lemma 4.7 we have that

$$\mathscr{V}^\lambda(\tilde{X}^\lambda(t\wedge\tau_u^\lambda)) \le a + b(y + E(\lambda\tau_u^\lambda))^m,$$

for some (not necessarily absolute) constants a, b. As in the proof of Lemma 4.5, we then have that $\mathscr{V}^\lambda(\tilde{X}^\lambda(t\wedge\tau_u^\lambda))$ is uniformly integrable in t. By Lemma 4.1, $\mathscr{V}^\lambda(\tilde{X}^\lambda(t\wedge\tau_u^\lambda)) \to 0$ as $t\to\infty$ almost surely, combined with the uniform integrability yields (43).

Next we prove (44). Taking $g(x) = x^l$ in Lemma 4.3 we have

$$\mathbb{E}_y\left[\sum_{s\le t\wedge\tau_u^\lambda:\,|\Delta\tilde{X}^\lambda(s)|>0} (\tilde{X}^\lambda(s-))^l\right] = \mathbb{E}_y\left[\int_0^{t\wedge\tau_u^\lambda} (2\lambda + \ell^\lambda(\tilde{X}^\lambda(s-)))(\tilde{X}^\lambda(s))^l\,ds\right].$$

As $\tilde{X}^\lambda \ge 0$ on $[0,\tau_u^\lambda)$ and since $2\lambda + \ell^\lambda(\tilde{X}^\lambda)$ is nonnegative, the required convergence now follows from the monotone convergence theorem. □

Proof of Lemma 5.1. For each $\lambda, n \in \mathbb{R}_+$, denote $g(\lambda,n) = \mathbb{E}[\tilde{Q}_n^\lambda(\infty)]$. We must establish that for each λ, $g(\lambda,n)$ is continuous and nonincreasing in n, and that

$$g(\lambda,n) \to 0, \quad \text{as } n\to\infty. \tag{B11}$$

In addition, we show $g(\lambda,0)/(\lambda/\theta) \to 1$ when $\lambda\to\infty$. These guarantee that for any $\alpha(\lambda) \le \alpha \in (0,1)$, there then exists $n(\lambda)$ such that $\theta\mathbb{E}[\tilde{Q}_{n(\lambda)}] = \lambda\alpha(\lambda)$ for all large λ.

Recall that

$$\mathbb{E}[\tilde{Q}_n^\lambda(\infty)] = \frac{\sqrt{\lambda}}{\sqrt{\theta}}[1 - p(\beta_n,\mu,\theta)][h(\beta_n/\sqrt{\theta}) - \beta_n/\sqrt{\theta}], \tag{B12}$$

where $\beta_n = (n\mu - \lambda)/\sqrt{\lambda}$. Continuity of $g(\lambda,n)$ follows trivially from the continuity of p and h, which, in turn, follows from the continuity of the normal density and distribution functions. To prove (B11), note that if $n\to\infty$ then $\beta_n\to\infty$. It is known that as $\beta_n\to\infty$, $1 - p(\beta_n,\mu,\theta)\to 0$ (see the proof of Garnett et al. [15, Theorem 4]) and $(h(\beta_n/\sqrt{\theta}) - \beta_n/\sqrt{\theta}) \to 0$ (see the proof of Mandelbaum and Zeltyn [28, Theorem 4.1]). Thus, $g(\lambda,n)\to 0$ as $n\to\infty$.

Next note that $g(\lambda,0) = (\sqrt{\lambda}/\sqrt{\theta})[1 - p(-\sqrt{\lambda},\mu,\theta)][h(-\sqrt{\lambda}/\sqrt{\theta}) + \sqrt{\lambda}/\sqrt{\theta}]$, in which

$$\lim_{\lambda\to\infty}\left[1 - p\left(-\sqrt{\lambda},\mu,\theta\right)\right] = 1 \quad \text{and} \quad \lim_{\lambda\to\infty}\left[h\left(-\sqrt{\lambda}/\sqrt{\theta}\right) + \sqrt{\lambda}/\sqrt{\theta}\right]/(\sqrt{\lambda/\theta}) = 1.$$

Hence $g(\lambda,0)/(\lambda/\theta)\to 1$ when $\lambda\to\infty$.

Finally, $n(\lambda)$ is unique follows from the monotonicity of the right-hand side of (B12) in β_n (see Mandelbaum and Zeltyn [28, Remark 4.2]) and, in turn, that of $g(\lambda,n)$ in n. □

Proof of Lemma 5.2. Fix two sequences $\{n_1^\lambda\}$ and $\{n_2^\lambda\}$ as in the statement of the lemma as well as a Brownian motion B. Let the sequences of diffusion processes $\{Y_1^\lambda\}$ and $\{Y_2^\lambda\}$ be defined as solutions to the two SDEs:

$$Y_1^\lambda(t) = Y_1^\lambda(0) + (\lambda - n_1^\lambda\mu)t + \mu\int_0^t (Y_1^\lambda(s))^-\,ds - \theta\int_0^t (Y_1^\lambda(s))^+\,ds + \sqrt{2\lambda}B(t),$$

and

$$Y_2^\lambda(t) = Y_2^\lambda(0) + (\lambda - n_2^\lambda\mu)t + \mu\int_0^t (Y_2^\lambda(s))^-\,ds - \theta\int_0^t (Y_2^\lambda(s))^+\,ds + \sqrt{2\lambda}B(t).$$

We must prove that if $n_1^\lambda - n_2^\lambda = \mathcal{O}(1)$, then $\mathbb{E}[(Y_1^\lambda(\infty))^+] - \mathbb{E}[(Y_2^\lambda(\infty))^+] = \mathcal{O}(1)$.

Define $\hat{Y}_1^\lambda(t) = Y_1^\lambda(t) + n_1^\lambda - \Delta_1^\lambda$ and $\hat{Y}_2^\lambda(t) = Y_2^\lambda(t) + n_2^\lambda - \Delta_2^\lambda$ with

$$\Delta_1^\lambda = \frac{\lambda}{\mu} - \left(\frac{\lambda}{\mu} - n_1^\lambda\right)^+\left(1 - \frac{\mu}{\theta}\right) \qquad \text{and} \qquad \Delta_2^\lambda = \frac{\lambda}{\mu} - \left(\frac{\lambda}{\mu} - n_2^\lambda\right)^+\left(1 - \frac{\mu}{\theta}\right).$$

We claim that

$$\mathbb{E}[(\hat{Y}_1^\lambda(\infty) - n_1^\lambda + \Delta_1^\lambda)^+] - \mathbb{E}[(\hat{Y}_2^\lambda(\infty) - n_2^\lambda + \Delta_2^\lambda)^+] = \mathcal{O}(1). \tag{B13}$$

Indeed, this follows from Remark 4.2 that guarantees that Theorem 1 and, in turn, Corollary 1 holds for both $Y_1^\lambda + n_1^\lambda$ and $Y_2^\lambda + n_2^\lambda$ relative to X^λ where X^λ is the B&D process with $\lceil n_1^\lambda \rceil$ as the number of servers. Namely, we have that both $\mathbb{E}[(\hat{Y}_1^\lambda(\infty) - n_1^\lambda + \Delta_1^\lambda)^+] - \mathbb{E}[(X^\lambda(\infty) - n_1^\lambda + \Delta_1^\lambda)^+] = \mathcal{O}(1)$ and $\mathbb{E}[(\hat{Y}_2^\lambda(\infty) - n_2^\lambda + \Delta_2^\lambda)^+] - \mathbb{E}[(X^\lambda(\infty) - n_1^\lambda + \Delta_1^\lambda)^+] = \mathcal{O}(1)$, which proves (B13). □

Proof of Lemma 5.3. Fix two sequences $\{n_1^\lambda\}$ and $\{n_2^\lambda\}$ as in the statement of the lemma and let $\beta_i^\lambda = (n_i^\lambda\mu - \lambda)/\sqrt{\lambda}$ for $i = 1, 2$. Using (16) we write

$$g^\lambda(\beta_i^\lambda) := \theta\,\mathbb{E}[\tilde{Q}_{n_i^\lambda}^\lambda] = \sqrt{\lambda}\sqrt{\theta}z_1(\beta_i^\lambda)z_2(\beta_i^\lambda),$$

where $z_2(\beta_i^\lambda) = [h(\beta_i^\lambda/\sqrt{\theta}) - \beta_i^\lambda/\sqrt{\theta}]$ and $z_1(\beta_i^\lambda) = 1 - p(\beta_i^\lambda, \mu, \theta)$. Note that $z_1(\cdot) \in [0, 1]$ and is, in fact, a decreasing function of its argument (see the proof of Garnett et al. [15, Theorem 4]). Also, if $\limsup_\lambda \beta_2^\lambda \le c_1$, then there exists c_2 such that $\liminf_{\lambda\to\infty}(1 - p(\beta_2^\lambda, \mu^\lambda, \theta)) \ge c_2 > 0$; see again Garnett et al. [15, Theorem 4]. Also, the function $z_2(\cdot)$ is nonnegative and strictly decreasing in its argument (see the proof of Mandelbaum and Zeltyn [28, Theorem 4.1]).

Assume first that $n_1^\lambda - n_2^\lambda \to \infty$. Then

$$\begin{aligned} g^\lambda(\beta_1^\lambda) - g^\lambda(\beta_2^\lambda) &= \frac{\sqrt{\lambda}}{\sqrt{\theta}}z_2(\beta_1^\lambda)z_1(\beta_1^\lambda) - \frac{\sqrt{\lambda}}{\sqrt{\theta}}z_2(\beta_2^\lambda)z_1(\beta_2^\lambda) \\ &\le \frac{\sqrt{\lambda}}{\sqrt{\theta}}(z_2(\beta_1^\lambda) - z_2(\beta_2^\lambda))z_1(\beta_2^\lambda) \\ &\le c_2\frac{\sqrt{\lambda}}{\sqrt{\theta}}(z_2(\beta_1^\lambda) - z_2(\beta_2^\lambda)). \end{aligned}$$

We claim that

$$\frac{\sqrt{\lambda}}{\sqrt{\theta}}(z_2(\beta_1^\lambda) - z_2(\beta_2^\lambda)) \to -\infty. \tag{B14}$$

Recall that, by assumption, $\limsup_\lambda \beta_2^\lambda < \infty$ so that since $z_2(\cdot)$ is strictly decreasing and convex (see the online appendix in Mandelbaum and Zeltyn [28]) there exists c_3 such that $\liminf_{\lambda\to\infty} z_2^{(1)}(\beta_2^\lambda) \le -c_3$. There are two cases to consider:

(i) $\beta_1^\lambda - \beta_2^\lambda \to 0$. In this case, by a Taylor expansion around β_2^λ we have that

$$\begin{aligned} \frac{\sqrt{\lambda}}{\sqrt{\theta}}(z_2(\beta_1^\lambda) - z_2(\beta_2^\lambda)) &\le -c_3\frac{\sqrt{\lambda}}{\sqrt{\theta}}(\beta_1^\lambda - \beta_2^\lambda + o(\beta_1^\lambda - \beta_2^\lambda)) \\ &= -c_3\frac{1}{\sqrt{\theta}}(n_1^\lambda - n_2^\lambda + o(n_1^\lambda - n_2^\lambda)) \to -\infty, \end{aligned}$$

where the divergence follows because $n_1^\lambda - n_2^\lambda \to \infty$.

(ii) If $\liminf_{\lambda\to\infty}(\beta_1^\lambda - \beta_2^\lambda) \ge c_4$ for an absolute constant c_4, then the result follows trivially because $z_2(\cdot)$ is strictly decreasing.

The case $n_1^\lambda - n_2^\lambda \to -\infty$ is treated similarly. Since $\beta_1^\lambda \le \beta_2^\lambda$ for all sufficiently large λ, we have that

$$\begin{aligned} g^\lambda(\beta_1^\lambda) - g^\lambda(\beta_2^\lambda) &= \frac{\sqrt{\lambda}}{\sqrt{\theta}}z_2(\beta_1^\lambda)z_1(\beta_1^\lambda) - \frac{\sqrt{\lambda}}{\sqrt{\theta}}z_2(\beta_2^\lambda)z_1(\beta_2^\lambda) \\ &\ge \frac{\sqrt{\lambda}}{\sqrt{\theta}}(z_2(\beta_1^\lambda) - z_2(\beta_2^\lambda))z_1(\beta_2^\lambda) \\ &\ge c_2\frac{\sqrt{\lambda}}{\sqrt{\theta}}(z_2(\beta_1^\lambda) - z_2(\beta_2^\lambda)), \end{aligned}$$

for all such λ. Similarly to the above, it is now verified that $(\sqrt{\lambda}/\sqrt{\theta})(z_2(\beta_1^\lambda) - z_2(\beta_2^\lambda)) \to \infty$ as required. □

Appendix C. The NDS regime. Thus far we assumed that $\mu^\lambda \equiv \mu$, which covers, in particular, the QED, ED, and QD regimes. In this section we focus on the NDS regime; namely, we assume that $\mu^\lambda = \bar{\mu}\sqrt{\lambda}$ for some $\bar{\mu} > 0$ and

$$\sqrt{\lambda}(1 - \rho^\lambda) = \mathcal{O}(1).$$

It is, in turn, a property of the NDS regime that

$$\Delta^\lambda - n^\lambda = \mathcal{O}(1) \qquad \text{and} \qquad \lambda - \mu^\lambda n^\lambda = \mathcal{O}(\sqrt{\lambda}). \tag{C1}$$

The arguments in this section prove that our universal approximation in (14) is indeed universal in that it covers also this (somewhat newer) regime. From a practical viewpoint, the only change is that the service rate μ should be replaced with μ^λ wherever it appears, particularly in the definition of the universal diffusion in (14) and in Δ^λ, which is now given by

$$\Delta^\lambda = \frac{\lambda}{\mu^\lambda} - \left(\frac{\lambda}{\mu^\lambda} - n^\lambda\right)^+ \left(1 - \frac{\mu^\lambda}{\theta}\right).$$

With these obvious changes, all the results stated in §3 apply to the NDS regime without exception.

Many of our proofs do not at all depend on whether or not μ^λ scales with λ. Most of the remaining proofs require only minor changes. Rather than repeating the proofs, we carefully point out the required adjustments. We regenerate our numerical examples for this regime in §C.4.

C.1. Changes to §2. The single mathematical result here is Lemma 2.1, which is argued for fixed λ and, in particular, does not depend on how (and whether) μ^λ scales with λ.

C.2. Changes to §3.

- *Theorem* 1: Theorem 1 is a direct corollary of Theorem 2. The bound $(\mathbb{E}_1[\tau_l^\lambda])^{-1} = \mathcal{O}(\sqrt{\lambda})$ does not appear in the NDS version of Theorem 2 because the lower excursion is, in fact, shorter here; see Proposition 9 below. The remainder of Theorem 2 persists in this regime and the required changes to its proof are detailed in §C.3 below.
- *Corollaries* 1 *and* 2: Given Theorem 1 and Lemma A.1, the proofs of these corollaries can be repeated without any change.
- *Lemma* A.1: The regenerative-structural based argument requires no change.

C.3. Changes to §4. The remainder of this appendix is dedicated to adjustments to the proof of Theorem 2 as it appears in §4. For the NDS regime we use n^λ (rather than Δ^λ) as the "center" of our regenerative structure. We redefine

$$\tilde{X}^\lambda(t) = X^\lambda(t) - n^\lambda \qquad \text{and} \qquad \tilde{Y}^\lambda(t) = Y^\lambda(t) - n^\lambda.$$

Given a sequence $\{f^\lambda\} \in \mathcal{S}_m$, define for each λ and x, $g^\lambda(x) = f^\lambda(x + n^\lambda - \Delta^\lambda)$. By (C1) it then holds that $\{g^\lambda\} \in \mathcal{S}_m$ so that the bound gaps for the redefined $\tilde{X}^\lambda$ and $\tilde{Y}^\lambda$ imply directly the bounds for $X^\lambda - \Delta^\lambda$ and $Y^\lambda - \Delta^\lambda$. The regenerative process is redefined for $\tilde{X}^\lambda$, $\tilde{Y}^\lambda$ in an obvious way together with the hitting times τ_u^λ, τ_l^λ and τ^λ for $\tilde{X}^\lambda$ and $\tilde{\tau}_u^\lambda$, $\tilde{\tau}_l^\lambda$ and $\tilde{\tau}^\lambda$ for $\tilde{Y}^\lambda$.

We note that though now $\tilde{X}^\lambda$ and $\tilde{Y}^\lambda$ are defined slightly differently, the proof of Lemma 4.1 will not change because we can use the same method to choose K there. We next consider separately each of the upper and lower excursions.

C.3.1. The upper excursion. With the redefined centering we replace (24) with

$$\ell^\lambda(x) = -\lambda + \mu^\lambda((x + n^\lambda) \wedge n^\lambda) + \theta x, \tag{C2}$$

for all $x \geq 0$. Using (C1) we then have that

$$\ell^\lambda(x) \leq \vartheta\sqrt{\lambda} + \theta x \qquad \text{and} \qquad \ell^\lambda(x) \geq \theta x - \vartheta\sqrt{\lambda}, \tag{C3}$$

for all $x \geq 0$ and an absolute constant ϑ. Let, as before, U^λ be a B&D process on $\{0, 1\ldots,\}$ with birth rate λ and death rate $n^\lambda\mu^\lambda + \theta x$ for all $x > 0$. Importantly, (C3) guarantees that (26) holds. Since having ℓ^λ nondecreasing and satisfying (26) are all that is required for Lemmas 4.4 and 4.7–4.8, these continue to hold without any changes to their respective proofs. Some adjustment is required, however, in the proof of Lemma 4.6.

Lemma 4.6: First we note that the proofs of the upper bounds in Lemma 4.6 (specifically, in (30) and (31)) only use (26). These bounds and their respective proof then require no changes.

We next argue the lower bound in (30), namely, that $\mathbb{E}[\ell^\lambda(U^\lambda(\infty))] \geq k\sqrt{\lambda}$ for some absolute constant k. The proof of this fact in Lemma 4.6 relies on the nonnegativity of ℓ^λ for $x \geq 0$. This nonnegativity does not necessarily hold for the redefined ℓ^λ in (C2). Fortunately, however, the result does follow from our existing results, as we outline next.

If $\Delta^\lambda \leq n^\lambda$, then ℓ^λ is indeed nonnegative for $x \geq 0$ and the proof requires no changes. Consequently, we must only consider the case $\Delta^\lambda > n^\lambda$. We rewrite (C2) as

$$\ell^\lambda(x) = -\kappa^\lambda + \theta x,$$

where $\kappa^\lambda = \lambda - \mu^\lambda n^\lambda \geq 0$. Let $k^\lambda = \lceil \kappa^\lambda/\theta \rceil$. Define $\hat{U}^\lambda$ to be a B&D process on the state space $\{0, 1, \dots\}$ with arrival rate λ and death rate $\lambda + \theta x$ in state x. All the conditions in the setting of §4.2 apply to $\hat{U}^\lambda$ so that by Lemma 4.6 we have that

$$\mathbb{E}[(\hat{U}^\lambda(\infty))^2] \leq \vartheta\lambda. \tag{C4}$$

Note that

$$\mathbb{E}[U^\lambda(\infty)^m] \leq c_1((k^\lambda)^2 + \mathbb{E}[(U^\lambda(\infty) - k^\lambda)^2 \mid U^\lambda(\infty) > k^\lambda]) \tag{C5}$$

for an absolute constant c and that by basic properties of B&D processes, conditional on being greater than k^λ, $U^\lambda(\infty) - k^\lambda$ has the law of $\hat{U}^\lambda$. By (C1) $k^\lambda = \mathcal{O}(\sqrt{\lambda})$ so that (C4) and (C5) imply that the family $\{U^\lambda(\infty)/\sqrt{\lambda}, \lambda \geq 0\}$ is uniformly integrable.

The process U^λ has the law of the headcount process in the $M/M/1+M$ queue with arrival rate λ, service rate $\mu^\lambda n^\lambda$ and patience rate θ. Assuming that $k^\lambda/\sqrt{\lambda} \to k \in \mathbb{R}$, it would then follow from Ward [38, Theorem 2.1] that

$$\frac{U^\lambda(\infty)}{\sqrt{\lambda}} \Longrightarrow U(\infty), \tag{C6}$$

where $U(\infty)$ has the density f_{ROU} in Ward [38, p. 6] (with θ replacing γ there and k replacing β). Since the density f_{ROU} is that of a truncated normal random variable with mean k and standard deviation 1, there exists $\vartheta > 0$ such that $\mathbb{E}[U(\infty) - k] = \vartheta$. By (C6) and the uniform integrability of $\{U^\lambda(\infty)/\sqrt{\lambda}, \lambda \geq 0\}$, we would conclude that $\mathbb{E}[(U^\lambda(\infty) - k^\lambda)] \geq \vartheta\sqrt{\lambda}/2$ for all sufficiently large λ. We could consequently have that $\mathbb{E}[\ell^\lambda(U^\lambda(\infty))] \geq \theta\vartheta\sqrt{\lambda}/2$.

It only remains to consider the case that $\{k^\lambda/\sqrt{\lambda}, \lambda \geq 0\}$ does not converge. In this case, however, since $k^\lambda = \mathcal{O}(\sqrt{\lambda})$ (see (C1)), we can find $\bar{k}^\lambda$ such that $k^\lambda \leq \bar{k}^\lambda$ for all sufficiently large λ and such that $\bar{k}^\lambda/\sqrt{\lambda} \to k \in (-\infty, \infty)$. Defining $\bar{U}^\lambda$ to be a B&D process with birth rate λ and death rate $\lambda - \bar{k}^\lambda\theta + \theta x$, a simple coupling argument shows that $U^\lambda(\infty) - k^\lambda \geq_{st} \bar{U}^\lambda(\infty) - \bar{k}^\lambda$. The arguments above apply directly to $\bar{U}^\lambda(\infty) - \bar{k}^\lambda$ so that the result follows.

C.3.2. The lower excursion. The lower excursion requires more elaborate adjustments. In the case $\mu^\lambda \equiv \mu$ treated in the main body of the paper, we relied on a certain symmetry between the upper and lower excursions; see §4.3. This symmetry is lost in the NDS regime. Informally, the lower excursion here is order-of-magnitude shorter than the upper excursion. Additionally, because of the way in which μ^λ scales with λ, the derivative bounds that we are able to establish for the ODE solution are somewhat weaker. Together, however, the correct ultimate bounds are achieved. Propositions 9 and 10 are the required analogues of Proposition 5 and 6 for the NDS regime. Because of the symmetry between the lower and upper excursion in the case $\mu^\lambda \equiv \mu$, the proof of Proposition 5 (respectively, 6) was identical to that of Proposition 3 (respectively, 4). Thus, we use the latter as our reference proofs.

Proposition 9 (Order Bounds).

$$\mathbb{E}_0[\tau_l^\lambda] = \mathcal{O}(\sqrt{\lambda}^{-3/2}) \qquad \textit{and} \qquad \mathbb{E}_0\left[\int_0^{\tau_l^\lambda} (\check{X}^\lambda(s))^m \, ds\right] = \mathcal{O}(\sqrt{\lambda}^{(m-3)/2}), \quad m \in \mathbb{N}.$$

Proof. The main step in adjusting the proof of Proposition 3 is in adapting the Lyapunov-function-argument in Lemma 4.6. In fact, the proof is almost identical, with the exception of the power of λ in various places being $1/4$ rather than $1/2$. We provide the details for completeness.

Let U^λ be a B&D process on the nonnegative integers with birth rate $\lambda - \check{\ell}^\lambda(x)$ and death rate λ, where we redefine $\check{\ell}^\lambda$ (see (46)) as

$$\check{\ell}^\lambda(x) = \lambda - n^\lambda\mu^\lambda + x\mu^\lambda. \tag{C7}$$

Let $\mathcal{U}^\lambda$ be the generator of the B&D process $(U^\lambda(t), t \geq 0)$. Let $f(x) = x^m$. Then, for all $x \in \mathbb{N}$,

$$\mathcal{U}^\lambda f(x) = \lambda(f(x-1) - f(x)) + (\lambda - \check{\ell}^\lambda(x))(f(x+1) - f(x)).$$

Since $f(x+1) - f(x) = \sum_{k \leq m,\, k \neq 0} \binom{m}{k} x^{m-k}$ and $f(x-1) - f(x) = \sum_{k \leq m,\, k \neq 0} \binom{m}{k} x^{m-k}(-1)^k$,

$$\mathcal{U}^\lambda f(x) = \sum_{k \leq m,\, k \neq 0} -\check{\ell}^\lambda(x) \binom{m}{k} x^{m-k} + \sum_{k \text{ even},\, k \neq 0} (2\lambda) \binom{m}{k} x^{m-k}.$$

Using (C1) we have that

$$\check{\ell}^\lambda(x) \geq \bar{\mu}\sqrt{\lambda}x - \vartheta\sqrt{\lambda} \tag{C8}$$

for some absolute constant ϑ and all $x \geq 0$ and we can then choose absolute constants c_1, c_2, and c_3 such that $\mathcal{U}^\lambda f(x) \leq -c_2\sqrt{\lambda}x^m$ for all $x \geq c_1\lambda^{1/4}$ and such that $\mathcal{U}^\lambda f(x) \leq c_3\lambda^{(1/4)m+1/2}$ for all $x \leq c_1\lambda^{1/4}$. Subsequently, there exist absolute constants c_4, c_5 such that

$$\mathcal{U}^\lambda f(x) \leq -c_4\sqrt{\lambda}x^m + c_5 x^{(1/4)m+1/2}.$$

Applying expectations we conclude that (see, e.g., Glynn and Zeevi [16, Corollary 1]),

$$\mathbb{E}[(U^\lambda(\infty))^m] \leq \frac{c_5}{c_4}\lambda^{(1/4)m}.$$

A Lyapunov function argument is also used to prove that $\mathbb{E}[\check{\ell}^\lambda(U^\lambda(\infty))] \geq \vartheta\lambda^{3/4}$ for an absolute constant ϑ. We consider in detail the case in which $\check{\ell}^\lambda$ is nonnegative (i.e., $\Delta^\lambda \geq n^\lambda$). Since $\Delta^\lambda - n^\lambda = \mathcal{O}(1)$ (see (C1)), the other case follows immediately by redefining the 0 point to be Δ^λ instead of n^λ.

Define (a sequence of) functions g^λ as follows (c_6 is to be determined):

$$g^\lambda(x) = \begin{cases} x^2; & x \leq \lceil c_6\lambda^{1/4} \rceil; \\ (2\lceil c_6\lambda^{1/4} \rceil + 1)x - (\lceil c_6\lambda^{1/4} \rceil + 1)\lceil c_6\lambda^{1/4} \rceil; & x > \lceil c_6\lambda^{1/4} \rceil. \end{cases}$$

Then

$$\mathcal{U}^\lambda g^\lambda(x) = \begin{cases} \lambda; & x = 0, \\ 2\lambda - (2x+1)\check{\ell}^\lambda(x); & 1 \leq x \leq \lceil c_6\lambda^{1/4} \rceil, \\ -(2\lceil c_6\lambda^{1/4} \rceil + 1)\check{\ell}^\lambda(x); & x > \lceil c_6\lambda^{1/4} \rceil. \end{cases}$$

We claim that we can choose c_6 in the definition of g^λ together with absolute constants c_7, c_8 such that

$$\mathcal{U}^\lambda g^\lambda(x) \geq c_7\lambda - c_8\lambda^{1/4}\check{\ell}^\lambda(x), \tag{C9}$$

in which case, since $\mathbb{E}[\mathcal{U}^\lambda g^\lambda(U^\lambda(\infty))] = 0$, we conclude that $\mathbb{E}[\check{\ell}^\lambda(U^\lambda(\infty))] \geq c_7\lambda^{3/4}/c_8$ as required. Since $\check{\ell}^\lambda$ is nonnegative and nondecreasing, to show the existence of such constants, it suffices to find $c_7 \leq 1$ and c_6, c_7, c_8 such that

$$\lambda - (2\lceil c_6\lambda^{1/4} \rceil + 1)\check{\ell}^\lambda(\lceil c_6\lambda^{1/4} \rceil) \geq 0, \tag{C10}$$

$$[c_8\lambda^{1/4} - (2\lceil c_6\lambda^{1/4} \rceil + 1)]\check{\ell}^\lambda(\lceil c_6\lambda^{1/4} \rceil) \geq c_7\lambda. \tag{C11}$$

Specifically, we choose c_6 sufficiently small so that (C10) holds. Since there exists an absolute constant c_9 such that $\check{\ell}^\lambda(\lceil c_6\lambda^{1/4} \rceil) \geq c_9\lambda^{3/4}$ we can subsequently choose c_8 to satisfy (C11). Equation (C9) follows and we conclude that $\mathbb{E}[\ell^\lambda(U^\lambda(\infty))] \geq c_7\lambda^{3/4}/c_8$ as required.

With these bounds, the proof of Proposition 9 follows exactly as that of Proposition 3 because for any nondecreasing function f,

$$\mathbb{E}[f(U^\lambda(\infty))] \leq \frac{\mathbb{E}_0[\int_0^{\tau_l^\lambda} f(\check{X}^\lambda(s))\, ds]}{\mathbb{E}_0[\tau_l^\lambda]} \leq \mathbb{E}[f(U^\lambda(\infty) + 1)],$$

which is proved identically to Lemma 4.4, and the inequalities

$$\frac{1}{\mathbb{E}[\check{\ell}^\lambda(U^\lambda(\infty) + 1)]} \leq \mathbb{E}_0[\tau_l^\lambda] \leq \frac{1}{\mathbb{E}[\check{\ell}^\lambda(U^\lambda(\infty))]},$$

$$\mathbb{E}_0\left[\int_0^{\tau_l^\lambda} f(\check{X}^\lambda(s))\, ds\right] \leq \frac{\mathbb{E}[f(U^\lambda(\infty) + 1)]}{\mathbb{E}[\check{\ell}^\lambda(U^\lambda(\infty))]},$$

which are, in turn, proved identically to Proposition 3. □

Proposition 10 (Gap Bounds). *Fix $m \in \mathbb{N}$ and $\{f^\lambda\} \in \mathcal{S}_m$,*

$$V_l^\lambda(f^\lambda, 0) - \mathcal{V}_l^\lambda(f^\lambda, 0) = \mathcal{O}(\sqrt{\lambda}^{m-2}).$$

The main difference here, relative to the proof of Proposition 4, is in the derivative bounds stated in Lemma C.1. These should be contrasted with those in Lemma 4.7. Given the derivative bounds and the order bounds in Proposition 9, the proof of Proposition 10 is similar to that of Proposition 4; we provide the complete details below. In what follows $\check{\ell}^\lambda$ is as in (C7). Also, when referring to (47), note that one should replace $\check{\ell}^\lambda$ there with the one defined in (C7).

Lemma C.1. *Fix $m \in \mathbb{N}$ and $\{f^\lambda\} \in \mathcal{S}_m$. Then for each λ, there exists an infinitely differentiable solution $\check{u}^\lambda_{\check{f}^\lambda}$ to (47) such that for $x \geq 0$,*

$$|(\check{u}^\lambda_{\check{f}^\lambda})^{(1)}(x)| \leq A_{1,m}\big(\sqrt{\lambda}^{(2m-3)/2} + \lambda^{(m-3)/4}x^m\big),$$
$$|(\check{u}^\lambda_{\check{f}^\lambda})^{(2)}(x)| \leq A_{2,m}\big(\sqrt{\lambda}^{m-2} + \lambda^{(m-5)/4}x^{m+1}\big),$$
$$|(\check{u}^\lambda_{\check{f}^\lambda})^{(3)}(x)| \leq A_{3,m}\big(\sqrt{\lambda}^{(2m-5)/2} + \lambda^{(m-7)/4}x^{m+2} + (1/\lambda)|(\check{f}^\lambda)^{(1)}(x)|\big),$$

where $A_{i,m}$, $i = 1, 2, 3$ are absolute constants. In addition, $\mathcal{V}_l^\lambda(y) =: \mathcal{V}_l^\lambda(f^\lambda, y)$ is the unique solution satisfying the above inequalities.

Proof. Using (C8) and recalling that $\mu^\lambda = \bar{\mu}\sqrt{\lambda}$, we have the existence of absolute constants $C_{0,m}, C_{1,m}$ such that

$$\begin{aligned}
|(\check{u}^\lambda_{\check{f}^\lambda})^{(1)}(x)| &= \left|\int_x^\infty \frac{a_1\sqrt{\lambda}^m + a_2 s^m}{\lambda} e^{-\int_x^s((\check{\ell}^\lambda(u))/\lambda)\,du}\,ds\right| \\
&\leq \int_x^\infty \frac{a_1\sqrt{\lambda}^m + a_2 s^m}{\lambda} e^{-\int_x^s((\bar{\mu}u - c_0)/\sqrt{\lambda})\,du}\,ds \\
&= a_1\sqrt{\lambda}^{m-2}\lambda^{1/4}e^{\bar{\mu}(x-c_0/\bar{\mu})^2/(2\sqrt{\lambda})}\int_{x-c_0/\bar{\mu}}^\infty e^{-(\bar{\mu}s^2/(2\sqrt{\lambda}))}\,d\frac{1}{\lambda^{1/4}}s \\
&\quad + \frac{a_2}{\lambda}\lambda^{(m+1)/4}e^{\bar{\mu}(x-c_0/\bar{\mu})^2/(2\sqrt{\lambda})}\int_{x-c_0/\bar{\mu}}^\infty (s + c_0/\bar{\mu})^m e^{-((\bar{\mu}s^2)/(2\sqrt{\lambda}))}\,d\frac{1}{\lambda^{1/4}}s \\
&\leq C_{0,m}\sqrt{\lambda}^{m-2}\lambda^{1/4}\left(1 + \frac{x}{\lambda^{1/4}}\right) + \frac{C_{1,m}}{\lambda}\lambda^{(m+1)/4}\left(1 + \left(\frac{x}{\lambda^{1/4}}\right)^m\right).
\end{aligned}$$

If $x \leq \lambda^{1/4}$, then the above is less than $(2C_{0,m} + 2C_{1,m})\lambda^{(2m-3)/4}$, whereas if $x \geq \lambda^{1/4}$, then the above is less than $2(C_{0,m} + C_{1,m})\lambda^{(m-3)/4}x^m$. As a result, we can define absolute constant $A_{1,m}$ appropriately for $|(\check{u}^\lambda_{\check{f}^\lambda})^{(1)}(x)|$. Using directly the ODE (47), we then obtain the bounds for $|(\check{u}^\lambda_{\check{f}^\lambda})^{(2)}(x)|$ and $|(\check{u}^\lambda_{\check{f}^\lambda})^{(3)}(x)|$ similarly to the proof of Lemma 4.7. Finally, the fact that $\mathcal{V}_l^\lambda(y)$ is the unique solution is proved identically to Lemma 4.7, noting that its proof is for fixed λ and, in particular, does not depend on whether or not μ^λ scales with λ. □

Proof of Proposition 10. The proof begins identically to that of Proposition 4 to obtain that

$$\begin{aligned}
&\left|\mathcal{V}_l^\lambda(y) - \mathbb{E}_y\left[\int_0^{t\wedge\tau_l^\lambda} \check{f}^\lambda(\check{X}^\lambda(s))\,ds\right]\right| \\
&\quad\leq \left|\mathbb{E}_y[\mathcal{V}_l^\lambda(\check{X}^\lambda(t\wedge\tau_l^\lambda))]\right| \\
&\qquad + \left|\mathbb{E}_y\left[\sum_{s\leq t\wedge\tau_l^\lambda:\,|\Delta\check{X}^\lambda(s)|>0} \big(\Delta\mathcal{V}_l^\lambda(\check{X}^\lambda(s)) - (\mathcal{V}_l^\lambda)^{(1)}(\check{X}^\lambda(s-))\Delta\check{X}^\lambda(s) - \tfrac{1}{2}(\mathcal{V}_l^\lambda)^{(2)}(\check{X}^\lambda(s-))(\Delta\check{X}^\lambda(s))^2\big)\right]\right| \\
&\qquad + \left|\mathbb{E}_y\left[\int_0^{t\wedge\tau_l^\lambda} \tfrac{1}{2}\check{\ell}^\lambda(\check{X}^\lambda(s))(\mathcal{V}_l^\lambda)^{(2)}(\check{X}^\lambda(s))\,ds\right]\right| \\
&\quad\leq \left|\mathbb{E}_y[\mathcal{V}_l^\lambda(\check{X}^\lambda(t\wedge\tau_l^\lambda))]\right| + \left|\mathbb{E}_y\left[\sum_{s\leq t\wedge\tau_l^\lambda:\,|\Delta\check{X}^\lambda(s)|>0} \tfrac{1}{2}\big|(\mathcal{V}_l^\lambda)^{(3)}(\check{X}^\lambda(s-) + \eta^{(1)}_{\check{X}^\lambda(s-)})\big|\right]\right| \\
&\qquad + \left|\mathbb{E}_y\left[\int_0^{t\wedge\tau_l^\lambda} \tfrac{1}{2}\check{\ell}^\lambda(\check{X}^\lambda(s))(\mathcal{V}_l^\lambda)^{(2)}(\check{X}^\lambda(s))\,ds\right]\right| \qquad \text{(C12)}
\end{aligned}$$

for some $\eta^{(1)}_{\check{X}^\lambda(s-)} \in (-1, 1)$.

Using Proposition 9, we have that for all $i \geq 0$,

$$\mathbb{E}_y\left[\int_0^{\tau_l^\lambda} |\check{\ell}^\lambda(\check{X}^\lambda(s))(\check{X}^\lambda(s))^i|\, ds\right] = \mathcal{O}(\sqrt{\lambda}^{i/2}). \tag{C13}$$

This, together with the bound for $(\mathcal{V}_l^\lambda)^{(2)}$ in Lemma C.1, gives

$$\mathbb{E}_y\left[\int_0^{\tau_l^\lambda} |\check{\ell}^\lambda(\check{X}^\lambda(s))(\mathcal{V}_l^\lambda)^{(2)}(\check{X}^\lambda(s))|\, ds\right] = \mathcal{O}(\sqrt{\lambda}^{m-2}).$$

Similarly, we can get the same order for the term involving $(\mathcal{V}_l^\lambda)^{(3)}$ in (C12) if we can prove

$$\left|\mathbb{E}_y\left[\sum_{s\leq t\wedge\tau_l^\lambda:\,|\Delta\check{X}^\lambda(s)|>0} |(\check{f}^\lambda)^{(1)}(\check{X}^\lambda(s-)+\eta^{(1)}_{\check{X}^\lambda(s-)})|\right]\right| = \mathcal{O}(\sqrt{\lambda}^{m}). \tag{C14}$$

Here, if $m \geq 1$, we can use (C13) to prove (C14). If $m = 0$, we have as in the proof for the upper excursion that

$$\mathbb{E}_y\left[\sum_{s\leq t\wedge\tau_l^\lambda:\,|\Delta\check{X}^\lambda(s)|>0} |(\check{f}^\lambda)^{(1)}(\check{X}^\lambda(s-)+\eta^\lambda_{\check{X}^\lambda(s-)})|\right] \leq a_3\,\mathbb{E}_y\left[\sum_{s\leq t\wedge\tau_l^\lambda:\,|\Delta\check{X}^\lambda(s)|>0} \mathbb{I}_{\{\check{X}^\lambda(s-)\in(a^\lambda-1,a^\lambda+2)\}}\right],$$

which is (for $y = 0$), by Proposition 5 and Lemma A.1, of the order of $\lambda \times \mathbb{E}_0[\tau_l^\lambda]\,\mathbb{P}\{\check{X}^\lambda(\infty) \in (a^\lambda - 1, a^\lambda + 2)\} = \mathcal{O}(1)$. Recall (Lemma 4.7) that $V_l^\lambda(f^\lambda, y) = \mathbb{E}_y[\int_0^{\tau_l^\lambda} \check{f}^\lambda(\check{X}^\lambda(s))\, ds]$. Thus, we conclude that

$$V_l^\lambda(f^\lambda, 0) - \mathcal{V}_l^\lambda(0) = \mathcal{O}(\sqrt{\lambda}^{m-2}),$$

as required. □

C.4. Numerical examples. In Figures C.1–C.3 we regenerate Example 3.1 where the single difference is that we replace $\mu = 1$ in the examples with a service rate that scales with λ, namely, with $\mu^\lambda = \sqrt{\lambda}$. In Figures C.4–C.6 we similarly regenerate Example 3.2.

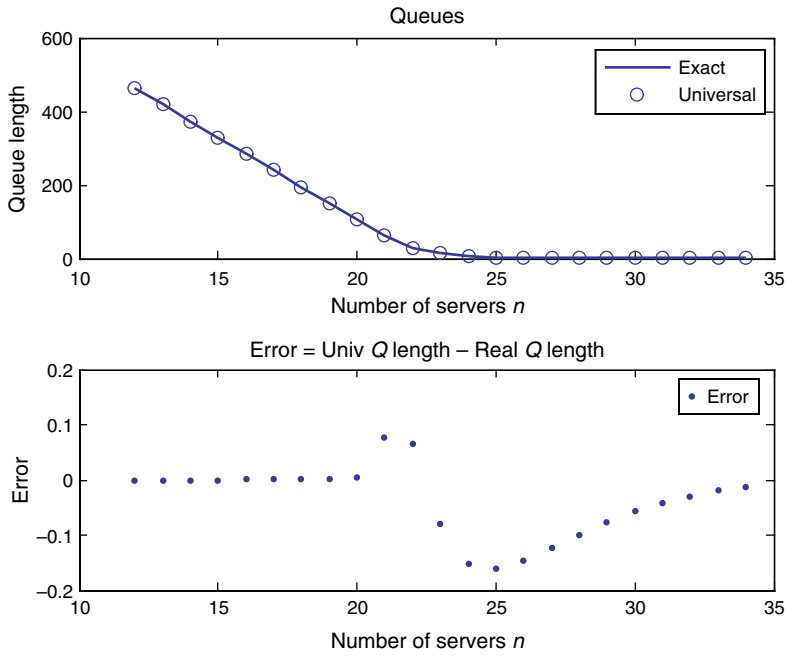

FIGURE C.1. Expected queue approximation: fixed λ, varying n ($11 \leq n \leq 34$).

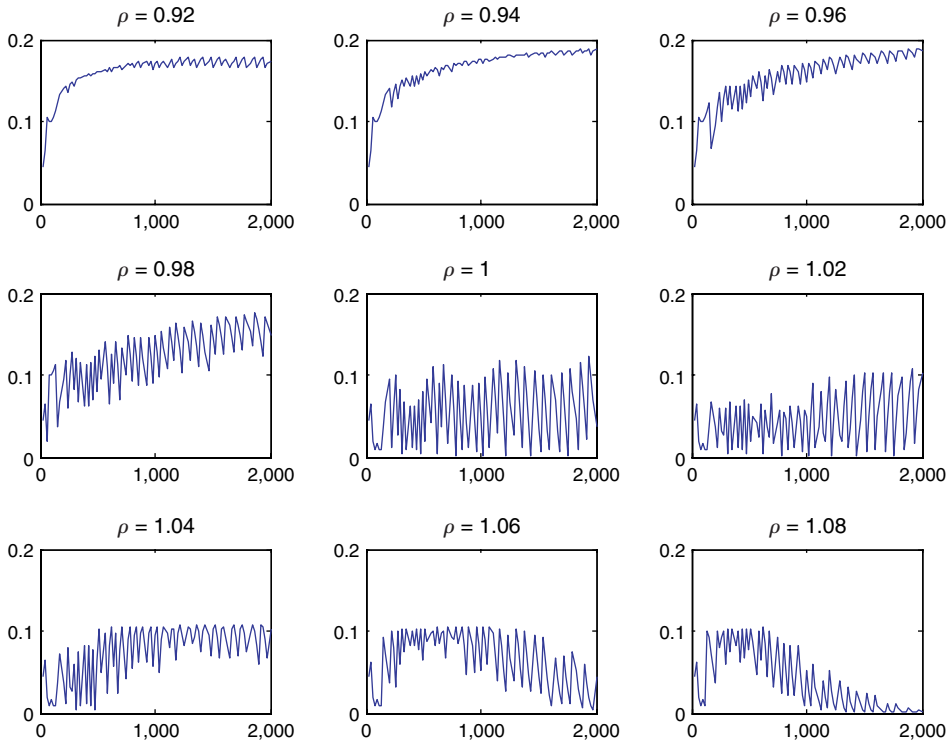

FIGURE C.2. Expected queue approximation: fixed ρ, varying λ ($20 \le \lambda \le 2{,}000$).

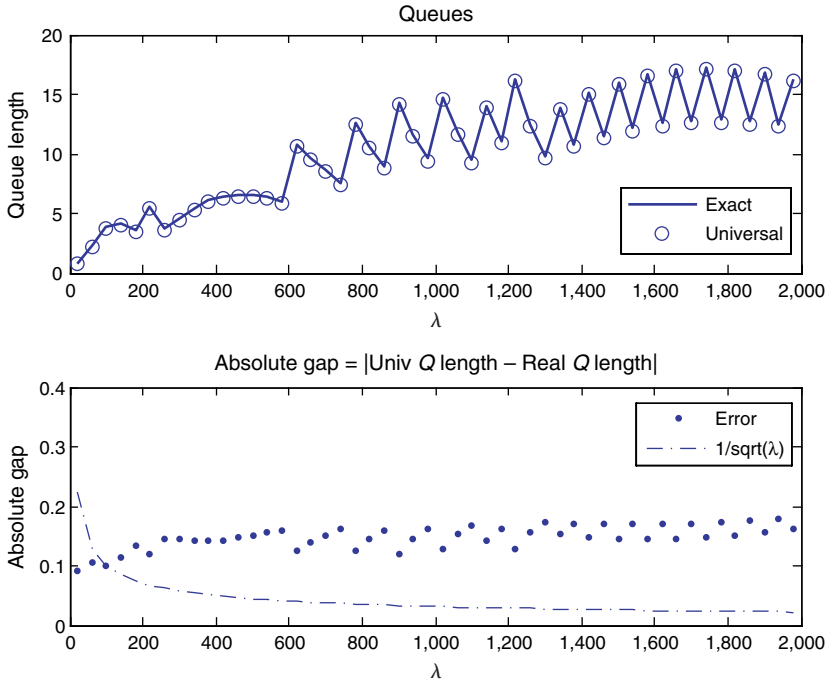

FIGURE C.3. Expected queue approximation: varying ρ^{λ} with λ ($20 \le \lambda \le 2{,}000$).

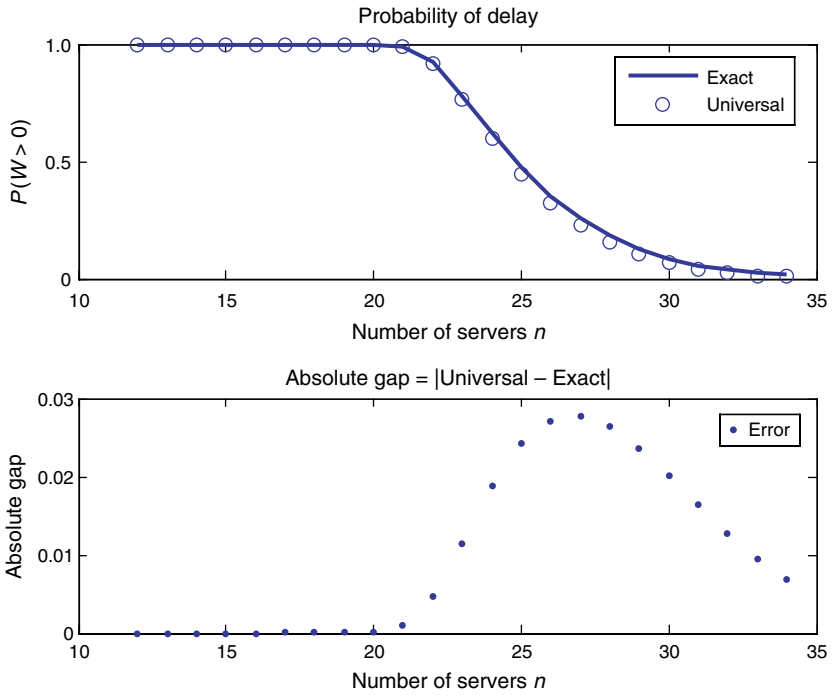

FIGURE C.4. Probability of delay: fixed λ, varying n ($11 \le n \le 34$).

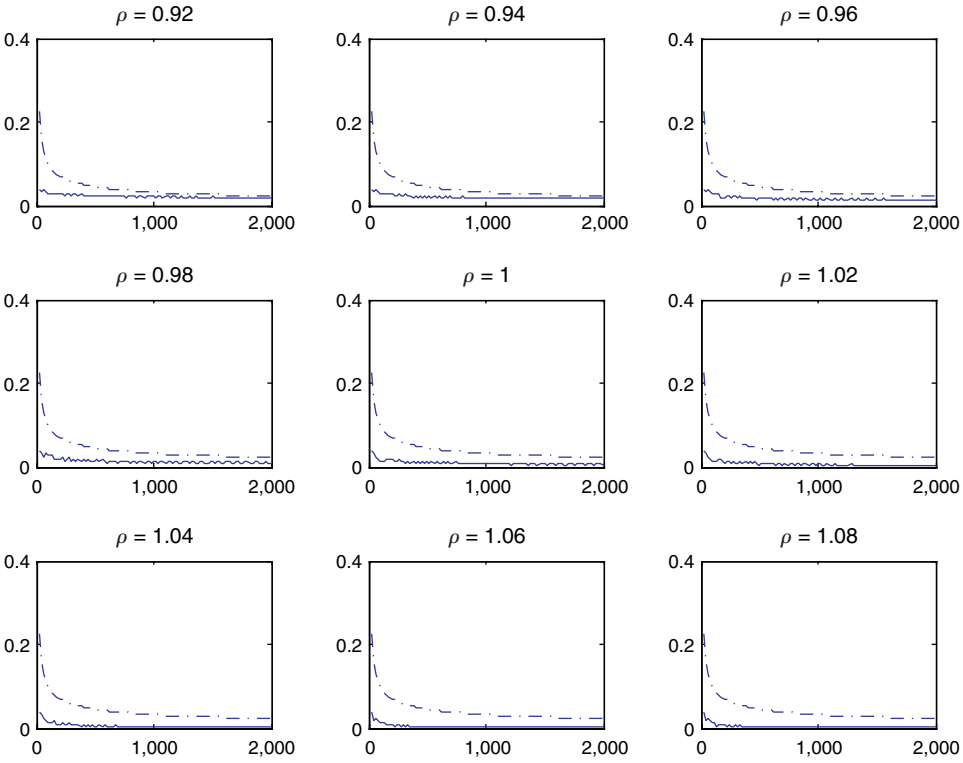

FIGURE C.5. Probability of delay: fixed ρ, varying λ ($20 \le \lambda \le 2{,}000$).

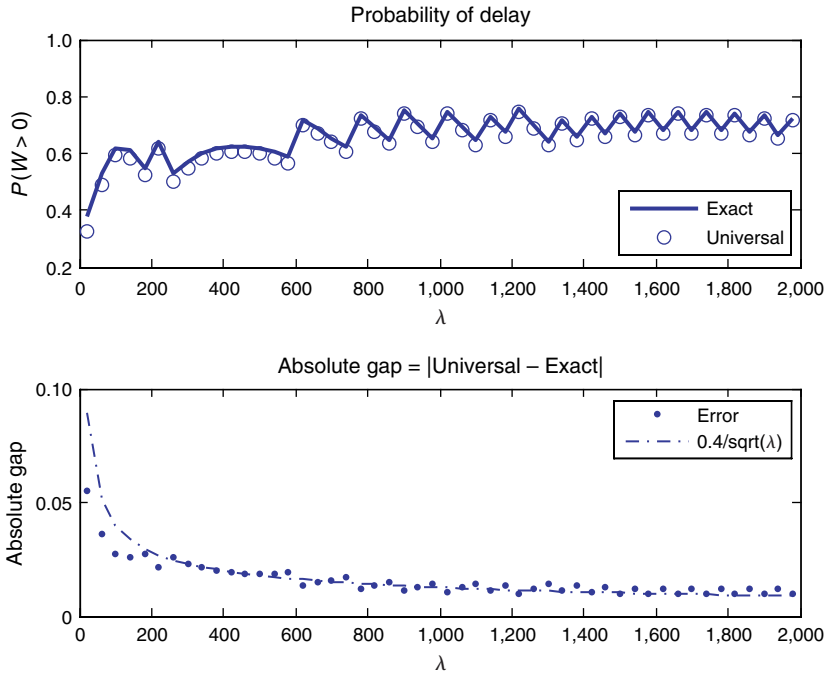

FIGURE C.6. Probability of delay: varying ρ^λ with λ $(20 \le \lambda \le 2{,}000)$.

References

[1] 4 call centers. Accessed March 2013, http://ie.technion.ac.il/serveng.
[2] Allon G, Deo S, Lin W (2013) The impact of size and occupancy of hospital on the extent of ambulance diversion: Theory and evidence. *Oper. Res.* Forthcoming.
[3] Asmussen S (2003) *Applied Probability and Queues* (Springer-Verlag, New York).
[4] Ata B, Gurvich I (2012) On optimality gaps in the Halfin-Whitt regime. *Ann. Appl. Probab.* 22(1):407–455.
[5] Atar R (2012) A diffusion regime with nondegenerate slowdown. *Oper. Res.* 60(2):490–500.
[6] Bassamboo A, Randhawa RS (2010) On the accuracy of fluid models for capacity sizing in queueing systems with impatient customers. *Oper. Res.* 58(5):1398–1413.
[7] Bassamboo A, Randhawa RS, Zeevi A (2010) Capacity sizing under parameter uncertainty: Safety staffing principles revisited. *Management Sci.* 56(10):1668–1686.
[8] Borst S, Mandelbaum A, Reiman M (2004) Dimensioning large call centers. *Oper. Res.* 52(1):17–34.
[9] Browne S, Whitt W (1995) Piecewise-linear diffusion processes. Dshalalow JH, ed. *Advances in Queueing: Theory, Methods, and Open Problems* (CRC Press, Boca Raton, FL), 463–480.
[10] Chen H (1996) Rate of convergence of the fluid approximation for generalized Jackson networks. *J. Appl. Probab.* 33(3):804–814.
[11] Chen H, Shen X (2000) Strong approximations for multiclass feedforward queueing networks. *Ann. Appl. Probab.* 10(3):828–876.
[12] Chen H, Yao DD (2001) *Fundamentals of Queueing Networks* (Springer-Verlag, New York).
[13] Feldman Z, Mandelbaum A, Massey WA, Whitt W (2008) Staffing of time-varying queues to achieve time-stable performance. *Management Sci.* 54(2):324–338.
[14] Gamarnik D, Zeevi A (2006) Validity of heavy traffic steady-state approximations in generalized Jackson networks. *Ann. Appl. Probab.* 16(1):56–90.
[15] Garnett O, Mandelbaum A, Reiman M (2002) Designing a call center with impatient customers. *Manufacturing Service Oper. Management* 4(3):208–227.
[16] Glynn PW, Zeevi A (2008) Bounding stationary expectations of Markov processes. *Markov Processes and Related Topics: A Festschrift for Thomas G. Kurtz. Selected papers of the Conference, Madison, WI, July*, Vol. 4, 195–214.
[17] Halfin S, Whitt W (1981) Heavy-traffic limits for queues with many exponential servers. *Oper. Res.* 29(3):567–588.
[18] Iglehart DL (1965) Limiting diffusion approximations for the many server queue and the repairman problem. *J. Appl. Probab.* 2(2):429–441.
[19] Karatzas I, Shreve S (1991) *Brownian Motion and Stochastic Calculus*, 2nd ed. (Springer-Verlag, New York).
[20] Kaspi H, Mandelbaum A (1992) Regenerative closed queueing networks. *Stochastics* 39(4):239–258.
[21] Kaspi H, Ramanan K (2013) SPDE limits of many-server queues. *Ann. Appl. Probab.* 23(1):145–229.
[22] Klebaner FC (2005) *Introduction to Stochastic Calculus with Applications* (Imperial College Press, London).
[23] Konstantopoulos T, Last G (1999) On the use of Lyapunov function methods in renewal theory. *Stochastic Processes Their Appl.* 79(1):165–178.
[24] Konstantopoulos T, Papadakis SN, Walrand J (1994) Functional approximation theorems for controlled renewal processes. *J. Appl. Probab.* 31:765–776.
[25] Liu Y, Whitt W (2012) Stabilizing customer abandonment in many-server queues with time-varying arrivals. *Oper. Res.* 60(6): 1551–1564.

[26] Loukianova D, Loukianov O, Song S (2009) Poincaré inequality and exponential integrability of hitting times for linear diffusions. ArXiv preprint ArXiv:0907.0762.
[27] Mandelbaum A, Zeltyn S (2007) Service engineering in action: The Palm/Erlang-A queue, with applications to call centers. Spath D, Fähnrich K-P, eds. *Advances in Services Innovations* (Springer-Verlag, Berlin, Heidelberg), 17–48.
[28] Mandelbaum A, Zeltyn S (2009) Staffing many-server queues with impatient customers: Constraint satisfaction in call centers. *Oper. Res.* 57(5):1189–1205.
[29] Mandelbaum A, Massey W, Reiman M (1998) Strong approximations for Markovian service networks. *Queueing Systems* 30(1–2):149–201.
[30] Meyn SP, Tweedie RI (2009) *Markov Chains and Stochastic Stability*, 2nd ed. (Springer-Verlag, New York).
[31] Pang G, Talreja R, Whitt W (2007) Martingale proofs of many-server heavy-traffic limits for Markovian queues. *Probab. Surveys* 4:193–267.
[32] Randhawa RS (2012) The optimality gap of asymptotically-derived prescriptions with applications to queueing systems. Working paper, University of Southern California, Los Angeles.
[33] Roberts GO, Rosenthal JS (1996) Quantitative bounds for convergence rates of continuous time Markov processes. *Electronic J. Probab.* 1:1–21.
[34] SEELab, Technion, http://ie.technion.ac.il/Labs/Serveng/.
[35] Talreja R, Whitt W (2009) Heavy-traffic limits for waiting times in many-server queues with abandonment. *Ann. Appl. Probab.* 19(6):2137–2175.
[36] Teschl G (2004) Ordinary differential equations and dynamical systems. Lecture notes, http://www.mat.univie.ac.at/gerald/ftp/book-ode/index.html.
[37] Van der Vaart AW (2006) Martingales, diffusions and financial mathematics. Lecture notes, http://www.math.vu.nl/sto/onderwijs/mdfm.
[38] Ward AR (2011) Asymptotic analysis of queueing systems with reneging: A survey of results for FIFO, single class models. *Surveys Oper. Res. Management Sci.* 16(1):1–14.
[39] Whitt W (1982) On the heavy-traffic limit theorem for $GI/G/\infty$ queues. *Adv. Appl. Probab.* 14(1):171–190.
[40] Whitt W (2004) Efficiency-driven heavy-traffic approximations for many-server queues with abandonments. *Management Sci.* 50(10):1449–1461.
[41] Zeltyn S, Mandelbaum A (2005) Call centers with impatient customers: Many-server asymptotics of the $M/M/n+G$ queue. *Queueing Systems* 51(3):361–402.
[42] Zhang B, van Leeuwaarden JSH, Zwart B (2012) Staffing call centers with impatient customers: Refinements to many-server asymptotics. *Oper. Res.* 60(2):461–474.