Stochastic Systems
2016, Vol. 6, No. 2, 301–366
DOI: 10.1214/15-SSY212

STEIN'S METHOD FOR STEADY-STATE DIFFUSION APPROXIMATIONS: AN INTRODUCTION THROUGH THE ERLANG-A AND ERLANG-C MODELS

By Anton Braverman, J. G. Dai, and Jiekun Feng

Cornell University

This paper provides an introduction to the Stein method framework in the context of steady-state diffusion approximations. The framework consists of three components: the Poisson equation and gradient bounds, generator coupling, and moment bounds. Working in the setting of the Erlang-A and Erlang-C models, we prove that both Wasserstein and Kolmogorov distances between the stationary distribution of a normalized customer count process, and that of an appropriately defined diffusion process decrease at a rate of $1/\sqrt{R}$, where R is the offered load. Futhermore, these error bounds are *universal*, valid in any load condition from lightly loaded to heavily loaded.

1. Introduction. In [10], the authors developed a framework based on Stein's method [53, 54] to prove the rates of convergence for steady-state diffusion approximations. Using their framework, they proved convergence rates for the steady-state approximation of the $M/Ph/n+M$ system, the many server queue with customer abandonment and phase-type service times, in the Halfin–Whitt regime [32]. The framework in [10] is modular and has four components: the Poisson equation and gradient bounds, generator coupling, moment bounds, and state space collapse (SSC). The purpose of this paper is to provide an accessible introduction to the Stein framework, focusing on two simple and yet fundamental systems. They are the $M/M/n+M$ system, known as the Erlang-A system, and $M/M/n$ system, known as the Erlang-C system. The accessibility is due to the fact that both systems can be represented by a one-dimensional continuous time Markov chain (CTMC). In addition, by focusing on these two systems we are able to present some sharp results that serve as benchmarks that future research should aspire to meet.

Stein's method is a powerful method used for studying approximations of probability distributions, and is best known for its ability to establish

Received December 2015.
MSC 2010 subject classifications: Primary 60K25; secondary 60F99, 60J60.
Keywords and phrases: Stein's method, steady-state, diffusion approximation, convergence rates, Erlang-A, Erlang-C.

convergence rates. It has been widely used in probability, statistics, and their wide range of applications such as bioinformatics; see, for example, the survey papers [52, 13], the recent book [15] and the references within. Applications of Stein's method always involve some unknown distribution to be approximated, and an approximating distribution. For instance, the first appearance of the method in [53] involved the sum of identically distributed dependent random variables as the unknown, and the normal as approximating distribution. Other approximating distributions include the Poisson [14], binomial [19], and multinomial [46] distributions, just to name a few. For each approximating distribution, one needs to establish separately gradient bounds, also known as Stein factors [60, 4], like those in Lemma 3 in Section 3. In this paper, the approximating distribution is the stationary distribution of a diffusion process, and the unknown is the stationary distribution of the CTMC introduced in (1.1) below.

Both Erlang-A and Erlang-C systems have n homogeneous servers that serve customers in a first-come-first-serve manner. Customers arrive according to a Poisson process with rate λ, and customer service times are assumed to be i.i.d. having exponential distribution with mean $1/\mu$. In the Erlang-A system, each customer has a patience time and when his waiting time in queue exceeds his patience time, he abandons the queue without service; the patience times are assumed to be i.i.d. having exponential distribution with mean $1/\alpha$. The Erlang-A system is a special case of the systems studied in [10], where SSC played an important role. The systems in this paper can be represented by a one-dimensional CTMC, meaning that there is no need to invoke SSC. Therefore, this paper illustrates only the first three components of the framework proposed in [10].

We will study the birth-death process

$$X = \{X(t), t \geq 0\}, \tag{1.1}$$

where $X(t)$ is the number of customers in the system at time t. In the Erlang-A system, α is assumed to be positive and therefore the mean patience time is finite. This guarantees that the CTMC X is positive recurrent. In the Erlang-C system, $\alpha = 0$, and in order for the CTMC to be positive recurrent we need to assume that the offered load to the system, defined as $R = \lambda/\mu$, satisfies

$$R < n. \tag{1.2}$$

For both Erlang-A and Erlang-C systems, we use $X(\infty)$ to denote the random variable having the stationary distribution of X.

Consider the case when $\alpha = 0$ and (1.2) is satisfied. Set

$$\tilde{X}(\infty) = (X(\infty) - R)/\sqrt{R},$$

and let $Y(\infty)$ denote a continuous random variable on $\mathbb{R}$ having density

$$\kappa \exp\Big(\frac{1}{\mu}\int_0^x b(y)dy\Big), \tag{1.3}$$

where $\kappa > 0$ is a normalizing constant that makes the density integrate to one,

$$b(x) = \big[(x+\zeta)^- - \zeta^-\big]\mu, \quad \text{and} \quad \zeta = \big(R - n\big)/\sqrt{R}. \tag{1.4}$$

Although our choice of notation does not make this explicit, we highlight that the random variable $Y(\infty)$ depends on λ, μ, and n, meaning that we are actually dealing with a *family* of random variables $\{Y^{(\lambda,\mu,n)}(\infty)\}_{(\lambda,\mu,n)}$. This plays a role in Lemma 3 in Section 3 for example, where we need to know exactly how the gradient bounds depend on λ, μ, and n. The following theorem illustrates the type of result that can be obtained by Stein's method.

THEOREM 1. *Consider the Erlang-C system ($\alpha = 0$). For all $n \geq 1$, $\lambda > 0$, and $\mu > 0$ satisfying $1 \leq R < n$,*

$$d_W(\tilde{X}(\infty), Y(\infty)) \equiv \sup_{h(x)\in \mathrm{Lip}(1)} \big|\mathbb{E}h(\tilde{X}(\infty)) - \mathbb{E}h(Y(\infty))\big| \leq \frac{205}{\sqrt{R}}, \tag{1.5}$$

where

$$\mathrm{Lip}(1) = \{h : \mathbb{R} \to \mathbb{R}, |h(x) - h(y)| \leq |x - y|\}.$$

The framework developed in [10] was inspired largely by the work of Gurvich in [28] who developed methodologies to prove statements similar to Theorem 1 for a broad class of multidimensional CTMCs. Along the way, he independently rediscovered many of the ideas central to Stein's method in the setting of steady-state diffusion approximations. See [10] for a more detailed discussion of Gurvich's results.

Several points are worth mentioning. First, we note that Theorem 1 is not a limit theorem. Steady-state approximations are usually justified by some kind of limit theorem. That is, one considers a sequence of queueing systems and proves that the corresponding sequence of steady-state distributions converges to some limiting distribution as traffic intensity approaches one, or as the number of servers goes to infinity. In contrast, our theorem holds for any finite parameter choices of λ, n, and μ satisfying (1.2) and $R \geq 1$.

TABLE 1
Comparing the error $\left|\mathbb{E}X(\infty) - \left(R + \sqrt{R}\mathbb{E}Y(\infty)\right)\right|$ *for different system configurations.*

$n = 5$			$n = 500$		
R	$\mathbb{E}X(\infty)$	Error	R	$\mathbb{E}X(\infty)$	Error
3	3.35	0.10	300	300.00	6×10^{-14}
4	6.22	0.20	400	400.00	2×10^{-6}
4.9	51.47	0.28	490	516.79	0.24
4.95	101.48	0.29	495	569.15	0.28
4.99	501.49	0.29	499	970.89	0.32

Second, the error bound in (1.5) is *universal*, as it does not assume any relationship between λ, n, and μ, other than the stability condition (1.2) and the condition that $R \geq 1$. The latter condition is a mere convenience, as Theorem 1 could be restated without it, but the error bound would then contain some terms involving $1/R$. One consequence of universality is that the error bound holds when parameters λ, n, and μ fall in one of the following asymptotic regimes:

$$n = \lceil R + \beta R \rceil, \quad n = \left\lceil R + \beta\sqrt{R} \right\rceil, \quad \text{or} \quad n = \lceil R + \beta \rceil,$$

where $\beta > 0$ is fixed, while $R \to \infty$. The first two parameter regimes above describe the quality-driven (QD), and quality-and-efficiency-driven (QED) regimes, respectively. The last regime is the nondegenerate-slowdown (NDS) regime, which was studied in [62, 1]. Universal approximations were previously studied in [59, 31]. Third, as part of the universality of Theorem 1, we see that

$$\left|\mathbb{E}X(\infty) - \left(R + \sqrt{R}\mathbb{E}Y(\infty)\right)\right| \leq 205. \tag{1.6}$$

For a fixed n, let $\rho = R/n \uparrow 1$. One expects that $\mathbb{E}X(\infty)$ be on the order of $1/(1-\rho)$. Conventional heavy-traffic limit theorems often guarantee that the left hand side of (1.6) is at most $o(1/\sqrt{1-\rho})$, whereas our error is bounded by a constant regardless of the load condition. This suggests that the diffusion approximation for the Erlang-C system is accurate not only as $R \to \infty$, but also in the heavy-traffic setting when $R \to n$. Table 1 contains some numerical results where we calculate the error on the left side of (1.6). The constant 205 in (1.6) is unlikely to be a sharp upper bound. In this paper we did not focus too much on optimizing the upper bound, as Stein's method is not known for producing sharp constants.

From Theorem 1 we know that the first moment of $\tilde{X}(\infty)$ can be approximated universally by the first moment of $Y(\infty)$. It is natural to ask what can be said about the approximation of higher moments. We performed

TABLE 2
Approximating the second and tenth moments of $\tilde{X}(\infty)$ with $n = 500$. The approximation error grows as R approaches n and suggests that the diffusion approximation of higher moments is not universal.

R	$\mathbb{E}(\tilde{X}(\infty))^2$	$\lvert\mathbb{E}(\tilde{X}(\infty))^2 - \mathbb{E}(Y(\infty))^2\rvert$	$\mathbb{E}(\tilde{X}(\infty))^{10}$	$\lvert\mathbb{E}(\tilde{X}(\infty))^{10} - \mathbb{E}(Y(\infty))^{10}\rvert$
300	1	4.55×10^{-15}	9.77×10^{2}	31.58
400	1	5.95×10^{-7}	9.70×10^{2}	24.44
490	6.96	0.11	7.51×10^{9}	7.01×10^{8}
495	31.56	0.27	9.10×10^{12}	4.34×10^{11}
499	9.47×10^{2}	1.59	1.07×10^{20}	1.03×10^{18}
499.9	9.94×10^{4}	16.50	1.13×10^{30}	1.09×10^{27}

some numerical experiments in which we approximate the second and tenth moments of $\tilde{X}(\infty)$ in a system with $n = 500$. The results are displayed in Table 2. One can see that the approximation errors grow as the offered load R gets closer to n. We will see in Section 6 that this happens because the $(m-1)$th moment appears in the approximation error of the mth moment. A similar phenomenon was first observed for the $M/GI/1 + GI$ model in Theorem 1 of [30].

Theorem 1 provides rates of convergence under the Wasserstein metric [52]. The Wasserstein metric is one of the most commonly studied metrics in the context of Stein's method. This is because the the space Lip(1) is relatively simple to work with, but is also rich enough so that convergence under the Wasserstein metric implies the convergence in distribution [25]. Another metric commonly studied in problems involving Stein's method is the Kolmogorov metric, which measures the distance between cumulative distribution functions of two random variables. The Kolmogorov distance between $\tilde{X}(\infty)$ and $Y(\infty)$ is

$$\sup_{h(x)\in\mathcal{H}_K} \left|\mathbb{E}h(\tilde{X}(\infty)) - \mathbb{E}h(Y(\infty))\right|, \quad \text{where} \quad \mathcal{H}_K = \{1_{(-\infty,a]}(x) : a \in \mathbb{R}\}.$$

Theorems 3 and 4 of Section 2 involve the Kolmogorov metric. A general trend in Stein's method is that establishing convergence rates for the Kolmogorov metric often requires much more effort than establishing rates for the Wasserstein metric, and our problem is no exception. The extra difficulty always comes from the fact that the test functions belonging to the class $\mathcal{H}_K$ are discontinuous, whereas the ones in Lip(1) are Lipschitz-continuous. In Section 5, we describe how to overcome this difficulty in our model setting.

The first paper to have established convergence rates for steady-state diffusion approximations was [31], which studied the Erlang-A system using an excursion based approach. Their approximation error bounds are universal. Although the authors in [31] did not study the Erlang-C system, their ap-

proach appears to be extendable to it as well. However, their method is not readily generalizable to the multi-dimensional setting.

We wish to point out that the results proved in this paper are quite sharp, and proving analogous results in a high dimensional setting is likely much more difficult. For instance, the constants in the error bounds in each theorem of this paper can be recovered explicitly, but this is not true in the problem studied in [10]. Moreover, the result of [10] is restricted to the Halfin–Whitt regime, and is not universal. All of this is because the model considered there is high dimensional.

1.1. *Related literature.* Diffusion approximations are a popular tool in queueing theory, and are usually "justified" by heavy traffic limit theorems. For example, a typical limit theorem would say that an appropriately scaled and centered version of the process X in (1.1) converges to some limiting diffusion process as the system utilization ρ tends to one. Proving such limit theorems has been an active area of research in the last 50 years; see, for example, [7, 8, 37, 38, 33, 51] for single-class queueing networks, [49, 9, 63] for multiclass queueing networks, [42, 64] for bandwidth sharing networks, [32, 50, 17] for many-server queues. The convergence used in these limit theorems is the convergence in distribution on the path space $\mathbb{D}([0,\infty),\mathbb{R}^d)$, endowed with Skorohod J_1-topology [20, 61]. The J_1-topology on $\mathbb{D}([0,\infty),\mathbb{R}^d)$ essentially means convergence in $\mathbb{D}([0,T],\mathbb{R}^d)$ for each $T>0$. In particular, it says nothing about the convergence at "∞". Therefore, these limit theorems do not justify steady-state convergence.

The jump from convergence on $\mathbb{D}([0,T],\mathbb{R}^d)$ to convergence of stationary distributions was first established in the seminal paper [22], where the authors prove an interchange of limits for generalized Jackson networks of single-server queues. The results in [22] were improved and extended by various authors for networks of single-servers [12, 68, 43], for bandwidth sharing networks [64], and for many-server systems [58, 21, 29]. These "interchange of limits" theorems are qualitative and thus do not provide rates of convergence as in Theorem 1.

The first uses of Stein's method for stationary distributions of Markov processes traces back to [5], where it is pointed out that Stein's method can be applied anytime the approximating distribution is the stationary distribution of a Markov proccess. That paper considers the multivariate Poisson, which is the stationary distribution of a certain multi-dimensional birth-death process. One of the major contributions of that paper was to show how viewing the Poisson distribution as the stationary distribution of a Markov chain could be exploited to establish gradient bounds using coupling

arguments; cf. the discussion around (3.20) of this paper. A similar idea was subsequently used for the multivariate normal distribution through its connection to the multi-dimensional Ornstein–Uhlenbeck process in [2, 27].

Of the papers that use the connection between Stein's method and Markov processes, [11] and the more recent [44] are the most relevant to this work. The former studies one-dimensional birth-death processes, with the focus being that many common distributions such as the Poisson, Binomial, Hypergeometric, Negative Binomial, etc., can be viewed as stationary distributions of a birth-death process. Although the Erlang-A and Erlang-C models are also birth-death processes, the focus in our paper is on how well these models can be approximated by diffusions, e.g. qualitative features of the approximation like the universality in Theorem 1. Diffusion approximations go beyond approximations of birth-death processes, with the real interest lying in cases when a higher-dimensional Markov chain collapses to a one-dimensional diffusion, e.g. [58, 55, 18], or when the diffusion approximation is multi-dimensional [33, 51, 49, 9, 63].

In [44], the authors apply Stein's method to one-dimensional diffusions. The motivation is again that many common distributions like the gamma, uniform, beta, etc., happen to be stationary distributions of diffusions. Their chief result is to establish gradient bounds for a very large class of diffusion processes, requiring only the mild condition that the drift of the diffusion be a decreasing function. However, their result cannot be applied here, because it is impossible to say how their gradient bounds depend on the parameters of the diffusion. Detailed knowledge of this dependence is crucial, because we are dealing with a *family* of approximating distributions; cf. (1.3) and the comments below (1.4).

Outside the diffusion approximation domain, Ying has recently successfully applied Stein's framework to establish error bounds for steady-state mean-field approximations [65, 66]. There is one additional recent line of work [6, 39, 40, 41, 67] that deserves mention, where the theme is corrected diffusion approximations using asymptotic series expansions. In particular, [41] considers the Erlang-C system and [67] considers the Erlang-A system. In these papers, the authors derive series expansions for various steady-state quantities of interest like the probability of waiting $\mathbb{P}(X(\infty) \geq n)$. These types of series expansions are very powerful because they allow one to approximate steady-state quantities of interest within arbitrary precision. However, while accurate, these expansions vary for different performance metrics (e.g. waiting probability, expected queue length), and require non-trivial effort to be derived. They also depend on the choice of parameter regime, e.g. Halfin-Whitt. In contrast, the results provided by the Stein

approach can be viewed as more robust because they capture multiple performance metrics and multiple parameter regimes at the same time.

1.2. *Notation.* For two random variables U and V, define their Wasserstein distance to be

$$d_W(U,V) = \sup_{h(x)\in \mathrm{Lip}(1)} \left|\mathbb{E}[h(U)] - \mathbb{E}[h(V)]\right|, \tag{1.7}$$

where

$$\mathrm{Lip}(1) = \{h : \mathbb{R} \to \mathbb{R}, |h(x) - h(y)| \le |x - y|\}.$$

It is known, see for example [52], convergence under the Wasserstein metric implies convergence in distribution. When Lip(1) in (1.7) is replaced by

$$\mathcal{H}_K = \{1_{(-\infty,a]}(x) : a \in \mathbb{R}\}, \tag{1.8}$$

the corresponding distance is the Kolmogorov distance, denoted by $d_K(U,V)$. For $a, b \in \mathbb{R}$, we use $a^+, a^-, a \wedge b$, and $a \vee b$ to denote $\max(a,0)$, $\max(-a,0)$, $\min(a,b)$, and $\max(a,b)$, respectively.

The rest of the paper is structured as follows. In Section 2 we state our main results. In Section 3 we present the Stein framework needed to prove our main results, Theorems 1–4, by introducing three ingredients central to our framework. Namely, the Poisson equation and gradient bounds, generator coupling, and moment bounds. In Section 4 we prove Theorem 1, which deals with the Wasserstein metric. The Kolmogorov metric presents an additional challenge, because the test functions are discontinuous. In Section 5 we address this new challenge. In Section 6, we discuss the approximation of higher moments in the Erlang-C model. Proofs of technical lemmas are left to the four appendices.

2. Main results. Recall the offered load $R = \lambda/\mu$. For notational convenience we define $\delta > 0$ as

$$\delta = \frac{1}{\sqrt{R}} = \sqrt{\frac{\mu}{\lambda}}.$$

Let $x(\infty)$ be the unique solution to the flow balance equation

$$\lambda = \big(x(\infty) \wedge n\big)\mu + \big(x(\infty) - n\big)^+\alpha. \tag{2.1}$$

Here, $x(\infty)$ is interpreted as the equilibrium number of customers in the corresponding fluid model, and is the point at which the arrival rate equals the departure rate. The latter is the sum of the service completion rate and

the customer abandonment rate with $x(\infty)$ customers in the system. One can check that the flow balance equation has a unique solution $x(\infty)$ given by

$$x(\infty) = \begin{cases} n + \frac{\lambda - n\mu}{\alpha} & \text{if } R \geq n, \\ R & \text{if } R < n. \end{cases} \tag{2.2}$$

By noting that the number of busy servers $x(\infty) \wedge n$ equals n minus the number of idle servers $(x(\infty) - n)^-$, the equation in (2.1) becomes

$$\lambda - n\mu = \big(x(\infty) - n\big)^+ \alpha - \big(x(\infty) - n\big)^- \mu. \tag{2.3}$$

We note that $x(\infty)$ is well-defined even when $\alpha = 0$, because in that case we always assume that $R < n$.

We consider the CTMC

$$\tilde{X} = \{\tilde{X}(t) \equiv \delta(X(t) - x(\infty)),\ t \geq 0\}, \tag{2.4}$$

and let the random variable $\tilde{X}(\infty)$ have its stationary distribution. Define

$$\zeta = \delta\big(x(\infty) - n\big), \tag{2.5}$$

and

$$b(x) = \big[(x + \zeta)^- - \zeta^-\big]\mu - \big[(x + \zeta)^+ - \zeta^+\big]\alpha \quad \text{for } x \in \mathbb{R}, \tag{2.6}$$

with convention that α is set to be zero in the Erlang-C system. For intuition about the quantity ζ, we note that in the Erlang-C system satisfying (1.2),

$$n = R - \zeta\sqrt{R}.$$

Thus, $-\zeta = |\zeta| > 0$ is precisely the "safety coefficient" in the square-root safety-staffing principle [24, equation (15)]. We point out that the event $\{\tilde{X}(t) = -\zeta\}$ corresponds to the event $\{X(t) = n\}$.

Throughout this paper, let $Y(\infty)$ denote a continuous random variable on $\mathbb{R}$ having density

$$\nu(x) = \kappa \exp\Big(\frac{1}{\mu}\int_0^x b(y)dy\Big), \tag{2.7}$$

where $\kappa > 0$ is a normalizing constant that makes the density integrate to one. Note that these definitions are consistent with (1.3) and (1.4).

THEOREM 2. *Consider the Erlang-A system ($\alpha > 0$). There exists an increasing function $C_W : \mathbb{R}_+ \to \mathbb{R}_+$ such that for all $n \geq 1, \lambda > 0, \mu > 0$, and $\alpha > 0$ satisfying $R \geq 1$,*

$$d_W(\tilde{X}(\infty), Y(\infty)) \leq C_W(\alpha/\mu)\delta. \tag{2.8}$$

REMARK 1. The proof of Theorems 1 and 2 uses the same ideas. Therefore, for the sake of brevity, we only give an outline for the proof of Theorem 2 and its ingredients in Appendix C.1, without filling in all the details. It is for this reason that we do not write out the explicit form of $C_W(\alpha/\mu)$, although it can be obtained from the proof. The same is true for Theorem 4 below.

Given two random variables U and V, [52, Proposition 1.2] implies that when V has a density that is bounded by $C > 0$,

$$d_K(U, V) \leq \sqrt{2C d_W(U, V)}. \tag{2.9}$$

At best, (2.9) and Theorems 1 and 2 imply a convergence rate of $\sqrt{\delta}$ for $d_K(\tilde{X}(\infty), Y(\infty))$. However, this bound is typically too crude, and the following two theorems show that convergence happens at rate δ. Theorem 3 is proved in Section 5.3. The proof of Theorem 4 is outlined in Appendix C.2.

THEOREM 3. *Consider the Erlang-C system ($\alpha = 0$). For all $n \geq 1$, $\lambda > 0$, and $\mu > 0$ satisfying $1 \leq R < n$,*

$$d_K(\tilde{X}(\infty), Y(\infty)) \leq 188\delta. \tag{2.10}$$

THEOREM 4. *Consider the Erlang-A system ($\alpha > 0$). There exists an increasing function $C_K : \mathbb{R}_+ \to \mathbb{R}_+$ such that for all $n \geq 1, \lambda > 0, \mu > 0$, and $\alpha > 0$ satisfying $R \geq 1$,*

$$d_K(\tilde{X}(\infty), Y(\infty)) \leq C_K(\alpha/\mu)\delta. \tag{2.11}$$

Theorems 1 and 3 are new, but versions of Theorems 2 and 4 were first proved in the pioneering paper [31] using an excursion based approach. However, our notion of universality in those theorems is stronger than the one in [31], because most of their results require μ and α to be fixed. The only exception is in Appendix C of that paper, where the authors consider the NDS regime with $\mu = \mu(\lambda) = \beta\sqrt{\lambda}$ and $\lambda = n\mu + \beta_1\mu$ for some $\beta > 0$ and $\beta_1 \in \mathbb{R}$.

We emphasize that both constants C_W and C_K are increasing in α/μ. That is, for an Erlang-A system with a higher abandonment rate with respect

to its service rate, our error bound becomes larger. The reader may wonder why these constants depend on α/μ, while the constant in the Erlang-C theorems does not depend on anything. Despite our best efforts, we were unable to get rid of the dependency on α/μ. The reason is that the Erlang-C model depends on only three parameters (λ, μ, n), while the Erlang-A model also depends on α. As a result, both the gradient bounds and moment bounds have an extra factor α/μ in the Erlang-A model. For example, compare Lemma 3 in Section 3.5 with Lemma 13 in Appendix B.2.

3. Outline of the Stein framework. In this section we introduce the main tools needed to prove Theorems 1–4. At this point we do not restrict ourselves to either the Erlang-A or Erlang-C systems, as the framework outlined here is generic and holds for both systems.

The following is an informal outline of the rest of this section. We know that $\tilde{X}(\infty)$ follows the stationary distribution of the CTMC $\tilde{X}$, and that this CTMC has a generator $G_{\tilde{X}}$. To $Y(\infty)$, we will associate a diffusion process with generator G_Y. We will start by fixing a test function $h : \mathbb{R} \to \mathbb{R}$ and deriving the identity

$$\big|\mathbb{E}h(\tilde{X}(\infty)) - \mathbb{E}h(Y(\infty))\big| = \big|\mathbb{E}G_{\tilde{X}} f_h(\tilde{X}(\infty)) - \mathbb{E}G_Y f_h(\tilde{X}(\infty))\big|, \tag{3.1}$$

where $f_h(x)$ is a solution to the Poisson equation

$$G_Y f_h(x) = \mathbb{E}h(Y(\infty)) - h(x), \quad x \in \mathbb{R}.$$

We then focus on bounding the right hand side of (3.1), which is easier to handle than the left hand side. This is done by performing a Taylor expansion of $G_{\tilde{X}} f_h(x)$ in Section 3.3. To bound the error term from the Taylor expansion, we require bounds on various moments of $\big|\tilde{X}(\infty)\big|$, as well as the derivatives of $f_h(x)$. We refer to the former as moment bounds, and the latter as gradient bounds. These are presented in Sections 3.4 and 3.5, respectively.

3.1. *The Poisson equation of a diffusion process.* The random variable $Y(\infty)$ in Theorems 1–4 is well-defined and its density is given in (2.7). It turns out that $Y(\infty)$ has the stationary distribution of a diffusion process $Y = \{Y(t), t \geq 0\}$, which we will define shortly. We do not prove this claim in this paper since it is not used anywhere in this paper. Nevertheless, it is helpful to think of $Y(\infty)$ in the context of diffusion processes. The diffusion process Y is the one-dimensional piecewise Ornstein–Uhlenbeck (OU) process. Its generator is given by

$$G_Y f(x) = b(x) f'(x) + \mu f''(x) \quad \text{for } x \in \mathbb{R},\ f \in C^2(\mathbb{R}), \tag{3.2}$$

where $b(x)$ is defined in (2.6). Clearly, $b(0) = 0$, and $b(x)$ is Lipschitz continuous. Indeed,

$$|b(x) - b(y)| \le (\alpha \vee \mu)\,|x - y| \qquad \text{for } x, y \in \mathbb{R}.$$

Since the diffusion process Y depends on parameters λ, n, μ, and α in an arbitrary way, there is no appropriate way to talk about the limit of $Y(\infty)$ in terms of these parameters. Therefore, we call Y a *diffusion model*, as opposed to a *diffusion limit*. Having a diffusion model whose input parameters are directly taken from the corresponding Markov chain model is critical to achieve *universal accuracy*. In other words, this diffusion model is accurate in any parameter regime, from underloaded, to critically loaded, and to overloaded. Diffusion models, not limits, of queueing networks with a given set of parameters have been advanced in [35, 34, 16, 59, 28, 31, 30].

The main tool we use is known as the Poisson equation. It allows us to say that $Y(\infty)$ is a good estimate for $\tilde{X}(\infty)$ if the generator of Y behaves similarly to the generator of $\tilde{X}$, where $\tilde{X}$ is defined in (2.4). Let $\mathcal{H}$ be a class of functions $h : \mathbb{R} \to \mathbb{R}$, to be specified shortly. For each function $h(x) \in \mathcal{H}$, consider the Poisson equation

$$G_Y f_h(x) = b(x) f_h'(x) + \mu f_h''(x) = \mathbb{E}h(Y(\infty)) - h(x), \quad x \in \mathbb{R}. \tag{3.3}$$

One may verify by differentiation that for all functions $h : \mathbb{R} \to \mathbb{R}$ satisfying $\mathbb{E}\big|h(Y(\infty))\big| < \infty$, the Poisson equation has a family of solutions of the form

$$f_h(x) = a_1 + \int_0^x \left[a_2 \frac{1}{\nu(u)} + \frac{1}{\nu(u)} \int_{-\infty}^u \frac{1}{\mu} \big(\mathbb{E}h(Y(\infty)) - h(y)\big)\nu(y)dy \right] du, \tag{3.4}$$

where $a_1, a_2 \in \mathbb{R}$ are arbitrary constants, and $\nu(x)$ is as in (2.7).

In this paper, we take $\mathcal{H} = \mathrm{Lip}(1)$ when we deal with the Wasserstein metric (Theorems 1 and 2), and we choose $\mathcal{H} = \mathcal{H}_K$ when we deal with the Kolmogorov metric (Theorems 3 and 4). We claim that $|\mathbb{E}h(Y(\infty))| < \infty$. Indeed, when $\mathcal{H} = \mathcal{H}_K$, this clearly holds. When $\mathcal{H} = \mathrm{Lip}(1)$, without loss of generality we take $h(0) = 0$ in (3.3), and use the Lipschitz property of $h(x)$ to see that

$$|\mathbb{E}h(Y(\infty))| \le \mathbb{E}\,|Y(\infty)| < \infty,$$

where the finiteness of $\mathbb{E}\,|Y(\infty)|$ will be proved in (B.16).

From (3.3), one has

$$\big|\mathbb{E}h(\tilde{X}(\infty)) - \mathbb{E}h(Y(\infty))\big| = \big|\mathbb{E}G_Y f_h(\tilde{X}(\infty))\big|. \tag{3.5}$$

In (3.5), $\tilde{X}(\infty)$ has the stationary distribution of the CTMC $\tilde{X}$, not necessarily defined on the same probability space of $Y(\infty)$. Actually, $\tilde{X}(\infty)$ in (3.5) can be replaced by any other random variable, although one does not expect the error on the right side to be small if this random variable has no relationship with the diffusion process Y.

3.2. *Comparing generators.* To prove Theorems 1–4, we need to bound the right side of (3.5). The CTMC $\tilde{X}$ defined in (2.4) also has a generator. We bound the right side of (3.5) by showing that the diffusion generator in (3.2) is similar to the CTMC generator.

For any $k \in \mathbb{Z}_+$, we define $x = x_k = \delta(k - x(\infty))$. Then for any function $f : \mathbb{R} \to \mathbb{R}$, the generator of $\tilde{X}$ is given by

$$G_{\tilde{X}} f(x) = \lambda(f(x+\delta) - f(x)) + d(k)(f(x-\delta) - f(x)), \tag{3.6}$$

where

$$d(k) = \mu(k \wedge n) + \alpha(k-n)^+, \tag{3.7}$$

is the departure rate corresponding to the system having k customers. One may check that

$$b(x) = \delta(\lambda - d(k)). \tag{3.8}$$

The relationship between $G_{\tilde{X}}$ and the stationary distribution of $\tilde{X}$ is illustrated by the following lemma.

LEMMA 1. *Let $f(x) : \mathbb{R} \to \mathbb{R}$ be a function such that $|f(x)| \le C(1+x)^2$ for some $C > 0$ (i.e. $f(x)$ is dominated by a quadratic function), and assume that the CTMC $\tilde{X}$ is positive recurrent. Then*

$$\mathbb{E}\big[G_{\tilde{X}} f(\tilde{X}(\infty))\big] = 0.$$

REMARK 2. We will see in Lemma 3 later this section, in Lemmas 4 and 5 of Section 5, and in Lemma 13 of Appendix B.2 that there is a family of solutions to the Poisson equation (3.3) whose first derivatives grow at most linearly in both the Wasserstein and Kolmgorov settings, meaning that these solutions satisfy the conditions of Lemma 1.

The proof of Lemma 1 is provided in Appendix D.1, and relies on Proposition 3 of [26]. Suppose for now that for any $h(x) \in \mathcal{H}$, the solution to the

Poisson equation $f_h(x)$ satisfies the conditions of Lemma 1. We can apply Lemma 1 to (3.5) to see that

$$\begin{aligned}\left|\mathbb{E}h(\tilde{X}(\infty)) - \mathbb{E}h(Y(\infty))\right| &= \left|\mathbb{E}G_Y f_h(\tilde{X}(\infty))\right| \\ &= \left|\mathbb{E}G_{\tilde{X}} f_h(\tilde{X}(\infty)) - \mathbb{E}G_Y f_h(\tilde{X}(\infty))\right| \\ &\le \mathbb{E}\left|G_{\tilde{X}} f_h(\tilde{X}(\infty)) - G_Y f_h(\tilde{X}(\infty))\right|. \end{aligned} \tag{3.9}$$

While the two random variables on the left side of (3.9) are usually defined on different probability spaces, the two random variables on the right side of (3.9) are both functions of $\tilde{X}(\infty)$. Thus, we have achieved a coupling through Lemma 1. Setting up the Poisson equation is a generic first step one performs any time one wishes to apply Stein's method to a problem. The next step is to bound the equivalent of our $\left|\mathbb{E}G_Y f_h(\tilde{X}(\infty))\right|$. This is usually done by using a coupling argument. However, this coupling is always problem specific, and is one of the greatest sources of difficulty one encounters when applying Stein's method. In our case, this generator coupling is natural because we deal with Markov processes $\tilde{X}$ and Y.

Since the generator completely characterizes the behavior of a Markov process, it is natural to expect that convergence of generators implies convergence of Markov processes. Indeed, the question of weak convergence was studied in detail, for instance in [20], using the martingale problem of Stroock and Varadhan [57]. However, (3.9) lets us go beyond weak convergence, both because different choices of $h(x)$ lead to different metrics of convergence, and also because the question of convergence rates can be answered. One interpretation of the Stein approach is to view $f_h(x)$ as a Lyapunov function that gives us information about $h(x)$. Instead of searching very hard for this Lyapunov function, the Poisson equation (3.3) removes the guesswork. However, this comes at the cost of $f_h(x)$ being defined implicitly as the solution to a differential equation.

3.3. *Taylor expansion.* To bound the right side of (3.9), we study the difference $G_{\tilde{X}} f_h(x) - G_Y f_h(x)$. For that we perform a Taylor expansion on $G_{\tilde{X}} f_h(x)$. To illustrate this, suppose that $f_h''(x)$ exists for all $x \in \mathbb{R}$, and is absolutely continuous. Then for any $k \in \mathbb{Z}_+$, and $x = x_k = \delta(k - x(\infty))$, we recall that $b(x) = \delta(\lambda - d(k))$ in (3.8) to see that

$$\begin{aligned} G_{\tilde{X}} f_h(x) &= \lambda(f_h(x+\delta) - f_h(x)) + d(k)(f_h(x-\delta) - f_h(x)) \\ &= f_h'(x)\delta(\lambda - d(k)) + \frac{1}{2}\delta^2 f_h''(x)(\lambda + d(k)) \\ &\quad + \frac{1}{2}\lambda\delta^2(f_h''(\xi) - f_h''(x)) + \frac{1}{2}d(k)\delta^2(f_h''(\eta) - f_h''(x)) \end{aligned}$$

$$
\begin{aligned}
&= f_h'(x)b(x) + \frac{1}{2}\delta^2(2\lambda - \frac{1}{\delta}b(x))f_h''(x) \\
&\quad + \frac{1}{2}\mu(f_h''(\xi) - f_h''(x)) + \frac{1}{2}(\lambda - \frac{1}{\delta}b(x))\delta^2(f_h''(\eta) - f_h''(x)) \\
&= G_Y f_h(x) - \frac{1}{2}\delta f_h''(x)b(x) + \frac{1}{2}\mu(f_h''(\xi) - f_h''(x)) \\
&\quad + \frac{1}{2}(\mu - \delta b(x))(f_h''(\eta) - f_h''(x)),
\end{aligned}
$$

where $\xi \in [x, x+\delta]$ and $\eta \in [x-\delta, x]$. We invoke the absolute continuity of $f_h''(x)$ to get

$$
\begin{aligned}
&\left|\mathbb{E}h(\tilde{X}(\infty)) - \mathbb{E}h(Y(\infty))\right| \\
&\le \frac{1}{2}\delta\mathbb{E}\Big[\big|f_h''(\tilde{X}(\infty))b(\tilde{X}(\infty))\big|\Big] + \frac{\mu}{2}\mathbb{E}\Big[\int_{\tilde{X}(\infty)}^{\tilde{X}(\infty)+\delta} \big|f_h'''(y)\big|\,dy\Big] \\
&\quad + \frac{\mu}{2}\mathbb{E}\Big[\int_{\tilde{X}(\infty)-\delta}^{\tilde{X}(\infty)} \big|f_h'''(y)\big|\,dy\Big] + \frac{1}{2}\delta\mathbb{E}\Big[\big|b(\tilde{X}(\infty))\big|\int_{\tilde{X}(\infty)-\delta}^{\tilde{X}(\infty)} \big|f_h'''(y)\big|\,dy\Big].
\end{aligned} \tag{3.10}
$$

As one can see, to show that the right hand side of (3.10) vanishes as $\delta \to 0$, we must be able to bound the derivatives of $f_h(x)$; we refer to these as gradient bounds. Furthermore, we will also need bounds on moments of $\big|\tilde{X}(\infty)\big|$; we refer to these as moment bounds. Both moment and gradient bounds will vary between the Erlang-A or Erlang-C setting, and the gradient bounds will be different for the Wasserstein, and Kolmogorov settings. Moment bounds will be discussed shortly, and gradient bounds in the Wasserstein setting will be presented in Section 3.5. We discuss the Kolmogorov setting separately in Section 5. In that case we face an added difficulty because $f_h''(x)$ has a discontinuity, and we cannot use (3.10) directly.

3.4. *Moment bounds.* The following lemma presents the necessary moment bounds to bound (3.10) in the Erlang-C model, and is proved in Appendix A.1. These moment bounds are used in both the Wasserstein and the Kolmogorov metric settings.

LEMMA 2. *Consider the Erlang-C model ($\alpha = 0$). For all $n \ge 1, \lambda > 0$, and $\mu > 0$ satisfying $0 < R < n$,*

$$
\mathbb{E}\Big[(\tilde{X}(\infty))^2 1(\tilde{X}(\infty) \le -\zeta)\Big] \le \frac{4}{3} + \frac{2\delta^2}{3}, \tag{3.11}
$$

$$
\mathbb{E}\Big[\big|\tilde{X}(\infty)1(\tilde{X}(\infty) \le -\zeta)\big|\Big] \le \sqrt{\frac{4}{3} + \frac{2\delta^2}{3}}, \tag{3.12}
$$

$$\mathbb{E}\Big[\big|\tilde{X}(\infty)1(\tilde{X}(\infty)\le -\zeta)\big|\Big] \le 2\,|\zeta| \tag{3.13}$$

$$\mathbb{E}\Big[\big|\tilde{X}(\infty)1(\tilde{X}(\infty)\ge -\zeta)\big|\Big] \le \frac{1}{|\zeta|}+\frac{\delta^2}{4\,|\zeta|}+\frac{\delta}{2}, \tag{3.14}$$

$$\mathbb{P}(\tilde{X}(\infty)\le -\zeta) \le (2+\delta)\,|\zeta|\,. \tag{3.15}$$

We see that (3.14) immediately implies that when $\delta \le 1$,

$$\begin{aligned} |\zeta|\,\mathbb{P}(\tilde{X}(\infty)\ge -\zeta) &\le |\zeta| \wedge \mathbb{E}\Big[\big|\tilde{X}(\infty)1(\tilde{X}(\infty)\ge -\zeta)\big|\Big] \\ &\le |\zeta| \wedge \Big(\frac{1}{|\zeta|}+\frac{\delta^2}{4\,|\zeta|}+\frac{\delta}{2}\Big) \\ &\le 7/4, \end{aligned} \tag{3.16}$$

where to get the last inequality we considered separately the cases where $|\zeta| \le 1$ and $|\zeta| \ge 1$. This bound will be used in the proofs of Theorems 1 and 3.

One may wonder why the bounds are separated using the indicators $1\{\tilde{X}(\infty) \le -\zeta\}$ and $1\{\tilde{X}(\infty) \ge -\zeta\}$. This is related to the drift $b(x)$ appearing in (3.10), and the fact that $b(x)$ takes different forms on the regions $x \le -\zeta$ and $x \ge -\zeta$. Furthermore, it may be unclear at this point why both (3.12) and (3.13) are needed, as the left hand side in both bounds is identical. The reason is that (3.12) is an $O(1)$ bound (we think of $\delta \le 1$), whereas (3.13) is an $O(|\zeta|)$ bound. The latter is only useful when $|\zeta|$ is small, but this is nevertheless an essential bound to achieve universal results. As we will see later, it negates $1/|\zeta|$ terms that appear in (3.10) from $f_h''(x)$ and $f_h'''(x)$.

For the Erlang-A model, we also require moment bounds similar to those stated in Lemma 2. Both the proof, and subsequent usage, of the Erlang-A moment bounds are similar to the proof and subsequent usage of the Erlang-C moment bounds. We therefore delay the precise statement of the Erlang-A bounds until Lemma 11 in Appendix A.2, to avoid distracting the reader with a bulky lemma.

3.5. *Wasserstein gradient bounds.* Recall the Poisson equation (3.3) and the family of solutions to this equation is given by (3.4); in particular, this family is parametrized by constants $a_1, a_2 \in \mathbb{R}$. Fix $h(x) \in \mathcal{H} = \mathrm{Lip}(1)$, and let $f_h(x)$ be a solution to the Poisson equation. The following lemma presents Wasserstein gradient bounds for the Erlang-C model. It is proved in Appendix B.1.

LEMMA 3. *Consider the Erlang-C model ($\alpha = 0$). The solution to the Poisson equation $f_h(x)$ is twice continuously differentiable, with an absolutely continuous second derivative. Fix a solution in (3.4) with parameter $a_2 = 0$. Then for all $n \geq 1, \lambda > 0$, and $\mu > 0$ satisfying $0 < R < n$,*

$$|f_h'(x)| \leq \begin{cases} \frac{1}{\mu}(6.5 + 4.2/|\zeta|), & x \leq -\zeta, \\ \frac{1}{\mu}\frac{1}{|\zeta|}(x + 1 + 2/|\zeta|), & x \geq -\zeta. \end{cases} \tag{3.17}$$

$$|f_h''(x)| \leq \begin{cases} \frac{32}{\mu}(1 + 1/|\zeta|), & x \leq -\zeta, \\ \frac{1}{\mu|\zeta|}, & x \geq -\zeta, \end{cases} \tag{3.18}$$

and for those $x \in \mathbb{R}$ where $f_h'''(x)$ exists,

$$|f_h'''(x)| \leq \begin{cases} \frac{1}{\mu}(23 + 13/|\zeta|), & x \leq -\zeta, \\ 2/\mu, & x \geq -\zeta. \end{cases} \tag{3.19}$$

REMARK 3. This lemma validates the Taylor expansion used to obtain (3.10) because $f_h''(x)$ is absolutely continuous. Furthermore, $f_h(x)$ satisfies the conditions of Lemma 1, because $f_h'(x)$ grows at most linearly.

Gradient bounds, also known as Stein factors, are central to any application of Stein's method. The problem of gradient bounds for diffusion approximations can be divided into two cases: the one-dimensional case, and the multi-dimensional case. In the former, the Poisson equation in (3.3) is an ordinary differential equation (ODE) corresponding to a one-dimensional diffusion process. In the latter, the Poisson equation is a partial differential equation (PDE) corresponding to a multi-dimensional diffusion process.

The one-dimensional case is simpler, because the explicit form of $f_h(x)$ is given to us by (3.4). To bound $f_h'(x)$ and $f_h''(x)$ we can analyze (3.4) directly, as we do in the proof of Lemma 3. This direct analysis can be used as a go-to method for one-dimensional diffusions, but fails in the multi-dimensional case, because closed form solutions for PDE's are not typically known. In this case, it helps to exploit the fact that $f_h(x)$ satisfies

$$f_h(x) = \int_0^\infty \Big(\mathbb{E}\big[h(Y(t)) \mid Y(0) = x\big] - \mathbb{E}h(Y(\infty))\Big)dt, \tag{3.20}$$

where $Y = \{Y(t), t \geq 0\}$ is a diffusion process with generator G_Y [48]. To bound derivatives of $f_h(x)$ based on (3.20), one may use coupling arguments to bound finite differences of the form $\frac{1}{s}(f_h(x + s) - f_h(x))$. For examples of coupling arguments, see [5, 3, 11, 4, 23]. A related paper to these types of gradient bounds is [56], where the author used a variant of (3.20) for the

fluid model of a flexible-server queueing system as a Lyapunov function. As an alternative to coupling, one may combine (3.20) with a-priori Schauder estimates from PDE theory, as was done in [28].

Just like we did with the moment bounds, we delay the Erlang-A gradient bounds to Lemma 13 in Appendix B.2. We are now ready to prove Theorem 1.

4. Proof of Theorem 1. In this section we prove Theorem 1. Fix $h(x) \in \mathrm{Lip}(1)$, and recall that the family of solutions to the Poisson equation is given by (3.4). For the remainder of Section 4, we fix one such solution $f_h(x)$ with $a_2 = 0$. Then by Lemma 3 , $f_h''(x)$ is absolutely continuous, implying that (3.10) holds; we recall it here as

$$\begin{aligned}
&\left|\mathbb{E}h(\tilde{X}(\infty)) - \mathbb{E}h(Y(\infty))\right| \\
&\le \frac{1}{2}\delta\mathbb{E}\Big[\big|f_h''(\tilde{X}(\infty))b(\tilde{X}(\infty))\big|\Big] + \frac{\mu}{2}\mathbb{E}\left[\int_{\tilde{X}(\infty)}^{\tilde{X}(\infty)+\delta} \big|f_h'''(y)\big|\,dy\right] \\
&+ \frac{\mu}{2}\mathbb{E}\left[\int_{\tilde{X}(\infty)-\delta}^{\tilde{X}(\infty)} \big|f_h'''(y)\big|\,dy\right] + \frac{1}{2}\delta\mathbb{E}\left[\big|b(\tilde{X}(\infty))\big| \int_{\tilde{X}(\infty)-\delta}^{\tilde{X}(\infty)} \big|f_h'''(y)\big|\,dy\right],
\end{aligned} \tag{4.1}$$

where $\delta = 1/\sqrt{R} = \sqrt{\mu/\lambda}$.

The proof of Theorem 1 simply involves applying the moment bounds and gradient bounds to show that the error bound in (4.1) is small.

Proof of Theorem 1. Throughout the proof we assume that $R \ge 1$, or equivalently, $\delta \le 1$. We bound each of the terms on the right side of (4.1) individually. We recall here that the support of $\tilde{X}(\infty)$ is a δ-spaced grid, and in particular this grid contains the point $-\zeta$. In the bounds that follow, we will often consider separately the cases where $\tilde{X}(\infty) \le -\zeta - \delta$, and $\tilde{X}(\infty) \ge -\zeta$. We recall that

$$b(x) = \mu\big[(x+\zeta)^- - \zeta^-\big] = \begin{cases} -\mu x, & x \le -\zeta, \\ \mu\zeta, & x \ge -\zeta, \end{cases} \tag{4.2}$$

and apply the moment bounds (3.12), (3.13), and the gradient bound (3.18), to see that

$$\begin{aligned}
\mathbb{E}\Big[\big|f_h''(\tilde{X}(\infty))b(\tilde{X}(\infty))\big|\Big] &\le 32(1+1/|\zeta|)\mathbb{E}\Big[\big|\tilde{X}(\infty)\big|1(\tilde{X}(\infty) \le -\zeta - \delta)\Big] \\
&\quad + \mathbb{P}(\tilde{X}(\infty) \ge -\zeta) \\
&\le 32(1+1/|\zeta|)\left(2|\zeta| \wedge \sqrt{\frac{4}{3} + \frac{2\delta^2}{3}}\right) + 1
\end{aligned}$$

$$\leq 32\Big(\sqrt{\frac{4}{3}+\frac{2\delta^2}{3}}+2\Big)+1$$
$$\leq 32\big(\sqrt{2}+2\big)+1 \leq 111.$$

Next, we use (3.15) and the gradient bound in (3.19) to get

$$\frac{\mu}{2}\mathbb{E}\bigg[\int_{\tilde{X}(\infty)}^{\tilde{X}(\infty)+\delta}\big|f_h'''(y)\big|\,dy\bigg]$$
$$\leq \frac{\delta}{2}\Big((23+13/|\zeta|)\mathbb{P}(\tilde{X}(\infty)\leq -\zeta-\delta)+2\mathbb{P}(\tilde{X}(\infty)\geq -\zeta)\Big)$$
$$\leq \frac{\delta}{2}\Big(23+\frac{13}{|\zeta|}(3|\zeta|)\Big) \leq 31\delta.$$

By a similar argument, we can show that

$$\frac{\mu}{2}\mathbb{E}\bigg[\int_{\tilde{X}(\infty)-\delta}^{\tilde{X}(\infty)}\big|f_h'''(y)\big|\,dy\bigg] \leq 31\delta,$$

with the only difference in the argument being that we consider the cases when $\tilde{X}(\infty)\leq -\zeta$ and $\tilde{X}(\infty)\geq -\zeta+\delta$, instead of $\tilde{X}(\infty)\leq -\zeta-\delta$ and $\tilde{X}(\infty)\geq -\zeta$. Lastly, we use the form of $b(x)$, the moment bounds (3.12), (3.13), and (3.16), and the gradient bound (3.19) to get

$$\frac{\delta}{2}\mathbb{E}\bigg[\big|b(\tilde{X}(\infty))\big|\int_{\tilde{X}(\infty)-\delta}^{\tilde{X}(\infty)}\big|f_h'''(y)\big|\,dy\bigg]$$
$$\leq \frac{\delta^2}{2}\Big((23+13/|\zeta|)\mathbb{E}\Big[\big|\tilde{X}(\infty)\big|1(\tilde{X}(\infty)\leq -\zeta)\Big]+2|\zeta|\,\mathbb{P}(\tilde{X}(\infty)\geq -\zeta+\delta)\Big)$$
$$\leq \frac{\delta^2}{2}\Big((23+13/|\zeta|)\Big(2|\zeta|\wedge\sqrt{\frac{4}{3}+\frac{2\delta^2}{3}}\Big)+14/4\Big)$$
$$\leq \frac{\delta^2}{2}\bigg(23\sqrt{2}+26+14/4\bigg) \leq 32\delta^2.$$

Hence, from (3.10) we conclude that for all $R\geq 1$, and $h(x)\in \mathrm{Lip}(1)$,

$$\big|\mathbb{E}h(\tilde{X}(\infty))-\mathbb{E}h(Y(\infty))\big| \leq \delta(111+31+31+32\delta)\leq 205\delta, \tag{4.3}$$

which proves Theorem 1. □

5. The Kolmogorov metric. In this section we prove Theorem 3, which is stated in the Kolmogorov setting. The biggest difference between

the Wasserstein and Kolmogorov settings is that in the latter, the test functions $h(x)$ used in the Poisson equation (3.3) are discontinuous. For this reason, the gradient bounds from Lemma 3 and Lemma 13 in Appendix B.2 do not hold anymore, and new gradient bounds need to be derived separately for the Kolmogorov setting; we present these new gradient bounds in Section 5.1. Furthermore, the solution to the Poisson equation no longer has a continuous second derivative, meaning that the Taylor expansion we used to derive the upper bound in (3.10) is invalid. We discuss an alternative to (3.10) in Section 5.2. This alternative bound contains a new error term that cannot be handled by the gradient bounds, nor the moment bounds. This term appears because the solution to the Poisson equation has a discontinuous second derivative, and to bound it we present Lemma 6. We then prove Theorem 3 in Section 5.3.

5.1. *Kolmogorov gradient bounds.* Recall that in the Kolmogorov setting, we take the class of test functions for the Poisson equation (3.3) to be $\mathcal{H}_K$ defined in (1.8). For the statement of the following two lemmas, we fix $a \in \mathbb{R}$ and set $h(x) = 1_{(-\infty,a]}(x)$. We use $f_a(x)$ instead of $f_h(x)$ to denote a solution to the Poisson equation. We recall that the family of solutions to the Poisson equation is parametrized by constants $a_1, a_2 \in \mathbb{R}$. The following lemmas state the gradient bounds in the Kolmogorov setting.

LEMMA 4. *Consider the Erlang-C model ($\alpha = 0$). Any solution to the Poisson equation $f_a(x)$ is continuously differentiable, with an absolutely continuous derivative. Fix a solution in (3.4) with parameter $a_2 = 0$. Then for all $n \geq 1, \lambda > 0$, and $\mu > 0$ satisfying $0 < R < n$,*

$$|f_a'(x)| \leq \begin{cases} 5/\mu, & x \leq -\zeta, \\ \frac{1}{\mu|\zeta|}, & x \geq -\zeta, \end{cases} \tag{5.1}$$

and for all $x \in \mathbb{R}$,

$$|f_a''(x)| \leq 3/\mu, \tag{5.2}$$

where $f_a''(x)$ is understood to be the left derivative at the point $x = a$.

LEMMA 5. *Consider the Erlang-A model ($\alpha > 0$). Any solution to the Poisson equation $f_a(x)$ is continuously differentiable, with an absolutely continuous derivative. Fix a solution in (3.4) with parameter $a_2 = 0$, and fix $n \geq 1, \lambda > 0, \mu > 0$, and $\alpha > 0$. If $0 < R \leq n$ (an underloaded system), then*

$$|f_a'(x)| \leq \begin{cases} \frac{1}{\mu}\sqrt{2\pi}e^{1/2}, & x \leq -\zeta, \\ \frac{1}{\mu}\left(\sqrt{\frac{\pi}{2}\frac{\mu}{\alpha}} \wedge \frac{1}{|\zeta|}\right), & x \geq -\zeta, \end{cases} \tag{5.3}$$

and if $n \le R$ (an overloaded system), then

$$|f_a'(x)| \le \begin{cases} \frac{1}{\mu}\sqrt{\frac{\pi}{2}}, & x \le -\zeta, \\ \frac{1}{\mu}\sqrt{\frac{\pi}{2}}\left(1+\sqrt{\frac{\mu}{\alpha}}\right), & x \ge -\zeta. \end{cases} \tag{5.4}$$

Moreover, for all $\lambda > 0, n \ge 1, \mu > 0$, and $\alpha > 0$, and all $x \in \mathbb{R}$,

$$|f_a''(x)| \le 3/\mu, \tag{5.5}$$

where $f_a''(x)$ is understood to be the left derivative at the point $x = a$.

Lemmas 4 and 5 are proved in Appendix B.3. Unlike the Wasserstein setting, these lemmas do not guarantee that $f_a''(x)$ is absolutely continuous. Indeed, for any $a \in \mathbb{R}$, substituting $h(x) = 1_{(-\infty,a]}(x)$ into (3.3) gives us

$$\mu f_a''(x) = \mathbb{P}(Y(\infty) \le a) - 1_{(-\infty,a]}(x) - b(x)f_a'(x).$$

Since $b(x)f_a'(x)$ is a continuous function, the above equation implies that $f_a''(x)$ is discontinuous at the point $x = a$. Thus, we can no longer use the error bound in (3.10), and require a different expansion of $G_{\tilde{X}}f_a(x)$.

5.2. *Alternative Taylor expansion.* To get an error bound similar to (3.10), we first define

$$\epsilon_1(x) = \int_x^{x+\delta} (x+\delta-y)(f_a''(y) - f_a''(x-))dy, \tag{5.6}$$

$$\epsilon_2(x) = \int_{x-\delta}^{x} (y-(x-\delta))(f_a''(y) - f_a''(x-))dy. \tag{5.7}$$

Now observe that

$$\begin{aligned} f_a(x+\delta) - f_a(x) &= f_a'(x)\delta + \int_x^{x+\delta} (x+\delta-y)f_a''(y)dy \\ &= f_a'(x)\delta + \frac{1}{2}\delta^2 f_a''(x-) \\ &\quad + \int_x^{x+\delta} (x+\delta-y)(f_a''(y) - f_a''(x-))dy \\ &= f_a'(x)\delta + \frac{1}{2}\delta^2 f_a''(x-) + \epsilon_1(x), \end{aligned}$$

and

$$(f_a(x-\delta) - f_a(x)) = -f_a'(x)\delta + \int_{x-\delta}^{x} (y-(x-\delta))f_a''(y)dy$$

$$
\begin{aligned}
&= -f_a'(x)\delta + \frac{1}{2}\delta^2 f_a''(x-) \\
&\quad + \int_{x-\delta}^{x} (y-(x-\delta))(f_a''(y) - f_a''(x-))dy \\
&= -f_a'(x)\delta + \frac{1}{2}\delta^2 f_a''(x-) + \epsilon_2(x).
\end{aligned}
$$

For $k \in \mathbb{Z}_+$ and $x = x_k = \delta(k - x(\infty))$, we recall the forms of $G_Y f_a(x)$ and $G_{\tilde{X}} f_a(x)$ from (3.2) and (3.6) to see that

$$
\begin{aligned}
G_{\tilde{X}} f_a(x) &= \lambda\delta f_a'(x) + \lambda\frac{1}{2}\delta^2 f_a''(x-) + \lambda\epsilon_1(x) \\
&\quad - d(k)\delta f_a'(x) + d(k)\frac{1}{2}\delta^2 f_a''(x-) + d(k)\epsilon_2(x) \\
&= b(x) f_a'(x) + \lambda\frac{1}{2}\delta^2 f_a''(x-) + \lambda\epsilon_1(x) \\
&\quad + (\lambda - \frac{1}{\delta}b(x))\frac{1}{2}\delta^2 f_a''(x-) + (\lambda - \frac{1}{\delta}b(x))\epsilon_2(x) \\
&= G_Y f(x) - b(x)\frac{1}{2}\delta f_a''(x-) + \lambda(\epsilon_1(x) + \epsilon_2(x)) - \frac{1}{\delta}b(x)\epsilon_2(x),
\end{aligned}
$$

where in the second equality we used the fact that $b(x) = \delta(\lambda - d(k))$, and in the last equality we use that $\delta^2\lambda = \mu$. Combining this with (3.9), we have an error bound similar to (3.10):

$$
\begin{aligned}
&\left|\mathbb{P}(\tilde{X}(\infty) \le a) - \mathbb{P}(Y(\infty) \le a)\right| \\
&\le \frac{1}{2}\delta\mathbb{E}\left[\left|f_a''(\tilde{X}(\infty)-)b(\tilde{X}(\infty))\right|\right] + \lambda\mathbb{E}\left[\left|\epsilon_1(\tilde{X}(\infty))\right|\right] \\
&\quad + \lambda\mathbb{E}\left[\left|\epsilon_2(\tilde{X}(\infty))\right|\right] + \frac{1}{\delta}\mathbb{E}\left[\left|b(\tilde{X}(\infty))\epsilon_2(\tilde{X}(\infty))\right|\right], \qquad (5.8)
\end{aligned}
$$

where $\epsilon_1(x)$ and $\epsilon_2(x)$ are as in (5.6) and (5.7). To bound the error terms in (5.8) that are associated with $\epsilon_1(x)$ and $\epsilon_2(x)$, we need to analyze the difference $f_a''(y) - f_a''(x-)$ for $|x-y| \le \delta$. Since $f_a(x)$ is a solution to the Poisson equation (3.3), we see that for any $x, y \in \mathbb{R}$ with $y \ne a$,

$$
f_a''(y) - f_a''(x-) = \frac{1}{\mu}\left[1_{(-\infty,a]}(x) - 1_{(-\infty,a]}(y) + b(x)f_a'(x) - b(y)f_a'(y)\right].
$$

Therefore, for any $y \in [x, x+\delta]$ with $y \ne a$,

$$
\begin{aligned}
&\left|f_a''(y) - f_a''(x-)\right| \\
&\le \frac{1}{\mu}\left[1_{(a-\delta,a]}(x) + |b(x)|\left|f_a'(x) - f_a'(y)\right| + |b(x) - b(y))|\left|f_a'(y)\right|\right]
\end{aligned}
$$

$$\leq \frac{1}{\mu}\big[1_{(a-\delta,a]}(x) + \delta\,|b(x)|\,\|f''\| + |b(x)-b(y))|\,\big|f_a'(y)\big|\,\big], \tag{5.9}$$

and likewise, for any $y \in [x-\delta, x]$ with $y \neq a$,

$$\begin{aligned}
&\big|f_a''(y) - f_a''(x-)\big| \\
&\leq \frac{1}{\mu}\big[1_{(a,a+\delta]}(x) + |b(x)|\,\big|f_a'(x) - f_a'(y)\big| + |b(x)-b(y))|\,\big|f_a'(y)\big|\,\big] \\
&\leq \frac{1}{\mu}\big[1_{(a,a+\delta]}(x) + \delta\,|b(x)|\,\|f''\| + |b(x)-b(y))|\,\big|f_a'(y)\big|\,\big]. \qquad (5.10)
\end{aligned}$$

The inequalities above contain the indicators $1_{(a-\delta,a]}(x)$ and $1_{(a,a+\delta]}(x)$. When we consider the upper bound in (5.8), these indicators will manifest themselves as probabilities $\mathbb{P}(a-\delta < \tilde{X}(\infty) \leq a)$ and $\mathbb{P}(a < \tilde{X}(\infty) \leq a+\delta)$. To this end we present the following lemma, which will be used in the proof of Theorem 3.

LEMMA 6. *Consider the Erlang-C model ($\alpha = 0$). Let W be an arbitrary random variable with cumulative distribution function $F_W : \mathbb{R} \to [0,1]$. Let $\omega(F_W)$ be the modulus of continuity of F_W, defined as*

$$\omega(F_W) = \sup_{\substack{x,y\in\mathbb{R}\\ x\neq y}} \frac{|F_W(x) - F_W(y)|}{|x-y|}.$$

Recall that $d_K(\tilde{X}(\infty), W)$ is the Kolmogorov distance between $X(\infty)$ and W. Then for any $a \in \mathbb{R}$, $n \geq 1$, and $0 < R < n$,

$$\mathbb{P}(a-\delta < \tilde{X}(\infty) \leq a+\delta) \leq \omega(F_W)2\delta + d_K(\tilde{X}(\infty), W) + 9\delta^2 + 8\delta^4.$$

This lemma is proved in Appendix D.2. We will apply Lemma 6 with $W = Y(\infty)$ in the proof of Theorem 3 that follows. The following lemma guarantees that the modulus of continuity of the cumulative distribution function of $Y(\infty)$ is bounded by a constant independent of λ, n, and μ. Its proof is provided in Appendix D.3.

LEMMA 7. *Consider the Erlang-C model ($\alpha = 0$), and let $\nu(x)$ be the density of $Y(\infty)$. Then for for all $n \geq 1, \lambda > 0$, and $\mu > 0$ satisfying $0 < R < n$,*

$$|\nu(x)| \leq \sqrt{\frac{2}{\pi}}, \quad x \in \mathbb{R}.$$

Lemmas 6 and 7 are stated for the Erlang-C model, but one can easily repeat the arguments in the proofs of those lemmas to prove analogues for the Erlang-A model. Therefore, we state the following lemmas without proof.

LEMMA 8. *Consider the Erlang-A model ($\alpha > 0$). Let W be an arbitrary random variable with cumulative distribution function $F_W : \mathbb{R} \to [0,1]$. Let $\omega(F_W)$ be the modulus of continuity of F_W. Then for any $a \in \mathbb{R}$, $\alpha > 0$, $n \geq 1$, and $R > 0$,*

$$\begin{aligned}&\mathbb{P}(a - \delta < \tilde{X}(\infty) \leq a + \delta)\\ &\leq \omega(F_W)2\delta + d_K(\tilde{X}(\infty), W) + 9\Big(\frac{\alpha}{\mu} \vee 1\Big)\delta^2 + 8\Big(\frac{\alpha}{\mu} \vee 1\Big)^2\delta^4.\end{aligned}$$

LEMMA 9. *Consider the Erlang-A model ($\alpha > 0$), and let $\nu(x)$ be the density of $Y(\infty)$. Fix $n \geq 1, \lambda > 0, \mu > 0$, and $\alpha > 0$. If $0 < R \leq n$, then*

$$|\nu(x)| \leq \sqrt{\frac{2}{\pi}}, \quad x \in \mathbb{R},$$

and if $n \leq R$, then

$$|\nu(x)| \leq \sqrt{\frac{2}{\pi}}\sqrt{\frac{\alpha}{\mu}}, \quad x \in \mathbb{R}.$$

5.3. *Proof of Theorem 3.*

PROOF OF THEOREM 3. Throughout the proof we assume that $R \geq 1$, or equivalently, $\delta \leq 1$. For $h(x) = 1_{(-\infty,a]}(x)$, we let $f_a(x)$ be a solution the Poisson equation (3.3) with parameter $a_2 = 0$. In this proof we will show that for all $a \in \mathbb{R}$,

(5.11)
$$\Big|\mathbb{P}(\tilde{X}(\infty) \leq a) - \mathbb{P}(Y(\infty) \leq a)\Big| \leq \frac{1}{2}\mathbb{P}(a - \delta < \tilde{X}(\infty) \leq a + \delta) + 75\delta,$$

The upper bound in (5.11) is similar to (4.3), however (5.11) has the extra term

$$\frac{1}{2}\mathbb{P}(a - \delta < \tilde{X}(\infty) \leq a + \delta). \tag{5.12}$$

The reason this term appears in the Kolmogorov setting but not in the Wasserstein setting is because $f_a''(x)$ is discontinuous in the Kolmogorov case, as opposed to the Wasserstein case where $f_h''(x)$ is continuous. Applying

Lemmas 6 and 7 to the right hand side of (5.11), and taking the supremum over all $a \in \mathbb{R}$ on both sides, we see that

$$d_K(\tilde{X}(\infty), Y(\infty)) \leq \frac{1}{2} d_K(\tilde{X}(\infty), Y(\infty)) + 2\sqrt{\frac{2}{\pi}}\delta + 9\delta^2 + 8\delta^4 + 75\delta,$$

or

$$d_K(\tilde{X}(\infty), Y(\infty)) \leq 188\delta.$$

We want to add that Lemma 6 makes heavy use of the birth-death structure of the Erlang-C model, and that it is not obvious how to handle (5.12) more generally.

To prove Theorem 3 it remains to verify (5.11), which we now do. The argument we will use is similar to the argument used to prove (4.3) in Theorem 1. We will bound each of the terms in (5.8), which we recall here as

$$\begin{aligned}
&\Big|\mathbb{P}(\tilde{X}(\infty) \leq a) - \mathbb{P}(Y(\infty) \leq a)\Big| \\
&\leq \frac{1}{2}\delta\mathbb{E}\Big[\big|f_a''(\tilde{X}(\infty)-)b(\tilde{X}(\infty))\big|\Big] + \lambda\mathbb{E}\Big[\big|\epsilon_1(\tilde{X}(\infty))\big|\Big] \\
&\quad + \lambda\mathbb{E}\Big[\big|\epsilon_2(\tilde{X}(\infty))\big|\Big] + \frac{1}{\delta}\mathbb{E}\Big[\big|b(\tilde{X}(\infty))\epsilon_2(\tilde{X}(\infty))\big|\Big].
\end{aligned}$$

We also recall the form of $b(x)$ from (4.2). We use the moment bounds (3.12) and (3.16), and the gradient bound (5.2) to see that

$$\begin{aligned}
&\mathbb{E}\Big[\big|f_a''(\tilde{X}(\infty)-)b(\tilde{X}(\infty))\big|\Big] \\
&\leq \frac{3}{\mu}\mathbb{E}\Big[\big|b(\tilde{X}(\infty))\big|\Big] \\
&= 3\mathbb{E}\Big[\big|\tilde{X}(\infty)1(\tilde{X}(\infty) \leq -\zeta - \delta)\big|\Big] + 3\,|\zeta|\,\mathbb{P}(\tilde{X}(\infty) \geq -\zeta) \\
&\leq 3\sqrt{\frac{4}{3} + \frac{2\delta^2}{3}} + 3\Big(\,|\zeta| \wedge \mathbb{E}\Big[\big|\tilde{X}(\infty)\big|1(\tilde{X}(\infty) \geq -\zeta)\Big]\Big) \\
&\leq 3\sqrt{2} + \frac{21}{4} \leq 10. \qquad (5.13)
\end{aligned}$$

Next, we use (5.9), (5.13), and the gradient bound (5.1) to get

$$\begin{aligned}
&\lambda\mathbb{E}\Big[\big|\epsilon_1(\tilde{X}(\infty))\big|\Big] \\
&= \lambda\mathbb{E}\Big[\int_{\tilde{X}(\infty)}^{\tilde{X}(\infty)+\delta} (\tilde{X}(\infty) + \delta - y)\big|f_a''(y) - f_a''(\tilde{X}(\infty)-)\big|dy\Big]
\end{aligned}$$

$$
\begin{aligned}
&\le \frac{\lambda}{\mu}\mathbb{E}\Big[1_{(a-\delta,a]}(\tilde{X}(\infty))\int_{\tilde{X}(\infty)}^{\tilde{X}(\infty)+\delta}(\tilde{X}(\infty)+\delta-y)dy\Big]\\
&\quad+\frac{\lambda}{\mu}\delta^3\mathbb{E}\Big[\big|b(\tilde{X}(\infty))\big|\Big]\|f_a''\|+\frac{\lambda}{\mu}\delta\mathbb{E}\Big[\int_{\tilde{X}(\infty)}^{\tilde{X}(\infty)+\delta}\big|b(\tilde{X}(\infty))-b(y)\big|\big|f_a'(y)\big|dy\Big]\\
&\le \frac{1}{2}\mathbb{P}(a-\delta<\tilde{X}(\infty)\le a)+10\delta+5\delta\\
&\equiv \frac{1}{2}\mathbb{P}(a-\delta<\tilde{X}(\infty)\le a)+15\delta,
\end{aligned}
$$

where in the last inequality we used the fact that for $y\in[\tilde{X}(\infty),\tilde{X}(\infty)+\delta]$,

$$b(\tilde{X}(\infty))-b(y)=\mu\delta 1(\tilde{X}(\infty)\le-\zeta-\delta).$$

By a similar argument, one can check that

$$\lambda\mathbb{E}\Big[\big|\epsilon_2(\tilde{X}(\infty))\big|\Big]\le\frac{1}{2}\mathbb{P}(a<\tilde{X}(\infty)\le a+\delta)+15\delta,$$

with the only difference in the argument being that we consider the cases when $\tilde{X}(\infty)\le-\zeta$ and $\tilde{X}(\infty)\ge-\zeta+\delta$, instead of $\tilde{X}(\infty)\le-\zeta-\delta$ and $\tilde{X}(\infty)\ge-\zeta$. Lastly, we use the first inequality in (5.10) to see that

$$
\begin{aligned}
&\frac{1}{\delta}\mathbb{E}\Big[\big|b(\tilde{X}(\infty))\epsilon_2(\tilde{X}(\infty))\big|\Big]\\
&\le\frac{1}{\mu}\mathbb{E}\Big[\big|b(\tilde{X}(\infty))\big|\int_{\tilde{X}(\infty)-\delta}^{\tilde{X}(\infty)}\Big[1_{(a,a+\delta]}(\tilde{X}(\infty))\\
&\qquad\qquad+\big|b(\tilde{X}(\infty))\big|\Big(\big|f_a'(\tilde{X}(\infty))\big|+\big|f_a'(y)\big|\Big)\\
&\qquad\qquad+\big|b(\tilde{X}(\infty))-b(y)\big|\big|f_a'(y)\big|\Big]dy\Big]\\
&\le\delta\frac{1}{\mu}\mathbb{E}\Big[\big|b(\tilde{X}(\infty))\big|\Big]+\delta\frac{1}{\mu}\mathbb{E}\Big[\big|b^2(\tilde{X}(\infty))f_a'(\tilde{X}(\infty))\big|\Big]\\
&\quad+\frac{1}{\mu}\mathbb{E}\Big[\big|b^2(\tilde{X}(\infty))\big|\int_{\tilde{X}(\infty)-\delta}^{\tilde{X}(\infty)}\big|f_a'(y)\big|\,dy\Big]+5\delta^2\mathbb{E}\Big[\big|\tilde{X}(\infty)1(\tilde{X}(\infty)\le-\zeta)\big|\Big]\\
&\le\frac{10}{3}\delta+\delta\frac{1}{\mu}\mathbb{E}\Big[\big|b^2(\tilde{X}(\infty))f_a'(\tilde{X}(\infty))\big|\Big]\\
&\quad+\frac{1}{\mu}\mathbb{E}\Big[\big|b^2(\tilde{X}(\infty))\big|\int_{\tilde{X}(\infty)-\delta}^{\tilde{X}(\infty)}\big|f_a'(y)\big|\,dy\Big]+5\sqrt{2}\delta^2,
\end{aligned}
$$

where in the last inequality we used (5.13) and the moment bound (3.12). Now by (3.11) and (3.16),

$$\delta\frac{1}{\mu}\mathbb{E}\Big[\big|b^2(\tilde{X}(\infty))f_a'(\tilde{X}(\infty))\big|\Big]$$

$$\leq 5\delta\mathbb{E}\big[\tilde{X}^2(\infty)1(\tilde{X}(\infty)\leq -\zeta)\big] + \delta\,|\zeta|\,\mathbb{P}(\tilde{X}(\infty)\geq -\zeta+\delta)$$
$$\leq 10\delta + \delta\frac{7}{4} \leq 12\delta,$$

and similarly,

$$\frac{1}{\mu}\mathbb{E}\left[\left|b^2(\tilde{X}(\infty))\right|\int_{\tilde{X}(\infty)-\delta}^{\tilde{X}(\infty)}\left|f_a'(y)\right|dy\right] \leq 12\delta.$$

Therefore,

$$\frac{1}{\delta}\mathbb{E}\Big[\left|b(\tilde{X}(\infty))\epsilon_2(\tilde{X}(\infty))\right|\Big] \leq \frac{10}{3}\delta + 24\delta + 5\sqrt{2}\delta^2 \leq 35\delta.$$

This verifies (5.11) and concludes the proof of Theorem 3. □

6. Extension: Erlang-C higher moments. In this section we consider the approximation of higher moments for the Erlang-C model. We begin with the following result.

THEOREM 5. *Consider the Erlang-C system ($\alpha = 0$), and fix an integer $m > 0$. There exists a constant $C = C(m)$, such that for all $n \geq 1, \lambda > 0$, and $\mu > 0$ satisfying $1 \leq R < n$,*

$$\left|\mathbb{E}(\tilde{X}(\infty))^m - \mathbb{E}(Y(\infty))^m\right| \leq (1 + 1/\left|\zeta\right|^{m-1})C(m)\delta, \tag{6.1}$$

where ζ is defined in (1.4).

The proof of this theorem follows the standard Stein framework in Section 3, but we do not provide it in this paper. The most interesting aspect of (6.1) is the appearance of $1/\left|\zeta\right|^{m-1}$ in the bound on the right hand side, which of course only matters when $|\zeta|$ is small. To check whether the bound is sharp, we performed some numerical experiments illustrated in Table 3. The results suggest that the approximation error does indeed grow like $1/\left|\zeta\right|^{m-1}$.

A better way to understand the growth parameter $1/\left|\zeta\right|^{m-1}$ is through its relationship with $\mathbb{E}(\tilde{X}(\infty))^{m-1}$. We claim that $\mathbb{E}(\tilde{X}(\infty))^{m-1} \approx 1/\left|\zeta\right|^{m-1}$ for small values of $|\zeta|$. The following lemma, which is proved in Appendix D.4, is needed.

LEMMA 10. *For any integer $m \geq 1$, and all $n \geq 1, \lambda > 0$, and $\mu > 0$ satisfying $R < n$,*

$$\lim_{\zeta\uparrow 0}\left|\zeta\right|^m \mathbb{E}(Y(\infty))^m = m!. \tag{6.2}$$

Multiplying both sides of (6.1) by $|\zeta|^m$ and applying Lemma 10, we see that for all $n \geq 1, \lambda > 0$, and $\mu > 0$ satisfying $1 \leq R < n$,

$$\lim_{\zeta \uparrow 0} |\zeta|^m \, \mathbb{E}(\tilde{X}(\infty))^m = m!.$$

In other words, we can rewrite (6.1) as

$$\begin{aligned} &\left|\mathbb{E}(\tilde{X}(\infty))^m - \mathbb{E}(Y(\infty))^m\right| \\ &\leq \left(1 + \frac{1}{|\zeta|^{m-1}\left|\mathbb{E}(\tilde{X}(\infty))^{m-1}\right|}\left|\mathbb{E}(\tilde{X}(\infty))^{m-1}\right|\right) C(m)\delta \\ &\leq \left(1 + \left|\mathbb{E}(\tilde{X}(\infty))^{m-1}\right|\right)\tilde{C}(m)\delta, \end{aligned}$$

where $\tilde{C}(m)$ is a redefined version of $C(m)$. That the approximation error in Table 3 increases is then attributed to the fact that $\mathbb{E}\tilde{X}(\infty)$ increases as $\zeta \uparrow 0$. As we mentioned before, the appearance of the $(m-1)$th moment in the approximation error of the mth moment was also observed recently in [30] for the virtual waiting time in the $M/GI/1+GI$ model, potentially suggesting a general trend.

TABLE 3

The error term above equals $\left|\mathbb{E}(\tilde{X}(\infty))^2 - \mathbb{E}(Y(\infty))^2\right|$ and grows as $R \to n$. The error term still grows when multiplied by $|\zeta|^{0.5}$, and the error term shrinks to zero when multiplied by $|\zeta|^{1.5}$. However, when multiplied by $|\zeta|$, the error term appears to converge to some limiting value, suggesting that the error does indeed grow at a rate of $1/|\zeta|$. We observed consistent behavior for higher moments of $\tilde{X}(\infty)$ as well.

R	$\lvert\zeta\rvert$	$\mathbb{E}(\tilde{X}(\infty))^2$	Error	$\lvert\zeta\rvert \times$Error	$\lvert\zeta\rvert^{0.5} \times$Error	$\lvert\zeta\rvert^{1.5} \times$Error
499	4.48×10^{-2}	9.47×10^{2}	1.59	7.10×10^{-2}	0.34	1.50×10^{-2}
499.9	4.50×10^{-3}	9.94×10^{4}	16.50	7.38×10^{-2}	1.10	4.94×10^{-3}
499.95	2.20×10^{-3}	3.99×10^{5}	33.08	7.40×10^{-2}	1.56	3.50×10^{-3}
499.99	4.47×10^{-4}	9.99×10^{6}	165.67	7.41×10^{-2}	3.50	1.57×10^{-3}

Acknowledgement. This research is supported in part by NSF Grants CNS-1248117, CMMI-1335724, and CMMI-1537795. The first two authors were stimulated from discussions with the participants of the 2015 Workshop on New Directions in Stein's Method held at the Institute for Mathematical Sciences at the National University of Singapore and they would like to thank the financial support from the Institute. The authors also wish to thank two anonymous referees for suggestions to improve the presentation of the paper.

APPENDICES

Appendix A handles the moment bounds, while Appendix B handles the gradient bounds. Appendix C contains outlines for the proofs of Theorems 2 and 4, and in Appendix D we prove several miscellaneous lemmas.

APPENDIX A: MOMENT BOUNDS

In Section A.1, we first prove Lemma 2, establishing the moment bounds for Erlang-C model. In Section A.2, we state and prove Lemma 11, establishing the moment bounds for Erlang-A model.

A.1. Erlang-C Moment Bounds.

PROOF OF LEMMA 2. We first prove (3.11), (3.12), and (3.14). Recalling the generator $G_{\tilde{X}}$ defined in (3.6), we apply it to the function $V(x) = x^2$ to see that for $k \in \mathbb{Z}_+$ and $x = x_k = \delta(k - x(\infty))$,

$$\begin{aligned} G_{\tilde{X}}V(x) &= \lambda(2x\delta + \delta^2) + \mu(k \wedge n)(-2x\delta + \delta^2) \\ &= 2x\delta(\lambda - n\mu + \mu(k-n)^-) + \mu + \delta^2\mu(k \wedge n) \\ &= 2x\mu(\zeta + (x+\zeta)^-) + \mu + \delta^2\mu(n - \frac{\lambda}{\mu} + \frac{\lambda}{\mu} - (k-n)^-) \\ &= 2x\mu(\zeta + (x+\zeta)^-) + \mu - \delta\mu\zeta + \mu - \delta\mu(x+\zeta)^- \\ &= 1(x \le -\zeta)\mu\big(-2x^2 + \delta x\big) + 1(x > -\zeta)\mu\big(2x\zeta - \delta\zeta\big) + 2\mu \\ &\le 1(x \le -\zeta)\mu\big(-\frac{3}{2}x^2 + \frac{\delta^2}{2}\big) + 1(x > -\zeta)\mu\big(2x\zeta - \delta\zeta\big) + 2\mu. \end{aligned} \tag{A.1}$$

Instead of splitting the last two lines into the cases $x \le -\zeta$ and $x > -\zeta$, we could have also considered $x < -\zeta$ and $x \ge -\zeta$ instead, and would have obtained

$$\begin{aligned} G_{\tilde{X}}V(x) &= 1(x < -\zeta)\mu\big(-2x^2 + \delta x\big) + 1(x \ge -\zeta)\mu\big(2x\zeta - \delta\zeta\big) + 2\mu \\ &\le 1(x < -\zeta)\mu\big(-\frac{3}{2}x^2 + \frac{\delta^2}{2}\big) + 1(x \ge -\zeta)\mu\big(2x\zeta - \delta\zeta\big) + 2\mu. \end{aligned} \tag{A.2}$$

We take expected values on both sides of (A.1) with respect to $\tilde{X}(\infty)$, and apply Lemma 1 to see that

$$\begin{aligned} 0 \le &-\frac{3}{2}\mu\mathbb{E}\big[(\tilde{X}(\infty))^2 1(\tilde{X}(\infty) \le -\zeta)\big] \\ &+ \mu\,|\zeta|\,\mathbb{E}\big[\big(-2\tilde{X}(\infty) + \delta\big)1(\tilde{X}(\infty) > -\zeta)\big] + 2\mu + \frac{\mu\delta^2}{2}. \end{aligned} \tag{A.3}$$

This implies that when $|\zeta| > \delta/2$,

$$0 \le -\frac{3}{2}\mu\mathbb{E}\big[(\tilde{X}(\infty))^2 1(\tilde{X}(\infty) \le -\zeta)\big] + 2\mu + \frac{\mu\delta^2}{2},$$

and when $|\zeta| \le \delta/2$,

$$0 \le -\frac{3}{2}\mu\mathbb{E}\big[(\tilde{X}(\infty))^2 1(\tilde{X}(\infty) \le -\zeta)\big] + 2\mu + \mu\delta^2.$$

Therefore,

$$\mathbb{E}\big[(\tilde{X}(\infty))^2 1(\tilde{X}(\infty) \le -\zeta)\big] \le \frac{4}{3} + \frac{2\delta^2}{3},$$

which proves (3.11). Jensen's inequality immediately gives us

$$\mathbb{E}\Big[\big|\tilde{X}(\infty)1(\tilde{X}(\infty) \le -\zeta)\big|\Big] \le \sqrt{\mathbb{E}\big[(\tilde{X}(\infty))^2 1(\tilde{X}(\infty) \le -\zeta)\big]},$$

which proves (3.12). Furthermore, (A.3) also gives us

$$\mathbb{E}\Big[\big|\tilde{X}(\infty)1(\tilde{X}(\infty) > -\zeta)\big|\Big] \le \frac{1}{|\zeta|} + \frac{\delta^2}{4\,|\zeta|} + \frac{\delta}{2},$$

which is not quite (3.14) because the inequality above has $1(\tilde{X}(\infty) > -\zeta)$ as opposed to $1(\tilde{X}(\infty) \ge -\zeta)$ as in (3.14). However, we can use (A.2) to get the stronger bound

$$\mathbb{E}\Big[\big|\tilde{X}(\infty)1(\tilde{X}(\infty) \ge -\zeta)\big|\Big] \le \frac{1}{|\zeta|} + \frac{\delta^2}{4\,|\zeta|} + \frac{\delta}{2},$$

which proves (3.14).

We now prove (3.13), or

$$\mathbb{E}\Big[\big|\tilde{X}(\infty)1(\tilde{X}(\infty) \le -\zeta)\big|\Big] \le 2\,|\zeta|\,. \tag{A.4}$$

We use the triangle inequality to see that

$$\mathbb{E}\Big[\big|\tilde{X}(\infty)1(\tilde{X}(\infty) \le -\zeta)\big|\Big] \le\ |\zeta| + \mathbb{E}\Big[\big|\tilde{X}(\infty) + \zeta\big|1(\tilde{X}(\infty) \le -\zeta)\Big].$$

The second term on the right hand side is just the expected number of idle servers, scaled by δ. We now show that this expected value equals $|\zeta|$. Applying the generator $G_{\tilde{X}}$ to the test function $f(x) = x$, one sees that for all $k \in \mathbb{Z}_+$ and $x = x_k = \delta(k - x(\infty))$,

$$G_{\tilde{X}} f(x) = \delta\lambda - \delta\mu(k \wedge n) = \mu\big[\zeta + (x + \zeta)^-\big].$$

Taking expected values with respect to $\tilde{X}(\infty)$ on both sides, and applying Lemma 1, we arrive at

$$\mathbb{E}\Big[\big|(\tilde{X}(\infty)+\zeta)1(\tilde{X}(\infty)\le -\zeta)\big|\Big] = |\zeta|\,, \tag{A.5}$$

which proves (3.13).

We move on to prove (3.15), or

$$\mathbb{P}(\tilde{X}(\infty)\le -\zeta) \le (2+\delta)\,|\zeta|\,. \tag{A.6}$$

Let I be the unscaled expected number of idle servers. Then by (A.5),

$$I = \mathbb{E}(X(\infty)-n)^- = \frac{1}{\delta}\mathbb{E}\Big[\big|(\tilde{X}(\infty)+\zeta)1(\tilde{X}(\infty)\le -\zeta)\big|\Big] = \frac{1}{\delta}\,|\zeta|\,.$$

Now let $\{\nu_k\}_{k=0}^{\infty}$ be the stationary distribution of X (the unscaled CTMC). We want to prove an upper bound on the probability

$$\mathbb{P}(\tilde{X}(\infty)\le -\zeta) = \sum_{k=0}^{n}\nu_k \le \sum_{k=0}^{\lfloor n-\sqrt{R}\rfloor}\nu_k + \sum_{k=\lceil n-\sqrt{R}\rceil}^{n}\nu_k.$$

Observe that

$$I = \sum_{k=0}^{n}(n-k)\nu_k \ge \sqrt{R}\sum_{k=0}^{\lfloor n-\sqrt{R}\rfloor}\nu_k.$$

Now let k^* be the first index that maximizes $\{\nu_k\}_{k=0}^{\infty}$, i.e.

$$k^* = \inf\{k\ge 0 : \nu_k \ge \nu_j,\ \text{for all } j\ne k\}.$$

Then

$$\begin{aligned}\mathbb{P}(\tilde{X}(\infty)\le -\zeta) = \sum_{k=0}^{\lfloor n-\sqrt{R}\rfloor}\nu_k + \sum_{k=\lceil n-\sqrt{R}\rceil}^{n}\nu_k &\le \frac{I}{\sqrt{R}} + (\sqrt{R}+1)\nu_{k^*}\\ &= |\zeta| + (\sqrt{R}+1)\nu_{k^*}.\end{aligned} \tag{A.7}$$

Applying $G_{\tilde{X}}$ to the test function $f(x) = (k\wedge k^*)$, we see that for all $k\in\mathbb{Z}_+$ and $x = x_k = \delta(k-x(\infty))$,

$$G_{\tilde{X}}f(x) = \delta\lambda 1(k<k^*) - \delta\mu(k\wedge n)1(k\le k^*).$$

Taking expected values with respect to $X(\infty)$ on both sides and applying Lemma 1, we see that

$$\mathbb{P}(X(\infty)\le k^*)=\frac{\mu}{n\mu-\lambda}\mathbb{E}\big[(X(\infty)-n)^-1(X(\infty)\le k^*)\big]-\nu_{k^*}\frac{\lambda}{n\mu-\lambda}\ge 0.$$

Using the inequality above, together with the fact that $k^*\le n$, we see that

$$\begin{aligned}\nu_{k^*}&\le\frac{\mu}{\lambda}\mathbb{E}\big[(X(\infty)-n)^-1(X(\infty)\le k^*)\big]\\&\le\frac{\mu}{\lambda}\mathbb{E}\big[(X(\infty)-n)^-1(X(\infty)\le n)\big]=\frac{I}{R}=\frac{|\zeta|}{\sqrt{R}}.\end{aligned}$$

The fact that $k^*\le n$ is a consequence of $\lambda<n\mu$, and can be verified through the flow balance equations of the CTMC X. We combine the bound above with (A.7) to arrive at (3.15), which concludes the proof of this lemma. □

A.2. Erlang-A moment bounds. The following lemma states the necessary moment bounds for the Erlang-A model. The underloaded and overloaded cases have to be handled separately. Since the drift $b(x)$ is different between the Erlang-A and Erlang-C models, the quantities bounded in the following lemma will resemble those in Lemma 2, but will not be identical.

LEMMA 11. *Consider the Erlang-A model ($\alpha>0$). Fix $n\ge 1,\lambda>0$, $\mu>0$, and $\alpha>0$. If $0<R\le n$ (an underloaded system), then*

$$\mathbb{E}\Big[\big(\tilde X(\infty)\big)^2 1(\tilde X(\infty)\le-\zeta)\Big]\le\frac{1}{3}\Big(\frac{\alpha}{\mu}\delta^2+\delta^2+4\Big),\tag{A.8}$$

$$\mathbb{E}\Big[\big|\tilde X(\infty)1(\tilde X(\infty)\le-\zeta)\big|\Big]\le\sqrt{\frac{1}{3}\Big(\frac{\alpha}{\mu}\delta^2+\delta^2+4\Big)},\tag{A.9}$$

$$\mathbb{E}\Big[\big|\tilde X(\infty)1(\tilde X(\infty)\le-\zeta)\big|\Big]\le 2\,|\zeta|+\frac{\alpha}{\mu}\sqrt{\frac{1}{3}\Big(\frac{\mu}{\alpha}\delta^2+\frac{\mu}{\alpha}4+\delta^2\Big)},\tag{A.10}$$

$$\begin{aligned}&\mathbb{E}\Big[\big|\tilde X(\infty)1(\tilde X(\infty)\ge-\zeta)\big|\Big]\\&\le\left(1+\frac{\delta^2}{4}+\frac{\delta}{2}\sqrt{\frac{1}{3}\Big(\frac{\alpha}{\mu}\delta^2+\delta^2+4\Big)}\right)\Big(\frac{\mu}{\mu\wedge\alpha}\wedge\frac{1}{|\zeta|}\Big),\end{aligned}\tag{A.11}$$

$$\mathbb{E}\Big[(\tilde X(\infty)+\zeta)^2 1(\tilde X(\infty)\ge-\zeta)\Big]\le\frac{1}{3}\Big(\frac{\mu}{\alpha}\delta^2+\frac{\mu}{\alpha}4+\delta^2\Big),\tag{A.12}$$

$$\mathbb{E}\Big[(\tilde X(\infty)+\zeta)1(\tilde X(\infty)\ge-\zeta)\Big]\le\sqrt{\frac{1}{3}\Big(\frac{\mu}{\alpha}\delta^2+\frac{\mu}{\alpha}4+\delta^2\Big)},\tag{A.13}$$

$$\mathbb{E}\Big[(\tilde X(\infty)+\zeta)1(\tilde X(\infty)\ge-\zeta)\Big]\le\frac{1}{|\zeta|}\Big(\frac{\delta^2}{4}\frac{\alpha}{\mu}+\frac{\delta^2}{4}+1\Big),\tag{A.14}$$

$$\mathbb{P}(\tilde{X}(\infty) \le -\zeta) \le (2+\delta)\left(|\zeta| + \frac{\alpha}{\mu}\sqrt{\frac{1}{3}\left(\frac{\mu}{\alpha}\delta^2 + \frac{\mu}{\alpha}4 + \delta^2\right)}\right). \tag{A.15}$$

and if $n \le R$ (an overloaded system), then

$$\mathbb{E}\left[\left|\tilde{X}(\infty)1(\tilde{X}(\infty) \le -\zeta)\right|\right] \le \sqrt{\frac{1}{\alpha \wedge \mu}\left(\alpha\frac{\delta^2}{4} + \mu\right)}, \tag{A.16}$$

$$\mathbb{E}\left[\left|\tilde{X}(\infty)1(\tilde{X}(\infty) \le -\zeta)\right|\right] \le \frac{1}{|\zeta|}\left(\frac{\delta^2}{4} + \frac{\mu}{\alpha}\right), \tag{A.17}$$

$$\mathbb{E}\left[(\tilde{X}(\infty))^2 1(\tilde{X}(\infty) \ge -\zeta)\right] \le \frac{1}{3}\left(\delta^2 + 4\frac{\mu}{\alpha}\right), \tag{A.18}$$

$$\mathbb{E}\left[\left|\tilde{X}(\infty)1(\tilde{X}(\infty) \ge -\zeta)\right|\right] \le \sqrt{\frac{1}{3}\left(\delta^2 + 4\frac{\mu}{\alpha}\right)}, \tag{A.19}$$

$$\mathbb{E}\left[\left|(\tilde{X}(\infty) + \zeta)1(\tilde{X}(\infty) \le -\zeta)\right|\right] \le \frac{1}{|\zeta|}\left(\frac{\delta^2}{4} + 1\right), \tag{A.20}$$

$$\mathbb{E}\left[(\tilde{X}(\infty) + \zeta)^2 1(\tilde{X}(\infty) \le -\zeta)\right] \le \frac{\delta^2}{4}\frac{\alpha}{\mu} + 1, \tag{A.21}$$

$$\mathbb{E}\left[\left|(\tilde{X}(\infty) + \zeta)1(\tilde{X}(\infty) \le -\zeta)\right|\right] \le \sqrt{\frac{\delta^2}{4}\frac{\alpha}{\mu} + 1}, \tag{A.22}$$

$$\mathbb{E}\left[\left|(\tilde{X}(\infty) + \zeta)1(\tilde{X}(\infty) \le -\zeta)\right|\right] \le \frac{\alpha}{\mu}\sqrt{\frac{1}{3}\left(\delta^2 + 4\frac{\mu}{\alpha}\right)}, \tag{A.23}$$

$$\mathbb{P}(\tilde{X}(\infty) \le -\zeta) \le (3+\delta)\frac{16}{\sqrt{2}}\left(\frac{\delta^2}{4} + 1\right)\left(\left(\frac{1}{\zeta} \vee \frac{\alpha}{\mu}\right) \wedge \sqrt{\frac{\alpha}{\mu}}\right). \tag{A.24}$$

A.2.1. *Proof outline for Lemma 11: The underloaded system.* The proof of the underloaded case of Lemma 11 is very similar to that of Lemma 2. Therefore, we only outline some key intermediate steps needed to obtain the results. We remind the reader that when $R \le n$, then $\zeta \le 0$. We first show how to establish (A.8), which is proved in a similar fashion to (3.11) of Lemma 2 – by applying the generator $G_{\tilde{X}}$ to the Lyapunov function $V(x) = x^2$. The following are some useful intermediate steps for any reader wishing to produce a complete proof. The first step to prove (A.8) is to get an analogue of (A.1). Namely, when $x \le -\zeta$,

$$G_{\tilde{X}}V(x) = -2\mu x^2 + \mu\delta x + 2\mu \le -\frac{3}{2}\mu x^2 + \mu\delta^2/2 + 2\mu,$$

and when $x \ge -\zeta$,

$$G_{\tilde{X}}V(x) = -2\alpha(x+\zeta)^2 + \alpha\delta(x+\zeta) - 2\mu|\zeta|(x+\zeta)$$

$$\begin{aligned} &-2\,|\zeta|\,\alpha(x+\zeta)+\mu\,|\zeta|\,(\delta-2\,|\zeta|)+2\mu \\ &\le -\frac{3}{2}\alpha(x+\zeta)^2-2\mu\,|\zeta|\,(x+\zeta)+\delta^2\alpha/2+\delta^2\mu/8+2\mu. \end{aligned} \tag{A.25}$$

From here, we use Lemma 1 to get a statement similar to (A.3), from which we can infer (A.8) and by applying Jensen's inequality to (A.8), we get (A.9). Observe that this procedure yields (A.12), (A.13), and (A.14) as well. We now describe how to prove (A.11), which requires only a slight modification of (A.25). Namely, for $x \ge -\zeta$,

$$G_{\tilde{X}}V(x) = 2x\big(-\alpha(x+\zeta)+\mu\zeta\big)-\delta\big(-\alpha(x+\zeta)+\mu\zeta\big)+2\mu.$$

From this, we can deduce that since $x \ge -\zeta$,

$$G_{\tilde{X}}V(x) \le -2(\mu\wedge\alpha)x^2-\delta\big(-\alpha(x+\zeta)+\mu\zeta\big)+2\mu,$$

and also

$$G_{\tilde{X}}V(x) \le -2\mu\,|\zeta|\,x-\delta\big(-\alpha(x+\zeta)+\mu\zeta\big)+2\mu.$$

Then Lemma 1 can be applied as before to see that both

$$2\mu\,|\zeta|\,\mathbb{E}\Big[\big|\tilde{X}(\infty)1(\tilde{X}(\infty)\ge-\zeta)\big|\Big] \text{ and } 2(\mu\wedge\alpha)\mathbb{E}\Big[\big(\tilde{X}(\infty)\big)^2 1(\tilde{X}(\infty)\ge-\zeta)\Big] \tag{A.26}$$

are bounded by

$$2\mu+\mu\delta^2/2-\delta\mathbb{E}\Big[\big(-\alpha(\tilde{X}(\infty)+\zeta)+\mu\zeta\big)1(\tilde{X}(\infty)\ge-\zeta)\Big].$$

Applying the generator $G_{\tilde{X}}$ to the test function $f(x)=x$ and taking expected values with respect to $\tilde{X}(\infty)$, we get $\mathbb{E}b(\tilde{X}(\infty))=0$, or

$$\mathbb{E}\Big[\big(-\alpha(\tilde{X}(\infty)+\zeta)+\mu\zeta\big)1(\tilde{X}(\infty)\ge-\zeta)\Big] = \mu\mathbb{E}\Big[\tilde{X}(\infty)1(\tilde{X}(\infty)<-\zeta)\Big]. \tag{A.27}$$

When combined with (A.9), this implies that

$$\begin{aligned} &2\mu+\mu\delta^2/2-\delta\mathbb{E}\Big[\big(-\alpha(\tilde{X}(\infty)+\zeta)+\mu\zeta\big)1(\tilde{X}(\infty)\ge-\zeta)\Big] \\ &\le 2\mu+\mu\delta^2/2+\mu\delta\sqrt{\frac{1}{3}\Big(\frac{\alpha}{\mu}\delta^2+\delta^2+4\Big)}, \end{aligned}$$

which proves (A.11), because the quantity above is an upper bound for (A.26). To prove (A.10), we manipulate (A.27) to get

$$\mathbb{E}\Big[\big|(\tilde{X}(\infty)+\zeta)1(\tilde{X}(\infty)\le -\zeta)\big|\Big] = |\zeta| + \frac{\alpha}{\mu}\mathbb{E}\Big[\big|(\tilde{X}(\infty)+\zeta)1(\tilde{X}(\infty)> -\zeta)\big|\Big],$$

to which we can apply the triangle inequality and (A.13) to conclude (A.10). Lastly, the proof of (A.15) is nearly identical to the proof of (3.15) in Lemma 2. The key step is to obtain an analogue of (A.7).

A.2.2. *Proof outline for Lemma 11: The overloaded system.* The proof of the overloaded case of Lemma 11 is also similar to that of Lemma 2. Therefore, we only outline some key intermediate steps needed to obtain the results; the bounds in this lemma are not proved in the order in which they are stated. We remind the reader that when $R \ge n$, then $\zeta \ge 0$. We start by proving (A.18). Although the left hand side of (A.18) is slightly different from (3.11) of Lemma 2, it is proved using the same approach – by applying the generator $G_{\tilde{X}}$ to the Lyapunov function $V(x) = x^2$. The following are some useful intermediate steps for any reader wishing to produce a complete proof. The first step to prove (A.18) is to get analogue of (A.1). Namely, when $x \le -\zeta$,

$$\begin{aligned} G_{\tilde{X}}V(x) = & -2\mu(x+\zeta)^2 + \mu\delta(x+\zeta) \\ & + 2(\mu+\alpha)\,|\zeta|\,(x+\zeta) - 2\alpha\zeta^2 - \alpha\delta\zeta + 2\mu \\ \le & -2\mu(x+\zeta)^2 + 2(\mu+\alpha)\,|\zeta|\,(x+\zeta) + 2\mu, \end{aligned} \tag{A.28}$$

and when $x \ge -\zeta$,

$$G_{\tilde{X}}V(x) = -2\alpha x^2 + \alpha\delta x + 2\mu \le -\frac{3}{2}\alpha x^2 + \alpha\delta^2/2 + 2\mu.$$

From here, we use Lemma 1 to get a statement similar to (A.3), which implies (A.18). Applying Jensen's inequality to (A.18) yields (A.19). The procedure used to get (A.18) also yields (A.20), (A.21), and (A.22).

We now describe how to prove (A.16) and (A.17), which requires only a slight modification of (A.28). Namely, we use the fact that for $x \le -\zeta$,

$$G_{\tilde{X}}V(x) = 2x\big(-\mu(x+\zeta)+\alpha\zeta\big) - \delta\big(-\mu(x+\zeta)+\alpha\zeta\big) + 2\mu.$$

From this, one can deduce that since $x \le -\zeta$,

$$G_{\tilde{X}}V(x) \le -2(\mu\wedge\alpha)x^2 + 2\mu,$$

and also

$$G_{\tilde{X}}V(x) \le -2\alpha\,|\zeta|\,|x| + 2\mu.$$

Then Lemma 1 and Jensen's inequality can be applied as before to get both (A.16) and (A.17).

We now prove (A.23). Observe that

$$\begin{aligned}
&\mathbb{E}\Big[\big|\tilde{X}(\infty)1(\tilde{X}(\infty) \ge -\zeta)\big|\Big] \\
&= \mathbb{E}\Big[\big|(\tilde{X}(\infty) + \zeta - \zeta)1(\tilde{X}(\infty) \ge -\zeta)\big|\Big] \\
&\ge \mathbb{E}\Big[\big|(\tilde{X}(\infty) + \zeta)1(\tilde{X}(\infty) > -\zeta)\big| - \zeta 1(\tilde{X}(\infty) > -\zeta)\Big] \\
&\ge \mathbb{E}\Big[\big|(\tilde{X}(\infty) + \zeta)1(\tilde{X}(\infty) > -\zeta)\big|\Big] - \zeta \\
&= \frac{\mu}{\alpha}\mathbb{E}\Big[\big|(\tilde{X}(\infty) + \zeta)1(\tilde{X}(\infty) \le -\zeta)\big|\Big],
\end{aligned}$$

where the last equality comes from applying the generator $G_{\tilde{X}}$ to the function $f(x) = x$ and taking expected values with respect to $\tilde{X}(\infty)$ to see that $\mathbb{E}b(\tilde{X}(\infty)) = 0$, or

(A.29)
$$\mathbb{E}\Big[\big(-\mu(\tilde{X}(\infty) + \zeta) + \alpha\zeta\big)1(\tilde{X}(\infty) \le -\zeta)\Big] = \alpha\mathbb{E}\Big[\tilde{X}(\infty)1(\tilde{X}(\infty) > -\zeta)\Big].$$

Therefore,

$$\mathbb{E}\Big[\big|(\tilde{X}(\infty) + \zeta)1(\tilde{X}(\infty) \le -\zeta)\big|\Big] \le \frac{\alpha}{\mu}\mathbb{E}\Big[\big|\tilde{X}(\infty)1(\tilde{X}(\infty) \ge -\zeta)\big|\Big],$$

and we can invoke (A.19) to conclude (A.23).

We now prove (A.24), which requires additional arguments that we have not used in the proof of Lemma 2. We assume for now that

$$\lambda \le n\mu + \frac{1}{2}\sqrt{n}\mu. \tag{A.30}$$

Fix $\gamma \in (0, 1/2)$, and define

$$J_1 = \sum_{k=0}^{\lfloor n-\gamma\sqrt{R}\rfloor} \nu_k, \quad J_2 = \sum_{k=\lceil n-\gamma\sqrt{R}\rceil}^{n} \nu_k, \tag{A.31}$$

where $\{\nu_k\}_{k=0}^{\infty}$ is the stationary distribution of X. We note that by (A.30),

$$n/\sqrt{R} \ge \sqrt{R} - \frac{1}{2}\sqrt{n/R} \ge \sqrt{R} - 1/2 \ge 1/2,$$

which implies that $n-\gamma\sqrt{R}>0$. Then

$$\mathbb{P}(\tilde{X}(\infty)\le -\zeta)=\mathbb{P}(X(\infty)\le n)\le J_1+J_2.$$

To bound J_1 we observe that

$$\mathbb{E}\left[\left|\widetilde{X}(\infty)+\zeta\right|1_{\{\widetilde{X}(\infty)\le-\zeta\}}\right]=\frac{1}{\sqrt{R}}\sum_{k=0}^{n}(n-k)\nu_k\ge\gamma\sum_{k=0}^{\lfloor n-\gamma\sqrt{R}\rfloor}\nu_k=\gamma J_1.$$

Combining (A.20)–(A.23), we conclude that

$$\begin{aligned}J_1&\le\frac{1}{\gamma}\frac{2}{\sqrt{3}}\Big(\frac{\delta^2}{4}+1\Big)\Big(\frac{1}{\zeta}\wedge\sqrt{\frac{\alpha}{\mu}\vee 1}\wedge\frac{\alpha}{\mu}\sqrt{\frac{\mu}{\alpha}\vee 1}\Big)\\&\le\frac{1}{\gamma}\frac{2}{\sqrt{3}}\Big(\frac{\delta^2}{4}+1\Big)\Big(\frac{1}{\zeta}\wedge\sqrt{\frac{\alpha}{\mu}}\Big).\end{aligned}\tag{A.32}$$

Now to bound J_2, we apply $G_{\tilde{X}}$ to the test function $f(x)=k\wedge n$, where $x=\delta(k-x(\infty))$, and take the expectation with respect to $\tilde{X}(\infty)$ to see that

$$0=-\lambda\nu_n+(\lambda-n\mu)\mathbb{P}(X(\infty)\le n)+\mu\mathbb{E}\left[(X(\infty)-n)^-1_{\{X(\infty)\le n\}}\right].$$

Noticing that

$$\mathbb{E}\left[(X(\infty)-n)^-1_{\{X(\infty)\le n\}}\right]=\frac{1}{\delta}\mathbb{E}\left[\left|\widetilde{X}(\infty)+\zeta\right|1_{\{\widetilde{X}(\infty)\le-\zeta\}}\right],$$

we arrive at

$$\nu_n\le\delta\frac{2}{\sqrt{3}}\Big(\frac{\delta^2}{4}+1\Big)\left(\frac{1}{\zeta}\wedge\sqrt{\frac{\alpha}{\mu}}\right)+\frac{\lambda-n\mu}{\lambda}\mathbb{P}(X(\infty)\le n).\tag{A.33}$$

The flow balance equations

$$\lambda\nu_{k-1}=k\mu\nu_k,\quad k=1,2,\cdots,n$$

imply that $\nu_0<\nu_1<\cdots<\nu_{n-2}<\nu_{n-1}\le\nu_n$, and therefore

$$\begin{aligned}&J_2\le(\gamma\sqrt{R}+1)\nu_n\\&\le(\gamma\sqrt{R}+1)\Big[\delta\frac{2}{\sqrt{3}}\Big(\frac{\delta^2}{4}+1\Big)\left(\frac{1}{\zeta}\wedge\sqrt{\frac{\alpha}{\mu}}\right)+\frac{\lambda-n\mu}{\lambda}\mathbb{P}(X(\infty)\le n)\Big]\\&=(\gamma+\delta)\frac{2}{\sqrt{3}}\Big(\frac{\delta^2}{4}+1\Big)\left(\frac{1}{\zeta}\wedge\sqrt{\frac{\alpha}{\mu}}\right)\\&\quad+(\gamma\sqrt{R}+1)\frac{\lambda-n\mu}{\lambda}J_1+(\gamma\sqrt{R}+1)\frac{\lambda-n\mu}{\lambda}J_2\end{aligned}\tag{A.34}$$

We use (A.30), the fact that $\gamma \in (0, 1/2)$, and that $R \geq n \geq 1$ to see that

$$(\gamma\sqrt{R}+1)\frac{\lambda - n\mu}{\lambda} \leq (\gamma\sqrt{R}+1)\frac{\sqrt{n}}{2R} \leq \frac{1}{2}(\gamma + 1/\sqrt{R}) = \frac{1}{2}(\gamma+1) < 3/4.$$

Then by rearranging terms in (A.34) and applying (A.32) we conclude that

$$\begin{aligned}\frac{1}{4}J_2 &\leq (\gamma+\delta)\frac{2}{\sqrt{3}}\Big(\frac{\delta^2}{4}+1\Big)\left(\frac{1}{\zeta}\wedge\sqrt{\frac{\alpha}{\mu}}\right) + \frac{3}{4}\frac{1}{\gamma}\frac{2}{\sqrt{3}}\Big(\frac{\delta^2}{4}+1\Big)\Big(\frac{1}{\zeta}\wedge\sqrt{\frac{\alpha}{\mu}}\Big)\\ &= \Big(\gamma+\delta+\frac{3}{4}\frac{1}{\gamma}\Big)\frac{2}{\sqrt{3}}\Big(\frac{\delta^2}{4}+1\Big)\left(\frac{1}{\zeta}\wedge\sqrt{\frac{\alpha}{\mu}}\right).\end{aligned}$$

Hence, we have just shown that under assumption (A.30),

$$\begin{aligned}\mathbb{P}(\tilde{X}(\infty)\leq -\zeta) \leq J_1 + J_2 &\leq \frac{1}{\gamma}\frac{2}{\sqrt{3}}\Big(\frac{\delta^2}{4}+1\Big)\Big(\frac{1}{\zeta}\wedge\sqrt{\frac{\alpha}{\mu}}\Big)\\ &\quad + 4\Big(\gamma+\delta+\frac{3}{4}\frac{1}{\gamma}\Big)\frac{2}{\sqrt{3}}\Big(\frac{\delta^2}{4}+1\Big)\left(\frac{1}{\zeta}\wedge\sqrt{\frac{\alpha}{\mu}}\right)\\ &\leq (3+\delta)\frac{8}{\sqrt{3}}\Big(\frac{\delta^2}{4}+1\Big)\left(\frac{1}{\zeta}\wedge\sqrt{\frac{\alpha}{\mu}}\right),\end{aligned}$$

where to get the last inequality we fixed $\gamma \in (0, 1/2)$ that solves $\gamma + 1/\gamma = 3$.

We now wish to establish the same result without assumption (A.30), i.e. when $\lambda > n\mu + \frac{1}{2}\sqrt{n}\mu$. For this, we rely on the following comparison result. Fix n, μ and α and let $X^{(\lambda)}(\infty)$ be the steady-state customer count in an Erlang-A system with arrival rate λ, service rate μ, number of servers n, and abandonment rate α. Then for any $0 < \lambda_1 < \lambda_2$,

$$\mathbb{P}(X^{(\lambda_2)}(\infty) \leq n) \leq \mathbb{P}(X^{(\lambda_1)}(\infty) \leq n). \tag{A.35}$$

This says that with all other parameters being held fixed, an Erlang-A system with a higher arrival rate is less likely to have idle servers. For a simple proof involving a coupling argument, see page 163 of [45].

Therefore, for $\lambda > n\mu + \frac{1}{2}\sqrt{n}\mu$,

$$\begin{aligned}&\mathbb{P}(X^{(\lambda)}(\infty)\leq n) \leq \mathbb{P}(X^{(n\mu+\frac{1}{2}\sqrt{n}\mu)}(\infty)\leq n)\\ &\leq (3+\delta)\frac{8}{\sqrt{3}}\Big(\frac{\delta^2}{4}+1\Big)\left(\frac{1}{\zeta^{(n\mu+\frac{1}{2}\sqrt{n}\mu)}}\wedge\sqrt{\frac{\alpha}{\mu}}\right)\end{aligned}$$

where $\zeta^{(n\mu+\frac{1}{2}\sqrt{n}\mu)}$ is the ζ corresponding to $X^{(n\mu+\frac{1}{2}\sqrt{n}\mu)}(\infty)$, and satisfies

$$\frac{1}{\zeta^{(n\mu+\frac{1}{2}\sqrt{n}\mu)}} = \frac{2\alpha}{\mu}\sqrt{\frac{n+\sqrt{n}/2}{n}} \leq \frac{2\alpha}{\mu}\sqrt{\frac{3}{2}}.$$

This concludes the proof of (A.24).

APPENDIX B: GRADIENT BOUNDS

In Section B.1, we first prove Lemma 3, establishing the Wasserstein gradient bounds for Erlang-C model. In Section B.2, we state and prove Lemma 13, establishing the Wasserstein gradient bounds for Erlang-A model. In Section B.3 we prove Lemmas 4 and 5, establishing the Kolmogorov gradient bounds for both Erlang-C and Erlang-A models.

The arguments below follow the proof of [15, Lemma 13.1]. We recall the Poisson equation (3.3), or

$$G_Y f_h(x) = b(x) f_h'(x) + \mu f_h''(x) = \mathbb{E}h(Y(\infty)) - h(x), \quad x \in \mathbb{R},\ h(x) \in \mathcal{H}.$$

Furthermore, recall that $\nu(x)$ is the density of $Y(\infty)$, and satisfies

$$\nu(x) = \kappa \exp\Big(\frac{1}{\mu}\int_0^x b(y)dy\Big),$$

where κ is a normalizing constant. Now recall that the family of solutions to the Poisson equation is given by (3.4), and is parametrized by constants $a_1, a_2 \in \mathbb{R}$. We fix a solution $f_h(x)$ with $a_2 = 0$, and see that for this solution

$$f_h'(x) = \frac{1}{\nu(x)}\int_{-\infty}^x \frac{1}{\mu}\big(\mathbb{E}h(Y(\infty)) - h(y)\big)\nu(y)dy, \quad x \in \mathbb{R}. \tag{B.1}$$

Observe that since $\nu(x)$ is the density of $Y(\infty)$,

$$\int_{-\infty}^{\infty} \frac{1}{\mu}\big(\mathbb{E}h(Y(\infty)) - h(y)\big)\nu(y)dy = 0,$$

which implies that

$$f_h'(x) = -\frac{1}{\nu(x)}\int_x^{\infty} \frac{1}{\mu}\big(\mathbb{E}h(Y(\infty)) - h(y)\big)\nu(y)dy. \tag{B.2}$$

We will see that to establish gradient bounds, we will have to make use of both expressions for $f_h'(x)$ in (B.1) and (B.2), depending on the value of x. It is here that the relationship between the diffusion process Y and the random variable $Y(\infty)$ surfaces. If $\mathbb{E}h(Y(\infty))$ were to be replaced by any other constant, then (B.2) would not hold. The reason $\mathbb{E}h(Y(\infty))$ is the "correct" constant is because $Y(\infty)$ has the stationary distribution of the diffusion process Y.

We now proceed to prove the gradient bounds. We do this first in case when $\mathcal{H} = \mathrm{Lip}(1)$ (the Wasserstein setting), and then when $\mathcal{H} = \mathcal{H}_K$ (the Kolmogorov setting).

B.1. Wasserstein gradient bounds. Fix $h(x) \in \mathrm{Lip}(1)$; without loss of generality we assume that $h(0) = 0$. We now derive some equations that will be useful to prove Lemma 3. From the form of $f_h'(x)$ in (B.1) and (B.2), we have

$$f_h'(x) \le \frac{1}{\mu\nu(x)} \int_{-\infty}^{x} |y|\, \nu(y)dy + \frac{\mathbb{E}\big|Y(\infty)\big|}{\mu\nu(x)} \int_{-\infty}^{x} \nu(y)dy,$$
$$f_h'(x) \le \frac{1}{\mu\nu(x)} \int_{x}^{\infty} |y|\, \nu(y)dy + \frac{\mathbb{E}\big|Y(\infty)\big|}{\mu\nu(x)} \int_{x}^{\infty} \nu(y)dy.$$

Now by differentiating the Poisson equation (3.3), we see that for those $x \in \mathbb{R}$ where $h'(x)$ and $b'(x)$ are defined,

$$f_h'''(x) = \frac{1}{\mu}\big[- h'(x) - b'(x)f_h'(x) - b(x)f_h''(x)\big].$$

By observing that $b(x) = \mu\nu'(x)/\nu(x)$, we can rearrange the terms above and multiply both sides by $\nu(x)$ to get

(B.3)
$$\frac{\nu(x)}{\mu}(-h'(x) - f_h'(x)b'(x)) = \nu(x)f_h'''(x) + \nu'(x)f_h''(x) = \big(f_h''(x)\nu(x)\big)'.$$

Suppose we know that

$$\lim_{x\to\pm\infty} \nu(x)f_h''(x) = 0. \tag{B.4}$$

Since $\lim_{x\to-\infty} \nu(x)f_h''(x) = 0$, we integrate (B.3) to get

$$f_h''(x) = \frac{1}{\nu(x)} \int_{-\infty}^{x} \frac{1}{\mu}(-h'(y) - f_h'(y)b'(y))\nu(y)dy,$$

and since $\lim_{x\to\infty} \nu(x)f_h''(x) = 0$, it is also true that

$$f_h''(x) = \ -\frac{1}{\nu(x)} \int_{x}^{\infty} \frac{1}{\mu}(-h'(y) - f_h'(y)b'(y))\nu(y)dy.$$

To summarize, we have just shown that provided (B.4) holds,

$$\big|f_h'(x)\big| \le \frac{1}{\mu\nu(x)} \int_{-\infty}^{x} |y|\, \nu(y)dy + \frac{\mathbb{E}\big|Y(\infty)\big|}{\mu\nu(x)} \int_{-\infty}^{x} \nu(y)dy, \tag{B.5}$$
$$\big|f_h'(x)\big| \le \frac{1}{\mu\nu(x)} \int_{x}^{\infty} |y|\, \nu(y)dy + \frac{\mathbb{E}\big|Y(\infty)\big|}{\mu\nu(x)} \int_{x}^{\infty} \nu(y)dy, \tag{B.6}$$

$$f_h''(x) = \frac{1}{\nu(x)} \int_{-\infty}^{x} \frac{1}{\mu}(-h'(y) - f_h'(y)b'(y))\nu(y)dy \tag{B.7}$$

$$= -\frac{1}{\nu(x)} \int_{x}^{\infty} \frac{1}{\mu}(-h'(y) - f_h'(y)b'(y))\nu(y)dy, \tag{B.8}$$

$$f_h'''(x) = \frac{1}{\mu}\big[-h'(x) - f_h''(x)b(x) - f_h'(x)b'(x)\big], \tag{B.9}$$

where $f_h'''(x)$ is defined for all $x \in \mathbb{R}$ such that $h'(x)$ and $b'(x)$ exist. The multiple bounds for $f_h'(x)$ are convenient, because we can choose which one to use based on the value of x. For example, suppose $\zeta \le 0$. Then when $x \le 0$, we will use (B.5), when $x \ge -\zeta$ we will use (B.6), and when $x \in [0, -\zeta]$ we will use both (B.5) and (B.6) and choose the better bound. The same applies to $f_h''(x)$.

B.1.1. *Proof of Lemma 3.* The following lemma presents several bounds that will be used to prove Lemma 3. We prove it at the end of this section.

LEMMA 12. *Consider the Erlang-C model ($\alpha = 0$), and let $\nu(x)$ be the density of $Y(\infty)$. Then*

$$\frac{1}{\nu(x)} \int_{-\infty}^{x} \nu(y)dy \le \begin{cases} \sqrt{\frac{\pi}{2}}, & x \le 0, \\ \sqrt{2\pi} e^{\frac{1}{2}\zeta^2}, & x \in [0, -\zeta], \end{cases} \tag{B.10}$$

$$\frac{1}{\nu(x)} \int_{x}^{\infty} \nu(y)dy \le \begin{cases} \sqrt{\frac{\pi}{2}} + \frac{1}{|\zeta|}, & x \in [0, -\zeta], \\ \frac{1}{|\zeta|}, & x \ge -\zeta, \end{cases} \tag{B.11}$$

$$\frac{1}{\nu(x)} \int_{-\infty}^{x} |y|\,\nu(y)dy \le \begin{cases} 1, & x \le 0, \\ 2e^{\frac{1}{2}\zeta^2} - 1, & x \in [0, -\zeta], \end{cases} \tag{B.12}$$

$$\frac{1}{\nu(x)} \int_{x}^{\infty} |y|\,\nu(y)dy \le \begin{cases} 2 + \frac{1}{\zeta^2}, & x \in [0, -\zeta], \\ \frac{x}{|\zeta|} + \frac{1}{\zeta^2}, & x \ge -\zeta, \end{cases} \tag{B.13}$$

$$\frac{|b(x)|}{\mu\nu(x)} \int_{-\infty}^{x} \nu(y)dy \le 1, \quad x \le 0, \tag{B.14}$$

$$\frac{|b(x)|}{\mu\nu(x)} \int_{x}^{\infty} \nu(y)dy \le 2, \quad x \ge 0, \tag{B.15}$$

$$\mathbb{E}\,|Y(\infty)| \le \frac{1}{|\zeta|} + 1. \tag{B.16}$$

PROOF OF LEMMA 3. We begin by bounding $f_h'(x)$ by applying Lemma 12 to (B.5) and (B.6). For $x \le -\zeta$, we apply (B.10), (B.12), and (B.16) to (B.5),

and for $x \geq 0$, we apply (B.11), (B.13), and (B.16) to (B.6) to see that

$$\begin{aligned}
\mu\left|f_h'(x)\right| &\leq 1+\sqrt{\frac{\pi}{2}}\Big(1+\frac{1}{|\zeta|}\Big) \leq 2.3+1.3/\left|\zeta\right|, \quad x \leq 0,\\
\mu\left|f_h'(x)\right| &\leq \min\Big\{2e^{\frac{1}{2}\zeta^2}-1+\sqrt{2\pi}e^{\frac{1}{2}\zeta^2}\Big(1+\frac{1}{|\zeta|}\Big),\\
&\qquad 2+\frac{1}{\zeta^2}+\Big(\sqrt{\frac{\pi}{2}}+\frac{1}{|\zeta|}\Big)\Big(1+\frac{1}{|\zeta|}\Big)\Big\}, \quad x \in [0,-\zeta],
\end{aligned} \tag{B.17}$$

$$\mu\left|f_h'(x)\right| \leq \frac{x}{|\zeta|}+\frac{1}{\zeta^2}+\frac{1}{|\zeta|}\Big(1+\frac{1}{|\zeta|}\Big) \leq \frac{1}{|\zeta|}(x+1+2/\left|\zeta\right|), \quad x \geq -\zeta.$$

For $x \in [0,-\zeta]$, observe that when $|\zeta| \leq 1$, then

$$2e^{\frac{1}{2}\zeta^2}-1+\sqrt{2\pi}e^{\frac{1}{2}\zeta^2}\Big(1+\frac{1}{|\zeta|}\Big) \leq 3.3-1+4.2\Big(1+\frac{1}{|\zeta|}\Big) = 6.5+4.2/\left|\zeta\right|,$$

and when $|\zeta| \geq 1$, then

$$2+\frac{1}{\zeta^2}+\Big(\sqrt{\frac{\pi}{2}}+\frac{1}{|\zeta|}\Big)\Big(1+\frac{1}{|\zeta|}\Big) \leq 3+(1.3+1)\Big(1+\frac{1}{|\zeta|}\Big) = 5.3+2.3/\left|\zeta\right|.$$

Therefore,

$$\left|f_h'(x)\right| \leq \begin{cases} \frac{1}{\mu}(6.5+4.2/\left|\zeta\right|), & x \leq -\zeta,\\ \frac{1}{\mu}\frac{1}{|\zeta|}(x+1+2/\left|\zeta\right|), & x \geq -\zeta. \end{cases} \tag{B.18}$$

Before proceeding to bound $|f_h''(x)|$ and $|f_h'''(x)|$, we first verify (B.4), or

$$\lim_{x\to\pm\infty}\nu(x)f_h''(x)=0.$$

To do so, we rearrange the Poisson equation (3.3) to get

$$\nu(x)f_h''(x) = -\nu(x)\frac{b(x)}{\mu}f_h'(x)+\frac{\nu(x)}{\mu}\big[\mathbb{E}h(Y(\infty))-h(x)\big].$$

It is now obvious that

$$\lim_{x\to\pm\infty}\nu(x)f_h''(x) \to 0 = 0,$$

because $h(x) \in \mathrm{Lip}(1)$, the drift $b(x)$ is piecewise linear, and $f_h'(x)$ is bounded as in (B.18), but on the other hand $\nu(x)$ decays exponentially fast as $x \to \infty$,

and decays even faster as $x \to -\infty$. To bound $|f_h''(x)|$, we use (B.7) and (B.8), together with the facts that $\|h'\| \le 1$ and

$$b'(x) = -\mu 1(x < -\zeta), \quad x \in \mathbb{R},$$

to see that

$$|f_h''(x)| \le \frac{1}{\nu(x)} \int_{-\infty}^{x} \frac{1}{\mu}\big(1 + \mu\,|f_h'(y)|\,1(y < -\zeta)\big)\nu(y)dy \tag{B.19}$$

$$|f_h''(x)| \le \frac{1}{\nu(x)} \int_{x}^{\infty} \frac{1}{\mu}\big(1 + \mu\,|f_h'(y)|\,1(y < -\zeta)\big)\nu(y)dy. \tag{B.20}$$

We know $|f_h'(x)|$ is bounded as in (B.18). For $x \le -\zeta$, we apply (B.10) to (B.19) and for $x \ge 0$ we apply (B.11) to (B.20) to conclude that

(B.21)
$$\mu\,|f_h''(x)| \le \begin{cases} \sqrt{\frac{\pi}{2}}\big(1 + 6.5 + 4.2/\,|\zeta|\,\big), & x \le 0, \\ \min\big\{\sqrt{2\pi}e^{\frac{1}{2}\zeta^2}, \sqrt{\frac{\pi}{2}} + \frac{1}{|\zeta|}\big\}\big(1 + 6.5 + 4.2/\,|\zeta|\,\big), & x \in [0, -\zeta], \\ \frac{1}{|\zeta|}, & x \ge -\zeta. \end{cases}$$

By considering separately the cases when $|\zeta| \le 1$ and $|\zeta| \ge 1$, we see that

$$\min\left\{\sqrt{2\pi}e^{\frac{1}{2}\zeta^2}, \sqrt{\frac{\pi}{2}} + \frac{1}{|\zeta|}\right\} \le 4.2, \tag{B.22}$$

and therefore,

$$|f_h''(x)| \le \begin{cases} \frac{32}{\mu}(1 + 1/\,|\zeta|), & x \le -\zeta, \\ \frac{1}{\mu|\zeta|}, & x \ge -\zeta. \end{cases} \tag{B.23}$$

Lastly, we bound $|f_h'''(x)|$, which exists for all $x \in \mathbb{R}$ where $h'(x)$ and $b'(x)$ exist. The fact that $h(x) \in \mathrm{Lip}(1)$ together with (B.9) tells us that

$$|f_h'''(x)| \le \frac{1}{\mu}\big(1 + |f_h''(x)b(x)| + |f_h'(x)b'(x)|\,\big).$$

For $x \ge -\zeta$, we use the forms of $b(x)$ and $b'(x)$ together with the bounds on $|f_h'(x)|$ and $|f_h''(x)|$ in (B.18) and (B.23) to see that

$$|f_h'''(x)| \le \frac{1}{\mu}\Big(1 + \mu\,|\zeta|\,\frac{1}{\mu\,|\zeta|}\Big).$$

Although tempting, it is not sufficient to use the bound on $|f_h''(x)|$ in (B.21) and the form of $b(x)$ to bound $|f_h''(x)b(x)|$ for all $x \le -\zeta$. Instead, we

multiply both sides of (B.19) and (B.20) by $|b(x)|$ to see that

$$\|f_h''(x)b(x)\| \le \big(1+\sup_{y\le x}\mu\,|f_h'(y)|\,\big)\frac{|b(x)|}{\mu\nu(x)}\int_{-\infty}^{x}\nu(y)dy, \quad x\le 0,$$

(B.24)

$$\|f_h''(x)b(x)\| \le \big(1+\sup_{y\in[x,-\zeta]}\mu\,|f_h'(y)|\,\big)\frac{|b(x)|}{\mu\nu(x)}\int_{x}^{\infty}\nu(y)dy, \quad x\in[0,-\zeta].$$

By invoking (B.14) and (B.15) together with the bound on $|f_h'(x)|$ from (B.18), we conclude that

$$|f_h'''(x)| \le \frac{1}{\mu}\big(1+2(1+6.5+4.2/|\zeta|)+6.5+4.2/|\zeta|\,\big), \quad x\le -\zeta.$$

Therefore, for those $x\in\mathbb{R}$ where $h'(x)$ and $b'(x)$ exist,

$$|f_h'''(x)| \le \begin{cases}\frac{1}{\mu}(23+13/|\zeta|), & x\le -\zeta,\\ \frac{2}{\mu}, & x\ge -\zeta.\end{cases}$$

This concludes the proof of Lemma 3. □

Proof of Lemma 12 . Recall that in the Erlang-C model, we set $\alpha=0$, which makes

$$\zeta=\delta(x(\infty)-n)=\delta(\frac{\lambda}{\mu}-n)<0, \quad b(x)=\begin{cases}-\mu x, & x\le-\zeta,\\ \mu\zeta, & x\ge-\zeta.\end{cases}$$

The density of $Y(\infty)$ has the form

$$\nu(x)=\begin{cases}a_- e^{-\frac{1}{2}x^2}, & x\le-\zeta,\\ a_+ e^{-|\zeta|x}, & x\ge-\zeta,\end{cases} \tag{B.25}$$

where a_- and a_+ are normalizing constants that make $\nu(x)$ continuous at the point $x=-\zeta$.

In the proof of this lemma we often rely on the fact that for any $c>0$ and $x\ge 0$,

$$e^{\frac{1}{2}cx^2}\int_x^\infty e^{-\frac{1}{2}cy^2}dy \le \int_0^\infty e^{-\frac{1}{2}cy^2}dy=\sqrt{\frac{\pi}{2c}}. \tag{B.26}$$

One can verify that the left hand side of (B.26) peaks at $x=0$ by using the bound

$$\int_x^\infty e^{-\frac{1}{2}cy^2}dy \le \int_x^\infty \frac{cy}{cx}e^{-\frac{1}{2}cy^2}dy=\frac{e^{-cx^2/2}}{cx} \tag{B.27}$$

to see that the derivative of the left side of (B.26) is negative for $x > 0$. We now prove (B.10). When $x \leq 0$, we use (B.26) and the symmetry of the function $e^{-\frac{1}{2}y^2}$ to see that

$$\frac{1}{\nu(x)} \int_{-\infty}^{x} \nu(y) = e^{\frac{1}{2}x^2} \int_{-\infty}^{x} e^{-\frac{1}{2}y^2} dy \leq \sqrt{\frac{\pi}{2}},$$

and when $x \in [0, -\zeta]$, we use (B.26) again to get

$$\begin{aligned} \frac{1}{\nu(x)} \int_{-\infty}^{x} \nu(y) &= e^{\frac{1}{2}x^2} \int_{-\infty}^{0} e^{-\frac{1}{2}y^2} dy + e^{\frac{1}{2}x^2} \int_{0}^{x} e^{-\frac{1}{2}y^2} dy \\ &\leq 2e^{\frac{1}{2}\zeta^2} \int_{-\infty}^{0} e^{-\frac{1}{2}y^2} dy = 2\sqrt{\frac{\pi}{2}} e^{\frac{1}{2}\zeta^2}. \end{aligned}$$

We now prove (B.11). Observe that when $x \geq -\zeta$,

$$\frac{1}{\nu(x)} \int_{x}^{\infty} \nu(y) dy = e^{|\zeta|x} \int_{x}^{\infty} e^{-|\zeta|y} dy = \frac{1}{|\zeta|},$$

and when $x \in [0, -\zeta]$, we use the fact that $a_- = a_+ e^{-\frac{1}{2}\zeta^2}$ together with (B.26) to see that

$$\begin{aligned} \frac{1}{\nu(x)} \int_{x}^{\infty} \nu(y) dy &= \frac{a_+}{a_-} e^{\frac{1}{2}x^2} \int_{-\zeta}^{\infty} e^{-|\zeta|y} dy + e^{\frac{1}{2}x^2} \int_{x}^{-\zeta} e^{-\frac{1}{2}y^2} dy \\ &\leq e^{\frac{1}{2}\zeta^2} e^{\frac{1}{2}x^2} \left(\frac{1}{|\zeta|} e^{-\zeta^2} \right) + \sqrt{\frac{\pi}{2}} \\ &\leq \frac{1}{|\zeta|} + \sqrt{\frac{\pi}{2}}. \end{aligned}$$

Moving on to show (B.12), when $x \leq 0$ we have

$$\frac{1}{\nu(x)} \int_{-\infty}^{x} |y| \, \nu(y) dy = e^{\frac{1}{2}x^2} \int_{-\infty}^{x} -ye^{-\frac{1}{2}y^2} dy = 1,$$

and when $x \in [0, -\zeta]$,

$$\begin{aligned} \frac{1}{\nu(x)} \int_{-\infty}^{x} |y| \, \nu(y) dy &= e^{\frac{1}{2}x^2} \int_{-\infty}^{0} -ye^{-\frac{1}{2}y^2} dy + e^{\frac{1}{2}x^2} \int_{0}^{x} ye^{-\frac{1}{2}y^2} dy \\ &= e^{\frac{1}{2}x^2} + e^{\frac{1}{2}x^2}(1 - e^{-\frac{1}{2}x^2}). \end{aligned}$$

We now prove (B.13). When $x \in [0, -\zeta]$, we again use that $a_- = a_+ e^{-\frac{1}{2}\zeta^2}$ to obtain

$$\frac{1}{\nu(x)} \int_{x}^{\infty} |y| \, \nu(y) dy = e^{\frac{1}{2}x^2} \int_{x}^{-\zeta} ye^{-\frac{1}{2}y^2} dy + \frac{a_+}{a_-} e^{\frac{1}{2}x^2} \int_{-\zeta}^{\infty} ye^{-|\zeta|y} dy$$

$$= e^{\frac{1}{2}x^2}(e^{-\frac{1}{2}x^2} - e^{-\frac{1}{2}\zeta^2}) + e^{\frac{1}{2}\zeta^2}e^{\frac{1}{2}x^2}\Big(1 + \frac{1}{\zeta^2}\Big)e^{-\zeta^2}$$
$$\leq 2 + \frac{1}{\zeta^2}.$$

When $x \geq -\zeta$,

$$\frac{1}{\nu(x)}\int_x^\infty |y|\,\nu(y)dy = e^{|\zeta|x}\int_x^\infty ye^{-|\zeta|y}dy = \frac{x}{|\zeta|} + \frac{1}{\zeta^2}.$$

We move on to prove (B.14) and (B.15). When $x \leq 0$, we use (B.27) to see that

$$\frac{|b(x)|}{\mu\nu(x)}\int_{-\infty}^x \nu(y)dy = -xe^{\frac{x^2}{2}}\int_{-\infty}^x e^{-\frac{1}{2}y^2}dy \leq -xe^{\frac{x^2}{2}}\Big(-\frac{1}{x}e^{-\frac{1}{2}x^2}\Big) = 1.$$

When $x \in [0, -\zeta]$, we also use the fact that $a_+ = a_-e^{\frac{1}{2}\zeta^2}$ to get

$$\frac{|b(x)|}{\mu\nu(x)}\int_x^\infty \nu(y)dy = xe^{\frac{x^2}{2}}\int_x^{-\zeta} e^{-\frac{1}{2}y^2}dy + \frac{a_+}{a_-}xe^{\frac{x^2}{2}}\int_{-\zeta}^\infty e^{-|\zeta|y}dy$$
$$\leq xe^{\frac{x^2}{2}}\Big(\frac{1}{x}e^{-\frac{1}{2}x^2}\Big) + e^{\frac{1}{2}\zeta^2}xe^{\frac{1}{2}x^2}\Big(\frac{1}{|\zeta|}e^{-\zeta^2}\Big)$$
$$= 1 + \frac{x}{|\zeta|}e^{\frac{1}{2}(x^2-\zeta^2)} \leq 2.$$

When $x \geq -\zeta$,

$$\frac{|b(x)|}{\mu\nu(x)}\int_x^\infty \nu(y)dy = |\zeta|\,e^{|\zeta|x}\int_x^\infty e^{-|\zeta|y}dy = 1.$$

Lastly we prove (B.16). Letting $V(x) = x^2$, and recalling the form of G_Y from (3.2), we consider

$$G_YV(x) = 2x\mu(\zeta + (x+\zeta)^-) + 2\mu$$
$$= -2\mu x^2 1(x < -\zeta) - 2x\mu\,|\zeta|\,1(x \geq -\zeta) + 2\mu. \tag{B.28}$$

By the standard Foster-Lyapunov condition (see [47, Theorem 4.3] for example), this implies that

$$2\mathbb{E}\big[(Y(\infty))^2 1(Y(\infty) < -\zeta)\big] + 2\,|\zeta|\,\mathbb{E}\big[Y(\infty)1(Y(\infty) \geq -\zeta)\big] \leq 2,$$

and in particular,

$$\mathbb{E}\big[Y(\infty)1(Y(\infty) \geq -\zeta)\big] \leq \frac{1}{|\zeta|},$$
$$\mathbb{E}\Big[\big|Y(\infty)1(Y(\infty) < -\zeta)\big|\Big] \leq \sqrt{\mathbb{E}\big[(Y(\infty))^2 1(Y(\infty) < -\zeta)\big]} \leq 1,$$

where we applied Jensen's inequality in the second set of inequalities. This concludes the proof of Lemma 12. □

B.2. Erlang-A Wasserstein gradient bounds. Below we state the Erlang-A gradient bounds, the proof of which is similar to that of Lemma 3. We only outline the necessary steps needed for a proof, and emphasize all the differences with the proof of Lemma 3. Furthermore, we do not seek an explicit value for the constant C below, although it can certainly be recovered.

LEMMA 13. *Consider the Erlang-A model ($\alpha > 0$). The solution to the Poisson equation $f_h(x)$ is twice continuously differentiable, with an absolutely continuous second derivative. Fix a solution in (3.4) with parameter $a_2 = 0$. Then there exists a constant $C > 0$ independent of λ, n, μ, and α such that for all $n \geq 1, \lambda > 0, \mu > 0$, and $\alpha > 0$ satisfying $0 < R \leq n$ (an underloaded system),*

(B.29)
$$\left|f_h'(x)\right| \leq \begin{cases} C\left(\sqrt{\frac{\mu}{\alpha}} \wedge \frac{1}{|\zeta|} + 1\right)\frac{1}{\mu}, & x \leq -\zeta, \\ C\left(\frac{\mu}{\alpha} + \sqrt{\frac{\mu}{\alpha}} \wedge \frac{1}{|\zeta|} + 1\right)\frac{1}{\mu}, & x \geq -\zeta, \end{cases}$$

(B.30)
$$\left|f_h''(x)\right| \leq \begin{cases} C\left(\sqrt{\frac{\mu}{\alpha}} \wedge \frac{1}{|\zeta|} + 1\right)\frac{1}{\mu}, & x \leq 0, \\ C\left[\left(\frac{\alpha}{\mu} + \sqrt{\frac{\alpha}{\mu}} + 1\right)\left(\sqrt{\frac{\mu}{\alpha}} \wedge \frac{1}{|\zeta|}\right) + 1\right]\frac{1}{\mu}, & x \in [0, -\zeta], \\ C\left(\frac{\alpha}{\mu} + \sqrt{\frac{\alpha}{\mu}} + 1\right)\left(\sqrt{\frac{\mu}{\alpha}} \wedge \frac{1}{|\zeta|}\right)\frac{1}{\mu}, & x \geq -\zeta, \end{cases}$$

and for those $x \in \mathbb{R}$ where $f_h'''(x)$ exists,

(B.31)
$$\left|f_h'''(x)\right| \leq \begin{cases} C\left(\sqrt{\frac{\mu}{\alpha}} \wedge \frac{1}{|\zeta|} + 1\right)\frac{1}{\mu}, & x \leq 0, \\ C\left(\sqrt{\frac{\mu}{\alpha}} \wedge \frac{1}{|\zeta|} + \frac{\alpha}{\mu} + \sqrt{\frac{\alpha}{\mu}} + 1\right)\frac{1}{\mu}, & x \in [0, -\zeta], \\ C\left(\frac{\alpha}{\mu} + \sqrt{\frac{\alpha}{\mu}} + 1\right)\frac{1}{\mu}, & x \geq -\zeta, \end{cases}$$

and for all $n \geq 1, \lambda > 0, \mu > 0$, and $\alpha > 0$ satisfying $n \leq R$ (an overloaded system),

(B.32)
$$\left|f_h'(x)\right| \leq \begin{cases} C\left(\frac{1}{\mu} + \frac{1}{\sqrt{\alpha}}\frac{1}{\sqrt{\mu}} + \frac{\zeta}{\mu} \wedge \frac{1}{\alpha}\right), & x \leq -\zeta, \\ C\left(\frac{1}{\mu} + \frac{1}{\sqrt{\alpha}}\frac{1}{\sqrt{\mu}} + \frac{1}{\alpha}\right), & x \geq -\zeta, \end{cases}$$

(B.33)
$$\left|f_h''(x)\right| \leq \begin{cases} C\left(\frac{1}{\mu} + \frac{1}{\sqrt{\alpha}}\frac{1}{\sqrt{\mu}} + \frac{\zeta}{\mu} \wedge \frac{1}{\alpha}\right), & x \leq -\zeta, \\ C\left(\frac{\alpha}{\mu} + \sqrt{\frac{\alpha}{\mu}} + 1\right)\frac{1}{\mu}|x| + C\left(\frac{1}{\mu} + \frac{1}{\sqrt{\alpha}}\frac{1}{\sqrt{\mu}}\right), & x \geq -\zeta, \end{cases}$$

and for those $x \in \mathbb{R}$ where $f_h'''(x)$ exists,

(B.34)
$$\left|f_h'''(x)\right| \le \frac{C}{\mu}\Big(1+\sqrt{\frac{\mu}{\alpha}}+\zeta\wedge\frac{\mu}{\alpha}\Big), \quad x \le -\zeta,$$

(B.35)
$$\left|f_h'''(x)\right| \le \frac{C}{\mu}\Big(\frac{\alpha}{\mu}+\sqrt{\frac{\alpha}{\mu}}+1\Big)\Big(1+\frac{\alpha}{\mu}x^2\Big)+\frac{C}{\mu}\Big(\frac{\alpha}{\mu}+\sqrt{\frac{\alpha}{\mu}}\Big)|x|, \quad x \ge -\zeta,$$

(B.36)
$$\left|f_h'''(x)\right| \le \frac{C}{\mu}\Big(\frac{\alpha}{\mu}+\sqrt{\frac{\alpha}{\mu}}+1\Big)+\frac{C}{\mu}\Big(\frac{\alpha}{\mu}+\sqrt{\frac{\alpha}{\mu}}+1\Big)^2|x|, \quad x \ge -\zeta.$$

B.2.1. *Proof outline for Lemma 13: The underloaded system.* To prove Lemma 13, we need the following version of Lemma 12.

LEMMA 14. *Consider the Erlang-A model ($\alpha > 0$) with $0 < R \le n$, and let $\nu(x)$ be the density of $Y(\infty)$. Then*

(B.37)
$$\frac{1}{\nu(x)}\int_{-\infty}^{x}\nu(y)dy \le \begin{cases}\sqrt{\frac{\pi}{2}}, & x \le 0,\\ \sqrt{2\pi}e^{\frac{\zeta^2}{2}}, & x \in [0,-\zeta],\end{cases}$$

(B.38)
$$\frac{1}{\nu(x)}\int_{x}^{\infty}\nu(y)dy \le \begin{cases}\sqrt{\frac{\pi}{2}}+\sqrt{\frac{\pi}{2}\frac{\mu}{\alpha}}\wedge\frac{1}{|\zeta|}, & x \in [0,-\zeta],\\ \sqrt{\frac{\pi}{2}\frac{\mu}{\alpha}}\wedge\frac{1}{|\zeta|}, & x \ge -\zeta,\end{cases}$$

(B.39)
$$\frac{1}{\nu(x)}\int_{-\infty}^{x}|y|\,\nu(y)dy \le \begin{cases}1, & x \le 0,\\ 2e^{\frac{\zeta^2}{2}}-1, & x \in [0,-\zeta],\end{cases}$$

(B.40)
$$\frac{1}{\nu(x)}\int_{x}^{\infty}|y|\,\nu(y)dy \le \begin{cases}2+\frac{1}{\zeta^2}, & x \in [0,-\zeta],\\ 1+\frac{\mu}{\alpha}, & x \ge -\zeta,\end{cases}$$

(B.41)
$$\frac{|b(x)|}{\mu\nu(x)}\int_{-\infty}^{x}\nu(y)dy \le 1, \quad x \le 0,$$

(B.42)
$$\frac{|b(x)|}{\mu\nu(x)}\int_{x}^{\infty}\nu(y)dy \le 2, \quad x \ge 0,$$

(B.43)
$$\mathbb{E}\,|Y(\infty)| \le 1+\sqrt{\frac{\mu}{\alpha}}\wedge\frac{1}{|\zeta|}.$$

To prove this lemma, we first observe that

$$b(x) = \begin{cases}-\mu x, & x \le -\zeta,\\ -\alpha(x+\zeta)+\mu\zeta, & x \ge -\zeta,\end{cases}$$

and

$$\nu(x) = \begin{cases} a_- e^{-\frac{1}{2}x^2}, & x \le -\zeta, \\ a_+ e^{-\frac{\alpha}{2\mu}\left(x+\zeta-\frac{\mu}{\alpha}\zeta\right)^2}, & x \ge -\zeta, \end{cases} \tag{B.44}$$

where a_- and a_+ are normalizing constants that make $\nu(x)$ continuous and integrate to one. By comparing the density in (B.44) to (B.25) for the region $x \le -\zeta$, we immediately see that (B.37), (B.39) and (B.41) are restatements of (B.10), (B.12), and (B.14) from Lemma 12, and hence have already been established. The proof of (B.43) involves applying G_Y to the Lyapunov function $V(x) = x^2$ to see that

$$\begin{aligned} G_Y V(x) = & -2\mu x^2 1(x < -\zeta) + 2\big(-\alpha x^2 + x\zeta(\mu-\alpha)\big)1(x \ge -\zeta) + 2\mu \\ \le & -2\mu x^2 1(x < -\zeta) - 2(\alpha \wedge \mu)x^2 1(x \ge -\zeta) + 2\mu, \end{aligned}$$

and

$$\begin{aligned} G_Y V(x) = & -2\mu x^2 1(x < -\zeta) + 2\big(-\alpha x(x+\zeta) - \mu\,|\zeta|\,x\big)1(x \ge -\zeta) + 2\mu \\ \le & -2\mu x^2 1(x < -\zeta) - 2\mu\,|\zeta|\,x 1(x \ge -\zeta) + 2\mu. \end{aligned}$$

One can compare these inequalities to (B.28) in the proof of Lemma 12 to see that (B.43) follows by the Foster-Lyapunov condition.

We now go over the proofs of (B.38), (B.40) and (B.42). We first prove (B.38) when $x \in [0, -\zeta]$. Observe that

$$\begin{aligned} \frac{1}{\nu(x)}\int_x^\infty \nu(y)dy &= e^{\frac{1}{2}x^2}\int_x^{|\zeta|} e^{-\frac{1}{2}y^2}dy + \frac{a_+}{a_-}e^{\frac{1}{2}x^2}\int_{|\zeta|}^\infty e^{-\frac{\alpha}{2\mu}\left(y+\zeta-\frac{\mu}{\alpha}\zeta\right)^2}dy \\ &\le e^{\frac{1}{2}x^2}\int_x^\infty e^{-\frac{1}{2}y^2}dy + e^{\frac{1}{2}(x^2-\zeta^2)}e^{\frac{\alpha}{2\mu}\left(\frac{\mu}{\alpha}\zeta\right)^2}\int_{\frac{\mu}{\alpha}|\zeta|}^\infty e^{-\frac{\alpha}{2\mu}y^2}dy \\ &\le \sqrt{\frac{\pi}{2}} + \sqrt{\frac{\pi}{2}\frac{\mu}{\alpha}} \wedge \frac{1}{|\zeta|} \end{aligned} \tag{B.45}$$

where in the first inequality we used a change of variables and the fact that $a_+/a_- = e^{-\zeta^2/2}e^{\frac{\alpha}{2\mu}(\frac{\mu}{\alpha}\zeta)^2}$, and in the last inequality we used both (B.26) and (B.27), and the fact that $e^{\frac{1}{2}(x^2-\zeta^2)} \le 1$. The rest of (B.38) is proved identically. We now prove (B.40). When $x \in [0, -\zeta]$,

$$\begin{aligned} &\frac{1}{\nu(x)}\int_x^\infty |y|\,\nu(y)dy \\ &= e^{\frac{1}{2}x^2}\int_x^{-\zeta} y e^{-\frac{1}{2}y^2}dy + \frac{a_+}{a_-}e^{\frac{1}{2}x^2}\int_{-\zeta}^\infty y e^{-\frac{\alpha}{2\mu}(y+\zeta-\frac{\mu}{\alpha}\zeta)^2}dy \end{aligned}$$

$$= (1 - e^{\frac{1}{2}(x^2-\zeta^2)}) + e^{\frac{\alpha}{2\mu}(\frac{\mu}{\alpha}\zeta)^2} e^{\frac{1}{2}(x^2-\zeta^2)} \int_{-\zeta}^{\infty} y e^{-\frac{\alpha}{2\mu}(y+\zeta-\frac{\mu}{\alpha}\zeta)^2} dy$$

$$\leq 1 + e^{\frac{\alpha}{2\mu}(\frac{\mu}{\alpha}\zeta)^2} \int_{-\zeta}^{\infty} y e^{-\frac{\alpha}{2\mu}\left((y+\zeta)^2 - 2\frac{\mu}{\alpha}(y+\zeta)\zeta + (\frac{\mu}{\alpha}\zeta)^2\right)} dy$$

$$\leq 1 + \int_{-\zeta}^{\infty} y e^{(y+\zeta)\zeta} dy = 1 + 1 + \frac{1}{\zeta^2},$$

and when $x \geq -\zeta$,

$$\begin{aligned}
\frac{1}{\nu(x)} \int_x^{\infty} |y|\, \nu(y) dy &= e^{\frac{\alpha}{2\mu}(x+\zeta-\frac{\mu}{\alpha}\zeta)^2} \int_x^{\infty} y e^{-\frac{\alpha}{2\mu}(y+\zeta-\frac{\mu}{\alpha}\zeta)^2} dy \\
&= e^{\frac{\alpha}{2\mu}(x+\zeta-\frac{\mu}{\alpha}\zeta)^2} \int_{x+\zeta-\frac{\mu}{\alpha}\zeta}^{\infty} y e^{-\frac{\alpha}{2\mu}y^2} dy \\
&\quad + (1-\mu/\alpha)\,|\zeta|\, e^{\frac{\alpha}{2\mu}(x+\zeta-\frac{\mu}{\alpha}\zeta)^2} \int_{x+\zeta-\frac{\mu}{\alpha}\zeta}^{\infty} e^{-\frac{\alpha}{2\mu}y^2} dy \\
&\leq \frac{\mu}{\alpha} + |\zeta| \frac{1}{\frac{\alpha}{\mu}(x+\zeta-\frac{\mu}{\alpha}\zeta)} \leq \frac{\mu}{\alpha} + 1,
\end{aligned}$$

where in the last inequality we used (B.27). Lastly, the argument for (B.42) is very similar to the chain of inequalities in (B.45), and we leave the details to the reader.

We now describe how to prove Lemma 13. To prove (B.29), we apply Lemma 14 to (B.5) and (B.6) just like in (B.17) of Lemma 3. Using these bounds on $f_h'(x)$, we argue (B.4) just like in the proof of Lemma 3. We now describe how to prove (B.30). When $x \leq 0$, we apply (B.29) and (B.37) to (B.7), and when $x \geq -\zeta$ we apply (B.29) and (B.38) to (B.8). The last region, when $x \in [0, -\zeta]$, has to be handled differently depending on the size of $|\zeta|$. When $|\zeta| \leq 1$, we just apply (B.29) and (B.37) to (B.7). However, when $|\zeta| \geq 1$, we manipulate (B.8) to see that

$$f_h''(x) = -\frac{1}{\nu(x)} \int_x^{-\zeta} \frac{1}{\mu}(-h'(y) + \mu f_h'(y))\nu(y) dy \tag{B.46}$$

$$- \frac{\nu(-\zeta)}{\nu(x)} \frac{1}{\nu(-\zeta)} \int_{-\zeta}^{\infty} \frac{1}{\mu}(-h'(y) + \alpha f_h'(y))\nu(y) dy. \tag{B.47}$$

We then apply (B.29), (B.38), and the fact that $\nu(-\zeta)/\nu(x) \leq 1$ to conclude (B.30). The proof of (B.31) relies on (B.9), which tells us that

$$\left|f_h'''(x)\right| \leq \frac{1}{\mu}\left[1 + \left|f_h''(x)b(x)\right| + \left|f_h'(x)b'(x)\right|\right].$$

Bounding $|f_h'(x)b'(x)|$ only relies on (B.29). The term $|f_h''(x)b(x)|$ is bounded similarly to the way it is done in Lemma 3; see for instance (B.24). Namely, for $x \le 0$ we multiply both sides of (B.7) by $b(x)$ and apply (B.29) with (B.41), and when $x \ge -\zeta$ we multiply both sides of (B.8) by $b(x)$ and apply (B.29) with (B.42). Lastly, when $x \in [0, -\zeta]$, we manipulate (B.8) to get

$$\begin{aligned} b(x)f_h''(x) = & -\frac{b(x)}{\nu(x)}\int_x^{-\zeta}\frac{1}{\mu}(-h'(y)+\mu f_h'(y))\nu(y)dy \\ & -\frac{b(x)\nu(-\zeta)}{\nu(x)b(-\zeta)}\frac{b(-\zeta)}{\nu(-\zeta)}\int_{-\zeta}^{\infty}\frac{1}{\mu}(-h'(y)+\alpha f_h'(y))\nu(y)dy. \end{aligned}$$

We then apply (B.29), (B.42), and the fact that $\left|\frac{b(x)\nu(-\zeta)}{\nu(x)b(-\zeta)}\right| \le 1$ to get the required bounds on $|b(x)f_h''(x)|$ and conclude (B.31). This concludes the proof outline for Lemma 13.

B.2.2. *Proof outline for Lemma 13: The overloaded system.* For the overloaded case in Lemma 13, we again need the following version of Lemma 12.

LEMMA 15. *Consider the Erlang-A model ($\alpha > 0$) with $R \ge n$, and let $\nu(x)$ be the density of $Y(\infty)$. Then*

$$\frac{1}{\nu(x)}\int_{-\infty}^{x}\nu(y)dy \le \begin{cases} \sqrt{\frac{\pi}{2}} \wedge \frac{\mu}{\alpha\zeta}, & x \le -\zeta, \\ \sqrt{\frac{\pi}{2}} + \sqrt{\frac{\pi}{2}\frac{\mu}{\alpha}} \wedge \zeta, & x \in [-\zeta, 0], \end{cases} \tag{B.48}$$

$$\frac{1}{\nu(x)}\int_{x}^{\infty}\nu(y)dy \le \begin{cases} \sqrt{2\pi\frac{\mu}{\alpha}}e^{\frac{\alpha}{2\mu}\zeta^2}, & x \in [-\zeta, 0], \\ \sqrt{\frac{\pi}{2}\frac{\mu}{\alpha}}, & x \ge 0, \end{cases} \tag{B.49}$$

$$\frac{1}{\nu(x)}\int_{-\infty}^{x}|y|\,\nu(y)dy \le \begin{cases} 1 + \sqrt{\frac{\pi}{2}}\zeta \wedge \frac{\mu}{\alpha}, & x \le -\zeta, \\ \frac{\mu}{\alpha} + 1, & x \in [-\zeta, 0], \end{cases} \tag{B.50}$$

$$\frac{1}{\nu(x)}\int_{x}^{\infty}|y|\,\nu(y)dy \le \begin{cases} 2\frac{\mu}{\alpha}e^{\frac{\alpha}{2\mu}\zeta^2}, & x \in [-\zeta, 0], \\ \frac{\mu}{\alpha}, & x \ge 0, \end{cases} \tag{B.51}$$

$$\frac{|b(x)|}{\mu\nu(x)}\int_{-\infty}^{x}\nu(y)dy \le 2, \quad x \le 0, \tag{B.52}$$

$$\frac{|b(x)|}{\mu\nu(x)}\int_{x}^{\infty}\nu(y)dy \le 1, \quad x \ge 0, \tag{B.53}$$

$$\mathbb{E}\,|Y(\infty)| \le \sqrt{\frac{\mu}{\alpha}} + 1. \tag{B.54}$$

To prove this lemma, we first observe that

$$b(x) = \begin{cases} -\mu(x+\zeta) + \alpha\zeta, & x \le -\zeta, \\ -\alpha x, & x \ge -\zeta, \end{cases}$$

and

$$\nu(x) = \begin{cases} a_- e^{-\frac{1}{2}\left(x+\zeta-\frac{\alpha}{\mu}\zeta\right)^2}, & x \le -\zeta, \\ a_+ e^{-\frac{\alpha}{2\mu}x^2}, & x \ge -\zeta, \end{cases} \tag{B.55}$$

where a_- and a_+ are normalizing constants that make $\nu(x)$ continuous and integrate to one. Observe that in the region $x \ge -\zeta$, the density in (B.55) looks very similar to the density in (B.25) in the region $x \le -\zeta$. Hence, one can check that the arguments needed to prove Lemma 15's (B.49), (B.51), and (B.53) are nearly identical to the arguments used to prove Lemma 12's (B.10), (B.12), and (B.14), respectively.

The proof of (B.54) involves applying G_Y to the Lyapunov function $V(x) = x^2$ to see that

$$\begin{aligned} G_Y V(x) = & -2\alpha x^2 1(x > -\zeta) + 2\big(-\mu x^2 + x\zeta(\alpha-\mu)\big)1(x \le -\zeta) + 2\mu \\ \le & -2\alpha x^2 1(x > -\zeta) - 2(\alpha \wedge \mu)x^2 1(x \le -\zeta) + 2\mu. \end{aligned}$$

One can compare this inequality to (B.28) in the proof of Lemma 12 to see that (B.54) follows by the Foster-Lyapunov condition.

We now describe how to prove (B.48), (B.50), and (B.52). To prove (B.48) we use a series of arguments similar to those in (B.45), where we proved (B.38) of Lemma 14. We now prove (B.50). When $x \le -\zeta$,

$$\begin{aligned} \frac{1}{\nu(x)} \int_{-\infty}^{x} |y|\, \nu(y) dy &= e^{\frac{1}{2}(x+\zeta-\frac{\alpha}{\mu}\zeta)^2} \int_{-\infty}^{x} -y e^{-\frac{1}{2}(y+\zeta-\frac{\alpha}{\mu}\zeta)^2} dy \\ &= e^{\frac{1}{2}(x+\zeta-\frac{\alpha}{\mu}\zeta)^2} \int_{-\infty}^{x+\zeta-\frac{\alpha}{\mu}\zeta} -y e^{-\frac{1}{2}y^2} dy \\ &\quad + (1-\alpha/\mu)\zeta e^{\frac{1}{2}(x+\zeta-\frac{\alpha}{\mu}\zeta)^2} \int_{-\infty}^{x+\zeta-\frac{\alpha}{\mu}\zeta} e^{-\frac{1}{2}y^2} dy \\ &\le 1 + \zeta\Big(\sqrt{\frac{\pi}{2}} \wedge \frac{1}{\frac{\alpha}{\mu}\zeta - x - \zeta}\Big) \le 1 + \sqrt{\frac{\pi}{2}}\zeta \wedge \frac{\mu}{\alpha}, \end{aligned} \tag{B.56}$$

where in the last inequality we used both (B.26) and (B.27). For $x \in [-\zeta, 0]$,

$$\frac{1}{\nu(x)} \int_{-\infty}^{x} |y|\, \nu(y) dy$$

$$= \frac{a_-}{a_+} e^{\frac{\alpha}{2\mu}x^2} \int_{-\infty}^{-\zeta} -y e^{-\frac{1}{2}(y+\zeta-\frac{\alpha}{\mu}\zeta)^2} dy + e^{\frac{\alpha}{2\mu}x^2} \int_{-\zeta}^{x} -y e^{\frac{\alpha}{2\mu}y^2} dy.$$

Repeating arguments from (B.56) and using $a_-/a_+ = e^{-\frac{\alpha}{2\mu}\zeta^2} e^{\frac{1}{2}(\frac{\alpha}{\mu}\zeta)^2}$, we can show that the first term above satisfies

$$\frac{a_-}{a_+} e^{\frac{\alpha}{2\mu}x^2} \int_{-\infty}^{-\zeta} -y e^{-\frac{1}{2}(y+\zeta-\frac{\alpha}{\mu}\zeta)^2} dy \le e^{\frac{\alpha}{2\mu}(x^2-\zeta^2)} \Big(1 + \frac{\mu}{\alpha}\Big),$$

and by computing the second term explicitly, we conclude that

$$\begin{aligned} \frac{1}{\nu(x)} \int_{-\infty}^{x} |y|\, \nu(y) dy &\le e^{\frac{\alpha}{2\mu}(x^2-\zeta^2)} \Big(1 + \frac{\mu}{\alpha}\Big) + \frac{\mu}{\alpha}\big(1 - e^{\frac{\alpha}{2\mu}(x^2-\zeta^2)}\big) \\ &\le 1 + \frac{\mu}{\alpha}, \end{aligned}$$

which proves (B.50). Lastly, it is not hard to see that (B.52) follows from a straightforward application of (B.27).

Having argued Lemma 15, we now use it to prove the bounds in (B.32)–(B.36). To prove (B.32), we apply Lemma 15 to (B.5) and (B.6) just like in (B.17) of Lemma 3. Using these bounds on $f_h'(x)$, we argue (B.4) just like in the proof of Lemma 3. We now describe how to prove (B.33). When $x \le -\zeta$, we apply (B.32) and (B.48) to (B.7). When $x \ge -\zeta$, instead of using the expressions for $f_h''(x)$ in (B.7) and (B.8) like we would usually do, we instead apply (B.32) to the bound

$$\big|f_h''(x)\big| \le \frac{1}{\mu}\big|f_h'(x)\big|\,|b(x)| + \frac{1}{\mu}\big(\,|x| + \mathbb{E}\,|Y(\infty)|\,\big), \quad x \in \mathbb{R},$$

which follows by rewriting the Poisson equation (3.3) and using the Lipschitz property of $h(x)$. We now prove (B.34)–(B.36). We recall (B.9) to see that

$$\big|f_h'''(x)\big| \le \frac{1}{\mu}\big[1 + \big|f_h''(x)b(x)\big| + \big|f_h'(x)b'(x)\big|\,\big].$$

Bounding $|f_h'(x)b'(x)|$ is simple, and only relies on (B.32). The other term, $|f_h''(x)b(x)|$, is bounded as follows. To prove (B.34), i.e. when $x \le -\zeta$, we multiply both sides of (B.7) by $b(x)$, and apply (B.32) and (B.52) to the result. When $x \ge -\zeta$ then

$$\big|f_h''(x)b(x)\big| = \alpha\,|x|\,\big|f_h''(x)\big|,$$

and the difference between (B.35) and (B.36) lies in the way that the quantity above is bounded. To get (B.35), we simply apply the bounds on $f_h''(x)$ from (B.33) to the right hand side above.

To prove (B.36), we will first argue that

(B.57)
$$\left|f_h'''(x)\right| \le \begin{cases} \frac{C}{\mu}\left(\frac{\alpha}{\mu}+\sqrt{\frac{\alpha}{\mu}}+1\right)+\frac{C}{\mu}\left(\frac{\alpha}{\mu}+\sqrt{\frac{\alpha}{\mu}}+1\right)^2|x|, & x\in[-\zeta,0], \\ \frac{C}{\mu}\left(\frac{\alpha}{\mu}+\sqrt{\frac{\alpha}{\mu}}+1\right), & x\ge 0, \end{cases}$$

where C is some positive constant independent of everything else; this will imply (B.36). The only difference between the proof of (B.57) and the bound on $f_h'''(x)$ in (B.35) is in how $|f_h''(x)b(x)|$ is bounded; we now describe the different way to bound $|f_h''(x)b(x)|$. When $x \ge 0$, we multiply both sides of (B.8) by $b(x)$ and use the bounds in (B.32) and (B.53) to bound $|f_h''(x)b(x)|$. When $x \in [-\zeta, 0]$, we want to prove that

$$\left|f_h''(x)\right| \le \frac{C}{\mu}\left(\frac{\alpha}{\mu}+\sqrt{\frac{\alpha}{\mu}}+1\right)\left(1+\sqrt{\frac{\mu}{\alpha}}\right)+\frac{C}{\mu}\left(\zeta\wedge\frac{\mu^2}{\alpha^2\zeta}\right), \tag{B.58}$$

which, after considering separately the cases when $\zeta \le \mu/\alpha$ and $\zeta \ge \mu/\alpha$, implies that

$$\left|f_h''(x)\right| \le \frac{C}{\mu}\left(\frac{\alpha}{\mu}+\sqrt{\frac{\alpha}{\mu}}+1\right)\left(1+\sqrt{\frac{\mu}{\alpha}}\right)+\frac{C}{\alpha}.$$

We can then use this fact to bound $|f_h''(x)b(x)| = \alpha|x|\,|f_h''(x)|$. To prove (B.58) for $\zeta \le \sqrt{\mu/\alpha}$, we bound (B.8) using (B.32) and (B.49). To prove (B.58) for $\zeta \ge \sqrt{\mu/\alpha}$, we bound (B.7) using (B.32) and (B.48). We point out that to bound (B.7) we need to perform a manipulation similar to the one in (B.47). This concludes the proof outline for the overloaded case.

B.3. Kolmogorov gradient bounds: Proof of Lemmas 4 and 5. For the remainder of this section, we take $\mathcal{H} = \mathcal{H}_K$ in (3.3). With this choice of test functions, any solution to the Poisson equation will have a discontinuity in its second derivative, which makes the gradient bounds for it differ from the Wasserstein setting. Fix $a \in \mathbb{R}$ and consider the Poisson equation

$$G_Y f_a(x) = b(x)f_a'(x) + \mu f_a''(x) = F_Y(a) - 1_{(-\infty,a]}(x),$$

where $F_Y(x)$ is the distribution function of $Y(\infty)$. If $f_a(x)$ is a solution the Poisson equation with $a_2 = 0$, then just as in (B.1) and (B.2),

$$f_a'(x) = \frac{1}{\mu\nu(x)}\int_{-\infty}^{x}\left(F_Y(a) - 1_{(-\infty,a]}(y)\right)\nu(y)dy,$$

$$f_a'(x) = -\frac{1}{\mu\nu(x)}\int_x^\infty \big(F_Y(a) - 1_{(-\infty,a]}(y)\big)\nu(y)dy,$$

which immediately implies that

$$\big|f_a'(x)\big| \le \frac{1}{\mu\nu(x)}\min\Big\{\int_{-\infty}^x \nu(y)dy, \int_x^\infty \nu(y)dy\Big\}.$$

Furthermore,

$$f_a''(x) = \frac{1}{\mu}\big[F_Y(a) - 1_{(-\infty,a]}(x) - b(x)f_a'(x)\big].$$

We now prove the Kolmogorov gradient bounds for the Erlang-C model.

Proof of Lemma 4. First of all, by (B.10) and (B.11),

$$\mu\big|f_a'(x)\big| \le \begin{cases} \sqrt{\frac{\pi}{2}}, & x \le 0, \\ \min\big\{\sqrt{2\pi}e^{\frac{1}{2}\zeta^2}, \sqrt{\frac{\pi}{2}} + \frac{1}{|\zeta|}\big\}, & x \in [0, -\zeta], \\ \frac{1}{|\zeta|}, & x \ge -\zeta, \end{cases} \tag{B.59}$$

and (B.22) implies that

$$\min\Big\{\sqrt{2\pi}e^{\frac{1}{2}\zeta^2}, \sqrt{\frac{\pi}{2}} + \frac{1}{|\zeta|}\Big\} \le 5,$$

which proves the bounds for $f_a'(x)$. Second, (B.14) and (B.15) imply that for all $x \in \mathbb{R}$,

$$\big|f_a''(x)\big| \le \frac{1}{\mu}\left[1 + \frac{|b(x)|}{\mu\nu(x)}\min\Big\{\int_{-\infty}^x \nu(y)dy, \int_x^\infty \nu(y)dy\Big\}\right] \le 3/\mu, \tag{B.60}$$

where $f_a''(x)$ is understood to be the left derivative at the point $x = a$. □

Proof of Lemma 5. The proof of this lemma is almost identical to the proof of Lemma 4. By using the analogues of (B.14) and (B.15) from Lemmas 14 and 15, its not hard to check that (B.60) holds for the Erlang-A model as well. To prove the bounds on $f_a'(x)$, we obtain inequalities similar to (B.59) by using analogues of (B.10) and (B.11) from Lemmas 14 and 15. These inequalities will imply (5.3) and (5.4) once we consider in them separately the cases when $|\zeta| \le 1$ and $|\zeta| \ge 1$. □

APPENDIX C: PROOF OUTLINES OF ERLANG-A THEOREMS

Sections C.1 and C.2 contain an outline for the proofs of Theorems 2 and 4, respectively.

C.1. Outline for Theorem 2. Proving Theorem 2 consists of bounding the four error terms in (3.10). Since the procedure is very similar to the proof of Theorem 1, we will only outline which gradient and moment bounds need to be used to bound each error term.

We start with the underloaded case, when $R \leq n$. To bound the first term in (3.10), we use moment bounds (A.9), (A.10), and (A.13), together with the gradient bounds in (B.30). For the second and third terms, we use moment bound (A.15) and the gradient bounds in (B.31). For the fourth term, we use moment bounds (A.9)–(A.13), and the gradient bounds in (B.31).

We now prove the overloaded case, when $R \geq n$. To bound the first term in (3.10), we use moment bounds (A.16)–(A.23), together with the gradient bounds in (B.33). For the second and third terms, we use moment bounds (A.18),(A.19), and (A.24), together with the gradient bounds in (B.34) and (B.35). For the fourth term, we use moment bounds (A.16)–(A.23), and gradient bounds in (B.34) and (B.36).

C.2. Outline for Theorem 4. The proof of this theorem is nearly identical to the proof of Theorem 3. Therefore, we only outline the key steps and differences. The goal is to obtain a version of (5.11), from which the theorem follows by applying Lemmas 8 and 9. To get a version of (5.11), we bound each of the terms in (5.8), just like we did in the proof of Theorem 3. The proof varies between the underloaded and overloaded cases.

We begin with the underloaded case ($1 \leq R \leq n$). To bound the first term in (5.8), we use moment bounds (A.9), (A.11), and (A.13), together with gradient bound (5.5). For the second and third terms in (5.8) we use the gradient bound in (5.3). For the fourth error term, we use gradient bound (5.3), and moment bounds (A.8), (A.11), and

$$\begin{aligned}
&\mathbb{E}\Big[\big(b(\tilde{X}(\infty))\big)^2 1(\tilde{X}(\infty) \geq -\zeta)\Big] \\
&= \alpha^2 \mathbb{E}\Big[\big(\tilde{X}(\infty)+\zeta\big)^2 1(\tilde{X}(\infty) \geq -\zeta)\Big] + \mu^2\zeta^2 \mathbb{P}(\tilde{X}(\infty) \geq -\zeta) \\
&\quad + 2\alpha\mu\,|\zeta|\,\mathbb{E}\Big[(\tilde{X}(\infty)+\zeta) 1(\tilde{X}(\infty) \geq -\zeta)\Big] \\
&\leq \alpha^2 \frac{1}{3}\Big(\frac{\mu}{\alpha}\delta^2 + \frac{\mu}{\alpha}4 + \delta^2\Big) + \mu^2\zeta^2 \mathbb{P}(\tilde{X}(\infty) \geq -\zeta) \\
&\quad + 2\alpha\mu\Big(\frac{\delta^2}{4}\frac{\alpha}{\mu} + \frac{\delta^2}{4} + 1\Big),
\end{aligned}$$

where the last inequality follows from moment bounds (A.12) and (A.14).

In the overloaded case ($n \leq R$), to bound the first term in (5.8) we use moment bounds (A.16), (A.19), and (A.22) with gradient bound (5.5). To

bound the second and third terms in (5.8) we use gradient bound (5.4). To bound the fourth term in (5.8), we use gradient bound (5.5), with moment bounds (A.18) and

$$\begin{aligned}
&\mathbb{E}\Big[\big(b(\tilde{X}(\infty))\big)^2 1(\tilde{X}(\infty) \le -\zeta)\Big] \\
&= \mu^2 \mathbb{E}\Big[\big(\tilde{X}(\infty)+\zeta\big)^2 1(\tilde{X}(\infty) \le -\zeta)\Big] + \alpha^2\zeta^2 \mathbb{P}(\tilde{X}(\infty) \le -\zeta) \\
&\quad + 2\alpha\mu\zeta \mathbb{E}\Big[\big|(\tilde{X}(\infty)+\zeta) 1(\tilde{X}(\infty) \le -\zeta)\big|\Big] \\
&\le \mu^2\Big(\frac{\delta^2}{4}\frac{\alpha}{\mu}+1\Big) + \alpha^2\Big(\frac{\delta^2}{4}+\frac{\mu}{\alpha}\Big) + 2\alpha\mu\Big(\frac{\delta^2}{4}+1\Big),
\end{aligned}$$

where the last inequality follows from moment bounds (A.17), (A.20), and (A.21).

APPENDIX D: MISCELLANEOUS LEMMAS

This appendix proves Lemmas 1, 6, 7, and 10.

D.1. Proof of Lemma 1.

PROOF OF LEMMA 1. Let $f(x) : \mathbb{R} \to \mathbb{R}$ satisfy $|f(x)| \le C(1+x)^2$. A sufficient condition to ensure that

$$\mathbb{E}\big[G_{\tilde{X}} f(\tilde{X}(\infty))\big] = 0$$

is given by [36, Proposition 1.1] (alternatively, see [26, Proposition 3]). Namely, we require that

$$\mathbb{E}\Big[\big|G_{\tilde{X}}(\tilde{X}(\infty), \tilde{X}(\infty)) f(\tilde{X}(\infty))\big|\Big] < \infty, \tag{D.1}$$

where $G_{\tilde{X}}(x,x)$ is the diagonal entry of the generator matrix $G_{\tilde{X}}$ corresponding to state x.

We begin with the Erlang-C model, where the transition rates of the CTMC are bounded by $\lambda + n\mu$. Since $|f(x)| \le C(1+x)^2$, it suffices to show that $\mathbb{E}(\tilde{X}(\infty))^2 < \infty$, or that $\mathbb{E}(X(\infty))^2 < \infty$, where $X(\infty)$ has the stationary distribution of the CTMC X. Consider the function $V(k) = k^3$, where $k \in \mathbb{Z}_+$. Let G_X be the generator of X, which is a simple birth death process with constant birth rate λ and departure rate $\mu(k \wedge n)$ in state $k \in \mathbb{Z}_+$. Then for $k \ge n$,

$$G_X V(k) = \lambda((k+1)^3 - k^3) + n\mu((k-1)^3 - k^3)$$

$$
\begin{aligned}
&= \lambda(3k^2+3k+1) + n\mu(-3k^2+3k-1) \\
&= -3k^2(n\mu-\lambda) + 3k(\lambda+n\mu) + (\lambda-n\mu). \qquad \text{(D.2)}
\end{aligned}
$$

It is not hard to see that there exists some $k_0 \in \mathbb{Z}_+$, and a constant $c > 0$ (that depends on λ, n, and μ), such that for all $k \geq k_0$,

$$
-3k^2(n\mu-\lambda) + 3k(\lambda+n\mu) \leq -ck^2. \qquad \text{(D.3)}
$$

We combine (D.2)–(D.3) to conclude that there exists some constant $d > 0$ (that depends on λ, n, and μ) satisfying

$$
G_X V(x) \leq -cx^2 + d1(k < (k_0 \vee n)),
$$

and invoking [47, Theorem 4.3], we see that $\mathbb{E}(X(\infty))^2 < \infty$.

The case of the Erlang-A model is not very different. When $\alpha > 0$, the transition rates of the CTMC depend linearly on its state. Hence, to satisfy (D.1) we need to show that $\mathbb{E}(X(\infty))^3 < \infty$. This is readily proven by repeating the procedure above with the Lyapunov function $V(k) = k^4$, and we omit the details. □

D.2. Proof of Lemma 6.

Proof of Lemma 6. We let $F_W(w)$ and $F_{\tilde{X}}(x)$ be the distribution functions of W and $\tilde{X}(\infty)$, respectively. For any $a \in \mathbb{R}$, let $\tilde{a} = \delta(a - x(\infty))$. We want to show that

$$
\begin{aligned}
\mathbb{P}(\tilde{a}-\delta < \tilde{X}(\infty) \leq \tilde{a}+\delta) &= F_{\tilde{X}}(\tilde{a}+\delta) - F_{\tilde{X}}(\tilde{a}-\delta) \\
&\leq 2\delta\omega(F_W) + d_K(\tilde{X}(\infty), W) + 9\delta^2 + 8\delta^4. \qquad \text{(D.4)}
\end{aligned}
$$

Define

$$
k^* = \inf\{k \geq 0 : \nu_k \geq \nu_j, \text{ for all } j \neq k\}.
$$

Then for any $\tilde{a} \in \mathbb{R}$,

$$
F_{\tilde{X}}(\tilde{a}+\delta) - F_{\tilde{X}}(\tilde{a}-\delta) \leq 2\nu_{k^*},
$$

because $\tilde{X}(\infty)$ takes at most two values in the interval $(\tilde{a}-\delta, \tilde{a}+\delta]$. Observe that by the flow balance equations, we know that for any $k \in \mathbb{Z}_+$,

$$
\nu_k = \frac{d(k+1)}{\lambda}\nu_{k+1}.
$$

Since k^* is the maximizer of $\{\nu_k\}$, we know that

$$d(k^*) \le \lambda \le d(k^*+1) \le \lambda + \mu,$$

where in the last inequality we have used the fact that the increase in departure rate between state k^* and k^*+1 is at most μ. Likewise, $d(k^*+i) \le \lambda + i\mu$ for $i = 2, 3$. Hence,

$$\begin{aligned}
\nu_{k^*} &= \frac{d(k^*+1)}{\lambda}\nu_{k^*+1} \le \Big(1+\frac{\mu}{\lambda}\Big)\nu_{k^*+1} \le \nu_{k^*+1} + \delta^2,\\
\nu_{k^*} &= \frac{d(k^*+1)}{\lambda}\frac{d(k^*+2)}{\lambda}\nu_{k^*+2}\\
&\le (1+\delta^2)(1+2\delta^2)\nu_{k^*+2} \le \nu_{k^*+2} + 3\delta^2 + 2\delta^4,\\
\nu_{k^*+1} &= \frac{d(k^*+2)}{\lambda}\frac{d(k^*+3)}{\lambda}\nu_{k^*+3}\\
&\le (1+2\delta^2)(1+3\delta^2)\nu_{k^*+3} \le \nu_{k^*+3} + 5\delta^2 + 6\delta^4,
\end{aligned}$$

which implies that for any $\tilde{a} \in \mathbb{R}$,

$$\begin{aligned}
F_{\tilde{X}}(\tilde{a}+\delta) - F_{\tilde{X}}(\tilde{a}-\delta) \le 2\nu_{k^*} &\le \nu_{k^*} + \nu_{k^*+1} + \delta^2\\
&= F_{\tilde{X}}(\tilde{k}^*+\delta) - F_{\tilde{X}}(\tilde{k}^*-\delta) + \delta^2.
\end{aligned}$$

There are now 4 cases to consider, with the first three being simple to handle. Recall that $\omega(F_W)$ is the modulus of continuity of $F_W(w)$.

1. If $F_W(\tilde{k}^*-\delta) \le F_{\tilde{X}}(\tilde{k}^*-\delta)$ and $F_W(\tilde{k}^*+\delta) \ge F_{\tilde{X}}(\tilde{k}^*+\delta)$, then

   (D.5)
   $$F_{\tilde{X}}(\tilde{k}^*+\delta) - F_{\tilde{X}}(\tilde{k}^*-\delta) \le F_W(\tilde{k}^*+\delta) - F_W(\tilde{k}^*-\delta) \le 2\delta\omega(F_W).$$

2. If $F_W(\tilde{k}^*-\delta) \le F_{\tilde{X}}(\tilde{k}^*-\delta)$ but $F_W(\tilde{k}^*+\delta) < F_{\tilde{X}}(\tilde{k}^*+\delta)$, then

   $$\begin{aligned}
   &F_{\tilde{X}}(\tilde{k}^*+\delta) - F_{\tilde{X}}(\tilde{k}^*-\delta)\\
   &\le F_{\tilde{X}}(\tilde{k}^*+\delta) - F_W(\tilde{k}^*+\delta) + F_W(\tilde{k}^*+\delta) - F_W(\tilde{k}^*-\delta)\\
   &\le 2\delta\omega(F_W) + d_K(\tilde{X}(\infty), W). \qquad \text{(D.6)}
   \end{aligned}$$

3. Similarly, if $F_W(\tilde{k}^*-\delta) > F_{\tilde{X}}(\tilde{k}^*-\delta)$ and $F_W(\tilde{k}^*+\delta) \ge F_{\tilde{X}}(\tilde{k}^*+\delta)$, then

   $$\begin{aligned}
   &F_{\tilde{X}}(\tilde{k}^*+\delta) - F_{\tilde{X}}(\tilde{k}^*-\delta)\\
   &\le F_W(\tilde{k}^*+\delta) - F_W(\tilde{k}^*-\delta) + F_W(\tilde{k}^*-\delta) - F_{\tilde{X}}(\tilde{k}^*-\delta)\\
   &\le 2\delta\omega(F_W) + d_K(\tilde{X}(\infty), W). \qquad \text{(D.7)}
   \end{aligned}$$

4. Suppose $F_W(\tilde{k}^* - \delta) > F_{\tilde{X}}(\tilde{k}^* - \delta)$ and $F_W(\tilde{k}^* + \delta) < F_{\tilde{X}}(\tilde{k}^* + \delta)$, then we need to use a different approach. We know that

$$\begin{aligned} F_{\tilde{X}}(\tilde{k}^* + \delta) - F_{\tilde{X}}(\tilde{k}^* - \delta) &= \nu_{k^*} + \nu_{k^*+1} \\ &\le \nu_{k^*+2} + \nu_{k^*+3} + 8\delta^2 + 8\delta^4 \\ &= F_{\tilde{X}}(\tilde{k}^* + 3\delta) - F_{\tilde{X}}(\tilde{k}^* + \delta) + 8\delta^2 + 8\delta^4. \end{aligned}$$

Since $F_W(\tilde{k}^* + \delta) \le F_{\tilde{X}}(\tilde{k}^* + \delta)$, we are either in case 1 or 2 for $F_{\tilde{X}}(\tilde{k}^* + 3\delta) - F_{\tilde{X}}(\tilde{k}^* + \delta)$, and hence we have

$$F_{\tilde{X}}(\tilde{k}^* + 3\delta) - F_{\tilde{X}}(\tilde{k}^* + \delta) \le 2\delta\omega(F_W) + d_K(\tilde{X}(\infty), W).$$

This proves (D.4), concluding the proof of this lemma. □

D.3. Proof of Lemma 7 .

Proof of Lemma 7. In the Erlang-C model,

$$\nu(x) = \begin{cases} a_- e^{-\frac{1}{2}x^2}, & x \le -\zeta, \\ a_+ e^{-|\zeta|x}, & x \ge -\zeta. \end{cases} \tag{D.8}$$

To bound this density, we need to bound a_- and a_+. We know that $\nu(x)$ must integrate to one, which implies that

$$a_- \int_{-\infty}^{-\zeta} e^{-\frac{1}{2}y^2} dy + a_+ \int_{-\zeta}^{\infty} e^{-|\zeta|y} dy = 1$$

Furthermore, since $\nu(x)$ is continuous at $x = -\zeta$,

$$a_- e^{-\frac{1}{2}\zeta^2} = a_+ e^{-\zeta^2}.$$

Combining these two facts, we see that

$$a_- = \frac{1}{\int_{-\infty}^{-\zeta} e^{-\frac{1}{2}y^2} dy + e^{\frac{1}{2}\zeta^2} \int_{-\zeta}^{\infty} e^{-|\zeta|y} dy} \le \frac{1}{\int_{-\infty}^{0} e^{-\frac{1}{2}y^2} dy} = \sqrt{\frac{2}{\pi}}, \tag{D.9}$$

and

$$a_+ = \frac{1}{e^{-\frac{1}{2}\zeta^2} \int_{-\infty}^{-\zeta} e^{-\frac{1}{2}y^2} dy + \int_{-\zeta}^{\infty} e^{-|\zeta|y} dy} \le \frac{1}{e^{-\frac{1}{2}\zeta^2} \int_{-\infty}^{0} e^{-\frac{1}{2}y^2} dy} = e^{\frac{1}{2}\zeta^2} \sqrt{\frac{2}{\pi}}. \tag{D.10}$$

Therefore, for $x \le -\zeta$,

$$|\nu(x)| \le a_- \le \sqrt{\frac{2}{\pi}},$$

and for $x \ge -\zeta$, we recall that $\zeta < 0$ to see that

$$|\nu(x)| \le a_+ e^{-|\zeta|x} \le \sqrt{\frac{2}{\pi}} e^{\frac{1}{2}\zeta^2} e^{-|\zeta|x} \le \sqrt{\frac{2}{\pi}}.$$

□

D.4. Proof of Lemma 10.

PROOF OF LEMMA 10. The density of $Y(\infty)$ is given in (D.8), and so

$$\mathbb{E}(Y(\infty))^m = a_- \int_{-\infty}^{-\zeta} y^m e^{-\frac{1}{2}y^2} dy + a_+ \int_{-\zeta}^{\infty} y^m e^{-|\zeta|y} dy,$$

where a_- and a_+ are as in (D.9) and (D.10). In particular,

$$a_- = \frac{1}{\int_{-\infty}^{-\zeta} e^{-\frac{1}{2}y^2} dy + e^{\frac{1}{2}\zeta^2} \int_{-\zeta}^{\infty} e^{-|\zeta|y} dy} = \frac{1}{\int_{-\infty}^{-\zeta} e^{-\frac{1}{2}y^2} dy + \frac{1}{|\zeta|} e^{-\frac{1}{2}\zeta^2}},$$

which implies that

$$\lim_{\zeta \uparrow 0} |\zeta|^m a_- \int_{-\infty}^{-\zeta} y^m e^{-\frac{1}{2}y^2} dy = 0.$$

Furthermore,

$$a_+ = \frac{1}{e^{-\frac{1}{2}\zeta^2} \int_{-\infty}^{-\zeta} e^{-\frac{1}{2}y^2} dy + \int_{-\zeta}^{\infty} e^{-|\zeta|y} dy} = \frac{1}{e^{-\frac{1}{2}\zeta^2} \int_{-\infty}^{-\zeta} e^{-\frac{1}{2}y^2} dy + \frac{1}{|\zeta|} e^{-\zeta^2}},$$

and using integration by parts,

$$\begin{aligned} \int_{-\zeta}^{\infty} y^m e^{-|\zeta|y} dy &= e^{-\zeta^2} \sum_{j=0}^{m} \frac{m!}{(m-j)!} \frac{1}{|\zeta|^{j+1}} |\zeta|^{m-j} \\ &= e^{-\zeta^2} \sum_{j=0}^{m-1} \frac{m!}{(m-j)!} \frac{1}{|\zeta|^{j+1}} |\zeta|^{m-j} + e^{-\zeta^2} \frac{m!}{|\zeta|^{m+1}}. \end{aligned}$$

Hence,

$$\lim_{\zeta \uparrow 0} |\zeta|^m a_+ \int_{-\zeta}^{\infty} y^m e^{-|\zeta|y} dy = m!.$$

□

REFERENCES

[1] Atar, R. (2012). A diffusion regime with nondegenerate slowdown. *Operations Research*, **60** 490–500. URL http://dx.doi.org/10.1287/opre.1110.1030. MR2935073

[2] Barbour, A. (1990). Stein's method for diffusion approximations. *Probability Theory and Related Fields*, **84** 297–322. URL http://dx.doi.org/10.1007/BF01197887. MR1035659

[3] Barbour, A. and Brown, T. (1992). Stein's method and point process approximation. *Stochastic Processes and their Applications*, **43** 9 – 31. URL http://dx.doi.org/10.1016/0304-4149(92)90073-Y. MR1190904

[4] Barbour, A. and Xia, A. (2006). On Stein's factors for Poisson approximation in Wasserstein distance. *Bernoulli*, **12** 943–954. URL http://dx.doi.org/10.3150/bj/1165269145.

[5] Barbour, A. D. (1988). Stein's method and Poisson process convergence. *Journal of Applied Probability*, **25** pp. 175–184. URL http://www.jstor.org/stable/3214155. MR0974580

[6] Blanchet, J. and Glynn, P. (2007). Uniform renewal theory with applications to expansions of random geometric sums. *Advances in Applied Probability*, **39** 1070–1097. URL https://doi.org/10.1017/S000186780000224X. MR2381589

[7] Borovkov, A. (1964). Some limit theorems in the theory of mass service, I. *Theory of Probability and its Applications*, **9** 550–565. URL http://dx.doi.org/10.1137/1109078.

[8] Borovkov, A. (1965). Some limit theorems in the theory of mass service, II. *Theory of Probability and its Applications*, **10** 375–400. URL http://dx.doi.org/10.1137/1110046.

[9] Bramson, M. (1998). State space collapse with application to heavy traffic limits for multiclass queueing networks. *Queueing Systems*, **30** 89–140. URL http://dx.doi.org/10.1023/A:1019160803783. MR1663763

[10] Braverman, A. and Dai, J. G. (2017). Stein's method for steady-state diffusion approximations of $M/Ph/n+M$ systems. *Ann. Appl. Probab.* URL http://arxiv.org/abs/1503.00774.

[11] Brown, T. C. and Xia, A. (2001). Stein's method and birth-death processes. *Ann. Probab.*, **29** 1373–1403. URL http://dx.doi.org/10.1214/aop/1015345606.

[12] Budhiraja, A. and Lee, C. (2009). Stationary distribution convergence for generalized Jackson networks in heavy traffic. *Mathematics of Operations Research*, **34** 45–56. URL http://dx.doi.org/10.1287/moor.1080.0353.

[13] Chatterjee, S. (2014). A short survey of Stein's method. To appear in Proceedings of ICM 2014, URL http://arxiv.org/abs/1404.1392.

[14] Chen, L. H. Y. (1975). Poisson approximation for dependent trials. *Ann. Probab.*, **3** 534–545. URL http://dx.doi.org/10.1214/aop/1176996359.

[15] Chen, L. H. Y., Goldstein, L. and Shao, Q.-M. (2011). *Normal approximation by Stein's method*. Probability and its Applications (New York), Springer, Heidelberg. URL http://dx.doi.org/10.1007/978-3-642-15007-4.

[16] Dai, J. G. and He, S. (2013). Many-server queues with customer abandonment: Numerical analysis of their diffusion model. *Stochastic Systems*, **3** 96–146. URL http://dx.doi.org/10.1214/11-SSY029.

[17] Dai, J. G., He, S. and Tezcan, T. (2010). Many-server diffusion limits for $G/Ph/n+GI$ queues. *Annals of Applied Probability*, **20** 1854–1890. URL http://projecteuclid.org/euclid.aoap/1282747403.

[18] Dai, J. G. and Lin, W. (2008). Asymptotic optimality of maximum pressure policies in stochastic processing networks. *Annals of Applied Probability*, **18** 2239–2299. URL http://projecteuclid.org/euclid.aoap/1227708918.

[19] Ehm, W. (1991). Binomial approximation to the Poisson binomial distribution. *Statistics & Probability Letters*, **11** 7 – 16. URL http://www.sciencedirect.com/science/article/pii/016771529190170V.

[20] Ethier, S. N. and Kurtz, T. G. (1986). *Markov Processes: Characterization and Convergence*. Wiley, New York. URL http://onlinelibrary.wiley.com/book/10.1002/9780470316658.

[21] Gamarnik, D. and Stolyar, A. L. (2012). Multiclass multiserver queueing system in the Halfin-Whitt heavy traffic regime: asymptotics of the stationary distribution. *Queueing Systems*, **71** 25–51. URL http://dl.acm.org/citation.cfm?id=2339029.

[22] Gamarnik, D. and Zeevi, A. (2006). Validity of heavy traffic steady-state approximation in generalized Jackson networks. *Ann. Appl. Probab.*, **16** 56–90. URL http://projecteuclid.org/euclid.aoap/1141654281. MR2209336

[23] Gan, H. and Xia, A. (2015). Stein's method for conditional compound Poisson approximation. *Statistics & Probability Letters*, **100** 19 – 26. URL http://www.sciencedirect.com/science/article/pii/S0167715215000486.

[24] Gans, N., Koole, G. and Mandelbaum, A. (2003). Telephone call centers: Tutorial, review, and research prospects. *Manufacturing & Service Operations Management*, **5** 79–141. http://msom.journal.informs.org/cgi/reprint/5/2/79.pdf, URL http://dx.doi.org/10.1287/msom.5.2.79.160719.

[25] Gibbs, A. L. and Su, F. E. (2002). On choosing and bounding probability metrics. *International Statistical Review / Revue Internationale de Statistique*, **70** pp. 419–435. URL http://www.jstor.org/stable/1403865.

[26] Glynn, P. W. and Zeevi, A. (2008). Bounding stationary expectations of Markov processes. In *Markov processes and related topics: a Festschrift for Thomas G. Kurtz*, vol. 4 of *Inst. Math. Stat. Collect.* Inst. Math. Statist., Beachwood, OH, 195–214. URL http://dx.doi.org/10.1214/074921708000000381.

[27] Götze, F. (1991). On the rate of convergence in the multivariate CLT. *Ann. Probab.*, **19** 724–739. URL http://dx.doi.org/10.1214/aop/1176990448.

[28] Gurvich, I. (2014). Diffusion models and steady-state approximations for exponentially ergodic Markovian queues. *The Annals of Applied Probability*, **24** 2527–2559. URL http://dx.doi.org/10.1214/13-AAP984.

[29] Gurvich, I. (2014). Validity of heavy-traffic steady-state approximations in multiclass queueing networks: the case of queue-ratio disciplines. *Mathematics of Operations Research*, **39** 121–162. URL http://dx.doi.org/10.1287/moor.2013.0593.

[30] Gurvich, I. and Huang, J. (2016). Beyond heavy-traffic regimes: universal bounds and controls for the single-server queue. Working paper, URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2784752.

[31] Gurvich, I., Huang, J. and Mandelbaum, A. (2014). Excursion-based universal approximations for the Erlang-A queue in steady-state. *Mathematics of Operations Research*, **39** 325–373. URL http://dx.doi.org/10.1287/moor.2013.0606.

[32] Halfin, S. and Whitt, W. (1981). Heavy-traffic limits for queues with many exponential servers. *Oper. Res.*, **29** 567–588. URL http://www.jstor.org/stable/170115.

[33] Harrison, J. M. (1978). The diffusion approximation for tandem queues in heavy traffic. *Advances in Applied Probability*, **10** 886–905. URL http://www.jstor.org/stable/1426665.

[34] Harrison, J. M. and Nguyen, V. (1993). Brownian models of multiclass queue-

ing networks: Current status and open problems. *Queueing Systems: Theory and Applications*, **13** 5–40. URL http://dx.doi.org/10.1007/BF01158927.

[35] Harrison, J. M. and Williams, R. J. (1987). Brownian models of open queueing networks with homogeneous customer populations. *Stochastics*, **22** 77–115. URL http://dx.doi.org/10.1080/17442508708833469.

[36] Henderson, S. G. (1997). *Variance reduction via an approximating Markov process*. Ph.D. thesis, Department of Operations Research, Stanford University. http://people.orie.cornell.edu/shane/pubs/thesis.pdf.

[37] Iglehart, D. L. and Whitt, W. (1970). Multiple channel queues in heavy traffic I. *Advances in Applied Probability*, **2** 150–177. URL http://www.jstor.org/stable/3518347.

[38] Iglehart, D. L. and Whitt, W. (1970). Multiple channel queues in heavy traffic II: sequences, networks, and batches. *Advances in Applied Probability*, **2** 355–369. URL http://www.jstor.org/stable/1426324.

[39] Janssen, A. J. E. M., van Leeuwaarden, J. S. H. and Zwart, B. (2008). Corrected asymptotics for a multi-server queue in the Halfin-Whitt regime. *Queueing Syst.*, **58** 261–301. URL http://dx.doi.org/10.1007/s11134-008-9070-0.

[40] Janssen, A. J. E. M., van Leeuwaarden, J. S. H. and Zwart, B. (2008). Gaussian expansions and bounds for the Poisson distribution applied to the Erlang B formula. *Adv. in Appl. Probab.*, **40** 122–143. URL http://dx.doi.org/10.1239/aap/1208358889.

[41] Janssen, A. J. E. M., van Leeuwaarden, J. S. H. and Zwart, B. (2011). Refining square-root safety staffing by expanding Erlang C. *Operations Research*, **59** 1512–1522. URL http://dx.doi.org/10.1287/opre.1110.0991.

[42] Kang, W., Kelly, F., Lee, N. and Williams, R. (2009). State space collapse and diffusion approximation for a network operating under a fair bandwidth sharing policy. *The Annals of Applied Probability*, **19** 1719–1780. URL http://www.jstor.org/stable/25662521. MR2569806

[43] Katsuda, T. (2010). State-space collapse in stationarity and its application to a multiclass single-server queue in heavy traffic. *Queueing Systems: Theory and Applications*, **65** 237–273. URL http://dx.doi.org/10.1007/s11134-010-9178-x.

[44] Kusuoka, S. and Tudor, C. A. (2012). Stein's method for invariant measures of diffusions via malliavin calculus. *Stochastic Processes and their Applications*, **122** 1627 – 1651. URL http://www.sciencedirect.com/science/article/pii/S0304414912000270.

[45] Lindvall, T. (1992). *Lectures on the coupling method*. Wiley series in probability and mathematical statistics, Wiley, New York. A Wiley-Interscience publication.

[46] Loh, W.-L. (1992). Stein's method and multinomial approximation. *Ann. Appl. Probab.*, **2** 536–554. URL http://dx.doi.org/10.1214/aoap/1177005648. MR1177898

[47] Meyn, S. P. and Tweedie, R. L. (1993). Stability of Markovian processes III: Foster-Lyapunov criteria for continuous time processes. *Adv. Appl. Probab.*, **25** 518–548. URL http://www.jstor.org/stable/1427522.

[48] Pardoux, E. and Veretennikov, Y. (2001). On the Poisson equation and diffusion approximation. I. *Ann. Probab.*, **29** 1061–1085. URL http://dx.doi.org/10.1214/aop/1015345596.

[49] Peterson, W. P. (1991). A heavy traffic limit theorem for networks of queues with multiple customer types. *Mathematics of Operations Research*, **16** 90–118. URL http://www.jstor.org/stable/3689851.

[50] Reed, J. (2009). The $G/GI/N$ queue in the Halfin-Whitt regime. *Annals of Ap-*

plied Probability, **19** 2211–2269. URL http://projecteuclid.org/euclid.aoap/1259158771.

[51] Reiman, M. I. (1984). Open queueing networks in heavy traffic. *Mathematics of Operations Research*, **9** 441–458. URL http://www.jstor.org/stable/3689532.

[52] Ross, N. (2011). Fundamentals of Stein's method. *Probab. Surv.*, **8** 210–293. URL http://dx.doi.org/10.1214/11-PS182.

[53] Stein, C. (1972). A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Probability Theory*. University of California Press, Berkeley, Calif., 583–602. URL http://projecteuclid.org/euclid.bsmsp/1200514239.

[54] Stein, C. (1986). Approximate computation of expectations. *Lecture Notes-Monograph Series*, **7**. URL http://www.jstor.org/stable/4355512.

[55] Stolyar, A. L. (2004). Maxweight scheduling in a generalized switch: state space collapse and workload minimization in heavy traffic. *Ann. Appl. Probab.*, **14** 1–53. URL http://dx.doi.org/10.1214/aoap/1075828046. MR2023015

[56] Stolyar, A. L. (2015). Tightness of stationary distributions of a flexible-server system in the Halfin-Whitt asymptotic regime. *Stoch. Syst.*, **5** 239–267. URL http://dx.doi.org/10.1214/14-SSY139.

[57] Stroock, D. W. and Varadhan, S. R. S. (1979). *Multidimensional Diffusion Processes*. Springer, New York. URL http://link.springer.com/book/10.1007%2F3-540-28999-2.

[58] Tezcan, T. (2008). Optimal control of distributed parallel server systems under the Halfin and Whitt regime. *Mathematics of Operations Research*, **33** 51–90. URL http://search.proquest.com/docview/212618995?accountid=10267.

[59] Ward, A. R. and Glynn, P. W. (2003). A diffusion approximation for a Markovian queue with reneging. *Queueing Systems*, **43** 103–128. URL http://dx.doi.org/10.1023/A%3A1021804515162.

[60] Weinberg, G. V. (2000). Stein factor bounds for random variables. *Journal of Applied Probability*, **37** 1181–1187. URL http://www.jstor.org/stable/3215511.

[61] Whitt, W. (2002). *Stochastic-process limits*. Springer, New York. URL http://link.springer.com/book/10.1007%2Fb97479.

[62] Whitt, W. (2003). How multiserver queues scale with growing congestion-dependent demand. *Operations Research*, **51** 531–542. URL http://pubsonline.informs.org/doi/abs/10.1287/opre.51.4.531.16093. MR1991969

[63] Williams, R. J. (1998). Diffusion approximations for open multiclass queueing networks: sufficient conditions involving state space collapse. *Queueing Systems*, **30** 27–88. URL http://dx.doi.org/10.1023/A:1019108819713.

[64] Ye, H.-Q. and Yao, D. D. (2012). A stochastic network under proportional fair resource control—diffusion limit with multiple bottlenecks. *Operations Research*, **60** 716–738. URL http://dx.doi.org/10.1287/opre.1120.1047.

[65] Ying, L. (2016). On the approximation error of mean-field models. In *Proceedings of the 2016 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Science*. ACM, Antibes Juan-les-Pins, France, 285–297. URL http://dx.doi.org/10.1145/2964791.2901463.

[66] Ying, L. (2016). On the rate of convergence of the power-of-two-choices to its mean-field limit. URL http://arxiv.org/abs/1605.06581.

[67] Zhang, B., van Leeuwaarden, J. and Zwart, B. (2012). Staffing call centers with impatient customers: refinements to many-server asymptotics. *Operations Research*, **60** 461–474. URL http://dx.doi.org/10.1287/opre.1110.1016.

[68] Zhang, J. and Zwart, B. (2008). Steady state approximations of limited processor sharing queues in heavy traffic. *Queueing Systems: Theory and Applications*, **60** 227–246. URL http://dx.doi.org/10.1007/s11134-008-9095-4.

Anton Braverman
School of Operations Research
and Information Engineering
Cornell University
Ithaca, New York 14853, USA
E-mail: ab2329@cornell.edu

J. G. Dai
School of Operations Research
and Information Engineering
Cornell University
Ithaca, New York 14853, USA
E-mail: jd694@cornell.edu

Jiekun Feng
Department of Statistical Science
Cornell University
Ithaca, New York 14853, USA
E-mail: jf646@cornell.edu