We show that the steady-state distribution of the join-the-shortest-queue (JSQ) system converges, in the Halfin-Whitt regime, to its diffusion limit at a rate of at least $1/\sqrt{n}$, where $n$ is the number of servers. Our proof uses Stein’s method, and, specifically, the recently proposed \textit{prelimit} generator comparison approach. The JSQ system is non-trivial, high-dimensional, and has a state-space collapse component, and our analysis may serve as a helpful example to readers wishing to apply the approach to their own setting.

\textit{Key words}: Stein’s method, generator comparison, join the shortest queue, load balancing, diffusion approximation
policy specifies the manner in which arriving customers are assigned to the servers. In this paper, we consider the classical join-the-shortest-queue (JSQ) policy. Under JSQ, an arriving customer enters service immediately if at least one server is idle; if not, they get routed to the server with the smallest number of customers in its buffer. Ties are broken arbitrarily. We refer to this as the JSQ system.

Parallel-server systems have generated immense interest in recent years, and the JSQ policy is fundamental because it minimizes the expected customer delay and maximizes, with respect to stochastic order, the number of customers served in a given time interval; see, for instance, Winston (1977), Weber (1978). For a sample of recent work on the JSQ policy, we refer readers to Eryilmaz and Srikant (2012), Mukherjee et al. (2016), Eschenfeldt and Gamarnik (2018), Gupta and Walton (2019), Banerjee and Mukherjee (2019), Liu and Ying (2019), Banerjee and Mukherjee (2020), Braverman (2020), Zhou and Shroff (2020a,b), Zhao et al. (2021), Hurtado-Lange and Maguluri (2021), Cao et al. (2021). Other popular load-balancing policies include the join-the-idle-queue policy (Stolyar (2015), Mukherjee et al. (2016)), the idle-one-first policy (Gupta and Walton (2019)), and of course the power-of-$d$ policy (Vvedenskaya et al. (1996), Mitzenmacher (2001)), but in this paper we focus on the JSQ policy. We make no attempt to give a comprehensive review of the literature on parallel-server systems, instead referring the reader to van der Boor et al. (2021) for a recent survey.

Understanding the exact performance of the system is known to be difficult and much attention has been devoted over the past decade to heavy-traffic asymptotics. The term “heavy traffic” refers to parameter regimes where the system utilization tends to one. “Conventional heavy traffic” assumes that the number of servers $n$ is fixed and $\lambda \uparrow 1$, while “many-server heavy traffic” assumes that $n \to \infty$ and $\lambda \uparrow 1$ jointly. For two examples of work in the conventional heavy-traffic setting, see Eryilmaz and Srikant (2012) and Zhou and Shroff (2020b). In this paper, we use the term “heavy-traffic” to refer to the many-server setting – the setting considered in most of the papers mentioned in the previous paragraph.

There are multiple many-server heavy-traffic regimes, depending on how $n$ and $\lambda$ jointly converge to their limit. For example, assuming that $\lambda = 1 - 1/n$ yields entirely different
asymptotic behavior compared to when $\lambda = 1 - 1/\sqrt{n}$. To capture all the possible heavy-traffic regimes, it is common practice to assume that the per-server load $\lambda$ is related to the number of servers $n$ through $\lambda = 1 - \beta/n^\alpha \in (0,1)$ for some $\alpha \geq 0$ and $\beta > 0$. In this paper we focus on the case when $\alpha = 1/2$; i.e., $\lambda = 1 - \beta/\sqrt{n}$. This regime is known as the Halfin-Whitt regime and is ubiquitous across the queueing theory literature. It derives from the work of Halfin and Whitt (1981) and is also known as the quality-and-efficiency-driven regime because it achieves reasonable customer wait times while maintaining high utilization of servers. The full list of parameter regimes is found in Figure 1.

![Figure 1](image-url)  
**Figure 1**  
The various many-server heavy-traffic regimes. Higher values of $\alpha$ represent heavier loads. Existing work across the different parameter regimes is reviewed in Section 1.1.

We now state and discuss our main results. Let $Q_i(t)$ be the number of servers with $i$ or more customers at time $t \geq 0$, noting that $Q_i(t) = 0$ for $i > b + 1$. The process $\{Q(t) = (Q_1(t), \ldots, Q_{b+1}(t))\}$ is an irreducible continuous-time Markov chain (CTMC) on a finite state space and therefore possesses a unique stationary distribution. We let $Q = (Q_1, \ldots, Q_{b+1})$ be the random vector having the stationary distribution of the CTMC. To describe the asymptotic behavior of $Q$, we let $\delta = 1/\sqrt{n}$ and define the diffusion-scaled random vector $X = (X_1, \ldots, X_{b+1})$ by $X_1 = \delta(n - Q_1)$, and $X_i = \delta Q_i$ for $2 \leq i \leq b + 1$. The results of Eschenfeldt and Gamarnik (2018) and Braverman (2020) imply that $X$ converges in distribution to some limiting $\mathbb{R}^{b+1}$-valued random vector $Y$ as $n \to \infty$. In this paper we establish an upper bound of order $1/\sqrt{n}$ on the rate of convergence to $Y$.

The random variable $Y$ is distributed according to the stationary distribution of the diffusion process $\{Y(t) \in \mathbb{R}^{b+1}_+\}$, which satisfies

$$Y_1(t) = Y_1(0) + \sqrt{2}W(t) + \beta t - \int_0^t (Y_1(s) + Y_2(s))ds + U(t),$$
\[ Y_2(t) = Y_2(0) + U(t) - \int_0^t Y_2(s) ds, \quad Y_3(t) = \cdots = Y_{b+1}(t) = 0, \] (1)

where \( \{W(t)\} \) is standard Brownian motion and \( \{U(t)\} \) is the unique nondecreasing, non-negative process in the space of càdlàg functions \( D[0, \infty) \) satisfying \( \int_0^\infty 1(Y_1(t) > 0)dU(t) = 0 \). The diffusion \( \{Y(t)\} \) was shown to be positive recurrent; see Banerjee and Mukherjee (2019) or Braverman (2020). Furthermore, (1) implies that \( Y_3 = \cdots = Y_{b+1} = 0 \).

Our main result is that there exists a constant \( C(b, \beta) \) such that for all \( n \geq 1 \), and any function \( h: \mathbb{R}^{b+1}_+ \to \mathbb{R} \) whose first-order and second-order partial derivatives are bounded in magnitude by one,

\[ |\mathbb{E}h(X) - \mathbb{E}h(Y)| \leq C(b, \beta)/\sqrt{n}. \] (2)

The assumption that \( b \) is finite is used frequently in the proof of (2) and, specifically, in the proof of Proposition 2. We deem the finite buffer assumption to be acceptable because it was shown by Braverman (2020) that even with infinite-sized buffers, \( \mathbb{E}Q_3 \leq C(\beta) \) for all \( n \geq 1 \) in the Halfin-Whitt regime, implying that \( X_3 \Rightarrow 0 \), or that the mass concentrates on those states with at most one customer waiting. Moreover, Liu and Ying (2020) showed that assuming finite buffers, \( \mathbb{E}Q_3 \to 0 \) as \( n \to \infty \) in the even busier super-Halfin-Whitt regime \( (1/2 < \alpha < 1) \).

In addition to the novelty of our result, this paper makes a methodological contribution. We prove (2) using Stein’s method, a framework introduced by Stein (1972) that allows one to study the rate of convergence of a sequence of random variables to its limit. Popularized in the area of queueing systems by Gurvich (2014), Ying (2017), Braverman and Dai (2017), Gast (2017), the generator comparison approach of Stein’s method, attributed to Barbour (1988, 1990) and Götze (1991), is used to study convergence rates of steady-state Markov chain distributions to their diffusion, fluid, or mean-field limits. For a few recent applications of the generator comparison approach in queueing, we refer the reader to Gaunt and Walton (2020), Hurtado-Lange and Maguluri (2021), Lu (2021), Liu et al. (2022); this list is by no means comprehensive. In this paper, we restrict our attention to the case when the limit is the stationary distribution of a diffusion process, referring the reader to Ying (2017) for a treatment of fluid and mean-field limits.
The generator approach requires bounds on various moments of the prelimit, known as moment bounds, and bounds on the derivatives of the solution to the Poisson equation for the limiting distribution. The latter are called gradient bounds in Braverman and Dai (2017), but in this paper we stick with the original term “Stein factors”, or “Stein factor bounds”; e.g., Ross (2011). While moment bounds can be difficult to obtain in some applications, Stein factor bounds are typically the bigger problem. When the limit is one-dimensional, Stein factors are bounded using the explicit form of the solution to the Poisson equation — an ordinary differential equation. When the limit is multidimensional, the Poisson equation is a partial differential equation (PDE) that generally does not have an explicit solution, making Stein factor bounds harder to establish. Techniques proposed to obtain multidimensional Stein factor bounds include using a priori Schauder estimates from elliptic PDE theory as in Gurvich (2014), using couplings to analyze and bound the sensitivity of the diffusion to its initial condition as in Barbour (1988) and Mackey and Gorham (2016), and bounding the Stein factors using Malliavin calculus as in Fang et al. (2018) and Jin et al. (2021). A detailed description of these techniques can be found in Section 1.1 of Braverman (2022). However, despite progress on multidimensional Stein factor bounds, the JSQ system is not covered by existing results because our limiting diffusion in (1) is constrained to the nonnegative orthant via reflecting boundary conditions.

To deal with the Stein factor bound problem, this paper promotes the use of the prelimit generator comparison approach, which was recently proposed by Braverman (2022) as an alternative to the generator comparison approach. The prelimit approach is the mirror image of the classical generator approach. Whereas the latter requires moment bounds on the prelimit $X$ and Stein factor bounds for limit $Y$, the former needs moment bounds on $Y$ and Stein factor bounds for the prelimit $X$. For the moment bounds used in this paper, the result that all moments of $Y$ are finite, proved by Banerjee and Mukherjee (2019), is sufficient because our limit $Y$ does not depend on $n$. The Stein factor bounds pose a bigger challenge, and we deal with them in Section 3. It was noted in Braverman (2022) that the prelimit and classical generator comparison approaches should be equivalent, in theory, in the sense that any bound on $|\mathbb{E}h(X) - \mathbb{E}h(Y)|$ obtained using one of them should be attainable using the other. However, in practice, one approach could be more tractable, or
convenient, to work with; see, for instance, the example in Section 4 of Braverman (2022). In the case of the JSQ system, we discuss in Remark 1 of Section 3.2.3 how the discrete state space simplifies the analysis of the couplings we use to establish Stein factor bounds, because the initial spacing of the coupled systems is preserved until coupling.

The introduction of the prelimit approach in Braverman (2022) was intended to be gentle, with the only example used there being the $M/M/1$ system. Our application of the approach to the JSQ system exposes all of its moving pieces and can be useful to those who want to apply the prelimit approach to their own setting. For example, some of the technical components of this paper that could be useful in other settings include: the regenerative argument used to establish first-order Stein factor bounds in Section 3.1, the approach we use to bound $E\lvert X \rvert$ in Section 3.2.1, and our treatment of reflecting boundary conditions in Appendix A.2.2.

It should be noted that Hurtado-Lange and Maguluri (2021) and Zhou and Shroff (2020a) used the classical generator comparison approach to obtain rates of convergence of the steady-state total customer count to an exponential random variable for $\alpha > 2$. The former paper was in the continuous-time setting, while the latter considered the discrete-time system, and the results in both papers also hold for routing policies other than JSQ, such as the power-of-$d$ policy. Since the limiting random variable in both papers is one-dimensional, the Stein factors bounds do not pose a challenge there.

1.1. Literature Review

Let us first review the literature on the analysis of the JSQ system in the various many-server heavy-traffic regimes. Most of the work has been done in the setting with infinite buffer sizes, so, unless otherwise noted, we assume that $b = \infty$. In Eschenfeldt and Gamarnik (2018), the authors established the process-level convergence of $\{X(t)\}_{t=1}^{\infty}$ to its diffusion limit in the Halfin-Whitt regime ($\alpha = 1/2$). That paper triggered a wave of interest in the many-server heavy-traffic asymptotics of the JSQ system. Convergence of the stationary distributions was later established by Braverman (2020), and the behavior of the stationary distribution of the limiting diffusion was studied by Banerjee and Mukherjee (2019, 2020). Our work fits with this group of papers, elevating the steady-state convergence result to one with rates of convergence.
Outside the Halfin-Whitt regime, Mukherjee et al. (2016) studied the transient and steady-state behavior of the JSQ system’s fluid limit when $\lambda = 1 - \beta < 1$ is a fixed constant ($\alpha = 0$), and Gupta and Walton (2019) established process-level convergence to the diffusion limit when $\alpha = 1$; known as the non-degenerate slowdown (NDS) regime and introduced by Atar (2012). In the sub-Halfin-Whitt regime when $\alpha \in (0, 1/2)$, Liu and Ying (2019) assumed finite buffers and obtained bounds on the steady-state total customer count in the system. A similar result was obtained for Coxian-2 service times by Liu et al. (2022), and by Liu and Ying (2020) for the super-Halfin-Whitt regime $\alpha \in (1/2, 1)$. Another recent work in the super-Halfin-Whitt regime was by Zhao et al. (2021), who worked with infinite buffers and established transient and steady-state diffusion limits for the normalized total queue length process. Their analysis exploited the regenerative structure of the JSQ system and contained several hitting-time estimates very close to our own estimates needed for the Stein factor bounds in Section 3. Lastly, both Hurtado-Lange and Maguluri (2021) and Zhou and Shroff (2020a) established rates of convergence to the exponential distribution for the steady-state normalized total customer count. Their results covered the case when $\alpha > 2$.

Other works have used Stein’s method in the setting of parallel-server systems beyond Hurtado-Lange and Maguluri (2021) and Zhou and Shroff (2020a). In Liu and Ying (2019, 2020), Liu et al. (2022), the authors used Stein’s method for mean-field analysis to obtain bounds on steady-state performance metrics of interest, like $E Q_2$ for instance, for the power-of-$d$ system. Another line of work on power-of-$d$ systems was by Gast (2017), Gast and Van Houdt (2017), Gast et al. (2019), where the authors showed how to derive refined mean-field models for improved steady-state approximations. More recently, Hairi et al. (2021) provide calculable error bounds for the mean-field approximation of the power-of-two-choices model.

1.2. Notation
We use $\mathbb{Z}$ to denote the set of integers and let $\mathbb{N} = \{0, 1, 2, \ldots\}$. For any $k \in \mathbb{N}$ and $B \subset \mathbb{R}^d$, we let $C^k(B)$ be the set of all $k$-times continuously differentiable functions $f : B \to \mathbb{R}$. We let $e \in \mathbb{R}^d$ be the vector whose elements all equal 1 and let $e^{(i)}$ be the element with 1 in the
ith entry and zeros otherwise. For any \( \delta > 0 \) and integer \( d > 0 \), we let \( \delta \mathbb{Z}^d = \{ \delta k : k \in \mathbb{Z}^d \} \) and define \( \delta \mathbb{N}^d \) similarly. For any function \( f : \delta \mathbb{Z}^d \to \mathbb{R} \), we define the forward difference operator in the \( i \)th direction as

\[
\Delta_i f(\delta k) = f(\delta (k + e^{(i)})) - f(\delta k), \quad k \in \mathbb{Z}^d, \ 1 \leq i \leq d,
\]

and for \( j \geq 0 \), we define

\[
\Delta^{j+1}_i f(\delta k) = \Delta^j_i f(\delta (k + e^{(i)})) - \Delta^j_i f(\delta k),
\]

with the convention that \( \Delta^0_i f(\delta k) = f(\delta k) \). For a vector \( a \in \mathbb{N}^d \), we also let

\[
\Delta^a f(\delta k) = \Delta^a_1 f(\delta k) \cdots \Delta^a_d f(\delta k),
\]

and if \( f : \mathbb{R}^d \to \mathbb{R} \), then

\[
\frac{\partial^a}{\partial x^a} f(x) = \frac{\partial^{a_1}}{\partial x_1^{a_1}} \cdots \frac{\partial^{a_d}}{\partial x_d^{a_d}} f(x),
\]

and we adopt the convention that \( \frac{\partial^0}{\partial x^0} f(x) = f(x) \). For any \( x \in \mathbb{R}^d \), we define \( \|x\|_1 = \sum_{i=1}^d |x_i| \) and use \( |x| \) to denote the Euclidean norm. For any \( f : \mathbb{R}^d \to \mathbb{R} \), we let \( \|f\|_\infty = \sup_{x \in \mathbb{R}^d} |f(x)| \). Throughout the paper, we will often use \( C \) to denote a generic positive constant that may change from line to line and that is independent of any parameters not explicitly specified.

2. Main Result

Recall that \( Q_i(t) \) is the number of servers with \( i \) or more customers at time \( t \geq 0 \) and that \( \{Q(t) = (Q_i(t))_{i=1}^{b+1} : t \geq 0\} \) is an irreducible CTMC with state space given by

\[
S_Q = \{ q \in \{0, \ldots, n\}^{b+1} : q_i \geq q_{i+1} \}. \tag{4}
\]

Figure 2 gives an example of a state \( q \in S_Q \). We assume that \( \lambda = 1 - \beta/\sqrt{n} \) for some fixed \( \beta > 0 \). Let \( \delta = 1/\sqrt{n} \) and define the diffusion-scaled CTMC \( \{X(t)\} \) by

\[
X_1(t) = \delta (n - Q_1(t)), \quad X_i(t) = \delta Q_i(t), \quad 2 \leq i \leq b + 1,
\]
which takes values on the state space
\[ S = \{ (x_1^q, x_2^q, \ldots, x_{b+1}^q) = (\delta(n - q_1), \delta q_2, \ldots, \delta q_{b+1}) : q \in S_Q \}. \]

We will often use \( x^q \in S \) and \( q \in S_Q \) interchangeably. Recalling that \( \Delta_i f(x^q) = f(x^q + \delta e^{(i)}) - f(x^q) \), for any \( f : S \to \mathbb{R} \), the infinitesimal generator of \( \{X(t)\} \) satisfies
\[
G_X f(x^q) = -1(1(q_1 < n)n \lambda \Delta_1 f(x^q - \delta e^{(1)}) + n \lambda \sum_{j=1}^b 1(q_1 = \ldots = q_j = n, q_{j+1} < n) \Delta_{j+1} f(x^q) + (q_1 - q_2) \Delta_1 f(x^q) - \sum_{j=2}^b (q_j - q_{j+1}) \Delta_j f(x^q - \delta e^{(j)}) - q_{b+1} \Delta_{b+1} f(x^q - \delta e^{(b+1)}). 
\]

The first line of transitions in (5) correspond to arrivals. We see that for \( j \geq 2 \), the \( j \)th component of \( x_j^q \) only grows provided the preceding \( j - 1 \) horizontal levels, as depicted in Figure 2, are full. The transitions in the second line of (5) correspond to service completions. Using Figure 2 again, we interpret \( (q_j - q_{j+1}) \) as the number of servers (vertical columns) with exactly \( j \) customers.

Recall that \( X = (X_1, \ldots, X_{b+1}) \) and \( Y = (Y_1, Y_2, 0, \ldots, 0) \) are distributed according to the stationary distributions of the scaled CTMC and the diffusion \( \{Y(t) \in \mathbb{R}_{+}^{b+1}\} \) defined in (1), respectively. Going forward, we note that unless explicitly stated, all expectations are with respect to the stationary distribution at hand; i.e., either \( X \) or \( Y \). To state our main result, we define
\[
\mathcal{M}_j = \left\{ h^* : \mathbb{R}^{b+1} \to \mathbb{R}, \left\| \frac{\partial^a}{\partial x^a} h^*(x) \right\|_{\infty} \leq 1, \ 1 \leq \|a\|_1 \leq j \right\},
\]
and $d_{M_j}(X, Y) = \sup_{h^* \in M_j} |E h^*(X) - E h^*(Y)|$. We use an asterisk to emphasize that $h^*(x)$ is defined on the continuum $\mathbb{R}^{b+1}$. Later we will drop the asterisk to refer to functions defined only on the grid $\delta \mathbb{Z}^{b+1}$. It was shown in Lemma 2.2 of Mackey and Gorham (2016) that $M_3$ is a convergence-determining class; i.e., $d_{M_3}(U, V) \to 0$ implies $U$ and $V$ converge in distribution. The following is our main result.

**Theorem 1.** For any $0 < b < \infty$, there exists a constant $C(b, \beta)$ such that for all $n \geq 1$,

$$d_{M_2}(X, Y) = \sup_{h^* \in M_2} |E h^*(X) - E h^*(Y)| \leq \frac{C(b, \beta)}{\sqrt{n}}.$$  \hspace{1cm} (6)

Note that $M_2$ is also a convergence-determining class because $M_3 \subset M_2$. We prove Theorem 1 in Section 2.1 using the prelimit generator approach of Stein’s method. Multiple parts of the proof assume that $n$ is large enough, say, $n > N(\beta)$ for some $N(\beta) > 0$. We can make this assumption without loss of generality by redefining $C(b, \beta)$ to be larger than $\max_{1 \leq n \leq N(\beta)} d_{M_2}(X, Y)$.

2.1. Proving Theorem 1

Central to our proof is the ability to extend any grid-valued function to be defined on all of $\mathbb{R}_+^{b+1}$. Although there are infinitely many such extensions, we use a polynomial spline $A$ that extends grid-valued functions $f : \delta \mathbb{N}^{b+1} \to \mathbb{R}$ to functions $A f : \mathbb{R}_+^{b+1} \to \mathbb{R}$. We leave the detailed construction to Appendix A.2 because for this section, it suffices to know that $A$ is a linear operator, that $A f \in C^2(\mathbb{R}_+^{b+1})$, and that $A$ applied to a constant equals that constant. Recalling that $\delta = 1/\sqrt{n}$, the following auxiliary lemma is needed.

**Lemma 1.** Define

$$M_{\text{disc}, j}(c) = \{ h : \delta \mathbb{N}^{b+1} \to \mathbb{R}, \ |\Delta^a h(\delta k)| \leq c \delta^{\|a\|_1}, \ 1 \leq \|a\|_1 \leq j, \ \delta k \in \delta \mathbb{N}^{b+1} \}.$$  

There exist some $C, C' > 0$ independent of any JSQ model parameters such that

$$d_{M_2}(X, Y) \leq \sup_{h \in M_{\text{disc}, 2}(C)} |E h(X) - E Ah(Y)| + C' \delta.$$  \hspace{1cm} (7)

**Proof of Lemma 1** The result follows by repeating the arguments used in the proof of Lemma 1 in Braverman (2022). □
Going forward, when we write $\mathcal{M}_{\text{disc}, 2}(C)$, the constant $C$ is assumed to be the one in Lemma 1. Furthermore, note that if $h(0) \neq 0$, then the linearity of $A$ and the fact that $A$ applied to a constant equals that constant implies that $\tilde{h}(x) = h(x) - h(0)$ satisfies $\mathbb{E}\tilde{h}(X) - \mathbb{E}Ah(Y) = \mathbb{E}h(X) - \mathbb{E}Ah(Y)$. We therefore, without loss of generality, consider only those $h \in \mathcal{M}_{\text{disc}, 2}(C)$ such that $h(0) = 0$.

To prove Theorem 1, we bound the right-hand side of (7) with the help of the following two ingredients. The first ingredient is a rate-conservation law for $\{Y(t)\}$, proved in Appendix A.

**Lemma 2.** Given $f \in C^2(\mathbb{R}_+^{b+1})$, define

$$G_Y f(x) = (\beta - (x_1 + x_2)) \frac{\partial}{\partial x_1} f(x) - x_2 \frac{\partial}{\partial x_2} f(x) + \frac{\partial^2}{\partial x_1^2} f(x), \quad x \in \mathbb{R}_+^{b+1}. \quad (8)$$

If $\mathbb{E}|f(Y)| < \infty$ and $\mathbb{E}|G_Y f(Y)| < \infty$, and if $Y(0)$ is initialized according to $Y$, then

$$\mathbb{E}G_Y f(Y) + \mathbb{E} \left( \int_0^1 \left( \frac{\partial}{\partial x_1} f(Y(s)) + \frac{\partial}{\partial x_2} f(Y(s)) \right) 1(Y_1(s) = 0) dU(s) \right) = 0. \quad (9)$$

The second ingredient is the Poisson equation. For $h : \delta\mathbb{N}^{b+1} \to \mathbb{R}$ and $c \in \mathbb{R}$, let

$$f_h^{(c)}(x^q) = c + \int_0^\infty (\mathbb{E}_{x^q} h(X(t)) - \mathbb{E} h(X)) dt, \quad x^q \in S,$$

which is well defined because the CTMC has a finite state space and is therefore exponentially ergodic. Furthermore, Lemma 2 of Braverman (2022) (see also Lemma 1 of Barbour (1988)) implies that

$$G_X f_h^{(c)}(x^q) = \mathbb{E} h(X) - h(x^q), \quad x^q \in S. \quad (10)$$

Most applications of Stein’s method have $c = 0$, but we choose $c = c^* = -f_h^{(0)}(0)$ and define

$$f_h(x^q) = f_h^{(c^*)}(x^q) = \int_0^\infty (\mathbb{E}_{x^q} h(X(t)) - \mathbb{E} h(X)) dt - \int_0^\infty (\mathbb{E}_0 h(X(t)) - \mathbb{E} h(X)) dt$$

$$= \int_0^\infty (\mathbb{E}_{x^q} h(X(t)) - \mathbb{E}_0 h(X(t))) dt, \quad x^q \in S. \quad (11)$$

Our choice of $c$ yields $f_h(0) = 0$, which comes in handy later when we need to bound $|f_h(x^q)|$ in Proposition 2. Going forward, we assume that $c = c^*$ when referring to (10).
Let us give an informal roadmap for bounding (7), with the formal statement of the bounds left to Proposition 1 below. We bound (6) by comparing the CTMC and diffusion generators. However, the former is defined only on a subset of $\mathbb{R}_+^{b+1}$, which requires the following workaround. Suppose that we are given a set $B \subset \mathbb{R}_+^{b+1}$ such that (a) $Eh(X) - Ah(x) = AG_X f_h(x)$ for $x \in B$ and (b) the probability that $Y \notin B$ goes to zero rapidly (we will make this precise) as $n \to \infty$. We decompose $Eh(X) - Ah(x)$ as

$$Eh(X) - Ah(x) = AG_X f_h(x)1(x \in B) + (Eh(X) - Ah(x))1(x \notin B)$$

and take expected values with respect to $Y$ (we will show that these are finite) to get

$$Eh(X) - EAh(Y) = E(AG_X f_h(Y)1(Y \in B)) + E((Eh(X) - Ah(Y))1(Y \notin B)).$$

Now extend $f_h(x^q)$ to $\delta \mathbb{N}_+^{b+1}$ by defining $f_h(x^q) = 0$ for $x^q \in \delta \mathbb{N}_+^{b+1} \setminus S$ and consider $Af_h(x)$. Provided that $E|Af_h(Y)| < \infty$ and $E|G_Y Af_h(Y)| < \infty$, we can invoke Lemma 2 with $f(x) = Af_h(x)$ there to conclude that

$$Eh(X) - EAh(Y) = E\left((AG_X f_h(Y) - G_Y Af_h(Y))1(Y \in B)\right)$$

$$+ E\left((Eh(X) - Ah(Y) - G_Y Af_h(Y))1(Y \notin B)\right)$$

$$- E\left(\int_0^1 \left(\frac{\partial}{\partial x_1} Af_h(Y(s)) + \frac{\partial}{\partial x_2} Af_h(Y(s))\right)1(Y_1(s) = 0)dU(s)\right), \quad (12)$$

where $Y(0)$ in the third line is initialized according to $Y$. We bound the first line by showing that $G_X$ and $G_Y$ are close to one another. The middle term is small due to our choice of $B$ and the last term can be bounded because the JSQ system exhibits reflecting behavior similar to $\{Y(t)\}$ at the boundary $\{x \in S : x^q_1 = 0\}$. As a final remark, our choice of $f_h(x^q) = 0$ for $x^q \in \delta \mathbb{N}_+^{b+1} \setminus S$ is made for convenience and is not essential to the proof, because the probability that $Y \notin B$ shrinks rapidly as $n \to \infty$.

To state the following proposition, define $k : \mathbb{R}_+^{b+1} \to \mathbb{Z}_+^{b+1}$ elementwise by $k_j(x) = \lfloor x_j / \delta \rfloor$. For notational convenience, we also define $I = \{i = (i_1, i_2, 0, \ldots, 0) \in \mathbb{N}_+^{b+1} : 0 \leq i_1, i_2 \leq 4\}$. The following proposition is proved in Appendix A.2.
Proposition 1. If \( h \in \mathcal{M}_{\text{disc},2}(C) \), then \( Ah(Y) \), \( Af_h(Y) \), and \( G_Y Af_h(Y) \) are integrable, and \((12)\) holds. Furthermore, suppose that \( n > 16 \), define

\[
B = \{(x_1, x_2, 0, \ldots, 0) \in \mathbb{R}_{+}^{k+1} : x_2 + x_1 \leq \delta(n/2 - 8) = (n/2 - 8)/\sqrt{n}\},
\]

and let

\[
\begin{align*}
\varepsilon_1(Y) &= (AG_X f_h(Y) - G_Y Af_h(Y))1(Y \in B), \\
\varepsilon_2(Y) &= (Eh(X) - Ah(Y) - G_Y Af_h(Y))1(Y \notin B), \\
\varepsilon_3(Y) &= \left( \frac{\partial}{\partial x_1} Af_h(Y) + \frac{\partial}{\partial x_2} Af_h(Y) \right)1(Y \in B), \text{ and} \\
\varepsilon_4(Y) &= \left( \frac{\partial}{\partial x_1} Af_h(Y) + \frac{\partial}{\partial x_2} Af_h(Y) \right)1(Y \notin B).
\end{align*}
\]

There exist \( C(\beta), C(b, \beta) > 0 \) independent of \( h(x) \) and \( n \) such that

\[
|\varepsilon_1(Y)| \leq C(\beta)(1 + \delta^{-1}Y_2) \max_{i \in I} |\Delta_i^{a_1} \Delta_i^{a_2} f_h(\delta(k(Y) + i))| + C(\beta)\delta^{-2} \max_{i \in I} |\Delta_i^{a_1} f_h(\delta(k(Y) + i))| \\
+ C(\beta)\delta^{-2}1(Y_1 \leq \delta) \max_{i_1 = 0} |\Delta_i^{a_1} f_h(\delta(k(Y) + i))|,
\]

\[
|\varepsilon_2(Y)| \leq C(b, \beta)1(Y \notin B)\delta^{-2}(1 + Y_1 + Y_2) \max_{i \in I} |f_h(\delta(k(Y) + i))|,
\]

\[
|\varepsilon_3(Y)| \leq C(\beta)\delta^{-1}1(Y \in B) \left( |(\Delta_1 + \Delta_2)f_h(\delta k(Y))| + \max_{a_1 + a_2 = 2} |\Delta_i^{a_1} \Delta_i^{a_2} f_h(\delta(k(Y) + i))| \right),
\]

\[
|\varepsilon_4(Y)| \leq C(\beta)\delta^{-1}1(Y \notin B) \max_{i \in I} |f_h(\delta(k(Y) + i))|.
\]

Note that \( \varepsilon_1(Y) \) and \( \varepsilon_2(Y) \) are related to the first and second lines of \((12)\), respectively, while \( \varepsilon_3(Y) \) and \( \varepsilon_4(Y) \) are related to the last line there. From the bounds in Proposition 1, we see that the bound on \((12)\) depends on the CTMC through the function \( f_h(x^q) \) and its differences, and on the diffusion through the distribution of \( Y \). The differences of \( f_h(x^q) \) are commonly known as Stein factors, and the following proposition, proved in Section 3, exhibits the Stein factor bounds we need to prove Theorem 1.

Proposition 2. There exists \( C(\beta, b) > 0 \) such that for any \( n \geq 1 \) and \( h \in \mathcal{M}_{\text{disc},2}(C) \),

\[
|\Delta_i^{a_1} \Delta_i^{a_2} f_h(x^q)| \leq C(\beta, b)\delta^{a_1 + a_2}(1 + x^q)^{a_1 + a_2},
\]
for all $a_1, a_2 \geq 0$ with $1 \leq a_1 + a_2 \leq 2$, and all $x^q \in S$ with $x^q_1 \leq \delta(n - a_1), x^q_2 \leq \delta(n - a_2)$, and $x^q_3 = 0$. Furthermore,

$$|f_h(x^q)| \leq C(\beta, b)(1 + x^q_2)(x^q_1 + x^q_2)/\delta, \quad x^q \in S, \ x^q_3 = 0,$$

$$|\Delta^3_h f_h(x^q)| \leq C(\beta, b)\delta^3(1 + x^q_2)^3, \quad x^q \in S, \ x^q_1 \leq \delta(n - 3), \ x^q_3 = 0,$$

and for all $x^q \in S$ with $x^q_1 = 0, 0 \leq x^q_2 \leq \delta(n - 1)$, and $x^q_3 = 0$,

$$|(\Delta_1 + \Delta_2)f_h(x^q)| \leq C(\beta, b)\delta^2(1 + x^q_2)^2 \quad \text{and} \quad |(\Delta_1^2 - (\Delta_1 + \Delta_2)) f_h(x^q)| \leq C(\beta, b)\delta^3(1 + x^q_2)^3.$$

The last component needed for the proof of Theorem 1 is the following lemma.

**Lemma 3.** All moments of $Y_1$ and $Y_2$ are finite. Furthermore, suppose that $Y(0)$ is initialized according to $Y$. Then for any $j > 0$,

$$\mathbb{E}Y_2^{j+1} = \left( \int_0^1 (Y_2(s))^{j}1(Y_1(s) = 0)dU(s) \right). \quad (13)$$

**Proof of Lemma 3** The finiteness of the moments follows from Theorem 2.1 of Banerjee and Mukherjee (2019) and (13) is implied by (9) of Lemma 2 with $f(y) = y_2^{j+1}$ there.

**Proof of Theorem 1** Initialize $Y(0)$ according to $Y$. Using (12) and the definitions of $\varepsilon_1(Y), \ldots, \varepsilon_4(Y)$, it follows that

$$\mathbb{E}h(X) - \mathbb{E}A h(Y) = \mathbb{E}\varepsilon_1(Y) + \mathbb{E}\varepsilon_2(Y) - \mathbb{E}\left( \int_0^1 (\varepsilon_3(Y(s)) + \varepsilon_4(Y(s)))1(Y_1(s) = 0)dU(s) \right).$$

We argue that $|\mathbb{E}h(X) - \mathbb{E}A h(Y)| \leq C(b, \beta)\delta$ for any $h \in \mathcal{M}_{disc,2}(C)$, which implies Theorem 1 when combined with Lemma 1. Since $\delta(k_2(Y) + i_2) \leq Y_2 + 4\delta$ for $i \in I$, applying the Stein factor bounds in Proposition 2 with the bounds on $\varepsilon_1(Y)$ and $\varepsilon_2(Y)$ in Proposition 1 yields

$$|\varepsilon_1(Y)| \leq C(b, \beta)1(Y \in B)\delta(1 + Y_2)^3, \quad |\varepsilon_2(Y)| \leq 1(Y \notin B)C(\beta)\delta^{-3}(1 + Y_1 + Y_2)^3. \quad (14)$$

We point out that

$$\delta^{-1} \leq C(Y_1 + Y_2), \quad \text{for any } Y \notin B, \quad (15)$$
which follows from the facts that $Y_1 + Y_2 \geq \delta(n/2 - 8) = \delta^{-1}/2 - \delta$ for $Y \notin B$, that $\delta = 1/\sqrt{n}$, and that $n > 16$. Combining (14), (15), and the fact that the moments of $Y_i$ are finite yields

$$\mathbb{E} |\varepsilon_1(Y)| + \mathbb{E} |\varepsilon_2(Y)| \leq C(b, \beta)\delta \mathbb{E}(1 + Y_1 + Y_2)^7 \leq C(b, \beta)\delta.$$ 

Furthermore, applying the Stein factor bounds in Proposition 2 to the bounds on $\varepsilon_3(Y)$ and $\varepsilon_4(Y)$ in Proposition 1, and using (15), we get

$$|\varepsilon_3(Y) + \varepsilon_4(Y)| \leq C(b, \beta)1(Y \in B)\delta(1 + Y_1 + Y_2)^2 + C(b, \beta)1(Y \notin B)\delta^{-2}(1 + Y_1 + Y_2)^2$$

$$\leq C(b, \beta)\delta(1 + Y_1 + Y_2)^5.$$ 

Thus, (13) of Lemma 3 implies that

$$\mathbb{E} \left( \int_0^1 |\varepsilon_3(Y(s)) + \varepsilon_4(Y(s))| 1(Y_1(s) = 0)dU(s) \right) \leq C(b, \beta)\delta.$$

### 3. Stein Factor Bounds

In this section we prove Proposition 2. We bound the first-order differences in Section 3.1. This requires the most effort. The second-order differences are bounded at the start of Section 3.2, with Section 3.2.1 showing how they can be used to bound $\mathbb{E}|h(X)|$, which may be of independent interest. Section 3.2.2 contains the third-order bounds and Section 3.2.3 proves two technical lemmas needed for the second-order bounds.

#### 3.1. First-Order Differences

In this section we bound

$$\Delta_i f_h(x^q) = \int_0^\infty \mathbb{E}_{x^q + d_i}\{h(X(t)) - h(x^q)\}dt$$

by coupling two copies of the JSQ model initialized one customer apart. The coupling is introduced in the following lemma, which is stated in terms of the unscaled CTMC $\{Q(t)\}$.

**Lemmas 4.** For $1 \leq i \leq b + 1$, define $\Theta_i^Q = \{(q, \bar{q}) \in S_Q \times S_Q : q_i < n, \bar{q}_i = q_i + 1\}$. There exists a coupling $\{\bar{Q}(t)\}$ of $\{Q(t)\}$ whose transient distribution satisfies

$$\{\bar{Q}(t) | (Q(0), \bar{Q}(0)) \in \Theta_i^Q, Q(0) = q\}_{t \geq 0} \overset{d}{=} \{Q(t) | Q(0) = (q + e^{(i)})\}. \quad (16)$$

Furthermore, if $(Q(0), \bar{Q}(0)) \in \bigcup_{i=1}^{b+1} \Theta_i^Q$, then
Figure 3  Two possible states of the joint chain $(Q(t), \tilde{Q}(t))$ are depicted. The red circles correspond to customers in $Q(t)$, while the blue circle is the extra customer in $\tilde{Q}(t)$. In the figure on the left, the joint chain is in $\Theta_Q^1$, meaning the blue customer is in service and will leave the system after an exponentially distributed amount of time, coupling the joint chain. In the figure on the right, the joint chain is in $\Theta_Q^3$ because the blue customer is assigned to a server with a total of three customers.

(a) $\tilde{Q}(t) = Q(t)$ for all times $t \geq \tau_C$, where $\tau_C = \inf \{ t \geq 0 : Q(t) = \tilde{Q}(t) \}$.

(b) The pair $(Q(t), \tilde{Q}(t))$ belongs to $\bigcup_{i=1}^{b+1} \Theta_Q^i$ for all times $t < \tau_C$.

(c) Let $V$ be a unit-mean exponentially distributed random variable independent of $\{Q(t)\}$.

Then
\[
\tau_C \overset{d}{=} \min \left\{ \inf_{t \geq 0} \left\{ \int_0^t 1((Q(s), \tilde{Q}(s)) \in \Theta_Q^i) \, ds = V \right\}, \inf_{t \geq 0} \left\{ Q_{b+1}(t) = n \right\} \right\}. \tag{17}
\]

Proof of Lemma 4  Let us construct a joint CTMC $\{(Q(t), \tilde{Q}(t))\}$ by specifying its transitions. For simplicity, we refer to $\{Q(t)\}$ as system 1 and to $\{\tilde{Q}(t)\}$ as system 2. We think of system 2 as a copy of system 1 but with an additional low-priority customer following a preemptive resume rule. That is, service is interrupted, and the extra customer moves to the back of its buffer when a regular customer joins, even if the low-priority customer is currently in service.

Any state in $\Theta_Q^1$ is one where the low-priority customer is in service. The remaining $\Theta_Q^i$ correspond to states where the low-priority customer is assigned to a server with a total of $i$ customers; Figure 3 contains an example of a states in $\Theta_Q^1$ and $\Theta_Q^3$. Assuming $(Q(0), \tilde{Q}(0)) = (q, \tilde{q}) \in \Theta_Q^i$ for some $1 \leq i \leq b+1$, we now describe the possible transitions of the joint chain.

If $i = 1$, then the low-priority customer is in service. After a unit-mean exponentially distributed amount of time, he leaves system 2 and both systems couple. After coupling,
systems 1 and 2 are identical in terms of current and future customers, so they coincide on every sample path. All other transitions of the joint chain are based on the standard transitions of the JSQ model. In other words, a service completion by any of the $q_1$ servers working in system 1 results in a customer departure from both systems.

Figure 4 illustrates the effect of arrivals when $(q, \tilde{q}) \in \Theta^Q_1$. Namely, when $q_1 \leq n - 2$, a new arrival is assigned to the same idle server in both systems. If a customer arrives when $q_1 = n - 1$, then system 1 has only one idle server and system 2 has none. In system 1, that customer will be assigned to the last remaining idle server. Recall that when defining our JSQ model, we allowed for an arbitrary tie-breaking decision in routing arrivals. Therefore, in system 2, we assign that customer to the server working on the low-priority customer, causing a service preemption and pushing the low-priority customer to the back of the buffer. An arrival when $q_1 = n - 1$ transitions the joint chain from $\Theta^Q_1$ to $\Theta^Q_2$.

If $2 \leq i \leq b$, then the low-priority customer is in the back of some server’s buffer. A service completion by any of the $q_1$ servers working in system 1 results in a customer departure from both systems. If, however, the service completion happens at the server containing the low-priority customer, then the chain transitions from $\Theta^Q_i$ to $\Theta^Q_{i-1}$ because the low-priority customer is now assigned to a server with $i - 1$ customers; see Figure 5 for a depiction of such a transition. All new arrivals get assigned to the same server in each system. Note that if an arrival happens when $q_i = n - 1$ and $q_1 = \cdots = q_{i-1} = n$, then the system transitions from $\Theta^Q_i$ to $\Theta^Q_{i+1}$.

The final case is when $i = b + 1$. All transitions are identical to the $2 \leq i \leq b$ case, except for a customer arrival to a system where $q_{b+1} = n - 1$ and $q_1 = \cdots = q_b = n$. In that case,
system 1 assigns the customer to the last available slot, but system 2 blocks the customer because it is already full. This transition causes the two systems to couple. Note that our construction immediately implies the three claims in Lemma 4.

Let \( \tilde{X}(t) = (\delta(n - \tilde{Q}_1(t)), \delta \tilde{Q}_2(t), \ldots, \delta \tilde{Q}_{b+1}(t)) \) be the scaled version of \( \tilde{Q}(t) \). For any \( x^q \in S \) with \( x^q_1 > 0 \), and any \( h \in \mathcal{M}_{disc,2}(C) \),

\[
\int_0^\infty \mathbb{E}_{x - \delta e^{(1)}}[h(X(t)) - h(X(t))] \, dt = \left| \int_0^\infty \mathbb{E}_{(x,x - \delta e^{(1)})}(h(\tilde{X}(t)) - h(X(t))) \, dt \right| \\
\leq \int_0^\infty \mathbb{E}_{(x,x - \delta e^{(1)})}(\delta 1(t \leq \tau_C)) \, dt = \delta \mathbb{E}_{(x,x - \delta e^{(1)})}\tau_C,
\]

where \( \mathbb{E}_{(x,x - \delta e^{(1)})}(\cdot) \) denotes the expectation given \( (X(0), \tilde{X}(0)) = (x, x - \delta e^{(1)}) \). The inequality above is true because the gap between \( \{X(t)\} \) and \( \{\tilde{X}(t)\} \) never increases beyond one customer. The same argument implies that \( |\Delta_i f_h(x^q)| \leq \delta \mathbb{E}_{(x,x + \delta e^{(1)})}\tau_C \) for \( i \geq 2 \), and we see that bounding the first-order Stein factors amounts to bounding the expected coupling time \( \tau_C \). The following lemma provides the necessary bound. It is worth highlighting that proving this result requires a large amount of effort and JSQ-model-specific insight.

**Lemma 5.** For any \( (q, \tilde{q}) \in \bigcup_{i=1}^{b+1} \Theta_i^Q \),

\[
\mathbb{E}_{(q, \tilde{q})}\tau_C \leq C(b, \beta)(1 + \delta q_2).
\]

Before proving the lemma, we note that the first-order bounds in Proposition 2 are a consequence of (18) and Lemma 5; i.e.,

\[
|\Delta_i f_h(x^q)| \leq C(b, \beta)(1 + x^q_i), \quad i = 1, \ldots, b+1.
\]
Furthermore, note that for any \( x^q \in S \) with \( x^q_3 = 0 \),
\[
f_h(x^q) = f_h(0) + \sum_{j_1=0}^{x^q_1/\delta-1} \Delta_1 f_h(\delta j_1, 0, \ldots, 0) + \sum_{j_2=0}^{x^q_2/\delta-1} \Delta_2 f_h(x^q_1, \delta j_2, 0, \ldots, 0).
\]
Recall that \( f_h(0) = 0 \), and that the definition of \( S \) implies that \( \delta(j_1, j_2, 0, \ldots, 0) \in S \) for any \( 0 \leq j_1 \leq x^q_1/\delta \) and \( 0 \leq j_2 \leq x^q_2/\delta \). Combining these facts with (19) yields
\[
|f_h(x^q)| \leq C(b, \beta)(1 + x^q_2)(x^q_1 + x^q_2)/\delta, \quad x^q \in S, \ x^q_3 = 0,
\]
which proves one of the claims from Proposition 2.

We now describe the main idea and introduce several auxiliary lemmas used to prove Lemma 5. Our discussion communicates the main intuition behind the proof, leaving the technical details to Appendix B. Let \( \gamma > 0 \) be a constant independent of \( n \) whose precise value will be specified later, and define
\[
\theta_1 = n - \lfloor \sqrt{n\beta}/2 \rfloor, \quad \text{and} \quad \theta_2 = \lceil \gamma \sqrt{n} \rceil.
\]
Additionally, we define the stopping times
\[
\tau_i(q_i) = \inf\{t \geq 0 : Q_i(t) = q_i\}, \quad q_i \in \{0, 1, \ldots, n\}, \ i = 1, 2.
\]
We now describe a sequence of cycles, or attempts, such that in each cycle, the probability of the joint chain coupling is bounded from below by a constant independent of \( n \). Given an initial state \((Q(0), \tilde{Q}(0)) = (q, \tilde{q})\) belonging to some \( \Theta^Q_i \), we wait until \( \tau_2(\theta_2) \), which marks the start of the first cycle. From that point, we wait until \( \min(\tau_1(\theta_1), \tau_2(2\theta_2)) \). If \( \tau_1(\theta_1) \geq \tau_2(2\theta_2) \), then we give up trying to couple this cycle, and wait until \( \tau_2(\theta_2) \) to start a fresh cycle. If \( \tau_1(\theta_1) < \tau_2(2\theta_2) \), then there are \( \lfloor \sqrt{n\beta}/2 \rfloor \) idle servers and at most \( 2\theta_2 \) non-empty buffers. From such a state, we are guaranteed that coupling happens if the joint CTMC enters \( \Theta^Q_i \) and spends an exponentially distributed amount of time there before all servers in \( \{Q(t)\} \) become busy; i.e., \( \tau_C < \tau_1(n) \). If \( \tau_C \geq \tau_1(n) \), we give up trying to couple this cycle and wait until \( \tau_2(\theta_2) \) for the next cycle to restart the coupling attempt. Note that this cycle sequence resembles a renewal sequence, but the new cycle times are not
renewal times because the values of $Q_3(\cdot), \ldots, Q_{b+1}(\cdot)$ can vary at the start of each new cycle.

From our discussion, it follows that coupling is guaranteed in any given cycle if, starting from a state with $q_2 = \theta_2$, the events $\{\tau_1(\theta_1) < \tau_2(2\theta_2)\}$ and $\{\tau_C < \tau_1(n)\}$ occur. In Appendix B we derive a lower bound, uniform in $n$, on the probability of coupling in a given cycle, implying that coupling is guaranteed to happen after a geometrically distributed number of cycles. We also derive an upper bound, uniform in $n$, on the expected time until the start of the first cycle, as well as the expected cycle duration, and then combine these bounds and prove Lemma 5.

3.2. Higher-Order Bounds

To prove the higher-order bounds, we first use the Poisson equation to write $\Delta_1^2 f_h(x^q)$ in terms of $h(x^q)$, $\mathbb{E} h(X)$, and first-order differences of $f_h(x^q)$. With the help of this expression, we use the dynamics of the JSQ model to relate all the second-order differences to each other and prove that

$$|\Delta_1^{a_1} \Delta_2^{a_2} f_h(x_1^q, x_2^q, 0, \ldots, 0)| \leq \delta^2 \sum_{i=1}^{b+1} \mathbb{E} X_i + C(b, \beta) \delta^2 (1 + x_2^q)^2 \tag{21}$$

for $\|a\|_1 = 2$, $x_1^q \leq \delta(n - a_1)$ and $x_2^q \leq \delta(n - a_2)$, followed by a similar bound for $|\Delta_1 + \Delta_2| f_h(0, x_2^q, 0, \ldots, 0)|$. We then bound $\sum_{i=1}^{b+1} \mathbb{E} X_i$ using the Poisson equation in Section 3.2.1 and bound $|\Delta_1^2 f_h(x^q)|$ and $|\Delta_1^2 (\Delta_1 + \Delta_2) f_h(x^q)|$ in Section 3.2.2. In Section 3.2.3, we prove two technical lemmas needed to establish (21). We also briefly discuss (see Remark 1 there) the advantage of using the prelimit generator approach and working with finite differences of $f_h(x^q)$, as opposed to using the classical generator approach and working with the derivatives of the solution to the Poisson equation for the diffusion.

For the following discussion, we assume that $x^q \in S$ with $x_3^q = 0$. Recall from (5) that

$$G_X f_h(x^q) = 1(q_1 < n) n \lambda \Delta_1^2 f_h(x^q - \delta e(1)) + 1(q_1 = n, q_2 < n) n \lambda (\Delta_2 + \Delta_1) f_h(x^q)$$

$$+ \frac{1}{\delta} (\beta - (x_1^q + x_2^q)) \Delta_1 f_h(x^q) - \frac{1}{\delta^2} x_2^q \Delta_2 f_h(x^q - \delta e(2)). \tag{22}$$
We rearrange the Poisson equation \( G_x f_h(x^q) = \mathbb{E} h(X) - h(x^q) \) to see that when \( 0 < q_1 < n \), or alternatively \( 0 < x_1^q < \delta n \),

\[
\Delta_1^2 f_h(x^q - \delta e^{(1)}) = \frac{1}{n\lambda} (\mathbb{E} h(X) - h(x^q)) + \frac{1}{n\lambda \delta} (\beta - (x^q_1 + x^q_2)) \Delta_1 f_h(x^q)
+ \frac{1}{n\lambda \delta} x^q_2 \Delta_2 f_h(x^q - \delta e^{(2)}).
\]

(23)

Note that \( \mathbb{E} |h(X)| \leq C \mathbb{E} (X_1 + \cdots + X_{b+1}) \) since \( h(0) = 0 \) and \( h \in \mathcal{M}_{\text{disc},2}(C) \). Together with the bound on \( \Delta_1 f_h(x^q) \) from (19), this implies that

\[
|\Delta_1^2 f_h(x^q)| \leq \delta^2 C \sum_{i=1}^{b+1} \mathbb{E} X_i + \delta^2 x_1^q + C(b, \beta) \delta^2 (1 + x_2^q)^2, \quad x_1^q < \delta (n - 1), \quad x_2^q = 0.
\]

(24)

Similarly, if \( x_1^q = 0 \),

\[
(\Delta_2 + \Delta_1) f_h(x^q) = \frac{1}{n\lambda} (\mathbb{E} h(X) - h(x^q)) + \frac{1}{n\lambda \delta} (\beta - x_2^q) \Delta_1 f_h(x^q) + \frac{1}{n\lambda \delta} \sum_{b} \Delta_2 f_h(x^q - \delta e^{(2)}),
\]

and therefore

\[
|\Delta_2 + \Delta_1| f_h(0, x_2, 0, \ldots, 0) \leq \delta^2 C \sum_{i=1}^{b+1} \mathbb{E} X_i + C(b, \beta) \delta^2 (1 + x_2^q)^2, \quad x_2^q < \delta n.
\]

(26)

Not all second-order differences can be bounded like this. For example, the equation for \( \Delta_2^2 f_h(x^q) \) would involve the third-order difference \( \Delta_2 \Delta_1^2 f_h(x^q) \), which we have not bounded. Instead, the following lemma relates the remaining second-order differences to \( \Delta_1^2 f_h(x^q) \) and \( (\Delta_2 + \Delta_1) f_h(0, x_2^q, 0, \ldots, 0) \) using the structure of the JSQ system. The proof is postponed to Section 3.2.3.

**Lemma 6.** Fix \( h \in \mathcal{M}_{\text{disc},2}(C) \). Then for any \( x^q \in S \) with \( x_3^q = 0 \),

\[
\begin{align*}
|\Delta_1^2 f_h(x^q)| &\leq C \delta^2 + \max_{0 \leq y^q_2 \leq x_2^q} |\Delta_1^2 f_h(0, y^q_2, 0, \ldots, 0)|, \quad \text{provided } x^q + 2\delta e^{(1)} \in S, \\
|\Delta_2 \Delta_1 f_h(x^q)| &\leq C \delta^2 + \max_{0 \leq y^q_2 \leq x_2^q} |\Delta_2 \Delta_1 f_h(0, y^q_2, 0, \ldots, 0)|, \quad \text{provided } x^q + \delta e^{(1)} + \delta e^{(2)} \in S, \\
|\Delta_2^2 f_h(x^q)| &\leq C \delta^2 + \max_{0 \leq y^q_2 \leq x_2^q} |\Delta_2^2 f_h(0, y^q_2, 0, \ldots, 0)|, \quad \text{provided } x^q + 2\delta e^{(2)} \in S.
\end{align*}
\]
We see from Lemma 6 that to bound the second-order differences, we only need bounds on $|\Delta^2 h(x,0)|$ and $|\Delta^2 h(x,0)|$. The former is bounded in (24), and for the latter term, we note that for any $x^q \in S$ with $x^q_1 = x^q_2 = 0$,

$$
|\Delta^2 h(x^q)| = |\Delta^2 h(x^q + \delta e^{(2)}) - \Delta^2 h(x^q)|
$$

$$
= |(\Delta^2 + \Delta_1) h(x^q + \delta e^{(2)}) - \Delta_1 h(x^q + \delta e^{(2)}) - \Delta^2 h(x^q)|
$$

$$
= |(\Delta^2 + \Delta_1) h(x^q + \delta e^{(2)}) + (\Delta h(x^q) - \Delta h(x^q + \delta e^{(1)} + \delta e^{(2)}))|
$$

$$
\leq \delta^2 C \sum_{i=1}^{b+1} \mathbb{E}X_i + C(b, \beta) \delta^2 (1 + x^q_2)^2 + |\Delta h(0, x^q_2, 0, \ldots, 0) - \Delta h(\delta, x^q_2 + \delta, 0, \ldots, 0)|,
$$

(27)

where the inequality follows from (26). The following lemma bounds the last term on the right-hand side, implying that $|\Delta^2 h(0, x^q_2, 0, \ldots, 0)| \leq \delta^2 C \sum_{i=1}^{b+1} \mathbb{E}X_i + C(b, \beta) \delta^2 (1 + x^q_2)^2$, and, consequently, (21). It is proved in Section 3.2.3.

**Lemma 7.** For all $n \geq 1$,

$$
|\Delta h(0, x^q_2, 0, \ldots, 0) - \Delta h(\delta, x^q_2 + \delta, 0, \ldots, 0)| \leq C(b, \beta) \delta^2 (1 + x^q_2), \quad 0 \leq x^q_2 < \delta n.
$$

(28)

**3.2.1. Bounding $\sum_{i=1}^{b+1} \mathbb{E}X_i$.** The bounds in (21) and (26) do not yet look like the stated bounds in Proposition 2 because the term $\sum_{i=1}^{b+1} \mathbb{E}X_i$ is present. However, we can bound this expectation using the Poisson equation as follows. Recall that $\lambda = 1 - \beta/\sqrt{n}$, let $x(\infty) = (\delta(n - [n\lambda]), 0, \ldots, 0) = (\beta + \delta(n\lambda - [n\lambda]), 0, \ldots, 0)$, and observe that this point is in $S$. In fact, it is the closest point in $S$, when rounded up, to the fluid equilibrium of the JSQ system, which happens to be $(\beta, 0, \ldots, 0)$; cf. Braverman (2020). From (22) we have

$$
G_x h(x(\infty)) = n\lambda \Delta^2 h(x(\infty) - \delta e^{(1)}) + (n\lambda - [n\lambda]) \Delta h(x(\infty)) = \mathbb{E}h(X) - h(x(\infty)).
$$

Choosing $h(x^q) = \sum_{i=1}^{b+1} x^q_i$ and noting that $h(x(\infty)) = \beta + \delta(n\lambda - [n\lambda])$ yields

$$
n\lambda \Delta^2 h(x(\infty) - \delta e^{(1)}) + (n\lambda - [n\lambda]) \Delta h(x(\infty)) - \beta - \delta(n\lambda - [n\lambda]) = \sum_{i=1}^{b+1} \mathbb{E}X_i.
$$

(29)

To bound $\sum_{i=1}^{b+1} \mathbb{E}X_i$ we need only bound $\Delta^2 f_h(x(\infty) - \delta e^{(1)})$, because $|\Delta^1 h(x(\infty))| \leq \delta C(b, \beta)$ due to (19). Note that we cannot use (24) for the second-order difference bound
because \( \sum_{i=1}^{b+1} E X_i \) is present on the right-hand side there. Instead, we exploit the structure of the JSQ model to bound \( \Delta_1^2 f(x(\infty) - \delta e^{(1)}) \) as follows.

Define \( \tau^-(x_1^q) = \inf_{t \geq 0} \{ X(t) = (x_1^q - \delta, 0, \ldots, 0) \mid X(0) = (x_1^q, 0, \ldots, 0) \} \), let \( (X(t), \tilde{X}(t)) \) be the scaled version of the coupling defined in Lemma 4, and let \( V \) be the unit-rate exponentially distributed random variable defined in the same lemma. Fix \( x^q = (x_1^q, 0, \ldots, 0) \) with \( x_1^q \geq 2 \delta \), and suppose \( X(0) = x^q \) and \( \tilde{X}(0) = x^q - \delta e^{(1)} \). Consider the evolution of \( (X(t), \tilde{X}(t)) \) for \( t \in [0, V \wedge \tau^-(x_1^q)] \). If \( V < \tau^-(x_1^q) \), the two processes couple and become identical. Otherwise, the joint process is in state \((x^q - \delta e^{(1)}, x^q - 2\delta e^{(1)})\). Using the strong Markov property, we conclude that

\[
\Delta_1 f_h(x^q - \delta e^{(1)}) = \int_0^\infty \mathbb{E}_{x^q} \left[ \left( h(X(t)) - h(X(t) - \delta e^{(1)}) \right) 1(t \leq (V \wedge \tau^-(x_1^q))) \right] dt \\
+ \mathbb{P}( V \geq \tau^-(x_1^q) ) \Delta_1 f_h(x^q - 2\delta e^{(1)}).
\]

Choosing \( x_1^q = x_1(\infty) \), we see that

\[
\Delta_1 f_h(x(\infty) - \delta e^{(1)}) - \Delta_1 f_h(x(\infty) - 2\delta e^{(1)}) \\
= \int_0^\infty \mathbb{E}_{x(\infty)} \left[ \left( h(X(t)) - h(X(t) - \delta e^{(1)}) \right) 1(t \leq (V \wedge \tau^-(x_1(\infty)))) \right] dt \\
- \mathbb{P}( V < \tau^-(x_1(\infty)) ) \Delta_1 f_h(x(\infty) - 2\delta e^{(1)}).
\]

Choosing \( h(x^q) = \sum_{i=1}^{b+1} x_i^q \) and using \( |\Delta_1 f(x(\infty))| \leq \delta C(b, \beta) \), we arrive at

\[
|\Delta_1^2 f_h(x(\infty) - \delta e^{(1)})| \leq \delta E \tau^-(x_1(\infty)) + C(b, \beta) \delta \mathbb{P}( V < \tau^-(x_1(\infty))).
\] (30)

The quantities involving \( \tau^-(x_1(\infty)) \) are bounded in the following lemma.

**Lemma 8.** There exists a constant \( C(\beta) > 0 \) such that for all \( n \geq 1 \),

\[
E \tau^-(x_1^q) \leq C(\beta)\delta, \quad \text{and} \quad \mathbb{P}( V \leq \tau^-(x_1^q) ) \leq C(\beta)\delta, \quad \text{for} \quad x_1^q \in \{x_1(\infty), \delta, 2\delta\}.
\] (31)

Lemma 8 is proved in Appendix B.5. It implies that \(|\Delta_1^2 f_h(x(\infty) - \delta e^{(1)})| \leq C(b, \beta)\delta^2 \), and therefore

\[
\sum_{i=1}^{b+1} E X_i \leq C(b, \beta).
\] (32)
Combining (32) with (21) proves the second-order bounds in Proposition 2.

Before moving on, let us make a few remarks. The bound in (32) implies that the sequence of steady-state distributions \( \{X_n\}_{n=1}^{\infty} \) is tight and, when combined with process-level convergence of \( \{X(t)\} \) to the diffusion \( \{Y(t)\} \), tightness can be used to imply convergence of the steady-state distributions via a limit-interchange argument; for an example of this applied to the JSQ model, see Braverman (2020). Alternatively, (32) can be recast into a result about the convergence rate to the mean-field equilibrium.

Let \( h(x) = |x_1 + \ldots + x_{b+1} - \beta| - \beta \), noting that \( h \in M_{\text{disc},1}(1) \) and that \( h(0) = 0 \), and suppose for the sake of exposition that \( [n\lambda] = n\lambda \). One may check that the bound in (30) holds even when \( h \in M_{\text{disc},1}(1) \), in which case (29) implies that

\[
\mathbb{E} \left| \sum_{i=1}^{b+1} X_i - \beta \right| = \beta + n\lambda \Delta_1^2 f_h(x(\infty) - \delta e^{(1)}) \leq C(b, \beta).
\]

If we divide both sides by \( \sqrt{n} \) to consider the mean-field scaled version of \( \sum_{i=1}^{b+1} X_i \), we get

\[
\mathbb{E} \left| (n - Q_1)/n + \sum_{i=2}^{b+1} Q_i/n - \beta \right| \leq C(b, \beta)/\sqrt{n}.
\]

Thus, we recover the \( 1/\sqrt{n} \) rate of convergence to the mean field equilibrium that one typically obtains using Stein’s method for the mean-field model, like in Ying (2017). The approach used to show tightness in this section can offer an alternative to the one proposed by Ying (2017), but the difficulty of implementing our approach is directly related to the difficulty of obtaining the relevant Stein factor bounds.

As a final remark, in this section we have shown that establishing tightness, or rates of convergence to the mean-field equilibrium, is equivalent to bounding the first- and-second-order differences of \( f_h(x^q) \) at a single point near the fluid equilibrium of the CTMC. In contrast, establishing rates of convergence to the diffusion requires bounds on the second- and-third-order differences at all points in the support of \( Y \).

### 3.2.2. Third-Order Bounds.

To bound \( \Delta_3^2 f_h(x^q) \), we recall (23), which says that for \( 0 < x_1^q < \delta n \) with \( x_3 = 0 \),

\[
\Delta_3^2 f_h(x^q - \delta e^{(1)}) = \frac{1}{n\lambda} \left( \mathbb{E} h(X) - h(x^q) \right) - \frac{1}{n\lambda \delta} (\beta - (x_1^q + x_2^q)) \Delta_1 f_h(x^q)
\]

\[
+ \frac{1}{n\lambda \delta} x_2^q \Delta_2 f_h(x^q - \delta e^{(2)}).
\]
Applying $\Delta_1$ to both sides yields

$$
\Delta_1^3 f_h(x^q - \delta e^{(1)}) = -\frac{1}{n\lambda} \Delta_1 h(x^q) - \frac{1}{n\lambda} \left( \beta - (x_1^q + x_2^q) \right) \Delta_1^2 f_h(x^q) + \frac{1}{n\lambda} \Delta_1 f_h(x^q + \delta e^{(1)}) \\
+ \frac{1}{n\lambda} \frac{x^q}{\delta} \Delta_1 \Delta_2 f_h(x^q - \delta e^{(2)}), \quad 0 < x_1^q < \delta(n - 1).
$$

The bounds on the first- and second-order differences of $f_h(x^q)$, together with the fact that $h \in M_{disc,2}(C)$, imply that

$$
|\Delta_1^3 f_h(x^q)| \leq C(b, \beta) \delta^3 (1 + x_2^q)^3, \quad x^q \in S, \ x_1^q \leq \delta(n - 3),
$$

which matches the inequality in Proposition 2. The bound on $|\Delta_2^2 - (\Delta_1 + \Delta_2)| f_h(x^q)$ when $x_1^q = 0$ is proved identically by subtracting $(\Delta_1 + \Delta_2) f_h(x^q)$ in (25) from $\Delta_2^2 f_h(x^q)$ in (23). This concludes the proof of Proposition 2.

### 3.2.3. Proving Lemmas 6 and 7

To conclude the section, we prove the auxiliary lemmas from Section 3.2.

**Proof of Lemma 6** Our first task is to bound

$$
\Delta_1^2 f_h(x^q) = \int_0^\infty \left( E_{x^q + 2\delta e^{(1)}} h(X(t)) - 2 E_{x^q + \delta e^{(1)}} h(X(t)) + E_{x^q} h(X(t)) \right) dt.
$$

Note that $x^q \in S$ with $x^q + 2\delta e^{(1)} \in S$ implies that $q_2 \leq q_1 - 2$. Working with the unscaled CTMC, we now construct four processes $\{\tilde{Q}^{(1)}(t)\}, \ldots, \{\tilde{Q}^{(4)}(t)\}$ defined on the time interval $[0, \tau_1(n)]$, where

$$
\tau_1(n) = \inf \{ \tilde{Q}^{(1)}(t) = n \} = \inf \{ \tilde{X}^{(1)}(t) = 0 \}.
$$

We refer to $\{\tilde{Q}^{(i)}(t)\}$ as the $i$th process. Process four is a copy of $\{Q(t)\}$. Numbers two and three are copies of four, but with one extra customer, who is assigned to a server with an empty buffer. The extra customer in two is different from the one in three. Lastly, process one is a copy of four, but with two extra customers. The extra customers are the same as those in two and three. Figure 6 visualizes the initial condition of the processes.

Let $\{\tilde{X}^{(1)}(t)\}, \ldots, \{\tilde{X}^{(4)}(t)\}$ be the scaled counterparts of these processes. Note that

$$
\Delta_1^2 f_h(x^q) = \int_0^\infty \left( E_{x^q + 2\delta e^{(1)}} h(X(t)) - 2 E_{x^q + \delta e^{(1)}} h(X(t)) + E_{x^q} h(X(t)) \right) dt
$$

$$
= \int_0^\infty E_{\tilde{X}^{(1)}(0) = x^q} \left( \left( h(\tilde{X}^{(4)}(t)) - h(\tilde{X}^{(3)}(t)) \right) - \left( h(\tilde{X}^{(2)}(t)) - h(\tilde{X}^{(1)}(t)) \right) \right) dt.
$$
We refer to the different customers according to their shapes in Figure 6. Define \( \tau_s \) and \( \tau_d \) to be the service times of the server with the star and diamond customer, respectively. Both are exponentially distributed with unit mean. Setting \( \tau_m = \min\{\tau_s, \tau_d, \tau_1(n)\} \), we observe that if \( \tau_m = \tau_s \), then

\[
\tilde{X}^{(1)}(t) = \tilde{X}^{(3)}(t), \quad \tilde{X}^{(2)}(t) = \tilde{X}^{(4)}(t), \quad t \geq \tau_m,
\]

and if \( \tau_m = \tau_d \), then

\[
\tilde{X}^{(1)}(t) = \tilde{X}^{(2)}(t), \quad \text{and} \quad \tilde{X}^{(3)}(t) = \tilde{X}^{(4)}(t), \quad t \geq \tau_m.
\]

Therefore,

\[
\int_0^\infty \mathbb{E}_{X^{(1)}(0)=x} \left( (h(\tilde{X}^{(4)}(t)) - h(\tilde{X}^{(3)}(t))) - (h(\tilde{X}^{(2)}(t)) - h(\tilde{X}^{(1)}(t))) \right) dt \\
= \mathbb{E}_{\tilde{X}^{(1)}(0)=x} \int_0^{\tau_m} \left( (h(\tilde{X}^{(4)}(t)) - h(\tilde{X}^{(3)}(t))) - (h(\tilde{X}^{(2)}(t)) - h(\tilde{X}^{(1)}(t))) \right) dt \\
+ \mathbb{P}_{X^{(1)}(0)=x}(\tau_m = \tau_1(n)) \mathbb{E}_{X^{(1)}(0)=x} \left[ \Delta^2 f_h(0, X^{(1)}_2(\tau_1(n)), 0, \ldots, 0) \big| \tau_m = \tau_1(n) \right].
\]

(34)

Since \( \tilde{X}^{(4)}(t) = \tilde{X}^{(3)}(t) + \delta e^{(1)} = \tilde{X}^{(2)}(t) + \delta e^{(1)} = \tilde{X}^{(1)}(t) + 2\delta e^{(1)} \) for \( 0 \leq t \leq \tau_m \),

\[
\left| (h(\tilde{X}^{(4)}(t)) - h(\tilde{X}^{(3)}(t))) - (h(\tilde{X}^{(2)}(t)) - h(\tilde{X}^{(1)}(t))) \right| = \left| \Delta^2 h(\tilde{X}^{(1)}(t)) \right| \leq C\delta^2,
\]

where the last inequality follows from \( h \in \mathcal{M}_{disc,2}(C) \). Combining this with the facts that \( \tilde{X}^{(1)}_2(\tau_1(n)) \leq \tilde{X}^{(1)}_2(0) \) and \( \mathbb{E}_x \tau_m \leq \mathbb{E}_x \tau_s = 1 \), we conclude that the right-hand side of (34) is bounded by \( C\delta^2 + \left| \Delta^2 f_h(0, x^{(q)}, 0, \ldots, 0) \right| \), which proves the bound on \( \left| \Delta^2 f_h(x^{(q)}) \right| \).
The remaining bounds are proved similarly, starting with $|\Delta_2 \Delta_1 f_h(x^q)|$. Fix $x^q \in S$ with $x^q_3 = 0$, and consider

$$\Delta_2 \Delta_1 f_h(x^q) = \left( f_h(x^q + \delta e^{(1)} + \delta e^{(2)}) - f_h(x^q + \delta e^{(1)}) \right) - \left( f_h(x^q + \delta e^{(2)}) - f_h(x^q) \right).$$

We again construct a coupling $\{\tilde{X}^{(1)}(t)\}, \ldots, \{\tilde{X}^{(4)}(t)\}$ corresponding to the four initial states on the right-hand side above. The initial conditions of the unscaled processes are visualized in Figure 7. Our construction yields

$$\tilde{Q}_1^{(0)}(0) = \tilde{Q}_2^{(0)}(0) = \tilde{Q}_3^{(0)}(0) = \tilde{Q}_4^{(0)}(0)$$

Figure 7 The initial state of the four systems. The red customers represent those common to all four systems.

$$\Delta_2 \Delta_1 f_h(x^q) = \int_0^\infty \mathbb{E}_{\tilde{X}^{(1)}(0) = x^q} \left( h(\tilde{X}^{(4)}(t)) - h(\tilde{X}^{(3)}(t)) \right) - \left( h(\tilde{X}^{(2)}(t)) - h(\tilde{X}^{(1)}(t)) \right) dt. \tag{35}$$

Let $\nu_1 = \inf_{t \geq 0} \{\tilde{Q}_1^{(3)}(t) = n\}$. We again let $\tau_s$ and $\tau_d$ be the remaining service time of the server with the star and diamond customer, respectively, and set $\tau_m = \min\{\tau_s, \tau_d, \nu_1\}$. Just like we argued before, if $\tau_m = \tau_s$, then the integrand in (35) is zero after $\tau_m$. If, however, $\tau_m = \tau_d$, then

$$\tilde{X}_2^{(1)}(\tau_m) = \tilde{X}_2^{(2)}(\tau_m) = \tilde{X}_2^{(3)}(\tau_m) = \tilde{X}_2^{(4)}(\tau_m),$$

$$\tilde{X}_1^{(2)}(\tau_m) + 2\delta = \tilde{X}_1^{(1)}(\tau_m) + \delta = \tilde{X}_1^{(4)}(\tau_m) + \delta = \tilde{X}_1^{(3)}(\tau_m),$$

and if $\tau_m = \nu_1$, then $\tilde{X}_1^{(i)}(\tau_m) = 0$ for $1 \leq i \leq 4$ and

$$\tilde{X}_2^{(2)}(\tau_m) = \tilde{X}_2^{(1)}(\tau_m) + \delta = \tilde{X}_2^{(4)}(\tau_m) + \delta = \tilde{X}_2^{(3)}(\tau_m) + 2\delta.$$
Therefore,
\[
\Delta_2 \Delta_1 f_h(x^q) = \mathbb{E}_{\hat{X}^{(1)}(0)=x^q} \int_0^{\tau_m} \left( (h(\hat{X}^{(4)}(t)) - h(\hat{X}^{(3)}(t))) - (h(\hat{X}^{(2)}(t)) - h(\hat{X}^{(1)}(t))) \right) dt \\
+ \mathbb{P}_{\hat{X}^{(1)}(0)=x^q} (\tau_m = \tau_d) \mathbb{E}_X \left[ -\Delta_1^2 f_h(\hat{X}^{(2)}(\tau_d)) \bigg| \tau_m = \tau_d \right] \\
+ \mathbb{P}_{\hat{X}^{(1)}(0)=x^q} (\tau_m = \nu_1) \mathbb{E}_X \left[ \Delta_2^2 f_h(0, \hat{X}^{(3)}(\nu_1), 0, \ldots, 0) \bigg| \tau_m = \nu_1 \right] \\
\leq C \delta^2 + |\Delta_1^2 f_h(0, x^q_2, 0, \ldots, 0)| + |\Delta_2^2 f_h(0, x^q_2, 0, \ldots, 0)|. 
\]

Figure 8 illustrates the coupling needed to bound $|\Delta_2^2 f_h(x^q)|$. The idea of the proof is again to wait until $\tau_1(n)$ and analyze what could happen if one of the servers containing the star or diamond customer completes service before $\tau_1(n)$. We leave the details to the reader. □

**Figure 8** The coupling needed to bound $|\Delta_2^2 f_h(x^q)|$.

**Remark 1.** Let us say a few words on the advantage of using the prelimit generator comparison approach over the classical generator comparison approach. Lemma 6 is proved using a synchronous coupling of four JSQ systems. The four systems are initialized one or two customers apart from one another and due to the discrete state space of the CTMC, all four systems stay one or two customers apart until they couple. Had we used the classical generator comparison approach, we would have needed to carry out a similar analysis by coupling four copies of the diffusion $\{Y(t)\}$. However, unlike the JSQ coupling, the four diffusions would not maintain their initial spacing relative to each other because $\{Y(t)\}$ takes values in a continuous state space. This would further complicate the analysis as we would now need to keep track of the positions of the four diffusions relative to each other.
Proof of Lemma 7  We want to bound

\[
|f_h(0, x_2^2, 0, \ldots, 0) - f_h(\delta, x_2^2 + \delta, 0, \ldots, 0)| = \left| \int_0^\infty (\mathbb{E}_{(0, x_2^2, 0, \ldots, 0)} h(X(t)) - \mathbb{E}_{(\delta, x_2^2 + \delta, 0, \ldots, 0)} h(X(t))) dt \right|
\]

As we are accustomed to doing by now, let us construct a coupling \( \{\tilde{Q}^{(1)}(t), \tilde{Q}^{(2)}(t)\} \) with

\[
\tilde{Q}^{(1)}(0) = (n, q_2, 0, \ldots, 0), \quad \text{and} \quad \tilde{Q}^{(2)}(0) = (n - 1, q_2 + 1, 0, \ldots, 0).
\]

System two has one less idle server and one more customer waiting in a buffer compared to system one, but the total initial customer count is identical across both systems. The initial condition of both systems is visualized in Figure 9. We assume that the diamond and star customers are independent of each other, that the systems see identical arrivals, and that the rest of the customers are identical across both systems.

![Figure 9](image-url)  
**Figure 9**  The initial state of the two systems in an example where \( n = 6 \). The red circles represent customers common to both systems.

Now define \( \tau_d \) and \( \tau_s \) to be the remaining service times of the server that has the diamond and star customer, respectively; let \( \nu_1 = \inf_{t \geq 0} \{ \tilde{Q}^{(2)}_1(t) = n \} \); and set \( \tau_m = \min \{ \tau_s, \tau_d, \nu_1 \} \).

If \( \tau_m = \nu_1 \) or \( \tau_m = \tau_d \), then \( \tilde{Q}^{(1)}(t) \overset{d}{=} \tilde{Q}^{(2)}(t) \) for \( t \geq \tau_m \). Letting \( \{ \tilde{X}^{(i)}(t) \} \) be the scaled version of \( \{ \tilde{Q}^{(i)}(t) \} \), it follows that

\[
\begin{align*}
&f_h(0, x_2^2, 0, \ldots, 0) - f_h(\delta, x_2^2 + \delta, 0, \ldots, 0) \\
&= \mathbb{E}_{\tilde{X}^{(1)}(0) = (0, x_2^2, 0, \ldots, 0)} \int_0^{\tau_m} (h(\tilde{X}^{(1)}(t)) - h(\tilde{X}^{(2)}(t))) dt \\
&\quad + \mathbb{P}_{\tilde{X}^{(1)}(0) = (0, x_2^2, 0, \ldots, 0)}(\tau_m = \tau_s) \mathbb{E}_{x_2} \left[ -\Delta_2 f_h(\tilde{X}^{(1)}(\tau_s)) \bigg| \tau_m = \tau_s \right].
\end{align*}
\]

To bound the first term on the right-hand side, note that

\[
\left| \mathbb{E}_{\tilde{X}^{(1)}(0) = (0, x_2^2, 0, \ldots, 0)} \int_0^{\tau_m} (h(\tilde{X}^{(1)}(t)) - h(\tilde{X}^{(2)}(t))) dt \right| \leq C\delta \mathbb{E}_{\tilde{X}^{(2)}(0) = (\delta, x_2^2 + \delta, 0, \ldots, 0)} \nu_1 \leq C(\beta)\delta^2.
\]
The first inequality is true because $h \in \mathcal{M}_{\text{disc},2}(C)$, and the last inequality follows from Lemma 8 with $x^q_1 = \delta$ there. Furthermore,

$$
P_{\tilde{X}^{(1)}(0)= (0, x^q_2, 0, \ldots, 0)}(\tau_m = \tau_s) \left| \mathbb{E}_x \left[ -\Delta_2 f_h(\tilde{X}^{(1)}(\tau_s)) \gvd \tau_m = \tau_s \right] \right| 
\leq P_{\tilde{X}^{(1)}(0)= (0, x^q_2, 0, \ldots, 0)}(\tau_s < \nu_1) C(b, \beta) \delta^2 (1 + x^2_2).
$$

The first inequality follows from the bound on the first-order difference in (19) together with the fact that $\tilde{X}_2^{(1)}(t) \leq x^2_2$ for all $t \in [0, \tau_m]$. The second inequality follows by noting that $\tau_s$ is independent of $\nu_1$ and using Lemma 8 with $x^q_1 = \delta$, $\tau^{-}(x^q_1) = \nu_1$, and $V = \tau_s$ there.

□

4. Conclusion

As stated in the introduction, the Stein factor bounds require the bulk of our efforts. Proving the first-order bounds in Section 3.1 amounts to considering two coupled JSQ systems, initialized with a difference of one customer, and bounding the expected coupling time of this joint chain. We bound the coupling time by considering a sequence of coupling attempts where the probability of coupling in a single attempt is bounded away from zero uniformly in $n$, and the expected inter-attempt times are also bounded from above, uniformly in $n$. The coupling time can then be bounded by a sum of a geometrically distributed number of random variables representing the inter-attempt durations. This renewal-like argument applies more generally to settings where (a) there is a region of the state space where the joint chain is guaranteed to couple provided it spends enough time there and (b) one can control the expected time to reach this region and the probability of coupling in the region before leaving it.

With the first-order Stein factor bounds in hand, the higher-order bounds require less effort. Our proofs of the high-order bounds make heavy use of the transition structure of the JSQ system, and, in particular, that $Q_2(t), \ldots, Q_{b+1}(t)$ increase only at those times when $Q_1(t) = n$. Readers should not be mislead into thinking that high-order Stein factor bounds require less effort than first-order bounds for all models. Indeed, in the classical generator comparison approach, high-order bounds require much more effort; e.g., Mackey and Gorham (2016), Erdogdu et al. (2019), Jin et al. (2021).
Regarding extending our results, we note that Proposition 1, which compares $G_X$ to $G_Y$, can be easily adjusted to hold for other parameter regimes and load-balancing policies. The main difficulty would be establishing Stein factor bounds. As mentioned in the introduction, Zhao et al. (2021) considered the super-Halfin-Whitt regime ($\frac{1}{2} < \alpha < 1$) and established several hitting-time estimates similar to the ones we use in the proof of Lemma 5 to bound the first-order Stein factors. It may be possible to build on their results and obtain rates of convergence for the super-Halfin-Whitt regime too.

Furthermore, it seems that the sub-Halfin-Whitt regime ($0 < \alpha < \frac{1}{2}$) should present less of a challenge than our own setting. Recall from the discussion in Section 3.1 that coupling of the joint CTMC is guaranteed provided it enters $\Theta_1^Q$ and spends an exponentially distributed amount of time there before all servers become busy. Compared to the Halfin-Whitt regime, the rate at which customers arrive in the sub-Halfin-Whitt regime is much smaller, so the event that all servers are busy should happen less frequently. Indeed, Liu and Ying (2020) showed that the steady-state probability that all servers are busy tends to zero in the sub-Halfin-Whitt regime. Consequently, the Stein factor bounds should be simpler to establish.

**Appendix A: Supporting Proofs for Section 2**

We first prove Lemma 2 and then introduce the operator $A$ in Appendix A.1. Once $A$ is introduced, we prove Proposition 1 in Appendix A.2.

**Proof of Lemma 2** Initialize $Y(0)$ according to $Y$. Since $\{Y(t)\}$ satisfies (1), for any $f \in C^2(\mathbb{R}^{k+1})$ with $E|f(Y)| < \infty$, Itô’s lemma implies that

$$0 = E[f(Y(1))] - E[f(Y(0))] = E \int_0^1 G_Y f(Y(s))ds + E \int_0^1 \left( \frac{\partial}{\partial x_1} f(Y(s)) + \frac{\partial}{\partial x_2} f(Y(s)) \right) 1(Y_1(s) = 0)dU(s).$$

(37)

If $E|G_Y f(Y)| < \infty$, then $E \int_0^1 G_Y f(Y(s))ds = E G_Y f(Y)$ follows from the Fubini-Tonelli theorem. □

**A.1. The Interpolator $A$**

The operator $A$ discussed in this section is identical to the one introduced in Appendix A of Braverman (2022), but we repeat its key properties here as they are needed for the proof of Proposition 1. Consider a one-dimensional function $f : \delta \mathbb{Z} \rightarrow \mathbb{R}$. We can extend it to $\mathbb{R}$ by defining

$$Af(x) = \sum_{i=0}^{\delta} \delta^{k(x)} f(\delta(k(x) + i)).$$
where \( k(x) = \lfloor x/\delta \rfloor \) and \( \alpha^k_{i+1} : \mathbb{R} \to \mathbb{R} \) are weights defined for all \( k \in \mathbb{Z} \) and \( i = 0, \ldots, 4 \). The function \( Af(x) \) is a weighted sum of the five points \( f(\delta k(x)), \ldots, f(\delta (k(x) + 4)) \). We mention the reason for using five points after stating Theorem 2. Note that if \( f(x) \) is defined only on a subset of \( \delta \mathbb{Z} \), then \( Af(x) \) can still be defined, provided that \( f(\delta k(x)), \ldots, f(\delta (k(x) + 4)) \) are defined. Braverman (2022) described how to choose these weights to make \( Af(x) \) coincide with \( f(\cdot) \) on grid points, and also to make it a differentiable function whose derivatives behave like the corresponding finite differences of \( f(\cdot) \). The idea can be applied to multidimensional grid-valued functions as well.

The following result is Theorem 2 of Braverman (2022). We use this as an interface that contains the important properties of \( A \) without delving into the low-level details behind its construction.

**Theorem 2.** Given a convex set \( K \subset \mathbb{R}^d \), define

\[
K_4 = \{ x \in K \cap \delta \mathbb{Z}^d : \delta(k(x) + i) \in K \cap \delta \mathbb{Z}^d \text{ for all } 0 \leq i \leq 4\epsilon \},
\]

let \( \text{Conv}(K_4) \) be the convex hull of \( K_4 \), and, for \( x \in \mathbb{R}^d \), define \( k(x) \) by \( k(x) = \lfloor x/\delta \rfloor \). There exist weights \( \{ \alpha^k_{i+1} : \mathbb{R} \to \mathbb{R}, k \in \mathbb{Z}, i = 0, 1, 2, 3, 4 \} \) such that for any \( f : K \cap \delta \mathbb{Z}^d \to \mathbb{R} \), the function

\[
Af(x) = \sum_{i_d=0}^{4} \alpha^k_{k_d(x)+i_d}(x_d) \cdots \sum_{i_1=0}^{4} \alpha^k_{k_1(x)+i_1}(x_1) f(\delta(k(x) + i))
\]

\[
= \sum_{i_1,\ldots,i_d=0}^{4} \left( \prod_{j=1}^{d} \alpha^k_{k_j(x)+i_j}(x_j) \right) f(\delta(k(x) + i)), \quad x \in \text{Conv}(K_4)
\]

(38)

satisfies \( Af(x) \in C^3(\text{Conv}(K_4)) \), where \( i = (i_1, \ldots, i_d) \) in (38). Additionally, \( Af(x) \) is infinitely differentiable almost everywhere on \( \text{Conv}(K_4) \),

\[
Af(\delta k) = f(\delta k), \quad \delta k \in K_4,
\]

(39)

and there exists a constant \( C(d) > 0 \) independent of \( f(\cdot), x, \) and \( \delta \), such that

\[
\left| \frac{\partial^r}{\partial x^a} Af(x) \right| \leq C(d)\delta^{-\|a\|_1} \max_{0 \leq j_1,\ldots,j_d \leq 4} |\Delta x_{j_1}^{a_1} \cdots \Delta x_{j_d}^{a_d} f(\delta(k(x) + i))|, \quad x \in \text{Conv}(K_4),
\]

(40)

for \( 0 \leq \|a\|_1 \leq 3 \), and (40) also holds when \( \|a\|_1 = 4 \) for almost all \( x \in \text{Conv}(K_4) \). Additionally, the weights \( \{ \alpha^k_{i+1} : \mathbb{R} \to \mathbb{R}, k \in \mathbb{Z}, i = 0, 1, 2, 3, 4 \} \) are degree-7 polynomials in \( (x - \delta k)/\delta \) whose coefficients do not depend on \( k \) or \( \delta \). They satisfy

\[
\alpha^k_{0}(\delta k) = 1, \quad \text{and} \quad \alpha^k_{i+1}(\delta k) = 0, \quad k \in \mathbb{Z}, \ i = 1, 2, 3, 4,
\]

(41)

\[
\sum_{i=0}^{4} \alpha^k_{i+1}(x) = 1, \quad k \in \mathbb{Z}, \ x \in \mathbb{R},
\]

(42)

and also the following translational invariance property:

\[
\alpha^k_{i+j+1}(x + \delta j) = \alpha^k_{i+1}(x), \quad i, j, k \in \mathbb{Z}, \ x \in \mathbb{R}.
\]

(43)
The proof is identical for all indices, so we assume that such that for any $x$

Furthermore, there exists some $\varepsilon$ the grid, we refer to $A$ as an interpolator. For the interested reader, $A$ is a degree-7 polynomial spline. From (39) we see that $A$ is a linear operator, and (42) implies that $A$ applied to a constant simply equals that constant. Before we can prove Proposition 1, we require one more lemma.

**Remark 2.** The bound in (40) holds almost everywhere when $\|a\|_1 = 4$. This bound is the reason we need to use $f(\delta k(x))$ and the four points to the right of it (in each dimension). By using more (fewer) points, one can alter the theorem so that (40) holds for larger (smaller) values of $\|a\|_1$. It is worth noting that to prove the results in this paper, we do not go beyond $\|a\|_1 = 3$.

Going forward, we let $A$ be the operator described in Theorem 2. Since $Af$ coincides with $f$ on the grid, we refer to $A$ as an interpolator. For the interested reader, $A$ is a degree-7 polynomial spline. From (39) we see that $A$ is a linear operator, and (42) implies that $A$ applied to a constant simply equals that constant. Before we can prove Proposition 1, we require one more lemma.

**Lemma 9.** In the setting of Theorem 2, for any $k \in K_4$ and $1 \leq j \leq d$,

$$\frac{\partial}{\partial x_j} Af(x) \bigg|_{x = \delta k} = \delta^{-1} \left( \Delta_j - \frac{1}{2} \Delta_j^2 + \frac{1}{3} \Delta_j^3 \right) f(\delta k).$$

(44)

Furthermore, there exists some $\varepsilon : \text{Conv}(K_4) \to \mathbb{R}$ satisfying

$$|\varepsilon(x)| \leq C(d) \delta^{-1} \max_{0 \leq i \leq 4e} |\Delta_i^{a_i} \cdots \Delta_i^{a_d} f(\delta(k(x) + i))|$$

such that for any $x \in \text{Conv}(K_4)$,

$$\frac{\partial}{\partial x_j} Af(x) = \delta^{-1} \Delta_j f(\delta k(x)) + \varepsilon(x).$$

**Proof of Lemma 9** The proof is identical for all indices, so we assume that $j = 1$. Fix $\delta k \in K_4$ and let $g(x_1) = Af(x_1, \delta k_2, \ldots, \delta k_d)$ be a function in $x_1$ only. The form of $Af(x)$ in (38), together with (41), implies that

$$\frac{\partial}{\partial x_1} Af(x) \bigg|_{x = \delta k} = g'(\delta k_1).$$

It follows that $g'(\delta k_1) = P'_{k_1}(\delta k_1)$, where $P_{k_1}(x)$ is a polynomial defined in (A.1) of Braverman (2022). Furthermore, (A.1) implies that

$$P'_{k_1}(\delta k_1) = \delta^{-1} \left( \Delta_1 - \frac{1}{2} \Delta_1^2 + \frac{1}{3} \Delta_1^3 \right) g(\delta k_1) = \delta^{-1} \left( \Delta_1 - \frac{1}{2} \Delta_1^2 + \frac{1}{3} \Delta_1^3 \right) f(\delta k),$$

from which (44) follows. To prove the second claim of the lemma, we write

$$\frac{\partial}{\partial x_j} Af(x) = \delta^{-1} \left( \Delta_j - \frac{1}{2} \Delta_j^2 + \frac{1}{3} \Delta_j^3 \right) f(\delta k(x)) + \frac{\partial}{\partial x_j} Af(x) - \frac{\partial}{\partial x_j} Af(x) \bigg|_{x = \delta k(x)}.$$

Now $|\Delta_j^3 f(\delta k(x))| \leq \max \{|\Delta_i^{a_i} \cdots \Delta_i^{a_d} f(\delta(k(x) + i))| : 0 \leq i \leq 4e, \|a\|_1 = 2\}$,

$$|\Delta_j^3 f(\delta k(x))| = |\Delta_j^3 f(\delta(k(x) + e(i)) - \Delta_j^3 f(\delta k(x))| \leq \max_{|\alpha|_1 = 2} |\Delta_i^{a_i} \cdots \Delta_i^{a_d} f(\delta(k(x) + i))|,$$
and
\[
\left| \frac{\partial}{\partial x_j} Af(x) - \frac{\partial}{\partial x_j} Af(x) \right|_{x=\delta k(x)} \leq \sum_{j'=1}^d |x_{j'} - \delta k_{j'}(x)| \left| \frac{\partial^2}{\partial x_j \partial x_{j'}} Af(\xi) \right| \leq C(d)\delta^{-1} \max_{0 \leq \varepsilon \leq 4} \left| \Delta_1 \ldots \Delta_d f(\delta(k(x) + i)) \right|,
\]
where $\xi$ is some point between $\delta k(x)$ and $x$. The last inequality follows from (40) and the fact that $|x_j - \delta k_j(x)| \leq \delta$.

Note that some of the bounds in Theorem 2 and Lemma 9 have a constant $C(d)$ depending on the dimension $d$ of the function; e.g., (40). In the JSQ model $d = b + 1$, but when proving Proposition 1 in the next section we can assume that $d = 2$ because of the following. Given a function $f : \delta \mathbb{N}_{b+1} \to \mathbb{R}$, we can use (38) and (41) of Theorem 2, and the fact that $Y_i = 0$ for $i > 2$, to see that
\[
Af(Y) = \sum_{i_1=0}^4 \alpha_{b+1}^0 (0) \sum_{i_2=0}^4 \alpha_{i_2}^0 (0) \sum_{i_1=0}^4 \alpha_{k_1(Y)+i_2}^1 (Y_2) \sum_{i_1=0}^4 \alpha_{k_1(Y)+i_2}^1 (Y_1) f(\delta(k(Y) + i))
\]
\[
= \sum_{i_2=0}^4 \alpha_{k_2(Y)+i_2}^1 (Y_2) \sum_{i_1=0}^4 \alpha_{k_1(Y)+i_2}^1 (Y_1) f(\delta(k_1(Y) + i_1), \delta(k_2(Y) + i_2), 0, \ldots, 0).
\]
Since $k_j(Y)$ depends only on $Y_j$, we see that $Af(Y)$ is actually a bivariate function. In Appendix A.2, we treat any function of the form $Af(Y)$ as a function of two variables.

### A.2. Proving Proposition 1

Fix $h \in \mathcal{M}_{\text{disc},2}(C)$. We recall from (5) that for $x^q \in S$,
\[
G_X f(x^q) = -1(q_1 < n)n\lambda \Delta_1 f(x^q - \delta e^{(1)}) + n\lambda \sum_{j=1}^b 1(q_1 = \ldots = q_j = n, q_{j+1} < n) \Delta_{j+1} f(x^q)
\]
\[
+ (q_1 - q_2)\Delta_1 f(x^q) - \sum_{j=2}^b (q_j - q_{j+1}) \Delta_j f(x^q - \delta e^{(j)}) - q_{b+1} \Delta_{b+1} f(x^q - \delta e^{(b+1)})
\]
and $f_h(x^q)$ is the unique solution to the Poisson equation
\[
G_X f_h(x^q) = \mathbb{E}h(X) - h(x^q), \quad x^q \in S
\]
with $f_h(0) = 0$. Also recall that we extended $f_h(x^q)$ by setting $f_h(x^q) = 0$ for $x^q \in \delta \mathbb{N}_{b+1} \setminus S$, and defined
\[
B = \{(x_1, x_2, 0, \ldots, 0) \in \mathbb{R}_{+}^{b+1} : x_2 + x_1 \leq \delta(n/2 - 8) = (n/2 - 8)/\sqrt{n}\}
\]
and
\[
I = \{i = (i_1, i_2, 0, \ldots, 0) \in \mathbb{N}^{b+1} : 0 \leq i_1, i_2 \leq 4\}.
\]
We first argue that $\mathbb{E}|Ah(Y)| < \infty$, $\mathbb{E}|Af_h(Y)| < \infty$, and $\mathbb{E}|G_Y Af_h(Y)| < \infty$, which together imply that (12) holds. The latter two statements follow immediately from the fact that $f_h(x^q)$, and therefore $Af_h(x)$, have compact support. Since $h \in \mathcal{M}_{\text{disc,2}}(C)$, inequality (40) of Theorem 2 implies that $Ah(Y)$ is Lipschitz and therefore, $\mathbb{E}|Ah(Y)| < \infty$ due to Lemma 3, which states that the moments of $Y_i$ are finite.

Next we argue that $AG_x f_h(x) = Eh(X) - Ah(x)$ for all $x \in B$. Given the Poisson equation (45) and the definition of $A$ in (38) of Theorem 2, it suffices to show that $\delta(k(x) + i) \in S$ for all $i \in I$.

From the definition of $S_Q$ in (4) we know that any point $q \in S_Q$ satisfies $0 \leq q_2 \leq q_1 \leq n$. The corresponding points $x^q \in S$ satisfy $x^q_i \geq 0$, $x^q_i \geq 0$, and $x^q_i + x^q_i = \delta(n - q_1) + \delta q_2 \leq \delta n$. The latter inequality says that the combined number of idle servers and servers with at least one person waiting in the buffer cannot exceed $n$. Now, provided that $n > 16$, any point $\delta k$ in

$$B \cap \delta \mathbb{N}^{b+1} = \{(x_1, x_2, 0, \ldots, 0) \in \mathbb{R}_+^{b+1} : x_2 + x_1 \leq \delta(n/2 - 8)\} \cap \delta \mathbb{N}^{b+1},$$

must satisfy $\delta(k + i) \in S$ for all $i \in I$ because $\delta(k_1 + i_1) + \delta(k_2 + i_2) \leq \delta n / 2$. Finally, recall that

$$\varepsilon_1(Y) = (AG_x f_h(Y) - G_Y Af_h(Y))1(Y \in B),$$
$$\varepsilon_2(Y) = (Eh(X) - Ah(Y) - G_Y Af_h(Y))1(Y \notin B),$$
$$\varepsilon_3(Y) = \left(\frac{\partial}{\partial x_1} Af_h(Y) + \frac{\partial}{\partial x_2} Af_h(Y)\right)1(Y \in B),$$
$$\varepsilon_4(Y) = \left(\frac{\partial}{\partial x_1} Af_h(Y) + \frac{\partial}{\partial x_2} Af_h(Y)\right)1(Y \notin B).$$

We bound $\varepsilon_2(Y)$, $\varepsilon_3(Y)$, and $\varepsilon_4(Y)$ in Appendix A.2.1 and bound $\varepsilon_1(Y)$ in Appendix A.2.2.

A.2.1. Bounding $\varepsilon_2(Y)$ through $\varepsilon_4(Y)$. We begin with the bound on

$$|\varepsilon_2(Y)| \leq |Ah(Y)|1(Y \notin B) + 1(Y \notin B)\mathbb{E}|h(X)| + |G_Y Af_h(Y)|1(Y \notin B).$$

The facts that $Ah(Y)$ is Lipschitz, that $Ah(0) = h(0) = 0$, and that $h \in \mathcal{M}_{\text{disc,2}}(C)$ imply that

$$|Ah(Y)|1(Y \notin B) \leq C(Y_1 + Y_2)1(Y \notin B)$$
and
$$1(Y \notin B)\mathbb{E}|h(X)| \leq 1(Y \notin B)C\mathbb{E}(X_1 + \cdots + X_{b+1}) \leq 1(Y \notin B)C(b, \beta),$$

where the last inequality follows from inequality (32). To bound the remaining term, we recall (40) of Theorem 2, which says that

$$\left|\frac{\partial^n}{\partial x^a} Af(Y)\right| \leq C\delta^{-\|a\|_1} \max_{0 \leq i_j \leq 4 - a_j} |\Delta_1 a_1 \Delta_2^2 f(\delta(k(Y) + i))| \leq C\delta^{-\|a\|_1} \max_{i \in I} |f(\delta(k(Y) + i))| \quad (46)$$
for $1 \leq \|a\|_1 \leq 3$. Combined with this bound, the definition of $G_Y$ in (8) implies that
\[
|G_Y A f_h(Y)| = \left| (\beta - (Y_1 + Y_2)) \frac{\partial}{\partial x_1} A f_h(Y) - Y_2 \frac{\partial}{\partial x_2} A f_h(Y) + \frac{\partial^2}{\partial x_1^2} A f_h(Y) \right| \\
\leq C(\beta) \delta^{-2} (1 + Y_1 + Y_2) \max_{i \in I} |f(\delta(k(Y) + i))|.
\]
Combining the bounds on the three terms yields the bound on $\varepsilon_2(Y)$. Lemma 9 implies the bound on $\varepsilon_3(Y)$ and (46) implies the bound on $\varepsilon_4(Y)$.

**A.2.2. Bounding $\varepsilon_1(Y)$.** Bounding $\varepsilon_1(Y)$ requires more effort. The first thing to note is that the weighted sum representation of $A G_X f_h(Y)$ is difficult to work with. Our first task is therefore to write it in a form that is more amenable to analysis. To this end, we extend the domain of $f_h(x^q)$ to allow either the first or second coordinate to take the value $-\delta$ by defining

\[
\hat{f}_h(x^q) = f_h(x^q), \quad x^q \in \delta N^{b+1},
\]

\[
\hat{f}_h(-\delta, x^q_2, \ldots, x^q_{b+1}) = f_h(0, x^q_2 + \delta, x^q_3, \ldots, x^q_{b+1}),
\]

\[
\hat{f}_h(x^q, -\delta, x^q_2, \ldots, x^q_{b+1}) = (1 - \Delta_2) f_h(x^q_1, 0, x^q_3, \ldots, x^q_{b+1}),
\]

\[
\hat{f}_h(x^q, -\delta, x^q_3, \ldots, x^q_{b+1}) = (1 - \Delta_2) f_h(x^q_1, 0, x^q_3, \ldots, x^q_{b+1}).
\]

The form of $\hat{f}_h(x^q)$ is tied to the transition structure of the JSQ model, and specifically to the “reflection” that occurs near the boundaries $\{x^q_1 = 0\}$ and $\{x^q_2 = 0\}$. Furthermore, the definition of $A$ in Theorem 2 implies that $A f_h(x) = A \hat{f}_h(x)$ for $x \in \mathbb{R}^{b+1}_+$ because $\hat{f}_h = f_h$ on $\delta N^{b+1}$. Having defined $\hat{f}_h(x^q)$, we present the following lemma, which is proved in Appendix A.2.3.

**Lemma 10.** For any $x^q \in B \cap \delta N^{b+1}$,
\[
G_X f_h(x^q) = n \lambda (\hat{f}_h(x^q - \delta e^{(1)}) - \hat{f}_h(x^q)) + (n - (x^q_1 + x^q_2)/\delta) (\hat{f}_h(x^q + \delta e^{(1)}) - \hat{f}_h(x^q))
\]
\[
+ \frac{1}{\delta} x^q_2 (\hat{f}_h(x^q - \delta e^{(2)}) - \hat{f}_h(x^q)).
\]

Consequently, for any $x \in B$,
\[
A G_X f_h(x) = n \lambda (A \hat{f}_h(x - \delta e^{(1)}) - A \hat{f}_h(x)) + (n - (x_1 + x_2)/\delta) (A \hat{f}_h(x + \delta e^{(1)}) - A \hat{f}_h(x))
\]
\[
+ \frac{1}{\delta} x_2 (A \hat{f}_h(x - \delta e^{(2)}) - A \hat{f}_h(x)) + \varepsilon_5(x),
\]
where
\[
\varepsilon_5(x) = \sum_{i=0}^{4} \alpha_{k_2(x)+i_2} \sum_{i_1=0}^{4} \alpha_{k_1(x)+i_1} (x_1) \frac{1}{\delta} (\delta(k_2(x) + i_2) - x_2)
\]
\[
\times \left( - \Delta_2 \hat{f}_h(\delta(k(x) + i - e^{(2)))} + \Delta_2 \hat{f}_h(\delta(k(x) - e^{(2)}) \right)
\]
\[
+ \sum_{i_2=0}^{4} \alpha_{k_2(x)+i_2} \sum_{i_1=0}^{4} \alpha_{k_1(x)+i_1} (x_1) \frac{1}{\delta} (\delta(k_1(x) + i_1 + k_2(x) + i_2) + x_1 + x_2)
\]
\[
\times \left( \Delta_1 \hat{f}_h(\delta(k(x) + i)) - \Delta_1 \hat{f}_h(\delta(k(x)) \right).
\]
We now bound $\varepsilon_1(Y)$ using Lemma 10. Applying Taylor expansion to (49), we have

$$AG_X f_h(Y) = n\lambda \left( -\delta \frac{\partial}{\partial x_1} A\hat{f}_h(Y) + \frac{1}{2}\delta^2 \frac{\partial^2}{\partial x_1^2} A\hat{f}_h(Y) - \frac{1}{6}\delta^3 \frac{\partial^3}{\partial x_1^3} A\hat{f}_h(\xi^1) \right)$$

$$+ (n - (Y_1 + Y_2)/\delta) \left( \delta \frac{\partial}{\partial x_1} A\hat{f}_h(Y) + \frac{1}{2}\delta^2 \frac{\partial^2}{\partial x_1^2} A\hat{f}_h(Y) + \frac{1}{6}\delta^3 \frac{\partial^3}{\partial x_1^3} A\hat{f}_h(\xi^2) \right)$$

$$+ \frac{1}{2}Y_2 \left( -\delta \frac{\partial}{\partial x_2} A\hat{f}_h(Y) + \frac{1}{2}\delta^2 \frac{\partial^2}{\partial x_2^2} A\hat{f}_h(\xi^3) \right) + \varepsilon_5(Y),$$

where $\xi^1, \xi^2,$ and $\xi^3$ are points strictly between $Y - \delta \varepsilon^{(1)}$ and $Y$, $Y + \delta \varepsilon^{(1)}$, and $Y - \delta \varepsilon^{(2)}$ and $Y$, respectively. Recall that $\delta^2 = 1/n, \delta(n - n\lambda) = \beta$, and $G_Y$ from (8), which imply that

$$AG_X f_h(Y) - G_Y A\hat{f}_h(Y) = -\frac{1}{6}\delta^3 \frac{\partial^3}{\partial x_1^3} A\hat{f}_h(\xi^1) + \frac{1}{6}\delta(1 - \delta(Y_1 + Y_2)) \frac{\partial^2}{\partial x_1^2} A\hat{f}_h(\xi^2) + \frac{1}{6}\delta^2 \frac{\partial^2}{\partial x_2^2} A\hat{f}_h(\xi^3) + \varepsilon_5(Y).$$

Note that $A\hat{f}_h(Y) = A f_h(Y)$ because $Y \geq 0$, so $G_Y A\hat{f}_h(Y) = G_Y A f_h(Y)$. We now prove the following four bounds, which together imply the bound on $\varepsilon_1(Y)$:

$$\left| \frac{1}{6}(1 - \delta(Y_1 + Y_2)) \frac{\partial^3}{\partial x_1^3} A\hat{f}_h(\xi^2) \right| \leq C\delta^{-2} \max_{i \in I} \left| \Delta_i^2 f_h(\delta(k(Y) + i)) \right|,$$  \hspace{1cm} (51)

$$\left| \frac{1}{6}\delta \frac{\partial^2}{\partial x_1^2} A\hat{f}_h(\xi^1) \right| \leq C\delta^{-2} \max_{i \in I} \left| \Delta_i^2 f_h(\delta(k(Y) + i)) \right| + C\delta^{-2} (1 \leq \delta) \max_{i \in I} \left| \Delta_i^2 f_h(\delta(k(Y) + i)) \right|,$$ \hspace{1cm} (52)

$$\left| \frac{1}{2}\delta^2 \frac{\partial^2}{\partial x_2^2} A\hat{f}_h(\xi^3) \right| \leq C\delta^{-1} Y_2 \max_{i \in I} \left| \Delta_i^2 f_h(\delta(k(Y) + i)) \right|, \hspace{1cm} \text{and} \hspace{1cm} (53)$$

$$\left| \varepsilon_5(Y) \right| \leq C \max_{a_1 + a_2 = 2} \left| \Delta_i^2 \Delta_i^2 f_h(\delta(k(Y) + i)) \right|.$$ \hspace{1cm} (54)

We begin with (51). Observe that $(1 - \delta(Y_1 + Y_2)) \in (0, 1/2)$ because $Y \in B$. Furthermore, $Y < \xi^2 < Y + \delta \varepsilon^{(1)}$ implies $k(\xi^2) = k(Y) \geq 0$. Combining this with (40) of Theorem 2, we get

$$\left| \frac{1}{6}(1 - \delta(Y_1 + Y_2)) \frac{\partial^3}{\partial x_1^3} A\hat{f}_h(\xi^2) \right| \leq C\delta^{-2} \max_{i \in I} \left| \Delta_i^2 f_h(\delta(k(Y) + i)) \right| = C\delta^{-2} \max_{i \in I} \left| \Delta_i^2 f_h(\delta(k(Y) + i)) \right|.$$

We now prove (52). As before, $Y - \delta \varepsilon^{(1)} < \xi^1 < Y$ implies that $k(\xi^1) = k(Y - \delta \varepsilon^{(1)}) = k(Y) - \varepsilon^{(1)}$, so

$$\left| \frac{1}{6}\delta \frac{\partial^2}{\partial x_1^2} A\hat{f}_h(\xi^1) \right| \leq C\delta^{-2} \max_{i \in I} \left| \Delta_i^2 f_h(\delta(k(Y) - \varepsilon^{(1)} + i)) \right|.$$ \hspace{1cm} (55)

Now when $Y \in [0, \delta)$ and $i_1 = 0$, the definition of $\hat{f}_h(x^n)$ in (47) implies that

$$\hat{f}_h(\delta(k(Y) - \varepsilon^{(1)} + i)) = \hat{f}_h(-\delta, \delta(k_2(Y) + i_2), 0, \ldots, 0) = f_h(0, \delta(k_2(Y) + i_2 + 1), 0, \ldots, 0),$$

from which we see that $\Delta_1 \hat{f}_h(-\delta, \delta(k_2(Y) + i_2), 0, \ldots, 0) = -\Delta_2 f_h(0, \delta(k_2(Y) + i_2), 0, \ldots, 0)$, and therefore

$$\Delta_2 \hat{f}_h(-\delta, \delta(k_2(Y) + i_2), 0, \ldots, 0) = (\Delta_1^2 - (\Delta_1 + \Delta_2)) f_h(0, \delta(k_2(Y) + i_2), 0, \ldots, 0).$$
Combining this with (55) implies (52). To prove (53), we note that $k(\xi^3) = k(Y) - e^{(2)}$ because $Y - \delta e^{(2)} < \xi^3 < Y$, so
\[
\left| \delta Y_2 \frac{\partial^2}{\partial x_2^2} A f_h(\xi^3) \right| \leq C \delta^{-1} Y_2 \max_{i \in I} \left| \Delta_2^2 \hat{f}_h(\delta(k(Y) - e^{(2)} + i)) \right|.
\]
The definition of $\hat{f}_h(x^0)$ in (47) says that $\Delta_2 \hat{f}_h(Y_1, -\delta, 0, \ldots, 0) = \Delta_2 \hat{f}_h(Y_1, 0, \ldots, 0)$, so $\Delta_2^2 \hat{f}_h(Y_1, -\delta, 0, \ldots, 0) = 0$, implying (53). Lastly, we prove (54). Theorem 2 tells us that $\alpha_{k_j + i_j}(x_j)$ are degree-7 polynomials in $(x_j - \delta k_j)/\delta$ whose coefficients do not depend on $k_j$ or $\delta$, so there exists a constant $C > 0$ such that $\left| \alpha_{k_j + i_j}(x_j) \right| \leq C$ for $j = 1, 2$, so
\[
\left| \sum_{i_2=0}^{4} \alpha_{k_2(x) + i_2}(x_2) \sum_{i_1=0}^{4} \alpha_{k_1(x) + i_1}(x_1) \frac{1}{\delta} \left( \delta(k_2(x) + i_2) - x_2 \right) \right|
\]
\[
\times \left( - \Delta_2 \hat{f}_h(\delta(k(x) + i - e^{(2)})) + \Delta_2 \hat{f}_h(\delta(k(x) - e^{(2)})) \right)
\]
\[
\leq C \max_{i \in I} \left| \Delta_2^2 \hat{f}_h(\delta(k(x) + i - e^{(2)})) - \Delta_2 \hat{f}_h(\delta(k(x) - e^{(2)})) \right|.
\]
Now
\[
\Delta_2 \hat{f}_h(\delta(k(x) + i - e^{(2)})) - \Delta_2 \hat{f}_h(\delta(k(x) - e^{(2)}))
\]
\[
= \Delta_1 \Delta_2 \hat{f}_h(\delta(k(x) + i_1 e^{(1)} + i_2 e^{(2)} - e^{(2)})) + \sum_{i_2=0}^{i_1-1} \Delta_2^2 \hat{f}_h(\delta(k(x) + i_2 e^{(2)} - e^{(2)})),
\]
implying that
\[
C \max_{i \in I} \left| \Delta_2^2 \hat{f}_h(\delta(k(x) + i - e^{(2)})) - \Delta_2 \hat{f}_h(\delta(k(x) - e^{(2)})) \right|
\]
\[
\leq C \max_{i \in I} \left| \Delta_2^2 \hat{f}_h(\delta(k(Y) - e^{(2)} + i)) \right| + C \max_{i \in I} \left| \Delta_1 \Delta_2 \hat{f}_h(\delta(k(Y) - e^{(2)} + i)) \right|.
\]
An identical argument allows us to bound the second term on the right-hand side of (50), yielding
\[
\varepsilon_5(Y) \leq C \max_{i \in I} \left| \Delta_2^2 \hat{f}_h(\delta(k(Y) - e^{(2)} + i)) \right| + C \max_{i \in I} \left| \Delta_1 \Delta_2 \hat{f}_h(\delta(k(Y) - e^{(2)} + i)) \right|
\]
\[
+ C \max_{i \in I} \left| \Delta_1^2 \hat{f}_h(\delta(k(Y) + i)) \right| + C \max_{i \in I} \left| \Delta_1 \Delta_2 \hat{f}_h(\delta(k(Y) + i)) \right|.
\]
Using $\Delta_2 \hat{f}_h(Y_1, -\delta, 0, \ldots, 0) = \Delta_2 \hat{f}_h(Y_1, 0, \ldots, 0)$ and $\Delta_2^2 \hat{f}_h(Y_1, -\delta, 0, \ldots, 0) = 0$, we conclude (54).
A.2.3. Proving Lemma 10  To prove Lemma 10, we need the following result.

Lemma 11. Suppose \( P \subset \delta \mathbb{Z}^d \) and let \( f, g : P \to \mathbb{R} \). Given \( \ell \in \mathbb{Z}^d \), for those \( k \) such that \( \delta k \in P \) and \( \delta (k + \ell) \in P \), we define

\[
F(\delta k) = g(\delta k)(f(\delta (k + \ell)) - f(\delta k)).
\]

Then \( AF(x) \) is well defined for those \( x \in \mathbb{R}^d \) such that \( \delta (k(x) + i) \in P \) and \( \delta (k(x) + \ell + i) \in P \) for all \( 0 \leq i \leq 4e \), where \( k_i(x) = \lfloor x_i / \delta \rfloor \). Furthermore, for all such \( x \),

\[
AF(x) = Ag(x)(Af(x + \delta) - Af(x)) + \sum_{\ell_1, \ldots, \ell_d = 0} 4 \left( \prod_{j=1}^d \alpha_{k_j(x) + \ell_j}(x_j) \right) \left( g(\delta (k(x) + i)) - Ag(x) \right) \times \left( f(\delta (k(x) + \ell + i)) - f(\delta (k(x) + i)) - (f(\delta (k(x) + \ell)) - f(\delta k(x))) \right).
\]

Proof of Lemma 11  The proof is identical to the proof of Proposition 3 of Braverman (2022). □

Proof of Lemma 10  First, we prove (48). Any \( x^q \in \mathbb{B} \cap \delta \mathbb{N}^{b+1} \) satisfies \( x^q_2 \leq \delta (n/2 - 8) \), or \( q_2 \leq n/2 - 8 \). It follows from the definition of \( G_X \) in (5) that for \( x^q \in \mathbb{B} \cap \delta \mathbb{N}^{b+1} \),

\[
G_X f_k(x^q) = -1(q_1 < n)n\lambda_{1} f_k(x^q - \delta e^{(1)}) + 1(q_1 = n)n\lambda_{2} f(x^q)
\]

\[
+ (q_1 - q_2) \Delta f_k(x^q - \delta e^{(2)}) - q_2 \Delta f_k(x^q - \delta e^{(2)}).
\]

Note that \( q_1 - q_2 = n - (n - q_1) - q_2 = n - (x^q_1 + x^q_2) / \delta \), and \( q_2 = x^q_2 / \delta \). Although \( \Delta f_k(x^q - \delta e^{(2)}) \) is technically not defined when \( x^q_2 = 0 \), we adopt the convention that \( 1(q_2 = 0)q_2 \Delta f_k(x^q - \delta e^{(2)}) = 0 \).

Using the definition of \( \hat{f}(x^q) \) in (47), we have

\[
1(q_2 = 0)q_2 \Delta f(x^q - \delta e^{(2)}) = 0 = 1(q_2 = 0)q_2 \Delta \hat{f}(x^q - \delta e^{(2)}).
\]

Similarly, since \( q_1 = n \) corresponds to \( x^q_1 = 0 \),

\[
n\lambda_{1}(q_1 = n) \Delta_{2} f(x^q) = -1(q_1 = n)n\lambda_{1} \hat{f}(x^q - \delta e^{(1)}),
\]

which proves (48). To prove (49), note that if \( g(x^q) = (n - x^q_1 - x^q_2) / \delta \), then \( Ag(x) = n - (x^q_1 + x^q_2) / \delta \).

To see why, note that \( \Delta_{i,j} g(x^q) = 0 \) for any \( i, j \), so Theorem 2 implies that all second-order partial derivatives of \( Ag(x) \) are zero. Since \( Ag(x) \) is twice continuously differentiable, it must be a linear function, and the only linear function that coincides with \( g(x^q) \) on the grid is \( Ag(x) = n - (x^q_1 + x^q_2) / \delta \).

Similarly, if \( g(x^q) = q_2 = x^q_2 / \delta \), then \( Ag(x) = x_2 / \delta \). Applying Lemma 11 to each of the three terms on the right-hand side of (48) proves (49). □
Appendix B: Supporting Proofs for Section 3

Apart from the short proof of Lemma 8 in Appendix B.5, this appendix is devoted to the proof of Lemma 5. Going forward, we fix \( \gamma = 2(17/\beta + \beta + 1) \), and recall from Section 3.1 that

\[
\theta_1 = n - \lfloor \sqrt{n} \beta / 2 \rfloor, \quad \theta_2 = \lfloor \gamma \sqrt{n} \rfloor,
\]

\[
\tau_i(q_i) = \inf\{ t \geq 0 : Q_i(t) = q_i \}, \quad q_i \in \{0, 1, \ldots, n\}, \ i = 1, 2.
\]

Following the proof roadmap of Lemma 5, we need an upper bound on the expected start of the first cycle and the expected duration of a single cycle. The following two lemmas provide the ingredients for these bounds and are proved in Appendices B.1 and B.2, respectively.

**Lemma 12.** For all \( n \geq 1 \),

\[
\max_{\theta_1 \leq q_1 \leq n} \max_{q_2 = \theta_2, \ q \in S_Q} \mathbb{E}_q\tau_2(2\theta_2) \leq C(b, \beta).
\]

**Lemma 13.** For all \( n \geq 1 \) and \( q \in S_Q \) with \( q_2 > \theta_2 \),

\[
\mathbb{E}_q\tau_2(\theta_2) \leq C(b, \beta)(1 + \delta q_2) = C(b, \beta)(1 + x_q^2), \quad q \in S_Q \text{ with } q_2 > \theta_2.
\]

To bound the probability of coupling in a given cycle, we require the following two lemmas.

**Lemma 14.** There exists a constant \( p_1(\beta) \in (0, 1) \) such that for all \( n \geq 1 \),

\[
\min_{\theta_1 \leq q_1 \leq n, q_2 = \theta_2, \ q \in S_Q} \mathbb{P}_q(\tau_1(\theta_1) < \tau_2(2\theta_2)) \geq p_1(\beta).
\]

**Lemma 15.** There exists a constant \( p_2(b, \beta) \in (0, 1) \) such that for all \( n \geq 1 \),

\[
\min_{0 \leq q_1 \leq \theta_1, \ 0 \leq q_2 \leq 2\theta_2, \ q \in S_Q} \mathbb{P}\left(\tau_C < \tau_1(n) \mid Q(0) = q, (Q(0), \tilde{Q}(0)) \in \bigcup_{i=1}^{b+1} \Theta_i^Q\right) \geq p_2(b, \beta).
\]

Lemmas 14 and 15 are proved in Appendices B.3 and B.4, respectively.

**Proof of Lemma 5** Throughout the proof, we use \( C \) to denote a positive constant that may change from line to line but depends only on \( \beta \) and \( b \). Given any initial condition \((q, \bar{q}) \in \bigcup_{i=1}^{b+1} \Theta_i^Q\),

\[
\mathbb{E}_{(q, \bar{q})}\tau_C \leq C(b, \beta)(1 + \delta q_2).
\]

For convenience, we abuse notation and adopt the convention that

\[
\mathbb{E}_q\tau_C = \max_{(q, \bar{q}) \in \bigcup_{i=1}^{b+1} \Theta_i^Q} \mathbb{E}[\tau_C(\mid Q(0), \tilde{Q}(0) = (q, \bar{q}))], \quad q \in S_Q.
\]
but \( \mathbb{E}_q(W) = \mathbb{E}(W|Q(0) = q) \) for any random variable \( W \) other than \( \tau_C \). We also assume that every max operator in this proof automatically considers the maximum over all \( q \in S_Q \): i.e.,

\[
\max_{q_2 = \theta_2} \mathbb{E}_q \tau_C = \max_{q \in S_Q} \mathbb{E}_q \tau_C.
\]

Lemma 13 implies that for any \( q \in S_Q \),

\[
\mathbb{E}_q \tau_C \leq \mathbb{E}_q \tau_2(\theta_2) + \max_{q_2 = \theta_2} \mathbb{E}_q \tau_C \leq C(1 + \delta q_2) + \max_{q_2 = \theta_2} \mathbb{E}_q \tau_C. \tag{57}
\]

We will argue that if \( p_1 = p_1(\beta) \) and \( p_2 = p_2(b, \beta) \) are the constants from Lemmas 14 and 15, then

\[
\max_{\theta_1 \leq q_1 \leq n, q_2 = \theta_2} \mathbb{E}_q \tau_C \leq C + (1 - p_1 p_2) \max_{q_2 = \theta_2} \mathbb{E}_q \tau_C, \tag{58}
\]

and therefore

\[
\max_{q_2 = \theta_2} \mathbb{E}_q \tau_C \leq C(1 - p_3)^{-1} \leq C. \tag{59}
\]

As a result, choosing \( p_3 = \max\{1 - p_1 p_2, (1 - p_2)\} \in (0, 1) \) implies that

\[
\max_{q_2 = \theta_2} \mathbb{E}_q \tau_C = \max \left\{ \max_{0 \leq q_1 < \theta_1, \theta_1 \leq q_1 \leq n, q_2 = \theta_2} \mathbb{E}_q \tau_C, \max_{0 \leq q_1 \leq n, q_2 = \theta_2} \mathbb{E}_q \tau_C \right\} \leq C + p_3 \max_{q_2 = \theta_2} \mathbb{E}_q \tau_C,
\]

and therefore

\[
\mathbb{E}_q \tau_C \leq C + p_3 \max_{q_2 = \theta_2} \mathbb{E}_q \tau_C \leq C + p_3 \max_{q_2 = \theta_2} \mathbb{E}_q \tau_C. \tag{60}
\]

where in the second inequality we used (56) of Lemma 13. To bound the right-hand side, let us define the events

\[
E_1 = \{ \tau_1(\theta_1) < \tau_2(2\theta_2) \}, \quad \text{and} \quad E_2 = \{ \tau_C < \tau_1(n) \},
\]

and their complements \( E_1^c \) and \( E_2^c \), respectively. Note that if \( Q_2(0) < 2\theta_2 \), then the event \( E_1^c \) implies that \( Q(\tau_M) = (n, 2\theta_2) \) because \( Q_2(t) \) increases only at times when \( Q_1(t) = n \). Using the law of total probability,

\[
\max_{\theta_1 \leq q_1 \leq n, q_2 = \theta_2} \mathbb{E}_q \left[ \mathbb{E}_{Q(\tau_M)} \tau_C \right] \leq \max_{\theta_1 \leq q_1 \leq n, q_2 = \theta_2} \left\{ \mathbb{P}_q(E_1^c) \max_{q_1 = \theta_1} \mathbb{E}_q \tau_C + \mathbb{P}_Q(E_1) \max_{q_1 = \theta_1} \mathbb{E}_q \tau_C \right\} \tag{61}
\]

We note that

\[
\max_{q_1 = \theta_1} \mathbb{E}_q \tau_C \leq \max_{0 \leq q_2 \leq 2\theta_2} \mathbb{E}_q \tau_2(\theta_2) + \max_{q_2 = \theta_2} \mathbb{E}_q \tau_C \leq C + \max_{q_2 = \theta_2} \mathbb{E}_q \tau_C, \tag{62}
\]
where we used Lemma 13 in the last inequality, so

\[ \mathbb{P}_q(E_1) \max_{\theta_1 = \theta_1} E_q \tau C \leq \mathbb{P}_q(E_1) (C + \max_{\theta_2 = \theta_2} E_q \tau C). \]  \hfill (63)

Provided we can show that

\[ \mathbb{P}_q(E_1) \max_{\theta_1 = \theta_1} E_q \tau C \leq \mathbb{P}_q(E_1) (C + (1-p_2) \max_{\theta_2 = \theta_2} E_q \tau C), \]  \hfill (64)

we can combine (63) and (64) with (61) to get

\[ \max_{\theta_1 \leq \theta_1 \leq \theta_2} E_q [E_q(\tau C)\tau C] \leq \max_{\theta_1 \leq \theta_1 \leq \theta_2} \left\{ \mathbb{P}_q(E_1) (C + \max_{\theta_2 = \theta_2} E_q \tau C) + \mathbb{P}_q(E_1) (C + (1-p_2) \max_{\theta_2 = \theta_2} E_q \tau C) \right\} \]

\[ = C + \left( \max_{\theta_2 = \theta_2} E_q \tau C \right) \max_{\theta_1 \leq \theta_1 \leq \theta_2} \left\{ 1 - p_2 \mathbb{P}_q(E_1) \right\} \]

\[ \leq C + (1-p_2 p_1) \max_{\theta_2} E_q \tau C, \]

where the last inequality follows from the lower bound on \( \mathbb{P}_q(E_1) \) in Lemma 14. Combining this bound with (60) proves (58). We now prove (64). Recall that \( E_2 = \{ \tau C < \tau_1(n) \} \) and observe that

\[ \max_{\theta_1 = \theta_1} E_q \tau C \]

\[ \leq \max_{\theta_1 = \theta_1} E_q \left( \left[ \tau C \land \tau_1(n) \right] (0) \right) + \max_{\theta_1 = \theta_1} E_q \left( \left[ \tau C \land \tau_1(n) \right] + E_q(\tau(n)) \tau C \right) \left( 0 \right) \]

\[ \leq 2 \max_{\theta_1 = \theta_1} E_q \left[ \tau C \land \tau_1(n) \right] + \text{max}_{\theta_1 = \theta_1} \left( \mathbb{P}_q \left( E_2 \mid Q(0) = \theta_1, (Q(0), \tilde{Q}(0)) \in \bigcup_{i=1}^{b+1} \Theta_i Q \right) \right) \]

\[ \leq 2 \max_{\theta_1 = \theta_1} E_q \left[ \tau C \land \tau_1(n) \right] + (1-p_2) \max_{\theta_2 = \theta_2} E_q \tau C, \]

where in the last inequality we used Lemma 15 and the fact that \( Q_2(\tau_1(n)) \leq Q_2(0) \) because \( Q_2(t) \) increases only at times when \( Q_1(t) = n \). Applying (62) to the right-hand side, we arrive at

\[ \max_{\theta_1 = \theta_1} E_q \tau C \leq 2 \max_{\theta_1 = \theta_1} E_q \left[ \tau C \land \tau_1(n) \right] + C + (1-p_2) \max_{\theta_2 = \theta_2} E_q \tau C. \]

To conclude, we argue that

\[ \max_{\theta_1 = \theta_1} E_q \left[ \tau C \land \tau_1(n) \right] \leq b + 1. \]  \hfill (65)

If \( (Q(0), \tilde{Q}(0)) \in \Theta_i Q \), then \( (Q(t), \tilde{Q}(t)) \in \Theta_i Q \) for all \( t \in [0, \tau_1(n)] \) by construction. The joint CTMC couples before \( \tau_1(n) \) if \( \tau_1(n) > V \), where \( V \) is as in (17). If \( (Q(0), \tilde{Q}(0)) \in \Theta_i Q \) for \( i \geq 2 \), coupling will happen before \( \tau_1(n) \) if the joint CTMC transitions to \( \Theta_i Q \) and then spends \( V \) time units there, all
before $\tau_1(n)$. From the construction of $\tilde{Q}(\cdot)$, we know that the time taken to get from $\Theta_1^Q$ to $\Theta_1^Q$ equals the sum of $i-1$ unit-mean exponentially distributed random variables, so the worst case is when $i = b + 1$. Letting $\Gamma_{b+1}$ represent this sum, it follows that
\[
\max_{q_i = \theta_1} \mathbb{E}_q [\tau_C \wedge \tau_1(n)] \leq \max_{0 \leq q_2 \leq 2\theta_2} \mathbb{E}_q [\Gamma_{b+1} \wedge \tau_1(n)] \leq \mathbb{E}(\Gamma_{b+1}) \leq b + 1,
\]
which proves (65). Our argument for (64) can be repeated to prove (59).

B.1. Proving Lemma 12

Proof of Lemma 12 Define $V(x^q) = \sum_{i=1}^{b+1} q_i$ and observe that
\[
G_x V(x^q) = n\lambda_1(q_{b+1} < n) - q_1, \quad x^q \in S.
\]
Since $\theta_1 = n - \lfloor \sqrt{n} \beta / 2 \rfloor$, it follows that for any $q \in S_Q$ with $\theta_1 < q_1 \leq n$,
\[
G_x V(x^q) = n\lambda - q_1 \leq n\lambda - (n - \lfloor \sqrt{n} \beta / 2 \rfloor) = b\sqrt{n} + \lfloor \sqrt{n} \beta / 2 \rfloor \leq -\sqrt{n} \beta / 2.
\]
Let $M > 0$, $t^{(M)} = \min\{\tau_1(\theta_1), \tau_2(\theta_2), M\}$, and note that $Q_1(t) \geq n - \lfloor \sqrt{n} \beta / 2 \rfloor$ for $t \leq t^{(M)}$. Dynkin’s formula, e.g., Lemma 17.2 in Kallenberg (2001), then implies that for any $q \in S_Q$ with $\theta_1 < q_1 \leq n$ and $q_2 = \theta_2$,
\[
\mathbb{E}_x V(X(t^{(M)})) - V(x^q) = \mathbb{E}_x \int_0^{t^{(M)}} G_x V(X(s)) ds = -\frac{\sqrt{n} \beta}{2} \mathbb{E}_x t^{(M)}.
\]
Since $Q_1(t^{(M)}) \geq n - \lfloor \sqrt{n} \beta / 2 \rfloor$ and $\theta_1 < q_1 \leq n$, it follows that $q_1 - Q_1(t^{(M)}) \leq \lfloor \sqrt{n} \beta / 2 \rfloor$, so
\[
\frac{\sqrt{n} \beta}{2} \mathbb{E}_x t^{(M)} \leq V(x^q) - \mathbb{E}_x V(X(t^{(M)})) \leq q_1 - \mathbb{E}_x V(X(t^{(M)})) + \sum_{i=2}^{b+1} q_i \leq \lfloor \sqrt{n} \beta / 2 \rfloor + b\theta_2,
\]
where in the last inequality we used $q_2 \geq q_3 \geq \ldots \geq q_{b+1}$. Dividing both sides by $\sqrt{n}$, and noting that $\theta_2 / \sqrt{n} \leq \gamma = 2(17/\beta + \beta + 1)$, yields $\mathbb{E}_x t^{(M)} \leq C(b, \beta)$. We conclude by taking $M \to \infty$ and using the monotone convergence theorem.

B.2. Proving Lemma 13

Recall that $\theta_2 = \lfloor \gamma \sqrt{n} \rfloor$ and $\gamma = 2(17/\beta + \beta + 1)$. In this section we show that $\mathbb{E}_x \tau_2(\theta_2) \leq C(b, \beta)(1 + \delta q_2)$ if $q_2 > \theta_2$. Our proof is based on a Lyapunov function characterized by the following proposition, proved in Appendix B.2.1.

Lemma 16. There exists a function $V : \mathbb{R}_{+}^{b+1} \to \mathbb{R}$ such that for any $n \geq 1$ and any $x^q \in S$ with $x_2^q \geq 2(17/\beta + \beta + \delta)$,
\[
G_x V(x^q) \leq -3/17 + \frac{\delta}{\beta} q_3 1(b > 1) - n\lambda_1(q_1 = q_2 = n).
\]
Furthermore, there exists a constant $C(\beta) > 0$ such that for any $n \geq 1$,
\[
0 \leq V(x) \leq C(\beta)(1 + x_2), \quad x \in \mathbb{R}^{b+1} \text{ with } x_2 \geq 2(17/\beta + \beta).
\]
Proof of Proposition 13 Let $V(x)$ be the function in Lemma 16, fix $X(0) = x^g \in S$ with $x^g \geq \delta \theta_2$, $M > 0$, and define $\tau^M_2(\theta_2) = M \wedge \tau_2(\theta_2)$. Dynkin’s formula says that

$$E_{x^g} V(X(\tau^M_2(\theta_2))) - V(x^g) = E_{x^g} \int_0^{\tau^M_2(\theta_2)} G_X V(X(t))dt.$$  \hspace{1cm} (67)

Since $X_2(t) \geq \delta \theta_2 \geq 2(17/\beta + \beta) + \delta$ for all $t \in [0, \tau^M_2(\theta_2)]$, Lemma 16 implies that

$$G_X V(X(t)) \leq -3/17 + \frac{\delta}{\beta} (Q_3(t)1(b > 1) - n\lambda_1(Q_1(t) = Q_2(t) = n)), \hspace{0.5cm} t \in [0, \tau^M_2(\theta_2)].$$

Combining this inequality with (67) and that $V(X(\tau^M_2(\theta_2))) \geq 0$ and $V(x^g) \leq C(\beta)x^g_3$ yields

$$\frac{3}{17}E_{x^g} (\tau^M_2(\theta_2)) \leq C(\beta)x^g_3 + \frac{\delta}{\beta} E_{x^g} \int_0^{\tau^M_2(\theta_2)} (Q_3(t)1(b > 1) - n\lambda_1(Q_1(t) = Q_2(t) = n))dt.$$  

If $b = 1$, the lemma follows trivially, so we assume that $b > 1$. It suffices to show that

$$E_{x^g} \int_0^{M} (Q_3(t)1(b > 1) - n\lambda_1(Q_1(t) = Q_2(t) = n))dt \leq \sum_{i=3}^{b+1} q_i,$$

because $\sum_{i=3}^{b+1} q_i \leq bq_2$. Since $Q_3(t)$ is the number of servers with at least two customers in their buffers, it is also the number of customers that are second in line at time $t$. Thus, $\int_0^{M} Q_3(t)dt$ is the cumulative time spent by customers being second in line. This cumulative time is contributed to by customers already in the system at time $t = 0$ and by new arrivals after $t = 0$. Of those customers present in the system at $t = 0$, the number that are, or could at some point become, second in line is $\sum_{i=3}^{b+1} q_i$, and each will spend at most one unit of time being second in line, in expectation.

Let $N$ be the number of customers in the interval $[0,M]$ that arrive when all servers are busy and all queues have at least one customer in them; i.e., $Q_1(t) = Q_2(t) = n$. For $1 \leq i \leq N$, let $\xi_i$ be the time customer $i$ spends being second in line, even if that customer becomes second in line after time $M$. We argue that conditioned on $\{N \geq i\}$, each $\xi_i$ is exponentially distributed with unit mean. Upon entry into the system, if customer $i$ is routed to a busy server with only one other customer waiting in the buffer, then $\xi_i$ is distributed according to the remaining service time of the server, which is exponentially distributed with unit mean. If the buffer has more than one other customer waiting, then $\xi_i$ equals the service time of the customer two spots ahead of customer $i$, which is also exponentially distributed with unit mean. Further note that the CTMC can be constructed in such a way that the value of $\xi_i$ is determined at the instant when customer $i$ enters the system, so

$$E_{x^g} \int_0^{M} Q_3(t)dt \leq \sum_{i=3}^{b+1} q_i + E_{x^g} \sum_{i=1}^{\infty} \xi_i 1(N \geq i) = \sum_{i=3}^{b+1} q_i + \sum_{i=1}^{\infty} E_{x^g} (\xi_i 1(N \geq i) \mathbb{P}(N \geq i)) = \sum_{i=3}^{b+1} q_i + E_{x^g} N.$$

Let $\eta_i$ be the time spent by the CTMC in a state with $Q_1(t) = Q_2(t) = n$ before customer $i$’s arrival. Since the arrivals to the JSQ system are governed by a rate-$n\lambda$ Poisson process, the arrival
of customer \(i\) corresponds to a time when \(\eta_i\) accumulates to equal an exponentially distributed 
random variable with rate \(n\lambda\), and therefore

\[
\mathbb{E}_{x_0} \int_0^M 1(Q_1(t) = Q_2(t) = n)dt \geq \mathbb{E}_{x_0} \sum_{i=1}^{\infty} \eta_i 1(N \geq i) = \sum_{i=1}^{\infty} \mathbb{E}_{x_0}(\eta_i|N \geq i)P(N \geq i) = \frac{1}{n\lambda} \mathbb{E}_{x_0}N.
\]

\[\square\]

**B.2.1. Proving Lemma 16.** The Lyapunov function in Lemma 16 is based on the fluid limit of 
the JSQ system, studied in Braverman (2020). Lemma 16 was, unfortunately, not proved there, but 
that paper contains all the necessary ingredients for the proof. We now recall them, using notation 
from Braverman (2020).

Consider the two-dimensional process \(\{(Q_1(t) - n)/n, Q_2(t)/n\}\). Note that the first coordinate 
is nonpositive, whereas so far we have been using a nonnegative first coordinate. Section 4.1 of 
Braverman (2020) described the fluid limit of this process. Letting 
\(\Omega = \{x \in \mathbb{R}^2 : x_1 \leq 0, x_2 \geq 0\}\), 
the fluid limit is a dynamical system \(v : \mathbb{R}_+ \rightarrow \Omega\) with initial condition 
\(v(0) = x \in \Omega\); we write \(v_x(t)\) to emphasize the relationship on \(x\). Postponing the discussion of the behavior of 
\(v_x(t)\), for \(\ell, u \in \mathbb{R}\) with \(\ell < u\) define the smoothed indicator function \(\phi^{(\ell,u)} : \mathbb{R} \rightarrow [0,1]\) by

\[
\phi^{(\ell,u)}(x) = \begin{cases} 
0, & x \leq \ell, \\
\frac{(x-\ell)^2}{(u+\ell/2-u/2)^2(u-\ell)} + \frac{2}{((u+\ell)/2-\ell/2)^2(u-\ell)}, & x \in [\ell,(u+\ell)/2], \\
1 - (x-u)^2 \left(\frac{(x-\ell)^2}{(u+\ell/2-\ell/2)^2(u-\ell)} + \frac{2}{((u+\ell)/2-\ell/2)^2(u-\ell)}\right), & x \in [(u+\ell)/2,u], \\
1, & x \geq u,
\end{cases}
\]

and let

\[
f^{(2)}(x) = \int_0^\infty \phi^{(\kappa_1,\kappa_2)}(v^x(t))dt, \quad x \in \Omega,
\]

where \(\delta = 1/\sqrt{n}\) and \(\kappa_1, \kappa_2 \in \mathbb{R}\) are to be determined. The function \(f^{(2)}(x)\) appeared in Section 5.1 
of Braverman (2020), where it was used as a Lyapunov function for the diffusion limit of the JSQ 
system; i.e., the process \(\{Y(t)\}\) in (1). We show that this is also a Lyapunov function for the CTMC. 
Define

\[
V(x) = f^{(2)}(-\delta x_1, \delta x_2), \quad x \in \mathbb{R}_+^{b+1}.
\]

The following result proved in Appendix B.2.2 gives us control over the derivatives of \(V(x)\).
LEMMA 17. For any $x \in \mathbb{R}_+^{b+1}$ with $x_2 \geq \kappa_2$,
\[
(\beta - (x_1 + x_2)) \frac{\partial V(x)}{\partial x_1} - \delta x_2 \frac{\partial V(x)}{\partial x_2} = -1, \quad \text{and} \quad 1(x_1 = 0)\left( \frac{\partial}{\partial x_1} + \frac{\partial}{\partial x_2} \right)V(x) = 0. \tag{70}
\]

Furthermore, if we choose $\kappa_1 = 17/\beta + \beta$ and $\kappa_2 = 2\kappa_1$, then for any $x \in \mathbb{R}_+^{b+1}$ with $x_2 \geq \kappa_2$, and any $x_2' \geq x_2$,
\[
\frac{\partial^2}{\partial x_1^2} V(x) \leq 9/17, \tag{71}
\]
\[
\frac{\partial}{\partial x_2} V(x) \leq \frac{\partial}{\partial x_2} V(0, x_2') = \frac{1}{\beta}, \quad \frac{\partial^2}{\partial x_2^2} V(x) \leq 5/17, \tag{72}
\]

and there exists a constant $C(\beta)$ such that $0 \leq V(x) \leq C(\beta)(1 + x_2)$.

Proof of Lemma 16  Let $\kappa_1 = 17/\beta + \beta$ and $\kappa_2 = 2\kappa_1$, and $V(x)$ be the function from Lemma 17, and recall $G_X$ defined in (5). Since $V(x)$ depends only on $x_1$ and $x_2$,

\[
G_X V(x^q) = 1(q_1 < n)n\lambda (-\Delta_1 V(x^q - \delta e^{(1)})) + n\lambda 1(q_1 = n, q_2 < n)\Delta_2 V(x^q)
\]
\[
+ (q_1 - q_2)\Delta_1 V(x^q) + (q_2 - q_3 1(b > 1))(-\Delta_2 V(x^q - \delta e^{(2)})), \quad x^q \in S.
\]

Using Taylor expansion, we get
\[
-\Delta_1 V(x^q - \delta e^{(1)}) = V(x^q - \delta e^{(1)}) - V(x^q) = -\delta \frac{\partial}{\partial x_1} V(x^q) + \int_{x_1}^{x_1 - \delta} (u - (x_1^q - \delta)) \frac{\partial^2}{\partial x_1^2} V(u, x_2^q)du,
\]
\[
\Delta_1 V(x^q) = V(x + \delta e^{(1)}) - V(x^q) = \delta \frac{\partial}{\partial x_1} V(x^q) + \int_{x_1}^{x_1 + \delta} (x_1^q + \delta - u) \frac{\partial^2}{\partial x_1^2} V(u, x_2^q)du,
\]

and a similar expression holds for $\Delta_2 V(x^q)$ and $-\Delta_2 V(x^q - \delta e^{(2)})$. Therefore,
\[
G_X V(x^q) = -\delta (1(q_1 < n)n\lambda - (q_1 - q_2)) \frac{\partial}{\partial x_1} V(x^q) + \delta (1(q_1 = n, q_2 < n)n\lambda - q_2) \frac{\partial}{\partial x_2} V(x^q)
\]
\[
- q_3 1(b > 1)(-\Delta_2 V(x^q - \delta e^{(2)})) + \psi(x^q), \tag{73}
\]

where
\[
\psi(x^q) = n\lambda 1(q_1 < n) \int_{x_1^q - \delta}^{x_1^q} (u - (x_1^q - \delta)) \frac{\partial^2}{\partial x_1^2} V(u, x_2^q)du
\]
\[
+ n\lambda 1(q_1 = n, q_2 < n) \int_{x_2^q}^{x_2^q + \delta} (x_2^q + \delta - u) \frac{\partial^2}{\partial x_2^2} V(x_1^q, u)du
\]
\[
+ (q_1 - q_2) \int_{x_1^q}^{x_1^q + \delta} (x_1^q + \delta - u) \frac{\partial^2}{\partial x_1^2} V(u, x_2^q)du + q_2 \int_{x_2^q - \delta}^{x_2^q} (u - (x_2^q - \delta)) \frac{\partial^2}{\partial x_2^2} V(x_1^q, u)du.
\]
Now suppose $x_2^* \geq \kappa_2 + \delta$. The bounds on the second-order derivatives of $V(x)$ from Lemma 17, together with the facts that $q_1 - q_2 \geq 0$, $q_2 \geq 0$, $\delta^2 n \lambda \leq 1$, and $\delta^2 q_1 \leq 1$, imply that $\psi(x^*) \leq 14/17$.

Next, we rewrite the first line on the right-hand side of (73), for which we note that

$$
\lambda = 1 - \beta/\sqrt{n}, \quad x_i^* = \delta(n - q_1), \quad 1(q_1 = n) = 1(x_i^* = 0),
$$

$$
1(q_1 < n) = 1 - 1(x_i^* = 0), \quad 1(q_1 = n, q_2 < n) = 1(x_i^* = 0) - 1(q_1 = q_2 = n),
$$

so

$$
- \delta(1(q_1 < n)n\lambda - (q_1 - q_2)) \frac{\partial}{\partial x_1} V(x^*) + \delta(1(q_1 = n, q_2 < n)n\lambda - q_2) \frac{\partial}{\partial x_2} V(x^*)
$$

$$
= \left(\beta - (x_1^* + x_2^*)\right) \frac{\partial}{\partial x_1} V(x^*) - x_2^* \frac{\partial}{\partial x_2} V(x) - 1(q_1 = q_2 = n)\delta n\lambda \frac{\partial}{\partial x_2} V(x^*) + \delta n\lambda(1(x_i^* = 0) - \frac{\partial}{\partial x_1} + \frac{\partial}{\partial x_2}) V(x^*)
$$

$$
= - 1 - 1(q_1 = q_2 = n)\delta n\lambda \frac{\partial}{\partial x_2} V(x^*),
$$

where the last equality is due to (70) from Lemma 17. We have thus shown that

$$
G_x V(x^*) \leq - 1 + 14/17 - 1(q_1 = q_2 = n)\delta n\lambda \frac{\partial}{\partial x_2} V(x^*) - q_3 1(b > 1)\left(- \Delta_2 V(x^* - \delta e^{(2)})\right).
$$

Now $V(x^*) = V(0, \sqrt{n})$ when $q_1 = q_2 = n$, so (72) in Lemma 17 tells us that $V(0, \sqrt{n}) = 1/\beta$ provided that $\sqrt{n} \geq \kappa_2 = 2(\beta/17 + \beta)$, which we assume, so

$$
- 1(q_1 = q_2 = n)\delta n\lambda \frac{\partial}{\partial x_2} V(x) - q_3 1(b > 1)\left(- \Delta_2 V(x^* - \delta e^{(2)})\right)
$$

$$
= - 1(q_1 = q_2 = n)\delta n\lambda \left(\frac{1}{\beta} + q_3 1(b > 1) \int_{x_i^* - \delta}^{x_i^*} \frac{\partial}{\partial x_2} V(x_i^*, u) du\right)
$$

$$
\leq \frac{\delta}{\beta} (q_3 1(b > 1) - n\lambda(1(q_1 = q_2 = n)),
$$

where the inequality follows from (72) in Lemma 17.

\[ \square \]

**B.2.2. Proof of Lemma 17.** Fix $\kappa_1 = 17/\beta + \beta$ and $\kappa_2 = 2\kappa_1$. The function $f^{(2)}(x)$ was considered in Lemma 8 of Braverman (2020), which tells us that that

$$
-(\beta \delta + x_1 - x_2) \frac{\partial}{\partial x_1} f^{(2)}(x) - x_2 \frac{\partial}{\partial x_2} f^{(2)}(x) = -1, \quad x \in \Omega \text{ with } x_2 > \kappa_2/\sqrt{n},
$$

$$
\frac{\partial}{\partial x_1} f^{(2)}(x) = \frac{\partial}{\partial x_2} f^{(2)}(x), \quad x \in \Omega \text{ with } x_1 = 0.
$$

Combining this with

$$
\frac{\partial}{\partial x_1} V(x) = -\delta \frac{\partial}{\partial x_1} f^{(3)}(-\delta x_1, \delta x_2), \quad \frac{\partial}{\partial x_2} V(x) = \delta \frac{\partial}{\partial x_2} f^{(2)}(-\delta x_1, \delta x_2)
$$

(74)
gives us (70). Going forward, we assume that \( x \in \Omega \). Let us bound the derivatives of \( V(x) \). On page 1100 of Braverman (2020), it was shown that
\[
\frac{\partial^2 f^{(2)}(x)}{\partial x_1^2} \leq \frac{n}{\beta(\kappa_1 - \beta)} + \frac{\kappa_1}{\kappa_1 - \beta} \frac{4n}{\beta(\kappa_2 - \kappa_1)} = \frac{n}{17} + \frac{17/\beta + \beta}{\beta(17/\beta + \beta)} \frac{4n}{17} = \frac{5n}{17}, \quad x \in \Omega,
\]
implying the bound on \( \partial^2 V(x)/\partial x_2^2 \) in (71). We now prove (72), followed by the bound on \( V(x) \). Unfortunately, \( \partial f^{(2)}(x)/\partial x_2 \) and \( \partial^2 f^{(2)}(x)/\partial x_2^2 \) are not bounded in Braverman (2020), so we must bound these partial derivatives ourselves.

We write the equation for \( \partial f^{(2)}(x)/\partial x_2 \) in (77) below, but writing it requires us to introduce some nontrivial objects from Braverman (2020). The first object we need is the family of curves \( \{ \Gamma^{(\kappa)} \subset \Omega \}_{\kappa \geq \beta} \), where \( \Gamma^{(\kappa)} \) is the graph of the unique fluid-limit trajectory that intersects the \( x_2 \) axis at the point \((0, \kappa/\sqrt{n})\). For the purposes of this proof, it suffices to treat \( \Gamma^{(\kappa)} \) as a two-dimensional geometric object satisfying the following properties:

1. \( \Gamma^{(\kappa)} \) is a graph of a continuous function; i.e. \( \Gamma^{(\kappa)} = \{(x_1, f(x_1))\} \) for some continuous function \( f: \mathbb{R}_+ \rightarrow \mathbb{R}_+ \).
2. \( \Gamma^{(\kappa)} \cap \{x \in \Omega : x_1 = 0\} = \{(0, \kappa/\sqrt{n})\} \).
3. If \( x \in \Gamma^{(\kappa)} \) and \( x_2 < 0 \), then \( x_2 > \kappa/\sqrt{n} \).
4. If \( \kappa' > \kappa \), then \( \Gamma^{(\kappa)} \cap \Gamma^{(\kappa')} = \emptyset \) and \( \Gamma^{(\kappa')} \) lies above \( \Gamma^{(\kappa)} \).

The first three properties are implied by Lemma 5 of Braverman (2020), and the fourth one follows from (39) there. Since \( \Gamma^{(\kappa)} \) is a graph, sets of the form \( \{x < \Gamma^{(\kappa)}\} \), \( \{x \leq \Gamma^{(\kappa)}\} \), etc., are well defined. Let us use \( \Gamma^{(\kappa_1)} \) and \( \Gamma^{(\kappa_2)} \) to partition \( \Omega \) into the four sets
\[
S_0 = \{x \in \Omega : x_2 \leq \kappa_1/\sqrt{n}\}, \quad S_1 = \{x \in \Omega : x_2 \geq \kappa_1/\sqrt{n}, x \leq \Gamma^{(\kappa_1)}\}, \\
S_2 = \{x \in \Omega : \Gamma^{(\kappa_1)} \leq x \leq \Gamma^{(\kappa_2)}\}, \quad S_3 = \{x \in \Omega : x \geq \Gamma^{(\kappa_2)}\}.
\]
The four properties of \( \Gamma^{(\kappa)} \) are sufficient to argue that \( S_0 \cup S_1 \cup S_2 \cup S_3 = \Omega \) and that the interiors of \( S_i \) and \( S_j \) are disjoint when \( i \neq j \); we refer the reader to Section C.2 of Braverman (2020) for more details.

The last object we need is the function \( \tau(x) \), which represents the first time that the fluid limit hits the \( x_2 \) axis starting from a state \( x > \Gamma^{(\beta)} \). The precise definition of \( \tau(x) \) is bulky and involves the Lambert-W function, but we can get by with only a few of its properties. Namely, for any \( \kappa > \beta \), Lemma 6 of Braverman (2020) introduces a nonnegative function \( \tau : \{x \in \Omega : x \geq \Gamma^{(\kappa)}\} \rightarrow \mathbb{R}_+ \) with \( \tau(0, x_2) = 0 \), which is differentiable for all \( x \in \{x \in \Omega : x \geq \Gamma^{(\kappa)}\} \) and satisfies
\[
\frac{\partial}{\partial x_1} \tau(x) = -\frac{e^{-\tau(x)}}{x_2 e^{-\tau(x)} - \beta/\sqrt{n}} \leq 0, \quad \frac{\partial}{\partial x_2} \tau(x) = \tau(x) \frac{\partial}{\partial x_1} \tau(x) \leq 0, \quad x \in \{x \in \Omega : x \geq \Gamma^{(\kappa)}\}. \quad (75)
\]
By choosing \( \kappa = \kappa_1 = 17/\beta + \beta \), we are assured that \( \tau(x) \) is defined on the set \( \{ x \in \Omega : x \geq \Gamma^{(\kappa_1)} \} = S_2 \cup S_3 \). Item 1 of Lemma 6 in Braverman (2020) tells us that \( \tau(x) \) is tied to \( \Gamma^{(\kappa)} \) for any \( \kappa > \beta \) via

\[
x_2 e^{-\tau(x)} \geq \kappa / \sqrt{n}, \quad x \geq \Gamma^{(\kappa)}.
\] (76)

We are now ready to bound the derivatives of \( f^{(2)}(x) \). Equation (C.9) of Braverman (2020) tells us that

\[
\frac{\partial}{\partial x_2} f^{(2)}(x) = \begin{cases} 
0, & x \in S_0, \\
\frac{1}{x_2^2} \phi(x_2) - \phi(x_2 e^{-\tau(x)}) + \phi(x_2 e^{-\tau(x)}) \frac{4 \sqrt{n}}{\beta} e^{-\tau(x)} (\tau(x) + 1), & x \in S_1, \\
\frac{2}{x_2^2} e^{-\tau(x)} (\tau(x) + 1), & x \in S_2, \\
\phi(x) = \phi^{(\kappa_1, \kappa_2)}(x) \) is the smoothed indicator defined in (68). By differentiating both sides of (68), it is straightforward to check that \( \phi(x) \) is non-decreasing, and

\[
\phi'(x) \leq \frac{4}{\delta(\kappa_2 - \kappa_1)} = \frac{4 \sqrt{n}}{17/\beta + \beta}.
\] (78)

Let us now argue that \( \partial f^{(2)}(x)/\partial x_2 \leq \sqrt{n}/\beta \) for any \( x \in \Omega \). If \( x \in S_3 \), this bound is implied by the inequality \( e^{-t(t+1)} \leq 1 \) for \( t \geq 0 \). If \( x \in S_1 \), the bound is implied by the facts that \( \phi(x_2) \leq 1 \) and \( 1/x_2 \leq \sqrt{n}/\kappa_1 \leq \sqrt{n}/\beta \). If \( x \in S_2 \), we note that \( \phi(x_2) - \phi(x_2 e^{-\tau(x)}) \geq 0 \), and \( 1/x_2 \leq \sqrt{n}/\beta \), meaning that

\[
\frac{\partial}{\partial x_2} f^{(2)}(x) = \frac{1}{x_2^2} \phi(x_2) - \phi(x_2 e^{-\tau(x)}) + \phi(x_2 e^{-\tau(x)}) \frac{4 \sqrt{n}}{\beta} e^{-\tau(x)} (\tau(x) + 1)
\leq \frac{\sqrt{n}}{\beta} (\phi(x_2) - \phi(x_2 e^{-\tau(x)}) + \phi(x_2 e^{-\tau(x)}) \frac{2 \sqrt{n}}{\beta} = \sqrt{n}/\beta.
\]

Observe that \( \partial f^{(2)}(x)/\partial x_2 = \sqrt{n}/\beta \) when \( \tau(x) = 0 \), which is true for any \( x \in S_2 \cup S_3 \) with \( x_1 = 0 \), implying the claim about \( \partial V(x)/\partial x_2 \) in (72). To conclude the proof, it remains to show \( \partial^2 V(x)/\partial x_2^2 \leq 9/17 \) by differentiating both sides in (77). Note that \( \partial^2 f^{(2)}(x)/\partial x_2^2 = 0 \) for \( x \in S_0 \).

When \( x \in S_1 \), we use the bound on \( \phi'(x) \) in (78), as well as the fact that \( 1/x_2 \leq \sqrt{n}/\beta \), to see that

\[
\frac{\partial^2}{\partial x_2^2} f^{(2)}(x) = -\frac{\sqrt{n}}{\beta} e^{-\tau(x)} (\tau(x) + 1) \frac{\partial}{\partial x_2} \tau(x) + \frac{\sqrt{n}}{\beta} e^{-\tau(x)} \frac{\partial}{\partial x_2} \tau(x) = -\frac{\sqrt{n}}{\beta} e^{-\tau(x)} \tau(x) \frac{\partial}{\partial x_2} \tau(x).
\]

Using the expression for \( \partial \tau(x)/\partial x_2 \) in (75), we see that

\[
\frac{\partial^2}{\partial x_2^2} f^{(2)}(x) = \frac{e^{-\tau(x)}}{x_2 e^{-\tau(x)} - \beta / \sqrt{n}} \frac{\sqrt{n}}{\beta} e^{-\tau(x)} \leq \frac{n}{7\beta(\kappa_2 - \beta)} \leq \frac{n}{7\beta(34/\beta + \beta)} \leq \frac{4n}{17}, \quad x \in S_3.
\] (79)
The first inequality follows from \( x_2 e^{-\tau(x)} \geq \kappa_2 / \sqrt{n} \) due to (76) and the fact that \( t^2 e^{-2t} \leq 1/7 \) for \( t \geq 0 \). Lastly, we consider the case when \( x \in S_2 \), for which we recall that

\[
\frac{\partial}{\partial x_2} f^{(2)}(x) = \frac{1}{x_2} \left( \phi(x_2) - \phi(x_2 e^{-\tau(x)}) \right) + \phi(x_2 e^{-\tau(x)}) \frac{\sqrt{n}}{\beta} e^{-\tau(x)} (\tau(x) + 1), \quad x \in S_2. \tag{80}
\]

To help organize terms, let \( g(x_2) = x_2 e^{-\tau(x)} \) and note from (75) that

\[
g'(x_2) = e^{-\tau(x)} \left( 1 - x_2 \frac{\partial}{\partial x_2} \tau(x) \right) = e^{-\tau(x)} \left( 1 + \frac{x_2 e^{-\tau(x)}}{x_2 e^{-\tau(x)} - \beta / \sqrt{n}} \tau(x) \right) = e^{-\tau(x)} \left( 1 + \tau(x) + \frac{\beta / \sqrt{n}}{x_2 e^{-\tau(x)} - \beta / \sqrt{n}} \right).
\]

We see that \( g'(x_2) \geq 0 \) because \( \tau(x) \geq 0 \) and \( x_2 e^{-\tau(x)} \geq \kappa_1 / \sqrt{n} \) for \( x \in S_2 \) due to (76). Furthermore, since \( e^{-t} \leq 1 \) and \( e^{-t} (t + 1) \leq 1 \) for \( t \geq 0 \), we conclude that

\[
0 \leq g'(x_2) \leq 1 + \frac{\beta / \sqrt{n}}{x_2 e^{-\tau(x)} - \beta / \sqrt{n}} \leq 1 + \frac{\beta}{\kappa_1 - \beta} = 1 + \frac{\beta^2}{17}. \tag{81}
\]

Let us now differentiate and bound each term on the right-hand side of (80) individually. First,

\[
\frac{\partial}{\partial x_2} \left( \frac{1}{x_2} \left( \phi(x_2) - \phi(x_2 e^{-\tau(x)}) \right) \right) = - \frac{1}{x_2} \left( \phi(x_2) - \phi(x_2 e^{-\tau(x)}) \right) + \frac{1}{x_2} \left( \phi'(x_2) - g'(x_2) \phi'(x_2 e^{-\tau(x)}) \right) \leq \frac{1}{x_2} \phi'(x_2) \leq \frac{\sqrt{n}}{\beta} \frac{4 \sqrt{n}}{17 / \beta + \beta} = 4n / 17.
\]

The first inequality is due to the facts that \( \phi(x) \) is non-decreasing and \( g'(x_2) \geq 0 \), and the second inequality follows from the fact that \( 1/x_2 \leq \sqrt{n}/\beta \) and the bound on \( \phi'(x) \) in (78). Differentiating the second term in (80), we get

\[
\frac{\partial}{\partial x_2} \left( \phi(x_2 e^{-\tau(x)}) \frac{\sqrt{n}}{\beta} e^{-\tau(x)} (\tau(x) + 1) \right) = \phi(x_2 e^{-\tau(x)}) \frac{\sqrt{n}}{\beta} e^{-\tau(x)} (\tau(x) + 1) + \phi'(x_2 e^{-\tau(x)}) \frac{\sqrt{n}}{\beta} e^{-\tau(x)} (\tau(x) + 1).
\]

To bound the first term, we use the fact that \( \phi(x) \leq 1 \), and we repeat the argument used to prove (79) to see that

\[
\phi(x_2 e^{-\tau(x)}) \frac{\sqrt{n}}{\beta} e^{-\tau(x)} (\tau(x) + 1) \leq \frac{n}{7 \beta (\kappa_1 - \beta)} = \frac{n}{119}.
\]

Furthermore, the bounds on \( \phi'(x) \) and \( g'(x_2) \) in (78) and (81), together with the fact that \( e^{-t} (t + 1) \leq 1 \), imply that

\[
\phi'(x_2 e^{-\tau(x)}) g'(x_2) \frac{\sqrt{n}}{\beta} e^{-\tau(x)} (\tau(x) + 1) \leq \frac{4 \sqrt{n}}{17 / \beta + \beta} \frac{1 + \beta^2}{17} = \frac{4 \sqrt{n} \beta}{17 + \beta^2} \frac{17}{\beta} = 4n / 17.
\]

Combining the pieces yields \( \partial^2 f^{(2)}(x) / \partial x_2^2 \leq 9n / 17 \), proving (72).
To conclude, we prove that $0 \leq V(x) \leq C(\beta)(1 + x_2)$ for $x_2 \geq \kappa_2$ by proving that $0 \leq f^{(2)}(x) \leq C(\beta)(1 + \sqrt{n}x_2)$ for $x_2 \geq \kappa_2/\sqrt{n}$. The form of $f^{(2)}(x)$ below can be found in Lemma 12 of Braverman (2020):

$$f^{(2)}(x) = \begin{cases} 0, & x \in S_0, \\ \int_0^{\log(\sqrt{n}x_2/\kappa_1)} \phi(x_2e^{-t})dt, & x_2 \leq \kappa_2/\sqrt{n}, \; x \in S_1, \\ \log(\sqrt{n}x_2/\kappa_2) + \int_0^{\log(\kappa_2/\kappa_1)} \phi(\frac{\sqrt{n}x_2}{\kappa_2}e^{-t})dt, & x_2 \geq \kappa_2/\sqrt{n}, \; x \in S_1, \\ \int_0^{\tau(x)} \phi(x_2e^{-t})dt + \frac{\sqrt{n}}{\beta} \int_{\kappa_1/\sqrt{n}}^{\kappa_2} \phi(t)dt, & x_2 \leq \kappa_2/\sqrt{n}, \; x \in S_2, \\ \log(\sqrt{n}x_2/\kappa_2) + \int_0^{\tau(x)} \phi(x_2e^{-t})dt + \frac{\sqrt{n}}{\beta} \int_{\kappa_1/\sqrt{n}}^{\kappa_2} \phi(t)dt, & x_2 \geq \kappa_2/\sqrt{n}, \; x \in S_2, \\ \tau(x) + \frac{x_2e^{-\tau(x)} - \kappa_2/\sqrt{n}}{\beta} + \frac{\sqrt{n}}{\beta} \int_{\kappa_1/\sqrt{n}}^{\kappa_2} \phi(t)dt, & x \in S_3. \end{cases}$$

The fact that $f^{(2)}(x) \geq 0$ follows from $\phi(x), \tau(x) \geq 0$, the definitions of $S_1, S_2,$ and $S_3$, and (76).

We combine all the cases above into the single upper bound

$$f^{(2)}(x) \leq \log(\sqrt{n}x_2/\kappa_1)1(x \in S_1 \cup S_2 \cup S_3) + \log(\kappa_2/\kappa_1) + \tau(x)1(x \in S_2 \cup S_3) + \frac{\sqrt{n}}{\beta} \left( x_2e^{-\tau(x)} - \kappa_1/\sqrt{n} \right)1(x \in S_2) + \frac{\sqrt{n}}{\beta} \left( x_2e^{-\tau(x)} - \kappa_2/\sqrt{n} \right)1(x \in S_3) + \frac{\kappa_2 - \kappa_1}{\beta}. \tag{82}$$

Using the inequality $\log(t) \leq 1 + t$ for $t \geq 0$, and the facts that $\kappa_1 = 17/\beta + \beta$ and $\kappa_2 = 2\kappa_1$, we see that $\log(\kappa_2/\kappa_1) = \log(2), (\kappa_2 - \kappa_1)/\beta = 1 + 17/\beta^2$,

$$\log(\sqrt{n}x_2/\kappa_1)1(x \in S_1 \cup S_2 \cup S_3) \leq 1 + \frac{\sqrt{n}x_2}{\kappa_1} \leq 1 + \frac{\sqrt{n}x_2}{\beta},$$

$$\frac{\sqrt{n}}{\beta} \left( x_2e^{-\tau(x)} - \kappa_1/\sqrt{n} \right)1(x \in S_2) \leq \frac{\sqrt{n}x_2}{\beta}, \quad \text{and} \quad \frac{\sqrt{n}}{\beta} \left( x_2e^{-\tau(x)} - \kappa_2/\sqrt{n} \right)1(x \in S_3) \leq \frac{\sqrt{n}x_2}{\beta}.$$

Furthermore, (76) and the definitions of $S_2$ and $S_3$ imply that

$$\tau(x)1(x \in S_2 \cup S_3) \leq \log(x_2/\sqrt{n}/\kappa_1) \leq 1 + \frac{\sqrt{n}x_2}{\kappa_1} = 1 + \frac{\sqrt{n}x_2}{\beta}. \tag{83}$$

We conclude by combining all of these bounds with (82).

B.3. Proof of Lemma 14

Assume without loss of generality that $Q_1(0) = n$ and $Q_2(0) = \theta_2$, because starting from $(q_1, \theta_2, q_3, \ldots, q_{k+1}) \in S_Q$, a state with $q_1 = n$ and $q_2 = \theta_2$ must be visited before $\tau_2(2\theta_2)$, so

$$\min_{q_2 = \theta_2, q_3 = n} \mathbb{P}_q(\tau_1(\theta_1) < \tau_2(2\theta_2)) \geq \min_{q_2 = \theta_2, q_3 = n} \mathbb{P}_q(\tau_1(\theta_1) < \tau_2(2\theta_2)).$$

We bound the right-hand side by relating it to the ruin probability in a certain gambler’s ruin problem. Namely, we construct a random walk $\{\mathcal{R}(t)\}$ with $\mathcal{R}(0) = 0$ that satisfies

$$\min_{q_2 = \theta_2, q_3 = n} \mathbb{P}_q(\tau_1(\theta_1) < \tau_2(2\theta_2)) \geq \mathbb{P} \left( \inf_{t \geq 0} \{\mathcal{R}(t) = |\gamma\sqrt{n}| \} > \inf_{t \geq 0} \{\mathcal{R}(t) = -|\sqrt{n}\beta/2| \} \right). \tag{83}$$
Jumps in the random walk are governed by a Poisson process with rate $n\lambda + \theta_1 - 3\theta_2$, and the up-step and down-step probabilities are

$$\frac{n\lambda}{n\lambda + \theta_1 - 3\theta_2}, \quad \text{and} \quad \frac{\theta_1 - 3\theta_2}{n\lambda + \theta_1 - 3\theta_2};$$

respectively. Note that we implicitly assume $n$ is large enough so that $\theta_1 - 3\theta_2 > 0$. The right-hand side in (83) is therefore the ruin probability in a gambler’s ruin problem with initial wealth $\lfloor \sqrt{n}\beta/2 \rfloor$ and opponent’s wealth $\sqrt{n}\beta/2 + \lfloor \gamma\sqrt{n} \rfloor$. A formula for the ruin probability was given by equation (2.4) in Section XIV.2 of Feller (1968):

$$P\left(\inf_{t \geq 0} (\bar{R}(t) = \lfloor \gamma\sqrt{n} \rfloor) > \inf_{t \geq 0} (\bar{R}(t) = -\lfloor \sqrt{n}\beta/2 \rfloor) \right) = 1 - \frac{1 - ((\theta_1 - 3\theta_2)/n\lambda)^{\lfloor \sqrt{n}\beta/2 \rfloor}}{1 - ((\theta_1 - 3\theta_2)/n\lambda)^{\lfloor \sqrt{n}\beta/2 \rfloor + \lfloor \gamma\sqrt{n} \rfloor}}.$$

Recalling the values of $\theta_1$ and $\theta_2$ and the fact that $\gamma > \beta$, we see that

$$\frac{\theta_1 - 3\theta_2}{n\lambda} = n - \lfloor \sqrt{n}\beta/2 \rfloor - 3\lfloor \gamma\sqrt{n} \rfloor = 1 - \frac{-\beta\sqrt{n} + \lfloor \sqrt{n}\beta/2 \rfloor + 3\lfloor \gamma\sqrt{n} \rfloor}{n\lambda} < 1,$n\lambda$$

and therefore,

$$\lim_{n \rightarrow \infty} \frac{1 - ((\theta_1 - 3\theta_2)/n\lambda)^{\lfloor \sqrt{n}\beta/2 \rfloor}}{1 - ((\theta_1 - 3\theta_2)/n\lambda)^{\lfloor \sqrt{n}\beta/2 \rfloor + \lfloor \gamma\sqrt{n} \rfloor}} = \lim_{n \rightarrow \infty} \frac{1 - (1 - \frac{-\beta\sqrt{n} + \lfloor \sqrt{n}\beta/2 \rfloor + 3\lfloor \gamma\sqrt{n} \rfloor}{n\lambda})^{\lfloor \sqrt{n}\beta/2 \rfloor}}{1 - (1 - \frac{-\beta\sqrt{n} + \lfloor \sqrt{n}\beta/2 \rfloor + 3\lfloor \gamma\sqrt{n} \rfloor}{n\lambda})^{\lfloor \sqrt{n}\beta/2 \rfloor + \lfloor \gamma\sqrt{n} \rfloor}} < 1,$$

implying Lemma 14. It remains to construct $\{\bar{R}(t)\}$.

Recall that $Q_1(0) = n$ and $Q_2(0) = \theta_2$, and let $\{\hat{Q}(t)\}$ be a copy of $\{Q(t)\}$, but with the modification that any server with a nonempty buffer permanently halts all its their work. Then $\hat{Q}_i(t) \geq Q_i(t)$ for all $t \geq 0$ and all $1 \leq i \leq b + 1$, because this modified system has the same arrival stream as $\{Q(t)\}$ but serves fewer customers. It follows that

$$\tau_1(\theta_1) = \inf_{t \geq 0} \{Q_1(t) = \theta_1\} \leq \inf_{t \geq 0} \{\hat{Q}_1(t) = \theta_1\},$$

$$\tau_2(2\theta_2) = \inf_{t \geq 0} \{Q_2(t) = 2\theta_2\} \geq \inf_{t \geq 0} \{\hat{Q}_2(t) = 2\theta_2\}, \quad \text{and}$$

$$\min_{q_2 = \theta_2, \ q_1 = n} P_q(\tau_1(\theta_1) < \tau_2(2\theta_2)) \geq \min_{q_2 = \theta_2, \ q_1 = n} P_q\left(\inf_{t \geq 0} \{\hat{Q}_2(t) = 2\theta_2\} > \inf_{t \geq 0} \{\hat{Q}_1(t) = \theta_1\}\right).$$

Now consider the process

$$R(t) = \hat{Q}_1(t) + \hat{Q}_2(t) - \hat{Q}_1(0) - \hat{Q}_2(0) = \hat{Q}_1(t) + \hat{Q}_2(t) - (n + \theta_2).$$

Note that $R(t) \geq \hat{Q}_1(t) - \hat{Q}_1(0) = \hat{Q}_1(t) - n$ since $\hat{Q}_2(t)$ is non-decreasing in $t$, which implies that

$$\inf_{t \geq 0} \{R(t) = -\lfloor \sqrt{n}\beta/2 \rfloor\} \geq \inf_{t \geq 0} \{\hat{Q}_1(t) = n - \lfloor \sqrt{n}\beta/2 \rfloor\} = \inf_{t \geq 0} \{\hat{Q}_1(t) = \theta_1\}.$$
Note also that \( \inf_{t \geq 0} \{ \hat{Q}_2(t) = 2\theta_2 \} = \inf_{t \geq 0} \{ R(t) = \theta_2 \} \) because \( \hat{Q}_2(t) \) is non-decreasing in \( t \) and \( \hat{Q}_2(t) \) increases only when \( \hat{Q}_1(t) = n \). Hence,

\[
\min_{q_2 = \theta_2, q_1 = n} \mathbb{P}_q \left( \inf_{t \geq 0} \{ R(t) = \theta_2 \} > \inf_{t \geq 0} \{ R(t) = -\sqrt{n\beta / 2} \} \right) \\
\leq \min_{q_2 = \theta_2, q_1 = n} \mathbb{P}_q \left( \inf_{t \geq 0} \{ \hat{Q}_2(t) = 2\theta_2 \} > \inf_{t \geq 0} \{ \hat{Q}_1(t) = \theta_1 \} \right) \\
\leq \min_{q_2 = \theta_2, q_1 = n} \mathbb{P}_q (r_1(\theta_1) < r_2(2\theta_2)).
\]

An arrival to \( \{ \hat{Q}(t) \} \) increases the value of \( \{ R(t) \} \), and a service completion by a server with an empty buffer decreases its value. However, \( \{ R(t) \} \) is still not the random walk we desire because the rate at which it decreases depends on the state of \( \hat{Q}(t) \). Instead, we want a random walk with a constant downward rate.

To construct this random walk, for \( 0 \leq t \leq \inf_{t \geq 0} \{ \hat{Q}_2(t) = 2\theta_2 \} \) let us define \( \{ \overline{Q}(t) = (\overline{Q}_1(t), \overline{Q}_2(t)) \} \) by setting \( \overline{Q}(0) = \hat{Q}(0) \) and defining the transitions of the joint process \( \{ (\hat{Q}(t), \overline{Q}(t)) \} \) in Tables 1, 2, and 3 below. Since we are defining \( \overline{Q}(t) \) only until the time \( \hat{Q}_2(t) \) hits \( 2\theta_2 \), we do not need to specify the transitions for states where \( \hat{Q}_2(t) > 2\theta_2 \). The intuition for the transition structure is as follows. Since arrivals occur at the constant rate of \( n\lambda \), we want any arrival to \( \{ \hat{Q}(t) \} \) to also occur in \( \{ \overline{Q}(t) \} \). However, we want to keep the rate at which \( \{ \overline{Q}(t) \} \) decreases a constant value of \( \theta_1 - 3\theta_2 \). To accomplish this, when \( \hat{Q}_1(t) \geq \theta_1 - \theta_2 \), the transitions in Table 2 have \( \{ \overline{Q}(t) \} \) ignore some departures from \( \{ \hat{Q}(t) \} \), and when \( \hat{Q}_1(t) < \theta_1 - \theta_2 \), we supplement the departures from \( \{ \hat{Q}(t) \} \); e.g., see transition #8 in Table 3.

### Table 1. Arrival transitions for the joint process in state \( ((\hat{u}_1, \hat{u}_2), (\overline{u}_1, \overline{u}_2)) \).

| #   | Rate                      | Transition                                      |
|-----|---------------------------|-------------------------------------------------|
| 1   | \( n\lambda(\hat{u}_1 < n, \overline{u}_1 < n) \) | \( ((\hat{u}_1 + 1, \hat{u}_2), (\overline{u}_1 + 1, \overline{u}_2)) \) |
| 2   | \( n\lambda(\hat{u}_1 = n, \overline{u}_1 < n) \) | \( ((\hat{u}_1, \hat{u}_2 + 1), (\overline{u}_1 + 1, \overline{u}_2)) \) |
| 3   | \( n\lambda(\hat{u}_1 < n, \overline{u}_1 = n) \) | \( ((\hat{u}_1 + 1, \hat{u}_2), (\overline{u}_1, \overline{u}_2 + 1)) \) |
| 4   | \( n\lambda(\hat{u}_1 = n, \overline{u}_1 = n) \) | \( ((\hat{u}_1, \hat{u}_2 + 1), (\overline{u}_1, \overline{u}_2 + 1)) \) |

### Table 2. Departure transitions for the joint process in state \( ((\hat{u}_1, \hat{u}_2), (\overline{u}_1, \overline{u}_2)) \) with \( \hat{u}_2 \leq 2\theta_2 \) and \( \hat{u}_1 \geq \theta_1 - \theta_2 \).

| #   | Rate                      | Transition                                      |
|-----|---------------------------|-------------------------------------------------|
| 5   | \( \theta_1 - 3\theta_2 \) | \( ((\hat{u}_1 - 1, \hat{u}_2), (\overline{u}_1 - 1, \overline{u}_2)) \) |
| 6   | \( \hat{u}_1 - \hat{u}_2 - (\theta_1 - 3\theta_2) \) | \( ((\hat{u}_1 - 1, \hat{u}_2), (\overline{u}_1, \overline{u}_2)) \) |
show that any decrease in \( \theta_1 \) tells us that a decrease in \( \theta_2 \) must be accompanied by a decrease in \( \theta_1 \).

To see why (86) is true, let us study the transitions in Tables 1–3. Table 1 tells us that \( \text{\overline{Q}}(t) \) and \( R(t) \) increase at the same times. The transitions in Table 2 show that any decrease in \( \text{\overline{Q}}_1(t) \), and consequently \( \text{\overline{Q}}(t) \), must be accompanied by a decrease in \( \hat{Q}_1(t) \) and \( R(t) \), but not vice versa. The only way \( \text{\overline{Q}}_1(t) \) can ever drop below \( \hat{Q}_1(t) \) is via transition 8, which can happen only if \( \hat{Q}_1(t) < \theta_1 - \theta_2 \), so the first intersection of \( \text{\overline{Q}}_1(t) \) and \( \hat{Q}_1(t) \) has to occur below \( \theta_1 - \theta_2 \). Therefore, \( \text{\overline{U}}(t) \geq R(t) \) for all times

\[
t \leq \min \left\{ \inf_{t \geq 0} \{ \text{\overline{U}}_1(t) = \theta_1 - \theta_2 \}, \inf_{t \geq 0} \{ \hat{Q}_2(t) = 2\theta_2 \} \right\} = \min \left\{ \inf_{t \geq 0} \{ \text{\overline{U}}_1(t) = \theta_1 - \theta_2 \}, \inf_{t \geq 0} \{ R(t) = \theta_2 \} \right\}.
\]

Let us now prove (86) by showing that the right-hand side is greater than

\[
\min \left\{ \inf_{t \geq 0} \{ \text{\overline{U}}(t) = \theta_1 - \theta_2 \}, \inf_{t \geq 0} \{ R(t) = \theta_2 \} \right\}.
\]
Since $\mathcal{R}(t) \geq R(t)$,
\[
\min\left\{ \inf_{t \geq 0} \{Q_1(t) = \theta_1 - \theta_2\}, \inf_{t \geq 0} \{R(t) = \theta_2\} \right\} \geq \min\left\{ \inf_{t \geq 0} \{Q_1(t) = \theta_1 - \theta_2\}, \inf_{t \geq 0} \{\mathcal{R}(t) = \theta_2\} \right\}.
\]
Furthermore, since $Q_2(t)$ is non-decreasing and increases only at those times when $Q_1(t) = n$, it follows that for all $t \leq \inf_{t \geq 0} \{\mathcal{R}(t) = \theta_2\}$,
\[
\mathcal{R}(t) = Q_1(t) + Q_2(t) - n - \theta_2 \leq Q_1(t) - n + \theta_2,
\]
and therefore
\[
\min\left\{ \inf_{t \geq 0} \{Q_1(t) = \theta_1 - \theta_2\}, \inf_{t \geq 0} \{\mathcal{R}(t) = \theta_2\} \right\} = \min\left\{ \inf_{t \geq 0} \{Q_1(t) = n - \lfloor \sqrt{n\beta}/2 \rfloor - \theta_2\}, \inf_{t \geq 0} \{\mathcal{R}(t) = \lfloor \gamma \sqrt{n} \rfloor\} \right\} \geq \min\left\{ \inf_{t \geq 0} \{\mathcal{R}(t) = -\lfloor \sqrt{n\beta}/2 \rfloor\}, \inf_{t \geq 0} \{\mathcal{R}(t) = \lfloor \gamma \sqrt{n} \rfloor\} \right\}.
\]

\[\Box\]

**B.4. Proving Lemma 15**

Central to our argument is a result about the moment-generating function of the duration of a gambler’s ruin game. We now describe this result and then prove Lemma 15. Consider a discrete-time gambler's ruin problem where the initial player's wealth is $z$, the win probability is $p$, the loss probability is $q$, and the player keeps playing until they go broke or accumulate a total wealth of $a$. Let $D_z \in \mathbb{Z}_+$ be the number of turns until the game ends, given an initial wealth of $z$. An expression for the generating function $\mathbb{E}_s^{D_z}$ was given in (4.11) and (4.12) in Section XIV.4 of Feller (1968):
\[
\mathbb{E}_s^{D_z} = \frac{\lambda_1(s)\lambda_2^2(s) - \lambda_1^2(s)\lambda_2(s)}{\lambda_1(s) - \lambda_2(s)} + \frac{\lambda_1^2(s) - \lambda_2^2(s)}{\lambda_1(s) - \lambda_2(s)}, \quad s \in (0, 1),
\]
(87)
where
\[
\lambda_1(s) = \frac{1 + \sqrt{1 - 4pq^2s^2}}{2ps}, \quad \text{and} \quad \lambda_2(s) = \frac{1 - \sqrt{1 - 4pq^2s^2}}{2ps}, \quad s \in (0, 1).
\]

Now consider the continuous-time gambler’s ruin problem, where the durations between turns are governed by an i.i.d. sequence $\{E_i\}$ of rate $r$ exponentially distributed random variables. Given initial wealth $z$, the duration of the continuous game equals $\sum_{i=1}^{D_z} E_i$. Since the $E_i$ are independent of $D_z$, it follows that
\[
\mathbb{E}e^{-\sum_{i=1}^{D_z} E_i} = \mathbb{E}(\mathbb{E}e^{-E_1})^{D_z} = \mathbb{E}\left(\frac{r}{r + 1}\right)^{D_z},
\]
so $\mathbb{E}e^{-\sum_{i=1}^{D_z} E_i}$ is related to (87). The following result proved in Appendix B.4.1 is needed to prove Lemma 15.
Let $i$ and $q_2$ be integers such that $1 \leq i \leq b + 1$ and $0 \leq q_2 \leq \theta_2$, and define

$$q^{(B,i)} = n - q_2 - 1 - \lfloor \sqrt{n\beta}/2 \rfloor + \lfloor \sqrt{n\beta}/2 \rfloor (b + 1).$$

Consider the continuous-time gambler’s ruin problem with probabilities

$$p = \frac{n\lambda}{n\lambda + q^{(B,i)} - \lfloor \sqrt{n\beta}/2 \rfloor}, \quad \text{and} \quad q = \frac{q^{(B,i)} - \lfloor \sqrt{n\beta}/2 \rfloor}{n\lambda + q^{(B,i)} - \lfloor \sqrt{n\beta}/2 \rfloor},$$

rate $r = n\lambda + q^{(B,i)} - \lfloor \sqrt{n\beta}/2 \rfloor$, initial wealth $z$ and terminal wealth $a$ given by

$$z = \lfloor \sqrt{n\beta}/2 \rfloor, \quad \text{and} \quad a = \lfloor \sqrt{n\beta}/2 \rfloor + \lfloor \sqrt{n\beta}/2 \rfloor (b + 1),$$

and game duration $\sum_i D_i E_i$. Then

$$\lim_{n \to \infty} \max_{0 \leq q_2 \leq 2n^{1/2}} \mathbb{E} e^{-\sum_i D_i E_i} < 1. \quad (89)$$

**Proof of Lemma 15** As discussed below (65), $\{\tau_C < \tau_1(n)\} \supset \{\Gamma_{b+1} < \tau_1(n)\}$, where $\Gamma_{b+1}$ is the sum of $b + 1$ unit-mean exponentially distributed random variables. The same discussion says that $\Gamma_{b+1}$ represents the time needed by the joint CTMC $(Q(t), \bar{Q}(t))$ to transition from $\Theta_{b+1}^2$ to $\Theta_1^2$, and to then couple by spending an exponentially distributed amount of time in $\Theta_1^2$. Thus,

$$\min_{0 \leq q_2 \leq 2n^{1/2}} \mathbb{P}\left(\tau_C < \tau_1(n) \mid Q(0) = q, (Q(0), \bar{Q}(0)) \in \bigcup_{i=1}^{b+1} \Theta_i^2\right) \geq \min_{0 \leq q_2 \leq 2n^{1/2}} \mathbb{P}\left(\Gamma_{b+1} < \tau_1(n) \mid Q(0) = q\right).$$

Let us analyze the probability above. At time $t = 0$, there are $q_2$ servers with nonempty buffers and another server containing the extra customer in $\{\bar{Q}(t)\}$. We group these $q_2 + 1$ servers together into group $A$, and the remaining $n - q_2 - 1$ servers into group $B$. Let $Q_1^A(t)$ and $Q_1^B(t)$ be the number of busy group $A$ and $B$ servers, respectively. Since

$$Q_1^A(0) = q_2 + 1, \quad \text{and} \quad Q_1^A(0) + Q_1^B(0) = Q_1(0) \leq n - \lfloor \sqrt{n\beta}/2 \rfloor,$$

it follows that $Q_1^B(0) \leq n - q_2 - 1 - \lfloor \sqrt{n\beta}/2 \rfloor$. We are implicitly assuming that $n$ is large enough so $n - q_2 - 1 - \lfloor \sqrt{n\beta}/2 \rfloor \geq 0$. Note that the buffer of any group $B$ server is empty for all $t \leq \tau_1(n)$.

If a customer arrives when more than one server is idle, we prioritize assigning this customer to servers in group $B$ over group $A$. Note that this tie-breaking rule is consistent with the tie-breaking rule we imposed in the proof of Lemma 4. Let $\tau_B = \inf_{t \geq 0} \{Q_1^B(t) = n - q_2 - 1\}$ be the first time that all servers in group $B$ are busy. By construction, $\tau_B \leq \tau_1(n)$, so

$$\min_{0 \leq q_2 \leq 2n^{1/2}} \mathbb{P}\left(\Gamma_{b+1} < \tau_1(n) \mid Q(0) = q\right) \geq \min_{0 \leq q_2 \leq 2n^{1/2}} \mathbb{P}\left(\Gamma_{b+1} < \tau_B \mid Q_1^B(0) = q\right) \geq \min_{0 \leq q_2 \leq 2n^{1/2}} \mathbb{P}\left(\Gamma_{b+1} < \tau_B \mid Q_1^B(0) = n - q_2 - 1 - \lfloor \sqrt{n\beta}/2 \rfloor\right).$$
The last inequality is true because increasing the value of the initial condition \( Q_1^{(B)}(0) \) does not increase the chance that \( \Gamma_{b+1} < \tau_B \). We now relate the right-hand side to the moment-generating function considered in Lemma 18 and use that lemma to conclude the proof. We can write \( \Gamma_{b+1} = \sum_{i=1}^{b+1} G_i \), where \( G_i \) are i.i.d. unit-mean exponentially distributed random variables independent of \( Q_1^{(B)}(t) \) for \( t \in [0, \tau_B] \), because they correspond to service times of the server containing the additional customer in \( \{ \bar{Q}(t) \} \), which is a server in group A.

Fixing \( 0 \leq q_2 \leq 2\theta_2 \) and \( Q_1^{(B)}(0) = n - q_2 - 1 - \lfloor \sqrt{n} \beta/2 \rfloor \), for \( 0 \leq i \leq b + 1 \) we define

\[
q^{(B,i)} = n - q_2 - 1 - \lfloor \sqrt{n} \beta/2 \rfloor + \frac{i}{b + 1},
\]

and

\[
\tau_{B,i} = \inf_{t \geq 0} \{ Q_1^{(B)}(t) - Q_1^{(B)}(0) = \left\lfloor \frac{\sqrt{n} \beta/2}{b + 1} \right\rfloor \} = \inf_{t \geq 0} \left\{ Q_1^{(B)}(t) = q^{(B,i)} \right\},
\]

and note that \( \tau_B = \tau_{B,b+1} \). We are guaranteed that \( \Gamma_{b+1} < \tau_B \) if for each \( 1 \leq i \leq b + 1 \), the exponentially distributed \( G_i \) is smaller than the time it takes for \( Q_1^{(B)}(t) \) to reach \( q^{(B,i)} \) if started from \( q^{(B,i-1)} \), so

\[
\mathbb{P} \left( \Gamma_{b+1} < \tau_B \mid Q_1^{(B)}(0) = n - q_2 - 1 - \lfloor \sqrt{n} \beta/2 \rfloor \right) \geq \prod_{i=1}^{b+1} \mathbb{P} \left( G_i < \tau_{B,i} \mid Q_1^{(B)}(0) = q^{(B,i-1)} \right).
\]

We now show that \( \tau_{B,i} \) can be bounded from below by the duration of a gambler’s ruin game, which allows us to apply Lemma 18. Fix \( 1 \leq i \leq b + 1 \), and consider the time interval \( t \in [0, \tau_{B,i}] \), on which we construct the coupling \( \{ (Q_1^{(B)}(t), \bar{Q}_1^{(B)}(t)) \} \) by setting

\[
\bar{Q}_1^{(B,i)}(0) = Q_1^{(B)}(0) = q^{(B,i-1)}
\]

and defining the transitions of the joint process in Tables 4 and 5 below. We implicitly assume that \( n \) is large enough that \( q^{(B,i)} - \lfloor \sqrt{n} \beta/2 \rfloor > 0 \).

**Table 4** Transition rates in state \((u, \bar{u})\) with \( u \geq q^{(B,i-1)} - \lfloor \sqrt{n} \beta/2 \).

| Rate          | Transition                  |
|---------------|-----------------------------|
| \( n \lambda \) | \((u + 1, \bar{u} + 1)\)   |
| \( q^{(B,i-1)} - \lfloor \sqrt{n} \beta/2 \) | \((u - 1, \bar{u} - 1)\)   |
| \( u - (q^{(B,i-1)} - \lfloor \sqrt{n} \beta/2 \) | \((u - 1, \bar{u})\)       |

Note that the only time \( \bar{Q}_1^{(B,i)}(t) \) decreases but \( Q_1^{(B)}(t) \) does not is when the latter is smaller than \( q^{(B,i-1)} - \lfloor \sqrt{n} \beta/2 \), so we are guaranteed that

\[
\bar{Q}_1^{(B,i)}(t) \geq Q_1^{(B)}(t), \quad \text{for all } t \leq \min \left\{ \tau_{B,i}, \inf_{t \geq 0} \{ \bar{Q}_1^{(B,i)}(t) = q^{(B,i-1)} - \lfloor \sqrt{n} \beta/2 \} \right\}. \quad (90)
\]
Recalling the definitions of $\tau_{B,i}$ and $q^{(B,i)}$, we have

$$
\min \left\{ \tau_{B,i}, \inf_{t \geq 0} \{ Q_1^{(B,i)}(t) = q^{(B,i-1)} - \lfloor \sqrt{n} \beta / 2 \rfloor \} \right\} = \min \left\{ \inf_{t \geq 0} \{ Q_1^{(B,i)}(t) = q^{(B,i)} \}, \inf_{t \geq 0} \{ Q_1^{(B,i)}(t) = q^{(B,i-1)} - \lfloor \sqrt{n} \beta / 2 \rfloor \} \right\} = \min \left\{ \inf_{t \geq 0} \{ Q_1^{(B,i)}(t) = q^{(B,i)} + \lfloor \sqrt{n} \beta / 2 \rfloor /(b+1) \}, \inf_{t \geq 0} \{ Q_1^{(B,i)}(t) = q^{(B,i-1)} - \lfloor \sqrt{n} \beta / 2 \rfloor \} \right\} \geq \min \left\{ \inf_{t \geq 0} \{ Q_1^{(B,i)}(t) = q^{(B,i)} + \lfloor \sqrt{n} \beta / 2 \rfloor /(b+1) \}, \inf_{t \geq 0} \{ Q_1^{(B,i)}(t) = q^{(B,i-1)} - \lfloor \sqrt{n} \beta / 2 \rfloor \} \right\},
$$

where the last inequality follows from (90). Let $\tau_{B,i}$ equal the right-hand side and note that

$$
\tau_{B,i} = \inf_{t \geq 0} \{ (Q_1^{(B,i)}(t) - Q_1^{(B,i)}(0)) \in \{ - \lfloor \sqrt{n} \beta / 2 \rfloor, \lfloor \sqrt{n} \beta / 2 \rfloor /(b+1) \} \}
$$

because $Q_1^{(B,i)}(0) = q^{(B,i-1)}$. Since $\tau_{B,i} \leq \tau_{B,i}$, it follows that

$$
\min_{0 \leq q \leq 2 \lfloor \sqrt{n} \beta \rfloor} \prod_{i=1}^{b+1} \mathbb{P}(G_t < \tau_{B,i} \mid Q_1^{(B,i)}(0) = q^{(B,i-1)}) \geq \min_{0 \leq q \leq 2 \lfloor \sqrt{n} \beta \rfloor} \prod_{i=1}^{b+1} \mathbb{P}(G_t < \tau_{B,i} \mid Q_1^{(B,i)}(0) = q^{(B,i-1)}).
$$

Recall that $G_i$ corresponds to the service time of a group $A$ server and is therefore independent of $\tau_{B,i}$. Furthermore, since $G_i$ is exponentially distributed with unit mean, conditioning on the value of $\tau_{B,i}$ yields

$$
\min_{0 \leq q \leq 2 \lfloor \sqrt{n} \beta \rfloor} \mathbb{P}(G_t < \tau_{B,i} \mid Q_1^{(B,i)}(0) = q^{(B,i-1)}) = \min_{0 \leq q \leq 2 \lfloor \sqrt{n} \beta \rfloor} \left( 1 - \mathbb{E}(e^{-\tau_{B,i}} \mid Q_1^{(B,i)}(0) = q^{(B,i-1)}) \right) = 1 - \max_{0 \leq q \leq 2 \lfloor \sqrt{n} \beta \rfloor} \mathbb{E}(e^{-\tau_{B,i}} \mid Q_1^{(B,i)}(0) = q^{(B,i-1)}).
$$

Applying (89) of Lemma 18 concludes, because our construction of $\{Q_1^{(B,i)}(t)\}$ implies that $\tau_{B,i}$ is the duration of a gambler’s ruin game with initial wealth $z = \lfloor \sqrt{n} \beta / 2 \rfloor$, terminal wealth $a = \lfloor \sqrt{n} \beta / 2 \rfloor + \lfloor \sqrt{n} \beta / 2 \rfloor /(b+1)$, rate $n \lambda + q^{(B,i)} - \lfloor \sqrt{n} \beta / 2 \rfloor$, and up-step and down-step probabilities $n \lambda + q^{(B,i)} - \lfloor \sqrt{n} \beta / 2 \rfloor$ and $n \lambda + q^{(B,i)} - \lfloor \sqrt{n} \beta / 2 \rfloor$, respectively.
B.4.1. Proving the Gambler’s Ruin Result. We require the following auxiliary lemma.

**Lemma 19.** Assume \( \{x_n \in \mathbb{R} \} \) is a sequence that converges to \( \bar{x} \). Then

\[
\lim_{n \to \infty} \left(1 + \frac{x_n}{n}\right)^n \to e^{\bar{x}}.
\]

**Proof of Lemma 19** Let \( f(x) = e^x \) and \( f_n(x) = \left(1 + \frac{x}{n}\right)^n \), and note that for any \( n \geq 0 \),

\[
|f_n(x_n) - e^{\bar{x}}| \leq |f_n(x_n) - f_n(\bar{x})| + |f_n(\bar{x}) - e^{\bar{x}}|.
\]

From the mean-value theorem, we know that there exists some \( c_n \) between \( x_n \) and \( \bar{x} \) such that

\[
|f_n(x_n) - f_n(\bar{x})| \leq |x_n - \bar{x}| f'_n(c_n) = |x_n - \bar{x}| \left(1 + \frac{c_n}{n}\right)^{n-1}.
\]

Since \( x_n \to \bar{x} \), it follows that \( (1 + c_n/n)^{n-1} \leq (1 + 2|\bar{x}|/n)^{n-1} \) for \( n \) large enough, and therefore,

\[
|f_n(x_n) - e^{\bar{x}}| \leq |x_n - \bar{x}| \left(1 + \frac{2|\bar{x}|}{n}\right)^{n-1} + |f_n(\bar{x}) - e^{\bar{x}}|.
\]

We can make the right-hand side arbitrarily small by increasing \( n \). \( \square \)

**Proof of Lemma 18** Recall that \( \mathbb{E} e^{-\sum_{i=1}^{D_x} s_i} = \mathbb{E}(r/(r+1))^{D_x} \), and that

\[
\mathbb{E}s^{D_x} = \frac{\lambda_1^s(s)\lambda_2^s(s) - \lambda_1^s(s)\lambda_2^s(s)}{\lambda_1^s(s) - \lambda_2^s(s)} + \frac{\lambda_1^s(s) - \lambda_2^s(s)}{\lambda_1^s(s) - \lambda_2^s(s)} = \frac{\lambda_1^s(s)(\lambda_1^s(s) - 1) - \lambda_1^s(s)(\lambda_2^s(s) - 1)}{\lambda_1^s(s) - \lambda_2^s(s)},
\]

where

\[
\lambda_1(s) = \frac{1 + \sqrt{1 - 4ps^2}}{2ps} \quad \text{and} \quad \lambda_2(s) = \frac{1 - \sqrt{1 - 4ps^2}}{2ps}, \quad s \in (0, 1).
\]

Fix \( s = r/(r+1) \). To show that \( \lim_{n \to \infty} \mathbb{E}s^{D_x} < 1 \), we derive expressions for \( \lim_{n \to \infty} \lambda_1^s(s) \) and \( \lim_{n \to \infty} \lambda_2^s(s) \). For notational economy, we let \( \theta_3 = [\sqrt{n}\beta/2] \). We can write \( p \) and \( q \) as

\[
p = \frac{n\lambda}{n\lambda + q(B,i) - \theta_3} = \frac{1}{2} + \frac{1}{2} \frac{n\lambda - (q(B,i) - \theta_3)}{n\lambda + q(B,i) - \theta_3}, \quad q = \frac{1}{2} - \frac{1}{2} \frac{n\lambda - (q(B,i) - \theta_3)}{n\lambda + q(B,i) - \theta_3},
\]

and

\[
pq = \frac{1}{4} - \frac{1}{4} \left( \frac{n\lambda - (q(B,i) - \theta_3)}{n\lambda + q(B,i) - \theta_3} \right)^2.
\]

Let us first consider \( \lambda_1(s) \), which satisfies

\[
\lambda_1(s) = \left(1 + \sqrt{1 - s^2 + \left( \frac{n\lambda - (q(B,i) - \theta_3)}{n\lambda + q(B,i) - \theta_3} \right)^2 s^2} \right)^{-1} \left(1 + \frac{n\lambda - (q(B,i) - \theta_3)}{n\lambda + q(B,i) - \theta_3} \right)^{-1}
\]

\[
= \left(1 + \frac{1}{\sqrt{n}} \left[ \sqrt{1 - s^2} + n \left( \frac{n\lambda - (q(B,i) - \theta_3)}{n\lambda + q(B,i) - \theta_3} \right)^2 s^2 \right] \right)^{-1} \left(1 + \frac{1}{\sqrt{n}} \left[ \sqrt{1 - s^2} + n \left( \frac{n\lambda - (q(B,i) - \theta_3)}{n\lambda + q(B,i) - \theta_3} \right)^2 s^2 \right] \right)^{-1}.
\]

(91)
We now show that the terms inside the square brackets have limits \( \overline{x}, \overline{y} \in \mathbb{R} \) as \( n \to \infty \); i.e.,
\[
\lim_{n \to \infty} \sqrt{n(1-s^2) + n\left(\frac{n\lambda - (q(B,i) - \theta_3)}{n\lambda + q(B,i) - \theta_3}\right)^2} = \overline{x} \quad \text{and} \quad \lim_{n \to \infty} \sqrt{n\frac{n\lambda - (q(B,i) - \theta_3)}{n\lambda + q(B,i) - \theta_3}} = \overline{y}. \tag{92}
\]
Note that \( \lim_{n \to \infty} s^2 = 1 \), and recall the definition of \( r \) to see that \( \lim_{n \to \infty} r/n = 1 + \lambda \), so
\[
\lim_{n \to \infty} (1-s^2) = \lim_{n \to \infty} \frac{n(2r+1)}{1+2r+r^2} = \lim_{n \to \infty} \frac{(2r+1)/n}{(1+2r+r^2)/n^2} = \lim_{n \to \infty} \frac{2}{1+\lambda}.\]
Furthermore, recalling the definition of \( q(B,i) \), we have
\[
\lim_{n \to \infty} \left(\frac{n\lambda - (q(B,i) - \theta_3)}{n\lambda + q(B,i) - \theta_3}\right)^2 = \lim_{n \to \infty} \left(\frac{-\beta\sqrt{n} + q_2 + 1 + 2\left[\sqrt{n}\beta/2\right] - \left[\sqrt{n}\beta/2\right] + \left[\sqrt{n}\beta/2\right] + \left[\sqrt{n}\beta/2\right] + \left[\sqrt{n}\beta/2\right] + \left[\sqrt{n}\beta/2\right]}{\lambda + n - q_2 - 1 - 2\left[\sqrt{n}\beta/2\right] + \left[\sqrt{n}\beta/2\right] + \left[\sqrt{n}\beta/2\right] + \left[\sqrt{n}\beta/2\right] + \left[\sqrt{n}\beta/2\right] + \left[\sqrt{n}\beta/2\right]}\right)^2.
\]
We know that \( \lim_{n \to \infty} q_2/\sqrt{n} \) exists because \( q_2 \) is fixed between zero and \( 2[\gamma\sqrt{n}] \). This proves (92).
Recall that \( z = \left[\sqrt{n}\beta/2\right] \) and \( a = \left[\sqrt{n}\beta/2\right] + \left[\sqrt{n}\beta/2\right] + 1/\gamma^{1/2} \). Since \( r/n \to 1 + \lambda \), it follows that
\[
\lim_{n \to \infty} s^a = \lim_{n \to \infty} \left(1 - 1/(r+1)\right)^{\left[\sqrt{n}\beta/2\right] + \left[\sqrt{n}\beta/2\right] + 1} = 1 \quad \text{and} \quad \lim_{n \to \infty} s^a = \lim_{n \to \infty} s^{\left[\sqrt{n}\beta/2\right]} = 1,
\]
and combined with (91), (92), and Lemma 19, this implies that
\[
\lim_{n \to \infty} \lambda_1(s) = \lim_{n \to \infty} \lambda_1^{\left[\sqrt{n}\beta/2\right]}(s) = \exp\left(\frac{\overline{x}\beta}{2}\right) \exp\left(-\frac{\overline{y}\beta}{2}\right),
\]
\[
\lim_{n \to \infty} \lambda_2(s) = \lim_{n \to \infty} \lambda_1^{\left[\sqrt{n}\beta/2\right] + \left[\sqrt{n}\beta/2\right] + 1}(s) = \exp\left(\frac{\overline{x}\beta + 2}{2}\right) \exp\left(-\frac{\overline{y}\beta + 2}{2}\right).
\]
The expressions for \( \lim_{n \to \infty} \lambda_2^a(s) \) and \( \lim_{n \to \infty} \lambda_2^a(s) \) follow similarly. Comparing
\[
\lambda_2(s) = \left(1 - \sqrt{1-s^2} + \frac{n\lambda - (q(B,i) - \theta_3)}{n\lambda + q(B,i) - \theta_3}\right)^2 s^{-1} \left(n\lambda - (q(B,i) - \theta_3)\right)^{-1}
\]
to the form of \( \lambda_1(s) \) in (91), we see that we can use (92) and Lemma 19 again to conclude that
\[
\lim_{n \to \infty} \lambda_2(s) = \exp\left(-\frac{\overline{x}\beta}{2}\right) \exp\left(-\frac{\overline{y}\beta}{2}\right), \quad \text{and} \quad \lim_{n \to \infty} \lambda_2(s) = \exp\left(-\frac{\overline{x}\beta + 2}{2}\right) \exp\left(-\frac{\overline{y}\beta + 2}{2}\right).
\]
For convenience, we define \( x = (\overline{x} - \overline{y})\beta/2 \) and \( y = (\overline{x} + \overline{y})\beta/2 \), so that
\[
\lim_{n \to \infty} \lambda_1(s) = e^x, \quad \lim_{n \to \infty} \lambda_2(s) = e^{x+2}/(b+1), \quad \lim_{n \to \infty} \lambda_2(s) = e^{-y} = e^{-(b+2)/(b+1)}, \quad \text{and} \quad \lim_{n \to \infty} \lambda_1(s) = e^{-y}/(b+1).
\]
It is straightforward to check that \( x, y > 0 \) using (92). Let us now prove that \( \lim_{n \to \infty} \mathbb{E} s^{D^x} < 1 \).
Using the definition of \( \mathbb{E} s^{D^x} \), we have
\[
\lim_{n \to \infty} \mathbb{E} s^{D^x} = \lim_{n \to \infty} \frac{\lambda_2^x(s)(\lambda_2^y(s) - 1) - \lambda_2^y(s)(\lambda_2^x(s) - 1)}{\lambda_2^x(s) - \lambda_2^y(s)} = \frac{e^y - e^{-y}}{e^x - e^{-y}}.
\]
Set \( c = (b + 2)/(b + 1) \). We want to show that for any \( x, y > 0 \),
\[
e^{-y}(e^{xc} - 1) - e^{x}(e^{-yc} - 1) < e^{xc} - e^{-yc}, \quad \text{or} \quad e^{-y}e^{xc} - e^{x}e^{-yc} < e^{xc} - e^{-yc} + e^{-y} - e^{x}.
\]
Rearranging terms, this is equivalent to
\[
e^{xc}(e^{-y} - 1) - e^{x}(e^{-yc} - 1) < -e^{-yc} + e^{-y}.
\]
Fix \( y > 0 \) and treat the left-hand side as a function of \( x \). Both sides are equal when \( x = 0 \), so it suffices to show that the derivative of the left-hand side with respect to \( x \) is negative. Now
\[
\frac{\partial}{\partial x} \left( e^{xc}(e^{-y} - 1) - e^{x}(e^{-yc} - 1) \right) = cce^{xc}(e^{-y} - 1) - e^{x}(e^{-yc} - 1). \tag{94}
\]
For the right-hand side to be negative, we must have
\[
cc e^{xc} > \frac{1 - e^{-yc}}{1 - e^{-y}}.
\]
Since \( c = (b + 2)/(b + 1) > 1 \), the left-hand side is bounded from below by \( c \) provided that \( x \geq 0 \). The right-hand side converges to \( c \) as \( y \downarrow 0 \), so we must show that the derivative of the right-hand side is negative. Differentiating yields
\[
\frac{\partial}{\partial y} \left( 1 - e^{-yc} \right) = cc e^{yc} - e^{-y}e^{-yc} - 1 + e^{-yc}.
\]
The numerator \( cc e^{yc} - (c - 1)e^{-yc} - 1 \) equals 0 when \( y = 0 \). Its derivative equals
\[
-c(c - 1)e^{-yc} + (c - 1)^2 e^{-yc} < -c(c - 1)e^{-yc} + (c - 1)^2 e^{-yc} \frac{c}{c - 1} = 0, \quad y \geq 0,
\]
where the inequality is due to \( e^{-y} \leq 1 < c/(c - 1) \). Therefore, the numerator is strictly negative for \( y > 0 \), meaning that (94) holds. \( \square \)

### B.5. Proof of Lemma 8

It suffices to show that \( \mathbb{E} \tau^{-}(x_{i}^{\uparrow}) \leq C(\beta)\delta \), because
\[
\mathbb{P}(V \leq \tau^{-}(x_{i}^{\uparrow})) = \int_{0}^{\infty} \mathbb{P}(V \leq t) dF(t) = \int_{0}^{\infty} (1 - e^{-t}) dF(t) = 1 - \mathbb{E} e^{-\tau^{-}(x_{i}^{\uparrow})} \leq \mathbb{E} \tau^{-}(x_{i}^{\uparrow}),
\]
where \( F(t) \) is the distribution function of \( \tau^{-}(x_{i}^{\uparrow}) \). Define
\[
\tau^{+}(q_{1}) = \inf_{t \geq 0} \{ Q(t) = (q_{1} + 1, 0, \ldots, 0) | Q(0) = (q_{1}, 0, \ldots, 0) \}, \quad 0 \leq q_{1} \leq n - 1,
\]
and note that \( \tau^{+}(q_{1}) = \tau^{-}(x_{i}^{\uparrow}) \). If we let \( \{ \pi_{q} \}_{q \in S_{Q}} \) be the stationary distribution of the unscaled CTMC, it follows from (2.11) of Brown and Xia (2001) that
\[
\mathbb{E} \tau^{+}(q_{1}) = \frac{\sum_{i=0}^{q_{1}} \pi_{i,0,\ldots,0}}{n \lambda \pi_{q_{1},0,\ldots,0}}.
\]
Letting $f(x^q) = 1(x^q_i \leq i)$ and using $E X f(X) = 0$ yields $n\lambda \pi_{i,0,...,0} = (i+1)\pi_{i+1,0,...,0}$, which implies that $\pi_{i,0,...,0} = \pi_{0,...,0}(n\lambda)^i/i!$, so

$$E r^+(q_1) = \frac{\sum_{k=0}^{q_1} (n\lambda)^k}{n\lambda^{(n\lambda)^n_{q_1}}!} = \frac{q_1!}{(n\lambda)^{q_1+1}} \sum_{k=0}^{q_1} \frac{(n\lambda)^k}{k!}. $$

Note that $x^q_i = \beta + \delta(n\lambda - [n\lambda])$ is equivalent to $q_1 = [n\lambda]$. If $[n\lambda] = 0$, we observe that the right-hand side equals $1/(n\lambda)$, which verifies (31) when $x^q_i = \beta + \delta(n\lambda - [n\lambda])$. If, however, $[n\lambda] > 0$, we may use Stirling's approximation to see that for $q_1 > 0$,

$$\frac{q_1!}{(n\lambda)^{q_1+1}} \sum_{k=0}^{q_1} \frac{(n\lambda)^k}{k!} \leq \frac{3q_1^{q_1+1/2} e^{-q_1}}{(n\lambda)^{q_1+1}} \sum_{k=0}^{q_1} \frac{(n\lambda)^k}{k!} \leq \frac{3q_1^{q_1+1/2} e^{-q_1}}{(n\lambda)^{q_1+1}} e^{n\lambda}. $$

Setting $q_1 = [n\lambda]$ proves (31) when $x_1 = \beta + \delta(n\lambda - [n\lambda])$. To prove (31) when $x^q_i = \delta$ and $x^q_i = 2\delta$ requires just a little more work. Setting $q_1 = n - 1$,

$$E r^+_{n-1} \leq \frac{3(n-1)^{n-1/2} e^{-(n-1)}}{(n\lambda)^n} e^{n\lambda} \leq \frac{3e}{\sqrt{n-1}} \frac{e^n}{(n-\beta\sqrt{n})^n} e^{n\lambda} = \frac{3e}{\sqrt{n-1}} \left(1 - \frac{\beta}{\sqrt{n}}\right)^{-n} e^{-\beta\sqrt{n}}. $$

To conclude, we need to bound

$$\left(\left(1 - \frac{\beta}{\sqrt{n}}\right)^{-\sqrt{n}} e^{-\beta}\right)^{\sqrt{n}} = \left(\exp\left(-\sqrt{n}\log\left(1 - \frac{\beta}{\sqrt{n}}\right) - \beta\right)\right)^{\sqrt{n}}. $$

Using Taylor expansion,

$$\log\left(1 - \frac{\beta}{\sqrt{n}}\right) = -\frac{\beta}{\sqrt{n}} - \frac{1}{2} \left(\frac{\beta}{\sqrt{n}}\right)^2 \frac{1}{(1 + \xi(\beta/\sqrt{n}))^2} $$

where $\xi(\beta/\sqrt{n}) \in [-\beta/\sqrt{n}, 0]$. Therefore,

$$\left(\exp\left(-\sqrt{n}\log\left(1 - \frac{\beta}{\sqrt{n}}\right) - \beta\right)\right)^{\sqrt{n}} = \exp\left(\frac{\beta^2/2}{(1 + \xi(\beta/\sqrt{n}))^2}\right), $$

and we conclude that

$$\sup_{n \geq 0} \left(\left(1 - \frac{\beta}{\sqrt{n}}\right)^{-\sqrt{n}} e^{-\beta}\right)^{\sqrt{n}} = \infty. $$

The argument when $q_1 = n - 2$ is identical. This proves (31) when $x_1 = \delta, 2\delta$. \qed

References

Atar R (2012) A diffusion regime with nondegenerate slowdown. Operations Research 60(2):490–500, URL http://dx.doi.org/10.1287/opre.1110.1030.

Banerjee S, Mukherjee D (2019) Join-the-shortest queue diffusion limit in Halfin-Whitt regime: Tail asymptotics and scaling of extrema. Ann. Appl. Probab. 29(2):1262–1309, URL http://dx.doi.org/10.1214/18-AAP1436.
Banerjee S, Mukherjee D (2020) Join-the-Shortest Queue diffusion limit in Halfin–Whitt regime: Sensitivity on the heavy-traffic parameter. *The Annals of Applied Probability* 30(1):80 – 144, URL http://dx.doi.org/10.1214/19-AAP1496.

Barbour A (1990) Stein’s method for diffusion approximations. *Probab. Theory and Related Fields* 84(3):297–322, ISSN 0178-8051, URL http://dx.doi.org/10.1007/BF01197887.

Barbour AD (1988) Stein’s method and Poisson process convergence. *Journal of Appl. Probab.* 25:175–184, ISSN 00219002, URL http://www.jstor.org/stable/3214155.

Braverman A (2020) Steady-state analysis of the join the shortest queue model in the Halfin-Whitt regime. *Math. Oper. Res.* 45(3):1069–1103, URL https://doi.org/10.1287/moor.2019.1023.

Braverman A (2022) The prelimit generator comparison approach of Stein’s method. *Stochastic Systems* 12(2):181–204, URL http://dx.doi.org/10.1287/stsy.2021.0085.

Braverman A, Dai JG (2017) Stein’s method for steady-state diffusion approximations of $M/Ph/n + M$ systems. *Ann. of Appl. Probab.* 27(1):550–581, ISSN 1050-5164, URL http://dx.doi.org/10.1214/16-AAP1211.

Brown TC, Xia A (2001) Stein’s method and birth-death processes. *Ann. Probab.* 29(3):1373–1403, URL http://dx.doi.org/10.1214/aop/1015345606.

Cao P, He S, Huang J, Liu Y (2021) To pool or not to pool: Queueing design for large-scale service systems. *Operations Research* 69(6):1866–1885, URL http://dx.doi.org/10.1287/opre.2019.1976.

Erdogdu MA, Mackey L, Shamir O (2019) Global non-convex optimization with discretized diffusions. URL https://arxiv.org/abs/1810.12361v1, working paper.

Eryilmaz A, Srikant R (2012) Asymptotically tight steady-state queue length bounds implied by drift conditions. *Queueing Systems* 72(3-4):311–359, ISSN 0257-0130, URL http://dx.doi.org/10.1007/s11134-012-9305-y.

Eschenfeldt P, Gamarnik D (2018) Join the shortest queue with many servers. the heavy-traffic asymptotics. *Math. Oper. Res.* 43(3):867–886, URL http://dx.doi.org/10.1287/moor.2017.0887.

Fang X, Shao QM, Xu L (2018) Multivariate approximations in Wasserstein distance by Stein’s method and Bismut’s formula. URL https://arxiv.org/abs/1801.07815.
Feller W (1968) *An introduction to probability theory and its applications. vol. I.* Third edition (New York: John Wiley & Sons Inc.).

Gast N (2017) Expected values estimated via mean-field approximation are 1/n-accurate. *Proc. ACM Meas. Anal. Comput. Syst.* 1(1), URL http://dx.doi.org/10.1145/3084454.

Gast N, Bortolussi L, Tribastone M (2019) Size expansions of mean field approximation: Transient and steady-state analysis. *Performance Evaluation* 129:60–80, ISSN 0166-5316, URL http://dx.doi.org/https://doi.org/10.1016/j.peva.2018.09.005.

Gast N, Van Houdt B (2017) A refined mean field approximation. *Proc. ACM Meas. Anal. Comput. Syst.* 1(2), URL http://dx.doi.org/10.1145/3154491.

Gaunt RE, Walton N (2020) Stein’s method for the single server queue in heavy traffic. *Statistics & Probability Letters* 156:108566, ISSN 0167-7152, URL http://dx.doi.org/https://doi.org/10.1016/j.spl.2019.108566.

Götze F (1991) On the rate of convergence in the multivariate CLT. *Ann. Probab.* 19(2):724–739, URL http://dx.doi.org/10.1214/aop/1176990448.

Gupta V, Walton N (2019) Load balancing in the nondegenerate slowdown regime. *Operations Research* 67(1):281–294, URL http://dx.doi.org/10.1287/opre.2018.1768.

Gurvich I (2014) Diffusion models and steady-state approximations for exponentially ergodic Markovian queues. *Ann. Appl. Probab.* 24(6):2527–2559, URL http://dx.doi.org/10.1214/13-AAP984.

Hairi, Liu X, Ying L (2021) Beyond scaling: Calculable error bounds of the power-of-two-choices mean-field model in heavy-traffic. *Proceedings of the Twenty-Second International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing*, 1–10, MobiHoc ’21 (New York, NY, USA: Association for Computing Machinery), ISBN 9781450385589, URL http://dx.doi.org/10.1145/3466772.3467029.

Halfin S, Whitt W (1981) Heavy-traffic limits for queues with many exponential servers. *Oper. Res.* 29(3):567–588, ISSN 0030-364X.

Hurtado-Lange D, Maguluri ST (2021) Load balancing system under join the shortest queue: Many-server-heavy-traffic asymptotics.

Jin X, Pang G, Xu L, Xu X (2021) An approximation to steady-state of m/ph/n+m queue.
Kallenberg O (2001) Foundations of Modern Probability. Springer Series in Statistics, Probability and its applications (New York: Springer), 2nd edition.

Liu X, Gong K, Ying L (2022) Steady-state analysis of load balancing with coxian-2 distributed service times. Naval Research Logistics (NRL) 69(1):57–75, URL http://dx.doi.org/https://doi.org/10.1002/nav.21986.

Liu X, Ying L (2019) A simple steady-state analysis of load balancing algorithms in the sub-halfin-whitt regime. SIGMETRICS Perform. Eval. Rev. 46(2):15–17, ISSN 0163-5999, URL http://dx.doi.org/10.1145/3305218.3305225.

Liu X, Ying L (2020) Steady-state analysis of load-balancing algorithms in the sub-halfin–whitt regime. Journal of Applied Probability 57(2):578–596, URL http://dx.doi.org/10.1017/jpr.2020.13.

Lu Y (2021) On a stein method based approximation for a two-dimensional markov chain.

Mackey L, Gorham J (2016) Multivariate Stein factors for a class of strongly log-concave distributions. Electron. Commun. Probab. 21:14, URL http://dx.doi.org/10.1214/16-ECP15.

Mitzenmacher M (2001) The power of two choices in randomized load balancing. IEEE Transactions on Parallel and Distributed Systems 12(10):1094–1104, URL http://dx.doi.org/10.1109/71.963420.

Mukherjee D, Borst SC, van Leeuwaarden JSH, Whiting PA (2016) Universality of load balancing schemes on the diffusion scale. J. Appl. Probab. 53(4):1111–1124, URL https://projecteuclid.org:443/euclid.jap/1481132840.

Ross N (2011) Fundamentals of Stein’s method. Probab. Surv. 8:210–293, ISSN 1549-5787, URL http://dx.doi.org/10.1214/11-PS182.

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

Stolyar AL (2014) Tightness of stationary distributions of a flexible-server system in the Halfin-Whitt asymptotic regime URL http://arxiv.org/abs/1403.4896v2.

Stolyar AL (2015) Pull-based load distribution in large-scale heterogeneous service systems. Queueing Systems 80(4):341–361, ISSN 1572-9443, URL http://dx.doi.org/10.1007/s11134-015-9448-8.
van der Boor M, Borst SC, van Leeuwaarden JSH, Mukherjee D (2021) Scalable load balancing in networked systems: A survey of recent advances.

Vvedenskaya N, Dobrushin R, Karpelevich F (1996) Queueing system with selection of the shortest of two queues: An asymptotic approach. *Problems Inform. Transmission* 32(1):15–27.

Weber RR (1978) On the optimal assignment of customers to parallel servers. *Journal of Applied Probability* 15(2):406–413, URL [http://dx.doi.org/10.2307/3213411](http://dx.doi.org/10.2307/3213411).

Winston W (1977) Optimality of the shortest line discipline. *Journal of Applied Probability* 14(1):181–189, ISSN 00219002, URL [http://www.jstor.org/stable/3213271](http://www.jstor.org/stable/3213271).

Ying L (2017) Stein’s method for mean field approximations in light and heavy traffic regimes. *Proc. ACM Meas. Anal. Comput. Syst.* 1(1):12:1–12:27, ISSN 2476-1249, URL [http://dx.doi.org/10.1145/3084449](http://dx.doi.org/10.1145/3084449).

Zhao Z, Banerjee S, Mukherjee D (2021) Many-server asymptotics for join-the-shortest queue in the super-Halfin-Whitt scaling window.

Zhou X, Shroff N (2020a) A note on load balancing in many-server heavy-traffic regime.

Zhou X, Shroff N (2020b) A note on Stein’s method for heavy-traffic analysis.