Stein’s Method for Mean-Field Approximations in Light and Heavy Traffic Regimes

LEI YING, School of Electrical, Computer and Energy Engineering, Arizona State University

Mean-field analysis is an analytical method for understanding large-scale stochastic systems such as large-scale data centers and communication networks. The idea is to approximate the stationary distribution of a large-scale stochastic system using the equilibrium point (called the mean-field limit) of a dynamical system (called the mean-field model). This approximation is often justified by proving the weak convergence of stationary distributions to its mean-field limit. Most existing mean-field models concerned the light-traffic regime where the load of the system, denote by $\rho$, is strictly less than one and is independent of the size of the system. This is because a traditional mean-field model represents the limit of the corresponding stochastic system. Therefore, the load of the mean-field model is $\rho = \lim_{N \to \infty} \rho(N)$, where $\rho(N)$ is the load of the stochastic system of size $N$. Now if $\rho(N) \to 1$ as $N \to \infty$ (i.e., in the heavy-traffic regime), then $\rho = 1$. For most systems, the mean-field limits when $\rho = 1$ are trivial and meaningless. To overcome this difficulty of traditional mean-field models, this paper takes a different point of view on mean-field models. Instead of regarding a mean-field model as the limiting system of large-scale stochastic system, it views the equilibrium point of the mean-field model, called a mean-field solution, simply as an approximation of the stationary distribution of the finite-size system. Therefore both mean-field models and solutions can be functions of $N$. This paper first outlines an analytical method to bound the approximation error based on Stein’s method and the perturbation theory. We further present two examples: the $M/M/N$ queueing system and the supermarket model under the power-of-two-choices algorithm. For both applications, the method enables us to characterize the system performance under a broad range of traffic loads. For the supermarket model, this is the first paper that rigorously quantifies the steady-state performance of the power-of-two-choices in the heavy-traffic regime. These results in the heavy-traffic regime cannot be obtained using the traditional mean-field analysis and the interchange of the limits.

ACM Reference format:

1 INTRODUCTION

From a hospital with hundreds of beds, to a city with thousands of Uber drivers, and to a data center with tens of thousands of servers, the scale and complexity of today’s systems one needs to manage and engineer have exploded. However, design and analysis of large-scale stochastic systems remain to be challenging problems due to "the curse of dimensionality". For large-scale stochastic systems, performance analysis beyond stability (i.e. the existence of stationary distributions) is notoriously hard and analytical tools beyond stability analysis are very limited. One such analytical tool, emerging from statistical physics, is the so called mean-field method, where the idea is to approximate the time evolution of a stochastic system, which is a random process, using the trajectory of a dynamical system, called a mean-field model; and approximate the stationary distribution of the stochastic system using the equilibrium point of the dynamical system. For many applications (e.g. [3, 8, 9, 18, 23, 25, 26, 28, 30]),

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2017 Copyright held by the owner/author(s). Publication rights licensed to ACM. 2476-1249/2017/1-ART1 $15.00
DOI: http://dx.doi.org/10.1145/3084449

it can be shown that the stationary distributions indeed converge (weakly) to their mean-field limits as the system size increases, which justifies the use of mean-field limits for analyzing large-scale stochastic systems. However, most existing results were established in the light-traffic regime where the load of the system, denote by \( \rho \), is strictly less than one and independent of the size of the system. This is because a traditional mean-field model represents the limit of the corresponding stochastic system. Therefore, the load of the mean-field model is \( \rho = \lim_{N \to \infty} \rho^{(N)} \), where \( \rho^{(N)} \) is the load of the stochastic system with size \( N \). Now if \( \rho^{(N)} \to 1 \) as \( N \to \infty \) (i.e. the heavy-traffic regime), then \( \rho = 1 \). For most systems, the mean-field limits when \( \rho = 1 \) are trivial and meaningless.

For example, consider the supermarket model under the randomized load-balancing algorithm called the-power-of-two-choices [18, 26]. The system consists of \( N \) identical servers and each maintains a separate queue. Each incoming job samples two queues and joins the shorter of the two. Assume Poisson arrivals with rate \( \lambda N \) and exponential service times with mean \( 1 \).

In the mean-field limit, the average queue length per server is finite-size system. [29] shows that for a large class of systems, we have

\[
\lambda^{(N)} = 1 - \gamma N^{-\alpha} \quad \text{for} \quad 1 > \alpha > 0 \quad \text{and} \quad 1 > \gamma > 0.
\]

While it seems reasonable to believe that the average queue length per server should be approximately \( \alpha \log_2 N \). This fact cannot be proved using the traditional mean-field model above because when \( \lambda = \lim_{N \to \infty} \lambda^{(N)} = 1 \), the limiting system becomes

\[
\dot{s}_k = (s_{k-1}^2 - s_k^2) - (s_k - s_{k+1}) \quad \forall \ k \geq 1
\]

The equilibrium of the system is \( s_k = 1 \) for all \( k \), which means all queues are infinite. This solution has little use for understanding the behavior of the original system in the heavy-traffic regime. The same phenomenon occurs when the mean-field analysis is used to analyze other queueing systems in the heavy-traffic regime because the mean-field limits trivialize when \( \rho = 1 \). Most existing approaches of mean-field analysis, except [29], are asymptotic in nature and based on the interchange of the limits. Let \( X^{(N)}(t) \) denote the state of the stochastic system at time \( t \), \( X^{(N)}(\infty) \) denote its steady-state, \( x(t) \) denote the state of the mean-field model at time \( t \), and \( x^* \) denote the corresponding mean-field limit. The idea is to first prove \( X^{(N)}(t) \overset{N \to \infty}{\longrightarrow} x(t) \) and \( x(t) \overset{t \to \infty}{\longrightarrow} x^* \), and then show \( X^{(N)}(\infty) \overset{N \to \infty}{\longrightarrow} x^* \) by exchanging the two limits. Therefore, these existing approaches have to consider a nontrivial mean-field model for \( N = \infty \) (for proving the first limiting result) and cannot be applied to the heavy-traffic regime.

To overcome this inherent difficulty of the traditional mean-field analysis, we adopt the approach based on Stein’s method used in [29], which directly bounds the approximation error of using the mean-field model for the finite-size system. [29] shows that for a large class of systems, we have

\[
E \left[ \left\| X^{(N)}(\infty) - x^* \right\|^2 \right] = O \left( \frac{1}{N} \right),
\]

in the heavy-traffic regime.
where \( \| \cdot \| \) denotes the 2-norm. Equation (3) implies the weak convergence of \( X^{(N)}(\infty) \) to \( x^* \). More importantly, it applies to finite (maybe large) size systems and quantifies the approximation error (the rate of convergence).

In this paper, we take this idea one step further. Instead of using \( x^* \) to approximate \( X^{(N)}(\infty) \) for all \( N \), we define an \( N \)-dependent mean-field model with an \( N \)-dependent \( x^{(N)} \) to approximate \( X^{(N)}(\infty) \). We then analyze the approximation error

\[
E \left[ d \left( X^{(N)}(\infty), x^{(N)} \right) \right]
\]

for some distance measure \( d(x, y) \) such that \( d(x, x) = 0 \) (e.g. \( d(x, y) \) can be the 2-norm as in [29]). Consider the supermarket model under the-power-of-two-choices with arrival rate \( (1 - \gamma N^{-\alpha})N \). We can define the mean-field model to be

\[
\begin{align*}
\bar{s}_k^{(N)} &= (1 - \gamma N^{-\alpha}) \left( \left( \bar{s}_{k-1}^{(N)} \right)^2 - \left( \bar{s}_k^{(N)} \right)^2 \right) - \left( \bar{s}_k^{(N)} - \bar{s}_k^{(N)} \right) \quad \forall k \geq 1 \\
\bar{s}_0^{(N)} &= 1,
\end{align*}
\]

which has a unique equilibrium

\[
\bar{s}_k^{(N)} = (1 - \gamma N^{-\alpha})^{k-1}.
\]

Different from the traditional mean-field model in (1), which is independent of \( N \), the mean-field model above is \( N \)-dependent.

In this paper, we call \( x^{(N)} \) a mean-field solution instead of mean-field limit because it is for a finite \( N \) instead of for \( N = \infty \). The proposed approach takes a different point of view on mean-field models. Instead of viewing a mean-field model as the limit of some stochastic system, it views the mean-field solution as just an approximation of the stationary distribution of the finite-size system. The method then focuses on quantifying the approximation error. If the approximation error is small (as we will show in two applications), then we can conclude that the mean-field solution is a good approximation of the stationary distribution. For example, if (4) is a good approximation, then it suggests \( S^{(N)}(\infty) \) concentrates around the point mass \( x^{(N)} \) and the average queue length per server is approximately \( \log_2 N \) in the heavy-traffic regime.

In this paper, we will outline a recipe for bounding the approximation error

\[
E \left[ d \left( X^{(N)}(\infty), x^{(N)} \right) \right]
\]

based on Stein’s method [7, 29] and the perturbation theory [17]. After outlining the framework, we apply this method to two queueing systems: the \( M/M/N \) queueing system and the supermarket market model under the-power-of-two-choices. The main results of this paper are summarized below.

1.1 Main Results

In the remainder of the paper, we drop the superscript \( (N) \) because \( N \) is given and fixed from now on. However, we emphasize that both mean-field models and solutions are \( N \)-dependent.

**Main Result 1**: In Section 2, we outline the proposed approach, which synthesizes mean-field analysis, Stein’s method and the perturbation theory. In particular, it is based on the following fundamental equation, which we call **Stein’s equation for mean-field models**

\[
E \left[ d (X(\infty), x^*) \right] = -E \sum_y R_{X(\infty), y} \Gamma(X(\infty), y),
\]

where \( R_{x, y} \) is the transition rate of the Continuous-Time Markov Chain (CTMC) from state \( x \) to state \( y \),

\[
\Gamma(x, y) = g(y) - g(x) - \nabla g(x) \cdot (y - x),
\]
$g(x)$ is the solution of the Poisson equation (see details in Section 2) such that

$$g(x) = -\int_0^\infty d(x(t), x') \, dt, \quad x(0) = x,$$

and $\nabla g(x)$ is the gradient of $g(x)$.

When the closed-form expression of $g(x)$ is available, obtaining the error bound is straightforward as we will show in the $M/M/N$ queueing system. However, for general stochastic systems, it is difficult (if not impossible) to get the closed-form expression of $g(x)$. In such cases, we present a lemma based on the perturbation theory to bound $\Gamma(x, y)$, where the idea is to bound the distance between the mean-field model starting from initial state $x$ and a system starting from a perturbed initial state $y$, which eventually leads to a bound on $\Gamma(x, y)$. We will see that the approximation error is determined by the convergence rate of the mean-field model (the dynamical system), which is a function of the load of the system. The heavier the load, the slower the convergence and the larger the approximation error.

![Diagram](image)

Fig. 1. The proposed approach can be used to analyze large-scale stochastic systems under a broad range of load regimes.

As shown in Figure 1, the traditional mean-field analysis for queueing systems only concerns large systems with light load (the blue star point). The approach in this paper enables us to study large-scale stochastic systems under a broad range of load conditions (the green rectangle area). Furthermore, our result identifies the fundamental connection between the convergence rate of the mean-field model and the approximation error of the mean-field solution. From the best of our knowledge, this is the first paper that explicitly identifies this connection.

**Main Result 2:** In the first application, we consider the $M/M/N$ queueing system (or Erlang-C queueing system), which consists of $N$ servers and a single queue. For this system, the closed-form expression of $g(x)$ can be obtained. Assuming the arrival rate is $\lambda N$ and service times are exponential with mean 1, we establish the following results. For any $N$ and $\lambda$ such that $N(1 - \lambda) \geq 4$, we have

$$E \left[ \left( \frac{Q(\infty)}{N} - \lambda \right) \right] \leq \frac{6(\lambda + 1)}{N} + \frac{36(1 + \lambda)}{N^2(1 - \lambda)^2},$$

and

$$\lambda \leq E \left[ \frac{Q(\infty)}{N} \right] \leq \lambda + \frac{(1 + \lambda)}{N - \lambda N},$$

where $Q(\infty)$ is the number of jobs in the system at the steady-state. This bound is universal. To see this, let $\lambda = 1 - \gamma N^{-\alpha}$, so $\gamma N^{1-\alpha}$ is the gap between the arrival rate and the total service rate. When $\alpha$ varies from 0 to 1, the bound $E \left[ \left( \frac{Q(\infty)}{N} - \lambda \right) \right]$ gets tighter as $\alpha$ increases.
1, the gap $\gamma^N N^{1-\alpha}$ varies from a constant fraction of $N$ (i.e., light load) to a constant independent of $N$ (i.e., very heavy traffic load). When $\alpha < 0.5$, (6) implies that

$$\Pr (Q(\infty) > N) \to 0 \text{ as } N \to \infty,$$

which recovers a Halfin-Whitt result [16] and implies almost all jobs are immediately served upon arrival when the difference between the total service rate $N$ and the arrival rate $\lambda N$ is $\omega(\sqrt{N})$. The details can be found in Section 3.

We remark that the $M/M/N$ queueing system has been well studied in the literature. For example, more accurate estimations have been obtained using diffusion approximation and Stein’s method [5, 7]. The purpose of this example is not to provide a better estimation. Rather, it demonstrates the proposed approach in the case where we have the closed-form expression of $g(x)$. Also the analysis for obtaining the universal bounds and the Halfin-Whitt result is remarkably simple (about 2 pages in total), which itself could be interesting.

We also like to comment that the mean-field model in this case is

$$\hat{q} = \lambda - \min\{1, q\} \quad q(0) \geq 0,$$

where $q = Q/N$. Note that $\lambda - \min\{1, q\}$ is not differentiable and $q$ is not bounded. Hence, this system does not satisfy the conditions required in [29]. Therefore, this example also shows the proposed approach can be applied to systems beyond those considered in [29].

**Main Result 3:** In the second application, we consider the supermarket model under the-power-of-two-choices. We assume $\lambda = 1 - \gamma N^{-\alpha}$ and have the following result. For any $0 \leq \alpha < 0.2$ and an arbitrarily small $\xi' > 0$, the following inequalities hold for sufficiently large $N$,

$$E \left[ \|S(\infty) - s^*\|^2 \right] \leq \frac{1}{N^{1-2\alpha-\xi'}}$$

and

$$E \left[ \sum_k S_k(\infty) - \sum_k s_k^* \right] \leq \frac{1}{N^{1-2\alpha-\xi'}}.$$

Note that when $\alpha > 0$, this quantifies the steady-state performance of the-power-of-two-choices in the heavy traffic regime. The most related work is a recent result independently developed in [12]. In [12], the authors studied the performance of the-power-of-d-choices in the heavy-traffic regime. They defined a rescaled process $T^{(N)}(t) = \frac{1}{1-S^{(N)}(t)}$, and established the process-level convergence of $T^{(N)}(t)$ to $\bar{T}(t)$ for fixed $t$ as $N \to \infty$, where $\bar{T}(t)$ is the unique solution of a deterministic system defined by linear differential equations in [12]. The deterministic system has a unique equilibrium point $T_k^* = \frac{d k - 1}{d-1}$. From this result, the authors conjectured that $S_k(\infty) \approx 1 - (1 - \lambda) d^{k-1}$. Note that this is a conjecture because to prove the result, it needs to be shown that $T^{(N)}(\infty)$ is tight, which was left as an open problem in [12]. The authors also obtained a lower bound on the expected value of $S_k(\infty)$

$$E [S_k(\infty)] \geq 1 - (1 - \lambda) \frac{d^k - 1}{d - 1}. $$

When $d = 2$ (the-power-of-two-choices) and $k \leq \log_2 (1 - \lambda) - \omega(1)$, the two results above suggest that $S_k(\infty) \approx 1$. They further used the light-traffic result in [18, 26] as a heuristic to argue that $S_k(\infty) = 0$ for $k \geq \log_2 (1 - \lambda) + \omega(1)$. From these observations and heuristics, the authors conjectured that at the steady-state, most queues are of size $\log_2 (1 - \lambda) + O(1)$, which equals to $\alpha \log_2 N + O(1)$ given $\lambda = 1 - \gamma N^{-\alpha}$ ($0 < \alpha < 1$). For $0 < \alpha < 0.2$, this paper studies the steady-state queue lengths of the-power-of-two-choices and bounds both the expected value and the mean-square error of using the mean-field solution for approximating the stationary
distribution. Our result in (7) shows that for a sufficiently large $N$, the stationary distribution concentrates around the point mass $s^*$ such that $s^*_k = \lambda^{2^k-1}$. It is easy to verify that $s^*_k \approx 1$ for $k \leq \left| \log_2(1 - \lambda) \right| - \omega(1)$ and $s^*_k \approx 0$ for $k \geq \left| \log_2(1 - \lambda) \right| + \omega(1)$. Therefore, our result resolves the open conjecture in [12] for $0 < \alpha < 0.2$, and is the first result on the steady-state approximation of the-power-of-two-choices in the heavy-traffic regime.

1.2 Related Work

Mean-field analysis originated from statistical physics has a broad range of applications in many areas. In queueing theory, mean-field models are equivalent to fluid models [10] for many-server queues. Mean-field approach often models a stochastic system as a population process, and then approximates the population process using a dynamical system, which is deterministic. In large-scale systems (e.g. many-server queues), this approximation often turns out to be very accurate, and enables us to quantitatively study the stationary distribution of a large-scale stochastic system. The method to bound the approximation error follows the one used in [29], which is based on Stein’s method [2, 21, 22], a method to bound the distance of two probability distributions. In particular, it follows the framework proposed in [6, 15] to couple the stationary distributions of two stochastic systems via their generators. Please see [7] for an introduction of Stein’s method for steady-state diffusion approximation and its connection to classical applications of Stein’s method. The key difference between this paper and the works mentioned above is that we use mean-field (fluid) approximation instead of diffusion approximation [7, 15]. Diffusion approximation is more accurate and yields tighter bounds, but analyzing a high-dimensional diffusion process is often very difficult (if not impossible). We also note that the same method has been used in [24] to prove the tightness of stationary distributions of a flexible-server system in the Halfin-Whitt regime, which proves the interchange of the limits and justifies the diffusion approximation. While the approaches are identical, the goals are very different. [24] uses the method to prove the tightness and justifies the diffusion approximation. This paper demonstrates that mean-field (fluid) approximation itself can be very accuracy for large-scale stochastic systems under a broad range of traffic loads, including the Halfin-Whitt regime.

The contributions compared with the traditional mean-field models have been discussed in the introduction. In summary, we propose $N$-dependent mean-field models and mean-field solutions. Together with Stein’s method and the perturbation theory, it enables us to study large-scale stochastic systems under a broad range of traffic loads and further quantify approximation errors, both are not possible using the traditional approaches based on the interchange of the limits. There are some recent work on heavy-traffic analysis of randomized load balancing in many server systems [11, 12, 19], where the focus is on process-level (i.e. finite $t$) approximation, not the steady-state approximation. Approximation errors of mean-field models have also been studied at the process-level (finite $t$) [13] and at the steady-state [29], but both are in the light traffic regime. Finally, the use of $N$-dependent dynamical systems for approximating stochastic systems has been suggested in [4], but it does not provide any theoretical analysis on the approximation error.

2 STEIN’S METHOD AND MEAN-FIELD APPROXIMATIONS

We first overview the analytical framework based on Stein’s method for mean-field models [29]. Let $X$ be a continuous time Markov chain (CTMC) with a unique stationary distribution. We use $X(\infty)$ to denote the random vector that has the stationary distribution of $X$. Define $\hat{x} = f(x)$ to be the mean-field model with a unique equilibrium point $x^*$ such that $f(x^*) = 0$. Our proposed method is to approximate $X(\infty)$ with $x^*$ and quantify the approximation error $E[d(X(\infty), x^*)]$ based on the following lemma.

**Lemma 2.1 (Stein’s equation for mean-field models).** Let $X$ denote a CTMC with a countable state space $\mathcal{X}$, and $R_{x,y}$ denote the transition rate of the CTMC from state $x$ to state $y$. Define the mean-field model to be

$$\hat{x} = f(x) = \sum_{y \neq x} R_{x,y} (y - x),$$
and define
\[ g(x) = - \int_0^\infty d(x(t), x^*) \, dt, \quad x(0) = x. \]  

(10)

Under the assumptions that

1. the CTMC has a unique stationary distribution \( X(\infty) \),
2. the mean-field model has a unique equilibrium point \( x^* \) such that \( f(x^*) = 0 \), and
3. given distance measure \( d(x, x^*) \), \( |g(x)| < \infty \) for any \( x \in X \),

the following Stein’s equation holds
\[ E[d(X(\infty), x^*)] = -E \left[ \sum_y R_{X(\infty), y} \Gamma(X(\infty), y) \right], \]

where
\[ \Gamma(x, y) = g(y) - g(x) - \nabla g(x) \cdot (y - x). \]

Proof. The first key idea is Stein’s method, which combines the Poisson equation (also known as the Stein equation) and the steady-state equation (also called the Basic Adjoint Relationship (BAR) [14]):
\[ \nabla g(x) \cdot f(x) = d(x, x^*) \quad \forall x \quad \text{(Poisson Equation)} \]  
\[ E[Gg(X(\infty))] = 0, \quad \text{(Steady-State Equation)} \]  

where \( g(x) \) is called the solution of the Poisson equation and \( G \) denotes the generator of the CTMC \( X \). Consider the \( g(x) \) defined in (10). Assume the mean-field model starts from \( t = \tau \), then we have
\[ g(x(\tau)) = - \int_\tau^\infty d(x(t), x^*) \, dt. \]  

(14)

When assumptions (2) and (3) hold, i.e. \( g(x) \) is well defined for all \( x \in X \), we can verify the Poisson equation by taking derivative at both sides of equation (14) with respect to \( \tau \), and verify the steady-state equation based on the global balance equation [20, p. 234].

Now combining (12) and (13), we obtain
\[ E[d(X(\infty), x^*)] = E[\nabla g(X(\infty)) \cdot f(X(\infty)) - Gg(X(\infty))] \]
\[ = E \left[ \nabla g(X(\infty)) \cdot \sum_{y, y \neq X(\infty)} R_{X(\infty), y} (y - X(\infty)) - \sum_{y, y \neq X(\infty)} R_{X(\infty), y} (g(y) - g(X(\infty))) \right] \]
\[ = - E \left[ \sum_{y, y \neq X(\infty)} R_{X(\infty), y} (g(y) - g(X(\infty))) - \nabla g(X(\infty)) \cdot (y - X(\infty)) \right]. \]

Remark: This lemma follows the idea of Stein’s method for steady-state approximation [7], and can also be obtained from equation (4) in [24].

The lemma above shows that the approximation error, \( E[d(X(\infty), x^*)] \), can be calculated using the approximation error of the first-order Taylor expansion of \( g(x) \), i.e.
\[ \Gamma(x, y) = g(y) - g(x) - \nabla g(x) \cdot (y - x). \]  

(15)

Quantifying the difference in (15) is known as establishing gradient bounds for the Poisson equation solution \( g \). While only limited results on gradient bounds are known for high-dimensional diffusion processes, gradient

bounds for mean-field models are more tractable because high-dimensional deterministic systems are easier to track and analyze as shown in the following discussion.

If \( g(x) \) has a closed-form expression, then it is straightforward to obtain a bound on the approximation error by bounding \( \Gamma(x, y) \), as we will see in the \( M/M/N \) queueing system. However, in many cases, the closed-forms may not exist. In such cases, we leverage the perturbation theory to bound \( \Gamma(x, y) \). Following the analysis in [29], we consider the following collection of dynamical systems:

\[
\dot{x}(t) = f(x(t), y) - \frac{1}{N} \frac{\partial f}{\partial x}(x(t), y) x^{(1)}(t)
\]

(16)

\[
\dot{x}^{(1)}(t) = \frac{\partial f}{\partial x}(x(t), y) x^{(1)}(t)
\]

(17)

\[
\dot{x}(t, y) = f(x(t), y)
\]

(18)

\[
\dot{x}(t, z) = f(x(t, z))
\]

(19)

with initial conditions \( e(0) = 0, x^{(1)}(0) = N(z - y), x(0, z) = z \) and \( x(0, y) = y \), where \( \frac{\partial f}{\partial x} \) denotes the Jacobian matrix.

**Lemma 2.2 (Gradient Bound for Mean-Field Models).** Assume the following conditions hold:

**C1** Given initial condition \( z \) and any positive constant \( d \), there exists \( \tilde{t}_{d, z} \) such that

\[
|x(t, z)| \leq \tilde{d} \quad \forall t \geq \tilde{t}_{d, z}.
\]

**C2** There exists constant \( c_1 \) such that

\[
|x^{(1)}(t)| \leq c_1|x^{(1)}(0)| \quad \forall t.
\]

(20)

**C3** There exist Lyapunov function \( V_1(x^{(1)}) \) and positive constants \( c_{u1}, c_{i1}, d_1 \) and \( \delta_1 \) such that

1. \( c_{i1}|x^{(1)}| \leq V_1(x^{(1)}) \leq c_{u1}|x^{(1)}| \), and
2. when \( |x(t, z)| \leq d_1 \),

\[
V_1(x^{(1)}(t)) \leq -\delta_1 V_1(x^{(1)}(t)).
\]

**C4** There exist a positive constant \( c_e \) such that given \( |e(t)| \leq \frac{1}{N} \),

\[
\frac{d|e(t)|}{dt} \leq c_e \frac{N^2}{N^2}.
\]

**C5** There exist Lyapunov function \( V_e(e) \) and positive constants \( c_{er}, \delta_e, d_e \) and \( \alpha \) such that

1. \( c_{e1}|e| \leq V_e(e) \leq c_{eu}|e| \), and
2. when \( |x(t, z)| \leq d_e \),

\[
V_e(e(t)) \leq -\delta_e V_e(e(t)) + \frac{c_{er}}{N^2} |x^{(1)}(t)|^\alpha.
\]

**C6** There exists constant \( b \) such that \( |x^{(1)}(0)| = N|z - y| \leq b \) for any \( y, z \in X \). Furthermore, \( \tilde{t}_{d, x} = o(N) \) and \( \tilde{t}_{d, x} = o(N) \) for any \( x \in X \).

**C7** The following constants are independent of \( N : c_1, c_{i1}, c_{u1}, c_{ie}, c_{eu}, c_e, \alpha \), and \( b \). Then there exists a positive constant \( \kappa \), independent of \( N \), such that when \( N \) is sufficiently large,

\[
\int_0^\infty |x^{(1)}(t)|^2 dt \leq \kappa \left( \tilde{t}_{d, z} + \frac{1}{\delta_1} \right)
\]

(21)

\[
\int_0^\infty |e(t)| dt \leq \kappa \left( \tilde{t}_{d, z} \frac{1}{\delta_e} + \frac{1}{\delta_1 \delta_e} \tilde{t}_{d, z}^2 \right) \frac{1}{N^2}.
\]

(22)
where \( d = \min \{ d_1, d_e \} \).

The lemma above provides bounds on \( \int_0^\infty |x^{(1)}(t)|^2 \, dt \) and \( \int_0^\infty |e(t)| \, dt \), which can be used to bound \( |\Gamma(z, y)| \). For example, if we use 2-norm as the distance measure, we will have

\[
|\Gamma(z, y)| \leq \int_0^\infty \left( 3x_{\text{max}}|e(t)| + \frac{1}{N^2}|x^{(1)}(t)|^2 \right) \, dt;
\]

(23)

where \( x_{\text{max}} \) is a constant such that \( x_i \leq x_{\text{max}} \) for any \( x \in \mathcal{X} \) and any \( i \); and if we use 1-norm as the distance measure, we will have

\[
|\Gamma(z, y)| \leq \int_0^\infty |e(t)| \, dt.
\]

(24)

The proof of this lemma is presented in Appendix A and is based on the analysis in [29] for bounding \( |\Gamma(z, y)| \), which, however, is for the light traffic regime. Lemma 2.2 summarizes the set of conditions under which the gradient bound can be established. Furthermore, we state \( \tilde{i}_d, \tilde{z}, \delta_1, \) and \( \delta_e \) explicitly in the bound because the values of these quantities are determined by the convergence rates of the mean-field model (both global convergence rate and local convergence rate around the equilibrium). The convergence rates, as we will see in the example of power-of-two-choices, depend on the load of the system. Therefore, Lemma 2.2 can be used to obtain gradient bounds under a broad range of load conditions.

3 THE \( M/M/N \) QUEUEING SYSTEM

An \( M/M/N \) queueing system is a system with \( N \) identical servers and a single queue as shown in Figure 2. We assume a Poisson arrival process with rate \( \lambda N \) such that \( N - \lambda N \geq 4 \), and exponentially distributed service times with mean 1. Furthermore, jobs are served in a First-In-First-Out (FIFO) fashion. This example represents the scenario where the closed-form expression of \( g(x) \) can be obtained. Letting \( Q(t) \) denote the number of jobs in the system, and \( Q(\infty) \) denote the steady-state of the system, we have the following theorem.

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{queueing_system.png}
\caption{The \( M/M/N \) Queueing System}
\end{figure}

\textbf{Theorem 3.1.} Consider the \( M/M/N \) queueing system. For any \( N > 0 \) and \( 0 < \lambda < 1 \) such that \( N(1 - \lambda) \geq 4 \), we have

\[
E \left[ \left( \frac{Q(\infty)}{N} - \lambda \right)^2 \right] \leq \frac{6(\lambda + 1)}{N} + \frac{36(1 + \lambda)}{N^2(1 - \lambda)^2}.
\]

(25)

Furthermore, for any \( N \) and \( \lambda \),

\[
\lambda \leq E \left[ \frac{Q(\infty)}{N} \right] \leq \lambda + \frac{(1 + \lambda)}{N - \lambda N}.
\]

(26)

\[\square\]
Define $q(t) = \frac{Q(t)}{N}$. We have the following mean-field model (fluid-model) for this system:

$$\dot{q} = \lambda - \min\{1, q\} \quad q(0) \geq 0.$$  

(27)

It is easy to verify that the unique equilibrium point of the system is $q = \lambda$. Define $x(t) = q(t) - \lambda$. We get

$$\dot{x} = \lambda - \min\{1, x + \lambda\} = \max\{\lambda - 1, -x\} \quad x(0) \geq -\lambda.$$  

(28)

To obtain (25), we define $d(x, x^*) = x^2$. From the Poisson equation $\nabla g(x) \dot{x} = x^2$, we have

$$\nabla g(x) = \left\{ \begin{array}{ll}
-x^2 & -1 \leq x \leq 1 - \lambda \\
\frac{x^2}{(1 - \lambda)^2} - \frac{\lambda}{2(1 - \lambda)} & \lambda \leq x \leq 1 - \lambda \\
x > 1 - \lambda & \end{array} \right.$$  

which implies that

$$g(x) = \left\{ \begin{array}{ll}
-x & x \leq 1 - \lambda \\
\frac{x}{(1 - \lambda)^2} - \frac{\lambda}{2(1 - \lambda)} & \lambda \leq x \leq 1 - \lambda \\
x > 1 - \lambda & \end{array} \right.$$  

From this closed-form expression of $g(x)$, we can bound $\Gamma(x, y)$ and obtain (25). The details are given in Appendix B.

To obtain (26), we define $d(x, x^*) = x$. From the Poisson equation $\nabla g(x) \dot{x} = x$, we have

$$\nabla g(x) = \left\{ \begin{array}{ll}
-1 & -1 \leq x \leq 1 - \lambda \\
-x & \lambda \leq x \leq 1 - \lambda \\
x > 1 - \lambda & \end{array} \right.$$  

which implies that

$$g(x) = \left\{ \begin{array}{ll}
-x & x \leq 1 - \lambda \\
\frac{x}{(1 - \lambda)^2} - \frac{\lambda}{2(1 - \lambda)} & \lambda \leq x \leq 1 - \lambda \\
x > 1 - \lambda & \end{array} \right.$$  

From this closed-form expression of $g(x)$, we can first bound $\Gamma(x, y)$ which leads to the upper bound in (26). The details are given in Appendix B. The lower bound in (26) holds because the system has to have at least $\lambda N$ busy servers on average at each time instance.

**Remark:** Inequality (26) implies that

$$\lambda N \leq E[Q(\infty)] \leq \lambda N + \frac{(1 + \lambda)}{1 - \lambda}.$$  

In other words, the average queue length is approximately $\lambda N$ unless $1 - \lambda = O\left(\frac{1}{N}\right)$. Note that $1 - \lambda = O\left(\frac{1}{N}\right)$ is equivalent to $N - \lambda N = O(1)$, i.e., the arrival rate and the total service rate differ by a constant only, which is a very heavily loaded regime, called the nondegenerate slowdown regime (NDS) [1, 7, 27].

**Remark:** From (25), we can also bound the probability that all servers are busy and recovers a Halfin-Whitt result [16]. According to (25) and the Markov inequality, we have

$$\Pr\{Q(\infty) \geq N\} \leq \frac{\Pr\{(Q(\infty) - \lambda N)^2 \geq (1 - \lambda)^2 N^2\}}{E[Q(\infty) - \lambda N]^2} \leq \frac{6(\lambda + 1)N + \frac{36(1 + \lambda)}{(1 - \lambda)^2}}{(1 - \lambda)^2 N^2}.$$  

Substituting $\lambda = 1 - \gamma N^{-\alpha}$ for $0 \leq \alpha < 1$, we get

$$\Pr\{Q(\infty) \geq N\} \leq \frac{6N + 36N^{2\alpha}}{N^{2\alpha}}.$$
When \( \alpha < 0.5 \),
\[
\Pr (Q(\infty) \geq N) = O \left( \frac{1}{N^{1-2\alpha}} \right) \to 0 \text{ as } N \to \infty,
\]
which recovers a celebrated Halfin-Whitt result [16] and implies that almost all jobs are immediately served upon arrival when the difference between the total service rate \( N \) and the arrival rate \( \lambda N \) is \( \omega \left( \sqrt{N} \right) \).

4 THE-POWER-OF-TWO-CHOICES

In this section, we consider the supermarket model introduced in [18, 26] with \( N \) identical servers. Each server has a separate queue as shown in Figure 3. Assume jobs arrive at the system following a Poisson process with rate \( \lambda N \) and the processing time of each job is exponentially distributed with mean \( \mu = 1 \).

Note that the results in [18, 26] were obtained under the assumption that \( \lambda < 1 \) and is independent of \( N \). We do not make such an assumption and assume \( \lambda = 1 - \gamma N^{-\alpha} \) and \( \gamma > 0.75 \). When \( \alpha = 0 \) and \( 0 < \gamma < 1 \), it is the traditional mean-field model. For any \( \alpha > 0 \), \( \lambda \to 1 \) as \( N \to \infty \), which represents the heavy-traffic regime with different rates of approaching the system capacity. Let \( Q_m(t) \) denote the queue size of server \( m \) at time \( t \). For each incoming job, the router (or called a scheduler) routes the job using the-power-of-two-choices algorithm, which randomly samples two servers and dispatches the job to the server with a smaller queue size. In this setting, \( Q(t) \) is a CTMC and has a unique stationary distribution when \( \lambda < 1 \) [26].

Let \( S_k(t) \) denote the fraction of servers with queue size at least \( k \) at time instant \( t \). \( S \) is also a CTMC with the following transition rates:
\[
R_{s,s'} = \begin{cases} 
N(s_k - s_{k+1}), & \text{if } s' = s - \frac{1}{N} \\
\lambda N \left( s_{k-1}^2 - s_k^2 \right), & \text{if } s' = s + \frac{1}{N} \\
\sum_{k=1}^{\infty} -\lambda N \left( s_{k-1}^2 - s_k^2 \right) - N(s_k - s_{k+1}), & \text{if } s' = s \\
0, & \text{otherwise.} 
\end{cases} \tag{29}
\]

where \( 1_k \) is an infinite-dimensional vector such that the \( k \)th element is 1 and the others are 0. Note that the first term is for the event that a departure occurs at a queue with size \( k \) so \( s_k \) decreases by \( 1/N \). The second term is for the event that an arrival occurs and it is routed to a queue with size \( k - 1 \). Recall that the mean-field model for this system is [18, 26]
\[
\begin{align*}
\dot{s}_k &= \lambda \left( s_{k-1}^2 - s_k^2 \right) - (s_k - s_{k+1}) \quad \forall \ k \geq 1 \\
\dot{s}_0 &= \lambda \left( s_{k-1}^2 - s_k^2 \right) - (s_k - s_{k+1}) \quad \forall \ k \geq 1 \\
s_0 &= 1,
\end{align*} \tag{30}
\]
which has a unique equilibrium point

$$s^*_k = 2^{k-1}. \quad (31)$$

Let $| \cdot |$ denote 1-norm. The following theorem quantifies the error of using $s^*$ for approximating the steady state distribution $S(\infty)$ in both the light and heavy traffic regimes.

**Theorem 4.1.** Consider the supermarket model under the power-of-two-choices with $\lambda = 1 - \gamma N^{-\alpha}$ and $\lambda > 0.75$. Given $0 \leq \alpha < 0.2$, and $\xi' > 0$ that can be chosen arbitrarily small, the following inequalities hold when $N$ is sufficiently large

$$E \left[ \| S(\infty) - s^* \|^2 \right] \leq \frac{1}{N^{1-2\alpha-\xi}}, \quad (32)$$

and

$$| E \left[ | S(\infty) | - | s^* | \right] \leq \frac{1}{N^{1-2\alpha-\xi}}. \quad (33)$$

**Proof.** We first prove (32), the bound on the mean-square error. Note that the mean-field model (30) is an infinite dimensional system. Analysis of perturbed infinite-dimensional nonlinear systems is difficult. To overcome this difficulty, we consider the following $n$-dimensional truncated system,

$$\tilde{s}_k = \begin{cases} 
\lambda (\tilde{s}_k^2 - \tilde{s}_k^1) - (\tilde{s}_k - \tilde{s}_{k+1}), & n - 1 \leq k \geq 1; \\
\lambda (\tilde{s}_{n-1}^2 - \tilde{s}_n^1) - (\tilde{s}_n - \tilde{s}_{n+1}), & k = n.
\end{cases} \quad (34)$$

where $\tilde{s}_k(t) = 1$ for $t \geq 0$. It is easy to verify that the unique equilibrium point of this truncated system. We next establish the upper bound on

$$E \left[ \| X(\infty) \|^2 \right],$$

where $X(\infty) = S(\infty) - s^*$, by bounding $E \left[ \sum_{k=1}^{n} X_k^2(\infty) \right]$ and $E \left[ \sum_{k=n+1}^{\infty} X_k^2(\infty) \right]$ separately, where we choose $n = N^{\alpha + \xi}$ with an arbitrarily small $\xi > 0$.

In Lemma 4.2, we use a minor variation of Stein’s equation in Lemma 2.1 and the truncated mean-field model (34) to show that for sufficiently large $N$,

$$E \left[ \sum_{k=1}^{n} X_k^2(\infty) \right] \leq \left( \frac{1}{N^{1-5\alpha-7\xi}} \right) E \left[ \sum_{k=1}^{n} X_k^2(\infty) \right] + \frac{1}{N^{1-2\alpha-6\xi}}.$$

Then in Lemma 4.3, we use the fact $E [S_k(\infty) ] \leq \lambda^k$ (equation (5.3) in [26]) to show that when $N$ is sufficiently large,

$$E \left[ \sum_{k=n+1}^{\infty} X_k^2(\infty) \right] \leq \frac{1}{N^{1-2\alpha}}.$$

Combining the two inequalities above, we have that when $N$ is sufficiently large,

$$E \left[ \sum_{k=1}^{n} X_k^2(\infty) \right] \leq \left( \frac{1}{N^{1-5\alpha-7\xi}} \right) E \left[ \sum_{k=1}^{n} X_k^2(\infty) \right] + \frac{1}{N^{1-2\alpha-6\xi}} + \frac{1}{N^{1-2\alpha}} \leq \left( \frac{1}{N^{1-5\alpha-7\xi}} \right) E \left[ \sum_{k=1}^{\infty} X_k^2(\infty) \right] + \frac{1}{N^{1-2\alpha-7\xi}}.$$

By moving the first term to the left-hand-size and then dividing both sides by $1 - \frac{1}{N^{1-5\alpha-7\xi}}$, we get

$$E \left[ \sum_{k=1}^{\infty} X_k^2(\infty) \right] \leq \frac{1}{1 - \frac{1}{N^{1-5\alpha-7\xi}}} \frac{1}{N^{1-2\alpha-7\xi}}.$$
Recall that $0 \leq \alpha < 0.2$, so \( \frac{1}{N^{1-2\alpha-\varepsilon}} \) can be made arbitrarily small when choosing sufficiently large $N$. Therefore, (32) holds when $N$ is sufficiently large and by choosing $\xi' = 8\xi$.

To prove (33), we again consider the truncated mean-field model and let $d(x, x') = \sum_{k=1}^n x_k$. The upper bound in (33) is established by applying the bound on \( \int_0^\infty \sum_{k=1}^n |\epsilon_k(t)|\, dt \) and the fact that $E[S_N(\infty)] \leq \lambda^k$. □

**Lemma 4.2.** Consider the supermarket model under the power-of-two choices. Given $n = N^{\alpha + \xi}$, where $\xi$ can be chosen arbitrarily small, the following inequality holds when $N$ is sufficiently large,

\[
E \left[ \sum_{k=1}^n X_k^2(\infty) \right] \leq \frac{1}{N^{1-5\alpha - \xi}} \left[ E \left[ \sum_{k=1}^n (X_k(\infty))^2 \right] + \frac{1}{N^{1-2\alpha-\varepsilon}} \right].
\]  

(35)

**Proof.** Note that the mean-field model is a truncated system, it is an imperfect mean-field model [29]. The Stein’s equation differs slightly from the one in Lemma 2.1. We first define $x_k = \bar{x}_k - s_k^*$ for $1 \leq k \leq n$, so

\[
\dot{x}_k = f_k(x) := \begin{cases}
\lambda((x_{k-1} + s_{k-1}^*)^2 - (x_k + s_k^* )^2) - (x_k + s_k^* - x_{k+1}^* - s_{k+1}^*), & 1 \leq k \leq n - 1; \\
\lambda((x_{n-1} + s_{n-1}^*)^2 - (x_n + s_n^* )^2) - (x_n + s_n^* - s_{n+1}^*), & k = n \\
-\frac{\lambda}{2} (x_k^2 + 2s_k^* x_k) - (x_n - x_k), & k = 1; \\
\lambda \left( (x_{k-1}^2 + 2s_{k-1}^* x_{k-1}) - (x_k^2 + 2s_k^* x_k) \right) - (x_k - x_{k+1}), & 2 \leq k \leq n - 1; \\
\lambda \left( (x_{n-1}^2 + 2s_{n-1}^* x_{n-1}) - (x_n^2 + 2s_n^* x_n) \right) - x_n, & k = n
\end{cases}
\]  

(36)

The unique equilibrium point for the system is $x^* = 0$.

Consider $d(x, x') = \sum_{k=1}^n x_k^2$. In this case,

\[
g(x) = -\int_0^\infty \sum_{k=1}^n (x_k(t, x))^2\, dt
\]

where $x_k(t, x)$ is the solution of the $n$-dimensional truncated dynamical system (36). By combining the Poisson equation and the steady-state equation, we have

\[
E \left[ \sum_{k=1}^n X_k^2(\infty) \right] = E \left[ \nabla g(X(\infty)) \cdot f(X(\infty)) - Gg(X(\infty)) \right].
\]

Note that $X(\infty) = S(\infty) - s^*$ is an infinite dimensional vector. $f(\cdot)$ and $g(\cdot)$ are defined by the $n$-dimensional truncated system, so are functions of the truncated $X(\infty)$.

From the definition of $f_k$ in (36) and $R_{x,y}$ in (29), we have

\[
f_k(x) = \begin{cases}
\sum_{y \neq x} R_{x,y}(y_k - x_k), & \text{if } 1 \leq k < n \\
\sum_{y \neq x} R_{x,y}(y_n - x_n - x_{n+1}), & \text{if } k = n
\end{cases}
\]

so

\[
\nabla g(x) \cdot f(x) = \sum_{k=1}^n \frac{\partial g}{\partial x_k} \left( \sum_{y \neq x} R_{x,y}(y_k - x_k) \right) - \frac{\partial g}{\partial x_n} (x_{n+1})
\]

\[
= \sum_{y \neq x} R_{x,y} \sum_{k=1}^n \frac{\partial g}{\partial x_k} (y_k - x_k) - \frac{\partial g}{\partial x_n} (x_{n+1})
\]

\[
= \sum_{y \neq x} R_{x,y} \nabla g(x) \cdot (y - x) - \frac{\partial g}{\partial x_n} (x_{n+1})
\]

where the last equality holds because $\frac{\partial g}{\partial x_k} = 0$ for $k > n$. Therefore, we have the following equation, which is a minor variation of Stein’s equation in Lemma 2.1,

$$
E \left[ \sum_{k=1}^{n} X_k^2(\infty) \right] = E \left[ \frac{\partial g}{\partial x_n} (X(\infty)) X_{n+1}(\infty) - \sum_{y \neq X(\infty)} R_{X(\infty), y} \Gamma(X(\infty), y) \right].
$$

Now according to inequalities (53) and (54) in Appendix C, and inequality (23), we have that for sufficiently large $N$,

$$
|\Gamma(z, y)| \leq \left( \frac{1}{N^{2-4\alpha-5\xi}} \right) \left( \max \left\{ \log 24 \sum_{k=1}^{n} |z_k|, 0 \right\} \right)^2 + \frac{1}{N^{2-2\alpha-5\xi}}.
$$

Furthermore, from Lemmas C.6 and C.8, we have

$$
\left| \frac{\partial g}{\partial z_n}(z) \right| = \left| \int_0^\infty \sum_{k=1}^{n} 2x_k(t, z) \frac{\partial}{\partial z_n} x_k(t, z) \, dt \right|
\leq 2 \left| \int_0^\infty \sum_{k=1}^{n} \left( \frac{\partial}{\partial z_n} x_k(t, z) \right) \, dt \right|
\leq N^{2\alpha+3\xi} \max \left\{ \log 24 \sum_{k=1}^{n} |z_k|, 0 \right\} + N^{\alpha+3\xi},
$$

where the last inequality holds when $N$ is sufficiently large. It has been shown in [26] that $E \left[ S_n(\infty) \right] \leq \lambda^n$ for any $n \geq 0$, so when $n$ is sufficiently large,

$$
E \left[ |X_{n+1}(\infty)| \right] \leq E \left[ |S_{n+1}(\infty)| \right] + s_{n+1}^* \leq \lambda^{n+1} + \lambda^{2^n+1-1} \leq \lambda^n.
$$

Therefore, we conclude

$$
E \left[ \sum_{k=1}^{n} X_k^2(\infty) \right] \leq E \left[ \lambda^n \left( N^{2\alpha+3\xi} \max \left\{ \log 24 \sum_{k=1}^{n} |X_k(\infty)|, 0 \right\} + N^{\alpha+3\xi} \right) \right]
+ E \left[ \left( \frac{1}{N^{2-4\alpha-5\xi}} \left( \max \left\{ \log 24 \sum_{k=1}^{n} |X_k(\infty)|, 0 \right\} \right)^2 + \frac{1}{N^{2-2\alpha-5\xi}} \right) \left( \sum_{y \neq X(\infty)} R_{X(\infty), y} \right) \right].
$$

Now note that

$$
\sqrt{n} \sum_{k=1}^{n} X_k^2(\infty) \geq \sum_{k=1}^{n} |X_k(\infty)|,
$$

which implies that

$$
\left( \max \left\{ \log 24 \sum_{k=1}^{n} |X_k(\infty)|, 0 \right\} \right)^2 \leq \left( \sum_{k=1}^{n} |X_k(\infty)| \right)^2 \leq 576 n \sum_{k=1}^{n} X_k^2(\infty),
$$

where the first inequality holds because $\log x \leq x$ when $x \geq 1$. Recall that $n = N^{\alpha+\xi}$, $\lambda = 1 - y N^{-\alpha}$, and $\sum_{y \neq X(\infty)} R_{X(\infty), y} \leq 2N$. Hence we conclude that when $N$ is sufficiently large,

$$
E \left[ \sum_{k=1}^{n} X_k^2(\infty) \right] \leq \left( \frac{1}{N^{1-5\alpha-6\xi}} \right) E \left[ \sum_{k=1}^{n} X_k^2(\infty) \right] + \frac{1}{N^{1-2\alpha-6\xi}}.
$$

□
Lemma 4.3. Consider the supermarket model under the-power-of-two-choices. The following inequality holds when \( n = N^{\alpha+\xi} \) and \( N \) is sufficiently large,

\[
E \left[ \sum_{k=n+1}^{\infty} X_k^2(\infty) \right] \leq \frac{1}{N^{1-2\alpha}}.
\]

Proof. Note that \( 0 \leq s_k \leq 1 \) for all \( k \), and \( E[S_k(\infty)] \leq \lambda^k \) (equation (5.3) in [26]). Therefore, we obtain

\[
E \left[ \sum_{k=n+1}^{\infty} X_k^2(\infty) \right] \leq E \left[ \sum_{k=n+1}^{\infty} S_k^2(\infty) + (s_k^*)^2 \right] \leq E \left[ \sum_{k=n+1}^{\infty} S_k(\infty) + (s_k^*)^2 \right] \leq \frac{\lambda^n}{1-\lambda},
\]

where the last inequality holds when \( n \) is sufficiently large. When \( n = N^{\alpha+\xi} \), we further have that for sufficiently large \( N \),

\[
\frac{\lambda^n}{1-\lambda} = \gamma N^\alpha (1 - \gamma N^{-\alpha}) N^{\alpha+\xi} \leq \frac{1}{N^{1-2\alpha}}.
\]

\[\square\]

5 CONCLUSION

This paper presented a method based on mean-field models, Stein’s method and the perturbation theory to study the performance of large-scale stochastic systems under a broad range of load conditions. Two examples, the \( M/M/N \) queueing system and the supermarket model under the-power-of-two-choices, were presented to demonstrate the novelty of the proposed method. For the \( M/M/N \) queueing system, we obtained universal bounds with simple and elementary analysis. For the-power-of-two-choices, we established the bounds on the expected queue length per server and the mean-square-error of the mean-field approximation, both of which are new results for the-power-of-two-choices in the heavy-traffic regime.

ACKNOWLEDGEMENT

The author sincerely thanks Jim Dai for the numerous discussions that motivated this work and inspired many results in this paper. The author also thanks R. Srikant for his invaluable comments and feedback.

This work was supported in part by NSF Grants CNS-1262329, ECCS-1547294, ECCS-1609202, and the U.S. Office of Naval Research (ONR Grant No. N00014-15-1-2169).

REFERENCES

We first calculate
\[ \int_0^\infty |x^{(1)}(t)|^2 \, dt. \]

Recall that \( d = \min\{d_1, d_c\}. \) According to C1, \( |x(t, z)| \leq d \) for \( t \geq \tilde{t}_{d,z}. \) From C2 and C6, we have
\[ \int_0^\tilde{t}_{d,z} \left( |x^{(1)}(t)| \right)^2 \, dt \leq (c_1 b)^2 \tilde{t}_{d,z}. \tag{38} \]

According to C3, we have that for \( t \geq \tilde{t}_{d,z}, \)
\[ c_1 |x^{(1)}(t)| \leq V_1(x^{(1)}(t)) \leq e^{-\delta_1 (t - \tilde{t}_{d,z})} V_1(x^{(1)}(\tilde{t}_{d,z})) \leq e^{-\delta_1 (t - \tilde{t}_{d,z})} c_{u1} |x^{(1)}(\tilde{t}_{d,z})|. \]

From C2 and C6, we can further obtain that for \( t \geq \tilde{t}_{d,z}, \)
\[ |x^{(1)}(t)| \leq \frac{c_{u1} c_1 b}{c_1} e^{-\delta_1 (t - \tilde{t}_{d,z})}. \tag{39} \]
which implies
\[
\int_{t_0}^{\infty} \left( |x^{(1)}(t)| \right)^2 dt \leq \left( \frac{c \alpha N_2}{\delta_e} \right)^2 \left( \frac{1}{2\delta_e} \right).
\] (40)

Combining (38) and (40), we obtain
\[
\int_{0}^{\infty} \left( |x^{(1)}(t)| \right)^2 dt \leq (c_1 b)^2 \tau_{d,z} + \left( \frac{c \alpha N_2}{\delta_e} \right)^2 \left( \frac{1}{2\delta_e} \right).
\]

Next, we bound
\[
\int_{0}^{\infty} |e(t)| dt.
\]

According to C4 and C6 and the fact \( e(t) = 0 \), we have that for \( t \leq \tilde{t}_{d,z} = o(N) \),
\[
|e(t)| \leq \frac{c_e}{N^2} t
\]
and
\[
\int_{0}^{\tilde{t}_z} |e(t)| dt \leq \frac{c_e}{N^2} \tilde{t}_{d,z} \frac{1}{N^2}.
\] (41)

Furthermore, from C5 and (39), we have that for \( t \geq \tilde{t}_{d,z} \),
\[
V_e(e(t)) \leq -\delta_e V_e(e(t)) + \frac{c}{N^2} e^{-\delta_e (t - \tilde{t}_{d,z})},
\]
where \( c = c_{er} \left( \frac{c_1 b}{\alpha} \right)^\alpha \). Based on the comparison principle [17], we obtain
\[
V_e(e(t)) \leq V_e(e(\tilde{t}_{d,z})) e^{-\delta_e (t - \tilde{t}_{d,z})} + \frac{c}{N^2} \int_{0}^{\tilde{t}_{d,z}} e^{-\delta_e (t - \tilde{t}_{d,z} - \tau)} - \delta_e \tau d\tau
\]
\[
= V_e(e(\tilde{t}_{d,z})) e^{-\delta_e (t - \tilde{t}_{d,z})} + \frac{c}{N^2} \int_{0}^{\tilde{t}_{d,z}} \frac{1}{\alpha \delta_e - \delta_e} \left( 1 - \exp \left( -\left( \alpha \delta_e - \delta_e \right) (t - \tilde{t}_{d,z}) \right) \right),
\]
which implies that
\[
\int_{\tilde{t}_{d,z}}^{\infty} |e(t)| dt \leq \frac{1}{c_{le}} \int_{\tilde{t}_{d,z}}^{\infty} V_e(e(t)) dt
\]
\[
\leq \frac{1}{c_{le}} e^{-\delta_e (t - \tilde{t}_{d,z})} + \frac{1}{c_{le} N^2} \delta_e \frac{1}{\alpha \delta_e - \delta_e}
\]
\[
\leq \frac{c_e}{c_{le}} \tilde{t}_{d,z} \frac{1}{\delta_e N^2} + \frac{c}{c_{le} e \delta_e N^2}.
\] (42)

Now combining (41) and (42), we obtain
\[
\int_{0}^{\infty} |e(t)| dt \leq \frac{c_e}{2 N^2} \tilde{t}_{d,z} \frac{1}{\delta_e N^2} + \frac{c}{c_{le} e \delta_e N^2}.
\]
Therefore, the lemma holds by choosing
\[ \kappa = \max \left\{ (c_1 b)^2, \left( \frac{c_1 c_2 b}{c_1} \right)^2, \frac{c_6 - \frac{c}{2}}{c_6 - \frac{c}{2} \cdot c_1} \right\}, \]
which is independent of \( N \) according to C7.

B ANALYSIS OF THE M/M/N QUEUEING SYSTEM

We first prove (25). Recall that for \( d(x, x^*) = x^2 \), we have
\[ \nabla g(x) = \begin{cases} \frac{-x}{1-x} & -\lambda \leq x \leq 1 - \lambda \\ -\frac{x^2}{1-x} & x > 1 - \lambda \end{cases}, \]
and
\[ g(x) = \begin{cases} \frac{-x^3}{2} & -\lambda \leq x \leq 1 - \lambda \\ -\frac{x^3}{6(1-\lambda)} & x > 1 - \lambda \end{cases}. \]
Note that \( \nabla g(x) \) is differentiable at all \( x \) except \( x = 1 - \lambda \). We next calculate \( \Gamma(x, y) \) case by case.

- Case 1: When \( x \leq 1 - \lambda + \frac{1}{N} \), based on Taylor’s theorem and the fact that \( \nabla g(x) \) is a decreasing function
  \[ -\left( g \left( x + \frac{1}{N} \right) - g(x) - \frac{1}{N} \nabla g(x) \right) \leq \frac{1}{N} \left| \nabla g \left( x + \frac{1}{N} \right) - \nabla g(x) \right| \leq \frac{1}{N} \left( \frac{2(1-\lambda + \frac{2}{N})}{(1-\lambda)} \right) \leq \frac{3}{N^2}, \]
  where the last inequality holds because \( N(1-\lambda) \geq 4 \). Similarly,
  \[ -\left( g \left( x - \frac{1}{N} \right) - g(x) + \frac{1}{N} \nabla g(x) \right) \leq \frac{2.5}{N^2}. \]

- Case 2: When \( 1 - \lambda + \frac{1}{N} \leq x \leq \frac{8}{N(1-\lambda)} \), based on Taylor’s theorem, we have
  \[ -\left( g \left( x + \frac{1}{N} \right) - g(x) - \frac{1}{N} \nabla g(x) \right) \leq \frac{2}{N} + \frac{1}{N} \frac{1}{1-\lambda} \frac{2}{N^2} \leq \frac{16}{N^3(1-\lambda)^2} + \frac{2}{N^3(1-\lambda)^2}, \]
  where the last inequality holds because \( x \leq \frac{8}{N(1-\lambda)} \). Similarly,
  \[ -\left( g \left( x - \frac{1}{N} \right) - g(x) + \frac{1}{N} \nabla g(x) \right) \leq \frac{16}{N^3(1-\lambda)^2}. \]
  \textit{Note that this case does not exist if} \( 1 - \lambda + \frac{1}{N} > \frac{8}{N(1-\lambda)} \).

- Case 3: When \( x \geq \max \left\{ \frac{8}{N(1-\lambda)}, 1 - \lambda + \frac{1}{N} \right\} \), based on Taylor’s theorem, we have
  \[ -\left( g \left( x + \frac{1}{N} \right) - g(x) - \frac{1}{N} \nabla g(x) \right) \leq 2 \frac{x + \frac{1}{N}}{1 - \lambda} \frac{1}{N^2} \leq a \frac{2x}{1 - \lambda} \frac{1}{N^2} + \frac{2}{1 - \lambda} \frac{1}{N^2} \leq b \frac{x^2}{4N} + \frac{1}{2N^2}, \]
  where inequality \( a \) holds because \( x \geq \frac{8}{N(1-\lambda)} \) and inequality \( b \) holds because \( N - \lambda N \geq 4 \).
  Similarly, we have
  \[ -\left( g \left( x - \frac{1}{N} \right) - g(x) + \frac{1}{N} \nabla g(x) \right) \leq \frac{x^2}{4N}. \]
Let $G$ denote the generator of the CTMC. Let $S$ denote the range of $x$. Then for any $x, y \in S$, we have
\[
R_{x,y} = \begin{cases} 
\lambda N, & \text{if } y = x + \frac{1}{N} \\
N \min \{1, x + \lambda\}, & \text{if } y = x - \frac{1}{N} \\
-\lambda N - N \min \{1, x + \lambda\}, & \text{if } y = x \\
0, & \text{otherwise}
\end{cases}
\]

Then,
\[
G g(x) = \sum_{y \in S} R_{x,y} (g(y) - g(x)) = \lambda N \left( g\left(x + \frac{1}{N}\right) - g(x) \right) + N \min \{1, x + \lambda\} \left( g\left(x - \frac{1}{N}\right) - g(x) \right).
\]

By combining the cases above and defining
\[
X(\infty) = \frac{Q(\infty)}{N} - \lambda,
\]
we obtain
\[
E \left[ X^2(\infty) \right] = -E \left[ N \lambda \left( g\left(x + \frac{1}{N}\right) - g(x) - \frac{1}{N} \nabla g(x) \right) + N \min \{1, x + \lambda\} \left( g\left(x - \frac{1}{N}\right) - g(x) + \frac{1}{N} \nabla g(x) \right) \right]
\]
\[
\leq \frac{3(1 + \lambda)}{N} E \left[ X(\infty) \leq 1 - \lambda + \frac{1}{N} \right] \\
+ \frac{18(1 + \lambda)}{N^2(1 - \lambda)^2} E \left[ 1 - \lambda + \frac{1}{N} \leq X(\infty) \leq \frac{8}{N(1 - \lambda)} \right] \\
+ \frac{\lambda + 1}{4} E \left[ X^2(\infty) \mathbb{1}_{X(\infty) \geq \max \left\{ \frac{8}{N(1 - \lambda)}, 1 - \lambda + \frac{1}{N} \right\}} \right] \\
+ \frac{\lambda + 1}{2N} E \left[ X(\infty) \geq \max \left\{ \frac{8}{N(1 - \lambda)}, 1 - \lambda + \frac{1}{N} \right\} \right] \\
\leq \frac{3(1 + \lambda)}{N} + \frac{18(1 + \lambda)}{N^2(1 - \lambda)^2} + \frac{1}{2} E \left[ X^2(\infty) \right],
\]
where in the last inequality, we combined (43) and (44). By moving the last term to the left-hand-side of the equation, we further have
\[
E \left[ X^2(\infty) \right] = E \left[ \left( \frac{Q(\infty)}{N} - \lambda \right)^2 \right] \leq \frac{6(\lambda + 1)}{N} + \frac{36(1 + \lambda)}{N^2(1 - \lambda)^2}.
\]

We now consider $d(x, x^*) = x$. Recall that
\[
\nabla g(x) = \begin{cases} 
-1, & -\frac{1}{2} \leq x \leq 1 - \lambda \\
-\frac{1}{x - 1 - \lambda}, & x > 1 - \lambda
\end{cases}
\]
and
\[
g(x) = \begin{cases} 
-\frac{x}{1 - \lambda - \frac{x^2}{2(1 - \lambda)}}, & -\frac{1}{2} \leq x \leq 1 - \lambda \\
-\frac{1}{x - 1 - \lambda}, & x > 1 - \lambda
\end{cases}
\]
Based on Taylor’s theorem and the fact that $\nabla g(x)$ is a decreasing function
\[
- \left( g\left(x + \frac{1}{N}\right) - g(x) - \frac{1}{N} \nabla g(x) \right) \leq \frac{1}{N} \left| g\left(x + \frac{1}{N}\right) - \nabla g(x) \right| \leq \frac{1}{N^2(1 - \lambda)}
\]
Therefore, we have
\[- \left( g \left( x - \frac{1}{N} \right) - g(x) + \frac{1}{N} \nabla g(x) \right) = \frac{1}{N^2(1-\lambda)}.\]

Therefore, we have
\[
E \left[ \frac{Q(\infty)}{N} - \lambda \right] = - E \left[ N \lambda \left( g(x + \frac{1}{N}) - g(x) - \frac{1}{N} \nabla g(x) \right) + N \min\{1, x + \lambda\} \left( g(x - \frac{1}{N}) - g(x) + \frac{1}{N} \nabla g(x) \right) \right]
\leq \frac{(1 + \lambda)}{N - \lambda N}.
\]

C APPLICATION OF LEMMA 2.2 TO THE-POWER-OF-TWO-CHOICES

In this section, we establish the gradient bound for the-power-of-two-choices by applying Lemma 2.2 to the \(n\)-dimensional truncated system defined in (36). In this section, all vectors are \(n\)-dimensional instead of infinite-dimensional.

**Lemma C.1.** Consider the dynamical system defined in (34). Given that the initial condition satisfies \(1 = \tilde{s}_0(0) \geq \tilde{s}_1(0) \geq \tilde{s}_2(0) \geq \cdots \geq \tilde{s}_n(0) \geq 0\), we have \(0 \leq \tilde{s}_k(t) \leq 1\) for any \(t \geq 0\) and \(0 \leq k \leq n\).

**Proof.** Note that \(\max_{0 \leq k \leq n} \tilde{s}_k(0) \leq 1\). Furthermore, if
\[\tilde{s}_j(t) = \max_{0 \leq k \leq n} \tilde{s}_k(t) \text{ and } \sup_{0 \leq t \leq t} \max_{0 \leq k \leq n} \tilde{s}_k(t) \leq 1\]
then \(\tilde{s}_j(t) \leq 0\). Therefore, we conclude that for any \(t \geq 0\),
\[\max_{0 \leq k \leq n} \tilde{s}_k(t) \leq 1.\]

Furthermore, when \(\tilde{s}_j(t) = \min_k \tilde{s}_k(t) = 0\),
\[\tilde{s}_j(t) = \lambda \tilde{s}_{j-1}^2(t) + \tilde{s}_{j+1}(t) \geq 0.\]
So the lemma holds. \(\square\)

**Lemma C.2.** For the mean-field model defined in (36), \(-s_k^{\ast} \leq x_k(t) \leq 1 - s_k^{\ast} \text{ for } 1 \leq k \leq n \text{ and } t \geq 0\).

**Proof.** According to Lemma C.1, \(0 \leq x_k(t) + s_k^{\ast} = s_k(t) \leq 1\) for any \(t \geq 0\), so the lemma holds. \(\square\)

Define
\[\tilde{k} = (\alpha + \xi) \log_2 N.\]  

(45)

**Lemma C.3.** According to the definition of \(\tilde{k}\), for sufficiently large \(N\), we have for any \(k \geq \tilde{k}\)
\[\lambda \left( s_k^{\ast} + 1 \right) = \lambda \left( \tilde{k}^{2k} + 1 \right) \leq \sqrt{\lambda}.\]

**Proof.** Note that
\[\lambda^{2k} \leq \lambda^{2k} = \lambda^{N^{\alpha+\xi}} = (1 - \gamma N^{-\alpha})^{N^{\alpha+\xi}},\]
so
\[\log \lambda^{2k} \leq N^{\alpha+\xi} \log (1 - \gamma N^{-\alpha}) \leq (a) -\gamma N^{\xi} = -\Theta \left( N^{\xi} \right),\]
where inequality (a) is a result of the Taylor expansion. Furthermore,
\[
\log \left( \sqrt{\lambda} - \lambda \right) = \log \sqrt{\lambda} + \log(1 - \sqrt{\lambda}) = -\Theta(\alpha \log N).
\]
Therefore, for sufficiently large \( N \), we have
\[
\lambda^{2^k} \leq \sqrt{\lambda} - \lambda,
\]
and the lemma holds. \( \square \)

Now define a sequence \( \{w_k\} \) such that for some \( \epsilon > 0 \)
\[
w_0 = 0, \\
w_1 = 1, \\
w_k = \left( 1 + \frac{1}{2} \sum_{j=1}^{k} \frac{1}{(2\lambda + \epsilon)^{j-1}} \right), \quad 2 \leq k \leq \tilde{k}, \\
w_k = w_{\tilde{k}} + \frac{k - \tilde{k}}{2(2\lambda + \epsilon)^{\tilde{k}}}, \quad \tilde{k} < k \leq n.
\]
Given \( \xi \), we choose a constant \( \epsilon \) independent of \( N \) such that
\[
\min \left\{ 0.5, 2^{\frac{\alpha \xi}{1 + \epsilon}} - 2\lambda \right\} > \epsilon > 2 - 2\lambda.
\]
Such an \( \epsilon \) exists when \( \lambda > 0.75 \). We further define
\[
\delta_0 = \frac{1 - \sqrt{\lambda}}{6(2\lambda + \epsilon)^{\tilde{k}}},
\]

**Lemma C.4.** For any \( 1 \leq k \leq n \), \( 1 \leq w_k \leq 3 \). When \( N \) is sufficiently large,
\[
\delta_0 \geq \frac{1}{N^{2\alpha + 2\xi}}.
\]

**Proof.** To prove the first result, we note that for \( k \leq \tilde{k} \),
\[
w_k \leq w_{\tilde{k}} \leq 1 + \frac{1}{2} \frac{1}{1 - \frac{1}{2\lambda + \epsilon}} \leq 2,
\]
where the last inequality holds because \( 2\lambda + \epsilon > 2 \). For \( k > \tilde{k} \),
\[
w_k \leq w_{\tilde{k}} + \frac{n}{2(2\lambda + \epsilon)^{\tilde{k}}} \leq 2 + 0.5 < 3,
\]
where the second inequality holds because
\[
(2\lambda + \epsilon)^{\tilde{k}} > 2^k = N^{\alpha + \xi} = n.
\]
To prove the second inequality, note that
\[
\log \delta_0 = \log(1 - \sqrt{\lambda}) - \log 6 - \tilde{k} \log(2\lambda + \epsilon) = -\alpha \log N - (\alpha + \xi) \log(2\lambda + \epsilon) \log N + O(1).
\]
According to the upper bound on \( \epsilon \) in (46), we have
\[
\log \delta_0 \geq -2(\alpha + \xi) \log N.
\]
\( \square \)
In the following analysis, we define $V(x) = \sum_{k=1}^{n} w_k |x_k(t)|$.

**Lemma C.5 (Proof of C1).** For the dynamical system defined in (36), we have

$$\dot{V}(x) \leq -\delta_0 V(x),$$

which implies that

$$|x(t)| \leq V(x(t)) \leq 3|x(0)|e^{-\delta_0 t}.$$ 

**Proof.** The proof follows the idea of Theorem 3.6 in [18]. Note that $V(t)$ is Lipschitz continuous because $|x_k(t)|$ is Lipschitz continuous for all $k$. We now consider regular points [10] such that $\frac{d|x_k(t)|}{dt}$ exists for all $k$ at $t$. Define $\dot{V}(t) = \sum_{k=1}^{n} W_k(t)$ such that $W_k(t)$ includes all the terms involving $x_k(t)$. The lemma is proved by showing that

$$W_k(t) \leq -\delta_0 w_k |x_k(t)|.$$  \hspace{1cm} (48)

When $x_k(t) > 0$, we have

$$W_k(t) \leq w_{k+1} \lambda |x_k| (x_k + 2s^*_k) - w_k \lambda |x_k| (x_k + 2s^*_k) - w_k |x_k| + w_{k-1} |x_k|$$

and the last term comes from\(\frac{d|x_k(t)|}{dt}\). The same inequality holds when $x_k(t) < 0$. So (48) holds if

$$w_{k+1} \lambda |x_k| (x_k + 2s^*_k) - w_k \lambda |x_k| (x_k + 2s^*_k) - w_k |x_k| + w_{k-1} |x_k| \leq -\delta_0 w_k |x_k|.$$ 

in other words, if

$$w_{k+1} - w_k \leq \frac{(1 - \delta_0)w_k - w_{k-1}}{\lambda (x_k + 2s^*_k)}. \hspace{1cm} (49)$$

We now prove (49) by considering the following three cases.

When $1 \leq k \leq \bar{k}$, we have

$$w_{k+1} - w_k = \frac{1}{2(2\lambda + \epsilon)^k} \frac{(1 - \delta_0)w_k - w_{k-1}}{\lambda (x_k + 2s^*_k)} \geq \frac{w_k - w_{k-1} - \delta_0 w_k}{\lambda (1 + \lambda)} \geq \frac{1}{2(2\lambda + \epsilon)^k} - \frac{\delta_0 w_k}{2\lambda}.$$ 

So inequality (49) holds if

$$2\lambda \leq 2\lambda + \epsilon - 2\delta_0 w_k (2\lambda + \epsilon)^k,$$

which can be established by proving

$$2\delta_0 w_k (2\lambda + \epsilon)^k \leq \epsilon.$$ 

It can be verified that the inequality holds according to the definition of $\delta_0$ and the fact that $\epsilon > 2 - 2\lambda$.

When $n \geq k \geq \bar{k} + 1$, according to Lemma C.3,

$$\lambda (x_k + 2s^*_k) \leq \lambda (1 + s^*_k) \leq \sqrt{\lambda}.$$
Therefore, we have

\[ w_{k+1} - w_k = \frac{1}{2(2\lambda + \epsilon)k} \]

\[ \frac{(1 - \delta_0)w_k - w_{k-1}}{\lambda(x_k + 2s^*_k)} \geq \frac{w_k - w_{k-1} - \delta_0 w_k}{\sqrt{\lambda}} = \frac{1}{2(2\lambda + \epsilon)k} - \delta_0 w_k. \]

So inequality (49) holds if

\[ \sqrt{\lambda} \leq 1 - 2\delta_0 w_k (2\lambda + \epsilon)^k, \]

in other words, if

\[ w_k \leq \frac{(1 - \sqrt{\lambda})}{2\delta_0 (2\lambda + \epsilon)^k}, \]

which holds because \( w_k \leq 3 \) according to Lemma C.4 and \( \frac{(1 - \sqrt{\lambda})}{2\delta_0 (2\lambda + \epsilon)^k} = 3 \) according to the definition of \( \delta_0 \).

From the discussion above, we conclude that

\[ \hat{V}(t) \leq -\sum_{k=1}^{n} \delta w_k |x_k(t)| = -\delta_0 V(t), \]

so the lemma holds.

We further have the following first-order system for the truncated system (36):

\[ x^{(1)}_k = g_k(x^{(1)}) = 2\lambda (x_{k-1} + s^*_{k-1}) x^{(1)}_{k-1} - 2\lambda (x_k + s^*_k) x^{(1)}_k - x^{(1)}_k + x^{(1)}_{k+1}, \]

where \( 0 < k < n \) and we define \( x^{(1)}_0 = x^{(1)}_n = 0 \).

**Lemma C.6 (Proof of C2).** Under the dynamical system defined by (50),

\[ |x^{(1)}(t)| \leq |x^{(1)}(0)|. \]

**Proof.** First recall that \( x_k(t) + s^*_k \geq 0 \) for any \( t \geq 0 \) and \( k \) according to Lemma C.2. Define

\[ V(t) = \sum_{k=1}^{n} |x^{(1)}_k(t)|. \]

Note that

\[ \frac{d|x^{(1)}_k(t)|}{dt} \leq 2\lambda (x_{k-1} + s^*_{k-1}) |x^{(1)}_{k-1}| - 2\lambda (x_k + s^*_k) |x^{(1)}_k| - |x^{(1)}_k| + |x^{(1)}_{k+1}|, \]

so

\[ \hat{V}(t) \leq -2\lambda (x_n + s^*_n) |x^{(1)}_n| - |x^{(1)}_1| \leq 0, \]

and the lemma holds.

Define

\[ \tilde{\delta} = \frac{\epsilon}{6(2\lambda + \epsilon)^k}. \]

Similar to Lemma C.4, the following lemma also holds.
Lemma C.7. When $N$ is sufficiently large,

$$\hat{\delta} \geq \frac{1}{N^{a+2\xi}}.$$ 

Lemma C.8 (Proof of C3). For sufficiently large $N$, we have

$$|x^{(i)}| \leq V(x^{(i)}) \leq 3|x^{(i)}|$$

$$\dot{V}(x^{(i)}(t)) \leq -\hat{\delta}V(x^{(i)})(t), \quad \text{if } |x(t)| \leq \frac{1}{8}.$$ 

Proof. The first inequality holds due to the definition of $w_k$. Next we define the following Lyapunov function

$$V(x^{(i)}) = \sum_{k=1}^{n} w_k |x_k^{(i)}|.$$ 

Following the proof of Lemma C.5, we obtain that

$$\dot{V}(x^{(i)}) \leq \sum_{k=1}^{n} - (2\lambda w_k (x_k + s_k^*) + w_k - 2\lambda w_{k+1} (x_k + s_k^*) - w_{k-1}) |x_k^{(i)}|.$$ 

So the lemma holds by proving

$$- (2\lambda w_k (x_k + s_k^*) + w_k - 2\lambda w_{k+1} (x_k + s_k^*) - w_{k-1}) \leq -\hat{\delta}w_k,$$

i.e. by proving

$$w_{k+1} - w_k \leq \frac{w_k - w_{k-1} - \hat{\delta}w_k}{2\lambda (x_k + s_k^*)}.$$ 

We now prove (51) by considering the following cases.

When $1 \leq k \leq \hat{k}$, we have

$$w_{k+1} - w_k = \frac{1}{2(2\lambda + \epsilon)k},$$

$$\frac{w_k - w_{k-1} - \hat{\delta}w_k}{2\lambda (x_k + s_k^*)} \geq \frac{1}{2(2\lambda + \epsilon)^{k-1}} - \frac{\hat{\delta}w_k}{2\lambda}.$$ 

So inequality (49) holds if

$$2\lambda \leq 2\lambda + \epsilon - \hat{\delta}2w_k(2\lambda + \epsilon)^k,$$

which can be established by proving

$$\hat{\delta}2w_k(2\lambda + \epsilon)^k \leq \epsilon,$$

which holds according to the definition of $\hat{\delta}$ and the fact $1 \leq w_k \leq 3$.

When $n \geq k \geq \hat{k} + 1$, according to the definition of $\hat{k}$,

$$s_k^* \leq s_k^* = \lambda^{N^a + \xi - 1}.$$ 

If $\lambda \geq \frac{64}{87}$, then

$$s_k^* \leq \frac{1}{\sqrt{\lambda}} - 1 \leq \frac{1}{8}.$$
Stein’s Method for Mean-Field Approximations in Light and Heavy Traffic Regimes

according to Lemma C.3; otherwise, we can find a sufficiently large $N$ such that

$$s_k^* \leq \lambda^{N^{x+\epsilon}} - 1 \leq \frac{1}{8}.$$  

Now given $|x_k| \leq |x| \leq \frac{1}{8}$, we have

$$w_{k+1} - w_k = \frac{1}{2(2\lambda + \epsilon)^k},$$

$$\frac{(1 - \delta)w_k - w_{k-1}}{2\lambda(|x_k| + s_k^*)} \geq \frac{w_k - w_{k-1} - \delta w_k}{\frac{1}{2}} = \frac{1}{\frac{2(2\lambda + \epsilon)^k}{\delta}} - \frac{\delta w_k}{\frac{1}{2}}.$$  

So inequality (51) holds if

$$\frac{1}{2} \lambda \leq 1 - \delta 2w_k (2\lambda + \epsilon)^k.$$  

Note that according to the definition of $\delta$,

$$\delta 2w_k (2\lambda + \epsilon)^k \leq 6\delta (2\lambda + \epsilon)^k = \epsilon.$$  

So the inequality above holds for $\epsilon \leq 0.5$. From the analysis above, we conclude $V(t) \leq -\delta V(t)$. \hfill \Box

**Lemma C.9 (Proof of C4).** Given $|e(t)| \leq \frac{1}{N}$, we have

$$\frac{d|e(t)|}{dt} \leq (4\lambda + 4) \frac{1}{N^2}.$$  

Proof. We first have for $1 < k < n$,

$$\dot{e}_k(t) = \frac{1}{N} x_k(t) + e(t) - \frac{1}{N} \sum_{j=1}^{n} \frac{\partial f_k}{\partial x_j}(x(t)) x_j(t)$$

$$= \lambda \left( x_{k-1}(t) + \frac{1}{N} \delta^{(1)}_{k-1} \right)^2 + 2 \delta^{(k-1)} - \left( x_{k-1}(t) + \frac{1}{N} \delta^{(1)}_{k-1} \right)$$

$$+ \frac{2}{N} \left( x_{k-1} + \delta^{(k-1)} \right)^2 + \frac{1}{N} \delta^{(1)}_{k-1} - \left( x_{k-1} + \frac{1}{N} \delta^{(1)}_{k-1} \right)$$

$$= \lambda \left( e^2_k + 2 \left( x_{k-1} + \delta^{(k-1)} \right) - e^2_k - 2 \left( x_{k-1} + \frac{1}{N} \delta^{(1)}_{k-1} \right) \right)$$

$$+ \frac{1}{N^2} \left( x^{(1)}_{k-1} - \left( x^{(1)}_k \right)^2 \right) \quad (52)$$

where the last equality holds according to the definition of $g_k(\cdot)$ in (50). The same equation holds for $k = 1$ and $k = n$. From the inequality above and following the proof of Lemma C.6, we can further obtain

$$\frac{d|e(t)|}{dt} \leq \sum_{k=1}^{n} 2 \lambda \left( e^2_k + 2 \frac{1}{N} \left| x^{(1)}_{k-1} \right| \right) + 2 \lambda \frac{1}{N^2} \left( x^{(1)}_k \right)^2$$

$$\leq 2 \lambda |e(t)| + \frac{4}{N} |e(t)| + \frac{2\lambda}{N^2}.$$ 

Given $|e(t)| \leq \frac{1}{N}$, we conclude

$$\frac{d|e(t)|}{dt} \leq (4\lambda + 4) \frac{1}{N^2}.$$ 

\[ \square \]

Lemma C.10 (Proof of C5). When $|x(t)| \leq 1/8$, we have

$$V(e(t)) \leq -\delta V(e(t)) + \frac{6\lambda}{N^2} \left( |x^{(1)}| \right)^2.$$
Proof. Now assume $|x(t)| \leq 1/8$. Recall (52)
\[
\hat{e}_k(t) = \lambda \left( 2 \left( x_{k-1} + s_{k-1} \right) + \frac{e_{k-1}}{2} + \frac{1}{N} x_{k-1}^{(1)} \right) e_{k-1} - 2 \left( x_k + s_k \right) + \frac{e_k}{2} + \frac{1}{N} x_k^{(1)} e_k
\]
\[
- (e_k - e_{k+1}) + \left( \frac{1}{N^2} \left( x_k^{(1)} \right)^2 - \frac{1}{N} \frac{1}{2} \left( x_k \right)^2 \right).
\]
Again consider
\[
\dot{V}(e(t)) = \sum_{k=1}^n W_k(t) + W(t),
\]
where $W_k(t)$ includes all the terms involving $e_k(t)$ and $W(t)$ includes all the remaining terms. Note that both $|e_k(t)|$ and $\frac{1}{N} |x_k^{(1)}|$ can be made arbitrarily small by choosing sufficiently large $N$. Therefore, following the analysis of Lemma C.8, we have
\[
\sum_{k=1}^n W_k(t) \leq - \tilde{\delta} V(t).
\]
Since $1 \leq w_k \leq 3$ for all $k \geq 1$,
\[
\dot{V}(e(t)) \leq - \tilde{\delta} V(e(t)) + \max_k W_k \frac{2\lambda}{N^2} \|x^{(1)}\|^2 \leq - \tilde{\delta} V(e(t)) + \frac{6\lambda}{N^2} \left( |x^{(1)}| \right)^2.
\]
\[
\Box
\]
The analysis above verifies condition C1-C5 in Lemma 2.2 with $c_1 = c_{e1} = c_{ue} = 1$, $c_{a1} = c_{ue} = 3$, $d_1 = d_e = \frac{1}{8}$, $\delta_1 = \delta_e = \tilde{\delta}$, and $c_e = 4\lambda + 4$. Furthermore, $b = 1$ according to (29) and $\tilde{t}_{a, z} = \frac{1}{\delta_0} \max \{0, \log 24 |x(0)|\}$ according to Lemma C.5. Therefore, both C6 and C7 hold. Hence, we conclude that there exists a constant $\kappa$ such that when $N$ is sufficiently large, the following two inequalities hold
\[
\int_0^\infty |x^{(1)}(t)|^2 \, dt \leq \kappa \left( \max \{0, \log 24 |x(0)|\} N^{2\alpha + 2\xi} + N^{\alpha + 2\xi} \right)
\]
\[
\int_0^\infty |e(t)| \, dt \leq \kappa \left( \max \{0, \log 24 |x(0)|\} N^{3\alpha + 4\xi} + N^{2\alpha + 4\xi} + (\max \{0, \log 24 |x(0)|\})^2 N^{4\alpha + 4\xi} \right) \frac{1}{N^2}.
\]