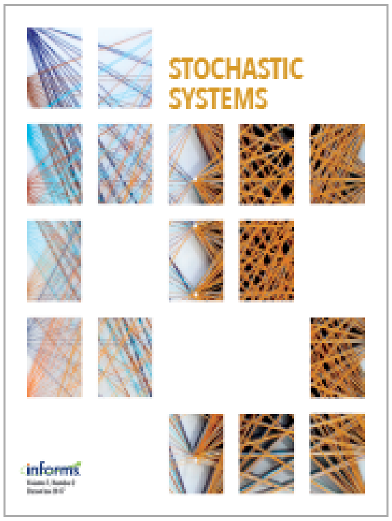


Stochastic Systems

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

On the Approximation Error of Mean-Field Models

Lei Ying

To cite this article:
Lei Ying (2018) On the Approximation Error of Mean-Field Models. Stochastic Systems 8(2):126-142. https://doi.org/10.1287/stsy.2018.0012

informs
http://pubsonline.informs.org/journal/stsy/

STOCHASTIC SYSTEMS
Vol. 8, No. 2, June 2018, pp. 126–142
ISSN 1946-5238 (print), ISSN 1946-5238 (online)

On the Approximation Error of Mean-Field Models

Lei Ying[a]

[a] ECEE, Arizona State University, Tempe, Arizona 85287

Contact: lei.ying.2@asu.edu (LY)

Received: May 1, 2016
Revised: August 28, 2017
Accepted: January 27, 2018
Published Online in Articles in Advance: May 11, 2018

MSC2010 Subject Classification: 60J27, 60B10

https://doi.org/10.1287/stsy.2018.0012

Abstract. Mean-field models have been used to study large-scale and complex stochastic systems, such as large-scale data centers and dense wireless networks, using simple deterministic models (dynamical systems). This paper analyzes the approximation error of mean-field models for continuous-time Markov chains (CTMCs) and focuses on mean-field models that are represented as finite-dimensional dynamical systems with a unique equilibrium point. By applying Stein's method and the perturbation theory, the paper shows that under some mild conditions, if the mean-field model is *globally asymptotically stable* and *locally exponentially stable*, the *mean-square difference* between the stationary distribution of the stochastic system of size M (i.e., with M nodes/particles) and the equilibrium point of the corresponding mean-field system is $O(1/M)$. The result of this paper establishes a general theorem for establishing the convergence and the approximation error (i.e., the rate of convergence) of a large class of CTMCs to their mean-field limits by mainly looking into the stability of the mean-field model, which is a deterministic system and is often easier to analyze than the CTMCs. Two applications of mean-field models in data center networks are presented to demonstrate the novelty of our results.

Funding: This work was supported in part by the National Science Foundationi [Grants ECCS-1255425, ECCS-1547294, ECCS-1739344] and the U.S. Office of Naval Research [ONR Grant N00014-15-1-2169].

Keywords: mean-field models • Stein's method • approximation error • perturbation theory

1. Introduction

Mean-field analysis is an analytical method to study large-scale and complex stochastic systems using simple deterministic models. The idea of the mean-field analysis is to assume the states of nodes in a large-scale system are independently and identically distributed (i.i.d.). Based on this i.i.d. assumption, in a large-scale system, the interaction of a node to the rest of the system can be replaced with an "average" interaction, and the evolution of the system can then be modeled as a deterministic dynamical system, called a *mean-field model*. Then, the macroscopic behaviors of the stochastic system can be approximated using the mean-field model, e.g., the stationary distribution of the stochastic system may be approximated using the equilibrium point of the mean-field model.

The mean-field analysis has important applications in various areas, including statistical physics, epidemiology, communication networks, queueing theory, and game theory (e.g., Kadanoff 2009, Bailey 1975, Baccelli et al. 1992, Vvedenskaya et al. 1996, Mitzenmacher 1996, Lasry and Lions 2007, Chaintreau et al. 2009, Bordenave et al. 2010, Adlakha and Johari 2013, Manjrekar et al. 2014, Cecchi et al. 2015). In particular, over the last few years, it has emerged as a powerful method for analyzing large-scale cloud computing systems and data center networks. For example, in Mitzenmacher (1996) and Vvedenskaya et al. (1996), the mean-field analysis has been used to show that routing each incoming task to the shorter of two randomly sampled servers can significantly reduce queueing delays, a phenomenon called *the power of two choices* (*Po*2). The result has been extended to heavy-tailed service-time distributions (Bramson et al. 2012), and to heterogeneous servers (Mukhopadhyay and Mazumdar 2013). In Tsitsiklis and Xu (2012), a mean-field model has been used to quantify the significant benefit of resource pooling. Stolyar (2015a) established the asymptotic optimality of the join-idle-queue (JIQ), proposed in Lu et al. (2011), using the mean-field model. In Ying et al. (2015), a novel randomized load balancing algorithm, named Batch-Filling, has been developed for cloud computing systems with batch arrivals. The algorithm achieves similar delay performance as the power of two choices with a sampling ratio slightly larger than one (i.e., it only samples slightly more than one server on average for each incoming task). In Xie et al. (2015) and Mukhopadhyay et al. (2015), the mean-field analysis has been used to study the loss model, which is related to the virtual machine placement problem in data center networks. In these applications, the systems under consideration are modeled as CTMCs, and the solutions (equilibrium points) of the corresponding mean-field models are then used to approximate the stationary distributions of the CTMCs.

To justify the mean-field analysis, a critical step is to prove that the stationary distribution of the CTMC indeed converges to the equilibrium point of the mean-field model as the size of the system increases. Consider a family of CTMCs. The Mth CTMC is an M-dimensional continuous-time Markov chain $\mathbf{W}^{(M)} \in \mathcal{U}^M$, where the superscript M is the number of nodes (or called particles) in the system and $\mathcal{U}^M \subseteq \mathbf{R}^M$ is the state space of the CTMC. We assume $\mathcal{U}$ is a finite state space and the CTMC is irreducible. Without loss of generality, let $\mathcal{U} = \{1, \ldots, n\}$. We further define

$$X_i^{(M)}(t) = \frac{1}{M}\sum_{m=1}^{M} \mathbf{1}_{W_m^{(M)}(t)=i}$$

where $\mathbf{1}$ is the indicator function, so $X_i^{(M)}(t)$ is the fraction of nodes in state i at time t, which itself is a random variable. This paper focuses on the case such that $\mathbf{X}^{(M)} = \{\mathbf{X}^{(M)}(t), t \geq 0\}$ is also an (n-dimensional) CTMC, i.e., the CTMC is a population process (Kurtz 1971, 1981). We remark that many applications of the mean-field analysis such as those in queueing networks and epidemiology are for population processes.

Now let $\mathbf{X}^{(M)}(\infty)$ denote the steady-state of the Mth CTMC. Furthermore, let $\mathbf{x}(t)$ denote the solution of an associated mean-field model and $\mathbf{x}^*$ denote its equilibrium point. Existing approaches for proving the convergence of $\mathbf{X}^{(M)}(\infty)$ to $\mathbf{x}^*$ often involve the following three components.

(1) The first component is to show the convergence of the CTMCs to the trajectory of the mean-field model over any finite time interval $[0, t]$, i.e.,

$$\lim_{M\to\infty} \sup_{0\leq s\leq t} d(\mathbf{X}^{(M)}(s), \mathbf{x}(s)) = 0, \tag{1}$$

where $d(\cdot,\cdot)$ is some measure of distance. This can be proved using different techniques including Kurtz's theorem (Kurtz 1971, 1981; Mitzenmacher 1996; Ying et al. 2015), propagation of chaos (Sznitman 1991, Anantharam and Benchekroun 1993, Bramson et al. 2012), or the convergence of the transition semigroup of CTMCs (Vvedenskaya et al. 1996, Mukhopadhyay and Mazumdar 2013).

(2) The second component is to prove the asymptotic stability of the mean-field model, i.e.,

$$\lim_{t\to\infty} \mathbf{x}(t) = \mathbf{x}^*.$$

Lyapunov theorem or LaSalle invariance principle can often be used for proving the stability.

(3) After establishing the previous two results, we obtained

$$\lim_{t\to\infty}\lim_{M\to\infty} \mathbf{x}^{(M)}(t) = \lim_{t\to\infty} \mathbf{x}(t) = \mathbf{x}^*.$$

The convergence of the stationary distributions can then be concluded if we can prove the interchange of the limits, i.e.,

$$\lim_{M\to\infty} \mathbf{X}^{(M)}(\infty) = \lim_{M\to\infty}\lim_{t\to\infty} \mathbf{X}^{(M)}(t) =_{(a)} \lim_{t\to\infty}\lim_{M\to\infty} \mathbf{X}^{(M)}(t) = \mathbf{x}^*,$$

where step (a) is called the interchange of the limits.

Since these approaches are all based on the interchange of the limits and use the finite-time convergence (equality (1)) as the stepping stone, they are *indirect* methods of proving

$$\lim_{M\to\infty} \mathbf{X}^{(M)}(\infty) = \mathbf{x}^*.$$

Because of this reason, these approaches can only establish the convergence of mean-field models and the asymptotic behaviors of the systems (i.e., for $M = \infty$). The approximation error (or called the rate of convergence) of mean-field models for finite-size systems (e.g., $E[\|\mathbf{X}^{(M)}(\infty) - \mathbf{x}^*\|]$ for a fixed M) is difficult to obtain using these indirect methods.

This paper tackles this fundamental problem and directly studies the approximation error of a large class of mean-field models using Stein's method (Stein 1972, 1986; Barbour and Chen 2005), which is a method to bound the distance of two probability distributions. Our use of Stein's method for the rate of convergence was inspired by the work by Braverman and Dai (2017), in which they developed a modular framework with three components for steady-state diffusion approximations and established the rate of convergence to diffusion models for $M/Ph/n+M$ queuing systems. Please see Braverman et al. (2016) for an introduction of Stein's method for steady-state diffusion approximation and its connection to classical applications of Stein's method. The results in this paper also share similar spirit with the work by Gurvich et al. (2014), which establishes the rate of convergence of diffusion models for steady-state approximations for exponentially ergodic Markovian queues. This paper differs from those works in that it considers mean-field approximations instead of diffusion approximations. Furthermore, while not naming it Stein's method, the same approach has been used in an earlier paper (Stolyar 2015b) to prove the tightness of diffusion-scaled stationary distributions for a two-queue system with many servers. The tightness result in Stolyar (2015b) establishes an approximation error of the fluid-limit, which is at the same order of the approximation error established in this paper.

To establish the approximation error, the paper identifies a fundamental connection between the perturbation theory for nonlinear systems and the convergence of mean-field models. The perturbation theory shows that for a stable nonlinear system with exponentially stable equilibrium point, the error of the first-order approximation of the nonlinear system is at the order of $O(\epsilon^2)$, where ϵ is the scaling factor of the perturbation. It turns out that the mean-square difference between the stationary distribution of the Mth CTMC and the equilibrium point of the mean-field model is related to the *cumulative* error (integrated over infinite time horizon) of this first-order approximation. After quantifying the cumulative error, we establish the following results for finite-dimensional mean-field models.

• If the mean-field model is perfect (see definition in Section 2); globally asymptotically stable and locally exponentially stable, then the stationary distributions of the CTMCs converges in the mean-square sense to the equilibrium point of the mean-field model with rate $O(1/M)$ (Theorem 1), specifically, we have the following result on the approximation error:

$$E[\|\mathbf{X}^{(M)}(\infty)-\mathbf{x}^*\|^2]=O\left(\frac{1}{M}\right). \tag{2}$$

• If the mean-field model is not perfect, sufficient conditions that guarantee the convergence of the stationary distributions have also been introduced in Remark 3.

We remark that the $O(1/M)$ convergence rate in terms of mean-square distance is intuitive and is consistent with the central limit theorem. The challenges of establishing this result are (1) $W_1^{(M)},\ldots,W_M^{(M)}$ are correlated, and (2) both the marginal and joint stationary distributions are hard to compute. We also remark that these results are different from the celebrated law of large numbers for Markov chains established by Kurtz (1971, 1981), where the convergence is established for sample paths of the CTMCs over a finite time interval or for a sequence of t_M that increases M increases (Norman 1974), not for the stationary distributions of the CTMCs. The contributions of our results are two-fold: First, our results provide a *direct* method of studying the convergence of stationary distributions of stochastic systems to their mean-field limits. The method connects the convergence of CTMCs with the stability of the mean-field model. Note that the mean-field model is a deterministic system, so it is often easier to analyze than the CTMCs. Second, the method quantifies the rate of convergence, and provides bounds on the approximation error when using the mean-field limit for approximating the performance of finite-size systems.

We finally comment that the convergence of stationary distributions of one-dimensional discrete-time Markov chains has been studied in Norman (1975). The approximation error of mean-field models for discrete-time Markov chains has been studied in Bortolussi and Hayden (2013), which, however, focuses on numerical methods to compute the error bounds and does not establish a general analytic answer like (2).

2. Mean-Field Models

The results in this paper are finite-M results instead of asymptotic results. In the rest of the paper, we assume M is fixed and drop the superscript (M) in the notations. In this paper, we consider an M-dimensional continuous-time Markov chain $\mathbf{W}\in\mathcal{U}^M$ with M nodes (or called particles) and $\mathcal{U}^M\subseteq\mathbf{R}^M$ is the state space of the CTMC. We assume $\mathcal{U}$ is a finite state space and the CTMC is irreducible. Without loss of generality, we assume $\mathcal{U}=\{1,\ldots,n\}$. We further define

$$X_i(t)=\frac{\sum_{m=1}^{M}\mathbf{1}_{W_m(t)=i}}{M}$$

where $\mathbf{1}$ is the indicator function, so $X_i(t)$ is the *fraction* of nodes in state i at time t. In this paper, we assume $\mathbf{X}=\{\mathbf{X}(t),\ t\geq 0\}$ is an (n-dimensional) CTMC. We use $\mathbf{X}(\infty)$ to denote its stationary distribution.

Furthermore, we have a mean-field model described by the following autonomous dynamical system:

$$\dot{\mathbf{x}} \triangleq \frac{d}{dt}x(t) = f(\mathbf{x}(t)), \quad \mathbf{x}(0) = \mathbf{x} \text{ and } \mathbf{x}(t) \in \mathscr{D} \subseteq [0,1]^n, \tag{3}$$

where $\mathscr{D}$ is a compact set. Note that while the CTMC with size M can only take a subset of values in $\mathbf{R}^M$, the mean-field model defined above is a dynamical system with domain $\mathscr{D}$, and can take any value in $\mathscr{D}$ by definition. Here, we abuse the notation and use $\mathbf{x}$ to denote the initial condition, which simplifies the notation in later analysis without causing too much confusion. Assume the system has a unique equilibrium point and let $\mathbf{x}^*$ denote the equilibrium point. The key idea of the mean-field analysis is to use the solution of this deterministic dynamical system to approximate the behavior of the CTMC when M is large; for example, use $\mathbf{x}^*$ to approximate $\mathbf{X}(\infty)$.

Let $Q_{x,y}$ denote the transition rate of the CTMC $\mathbf{X}$ from state $\mathbf{x}$ to state $\mathbf{y}$. A family of CTMCs is called a density-dependent family of CTMCs if the normalized transition rate

$$q_{x,y} = \frac{1}{M}Q_{x,y}$$

only depends on $\mathbf{x}$ and $\mathbf{y}$ but is independent of M (see a detailed definition in Mitzenmacher 1996). For a density-dependent family of CTMCs, we can often have a perfect mean-field model:

Perfect mean-field model: $f(\mathbf{x}) = \sum_{y:\, y\neq x} Q_{x,y}(\mathbf{y}-\mathbf{x})$, which is independent of M.

We next illustrate the idea using an SIS (susceptible-infected-susceptible) model with an external infection source, which is a variation of the original SIS model.

Example. Let W_m denote the state of an individual such that $W_m = 0$ if the individual is susceptible and $W_m = 1$ if the individual is infected. So X_0 is the fraction of susceptible individuals and X_1 is the fraction of infected individuals. We assume the recover time of an individual follows an exponential distribution with mean 1. Each infected node randomly selects a node after waiting for a random time period exponentially distributed with mean $1/\beta$. If the selected node is an susceptible node, it gets infected. Each susceptible node, after it becomes susceptible, gets infected by an external infection source after a random time period that is exponentially distributed with mean $1/\alpha$. Therefore, $\mathbf{W}$ and $\mathbf{X}$ are CTMCs. Specifically, $\mathbf{X}$ has the following transition rates

$$Q_{x,y} = \begin{cases} M\alpha x_0 + M\beta x_0^{(M)} x_1, & \text{if } \mathbf{y} = \mathbf{x} + \frac{1}{M}\begin{pmatrix}-1\\1\end{pmatrix}, \\ Mx_1, & \text{if } \mathbf{y} = \mathbf{x} + \frac{1}{M}\begin{pmatrix}1\\-1\end{pmatrix} \\ -M\alpha x_0 - M\beta x_0 x_1 - Mx_1, & \text{if } \mathbf{y} = \mathbf{x}, \\ 0, & \text{otherwise.} \end{cases}$$

Note for a given M, computing the stationary distribution of $\mathbf{X}$ is not easy because it has a large state space

$$\left\{\frac{1}{M}, \frac{2}{M}, \ldots, 1\right\}^2$$

and the transition rates are nonlinear functions of the states.

However, the SIS considered above is a density-dependent CTMC, so we consider the following mean-field model

$$\begin{pmatrix}\dot{x}_0\\ \dot{x}_1\end{pmatrix} = f(\mathbf{x}) = \sum_{y:\, y\neq x} Q_{x,y}(\mathbf{y}-\mathbf{x}) = (\alpha x_0 + \beta x_0 x_1)\begin{pmatrix}-1\\1\end{pmatrix} + x_1\begin{pmatrix}1\\-1\end{pmatrix}.$$

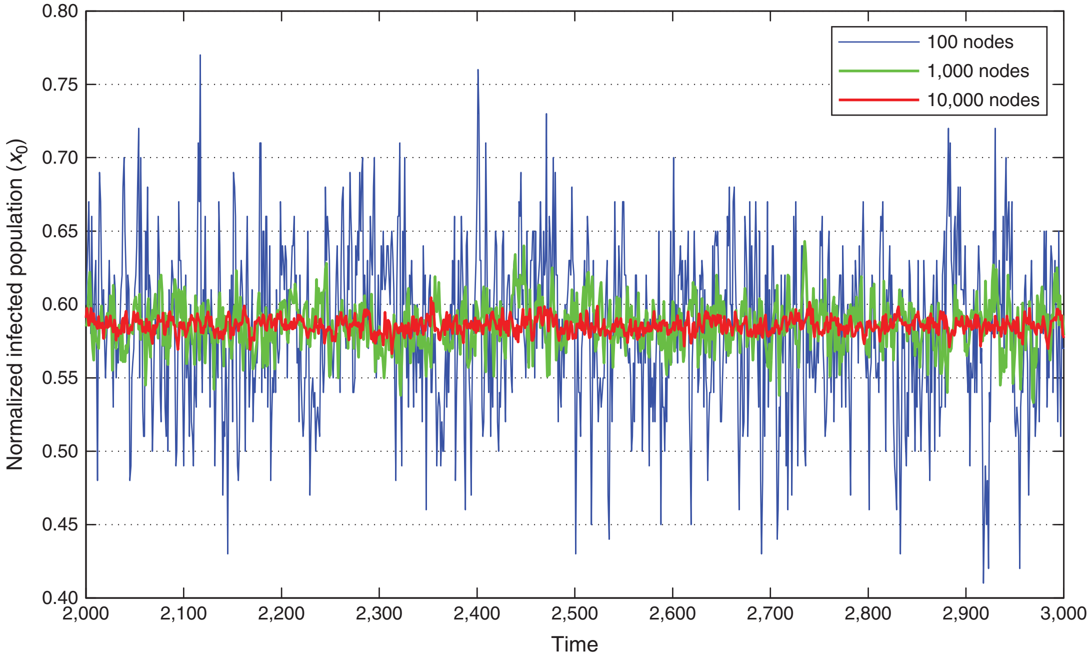
To solve the mean-field model above, we notice that $x_0 + x_1 = 1$ always holds, so we only need to consider

$$\dot{x}_0 = -\alpha x_0 - \beta x_0(1-x_0) + (1-x_0).$$

The equilibrium point can then be obtained by solving

$$0 = -\alpha x_0 - \beta x_0(1-x_0) + (1-x_0).$$

Figure 1. Simulation results of the fraction of susceptible population with $M = 100$, 1,000, and 100,000. We set $\alpha = \beta = 0.5$ in these simulations. The CTMCs were simulated using the uniformization method. The time is the scaled discrete-time used in the uniformization (scaled with M). Specifically, each time slot in the simulation includes M jumps on average.



For example, if $\alpha = \beta = 0.5$, then

$$x_0^* = 2 - \sqrt{2} \qquad \text{and} \qquad x_1^* = \sqrt{2} - 1,$$

which can be used to approximate the fractions of susceptible and infected populations when M is large, i.e., the stationary distribution of $\mathbf{X}$. The simulation results of the fraction of susceptible population with $M = 100$, 1,000 and 100,000 are shown in Figure 1, from which the convergence of X_0 to $2 - \sqrt{2}$ can be seen clearly.

3. Stein's Method for Quantifying the Approximation Error

In this section, we study the convergence and the approximation error (the rate of convergence) of the CTMCs to a mean-field model using Stein's method and the perturbation theory. Throughout this paper, $\|\cdot\|$ denotes the 2-norm, i.e., $\|\mathbf{x}\| = \sqrt{\sum_i x_i^2}$, and $|\cdot|$ denotes the absolute value. For two vectors $\mathbf{a}, \mathbf{b} \in \mathbf{R}^n$, $\mathbf{a} \cdot \mathbf{b}$ is the dot product. Furthermore, $(\partial g/\partial \mathbf{x})(\mathbf{x})$ denotes the gradient of $g(\mathbf{x})$ at $\mathbf{x}$, i.e.,

$$\frac{\partial g}{\partial \mathbf{x}} = \left(\frac{\partial g}{\partial x_1}, \ldots, \frac{\partial g}{\partial x_n}\right).$$

Furthermore, $(\partial x_i/\partial \mathbf{y})(t, \mathbf{y})$ refers to first-order partial derivatives of $x_i(t, \mathbf{y})$ with respect to the initial condition $\mathbf{y}$,

$$\frac{\partial \mathbf{x}}{\partial \mathbf{y}}(t, \mathbf{y}) = \begin{pmatrix} \dfrac{\partial x_1}{\partial \mathbf{y}}(t, \mathbf{y}) \\ \vdots \\ \dfrac{\partial x_n}{\partial \mathbf{y}}(t, \mathbf{y}) \end{pmatrix},$$

and $\dot{\mathbf{x}}(t, \mathbf{y}) = (\partial \mathbf{x}/\partial t)(t, \mathbf{y})$ is the derivative with respect to time and is an n-dimensional vector.

Recall the mean-field model defined in Equation (3):

$$\dot{\mathbf{x}} = f(\mathbf{x}(t)), \quad \mathbf{x}(0) = \mathbf{x} \text{ and } \mathbf{x}(t) \in \mathcal{D} \subseteq [0, 1]^n.$$

The mean-field model is said to be *globally asymptotically stable* if given $\epsilon > 0$, there exists t_ϵ such that starting from any initial condition $\mathbf{x}(0) \in \mathcal{D}$,

$$\|\mathbf{x}(t) - \mathbf{x}^*\| \le \epsilon, \quad \forall\, t \ge t_\epsilon.$$

The mean-field model is said to be *locally exponentially stable* if there exist positive constants ϵ, α and κ such that starting from any initial condition $\|\mathbf{x}(0)-\mathbf{x}^*\| \le \epsilon$,

$$\|\mathbf{x}(t)-\mathbf{x}^*\| \le \kappa \|\mathbf{x}(0)\| \exp(-\alpha t).$$

Let $g(\mathbf{x})$ be the solution to the following Poisson equation (also called Stein's equation)

$$\frac{\partial g}{\partial \mathbf{x}}(\mathbf{x})\dot{\mathbf{x}} = \frac{\partial g}{\partial \mathbf{x}}(\mathbf{x}) f(\mathbf{x}) = \|\mathbf{x}-\mathbf{x}^*\|^2. \tag{4}$$

Then, the solution has the following form

$$g(\mathbf{x}) = -\int_0^\infty \|\mathbf{x}-\mathbf{x}^*\|^2\,dt,$$

when the integral is finite (see Barbour 1988, Gotze 1991), where $\mathbf{x}(t,\mathbf{x})$ is the trajectory of the dynamical system with $\mathbf{x}$ as the initial condition. The integral is finite when the mean-field model is asymptotically stable and locally exponentially stable because it can be shown that there exists time τ such that $\int_0^\tau \|\mathbf{x}-\mathbf{x}^*\|^2\,dt$ is bounded due to stability and $\int_\tau^\infty \|\mathbf{x}-\mathbf{x}^*\|^2\,dt$ is bounded due to local exponential stability. These will become clear in Section A. Note that $-g(\mathbf{x})$ can be viewed as the cumulative square-deviation of the system state from the equilibrium point when the initial condition is $\mathbf{x}$.

Now let G denote the generator for the Mth CTMC, then

$$Gg(\mathbf{X}) = \sum_{y:\, y\neq X} Q_{X,y}(g(\mathbf{y})-g(\mathbf{X})) = M \sum_{y:\, y\neq X} q_{X,y}(g(\mathbf{y})-g(\mathbf{X})).$$

Since $\mathbf{X}$ is irreducible and has finite state space, it has a unique stationary distribution. Initializing $\mathbf{X}(0)$ according to its stationary distribution, and taking the expectation over the stationary distribution $\mathbf{X}(\infty)$, we have

$$E[Gg(\mathbf{X}(\infty))] = E\left[M \sum_{y:\, y\neq X(\infty)} q_{X(\infty),y}(g(\mathbf{y})-g(\mathbf{X}(\infty)))\right] = 0. \tag{5}$$

Then by taking expectation of the Poisson Equation (4) over the stationary distribution $\mathbf{x}(\infty)$ and then adding (5) to the equation, we obtain

$$E[\|\mathbf{X}(\infty)-\mathbf{x}^*\|^2] = E\left[\frac{\partial g}{\partial \mathbf{x}}(\mathbf{X}(\infty)) f(\mathbf{X}(\infty)) - M \sum_{y:\, y\neq X(\infty)} q_{X(\infty),y}(g(\mathbf{y})-g(\mathbf{X}(\infty)))\right].$$

Now adding and subtracting $(\partial g/\partial \mathbf{x})(\mathbf{x}) \cdot \sum_{y:\, y\neq x} q_{x,y} M(\mathbf{y}-\mathbf{x})$ yields

$$\begin{aligned}
E[\|\mathbf{X}(\infty)-\mathbf{x}^*\|^2] &= E\Bigg[\frac{\partial g}{\partial \mathbf{x}}(\mathbf{X}(\infty)) f(\mathbf{X}(\infty)) - \frac{\partial g}{\partial \mathbf{x}}(\mathbf{X}(\infty)) \sum_{y:\, y\neq X(\infty)} q_{X(\infty),y} M(\mathbf{y}-\mathbf{X}(\infty)) \\
&\quad - M \sum_{y:\, y\neq X(\infty)} q_{X(\infty),y}\left(g(\mathbf{y})-g(\mathbf{X}(\infty)) - \frac{\partial g}{\partial \mathbf{x}}(\mathbf{X}(\infty))(\mathbf{y}-\mathbf{X}(\infty))\right)\Bigg] \\
&= E\Bigg[\frac{\partial g}{\partial \mathbf{x}}(\mathbf{X}(\infty))\left(f(\mathbf{X}(\infty)) - \sum_{y:\, y\neq X(\infty)} q_{X(\infty),y} M(\mathbf{y}-\mathbf{X}(\infty))\right) \\
&\quad - \sum_{y:\, y\neq X(\infty)} q_{X(\infty),y} M\left(g(\mathbf{y})-g(\mathbf{X}(\infty)) - \frac{\partial g}{\partial \mathbf{x}}(\mathbf{X}(\infty))(\mathbf{y}-\mathbf{X}(\infty))\right)\Bigg].
\end{aligned} \tag{6}$$

From the equality above, *intuitively*, that

$$E[\|\mathbf{X}(\infty)-\mathbf{x}^*\|^2]$$

converges to zero as $M\to\infty$ can be established if the following conditions are true:

- Bounded gradient of $g(\mathbf{x})$: $\|(\partial g/\partial \mathbf{x})(\mathbf{x})\|$ is bounded by a constant independent of M.
- Convergence of the generator:

$$f(\mathbf{x}) - \sum_{y:\, y\neq x} q_{x,y} M(\mathbf{y}-\mathbf{x}) = 0.$$

- Bounded transition-rate of the CTMC: $E[\sum_{y:\, y\neq X(\infty)} q_{X(\infty),\,y}]$ is bounded.
- Diminishing first-order approximation error:

$$\left\|g(\mathbf{y}) - g(\mathbf{x}) - \frac{\partial g}{\partial \mathbf{x}}(\mathbf{x})(\mathbf{y}-\mathbf{x})\right\| = O\left(\frac{1}{M^2}\right).$$

Note that $g(\mathbf{x}) + (\partial g/\partial \mathbf{x})(\mathbf{x})(\mathbf{y}-\mathbf{x})$ is the first-order Taylor approximation of $g(\mathbf{y})$.

For many CTMCs and the associated mean-field models, the first three conditions mentioned above can be easily verified. In the following theorem, we will prove that the last condition holds when the mean-field model is globally asymptotically stable and locally exponentially stable (see inequality (11)), and then establish the rate of convergence based on that. The following theorem presents the main result of this paper.

Theorem 1. *The mean-square error of approximating the stationary distribution of the CTMC* $(\mathbf{X}(\infty))$, *defined in Section 2, using the equilibrium point* $(\mathbf{x}^*)$ *of the mean-field model* (3) *is* $O(1/M)$, *i.e.,*

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

when the following conditions hold:

- Bounded transition-rate condition: *There exists a constant* $c > 0$ *independent of* M *such that*

$$E\left[\sum_{y:\, y\neq X(\infty)} q_{X(\infty),\,y}\right] \le c.$$

- Bounded state transition condition: *There exists a constant* $\tilde{c}$ *independent of* M *such that* $\|\mathbf{x}-\mathbf{y}\| \le \tilde{c}/M$ *for any* $\mathbf{x}$ *and* $\mathbf{y}$ *such that* $q_{x,y} \neq 0$.
- Perfect mean-field model condition: *The mean-field model* (3) *is given by*

$$f(\mathbf{x}) = \sum_{y:\, y\neq x} q_{x,y} M(\mathbf{y}-\mathbf{x}), \quad \forall\, \mathbf{x}.$$

- Partial derivative condition: *First-order partial derivatives* $(\partial f_i/\partial x_j)(\mathbf{x})$ *exist and are Lipschitz.*
- Stability condition: *The mean-field model is globally asymptotically stable and locally exponentially stable.*

Remark 1. The first four conditions are easy to verify, so only the stability condition requires nontrivial work. Since a dynamical system has an exponentially stable equilibrium point if and only if the linearized system (at the equilibrium) is exponentially stable (see Khalil 2001, Theorem 4.15), the local exponential stability can be verified by proving the linearized system is exponentially stable (e.g., using Lyapunov method) or numerically verified by calculating the eigenvalues of the state matrix of the mean-field model. The global asymptotical stability in general is studied using the Lyapunov theorem. We also would like to point out that when the CTMC is density dependent, the perfect mean-field model is independent of M, so the convergence rate of the dynamical system is also independent of M. Two applications of this theorem in data center networks will be presented in Section 4.

Remark 2. It is worth pointing out that Kurtz's theorem (Kurtz 1971, 1981) states that the sample paths of the CTMCs converge to the trajectory of the mean-field model over any finite time interval. Therefore, if the mean-field model is unstable and the perfect mean-field model assumption holds, then the CTMCs are "unstable" as well.

Proof. Under the perfect mean-field model assumption, Equation (6) becomes

$$E[\|\mathbf{X}(\infty) - \mathbf{x}^*\|^2] = E\left[-\sum_{y:\, y\neq X(\infty)} q_{X(\infty),\,y} M\left(g(\mathbf{y}) - g(\mathbf{X}(\infty)) - \frac{\partial g}{\partial \mathbf{x}}(\mathbf{X}(\infty))(\mathbf{y}-\mathbf{X}(\infty))\right)\right], \tag{7}$$

where

$$g(\mathbf{x}) = -\int_0^\infty \|\mathbf{x}(t,\mathbf{x}) - \mathbf{x}^*\|^2\, dt.$$

We next focus on the following term

$$g(\mathbf{y}) - g(\mathbf{x}) - \frac{\partial g}{\partial \mathbf{x}}(\mathbf{x})(\mathbf{y}-\mathbf{x}). \tag{8}$$

Defining

$$\mathbf{x}^{(1)}(t,\mathbf{x},\mathbf{y}) = M\frac{\partial \mathbf{x}}{\partial \mathbf{x}}(t,\mathbf{x})(\mathbf{y}-\mathbf{x}), \tag{9}$$

we first have

$$\int_0^\infty 2(x_i(t,\mathbf{x})-x_i^*)\frac{\partial x_i}{\partial \mathbf{x}}(t,\mathbf{x})(\mathbf{y}-\mathbf{x})\,dt \le \frac{2}{M}\int_0^\infty |x_i^{(1)}(t,\mathbf{x},\mathbf{y})|\,dt \le \frac{2}{M}\int_0^\infty \|\mathbf{x}^{(1)}(t,\mathbf{x},\mathbf{y})\|\,dt \le \frac{2c}{\sigma}\frac{1}{M},$$

where the first inequality holds due to the definition of $x_i^{(1)}$ and the fact that $x_i(t,\mathbf{x})$, $x_i^*\in[0,1]$, and the last inequality holds due to inequality (14) in Lemma 1. Due to this result, we can exchange the order of integration and differentiation for the third term in (8) and get

$$-\left(g(\mathbf{y})-g(\mathbf{x})-\frac{\partial g}{\partial \mathbf{x}}(\mathbf{x})(\mathbf{y}-\mathbf{x})\right) = \int_0^\infty \sum_i \left((x_i(t,\mathbf{y})-x_i^*)^2-(x_i(t,\mathbf{x})-x_i^*)^2-2(x_i(t,\mathbf{x})-x_i^*)\frac{\partial x_i}{\partial \mathbf{x}}(t,\mathbf{x})(\mathbf{y}-\mathbf{x})\right)dt.$$

We next define

$$\mathbf{e}(t,\mathbf{x},\mathbf{y}) = \mathbf{x}(t,\mathbf{y})-\mathbf{x}(t,\mathbf{x})-\frac{1}{M}\mathbf{x}^{(1)}(t,\mathbf{x},\mathbf{y}),$$

i.e.,

$$x_i(t,\mathbf{y}) = e_i(t,\mathbf{x},\mathbf{y})+x_i(t,\mathbf{x})+\frac{1}{M}x_i^{(1)}(t,\mathbf{x},\mathbf{y}),$$

so

$$\begin{aligned}
&(x_i(t,\mathbf{y})-x_i^*)^2-(x_i(t,\mathbf{x})-x_i^*)^2-\frac{2}{M}(x_i(t,\mathbf{x})-x_i^*)x_i^{(1)}(t,\mathbf{x},\mathbf{y})\\
&\quad=\left(e_i(t,\mathbf{x},\mathbf{y})+x_i(t,\mathbf{x})-x_i^*+\frac{1}{M}x_i^{(1)}(t,\mathbf{x},\mathbf{y})\right)^2-(x_i(t,\mathbf{x})-x_i^*)^2-\frac{2}{M}(x_i(t,\mathbf{x})-x_i^*)x_i^{(1)}(t,\mathbf{x},\mathbf{y})\\
&\quad=e_i^2(t,\mathbf{x},\mathbf{y})+\frac{1}{M^2}(x_i^{(1)}(t,\mathbf{x},\mathbf{y}))^2+\frac{2}{M}e_i(t,\mathbf{x},\mathbf{y})x_i^{(1)}(t,\mathbf{x},\mathbf{y})+2e_i(t,\mathbf{x},\mathbf{y})(x_i(t,\mathbf{x})-x_i^*)\\
&\quad=e_i(t,\mathbf{x},\mathbf{y})\left(e_i(t,\mathbf{x},\mathbf{y})+\frac{2}{M}x_i^{(1)}(t,\mathbf{x},\mathbf{y})+2(x_i(t,\mathbf{x})-x_i^*)\right)+\frac{1}{M^2}(x_i^{(1)}(t,\mathbf{x},\mathbf{y}))^2.
\end{aligned}$$

According to inequality (15) in Lemma 1, we have

$$|\mathbf{e}(t,\mathbf{x},\mathbf{y})| = O\left(\frac{1}{M^2}\right).$$

According to the bounded state transition condition,

$$\|\mathbf{x}-\mathbf{y}\| \le \frac{\tilde{c}}{M}.$$

Furthermore, $|x_i^{(1)}(t,\mathbf{x},\mathbf{y})|$ is bounded by a constant c independent of M and t according to inequality (14) in Lemma 1. Therefore, there exists $\tilde{M}$ such that for any $M\ge\tilde{M}$,

$$\left|e_i(t,\mathbf{x},\mathbf{y})+\frac{2}{M}x_i^{(1)}(t,\mathbf{x},\mathbf{y})+2(x_i(t,\mathbf{x})-x_i^*)\right|\le 3,$$

which implies that

$$\left|g(\mathbf{y})-g(\mathbf{x})-\frac{\partial g}{\partial \mathbf{x}}(\mathbf{x})(\mathbf{y}-\mathbf{x})\right| \le 3\int_0^\infty |\mathbf{e}(t)|\,dt+\frac{1}{M^2}\int_0^\infty \|\mathbf{x}^{(1)}(t,\mathbf{x},\mathbf{y})\|^2\,dt. \tag{10}$$

According to inequalities (16) and (17) in Lemma 1, we have

$$\int_0^\infty |\mathbf{e}(t)|\,dt = O\left(\frac{1}{M^2}\right)$$

and

$$\int_0^\infty \|\mathbf{x}^{(1)}(t,\mathbf{x},\mathbf{y})\|^2\,dt = \Theta(1),$$

which implies

$$\left| g(\mathbf{y}) - g(\mathbf{x}) - \frac{\partial g}{\partial \mathbf{x}}(\mathbf{x})(\mathbf{y}-\mathbf{x}) \right| = O\left(\frac{1}{M^2}\right). \tag{11}$$

Substituting the result above into (7) yields

$$E[\|\mathbf{X}(\infty) - \mathbf{x}^*\|^2] = O\left(\frac{1}{M}\right) E\left[\sum_{y:\, y \neq X(\infty)} q_{X(\infty),\, y}\right]. \tag{12}$$

Finally, based on the bounded transition-rate condition, we conclude

$$E[\|\mathbf{X}(\infty) - \mathbf{x}^*\|^2] = O\left(\frac{1}{M}\right). \quad \square \tag{13}$$

Lemma 1. *Consider an n-dimensional dynamical system*

$$\dot{\mathbf{x}} = f(\mathbf{x}), \quad \mathbf{x} \in [0,1]^n,$$

which satisfies the partial derivative condition and the stability condition in Theorem 1. Consider $\mathbf{y}$ *and* $\mathbf{z}$ *such that* $\|\mathbf{y} - \mathbf{z}\| = \kappa/M$ *for some* $\kappa > 0$*, then there exist positive constants* c *and* σ *independent of* M *such that the following inequalities hold*:

$$\|\mathbf{x}^{(1)}(t, \mathbf{y}, \mathbf{z})\| \le c \exp(-\sigma t), \tag{14}$$

$$|\mathbf{e}(t, \mathbf{y}, \mathbf{z})| \le \frac{c}{M^2} \exp(-\sigma t), \tag{15}$$

which implies that

$$\int_0^\infty \|\mathbf{x}^{(1)}(t, \mathbf{y}, \mathbf{z})\|^2 \, dt \le \frac{c^2}{2\sigma}, \tag{16}$$

$$\int_0^\infty |\mathbf{e}(t, \mathbf{y}, \mathbf{z})| \, dt \le \frac{c}{\sigma M^2}. \quad \square \tag{17}$$

The proof of this lemma is based on the perturbation theory and is presented in the appendix.

Example. Let us go back to the SIS model introduced in Section 2. A closed-form solution can be obtained for $x_0(t)$. Again assume $\alpha = \beta = 0.5$, then the solution of the ordinary differential equation is

$$x_0(t) = \frac{-e^{-\sqrt{2}t}(2+\sqrt{2})(x_0(0) - 2 + \sqrt{2}) + (2-\sqrt{2})x_0(0) - 2}{-e^{-\sqrt{2}t}(x_0(0) - 2 + \sqrt{2}) + x_0(0) - 2 - \sqrt{2}},$$

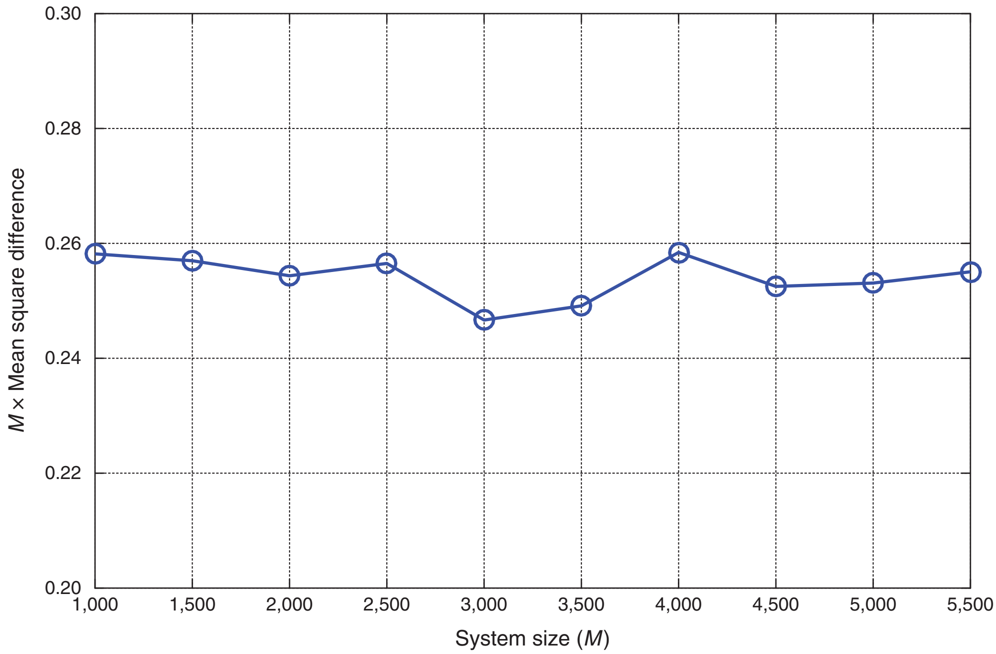
which converges to $2 - \sqrt{2}$ as $t \to \infty$ independent of $x_0(0)$. Therefore, it is easy to verify that the system is globally, asymptotically stable. Furthermore, the linearized system at the equilibrium is

$$\dot{\epsilon}_0 = -(\alpha + \beta(1 - x_0^*) + 1)\epsilon_0,$$

where $\epsilon_0 = x_0 - x_0^*$ and x_0^* is the equilibrium value, so the equilibrium point is locally exponentially stable. Furthermore, the mean-field model is perfect in this case and it can be easily verified that all other conditions in Theorem 1 hold. So, in the mean-square sense, stationary distributions converge to the $x_0 = 2 - \sqrt{2}$ and $x_1 = \sqrt{2} - 1$ with rate $O(1/M)$. Numerical evaluation of $ME[(x_0 - x_0^*)^2]$ versus M is shown in Figure 2, where M varies from 1,000 to 5,500. We can see that $ME[\sum_{i=1}^n (x_i - x_i^*)^2]$ varies within the interval [0.24, 0.26].

From this example, we can see that the mean-field limit is a good approximation of the system when the size of the system is moderate large and the mean-square approximation error of this example is around $1/(4M)$.

Figure 2. Numerical evaluation of $ME_{\mathbf{x}}[(x_0 - x_0^*)^2]$ versus M.



Remark 3. Theorem 1 requires a *perfect mean-field model* and *bounded state transitions*. Both conditions can be relaxed. Assume partial derivative condition and the stability condition in Theorem 1 hold. The stationary distributions of the CTMCs converge (in the mean square sense) to equilibrium point of the mean-field model, i.e.,

$$E[\|\mathbf{X}(\infty) - \mathbf{x}^*\|^2] = O(\max\{h_1(M), h_2(M)\})$$

when the following conditions also hold:

$$E\left[\left\|f(\mathbf{X}(\infty)) - \sum_{y:\, y \neq X(\infty)} q_{X(\infty),\, y} M(\mathbf{y} - \mathbf{X}(\infty))\right\|\right] = O(h_1(M)), \tag{18}$$

$$E\left[\sum_{y:\, y \neq X(\infty)} q_{X(\infty),\, y} M\|\mathbf{y} - \mathbf{X}(\infty)\|^2\right] = O(h_2(M)), \tag{19}$$

$$\max_{x,\, y:\, q_{x,y} > 0} \|\mathbf{y} - \mathbf{x}\| = o(1). \tag{20}$$

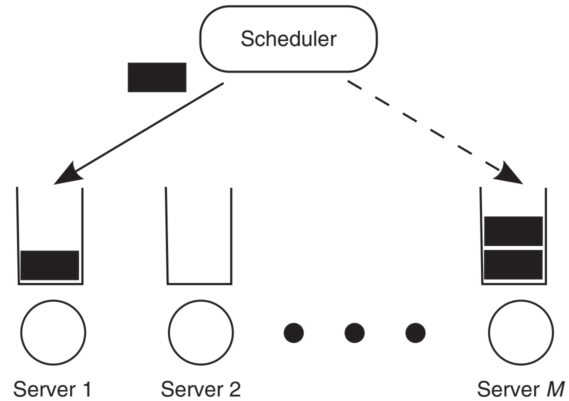
We say that the mean-field model is asymptotically accurate when condition (18) holds, which replaces the perfect mean-field model condition. Conditions (19) and (20) replace the bounded state transition condition. The proof can be found in Appendix B.

By relaxing this perfect mean-field model assumption, the CTMC is no longer required to be density dependent. Furthermore, relaxing the bounded state transition condition makes the result applicable to CTMCs for which the number of "jumps" during a transition is a function of M instead of a constant. The rate of convergence in these cases depends on the distance between the generator of the CTMC with size M and the mean-field model (see (18)), and depends on the mean-square jump size of the CTMC $\mathbf{X}$ at steady-state (see (19)).

4. Applications in Data Center Networks

In this section, we will demonstrate the novelty of Theorem 1 by considering two applications in data center networks: the power of two choices (Mitzenmacher 1996, Vvedenskaya et al. 1996) and the virtual machine placement problem (Xie et al. 2015). For both problems, mean-field models have been used to analyze the performance of the systems in an infinite server regime, but the approximation errors of the mean-field limits for systems with finite number of servers were unknown.

Figure 3. The system has M servers. When a task comes in, the scheduler samples two servers and routes the task to the server with shorter queue. In this example, the scheduler probes server 1 and server M and routes the task to server 1.



4.1. The Power of Two Choices for Servers with Finite Buffer

In Mitzenmacher (1996), Vvedenskaya et al. (1996), the authors considered a data center network with M identical servers as shown in Figure 3. Assume tasks arrive at the data center following a Poisson process with rate λM and the processing time of each task is exponentially distributed with mean processing time $\mu = 1$. Each server maintains a queue and $Q_m(t)$ denotes the queue size of server m at time t. For each incoming task, the router (or called scheduler) randomly samples two servers and dispatches the task to the server with a smaller queue size. In this setting, $\mathbf{Q}(t)$ is a CTMC and is a population process.

Let $s_k(t)$ denote the fraction of servers with queue size *at least* k. Based on the mean-field analysis, it has been shown in Mitzenmacher (1996) and Vvedenskaya et al. (1996) that $s_k(\infty)$ weakly converges to s_k^*, where s_k^* is the equilibrium point of the following mean-field system:

$$\dot{s}_k = \begin{cases} \lambda(s_{k-1}^2 - s_k^2) - (s_k - s_{k+1}), & k \geq 1, \\ 1, & k = 0. \end{cases}$$

The mean-field model above is an infinite-dimensional system, so Theorem 1 does not apply. We instead consider finite-buffer servers with buffer size B, for which the following mean-field model is a perfect mean-field model for the finite-buffer system:

$$\dot{s}_k = \begin{cases} 1, & k = 0, \\ \lambda(s_{k-1}^2 - s_k^2) - (s_k - s_{k+1}), & B-1 \geq k \geq 1, \\ \lambda(s_{k-1}^2 - s_k^2) - s_k, & k = B; \end{cases}$$

and the equilibrium point satisfies the conditions:

$$\begin{aligned} s_0^* &= 1, \\ \lambda({s_{k-1}^*}^2 - {s_k^*}^2) - (s_k^* - s_{k+1}^*) &= 0, \quad B-1 \geq k \geq 1, \\ \lambda({s_{k-1}^*}^2 - {s_k^*}^2) - s_k^* &= 0, \quad k = B. \end{aligned}$$

The existence and uniqueness of the solution have been proved in Mitzenmacher (1996).

Define the Lyapunov function to be

$$V(\mathbf{s}(t)) = \sum_{k=1}^{B} w_k |s_k(t) - s_k^*|$$

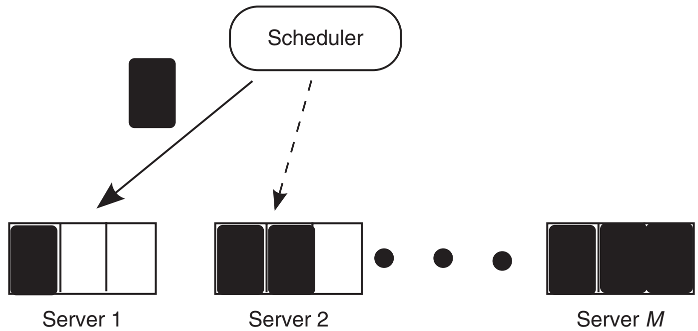
for w_k satisfies

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

for some $\delta > 0$. The existence of such $w_k > 0$ and $\delta > 0$ for the infinite-dimensional mean-field model has been proved in Mitzenmacher (1996). The same w_k and δ can be used in the finite-dimensional system as well. Following a similar analysis in Mitzenmacher (1996), we obtain

$$\dot{V}(t) \leq -\delta V(t),$$

Figure 4. The system has M servers, and each server can host at most three three virtual machines (VMs). The scheduler samples two servers and routes the VM request to the server with a smaller number of VMs. In this example, the scheduler routes the VM request to server 1. The request is blocked if both servers are full.



which implies that

$$\|\mathbf{s}(t)-\mathbf{s}^*\| \leq \frac{1}{w_1}V(t) = \frac{1}{w_1}\sum_{k=1}^{B} w_k|s_k(t)-s_k^*| \leq \frac{V(0)}{w_1}e^{-\delta t}.$$

So the system is globally, exponentially stable. Other conditions in Theorem 1 can be easily verified. So the approximation error in Theorem 1 applies.

4.2. Virtual Machine Placement in Cloud Computing Systems

In Xie et al. (2015), the authors considered a data center network with M identical servers where each server has B units of resources and can host at most B virtual machines (VMs) as shown in Figure 4. Assume VM requests arrive according to a Poisson process with rate λM and the lifetime of each VM is exponentially distributed with mean $\mu = 1$. Let $Q_m(t)$ denotes the number of VMs hosted at server m at time t. For each incoming request, the router (or called scheduler) randomly samples two servers and dispatches to the server with a smaller number of VMs. If both servers have already hosted B VMs, the request is blocked. In this setting, $\mathbf{Q}(t)$ again is a CTMC and is a population process.

Let $s_k(t)$ denote the fraction of servers with *at least* k VMs. Based on the mean-field analysis, it has been shown in Xie et al. (2015) that $s_k(\infty)$ weakly converges to s_k^*, where s_k^* is the equilibrium point of the following mean-field system:

$$\dot{s}_k = \begin{cases} 1, & k=0,\\ \lambda(s_{k-1}^2 - s_k^2) - k(s_k - s_{k+1}), & B-1 \geq k \geq 1,\\ \lambda(s_{k-1}^2 - s_k^2) - Bs_k, & k = B. \end{cases}$$

This is a finite-dimensional mean-field model, so Theorem 1 can be applied. The equilibrium point in this case can be recursively solved but the closed-form expression is difficult to obtain. The asymptotic stability of the system has been proved in Xie et al. (2015). We now consider the linearized system at the equilibrium, which is

$$\dot{x}_k = \begin{cases} 0, & k=0,\\ 2\lambda(s_{k-1}^* x_{k-1} - s_k^* x_k) - k(x_k - x_{k+1}), & B-1 \geq k \geq 1,\\ 2\lambda(s_{k-1}^* x_{k-1} - s_k^* x_k) - Bx_k, & k = B; \end{cases}$$

where $x_k = s_k - s_k^*$.

Define the Lyapunov function to be

$$V(t) = \sum_{k=1}^{B} |x_k|.$$

Note that when $x_k > 0$

$$\dot{\overline{|x_k|}} = \dot{x}_k = \begin{cases} 0, & k=0,\\ 2\lambda(s_{k-1}^* x_{k-1} - s_k^* x_k) - k(x_k - x_{k+1}), & B-1 \geq k \geq 1,\\ 2\lambda(s_{k-1}^* x_{k-1} - s_k^* x_k) - Bx_k, & k = B \end{cases}$$
$$\leq \begin{cases} 0, & k=0,\\ 2\lambda(s_{k-1}^* |x_{k-1}| - s_k^* |x_k|) - k(|x_k| - |x_{k+1}|), & B-1 \geq k \geq 1,\\ 2\lambda(s_{k-1}^* |x_{k-1}| - s_k^* |x_k|) - B|x_k|, & k = B; \end{cases}$$

and when $x_k < 0$,

$$\dot{|x_k|} = -\dot{x}_k = \begin{cases} 0, & k = 0, \\ -2\lambda(s^*_{k-1}x_{k-1} - s^*_k x_k) + k(x_k - x_{k+1}), & B-1 \geq k \geq 1, \\ -2\lambda(s^*_{k-1}x_{k-1} - s^*_k x_k) + Bx_k, & k = B \end{cases}$$
$$\leq \begin{cases} 0, & k = 0, \\ 2\lambda(s^*_{k-1}|x_{k-1}| - s^*_k|x_k|) - k(|x_k| - |x_{k+1}|), & B-1 \geq k \geq 1, \\ 2\lambda(s^*_{k-1}|x_{k-1}| - s^*_k|x_k|) - B|x_k|, & k = B. \end{cases}$$

It is not difficult to verify that the same inequalities hold when $x_k = 0$. Therefore, we have

$$\dot{V}(t) = \sum_{k=1}^{B} \dot{x}_k = -\sum_{k=1}^{B} |x_k| - 2\lambda s^*_B |x_B| \leq -\sum_{k=1}^{B} |x_k| = -V(t),$$

which implies that

$$\sum_{k=1}^{B} |x_k(t)| = V(t) \leq V(0)e^{-t}.$$

Therefore, the equilibrium point is (locally) exponentially stable. Other conditions in Theorem 1 again can be easily checked, so the approximation bound applies.

For both systems, the convergence of the stationary distributions to the mean-field limits have been proved in the literature based on the interchange of the limits, but the approximation errors (or the rate of convergence) were unknown. The result in this paper not only establishes the approximation errors, but also significantly reduces the additional analysis; in particular, both the convergence in finite time and the interchange of the limits are no longer needed. The purpose of presenting these two applications is to demonstrate the novelty of our result. The mean-field analysis of these two systems has been established in the literature and is not the focus of this paper. We therefore ignored the details and only presented the key steps. The simulation results that demonstrate the convergence of the finite systems can also be found in the original papers.

5. Conclusion

This paper studies the approximation error of a large-class of mean-field models. When the mean-field model is perfect, the mean-square difference (also called the rate of convergence) has been proved to be $O(1/M)$. Based on Stein's method for bounding the distance of probability distributions and the perturbation theory for nonlinear systems, a fundamental connection between the convergence to the mean-field limit and the stability of the mean-field model has been established. Two applications of mean-field models for large-scale data center networks were discussed to demonstrate the novelty of our results.

Acknowledgments

The author is very grateful to Prof. Jim Dai and Anton Braverman. Jim's seminar on Stein's method for the steady-state diffusion approximations inspired this work. The discussions with Jim and Anton had continuously stimulated the author during the writing of this paper. The author also wishes to thank Prof. Peter Glynn for very helpful discussions.

Appendix A. Proof of Lemma 1 Using the Perturbation Theory

Recall that

$$\mathbf{e}(t, \mathbf{y}, \mathbf{z}) = \mathbf{x}(t, \mathbf{z}) - \mathbf{x}(t, \mathbf{y}) - \frac{1}{M}\mathbf{x}^{(1)}(t, \mathbf{y}, \mathbf{z})$$

and

$$\mathbf{x}^{(1)}(t, \mathbf{y}, \mathbf{z}) = M\frac{\partial \mathbf{x}}{\partial \mathbf{y}}(t, \mathbf{y})(\mathbf{z} - \mathbf{y}).$$

The following analysis assumes fixed initial conditions $\mathbf{y}$ and $\mathbf{z}$. To simplify the notation, we drop $\mathbf{y}$ and $\mathbf{z}$ in $\mathbf{e}(t,\mathbf{y},\mathbf{z})$ and $\mathbf{x}^{(1)}(t,\mathbf{y},\mathbf{z})$.

Note that the derivative of $\mathbf{x}^{(1)}(t)$ with respect to time t is

$$\dot{\mathbf{x}}^{(1)}(t)=M\frac{\partial f}{\partial \mathbf{y}}(\mathbf{x}(t,\mathbf{y}))(\mathbf{z}-\mathbf{y})=M\frac{\partial f}{\partial \mathbf{x}}(\mathbf{x}(t,\mathbf{y}))\frac{\partial \mathbf{x}}{\partial \mathbf{y}}(t,\mathbf{y})(\mathbf{z}-\mathbf{y})=\frac{\partial f}{\partial \mathbf{x}}(\mathbf{x}(t,\mathbf{y}))\mathbf{x}^{(1)}(t).$$

Therefore, taking derivative of $\mathbf{e}(t)$ with respect to t yields

$$\dot{\mathbf{e}}(t)=\dot{\mathbf{x}}(t,\mathbf{z})-\dot{\mathbf{x}}(t,\mathbf{y})-\frac{1}{M}\dot{\mathbf{x}}^{(1)}(t) \tag{A.1}$$

$$=f(\mathbf{x}(t,\mathbf{z}))-f(\mathbf{x}(t,\mathbf{y}))-\frac{1}{M}\frac{\partial f}{\partial \mathbf{x}}(\mathbf{x}(t,\mathbf{y}))\mathbf{x}^{(1)}(t). \tag{A.2}$$

Note that $\mathbf{e}(t)=0$ according to its definition.

We first bound $\|\mathbf{x}^{(1)}(t)\|$ and $\int_0^\infty \|\mathbf{x}^{(1)}(t)\|^2\,dt$. From the stability condition of the lemma, we have the following results (Khalil 2001, Theorem 4.14):

- From the local exponential stability condition, we have that there exist constant d, c_l, c_u, c_p, φ and a Lyapunov function $V(\cdot)$ such that

$$c_l\|\mathbf{x}^{(1)}\|^2\le V(\mathbf{x}^{(1)})\le c_u\|\mathbf{x}^{(1)}\|^2, \tag{A.3}$$

$$\left\|\frac{\partial V}{\partial \mathbf{x}^{(1)}}(\mathbf{x}^{(1)})\right\|\le c_p\|\mathbf{x}^{(1)}\|, \tag{A.4}$$

and given $\|\mathbf{x}(t)\|\le d$,

$$\dot{V}(\mathbf{x}^{(1)}(t))\le -\varphi V(\mathbf{x}^{(1)}(t)). \tag{A.5}$$

- From the global stability condition, we have that given a positive constant d, there exists t_d such that $\|\mathbf{x}(t)\|\le d$ for any $t\ge t_d$.

We first calculate $\|\mathbf{x}^{(1)}(t)\|$ for $t\le t_d$. We know that

$$\frac{d}{dt}\|\mathbf{x}^{(1)}(t)\|^2=2(\mathbf{x}^{(1)}(t))^T\frac{\partial f}{\partial \mathbf{x}}(\mathbf{x}(t,\mathbf{y}))\mathbf{x}^{(1)}(t),$$

where the superscript T denotes the transpose. Since $\mathscr{D}$ is bounded and the first-order partial-derivatives are Lipschitz, the first-order partial-derivatives are bounded and there exists a constant b such that

$$\left|\frac{d}{dt}\|\mathbf{x}^{(1)}(t)\|^2\right|\le b\|\mathbf{x}^{(1)}(t)\|^2,$$

which implies that

$$\|x^{(1)}(t)\|^2\le \exp(bt)\|x^{(1)}(0)\|^2=\exp(bt)\|\mathbf{z}-\mathbf{y}\|^2,$$

i.e.,

$$\|x^{(1)}(t)\|\le \exp\left(\tfrac{1}{2}bt\right)\|\mathbf{z}-\mathbf{y}\|.$$

From (A.5), we have

$$V(\mathbf{x}^{(1)}(t))\le \exp(-\varphi(t-t_d))V(\mathbf{x}^{(1)}(t_d)),$$

which implies that

$$c_l\|\mathbf{x}^{(1)}(t)\|^2\le V(\mathbf{x}^{(1)}(t))\le \exp(-\varphi(t-t_d))V(\mathbf{x}^{(1)}(t_d))\le c_u\exp(-\varphi(t-t_d))\|\mathbf{x}^{(1)}(t_d)\|^2,$$

and

$$\|\mathbf{x}^{(1)}(t)\|\le\sqrt{\frac{c_u}{c_l}}\exp\left(-\frac{\varphi}{2}(t-t_d)\right)\|\mathbf{x}^{(1)}(t_d)\|\le\sqrt{\frac{c_u}{c_l}}\exp\left((b+\varphi)\frac{t_d}{2}\right)\|\mathbf{z}-\mathbf{y}\|\exp\left(-\frac{\varphi}{2}t\right), \tag{A.6}$$

which proves (14) in Lemma 1.

We next bound $|\mathbf{e}(t)|$. Recall that

$$\dot{e}_i(t)=f_i(\mathbf{x}(t,\mathbf{z}))-f_i(\mathbf{x}(t,\mathbf{y}))-\frac{1}{M}\frac{\partial f_i}{\partial \mathbf{x}}(\mathbf{x}(t,\mathbf{y}))\mathbf{x}^{(1)}(t) \tag{A.7}$$

$$=\frac{\partial f_i}{\partial \mathbf{x}}\left(\mathbf{x}(t,\mathbf{y})+\frac{\xi_i}{M}\mathbf{x}^{(1)}(t)+\xi_i\mathbf{e}(t)\right)\left(\frac{1}{M}\mathbf{x}^{(1)}(t)+\mathbf{e}(t)\right)-\frac{1}{M}\frac{\partial f_i}{\partial \mathbf{x}}(\mathbf{x}(t,\mathbf{y}))\mathbf{x}^{(1)}(t) \tag{A.8}$$

$$= \frac{\partial f_i}{\partial \mathbf{x}}\left(\mathbf{x}(t,\mathbf{y}) + \frac{\xi_i}{M}\mathbf{x}^{(1)}(t) + \xi_i\mathbf{e}(t)\right)\mathbf{e}(t) \tag{A.9}$$

$$+ \left(\frac{\partial f_i}{\partial \mathbf{x}}\left(\mathbf{x}(t,\mathbf{y}) + \frac{\xi_i}{M}\mathbf{x}^{(1)}(t) + \xi_i\mathbf{e}(t)\right) - \frac{\partial f_i}{\partial \mathbf{x}}(\mathbf{x}(t,\mathbf{y}))\right)\frac{1}{M}\mathbf{x}^{(1)}(t), \tag{A.10}$$

where $\xi_i \in [0,1]$ and the second equality holds due to the mean-value theorem. Since the first-order partial-derivatives are Lipschitz, there exists a constant $L > 0$ such that

$$|\dot{e}_i(t)| \leq \left|\frac{\partial f_i}{\partial \mathbf{x}}\left(\mathbf{x}(t,\mathbf{y}) + \frac{\xi_i}{M}\mathbf{x}^{(1)}(t) + \xi_i\mathbf{e}(t)\right)\mathbf{e}(t)\right| \tag{A.11}$$

$$+ \frac{L\xi_i}{M^2}\|\mathbf{x}^{(1)}(t)\|^2 + \frac{L\xi_i}{M^2}\|\mathbf{x}^{(1)}(t)\|\|\mathbf{e}(t)\|. \tag{A.12}$$

Furthermore, since the first-order partial-derivatives are also bounded (due to the Lipschitz condition), and $\mathbf{x}^{(1)}(t)$ is bounded by a constant independent of M as shown earlier, there exists a constant $\tilde{b}$ independent M such that

$$|\dot{e}_i(t)| \leq \frac{\tilde{b}}{M^2} + \tilde{b}|\mathbf{e}(t)|, \tag{A.13}$$

which implies that

$$\frac{d}{dt}|\mathbf{e}(t)| \leq \frac{\tilde{b}n}{M^2} + \tilde{b}n|\mathbf{e}(t)|. \tag{A.14}$$

From the inequality above, we have

$$|\mathbf{e}(t)| \leq (\exp(\tilde{b}nt) - 1)\frac{1}{M^2}. \tag{A.15}$$

Now we consider $|\mathbf{e}(t)|$ for $t \geq t_d$. We first have

$$\dot{V}(\mathbf{e}(t)) = \frac{\partial V}{\partial \mathbf{e}}(\mathbf{e}(t))\dot{\mathbf{e}}(t) = \frac{\partial V}{\partial \mathbf{e}}(\mathbf{e}(t))\left(f(\mathbf{x}(t,\mathbf{z})) - f(\mathbf{x}(t,\mathbf{y})) - \frac{1}{M}\frac{\partial f}{\partial \mathbf{x}}(\mathbf{x}(t,\mathbf{y}))\mathbf{x}^{(1)}(t)\right).$$

From (A.5), we know that for $\|\mathbf{x}(t)\| \leq d$,

$$\dot{V}(\mathbf{x}^{(1)}(t)) = \frac{\partial V}{\partial \mathbf{x}^{(1)}}(\mathbf{x}^{(1)}(t))\frac{\partial f}{\partial \mathbf{x}}(\mathbf{x}(t,\mathbf{y}))\mathbf{x}^{(1)}(t) \leq -\varphi V(\mathbf{x}^{(1)}(t)).$$

By replacing $\mathbf{x}^{(1)}(t)$ with $\mathbf{e}(t)$, we obtain

$$\frac{\partial V}{\partial \mathbf{e}}(\mathbf{e}(t))\frac{\partial f}{\partial \mathbf{x}}(\mathbf{x}(t,\mathbf{y}))\mathbf{e}(t) \leq -\varphi V(\mathbf{e}(t)),$$

which implies

$$\begin{aligned}
\dot{V}(\mathbf{e}(t)) &\leq -\varphi V(\mathbf{e}(t)) + \frac{\partial V}{\partial \mathbf{e}}(\mathbf{e}(t))\left(f(\mathbf{x}(t,\mathbf{z})) - f(\mathbf{x}(t,\mathbf{y})) - \frac{1}{M}\frac{\partial f}{\partial \mathbf{x}}(\mathbf{x}(t,\mathbf{y}))\mathbf{x}^{(1)}(t) - \frac{\partial f}{\partial \mathbf{x}}(\mathbf{x}(t,\mathbf{y}))\mathbf{e}(t)\right) \\
&\underset{(a)}{\leq} -\varphi V(\mathbf{e}(t)) + \sum_i \frac{\partial V}{\partial e_i}(\mathbf{e}(t))\left(\frac{\partial f_i}{\partial \mathbf{x}}\left(\mathbf{x}(t,\mathbf{y}) + \frac{\xi_i}{M}\mathbf{x}^{(1)}(t) + \xi_i\mathbf{e}(t)\right) - \frac{\partial f_i}{\partial \mathbf{x}}(\mathbf{x}(t,\mathbf{y}))\right) \times \left(\frac{1}{M}\mathbf{x}^{(1)}(t) + \mathbf{e}(t)\right) \\
&\underset{(b)}{\leq} -\varphi V(\mathbf{e}(t)) + L\left(\left\|\frac{1}{M}\mathbf{x}^{(1)}(t)\right\|^2 + \|\mathbf{e}(t)\|^2\right)\sum_i \frac{\partial V}{\partial e_i}(\mathbf{e}(t)) \\
&\underset{(c)}{\leq} -\varphi V(\mathbf{e}(t)) + \frac{L\sqrt{n}c_p}{M^2}\|\mathbf{x}^{(1)}(t)\|^2\|\mathbf{e}(t)\| + L\sqrt{n}c_p\|\mathbf{e}(t)\|^3 \\
&\underset{(d)}{\leq} -\left(\varphi - \frac{L\sqrt{n}c_p}{c_l}\|\mathbf{e}(t)\|\right)V(\mathbf{e}(t)) + \frac{L\sqrt{n}c_p}{M^2}\|\mathbf{x}^{(1)}(t)\|^2\|\mathbf{e}(t)\| \\
&\underset{(e)}{\leq} -\left(\varphi - \frac{L\sqrt{n}c_p}{c_l}\|\mathbf{e}(t)\|\right)V(\mathbf{e}(t)) + \frac{1}{M^2}\frac{L\sqrt{n}c_p}{\sqrt{c_l}}\|\mathbf{x}^{(1)}(t)\|^2\sqrt{V(\mathbf{e}(t))} \\
&\underset{(f)}{\leq} -\left(\varphi - \frac{L\sqrt{n}c_p}{c_l}\|\mathbf{e}(t)\|\right)V(\mathbf{e}(t)) + \frac{\tilde{c}}{M^2}\exp(-\varphi t)\sqrt{V(\mathbf{e}(t))},
\end{aligned}$$

where

$$\tilde{c} = \frac{L\sqrt{n}c_p}{\sqrt{c_l}}\left(\sqrt{\frac{c_u}{c_l}}\exp\left((b+\varphi)\frac{t_d}{2}\right)\|\mathbf{z}-\mathbf{y}\|\right)^2.$$

In the inequalities above, (a) yields from the mean-value theorem, (b) holds because the first-order partial-derivatives are Lipschitz, (c) is based on inequality (A.4), (d) and (e) hold due to inequality (A.3), and (f) holds by substituting (A.6).

Define $W(t)=\sqrt{V(t)}$, then we have

$$\dot{W}(\mathbf{e}(t)) \le -\frac{1}{2}\left(\varphi - \frac{\sqrt{n}cc_p}{c_l}\|\mathbf{e}(t)\|\right)W(\mathbf{e}(t)) + \frac{\tilde{c}}{M^2}\exp(-\varphi t).$$

Considering the set such that $\|\mathbf{e}(t)\| \le 2c_l/(\sqrt{n}cc_p)$, we get

$$\dot{W}(\mathbf{e}(t)) \le -\frac{\varphi}{4}W(\mathbf{e}(t)) + \frac{\tilde{c}}{M^2}\exp(-\varphi t).$$

By the comparison lemma in Khalil (2001), we have

$$W(t) \le \exp\left(-\frac{\varphi}{4}(t-t_d)\right)W(t_d) + \frac{\tilde{c}}{M^2}\int_{t_d}^{t}\exp\left(-\frac{\varphi}{4}(t-\tau)-\varphi\tau\right)d\tau \le \frac{\hat{c}}{M^2}\exp\left(-\frac{\varphi}{4}t\right),$$

where $\hat{c}$ is a constant independent of M and the inequality holds because $|\mathbf{e}(t_d)| = O(1/M^2)$ due to inequality (A.15). From the inequality above, we have

$$|\mathbf{e}(t)| \le \sqrt{n}\|\mathbf{e}(t)\| \le \frac{\sqrt{n}}{\sqrt{c_l}}W(t) \le \frac{\sqrt{n}\hat{c}}{\sqrt{c_l}}\frac{1}{M^2}\exp\left(-\frac{\varphi}{4}t\right).$$

We therefore conclude (15) in Lemma 1.

Appendix B. Proof Related to Remark 3

First, recall that we have

$$\begin{aligned}
E[\|\mathbf{X}(\infty)-\mathbf{x}^*\|^2] &= E\left[\frac{\partial g}{\partial \mathbf{x}}(\mathbf{X}(\infty))\left(f(\mathbf{X}(\infty)) - \sum_{y:\,y\neq X(\infty)} q_{X(\infty),y}M(\mathbf{y}-\mathbf{X}(\infty))\right)\right.\\
&\quad\left. - \sum_{y:\,y\neq X(\infty)} q_{X(\infty),y}M\left(g(\mathbf{y}) - g(\mathbf{X}(\infty)) - \frac{\partial g}{\partial \mathbf{x}}(\mathbf{X}(\infty))(\mathbf{y}-\mathbf{X}(\infty))\right)\right]\\
&\le \left(\max_{\mathbf{x}}\left\|\frac{\partial g}{\partial \mathbf{x}}(\mathbf{x})\right\|\right)E\left[\left\|f(\mathbf{X}(\infty)) - \sum_{y:\,y\neq X(\infty)} q_{X(\infty),y}M(\mathbf{y}-\mathbf{X}(\infty))\right\|\right]\\
&\quad + E\left[\sum_{y:\,y\neq X(\infty)} q_{X(\infty),y}M\left|g(\mathbf{y}) - g(\mathbf{X}(\infty)) - \frac{\partial g}{\partial \mathbf{x}}(\mathbf{X}(\infty))(\mathbf{y}-\mathbf{X}(\infty))\right|\right].
\end{aligned}$$

Note that

$$\frac{\partial g}{\partial x_i}(\mathbf{y}) = -\int_0^\infty (x_i(t,\mathbf{y}) - x_i^*)x_i^{(1)}(t)\,dt$$

by choosing $\mathbf{z}=\mathbf{y}+(1/M)\mathbf{1}_i$. It is easy to verify $|(\partial g/\partial x_i)(\mathbf{y})|$ is bounded by a constant because both $x_i(t,\mathbf{y})$ and $x_i^{(1)}(t)$ can be bounded by $ce^{-\sigma t}$ for some $c>0$ and $\sigma>0$ according to the proof of Theorem 1.

Furthermore, under condition (20), we have

$$\int_0^\infty |\mathbf{e}(t,\mathbf{y},\mathbf{z})|\,dt = O(\|\mathbf{z}-\mathbf{y}\|^2),$$
$$\int_0^\infty |\mathbf{x}^{(1)}(t,\mathbf{y},\mathbf{z})|\,dt = O(\|\mathbf{z}-\mathbf{y}\|).$$

Both equations can be easily verified based on the proof of Lemma 1 by simply replacing M with $1/\|\mathbf{z}-\mathbf{y}\|$.

When condition (20) holds, following the analysis that leads to inequality (10), we can again show that there exists a constant b independent M such that

$$\left|g(\mathbf{z}) - g(\mathbf{y}) - \frac{\partial g}{\partial \mathbf{y}}(\mathbf{z}-\mathbf{y})\right| \le b\int_0^\infty |\mathbf{e}(t,\mathbf{y},\mathbf{z})|\,dt + \frac{1}{M^2}\int_0^\infty \|\mathbf{x}^{(1)}(t,\mathbf{y},\mathbf{z})\|^2\,dt = O(\|\mathbf{y}-\mathbf{z}\|^2).$$

Therefore, we have

$$\begin{aligned}
E[\|\mathbf{X}(\infty)-\mathbf{x}^*\|^2] &= O\left(E\left[\left\|f(\mathbf{X}(\infty)) - \sum_{y:\,y\neq X(\infty)} q_{X(\infty),y}M(\mathbf{y}-\mathbf{X}(\infty))\right\|\right]\right) + O\left(E\left[\sum_{y:\,y\neq X(\infty)} q_{X(\infty),y}M\|\mathbf{y}-\mathbf{X}(\infty)\|^2\right]\right)\\
&= O(\max\{h_1(M), h_2(M)\}).
\end{aligned}$$

References

Adlakha S, Johari R (2013) Mean field equilibrium in dynamic games with strategic complementarities. *Oper. Res.* 61(4):971–989.

Anantharam V, Benchekroun M (1993) A technique for computing sojourn times in large networks of interacting queues. *Probab. Engrg. Inform. Sci.* 7(4):441–464.

Baccelli F, Karpelevich F, Kelbert MY, Puhalskii A, Rybko A, Suhov YM (1992) A mean-field limit for a class of queueing networks. *J. Statist. Phys.* 66(3–4):803–825.

Bailey NTJ (1975) *The Mathematical Theory of Infectious Diseases and Its Applications* (Hafner Press, New York).

Barbour AD (1988) Steins's method and Poisson process convergence. *J. Appl. Probab.* 25(A):175–184.

Barbour AD, Chen LH (2005) *An Introduction to Stein's Method*, Vol. 4 (World Scientific, Singapore).

Bordenave C, Mcdonald D, Proutière A (2010) A particle system in interaction with a rapidly varying environment: Mean field limits and applications. *Networks and Heterogeneous Media* 5(1):31–62.

Bortolussi L, Hayden RA (2013) Bounds on the deviation of discrete-time markov chains from their mean-field model. *Performance Evaluation* 70(10):736–749.

Bramson M, Lu Y, Prabhakar B (2012) Asymptotic independence of queues under randomized load balancing. *Queueing Systems* 71(3): 247–292.

Braverman A, Dai JG (2017) Stein's method for steady-state diffusion approximations of $m/Ph/n+m$ systems. *Ann. Appl. Probab.* 27(1): 550–581.

Braverman A, Dai JG, Feng J (2016) Steins method for steady-state diffusion approximations: An introduction through the Erlang-A and Erlang-C models. *Stoch. Syst.* 6(2):301–366.

Cecchi F, Borst SC, van Leeuwaardena JSH (2015) Mean-field analysis of ultra-dense csma networks. *ACM SIGMETRICS Performance Evaluation Rev.* 43(2):13–15.

Chaintreau A, Le Boudec J-Y, Ristanovic N (2009) The age of gossip: Spatial mean field regime. Lin B, Xu J, Sengupta S, Shah D, eds. *Proc. ACM SIGMETRICS Internat. Conf. Measurement Model. Comput. Systems* (ACM, New York), 109–120.

Gotze F (1991) On the rate of convergence in the multivariate CLT. *Ann. Probab.* 19(2):724–739.

Gurvich I, et al. (2014) Diffusion models and steady-state approximations for exponentially ergodic markovian queues. *Adv. Appl. Probab.* 24(6):2527–2559.

Kadanoff LP (2009) More is the same; phase transitions and mean field theories. *J. Statist. Phys.* 137(5–6):777–797.

Khalil HK (2001) *Nonlinear Systems* (Prentice Hall, Upper Saddle River, NJ).

Kurtz TG (1971) Limit theorems for sequences of jump markov processes approximating ordinary differential processes. *J. Appl. Probab.* 8(2):344–356.

Kurtz TG (1981) *Approximation of Population Processes*, Vol. 36 (SIAM, Philadelphia).

Lasry J-M, Lions P-L (2007) Mean field games. *Japanese J. Math.* 2(1):229–260.

Lu Y, Xie Q, Kliot G, Geller A, Larus JR, Greenberg A (2011) Join-Idle-Queue: A novel load balancing algorithm for dynamically scalable web services. *Performance Evaluation* 68(11):1056–1071.

Manjrekar M, Ramaswamy V, Shakkottai S (2014) A mean field game approach to scheduling in cellular systems. *Proc. IEEE Conf. Comput. Comm.* (*INFOCOM*) (IEEE, Piscataway, NJ), 1554–1562.

Mitzenmacher M (1996) The power of two choices in randomized load balancing. Ph.D. thesis, University of California at Berkeley.

Mukhopadhyay A, Mazumdar RR (2013) Analysis of load balancing in large heterogeneous processor sharing systems. Preprint https://arxiv.org/abs/1311.5806.

Mukhopadhyay A, Karthik A, Mazumdar RR, Guillemin F (2015) Mean field and propagation of chaos in multi-class heterogeneous loss models. *Perform. Eval.* 91:117–131.

Norman MF (1974) A central limit theorem for Markov processes that move by small steps. *Ann. Probab.* 2(6):1065–1074.

Norman MF (1975) Limit theorems for stationary distributions. *Ann. Appl. Probab.* 15(1):561–575.

Stein C (1972) A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. *Proc. Sixth Berkeley Symp. Math. Stat. Prob.* (University of California Press, Berkeley, CA), 583–602.

Stein C (1986) *Approximate Computation of Expectations*. Lecture Notes-Monograph Ser., Vol. 7 (Institute of Mathematical Statistics, Hayward, CA).

Stolyar A (2015a) Pull-based load distribution in large-scale heterogeneous service systems. *Queueing Systems* 80(4):341–361.

Stolyar A (2015b) Tightness of stationary distributions of a flexible-server system in the Halfin-Whitt asymptotic regime. *Stochastic Systems* 5(2):239–267.

Sznitman A-S (1991) Topics in propagation of chaos. Hennequin PL, ed. *Ecole d'Eté de Probabilités de Saint-Flour* XIX–1989 Lecture Notes in Mathematics, Vol. 1464 (Springer, Berlin), 165–251.

Tsitsiklis JN, Xu K (2012) On the power of (even a little) resource pooling. *Stochastic Systems* 2(1):1–66.

Vvedenskaya ND, Dobrushin RL, Karpelevich FI (1996) Queueing system with selection of the shortest of two queues: An asymptotic approach. *Problemy Peredachi Informatsii* 32(1):20–34.

Xie Q, Dong X, Lu Y, Srikant R (2015) Power of d choices for large-scale bin packing: A loss model. *Proc. ACM SIGMETRICS Internat. Conf. Measuring Model. Comput. Systems* (ACM, New York), 321–334.

Ying L, Srikant R, Kang X (2015) The power of slightly more than one sample in randomized load balancing. *Proc. IEEE Conf. Comput. Comm.* (*INFOCOM*) (IEEE, Piscataway, NJ), 1131–1139.