On the Approximation Error of Mean-Field Models

Lei Ying
School of Electrical, Computer and Energy Engineering
Arizona State University
Tempe, AZ 85287
lei.ying.2@asu.edu

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 (CTMC), 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 with size M 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 limit 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.

1. INTRODUCTION

The mean-field method is to study large-scale and complex stochastic systems using simple deterministic models. The idea of the mean-field method 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 method has important applications in various areas including statistical physics, epidemiology, communication networks, queueing theory, and game theory (e.g., [15, 4, 3, 32, 22, 19, 12, 7, 1, 21, 11]). In particular, over the last few years, it also has emerged as a powerful method for analyzing large-scale cloud computing systems and data center networks. For example, in [22, 32], 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 (Po2)*. The result has been extended to heavy-tailed service-time distributions [9], and to heterogeneous servers [23]. In [31], a mean-field model has been used to quantify the significant benefit of resource pooling. [28] established the asymptotic optimality of the join-idle-queue (JIQ), proposed in [20], using the mean-field model. In [34], 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 [33], the mean-field analysis has been used for studying 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 $\mathcal{U} = \{1, \cdots, n\}$. 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. This paper focuses on the case such that $\mathbf{x}^{(M)} = \left\{\mathbf{x}^{(M)}(t), t \geq 0\right\}$ is also an (n-dimensional) CTMC, i.e., the CTMC is a population process

SIGMETRICS '16, June 14 - 18, 2016, Antibes Juan-Les-Pins, France

ISBN 978-1-4503-4266-7/16/06…$15.00

DOI: http://dx.doi.org/10.1145/2896377.2901463

[17, 18]. We remark that many applications of the mean-field method such as those in queueing networks and epidemiology are for population processes.

Now let $\mathbf{x}^{(M)}(\infty)$ denote the stationary distribution 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 CTMCs to the trajectory of the mean-field model for any finite time interval $[0, t]$, i.e.,

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

where $d(\cdot,\cdot)$ is some measure of distance. This can be proved using different techniques including Kurtz's theorem [17, 18, 22, 34], propagation of chaos [30, 2, 9], or the convergence of the transition semigroup of CTMCs [32, 23].

(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 behavior 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., $\|\mathbf{x}^{(M)}(\infty) - \mathbf{x}^*\|$ 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 [26, 27, 6], 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 [10], 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. The results in this paper also share similar spirit with the work by Gurvich [14], which establishes the rate of convergence of diffusion models for steady-state approximations for exponentially ergodic Markovian queues. This paper differs from both work in that it considers mean-field models instead of diffusion approximations.

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 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 the first-order approximation of the mean-field model. 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

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

- If the mean-field model is not perfect, sufficient conditions that guarantee the convergence of the stationary distributions have been obtained in Corollary 1.

We remark that these results are different from the celebrated law of large numbers for Markov chains established by Kurtz [17, 18], where the convergence is established for sample paths of the CTMCs over a finite time interval or for a sequence of t_M which increases M increases [24], not for the stationary distributions of the CTMCs. The contributions of our results are two-fold: First, it provides 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 [25]. The approximation error of mean-field models for discrete-time Markov chains has been studied in [8], which, however, focuses on numerical methods to compute the error bounds and does not establish a general analytic answer like (2). Furthermore, an approach similar to Stein's method has been used in [29] to prove the tightness of diffusion-scaled stationary distributions for a two-queue system with many servers. The tightness result in [29] establishes an approximation error of the fluid-limit, which is at the same order of the approximation error established in this paper. The key differences are that this paper considers mean-field models for population processes (i.e., with many queues instead of two queues) and establishes sufficient con-

ditions for the convergence and the rate of convergence of a large-class of systems instead of only for a specific system.

2. MEAN-FIELD MODELS

Consider 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, we assume $\mathcal{U} = \{1, \cdots, n\}$. We further define

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

where $\mathbf{1}$ is the indicator function, so $X_i^{(M)}(t)$ is the number of nodes in state i at time t. We further define

$$\mathbf{x}^{(M)}(t) = \frac{\mathbf{X}^{(M)}(t)}{M},$$

so $x_i^{(M)}(t) \in [0,1]$ represents the *fraction* of nodes in state i at time t. In this paper, we assume $\mathbf{x}^{(M)} = \left\{\mathbf{x}^{(M)}(t), t \geq 0\right\}$ is an (n-dimensional) CTMC. We use $\mathbf{x}^{(M)}(\infty)$ to denote its stationary distribution.

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

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

where $\mathcal{D}$ is a compact set. Here, we abuse the notation and use $\mathbf{x}$ to denote the initial condition, which simplifies the notation in the analysis later 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 $\mathbf{x}^{(M)}(\infty)$.

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

$$q_{\mathbf{x}^{(M)},\mathbf{y}^{(M)}} = \frac{1}{M} Q_{\mathbf{x}^{(M)},\mathbf{y}^{(M)}}$$

only depends on $\mathbf{x}^{(M)}$ and $\mathbf{y}^{(M)}$ but is independent of M (see a detailed definition in [22]). For a density-dependent family of CTMCs, the mean-field model can often be obtained by choosing

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

because $q_{\mathbf{x},\mathbf{y}}$ is the transition rate from $\mathbf{x}$ to $\mathbf{y}$ and $\mathbf{y} - \mathbf{x}$ is the change of system state when such a transition occurs.

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 $W_m^{(M)}$ denote the state of an individual such that $W_m^{(M)} = 0$ if the individual is susceptible and $W_m^{(M)} = 1$ if the individual is infected. So $x_0^{(M)}$ is the fraction of susceptible individuals and $x_1^{(M)}$ 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 that is 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, $\mathbf{W}^{(M)}$, $\mathbf{X}^{(m)}$ and $\mathbf{x}^{(M)}$ are CTMCs. Specifically, $\mathbf{x}^{(M)}$ has the following transition rates, where $x_0^{(M)}$ is the fraction of susceptible individuals and $x_1^{(M)}$ is the fraction of infected individuals:

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

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

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

and the transition rates are nonlinear functions of the states.

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

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

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

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

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

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

which can be used to approximate the fractions of susceptible and infected populations when M is large, i.e., the stationary distribution of $\mathbf{x}^{(M)}$. 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 $x_0^{(M)}$ 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

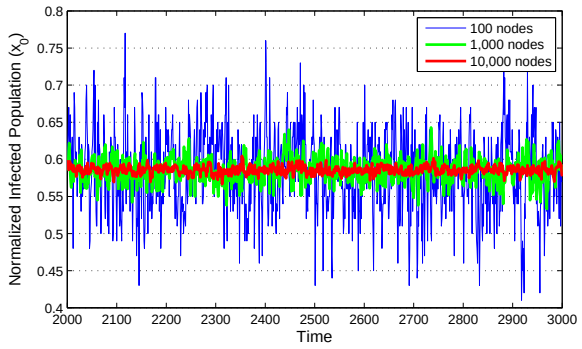

Figure 1: Simulation results of the fraction of susceptible population with $M = 100,\ 1,000$, and $100,000$. $\alpha = \beta = 0.5$ in these simulations. The CTMCs were simulated using the uniformization method. The time is the scaled discrete-time used in the uniformization (scaled with M). Specifically,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

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, $\nabla g(\mathbf{x})$ denotes the gradient of $g(\mathbf{x})$, and $\nabla x_i(t, \mathbf{x})$ refers to differentiating with respect to the location $\mathbf{x}$, and $\dot{\mathbf{x}}$ is the derivative with respect to time.

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 any initial condition $\mathbf{x}(0) \in \mathcal{D}$ and any $\epsilon > 0$, there exists $t(\mathbf{x}(0), \epsilon)$ such that

$$\|\mathbf{x}(t) - \mathbf{x}^*\| \leq \epsilon \quad \forall t \geq t(\mathbf{x}(0), \epsilon).$$

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 $\|\mathbf{x}(0) - \mathbf{x}^*\| \leq \epsilon$,

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

Let $g(\mathbf{x})$ be the solution to the Poisson equation

$$\nabla g(\mathbf{x}) \cdot \dot{\mathbf{x}} = \nabla g(\mathbf{x}) \cdot f(\mathbf{x}) = \sum_{i=1}^{n} (x_i - x_i^*)^2 \tag{4}$$

Then, the solution has the following form

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

when the integral is finite (see [5, 13]), 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, which will become clear in Section 5. 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 $G_{\mathbf{x}^{(M)}}$ denote the generator for the Mth CTMC, then

$$\begin{aligned} G_{\mathbf{x}^{(M)}} g(\mathbf{x}) &= \sum_{\mathbf{y}:\mathbf{y}\neq\mathbf{x}} Q_{\mathbf{x},\mathbf{y}}(\mathbf{x}) \left(g(\mathbf{y}) - g(\mathbf{x})\right) \\ &= M \sum_{\mathbf{y}:\mathbf{y}\neq\mathbf{x}} q_{\mathbf{x},\mathbf{y}}(\mathbf{x}) \left(g(\mathbf{y}) - g(\mathbf{x})\right). \end{aligned}$$

Since $\mathbf{x}^{(M)}$ is irreducible and has finite state space, $\mathbf{x}^{(M)}$ has a stationary distribution. Initializing $\mathbf{x}^{(M)}(0)$ according to its stationary distribution, and using $E_{\mathbf{x}^{(M)}}[\cdot]$ throughout to the expectation taking over the stationary distribution $\mathbf{x}^{(M)}(\infty)$, we have

$$\begin{aligned} &E_{\mathbf{x}^{(M)}}\left[G_{\mathbf{x}^{(M)}} g(\mathbf{x})\right] \\ =&E_{\mathbf{x}^{(M)}}\left[M \sum_{\mathbf{y}:\mathbf{y}\neq\mathbf{x}} q_{\mathbf{x},\mathbf{y}}(\mathbf{x}) \left(g(\mathbf{y}) - g(\mathbf{x})\right)\right] = 0. \end{aligned} \tag{5}$$

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

$$\begin{aligned} &E_{\mathbf{x}^{(M)}}\left[\sum_{i=1}^{n} (x_i - x_i^*)^2\right] \\ =\ & E_{\mathbf{x}^{(M)}}\left[\nabla g(\mathbf{x}) \cdot f(\mathbf{x}) - M \sum_{\mathbf{y}:\mathbf{y}\neq\mathbf{x}} q_{\mathbf{x},\mathbf{y}} \left(g(\mathbf{y}) - g(\mathbf{x})\right)\right] \end{aligned} \tag{6}$$

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

$$\begin{aligned} &E_{\mathbf{x}^{(M)}}\left[\sum_{i=1}^{n} (x_i - x_i^*)^2\right] \\ =\ & E_{\mathbf{x}^{(M)}}\Bigg[\nabla g(\mathbf{x}) \cdot f(\mathbf{x}) - \nabla g(\mathbf{x}) \cdot \sum_{\mathbf{y}:\mathbf{y}\neq\mathbf{x}} q_{\mathbf{x},\mathbf{y}} M(\mathbf{y} - \mathbf{x}) \\ &-M \sum_{\mathbf{y}:\mathbf{y}\neq\mathbf{x}} q_{\mathbf{x},\mathbf{y}} \left(g(\mathbf{y}) - g(\mathbf{x}) - \nabla g(\mathbf{x}) \cdot (\mathbf{y} - \mathbf{x})\right)\Bigg] \\ =\ & E_{\mathbf{x}^{(M)}}\Bigg[\nabla g(\mathbf{x}) \cdot \left(f(\mathbf{x}) - \sum_{\mathbf{y}:\mathbf{y}\neq\mathbf{x}} q_{\mathbf{x},\mathbf{y}} M(\mathbf{y} - \mathbf{x})\right) \\ &- \sum_{\mathbf{y}:\mathbf{y}\neq\mathbf{x}} q_{\mathbf{x},\mathbf{y}} M \left(g(\mathbf{y}) - g(\mathbf{x}) - \nabla g(\mathbf{x}) \cdot (\mathbf{y} - \mathbf{x})\right)\Bigg]. \end{aligned} \tag{7}$$

From the equality above, *intuitively*, that

$$E_{\mathbf{x}^{(M)}}\left[\sum_{i=1}^{n} (x_i - x_i^*)^2\right]$$

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

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

$$\lim_{M\to\infty} E_{\mathbf{x}^{(M)}}\left[\left\|f(\mathbf{x}) - \sum_{\mathbf{y}:\mathbf{y}\neq\mathbf{x}} q_{\mathbf{x},\mathbf{y}} M(\mathbf{y} - \mathbf{x})\right\|\right] = 0.$$

- Bounded transition-rate of the CTMC: $E_{\mathbf{x}^{(M)}}\left[\sum_{\mathbf{y}:\mathbf{y}\neq\mathbf{x}} q_{\mathbf{x},\mathbf{y}}\right]$ is bounded.

- Diminishing first-order approximation error:

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

  Note that $g(\mathbf{x}) + \nabla g(\mathbf{x}) \cdot (\mathbf{y} - \mathbf{x})$ 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 stationary distributions of the CTMCs ($\mathbf{x}^{(M)}(\infty)$), defined in Section 2, converge to the equilibrium point ($\mathbf{x}^*$) of the mean-field model (3) in the mean-square sense with rate $1/M$, i.e.,*

$$E_{\mathbf{x}^{(M)}}\left[\sum_{i=1}^{n} (x_i - x_i^*)^2\right] = O\left(\frac{1}{M}\right)$$

when the following conditions hold:

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

$$E_{\mathbf{x}^{(M)}}\left[\sum_{\mathbf{y}:\mathbf{y}\neq\mathbf{x}} q_{\mathbf{x},\mathbf{y}}\right] \leq c.$$

- **Bounded state transition condition:** *There exists a constant $\tilde{c}$ independent of M such that $\|\mathbf{x} - \mathbf{y}\| \leq \frac{\tilde{c}}{M}$ for any $\mathbf{x}$ and $\mathbf{y}$ such that $q_{\mathbf{x},\mathbf{y}} \neq 0$.*

- **Perfect mean-field model condition:** *The mean-field model (3) is given by*

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

- **Partial derivative condition:** *The function $f(\mathbf{x})$ is twice continuously differentiable.*

- **Stability condition:** *The mean-field model is globally asymptotically stable and is 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 Theorem 4.15 in [16]), 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. Two applications of this theorem in data center networks will be presented in Section 4.*

REMARK 2. *It is worth to pointing out that if the mean-field model is unstable but the perfect mean-field model assumption holds. Kurtz's theorem [17, 18] indicates that the sample paths of the CTMCs converge to the trajectory of the mean-field model for any finite time interval, which implies that the CTMCs are "unstable" as well.*

PROOF. We first prove the theorem assuming the mean-field model is globally exponentially stable, and then extend it to the general case. Under the perfect mean-field model assumption, equation (7) becomes

$$\begin{aligned} & E_{\mathbf{x}^{(M)}}\left[\sum_{i=1}^{n} (x_i - x_i^*)^2\right] \\ = \ & E_{\mathbf{x}^{(M)}}\left[-\sum_{\mathbf{y}:\mathbf{y}\neq\mathbf{x}} q_{\mathbf{x},\mathbf{y}} M\left(g(\mathbf{y}) - g(\mathbf{x}) - \nabla g(\mathbf{x}) \cdot (\mathbf{y} - \mathbf{x})\right)\right], \end{aligned}$$

where

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

We next focus on the following term,

$$\begin{aligned} & -\left(g(\mathbf{y}) - g(\mathbf{x}) - \nabla g(\mathbf{x}) \cdot (\mathbf{y} - \mathbf{x})\right) \\ = \ & \int_0^\infty \sum_i \Big(\left(x_i(t, \mathbf{y}) - x_i^*\right)^2 - \left(x_i(t, \mathbf{x}) - x_i^*\right)^2 \\ & -2\left(x_i(t, \mathbf{x}) - x_i^*\right) \nabla x_i(t, \mathbf{x}) \cdot (\mathbf{y} - \mathbf{x}) \Big)\, dt \end{aligned} \tag{8}$$

Note that we exchanged the order of integration and differentiation for the third term. This is can be done because

$$\int_0^\infty 2\left(x_i(t, \mathbf{x}) - x_i^*\right) \nabla x_i(t, \mathbf{x}) \cdot (\mathbf{y} - \mathbf{x})\, dt$$

is finite, which can be proved using the fact that both $(x_i(t, \mathbf{x}) - x_i^*)$ and $\|\nabla x_i(t, \mathbf{x})\|$ decay exponentially fast to zero as t increases (apply inequalities (20) and (31) with $\mathbf{z} = \mathbf{1}$), and the fact that $\|\mathbf{y} - \mathbf{x}\|$ is bounded due to the bounded state transition condition.

We next define

$$e_i(t) = x_i(t, \mathbf{y}) - x_i(t, \mathbf{x}) - \nabla x_i(t, \mathbf{x}) \cdot (\mathbf{y} - \mathbf{x}),$$

i.e.,

$$x_i(t, \mathbf{y}) = e_i(t) + x_i(t, \mathbf{x}) + \nabla x_i(t, \mathbf{x}) \cdot (\mathbf{y} - \mathbf{x}),$$

so

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

According to the perturbation theory, in particular, inequality (34), when the system is exponentially stable, we have that

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

According to the bounded state transition condition,

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

Furthermore, both $\nabla x_i(t, \mathbf{x})$ and $x_i(t, \mathbf{x})$ are bounded (see inequalities (20) and (31)) by constants independent of M

and t. Therefore, we can choose a constant b and a sufficiently large $\tilde{M}$ such that for any $M \geq \tilde{M}$,

$$|e_i(t) + 2\nabla x_i(t,\mathbf{x}) \cdot (\mathbf{y}-\mathbf{x}) + 2\left(x_i(t,\mathbf{x}) - x_i^*\right)| \leq b,$$

which implies that

$$\begin{aligned} &|g(\mathbf{y}) - g(\mathbf{x}) - \nabla g(\mathbf{x}) \cdot (\mathbf{y}-\mathbf{x})| \\ \leq & b\int_0^\infty \sum_i |e_i(t)|\, dt + \int_0^\infty \sum_i \left(\nabla x_i(t,\mathbf{x}) \cdot (\mathbf{y}-\mathbf{x})\right)^2 dt, \qquad (9) \\ \leq & b\sqrt{n}\int_0^\infty \|\mathbf{e}(t)\|\, dt + \int_0^\infty \sum_i \left(\nabla x_i(t,\mathbf{x}) \cdot (\mathbf{y}-\mathbf{x})\right)^2 dt, \qquad (10) \end{aligned}$$

where the last inequality is based on the following relation between 1-norm and 2-norm: $\sum_i |e_i(t)| \leq \sqrt{n}\|\mathbf{e}(t)\|$. In Section 5 (in particular, inequality (35)), we will show that under the exponential stability assumption,

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

From the bounded state transition condition, $\|\mathbf{y}-\mathbf{x}\|^2 \leq \frac{\tilde{c}^2}{M^2}$. Therefore,

$$\begin{aligned} &\int_0^\infty \sum_i \left(\nabla x_i(t,\mathbf{x}) \cdot (\mathbf{y}-\mathbf{x})\right)^2 dt \\ &\leq \frac{\tilde{c}^2}{M^2}\int_0^\infty \sum_i \|\nabla x_i(t,\mathbf{x})\|^2\, dt. \end{aligned}$$

Now according to inequality (31), there exist positive constants b_1 and b_2, both independent M, such that

$$\begin{aligned} &\int_0^\infty \sum_i \left(\nabla x_i(t,\mathbf{x}) \cdot (\mathbf{y}-\mathbf{x})\right)^2 dt \\ \leq \quad & \frac{\tilde{c}^2}{M^2}\int_0^\infty b_1 \exp(-b_2 t)\, dt \leq \frac{b_1}{b_2}\frac{1}{M^2}. \end{aligned}$$

Therefore, we can conclude that

$$|g(\mathbf{y}) - g(\mathbf{x}) - \nabla g(\mathbf{x}) \cdot (\mathbf{y}-\mathbf{x})| = O\left(\frac{1}{M^2}\right), \qquad (11)$$

which implies that

$$E_{\mathbf{x}^{(M)}}\left[\sum_{i=1}^n (x_i - x_i^*)^2\right] = O\left(\frac{1}{M}\right) E_{\mathbf{x}^{(M)}}\left[\sum_{\mathbf{y}:\mathbf{y}\neq\mathbf{x}} q_{\mathbf{x},\mathbf{y}}\right]. \qquad (12)$$

Finally, using the bounded transition rate condition, we conclude

$$E_{\mathbf{x}^{(M)}}\left[\sum_{i=1}^n (x_i - x_i^*)^2\right] = O\left(\frac{1}{M}\right). \qquad (13)$$

Now consider the case that the mean-field model is not globally exponentially stable, but is globally asymptotically stable and locally exponentially stable. Recall that $\mathcal{D} \subseteq [0,1]^n$ is compact. According to the definition of global asymptotic stability (Definition 4.4 in [16]), given any $\epsilon > 0$, there exists a finite time t' such that

$$\|\mathbf{x}(t) - \mathbf{x}^*\| \leq \epsilon$$

for any $t \geq t'$. For any finite t, following a similar analysis as in Section 5 (or Section 10.1 in [16]), $\|\mathbf{e}(t,\mathbf{x})\| = O(1/M^2)$ holds.[1] Therefore, we can bound the term (8) by separating the integration into two intervals: from 0 to t', and from t' to ∞, where t' is chosen such that $\mathbf{x}(t)$ converges exponentially to the equilibrium point after t'. Since $\|\mathbf{e}(t',\mathbf{x})\| = O(1/M^2)$, the analysis above applies to the integration over $[t',\infty)$. Hence, the result holds. □

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

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

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

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

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 $ME_{\mathbf{x}^{(M)}}\left[(x_0 - x_0^*)^2\right]$ versus M is shown in Figure 2, where M varies from 100 to 1,000. We can see that $ME_{\mathbf{x}^{(M)}}\left[\sum_{i=1}^n (x_i - x_i^*)^2\right]$ varies within the interval [0.21, 0.27] while the size of the system increases by 10 times (from 100 to 1,000). The standard deviation (deviation from $2-\sqrt{2}$) is 0.02177 ($3.72\% \times (2-\sqrt{2})$)) when $M = 100$ and is 0.0068 ($1.16\% \times (2-\sqrt{2})$) when $M = 1,000$.

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 $\frac{1}{4M}$. □

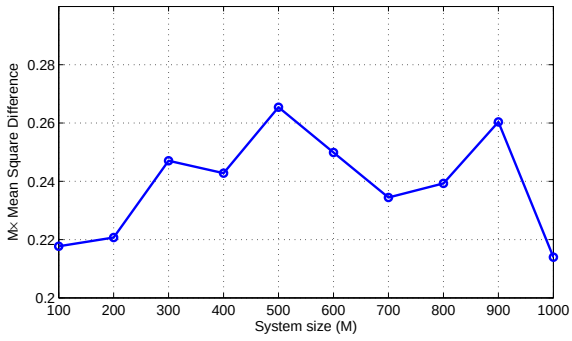

Figure 2: Numerical evaluation of $ME_{\mathbf{x}^{(M)}}\left[(x_0 - x_0^*)^2\right]$ versus M.

Theorem 1 requires a *perfect mean-field model* and *bounded state transitions*. Both conditions can be relaxed, but the rate of convergence will be different.

[1]This holds without exponential stability, but the constant in $O(1/M^2)$ may be a function of t if the system is not exponentially stable.

COROLLARY 1. *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.,*

$$\lim_{M\to\infty} E_{\mathbf{x}^{(M)}}\left[\sum_{i=1}^{n}(x_i - x_i^*)^2\right] = 0$$

when the following conditions also hold:

$$\lim_{M\to\infty} E_{\mathbf{x}^{(M)}}\left[\left\|f(\mathbf{x}) - \sum_{\mathbf{y}:\mathbf{y}\neq\mathbf{x}} q_{\mathbf{x},\mathbf{y}} M(\mathbf{y}-\mathbf{x})\right\|\right] = 0. \quad (14)$$

$$\lim_{M\to\infty} E_{\mathbf{x}^{(M)}}\left[\sum_{\mathbf{y}:\mathbf{y}\neq\mathbf{x}} q_{\mathbf{x},\mathbf{y}} M\|\mathbf{y}-\mathbf{x}\|^2\right] = 0 \quad (15)$$

$$\lim_{M\to\infty} \max_{\mathbf{x},\mathbf{y}:q_{\mathbf{x},\mathbf{y}}>0} \|\mathbf{y}-\mathbf{x}\| = 0 \quad (16)$$

We say that the mean-field model is asymptotically accurate when condition (14) holds, which replaces the perfect mean-field model condition. Conditions (15) and (16) replace the bounded state transition condition.

PROOF. First recall that we have

$$\begin{aligned}
&E_{\mathbf{x}^{(M)}}\left[\sum_{i=1}^{n}(x_i - x_i^*)^2\right]\\
=&E_{\mathbf{x}^{(M)}}\left[\nabla g(\mathbf{x})\cdot\left(f(\mathbf{x}) - \sum_{\mathbf{y}:\mathbf{y}\neq\mathbf{x}} q_{\mathbf{x},\mathbf{y}} M(\mathbf{y}-\mathbf{x})\right)\right.\\
&\left. - \sum_{\mathbf{y}:\mathbf{y}\neq\mathbf{x}} q_{\mathbf{x},\mathbf{y}} M\left(g(\mathbf{y}) - g(\mathbf{x}) - \nabla g(\mathbf{x})\cdot(\mathbf{y}-\mathbf{x})\right)\right]\\
\leq&\left(\max_{\mathbf{x}}\|\nabla g(\mathbf{x})\|\right) E_{\mathbf{x}^{(M)}}\left[\left\|f(\mathbf{x}) - \sum_{\mathbf{y}:\mathbf{y}\neq\mathbf{x}} q_{\mathbf{x},\mathbf{y}} M(\mathbf{y}-\mathbf{x})\right\|\right] + \quad (17)\\
&E_{\mathbf{x}^{(M)}}\left[\sum_{\mathbf{y}:\mathbf{y}\neq\mathbf{x}} q_{\mathbf{x},\mathbf{y}} M\left|g(\mathbf{y}) - g(\mathbf{x}) - \nabla g(\mathbf{x})\cdot(\mathbf{y}-\mathbf{x})\right|\right]. \quad (18)
\end{aligned}$$

By choosing $\mathbf{z} = \mathbf{1}$ in Section 5, it is easy to verify according to inequality (31) that $(\max_{\mathbf{x}}\|\nabla g(\mathbf{x})\|)$ is upper bounded by a constant independent of M. Therefore, under condition (14), (17) $\to 0$ as $M\to\infty$.

A careful examination of inequality (35) shows that

$$\int_0^\infty \|\mathbf{e}(t)\|\,dt = O\left(\|\mathbf{y}-\mathbf{x}\|^2 \exp\left(\alpha_3'\|\mathbf{y}-\mathbf{x}\|\right)\right).$$

So under condition (16), we have

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

When condition (16) holds, following the analysis that leads to inequality (10), we can again show that there exists a constant $\tilde{b}$ independent M such that

$$\begin{aligned}
&|g(\mathbf{y}) - g(\mathbf{x}) - \nabla g(\mathbf{x})\cdot(\mathbf{y}-\mathbf{x})|\\
\leq\ & \tilde{b}\sqrt{n}\int_0^\infty \|\mathbf{e}(t)\|\,dt + \int_0^\infty \sum_i \left(\nabla x_i(t,\mathbf{x})\cdot(\mathbf{y}-\mathbf{x})\right)^2 dt.
\end{aligned}$$

According to inequality (31), we also have

$$\int_0^\infty \sum_i \left(\nabla x_i(t,\mathbf{x})\cdot(\mathbf{y}-\mathbf{x})\right)^2 dt = O\left(\|\mathbf{y}-\mathbf{x}\|^2\right).$$

Therefore, we have

$$(18) = O\left(E_{\mathbf{x}^{(M)}}\left[\sum_{\mathbf{y}:\mathbf{y}\neq\mathbf{x}} q_{\mathbf{x},\mathbf{y}} M\|\mathbf{y}-\mathbf{x}\|^2\right]\right),$$

which converges to zero according to condition (15). Hence, the corollary holds. □

REMARK 3. *When the mean-field model is asymptotically accurate, the convergence rate depends on the convergence rates of (14) and (15).*

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 [22, 32] and the virtual machine placement problem [33]. For both problems, mean-field models have been used to analyze the performance of the systems in infinite server regime, but the approximation errors of the mean-field limits for systems with finite number of servers were unknown.

4.1 The power-of-two-choices for servers with finite buffer

In [22, 32], 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.

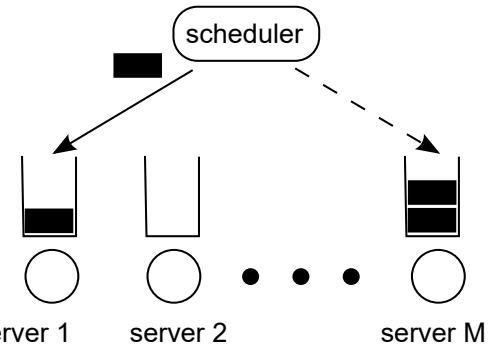

Figure 3: The system has M servers. When a task comes in, the scheduler samples two servers and routes the task to the server with shorter queue. In this example, the scheduler probes server 1 and server M and routes the task to server 1.

Let $s_k^{(M)}(t)$ denote the fraction of servers with queue size *at least* k. Based on the mean-field analysis, it has been shown in [22, 32] that $s_k^{(M)}(\infty)$ 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 has been proved in [22].

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 [22]. The same w_k and δ can be used in the finite-dimensional system as well. Following a similar analysis in [22], we obtain

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

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 [33], the authors considered a data center network with M identical servers where each server have 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 is exponentially distributed with mean lifetime $\mu = 1$. 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 $s_k^{(M)}(t)$ denote the fraction of servers with *at least* k VMs. Based on the mean-field analysis, it has been shown

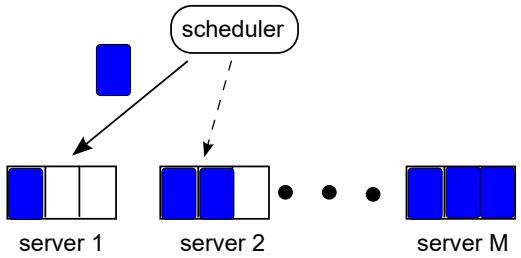

Figure 4: The system has M servers, and each server can host at most three VMs. the scheduler samples two servers and routes the VM request to the server with a smaller number of VMs. In this example, the scheduler routes the VM request to server 1. The request is blocked if both servers are full.

in [33] that $s_k^{(M)}(\infty)$ 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 [33]. 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 $x_k > 0$

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

and when $x_k < 0$,

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

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, so were not presented due to page limit.

5. THE PERTURBATION THEORY

In this section, we summarize the results of the perturbation theory for nonlinear systems. These results are special cases of those results presented in [16] because we only need to consider a perturbation to the initial condition. Furthermore, the mean-field model considered in this paper is an autonomous system, which again is a special case of the nonlinear system considered in [16]. For these reasons, the analysis of the perturbation results can be simplified. On the other hand, the perturbation method introduced in [16] only states that the 2-norm of the following error is is at the order of $\|\mathbf{y}-\mathbf{x}\|^2$ independent of t (under certain conditions)

$$\mathbf{e}(t) = \mathbf{x}(t, \mathbf{y}) - \mathbf{x}(t, \mathbf{x}) - \nabla \mathbf{x}(t, \mathbf{x}) \cdot (\mathbf{y} - \mathbf{x})$$

Our result on the rate of convergence, however, requires such an upper bound on the cumulative error, i.e., an upper bound on

$$\int_0^\infty \mathbf{e}(t)\, dt.$$

Therefore, it is necessary to go through the detailed analysis for the system considered in this paper to establish the result for the cumulative error. For the completeness of the paper and the easy reference of the reader, we next introduce the perturbation results in [16] with a more detailed calculation of $\|\mathbf{e}(t)\|$, which shows that not only the approximation error is bounded, but the upper bound decays exponentially to zero as t increases. The analysis closely follows [16].

Consider the system

$$\dot{\mathbf{x}} = f(\mathbf{x}) \tag{19}$$

where $f : \mathcal{D} \subseteq [0,1]^n \to R^n$. *Without the loss of generality, we assume* $\mathbf{x}^* = \mathbf{0}$. We are interested in comparing the solution of this nominal system with the system with a perturbation on the initial condition $\mathbf{x}(0) = \mathbf{x} + \epsilon\mathbf{z}$, where $\mathbf{z} = \frac{1}{\epsilon}(\mathbf{y} - \mathbf{x})$ and is an n-dimensional vector. For the mean-field analysis considered in this paper, $\epsilon = 1/M$. Under the condition of Theorem 1, for any neighboring states $\mathbf{x}$ and $\mathbf{y}$,

$$\|\mathbf{z}\| = \frac{1}{\epsilon}\|(\mathbf{y} - \mathbf{x})\| \leq \tilde{c}.$$

Let $\mathbf{x}(t, \epsilon)$ to denote the solution of the dynamical system with initial perturbation ϵ. Note that the dependence of the solution on $\mathbf{y} - \mathbf{x}$ is omitted to simplify the notation. The analysis holds for any $\mathbf{y}$ and $\mathbf{x}$. We next first repeat the assumptions on the nominal dynamical system.

ASSUMPTION 1. *For any i, the function $f_i(\mathbf{x})$ is twice continuously differentiable. Therefore, the Jacobian matrix of $f(\mathbf{x})$, denoted by $\frac{\partial f}{\partial x}$, is Lipschitz. In other words, there exists a constant $L > 0$ such that*

$$\left\| \frac{\partial f}{\partial x}(\mathbf{x}) - \frac{\partial f}{\partial x}(\mathbf{y}) \right\| \quad \leq \quad L\|\mathbf{x} - \mathbf{y}\|.$$

□

ASSUMPTION 2. *The dynamical system (19) has a unique equilibrium point and is exponentially stable. In other words, there exist positive constants α and κ such that starting from any initial condition $\mathbf{x}(0) \in \mathcal{D}$,*

$$\|\mathbf{x}(t)\| \leq \kappa \|\mathbf{x}(0)\| \exp(-\alpha t). \tag{20}$$

Under this assumption, according to Theorem 4.14 in [16], there exist a Lyapunov function $V(\mathbf{x})$ and positive constants c_u, c_l, c_d, and c_p such that for any $\mathbf{x} \in \mathcal{D}$, the following inequalities hold

$$\begin{aligned} c_l\|\mathbf{x}\|^2 \leq V(\mathbf{x}) &\leq c_u\|\mathbf{x}\|^2 \\ \dot{V}(\mathbf{x}) &\leq -c_d\|\mathbf{x}\|^2 \\ \|\nabla V(\mathbf{x})\| &\leq c_p\|\mathbf{x}\|. \end{aligned}$$

□

We first consider the finite Taylor series for $\mathbf{x}(t, \epsilon)$ in terms of ϵ :

$$\mathbf{x}(t, \epsilon) = \mathbf{x}^{(0)}(t) + \epsilon \mathbf{x}^{(1)}(t) + \mathbf{e}(t), \tag{21}$$

and

$$\mathbf{x}(0, \epsilon) = \mathbf{x} + \epsilon\mathbf{z}, \tag{22}$$

where

$$\mathbf{x}^{(0)}(t) = \mathbf{x}(t, 0) \text{ and } \mathbf{x}^{(1)}(t) = \left.\frac{d\mathbf{x}}{d\epsilon}(t, \epsilon)\right|_{\epsilon=0}.$$

Substituting (21) into the dynamical system equation, we get

$$\dot{\mathbf{x}}(t, \epsilon) = \dot{\mathbf{x}}^{(0)}(t) + \epsilon\dot{\mathbf{x}}^{(1)}(t) + \dot{\mathbf{e}}(t) = f(\mathbf{x}(t, \epsilon)) \tag{23}$$

$$= h^{(0)}(\mathbf{x}^{(0)}(t)) + h^{(1)}(\mathbf{x}^{(\leq 1)}(t))\epsilon + R_{\mathbf{e}}(t, \epsilon), \tag{24}$$

where $\mathbf{x}^{(\leq 1)} = \left(\mathbf{x}^{(0)}, \mathbf{x}^{(1)}\right)$. The zero-order term $h^{(0)}$ is given by

$$\dot{x}^{(0)}(t) = h^{(0)}\left(\mathbf{x}^{(0)}(t)\right) = f\left(\mathbf{x}^{(0)}(t)\right) \quad \text{with} \quad \mathbf{x}^{(0)}(0) = \mathbf{x},$$

which is the nominal system without the perturbation on the initial condition. The first-order term is given by

$$\begin{aligned} h^{(1)}\left(\mathbf{x}^{(\leq 1)}(t)\right) &= \frac{d}{d\epsilon} f(\mathbf{x}(t,\epsilon))\Big|_{\epsilon=0} \\ &= \frac{\partial f}{\partial x}(\mathbf{x}(t,\epsilon))\frac{d\mathbf{x}}{d\epsilon}(t,\epsilon)\Big|_{\epsilon=0} \\ &= \frac{\partial f}{\partial x}(\mathbf{x}^{(0)}(t))\mathbf{x}^{(1)}(t). \end{aligned}$$

Recall that $\frac{\partial f}{\partial x}$ is the Jacobian matrix. Therefore, we have

$$\dot{\mathbf{x}}^{(1)}(t) = \frac{\partial f}{\partial x}(\mathbf{x}^{(0)}(t))\mathbf{x}^{(1)}(t) \quad \text{with} \quad \mathbf{x}^{(1)}(0) = \mathbf{z}. \tag{25}$$

We next study $\mathbf{e}(t) = \mathbf{x}(t,\epsilon) - \mathbf{x}^{(0)}(t) - \epsilon\mathbf{x}^{(1)}(t)$. Combining the results above, we have

$$\begin{aligned} \dot{\mathbf{e}}(t) &= f\left(\mathbf{x}(t,\epsilon)\right) - f\left(\mathbf{x}^{(0)}(t)\right) - \epsilon\frac{\partial f}{\partial x}(\mathbf{x}^{(0)}(t))\mathbf{x}^{(1)}(t) \\ \mathbf{e}(0) &= 0. \end{aligned}$$

Now by defining

$$\begin{aligned} &\rho\left(\mathbf{x}^{(\leq 1)}(t), \mathbf{e}(t), \epsilon\right) \\ =& f\left(\mathbf{e}(t) + \mathbf{x}^{(0)}(t) + \epsilon\mathbf{x}^{(1)}(t)\right) - f\left(\mathbf{x}^{(0)}(t) + \epsilon\mathbf{x}^{(1)}(t)\right) \\ &- \frac{\partial f}{\partial x}(\mathbf{x}^{(0)}(t))\mathbf{e}(t), \end{aligned}$$

and

$$\begin{aligned} &\gamma\left(\mathbf{x}^{(\leq 1)}(t), \epsilon\right) \\ =& f\left(\mathbf{x}^{(0)}(t) + \epsilon\mathbf{x}^{(1)}(t)\right) - f\left(\mathbf{x}^{(0)}(t)\right) - \epsilon\frac{\partial f}{\partial x}(\mathbf{x}^{(0)}(t))\mathbf{x}^{(1)}(t), \end{aligned}$$

we obtain

$$\dot{\mathbf{e}}(t) = \frac{\partial f}{\partial x}(\mathbf{x}^{(0)}(t))\mathbf{e}(t) + \rho\left(\mathbf{x}^{(\leq 1)}(t), \mathbf{e}(t), \epsilon\right) + \gamma\left(\mathbf{x}^{(\leq 1)}(t), \epsilon\right). \tag{26}$$

Note that both ρ and γ are n-dimensional vectors. It is easy to see that

$$\rho\left(\mathbf{x}^{(\leq 1)}(t), 0, \epsilon\right) = \mathbf{0}. \tag{27}$$

According to Taylor's theorem and the mean value theorem, we have

$$\begin{aligned} \gamma_l\left(\mathbf{x}^{(\leq 1)}(t), \epsilon\right) &= \epsilon^2 \mathbf{x}^{(1)}(t)^T \mathbf{H}(f_l)(\xi)\mathbf{x}^{(1)}(t) \\ &= \epsilon^2 \sum_{i,j} \frac{\partial^2 f_l}{\partial x_i \partial x_j}(\xi)\left(x_i^{(1)}(t)x_j^{(1)}(t)\right) \end{aligned}$$

for $\xi = \mathbf{x}^{(0)}(t) + \alpha\epsilon\mathbf{x}^{(1)}(t)$ for some $0 \leq \alpha \leq 1$. $\mathbf{H}(f_l)$ is the Hessian matrix of function $f_l(\mathbf{x})$. Then we have

$$\left|\gamma_l\left(\mathbf{x}^{(\leq 1)}(t), \epsilon\right)\right| = \epsilon^2 \left|\sum_{i,j} \frac{\partial^2 f_l}{\partial x_i \partial x_j}(\xi)\left(x_i^{(1)}(t)x_j^{(1)}(t)\right)\right|.$$

Furthermore, we have

$$\begin{aligned} &\frac{\partial \rho_l}{\partial e_i}\left(\mathbf{x}^{(\leq 1)}(t), \mathbf{e}(t), \epsilon\right) \\ =& \frac{\partial f_l}{\partial x_i}(\mathbf{e}(t) + \mathbf{x}^{(0)}(t) + \epsilon\mathbf{x}^{(1)}(t)) - \frac{\partial f_l}{\partial x_i}(\mathbf{x}^{(0)}(t)). \end{aligned}$$

According to the mean-value theorem and (27), we have that

$$\begin{aligned} &\rho\left(\mathbf{x}^{(\leq 1)}(t), \mathbf{e}(t), \epsilon\right) \\ =& \left(\frac{\partial f}{\partial x}(\tilde{\mathbf{e}}(t) + \mathbf{x}^{(0)}(t) + \epsilon\mathbf{x}^{(1)}(t)) - \frac{\partial f}{\partial x}(\mathbf{x}^{(0)}(t))\right)\mathbf{e}(t), \end{aligned}$$

where $\tilde{\mathbf{e}}(t) = a\mathbf{e}(t)$ for some $0 \leq a \leq 1$. According to the Lipschitz condition in Assumption (1) and the Cauchy-Schwarz inequality, we have

$$\left\|\rho\left(\mathbf{x}^{(\leq 1)}(t), \mathbf{e}(t), \epsilon\right)\right\| \leq L\left(\|\mathbf{e}(t)\| + \epsilon\|\mathbf{x}^{(1)}(t)\|\right)\|\mathbf{e}(t)\|.$$

Now we utilize the assumption that the nominal system (19) converges to the equilibrium point exponentially fast from any initial condition in the domain. We use the Lyapunov function in Assumption (2) to bound $\|\mathbf{e}(t)\|$. We start from

$$\begin{aligned} &\dot{V}(\mathbf{e}(t)) \\ =& \nabla V(\mathbf{e}(t)) \cdot \dot{\mathbf{e}}(t) \\ =& \nabla V(\mathbf{e}(t)) \cdot f(\mathbf{e}(t)) + \nabla V(\mathbf{e}(t)) \cdot (\dot{\mathbf{e}}(t) - f(\mathbf{e}(t))) \\ \leq_{(a)}& -c_d V(\mathbf{e}(t)) + \nabla V(\mathbf{e}(t)) \cdot (\dot{\mathbf{e}}(t) - f(\mathbf{e}(t))) \\ =& -c_d V(\mathbf{e}(t)) + \nabla V(\mathbf{e}(t)) \cdot \\ &\left(\frac{\partial f}{\partial \mathbf{x}}(\mathbf{x}^{(0)}(t))\mathbf{e}(t) - \frac{\partial f}{\partial \mathbf{x}}(0)\mathbf{e}(t) + \frac{\partial f}{\partial \mathbf{x}}(0)\mathbf{e}(t) - f(\mathbf{e}(t))\right. \\ &\left. + \rho\left(\mathbf{x}^{(\leq 1)}(t), \mathbf{e}(t), \epsilon\right) + \gamma\left(\mathbf{x}^{(\leq 1)}(t), \epsilon\right)\right) \\ \leq& -c_d V(\mathbf{e}(t)) \\ &+ \|\nabla V(\mathbf{e}(t))\| \left(\left\|\frac{\partial f}{\partial \mathbf{x}}(\mathbf{x}^{(0)}(t))\mathbf{e}(t) - \frac{\partial f}{\partial \mathbf{x}}(0)\mathbf{e}(t)\right\|\right. \\ &+ \left\|\frac{\partial f}{\partial \mathbf{x}}(0)\mathbf{e}(t) - f(\mathbf{e}(t))\right\| \\ &\left. + \left\|\rho\left(\mathbf{x}^{(\leq 1)}(t), \mathbf{e}(t), \epsilon\right)\right\| + \left\|\gamma\left(\mathbf{x}^{(\leq 1)}(t), \epsilon\right)\right\|\right) \end{aligned}$$

where inequality (a) is due to assumption (2) and the last inequality is a result of the Cauchy-Schwarz inequality. Note that based on Assumption (1) and the mean-value theorem, we have

$$\begin{aligned} \left\|\frac{\partial f}{\partial \mathbf{x}}(\mathbf{x}^{(0)}(t))\mathbf{e}(t) - \frac{\partial f}{\partial \mathbf{x}}(0)\mathbf{e}(t)\right\| &\leq L\|\mathbf{x}^{(0)}(t)\|\|\mathbf{e}(t)\| \\ \left\|\frac{\partial f}{\partial \mathbf{x}}(0)\mathbf{e}(t) - f(\mathbf{e}(t))\right\| &\leq L\|\mathbf{e}(t)\|^2. \end{aligned}$$

We also know that

$$\begin{aligned} \left\|\rho\left(\mathbf{x}^{(\leq 1)}(t), \mathbf{e}(t), \epsilon\right)\right\| &\leq L\left(\|\mathbf{e}(t)\| + \epsilon\|\mathbf{x}^{(1)}(t)\|\right)\|\mathbf{e}(t)\| \\ \left\|\gamma\left(\mathbf{x}^{(\leq 1)}(t), \epsilon\right)\right\| &= \epsilon^2 A(t), \end{aligned}$$

where we define

$$A(t) = \sqrt{\sum_l \left(\sum_{i,j} \frac{\partial^2 f_l}{\partial x_i \partial x_j}(\xi)\left(x_i^{(1)}(t)x_j^{(1)}(t)\right)\right)^2}$$

to simplify the notation.

Summarizing the results above, we get

$$\begin{aligned} &\dot{V}(\mathbf{e}(t)) \\ \leq& -c_d V(\mathbf{e}(t)) \\ &+ L\|\nabla V(\mathbf{e}(t))\|\left(\|\mathbf{x}^{(0)}(t)\| + \epsilon\|\mathbf{x}^{(1)}(t)\| + 2\|\mathbf{e}(t)\|\right)\|\mathbf{e}(t)\| \end{aligned}$$

$$
\begin{aligned}
&+\|\nabla V(\mathbf{e}(t))\|\, A(t)\epsilon^2\\
\leq& -c_d V(\mathbf{e}(t))\\
&+Lc_p\left(\|\mathbf{x}^{(0)}(t)\|+\epsilon\|\mathbf{x}^{(1)}(t)\|+2\|\mathbf{e}(t)\|\right)\|\mathbf{e}(t)\|^2\\
&+c_pA(t)\epsilon^2\|\mathbf{e}(t)\|\\
\leq& -c_d V(\mathbf{e}(t))\\
&+L\frac{c_p}{c_l}\left(\|\mathbf{x}^{(0)}(t)\|+\epsilon\|\mathbf{x}^{(1)}(t)\|+2\|\mathbf{e}(t)\|\right)V(\mathbf{e}(t))\\
&+\frac{c_p}{\sqrt{c_l}}A(t)\epsilon^2\sqrt{V(\mathbf{e}(t))}.
\end{aligned}
$$

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

$$
\begin{aligned}
&\dot{W}(\mathbf{e}(t))\\
\leq& -\frac{c_d}{2}W(\mathbf{e}(t))\\
&+\frac{L}{2}\frac{c_p}{c_l}\left(\|\mathbf{x}^{(0)}(t)\|+\epsilon\|\mathbf{x}^{(1)}(t)\|+2\|\mathbf{e}(t)\|\right)W(\mathbf{e}(t))\\
&+\frac{c_p}{\sqrt{c_1}}A(t)\epsilon^2.
\end{aligned}
$$

By the comparison lemma in [16], we have

$$
\begin{aligned}
W(t) &\leq \phi(t,0)W(0)+\frac{c_p}{\sqrt{c_1}}\epsilon^2\int_0^t\phi(t,\tau)A(\tau)\,d\tau \quad (28)\\
&= \frac{c_p}{\sqrt{c_1}}\epsilon^2\int_0^t\phi(t,\tau)A(\tau)\,d\tau \quad (29)
\end{aligned}
$$

where the transition function $\phi(t,\tau)$ is

$$
\begin{aligned}
&\phi(t,\tau)=\exp\Big(-\frac{c_d}{2}(t-\tau)\\
&+\frac{L}{2}\frac{c_p}{c_l}\int_\tau^t\left(\|\mathbf{x}^{(0)}(\gamma)\|+\epsilon\|\mathbf{x}^{(1)}(\gamma)\|+2\|\mathbf{e}(\gamma)\|\right)d\gamma\Big),
\end{aligned}
$$

and the equality holds because $\mathbf{e}(0)=0$.

The following lemma proves that the first-order system (25) converges exponentially starting from any initial condition in $\mathcal{D}$.

LEMMA 1. *The first-order system (25) is exponentially stable for any solution $\mathbf{x}^{(0)}(t)$ that starts from $\mathcal{D}$.*

PROOF. That the nominal system is exponentially stable implies that the following linear system

$$\dot{\mathbf{x}}=\frac{\partial f}{\partial x}(0)\mathbf{x}$$

is (globally) exponentially stable, and $\frac{\partial f}{\partial x}(0)$ is Hurwitz (Corollary 4.3 in [16]), which further implies that there exists a positive definite symmetric matrix $\mathbf{P}$ such that

$$V(\mathbf{x})=\mathbf{x}^T\mathbf{P}\mathbf{x}$$

is a Lyapunov function for the linear system such that

$$\dot{V}(\mathbf{x})\leq-\|\mathbf{x}\|^2. \quad (30)$$

We start from

$$
\begin{aligned}
&\dot{V}(\mathbf{x}^{(1)}(t))\\
=&\nabla V(\mathbf{x}^{(1)}(t))\cdot\dot{\mathbf{x}}^{(1)}(t)\\
=&\nabla V(\mathbf{x}^{(1)}(t))\cdot\frac{\partial f}{\partial x}(0)\mathbf{x}^{(1)}(t)+\nabla V(\mathbf{x}^{(1)}(t))\cdot\\
&\left(\frac{\partial f}{\partial x}(\mathbf{x}^{(0)}(t))\mathbf{x}^{(1)}(t)-\frac{\partial f}{\partial x}(0)\mathbf{x}^{(1)}(t)\right)\\
\leq_{(a)}&-\left\|\mathbf{x}^{(1)}(t)\right\|^2+2\lambda_{\max}(\mathbf{P})\,L\left\|\mathbf{x}^{(0)}(t)\right\|\left\|\mathbf{x}^{(1)}(t)\right\|^2\\
\leq&-\frac{1}{\lambda_{\max}(\mathbf{P})}V(\mathbf{x}^{(1)}(t))+\frac{2\lambda_{\max}(\mathbf{P})}{\lambda_{\min}(\mathbf{P})}L\left\|\mathbf{x}^{(0)}(t)\right\|V\left(\mathbf{x}^{(1)}(t)\right)\\
\leq&-\left(\frac{1}{\lambda_{\max}(\mathbf{P})}-\frac{2\lambda_{\max}(\mathbf{P})}{\lambda_{\min}(\mathbf{P})}L\left\|\mathbf{x}^{(0)}(t)\right\|\right)V\left(\mathbf{x}^{(1)}(t)\right)
\end{aligned}
$$

where inequality (a) is based on (30) and the definition of $V(\mathbf{x})$, Assumption (1) and the mean-value theorem, and $\lambda_{\max}(\mathbf{P})$ is the largest eigenvalue of matrix $\mathbf{P}$.

By the comparison lemma, we have

$$
\begin{aligned}
&V(t)\\
\leq&\exp\left(-\frac{1}{\lambda_{\max}(\mathbf{P})}t+\frac{2\lambda_{\max}(\mathbf{P})}{\lambda_{\min}(\mathbf{P})}L\int_0^t\left\|\mathbf{x}^{(0)}(\tau)\right\|d\tau\right)V(0)\\
\leq&\exp\left(\frac{2\lambda_{\max}(\mathbf{P})\,L\kappa\|\mathbf{x}(0)\|}{\alpha\lambda_{\min}(\mathbf{P})}\right)\exp\left(-\frac{1}{\lambda_{\max}(\mathbf{P})}t\right)V(0),
\end{aligned}
$$

where the last inequality holds because the exponential convergence assumption (2) yields

$$\int_0^t\left\|\mathbf{x}^{(0)}(\tau)\right\|d\tau\leq\int_0^\infty\left\|\mathbf{x}^{(0)}(\tau)\right\|d\tau\leq\frac{\kappa\|\mathbf{x}(0)\|}{\alpha}.$$

Recall that $\mathbf{x}^{(1)}(0)=\mathbf{z}$, so $V(0)=\mathbf{z}^T\mathbf{P}\mathbf{z}=p$ and

$$
\begin{aligned}
&\|\mathbf{x}^{(1)}(t)\|^2\\
\leq&\frac{p}{\lambda_{\min}(\mathbf{P})}\exp\left(\frac{2\lambda_{\max}(\mathbf{P})\,L\kappa\|\mathbf{x}(0)\|}{\alpha\lambda_{\min}(\mathbf{P})}\right)\exp\left(-\frac{1}{\lambda_{\max}(\mathbf{P})}t\right). \quad (31)
\end{aligned}
$$

□

From the lemma above and assumption (1), we have that there exists a constant μ such that

$$
\begin{aligned}
&A(t)\\
\leq&\mu\|\mathbf{x}^{(1)}(t)\|^2\\
\leq&\frac{\mu p}{\lambda_{\min}(\mathbf{P})}\exp\left(\frac{2\lambda_{\max}(\mathbf{P})\,L\kappa\|\mathbf{x}(0)\|}{\alpha\lambda_{\min}(\mathbf{P})}\right)\exp\left(-\frac{1}{\lambda_{\max}(\mathbf{P})}t\right).
\end{aligned}
$$

Consider the set such that $\|\mathbf{e}(t)\|\leq\frac{c_dc_l}{4c_pL}$, we have

$$
\begin{aligned}
&\phi(t,\tau)\\
\leq&\exp\left(-\frac{c_d}{2}(t-\tau)+\frac{L}{2}\frac{c_p}{c_l}\int_\tau^t\left(\|\mathbf{x}^{(0)}(\gamma)\|+\epsilon\|\mathbf{x}^{(1)}(\gamma)\|\right)d\gamma\right)\\
\leq&\exp\Bigg(-\frac{c_d}{4}(t-\tau)+\frac{L}{2}\frac{c_p}{c_l}\Bigg(\frac{\kappa\|\mathbf{x}(0)\|}{\alpha}\\
&+2\epsilon\frac{\lambda_{\max}(\mathbf{P})\sqrt{p}}{\sqrt{\lambda_{\min}(\mathbf{P})}}\exp\left(\frac{\lambda_{\max}(\mathbf{P})L\kappa\|\mathbf{x}(0)\|}{\alpha\lambda_{\min}(\mathbf{P})}\right)\Bigg)\Bigg),
\end{aligned}
$$

where last inequality yields from (20) and (31).

Recall we have inequality (29)

$$W(t)\leq\frac{c_p}{\sqrt{c_1}}\epsilon^2\int_0^t\phi(t,\tau)A(\tau)\,d\tau. \quad (32)$$

Substituting the bounds on $\phi(t,\tau)$ and $A(\tau)$, we obtain

$$W(t)\leq\epsilon^2\sqrt{c_1}Z(\|\mathbf{x}(0)\|)\int_0^t\exp\left(-\frac{c_d}{4}(t-\tau)-\frac{1}{\lambda_{\max}(\mathbf{P})}\tau\right)d\tau,$$

where

$$Z(\|\mathbf{x}(0)\|)$$

$$
\begin{aligned}
=&\frac{c_p}{c_1}\exp\left(\frac{L}{2}\frac{c_p}{c_l}\left(\frac{\kappa\|\mathbf{x}(0)\|}{\alpha}\right.\right.\\
&\left.\left.+2\epsilon\frac{\lambda_{\max}(\mathbf{P})\sqrt{p}}{\sqrt{\lambda_{\min}(\mathbf{P})}}\exp\left(\frac{\lambda_{\max}(\mathbf{P})L\kappa\|\mathbf{x}(0)\|}{\alpha\lambda_{\min}(\mathbf{P})}\right)\right)\right)\times\\
&\frac{\mu p}{\lambda_{\min}(\mathbf{P})}\exp\left(\frac{2\lambda_{\max}(\mathbf{P})L\kappa\|\mathbf{x}(0)\|}{\alpha\lambda_{\min}(\mathbf{P})}\right)\\
=&\frac{c_p\mu p}{c_1\lambda_{\min}(\mathbf{P})}\exp\left(\frac{\kappa L}{\alpha}\left(\frac{c_p}{2c_l}+\frac{2\lambda_{\max}(\mathbf{P})}{\lambda_{\min}(\mathbf{P})}\right)\|\mathbf{x}(0)\|\right.\\
&\left.+\epsilon\sqrt{p}\frac{c_pL}{c_l}\frac{\lambda_{\max}(\mathbf{P})}{\sqrt{\lambda_{\min}(\mathbf{P})}}\exp\left(\frac{\lambda_{\max}(\mathbf{P})L\kappa\|\mathbf{x}(0)\|}{\alpha\lambda_{\min}(\mathbf{P})}\right)\right).
\end{aligned}
$$

In other words, we have

$$
\begin{aligned}
&\|\mathbf{e}(t)\|\\
\le&\epsilon^2 Z(\|\mathbf{x}(0)\|)\int_0^t\exp\left(-\frac{c_d}{4}(t-\tau)-\frac{1}{\lambda_{\max}(\mathbf{P})}\tau\right)d\tau \quad (33)\\
=&\begin{cases}\epsilon^2 Z(\|\mathbf{x}(0)\|)\frac{1}{\frac{c_d}{4}-\frac{1}{\lambda_{\max}(\mathbf{P})}}\left(\exp\left(-\frac{1}{\lambda_{\max}(\mathbf{P})}t\right)-\exp\left(-\frac{c_d}{4}t\right)\right), & \text{if } \frac{c_d}{4}\neq\frac{1}{\lambda_{\max}(\mathbf{P})}\\ \epsilon^2 Z(\|\mathbf{x}(0)\|)t\exp\left(-\frac{c_d}{4}t\right), & \text{otherwise.}\end{cases} \quad (34)
\end{aligned}
$$

Then

$$
\begin{aligned}
&\int_0^\infty\|\mathbf{e}(t)\|\,dt\\
\le&\begin{cases}\epsilon^2 Z(\|\mathbf{x}(0)\|)\frac{4\lambda_{\max}(\mathbf{P})}{c_d} & \text{if } \frac{c_d}{4}\neq\frac{1}{\lambda_{\max}(\mathbf{P})}\\ \epsilon^2 Z(\|\mathbf{x}(0)\|)\frac{16}{c_d^2}, & \text{otherwise.}\end{cases}.\\
=&\epsilon^2 Z(\|\mathbf{x}(0)\|)\frac{4\lambda_{\max}(\mathbf{P})}{c_d}.
\end{aligned}
$$

It is easy to see that with properly defined α_1, α_2, α_3 and α_4, we have

$$
\begin{aligned}
&\int_0^\infty\|\mathbf{e}(t)\|\,dt\\
\le&\epsilon^2 p\alpha_1\exp\left(\alpha_2\frac{\|\mathbf{x}(0)\|}{\alpha}+\epsilon\sqrt{p}\alpha_3\exp\left(\alpha_4\frac{\|\mathbf{x}(0)\|}{\alpha}\right)\right). \quad (35)
\end{aligned}
$$

We keep the terms $\|\mathbf{x}(0)\|$ and α to show that the cumulative error depends on the initial condition and the convergence rate of the mean-field model. Furthermore, $p=\mathbf{z}^T\mathbf{P}\mathbf{z}\le\lambda_{\max}(\mathbf{P})\|\mathbf{z}\|^2$.

6. 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.

Acknowledgement

The author is very grateful to 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. This work was supported in part by the NSF under Grant ECCS-1255425.

7. REFERENCES

[1] S. Adlakha and R. Johari. Mean field equilibrium in dynamic games with strategic complementarities. *Operations Research*, 61(4):971–989, 2013.

[2] V. Anantharam and M. Benchekroun. A technique for computing sojourn times in large networks of interacting queues. *Probability in the engineering and informational sciences*, 7(04):441–464, 1993.

[3] F. Baccelli, F. Karpelevich, M. Y. Kelbert, A. Puhalskii, A. Rybko, and Y. M. Suhov. A mean-field limit for a class of queueing networks. *Journal of statistical physics*, 66(3-4):803–825, 1992.

[4] N. T. J. Bailey. *The mathematical theory of infectious diseases and its applications*. Hafner Press, 1975.

[5] A. D. Barbour. Stein's method and Poisson process convergence. *J. Appl. Probab.*, pages 175–184, 1988.

[6] A. D. Barbour and L. H. Chen. *An Introduction to Stein's Method*, volume 4. World Scientific, 2005.

[7] C. Bordenave, D. Mcdonald, and A. Proutiere. A particle system in interaction with a rapidly varying environment: Mean field limits and applications. *Networks and Heterogeneous Media (NHM)*, 2010.

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

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

[10] A. Braverman and J. Dai. Stein's method for steady-state diffusion approximations of $m/ph/n+m$ systems. *arXiv preprint arXiv:1503.00774*, 2015.

[11] F. Cecchi, S. C. Borst, and J. S. H. van Leeuwaardena. Mean-field analysis of ultra-dense csma networks. *ACM SIGMETRICS Performance Evaluation Review*, 43(2):13–15, 2015.

[12] A. Chaintreau, J.-Y. Le Boudec, and N. Ristanovic. The age of gossip: spatial mean field regime. In *Proc. Ann. ACM SIGMETRICS Conf.*, pages 109–120, Seattle, Washington, USA, 2009.

[13] F. Gotze. On the rate of convergence in the multivariate clt. *Ann. Probab.*, pages 724–739, 1991.

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

[15] L. P. Kadanoff. More is the same; phase transitions and mean field theories. *Journal of Statistical Physics*, 137(5-6):777–797, 2009.

[16] H. K. Khalil. *Nonlinear systems*. Prentice Hall, 2001.

[17] T. G. Kurtz. Limit theorems for sequences of jump markov processes approximating ordinary differential processes. *J. Appl. Probab.*, 8(2):344–356, 1971.

[18] T. G. Kurtz. *Approximation of population processes*, volume 36. SIAM, 1981.

[19] J.-M. Lasry and P.-L. Lions. Mean field games. *Japanese Journal of Mathematics*, 2(1):229–260, 2007.

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

[21] M. Manjrekar, V. Ramaswamy, and S. Shakkottai. A mean field game approach to scheduling in cellular systems. In *Proc. IEEE Int. Conf. Computer Communications (INFOCOM)*, pages 1554–1562, Toronto, Canada, 2014.

[22] M. Mitzenmacher. *The Power of Two Choices in Randomized Load Balancing*. PhD thesis, University of California at Berkeley, 1996.

[23] A. Mukhopadhyay and R. R. Mazumdar. Analysis of load balancing in large heterogeneous processor sharing systems. *arXiv preprint arXiv:1311.5806*, 2013.

[24] M. F. Norman. A central limit theorem for Markov processes that move by small steps. *Ann. Probab.*, pages 1065–1074, 1974.

[25] M. F. Norman. Limit theorems for stationary distributions. *Ann. Appl. Probab.*, pages 561–575, 1975.

[26] C. Stein. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proc. Sixth Berkeley Symp. Math. Stat. Prob.*, pages 583–602, 1972.

[27] C. Stein. Approximate computation of expectations. *Lecture Notes-Monograph Series*, 7:i–164, 1986.

[28] A. L. Stolyar. Pull-based load distribution in large-scale heterogeneous service systems. *arXiv preprint arXiv:1407.6343*, 2014.

[29] A. L. Stolyar. Tightness of stationary distributions of a flexible-server system in the halfin-whitt asymptotic regime. *arXiv preprint arXiv:1403.4896*, 2014.

[30] A.-S. Sznitman. Topics in propagation of chaos. In *Ecole d'été de probabilités de Saint-Flour XIXâĂŤ1989*, pages 165–251. 1991.

[31] J. N. Tsitsiklis and K. Xu. On the power of (even a little) resource pooling. *Stochastic Systems*, 2(1):1–66, 2012.

[32] N. D. Vvedenskaya, R. L. Dobrushin, and F. I. Karpelevich. Queueing system with selection of the shortest of two queues: An asymptotic approach. *Problemy Peredachi Informatsii*, 32(1):20–34, 1996.

[33] Q. Xie, X. Dong, Y. Lu, and R. Srikant. Power of d choices for large-scale bin packing: A loss model. In *Proc. Ann. ACM SIGMETRICS Conf.*, 2015.

[34] L. Ying, R. Srikant, and X. Kang. The power of slightly more than one sample in randomized load balancing. In *Proc. IEEE Int. Conf. Computer Communications (INFOCOM)*, Hong Kong, 2015.