Beyond Scaling: Calculable Error Bounds of the Power-of-Two-Choices Mean-Field Model in Heavy-Traffic

Hairi
Arizona State University
Tempe, Arizona, USA
fhairi@asu.edu

Xin Liu
University of Michigan, Ann Arbor
Ann Arbor, Michigan, USA
xinliuee@umich.edu

Lei Ying
University of Michigan, Ann Arbor
Ann Arbor, Michigan, USA
leiying@umich.edu

ABSTRACT

This paper provides a recipe for deriving calculable approximation errors of mean-field models in heavy-traffic with the focus on the well-known load balancing algorithm — power-of-two-choices (Po2). The recipe combines Stein's method for linearized mean-field models and State Space Concentration (SSC) based on geometric tail bounds. In particular, our approach divides the state space into two regions, a neighborhood near the mean-field equilibrium and the complement of that. We first use a tail bound to show that the steady-state probability being outside the neighborhood is small. Then, we use a linearized mean-field model and Stein's method to characterize the generator difference, which provides the dominant term of the approximation error. From the dominant term, we are able to obtain an asymptotically-tight bound and a nonasymptotic upper bound, both are calculable bounds, not order-wise scaling results like most results in the literature. Finally, we compare the theoretical bounds with numerical evaluations to show the effectiveness of our results. We note that the simulation results show that both bounds are valid even for small size systems such as a system with only ten servers.

CCS CONCEPTS

• **Networks → Network performance analysis**; • **Mathematics of computing → Markov processes**.

KEYWORDS

Mean-field models, power-of-two-choices, heavy-traffic analysis, error bounds, Stein's method, state-space-concentration (SSC)

ACM Reference Format:
Hairi, Xin Liu, and Lei Ying. 2021. Beyond Scaling: Calculable Error Bounds of the Power-of-Two-Choices Mean-Field Model in Heavy-Traffic. In *The Twenty-second International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing (MobiHoc '21), July 26–29, 2021, Shanghai, China.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3466772.3467029

1 INTRODUCTION

Large-scale and complex stochastic systems have become ubiquitous, including large-scale data centers, the Internet of Things, and city-wide ride-hailing systems. Queueing theory has been a fundamental mathematical tool to model large-scale stochastic systems, and to analyze their steady-state performance. For example, steady-state analysis of load balancing algorithms in many-server systems is one of the most fundamental and widely-studied problems in queueing theory [11]. When the stationary (steady-state) distribution is known, the mean queue length and waiting time can be easily calculated, which reveal the performance of the system and can be used to guide the design of load balancing algorithms. However, for large-scale stochastic systems in general, it is extremely challenging (if not impossible) to characterize a system's stationary distribution due to the curse of dimensionality. For example, in a queueing system with N servers, each with a buffer of size b, the size of state space is at the order of N^b. Moreover, in such a system, the transition rate from a state to another may be state-dependent, so it becomes almost impossible to characterize its stationary distribution unless the system has some special properties such as when the stationary distribution has a product form [20]. To address these challenges, approximation methods such as mean-field models, fluid models, or diffusion models have been developed to study the stationary distributions of large-scale stochastic systems.

This paper focuses on stochastic systems with many agents, in particular, a many-server, many-queue system such as a large-scale data center. For such systems, mean-field models have been successfully used to approximate the stationary distribution of the system in the large-system limit (when the system size becomes infinity), see e.g. the seminal papers on power-of-two-choices [18, 22]. These earlier results, however, are asymptotic in nature by showing that the stationary distribution of the stochastic system weakly converges to the equilibrium point of the corresponding mean-field model as the system size becomes infinity. So these results do not provide the rate of convergence or the approximation error for finite-size stochastic systems. Furthermore, because of the asymptotic nature of the traditional mean-field model, it only applies to the light-traffic regime where the normalized load (load per server) is strictly less than the per-server capacity in the limit.

Both issues have been recently addressed using Stein's method. [24] studied the approximation error of the mean-field models in the light-traffic regime using Stein's method, and showed that for a large-class of mean-field models, the mean-square error is $O\left(\frac{1}{N}\right)$, where N is the number of agents in the system. [25] further extended Stein's method to mean-field models for heavy-traffic systems and developed a framework for quantifying the approximation

MobiHoc '21, July 26–29, 2021, Shanghai, China

ACM ISBN 978-1-4503-8558-9/21/07…$15.00
https://doi.org/10.1145/3466772.3467029

errors by connecting them to the local and global convergence of the mean-field models. While these results overcome the weakness of the traditional mean-field analysis, they only provide order-wise results, i.e. the scaling of the approximation errors in terms of the system size. Later, a refined mean-field analysis was developed in [6, 7]. In particular, [7] established the coefficient of the $\frac{1}{N}$ approximation error for light-traffic mean-field models, which provides an asymptotically exact characterization of the approximation error. However, the refined result in [7] is based on the light-traffic mean-field model and the analysis uses a limiting approach. Therefore, the result and method do not apply to heavy-traffic mean-field models.

This paper obtains calculable error bounds of heavy-traffic mean-field models. We consider the supermarket model [18], and assume jobs are allocated to servers according to a load balancing algorithm called power-of-two-choices [18, 22]. While the approximation error of this system has been studied in [25], it only characterizes the order of the error (in terms of the number of servers) and does not provide a calculable error bound. The difficulty of obtaining a calculable error bound is that the mean-field model of power-of-two-choices is a nonlinear system, so quantifying the convergence rate explicitly is difficult. In this paper, we overcome this difficulty by focusing on a linearized heavy-traffic mean-field model, linearized around its equilibrium point, so that we can explicitly solve Stein's equation. The linearized model cannot approximate the system well when the state of the system is not near the equilibrium point, which is further taken care of by using a geometric tail bound to show that such deviation only occurs with a small probability. The main results of this paper are summarized below.

- For the supermarket model with power-of-two-choices, we obtain two error bounds for the system in heavy-traffic. We first characterize the dominant term in the approximation error, which is a function of the Jacobian matrix of the mean-field model at its equilibrium and is asymptotically accurate. We then obtain a general upper bound which holds for finite size systems. We obtain the explicit forms of both bounds so they are calculable given the load and the system size.
- From the methodology perspective, the combination of state-space-concentration (SSC) and the linearized mean-field model provides a recipe for studying other mean-field models in heavy-traffic. The most difficult part of applying Stein's method for mean-field models is to establish the derivative bounds. While perturbation theory [12] provides a principled approach, we can only obtain order-wise results when facing nonlinear mean-field models (see e.g. [25]). Our approach is based on a basic hypothesis that if the mean-field solution would well approximate the steady-state of the stochastic system, then the steady-state should concentrate around the equilibrium point. Therefore, the focus should be on the mean-field system around its equilibrium point, which can be reduced to a linearized version. This basic hypothesis can be supported analytically using SSC based on the geometric tail bound [1]. In other words, the state-space concentration result leads to a linear system with a "solvable" Stein's equation, which is the key to applying Stein's method for steady-state approximation.

2 RELATED WORK

This section summarizes the related results in two categories. From the methodology perspective, this paper follows the line of research on using Stein's method for steady-state approximation of queueing systems introduced in [2, 3]. This paper uses Stein's method for mean-field approximations, which has been introduced in [24] and extended in [6, 7, 14, 25]. Stein's method for mean-field models (or fluid models) can also be interpreted as drift analysis based on integral Lyapunov functions, which was introduced in an earlier paper [21]. The combination of SSC and Stein's method was used in [2], which introduces Stein's method for steady-state diffusion approximation of queueing systems. The framework is later applied to mean-field (fluid) models where Stein's equation for a simplified one-dimensional mean-field models can be solved [15, 16]. In this paper, the linearized system is still a multi-dimensional system. SSC has been used in heavy-traffic analysis based on the Lyapunov-drift method, which was developed in [4] and used for analyzing computer systems and communication systems (see e.g. [17, 23]).

From the perspective of the power-of-two-choices load-balancing algorithm, for the light-traffic regime, [22] proved the weak convergence of the stationary distribution of power-of-two-choices to its mean field limit, the order-wise rate of convergence was established in [24], and [7] proposed a refined mean-field model with significantly smaller approximation errors. The scaling of queue lengths of power-of-two-choices in heavy-traffic has only been studied recently, first in [5] for finite-time analysis (transient analysis) and then in [25] for steady-state analysis. Our result was inspired by [7], which refines the mean-field model using the Jacobian matrix of the light-traffic mean-field equilibrium. Different from [7], based on state-space-concentration and linearized mean-field model, we established calculable error bounds for heavy-traffic mean-field models where the mean-field equilibrium and the associated Jacobian matrix are both functions of the system load and system size, which prevented us from using the asymptotic approach used in [7].

3 SYSTEM MODEL

In this section, we first introduce the well-known supermarket model under the power-of-two-choices load balancing algorithm. Our focus is the stationary distribution of such a system in the heavy-traffic regime (i.e. the load approaches to one as the number of servers increases). Then, we present the mean-field model, tailored for the N-server system [25] and the exact load of the N-server system. The solution to the mean-field model is an approximation of the stationary distribution of the stochastic system. We will then present the approach to characterize the approximation error based on Stein's equation.

Consider a many-server system with N homogeneous servers, where job arrivals follow a Poisson process with rate λN and service times are i.i.d. exponential random variables with rate one. Each server can hold at most b jobs, including the one in service. We consider $\lambda = 1 - \frac{\gamma}{N^\alpha}$ for some $0 < \gamma \leq 1$ and $\alpha \geq 0$. When $\alpha = 0$, λ is a constant independent of N which we call the light-traffic regime. When $\alpha > 0$, the arrival rate depends on N and approaches to one as $N \to \infty$, which we call the heavy-traffic regime. We assume the system is operated by a load balancing algorithm called power-of-two-choices [18, 22].

Power-of-Two-Choices (Po2): When a job arrives, Po2 samples two servers uniformly at random among N servers and dispatches the incoming job to the server with the shorter queue size. Ties are broken uniformly at random.

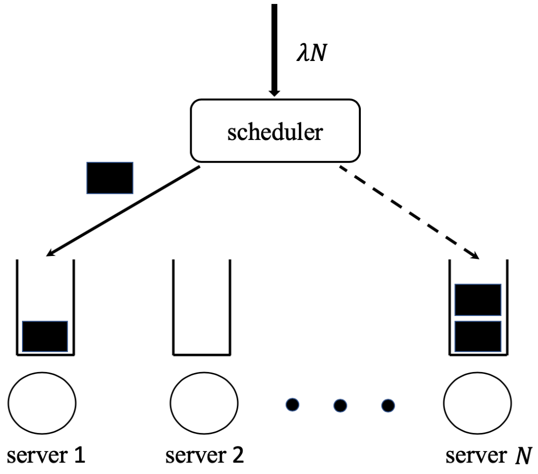

Figure 1: Power-of-Two-Choices

Let $S_i(t)$ denote the fraction of servers with queue size at least i at time t. The term $S_0(t) = 1, \forall t$ by definition. Under the finite buffer assumption with buffer size b, $S_i(t) = 0, \forall i \geq b+1, \forall t$. Throughout the paper, we assume that the buffer size b can be up to the order of $\log N$, i.e. $b = O(\log N)$. Define set $\mathcal{S}$ to be

$$\mathcal{S} = \{s \mid 1 \geq s_1 \geq \cdots \geq s_b \geq 0\},$$

and b-dimensional vector $S(t) = [S_1(t), S_2(t), \cdots, S_b(t)]$. It is easy to verify that the state $S(t)$ is a continuous time Markov chain (CTMC). Define e_k to be a b-dimensional vector such that the kth entry is one and all other entries are zero. Under Po2, the transition rate from state s and s' is as follows:

$$R_{s,s'} = \begin{cases} N(s_k - s_{k+1}), \text{if } s' = s - \frac{e_k}{N} \text{ and } 1 \leq k \leq b-1 \\ Ns_b, \text{if } s' = s - \frac{e_b}{N} \\ \lambda N(s_{k-1}^2 - s_k^2), \text{if } s' = s + \frac{e_k}{N} \\ \sum_{k=1}^{b} -\lambda N(s_{k-1}^2 - s_k^2) - N(s_k - s_{k+1}), \text{if } s' = s \\ 0, \text{otherwise} \end{cases}.$$

The first and second terms correspond to the event that a job departs from a server with queue size k so s_k decreases by $\frac{1}{N}$, and the third term corresponds to the event that a job arrives and joins a server with queue size $k-1$. We define a normalized transition rate to be

$$q_{s,s'} = \frac{R_{s,s'}}{N}.$$

We focus on the steady-state analysis of the system, i.e. the distribution of $S(\infty)$. At the steady-state, $S(\infty)$ is a b-dimensional random vector. For simplicity, let S denote $S(\infty)$. In this paper, we use uppercase letters for random variables and lowercase letters for deterministic values.

The mean-field model [18, 22, 25] for this system is

$$\dot{s} = f(s) = \sum_{s':s' \neq s} R_{s,s'}(s' - s) = N \sum_{s':s' \neq s} q_{s,s'}(s' - s).$$

According to the definition of $R_{s,s'}$ and $q_{s,s'}$, we have

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

The equilibrium point of this mean-field model, denoted by s^*, satisfies the following conditions:

$$s_0^* = 1 \tag{1a}$$

$$\lambda\left((s_{k-1}^*)^2 - (s_k^*)^2\right) - (s_k^* - s_{k+1}^*) = 0, \quad 1 \leq k \leq b-1 \tag{1b}$$

$$\lambda\left((s_{b-1}^*)^2 - (s_b^*)^2\right) - s_b^* = 0. \tag{1c}$$

The existence and uniqueness of the equilibrium point have been proved in [18]. Define

$$g(s) = -\int_0^\infty d(s(t), s^*)dt, \quad s(0) = s.$$

where $d(s(t), s^*)$ is a distance function. Then, by the definition of $g(s)$, we have

$$\nabla g(s) \cdot f(s) = d(s, s^*). \tag{2}$$

Equation (2) is called the Poisson equation or Stein's equation. For any bounded g, we have the following steady state equation (Basic Adjoint Relationship (BAR) [8])

$$\mathbb{E}[Gg(S)] = 0, \tag{3}$$

where the expectation is taken with respect to the steady state distribution of S and G is the generator of the CTMC. Combining (2) and (3), we have

$$\begin{aligned} \mathbb{E}\left[d(S, s^*)\right] =& \mathbb{E}\left[\nabla g(S) \cdot f(S) - Gg(S)\right] \\ =& -\mathbb{E}\left[\sum_{s'} R_{S,s'} \Gamma(S, s')\right], \end{aligned} \tag{4}$$

where $\Gamma(s, s') = g(s') - g(s) - \nabla g(s) \cdot (s' - s)$. From (4), Stein's method provides us a way to study the approximation error, defined by $\mathbb{E}[d(S, s^*)]$, by bounding the generator difference between the original system and the mean-field model.

4 MAIN RESULTS AND METHODOLOGY

This section summarizes our main results, which include an asymptotically tight approximation error bound and an upper bound that holds for finite N. We remark again these bounds can be calculated numerically and are not order-wise results as in most earlier papers.

Theorem 4.1 (**Asymptotically Tight Bound**). *For $0 < \alpha < \frac{1}{18}$, we have that*

$$\mathbb{E}[||S - s^*||^2] = -\frac{1}{N} \sum_{i=1}^{b} [J^T(s^*)]_{ii}^{-1} \tilde{f}_i(s^*) + o\left(\frac{1}{N^{1+\alpha}}\right) \tag{5}$$

where

$$J(s^*) = \begin{bmatrix} -2\lambda s_1^* - 1 & 1 & & 0 \\ 2\lambda s_1^* & \ddots & \ddots & \\ & \ddots & \ddots & 1 \\ 0 & & 2\lambda s_{b-1}^* & -2\lambda s_b^* - 1 \end{bmatrix}$$

is the Jacobian matrix of the mean-field model $f(s)$ at equilibrium point s^*,

$$\tilde{f}_i(s^*) = \frac{1}{2}\left(\lambda\left((s_{i-1}^*)^2 - (s_i^*)^2\right) + (s_i^* - s_{i+1}^*)\right)$$

for $i = 1, 2, \cdots, b$, and $[A]_{ij}^{-1}$ denotes the (i, j)th entry of the inverse of matrix A. □

The theorem states that the mean square error $\mathbb{E}[||S - s^*||^2]$ has an asymptotic dominant term $-\frac{1}{N}\sum_{i=1}^{b}[J^T(s^*)]_{ii}^{-1}\tilde{f}_i(s^*)$. Therefore, we have

$$\lim_{N\to\infty} N\mathbb{E}[||S - s^*||^2] = -\sum_{i=1}^{b}[J^T(s^*)]_{ii}^{-1}\tilde{f}_i(s^*). \tag{6}$$

Note that $\sum_{i=1}^{b}[J^T(s^*)]_{ii}^{-1}\tilde{f}_i(s^*)$ is negative, so the dominating term is positive.

Corollary 4.2 **(General Upper Bound)**. *For $0 < \alpha < \frac{1}{18}$ and a sufficiently large N, we have that*

$$\mathbb{E}[||S - s^*||^2] \leq -\frac{4}{N}\sum_{i=1}^{b}[J^T(s^*)]_{ii}^{-1}\tilde{f}_i(s^*). \tag{7}$$

□

This result tells us that we can have a calculable upper bound for heavy-traffic which holds for finite N.

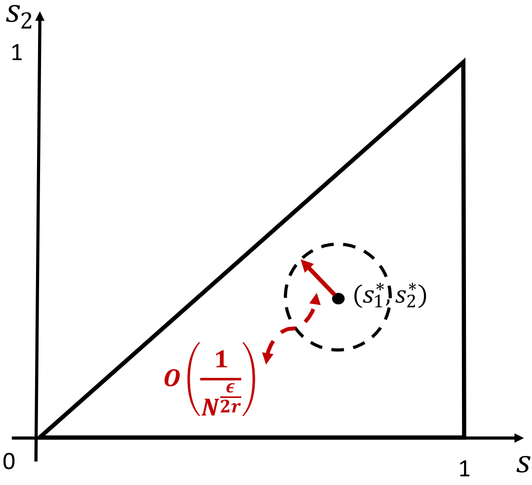

Figure 2: Illustration of the Inside and Outside Regions for $b = 2$

Our analysis combines Stein's method with a linear dynamical system and SSC. We divide the state space into two regions based on the mean-field solution: one region including those states "close" to the equilibrium point s^*, and the other region that includes all other states. For example, considering the case of $b = 2$, the state space is a two-dimensional region

$$\left\{(s_1, s_2) = \left(\frac{p}{N}, \frac{q}{N}\right)\middle| p \geq q \in \{0, 1, \cdots, N\}\right\} \subset [0, 1]^2.$$

As shown in Figure 2, we divide the state space into two regions separated by the dashed circle. The size of the circle is small and depends on N, in particular, the radius is $O\left(\frac{1}{N^{\frac{\epsilon}{2r}}}\right)$ where both ϵ and r are positive values (the choices of these two values will become clear in the analysis).

For the two different regions, we apply different techniques:

1. We first establish higher moment bounds that upper bound the probability that the steady state is outside the dashed circle. The proof is based on the geometric tail bound in [1, 10] and by showing that there is a "significant" negative drift that moves the system closer to the equilibrium point when the system is outside of the dashed circle.
2. For the states close to the equilibrium point, i.e, inside the dashed circle, from the control theory, we know that the mean-field nonlinear system behavior can be well approximated by the linearized dynamical system. By carefully choosing the parameters, we can look into the generator difference and calculate the dominant term of the approximation error by using the linearized mean-field model. The linearity enables us to solve Stein's equation, which is a key obstacle in applying Stein's method.

5 SIMULATIONS

Given $\alpha = 0.05$, we performed simulations for two different choices of γ and different system sizes. The purpose of these simulations is to compare the approximation errors calculated from the simulations with the asymptotically tight bound and the general upper bound. The results are based on the average of 10 runs, where each run simulates 10^9 time steps. We averaged over the last 9×10^8 time slots of each run to compute the steady state values.

For each run, we calculated the empirical mean square error multiplied by the system size N. Recall that the asymptotically tight bound and upper bound are

$$-\sum_{i=1}^{b}[J^T(s^*)]_{ii}^{-1}\tilde{f}_i(s^*)$$

and

$$-4\sum_{i=1}^{b}[J^T(s^*)]_{ii}^{-1}\tilde{f}_i(s^*),$$

respectively. Note that the two bounds only differ by a factor of four.

Table 1: $\gamma = 0.1, \alpha = 0.05$

N	10	100	1,000	10,000
λ	0.9109	0.9206	0.9292	0.9369
Simulation	4.2975	3.6884	3.9553	4.4068
Asymptotic Bound	3.2773	3.6411	4.0455	4.4955
Upper Bound	13.1092	14.5644	16.1820	17.9820

Table 2: $\gamma = 0.01, \alpha = 0.05$

N	10	100	1,000	10,000
λ	0.9911	0.9921	0.9929	0.9937
Simulation	77.9532	46.4641	38.7702	40.2093
Asymptotic Bound	28.2972	31.6293	35.3629	39.5457
Upper Bound	113.1888	126.5172	141.4516	158.1828

Tables 1 and 2 summarize the results with $\gamma = 0.1, \alpha = 0.05$ and $\gamma = 0.01, \alpha = 0.05$. We varied the size of the system in both cases. Note that the arrival rate is a function of the system size and approaches one as N increases. As N increases, the simulation results are in the same order with the dominant terms and are bounded by the upper bounds.

Our numerical results show that the asymptotic bound matches the empirical error very well, and approaches the empirical error as N increases. In particular, for $\gamma = 0.1$ and $\alpha = 0.05$, the results are close even when $N = 100$; and for $\gamma = 0.01$ and $\alpha = 0.05$, the results are close when $N = 1,000$.

As we can see, the upper bound is valid even for small size systems, e.g. $N = 10$, which shows the effectiveness of our results. From a practical point of view, both bounds are calculable, so together, they provide good estimates of the mean-square error.

6 PROOFS

In this section, we assume arrival rate is in the form of $\lambda = 1 - \frac{\gamma}{N^\alpha}$, which means the arrival rate is the function of system size N. As a result, generally the equilibrium point is also a function of N, a notation like $s^{*(N)}$ is a more proper way of describing the dependency on the system size. But for convenience, we still use s^* to denote the equilibrium point.

As we mentioned earlier, the results are established by looking at the system in two different regions, near the equilibrium point and outside. We next present our proof following this idea.

6.1 State Space Concentration

First, we present some preliminary convergence results in heavy-traffic for finite buffer size $b = O(\log N)$.

Lemma 6.1. *For any $0 < \alpha < 0.25$ and a sufficiently large N, we have*

$$\mathbb{E}[||S - s^*||^2] \leq \frac{1}{N^{1-4\alpha-7\xi}}$$

where $\xi > 0$ is an arbitrarily small number. □

Lemma 6.2 (Higher Moment Bounds). *For $r \in \mathbb{N}$ and a sufficiently large N, we have*

$$\mathbb{E}\left[||S - s^*||^{2r}\right] \leq \frac{1}{N^{r(1-4\alpha-7\xi)}}.$$

□

The proofs for both lemmas can be found in our technical report [9].

Lemma 6.3 (State Space Concentration). *Letting $\epsilon > 0$ and $r \in \mathbb{N}$, for a sufficiently large N, we have*

$$\mathbb{P}\left(||S - s^*||^{2r} \geq \frac{1}{N^\epsilon}\right) \leq \frac{1}{N^{r(1-4\alpha-7\xi)-\epsilon}}.$$

□

Proof. Applying the Markov inequality to the result in Lemma 6.2, we have

$$\begin{aligned}\mathbb{P}(||S - s^*||^{2r} \geq \frac{1}{N^\epsilon}) &\leq \frac{\mathbb{E}[||S - s^*||^{2r}]}{\frac{1}{N^\epsilon}} \\ &\leq \frac{N^\epsilon}{N^{r(1-4\alpha-7\xi)-\epsilon}} \\ &= \frac{1}{N^{r(1-4\alpha-7\xi)-\epsilon}}.\end{aligned}$$

□

6.2 Linear Mean-Field Model

Define a set of states to be $\mathcal{B} = \{s \mid ||s - s^*||^{2r} \leq \frac{1}{N^\epsilon}\}$, which are the states close to the equilibrium point. Let $d(s, s^*) = ||s - s^*||^2$ be the distance function. We consider a simple linear system

$$\dot{s} = l(s) = J(s^*)(s - s^*), \tag{8}$$

where $J(s^*)$ is the Jacobian matrix of $f(s)$ at the equilibrium point s^*. In heavy-traffic, the entries of $J(s^*)$ are functions of N as well when s^* is a function of N. The Jacobian matrix at s is

$$J(s) = \begin{bmatrix} -2\lambda s_1 - 1 & 1 & & 0 \\ 2\lambda s_1 & \ddots & \ddots & \\ & \ddots & \ddots & 1 \\ 0 & & 2\lambda s_{b-1} & -2\lambda s_b - 1 \end{bmatrix}.$$

We first introduce a lemma stating that matrix $J(s^*)$ is invertible, i.e. $J(s^*)^{-1}$ exists.

Lemma 6.4 (Invertibility). *For any $s \in \mathcal{S}$, the Jacobian matrix $J(s)$ is invertible.*

Proof. Since it is a tridiagonal matrix, we can write down the determinant in a recursive form for $i = 1, \cdots, b$,

$$P_i = -(2\lambda s_i + 1)P_{i-1} - 2\lambda s_{i-1}P_{i-2}$$

with initial values $P_0 = 1$ and $P_{-1} = 0$, where

$$P_i = \begin{vmatrix} -2\lambda s_1 - 1 & 1 & & 0 \\ 2\lambda s_1 & \ddots & \ddots & \\ & \ddots & \ddots & 1 \\ 0 & & 2\lambda s_{i-1} & -2\lambda s_i - 1 \end{vmatrix}.$$

Furthermore, we can verify that in fact, P_i can be written in the following form

$$P_i = (-1)^i - 2\lambda s_i P_{i-1} \tag{9}$$

with $P_1 = -(2\lambda s_1 + 1)$. We can draw two conclusions from equation (9), for any $s \in \mathcal{S}$:

- The sign of P_i alternates, i.e. when i is odd, $P_i < 0$; and when i is even, $P_i > 0$.
- The absolute value of P_i is no less than 1, i.e. $|P_i| \geq 1$.

Because the determinant is nonzero, $J(s)$ is invertible. □

Next we introduce a lemma on the solution to Stein's equation (the Poisson equation) for the linear mean-field system. Consider a function $g : \mathcal{S} \to \mathcal{S}$ such that it satisfies the following equation

$$Lg(s) \doteq \frac{dg(s)}{dt} = \nabla g(s) \cdot l(s) = ||s - s^*||^2. \tag{10}$$

According to the definition of the linear mean-field model in (8), we have

$$\nabla g(s) \cdot J(s^*)(s - s^*) = ||s - s^*||^2. \tag{11}$$

Lemma 6.5 (Solution to Stein's Equation). *The solution to the Poisson equation* (11) *satisfies*

$$\nabla g(s) = [J^T(s^*)]^{-1}(s - s^*), \tag{12}$$

and furthermore

$$\nabla^2 g(s) = [J^T(s^*)]^{-1} \quad \textit{and} \quad \nabla^3 g(s) = 0.$$

Proof. According to Stein's equation (11), we have

$$\nabla g(s)^T J(s^*)(s - s^*) = (s - s^*)^T (s - s^*),$$

which implies

$$\left[\nabla g(s)^T J(s^*) - (s - s^*)^T\right](s - s^*) = 0.$$

Since the equation has to hold for any s, we have

$$\nabla g(s)^T J(s^*) - (s - s^*)^T = 0,$$

which implies

$$\nabla g(s) = [J^T(s^*)]^{-1}(s - s^*).$$

The higher-order derivatives follow because $\nabla g(s)$ is a linear function of s. □

6.3 Proof of Theorem 4.1

We start from analyzing the generator difference when state S is close to s^*. In particular, we focus on

$$\mathbb{E}\left[Lg(S) - Gg(S) \,|S \in \mathcal{B}\right]. \tag{13}$$

Lemma 6.6. *The generator applying to function $g(s)$ satisfies*

$$Gg(s) = \nabla g(s) \cdot f(s) + \frac{1}{N}\sum_{i=1}^{b} \nabla^2 g(s)_{ii} \tilde{f}_i(s) \tag{14}$$

where $\nabla^2 g(s)_{ii}$ is the i-th diagonal element of the Hessian matrix $\nabla^2 g(s)$ and

$$\tilde{f}_i(s) = \frac{1}{2}[\lambda(s_{i-1}^2 - s_i^2) + (s_i - s_{i+1})].$$

Proof. According to the definition of generator G, we have

$$\begin{aligned} Gg(s) = &\sum_{i=1}^{b} \lambda N(s_{i-1}^2 - s_i^2)[g(s + e_i) - g(s)] \\ &+ N(s_i - s_{i+1})[g(s - e_i) - g(s)]. \end{aligned}$$

By the Taylor expansion at the state s, we have

$$\begin{aligned} Gg(s) = &\sum_{i=1}^{b} \lambda N(s_{i-1}^2 - s_i^2)[\nabla g(s) \cdot e_i + \frac{1}{2}e_i^T \nabla^2 g(s) e_i] \\ &+ N(s_i - s_{i+1})[\nabla g(s) \cdot (-e_i) + \frac{1}{2}e_i^T \nabla^2 g(s) e_i] \\ = &\sum_{i=1}^{b} \nabla g(s) \cdot [\lambda(s_{i-1}^2 - s_i^2) - (s_i - s_{i+1})]Ne_i \\ &+ \frac{1}{2}Ne_i^T \nabla^2 g(s) e_i [\lambda(s_{i-1}^2 - s_i^2) + (s_i - s_{i+1})] \\ = &\nabla g(s) \cdot f(s) + \frac{1}{N}\sum_{i=1}^{b} \nabla^2 g(s)_{ii} \tilde{f}_i(s). \end{aligned}$$

The first equality holds because $\nabla^3 g(s) = 0$ according to Lemma 6.5. □

Assume the state s is close to the equilibrium point s^* such that $||s - s^*||^{2r} \leq \frac{1}{N^\epsilon}$. We define

$$x_i = s_i - s_i^*$$

and obtain the Taylor expansion of $\tilde{f}_i(s)$ at the equilibrium point s^* as follows:

$$\begin{aligned} \tilde{f}_i(s) = &\frac{1}{2}[\lambda(s_{i-1}^2 - s_i^2) + (s_i - s_{i+1})] \\ = &\frac{\lambda}{2}[(s_{i-1}^* + x_{i-1})^2 - (s_i^* + x_i)^2] \\ &+ \frac{1}{2}(s_i^* + x_i - s_{i+1}^* - x_{i+1}) \\ = &\frac{\lambda}{2}[(s_{i-1}^*)^2 + 2x_{i-1}s_{i-1}^* + x_{i-1}^2 - (s_i^*)^2 - 2x_i s_i^* - x_i^2] \\ &+ \frac{1}{2}(s_i^* + x_i - s_{i+1}^* - x_{i+1}) \\ = &\frac{\lambda}{2}[(s_{i-1}^*)^2 - (s_i^*)^2] + \frac{1}{2}(s_i^* - s_{i+1}^*) + O(\frac{1}{N^{\frac{\epsilon}{2r}}}) \\ = &\tilde{f}_i(s^*) + O(\frac{1}{N^{\frac{\epsilon}{2r}}}), \end{aligned} \tag{15}$$

where the last equality holds because $||s - s^*||^{2r} \leq \frac{1}{N^\epsilon}$ implies $|x_i| \leq \frac{1}{N^{\frac{\epsilon}{2r}}}$.

Consider a state s, which is close to the equilibrium point, i.e. $||s - s^*||^{2r} \leq \frac{1}{N^\epsilon}$. According to Stein's equation (11) and the previous lemma, we have

$$\begin{aligned} &Lg(s) - Gg(s) \\ = &\nabla g(s) \cdot J(s^*)(s - s^*) - \nabla g(s) \cdot f(s) \\ &- \frac{1}{N}\sum_{i=1}^{b} \nabla^2 g(s)_{ii} \tilde{f}_i(s) \\ = &\nabla g(s) \cdot \left(J(s^*)(s - s^*) - f(s)\right) \\ &- \frac{1}{N}\sum_{i=1}^{b} \nabla^2 g(s)_{ii} \left(\tilde{f}_i(s^*) + O\left(\frac{1}{N^{\frac{\epsilon}{2r}}}\right)\right) \\ = &\nabla g(s) \cdot \left(J(s^*)(s - s^*) - f(s)\right) - \frac{1}{N}\sum_{i=1}^{b} \nabla^2 g(s)_{ii} \tilde{f}_i(s^*) \\ &- \frac{1}{N}\sum_{i=1}^{b} \nabla^2 g(s)_{ii} O\left(\frac{1}{N^{\frac{\epsilon}{2r}}}\right). \end{aligned}$$

According to Lemma 6.5, we have

$$\nabla g(s) = [J^T(s^*)]^{-1}(s - s^*) \text{ and } \nabla^2 g(s) = [J^T(s^*)]^{-1},$$

which are the functions of $J(s^*)$.

Since the mean-field is a second-order system, we have

$$\begin{aligned} f(s) =& f(s^*) + J(s^*)(s - s^*) \\ &+ \frac{1}{2} < s - s^*, \nabla^2 f(s^*)(s - s^*) >, \end{aligned}$$

where $\nabla^2 f(s^*)$ is the Hessian of $f(s)$ at equilibrium point. For any $s \in \mathcal{S}$ and $i = 1, \cdots, b$, the Hessian has the following form for $f_i(s)$

$$\nabla^2 f_i(s)_{kj} = \frac{\partial^2 f_i(s)}{\partial s_j \partial s_k} = \begin{cases} -2\lambda, & \text{if } j = k = i, \\ 2\lambda, & \text{if } j = k = i - 1, \\ 0, & \text{otherwise.} \end{cases}$$

Substituting it into the generator difference, we obtain

$$\begin{aligned} &\mathbb{E}\left[Lg(S) - Gg(S) \mid S \in \mathcal{B}\right] \\ =&\mathbb{E}\Big[[J^T(s^*)]^{-1}(S - s^*) \cdot (\frac{1}{2} < S - s^*, \nabla^2 f(s^*)(S - s^*) >) \\ &- \frac{1}{N}\sum_{i=1}^{b} \nabla^2 g(S)_{ii} \tilde{f}_i(s^*) - \frac{1}{N}\sum_{i=1}^{b} \nabla^2 g(S)_{ii} O(\frac{1}{N^{\frac{\epsilon}{2r}}}) \\ &\Big| S \in \mathcal{B}\Big]. \end{aligned} \tag{16}$$

This generator difference includes three terms. Note that $\nabla^2 g(s) = [J^T(s^*)]^{-1} = [J^{-1}(s^*)]^T$ according to Lemma 6.5.

We next introduce two lemmas about matrix $J^{-1}(s^*)$ that is involved in all three terms in Equation (16).

Lemma 6.7 (Upper Bound on the Entries of Matrix $J^{-1}(s^*)$). *For all $i, j = 1, \cdots, b$ and a sufficiently large N, we have*

$$|[J(s^*)]^{-1}_{ij}| \leq \frac{12}{\gamma} N^{2\alpha + 2\xi}.$$

Proof. First, we show that for any $\Phi \in R^b \setminus \{0\}$, we have

$$\frac{||J(s^*)\Phi||}{||\Phi||} \geq \delta_0$$

where $\delta_0 \geq \frac{\gamma}{12N^{2\alpha+2\xi}}$ is the absolute value of the negative drift of the original mean-field model, by Lemma 18 in [9].

Since $J(s^*)$ is a tridiagonal matrix that satisfies $J(s^*)_{i,i+1}J(s^*)_{i+1,i} > 0$ for all i, we know that $J(s^*)$ can be diagonalized and the eigenvalues are all real. Also, we know the eigenvalues are negative from the fact that $J(s^*)$ is a Hurwitz matrix.

Define the following Lyapunov functions

$$L_2(s) = \sqrt{\sum_{k=1}^{b}(s_k - s_k^*)^2}$$

$$L_w(s) = \sum_{k=1}^{b} w_k |s_k - s_k^*|$$

where $w_k \geq 1, k = 1, \cdots, b$ are defined in technical report [9]. First, it is easy to verify that the following inequality holds:

$$L_2(s) \leq L_w(s).$$

For the linear mean-field model $\dot{s}(t) = J(s^*)(s - s^*)$, we have the following exponential convergence result

$$L_2(s(t)) = \sqrt{\sum_{k=1}^{b}(s_k(t) - s_k^*)^2} \leq L_w(s(t)) \leq 3\exp(-\delta_0 t)$$

for $t \geq 0$. The proof for the second inequality is similar to the exponential convergence of the original mean-field system for power-of-two-choices, which can be found in Lemma 18 of our technical report [9].

Since $J(s^*)$ is diagonalizable, any vector in a b-dimensional space can be represented by a linear combination of the orthonormal eigenvectors r_k, for $k = 1, \cdots, b$, of the matrix $J(s^*)$. Suppose the eigenvalues are $\mu_1 \leq \mu_2 \leq \cdots \leq \mu_b < 0$. We can write the initial condition as

$$x \doteq s - s^* = \sum_{i=1}^{b} \alpha_i r_i$$

for some $\alpha_i \in R$ and $i = 1, \cdots, b$. Therefore, the general solution $s(t)$ of linear dynamical system $\dot{s}(t) = J(s^*)(s(t) - s^*)$ is a linear combination of the eigenvectors, i.e.

$$s(t) - s^* = \sum_{i=1}^{b} \alpha_i r_i \exp(\mu_i t).$$

So

$$L_2(s(t)) = ||\sum_{i=1}^{b} \alpha_i r_i \exp(\mu_i t)|| \leq 3\exp(-\delta_0 t).$$

Since this is true for all $x \in \mathbb{R}^b$, we can choose an initial condition such that $\alpha_i = 0$ for $i = 1, \cdots, b-1$ such that for all $t \geq 0$

$$L_2(s(t)) = ||\alpha_b \exp(\mu_b t)|| \leq 3\exp(-\delta_0 t).$$

Thus we conclude

$$\mu_b \leq -\delta_0.$$

As a result, for any $\Phi \in R^b \setminus \{0\}$, for some $\beta_i \in \mathbb{R}$ and $i = 1, \cdots, b$, we have

$$\begin{aligned} \Phi &= \beta_1 r_1 + \beta_2 r_2 + \cdots + \beta_b r_b \\ J(s^*)\Phi &= \beta_1 J(s^*) r_1 + \beta_2 J(s^*) r_2 + \cdots + \beta_b J(s^*) r_b \\ &= \beta_1 \mu_1 r_1 + \beta_2 \mu_2 r_2 + \cdots + \beta_b \mu_b r_b \end{aligned}$$

so

$$\frac{||J(s^*)\Phi||}{||\Phi||} = \frac{\sqrt{\sum_{i=1}^{b}\beta_i^2\mu_i^2}}{\sqrt{\sum_{i=1}^{b}\beta_i^2}} \geq \frac{\sqrt{\mu_b^2\sum_{i=1}^{b}\beta_i^2}}{\sqrt{\sum_{i=1}^{b}\beta_i^2}} = |\mu_b| \geq \delta_0.$$

Next, based on the results in [19] (in particular, by letting $x = y = 0$ for both diagonal elements Eq.(4.5) [19] and non-diagonal elements Eq.(4.7) [19]), we obtain an upper bound for any $i, j = 1, \cdots, b$

$$|[J(s^*)]^{-1}_{ij}| \leq \frac{1}{\delta_0} \leq \frac{12}{\gamma} N^{2\alpha+2\xi}.$$

□

Lemma 6.8 (Lower Bound on a Diagonal Entry of Matrix $J^{-1}(s^*)$). *For tridiagonal matrix $J^{-1}(s^*)$, we have that*

$$|J_{11}^{-1}(s^*)| \geq \frac{1}{3} \tag{17}$$

and for all $i = 1, \cdots, b$, we have $J_{ii}^{-1}(s^*) < 0$.

Proof. Suppose we have an $n \times n$ tridiagonal matrix G_n with entries denoted as follows

$$G_n = \begin{bmatrix} x_1 & y_1 & & 0 \\ z_1 & x_2 & \ddots & \\ & \ddots & \ddots & y_{n-1} \\ 0 & & z_{n-1} & x_n \end{bmatrix}.$$

We can define a backward continued fraction C_n [13] by the entries of G_n as follows

$$\begin{aligned} C_n &= [x_1 + \frac{-y_1 z_1}{x_2 +} \frac{-y_2 z_2}{x_3 +} \cdots \frac{-y_{n-1} z_{n-1}}{x_n}] \\ &= x_n + \cfrac{-y_{n-1} z_{n-1}}{x_{n-1} + \cfrac{-y_{n-2} z_{n-2}}{\ddots_{x_2 + \frac{-y_1 z_1}{x_1}}}}. \end{aligned}$$

Define sequence $\{P_n\}$ such that for $1 \leq k \leq n-1$

$$P_{k+1} = x_{k+1} P_k - y_k z_k P_{k-1}$$

and $P_0 = 1$ and $P_1 = x_1$. From the proof of Lemma 6.4, we know the sequence is also the iterative equation for the determinant of $J(s^*)$.

We introduce the following theorems in [13] to apply to our case.

Theorem 6.9. *Let the $n \times n$ tridiagonal matrix G_n have the form above. Let $G_n^{-1} = [w_{ij}]$ denote the inverse of G_n. Then*

$$w_{ii} = \frac{1}{C_i} + \sum_{k=i+1}^{n} (\frac{1}{C_k} \prod_{t=i}^{k-1} \frac{y_t z_t}{(C_t)^2}).$$

□

Theorem 6.10. *Let the matrix G_n be as above. Then for $n \geq 1$*

$$detG_n = P_n.$$

□

Theorem 6.11. *Consider a general backward continued function $A = [a_0 + \frac{b_1}{a_1 +} \frac{b_2}{a_2 +} \cdots \frac{b_n}{a_n}]$. If $0 \leq k \leq n$ and C_k is the kth backward convergent to A, i.e. $C_k = [a_0 + \frac{b_1}{a_1 +} \frac{b_2}{a_2 +} \cdots \frac{b_k}{a_k}]$, then $C_k = \frac{P_k}{P_{k-1}}$.* □

Thus some of the convergents of C_n are

$$C_1 = [x_1] = \frac{P_1}{P_0} = x_1,$$

$$C_2 = [x_1 + \frac{-y_1 z_1}{x_2}] = \frac{P_2}{P_1} = \frac{x_1 x_2 - y_1 z_1}{x_1}.$$

So in our case, we have that for $i = 1, \cdots, b-1$

$$y_i = 1$$
$$z_i = 2\lambda s_i^*$$

and for $i = 1, \cdots, b$

$$x_i = -2\lambda s_i^* - 1.$$

Therefore, we have

$$C_1 = x_1 = -2\lambda s_1^* - 1,$$

$$C_2 = \frac{x_1 x_2 - y_1 z_1}{x_1} = -2\lambda s_2^* - 1 + \frac{2\lambda s_1^*}{2\lambda s_1^* + 1} = -2\lambda s_2^* - \frac{1}{2\lambda s_1^* + 1}.$$

Note that sequence $\{P_n\}$ is the determinant of an $n \times n$ Jacobian matrix and we already know that the sign of P_n alternates, so $C_k = \frac{P_k}{P_{k-1}} < 0$ for all $k = 1, \cdots, b$ according to Theorem 6.11. Furthermore, from Theorem 6.9, we conclude that $J^{-1}(s^*)_{ii} < 0$ for all $i = 1, \cdots, b$. Furthermore, we have

$$J^{-1}(s^*)_{11} = \frac{1}{C_1} + \sum_{k=2}^{b} (\frac{1}{C_k} \prod_{t=1}^{k-1} \frac{2\lambda s_t^*}{(C_t)^2}) < 0$$

and

$$|J^{-1}(s^*)_{11}| \geq \frac{1}{|C_1|} \geq \frac{1}{3}$$

where the last inequality holds because $0 \leq s_1^* \leq 1$. This concludes the proof of Lemma 6.8. □

Based on Lemmas 6.7 and 6.8, we obtain the following lemmas to bound the terms in (16).

Lemma 6.12. *Given $||s - s^*||^{2r} \leq \frac{1}{N^\epsilon}$, we have*

$$\begin{aligned} &||[J^T(s^*)]^{-1}(s - s^*) \cdot < s - s^*, \nabla^2 f(s^*)(s - s^*) > || \\ &= O\left(\frac{1}{N^{\frac{3\epsilon}{2r} - 2\alpha - 3\xi}}\right). \end{aligned} \tag{18}$$

Proof. Consider the 2-norm of the first term in (16). We have

$$\begin{aligned} &||[J^T(s^*)]^{-1}(s - s^*) \cdot < s - s^*, \nabla^2 f(s^*)(s - s^*) > || \\ \leq &||[J^T(s^*)]^{-1}(s - s^*)|| || < s - s^*, \nabla^2 f(s^*)(s - s^*) > || \\ \leq &||[J^T(s^*)]^{-1}|| ||s - s^*|| || < s - s^*, \nabla^2 f(s^*)(s - s^*) > || \\ \leq &2\sqrt{2}\lambda ||[J^T(s^*)]^{-1}|| ||s - s^*||^3, \end{aligned} \tag{19}$$

where the third inequality holds because

$$\begin{aligned} &|| < s - s^*, \nabla^2 f(s^*)(s - s^*) > || \\ = &\sqrt{\sum_{i=1}^{b} [(s - s^*) \nabla^2 f_i(s^*)(s - s^*)]^2} \\ = &\sqrt{\sum_{i=1}^{b} \left(2\lambda[(s_{i-1} - s_{i-1}^*)^2 - (s_i - s_i^*)^2]\right)^2} \\ = &2\lambda \sqrt{\sum_{i=1}^{b} [(s_{i-1} - s_{i-1}^*)^2 - (s_i - s_i^*)^2]^2} \\ \leq &2\lambda \sqrt{\sum_{i=1}^{b} (s_{i-1} - s_{i-1}^*)^4 + (s_i - s_i^*)^4} \\ \leq &2\sqrt{2}\lambda \sqrt{\sum_{i=1}^{b} (s_i - s_i^*)^4} \\ \leq &2\sqrt{2}\lambda \sqrt{[\sum_{i=1}^{b} (s_i - s_i^*)^2]^2} = 2\sqrt{2}\lambda ||s - s^*||^2. \end{aligned}$$

Furthermore, from Lemma 6.7, for sufficiently large N, we have

$$\begin{aligned}||[J^T(s^*)]^{-1}|| = ||[J(s^*)]^{-1}|| \leq & \max_{ij} |[J(s^*)]^{-1}_{ij}| \times b \\ =& O(N^{2\alpha+2\xi}) \times O(\log N) \\ =& O(N^{2\alpha+3\xi}). \end{aligned} \tag{20}$$

Since $||s - s^*||^{2r} \leq \frac{1}{N^\epsilon}$, combining inequalities (19) and (20), we have

$$\begin{aligned} & ||[J^T(s^*)]^{-1}(s - s^*) \cdot < s - s^*, \nabla^2 f(s^*)(s - s^*) > || \\ \leq & 2\sqrt{2}\lambda \times O\left(N^{2\alpha+3\xi}\right) \times \frac{1}{N^{\frac{3\epsilon}{2r}}} = O\left(\frac{1}{N^{\frac{3\epsilon}{2r}-2\alpha-3\xi}}\right). \end{aligned}$$

□

Lemma 6.13. *Given $||s - s^*||^{2r} \leq \frac{1}{N^\epsilon}$, we have*

$$-\frac{1}{N}\sum_{i=1}^{b} \nabla^2 g_{ii}(s)\tilde{f}_i(s^*) \geq \frac{\lambda\gamma}{3N^{1+\alpha}}. \tag{21}$$

Proof. Recall that $\nabla^2 g(s) = [J^T(s^*)]^{-1}$ and for $i = 1, \cdots, b$, $J^{-1}(s^*)_{ii} < 0$ according to Lemma 6.8. It is easy to check that for $i = 1, \cdots, b$, $\tilde{f}_i(s^*) \geq 0$. Therefore, for $i = 1, \cdots, b$, we have

$$-\nabla^2 g(s)_{ii}\tilde{f}_i(s^*) \geq 0.$$

Furthermore, we also have

$$\begin{aligned}\tilde{f}_i(s^*) =& \frac{1}{2}[\lambda((s^*_{i-1})^2 - (s^*_i)^2) + (s^*_i - s^*_{i+1})] \\ =& \lambda[(s^*_{i-1})^2 - (s^*_i)^2],\end{aligned}$$

where the second equality holds because s^* is the equilibrium point. Thus, for $i = 1$ by equation (1b), we have

$$\tilde{f}_1(s^*) = \lambda[1 - (s^*_1)^2] \geq \lambda(1 - \lambda^2) \geq \lambda(1 - \lambda) = \frac{\lambda\gamma}{N^\alpha}$$

which implies

$$\frac{1}{N}\sum_{i=1}^{b} \nabla^2 g_{ii}(s)\tilde{f}_i(s^*) \geq -\frac{1}{N}J^{-1}_{11}(s^*)\tilde{f}_1(s^*) \geq \frac{\lambda\gamma}{3N^{1+\alpha}}.$$

□

Lemma 6.14. *Given $||s - s^*||^{2r} \leq \frac{1}{N^\epsilon}$, we have*

$$-\frac{1}{N}\sum_{i=1}^{b} \nabla^2 g_{ii}(s)\tilde{f}_i(s^*) = O(\frac{1}{N^{1-2\alpha-3\xi}}). \tag{22}$$

Proof. It is easy to check that $\tilde{f}_i(s^*) \leq 1$ for $i = 1, \cdots, b$. Recall that $|\nabla^2 g(s)_{ii}| \leq O(N^{2\alpha+2\xi})$. Therefore, we have

$$-\frac{1}{N}\sum_{i=1}^{b} \nabla^2 g_{ii}(s)\tilde{f}_i(s^*) = \frac{b}{N}O(N^{2\alpha+2\xi}) = O(\frac{1}{N^{1-2\alpha-3\xi}}).$$

□

Lemma 6.15. *Given $||s - s^*||^{2r} \leq \frac{1}{N^\epsilon}$, we have that for a sufficiently large N,*

$$\left\|-\frac{1}{N}\sum_{i=1}^{b} \nabla^2 g(s)_{ii} O(\frac{1}{N^{\frac{\epsilon}{2r}}})\right\| = O\left(\frac{1}{N^{1+\frac{\epsilon}{2r}-2\alpha-3\xi}}\right). \tag{23}$$

Proof. Recall that $|\nabla^2 g(s)_{ii}| = O(N^{2\alpha+2\xi})$ for $i = 1, \cdots, b$. Thus, we have

$$\begin{aligned} & || -\frac{1}{N}\sum_{i=1}^{b} \nabla^2 g(s)_{ii} O(\frac{1}{N^{\frac{\epsilon}{2r}}})|| \\ \leq & \frac{b}{N}O(N^{2\alpha+2\xi}) \cdot O(\frac{1}{N^{\frac{\epsilon}{2r}}}) = O\left(\frac{1}{N^{1+\frac{\epsilon}{2r}-2\alpha-3\xi}}\right). \end{aligned}$$

□

Based on these lemmas, we are now able to characterize the generator difference when state S is close to s^*.

Lemma 6.16. *For $0 < \alpha < \frac{1}{18}$ and a sufficiently large N, we have*

$$\begin{aligned} & \mathbb{E}\left[Lg(S) - Gg(S) \mid S \in \mathcal{B}\right] \\ =& -\frac{1}{N}\sum_{i=1}^{b}[J^T(s^*)]^{-1}_{ii}\tilde{f}_i(s^*) + o\left(\frac{1}{N^{1+\alpha}}\right), \end{aligned} \tag{24}$$

with the following choice of parameters

$$\frac{3(1+\alpha+\xi)}{1-18\alpha-27\xi} < r, \tag{25}$$

$$\frac{2r(1+3\alpha+3\xi)}{3} < \epsilon < r(1-4\alpha-7\xi) - 1 - \alpha - \xi. \tag{26}$$

Proof. Under the conditions of the lemma, it is easy to check that the upper bounds on the first and third terms in equation (16) are order-wise smaller than the lower bound on the second term, i.e.

$$\frac{3\epsilon}{2r} - 2\alpha - 3\xi > 1 + 3\alpha + 3\xi - 2\alpha - 3\xi = 1 + \alpha$$

and

$$\begin{aligned} 1 + \frac{\epsilon}{2r} - 2\alpha - 3\xi >& 1 + \frac{1}{3} + \alpha + \xi - 2\alpha - 3\xi \\ >& (1+\alpha) + (\frac{2}{9} - 2\xi), \end{aligned}$$

where the last inequality is due to the fact $0 < \alpha < \frac{1}{18}$. Therefore, the lemma holds. □

We also remark that there exist parameters that satisfy the conditions in the lemma because the right-hand side of ϵ in (26) is larger than the left-hand side given that the r satisfies (25), where r has to be large enough. For example, when $\alpha = 0.05$, r needs to at least 32 and ϵ can be 24.54. It is easy to check that we can find a small enough ξ.

6.3.1 *Proof of Theorem 4.1.*

We again choose parameters that satisfy the following conditions:

$$\frac{3(1+\alpha)}{1-18\alpha-27\xi} < r$$

$$\frac{2r(1+3\alpha+3\xi)}{3} < \epsilon < r(1-4\alpha-7\xi) - 1 - \alpha - \xi$$

and $\xi > 0$ is arbitrarily small. Then, for sufficiently large N, the mean square distance is

$$\begin{aligned} & \mathbb{E}[||S - s^*||^2] \\ =& \mathbb{E}\left[||S - s^*||^2 \middle| S \notin \mathcal{B}\right]\mathbb{P}(S \notin \mathcal{B}) + \mathbb{E}\left[||S - s^*||^2 \middle| S \in \mathcal{B}\right]\mathbb{P}(S \in \mathcal{B}) \end{aligned}$$

$$
\begin{aligned}
=&O(\log N) \times O(\frac{1}{N^{r(1-4\alpha-7\xi)-\epsilon}}) + \left(-\frac{1}{N}\sum_{i=1}^{b} \nabla^2 g(s)_{ii}\tilde{f}_i(s^*)\right. \\
&\left.+o(\frac{1}{N^{1+\alpha}})\right) \times \left(1 - O(\frac{1}{N^{r(1-4\alpha-7\xi)-\epsilon}})\right) \\
=&O\left(\frac{1}{N^{r(1-4\alpha-7\xi)-\epsilon-\xi}}\right) - \frac{1}{N}\sum_{i=1}^{b} \nabla^2 g(s)_{ii}\tilde{f}_i(s^*) \\
&+ O\left(\frac{1}{N^{1-2\alpha-3\xi}}\right) \times O\left(\frac{1}{N^{r(1-4\alpha-7\xi)-\epsilon}}\right) + o\left(\frac{1}{N^{1+\alpha}}\right) \\
=&-\frac{1}{N}\sum_{i=1}^{b}[J^T(s^*)]_{ii}^{-1}\tilde{f}_i(s^*) + o\left(\frac{1}{N^{1+\alpha}}\right),
\end{aligned}
$$

where the second equality holds because $||s-s^*||^2 \leq b = O(\log N)$. Note that with the choice of parameters r, ϵ and $0 < \alpha < \frac{1}{18}$, the lower bound on the term $-\frac{1}{N}\sum_{i=1}^{b}[J^T(s^*)]_{ii}^{-1}\tilde{f}_i(s^*) + o(\frac{1}{N^{1+\alpha}})$ is $O(\frac{1}{N^{1+\alpha}})$ while other terms are strictly upper bounded by this order for sufficiently large N.

6.4 Proof of Corollary 4.2

From Lemma 6.16 with the same parameter choices, it is easy to check that for a sufficiently large N, we have

$$\mathbb{E}\left[Lg(S) - Gg(S)\,|S \in \mathcal{B}\right] \leq -\frac{3}{N}\sum_{i=1}^{b}[J^T(s^*)]_{ii}^{-1}\tilde{f}_i(s^*).$$

Also, the following inequality holds for sufficiently large N

$$\mathbb{P}\left(||S - s^*||^{2r} \geq \frac{1}{N^{\epsilon}}\right) \leq \frac{1}{N^{r(1-4\alpha-\xi')-\epsilon}} \leq \frac{1}{N^{1+\alpha+\xi}}.$$

Then from the above two inequalities, for a sufficiently large N, the mean square distance is

$$
\begin{aligned}
&\mathbb{E}[||S - s^*||^2] \\
=&\mathbb{E}\left[Lg(S) - Gg(S)\middle|S \notin \mathcal{B}\right]\mathbb{P}\left(S \notin \mathcal{B}\right) \\
&+ \mathbb{E}\left[Lg(S) - Gg(S)\middle|S \in \mathcal{B}\right]\mathbb{P}\left(S \in \mathcal{B}\right) \\
\leq&\frac{b}{N^{1+\alpha+\xi}} - \frac{3}{N}\sum_{i=1}^{b}[J^T(s^*)]_{ii}^{-1}\tilde{f}_i(s^*) \\
\leq&\frac{1}{N^{1+\alpha}} - \frac{3}{N}\sum_{i=1}^{b}[J^T(s^*)]_{ii}^{-1}\tilde{f}_i(s^*) \\
\leq&-\frac{4}{N}\sum_{i=1}^{b}[J^T(s^*)]_{ii}^{-1}\tilde{f}_i(s^*),
\end{aligned}
$$

where the second from the last inequality holds because the first term is larger than the right-hand side of inequality (21).

7 CONCLUSION

In this paper, we established calculable bounds on the mean-square errors of the power-of-two-choices mean-field model in heavy-traffic. Our approach combined SSC and Stein's method with a linearized mean-field models, and characterized the dominant term of the mean square error. Our simulation results confirmed the theoretical bounds and showed that the bounds are valid even for small size systems such as when $N = 10$. This recipe of combining SSC and Stein's method for linearized mean-field model can be applied to other mean-field models beyond the power-of-two-choices load balancing algorithm.

ACKNOWLEDGEMENTS

The authors are very grateful to Nicolas Gast for his valuable comments. This work was supported in part by NSF ECCS 1739344, CNS 2002608 and CNS 2001687.

REFERENCES

[1] D. Bertsimas, D. Gamarnik, and J. N. Tsitsiklis. 2001. Performance of Multiclass Markovian Queueing Networks Via Piecewise Linear Lyapunov Functions. *Adv. in Appl. Probab.* (2001).

[2] Anton Braverman and J. G. Dai. 2017. Stein's method for steady-state diffusion approximations of $M/Ph/n + M$ systems. *Ann. Appl. Probab.* 27, 1 (02 2017), 550–581. https://doi.org/10.1214/16-AAP1211

[3] A. Braverman, J. G. Dai, and J. Feng. 2016. Stein's method for steady-state diffusion approximations: an introduction through the Erlang-A and Erlang-C models. *Stochastic Systems* 6 (2016), 301–366.

[4] Atilla Eryilmaz and R. Srikant. 2012. Asymptotically tight steady-state queue length bounds implied by drift conditions. *Queueing Syst.* 72, 3-4 (Dec. 2012), 311–359.

[5] Patrick Eschenfeldt and David Gamarnik. 2016. Supermarket queueing system in the heavy traffic regime. Short queue dynamics. *arXiv preprint arXiv:1610.03522* (2016).

[6] Nicolas Gast. 2017. Expected Values Estimated via Mean-Field Approximation Are 1/N-Accurate. *Proc. ACM Meas. Anal. Comput. Syst.* 1, 1 (June 2017), 17:1–17:26. https://doi.org/10.1145/3084454

[7] Nicolas Gast and Benny Van Houdt. 2018. A Refined Mean Field Approximation. In *Proc. Ann. ACM SIGMETRICS Conf.* Irvien, CA. https://doi.org/10.1145/3152542

[8] Peter W Glynn and Assaf Zeevi. 2008. Bounding stationary expectations of Markov processes. In *Markov processes and related topics: a Festschrift for Thomas G. Kurtz.* Institute of Mathematical Statistics, 195–214.

[9] Hairi, Xin Liu, and Lei Ying. 2020. Beyond Scaling: Calculable Error Bounds of the Power-of-Two-Choices Mean-Field Model in Heavy-Traffic. (2020). arXiv:2012.06613 [cs.PF]

[10] B. Hajek. 1982. Hitting-time and occupation-time bounds implied by drift analysis with applications. *Ann. Appl. Prob.* (1982), 502–525.

[11] Mor Harchol-Balter. 2013. *Performance Modeling and Design of Computer Systems: Queueing Theory in Action.* Cambridge University Press. https://doi.org/10.1017/CBO9781139226424

[12] Hassan K Khalil. 2001. *Nonlinear systems.* Prentice Hall.

[13] Emrah Kılıç. 2008. Explicit formula for the inverse of a tridiagonal matrix by backward continued fractions. *Appl. Math. Comput.* 197, 1 (2008), 345–357.

[14] Xin Liu and Lei Ying. 2018. On Achieving Zero Delay with Power-of-d-Choices Load Balancing. In *Proc. IEEE Int. Conf. Computer Communications (INFOCOM).* Honolulu,Hawaii.

[15] Xin Liu and Lei Ying. 2019. On universal scaling of distributed queues under load balancing. *arXiv preprint arXiv:1912.11904* (2019).

[16] Xin Liu and Lei Ying. 2020. Steady-State Analysis of Load Balancing Algorithms in the Sub-Halfin-Whitt Regime. *J. Appl. Probab.* 57, 2 (June 2020), 578 – 596.

[17] Siva Theja Maguluri and R. Srikant. 2016. Heavy traffic queue length behavior in a switch under the MaxWeight algorithm. *Stoch. Syst.* 6, 1 (2016), 211–250.

[18] M. Mitzenmacher. 1996. *The Power of Two Choices in Randomized Load Balancing.* Ph.D. Dissertation. University of California at Berkeley.

[19] Peter D Robinson and Andrew J Wathen. 1992. Variational bounds on the entries of the inverse of a matrix. *IMA journal of numerical analysis* 12, 4 (1992), 463–486.

[20] R. Srikant and Lei Ying. 2014. *Communication Networks: An Optimization, Control and Stochastic Networks Perspective.* Cambridge University Press.

[21] Alexander Stolyar. 2015. Tightness of stationary distributions of a flexible-server system in the Halfin-Whitt asymptotic regime. *Stoch. Syst.* 5, 2 (2015), 239–267.

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

[23] Weina Wang, Siva Theja Maguluri, R Srikant, and Lei Ying. 2018. Heavy-traffic delay insensitivity in connection-level models of data transfer with proportionally fair bandwidth sharing. *ACM SIGMETRICS Performance Evaluation Review* 45, 3 (2018), 232–245.

[24] Lei Ying. 2016. On the Approximation Error of Mean-Field Models. In *Proc. Ann. ACM SIGMETRICS Conf.* Antibes Juan-les-Pins, France.

[25] Lei Ying. 2017. Stein's Method for Mean Field Approximations in Light and Heavy Traffic Regimes. *Proc. ACM Meas. Anal. Comput. Syst.* 1, 1, Article 12 (June 2017), 12:1–12:27 pages. https://doi.org/10.1145/3084449