# On Achieving Zero Delay with Power-of-$d$-Choices Load Balancing

Xin Liu and Lei Ying

School of Electrical, Computer and Energy Engineering,
Arizona State University, Tempe, AZ, United States, 85287
Email: {xliu272, lei.ying.2}@asu.edu

***Abstract*—Power-of-$d$-choices is a popular load balancing algorithm for many-server systems such as large-scale data centers. For each incoming job, the algorithm probes $d$ servers, chosen uniformly at random from a total of $N$ servers, and routes the job to the least loaded one. It is well known that power-of-$d$-choices reduces queueing delays by orders of magnitude compared to the policy that routes each incoming job to a randomly selected server. The question to be addressed in this paper is how large $d$ needs to be so that power-of-$d$-choices achieves asymptotic zero delay like the join-the-shortest-queue (JSQ) algorithm, which is a special case of power-of-$d$-choices with $d = N$. We are interested in the heavy-traffic regime where the load of the system, denoted by $\lambda$, approaches to one as $N$ increases, and assume $\lambda = 1 - \gamma N^{-\alpha}$ for $0 < \gamma < 1$ and $0 \leq \alpha < 1/6$. This paper establishes that when $d = \omega\left(\frac{1}{1-\lambda}\right)$, the probability that an incoming job is routed to a busy server is asymptotically zero, i.e. a job experiences zero queueing delay with probability one asymptotically; and when $d = O\left(\frac{1}{1-\lambda}\right)$, the probability that a job is routed to a busy server is lower bounded by a positive constant independent of $N$. Therefore, our results show that $d = \omega(\frac{1}{1-\lambda})$ is sufficient and almost necessary for achieving zero delay with the power-of-$d$-choices load balancing policy.**

## I. Introduction

Load balancing in many-server systems (such as data centers) routes incoming jobs to servers to balance the load across servers and to minimize response times to improve user experience. Small-scale data centers often use centralized load-balancing algorithms based on complete information, such as the join-the-shortest-queue (JSQ) algorithm. JSQ continuously monitors the states of all servers and routes each incoming job to the least loaded server in the system. While JSQ has been shown to be delay optimal in a number of different settings (see [21], [20], [5] and references within); continuously collecting information from all servers incurs prohibitive messaging and computational overhead and is not feasible in large-scale data centers.

As cloud computing and big-data analytics, both relying on large-scale data centers, are playing increasingly important roles in enterprise and personal computing, there has been a renewed interest in load balancing algorithms with low messaging overhead for large-scale many-server systems. A popular load balancing algorithm based on incomplete information and with low messaging overhead is the so called power-of-$d$-choices (also known as JSQ($d$)), proposed in [13], [19]. When a job arrives, power-of-$d$-choices probes $d$ servers uniformly at random and dispatches the job to the server with the shortest queue among the $d$ servers. Compared with randomized routing that dispatches a task to a randomly chosen server (named RAND in this paper), power-of-$d$-choices reduces the mean sojourn time from $\frac{1}{1-\lambda}$ to $\log_d \frac{1}{1-\lambda}$, so reducing response times by orders of magnitude.

While power-of-$d$-choices performs much better than RAND with low messaging overhead; for a constant $d$ (i.e. independent of $N$), queueing delay, which is the time from which a job enters the system to the time at which the job starts to be processed, is lower bounded by a positive constant for any $\lambda$ and $N$. Low delay (i.e. short response time) is very important in modern data centers. It has been reported [16] that an extra delay of 500 ms led to 1.2% loss of users and revenue. The importance of low delay has motivated the following question: *can power-of-d-choices, with carefully chosen* $d$, *achieve zero delay?* Assume $\lambda = 1-\gamma N^{-\alpha}$. In a recent paper [14], it has been shown that power-of-$d$-choices becomes JSQ at the fluid level when $d = \omega(1)$, and at the diffusion level with $d = O(\sqrt{N} \log N)$. According to [5], in the large-system limit, JSQ achieves zero queueing delay at both the fluid and diffusion levels. Motivated by this recent work [14], this paper studies the fundamental requirement on $d$, as a function of the load of the system ($\lambda$), to achieve asymptotic zero delay. The main results are summarized below.

### A. Main Results

Consider power-of-$d$-choices load balancing in a system with $N$ homogeneous servers. Assume Poisson arrival and exponential service times. Let $\mathcal{W}_N$ denote the event that an incoming job is routed to a busy server in a system with $N$ servers, and $p_{\mathcal{W}_N}$ denote the probability of this event at the steady-state. Assuming $0 \leq \alpha < 1/6$ and $d = \omega\left(\frac{1}{1-\lambda}\right)$, we will prove that

$$\lim_{N\to\infty} p_{\mathcal{W}_N} = 0. \tag{1}$$

In other words, almost all jobs are served immediately upon arrival at the steady-state and experience zero queueing delay. This main result is proved using mean-field analysis (fluid-limit analysis) based on Stein's method [18], [22], [23], [8], [9].

Assuming $0 \leq \alpha < 1/6$ and $d = O\left(\frac{1}{1-\lambda}\right)$, we further show that

$$\lim_{N\to\infty} p_{\mathcal{W}_N} > 0. \tag{2}$$

In other words, the probability of being queued is nonzero. In summary, our results show that to achieve asymptotic zero delay under power-of-$d$-choices, $d$ should scale super-linearly with $\frac{1}{1-\lambda}$. The main results are also summarized in Figure 1.

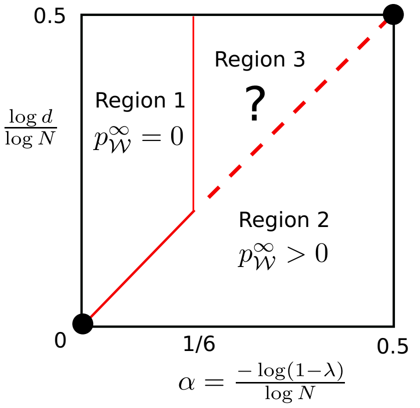

0.5
$\frac{\log d}{\log N}$
Region 1
$p_{\mathcal{W}}^{\infty} = 0$
Region 3
?
Region 2
$p_{\mathcal{W}}^{\infty} > 0$
0
1/6
0.5
$\alpha = \frac{-\log(1-\lambda)}{\log N}$

Fig. 1: This figure illustrates the main results in a two-dimensional region, where $x$-axis represents $\alpha = \frac{-\log(1-\lambda)}{\log N}$, $y$-axis represents $\frac{\log d}{\log N}$, and $p_{\mathcal{W}}^{\infty} = \lim_{N\to\infty} p_{\mathcal{W}_N}$ is the asymptotic probability of being routed to a busy server. Our results show that the asymptotic delay is zero in regime 1 (i.e. when $d = \omega\left(\frac{1}{1-\lambda}\right)$) and nonzero in regime 2 (i.e. when $d = O\left(\frac{1}{1-\lambda}\right)$). We also conjecture that the same results hold for regime 3, which, however, remain to be proved.

We would like to remark that the scaling $d = \omega\left(\frac{1}{1-\lambda}\right)$ is quite intuitive. Note that at steady-state, $1-\lambda$ fraction of servers are expected to be idle due to the work conservation law. Suppose exactly $1-\lambda$ fraction of servers are idle when a job arrives. Then the probability that at least one of the $d$ probed server is idle is

$$1 - \lambda^d = 1 - \left(1 - \gamma N^{-\alpha}\right)^d.$$

Therefore, it is easy to see that $d = \omega(\frac{1}{1-\lambda}) = \omega\left(N^{\alpha}\right)$ probes are necessary to find an idle server with a high probability. However, a rigorous proof is highly nontrivial because the fraction of servers that are idle is a stochastic process instead of a constant. To see this, consider the $M/M/N$ queue (Erlang-C model) consisting of $N$ servers and a global FIFO queue. When a job arrives, it joins the global queue; and when a server completes a job, it fetches a new job from the global queue (if the queue is nonempty). The delay of the $M/M/N$ queue provides a universal lower bound on the delay of the systems that maintain distributed queues (one per server) like JSQ and power-of-$d$-choices. A celebrated result by Halfin and Whitt [10] shows that the asymptotic delay is lower bounded by a positive constant when $\alpha \geq 0.5$. Therefore, when $\lambda = 1 - \gamma N^{-\alpha}$ and $\alpha \geq 0.5$, no load balancing policy can achieve asymptotic zero delay. This result demonstrates that while the scaling $d = O(\frac{1}{1-\lambda})$ is intuitive, it is not sufficient. In fact, from the best of our knowledge, This paper is the *first* one that shows power-of-$d$-choices can achieve asymptotic zero delay *at steady-state in the heavy-traffic regime.* [14] considered the Halfin-Whitt heavy-traffic regime, but the result only proves when considering a *finite* time interval, the diffusion limit of power-of-$d$-choices coincides with that of JSQ when $d = O\left(\sqrt{N}\log N\right)$. So it is for the transient region, not a steady-state result. Diffusion analysis of JSQ [5] suggests that most queues are either zero or one. The result [14] strongly suggests that power-of-$d$-choices achieves zero delay; but technically, it does not result in any bound on the queueing delay at steady-state.

### B. Related Work

Power-of-$d$-choices was first studied in [13], [19] assuming Poisson arrivals, exponential service times and homogeneous servers. The analysis has been extended to general service distributions [1] and heterogeneous servers [15]. These results were established in the light traffic regime (i.e. $\alpha = 0$ in $\lambda = 1 - \gamma N^{-\alpha}$). In the light traffic regime, it has been shown in the recent paper [14] that the fluid limit of power-of-$d$-choices coincides with that of JSQ when $d = \omega(1)$. Heavy traffic analysis of power-of-$d$-choices ($\alpha > 0$) started only very recently. [6] established the process-level convergence of power-of-$d$-choices in heavy traffic over a *finite* time interval when $d$ is a constant (i.e. $d = O(1)$), and conjectured that most queues are of size $\alpha \log_d N$. [23] resolved the conjecture in [5] for power-of-2-choices (i.e. $d = 2$) with $0 < \alpha < 0.2$. [4] also considered power-of-$d$-choices in the heavy-traffic regime but assumed $d = o\left(\frac{1}{1-\lambda}\right)$, which is not sufficient to achieve zero delay. [14] is the first paper that shows the equivalence of power-of-$d$-choices and JSQ at the diffusion scale when $d = O(\sqrt{N}\log N)$. As we mentioned earlier, the result was proved for a finite time interval, not for steady-state distributions. This paper was motivated by [14] and shows that $d = \omega\left(\frac{1}{1-\lambda}\right)$ is a sufficient and almost necessary condition for achieving asymptotic zero delay for $0 \leq \alpha < 1/6$. The proof is based on Stein's method for mean-field (fluid) models used in [18], [22], [23], [8], [9]. We also refer the reader to [3], [2] for their pioneering work on Stein's method for steady-state diffusion approximation.

We finally remark that a number of new load balancing algorithms have been developed recently. For example, another class of load-balancing policies that have superior delay performance are PULL-based policy [17], [12] such as the Join-the-Idle-Queue (JIQ) policy. In the light-traffic regime, JIQ has similar performance with JSQ in the mean-field limit. A comprehensive analysis of the delay, memory, and messaging tradeoffs of PULL-based policies can be found in [7]. The focus of this paper is not to develop new load balancing algorithms but to study the fundamental limit of the popular power-of-$d$-choices policy for achieving asymptotic zero delay.

## II. Notation and Assumptions

- $|x| = \sum_{k=1}^{b} |x_k|$ and $\|x\| = \sqrt{\sum_{k=1}^{b} x_k^2}$ are 1-norm and 2-norm of $b$-dimensional vector $x$, respectively.
- $l(N) = \omega(1)$ and $l(N) = o(\log N)$, i.e. $l(N)$ diverges to $\infty$ as $N$ increases, but the divergence rate can be arbitrarily slow and is slower than $\log N$.
- The arrival process is Poisson with rate $\lambda N$ with $\lambda = 1 - \gamma N^{-\alpha}$, where $0 < \gamma < 1$.
- The process time of each job is exponentially distributed with rate one.
- Finite buffer at each server with buffer size $b$ which is independent of $N$.
- $1_k$ : a $b$-dimensional vector with only the $k$th entry is 1 and the rest are 0.
- $d = \frac{l(N)}{1-\lambda} = \gamma^{-1} l(N) N^{\alpha}$.
- $n = N^{\alpha+\xi}$ with an arbitrarily small $\xi > 0$.

## III. The Mean-Field Model of Power-of-$d$-Choices

We consider power-of-$d$-choices for many-server systems introduced in [13], [19]. The system consists of $N$ identical servers, and each server has a separate queue as shown in Figure 2, and each queue can hold at most $b$ jobs (i.e. finite buffer systems), including the one in service. Assume jobs arrive at the system following a Poisson process with rate $\lambda N$ and the processing time of each job is exponentially distributed with mean $\mu = 1$. Let $Q_k(t)$ denote the queue size of server $k$ at time $t$. For each incoming job, the load balancer randomly samples $d$ servers independently and dispatches the job to the least loaded server among the $d$ servers. The job is discarded permanently if the buffers of the $d$ servers are all full. In this setting, $Q(t)$ is a continuous-time Markov chain (CTMC) and has a unique stationary distribution for any $\lambda$. Note that the existence of the stationary distribution does not require $\lambda < 1$ since it is a finite-buffer system (i.e. the CTMC has a finite state space). We assume $\lambda = 1 - \gamma N^{-\alpha}$ in this paper so that the fraction of jobs discarded diminishes as $N$ increases.

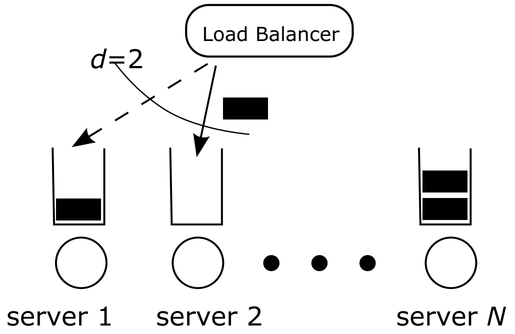


Fig. 2: Power-of-$d$-Choices for Many-Server Systems.

Let $S_k(t)$ denote the fraction of servers with queue size *at least* $k$ at time $t$. $S$ is a $b$-dimensional vector in $\mathcal{S}$, where

$$\mathcal{S} = \left\{ s \in [0,1]^b : s_k = \frac{i_k}{N} \text{ for some integer } 0 \le i_k \le N, \text{ and } s_1 \ge s_2 \ge \cdots \ge s_b \right\}.$$

$S$ is a CTMC. Given $s, s' \in \mathcal{S}$, the transition rate of the CTMC from state $s$ to $s'$ is

$$R_{s,s'} = \tag{3}$$
$$\begin{cases} N(s_k - s_{k+1}), & \text{if } s' = s - \frac{1_k}{N} \\ \lambda N \left(s_{k-1}^d - s_k^d\right), & \text{if } s' = s + \frac{1_k}{N} \\ N \sum_{k=1}^{\infty} -\lambda \left(s_{k-1}^d - s_k^d\right) - (s_k - s_{k+1}), & \text{if } s' = s \\ 0, & \text{otherwise.} \end{cases},$$

where we define $s_0 \equiv 1$ for convenience.

Now consider a sufficiently small time interval $\delta$. According to the transition rates above and a standard argument for CTMC, we have

$$\begin{aligned} & E[S_k(t+\delta) - S_k(t) | S(t) = s] \\ = \; & \lambda N \left(s_{k-1}^d - s_k^d\right) \delta - N(s_k - s_{k+1})\delta + O(\delta^2), \end{aligned} \tag{4}$$

where $\lambda N \left(s_{k-1}^d - s_k^d\right) \delta$ is the probability that during $[t, t+\delta]$, a new job arrives and is routed to a server with $k-1$ jobs already, and $N(s_k - s_{k+1})\delta$ is the probability that during $[t, t+\delta]$, one of the servers with $k$ jobs completes the job in service. Defining

$$\dot{s}_k = \lim_{\delta \to 0} \frac{E[S_k(t+\delta) - S_k(t) | S(t) = s]}{N\delta}$$

leads to the following mean-field model [13], [19]:

$$\begin{aligned} \dot{s}_k &= f_k(s) \\ &= \begin{cases} \lambda(s_{k-1}^d - s_k^d) - (s_k - s_{k+1}), & b-1 \ge k \ge 1, \\ \lambda(s_{b-1}^d - s_b^d) - (s_b - s_{b+1}^*), & k = b. \end{cases} \end{aligned} \tag{5}$$

where $s_{b+1}^* = \lambda^{\frac{d^{b+1}-1}{d-1}}$. The mean-field model is a dynamical system that approximates the original stochastic system by using the expected drift (4) as the system dynamic. We expect the mean-field approximation to be accurate when $N$ is large because each transition leads only a small change $(1/N)$ of the stochastic system. In such a case, the equilibrium point of the mean field model is expected to be "close" to the stationary distribution of the stochastic system.

Note that we added the extra term $s_{b+1}^*$ into the mean-field model above so that the unique equilibrium point of this system has a closed-form:

$$s_k^* = \lambda^{\frac{d^k - 1}{d-1}}. \tag{6}$$

If this extra term is removed in (5), we can only have a recursive expression of the unique equilibrium point [13], which complicates the notation. Nevertheless, in the heavy-traffic regime with $d = \omega\left(\frac{1}{1-\lambda}\right)$, $s_{b+1}^* = O\left((e^{-l(N)})^b\right)$, which is very small. So we chose to add this extra term.

Suppose $d = \frac{l(N)}{1-\lambda}$, then we have $s_1^* = \lambda$ and

$$s_2^* = \lambda^{d+1} = \left(1 - \gamma N^{-\alpha}\right)^{\gamma^{-1} N^{\alpha} l(N) + 1} = O\left(e^{-l(N)}\right).$$

Therefore, *suppose that the mean-field solution is an accurate approximation of the stationary distribution of the system* *under power-of-d-choices,* then the solution suggests that, approximately, $1-\lambda$ fraction of the servers are idle, $\lambda$ fraction of the servers have exactly one job, and the fraction of the servers with more than one job is close to zero. This is similar to JSQ as shown in [5]. We will further show that most jobs are served upon arrival and experience zero delay.

## IV. Main Results

The discussion in the section above serves as a heuristic argument. The following theorem shows that $S(\infty)$ is close to $s^*$ (in the mean-square-sense), from which we will further prove that the asymptotic delay is zero.

**Theorem 1.** *Consider any $0 \leq \alpha < 1/6$ and $d = \frac{l(N)}{1-\lambda}$. The stationary distribution $S(\infty)$ under power-of-d-choices satisfies*

$$E\left[\|S(\infty) - s^*\|^2\right] = O\left(\frac{1}{N}\right). \tag{7}$$

The proof of this theorem is based on Stein's method for mean-field models used in [18], [22], [23]. In particular, it is based on the idea in [23], which views a mean-field model as an approximation of the $N$-server system instead of its limit, which enables us to use $N$-dependent $d$ and $\lambda$ in the mean-field model. The proof of the main theorem is based on a variation of Stein's equation for mean-field models, and application of the lemma for gradient bounds for mean-field models [23] which are both presented below.

**Lemma 1.** *Let $s(t, y)$, a $b$-dimensional vector, denote the solution of the mean-field model (5) with initial condition $y$, and define*

$$g(y) = -\int_0^\infty \|s(t,y) - s^*\|^2 \, dt.$$

*Then we have*

$$\begin{aligned} &E\left[\|S(\infty) - s^*\|^2\right] \\ &= E\left[-\frac{\partial g}{\partial s_b}(S(\infty))s^*_{b+1} - \sum_{y\in\mathcal{S}} R_{S(\infty),y}\Gamma(S(\infty),y)\right], \end{aligned} \tag{8}$$

*where*

$$\Gamma(S(\infty), y) = g(y) - g(S(\infty)) - \nabla g(S(\infty)) \cdot (y - S(\infty)).$$

*Proof.* Please refer to Lemma 4.2 in [23]. □

**Lemma 2.** *Given any $x, y \in \mathcal{S}$ and $R_{x,y} \neq 0$, we have*

$$\begin{aligned} &|g(y) - g(x) - \nabla g(x) \cdot (y - x)| \\ =&O\left(\frac{\log^4(N)}{N^{2-4\alpha-2\epsilon}}\right) 1_{\{x\notin\mathcal{D}^e\}} + O\left(\frac{1}{N^2}\right) 1_{\{x\in\mathcal{D}^e\}}, \end{aligned}$$

*where $\mathcal{D}^e = \{s \mid |s - s^*| \leq \frac{\gamma}{2}N^{-\alpha}\}$ and*

$$\left|\frac{\partial g(s)}{\partial s_b}\right| = O\left(\frac{1}{1-\lambda}\right) = O\left(N^\alpha\right).$$

*Proof.* Please refer to Appendix. □

*Proof of Theorem 1.* According to the two lemmas above, we have

$$\begin{aligned} &E\left[\|S(\infty) - s^*\|^2\right] \\ =&E\left[-\frac{\partial g}{\partial s_b}(S(\infty))s^*_{b+1} - \sum_{y\in\mathcal{S}} R_{S(\infty),y}\Gamma\left(S(\infty),y\right)\right] \\ =&O\left(N^\alpha(1-\gamma N^{-\alpha})^{\frac{d^{b+1}-1}{d-1}}\right) + O\left(\frac{1}{N}\right) + \\ &\frac{4b}{\gamma^2}E\left[\|S(\infty) - s^*\|^2\right] O\left(\frac{\log^4(N)}{N^{1-6\alpha-2\epsilon}}\right). \end{aligned} \tag{9}$$

In the equation above, the first term holds because $\left|\frac{\partial g(s)}{\partial s_b}\right| = O\left(N^\alpha\right)$ and $s^*_{b+1} = (1-\gamma N^{-\alpha})^{\frac{d^{b+1}-1}{d-1}}$, the second term holds because $E\left[1_{\{S(\infty)\in\mathcal{D}^e\}}\right] \leq 1$, and the third term holds because

$$\begin{aligned} E\left[1_{\{S(\infty)\notin\mathcal{D}^e\}}\right] =&E\left[1_{|S(\infty)-s^*|^2 \geq \frac{(1-\lambda)^2}{4}}\right] \\ \leq&E\left[1_{||S(\infty)-s^*||^2 \geq \frac{(1-\lambda)^2}{4b}}\right] \\ \leq&\frac{4b}{(1-\lambda)^2}E\left[||S(\infty) - s^*||^2\right]. \end{aligned}$$

From (9), we have

$$\begin{aligned} &\left(1 - O\left(\frac{\log^4(N)}{N^{1-6\alpha-2\epsilon}}\right)\right) E\left[\|S(\infty) - s^*\|^2\right] \\ =&O\left(N^\alpha(1-\gamma N^{-\alpha})^{\frac{d^{b+1}-1}{d-1}}\right) + O\left(\frac{1}{N}\right). \end{aligned}$$

Note that by choosing $\epsilon = \frac{1-6\alpha}{4}$, we have

$$\lim_{N\to\infty} \frac{\log^4(N)}{N^{1-6\alpha-2\epsilon}} = 0,$$

because $0 \leq \alpha < 1/6$. Further

$$N^\alpha\left(1-\gamma N^{-\alpha}\right)^{\frac{d^{b+1}-1}{d-1}} = o\left(\frac{1}{N}\right)$$

because

$$\begin{aligned} (1-\gamma N^{-\alpha})^{\frac{d^{b+1}-1}{d-1}} \leq&(1-\gamma N^{-\alpha})^{d^b} \\ =&\left((1-\gamma N^{-\alpha})^{\gamma^{-1}N^\alpha l(N)}\right)^b \\ =&O\left((e^{-l(N)})^b\right), \end{aligned}$$

which concludes the proof. □

Now recall that $\mathcal{W}_N$ denotes the event that an incoming job is routed to a busy server, and $p_{\mathcal{W}_N}$ denotes the probability of this event at steady-state. We have the following corollary based on Theorem 1.

**Corollary 1.** *Consider any $0 \leq \alpha < 1/6$ and $d = \omega\left(\frac{1}{1-\lambda}\right)$. Under power-of-d-choices, we have*

$$\lim_{N\to\infty} p_{\mathcal{W}_N} = 0. \tag{10}$$

*In other words, almost all jobs are served immediately upon arrival at steady-state and experience zero waiting time.*

*Proof.* We first assume that $d = \frac{l(N)}{1-\lambda}$. Note that an incoming job is routed to an idle server if one of the $d$ probed servers is an idle server. From Theorem 1, the probability that the system has at most $\frac{\gamma}{2}N^{-\alpha}$ idle servers is

$$\begin{aligned} & \Pr\left(S_1(\infty) \geq 1 - \frac{\gamma}{2}N^{-\alpha}\right) \\ \leq\ & \Pr\left((S_1(\infty) - s_1^*)^2 \geq \frac{\gamma^2}{4}N^{-2\alpha}\right) \\ \leq\ & \frac{E\left[\|S(\infty) - s^*\|^2\right]}{\frac{\gamma^2}{4}N^{-2\alpha}} \\ =\ & O\left(\frac{1}{N^{1-2\alpha}}\right). \end{aligned}$$

Now conditioned on that the system has at least $\frac{\gamma}{2}N^{-\alpha}$ idle servers, the probability that one of $d$ servers sampled is idle is at least

$$\begin{aligned} 1 - \left(1 - \frac{\gamma}{2}N^{-\alpha}\right)^d &= 1 - \left(1 - \frac{\gamma}{2}N^{-\alpha}\right)^{l(N)N^{\alpha}} \\ &= 1 - O\left(e^{-\frac{\gamma}{2}l(N)}\right). \end{aligned}$$

Therefore, we have

$$\begin{aligned} & p_{\mathcal{W}_N} \\ & = \Pr(\mathcal{W}_N | S_1(\infty) \geq 1 - \frac{\gamma}{2}N^{-\alpha}) \Pr\left(S_1(\infty) \geq 1 - \frac{\gamma}{2}N^{-\alpha}\right) \\ & + \Pr(\mathcal{W}_N | S_1(\infty) < 1 - \frac{\gamma}{2}N^{-\alpha}) \Pr\left(S_1(\infty) < 1 - \frac{\gamma}{2}N^{-\alpha}\right) \\ & \leq \Pr(S_1(\infty) \geq 1 - \frac{\gamma}{2}N^{-\alpha}) + \Pr\left(\mathcal{W}_N | S_1(\infty) < 1 - \frac{\gamma}{2}N^{-\alpha}\right) \\ & = O\left(\frac{1}{N^{1-2\alpha}}\right) + O\left(e^{-\frac{\gamma}{2}l(N)}\right), \end{aligned}$$

which proves the corollary for $d = \frac{l(N)}{1-\lambda}$. For fixed $N$ and $\lambda$, using stochastic coupling, it can be shown that $p_{\mathcal{W}_N}$ decreases as $d$ increases, so the result holds for $d = \omega\left(\frac{1}{1-\lambda}\right)$. □

The following theorem shows that $d = \omega\left(\frac{1}{1-\lambda}\right)$ is almost necessary.

**Theorem 2.** *Consider any $0 \leq \alpha < 1$ and $d = O\left(\frac{1}{1-\lambda}\right)$. Under power-of-d-choices with $b = \infty$, we have*

$$\lim_{N\to\infty} p_{\mathcal{W}_N} > \epsilon. \tag{11}$$

*for some $\epsilon > 0$.*

*Proof.* According to work conservation law, we have

$$E[S_1(\infty)] = \lambda,$$

because the incoming workload is $\lambda$. We also know that $0 \leq S_1(\infty) \leq 1$. Now define $\beta = 1 - \lambda = \gamma N^{-\alpha}$. It is easy to show that

$$\Pr\left(S_1(\infty) \geq \lambda - \beta\right) \geq \frac{1}{2}$$

because otherwise,

$$E[S_1(\infty)] < \frac{1}{2}(\lambda - \beta) + \frac{1}{2} = \lambda,$$

which contradicts $E[S_1(\infty)] = \lambda$. Now condition on the system has at most

$$N(1 - \lambda + \beta) = 2\gamma N^{1-\alpha}$$

idle servers, the probability of having an idle server in $d$ probes is at most

$$1 - \left(1 - 2\gamma N^{-\alpha}\right)^d,$$

which converges to a positive constant as $N \to \infty$. □

**Theorem 3.** *Consider $b = 1$ and $d = \frac{l(N)}{1-\lambda}$. The stationary distribution $S(\infty)$ under power-of-d-choices satisfies*

$$E\left[\|S(\infty) - s^*\|^2\right] = O\left(\frac{1}{N}\right).$$

*Proof.* Please refer to our technical report [11]. □

Based on Theorem 3, we have the following corollary.

**Corollary 2.** *For loss system with $b = 1$. Consider any $0 \leq \alpha < 0.5$ and $d = \omega\left(\frac{1}{1-\lambda}\right)$. Under power-of-d-choices, we have*

$$\lim_{N\to\infty} p_{\mathcal{W}_N} = 0. \tag{12}$$

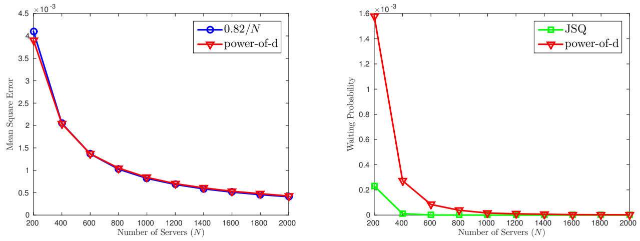
*Proof.* Following the steps in the proof of Corollary 1. □

## V. Simulations

In this section, we use simulations to evaluate the performance of power-of-$d$-choices with $d = 2\log(N)/(1-\lambda)$. In particular, we compared its waiting probability $p_{\mathcal{W}_N}$ with that under the JSQ policy because under our modeling assumptions (Poisson arrivals, exponential service times, homogeneous servers and per-server queues), JSQ minimizes $p_{\mathcal{W}_N}$ among all load balancing policies.

### *A. Performance under different system sizes (i.e. different $N$)*

In the first set of simulations, we fixed $\alpha = 1/6$ and $\gamma = 0.5$ (so $\lambda = 1 - 0.5N^{-\frac{1}{6}}$) and varied the number of servers $N$ from 200 to $2,000$ with a step size of 200. Figure 3a shows the mean square error $E\left[\|S(\infty) - s^*\|^2\right]$ versus $N$. From the figure, we can see that the mean-square error decreases as $N$ increases. and the curve showed that $0.82/N$ best fits the mean-square error. Figure 3b shows the waiting probability $p_{\mathcal{W}_N}$ versus the number of servers $N$ for both power-of-$d$-choices and JSQ. As we showed in Corollary 1, $p_{\mathcal{W}_N}$ is close to zero for large $N$. Figure 3b confirmed this result. We can see that the waiting probability quickly converges to zero and the wait probability of JSQ is smaller than that of power-of-$d$-choices for all $N$ since JSQ samples all the queues for each incoming job.

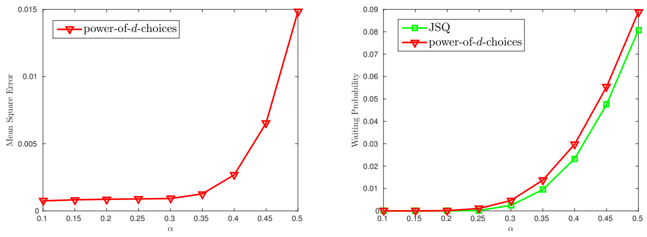


(a) Mean Square Error (b) Waiting probability

Fig. 3: Performance of power-of-$d$-choices and JSQ policies with $\alpha = 1/6$ and $\gamma = 0.5$ under different $N$.

### B. Performance in different heavy-traffic regimes (i.e. different $\alpha$)

In this set of simulations, we fixed $N = 1,000$ and $\gamma = 0.5$, and varied $\alpha$ from 0.1 to 0.5 with step size 0.05. Figure 4a shows the mean-square error versus $\alpha$. As we can expect, the mean-square error increases as $\alpha$ increases. One explanation is that larger $\alpha$ means heavier the traffic load, which leads to higher variance at the steady state (note that the mean-square error is closely related to the variance of $S(\infty)$). Figure 4b illustrates the waiting probability $p_{\mathcal{W}_N}$ versus the load $\alpha$. Interestingly, we can see that when $\alpha \leq 0.25$, the waiting probability of power-of-$d$-choices and JSQ is close to zero. When $\alpha > 0.25$, the waiting probability of both policies increases significantly. For $\alpha = 0.5$, the waiting probability is close to 8% under both policies.



(a) Mean Square Error (b) Waiting probability

Fig. 4: Performance of power-of-$d$-choices and JSQ policies with $N = 1,000$ and $\gamma = 0.5$ under different $\alpha$.

Figure 5 shows the mean square error versus $N$ with $\alpha = \{0.3, 0.4, 0.5\}$, which complements Figure 3a with $\alpha = 1/6$. For $\alpha = \{0.3, 0.4, 0.5\}$, the curve $c/N$ fits the mean-square error well with $c = \{1.1, 2.8, 14.2\}$, respectively. These results strongly suggest that the actual mean-square error in fact decreases as $O(1/N)$ for $0 \leq \alpha \leq 0.5$.

## VI. Conclusions

In this paper, we studied power-of-$d$-choices, a popular load balancing algorithm for many-server system. Motivated by [14], we considered the fundamental requirement on $d$ to achieve asymptotic zero delay in the large-system limit and in the heavy traffic regime. Assuming Poisson arrivals, exponential service times and homogeneous servers, and assuming the load of the system is $\lambda = 1 - \gamma N^{-\alpha}$, we proved that $d = \omega\left(\frac{1}{1-\lambda}\right)$ is sufficiently and almost necessary to achieve zero queueing delay for $0 \leq \alpha < 1/6$. Our simulation results confirmed this result. It is known that when $\alpha \geq 0.5$, no load balancing algorithm can achieve asymptotic zero delay. An open question remains to be addressed is whether $d = \omega\left(\frac{1}{1-\lambda}\right)$ is sufficient for power-of-$d$-choices to achieve asymptotic zero delay when $1/6 \leq \alpha < 0.5$.

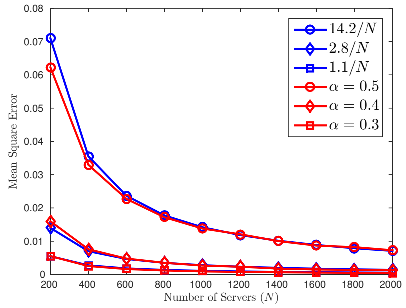


Fig. 5: Mean square error of power-of-$d$-choices policy with $\alpha = \{0.3, 0.4, 0.5\}$.

## Acknowledgment

This work was supported in part by NSF ECCS-1547294, ECCS-1609202, ECCS-1739344 and the U.S. Office of Naval Research (ONR Grant No. N00014-15-1-2169).

## References

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

[2] A. Braverman and J. G. Dai. Steins method for steady-state diffusion approximations of $m/Ph/n+m$ systems. *Ann. Appl. Probab.*, 27(1):550–581, 02 2017.

[3] A. Braverman, J. G. Dai, and J. Feng. Steins method for steady-state diffusion approximations: An introduction through the erlang-a and erlang-c models. *Stoch. Syst.*, 6(2):301–366, 2016.

[4] G. Brightwell and M. Luczak. The supermarket model with arrival rate tending to one. *arXiv preprint arXiv:1201.5523*, 2012.

[5] P. Eschenfeldt and D. Gamarnik. Join the shortest queue with many servers. the heavy traffic asymptotics. *arXiv preprint arXiv:1502.00999*, 2015.

[6] P. Eschenfeldt and D. Gamarnik. Supermarket queueing system in the heavy traffic regime. Short queue dynamics. *arXiv preprint arXiv:1610.03522*, 2016.

[7] D. Gamarnik, J. N. Tsitsiklis, and M. Zubeldia. Delay, memory, and messaging tradeoffs in distributed service systems. In *Proc. Ann. ACM SIGMETRICS Conf.*, pages 1–12, 2016.

[8] N. Gast. Expected values estimated via mean-field approximation are 1/n-accurate. *Proc. ACM Meas. Anal. Comput. Syst.*, 1(1):17:1–17:26, June 2017.

[9] N. Gast and B. Van Houdt. A refined mean field approximation. In *Proc. Ann. ACM SIGMETRICS Conf.*, Irvien, CA, 2018.

[10] S. Halfin and W. Whitt. Heavy-traffic limits for queues with many exponential servers. *Operations Research*, 29(3):567–588, 1981.

[11] X. Liu and L. Ying. On achieving zero delay with power-of-$d$-choices load balancing. 2017. Arizona State University Technical Report.

[12] 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.

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

[14] D. Mukherjee, S. C. Borst, J. S. van Leeuwaarden, and P. A. Whiting. Universality of power-of-$d$ load balancing in many-server systems. *arXiv preprint arXiv:1612.00723*, 2016.

[15] A. Mukhopadhyay and R. R. Mazumdar. Analysis of randomized join-the-shortest-queue (JSQ) schemes in large heterogeneous processor-sharing systems. *IEEE Trans. Cont. Netw. Sys.*, pages 116–126, June 2016.

[16] E. Schurman and J. Brutlag. The user and business impact of server delays, additional bytes, and http chunking in web search. In *O'Reilly Velocity Web Performance and Operations Conf.*, June 2009.

[17] A. Stolyar. Pull-based load distribution in large-scale heterogeneous service systems. *Queueing Syst.*, 80(4):341–361, 2015.

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

[19] 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.

[20] R. R. Weber. On the optimal assignment of customers to parallel servers. *J. Appl. Probab.*, 15(2):406–413, 1978.

[21] W. Winston. Optimality of the shortest line discipline. *J. Appl. Probab.*, 14(1):181–189, 1977.

[22] L. Ying. On the approximation error of mean-field models. In *Proc. Ann. ACM SIGMETRICS Conf.*, Antibes Juan-les-Pins, France, 2016.

[23] L. Ying. Stein's method for mean field approximations in light and heavy traffic regimes. *Proc. ACM Meas. Anal. Comput. Syst.*, 1(1):12:1–12:27, June 2017.

## APPENDIX

## PROOF OF LEMMA 2

The proof of this lemma is based on the gradient bound for mean-field models (Lemma 2.2 in [23]). Recall the mean-field model

$$\dot{s}_k = f_k(s) \tag{13}$$

$$= \begin{cases} \lambda(s_{k-1}^d - s_k^d) - (s_k - s_{k+1}), & 1 \le k \le b-1, \\ \lambda(s_{b-1}^d - s_b^d) - (s_b - s_{b+1}^*), & k = b. \end{cases} \tag{14}$$

Define $x_k = s_k - s_k^*$ and

$$A_k = \frac{s_k^d - (s_k^*)^d}{s_k - s_k^*} = \sum_{j=0}^{d-1} (s_k)^{d-1-j}(s_k^*)^j.$$

Then we obtain the dynamics for $x(t)$ as follows

$$\dot{x}_k = \begin{cases} -\lambda A_1 x_1 - (x_1 - x_2), & k = 1 \\ \lambda A_{k-1} x_{k-1} - \lambda A_k x_k - (x_k - x_{k+1}), & 2 \le k \le b-1 \\ \lambda A_{b-1} x_{b-1} - \lambda A_b x_b - x_b, & k = b \end{cases},$$

where the second equality holds because $s^*$ is the equilibrium point of (13).

The first-order system for this mean-field model is

$$\dot{x}_k^{(1)} = g_k(x^{(1)}) = \frac{\partial f_k}{\partial s}(s) \cdot x^{(1)}$$
$$= \lambda d \left( s_{k-1}^{d-1} x_{k-1}^{(1)} - s_k^{d-1} x_k^{(1)} \right) - (x_k^{(1)} - x_{k+1}^{(1)}), \tag{15}$$

where we define $x_0^{(1)} \equiv 0$ and $x_{b+1}^{(1)} \equiv 0$ for convenience.

**Lemma 3.** *Consider the dynamic system defined in* (13)*. Given that the initial condition satisfies* $1 = s_0(0) \ge s_1(0) \ge s_2(0) \ge \cdots \ge s_n(0) \ge 0$, *we have* $0 \le s_k(t) \le 1$ *for any* $t \ge 0$ *and* $0 \le k \le b$.

*Proof.* Please refer to Lemma C.1 in [23]. □

**Lemma 4.** *Consider Lyapunov function*

$$V(x(t)) = \sum_{i=k}^{b} w_k \left| x_k(t) \right|,$$

*where* $w_k$ *are defined to be*

$$w_0 = 0, w_1 = 1,$$
$$w_2 = 1 + (1-\delta)(1-\lambda),$$
$$w_{k+1} = 2w_k - w_{k-1} - \delta w_k, \ 2 \le k \le b.$$

*For* $\delta = N^{-(\alpha+\varepsilon)}$ *such that* $\varepsilon > 0$, *and sufficiently large* $N$, *we have*

$$\dot{V}(x(t)) \le -\delta V(x(t)).$$

*Proof.* We first have

$$\dot{V}(x(t)) \le \sum_{k=1}^{b} -(\lambda A_k w_k + w_k - \lambda A_k w_{k+1} - w_{k-1}) \left| x_k(t) \right|.$$

So the lemma holds by proving

$$\lambda A_k w_k + w_k - \lambda A_k w_{k+1} - w_{k-1} \ge \delta w_k,$$

i.e. by proving

$$w_{k+1} - w_k \le \frac{w_k - w_{k-1} - \delta w_k}{\lambda A_k}. \tag{16}$$

Note $w_k$ is an increasing sequence when $N$ is sufficiently large because

$$w_k - w_{k-1} = \Theta\left(N^{-\alpha}\right) \text{ and } \delta w_k = O\left(N^{-(\alpha+\varepsilon)}\right).$$

We now prove (16) by considering the following cases.

For $k = 1$, we have $w_{k-1} = 0$. Therefore, we need to prove

$$w_2 \le w_1 + \frac{w_1 - \delta w_1}{\lambda A_1} = 1 + \frac{1-\delta}{\lambda A_1}.$$

This holds according to the definition of $w_2$ and

$$A_1 = \sum_{j=0}^{d-1} (s_1)^{d-1-j}(s_1^*)^j \le \sum_{j=0}^{d-1} \lambda^j \le \frac{1}{1-\lambda},$$

where we use the facts $0 \le s_1 \le 1$ and $s_1^* = \lambda < 1$.

For $k = 2$, we need to prove

$$w_3 - w_2 \le \frac{w_2 - w_1 - \delta w_2}{\lambda A_2}$$

The inequality holds because

$$\begin{aligned} w_2 - w_1 - \delta w_2 &= (1-\delta)(1-\lambda) - \delta\left(1 + (1-\delta)(1-\lambda)\right) \\ &= (1-\delta)^2(1-\lambda) - \delta \\ &= (1-\delta)^2 \frac{\gamma}{N^\alpha} - \frac{1}{N^{\alpha+\varepsilon}} > 0, \end{aligned}$$

and

$$\lambda A_2 = \lambda \sum_{j=0}^{d-1} (s_2)^{d-1-j}(s_2^*)^j \le \lambda \sum_{j=0}^{d-1} (s_2^*)^j \le \frac{\lambda}{1-s_2^*} < 1,$$

when $N$ is sufficiently large.

For $k \geq 3$, (16) holds by the definition of $w_{k+1} - w_k$ and the fact $\lambda A_k < 1$ for sufficiently large $N$. □

**Lemma 5.** *Define $\tilde{t} = (\alpha + \varepsilon) N^{\alpha+\varepsilon} \log N$. Starting from any initial condition, $s(t) \in \mathcal{D}^e$ for $t \geq \tilde{t}$, which implies that $s_k(t) \leq 1 - \frac{\gamma}{2} N^{-\alpha}$ for $k \geq 1$.*

The proof of this lemma is based on Lemma 4. The details can be found in [11].

**Lemma 6.** *(***Proof of C2 of Lemma 2.2 in [23]***) Under the dynamical system defined by (15),*

$$\left| x^{(1)}(t) \right| \leq \left| x^{(1)}(0) \right|.$$

*Proof.* Define $V(t) = \left| x^{(1)}(t) \right|$, and we have

$$\frac{d|x_k^{(1)}(t)|}{dt} \leq \lambda d s_{k-1}^{d-1} \left| x_{k-1}^{(1)} \right| - \lambda d s_k^{d-1} \left| x_k^{(1)} \right| - \left| x_k^{(1)} \right| + \left| x_{k+1}^{(1)} \right|,$$

where $x_{b+1}^{(1)}(t) = 0$ for all $t$. So

$$\dot{V}(t) \leq -\lambda d x_b^{d-1} \left| x_b^{(1)} \right| - \left| x_1^{(1)} \right| \leq 0,$$

and the lemma holds. □

**Lemma 7 (Proof of C3 of Lemma 2.2 in [23]).** *Consider Lyapunov function*

$$V(x^{(1)}) = \sum_{k=1}^{b} w_k \left| x_k^{(1)} \right|,$$

*where $w_k$ are defined to be*

$$w_0 = 0, w_1 = 1,$$
$$w_k = w_{k-1} + \tfrac{1}{b}, 2 \leq k \leq b.$$

*For sufficiently large $N$ and $\delta_1 = \frac{1}{4b}$, we have*

$$|x^{(1)}| \leq V(x^{(1)}) \leq 2|x^{(1)}|,$$
$$\dot{V}(x^{(1)}) \leq -\delta_1 V(x^{(1)}), \quad \textit{given } s \in \mathcal{D}^e.$$

*Proof.* The first inequality holds due to the definition of $w_k$. Similar with the proof in Lemma 4, we have

$$\dot{V}(x^{(1)}) \leq \sum_{k=1}^{b} - \left( \lambda d w_k s_k^{d-1} + w_k - \lambda d w_{k+1} s_k^{d-1} - w_{k-1} \right) \left| x_k^{(1)} \right|.$$

So the lemma holds by proving

$$\lambda d w_k s_k^{d-1} + w_k - \lambda d w_{k+1} s_k^{d-1} - w_{k-1} \geq \delta_1 w_k,$$

i.e. by proving

$$w_{k+1} - w_k \leq \frac{w_k - w_{k-1} - \delta_1 w_k}{\lambda d s_k^{d-1}}. \tag{17}$$

We prove (17) by considering the following cases.

For $k = 1$, we have $w_{k-1} = 0$. Therefore, we need to prove

$$w_2 \leq w_1 + \frac{w_1 - \delta_1 w_1}{\lambda d s_1^{d-1}} = 1 + \frac{1 - \frac{1}{4b}}{\lambda d s_1^{d-1}}. \tag{18}$$

This holds according to the definition of $w_2$ and $0 \leq s_1 \leq 1 - \frac{\gamma}{2} N^{-\alpha}$ from Lemma 5.

For $k \geq 2$, we need to prove

$$\frac{1}{b} \leq \frac{\frac{1}{b} - \frac{1}{4b} w_k}{\lambda d s_k^{d-1}}. \tag{19}$$

Again this holds according to the fact $0 \leq s_k \leq 1 - \frac{\gamma}{2} N^{-\alpha}$. □

The error system for this mean-field model is

$$\dot{e}_k = f_k \left( s + \frac{1}{N} x^{(1)} + e \right) - f_k(s) - \frac{1}{N} \frac{\partial f_k}{\partial s}(s) \cdot x^{(1)},$$

with initial condition $e(0) = 0$.

**Lemma 8 (Proof of C4 of Lemma 2.2 in [23]).** *Given $|e(t)| \leq \frac{1}{N}$, we have*

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

*Further, for any $t \leq \tilde{t}$, and $\alpha < 1/3$,*

$$|e(t)| \leq \frac{1}{N}.$$

*Proof.* We first have

$$\begin{aligned} \dot{e}_k \overset{(a)}{=} & \; g_k(e) + f_k \left( s + \frac{1}{N} x^{(1)} + e \right) - f_k(s) - \\ & \frac{\partial f_k}{\partial s}(s) \cdot \left( \frac{1}{N} x^{(1)} + e \right) \\ \overset{(b)}{=} & \; g_k(e) + \frac{\lambda d(d-1)}{2} \tilde{s}_{k-1}^{d-2} \left( \frac{1}{N} x_{k-1}^{(1)} + e_{k-1} \right)^2 \\ & - \frac{\lambda d(d-1)}{2} \tilde{s}_k^{d-2} \left( \frac{1}{N} x_k^{(1)} + e_k \right)^2, \end{aligned}$$

where (a) holds due to the definition $g_k(e) = \frac{\partial f_k}{\partial s}(s) \cdot e$, (b) holds due to the Taylor expansion of $f_k(s + \frac{1}{N} x^{(1)} + e)$ at $s$, and $\tilde{s}$ is between $s + \frac{1}{N} x^{(1)} + e$ and $s$. Define Lyapunov function

$$V(t) = |e(t)|,$$

and $\hat{g}_k(e) = \lambda d s_{k-1}^{d-1} |e_{k-1}| - \lambda d s_k^{d-1} |e_k| - |e_k| + |e_{k+1}|$.

Note that

$$\begin{aligned} \frac{d|e_k(t)|}{dt} \leq & \hat{g}_k(e) + \frac{\lambda d(d-1)}{2} |\tilde{s}_{k-1}^{d-2}| \left( \frac{1}{N} x_{k-1}^{(1)} + e_{k-1} \right)^2 \\ & + \frac{\lambda d(d-1)}{2} |\tilde{s}_k^{d-2}| \left( \frac{1}{N} x_k^{(1)} + e_k \right)^2. \end{aligned}$$

By following the proof of Lemma 6, we know $\sum_{k=1}^{b} \hat{g}_k(e) \leq 0$. Then we immediately obtain

$$\frac{d|e(t)|}{dt} \leq \frac{\lambda d(d-1)}{2} \sum_{k=1}^{b} |\tilde{s}_{k-1}^{d-2}| \left( \frac{1}{N} x_{k-1}^{(1)} + e_{k-1} \right)^2$$

$$+ \frac{\lambda d(d-1)}{2} \sum_{k=1}^{b} |\tilde{s}_k^{d-2}| \left(\frac{1}{N} x_k^{(1)} + e_k\right)^2,$$

which implies

$$\frac{d|e(t)|}{dt} \le \lambda d(d-1) \left\| \frac{1}{N} x^{(1)} + e \right\|^2.$$

Given Lemma 6 and $|e(t)| \le \frac{1}{N}$, we conclude

$$\frac{d|e(t)|}{dt} \le \frac{d^2}{N^2}.$$

Further, for any $t \le \tilde{t}$ and $\alpha < 1/3$,

$$|e(t)| \le \tilde{t} \frac{d^2}{N^2} \le \frac{1}{N}.$$

□

**Lemma 9 (Proof of C5 of Lemma 2.2 in [23]).** *Consider Lyapunov function*

$$V(e(t)) = \sum_{k=1}^{b} w_k \left| e_k(t) \right|,$$

*where $w_k$ are defined to be*

$$\begin{gathered} w_0 = 0, w_1 = 1, \\ w_k = w_{k-1} + \tfrac{1}{b}, 2 \le k \le b. \end{gathered}$$

*For sufficiently large $N$, we have*

$$\dot{V}(e) \le -\delta_1 V(e) + \frac{|x^{(1)}|^2}{N^2}, \qquad givens(t) \in \mathcal{D}^e.$$

*Proof.* Similar to the proof of the previous lemma, we have

$$\begin{aligned} \frac{d|e_k(t)|}{dt} \le& \hat{g}_k(e) + \frac{\lambda d(d-1)}{2} |\tilde{s}_{k-1}^{d-2}| \left(\frac{1}{N} x_{k-1}^{(1)} + e_{k-1}\right)^2 \\ &+ \frac{\lambda d(d-1)}{2} |\tilde{s}_k^{d-2}| \left(\frac{1}{N} x_k^{(1)} + e_k\right)^2. \end{aligned}$$

Assume $|e(t)| \le \frac{2}{N}$ and from Lemma 8, we have

$$|\tilde{s}_k| \le 1 - \frac{\gamma}{2} N^{-\alpha} + \frac{1}{N} |x_k^{(1)}| + |e_k| \le 1 - \frac{\gamma}{3} N^{-\alpha}.$$

which implies $|\tilde{s}_k^{d-2}| = O(e^{-l(N)})$. So we have

$$\frac{d|e_k(t)|}{dt} \le \hat{g}_k(e) + \frac{|x_k^{(1)}|^2}{bN^2},$$

for sufficient large $N$. By following the proof in Lemma 7, we have

$$\dot{V}(e(t)) \le -\delta_1 V(e(t)) + \frac{|x^{(1)}(t)|^2}{N^2}.$$

Note the assumption $|e(t)| \le \frac{2}{N}$ holds because

$$|e(t)| \le V(e(t)) \le \frac{2}{N}.$$

□

Now we are ready to prove Lemma 2. First we compute

$$\int_0^\infty |x^{(1)}(t)|^2 \, dt.$$

According to Lemma 6, we have for $t \le \tilde{t}$

$$\int_0^{\tilde{t}} |x^{(1)}(t)|^2 \, dt \le |x^{(1)}(0)|^2 \tilde{t} \le \tilde{t}.$$

According to Lemma 5, we have for $t \ge \tilde{t}$

$$\begin{aligned} \int_{\tilde{t}}^\infty |x^{(1)}(t)|^2 \, dt \le& \int_{\tilde{t}}^\infty \left(V(x^{(1)}(t))\right)^2 \, dt \\ \le& \int_{\tilde{t}}^\infty V(x^{(1)}(\tilde{t}))^2 e^{-2\delta_1(t-\tilde{t})} \, dt \\ =& \frac{V(x^{(1)}(\tilde{t}))^2}{2\delta_1} \le \frac{4|x^{(1)}(t)|}{2\delta_1} \le \frac{2}{\delta_1}. \end{aligned}$$

Therefore, we have

$$\int_0^\infty |x^{(1)}(t)|^2 \, dt \le \tilde{t} + \frac{2}{\delta_1}.$$

Then we compute

$$\int_0^\infty |e(t)| \, dt.$$

According to Lemma 8, we have for $t \le \tilde{t}$

$$\int_0^{\tilde{t}} |e(t)| \, dt \le \int_0^{\tilde{t}} \frac{td^2}{N^2} \, dt \le \frac{\tilde{t}^2 d^2}{2N^2}.$$

According to Lemma 9, we have for $t \ge \tilde{t}$

$$\begin{aligned} \dot{V}(e(t)) \le& -\delta_1 V(e(t)) + \frac{1}{N^2} \left| x^{(1)} \right|^2 \\ \le& -\delta_1 V(e(t)) + \frac{1}{N^2} e^{-2\delta_1(t-\tilde{t})}. \end{aligned}$$

Based on comparison principle, we obtain for $t \ge \tilde{t}$

$$V(e(t)) \le V(e(\tilde{t})) e^{-\delta_1(t-\tilde{t})} + \frac{1}{\delta_1 N^2} \left( e^{-\delta_1(t-\tilde{t})} - e^{-2\delta_1(t-\tilde{t})} \right),$$

which implies that

$$\begin{aligned} &\int_{\tilde{t}}^\infty |e(t)| \, dt \le \int_{\tilde{t}}^\infty V(e(t)) \, dt \\ &\le \frac{V(e(\tilde{t}))}{\delta_1} + \frac{1}{2\delta_1^2 N^2} \le \frac{2\tilde{t} d^2}{\delta_1 N^2} + \frac{1}{2\delta_1^2 N^2}. \end{aligned}$$

Finally we have

$$\begin{aligned} \frac{1}{N^2} \int_0^\infty |x^{(1)}(t)|^2 \, dt =& O\left(\frac{\tilde{t}}{N^2}\right) 1_{\{s \notin \mathcal{D}^e\}} + O\left(\frac{1}{N^2}\right) 1_{\{s \in \mathcal{D}^e\}} \\ \int_0^\infty |e(t)| \, dt =& O\left(\frac{\tilde{t}^2 d^2}{N^2}\right) 1_{\{s \notin \mathcal{D}^e\}} + O\left(\frac{1}{N^2}\right) 1_{\{s \in \mathcal{D}^e\}}. \end{aligned}$$

which implies from Lemma 2.2 in [23]

$$\begin{aligned} &|g(y) - g(x) - \nabla g(x) \cdot (y - x)| \\ =& O\left(\frac{\log^4 N}{N^{2-4\alpha-2\epsilon}}\right) 1_{\{s \notin \mathcal{D}^e\}} + O\left(\frac{1}{N^2}\right) 1_{\{s \in \mathcal{D}^e\}}. \end{aligned}$$