A Note on Load Balancing in Many-Server Heavy-Traffic Regime

Xingyu Zhou
Department of ECE
The Ohio State University
zhou.2055@osu.edu

Ness Shroff
Department of ECE and CSE
The Ohio State University
shroff.11@osu.edu

Abstract

In this note, we apply Stein's method to analyze the performance of general load balancing schemes in the many-server heavy-traffic regime. In particular, consider a load balancing system of N servers and the distance of arrival rate to the capacity region is given by $N^{1-\alpha}$ with $\alpha > 1$. We are interested in the performance as N goes to infinity under a large class of policies. We establish different asymptotics under different scalings and conditions. Specifically, (i) if the second moments of total arrival and total service processes approach to constants σ_Σ^2 and ν_Σ^2 as $N \to \infty$, then for any $\alpha > 3$, the distribution of the sum queue length scaled by $N^{1-\alpha}$ converges to an exponential random variable with rate $\frac{\sigma_\Sigma^2+\nu_\Sigma^2}{2}$. (2) If the second moments linearly increase with N with coefficients σ_a^2 and ν_s^2, then for any $\alpha > 2$, the distribution of the sum queue length scaled by $N^{-\alpha}$ converges to an exponential random variable with rate $\frac{\sigma_a^2+\nu_s^2}{2}$. (3) If the second moments quadratically increase with N with coefficients $\tilde{\sigma}_a^2$ and $\tilde{\nu}_s^2$, then for any $\alpha > 1$, the distribution of the sum queue length scaled by $N^{-\alpha-1}$ converges to an exponential random variable with rate $\frac{\tilde{\sigma}_a^2+\tilde{\nu}_s^2}{2}$. All the results are simple applications of our previously developed framework of Stein's method for heavy-traffic analysis in [9].

1 Introduction

Load balancing has attracted increasing attention recently due to its application in cloud computing and data centers. In this note, we consider a system consisting of one load balancer and N servers each with an infinite buffer queue. The arrival is immediately dispatched to one of the servers based on a certain load balancing policy. In particular, we consider a set of systems where the distance of arrival rate to the capacity is given by $N^{1-\alpha}$ with $\alpha > 1$ and let N go to infinity, which is often called the many-server heavy-traffic regime.

Many previous works have investigated the system performance under different values of α. For example, if $\alpha = \frac{1}{2}$, i.e., Halfin-Whitt regime, Join-Shortest-Queue (JSQ) has been extensively studied [1, 2, 3]. In the Sub-Halfin-Whitt regime where $\alpha \in (0, \frac{1}{2})$, several load balancing policies are investigated [8]. Recently, the authors also extend the analysis to the case when $\alpha \in (\frac{1}{2}, 1)$ [7]. In [5], load balancing policies in Nondegenerate Slowdown regime (NDS) (i.e., $\alpha = 1$) are studied. More recently, [6] studied JSQ in the regime when $\alpha > 2$ and they show that the total queue length scaled by $N^{-\alpha}$ converges to an exponential random variable via transform method and Stein's method.

In this paper, instead of only focusing on JSQ policy under one particular scaling situation as in [6], we investigate a large class of load balancing policies and establish their asymptotic performance for different values of α under different scalings. This is possible because we adopt the framework of Stein's method for heavy-traffic analysis developed in our early work [9] for general load balancing and scheduling problems. In details, we have made the following key contributions.

First, we present the asymptotic performance for a large class of load balancing schemes. For any policy in this class, we show that the asymptotic performance depends on the scaling properties of the second moments of total arrival and service processes, i.e., $\sigma_\Sigma^{(N)}$ and $\nu_\Sigma^{(N)}$. In particular, if both of

them converge to constants σ_Σ and μ_Σ respectively as $N \to \infty$, then for any $\alpha > 3$, the distribution of the sum queue length scaled by $N^{1-\alpha}$ converges to an exponential random variable with rate $\frac{\sigma_\Sigma^2+\nu_\Sigma^2}{2}$. If $\sigma_\Sigma^{(N)} = N\sigma_s^2$ and $\nu_\Sigma^{(N)} = N\nu_s^2$, then for any $\alpha > 2$, the distribution of the sum queue length scaled by $N^{-\alpha}$ converges to an exponential random variable with rate $\frac{\sigma_a^2+\nu_s^2}{2}$ (which recovers the special JSQ case as in [6]). If $\sigma_\Sigma^{(N)} = N^2\tilde{\sigma}_s^2$ and $\nu_\Sigma^{(N)} = N^2\tilde{\nu}_s^2$, then for any $\alpha > 1$, the distribution of the sum queue length scaled by $N^{-\alpha-1}$ converges to an exponential random variable with rate $\frac{\tilde{\sigma}_a^2+\tilde{\nu}_s^2}{2}$. It is worth noting that this class not only includes policies that achieve a single-dimensional state-space collapse (e.g., JSQ, Power-of-d, p-JSQ as in [11], and many others in [12]), but also includes all the policies under which the state-space collapse region is multi-dimensional as long as it can be covered by a cone. On one hand, this directly indicates that a single-dimensional state-space collapse is not necessary for the asymptotic performance as in [6]. On the other hand, it also allows us to explore the trade-off between flexibility and performance.

Second, although Stein's method serves as the key idea behind both [6] and our work, the execution in our work is totally different from [6]. In particular, our analysis is purely based on the general framework of Stein's method developed in our early work [9]. This framework of Stein's method for heavy-traffic analysis can be used to analyze single-server system, general load balancing problems and scheduling problems. The result in this paper is just another application of our early framework with a very simple proof. By using this framework, we are not only able to establish asymptotic performance for a large class of policies, but also obtain different asymptotics under different scalings. The simplicity and broader applicability of our framework comes from the fact that it inherits the same intuitions and mathematical bounds as in the drift-based method. As a result, we can directly plug in previously well-known bounds established by drift-method into this framework, and hence easily establish new asymptotic performance beyond first moment result (e.g., convergence in distribution) without analyzing each policy by going through all the details repeatedly. For interesting readers, please refer to [9] for more details.

2 System model and preliminaries

We consider a single-hop queueing system in the discrete time, i.e., a time-slotted system. There are N separate servers, each of them maintains an infinite capacity FIFO queue. Once a task or job is in a queue, it remains in that queue until its service is completed. Each server is assumed to be work conserving, i.e., a server is idle if and only if its corresponding queue is empty.

Let $Q_n(t)$ be the queue length (i.e., tasks in the queue and the server) of server n at the beginning of time-slot t. Let $A_\Sigma(t)$ denote denote the number of exogenous tasks that arrive at the beginning of time-slot t. We assume that $A_\Sigma(t)$ is an integer-valued random variable with mean of λ_Σ, which is i.i.d. across time-slots. We further assume that there is a positive probability for $A_\Sigma(t)$ to be zero. We assume that $S_n(t)$ is also an integer-valued random variable with mean μ_n, which is i.i.d. across time-slots. We also assume that $S_n(t)$ is independent across different servers as well as the arrival process. Let $S_\Sigma(t) \triangleq \sum_{n=1}^N S_n(t)$ denote the hypothetical total service process with mean of $\mu_\Sigma \triangleq \sum_{n=1}^N \mu_n$. We assume that both arrival and service processes have a finite support, i.e., $A_\Sigma(t) \leq A_{max}$ and $S_n(t) \leq S_{max}$ for all t.

We consider a set of load balancing systems parameterized by $\epsilon \triangleq N^{1-\alpha}$ such that $\lambda_\Sigma^{(\epsilon)} = \mu_\Sigma - \epsilon$ and $\mu_\Sigma = \theta(N)$. In particular, we have $\lambda_\Sigma^{(\epsilon)} = \mathbb{E}\left[\overline{A}_\Sigma\right]$, $(\sigma_\Sigma^{(\epsilon)})^2 = \mathrm{Var}(\overline{A}_\Sigma)$, $\mu_\Sigma = \mathbb{E}\left[\overline{S}_\Sigma\right]$ and $\nu_\Sigma^2 = \mathrm{Var}(\overline{S}_\Sigma)$. A load balancing policy is adopted by the dispatcher to determine to which queue the new arrivals should be sent.

In each time-slot, the order of events is as follows. First, queue lengths (or partial queue lengths) are observed. Based on these observations, a control problem is solved (i.e., the load balancing problem or the scheduling problem). Then, arrivals happen and the server processes tasks at the end of each

time slot. In particular, the evolution of the length of queue n is given by

$$Q_n(t+1) = Q_n(t) + A_n(t) - S_n(t) + U_n(t), \tag{1}$$

where $U_n(t) = \max(S_n(t) - A_n(t) - Q_n(t), 0)$ is the unused service due to an empty queue.

In this paper, we add a line on top of variables and vectors to denote steady-state (e.g., $\overline{\mathbf{Q}}$, $\overline{\mathbf{A}}$ and $\overline{\mathbf{S}}$). In order to perform our heavy-traffic analysis, we consider a set of systems parametrized by a positive parameter ϵ (or equivalently by N). In particular, the parameter ϵ captures the distance of arrival vector to a particular point on the capacity region, i.e., a smaller ϵ means a heavier load.

Definition 1. *A control policy is said to be throughput optimal if for any $\epsilon > 0$, the system is positive recurrent and all the moments of $\|\overline{\mathbf{Q}}^{(\epsilon)}\|$ are finite.*

The main convergence metric used in this paper is the Wasserstein distance metric, which is defined as follows for non-negative random variables.

$$d_W(X, Y) = \sup_{h \in \text{Lip}(1)} |\mathbb{E}[h(X)] - \mathbb{E}[h(Y)]|$$

where for a metric space $(\mathcal{S}, d)$, $\text{Lip}(1) = \{h : \mathcal{S} \to \mathbb{R}, |h(x) - h(y)| \leq d(x, y)\}$. The class Lip(1) is simple to work with but at the same time rich enough so that convergence under the Wasserstein metric implies the convergence in distribution [4].

3 Main Results

In this section, we directly apply the framework of Stein's method for heavy-traffic analysis developed in our early work [9] to study load balancing in many-server heavy-traffic regime. As can be seen from the proof, all we need to do is basically replace ϵ by $N^{1-\alpha}$ and plug in previous bounds obtained via drift-based method. This directly implies the simplicity and general applicability of our framework.

Lemma 1. *Consider a set of load balancing systems parameterized by N such that $\epsilon = N^{1-\alpha}$, $\alpha > 1$. Suppose that the load balancing policy is throughput optimal and there exists a function $g(N)$ such that*

$$\mathbb{E}\left[\|\overline{\mathbf{Q}}^{(N)}(t+1)\|_1 \|\overline{\mathbf{U}}^{(N)}\|_1\right] = O(g(N)). \tag{2}$$

Then, we have

$$d_W(N^{1-\alpha} \sum_{n=1}^{N} \overline{Q}_n^{(N)}, Z) = O(\max(g(N), N^{2-\alpha})).$$

where $Z \sim Exp(\frac{(\sigma_\Sigma^{(N)})^2 + (\nu_\Sigma^{(N)})^2}{2})$.

Proof. This result directly follows from Theorem 3 in [9] by replacing ϵ with $N^{1-\alpha}$. The full proof is presented in Appendix A. □

Lemma 2. *Consider a set of load balancing systems parameterized by N such that $\epsilon = N^{1-\alpha}$, $\alpha > 1$ with $(\sigma_\Sigma^{(N)})^2 = N\sigma_a^2$ and $(\nu_\Sigma^{(N)})^2 = N\sigma_s^2$. Suppose that the load balancing policy is throughput optimal and there exists a function $g(N)$ such that*

$$\frac{1}{N}\mathbb{E}\left[\|\overline{\mathbf{Q}}^{(N)}(t+1)\|_1 \|\overline{\mathbf{U}}^{(N)}\|_1\right] = O(g(N)). \tag{3}$$

Then, we have

$$d_W(N^{-\alpha} \sum_{n=1}^{N} \overline{Q}_n^{(N)}, Z) = O(\max(g(N), N^{1-\alpha})).$$

where $Z \sim Exp(\frac{\sigma_a^2 + \nu_s^2}{2})$.

Proof. The proof is a direct application of the framework of Stein's method developed in [9]. The full proof is presented in Appendix B. □

Lemma 3. *Consider a set of load balancing systems parameterized by N such that $\epsilon = N^{1-\alpha}$, $\alpha > 1$ with $(\sigma_\Sigma^{(N)})^2 = N^2\tilde{\sigma}_a^2$ and $(\nu_\Sigma^{(N)})^2 = N^2\tilde{\sigma}_s^2$. Suppose that the load balancing policy is throughput optimal and there exists a function $g(N)$ such that*

$$\frac{1}{N^2}\mathbb{E}\left[\|\overline{\mathbf{Q}}^{(N)}(t+1)\|_1\|\overline{\mathbf{U}}^{(N)}\|_1\right] = O(g(N)). \tag{4}$$

Then, we have

$$d_W(N^{-\alpha-1}\sum_{n=1}^{N}\overline{Q}_n^{(N)}, Z) = O(\max(g(N), N^{-\alpha})).$$

where $Z \sim Exp(\frac{\tilde{\sigma}_a^2+\tilde{\nu}_s^2}{2})$.

Proof. The proof is nearly the same as that of Lemma 2. See Appendix C □

Now, armed with the lemmas above, we can directly analyze a class of load balancing schemes in the many-server heavy-traffic regime. In particular, we focus on the class introduced in one of our early works [10], which have been well-studied via drift-based method. Based on our framework, we can directly plug in the bounds obtained in the previous work to establish new asymptotic performance.

We first summarize the key ideas behind this class as follows. More details can be found in [10].

Consider an N-dimensional cone $\mathcal{K}_\gamma$, which is finitely generated by a set of N vectors $\{\mathbf{b}^{(n)}, n \in \mathcal{N}\}$, i.e.,

$$\mathcal{K}_\gamma = \left\{\mathbf{x} \in \mathbb{R}^N : \mathbf{x} = \sum_{n\in\mathcal{N}} w_n\mathbf{b}^{(n)}, w_n \geq 0 \text{ for all } n \in \mathcal{N}\right\}, \tag{5}$$

where $\mathbf{b}^{(n)}$ is an N-dimensional vector with the nth component being 1 and γ everywhere else for some $\gamma \in [0, 1]$. It follows that, if $\gamma = 0$, the cone $\mathcal{K}_\gamma$ is the non-negative orthant of $\mathbb{R}^N$, and if $\gamma = 1$, the cone $\mathcal{K}_\gamma$ reduces to the single-dimensional line in which all the components are equal.

For a given cone $\mathcal{K}_\gamma$, we decompose $\overline{\mathbf{Q}}$ into two parts as follows

$$\overline{\mathbf{Q}} = \overline{\mathbf{Q}}_\parallel + \overline{\mathbf{Q}}_\perp,$$

where $\overline{\mathbf{Q}}_\parallel$ is the projection onto the cone $\mathcal{K}_\gamma$, referred to as the parallel component, and $\overline{\mathbf{Q}}_\perp$ is the remainder, referred to as the perpendicular component

Given a load balancing policy $\eta(t)$, we define the dispatching preference as

$$\Delta_{\eta(t)}(t) = \mathbf{P}_{\eta(t)}(t) - \mathbf{P}_{rand}(t),$$

where $\mathbf{P}_{\eta(t)}(t)$ is the dispatching distribution vector and the nth component is the probability of selecting the nth shortest queue under $\eta(t)$. $\mathbf{P}_{rand}(t)$ is the dispatching distribution under (weighted) random routing.

Definition 2 (Flexible Class Π_1)**.** *A load balancing scheme is said to be in the class Π_1 if there exists a cone $\mathcal{K}_\gamma$ such that for all $\mathbf{Q}(t) \notin \mathcal{K}_\gamma$,*

1. *there exists a $k \in \{2, 3, \ldots, N\}$ such that $\Delta_n \geq 0$ for all $n < k$ and $\Delta_n \leq 0$ for all $n \geq k$.*
2. *$\min(|\Delta_1|, |\Delta_N|) \geq \delta$ for some constant δ.*

Remark 1. *The flexibility of this class comes from three dimensions: (a) it includes JSQ and Power-of-d as special cases. Moreover, it also include many other useful policies as discussed in [12, 10]. (b) it does not require that the state-space collapse onto the line* $\mathbf{c} = \{1, 1, \ldots, 1\}$ *as in previous policies.(c) it also enables us to study the trade-off between flexibility and performance by scaling the constant* δ *and* α *with the load or the number of servers.*

Theorem 1. *Given any load balancing scheme in class* Π_1*. Consider a set of load balancing systems parameterized by* N *such that* $\epsilon = N^{1-\alpha}$*.*

1. *For any* $r \geq 2$

$$d_W(N^{1-\alpha}\sum_{n=1}^{N}\overline{Q}_n^{(N)}, Z) = O(N^{3-\alpha+\frac{\alpha-1}{r}}).$$

where $Z \sim Exp(\frac{(\sigma_\Sigma^{(N)})^2+(\nu_\Sigma^{(N)})^2}{2})$*. Thus, given* $(\sigma_\Sigma^{(N)})^2 \to \sigma_\Sigma^2$ *and* $(\nu_\Sigma^{(N)})^2 \to \nu_\Sigma^2$*, then for any* $\alpha > 3$*, the distribution of the sum queue length scaled by* $N^{1-\alpha}$ *converges to an exponential random variable with mean* $\frac{\sigma_\Sigma^2+\nu_\Sigma^2}{2}$*.*

2. *If* $(\sigma_\Sigma^{(N)})^2 = N\sigma_a^2$ *and* $(\nu_\Sigma^{(N)})^2 = N\sigma_s^2$*, we have for any* $r \geq 2$

$$d_W(N^{-\alpha}\sum_{n=1}^{N}\overline{Q}_n^{(N)}, Z) = O(N^{2-\alpha+\frac{\alpha-1}{r}}).$$

where $Z \sim Exp(\frac{\sigma_a^2+\nu_s^2}{2})$*. Thus, for any* $\alpha > 2$*, the distance approaches zero as* $N \to \infty$ [1]*.*

3. *If* $(\sigma_\Sigma^{(N)})^2 = N^2\tilde{\sigma}_a^2$ *and* $(\nu_\Sigma^{(N)})^2 = N\tilde{\sigma}_s^2$*, we have for any* $r \geq 2$

$$d_W(N^{-(\alpha+1)}\sum_{n=1}^{N}\overline{Q}_n^{(N)}, Z) = O(N^{1-\alpha+\frac{\alpha-1}{r}}).$$

where $Z \sim Exp(\frac{\tilde{\sigma}_a^2+\tilde{\nu}_s^2}{2})$*. Thus, for any* $\alpha > 1$*, the distance approaches zero as* $N \to \infty$

Proof. Based on Lemmas 1, 2 and 3, all we need to study is the term $\mathbb{E}\left[\|\overline{\mathbf{Q}}^{(N)}(t+1)\|_1\|\overline{\mathbf{U}}^{(N)}\|_1\right]$. In particular, it follows from the proof in [10] that for any scheme in class Π_1 and any $r \geq 2$

$$\begin{aligned}
&\mathbb{E}\left[\|\overline{\mathbf{Q}}^{(N)}(t+1)\|_1\|\overline{\mathbf{U}}^{(N)}(t)\|_1\right]\\
\leq&\frac{N}{\gamma}\mathbb{E}\left[\langle\overline{\mathbf{U}}, -\overline{\mathbf{Q}}_\perp^+\rangle\right]\\
\leq&\frac{N}{\gamma}\left(\mathbb{E}\left[\|\overline{\mathbf{U}}\|_{r'}^{r'}\right]\right)^{\frac{1}{r'}}\left(\mathbb{E}\left[\left\|\overline{\mathbf{Q}}_\perp^+\right\|_r^r\right]\right)^{\frac{1}{r}}.\\
\leq&\frac{N}{\gamma}\left(c_{r'}\epsilon\right)^{\frac{1}{r'}}\left(\mathbb{E}\left[\left\|\overline{\mathbf{Q}}_\perp^+\right\|_2^r\right]\right)^{\frac{1}{r}}.\\
\leq&\frac{N}{\gamma}\left(c_{r'}\epsilon\right)^{\frac{1}{r'}}\left(\mathbb{E}\left[\left\|\overline{\mathbf{Q}}_\perp\right\|_2^r\right]\right)^{\frac{1}{r}}\\
\leq&\frac{N}{\gamma\delta}(S_{max})^{\frac{1}{r}}M_rN\epsilon^{1-1/r}\\
=&\frac{L_r}{\gamma\delta}N^{3-\alpha-\frac{1-\alpha}{r}},
\end{aligned} \tag{6}$$

[1]Note that in Theorem 3 of [6], it states that the distance is upper bounded by a finite constant K times $N^{2-\alpha}$ by letting $r \to \infty$. However, this is not rigorous since K is not a finite constant. It is unbounded as $r \to \infty$.

in which L_r is a constant independent of N. Thus, we have for any $r \geq 2$

$$\mathbb{E}\left[\|\overline{\mathbf{Q}}^{(N)}(t+1)\|_1 \|\overline{\mathbf{U}}^{(N)}(t)\|_1\right] = O(N^{3-\alpha+\frac{\alpha-1}{r}}).$$

Then, the results of Theorem 1 directly follow from Lemmas 1, 2 and 3. □

4 Conclusion

In this note, we apply the recently developed framework of Stein's method for heavy-traffic analysis to study asymptotic performance of general load balancing schemes in many-server heavy-traffic regime. The main results can be easily obtained by plugging in well-known bounds obtained by drift-based method.

References

[1] Sayan Banerjee, Debankur Mukherjee, et al. Join-the-shortest queue diffusion limit in halfin–whitt regime: Tail asymptotics and scaling of extrema. *The Annals of Applied Probability*, 29(2):1262–1309, 2019.

[2] Anton Braverman. Steady-state analysis of the join-the-shortest-queue model in the halfin–whitt regime. *Mathematics of Operations Research*, 2020.

[3] Patrick Eschenfeldt and David Gamarnik. Join the shortest queue with many servers. the heavy-traffic asymptotics. *Mathematics of Operations Research*, 43(3):867–886, 2018.

[4] Alison L Gibbs and Francis Edward Su. On choosing and bounding probability metrics. *International statistical review*, 70(3):419–435, 2002.

[5] Varun Gupta and Neil Walton. Load balancing in the nondegenerate slowdown regime. *Operations Research*, 67(1):281–294, 2019.

[6] Daniela Hurtado-Lange and Siva Theja Maguluri. Load balancing system under join the shortest queue: Many-server-heavy-traffic asymptotics. *arXiv preprint arXiv:2004.04826*, 2020.

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

[8] Xin Liu and Lei Ying. A simple steady-state analysis of load balancing algorithms in the sub-halfin-whitt regime. *ACM SIGMETRICS Performance Evaluation Review*, 46(2):15–17, 2019.

[9] Xingyu Zhou and Ness Shroff. A note on stein's method for heavy-traffic analysis. *arXiv preprint arXiv:2003.06454*, 2020.

[10] Xingyu Zhou, Jian Tan, and Ness Shroff. Flexible load balancing with multi-dimensional state-space collapse: Throughput and heavy-traffic delay optimality. *Performance Evaluation*, 127:176–193, 2018.

[11] Xingyu Zhou, Fei Wu, Jian Tan, Kannan Srinivasan, and Ness Shroff. Degree of queue imbalance: Overcoming the limitation of heavy-traffic delay optimality in load balancing systems. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 2(1):1–41, 2018.

[12] Xingyu Zhou, Fei Wu, Jian Tan, Yin Sun, and Ness Shroff. Designing low-complexity heavy-traffic delay-optimal load balancing schemes: Theory to algorithms. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 1(2):39, 2017.

Appendix

A Proof of Lemma 1

Proof. Compared to the template proof in [9] of Theorem 3, we only need to update the term $\mathcal{T}_1$. In particular, we have $\mathcal{T}_1 = O(N^{2-\alpha})$ since $\epsilon = N^{1-\alpha}$. □

B Proof of Lemma 2

Proof. Replace $\epsilon\|\overline{\mathbf{Q}}^{(\epsilon)}\|_1$ in Eq.(6) of [9] by $\hat{\epsilon}\|\overline{\mathbf{Q}}^{(\epsilon)}\|_1$ with $\epsilon = N^{1-\alpha} = \mu_\Sigma - \lambda_\Sigma$ and $\hat{\epsilon} = N^{-\alpha}$. Taking expectation of both sides, yields

$$\left|\mathbb{E}\left[h(\hat{\epsilon}\|\overline{\mathbf{Q}}^{(\epsilon)}\|_1)\right] - \mathbb{E}\left[h(Z)\right]\right| = \left|\mathbb{E}\left[\frac{1}{2}\sigma^2 f_h''\left(\hat{\epsilon}\|\overline{\mathbf{Q}}^{(\epsilon)}\|_1\right) - \theta f_h'\left(\hat{\epsilon}\|\overline{\mathbf{Q}}^{(\epsilon)}\|_1\right)\right]\right| \tag{7}$$

Now, we focus on the RHS. In particular, we have

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{2}\sigma^2 f_h''\left(\hat{\epsilon}\|\overline{\mathbf{Q}}^{(\epsilon)}\|_1\right) - \theta f_h'\left(\hat{\epsilon}\|\overline{\mathbf{Q}}^{(\epsilon)}\|_1\right)\right] \\
\overset{(a)}{=}&\mathbb{E}\left[\frac{1}{2}\sigma^2 f_h''\left(\hat{\epsilon}\|\overline{\mathbf{Q}}^{(\epsilon)}\|_1\right) - \theta f_h'\left(\hat{\epsilon}\|\overline{\mathbf{Q}}^{(\epsilon)}\|_1\right) - \left(f_h\left(\hat{\epsilon}\|\overline{\mathbf{Q}}^{(\epsilon)}(t+1)\|_1\right) - f_h\left(\hat{\epsilon}\|\overline{\mathbf{Q}}^{(\epsilon)}(t)\|_1\right)\right)\right] \\
=&\mathbb{E}\left[\frac{1}{2}\sigma^2 f_h''\left(\hat{\epsilon}\|\overline{\mathbf{Q}}\|_1\right) - \theta f_h'\left(\hat{\epsilon}\|\overline{\mathbf{Q}}\|_1\right)\right] - \mathbb{E}\left[f_h\left(\hat{\epsilon}(\|\overline{\mathbf{Q}}(t)\|_1 + \|\overline{\mathbf{A}}(t)\|_1 - \|\overline{\mathbf{S}}(t)\|_1 + \|\overline{\mathbf{U}}(t)\|_1)\right) - f_h\left(\hat{\epsilon}\|\overline{\mathbf{Q}}\|_1\right)\right]
\end{aligned}$$

where (a) holds since the policy is throughput optimal and the result (a) in Lemma 1 of [9].

For the second expectation, we have

$$\begin{aligned}
&\mathbb{E}\left[f_h\left(\hat{\epsilon}(\|\overline{\mathbf{Q}}(t)\|_1 + \|\overline{\mathbf{A}}(t)\|_1 - \|\overline{\mathbf{S}}(t)\|_1 + \|\overline{\mathbf{U}}(t)\|_1)\right) - f_h\left(\hat{\epsilon}\|\overline{\mathbf{Q}}\|_1\right)\right] \\
=&\mathbb{E}\left[\hat{\epsilon}^2\frac{f_h''(\hat{\epsilon}\|\overline{\mathbf{Q}}\|_1)}{2}\left(\|\overline{\mathbf{A}}\|_1 - \|\overline{\mathbf{S}}\|_1\right)^2 + \hat{\epsilon}f_h'(\hat{\epsilon}\|\overline{\mathbf{Q}}\|_1)\left(\|\overline{\mathbf{A}}\|_1 - \|\overline{\mathbf{S}}\|_1\right)\right] \\
&+\mathbb{E}\left[\hat{\epsilon}^3\frac{f_h'''(\eta)}{6}\left(\|\overline{\mathbf{A}}\|_1 - \|\overline{\mathbf{S}}\|_1\right)^3 + \hat{\epsilon}\|\overline{\mathbf{U}}\|_1 f_h'(\hat{\epsilon}\|\overline{\mathbf{Q}}(t+1)\|_1) - \hat{\epsilon}^2\frac{f_h''(\xi)}{2}\|\overline{\mathbf{U}}\|_1^2\right] \\
=&\mathbb{E}\left[\hat{\epsilon}^2\frac{f_h''(\hat{\epsilon}\|\overline{\mathbf{Q}}\|_1)}{2}\left(N\sigma_a^2 + N\nu_s^2\right) - N\hat{\epsilon}^2 f_h'(\hat{\epsilon}\|\overline{\mathbf{Q}}\|_1)\right] \\
&+\mathbb{E}\left[\hat{\epsilon}^4\frac{f_h''(\hat{\epsilon}\|\overline{\mathbf{Q}}\|_1)}{2} + \hat{\epsilon}^3\frac{f_h'''(\eta)}{6}\left(\|\overline{\mathbf{A}}\|_1 - \|\overline{\mathbf{S}}\|_1\right)^3 + \hat{\epsilon}\|\overline{\mathbf{U}}\|_1 f_h'(\hat{\epsilon}\|\overline{\mathbf{Q}}(t+1)\|_1) - \hat{\epsilon}^2\frac{f_h''(\xi)}{2}\|\overline{\mathbf{U}}\|_1^2\right]
\end{aligned}$$

Now, let $\sigma^2 = N\hat{\epsilon}^2\left(\sigma_a^2 + \nu_s^2\right)$ and $\theta = N\hat{\epsilon}^2$ in Eq. (7), we have

$$\begin{aligned}
\left|\mathbb{E}\left[h(\hat{\epsilon}\|\overline{\mathbf{Q}}^{(\epsilon)}\|_1)\right] - \mathbb{E}\left[h(Z)\right]\right| \leq &\underbrace{\mathbb{E}\left[\left|\hat{\epsilon}^3\frac{f_h'''(\eta)}{6}\left(\|\overline{\mathbf{A}}\|_1 - \|\overline{\mathbf{S}}\|_1\right)^3\right| + \left|\hat{\epsilon}^2\frac{f_h''(\xi)}{2}\|\overline{\mathbf{U}}\|_1^2\right| + \left|\hat{\epsilon}^4\frac{f_h''(\hat{\epsilon}\|\overline{\mathbf{Q}}\|_1)}{2}\right|\right]}_{\mathcal{T}_1} \\
&+\underbrace{\mathbb{E}\left[\left|\hat{\epsilon}\|\overline{\mathbf{U}}\|_1 f_h'(\hat{\epsilon}\|\overline{\mathbf{Q}}(t+1)\|_1)\right|\right]}_{\mathcal{T}_2}
\end{aligned}$$

For $\mathcal{T}_1$, we have

$$\begin{aligned}
\mathcal{T}_1 &\leq \hat{\epsilon}^3 \frac{\|f_h'''\|}{6} \mathbb{E}\left[\overline{A}_\Sigma^3 + \overline{S}_\Sigma^3 + 3\mu_\Sigma(\overline{A}_\Sigma^2 + \overline{S}_\Sigma^2)\right] + \hat{\epsilon}^2 \frac{\|f_h''\|}{2} \mathbb{E}\left[\|\overline{\mathbf{U}}\|_1^2\right] + \hat{\epsilon}^4 \frac{\|f_h''\|}{2} \\
&\leq \frac{2\hat{\epsilon}}{3\left(N\sigma_a^2 + N\nu_s^2\right)} \mathbb{E}\left[\overline{A}_\Sigma^3 + \overline{S}_\Sigma^3 + 3\mu_\Sigma(\overline{A}_\Sigma^2 + \overline{S}_\Sigma^2)\right] + \frac{1}{2N} \mathbb{E}\left[\|\overline{\mathbf{U}}\|_1^2\right] + \frac{1}{2N}\hat{\epsilon}^2 \\
&\leq O(\hat{\epsilon}) + S_{max} \mathbb{E}\left[\|\overline{\mathbf{U}}\|_1\right] \\
&= O(N^{1-\alpha})
\end{aligned}$$

For $\mathcal{T}_2$, we have

$$\begin{aligned}
\mathcal{T}_2 &= \mathbb{E}\left[\left|\hat{\epsilon}\|\overline{\mathbf{U}}\|_1 f_h'(\hat{\epsilon}\|\overline{\mathbf{Q}}(t+1)\|_1) - \hat{\epsilon}\|\overline{\mathbf{U}}\|_1 f_h'(0)\right|\right] \\
&= \mathbb{E}\left[\left|\hat{\epsilon}^2 \|\overline{\mathbf{Q}}(t+1)\|_1 \|\overline{\mathbf{U}}\|_1 f_h''(\zeta)\right|\right] \\
&\leq \frac{1}{N} \mathbb{E}\left[\|\overline{\mathbf{Q}}(t+1)\|_1 \|\overline{\mathbf{U}}\|_1\right] \\
&= O(g(N))
\end{aligned}$$

Thus, we have

$$\left|\mathbb{E}\left[h(\epsilon\|\overline{\mathbf{Q}}^{(\epsilon)}\|_1)\right] - \mathbb{E}\left[h(Z)\right]\right| \leq \mathcal{T}_1 + \mathcal{T}_2 = O(\max(g(N), N^{1-\alpha})),$$

which completes the proof of Lemma 2. □

C Proof of Lemma 2

Proof. It follows exactly the same procedure as the proof of Lemma 2 with $\sigma^2 = N^2\hat{\epsilon}^2\left(\tilde{\sigma}_a^2 + \tilde{\nu}_s^2\right)$ and $\theta = N^2\hat{\epsilon}^2$. For $\mathcal{T}_1$, we have now $\mathcal{T}_1 = O(N^{-\alpha})$. For $\mathcal{T}_2$, we have

$$\mathcal{T}_2 \leq \frac{1}{N^2} \mathbb{E}\left[\|\overline{\mathbf{Q}}(t+1)\|_1 \|\overline{\mathbf{U}}\|_1\right] = O(g(N)).$$

□