Stein's Method for Heavy-Traffic Analysis: Load Balancing and Scheduling

Xingyu Zhou
Wayne State University

(YEQT) XIV workshop, 2021

Backgrounds: heavy-traffic analysis of queueing systems

Diffusion approximations: process-level convergence to a (regulated) Brownian motion

- A large amount of works. To name a few: [Kingman'62,Foschini and Salz'78, Reiman'84, Kelly and Laws'93, Bramson'98, Kang and Williams'12]
- It can capture the transient behavior of the queueing systems
- However, steady-state distribution convergence needs more care, i.e., *interchange-of-limits*

Can we directly work on steady state?

Backgrounds: heavy-traffic analysis of queueing systems

Drift method: set the mean drift of a test function to zero in steady state

- Introduced in [Eryilmaz and Srikant'12] with many recent follow-ups and extensions, see [Maguluri and Srikant'16, Wang et al'18, Xie and Lu'15, Wang et al'16, Zhou et al'19]
- Combined with *state space collapse*, establish first moment (and in general nth moment) optimality in steady state ☺
- However, no explicit characterization of the steady-state distribution ☹

Can we directly say something about steady-state distribution?

Backgrounds: heavy-traffic analysis of queueing systems

Transform method: choose exponential function as the test function

- Introduced in [Hurtado-Lange and Maguluri'18]
- Convergence of MGF implies convergence of stationary distribution 🙂
- However, it needs more work and no explicit characterization of convergence rate 🙁

Motivations

We are particularly interested in the following questions:

Q1: Can we directly establish convergence of stationary distribution and convergence rate in heavy traffic?

Q2: Can we maintain the same simplicity of drift method in the analysis?

Q3: Can the same analysis be applied to various systems, e.g., load balancing and scheduling?

Main Results

Stein's method allows us to address all the questions:

Q1: Can we directly establish convergence of stationary distribution and convergence rate?
- $d_W(f(\bar{Q}^{(\varepsilon)}), Z) = O(g(\varepsilon))$, convergence in Wasserstein distance

Q2: Can we maintain the same simplicity of drift method in the analysis?
- key established bounds in drift method + routine Stein's method
- i.e., strong results come for free

Q3: Can the same analysis be applied to various systems, e.g., load balancing and scheduling?
- LB: traditional heavy-traffic, many-server heavy-traffic
- Scheduling: Max-Weight

The punchline...

The punchline...

Bounds from drift method

Stein's method

Convergence of stationary distribution
with convergence rates

A gentle start: single-server system

- Consider a discrete-time single server system
- $a(t)$ $i.i.d$ integer arrival (mean λ) and $s(t)$ $i.i.d$ integer potential service (mean μ)
- $q(t+1) = q(t) + a(t) - s(t) + u(t)$
- Let $\varepsilon = \mu - \lambda$ and denote ε-parameterized system $\{q^{(\varepsilon)}(t)\}$
- Let $\bar{q}^{(\varepsilon)}$, $\bar{a}^{(\varepsilon)}$ and $\bar{s}$ be random variables in steady state
- Statistics: $\mathbb{E}\left[\bar{a}^{(\varepsilon)}\right] = \lambda^{(\varepsilon)}$, $\mathrm{Var}[\bar{a}^{(\varepsilon)}] = (\sigma_a^{(\varepsilon)})^2$, $\mathbb{E}\left[\bar{s}\right] = \mu$ and $\mathrm{Var}[\bar{s}] = \sigma_s^2$

The goal: show that $\varepsilon\bar{q}^{(\varepsilon)}$ converges to an exponential distribution as $\varepsilon \to 0$ with rate $g(\varepsilon)$

Note 1: For continuous-time systems ($M/G/1$, $G/G/1$), Stein's method was first adopted in [Gaunt and Walton'20]
Note 2: Our analysis is mainly based on the framework of Stein's method developed in [Braverman et al' 17]

A gentle start: single-server system

Theorem

Consider the single-server system as described above with $a(t) \leq A_{max}$, $s(t) \leq S_{max}$ and $Z \sim Exp(\frac{2}{(\sigma_a^{(\varepsilon)})^2 + \sigma_s^2})$. Then, there exists a constant K such that

$$d_W(\varepsilon \bar{q}^{(\varepsilon)}, Z) \leq K\varepsilon,$$

where

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

and for a metric space, $Lip(1) = \{h : \mathcal{S} \to \mathbb{R}, |h(x) - h(y)| \leq d(x, y)\}$.

Note: Convergence under Wasserstein distance implies the convergence in distribution

A routine analysis: 4 steps

Step 1: Stein's equation (or Poisson equation). $f_h'(0) = 0$ and

$$\frac{1}{2}\sigma^2 f_h''(x) - \theta f_h'(x) = h(x) - \mathbb{E}\left[h(Z)\right]$$

Intuitions: two views

- *characterizing equation* for exponential distribution: $Z \sim \text{Exp}(\frac{2\theta}{\sigma^2})$, i.e., with mean of $\frac{\sigma^2}{2\theta}$, then

$$\mathbb{E}\left[\frac{1}{2}\sigma^2 f''(Z) - \theta f'(Z) + \theta f'(0)\right] = 0 \qquad (1)$$

  holds for all functions $f : \mathbb{R}^+ \to \mathbb{R}$ with Lipschitz derivative
- *generator of RBM*: $Z \sim \text{Exp}(\frac{2\theta}{\sigma^2})$ is stationary distribution of RBM with drift θ and variance σ^2 with generator being

$$Gf(x) = \frac{1}{2}\sigma^2 f''(x) - \theta f'(x) \text{ for } x \geq 0 \text{ and } f'(0) = 0 \qquad (2)$$

  In steady-state, $\mathbb{E}_{x\sim Z} Gf(x) = 0$

A routine analysis: 4 steps

Step 2: Generator coupling. replace x in Stein's equation by $\varepsilon\bar{q}^{(\varepsilon)}$

$$\mathbb{E}\left[h(\varepsilon\bar{q})\right] - \mathbb{E}\left[h(Z)\right] = \mathbb{E}\left[\frac{1}{2}\sigma^2 f_h''(\varepsilon\bar{q}) - \theta f_h'(\varepsilon\bar{q})\right]$$

Add the 'generator' (or drift) of the single-server system (which is zero in steady state) to RHS, i.e.,

$$\mathbb{E}\left[h(\varepsilon\bar{q})\right] - \mathbb{E}\left[h(Z)\right] = \mathbb{E}\left[\frac{1}{2}\sigma^2 f_h''(\varepsilon\bar{q}) - \theta f_h'(\varepsilon\bar{q}) - \left(f_h(\varepsilon\bar{q}(t+1)) - f_h(\varepsilon\bar{q}(t))\right)\right]$$

Intuitions: reduces to the distance between two generators – one is generator for RBM, the other is our single-server system

A routine analysis: 4 steps

Step 3: Taylor expansion. over the generator of single-server system in the hope to recover the structure of generator of RBM.

$$
\begin{aligned}
&(f_h(\varepsilon\bar{q}(t+1)) - f_h(\varepsilon\bar{q}(t))) \\
=&\mathbb{E}\left[\varepsilon^2 \frac{f_h''(\varepsilon\bar{q})}{2}\left((\sigma_a^{(\varepsilon)})^2 + \sigma_s^2\right) - \varepsilon^2 f_h'(\varepsilon\bar{q})\right] \\
&+ \mathbb{E}\left[\varepsilon^3 \frac{f_h'''(\eta)}{6}(\bar{a} - \bar{s})^3 + \varepsilon\bar{u} f_h'(\varepsilon\bar{q}(t+1)) - \varepsilon^2 \frac{f_h''(\xi)}{2}\bar{u}^2\right] \\
&+ \mathbb{E}\left[\varepsilon^4 \frac{f_h''(\varepsilon\bar{q})}{2}\right]
\end{aligned}
$$

Idea: set $\sigma^2 = \varepsilon^2\left((\sigma_a^{\varepsilon})^2 + \sigma_s^2\right)$ and $\theta = \varepsilon^2$ in Stein's equation and hence green term cancels

A routine analysis: 4 steps

Step 4: Gradient bounds. Now we have

$$
\begin{aligned}
&|\mathbb{E}\left[h(\varepsilon\bar{q})\right] - \mathbb{E}\left[h(Z)\right]| \\
&\leq \underbrace{\mathbb{E}\left[\left|\varepsilon^4 \frac{f_h''(\varepsilon\bar{q})}{2}\right| + \left|\varepsilon^3 \frac{f_h'''(\eta)}{6}(\bar{a}-\bar{s})^3\right| + \left|\varepsilon^2 \frac{f_h''(\xi)}{2}\bar{u}^2\right|\right]}_{\mathcal{T}_1} \\
&+ \underbrace{\mathbb{E}\left[|\varepsilon\bar{u}f_h'(\varepsilon\bar{q}(t+1))|\right]}_{\mathcal{T}_2}
\end{aligned}
$$

Tools: standard gradient bounds for the solution of Stein's equation, i.e., $\|f_h''\| \leq \frac{\|h'\|}{\theta}$ and $\|f_h'''\| \leq \frac{4\|h'\|}{\sigma^2}$ (noting that $\|h'\| \leq 1$)

Results:

- $\mathcal{T}_1 \leq K\varepsilon$ by gradient bounds and boundedness assumption
- $\mathcal{T}_2 \overset{(a)}{=} \mathbb{E}\left[|\varepsilon\bar{u}f_h'(\varepsilon\bar{q}(t+1)) - \varepsilon\bar{u}f_h'(0)|\right] = \mathbb{E}\left[|\varepsilon\bar{u}(t)f_h''(\zeta)\varepsilon\bar{q}(t+1)|\right] = 0$, where (a) holds since $f_h'(0) = 0$

A generalization

Assumption (Light-tail assumption)

The arrival process $a(t)$ and service process $s(t)$ satisfy that

$$\mathbb{E}\left[e^{\theta_1 a(t)}\right] \leq D_1 \text{ and } \mathbb{E}\left[e^{\theta_2 s(t)}\right] \leq D_2,$$

for some constants $\theta_1 > 0$, $\theta_2 > 0$, $D_1 < \infty$ and $D_2 < \infty$ that are all independent of ε.

Theorem

Consider a single-server system that satisfies the light-tail assumption. Let $Z \sim Exp(\frac{2}{(\sigma_a^{(\varepsilon)})^2+\sigma_s^2})$, then

$$d_W(\varepsilon\bar{q}^{(\varepsilon)}, Z) = O(\varepsilon \log \frac{1}{\varepsilon}).$$

A particular case: $M/M/1$

Theorem

Consider an $M/M/1$ system with $\lambda = \mu - \varepsilon$. Let $Z \sim Exp(\frac{1}{\lambda})$, then

$$d_W(\varepsilon \bar{q}^{(\varepsilon)}, Z) \leq \frac{2}{3}\varepsilon.$$

Idea: follow the same routine analysis and use the generator of $M/M/1$ system instead

Load balancing

A discrete-time LB model with 1 dispatcher and N queues

- $A_\Sigma(t)$ $i.i.d$ total arrival at time t
- $S_\Sigma(t) := \sum_{n=1} S_n(t)$, each n $i.i.d$ potential service for queue n
- At each time t, one queue is selected
- $Q_n(t+1) = Q_n(t) + A_n(t) - S_n(t) + U_n(t)$
- Statistics: $\lambda_\Sigma^{(\varepsilon)} = \mu_\Sigma - \varepsilon$, $\lambda_\Sigma^{(\varepsilon)} = \mathbb{E}\left[\overline{A}_\Sigma\right]$, $(\sigma_\Sigma^{(\varepsilon)})^2 = \mathrm{Var}(\overline{A}_\Sigma)$, $\mu_\Sigma = \mathbb{E}\left[\overline{S}_\Sigma\right]$ and $\nu_\Sigma^2 = \mathrm{Var}(\overline{S}_\Sigma)$

The goal: show that $\varepsilon \sum_{n=1}^{N} \overline{Q}_n^{(\varepsilon)}$ converges to an exponential distribution as $\varepsilon \to 0$ with rate $g(\varepsilon)$ under a class of policies

Load balancing: general results

Theorem

Consider a set of load balancing systems parameterized by ε. Suppose that the load balancing policy is throughput optimal and there exists a function $g(\varepsilon)$ such that

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

Then, we have

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

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

Implication: the key is to bound the cross term, which is in fact the key term in drift method, i.e., *state-space collapse*

LB in classical heavy-traffic regime

We consider N **is fixed and** $\varepsilon \to 0$

Theorem

For a class of LB policies (including JSQ, Pod). We have for all $\varepsilon \leq \varepsilon_0$, $\varepsilon_0 \in (0, \mu_\Sigma)$

$$\mathbb{E}\left[\|\overline{\mathrm{Q}}^{(\varepsilon)}(t+1)\|_1 \|\overline{\mathrm{U}}^{(\varepsilon)}\|_1\right] \leq K\varepsilon \log(1/\varepsilon), \tag{4}$$

and

$$d_W(\varepsilon \sum_{n=1}^{N} \overline{Q}_n^{(\varepsilon)}, Z) \leq K\varepsilon \log(1/\varepsilon).$$

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

Note 1: one can directly utilize the bounds on the cross term for specific policy, e.g., JSQ in [Hurtado-Lange and Maguluri'20]
Note 2: we establish the bounds for general policies

LB in many-server heavy-traffic regime

We consider $\varepsilon = N^{1-\alpha}$ with $\alpha > 1$ and $\mu_\Sigma = cN$ for some $c > 0$

- One example: N homogeneous servers with rate 1, then in the regime above, $\rho = 1 - N^{-\alpha}$

We will replace ε by N in our parameterized systems and consider two scalings:

- $(\sigma_\Sigma^{(N)})^2 = N\sigma_a^2$ and $(\nu_\Sigma^{(N)})^2 = N\sigma_s^2$: 'independent' sum
- $(\sigma_\Sigma^{(N)})^2 = N^2\tilde{\sigma}_a^2$ and $(\nu_\Sigma^{(N)})^2 = N^2\tilde{\sigma}_s^2$: 'correlated' sum

The goal: show that $N^{f(\alpha)} \sum_{n=1}^{N} \overline{Q}_n^{(N)}$ converges to an exponential distribution as $N \to \infty$ with rate $g(N)$ under a class of policies

LB in many-server heavy-traffic regime

Lemma (Independent case)

Consider a set of load balancing systems parameterized by N such that $\varepsilon = N^{1-\alpha}$, $\alpha > 1$ with $\mu_\Sigma = \theta(N)$ and $A_{max} = \theta(N)$. Assume that $(\sigma_\Sigma^{(N)})^2 = N\sigma_a^2$ 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

$$\frac{1}{N}\mathbb{E}\left[\|\overline{\mathrm{Q}}^{(N)}(t+1)\|_1\|\overline{\mathrm{U}}^{(N)}\|_1\right] = O(g(N)). \tag{5}$$

Then, we have

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

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

LB in many-server heavy-traffic regime

Lemma (Correlated case)

Consider a set of load balancing systems parameterized by N such that $\varepsilon = N^{1-\alpha}$, $\alpha > 1$ with $\mu_\Sigma = \theta(N)$ and $A_{max} = \theta(N)$. Assume that $(\sigma_\Sigma^{(N)})^2 = N^2\tilde{\sigma}_a^2$ 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

$$\frac{1}{N^2}\mathbb{E}\left[\|\overline{\mathrm{Q}}^{(N)}(t+1)\|_1\|\overline{\mathrm{U}}^{(N)}\|_1\right] = O(g(N)). \tag{6}$$

Then, we have

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

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

LB in many-server heavy-traffic regime: JSQ and Pod

Theorem (Independent case)

Consider a set of load balancing systems parameterized by N such that $\varepsilon = N^{1-\alpha}$, $\mu_\Sigma = \theta(N)$, $A_{max} = \theta(N)$. Assume that $(\sigma_\Sigma^{(N)})^2 = N\sigma_a^2$ and $(\nu_\Sigma^{(N)})^2 = N\sigma_s^2$. Let $Z \sim Exp(\frac{2}{\sigma_a^2+\nu_s^2})$.
Then, under JSQ, we have

$$d_W(N^{-\alpha}\sum_{n=1}^{N}\overline{Q}_n^{(N)}, Z) = O(N^{4-\alpha}\log N).$$

Under Power-of-d with homogeneous servers, we have

$$d_W(N^{-\alpha}\sum_{n=1}^{N}\overline{Q}_n^{(N)}, Z) = O(N^{4.5-\alpha}\log N).$$

Note: similar results are also obtained in [Hurtado-Lange and Maguluri'20]

LB in many-server heavy-traffic regime: JSQ and Pod

Theorem (Correlated case)

Consider a set of load balancing systems parameterized by N such that $\varepsilon = N^{1-\alpha}$, $\mu_\Sigma = \theta(N)$, $A_{max} = \theta(N)$. Assume that $(\sigma_\Sigma^{(N)})^2 = N^2 \tilde{\sigma}_a^2$ and $(\nu_\Sigma^{(N)})^2 = N^2 \tilde{\sigma}_s^2$. Let $Z \sim Exp(\frac{2}{\tilde{\sigma}_a^2 + \tilde{\nu}_s^2})$.
Then, under JSQ, we have

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

Under Power-of-d with homogeneous servers, we have

$$d_W(N^{-\alpha-1} \sum_{n=1}^{N} \overline{Q}_n^{(N)}, Z) = O(N^{3.5-\alpha} \log N).$$

Comparison of two heavy-traffic regimes

- **Classical heavy-traffic regime**: N fixed, $\varepsilon \to 0$: JSQ and Pod have the same convergence rate, i.e.,

$$d_W(\varepsilon \sum_{n=1}^{N} \overline{Q}_n^{(\varepsilon)}, Z) \leq K\varepsilon \log(1/\varepsilon).$$

- **Many-server heavy-traffic regime**: JSQ and Pod have different convergence rates, i.e.,

$$\text{(JSQ)} \quad d_W(N^{-\alpha} \sum_{n=1}^{N} \overline{Q}_n^{(N)}, Z) = O(N^{4-\alpha} \log N)$$

$$\text{(Pod)} \quad d_W(N^{-\alpha} \sum_{n=1}^{N} \overline{Q}_n^{(N)}, Z) = O(N^{4.5-\alpha} \log N),$$

Implication: many-server heavy-traffic regime is better at differentiating the *strongness* of state-space collapse

Scheduling: Max-Weight

A discrete-time N-queue model...

- $\boldsymbol{\lambda} = (\lambda_n)_n$ and $\boldsymbol{\sigma^2} = (\sigma_n^2)_n$ for arrival and $\boldsymbol{\mu} = (\mu_n)_n$ and $\boldsymbol{\nu^2} = (\nu_n^2)_n$ for the service
- Capacity region: $\mathcal{R} = \{\mathrm{r} \geq 0 : \langle \mathrm{c}^{(k)}, \mathrm{r} \rangle \leq b^{(k)}, k = 1, 2, \ldots, K\}$
- kth face: $\mathcal{F}^{(k)} \triangleq \{\mathrm{r} \in \mathcal{R} : \langle \mathrm{c}^{(k)}, \mathrm{r} \rangle = b^{(k)}\}$
- We fix a particular $\mathcal{F}^{(k)}$ and a point $\boldsymbol{\lambda}^{(k)} \in \mathrm{Relint}(\mathcal{F}^{(k)})$
- Let $\boldsymbol{\lambda}^{(\varepsilon)} \triangleq \boldsymbol{\lambda}^{(k)} - \varepsilon \mathrm{c}^{(k)}$

The goal: show that $\varepsilon \langle \mathrm{c}^{(k)}, \overline{\mathrm{Q}}^{(\varepsilon)} \rangle$ converges to an exponential distribution as $\varepsilon \to 0$ with rate $g(\varepsilon)$ under Max-Weight

Scheduling: Max-Weight

Theorem

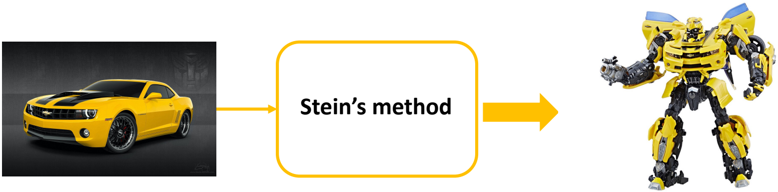
Consider a set of scheduling systems described above that are parametrized by ε defined above. Suppose the scheduling policy is MaxWeight and $Z \sim Exp(\frac{2}{\langle (\mathrm{c}^{(k)})^2, (\boldsymbol{\sigma}^{(\varepsilon)})^2 \rangle})$, then

$$d_W(\varepsilon \langle \mathrm{c}^{(k)}, \overline{\mathrm{Q}}^{(\varepsilon)} \rangle, Z) = O\left(\varepsilon \log \frac{1}{\varepsilon}\right).$$

Proof idea: 4 steps Stein's method (routine) + key bounds from drift method (e.g., [Eryilmaz and Srikant'12, Hurtado-Lange and Maguluri'20])

Conclusion

- Stein's method provides a powerful way of obtaining stronger results by utilizing results of drift method
- This can be readily applied to LB: classical heavy-traffic regime and many-server heavy-traffic regime
- This can be readily applied to scheduling: Max-Weight
- **Open problem**: what if $1 < \alpha \leq 4$ in the many-server heavy-traffic regime?



Bounds from drift method

Convergence of stationary distribution with convergence rates

Thank you!

Q & A