

Advanced Stochastic Processes.

David Gamarnik

LECTURE 25

Final notes and ongoing research questions and resources

26.1. GJN and open questions

26.1.1. A brief summary of GJN heavy-traffic theory

We have described in previous lecture the GJN model. It is called "Generalized" because original (Jackson) network assumes exponential interarrival times and exponential service times. As we will mention shortly this dramatically simplifies the picture.

From the arrival rates λ_j we first formulate the vector of so-called "effective" arrival rates, $\bar{\lambda}_j$ which is defined to be the unique solution to the traffic equation

$$\bar{\lambda}_j = \lambda_j + \sum_k p_{jk} \bar{\lambda}_k$$

Can you interpret the meaning of $\bar{\lambda}_j$? In vector form

$$\bar{\lambda} = \lambda + P^t \bar{\lambda}$$

or

$$\bar{\lambda} = [I - P^t]^{-1} \lambda.$$

From this we formulate workload ρ_j in every server j as $\rho_j = \bar{\lambda}_j / \mu_j$. It turns out that if $\rho_j < 1$ in every server, then the system converges to a steady-state and in steady-state the proportion of time server j is working is ρ_j .

Matrix $R = I - P^t$ is the key for formulating the corresponding Skorohod mapping problem: given $x \in D^J[0, \infty) \triangleq D$ find a unique pair $y, z \in D$ such that

$$\begin{aligned} z &= x + Ry \geq 0 \\ dy &\geq 0, y(0) = 0 \\ z_j dy_j &= 0. \end{aligned}$$

Theorem 26.1. *The Skorohod mapping problem has a unique solution $y = \Psi(x), z = \Phi(x)$. Moreover, the mappings Ψ, Φ are Lipschitz continuous.*

How is this used in GJN to formulate Fluid Model and Heavy-Traffic approximations? Again a corresponding input process $X(t)$ is identified such that $Z(t) = X(t) + [I - P^t]I(t)$. To obtain the fluid model we need to consider

$$\frac{Z(nt)}{n} = \frac{X(nt)}{n} + [I - P^t] \frac{I(nt)}{n}$$

Theorem 26.2. *Assume $Q(0) = nq$ for some vector $q \geq 0$. $\frac{X(nt)}{n}$ converges to a linear function x with rate $(\rho - e)t$, where $\rho = (\rho_j)$ and e is the J -dimensional unit vector. The unique solution $y = \Psi(x)$, $z = \Phi(x)$ are piece-wise linear functions. Moreover $z(t)$ becomes a zero function after some finite time period.*

A heavy-traffic approximation is obtained similarly to G/G/1 queue as well. First when we take as a input to a multi-dimensional Skorohod problem a J -dimensional Brownian motion $W(t)$, on the output we obtain $\Phi(W)$ which is also called a (multidimensional) Reflected Brownian Motion. Then we consider a critically loaded system with $\rho_j = 1$ for all servers j . For each external arrival process j we again rescale interarrival times so that the resulting $\rho_j = 1 - \frac{\theta_j}{\sqrt{n}}$ for some constants θ_j . It is then showed that $X(nt)/\sqrt{n}$ converges weakly to a multidimensional Brownian motion, and, as a result, $Z(nt)/\sqrt{n}$ and $Q(nt)/\sqrt{n}$ converge weakly to a Reflected Brownian motion. The drift and covariance matrix of this RBM is explicitly computed knowing θ_j and parameters of the networks: arrival rates, service rates, coefficients of variations for interarrival and service times and matrix P . This development is primarily due to Harrison and Reiman [3] and Reiman [6]. The second paper, which is the paper developing heavy-traffic approximation of GJN is included in the reading materials.

What can be said about the distribution of the RBM as $t \rightarrow \infty$? Recall that in G/G/1 case this is simply the exponential distribution. Things are far more complicated here. It is known Harrison and Williams [4] (I only have hard-copy) that the stationary distribution exists if and only if $\rho_j < 1$, but it does not necessarily have exponential distribution. It turns out that there are special cases when it does. This was established also by Harrison and Williams [5].

Theorem 26.3. *Suppose the coefficient of variations of interarrival and service times are equal to unity (for example when arrival and service processes are Poisson). Then the stationary distribution of the RBM is exponential, with rates explicitly computable from the parameters of the network.*

As in the case of G/G/1 queueing system, the GJN itself converges to a steady-state when $\rho_j < 1$. Thus it has some corresponding steady-state random vectors $Q(\infty)$, $Z(\infty)$ corresponding to $t = \infty$. Is there a relationship between these vectors and steady-state of the RBM? We have established this only recently jointly with Zeevi [2] (included in the reading materials). Namely that $Q^n(\infty)/\sqrt{n}$, $Z^n(\infty)/\sqrt{n}$ converge weakly to the stationary RBM. In the special case when stationary RBM has exponential distributions, this implies exponential distribution for the limits of queue length and workload. One important "side result" of our analysis was exponential decay for these quantities: $\mathbb{P}(\|Q(\infty)\| > x) \approx e^{-c(1-\rho)x}$.

26.1.2. Open problems

As we have mentioned, simulations show that the predictions of the heavy-traffic theory are pretty good. But there is not good theory around it.

Open problem 1. [Quality of heavy-traffic approximations]. *What is the quality of heavy-traffic approximations in $G/G/1$ and more interestingly GJN. Specifically, try to identify the rate with which*

$$\frac{Q^n(nt)}{\sqrt{n}} - RBM$$

converges to zero in various senses. Even better obtain an explicit bound on $\mathbb{E}[\frac{Q^n(nt)}{\sqrt{n}} - RBM]$ which converges to zero. The corresponding question in stationary regime. With which rate does

$$\mathbb{E}[(1 - \rho)Q(\infty)] - \mathbb{E}[RBM]$$

converges to zero?

Quite often the only problem to solve some performance analysis problem related to a queueing system is via simulations. But how long do we need to simulate in order to get a good confidence? This critically depends on what time does it take for a queueing network to converge to stationarity. It is known that simulations becomes harder and harder when system is in heavy-traffic $\rho \approx 1$.

Open problem 2. [Relaxation times of queueing networks] *What is the rate of convergence to stationarity for GJN? For example, identify the rate of the convergence*

$$\left| \mathbb{E}[Z(t)] - \mathbb{E}[Z(\infty)] \right| \xrightarrow{t \rightarrow \infty} 0,$$

or in distribution sense

$$\left| \mathbb{P}(Z(t) \leq x) - \mathbb{P}(Z(\infty) \leq x) \right| \xrightarrow{t \rightarrow \infty} 0.$$

The problem is especially important in heavy-traffic regime when $\rho \approx 1$. Heuristic arguments show that the relaxation times is $\approx 1/(1 - \rho)^3$, but there is no solid theory.

A lot of research is devoted to studying large deviations theory of queueing systems. Here the question is: establish that $\mathbb{P}(\|Q\| > x) \approx e^{-cx}$ and identify the leading constant c . The full theory is developed for *acyclic* GJN, but not in general case.

Open problem 3. [Large deviations for GJN] *Establish whether $\mathbb{P}(\|Q\| > x) \approx e^{-cx}$ in GJN and identify the leading constant c . The problem can be formulated both as a process level large deviations problem and as large deviations for steady-state.*

26.2. "Hot" research area in stochastic processes – Call centers

Call centers can be modeled as queues which tend to operate in heavy-traffic regime. Back in 80's Halfin and Whitt showed that the appropriate model is a queueing model with n parallel servers and arrival rate $\approx \lambda n$. The service rate is assume to "match" the arrival rate $\mu = \lambda$, but the actual number of servers needs to be increased by a safety stock $\beta\sqrt{n}$. This results in a heavy-traffic regime with utilization $\rho \approx 1 - \beta/\sqrt{n}$. Let $Q(t)$ be the number of jobs in the system *excluding* those in service. Namely the total number is $Q(t) + n + \beta\sqrt{n}$. Let $I(t)$ be the number of idle servers.

Theorem 26.4 (Halfin-Whitt Theory). *Suppose the arrival and service processes are Poisson. Then in steady-state*

$$\mathbb{P}\left(\frac{Q(\infty)}{\sqrt{n}} > x | Q(\infty) > 0\right) \rightarrow e^{-\beta x},$$

$$\mathbb{P}\left(\frac{I(\infty)}{\sqrt{n}} > x | I(\infty) > 0\right) \rightarrow \frac{\Phi(\beta - x)}{\Phi(\beta)}.$$

Moreover, on the process level, $Q(t)/\sqrt{n}$ converges weakly to a Brownian motion with a negative drift $-\mu\beta$, and $I(t)/\sqrt{n}$ converges weakly to a diffusion with drift $-\mu(x + \beta)$.

Halfin-Whitt theory is not developed for non-Poisson case. Moreover, in multitype case (several streams of calls with different service requirements) the question is optimal control of the call center. Finally, these days many call-centers have agents with different skills. Skill-based routing is one of the most "burning" practical problem of modern days call centers.

26.3. Other applications of heavy-traffic theory

Production processes (manufacturing, specifically semiconductor wafer fabrication labs, inventory, supply-chains); communication theory (switches, TCP protocols, web servers), service processes (real queues).

26.4. Other topics in stochastic processes not touched in this course

Regenerative processes, Markov processes in general state space, stochastic differential equations, rare-event simulations. For this topic see course Stat. 271 in Harvard taught by Jose Blanchet in Statistics department.

Mathematical finance. Some courses here in Sloan and Harvard.

Heavy-tailed processes. CLT does not hold when the variance is infinite (for example power law distribution $\mathbb{P}(X > x) \approx c/x^\alpha, \alpha \leq 2$). Instead, the sum of i.i.d. sequence appropriately rescaled converges to a so-called stable process, with a limiting distribution which is not Gaussian, but something else called stable distribution. A special case of a stable distribution is Cauchy distribution

$$\mathbb{P}(X \leq x) = \int_{-\infty}^x \frac{dt}{\pi(1+t^2)}$$

The functional limit theorems also hold. Instead of a Brownian motion, one obtains a process called *Fractional Brownian motion*, which has variance growing at a non-linear rate $O(t^H)$, H is called the Hurst parameter.

Strong approximations. Donsker Theorem states that a random walk suitably rescaled "looks" increasingly like a Brownian motion in distribution. It turns out that one can "embed" a Brownian motion into a random walk.

Theorem 26.5. *Suppose $X_n, n \geq 1$ is an i.i.d. sequence such that $\mathbb{E}[e^{\theta X_1}] < \infty$ for some $\theta > 0$. Then there exists a Brownian motion with drift $m = \mathbb{E}[X_1]$ and variance $\sigma^2 = \text{Var}(X_1^2)$, such*

that for $S_n = \sum_{1 \leq k \leq n}$, for every $T > 0$

$$\|S_n(t) - W(t)\|_T \stackrel{a.s.}{=} O(\log T).$$

In other words, a Brownian motion can be embedded into a random walk such that the distance between the two grows only logarithmically with time.

Branching processes Sub-critical, critical and super-critical branching processes. Multi-type branching processes. Applications to error-correction, population biology and statistical physics. Reconstruction problems.

26.5. Future courses

26.5.1. Queues: theory and applications. 15.072

26.5.2. A special seminar on applied probability, 15.098/6.979

This is a doctoral student seminar covering current topics in applied probability and stochastic processes. For the offering of Spring 2006, the topics covered will be at the interface of statistical physics (theory of spin glasses), probability (local weak convergence), artificial intelligence (belief propagation), operations research and computer science (random integer programming, random 3-SAT) and electrical engineering (Low density parity check codes).

26.6. Resources for Applied Probability and Stochastic Process

26.6.1. Web resources

- Applied Probability Society of INFORMS.
<http://appliedprob.society.informs.org/>
(jobs, publications, conferences, etc.) Student membership is \$10, you DO NOT have to be a member of INFORMS.
- Probability Web
<http://www.mathcs.carleton.edu/probweb/probweb.html>
(jobs, publications, conferences, etc.)
- Probability ArXive
<http://front.math.ucdavis.edu/math.PR> and <http://arxiv.org/list/math.PR/recent>
– online papers before they appear in journals. Many journals also publish their final versions.

26.6.2. Key conferences

- Applied probability cluster of INFORMS meeting.
- INFORMS Applied Probability Conference (bi-annual) (next will be in Eindhoven in 2007)
<http://appliedprob.society.informs.org/INFORMS2007/Index.html>
- Conference on Stochastic Networks
<http://www.comm.csl.uiuc.edu/~srikant/stochnet.htm>
- Conference on Stochastic Processes and Applications
<http://www.proba.jussieu.fr/spa06/index.php>
- Bernoulli Society meeting
<http://www.imub.ub.es/events/wc2004/main.html>

26.6.3. Key journals

Annals of Applied Probability
Annals of Statistics
Stochastic Processes and Their Applications
Probability Theory and the Related Fields

26.6.4. Societies

- INFORMS and Applied Probability Section of INFORMS
- Institute for Mathematical Statistics
<http://www.imstat.org/>
- Bernoulli Society
<http://isi.cbs.nl/BS/bshome.htm>

26.7. Additional reading materials

- Chapter 6 of Chen & Yao book [1] from the course packet.

BIBLIOGRAPHY

1. H. Chen and D. Yao, *Fundamentals of queueing networks: Performance, asymptotics and optimization*, Springer-Verlag, 2001.
2. D. Gamarnik and A. Zeevi, *Validity of heavy traffic steady-state approximations in open queueing networks*, Submitted to Ann. Appl. Prob.
3. J. M. Harrison and M. I. Reiman, *Reflected Brownian motion on an orthant*, Annals of Probability **9** (1981), 302–308.
4. J. M. Harrison and R. J. Williams, *Brownian models of open queueing networks with homogeneous customer populations*, Stochastics **22** (1987), 77–115.
5. ———, *Multidimensional reflected brownian motions having exponential stationary distributions*, Annals of Applied Probability **15** (1987), no. 1, 115–137.
6. M. I. Reiman, *Open queueing networks in heavy traffic*, Mathematics of Operations Research **9** (1984), 441–458.