Design of Statistical Experiments

Exploiting Heavy-Traffic Limits for Queueing Processes

IEOR 4615, Service Engineering, Professor Whitt

Lecture 15, March 12, 2015

OUTLINE

- Estimation for Queueing Processes: Dependence!
- Exploit CLT to Estimate Confidence Intervals
- The Heavy-Traffic Limit for the $GI/GI/s/\infty$ Model
- How Scaling Affects the Asymptotic Variance
- Putting It All Together:
    - Approximations for the Steady-State Mean and the Asymptotic Variance.
    - Approximations for the Required Interval Length
        - Run length in a simulation.
        - Measurement Interval (Sample Size) with System Data

Estimating the Expected Number in the System

- Given **stationary stochastic process:** $\{Q(t) : t \geq 0\}$
- Our goal: **Estimate the unknown mean** $E[Q(0)]$, using
  - **sample mean:** $\bar{Q}(t) \equiv \frac{1}{t} \int_0^t Q(s)\, ds$
- But the observations are typically **highly dependent**!
- Use **batch means** as in Lecture 4.
- **How much data do we need?**

Exploit CLT to Construct Confidence Intervals for Estimates

- Given **stationary stochastic process:** $\{Q(t) : t \geq 0\}$
- Our goal: **Estimate the unknown mean** $E[Q(0)]$, using
  - **sample mean:** $\bar{Q}(t) \equiv \frac{1}{t} \int_0^t Q(s)\, ds$
- **Use CLT:** $t^{1/2} \left(\bar{Q}(t) - E[Q(0)]\right) \Rightarrow N(0, \sigma^2)$, where
  - $\sigma^2 \equiv \lim_{t\to\infty} tVar(\bar{Q}(t)) = 2 \int_0^\infty C(s)\, ds$ (**asymptotic variance**)
  - $C(t) \equiv Cov(Q(0), Q(t)) \equiv E[Q(0)Q(t)] - E[Q(0)]E[Q(t)]$
- **Use normal approximation:** $\bar{Q}(t) \approx N(E[Q(0)], \sigma^2/t)$.
  - Use HT limit to estimate both the mean $E[Q(0)]$ and the asymptotic variance σ^2.
  - (With data, use the method of batch means, as in Lecture 4.)
  - See §2 of "Planning Queueing Simulations," 1989.

95% Confidence Intervals (CI's)

- **Use normal approximation:** $\bar{Q}(t) \approx N(E[Q(0)], \sigma^2/t)$.
- Given **asymptotic variance** σ^2 and **interval length** t,
- **Confidence Interval:** $[\bar{Q}(t) - z_{\beta/2}\sqrt{\sigma^2/t}, \bar{Q}(t) + z_{\beta/2}\sqrt{\sigma^2/t}]$
  - where $P(-z_{\beta/2} < N(0,1) < z_{\beta/2}) = 1 - \beta = 0.95 \quad (\beta = 0.05)$
- **absolute width of CI:** $w_a(\beta) \equiv \frac{2z_{\beta/2}\sigma}{\sqrt{t}}$;
- **relative width of CI:** $w_r(\beta) \equiv \frac{w_a(\beta)}{E[Q(0)]} = \frac{2z_{\beta/2}\sigma}{\sqrt{t}E[Q(0)]}$.
  - **required values of** t: $t_a(\epsilon, \beta) \equiv \frac{4\sigma^2 z_{\beta/2}^2}{\epsilon^2}$ and $t_r(\epsilon, \beta) \equiv \frac{4\sigma^2 z_{\beta/2}^2}{\epsilon^2 E[Q(0)]^2}$.
- Use HT limit to estimate both $E[Q(0)]$ and the asymptotic variance σ^2.

Review: Conventional HT Limit in $GI/GI/s/\infty$ Model

- i.i.d. interarrival times T_k: $E[T] \equiv \frac{1}{\lambda}$, $c_a^2 \equiv \frac{Var(T)}{E[T]^2}$
- i.i.d. service times S_k: $E[S] \equiv \frac{1}{\mu}$, $c_s^2 \equiv \frac{Var(S)}{E[S]^2}$
- Let traffic intensity $\rho \equiv \lambda/s\mu \uparrow 1$ (by multiplying T_k be constants).
- Let $W(\rho)$ be the steady-state waiting time before starting service.
- **The distribution of $W(\rho)$ is complicated except for special cases.**
- **HT limit:** $(1-\rho)W(\rho) \Rightarrow W^*$ (with exponential distribution)
- $E[W(\rho)] \approx \frac{\rho E[S](c_a^2+c_s^2)}{2(1-\rho)}$ and $P(W(\rho) > x) \approx e^{-2(1-\rho)x/(c_a^2+c_s^2)}$, $x \geq 0$.
- (The mean is exact for $M/M/1/\infty$ and $M/GI/1/\infty$ special cases.)
- Refined approx.: $E[W(\rho)] \approx \left(\frac{c_a^2+c_s^2}{2}\right) E[W(\rho; M/M/s/\infty)]$ (QNA)

Stochastic-Process Limit for $GI/GI/s/\infty$

- i.i.d. interarrival times T_k: $E[T] \equiv \frac{1}{\lambda}$, $c_a^2 \equiv \frac{Var(T)}{E[T]^2}$
- i.i.d. service times S_k: $E[S] \equiv \frac{1}{\mu}$, $c_s^2 \equiv \frac{Var(S)}{E[S]^2}$
- Let traffic intensity $\rho \equiv \lambda / s\mu \uparrow 1$ (by multiplying T_k be constants).
- Let $Q_\rho(t)$ be the number in queue at time t.
- **HT limit:** $(1-\rho)Q_\rho(t(1-\rho)^{-2}) \Rightarrow R(t; a, b)$ as $\rho \uparrow 1$
  - Both **time scaling** and **space scaling**!
  - The limit process is **reflected Brownian motion (RBM)**.
  - The drift is $a = -s$; the variance constant is $b = s(c_a^2 + c_s^2)$.
  - $Q_\rho(t) \approx \left(\frac{b}{|a|(1-\rho)}\right) R\left(a^2(1-\rho)^2 t; -1, 1\right), \quad t \geq 0.$
  - (See §4.3 and equation (34) in "Planning Queueing Simulations," 1989.)

How does scaling affect the asymptotic variance?

- Recall that $\sigma^2 = 2 \int_0^\infty C(s)\, ds$.
- If $Q_{y,z}(t) \equiv yQ(zt), \quad t \geq 0$, then
- $E[Q_{y,z}(t) = yE[Q(t)], C_{y,z}(t) = y^2 C(zt)$
- and $\sigma_{y,z}^2 \equiv y^2\sigma^2/z$.
  - For $\sigma_{y,z}^2(t)$, we do the change of variables $u = zs$ in the integral:
  - $\int_0^\infty C(zs)\, ds = \int_0^\infty C(u)\, du/z = (1/z) \int_0^\infty C(u)\, du$

(See §4.2 of “Planning Queueing Simulations.”)

Approximations for $E[Q(0)]$ and σ^2 in $GI/GI/s/\infty$

- steady-state mean: $E[Q_\rho(0)] \approx \frac{\rho^2(c_a^2+c_s^2)}{2(1-\rho)}$
- asymptotic variance: $\sigma_\rho^2 \approx \frac{\rho^2(c_a^2+c_s^2)^3}{2s(1-\rho)^4}$
- ratio: $\frac{\sigma_\rho^2}{E[Q_\rho(0)]^2} \approx \frac{2(c_a^2+c_s^2)}{s\rho^2(1-\rho)^2}$
- (See equation (42) in “Planning Queueing Simulations,” 1989.)
- How do these depend on the traffic intensity ρ and on the overall variability $(c_a^2 + c_s^2)$?

Required Interval Length

- As a consequence, the required run length based on a specified ϵ **absolute error** is
  - $t_a(\epsilon, \beta) \equiv \frac{4\sigma^2 z_{\beta/2}^2}{\epsilon^2} = \frac{4\rho^2 (c_a^2 + c_s^2)^3 z_{\beta/2}^2}{\epsilon^2 2s(1-\rho)^4}$,
- while the required run length based on a specified ϵ **relative error** is
  - $t_r(\epsilon, \beta) \equiv \frac{4\sigma^2 z_{\beta/2}^2}{\epsilon^2 E[Q(0)]^2} = \frac{4(c_a^2 + c_s^2) z_{\beta/2}^2}{\epsilon^2 2s\rho^2(1-\rho)^2}$.
- How do these required values of t depend on the traffic intensity ρ and on the overall variability $(c_a^2 + c_s^2)$?

References

- W^2. **Planning Queueing Simulations.** Management Science 35(11) (1989) 1341–1366.
- W^2. **Analysis for the Design of Simulation Experiments**. Chapter 13 in *Simulation*, Volume 13 in the Elsevier series of *Handbooks in Operations Research and Management Science*, 2006, edited by Shane Henderson and Barry Nelson, 381–413.
- R. Srikant, W^2. **Simulation Run Lengths to Estimate Blocking Probabilities.** *ACM Transactions on Modeling and Computer Simulation* (TOMACS) 6(1) (1996) 7–52.