Steady-State Simulation

Reading: Ch. 9 in Law & Ch. 15 in Handbook of Simulation

Peter J. Haas

CS 590M: Simulation
Spring Semester 2020

Steady-State Simulation

Why do it?

- Quick approximation for cumulative cost $C(t) = \int_0^t Y(s)\,ds$
  - $Y(s)$ is output process of the simulation, e.g., $Y(t) = f\big(X(t)\big)$ where $X(t)$ is system state and f is a real-valued function
  - If time-average limit $\alpha = \lim_{t\to\infty}(1/t)\int_0^t Y(s)\,ds$ exists, then $C(t) \approx t\alpha$ for large t
- Avoids arbitrary choice of time horizon
- Avoids arbitrary choice of initial conditions

Appropriate if

- No "natural" termination time for simulation
- No "natural" initial conditions
- Rapid convergence to (quasi-)steady state
  (e.g., telecom w. nanosecond timescale observed every 5 min.)

Steady-State Performance Measures

The setup for GSMP $\big(X(t) : t \geq 0\big)$ with state space S

- Output process $Y(t) = f\big(X(t)\big)$
  where f is a real-valued function
- Let μ = initial distribution of GSMP

A reminder: General notion of convergence in distribution

- Discrete-state case:
  - discrete time → $X_n \Rightarrow X$ if $\lim_{n\to\infty} P(X_n = s) = P(X = s)$ for all $s \in S$
  - cont. time → $X(t) \Rightarrow X$ if $\lim_{t\to\infty} P\big(X(t) = s\big) = P(X = s)$ for all $s \in S$
    - Note: $E[f(X)] = \sum_{s\in S} f(s)\pi(s)$, where $\pi(s) = \lim_{t\to\infty} P\big(X(t) = s\big) = P(X = s)$
- Continuous-state case:
  - $Z_n \Rightarrow Z$ if $\lim_{n\to\infty} P(Z_n \leq x) = P(Z \leq x)$ for all x where F_Z is continuous
  - Ex: $\frac{\sqrt{n}}{\sigma}\big(\bar{X}_n - \mu_X\big) \Rightarrow N(0,1)$ by CLT

Steady-State Performance Measures, Continued

Time-Average Limit of $Y(t)$ process

α such that $P_\mu\left\{\lim_{t\to\infty}(1/t)\int_0^t Y(u)\,du = \alpha\right\} = 1$ for any μ

Steady-State Mean of $Y(t)$ process

$\alpha = E[f(X)]$, where, for any μ, $X(t) \Rightarrow X$ and $E[f(X)]$ exists

Limiting Mean of $Y(t)$ process

$\alpha = \lim_{t\to\infty} E\left[f\left(X(t)\right)\right]$ for any μ

- "for any μ" $=$ for any member of GSMP family indexed by μ (with other building blocks the same)
- $E[f(X)]$ exists if and only if $E[|f(X)|] < \infty$
- If f is bounded or S is finite, then $X(t) \Rightarrow X$ implies $\lim_{t\to\infty} E\left[f\left(X(t)\right)\right] = E[f(X)]$ (s-s mean $=$ limiting mean)

Steady-State Simulation Challenges

Autocorrelation problem

- For time-average limit,
  $\alpha = \lim_{t\to\infty} \bar{Y}(t) = \lim_{t\to\infty}(1/t)\int_0^t Y(u)\,du$
- Natural estimator of α is $\bar{Y}(t)$ for some large t
  (obtained from one long observation of system)
- But $Y(t)$ and $Y(t+\Delta t)$ highly correlated if Δt is small
- So estimator is average of autocorrelated observations
- Techniques based on i.i.d. observations don't work

Initial-Transient Problem

- Steady-state distribution unknown,
  so initial dist'n is not typical of steady-state behavior
- Autocorrelation implies that initial bias will persist
- Very hard to detect "end of initial-transient period"

Estimation Methods

Many alternative estimation methods

- Regenerative method
- Batch-means method
- Autoregressive method
- Standardized-time-series methods
- Integrated-path method
- . . .

We will focus on:

- Regenerative method: clean and elegant
- Batch means: simple, widely used and the basis for other methods

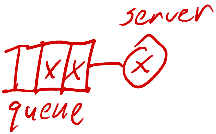
The Regenerative Method

References:

- Shedler [Ch. 2 & 3], Haas [Ch. 5 & 6]
- Recent developments: *ACM TOMACS* 25(4), 2015

Regenerative Processes

- Intuitively: $\big(X(t) : t \geq 0\big)$ is regenerative if process "probabilistically restarts" infinitely often
- Restart times $T(0), T(1), \ldots$ called regeneration times or regeneration points
  - Regeneration points are random
  - Must be almost surely (a.s.) finite
  - Ex: Arrivals to empty GI/G/1 queue



Regenerative Processes: Formal Definition

Definition: Stopping time

A random variable T is a stopping time with respect to $\big(X(t) : t \geq 0\big)$ if occurrence or non-occurrence of event $\{T \leq t\}$ is completely determined by $\big(X(u) : 0 \leq u \leq t\big)$

Definition: Regenerative process

The process $\big(X(t) : t \geq 0\big)$ is regenerative if there exists an infinite sequence of a.s. finite stopping times $\big(T(k) : k \geq 0\big)$ s.t. for $k \geq 1$

1. $\big(X(t) : t \geq T(k)\big)$ is distributed as $\big(X(t) : t \geq T(0)\big)$
2. $\big(X(t) : t \geq T(k)\big)$ is independent of $\big(X(t) : t < T(k)\big)$

- If $T(0) = 0$, process is non-delayed (else delayed)
- Can drop stopping-time requirement, (more complicated def.)
- $\big\{X(t) : t \geq 0\big\}$ regen. $\Rightarrow$ $\big\{f\big(X(t)\big) : t \geq 0\big\}$ regen.
- Analogous definition for discrete-time processes

Regenerative Processes: Examples

Ex 1: Successive times that CTMC hits a fixed state x

- Formally, $T(0) = 0$ and (start in state x)
  $T(k) = \min\{t > T(k-1) : X(t-) \neq x \text{ and } X(t) = x\}$
- Observe that $X\big(T(k)\big) = x$ for all k
- The two regenerative criteria follow from Markov property

Ex 2: Successive times that CTMC leaves a fixed state x

- $X\big(T(k)\big)$ distributed according to $P(x, \cdot)$ for each k
- Second criterion follows from Markov property

Q: Is a semi-Markov process regenerative? Yes

same definitions as examples 1+2 above
because of Markov property at state transition times

Regenerative GSMPs

Ex 3: GSMP with a single state

- $\bar{s} \in S$ is a single state if $E(\bar{s}) = \{\bar{e}\}$ for some $\bar{e} \in E$
- Regeneration points: successive times that $\bar{e}$ occurs in $\bar{s}$
- Observe that for each $k \geq 1$,
  - New state s' at $T(k)$ distributed according to $p(\,\cdot\,; \bar{s}, \bar{e})$
  - No old clocks
  - Clock for new event e' distributed as $F(\,\cdot\,; s', e', \bar{s}, \bar{e})$
- Regenerative property follows from Markov property for $\big((S_n, C_n) : n \geq 0\big)$

(In a semi-Markov process, every state is a single state)

Ex 4: GI/G/1 queue

- $X(t)$ = number of jobs in system at time t
- $\big(X(t) : t \geq 0\big)$ is a GSMP
- $T(k)$ = time of kth arrival to empty system (why?)

because 0 is a single-state

Regenerative GSMPs, Continued

Ex 5: Cancellation

- Suppose there exist $\bar{s}', \bar{s} \in S$ and $\bar{e} \in E(\bar{s})$ with $p(\bar{s}'; \bar{s}, \bar{e}) r(\bar{s}, \bar{e}) > 0$ such that $O(\bar{s}'; \bar{s}, \bar{e}) = \emptyset$
- $T(k) = k$th time that $\bar{e}$ occurs in $\bar{s}$ and new state is $\bar{s}'$

Ex 6: Exponential clocks

new state chosen according to $p(\cdot; \bar{s}, \bar{e})$

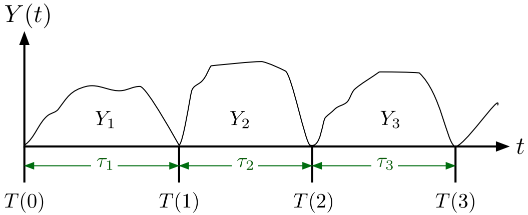
- Suppose that
  - There exists $\tilde{E} \subseteq E$ such that each $e \in \tilde{E}$ is a simple event with $F(x; e) = 1 - e^{-\lambda(e)x}$
  - There exists $\bar{s} \in S$ and $\bar{e} \in E(\bar{s})$ s.t. $E(\bar{s}) - \{\bar{e}\} \subseteq \tilde{E}$
- $T(k) = k$th time that $\bar{e}$ occurs in $\bar{s}$ (memoryless property)

old clock reading for $e \in \tilde{E}$ is $\exp(\lambda(e))$

Other (fancier) regeneration point constructions are possible

- E.g., if clock-setting distn's have heavier-than-exponential tails or bounded hazard rates $h(t) = f(t)/\big(1 - F(t)\big)$

Regenerative Simulation: Cycles



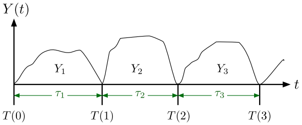
Regeneration points decompose process into i.i.d. cycles

- kth cycle: $\big(X(t) : T(k-1) \le t < T(k)\big)$
- Length of kth cycle: $\tau_k = T(k) - T(k-1)$
- Set $Y_k = \int_{T(k-1)}^{T(k)} Y(u)\, du$
- The pairs $(Y_1, \tau_1), (Y_2, \tau_2), \ldots$ are i.i.d as (Y, τ)

Initial transient is not a problem!

Regenerative Simulation: Time-Average Limits

- Recall: $\bar{Y}(t) = (1/t)\int_0^t Y(u)\,du$



Theorem

Suppose that $E[|Y_1|] < \infty$ and $E[\tau_1] < \infty$. Then $\lim_{t\to\infty} \bar{Y}(t) = \alpha$ a.s., where $\alpha = E[Y]/E[\tau]$.

- So estimating time-average limit reduces to a ratio-estimation problem (can use delta method, jackknife, bootstrap)

(Most of) Proof

$$\bar{Y}\big(T(n)\big) = \frac{1}{T(n)}\int_0^{T(n)} Y(u)\,du = \frac{\sum_{j=1}^{n}\int_{T(j-1)}^{T(j)} Y(u)\,du}{\sum_{j=1}^{n}\big(T(j) - T(j-1)\big)} = \frac{\sum_{j=1}^{n} Y_j}{\sum_{j=1}^{n} \tau_j}$$

$$\Rightarrow \lim_{n\to\infty} \bar{Y}\big(T(n)\big) = \alpha \text{ a.s. by SLLN}$$

Regenerative Simulation: Steady-State Means

Definition

A real-valued random variable τ is said to be periodic with period d if d is the largest real number such that, w.p.1, τ assumes values in the set $\{0, d, 2d, 3d, \dots\}$. If no such number exists, then τ is aperiodic. (A discrete random variable is aperiodic if $d = 1$.)

Theorem

Suppose that $\big(X(t) : t \geq 0\big)$ is regenerative with finite state space S and τ is aperiodic with $E[\tau] < \infty$. Then $X(t) \Rightarrow X$ and $E[f(X)] = E[Y_1(f)]/E[\tau_1]$ for any real-valued function f on S, where $Y_1(f) = \int_{T(0)}^{T(1)} f\big(X(u)\big)\,du$ and $\tau_1 = T(1) - T(0)$

- Under conditions of theorem, time avg limit is also a steady-state mean (and a limiting mean)

time avg. limit: $\lim_{t\to\infty} E[f(X(t))]$

Regenerative Simulation: Other Performance Measures

Important observation:
Y_k and τ_k can be any quantities determined by a cycle

- Ex 1: Long-run avg rate at which GSMP jumps from s to s'
  - $Y_k =$ number of jumps from s to s' in kth cycle
  - $\tau_k =$ length of kth cycle
- Ex 2: Long-run fraction of jumps from s to s'
  - $Y_k =$ number of jumps from s to s' in kth cycle
  - $\tau_k =$ total number of jumps in kth cycle
- Ex 3: Long-run frac. occurrences of e where new state $\in A$
  - $Y_k =$ number of occurrences of e in kth cycle where $s' \in A$
  - $\tau_k =$ total number of occurrences of e in kth cycle
  - E.g., frac. ambulance arrivals that find emergency room full

Validity of Regenerative Method

Usually not hard to show probabilistic restart

But must also show:

- Regeneration points are a.s. finite
  (i.e., infinitely many regenerations w.p.1)
- $E[\tau] < \infty$
- $\sigma^2 < \infty$ for confidence intervals
  - If S is finite, suffices to show that $E[\tau^2] < \infty$
- Nontrivial!

$r = \frac{E[Y]}{E[\tau]}$

regen. variance estimator:

$$\frac{\hat{Var}[Y] - 2r\,\hat{Cov}[Y, \tau] + r^2\,\hat{Var}[\tau]}{\bar{\tau}^2}$$

See my book for techniques to prove validity

Regenerative Method: Delays

Formal definition of delays in a GSMP

- Sequences of starts $(U_n : n \geq 0)$ and terminations $(V_n : n \geq 0)$
- Assume $U_0 \leq U_1 \leq \cdots$ (delays enumerated in start order)
- nth delay is then $D_n = V_n - U_n$

Regular delay sequence

- $\big(D_n : n \geq 0\big)$ is regular with respect to $\big(X(t) : t \geq 0\big)$ if
  - Occurrence or non-occurrence of event $\{U_{N(t)+1} - t \leq x\}$ determined by $\big(X(t) : t \leq u \leq t + x\big)$
  - Occurrence or non-occurrence of event $\{V_n \leq U_n + v\}$ determined by $\big(X(t) : U_n \leq t \leq U_n + v\big)$

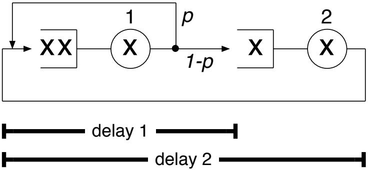
  where $N(t) =$ number of starts in $[0, t]$
- Intuition: A regular delay sequence is "determined" by $\big(X(t) : t \geq 0\big)$ in a reasonable way

Delays, Continued

Example of regular delays in a GSMP [Shedler, Sec. 5.5]

- Assume at most one ongoing delay at any time point
- $U_n =$ time of nth jump from a state $s \in A_1$ to a state $s' \in A_2$
- $V_n =$ time of nth jump from a state $s \in B_1$ to a state $s' \in B_2$

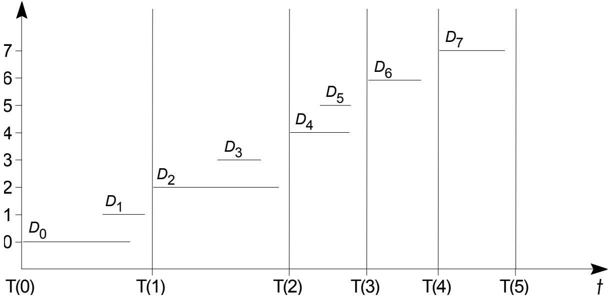
Delays, Continued



Case 1: Delays bounded by regenenerative cycles of GSMP

- GSMP $(X(t) : t \geq 0)$: $X(t) = \#$ of jobs at center 1 at time t
- $T(k) = k$th time GSMP jumps out of a single state $\bar{s} = 0$
- For delay 1: at each $T(k)$ a single delay starts
  (no other delays in progress)
- Thus every delay starts and ends in the same regen. cycle
  $U_n \in [T(k-1), T(k)] \Rightarrow V_n \in [T(k-1), T(k)]$
- Sequence of delays is decomposed into i.i.d. cycles

Delays, Continued



Regeneration points for $(D_n : n \geq 0)$:
$N_0 = 0, N_1 = 2, N_2 = 4, N_3 = 6,$ and $N_4 = 7$

Delays, Continued

Ex: Estimate α = long-run fraction of delays ≥ 2 time units

- Set $f(x) = 1$ if $x \geq 2$ and $f(x) = 0$ otherwise
- By discrete-time version of prior results,

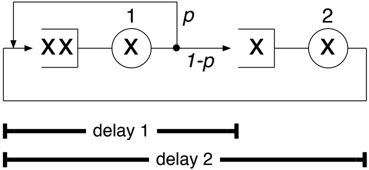
$$\alpha \stackrel{\Delta}{=} \lim_{n\to\infty} (1/n) \sum_{j=0}^{n-1} f(D_j) = E[Y_1]/E[\tau_1] \text{ a.s.}$$

where $Y_k = \sum_{n=N_{k-1}}^{N_k - 1} f(D_n)$ and $\tau_k = N_k - N_{k-1}$ = # of delays in k^{th} regen. cycle

- In example, $Y_1 = f(D_0) + f(D_1)$ and $\tau_1 = 2$
- The (Y_k, τ_k) pairs are i.i.d., so use ratio estimation methods
- If τ_1 has period 1 (i.e., aperiodic in discrete time), then
  - $D_n \Rightarrow D$
  - $\alpha = E[f(D)]$ = steady state probability that a delay is ≥ 2

$= P(D \geq 2)$

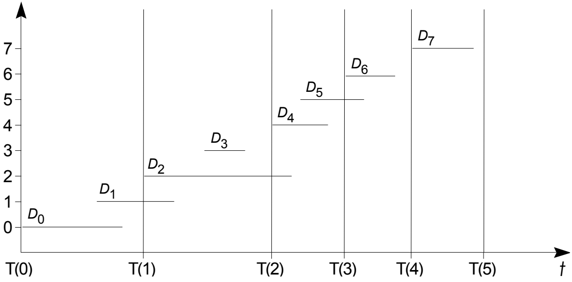
Delays, Continued



Case 2: Delays span regenerative cycles

- Same GSMP as before: $X(t) = \#$ of jobs at center 1 (N jobs total)
- Same regeneration points $T(k)$ as before: Jumps out of $\bar{s} = 0$
- For delay 2: At each $T(k)$, one delay starts but $N - 1$ delays are in progress
- Thus delays span regenerative cycles

Delays, Continued



Case 2, continued

- Take subset of regeneration points so that delay spans at most two cycles
- (Y_k, τ_k) pairs are now one-dependent
- Variant of regenerative method works [see my book]
  - Same point estimate
  - CLT variance accounts for dependence between adjacent cycles

Batch Means

A method for estimating time-average limits when we can't find regeneration points

To estimate $\alpha = \lim_{t\to\infty} \frac{1}{t}\int_0^t Y(u)\,du$:

Basic Batch Means

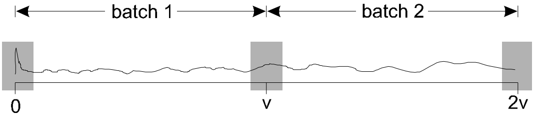
1. Choose small integer m and large number v.
2. Set $t_{m-1,\delta} = 1-(\delta/2)$ Student-t quantile, $m-1$ d.o.f.
3. Simulate $\big(Y(t) : t \geq 0\big)$ up to time $t = mv$
4. Compute batch mean $\bar{Y}_j = \frac{1}{v}\int_{(j-1)v}^{jv} Y(u)\,du$ for $1 \leq j \leq m$
5. Compute point estimator $\alpha_m = (1/m)\sum_{j=1}^{m} \bar{Y}_j$
6. Compute $s_m^2 = \frac{1}{m-1}\sum_{j=1}^{m}(\bar{Y}_j - \alpha_m)^2$
7. Compute $100(1-\delta)\%$ confidence interval
   $\left[\alpha_m - \frac{t_{m-1,\delta}s_m}{\sqrt{m}}, \alpha_m + \frac{t_{m-1,\delta}s_m}{\sqrt{m}}\right]$

Batch Means, Continued

Batch means

Why Does Batch Means Work?

- Intuition: Batches look like i.i.d. normal random variables
- See my book for conditions on GSMP ensuring validity



Many variants and generalizations (Handbook of Simulation)

- Overlapping batch means
- Sequential batch means
- Standardized time series
- ...

Comparison to regenerative method

- When both are applicable, regenerative yields shorter CIs when run length is long

Jackknifed Batch Means

Ex: Nonlinear functions of time-average limits

- Estimate $\alpha = g(\mu_1, \mu_2)$, where $\mu_i = \lim_{t\to\infty}(1/t)\int_0^t f_i\big(X(u)\big)\,du$
- For $i = 1, 2$ set
  - $\bar{Y}_j^{(i)} = \frac{1}{v}\int_{(j-1)v}^{jv} f_i\big(X(u)\big)\,du$ for $j = 1, \ldots, m$
  - $\bar{\bar{Y}}^{(i)} = \mathrm{avg}(\bar{Y}_1^{(i)}, \ldots, \bar{Y}_m^{(i)})$
  - $\bar{\bar{Y}}_{-k}^{(i)} = \mathrm{avg}(\bar{Y}_1^{(i)}, \ldots, \bar{Y}_{k-1}^{(i)}, \bar{Y}_{k+1}^{(i)}, \ldots, \bar{Y}_m^{(i)})$

Jackknifed Batch Means (JBM)

1. Simulate $\big(X(t) : t \geq 0\big)$ up to time $t = mv$
2. Compute batch means $\bar{Y}_1^{(i)}, \ldots, \bar{Y}_m^{(i)}$ for $i = 1, 2$
3. For $1 \leq k \leq m$, compute pseudovalue

   $$\alpha_m(k) = mg(\bar{\bar{Y}}^{(1)}, \bar{\bar{Y}}^{(2)}) - (m-1)g(\bar{\bar{Y}}_{-k}^{(1)}, \bar{\bar{Y}}_{-k}^{(2)})$$

4. Compute point estimator $\alpha_m^J = (1/m)\sum_{k=1}^m \alpha_m(k)$
5. Compute $100(1-\delta)$ CI

   $$\Big[\alpha_m^J - t_{m-1,\delta}(v_m/m)^{1/2}, \alpha_m^J + t_{m-1,\delta}(v_m/m)^{1/2}\Big]$$

   where v_m = sample variance of $\alpha_m(1), \ldots, \alpha_m(m)$

Jackknifed Batch Means, Continued

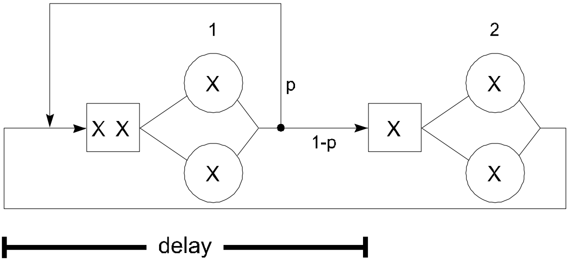
Can apply JBM to obtain low-bias estimator for ordinary time-average limits in a GSMP

- Goal: Estimate $\alpha = \lim_{t\to\infty}(1/t)\int_0^t f\big(X(u)\big)\,du$
- Can show that $\alpha = g(\mu_1, \mu_2)$, where $g(x,y) = x/y$ and
  - $\mu_1 = \lim_{n\to\infty}(1/n)\sum_{i=0}^{n-1} f(S_n)t^*(S_n, C_n)$
  - $\mu_2 = \lim_{n\to\infty}(1/n)\sum_{i=0}^{n-1} t^*(S_n, C_n)$

  where $\big((S_n, C_n) : n \geq 0\big)$ is underlying GSSMC of GSMP and t^* is holding time function
- Partial proof: $\frac{1}{\zeta_n}\int_0^{\zeta_n} f\big(X(u)\big)\,du = \frac{\sum_{i=0}^{n-1} f(S_n)t^*(S_n,C_n)}{\sum_{i=0}^{n-1} t^*(S_n,C_n)} \to \frac{\mu_1}{\mu_2}$
- So can apply discrete-time version of JBM with batches:
  - $\bar{Y}_j^{(1)} = (1/v)\sum_{i=(j-1)v}^{jv-1} f(S_i)t^*(S_i, C_i)$
  - $\bar{Y}_j^{(2)} = (1/v)\sum_{i=(j-1)v}^{jv-1} t^*(S_i, C_i)$

  for $j = 1, \ldots, m$

Jackknifed Batch Means, Continued



Can apply discrete-time version of JBM to analyze delays

- To estimate $\alpha = \lim_{n\to\infty}(1/n)\sum_{j=0}^{n-1} f(D_j)$
- Simulate $D_1, D_2, \ldots, D_{vm}$ (here v is an integer)
- Batch means: $\bar{Y}_j = (1/v)\sum_{i=(j-1)v}^{jv-1} f(D_i)$ for $j = 1, \ldots, m$
- Ex: Cyclic queues with multiple servers per station

m batches
v delays per batch