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

Overview

The Regenerative Method
  Regenerative processes
  Regenerative Simulation
  Delays

The Batch Means Method
  Time-average limits
  Jackknifed Batch Means
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., \( Y(t) = f(X(t)) \)
    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 \( (X(t) : t \geq 0) \) with state space \( S \)

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

A reminder: General notion of convergence in distribution

- Discrete-state case:
  \( X_n \Rightarrow X \) if \( \lim_{n \to \infty} P(X_n = s) = P(X = s) \) for all \( s \in S \)
  \( X(t) \Rightarrow X \) if \( \lim_{t \to \infty} P(X(t) = s) = P(X = s) \) for all \( s \in S \)
  - Note: \( E[f(X)] = \sum_{s \in S} f(s)\pi(s) \), where \( \pi(s) = \lim_{t \to \infty} P(X(t) = s) = P(X = s) \)

- 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: \( \frac{\sqrt{n}}{\sigma} (\bar{X}_n - \mu_X) \Rightarrow N(0, 1) \) by CLT
Steady-State Performance Measures, Continued

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

\[ \alpha \text{ such that } P_\mu \left\{ \lim_{t \to \infty} \frac{1}{t} \int_0^t Y(u) \, du = \alpha \right\} = 1 \text{ for any } \mu \]

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

\[ \alpha = E[f(X)], \text{ where, for any } \mu, \ X(t) \Rightarrow X \text{ and } E[f(X)] \text{ exists} \]

Limiting Mean of $Y(t)$ process

\[ \alpha = \lim_{t \to \infty} E[f(X(t))] \text{ for any } \mu \]

- “for any $\mu$” = for any member of GSMP family indexed by $\mu$ (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 $\lim_{t \to \infty} E[f(X(t))] = E[f(X)]$ (s-s mean = limiting mean)
Steady-State Simulation Challenges

Autocorrelation problem

- For time-average limit,
  \[ \alpha = \lim_{t \to \infty} \bar{Y}(t) = \lim_{t \to \infty} \left( \frac{1}{t} \int_0^t Y(u) \, du \right) \]
- Natural estimator of \( \alpha \) 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 \( \Delta t \) 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
Steady-State Simulation

Overview

The Regenerative Method
  Regenerative processes
  Regenerative Simulation
  Delays

The Batch Means Method
  Time-average limits
  Jackknifed Batch Means
The Regenerative Method

References:

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

Regenerative Processes

- Intuitively: \( (X(t) : t \geq 0) \) 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 $(X(t) : t \geq 0)$ if occurrence or non-occurrence of event $\{T \leq t\}$ is completely determined by $(X(u) : 0 \leq u \leq t)$.

Definition: Regenerative process

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

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

- If $T(0) = 0$, process is non-delayed (else delayed)
- Can drop stopping-time requirement, (more complicated def.)
- $\{X(t) : t \geq 0\}$ regen. $\Rightarrow \{f(X(t)) : t \geq 0\}$ 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 $T(k) = \min\{t > T(k - 1) : X(t-) \neq x \text{ and } X(t) = x\}$
- Observe that $X(T(k)) = x$ for all $k$
- The two regenerative criteria follow from Markov property

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

- $X(T(k))$ 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 and 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 $((S_n, C_n) : n \geq 0)$

Ex 4: GI/G/1 queue

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

(In a semi-Markov process, every state 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

- 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)

Other (fancier) regeneration point constructions are possible

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

Overview

The Regenerative Method

  Regenerative processes
  Regenerative Simulation
  Delays

The Batch Means Method

  Time-average limits
  Jackknifed Batch Means
Regenerative Simulation: Cycles

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

- *kth cycle:* \( (X(t) : T(k - 1) \leq t < T(k)) \)
- 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 \( (Y, \tau) \)

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

\[
\lim_{t \to \infty} \bar{Y}(t) = \alpha \text{ a.s., where } \alpha = \frac{E[Y]}{E[\tau]}.
\]

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

**(Most of) Proof**

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

\[
\Rightarrow \lim_{n \to \infty} \bar{Y}(T(n)) = \alpha \text{ a.s. by SLLN}
\]
Regenerative Simulation: Steady-State Means

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

Theorem
Suppose that $(X(t) : t \geq 0)$ is regenerative with finite state space $S$ and $\tau$ 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 $Y_1(f) = \int_{T(0)}^{T(1)} f(X(u)) \, du$ and $\tau_1 = T(1) - T(0)$.

- Under conditions of theorem, time avg limit is also a steady-state mean (and a limiting mean):
\[
\text{time avg. limit: } \lim_{b \to \infty} E[P(X(b))] \]
Regenerative Simulation: Other Performance Measures

Important observation:
$Y_k$ and $\tau_k$ 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 $k$th cycle
  - $\tau_k =$ length of $k$th cycle

- **Ex 2:** Long-run fraction of jumps from $s$ to $s'$
  - $Y_k =$ number of jumps from $s$ to $s'$ in $k$th cycle
  - $\tau_k =$ total number of jumps in $k$th cycle

- **Ex 3:** Long-run frac. occurrences of $e$ where new state $\in A$
  - $Y_k =$ number of occurrences of $e$ in $k$th cycle where $s' \in A$
  - $\tau_k =$ total number of occurrences of $e$ in $k$th 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!

See my book for techniques to prove validity
Steady-State Simulation

Overview

The Regenerative Method
  Regenerative processes
  Regenerative Simulation
  Delays

The Batch Means Method
  Time-average limits
  Jackknifed Batch Means
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)
- \(n\)th delay is then \(D_n = V_n - U_n\)

Regular delay sequence

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

where \(N(t) = \text{number of starts in } [0, t]\)

- Intuition: A regular delay sequence is “determined” by \((X(t) : t \geq 0)\) in a reasonable way
Example of regular delays in a GSMP [Shedler, Sec. 5.5]

- Assume at most one ongoing delay at any time point
- \( U_n = \text{time of } n\text{th jump from a state } s \in A_1 \text{ to a state } s' \in A_2 \)
- \( V_n = \text{time of } n\text{th jump from a state } s \in B_1 \text{ to a state } s' \in B_2 \)
Case 1: Delays bounded by regenerative 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
Regeneration points for \((D_n : n \geq 0)\):
\[N_0 = 0, \; N_1 = 2, \; N_2 = 4, \; N_3 = 6, \; \text{and} \; N_4 = 7\]
Delays, Continued

Ex: Estimate $\alpha = \text{long-run fraction of delays} \geq 2$ time units

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

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

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

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

$\Rightarrow P(D \geq 2)$
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
Case 2, continued

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

Overview

The Regenerative Method
  Regenerative processes
  Regenerative Simulation
  Delays

The Batch Means Method
  Time-average limits
  Jackknifed Batch Means
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 \( \nu \).
2. Set \( t_{m-1, \delta} = 1 - (\delta/2) \) Student-t quantile, \( m - 1 \) d.o.f.
3. Simulate \( (Y(t) : t \geq 0) \) up to time \( t = mv \)
4. Compute batch mean \( \bar{Y}_j = \frac{1}{\nu} \int_{(j-1)\nu}^{j\nu} Y(u) \, du \) 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**

**Why Does Batch Means Work?**
- **Intuition:** Batches look like i.i.d. normal random variables
- **See my book for conditions on GSMP ensuring validity**

![Diagram of batch 1 and batch 2](image)

**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
Steady-State Simulation

Overview
The Regenerative Method
  Regenerative processes
  Regenerative Simulation
  Delays
The Batch Means Method
  Time-average limits
  Jackknifed Batch Means
Jackknifed Batch Means

Ex: Nonlinear functions of time-average limits

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

Jackknifed Batch Means (JBM)

1. Simulate $(X(t) : t \geq 0)$ up to time $t = mv$
2. Compute batch means $\tilde{Y}_1^{(i)}, \ldots, \tilde{Y}_m^{(i)}$ for $i = 1, 2$
3. For $1 \leq k \leq m$, compute pseudo-value
   \[
   \alpha_m(k) = mg(\tilde{Y}^{(1)}, \tilde{Y}^{(2)}) - (m - 1)g(\tilde{Y}_{-1}^{(1)}, \tilde{Y}_{-1}^{(2)})
   \]
4. Compute point estimator $\alpha^J_m = (1/m) \sum_{k=1}^{m} \alpha_m(k)$
5. Compute $100(1 - \delta)$ CI
   \[
   \left[\alpha^J_m - t_{m-1, \delta} (\nu_m/m)^{1/2}, \alpha^J_m + t_{m-1, \delta} (\nu_m/m)^{1/2}\right]
   \]
   where $\nu_m = \text{sample variance of } \alpha_m(1), \ldots, \alpha_m(m)$
Can apply JBM to obtain low-bias estimator for ordinary time-average limits in a GSMP

- **Goal:** Estimate \( \alpha = \lim_{t \to \infty} \left( \frac{1}{t} \int_0^t f(X(u)) \, du \right) \)

- Can show that \( \alpha = g(\mu_1, \mu_2) \), where \( g(x, y) = \frac{x}{y} \) and
  - \( \mu_1 = \lim_{n \to \infty} \left( \frac{1}{n} \sum_{i=0}^{n-1} f(S_n) t^*(S_n, C_n) \right) \)
  - \( \mu_2 = \lim_{n \to \infty} \left( \frac{1}{n} \sum_{i=0}^{n-1} t^*(S_n, C_n) \right) \)

where \( (S_n, C_n) : n \geq 0 \) is underlying GSSMC of GSMP and \( t^* \) is holding time function.

- **Partial proof:** \( \frac{1}{n} \int_0^{\zeta_n} f(X(u)) \, du = \frac{\sum_{i=0}^{n-1} f(S_n) t^*(S_n, C_n)}{\sum_{i=0}^{n-1} t^*(S_n, C_n)} \to \frac{\mu_1}{\mu_2} \)

- So can apply discrete-time version of JBM with batches:
  - \( \tilde{Y}^{(1)}_j = \frac{1}{v} \sum_{i=(j-1)v}^{jv-1} f(S_i) t^*(S_i, C_i) \)
  - \( \tilde{Y}^{(2)}_j = \frac{1}{v} \sum_{i=(j-1)v}^{jv-1} t^*(S_i, C_i) \)

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