On the Initial Transient Problem for Steady-State Simulation

Peter W. Glynn

Stanford University

Lunteren Conference

January 17th, 2007

Outline of Talk

1. The Steady-State Simulation Problem
   - What is it?
   - Why is it relevant?
   - Challenges
2. The Initial Transient Problem
   - Is it a serious problem?
   - When is it a serious problem?
3. The Basic Approaches to the Initial Transient Problem
   - Simulating a stationary version
   - Identifying the initial transient period
   - Low Bias estimators

Steady-State Simulation

$Y(t)$ = rate at which "reward" increases at time t

Assume there exists a (deterministic) constant α such that

$$\frac{1}{t}\int_0^t Y(s)ds \Rightarrow \alpha$$

as $t \longrightarrow \infty$.

Goal: Compute α

Examples

$Y(t) =$ **number-in-system (in queue) at time** t

$Y(t) =$ **work-in-process (in manufacturing system) at time** t

$$Y(t) = I(Q(t) \geq b)$$

Why the emphasis on steady-state simulation?

To compute $\alpha(t,x) = \mathrm{E}_x f(X(t))$**:**

$$\alpha'(t) = A\alpha(t)$$

$$\alpha(0) = f$$

To compute $\alpha = \sum_x \pi(x) f(x)$**:**

$$0 = \pi A$$

s/t $\sum_x \pi(x) = 1$

$$\alpha = \pi f$$

Analytical and computational tractability

Modeling Issues

(favoring use of a steady-state formulation)

- no need to specify time horizon
- no need to specify initial distribution
- appropriate in many applications

Steady-State Simulation: Challenge 1

- **The steady state mean α involves the "infinite time behavior" of $(Y(s) : s \geq 0)$**

- **How do we compute α based on a finite-horizon simulation?**

- **Mathematically:** $\mathrm{E}\,\frac{1}{t}\int_0^t Y(s)ds \neq \alpha$

- **Conceptually:** **Initial condition is atypical of steady-state**

$$\Downarrow$$

"initial transient"

$$\Downarrow$$

"initial bias"

Steady-State Simulation: Challenge 2

In view of Challenge 1, choose time horizon t large

$$\Downarrow$$

One observation of process Y over $[0, t]$

$$\Downarrow$$

How to compute an estimator for the variance of

$$\frac{1}{t}\int_0^t Y(s)ds\,?$$

A Connection between the Two Challenges

$$E_x f(X(t)) = \alpha + a(x)e^{-\eta t}(1 + o(1))$$

and

$$\begin{aligned} & cov(f(X^*(0)), f(X^*(t))) \\ & = \; E_\pi f_c(X(0)) f_c(X(t)) \\ & = \; E_\pi f_c(X(0)) \cdot E[f_c(X(t)) | X(0)] \\ & = \; E_\pi f_c(X(0)) \cdot a(X(0)) e^{-\eta t}(1 + o(1)) \end{aligned}$$

- **used intensively**
- **doesn't work in presence of non–Markov processes**

Today: We focus on Challenge 1

"The Initial Transient Problem"

How Serious a problem is the Initial Transient?

For geometrically ergodic (Markov) processes,

$Y(t) \Rightarrow Y(\infty)$ **as** $t \to \infty$ **and**

$$\mathrm{E}\, Y(t) = \mathrm{E}\, Y(\infty) + O(e^{-\eta t})$$

so

$$\begin{aligned}
&\mathrm{E}\, \frac{1}{t} \int_0^t Y(s) ds - \alpha \\
&\quad = \quad \frac{1}{t} \int_0^\infty \{\mathrm{E}\, Y(s) - \mathrm{E}\, Y(\infty)\} ds - \frac{1}{t} \int_t^\infty \{\mathrm{E}\, Y(s) - \mathrm{E}\, Y(\infty)\} ds \\
&\quad = \quad b/t + O(e^{-\eta t})
\end{aligned}$$

Also, $\quad t^{1/2} \left(\frac{1}{t} \int_0^t Y(s) ds - \alpha \right) \Rightarrow \sigma N(0, 1)$

$$\mathrm{var} \left(\frac{1}{t} \int_0^t Y(s) ds \right) \sim \sigma^2 / t$$

Conclusion

- **Initial bias effect is small relative to sampling variability**

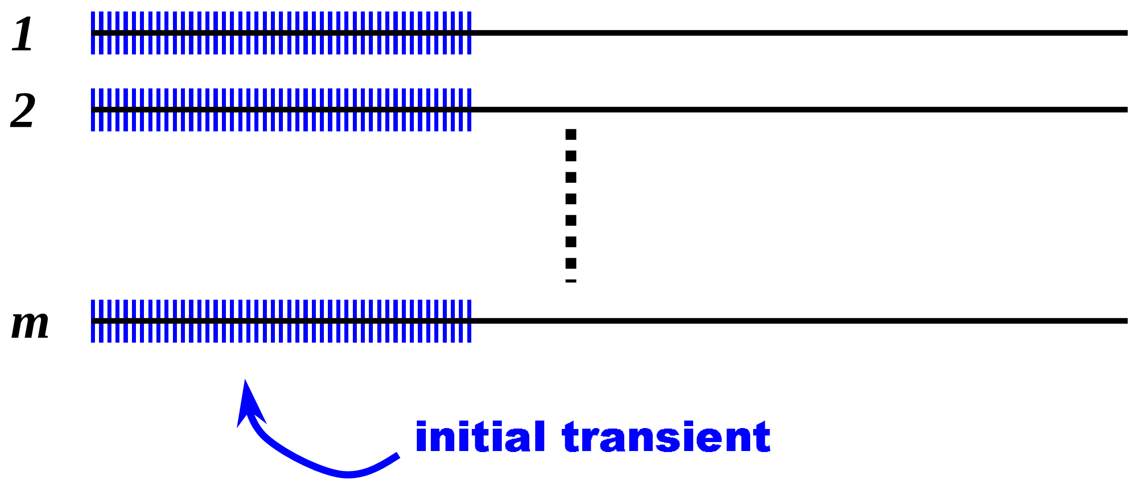
- **For most OR-related steady-state simulations, initial transient is not serious and can be ignored**

Why: Any time-horizon large enough to make the sampling variability small effectively eliminates the effect of the initial transient.

There are exceptions

- parallel simulation
- long-range dependent processes
- multi-modal behavior / nearly decomposable systems
- high dimensional systems

Parallel Simulation



- **MSE analysis:**

$$\textbf{variance}\left(\frac{1}{m}\sum_{i=1}^{m}\frac{1}{t}\int_0^t Y_i(s)ds\right) \sim \frac{\sigma^2}{mt}$$

$$\textbf{bias}\left(\frac{1}{m}\sum_{i=1}^{m}\frac{1}{t}\int_0^t Y_i(s)ds\right) \sim \frac{b}{t}$$

- **If $m \gg t$, initial transient becomes dominant effect.**

GH 90's

Long-range dependence

- In the presence of long-range dependence,

$$\mathrm{var}\left(\frac{1}{t}\int_0^t Y(s)ds\right) \sim \frac{\sigma^2}{t^{2-2H}} \qquad (1/2 < H < 1)$$

- How to allocate c units of computer time in single processor setting?

m independent replications of length c/m

$$\mathrm{var}\left(\frac{1}{m}\sum_{i=1}^{m}\frac{1}{c/m}\int_0^{c/m} Y_i(s)ds\right) \approx \frac{\sigma^2 m^{1-2H}}{c^{2-2H}}$$

- Choose m large ...

- Same initial bias problem as in parallel simulation setting

Nearly Decomposable Systems

$$P = \begin{pmatrix} P_{11} & & & \\ & P_{22} & \varepsilon & \\ & & \ddots & \\ \varepsilon & & \ddots & \\ & & & P_{dd} \end{pmatrix}$$

- Particularly challenging context . . .
- Any statistically based "detection rule" can be fooled
- Arises in Markov chain Monte Carlo context
- Random re-start

High-Dimensional Systems

$$\overrightarrow{Y}(t) = (Y_1(t), \ldots, Y_d(t))$$

independent

$$\|P(Y_i(t) \in \cdot) - P(Y_i(\infty) \in \cdot)\| y \sim ce^{-\lambda t}$$

$$\|P(\overrightarrow{Y}(t) \in \cdot) - P(\overrightarrow{Y}(\infty) \in \cdot)\| \sim dce^{-\lambda t} \qquad \text{G05}$$

Slows down convergence rate...

There are models for which initial transient is significant

3 Approaches to Dealing with the Initial Transient

- **Approach 1** Simulate a stationary version of Y
  exact simulation / perfect simulation

- **Approach 2** Identify the initial transient interval $[0, t]$

- **Approach 3** Modify the estimator so as to reduce the "initial bias" effect

Approach 1: Simulate a stationary version of Y

$$Y(t) = f(X(t))$$

- **Choose** $X(0) \sim \pi$
- **Don't know** π**!**
- **Can we sample from** π **based on the ability to do a dynamic simulation of** X**?**
- **Exact/ perfect simulation**

$$\begin{aligned}\pi(\cdot) &= \frac{\mathrm{E}\int_0^\tau I(X(s)\in\cdot)ds}{\mathrm{E}\,\tau}\\ &= \int_0^\infty \mathrm{P}\left(X(s)\in\cdot \mid \tau > s\right)\frac{\mathrm{P}(\tau>s)ds}{\mathrm{E}\,\tau}\\ &= \mathrm{P}(X(Z)\in\cdot)\end{aligned}$$

AGT 92

$$Z \sim \frac{\mathrm{P}(\tau>\cdot)}{\mathrm{E}\,\tau}$$

- **Are there any interesting problem classes for which**

$$\mathrm{P}(\tau > \cdot)/\,\mathrm{E}\,\tau$$

is known, but simulating from $\pi(\cdot)$ directly is hard?

- **Yes . . .**

$$P^m(x,y) \geq \varepsilon\phi(y)$$

$$P^m(x,y) = \varepsilon\phi(y) + (1-\varepsilon)q(x,y)$$

$$P(\tau = mk) = \varepsilon(1-\varepsilon)^{k-1}$$

$\frac{\tau}{m}$ **geometric**

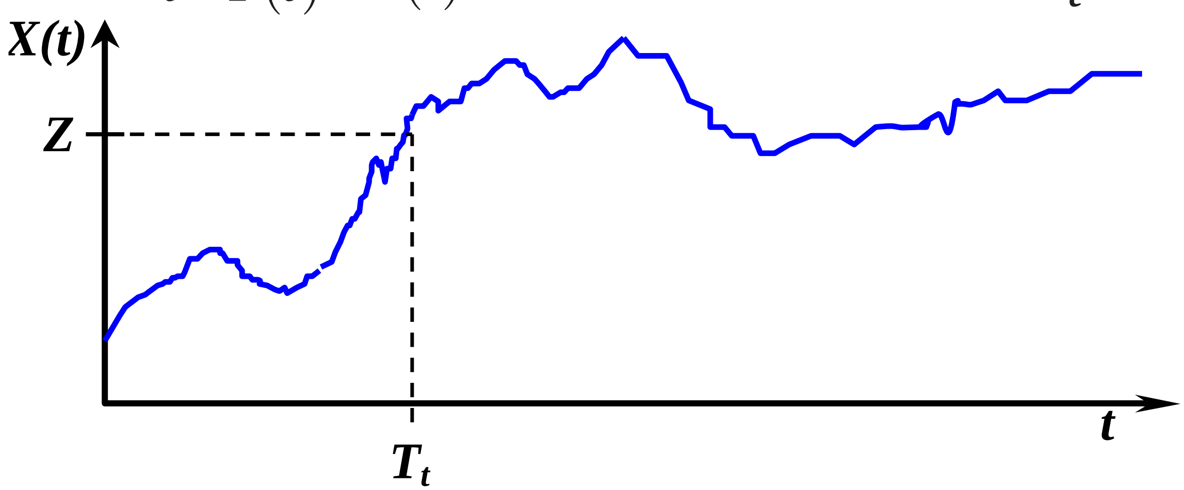
AGT (1992)
Propp/Wilson (1996)
⋮

A Statistically-based alternative

- Simulate Markov process X over $[0, t]$
- Compute $\qquad \widehat{\pi}_t(x) = \frac{1}{t}\int_0^t I(X(s) = x)ds$
- Generate $Z_t \sim \widehat{\pi}_t$ ▸ **Put** $T_t = \inf\{s \geq 0 : X(s) = Z_t\}$
- Then,

$$\|(X(s) : T_t \leq s \leq T_t + t) - (X^*(s) : 0 \leq s \leq t)\| \to 0 \quad \text{G93}$$

- **E** $\left(\frac{1}{t - T(t)}\int_{T(t)}^t f(X(s))ds\right) = \alpha + O(\frac{1}{t^2}) \quad$ (A + G 06)



Approach 2: Identify initial transient interval $[0, t_0]$

- Use analytic bounds to compute t_0 for which

$$||\mathrm{P}(Y(t_0) \in \cdot) - \mathrm{P}(Y(\infty) \in \cdot)|| < \varepsilon$$

For many geometrically ergodic processes,

$$\mathrm{E}_x f(X(t)) = \alpha + a(x)e^{-\gamma t + i\theta t}(1 + o(1))$$

where $\gamma > 0$. Here, $z = -\gamma + i\theta$ is second largest (in modulus) eigenvalue of A.

  - For highly structured models, one can sometimes compute asymptotics for z (Diaconis)
  - For reversible systems, one can compute analytic bounds on z
  - For regenerative systems, one can obtain bounds on

$$||\mathrm{P}(Y(t) \in \cdot) - \mathrm{P}(Y(\infty) \in \cdot)||$$

- **Use an approximating analytic model to identify t_0:**

$$Y(t) \overset{\mathscr{D}}{\approx} \widetilde{Y}(t)$$

Single-server FIFO queue in "heavy-traffic":

$$Y(t) \overset{\mathscr{D}}{\approx} \mathrm{RBM}(t)$$

$$t_0 = c(1-\rho)^{-2}$$

For infinite-server queue in "heavy-traffic":

$$Y(t) \overset{\mathscr{D}}{\approx} \text{Gaussian process}$$

$$t_0 \approx F^{-1}(1-\varepsilon)$$

Problems

- good approximations hard to compute for most models (networks)

  Possible remedy:

  - simulate simplified model first; determine t_0 based on simulation
  - then simulate real model

- note that time to stationarity t_0 depends on functional being computed

$$\begin{aligned} & t_0 \text{ for ``number-in-system''} \\ & \neq t_0 \text{ for ``buffer loss''} \end{aligned}$$

$$\begin{aligned} & \gamma \text{ for ``number-in-system''} \\ & \neq \widetilde{\gamma} \text{ for exponential moments of ``number-in-system''} \end{aligned}$$

Statistically Based Methods

- **Many proposals have been made**

  e.g. discard initial observations until the first one left is neither the maximum nor the minimum of the remaining observations (Conway 1963)

- **Problematic both in theory and in practice**

A Basic Difficulty

- **In general, one has only one realization of the initial transient on which to base one's statistical analysis.**

- **Any rigorous rule needs to "model" the initial transient**
  e.g., autoregressive process

$$X_{n+1} = \rho X_n + \varepsilon_{n+1}$$

Regenerative Processes

$$a(t) = \mathrm{E}\, Y(t)$$

$$a = b + F * a \qquad \textbf{(renewal eqn)}$$

Solution is

$$a = U * b$$

If b, F are known, choose t_0 such that $|a(t_0) - a(\infty)| < \varepsilon$

If b, F are unknown, estimate these from cycles simulated:

$$\widehat{F}_n(t) = \tfrac{1}{n} \sum_{i=1}^{n} I(\tau_i \le t)$$

$$\widehat{b}_n(t) = \tfrac{1}{n} \sum_{i=1}^{n} b_i(t)$$

Choose $\widehat{t}$ so that

$$\left|(\widehat{U} + \widehat{b})(\widehat{t}) - (\widehat{U} + \widehat{b})(\infty)\right| < \varepsilon \qquad \text{BG06, KG06}$$

Approach 3: Low Bias Estimators

Attempt 1

$$\mathrm{E}\,\frac{1}{t}\int_0^t Y(s)ds = \alpha + \frac{b}{t} + O(e^{-\lambda t})$$

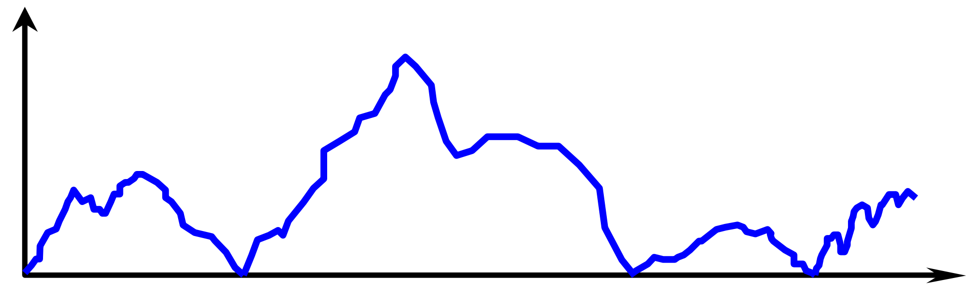
Estimate α, b via a linear regression

Use $\widehat{\alpha}$ from linear regression

Doesn't work ...

As for Approach 2, must "model" initial transient

Regenerative Processes



$$Y = (Y(s) : 0 \leq s \leq T_n)$$

$$\left(\int_{T_{i-1}}^{T_i} Y(s)ds, T_i - T_{i-1} : 1 \leq i \leq n\right)$$

$$iid$$

- **No initial transient on time scale of regenerative cycles**
- **Structure of initial transient highly dependent on time scale used**

How does initial bias manifest itself on time scale of regenerative cycles?

$$\alpha = \frac{\mathrm{E}\int_0^{T_1} Y(s)ds}{\mathrm{E}\,T_1} \triangleq \frac{\mathrm{E}\,W}{\mathrm{E}\,T}$$

$$\alpha_n = \frac{\overline{W}_n}{\overline{T}_n} = g(\overline{W}_n, \overline{T}_n)$$

$$\mathrm{E}\alpha_n \neq g(\mathrm{E}\overline{W}_n, \mathrm{E}\overline{T}_n)$$

non-linearity bias

Solutions:

- **Taylor expand ...**

estimate

$$\mathrm{E}\alpha_n = \alpha - \frac{1}{n}\frac{\mathrm{E}[(W-\alpha T)T]}{(\mathrm{E}\,T)^2} + O(n^{-2}) \qquad \text{CL75}$$

- **Jack-knife**

On time scale of simulated time:

$$\mathrm{E}\,\frac{1}{t}\int_0^t Y(s)ds = \alpha + \frac{b}{t} + O(e^{-\lambda t})$$

where

$$b = \frac{\mathrm{E}\int_0^{T_1} s[Y(s)-\alpha]ds}{\mathrm{E}\,T_1}$$

estimate from ↗
cycles completed by t

G 90

Typical Theoretical Analysis

$$\mathrm{E}\,\widehat{\alpha}_L(t) = \alpha + o(1/\tau)$$

$$\mathrm{var}\,\widehat{\alpha}_L(t) \sim \mathrm{var}\,\widehat{\alpha}(t) \qquad \textbf{as } t \longrightarrow \infty.$$

Hence, $\widehat{\alpha}_L(t)$ is better...

A More Careful Analysis

$$
\begin{aligned}
\mathrm{E}\,\widehat{\alpha}(t) &= \alpha + \frac{b}{t} + o(1/t) \\
\mathrm{E}\,\widehat{\alpha}_L(t) &= \alpha + o(1/t) \\
\mathrm{var}\,\widehat{\alpha}(t) &= \frac{\sigma^2}{t} + \frac{c_1}{t^2} + o(1/t^2) \\
\mathrm{var}\,\widehat{\alpha}_L(t) &= \frac{\sigma^2}{t} + \frac{c_2}{t^2} + o(1/t^2) \\
\mathrm{MSE}(\widehat{\alpha}(t)) &= \frac{\sigma^2}{t} + \frac{c_1 + b^2}{t^2} + o(1/t^2) \\
\mathrm{MSE}(\widehat{\alpha}_L(t)) &= \frac{\sigma^2}{t} + \frac{c_2}{t^2} + o(1/t^2)
\end{aligned}
$$

Need to check that $c_2 < c_1 + b^2$ **in single replication setting; When a large number of processors are used,** $\alpha_L(t)$ **is clearly better** A+G 07

Conclusions

- **There are problems in which initial transient can be a serious issue**

- **There are rigorously supported methods for reducing the effect of the initial transient**

  even in single replication setting!

- **Ideas apply even in nonstationary settings, e.g., when to initialize simulations focused on "rush hours"**