7 Topics in Brief

7.1 Stochastic Systems
7.2 Sensitivity Analysis and Parametric Optimization
7.3 Mixing Times of Markov chains
7.4 Splitting Methods

7.1 Stochastic System Simulation

Up till now we have considered the simulation of random variables, without specific reference to their origin. An important application of MC method is to the performance evaluation and optimization of *stochastic systems*. Such systems arise naturally in all areas of science and engineering, and due to their complexity simulation and Monte Carlo methods are widely used in their analysis.

7.1.1 Discrete-Event Dynamic Systems

Familiar models for stochastic systems include continuous time and discrete time systems, often modeled as state-space systems driven by noise. Another important class is *discrete-event dynamic systems* (DEDSs), where the observed system state changes only at certain time instances, which need need be uniformly spaced.

Some discrete event systems can be modeled as Markov chains (in discrete or continuous time), but most cannot. Consider the following examples:
- An M/M/1 queue.
- A GI/G/1 queue.
- A tandem queueing network.

The area of DEDS simulation is wide and covered by many textbooks as well as various simulation environments and languages. The common approach to simulation of DEDSs employs an *event list*, on which pending events (such as the next arrival, the next machine failure) are listed along with their occurrence time. This events are ordered chronologically according to their future occurrence time, and the next event is chosen from the top of the list. Processing of an event includes adding new pending events to the list.

7.1.2 Statistical Analysis of Dynamic Systems

Dynamic simulation deals with systems that evolve stochastically over time. Our goal is to estimate the expected system performance with respect to pre-defined performance criteria. Let the state of the system be described by a stochastic process X_t. We can identify two types of system simulation types, according to the performance criteria of interest:

1. Finite time (or single-shot) simulation
2. Steady-state simulation.

The first involves simulating the system up to a given time or event. Typical quantities that can be estimated in this framework include:

- $E(g(X_T)|X_0 = x_0)$.
- $P(X_t \geq \gamma \text{ for some } t \in [0,T]|X_0 = x_0)$
- $E(\tau_A|X_0 = x_0)$, where $\tau_A = \inf\{t > 0 : X_t \in A\}$

All these examples can be cast in the familiar form of estimating $\ell = E(Z)$, and the methods described before for this problem are applicable. Basically, N independent simulation runs of the system from the specified initial conditions x_0 are carried out, the value of Z is obtained in each, and the estimate is obtained as the average $\frac{1}{N}\sum_{i=1}^{N} Z_i$. Various variance reduction methods such as importance sampling can be employed here as well.

Steady state simulation, as the name implies, is used to estimate steady-state performance measures. To be specific, suppose (X_t) is a continuous-time Markov process with a unique stationary distribution $\pi(x)$. We wish to estimate

$$\ell = E(H(X)), \quad X \sim \pi$$

Assuming π is also the limiting distribution, this is equivalent to estimating, for some initial condition X_0,

$$\ell = \lim_{t\to\infty} E(H(X_t))$$

or, under ergodicity conditions,

$$\ell = \lim_{T\to\infty} \frac{1}{T}\int_0^T H(X_t)dt\,.$$

More generally, assume that $Z_t = H(X_t)$ is a random process for which the last limit exists. We wish to estimate ℓ.

A steady state quantity (ℓ) is typically estimated using one long simulation run of the system. A natural estimator is

$$\hat{\ell}_T = \frac{1}{T}\int_0^T H(X_t)dt\,.$$

Under some mild condition, the LLN convergence $\hat{\ell}_T \to \ell$ is accompanies by a CLT of the form

$$\sqrt{T}(\hat{\ell}_T - \ell) \to N(0, \sigma^2)$$

(convergence in distribution to a normal RV), implying

$$\ell_T \approx \ell + \frac{\sigma}{\sqrt{T}}V, \quad V \sim N(0,1)\,,$$

from which confidence intervals can be deduced as usual.

The problem is estimating σ^2, which is mostly done empirically. There are two popular approaches:

a. The batch means method: Given the simulation run $(X_t, 0 \le t \le T)$, we may form the straightforward estimate

$$\hat{\ell} = \frac{1}{T-B}\int_B^T H(X_t)dt\,.$$

The problem is how to estimate the variance. For that purpose, the interval $[B, T]$ is divided into N sub-intervals (or batches) of length $T_1 = (T-B)/N$ each, and the partial estimators

$$\hat{\ell}_k = \frac{1}{T_1}\int_{B+(k-1)T_1}^{B+kT_1} H(X_t), \quad k = 1, \ldots, N$$

are computed. Note that $\hat{\ell} = \frac{1}{N}\sum_{k=1}^N \hat{\ell}_k$.

The batch size T_1 is ideally chosen to be long enough so that the estimates $\hat{\ell}_k$ are approximately independent. An estimate of the variance of ℓ_k with corresponding confidence interval can now be obtained empirically as in CMC. The number of batches N should typically be chosen in the range 20-30 to obtain a reasonable estimate of the variance.

There also exist formulas that correct for dependence among batches in the empirical calculation of the variance and confidence intervals.

b. The regenerative method: A stochastic process (X_t) is regenerative if there exist random times points $T_0 < T_1 < \ldots$ such that, essentially, the subprocesses $(X_t,\, T_{i-1} \leq t < T_i)$ are independent and identically distributed. Typically the times T_i are arrival times to a particular state. For example, in a GI/G/1 queue, the regeneration times may be taken as the arrival times to an empty system.

We further assume that $E(T_i - T_{i-1})$is finite.

Under some additional mild conditions, the process (X_t) has a limiting distribution π, which we represent by a random variable $X \sim \pi$. Let (R, τ) be random variables distributed as (R_i, τ_i). It is known, by the reward-renewal theorem, that

$$\ell = E(H(X)) = \lim_{t\to\infty} E(H(X_t)) = \frac{E(R)}{E(\tau)}\,.$$

Let

$$R_i = \int_{T_{i-1}}^{T_i} H(X_t)dt, \quad \tau_i = T_i - T_{i-1}$$

An estimate for ℓ can now be formed be simulating the process over N regenerative cycles, and computing

$$\hat{\ell}_N = \frac{\hat{R}}{\hat{\tau}} \triangleq \frac{\frac{1}{N}\sum_{i=1}^N R_i}{\frac{1}{N}\sum_{i=1}^N \tau_i}$$

We note that this estimate is biases, but converges to ℓ as N is increased. An empirical estimate of the variance can also be formed from the pairs (R_i, τ_i), by observing the CLT:

$$\sqrt{N}(\hat{\ell}_N - \ell) \to N(0, \eta^2)$$

where

$$\eta^2 = \frac{E(R - \ell\tau)^2}{(E\tau)^2}\,.$$

This variance can be estimated empirically using

$$(\hat{\eta_N})^2 = \frac{\frac{1}{N-1}\sum_{i=1}^N (R_i - \hat{\ell}_N\tau_i)^2}{\left(\frac{1}{N}\sum_{i=1}^N \tau_i\right)^2}\,,$$

with corresponding δ-confidence intervals

$$\hat{\ell}_N \pm z_{1-\delta/2}\hat{\eta}_N/\sqrt{N}\,.$$

7.2 Sensitivity Analysis and Parametric Optimization

Consider a stochastic system, whose performance measure depends a system parameter u. We wish to optimize this performance by selecting an optimal parameter.

In general we may write a expected performance measure as

$$\ell(u) = \mathbb{E}_{u_1}[H(X; u_2)] = \int H(x; u_2) f(x; u_1) dx \,,$$

where $u = (u_1, u_2)$ is the parameter vector. We refer to the parameters in u_1 as *distributional parameters*, es they affect the distribution of the basic random variables that drive the system: For example, the arrival rate or service distribution in a queueing system. The parameters in u_2 are *structural parameters*, for example the buffer size or routing algorithm.

Our goal is to find parameters u that optimize the system performance:

$$\min_{u \in U} \ell(u) \,.$$

A more general problem may consider multiple performance criteria, e.g., in the constrained form:

$$\min \ell_0(u), \quad \text{s.t. } \ell_1(u) \leq 0 \,.$$

For complex systems, where analytical solutions are not available, Monte Carlo simulation provides a major tool for system optimization.

A key step in optimizing over continuous parameter is of course evaluation of the gradient, $\Delta_u \ell(u$.

7.2.1 The Score Function Method

We focus here on the case of distributional parameters: $\ell(u) = \mathbb{E}_u(H(X)$, and, for simplicity of presentation, on the case of a scalar parameter u.

The Score Function method allows to estimate the gradient $\Delta\ell(u)$ (as well as higher derivatives) for a given value of the parameter u, using a *single* simulation run of the system.

The following derivation holds under appropriate regularity conditions, that allow in particular changing the order of differentiation and expectation. We assume that these

hold without further mention. We then have

$$\begin{aligned}\nabla\ell(u) &= \frac{\mathrm{d}}{\mathrm{d}u}\int H(x)f(x;u)dx = \int H(x)\frac{\mathrm{d}}{\mathrm{d}u}f(x;u)dx \\ &= ... = E_u[H(x)S(u;x)]\,,\end{aligned}$$

where

$$S(u;x) = \frac{\mathrm{d}}{\mathrm{d}u}\ln f(x;u)$$

is the *score function*. The score function is easy to compute explicitly for common distributions, such as Exponential, Normal, Binomial, Poisson, etc.

The last expectation may now be evaluated using standard MC:

$$\widehat{\nabla\ell(u)} = \frac{1}{N}\sum_{i=1}^{N} H(X_i)S(u;X_i) \quad X_i \sim f(x;u)\,.$$

We observe that the function $\ell(u)$ as well as the derivative with respect to different components of u can all be estimated from a single run of the system, with $X_i \sim f(x;u)$.

We can further employ *importance sampling* to evaluate gradients of different parameter values u from a single run of the system with a fixed parameter v. That is

$$\widehat{\nabla\ell(u)} = \frac{1}{N}\sum_{i=1}^{N} H(X_i)S(u;X_i)\frac{f(X_i;u)}{f(X_i;v)}\,, \quad X_i \sim f(x;v)$$

We mention that the later method caries the advantages and disadvantages of Importance Sampling – in particular, the samples may degenerate when $f(x;u)$ and $f(x;v)$ are different, especially in problems with a large number of parameters ($n > 10$).

The score function method (and similar sensitivity analysis methods) can be extended to stochastic systems (e.g., Markov chains and regenerative processes).

7.2.2 The Stochastic Counterpart Method

Consider the optimization problem

$$\min_u \ell(u) = \min_{tu} E_{u_1} H(X;u_2)\,.$$

The underlying idea in the stochastic counterpart approach is to replace the expected value with sample means, using IS, and then solve the resulting deterministic optimization problem.

Let $X_i \sim f(x; v_1)$. Then

$$\widehat{\ell(\mathrm{u})} = \frac{1}{N} \sum_{i=1}^{N} H(X_i; u_2) \frac{f(X_i; u_1)}{f(X_i; v_1)}$$

The above sum is a deterministic function of $u = (u_1, u_2)$, which can be optimized using iterative optimization algorithms.

7.3 Mixing Times of Markov Chains

We provide a brief glimpse at bounds on the mixing times of (finite, reversible) Markov chains. A lucid introduction to the subject can be found in the textbook

D. Levine, Y. Peres, and E. Wilmer, *Markov Chains and Mixing Times,* AMS, 2008.

A number of probabilistic techniques are available, we only quote here some results that are based on bounding the eigenvalues of the transition matrix.

7.3.1 Basic Definitions

For two probability distributions μ and ν on a finite set Ω, the *total variation distance* is defined as

$$\begin{aligned}\|\mu-\nu\|_{TV} &\stackrel{\triangle}{=} \max_{A\subset\Omega} |\mu(A)-\nu(A)| \\ &= \frac{1}{2}\sum_{x\in\Omega} |\mu(x)-\nu(x)| \,.\end{aligned}$$

Let $P=(P(x,y))$ denote the transition matric of a finite Markov chain. Then P^t is the associated t-stage transition matrix, and $P^t(x,\cdot)$ is the state distribution after t steps, starting from state $X_0=x$.

Theorem 7.1 (Exponential Convergence) *Suppose that P is irreducible and aperiodic, with (unique) stationary distribution π. Then there exist constants $\alpha\in(0,1)$ and $C>0$ such that*

$$\max_x \|P^t(x,\cdot)-\pi\|_{TV} \le C\alpha^t \,.$$

Define the following *distance from stationarity* measure:

$$d(t) \stackrel{\triangle}{=} \max_x \|P^t(x,\cdot)-\pi\|_{TV}$$

The ϵ-mixing time can now be defined as

$$t_{\rm mix}(\epsilon) = \min\{t\ge 0 : d(t)\le\epsilon\} \,.$$

and the *mixing-time* as

$$t_{\rm mix} = t_{\rm mix}(\tfrac{1}{4}) \,.$$

It can be shown that $t_{\rm mix}(\epsilon^k)\le k\cdot t_{\rm mix}(\frac{\epsilon}{2})$. In particular, $t_{\rm mix}(0.5^k)\le k\cdot t_{\rm mix}$.

Recall that the eigenvalues $\{\lambda\}$ of the transition matrix P satisfy $|\lambda| \leq 1$, and 1 is always an eigenvalue. If P is irreducible, then 1 is a *simple* eigenvalue. If P is irreducible and aperiodic, then 1 is the only eigenvalue with $|\lambda| = 1$.

7.3.2 Reversible Chains

Recall that P is *reversible* if there exists a probability distribution $\pi = (\pi(x))$ so that $\pi(x)P(x,y) = \pi(y)P(y,x)$, in which case π is a stationary distribution of P. It is easy to verify the following property (e.g., by noting that P is similar to a symmetric matrix):

Proposition 7.1 *If P is reversible, then all its eigenvalues are real.*

We can therefore arrange these eigenvalues in decreasing order:

$$1 = \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq -1\,.$$

If P is irreducible then $\lambda_2 < 1$, and if P is aperiodic then $\lambda_n > -1$.

Let $\lambda_* = \max_{i \geq 2} |\lambda_i|$. The difference $\gamma_* = 1 - \lambda_*$ is called the *absolute spectral gap*, and $\gamma = 1 - \lambda_2$ is the *spectral gap*. We refer to $t_{rel} = 1/\gamma_*$ as the *relaxation time*.

Remark: If $|\lambda_n|$ is close to one, we can modify P to $\tilde{P} = (1-\epsilon)P + \epsilon I$ for $\epsilon > 0$. The latter remains irreducible and reversible, and is also aperiodic with $\lambda_n > -1 + \epsilon$. Therefore, we may focus on λ_2 as the critical eigenvalue.

If P is aperiodic, it can be shown that $\lim_{t\to\infty} d(t)^{1/t} = \lambda_*$, i.e., $d(t) \approx O(\lambda_*^t)$. More specifically, the following bounds relate the mixing time to the absolute spectral gap.

Theorem 7.2 *Let P be the transition matrix of a reversible, irreducible Markov chain, and let $\pi_{\min} = \min_x \pi(x)$. Then*

$$\log(\frac{1}{2\epsilon})(t_{rel} - 1) \leq t_{mix}(\epsilon) \leq \log\left(\frac{1}{\epsilon \pi_{\min}}\right) t_{rel}\,.$$

7.3.3 The Bottleneck Ratio

Let P be an irreducible and aperiodic transition matrix with stationary distribution π. Denote

$$Q(x,y) = \pi(x)P(x,y), \quad Q(A,B) = \sum_{x\in A, y\in B} Q(A,B)\,.$$

Here $Q(A, B)$ is the probability of moving from A to B in one step when starting from the stationary distribution.

The *bottleneck ratio* of a set S is defined as

$$\Phi(S) = \frac{Q(S, S^c)}{\pi(S)},$$

and the bottleneck ratio of the whole chain is

$$\Phi_* = \min_{S:\pi(S)\leq 0.5} \Phi(S).$$

If $\Phi(S)$ is small, it is "hard" to exit from the set S.

Theorem 7.3 (General lower bound) *Let P be irreducible and aperiodic. Then*

$$t_{mix} \geq \frac{1}{4\Phi_*}.$$

Theorem 7.4 (Reversible chain) *Suppose that P be also reversible. Then the spectral gap $\gamma = 1 - \lambda_2$ satisfies*

$$\frac{\Phi_*^2}{2} \leq \gamma \leq 2\Phi_*.$$

7.3.4 The Path Method

Let P be reversible, and define the *connectivity graph* of P as the graph with the state as the vertices and edges $E = \{(x, y) : P(x, y) > 0\}$. An E-path from x to y is defined in the usual way. Let $Q(x, y) = \pi(x)P(x, y)$.

For each x, y, let Γ_{xy} denote a choice of some path from x to y, of length $|\Gamma_{xy}|$. Define

$$B = \max_{e\in E} \frac{1}{Q(e)} \sum_{(x,y):e\in\Gamma_{xy}} \pi(x)\pi(y)|\Gamma_{xy}|.$$

Roughly, the edge that determines B is *central*, in the sense that many paths go through it.

Theorem 7.5 *Let P be a reversible and irreducible transition matrix with stationary distribution π. Then the spectral gap $\gamma = 1 - \lambda_2$ satisfies $\gamma \geq B^{-1}$.*

This result is actually a special case of the following comparison result, which allows to obtain bounds for perturbations of "nice" chains.

Theorem 7.6 (Comparison Theorem) *Let P and $\tilde{P}$ be reversible transition matrices, with respective stationary distributions π and $\tilde{\pi}$. Suppose that for each each $(x, y) \in \tilde{E}$ there is an E-path from x to y, choose one and denote it by Γ_{xy}. Define the corresponding* congestion ratio *B by*

$$B = \max_{e \in E} \frac{1}{Q(e)} \sum_{x,y: e \in \Gamma_{xy}} \tilde{Q}(x, y) |\Gamma_{xy}| \,.$$

Then

$$\tilde{\gamma} \leq (\max_{x} \frac{\pi(x)}{\tilde{\pi}(x)}) B \gamma \,.$$

7.4 Splitting Methods

The splitting method is essentially used for estimating rare events. It can also be adapted to counting problems and others.

Suppose we want to estimate the probability $P(E)$ of a rare event E (i.e., $P(E) << 1$) for some Markov process (X_t). The standard MC approach is to simulate N independent copies of (X_t), count the number N_1 that satisfied E, and estimate $P(E) \approx \frac{N_1}{N}$. The problem of course is the relative variance.

Define a sequence of events $E_1, \ldots, E_n$ so that

$$E_1 \supset E_2 \ldots E_n = E \,.$$

Then

$$P(E_n) = P(E_1) \cdot P(E_2|E_1) \cdot \ldots \cdot P(E_n|E_{n-1}) \,.$$

Presumably each of this terms is much larger (less rare) than $P(E_n)$, and can be estimated more easily.

Example: Suppose $(X_t, t \geq 0)$ is a Markov chain, starting from x_0. Let E be the event that $f(X_t)$ hits a level $\gamma >> 1$ before it hits 0.
Define E_i as the event that X_t hits γ_i before it hits 0, where $\gamma_1 < \gamma_2 < \cdots < \gamma_n = \gamma$.

Outline:

- Stage 1: Simulate N_0 independent copies of the process (X_t), until event E_1 or its complement occur. Let N_1 be the number of 'positives'. Record the finishing state $x_k, k = 1, \ldots, N_1$ of each of these N_1 processes.
- Stage 2: For each of the N_1 'positives', and run s_1 independent copies of the chain X_t starting from x^k, until E_2 or its complement occur. Let N_2 be the number of "positives" of these $s_1 N_1$ trials.
- Repeat the above until E_n is reached, with N_n 'positives'.
- Set

$$\hat{p}(E_n) = \frac{N_1}{N_0} \prod_{i=1}^{n-1} \frac{N_{i+1}}{s_i N_i} = \frac{N_n}{N_0} \frac{1}{\prod_{i=1}^{n-1} s_i} \,.$$

Under certain conditions, this turns out to be an unbiased estimate of $p(E_n)$.

Application to counting:

Consider a problem whose solution is defined by a set $C_n = (c_1, \ldots, c_n)$ of conditions, such as equations, inequalities, logical expressions, etc. The is, an assignment $s \in S$ is a solution if it satisfies all the conditions. The set S may be discrete or continuous. Let $\mathcal{X}_\setminus$ denote the set of solutions.

Suppose S is discrete (finite), and we wish to find the *number* of solutions, $|\mathcal{X}_n|$. Many (hard) combinatorial problems fall into this framework. (For a concrete example, consider the number of satisfying solutions of a predicate in DNF form. This problem is in the so called sharp-P complexity class.)

As before, we could draw N random elements s uniformly from S, and count the number N_1 of solutions, and estimate $|\mathcal{X}_n| \approx |S|\frac{N_1}{N}$. And again this is not practical if the relative number of solution is small.

Define then X_i as the set of assignments $s \in S$ that satisfy $C_i = (c_1 \ldots, c_i)$, with $X_0 = S$. Note that

$$|\mathcal{X}_n| = |\mathcal{X}_0| \prod_{i=1}^{n} \frac{|\mathcal{X}_i|}{|\mathcal{X}_{i-1}|} \,.$$

We can now apply a similar splitting approach as before:
– Start by sampling uniformly from $S = \mathcal{X}_0$.
– Keep only the solutions in $\mathcal{X}_1$, and duplicate each s_1 times.
– Continue sampling uniformly from X_1 (typically, by using an MCMC/Gibbs sampler which starts from the duplicated samples).
– Etc.

This approach may also be applied to estimation the volume of a continuous set, defined by (a large number of) inequalities.