JHU 553.633/433: Monte Carlo Methods
J. C. Spall
5 September 2018

SELECTED FACTS IN PROBABILITY AND STATISTICS (Chapter 1 in textbook and other sources)

- Probability, Bayes' rule, and expectation
- Some popular distributions (continuous and discrete)
- Properties of variance and covariance
- Limiting theorems
- Poisson processes
- Markov chains
- Basic statistics

Six (plus one) Broad Areas of 553.633 (Per Syllabus)

1. Random number generation
   - Linear generators and other methods
2. Simulation of stochastic differential equations
3. Models, estimation, and statistical inference
   - Uncertainty bounds
   - Applications in reliability, etc.
   - Bootstrap
4. Variance reduction in simulation
5. Markov chain Monte Carlo
6. Dynamical systems: Kalman filter and particle filters
7. Sensitivity analysis and optimization [Maybe!]

All areas above rely heavily on understanding of probability and statistics!

2

Probability Theory

…you can never know too much probability theory. If you are well grounded in probability theory, you will find it easy to integrate results from theoretical and applied statistics into the analysis of your applications.—Daniel McFadden, 2000 Nobel Prize in Economics

- Random variables, distribution functions, and expectations are critical
  - Central tools in Monte Carlo simulation
- Some results based on notions of probabilistic convergence
  - Laws of large numbers
  - Central limit theorems
- Many theoretical results for practical algorithms rely on asymptotic arguments (i.e., convergence)

3

Probability Mass and Density Functions

- Will be concerned with random variables that take on discrete values and continuous values.
  - E.g., Discrete is counting process: 1, 2, 3, ….
  - E.g., Continuous is time to event: $[0, \infty)$
- Probabilistic characterization (e.g., frequency) for discrete and continuous random variables is via probability mass function (pmf) and probability density function (pdf) $f(x) \geq 0$
- Occasional notational shorthand: **argument defines function**, as in $f(x) = f_X(x)$ and $f(y) = f_Y(y)$
  - This shorthand is sometimes used in textbook and also widely in other probability and statistics literature
  - Can be misleading if not properly interpreted: $f(x)$ and $f(y)$ are different functions!
- Key fact (pdf version): $\int f(x)dx = 1$ (integral over domain of X)

4

Bayes' Rule

- A direct consequence of law of total probability is Bayes' rule
- Continuous version:

$$f_{X|Y}(x \mid y) = \frac{f_{Y|X}(y \mid x) f_X(x)}{f_Y(y)},$$

$$\text{where } f_Y(y) = \int f_{Y|X}(y \mid x) f_X(x) dx$$

- Discrete version in textbook p. 3
- Bayes' rule has large role in practical applications of simulation (e.g., MCMC)
- $f(x)$ above (or analogous discrete probability) is called **prior distribution**
  - Prior distribution is source of both controversy and power in a Bayesian analysis

5

Expectation

- Let $\boldsymbol{X} \in \mathbb{R}^m$, $m \in \{1, 2, \ldots\}$, be distributed according to density function $f(\boldsymbol{x})$
- If $E(\|\boldsymbol{h}(\boldsymbol{X})\|) < \infty$, **expected value** of a function $\boldsymbol{h}(\boldsymbol{X})$ is

$$E[\boldsymbol{h}(\boldsymbol{X})] = \int_{\mathbb{R}^m} \boldsymbol{h}(\boldsymbol{x}) f(\boldsymbol{x}) d\boldsymbol{x}$$

- Obvious analogue to above for discrete random vectors
- Important special cases for expected value:
  - Mean: $\boldsymbol{h}(\boldsymbol{X}) = \boldsymbol{X}$
  - Covariance matrix: $\boldsymbol{h}(\boldsymbol{X}) = [\boldsymbol{X} - E(\boldsymbol{X})]\,[\boldsymbol{X} - E(\boldsymbol{X})]^T$

6

Important Continuous Distributions

Name	Notation	$f(x)$	$x \in$	Params.
Uniform	$\mathsf{U}[\alpha, \beta]$	$\frac{1}{\beta - \alpha}$	$[\alpha, \beta]$	$\alpha < \beta$
Normal	$\mathsf{N}(\mu, \sigma^2)$	$\frac{1}{\sigma\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$	$\mathbb{R}$	$\sigma > 0,\ \mu \in \mathbb{R}$
Gamma	$\mathsf{Gamma}(\alpha, \lambda)$	$\frac{\lambda^\alpha x^{\alpha-1} \mathrm{e}^{-\lambda x}}{\Gamma(\alpha)}$	$\mathbb{R}_+$	$\alpha, \lambda > 0$
Exponential	$\mathsf{Exp}(\lambda)$	$\lambda \mathrm{e}^{-\lambda x}$	$\mathbb{R}_+$	$\lambda > 0$
Beta	$\mathsf{Beta}(\alpha, \beta)$	$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$	$[0, 1]$	$\alpha, \beta > 0$
Weibull	$\mathsf{Weib}(\alpha, \lambda)$	$\alpha\lambda\,(\lambda x)^{\alpha-1} \mathrm{e}^{-(\lambda x)^\alpha}$	$\mathbb{R}_+$	$\alpha, \lambda > 0$
Pareto	$\mathsf{Pareto}(\alpha, \lambda)$	$\alpha\lambda\,(1+\lambda x)^{-(\alpha+1)}$	$\mathbb{R}_+$	$\alpha, \lambda > 0$

where: $\Gamma(\alpha) = \int_0^\infty \mathrm{e}^{-x} x^{\alpha-1}\, dx\,, \quad \alpha > 0$

Source: Table 1.1 in Rubinstein and Kroese (2017)

7

Important Discrete Distributions

Name	Notation	$f(x)$	$x \in$	Params.
Bernoulli	$\mathsf{Ber}(p)$	$p^x(1-p)^{1-x}$	$\{0, 1\}$	$0 \leqslant p \leqslant 1$
Binomial	$\mathsf{Bin}(n, p)$	$\binom{n}{x} p^x(1-p)^{n-x}$	$\{0, 1, \ldots, n\}$	$0 \leqslant p \leqslant 1$, $n \in \mathbb{N}$
Discrete uniform	$\mathsf{DU}\{1, \ldots, n\}$	$\frac{1}{n}$	$\{1, \ldots, n\}$	$n \in \{1, 2, \ldots\}$
Geometric	$\mathsf{G}(p)$	$p(1-p)^{x-1}$	$\{1, 2, \ldots\}$	$0 \leqslant p \leqslant 1$
Poisson	$\mathsf{Poi}(\lambda)$	$\mathrm{e}^{-\lambda} \frac{\lambda^x}{x!}$	$\mathbb{N}$	$\lambda > 0$

Source: Table 1.2 in Rubinstein and Kroese (2017)

8

Important Special Case: Multivariate Normal Distribution

- Most popular distribution for random vector $\boldsymbol{X}$ is multivariate normal (MVN)
  - Defining property is that $\boldsymbol{t}^T\boldsymbol{X}$ must be scalar normal for *all* deterministic $\boldsymbol{t}$ ($\boldsymbol{0}$ vector is special case of normal to allow $\boldsymbol{t} = \boldsymbol{0}$)
- MVN written as $N(\boldsymbol{\mu}, \Sigma)$ for mean vector $\boldsymbol{\mu}$ and covariance matrix Σ
- PDF for MVN is

$$f(\boldsymbol{x}) = \frac{1}{\sqrt{\det(2\pi\Sigma)}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}$$

- Video illustrating why having each component of $\boldsymbol{X}$ normally distributed does ***not*** imply $\boldsymbol{X}$ is MVN: https://www.youtube.com/watch?v=TbAbwtnTbZM (watch from 2:00)

9

Properties of Variances and Covariances

The *covariance* of two random variables X and Y with expectations $\mathbb{E}[X] = \mu_X$ and $\mathbb{E}[Y] = \mu_Y$, respectively, is defined as

$$\mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)] .$$

1	$\mathrm{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$
2	$\mathrm{Var}(aX + b) = a^2\mathrm{Var}(X)$
3	$\mathrm{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]$
4	$\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$
5	$\mathrm{Cov}(aX + bY, Z) = a\,\mathrm{Cov}(X, Z) + b\,\mathrm{Cov}(Y, Z)$
6	$\mathrm{Cov}(X, X) = \mathrm{Var}(X)$
7	$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X, Y)$
8	X and Y indep. $\Longrightarrow \mathrm{Cov}(X, Y) = 0$

Source: Excerpts from p. 10 in Rubinstein and Kroese (2017)

10

Law of Large Numbers

- Consider set of i.i.d. random vectors $\{\boldsymbol{X}_i\}$ with mean $\boldsymbol{\mu} \equiv E(\boldsymbol{X}_i)$
- Let $\bar{\boldsymbol{X}}_n$ be the sample mean based on n values of $\boldsymbol{X}_i$
- The **strong law of large numbers** (SLLN) states that:

$$P\left(\lim_{n\to\infty} \bar{\boldsymbol{X}}_n = \boldsymbol{\mu}\right) = 1$$

- The **weak law of large numbers** (WLLN) states that:

$$\lim_{n\to\infty} P\left(\left\|\bar{\boldsymbol{X}}_n - \boldsymbol{\mu}\right\| \geq \varepsilon\right) = 0 \quad \text{for any } \varepsilon > 0$$

- **SLLN** $\Rightarrow$ **WLLN**
- Practical meaning is that sample mean of i.i.d. random vectors is close to true (unknown) mean if n is large

11

Convergence in Distribution and Central Limit Theorem

- Finite-sample results are usually hopeless in analyzing most practical simulation results
- Notion of asymptotic (large-sample) distribution arises frequently in simulation; e.g.:
  - Central limit theorem
  - Stationary distribution for MCMC
  - Estimation bounds and Fisher information
- **Definition.** Random vector $\boldsymbol{Z}_k$ converges in distribution to $\boldsymbol{Z}$ (written $\boldsymbol{Z}_k \xrightarrow{\text{dist.}} \boldsymbol{Z}$ or $\boldsymbol{Z}_k \xrightarrow{\text{dist.}} F_{\boldsymbol{Z}}(\boldsymbol{z})$) if

$$\lim_{k\to\infty} F_{\boldsymbol{Z}_k}(\boldsymbol{z}) = F_{\boldsymbol{Z}}(\boldsymbol{z})$$

at every point $\boldsymbol{z}$ where $F_{\boldsymbol{Z}}(\boldsymbol{z})$ is continuous.

12

Convergence in Distribution and Central Limit Theorem (cont'd)

- Intuitively, when $\boldsymbol{Z}_k \xrightarrow{\text{dist.}} \boldsymbol{Z}$, then we know that the cdf of $\boldsymbol{Z}_k$ starts to look like the cdf for $\boldsymbol{Z}$ when k is large
  - Additional conditions required to ensure that the pdfs start to look similar for large k
- Example of convergence in distribution is "the" central limit theorem
- Let $\bar{\boldsymbol{X}}_k$ denote sample mean of k random vectors, each having true mean μ and covariance matrix Σ
- Then from the CLT

$$\boldsymbol{Z}_k \equiv \sqrt{k}(\bar{\boldsymbol{X}}_k - \mu) \xrightarrow{\text{dist.}} N(\boldsymbol{0}, \Sigma)$$

13

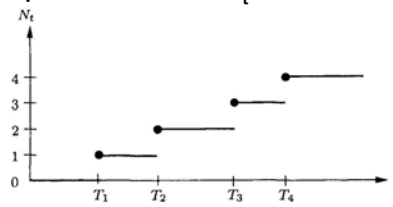
Hypothetical Application of Simulation and Central Limit Theorem

- Suppose have large Monte Carlo simulation that simulates cost of traffic in downtown Baltimore; simulation output has unknown pdf but known variance σ^2
- City planners and traffic engineers want to consider (costly) modification in traffic signal timings and street configuration
- Hypothetical scenario: Monetary value of people's time waiting in traffic and current daily cost of road maintenance is $1 million/day (made up number!!)
- Run simulation N times for modified system and produce sample mean cost of $900K/day, including amortized (per day) construction costs for changing system
- Use of CLT: If the value $900K/day lies in "extreme" lower tail of $N(0, \sigma^2/N)$ distribution, then have indication that changes are worth doing
  - I.e., there is indication of "real" cost savings (statistical test)
- May suggest Baltimore should move forward with modification

14

Poisson Processes

- Poisson distribution fundamental for characterizing number of random events in a given time period
- Key assumption is that number of events in one time interval is independent of number of events in a different non-overlapping interval
- Poisson distribution hugely useful in practice
  - Many practical systems have approximate independence of events in non-overlapping intervals (e.g., Chaps. 3−4 of textbook)
- Example sample path of process with $N_t \equiv$ # of events in time interval [0, t]



(Source: Fig. 1.3 in Rubinstein and Kroese, 2017)

15

Poisson Processes (cont'd)

- Poisson and exponential distributions are directly connected: Poisson (discrete) for frequency and exponential (continuous) for time between events
- Counting process $\{N_t\}$ is Poisson with rate λ iff the inter-arrival times $T_1 - 0$, $T_2 - T_1$, $T_3 - T_2$,…. are i.i.d. with exponential distribution having mean $1/\lambda$
- Poisson and exponential probability functions:

$$\text{Poisson: } P(N_t = k) = e^{-\lambda t}\frac{(\lambda t)^k}{k!}$$

$$\text{Exponential: } f(t) = \lambda e^{-\lambda t}$$

- Means of Poisson and exponential (w/ T = arbitrary inter-arrival time):

$$E(N_t) = \lambda t \text{ and } E(T) = \frac{1}{\lambda}$$

16

Markov Processes and Markov Chains

- Consider stochastic process $\boldsymbol{X}_0, \boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$. Sequence is *Markov process* if:

$$P(\boldsymbol{X}_{t+1} \in \Lambda | \boldsymbol{X}_0, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_t) = P(\boldsymbol{X}_{t+1} \in \Lambda | \boldsymbol{X}_t)$$

  for all dimensionally appropriate sets Λ.
- **Equivalently: Conditional probability of process depends only on most recent value of process, not on full collection of past values**
- Example of Markov process: "State equation" for linear system (e.g., model for Kalman filter)
- Important special case of Markov process is discrete *Markov chain*
  - $\boldsymbol{X}_t$ may take on only a discrete number of values
  - Number of possible values may be finite or infinite

17

Markov Chains

- Let {1, 2,…,*m*} represent labels for possible values in Markov chain (sometimes called *state space* for chain)
  - That is, each X_t may take on one of values {1, 2,…, *m*} (actual values of random process may be something else, but we use integer labels for convenience)
- Transition probabilities: $p_{ij} \equiv P(X_{t+1} = j \mid X_t = i)$
- Transition probabilities collected into transition matrix:

$$\boldsymbol{P} \equiv \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1m} \\ p_{21} & p_{22} & \cdots & p_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ p_{m1} & p_{m2} & \cdots & p_{mm} \end{bmatrix}$$

- Transition matrix $\boldsymbol{P}$ contains all information for Markov chain analysis

18

Markov Chains (cont'd)

- Distribution vector for X_t is key probability of interest:

$$\pi^{(t)} \equiv [P(X_t = 1), P(X_t = 2), \ldots, P(X_t = m)]$$

- Key relationship between $\pi^{(t)}$ and $\boldsymbol{P}$:

$$\pi^{(t+1)} \equiv \pi^{(t)}\boldsymbol{P} \quad \text{(user specified } \pi^{(0)}\text{)}$$

- Simple recursion yields:

$$\pi^{(t)} \equiv \pi^{(0)}\boldsymbol{P}^t \quad (\boldsymbol{P}^0 = \text{identity matrix, } \boldsymbol{I})$$

19

Classification of States in Markov Chains

- Let $P^t(i,j)$ be the ijth element of $\boldsymbol{P}^t$
- State j is *accessible* from state i if $P^t(i,j) > 0$ for some $t \geq 1$ (textbook uses "i leads to j")
  - $\Rightarrow$ Beginning in state i, it is possible that process will eventually reach state j
- States i and j ***communicate*** if each is accessible from the other
- Chain is ***irreducible*** if all m states communicate
- If there are two consecutive numbers s and $s + 1$ such that process can be in state i at times s and $s + 1$, then state is called ***aperiodic***
  - When state i is aperiodic, it is possible to return to state i at irregular times
- State i is ***positive recurrent*** if, starting in i, the expected time for the process to reenter i is finite.
- Positive recurrent states that are aperiodic are ***ergodic***; chain is ergodic if relationship holds for all states

20

Convergence Theorem for Markov Chains

- **Theorem.** Consider discrete Markov chain with finite state space. If chain is irreducible and ergodic (i.e., positive recurrent states and aperiodic), then π is unique solution to

$$\pi = \pi \boldsymbol{P}, \qquad (*)$$

  where components of π satisfy

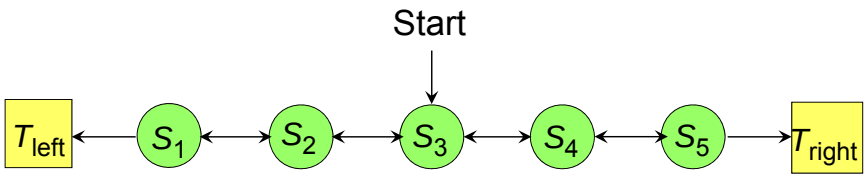
$$\pi_j = \lim_{t \to \infty} P^t(i, j)$$

- Eqn. (*) if of fundamental importance in Markov chain theory and applications
  - Says that long-run probabilities for $\boldsymbol{X}_t$ are independent of initial probabilities (i.e., $\pi^{(t)} \to \pi$ as $t \to \infty$)
  - Allows for determination of steady-state (stationary) probabilities of system
  - Reduces probability problem to linear algebra problem (solving (*) for π)

21

Example Markov Chain: Random-Walk Model
(Example 11.3 in Spall, 2003)

- All walks begin in state S_3
- Each step involves 50–50 chance of moving left or right until terminal state T_{left} or T_{right} is reached



- 7 states (two are absorbing states)
- Transition matrix has banded structure with entries ½ above and below diagonal

22

Basic Statistics: The Standard One-Sample Test

- One set of i.i.d. data $\{X_i\}$ for testing on $\mu \equiv E(X_i)$
- Famous test statistics

$$z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \quad \text{or} \quad t = \frac{\bar{X} - \mu}{s/\sqrt{n}},$$

  where σ^2 is variance of X_i and s^2 is standard unbiased estimate of σ^2
- z and t have a $N(0, 1)$ and t-distribution, respectively
- t-statistic useful in small samples; both z and t often used with non-normal samples
- Large values of $|z|$ or $|t|$ indicate ***rejection*** of null hypothesis that μ is some chosen value (commonly $\mu = 0$)

23

P-Values

- ***P*-value: Probability that future experiment would have value of test statistic *at least as extreme* as that observed in the current experiment**
- Provides info. beyond binary accept/reject null hypothesis
  - Useful as indicator of strength of rejection
- **Example:** If $z = 2.15$, *P*-value is 0.016 based on null hypothesis that $\mu \leq 0$
  - Fairly strong evidence that $\mu > 0$

24

Two-Sample Tests

- Two sets of data $\{X_i\}$ and $\{Y_i\}$ for testing $\mu_X = \mu_Y$
  - E.g., X_i and Y_i represent simulation outputs under two scenarios
- Generic test statistic form

$$t = \frac{\bar{X} - \bar{Y}}{\sqrt{(\cdot)}}$$

  where $(\cdot)$ denotes appropriate variance estimate
- Three basic categories of tests affecting $(\cdot)$ in denominator of t
  - matched pairs
  - unmatched pairs; identical variances ($\sigma_X^2 = \sigma_Y^2$)
  - unmatched pairs; non-identical variances ($\sigma_X^2 \neq \sigma_Y^2$)
- Large values of $|t|$ indicate ***rejection*** of null hypothesis that $\mu_X = \mu_Y$

25