Applied Probability and Statistics

Module-5, Lecture-1

By: Dr. A K Yadav (9911375598)
Dept of CSE, ASET, AUUP

- This module addresses the simulation of random draws $X_1, \ldots, X_n$ from a target distribution f.
- The most frequent use of such draws is to perform Monte Carlo integration, which is the statistical estimation of the value of an integral using evaluations of an integrand at a set of points drawn randomly from a distribution with support over the range of integration.
- Estimation of integrals via Monte Carlo simulation can be useful in a wide variety of settings.

- In Bayesian analyses, posterior moments can be written in the form of an integral but typically cannot be evaluated analytically. Posterior probabilities can also be written as the expectation of an indicator function with respect to the posterior.
- The calculation of risk in Bayesian decision theory relies on integration.
- Integration is also an important component in frequentist likelihood analyses.
- For example, marginalization of a joint density relies upon integration.

- Aside from its application to Monte Carlo integration, simulation of random draws from a target density f is important in many other contexts.
- Markov chain Monte Carlo, Bootstrap methods, stochastic search algorithms, and a wide variety of other statistical tools also rely on generation of random deviates.
- Monte Carlo integration motivates our focus on simulation of random variables that do not follow a familiar parametric distribution.

- We refer to the desired sampling density f as the target distribution.
- When the target distribution comes from a standard parametric family, abundant software exists to easily generate random deviates.
- At some level, all code relies on the generation of standard uniform random deviates.
- Given the deterministic nature of the computer, such draws are not really random, but a good generator will produce a sequence of values that are statistically indistinguishable from independent standard uniform variates.

- Rather than rehash the theory of uniform random number generation, we focus on the practical quandary faced by those with good software: what should be done when the target density is not one easily sampled using the software?
- For example, nearly all Bayesian posterior distributions are not members of standard parametric families.
- Posteriors obtained when using conjugate priors in exponential families are exceptions.

- There can be additional difficulties beyond the absence of an obvious method to sample f .
- In many cases, especially in Bayesian analyses, the target density may be known only up to a multiplicative proportionality constant.
- In such cases, f cannot be sampled and can only be evaluated up to that constant.
- Fortunately, there are a variety of simulation approaches that still work in this setting.

- Finally, it may be possible to evaluate f, but computationally expensive.
- If each computation of $f(x)$ requires an optimization, an integration, or other time-consuming computations, we may seek simulation strategies that avoid direct evaluation of f as much as possible.
- Simulation methods can be categorized by whether they are exact or approximate.

Rejection Sampling

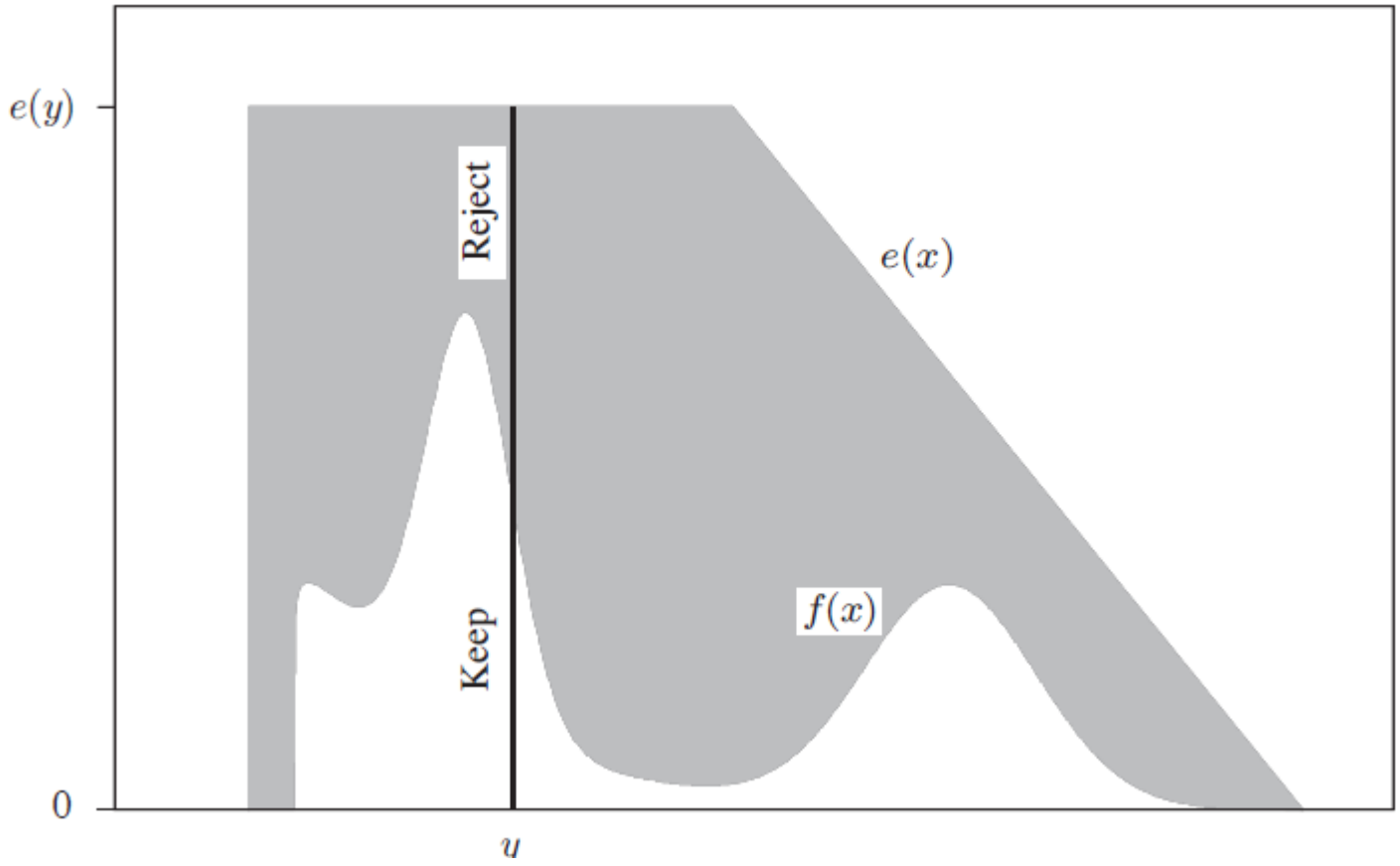
- If $f(x)$ can be calculated, at least up to a proportionality constant, then we can use rejection sampling to obtain a random draw from exactly the target distribution.
- This strategy relies on sampling candidates from an easier distribution and then correcting the sampling probability through random rejection of some candidates.
- Let g denote another density from which we know how to sample and for which we can easily calculate $g(x)$.

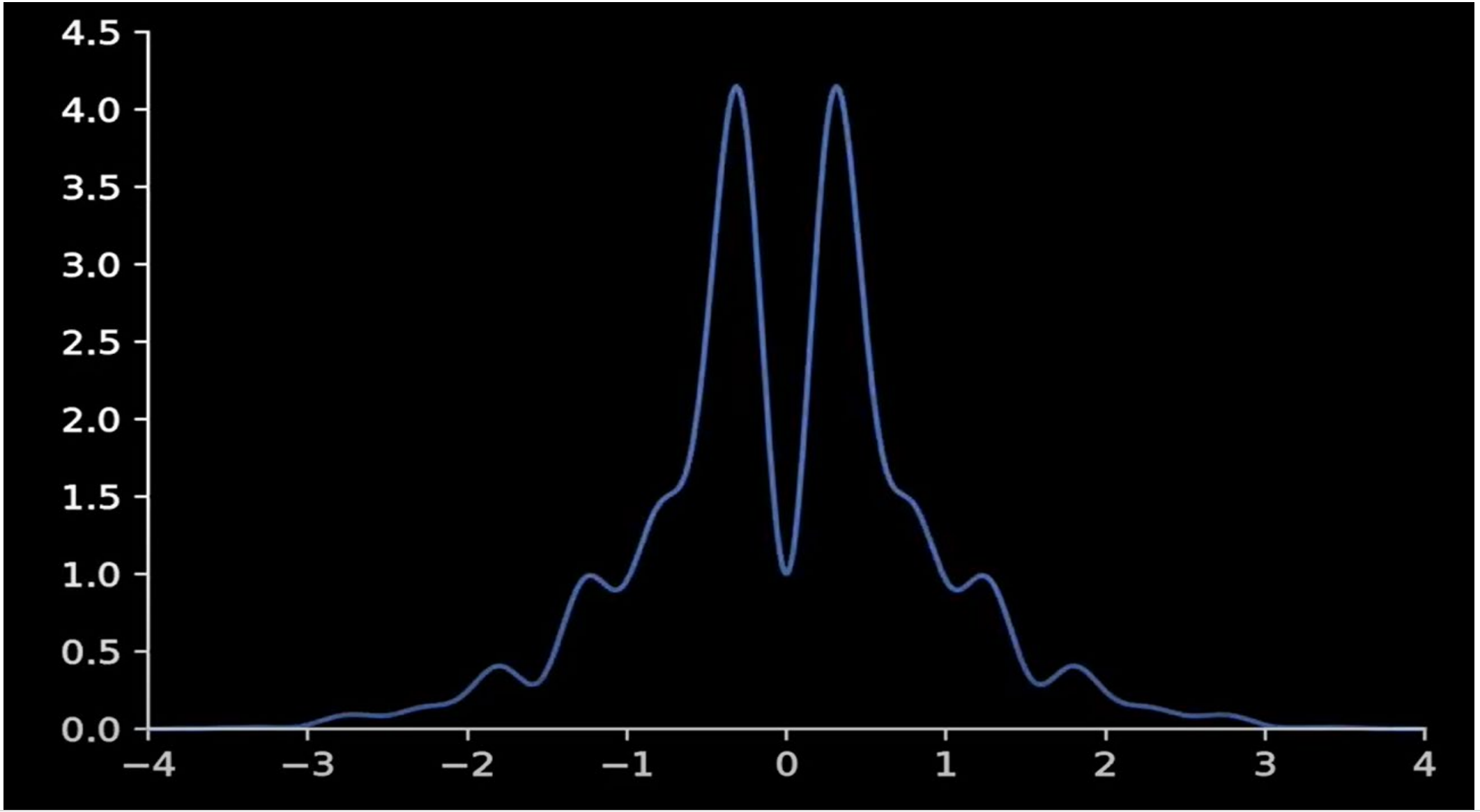
- Let $e(\cdot)$ denote an envelope, having the property $e(x) = \frac{g(x)}{\alpha} \geq f(x)$ for all x for which $f(x) > 0$ for a given constant $\alpha \leq 1$.
- Rejection sampling proceeds as follows:

1. Sample $Y \sim g$.
2. Sample $U \sim \text{Unif}(0, 1)$.
3. Reject Y if $U > f(Y)/e(Y)$. In this case, do not record the value of Y as an element in the target random sample. Instead, return to step 1.
4. Otherwise, keep the value of Y. Set $X = Y$, and consider X to be an element of the target random sample. Return to step 1 until you have accumulated a sample of the desired size.

- The draws kept using this algorithm constitute an i.i.d. sample from the target density f; there is no approximation involved.
- Thus, the sampling distribution is exact, and α can be interpreted as the expected proportion of candidates that are accepted.
- Hence α is a measure of the efficiency of the algorithm.
- We may continue the rejection sampling procedure until it yields exactly the desired number of sampled points, but this requires a random total number of iterations that will depend on the proportion of rejections

- Illustration of rejection sampling for a target distribution f using a rejection sampling envelope e.




4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
−4
−3
−2
−1
0
1
2
3
4

Applied Probability and Statistics

Module-5, Lecture-2

By: Dr. A K Yadav (9911375598)
Dept of CSE, ASET, AUUP

- The simple Monte Carlo estimator of $\int h(x)f(x)dx$ is

$\hat{\mu}_{MC} = \frac{1}{N}\sum_{i=1}^{N} h(X_i)$ where the variables $X_1, \ldots, X_N$ are randomly sampled from f.

- This approach is intuitively appealing, and we have thus far focused on methods to generate draws from f.
- However, in some situations better Monte Carlo estimators can be derived.

- These approaches are still based on the principle of averaging Monte Carlo draws, but they employ clever sampling strategies and different forms of estimators to yield integral estimates with lower variance than the simplest Monte Carlo approach.

- Suppose we wish to estimate the probability that a die roll will yield a one.
- If we roll the die N times, we would expect to see about $N/6$ ones, and our point estimate of the true probability would be the proportion of ones in the sample.
- The variance of this estimator is 5/36N if the die is fair.
- To achieve an estimate with a coefficient of variation of, say, 5%, one should expect to have to roll the die 2000 times.
- To reduce the number of rolls required, consider biasing the die by replacing the faces bearing 2 and 3 with additional 1 faces.

- This increases the probability of rolling a one to 0.5, but we are no longer sampling from the target distribution provided by a fair die.
- To correct for this, we should weight each roll of a one by 1/3.
- In other words, let $Y_i = 1/3$ if the roll is a one and $Y_i = 0$ otherwise.
- Then the expectation of the sample mean of the Y_i is 1/6, and the variance of the sample mean is 1/36N.
- To achieve a coefficient of variation of 5% for this estimator, one expects to need only 400 rolls.

- This improved accuracy is achieved by causing the event of interest to occur more frequently than it would in the naïve Monte Carlo sampling framework, thereby enabling more precise estimation of it.
- Using importance sampling terminology, the die-rolling example is successful because an importance sampling distribution (corresponding to rolling the die with three ones) is used to oversample a portion of the state space that receives lower probability under the target distribution (for the outcome of a fair die).

- An importance weighting corrects for this bias and can provide an improved estimator.
- For very rare events, extremely large reductions in Monte Carlo variance are possible.

Applied Probability and Statistics

Module-5, Lecture-3

By: Dr. A K Yadav (9911375598)
Dept of CSE, ASET, AUUP

- What is variance reduction?
  - A variance reduction technique is a statistical technique for improving the precision of a simulation out-put performance measure without using more simulation, or alternatively achieve a desired precision with less simulation effort" (Kleijnen 1974)
- Why do we need variance reduction?
  - In order to make a simulation statistically efficient, i.e., to obtain a greater precision and smaller confidence intervals for the output random variable of interest, variance reduction techniques can be used.

We have been considering the estimation of $\mu = E\{h(\mathbf{X})\}$ using a random sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ drawn from f. Suppose that each $\mathbf{X}_i = (\mathbf{X}_{i1}, \mathbf{X}_{i2})$ and that the conditional expectation $E\{h(\mathbf{X}_i)|\mathbf{x}_{i2}\}$ can be solved for analytically. To motivate an alternate estimator to $\hat{\mu}_{\text{MC}}$, we may use the fact that $E\{h(\mathbf{X}_i)\} = E\{E\{h(\mathbf{X}_i)|\mathbf{X}_{i2}\}\}$, where the outer expectation is taken with respect to the distribution of $\mathbf{X}_{i2}$. The *Rao–Blackwellized estimator* can be defined as

$$\hat{\mu}_{\text{RB}} = \frac{1}{n}\sum_{i=1}^{n} E\{h(\mathbf{X}_i)|\mathbf{X}_{i2}\} \tag{6.80}$$

and has the same mean as the ordinary Monte Carlo estimator $\hat{\mu}_{\mathrm{MC}}$. Notice that

$$\mathrm{var}\{\hat{\mu}_{\mathrm{MC}}\} = \frac{1}{n}\mathrm{var}\{E\{h(\mathbf{X}_i)|\mathbf{X}_{i2}\}\} + \frac{1}{n}E\{\mathrm{var}\{h(\mathbf{X}_i)|\mathbf{X}_{i2}\}\} \geq \mathrm{var}\{\hat{\mu}_{\mathrm{RB}}\} \qquad (6.81)$$

follows from the conditional variance formula. Thus, $\hat{\mu}_{\mathrm{RB}}$ is superior to $\hat{\mu}_{\mathrm{MC}}$ in terms of mean squared error. This conditioning process is often called Rao–Blackwellization due to its use of the Rao–Blackwell theorem, which states that one can reduce the variance of an unbiased estimator by conditioning it on the sufficient statistics [96]. Further study of Rao–Blackwellization for Monte Carlo methods is given in [99, 216, 507, 542, 543].

Example 6.14 (Rao–Blackwellization of Rejection Sampling) A generic approach that Rao–Blackwellizes rejection sampling is described by Casella and Robert [99]. In ordinary rejection sampling, candidates $Y_1, \ldots, Y_M$ are generated sequentially, and some are rejected. The uniform random variables $U_1, \ldots, U_M$ provide the rejection decisions, with Y_i being rejected if $U_i > w^*(Y_i)$, where $w^*(Y_i) = f(Y_i)/e(Y_i)$. Rejection sampling stops at a random time M with the acceptance of the nth draw, yielding $X_1, \ldots, X_n$. The ordinary Monte Carlo estimator of $\mu = E\{h(X)\}$ can then be reexpressed as

$$\hat{\mu}_{\mathrm{MC}} = \frac{1}{n} \sum_{i=1}^{M} h(Y_i) 1_{\{U_i \le w^*(Y_i)\}}, \tag{6.82}$$

which presents the intriguing possibility that $\hat{\mu}_{\mathrm{MC}}$ somehow can be improved by using all the candidate Y_i draws (suitably weighted), rather than merely the accepted draws.

Rao–Blackwellization of (6.82) yields the estimator

$$\hat{\mu}_{\mathrm{RB}} = \frac{1}{n} \sum_{i=1}^{M} h(Y_i) t_i(\mathbf{Y}), \tag{6.83}$$

where the $t_i(\mathbf{Y})$ are random quantities that depend on $\mathbf{Y} = (Y_1, \ldots, Y_M)$ and M according to

$$\begin{aligned} t_i(\mathbf{Y}) &= E\left\{1_{\{U_i \le w^*(Y_i)\}} | M, Y_1, \ldots, Y_M\right\} \\ &= P[U_i < w^*(Y_i) | M, Y_1, \ldots, Y_M]. \end{aligned} \tag{6.84}$$

Now $t_M(\mathbf{Y}) = 1$ since the final candidate was accepted. For previous candidates, the probability in (6.84) can be found by averaging over permutations of subsets of the realized sample [99]. We obtain

$$t_i(\mathbf{Y}) = \frac{w^*(Y_i)\sum_{A\in\mathcal{A}_i}\prod_{j\in A} w^*(Y_j)\prod_{j\notin A}[1-w^*(Y_j)]}{\sum_{B\in\mathcal{B}}\prod_{j\in B} w^*(Y_j)\prod_{j\notin B}[1-w^*(Y_j)]}, \tag{6.85}$$

where $\mathcal{A}_i$ is the set of all subsets of $\{1,\ldots,i-1,i+1,\ldots,M-1\}$ containing $n-2$ elements, and $\mathcal{B}$ is the set of all subsets of $\{1,\ldots,M-1\}$ containing $n-1$ elements. Casella and Robert [99] offer a recursion formula for computing the $t_i(\mathbf{Y})$, but it is difficult to implement unless n is fairly small.

Notice that the conditioning variables used here are statistically sufficient since the conditional distribution of $U_1, \ldots, U_M$ does not depend on f. Both $\hat{\mu}_{\mathrm{RB}}$ and $\hat{\mu}_{\mathrm{MC}}$ are unbiased; thus, the Rao–Blackwell theorem implies that $\hat{\mu}_{\mathrm{RB}}$ will have smaller variance than $\hat{\mu}_{\mathrm{MC}}$. □

Stratified sampling

- Stratified random sampling is a method of sampling that involves the division of a population into smaller sub-groups known as strata.
- In stratified random sampling, or stratification, the strata are formed based on members' shared attributes or characteristics such as income or educational attainment, sex, religion etc.
- Stratified random sampling is also called proportional random sampling or quota random sampling.

- Stratified sampling strategies
  - Proportionate allocation
  - Optimum allocation or disproportionate allocation
- Advantages
  - Stratified random sampling allows researchers to obtain a sample population that best represents the entire population being studied.
  - Stratified random sampling involves dividing the entire population into homogeneous groups.

- If measurements within strata have lower standard deviation (as compared to the overall standard deviation in the population), stratification gives smaller error in estimation.
- For many applications, measurements become more manageable and/or cheaper when the population is grouped into strata.
- When it is desirable to have estimates of population parameters for groups within the population - stratified sampling verifies we have enough samples from the strata of interest.

Disadvantages

- Stratified sampling is not useful when the population cannot be exhaustively partitioned into disjoint subgroups.

- It would be a misapplication oft he technique to make subgroups' sample sizes proportional to the amount of data available from the subgroups, rather than scaling sample sizes to subgroup sizes or to their variances, if known to vary significantly.
- Data representing each subgroup are taken to be of equal importance if suspected variation among them justify stratified sampling.
- If subgroup variances differ significantly and the data needs to be stratified by variance, it is not possible to simultaneously make each subgroup sample size proportional to subgroup size within the total population.

- For an efficient way to partition sampling resources among groups that vary in their means, variance and costs, then try for “optimum allocation“
- The problem of stratified sampling in the case of unknown class priors (ratio of subpopulations in the entire population) can have deleterious effect on the performance of any analysis on the dataset, e.g., classification.
- In that case, minimax sampling ratio can be used to make the dataset robust with respect to uncertainty in the underlying data generating process

Applied Probability and Statistics

Module-5, Lecture-4

By: Dr. A K Yadav (9911375598)
Dept of CSE, ASET, AUUP

- Thus far we have treated $X^{(t)}$ with little regard to its dimensionality.
- The Gibbs sampler is specifically adapted for multidimensional target distributions.
- The goal is to construct a Markov chain whose stationary distribution equals the target distribution f .
- The Gibbs sampler does this by sequentially sampling from univariate conditional distributions, which are often available in closed form.

Recall $\mathbf{X} = \left(X_1, \ldots, X_p\right)^{\mathrm{T}}$, and denote $\mathbf{X}_{-i} = (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_p)^{\mathrm{T}}$. Suppose that the univariate conditional density of $X_i|\mathbf{X}_{-i} = \mathbf{x}_{-i}$, denoted $f\left(x_i|\,\mathbf{x}_{-i}\right)$, is easily sampled for $i = 1, \ldots, p$. A general Gibbs sampling procedure can be described as follows:

1. Select starting values $\mathbf{x}^{(0)}$, and set $t = 0$.

2. Generate, in turn,

$$X_1^{(t+1)}|\cdot \sim f\left(x_1|x_2^{(t)}, \ldots, x_p^{(t)}\right),$$

$$X_2^{(t+1)}|\cdot \sim f\left(x_2|x_1^{(t+1)}, x_3^{(t)}, \ldots, x_p^{(t)}\right),$$

$$\vdots$$

$$X_{p-1}^{(t+1)}|\cdot \sim f\left(x_{p-1}|x_1^{(t+1)}, x_2^{(t+1)}, \ldots, x_{p-2}^{(t+1)}, x_p^{(t)}\right),$$

$$X_p^{(t+1)}|\cdot \sim f\left(x_p|x_1^{(t+1)}, x_2^{(t+1)}, \ldots, x_{p-1}^{(t+1)}\right),$$

where $|\cdot$ denotes conditioning on the most recent updates to all other elements of $\mathbf{X}$.

3. Increment t and go to step 2.

- The completion of step 2 for all components of X is called a cycle.
- Several methods for improving and generalizing the basic Gibbs sampler are discussed.
- The term $x_{-i}^{(t)}$, which represents all the components of x, except for x_i, at their current values:

$$\mathbf{x}_{-i}^{(t)} = \left(x_1^{(t+1)}, \ldots, x_{i-1}^{(t+1)}, x_{i+1}^{(t)}, \ldots, x_p^{(t)}\right)$$

LINUX FOR DEVICES

By: Dr. A K Yadav (9911375598)
Dept of CSE, ASET, AUUP

Objective and Learning Outcome

Objective:

- To be able to understand the basics of Linux
- To explores the basic characteristics of Linux Networking
- To helps in learning about Linux Shell, File Structure and Network Administration Services
- It gives overview about the Linux Security Techniques

Learning Outcomes

Student will be able to

- Perform the basic operations for Linux.
- Compare the various Linux security techniques.
- Implement the Docker in Linux.
- Execute the shell scripts on Linux.
- Devise the network administration services.
- Able to design device drivers

Recommended Reading

- Christopher Negus, "Linux Bible: The comprehensive Tutorial, Resource ", 8nd Edition, John Wiley
- Richard Petersen "Linux: The Complete Reference," 6th Edition, Tata Mc Graw Hill
- Jonathan Corbet, Alessandro Rubini, Greg Kroah-Hartman, "Linux Device Drivers", 3rd ed, O'Reilly

Module Assessment and Self Work

Module Assessment

- Quiz
- Assignment

PSDA (Self Work)

- Minor Experiment
- Group Discussion
- Case study

Module I - Topics to be covered

- Introduction to Linux
- File System of the Linux
- General usage of Linux kernel & basic commands
- Linux users and group
- Permissions for file, directory and users
- Searching a file & directory
- zipping and unzipping concepts
- Linux for the Industry 4.0 Era,
- OPENIL and its advantages, Features of OPENIL

About Linux, Linux System Architecture

Linux Kernel

Applied Probability and Statistics

Module-5, Lecture-6

By: Dr. A K Yadav (9911375598)
Dept of CSE, ASET, AUUP

- In computational physics and statistics , the Hamiltonian Monte Carlo algorithm (also known as hybrid Monte Carlo), is a Markov chain Monte Carlo method
- Used for obtaining a sequence of random samples which converge to being distributed according to a target probability distribution for which direct sampling is difficult.
- This sequence can be used to estimate integrals with respect to the target distribution (expected values).

Hamiltonian dynamics operates on a d-dimensional *position* vector, q, and a d-dimensional *momentum* vector, p, so that the full state space has $2d$ dimensions. The system is described by a function of q and p known as the *Hamiltonian*, $H(q, p)$.

Equations of motion. The partial derivatives of the Hamiltonian determine how q and p change over time, t, according to Hamilton's equations:

$$\frac{dq_i}{dt} = \frac{\partial H}{\partial p_i} \tag{2.1}$$

$$\frac{dp_i}{dt} = -\frac{\partial H}{\partial q_i} \tag{2.2}$$

for $i = 1, \ldots, d$. For any time interval of duration s, these equations define a mapping, T_s, from the state at any time t to the state at time $t + s$. (Here, H, and hence T_s, are assumed to not depend on t.)

Alternatively, we can combine the vectors q and p into the vector $z = (q, p)$ with $2d$ dimensions, and write Hamilton's equations as

$$\frac{dz}{dt} = J\nabla H(z) \tag{2.3}$$

where ∇H is the gradient of H (ie, $[\nabla H]_k = \partial H/\partial z_k$), and

$$J = \begin{bmatrix} 0_{d\times d} & I_{d\times d} \\ -I_{d\times d} & 0_{d\times d} \end{bmatrix} \tag{2.4}$$

is a $2d \times 2d$ matrix whose quadrants are defined above in terms identity and zero matrices.

Potential and kinetic energy. For Hamiltonian Monte Carlo, we usually use Hamiltonian functions that can be written as follows:

$$H(q,p) = U(q) + K(p) \tag{2.5}$$

Here, $U(q)$ is called the *potential energy*, and will be defined to be minus the log probability density of the distribution for q that we wish to sample, plus any constant that is convenient. $K(p)$ is called the *kinetic energy*, and is usually defined as

$$K(p) = p^T M^{-1} p \,/\, 2 \tag{2.6}$$

Here, M is a symmetric, positive-definite "mass matrix", which is typically diagonal, and is often a scalar multiple of the identity matrix. This form for $K(p)$ corresponds to minus the log probability density (plus a constant) of the zero-mean Gaussian distribution with covariance matrix M.

With these forms for H and K, Hamilton's equations, (2.1) and (2.2), can be written as follows, for $i = 1, \ldots, d$:

$$\frac{dq_i}{dt} = [M^{-1}p]_i \tag{2.7}$$

$$\frac{dp_i}{dt} = -\frac{\partial U}{\partial q_i} \tag{2.8}$$

A one-dimensional example. Consider a simple example in one dimension (for which q and p are scalars and will be written without subscripts), in which the Hamiltonian is defined as follows:

$$H(q,p) = U(q) + K(p), \quad U(q) = q^2/2, \quad K(p) = p^2/2 \tag{2.9}$$

As we'll see later in Section 3.1, this corresponds to a Gaussian distribution for q with mean zero and variance one. The dynamics resulting from this Hamiltonian (following equations (2.7) and (2.8)) is

$$\frac{dq}{dt} = p, \quad \frac{dp}{dt} = -q, \tag{2.10}$$

Solutions have the following form, for some constants r and a:

$$q(t) = r\cos(a+t), \quad p(t) = -r\sin(a+t) \tag{2.11}$$

Hence the mapping T_s is a rotation by s radians clockwise around the origin in the (q,p) plane. In higher dimensions, Hamiltonian dynamics generally does not have such a simple periodic form, but this example does illustrate some important properties that we will look at next.

Applied Probability and Statistics

Module-5, Lecture-7

By: Dr. A K Yadav (9911375598)
Dept of CSE, ASET, AUUP

- In computational physics and statistics , the Hamiltonian Monte Carlo algorithm (also known as hybrid Monte Carlo), is a Markov chain Monte Carlo method
- Used for obtaining a sequence of random samples which converge to being distributed according to a target probability distribution for which direct sampling is difficult.
- This sequence can be used to estimate integrals with respect to the target distribution (expected values).

Hamiltonian dynamics operates on a d-dimensional *position* vector, q, and a d-dimensional *momentum* vector, p, so that the full state space has $2d$ dimensions. The system is described by a function of q and p known as the *Hamiltonian*, $H(q, p)$.

Equations of motion. The partial derivatives of the Hamiltonian determine how q and p change over time, t, according to Hamilton's equations:

$$\frac{dq_i}{dt} = \frac{\partial H}{\partial p_i} \tag{2.1}$$

$$\frac{dp_i}{dt} = -\frac{\partial H}{\partial q_i} \tag{2.2}$$

for $i = 1, \ldots, d$. For any time interval of duration s, these equations define a mapping, T_s, from the state at any time t to the state at time $t + s$. (Here, H, and hence T_s, are assumed to not depend on t.)

Alternatively, we can combine the vectors q and p into the vector $z = (q, p)$ with $2d$ dimensions, and write Hamilton's equations as

$$\frac{dz}{dt} = J\nabla H(z) \tag{2.3}$$

where ∇H is the gradient of H (ie, $[\nabla H]_k = \partial H/\partial z_k$), and

$$J = \begin{bmatrix} 0_{d\times d} & I_{d\times d} \\ -I_{d\times d} & 0_{d\times d} \end{bmatrix} \tag{2.4}$$

is a $2d \times 2d$ matrix whose quadrants are defined above in terms identity and zero matrices.

Potential and kinetic energy. For Hamiltonian Monte Carlo, we usually use Hamiltonian functions that can be written as follows:

$$H(q,p) = U(q) + K(p) \tag{2.5}$$

Here, $U(q)$ is called the *potential energy*, and will be defined to be minus the log probability density of the distribution for q that we wish to sample, plus any constant that is convenient. $K(p)$ is called the *kinetic energy*, and is usually defined as

$$K(p) = p^T M^{-1} p \,/\, 2 \tag{2.6}$$

Here, M is a symmetric, positive-definite "mass matrix", which is typically diagonal, and is often a scalar multiple of the identity matrix. This form for $K(p)$ corresponds to minus the log probability density (plus a constant) of the zero-mean Gaussian distribution with covariance matrix M.

With these forms for H and K, Hamilton's equations, (2.1) and (2.2), can be written as follows, for $i = 1, \ldots, d$:

$$\frac{dq_i}{dt} = [M^{-1}p]_i \tag{2.7}$$

$$\frac{dp_i}{dt} = -\frac{\partial U}{\partial q_i} \tag{2.8}$$

A one-dimensional example. Consider a simple example in one dimension (for which q and p are scalars and will be written without subscripts), in which the Hamiltonian is defined as follows:

$$H(q,p) = U(q) + K(p), \quad U(q) = q^2/2, \quad K(p) = p^2/2 \tag{2.9}$$

As we'll see later in Section 3.1, this corresponds to a Gaussian distribution for q with mean zero and variance one. The dynamics resulting from this Hamiltonian (following equations (2.7) and (2.8)) is

$$\frac{dq}{dt} = p, \quad \frac{dp}{dt} = -q, \tag{2.10}$$

Solutions have the following form, for some constants r and a:

$$q(t) = r\cos(a+t), \quad p(t) = -r\sin(a+t) \tag{2.11}$$

Hence the mapping T_s is a rotation by s radians clockwise around the origin in the (q,p) plane. In higher dimensions, Hamiltonian dynamics generally does not have such a simple periodic form, but this example does illustrate some important properties that we will look at next.

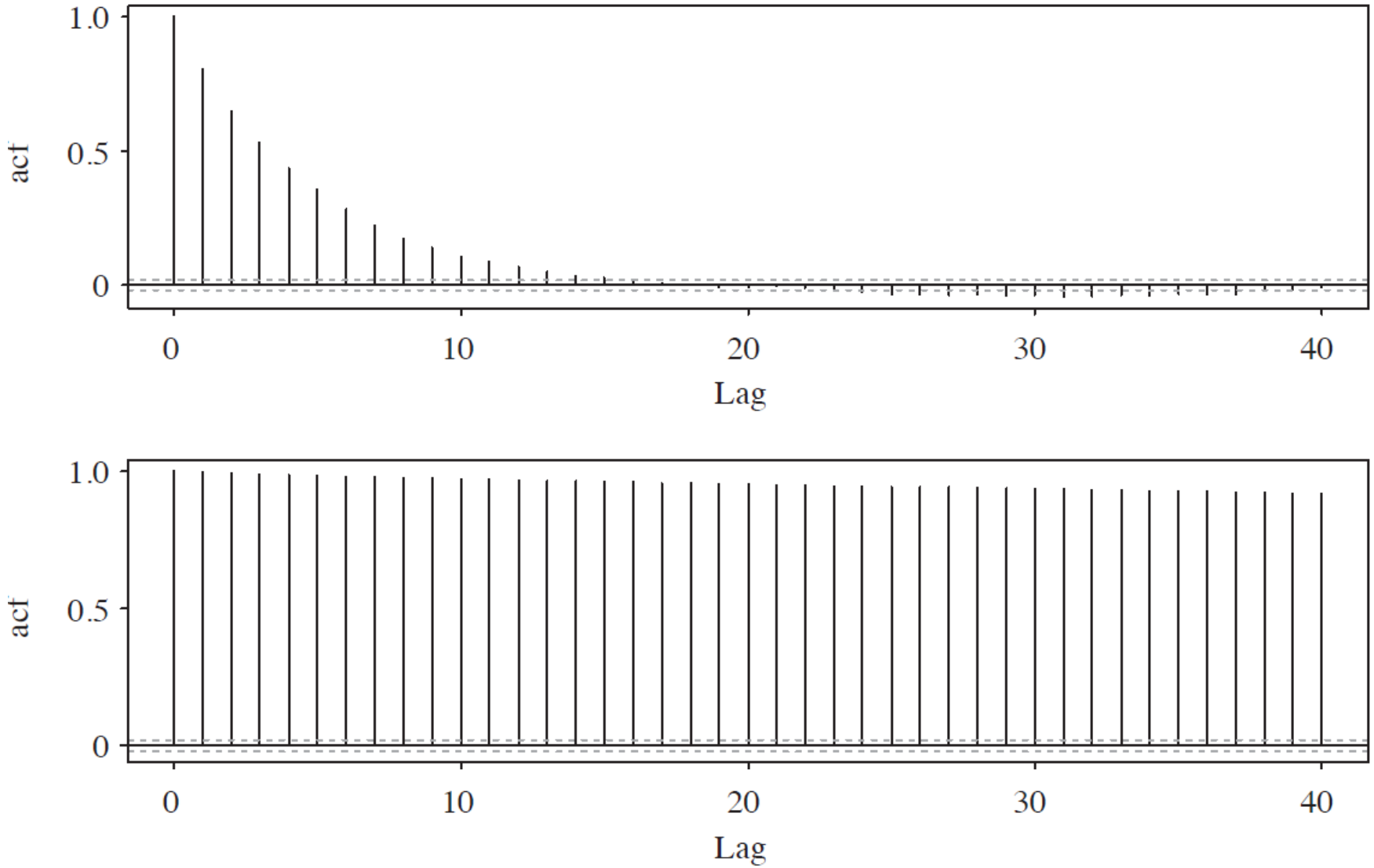
Applied Probability and Statistics

Module-5, Lecture-8

By: Dr. A K Yadav (9911375598)
Dept of CSE, ASET, AUUP

- Has the chain run long enough?
- Is the first portion of the chain highly influence by the starting value?
- Should the chain be run from several different starting values?
- Has the chain traversed all portions of the region of support of f?
- Are the sampled values approximate draws from f?
- How shall the chain output be used to produce f estimates and assess their precision?

- Ensuring Good Mixing and Convergence
  - ✓ Simple Graphical Diagnostics
    - ➢ sample path
    - ➢ cumulative sum
    - ➢ autocorrelation plot
  - ✓ Burn-in and Run Length
  - ✓ Choice of Proposal
  - ✓ Reparameterization
  - ✓ Comparing Chains: Effective Sample Size
  - ✓ Number of Chains
- Practical Implementation Advice
- Using the Results


1.0
0.5
0
acf
0
10
20
30
40
Lag
1.0
0.5
0
acf
0
10
20
30
40
Lag

- ✓Burn-in and Run Length:- This method is based on a statistic motivated by an analysis of variance (ANOVA): The burn-in period or MCMC run-length should be increased if a between-chain variance is considerably larger than the within-chain variance.

Let L denote the length of each chain after discarding D burn-in iterates. Suppose that the variable (e.g., parameter) of interest is X, and its value at the tth iteration of the jth chain is $x_j^{(t)}$. Thus, for the jth chain, the D values $x_j^{(0)}, \ldots, x_j^{(D-1)}$ are discarded and the L values $x_j^{(D)}, \ldots, x_j^{(D+L-1)}$ are retained. Let

$$\bar{x}_j = \frac{1}{L}\sum_{t=D}^{D+L-1} x_j^{(t)} \quad \text{and} \quad \bar{x}_. = \frac{1}{J}\sum_{j=1}^{J} \bar{x}_j, \tag{7.19}$$

and define the between-chain variance as

$$B = \frac{L}{J-1}\sum_{j=1}^{J} \left(\bar{x}_j - \bar{x}_.\right)^2. \tag{7.20}$$

Next define

$$s_j^2 = \frac{1}{L-1}\sum_{t=D}^{D+L-1} \left(x_j^{(t)} - \bar{x}_j\right)^2$$

to be the within-chain variance for the jth chain. Then let

to be the within-chain variance for the jth chain. Then let

$$W = \frac{1}{J}\sum_{j=1}^{J} s_j^2 \tag{7.21}$$

represent the mean of the J within-chain estimated variances. Finally, let

$$R = \frac{[(L-1)/L]W + (1/L)B}{W}. \tag{7.22}$$

If all the chains are stationary, then both the numerator and the denominator should estimate the marginal variance of X. If, however, there are notable differences between the chains, then the numerator will exceed the denominator.

In theory, $\sqrt{R} \to 1$ as $L \to \infty$. In practice, the numerator in (7.22) is slightly too large and the denominator is slightly too small. An adjusted estimator is given by

$$\hat{R} = \frac{J+1}{J} R - \frac{L-1}{JL}.$$

Some authors suggest that $\sqrt{\hat{R}} < 1.1$ indicates that the burn-in and chain length are sufficient [544]. Another useful convergence diagnostic is a plot of the values of $\hat{R}$ versus the number of iterations. When $\hat{R}$ has not stabilized near 1, this suggests lack of convergence. If the chosen burn-in period did not yield an acceptable result, then D should be increased, L should be increased, or preferably both. A conservative choice is to use one-half of the iterations for burn-in. The performance of this diagnostic is improved if the iterates $x_j^{(t)}$ are transformed so that their distribution is approximately normal. Alternatively, a reparameterization of the model could be undertaken and the chain rerun.

- There are several potential difficulties with this approach.
- Selecting suitable starting values in cases of multimodal f may be difficult, and the procedure will not work if all of the chains become stuck in the same subregion or mode.
- Due to its unidimensionality, the method may also give a misleading impression of convergence for multidimensional target distributions.
- Raftery and Lewis proposed a very different quantitative strategy for estimating run length and burn-in period.
- Some researchers advocate no burn-in

- ✓ Choice of Proposal
- ✓ Reparameterization