EE 503 – Week 13 [18-Nov-2013]

Dr. Osonde Osoba

This week:

I. Markov Chain Recap
- Limit Theorems for Ergodic MCs
- Advanced MC applications

II. Introduction to Monte Carlo Simulation

H/W:

None

I. Markov Chains (Recap):

Aperiodic Irreducible (AI) Markov Chains have the following property:

i. $\exists$ unique $\vec{\pi}^*$: $\vec{\pi}^* \underline{P} = \vec{\pi}^*$

ii. $\vec{\pi}_n \longrightarrow \vec{\pi}^*$ $\forall$ initial π_0 $\left(\vec{\pi}_n = \pi_0 \underline{P}^n\right)$

Thus the Markov Chain $\{X_n\}_{n \geq 1}$ is converging in distribution: $X_n \xrightarrow{d} X^*$ where X^* has pdf vector $\vec{\pi}^*$.

[Note that $\{X_n\}_n$ is not independent!!!]

This suggests a scheme for generating samples from the distribution $\vec{\pi}^*$: Simulate an AI MC whose $\lim_n \pi_n = \vec{\pi}^*$ and then take samples of $\{X_n\}_n$ for n "sufficiently" large as samples from the desired distribution, $\vec{\pi}^*$. This is the basic idea behind Markov Chain Monte Carlo (MCMC) schemes.

We can also talk about limit-theorems for Markov Chain similar to LLN and CLT. even though $\{X_n\}_n$ are not independent for MCs.

These MC limit theorems (MC-LLN and MC-CLT) require a condition stronger than AI: Ergodicity

MC Ergodic $\iff$ { i/ Aperiodic and Irreducible [AI]; ii/ "Positive Recurrent" }

Positive recurrence means that the average return time for each state i is finite. i.e.

$$E[T_1(i) \mid X_0 = i] < \infty$$

Theorem: [Positive recurrence for finite-state MCs]

If $\{X_n\}_{n \geq 0}$ is a MC on a finite state space Λ then every recurrent state is positive-recurrent.

This means that Ergodicity for finite state MC reduces to just the AI condition. This is not true for MCs on, say, $\Lambda = \mathbb{N} = \{0, 1, \ldots\}$ or $\Lambda = \mathbb{R}^d$. For example, all states in the 1-d symmetric ($p = 1/2$) random walk (last week) are recurrent. $\Lambda = \mathbb{Z}$ in that example. It turns out all states were not positive recurrent (i.e null recurrent).

Next, we state the main limit theorems for MCs in the more general case where π is a pdf on $\Lambda = \mathbb{R}^d$.

Limit Theorems for Markov Chains:

Suppose:

- $\{X_n\}_{n\geq 0}$ is an ergodic Markov chain with stationary pdf $\pi(x)$.
- $f: \mathbb{R}^d \longrightarrow \mathbb{R}$ is a measurable function.
- $E_\pi[|f|] = \int |f(x)| \cdot \pi(x)\, dx < \infty$.

Then:

(a) [MC-LLN]

$$\overline{f_n(X)} = \frac{1}{n}\sum_{k=1}^{n} f(X_k) \xrightarrow{a.s} E_\pi[f(X)]$$

(b) [MC-CLT]

$$\sqrt{n}\left(\overline{f_n(X)} - E_\pi[f(X)]\right) \xrightarrow{d} W \sim N(0, \sigma_f^2)$$

where:

$$\sigma_f^2 = V_\pi[f(X_1)] + Cov_\pi[f(X_1), f(X_k)]$$

Advanced Applications of MCs:

i/ Markov Random Fields (MRF):

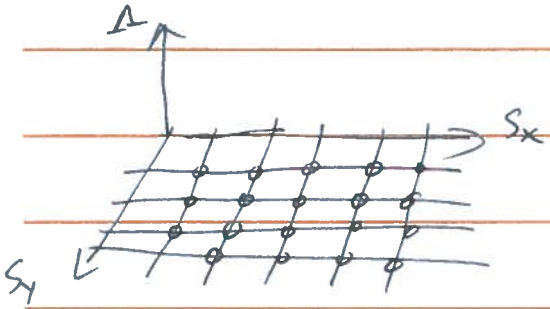
MRFs extend MCs by replacing the "time-like" (or "graph-like") index set $I = \mathbb{N}$ with a "space-like" index set such as $S = \mathbb{N} \times \mathbb{N}$. So the MRF is $\{X_{\vec{s}}\}_{\vec{s} \in \mathbb{N} \times \mathbb{N}}$ instead of the MC $\{X_i\}_{i \in \mathbb{N}}$. This change in index set requires a change the definition of "locality" for the Markov property. Define the local neighborhood of $\vec{s}$, $N(\vec{s})$ as the set of points in S "close" to $\vec{s}$

Then the Markov property say:

$$P(X_{\vec{s}} = t_{\vec{s}} \mid X(S - \{\vec{s}\})) = P(X_{\vec{s}} = t_{\vec{s}} \mid X(N(\vec{s})))$$

i.e. $X_{\vec{s}}$ only depends on the state of $X(\vec{u})$ for $\vec{u}$ close to $\vec{s}$ or $\vec{u} \in N(\vec{s})$

Examples of neighborhoods include: 



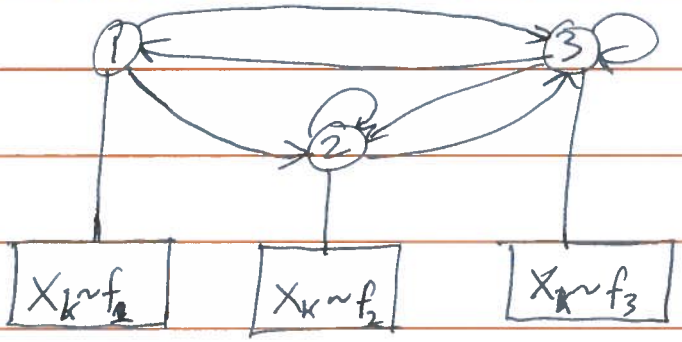
$S = S_x \times S_y$

$\Rightarrow$ A MC is just an MRF with $S = \mathbb{N}$ and $N(s) = \{s-1, s, s+1\}$

ii) Hidden Markov Models (HMMs)

HMMs are useful for modelling non-stationary sequential data (rv samples with changing pdf). The basic approach is to treat the data as samples from one of a handful of pdfs $\{f_i(x)\}_i$. And model the switch between distributions as a MC.

e.g.



The modeller assumes the MC is unobservable/hidden.

The goal is to estimate all model parameters $(\underline{P}, \vec{\pi}, \{f_i\}_i)$ from observed samples $\{X_m\}_{m \geq 1}$.

The standard solution is the Baum-Welch Algorithm.

The algorithm is an ML/EM estimation method.

II Monte Carlo Simulation:

- Theme: use rv sample to estimate population characteristics. Rely on s/m/w LLN for convergence guarantee. Rely on CLT for measurement of estimation errors.

$$\overline{g(X)} \xrightarrow{\text{mgf}} E[g(X)]$$

$(1-\alpha)100\%$ for $E[g(X)]$: $\overline{g(X)} \pm z_{\alpha/2} \cdot \frac{\sigma_{\overline{g(X)}}}{\sqrt{n}}$

- Monte Carlo Simulation is increasingly popular because the complexity of calculating $E[g(X)]$ analytically does not change. But the ease of computing $\overline{g(X)} = \frac{1}{n}\sum_i g(X_i)$ reduces as computers become more powerful.

- Computational Asymmetry between Samples and Populations:

"Samples are easy" vs. "Populations are hard"

$\Rightarrow$ Use sample estimation whenever possible.

Monte Carlo Integration:

Goal: Estimate $I = \int_V g(x)\,dx$

$$I = \int_V g(x)\,dx = \int_V g(x)\,\frac{\text{vol}(V)}{\text{vol}(V)}\,dx$$

$$\Rightarrow I = \text{vol}(V) \cdot \int g(x) \cdot f_X(x)\,dx$$

where $f_X(x) = \begin{cases} 0 & x \notin V \\ 1/\text{vol}(V) & x \in V \end{cases}$

i.e. $X \sim \text{Uniform}(V)$

$$\Rightarrow I = \text{vol}(V) \cdot E[g(X)]$$

$\therefore$ by LLN $\quad \frac{1}{n}\sum_{i=1}^{n} g(X_i) \xrightarrow{\text{a.s.}} E[g(X)]$.

---

e.g. (a) $g(x) = I_D(x)$ [see example in L7, pg 3]

$$\Rightarrow E[g(X)] = E[I_D(X)] = P(X \in D)$$

$$\Rightarrow \frac{1}{n}\sum_{i=1}^{n} g(X_i) = \frac{|\{X_k : X_k \in D\}|}{n} \xrightarrow{\text{a.s.}} P(X \in D)$$

(b) $g(x) = x^p$

$$\Rightarrow \overline{g(X)} \xrightarrow{\text{a.s.}} E[X^p]$$

As an estimator for $E[g(X)]$, $\overline{g(X)}$ is unbiased (since $E[\frac{1}{n}\sum_i g(X_i)] = E[g(X)]$) and

$$\boxed{\text{Var}(\overline{g(X)}) = \frac{1}{n}\,\text{Var}(g(X))}$$

i. MC example : π - estimation

Qu: Find an estimate for $\pi/4$ using MC

Soln:

Consider the 2-dim (standard) uniform pdf, $f(x,y)$:

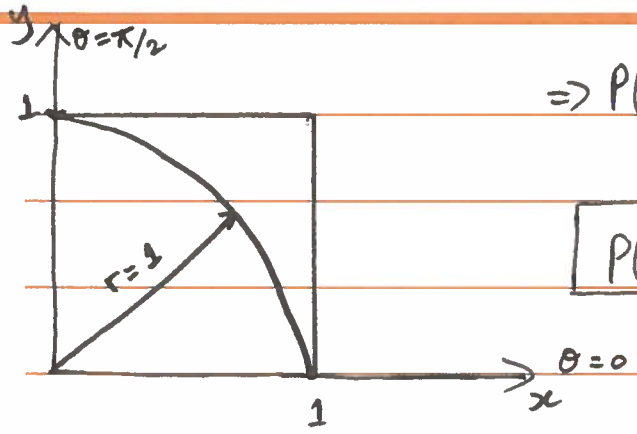
$$f(x,y) = \begin{cases} 1 & x \in (0,1) \text{ and } y \in (0,1) \\ 0 & \text{otherwise} \end{cases}$$

Note: $\text{Vol}(V) = 1$

$$\Rightarrow P(x^2+y^2 \leq 1) = \iint_D f(x,y)\, dx\, dy$$

$$D = \{(x,y) \mid x^2+y^2 \leq 1,\ x \in (0,1),\ y \in (0,1)\}$$

$$\Rightarrow D = \{(r,\theta) \mid 0 \leq r \leq 1,\ 0 \leq \theta \leq \pi/2\}$$



$$\Rightarrow P(X^2+Y^2 \leq 1) = \int_0^{\pi/2} \int_0^1 r\, dr\, d\theta$$

$$= \pi/2 \left[r^2/2 \right]_0^1$$

$$\boxed{P(X^2+Y^2 \leq 1) = \pi/4}$$

Monte Carlo Estimate:

- define $U = \{(x,y) \mid 0 < x < 1,\ 0 < y < 1\}$
- generate N random samples of (x_i, y_i)
- $Z_N \triangleq \dfrac{\#D}{\#U} = \dfrac{\#D}{N} \Rightarrow \boxed{Z_N \to \pi/4}$

Precision of the Monte Carlo Estimate

$n \longrightarrow \infty$ so we can invoke CLT under

$\sigma_x^2 < \infty$ assumption.

i.e $\left(\overline{g(X)} - E[g(X)]\right) \overset{d}{\approx} N\left(0, \sigma^2_{\overline{g(X)}}\right)$

$$\sigma^2_{\overline{g(X)}} = 1/n \; \sigma^2_{g(X)}$$

But in practice we do not know $(\mu_{g(X)}, \sigma^2_{g(X)})$.

So we use the Sample Variance S_n^2

$$\boxed{S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(g(X_i) - \overline{g(X_i)}\right)^2}$$

Thus the $(1-\alpha)100\%$ C-I for the Monte Carlo Estimate is:

$$\boxed{\overline{g(X)} \pm z_{\alpha/2} \cdot \frac{\sigma_{g(X)}}{\sqrt{n}}}$$

$$\approx \overline{g(X)} \pm z_{\alpha/2} \cdot \frac{S_n}{\sqrt{n}}$$

The size of the C-I $\left(|\varepsilon| = z_{\alpha/2}\sigma/\sqrt{n}\right)$ is inversely proportional to the precision of the Monte Carlo estimate.

$\varepsilon \propto 1/\sqrt{n} \Rightarrow$ **Precision $\propto \sqrt{n}$**

Thus Monte Carlo estimates converge rather slowly ($O(\sqrt{n})$) to their estimand. So Monte Carlo estimation is not a good idea unless the problem is hard enough to justify slowly converging estimates.

There are a number of tricks for controlling the variance/precision of the Monte Carlo Estimate:

(a) Stratified Sampling:

$$I = \int_V g(x)\,dx = \sum_{k=1}^{m} \int_{V_k} g(x)\,dx$$

[Assume $\text{vol}(V)=1$]

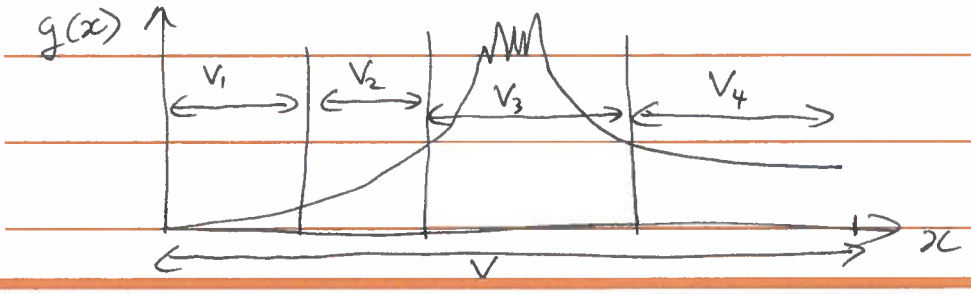
Stratification breaks V into a partition $\{V_k\}_{k=1}^m$.

The Monte Carlo Estimates:

$$MC(I,n) = MC(I,n \mid X \in V) = \frac{1}{n}\sum_{i=1}^{n} g(X_i) \quad \text{[original MC]}$$

vs $$MC_s(I,n) = \sum_{k=1}^{m} MC(I, n_k \mid X \in V_k) \quad \text{[stratified MC]}$$

Converge to I. But the stratified estimator has lower variance and $\therefore$ higher precision. This is especially useful for estimations were $g(x)$ has localized discontinuities.



(b) Importance Sampling:

$$I = \int_V g(x)\,dx$$

We can rewrite I as

$$I = \int_V \left(g(x)/f(x)\right) \cdot f(x)\,dx$$

where $f(x)$ is also a pdf supported over V.

$$\Rightarrow I = E_f\left[\frac{g(X)}{f(X)}\right]$$

and $$MC(I,n) = \frac{1}{n}\sum_{i=1}^{n}\frac{g(X_i)}{f(X_i)}$$

where $X_i \sim f(x)$ [instead of uniform samples]

The right choice of pdf $f(x)$ can reduce the variance of $MC(I,n)$. The optimal importance sampling pdf $f^*(x)$ matches the shape of $g(x)$ roughly.

$$f^*(x) \propto |g(x)|$$

(i.e. fewer samples where $g(x) \approx 0$ and more sample where $g(x) \gg 0$)

General Monte Carlo:

The more general Monte Carlo setting does not use uniform samples: e.g.

$$I = \int g(x)\,dx = \int h(x) \cdot f(x)\,dx$$

where $f(x)$ is a pdf of X

$$\Rightarrow I = E_f[h(X)]$$

Giving our Monte Carlo estimate

$$\boxed{MC(I,n) = \frac{1}{n}\sum_{i=1}^{n} h(X_i)}$$ where $X_i \sim f(x)$ (has pdf)

General Monte Carlo with importance sampling

$$\int h(x) f(x)\,dx = \int \left(h(x) f(x)/f_2(x)\right) \cdot f_2(x)\,dx$$

$$\Rightarrow \int h(x) f(x)\,dx = E_{f_2}\left[\frac{h(X) f(X)}{f_2(X)}\right]$$

$$\Rightarrow MC(I,n) = \frac{1}{n}\sum_{i=1}^{n} \frac{h(X_i) f(X_i)}{f_2(X_i)}$$

where $X_i \sim f_2(x)$