This week: Dr. Osonde Osoba

Markov Chain Monte Carlo Methods (MCMC)

I – Random sampling via $U \sim U(a,b)$
- $F^{-1}(\cdot)$ method
- Rejection sampling
- Monte Carlo Applications

II – MCMC sampling methods
- Gibbs' Sampling
- Metropolis(-Hastings) Algorithm
  - Balance and Detailed Balance

III – MCMC for Parameter Estimation
- Data Augmentation (DA)
- DA vs EM
- Simulated Annealing.

I. Random sampling via $U \sim \text{Uniform}(a,b)$

Suppose we have a random number generator (RNG) that gives iid samples of $U \sim U(0,1)$. We want to generate iid samples of $X \sim f_X(x)$ using U. (abuse of notation: X has pdf $f_X(x)$) If $X \sim U(a,b)$ then a simple scale + shift $aU+b = X$ works. We need more advanced machinery for more complicated distributions.

(a) $F_X^{-1}(\cdot)$ method:

Claim: if $U \sim U(0,1)$ and F_X is absolutely continuous

$\Rightarrow F_X^{-1}(U)$ has pdf $f_X(x)$

Proof: [show that the cdf of $Y = F_X^{-1}(U)$ is $F_X(x)$]

$$F_Y(x) = P(Y \le x)$$

$$Y = F_X^{-1}(U)$$

$$\Rightarrow F_Y(x) = P(F_X^{-1}(U) \le x)$$

$$= P(U \le F_X(x))$$

$$= F_U(F_X(x))$$

$$F_Y(x) = F_X(x) \quad //$$

$$\left[\because F_U(u) = \begin{cases} 0 & u \le 0 \\ u & u \in (0,1) \\ 1 & u \ge 1 \end{cases}\right]$$

e.g. (i) $X \sim \exp(\theta)$

$\Rightarrow F_x(x) = 1 - \exp(-x/\theta) = u$

$\Rightarrow 1 - u = \exp(-x/\theta)$

$\Rightarrow -\theta \ln(1-u) = x$

$\therefore F_x^{-1}(u) = -\theta \ln(1-u) \sim \exp(\theta)$ // [note: $u \overset{d}{=} (1-u)$]

(ii) $X \sim C(0, d)$

$\Rightarrow F_x(x) = \frac{1}{2} + \frac{1}{\pi} \arctan(x/d) = u$

$\Rightarrow \pi(u - \frac{1}{2}) = \arctan(x/d)$

$\Rightarrow d \cdot \tan(\pi(u - \frac{1}{2})) = x$

$F_x^{-1}(u) = d \cdot \tan\left(\pi(u - \frac{1}{2})\right)$ //

(b) Rejection Method: (or "Accept-Reject" method)

- Use 2 random number generators (RNGs) for (1) a simpler proposal pdf $g(x)$ (usually Uniform) and (2) another uniform pdf. $g(x)$ "proposes" samples $X = x$ as samples from the target pdf $f(x)$. We accept or reject $X = x$ based on the value of the (2^{nd}) Uniform sample.

Goal: Samples from target pdf $f(x)$.

Requires: i/ proposal pdf $g(x)$ with larger support than $f(x)$.

ii/ Constant α such that $\alpha \geq [f(x)/g(x)] \geq 0$.

Steps:

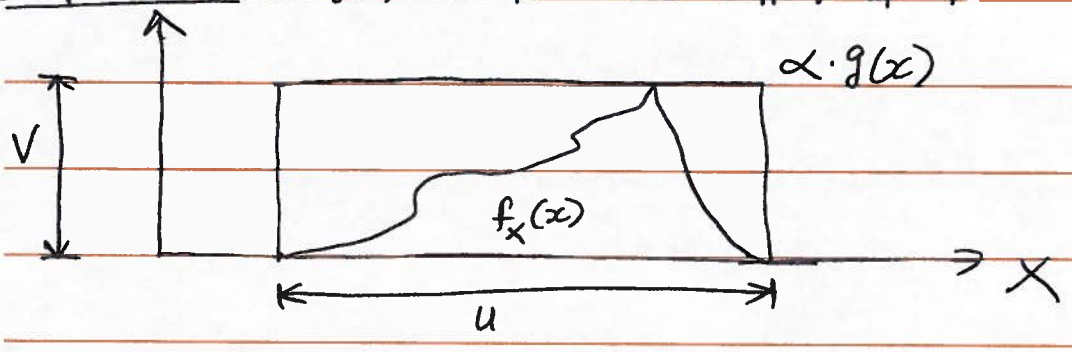
i/ Sample $U \sim g(x)$

ii/ Sample $V \sim \text{Uniform}(0, \alpha \cdot g(U))$ | $\alpha = \max[f(x)/g(x)]$

iii/ [Accept] $X \leftarrow U$ if $V \leq f(U)$

else reject U.

Simplest Case: $g(x)$ uniform over support of f



These sampling methods produce samples of $f(x)$ for use in Monte Carlo estimations of $E_{f(x)}[h(\vec{X})]$.

Ex: An actuary believes that lifetime healthcare costs $L(\vec{X})$ is a function of a set of jointly random factors $\vec{X} = \{X_i\}_{i=1}^{p}$. He wants an estimate of $E[L(\vec{X})]$ for a specific subpopulation. So

$\vec{X}_k \sim f_{\vec{X}}(\vec{x})$

and $\frac{1}{n}\sum_{k=1}^{n} L(\vec{X}_k) \longrightarrow E[L(\vec{X})]$.

He can also simulate "what-if" scenarios for different $f(\vec{x})$ by sampling from the appropriate $f(\vec{x})$

II MCMC Sampling Methods

Both previous methods fail for complicated pdfs: $F^{-1}(\cdot)$ fails for discontinuous cdfs and rejection methods are very inefficient in higher dimensions.

More successful sampling schemes use the following theme:

i/ Design an Ergodic Markov chain with the target pdf $f(x)$ as the equilibrium pdf.

ii/ Run the M.chain "long enough" to reach equilibrium.

$\Rightarrow \{X_n\}_{n \geq \text{burn-in time}}$ are approximately $\sim f(x)$.

[MC Convergence time (or Burn-in time) is a key consideration]

Issues:

i/ Designing such M. Chains for arbitrary pdfs is not trivial. MCMC algorithms typically use ideas of reversibility/Detailed Balance.

ii/ The samples are no longer iid since $\{X_n\}_{n \geq 1}$ is a Markov Chain. There will be some serial correlation even if the samples individually have the right distribution.

We need some new concepts on M. Chains first-

Definition: Reversible distributions / Markov Chains

A distribution $\vec{\pi}$ on a MC is reversible

$\iff \forall i, j \in \Lambda, \quad \boxed{\pi_i P_{ij} = \pi_j P_{ji}}$

("Detailed Balance" Condition)

A MC is reversible if its initial distribution is reversible

Claim:

$$\left[\vec{\pi} \text{ reversible}\right] \Rightarrow \left[\vec{\pi} \underline{P} = \vec{\pi}\right]$$

Proof:

$$\sum_{i \in \Lambda} P_{ji} = 1 \qquad \left[\text{rows sum to } 1\right]$$

$$\Rightarrow \forall j \in \Lambda, \quad \pi_j = \pi_j \sum_{i \in \Lambda} P_{ji}$$

$$\overset{\forall j \in \Lambda}{=} \sum_i \pi_i P_{ji}$$

$$\Rightarrow \pi_j = \sum_i \pi_i P_{ij} \qquad \left[\vec{\pi} \text{ reversible}\right]$$

$$\therefore \vec{\pi} = \vec{\pi} \cdot \underline{P} \quad //$$

A MC in its reversible $\vec{\pi}$ is Physical system in thermodynamic equilibrium. The amount of "energy" flowing in is equal to the amount "energy" flowing out. $\pi_i P_{ij}$ is the probab. mass flowing from $i \to j$ which is equal to $\pi_j P_{ji}$ i.e mass flow from $j \to i$. $\therefore$ All states are "mass-balanced."

Continuous-Valued Markov Chains:

General MCMC methods produce samples $\{X_n\}$ with continuous values. i.e $X_i: \Omega \to \mathbb{R}$, $\Lambda = \mathbb{R}$.

Λ is no longer finite/countable. Markov chains on uncountable state spaces have a more general characterization. [Assume time-homogeneity].

$\Lambda = \{1, \dots, n\}$	$\lvert\Lambda\rvert = \lvert\mathbb{R}\rvert$
i/ $\underline{\underline{P}} = ((p_{ij}))_{ij}$ $p_{ij} = P(X_{n+1} = j \mid X_n = i)$ $(P_{i \to j})$	$P(x, A) = P(X_{n+1} \in A \mid X_n = x)$ [Transition kernel] $(P(x \to A))$
ii/ $\sum_{j \in \Lambda} p_{ij} = 1$	$\int P(x, y)\, dy = \int p(y \mid x)\, dy = 1$ [AC/pdf assumption]
iii/ $\underline{\underline{P}}^{n+1} = (\underline{\underline{P}}^n) \cdot \underline{\underline{P}}$ [Chapman-Kolmogorov]	$P^{n+1}(x, A) = \int P^n(x, y) \cdot P(y, A)\, dy$
iv/ [Balance Eqn] $\vec{\pi} = \vec{\pi} \cdot \underline{\underline{P}}$	$f(y) = \int P(x, y)\, f(x)\, dx$ $= \int P(y \mid x) \cdot f(x)\, dx$
v/ [Detailed Balance] $\pi_i p_{ij} = \pi_j p_{ji}$	$f(x)\, P(x, y) = f(y) P(y, x)$
In general: matrix mult $\longleftrightarrow$ vector mult. $\longleftrightarrow$	$\int P(\cdot, \kappa) \cdot P(\kappa, \cdot)\, d\kappa$ $\int f(\kappa) \cdot P(\kappa, \cdot)\, d\kappa$

MCMC I : Gibbs Sampler

Goal: Samples from $f_{\vec{X}}(\vec{x})$, $\vec{X} = \{X_i\}_{i=1}^{p}$

Problem: $f_{\vec{X}}$ complicated

but $f(x_k | \{x_i\}_{i \neq k})$ simple.

Steps:

Pick first $\vec{x}(0)$ such that $f_{\vec{X}}(\vec{x}(0)) > 0$

1// (1) generate $x_1(t) \sim f(x_1 | \{x_i(t-1)\}_{i \neq 1})$

(2) generate $x_2(t) \sim f(x_2 | x_1(t), \{x_i(t-1)\}_{i>2})$

$\vdots$

(p) generate $x_p(t) \sim f(x_p | \{x_i(t)\}_{i \neq p})$

2// $\vec{x}(t) = \{x_i(t)\}_{i=1}^{p}$

3// $t \leftarrow t+1$, go to 1// until $t = n$

The sample $\{\vec{x}(t)\}_{t=1}^{n}$ are samples of a M-Chain with stationary pdf $f_{\vec{X}}(\vec{x})$.

This is the simplest MCMC method. Usually for high dimensional data (e.g. images). Variations include combining conditional ("blocking") some generation steps and partial marginalizations ("collapsing") depending on structure of conditional pdfs.

Example:

$$\vec{X} = N(\vec{\mu}, \Sigma)$$

$$\mu = \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix} \quad ; \quad \Sigma = \begin{pmatrix} 1 & 1/2 & 0 \\ 1/2 & 1 & 0 \\ 0 & 0 & 2 \end{pmatrix}$$

$\Rightarrow X_3 \perp\!\!\!\perp \{X_1, X_2\} \Rightarrow X_3 \sim N(1, 2)$

$(X_1, X_2) \sim N\left(\begin{pmatrix} -1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 1/2 \\ 1/2 & 1 \end{pmatrix}\right)$; $\sigma_{X_1 X_2} = 1/2 \Rightarrow \ell = 1/2$

$X_1 | X_2 \sim N(-1 + 1/2(x_2 - 0), \ 3/4)$

i.e $X_1 | X_2 = x_2 \sim N(x_2/2 - 1, \ 3/4)$

$X_2 | X_1 = x_1 \quad N(0 + 1/2(x_1 - -1), \ 3/4)$

i.e $X_2 | X_1 \sim N(1/2(x_1 + 1), \ 3/4)$

$$\left[X_1 | (X_2 = x_2) \sim N\left(\mu_1 + \frac{\sigma_1}{\sigma_2} \ell (x_2 - \mu_2), \ \sigma_1^2 (1 - \ell^2)\right) \right]$$

Gibbs steps:

(0) $\vec{x}(0) = \mu = \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix} \quad [\because f_{\vec{X}}(\vec{\mu}) > 0]$

(1) $x_1(t) \sim f_{X_1 | X_2(t-1) X_3(t-1)} = f_{X_1 | X_2(t-1)} = N(x_2/2 - 1, \ 3/4)$

$x_2(t) \sim f_{X_2 | X_1(t) X_3} = f_{X_2 | X_1(t)} = N(1/2(x_1 + 1), \ 3/4)$

$x_3(t) \sim f_{X_3 | X_1(t), X_2(t)} = f_{X_3} = N(1, 2)$

[Repeat step (1)]

(A) shows "collapsing" idea $f_{X_1 | X_2, X_3} = f_{X_1 | X_2}$.

(B) Blocking: (1) $(x_1, x_2) \rightarrow$ then $x_3 | (x_1, x_2)$

MCMC II : Metropolis and Metropolis–Hastings Algorithm

Goal : Samples from $f(x)$

- Modifies Rejection method to include Markovian dynamics.
- Requires a conditional proposal pdf $q(y|x)$.

Steps : [Metropolis Algorithm, $q(\cdot|\cdot)$ symmetric]

0/ $x(0) = x$ such that $f(x) > 0$

1/ Sample $y \sim q(y|x_t)$

2/ Calculate acceptance probability

$$\alpha(x_t, y) = \min\left[1, \frac{f(y)}{f(x_t)}\right] \quad \text{(Hastings Ratio)}$$

3/ Sample $U \sim U(0,1)$

4/ Accept : $x_{t+1} \leftarrow y$ if $U \leq \alpha(x_t, y)$

else : $x_{t+1} \leftarrow x_t$

Metropolis–Hastings adjustment :

i/ $q(y|x_t)$ not symmetric

ii/ $$\alpha(x_t, y) = \min\left[1, \frac{f(y) \cdot q(x_t|y)}{f(x_t) \cdot q(y|x_t)}\right]$$

The transition kernel for M/M-H algorithm:

$$P(x,y) = q(y|x)\cdot\alpha(x,y) + r(x)\,\delta(y-x)$$

where $r(x) = 1 - \int q(y|x)\,\alpha(x,y)\,dy$ [rejection probab.]

$f(x)$, the target pdf, is the reversible pdf for $P(x,y)$ and there for an equilibrium pdf for $P(x,y)$.

Since

$$f(x)\,q(y|x)\cdot\alpha(x,y) = f(y)\,q(x|y)\cdot\alpha(y,x)$$

$$\Rightarrow \quad f(x)\cdot P(x,y) = f(y)\,P(y,x)$$

[Detailed Balance]

Modifications to M/M-H:

- Independence samples:

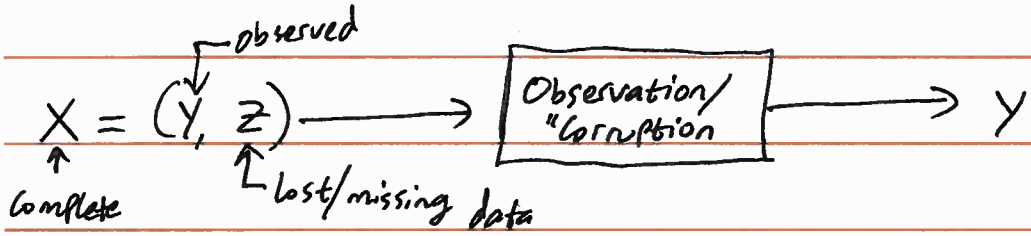
$$q(y|x) = q(y)$$

- Random Walk M-H:

$$q(x,y) = q(y-x)$$

III. MCMC for Parameter Estimation:

MCMC is useful for dealing with "missing information" problems. Typically MCMC provides a method for optimally "filling in" lost info. The scheme goes thus:



$\{X_i\}_{i\geq 1}$: iid common pdf $f(x|\theta) = f(y,z|\theta)$

$\{Y_i\}_{i\geq 1}$: iid common pdf $g(y|\theta) = \int_z f(y,z|\theta)\,dz$

Usually g is very complicated and $f(x|\theta)$ is simple.

Goal: find optimal θ based on observed $\{Y_i\}_{i\geq 1}$

2 Approaches

i/ Impute/Guess good values for missing info Z and use guesses in estimation procedure. e.g

- Single/Multiple Imputation techniques
- Data Augmentation (DA) algorithm

ii/ Use statistics of $Z|Y$ (i.e $f(z|y,\theta)$) to guess at good/compatible likelihood function for Y.

- Expectation-Maximization Algorithm.

i/ **Data Augmentation Algorithm:**

Consider the pdf of the missing data

$$Z \sim p(z|y,\theta) = \frac{f(y,z|\theta)}{g(y|\theta)}$$

and the posterior $\theta \sim f(\theta|y,z)$

Steps:

θ_0 = initial viable estimate for θ

I-step: Generate $Z_t \sim p(z|\theta_{t-1}, Y=y)$

P-step: Generate $\theta_t \sim f(\theta|Y=y, Z_t)$

[repeat I- and P- steps]

This is a Gibbs Sampler for (θ_t, Z_t) consistent with observed samples $Y=y$.

The samples $\{\theta_t\}_t \sim f(\theta|Y, Z_t)$ give increasingly better stochastic estimates for $\hat{\theta}$ after the "burn-in" period. $\{\theta_t\}$ also provides a method for estimating the posterior. So we can use DA/Gibbs for any kind of Bayesian Estimation method.

ii/ The likelihood approach to missing info is the EM algorithm. It still uses the $Z \sim P(Z|y,\theta)$ characterization. But no sampling involved. We just use $P(Z|y,\theta)$ to average out the effects of the missing info.

Steps:

θ_0 = initial estimate

E-Step: $Q(\theta|\theta_t) = E_{Z|y,\theta_t}[\ln f(y,Z|\theta)]$

$= \int_Z \ln f(y,z|\theta) \cdot p(z|y,\theta_t)\, dz$

M-Step: $\theta_{t+1} = \arg\max_\theta \{Q(\theta|\theta_t)\}$

[repeat E & M - Until convergence]

This is like a deterministic version of the DA algorithm/Gibbs sampler. But EM is limited to MLE and MAP while DA does full Bayesian Estimation.

Eg: [GMM] $Y \sim (1-\alpha)\cdot N(\mu_0, \sigma_0^2) + \alpha \cdot N(\mu_1, \sigma_1^2)$

$Y|Z=i \sim N(\mu_i, \sigma_i^2)$

$Z \sim \text{Bernoulli}(\alpha) \Rightarrow Z = \begin{cases} 0 & (1-p)=1-\alpha \\ 1 & p=\alpha \end{cases}$

$\vec{\theta} = (\alpha, \mu_0, \sigma_0^2, \mu_1, \sigma_1^2)$

$\Rightarrow f(y,z|\vec{\theta}) = f(y|z,\vec{\theta}) \cdot p(z|\vec{\theta})$

$= N(\mu_0,\sigma_0^2)\,\delta(z)\cdot(1-\alpha) + N(\mu_1,\sigma_1^2)\,\delta(z-1)\cdot\alpha$

$$P(z|y,\vec{\theta}) = \frac{f(y,z|\vec{\theta})}{f(y|\vec{\theta})}$$

$$= \begin{cases} \dfrac{\alpha\, N(\mu_1, \sigma_1^2)}{\sum P(z=i)\, N(\mu_i, \sigma_i^2)} & z=1 \\ \dfrac{(1-\alpha)\, N(\mu_0\ \sigma_0^2)}{\sum P(z=i)\, N(\mu_i, \sigma_i^2)} & z=0 \end{cases}$$

$$f(\theta|y,z) = \frac{f(y,z|\theta)\cdot h(\theta)}{\int f(y,z|\theta)\cdot h(\theta)\, d\theta}$$

EM:

$$Q(\theta|\theta_t) = \sum_{z\in\{0,1\}} \ln f(y,z|\vec{\theta})\cdot P(z|y,\vec{\theta}_t)$$

↳ $\theta_{MM} = \operatorname{argmax} Q(\theta|\theta_t)$

DA:

$$z_t \sim P(z|y,\vec{\theta}_{t-1})$$

$$\vec{\theta}_t \sim f(\vec{\theta}|y,z_t)$$

Other applications of MCMC: Optimization

The key feature of Monte Carlo methods is the conversion of a hard "Population" problem into a simpler "sampling" problem. The cost is typically slower convergence and some estimation error.

Sometimes it makes sense to convert a fully deterministic problem into a stochastic problem and solve using MC/probabilistic techniques. This typically applies to NP-Hard problems (e.g. TSP, Combinatorial opt'n)

One such problem is the Simulated Annealing method.

Goal: find global minimum of $H(x)$: x^*

Approach:

i/ Specify pdf $f(x|T)$ (T = temp) such that:

$$\lim_{T \to 0} f(x^*|T) = 1$$

ii/ Specify "cooling schedule" $\{T_k\}_{k \geq 1}$ such that

$$T_k \longrightarrow 0 \quad \text{as } k \to \infty$$

iii/ at each step k, get samples of $f(x|T_k)$ using MH/MCMC (starting w/ last sample from previous (k-1) batch)

$\therefore \lim x_k \longrightarrow x^*$ under some conditions.

Typically:

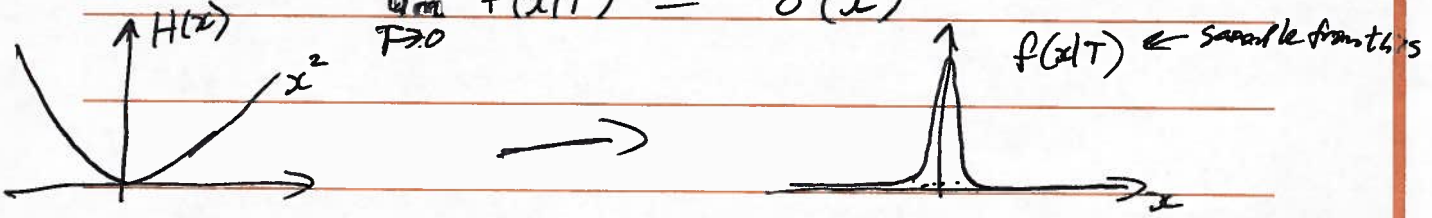
$$f(x|T) \propto \exp(-H(x)/T) \quad \text{[Boltzmann Distribution]}$$

eg $H(x) = x^2$ $\left[\Rightarrow x^* = 0\right]$

$f(x|T) \propto \exp(-x^2/T)$

i.e $f(x|T) = N(0, T/2)$

$$\lim_{T \to 0} f(x|T) = \delta(x)$$



If $\left[T_k \propto 1/\log(k)\right] \Rightarrow \left[x_k \xrightarrow{a.s.} x^*\right]$

[Logarithmic Cooling schedule]

But too slow.

Recall Hastings ratio:

$$\alpha = \min\left\{1, \frac{f(y_{cand})}{f(x_t)}\right\}$$

So even if y_{cand} is worse than x_t, SA can still accept