LECTURE NOTES ON SAMPLING AND MARKOV CHAINS *

A. A. ERGÜR, B. SAMUEL

Abstract. This document contains lecture notes on Markov chains and sampling techniques, with a focus on Markov Chain Monte Carlo (MCMC) methods. Topics covered include random walks, stationary distributions, conductance and mixing times, rejection and importance sampling, and advanced algorithms such as Metropolis–Hastings, Gibbs sampling, and Hamiltonian Monte Carlo. The notes also provide algorithmic derivations, theoretical results, and exercise-based applications related to convergence, marginal estimation, and efficient sampling over discrete and continuous domains.

Key words. Markov chains, Monte Carlo methods, MCMC, Metropolis–Hastings, Gibbs sampling, Hamiltonian Monte Carlo, stationary distribution, mixing time, conductance, rejection sampling, importance sampling, marginal estimation, random walks, symmetric Markov chains

AMS subject classifications. safely ignore

1. Sampling.

1.1. CDF Sampling.

Let $F(t) = \Pr(x \leq t)$.

THEOREM 1.1 (Theorem 11.3.1). *If $U \sim \mathcal{U}(0,1)$ is a uniform random variable, then $F^{-1}(U) \sim F$.*

Proof.

$$\Pr(F^{-1}(U) \leq x) = \Pr(U \leq F(x)) \quad \text{(Applying } F \text{ to both sides)} \tag{1.1}$$

$$= F(x) \quad \text{(since } \Pr(U \leq y) = y \text{ for } U \sim \mathcal{U}(0,1)) \tag{1.2}$$

Where the first line follows since F is a monotonic function, and the second line follows because U is uniform on the unit interval. ∎

1.2. Example: Exponential Distribution.

Let

$$\text{Expon}(x \mid \lambda) = \lambda e^{-\lambda x}\, \mathbb{I}(x \geq 0)$$

The CDF is:

$$F(x) = 1 - e^{-\lambda x}\, \mathbb{I}(x \geq 0)$$

The inverse (quantile function) is:

$$F^{-1}(p) = -\frac{\ln(1-p)}{\lambda}$$

1.3. Rejection Sampling.

Setup:

- Target: $p(x) = \frac{\hat{p}(x)}{Z_p}$
- Proposal: $q(x)$
- Upper bound: $p(x) \leq Cq(x)$

Algorithm:

1. Sample $x_0 \sim q(x)$
2. Sample $\alpha \sim \mathcal{U}(0, Cq(x_0))$

*We thank Robbins family for supporting the Algorithmic Foundations of Data Science Course

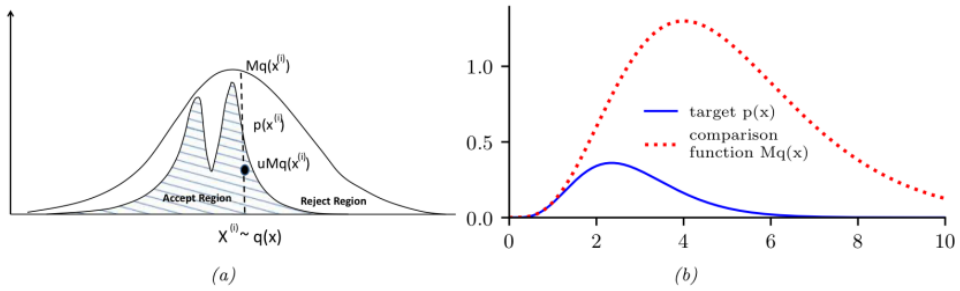

FIG. 1. *(a) Schematic illustration of rejection sampling. From Figure 2. Used with kind permission of Nando de Freitas. (b) Rejection sampling from a* $Ga(\alpha = 5.7, \lambda = 2)$ *distribution (solid blue) using a proposal of the form* $MGa(k, \lambda - 1)$ *(dotted red), where* $k = \lfloor 5.7 \rfloor = 5$*. The curves touch at* $\alpha - k = 0.7$*. Generated by* `rejection_sampling_demo.ipynb`*.*

3. If $p(x_0) \leq \alpha$, reject; otherwise accept.

Analysis:

$$q(\text{accept} \mid x_0) = \int_0^{\hat{p}(x_0)} \frac{1}{Cq(x_0)}\, du = \frac{\hat{p}(x_0)}{Cq(x_0)}$$

$$q(\text{propose and accept } x_0) = q(x_0) \cdot \frac{\hat{p}(x_0)}{Cq(x_0)} = \frac{\hat{p}(x_0)}{C}$$

Integrating both sides:

$$q(\text{accept}) = \int \frac{\hat{p}(x_0)}{C}\, dx_0 = \frac{Z_p}{C}$$

Therefore, the distribution of accepted samples is:

$$q(x_0 \mid \text{accept}) = \frac{q(x_0, \text{accept})}{q(\text{accept})} = \frac{\frac{\hat{p}(x_0)}{C}}{\frac{Z_p}{C}} = \frac{\hat{p}(x_0)}{Z_p} = p(x_0)$$

1.4. Importance Sampling. Goal: Estimate an expectation with respect to $\pi(x)$:

$$\mathbb{E}_\pi[\phi(x)] = \int \phi(x)\pi(x)\, dx = \int \phi(x)\frac{\pi(x)}{q(x)} q(x)\, dx$$

Approximation using samples $x_1, \ldots, x_{N_s} \sim q(x)$:

$$\mathbb{E}_\pi[\phi(x)] \approx \frac{1}{N_s} \sum_{n=1}^{N_s} \left(\frac{\pi(x_n)}{q(x_n)} \phi(x_n) \right) = \frac{1}{N_s} \sum_{n=1}^{N_s} \hat{w}_n \phi(x_n)$$

Where the importance weights are:

$$\hat{w}_n = \frac{\pi(x_n)}{q(x_n)}$$

2. Monte Carlo Method.

2.1. Monte Carlo Estimation. We wish to estimate an expectation with respect to a distribution μ:

$$\mathbb{E}[f] = \int f(x)\mu(x)\,dx = \sum_i f(i)\mu(i)$$

Let $x_1, x_2, \ldots, x_N \sim \mu$. Then the empirical estimate is:

$$\hat{\mathbb{E}}[f] = \frac{1}{N}\sum_{i=1}^{N} f(x_i)$$

We are interested in the concentration:

$$\Pr\left\{\left|\mathbb{E}[f] - \frac{1}{N}\sum_{i=1}^{N} f(x_i)\right| \geq t\right\}$$

This is equivalent to:

$$\Pr\left\{\left|\frac{1}{N}\sum_{i=1}^{N}(f(x_i) - \mathbb{E}[f])\right| \geq t\right\} \leq \frac{\sqrt{\mathbb{V}_{c_1}(f)}}{\sqrt{N}\cdot t}$$

The variance of the empirical mean is:

$$\mathbb{V}\left(\frac{1}{N}\sum_{i=1}^{N}(f(x_i) - \mathbb{E}[f])\right) = \frac{\mathbb{V}(f(x))}{N}$$

So the standard deviation is:

$$\sqrt{\mathbb{V}(\alpha)} = \frac{\sqrt{\mathbb{V}(f)}}{\sqrt{N}}$$

2.2. Hamiltonian Monte Carlo. Define the **Hamiltonian**:

$$H(\Theta, v) := E(\Theta) + K(v)$$

where:

- H is the total energy
- Θ is the position (parameter)
- v is the momentum
- $E(\Theta)$ is the potential energy (usually the negative log-density)
- $K(v)$ is the kinetic energy (often $\frac{1}{2}v^\top M^{-1}v$)

The dynamics are given by Hamilton's equations:

$$\frac{d\Theta}{dt} = \frac{\partial H}{\partial v} = \frac{\partial K}{\partial v} \tag{2.1}$$

$$\frac{dv}{dt} = -\frac{\partial H}{\partial \Theta} = -\frac{\partial E}{\partial \Theta} \tag{12.65}$$

To see why energy is conserved:

$$\frac{dH}{dt} = \sum_{i=1}^{D}\left(\frac{\partial H}{\partial \Theta_i}\cdot\frac{d\Theta_i}{dt} + \frac{\partial H}{\partial v_i}\cdot\frac{dv_i}{dt}\right) \tag{2.2}$$

$$= \sum_{i=1}^{D}\left(\frac{\partial H}{\partial \Theta_i}\cdot\frac{\partial H}{\partial v_i} - \frac{\partial H}{\partial \Theta_i}\cdot\frac{\partial H}{\partial v_i}\right) = 0 \tag{12.66}$$

Leapfrog Integrator. Let η be the step size. Then the leapfrog updates are:

$$v_{t+1/2} = v_t - \frac{\eta}{2} \cdot \frac{\partial E(\Theta_t)}{\partial \Theta} \tag{12.72}$$

(2.3)

$$\Theta_{t+1} = \Theta_t + \eta \cdot \frac{\partial K(v_{t+1/2})}{\partial v} \tag{12.73}$$

(2.4)

$$v_{t+1} = v_{t+1/2} - \frac{\eta}{2} \cdot \frac{\partial E(\Theta_{t+1})}{\partial \Theta} \tag{12.74}$$

Algorithm 2.1 Hamiltonian Monte Carlo (HMC)

Require: Initial state Θ_0, step size η, number of leapfrog steps L, covariance matrix Σ, potential energy function $E(\Theta)$

for $t = 1$ to T **do**
  Sample momentum $v_{t-1} \sim \mathcal{N}(0, \Sigma)$
  Initialize $(\Theta'_0, v'_0) \leftarrow (\Theta_{t-1}, v_{t-1})$
  $v'_{1/2} \leftarrow v'_0 - \frac{\eta}{2} \nabla E(\Theta'_0)$ // Half step for momentum
  for $\ell = 1$ to $L - 1$ **do**
    $\Theta'_\ell \leftarrow \Theta'_{\ell-1} + \eta \cdot \Sigma^{-1} v'_{\ell-1/2}$ // Full step for position
    $v'_{\ell+1/2} \leftarrow v'_{\ell-1/2} - \eta \nabla E(\Theta'_\ell)$ // Full step for momentum
  end for
  $\Theta'_L \leftarrow \Theta'_{L-1} + \eta \cdot \Sigma^{-1} v'_{L-1/2}$ // Final position update
  $v'_L \leftarrow v'_{L-1/2} - \frac{\eta}{2} \nabla E(\Theta'_L)$ // Final momentum update
  Set proposal $(\Theta^*, v^*) \leftarrow (\Theta'_L, v'_L)$
  Compute acceptance probability:

$$\alpha = \min\left(1, \exp\left[-H(\Theta^*, v^*) + H(\Theta_{t-1}, v_{t-1})\right]\right)$$

  Sample $u \sim \mathcal{U}(0, 1)$
  if $u < \alpha$ **then**
    $\Theta_t \leftarrow \Theta^*$
  else
    $\Theta_t \leftarrow \Theta_{t-1}$
  end if
end for

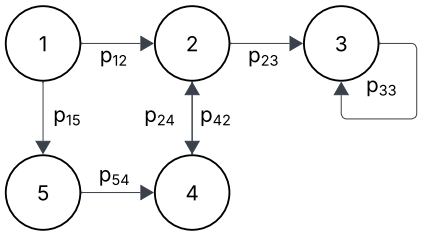

FIG. 2.

3. A Random Walk.

$$(1, 0, 0, 0, 0)$$

$$(0, p_{12}, 0, 0, p_{15})$$

$$p_{12} + p_{15} = 1, \quad p_{42} = 1$$

3.1. Markov Chain.

- A Markov chain consists of:
  - n states
  - $P \in \mathbb{R}^{n \times n}$: the transition matrix
- Example matrix:

$$P = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ p_{n1} & p_{n2} & \cdots & p_{nn} \end{bmatrix}$$

- Let $p(t) = (p_1(t), p_2(t), \ldots, p_n(t))$ be the distribution at time t.

$$p(t+1) = p(t)P$$

Specifically:

$$p_1(t+1) = p_1(t)p_{11} + p_2(t)p_{21} + \cdots + p_n(t)p_{n1}$$

- A graph is **strongly connected** if for all $x, y \in V$, there exists a path from x to y. This corresponds to a **connected Markov chain**.
- A state is **persistent** if, once visited, it is guaranteed to be revisited (with probability 1).

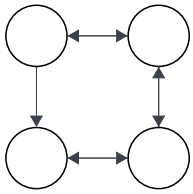

FIG. 3.

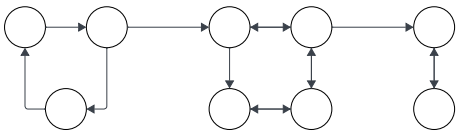

FIG. 4.

3.2. Stationary Distribution.

- A distribution $\pi \in \mathbb{R}^n$ is stationary if:

$$\pi P = \pi$$

- Define the time-average distribution:

$$a(t) = \frac{1}{t} \sum_{i=0}^{t-1} p(i)$$

- **Theorem:** For any connected Markov chain, the stationary distribution π is unique, and:

$$\lim_{t \to \infty} a(t) = \pi \quad \text{for any initial distribution } p(0)$$

- **Claim:** Let $B = [P - I \mid \mathbf{1}] \in \mathbb{R}^{n \times (n+1)}$, where:
  - I: identity matrix
  - $\mathbf{1}$: column vector of ones

  Then:

$$a(t)P = a(t) \Rightarrow a(t)(P - I) = 0$$

$$= a(t)P - a(t)$$

$$= \frac{1}{t} \left(\sum_{i=0}^{t-1} p(i)P - \sum_{i=0}^{t-1} p(i) \right)$$

$$= \frac{1}{t}(p(t) - p(0))$$

Therefore, in the limit as $t \to \infty$, $a(t)(P - I) \to 0$.

- To solve for π, solve:

$$\pi(P - I) = 0, \quad \pi\mathbf{1} = 1 \Rightarrow \pi[P - I \mid \mathbf{1}] = [0 \ \cdots \ 0 \ 1]$$

Let $B \in \mathbb{R}^{n\times(n+1)}$ be the matrix above, then:

$$\pi B = [0 \ \cdots \ 0 \ 1] \quad \text{and} \quad \pi = [0 \ \cdots \ 0 \ 1] \cdot B^*$$

where $B^* \in \mathbb{R}^{(n+1)\times n}$ is a right pseudo-inverse of B:

$$B^* = B^\top (BB^\top)^{-1}$$

3.3. Expected Return Times and Reversibility.

- **Theorem:** Let h_{ij} be the expected number of steps to go from state i to j. Then for the stationary distribution $\pi = (\pi_1, \ldots, \pi_n)$:

$$\pi_i = \frac{1}{h_{ii}}$$

- **Lemma:** Let $\pi \in \mathbb{R}^n$ with $\pi_i \geq 0$ and $\sum_{i=1}^n \pi_i = 1$. If for all i, j:

$$\pi_i p_{ij} = \pi_j p_{ji}$$

then π is the stationary distribution (this is called detailed balance).

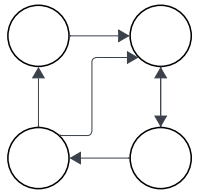

FIG. 5.

3.4. Markov Chain Monte Carlo (MCMC).

Let $w(0), w(1), \ldots, w(t+1)$ be a trajectory of a Markov chain.

- Monte Carlo estimate of $\mathbb{E}[f]$:

$$\frac{1}{t}\sum_{s=0}^{t-1} f(w(s))$$

- Define the time-average distribution:

$$a(t) = \frac{1}{t}\sum_{i=0}^{t-1} p(i) \quad \text{with} \quad a(t) = (a(t,1), a(t,2), \ldots, a(t,n))$$

- Time-averaged expectation:

$$\mathbb{E}_a[f] = \sum_{i=1}^{n} f(i)\cdot a(t,i) = \sum_i f(i)\cdot\left(\frac{1}{t}\sum_{j=0}^{t-1} p(j,i)\right)$$

- Stationary expectation:

$$\mathbb{E}_\pi[f] = \sum_i f(i)\cdot\pi(i)$$

- Total deviation:

$$|\mathbb{E}_a[f] - \mathbb{E}_\pi[f]| \le \sum_{i=1}^{n} |f(i)|\cdot|\pi(i) - a(t,i)|$$

$$\le \max_i |f(i)| \cdot \sum_{i=1}^{n} |\pi(i) - a(t,i)|$$

$$= \max(f)\cdot\|\pi - a(t)\|_1$$

Definition (Mixing Time): Let $\varepsilon > 0$. A time t is called an ε-mixing time if:

$$\|\pi - a(t)\|_1 \le \varepsilon$$

How do we determine ε-mixing time t?

- See: *The Nature of Computation* by Moore and Mertens [1]
- See: *Information, Physics, and Computation* by Mézard and Montanari [2]

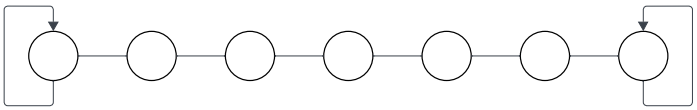

FIG. 6.

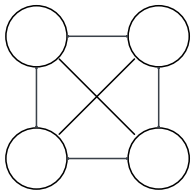

FIG. 7.

3.5. Conductance. Definition: Let $S \subseteq V = \{1, \ldots, n\}$. Define:

$$\pi(S) = \sum_{i \in S} \pi(i)$$

$$\tau(S,T) = \frac{\sum_{i \in S, j \in T} \pi_i P_{ij}}{\min(\pi(S), \pi(T))}$$

$$\tau(S) = \min_{T = V \setminus S} \tau(S,T)$$

$$\Phi := \min_{\substack{S \subseteq V \\ S \neq \emptyset,\, \pi(S) \leq 1/2}} \tau(S)$$

Theorem: For an undirected Markov chain with minimum stationary probability $\pi_{\min}$, the ε-mixing time satisfies:

$$t_{\text{mix}}(\varepsilon) = O\left(\frac{\log(1/\pi_{\min})}{\Phi^2 \varepsilon^3}\right)$$

Example:

3.6. Metropolis–Hastings Algorithm. Let $\pi = (\pi_1, \pi_2, \ldots, \pi_n)$ be the target distribution.

- Step 1: From current state i, propose a neighbor j with probability $1/r$.
- Step 2:
  - If $\pi_j \geq \pi_i$, move to j.
  - Otherwise, move to j with probability π_j/π_i, and stay in i with probability $1 - \pi_j/\pi_i$.
- Transition probabilities:

$$P_{ij} = \frac{1}{r} \min\left(1, \frac{\pi_j}{\pi_i}\right) \quad P_{ii} = 1 - \sum_{j \neq i} P_{ij}$$

- Satisfies detailed balance:

$$P_{ij}\pi_i = P_{ji}\pi_j$$

3.7. Gibbs Sampling.

[Figure 7: Gibbs sampling — placeholder]

Let the state be $x = (x_1, x_2, \ldots, x_n)$. Define a transition $x \to y$ where $y = (y_1, x_2, \ldots, x_n)$, i.e., only one component changes.

- Transition probability:

$$P_{xy} = \frac{1}{n} \cdot p(y_1 \mid x_2, \ldots, x_n)$$

- Using joint probabilities:

$$P_{xy} = \frac{1}{n} \cdot \frac{p(y_1, x_2, \ldots, x_n)}{p(x_2, \ldots, x_n)}$$

- Satisfies:

$$P_{xy}p(x) = \frac{1}{n} \cdot \frac{p(y_1, \ldots, x_n)}{p(x_2, \ldots, x_n)} \cdot p(x_1, \ldots, x_n) = \frac{1}{n} \cdot p(y) \cdot p(x_1 \mid x_2, \ldots, x_n)$$

- Therefore:

$$P_{xy}p(x) = P_{yx}p(y)$$

Exercises.

1. A Markov chain is said to be symmetric if for all i and j we have $p_{ij} = p_{ji}$. What is the stationary distribution in a symmetric Markov Chain? Prove your answer.
2. Suppose we have a multivariate probability distribution with density

$$p(x_1, x_2, \ldots, x_n) \quad \text{where } x_i \in \{0, 1\}.$$

   Please design an algorithm to estimate the marginal distribution on the first two coordinates:

$$p(t_1, t_2) = \sum_{x_3, x_4, \ldots, x_n} p(t_1, t_2, x_3, x_4, \ldots, x_n).$$

   Also design a variant of your algorithm for the case $x_i \in [0, 1]$.
3. We want to create a uniform sample from the set

$$\{1, 2, 3, \ldots, n\}.$$

   Design a Metropolis–Hastings algorithm for this, please.
4. Consider the vertices of the three-dimensional cube:

$$\{0, 1\}^3.$$

   We want to create a sample from the following distribution on these vertices:

$$p(x_1, x_2, x_3) = \begin{cases} 0 & \text{if } x_3 = 0, \\ \frac{1}{4} & \text{otherwise.} \end{cases}$$

   Create two connected graphs:
   - The first one should be very good for using Metropolis–Hastings to sample from p,
   - The second should be very bad for using Metropolis–Hastings to sample from p.

REFERENCES

[1] Cristopher Moore and Stephan Mertens. *The Nature of Computation*. Oxford University Press, 2011.

[2] Marc Mézard and Andrea Montanari. *Information, Physics, and Computation*. Oxford University Press, 2009.

Presentation follows this book:

[3] Avrim Blum, John Hopcroft, and Ravindran Kannan. *Foundations of Data Science*. Toyota Technological Institute at Chicago, Cornell University, and Microsoft Research India, 2020. https://www.cs.cornell.edu/jeh/book.pdf