Sampling Methods for Probability Density Estimation

Sree Rohith Pulipaka

April 2024

1 Introduction

How do we generate samples from a probability distribution whose analytical form is potentially unknown? As an example, this question surfaces in many cases in Posterior density approximation in Bayesian Inference.

Even for the cases in which the analytical form of the distribution is known- the exponential distribution, for example - how is the distribution sampled? The aim of this article Series is to provide an overview of three popular sampling methods for generating samples for known/unknown distributions:

- Inverse Sampling
- Accept-Reject Sampling
- Markov Chain Monte Carlo Methods

The reader should also note that to approximate a posterior distribution, another class of methods called Variational Inference also exists. Variational Auto-Encoders for example, fall under this category and unlike sampling, they estimate the posterior using a parametrized density function and find the parameters through optimization.

2 Markov Chain Monte Carlo (MCMC)

At the outset, the main idea of MCMC is to construct a Markov Chain such that its steady-state distribution is the probability distribution we would like to sample from. Sections 2.1 and 2.2 motivate this method listing a few theorems that guarantee this steady-state distribution. Section 2.3 applies the theory to a neat algorithm called as the Metropolis Algorithm that, under certain conditions, constructs a Markov chain for any steady-state distribution with the guarantee of its existence. A more general version called as the Metropolis-Hanstings algorithm is left for future discussion.

2.1 Markov Chains

A Markov chain is characterized by a set of n states $S = \{s_0, s_1, s_2 \dots s_n\}$ and transition probabilities taking us from one state to the other. We initially start at some state s_i and at each progressing time step, move from one state to the other based on the tranisition probabilities. That means we have a sequence of random variables denoted by S: $X_0, X_1, \dots X_\infty$ where the subscript of X denotes its time-step.

The transition probability $p_{ij} = P(X_{t+1} = s_j | X_t = s_i)$ denotes the probability of going to the state s_j in the next time step if we are in s_i in the current time step.

At any given time step we can define the state probability vector as the following column vector:

$$\mathbf{v_t} = [P(X_t = s_0) \quad P(X_t = s_1) \quad P(X_t = s_2) \quad \dots \quad P(X_t = s_n)]^T$$

This vector gives the probability distribution for the random variable X_t for all the states at time t.

So this random vector takes various values for each time step. It would an interesting situation if $\mathbf{v}_\infty$ was a constant vector. That would mean that eventually, this distribution is going to stabilize to some stationary value and that is the distribution of our interest. In the context of MCMC methods, we are going to construct a Markov chain whose stationary distribution is the probability distribution that we would like to sample from.
How do we find this distribution for a given Markov Chain? Is it unique? Does it even exist? Before we answer these questions, we define the transition probability matrix:

$$\mathbf{Pr} = \begin{bmatrix} p_{00} & p_{01} & \dots & p_{0n} \\ p_{10} & p_{11} & \dots & p_{1n} \\ \vdots & & & \vdots \\ p_{n0} & p_{n1} & \dots & p_{nn} \end{bmatrix}$$

The transition probabilities should add upto 1 and in $\mathbf{Pr}$, this means that each row sums up to 1. We define $\mathbf{P} = \mathbf{Pr}^T$ and in this matrix satisfies all the conditions of a Markov Matrix:

- Each element of the matrix ≥ 0
- Each column sums to 1.

We can then write the following equation based on law of total probability:

$$\mathbf{v_{t+1}} = \mathbf{P}\mathbf{v} \tag{1}$$

The steady state vector $\mathbf{v}_\infty$, if it exists satisfies the following equation:

$$\mathbf{v}_\infty = \mathbf{P}\mathbf{v}_\infty \tag{2}$$

The key caveat here is that such a vector needs to exist for $\mathbf{P}$ as $t \to \infty$ and does not depend on any starting distribution $\mathbf{v_0}$.

We can also define the notion of a stationary vector π:

$$\mathbf{P}\pi = \pi$$

It is important to note the difference between a steady-state distribution and stationary distribution: A stationary distribution is one, if the chain arrives at, stays the same forever. However, the chain may or may not arrive at it.
On the other hand, a steady-state distribution is the stationary distribution that a Markov chain is guaranteed to arrive at eventually, no matter what starting state we begin with.

From (2), we can see that $\mathbf{v}_\infty$ is the eigenvector of $\mathbf{P}$ with eigenvalue $\lambda = 1$. Does $\mathbf{P}$ have an eigenvalue of 1 in the first place? The answer is yes!

Theorem 2.1. *$\lambda = 1$ is the eigenvalue of any Markov Matrix $\mathbf{P}$*

Proof. The characteristic equation for $\lambda = 1$ is:

$$|\mathbf{P} - \lambda\mathbf{I}| = 0 \tag{3}$$

For $\lambda = 1$, we see that each column of $\mathbf{P} - \mathbf{I}$ sums to 0, instead of 1.That is, the sum of all rows is zero row, which means the rows are not linearly independent. Therefore, the matrix $\mathbf{P} - \mathbf{I}$ is singular and its determinant 0. □

What can we say about other values of λ in (3)?

Theorem 2.2. *Any eigenvalue λ of $\mathbf{P}$ satisfies $|\lambda| \leq 1$ i.e. 1 is the spectral radius of $\mathbf{P}$.*

Proof. Let $\mathbf{v} \in \mathbb{C}$ be the corresponding eigenvector of λ. Also, let j, k such that the magnitude of the j^{th} component of $\mathbf{v}$, $|\mathbf{v}_j| \leq |\mathbf{v}_k| \forall j \in \{1 \dots n\}$. Then,

$$\begin{aligned}
|\lambda\mathbf{v}_k| &= |[\mathbf{P}\mathbf{v}]_k| && (\lambda, \mathbf{v}) \text{ are eigen value-vector pair} \\
&= |\sum_j p_{kj}\mathbf{v}_j| \\
&\leq \sum_j p_{kj}|\mathbf{v}_j| \\
&\leq \sum_j p_{kj}|\mathbf{v}_k| \\
&= |\mathbf{v}_k|
\end{aligned}$$

$\Rightarrow |\lambda| \leq 1$ □

The role of these eigenvalues becomes clear when $\mathbf{P}$ is diagonalizable (i.e. it has all distinct eigenvalues or when geometric multiplicity of eigenvectors is same as algebraic multiplicity of the eigenvalue). In such a case $\mathbf{P} = \mathbf{S\Lambda S^{-1}}$. More explicityly, we can write this as :

$$\mathbf{P} = \underbrace{\begin{bmatrix} \vdots & \vdots & & \vdots \\ \mathbf{w_1} & \mathbf{w_2} & \dots & \mathbf{w_n} \\ \vdots & \vdots & & \vdots \end{bmatrix}}_{\mathbf{S}} \underbrace{\begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix}}_{\mathbf{\Lambda}} \underbrace{\begin{bmatrix} \dots & \mathbf{u_1^T} & \dots \\ \dots & \mathbf{u_2^T} & \dots \\ & \vdots & \\ \dots & \mathbf{u_n^T} & \dots \end{bmatrix}}_{\mathbf{S^{-1}}} \tag{4}$$

Here the set $\{\mathbf{w_1}, \dots, \mathbf{w_n}\}$ which are the eigenvectors of $\mathbf{P}$ are the column vectors of matrix $\mathbf{S}$ and $\{\mathbf{u_1^T}, \dots, \mathbf{u_n^T}\}$ are row vectors of matrix $\mathbf{S^{-1}}$ Note that these sets need not be orthonormal since that is possible only when $\mathbf{P}$ is symmetric. The set $\{\lambda_1, \dots, \lambda_n\}$ are the eigenvalues of $\mathbf{P}$ written in decreasing order of magnitude.
When we write (1) recursively starting with $\mathbf{v_0}$ at $t = 0$, we have :

$$\mathbf{v}^{t+1} = \mathbf{P}^t \mathbf{v_0}$$

The eigenvectors of $\mathbf{P}^t$ are the same as that of $\mathbf{P}$ and the eigenvalues are the set $\{\lambda_1^t, \dots \lambda_n^t\}$. Using this fact and expanding the equation in (4), we have $\mathbf{v^{t+1}}$ expanded with eigenvectors as basis:

$$\mathbf{v}^{t+1} = \lambda_1^t(\mathbf{u_1^T v_0})\mathbf{w_1} + \lambda_2^t(\mathbf{u_2^T v_0})\mathbf{w_2} + \dots + \lambda_n^t(\mathbf{u_n^T v_0})\mathbf{w_n} \tag{5}$$

As $t \to \infty$, λ's are exponentiated. This is where Theorem 2.2 converges the sum to a finite value with $|\lambda_i| \leq 1 \forall i \in \{1, \dots, n\}$. We however are not just interested in finite, we are also interested in a stationary and unique distribution in $\mathbf{v}^\infty$. Equation (5) however does not immediately guarantee that. If $|\lambda_j| < 1 \forall j \neq 1$, then we only have the first term of (5) giving us a unique (upto a scaling factor) steady state ($t \to \infty$) distribution.

However, if more than one eigenvalue has a magnitude equal to 1, we have a resulting linear combination whose coefficients are dependent on initial state and therefore we may not have a unique distribution at $t \to \infty$. Moreover, the distribution in such a case might also not be stationary if we have $\lambda = -1$. In such a case we may find a solution that oscillates based on if t is even or odd. For example the matrix $\mathbf{A} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ has eigenvalues $\lambda_1 = 1, \lambda_2 = -1$. The corresponding eigenvectors are $\mathbf{w_1} = [1, 1]^T$ and $\mathbf{w_2} = [-1, 1]^T$ and substituting these values in equation (5) shows us the non-stationary oscillatory behaviour.

For which $\mathbf{A}$, can we guarantee a unique steady state distribution? The following theorem lays a foundation for when we shall not have a repeated $\lambda = 1$:

Perron-Frobenius Theorem For a positive matrix $\mathbf{A} > 0$ i.e. all its entries are strictly greater than 0, there exists a positive eigenvalue r which is the spectral radius and whose algebraic and geometric multiplicity is 1. The corresponding eigenvector is positive.

Proof. **Todo** □

We apply this theorem to Markov Matrices, when $r = 1$.

2.2 The Aperiodic Irreducible case

When the transition matrix $\mathbf{P}$ has all positive entries, multiplication with itself will result in another positive matrix. Therefore, at any given time t, we have that $\mathbf{P}^t$ is a positive matrix and therefore has an algebraic and geometric multiplicity of 1 for $\lambda = 1$ from the theorem **Perron-Frobenius Theorem** .
Such a transition matrix would mean that each state in the Markov chain can transition to any other state with some probability *immediately*. A more general case would be that in the Markov chain, starting from a given state, we can transition to any other state *eventually* i.e at a future time-step which may not be the immediate next step. Such a chain is graph that is strongly connected i.e. there exists a directed path between any two given nodes. Such a Markov chain is called **irreducible**. Note that the transition matrix of such a chain may contain some entries which are zeroes. Formally, a Markov Chain is said to be irreducible if:

$$\sum_{t \geq 0} P(\mathbf{X}_t = j | \mathbf{X}_0 = i) > 0$$

Theorem 2.3. *For an irreducible Markov chain, there exists a unique stationary distribution.*

Proof. ***Todo*** □

We need another condition called **aperiodicity** for this unique stationary distribution to also be the steady state distribution. A markov chain is said to be periodic if a state can only be visited at timesteps which are not multiples of 1. So for it to be periodic, we have:

$$gcd(\{t | t \geq 0, P(\mathbf{X}_t = j | \mathbf{X}_0 = i) > 0\}) \neq 1$$

where *gcd is greatest common divisor* and if it doesn not statisfy the above condition, it is *aperiodic*.

Theorem 2.4. *For a finite state, irreducible, aperiodic Markov Chain, there exists a unique stationary distribution which is also the steady-state distribution.*

Proof. ***Todo*** □

Proof Sketch We require the additional aperiodicity condition to show that after a certain number of time steps n, the transition matrix $\mathbf{P}^n$ becomes strictly positive i.e. $\mathbf{P}^n > 0$. We can then use the *Perron-Frobenius Theorem* to complete the proof.

2.3 Metropolis Algorithm

So far, we have discussed what kind of Markov Chains guarantee a steady-state distribution. In this section, we discuss about an algorithm to explicitly construct a Markov chain whose steady-state distribution is of our interest. We assume to following about our target distribution π, with support S:

- We can compute the ratio $\frac{\pi(i)}{\pi(j)}$ for $i, j \in S$.
- S is finite.

The second condition is important since Theorem 2.4 requires the same. We shall construct a markov chain whose state space is given by the elements of S and whose transition probabilities of the matrix **Pr** (the rows of this matrix sum to 1) by:

$$Pr(i,j) = \begin{cases} 0 & \text{i, j are not neighbors} \\ \frac{1}{d}\min(1, \frac{\pi(j)}{\pi(i)}) & \text{i, j are neighbors} \\ 1 - \sum_{k \neq i} Pr(i,k) & i = j \end{cases} \tag{6}$$

Here d is any positive number that is strictly greater than the maximum out degree of the graph. We claim the following:

Theorem 2.5. *The Markov Chain with the transition probabilities defined in Equation* (6) *with a strongly connected state space given by the suport S is aperiodic, irreducible and has the stationary, steady state distribution π.*

Proof. If a chain has self-loops for all nodes, it is aperiodic. So we show that $Pr(i,i) > 0$. Let the i^{th} node have a degree d_i.

$$\begin{aligned} \sum_{k \neq i} Pr(i,k) &= \sum_{k \neq i} \frac{1}{d}\min(1, \frac{\pi(k)}{\pi(i)}) \\ &\leq \sum_{k \neq i} \frac{1}{d}.1 \\ &= \frac{d_i}{d} < 1 \\ \Rightarrow 1 - \sum_{k \neq i} Pr(i,k) &> 0 \end{aligned}$$

It is irreducible since the graph is strongly connected. It remains to show that π is indeed the stationary distribution of the chain.To show $\pi^T \mathbf{Pr} = \pi^T$, we define the following sets: $N^+(j) = \{i | \pi(i) \leq \pi(j)$, j is a neigbor of $j\}$ and $N^-(j) = \{i | \pi(i) > \pi(j)$, j is a neigbor of $j\}$. The j^{th} entry $(\pi^T P)_j$ is:

$$\sum_{l \in N^-(j)} \pi(l) Pr(l,j) + \sum_{i \in N^+(j)} \pi(i) Pr(i,j) + \pi(j)\big(1 - \sum_{r \neq j} Pr(j,r)\big)$$

Using the definition of the sets above and the definition of the transition matrix for each case, we have:

$$(\pi^T Pr)_j = \sum_{l \in N^-(j)} \pi(l) \frac{\pi(j)}{d.\pi(l)} + \sum_{i \in N^+(j)} \pi(i) \frac{1}{d} + \pi(j)\big(1 - \sum_{r \in N^-(j)} \frac{1}{d} - \sum_{r \in N^+(j)} \frac{\pi(r)}{d.\pi(j)}\big)$$

Cancelling the terms, we see:

$$(\pi^T Pr)_j = \pi(j)$$

which is the condition for stationarity. Since the stationary distribution is unique and steady state for the given chain, we prove the theorem. □