Sampling 2

A self-study metarials for PRML [1]

Jisung Lim[1]

[1] B.S. Candidate of Industrial Engineering
Yonsei University, South Korea.

24th June, 2017

Christopher M. Bishop. **Pattern Recognition and Machine Learning**. Springer, 2006.

Summary

1. Introduction
   - Metropolis Algorithm
2. Markov Chain
   - Markov Chain
   - Markov Chain for MCMC
3. Markov Chain Monte Carlo
   - MCMC: Introduction
   - MCMC Algorithms
   - MCMC: Metropolis-Hastings
   - MCMC: Gibbs Sampling
   - MCMC: Slice Sampling
4. Hybrid Monte Carlo Algorithm

Metropolis Algorithm

State:

- $\boldsymbol{z}^{(1)}, \boldsymbol{z}^{(2)}, \ldots$ forms Markov chain

$$\text{current state: } \boldsymbol{z}^{(\tau)} \tag{1}$$

Proposal distribution at τ:

- It should be simple enough to draw sample directly.
- Symmetry condition: $q(\boldsymbol{z}|\boldsymbol{z}') = q(\boldsymbol{z}'|\boldsymbol{z})$

$$\text{proposal distribution: } q(\boldsymbol{z}|\boldsymbol{z}^{(\tau)}) \tag{2}$$

True distribution:

- tractable up to its normalization term

$$\text{true distribution: } p(\boldsymbol{z}) = \tilde{p}(\boldsymbol{z})/Z_p \tag{3}$$

Metropolis Algorithm

Algorithm: At each cycle of the algorithm,

1. We generate a candidate sample $\boldsymbol{z}^{(*)}$ from the proposal distribution.
2. and then accept the sample according to an appropriate criterion.

Acceptance function:

$$A(\boldsymbol{z}^*, \boldsymbol{z}^{(\tau)}) = \min(1, a(\tau)) \quad \text{where} \quad a(\tau) = \frac{\tilde{p}(\boldsymbol{z}^*)}{\tilde{p}(\boldsymbol{z}^{(\tau)})} \tag{4}$$

Rejection criterion:

$$\begin{aligned} \text{Reject: } & A(\boldsymbol{z}^*, \boldsymbol{z}^{(\tau)}) \leq u(0,1) \\ \text{Accept: } & A(\boldsymbol{z}^*, \boldsymbol{z}^{(\tau)}) > u(0,1) \end{aligned} \tag{5}$$

where u is randomly picked from uniform distribution $U(0,1)$.

Markov Chain; Definition

Before discussing MCMC methods, it is fair to study some general and useful properties of Markov chains in some detail. We only deals with first-order Markov chain in this chapter which is defined by

Definition

A first-order Markov Chain is defined to be a series of **random variables**

$$\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(M)} \tag{6}$$

satisfying the Markov property given by

$$p(\boldsymbol{z}^{(m+1)}|\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(m)}) = p(\boldsymbol{z}^{(m+1)}|\boldsymbol{z}^{(m)})$$

which also can be represented as a directed graph in the form of a chain. And the finite-dimensional (here, it stands for time dimension) is given by

$$P(\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(M)}) \tag{7}$$

Markov Chain; Examining the state space

General definition of Stochastic process

$$X : (\Omega, \mathcal{F}) \to (S^{\infty}, \mathcal{E}^{\infty}) \tag{8}$$

where $(S, \mathcal{E})$ is state space on which $X^{(n)}$'s are defined as follows

$$X^{(n)} : (\Omega, \mathcal{F}) \to (S, \mathcal{E}) \tag{9}$$

The random variable from State space to Sampling space The stochastic process, or discrete-time Markov chain, have countable state space S. To simplify the notation and focus on the sample itself we introduce the random vector maps original state S to $\mathbf{z}$ which is equiprobable, D-dimensional space.

$$\mathbf{z}^{(n)} : (S, \mathcal{E}) \to (\mathbf{z}, \mathcal{A}) \tag{10}$$

where $\mathcal{A}$ is a σ-algebra of $\mathbf{z}$.

Markov Chain; Properties

- *Initial state*: $\boldsymbol{z}^{(0)}$
- *Transition probability*:

$$T_m(\boldsymbol{z}^{(m)}, \boldsymbol{z}^{(m+1)}) := p(\boldsymbol{z}^{(m+1)} | \boldsymbol{z}^{(m)}) \tag{11}$$

- *Time homogenuity*:

$$T_m(\boldsymbol{z}^{(m)}, \boldsymbol{z}^{(m+1)}) = T_n(\boldsymbol{z}^{(n)}, \boldsymbol{z}^{(n+1)}) \quad \forall m, n \tag{12}$$

- *Invariant distribution:*

$$p(\boldsymbol{z}) = \sum_{\boldsymbol{z}'} T(\boldsymbol{z}', \boldsymbol{z}) p(\boldsymbol{z}') \tag{13}$$

- *Detailed balance:*

$$p(\boldsymbol{z}) T(\boldsymbol{z}, \boldsymbol{z}') = p(\boldsymbol{z}') T(\boldsymbol{z}', \boldsymbol{z}) \tag{14}$$

Markov Chain; properties

Properties From the definition and the specification of the first-order Markov chain, we can examine some useful properties.

- *Reducibility*: A Markov chain is said to be irreducible if it is possible to get to any state from any state. The accessibility can be defined by

$$P(\boldsymbol{z}_n|\boldsymbol{z}_m) > 0 \quad \forall n, m \tag{15}$$

- *Periodicity*: A state n has period p if any return to state n must occur in multiples of p time steps. If $p = 1$, then the state is said to be *aperiodic*.
- *Recurrent*: A state n is said to be *transient* if, given that we start in state n, there is a non-zero probability that we will never return to n. State n is *recurrent* (or persistent) if it is not transient.
- *Ergodicity*: The Markov chain is called ergodic if it is irreducible, and its states are positive recurrent and aperiodic.

Markov Chain for MCMC

Our goal is still to generate a sample from $p(\boldsymbol{z})$.

- A standard Markov chain Monte Carlo approach is to construct an ***ergodic*** Markov chain with a stationary distribution.

The main properties underpinning the MCMC method, to accomplish our goal, are followings.

1. **Existence of invariant distribution:**

$$p^*(\boldsymbol{z}^{(\tau+1)}) = \sum_{\tau} T(\boldsymbol{z}^{(\tau)}, \boldsymbol{z}^{(\tau+1)}) p^*(\boldsymbol{z}^{(\tau)}) \quad \forall \tau \tag{16}$$

   - sufficient condition: detailed balance

2. **Uniqueness of invariant distribution:**

$$p(\boldsymbol{z}^{(m)}) \to p^*(\boldsymbol{z}) \quad \text{as} \quad m \to \infty \quad \forall p \tag{17}$$

   - sufficient condition: ergodicity

MCMC: Introduction

Motivation: Limitation of Basic Samplings

- In the previous section, we discussed the rejection sampling and importance sampling strategies for evaluating expectations of functions, and we saw that they suffer from severe limitations particularly in spaces of high dimensionality.

Intuition: Markov Chain Monte Carlo

- General strategy which allows sampling from a large class of distribution (based on the mechanism of Markov chains)
- It also uses a proposal distribution to generate samples from another distribution.
- Introduces state, denoted by τ, and remember the previous informration, a sample, $\boldsymbol{z}^{(\tau)}$
- The proposal distribution then depends on the current state: $q(\boldsymbol{z}|\boldsymbol{z}^{(\tau)})$

MCMC Algorithms

Goal and Requirement

- **Goal:** to generate a sample from $p(\boldsymbol{z})$
- **Requirement 1: invariant distribution**
  A markov chain where $p^*(\boldsymbol{z})$ is invariant.
- **Requirement 2: ergodicity property**
  With requirement 1, $p(\boldsymbol{z}^{(m)}) \to p^*(\boldsymbol{z})$ as $m \to \infty$

Sampling Step

1. Remembering the current sample $\boldsymbol{z}^{(\tau)}$, generate a candidate sample $\boldsymbol{z}^*$ from a proposal distribution $q(\boldsymbol{z}|\boldsymbol{z}^{(\tau)})$.
2. Accept the sample according to the criterion.
3. If the candidate sample is accepted, then $\boldsymbol{z}^{(\tau+1)} \leftarrow \boldsymbol{z}^*$, otherwise $\boldsymbol{z}^{(\tau+1)} \leftarrow \boldsymbol{z}^{(\tau)}$.

Note that

- The sequence of samples $\boldsymbol{z}^{(1)}, \boldsymbol{z}^{(1)}, \ldots$ are strongly dependent to previous sample due to its sequential sampling. We therefore chose only Mth samples to ensure the independence.

Metropolis-Hastings Algorithm

2 Conditions:

1. detailed balance (*Existence*)

$$p(\boldsymbol{z})T(\boldsymbol{z},\boldsymbol{z}') = p(\boldsymbol{z}')T(\boldsymbol{z}',\boldsymbol{z}) \tag{18}$$

2. ergodicity (*Uniqueness*)

Separate transition in 2 sub-steps:

$$T(\boldsymbol{z}^{(\tau)},\boldsymbol{z}^{(\tau+1)})\ \underbrace{q(\boldsymbol{z}^*|\boldsymbol{z}^{(\tau)})}_{\text{proposal dist.}}\ \underbrace{A_k(\boldsymbol{z}^*,\boldsymbol{z}^{(\tau)})}_{\text{acceptance dist.}} \tag{19}$$

Acceptance distribution:

$$A_k(\boldsymbol{z}^*,\boldsymbol{z}^{(\tau)}) = \min(1, a(\tau)) \quad \text{where} \quad a(\tau) = \frac{\tilde{p}(\boldsymbol{z}^*)q_k(\boldsymbol{z}^{(\tau)}|\boldsymbol{z}^*)}{\tilde{p}(\boldsymbol{z}^{(\tau)})q_k(\boldsymbol{z}^*|\boldsymbol{z}^{(\tau)})} \tag{20}$$

Here k labels the members of the set of possible transitions being considered.

Validation of limiting distribution

$$\begin{aligned} p(\boldsymbol{z})q_k(\boldsymbol{z}'|\boldsymbol{z})A_k(\boldsymbol{z}',\boldsymbol{z}) &= \min(p(\boldsymbol{z})q_k(\boldsymbol{z}'|\boldsymbol{z}), p(\boldsymbol{z}')q_k(\boldsymbol{z}|\boldsymbol{z}')) \\ &= \min(p(\boldsymbol{z}')q_k(\boldsymbol{z}|\boldsymbol{z}'), p(\boldsymbol{z})q_k(\boldsymbol{z}'|\boldsymbol{z})) \\ &= p(\boldsymbol{z}')q_k(\boldsymbol{z}|\boldsymbol{z}')A_k(\boldsymbol{z},\boldsymbol{z}') \end{aligned} \tag{21}$$

Metropolis-Hastings Algorithm

Step-by-step examination:

1. generate candidate sample $\boldsymbol{z}^*$ from proposal distribution
$$q(\boldsymbol{z}|\boldsymbol{z}^{(\tau)}) \tag{22}$$
2. calculate $a(\tau)$
$$a(\tau) = \frac{\tilde{p}(\boldsymbol{z}^*)q_k(\boldsymbol{z}^{(\tau)}|\boldsymbol{z}^*)}{\tilde{p}(\boldsymbol{z}^{(\tau)})q_k(\boldsymbol{z}^*|\boldsymbol{z}^{(\tau)})} \tag{23}$$
3. determine to accept the sample based on criterion
   - If $a \geq 1$,
$$\boldsymbol{z}^{(\tau+1)} = \boldsymbol{z}^* \tag{24}$$
   - Else,
$$\boldsymbol{z}^{(\tau+1)} = \begin{cases} \boldsymbol{z}^* & \text{with probability } a(\tau) \\ \boldsymbol{z}^{(\tau)} & \text{with probability } (1 - a(\tau)) \end{cases} \tag{25}$$

Metropolis-Hastings Algorithm

Gaussian proposal distribution:

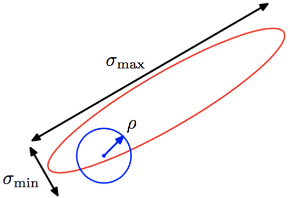
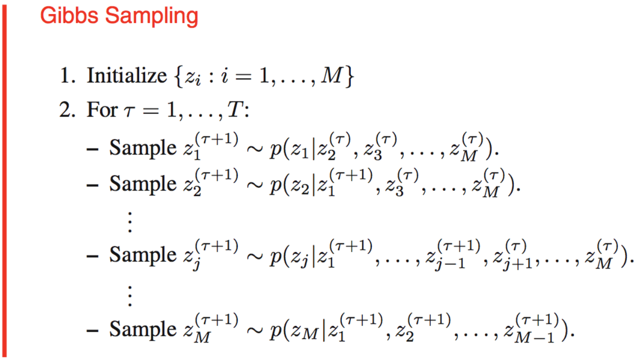


Figure 1: Gaussian proposal distribution.

- If the variance is small, then the acceptance rate will be high, but the progress will be slow.
- However, if the variance is large, then the rejection rate will be high.
- The scale ρ of the proposal distribution should be as large as possible without incurring high rejection rates.
- The number of steps needed to obtain independent samples will be of order

$$(\sigma_{\max}/\sigma_{\min})^2 \tag{26}$$

Gibbs Sampling



Gibbs Sampling

```
1. Initialize $\{z_i : i = 1, \ldots, M\}$
2. For $\tau = 1, \ldots, T$:
   – Sample $z_1^{(\tau+1)} \sim p(z_1 | z_2^{(\tau)}, z_3^{(\tau)}, \ldots, z_M^{(\tau)})$.
   – Sample $z_2^{(\tau+1)} \sim p(z_2 | z_1^{(\tau+1)}, z_3^{(\tau)}, \ldots, z_M^{(\tau)})$.
     $\vdots$
   – Sample $z_j^{(\tau+1)} \sim p(z_j | z_1^{(\tau+1)}, \ldots, z_{j-1}^{(\tau+1)}, z_{j+1}^{(\tau)}, \ldots, z_M^{(\tau)})$.
     $\vdots$
   – Sample $z_M^{(\tau+1)} \sim p(z_M | z_1^{(\tau+1)}, z_2^{(\tau+1)}, \ldots, z_{M-1}^{(\tau+1)})$.
```

Figure 2: Gibbs Sampling.

Gibbs Sampling

Acceptance function:

$$A_(\boldsymbol{z}^*, \boldsymbol{z}) = \frac{p(\boldsymbol{z}^*)q_k(\boldsymbol{z}|\boldsymbol{z}^*)}{p(\boldsymbol{z}^{(\tau)})q_k(\boldsymbol{z}^*|\boldsymbol{z})} = \frac{p(z_k^*|\boldsymbol{z}^*_{-k})p(\boldsymbol{z}^*_{-k})p(z_k|\boldsymbol{z}^*_{-k})}{p(z_k|\boldsymbol{z}_{-k})p(\boldsymbol{z}_{-k})p(z_k^*|\boldsymbol{z}_{-k})} \tag{27}$$

where $\boldsymbol{z}_{-k} = \boldsymbol{z}^*_{-k}$

Validation of equilibrium distribution p

$$p(z_i = z_i^{(\tau+1)}) = T(z_i = z_i^{(\tau)}, z_i = z_i^{(\tau+1)})p(z_i = z_i^{(\tau)}) \tag{28}$$

where $T = p(z_i|\boldsymbol{z}_{-i})$. From this fact,

$$A_(\boldsymbol{z}^*, \boldsymbol{z}) = 1 \tag{29}$$

i.e., the Metropholis-Hastings steps are always accepted.

Gibbs Sampling

Random walk:

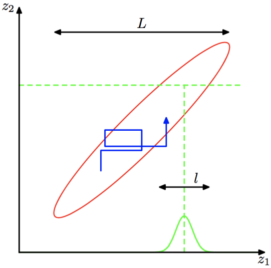


Figure 3: Gibbs random walk

- Because the state evolves according to a random walk, the number of steps needed to obtain independent samples from the distribution will be of order $(L/l)^2$.
- Reducing random walk: *Over-relaxation*

Slice Sampling

Motivation

- In Metropolis-Hastings Algorithm, If the step size is too small, the result is slow decorrelation due to random walk behaviour, whereas if it is too large the result is inefficiency due to a high rejection rate.

Concept

- adaptive step size that is automatically adjusted to match the characteristics of the distribution.

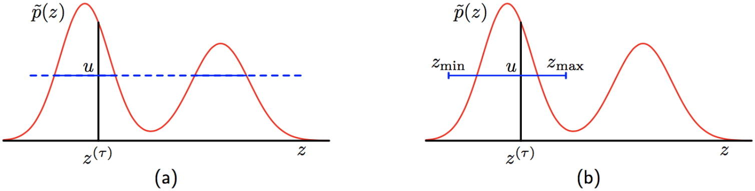


Figure 4: Slice Sampling.

Slice Sampling

Augmented space (z, u)

- Introduce additional variable u so as to augment z with u.
- Drawing samples from the joint (z, u) space.

$$\hat{p}(z,u) = \begin{cases} 1/Z_p & \text{if } 0 \le u \le \tilde{p}(z) \\ 0 & \text{otherwise.} \end{cases} \tag{30}$$

Marginal distribution over z:

$$\int \hat{p}(z,u)\mathrm{d}u = \int_0^{\tilde{p}(z)} \frac{1}{Z_p}\mathrm{d}u = \frac{\tilde{p}(z)}{Z_p} = p(z) \tag{31}$$

- We can sample from $p(z)$ by sampling from $\hat{p}(z,u)$ and then ignoring the u values.
- This can be achieved by alternately sampling z and u.

Alternate sampling:

1. Given z, evaluate $\tilde{p}(z)$.
2. draw u from $U(0, \tilde{p}(z))$.
3. fix u, sample z from $U(\{z : \tilde{p}(z) > u\})$
   (In practice, we can find the set $\{z\}$ by iterative extension.)

Dynamical systems

Hamiltonian function:

$$H(\boldsymbol{z}, \boldsymbol{r}) = E(\boldsymbol{z}) + K(\boldsymbol{r}) \tag{32}$$

- H is constant: $\frac{\mathrm{d}H}{\mathrm{d}r} = 0$

Flow field:

$$\mathbf{V} = \left(\frac{\mathrm{d}\mathbf{z}}{\mathrm{d}\tau}, \frac{\mathrm{d}\mathbf{r}}{\mathrm{d}\tau} \right) \tag{33}$$

- The volume is invariant: $\mathrm{div}\mathbf{V} = 0$

Joint distribution of $\mathbf{z}, \mathbf{r}$:

$$p(\mathbf{z}, \mathbf{r}) = \frac{1}{Z_H} \exp\left(-H(\boldsymbol{z}, \boldsymbol{r})\right) \tag{34}$$

- From the two results of conservation of volume and H, it follows that the Hamiltonian dynamics will leave $p(\mathbf{z}, \mathbf{r})$ ***invariant***.

Dynamical systems

- Evolution under the Hamiltonian dynamics will not, however, sample ergodically from $p(\mathbf{z}, \mathbf{r})$ because the value of H is constant.
- Noting that $\mathbf{z}$ and $\mathbf{r}$ are independent in the distribution and the conditional distribution $P(\mathbf{r}|\mathbf{z})$ is a Gaussian.