Sampling 1

A self-study metarials for PRML [1]

Jisung Lim[1]

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

8th February, 2017


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


Stephen P. Brooks. "Markov Chain Monte Carlo Method and Its Application". In: **Journal of the Royal Statistical Society. Series D (The Statistician)** 47.1 (1998), pp. 69–100. ISSN: 00390526, 14679884. URL: http://www.jstor.org/stable/2988428.

Summary

Bayesian and Intractability

Bayesian and Intractability

- **Yes, it is the Bayesian.**
  As we've seen in the previous chapters, the Bayesian approach have been more strongly focused than the frequentist veiwpoint, reflecting the huge growth in the practical importance of Bayesian methods.
- **Marginalization and Intractability**
  Full Bayesian approach requires intensive marginalization over the posterior distribution $p(\boldsymbol{\theta}|\mathcal{D})$ of parameters $\boldsymbol{\theta}$. For instance, when we solve regression problem with a full bayesian procedure, we encountered the equation such as

$$p(t|\boldsymbol{x}, \mathcal{D}) = \int_{\boldsymbol{\theta}} \int_{\boldsymbol{w}} p(t|\boldsymbol{x}, \boldsymbol{w}, \boldsymbol{\theta}) p(\boldsymbol{w}|\mathcal{D}, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathcal{D}) \mathrm{d}\boldsymbol{w} \mathrm{d}\boldsymbol{\theta}$$

  which is analytically intractable.

Intractability Relaxation

Two Main Paradigm of Intractability Relaxation

- **Intractability is around us.**
  Throughout other chapters, We've also seen this kind of intractable integrals in other chapters such as: 'Bayesian Logistic Regression', 'Bayesian Neural Network', 'Gaussian Process', 'Mixture Model', etc.
- **Sampling**
  The development of sampling methods, such as Markov chain Monte Carlo (discussed in Chapter 11) along with dramatic improvements in the speed and memory capacity of computers, opened the door to the practical use of Bayesian techniques in an impressive range of problem domains. Monte Carlo methods are very flexible and can be applied to a wide range of models. However, they are computationally intensive and have mainly been used for small-scale problems.
- **Variational Inference**
  More recently, highly efficient deterministic approximation schemes such as variational Bayes and expectation propagation (discussed in Chapter 10) have been developed. These offer a complementary alternative to sampling methods and have allowed Bayesian techniques to be used in large-scale applications.

Introduction to Sampling

Our Goal

In many cases, the posterior distribution is required primarily for the purpose of evaluating expectations. Hence our fundamental goal is to evaluate the expectation of some function $f(\boldsymbol{z})$ with respect to a probability distribution $p(\boldsymbol{z})$ i.e.,

$$\mathbb{E}[f] = \int f(\boldsymbol{z})p(\boldsymbol{z})\mathrm{d}\boldsymbol{z}, \tag{1}$$

or in the case of discrete variables,

$$\mathbb{E}[f] = \sum_{\boldsymbol{z}} f(\boldsymbol{z})p(\boldsymbol{z}). \tag{2}$$

Analytical intractability and approximation

Let us suppose that such expectations $\mathbb{E}[f]$ are too complex to be evaluated exactly using anlytical techniques.

For random variable $\mathbf{z}$, iid random samples $\{\mathbf{z}^{(l)}\}_{l=1}^{L}$ allow us to approximate the expectation by a statistic given by

$$\bar{f} = \frac{1}{L}\sum_{l=1}^{L} f(\mathbf{z}^{(l)}). \tag{3}$$

Introduction to Sampling

Nice Properties of Statistic $\bar{f}$

- $\bar{f}$ is unbiasedness estimator for $\mathbb{E}[f]$

$$\mathbb{E}[f] = \mathbb{E}[\bar{f}]$$

  i.e., the estimator $\bar{f}$ has the correct mean.

- The variance of $\bar{f}$ is given by

$$\mathrm{Var}[\bar{f}] = \frac{1}{L^2}(\mathrm{Var}[f(\boldsymbol{z}^{(1)})] + \cdots + \mathrm{Var}[f(\boldsymbol{z}^{(L)})]) = \frac{1}{L}\mathrm{Var}[f] = \frac{1}{L}\mathbb{E}[(f - \mathbb{E}[f])^2]$$

  i.e., the accuracy of the estimator does not depend on the dimensionality of $\boldsymbol{z}$ but depends on the number of samples L. That is, in principle, high accuracy may be achiveable with a relative small samples.

Common Problems

- The samples $\{\mathbf{z}^{(l)}\}_{l=1}^{L}$ might not be independent. That is, the effective sample size might be much smaller than the apparent sample size.
- Some functions f have large value with low p at some point $\boldsymbol{z}$, which can cause biased result evaluating expectation. That is, the approximation of expectations is undesirably dominated by some specific tendency.
- In practical, we therefore need relatively large sample size than we expected to achieve sufficient accuracy.

Ancestral Sampling

Sampling from the joint probability distribution

- The joint distribution can be specified by

$$p(\boldsymbol{z}) = \prod_{i=1}^{M} p(\boldsymbol{z}_i|\mathrm{pa}_i) \tag{4}$$

- Let the graph has partial order between parent(low) and child(high).
- Get sample $\hat{\boldsymbol{z}}_i$ from the pdf $p(\boldsymbol{z}_i|\mathrm{pa}_i)$ in topological order, from lowest to highest (i.e., parent to child).
- This is always possible because at each step all of the parent values will have been instantiated.

Logic Sampling (w/ evidence)

What if some data (evidence) is already taken?

- Now we consider the case of a directed graph in which some of the nodes are instantiated with observed values.
- The most simple, straight forward algorithm is *logic sampling*
- Do *Ancestral Sampling* repetitively until the sample satisfies the evidence.
- However, this algorithm is impractical because the sample is highly likely to be disagreed so as to be discarded.
- There are better algorithms such as *rejection sampling* and *adaptive rejection sampling* and *importance sampling*.

Logic Sampling (w/ evidence)

How to deal with indirected graph model?

- In the case of probability distributions defined by an undirected graph, there is no one-pass sampling strategy that will sample even from the prior distribution with no observed variables.
- Instead, computationally more expensive techniques must be employed, such as Gibbs sampling, which is discussed in Section 11.3.

How to deal with the sampling problem from marginal distribution

- As well as sampling from conditional distributions, we may also require samples from a marginal distribution.
- If we already have a strategy for sampling from a joint distribution $p(u, v)$, then it is straightforward to obtain samples from the marginal distribution $p(u)$ simply by ignoring the values for v in each sample.

Generating Random Samples

- **Goal:** How to generate random numbers from simple non-uniform distributions?

$$\text{random variable: } \mathrm{y} \tag{5}$$

- **Given:** A source of uniformly distributed random numbers and a function maps z to y.

$$\mathrm{z} \sim U(0,1) \quad \text{and} \quad \mathrm{y} = f(\mathrm{z}) \tag{6}$$

- **find** f**:** By change of variables

$$p(y) = p(z) \left| \frac{\mathrm{d}z}{\mathrm{d}y} \right| \tag{7}$$

and integrating each side

$$h(y) := \int_{-\infty}^{y} p(y)\, \mathrm{d}y = z \tag{8}$$

thus

$$y = h^{-1}(z) \quad \text{i.e.} \quad f = h^{-1} \tag{9}$$

Generating Random Samples

Generalization to Multiple Variables
The generalized form for multiple variables is started with

$$p(y_1, y_2, \ldots, y_M) = p(z_1, z_2, \ldots, z_M) \left| \frac{\partial(z_1, z_2, \ldots, z_M)}{\partial(y_1, y_2, \ldots, y_M)} \right| \tag{10}$$

cf) The Box-Muller method.

Limitation

- Obviously, the transformation technique depends for its success on the ability to calculate and then invert the indefinite integral of the required distribution.
- That is, such operations will only be feasible for a limited number of simple distributions.
- and so we must turn to alternative approaches in search of a more general strategy: **rejection sampling** and **importance sampling**.

Rejection Sampling

- **Goal:** Sample from relatively complex distribution, subject to certain constraints.
- **Given 1:** The distribution $p(z)$ is not a simple, standard distributions considered in the previous slides. That is, it is too complex to find

$$h(z) := \int_{-\infty}^{z} p(t)\,\mathrm{d}t = u \tag{11}$$

  inverse function of which is the function $f : u \to z$ where $u \sim U(0, 1)$.
- **Given 2:** We can easily evaluate evaluate $p(z)$ for any given z up to some normalizing constant Z_p as follows

$$p(z) = \frac{1}{Z_p}\tilde{p}(z) \tag{12}$$

  where $\tilde{p}(z)$ can readily be evaluated but Z_p is unknown.
- **Main concept:** The main concept of rejection sampling is using distribution cover $q(z)$ which is much simpler than $p(z)$. That envelope distribution is called 'proposal distribution'.

Rejection Sampling

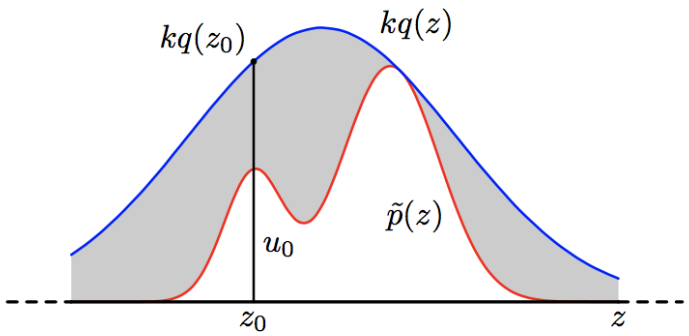

Figure 1: Rejection Sampling.

- **Proposal distribution $q(z)$:** We first introduce *proposal distribution*, from which we can readily draw samples.

$$q(z) \tag{13}$$

- **Bound Constant k:** We next introduce a constant k whose value is chosen such that

$$kq(z) \geq \tilde{p}(z) \quad \forall z \tag{14}$$

Rejection Sampling

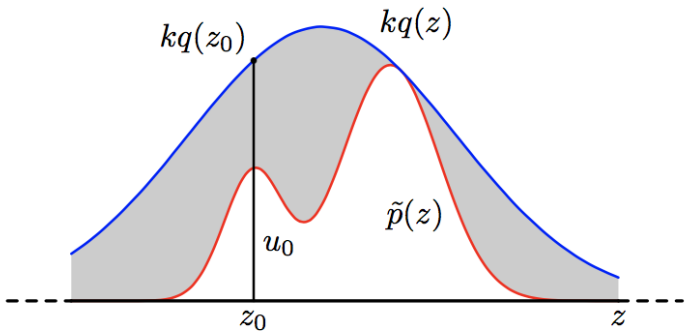

Figure 2: Rejection Sampling.

Drawing sample

1. Generate a number z_0 from the distribution $q(z)$.
2. Generate a number u_0 from the uniform distribution over $[0, kq(z_0)]$.
3. If $u_0 > \tilde{p}(z_0)$ then the sample is rejected, otherwise accepted.

The acceptance rate

$$\begin{aligned} p(\text{accept}) &= \int \{\tilde{p}(z)/kq(z)\}q(z)\mathrm{d}z \\ &= \frac{1}{k}\int \tilde{p}(z)\mathrm{d}z \end{aligned} \tag{15}$$

Adaptive Rejection Sampling

Motivation

- In many instances where we might wish to apply rejection sampling, it proves difficult to determine a suitable analytical form for the envelope distribution $q(z)$.
- How about constructing the envelope function $q(z)$ on the fly based on measured values of the distribution $p(z)$.

Settings

- $p(z)$ is log-concave.
- The function $\ln p(z)$ and its gradient are evaluated at some initial set of grid points, and the intersections of the resulting tangent lines are used to construct the envelope function.

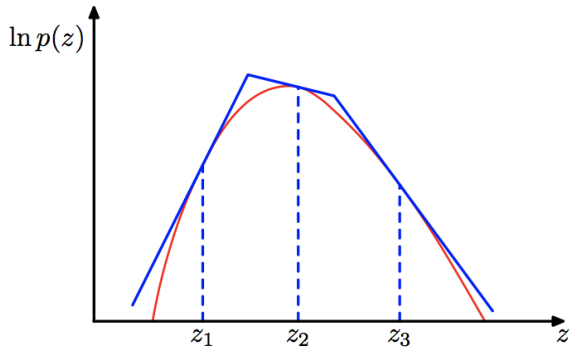

Figure 3: log envelope function.

Adaptive Rejection Sampling

envelope function

- Because the log envelope distribution is a succession of linear functions, and hence the envelope distribution itself comprises a piecewise, smooth exponential distribution of the form

$$q(z) = k_i \lambda_i \exp\{-\lambda_i(z - z_i)\} \quad \text{where} \quad \text{interval}(z_i) \tag{16}$$

Sampling step

1. A sample valuse is drawn from the envelope distribution, let the sample be drawn at z'.
2. If accepted, it will be a sample drawn from the desired distribution.
3. If the sample is rejected, a new line is evaluated at the corresponding position z', and the log envelope function (and eventually envelope function) is thereby refined.

Conclusion to Rejection Sampling

Conclusions

- The log concave constraint can be vanished simply by following each rejection sampling step with a Metropolis-Hastings step (section 11.2), giving rise to *adaptive rejection metropolis* sampling.
- That is, clearly, for rejection sampling to be of practical value, we required that the comparison function be close to the required distribution so that the rate of rejection is kept to minimum.

Limitations

- For more practical examples, where the desired distribution may be multimodal and sharply peaked, it will be extremely difficult to find a good proposal distribution and comparison function. (e.g. Gaussian cover)
- Furthermore, the exponential decrease of acceptance rate with dimensionality is a generic feature of rejection sampling.

Importance Sampling

Motivation

- Only wish to evaluate the expectation, not the samples itselves.
- Discretize whole z-space into a uniform grid and evaluate the integration. (c.f. Riemannian sum)

$$\mathbb{E}[f] \approx \sum_{l=1}^{L} p(\boldsymbol{z}^{(l)}) f(\boldsymbol{z}^{(l)}). \tag{17}$$

Concerns

- An obvious problem, however, is that the number of terms in the summation grows exponentially with the dimensionality of z.
- Only relatively small regions of z space has probability.
- So this first motivation is very inefficient.
- We therefore introduce a proposal distribution $q(z)$.

Importance Sampling

Approximation

- We can appoximate $\mathbb{E}[f]$ as follows

$$\begin{aligned}\mathbb{E}[f] &= \int f(\boldsymbol{z})p(\boldsymbol{z})\mathrm{d}\boldsymbol{z} = \int f(\boldsymbol{z})\frac{p(\boldsymbol{z})}{q(\boldsymbol{z})}q(\boldsymbol{z})\mathrm{d}\boldsymbol{z} \\ &\approx \frac{1}{L}\sum_{l=1}^{L}\frac{p(\boldsymbol{z}^{(l)})}{q(\boldsymbol{z}^{(l)})}f(\boldsymbol{z}^{(l)}) = \frac{1}{L}\sum_{l=1}^{L} r_l f(\boldsymbol{z}^{(l)})\end{aligned} \tag{18}$$

where $r_l = \frac{p(\boldsymbol{z}^{(l)})}{q(\boldsymbol{z}^{(l)})}$ are knwon as *importance weights*.

Normalzation constant

- In practical sense,

$$\begin{aligned}p(\boldsymbol{z}) &= \tilde{p}(\boldsymbol{z})/Z_p \\ q(\boldsymbol{z}) &= \tilde{q}(\boldsymbol{z})/Z_q\end{aligned} \tag{19}$$

where Z_p and Z_q are unknown.

- The approximation is then given by

$$\mathbb{E}[f] \approx \frac{Z_q}{Z_p}\frac{1}{L}\sum_{l=1}^{L}\frac{\tilde{p}(\boldsymbol{z}^{(l)})}{\tilde{q}(\boldsymbol{z}^{(l)})}f(\boldsymbol{z}^{(l)}) = \frac{Z_q}{Z_p}\frac{1}{L}\sum_{l=1}^{L}\tilde{r}_l f(\boldsymbol{z}^{(l)}) \tag{20}$$

Importance Sampling

- The ratio of normalization constants Z_p and Z_q can be approximated with same sample set

$$\frac{Z_p}{Z_q} = \frac{1}{Z_q}\int \tilde{p}(\boldsymbol{z})\mathrm{d}\boldsymbol{z} = \int \frac{\tilde{p}(\boldsymbol{z})}{\tilde{q}(\boldsymbol{z})} q(\boldsymbol{z})\mathrm{d}\boldsymbol{z} \approx \frac{1}{L}\sum_{l=1}^{L} \tilde{r}_l \tag{21}$$

- The approximation of $\mathbb{E}[f]$ is therefore given by

$$\mathbb{E}[f] \approx \sum_{l=1}^{L} w_l f(\boldsymbol{z}^{(l)}) \tag{22}$$

where

$$w_l = \frac{\tilde{r}_l}{\sum_m \tilde{r}_m} = \frac{\tilde{q}(\boldsymbol{z}^{(l)})/q(\boldsymbol{z}^{(l)})}{\sum_m \tilde{q}(\boldsymbol{z}^{(m)})/q(\boldsymbol{z}^{(m)})} \tag{23}$$

which implies that the success of the importance sampling approach depends crucially on how well the sampling distribution $q(\boldsymbol{z})$ matches the desired distribution $p(\boldsymbol{z})$.

Importance Sampling

Limitations and Hazards

- If, as is often the case,
  1. $p(z)f(z)$ is strongly varying i.e., $q(z)$ is not tight as much as we desired.
  2. Significant proportion of its mass concentrated over relatively small regions of z space.
  3. The set $\{r_l\}$ may be therefore dominated by a few weights having large values.
  4. Thus the effective sample size can be much smaller than the apparent sample size L.
- The extreme case is where none of the samples falls in the regions where $p(z)f(z)$ is large. Just consider the case where the function $p(z)f(z)$ is sharply peaked in some small regions of z space.

Conclusion

- Hence a major drawback of the importance sampling method is the potential to produce results that are arbitrarily in error and with no diagnostic indication.
- This also highlights a key requirement for the sampling distribution $q(z)$, namely that it should not be small or zero in regions where $p(z)$ may be significant.

Further study

- Likelihood weighted sampling
- Self-importance sampling

Sampling-importance-resampling

Motivation

- Importance sampling approach only gives approximation to expectation, not samples.
- Rejection sampling approach gives samples but it requires to determine the constant k with desirable acceptance rate.

Concepts: There are two stages to the scheme.

1. L samples $\boldsymbol{z}^{(l)}$ are drawn from $q(\boldsymbol{z})$.
2. w_l are constructed using (23).
3. A second set of L samples is drawn from the discrete distribution $(\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(L)})$ with probability given by the weights $(w_1, \ldots, w_L)$.

Intuition

- Let us consider the approximation

$$\int f(\boldsymbol{z}) \frac{p(\boldsymbol{z})}{q(\boldsymbol{z})} q(\boldsymbol{z}) \mathrm{d}\boldsymbol{z} \approx \frac{1}{L} \sum_{l=1}^{L} \frac{p(\boldsymbol{z}^{(l)})}{q(\boldsymbol{z}^{(l)})} f(\boldsymbol{z}^{(l)}) \approx \int f(\boldsymbol{z}) \frac{p(\boldsymbol{z})}{q(\boldsymbol{z})} \mathrm{d}\boldsymbol{z} \tag{24}$$

  where $r_l = \frac{p(\boldsymbol{z}^{(l)})}{q(\boldsymbol{z}^{(l)})}$.

- Consider then what is the role of w_l

$$w_l = \frac{\tilde{r}_l}{\sum_m \tilde{r}_m} \tag{25}$$

Monte Carlo EM algorithm

- Between E step and M step we must evalute the complete log likelihood given by

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = \int p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \ln p(\mathbf{Z}, \mathbf{X}|\boldsymbol{\theta}) \mathrm{d}\mathbf{Z} \tag{26}$$

  which can be approximated by sampling methods as follows

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = \frac{1}{L} \sum_{l=1}^{L} \ln p(\mathbf{Z}^{(l)}, \mathbf{X}|\boldsymbol{\theta}) \tag{27}$$

- This procedure is called Monte Carlo EM algorithm.
- A particular instance of the Monte Carlo EM algorithm, called stochastic EM, arises if we consider a finite mixture model, and draw just one sample at each E step.

IP algorithm

- When we move from a maximum likelihood approach to a full Bayesian treatment, marginalization is required over the posterior distribution $p(\boldsymbol{\theta}|\mathcal{D})$ of the parameter.
- Furthermore, we shall suppose that drawing samples from the posterior distribution $p(\boldsymbol{\theta}|\mathcal{D})$ is computationally difficult.
- **I-step:** For some case where there exists latent variable $\mathbf{Z}$ as follows

$$p(\mathbf{Z}|\mathcal{D}) = \int p(\mathbf{Z}|\boldsymbol{\theta}, \mathcal{D}) p(\boldsymbol{\theta}|\mathcal{D}) \mathrm{d}\boldsymbol{\theta}, \tag{28}$$

  let first draw a sample $\boldsymbol{\theta}^{(l)}$ from the current estimate for $p(\boldsymbol{\theta}|\mathcal{D})$, and then draw a sample $\mathbf{Z}^{(l)}$ from $p(\mathbf{Z}|\boldsymbol{\theta}^{(\mathbf{l})}, \mathcal{D})$.
- **P-step:** Using the sample $\mathbf{Z}^{(l)}$, approximate $p(\boldsymbol{\theta}|\mathcal{D})$ as follows

$$p(\boldsymbol{\theta}|\mathcal{D}) \approx \frac{1}{L} \sum_{l=1}^{L} p(\boldsymbol{\theta}|\mathbf{Z}^{(l)}, \mathcal{D}). \tag{29}$$

- Note that we are making a (somewhat artificial) distinction between parameters $\boldsymbol{\theta}$ and hidden variables $\mathbf{Z}$. From now on, we blur this distinction and focus simply on the problem of drawing samples from a given posterior distribution.

Summary and Limitations

- At the heart of the Bayesian inference is “marginalization.” Therfore, the ability to integrate analytically complex and high dimensional functions is extremely important in Bayesian statistics.
- Sampling methods allows us to sample from posterior distribution directly, and obtain sample estimates of the quantities of interests.
- With high dimensionality, however, the sampling methods what we’ve discussed so far still have severe limitations.
- e.g., guassian enveloping function in rejection sampling.
- e.g., Scattered clusters in importance sampling.

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**

$$\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(M)} \tag{30}$$

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

$$P(\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(M)}) \tag{31}$$

Markov Chain; Examining the state space

General definition of Stochastic process

$$X : (\Omega, \mathcal{F}) \rightarrow (S^{\infty}, \mathcal{E}^{\infty}) \tag{32}$$

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

$$X^{(n)} : (\Omega, \mathcal{F}) \rightarrow (S, \mathcal{E}) \tag{33}$$

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.

$$\mathbf{z}^{(n)} : (S, \mathcal{E}) \rightarrow (\mathbf{z}, \mathcal{A}) \tag{34}$$

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

Markov Chain; Properties

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

$$T_m(\boldsymbol{z}^{(m)}, \boldsymbol{z}^{(m+1)}) := p(\boldsymbol{z}^{(m+1)}|\boldsymbol{z}^{(m)}) \tag{35}$$

- *Time homogenuity*: Time-homogeneous Markov chains (or stationary Markov chains) are processes where

$$T_m(\boldsymbol{z}^{(m)}, \boldsymbol{z}^{(m+1)}) = T_n(\boldsymbol{z}^{(n)}, \boldsymbol{z}^{(n+1)}) \quad \forall m, n \tag{36}$$

- *Invariant distribution:*

$$p(\boldsymbol{z}) = \sum_{\boldsymbol{z}'} T(\boldsymbol{z}', \boldsymbol{z}) p(\boldsymbol{z}') \tag{37}$$

- *Detailed balance:*

$$p(\boldsymbol{z}) T(\boldsymbol{z}, \boldsymbol{z}') = p(\boldsymbol{z}') T(\boldsymbol{z}', \boldsymbol{z}) \tag{38}$$

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

$$P(\boldsymbol{z}_n|\boldsymbol{z}_m) > 0 \quad \forall n, m \tag{39}$$

- *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.

- *Time reversible property*

$$\sum_{\boldsymbol{z}'} p(\boldsymbol{z}')T(\boldsymbol{z}', \boldsymbol{z}) = \sum_{\boldsymbol{z}'} p(\boldsymbol{z})T(\boldsymbol{z}, \boldsymbol{z}') = p(\boldsymbol{z}) \sum_{\boldsymbol{z}'} T(\boldsymbol{z}, \boldsymbol{z}') = p(\boldsymbol{z}) \tag{40}$$

  where $p(\boldsymbol{z})$ is invariant distribution.

- *Existing and unique stationary distribution*
  Any chain which is *irreducible* and *aperiodic* will have a unique stationary distribution.

- *Ergodic property*

$$p(\boldsymbol{z}^{(m)}) \to p(\boldsymbol{z}) \quad \text{as} \quad m \to \infty \tag{41}$$

Markov Chain for MCMC[2]

1. A Markov chain with stationary distribution $p(\boldsymbol{z})$ is called *time reversible* (or simply *reversible*) if its transition kernel $T_m(\boldsymbol{z}^{(m)}, \boldsymbol{z}^{(m+1)})$ is such that it exhibits *detailed balance*, i.e.

$$p(\boldsymbol{z})T(\boldsymbol{z}, \boldsymbol{z}') = p(\boldsymbol{z}')T(\boldsymbol{z}', \boldsymbol{z}). \tag{42}$$

   This essentially means that the chain would look the same whether you ran it forwards in time or backwards. The behaviour of reversible chains is well understood, and it is a desirable property for any MCMC transition kernel to have, since any transition kernel for which *detailed balance* holds will have stationary distribution $p(\boldsymbol{z})$.

2. Any chain which is *irreducible* and *aperiodic* will have a unique stationary distribution, and that the t-step transition kernel will 'converge' to that stationary distribution as $t \to \infty$. See Meyn and Tweedie (1993).

3. A Markov chain is *ergodic* if, given realizations $\{\boldsymbol{z}^{(m)} : m = 0, 1, \ldots\}$ from the chain, typical asymptotic results include the distributional convergence of the realizations. That is, the distribution of the state of the chain at time m converges to the stational distribution $p(\boldsymbol{z})$ as $m \to \infty$.

$$p(\boldsymbol{z}^{(m)}) \to p(\boldsymbol{z}) \quad \text{as} \quad m \to \infty \tag{43}$$

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

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

Acceptance function:

$$A_k(\boldsymbol{z}^*, \boldsymbol{z}^{(\tau)}) = \min_k \left(1, \frac{\tilde{p}(\boldsymbol{z}^*) q_k(\boldsymbol{z}^{(\tau)}|\boldsymbol{z}^*)}{\tilde{p}(\boldsymbol{z}^{(\tau)}) q_k(\boldsymbol{z}^*|\boldsymbol{z}^{(\tau)})}\right) \tag{44}$$

Here k labels the members of the set of possible transitions being considered.
Rejection criteria: For $\boldsymbol{z}^*$ from the proposal distribution $q(\boldsymbol{z}^*|\boldsymbol{z}^{(\tau)})$, Consider the function $A_k(\boldsymbol{z}^*, \boldsymbol{z}^{(\tau)})$. Here the transition probabilit is given by

$$q(\boldsymbol{z}^*|\boldsymbol{z}^{(\tau)}) A_k(\boldsymbol{z}^*, \boldsymbol{z}^{(\tau)}) \tag{45}$$

Validation of limiting distribution

$$\begin{aligned} p(\boldsymbol{z}) q_k(\boldsymbol{z}'|\boldsymbol{z}) A_k(\boldsymbol{z}', \boldsymbol{z}) &= \min(p(\boldsymbol{z}) q_k(\boldsymbol{z}'|\boldsymbol{z}), p(\boldsymbol{z}') q_k(\boldsymbol{z}|\boldsymbol{z}')) \\ &= \min(p(\boldsymbol{z}') q_k(\boldsymbol{z}|\boldsymbol{z}'), p(\boldsymbol{z}) q_k(\boldsymbol{z}'|\boldsymbol{z})) \\ &= p(\boldsymbol{z}') q_k(\boldsymbol{z}|\boldsymbol{z}') A_k(\boldsymbol{z}, \boldsymbol{z}') \end{aligned} \tag{46}$$

Metropolis Algorithm

Symmetry specification: $q_k(\boldsymbol{z}|\boldsymbol{z}') = q_k(\boldsymbol{z}'|\boldsymbol{z})$ **Acceptance function:**

$$A(\boldsymbol{z}^*, \boldsymbol{z}^{(\tau)}) = \min\left(1, \frac{\tilde{p}(\boldsymbol{z}^*)}{\tilde{p}(\boldsymbol{z}^{(\tau)})}\right) \tag{47}$$

where $q(\boldsymbol{z}_A|\boldsymbol{z}_B) = q(\boldsymbol{z}_B|\boldsymbol{z}_A)$ for all A and B.

Rejection criterion

$$\begin{aligned} \text{Reject: } & A_(\boldsymbol{z}^*, \boldsymbol{z}^{(\tau)}) \leq u \\ \text{Accept: } & A_(\boldsymbol{z}^*, \boldsymbol{z}^{(\tau)}) > u \end{aligned} \tag{48}$$

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

$$\begin{aligned} p(\boldsymbol{z})q_k(\boldsymbol{z}'|\boldsymbol{z})A_k(\boldsymbol{z}', \boldsymbol{z}) &= q_k(\boldsymbol{z}'|\boldsymbol{z}) \min\left\{p(\boldsymbol{z}), p(\boldsymbol{z}')\right\} \\ &= q_k(\boldsymbol{z}|\boldsymbol{z}') \min\left\{p(\boldsymbol{z}'), p(\boldsymbol{z})\right\} \\ &= p(\boldsymbol{z}')q_k(\boldsymbol{z}|\boldsymbol{z}')A_k(\boldsymbol{z}, \boldsymbol{z}') \end{aligned} \tag{49}$$

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 4: Gibbs Sampling.

Gibbs Sampling

Acceptance function:

$$A_(\boldsymbol{z}^*, \boldsymbol{z}) = \frac{p(\boldsymbol{z}^*)q_k(\boldsymbol{z}|\boldsymbol{z}^*)}{p(\boldsymbol{z}^{(\tau)})q_k(\boldsymbol{z}^*|\boldsymbol{z})} = \frac{p(z_k^*|\boldsymbol{z}_{-k}^*)p(\boldsymbol{z}_{-k}^*)p(z_k|\boldsymbol{z}_{-k}^*)}{p(z_k|\boldsymbol{z}_{-k})p(\boldsymbol{z}_{-k})p(z_k^*|\boldsymbol{z}_{-k})} \tag{50}$$

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

Validation of equilibrium distribution p

$$p(z_i = z_i^{(\tau+1)}) = T(z_i = z_i^{(\tau)}, z_i = z_i^{(\tau+1)})p(z_i = z_i^{(\tau)}) \tag{51}$$

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

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.

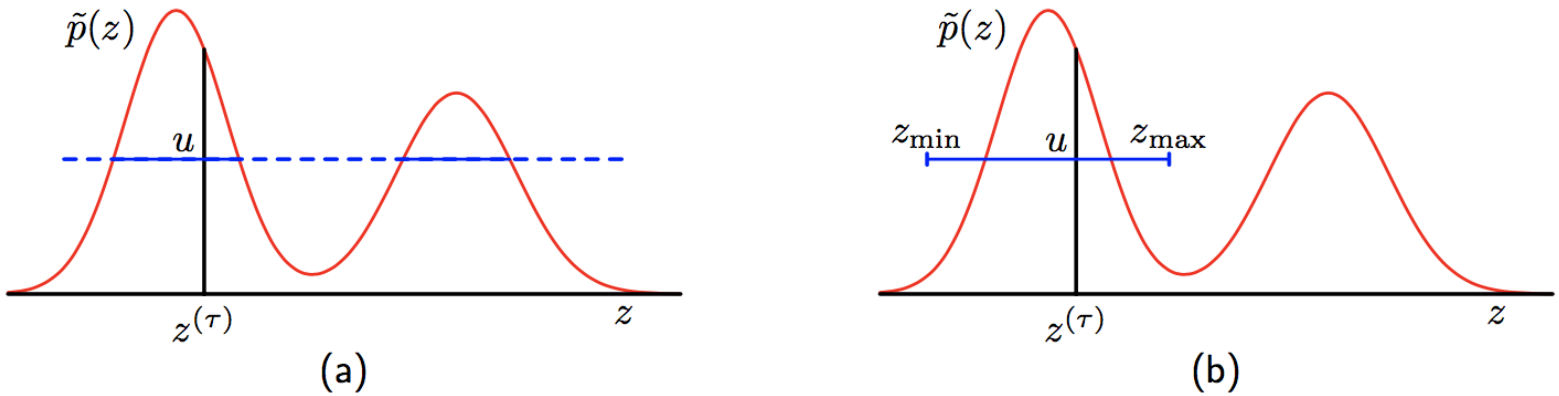

Figure 5: Slice Sampling.