CS5339 Project - Sampling Methods Monte Carlo Markov Chain

Sean Ng

15th March 21

1 Introduction

Modern history of Monte Carlo techniques began with Stan Ulam in 1946. Ulam was playing Solitaire and was trying to compute the probability that a randomly initialized solitaire will have a successful game. He found that it was easier to lay out the solitaire at random and count the plays that completed rather than to compute it directly. [1,4] *The idea is to use a statistical sample to approximate a hard combinatorial problem.*

These algorithms were the significant focus of statistical methods in the 1960s and these methods are considered to be some of the most general methods that can solve a wide array of high-dimensional problems. [2, 3, 5]

Before describing the algorithms, we will begin by describing the problems that Monte-Carlo Markov Chain (MCMC) methods aims to solve.

Here are common computations that are addressed by MCMC methods:

- **Normalization:** To obtain posterior given the prior and likelihood.

$$p(x|y) = \frac{p(y|x)p(x)}{\int_X p(y|x')p(x')dx'}$$

- **Marginalization:** To compute the marginal posterior given joint posterior of $(X, Z) \in \chi \times \zeta$
- **Expectation:** To obtain a summary statistics of the form

$$E_p(x|y)(f(x)) = \int_\chi f(x)p(x|y)dx$$

In each of these computations, integrating or summing across the entire space is computationally expensive especially in *higher dimensional datasets*.

In this report, we will use ideas such as dirac-delta expressions to describe basic sampling methods, before introducing Metropolis-Hastings and Gibbs algorithm. The sections of particular importance are the key ideas of MCMC in section 2.1 and the MCMC methods from section 4 onwards.

2 Key Ideas

We will first talk about the ideas behind estimating a value using a sample using two complementary ideas.

2.1 Law of Large Numbers

Assume that we would like to calculate the following value:

$E_\pi(f(x)) = \int_\chi f(x)\pi(x)dx$, where f : X $\rightarrow$ R^{n_f} and π is a probability distribution on $X \subset \mathbb{R}^{n_x}$, but it is difficult to obtain an analytical solution for it.

An estimator (S_N) that we would naturally think of is:

$$S_N(f) = \frac{1}{N}\sum_{i=1}^{N} f(X_i)$$

By *Law of Large Numbers*, we can show that the estimator is consistent. [1, 2]

$$\lim_{N\to\infty} \frac{1}{N} f(X_i) \rightarrow E_\pi(f(x))$$

We can also show that the rate of convergence is independent of the size of the dimension of the data.

$$var_\pi[S_N(f)] = var_\pi[\frac{1}{N}\sum_{i=1}^{N} f(X_i)] = \frac{var_\pi[f(X)]}{N}$$

The rate of convergence of the estimator when considering the mean squared error, since it is an unbiased estimator, is proportional to the variance of the sample mean. The rate of convergence is thus independent of the dimension of the data n_x. This would not be the case if we used a deterministic method where we evaluate the integral over a grid of regularly spaced points. Thus, MCMC methods are favoured when n_x is large.

2.2 Description using Sampling Representation

We will introduce the sampling representation to provide an alternative view of the sampling problem that is different from the way that it is usually presented. [2]

We first introduce the idea of a dirac-delta function, that applies for any $f : \chi \rightarrow \mathbb{R}$

$$\int_\chi f(x)\delta_{x_0}(x) = f(x_0)$$

Hence,

$$\int_\chi I_A(x)\delta_{x_0}(x) = \int_A f(x)\delta_{x_0}(x) = I_A(x_0)$$

Consider $A \subset \chi$, where A is the set of elements of interest. Assume that we are interested in computing statistics related to the set A.

$$\pi(A) \simeq \frac{number\ of\ samples\ in\ A}{total\ number\ of\ samples}$$

Consider the following mixture of dirac-delta functions:

$$\hat{\pi}_N(x) = \frac{1}{N}\sum_{i=1}^{N}\delta_{X_i}(x)$$

This equation can be seen as the empirical measurement of the sample, where the concentration of these realisations in the space now represents the distribution π.

When returning back to the problem of estimating $E_\pi(f)$, we can replace $\pi(x)$ with the sample representation $\hat{\pi}_N(x)$ to show that $E_\pi(f)$ can be estimated using a sample mean.

$$E_\pi(f) \simeq \int_\chi f(x)\frac{1}{N}\sum_{i=1}^{N}\delta_{X_i}(x)dx = \sum_{i=1}^{N}\frac{1}{N}\int_\chi f(x)\delta_{X_i}dx = \frac{1}{N}\sum_{i=1}^{N}f(X_i)$$

3 Basic Sampling Algorithms

We will first introduce some basic sampling algorithms.

Suppose that we would like to sample n samples from a target distribution. *The fundamental idea is that from a computational perspective, anything except a uniform distribution is not easily generated.*

An exact simulation method is the inverse cdf method, which is to uniformly sample a value, y, from 0 to 1, and mapping this value y to the corresponding sample value, x, where cdf(x) = y, and take x as the sample. However, this requires one to compute the inverse cdf which does not readily exist for many distributions.

3.1 Rejection Sampling

An intuitive way that works is to use another distribution that is easy to sample from, and accept if it falls within an acceptance probability. [1, 2]

3.1.1 Algorithm:

Let the target distribution be $\pi(\theta)$ and let the proposed distribution(the distribution that we sample from) be $q(\theta)$.

1. For i till n:
   (a) Propose a candidate $\theta \sim q(\theta)$
   (b) Sample uniformly from 0 up to the maximum $q(\theta)$ value, assign that value as u
   (c) Accept θ as an element in the sample if $u < \pi(\theta)$ else repeat from step (a)

Good: The proposal is universal. Unlike inverse cdf, it does not rely on algebraic properties.
Bad: The sampling might be inefficient.

3.2 Importance Sampling

Importance sampling [1,2] evolved from the idea that instead of taking a simple sample mean, we can use a weighted sample mean where the weight is proportional to $\pi(x)$ where π is the target distribution. Similar to rejection sampling, we will use another distribution, q is easier to sample from. [2]

Consider a probability distribution, q, such that $\pi(x) > 0 \rightarrow q(x) > 0$. We can define an importance weight $w_{X_i} = \frac{\pi(x)}{q(x)}$

$$E_\pi(f(x)) = \int_\chi f(x)\pi(x)dx = \int_\chi f(x)\frac{\pi(x)}{q(x)}q(x)dx = E_q(w(x)f(x))$$

We can introduce the follow dirac-delta functions.

$$\hat{\pi}_N(x) = \frac{1}{N}\sum_{i=1}^{N} w_{X_i}\delta_{X_i}(x)$$

Sampling representation gives us an interpretable result. Large w_i indicates under-representation of π by samples of q around X_i, small w_i indicates over-representation of π by samples of q around X_i.

Claim 1. *$\pi(z)$ and $q(z)$ can be evaluated up to a normalization constant (i.e. use $\tilde{\pi}(x)$ where $\tilde{\pi}(x)$ is proportional to $\pi(x)$).*

Proof. Proof in appendix in section B. □

Benefits: Good if proposal distribution is selected well. Intuitively it is a good solution when trying to sample from a subset of a distribution that is small (tail distribution for example).

Drawbacks: Effectiveness depends on choice of distribution.

4 Monte Carlo Markov Chains

The appendix section A contains a summary of key results in Markov chains, which is necessary to understand the results in this section.

4.1 Metropolis-Hastings

When discussing about bayesian inference, we consider the generation of samples from distribution π, such that $P(x) = \frac{f(x)}{Z}, f : \chi \geqslant 0$. f can be seen as a score of the desirability to be in that state.

Metropolis-Hastings allows us to sample from π without having to compute Z, through a Markov Chain that has stationary distribution π. Instead, the sampling is facilitated through an alternative distribution q.

4.1.1 Algorithm

1. Initialization $i = 0$. Set randomly or deterministically θ_0.
2. For i till n:
   (a) Propose a candidate $\theta \sim q(\theta)$

(b) Evaluate acceptance probability

$$\alpha(\theta^{i-1}, \theta) = \min\left(1, \frac{\pi(\theta)/q(\theta^{i-1}, \theta)}{\pi(\theta^{i-1})/q(\theta, \theta^{i-1})}\right)$$

(c) Set $\theta' = \theta$ with probability $\alpha(\theta^{i-1}, \theta)$ otherwise $\theta = \theta^{i-1}$

4.1.2 Proof

Claim 2. *Stationary distribution of Metropolis-Hastings Markov chain matrix is the target distribution.*

Let's define a transition matrix P:

$$P_{ij} = P(X_n = j | X_{n-1} = i) = \begin{cases} q(i,j)\alpha(i,j) & \text{if } j \neq i \\ q(i,i) + \sum_{k \neq i} q(i,k)(1 - \alpha(i,k)) & \text{otherwise} \end{cases}$$

Proof. Using the detailed balanced equation, we can show that the target distribution is the stationary distribution.

$$\begin{aligned} \pi(x)P(x,y) &= \pi(x)\alpha(x,y)q(x,y) \\ &= \min\left(1, \frac{\pi(y)q(y,x)}{\pi(x)q(x,y)})\pi(x)q(x,y)\right) \\ &= \min(\pi(x)q(x,y), \pi(y)q(y,x)) \\ &= \pi(y)q(y,x)\min(\pi(x)q(x,y), 1) \\ &= \pi(y)q(y,x)\alpha(y,x) \\ &= \pi(y)P(y,x) \end{aligned}$$

By lemma A.1, since the detailed balance equation is fulfilled, $\pi(x)$ is the stationary distribution. □

Claim 3. *The matrix P converges to the stationary distribution if q(x,y) is continuous and strictly positive on the support of π*

Proof. If q(x,y) is continuous and strictly positive on the support of π, alpha(x, y) is greater than 0. Hence, every state is reachable from every other state, and the graph is irreducible.

Aperiodicity is shown since π is positive for every x.

Hence, by Fundamental Theorem of Markov Chain (theorem 1 in appendix), the Markov chain converges to the stationary distribution. □

4.1.3 Analysis

The choice of the proposal distribution q has significant effects on the performance of the algorithm. For continuous state spaces, having a Gaussian centered at a state. If the variance is low, proportion of accepted transitions will be high but this would lead to long correlation times. If variance is high, the opposite will happen.

The number of steps to get independent samples is shown to be of the order $(\sigma_{max}/\sigma_{min})^2$. [1]

4.2 Gibbs Sampling

Gibbs sampling [1,2] is a simple and widely applicable MCMC algorithm and it can be seen as a special case of the Metropolis-Hastings algorithm.

Intuitively, each step of the Gibbs sampling is constrained to updating a single variable conditioned on the values of the remaining variables (i.e. $p(z_d|z_{\setminus d})$)

We define new notations:

$p(\theta_d|\theta_{\setminus d})$, where θ_d denotes the d^{th} component of θ, and $z_{\setminus d}$ denotes the set of components that is not the d^{th} component of θ.

4.2.1 Algorithm

1. Initialization $i = 0$. Set randomly or deterministic-ally θ_0.
2. For i till n:
   (a) Pick index d uniformly and at random from 1,...,k
   (b) Propose a candidate $\theta \sim q(\theta'_d, \theta^{i-1}_{\setminus d})$ where $\theta^{i-1}_{\setminus d}$ is the set of all variables in θ^{i-1} except for the d^{th} variable.
   (c) Set $\theta^i = (\theta^{i-1}_1, \theta^{i-1}_2, \theta^{i-1}_3, ..., \theta'_d, ..., \theta^{i-1}_k)$

Gibbs is specialized case of the Metropolis-Hastings algorithm where the acceptance probability is always 1. The proposal distribution for $j = 1, ..., N$:

$$q(\theta, \theta') = \begin{cases} \pi(\theta'_d|\theta_{\setminus d}) & \text{if } \theta_{\setminus d} = \theta_{\setminus d} \\ 0 & \text{otherwise} \end{cases}$$

4.2.2 Proofs

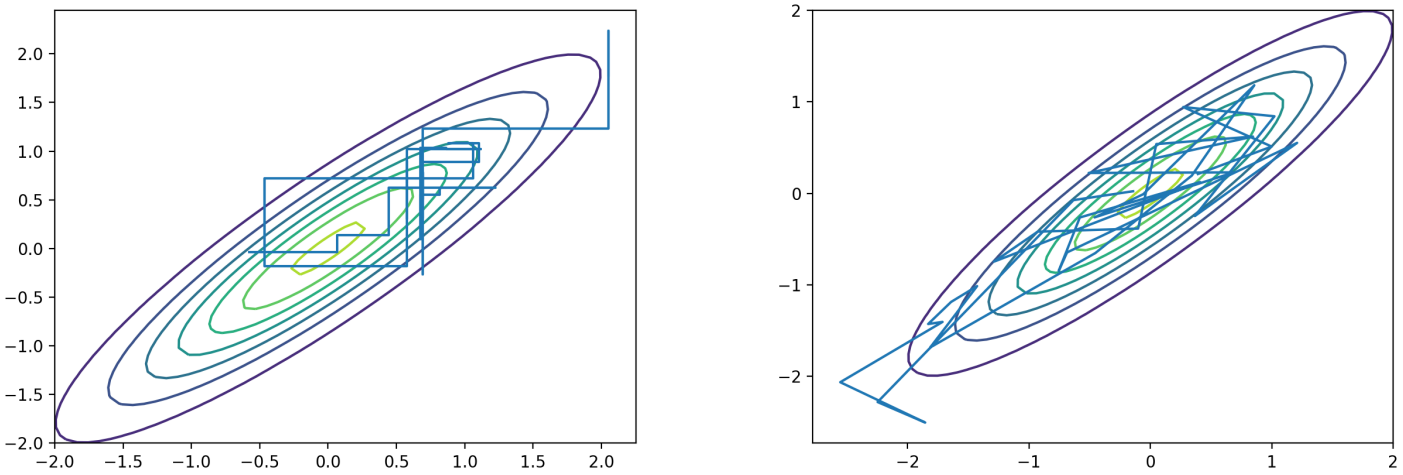
Claim 4. *Gibbs is a specialized Metropolis-Hastings where the acceptance probability evaluates to 1.*

Proof. The target distribution function can be expressed as follows:

$$\pi(\theta) = \pi(\theta_d, \theta_{\setminus d}) = \pi(\theta_d|\theta_{\setminus d})\pi(\theta_{\setminus d})$$

$$\begin{aligned} \alpha(\theta^{i-1}, \theta) &= \min\Big(1, \frac{\pi(\theta)/q(\theta^{i-1}, \theta)}{\pi(\theta^{i-1})/q(\theta, \theta^{i-1})}\Big) \\ &= \min\Big(1, \frac{\pi(\theta)q(\theta, \theta^{i-1})}{\pi(\theta^{i-1})q(\theta^{i-1}, \theta)}\Big) \\ &= \min\Big(1, \frac{\pi(\theta_d|\theta_{\setminus d})\pi(\theta_{\setminus d})\pi(\theta^{i-1}_d|\theta_{\setminus d})}{\pi(\theta^{i-1}_d|\theta^{i-1}_{\setminus d})\pi(\theta^{i-1}_{\setminus d})\pi(\theta_d|\theta^{i-1}_{\setminus d})}\Big) \\ &= 1[Since \pi(\theta_{\setminus d}) = \pi(\theta^{i-1}_{\setminus d})] \end{aligned}$$

□



(a) Results of Gibbs sampling

(b) Results of Metropolis-Hastings sampling

Figure 1: Samples retrieved from Gibbs sampling vs Metropolis-Hastings on a highly correlated multinomial distribution. (300 iterations, 100 iterations burn-in)

Benefits: It is best used with exponential families of distributions, when it is clear that the conditional distribution exists.

Drawbacks: Gibbs sampling has a low speed of convergence when the different dimensions are highly correlated. Intuitively, it is because it is constrained by the algebraic properties of the update. We will demonstrate this in the experimental section 5.1.

5 Experimental Examples

5.1 Metropolis-Hastings vs Gibbs experiment

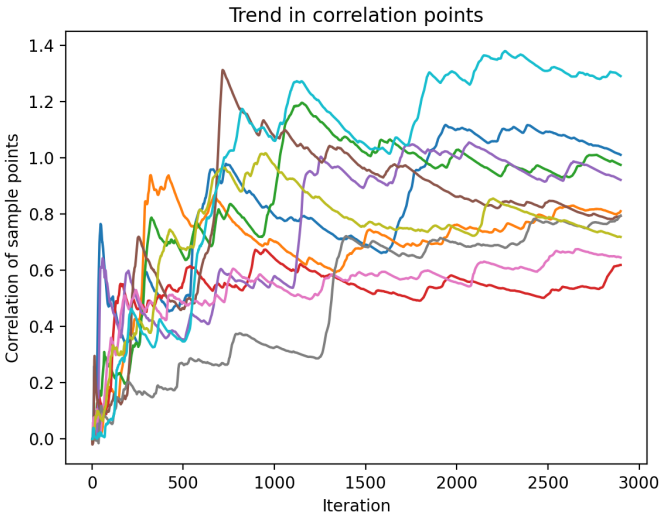
We applied a simple Metropolis-Hastings and Gibbs using a gaussian transition function. To illustrate the difference between Gibbs sampling and Metropolis-Hastings, we chose to model a highly correlated multinomial normal distribution of mean (0,0) and a covariance matrix of [[1, 0.9], [0.9,1]].

In figure 1, Metropolis-Hastings better explored the space as compared to Gibbs sampling. Both experiments were ran for 300 iterations with the first 100 iterations as the burn-in period(results discarded).

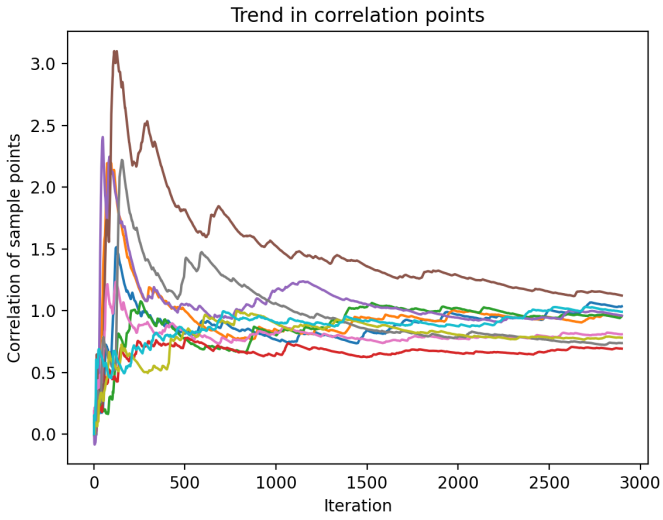
In figure 2, it is clear that Metropolis-Hastings modelled the distribution better than Gibbs as it is closer to the true value of 0.9 with smaller number of iterations. We can see the convergence of the line towards 0.9 correlation. The thick pencil pattern is an indication that the algorithm converges almost immediately. [6]

5.2 Simulated Annealing Experiment

Simulated annealing [5] is a type of MCMC that is based on the the annealing of metals, where the temperature is control its ductility and hardness. Simulated annealing can be considered an adaptation of the



(a) Results of Gibbs sampling



(b) Results of Metropolis-Hastings sampling

Figure 2: Correlation of samples retrieved from Gibbs sampling vs Metropolis-Hastings on a highly correlated multinomial distribution. (3000 iterations, 100 iterations burn-in). The original correlation is 0.9.

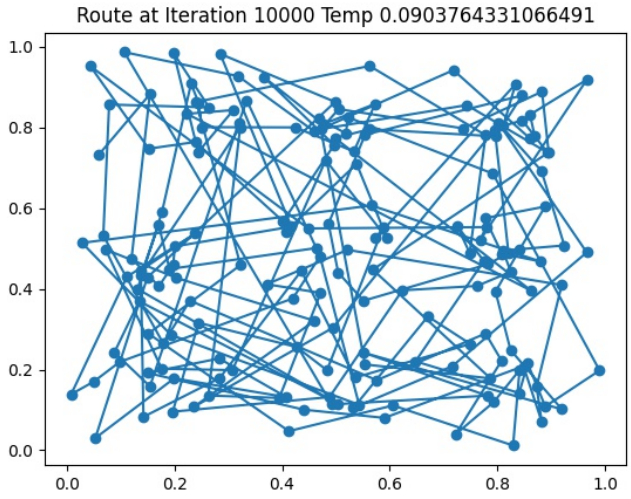
Metropolis-Hastings algorithm.

We applied Simulated Annealing to approximately come up with a better improvement to the greedy solution of the iconic travelling salesman problem(TSP). We referred to ideas presented in the original paper by Kirkpatric. [3] We used a simple temperature schedule that decreases the temperature by a factor of $\alpha = 0.95$. A "neighbour" of a state is generated by inverting a random subsection of the route.

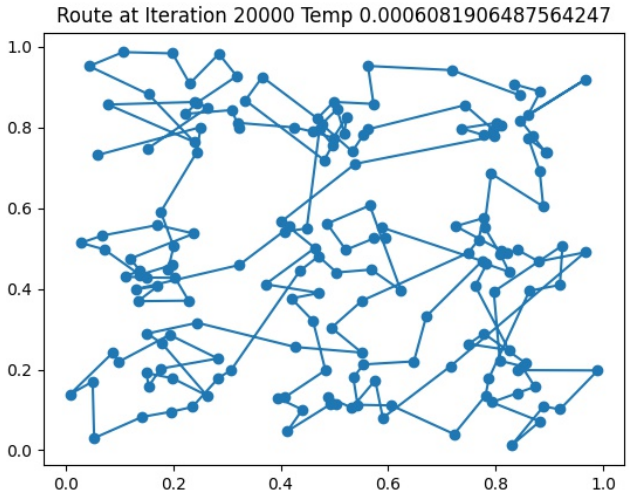
As with normal cities, we would expect there to be regions of high density and regions of low density. We simulated this by dividing into 9 areas and using a normal distribution that's centered at the center of each area.

We were able to improve on the greedy solution to the TSP problem by 17% by using Simulated Annealing.

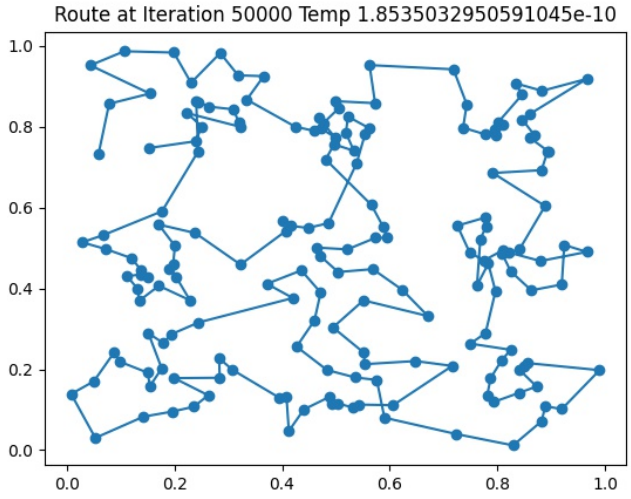
Figure 3 shows the intermediate results at different temperatures. As noted by Kirkpatric, the connection with statistical mechanics offers an interesting perspective . As the "temperature" of the system decreases, we can see the coarse structure of the system refining itself before the smaller parts of the system are defined.



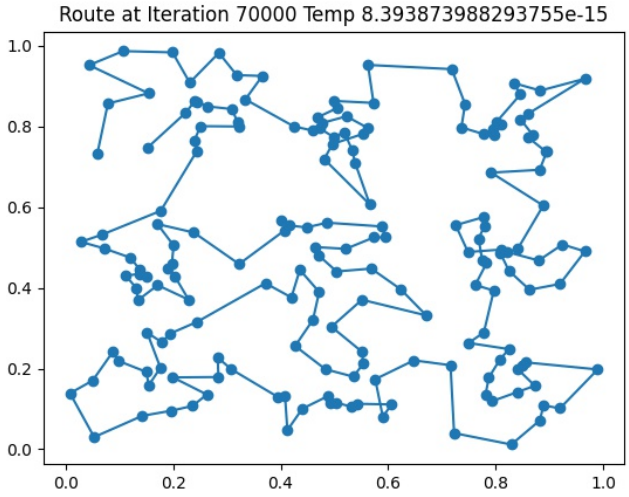
(a) At iteration 10000: no clear structure in route



(b) At iteration 20000: clear routes begin forming



(c) At iteration 50000: route is distinct but some routes cross each other



(d) At iteration 70000: crossing routes resolved

Figure 3: Evolution of route change over as iteration increases and temp decreases

References

[1] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

[2] Olivier Bousquet, Ulrike Luxburg, and Gunnar RÃ€tsch. Advanced lectures on machine learning, ml summer schools 2003, canberra, australia, february 2-14, 2003, tÃŒbingen, germany, august 4-16, 2003, revised lectures. 01 2004.

[3] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220:671–680, 1983.

[4] Scott M. Lynch. *Introduction to Applied Bayesian Statistics and Estimation for Social Scientists*. Springer Science & Business Media, LLC., New York, NY, April 2007.

[5] Christian P. Robert and George Casella. *Monte Carlo Statistical Methods (Springer Texts in Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2005.

[6] Ilker Yildirim. Bayesian inference: Metropolis-hastings sampling. 08 2012.

Appendix

A Markov Chains Primer

Definition A.1 (Markov Chain)**.** A Markov chain over a space χ is a series of random variables $X_0, X_1, X_2...$ such that the following properties hold:

1. The conditional independence of a term on the chain only depends on the most recent term:

$$\forall x_0, x_1, ... \in \chi, P(X_{t+1}|X_t = x_t, X_{t-1} = x_{t-1}, ..., X_0 = x_0) = P(X_{t+1}|X_t = x_t)$$

2. $P(X_t = x_t|X_{t-1} = x_{t-1})$ is independent of t

As a result of this properties, a Markov chain can be defined by a matrix $M(x, y)$ with $M(x, y) \geqslant 0$, $\forall x, \sum_y M(x, y) = 1$.

Definition A.2 (Stationary Distribution)**.** A distribution, π, is said to be stationary with respect to a Markov chain, M if

$$\pi = \pi M$$

All Markov chains have stationary distributions but not all converge to it.

A finite regular Markov chain has a unique stationary distribution π that it converges to if $\forall x \in \chi, \lim_{x\to\infty} P(X_i = x) = \pi(x)$.

There exists a sufficient (but not necessary) condition to prove that a distribution is a stationary distribution for a Markov chain.

Lemma A.1 (Reversible Markov Chains)**.** A distribution, π, is said to be stationary with respect to a Markov chain, M, if it satisfies the Detailed Balanced Equation

$$\pi(x)M(x, x') = \pi(x')M(x', x)$$

Definition A.3 (Irreducible)**.** A Markov chain, M, is said to be irreducible if for every two states in χ, they are reachable from each other.

Definition A.4 (Aperiodic)**.** A Markov chain, M, with state space $x_i, ..., x_n$,is aperiodic if $\exists n$ such that $\forall k \geqslant n$, $M_{ii}^k > 0$

Theorem 1 (Fundamental Theorem of Markov Chains)**.** *If a Markov chain P is* ***irreducible*** *and* ***aperiodic*** *then it has a unique stationary distribution π that it converges to.*

B Proof for claim 1

Proof. By expressing $\pi(x) = \tilde{\pi}(x)/Z_p$ and $q(z) = \tilde{q}(x)/Z_q$, we can define a new expression $\tilde{r}_l = \tilde{q}/\tilde{\pi}$.

We can re-express the expectation term with the following equation:

$$E(f) = \int f(z)\pi(z)dz = \frac{1}{Z_\pi/Z_q}\int \frac{\tilde{\pi}(z)}{\tilde{q}(z)} f(z(x))dz \simeq \frac{1}{Z_\pi/Z_q}\frac{1}{L}\sum_{l=1}^{L} \tilde{r}_l f(z^{(l)})$$

We can re-express the normalization constants:

$$\frac{Z_\pi}{Z_q} = \frac{1}{Z_q}\int \tilde{\pi}(z)dz = \int \frac{\tilde{\pi}(z)}{\tilde{q}z} q(z)dz \simeq \frac{1}{L}\sum_{l=1}^{L} \tilde{r}_l$$

And since

$$w_l \simeq \frac{r_l}{\sum_{l=1}^{L} \tilde{r}_l}$$

Which means

$$E(f) = \frac{1}{L}\sum_{l=1}^{L} w_l f(z^{(l)})dz$$

□