
Output Analysis

Nicholas Zabararas

***Materials Process Design and Control Laboratory
Sibley School of Mechanical and Aerospace Engineering***

188 Rhodes Hall

Cornell University

Email: zabararas@cornell.edu

URL: <http://mpdc.mae.cornell.edu/>

Introduction

- Let $X_0 = x^{(0)}, X_1 = x^{(1)}, \dots, X_N = x^{(N)}$ be a realization of a homogeneous and reversible Markov chain, output by some MCMC algorithm which was set up to sample a distribution $Q(x)$ on state space Ω .
- Suppose that $X \sim Q(x)$ is a random variable. A function $f(X)$ is called a statistic.
- Estimator
 - e.g. Consider the expected value of a statistic $f(X)$, i.e.

$$\langle f(X) \rangle \equiv \sum_{x \in \Omega} f(x) \Pr(x)$$

We can estimate this from the output sample set $\{x^{(n)}\}_{n=0}^N$.

The quantity

$$\bar{f}_N \left(\{x^{(n)}\}_{n=0}^N \right) \equiv \frac{1}{N} \sum_{n=1}^N f(x^{(n)})$$

is called the estimator of $\langle f(X) \rangle$.

- Since the samples $\{x^{(n)}\}_{n=0}^N$ are realizations of random variables $\{X_n\}_{n=0}^N$, the estimator is itself a random variable.

Introduction

- Consider the quality of an estimator, there are two issues
 - Systematic error due to initialization bias
Although $\pi^{(n)}$ may tend to some desired unique equilibrium distribution $Q(x)$, we should consider how large n need to be for $\pi^{(n)} \approx Q(x)$
 - Autocorrelation in equilibrium
Given $\pi^{(n)} = Q(x)$, for some n , how accurate is \bar{f}_N as an estimator for $\langle f(X) \rangle$?
Samples $x^{(n)}, x^{(n+1)}$ are highly correlated. How many samples should we take to achieve a given level of accuracy?

Autocorrelation in equilibrium

- If $x \sim Q(x)$, and $\mu_f \equiv \langle f(X) \rangle$,

$$\text{var}(f) = \langle f(X)^2 - \mu_f^2 \rangle$$

measures the variance of statistic f in samples x distributed like $Q(x)$.

- Given \bar{f}_N as the estimator of f . According to the “central limit theorem”, for N sufficiently large, the estimator

$$\bar{f} \sim \text{Normal}(\langle f \rangle, \text{var}(\bar{f}_N))$$

i.e. when N is large, our estimate \bar{f}_N is normally distributed with mean $\langle f(X) \rangle$ and some variance $\text{var}(\bar{f}_N)$.

- If $\{x^{(n)}\}_{n=0}^N$ were independent samples and $\bar{f}_N \equiv \frac{1}{N} \sum_{n=1}^N f(x^{(n)})$, then

$$\text{var}(\bar{f}_N) = \frac{1}{N} \text{var}(f)$$

Autocorrelation in equilibrium

- If $\{x^{(n)}\}_{n=0}^N$ is a sequence of correlated samples from a Markov chain, we have

$$\text{var}(\bar{f}_N) = \frac{\tau_f}{N} \text{var}(f)$$

where τ_f is a number characteristic of the transition matrix of the Markov chain used to generate the sequence $\{x^{(n)}\}_{n=0}^N$.

- The quantity τ_f is called the **integrated autocorrelation time** (IACT, in physics literature) or **autocovariance time** (statistics literature).
- For a given equilibrium distribution $Q(x)$, we would like to design a chain for which τ_f is as small as possible, so that we get accurate estimates without needing large sample sizes N .

Calculate Autocorrelation Time

- Let $\{X_n\}_{n=0}^{N-1}$ be a sequence of N stationary random variables in some given homogeneous Markov chain with equilibrium distribution Q . In terms of some statistic $f(x)$, let

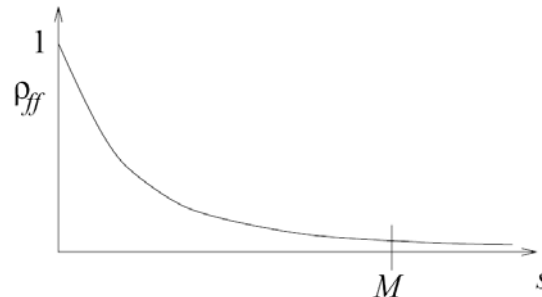
$$C_{ff}(s) \equiv \text{cov}(f(X_n), f(X_{n+s})) \equiv \langle f(X_n), f(X_{n+s}) \rangle - \mu_f^2$$

be the autocovariance function at lag s .

- $C_{ff}(s)$ is the covariance between the values taken by f for two random variables X_n and X_{n+s} in the chain.
 - $C_{ff}(s)$ depends only on s and not on n .
- Define the normalized autocovariance function

$$\rho_{ff}(s) = C_{ff}(s) / C_{ff}(0) = C_{ff}(s) / \text{var}(f)$$

We expect $\rho_{ff}(s) \rightarrow 0$ monotonically as $s \rightarrow \infty$



Calculate Autocovariance Time

$$\text{var}(\bar{f}_N) = \langle \bar{f}_N^2 \rangle - \langle \bar{f}_N \rangle^2 = \frac{1}{N^2} \sum_{n=1}^N \sum_{m=1}^N \langle f(X_m) f(X_n) \rangle - \langle f^2 \rangle$$

– Assume that

- for some M sufficiently large $\rho_{ff}(s) \approx 0$ when $s \geq M$
- $N \gg M$ so that the first $x^{(0)}$ and last $x^{(N)}$ samples are totally uncorrelated.

– Thus,

$$\begin{aligned} \sum_{n=1}^N \sum_{m=1}^N \langle f(X_m) f(X_n) \rangle &\approx \sum_{n=1}^N \left[\langle f(X_n) f(X_n) \rangle + 2 \sum_{s=1}^{N-M} \langle f(X_n) f(X_{n+s}) \rangle \right] \\ &= \sum_{n=1}^N \left[C_{ff}(0) + \langle f \rangle^2 + 2 \sum_{s=1}^{N-M} C_{ff}(s) + \langle f \rangle^2 \right] \\ \text{var}(\bar{f}_N) &= \frac{1}{N^2} \sum_{n=1}^N \text{var}(f) + 2 \sum_{s=1}^{N-M} C_{ff}(s) \approx \frac{\text{var}(f)}{N^2} \sum_{n=1}^N \left[1 + 2 \sum_{s=1}^M \rho_{ff}(s) \right] \end{aligned}$$

– we have

$$\begin{aligned} &\approx \frac{\text{var}(f) \tau_f}{N^2} \\ \tau_f &\equiv 1 + 2 \sum_{s=1}^{\infty} \rho_{ff}(s) \end{aligned}$$

Calculate Autocovariance Time

- To estimate τ_f from the output $\{x^{(n)}\}_{n=1}^N$ of an MCMC algorithm, we estimate $C_{ff}(s)$ by

$$\bar{C}_{ff}(s) = \frac{1}{N} \sum_{n=1}^N f(x^{(n)}) f(x^{(n+s)}) - \frac{1}{N^2} \left[\sum_{n=1}^N f(x^{(n)}) \right]^2$$

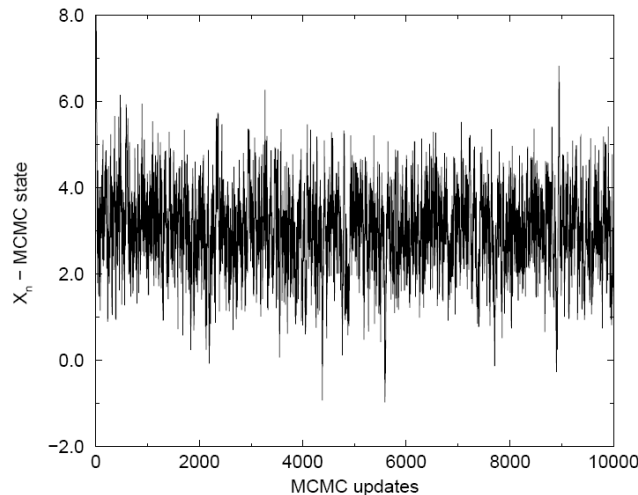
Calculate Autocovariance Time

➤ Example 1

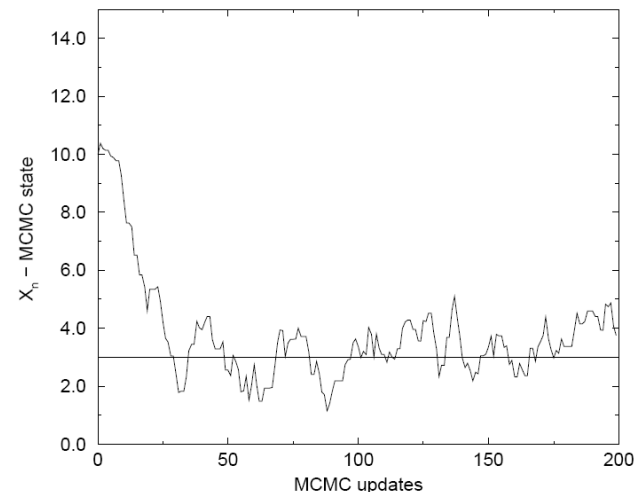
Illustrate output analysis on the output of the MCMC algorithm for sampling the normal distribution with a mean $\mu=3$ and standard deviation $\sigma=1$.

- According to the sequence of MCMC updates presented in the figure, the first 200 updates were discarded to allow time for the distribution of the Markov chain to converge to its equilibrium distribution.

sequence of 10000 MCMC updates output



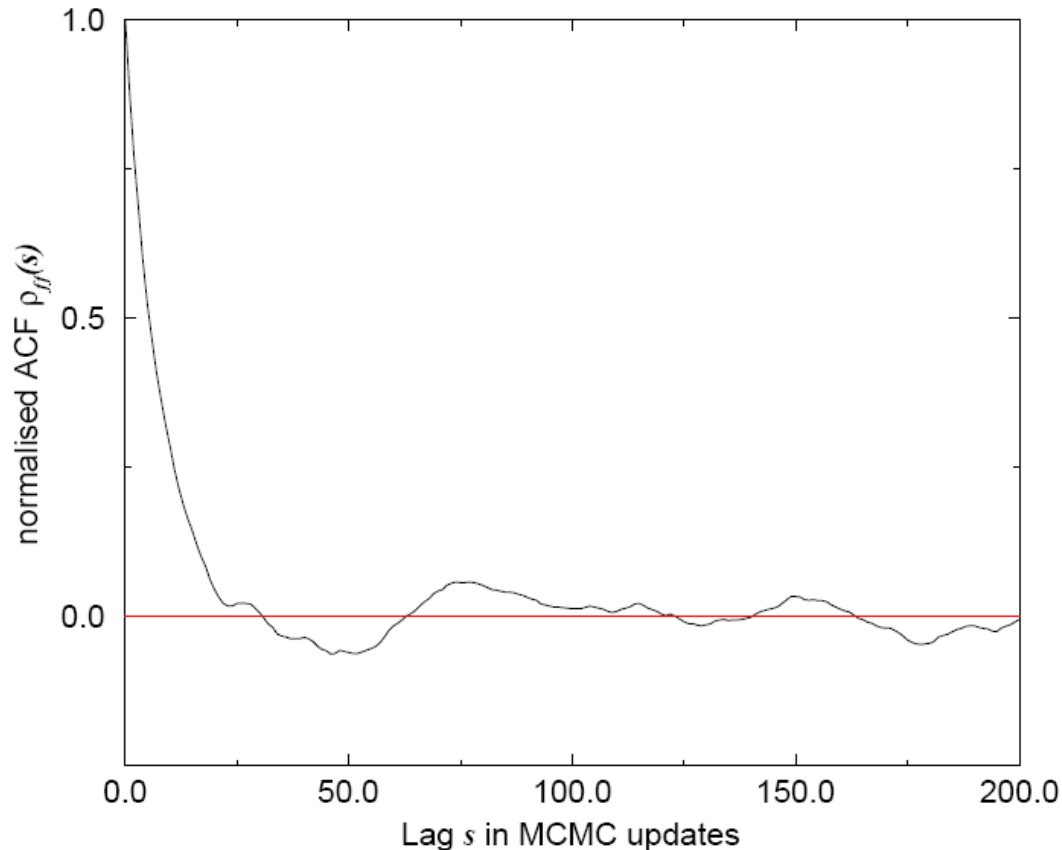
Converging MCMC for Normal distribution



Calculate Autocovariance Time

➤ Example 1

- Thus we have a sample sequence $\{x^{(n)}\}_{n=200}^{n=10000}$ of length 9800. The estimated normalized autocovariance function is

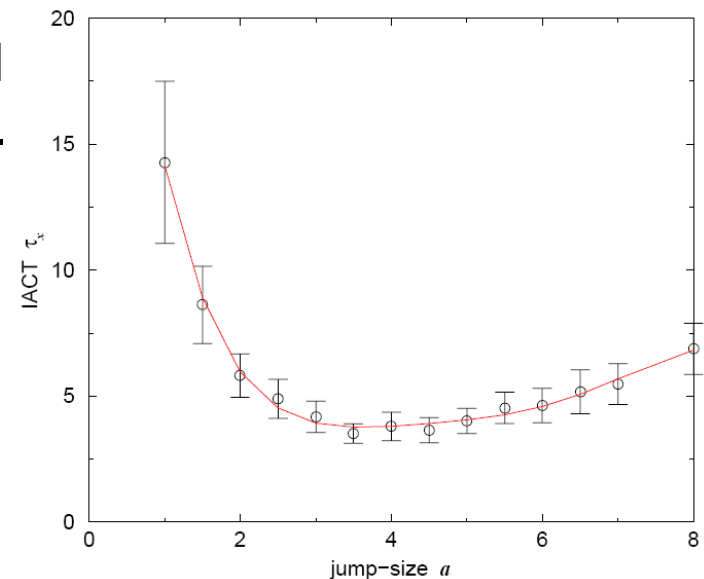


Calculate Autocovariance Time

➤ Example 2

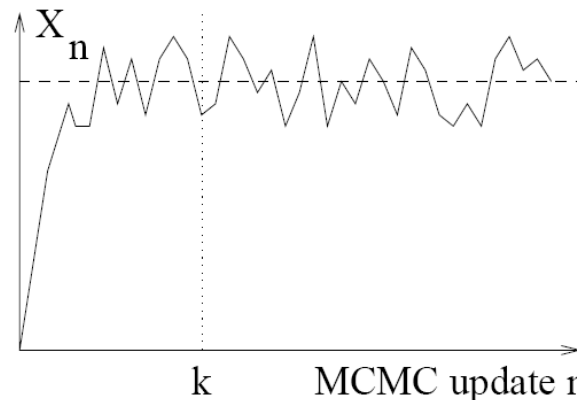
In the MCMC algorithm sampling the normal distribution, we generate candidate states by choosing new state uniformly on the interval $[x-a, x+a]$ where a is a constant.

- Can we reduce τ_f by using some other value of a ?
- Suppose the mean $\mu=3$ and standard deviation $\sigma=1$. For each of a sequence of a -values we run the MCMC and analyze the output, computing $\tau_x(a)$
- The autocorrelation time is minimized by a value of jumpsize of around $a=3$.
- The chain outputs the greatest number of effectively independent samples per N updates when the jumpsize is about 3.5



Initialization Bias

- Consider an MCMC algorithm with intended equilibrium distribution $Q(x)$. The initial samples generated by MCMC are not representative of the equilibrium distribution.
- How long should we simulate before beginning to take samples?
 - No general solution to this problem !
 - The usual approach is to monitor the behavior of some statistic
 - e.g. $q(x^{(n)})$ where $Q(x) = \exp(-q(x)) / Z$ and drop samples from the part of the run where $q(x)$ is converging to its equilibrium range.



A MCMC algorithm started at an unrepresentative state converges to its equilibrium range

Initialization Bias

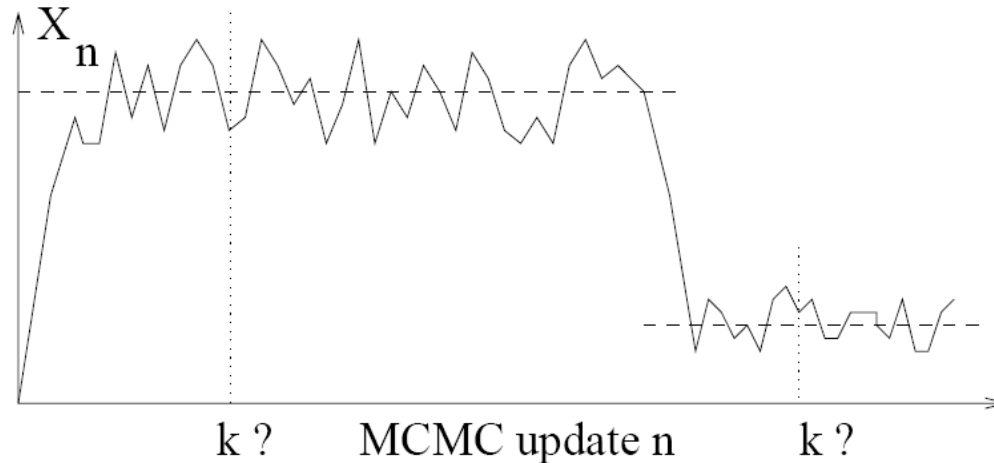
- We could make the output sample would consist of $x^{(n)}$ for $n \geq k$ only.
- Assuming that, for $n \geq k$ $X_n \sim Q$. If the total run length $N \gg k$, this has a negligible effect on estimates, i.e.

$$\bar{f}_{N-k} = \frac{1}{N-k} \sum_{n=k}^N f(x^{(n)})$$

when the first k samples are dropped.

Sticking and Multimodality

- Generally, it is unknown that when the Markov chain has reached equilibrium. (see the figure below)



- It may be that the generation probability $g(x'|x)$ can take the chain to any state in the state space Ω , but sections of Ω communicate only through states of very low acceptance probability. The chain may be stuck in the vicinity of one mode if $Q(x)$ is a multimodal distribution.

Sticking and Multimodality

➤ Example

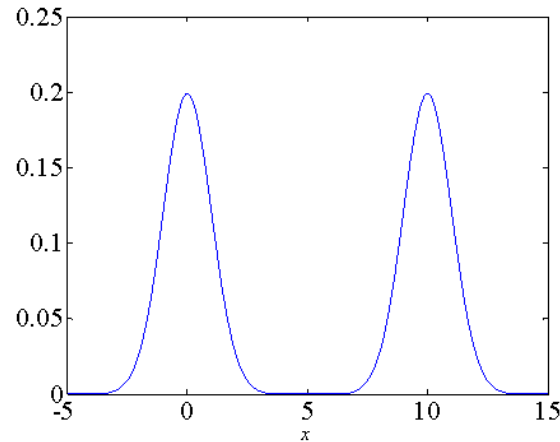
Suppose $Q(X \in dx)$ is the sum of two normal distribution with means μ_1 and μ_2 ($\mu_2 > \mu_1$) and each having the same standard deviation σ . If

$Q(X \in dx) = q(x)dx$, the density is

$$q(x) = \frac{1}{2\sqrt{2\pi}\sigma} \left\{ \exp\left(-\frac{(x-\mu_1)^2}{2\sigma^2}\right) + \exp\left(-\frac{(x-\mu_2)^2}{2\sigma^2}\right) \right\}$$

The two normal distributions are separated in the sense that $\mu_2 - \mu_1 \gg \sigma$

➤ e.g. $\mu_1 = 0$, $\mu_2 = 10$, $\sigma = 1$,



two regions of higher probability are separated by a region of low probability

Sticking and Multimodality

- MCMC algorithm with equilibrium $Q(x)$

Let $X_n=x$, X_{n+1} is determined in the following way

1. choose x' uniformly in an interval $[x-a, x+a]$ centered at x . Thus $x' \sim dx' / 2a$ and the density $1 / 2a$ is symmetric in x and x' .

2. With probability

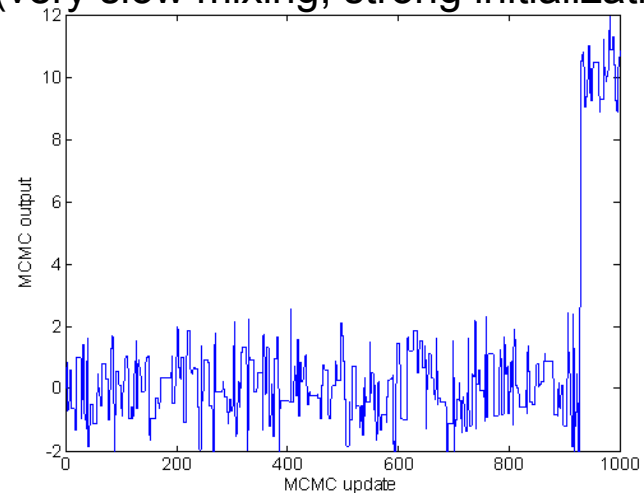
$$\alpha = \min \left\{ 1, \frac{\exp\left(-\frac{(x' - \mu_1)^2}{2\sigma^2}\right) + \exp\left(-\frac{(x' - \mu_2)^2}{2\sigma^2}\right)}{\exp\left(-\frac{(x - \mu_1)^2}{2\sigma^2}\right) + \exp\left(-\frac{(x - \mu_2)^2}{2\sigma^2}\right)} \right\}$$

set $X_{n+1}=x'$. Otherwise we set $X_{n+1}=x$.

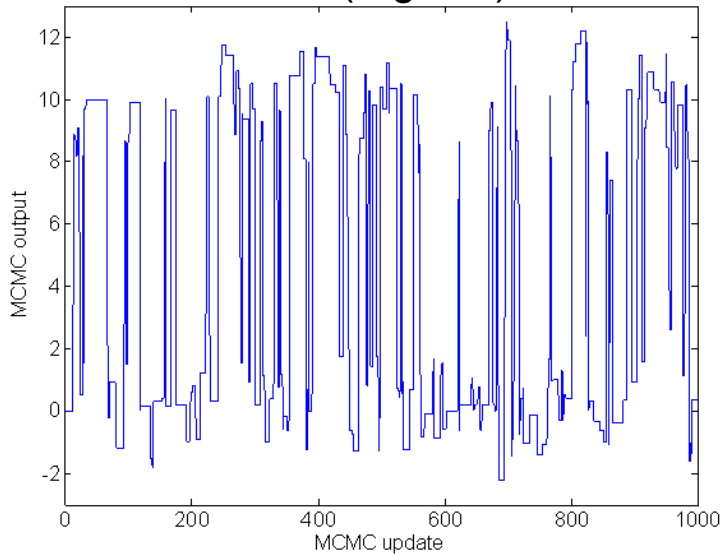
Sticking and Multimodality

- $\mu_1 = 0, \mu_2 = 10, \sigma = 1$
- $a = 1, 5, 10$

$a=5$ (very slow mixing, strong initialization bias)



$a=10$ (ergodic)



$a=1$ (stuck)

