CONSISTENCY AND CONVERGENCE RATE OF MARKOV CHAIN QUASI MONTE CARLO WITH EXAMPLES

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF STATISTICS
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Su Chen
August 2011

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Art. B. Owen) Principal Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Wing. H. Wong)

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Tze. L. Lai)

Approved for the University Committee on Graduate Studies

Abstract

Markov Chain Monte Carlo methods have been widely used in various science areas for generation of samples from distributions that are difficult to simulate directly. The random numbers driving Markov Chain Monte Carlo algorithms are modeled as independent $\mathcal{U}[0,1)$ random variables. By constructing a Markov Chain, we are able to sample from its stationary distribution if irreducibility and Harris recurrence conditions are satisfied. The class of distributions that could be simulated are largely broadened by using Markov Chain Monte Carlo.

Quasi-Monte Carlo, on the other hand, aims to improve the accuracy of estimation of an integral over the unit cube $[0,1)^k$. By using more carefully balanced inputs, under some smoothness conditions the estimation error can be shown to converge at a speed of $O\left(\frac{\log^k n}{n}\right)$, while the plain Monte Carlo method can only give a convergence rate of $O_p\left(\frac{1}{\sqrt{n}}\right)$. The improvement is significant when n is large.

In other words, Markov Chain Monte Carlo is creating a larger world, and quasi-Monte Carlo is creating a better world. Ideally we would like to combine these two techniques, so that we can sample more accurately from a larger class of distributions. This method, called Markov Chain quasi-Monte Carlo (MCQMC), is the main topic of this work.

The idea is simple: we are going to replace the IID driving sequence used in MCMC algorithms by a deterministic sequence which is designed to be more uniform. The

sequence we are using to substitute the IID sequence is a completely uniformly distributed sequence (CUD). Then two natural questions arisen: is MCQMC consistent? If it is, what is the convergence rate if we seek to estimate $\mathbf{E}_\pi f$ for some test function f?

Previously the justification for MCQMC is proved only for finite state space case. We are going to extend those results to some Markov Chains on continuous state spaces. The conditions under which the consistency holds will be given as well as some corresponding examples. We will also demonstrate the necessity of some of the conditions by counterexamples. We will show that, without these conditions, even for an IID sampling on a nice bounded region $\Omega \subseteq \mathbb{R}^2$, using a CUD sequence in place of random sequence could be highly biased.

We also explore the convergence rate of MCQMC under stronger assumptions. We adopt the Weighted Sobolev Space technique to show that a convergence rate of $O\left(\frac{1}{n^{1-\delta}}\right)$ can be achieved if the function f whose mean we seek to estimate composes the n step update function ϕ_n stays in the Weighted Sobolev Space $\mathcal{H}$. As an example, ARMA process is shown to satisfy these conditions.

Lastly we present some numerical results for demonstration of MCQMC's performance. From these examples, the empirical benefits of more balanced sequences are significant.

Acknowledgement

When I first started to read the books and articles about the lives of great mathematicians and scientists, I got a feeling that scientific researches are carried out by a group of siloed whizzes, who stay alone by him or herself with a pen and a piece of paper and create their groundbreaking works from nowhere. As I became a Ph.D student and began to do my own research, I figured out that my previous impression was highly biased. There could be some scientists, very few I suspect, who are able to work everything out with little support from others. This definitely does not apply to me.

This thesis could never be what it is without the guidance, help, support, and encouragement from many people whom I would like to thank. First and foremost, I would like to thank my Ph.D advisor Professor Art Owen. Art guided me through the past three years, introduced me to the world of Markov Chain Monte Carlo and Quasi Monte Carlo, and directed me to explore the areas that are still unknown. Art has great intuition for math and statistics, and he is a thinker with so many great ideas. Whenever I felt it is already a dead end, Art is always able to point me to a new direction. Art is a statistician with genuine love and passion for research, which has a long term impact on me. To me, Art is not only an advisor but also a mentor. I am deeply thankful for him dedicating so much time on me with great kindness, teaching me how to do research, how to write papers, correcting my English mistakes, and helping me adapt myself to the very different culture.

I would like to thank Professor Wing Wong and Professor Tze Lai for reading

my thesis and giving me lots of extremely helpful suggestions. I am very grateful to Professor Monic Sun and Professor Kay Giesecke for being on my oral committee. I would also like to thank Professor Persi Diaconis for his excellent class and educational discussions about Modern Markov Chain.

My collaborators: Josef Dick, Makoto Matsumoto, and Takuji Nishimura, have been a powerful propeller for me to proceed my research. I feel so lucky to meet Professor Josef Dick right at a time when I just started my research, and he has been an enormous help since then. Makoto Matsumoto and Takuji Nishimura have greatly deepened my understanding of random number generators. They always answer my questions and reply to my emails with great patience. I feel trully grateful to them.

My peers at Stanford have been a stable source for new ideas and fun. Special thanks to Wai Wai Liu, Xiaowei Zhang and John Jiang, who have provided a wide range of perspective on my work. Ya Xu, Justin Dyer, Li Ma and Yunting Sun have given me so many precious suggestions during the group meetings. Zongming Ma, Ling Chen, Feng Zhang, Shaojie Deng, Camilo Rivera, Jacob Bien, Ryan Tibshirani, Rahul Mazumder, Yi Liu, Victor Hu, Jeremy Shen, Brad Klingenberg, Zhen Zhu, Luo Lu, Vec Su and Jun Li are also due many thanks for their friendship.

Finally, my family have always been there and I can not thank them enough. My parents, although they are thousands of miles away from me, are always supporting me and encouraging me. I would like to thank my dear wife, Ruixue Fan, for her patience, understanding and caring. Although being very busy with her Ph.D work at Columbia University, Ruixue kept flying back and forth coast-to-coast to visit me. Without her company, my Ph.D life would have been very different. Thank you.

This thesis is dedicated to my advisor, my collaborators, my professors, my friends, and my family.

Contents

Abstract **iv**

Acknowledgement **vi**

1 Introduction **1**

1.1 Literature Review 2

1.2 Outline 3

1.3 New Results 4

2 Background **5**

2.1 Quasi-Monte Carlo 5

2.2 Plain Monte Carlo 8

2.3 MCMC iterations 9

2.4 Markov Chain quasi-Monte Carlo 12

3 Jordan Measurability **15**

3.1 Definitions 15

3.2 Charaterization of Riemann Integrable Functions 16

3.3 Importance of Being Riemann Integrable 18

3.4 QMC for Unsmooth Functions 19

4 Consistency of MCQMC **24**

4.1 Main Result for General Markov Chain 25

4.2 Main Results for Finite State Space Markov Chain 27

4.3 Main Results for Markov Chain on Continuous State Spaces 29
4.3.1 With Coupling Region 30
4.3.2 Global Contracting Mapping 32
4.4 Examples 36

5 Relaxation of Global Contracting Mapping **38**
5.1 Global Non-Expansive Mapping 38
5.2 Contracting on Average 44

6 Convergence Rate of MCQMC **59**
6.1 Reducing the Dimension 60
6.2 Weighted Sobolev Space 64
6.3 Example: ARMA Process 68

7 Consistent MCQMC Examples **76**
7.1 Coupling Region 76
7.2 Gibbs Sampling - Global Contracting Mapping 78
7.3 Global non-expansive Mapping 82
7.4 Contracting on Average 84

8 MCQMC in Practice **86**
8.1 Linear Feedback Shift Register 86
8.2 Algorithm Implementation 91
8.2.1 Construction of Variate Matrix 91
8.2.2 Randomization 92
8.3 Examples 93
8.3.1 2 Dimensional Gibbs Sampling 94
8.3.2 M/M/1 Queuing system 96
8.3.3 Dickman's Distribution 98
8.3.4 Garch Model 99
8.3.5 Heston's Stochastic Volatility Model 101

9 Conclusion and Future Directions **105**

9.1 Future Directions 106

9.1.1 Bias-Variance Tradeoff 106

9.1.2 Coupling Chains from Different Initial Values 107

Bibliography **108**

List of Tables

8.1 LFSR generators: Polynomials and Offsets 90
8.2 Digit Shift 93
8.3 MCQMC Algorithm 94
8.4 $\log_2$(Root Mean Squared Error) for Gibbs mean 95
8.5 $\log_2$(Root Mean Squared Error) for Gibbs correlation 97
8.6 $\log_2$(RMSE) for Average Waiting Time 98
8.7 $\log_2$(RMSE) for Dickman's Distribution 99
8.8 $\log_2$(Root Mean Squared Error) for Garch option pricing 101
8.9 $\log_2$(Root Mean Squared Error) for Stochastic Volatility Model 102

List of Figures

8.1 Numerical results for bivariate Gaussian Gibbs sampling. LFSR = solid and IID = dashed. The goal is to estimate the mean. The correlation is marked at the right. Y axis = $\log_2 \sqrt{\text{Mean Squared Error}}$ . . 96

8.2 Numerical results for bivariate Gaussian Gibbs sampling. LFSR = solid and IID = dashed. The goal is to estimate the correlation. The true correlation is marked at the right. Y axis = $\log_2(\sqrt{\text{Mean Squared Error}})$ 97

8.3 Numerical results for M/M/1 queue average waiting time. LFSR = solid and IID = dashed. 98

8.4 Numerical results for Dickman's Distribution. LFSR = solid and IID = dashed. The goal is to estimate the first and sixth moments of the stationary distribution. 100

8.5 Numerical results for Garch(1,1) pricing model. LFSR = solid and IID = dashed. The goal is to estimate the price of European Option. The initial price is marked at the right. The strike price K = 1. 103

8.6 Numerical results for Stochastic Volatility Model. LFSR = solid and IID = dashed. The goal is to estimate the price of European Option. The initial price = 100 = Strike price. 104

Chapter 1

Introduction

Markov Chain Monte Carlo (MCMC) is an important tool for generating samples from distributions that are hard to generate directly. The method was invented by Metropolis et al. [27] in 1953 and then greatly generalized by Hastings [15] in 1970. As the computing power grows rapidly, Markov Chain Monte Carlo has become more and more important and familiar to those who want to generate data from high dimensional space or large state space, in various areas such as statistical physics and computational chemistry. The justification of Markov Chain Monte Carlo relies on the ergodicity and Law of Large Numbers of Markov Chain, which requires the driving sequence to be IID from $\mathcal{U}[0,1)$. By assuming IID, the future of the Markov Chain conditioning on current state is independent of the past.

Quasi-Monte Carlo comes out from a very different motivation. It is trying to replace a randomly sampled sequence with a carefully designed sequence to get a better approximation to the uniform distribution on the unit cube. The points being constructed are usually highly correlated, while as a whole they are more uniformly distributed. The main goal of Quasi Monte Carlo is to more accurately estimate the integral of a certain function over the unit cube $[0,1)^k$. There were very sparse research on combining Markov Chain Monte Carlo and quasi-Monte Carlo until recently.

The motivation for replacing IID $\mathcal{U}[0,1)$ points in Markov Chain Monte Carlo

by a deterministic sequence is that by carefully designing the points, we can get a more accurate estimation by taking sample average. To be more specific, suppose we want to estimate $\mathbf{E}_\pi f$ for some function f where π is the stationary distribution of the Markov Chain. The MCMC algorithm takes the sample average $\frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{x}_i)$ to be an estimate of the quantity, where $\boldsymbol{x}_i = \phi(\boldsymbol{x}_{i-1}, \boldsymbol{u}_i)$ for $\boldsymbol{u}_i \in \mathcal{U}[0,1)^d$. Assuming irreducibility and Harris recurrence of the Markov Chain, the estimate will converge to the true value with probability 1 by the Law of Large Numbers. If we replace the IID sequence by a more balanced quasi-Monte Carlo sequence, we are hoping to benefit from the supreme uniformity. However, it is easy to doubt whether the consistency still maintains through a deterministic sequence, since the generated sample path $(\boldsymbol{x}_i)_{i\geq 1}$ is no longer Markovian, which means a small perturbation might cause a nonvanishing impact to the future.

The main goal of this work is the development of a theory that justifies the use of a deterministic sequence in the MCMC algorithm, which we call the Markov Chain quasi- onte Carlo (MCQMC) method. Owen and Tribble [36] proved the consistency result when the state space is finite. This work is going to extend the results to continuous state spaces, as well as give a unified approach for proving similar results in general case. Besides, we are going to discuss the convergence rate of MCQMC which partly justifies our motivation for using a more uniform but non-IID sequence.

1.1 Literature Review

There have been few attempts to combine Markov Chain Monte Carlo and Quasi Monte Carlo until recently. Chentsov [5] in 1967 replaced IID sequence by a CUD sequence which we will explain in Chapter 2. He proved the consistency of MCQMC under the assumptions that the state space is discrete, the transitions are sampled by inversion, and the transition probability from any state to any other state is strictly positive. Sobol [44] simulated the finite state Markov Chain using a $n \times \infty$ matrix. The algorithm starts with using the numbers in the first row. Every time the chain hits back the initial point $\boldsymbol{x}_0$, the numbers in the next row will be used. Liao [25]

simulated a Gibbs sampler using a set of Quasi Monte Carlo points after shuffling. He reported a variance reduction of 4 to 25 folds compared to usual MCMC. Lemieux and L'Ecuyer [22] used quasi-Monte Carlo points to generate a surplus process of a insurance company and estimate the ruin probability. L'Ecuyer, Lecot and Tuffin[21] adopted an Array-Randomized Quasi-Monte Carlo approach, constructed multiple copies of the same Markov Chain in parallel and carefully balanced the points in order to reduce the variance.

Only recently has there been a large progress in combination of MCMC and QMC, in both theory and practice. Owen and Tribble [36] proved the consistency of MCQMC on finite state spaces, under certain Jordan measurability assumptions. Tribble [45] demonstrates significant improvement from using a Mini Linear Feedback Shift Register, which rendered a variance reduction from hundreds to thousands folds. Those improvements have mainly arisen for continuous state space Markov Chains, which serves as a major motivation for us to investigate the consistency and convergence speed of MCQMC on continuous spaces.

1.2 Outline

This work begins with a brief overview of the background of Markov Chain Monte Carlo, quasi-Monte Carlo and Randomized quasi-Monte Carlo in Chapter 2. Chapter 3 provides some useful results about Jordan measurability which will be needed for the proofs in the following chapters. Chapter 3 also gives a proof of the L_2 consistency of Randomized quasi-Monte Carlo when evaluating $\int_{[0,1)^k} f(\boldsymbol{u})\,\mathrm{d}\boldsymbol{u}$ and the integrand is unsmooth. Chapter 4 establishes the theory of consistency of MCQMC on general state spaces, including a new proof for Owen and Tribble [36]'s result on finite state spaces. The theorem of consistency in the continuous state space case is based on Chen, Dick and Owen [3] with slight modifications. Chapter 5 continues to develop the consistency results on continuous state spaces, meanwhile relaxing the conditions imposed in Chapter 4. Chapter 6 discusses the convergence rate of MCQMC under

certain assumptions, and includes ARMA process as a special case. Chapter 7 gives several examples that satisfy the conditions for consistency to hold. As an application of the consistency Theorems, we prove that the partial sum of a CUD sequence is still CUD. Some of the examples have appeared in Chen et al. [3].

All experiments and numerical results are stated in Chapter 8, where we discuss in detail how to generate suitable quasi-Monte Carlo sequence and how to do randomization. A couple of examples are given in this chapter to compare the performance of MCQMC and usual MCMC algorithms. Much of this chapter is a restatement of [4].

1.3 New Results

Some of the results have appeared in [3] and [4]. The following new results are most significant:

- Theorem of L_2 consistency of quasi-Monte Carlo estimation when integrand is unsmooth, in Chapter 3.
- The entire discussion of Jordan measurability in Chapter 3.
- A unified theorem of consistency of MCQMC on general state spaces, in Chapter 4.
- Theorems of consistency of MCQMC on continuous state spaces under weaker assumptions, in Chapter 5.
- Discussion of the convergence rate of MCQMC, in Chapter 6.
- Proof of the theorem that the partial sum of a CUD sequence is also CUD, in Chapter 7.
- Some new simulation experiments, in Chapter 8.

Chapter 2

Background

In this chapter we give a brief overview of quasi-Monte Carlo and Markov Chain Monte Carlo. The aim is to make the reader familiar with the basic conceptions in these two areas, as well as to provide some notations which will be used throughout the paper. The central goal of the various algorithms covered in this chapter and in this paper, is to estimate $\mathbf{E}_\pi f$ where π is a distributuion defined on Ω. A large part of this chapter has appeared in [3].

2.1 Quasi-Monte Carlo

In this section we give a short summary of quasi-Monte Carlo. Further information may be found in the monograph by Niederreiter [32].

QMC is ordinarily used to approximate integrals over the unit cube $[0,1)^k$, for $k \in \mathbb{N}$. Let $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n \in [0,1)^k$. The QMC estimate of $\theta(f) = \int_{[0,1]^k} f(\boldsymbol{u})\,\mathrm{d}\boldsymbol{u}$ is $\hat{\theta}_n(f) = \frac{1}{n}\sum_{i=1}^n f(\boldsymbol{u}_i)$, just as we would use in plain Monte Carlo. The difference is that in QMC, distinct points $\boldsymbol{u}_i$ are chosen deterministically to make the discrete probability distribution with an atom of size $1/n$ at each $\boldsymbol{u}_i$ close to the continuous $\mathcal{U}[0,1)^k$ distribution.

The distance between these distributions is quantified by discrepancy measures.

The local discrepancy of $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$ at $\boldsymbol{a} \in [0,1)^k$ is

$$\delta(\boldsymbol{a}) = \delta(\boldsymbol{a}; \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = \frac{1}{n} \sum_{i=1}^{n} 1_{[\boldsymbol{0}, \boldsymbol{a})}(\boldsymbol{x}_i) - \prod_{j=1}^{k} a_j. \tag{2.1.1}$$

The star discrepancy of $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$ in dimension k is

$$D_n^{*k} = D_n^{*k}(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n) = \sup_{\boldsymbol{a} \in [0,1)^k} |\delta(\boldsymbol{a}; \boldsymbol{u}_1, \ldots, \boldsymbol{u}_n)|. \tag{2.1.2}$$

For $k = 1$, the star discrepancy reduces to the Kolmogorov-Smirnov distance between a discrete and a continuous uniform distribution.

Definition 2.1.1. *A sequence $(\boldsymbol{u}_i)_{i \geq 1}$ is called uniformly distributed, if*

$$D_n^{*k} \to 0, \quad \text{as } n \to \infty.$$

How fast D_n^{*k} goes to 0 as $n \to \infty$ indicates how "uniform" this sequence is. Under the null hypothesis, for an IID sequence, it can be shown (see [38]) that its star discrepancy converges at a speed of $O_p\left(n^{-\frac{1}{2}}\right)$:

$$D_n^{*k} = O_p\left(n^{-\frac{1}{2}}\right)$$

Such a bound is under the assumption that we are putting the points randomly inside of the unit cude. It is easy to imagine that we can do a better job by careful construction of the points. The large deviations of the empirical measure from uniformity occur when there are big clusters or voids of points in the cube. Hence intuitively to get a small star discrepancy, we should put points in some more regular way. People have found constructions of $\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots, \boldsymbol{u}_n, \cdots \in [0,1)^k$ such that $D_n^{*k} = O\left(\frac{(\log n)^k}{n}\right)$. Such constructions include, not exhaustively, Sobol's sequence, Halton's sequence, Niederreiter's sequence, and Lattice rules. A sequence having star discrepancy $O\left(\frac{(\log n)^k}{n}\right)$ is called a Low Discrepancy Sequence.

The star discrepancy plays an important role in determining the estimation error of an integral over the unit cube $[0,1)^k$. The following is the well-known Koksma-Hlawka inequality:

Theorem 2.1.2 (Koksma Hlawka inequality)**.**

$$|\hat{\theta}_n(f) - \theta(f)| = \left| \frac{1}{n} \sum_{i=1}^{n} f(\boldsymbol{u}_i) - \int_{[0,1)^k} f(\boldsymbol{u}) \, \mathrm{d}\boldsymbol{u} \right| \leq D_n^{*k} V_{\mathrm{HK}}(f), \tag{2.1.3}$$

where $V_{\mathrm{HK}}(f)$ is the total variation of f in the sense of Hardy and Krause. For properties of V_{HK} and other multidimensional variation measures see [35].

Equation (2.1.3) gives a deterministic upper bound on the integration error, and it factors into a measure of the points' quality and a measure of the integrand's roughness. As we have discussed above, the IID sequence will give a star discrepancy of order $O_p\left(n^{-\frac{1}{2}}\right)$ while a Low Discrepancy sequence could give a star discrepancy of order $O\left(\frac{(\log n)^k}{n}\right)$, therefore for functions of finite variation, they can be integrated at a much better rate by QMC than by plain Monte Carlo. This is the main motivation for people to use quasi-Monte Carlo. We would like to point out that under higher smoothness assumptions the rates of convergence of $O(n^{-\alpha}(\log n)^{d\alpha})$, where $\alpha \geq 1$ denotes the smoothness of the integrand which can therefore be arbitrarily large, can also be achieved [8].

In Theorem 2.1.2 we require the function have finite total variation. It will rule out most of the functions with discontinuity. On the other hand, for Monte Carlo estimation, as long as f has a finite second moment, by the Central Limit Theorem a MC estimation will have error bound $O_p(n^{-\frac{1}{2}})$. This means that Monte Carlo does not rely on the smoothness of the function while quasi-Monte Carlo does.

Equation (2.1.3) is not usable for error estimation. Computing the star discrepancy is very difficult [13], and computing $V_{\mathrm{HK}}(f)$ is harder than integrating f. Practical error estimates for QMC may be obtained using randomized quasi-Monte Carlo (RQMC). In RQMC each $\boldsymbol{u}_i \sim U[0,1)^k$ individually while the ensemble $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$

has $\mathbf{Pr}(D_n^{*k}(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n) < C(\log n)^k/n) = 1$ for some $C < \infty$. For an example see [34]. A small number of independent replicates of the RQMC estimate can be used to get an error estimate. RQMC has the further benefit of making QMC unbiased. For a survey of RQMC, see [23]. The randomization scheme we will use in this paper are Cranley-Patterson Rotation and Digital Shift, which will be defined when invoked.

Quasi-Monte Carlo is not restricted to computing expectation of a function over a unit cube. For arbitrary distribution π on Ω, if we can find a transformation $\psi : [0,1)^k \mapsto \Omega$ such that $\boldsymbol{u} \sim \mathcal{U}[0,1)^k \Rightarrow \psi(\boldsymbol{u}) \sim \pi$, then $\mathbf{E}_\pi f = \int_{[0,1)^k} f(\psi(\boldsymbol{u}))d\boldsymbol{u}$. Therefore, if we can find a transformation to turn a finite number of IID $\mathcal{U}[0,1)$ random numbers into a certain distribution π, then we can use QMC to estimate the expectation of a function under π.

However, in a lot of cases we can not find such transformations. Markov Chain Monte Carlo provides us a powerful tool to generate samples from a much wider class of distributions.

2.2 Plain Monte Carlo

In some situations we do know how to generate a distribution π through k IID $\mathcal{U}[0,1)$ random variables. Notice we require the number of inputs are fixed, so acceptance-rejection method is not directly covered.

For an encyclopedic presentation of methods to generate non-uniform random vectors see Devroye [6]. Here we limit ourselves to the inversion method and its generalization culminating in the Rosenblatt-Chentsov transformation introduced below.

Let F be the CDF of $x \in \mathbb{R}$, and for $0 < u < 1$ define

$$F^{-1}(u) = \inf\{\boldsymbol{x} \mid F(\boldsymbol{x}) \geq u\}.$$

Take $F^{-1}(0) = \lim_{u\to 0^+} F^{-1}(u)$ and $F^{-1}(1) = \lim_{u\to 1^-} F^{-1}(u)$, using extended reals if necessary. Then $\boldsymbol{x} = F^{-1}(u)$ has distribution F on $\mathbb{R}$ when $u \sim \mathcal{U}[0,1)$.

Multidimensional inversion is based on inverting the Rosenblatt transformation [41]. Let F be the joint distribution of $\boldsymbol{x} \in \mathbb{R}^k$. Let F_1 be the marginal CDF of x_1 and for $j = 2, \ldots, s$, let $F_j(\cdot\,; \boldsymbol{x}_{1:(j-1)})$ be the conditional CDF of x_j given $x_1, \ldots, x_{j-1}$. The inverse Rosenblatt transformation ψ_R of $\boldsymbol{u} \in [0,1)^k$ is $\psi_R(\boldsymbol{u}) = \boldsymbol{x} \in \mathbb{R}^k$ where

$$\begin{aligned} x_1 &= F_1^{-1}(u_1), \quad \text{and,} \\ x_j &= F_j^{-1}(u_j; \boldsymbol{x}_{1:(j-1)}), \quad j \geq 1. \end{aligned}$$

If $\boldsymbol{u} \sim \mathcal{U}[0,1)^k$, then $\psi_R(\boldsymbol{u}) \sim F$.

As we have mentioned in the previous chapter, if we can find the transformation ψ to generate π on Ω, we could use quasi-Monte Carlo to estimate $\mathbf{E}_\pi f(\boldsymbol{x})$ by replacing an IID sequence with a Low Discrepancy sequence $\in [0,1)^k$ where k is the number of $[0,1)$ random variables needed. If $f(\psi(\boldsymbol{u}))$ has finite total variation, by the Koksma-Hlawka inequality, quasi-Monte Carlo will give us a higher convergence rate than plain Monte Carlo. Both f and the transformation ψ can cause an infinite total variation. Therefore although there could be many different transformations generating the same distribution, we would prefer the one has the best smoothness. The inverse method and its multidimensional generalization are generally good in this sense, because they usually lead to smooth transformations if π is smooth and strictly nonzero on a convex support.

2.3 MCMC iterations

In this section we briefly review Markov Chain Monte Carlo methods. For a full description of MCMC see the monographs by Liu [26] or Robert and Casella [39]. Suppose that we want to sample $\boldsymbol{x} \sim \pi$ for a density function π defined with respect to Lebesgue measure on $\Omega \subseteq \mathbb{R}^s$. A MCMC algorithm on discrete state space is

similar, by replacing the density function π with a probability mass function. Here we describe the MCMC algorithm under the continuous state space setting.

For definiteness, we will seek to approximate $\theta(f) = \int_\Omega f(\boldsymbol{x})\pi(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$. In an MCMC simulation, we choose an arbitrary $\boldsymbol{x}_0 \in \Omega$ with $\pi(\boldsymbol{x}_0) > 0$ and then for $i \geq 1$ update via

$$\boldsymbol{x}_i = \phi(\boldsymbol{x}_{i-1}, \boldsymbol{u}_i) \tag{2.3.1}$$

where $\boldsymbol{u}_i \in \mathcal{U}[0,1)^d$ and ϕ is called an update function. The distribution of $\boldsymbol{x}_i$ depends on $\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{i-1}$ only through $\boldsymbol{x}_{i-1}$ and so these random variables have the Markov property. The function ϕ is chosen so that the stationary distribution of $\boldsymbol{x}_i$ is π. Then we estimate $\theta(f)$ by $\hat{\theta}_n(f) = \frac{1}{n}\sum_{i=1}^n f(\boldsymbol{x}_i)$ as before. If a burn-in period was used, we assume that $\boldsymbol{x}_0$ is the last point of it.

To ensure that $\hat{\theta}_n(f) \to \theta(f)$ *a.s.*, we need the Markov Chain to be irreducible and Harris recurrent (cf. Meyn and Tweedie [28]):

Definition 2.3.1. *A Markov Chain X_n defined on $(\Omega, \mathscr{F})$ is Harris recurrent, if for any $\boldsymbol{x}_0 \in \Omega$ and $A \in \mathcal{B}^+(\Omega)$:*

$$\mathbf{Pr}_{\boldsymbol{x}_0}(X_n \in A \text{ } i.o.) = 1$$

where $\mathcal{B}^+(\Omega) = \{A \in \mathscr{F} : \psi(A) > 0\}$ and ψ is the maximal irreducibility measure.

Under the assumption that the corresponding Markov Chain is irreducible and Harris recurrent, we have the Strong Law of Large Numbers (see, for example, Meyn and Tweedie [28]):

$$\hat{\theta}_n(f) = \frac{1}{n}\sum_{i=1}^n f(\boldsymbol{x}_i) \to \mathbf{E}_\pi f(\boldsymbol{x}), \qquad \text{a.s.}$$

for any starting point $\boldsymbol{x}_0$ if $\mathbf{E}_\pi|f(\boldsymbol{x})| < \infty$.

The Metropolis-Hastings algorithm provides a generic way of constructing such update functions. At any state $\boldsymbol{x} \in \Omega$, it begins with a proposal $\boldsymbol{y}$ taken from a transition kernel $P(\boldsymbol{x}, \mathrm{d}\boldsymbol{y})$. With genuinely random proposals, the transition kernel gives a complete description. But for either quasi-Monte Carlo or pseudo-random sampling, it matters how we actually generate the proposal. We will assume that $d-1$ $\mathcal{U}[0,1)$ random variables are used to generate $\boldsymbol{y}$ via $\boldsymbol{y} = \psi_{\boldsymbol{x}}(u_{1:(d-1)})$. Then the proposal $\boldsymbol{y}$ is either accepted or rejected with probability $A(\boldsymbol{x}, \boldsymbol{y})$. The decision is typically based on whether the d'th random variable u_d is below A.

Definition 2.3.2 (Generator)**.** *The function $\psi : [0,1)^d \to \mathbb{R}^s$ is a generator for the distribution F on $\mathbb{R}^s$ if $\psi(\boldsymbol{u}) \sim F$ when $\boldsymbol{u} \sim \mathcal{U}[0,1)^d$.*

Definition 2.3.3 (Metropolis-Hastings update)**.** *For $\boldsymbol{x} \in \Omega$ let $\psi_{\boldsymbol{x}} : [0,1)^{d-1}$ be a generator for the transition kernel $P(\boldsymbol{x}, \mathrm{d}\boldsymbol{y})$ with conditional density $p(\cdot \mid \boldsymbol{x})$. The Metropolis-Hastings sampler has*

$$\phi(\boldsymbol{x}, \boldsymbol{u}) = \begin{cases} \boldsymbol{y}(\boldsymbol{x}, \boldsymbol{u}), & u_d \le A(\boldsymbol{x}, \boldsymbol{u}) \\ \boldsymbol{x}, & u_d > A(\boldsymbol{x}, \boldsymbol{u}) \end{cases}$$

where $\boldsymbol{y}(\boldsymbol{x}, \boldsymbol{u}) = \psi_{\boldsymbol{x}}(\boldsymbol{u}_{1:(d-1)})$ and

$$A(\boldsymbol{x}, \boldsymbol{u}) = \min\left(1, \frac{\pi(\boldsymbol{y}(\boldsymbol{x}, \boldsymbol{u}))\, p(\boldsymbol{x} \mid \boldsymbol{y}(\boldsymbol{x}, \boldsymbol{u}))}{\pi(\boldsymbol{x})\, p(\boldsymbol{y}(\boldsymbol{x}, \boldsymbol{u}) \mid \boldsymbol{x})}\right).$$

Here we give some special cases of Metropolis-Hastings algorithms.

Example 2.3.4 (Metropolized independence sampler (MIS))**.** The MIS update is a special case of the Metropolis-Hastings update in which $\boldsymbol{y}(\boldsymbol{x}, \boldsymbol{u}) = \psi(\boldsymbol{u}_{1:(d-1)})$ does not depend on $\boldsymbol{x}$.

Example 2.3.5 (Random walk Metropolis (RWM))**.** The RWM update is a special case of the Metropolis-Hastings update in which $\boldsymbol{y}(\boldsymbol{x}, \boldsymbol{u}) = \boldsymbol{x} + \psi(\boldsymbol{u}_{1:(d-1)})$ for some generator ψ not depending on $\boldsymbol{x}$.

Definition 2.3.6 (Systematic scan Gibbs sampler)**.** *Let $\boldsymbol{x} = (x_1, \ldots, x_s) \in \mathbb{R}^d$ with $x_j \in \mathbb{R}^{k_j}$ and $d = \sum_{j=1}^s k_j$. To construct the systematic scan Gibbs sampler let*

$\psi_{j,\boldsymbol{x}_{-j}}(\boldsymbol{u}_j)$ be a k_j–dimensional generator of the full conditional distribution of x_j given x_ℓ for all $\ell \neq j$. This Gibbs sampler generates the new point using $\boldsymbol{u} \in [0,1)^d$. Write $\boldsymbol{u} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_s)$ with $\boldsymbol{u}_j \in [0,1)^{k_j}$. The systematic scan Gibbs sampler has

$$\phi(\boldsymbol{x}, \boldsymbol{u}) = (\phi_1(\boldsymbol{x}, \boldsymbol{u}), \phi_2(\boldsymbol{x}, \boldsymbol{u}), \ldots, \phi_s(\boldsymbol{x}, \boldsymbol{u}))$$

where, for $1 \le j \le s$,

$$\phi_j(\boldsymbol{x}, \boldsymbol{u}) = \psi_{j,\,\boldsymbol{x}_{[j]}}(\boldsymbol{u}_j)$$

and $\boldsymbol{x}_{[j]} = (\phi_1(\boldsymbol{x}, \boldsymbol{u}), \ldots, \phi_{j-1}(\boldsymbol{x}, \boldsymbol{u}), x_{j+1}, \ldots, x_d)$.

Example 2.3.7 (Inversive slice sampler)**.** Let π be a probability density function on $\Omega \subseteq \mathbb{R}^s$. Let $\Omega' = \{(y, \boldsymbol{x}) \mid \boldsymbol{x} \in \Omega, 0 \le y \le \pi(\boldsymbol{x})\} \subseteq \mathbb{R}^{s+1}$ and let π' be the uniform distribution on Ω'. The inversive slice sampler is the systematic scan Gibbs sampler for π' with each $k_j = 1$ using inversion for every $\psi_{j,\boldsymbol{x}_{[j]}}$.

There are many other slice samplers. See [29]. It is elementary that $(y, \boldsymbol{x}) \sim \pi'$ implies $\boldsymbol{x} \sim \pi$. It is more usual to use $(\boldsymbol{x}, y)$, but our setting simplifies when we assume y is updated first.

2.4 Markov Chain quasi-Monte Carlo

For plain Monte Carlo, we can easily replace an IID sequence by a Low Discrepancy sequence and gain a better convergence rate for estimating $\mathbf{E}_\pi f$ if $f \circ \psi$ is smooth. For MCMC algorithms, we would also like to replace the IID driving sequence by a more balanced sequence. In this context, we need a lesser known QMC concept as follows:

Definition 2.4.1. *A sequence $v_1, v_2, \cdots \in [0,1)$ is completely uniformly distributed (CUD) if for any $d \ge 1$ the points $\boldsymbol{u}_i^{(d)} = (v_i, \ldots, v_{i+d-1})$ satisfy*

$$D_n^{*d}(\boldsymbol{u}_1^{(d)}, \ldots, \boldsymbol{u}_n^{(d)}) \to 0$$

as $n \to \infty$.

This is one of the definitions of a random sequence from Knuth [17], and it is an important property for modern random number generators.

Definition 2.4.2. *A triangle array* $(v_i^j) : j = 1, 2, \cdots ; i = 1, 2, \cdots N_j$ *where* $N_j \to \infty$ *is called array-completely uniformly distributed, if for any* $d \geq 1$ *the points* $\boldsymbol{u}_{i,j}^{(d)} = (v_i^j, \ldots, v_{i+d-1}^j)$ *satisfy*

$$D^{*d}_{N_j-d+1}(\boldsymbol{u}^{(d)}_{1,j}, \ldots, \boldsymbol{u}^{(d)}_{N_j-d+1,j}) \to 0$$

as $j \to \infty$.

The Markov Chain quasi-Monte Carlo method, is to replace the IID driving squence by a CUD sequence $(v_i)_{i\geq 1}$ or an array-CUD sequence. Using a CUD (array-CUD) sequence in a MCMC algorithm is akin to using up the entire period of a random number generator, as remarked by [31] in 1986. It is then necessary to use a small random number generator. The CUD (array-CUD) sequences used by this work are miniature versions of linear feedback shift register generators (LFSR). As such they are no slower than ordinary pseudo-random numbers.

Consider non-overlapping d-tuples $\widetilde{\boldsymbol{u}}_i^{(d)} = (v_{di-d+1}, \ldots, v_{di})$ for $i \geq 1$. It is known, Chenstov [5], that

$$\begin{aligned} & D_n^{*d}(\boldsymbol{u}_1^{(d)}, \ldots, \boldsymbol{u}_n^{(d)}) \to 0, \quad \forall d \geq 1 \\ \iff & \\ & D_n^{*d}(\widetilde{\boldsymbol{u}}_1^{(d)}, \ldots, \widetilde{\boldsymbol{u}}_n^{(d)}) \to 0, \quad \forall d \geq 1. \end{aligned} \tag{2.4.1}$$

We will need one new technical Lemma about CUD points. Consider overlapping blocks of dk–tuples from $(v_i)_{i\geq 1}$, with starting indices d units apart. If v_i are CUD then these overlapping blocks are uniformly distributed. The proof works by embedding the dk–tuples into non-overlapping rdk–tuples. For large r the boundary effect between adjacent blocks becomes negligible. This result is needed for the consistency Theorems of MCQMC.

Lemma 2.4.3. *For $j \geq 1$ let $(v_j) \in [0,1)$ be a completely uniformly distributed sequence. For integers $d, i, k \geq 1$ let $\boldsymbol{u}_i = (v_{d(i-1)+1}, \ldots, v_{d(i-1)+dk})$. Then, $\boldsymbol{u}_i \in [0,1)^{dk}$ are uniformly distributed.*

Proof. Choose any $\boldsymbol{c} \in [0,1)^{dk}$. Let $v = \prod_{j=1}^{dk} c_j$ be the volume of $[\boldsymbol{0}, \boldsymbol{c})$. For integers $r \geq 1$ define f_r on $[0,1]^{rdk}$ by $f_r(\boldsymbol{u}) = \sum_{j=0}^{(r-1)k} 1_{[\boldsymbol{0},\boldsymbol{c})}(u_{jd+1}, \ldots, u_{jd+dk})$. Each f_r has integral $((r-1)k+1)v$. We use f_r on non-overlapping blocks of length rdk from u_j:

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^{n} 1_{[\boldsymbol{0},\boldsymbol{c})}(\boldsymbol{x}_i) \geq \frac{1}{n}\sum_{i=1}^{\lfloor n/(rk)\rfloor} f_r(u_{(i-1)rdk+1}, \ldots, u_{irdk}) \\ \to\ &\frac{(r-1)k+1}{rk}v > \frac{r-1}{r}v, \end{aligned}$$

after using (2.4.1). Taking r as large as we like, we get $\liminf_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} 1_{[\boldsymbol{0},\boldsymbol{c})}(\boldsymbol{x}_i) \geq v$. It follows that $\liminf_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} 1_{[\boldsymbol{a},\boldsymbol{b})}(\boldsymbol{x}_i) \geq \mathrm{Vol}[\boldsymbol{a}, \boldsymbol{b})$ for any rectangular subset $[\boldsymbol{a}, \boldsymbol{b}) \subset [0,1]^{dk}$. Therefore $\limsup_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} 1_{[\boldsymbol{0},\boldsymbol{c})}(\boldsymbol{x}_i) \leq v$ too, for otherwise some rectangle $[\boldsymbol{a}, \boldsymbol{b})$ would get too few points. □

In section 3.2 of Tribble [45], several examples were given to demonstrate why we can't just use a fixed dimension Low Discrepancy sequence. In Chapter 4 and Chapter 5 we will justify the practice of replacing IID sequence by a CUD or array-CUD sequence in MCMC algorithms.

Chapter 3

Jordan Measurability

In this chapter we will discuss the necessary regularity conditions for quasi-Monte Carlo to be consistent. By Koksma-Hlawka inequality, a QMC estimation is consistent as long as the function defined on unit cube is of finite total variation. However, such condition is unnecessarily strong. The most important concepts are Jordan measurability and Riemann integrability, which are closely related to each other. As we will see in this chapter, Riemann integrability is sufficient and necessary for quasi-Monte Carlo to work. In Chapter 4 it will be shown that Jordan measurability is critical for the consistency of Markov Chain quasi-Monte Carlo.

Throughout the section we assume $f(\boldsymbol{x})$ is a function defined on $[0,1)^k$ and λ_k is the k dimensional Lebesgue measure.

3.1 Definitions

Definition 3.1.1. *A bounded set $A \subseteq \mathbb{R}^k$ is called Jordan Measurable, if $f = \mathbf{1}_A(\boldsymbol{x})$ is Riemann Integrable.*

Definition 3.1.2. *For a function $f : \mathbb{R}^k \to \mathbb{R}$ the discontinuity set of f is*

$$D(f) = \{\boldsymbol{x} \in \mathbb{R}^k \mid f \text{ discontinuous at } \boldsymbol{x}\}.$$

If f is only defined on $A \subset \mathbb{R}^k$ then $D(f) = D(f_0)$ where $f_0(\boldsymbol{x}) = f(\boldsymbol{x})$ for $\boldsymbol{x} \in A$ and $f_0(\boldsymbol{x}) = 0$ for $\boldsymbol{x} \notin A$.

We have the following two useful propositions, for which the proofs can be found in [45], Theorem A.1.4 and Theorem A.1.5.

Proposition 3.1.3. *The collection of Jordan measurable sets in $[0,1)^k$ forms an algebra.*

Proposition 3.1.4. *The Cartesian product of two Jordan measurable sets is still Jordan measurable.*

3.2 Charaterization of Riemann Integrable Functions

First we have the famous theorem from Lebesgue which gives us a convenient way of checking a function's Riemann integrability:

Theorem 3.2.1 (Lebesgue's theorem)**.** *Let $A \subseteq \mathbb{R}^k$ be bounded and let $f : A \to \mathbb{R}$ be a bounded function. Then f is Riemann integrable iff $\lambda_k(D(f)) = 0$.*

Proof. Marsden and Hoffman (1993, page 455). □

Corollary 3.2.2. *A bounded set $A \subseteq \mathbb{R}^k$ is Jordan measurable, if and only if $\lambda_k(\partial(A)) = 0$.*

When discussing the consistency of MCQMC, many times we need to pay attention to the Jordan measurablity of the set $\{\boldsymbol{x} : f(\boldsymbol{x}) > C\}$ for some real number C. Here is a lemma which will be very useful later:

Lemma 3.2.3. *f is a Borel measurable function defined on $[0,1)^k$, then the following conditions are equivalent:*

1. *f is continuous almost everywhere.*
2. *For any bounded continuous function $g : \mathbb{R} \to \mathbb{R}$, $g \circ f$ is Riemann Integrable.*

3. *The set* $\{\boldsymbol{x} : f(\boldsymbol{x}) > C\}$ *is Jordan measurable for any* $C \in \mathbb{R} \setminus \mathcal{N}$ *where* $\mathcal{N}$ *is a null set w.r.t Lebesgue measure.*

Proof. 1) $\rightarrow$ 2): If g is bounded and continuous, then $g \circ f$ is also bounded and continuous almost everywhere, therefore by Lebesgue's Theorem it is Riemann integrable.

2) $\rightarrow$ 1): If for any bounded continuous function g, $g \circ f$ is Riemann Integrable, we can choose $g = \tan^{-1}$ and then we will get $\tan^{-1}(f)$ is Riemann Integrable. By Lebesgue's Theorem, $\tan^{-1}(f)$ is continuous almost everywhere. Thus $f = \tan(\tan^{-1}(f))$ is continuous almost everywhere.

1) $\rightarrow$ 3): In order to show $\{\boldsymbol{x} : f(\boldsymbol{x}) > C\}$ is Jordan Measurable, we just need to show $\mathbf{1}\{f(\boldsymbol{x}) > C\}$ is Riemann integrable. The discontinuity set of $\mathbf{1}\{f(\boldsymbol{x}) > C\}$ is a subset of $f^{-1}(C) \bigcup D(f) \bigcup \partial([0,1)^k)$. By assuming f to be continuous almost everywhere, we know $\lambda_k(D(f)) = 0$. Also $\lambda_k(\partial([0,1)^k) = 0$. Hence $\lambda_k(D(\mathbf{1}_{f(\boldsymbol{x})>C})) = 0$ if $\lambda_k(f^{-1}(C)) = 0$. We know $f^{-1}(C)$ is Borel measurable for any $C \in \mathbb{R}$, therefore there could only be at most countable C_i's such that $\lambda_k(f^{-1}(C_i)) > 0$, which proves the desired result.

3) $\rightarrow$1): Without loss of generality we can assume $f \in (-1, 1)$, otherwise we can work with $\frac{2}{\pi}\tan^{-1}(f)$. For $i = 1, 2, \cdots$, choose

$$-1 = C_i(0) < C_i(1) < \cdots < C_i(j) < \cdots C_i(2^i) = 1 \text{ for j} = 0 \cdots 2^i$$

be a partition of $[-1, 1]$ such that $\{\boldsymbol{x} : f(\boldsymbol{x}) > C_i(j)\}$ is Jordan measurable for any $C_i(j)$. This is doable since the set of C's that do not satisfy this condition forms a null set. Also we require partition C_{i+1} be a refinement of C_i such that $\lim_{i\to\infty} \max_{0\le j\le 2^i-1} |C_i(j+1) - C_i(j)| \to 0$. Then we can define

$$\underline{f_i}(\boldsymbol{x}) = C_i(j), \tilde{f}_i(\boldsymbol{x}) = C_i(j+1) \text{ if } f(\boldsymbol{x}) \in (C_i(j), C_i(j+1)]$$

By construction, we know

$$\underline{f_1} \leq \underline{f_2} \leq \cdots \leq f \leq \cdots \leq \tilde{f}_2 \leq \tilde{f}_1.$$

Since $\{\boldsymbol{x} : f(\boldsymbol{x}) > C_i(j)\}$ is Jordan measurable for any $C_i(j)$, we know $\underline{f_i}$ and $\tilde{f}_i$ are all Riemann Integrable, and $\sup|\tilde{f}_i - \underline{f_i}| \to 0$. By some simple calculus trick, we know f itself is Riemann integrable, hence continuous almost everywhere. □

Remark. A Riemann integrable function is NOT necessarily Borel Measurable.

Example 3.2.4. f is Borel measurable and continuous almost everywhere does not necessarily lead to Jordan measurability of $\{f(x) > C\}$. Here is the famous Thomae's function defined on $[0, 1]$:

$$f(x) = \begin{cases} \frac{1}{q} & x = \frac{p}{q}, \quad \gcd(p, q) = 1 \\ 0 & \text{else.} \end{cases}$$

This function is continuous on all the irrational numbers, therefore continuous almost everywhere. However, $\{f(x) > 0\} = \mathbb{Q} \bigcap [0, 1]$ which is not Jordan measurable.

3.3 Importance of Being Riemann Integrable

Riemann integrability is important in the nature of quasi-Monte Carlo. The Lebesgue integral was invented at the beginning of 20th century [20] and it turned out to be a much more delicate theory for integration than the Riemann integral. But pseudo-random number generators are now typically designed to meet an equidistribution criterion over rectangular regions. Other times they are designed with a spectral condition in mind. Therefore they are not "really" random, and most times they are only taking values on rational numbers. So in order for QMC to work, unless someone can use "Physical Random" numbers, Riemann integrability is crucial, which is illustrated in the following theorem and example:

Theorem 3.3.1. *Assume f is a Riemann Integrable function defined on $[0, 1)^k$, and $(\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots, \boldsymbol{u}_n, \cdots)$ is a uniformly distributed sequence $\in [0, 1)^k$, then we have the*

consistency result:

$$\frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{u}_i) \to \int_{[0,1)^k} f(\boldsymbol{u})\,\mathrm{d}\boldsymbol{u}$$

Proof. See Kuipers and Niederreiter [18]. □

Example 3.3.2 (Dirichlet Function)**.** The famous Dirichilet Function $f : [0,1] \to \mathbb{R}$ has the following form:

$$f(x) = \begin{cases} 1 & \text{if } x \in \mathbb{Q} \\ 0 & \text{Otherwise} \end{cases}$$

$\int_{[0,1]} f(x)dx = 0$. However, if we choose a sequence $(u_1, u_2, \cdots, u_n, \cdots)$ which only contains rational numbers to estimate the integral, we will reach a wrong conclusion that $\int_{[0,1]} f(x)\,\mathrm{d}x \approx 1$.

3.4 QMC for Unsmooth Functions

As we have seen in the previous section, quasi-Monte Carlo may fail when the integrand is not Riemann integrable. However, if we randomize the sequence, we might be able to avoid the extreme case. The randomization scheme we use is Cranley-Patterson rotation, as defined below:

Definition 3.4.1. *The Cranley Patterson rotation applied on a sequence $\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots \in [0,1)^k$ is defined as follows: Pick $\boldsymbol{v} \in \mathcal{U}[0,1)^k$, and let*

$$\boldsymbol{u}_i \mapsto \boldsymbol{u}_i + \boldsymbol{v} \triangleq (\boldsymbol{u}_i + \boldsymbol{v}) \bmod 1$$

Here is the main theorem:

Theorem 3.4.2. *Assume $f : [0,1)^k \to \mathbb{R}$ is a square integrable Lebesgue measurable function. $(\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots, \boldsymbol{u}_n, \cdots) \in [0,1)^k$ is a uniformly distributed sequence. $\boldsymbol{v} \in [0,1)^k$. We randomize the QMC sequence by Cranley Patterson rotation. I.e., we add $\boldsymbol{v}$ to each of the $\boldsymbol{u}_i$'s. Here $\boldsymbol{u}_i + \boldsymbol{v} \triangleq (\boldsymbol{u}_i + \boldsymbol{v}) \bmod 1$. Then we have the following*

convergence result:

$$\frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{u}_i + \boldsymbol{v}) \to \int_{[0,1)^k} f(\boldsymbol{u})\,\mathrm{d}\boldsymbol{u} \quad \textit{in } L_2([0,1)^k, \lambda_k). \tag{3.4.1}$$

Proof. Denote by $\mathscr{B}[0,1)^k$ the collection of Borel measurable sets on $[0,1)^k$. We use the "Standard Machine" in real analysis to prove the statement.

Step 1: We first prove that equation (3.4.1) holds for $f = \mathbf{1}_A$ for an interval $A = [\boldsymbol{a}, \boldsymbol{b})$ where $\boldsymbol{a} \leq \boldsymbol{b}$. This is easy because we know the randomized sequence $(\boldsymbol{u}_1+\boldsymbol{v}, \cdots, \boldsymbol{u}_n+\boldsymbol{v}, \cdots)$ is still uniformly distributed for any $\boldsymbol{v}$, and $\mathbf{1}_{[\boldsymbol{a},\boldsymbol{b})}$ is Riemann Integrable. Therefore by Theorem 3.3.1,

$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{[\boldsymbol{a},\boldsymbol{b})}(\boldsymbol{u}_i + \boldsymbol{v}) \to \int_{[0,1)^k} \mathbf{1}_{[\boldsymbol{a},\boldsymbol{b})}(\boldsymbol{u})\,\mathrm{d}\boldsymbol{u} = \lambda_k([\boldsymbol{a}, \boldsymbol{b}))$$

for all $\boldsymbol{v} \in [0,1)^k$. Since the indicator function is bounded, convergence almost surely leads to convergence in L_2.

Step 2: We secondly show that equation (3.4.1) holds for $f = \mathbf{1}_A$ where A is Borel measurable. Let the set collection

$$\mathscr{C} = \{A \subseteq [0,1)^k | A \text{ is Borel Measurable and } \mathbf{1}_A \text{ satisfies (3.4.1)}\}$$

By step 1, we have shown that $[\boldsymbol{a}, \boldsymbol{b}) \in \mathscr{C}$ for any $\boldsymbol{a} \leq \boldsymbol{b}$, thus we just need to show that $\mathscr{C}$ is a $\sigma-$ algebra. By Dynkin's $\pi - \lambda$ Theorem [1] , it suffices to show that $\mathscr{C}$ is a λ system. Obviously $\emptyset \in \mathscr{C}$ and $[0,1)^k \in \mathscr{C}$. If $A \subset B \in \mathscr{C}$, then easily we can see $B \setminus A \in \mathscr{C}$. The last thing we need to prove is, if $A_j \in \mathscr{C}$, $A_j \subset A_{j+1}$, then $A_\infty = \bigcup A_j \in \mathscr{C}$. For simplicity of notation let's define $f_j = \mathbf{1}_{A_j}(\boldsymbol{u})$. By the

monotone convergence theorem, $\int_{[0,1)^k} f_j \,\mathrm{d}\boldsymbol{u} \to \int_{[0,1)^k} f_\infty \,\mathrm{d}\boldsymbol{u}$. Hence we have

$$\frac{1}{n}\sum_{i=1}^{n} f_\infty(\boldsymbol{u}_i+\boldsymbol{v}) = \frac{1}{n}\sum_{i=1}^{n} f_j(\boldsymbol{u}_i+\boldsymbol{v}) + \frac{1}{n}\sum_{i=1}^{n}(f_\infty - f_j)(\boldsymbol{u}_i+\boldsymbol{v}) \tag{3.4.2}$$

$$\overset{n\to\infty}{\to} \int_{[0,1)^k} f_j \,\mathrm{d}\boldsymbol{u} + \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}(f_\infty - f_j)(\boldsymbol{u}_i+\boldsymbol{v}) \tag{3.4.3}$$

Then, by the Triangle Inequality:

$$\begin{aligned}
&\limsup_{n\to\infty}\left\|\frac{1}{n}\sum_{i=1}^{n} f_\infty(\boldsymbol{u}_i+\boldsymbol{v}) - \int_{[0,1)^k} f_\infty \,\mathrm{d}\boldsymbol{u}\right\|_{L_2} \\
\le &\limsup_{n\to\infty}\left\|\frac{1}{n}\sum_{i=1}^{n}(f_\infty - f_j)(\boldsymbol{u}_i+\boldsymbol{v})\right\|_{L_2} + \left|\int_{[0,1)^k}(f_j - f_\infty)\,\mathrm{d}\boldsymbol{u}\right| \\
= &\Sigma_1 + \Sigma_2
\end{aligned}$$

$$\Sigma_2 = \left|\int_{[0,1)^k}(f_j - f_\infty)\,\mathrm{d}\boldsymbol{u}\right| = \lambda_k(A_\infty \setminus A_j) \to 0 \text{ as } j\to\infty \tag{3.4.4}$$

Therefore we just need to show that $\limsup_{n\to\infty}\|\frac{1}{n}\sum_{i=1}^{n}(f_\infty - f_j)(\boldsymbol{u}_i+\boldsymbol{v})\|_{L_2} \to 0$. Still by the Triangle Inequality of L_2 norm, we have

$$\begin{aligned}
\Sigma_1 &\le \limsup_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\|(f_\infty - f_j)(\boldsymbol{u}_i+\boldsymbol{v})\|_{L_2} \\
&= \limsup_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\|f_\infty - f_j\|_{L_2} \\
&= \|f_\infty - f_j\|_{L_2} \\
&= \lambda_k(A_\infty \setminus A_j) \\
&\to 0
\end{aligned}$$

as $j\to\infty$. Therefore we complete the proof that $\mathscr{C}$ is a $\lambda-$ system. By the $\pi-\lambda$ theorem, $\mathscr{C} = \mathscr{B}[0,1)^k$.

Step 3: The above analysis shows that if $f = \mathbf{1}_A$ for A Borel measurable, equation (3.4.1) holds. It is easy to extend the result to $f = \mathbf{1}_A$ where A is Lebesgue measurable. To see that, we notice that Lebesgue measure is the completion of Borel measure, therefore $\forall A$ Lebesgue measurable, there exist A_1 and A_2 which are Borel measurable, such that $A_1 \subseteq A \subseteq A_2$ and $\lambda_k(A_2 \setminus A_1) = 0$. Then

$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{A_1}(\boldsymbol{u}_i + v) \leq \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{A}(\boldsymbol{u}_i + v) \leq \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{A_2}(\boldsymbol{u}_i + v)$$

and

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{A_1}(\boldsymbol{u}_i + v) = \lambda_k(A_1) = \lambda_k(A_2) = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{A_2}(\boldsymbol{u}_i + v)$$

which gives us the desired result.

Step 4: We want to show (3.4.1) holds for any $f \geq 0, f \in L_2([0,1)^k, \lambda_k)$. By the linearity of both sides of the equation (3.4.1), we know it holds for simple functions. For any $f \geq 0, f \in L_2([0,1)^k, \lambda_k)$, we can find $f_j : 0 \leq j < \infty$ simple such that

$$0 = f_0 \leq f_1 \leq f_2 \cdots \leq f_j \leq \cdots \leq f \stackrel{\Delta}{=} f_\infty$$

and $f_j \to f$ almost surely. By the Dominated Convergence Theorem, $f - f_j \to 0$ in L_2. Adopting a similar argument as before, we have the triangle inequality:

$$\begin{aligned}
&\limsup_{n\to\infty} \left\| \frac{1}{n}\sum_{i=1}^{n} f_\infty(\boldsymbol{u}_i + \boldsymbol{v}) - \int_{[0,1)^k} f_\infty d\boldsymbol{u} \right\|_{L_2} \\
\leq &\limsup_{n\to\infty} \left\| \frac{1}{n}\sum_{i=1}^{n} (f_\infty - f_j)(\boldsymbol{u}_i + \boldsymbol{v}) \right\|_{L_2} + \left| \int_{[0,1)^k} (f_j - f_\infty) d\boldsymbol{u} \right| \\
= &\Sigma_1 + \Sigma_2
\end{aligned}$$

$$\Sigma_2 = \left| \int_{[0,1)^k} (f_j - f_\infty) d\boldsymbol{u} \right| \to 0 \quad \text{as } j \to \infty \text{ by Dominated Convergence Theorem.}$$

$$\begin{aligned}
\Sigma_1 &= \limsup_{n\to\infty} \left\| \frac{1}{n} \sum_{i=1}^{n} (f_\infty - f_j)(\boldsymbol{u}_i + \boldsymbol{v}) \right\|_{L_2} \\
&\leq \limsup_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \|(f_\infty - f_j)(\boldsymbol{u}_i + \boldsymbol{v})\|_{L_2} \\
&= \limsup_{n\to\infty} \sum_{i=1}^{n} \|f_\infty - f_j\|_{L_2} \\
&= \|f_\infty - f_j\|_{L_2} \\
&\to 0
\end{aligned}$$

as $j \to \infty$. Therefore equation (3.4.1) holds for non-negative $f \in L_2$. Using the linearity again, we proved the theorem for all $f \in L_2$

□

Chapter 4

Consistency of MCQMC

When we use a MCMC algorithm to simulate samples from stationary distribution π, the first thing we are concerned about is whether the algorithm gives us a consistent result. I.e., whether the simulated sample path $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n, \cdots$ form an empirical distribution which converges to the stationary distribution. The answer is yes, with probability one, if the driving sequence is IID, given that the Markov Chain is irreducible and Harris recurrent(cf. Meyn and Tweedie [28]). In this chapter we would like to consider the consistency of MCQMC, i.e, replacing the IID driving sequence by a CUD (array-CUD) sequence. We do not expect CUD sequence to correct for a MCMC algorithm that would not satisfy Law of Large Numbers when using IID sequence. In light of this, throughout this chapter, we always assume irreducibility and Harris recurrence of the Markov Chain. To be more precise, here is the definition of consistency:

Definition 4.0.3. *$\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n, \cdots$ is said to Consistently Samples π, if for a distribution determining class $\mathcal{F}$ and all $f \in \mathcal{F}$,*

$$\frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{x}_i) \to \mathbf{E}_\pi f(\boldsymbol{x}) \tag{4.0.1}$$

Definition 4.0.4. *For a Markov Chain update function $\phi(\boldsymbol{x}, \boldsymbol{u})$, define the nth step*

iteration by:

$$\phi_1(\boldsymbol{x};\boldsymbol{u}) = \phi(\boldsymbol{x},\boldsymbol{u}) \text{ and } \phi_n(\boldsymbol{x};\boldsymbol{u}_n,\boldsymbol{u}_{n-1},\cdots,\boldsymbol{u}_1) = \phi(\phi_{n-1}(\boldsymbol{x};\boldsymbol{u}_{n-1},\cdots,\boldsymbol{u}_1),\boldsymbol{u}_n).$$

All of the results in this chapter apply to both CUD sequence and array-CUD sequences. For simplicity we focus on CUD sequence.

Most results in this chapter have appeared in [3] and [36]. Nevertheless, Theorem 4.1.1 is new and unifies the proofs for Markov Chains on different kinds of state spaces.

4.1 Main Result for General Markov Chain

Theorem 4.1.1. *Assume the sample space is Ω and let $(v_i)_{i\geq 1}$ be a CUD sequence. Let $\boldsymbol{x}_0 \in \Omega$ and $\boldsymbol{x}_i = \phi(\boldsymbol{x}_{i-1}, \widetilde{\boldsymbol{u}}_i)$ for $i \geq 1$. ϕ is a Markov Chain update function with stationary distribution π, $\widetilde{\boldsymbol{u}}_i = (v_{di-d+1},\cdots,v_{di})$. Assume for a class of bounded distribution determining functions $\mathcal{F}$, $\forall f \in \mathcal{F}$ and $\forall \epsilon > 0$, we can find $\mathcal{B}_m(\epsilon) \subseteq [0,1)^{dm}$ Jordan measurable with $Vol(\mathcal{B}_m(\epsilon)) \to 0$, such that*

$$\mathcal{A}_m(\epsilon) \triangleq \left\{ (\boldsymbol{u}_1,\cdots,\boldsymbol{u}_m) \in [0,1)^{dm} : \sup_{\widetilde{\boldsymbol{x}}_0 \in \Omega} \left| \frac{1}{m}\sum_{i=1}^{m} f(\widetilde{\boldsymbol{x}}_i) - \mathbf{E}_\pi f \right| > \epsilon \right\} \subseteq \mathcal{B}_m(\epsilon) \tag{4.1.1}$$

where $\widetilde{\boldsymbol{x}}_i = \phi(\widetilde{\boldsymbol{x}}_0;\boldsymbol{u}_i,\cdots,\boldsymbol{u}_1)$, then $\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n,\cdots$ consistently samples π.

Remark. Notice that $(\boldsymbol{x}_i)_{i\geq 1}$ is a deterministic sequence since $\boldsymbol{x}_0$ is fixed and $(\widetilde{\boldsymbol{u}}_i)_{i\geq 1}$ comes from a deterministic CUD sequence (v_i). Essentially $\mathcal{A}_m(\epsilon)$ defines a "Bad Set" of driving sequences that will crush the MCQMC algorithm. Condition (4.1.1) ensures that the volume of such "Bad Set" is negligible. As $(\widetilde{\boldsymbol{u}}_i)_{i\geq 1}$ is constructed from a CUD sequence, it will avoid the "Bad Set" eventually.

Proof. We just need to show that $\forall f \in \mathcal{F}$ and $\forall \epsilon > 0$

$$\limsup_{n\to\infty} \left| \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{x}_i) - \mathbf{E}_\pi f \right| \leq \epsilon. \tag{4.1.2}$$

Without loss of generality we assume $\sup|f| = 1$.

$$\begin{aligned}\sum_{i=1}^{n} f(\boldsymbol{x}_i) &= \sum_{j=1}^{\lfloor \frac{n}{m} \rfloor} \left(\sum_{k=1}^{m} f(\boldsymbol{x}_{jm-m+k}) \right) + \sum_{i=m\lfloor \frac{n}{m} \rfloor}^{n} f(\boldsymbol{x}_i) \\ &= \sum_{j=1}^{\lfloor \frac{n}{m} \rfloor} A_j + O(m) \\ \text{where } A_j &= \sum_{k=1}^{m} f(\boldsymbol{x}_{jm-m+k})\end{aligned}$$

A key observation is, $\boldsymbol{x}_{jm-m+k} = \phi_k(\boldsymbol{x}_{jm-m}; \widetilde{\boldsymbol{u}}_{jm-m+k}, \cdots, \widetilde{\boldsymbol{u}}_{jm-m+1})$. Therefore

$$\left| \frac{1}{m} A_j - \mathbf{E}_\pi f \right| \leq \epsilon \quad \text{if } (\widetilde{\boldsymbol{u}}_{jm-m+1}, \cdots, \widetilde{\boldsymbol{u}}_{jm-m+k}) \notin \mathcal{B}_m(\epsilon) \tag{4.1.3}$$

by condition 4.1.1. Thus,

$$\left| \frac{1}{n} \sum_{i=1}^{n} f(\boldsymbol{x}_i) - \mathbf{E}_\pi f \right| = \left| \frac{1}{n} \sum_{j=1}^{\lfloor \frac{n}{m} \rfloor} A_j + O\left(\frac{m}{n}\right) - \mathbf{E}_\pi f \right| \tag{4.1.4}$$

$$\leq \left| \frac{m}{n} \sum_{j=1}^{\lfloor \frac{n}{m} \rfloor} \left(\frac{A_j}{m} - \mathbf{E}_\pi(f) \right) \right| + O\left(\frac{m}{n}\right) \tag{4.1.5}$$

$$\leq \left| \frac{m}{n} \sum_{j=1}^{\lfloor \frac{n}{m} \rfloor} \mathbf{1}_{(\widetilde{\boldsymbol{u}}_{jm-m+1}, \cdots, \widetilde{\boldsymbol{u}}_{jm-m+k}) \in \mathcal{B}_m(\epsilon)} \right| + \epsilon + O(\frac{m}{n}) \tag{4.1.6}$$

$$\overset{n\to\infty}{\to} Vol(\mathcal{B}_m(\epsilon)) + \epsilon \tag{4.1.7}$$

(4.1.6) comes from assuming $|f| \leq 1$ and (4.1.3). by the definition of a CUD sequence. (4.1.7) comes from Lemma 2.4.3. Therefore

$$\limsup_{n\to\infty} \left| \frac{1}{n} \sum_{i=1}^{n} f(\boldsymbol{x}_i) - \mathbf{E}_\pi f \right| \leq Vol(\mathcal{B}_m(\epsilon)) + \epsilon$$

for any m and ϵ. Let $m \to \infty$ and $\epsilon \to 0$, by the assumption that $Vol(\mathcal{B}_m(\epsilon)) \to 0$, we proved the desired result. □

Remark. The distribution determining class $\mathcal{F}$ varies under different situations. For discrete state space Markov Chain, $\mathcal{F} = \{\mathbf{1}_{X=\omega} : \omega \in \Omega\}$. For Ω being a separable metric space, $\mathcal{F} = \{\text{all bounded continuous functions on } \Omega\}$.

4.2 Main Results for Finite State Space Markov Chain

Owen and Tribble (2005) in [36] proved an important result about consistency of MC-QMC algorithm under the assumption that the state space is finite. The original idea dates back to Chenstov [5] in 1967 from what he called a "standard construction" for Markov Chain simulations. The result is contained in [36], but here we are giving a new proof, which utilizes Theorem 4.1.1 proved in the previous section. We will see clearly from the following proof and in the next chapters that Theorem 4.1.1 provides us with a unified way of proving consistency of Markov Chain quasi-Monte Carlo in different settings.

Jordan measurability is still an issue in this case. The following Lemma shows that as long as one step transition function is Jordan measurable, multistep transition function will also be Jordan measurable.

Definition 4.2.1. *The Markov Chain update function ϕ on finite state space $\Omega = \{\omega_1, \omega_2, \cdots, \omega_{|\Omega|}\}$ is called Regular, if the set*

$$\left\{\boldsymbol{u} \in [0,1)^d : X_1 = \omega_j \text{ when } X_0 = \omega_i\right\} \tag{4.2.1}$$

is Jordan measurable for all (ω_i, ω_j)

Lemma 4.2.2. *If regularity of finite state space Markov Chain update function ϕ holds, then the m step transition sets defined as*

$$\mathcal{S}_{i,j}^m = \{(\boldsymbol{u}_1, \cdots, \boldsymbol{u}_m) \in [0,1)^{dm} : X_m = \omega_j \text{ when } X_0 = \omega_i\} \tag{4.2.2}$$

are Jordan measurable.

Proof. We prove the lemma by induction. The statement is true for $m = 1$, by assuming the regularity of ϕ. Assume the statement is true for $m \leq M$. Then:

$$\mathcal{S}_{i,j}^{M+1} = \bigcup_l \mathcal{S}_{i,l}^{M} \times \mathcal{S}_{l,j}^{1}$$

which is Jordan measurable by Proposition 3.1.3 and 3.1.4. □

In finite state space case, the class of bounded distribution determining functions $\mathcal{F}$ can be chosen as

$$\mathcal{F} = \{\mathbf{1}_{X=\omega} : \omega \in \Omega\}. \tag{4.2.3}$$

Theorem 4.2.3 (Owen, Tribble 2005)**.** *Suppose $\Omega = \{\omega_1, \cdots, \omega_{|\Omega|}\}$ is finite with Markov Chain update function $\phi(\boldsymbol{x}, \boldsymbol{u})$ which is regular. Assume the Markov Chain is recurrent with stationary distribution π. If $\boldsymbol{x}_0 \in \Omega$ and $\boldsymbol{x}_i = \phi(\boldsymbol{x}_{i-1}, \widetilde{\boldsymbol{u}}_i)$ where $\widetilde{\boldsymbol{u}}_i = (v_{di-d+1}, \cdots, v_{di})$ for a CUD sequence $(v_i)_{i\geq 1}$, then $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n, \cdots$ consistenly samples π.*

Proof. We choose the distribution determining class $\mathcal{F}$ as in (4.2.3). To prove the consistency, by Theorem 4.1.1, we just need to find $\mathcal{B}_m(\epsilon) \subseteq [0,1)^{dm}$ Jordan measurable with vanishing volume, such that:

$$\mathcal{A}_m(\epsilon) \triangleq \left\{ (\boldsymbol{u}_1, \cdots, \boldsymbol{u}_m) \in [0,1)^{dm} : \sup_{\widetilde{\boldsymbol{x}}_0 \in \Omega} \left| \frac{1}{m} \sum_{i=1}^{m} \mathbf{1}_{\widetilde{\boldsymbol{x}}_i = \omega} - \pi(\omega) \right| > \epsilon \right\} \subseteq \mathcal{B}_m(\epsilon) \tag{4.2.4}$$

Here we choose $\mathcal{B}_m(\epsilon) = \mathcal{A}_m(\epsilon)$. First we show that $\mathcal{A}_m(\epsilon)$ is Jordan measurable. This is because

$$\mathcal{A}_m(\epsilon) = \bigcup_l \left\{ (\boldsymbol{u}_1, \cdots, \boldsymbol{u}_m) \in [0,1)^{dm} : \left| \frac{1}{m} \sum_{i=1}^{m} \mathbf{1}_{\widetilde{\boldsymbol{x}}_i = \omega} - \pi(\omega) \right| > \epsilon \quad \middle| \widetilde{\boldsymbol{x}}_0 = \omega_l \right\} \tag{4.2.5}$$

$$= \bigcup_l \mathcal{A}_m^l(\epsilon) \tag{4.2.6}$$

All of these $\mathcal{A}_m^l(\epsilon)$'s are Jordan measurable, since they are finite unions of finite intersections of regions like (4.2.2), and we know by proposition 3.1.3, the Jordan

measurable sets form an algebra, thus are closed under finite intersection and union. Therefore $\mathcal{A}_m(\epsilon)$ is also Jordan measurable.

The remaining part is to show that $Vol(\mathcal{A}_m) \to 0$ as $m \to \infty$. Since $\mathcal{A}_m$ is Jordan measurable, its volume is the same as its Lebesgue measure. From equation (4.2.5) and (4.2.6), it suffices to show the measure of each of the $\mathcal{A}_m^l(\epsilon)$ converges to 0. It follows from the Weak Law of Large Numbers for a Recurrent finite state space Markov Chain. Hence we get the desired result. □

Remark. As we can see in the proof, finiteness is crucial for the proof to go through. For a Markov Chain on infinite state space, we are facing two major difficulties: first, Jordan measurable sets are not closed under infinite union or intersection. Secondly, showing $Vol(\mathcal{A}_m^l(\epsilon))$ does not automatically leads to $Vol(\mathcal{A}_m) \to 0$ since in this situation $\mathcal{A}_m(\epsilon)$ is an infinite union of the $\mathcal{A}_m^l(\epsilon)$'s.

4.3 Main Results for Markov Chain on Continuous State Spaces

In this section we assume that the sample space $\Omega \subseteq \mathbb{R}^s$ is equipped with Borel σ algebra $\mathscr{B}$. The results can be extended to certain metric spaces, but here we focus mainly on Euclidean Spaces. The results are contained in [3], but we here loosen the conditions and give a totally different proof. For continuous state space case, Jordan measurability issue is more difficult than finite state space case. One step transition being Riemann integrable is no longer sufficient, since the composition of two Riemann integrable functions are not necessarily Riemann integrable. Hence the definition of a Regular Markov Chain update function need to be modified as follows:

Definition 4.3.1 (Regularity of Markov Chain update function on Continuous State Space)**.** *Let $\boldsymbol{x}_m = \phi_m(\boldsymbol{x}_0; \boldsymbol{u}_m, \cdots, \boldsymbol{u}_1)$ be the last point generated by the Markov Chain update function $\phi(\boldsymbol{x}, \boldsymbol{u})$, viewed as a function on $[0,1)^{dm}$. The Markov Chain update function ϕ is called Regular at initial value $\boldsymbol{x}_0$, if $\phi_m(\boldsymbol{x}_0; \boldsymbol{u}_m, \cdots, \boldsymbol{u}_1)$ is continuous almost surely on $[0,1)^{dm}$. ϕ is called Regular if it is regular for all $\boldsymbol{x}_0 \in \Omega$.*

4.3.1 With Coupling Region

We begin with a definition:

Definition 4.3.2 (Coupling Region)**.** *Let $\mathcal{C} \subseteq [0,1)^d$ have positive Jordan measure. If $\boldsymbol{u} \in \mathcal{C}$ implies that $\phi(\boldsymbol{x}, \boldsymbol{u}) = \phi(\boldsymbol{x}', \boldsymbol{u})$ for all $\boldsymbol{x}, \boldsymbol{x}' \in \Omega$, then $\mathcal{C}$ is a coupling region.*

A Markov Chain with coupling region forgets about the past: as long as $\boldsymbol{u}_i \in \mathcal{C}$, $\boldsymbol{x}_i$ does not depend on $\boldsymbol{x}_{i-1}$. Notice here the "forgetting about the past" is not only in probablistic sense but also in a path by path sense. Below is the Theorem 2 in [3], and here we have a new proof using Theorem 4.1.1.

Theorem 4.3.3. *Let $\Omega \subseteq \mathbb{R}^s$ and let $\boldsymbol{x}_0 \in \Omega$, and for $i \geq 1$ let $\boldsymbol{x}_i = \phi(\boldsymbol{x}_{i-1}, \widetilde{\boldsymbol{u}}_i)$ where ϕ is a Markov Chain update function with a coupling region $\mathcal{C}$. Assume the Markov Chain is irreducible and Harris recurrent with stationary distribution π. If $\widetilde{\boldsymbol{u}}_i = (v_{d(i-1)+1}, \ldots, v_{di})$ for a CUD sequence $(v_i)_{i\geq 1}$, and ϕ is Regular at some initial state $\boldsymbol{x}_0^*$, then $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, \cdots$ consistently samples π.*

Proof. Pick $\epsilon > 0$. We choose $\mathcal{F}$ to be the collection of all bounded continuous functions on $\Omega \subseteq \mathbb{R}^s$. By Portmanteau Theorem (see [11]), $\mathcal{F}$ is a distribution determining class of $(\Omega, \mathscr{B})$. Thus we just need to find $\mathcal{B}_m(\epsilon)$ which is Jordan measurable, $Vol(\mathcal{B}_m(\epsilon)) \to 0$ and satisfies (4.1.1):

$$\mathcal{A}_m(\epsilon) \triangleq \left\{ (\boldsymbol{u}_1, \cdots, \boldsymbol{u}_m) \in [0,1)^{dm} : \sup_{\widetilde{\boldsymbol{x}}_0 \in \Omega} \left| \frac{1}{m} \sum_{i=1}^{m} f(\widetilde{\boldsymbol{x}}_i) - \mathbf{E}_\pi f \right| > \epsilon \right\} \subseteq \mathcal{B}_m(\epsilon) \tag{4.3.1}$$

where $\widetilde{\boldsymbol{x}}_i = \phi_i(\widetilde{\boldsymbol{x}}_0; \boldsymbol{u}_i, \cdots, \boldsymbol{u}_1)$. Without loss of generality, we assume $|f| \leq 1$.

Let $\boldsymbol{x}_i^* = \phi(\boldsymbol{x}_{i-1}^*, \boldsymbol{u}_i)$ be the point generated by the Markov Chain updating starting from initial value $\boldsymbol{x}_0^* \in \Omega$. Then,

$$\sup_{\widetilde{\boldsymbol{x}}_0 \in \Omega} \left| \frac{1}{m} \sum_{i=1}^{m} f(\widetilde{\boldsymbol{x}}_i) - \mathbf{E}_\pi f \right| \leq \left| \frac{1}{m} \sum_{i=1}^{m} f(\boldsymbol{x}_i^*) - \mathbf{E}_\pi f \right| + \sup_{\widetilde{\boldsymbol{x}}_0 \in \Omega} \left| \frac{1}{m} \sum_{i=1}^{m} \left(f(\widetilde{\boldsymbol{x}}_i) - f(\boldsymbol{x}_i^*) \right) \right| \tag{4.3.2}$$

$$= \Sigma_1 + \Sigma_2 \tag{4.3.3}$$

For Σ_1 we can use the Strong Law of Large Numbers for Markov chain, and for Σ_2 we resort to the existence of Coupling Region. Since we are assuming irreducibility and Harris recurrence, by Strong Law of Large Numbers, define

$$\mathcal{B}_m(\Sigma_1) \triangleq \left\{ (\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots, \boldsymbol{u}_m) \in [0,1)^{dm} : \left| \frac{1}{m} \sum_{i=1}^{m} f(\boldsymbol{x}_i^*) - \mathbf{E}_\pi f \right| > \frac{\epsilon}{2} \right\} \tag{4.3.4}$$

$$\lambda_{dm}(\mathcal{B}_m(\Sigma_1)) \to 0 \text{ when } m \to \infty \tag{4.3.5}$$

Notice by the definition of ϕ being regular at $\boldsymbol{x}_0^*$, $\mathcal{B}_m(\Sigma_1)$ is Jordan measurable for all $\epsilon > 0$ except for a Null set $\mathcal{N}$. We can for simplicity assume the Jordan measurability here, otherwise we can choose $0 < \tilde{\epsilon} < \epsilon$ such that Jordan measurability is satisfied for $\tilde{\epsilon}$.

For Σ_2, define

$$\mathcal{B}_m(\Sigma_2) \triangleq \left\{ (\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots, \boldsymbol{u}_m) : \boldsymbol{u}_j \notin \mathcal{C} \text{ for all } j \in [1, \lfloor \frac{\epsilon}{4} m \rfloor] \right\} \tag{4.3.6}$$

If $(\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots, \boldsymbol{u}_m) \notin \mathcal{B}_m(\Sigma_2)$, then $\exists j \in [1, \lfloor \frac{\epsilon}{4} m \rfloor]$ such that $\boldsymbol{u}_j \in \mathcal{C}$, which by definition of coupling region implies $\widetilde{\boldsymbol{x}}_i = \boldsymbol{x}_i^*$ for $i \geq \frac{\epsilon}{4} m$. Noticing the assumption that $|f| \leq 1$, we have for $(\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots, \boldsymbol{u}_m) \notin \mathcal{B}_m(\Sigma_2)$,

$$\Sigma_2 = \sup_{\widetilde{\boldsymbol{x}}_0 \in \Omega} \left| \frac{1}{m} \sum_{i=1}^{m} (f(\widetilde{\boldsymbol{x}}_i) - f(\boldsymbol{x}_i^*)) \right| \leq \frac{\epsilon}{4} m \times 2 \times \frac{1}{m} = \frac{\epsilon}{2} \tag{4.3.7}$$

$\mathcal{B}_m(\Sigma_2) = \prod_{j=1}^{\lfloor \frac{\epsilon}{4} m \rfloor} [0,1)^d \setminus \mathcal{C}$, which is Jordan measurable by Proposition 3.1.4. To compute the volume of $\mathcal{B}_m(\Sigma_2)$,

$$Vol(\mathcal{B}_m(\Sigma_2)) = (1 - Vol(\mathcal{C}))^{\lfloor \frac{\epsilon}{4} m \rfloor} \to 0 \tag{4.3.8}$$

as $m \to \infty$. Combining (4.3.2),(4.3.5) and (4.3.7), (4.3.8), we have:

$$\mathcal{A}_m(\epsilon) \subseteq \mathcal{B}_m(\Sigma_1) \bigcup \mathcal{B}_m(\Sigma_2) \triangleq \mathcal{B}_m(\epsilon)$$

$\mathcal{B}_m(\epsilon)$ is Jordan measurable with vanishing volume, hence we completed our proof. □

4.3.2 Global Contracting Mapping

In this section, we assume that the update function $\phi(\boldsymbol{x}, \boldsymbol{u})$ is jointly Borel measurable in $\boldsymbol{x}$ and $\boldsymbol{u}$ and that it is Lipschitz continuous in $\boldsymbol{x}$ for any $\boldsymbol{u}$. Lipschitz continuity is defined through a possibly different metric $d(\cdot, \cdot)$ on Ω. In another word, $d(\cdot, \cdot)$ is not necessarily equal to the Euclidean norm. However, we do require d give rise to the same topology on Ω as the Euclidean norm.

The Lipschitz constant, which depends on $\boldsymbol{u}$, is defined as:

$$\ell(\boldsymbol{u}) = \sup_{\boldsymbol{x} \neq \boldsymbol{x}'} \frac{d\big(\phi(\boldsymbol{x}, \boldsymbol{u}), \phi(\boldsymbol{x}', \boldsymbol{u})\big)}{d(\boldsymbol{x}, \boldsymbol{x}')} < \infty. \tag{4.3.9}$$

For each $\boldsymbol{u}_n \in [0,1)^d$ define $L_n = \ell(\boldsymbol{u}_n)$.

Lemma 4.3.4. *$\ell(\boldsymbol{u})$ is a Borel measurable function.*

Proof. Ω is separable therefore we can find a dense countable subset $(\boldsymbol{x}_i)_{i \geq 1} \in \Omega$. Using the continuity of ϕ in $\boldsymbol{x}$, we get

$$\ell(\boldsymbol{u}) = \sup_{\boldsymbol{x}_i \neq \boldsymbol{x}_j} \frac{d\big(\phi(\boldsymbol{x}_i, \boldsymbol{u}), \phi(\boldsymbol{x}_j, \boldsymbol{u})\big)}{d(\boldsymbol{x}_i, \boldsymbol{x}_j)}.$$

Therefore $\ell(\boldsymbol{u})$ is the supereme of countable Borel measurable functions, which is also Borel measurable. □

Lemma 4.3.5. *Assume $\ell(\boldsymbol{u})$ is continuous almost everywhere on $[0,1)^d$, then*

$$\{(\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots, \boldsymbol{u}_n) : \log(\ell(\boldsymbol{u}_1)) + \log(\ell(\boldsymbol{u}_2)) + \cdots + \log(\ell(\boldsymbol{u}_n)) > C\}$$

is Jordan Measurable for all $C \in \mathbb{R} \setminus \mathcal{N}$ where $\mathcal{N}$ is a null set w.r.t Lebesgue measure.

Proof. The sum of almost continuous functions is also almost continuous. Then it follows easily from Lemma 3.2.3. □

Definition 4.3.6. *An update function $\phi(\boldsymbol{x}, \boldsymbol{u})$ is called Global Contracting, if*

$$\int_{[0,1)^d} \log(\ell(\boldsymbol{u}))\,\mathrm{d}\boldsymbol{u} < 0.$$

The essence of the idea behind global contracting update function is to ensure that the Markov Chain will eventually forget about the past. We know that a geometric ergodic Markov Chain forgets about its past at a geometric rate in a probabilistic sense. Here our assumption is even stronger: not only the distribution of X_n does not depend on the initial value X_0 when n gets big, we require the realized value $\boldsymbol{x}_n = X_n(\omega)$ does not depend on the initial value $\boldsymbol{x}_0 = X_0(\omega)$ when $n \to \infty$.

Theorem 4.3.7. *Assume the sample space is $(\Omega, \mathscr{B})$ where $\Omega \subseteq \mathbb{R}^s$ and $\mathscr{B}$ is the Borel σ algebra and let $(v_i)_{i\geq 1}$ be a CUD sequence. Let $\boldsymbol{x}_0 \in \Omega$ and $\boldsymbol{x}_i = \phi(\boldsymbol{x}_{i-1}, \widetilde{\boldsymbol{u}}_i)$ for $i \geq 1$ where ϕ is a Markov Chain update function. Assume the Markov Chain is irreducible and Harris recurrent with stationary distribution π and $\widetilde{\boldsymbol{u}}_i = (v_{di-d+1}, \cdots, v_{di})$. Let $d(\cdot,\cdot)$ be a metric defined on Ω which gives the same topology as Euclidean norm. Assume Ω is bounded under $d(\cdot,\cdot)$ and $\ell(\boldsymbol{u})$ is continuous almost everywhere. If $\phi(\boldsymbol{x}, \boldsymbol{u})$ is global contracting and Regular at some initial value $\boldsymbol{x}_0^*$, then $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n, \cdots$ consistently samples π.*

Proof. We use Theorem 4.1.1 to prove the result. Let $\mathcal{F}$ be the collection of all bounded, uniformly continuous functions on Ω under the metric $d(\cdot,\cdot)$. We prove that $\forall f \in \mathcal{F}$, (4.1.1) is satisfied. By Portmanteau Theorem(cf. [11]), $\mathcal{F}$ is a distribution determining class on $(\Omega, \mathscr{B})$. Notice that the boundedness and continuity don't depend on which metric we are using, since both $d(\cdot,\cdot)$ and Euclidean norm give the same topology.

Let $\boldsymbol{x}_i^* = \phi(\boldsymbol{x}_{i-1}^*, \boldsymbol{u}_i)$ be the point generated by the Markov Chain updating starting from initial value $\boldsymbol{x}_0^*$. Without loss of generality we assume $|f| \leq 1$. The key idea

is to show that as time goes the initial value matters less and less.

$$\sup_{\widetilde{\boldsymbol{x}}_0\in\Omega}\left|\frac{1}{m}\sum_{i=1}^{m}f(\widetilde{\boldsymbol{x}}_i)-\mathbf{E}_\pi f\right|\le\left|\frac{1}{m}\sum_{i=1}^{m}f(\boldsymbol{x}_i^*)-\mathbf{E}_\pi f\right|+\sup_{\widetilde{\boldsymbol{x}}_0\in\Omega}\left|\frac{1}{m}\sum_{i=1}^{m}\left(f(\widetilde{\boldsymbol{x}}_i)-f(\boldsymbol{x}_i^*)\right)\right| \tag{4.3.10}$$

$$=\Sigma_1+\Sigma_2 \tag{4.3.11}$$

For Σ_1, as before we can use the Strong Law of Large Numbers. And for Σ_2 we resort to the global contracting property. Assuming irreducibility and Harris recurrence, by Strong Law of Large Numbers, $\forall\epsilon>0$

$$\mathcal{B}_m(\Sigma_1)\triangleq\left\{(\boldsymbol{u}_1,\boldsymbol{u}_2,\cdots,\boldsymbol{u}_m)\in[0,1)^{dm}:\left|\frac{1}{m}\sum_{i=1}^{m}f(\boldsymbol{x}_i^*)-\mathbf{E}_\pi f\right|>\frac{\epsilon}{2}\right\}, \tag{4.3.12}$$

$$\lambda_{dm}(\mathcal{B}_m(\Sigma_1))\to 0 \text{ when } m\to\infty.$$

Notice by the definition of ϕ being regular at $\boldsymbol{x}_0^*$, $\mathcal{B}_m(\Sigma_1)$ is Jordan measurable for all ϵ except for a null set $\mathcal{N}$. Using the same argument as in the proof of Theorem 4.3.3, we can assume $\mathcal{B}_m(\Sigma_1)$ to be Jordan measurable.

Next we look at Σ_2. Since we are assuming f to be uniformly continuous under $d(\cdot,\cdot)$, there exists $\delta>0$ such that $\forall\boldsymbol{y},\boldsymbol{z}\in\Omega, d(\boldsymbol{y},\boldsymbol{z})<\delta\Rightarrow|f(\boldsymbol{y})-f(\boldsymbol{z})|<\frac{\epsilon}{4}$. Recall $L_i=\ell(\boldsymbol{u}_i)$. By the Strong Law of Large Numbers,

$$\frac{1}{m}\sum_{i=1}^{m}\log(L_i)\to\int_{[0,1)^d}\log(\ell(\boldsymbol{u}))\,\mathrm{d}\boldsymbol{u}<0, a.s.$$

Let d_Ω be the diameter of Ω under $d(\cdot,\cdot)$. Then

$$\Sigma_2\le\frac{2}{m}\sum_{i=1}^{m}\mathbf{1}_{d(\widetilde{\boldsymbol{x}}_i,\boldsymbol{x}_i^*)\ge\delta}+\frac{\epsilon}{4}\frac{1}{m}\sum_{i=1}^{m}\mathbf{1}_{d(\widetilde{\boldsymbol{x}}_i,\boldsymbol{x}_i^*)<\delta} \tag{4.3.13}$$

$$\le\frac{2}{m}\sum_{i=1}^{m}\mathbf{1}_{d(\widetilde{\boldsymbol{x}}_i,\boldsymbol{x}_i^*)\ge\delta}+\frac{\epsilon}{4} \tag{4.3.14}$$

By the definition of $L_i = \ell(\boldsymbol{u}_i)$, it is easy to see that

$$d(\boldsymbol{x}_i, \boldsymbol{x}_i^*) \leq L_i d(\boldsymbol{x}_{i-1}, \boldsymbol{x}_{i-1}^*) \Rightarrow d(\boldsymbol{x}_i, \boldsymbol{x}_i^*) \leq d_\Omega \prod_{j=1}^{i} L_j.$$

Hence we can bound Σ_2 by:

$$\Sigma_2 \leq \frac{2}{m} \sum_{i=1}^{m} \mathbf{1}_{\left\{\prod_{j=1}^{i} L_j > \frac{\delta}{d_\Omega}\right\}} + \frac{\epsilon}{4} \tag{4.3.15}$$

$$= \frac{2}{m} \sum_{i=1}^{m} \mathbf{1}_{\left\{\sum_{j=1}^{i} \log(L_j) > \log(\frac{\delta}{d_\Omega})\right\}} + \frac{\epsilon}{4}. \tag{4.3.16}$$

Define

$$\mathcal{B}_m(\Sigma_2) = \left\{ (\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots, \boldsymbol{u}_m) : \sum_{j=1}^{i} \log\left(\ell(\boldsymbol{u}_j)\right) > \log\left(\frac{\delta}{d_\Omega}\right) \text{ for some } i \in [\frac{\epsilon}{8}m, m] \right\}. \tag{4.3.17}$$

By the Strong Law of Large Numbers, $\mathbf{Pr}\left(\mathcal{B}_m(\Sigma_2)\right) \to 0$ when $m \to \infty$. By the almost everywhere continuity of $\ell(\boldsymbol{u})$ and Lemma 4.3.5, we know $\mathcal{B}_m(\Sigma_2)$ is Jordan measurable for almost all δ. Here we simply assume the Jordan measurability to hold, otherwise we can choose a smaller δ to make it so.

Combining the bounds on Σ_1 and Σ_2, we have $\forall(\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots, \boldsymbol{u}_m) \notin \mathcal{B}_m\left(\Sigma_1\right) \bigcup \mathcal{B}_m\left(\Sigma_2\right)$,

$$\begin{aligned} \sup_{\widetilde{\boldsymbol{x}}_0 \in \Omega} \left| \frac{1}{m} \sum_{i=1}^{m} f(\widetilde{\boldsymbol{x}}_i) - \mathbf{E}_\pi f \right| &\leq \Sigma_1 + \Sigma_2 \\ &\leq \frac{\epsilon}{2} + \frac{2}{m} \times \frac{\epsilon}{8} m + \frac{\epsilon}{4} \\ &< \epsilon. \end{aligned}$$

Define $\mathcal{B}_m(\epsilon) = \mathcal{B}_m\left(\Sigma_1\right) \bigcup \mathcal{B}_m\left(\Sigma_2\right)$ which is Jordan measurable and $Vol(\mathcal{B}_m(\epsilon)) \to 0$, we proved 4.1.1 holds. Invoking Theorem 4.1.1, we proved that $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots$ consistently samples π. □

4.4 Examples

In this section we give twol examples to show the inconsistency of MCQMC if the conditions in the previous Theorems are not satisfied.

Example 4.4.1 (Why Boundedness is Important)**.** Consider the AR(1) process:

$$X_0 = 0, X_{i+1} = \alpha X_i + \epsilon_{i+1}$$

where $\epsilon_i \sim \mathcal{N}(0,1)$ is generated by inverting the CDF of standard normal distribution. I.e, $\epsilon_i = \Phi^{-1}(v_i)$. Assume (v_i) is a CUD sequence. It's easy to twist (v_i) a little bit while keeping the CUD property. For instance, we can change $v_{2^k} \to v'_{2^k}, k = 1, 2, \cdots$, and keep all the other numbers. Apparently it is not going to affect the CUD property. And we can make v'_{2^k} as small as possible so that $X_i \leq 0$ for all i, which implies (X_i) is not consistently sampling the stationary distribution. However, if we constrain the innovation ϵ_i to have compact support, by Theorem 4.3.7, we can show (X_i) consistently samples the stationary distribution.

Example 4.4.2 (Why Contracting Mapping is Important)**.** $\Omega = \{z \in \mathbb{C} : |z| = 1\}$ and $z_{i+1} = z_i^2 \times e^{2\pi\sqrt{-1}v_{i+1}}$ where $v_{i+1} \sim \mathcal{U}[0,1)$. From a Markov Chain point of view, z_i are IID and following uniform distribution on Ω. However, we can construct a CUD sequence which makes the simulation fail. For arbitrary CUD sequence $(v_i)_{i\geq 1}$, we are going to change $(v_{2^k+1})_{k\geq 0}$ and keep all the other numbers. Apparently such modification is not going to change the CUD property of the sequence. We construct the sequence $(z_i)_{i\geq 0}$ as follows:

For any CUD sequence $(v_i)_{i\geq 1}$, we would like to enforce $Re(z_i) \geq 0$ for all $i \geq 1$ by twisting the subsequence $(v_{2^k+1})_{k\geq 0}$. First we set $z_{2^k} = 1$ for all $k \geq 0$. Then by the equation $z_{i+1} = z_i^2 \times e^{2\pi\sqrt{-1}v_{i+1}}$, we can solve backwardly $z_{2^k-j} : 1 \leq j \leq 2^{k-1} - 1$. Notice for any z_{i+1}, the equation has two roots, and we can choose z_i to be the one with non-negative real part. For instance, by fixing $z_8 = 1$, we can get $z_7 = \pm\sqrt{e^{-2\pi\sqrt{-1}v_8}}$ and we choose z_7 to be the root with non-negative real part. After solving z_7, we can continue and solve z_6, and keep doing this backwardly until we find z_5. Notice in this

step we are not changing any of the driving sequence (v_i). Such an algorithm makes sure that $Re(z_i) \geq 0$ for all $i \geq 1$. Also, this algorithm guarantees that the equation $z_{i+1} = z_i^2 \times e^{2\pi\sqrt{-1}v_{i+1}}$ is satisfied for $i \neq 2^k, k \geq 0$. Now we only need to enforce $z_{2^k+1} = z_{2^k}^2 \times e^{2\pi\sqrt{-1}v_{2^k+1}} = e^{2\pi\sqrt{-1}v_{2^k+1}}$ by changing $v_{2^k+1} \mapsto v'_{2^k+1} = \frac{1}{2\pi}Arg(z_{2^k+1})$. Therefore, we only modified the values of a negeligible subsquence $(v_{2^k+1})_{k\geq 0}$ and we enforced $Re(z_i) \geq 0$ for all $i \geq 1$, while the stationary distribution is the uniform distribution on Ω. This example demonstrates that even for a MCMC algorithm on a nice region with a smooth update function which is essentially IID sampling, MCQMC could still fail without the contraction property.

Chapter 5

Relaxation of Global Contracting Mapping

In this chapter we continue our discussion about consistency of MCQMC algorithm on continuous state spaces. From last chapter we have seen, the consistency is guaranteed under the global contracting condition. However, this condition is too strong in two ways: first it requires the expectation of log of Lipshitz constant to be strictly less than zero. Secondly, for fixed $\boldsymbol{u}$, $\ell(\boldsymbol{u})$ is the global Lipshitz constant, measuring the universal continuity of $\phi(\boldsymbol{x}, \boldsymbol{u})$ for a given $\boldsymbol{u}$. In this chapter we will loosen the global contracting condition in both ways.

Throughout this chapter we assume the state space $\Omega \subseteq \mathbb{R}^s$ is a nice region equipped with Borel $\sigma-$ algebra $\mathscr{B}$.

5.1 Global Non-Expansive Mapping

Definition 5.1.1. *An update function $\phi(\boldsymbol{x}, \boldsymbol{u})$ is called Global Non Expansive, if*

$$\int_{[0,1)^d} \log(\ell(\boldsymbol{u}))\, \mathrm{d}\boldsymbol{u} \leq 0, \quad \ell(\boldsymbol{u}) \text{ as defined in (4.3.9).} \tag{5.1.1}$$

Definition 5.1.2. *For a metric space (Ω, d), the Bracketing Number $N(\delta)$ is defined*

as:

$$N(\delta) \triangleq \inf\{k: \text{ There exists } (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_k) \in \Omega \text{ such that } \bigcup_{j=1}^{k} B(\boldsymbol{x}_j, \delta) = \Omega\}$$

where $B(\boldsymbol{x}, \delta) = \{\boldsymbol{y} : d(\boldsymbol{x}, \boldsymbol{y}) < \delta\}$. In another word, $N(\delta)$ is the minimum number of balls with radius δ that can cover the whole space.

Remark. The Bracketing Number is in some sense describing the dimension of the metric space (Ω, d). A related but not the same concept is Hausdorff Dimension.

Theorem 5.1.3. *Assume the sample space is $(\Omega, \mathcal{B})$ where $\Omega \subseteq \mathbb{R}^s$ and $\mathscr{B}$ is the Borel σ algebra and let $(v_i)_{i\geq 1}$ be a CUD sequence. Let $\boldsymbol{x}_0 \in \Omega$ and $\boldsymbol{x}_i = \phi(\boldsymbol{x}_{i-1}, \widetilde{\boldsymbol{u}}_i)$ for $i \geq 1$ where ϕ is a Markov Chain update function. Assume the Markov Chain is irreducible and Harris recurrent with stationary distribution π and $\widetilde{\boldsymbol{u}}_i = (v_{di-d+1}, \cdots, v_{di})$. Let $d(\cdot, \cdot)$ be a metric defined on Ω which gives the same topology as Euclidean norm. Assume Ω is totally bounded under $d(\cdot, \cdot)$ and there exist $\gamma > 0$ such that $N(\delta) = O(\delta^{-\gamma})$. Furthermore assume the Markov Chain satisfies:*

$$\mathbf{Pr}_{\widetilde{\boldsymbol{y}}_0}\left(\phi_M(\widetilde{\boldsymbol{y}}_0; \boldsymbol{u}_M, \cdots, \boldsymbol{u}_1) \in \cdot\right) \geq \lambda\mu(\cdot), \text{ when } \boldsymbol{u}_j \sim \mathcal{U}[0,1)^d, \text{ for some } M \in \mathbb{N}, \lambda > 0. \tag{5.1.2}$$

for any $\widetilde{\boldsymbol{y}}_0 \in \Omega$ where $\mu(\cdot)$ is a probability measure on Ω. Assume $\ell(\boldsymbol{u})$ is continuous almost everywhere. If $\phi(\boldsymbol{x}, \boldsymbol{u})$ is global non expansive and regular, then $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n, \cdots$ consistently samples π.

Proof. We still resort to Theorem 4.1.1 for the proof. Here we choose, the same as the proof in the case of global contracting mapping, $\mathcal{F}$ = all uniform Lipshitz continuous functions on Ω with respect to $d(\cdot, \cdot)$. We just need to show that for any $f \in \mathcal{F}$ and any $\epsilon > 0$, we can find $\mathcal{B}_m(\epsilon) \subseteq [0,1)^{dm}$ Jordan measurable, $Vol(\mathcal{B}_m(\epsilon)) \to 0$ such

that

$$\mathcal{A}_m(\epsilon) \triangleq \left\{ (\boldsymbol{u}_1, \cdots, \boldsymbol{u}_m) \in [0,1)^{dm} : \sup_{\widetilde{\boldsymbol{x}}_0 \in \Omega} \left| \frac{1}{m} \sum_{i=1}^{m} f(\widetilde{\boldsymbol{x}}_i) - \mathbf{E}_\pi f \right| > \epsilon \right\} \subseteq \mathcal{B}_m(\epsilon)$$

where $\widetilde{\boldsymbol{x}}_i = \phi_i(\widetilde{\boldsymbol{x}}_0; \boldsymbol{u}_i, \cdots, \boldsymbol{u}_1)$. Without loss of generality, we assume $\mathbf{E}_\pi f = 0$, $|f| \leq 1$ and f has Lipschitz constant 1, i.e, $|f(\boldsymbol{x}) - f(\boldsymbol{y})| \leq d(\boldsymbol{x}, \boldsymbol{y})$.

By the definition of Bracketing number, there exists $\Omega_\delta = \{\boldsymbol{y}_0^1, \boldsymbol{y}_0^2, \cdots, \boldsymbol{y}_0^{N(\delta)} \in \Omega\}$ such that $\Omega = \bigcup_{j=1}^{N(\delta)} B(\boldsymbol{y}_0^j, \delta)$. In another word, $\forall \widetilde{\boldsymbol{x}}_0 \in \Omega$, there exists $\widetilde{\boldsymbol{y}}_0 \in \Omega_\delta$ such that $d(\widetilde{\boldsymbol{x}}_0, \widetilde{\boldsymbol{y}}_0) < \delta$. Thus we have the following inequality:

$$\sup_{\widetilde{\boldsymbol{x}}_0 \in \Omega} \left| \frac{1}{m} \sum_{i=1}^{m} f(\widetilde{\boldsymbol{x}}_i) \right| \leq \sup_{\widetilde{\boldsymbol{y}}_0 \in \Omega_\delta} \left| \frac{1}{m} \sum_{i=1}^{m} f(\widetilde{\boldsymbol{y}}_i) \right| + \sup_{d(\widetilde{\boldsymbol{x}}_0, \widetilde{\boldsymbol{y}}_0) < \delta} \left| \frac{1}{m} \sum_{i=1}^{m} (f(\widetilde{\boldsymbol{x}}_i) - f(\widetilde{\boldsymbol{y}}_i)) \right| \quad (5.1.3)$$

where $\widetilde{\boldsymbol{x}}_i = \phi_i(\widetilde{\boldsymbol{x}}_0; \boldsymbol{u}_i, \cdots, \boldsymbol{u}_1)$ and $\widetilde{\boldsymbol{y}}_i = \phi_i(\widetilde{\boldsymbol{y}}_0; \boldsymbol{u}_i, \cdots, \boldsymbol{u}_1)$.

Hence, we can decompose $\mathcal{A}_m(\epsilon)$ into the following two parts:

$$\begin{aligned} \mathcal{A}_m(\epsilon) \subseteq & \left\{ (\boldsymbol{u}_1, \cdots, \boldsymbol{u}_m) \in [0,1)^{dm} : \sup_{\widetilde{\boldsymbol{y}}_0 \in \Omega_\delta} \left| \frac{1}{m} \sum_{i=1}^{m} f(\widetilde{\boldsymbol{y}}_i) \right| > \frac{\epsilon}{2} \right\} \\ & \bigcup \left\{ (\boldsymbol{u}_1, \cdots, \boldsymbol{u}_m) \in [0,1)^{dm} : \sup_{d(\widetilde{\boldsymbol{x}}_0, \widetilde{\boldsymbol{y}}_0) < \delta} \left| \frac{1}{m} \sum_{i=1}^{m} (f(\widetilde{\boldsymbol{x}}_i) - f(\widetilde{\boldsymbol{y}}_i)) \right| > \frac{\epsilon}{2} \right\} \quad (5.1.4) \\ & = \mathcal{A}_m^1(\epsilon) \bigcup \mathcal{A}_m^2(\epsilon). \quad (5.1.5) \end{aligned}$$

We deal with $\mathcal{A}_m^1(\epsilon)$ and $\mathcal{A}_m^2(\epsilon)$ separately as below.

Step 1: First we would like to prove, by suitably choosing δ, $\lambda_{dm}(\mathcal{A}_m^1(\epsilon)) \to 0$ as $m \to \infty$. For any $\widetilde{\boldsymbol{y}}_0 \in \Omega_\delta$, we are going to use Hoeffding's inequality, and use Union Bound to handle the supreme. By assumption (5.1.2), there exists $\lambda > 0$ and a probability measure μ on Ω, such that for some $M \in \mathbb{N}$,

$$\mathbf{Pr}_{\widetilde{\boldsymbol{y}}_0}(\widetilde{\boldsymbol{y}}_M \in \cdot) \geq \lambda \mu(\cdot)$$

for any $\widetilde{\boldsymbol{y}}_0 \in \Omega$. This condition is closely related to Doeblin recurrence and uniform ergodicity. (cf. Tweedie and Meyn [28]). By Theorem 2 of Peter Glynn and Dirk Ormoneit (2002 [12]), we have the following Hoeffding's inequality:

$$\mathbf{Pr}_{\widetilde{\boldsymbol{y}}_0}\left(\sum_{i=1}^{m} f(\widetilde{\boldsymbol{y}}_i) - \mathbf{E}\left(\sum_{i=1}^{m} f(\widetilde{\boldsymbol{y}}_i)\right) \geq m\epsilon\right) \leq \exp\left(-\frac{\lambda^2(m\epsilon - 2M/\lambda)_+^2}{2mM^2}\right) \tag{5.1.6}$$

and by plugging $-f$ into the inequality above, it's easy to get:

$$\mathbf{Pr}_{\widetilde{\boldsymbol{y}}_0}\left(\left|\sum_{i=1}^{m} f(\widetilde{\boldsymbol{y}}_i) - \mathbf{E}\left(\sum_{i=1}^{m} f(\widetilde{\boldsymbol{y}}_i)\right)\right| \geq m\epsilon\right) \leq 2\exp\left(-\frac{\lambda^2(m\epsilon - 2M/\lambda)_+^2}{2mM^2}\right). \tag{5.1.7}$$

From this inequality, we can give an exponential decaying bound on the large deviation. By Rosenthal ([42]), we know that

$$\left|\mathbf{E}_{\widetilde{\boldsymbol{y}}_0} f(\widetilde{\boldsymbol{y}}_m)\right| \leq (1-\lambda)^{\lfloor m/M \rfloor}$$

Hence

$$\left|\sum_{i=1}^{m} \mathbf{E}_{\widetilde{\boldsymbol{y}}_0} f(\widetilde{\boldsymbol{y}}_i)\right| \leq \sum_{i=1}^{m} (1-\lambda)^{\lfloor i/M \rfloor} \leq C$$

where C is a constant only depend on (M, λ). Therefore,

$$\begin{aligned}
\mathbf{Pr}_{\widetilde{\boldsymbol{y}}_0}\left(\left|\sum_{i=1}^{m} f(\widetilde{\boldsymbol{y}}_i)\right| \geq m\epsilon\right) &\leq \mathbf{Pr}_{\widetilde{\boldsymbol{y}}_0}\left(\left|\sum_{i=1}^{m} f(\widetilde{\boldsymbol{y}}_i) - \mathbf{E}\left(\sum_{i=1}^{m} f(\widetilde{\boldsymbol{y}}_i)\right)\right| \geq m\epsilon - C\right) \\
&= \mathbf{Pr}_{\widetilde{\boldsymbol{y}}_0}\left(\left|\sum_{i=1}^{m} f(\widetilde{\boldsymbol{y}}_i) - \mathbf{E}\left(\sum_{i=1}^{m} f(\widetilde{\boldsymbol{y}}_i)\right)\right| \geq m(\epsilon - \frac{C}{m})\right) \\
&\leq 2\exp\left(-\frac{\lambda^2(m\epsilon - C - 2M/\lambda)_+^2}{2mM^2}\right) \\
&\leq \exp(-\alpha m)
\end{aligned}$$

when m is large enough if we choose $\alpha = \frac{\lambda^2\epsilon^2}{4M^2}$

Hence, there exists $\alpha > 0$ which only depends on (M, λ, ϵ) such that for large enough (which also only depends on (M, λ, ϵ)) m ,

$$\mathbf{Pr}_{\widetilde{\boldsymbol{y}}_0}\left(\left|\sum_{i=1}^{m} f(\widetilde{\boldsymbol{y}}_i)\right| \geq m\epsilon\right) \leq \exp(-\alpha m). \tag{5.1.8}$$

Plugging $\epsilon' = \frac{\epsilon}{2}$ into (5.1.8), we get an upper bound on $\lambda_{dm}(\mathcal{A}_m^1(\epsilon))$:

$$\begin{aligned}\lambda_{dm}(\mathcal{A}_m^1(\epsilon)) &= \mathbf{Pr}\left(\sup_{\widetilde{\boldsymbol{y}}_0 \in \Omega_\delta}\left|\frac{1}{m}\sum_{i=1}^{m} f(\widetilde{\boldsymbol{y}}_i)\right| > \frac{\epsilon}{2}\right) \\ &\leq |\Omega_\delta| \times \exp(-\alpha m) \\ &= O(\delta^{-\gamma}) \times \exp(-\alpha m)\end{aligned}$$

where $\alpha > 0$ depends on (M, λ, ϵ) and when m is large. If we choose $\delta^* = \exp(-\frac{\alpha m}{2\gamma})$, we have

$$\lambda_{dm}(\mathcal{A}_m^1(\epsilon)) = \mathbf{Pr}\left(\sup_{\widetilde{\boldsymbol{y}}_0 \in \Omega_{\delta^*}}\left|\frac{1}{m}\sum_{i=1}^{m} f(\widetilde{\boldsymbol{y}}_i)\right| > \frac{\epsilon}{2}\right) = O\left(\exp\left(-\frac{\alpha}{2}m\right)\right) \to 0. \tag{5.1.9}$$

Hence we have proved that $\mathcal{A}_m^1(\epsilon)$ has vanishing Lebesgue measure if $\delta^* = \exp(-\frac{\alpha m}{2\gamma})$. The remaining job is to show $\mathcal{A}_m^1(\epsilon)$ is actually Jordan measurable. By assuming ϕ to be regular, we know $\frac{1}{m}\sum_{i=1}^{m} f(\widetilde{\boldsymbol{y}}_i)$ is Riemann integrable for any initial state $\widetilde{\boldsymbol{y}}_0 \in \Omega_\delta$. Therefore by proposition 3.1.3, when δ^* is fixed and Ω_{δ^*} has been choosen, $\mathcal{A}_m^1(\epsilon)$ is Jordan measurable for all $\epsilon > 0$ except for a null set. For simplicity we can assume $\mathcal{A}_m^1(\epsilon)$ to be Jordan measurable, otherwise we can find $0 < \tilde{\epsilon} < \epsilon$ such that the Jordan measurability holds meanwhile keeping the volume vanishing:

$$\mathbf{Pr}\left(\sup_{\widetilde{\boldsymbol{y}}_0 \in \Omega_{\delta^*}}\left|\frac{1}{m}\sum_{i=1}^{m} f(\widetilde{\boldsymbol{y}}_i)\right| > \frac{\tilde{\epsilon}}{2}\right) \to 0.$$

Step 2: For this step we try to compute the volume of $\mathcal{A}_m^2(\epsilon)$. By assumption, we

have the global non expansive property. Let $L_i = \ell(\boldsymbol{u}_i)$, then $\mathbf{E}\log(L_i) \leq 0, \forall i \geq 1$.

$$\mathcal{A}_m^2(\epsilon) = \left\{ (\boldsymbol{u}_1, \cdots, \boldsymbol{u}_m) : \sup_{d(\widetilde{\boldsymbol{x}}_0, \widetilde{\boldsymbol{y}}_0) < \delta^*} \left| \frac{1}{m} \sum_{i=1}^{m} (f(\widetilde{\boldsymbol{x}}_i) - f(\widetilde{\boldsymbol{y}}_i)) \right| > \frac{\epsilon}{2} \right\} \tag{5.1.10}$$

$$\subseteq \left\{ (\boldsymbol{u}_1, \cdots, \boldsymbol{u}_m) : \sup_{d(\widetilde{\boldsymbol{x}}_0, \widetilde{\boldsymbol{y}}_0) < \delta^*} \frac{1}{m} \sum_{i=1}^{m} d(\widetilde{\boldsymbol{x}}_i, \widetilde{\boldsymbol{y}}_i) > \frac{\epsilon}{2} \right\} \tag{5.1.11}$$

by assuming f is uniformly Lipshitz continuous with Lipshitz constant 1. Noticing $d(\widetilde{\boldsymbol{x}}_i, \widetilde{\boldsymbol{y}}_i) \leq \delta^* \times \prod_{j=1}^{i} L_j$ and $\delta^* = \exp(-\frac{\alpha m}{2\gamma})$, we have:

$$\begin{aligned}
\mathcal{A}_m^2(\epsilon) &\subseteq \left\{ (\boldsymbol{u}_1, \cdots, \boldsymbol{u}_m) : \frac{1}{m} \sum_{i=1}^{m} \left(\delta^* \prod_{j=1}^{i} L_j \right) > \frac{\epsilon}{2} \right\} \\
&= \left\{ (\boldsymbol{u}_1, \cdots, \boldsymbol{u}_m) : \frac{1}{m} \sum_{i=1}^{m} \left(\prod_{j=1}^{i} L_j \right) > \frac{\epsilon}{2} \exp\left(\frac{\alpha m}{2\gamma} \right) \right\} \\
&\subseteq \left\{ (\boldsymbol{u}_1, \cdots, \boldsymbol{u}_m) : \exists 1 \leq i \leq m : \prod_{j=1}^{i} L_j > \frac{\epsilon}{2} \exp\left(\frac{\alpha m}{2\gamma} \right) \right\} \\
&= \left\{ (\boldsymbol{u}_1, \cdots, \boldsymbol{u}_m) : \exists 1 \leq i \leq m : \sum_{j=1}^{i} \log(L_j) > \log\left(\frac{\epsilon}{2} \right) + \frac{\alpha}{2\gamma} m \right\} \\
&\triangleq \mathcal{B}_m^2(\epsilon)
\end{aligned}$$

By the Strong Law of Large Numbers, since $\mathbf{E}(\log(L_j)) \leq 0$, we have $\lambda_{dm}\left(\mathcal{B}_m^2(\epsilon)\right) \to 0$ as $m \to \infty$. To show $\mathcal{B}_m^2(\epsilon)$ is Jordan measurable, we use the same argument as in Step 1. By assuming $\ell(\boldsymbol{u})$ to be continuous almost surely, $\sum_{i=1}^{m} \left(\prod_{j=1}^{i} L_j \right)$ is also continuous almost surely. Therefore by Lemma 3.1.3, $\mathcal{B}_m^2(\epsilon)$ is Jordan measurable for all ϵ except for a Null set. We can assume the Jordan measurablity of $\mathcal{B}_m^2(\epsilon)$, otherwise we can find $0 < \tilde{\epsilon} < \epsilon$ to make it so. Thus we have proved that $\mathcal{B}_m^2(\epsilon)$ is Jordan measurable and has vanishing volume.

Finally, let $\mathcal{B}_m(\epsilon) = \mathcal{A}_m^1(\epsilon) \bigcup \mathcal{B}_m^2(\epsilon)$, which is Jordan measurable. $\mathcal{A}_m(\epsilon) \subseteq \mathcal{B}_m(\epsilon)$ and $Vol(\mathcal{B}_m(\epsilon)) \to 0$ as $m \to \infty$. Therefore we completed our proof. □

5.2 Contracting on Average

The global contracting property defined in the previous Chapter essentially implies two features of the Markov Chain: first of all, no matter how far two chains start with, as long as they are updated through the same driving sequence, they will eventually be arbitrarily close to each other. Secondly, after getting close to each other, they will keep being close. Our motivation is, for the second feature to hold, we don't quite need the update function ϕ to be "Global" contracting. Instead, under a weaker version of contracting mapping: contracting on average, we will show in this section that the second feature still holds.

As for the first feature, our motivation comes from systematic Gibbs Sampler on a rectangle $\Omega = [a, b] \times [c, d]$. Assume that the stationary distribution π is bounded from below and above, and we are using the inverse method to sample from conditional distribution. The algorithm needs two random numbers v_1, v_2 for one iteration and goes as follows: $(x, y) \mapsto (x', y)$ where $x' = F_y^{-1}(v_1)$ and $(x', y) \mapsto (x', y')$ where $y' = F_{x'}^{-1}(v_2)$. A key observation is, as long as $v_1 \approx 1$ and $v_2 \approx 1$, no matter what value (x, y) takes, (x', y') will be very close to the upper right corner (b, d). Therefore, although generally such a Gibbs Sampler algorithm is not necessarily global contracting, two chains starting from different initial values will still get close to each other, because of the existence of a "Corner" (b, d). This observation motivates the following definition:

Definition 5.2.1. *X_n is a Markov Chain on $\Omega \subseteq \mathbb{R}^s$ with update function $\phi(\boldsymbol{x}, \boldsymbol{u})$. $\boldsymbol{x}^* \in \Omega$ is called a* **Corner** *if: for any neighborhood of $\boldsymbol{x}^*$: $B(\boldsymbol{x}^*, \delta) = \{\boldsymbol{y} : \|\boldsymbol{x} - \boldsymbol{y}\| < \delta\}$, there exists $\mathcal{C}_\delta \subseteq [0, 1)^d$ Jordan measurable with positive volume, such that*

$$\phi(\boldsymbol{x}, \boldsymbol{u}) \in B(\boldsymbol{x}^*, \delta), \forall \boldsymbol{u} \in \mathcal{C}_\delta, \boldsymbol{x} \in \Omega.$$

First we prove two useful Lemmas:

Lemma 5.2.2. *$X_n = \phi(X_{n-1}, U_n) : U_n \sim \mathcal{U}[0, 1)^d$ is an irreducible, Harris recurrent*

Markov Chain on $(\Omega, \mathscr{F})$ with stationary distribution π, where ϕ is the update function. Let $Y_n = (X_n, U_{n+1})$. Then Y_n is an irreducible, Harris recurrent Markov Chain on $(\Omega \times [0,1)^d, \mathscr{F} \otimes \mathscr{B}[0,1)^d)$ with stationary distribution $\pi \otimes \mathcal{U}[0,1)^d$.

Proof. Define $\widetilde{\Omega} \triangleq \Omega \times [0,1)^d$, $\widetilde{\mathscr{F}} \triangleq \mathscr{F} \otimes \mathcal{U}[0,1)^d$ and $\widetilde{\pi} \triangleq \pi \otimes \mathcal{U}[0,1)^d$. First we prove Y_n is a Markov Chain on $(\widetilde{\Omega}, \widetilde{\mathscr{F}})$. This is quite obvious, since conditionally on $(Y_n, Y_{n-1}, \cdots, Y_0)$, $X_{n+1} = \phi(X_n, U_{n+1})$ is fully determined, and $U_{n+2} \sim \mathcal{U}[0,1)^d$ is independent of the past. Therefore the law of $Y_{n+1} = (X_{n+1}, U_{n+2})$ is fully determined by $Y_n = (X_n, U_{n+1})$, which proves that $(Y_n)_{n \geq 0}$ is a Markov Chain.

Second we prove Y_n has stationary distribution $\widetilde{\pi} \triangleq \pi \otimes \mathcal{U}[0,1)^d$. Assuming $X_n \sim \pi$ and $U_{n+1} \sim \mathcal{U}[0,1)^d$, by the definition of π being invariant, we know $X_{n+1} = \phi(X_n, U_{n+1}) \sim \pi$. Since U_{n+2} is independent of (X_n, U_{n+1}), we proved that (X_{n+1}, U_{n+2}) also follows distribution $\pi \otimes \mathcal{U}[0,1)^d$, which means $\pi \otimes \mathcal{U}[0,1)^d$ is the stationary distribution.

Thirdly we prove Y_n is $\widetilde{\pi}$ irreducible. We just need to show that, for any $A \in \widetilde{\mathscr{F}}$ with $\widetilde{\pi}(A) > 0$, and any $(\boldsymbol{x}_0, \boldsymbol{u}_1) \in \widetilde{\Omega}$, there exists $n \geq 1$ such that:

$$\mathbf{Pr}\left(Y_n \in A \middle| Y_0 = (\boldsymbol{x}_0, \boldsymbol{u}_1)\right) > 0.$$

By Fubini's Theorem, there exists $B \in \mathscr{F}, \pi(B) > 0$ such that for any $\boldsymbol{x} \in B$:

$$\lambda_d\left(\{\boldsymbol{u} : (\boldsymbol{x}, \boldsymbol{u}) \in A\}\right) > \frac{1}{m}$$

for some $m \in \mathbb{N}$. Since X_n is irreducible, Harris recurrent with stationary distribution π, we know there exists $n \geq 2$ such that

$$\mathbf{Pr}\left(X_n \in B \middle| X_1 = \phi(\boldsymbol{x}_0, \boldsymbol{u}_1)\right) > 0.$$

Therefore

$$\begin{aligned}
&\mathbf{Pr}\left(Y_n \in A \middle| Y_0 = (\boldsymbol{x}_0, \boldsymbol{u}_1)\right) \\
\geq &\mathbf{Pr}\left(X_n \in B, (X_n, U_{n+1}) \in A \middle| Y_0 = (\boldsymbol{x}_0, \boldsymbol{u}_1)\right) \\
\geq &\frac{1}{m}\mathbf{Pr}\left(X_n \in B \middle| Y_0 = (\boldsymbol{x}_0, \boldsymbol{u}_1)\right) \\
= &\frac{1}{m}\mathbf{Pr}\left(X_n \in B \middle| X_1 = \phi(\boldsymbol{x}_0, \boldsymbol{u}_1)\right) > 0.
\end{aligned}$$

Lastly we prove Y_n is Harris recurrent. By Theorem 6 of Roberts and Rosenthal ([40]), it's enough to show that for any $A \in \widetilde{\mathscr{F}}$, $\widetilde{\pi}(A) = 1$, and for any $(\boldsymbol{x}_0, \boldsymbol{u}_1) \in \Omega \times [0,1)^d$, we have

$$\mathbf{Pr}\left(\tau_A < \infty | Y_0 = (\boldsymbol{x}_0, \boldsymbol{u}_1)\right) = 1 \tag{5.2.1}$$

where $\widetilde{\tau}_A = \inf\{n \geq 1 : Y_n \in A\}$. Since $\widetilde{\pi}(A) = 1$, by Fubini's Theorem, there exists $B \in \mathscr{F} : \pi(B) = 1$ such that:

$$\forall \boldsymbol{x} \in B, \lambda_d\left(\{\boldsymbol{u} : (\boldsymbol{x}, \boldsymbol{u}) \in A\}\right) = 1. \tag{5.2.2}$$

Define $\tau_B = \inf\{n \geq 1 : X_n \in B\}$. Then,

$$\begin{aligned}
&\mathbf{Pr}\left(\widetilde{\tau}_A < \infty \middle| Y_0 = (\boldsymbol{x}_0, \boldsymbol{u}_1)\right) \\
\leq &\mathbf{Pr}\left(\exists n \geq 1 : X_n \in B, (X_n, U_{n+1}) \in A \middle| X_0 = \boldsymbol{x}_0, U_1 = \boldsymbol{u}_1\right) \\
= &\sum_{n=1}^{\infty} \mathbf{Pr}\left(\tau_B = n, X_n \in B, (X_n, U_{n+1}) \in A \middle| (\boldsymbol{x}_0, \boldsymbol{u}_1)\right) \\
= &\sum_{n=1}^{\infty} \mathbf{Pr}\left(\tau_B = n, X_n \in B \middle| (\boldsymbol{x}_0, \boldsymbol{u}_1)\right) \qquad \text{by equation (5.2.2)} \\
= &\mathbf{Pr}\left(\tau_B < \infty \middle| X_1 = \phi(\boldsymbol{x}_0, \boldsymbol{u}_1)\right) \\
= &1
\end{aligned}$$

by assuming X_n to be Harris recurrent.

Through the analysis above we have shown that Y_n is an irreducible Harris recurrent Markov Chain with stationary distribution $\widetilde{\pi}$. Our proof is complete.

□

Lemma 5.2.3. *$\{d_n\}$ is a sequence of positive real numbers satisfying the following recursive relation:*

$$d_{n+1} \leq \rho_n d_n + C d_n^2 \tag{5.2.3}$$

where $\rho_n \geq \epsilon > 0$ is a positive sequence bounded from below. C is a positive constant number. Assuming the following condition is satisfied:

$$\rho_{i-1} ... \rho_0 < M\rho^i \quad \forall i \geq 1 \tag{5.2.4}$$

for some constant $M > 1$ and $\rho \in (0, 1)$, then we have the following exponentially decaying result: $\forall S \in (1, \frac{1}{\rho})$:

$$d_n \leq \frac{(S-1)\epsilon}{C}(S\rho)^n \tag{5.2.5}$$

if initially

$$d_0 \leq \frac{(S-1)\epsilon}{MC}.$$

Proof. Define $e_n = \frac{d_n}{\rho_{n-1}\rho_{n-2}\cdots\rho_0}$. Then we can rewrite (5.2.3) as follows:

$$\begin{aligned} e_{n+1} &\leq e_n + C e_n^2 \frac{\rho_{n-1}\rho_{n-2}\cdots\rho_0}{\rho_n} \\ &\leq e_n + \frac{C \times M}{\epsilon} e_n^2 \rho^n \quad \text{by (5.2.4)} \\ &= e_n + C_1 \rho^n e_n^2 \end{aligned}$$

where $C_1 \triangleq \frac{CM}{\epsilon}$. Assume $d_0 = e_0 \leq K = \frac{S-1}{C_1} = \frac{(S-1)\epsilon}{MC} > 0$. We claim:

$$e_n \leq KS^n \tag{5.2.6}$$

This inequality is obviously true when $n = 0$. Assume it is true for n. Then

$$\begin{aligned} e_{n+1} &\leq e_n + C_1 \rho^n e_n^2 \\ &\leq KS^n + C_1 \rho^n K^2 S^{2n} \\ &= KS^{n+1} \left(\frac{1}{S} + C_1 \frac{K}{S} (\rho S)^n \right) \\ &\leq KS^{n+1} \frac{1 + C_1 K}{S} \\ &\leq KS^{n+1}. \end{aligned}$$

Using the inequality above, we have

$$\begin{aligned} d_n &= e_n \times \rho_{n-1} \times \rho_{n-2} \ldots \rho_0 \\ &\leq KMS^n \rho^n \\ &= \frac{(S-1)\epsilon}{C} (S\rho)^n \end{aligned}$$

which completes our proof. □

With the previous Lemmas, we can now show the following "coupling" result, which will lead to a proof of the consistency of MCQMC. The basic idea is, if we start the Markov Chain from the Corner $\boldsymbol{x}^*$ and some $\boldsymbol{y} \in B(\boldsymbol{x}^*, \delta)$, under some relaxed version of contracting mapping condition, the two chains should remain close to each other for all n. And $\boldsymbol{x}^*$ being a corner guarantees that no matter where we start, after a geometrically distributed random time, the Markov Chain will hit any arbitrary small neighborhood of the Corner $\boldsymbol{x}^*$, after which we can "Couple" the original chain with a Chain starting from $\boldsymbol{x}^*$.

The following Theorem is totally probabilistic. We will take care of the Jordan measurability later.

Theorem 5.2.4. *Assume X_n is an irreducible, Harris recurrent Markov Chain on a bounded convex space $\Omega \in \mathbb{R}^s$ with update function $\phi(\boldsymbol{x}, \boldsymbol{u})$. Let π be its stationary*

distribution. Assume ϕ is twice continuously differentiable w.r.t $\boldsymbol{x}$. Define:

$$\rho(\boldsymbol{x}, \boldsymbol{u}) = \left\| \frac{\partial \phi(\boldsymbol{x}, \boldsymbol{u})}{\partial \boldsymbol{x}} \right\|_2$$

where $\|\cdot\|_2$ denotes the operator L_2 norm. And

$$C \triangleq \frac{1}{s^3} \sup_{\boldsymbol{u}, \boldsymbol{x}, i, j, k} \left| \frac{\partial^2 \phi^i(\boldsymbol{x}, \boldsymbol{u})}{\partial x^j \partial x^k} \right| < \infty, \quad s = \text{ Dimension of } \Omega \tag{5.2.7}$$

where ϕ^i is the ith component of ϕ and x^j is the jth component of $\boldsymbol{x} \in \mathbb{R}^s$. If:

- *the update function satisfies the following* **Contracting on Average** *property:*

$$\mathbf{E}_{\pi \otimes \mathcal{U}[0,1)^d}(\log(\rho(\boldsymbol{x}, \boldsymbol{u}))) < 0 \tag{5.2.8}$$

  where π is the stationary distribution, and

- *the Markov Chain contains a* **corner** $\boldsymbol{x}^*$*,*

then we have the **coupling** *property:*

$$\forall \delta > 0, \mathbf{Pr}\left(\sup_{\boldsymbol{x}_0, \boldsymbol{y}_0} \|\boldsymbol{x}_n - \boldsymbol{y}_n\| > \delta \right) \to 0 \text{ as } n \to \infty. \tag{5.2.9}$$

Proof. First we would like to show that the set $\{\sup_{\boldsymbol{x}_0, \boldsymbol{y}_0} \|\boldsymbol{x}_n - \boldsymbol{y}_n\| > \delta\}$ is measurable. Since Ω is separable, we can find a dense countable subset $\Omega' \subseteq \Omega$. Fixing the driving sequence $(\boldsymbol{u}_i)_{i=1}^n$, $\boldsymbol{x}_n = \phi_n(\boldsymbol{x}_0; \boldsymbol{u}_n, \cdots, \boldsymbol{u}_1)$ is a continuous function of $\boldsymbol{x}_0$. Therefore,

$$\{\sup_{\boldsymbol{x}_0, \boldsymbol{y}_0} \|\boldsymbol{x}_n - \boldsymbol{y}_n\| > \delta\} = \{\sup_{\boldsymbol{x}_0, \boldsymbol{y}_0 \in \Omega'} \|\boldsymbol{x}_n - \boldsymbol{y}_n\| > \delta\}$$

which is a countable union of measurable sets, hence also measurable.

By the contracting on average assumption,

$$\mathbf{E}_{\pi \otimes \mathcal{U}[0,1)^d} \log\left(\rho(\boldsymbol{x}, \boldsymbol{u})\right) < 0$$

we can find $\epsilon > 0$ small enough and $\bar{\rho} \in (0,1)$ such that

$$\mathbf{E}_{\pi \otimes \mathcal{U}[0,1)^d} \left(\log(\rho(\boldsymbol{x}, \boldsymbol{u}) \vee \epsilon) \right) < \log \bar{\rho} < 0. \tag{5.2.10}$$

Denote $\boldsymbol{x}_i^* = \phi(\boldsymbol{x}^*; \boldsymbol{u}_i, \boldsymbol{u}_{i-1}, \cdots, \boldsymbol{u}_1)$ be the ith step iteration starting at $\boldsymbol{x}_0^* = \boldsymbol{x}^*$. Let $\rho_i = \rho(\boldsymbol{x}_i^*, \boldsymbol{u}_{i+1})$ and $\rho_i^\epsilon = \rho(\boldsymbol{x}_i^*, \boldsymbol{u}_{i+1}) \vee \epsilon$. Let $1 < S < \frac{1}{\bar{\rho}}$ (For example, $S = \frac{2}{1+\bar{\rho}}$), $M > 1$ is arbitrary, and $K = \frac{(S-1)\epsilon}{MC}$.

Step 1: First we assume the two starting points $\boldsymbol{x}_0 = \boldsymbol{x}^*$ and $\boldsymbol{y}_0$ lies within a small neighborhood of $\boldsymbol{x}^*$. We would like to prove the following result:

$$\begin{aligned}
&\mathbf{Pr}\left(\sup_{\boldsymbol{x}_0 = \boldsymbol{x}^*, \boldsymbol{y}_0 \in B(\boldsymbol{x}^*, K)} \|\boldsymbol{x}_n - \boldsymbol{y}_n\| \geq \frac{(S-1)\epsilon}{C} (S\bar{\rho})^n \right) \\
\leq &\mathbf{Pr}\left(\rho_{i-1}^\epsilon \rho_{i-2}^\epsilon \cdots \rho_0^\epsilon > M\bar{\rho}^i \text{ for some } 0 \leq i < \infty \right) \\
\triangleq &h(M) \to 0 \text{ as M } \to \infty.
\end{aligned} \tag{5.2.11}$$

Remember the two chains starting from $\boldsymbol{x}_0$ and $\boldsymbol{y}_0$ are driven by the same update random numbers $(\boldsymbol{u}_i)_{i \geq 1}$. Now let's look at how the distance between $\boldsymbol{x}_n$ and $\boldsymbol{y}_n$ evolves over time. Since $\phi(\boldsymbol{x}, \boldsymbol{u})$ is twice continuously differentiable w.r.t $\boldsymbol{x}$, we can do the Lagrange expansion of ϕ at $\boldsymbol{x}_n$:

$$y_{n+1}^i - x_{n+1}^i = \sum_j \left. \frac{\partial \phi^i}{\partial x^j} \right|_{\boldsymbol{x}_n, \boldsymbol{u}_n} (y_n^j - x_n^j) + \frac{1}{2} \sum_{j,k} \left. \frac{\partial^2 \phi^i}{\partial x^j x^k} \right|_{\boldsymbol{z}_n, \boldsymbol{u}_{n+1}} (y_n^j - x_n^j)(y_n^k - x_n^k)$$

for some $\boldsymbol{z}_n$ lying between $\boldsymbol{x}_n$ and $\boldsymbol{y}_n$. Notice here we need to assume Ω to be convex, otherwise it's not necessary that we can do the Lagrange Expansion. By some easy

calculation:

$$\begin{aligned} d_{n+1} &\stackrel{\Delta}{=} \|\boldsymbol{y}_{n+1} - \boldsymbol{x}_{n+1}\| \\ &= \|\phi(\boldsymbol{y}_n, \boldsymbol{u}_{n+1}) - \phi(\boldsymbol{x}_n, \boldsymbol{u}_{n+1})\| \\ &\leq \left\| \left. \frac{\partial \phi(\boldsymbol{x}, \boldsymbol{u})}{\partial \boldsymbol{x}} \right|_{(\boldsymbol{x}_n, \boldsymbol{u}_{n+1})} (\boldsymbol{y}_n - \boldsymbol{x}_n) \right\| + \frac{1}{2} s^2 \times s \times \sup_{\boldsymbol{u}, \boldsymbol{x}, i, j, k} \left| \frac{\partial^2 \phi^i(\boldsymbol{x}, \boldsymbol{u})}{\partial x^j \partial x^k} \right| \|\boldsymbol{y}_n - \boldsymbol{x}_n\|^2 \\ &\leq \rho(\boldsymbol{x}_n, \boldsymbol{u}_{n+1}) d_n + C d_n^2 \quad \text{by assumption (5.2.7).} \end{aligned}$$

Combining the initial condition $\boldsymbol{y}_0 \in B(\boldsymbol{x}^*, K)$, we get:

$$d_{n+1} \leq \rho_n \times d_n + C d_n^2, \quad d_0 \leq K = \frac{(S-1)\epsilon}{MC}$$

which is very similar to Lemma 5.2.3's condition, except that in Lemma 5.2.3 we need ρ_n to be bounded from below. Obviously we can make it so by taking the maximum of ρ_n and ϵ and get:

$$d_{n+1} \leq (\rho_n \vee \epsilon) \times d_n + C d_n^2.$$

By Lemma 5.2.3, we have the following exponential decay result:

$$d_n \leq \frac{(S-1)\epsilon}{C} (S\bar{\rho})^n \text{ on the set } \{(\boldsymbol{u}_1, \cdots, \boldsymbol{u}_n) : \rho_{i-1}^\epsilon \rho_{i-2}^\epsilon \dots \rho_0^\epsilon \leq M \bar{\rho}^i \quad \forall 1 \leq i \leq n\}. \tag{5.2.12}$$

We have already shown in Lemma 5.2.2, the augmented chain $(\boldsymbol{x}_i, \boldsymbol{u}_{i+1})_{i \geq 0}$ is irreducible and Harris recurrent. Therefore by the Strong Law of Large Numbers, for almost every sequence of $(\boldsymbol{u}_1, \boldsymbol{u}_2 ...)$ we have $(\boldsymbol{x}_i, \boldsymbol{u}_{i+1})_{i \geq 0}$ samples $\pi \otimes \mathcal{U}[0,1)^d$. Therefore for almost every sequence of $(\boldsymbol{u}_1, \boldsymbol{u}_2, ...)$, we have:

$$\frac{1}{n} \sum_{i=0}^{n-1} \log(\rho_i^\epsilon) \to \mathbf{E}_{\pi \otimes \mathcal{U}[0,1)^d} (\log(\rho(\boldsymbol{x}, \boldsymbol{u}) \vee \epsilon)) < \log \bar{\rho} < 0$$

which implies that

$$\mathbf{Pr}\left(\rho_{i-1}^\epsilon \rho_{i-2}^\epsilon \dots \rho_0^\epsilon \leq M \bar{\rho}^i \quad \forall i \geq 1\right) = h(M) \to 0 \quad \text{as } M \to \infty.$$

Hence we have proved inequality 5.2.11.

Step 2: For general $\boldsymbol{x}_0, \boldsymbol{y}_0 \in B(\boldsymbol{x}^*, K)$, we can easily apply the Triangle inequality and get the following similar result:

$$\begin{aligned}
&\mathbf{Pr}\left(\sup_{\boldsymbol{x}_0,\boldsymbol{y}_0\in B(\boldsymbol{x}^*,K)} ||\boldsymbol{x}_n - \boldsymbol{y}_n|| \geq 2\frac{(S-1)\epsilon}{C}(S\bar{\rho})^n\right)\\
\leq& 2\mathbf{Pr}\left(\rho^\epsilon_{i-1}\rho^\epsilon_{i-2}\dots\rho^\epsilon_0 > M\bar{\rho}^i \text{ for some } 0\leq i<\infty\right)\\
=& 2h(M) \to 0 \text{ as M } \to \infty.
\end{aligned}$$

Step 3: Now we are ready to prove the Coupling property (5.2.9). For any $M > 1$, define $B(\boldsymbol{x}^*, \frac{(S-1)\epsilon}{MC}) = B(\boldsymbol{x}^*, K_M)$. It is a neighborhood of $\boldsymbol{x}^*$, therefore by the assumption that $\boldsymbol{x}^*$ is a Corner, there exists $\mathcal{C}_M \subseteq [0,1)^d$ Jordan measurable, such that $\boldsymbol{u}_i \in \mathcal{C}_M \Rightarrow \boldsymbol{x}_i \in B(\boldsymbol{x}^*, K_M)$. Also notice that $Vol(\mathcal{C}_M) > 0$ by the definition of Corner.

Define $\tau = \inf\{i \geq 1 : \boldsymbol{u}_i \in \mathcal{C}_M\}$ to be the first hitting time of region $\mathcal{C}_M$, then τ is a stopping time, following Geometric distribution with parameter $Vol(\mathcal{C}_M) > 0$.

Then we can decompose the probability of "Non-Coupling" by τ as follows: for any $k > 0, k \in \mathbb{N}$,

$$\begin{aligned}
&\mathbf{Pr}\left(\sup_{(\boldsymbol{x}_0,\boldsymbol{y}_0)} \|\boldsymbol{x}_n - \boldsymbol{y}_n\| \geq 2\frac{(S-1)\epsilon}{C}(S\bar{\rho})^{n-k}\right)\\
\leq&\sum_{i=1}^{k-1}\mathbf{Pr}\left(\|\boldsymbol{x}_n - \boldsymbol{y}_n\| \geq 2\frac{(S-1)\epsilon}{C}(S\bar{\rho})^{n-k}, \tau = i\right) + \mathbf{Pr}(\tau \geq k)\\
\leq&\sum_{i=1}^{k-1}\mathbf{Pr}(\tau = i)\mathbf{Pr}\left(\sup_{\boldsymbol{x}_i,\boldsymbol{y}_i\in B(\boldsymbol{x}^*,K)} \|\boldsymbol{x}_n - \boldsymbol{y}_n\| \geq 2\frac{(S-1)\epsilon}{C}(S\bar{\rho})^{n-k}\right) + \mathbf{Pr}(\tau \geq k)\\
\leq& 2h(M) + (1 - Vol(\mathcal{C}_M))^{k-1}.
\end{aligned}$$

Letting $n \to \infty$ on both sides and taking lim sup, we get

$$\limsup_{n\to\infty} \mathbf{Pr}\left(\sup_{(\boldsymbol{x}_0,\boldsymbol{y}_0)} \|\boldsymbol{x}_n - \boldsymbol{y}_n\| \geq 2\frac{(S-1)\epsilon}{C}(S\bar{\rho})^{n-k}\right) \leq 2h(M) + (1 - Vol(\mathcal{C}_M))^{k-1}$$

for any $M > 1$, $k \in \mathbb{N}$. Since we are enforcing $1 < S < \frac{1}{\bar{\rho}}$, we have $S\bar{\rho}^{n-k} \to 0$ as $n \to \infty$, which implies:

$$\begin{aligned}
&\limsup_{n\to\infty} \mathbf{Pr}\left(\sup_{(\boldsymbol{x}_0,\boldsymbol{y}_0)} \|\boldsymbol{x}_n - \boldsymbol{y}_n\| \geq \delta\right) \\
\leq &\limsup_{n\to\infty} \mathbf{Pr}\left(\sup_{(\boldsymbol{x}_0,\boldsymbol{y}_0)} \|\boldsymbol{x}_n - \boldsymbol{y}_n\| \geq 2\frac{(S-1)\epsilon}{C}(S\bar{\rho})^{n-k}\right) \\
\leq &2h(M) + (1 - Vol(\mathcal{C}_M))^{k-1}.
\end{aligned}$$

As proved in Step 1, $h(M) \to 0$ as $M \to \infty$. Choosing k large enough so that $(1 - Vol(\mathcal{C}_M))^{k-1}$ is arbitrarily small, we get our desired result.

□

The previous Theorem tells us that under the existence of a corner and contraction on average, after significantly long time, no matter where the Markov Chain starts, the end point $\boldsymbol{x}_n$ forgets about the initial value. With this Theorem proved, we are ready to show the consistency of MCQMC under contracting on average.

Theorem 5.2.5. *Assume all the conditions in Theorem 5.2.4 to hold, and let* $\widetilde{\boldsymbol{u}}_i = (v_{di-i+1}, \cdots, v_{di})$ *for a CUD sequence* $(v_i)_{i\geq 1}$. *The MCQMC algorithm has initial value* $\boldsymbol{x}_0$ *and* $\boldsymbol{x}_i = \phi(\boldsymbol{x}_{i-1}, \widetilde{\boldsymbol{u}}_i)$. *Further assume* ϕ *is Regular, and* $\sup_{\boldsymbol{x},\boldsymbol{u}} \rho(\boldsymbol{x}, \boldsymbol{u}) < \infty$. *Then* $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots$ *consistently samples* π.

Proof. We adopt the same strategy as before, resorting to Theorem 4.1.1 for the proof. Let $\mathcal{F}$ be the set of all uniform Lipshitz continuous functions on Ω with respect to the Euclidean metric. As before, we just need to show for any $f \in \mathcal{F}$ and any $\epsilon > 0$,

we can find $\mathcal{B}_m(\epsilon)$ Jordan measurable, $Vol(\mathcal{B}_m(\epsilon)) \to 0$ such that

$$\mathcal{A}_m(\epsilon) \triangleq \left\{ (\boldsymbol{u}_1, \cdots, \boldsymbol{u}_m) \in [0,1)^{dm} : \sup_{\widetilde{\boldsymbol{x}}_0 \in \Omega} \left| \frac{1}{m} \sum_{i=1}^{m} f(\widetilde{\boldsymbol{x}}_i) - \mathbf{E}_\pi f \right| > \epsilon \right\} \subseteq \mathcal{B}_m(\epsilon).$$

Without loss of generality, we assume $\mathbf{E}_\pi f = 0$, $|f| \leq 1$ and f has Lipschitz constant 1, i.e, $|f(\boldsymbol{x}) - f(\boldsymbol{y})| \leq \|\boldsymbol{x} - \boldsymbol{y}\|$. We would like to couple a chain from any starting point $\widetilde{\boldsymbol{x}}_0$ with a chain starting from the corner $\boldsymbol{x}^*$. Let $\boldsymbol{x}_0^* = \boldsymbol{x}^*$ (the corner) and $\boldsymbol{x}_n^* = \phi(\boldsymbol{x}_{n-1}^*, \boldsymbol{u}_n)$ be the nth iteration. Then,

$$\sup_{\widetilde{\boldsymbol{x}}_0 \in \Omega} \left| \frac{1}{m} \sum_{i=1}^{m} f(\widetilde{\boldsymbol{x}}_i) - \mathbf{E}_\pi f \right| \leq \left| \frac{1}{m} \sum_{i=1}^{m} f(\boldsymbol{x}_i^*) \right| + \sup_{\widetilde{\boldsymbol{x}}_0 \in \Omega} \left| \frac{1}{m} \sum_{i=1}^{m} (f(\widetilde{\boldsymbol{x}}_i) - f(\boldsymbol{x}_i^*)) \right| \quad (5.2.13)$$

$$= \Sigma_1 + \Sigma_2, \quad (5.2.14)$$

by assuming $\mathbf{E}_\pi f = 0$. For Σ_1, by the Strong Law of Large Numbers, $\forall \epsilon > 0$

$$\mathcal{B}_m(\Sigma_1) \triangleq \left\{ (\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots, \boldsymbol{u}_m) \in [0,1)^{dm} : \left| \frac{1}{m} \sum_{i=1}^{m} f(\boldsymbol{x}_i^*) \right| > \frac{\epsilon}{2} \right\}, \text{ and} \quad (5.2.15)$$

$$\lambda_{dm}(\mathcal{B}_m(\Sigma_1)) \to 0 \text{ as } m \to \infty.$$

For Σ_2, we use the Coupling property proved in Theorem 5.2.4.

$$\begin{aligned}
\mathbf{E}(\Sigma_2) &= \mathbf{E}\left(\sup_{\widetilde{\boldsymbol{x}}_0 \in \Omega} \left| \frac{1}{m} \sum_{i=1}^{m} (f(\widetilde{\boldsymbol{x}}_i) - f(\boldsymbol{x}_i^*)) \right| \right) \\
&\leq \frac{1}{m} \sum_{i=1}^{m} \mathbf{E}\left(\sup_{\widetilde{\boldsymbol{x}}_0 \in \Omega} |(f(\widetilde{\boldsymbol{x}}_i) - f(\boldsymbol{x}_i^*))| \right) \\
&\leq \frac{1}{m} \sum_{i=1}^{m} \left(\delta \mathbf{Pr}\left(\sup_{\widetilde{\boldsymbol{x}}_0 \in \Omega} \|\widetilde{\boldsymbol{x}}_i - \boldsymbol{x}_i^*\| \leq \delta \right) + 1 \times \mathbf{Pr}\left(\sup_{\widetilde{\boldsymbol{x}}_0 \in \Omega} \|\widetilde{\boldsymbol{x}}_i - \boldsymbol{x}_i^*\| > \delta \right) \right) \\
&\quad \text{by Lipschitz Continuity and Boundedness of f} \\
&\overset{m \to \infty}{\longrightarrow} \delta \quad \text{by Theorem 5.2.4}
\end{aligned}$$

for any $\delta > 0$. Therefore $\mathbf{E}(\Sigma_2) \to 0$ as $m \to \infty$. By Markov inequality:

$$\mathbf{Pr}\left(\Sigma_2 > \frac{\epsilon}{2}\right) \leq \frac{2}{\epsilon}\mathbf{E}(\Sigma_2) \to 0 \text{ as } m \to \infty. \tag{5.2.16}$$

Through the previous analysis of Σ_1 and Σ_2, we have shown that $\mathbf{Pr}(\Sigma_1 > \frac{\epsilon}{2}) \to 0$ as well as $\mathbf{Pr}(\Sigma_2 > \frac{\epsilon}{2}) \to 0$, which leads to

$$\mathbf{Pr}(\mathcal{A}_m(\epsilon)) \leq \mathbf{Pr}\left(\Sigma_1 > \frac{\epsilon}{2}\right) + \mathbf{Pr}\left(\Sigma_2 > \frac{\epsilon}{2}\right) \to 0 \text{ as } m \to \infty. \tag{5.2.17}$$

We have almost finished our proof, however we still need to resolve the Jordan measurability issue. $\mathcal{A}_m(\epsilon)$ itself is not necessarily Jordan measurable, however we will show that it is contained in a slightly larger set $\mathcal{B}_m(\epsilon)$ which is so. Here we need to invoke the assumption that $\sup_{\boldsymbol{x},\boldsymbol{u}} \rho(\boldsymbol{x},\boldsymbol{u}) < \infty$. Let $K = \sup_{\boldsymbol{x},\boldsymbol{u}} \rho(\boldsymbol{x},\boldsymbol{u})$. Recall $\boldsymbol{x}^*$ is a corner, therefore by definition there exists Jordan measurable set $\mathcal{C}_1$ such that $\forall \boldsymbol{u} \in \mathcal{C}_1, \boldsymbol{x} \in \Omega$, $\phi(\boldsymbol{x},\boldsymbol{u}) \in B(\boldsymbol{x}^*, 1)$, where $B(\boldsymbol{x}^*, 1)$ denotes the unit ball centered at $\boldsymbol{x}^*$. As we are assuming the state space to be embeded in a s dimension Euclidean space, for any $m \in \mathbb{N}$, we can find a set of finite points:

$$\Omega_m \stackrel{\Delta}{=} \left\{\boldsymbol{y}^1, \boldsymbol{y}^2, \cdots, \boldsymbol{y}^J\right\} \subseteq B(\boldsymbol{x}^*, 1) \tag{5.2.18}$$

such that

$$B(\boldsymbol{x}^*, 1) \subseteq \bigcup_{j=1}^{J} B(\boldsymbol{y}^j, \frac{\epsilon}{2K^m}). \tag{5.2.19}$$

In fact, we can choose the constant $J = \lceil \frac{2\sqrt{s}K^m}{\epsilon} \rceil^s$. Define $\tau = \inf\{i \geq 1 : \boldsymbol{u}_i \in \mathcal{C}_1\}$, then τ is a hitting time following Geometric distribution.

$$\begin{aligned}\mathcal{A}_m(\epsilon) &= \left\{(\boldsymbol{u}_1, \cdots, \boldsymbol{u}_m) \in [0,1)^{dm} : \sup_{\widetilde{\boldsymbol{x}}_0 \in \Omega} \left|\frac{1}{m}\sum_{i=1}^{m} f(\widetilde{\boldsymbol{x}}_i)\right| > \epsilon\right\} \\ &\subseteq \{\tau > M\} \bigcup \Big(\{\tau = 1\} \bigcap \mathcal{A}_m(\epsilon)\Big) \bigcup \cdots \bigcup \Big(\{\tau = M\} \bigcap \mathcal{A}_m(\epsilon)\Big)\end{aligned}$$

for any $M \in \mathbb{N}$. Fix M, define $\mathcal{A}_m^j(\epsilon) \stackrel{\Delta}{=} \{\tau = j\} \bigcap \mathcal{A}_m(\epsilon)$. Obviously $\{\tau > M\}$ is

Jordan measurable and has volume $(1 - Vol(\mathcal{C}_1))^M$. For any $1 \leq j \leq M$,

$$\begin{aligned}
\mathcal{A}_m^j(\epsilon) &\triangleq \{\tau = j\} \bigcap \mathcal{A}_m(\epsilon) \\
&\subseteq \left\{ (\boldsymbol{u}_1, \cdots, \boldsymbol{u}_m) \in [0,1)^{dm} : \boldsymbol{u}_j \in \mathcal{C}_1, \sup_{\widetilde{\boldsymbol{x}}_0 \in \Omega} \left| \frac{1}{m} \sum_{i=1}^{m} f(\widetilde{\boldsymbol{x}}_i) \right| > \epsilon \right\} \\
&\subseteq \left\{ (\boldsymbol{u}_1, \cdots, \boldsymbol{u}_m) \in [0,1)^{dm} : \widetilde{\boldsymbol{x}}_j \in B(\boldsymbol{x}^*, 1), \sup_{\widetilde{\boldsymbol{x}}_0 \in \Omega} \left| \frac{1}{m} \sum_{i=j+1}^{m} f(\widetilde{\boldsymbol{x}}_i) \right| > \epsilon - \frac{j}{m} \right\} \\
&\subseteq \left\{ (\boldsymbol{u}_1, \cdots, \boldsymbol{u}_m) \in [0,1)^{dm} : \sup_{\widetilde{\boldsymbol{x}}_j \in B(\boldsymbol{x}^*, 1)} \left| \frac{1}{m-j} \sum_{i=j+1}^{m} f(\widetilde{\boldsymbol{x}}_i) \right| > \epsilon - \frac{M}{m} \right\} \\
&= \left\{ (\boldsymbol{u}_1, \cdots, \boldsymbol{u}_j) \in [0,1)^{dj} \right\} \\
&\quad \otimes \left\{ (\boldsymbol{u}_{j+1}, \cdots, \boldsymbol{u}_m) \in [0,1)^{d(m-j)} : \sup_{\widetilde{\boldsymbol{x}}_j \in B(\boldsymbol{x}^*, 1)} \left| \frac{1}{m-j} \sum_{i=j+1}^{m} f(\widetilde{\boldsymbol{x}}_i) \right| > \epsilon - \frac{M}{m} \right\}
\end{aligned} \tag{5.2.20}$$

where $\otimes$ denotes the Cartesian product. The $\sup_{\widetilde{\boldsymbol{x}}_j \in B(\boldsymbol{x}^*, 1)}$ term causes the Jordan non-measurability, since it implies taking the union of infinite sets. We would like to replace $\sup_{\widetilde{\boldsymbol{x}}_j \in B(\boldsymbol{x}^*, 1)}$ with $\sup_{\widetilde{\boldsymbol{x}}_j \in \Omega_m}$, where Ω_m is defined in (5.2.18). We are able to do so because $\forall \widetilde{\boldsymbol{x}}_j \in B(\boldsymbol{x}^*, 1)$, we can find some $\boldsymbol{y} \in \Omega_m$ such that $\|\boldsymbol{y} - \widetilde{\boldsymbol{x}}_j\| < \frac{\epsilon}{2K^m}$. By assuming $\rho(\boldsymbol{x}, \boldsymbol{u}) \leq K$, it is easy to see $\forall i \in [j+1, m]$:

$$\|\phi_{i-j}(\widetilde{\boldsymbol{x}}_j; \boldsymbol{u}_i, \boldsymbol{u}_{i-1}, \cdots, \boldsymbol{u}_{j+1}) - \phi_{i-j}(\boldsymbol{y}^k; \boldsymbol{u}_i, \boldsymbol{u}_{i-1}, \cdots, \boldsymbol{u}_{j+1})\| \leq \frac{\epsilon}{2K^m} \times K^{i-j} \leq \frac{\epsilon}{2}$$

which leads to the following inequality by further assuming f is Lipshitz continuous with Lipshitz constant 1:

$$\sup_{\widetilde{\boldsymbol{x}}_j \in B(\boldsymbol{x}^*, 1)} \left| \frac{1}{m-j} \sum_{i=j+1}^{m} f(\widetilde{\boldsymbol{x}}_i) \right| \leq \sup_{\widetilde{\boldsymbol{x}}_j \in \Omega_m} \left| \frac{1}{m-j} \sum_{i=j+1}^{m} f(\widetilde{\boldsymbol{x}}_i) \right| + \frac{\epsilon}{2}. \tag{5.2.21}$$

Therefore we are able to replace $\sup_{\widetilde{\boldsymbol{x}}_j \in B(\boldsymbol{x}^*,1)}$ with $\sup_{\widetilde{\boldsymbol{x}}_j \in \Omega_m}$ as follows:

$$\left\{ (\boldsymbol{u}_{j+1}, \cdots, \boldsymbol{u}_m) \in [0,1)^{d(m-j)} : \sup_{\widetilde{\boldsymbol{x}}_j \in B(\boldsymbol{x}^*,1)} \left| \frac{1}{m-j} \sum_{i=j+1}^{m} f(\widetilde{\boldsymbol{x}}_i) \right| > \epsilon - \frac{M}{m} \right\} \quad (5.2.22)$$

$$\subseteq \left\{ (\boldsymbol{u}_{j+1}, \cdots, \boldsymbol{u}_m) \in [0,1)^{d(m-j)} : \sup_{\widetilde{\boldsymbol{x}}_j \in \Omega_m} \left| \frac{1}{m-j} \sum_{i=j+1}^{m} f(\widetilde{\boldsymbol{x}}_i) \right| > \frac{\epsilon}{2} - \frac{M}{m} \right\} \quad (5.2.23)$$

$$\triangleq \mathcal{B}_m^j(\epsilon). \quad (5.2.24)$$

Each of the $\mathcal{B}_m^j(\epsilon)$ is Jordan measurable for all ϵ except for a Null set, which we simply ignore here. Therefore combining (5.2.24) and (5.2.20):

$$\mathcal{A}_m(\epsilon) \subseteq \{\tau > M\} \bigcup_{j=1}^{M} \mathcal{A}_m^j(\epsilon) \subseteq \{\tau > M\} \bigcup_{j=1}^{M} [0,1)^{dj} \otimes \mathcal{B}_m^j(\epsilon) \triangleq \mathcal{B}_m^M(\epsilon) \quad (5.2.25)$$

and $\mathcal{B}_m^M(\epsilon)$ is Jordan measurable. For each $1 \leq j \leq M$, substitute m by $m-j$ in (5.2.17),

$$\begin{aligned} \limsup_{m\to\infty} Vol\left(\mathcal{B}_m^j(\epsilon)\right) &\leq \lim_{m\to\infty} \mathbf{Pr}\left(\sup_{\widetilde{\boldsymbol{x}}_j \in \Omega} \left| \frac{1}{m-j} \sum_{i=j+1}^{m} f(\widetilde{\boldsymbol{x}}_i) \right| > \frac{\epsilon}{2} - \frac{M}{m} \right) \\ &= \mathbf{Pr}\left(\mathcal{A}_{m-j}\left(\frac{\epsilon}{2} - \frac{M}{m} \right) \right) \to 0 \quad \text{as } m \to \infty \end{aligned}$$

Hence, for fixed $M > 0$,

$$\limsup_{m\to\infty} Vol(\mathcal{B}_m^M(\epsilon)) = \mathbf{Pr}(\tau > M) = (1 - Vol(\mathcal{C}_1))^M$$

which can be arbitrarily small as $M \to \infty$, therefore we get our desired result.

□

Remark (1)**.** Contracting on average is a relaxation of the global contracting property defined in Chapter 4. Under the same assumptions as in Theorem 5.2.4, global

contracting property can be restated as

$$\mathbf{E}_{\mathcal{U}[0,1)^d} \sup_{\boldsymbol{x}} \log(\rho(\boldsymbol{x}, \boldsymbol{u})) < 0$$

while the contracting on average only requires

$$\mathbf{E}_{\mathcal{U}[0,1)^d} \mathbf{E}_{\boldsymbol{x}\sim\pi} \log(\rho(\boldsymbol{x}, \boldsymbol{u})) < 0$$

which is clearly a weaker assumption. However, some additional assumptions are needed for the consistency to hold under contracting on average situation, among which the existence of a corner is the most crucial. The corner $\boldsymbol{x}^* \in \Omega$ guarantees us that any two chains will be very close at some point, and contracting on average property makes sure that they keep being close afterwards.

Remark (2)**.** The convexity of state space Ω is important in the proof, as it ensures that we can do the Taylor expansion between any two points $\boldsymbol{x}$ and $\boldsymbol{y}$. However it doesn't seem to be as critical as the existence of corner.

Remark (3)**.** In our definition of $\boldsymbol{x}^*$ being a corner, we require $\mathcal{C}_\delta \subseteq [0,1)^d$ be Jordan measurable, which is necessary in our proof to make sure $\{\tau > M\}$ is Jordan measurable. However, if we assume that Ω is bounded, we no longer need the Jordan measurability of $\mathcal{C}_\delta$, since in this situation we can w.o.l.g assume $\Omega \subseteq B(\boldsymbol{x}^*, 1)$ and then $\{\tau > M\} = \emptyset$.

Chapter 6

Convergence Rate of MCQMC

Quasi-Monte Carlo can attain a convergence rate of $O\left(\frac{1}{n^{1-\delta}}\right)$ for any $\delta > 0$ when we know the transformation to convert finite $\mathcal{U}[0,1)$ numbers to generate a certain distribution, assuming the function f we are evaluating composes the transformation function ψ has finite variation under the sense of Hardy-Krause. A key condition for this convergence rate to hold is that the sampling problem is in nature finite, i.e. it only requires finitely many random numbers $(v_1, v_2, \cdots, v_d)$ to generate $\pi \sim \Omega$. When we apply QMC to Markov Chain Monte Carlo, such condition is not satisfied. MCMC can be viewed as a sampling scheme using infinitely many random numbers $(v_1, v_2, \cdots)$. Therefore the convergence rate results for finite case could not be applied directly to MCQMC. However, in a lot of experiments we do observe a better convergence rate by using better balanced points, although the dimension is in principle infinite. Our intuition is, for a Markov Chain updating ϕ, $\boldsymbol{x}_n$ depends on $(\boldsymbol{u}_n, \boldsymbol{u}_{n-1}, \cdots, \boldsymbol{u}_1, \boldsymbol{x}_0)$, but usually it depends the greatest on the most recent vector $\boldsymbol{u}_n$, less on $\boldsymbol{u}_{n-1}$, etc, and least on $\boldsymbol{x}_0$. This is true if ϕ has the Strong Contracting Mapping property which will be defined later. Hence the "effective dimension" of the simulation problem is finite, instead of being infinite, which allows us to expect a higher convergence rate by using QMC.

Throughout this chapter, we assume the state space $\Omega \subseteq \mathbb{R}^s$ is a nice bounded region equiped with Euclidean norm $\|\cdot\|$. As before, we seek to evaluate $\mathbf{E}_\pi f$ for a certain

test function f and π is the stationary distribution. We would like to analyze how fast the estimation error

$$\left|\frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{x}_i) - \mathbf{E}_\pi f(\boldsymbol{x})\right|$$

converges to 0 as the sample size $n \to \infty$. $\boldsymbol{x}_i = \phi(\boldsymbol{x}_{i-1}, \boldsymbol{u}_i)$ is a sample path of the Markov chain. $\boldsymbol{u}_i$ can be IID, as in usual MCMC, or be constructed from a deterministic CUD sequence as in MCQMC. The baseline is the IID case, and under certain mixing or ergodicity conditions we have the Central Limit Theorem for it:(cf. Jones [16]):

$$\left|\frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{x}_i) - \mathbf{E}_\pi f(\boldsymbol{x})\right| = O_p\left(\frac{1}{\sqrt{n}}\right) \tag{6.0.1}$$

We would like to see how much we can improve by using MCQMC, i.e, using a more carefully designed more balanced sequence to replace the IID sequence.

6.1 Reducing the Dimension

As before, we assume that we have a Markov Chain starting from $\boldsymbol{x}_0 \in \Omega$ with an update function $\phi(\boldsymbol{x}, \boldsymbol{u})$: $\boldsymbol{x}_i = \phi(\boldsymbol{x}_{i-1}, \boldsymbol{u}_i)$. Since we are discussing MCQMC, we assume $(\boldsymbol{u}_i)_{i\geq 1}$ to be a deterministic sequence. Thus, we can write $\boldsymbol{x}_i$ as:

$$\begin{aligned}\boldsymbol{x}_i &= \phi_i(\boldsymbol{x}_0; \boldsymbol{u}_i, \boldsymbol{u}_{i-1}, \cdots, \boldsymbol{u}_{i-k+1}, \boldsymbol{u}_{i-k}, \boldsymbol{u}_{i-k-1}, \cdots, \boldsymbol{u}_1)\\ &= \phi_k(\boldsymbol{x}_{i-k}; \boldsymbol{w}_i^{(k)})\end{aligned}$$

where $\boldsymbol{w}_i^{(k)} = (\boldsymbol{u}_i, \boldsymbol{u}_{i-1}, \cdots, \boldsymbol{u}_{i-k+1}) \in [0,1)^{kd}$. Here we order the $\boldsymbol{u}$'s in descending index. It demonstrates that $\boldsymbol{x}_i$ only depends on $\boldsymbol{x}_{i-k}$ and the subsequent driving numbers. Our goal is to give an error bound on the estimator:

$$\mathbf{err}_n \triangleq \left|\frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{x}_i) - \mathbf{E}_\pi f(\boldsymbol{x})\right|$$

for a bounded smooth function $f : \Omega \mapsto \mathbb{R}$.

Fix $k \in \mathbb{N}$, define:

$$\bar{f}_k(\boldsymbol{w}^{(k)}) = \int_\Omega f\left(\phi_k(\boldsymbol{x}; \boldsymbol{w}^{(k)})\right) \pi(d\boldsymbol{x}), \text{ for } \boldsymbol{w}^{(k)} \in [0,1)^{dk} \tag{6.1.1}$$

i.e, $\bar{f}_k(\boldsymbol{w}_i^{(k)})$ is the conditional expectation of $f(\boldsymbol{x}_i)$ when assuming $\boldsymbol{x}_{i-k} \sim \pi$ and fixing $\boldsymbol{w}_i^{(k)} = (\boldsymbol{u}_i, \boldsymbol{u}_{i-1}, \cdots, \boldsymbol{u}_{i-k+1})$. It is integrating out the dependence of $f(\boldsymbol{x}_i)$ on $\boldsymbol{x}_{i-k}$.

Lemma 6.1.1.

$$\int_{[0,1)^{kd}} \bar{f}_k(\boldsymbol{w}^{(k)})\, \mathrm{d}\boldsymbol{w}^{(k)} = \mathbf{E}_\pi f(\boldsymbol{x}) \tag{6.1.2}$$

Proof. Plugging in the definition of $\bar{f}_k$ and using the Fubini's Theorem, we have:

$$\begin{aligned}\int_{[0,1)^{kd}} \bar{f}_k(\boldsymbol{w}^{(k)})\, \mathrm{d}\boldsymbol{w}^{(k)} &= \int_{[0,1)^{kd}} \int_\Omega f\left(\phi_k(\boldsymbol{x}; \boldsymbol{w}^{(k)})\right) \pi(d\boldsymbol{x})\, \mathrm{d}\boldsymbol{w}^{(k)} \\ &= \int_{[0,1)^{kd}\times\Omega} f\left(\phi_k(\boldsymbol{x}; \boldsymbol{w}^{(k)})\right) \pi(d\boldsymbol{x})\, \mathrm{d}\boldsymbol{w}^{(k)}\end{aligned}$$

where $\times$ denotes the Cartesian product. Notice that π is the stationary distribution of the Markov Chain, and $\boldsymbol{w}^{(k)} \sim \mathcal{U}[0,1)^{kd}$, therefore $\phi_k(\boldsymbol{x}; \boldsymbol{w}^{(k)}) \sim \pi$, which gives us the desired result. □

Definition 6.1.2. *A Markov Chain update function $\phi(\boldsymbol{x}, \boldsymbol{u})$ is called Strong Contracting Mapping, if for any $\boldsymbol{u}$, we have*

$$\|\phi(\boldsymbol{x}, \boldsymbol{u}) - \phi(\boldsymbol{x}', \boldsymbol{u})\| \leq \alpha \|\boldsymbol{x} - \boldsymbol{x}'\| \tag{6.1.3}$$

for some $0 \leq \alpha < 1$. Here $\|\cdot\|$ denotes the Euclidean norm.

The strong contracting mapping condition is much stronger than the global contracting property defined in (4.3.6), since it requires ϕ to contract by a factor of $\alpha \in (0,1)$ at each step with no exception, while global contracting mapping only requires ϕ to contract under an average sense. It is clear that strong contracting implies global contracting. Next we have a simple but useful Lemma showing that under strong contracting mapping condition, the Markov Chain forgets about the past at an exponential speed.

Lemma 6.1.3. *Assume the update function $\phi(\boldsymbol{x}, \boldsymbol{u})$ is strong contracting, and $\boldsymbol{w}^{(k)} = (\boldsymbol{u}_k, \boldsymbol{u}_{k-1}, \cdots, \boldsymbol{u}_1)$, then we have the following inequality:*

$$\|\phi_k(\boldsymbol{x}; \boldsymbol{w}^{(k)}) - \phi_k(\boldsymbol{x}'; \boldsymbol{w}^{(k)})\| \leq \alpha^k \|\boldsymbol{x} - \boldsymbol{x}'\|.$$

Proof. Let $\boldsymbol{w}_k^{(k)} = \boldsymbol{w}^{(k)}$ and $\boldsymbol{w}_j^{(k)} = (\boldsymbol{u}_j, \boldsymbol{u}_{j-1}, \cdots, \boldsymbol{u}_1)$ for $0 \leq j \leq k$. Then

$$\begin{aligned}
\|\phi_k(\boldsymbol{x}; \boldsymbol{w}^{(k)}) - \phi_k(\boldsymbol{x}'; \boldsymbol{w}^{(k)})\| &= \|\phi\left(\phi_{k-1}(\boldsymbol{x}; \boldsymbol{w}_{k-1}^{(k)}), \boldsymbol{u}_k\right) - \phi\left(\phi_{k-1}(\boldsymbol{x}'; \boldsymbol{w}_{k-1}^{(k)}), \boldsymbol{u}_k\right)\| \\
&\leq \alpha \|\phi_{k-1}(\boldsymbol{x}; \boldsymbol{w}_{k-1}^{(k)}) - \phi_{k-1}(\boldsymbol{x}'; \boldsymbol{w}_{k-1}^{(k)})\| \\
&\leq \cdots \\
&\leq \alpha^k \|\boldsymbol{x} - \boldsymbol{x}'\|.
\end{aligned}$$

□

Now we turn back to the orignal MCQMC sampling problem and see how the strong contracting mapping property can turn it into a more QMC-like finite dimension integration problem. Recall that the MCQMC algorithm starts at $\boldsymbol{x}_0$ and $\boldsymbol{x}_n$ is the nth iteration. First we show that $f(\boldsymbol{x}_i)$ is not too different from $\bar{f}_k(\boldsymbol{w}_i^{(k)})$:

Lemma 6.1.4. *Let the update function $\phi(\boldsymbol{x}, \boldsymbol{u})$ be jointly measurable with respect to $\boldsymbol{x}$ and $\boldsymbol{u}$, and strong contracting mapping with parameter $\alpha \in (0, 1)$. Further assume the sample space Ω is bounded with diameter d_Ω and f is uniformly Lipshitz continuous with Lipshitz constant C. I.e., $|f(\boldsymbol{x}) - f(\boldsymbol{y})| \leq C\|\boldsymbol{x} - \boldsymbol{y}\|$. Then for any $i \geq k$, we have*

$$\left|\bar{f}_k(\boldsymbol{w}_i^{(k)}) - f(\boldsymbol{x}_i)\right| \leq C\alpha^k d_\Omega.$$

Proof.

$$\begin{aligned}
\left|\bar{f}_k(\boldsymbol{w}_i^{(k)}) - f(\boldsymbol{x}_i)\right| &= \left|\int_\Omega f\left(\phi_k(\boldsymbol{x}; \boldsymbol{w}_i^{(k)})\right)\pi(d\boldsymbol{x}) - f\left(\phi_k(\boldsymbol{x}_{i-k}; \boldsymbol{w}_i^{(k)})\right)\right| \\
&\leq \int_\Omega \left|f\left(\phi_k(\boldsymbol{x}; \boldsymbol{w}_i^{(k)})\right) - f\left(\phi_k(\boldsymbol{x}_{i-k}; \boldsymbol{w}_i^{(k)})\right)\right| \pi(d\boldsymbol{x}).
\end{aligned}$$

For any $\boldsymbol{x} \in \Omega$:

$$\begin{aligned}\left|f\left(\phi_k(\boldsymbol{x}; \boldsymbol{w}_i^{(k)})\right) - f\left(\phi_k(\boldsymbol{x}_{i-k}; \boldsymbol{w}_i^{(k)})\right)\right| &\leq C\|\phi_k(\boldsymbol{x}; \boldsymbol{w}_i^{(k)}) - \phi_k(\boldsymbol{x}_{i-k}; \boldsymbol{w}_i^{(k)})\| \\ &\leq C\alpha^k\|\boldsymbol{x} - \boldsymbol{x}_{i-k}\|\end{aligned}$$

by Lemma 6.1.3. Hence we get:

$$\left|\bar{f}_k(\boldsymbol{w}_i^{(k)}) - f(\boldsymbol{x}_i)\right| \leq \int_\Omega C\alpha^k\|\boldsymbol{x} - \boldsymbol{x}_{i-k}\|\pi(d\boldsymbol{x}) \leq C\alpha^k d_\Omega$$

which completes our proof. □

The previous lemmas tell us, under strong contracting condition, $\bar{f}_k(\boldsymbol{w}_i^{(k)})$ is very close to $f(\boldsymbol{x}_i)$ up to an error bounded by $O(\alpha^k)$, which is of order $O(\frac{1}{n})$ if we choose $k = k_n^* = \lceil \frac{\log(n)}{-\log(\alpha)} \rceil$. Replacing $f(\boldsymbol{x}_i)$ with $\bar{f}_k(\boldsymbol{w}_i^{(k)})$, the MCQMC sample average $\frac{1}{n}\sum_{i=1}^n f(\boldsymbol{x}_i)$ now turns into a k dimensional numerical integration estimation, as described in the following:

Corollary 6.1.5. *Assume the update function $\phi(\boldsymbol{x}, \boldsymbol{u})$ is strong contracting with parameter α, Ω is bounded with diameter d_Ω and f is bounded by $|f|_\infty$. Further assume f is uniformly Lipshitz continuous with Lipshitz constant C and $k_n^* = \lceil \frac{\log(n)}{-\log(\alpha)} \rceil$. Then,*

$$\mathbf{err}_n \leq \left|\frac{1}{n - k_n^* + 1}\sum_{i=k_n^*}^n \bar{f}_{k_n^*}(\boldsymbol{w}_i^{(k_n^*)}) - \int_{[0,1)^{k_n^* d}} \bar{f}_{k_n^*}(\boldsymbol{w}^{(k_n^*)})d\boldsymbol{w}^{(k_n^*)}\right| + O\left(\frac{\log(n)}{n}\right). \tag{6.1.4}$$

Proof.

$$\begin{aligned}\mathbf{err}_n = \left|\frac{1}{n}\sum_{i=1}^n f(\boldsymbol{x}_i) - \mathbf{E}_\pi f(\boldsymbol{x})\right| &\leq \left|\frac{1}{n}\sum_{i=k_n^*}^n f(\boldsymbol{x}_i) - \mathbf{E}_\pi f(\boldsymbol{x})\right| + \frac{k_n^* - 1}{n}|f|_\infty \\ &\leq \left|\frac{1}{n}\sum_{i=k_n^*}^n \bar{f}_{k_n^*}(\boldsymbol{w}_i^{(k_n^*)}) - \mathbf{E}_\pi f(\boldsymbol{x})\right| + \frac{k_n^* - 1}{n}|f|_\infty + \left|\frac{1}{n}\sum_{i=k_n^*}^n \left(\bar{f}_{k_n^*}(\boldsymbol{w}_i^{(k_n^*)}) - f(\boldsymbol{x}_i)\right)\right|\end{aligned}$$

$$\leq \left| \frac{1}{n} \sum_{i=k_n^*}^{n} \bar{f}_{k_n^*}(\boldsymbol{w}_i^{(k_n^*)}) - \mathbf{E}_\pi f(\boldsymbol{x}) \right| + \frac{\log(n)}{-\log(\alpha) n} |f|_\infty + C \frac{1}{n} d_\Omega \sup_\Omega \tag{6.1.5}$$

$$\leq \left| \frac{1}{n - k_n^* + 1} \sum_{i=k_n^*}^{n} \bar{f}_{k_n^*}(\boldsymbol{w}_i^{(k_n^*)}) - \int_{[0,1)^{k_n^* d}} \bar{f}_{k_n^*}(\boldsymbol{w}^{(k_n^*)}) \, \mathrm{d}\boldsymbol{w}^{(k_n^*)} \right| + O\left(\frac{\log(n)}{n} \right) \tag{6.1.6}$$

where O depends on $(\alpha, d_\Omega, |f|_\infty, C)$. (6.1.6) comes from Lemma 6.1.1 and (6.1.5) comes from Lemma 6.1.4. □

Define

$$\widetilde{\mathbf{err}}_n \triangleq \left| \frac{1}{n - k_n^* + 1} \sum_{i=k_n^*}^{n} \bar{f}_{k_n^*}(\boldsymbol{w}_i^{(k_n^*)}) - \int_{[0,1)^{k_n^* d}} \bar{f}_{k_n^*}(\boldsymbol{w}^{(k_n^*)}) \, \mathrm{d}\boldsymbol{w}^{(k_n^*)} \right| \tag{6.1.7}$$

the previous Lemma tells us $\mathbf{err}_n \leq \widetilde{\mathbf{err}}_n + O\left(\frac{\log(n)}{n}\right)$. An important feature of $\widetilde{\mathbf{err}}_n$ is, it is in a form of integration estimation error of the function $\bar{f}_{k_n^*} : [0,1)^{k_n^* d} \mapsto \mathbb{R}$ using $n - k_n^* + 1$ points $(\boldsymbol{w}_{k_n^*}^{(k_n^*)}, \cdots, \boldsymbol{w}_n^{(k_n^*)})$. Thus, if $(\boldsymbol{w}_{k_n^*}^{(k_n^*)}, \cdots, \boldsymbol{w}_n^{(k_n^*)})$ are more balanced than IID sampling, we may hope that MCQMC can give us a better convergence rate. The remaining job of this chapter is to give an upper bound on $\widetilde{\mathbf{err}}_n$ under certain conditions.

6.2 Weighted Sobolev Space

Throughout this section we assume the state space is a nice bounded region $\Omega \subseteq \mathbb{R}^s$, and $\phi(\boldsymbol{x}, \boldsymbol{u})$ is infinitely differentiable. As we have discussed in the previous section, under strong contracting mapping condition, the estimation error $\mathbf{err}_n$ is the same as an numerical integration error $\widetilde{\mathbf{err}}_n$, up to a difference of $O\left(\frac{\log(n)}{n}\right)$. Therefore in order to bound $\mathbf{err}_n$, we just need to bound $\widetilde{\mathbf{err}}_n$, on which much research has been done.

Recall

$$\bar{f}_{k_n^*}(\boldsymbol{w}^{(k_n^*)}) = \int_\Omega f\left(\phi_{k_n^*}(\boldsymbol{x};\boldsymbol{w}^{(k_n^*)})\right)\pi(d\boldsymbol{x}),\ \text{for } \boldsymbol{w}^{(k_n^*)} \in [0,1)^{dk_n^*},\ \text{and} \tag{6.2.1}$$

$$\widetilde{\mathbf{err}}_n \triangleq \left| \frac{1}{n-k_n^*+1}\sum_{i=k_n^*}^{n} \bar{f}_{k_n^*}(\boldsymbol{w}_i^{(k_n^*)}) - \int_{[0,1)^{k_n^* d}} \bar{f}_{k_n^*}(\boldsymbol{w}^{(k_n^*)})\,\mathrm{d}\boldsymbol{w}^{(k_n^*)} \right|. \tag{6.2.2}$$

$\widetilde{\mathbf{err}}_n$ is the estimation error of the integral of a $k_n^* d$ dimention function $\bar{f}_{k_n^*}$ using $n-k_n^*+1$ points. Noticing $k_n^* = \lceil \frac{\log(n)}{-\log(\alpha)} \rceil$, therefore the dimension of the integration goes to infinity as the sample size $n \to \infty$. Hence the usual Koksma - Hlawka inequality cannot be used to bound $\widetilde{\mathbf{err}}_n$. In fact, if we use blindly the Koksma-Hlawka inequality, we will get

$$\widetilde{\mathbf{err}}_n \le O\left(\frac{\log(n)^{k_n^*}}{n}\right) = O\left(\frac{n^{\log\log(n)/-\log(\alpha)}}{n}\right)$$

which is not very useful. However, $\bar{f}_{k_n^*}(\boldsymbol{w}_i^{(k_n^*)}) = \bar{f}_{k_n^*}(\boldsymbol{u}_i, \boldsymbol{u}_{i-1}, \cdots, \boldsymbol{u}_{i-k_n^*+1})$ depends greatest on $\boldsymbol{u}_i$, less on $\boldsymbol{u}_{i-1}$, and least on $\boldsymbol{u}_{i-k_n^*+1}$, by the strong contracting mapping assumption. (Here we can see why we order the $\boldsymbol{u}$'s in descending index). A lot of research has been done on the behavior of quasi-Monte Carlo methods when the integrand has such property, which partially explains why Quasi Monte Carlo still works very well when the dimension is very high. Many people have been working on this problem in different directions, including Weighted Sobolev Space (cf. Sloan, Kuo and Joe [43]) and Effective Dimension (cf. Owen [33]). In this paper we are going to adopt the Weighted Sobolev Space approach.

Definition 6.2.1. *f is a function $[0,1)^k \mapsto \mathbb{R}$, $f = f(x_1, x_2, \cdots, x_k)$. Let $\gamma_1 \ge \gamma_2 \ge \gamma_3 \ge \cdots \ge \gamma_j \ge \cdots > 0$. Define*

$$\|f\|_{k,\gamma}^2 = \sum_{A\subseteq\{1,\cdots,k\}} \frac{1}{\gamma_A}\int_{[0,1)^A}\left|\int_{[0,1)^{-A}} \frac{\partial f^{|A|}}{\partial x_A}(x_A, x_{-A})\,\mathrm{d}x_{-A}\right|^2 \mathrm{d}x_A \tag{6.2.3}$$

where $\gamma_\emptyset = 1$, $\gamma_A = \prod_{j\in A}\gamma_j$. For each subset $A \subseteq \{1,\cdots,k\}$, $x_A = (x_{j_1}, x_{j_2}, \cdots, x_{j_{|A|}})$: $j_i \in A$. This is a Reproducing Kernel Hilbert Space and is denoted by $\mathcal{H}_{k,\gamma}$.

It is easy to see that functions lying in the Sobolev space have vanishing higher order mixed derivatives, which implies that the functions are close to low dimension. In order to estimate the integral of a function $f \in \mathcal{H}_{k,\gamma}$, the difficulty will not grow as the dimension $k \to \infty$, as more precisely described below:

Theorem 6.2.2 (X. Wang 2002)**.** *If $\gamma = (\gamma_1, \gamma_2, \cdots)$ satisfies:*

$$\sum_{j=1}^{\infty} \sqrt{\gamma_j} j \log j \log\log j < \infty, \tag{6.2.4}$$

then the first n points of the k dimension Sobol Sequence $\boldsymbol{w}_1, \boldsymbol{w}_2, ...\boldsymbol{w}_n$ satisfy:

$$\forall \|f\|_{k,\gamma} < \infty : \quad \left| \frac{1}{n} \sum_{i=1}^{n} f(\boldsymbol{w}_i) - \int_{[0,1)^k} f(\boldsymbol{w}) \, \mathrm{d}\boldsymbol{w} \right| \le \frac{C_{\gamma,\delta} \|f\|_{k,\gamma}}{n^{1-\delta}}, \forall \delta > 0 \tag{6.2.5}$$

where $C_{\gamma,\delta}$ does not depend on dimension k.

Proof. see [46]. □

A similar result holds for Niederreiter sequences without the $\log\log j$ term(see [46]). We would like to point out that the conditions on the weights $\gamma = (\gamma_1, \cdots)$ are not best possible. If we randomize the Quasi Monte Carlo points and consider the Mean Squared Error of the estimation, in some cases we are able to relax the condition on γ to $\sum_{j=1}^{\infty} \sqrt{\gamma_j} < \infty$. Please refer to Kuo([19]) for more details.

Theorem 6.2.2 inspires us that, if we can show $\bar{f}_{k_n^*}(\boldsymbol{w}) \in \mathcal{H}_{k_n^* d,\gamma}$ for some γ, and there exists a sequence $\boldsymbol{u}_1, \cdots, \boldsymbol{u}_n$ such that $\boldsymbol{w}_i^{(k_n^*)} = (\boldsymbol{u}_i, \boldsymbol{u}_{i-1}, \cdots, \boldsymbol{u}_{i-k_n^*+1})$: $k_n^* \le i \le n$ satisfy (6.2.5), then by replacing IID sequence with this "Optimal" sequence $(\boldsymbol{u}_i)_{1 \le i \le n}$, we could achieve a convergence rate of $O\left(\frac{1}{n^{1-\delta}}\right)$ for any $\delta > 0$ when estimating $\mathbf{E}_\pi f$. Here is the main theorem:

Theorem 6.2.3 (Optimistic Convergence Rate)**.** *Assume the state space is $(\Omega, \mathscr{B})$ where $\Omega \subseteq \mathbb{R}^s$ is a nice bounded region with diameter d_Ω. Let $(v_i)_{i \ge 1}$ be the driving sequence and $\boldsymbol{u}_i = (v_{di-d+1}, \cdots, v_{di})$. The MCQMC algorithm starts at $\boldsymbol{x}_0 \in \Omega$ and $\boldsymbol{x}_i = \phi(\boldsymbol{x}_{i-1}, \boldsymbol{u}_i)$ for $i \ge 1$. Assume the Markov Chain is irreducible and Harris*

recurrent with stationary distribution π. Further assume ϕ is infinitely differentiable and $\|\frac{\partial \phi}{\partial \boldsymbol{x}}\|_2 \leq \alpha < 1$. f is a smooth function defined on Ω and satisfies $\|\bar{f}_k\|_{kd,\gamma} \leq M < \infty$ for all $k \in \mathbb{N}$, for some $\gamma = (\gamma_1, \gamma_2, \cdots)$. Let $k_n^* = \lceil \frac{\log n}{-\log \alpha} \rceil$, and $\boldsymbol{w}_i^{(k_n^*)} = (\boldsymbol{u}_i, \boldsymbol{u}_{i-1}, \cdots, \boldsymbol{u}_{i-k_n^*+1}), k_n^* \leq i \leq n$. Then if $\boldsymbol{w}_{k_n^*}^{(k_n^*)}, \cdots, \boldsymbol{w}_n^{(k_n^*)}$ satisfies (6.2.5), the sample average $\frac{1}{n}\sum_{i=1}^n f(\boldsymbol{x}_i)$ has estimation error $\mathbf{err}_n = O\left(\frac{1}{n^{1-\delta}}\right)$ for any $\delta > 0$.

Proof. By Corollary 6.1.5, for any $\delta > 0$,

$$\begin{aligned}
\mathbf{err}_n &\leq \widetilde{\mathbf{err}}_n + O\left(\frac{\log n}{n}\right) \\
&\leq \frac{C_{\gamma,\delta}\|\bar{f}_{k_n^*}\|_{k_n^* d,\gamma}}{(n-k_n^*)^{1-\delta}} + O\left(\frac{\log n}{n}\right) \\
&\quad \text{by assuming } \boldsymbol{w}_i^{(k_n^*)} \text{ satisfies (6.2.5)} \\
&= O\left(\frac{1}{n^{1-\delta}}\right).
\end{aligned}$$

□

Remark. As we have seen in the proof, we here are only giving an optimistic convergence rate, since we need to assume $(\boldsymbol{w}_i^{k_n^*})_{i=k_n^*}^n$ satisfies (6.2.5), which we don't know whether is plausible. The difficulty comes from the fact that $(\boldsymbol{w}_i^{k_n^*})_{i=k_n^*}^n$ are overlapping with each other, unless $k_n^* = 1$ which is not interesting. We haven't found in the literature any result about the existence of such overlapping sequence that satisfies (6.2.5).

The following is a constructive result, which only requires us know the strong contracting mapping parameter α. The idea is that we jump k_n^* steps ahead each time and only take the ik_n^*th samples for $i = 1, \cdots, \lfloor \frac{n}{k_n^*} \rfloor$. By doing this, we are losing the efficiency by a factor of $k_n^* = O(\log n)$, which is negligible as shown in the following Theorem:

Theorem 6.2.4 (Informed Convergence Rate)**.** *Assume the state space is $(\Omega, \mathscr{B})$ where $\Omega \subseteq \mathbb{R}^s$ is bounded with diameter d_Ω. Let $(v_i)_{i\geq 1}$ be the driving sequence and $\boldsymbol{u}_i = (v_{di-d+1}, \cdots, v_{di})$. The MCQMC algorithm starts at $\boldsymbol{x}_0 \in \Omega$ and $\boldsymbol{x}_i = \phi(\boldsymbol{x}_{i-1}, \boldsymbol{u}_i)$*

for $i \geq 1$. Assume the Markov Chain is irreducible and Harris recurrent with stationary distribution π. Further assume ϕ to be infinitely differentiable and $\|\frac{\partial \phi}{\partial \boldsymbol{x}}\|_2 \leq \alpha < 1$. f is a smooth function defined on Ω and satisfies $\|\bar{f}_k\|_{kd,\gamma} \leq M < \infty$ for all k, for some $\gamma = (\gamma_1, \gamma_2, \cdots)$. Let $k_n^* = \lceil \frac{\log n}{-\log \alpha} \rceil$, and $\boldsymbol{w}_i^{(k_n^*)} = (\boldsymbol{u}_i, \boldsymbol{u}_{i-1}, \cdots, \boldsymbol{u}_{i-k_n^*+1})$ for $i \geq k_n^*$. If we take $\boldsymbol{w}_{ik_n^*}^{(k_n^*)}$ to be the first ik_n^* point of dk_n^* dimension Sobol's sequence, and $\sum_{j=1}^{\infty} \sqrt{\gamma_j} j \log j \log\log j < \infty$, then we can take the average of the samples $f(\boldsymbol{x}_{ik_n^*}) : 1 \leq i \leq \lfloor \frac{n}{k_n^*} \rfloor$ as an estimate of $\mathbf{E}_\pi f$ which achieves an estimation error of $O\left(\frac{1}{n^{1-\delta}}\right)$ for any $\delta > 0$.

Proof. For simplicity we neglect the rounding issue here to avoid too many $\lfloor \rfloor$ signs.

$$\begin{aligned}
\left| \frac{k_n^*}{n} \sum_{i=1}^{\frac{n}{k_n^*}} f(\boldsymbol{x}_{ik_n^*}) - \mathbf{E}_\pi f \right| &\leq \left| \frac{k_n^*}{n} \sum_{i=1}^{\frac{n}{k_n^*}} \bar{f}_{k_n^*}(\boldsymbol{w}_{ik_n^*}^{(k_n^*)}) - \mathbf{E}_\pi f \right| + O(\alpha^{k_n^*}) \quad \text{by Lemma 6.1.4} \\
&\leq \frac{C_{\gamma,\delta} \|\bar{f}_{k_n^*}\|_{k_n^* d,\gamma}}{(\frac{n}{k_n^*})^{1-\delta}} + O\left(\frac{1}{n}\right) \text{ by Theorem 6.2.2} \qquad (6.2.6) \\
&= O\left(\frac{1}{n^{1-\delta'}}\right) \qquad (6.2.7)
\end{aligned}$$

for arbitrary $\delta' > \delta$. (6.2.6) comes from Theorem 6.2.2 and the assumptions. Since δ is arbitrary positive number, we get the desired result. □

Remark. This is a more realistic convergence rate, however we do need to know the strong contracting mapping parameter α, which could be very hard to compute in real problem. This is why we call it "Informed" convergence rate.

6.3 Example: ARMA Process

The ARMA process is an important model in Times Series Analysis to describe the linear dependence of current value on the historical data. It is being widely used in Econometrics, Social Science, Finance and many other areas. For a monograph of Time Series models, see Brockwell and Davis [2]. Here we focus on how to simulate from an ARMA process's stationary distribution. An ARMA process with general

innovations can be written as follows:

$$X_t = \epsilon_t + \sum_{j=1}^{p} \alpha_j X_{t-j} + \sum_{j=1}^{q} \beta_j \epsilon_{t-j} \tag{6.3.1}$$

$$\epsilon_t = F^{-1}(u_t), u_t \text{ IID} \sim \mathcal{U}[0,1), \quad \mathbf{E}\epsilon_t = 0, \mathbf{E}\epsilon_t^2 < \infty \tag{6.3.2}$$

where F is the CDF of ϵ_t. Given the initial values $X_0, X_{-1}, \cdots, X_{1-p}$ and driving sequence $(u_t)_{t \geq 1-q}$, we can easily simulate the whole time series. Assume f is a function on $\mathbb{R}$ and we seek to evaluate $\mathbf{E}_\pi f(X_t)$ where π is the stationary distribution. For simplicity, we assume the initial condition to be:

$$X_0 = X_{-1} = \cdots = X_{1-p} = 0, \text{ and } \quad \epsilon_0 = \epsilon_{-1} = \cdots = \epsilon_{1-q} = 0. \tag{6.3.3}$$

In order for the time series to be stationary, we need the following assumptions (see, for example, [2]):

Theorem 6.3.1. *An ARMA model* (6.3.1) *is stationary if and only if all the roots of polynomial:*

$$1 - \sum_{j=1}^{p} \alpha_j z^j = 0 \tag{6.3.4}$$

on $\mathbb{C}$ *are outside of the unit circle. In such case, if we further assume initial condition* (6.3.3) *to hold, then* X_t *has a Moving Average representation:*

$$X_t = \sum_{j=0}^{\infty} \varphi_j \epsilon_{t-j}, \quad \epsilon_0 = \epsilon_{-1} = \epsilon_{-2} = \cdots = 0, \tag{6.3.5}$$

where $\sum_{j=0}^{\infty} \varphi_j z^j = \frac{1+\sum_{j=1}^{q} \beta_j z^j}{1-\sum_{j=1}^{p} \alpha_j z^j}$, *for* $z \in \mathbb{C}, |z| < 1$.

In order to determine the strong contracting mapping parameter α, we need to find the decaying speed of $(\varphi_j)_{j \geq 0}$. We have the following Theorem:

Theorem 6.3.2. *Assume the ARMA model* (6.3.1) *is stationary, and* $z_1 \in \mathbb{C}, |z_1| > 1$ *is the smallest root of polynomial* $1 - \sum_{j=1}^{p} \alpha_j z^j = 0$. *Then for any* $|z_1|^{-1} < \rho < 1$, *there exists* $K_\rho > 0$ *such that* $|\varphi_j| \leq K_\rho \rho^j$.

Proof. By Theorem 6.3.1, the weights in the moving-average representation of X_t satisfiy:

$$\sum_{j=0}^{\infty} \varphi_j z^j = \frac{1 + \sum_{j=1}^{q} \beta_j z^j}{1 - \sum_{j=1}^{p} \alpha_j z^j}.$$

It is easy to see that φ_j is a polynomial of $(\alpha_j)_{j=1}^{p}$ and $(\beta_j)_{j=1}^{q}$ with positive coefficients. Let $\beta = \max |\beta_j|$ and define the majorant function:

$$\sum_{j=0}^{\infty} \widetilde{\varphi}_j z^j \triangleq \frac{1 + \sum_{j=1}^{q} |\beta| z^j}{\prod_{j=1}^{p} (1 - \frac{z}{|z_1|})} \tag{6.3.6}$$

we have $|\varphi_j| \leq |\widetilde{\varphi}_j|$. To compute $\widetilde{\varphi}_j$, by generalized binomial Theorem,

$$\prod_{j=1}^{p} \left(1 - \frac{z}{|z_1|}\right)^{-1} = \sum_{j=0}^{\infty} \begin{pmatrix} p+j-1 \\ p-1 \end{pmatrix} \left(\frac{z}{|z_1|}\right)^j \tag{6.3.7}$$

$$\begin{aligned} \Rightarrow \quad \widetilde{\varphi}_j &= \begin{pmatrix} p+j-1 \\ p-1 \end{pmatrix} |z_1|^{-j} + |\beta| \sum_{l=1}^{q \wedge j} \begin{pmatrix} p+j-l-1 \\ p-1 \end{pmatrix} |z_1|^{-j+l} \\ &= O\left(\rho^j\right) \end{aligned}$$

for any $\rho \in (|z_1|^{-1}, 1)$. Hence we completed our proof. □

With the previous Theorems, we are ready to show that by replacing the IID sequence with a suitable Sobol's sequence, MCQMC can achieve $O\left(\frac{1}{n^{1-\delta}}\right)$ convergence rate under mild conditions.

Theorem 6.3.3. *In an ARMA process* (6.3.1) *with initial condition* (6.3.3)*, assume* $\epsilon_t = F^{-1}(u_t)$ *has compact support and continuous density bounded from below:* $F' \geq \epsilon > 0$*, and the* r^{th} *derivative of* f*:*$|f^{(r)}| \leq M^r r!$ *for some constant* $M > 0$*. Let* $z_1 \in \mathbb{C}, |z_1| > 1$ *be the smallest root of polynomial* $1 - \sum_{j=1}^{p} \alpha_j z^j = 0$ *and* $\rho \in (|z_1|^{-1}, 1)$*. Define* $k_n^* = \lceil \frac{\log n}{-\log \rho} \rceil$*. If the driving sequence* $(u_t)_{t \geq 1}$ *is constructed in such a way that* $(u_{ik_n^*}, \cdots, u_{(i-1)k_n^*+1})$ *is the ith point of* k_n^* *dimension Sobol's sequence, then we can take the average of the samples* $f(X_{ik_n^*}) : 1 \leq i \leq \lfloor \frac{n}{k_n^*} \rfloor$ *as an estimate of* $\mathbf{E}_\pi f$ *which*

achieves an estimation error of $O\left(\frac{1}{n^{1-\delta}}\right)$ *for any* $\delta > 0$.

Proof. As shown in Theorem 6.3.1, assuming the initial condition (6.3.3), X_t has the following representation: $X_t = \sum_{j=0}^{t-1} \varphi_j \epsilon_{t-j} = \sum_{j=0}^{t-1} \varphi_j F^{-1}(u_{t-j})$. Therefore letting $t = ik_n^*, 1 \le i \le \lfloor \frac{n}{k_n^*} \rfloor$, we have

$$X_{ik_n^*} = \sum_{j=0}^{ik_n^*-1} \varphi_j F^{-1}(u_{ik_n^*-j}), \quad 1 \le i \le \lfloor \frac{n}{k_n^*} \rfloor.$$

Define the "truncated" $X_{ik_n^*}$ as:

$$\widetilde{X}_{ik_n^*} = \sum_{j=0}^{k_n^*-1} \varphi_j F^{-1}(u_{ik_n^*-j}), \quad 1 \le i \le \lfloor \frac{n}{k_n^*} \rfloor \tag{6.3.8}$$

$$|X_{ik_n^*} - \widetilde{X}_{ik_n^*}| \le \sum_{j=k_n^*}^{\infty} |\varphi_j| \max_u |F^{-1}(u)| = O\left(\rho^{k_n^*}\right) = O\left(\frac{1}{n}\right) \tag{6.3.9}$$

by assuming distribution $\epsilon_t \sim F$ has compact support and Theorem 6.3.2.

Define $\bar{f}_k : [0,1)^k \mapsto \mathbb{R} : \bar{f}_k(x_1, x_2, \cdots, x_k) \triangleq f\left(\sum_{j=1}^k \varphi_{j-1} F^{-1}(x_j)\right)$. Then $f(\widetilde{X}_{ik_n^*}) = \bar{f}_{k_n^*}(u_{ik_n^*}, \cdots, u_{ik_n^*-k_n^*+1})$. Notice this definition is a little bit different from section 1 (6.1.1). First we would like to prove that $\|\bar{f}_k\|_{k,\gamma} \le M_2 < \infty$ for all $k > 0$, for some $\gamma = (\gamma_1, \gamma_2, \cdots)$ satisfies $\sum_{j=1}^{\infty} \sqrt{\gamma_j} j \log j \log\log j < \infty$. To show this, we need to bound the magnitude of $\bar{f}_k$'s higher order mixed derivatives:

Lemma 6.3.4. *There exists* $M_1 > 0$ *such that:*

$$\left| \frac{\partial \bar{f}_k^{|A|}}{\partial x_A} \right| \le M_1^{|A|} |A|! \rho^{\sum_{j \in A} j} \tag{6.3.10}$$

for $A \subseteq \{1, 2, \cdots, k\}$, $A \neq \emptyset$.

Proof. By definition, $\bar{f}_k(x_1, x_2, \cdots, x_k) = f\left(\sum_{j=1}^k \varphi_{j-1} F^{-1}(x_j)\right)$. We prove this result by Method of Majorants, which could be traced back to Cauchy who used it to prove the existence of analytical solution to some certain ODEs or PDEs (cf. Cauchy -

Kovalevskaya Theorem). Using the chain rule of differentiation and Leibniz's product rule, it is easy to see that when $|A| > 0$, $\frac{\partial \bar{f}_k^{|A|}}{\partial x_A}$ is a polynomial of $\varphi_j : 0 \le j \le k-1$, $f^{(r)} : 1 \le r \le |A|$ and $(F^{-1})' = \frac{1}{F'}$ with non-negative coefficients. By Theorem 6.3.2, $|\varphi_j| \le K_\rho \rho^j$, and $|f^{(r)}| \le M^r r!$, $(F^{-1})' = \frac{1}{F'} \le \frac{1}{\epsilon}$ by assumption, thus we have:

$$\left| \frac{\partial \bar{f}_k^{|A|}}{\partial x_A} \right| \le \left. \frac{\partial h^{|A|}}{\partial x_A} \right|_{x_j = 0, 1 \le j \le k} \tag{6.3.11}$$

where h is defined as:

$$h(x_1, x_2, \cdots, x_k) \stackrel{\Delta}{=} 1 + \sum_{r=1}^{\infty} M^r \left(\sum_{j=1}^{k} K_\rho \rho^{j-1} \frac{x_j}{\epsilon} \right)^r \tag{6.3.12}$$

$$= \frac{1}{1 - M \sum_{j=1}^{k} K_\rho \rho^{j-1} \frac{x_j}{\epsilon}} \tag{6.3.13}$$

$$= \frac{1}{1 - \frac{MK_\rho}{\epsilon\rho} \sum_{j=1}^{k} \rho^j x_j}. \tag{6.3.14}$$

Define $M_1 = \frac{MK_\rho}{\epsilon\rho}$. Assume $A = \{j_1, j_2, \cdots, j_{|A|}\}$ and $1 \le j_1 < j_2, \cdots, < j_{|A|} \le k$, we now directly compute $\left. \frac{\partial h^{|A|}}{\partial x_A} \right|_{x_j = 0, 1 \le j \le k}$. By some easy calculation, we get

$$\left. \frac{\partial h^{|A|}}{\partial x_A} \right|_{x_j = 0, 1 \le j \le k} = |A|! \left(\frac{MK_\rho}{\epsilon\rho} \right)^{|A|} \rho^{\sum_{l=1}^{|A|} j_l} = M_1^{|A|} |A|! \rho^{\sum_{j \in A} j} \tag{6.3.15}$$

which proves that $\left| \frac{\partial \bar{f}_k^{|A|}}{\partial x_A} \right| \le M_1^{|A|} |A|! \rho^{\sum_{j \in A} j}$ when $A \neq \emptyset$.

□

Using the previous Lemma we are ready to embed $\bar{f}_k$ into the Weighted Sobolev Space $\mathcal{H}_{k,\gamma}$ for some $\gamma = (\gamma_1, \gamma_2, \cdots)$. We choose $\gamma_j = \rho^j, j = 1, 2, \cdots$, which satisfies the requirements:

$$\gamma_1 \ge \gamma_2 \ge \cdots \ge \gamma_j \ge \cdots > 0$$

$$\sum_{j=1}^{\infty} \sqrt{\gamma_j} j \log j \log\log j < \infty$$

Next we prove for any $k \in \mathbb{N}$, $\|\bar{f}_k\|_{k,\gamma} < M_2 < \infty$ for some constant M_2. By definition,

$$\|\bar{f}_k\|_{k,\gamma}^2 = \sum_{A \subseteq \{1,\cdots,k\}} \frac{1}{\gamma_A} \int_{[0,1)^A} \left| \int_{[0,1)^{-A}} \frac{\partial \bar{f}_k^{|A|}}{\partial x_A}(x_A, x_{-A}) \, \mathrm{d}x_{-A} \right|^2 \mathrm{d}x_A \tag{6.3.16}$$

where $\gamma_A = \prod_{j \in A} \gamma_j$ and $\gamma_\emptyset = 1$. We have already shown in the previous analysis that $\left| \frac{\partial \bar{f}_k^{|A|}}{\partial x_A} \right| \le M_1^{|A|} |A|! \rho^{\sum_{j \in A} j}$ when $A \neq \emptyset$. Hence,

$$\begin{aligned}
\|\bar{f}_k\|_{k,\gamma}^2 &= \sum_{A \subseteq \{1,\cdots,k\}} \frac{1}{\gamma_A} \int_{[0,1)^A} \left| \int_{[0,1)^{-A}} \frac{\partial \bar{f}_k^{|A|}}{\partial x_A}(x_A, x_{-A}) \, \mathrm{d}x_{-A} \right|^2 \mathrm{d}x_A \\
&\le \left(\sup |\bar{f}_k|\right)^2 + \sum_{A \subseteq \{1,\cdots,k\}, A \neq \emptyset} \frac{1}{\gamma_A} \left(M_1^{|A|} |A|! \rho^{\sum_{j \in A} j} \right)^2 \\
&= \left(\sup |\bar{f}_k|\right)^2 + \sum_{A \subseteq \{1,\cdots,k\}, A \neq \emptyset} \frac{1}{\prod_{j \in A} \gamma_j} \left(M_1^{|A|} |A|! \rho^{\sum_{j \in A} j} \right)^2 \\
&\le \left(\sup |\bar{f}_k|\right)^2 + \sum_{A \subseteq \{1,\cdots,k\}, A \neq \emptyset} \frac{1}{\prod_{j \in A} \gamma_j} \left(M_1^{|A|} \left(\prod_{j \in A} j \right) \rho^{\sum_{j \in A} j} \right)^2
\end{aligned}$$

since $|A|! \le \prod_{j \in A} j$. Therefore:

$$\begin{aligned}
\|\bar{f}_k\|_{k,\gamma}^2 &\le \left(\sup |\bar{f}_k|\right)^2 + \prod_{j=1}^{k} \left(1 + \frac{M_1^2 j^2 \rho^{2j}}{\gamma_j} \right) - 1 \\
&\le \left(\sup |\bar{f}_k|\right)^2 + \exp\left(\sum_{j=1}^{k} \frac{M_1^2 j^2 \rho^{2j}}{\gamma_j} \right) - 1 \\
&\le \left(\sup |\bar{f}_k|\right)^2 + \exp\left(\sum_{j=1}^{\infty} M_1^2 j^2 \rho^j \right) \quad \text{by plugging in } \gamma_j = \rho^j \\
&\triangleq M_2^2
\end{aligned}$$

which is a finite number. Therefore as the calculation shows $\|\bar{f}_k\|_{k,\gamma}$ is bounded by M_2 for any k, which indicates that the "effective dimension" of $\bar{f}_k(x_1, x_2, \cdots, x_k)$ is not growing to infinity as $k \to \infty$.

Secondly we want to show that $\left|\int_{[0,1)^k} \bar{f}_k \,\mathrm{d}x_1 \,\mathrm{d}x_2 \cdots \mathrm{d}x_k - \mathbf{E}_\pi f\right| \leq O\left(\rho^k\right)$. Let $\boldsymbol{x} = (x_1, x_2, \cdots)$ and $\boldsymbol{x}_{1:k} = (x_1, \cdots, x_k)$. Notice the distribution of X_∞ (the stationary distribution) has the same law as:

$$X_\infty \stackrel{d}{=} \sum_{j=1}^{\infty} \varphi_{j-1} F^{-1}(x_j): \quad x_j \text{ IID } \sim \mathcal{U}[0,1). \tag{6.3.17}$$

Therefore,

$$\begin{aligned}
&\left|\int_{[0,1)^k} f\left(\sum_{j=1}^{k} \varphi_{j-1} F^{-1}(x_j)\right) \mathrm{d}\boldsymbol{x}_{1:k} - \mathbf{E}_\pi f\right| \\
=&\left|\int_{[0,1)^\infty} f\left(\sum_{j=1}^{k} \varphi_{j-1} F^{-1}(x_j)\right) \mathrm{d}\boldsymbol{x} - \int_{[0,1)^\infty} f\left(\sum_{j=1}^{\infty} \varphi_{j-1} F^{-1}(x_j)\right) \mathrm{d}\boldsymbol{x}\right| \qquad (6.3.18)\\
\leq& \max|f'| \int_{[0,1)^\infty} \left|\sum_{j=k+1}^{\infty} \varphi_j F^{-1}(x_j)\right| \mathrm{d}\boldsymbol{x} \\
\leq& \max|f'| \max_u \left|F^{-1}(u)\right| \sum_{j=k+1}^{\infty} K_\rho \rho^j \qquad (6.3.19)\\
=& O\left(\rho^k\right). \qquad (6.3.20)
\end{aligned}$$

(6.3.18) comes from (6.3.17) and (6.3.19) comes from the assumption that F^{-1} is bounded as well as Theorem 6.3.2.

Lastly, if we take the sample average of $f(X_{ik_n^*}), 1 \leq i \leq \lfloor \frac{n}{k_n^*} \rfloor$ to estimate $\mathbf{E}_\pi f$, we can bound the estimation error as in the proof of Theorem 6.2.4(also ignoring the rounding issue):

$$\left|\frac{k_n^*}{n} \sum_{i=1}^{\frac{n}{k_n^*}} f(X_{ik_n^*}) - \mathbf{E}_\pi f\right| \leq \left|\frac{k_n^*}{n} \sum_{i=1}^{\frac{n}{k_n^*}} f(\widetilde{X}_{ik_n^*}) - \mathbf{E}_\pi f\right| + \sup|f'| \sup_{1 \leq i \leq \frac{n}{k_n^*}} |X_{ik_n^*} - \widetilde{X}_{ik_n^*}|. \tag{6.3.21}$$

$$
\begin{aligned}
&\left| \frac{k_n^*}{n} \sum_{i=1}^{\frac{n}{k_n^*}} f(\widetilde{X}_{ik_n^*}) - \mathbf{E}_\pi f \right| \\
\leq &\left| \frac{k_n^*}{n} \sum_{i=1}^{\frac{n}{k_n^*}} f(\widetilde{X}_{ik_n^*}) - \int_{[0,1)^{k_n^*}} \bar{f}_{k_n^*} \, \mathrm{d}\boldsymbol{x}_{1:k_n^*} \right| + \left| \int_{[0,1)^{k_n^*}} \bar{f}_{k_n^*} \, \mathrm{d}\boldsymbol{x}_{1:k_n^*} - \mathbf{E}_\pi f \right| && (6.3.22) \\
= &\left| \frac{k_n^*}{n} \sum_{i=1}^{\frac{n}{k_n^*}} \bar{f}_{k_n^*}(u_{ik_n^*}, \cdots, u_{ik_n^*-k_n^*+1}) - \int_{[0,1)^{k_n^*}} \bar{f}_{k_n^*} \, \mathrm{d}\boldsymbol{x}_{1:k_n^*} \right| + O\left(\frac{1}{n}\right) && (6.3.23) \\
\leq &\frac{C_{\gamma,\delta} \| \bar{f}_{k_n^*} \|_{k_n^*,\gamma}}{(\frac{n}{k_n^*})^{1-\delta}} + O\left(\frac{1}{n}\right) && (6.3.24) \\
= &O\left(\frac{1}{n^{1-\delta'}}\right) && (6.3.25)
\end{aligned}
$$

for any $\delta' > \delta > 0$. $\gamma = (\rho^j)_{j \geq 1}$ and O depends on $(\gamma, \delta, \delta', M_1, M_2, \rho)$. (6.3.23) comes from (6.3.20). (6.3.24) comes from Theorem 6.2.2 and the result that $\| \bar{f}_{k_n^*} \|_{k_n^*,\gamma} \leq M_2$. On the other hand,

$$
\sup |f'| \sup_{1 \leq i \leq \frac{n}{k_n^*}} |X_{ik_n^*} - \widetilde{X}_{ik_n^*}| = O\left(\frac{1}{n}\right) \tag{6.3.26}
$$

by inequality (6.3.9). Plugging (6.3.25) and (6.3.26) into (6.3.21), we completed our proof.

□

Remark. Compared with Theorem 6.2.3 and Theorem 6.2.4 in the previous section, this result about ARMA process is more "constructive" in the sense that we have full knowledge of how to construct the driving sequence as well as how do the simulation. On the contrary, Theorem 6.2.3 requires a CUD sequence which we don't know whether exists, and Theorem 6.2.4 requires the information about strong contracting parameter α to determine k_n^*. In our ARMA example, k_n^* is determined by the smallest root of polynomial (6.3.4).

Chapter 7

Consistent MCQMC Examples

In this chapter we give several examples which satisfy the conditions imposed in the previous chapters to make MCQMC consistent. Recall we have various types of conditions: coupling region, global contracting, global non-expansive, and contracting on average. For each type of condition we will provide one or more examples. Some of the examples have already appeared in [3].

7.1 Coupling Region

Theorem 4.3.3 used coupling regions. These are somewhat special. But they do exist for some realistic Markov Chain update functions.

Example 7.1.1. Let ϕ be the update for the Metropolized independence sampler on $\Omega \subseteq \mathbb{R}^s$ obtaining the proposal $\boldsymbol{y} = \psi(\boldsymbol{u}_{1:(d-1)})$, where ψ generates samples from the density p, which are accepted when

$$u_d \leq \frac{\pi(\boldsymbol{y})p(\boldsymbol{x})}{\pi(\boldsymbol{x})p(\boldsymbol{y})}.$$

Assume that the importance ratio is bounded above, i.e.,

$$\kappa \equiv \sup_{\boldsymbol{x}\in\Omega} \frac{\pi(\boldsymbol{x})}{p(\boldsymbol{x})} < \infty.$$

Suppose also that there is a rectangle $[\boldsymbol{a}, \boldsymbol{b}] \subset [0,1]^{d-1}$ of positive volume with

$$\eta \equiv \inf_{\boldsymbol{u} \in [\boldsymbol{a}, \boldsymbol{b}]} \frac{\pi(\psi(\boldsymbol{u}))}{p(\psi(\boldsymbol{u}))} > 0.$$

Then $\mathcal{C} = [\boldsymbol{a}, \boldsymbol{b}] \times [0, \eta/\kappa]$ is a coupling region.

Proof. The set $\mathcal{C}$ has positive Jordan measure. Suppose that $\boldsymbol{u} \in \mathcal{C}$. Then

$$\pi(\boldsymbol{y})p(\boldsymbol{x}) \geq \eta p(\boldsymbol{y}) \frac{1}{\kappa} \pi(\boldsymbol{x}) \geq u_d p(\boldsymbol{y}) \pi(\boldsymbol{x}),$$

and so $\phi(\boldsymbol{x}, \boldsymbol{u}) = \boldsymbol{y}$, regardless of $\boldsymbol{x}$. □

Example 7.1.2. Let π be a density on a bounded rectangular region $\Omega = [\boldsymbol{a}, \boldsymbol{b}] \subset \mathbb{R}^s$. Assume that $0 < \eta \leq \pi(\boldsymbol{x}) \leq \kappa < \infty$ holds for all $\boldsymbol{x} \in \Omega$. Let $\Omega' = \{(y, \boldsymbol{x}) \mid 0 \leq y \leq \pi(\boldsymbol{x})\} \subset [\boldsymbol{a}, \boldsymbol{b}] \times [0, \kappa]$ be the domain of the inversive slice sampler. Let $(y_i, \boldsymbol{x}_i) = \phi((y_{i-1}, \boldsymbol{x}_{i-1}), \boldsymbol{u}_i)$ for $\boldsymbol{u}_i \in [0,1]^{s+1}$ be the update for the inversive slice sampler and put $(y_i', \boldsymbol{x}_i') = \phi((y_{i-1}', \boldsymbol{x}_{i-1}'), \boldsymbol{u}_i)$. If $\boldsymbol{u}_i \in \mathcal{C} = [0, \eta/\kappa] \times [0,1]^s$, then $\boldsymbol{x}_i = \boldsymbol{x}_i'$.

Proof. If $u_{i,1} \leq \eta/\kappa$ then $y_i = u_{i1}\pi(\boldsymbol{x}_{i-1})$ and $y_i' = u_{i1}\pi(\boldsymbol{x}_{i-1}')$ are in the set $[0, \eta/\kappa]$. The distribution of $\boldsymbol{x}$ given y for any $y \in [0, \eta/\kappa]$ is $U[\boldsymbol{a}, \boldsymbol{b}]$. Therefore $\boldsymbol{x}_i = \boldsymbol{x}_i' = \boldsymbol{a} + u_{2:(s+1)}(\boldsymbol{b} - \boldsymbol{a})$ (componentwise). □

Remark. Example 7.1.2 does not couple the chains because y_i and y_i' are different in general. But because $\boldsymbol{x}_i = \boldsymbol{x}_i'$, a coupling will happen at the next step, that is $(y_{i+1}, \boldsymbol{x}_{i+1}) = (y_{i+1}', \boldsymbol{x}_{i+1}')$ when $\boldsymbol{u}_i \in [0, \eta/\kappa] \times [0,1]^s$. One could revise Theorem 4.3.3 to include couplings that happen within some number t of steps after $\boldsymbol{u} \in \mathcal{C}$ happens. In this case it is simpler to say that the chain whose update comprises two iterations of the inversive slice sampler satisfies Theorem 4.3.3. For a chain whose update is just one iteration the averages over odd and even numbered iterations both converge properly and so that chain is also consistent. Alternatively, we could modify the space of y values so that all $y \in [0, \eta/\kappa]$ are identified as one point. Then $\mathcal{C}$ is a coupling region.

The result of Lemma 7.1.2 also applies to slice samplers that sample $y \mid \boldsymbol{x}$ and then $\boldsymbol{x} \mid y \sim U\{\boldsymbol{x} \mid \pi(\boldsymbol{x}) \le y\}$ using an s dimensional generator that is not necessarily inversion.

7.2 Gibbs Sampling - Global Contracting Mapping

Here we illustrate how the Gibbs sampler yields a contraction for the probit model. In the Probit model:

$$\begin{aligned} Z_i &= \boldsymbol{x}_i^{\mathsf{T}}\beta + \epsilon_i \quad \text{and} \\ Y_i &= \mathbf{1}_{Z_i>0}, \end{aligned}$$

for $i = 1, \dots, n$ for independent $\epsilon_i \sim \mathcal{N}(0,1)$. The coefficient $\beta \in \mathbb{R}^p$ has a noninformative prior distribution. The predictors are $\boldsymbol{x}_i \in \mathbb{R}^p$. We define the matrix X with ij element x_{ij}. We assume that X has rank p.

The state of the Markov chain is $(\beta, \boldsymbol{Z}) \in \Omega = \mathbb{R}^{p+n}$, where $\boldsymbol{Z} = (Z_1, \dots, Z_n)^{\mathsf{T}}$. Given the observed data $(y_1, \dots, .y_n, \boldsymbol{x}_1, \dots, \boldsymbol{x}_n)$, we can use the Gibbs sampler to simulate the posterior distribution of β and $\boldsymbol{Z} = (Z_1, \dots, Z_n)^{\mathsf{T}}$. A single step of the Gibbs sampler makes the transition

$$\begin{pmatrix} \beta^{(k-1)} \\ \boldsymbol{Z}^{(k-1)} \end{pmatrix} \overset{u_1,\dots,u_n}{\longrightarrow} \begin{pmatrix} \beta^{(k-1)} \\ \boldsymbol{Z}^{(k)} \end{pmatrix} \overset{u_{n+1},\dots,u_{n+p}}{\longrightarrow} \begin{pmatrix} \beta^{(k)} \\ \boldsymbol{Z}^{(k)} \end{pmatrix}$$

for $k \ge 1$ using generators given explicitly below. The values $u_1, \dots, u_{n+p}$ are the components of $\boldsymbol{u}_k \in (0,1)^{n+p}$. We also write the transitions as

$$(\beta, \boldsymbol{Z}) \to \phi((\beta, \boldsymbol{Z}), \boldsymbol{u}) = \big(\phi^{(1)}((\beta, \boldsymbol{Z}), \boldsymbol{u}), \phi^{(2)}((\beta, \boldsymbol{Z}), \boldsymbol{u})\big)$$

where ϕ and its components $\phi^{(1)}$ and $\phi^{(2)}$ (for β and $\boldsymbol{Z}$ respectively) are given explicitly below.

Given β, the components of $\boldsymbol{Z}$ are independent, with

$$Z_i \sim \begin{cases} \mathcal{N}(\boldsymbol{x}_i^\mathsf{T}\beta, 1) | Z_i > 0 & \text{if } Y_i = 1 \\ \mathcal{N}(\boldsymbol{x}_i^\mathsf{T}\beta, 1) | Z_i \le 0 & \text{if } Y_i = 0. \end{cases}$$

We may generate them from $u_1, \dots, u_n \in (0,1)$ by

$$Z_i = \begin{cases} \boldsymbol{x}_i^\mathsf{T}\beta + \Phi^{-1}\big(\Phi(-\boldsymbol{x}_i^\mathsf{T}\beta) + u_i\Phi(\boldsymbol{x}_i^\mathsf{T}\beta)\big) & \text{if } Y_i = 1, \\ \boldsymbol{x}_i^\mathsf{T}\beta + \Phi^{-1}\big(u_i\Phi(-\boldsymbol{x}_i^\mathsf{T}\beta)\big) & \text{if } Y_i = 0. \end{cases} \tag{7.2.1}$$

Given $\boldsymbol{Z}$, the distribution of β is $\beta \sim \mathcal{N}((X^\mathsf{T}X)^{-1}X^\mathsf{T}\boldsymbol{Z}, (X^\mathsf{T}X)^{-1})$. We may generate it using $u_{n+1}, \dots, u_{n+p} \in (0,1)$ via

$$\beta = (X^\mathsf{T}X)^{-1}X^\mathsf{T}\boldsymbol{Z} + (X^\mathsf{T}X)^{-1/2}\begin{pmatrix} \Phi^{-1}(u_{n+1}) \\ \vdots \\ \Phi^{-1}(u_{n+p}) \end{pmatrix}. \tag{7.2.2}$$

Thus equation (7.2.2) defines $\phi^{(1)}$ while equation (7.2.1) defines $\phi^{(2)}$.

Theorem 4.3.7 allows one to pick a metric that conforms to the problem. We use the metric $d\left((\beta, \boldsymbol{Z}), (\beta', \boldsymbol{Z}')\right) = \max\big(d_1(\beta, \beta'), d_2(\boldsymbol{Z}, \boldsymbol{Z}')\big)$, where

$$d_1(\beta, \beta') = d_1(\beta - \beta') = \sqrt{(\beta - \beta')^\mathsf{T}(X^\mathsf{T}X)(\beta - \beta')} \quad \text{and,} \tag{7.2.3}$$

$$d_2(\boldsymbol{Z}, \boldsymbol{Z}') = d_2(\boldsymbol{Z} - \boldsymbol{Z}') = \sqrt{(\boldsymbol{Z} - \boldsymbol{Z}')^\mathsf{T}(\boldsymbol{Z} - \boldsymbol{Z}')}. \tag{7.2.4}$$

We show below that

$$d\big((\beta^{(k)}, \boldsymbol{Z}^{(k)}), (\beta'^{(k)}, \boldsymbol{Z}'^{(k)})\big) < d\big((\beta^{(k-1)}, \boldsymbol{Z}^{(k-1)}), (\beta'^{(k-1)}, \boldsymbol{Z}'^{(k-1)})\big) \tag{7.2.5}$$

for pairs $(\beta^{(k-1)}, \boldsymbol{Z}^{(k-1)}), (\beta'^{(k-1)}, \boldsymbol{Z}'^{(k-1)})$ of distinct points in Ω. Both metrics d_1 and d_2 are also norms, which simplifies our task.

Suppose first that $\beta^{(k-1)} = \beta'^{(k-1)}$. Then it follows easily that $\boldsymbol{Z}^{(k)} = \boldsymbol{Z}'^{(k)}$ and $\beta^{(k)} = \beta'^{(k)}$, so then the left side of (7.2.5) is 0.

As a result we may assume without loss of generality that $d_1(\beta^{(k-1)} - \beta'^{(k-1)}) > 0$. With this assumption we will use the bound

$$\begin{aligned}&\frac{d\big((\beta^{(k)}, \boldsymbol{Z}^{(k)}), (\beta'^{(k)}, \boldsymbol{Z}'^{(k)})\big)}{d\big((\beta^{(k-1)}, \boldsymbol{Z}^{(k-1)}), (\beta'^{(k-1)}, \boldsymbol{Z}'^{(k-1)})\big)}\\ &\leq \max\left(\frac{d_1\big(\beta^{(k)} - \beta'^{(k)}\big)}{d_1\big(\beta^{(k-1)} - \beta'^{(k-1)}\big)}, \frac{d_2\big(\boldsymbol{Z}^{(k)} - \boldsymbol{Z}'^{(k)}\big)}{d_1\big(\beta^{(k-1)} - \beta'^{(k-1)}\big)}\right).\end{aligned} \tag{7.2.6}$$

We begin by studying the update to $\boldsymbol{Z}$. Subtracting $\boldsymbol{x}_i^{\mathsf{T}}\beta$ from both sides of (7.2.1), applying $\Phi(\cdot)$, differentiating with respect to β and gathering up terms, we find that $\frac{\partial}{\partial\beta} Z_i = \lambda_i \boldsymbol{x}_i$ where

$$\lambda_i = \begin{cases} 1 - \dfrac{(1-u_i)\varphi(\boldsymbol{x}_i^{\mathsf{T}}\beta)}{\varphi(Z_i - \boldsymbol{x}_i^{\mathsf{T}}\beta)} & \text{if } Y_i = 1,\\ 1 - \dfrac{u_i\varphi(-\boldsymbol{x}_i^{\mathsf{T}}\beta)}{\varphi(Z_i - \boldsymbol{x}_i^{\mathsf{T}}\beta)} & \text{if } Y_i = 0, \end{cases} \tag{7.2.7}$$

and φ is the $\mathcal{N}(0,1)$ probability density function. It is clear that $\lambda_i < 1$. Next we show that $\lambda_i \geq 0$. We begin by inverting (7.2.1) to get:

$$u_i = \begin{cases} \dfrac{\Phi(Z_i - \boldsymbol{x}_i^{\mathsf{T}}\beta) - \Phi(-\boldsymbol{x}_i^{\mathsf{T}}\beta)}{\Phi(\boldsymbol{x}_i^{\mathsf{T}}\beta)} & \text{if } Y_i = 1,\\ \dfrac{\Phi(Z_i - \boldsymbol{x}_i^{\mathsf{T}}\beta)}{\Phi(-\boldsymbol{x}_i^{\mathsf{T}}\beta)} & \text{if } Y_i = 0. \end{cases} \tag{7.2.8}$$

Substituting (7.2.8) into (7.2.7) and simplifying yields

$$1 - \lambda_i = \begin{cases} \dfrac{\varphi(\boldsymbol{x}_i^{\mathsf{T}}\beta)\Phi(-Z_i + \boldsymbol{x}_i^{\mathsf{T}}\beta)}{\Phi(\boldsymbol{x}_i^{\mathsf{T}}\beta)\varphi(Z_i - \boldsymbol{x}_i^{\mathsf{T}}\beta)} & \text{if } Y_i = 1,\\ \dfrac{\varphi(-\boldsymbol{x}_i^{\mathsf{T}}\beta)\Phi(Z_i - \boldsymbol{x}_i^{\mathsf{T}}\beta)}{\Phi(-\boldsymbol{x}_i^{\mathsf{T}}\beta)\varphi(Z_i - \boldsymbol{x}_i^{\mathsf{T}}\beta)} & \text{if } Y_i = 0. \end{cases} \tag{7.2.9}$$

Now consider the function $\tau(x) = \varphi(x)/\Phi(x)$. This function is nonnegative and decreasing, using a Mill's ratio bound from [14]. When $Y_i = 1$, then $1 - \lambda_i = \tau(\boldsymbol{x}_i^\mathsf{T}\beta)/\tau(\boldsymbol{x}_i^\mathsf{T}\beta - Z_i) \le 1$ because then $Z_i \ge 0$. We also used symmetry of $\varphi(\cdot)$. If instead $Y_i = 0$, then $1 - \lambda_i = \tau(-\boldsymbol{x}_i^\mathsf{T}\beta)/\tau(-\boldsymbol{x}_i^\mathsf{T}\beta + Z_i) \le 1$ because then $Z_i \le 0$. Either way, $1 - \lambda_i \le 1$ and therefore $\lambda_i \in [0, 1)$ for all i.

Writing the previous results in a compact matrix form, we have

$$\frac{\partial \boldsymbol{Z}}{\partial \beta} = \left(\frac{\partial z_i}{\partial \beta_j}\right)_{ij} = \Lambda X$$

where $\Lambda = \Lambda(\beta, \boldsymbol{Z}) = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$. Similarly equation (7.2.2) yields

$$\frac{\partial \beta}{\partial \boldsymbol{Z}} = (X^\mathsf{T}X)^{-1}X^\mathsf{T}.$$

Thus for the $\boldsymbol{Z}$ update with any $\boldsymbol{u}_k \in (0,1)^{n+p}$, since d_1 and d_2 are all norms,

$$\begin{aligned}
\frac{d_2\big(\boldsymbol{Z}^{(k)} - \boldsymbol{Z}'^{(k)}\big)}{d_1\big(\beta^{(k-1)} - \beta'^{(k-1)}\big)} &\le \sup_{\widetilde{\beta}^{(k-1)}, \widetilde{\boldsymbol{Z}}^{(k)}} d_2\left(\frac{\partial \widetilde{\boldsymbol{Z}}^{(k)}}{\partial \widetilde{\beta}^{(k-1)}}\,\xi\right) \\
&\qquad d_1(\xi) = 1 \\
&\le \sup_{\substack{\beta, \boldsymbol{Z} \\ (X\xi)^\mathsf{T}X\xi = 1}} \|\Lambda(\beta, \boldsymbol{Z})X\,\xi\| \\
&< 1. \qquad (7.2.10)
\end{aligned}$$

For the β update, applying the chain rule gives

$$\frac{\partial \beta^{(k)}}{\partial \beta^{(k-1)}} = \frac{\partial \beta^{(k)}}{\partial \boldsymbol{Z}^{(k-1)}}\frac{\partial \boldsymbol{Z}^{(k-1)}}{\partial \beta^{(k-1)}} = (X^\mathsf{T}X)^{-1}X^\mathsf{T}\Lambda X$$

Since d_1 and d_2 are all norms, we have:

$$\frac{d_1\big(\beta^{(k)} - \beta'^{(k)}\big)}{d_1\big(\beta^{(k-1)} - \beta'^{(k-1)}\big)} \le \sup_{\widetilde{\beta},\, d_1(\xi)=1} d_1\left(\frac{\partial \widetilde{\beta}^{(k)}}{\partial \widetilde{\beta}^{(k-1)}}\,\xi\right)$$

and then,

$$\begin{aligned}
\frac{d_1\big(\beta^{(k)} - \beta'^{(k)}\big)}{d_1\big(\beta^{(k-1)} - \beta'^{(k-1)}\big)} &\leq \sup_{\widetilde{\beta},\, d_1(\xi)=1} d_1\left(\frac{\partial\widetilde{\beta}^{(k)}}{\partial\widetilde{\beta}^{(k-1)}}\,\xi\right) \\
&= \sup_{\substack{\beta,\boldsymbol{Z} \\ (X\xi)^{\mathsf{T}}X\xi=1}} d_1\left((X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}\Lambda X\xi\right) \\
&= \sup_{\beta,\boldsymbol{Z},\|\eta\|=1} d_1\left((X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}\Lambda\eta\right) \\
&= \sup_{\beta,\boldsymbol{Z},\|\eta\|=1} \left\|X(X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}\Lambda\eta\right\| \\
&\leq \max_{1\leq i\leq n} \lambda_i \\
&< 1, \qquad (7.2.11)
\end{aligned}$$

using the non-expansive property of the projection matrix $X(X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}$.

By combining (7.2.10) with (7.2.11) we establish the contraction (7.2.5).

Remark. This example fails the boundedness assumption of Theorem 4.3.7. To address this problem, we can force β to stay in a large ball with radius $M \gg 1$. We can do so by modifying the Gibbs Sampler a little bit as follows: for each step, update $\beta_i \mapsto \beta_{i+1}$ as usual. If $d_1(\beta_{i+1}) > M$, let $\beta_{i+1} = \frac{M}{d_1(\beta_{i+1})}\beta_{i+1}$. It is easy to see that it is still a Markov Chain, and since projecting onto a ball is non-expansive, we know this new Markov Chain algorithm is a global contracting mapping on the bounded space $\{\beta : d(\beta) \leq M\}$.

7.3 Global non-expansive Mapping

In this section we use Theorem 5.1.3 to prove the following Theorem about completely uniformly distributed sequences. It is a corollary of the Van der Corput Difference Theorem (cf. Kuipers and Niederreiter [18]). Here we are giving a totally different proof.

Theorem 7.3.1. *For any completely uniformly distributed (CUD) sequence* $(v_i)_{i\geq 1}$*, its partial sum* $s_i = \sum_{j=1}^{i} v_i(\ mod\ 1)$ *is also CUD.*

Proof. We need to show that, for any $k \in \mathbb{N}$, $\{(s_i, s_{i+1}, s_{i+k-1})_{i\geq 1}\}$ uniformly samples $\mathcal{U}[0,1)^k$. We construct a Markov Chain as follows. Define $\mathrm{T} = \{z \in \mathbb{C} : |z| = 1\}$ which is a unit circle, and the state space $\Omega = \mathrm{T}^k$ be the k dimensional Torus. The Markov Chain starts from $\boldsymbol{z}_0 = (1, 1, \cdots, 1) \in \Omega$ and has update function:

$$\phi(\boldsymbol{z}, u) = (z_2, z_3, \cdots, z_k, z_k \times e^{2\pi\sqrt{-1}u}), \quad \boldsymbol{z} = (z_1, z_2, \cdots, z_k), u \in [0,1).$$

The metric on Ω is defined as:

$$d(\boldsymbol{z}, \boldsymbol{z}') = \max\left(\|z_1 - z_1'\|, \cdots, \|z_k - z_k'\|\right), \quad \boldsymbol{z} = (z_1, \cdots, z_k), \boldsymbol{z}' = (z_1', \cdots, z_k')$$

which is equivalent with Euclidean norm. It is easy to see that the update function is global non-expansive. Since the update function is smooth, ϕ is regular. Also, from a probabilistic point of view, the k-step transition ϕ_k is indeed IID sampling from $\mathcal{U}(\Omega)$ if u_i is IID from $\mathcal{U}[0,1)$, which automatically satisfies condition (5.1.2). For any $\delta > 0$, it is easy to see that Ω can be covered by $\left(\lceil \frac{2\pi}{\delta} \rceil\right)^k$ balls with radius δ. Therefore the Bracketing Number $N(\delta) = O(\delta^{-k})$. Hence all the conditions in Theorem 5.1.3 are satisfied. By the conclusion of the Theorem, if we let $\boldsymbol{z}_i = \phi(\boldsymbol{z}_{i-1}, v_i)$ for a CUD sequence $(v_i)_{i\geq 1}$, $\boldsymbol{z}_i$ will consistently sample $\mathcal{U}(\Omega)$. Noticing that

$$\boldsymbol{z}_i = (e^{2\pi\sqrt{-1}s_{i-k+1}}, e^{2\pi\sqrt{-1}s_{i-k+2}}, \cdots, e^{2\pi\sqrt{-1}s_i}), \quad i \geq k$$

we completed our proof. □

Remark. This Theorem tells us, given a CUD sequence, we can get another CUD sequence for free by doing the partial sum. Obviously we can continue doing the partial sum and get another CUD sequence. It will be interesting to investigate how a CUD sequence's uniformity will affect its partial sum's uniformity.

7.4 Contracting on Average

Contracting on average condition is motivated by Gibbs Sampler on rectangle. In this section we discuss the equivalent condition for a two-dimensional systematic Gibbs Sampler to be contracting on average when the target distribution has uniform marginals.

Example 7.4.1 (Copula Model)**.** $\Omega = [0,1] \times [0,1]$. $\pi(x,y)$ is a density function on Ω that we would like to simulate from. Let π_x, π_y be the marginal density of x and y, respectively. Denote by $Q(x,y)$ the CDF of the distribution. Q is called a copula function if the marginal distributions are uniform. I.e., $\pi_x = \pi_y = 1$. Assume $\pi(x,y)$ is a smooth function defined on Ω, and bounded from below: $\pi(x,y) \geq \epsilon > 0$ for some ϵ. Assuming we use inverse method to generate from conditional distribution, the Gibbs Sampler has the following update function :

$$\phi(x,y;u_1,u_2) = \left(F_y^{-1}(u_1), G^{-1}_{F_y^{-1}(u_1)}(u_2)\right)$$

where F_y is the CDF of x conditioning on y, and G_x is the CDF of y conditioning on x. Noticing that ϕ actually does not depend on x, which means it's defining a Markov Chain on $y \in [0,1]$ by

$$\phi(y;u_1,u_2) = G^{-1}_{F_y^{-1}(u_1)}(u_2)$$

By assuming $\pi(x,y)$ to be smooth and bounded from below, we know ϕ is continuously differentiable w.r.t y, and we compute its first derivative as below.

Let $x = F_y^{-1}(u_1)$, then $\phi(y;u_1,u_2) = G_x^{-1}(u_2)$. By chain rule, we have:

$$\frac{\partial \phi}{\partial y} = \frac{\partial \phi}{\partial x}\frac{\partial x}{\partial y} \tag{7.4.1}$$

We just need to compute $\frac{\partial x}{\partial y}$, and $\frac{\partial \phi}{\partial x}$ will be very similar. x satisfies:

$$F_y(x) = u_1 \Leftrightarrow \frac{\int_0^x \pi(w,y)\,\mathrm{d}w}{\pi_y} = u_1 \Leftrightarrow \int_0^x \pi(w,y)\,\mathrm{d}w = u_1 \tag{7.4.2}$$

since y is marginally uniformly distributed. Taking derivative last expression w.r.t y on both sides, we get:

$$x'(y)\pi(x,y) + \int_0^x \frac{\partial \pi}{\partial y}(w,y)\,\mathrm{d}w = 0 \Rightarrow x'(y) = -\frac{\int_0^x \frac{\partial \pi}{\partial y}(w,y)\,\mathrm{d}w}{\pi(x,y)} \tag{7.4.3}$$

Equation (7.4.3) can be restated in terms of the copula function $Q(x,y)$ as follows:

$$x'(y) = -\frac{Q_{yy}}{Q_{xy}}(x,y)$$

Similarly, $\frac{\partial \phi}{\partial x} = -\frac{Q_{xx}}{Q_{xy}}(x,\phi)$. Therefore, substituting the expressions into (7.4.1), we get

$$\frac{\partial \phi}{\partial y} = \frac{Q_{yy}(x,y)Q_{xx}(x,\phi)}{Q_{xy}(x,y)Q_{xy}(x,\phi)} \tag{7.4.4}$$

The contracting on average property (5.2.8) thus can be written as:

$$\mathbf{E}_{\pi_y \otimes \mathcal{U}[0,1)^2} \log \left| \frac{\partial \phi}{\partial y} \right| < 0 \Leftrightarrow \mathbf{E}_{\pi_y \otimes \mathcal{U}[0,1)^2} \log \left| \frac{Q_{yy}(x,y)Q_{xx}(x,\phi)}{Q_{xy}(x,y)Q_{xy}(x,\phi)} \right| < 0 \tag{7.4.5}$$

where π_y denotes the marginal distribution of y (in our case, uniform distribution). A key observation is, if $y \sim \pi_y$, then $(x,y) \sim \pi(x,y)$ as well as $(x,\phi) \sim \pi(x,y)$. Therefore, the contracting on average condition is equivalent with:

$$\begin{aligned} &\mathbf{E}_{\pi_y} \log \left| \frac{Q_{yy}(x,y)Q_{xx}(x,\phi)}{Q_{xy}(x,y)Q_{xy}(x,\phi)} \right| < 0 \\ \Leftrightarrow &\int_0^1 \int_0^1 \log \left| \frac{Q_{xx}Q_{yy}}{Q_{xy}^2}(x,y) \right| Q_{xy}(x,y)\,\mathrm{d}x\,\mathrm{d}y < 0 \end{aligned} \tag{7.4.6}$$

Thus, for a given copula function Q, we can use inequality (7.4.6) to check whether the systematic Gibbs Sampler is contracting on average or not.

Chapter 8

MCQMC in Practice

The results from the previous chapters indicate that a MCQMC algorithm gives consistent result by replacing IID sequence with a CUD sequence under certain conditions. In this chapter we are going to discuss in more detail how in practice we implement the MCQMC algorithm. It involves two tasks: first, we need a CUD sequence or array - CUD sequence generator, which will be described in the first section. Second, having the CUD sequence generator at hand, how do we feed the numbers into a MCMC algorithm, which will be the main topic of the second section. We will also compare the performance of MCQMC and usual MCMC through some simulation examples. Most of the results in this chapter have appeared in [4]. The CUD sequence generator we used is contructed by Prof. Matsumoto and Prof. Nishimura.

8.1 Linear Feedback Shift Register

Levin [24] gives several different ways of constructing CUD sequences, but not convenient to implement. Tribble in [45] compared several different generators and he got the best results from Linear Feedback Shift Registers (LFSR), even though the LFSR sequences he used was limited. LSFR has some distinctive features that are beneficial to MCQMC algorithms. In this section we will discuss the (binary) LFSR method and describe the generator we used for our simulation experiments. Prof. Matsumoto and Prof. Nishimura take full credit for designing and implementing the

LFSR generators that we use .

Let GF(2) be the Galois field with two elements $\{0,1\}$. For notation simplicity, define $\mathbf{0} = (0,0,\cdots,0) \in \{0,1\}^m$. For any positive integer $m \in \mathbb{N}$, a LFSR of degree m generates a sequence of 0 or 1 bits through the initial states $(b_0, b_1, \cdots, b_{m-1}) \in \{0,1\}^m$ which are not all 0, and $(a_{m-1}, a_{m-2}, \cdots, a_0) \in \{0,1\}^m : a_0 = 1$. For $i \geq m$, b_i is generated recursively through the following equation:

$$b_i = a_{m-1}b_{i-1} + a_{m-2}b_{i-2} + \cdots + a_1 b_{i-m+1} + a_0 b_{i-m}, \quad i \geq m. \tag{8.1.1}$$

Since the algorithm is deterministic and the $m-$tuple $(b_i, b_{i+1}, \cdots, b_{i+m-1}) \in GF(2)^m$ can only take 2^m possible values, the sequence $(b_i)_{i\geq 0}$ is periodic with period at most 2^m. In fact, if for some $i \geq 0$, $(b_i, b_{i+1}, \cdots, b_{i+m-1}) = \mathbf{0}$, then by the recurrence relation (8.1.1), $b_{i'} = 0$ for all $i' \geq i+m$. It implies that the initial state $(b_0, \cdots, b_{m-1}) = \mathbf{0}$ which contradicts to our assumption. Therefore, if the initial states are not all zero, the LFSR will generate a sequence of numbers with period at most $2^m - 1$. And the following Theorem fully determines the condition for which the maximal period is achieved.

Theorem 8.1.1. *The linear feedback shift register algorithm attains the maximal length of period $2^m - 1$ if and only if its characteristic polynomial*

$$z^m + a_{m-1}z^{m-1} + a_{m-2}z^{m-2} + \cdots + a_1 z + a_0$$

is primitive over Galois Field GF(2).

The definition of a polynomial being primitive is beyond the scope of this paper. See [37] for a detailed discussion about primitive polynomials. We would like to point out that for each degree $m > 0, m \in \mathbb{N}$, there are a total of $\frac{1}{m}\phi(2^m - 1)$ primitive polynomials. Here ϕ is the Euler function. A simple corollary is, for any m, there is at least one primitive polynomial which by Theorem 8.1.1 generates a LFSR sequence of maximal period $2^m - 1$.

Assume $(a_{m-1}, a_{m-2}, \cdots, a_0)$ corresponds to a primitive polynomial. Then for any initial state $(b_0, b_1, \cdots, b_{m-1}) \neq \mathbf{0}$, we can generate $b_i : m \leq i \leq 2^m - 2$ using recurrence relation (8.1.1) and $b_{2^m-1} = b_0, b_{2^m} = b_1, \cdots$ etc. The consecutive $m-$blocks of this sequence $(b_i)_{i=0}^{2^m-2}$ has every $m-$ tuple of bits in the set $\{0,1\}^m \setminus \mathbf{0}$. Therefore, for any integer $g > 0$ such that $gcd(g, 2^m - 1) = 1$, which is called an offset, we can define

$$v_i = \sum_{j=0}^{m-1} b_{gi+j} 2^{-j-1}, \text{ for } i = 0, 1, \cdots, 2^m - 2 \tag{8.1.2}$$

Because $gcd(g, 2^m - 1) = 1$, it is easy to see that (v_i)s are taking all distinct values. If we partition $[0, 1)$ into 2^m equalsize subintervals, there is one v_i lying in each of them except the lowest one, since v_i is never 0. The offset g is usually chosen to be greater than one, because if $g = 1$, the second digit of v_i will be the same as the first digit of v_{i+1}, which renders $(v_i, v_{i+1})_{i=0}^{2^m-2}$ to be strongly correlated.

In order to make a sequence suitable for MCQMC algorithms, as discussed before one dimension uniformity is not sufficient. In ideal situation we would like to have $(v_i, v_{i+1}, \cdots, v_{i+k-1})_{i=0}^{2^m-2}$ being uniformly distributed on $k-$ dimension unit cube for all $k = 1, 2, \cdots$. We could use Star-Discrepancy to measure how uniformly they are distributed, but as pointed out in Chapter 2, computing Star Discrepancy of a certain point set is computationally very expensive for large k. We here give another criterion for uniformity on $[0, 1)^k$, which is related to Star Discrepancy but not exactly the same.

Definition 8.1.2. *For any integer $k > 0, \nu > 0$, we can divide interval $[0, 1)$ into 2^ν equal pieces, which yields a partition of $[0, 1)^k$ into $2^{k\nu}$ equalsize cells. A sequence $(x_i)_{i=0}^{2^m-2}$ is called $k-$ dimensionally equidistributed with ν bit accuracy, if each cell contains exactly same number of points of $(x_i, x_{i+1}, \cdots, x_{i+k-1})_{i=0}^{2^m-2}$ except for the cell at the origin that contains one less.*

Since there are a total of $2^{k\nu}$ cells while only $2^m - 1$ points, we have $k\nu \leq m \Rightarrow k \leq \lfloor \frac{m}{\nu} \rfloor$. Define

$$k(\nu) = \max\{\nu : (x_i) \text{ is } k \text{ dimensionally equidistributed with } \nu \text{ bit accuracy } \}$$

Definition 8.1.3 (Fully Equidistributed Sequence)**.** *A sequence* $(x_i)_{i=0}^{2^m-2} \in [0,1)$ *is called Fully Equidistributed (FE) if it is* k *dimensionally equidistributed with* ν *bit accuracy for all* $1 \le \nu \le m$ *and* $1 \le k \le \lfloor \frac{m}{\nu} \rfloor$.

In our situation, since we know v_i has binary expansion $v_i = 0.b_{gi}b_{gi+1}\cdots b_{gi+m-1}$, we can define Fully Equidistributed in a different but equivalent way. Let $M(k,\nu)$ denote the $k \times \nu$ binary matrices. The above definition is equivalent to that the multiset of $k \times \nu$ matrices

$$\Phi_{k,\nu} \stackrel{\Delta}{=} \left\{ (b_{g(i+l),j})_{l=0,\cdots,k-1;j=0,\cdots,\nu-1} \middle| 0 \le i \le 2^m - 2 \right\} \tag{8.1.3}$$

contains every element of $M(k,\nu)$ with the same multiplicity, except the 0 matrix with one less multiplicity, for any $\nu = 1\cdots, m$ and $1 \le k \le \lfloor \frac{m}{\nu} \rfloor$.

Our generator is constructed in the following way: for any $10 \le m \le 32$, we choose one primitive polynomial of degree m over GF(2), which will generate a sequence $(b_0, b_1, \cdots, b_{2^m-2})$ of period 2^m-1. Then we search in ascending order of $1 < g < 4000$ such that the FE condition is satisfied. The LFSR sequence we output is

$$(v_i)_{i=0}^{2^m-2} : v_i = \sum_{j=0}^{m-1} b_{gi+j} 2^{-j-1}, \text{ for } i = 0, 1, \cdots, 2^m - 2.$$

The Fully Equidistributed LFSR sequences are suitable for MCQMC algorithms, since the k dimensional vectors formed by taking consecutive numbers are of low discrepancy. In fact, assuming we are able to construct Fully Equidistributed LFSR sequence of any order m, piling them up will give us a CUD array, as proved by Theorem 4.5.4 of Tribble [45] :

Theorem 8.1.4. *For any integer* $m > 0$*, define a fully equidistributed LFSR sequence of length* $2^m - 1$*. Then the collection of LFSR sequences are CUD-array.*

Proof. See [45], page 49-50. □

Here we list the LFSR generators that we use. Each generator is indexed by its degree m and corresponds to a primitive polynomial over GF(2) and an offset

Table 8.1: LFSR generators: Polynomials and Offsets

m	Polynomial	Offset g	m	Polynomial	Offset g
10:	$t^{10}+t^3+1$	115	22:	$t^{22}+t+1$	1336
11:	$t^{11}+t^2+1$	291	23:	$t^{23}+t^5+1$	1236
12:	$t^{12}+t^6+t^4+t+1$	172	24:	$t^{24}+t^4+t^3+t+1$	1511
13:	$t^{13}+t^4+t^3+t+1$	267	25:	$t^{25}+t^3+1$	1445
14:	$t^{14}+t^5+t^3+t+1$	332	26:	$t^{26}+t^6+t^2+t+1$	1906
15:	$t^{15}+t+1$	388	27:	$t^{27}+t^5+t^2+t+1$	1875
16:	$t^{16}+t^5+t^3+t^2+1$	283	28:	$t^{28}+t^3+1$	2573
17:	$t^{17}+t^3+1$	514	29:	$t^{29}+t^2+1$	2633
18:	$t^{18}+t^7+1$	698	30:	$t^{30}+t^6+t^4+t+1$	2423
19:	$t^{19}+t^5+t^2+t+1$	706	31:	$t^{31}+t^3+1$	3573
20:	$t^{20}+t^3+1$	1304	32:	$t^{32}+t^7+t^6+t^2+1$	3632
21:	$t^{21}+t^2+1$	920			

$g > 0, g \in \mathbb{N}$.

One advantage of using a Mini-LFSR generator ($d \leq 32$) is that, we can actually run through the whole sequence instead of just visiting a short segment of it. The sequence is constructed so that as a whole the numbers are equidistributed, therefore completely running through the sequence is more preferrable than only using a small piece of it. In comparison, it is infeasible to run through the entire sequence produced by usual random number generators such as Mersenne Twister 19937.

A distinctive feature of a LFSR sequence is, the overlapping k dimensional blocks $(v_i, v_{i+1}, \cdots, v_{i+k-1})_{i=0}^{2^m-2}$ are close to uniformly distributed, while some of the other quasi-Monte Carlo generators only focus on the non-overlapping blocks. We have already seen in Chapter 6 Theorem 6.2.3, for a MCQMC algorithm to attain optimal convergence rate, we do require the overlapping blocks be more uniform than IID sequence. This feature might be one of the reasons why Tribble found the best results from using LFSR.

We would like to mention that, the existence of Fully Equidistributed LFSR sequence of arbitrary length $2^m - 1$ is still an open problem. For $10 \leq m \leq 32$, by searching we found the proper offset g as listed in Table 8.1. However it's still unknown whether for any $m \in \mathbb{N}$ such offsets can be found.

8.2 Algorithm Implementation

In this section we describe for a given Markov Chain update function ϕ, how we feed the LFSR sequence into the updating and generate the Markov Chain. Throughout this section we assume $\phi(\boldsymbol{x}, \boldsymbol{u})$ is the update function where $\boldsymbol{u} \in [0,1)^d$, and the MCQMC algorithm starts at $\boldsymbol{x}_0 \in \Omega$.

8.2.1 Construction of Variate Matrix

Let $N = 2^m - 1$ be the length of a fully equidistributed LFSR sequence $(v_i)_{i=1}^N$. It is natural to set $\boldsymbol{u}_i = (v_{di-d+1}, \cdots, v_{di})$. The problem with this strategy is, if $gcd(d, N) \neq 1$, we are not going to traverse all the possible d tuples formed by consecutive numbers in the sequence $(v_i)_{i=1}^N$. Therefore there is no guarantee on the uniformity of $(\boldsymbol{u}_i)_{i=1}^N$ constructed in this way. To address this problem, we find the smallest $y \in \mathbb{N}, y \geq d$ such that $gcd(y, N) = 1$, and form a matrix called variate matrix as follows:

$$\begin{pmatrix} v_1 & v_2 & \cdots & v_d \\ v_{y+1} & v_{y+2} & \cdots & v_{y+d} \\ v_{2y+1} & v_{2y+2} & \cdots & v_{2y+d} \\ \vdots & \vdots & \ddots & \vdots \\ v_{Ny-y+1} & v_{Ny-y+2} & \cdots & v_{Ny-y+d} \end{pmatrix} \tag{8.2.1}$$

The (i,j)th entry of this matrix is $v_{(i-1)y+j}$. We set $\boldsymbol{u}_i$ to be the ith row of this matrix, for $1 \leq i \leq N$. This strategy maintains a balance among $\boldsymbol{u}_i$s, since $(\boldsymbol{u}_i)_{i=1}^N$ will be a permutation of $(v_i, \cdots, v_{i+d-1})_{i=1}^N$, thus have the same equidistribution property.

Secondly, if y is not too big, $(\boldsymbol{u}_i, \boldsymbol{u}_{i+1}, \cdots, \boldsymbol{u}_{i+k-1})_{i=1}^N$ will also have certain equidistribution property if $ky \leq m$. This means not only $\boldsymbol{u}_i$ itself are well balanced, the vectors formed by taking k consecutive $\boldsymbol{u}_i$'s are also approximately uniformly distributed, which is desirable in MCQMC simulation. Noticing $N = 2^m - 1$ is an odd number, it is easy to see $gcd(2^r, N) = 1$ for any $r \in \mathbb{N}$. Hence the y we choose is at most as big as $2d$.

In the experiments we did in this paper, for simplicity we choose $y = 2^r$ for some $r \in \mathbb{N}$. The optimal situation is when $N = 2^m - 1$ being a prime number itself (usually called Mersenne Prime), in which case we can simply choose $y = d$.

8.2.2 Randomization

As we have discussed in Chapter 2, randomized quasi-Monte Carlo is an important tool in two aspects: first, it is able to eliminate the bias. Second, we can estimate the error much easier than using the Koksma-Hlawka and Star-Discrepancy. For MCQMC algorithms, we still hope randomization can help us in these two aspects. It is unlikely that we can fully eliminate the bias by adding randomness to the deterministic CUD sequence, since even if we are using IID squence as the driving sequence the bias is not equal to 0. However, we may still hope to reduce the bias by randomization. There hasn't been too much research done on this problem yet, and it could be an interesting topic for future research to investigate the bias reduction of Randomized MCQMC.

In our experiments we adopt the digit shift on our variate matrix. Recall the variate matrix being constructed by a degree m LFSR sequence $(v_i)_{i=1}^N$ is a $(2^m - 1) \times d$ matrix whose (i, j)th entry is $v_{y(i-1)+j}$. We will set $\boldsymbol{u}_i$ to be the ith row of this matrix and feed them into the Markov Chain update $\boldsymbol{x}_i = \phi(\boldsymbol{x}_{i-1}, \boldsymbol{u}_i)$. It seems reasonable to randomize the matrix so that each $(\boldsymbol{u}_i)_{i=1}^N$ is marginally uniformly distributed on $[0, 1)^d$ while preserving the equidistribution property among the columns. Therefore

Table 8.2: Digit Shift

Digit Shift	
1	For each column $1 \leq j \leq d$, pick an IID random number z_j
2	Write down z_j's binary expansion $0.z_j^1 z_j^2 \cdots$
3	For the (i,j)th entry, write down its binary expansion $0.a^1 a^2 \cdots$
4	$a^k \mapsto a^k + z_j^k \pmod 2$

the digit shift scheme fits our goals very well.

The digit shift scheme is summarized in Table 8.2. It is easy to see that after the randomization, each row $\boldsymbol{u}_i$ has marginal distribution $\mathcal{U}[0,1)^d$ while the equidistribution property is preserved. In real practice, since each entry of the variate matrix has length d binary expansion, we only do the digit shift up to the dth digit. As a result, the ith row $\boldsymbol{u}_i$ is not exactly following $\mathcal{U}[0,1)^d$ but uniformly distributed on a grid formed by dividing each edge $[0,1)$ into 2^d equalsize pieces. When d is large, it is very close to uniform distribution.

Our MCQMC algorithm can be summarized as in Table 8.3. Notice in Step 4 we add a row of all 0's to be the first row of the variate matrix. This is not going to affect the efficiency or consistency of MCQMC algorithm since one row out of $N = 2^m - 1$ rows is quite negeligible. On the other hand, as we know by the definition of a Fully-Distributed sequence, if we devide $[0,1)^d$ into 2^d equalsize cubes, each cube is going to contain exactly one row of the variate matrix (viewed as a d dimensional vector) except for the lowest cube. Therefore adding a row of all 0s will make the points more balanced. Adding such a row or not shouldn't affect the estimation a lot. In our experiments we choose to employ this practice.

8.3 Examples

In this section we give several simulation examples to compare the performance of MCQMC and usual MCMC. For MCQMC we use the algorithm discussed in the previous section, while the error is estimated by $K = 100$ times digit shift. The

Table 8.3: MCQMC Algorithm

MCQMC Algorithm	
1	Choose degree m of the LFSR generator, $N = 2^m - 1$.
2	Generate $(v_i)_{i=1}^N$ using the primitive polynomial and offset listed in Table 8.1.
3	Find $y \leq 2d$ such that $gcd(y, N) = 1$. Construct the variate matrix.
4	Add a row of all 0s to the top of the variate matrix.
5	Randomize the variate matrix by digit shift.
6	Let $\boldsymbol{u}_i$ be the ith row of the variate matrix after randomization.
7	Generate Markov Chain samples $\boldsymbol{x}_i = \phi(\boldsymbol{x}_{i-1}, \boldsymbol{u}_i)$.
8	Report $\frac{1}{N+1}\sum_{i=1}^{N+1} f(\boldsymbol{x}_i)$ to be an estimate of $\mathbf{E}_\pi f$.
9	Repeat 5-8 K times where K is a predetermined number.

random numbers needed in digit shift are generated by Mersenne Twister 19937. For usual MCMC, we also use Mersenne Twister 19937 to generate the sequence of IID numbers. To generate a standard normal distribution, we use the Inverse Method, i.e., inverting the CDF of standard normal distribution. Although the examples are easy, various aspects of the MCQMC algorithm can be illustrated.

8.3.1 2 Dimensional Gibbs Sampling

In this example we use systematic Gibbs Sampler to generate two dimensional Normal Distribution $N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$ for $\rho \in (-1, 1)$. In reality we can generate two independent normal distributed variables and then multiply on the left by a matrix to get such a distribution, so Gibbs Sampler is not necessary. But it serves great as a test for our algorithms. Here we fix the initial value of Gibbs Sampler to be $(x_0, y_0) = (1, 1)$ and for each step, we update $x_i \mapsto x_{i+1}$ first and then update $y_i \mapsto y_{i+1}$. The conditional distribution can be easily seen to follow a normal distribution therefore easy to generate through the Inverse Method. We compares the LFSR to IID sampling with $2^{12} = 4{,}096$ steps to $2^{20} = 1{,}048{,}576$ steps.

First we study the estimation of $\mathbf{E}_\pi x$. Table 8.4 and Figure 8.1 compares LFSR to IID when $\rho = 0$ and $\rho = 0.9$. When $\rho = 0.0$, it is essentially independent sampling, in which case we can expect a better performance by using MCQMC. Even in the

Table 8.4: $\log_2$(Root Mean Squared Error) for Gibbs mean

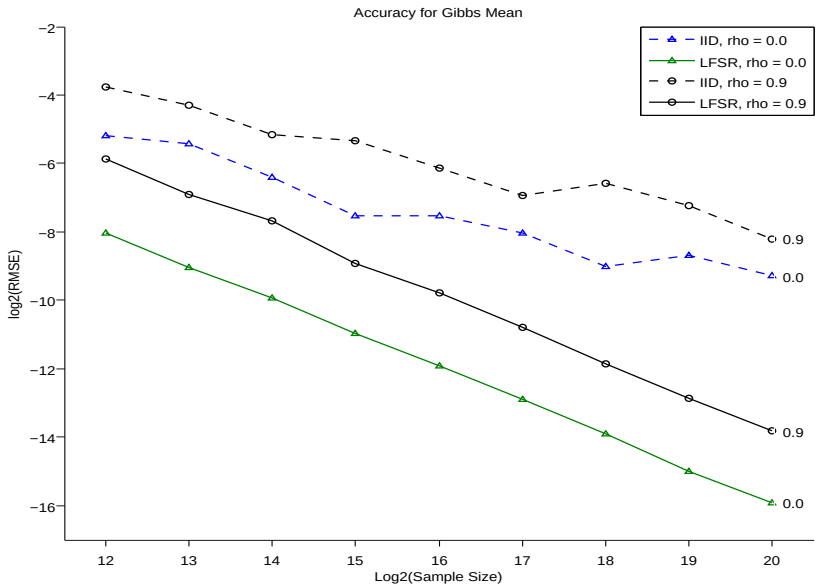
N = 2^{12} to 2^{20}, $\rho = 0.0$									
$\log_2(N)$	12	13	14	15	16	17	18	19	20
LFSR	-8.01	-9.03	-9.92	-10.97	-11.90	-12.90	-13.91	-15.00	-15.92
IID	-5.19	-5.41	-6.40	-7.51	-7.52	-8.04	-9.01	-8.67	-9.27
N = 2^{12} to 2^{20}, $\rho = 0.9$									
$\log_2(N)$	12	13	14	15	16	17	18	19	20
LFSR	-5.85	-6.90	-7.68	-8.91	-9.78	-10.79	-11.84	-12.87	-13.81
IID	-3.77	-4.29	-5.16	-5.33	-6.13	-6.93	-6.58	-7.22	-8.20

case of $\rho = 0.9$ when the Markov Chain is mixing slower, MCQMC still outperforms the usual MCMC by a large factor and demonstrates a better convergence rate. The reason is, a two dimensional Gaussian Gibbs sampling is equivalent with an AR(1) process, which we have shown in Chapter 6 that should give MCQMC a better convergence rate. We would like to point out that the bounded support assumption in Theorem 6.3.3 is not satisfied in this case, but we still observe an almost $\frac{1}{N}$ convergence rate.

The mean under Gibbs sampling is easier than most problems we will face. To make it more difficult now we consider estimating the correlation ρ itself. From Table 8.5 and Figure 8.2 we can see LFSR still outperforms IID in this case. Also, we can see the slope of the LFSR curves are higher than the slope of the IID curves, which implies that MCQMC might be giving a higher convergence rate.

The main feature of 2-dimensional Gaussian Gibbs sampling is that it is in the form of an AR(1) process. We could not expect a similar strong performance by MCQMC if we are running a Gibbs sampler to generate some other two dimensional distributions. Another feature is that the conditional distribution is Normal, which we can easily generate by Inverse Method. For general case the conditional distribution might require some Acceptance - Rejection step to generate, which could render MCQMC less efficient.

Figure 8.1: Numerical results for bivariate Gaussian Gibbs sampling. LFSR = solid and IID = dashed. The goal is to estimate the mean. The correlation is marked at the right. Y axis = $\log_2 \sqrt{\text{Mean Squared Error}}$



8.3.2 M/M/1 Queuing system

The M/M/1 model is the simplest one server queuing model. It is popular because of its simplicity and analytical tractability. It indicates a system where the customers arrive as a Poisson process with intensity $\lambda > 0$ and the service time is exponentially distributed with intensity $\mu > 0$. In order to make the system stable, we need the service time to be shorter on average than the customer arrival time, i.e, $\mu > \lambda$. Let W_n be the waiting time of the nth customer, S_n be the service time of the nth customer and τ_n be the time interval between the nth customer and the $(n-1)$th customer. Then we have the following recurrence relation:

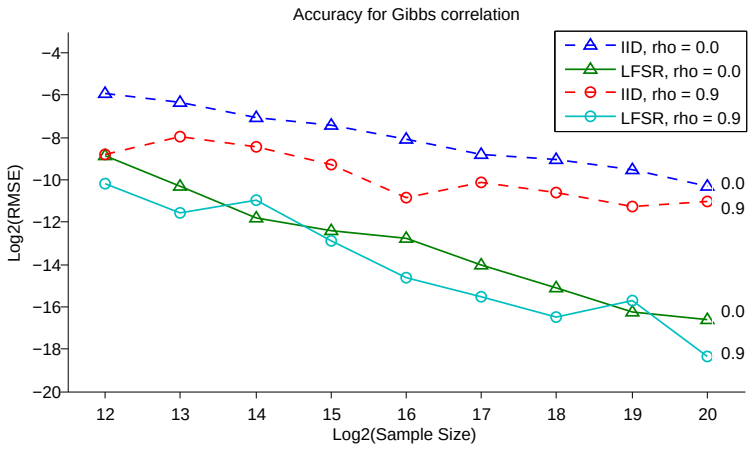
$$W_{n+1} = \max(W_n + S_n - \tau_n, 0), \quad S_n \sim \text{Exp}(\mu), \tau_n \sim \text{Exp}(\lambda) \tag{8.3.1}$$

Under stationarity the average waiting time $\mathbf{E}_\pi W_n = \frac{\lambda}{\mu(\mu-\lambda)}$. For example, see [30]. We would like to compare the performance of MCQMC and MCMC for estimating the average waiting time. The parameters we are using are $\lambda = 0.5$ and $\mu = 1$. We

Table 8.5: $\log_2$(Root Mean Squared Error) for Gibbs correlation

N = 2^{12} to 2^{20}, $\rho = 0.0$									
$\log_2(N)$	12	13	14	15	16	17	18	19	20
LFSR	-8.83	-10.28	-11.80	-12.41	-12.73	-14.04	-15.12	-16.23	-16.60
IID	-5.90	-6.30	-7.03	-7.40	-8.05	-8.80	-9.02	-9.50	-10.27
N = 2^{12} to 2^{20}, $\rho = 0.9$									
$\log_2(N)$	12	13	14	15	16	17	18	19	20
LFSR	-10.14	-11.57	-10.96	-12.89	-14.60	-15.53	-16.47	-15.68	-18.34
IID	-8.82	-7.93	-8.45	-9.29	-10.84	-10.09	-10.59	-11.25	-11.00

Figure 8.2: Numerical results for bivariate Gaussian Gibbs sampling. LFSR = solid and IID = dashed. The goal is to estimate the correlation. The true correlation is marked at the right. Y axis = $\log_2(\sqrt{\text{Mean Squared Error}})$



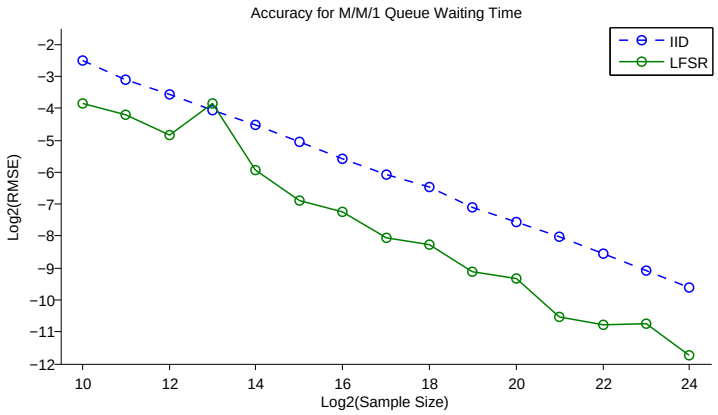
comepare LFSR to IID with 2^{10} steps to 2^{24} steps.

The results are shown in Table 8.6 and Figure 8.3. LFSR does better than IID by a factor of 4 to 8 in terms of reducing the Root Mean Squared Error. Assuming the bias is negligible, it means that LFSR is reducing the variance by a factor of 16 to 64, which is quite significant. The $\max(\cdot, 0)$ function is unsmooth at 0, which in theory could cause a big or even infinite Hardy-Krause variation. It is still unclear why MCQMC works very well under such unsmooth situations.

Table 8.6: $\log_2$(RMSE) for Average Waiting Time

N = 2^{10} to 2^{17}								
$\log_2(N)$	10	11	12	13	14	15	16	17
LFSR	-3.84	-4.18	- 4.85	-3.84	-5.94	-6.87	-7.24	-8.05
IID	-2.51	-3.10	-3.58	-4.04	-4.53	-5.05	-5.57	-6.06
N = 2^{18} to 2^{24}								
$\log_2(N)$	18	19	20	21	22	23	24	
LFSR	-8.26	-9.10	-9.32	-10.54	-10.76	-10.76	-11.73	
IID	-6.48	-7.09	-7.55	-8.02	-8.55	-9.08	-9.59	

Figure 8.3: Numerical results for M/M/1 queue average waiting time. LFSR = solid and IID = dashed.



8.3.3 Dickman's Distribution

The Dickman's Distribution comes out from analytic number theory. It is a special case of Vervaat perpetuity random variables which occur in various fields such as financial modeling, hydrology and number theory. See Dickman [9] and Devroye et al [7]. It has the following form:

$$X_{n+1} = (X_n + 1) \times U_{n+1}, \quad \text{where } U_{n+1} \sim \mathcal{U}[0,1) \tag{8.3.2}$$

We would like to compare MCQMC to MCMC for estimating the moments of X_n under stationarity. We chose moments of X_n to be the test functions because we can

Table 8.7: $\log_2$(RMSE) for Dickman's Distribution

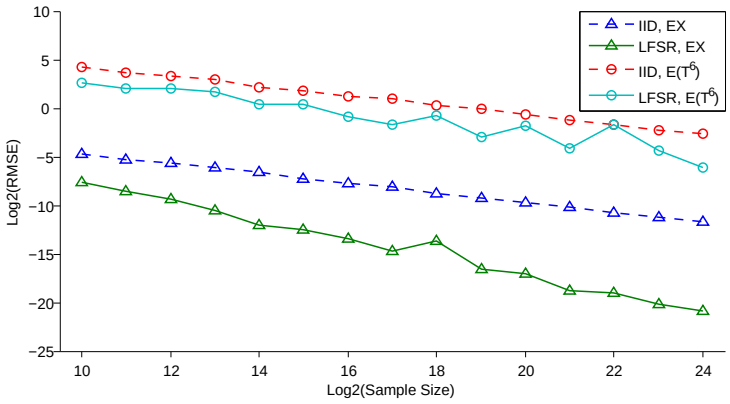
Estimating $\mathbf{E}_\pi X$								
$\log_2(N)$	10	11	12	13	14	15	16	17
LFSR	-7.66	-8.48	-9.34	-10.54	-12.08	-12.47	-13.46	-14.67
IID	-4.69	-5.32	-5.67	-6.15	-6.62	-7.25	-7.68	-8.13
$\log_2(N)$	18	19	20	21	22	23	24	
LFSR	-13.63	-16.52	-17.0815	-18.75	-19.06	-20.17	-20.86	
IID	-8.76	-9.24	-9.66	-10.15	-10.77	-11.16	-11.69	
Estimating $\mathbf{E}_\pi X^6$								
$\log_2(N)$	10	11	12	13	14	15	16	17
LFSR	2.60	1.99	2.02	1.69	0.37	0.42	-0.90	-1.63
IID	4.19	3.70	3.36	2.95	2.18	1.79	1.22	0.93
$\log_2(N)$	18	19	20	21	22	23	24	
LFSR	-0.71	-3.01	-1.81	-4.11	-1.69	-4.36	-6.12	
IID	0.23	-0.112	-0.6317	-1.20	-1.64	-2.24	-2.66	

find the value analytically so that we can compute the mean squared error of MCQMC and MCMC. Also, any smooth function can be well approximated by polynomials on bounded interval, therefore a better performance on estimating $\mathbf{E}_\pi X_n^k, k = 1, 2, \cdots$ is an evidence of superiority of one method over the other. As shown in Table 8.7 and Figure 8.4, we are comparing LFSR to IID for estimating the first and sixth moments. As we can see, when estimating the first moment, LFSR performs much stronger than IID. This is because the first moment is the expectation of a linear function, in which situation quasi-Monte Carlo performs the best. On the other hand, when estimating the sixth moment, LFSR does slightly better than IID but not very significant. The extra curvature brought by the test function x^6 makes the quasi-Monte Carlo harder to outperform.

8.3.4 Garch Model

This model is more challenging than the previous ones. We are going to study the problem of using Garch model to price a European Call Option. Let X_t be the stock price at time t and T be the maturity of the Call Option. Assume the interest rate is a constant $r \geq 0$. By Duan [10], the Garch(1,1) option pricing model under the risk

Figure 8.4: Numerical results for Dickman's Distribution. LFSR = solid and IID = dashed. The goal is to estimate the first and sixth moments of the stationary distribution.



neutral world can be written as:

$$\log\left(\frac{X_t}{X_{t-1}}\right) = r - \frac{1}{2}h_t + \epsilon_t, \quad 1 \le t \le T, \quad \text{where} \tag{8.3.3}$$

$$\epsilon_t \sim N(0, h_t), \quad \textit{and} \tag{8.3.4}$$

$$h_t = \alpha_0 + \alpha_1(\epsilon_{t-1} - \lambda\sqrt{h_{t-1}})^2 + \beta_1 h_{t-1}. \tag{8.3.5}$$

Here λ can be viewed as the market price of risk. The Garch model is a useful tool to model the phenomemon of "Volatility Clustering". Also it has a heavier tail than Log Normal distribution, which can be used to explain the Volatility Skew and Smile. The parameter values, from Duan [10] were $r = 0$, $\lambda = 7.452 \times 10^{-3}$, $T = 30$(days), $\alpha_0 = 1.525 \times 10^{-5}$, $\alpha_1 = 0.1883$ and $\beta_1 = 0.7162$. The process starts with $h_0 = 0.64\sigma^2$ where $\sigma^2 = 0.2413$ is the stationary variance of X_t.

The results are shown in Table 8.8 and Figure 8.5. We are simulating the value of a European call option with strike price $K = 1$. X_0 is chosen from $X_0 \in \{0.8, 0.9, 1.0, 1.2\}$. By Risk-Neutural Pricing theory, the value of a European call option is equal to the discounted expected payoff under the risk-neutral world. I.e., Price(call option) $= \mathbf{E}\left(e^{-rT}(X_T - K)^+\right)$.

Table 8.8: $\log_2$(Root Mean Squared Error) for Garch option pricing

N = 2^{11} to 2^{20}, $X_0 = 0.8$

$\log_2(N)$	11	12	13	14	15	16	17	18	19	20
LFSR	-13.3	-14.4	-15.1	-15.6	-16.1	-16.5	-17.1	-17.6	-18.1	-18.4
IID	-12.8	-14.6	-14.7	-14.4	-15.1	-15.9	-16.8	-17.6	-17.9	-18.1

N = 2^{11} to 2^{20}, $X_0 = 0.9$

$\log_2(N)$	11	12	13	14	15	16	17	18	19	20
LFSR	-12.0	-12.9	-12.6	-14.1	-15.2	-15.2	-15.9	-17.0	-17.2	-17.2
IID	-11.1	-12.3	-12.5	-13.0	-13.2	-14.4	-14.5	-15.1	-15.9	-15.6

N = 2^{11} to 2^{20}, $X_0 = 1.0$

$\log_2(N)$	11	12	13	14	15	16	17	18	19	20
LFSR	-11.4	-12.2	-11.4	-13.5	-14.5	-15.1	-15.2	-15.6	-16.8	-17.6
IID	-10.0	-10.3	-11.1	-11.5	-12.1	-12.2	-12.5	-13.5	-13.4	-13.7

N = 2^{11} to 2^{20}, $X_0 = 1.2$

$\log_2(N)$	11	12	13	14	15	16	17	18	19	20
LFSR	-9.5	-10.4	-10.2	-12.2	-13.4	-14.2	-15.4	-16.1	-16.9	-17.6
IID	-8.8	-9.4	-10.1	-10.3	-11.1	-11.2	-11.3	-12.6	-12.0	-13.3

In this example we see MCQMC performs better than MCMC by a large factor. An interesting feature is, when $X_0 = 1.2$ which means the call option is deep in the money, MCQMC presents a very significant improvement over MCMC. MCQMC does not perform as well when the call option is at the money or out of the money. One possible reason is, when it is deep in the money, the payoff function $(X_T - K)^+$ is essentially the same as $(X_T - K)$ which is linear. Quasi-Monte Carlo works the best when the fuction is or close to linear functions. This might explain why the best convergence rate comes from the deep-in-the-money case.

8.3.5 Heston's Stochastic Volatility Model

Heston's Stochastic Volatility Model (Zhu, [47]) is very similar to Garch model in the sense that they both assume a non-deterministic process for the volatility. Unlike Garch model being discrete, stochastic volatility model is a continuous time diffusion

Table 8.9: $\log_2$(Root Mean Squared Error) for Stochastic Volatility Model

N = 2^{11} to 2^{17}							
$\log_2(N)$	11	12	13	14	15	16	17
LFSR	-0.20	-1.22	-1.60	-2.19	-2.86	- 3.92	- 4.05
IID	0.66	-0.31	0.25	-1.36	-1.40	-2.64	-1.48

process with the following form:

$$\frac{dX}{X} = rdt + \sqrt{V}dW_1, \quad 0 < t < T, \tag{8.3.6}$$

$$dV = \kappa(\theta - V)dt + \sigma\sqrt{V}dW_2. \tag{8.3.7}$$

Marginally the volatility process V_t is following a CIR process which is always positive under the assumption $2\kappa\theta \geq \sigma^2$. In order to sample this continuous time process, we devide $[0, T]$ into $2^8 = 256$ equal size pieces and use the Euler Scheme for approximation.

The parameter we are using from Zhu [47] are as the following: $T = 6$(years), $r = 0.04$, $\theta = 0.04$, $\kappa = 2$ and $\sigma = 0.3$. The initial conditions were $X_0 = 100$ and $V_0 = 0.025$. The processes W_1 and W_2 driving the stock price and volatility are correlated Brownian motions with $dW_1 dW_2 = -0.5dt$. We want to price a European call option with strike price 100. That is, the option is at the money.

The results are shown in Table 8.9 and Figure 8.6. For this problem the dimension is very high - essentially it is a $2^8 \times 2 = 512$ dimension integration problem. Therefore for the Mini-LFSR generators we used we don't observe a significant higher convegence rate than IID sequence. Nevertheless, the MCQMC still reduce the root mean squred error by a factor of 2 to 4, which could be useful and important for finance practitioners.

Figure 8.5: Numerical results for Garch(1,1) pricing model. LFSR = solid and IID = dashed. The goal is to estimate the price of European Option. The initial price is marked at the right. The strike price K = 1.

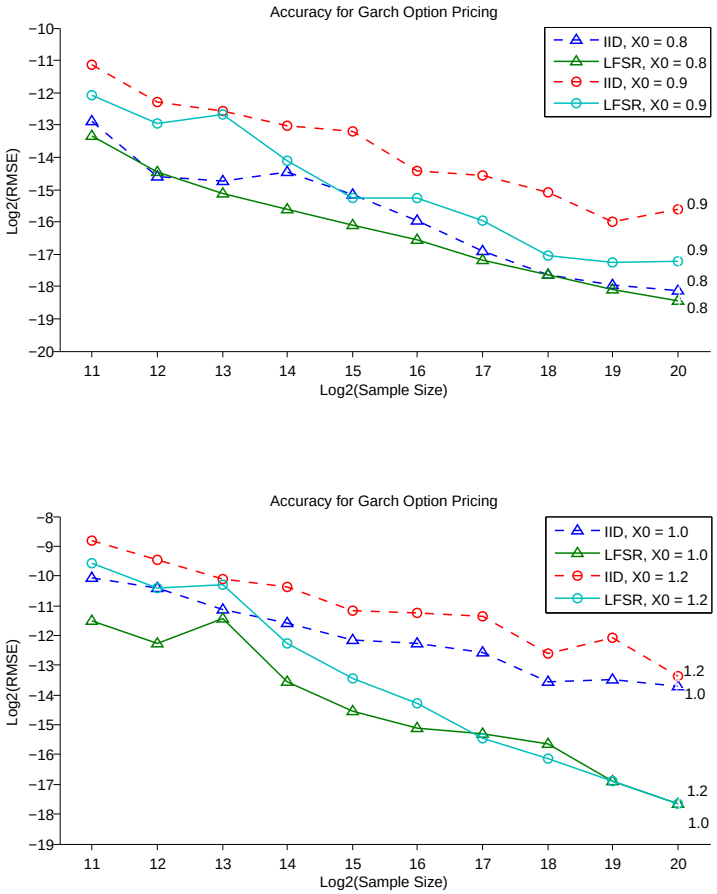
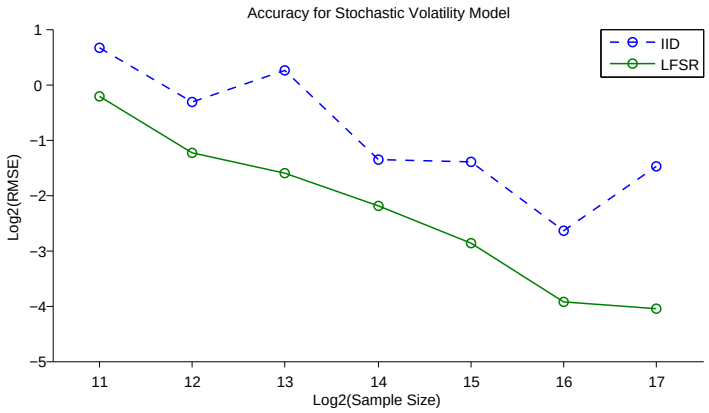


Figure 8.6: Numerical results for Stochastic Volatility Model. LFSR = solid and IID = dashed. The goal is to estimate the price of European Option. The initial price = 100 = Strike price.



Chapter 9

Conclusion and Future Directions

In this paper we have domonstrated that MCQMC algorithms formed by replacing IID driving sequence in a Markov Chain update by a completely uniformly distributed sequence can give a consistent result under certain assumptions. We require some regularity conditions, which hold for most of the Markov Chain update functions. If the state space is discrete, the consistency holds as long as the Markov Chain is recurrent. For continuous state space case, we require the Markov Chain update function to forget about the past - not only probabilistically but also in a path by path sense. On the other hand, we have given several examples showing the necessity of some of the conditions. The most surprising one might be example 4.4.2, in which case the Markov Chain is IID sampling from the unit circle with smooth update function ϕ, while we can crush any MCQMC algorithm by twisting a negligible subsequence of the driving CUD sequence $(v_i)_{i\geq 1}$. Also, we have shown that under certain conditions, MCQMC could have a convergence rate of $O\left(\frac{1}{n^{1-\delta}}\right)$, for any $\delta > 0$. We proved that simulating an ARMA process with bounded innovations can benefit from using MCQMC and achieve the optimal convergence rate $O\left(\frac{1}{n^{1-\delta}}\right)$.

In Chapter 8 we described in detail how MCQMC is implemented in practice, including the construction of a Mini-Linear Feedback Shift Register and digital Cranley-Patterson Randomization. The numerical experiments we did demonstrates the outperformance of MCQMC in various situations. For the 2 dimensional Gibbs Sampler

example, since it fits in the ARMA process framework, as we expected we observe a much better convergence by using MCQMC. For the other examples, the improvement on convergence rate is not as significant, but we still see variance reductions from 4 folds to thousands of folds. We do not expect that a substitution of more balanced sequence will always bring a large improvement, but we do expect that the expectation of a smooth function under stationary distribution can be more accurately estimated when the Markov Chain update function forgets about the past fast enough.

There are still a lot of unsolved problems in this area, some of which will be listed in the following section.

9.1 Future Directions

9.1.1 Bias-Variance Tradeoff

As we have seen, in some cases MCQMC is not consistent (Example 4.4.2 is a good counterexample). An interesting question will be, whether suitable randomization can make the MCQMC consistent almost surely. Of course if we completely randomize the CUD sequence by raplacing it with an IID sequence, we will get the consistency almost surely. But in this case we are giving up the better uniformity from the CUD sequence. There are a lot of different ways of randomization, some of them are adding "more" randomness but losing more uniformity while the others are adding "less" randomness and preserving more uniformity. There is a tradeoff between them and it will be interesting to ask how to balance.

To be more specific, assume we have a MCQMC algorithm starting at $\boldsymbol{x}_0 \in \Omega$ and $\boldsymbol{x}_i = \phi(\boldsymbol{x}_{i-1}, \boldsymbol{u}_i)$ where $\boldsymbol{u}_i$ is constructed from a CUD sequence. We estimate $\theta = \mathbf{E}_\pi f(\boldsymbol{x})$ by $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{x}_i)$.The mean squared error of the estimator $\hat{\theta}$ can be decomposed as:

$$MSE(\hat{\theta}) = Var(\hat{\theta}) + bias^2(\hat{\theta}) \tag{9.1.1}$$

If $(\boldsymbol{u}_i)_{i\geq 1}$ is totally deterministic, $\hat{\theta}$ will be also deterministic, therefore $Var(\hat{\theta}) =$

0. However, $bias(\hat{\theta}) \neq 0$. If we randomize $\boldsymbol{u}_i$, for instance, by Cranley-Patterson Rotation or a permutation, we will for sure increase the variance, but we hope the bias will at the same time be reduced.

9.1.2 Coupling Chains from Different Initial Values

In our proofs for the consistency of MCQMC, a lot of times we need to prove that two chains from different starting points will couple together or at least be close after sufficient long time. It can be described as follows. Let $X_0 \neq Y_0 \in \Omega$ are starting points for two Markov Chains X_n and Y_n. Assume they are updated through the same sequence $(\boldsymbol{u}_i)_{i\geq 1}$. I.e, $X_i = \phi(X_{i-1}, \boldsymbol{u}_i)$ and $Y_i = \phi(Y_{i-1}, \boldsymbol{u}_i)$. Then $Z_n = (X_n, Y_n), n \geq 0$ forms a Markov Chain defined on $\Omega \times \Omega$, assuming $\boldsymbol{u}_i$ IID $\sim \mathcal{U}[0,1)^d$. Properties of the Markov Chain (Z_n) can give us information about the dependence of X_n on its initial state X_0, which is critical for MCQMC to be consistent. Global Contracting mapping, Contracting on Average and Coupling region all guarantee that Z_n will converge to the diagonal line of $\Omega \times \Omega$ almost surely.

Bibliography

[1] P. Billingsley. *Convergence of probability measures*. John Wiley & Sons, Inc., New York, 1999.

[2] P.J. Brockwell and R.A. Davis. *Time Series: Theory and Methods*. Springer Verlag, New York, 2nd edition, 1996.

[3] S. Chen, J. Dick, and A.B.Owen. Consistency of Markov chain quasi-Monte Carlo on continuous state spaces. *Annals of Statistics*, 39(2):673–701, 2011.

[4] S. Chen, M. Matsumoto, T. Nishimura, and A. B. Owen. New inputs and methods for Markov chain quasi-Monte Carlo. In H. Wozniakowski and L. Plaskota, editors, *Monte Carlo and Quasi-Monte Carlo Methods 2010*. Springer-Verlag, 2011.

[5] N. N. Chentsov. Pseudorandom numbers for modelling Markov chains. *Computational Mathematics and Mathematical Physics*, 7:218–2332, 1967.

[6] L. Devroye. *Non-uniform Random Variate Generation*. Springer, 1986.

[7] L. Devroye and O. Fawzi. Simulating the Dickman distribution. *Statistics & Probability Letters*, 80(3-4), 2010.

[8] J. Dick. On quasi-Monte Carlo rules achieving higher order convergence. In P. L'Ecuyer and A. B. Owen, editors, *Monte Carlo and quasi-Monte Carlo Methods 2008*, 2009.

[9] K. Dickman. On the frequency of numbers containing prime factors of a certain relative magnitude. 1930.

[10] J.C. Duan. The Garch option pricing model. *Mathematical Finance*, 5(1), 1995.

[11] R. Durrett. *Probability: Theory and Examples*. Duxbury, California, 3rd edition, 2004.

[12] P. W. Glynn and D. Ormoneit. Hoeffding's inequality for uniformly recurrent Markov chains. *Statistics and Probability Letters*, 56(2):943–146, 2002.

[13] M. Gnewuch, A. Srivastav, and C. Winzen. Finding optimal volume subintervals with k points and computing the star discrepancy are NP-hard. *Journal of Complexity*, 24:154–172, 2008.

[14] R. D. Gordon. Value of Mill's ratio of area to bounding ordinate and of the normal probability integral for large values of the argument. *The Annals of Mathematical Statistics*, 18:364–366, 1941.

[15] W.K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[16] G.L. Jones. On the Markov chain central limit theorem. *Probability Surveys*, 1:299–320, 2004.

[17] D. E. Knuth. *The Art of Computer Programming*, volume 2: Seminumerical algorithms. Addison-Wesley, Reading MA, Third edition, 1998.

[18] L. Kuipers and H. Niederreiter. *Uniform distribution of sequences*. Wiley, New York, 1st edition, 1974.

[19] F.Y. Kuo. Component-by-component constructions achieve the optimal rate of convergence for multivariate integration in weighted Korobov and Sobolev spaces. *Journal of Complexity*, 19(3):301–320, 2003.

[20] H. L. Lebesgue. *Intégrale, longueur, aire*. PhD thesis, Université de Paris, 1902.

[21] P. L'Ecuyer, C. Lecot, and B. Tuffin. A randomized quasi-Monte Carlo simulation method for Markov chains. *Operations Research*, 56(4):958–975, 2008.

[22] P. L'Ecuyer and C. Lemieux. Lattice rules for the simulation of ruin problems. In P. A. Farrington, H. B. Nembhard, D. T. Sturrock, and G. W. Evans, editors, *Proceedings of the 1999 Winter Simulation Conference*. IEEE Press, 1999.

[23] P. L'Ecuyer and C. Lemieux. Quasi-Monte Carlo via linear shift-register sequences. In P. A. Farrington, H. B. Nembhard, D. T. Sturrock, and G. W. Evans, editors, *Proceedings of the 1999 Winter Simulation Conference*, pages 632–639. IEEE Press, 1999.

[24] M.B. Levin. Discrepancy estimates of completely uniformly distributed and pseudorandom number sequences. *International Mathematics Research Notice*, (22), 1999.

[25] L. G. Liao. Variance reduction in Gibbs sampler using quasi random numbers. *Journal of Computational and Graphical Statistics*, 7:253–266, 1998.

[26] J. S. Liu. *Monte Carlo strategies in scientific computing*. Springer, New York, 2001.

[27] N. Metropolis, A.W. Rossenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *Journal of chemical Physics*, 21(6):1087 – 1092, 1953.

[28] S.P. Meyn and R.L. Tweedie. *Markov Chains and Stochastic Stability*. Springer-Verlag, London, 2nd edition, 2005.

[29] R. M. Neal. Slice sampling. *Annals of Statistics*, 31(3):705–767, 2003.

[30] R. Nelson. *Probability, Stochastic Processes, and Queueing Theory: The mathemtics of Computer Performance Modeling*. Springer, 1st edition, 1995.

[31] H. Niederreiter. Multidimensional integration using pseudo-random numbers. *Mathematical Programming Study*, 27:17–38, 1986.

[32] H. Niederreiter. *Random number generation and quasi-Monte Carlo Methods*. S.I.A.M., Philadelphia, PA, 1992.

[33] A. Owen. Necessity of low effective dimension. 2002.

[34] A. B. Owen. Randomly permuted (t, m, s)-nets and (t, s)-sequences. In H. Niederreiter and P. Jau-Shyong Shiue, editors, *Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing*, pages 299–317, New York, 1995. Springer-Verlag.

[35] A. B. Owen. Multidimensional variation for quasi-Monte Carlo. In Jianqing Fan and Gang Li, editors, *International Conference on Statistics in honour of Professor Kai-Tai Fang's 65th birthday*, 2005.

[36] A. B. Owen and S. D. Tribble. A quasi-Monte Carlo Metropolis algorithm. *Proceedings of the National Academy of Sciences*, 102(25):8844–8849, 2005.

[37] W.W. Peterson and E.J. Weldon. *Error-Correcting Codes*. MIT Press, 2nd edition, 1972.

[38] W. Philipp. Empirical distribution functions and uniform distribution mod 1. *Diophantine approximation and its applications*, 1973.

[39] C. P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer, New York, 2nd edition, 2004.

[40] G.O. Roberts and J.S. Rosenthal. Harris recurrence of Metropolis-within-Gibbs and trans-dimensional Markov chains. *The Annals of Applied Probability*, 16(4):2123–2139, 2006.

[41] M. Rosenblatt. Remarks on a multivariate transformation. *Annals of Mathematical Statistics*, 23(3):470–472, 1952.

[42] J.S. Rosenthal. *Rates of Convergence for Gibbs Sampler and other Markov chains. Doctoral Dissertation*. PhD thesis, Harvard University, 1992.

[43] I.H. Sloan, F.Y. Kuo, and S. Joe. Constructing randomly shifted lattice rules in weighted Sobolev spaces. *SIAM Journal on Numerical Analysis*, 40(5):1650–1665, 2002.

[44] I. M. Sobol'. Pseudo-random numbers for constructing discrete Markov chains by the Monte Carlo method. *USSR Computational Mathematics and Mathematical Physics*, 14(1):36–45, 1974.

[45] S. D. Tribble. *Markov chain Monte Carlo algorithms using completely uniformly distributed driving sequences*. PhD thesis, Stanford University, 2007.

[46] X. Wang. A constructive approach to strong tractability using quasi-Monte Carlo. *Journal of Complexity*, 18:683–701, 2002.

[47] J. Zhu. A simple and exact simulation approach to Heston model. *Tech. rep*, 2008.