Advances in Monte Carlo methods:

exponentially tilted sequential proposal distributions and regenerative Markov chain samplers

Yi-Lung Chen

A thesis in fulfilment of the requirements for the degree of

Doctor of Philosophy

School of Mathematics & Statistics

Faculty of Engineering

The University of New South Wales

October 2021

THE UNIVERSITY OF NEW SOUTH WALES
Thesis/Dissertation Sheet

Surname or Family name: **Chen**

First name: **Yi-Lung**

Abbreviation for degree as given in the University calendar: **PhD**

School: **School of Mathematics & Statistics** Faculty: **Faculty of Science**

Title: Advances in Monte Carlo methods:
exponentially tilted sequential proposal distributions and regenerative Markov chain samplers

Abstract

Inference for Bayesian models often require one to simulate from some non-standard multivariate probability distributions. In the first part of the thesis, we successfully simulate exactly from certain Bayesian posteriors (the Tobit, the constrained linear regression, smoothing spline, and the Lasso) by applying rejection sampling using exponentially tilted sequential proposal distributions. This technique is typically efficient for posteriors which have the form of truncated multivariate normal/student. In this manner, we are able to simulate exactly from the posterior in hundreds of dimensions, which has until now being unattainable.

Due to the curse of dimensionality, these rejection schemes are unfortunately bound to fail as the dimensions of the problems grow. In such cases, one ultimately has to resort to approximate MCMC schemes. It is known that the sampling error of a Markov chain can be a lot easier if we can identify the regeneration times for the Markov chain. In particular, the convergence rate of a geometrically ergodic Markov chain can be estimated if one can identify the underlying regeneration events. While the idea of using regeneration in the error analysis of MCMC is not new, our contribution in the second part of the thesis is to provide simpler estimates of the total variation error, and a new graphical diagnostic with strong theoretical justification.

Finally, in the third part of the thesis, we consider the exponentially tilted sequential distributions in part one as proposal distributions for the MCMC samplers in part two. We introduce a novel Reject-Regenerate sampler, which combines the lessons learned about exact sampling and regenerative MCMC into a single framework. The resulting MCMC algorithm is a Markov chain with clearly demarcated regeneration events. Moreover, in the event of a regeneration, the Markov chain achieves a perfect draw with some probability.

Originality Statement

I hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by another person, or substantial proportions of material which have been accepted for the award of any other degree or diploma at UNSW or any other educational institution, except where due acknowledgement is made in the thesis. Any contribution made to the research by others, with whom I have worked at UNSW or elsewhere, is explicitly acknowledged in the thesis. I also declare that the intellectual content of this thesis is the product of my own work, except to the extent that assistance from others in the project's design and conception or in style, presentation and linguistic expression is acknowledged.

Yi-Lung Chen
02 November, 2021

Copyright Statement

Yi-Lung Chen
02 November, 2021

Authenticity Statement

I certify that the Library deposit digital copy is a direct equivalent of the final officially approved version of my thesis.

Yi-Lung Chen
02 November, 2021

INCLUSION OF PUBLICATIONS STATEMENT

UNSW is supportive of candidates publishing their research results during their candidature as detailed in the UNSW Thesis Examination Procedure.

Publications can be used in their thesis in lieu of a Chapter if:

- The candidate contributed greater than 50% of the content in the publication and is the "primary author", ie. the candidate was responsible primarily for the planning, execution and preparation of the work for publication
- The candidate has approval to include the publication in their thesis in lieu of a Chapter from their supervisor and Postgraduate Coordinator.
- The publication is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in the thesis

Please indicate whether this thesis contains published material or not:

☐ This thesis contains no publications, either published or submitted for publication
(if this box is checked, you may delete all the material on page 2)

☑ Some of the work described in this thesis has been published and it has been documented in the relevant Chapters with acknowledgement
(if this box is checked, you may delete all the material on page 2)

☐ This thesis has publications (either published or submitted for publication) incorporated into it in lieu of a chapter and the details are presented below

CANDIDATE'S DECLARATION

I declare that:

- I have complied with the UNSW Thesis Examination Procedure
- where I have used a publication in lieu of a Chapter, the listed publication(s) below meet(s) the requirements to be included in the thesis.

Candidate's Name	Signature	Date (dd/mm/yy)
Yi-Lung Chen		02/11/2021

POSTGRADUATE COORDINATOR'S DECLARATION *To only be filled in where publications are used in lieu of Chapters*

I declare that:

- the information below is accurate
- where listed publication(s) have been used in lieu of Chapter(s), their use complies with the UNSW Thesis Examination Procedure
- the minimum requirements for the format of the thesis have been met.

PGC's Name	PGC's Signature	Date (dd/mm/yy)

For each publication incorporated into the thesis in lieu of a Chapter, provide all of the requested details and signatures required

Details of publication #1:
Full title:
Authors:
Journal or book name:
Volume/page numbers:
Date accepted/ published:

Status	*Published*		*Accepted and In press*		*In progress (submitted)*	

The Candidate's Contribution to the Work
Insert text describing how the candidate has contributed to the work

Location of the work in the thesis and/or how the work is incorporated in the thesis:
Insert text

PRIMARY SUPERVISOR'S DECLARATION

I declare that:

- the information above is accurate
- this has been discussed with the PGC and it is agreed that this publication can be included in this thesis in lieu of a Chapter
- All of the co-authors of the publication have reviewed the above information and have agreed to its veracity by signing a 'Co-Author Authorisation' form.

Primary Supervisor's name	Primary Supervisor's signature	Date (dd/mm/yy)

Add additional boxes if required

Publications and Presentations

List of Publications

- Botev, Z., Chen, Y.L., L'Ecuyer, P., MacNamara, S. and Kroese, D.P., 2018, December. Exact posterior simulation from the linear LASSO regression. In 2018 Winter Simulation Conference (WSC) (pp. 1706-1717). IEEE.

List of Presentations

Poster presentations:

- Chen, Y.L. and Botev, Z.I., 2018, December. MCMC convergence diagnostics via regeneration with application to the bayesian lasso. In Proceedings of the 2018 Winter Simulation Conference (pp. 4224-4225).

Acknowledgments

Firstly, I would like to express my sincere gratitude to my supervisors Dr. Botev and Prof. Dick for the continuous support of my undergraduate study, honours year study, and my thesis, with their patience, motivation, and immense knowledge. I would also like to thank Dr. Botev again for providing me a project to work on during the summer break at the end of the second year of my degree. Indeed, approaching Dr. Botev at the end of the second year of my degree was surely the best decision I made in my university experience - it was not until then that I realized mathematical research was interesting and inspiring.

I would also like to thank my fellow mates, who cared to spend their precious time in discussing and proof-reading my thesis. Their generous constructive comments and criticisms were a lot helpful to me than they could have ever imagined. Finally, I would like to thank my family and friends for supporting me throughout my study. I could not imagine what it would be like without their support. I love them all, and I would like to express my gratitude to them in advance for their support in the future.

Abstract

Inference for Bayesian models often require one to simulate from some non-standard multivariate probability distributions. In the first part of the thesis, we successfully simulate exactly from certain Bayesian posteriors (the Tobit, the constrained linear regression, smoothing spline, and the Lasso) by applying rejection sampling using exponentially tilted sequential proposal distributions. This technique is typically efficient for posteriors which have the form of truncated multivariate normal/student. In this manner, we are able to simulate exactly from the posterior in hundreds of dimensions, which has until now being unattainable.

Due to the curse of dimensionality, these rejection schemes are unfortunately bound to fail as the dimensions of the problems grow. In such cases, one ultimately has to resort to approximate MCMC schemes. It is known that the sampling error of a Markov chain can be a lot easier if we can identify the regeneration times for the Markov chain. In particular, the convergence rate of a geometrically ergodic Markov chain can be estimated if one can identify the underlying regeneration events. While the idea of using regeneration in the error analysis of MCMC is not new, our contribution in the second part of the thesis is to provide simpler estimates of the total variation error, and a new graphical diagnostic with strong theoretical justification.

Finally, in the third part of the thesis, we consider the exponentially tilted sequential distributions in part one as proposal distributions for the MCMC samplers in part two. We introduce a novel Reject-Regenerate sampler, which combines the lessons learned about exact sampling and regenerative MCMC into a single framework. The resulting MCMC algorithm is a Markov chain with clearly demarcated regeneration events. Moreover, in the event of a regeneration, the Markov chain achieves a perfect draw with some probability.

Keywords: MCMC, regeneration, rejection sampling, total variation distance, independent sampler, perfect sampling, Bayesian Lasso, Bayesian constrained linear regression, Bayesian smoothing spline

Notation and Symbols

We make use of various typographical aids, and it will be beneficial for the reader to be aware of some of these.

- $\mathbb{E}$ denotes the expectation operator, Var as the variance operator
- $\mathbb{P}[\boldsymbol{X} \in A]$ is the probability of random vector $\boldsymbol{X}$ falling in set A
- $\mathbb{I}\{\boldsymbol{x} \in A\}$, and $\mathbb{I}_{\boldsymbol{x} \in A}$ denote the indicator function on the set A evaluated at $\boldsymbol{x}$
- $\mathsf{N}(\boldsymbol{\mu}, \Sigma)$ denotes the distribution of a Normal random vector (variable) centred at $\boldsymbol{\mu}$ and covariance matrix Σ
- $\mathsf{Exp}(\lambda)$ denotes the distribution of an exponential random variable with rate $\lambda > 0$
- $\mathsf{Unif}(a, b)$ denotes the distribution of a uniform random variable with rate with support on the interval (a, b)
- $\sim$ denotes 'obeys', for example $\boldsymbol{X} \sim \mathsf{N}(\boldsymbol{\mu}, \Sigma)$ means the random vector $\boldsymbol{X}$ has a distribution of $\mathsf{N}(\boldsymbol{\mu}, \Sigma)$
- $\dim(\boldsymbol{v})$ denotes the dimension of the vector $\boldsymbol{v}$
- $\mathrm{diag}(\boldsymbol{v})$ denotes the diagonal matrix with (i, i)-th entry being $\boldsymbol{v}_i$, the i-th component of the vector $\boldsymbol{v}$
- I denotes an identity matrix of appropriate dimension
- $\int f(\boldsymbol{x})\, d\boldsymbol{x}$ denotes the integral of the function f with respect to the Lebesgue measure
- $\int f(\boldsymbol{x})\, \pi(d\boldsymbol{x})$ denotes the integral of the function f with respect to the measure π
- We often omit brackets when it is clear what the argument is of a function or operator. For example, we prefer $\mathbb{E}X^2$ to $\mathbb{E}[X^2]$.
- We will occasionally use a Bayesian notation convention in which the *same* symbol is used to denote different (conditional) probability densities. In particular, instead of writing $f_X(x)$ and $f_{X \,|\, Y}(x \,|\, y)$ for the probability density function (pdf) of X and the conditional pdf of X given Y, we simply write $f(x)$ and $f(x \,|\, y)$. This particular style of notation can be of great descriptive value, despite its apparent ambiguity.

Publication included in this thesis

Part of the content in Chapter 3 is based on the published work

Botev, Z., Chen, Y.L., L'Ecuyer, P., MacNamara, S. and Kroese, D.P., 2018, December. Exact posterior simulation from the linear LASSO regression. In 2018 Winter Simulation Conference (WSC) (pp. 1706-1717). IEEE.

Contents

Chapter 1

Introduction

Let π be a probability measure defined on a measurable space $(\mathcal{X}, \mathscr{A})$, where $\mathcal{X} \subseteq \mathbb{R}^d$ and d can be large. Generating random[1] vectors $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$ whose probability laws coincide to π, that is $\mathbb{P}[\boldsymbol{X}_k \in A] = \pi(A)$ for all k and $A \in \mathscr{A}$, can be difficult for arbitrary π and $\mathcal{X}$. However, simulating draws from a (posterior) density (known up to some marginalizing constant) is a common theme in Bayesian inference.

Rejection sampling allows one to simulate iid random vectors $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots \sim \pi$ in the case where π exhibits a density (with respect to the Lebesgue measure). Formally, let $f = \tilde{f}/\ell$ denote the density of π, where ℓ is the normalizing constant for some positive function with finite integral $\tilde{f}$, and let g be a probability density for which $\tilde{f} \leq cg$ for some $c > 0$. If one retains a draw $\boldsymbol{X} \sim g$ only if $\tilde{f}(\boldsymbol{X})/cg(\boldsymbol{X}) \leq U$ for some independent $U \sim \mathsf{Unif}(0, 1)$, then the resulting random vector obeys π. (We refer readers to [70] and the reference therein for further details.)

In Chapter 2 we construct efficient rejection sampling proposal density for the posterior densities of the Bayesian constrained linear regression [17, 37], the Bayesian Tobit model [20], and the Bayesian smoothing spline [89]. Up until now, the standard approaches only simulate draws that approximately obey these posterior densities, however we show that it is possible to obtain exact draws from these posterior densities. The key insight is that these posterior densities, after some coordinate transformations, take the form of the proposal density studied in [12]. The original work there provides accurate importance sampling estimator for multivariate student probabilities on sets described by a linear system of inequalities. The technique is to constructs an exponentially tilted sequential proposal density. It turns out that this density is also an efficient rejection sampling proposal density that targets truncated multivariate student distribution even up to hundreds of dimensions. (Note that multivariate normal with linear constraint is a special case of the multivariate student studied in [11].)

The idea of sequential exponential tilting presented in Chapter 2 and applied to the truncated multivariate normal/student can also be extended to some other non-standard

[1] Algorithms that run on standard PC actually generate 'pseudo random' vectors.

multivariate distributions. We show that an example of such non-standard multivariate distributions amenable to our rejection sampling methods is the Bayesian Lasso posterior studied in Chapter 3.

The Lasso linear regression [107] and its Bayesian analogue [91] have appealed practitioners. However, the standard samplers in the current again only simulate approximate draws. In this chapter we construct a novel exact sampling algorithm for the Bayesian Lasso posterior which is presented in [10]. The idea is to construct an efficient proposal distribution for rejection sampler, again via an exponentially tilted sequential distribution.

The obvious advantage of our novel rejection samplers is that standard iid analyses hold for any simulated draws, making error analysis straightforward. Unfortunately, designing an efficient proposal density is not a routine task – it requires careful analysis for each problem. Moreover, as d increases, due to the curse of dimensionality, any proposal density is destined to lose its efficiency. For this reason we also study approximate sampling by MCMC (Markov chain Monte Carlo). Algorithms in this framework generate a Markov chain $\{\boldsymbol{X}_k^*, k \geq 1\}$ on $(\mathcal{X}, \mathscr{A})$ such that the limiting probability law, $\lim_{k\to\infty} \mathbb{P}[\boldsymbol{X}_k^* \in \cdot]$, coincides to the targeted π in some metric. The idea is that the sample path of the Markov chain itself corresponds to correlated draws that approximately obey π. Moreover, for any h from a large class of functions, the ergodic average, $\hat{q}_t := \frac{1}{t}\sum_{k=1}^{t} h(\boldsymbol{X}_k^*)$ converges to $q := \int_{\mathcal{X}} h(\boldsymbol{x})\pi(d\boldsymbol{x})$ as $t \to \infty$. Further along with some regularity conditions, there is also a Markov chain CLT that is, $\sqrt{t}(\hat{q}_t - q) \to \mathsf{N}(0, \sigma^2)$, for some σ^2. Common MCMC algorithms include Metropolis-Hastings samplers, independent samplers, Gibbs sampler, and hit-and-run sampler [82, 53, 39].

Despite being computationally attractive, evaluating the performance of a MCMC sampler is a difficult task due to the dependence between the MCMC draws. Systematically evaluating the performance of a given MCMC method naturally resorts to answering the following questions. For $t < \infty$, how well does $\hat{q}_t$ approximate q on average? How variable is it? How close is $\mathbb{P}[\boldsymbol{X}_k^* \in \cdot]$ to π for different values of k? How much closer does it become for each extra step? This motivates our study in Chapters 5 and 6. (Some preliminary results and notions are recounted in Chapter 4.)

Moving away from the ambitious rejecting sampling, in Chapter 5 we explore how one can evaluate the performance of a given MCMC by identifying its underlying regenerative times. Roughly speaking, these are instances where the Markov chain 'stochastically restarts, thereby allowing one to segment the chain into iid cycles [87, 86]. It is well established that if one can identify the regeneration times of a Markov chain sampler, then the variability of ergodic estimators can be quantified in an easier manner.

Our contributions in regenerative MCMC are as follows. In Chapter 5 we derive a novel total variation distance bounds between a geometric Markov chain and its limiting

distribution. The constants for these bounds can be easily estimated by identifying its underlying regenerative structure, thus we propose estimators for these constants using the output of the MCMC samplers. Such bounds are useful for assessing the convergence rate of a Markov chain and for estimating the size of a burn-in period [64].

Another contribution is in Chapter 5, where we show that the convergence of a regenerative MCMC can be summarized by an underlying one dimensional process, known as the *elapsed time process*. To this end, we propose a univariate diagnostic plot that assesses the global mixing. Consequently, practitioners no longer need to rely on arbitrary projection of high dimensional processes for the construction of diagnostic plots.

An interesting insight is that a carefully designed rejection sampling proposal naturally leads to a Metropolis-independence MCMC sampler with frequent regeneration. In particular, we consider how the sequential proposal densities studied in Chapters 2 and 3 perform when they are the proposal densities Metropolis-independence samplers. In applying our novel MCMC diagnostics on these independent samplers, we observe that these independent samplers converge quickly (they exhibit fast mixing). Consequently, we demonstrate the value of these carefully designed sequential proposals outside the framework of rejection sampling.

Finally, in Chapter 6 we describe our novel *Reject-Regenerate* Markov chain sampler. In this framework the simulated Markov chain is regenerative. Moreover, in the event of a regeneration, the Markov chain has some probability of achieving a perfect draw. The Reject-Regenerate algorithm can be thought of as a hybrid algorithm combining the best aspects of exact rejection sampling and approximate MCMC sampling.

Chapter 2

Exponentially tilted sequential proposal for the truncated multivariate student distribution

2.1 Introduction to this chapter

A random vector $\boldsymbol{Y} \in \mathbb{R}^d$, whose density (with respect to the Lebesgue measure) exists, is said to obey the multivariate student distribution, centered at $\boldsymbol{\mu} \in \mathbb{R}^d$ with covariance matrix Σ, denoted by $\boldsymbol{Y} \sim \mathsf{t}_\nu(\boldsymbol{\mu}, \Sigma)$, if its density[1] function (with respect to the Lebesgue measure) is

$$\frac{\Gamma((d+\nu)/2)}{|\Sigma|^{1/2}(\pi\nu)^{d/2}\Gamma(\nu/2)} \left(1 + \frac{1}{\nu}(\boldsymbol{y}-\boldsymbol{\mu})^\top \Sigma^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right)^{-(\nu+d)/2}, \quad \boldsymbol{y} \in \mathbb{R}^d.$$

For a given $m \times d$ real matrix C and vectors $\boldsymbol{l}, \boldsymbol{u} \in \overline{\mathbb{R}}^m$, denoting $\ell = \mathbb{P}[\boldsymbol{l} \leq \mathrm{C}\boldsymbol{Y} \leq \boldsymbol{u}]$, it follows that the density of $\boldsymbol{Y}$ conditioned on $\boldsymbol{Y} \in \{\boldsymbol{y} \,|\, \boldsymbol{l} \leq \mathrm{C}\boldsymbol{y} \leq \boldsymbol{u}\}$ is

$$h(\boldsymbol{y}) = \frac{\frac{\Gamma((d+\nu)/2)}{|\Sigma|^{1/2}(\pi\nu)^{d/2}\Gamma(\nu/2)} \left(1 + \frac{1}{\nu}(\boldsymbol{y}-\boldsymbol{\mu})^\top \Sigma^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right)^{-(\nu+d)/2} \mathbb{I}\{\boldsymbol{l} \leq \mathrm{C}\boldsymbol{y} \leq \boldsymbol{u}\}}{\ell}. \tag{2.1}$$

Here we take the convention $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, \infty\}$ to allow for certain components of $\mathrm{C}\boldsymbol{y}$ to have no lower or upper restrictions.

Estimating ℓ and simulating draws $\boldsymbol{Y} \sim h$ are two closely related problems. Both problems can be notoriously difficult for general C, $\boldsymbol{l}$ and $\boldsymbol{u}$, however they have many statistical applications (see [58, 42, 44] and the references therein).

In this chapter we consider the exponentially tilted sequential proposal density for the truncated multivariate student distribution derived in [12]. This proposal density gives an accurate importance sampling estimator for ℓ and an efficient rejection sampler to simulate draws from h.

Our contributions over and above the existing theory in [12] are as follows. In this chapter we shall prove a theoretical results concerning the asymptotic efficiency of this importance sampler. A by-product of this proof is a multivariate extension to the Mill's

[1] In this thesis, we only consider Σ is invertible so that the density exists.

ratio of the univariate student density [69, 84]. We then find applications for this rejection sampler in simulating draws from the posterior densities of the Bayesian constrained linear regression, Bayesian Tobit model, and the Bayesian smoothing spline. In other words, we construct exact sampling schemes for these Bayesian posterior densities whose default samplers at the current literature are approximate MCMC samplers.

Before we present a brief recount of notable related works, we note that it suffices to consider the case where $\boldsymbol{\mu} = \mathbf{0}$. This is because if $\boldsymbol{Y} \sim \mathsf{t}_\nu(\boldsymbol{\mu}, \Sigma)$ and $\boldsymbol{Y}' \sim \mathsf{t}_\nu(\mathbf{0}, \Sigma)$ then $\boldsymbol{Y} = \boldsymbol{Y}' + \boldsymbol{\mu}$, so that

$$\ell = \mathbb{P}[\boldsymbol{l} \leq \mathrm{C}\boldsymbol{Y} \leq \boldsymbol{u}] = \mathbb{P}[\boldsymbol{l}' \leq \mathrm{C}\boldsymbol{Y}' \leq \boldsymbol{u}']$$

where $\boldsymbol{l}' = \boldsymbol{l} - \mathrm{C}\boldsymbol{\mu}$ and $\boldsymbol{u}' = \boldsymbol{u} - \mathrm{C}\boldsymbol{\mu}$. Moreover, if $\boldsymbol{Y}'$ obeys $\mathsf{t}_\nu(0, \Sigma)$ restricted to the set $\{\boldsymbol{y}' \,|\, \boldsymbol{l}' \leq \mathrm{C}\boldsymbol{y}' \leq \boldsymbol{u}'\}$, then $\boldsymbol{Y} := \boldsymbol{Y}' + \boldsymbol{\mu}$ obeys $\mathsf{t}_\nu(\boldsymbol{\mu}, \Sigma)$ restricted to the set $\{\boldsymbol{y} \,|\, \boldsymbol{l} \leq \mathrm{C}\boldsymbol{y} \leq \boldsymbol{u}\}$. We henceforth refer to ℓ and h for the case $\boldsymbol{\mu} = \mathbf{0}$, for some matrices Σ, C and vectors $\boldsymbol{l}, \boldsymbol{u}$ of appropriate dimensions. That is

$$\ell = \mathbb{P}[\boldsymbol{l} \leq \mathrm{C}\boldsymbol{Y} \leq \boldsymbol{u}] = \int_{\boldsymbol{y}: \boldsymbol{l} \leq \mathrm{C}\boldsymbol{y} \leq \boldsymbol{u}} \frac{\Gamma((d+\nu)/2)}{|\Sigma|^{1/2}(\pi\nu)^{d/2}\Gamma(\nu/2)} \left(1 + \frac{1}{\nu}\boldsymbol{y}^\top \Sigma^{-1} \boldsymbol{y}\right)^{-(\nu+d)/2} d\boldsymbol{y},$$

and

$$h(\boldsymbol{y}) = \frac{\frac{\Gamma((d+\nu)/2)}{|\Sigma|^{1/2}(\pi\nu)^{d/2}\Gamma(\nu/2)} \left(1 + \frac{1}{\nu}\boldsymbol{y}^\top \Sigma^{-1} \boldsymbol{y}\right)^{-(\nu+d)/2} \mathbb{I}\{\boldsymbol{l} \leq \mathrm{C}\boldsymbol{y} \leq \boldsymbol{u}\}}{\ell}.$$

The exponentially tilted sequential proposal density derived in [12] is motivated by the 'separation of variable method', which concerns estimating ℓ by Monte Carlo, proposed in [42]. This method concerns the case where C is the $d \times d$ identity matrix so that ℓ is the probability of a $\mathsf{t}_\nu(\mathbf{0}, \Sigma)$ random vector falling in some (possibly unbounded) hyper rectangle $[\boldsymbol{l}, \boldsymbol{u}]$. The method begins by computing the Cholesky decomposition of Σ so that ℓ can be equivalently expressed as the probability of a $\mathsf{t}_\nu(\mathbf{0}, \mathrm{I}_d)$ random vector falling in a set described by a lower triangular system of linear inequalities.

Formally, let $\Sigma = \mathrm{L}\mathrm{L}^\top$ be the Cholesky decomposition of Σ. It follows that $\boldsymbol{Y} = \mathrm{L}\boldsymbol{X}$ in distribution, where $\boldsymbol{X} \sim \mathsf{t}_\nu(\mathbf{0}, \mathrm{I}_d)$, and

$$\ell = \mathbb{P}[\boldsymbol{l} \leq \boldsymbol{Y} \leq \boldsymbol{u}] = \mathbb{P}[\boldsymbol{l} \leq \mathrm{L}\boldsymbol{X} \leq \boldsymbol{u}].$$

That is,

$$\ell = \mathbb{P}[\boldsymbol{l} \leq \mathrm{L}\boldsymbol{X} \leq \boldsymbol{u}] = \int_{\boldsymbol{x}: \boldsymbol{l} \leq \mathrm{L}\boldsymbol{x} \leq \boldsymbol{u}} \frac{\Gamma((d+\nu)/2)}{(\pi\nu)^{d/2}\Gamma(\nu/2)} \left(1 + \frac{\boldsymbol{x}^\top \boldsymbol{x}}{\nu}\right)^{-(\nu+d)/2} d\boldsymbol{x}. \tag{2.2}$$

The lower triangular structure of L in (2.2) means that

$$\begin{aligned} l_1' := l_1/L_{11} \leq & x_1 \leq u_1/L_{11} := u_1' \\ l_2' := \frac{l_2 - L_{21}x_1}{L_{22}} \leq & x_2 \leq \frac{u_2 - L_{21}x_1}{L_{22}} := u_2' \\ & \vdots \\ l_d' := \frac{l_d - \sum_{i=1}^{d-1} L_{di}x_i}{L_{dd}} \leq & x_d \leq \frac{u_d - \sum_{i=1}^{d-1} L_{di}x_i}{L_{dd}} := u_d', \end{aligned}$$

so that the constraint on x_k only depends on x_j for $j < k$. Next, since

$$\left(1 + \frac{\boldsymbol{x}^\top \boldsymbol{x}}{\nu}\right) = \left(1 + \frac{x_1^2}{\nu}\right) \cdot \left(1 + \frac{x_2^2}{\nu + x_1^2}\right) \cdot \ldots \cdot \left(1 + \frac{x_d^2}{\nu + \sum_{j=1}^{d-1} x_j^2}\right),$$

applying the substitution $x_k = z_k \sqrt{\frac{\nu + x_1^2 + \ldots x_{k-1}^2}{\nu + k - 1}}$ for $k = d, d-1, \ldots, 2$ and $x_1 = z_1$, one after the other, we can write

$$\ell = \int_{[\hat{l}_1, \hat{u}_1]} t_\nu(z_1) \int_{[\hat{l}_2, \hat{u}_2]} t_{\nu+1}(z_2) \ldots \int_{[\hat{l}_d, \hat{u}_d]} t_{\nu+d-1}(z_d)\, d\boldsymbol{z}, \tag{2.3}$$

where $\hat{l}_k = l_k' \sqrt{\frac{\nu+k-1}{\nu + x_1^2 + \ldots x_{k-1}^2}}$, $\hat{u}_k = u_k' \sqrt{\frac{\nu+k-1}{\nu + x_1^2 + \ldots x_{k-1}^2}}$ for $k = 2, \ldots, 2$, $\hat{l}_1 = l_1'$, $\hat{u}_1 = u_1'$, and $t_{\nu+k-1}$ is the density of a univariate $\mathsf{t}_{\nu+k-1}(0,1)$ random variable. Readers should note that $\hat{l}_k$ and $\hat{u}_k$ still only implicitly depends on z_j for $j \leq k$ after these transformations.

These observations motivate a numerical scheme for estimating ℓ. This numerical scheme can be equivalently viewed as a sequential importance sampling estimation as follows. Define the density

$$g(\boldsymbol{z}) = g_1(z_1) g(z_2 \,|\, z_1) g_2(z_3 \,|\, z_1, z_2) \ldots g_d(z_d \,|\, z_1, z_2, \ldots, z_{d-1}),$$

where

$$g(\boldsymbol{z}_k \,|\, z_1, \ldots, z_{k-1}) = \frac{t_{\nu+k-1}(z_k)}{T_{\nu+k-1}(\hat{u}_k) - T_{\nu+k-1}(\hat{l}_k)} \mathbb{I}\{\hat{l}_k \leq z \leq \hat{u}_k\}, \quad k = 1, \ldots, d$$

is the density of a univariate $\mathsf{t}_{\nu+k-1}(0,1)$ random variable constrained to $[\hat{l}_k, \hat{u}_k]$ and that $T_{\nu+k-1}$ is the cdf of a univariate $\mathsf{t}_{\nu+k-1}(0,1)$ random variable. Denoting the integrand in (2.3) as

$$p(\boldsymbol{z}) = t_\nu(z_1) t_{\nu+1}(z_2) \ldots t_{\nu+d-1}(z_d),$$

and let $\boldsymbol{Z}_i \overset{\text{iid}}{\sim} g$ for $i = 1, \dots n$. It follows, by the virtue of importance sampling estimation,

$$\ell = \mathbb{E}\frac{p(\boldsymbol{Z}_1)}{g(\boldsymbol{Z}_1)} = \prod_{k=1}^{d}[T_{\nu+k-1}(\hat{l}(\boldsymbol{Z}_1)) - T_{\nu+k-1}(\hat{u}(\boldsymbol{Z}_1))].$$

Therefore, the estimator for ℓ is

$$\hat{\ell}_{\text{Genz}} = \frac{1}{n}\sum_{i=1}^{n}\prod_{k=1}^{d}[T_{\nu+k-1}(\hat{l}(\boldsymbol{Z}_i)) - T_{\nu+k-1}(\hat{u}(\boldsymbol{Z}_i))].$$

Remark. The original work in [42] introduces one more transformation so that the integral (2.3) *is on a* $[0,1]^{d-1}$ *hypercube. In this manner, instead of simulating univariate truncated students, one simulates random/quasi-random uniformly distributed points on* $[0,1]^{d-1}$ *to estimate the integral.*

Indeed, this technique, known as the 'separation of variable' has already been applied in earlier works on estimating multivariate normal probabilities over linear constraints [40, 112] and even for the case where the covariance matrix of the multivariate normal distribution is singular [43]. Other methods for computing ℓ are discussed in [41].

A problem closely related to computing ℓ is to compute $\mathbb{E}[\boldsymbol{Y}^k]$, where $\boldsymbol{Y} \sim h$ and the exponent k is taken coordinate-wise. Again, for the case where $\mathrm{C} = \mathrm{I}_d$, [55] derives analytic formula to compute $\mathbb{E}[\boldsymbol{Y}]$ and $\mathbb{E}[\boldsymbol{Y}\boldsymbol{Y}^\top]$. A recurrence relation for $\mathbb{E}[\boldsymbol{Y}^k]$ is also derived in [36].

The next related problem is simulating from h. The standard approach when d is large is Gibbs sampling which gives approximate draws from h. These Gibbs samplers leverage the fact that if $\boldsymbol{X} \sim \mathsf{N}(0, \Sigma)$, and independently $R \sim \mathsf{chi}_\nu$, then $Y = \sqrt{\nu}\boldsymbol{X}/R \sim \mathsf{t}_\nu(\boldsymbol{0}, \Sigma)$ where the density of a chi_ν random variable is

$$\frac{1}{2^{\nu/2-1}\Gamma(\nu/2)}\exp\left(-\frac{r^2}{2} + (\nu-1)\ln r\right), \quad r > 0.$$

In this manner, it suffices for one to consider simulating $(\boldsymbol{X}, R) \sim f$ where

$$f(\boldsymbol{x}, r) = \frac{\frac{1}{\sqrt{|\Sigma|}(2\pi)^{d/2}\times 2^{\nu/2-1}\Gamma(\nu/2)}\exp\left(-\frac{1}{2}\boldsymbol{x}^\top\Sigma^{-1}\boldsymbol{x} - \frac{r^2}{2} + (\nu-1)\ln r\right)\mathbb{I}\{r\boldsymbol{l} \le \sqrt{\nu}\mathrm{C}\boldsymbol{x} \le r\boldsymbol{u}\}}{\ell},$$

so that $\boldsymbol{Y} = \sqrt{\nu}\boldsymbol{X}/R \sim h$.

The Gibbs sampler proposed in [44] assumes that C is invertible and considers the transformation $\boldsymbol{z} = \mathrm{C}\boldsymbol{x}$. This yields the following density.

$$f_1(\boldsymbol{z}, r) = \frac{\frac{1}{\sqrt{|\mathrm{D}|}(2\pi)^{d/2}\times 2^{\nu/2-1}\Gamma(\nu/2)} \exp\left(-\frac{1}{2}\boldsymbol{z}^\top \mathrm{D}^{-1}\boldsymbol{z} - \frac{r^2}{2} + (\nu-1)\ln r\right) \mathbb{I}\{r\boldsymbol{l} \le \sqrt{\nu}\boldsymbol{z} \le r\boldsymbol{u}\}}{\ell},$$

where $\mathrm{D} = \mathrm{C}^\top \Sigma \mathrm{C}$. Here, the marginal density $f_1(r)$ is the density of a chi_ν random variable while the full conditional densities $f_1(z_k \,|\, r, z_1, \ldots, z_{k-1}, z_{k+1}, \ldots z_d)$ is the density of some univariate truncated normal random variable for $k = 1, \ldots, d$. The proposed Gibbs sampler then cycles between these densities to simulate $(\boldsymbol{Z}, R)$ which is approximately f_1 distributed.

An alternative Gibbs sampler proposed in [76] considers a different transformation on f. Again denoting the Cholesky decomposition $\Sigma = \mathrm{L}_1\mathrm{L}_1^\top$, the transformation $\boldsymbol{z} = \mathrm{L}_1^{-1}\boldsymbol{x}$ for f yields the following density.

$$f_2(\boldsymbol{z}, r) = \frac{\frac{1}{(2\pi)^{d/2}\times 2^{\nu/2-1}\Gamma(\nu/2)} \exp\left(-\frac{\|\boldsymbol{z}\|^2}{2} - \frac{r^2}{2} + (\nu-1)\ln r\right) \mathbb{I}\{r\boldsymbol{l} \le \sqrt{\nu}\mathrm{C}\mathrm{L}\boldsymbol{z} \le r\boldsymbol{u}\}}{\ell}.$$

Again, the marginal density $f_2(r)$ is the density of a chi_ν random variable while the full conditional densities $f_2(z_k \,|\, r, z_1, \ldots, z_{k-1}, z_{k+1}, \ldots z_d)$ is the density of some univariate truncated normal random variable for $k = 1, \ldots, d$. The proposed Gibbs sampler again cycles between these densities to simulate $(\boldsymbol{Z}, R)$ which is approximately f_2 distributed.

The Gibbs sampler in [44] has the advantage that the support of f_1 is a simple hyper rectangle $[\boldsymbol{l}, \boldsymbol{u}]$. This makes the implementation of algorithm simple. However this Gibbs sampler is restricted to the case where C is invertible. On the other hand, the support of f_2 can potentially be much more complicated, however its normal component does not have a complicated covariance structure unlike the matrix D in f_1. The authors of [76] argue that this renders a better mixing Gibbs sampler. Moreover, this Gibbs sampler does not require C to be invertible in the first place.

Finally, the exponentially tilted sequential proposal derived in [12] also considers a proposal density g that takes a sequential form. That is, (denoting $\boldsymbol{\theta} = (\boldsymbol{z}, r)$)

$$g(\boldsymbol{\theta}) = g_0(\theta_0) g_1(\theta_1 \,|\, \theta_0) g_2(\theta_2 \,|\, \theta_0, \theta_1) \ldots g_d(\theta_d \,|\, \theta_0, \ldots, \theta_{d-1}),$$

where each g_k belongs to a family of densities indexed by some 'tilting' parameter. An optimal tilting parameter is chosen such that the corresponding importance sampling estimator $\hat{\ell}$ has (approximately) minimal variance within that family. Moreover, it turns out this optimally chosen g may render an efficient exact sampling scheme for h.

The rest of the chapter is structured as follows. We first revisit the importance sampling estimation for ℓ and the exact simulation scheme for h proposed in [12]. We then establish a theoretical result concerning the asymptotic efficiency of this importance sampler. Finally we find applications for this rejection sampler in simulating draws from the posterior densities of the Bayesian constrained linear regression, Bayesian Tobit model, and the Bayesian smoothing spline. In other words, we construct efficient exact samplers for these posterior densities whose current default samplers in the literature are approximate MC samplers.

2.2 A sequential proposal density via optimal exponential tilting

We describe the methods proposed in [12] in this section. Again, let

$$f(\boldsymbol{x}, r) = \frac{\frac{1}{\sqrt{|\Sigma|}(2\pi)^{d/2}\times 2^{\nu/2-1}\Gamma(\nu/2)} \exp\left(-\frac{1}{2}\boldsymbol{x}^\top\Sigma^{-1}\boldsymbol{x} - \frac{r^2}{2} + (\nu-1)\ln r\right)\mathbb{I}\{r\boldsymbol{l} \le \sqrt{\nu}\mathrm{C}\boldsymbol{x} \le r\boldsymbol{u}\}}{\ell},$$

so that if $(\boldsymbol{X}, R) \sim f$ then $\boldsymbol{Y} = \sqrt{\nu}\boldsymbol{X}/R \sim h$.

Next, let $\Sigma = \mathrm{L}_1\mathrm{L}_1^\top$ be the Cholesky decomposition of Σ and $\mathrm{C}\mathrm{L}_1 = \mathrm{L}\mathrm{Q}$ be the LQ decomposition of $\mathrm{C}\mathrm{L}_1$ so that L_1, L are lower triangular while Q is orthogonal. It follows that the substitution $\boldsymbol{x} = \mathrm{L}_1\mathrm{Q}^\top\boldsymbol{z}$ yields the density

$$f(\boldsymbol{z}, r) = \frac{\frac{1}{\sqrt{|\Sigma|}(2\pi)^{d/2}\times 2^{\nu/2-1}\Gamma(\nu/2)} \exp\left(-\frac{\|\boldsymbol{z}\|_2^2}{2} - \frac{r^2}{2} + (\nu-1)\ln r\right)\mathbb{I}\{r\boldsymbol{l} \le \sqrt{\nu}\mathrm{L}\boldsymbol{z} \le r\boldsymbol{u}\}}{\ell}, \tag{2.4}$$

where $\|\cdot\|_p$ is the p-norm.

Exploiting the lower triangular structure of L and write $\mathscr{R} := \{(\boldsymbol{z}, r) \,|\, r\boldsymbol{l} \le \sqrt{\nu}\mathrm{L}\boldsymbol{z} \le r\boldsymbol{u}\}$ as

$$\begin{aligned}
\tilde{l}_1(r) := \frac{r\,l_1}{\sqrt{\nu}}/L_{11} \le z_1 &\le \frac{r\,u_1}{\sqrt{\nu}}/L_{11} := \tilde{u}_1(r)\\
\tilde{l}_2(r, z_1) := \frac{r\,l_2\nu^{-1/2} - L_{21}z_1}{L_{22}} \le z_2 &\le \frac{r\,u_2\nu^{-1/2} - L_{21}z_1}{L_{22}} := \tilde{u}_2(r, z_1)\\
&\vdots\\
\tilde{l}_d(r, z_1, \ldots, z_{d-1}) := \frac{\frac{r\,l_d}{\sqrt{\nu}} - \sum_{i=1}^{d-1} L_{di}z_i}{L_{dd}} \le z_d &\le \frac{\frac{r\,u_d}{\sqrt{\nu}} - \sum_{i=1}^{d-1} L_{di}z_i}{L_{dd}} := \tilde{u}_d(r, z_1, \ldots, z_{d-1}).
\end{aligned}$$

Observing that $\tilde{l}_k, \tilde{u}_k$ only depends on $r, z_1, z_2, \ldots, z_{k-1}$ for all k, it is therefore natural to construct proposal density g sequentially in the sense that

$$g(\boldsymbol{z}, r) = g_0(r)g_1(z_1 \,|\, r)g_2(z_2 \,|\, r, z_1)\ldots g_d(z_d \,|\, r, z_1, \ldots, z_{d-1}).$$

Denoting $\phi(\cdot;\theta,\sigma^2)$ as the density function for a $\mathsf{N}(\theta,\sigma^2)$ random variable, we may choose

$$g_0(r) = \frac{\phi(r;\eta,1)}{\Phi(\eta)},\ r>0$$
$$g_k(z_k \,|\, r, z_1,\ldots,z_{k-1}) = \frac{\phi(z_k;\mu_k,1)\mathbb{I}\{\tilde{l}_k \le z_k \le \tilde{u}_k\}}{\Phi(\tilde{u}_k-\mu_k)-\Phi(\tilde{l}_k-\mu_k)},\ k=1,2,\ldots,$$

for some $(\boldsymbol{\mu},\eta)$ that is yet to be specified. That is,

$$R \sim \mathsf{TN}_{(0,\infty)}(\eta,1)$$
$$Z_k \,|\, R, Z_1,\ldots,Z_{k-1} \sim \mathsf{TN}_{(\tilde{l}_k,\tilde{u}_k)}(\mu_k,1), \quad k=1,\ldots,d$$

where we denote $\mathsf{TN}_{(a,b)}(\theta,\sigma^2)$ as a $\mathsf{N}(\theta,\sigma^2)$ random variable, conditional on interval (a,b).

We can think of $(\boldsymbol{\mu},\eta)$ as some indexing parameter for the class of densities whose form is defined by g, since the shape of g depends on these parameters, it is also called 'tilting parameters'. The idea is now to choose the optimal tilting parameter for which g matches f as much as possible. Formally, let $\tilde{f}(\boldsymbol{z},r) = \ell f(\boldsymbol{z},r)$ denote the kernel part of f, for all $(\boldsymbol{z},r)\in\mathscr{R}$, so that $\ell = \int_{\mathscr{R}} \tilde{f}(\boldsymbol{z},r)\,d(\boldsymbol{z},r)$. Moreover let

$$\begin{aligned}\psi(\boldsymbol{z},r,\boldsymbol{\mu},\eta) &= \ln[\tilde{f}(\boldsymbol{z},r)/g(\boldsymbol{z},r)] \\ &= \frac{\|\boldsymbol{\mu}\|^2}{2} - \boldsymbol{z}^\top\boldsymbol{\mu} + \frac{\eta^2}{2} - r\eta + (\nu-1)\ln r + \ln\Phi(\eta) \\ &\quad + \sum_{k=1}^{d} \ln[\Phi(\tilde{u}_k-\mu_k)-\Phi(\tilde{l}_k-\mu_k)].\end{aligned}$$

The optimal tilting parameters $(\boldsymbol{\mu},\eta)$ is given by the solution to the optimization program

$$(\boldsymbol{z}^*, r^*, \boldsymbol{\mu}^*, \eta^*) = \underset{(\eta,\boldsymbol{\mu})}{\operatorname{argmin}} \underset{(\boldsymbol{z},r)\in\mathscr{R}}{\operatorname{argmax}} \psi(\boldsymbol{z},r;\eta,\boldsymbol{\mu}). \tag{2.5}$$

Here, since ψ is the log-likelihood ratio of f and g, intuitively, the maximization with respect to $(\boldsymbol{z},r)$ identifies the maximum deviance between f and g while the minimization with respect to $(\boldsymbol{\mu},\eta)$ then mitigates for this deviance. In other words, this program can be viewed as a minimax problem. Indeed, choosing g as a sequence of truncated normal guarantees ψ is concave in $(\boldsymbol{z},r)$ and is convex is $(\boldsymbol{\mu},\eta)$. This means that the unique solution to this program is given by solving for the system of equations $\nabla\psi = \mathbf{0}$, where $\nabla = (\partial_{\boldsymbol{z}},\partial_r,\partial_{\boldsymbol{\mu}},\partial_\eta)^\top$.

Moreover let g be the optimally tilted proposal density and let $(\boldsymbol{Z}_1,R_1),\ldots,(\boldsymbol{Z}_n,R_n) \overset{\text{iid}}{\sim} g$. It follows that

$$\ell = \mathbb{E}\frac{\tilde{f}(\boldsymbol{Z}_1,R_1)}{g(\boldsymbol{Z}_1,R_1)} = \mathbb{E}\exp(\psi(\boldsymbol{Z}_1,R_1;\boldsymbol{\mu}^*,\eta^*))$$

so that the importance sampling estimator is

$$\hat{\ell} = \frac{1}{n}\sum_{i=1}^{n} \exp(\psi(\boldsymbol{Z}_i, R_i; \boldsymbol{\mu}^*, \eta^*)).$$

A motivation for the program (2.5) given in [12] is that for all $(\boldsymbol{\mu}, \eta)$,

$$\mathrm{Var}(\hat{\ell}) \leq \exp[2 \sup_{(\boldsymbol{z},r)\in\mathscr{R}} \psi(\boldsymbol{z}, r; \boldsymbol{\mu}, \eta)].$$

Therefore, the program (2.5) minimizes an upper bound on the variance of $\hat{\ell}$, thereby giving a more accurate estimate (say for example in the MSE sense) for ℓ.

Next, denoting $c_{\boldsymbol{\mu},\eta} = \exp(\psi(\boldsymbol{z}^*, r^*; \boldsymbol{\mu}, \eta))$, since ψ is concave in $(\boldsymbol{z}, r)$, we have then $\tilde{f} \leq c_{\boldsymbol{\mu},\eta} g$. This means g is a valid rejection sampling proposal for f. Since the probability of retaining a proposal is $\ell/c_{\boldsymbol{\mu},\eta}$, maximizing the efficiency of this rejection sampling scheme is equivalent to minimizing $c_{\boldsymbol{\mu},\eta}$ with respect to $(\boldsymbol{\mu}, \eta)$. This again reduces to optimization program (2.5). To this end, we have following rejection sampler.

Algorithm 1 : Exact draw from (2.4)

Require: $(\boldsymbol{z}^*, r^*, \eta^*, \boldsymbol{\mu}^*)$

repeat

Draw $(\boldsymbol{Z}, R) \sim g$, the optimally tilted proposal density

Draw $E \sim \mathsf{Exp}(1)$

until $E \geq \psi(\boldsymbol{z}^*, r^*; \eta^*, \boldsymbol{\mu}^*) - \psi(\boldsymbol{Z}, R; \eta^*, \boldsymbol{\mu}^*)$

return $(\boldsymbol{Z}, R)$

Finally, given an exact draw $(\boldsymbol{Z}, R) \sim f$, one then computes $\boldsymbol{Z} = \mathrm{Q}\mathrm{L}_1^{-1}\boldsymbol{X}$ and $\boldsymbol{Y} = \sqrt{\nu}\boldsymbol{X}/R$ so that $\boldsymbol{Y} \sim h$. Note that L_1 is a lower diagonal matrix, so $\mathrm{L}_1^{-1}\boldsymbol{X}$ can be efficiently calculated by forward substitution.

2.3 Asymptotic efficiency of the importance sampling estimator

Suppose that ℓ is a function of a parameter γ in the following manner.

$$\ell(\gamma) = \mathbb{P}[\boldsymbol{Y} \geq \boldsymbol{l}(\gamma)], \quad \boldsymbol{Y} \sim \mathsf{t}_\nu(\boldsymbol{0}, \Sigma)$$

where $\max_i l_i > 0$, and at least one component of $\boldsymbol{l}(\gamma)$ diverges to ∞, that is, $\lim_{\gamma\uparrow\infty} \|\boldsymbol{l}(\gamma)\| = \infty$. We are interested in studying the asymptotic accuracy of the importance sampling estimator in the sense that $\gamma \uparrow \infty$. The key result for this section is the following theorem, whose proof is given later in the section.

Theorem 2.3.1 (Bounded Relative Error estimator). Suppose we wish to estimate the tail probability $\ell(\gamma) = \mathbb{P}[\boldsymbol{X} \geq \boldsymbol{l}(\gamma)]$, where $\boldsymbol{X} \sim \mathsf{t}_\nu(\boldsymbol{0}, \Sigma)$, and $\max_i l_i > 0$ with $\boldsymbol{l}(\gamma)/\gamma = \Theta(\boldsymbol{1})$ as $\gamma \uparrow \infty$. Then, the exponentially tilted estimator

$$\hat{\ell} = \exp(\psi(\boldsymbol{Z}, R; \boldsymbol{\mu}^*, \eta^*)), \qquad (\boldsymbol{Z}, R) \sim g(\boldsymbol{z}, r; \eta^*, \boldsymbol{\mu}^*),$$

where

$$\psi(r^*, \boldsymbol{z}^*; \eta^*, \boldsymbol{\mu}^*) = \inf_{\eta, \boldsymbol{\mu}} \sup_{(r, \boldsymbol{z}) \in \mathscr{R}} \psi(r, \boldsymbol{z}; \eta, \boldsymbol{\mu}),$$

is a bounded relative error estimator:

$$\limsup_{\gamma \uparrow \infty} \frac{\mathrm{Var}(\hat{\ell})}{\ell^2(\gamma)} < \infty.$$

Where we recall that for real-valued functions f and g, we write $g(\boldsymbol{x}) = \mathscr{O}(f(\boldsymbol{x}))$ if there are constants $M, C > 0$ such that $|f(\boldsymbol{x})| < C|g(\boldsymbol{x})|$ whenever $\|\boldsymbol{x}\|_\infty \geq M$. Furher, we write $g(\boldsymbol{x}) = \Theta(f(\boldsymbol{x}))$ whenever $g(\boldsymbol{x}) = \mathscr{O}(f(\boldsymbol{x}))$ and $f(\boldsymbol{x}) = \mathscr{O}(g(\boldsymbol{x}))$. Finally, in addition to $f(\boldsymbol{x}) = o(g(\boldsymbol{x}))$ being a shorthand notation for $\lim_{\|\boldsymbol{x}\|_\infty \uparrow \infty} f(\boldsymbol{x})/g(\boldsymbol{x}) = 0$, we use the notation $\lesssim$ for "asymptotically less than".

Theorem 2.3.1 states that our estimator $\hat{\ell}$ has bounded relative error even when the rarity parameter $\gamma \to +\infty$ so that at least one of the component of $\boldsymbol{l}$ diverges to $+\infty$. This is a desirable property but it is often unobtainable for naive estimators. For example, given $X_k \overset{\text{iid}}{\sim} \mathsf{Exp}(1)$, the naive estimator $\hat{\rho}(\gamma) := \frac{1}{n}\sum_{k=1}^n \mathbb{I}\{X_k > \gamma\}$ for the quantity $\rho(\gamma) := \mathbb{P}[X_1 > \gamma]$ has unbounded relative error as $\gamma \to +\infty$ (when n is held fixed):

$$\frac{\mathrm{Var}(\hat{\rho}(\gamma))}{\rho^2(\gamma)} = \frac{\rho(\gamma)(1 - \rho(\gamma))}{\rho^2(\gamma)} = \Theta(1/\rho(\gamma)) \uparrow \infty.$$

This is because the variance does not decay at a rate fast enough to offset the growth in $1/\rho(\gamma)$.

Quite often, practical applications are interested in the case $\boldsymbol{l}(\gamma) = (\gamma, \gamma, \ldots, \gamma)^\top$ so that $\ell(\gamma)$ is the probability of each coordinate exceeding some threshold γ. In such cases, the assumption $\boldsymbol{l}(\gamma)/\gamma = \Theta(\boldsymbol{1})$, which requires $\boldsymbol{l}$ to grow linearly with γ, is fulfilled.

We now present the required notations and lemmas for proving Theorem 2.3.1. The idea of the proof closely follows the proof of Theorem 1 in [11] – we shall analyze the set of equations derived from $\nabla\psi = \boldsymbol{0}$ as in the program (2.5), and this will give us an upper bound on $\mathrm{Var}(\hat{\ell})$, which then implies our desired result.

Let P be a permutation matrix which maps the vector $(1,\dots,d)^\top$ into the permutation $\boldsymbol{p} = (p_1,\dots,p_d)^\top$, that is, $\mathrm{P}(1,\dots,d)^\top = \boldsymbol{p}$. Note that $\ell(\gamma) = \mathbb{P}[\mathrm{P}\boldsymbol{Y} \geq \mathrm{P}\boldsymbol{l}(\gamma)]$ for any permutation $\boldsymbol{p}$, and $\mathrm{P}\boldsymbol{Y} \sim \mathsf{t}_\nu(\boldsymbol{0}, \mathrm{P}\Sigma\mathrm{P}^\top)$. We will specify $\boldsymbol{p}$ shortly.

Define the convex quadratic programming:

$$\begin{aligned} \min_{\boldsymbol{x}} \ & \frac{1}{2}\boldsymbol{x}^\top(\mathrm{P}\Sigma\mathrm{P}^\top)^{-1}\boldsymbol{x} \\ \text{subject to: } & \boldsymbol{x} \geq \mathrm{P}\boldsymbol{l}(\gamma) \end{aligned} \tag{2.6}$$

The Karush-Kuhn-Tucker conditions are a necessary and sufficient condition to find the unique solution:

$$\begin{aligned} (\mathrm{P}\Sigma\mathrm{P}^\top)^{-1}\boldsymbol{x} - \boldsymbol{\lambda} &= \boldsymbol{0} \\ \boldsymbol{\lambda} \geq \boldsymbol{0}, \ \ \mathrm{P}\boldsymbol{l} - \boldsymbol{x} &\leq \boldsymbol{0} \\ \boldsymbol{\lambda}^\top(\mathrm{P}\boldsymbol{l} - \boldsymbol{x}) &= 0\ , \end{aligned} \tag{2.7}$$

where $\boldsymbol{\lambda} \in \mathbb{R}^d$ is the Lagrange multiplier. Denote the number of active constraints in (2.6) by d_1 and the number of inactive constraints as d_2, so that $d_1 + d_2 = d$. Note that the number of active constraints $d_1 \geq 1$, because otherwise the solution is $\boldsymbol{x} = \boldsymbol{0}$, which implies $\mathrm{P}\boldsymbol{l} \leq \boldsymbol{0}$, thus reaching a contradiction.

Given the partition $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_1^\top, \boldsymbol{\lambda}_2^\top)^\top$ with $\dim(\boldsymbol{\lambda}_1) = d_1$ and $\dim(\boldsymbol{\lambda}_2) = d_2$, one can select the permutation vector $\boldsymbol{p}$ and the corresponding matrix P in such a way that all the active constraints in (2.7) correspond to $\boldsymbol{\lambda}_1 > \boldsymbol{0}$ and all the inactive ones to $\boldsymbol{\lambda}_2 = \boldsymbol{0}$. Henceforth, we assume that this reordering of the variables via the permutation operator P has been applied as a preprocessing step to both $\boldsymbol{l}$ and Σ so that $\mathrm{P}\boldsymbol{l} = \boldsymbol{l}$ and $\mathrm{P}\Sigma\mathrm{P}^\top = \Sigma$. If we partition $\boldsymbol{x}, \boldsymbol{l}$, and

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix},$$

then the KKT equations tell us that the optimal solution $\boldsymbol{x}^*$ is:

$$\begin{aligned} \boldsymbol{x}_1^* &= \Sigma_{11}\boldsymbol{\lambda}_1 = \boldsymbol{l}_1(\gamma) \\ \boldsymbol{x}_2^* &= \Sigma_{21}\boldsymbol{\lambda}_1 = \Sigma_{21}\Sigma_{11}^{-1}\boldsymbol{l}_1(\gamma) > \boldsymbol{l}_2(\gamma) \end{aligned}$$

with the global minimum $\frac{1}{2}(\boldsymbol{x}^*)^\top\Sigma^{-1}\boldsymbol{x}^* = \frac{1}{2}(\boldsymbol{x}_1^*)^\top\boldsymbol{\lambda}_1 = \frac{1}{2}\boldsymbol{l}_1^\top\Sigma_{11}^{-1}\boldsymbol{l}_1$.

Note that $\boldsymbol{x}^*(\gamma)$ is implicitly a function of γ and that in general the active constraint set of (2.6) and its size, d_1, also depends on the value of γ through $\boldsymbol{l}(\gamma)$. We henceforth assume that $\|\boldsymbol{l}(\gamma)\|$ diverges to infinity as $\gamma \uparrow \infty$ in such a way that, for large enough γ, the active constraint set of (2.6) becomes independent of γ.

One of our main contributions is to generalize the following result of [52].

Proposition 2.3.1 (Mill's Ratio For Multivariate Normal [52]). Under the conditions above, if $\boldsymbol{Z} \sim \mathsf{N}(\mathbf{0}, \Sigma)$, then as $\gamma \uparrow \infty$, we have:

$$\mathbb{P}[\boldsymbol{Z} \geq \boldsymbol{l}(\gamma)] = \frac{\mathbb{P}[\boldsymbol{Z}_2 \geq \boldsymbol{l}_\infty \,|\, \boldsymbol{Z}_1 = \mathbf{0}]}{(2\pi)^{d_1/2}|\Sigma_{11}|^{1/2}\prod_{k=1}^{d_1} \boldsymbol{e}_k^\top \Sigma_{11}^{-1}\boldsymbol{l}_1} \exp\left(-\frac{\boldsymbol{l}_1^\top \Sigma_{11}^{-1}\boldsymbol{l}_1}{2}\right)(1+o(1)),$$

where $\boldsymbol{l}_\infty := \lim_{\gamma\uparrow\infty}(\boldsymbol{l}_2(\gamma) - \boldsymbol{x}_2^*(\gamma))$ with $\boldsymbol{l}_\infty \leq \mathbf{0}$.

Using the notations introduced, we have the following lemma, which can be thought of as a multivariate extension to the Mill's ratio of the univariate student density [69, 84] and the Multivariate student version of Proposition 2.3.1.

Lemma 2.3.1 (Mill's Ratio For Multivariate Student). Suppose $\boldsymbol{Y} \sim \mathsf{t}_\nu(\mathbf{0}, \Sigma)$ with $\nu > 0$, and Σ and $\boldsymbol{l}$ satisfy the conditions imposed for the solution of (2.6). Then,

$$\mathbb{P}[\boldsymbol{Y} \geq \boldsymbol{l}(\gamma)] = (c + o(1)) \times \left(1 + \frac{\boldsymbol{l}_1(\gamma)^\top \Sigma_{11}^{-1}\boldsymbol{l}_1(\gamma)}{\nu}\right)^{-\nu/2}, \qquad \gamma \uparrow \infty,$$

where c is a constant, independent of γ, and is given by the expression:

$$c = \frac{2^{1-\nu/2}}{\Gamma(\nu/2)} \int_0^\infty r^{\nu-1}\mathbb{P}[\boldsymbol{Z} \geq r\boldsymbol{l}_\infty]\, \mathrm{d}r,$$

with $\boldsymbol{l}_\infty = \lim_{\gamma\uparrow\infty} \frac{\boldsymbol{l}(\gamma)}{\sqrt{\nu + \boldsymbol{l}_1(\gamma)^\top \Sigma_{11}^{-1}\boldsymbol{l}_1(\gamma)}}$ and $\boldsymbol{Z} \sim \mathsf{N}(\mathbf{0}, \Sigma)$.

Proof. First, we use the normal scale-mixture representation of $\boldsymbol{Y} \sim \mathsf{t}_\nu(\mathbf{0}, \Sigma)$ as $\boldsymbol{Y} = \sqrt{\nu}\boldsymbol{Z}/R$, where $\boldsymbol{Z} \sim \mathsf{N}(\mathbf{0}, \Sigma)$ is independent of

$$R \sim \mathrm{chi}_\nu(r) = \frac{\exp\left(-\frac{r^2}{2} + (\nu-1)\ln r\right)}{2^{\nu/2-1}\Gamma(\nu/2)}, \quad r > 0.$$

We can thus write ℓ as a conditional expectation:

$$\ell(\gamma) = \mathbb{P}\left[\frac{\sqrt{\nu}\boldsymbol{Z}}{R} \geq \boldsymbol{l}(\gamma)\right] = \mathbb{E}\mathbb{P}\left[\frac{\sqrt{\nu}\boldsymbol{Z}}{R} \geq \boldsymbol{l}(\gamma) \,\middle|\, R\right].$$

Next, condition on $R = r$, and let $\boldsymbol{\mu} = r\boldsymbol{x}^*/\sqrt{\nu}$, where $\boldsymbol{x}^*$ is the solution of (2.6). Denoting $\boldsymbol{t} = [\boldsymbol{t}_1^\top, \boldsymbol{t}_2^\top]^\top := r\boldsymbol{l}/\sqrt{\nu}$, and making a change of variable $\boldsymbol{z} \leftarrow \boldsymbol{z} - \boldsymbol{\mu}$, we obtain

$$\mathbb{P}\left[\frac{\sqrt{\nu}\boldsymbol{Z}}{R} \geq \boldsymbol{l}(\gamma) \,\middle|\, R = r\right] = \mathbb{P}[\boldsymbol{Z} \geq \boldsymbol{t}] =$$

$$\begin{aligned}
&= \mathbb{E}\exp(-\tfrac{\boldsymbol{\mu}^\top\Sigma^{-1}\boldsymbol{\mu}}{2} - \boldsymbol{Z}^\top\Sigma^{-1}\boldsymbol{\mu})\mathbb{I}\{\boldsymbol{Z} \geq \boldsymbol{t} - \boldsymbol{\mu}\} \\
&= \exp(-\tfrac{\boldsymbol{\mu}^\top\Sigma^{-1}\boldsymbol{\mu}}{2})\mathbb{E}\exp(-\boldsymbol{Z}_1^\top\Sigma_{11}^{-1}\boldsymbol{t}_1)\mathbb{I}\{\boldsymbol{Z}_1 \geq \boldsymbol{t}_1 - \boldsymbol{\mu}_1, \boldsymbol{Z}_2 \geq \boldsymbol{t}_2 - \boldsymbol{\mu}_2\} \\
&= \exp(-\boldsymbol{t}_1^\top\Sigma_{11}^{-1}\boldsymbol{t}_1/2)\mathbb{E}\exp(-\boldsymbol{Z}_1^\top\Sigma_{11}^{-1}\boldsymbol{t}_1)\mathbb{I}\{\boldsymbol{Z}_1 \geq \boldsymbol{0}, \boldsymbol{Z}_2 \geq \boldsymbol{t}_2 - \boldsymbol{\mu}_2\}.
\end{aligned}$$

In other words, we have:

$$\mathbb{P}[\boldsymbol{Z} \geq \boldsymbol{t}] = \exp(-\tfrac{r^2\boldsymbol{l}_1^\top\Sigma_{11}^{-1}\boldsymbol{l}_1}{2\nu})\mathbb{E}\exp(-\tfrac{r\boldsymbol{Z}_1^\top\Sigma_{11}^{-1}\boldsymbol{l}_1}{\sqrt{\nu}})\mathbb{I}\{\boldsymbol{Z}_1 \geq \boldsymbol{0}, \boldsymbol{Z}_2 \geq \tfrac{r(\boldsymbol{l}_2 - \Sigma_{21}\Sigma_{11}^{-1}\boldsymbol{l}_1)}{\sqrt{\nu}}\} \tag{2.8}$$

Let $\mathscr{D} \equiv \{\boldsymbol{z} : \boldsymbol{z}_1 \geq \boldsymbol{0}, \boldsymbol{z}_2 \geq \frac{r(\boldsymbol{l}_2 - \Sigma_{21}\Sigma_{11}^{-1}\boldsymbol{l}_1)}{\sqrt{\nu}}\}$. We can now rewrite (2.8) as an integral and integrate over r. This gives $\ell(\gamma) =:$

$$\begin{aligned}
&= \textstyle\int_0^\infty \int_{\mathscr{D}} \mathrm{chi}_\nu(r)\phi_\Sigma(\boldsymbol{z}) \exp\left(-r^2\boldsymbol{l}_1^\top\Sigma_{11}^{-1}\boldsymbol{l}_1/(2\nu) - r\boldsymbol{z}_1^\top\Sigma_{11}^{-1}\boldsymbol{l}_1/\sqrt{\nu}\right) d\boldsymbol{z}\, dr \\
&= \tfrac{2^{1-(\nu+d)/2}\pi^{-d/2}}{\Gamma(\frac{\nu}{2})|\Sigma|^{1/2}} \textstyle\int_0^\infty\int_{\mathscr{D}} \exp\left(-\frac{r^2}{2}\left(1 + \frac{\boldsymbol{l}_1^\top\Sigma_{11}^{-1}\boldsymbol{l}_1}{2\nu}\right) - \frac{\boldsymbol{z}^\top\Sigma^{-1}\boldsymbol{z}}{2} - \frac{r\boldsymbol{z}_1^\top\Sigma_{11}^{-1}\boldsymbol{l}_1}{\sqrt{\nu}} + (\nu-1)\ln r\right) d\boldsymbol{z}\, dr \\
&= \frac{2^{1-(\nu+d)/2}\pi^{-d/2}}{\Gamma(\frac{\nu}{2})|\Sigma|^{1/2}\left(1+\frac{\boldsymbol{l}_1^\top\Sigma_{11}^{-1}\boldsymbol{l}_1}{\nu}\right)^{\nu/2}} \textstyle\int_0^\infty\int_{\mathscr{D}} \exp\left(-\frac{u^2}{2} - \frac{\boldsymbol{z}^\top\Sigma^{-1}\boldsymbol{z}}{2} - \frac{u\,\boldsymbol{z}_1^\top\Sigma_{11}^{-1}\boldsymbol{l}_1}{\sqrt{\nu+\boldsymbol{l}_1^\top\Sigma_{11}^{-1}\boldsymbol{t}_1}} + (\nu-1)\ln u\right) d\boldsymbol{z}\, dr \\
&= \frac{1}{\left(1+\frac{\boldsymbol{l}_1^\top\Sigma_{11}^{-1}\boldsymbol{l}_1}{\nu}\right)^{\nu/2}} \textstyle\int_0^\infty\int_{\mathbb{R}^d} \mathrm{chi}_\nu(u)\phi_\Sigma(\boldsymbol{z}) \exp\left(-\frac{u\,\boldsymbol{z}_1^\top\Sigma_{11}^{-1}\boldsymbol{l}_1}{\sqrt{\nu+\boldsymbol{l}_1^\top\Sigma_{11}^{-1}\boldsymbol{l}_1}}\right) \mathbb{I}\left\{\boldsymbol{z}_1 \geq \boldsymbol{0}, \boldsymbol{z}_2 \geq \frac{u(\boldsymbol{l}_2 - \Sigma_{21}\Sigma_{11}^{-1}\boldsymbol{l}_1)}{\sqrt{\nu+\boldsymbol{l}_1^\top\Sigma_{11}^{-1}\boldsymbol{l}_1}}\right\} d\boldsymbol{z}\, dr \\
&= \left(1 + \frac{\boldsymbol{l}_1^\top\Sigma_{11}^{-1}\boldsymbol{l}_1}{\nu}\right)^{-\nu/2} \mathbb{E}\exp\left(-\frac{R\,\boldsymbol{Z}_1^\top\Sigma_{11}^{-1}\boldsymbol{l}_1}{\sqrt{\nu+\boldsymbol{l}_1^\top\Sigma_{11}^{-1}\boldsymbol{l}_1}}\right) \mathbb{I}\left\{\boldsymbol{Z}_1 \geq \boldsymbol{0}, \boldsymbol{Z}_2 \geq \frac{R(\boldsymbol{l}_2 - \Sigma_{21}\Sigma_{11}^{-1}\boldsymbol{l}_1)}{\sqrt{\nu+\boldsymbol{l}_1^\top\Sigma_{11}^{-1}\boldsymbol{l}_1}}\right\}
\end{aligned}$$

where the third line follows from the change of variable $u = r\sqrt{1 + \frac{\boldsymbol{l}_1^\top\Sigma_{11}^{-1}\boldsymbol{l}_1}{\nu}}$. Next, using formula (2.8) we rewrite the last expression as:

$$\left(1 + \frac{\boldsymbol{l}_1^\top\Sigma_{11}^{-1}\boldsymbol{l}_1}{\nu}\right)^{-\nu/2} \mathbb{E}\exp\left(\frac{R^2\boldsymbol{l}_1^\top\Sigma_{11}^{-1}\boldsymbol{l}_1}{2(\nu+\boldsymbol{l}_1^\top\Sigma_{11}^{-1}\boldsymbol{l}_1)}\right) \mathbb{P}\left[\boldsymbol{Z} \geq \frac{R\boldsymbol{l}}{\sqrt{\nu+\boldsymbol{l}_1^\top\Sigma_{11}^{-1}\boldsymbol{l}_1}} \,\middle|\, R\right]$$

We now seek to apply the dominated convergence theorem to the expectation in the last displayed equation. For this we need the upper bound (recall that $\Sigma_{11}^{-1}\boldsymbol{l}_1 \geq \mathbf{0}$)

$$\begin{aligned}
\exp\left(\frac{r^2\boldsymbol{l}_1^\top\Sigma_{11}^{-1}\boldsymbol{l}_1}{2(\nu+\boldsymbol{l}_1^\top\Sigma_{11}^{-1}\boldsymbol{l}_1)}\right)\mathbb{P}\left[\boldsymbol{Z}\geq\frac{r\boldsymbol{l}}{\sqrt{\nu+\boldsymbol{l}_1^\top\Sigma_{11}^{-1}\boldsymbol{l}_1}}\right] &\leq \exp(r^2/2)\mathbb{P}\left[\boldsymbol{Z}_1\geq\frac{r\boldsymbol{l}_1}{\sqrt{\nu+\boldsymbol{l}_1^\top\Sigma_{11}^{-1}\boldsymbol{l}_1}}\right]\\
&\leq \exp(r^2/2)\mathbb{P}\left[\boldsymbol{l}_1^\top\Sigma_{11}^{-1}\boldsymbol{Z}_1\geq\frac{r\boldsymbol{l}_1^\top\Sigma_{11}^{-1}\boldsymbol{l}_1}{\sqrt{\nu+\boldsymbol{l}_1^\top\Sigma_{11}^{-1}\boldsymbol{l}_1}}\right]\\
&= \exp(r^2/2)\overline{\Phi}\left[r\sqrt{\frac{\boldsymbol{l}_1^\top\Sigma_{11}^{-1}\boldsymbol{l}_1}{\nu+\boldsymbol{l}_1^\top\Sigma_{11}^{-1}\boldsymbol{l}_1}}\right]\\
&\leq \exp(r^2/2)\overline{\Phi}\left(r\right).
\end{aligned}$$

The last expression is integrable in the sense that

$$\begin{aligned}
\int_0^\infty \mathrm{chi}_\nu(r)\exp(r^2/2)\overline{\Phi}\left(r\right)\,\mathrm{d}r &= \frac{2^{1-\nu/2}}{\Gamma(\nu/2)}\int_0^\infty r^{\nu-1}\overline{\Phi}\left(r\right)\,dr\\
&= \frac{2^{1-\nu/2}}{\Gamma(\nu/2)2\nu}\int_{-\infty}^\infty |u|^\nu\phi(u)\,du\\
&= \frac{2^{1-\nu/2}\Gamma((\nu+1)/2)2^{\nu/2}}{\sqrt{\pi}\Gamma(\nu/2)2\nu} = \frac{\Gamma((\nu+1)/2)}{\sqrt{\pi}\Gamma(\nu/2)\nu} < \infty.
\end{aligned}$$

In addition, as $\gamma\uparrow\infty$, by Lemma 2.3.2 we have the pointwise limits:

$$\exp\left[\tfrac{r^2\boldsymbol{l}_1^\top\Sigma_{11}^{-1}\boldsymbol{l}_1}{2(\nu+\boldsymbol{l}_1^\top\Sigma_{11}^{-1}\boldsymbol{l}_1)}\right]\mathbb{P}\left[\boldsymbol{Z}\geq\tfrac{r\boldsymbol{l}}{\sqrt{\nu+\boldsymbol{l}_1^\top\Sigma_{11}^{-1}\boldsymbol{l}_1}}\right]\to\exp(r^2/2)\mathbb{P}[\boldsymbol{Z}\geq r\boldsymbol{l}_\infty].$$

Therefore, by the dominated convergence theorem

$$\lim_{\gamma\uparrow\infty}\mathbb{E}\exp\left(\tfrac{R^2\boldsymbol{l}_1^\top\Sigma_{11}^{-1}\boldsymbol{l}_1}{2(\nu+\boldsymbol{l}_1^\top\Sigma_{11}^{-1}\boldsymbol{l}_1)}\right)\mathbb{P}\left[\boldsymbol{Z}\geq\tfrac{R\boldsymbol{l}}{\sqrt{\nu+\boldsymbol{l}_1^\top\Sigma_{11}^{-1}\boldsymbol{l}_1}}\,\middle|\,R\right] = \tfrac{2^{1-\nu/2}}{\Gamma(\nu/2)}\int_0^\infty r^{\nu-1}\mathbb{P}[\boldsymbol{Z}\geq r\boldsymbol{l}_\infty]\,dr.$$

This concludes the proof. □

Lemma 2.3.2 (Continuity of Gaussian tail). Suppose that $\boldsymbol{Z}\sim\mathscr{N}(\mathbf{0},\Sigma)$ for some positive definite matrix Σ, and $\boldsymbol{a}_n\to\boldsymbol{a}$ as $n\uparrow\infty$. Then, the tail of the multivariate Gaussian is continuous:

$$\lim_{n\uparrow\infty}\mathbb{P}[\boldsymbol{Z}\geq\boldsymbol{a}_n] = \mathbb{P}[\boldsymbol{Z}\geq\boldsymbol{a}].$$

Proof. The proof is yet another application of the dominated convergence theorem to show that:

$$\int_{[\mathbf{0},\boldsymbol{\infty})} \phi_\Sigma(\boldsymbol{z}+\boldsymbol{a}_n)\,d\boldsymbol{z} \to \int_{[\mathbf{0},\boldsymbol{\infty})} \phi_\Sigma(\boldsymbol{z}+\boldsymbol{a})\,d\boldsymbol{z} = \mathbb{P}[\boldsymbol{Z}\geq \boldsymbol{a}].$$

Since Σ is a positive definite matrix, the $\|\boldsymbol{x}\|_\Sigma^2 := \boldsymbol{x}^\top\Sigma^{-1}\boldsymbol{x}$ is a norm satisfying $\|\boldsymbol{z}+\boldsymbol{a}_n\|_\Sigma^2 \leq 2(\|\boldsymbol{z}\|_\Sigma^2 + \|\boldsymbol{a}_n\|_\Sigma^2)$. Therefore,

$$\int_{[\mathbf{0},\boldsymbol{\infty})} \phi_\Sigma(\boldsymbol{z}+\boldsymbol{a}_n)\,d\boldsymbol{z} \leq \frac{\exp(-\|\boldsymbol{a}_n\|_\Sigma^2)}{2^{n/2}} \int_{[\mathbf{0},\boldsymbol{\infty})} \phi_{\Sigma/2}(\boldsymbol{z})\,d\boldsymbol{z} < \infty,$$

and the conditions for the dominated convergence theorem are met. □

Finally, we have the following proof for Theorem 2.3.1

Proof. First, note that the second moment is

$$\begin{aligned}\int g(\boldsymbol{z},r;\boldsymbol{\mu}^*,\eta^*)\exp(2\psi(\boldsymbol{z},r;\boldsymbol{\mu}^*,\eta^*))\,d\boldsymbol{z}\,dr &= \int_{\mathscr{R}} \mathrm{chi}_\nu(r)\phi_\Sigma(\boldsymbol{z})\exp(\psi(\boldsymbol{z},r;\boldsymbol{\mu}^*,\eta^*))\,d\boldsymbol{z}\,dr \\ &\leq \ell(\gamma)\exp(\psi(\boldsymbol{z}^*,r^*;\boldsymbol{\mu}^*,\eta^*)).\end{aligned}$$

Since the properties of ψ imply that

$$\psi(\boldsymbol{z}^*,r^*;\boldsymbol{\mu}^*,\eta^*) \leq \psi(\boldsymbol{z}^*,r^*;\mathbf{0},\eta^*) \leq \frac{(\eta^*)^2}{2} - r^*\eta^* + (\nu-1)\ln r^* + \ln\Phi(\eta^*),$$

bounded relative error will follow if we can show that

$$\frac{(r^*)^{\nu-1}\Phi(\eta^*)\exp(\frac{(\eta^*)^2}{2} - r^*\eta^*)}{\ell(\gamma)}$$

remains bounded in γ. The pair (r^*,η^*) is determined from the solution to the saddle-point problem: $\max_{r,\boldsymbol{z}}\min_{\eta,\boldsymbol{\mu}}\psi(\boldsymbol{z},r;\boldsymbol{\mu},\eta)$. This can be obtained by setting the gradient of ψ with respect to the vector $(\boldsymbol{z},r,\boldsymbol{\mu},\eta)$ to zero: $\nabla\psi = \mathbf{0}$. We now introduce the following notation that will allow us to express $\nabla\psi = \mathbf{0}$ explicitly. Let L be the lower triangular Cholesky factor of $\Sigma = \mathrm{L}\mathrm{L}^\top$. Define $\mathrm{D} = \mathrm{diag}(\mathrm{L})$, $\breve{\mathrm{L}} = \mathrm{D}^{-1}\mathrm{L}$,

$$\tilde{\boldsymbol{l}} = \frac{r}{\sqrt{\nu}}\mathrm{D}^{-1}\boldsymbol{l}(\gamma) - (\breve{\mathrm{L}} - \mathrm{I})\boldsymbol{z},$$

and vector $\boldsymbol{\Psi}$ with elements $\Psi_k = \phi(\tilde{l}_k - \mu_k)/\overline{\Phi}(\tilde{l}_k - \mu_k)$. Then, $\nabla\psi = \mathbf{0}$ can be written as

$$\begin{aligned} (\breve{\mathrm{L}}^\top - \mathrm{I})\boldsymbol{\Psi} - \boldsymbol{\mu} &= \mathbf{0} \\ \frac{\nu-1}{r} - \eta - \frac{1}{\sqrt{\nu}}\boldsymbol{\Psi}^\top \mathrm{D}^{-1}\boldsymbol{l}(\gamma) &= 0 \\ \boldsymbol{\mu} + \boldsymbol{\Psi} - \boldsymbol{z} &= \mathbf{0} \\ \eta + \frac{\phi(\eta)}{\Phi(\eta)} - r &= 0. \end{aligned} \tag{2.9}$$

Next, we verify via substitution that the solution of (2.9) as $\gamma \uparrow \infty$ satisfies

$$r^* = \mathscr{O}(\gamma^{-1}), \quad \boldsymbol{z}^* = \mathscr{O}(\mathbf{1}), \quad \eta^* = \mathscr{O}(-\gamma), \quad \boldsymbol{\mu}^* = \mathscr{O}(\mathbf{1}).$$

First, equations one and three in (2.9) are trivially satisfied and we can deduce that $\boldsymbol{\Psi} = \mathscr{O}(\mathbf{1})$. Second, since $\tilde{\boldsymbol{l}} = \mathscr{O}(r\boldsymbol{l}(\gamma)) = \mathscr{O}(\mathbf{1})$, it follows that equation two in (2.9) is equivalent to

$$r^*\eta^* = \nu - 1 - \frac{r^*}{\sqrt{\nu}}\boldsymbol{\Psi}^\top \mathrm{D}^{-1}\boldsymbol{l}(\gamma) = \mathscr{O}(1).$$

Finally, note that Mill's ratio

$$\frac{\Phi(\eta)}{\phi(\eta)} \simeq -\frac{1}{\eta} + \frac{1}{\eta^3}, \qquad \eta \downarrow -\infty,$$

implies that equation four is asymptotically equivalent to $r\eta^2 + \eta - r \simeq 0$. The solution of this quadratic equation in turn implies that $\eta \simeq (-1 - \sqrt{1+4r^2})/(2r) \simeq -1/r$. In other words, $\eta^* r^* = \mathscr{O}(1)$, as desired. Therefore, if $\tilde{\psi}$ denotes the value of ψ at the solution (2.9), we have

$$\begin{aligned} \tilde{\psi} &= \frac{\|\boldsymbol{\mu}^*\|^2}{2} - (\boldsymbol{z}^*)^\top\boldsymbol{\mu}^* + \frac{(\eta^*)^2}{2} - r^*\eta^* + (\nu-1)\ln r^* + \ln\Phi(\eta^*) + \sum_{k=1}^{d}\ln\overline{\Phi}(\tilde{l}_k - \mu_k^*) \\ &= \mathscr{O}(1) + \frac{(\eta^*)^2}{2} + (\nu-1)\ln r^* + \ln\overline{\Phi}(-\eta^*). \end{aligned}$$

By Mill's ratio inequality:

$$\ln\overline{\Phi}(-\eta) \leq -\eta^2/2 - \frac{1}{2}\ln(2\pi) - \ln(-\eta),$$

we obtain:

$$\tilde{\psi} \lesssim \mathscr{O}(1) - \ln(-\eta^*) - \frac{1}{2}\ln(2\pi) + (\nu-1)\ln r^* = -\nu\log(\gamma) + \mathscr{O}(1).$$

In other words, there exist constants $c_1, c_2 > 0$ such that $\exp(\tilde{\psi}) \leq c_1\gamma^{-\nu}$ for every $\gamma > c_2$. Therefore, $\mathrm{Var}(\hat{\ell}) = \mathbb{E}\exp(\psi(\boldsymbol{Z}, R; \boldsymbol{\mu}^*, \eta^*)) - \ell^2 \lesssim \exp(\tilde{\psi}) - \ell^2 \leq c_1\gamma^{-\nu} - \ell^2(\gamma)$ and since by Proposition 2.3.1

$$\ell(\gamma) \simeq c \times \left(1 + \gamma \times \underbrace{\frac{\boldsymbol{l}_1^\top \Sigma_{11}^{-1} \boldsymbol{l}_1}{\nu \times \gamma}}_{\Theta(1)}\right)^{-\nu/2} = \Theta(\gamma^{-\nu/2}), \quad \gamma \uparrow \infty,$$

we have $\limsup_{\gamma\uparrow\infty} \mathrm{Var}(\hat{\ell})/\ell^2 < \infty$. □

2.4 The rejection sampler

Although Algorithm 1 is motivated by simulating from h, in this section we show that Algorithm 2.1 actually renders exact sampling schemes for some Bayesian posterior densities. In particular, we study the posterior densities of the Bayesian constrained linear regression, Bayesian Tobit model, and the Bayesian smoothing spline.

2.4.1 Constrained Linear Regression

Consider the linear regression model

$$\boldsymbol{Y} = \mathrm{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \mathrm{X} \in \mathbb{R}^{m\times d}, \quad \boldsymbol{\epsilon} \sim \mathsf{N}(\mathbf{0}, \sigma^2\mathrm{I})$$

with the prior information $p(\boldsymbol{\beta}) \propto \mathbb{I}\{\boldsymbol{l} \leq \mathrm{C}\boldsymbol{\beta} \leq \boldsymbol{u}\}$ for some appropriate matrix C and vectors $\boldsymbol{l}, \boldsymbol{u}$. Assuming for simplicity a non-informative prior $p(\sigma) \propto \sigma^{-2}$ (the approach is straightforward to generalize to an inverse gamma prior), the posterior from which we wish to sample is:

$$\pi(\boldsymbol{\beta}, \sigma) \propto \exp\left(-\frac{\|\boldsymbol{y} - \mathrm{X}\boldsymbol{\beta}\|^2}{2\sigma^2} - (m+2)\ln\sigma\right) \times \mathbb{I}\{\boldsymbol{l} \leq \mathrm{C}\boldsymbol{\beta} \leq \boldsymbol{u}\}.$$

Readers should note that the posterior distribution is conditional on the observed vector of response $\boldsymbol{y}$. However for the ease of presentation, we have suppressed it from our notation and write $\pi(\cdot)$ instead of $\pi(\cdot \,|\, \boldsymbol{y})$. We do this for every examples in the remaining of this chapter. If $\mathrm{H} := \mathrm{X}(\mathrm{X}^\top\mathrm{X})^{-1}\mathrm{X}^\top$ is the hat matrix, $\hat{\boldsymbol{\beta}}$ is the least squares estimate, and $s^2 := \boldsymbol{y}^\top(\mathrm{I} - \mathrm{H})\boldsymbol{y} = \|\boldsymbol{y} - \mathrm{X}\hat{\boldsymbol{\beta}}\|^2$ is the norm of the residuals squared, then

$$\pi(\boldsymbol{\beta}, \sigma) \propto \exp\left(-\frac{s^2}{2\sigma^2} - (m+2)\ln\sigma - \frac{(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^\top \mathrm{X}^\top\mathrm{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})}{2\sigma^2}\right) \times \mathbb{I}\{\boldsymbol{l} \leq \mathrm{C}\boldsymbol{\beta} \leq \boldsymbol{u}\}. \tag{2.10}$$

Let $\mathrm{L}_1\mathrm{L}_1^\top = \mathrm{X}^\top\mathrm{X}$ be the lower triangular Cholesky decomposition of $\mathrm{X}^\top\mathrm{X}$ and $\mathrm{LQ} = \mathrm{CL}_1^{-\top}$ be the LQ decomposition of matrix $\mathrm{CL}_1^{-\top}$. Then, the bijective smooth transformation

$$\begin{aligned} r &= s/\sigma \\ \boldsymbol{z} &= \mathrm{Q}\,\mathrm{L}_1^\top(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})/\sigma \\ \boldsymbol{l} &\leftarrow \sqrt{\nu}(\boldsymbol{l} - \mathrm{C}\hat{\boldsymbol{\beta}})/s \\ \boldsymbol{u} &\leftarrow \sqrt{\nu}(\boldsymbol{u} - \mathrm{C}\hat{\boldsymbol{\beta}})/s, \end{aligned}$$

where $\nu \leftarrow (m - d + 1) \geq 1$, yields the density:

$$f(\boldsymbol{z}, r) \propto \exp\left(-\tfrac{1}{2}\|\boldsymbol{z}\|^2 - \tfrac{r^2}{2} + (\nu - 1)\ln r\right) \times \mathbb{I}\{r\boldsymbol{l} \leq \sqrt{\nu}\mathrm{L}\boldsymbol{z} \leq r\boldsymbol{u}\},$$

and thus amenable to Algorithm 1.

We consider the 'Apple dataset' [18, 17] which records 207 observations of the number apples produced (in cartons) along with the number of trees of each year of age from various growers. This can be modeled by the Bayesian constrained linear regression where the i-th response, $y_i \in \mathbb{R}$, is the number of apples produced and the i-th predictor vector, $\boldsymbol{x}_i \in \mathbb{R}^{10}$, records the number of trees of age '$j + 1$', $j = 1, \ldots, 10$ being the entry index within the vector $\boldsymbol{x}_i$. Note that here trees of year 1 is considered to have zero production and 'age 11' is considered as the mature age of an apple tree, so that any tree above an age of 11 is recorded as 'age 11'. Finally, the prior $\pi(\boldsymbol{\beta}) \propto \mathbb{I}\{\beta_1 \leq \beta_2 \leq \ldots \leq \beta_{10}\}$ captures the prior believe that a more mature tree produces more apples. In this example, the optimal tilting parameter is

$$\eta = -0.0425,\ \boldsymbol{\mu} = (-0.0425, -6.0946, -2.6257, -2.5290, -0.0017, -0.0170, 0, 0, 0, 13.4835)^\top.$$

This gives an acceptance probability for this rejection sampling scheme is about 0.637, suggesting that our rejection sampling scheme is efficient here. The results of the inference are summarized in figure 2.1 and table 2.1.

	mean	0.025-quantile	0.975-quantile	sample std.
Age 2	3.19×10^{-2}	6.11×10^{-3}	5.46×10^{-2}	1.20×10^{-2}
Age 3	4.85×10^{-2}	2.56×10^{-2}	7.77×10^{-2}	1.33×10^{-2}
Age 4	1.79×10^{-1}	1.53×10^{-1}	2.06×10^{-1}	1.38×10^{-2}
Age 5	2.79×10^{-1}	2.02×10^{-1}	3.69×10^{-1}	4.37×10^{-2}
Age 6	5.21×10^{-1}	3.54×10^{-1}	7.08×10^{-1}	9.19×10^{-2}
Age 7	7.02×10^{-1}	5.81×10^{-1}	8.28×10^{-1}	6.45×10^{-2}
Age 8	7.31×10^{-1}	6.16×10^{-1}	8.54×10^{-1}	6.26×10^{-2}
Age 9	8.62×10^{-1}	6.81×10^{-1}	1.20	1.32×10^{-1}
Age 10	9.57×10^{-1}	7.25×10^{-1}	1.31	1.56×10^{-1}
Age 11	1.16	8.28×10^{-1}	1.66	2.18×10^{-1}

Table 2.1: Estimated mean, 0.95 credible interval and standard deviation of the posterior distribution from the exact simulation.

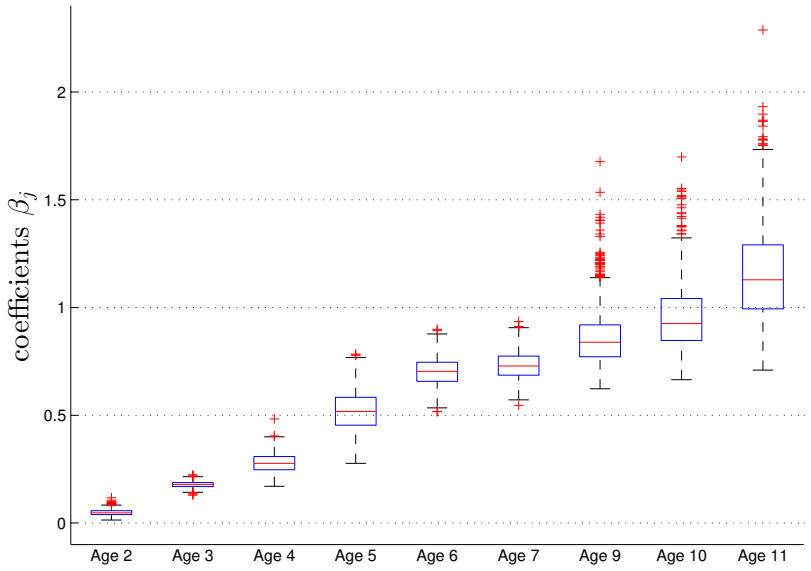

Figure 2.1: The empirical posterior distribution for the New Zealand apples dataset derived from $n = 10^4$ independent exact draws.

2.4.2 Tobit Model

Again, let X be the design (data) matrix and $\boldsymbol{Y}$ be the vector of response. The Tobit linear regression model concerns observations where the i-th response is (left) censored at some lower threshold u_i (known a priori). The Tobit regression model with normally distributed error, treats this problem by assuming that Y_i takes the following form:

$$Y_i = \begin{cases} W_i, & \text{if } u_i < W_i, \\ u_i, & \text{if } W_i \leq u_i, \end{cases}, \qquad \text{where} \qquad \boldsymbol{W} \sim \mathsf{N}(\mathrm{X}\boldsymbol{\beta}, \sigma^2 I).$$

The posterior, given for the data $\boldsymbol{y}$ and with uninformative priors, say $p(\boldsymbol{\beta}) \propto 1$ and $p(\sigma) \propto \sigma^{-2}$, is then of the form:

$$\pi(\boldsymbol{\beta}, \sigma) \propto \exp\left(- \sum_{i:y_i > u_i} \left(\tfrac{(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2}{2\sigma^2} + \ln \sigma \right) + \sum_{i:y_i = u_i} \ln \Phi((u_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})/\sigma) \right) \times \sigma^{-2}.$$

Let $\overline{\boldsymbol{y}}$ and $\underline{\boldsymbol{y}}$ be vectors that collect all $y_i > u_i$ and $y_i = u_i$, respectively. Denote the corresponding matrix with predictors via $\overline{\mathrm{X}}$ and $\underline{\mathrm{X}}$, respectively. Using a latent variable w_i for each $y_i = u_i$, we can write:

$$\pi(\boldsymbol{\beta}, \sigma, \boldsymbol{w}) \propto \exp\left(- \frac{\|\overline{\boldsymbol{y}} - \overline{\mathrm{X}}\boldsymbol{\beta}\|^2}{2\sigma^2} - \frac{\|\boldsymbol{w} - \underline{\mathrm{X}}\boldsymbol{\beta}\|^2}{2\sigma^2} - (m+2)\ln\sigma \right) \mathbb{I}\{\boldsymbol{w} \leq \boldsymbol{u}\}$$

so that the marginal of $(\boldsymbol{\beta}, \sigma)$ has the desired posterior pdf. Note that, conditional on $(\sigma, \boldsymbol{w})$, the distribution of $\boldsymbol{\beta}$ is $\mathsf{N}(\mathrm{C}(\overline{\mathrm{X}}^\top \overline{\boldsymbol{y}} + \underline{\mathrm{X}}^\top \boldsymbol{w}), \sigma^2 \mathrm{C})$, where $\mathrm{C}^{-1} = \overline{\mathrm{X}}^\top \overline{\mathrm{X}} + \underline{\mathrm{X}}^\top \underline{\mathrm{X}}$. Given a draw of $(\sigma, \boldsymbol{W})$ from the marginal, simulating from $\mathsf{N}(\mathrm{C}(\overline{\mathrm{X}}^\top \overline{\boldsymbol{y}} + \underline{\mathrm{X}}^\top \boldsymbol{w}), \sigma^2 \mathrm{C})$ is a routine task on standard mathematics/statistics toolbox, thus the only difficulty is to simulate from the marginal density of $(\sigma, \boldsymbol{W})$,

$$\pi(\sigma, \boldsymbol{w}) \propto \exp\left(-\tfrac{\|\boldsymbol{w}\|^2}{2\sigma^2} + \tfrac{(\overline{\mathrm{X}}^\top \overline{\boldsymbol{y}} + \underline{\mathrm{X}}^\top \boldsymbol{w})^\top \mathrm{C} (\overline{\mathrm{X}}^\top \overline{\boldsymbol{y}} + \underline{\mathrm{X}}^\top \boldsymbol{w})}{2\sigma^2} - \tfrac{\|\overline{\boldsymbol{y}}\|^2}{2\sigma^2}\right) \mathbb{I}\{\boldsymbol{w} \leq \boldsymbol{u}\} \times \sigma^{d-m-2}.$$

We may rewrite this as

$$\pi(\sigma, \boldsymbol{w}) \propto \exp\left(-\tfrac{\boldsymbol{w}^\top (\mathrm{I} - \underline{\mathrm{X}} \mathrm{C} \underline{\mathrm{X}}^\top) \boldsymbol{w}}{2\sigma^2} + \tfrac{\overline{\boldsymbol{y}}^\top \overline{\mathrm{X}} \mathrm{C} \underline{\mathrm{X}}^\top \boldsymbol{w}}{\sigma^2} - \tfrac{\overline{\boldsymbol{y}}^\top (\mathrm{I} - \overline{\mathrm{X}} \mathrm{C} \overline{\mathrm{X}}^\top) \overline{\boldsymbol{y}}}{2\sigma^2}\right) \mathbb{I}\{\boldsymbol{w} \leq \boldsymbol{u}\} \times \sigma^{d-m-2}$$

the identity $(\mathrm{I} - \underline{\mathrm{X}} \mathrm{C} \underline{\mathrm{X}}^\top)^{-1} = \mathrm{I} + \underline{\mathrm{X}} (\overline{\mathrm{X}}^\top \overline{\mathrm{X}})^{-1} \underline{\mathrm{X}}^\top$, then yields

$$\pi(\sigma, \boldsymbol{w}) \propto \exp\left(-\tfrac{(\boldsymbol{w} - \hat{\boldsymbol{w}})^\top (\mathrm{I} - \underline{\mathrm{X}} \mathrm{C} \underline{\mathrm{X}}^\top)(\boldsymbol{w} - \hat{\boldsymbol{w}})}{2\sigma^2} - \tfrac{s^2}{2\sigma^2} - (m - d + 2) \ln \sigma\right) \mathbb{I}\{\boldsymbol{w} \leq \boldsymbol{u}\},$$

where

$$\begin{aligned}
\hat{\boldsymbol{w}} &:= (\mathrm{I} - \underline{\mathrm{X}} \mathrm{C} \underline{\mathrm{X}}^\top)^{-1} \underline{\mathrm{X}} \mathrm{C} \overline{\mathrm{X}}^\top \overline{\boldsymbol{y}} = \underline{\mathrm{X}} (\overline{\mathrm{X}}^\top \overline{\mathrm{X}})^{-1} \overline{\mathrm{X}}^\top \overline{\boldsymbol{y}} \\
s^2 &:= \overline{\boldsymbol{y}}^\top (\mathrm{I} - \overline{\mathrm{X}} \mathrm{C} \overline{\mathrm{X}}^\top - \overline{\mathrm{X}} \mathrm{C} \underline{\mathrm{X}}^\top \underline{\mathrm{X}} (\overline{\mathrm{X}}^\top \overline{\mathrm{X}})^{-1} \overline{\mathrm{X}}^\top) \overline{\boldsymbol{y}} = \overline{\boldsymbol{y}}^\top (\mathrm{I} - \overline{\mathrm{X}} (\overline{\mathrm{X}}^\top \overline{\mathrm{X}})^{-1} \overline{\mathrm{X}}^\top) \overline{\boldsymbol{y}}.
\end{aligned}$$

Finally, let $\mathrm{L}\mathrm{L}^\top = \mathrm{I} + \underline{\mathrm{X}} (\overline{\mathrm{X}}^\top \overline{\mathrm{X}})^{-1} \underline{\mathrm{X}}^\top$ be the Cholesky decomposition, and $\nu \leftarrow m - d - \dim(\underline{\boldsymbol{y}}) + 1$, $\boldsymbol{l} \leftarrow \sqrt{\nu}(\hat{\boldsymbol{w}} - \boldsymbol{u})/s$, the transformation $r = s/\sigma$, $\boldsymbol{z} = \mathrm{L}^{-1}(\hat{\boldsymbol{w}} - \boldsymbol{w})/\sigma$, again gives

$$f(\boldsymbol{z}, r) \propto \exp\left(-\tfrac{\|\boldsymbol{z}\|^2}{2} - \tfrac{r^2}{2} + (\nu - 1) \ln r\right) \mathbb{I}\{\sqrt{\nu}\, \mathrm{L}\boldsymbol{z} \geq r\boldsymbol{l}\}.$$

The analysis above can be easily generalized to the general Tobit model in which we have both left and right censoring. That is the response is modeled as

$$Y_i = \begin{cases} W_i, & \text{if } u_i < W_i < l_i \\ l_i, & \text{if } W_i > l_i \\ u_i, & \text{if } W_i < u_i \end{cases}, \qquad \boldsymbol{W} \sim \mathsf{N}(\mathrm{X}\boldsymbol{\beta}, \sigma^2 I).$$

However, we do not pursue this further in this thesis.

2.4.2.1 Women's wages dataset

The Women's Wages dataset [85] consists of $m = 753$ observations on the number of hours (the response y_i) married women spend in the labor force. The seven predictor variables $(x_1, \ldots, x_7)$ are:

1. kidslt6: number of children of age less than 6

2. kidsge6: number of children of age between 6 and 18
3. age: age of the married woman
4. educ: level of education
5. experience: number of years worked since age 18
6. nwifeinc: income that does not come from the wife
7. expersq: square of the number of years the married woman has worked since age 18

The acceptance probability in this rejection sampling is about 0.41 this suggests our rejection sampling scheme is efficient in practical inferences. (We do not report the optimal tilting parameter here because its dimension is too large.) The results are summarized in Figure 2.2 and Table 2.2.

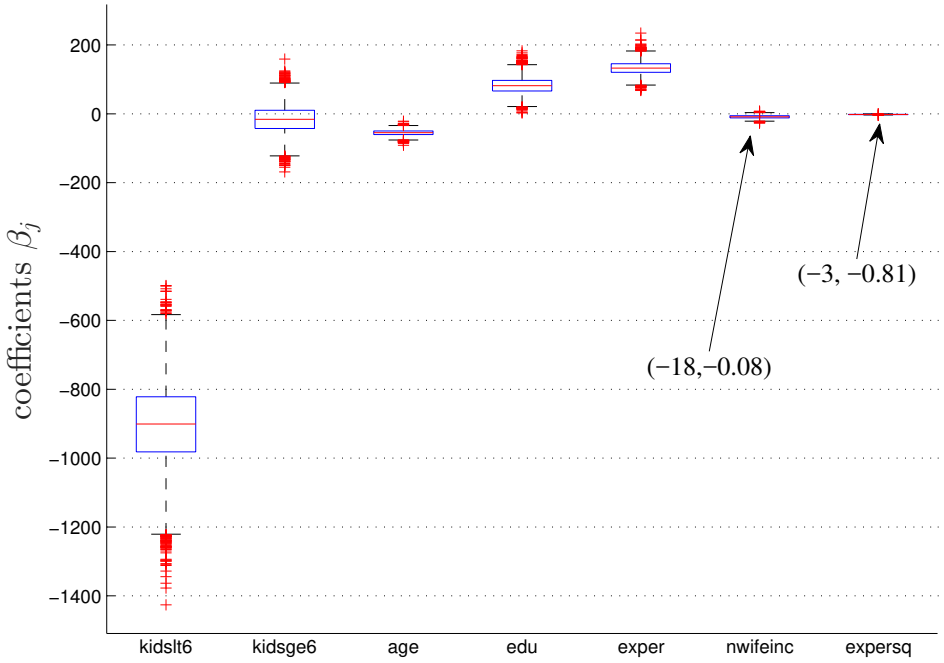

Figure 2.2: The empirical posterior distribution for the women's wage dataset derived from $n = 10^4$ independent exact draws.

We can see that the most important factor is the number of children of age less than 6 with a negative effect on the number of hours in the workforce. The experience in the work force is the second most important factor and it has a positive effect.

	mean	0.025-quantile	0.975-quantile	sample std.
β_0	9.59×10^2	2.29×10^1	1.84×10^3	4.64×10^2
kidslt6	-9.03×10^2	-1.15×10^3	-6.80×10^2	1.20×10^2
kidsge6	-1.62×10^1	-9.39×10^1	6.10×10^1	3.99×10^1
age	-5.50×10^1	-7.09×10^1	-4.00×10^1	7.86
educ	8.18×10^1	3.81×10^1	1.28×10^2	2.27×10^1
exper	1.33×10^2	9.80×10^1	1.70×10^2	1.85×10^1
nwifeinc	-8.92	-1.81×10^1	8.08×10^{-2}	4.65
expersq	-1.89	-3.00	-8.06×10^{-1}	5.59×10^{-1}

Table 2.2: Estimated mean, 0.95 credible interval and standard deviation of the posterior distribution from the exact simulation.

2.4.2.2 The affairs dataset

The affairs dataset consists of $m = 601$ independent observations. The response variable is a measure of time spent engaging in extramarital affairs, a variable that takes on a value for each individual of either zero or some positive number [28, 29].

The predictor variables include the following:

1. gender: a categorical variable (male being 1 and female being 0)
2. age: the age of the person
3. yearsmarried: the number of years being married
4. children: whether the couple has a children or not
5. religiousness: a rating on how religious the person is
6. education: the level of education
7. occupation: a measure of the socioeconomic status of the occupation
8. rating: how satisfied the person is with the current marriage

The acceptance probability in this rejection sampling is about 0.166 this again suggests our rejection sampling scheme is efficient in practical inferences. The results are summarized in figure 2.3 and table 2.3

	mean	approx. 0.025-quantile	approx. 0.975-quantile	sample std.
β_0	7.65	-1.14	1.64×10^1	4.40
gender	1.04	-1.28	3.68	1.25
age	-2.09×10^{-1}	-4.16×10^{-1}	-3.71×10^{-2}	9.60×10^{-2}
yearsmarried	5.65×10^{-1}	2.51×10^{-1}	9.42×10^{-1}	1.76×10^{-1}
children	1.17	-1.62	4.38	1.52
religiousness	-1.80	-2.93 o	-9.07×10^{-1}	5.14×10^{-1}
education	3.59×10^{-2}	-4.70×10^{-1}	5.61×10^{-1}	2.61×10^{-1}
occupation	2.14×10^{-1}	-5.12×10^{-1}	9.33×10^{-1}	3.66×10^{-1}
rating	-2.42	-3.73	-1.48	5.69×10^{-1}

Table 2.3: Estimated mean, 0.95 credible interval and standard deviation of the posterior distribution from the exact simulation.

The result shows that the most important factor is the level of satisfaction with the marriage. The posterior distribution of the rating is concentrated at the negative values. This means one is less likely to engage in extramarital affairs if one is satisfied with one's marriage.

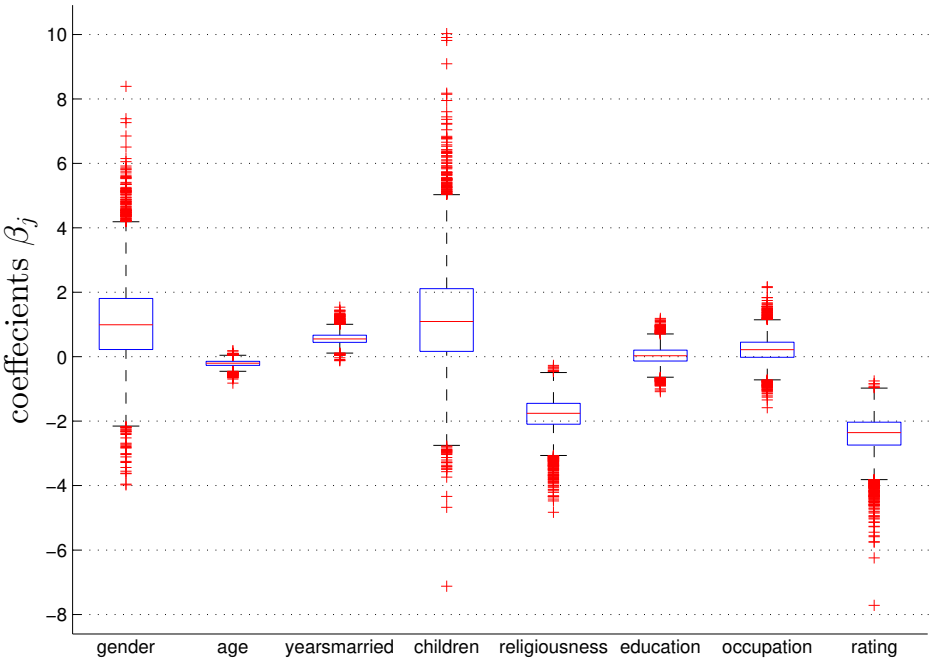

Figure 2.3: The empirical posterior distribution for the affairs dataset derived from $n = 10^4$ independent exact draws.

2.4.3 Bayesian splines

Consider noisy iid pairs $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ generated from some unknown positive function f over $x \in [0, h]$. Using $\{0, x_1, \ldots, x_n, h\}$ as knots and, the model for estimating f with a cubic smoothing splines is

$$y_i = \sum_{k=1}^{n+4} \beta_k s_k(x_i) + \epsilon_i, \qquad \epsilon_i \overset{\text{iid}}{\sim} \mathsf{N}(0, \sigma^2).$$

Here s_k is the k-th 4-th order B-spline basis for inner knots $\{x_1, \ldots, x_n\}$. (Recall that the 4-th order B-spline basis is a linear combination of cubic polynomials.) The goal is to estimate $\boldsymbol{\beta} := (\beta_1, \ldots, \beta_{n+4})^\top$, such that the model 'fits the data well' and is 'reasonably smooth'.

The classical treatment of the problem reduces down to solving the penalized regression optimization:

$$\underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i=1}^{n} \left(y_i - \sum_{k=1}^{n+4} \beta_k s_k(x_i) \right)^2 + \lambda \int_0^h \left(\sum_{k=1}^{n+4} \beta_k s_k''(x) \right)^2 dx,$$

where $\lambda > 0$ controls the smoothness of the estimated model.

Denoting $\boldsymbol{s}(x_i) = (s_1(x_i), s_2(x_i), \ldots, s_{n+4}(x_i))^\top$, its Bayesian analogue is therefore [89]:

$$\begin{aligned}
\pi(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\beta}, \sigma^2) &\propto \sigma^{-n} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \boldsymbol{\beta}^\top \boldsymbol{s}(x_i))^2\right) \\
\pi(\boldsymbol{\beta}|\sigma^2, \lambda) &\propto \lambda^{(n+4)/2}\sigma^{-(n+4)} \exp\left(-\frac{\lambda}{2\sigma^2}\int_0^h (\boldsymbol{\beta}^\top \boldsymbol{s}''(x))^2\, dx\right) \\
&\propto \lambda^{(n+4)/2}\sigma^{-(n+4)} \exp\left(-\frac{\lambda}{2\sigma^2}\boldsymbol{\beta}^\top \mathrm{K}\boldsymbol{\beta}\right) \\
p(\sigma^2) &\propto \sigma^{-2},
\end{aligned}$$

where K is a square matrix of size $n+4$ with entries

$$\mathrm{K}_{kl} = \int_0^h s_k''(x) s_l''(x)\, dx.$$

In practice, enforcing the positive definiteness of f reduces down to selecting points $\{0 \leq z_1 < z_2 < \ldots < z_m \leq h\}$ and imposing the constraint:

$$\boldsymbol{\beta}^\top \boldsymbol{s}(z_j) := \sum_{k=1}^{n+4} \beta_k s_k(z_j) > 0, \qquad j = 1, \ldots, m.$$

Consequently, Bayesian inference for this model requires one to sample from the posterior distribution:

$$\pi(\boldsymbol{\beta}, \sigma^2) \propto \pi(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\beta}, \sigma^2)\pi(\boldsymbol{\beta}|\sigma^2, \lambda)p(\sigma^2)$$

restricted to:

$$\boldsymbol{\beta}^\top \boldsymbol{s}(z_j) > 0, \qquad j = 1, \ldots, m.$$

By denoting the matrix $\mathrm{S} = \begin{pmatrix} \boldsymbol{s}(x_1)^\top \\ \vdots \\ \boldsymbol{s}(x_n)^\top \end{pmatrix}$ and completing the square, the posterior distribution reduces to:

$$\pi(\boldsymbol{\beta}, \sigma^2) \propto \exp\left(-\tfrac{s^2}{2\sigma^2} - (2n+6)\ln\sigma - \tfrac{(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^\top \mathrm{A}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})}{2\sigma^2}\right) \mathbb{I}\{\boldsymbol{\beta}^\top \boldsymbol{s}(z_j) > 0,\ j = 1, \ldots, m\}$$

where $\mathrm{A} = \mathrm{S}^\top\mathrm{S} + \lambda\mathrm{K}$, $\hat{\boldsymbol{\beta}} = \mathrm{A}^{-1}\mathrm{S}^\top\boldsymbol{y}$ and $s^2 = \boldsymbol{y}^\top\boldsymbol{y} - \boldsymbol{y}^\top\mathrm{S}\mathrm{A}^{-1}\mathrm{S}\boldsymbol{y}$. This again takes the form admissible to the sampling scheme since this posterior density takes the same as the constrained linear regression in (2.10). Figure 2.4 is an example of such smoothing problem.

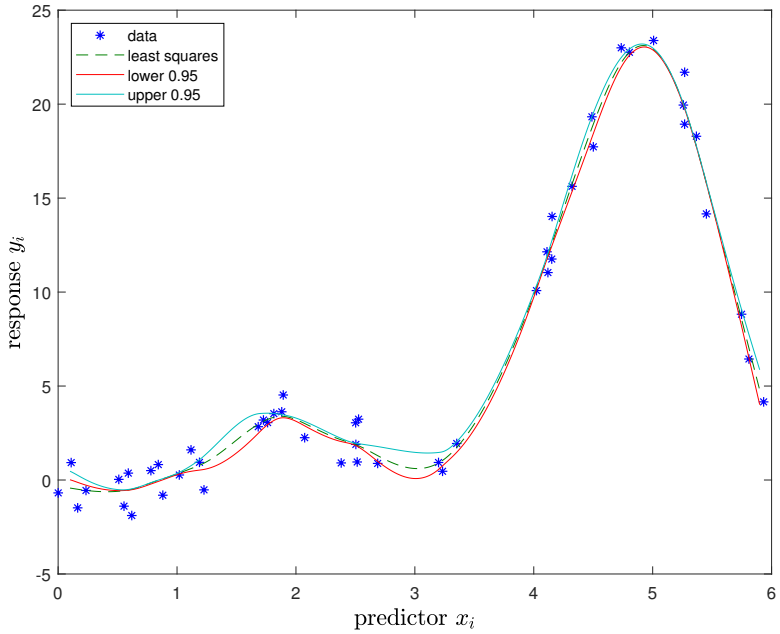

Figure 2.4: We consider the same example described in [89] in which they used a Gibbs sampler to sample the coefficients from the posterior distribution. There are 50 x_i uniformly distributed across $[0, 2\pi]$ and $y_i = x_i \sin^2(x_i) + \epsilon_i$, where $\epsilon_i \sim \mathsf{N}(0, 1)$. The dotted line and the bands are the mean and the empirical 95% function values obtained from sampling the posterior distribution exactly 1000 times.

2.5 Concluding remarks for this chapter

In this chapter we establish theoretical results concerning the asymptotic efficiency of the importance sampling estimator for ℓ derived in [12]. A byproduct of this proof is a novel multivariate extension to the Mill's ratio, currently known only for univariate student densities.

We also find novel applications of the exact samplers derived in [12]. In particular, we construct efficient rejection sampler for the posterior densities of the Bayesian constrained linear regression model, the Bayesian Tobit model and the Bayesian smoothing spline. Integrals with respect to these posterior distributions are intractable so practitioners call for Monte Carlo methods to estimate these integrals. The standard approach in the literature is to sample approximately from the posterior by MCMC samplers.

We have also tested these rejection samplers on real and synthetic datasets. Our simulation experience reveals that these samplers achieve valid posterior inferences and the probabilities of retaining samples are reasonably high.

In the next chapter we consider applying similar technique for simulating from the Bayesian Lasso linear regression model. The Lasso linear regression and its Bayesian analogue are popular extensions of the simple linear regression model. The standard

approaches for the posterior inference of the Bayesian Lasso linear regression model are approximate Markov chain samplings. It turns out that we can construct an optimally tilted sequential proposal for this posterior distribution too (though it will not be a simple standard family of densities), and in this manner we propose a novel efficient rejection sampler in the next chapter.

Chapter 3

Sequential proposal density via exponential tilting for the Bayesian Lasso linear regression posterior density

3.1 Introduction to this chapter

The Bayesian Lasso linear regression model can be summarized by the following hierarchical representation:

$$\begin{aligned} \boldsymbol{Y} \,|\, \boldsymbol{\beta}, \sigma &\sim \mathsf{N}(\mathrm{X}\boldsymbol{\beta}, \sigma^2 \mathrm{I}), \\ \beta_j &\overset{\text{iid}}{\sim} \mathsf{Lap}(\lambda/\sigma), \quad j = 1, \ldots, p, \\ \sigma &\sim p(\sigma). \end{aligned} \tag{3.1}$$

Here $\boldsymbol{Y} \in \mathbb{R}^n$ denotes the vector of some centered response variable, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)^\top \in \mathbb{R}^d$ is the random vector of coefficients, X is the $n \times d$ associated standardized design matrix, $p(\sigma)$ is some prior distribution for the noise parameter σ, and λ is called the 'Lasso parameter' which determines the level of shrinkage for β_j. Moreover, a $\mathsf{Lap}(\lambda)$ random variables are defined by the density

$$\frac{\lambda}{2} \exp(-\lambda |z|), \quad z \in \mathbb{R}.$$

It is common to choose $p(\sigma) \propto \sigma^{-2}$, the non-informative scale invariant prior [62], and in this way, inference for the Bayesian Lasso linear regression model ultimately requires computing integrals with respect to posterior densities of the form

$$\pi(\boldsymbol{\beta}, \sigma \,|\, \boldsymbol{y}, \lambda) = \frac{\sigma^{-2} \times (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \|\boldsymbol{y} - \mathrm{X}\boldsymbol{\beta}\|_2^2\right) \times \frac{\lambda^d}{(2\sigma)^d} \exp\left(-\frac{\lambda}{\sigma} \|\boldsymbol{\beta}\|_1\right)}{\ell(\lambda \,|\, \boldsymbol{y})}. \tag{3.2}$$

Here ℓ is marginal likelihood. (Note that, in this thesis, we consider $\boldsymbol{Y}$ being centralized during data-preprocessing so that we ignore the intercept term in the regression. One can alternatively assign an non-informative prior for the intercept and integrate it out, see [91] for detail on this matter.) For the simplicity of notation, we shall now drop the condition on $\boldsymbol{y}$ for the posterior density, that is we will write $\pi(\cdot)$ and $\ell(\cdot)$ instead of $\pi(\cdot \,|\, \boldsymbol{y})$, and

$\ell(\cdot \mid \boldsymbol{y})$. Since integrals with respect to this posterior density are intractable, in practice, one simulates draws from π and approximate the integrals with some average.

In this chapter we construct novel exact sampling algorithms to sample from the posterior density of the Bayesian Lasso. We begin by considering a simplified model in which we take σ as a fixed and known. We show that this simplified posterior density exhibits a simple rejection sampling algorithm by constructing an exponentially tilted normal proposal density. The tilting parameter here is optimally chosen by solving a quadratic programming problem. Moreover, it turns out this exponentially tilted normal density renders an accurate importance sampling estimator for the marginal likelihood ℓ.

We then consider the standard case where σ is random. In this case, the normal proposal density is no longer efficient. For this reason, we construct a more complex sequential proposal density, again based on the "separation of variable" technique in Chapter 2. This later construction is published in [10].

The rest of this section provides a discussion on the origin and the motivation for the Bayesian Lasso linear regression model. A brief recount of notable works in the literature is also presented. We find that MCMC is the standard approach to the problem as oppose to our rejection sampler.

The classical ordinary least squares (OLS) estimator, which historically traces back to Legendre and Gauss [103], considers solving the optimization

$$\hat{\boldsymbol{\beta}}_{\text{OLS}} = \text{argmin}_{\boldsymbol{\beta}} \|\boldsymbol{Y} - \text{X}\boldsymbol{\beta}\|_2^2.$$

This approach quickly loses appeal when one wishes to achieve a lower mean squared error (MSE) or to perform variable selection. Studies have shown that introducing regularization penalties can achieve lower MSE. For example, for some $\lambda > 0$, the ridge regression [59] which concerns

$$\hat{\boldsymbol{\beta}}_{\text{ridge}} = \text{argmin}_{\boldsymbol{\beta}} \|\boldsymbol{Y} - \text{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_2,$$

or the Lasso regression [107], which concerns

$$\hat{\boldsymbol{\beta}}_{\text{Lasso}} = \text{argmin}_{\boldsymbol{\beta}} \|\boldsymbol{Y} - \text{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1,$$

or in general, the bridge regression [34] that concerns

$$\hat{\boldsymbol{\beta}}_{\text{bridge}} = \text{argmin}_{\boldsymbol{\beta}} \|\boldsymbol{Y} - \text{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_p.$$

Here $\|\cdot\|_p$ denotes the p-norm (or the p-pseudo-norm for $0 \leq p < 1$). Note that the bridge regression naturally encompasses 'subset selection' when $p = 0$, the Lasso regression when $p = 1$ and the ridge regression when $p = 2$.

The Lasso is of particular interest to practitioners because $\|\cdot\|_1$ corresponds to the smallest p out of all $\|\cdot\|_p$ such that the programming remains a convex problem, and unlike the ridge regression, it can set some components of $\boldsymbol{\beta}$ to 0. Consequently, variable selection naturally embeds in the Lasso regression [108]. This is advantageous over earlier variable selection methods that involve repeated model fitting and calculation of statistics such as the Akaike Information criterion [1] and the Bayesian Information criterion [102].

A Bayesian interpretation of the Lasso linear regression is also given in [107]. Here $\boldsymbol{\beta}$ is assigned an independent $\mathsf{Lap}(\lambda)$ prior. The first systematic study on this Bayesian analogue is given in [91], however they argue that a $\mathsf{Lap}(\lambda/\sigma)$ prior should be imposed to achieve unimodality of the posterior distribution so as to ensure the proposed Gibbs sampler converges. This model can is summarized by hierarchy (3.1).

Bayesian inference for this model requires one to evaluate intractable integrals with respect to the posterior density (3.2). Numerical techniques such as importance sampling and MCMC sampling are almost the default methods in this context. The Gibbs sampler proposed in [91] exploits the normal scale mixture representation of the Laplace density as follows:

$$\frac{\lambda}{2\sigma}e^{-\lambda|z|/\sigma} = \int_0^\infty \frac{1}{\sqrt{2\pi s}}e^{-z^2/(2s)}\frac{\lambda^2}{2\sigma}e^{-\lambda^2 s/(2\sigma^2)}\,ds.$$

Consequently, one can consider

$$\pi(\boldsymbol{\beta},\boldsymbol{s},\sigma) = \frac{\sigma^{-2}\times(2\pi\sigma^2)^{-n/2}\exp\left(-\frac{1}{2\sigma^2}\|\boldsymbol{y}-\mathrm{X}\boldsymbol{\beta}\|_2^2\right)\times\prod_{j=1}^d \frac{1}{\sqrt{2\pi s_j}}e^{-\beta_j^2/(2s_j)}\frac{\lambda^2}{2\sigma}e^{-\lambda^2 s_j/(2\sigma^2)}}{\ell(\lambda)}$$

so that $\pi(\boldsymbol{\beta},\sigma) = \int_{\mathbb{R}_+^d}\pi(\boldsymbol{\beta},\boldsymbol{s},\sigma)\,d\boldsymbol{s}$ is recovered. Defining $\tau_j = 1/s_j$ for each j, and denote $\boldsymbol{\tau} = (\tau_1,\ldots,\tau_d)^\top$, one may rewrite the hierarchy as follows.

$$\begin{aligned}
\boldsymbol{Y}\,|\,\boldsymbol{\beta},\sigma &\sim \mathsf{N}(\mathrm{X}\boldsymbol{\beta},\sigma^2\mathrm{I}),\\
\boldsymbol{\beta}\,|\,\sigma,\boldsymbol{s} &\sim \mathsf{N}(\boldsymbol{0},\sigma^2\mathrm{diag}(\boldsymbol{\tau})),\\
p(\boldsymbol{s}\,|\,\sigma) &= \prod_{j=1}^d \frac{\lambda^2}{2\sigma}e^{-\lambda^2 s_j/(2\sigma^2)},\\
p(\sigma) &\propto \sigma^{-2}.
\end{aligned}$$

The idea is then to construct a (block) Gibbs sampler on the augmented state space, $\mathbb{R}^d\times\mathbb{R}_+^d\times\mathbb{R}_+$ for $(\boldsymbol{\beta},\boldsymbol{\tau},\sigma)$, that cycles through the following conditional distributions.

$$\begin{aligned}
\beta\,|\,\boldsymbol{\tau},\sigma &\sim \mathsf{N}(\mathrm{A}\mathrm{X}^\top\boldsymbol{y},\sigma^2\mathrm{A}),\\
\sigma^2\,|\,\boldsymbol{\beta},\boldsymbol{\tau} &\sim \mathsf{InvGamma}(n/2+d/2,\|\boldsymbol{y}-\mathrm{X}\boldsymbol{\beta}\|_2^2/2+\boldsymbol{\beta}^\top\mathrm{diag}(\boldsymbol{\tau})\boldsymbol{\beta}/2),\\
\tau_j := \frac{1}{s_j}\,|\,\boldsymbol{\beta},\sigma &\sim \mathsf{Wald}(\lambda^2,\sigma\lambda/|\beta_j|),\quad \text{independently for all } j.
\end{aligned}$$

Here $\mathrm{A}^{-1} = \mathrm{X}^\top\mathrm{X} + \mathrm{diag}(\boldsymbol{\tau})$ is symmetric and invertible, and the density of a $\mathsf{Wald}(\lambda, \mu)$[19] distribution is

$$\sqrt{\frac{\lambda}{2\pi}} x^{-3/2} \exp\left(\frac{\lambda(x-\mu)^2}{2\mu^2 x}\right), \quad x \in \mathbb{R}_+.$$

Since then, different approaches have been proposed. For example, [79] uses the relation

$$\frac{\lambda}{2\sigma} e^{-\frac{\lambda}{\sigma}|z|} = \frac{\lambda}{2\sigma} \int_0^\infty \lambda e^{-\lambda u} \mathbb{I}\{u > |z|/\sigma\}\, du = \int_0^\infty \frac{\lambda^2}{2\Gamma(2)\sigma u} u e^{-\lambda u}\, du$$

so that

$$\pi(\boldsymbol{\beta}, \boldsymbol{u}, \sigma) = \frac{\sigma^{-2} \times (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\|\boldsymbol{y} - \mathrm{X}\boldsymbol{\beta}\|_2^2\right) \times \prod_{j=1}^d \frac{\lambda^2}{2\Gamma(2)\sigma u_j} u_j e^{-\lambda u_j} \mathbb{I}\{u_j \geq |\beta_j|/\sigma\}}{\ell(\lambda)}$$

also satisfies $\pi(\boldsymbol{\beta}, \sigma) = \int_{\mathbb{R}_+^d} \pi(\boldsymbol{\beta}, \boldsymbol{u}, \sigma)\, d\boldsymbol{u}$. Hence one can study the Bayesian Lasso with the following equivalent hierarchy.

$$\begin{aligned}
\boldsymbol{Y} \,|\, \boldsymbol{\beta}, \sigma &\sim \mathsf{N}(\mathrm{X}\boldsymbol{\beta}, \sigma^2 \mathrm{I}) \\
\beta_j \,|\, \boldsymbol{u}, \sigma &\sim \mathsf{Unif}(-\sigma u_j, \sigma u_j), \quad \text{for } j = 1, \ldots, d \\
u_j &\overset{\text{iid}}{\sim} \mathsf{Gamma}(2, \lambda), \quad \text{for } j = 1, \ldots, d \\
p(\sigma) &\propto \sigma^{-2}.
\end{aligned}$$

A natural Gibbs sampler for $(\boldsymbol{\beta}, \boldsymbol{u}, \sigma)$ then cycles between the following full conditional densities.

$$\begin{aligned}
\boldsymbol{\beta} \,|\, \boldsymbol{u}, \sigma &\sim \mathsf{N}(\hat{\boldsymbol{\beta}}_{\text{OLS}}, \sigma^2(\mathrm{X}^\top\mathrm{X})^{-1}), \quad \text{conditional on } |\beta_j| < \sigma u_j \ \forall j \\
u_j \,|\, \boldsymbol{\beta}, \sigma &\sim \mathsf{Exp}(\lambda), \quad \text{conditional on } u_j > |\beta_j|/\sigma \ \forall j \\
\sigma^2 \,|\, \boldsymbol{\beta}, \boldsymbol{u} &\sim \mathsf{InvGamma}(n/2 + d/2, \|\boldsymbol{y} - \mathrm{X}\boldsymbol{\beta}\|_2^2/2), \quad \text{conditional on } \sigma^2 > \max_j \beta_j^2/u_j^2.
\end{aligned}$$

Both [91] and [79] construct Markov chain on some augmented state space. A noteworthy work along this framework is [92] which also injects the posterior density of the Bayesian bridge into some augmented state space, and proposes a Gibbs sampler from there. On the other hand [50] proposes a direct approach, which does not introduce auxiliary variables. This is done by leveraging the observation that $|\beta_j| = \beta_+ - \beta_-$, so that the density of $\boldsymbol{\beta} \,|\, \sigma$ is some mixture of normal densities, and a Gibbs sampler follows. (Here, $\cdot_+ = \max\{0, \cdot\}$ and $\cdot_- = \min\{0, \cdot\}$.)

Before we move on to the content of this chapter, we draw readers' attention to the matter of choosing λ. Here, one can take a fully Bayesian approach where λ is assigned a prior and is treated as random. Alternatively one can take the empirical Bayes approach

and find the λ for which ℓ is maximized. In this thesis, we take the latter approach in all our numerical examples. This optimization problem can be solved numerically using the approximate EM algorithm proposed in [15] whenever there is a Gibbs sampler targeting the posterior density.

3.2 Bayesian Lasso, with σ fixed

In this section we shall consider the case where σ is known and is not random. In the numerical experiment, the maximum likelihood estimator $\hat{\sigma}$ is substituted for σ.

Observe that the Pythagorean Identity

$$\|\boldsymbol{y} - \mathrm{X}\boldsymbol{\beta}\|_2^2 = \|\boldsymbol{y} - \mathrm{X}\hat{\boldsymbol{\beta}}\|_2^2 + \|\mathrm{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\|_2^2$$

yields

$$\pi(\boldsymbol{\beta}) \propto \exp\left(-\frac{1}{2\sigma^2}\|\mathrm{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\|_2^2 - \frac{\lambda}{\sigma}\|\boldsymbol{\beta}\|_1\right).$$

Denoting $\Sigma = \sigma^2(\mathrm{X}^\top\mathrm{X})^{-1}$ yields the target distribution:

$$\pi(\boldsymbol{\beta}) = \frac{\phi_\Sigma(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\frac{\lambda^p}{2^p}\exp(-\frac{\lambda}{\hat{\sigma}}\|\boldsymbol{\beta}\|_1)}{\ell(\lambda)}, \tag{3.3}$$

where $\phi_\Sigma(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})$ is the pdf of a d-dimensional normal random variable with covariance matrix Σ, and mean vector $\hat{\boldsymbol{\beta}}$.

Perhaps a natural choice of the proposal density for the rejection sampler is $\phi_\Sigma(\cdot - \hat{\boldsymbol{\beta}})$. In this case, we have the following algorithm.

Algorithm 2 : Naive rejection sampler for $\pi(\boldsymbol{\beta})$.

Require: Supremum of likelihood ratio $c = \sup_{\boldsymbol{\beta}} \ell\pi(\boldsymbol{\beta})/\phi_\Sigma(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})$.

Simulate $U \sim \mathsf{U}(0,1)$ and $\boldsymbol{\beta} \sim \phi_\Sigma(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})$, independently.

while $cU > \ell\pi(\boldsymbol{\beta})/\phi_\Sigma(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})$ **do**

Simulate $U \sim \mathsf{U}(0,1)$ and $\boldsymbol{\beta} \sim \phi_\Sigma(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})$, independently.

return $\boldsymbol{\beta}$, a perfect draw from the posterior.

This rejection scheme will only be useful if the probability of acceptance

$$\mathbb{P}\Big[cU \leq \ell\tfrac{\pi(\boldsymbol{\beta})}{\phi_\Sigma(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})}\Big] = \ell/c$$

is high. For this particular proposal, it is clear that

$$c = \sup_{\boldsymbol{\beta}} \frac{\ell\pi(\boldsymbol{\beta})}{\phi_\Sigma(\boldsymbol{\beta} - \boldsymbol{\beta})} = \sup_{\boldsymbol{\beta}} \frac{\lambda^p}{2^p}\exp(-\frac{\lambda}{\sigma}\|\boldsymbol{\beta}\|_1) = \frac{\lambda^p}{2^p}.$$

Moreover, the marginal likelihood ℓ can be estimated with importance sampling. Hence, we can estimate the acceptance probability, ℓ/c, of Algorithm 2. Take for example, the diabetes data set [27]. The empirical Bayes approach [15] estimates $\lambda = 0.237$ and plugging this value into the expression, along with the least squares estimator for σ, we get that a 95% asymptotic confidence interval for the acceptance probability is $(7.53 \pm 0.053) \times 10^{-6}$. Since the acceptance probability is too small, we next propose a more efficient alternative.

Suppose that, instead of $\phi_\Sigma(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})$ as a proposal density, we use $\phi_\Sigma(\boldsymbol{\beta} - \boldsymbol{\mu})$ where the parameter $\boldsymbol{\mu} \in \mathbb{R}^p$ is yet to be specified. In a similar to manner to Chapter 2, we define

$$\begin{aligned}\psi(\boldsymbol{\beta}; \boldsymbol{\mu}) := \ln \frac{\ell\, \pi(\boldsymbol{\beta}|\boldsymbol{y}, \lambda)}{\phi_\Sigma(\boldsymbol{\beta} - \boldsymbol{\mu})} &= -\frac{1}{2}\|\mathrm{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\|_2^2 - \frac{\lambda}{\sigma}\|\boldsymbol{\beta}\|_1 + \frac{1}{2}\|\mathrm{X}(\boldsymbol{\beta} - \boldsymbol{\mu})\|_2^2 + p\ln(\lambda/2) \\ &= \frac{\boldsymbol{\mu}^\top \Sigma^{-1}\boldsymbol{\mu} - \hat{\boldsymbol{\beta}}^\top \Sigma^{-1}\hat{\boldsymbol{\beta}}}{2} + \boldsymbol{\beta}^\top \Sigma^{-1}(\hat{\boldsymbol{\beta}} - \boldsymbol{\mu}) - \frac{\lambda}{\sigma}\|\boldsymbol{\beta}\|_1 + p\ln(\lambda/2),\end{aligned}$$

and solve the program $\inf_{\boldsymbol{\mu}} \sup_{\boldsymbol{\beta}} \psi(\boldsymbol{\beta}; \boldsymbol{\mu})$. Noting that we can exchange the order:

$$\inf_{\boldsymbol{\mu}} \sup_{\boldsymbol{\beta}} \psi(\boldsymbol{\beta}; \boldsymbol{\mu}) = \sup_{\boldsymbol{\beta}} \inf_{\boldsymbol{\mu}} \psi(\boldsymbol{\beta}; \boldsymbol{\mu}).$$

Since $\psi(\boldsymbol{\beta}; \boldsymbol{\mu})$ is a quadratic function in $\boldsymbol{\mu}$, one can easily verify via standard calculations that $\boldsymbol{\mu} = \boldsymbol{\beta}$. Eliminating $\boldsymbol{\mu}$ from the objective function, yields

$$\sup_{\boldsymbol{\beta}} \psi(\boldsymbol{\beta}, \boldsymbol{\beta}) = -\min_{\boldsymbol{\beta}} \frac{1}{2}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^\top \Sigma^{-1}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + \frac{\lambda}{\sigma}\|\boldsymbol{\beta}\|_1 + p\ln(\lambda/2). \tag{3.4}$$

Next, recall that

$$\|\boldsymbol{\beta}\|_1 = \max_{\boldsymbol{a}:\|\boldsymbol{a}\|_\infty = 1} \boldsymbol{a}^\top \boldsymbol{\beta}.$$

With this, we can transform (3.4) into a constrained quadratic programming in the following way.

$$\begin{aligned}\min_{\boldsymbol{\beta}} \frac{1}{2}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^\top \Sigma^{-1}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + \lambda\|\boldsymbol{\beta}\|_1 &= \min_{\boldsymbol{\beta}} \max_{\boldsymbol{a}:\|\boldsymbol{a}\|_\infty = 1} \frac{1}{2}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^\top \Sigma^{-1}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + \lambda \boldsymbol{a}^\top \boldsymbol{\beta} \\ &= \min_{\boldsymbol{\beta}} \max_{-\mathbf{1} \le \boldsymbol{a} \le \mathbf{1}} \frac{1}{2}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^\top \Sigma^{-1}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + \lambda \boldsymbol{a}^\top \boldsymbol{\beta} \\ &= \max_{-\mathbf{1} \le \boldsymbol{a} \le \mathbf{1}} \min_{\boldsymbol{\beta}} \frac{1}{2}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^\top \Sigma^{-1}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + \lambda \boldsymbol{a}^\top \boldsymbol{\beta} \\ &= \max_{-\mathbf{1} \le \boldsymbol{a} \le \mathbf{1}} \lambda \boldsymbol{a}^\top \hat{\boldsymbol{\beta}} - \frac{\lambda^2}{2}\boldsymbol{a}^\top \Sigma \boldsymbol{a}.\end{aligned} \tag{3.5}$$

The above calculation leverages the following. Firstly, the maximization over $\boldsymbol{a}$ with the constraint $\|\boldsymbol{a}\|_\infty = 1$ is equivalent to maximization over $\boldsymbol{a}$ with the box constraint

$-\mathbf{1} \leq \boldsymbol{a} \leq \mathbf{1}$, provided there is at least one active constraint. We assume that there is at least one active constraint, since otherwise we end up with an uninteresting Lasso solution of $\boldsymbol{\beta} = \mathbf{0}$. Further, we use $\min_{\boldsymbol{\beta}} \max_{-\mathbf{1} \leq \boldsymbol{a} \leq \mathbf{1}} \equiv \max_{-\mathbf{1} \leq \boldsymbol{a} \leq \mathbf{1}} \min_{\boldsymbol{\beta}}$, so that substituting the minimum over $\boldsymbol{\beta}$, which is $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}} - \lambda \Sigma \boldsymbol{a}$, we arrive at the last equality. The last expression can be solved by constrained quadratic programming packages. Plugging $\boldsymbol{a}^*$, the solution to the last expression, gives $\boldsymbol{\mu}^* = \boldsymbol{\beta}^* = \hat{\boldsymbol{\beta}} - \lambda \Sigma \boldsymbol{a}^*$.

Algorithm 3 : Rejection sampler for $\pi(\boldsymbol{\beta})$ with minimax tilting.

Require: Supremum of likelihood ratio $c = \inf_{\boldsymbol{\mu}} \sup_{\boldsymbol{\beta}} \ell\pi(\boldsymbol{\beta})/\phi_\Sigma(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})$., and μ^*, the solution to the minimax program

Simulate $U \sim \mathsf{U}(0,1)$ and $\boldsymbol{\beta} \sim \phi_\Sigma(\boldsymbol{\beta} - \boldsymbol{\mu}^*)$, independently.

while $cU > \ell\pi(\boldsymbol{\beta})/\phi_\Sigma(\boldsymbol{\beta} - \boldsymbol{\mu}^*)$ **do**

Simulate $U \sim \mathsf{U}(0,1)$ and $\boldsymbol{\beta} \sim \phi_\Sigma(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})$, independently.

return $\boldsymbol{\beta}$, a perfect draw from the posterior.

3.2.1 Estimating the marginal likelihood

Observe that the marginal likelihood, as a function of the Lasso parameter λ, is given by

$$\ell(\lambda) = \int_{\mathbb{R}^p} \phi_\Sigma(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \frac{\lambda^p}{2^p} \exp\left(-\|\boldsymbol{\beta}\|_1\right) d\boldsymbol{\beta}.$$

Such an integral clearly cannot be computed analytically so one needs to resort to numerical methods. In the context of importance sampling estimation, perhaps it is natural to consider a $\mathsf{N}(\hat{\boldsymbol{\beta}}, \Sigma)$ proposal. It follows that for any $\lambda > 0$,

$$\hat{\ell}_{\text{naive}}(\lambda) = \frac{\lambda^p}{2^p} \exp\left(-\lambda \|\boldsymbol{Z}\|_1\right), \qquad \boldsymbol{Z} \sim \mathsf{N}(\hat{\boldsymbol{\beta}}, \Sigma),$$

is an unbiased estimator, that is $\frac{\lambda^p}{2^p} \mathbb{E} \exp(-\lambda \|\boldsymbol{Z}\|_1) = \ell(\lambda)$. However, it turns out that $\mathsf{N}(\mu^*, \Sigma)$ is a more efficient importance sampling proposal for $\ell(\lambda)$. That is, we propose the estimator

$$\hat{\ell}(\lambda) = \exp(\psi(\boldsymbol{Z}; \boldsymbol{\mu}^*)), \qquad \boldsymbol{Z} \sim \mathsf{N}(\boldsymbol{\mu}^*, \Sigma). \tag{3.6}$$

(Notice that $\mathbb{E} \exp(\psi(\boldsymbol{Z}; \boldsymbol{\mu}^*)) = \ell(\lambda)$ as well.)

One reason is that $\hat{\ell}(\lambda)$ is more efficient is the following. Note that controlling the variance of (3.6), an unbiased estimator, is the same as controlling its second moment, and the following calculation shows that $\boldsymbol{\mu} = \boldsymbol{\beta}^*$ also controls the second moment. Set

$\boldsymbol{\nu} = \boldsymbol{\mu} - \hat{\boldsymbol{\beta}}$, then:

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\mu}} \exp(2\psi(\boldsymbol{Z};\boldsymbol{\mu})) &= (\lambda/2)^{2p} \mathbb{E}_{\hat{\boldsymbol{\beta}}} \exp(\tfrac{\boldsymbol{\mu}^\top \Sigma^{-1} \boldsymbol{\mu} - \hat{\boldsymbol{\beta}}^\top \Sigma^{-1} \hat{\boldsymbol{\beta}}}{2} + \boldsymbol{Z}^\top \Sigma^{-1}(\hat{\boldsymbol{\beta}} - \boldsymbol{\mu})) \exp(-2\lambda \|\boldsymbol{Z}\|_1) \\
&= (\lambda/2)^{2p} \exp(\boldsymbol{\nu}^\top \Sigma^{-1} \boldsymbol{\nu}) \mathbb{E}_{\hat{\boldsymbol{\beta}} - \boldsymbol{\nu}} \exp(-2\lambda \|\boldsymbol{Z}\|_1) \\
&\leq (\lambda/2)^{2p} \exp(\boldsymbol{\nu}^\top \Sigma^{-1} \boldsymbol{\nu}) \min_{\boldsymbol{a}:\|\boldsymbol{a}\|_\infty = 1} \mathbb{E}_{\hat{\boldsymbol{\beta}} - \boldsymbol{\nu}} \exp(-2\lambda \boldsymbol{a}^\top \boldsymbol{Z}) \\
&\leq (\lambda/2)^{2p} \exp(\boldsymbol{\nu}^\top \Sigma^{-1} \boldsymbol{\nu}) \min_{\boldsymbol{a}:\|\boldsymbol{a}\|_\infty = 1} \exp(-2\lambda \boldsymbol{a}^\top (\hat{\boldsymbol{\beta}} - \boldsymbol{\nu}) + 2\lambda^2 \boldsymbol{a}^\top \Sigma \boldsymbol{a}).
\end{aligned}$$

Minimization over ν requires us to solve:

$$\min_{\boldsymbol{a}:\|\boldsymbol{a}\|_\infty = 1} \min_{\boldsymbol{\nu}} \left\{ \boldsymbol{\nu}^\top \Sigma^{-1} \boldsymbol{\nu} - 2\lambda \boldsymbol{a}^\top (\hat{\boldsymbol{\beta}} - \boldsymbol{\nu}) + 2\lambda^2 \boldsymbol{a}^\top \Sigma \boldsymbol{a} \right\}.$$

Eliminating

$$\boldsymbol{\nu} = -\lambda \Sigma \boldsymbol{a},$$

the last is equivalent to

$$\min_{\boldsymbol{a}:\|\boldsymbol{a}\|_\infty = 1} \lambda^2 \boldsymbol{a}^\top \Sigma \boldsymbol{a} - 2\lambda \boldsymbol{a}^\top \hat{\boldsymbol{\beta}},$$

which is equivalent to the quadratic programming problem (3.5), up to some scaling constant. Thus, the tilted proposal density $\mathsf{N}(\boldsymbol{\mu}^*, \Sigma)$ is useful both in the acceptance rejection Algorithm 3 and also in the estimation of the marginal likelihood $\ell(\lambda)$.

Computing the marginal likelihood has practical importance in Bayesian inference such as model comparison with the Bayes factor. Further, suppose that one takes the empirical Bayes approach [15] to choose the Lasso parameter λ. To do so, one is required to find the value of λ for which the marginal likelihood is maximized. This can be done with pilot Gibbs sampling run. Our approach here provides an alternative to Gibbs sampling. An advantage of our importance sampler is that we have a analytic upper bound on the variance:

$$(\lambda/2)^{2p} \exp(\boldsymbol{\nu}^\top \Sigma^{-1} \boldsymbol{\nu}) \min_{\boldsymbol{a}:\|\boldsymbol{a}\|_\infty = 1} \exp(-2\lambda \boldsymbol{a}^\top (\hat{\boldsymbol{\beta}} - \boldsymbol{\nu}) + 2\lambda^2 \boldsymbol{a}^\top \Sigma \boldsymbol{a}).$$

Further, the importance sampler is based on iid replicates, and so standard iid analysis applies here.

3.2.2 Numerical experiments

We take the "diabetes dataset" on $n = 442$ patients from [27]. For each patient, we have a record of $p = 10$ predictor variables (age, sex, body mass index, blood pressure, and six blood serum measurements), and a response variable, which measures the severity of nascent diabetes. We then wish to determine which of the predictors are most relevant to the response variable using this simplified Lasso model by simulating from the posterior.

Here $\boldsymbol{y} \in \mathbb{R}^n$ is the vector of centralized response variables, and X is the standardized matrix of predictors. The Lasso parameter λ is chosen to be $\lambda = 0.237$, which is the value that maximizes the marginal likelihood; see [91], and σ is taken to be the maximum likelihood estimator $\hat{\sigma}$. The result of the sampler is presented below.

	Gibbs sampler [91]		Exact sampler via minimax tiliting	
	Median	95% credible interval	Median	95% credible interval
Age	-3.296	(-110.99, 102.54)	-3.4176	(-111.28,102.87)
sex	-213.90	(-333.41, -95.448)	-213.76	(-333.28,-94.658)
bmi	523.56	(393.45, 653.59)	523.72	(394.55,654.14)
map	307.81	(179.99,434.24)	307.30	(180.33,435.12)
tc	-171.95	(-576.12,125.90)	-172.05	(-578.95,125.82)
ldl	-2.7453	(-273.66, 332.13)	-2.2558	(-275.06,335.22)
hdl	-152.24	(-382.22,69.698)	-151.59	(-381.67,70.233)
tch	92.174	(-126.57, 351.57)	92.480	(-127.4, 349.79)
ltg	521.62	(333.88,725.86)	521.8	(333.04,728.47)
glu	63.007	(-50.496, 188.16)	63.032	(-51.175,189.20)

For the diabetes dataset we have estimated the marginal likelihood for different values of the Lasso parameter λ, as shown in Figure 3.1. For ease of comparison, we plotted λ against the estimated log-likelihood. We see that the maximum indeed occurs near $\lambda = 0.237$ as reported by [91]. Moreover, it is clear from the figure that importance sampling gives more accurate point-wise estimates.

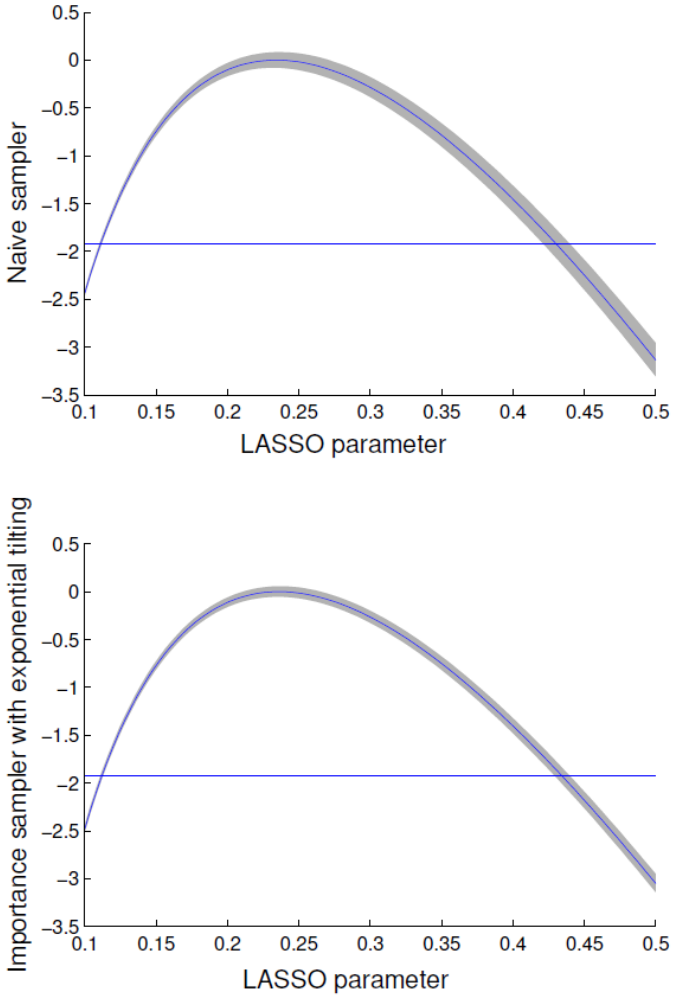

Figure 3.1: Comparing the estimated logarithm of the marginal likelihood ratio $\ln(\ell(\lambda))$ with point-wise approximate confidence interval using 10^4 Monte Carlo samples. The horizontal line is $-\chi^2_{1,95}$, the 0.95 quantile of the χ^2 distribution with 1 degree of freedom.

3.3 Bayesian Lasso

We now construct a more sophisticated exact sampler for the general case of the Bayesian Lasso in which we sample $\sigma \in \mathbb{R}_+$ along with $\boldsymbol{\beta}$ according to the density. Taking the uninformative scale-invariant prior $\pi(\sigma) \propto \sigma^{-2}$, the posterior density takes the form

$$\pi(\boldsymbol{\beta}, \sigma) \propto \frac{\lambda^p}{2^p} \sigma^{-p+2} \exp\left(-\frac{1}{2\sigma^2} \|\boldsymbol{y} - \mathrm{X}\boldsymbol{\beta}\|_2^2 - \frac{\lambda}{\sigma} \|\boldsymbol{\beta}\|_1 \right).$$

The idea here is again similar to Chapter 2 in which we shall introduce some coordinate transformations that motivates a sequential proposal density. The sequentially proposal is then optimally tilted so that rejection sampling remains practical for reasonably large

dimensions. We begin by introducing some probability distributions, which will be part of the sequential densities.

For $z \in \mathbb{R}$, the exponentially tilted Laplace distribution is defined by

$$\mathsf{Lap}_\mu(z) = \exp\left(-|z| + \mu z + \ln(1-\mu^2)\right), \qquad |\mu| < 1.$$

For $r > 0$, the chi distribution is defined by

$$\mathsf{chi}_\nu(r) = \frac{\exp(-r^2/2 + (\nu-1)\ln r)}{2^{\nu/2-1}\Gamma(\nu/2)}, \qquad r > 0.$$

Finally, the normal-Laplace density is

$$\mathsf{nl}(z;\lambda,\alpha) = \phi(z;0,1)\exp\left(-\lambda|z-\alpha| - \xi(\lambda,\alpha)\right),$$

where

$$\xi(\lambda,\alpha) = -\frac{\alpha^2 + \ln(2\pi)}{2} + \ln\left[\frac{\bar{\Phi}(\lambda+\alpha)}{\phi(\lambda+\alpha)} + \frac{\bar{\Phi}(\lambda-\alpha)}{\phi(\lambda-\alpha)}\right]$$

is the normalizing constant, and $\bar{\Phi}$ is the complementary cdf for a standard normal distribution. Note that it is not difficult to simulate from each of them. Standard statistical software packages can generate chi_ν random variables while Lap_μ and $\mathsf{nl}(\cdot;\lambda,\alpha)$ can be written as a mixture of exponential and a mixture of normal random variables respectively.

Next, observe that we can factorize π in the following way

$$\pi(\boldsymbol{\beta},\sigma) \propto \phi_{\sigma^2}(\boldsymbol{y} - \mathrm{X}\boldsymbol{\beta})\lambda^d \exp\left(-(2+d)\ln\sigma - \lambda\|\boldsymbol{\beta}\|_1/\sigma\right),$$

and let $\mathrm{X} = \mathrm{QL}$ be the QL decomposition of the matrix X. Then, the bijective smooth transformation

$$\begin{aligned} r = s/\sigma, \quad s^2 = \|\boldsymbol{y} - \mathrm{X}\hat{\boldsymbol{\beta}}\|^2, \quad \boldsymbol{z} = \boldsymbol{\beta}/\sigma \\ \nu = n - d + 1 + d = n + 1, \quad \boldsymbol{\gamma} = \mathrm{L}\hat{\boldsymbol{\beta}}/s \end{aligned}$$

yields

$$f(\boldsymbol{z}, r) \propto \lambda^d \mathsf{chi}_\nu(r) \exp\left(-\frac{1}{2}\sum_i \left(L_{ii}z_i - r\gamma_i + \sum_{j<i} L_{ij}z_j\right)^2 - \lambda\sum_j |z_j|\right).$$

Now set

$$\alpha_j(r, z_1, \ldots, z_{j-1}) := -r\gamma_j + \sum_{k<j} L_{jk}z_k,$$

and $\mathscr{J} := \{j : L_{jj} \neq 0\}$, so that we can write:

$$f(\boldsymbol{z}, r) \propto \lambda^d f_\nu(r) \exp\left(-\sum_{j \notin \mathscr{J}} \frac{\alpha_j^2}{2} - \sum_{j \in \mathscr{J}} \frac{(L_{jj} z_j + \alpha_j)^2}{2} - \lambda \sum_j |z_j| \right).$$

The key insight here is that $f(\boldsymbol{z}, r)$ can be 'nearly' written as a product of chi and a sequence of Lap or nl, and each term, as a function of z_j only depends on $z_1, \ldots z_{j-1}$. Consider for example $d = 3$, so that $\boldsymbol{z} = (z_1, z_2, z_3)$.

- If $\mathscr{J} = \emptyset$, then

  $$f(\boldsymbol{z}, r) \propto \mathsf{chi}_\nu(r) \times \exp\left(-\frac{\alpha_1^2}{2} - \lambda |z_1| \right) \times \exp\left(-\frac{\alpha_2^2}{2} - \lambda |z_2| \right) \times \exp\left(-\frac{\alpha_3^2}{2} - \lambda |z_3| \right).$$

  This looks like, but does not equal to

  $$\mathsf{chi}_\nu \times \mathsf{Lap} \times \mathsf{Lap} \times \mathsf{Lap}.$$

- If $\mathscr{J} = \{1, 2, 3\}$, then

  $$f(\boldsymbol{z}, r) \propto \mathsf{chi}_\nu(r) \times \exp\left(-\frac{(L_{11} z_1 + \alpha_1)^2}{2} - \lambda |z_1| \right) \times \exp\left(-\frac{(L_{22} z_2 + \alpha_2)^2}{2} - \lambda |z_2| \right) \times \exp\left(-\frac{(L_{33} z_3 + \alpha_3)^2}{2} - \lambda |z_3| \right).$$

  This looks like (and again does not equal to)

  $$\mathsf{chi}_\nu \times \mathsf{nl} \times \mathsf{nl} \times \mathsf{nl}.$$

- Finally, if $\mathscr{J} = \{1,3\}$, then

  $$f(\boldsymbol{z}, r) \propto \mathsf{chi}_\nu(r) \times \exp\left(-\frac{(L_{11} z_1 + \alpha_1)^2}{2} - \lambda |z_1| \right) \times \exp\left(-\frac{\alpha_2^2}{2} - \lambda |z_2| \right) \times \exp\left(-\frac{(L_{33} z_3 + \alpha_3)^2}{2} - \lambda |z_3| \right).$$

  This looks like (and again does not equal to)

  $$\mathsf{chi}_\nu \times \mathsf{nl} \times \mathsf{Lap} \times \mathsf{nl}.$$

In general, f always looks like a product of a sequence of chi_ν, Lap and nl, where the first term is always a chi_ν, and the $(j+1)$-term looks like a nl if $j \in \mathscr{J}$, otherwise it looks like

a Lap. Moreover, the α_j in each term is an expression that only depends on r and z_k, for $k < j$. This motivates a sequential proposal density $g(\boldsymbol{z}, r)$ on $\mathbb{R}^d \times \mathbb{R}_+$ by

$$R \to (Z_1 \,|\, R) \to (Z_2 \,|\, R, Z_1) \to (Z_3 \,|\, R, Z_1, Z_2) \to \dots.$$

Again, let $(\boldsymbol{\mu}, \eta) = (\mu_1, \dots, \mu_d, \eta)$ be some tilting parameter, we shall define $g(\boldsymbol{z}, r; \boldsymbol{\mu}, \eta)$ by

$$\begin{aligned} R &\sim \mathsf{TN}_{(0,\infty)}(\eta, 1) \\ Z_j \,|(R, Z_1, \dots, Z_{j-1}) &\sim \mathsf{nl}(L_{jj} z_j + \alpha_j - \mu_j; \lambda/|L_{jj}|, \alpha_j - \mu_j), \quad \text{if } j \in \mathscr{J} \\ Z_j \,|(R, Z_1, \dots, Z_{j-1}) &\sim \mathsf{Lap}_{\mu_j}(\lambda z_j), \quad \text{if } j \notin \mathscr{J}. \end{aligned}$$

Notice that $g(\boldsymbol{z}, r; \boldsymbol{\mu}, \eta)$ is chosen so that

$$\begin{aligned} \psi(\boldsymbol{z}, r; \boldsymbol{\mu}, \eta, \lambda) &:= \ln \frac{f(\boldsymbol{z}, r)}{g(\boldsymbol{z}, r; \boldsymbol{\mu}, \eta, \lambda)} \\ &= \sum_{j \in \mathscr{J}} \left(\frac{\mu_j^2}{2} - \mu_j (L_{jj} z_j + \alpha_j) \right) - \lambda \sum_{i \notin \mathscr{J}} \mu_j z_j - \sum_{j \neq \mathscr{J}} \frac{\alpha_j^2}{2} \\ &\quad + \frac{\eta^2}{2} - r\eta + (\nu - 1) \ln r + \ln \overline{\Phi}(-\eta) + \text{const.} + d \ln \lambda \\ &\quad + \sum_{j \in \mathscr{J}} \xi(\lambda / L_{jj}, \alpha_j - \mu_j) - \sum_{j \notin \mathscr{J}} \ln(1 - \mu_j^2) \end{aligned}$$

is concave in $(\boldsymbol{z}, r)$ and convex in $(\boldsymbol{\mu}, \eta)$. Consequently, we can seek for $\inf_{(\boldsymbol{\mu}, \eta)} \sup_{(\boldsymbol{z}, r)} \psi$ by solving $\nabla \psi = 0$.

Denoting $\tilde{\mathrm{L}} = \mathrm{L} - \mathrm{diag}(\mathrm{L})$ and $\xi_{2j} := \xi_2(\lambda / L_{jj}, \alpha_j - \mu_j)$, we have that for $\nabla \psi$,

$$\begin{aligned} \partial \psi / \partial z_i &= -\mu_i (L_{ii} \mathbb{I}_{\{i \in \mathscr{J}\}} + \mathbb{I}_{\{i \notin \mathscr{J}\}}) + \textstyle\sum_{j \in \mathscr{J}} \tilde{L}_{ji} \xi_{2j} - \sum_{j \notin \mathscr{J}} \mu_j \tilde{L}_{ji} \\ \partial \psi / \partial r &= -\eta + (\nu - 1)/r - \textstyle\sum_{j \in \mathscr{J}} \gamma_j \xi_{2j} + \sum_{j \notin \mathscr{J}} \mu_j \gamma_j \\ \partial \psi / \partial \mu_i &= (\mu_i - (L_{ii} z_i + \alpha_i) - \xi_{2j}) \mathbb{I}_{\{i \in \mathscr{J}\}} - (z_i - 2\mu_i / (\lambda^2 - \mu_i^2)) \mathbb{I}_{\{i \notin \mathscr{J}\}} \\ \partial \psi / \partial \eta &= \eta - r + \phi(\eta) / \Phi(\eta) \\ \partial \psi / \partial \lambda &= (d + |\mathscr{J}^c|)/\lambda - 2\lambda \sum_{j \notin \mathscr{J}} \frac{1}{\lambda^2 - \mu_j^2} + \sum_{j \in \mathscr{J}} \xi_{1j} / L_{jj}. \end{aligned}$$

In summary, we have the following algorithms.

Algorithm 4 : Defining the proposal density $g(\boldsymbol{z}, r; \boldsymbol{\mu}, \eta)$.

Require: tuning parameters η and $\boldsymbol{\mu}$

Compute QL decomposition X = QL; $\hat{\boldsymbol{\beta}} \leftarrow \text{argmin}_{\boldsymbol{\beta}} \|\boldsymbol{y} - \text{X}\boldsymbol{\beta}\|_2$; $s^2 \leftarrow \|\text{X}\hat{\boldsymbol{\beta}} - \boldsymbol{y}\|_2^2$;

$\boldsymbol{\gamma} \leftarrow \text{L}\hat{\boldsymbol{\beta}}/s$

$R \sim \text{chi}_{n+1}(r)$

for $j = 1, \ldots, p$ **do**

  if $j \notin \mathscr{J}$ **then**

    $Z_j \sim \mathsf{Lap}_{\mu_j}(\lambda z_j)$

  else

    $\alpha_j \leftarrow -R\gamma_j + \sum_{k<j} l_{jk} Z_k$

    $Z_j \sim \mathsf{nl}(l_{jj} z_j + \alpha_j - \mu_j; \lambda/|l_{jj}|, \alpha_j - \mu_j)$

return A draw $(\boldsymbol{Z}, R)$ from the proposal density.

Algorithm 5 : Simulating from posterior $\pi(\boldsymbol{\beta}, \sigma \,|\, \boldsymbol{y}, \lambda)$.

Require: shrinkage parameter $\lambda > 0$

Solve the nonlinear system $\nabla\psi = \boldsymbol{0}$ to obtain the solution $(\boldsymbol{z}^*, r^*; \boldsymbol{\mu}^*, \eta^*)$

$\psi^* \leftarrow \psi(\boldsymbol{z}^*, r^*; \boldsymbol{\mu}^*, \eta^*)$

repeat

  $E \sim \mathsf{Exp}(1)$, that is, E is an exponential r.v. with rate unity

  $(\boldsymbol{Z}, R) \sim g(\boldsymbol{z}, r; \boldsymbol{\mu}^*, \eta^*)$ using Algorithm 4

until $E > \psi^* - \psi(\boldsymbol{Z}, R; \boldsymbol{\mu}^*, \eta^*)$

$\sigma \leftarrow s/R$

$\boldsymbol{\beta} \leftarrow \sigma \boldsymbol{Z}$

return A draw $(\boldsymbol{\beta}, \sigma)$ from the posterior density.

3.4 Numerical studies for the general case

3.4.1 Diabetes Dataset

The first numerical experiment considers the same dataset as in Section 3.2.2, but we now include σ in the simulation scheme. The results of simulating from the posterior are given in the last two columns of Table 3.4.1, which also shows the ordinary least squares estimate. Figure 3.2 shows the estimated marginal distributions of each of the ten predictors with the ordinary least squares point-estimate superimposed as a (blue) dot on the boxplots.

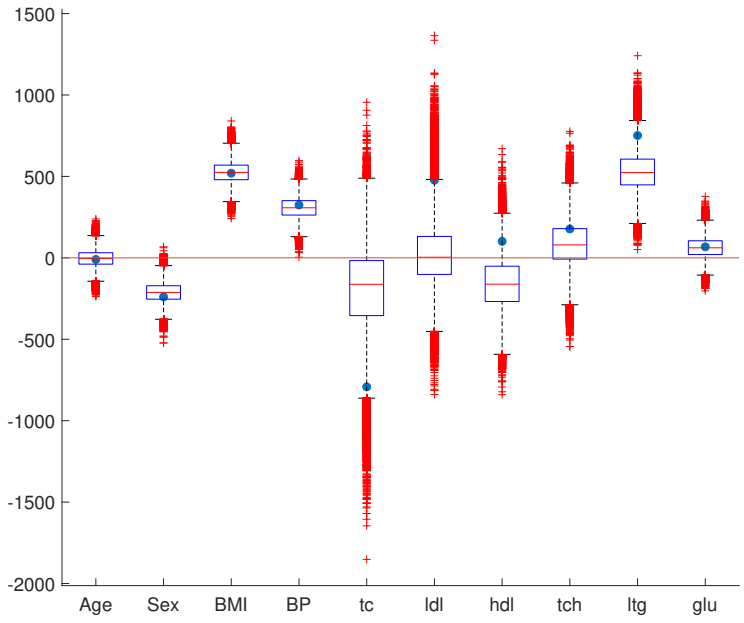

Figure 3.2: Marginal distributions of each predictor coefficient (as boxplots). The statistically significant predictors appear to be sex, body mass index (BMI), blood pressure (BP), and lgt.

	Ord. least sq.	Posterior Median	Estimated 95% credible interval
age of patient (age)	-10	-3.1	(-110,102)
Gender of patient (sex)	-239	-212	(-332,-91)
body mass index (BMI)	519	524	(394,655)
blood pressure (BP)	324	307	(179,435)
tc	-792	-161	(-807,232)
ldl	476	2.7	(-327,491)
hdl	101	-160	(-454,153)
tch	177	82	(-186, 374)
ltg	751	523	(309,791)
glu	67	61	(-53,188)

Table 3.1: Simulation results from the posterior density (diabetes dataset) using 10^5 iid draws.

The acceptance rate of the sequential sampling Algorithm 5 was estimated to be approximately 0.39. In contrast, the naive rejection Algorithm 2 has an acceptance probability smaller than 10^{-7} (making the event of drawing from the posterior a rare event, and the probability of acceptance a rare-event probability).

In addition, we compared the output of Algorithm 5 (taking about 7 seconds) with the output of the popular Park&Casella Gibbs sampler (taking about 16 seconds). We observed that the boxplots computed from the output of the Gibbs sampler (not shown

here) do not extend as much as the boxplots on Figure 3.2. This suggests that the exact sampler is better at exploring the tails of the posterior density.

3.4.2 Boston Housing Dataset

As another example, we consider the Boston housing dataset from [51], which attempts to explain housing prices in the Boston area from the $p = 13$ predictors given in Table 3.4.2 and $n = 506$ observations. The table confirms that Algorithm 5 simulates from the posterior accurately, because the MCMC simulations in [92] suggest the same list of statistically relevant predictors (crime levels, proximity to waterfront, number of rooms, distance to employment centers, etc) using the shrinkage value of $\lambda = 5.71$. For this example, the acceptance rate of the sequential rejection sampler was estimated to be approximately 0.67.

	Ord. least sq.	Posterior Median	Estimated 95% credible interval
per capita crime rate by town (crim)	-0.10	-0.098	(-0.16, -0.032)
proportion of zoned land (zn)	0.047	0.048	(0.021,0.076)
proportion of non-retail business (indus)	0.011	-0.034	(-0.15, 0.084)
Charles River exposure (chas)	2.75	1.75	(0.14, 3.47)
nitric oxide pollution (nox)	-12.0	-1.49	(-6.41 , 0.70)
number of rooms (rm)	4.61	4.079	(3.54, 4.62)
proportion built before 1940 (age)	-0.0023	-0.010	(-0.035 , 0.015)
distance to CBD (dis)	-1.28	-1.17	(-1.54, -0.80)
highway access (rad)	0.25	0.25	(0.13, 0.38)
property tax rate (tax)	-0.011	-0.013	(-0.020, -0.0059)
quality of schools (ptratio)	-0.73	-0.72	(-0.93, -0.51)
proportion of ethnic diversity (b)	0.011	0.010	(0.0056, 0.015)
economic status of residents (lstat)	-0.48	-0.53	(-0.63,-0.44)

Table 3.2: Posterior estimates for the Boston Housing dataset using 10^5 iid draws.

3.5 Concluding remarks for this chapter

In this chapter we have constructed two exact samplers for the posterior distribution of the Bayesian Lasso linear regression model for a simplified case where σ is considered fixed and for the standard case where σ is included for the posterior inference too. Both samplers are rejection samplers, however the proposal densities are optimally tilted to achieve efficiency even for real datasets with dimensions greater than 10.

An optimally tilted normal density is constructed for the first case. Unsurprisingly, it turns out that the optimal tilting parameter corresponds to the solution to the frequentist Lasso linear regression problem. We also show that this proposal density renders an importance sampling estimator for the marginal likelihood whose variance is better than a naive alternative, say a normal proposal density centred at the OLS estimator.

An optimally tilted sequential proposal density is constructed next, which is a more sophisticated alternative to the simpler Gaussian proposal. Two real datasets are tested against this samplers, and it appears that this sequential proposal density performs well for the posterior inference on the Bayesian Lasso linear regression model. This sequential density naturally gives an importance sampling estimator for the marginal likelihood as well (not presented here).

Despite this success, due to the curse of dimensionality, these rejection samplers are bound to be inefficient as the dimension of the posterior distributions increase. In this manner, no matter how well a proposal distribution can approximate a posterior distribution, say for example using the tilting techniques we have studied so far, rejection samplings will eventually be inefficient as the dimensions increase.

To this end, we shall move on to the study of Markov chain sampling. The error analysis of Markov chain sampling is known to be difficult. Nevertheless, in Chapter 5 we present novel diagnostics for geometrically ergodic Markov chain samplers whose 'regeneration times' are identifiable. Roughly speaking, regeneration times are instances where a stochastic process restarts. Some theoretical backgrounds on Markov chains and regenerative processes are given in Chapter 4.

We also study what happens when the proposal densities we have considered so far are used as proposal densities in the context of an independence (Markov chain) samplers, instead of rejection sampling. Our novel diagnostics reveal promising convergence acceleration in this setting, and thus these proposal densities have values outside the rejection sampling paradigm as well.

Chapter 4

Theoretical backgrounds on Markov chains and regenerative processes

4.1 Introduction to this chapter

In Chapters 2 and 3 we have seen how exponentially tilted sequential proposal densities result in efficient rejection samplers for some posterior densities π in Bayesian inferences. Rejection samplers gives exact[1] iid draws $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots \sim \pi$, so that for any measurable function h such that $\int |h(\boldsymbol{y})|\pi(d\boldsymbol{y}) < \infty$

$$\tilde{q}_t := \frac{1}{t}\sum_{k=1}^{t} h(\boldsymbol{Y}_k) \xrightarrow{\text{a.s.}} \int h(\boldsymbol{y})\pi(d\boldsymbol{y}), \quad \text{as } t \to \infty.$$

Moreover standard iid error analysis, such as computing the standard error of $\tilde{q}_t$ or analyzing $\tilde{q}_t$ asymptotically as some normally distributed random variable (i.e. the CLT for $\tilde{q}_t$), are valid here (provided that $\int [h(\boldsymbol{y})]^2\pi(d\boldsymbol{y}) < \infty$). However, these samplers are bound to fail when the dimension of the support of π increases, and in such an extreme setting one has to resort to approximate sampling schemes such as MCMC.

MCMC simulates a Markov chain $\{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots\}$ whose distribution limits towards π. In a similar way, one definines

$$\hat{q}_t := \frac{1}{t}\sum_{k=1}^{t} h(\boldsymbol{X}_k),$$

and hopes that $\hat{q}_t \xrightarrow{\text{a.s.}} \int h(\boldsymbol{x})\pi(d\boldsymbol{x})$ too. Moreover, one can wish for some CLT approximation to hold as well. However, these are not guaranteed for arbitrary Markov chains.

In this chapter we shall give a brief recount the theory of Markov chains on $\mathcal{X} \subseteq \mathbb{R}^p$. In particular, we state known results concerning the conditions under which a Markov chain converges to a limiting distribution. We also describe notions regarding a regenerative processes and their connections to Markov chains. Along the way, we shall establish notations and definitions we use in the study of Chapter 5, in which we describe our novel

[1] In this thesis we often refer to rejection sampling as 'exact sampling' because they obey the target density. This is different to some MCMC literature where the term 'exact MCMC' refers to a Markov chain whose initial distribution coincides with its limiting distribution.

MCMC error analyses. The definitions and results here can be found in related books such as [3, 83] and survey papers such as [96, 66].

4.2 Theoretical backgrounds on Markov chains and regenerative processes

Definition 4.2.1. A transition probability kernel κ on a measurable space $(\mathcal{X}, \mathscr{A})$ is a function $\kappa : \mathcal{X} \times \mathscr{A} \to [0, 1]$ such that

1. for every $\boldsymbol{x} \in \mathcal{X}$, $\kappa(\cdot \,|\, \boldsymbol{x}) := \int_{\cdot} \kappa(d\xi \,|\, \boldsymbol{x})$ is a probability measure; and
2. for every $A \in \mathscr{A}$, the function defined by $\kappa(A \,|\, \boldsymbol{x}) := \int_A \kappa(dy \,|\, \cdot)$ is measurable.

The idea of a probability transition kernel is to associate a probability measure for each point in $\mathcal{X}$ so that starting from some initial $\boldsymbol{X}_1 \sim \nu_1$, the sequence of random vectors $\{\boldsymbol{X}_k, k \geq 1\}$ such that $\mathbb{P}[\boldsymbol{X}_{t+1} \in A | \boldsymbol{X}_t = \boldsymbol{x}] = \kappa(A \,|\, \boldsymbol{x})$ for all $t = 1, 2, \ldots$ and $A \in \mathscr{A}$. Such a sequence of random vectors is called a time-homogeneous Markov chain on state space $(\mathcal{X}, \mathscr{A})$ and, for all $A_1, \ldots, A_t \in \mathscr{A}$, it satisfies the relation

$$\mathbb{P}[\boldsymbol{X}_1 \in A_1, \boldsymbol{X}_2 \in A_2, \ldots, \boldsymbol{X}_t \in A_t] = \int_{\boldsymbol{x}_1 \in A_1} \cdots \int_{\boldsymbol{x}_{t-1} \in A_{t-1}} \kappa(A_t \,|\, \boldsymbol{x}_{t-1}) \kappa(d\boldsymbol{x}_{t-1} \,|\, \boldsymbol{x}_{t-2}) \ldots \kappa(d\boldsymbol{x}_2 \,|\, \boldsymbol{x}_1) \nu_1(d\boldsymbol{x}_1).$$

Notice that κ naturally defines an operator on the set of probability measures on $(\mathcal{X}, \mathscr{A})$ via

$$(\nu\kappa)(A) := \int_{\boldsymbol{x} \in \mathcal{X}} \kappa(A \,|\, \boldsymbol{x}) \pi(d\boldsymbol{x}),$$

for any probability measure ν and $A \in \mathscr{A}$. In this manner, the t-step probability transition kernel, κ_t, can be defined by the recursion

$$\kappa_t(A \,|\, \boldsymbol{x}) = [\kappa(\cdot \,|\, \boldsymbol{x}) \kappa_{t-1}](A), \qquad \text{for } t = 2, 3, 4, \ldots.$$

Further if, π is invariant in the sense $\pi\kappa = \pi$, we say π is the stationary probability measure for κ. (We also call it the stationary probability distribution of a Markov chain κ induces.) In many applications of Bayesian inference, one designs a κ for which the posterior probability distribution π is the invariant one and in this way, the simulated Markov chain κ induces may perhaps be approximate draws from π.

Checking whether or not $\pi\kappa = \pi$ in practice can be difficult. Indeed, one often checks the following sufficient condition known as the reversibility condition.

Definition 4.2.2. Let κ be a transition probability kernel on a measurable space $(\mathcal{X}, \mathscr{A})$. The Markov chain κ induces is said to be reversible with respect to a probability measure π on $(\mathcal{X}, \mathscr{A})$, if for every $A, B \in \mathscr{A}$

$$\int_{\boldsymbol{x} \in B} \kappa(A \,|\, \boldsymbol{x}) \pi(d\boldsymbol{x}) = \int_{\boldsymbol{x} \in A} \kappa(B \,|\, \boldsymbol{x}) \pi(d\boldsymbol{x}).$$

This is sometimes stated in the following form.

$$\kappa(d\boldsymbol{y} \,|\, \boldsymbol{x}) \pi(d\boldsymbol{x}) = \kappa(d\boldsymbol{x} \,|\, \boldsymbol{y}) \pi(d\boldsymbol{y}).$$

Intuitively, the equation describes the situation where the probability flux of entering a set from another set is the same as the probability flux of the reverse transition. Formally one can check that if κ is reversible with respect to π, then π is stationary for κ with the following calculations. For any $A \in \mathscr{A}$,

$$(\pi\kappa)(A) = \int_{x \in \mathcal{X}} \kappa(A \,|\, \boldsymbol{x}) \pi(d\boldsymbol{x}) = \int_{x \in A} \kappa(\mathcal{X} \,|\, \boldsymbol{x}) \pi(d\boldsymbol{x}) = \int_{x \in A} \pi(d\boldsymbol{x}) = \pi(A).$$

The equation $\pi\kappa = \pi$ is sometimes called the 'global-balance' equation while the equation $\kappa(d\boldsymbol{y} \,|\, \boldsymbol{x}) \pi(d\boldsymbol{x}) = \kappa(d\boldsymbol{x} \,|\, \boldsymbol{y}) \pi(d\boldsymbol{y})$ is called the 'detail-balance' equation. One should also note that if π and $\kappa(\cdot \,|\, \boldsymbol{x})$ exhibits a density for every $\boldsymbol{x}$ (say with respect to the Lebesgue measure or the counting measure, depending on what $(\mathcal{X}, \mathscr{A})$ is), checking whether detail-balance equation is satisfied amounts to checking $\kappa(\boldsymbol{y} \,|\, \boldsymbol{x}) \pi(\boldsymbol{x}) = \kappa(\boldsymbol{x} \,|\, \boldsymbol{y}) \pi(\boldsymbol{y})$, where $\kappa(\boldsymbol{y} \,|\, \boldsymbol{x})$ and $\pi(\boldsymbol{x})$ are the density functions.

Next we introduce the notion of total variation distance. The total variation distance is a metric that can be defined on sets of (probability) measures.

Definition 4.2.3. Let π_1 and π_2 be two probability measures defined on $(X, \mathcal{F})$. The total variation distance between two probability measures is

$$\|\pi_1 - \pi_2\|_{\mathrm{TV}} = \sup_{A \in \mathcal{F}} |\pi_1(A) - \pi_2(A)|.$$

When approximating draws from π by a Markov chain $\{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots\}$, the least that one should require is having π as some 'limiting distribution' of the Markov chain. Formally, this means $\|\kappa_t(\cdot \,|\, \boldsymbol{x}_1) - \pi\|_{\mathrm{TV}} \to 0$ as $t \to \infty$ for π-a.e. $\boldsymbol{x}_1 \in \mathcal{X}$.

Unfortunately, reversibility by itself is not sufficient to guarantee the Markov chain to limit towards π, let alone $\hat{q}_t \to q$ and the existence of some CLT for $\hat{q}_t$. Reversibility simply ensures that if the Markov chain, for some reason, at some stage $\boldsymbol{X}_k$ is π distributed, then the marginal distributions of $\boldsymbol{X}_l$ will be π as well for $l > k$.

A concrete but naive example to illustrate why reversibility with respect to π by itself is insufficient is to consider $\kappa(d\boldsymbol{y} \,|\, \boldsymbol{x}) = \delta_{\boldsymbol{x}}(d\boldsymbol{y})$ for every $\boldsymbol{x} \in \mathcal{X}$. Here, for any $A \in \mathscr{A}$,

$$(\pi\delta_{\boldsymbol{x}})(A) = \int \delta_{\boldsymbol{x}}(A)\pi(d\boldsymbol{x}) = \pi(A),$$

so that π is stationary for κ. However this probability transition kernel renders an uninteresting Markov chain, $\{\boldsymbol{X}_1, \boldsymbol{X}_1, \boldsymbol{X}_1, \ldots\}$, which clearly cannot be a helpful approximation for most π. To this end, we need further assumptions on the probability transition kernel.

Definition 4.2.4. Suppose κ is a probability transition kernel with invariant probability measure π. The probability transition kernel κ and the Markov chain it induces are said to be

1. ϕ-irreducible if there is a σ-finite measure ϕ on $\mathcal{X}$, such that for all $A \subseteq \mathcal{X}$ with $\phi(A) > 0$, there is a positive integer t for which $\kappa_t(A \,|\, \boldsymbol{x}) > 0$ for all $\boldsymbol{x} \in \mathcal{X}$.
2. Aperiodic if there is no $d \geq 2$ and disjoint subsets $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_d \subseteq \mathcal{X}$ such that $\kappa(\mathcal{X}_{(i+1) \bmod d} \,|\, \boldsymbol{x}) = 1$ for all $\boldsymbol{x} \in \mathcal{X}_i$ for $i = 1, \ldots, d$.
   In the case where such a d exists, we call the largest such d the period of the Markov chain and the Markov chain is said to be periodic.

It is apparent how these notions mimic the notion of irreducibility and aperiodicity for the classical treatment of Markov chains on countable state spaces. Indeed, just like Markov chains on countable state spaces, we have the following well-known theorems in the MCMC literature.

Theorem 4.2.1. Suppose $\{\boldsymbol{X}_k, k \geq 1\}$ is a Markov chain on a state space with countably generated σ-algebra and denote its t-step transition kernel as κ_t. Further suppose that it is ϕ-irreducible, aperiodic and exhibits an stationary probability measure π, then

1. for π-a.e. $\boldsymbol{x} \in \mathcal{X}$, $\|\kappa_t(\cdot \,|\, \boldsymbol{x}) - \pi\|_{\mathrm{TV}} \to 0$ as $t \to \infty$;
2. $\hat{q}_t := \frac{1}{t}\sum_{k=1}^{t} h(\boldsymbol{X}_k) \to q := \int h(\boldsymbol{x})\pi(d\boldsymbol{x})$ almost surely, as $t \to \infty$, provided $\int |h(\boldsymbol{x})|\pi(d\boldsymbol{x}) < \infty$.

Theorem 4.2.1 works for most Markov chain samplers in practice, however since the statement holds only for π-a.e. $\boldsymbol{x} \in \mathcal{X}$, if one initializes the chain exactly at points for which the statement fails, the resulting Markov chain can have pathological behaviors. (See [16, 109] for discussion on this matter.) To improve from 'π-a.e. convergence' to 'convergence for all $\boldsymbol{x} \in \mathcal{X}$' we need a stronger condition known as Harris recurrence.

Definition 4.2.5. A ϕ-irreducible Markov chain with stationary probability measure π is Harris recurrent if for all $A \subseteq \mathcal{X}$ with $\phi(A) > 0$ and all $\boldsymbol{x} \in \mathcal{X}$, we have

$$\mathbb{P}[\tau_A < \infty \,|\, \boldsymbol{X}_1 = x] = 1$$

where $\tau_A = \inf\{k > 1 \,|\, \boldsymbol{X}_k \in A\}$.

It turns out that being Harris recurrent and aperiodic is equivalent to the convergence of κ_t to π for all initial $\boldsymbol{x} \in \mathcal{X}$ [95]. This can be formally stated as follows.

Theorem 4.2.2. For a ϕ-irreducible, aperiodic Markov chain with stationary probability measure π, Harris recurrence is equivalent to

$$\|\kappa_t(\cdot \,|\, \boldsymbol{x}) - \pi\|_{\mathrm{TV}} \to 0, \quad \text{as } t \to \infty,$$

for all $\boldsymbol{x} \in \mathcal{X}$ where κ_t is the t-step probability transition kernel of the Markov chain.

Fortunately, many practical Markov chain samplers, in particular the Markov chains we consider this thesis, are indeed Harris recurrent. Precisely, suppose that π exhibits a density (in this thesis, densities are always with respect to the Lebesgue measure on $\mathbb{R}^d$) and let $g(\boldsymbol{y} \,|\, \boldsymbol{x})$ be some transition density. Defining

$$\alpha(\boldsymbol{y} \,|\, \boldsymbol{x}) = \begin{cases} \min\left\{1, \frac{\pi(\boldsymbol{y})g(\boldsymbol{x} \,|\, \boldsymbol{y})}{\pi(\boldsymbol{x})g(\boldsymbol{y} \,|\, \boldsymbol{x})}\right\}, & \text{if } \pi(\boldsymbol{x})g(\boldsymbol{y} \,|\, x) > 0, \\ 1, & \text{if } \pi(\boldsymbol{x})g(\boldsymbol{y} \,|\, x) = 0, \end{cases}$$

the probability transition kernel of the well-known Metropolis-Hastings sampler is

$$\kappa(d\boldsymbol{y} \,|\, \boldsymbol{x}) = \alpha(\boldsymbol{y} \,|\, \boldsymbol{x})g(d\boldsymbol{y} \,|\, \boldsymbol{x})\, d\boldsymbol{y} + (1 - \alpha^*(\boldsymbol{x}))\delta_{\boldsymbol{x}}(d\boldsymbol{y})$$

where $\alpha^*(\boldsymbol{x}) = \int \alpha(\boldsymbol{u} \,|\, \boldsymbol{x})g(d\boldsymbol{u} \,|\, \boldsymbol{x})$. Here, $\alpha(\boldsymbol{y} \,|\, \boldsymbol{x})$ is chosen so that $\alpha(\boldsymbol{y} \,|\, \boldsymbol{x})g(\boldsymbol{y} \,|\, \boldsymbol{x})$ is reversible with respect to π, and in this manner, π is a stationary probability measure for κ here. The following result due to [109] ensures a Metropolis-Hastings sampler generates a Harris recurrent Markov chain.

Theorem 4.2.3. Let κ be the kernel of a Metropolis-Hastings sampler with stationary probability measure π. If κ is π-irreducible, then the Markov chain κ induces is Harris recurrent.

In this thesis, we are particularly interested in the independence (chain) sampler and the (block) Gibbs sampler. The independence sampler is a special case of the Metropolis-Hastings sampler the density g is a constant in $\boldsymbol{x}$ i.e. $g(\boldsymbol{y} \,|\, \boldsymbol{x}) = g(\boldsymbol{y})$ for all $\boldsymbol{x}$. (The general Metropolis-Hastings sampler, also known as the random-walk sampler is not considered in this thesis because there is no systematic way to identify its regeneration times, as discuss later.) Algorithmically, given the current state of the Markov chain $\boldsymbol{x}$, an independence sampler simulates a draw $\boldsymbol{Y}$ from another density g, which is called the proposal density. It is similar to rejection sampling but there are two differences. Firstly, the probability of retaining $\boldsymbol{Y}$ depends on $\boldsymbol{x}$, secondly, in the event of a rejection, the next state of the Markov chain is set to be the current state. Formally, the transition kernel of an independence sampler with proposal density g, targeting density π is

$$\kappa(d\boldsymbol{y} \,|\, \boldsymbol{x}) = \alpha(\boldsymbol{y} \,|\, \boldsymbol{x}) g(\boldsymbol{y})\, d\boldsymbol{y} + (1 - \alpha^*(\boldsymbol{x}))\delta_{\boldsymbol{x}}(d\boldsymbol{y}),$$

where $\alpha(\boldsymbol{y} \,|\, \boldsymbol{x}) = \min\left\{1, \frac{\pi(\boldsymbol{y})g(\boldsymbol{x})}{\pi(\boldsymbol{x})g(\boldsymbol{y})}\right\}$, and $\alpha^*(\boldsymbol{x}) = \int \alpha(\boldsymbol{u} \,|\, \boldsymbol{x}) g(\boldsymbol{u})\, d\boldsymbol{u}$. It is clear from the construction of this algorithm, Theorem 4.2.3 trivially holds for the independence sampler if the support of g encompasses the support of π.

On the other hand, a Gibbs sampler begins by segmenting $\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p)^\top$ and constructs the full conditional densities $\pi(\boldsymbol{y}_j \,|\, \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{j-1}, \boldsymbol{x}_{j+1}, \ldots \boldsymbol{x}_p)$ for $j = 1, \ldots, p$. Given the current state of the Markov chain $\boldsymbol{x}$, a Gibbs sampler cycles through each of the full conditionals by simulating draws $\boldsymbol{Y}_j$ from $\pi(\cdot \,|\, \boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots \boldsymbol{Y}_{j-1}, \boldsymbol{x}_{j+1}, \ldots, \boldsymbol{x}_p)$ for $j = 1, \ldots, d$. In this manner, its transition density is

$$\kappa(\boldsymbol{y} \,|\, \boldsymbol{x}) = \pi(\boldsymbol{y} \,|\, \boldsymbol{x}_2, \boldsymbol{x}_3 \ldots, \boldsymbol{x}_p)\pi(\boldsymbol{y}_2 \,|\, \boldsymbol{y}_1, \boldsymbol{x}_3, \ldots \boldsymbol{x}_p) \ldots \pi(\boldsymbol{y}_d \,|\, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_{d-1}).$$

It turns out that a Gibbs sampler is actually a special case of the Metropolis-Hastings algorithm as well, and moreover because it simply cycles through the full conditional densities of π, Theorem 4.2.3 again holds.

Although Harris reccurrence guarantees a limiting distribution π, it turns out that it still does not guarantee $\hat{q}_t$ to admit a CLT. Indeed, this actually needs assumptions on the rates with which $\kappa_t(\cdot \,|\, \boldsymbol{x})$ converges to π. To this end, we recall the following notions concerning the convergence rates of Markov chains.

Definition 4.2.6. A Markov chain having stationary distribution π is

1. *geometrically ergodic* if there is a function $M(\boldsymbol{x}) < \infty$ for π-a.e. $\boldsymbol{x} \in \mathcal{X}$ and some $r < 1$ such that for all $t = 1, 2, \ldots$

$$\|\kappa_t(\cdot \,|\, \boldsymbol{x}) - \pi\|_{\mathrm{TV}} \leq M(\boldsymbol{x}) r^t.$$

2. *uniformly ergodic* if there is a constant $M < \infty$ for π-a.e. $\boldsymbol{x} \in \mathcal{X}$ and some $r < 1$ such that for all $t = 1, 2, \ldots$

$$\|\kappa_t(\cdot \,|\, \boldsymbol{x}) - \pi\|_{\mathrm{TV}} \leq M r^t.$$

Clearly, uniform ergodicity is a stronger condition in the sense that uniform ergodicity implies geometric ergodicity. A version of Markov chain CLT for $\hat{q}_t$ is that if the Markov chain is Harris recurrent and geometrically ergodic, and that $\int [h(\boldsymbol{x})]^{2+\delta} \pi(d\boldsymbol{x}) < \infty$ for some $\delta > 0$, then $\sqrt{t}(\hat{q}_t - q) \to \mathsf{N}(0, \sigma^2)$ in distribution for some σ^2. Other versions of Markov chain CLT is surveyed in [66].

A result that is of particular interest for this thesis is the following theorem given in [81].

Theorem 4.2.4. The independence sampler is uniformly ergodic if there is a constant c such that

$$\frac{\pi(\boldsymbol{x})}{g(\boldsymbol{x})} \leq c, \quad \forall \boldsymbol{x} \in \mathcal{X},$$

where π is the stationary density and g is the proposal density.

It is apparent that the proposal densities described in Chapters 2 and 3 are bound to fail in the rejection sampling context, as the dimension of the problem grows. Nevertheless, this theorem suggests that we can perhaps use them as the proposal densities for independence sampling, and we are immediately guaranteed to enjoy uniform ergodicity. Indeed, we will present few numerical studies in this paradigm later in the thesis.

We now move on to regenerations in Markov chains. Firstly, recall that a strictly increasing sequence of random variables $\{T_k, k \geq 0 \,|\, 0 \leq T_0 < T_1 < T_2 < \ldots\}$ taking value on $\mathbb{R}_+$ or $\mathbb{N}$ is called a renewal process if the inter-arrival times $M_j = T_j - T_{j-1}$ are positive iid random variables.

Definition 4.2.7. A (discrete time) stochastic process $\{\boldsymbol{X}_k, k \geq 1\}$ is a zero-delayed regenerative process if there is a renewal process $\{T_k, k \geq 0\}$ with $T_0 = 0$, for which the process $\{X_{T_r+k}, k > 0\}$

1. has the same distribution as $\{\boldsymbol{X}_k, k \geq 1\}$; and
2. is independent of $\{\boldsymbol{X}_k, 1 < k \leq T_r - 1\}$

for all $r = 1, 2, \ldots$, and consequently, the cycles $\{\{\boldsymbol{X}_k, T_{r-1} < k \leq T_r\}, M_r\}$ are iid for all $r = 1, 2, \ldots$.

The cycles $\{\{\boldsymbol{X}_k, T_{r-1} < k \leq T_r\}, M_r\}$ can be formally understood as a killed process $\{\boldsymbol{Y}_k, k \geq 1\}$ where $\boldsymbol{Y}_k := \boldsymbol{X}_{T_{r-1}+k}$ for all $k \leq M_r$ and $\boldsymbol{Y}_k$ takes the 'coffin state', say for example $\mathbf{0}$, for all $k > M_r$ [104]. In this manner, in the context of a regenerative process, the random variables $M_r = T_r - T_{r-1}$ are often called the 'tour length' or the 'cycle length'.

Although the theory for renewal processes and regenerative processes have been developed for general index set $\mathbb{T}$, our focus is on Markov chains so that $\mathbb{T} = \mathbb{N}$ and M_j are random variables on the set of strictly positive integers $\mathbb{Z}_+$. Next, The random instances $T_0, T_1, T_2, \ldots$ are called regeneration times, and since the post regeneration processes are iid, the regeneration times are commonly understood as the instances for which the stochastic process has 'restarted'. Finally, in general, one can have $T_0 \geq 0$ with non-zero probability. Such a regenerative processes is called a delayed regenerative process, however we only consider zero-delayed regenerative processes in this thesis.

*Remark. We also note that many texts index a stochastic process from $k = 0$, in other words, they present their discrete-time stochastic process as the collection $\{\boldsymbol{X}_k, k \geq 0\}$. In this setting, the convention is to define the underlying regeneration times as the instances $T_0, T_1, T_2, \ldots$ that **initiate** a new regenerative cycle i.e. the instances such that $\{\{\boldsymbol{X}_k, T_{r-1} \leq k < T_r\}, M_r\}$ are iid for all r. This is in contrast to the presentation in this thesis, where we define regeneration times as the instances that **end** a regenerative cycle. This is because it appears that this convention is easier to present our contributions in later chapters.*

Clearly, for (time-homogenous) Markov chains regeneration happen whenever the process reaches any particular state. However, this is only useful for Markov chains on countable state space since in general uncountable state spaces, the probability of visiting the same state twice is usually zero. This leads to notion of an atom. Atoms are sets for which the transition out of the set is the same for any points. That is to say the following.

Definition 4.2.8. Let κ be the probability transition density of a Markov chain on $(\mathcal{X}, \mathscr{A})$. A set $C \in \mathscr{A}$ is called an atom for the Markov chain if there is a probability measure ν on $\mathscr{A}$ for which

$$\kappa(\cdot \,|\, \boldsymbol{x}) = \nu(\cdot), \quad \forall \boldsymbol{x} \in C.$$

The motivation of an atom C is that, should it exist, then any transition out of C is the same disregarding the exact position of the current state. That is $\boldsymbol{X}_{k+1} \sim \kappa(\cdot \,|\, X_k \in C) = \nu$. Consequently, if a Markov chain has an initial distribution of ν, that is $\boldsymbol{X}_1 \sim \nu$, then whenever the Markov chain falls into an atom, the next step always initiates a new regenerative cycle. To see this, suppose that $\boldsymbol{X}_k \in C$ and observe that $\boldsymbol{X}_{k+1} \sim \nu$ is actually independent of $\boldsymbol{X}_k$ and in fact the whole process $\{\boldsymbol{X}_k, k \geq 1\}$ must have the same distribution as that of $\{\boldsymbol{X}_{k+j}, j \geq 1\}$, again due to the fact that $\boldsymbol{X}_{k+1} \sim \nu$. Finally the tour lengths now corresponds to the hitting time τ_C starting from ν, and in this manner, the iid assumption on the tour lengths is fulfilled. Ultimately, if an atom C exists, the Markov chain is actually a zero-delayed regenerative process whenever the initial distribution is ν.

Interesting results concerning the moments of the tour lengths (i.e. the hitting times to C) can be found in [88]. In particular, all moments of regenerative tour lengths exist for a geometrically ergodic Markov chain.

The natural question that follows is whether an atom actually exists, and how does one identify one. Fortunately, the split-chain technique due to [4, 87] enables one to systematically construct atoms for many Markov chains by carefully injecting the Markov chain in an enlarged state space. We now present a brief summary of this technique.

For a given probability transition kernel κ, we begin by assuming that that there is a measurable function $s : \mathcal{X} \to [0, 1]$, and a probability measure ν for which

$$\kappa(d\boldsymbol{y} \,|\, \boldsymbol{x}) \geq s(\boldsymbol{x})\nu(d\boldsymbol{y})$$

for all $\boldsymbol{x} \in \mathcal{X}$. This inequality is known as the 'minorization condition', and in the case where $\kappa(\cdot \,|\, \boldsymbol{x})$ and ν exhibit densities for all $\boldsymbol{x}$, it simplifies to $\kappa(\boldsymbol{y} \,|\, \boldsymbol{x}) \geq s(\boldsymbol{x})\nu(\boldsymbol{y})$ for all $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}$.

It becomes apparent why the authors call this technique 'splitting' in the construction that follows. Let $\mathcal{X}^* = \mathcal{X} \times \{0, 1\}$ be an enlarged state space. For $i = 0, 1$ we denote $\boldsymbol{x}^i = \boldsymbol{x} \times \{i\}$, $A^i = A \times \{i\}$ for all $\boldsymbol{x} \in \mathcal{X}$, $A \in \mathscr{A}$, and equip $\mathcal{X}^*$ with $\mathscr{A}^*$, the σ-algebra generated by the family of sets $\{A^i \,|\, i = 0, 1 \text{ and } A \in \mathscr{A}\}$. Next, any probability

measure λ defined on $(\mathcal{X}, \mathscr{A})$ exhibits an unique extension λ^* defined $(\mathcal{X}^*, \mathscr{A}^*)$ for which λ^* satisfies

$$\lambda^*(A^0) = \int_A (1 - s(\boldsymbol{u}))\lambda(d\boldsymbol{u}), \quad \text{and} \quad \lambda^*(A^1) = \int_A s(\boldsymbol{u})\lambda(d\boldsymbol{u}),$$

for all $A \in \mathscr{A}$. Intuitively, the function s dictates how the 'volume' of a set $A \in \mathscr{A}$ splits into its corresponding 'branches' A^0 and A^1. Formally, for any $A \in \mathscr{A}$, observing that $A^0 \cup A^1 = A \times \{0, 1\}$, and $A^0 \cap A^1 = \emptyset$, we have

$$\lambda^*(A^0 \cup A^1) = \int_A (1 - s(\boldsymbol{u}))\lambda(d\boldsymbol{u}) + \int_A s(\boldsymbol{u})\lambda(d\boldsymbol{u}) = \lambda(A). \tag{4.1}$$

Consequently, $\lambda^*(\mathcal{X}^*) = \lambda(\mathcal{X}) = 1$ so that λ^* is a probability measure on $(\mathcal{X}^*, \mathscr{A})$.

In other words, if $(\boldsymbol{X}, Y)$ is a $\mathcal{X}^*$-valued random vector whose law coincides with λ^*, then the law of $\boldsymbol{X}$ coincides with λ. This observation will ultimately give as a Markov chain $\{(\boldsymbol{X}_k, B_k), k \geq 1\}$ such that the first coordinate process $\{\boldsymbol{X}_k, k \geq 1\}$ has the same distribution as the Markov chain κ induces, a regenerative cycle starts whenever $B_k = 1$. To achieve this, we need to carefully extend κ from $\mathcal{X} \times \mathscr{A}$ onto $\mathcal{X}^* \times \mathscr{A}^*$.

Extending κ to achieve the desired property is rather subtle. The construction begins by extending κ from defined on $\mathcal{X} \times \mathscr{A}$ to $\check{\kappa}$ defined on $\mathcal{X}^* \times \mathscr{A}$, where

$$\begin{aligned}
\check{\kappa}(A \,|\, \boldsymbol{x}^0) &= \frac{\kappa(A \,|\, \boldsymbol{x}) - s(\boldsymbol{x})\nu(A)}{1 - s(\boldsymbol{x})}, \\
\check{\kappa}(A \,|\, \boldsymbol{x}^1) &= \nu(A).
\end{aligned}$$

The next step is to extend $\check{\kappa}$ from $\mathcal{X}^* \times \mathscr{A}$ to $\mathcal{X}^* \times \mathscr{A}^*$. This is done by leveraging equation (4.1). For every $\boldsymbol{z} \in \mathcal{X}^*$, we extend the probability measure $\check{\kappa}(\cdot \,|\, \boldsymbol{z})$ defined on $(\mathcal{X}, \mathscr{A})$, to $\check{\lambda}^*(\cdot \,|\, \boldsymbol{z})$ defined on $(\mathcal{X}^*, \mathscr{A}^*)$. We shall henceforth suppress our notation and write κ^* instead of $\check{\kappa}^*$.

To see why κ^* can induce a Markov chain $\{(\boldsymbol{X}_k, B_k), k \geq 1\}$ on $\mathcal{X}^*$ such that the first coordinate process $\{\boldsymbol{X}_k, k \geq 1\}$ has the same distribution as a Markov chain κ induces, let ν_1 be some (initial) distribution on $(\mathcal{X}, \mathscr{A})$ and suppose $(\boldsymbol{X}_1, B_1) \sim \nu_1^*$ is an initial state, so that $\boldsymbol{X}_1 \sim \nu_1$. Observe that for any $A \in \mathscr{A}$

$$\begin{aligned}
(\nu_1^* \kappa^*)(A^0 \cup A^1) &= \int_{\mathcal{X}^*} \kappa^*(A^0 \cup A^1 \,|\, \boldsymbol{x})\nu_1^*(d\boldsymbol{x}) \\
&= \int_{\mathcal{X}} \frac{\kappa(A \,|\, \boldsymbol{x}) - s(\boldsymbol{x})\nu(A)}{1 - s(\boldsymbol{x})}(1 - s(\boldsymbol{x}))\nu_1(d\boldsymbol{x}) + \int_{\mathcal{X}} \nu(A)s(\boldsymbol{x})\nu_1(d\boldsymbol{x}) \\
&= \int_{\mathcal{X}} \kappa(A \,|\, \boldsymbol{x})\nu_1(d\boldsymbol{x}) \\
&= (\nu_1 \kappa)(A).
\end{aligned}$$

Consequently, whenever κ^* and initial distribution ν_1^* induce a Markov chain $\{(\boldsymbol{X}_k, B_k), k \geq 1\}$, the first coordinate process $\{\boldsymbol{X}_k, k \geq 1\}$ has the same distribution as a Markov chain κ and initial distribution ν_1 induce. Moreover, since the transition out of $\mathcal{X}^1$ is always governed by ν, disregarding the current state of the chain, the entire $\mathcal{X}^1 \in \mathscr{A}^*$ is an atom for κ^*, and in this manner, a new regenerative cycle starts whenever $B_k = 1$. Finally it is well-known if $\{\boldsymbol{Z}_k, k \geq 1\}$ is a regenerative process with regeneration times $T_0 < T_1 < T_2 < \ldots$, then so is the process $\{f(\boldsymbol{Z}_k), k \geq 1\}$ for any function f. Consequently, given the Markov chain $\{(\boldsymbol{X}_k, B_k), k \geq 1\}$ on $\mathcal{X}^*$, taking f as the projection onto the first coordinate, the process $\{\boldsymbol{X}_k, k \geq 1\}$ must have identical regeneration times. Of course, in this construction, to ensure the process zero-delayed, that is $T_0 = 0$, we need to initialize $\boldsymbol{X}_1 \sim \nu^*$.

In fact, Nummelin's original work proceeds further and shows that whenever κ is Harris recurrent, so is κ^* and in this manner, the atom $\mathcal{X}^1$ is visited infinitely often. In fact it is not difficult to see from [83, Theorem 15.0.1] that if κ is geometrically erdogic, then so is κ^*.

Applying the split chain technique directly as per described can be difficult for arbitrary κ on $\mathcal{X} \times \mathscr{A}$ since we often do not know how to simulate draws from $\check{\kappa}(A \,|\, \boldsymbol{x}^0)$. Luckily the 'retrospective' technique for identifying regeneration times introduced in [86] avoids this complication. Suppose that κ and ν exhibit densities, and so for any $\boldsymbol{x} \in \mathcal{X}$, we can write $\kappa(\boldsymbol{y} \,|\, \boldsymbol{x})$ as a mixture of densities in the following way.

$$\begin{aligned}\kappa(\boldsymbol{y} \,|\, \boldsymbol{x}) &= s(\boldsymbol{x})\nu(\boldsymbol{y}) + (1 - s(\boldsymbol{x}))\frac{\kappa(\boldsymbol{y} \,|\, \boldsymbol{x}) - s(\boldsymbol{x})\nu(\boldsymbol{y})}{1 - s(\boldsymbol{x})} \\ &:= s(\boldsymbol{x})\nu(\boldsymbol{y}) + (1 - s(\boldsymbol{x}))\bar{\nu}(\boldsymbol{y} \,|\, \boldsymbol{x}), \quad \forall \boldsymbol{y} \in \mathcal{X}.\end{aligned}$$

The insight is that given the current state of the chain $\boldsymbol{X}_k$, $B_{k+1} = 1$ if and only if $\boldsymbol{X}_{k+1} \,|\, \boldsymbol{X}_k \sim \nu$. Moreover, since $\kappa(\boldsymbol{y} \,|\, \boldsymbol{x}) \geq s(\boldsymbol{x})\nu(\boldsymbol{y})$, we can simulate $\boldsymbol{X}_{k+1} \sim \kappa(\cdot \,|\, \boldsymbol{X}_k)$ and 'retrospectively' decide whether $\boldsymbol{X}_{k+1} \sim \nu$ or $\boldsymbol{X}_{k+1} \sim \bar{\nu}(\cdot \,|\, \boldsymbol{X}_k)$, which is equivalently to deciding whether $B_k = 1$ or $B_k = 0$, by plugging $\boldsymbol{X}_k$ in to a rejection sampling targeting ν. That is, denoting $r(\boldsymbol{y} \,|\, \boldsymbol{x}) = s(\boldsymbol{x})\nu(\boldsymbol{y})/\kappa(\boldsymbol{y} \,|\, \boldsymbol{x})$, one proceeds with the following algorithm.

Algorithm 6 : MCMC with regeneration

Require: Current state of chain $(\boldsymbol{X}_k, B_k)$.

Set $B_k \leftarrow 0$

Simulate $\boldsymbol{X}_{k+1} \sim \kappa(\cdot \,|\, \boldsymbol{X}_k)$ and $U \sim \mathsf{Unif}(0, 1)$, independently.

if $U \leq r(\boldsymbol{X}_{k+1} \,|\, \boldsymbol{X}_k)$, **then**

$B_{k+1} \leftarrow 1$

return $(\boldsymbol{X}_{k+1}, B_{k+1})$ as the next state of the chain

To ensure $B_0 = 1$, in other words $\boldsymbol{X}_1 \sim \nu$, one can start with any initial $\boldsymbol{z} \in \mathcal{X}$ and proceed with the simulation until one observes a regeneration, and discard all previous draws. Algorithmically speaking, for an arbitrary initial $\boldsymbol{z}_1$, one simulates $\boldsymbol{Z}_2, \boldsymbol{Z}_3, \ldots$ whose dynamic is governed by κ until one gets a draw $\boldsymbol{Z}_* \sim \nu$ and set $\boldsymbol{X}_1 \leftarrow \boldsymbol{Z}_*$.

We now describe how one can systematically establish the minorization condition for the probability transition kernels for Gibbs samplers and independence samplers which are also introduced in [86].

We illustrate the technique for a two-staged Gibbs sampler, but it can be easily extended to a general one. The transition density of a two-staged Gibbs sampler is

$$\kappa(\boldsymbol{y} \,|\, \boldsymbol{x}) = \pi(\boldsymbol{y}_2 \,|\, \boldsymbol{y}_1)\pi(\boldsymbol{y}_1 \,|\, \boldsymbol{x}).$$

The construction begins by considering a hyper-rectangle of the form $[\boldsymbol{c}, \boldsymbol{d}]$ where $\boldsymbol{c}, \boldsymbol{d} \in$, and a choice of a point $\tilde{\boldsymbol{x}} \in [\boldsymbol{c}, \boldsymbol{d}]$ so that

$$\begin{aligned}
\kappa(\boldsymbol{y} \,|\, \boldsymbol{x}) &= \pi(\boldsymbol{y}_2 \,|\, \boldsymbol{y}_1)\pi(\boldsymbol{y}_1 \,|\, \boldsymbol{x}) \\
&= \pi(\boldsymbol{y}_2 \,|\, \boldsymbol{y}_1)\pi(\boldsymbol{y}_1 \,|\, \tilde{\boldsymbol{x}})\frac{\pi(\boldsymbol{y}_1 \,|\, \boldsymbol{x})}{\pi(\boldsymbol{y}_1 \,|\, \tilde{\boldsymbol{x}})} \\
&\geq \frac{\kappa(\boldsymbol{y} \,|\, \tilde{\boldsymbol{x}})}{\varepsilon}\mathbb{I}\{\boldsymbol{y} \in [\boldsymbol{c}, \boldsymbol{d}]\} \times \varepsilon \min_{\boldsymbol{y} \in [\boldsymbol{c}, \boldsymbol{d}]} \frac{\pi(\boldsymbol{y}_1 \,|\, \boldsymbol{x})}{\pi(\boldsymbol{y}_1 \,|\, \tilde{\boldsymbol{x}})}.
\end{aligned}$$

With this, we can choose $\nu(\boldsymbol{y}) = \frac{\kappa(\boldsymbol{y} \,|\, \tilde{\boldsymbol{x}})}{\varepsilon}\mathbb{I}\{\boldsymbol{y} \in [\boldsymbol{c}, \boldsymbol{d}]\}$ where $\varepsilon = \int_{\mathbb{R}^d} \kappa(\boldsymbol{y} \,|\, \tilde{\boldsymbol{x}})\mathbb{I}\{\boldsymbol{y} \in [\boldsymbol{c}, \boldsymbol{d}]\}\, d\boldsymbol{x}$ is the marginalizing constant and $s(\boldsymbol{x}) = \min_{\boldsymbol{y} \in [\boldsymbol{c}, \boldsymbol{d}]} \frac{\pi(\boldsymbol{y}_1 \,|\, \boldsymbol{x})}{\pi(\boldsymbol{y}_1 \,|\, \tilde{\boldsymbol{x}})}$.

Next, let us consider the independence sampler algorithm with a proposal probability density $g(\boldsymbol{y})$ so that the transition kernel is

$$\kappa(d\boldsymbol{y} \,|\, \boldsymbol{x}) = \alpha(\boldsymbol{y} \,|\, \boldsymbol{x})g(\boldsymbol{y})d\boldsymbol{y} + (1 - \alpha^*(\boldsymbol{x}))\delta_{\boldsymbol{x}}(d\boldsymbol{y}),$$

where $\alpha(\boldsymbol{y} \,|\, \boldsymbol{x}) = \min\left\{\frac{\pi(\boldsymbol{y})g(\boldsymbol{x})}{\pi(\boldsymbol{x})g(\boldsymbol{y})}, 1\right\}$. The idea is to find ν and s such that $\alpha(\boldsymbol{y} \,|\, \boldsymbol{x})g(\boldsymbol{y}) \geq \nu(\boldsymbol{y})s(\boldsymbol{x})$ for all $\boldsymbol{x}, \boldsymbol{y}$ and define

$$r(\boldsymbol{y} \,|\, \boldsymbol{x}) = \begin{cases} \frac{s(\boldsymbol{x})\nu(\boldsymbol{y})}{\alpha(\boldsymbol{y} \,|\, \boldsymbol{x})g(\boldsymbol{y})} & \text{if } \boldsymbol{x} \neq \boldsymbol{y} \\ 0, & \text{else.} \end{cases}$$

Since $\alpha(\boldsymbol{y} \,|\, \boldsymbol{x})g(\boldsymbol{y}) \geq \nu(\boldsymbol{y})s(\boldsymbol{x})$, we automatically have $\kappa(d\boldsymbol{y} \,|\, \boldsymbol{x}) \geq \nu(d\boldsymbol{y})s(\boldsymbol{x})$. In terms of practical implementation, we are nesting the regeneration event within the event of accepting a transition.

To find such s and ν, observe that, if we denote $w(\boldsymbol{y}) = \pi(\boldsymbol{y})/g(\boldsymbol{y})$ we have

$$\begin{aligned}\alpha(\boldsymbol{y}\,|\,\boldsymbol{x})g(\boldsymbol{y}) &= \min\left\{\frac{\pi(\boldsymbol{y})g(\boldsymbol{x})}{\pi(\boldsymbol{x})g(\boldsymbol{y})}, 1\right\} g(\boldsymbol{y}) \\ &= \min\left\{\frac{w(\boldsymbol{y})}{w(\boldsymbol{x})}, 1\right\} g(\boldsymbol{y}) \\ &\geq g(\boldsymbol{y}) \min\left\{\frac{w(\boldsymbol{y})}{c}, 1\right\} \min\left\{\frac{c}{w(\boldsymbol{x})}, 1\right\}\end{aligned}$$

for any $c > 0$. It follows that we can choose $\nu(\boldsymbol{y}) = \varepsilon^{-1} g(\boldsymbol{y}) \min\left\{\frac{w(\boldsymbol{y})}{c}, 1\right\}$ and $s(\boldsymbol{x}) = \varepsilon \min\left\{\frac{c}{w(\boldsymbol{x})}, 1\right\}$, where $\varepsilon = \int g(\boldsymbol{y}) \min\left\{\frac{w(\boldsymbol{y})}{c}, 1\right\} d\boldsymbol{y}$. Finally, given the current state $\boldsymbol{X}$, and condition on the event where $\boldsymbol{Y} \sim g$ is accepted as the next state, the probability that $\boldsymbol{Y} \sim \nu$ is

$$\frac{\min\left\{\frac{w(\boldsymbol{y})}{c}, 1\right\} \min\left\{\frac{c}{w(\boldsymbol{x})}, 1\right\}}{\min\left\{\frac{w(\boldsymbol{y})}{w(\boldsymbol{x})}, 1\right\}}.$$

Notice that the choice of s in both the Gibbs sampler and the independence sampler described above involves some unknown normalizing constant ε. However one does not actually need to compute it because in either cases the expression for r, which is all that is needed in the practical implementations, only depends on the product $s\nu$, which does not involve ε.

To this ends are now ready exploit the rich theory of regenerative processes can offer in the analysis of Markov chain sampling studied in this thesis. We shall conclude this section by giving few more results we will refer to in Chapter 5.

Firstly, we have the well-known Wald's identity and Wald's second moment identity [113, 114].

Theorem 4.2.5 (Wald's identities). Let τ be an a.s. finite stopping time with respect to a filtration $\{\mathscr{A}_k, k \geq 0\}$ and let $Z_1, Z_2, \ldots$ be iid random variables such that Z_k is $\mathscr{A}_k$-measurable and $Z_{k+1}, Z_{k+2}, \ldots$ are independent of $\mathscr{A}_k$. Denoting $S_n = Z_1 + \ldots + Z_n$, it follows that

1. if either $\mathbb{E}|Z_1| < \infty$ and $\mathbb{E}\tau < \infty$, or $Z_1 \geq 0$, then $\mathbb{E}[S_\tau] = \mathbb{E}Z_1 \mathbb{E}\tau$.
2. if $\mathrm{Var}(Z_1) < \infty$ and $\mathbb{E}\tau < \infty$, then $\mathbb{E}[S_\tau - \tau \mathbb{E}Z_1]^2 = \mathrm{Var}(Z_1)\mathbb{E}\tau$.

In our context, we will often use $N(t) := \inf\{n \,|\, T_n > t\}$ as our stopping time while Z_k will be some measurable functionals of the k-th cycle $\{\{\boldsymbol{X}_k, T_{j-1} < k \leq T_j\}, M_j\}$. The process $N(t)$ can be thought of as the minimum number of regeneration cycles needed to surpass time t. It follows that in our zero-delayed setting where $T_0 = 0$, we have $N(0) = 1$. In fact, the counting process $\{N(t) - 1\}$, which is the number of renewals (regenerations) up until time t is sometimes called 'the renewal process' while the process $\{T_k, k \geq 0\}$,

which is what we refer to as 'the renewal process', is sometimes called 'the renewal time process'. Finally, note that in our context, the time parameter t takes value on $\mathbb{N}$, so one may prefer the subscript notation N_t for the process $N(t)$. However, to avoid stacking of subscripts in the next chapter, we have decided to use the notation $N(t)$.

Next, given a renewal process $\{T_k, k \geq 0\}$, we define its residual life time process as $R(t) = T_{N(t)} - t$. In our context, it can be thought of as the the number of steps until the current regenerative cycle ends. The results due to [78] gives us bounds on the p-moments of $R(t)$.

Theorem 4.2.6 (Lorden's inequalities). Let $\{T_k, k \geq 0\}$ be a renewal process with inter-arrival times $M_1, M_2, \ldots$, and let $R(t)$ be its corresponding residual life time process. Denoting the p-th moment of the inter-arrival times of a renewal process by $m_p = \mathbb{E}M_1^p$, the moments of $R(t)$ satisfy:

$$\mathbb{E}R^p(t) \leq \begin{cases} \frac{m_2}{m_1}, & \text{if } p = 1, \\ \frac{p+2}{p+1}\frac{m_{p+1}}{m_1}, & \text{if } p > 1. \end{cases}$$

In our context, the inter-arrival times $M_1, M_2, \ldots$ are the regenerative cycle lengths. A fact that we will repeatedly use is that once we have observed $M_1, M_2, \ldots$ in our simulation output, there is a natural estimator $\hat{m}_p = \frac{1}{n}\sum_{k=1}^n M_k^p$ for m_p.

The next result is [3, Chapter 7,Corollary 1.5], which can be viewed as a statement concerning the relationship between the regenerative cycles and the limiting distribution of a Markov chain. Although the result holds for general regenerative processes (not necessarily Markov chains), we shall state it in terms of Markov chains.

Theorem 4.2.7. Let $\{\boldsymbol{X}_k, k \geq 1\}$ be a Markov chain with limiting distribution π with regeneration times $\{T_k, k \geq 0\}$, $T_0 = 0$. For any function f such that $\int f(\boldsymbol{x})\pi(d\boldsymbol{x}) < \infty$, we have for all $r = 1, 2, \ldots$,

$$\int f(\boldsymbol{x})\pi(d\boldsymbol{x}) = \frac{1}{m_1}\mathbb{E}\sum_{k=1}^{M_r} f(\boldsymbol{X}_{T_r+k})$$

where $m_1 = \mathbb{E}M_1$ is the expected length of the regenerative cycle.

In other words, the average of f evaluated along each regenerative cycle, on average, equals to the expectation of f with respect to the Markov chain's limiting distribution. In particular, in our proofs for Theorems 5.2.2 and 5.2.3, given in Appendices A.2 and A.3,

we will choose $f(\boldsymbol{x}) = \mathbb{I}\{\boldsymbol{x} \in A\}$ for some $A \in \mathscr{A}$ so that we have the relation

$$\pi(A)\mathbb{E}M_r = \mathbb{E}\sum_{k=1}^{M_r} \mathbb{I}\{\boldsymbol{X}_{T_r+k} \in A\}.$$

4.3 Concluding remarks for this chapter

In this chapter, we have introduced notions and notations that we shall use in the upcoming chapters. In particular, we have discussed the technique of split chain and retrospective identification of regeneration times of a Markov chain. In this manner, we can identify the regeneration times for many Markov chain samplers (at least for the examples we consider in this thesis). This is important to the novel Markov chain sampling diagnostics we propose in Chapter 5. The diagnostics use the observed regenerative cycle tour lengths to estimate error bounds for the Markov chain. Since our proposed methods use the simulation output to do these estimations, they still fall under the 'diagnostics framework'. However, we shall argue that these novel diagnostics have solid theoretical footings, and are much easier to implement than a purely analytical approach.

Chapter 5

Regenerative Markov chain sampler and error analysis

5.1 Introduction to this chapter

Most inference in Bayesian statistics requires one to take expectations with respect to π, the posterior distribution. These expectations are intractable and call for Monte Carlo statistical methods, in particular, Markov chain Monte Carlo (MCMC).

MCMC approximates draws from π by the sample path of a simulated Markov chain $\{\boldsymbol{X}_k, k \geq 1\}$ on $\mathbb{R}^d$ whose limiting distribution is π. Moreover, given a measurable function h, one often estimates $q := \int h(\boldsymbol{x})\pi(d\boldsymbol{x})$ with averages such as $\hat{q}_t := \frac{1}{t}\sum_{k=1}^{t} h(\boldsymbol{X}_k)$. Despite being computationally attractive, evaluating the performance of a MCMC method is difficult due to the dependence between each draw and the fact that often $\boldsymbol{X}_1 \not\sim \pi$.

To evaluate the performance of a given MCMC method, one naturally asks the following questions. How close is the probability law of $\boldsymbol{X}_t$ to π for different values of t? How much closer does it become for each extra step? For $t < \infty$, how well does $\hat{q}_t$ approximate q on average? How variable is it? These questions motivate our work in this chapter. In this chapter we propose novel assessment tools to address these questions for the cases where one can identify the underlying regeneration times of geometrically ergodic Markov chains. Specifically, let κ_t be a t-th step probability transition kernel on $\mathcal{X} \subseteq \mathbb{R}^d$, and we assume that the single step transition kernel satisfies the minorization condition, that is there is a function $s : \mathcal{X} \to [0, 1]$ and a probability measure on $\mathcal{X}$ for which $\kappa(d\boldsymbol{y} \,|\, \boldsymbol{x}) \geq s(\boldsymbol{x})\nu(d\boldsymbol{y})$ for all $\boldsymbol{x} \in \mathcal{X}$.

Firstly, we derive an asymptotic bound on $\delta_t := \|\mathbb{E}[\kappa_t(\cdot \,|\, \boldsymbol{X}_1)] - \pi\|_{\mathrm{TV}}$, where $\boldsymbol{X}_1 \sim \nu$ and $\boldsymbol{X}_t \sim \mathbb{E}[\kappa_t(\cdot \,|\, \boldsymbol{X}_1)] := \int \kappa_t(\cdot \,|\, \boldsymbol{x})\nu(d\boldsymbol{x})$. In other words, δ_t is the total variation distance between the t-th step Markov transition kernel marginalized over the initial distribution. Such a bound allows practitioners to quantify convergence speeds of Markov chains and provide estimates for burn-in periods [64].

Moreover, recall from Chapter 4 that we denote $\{T_k, k \geq 0\}$ as the regeneration times for the Markov chain $\{\boldsymbol{X}_k, k \geq 1\}$ and that $N(t) = \inf\{n \,|\, T_n > t\}$. Motivated by the results in [71], we also consider a regime where one simulates until $T_{N(t)}$ and resample with uniform probability from the path's history. In other words, given a pre-specified

$t > 0$, one simulates the collection of random variables $\{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_{T_{N(t)}}\}$ and resample $\boldsymbol{X}^{\text{reg-seq}} \in \{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_{T_{N(t)}}\}$ with uniform probability. In this setting, we derive two bounds (one asymptotic, and one non-asymptotic) on the total variation distance between the distribution of $\boldsymbol{X}^{\text{reg-seq}}$ and π. Motivated by the terminology *'regenerative-sequential estimator'* in [71], we denote this total variation distance by $\delta_t^{\text{reg-seq}}$, and it is formally defined by the formula

$$\begin{aligned} \delta_t^{\text{reg-seq}} &:= \|\mathbb{E}[\kappa_{T_{N(t)}}(\cdot \mid \boldsymbol{X}_1)] - \pi\|_{\text{TV}}, \\ \text{where } \mathbb{E}[\kappa_{T_{N(t)}}(\cdot \mid \boldsymbol{X}_1)] &:= \frac{1}{T_{N(t)}} \sum_{k=1}^{T_{N(t)}} \mathbb{E}[\kappa_k(\cdot \mid \boldsymbol{X}_1)]. \end{aligned} \tag{5.1}$$

These bounds can be thought of as extensions to the non-asymptotic MSE bound on the estimator $\hat{q}_{T_{N(t)}}$ derived in [71] and Markov chain analogues to the bounds on the generalized splitting algorithm derived in [13].

We point out in advance that our bounds depend on some unknown constants, and thus they are not computable exactly. However, these constants can be easily estimated from the sample path itself provided that we can identify the regeneration times.

Secondly, we propose a novel visual aid for assessing the convergence of a Markov chain whenever we can identify the regeneration times. Visual aids such as trace plots and autocorrelation plots of coordinate processes are very common in practice. (See for example [21, 38, 92, 79] and the popular software WinBUGS.) Such plots project a Markov process in high-dimensions to a univariate process and then display the autocorrelation/trace of this process to determine whether the sampler is working satisfactorily. Typically, the greatest difficulty is in determining what kind of projection to use to summarize the high-dimensional process via a one-dimensional one. Almost any projection will lose essential information about the process and thus convergence of the low-dimensional process is not a sufficient condition for the convergence of the high-dimensional process.

However, our novel visual aid provides a loss-less dimension-reduction projection, in the sense that the convergence of the low-dimensional process is sufficient to guarantee the convergence of the high-dimensional process. Thus, we can examine the convergence of the underlying Markov chain without the need of some arbitrary lossy projection such as the coordinate projections.

Finally, we also consider quantifying the variability of some point estimator. The authors of [71] derive a non-asymptotic bound on the MSE of $\hat{q}_{T_{N(t)}}$, and from there, they can estimate a non-asymptotic confidence region for $\hat{q}_{T_{N(t)}}$ by running parallel Markov chain. (We give a brief description to this later in the section.) We also recommend the estimator $\hat{q}_{T_{N(t)}}$ to practitioners. However, in contrast to the proposal in [71], we estimate

the non-asymptotic bound on the MSE from a single MCMC simulation output directly (no parallel Markov chains are necessary).

The problems we study in this Chapter are actually well studied in the literature, but existing works do not provide the same solution as ours. We now provide a brief review on existing works and compare them to our proposed methods.

Firstly, if the initial state is indexed by 0, an analytic bound on $\|\kappa(\cdot \,|\, \boldsymbol{X}_0) - \pi\|_{\mathrm{TV}}$ for all initial distribution ν_0 is derived in the highly cited work [98]. This bound requires one to establish the following two conditions.

1. There exists a function $V : \mathbb{R}^d \to [0, \infty)$, constants $\lambda < 1$, $b < \infty$ such that

$$\mathbb{E}(V(\boldsymbol{X}_1) \,|\, \boldsymbol{X}_0 = \boldsymbol{x}) \leq \lambda V(\boldsymbol{x}) + b, \quad x \in \mathbb{R}^d$$

2. The set $\{\boldsymbol{x} \in \mathbb{R}^d \,|\, V(\boldsymbol{x}) \leq c\}$ is small, for some $c > 2b/(1-\lambda)$, that is there exists a probability measure ν on $\mathbb{R}^d$ such that

$$\kappa(\cdot \,|\, \boldsymbol{x}) \geq \varepsilon \nu$$

   for all $x \in \{\boldsymbol{x} \in \mathbb{R}^d : V(\boldsymbol{x}) \leq d\}$.

The first condition is known as drift, and the second one is known as the associated minorization. With these, [98] asserts the following theorem.

Theorem 5.1.1. If κ satisfies the conditions mentioned, then for any $r \in (0, 1)$

$$\|\kappa_t(\cdot \,|\, \boldsymbol{X}_0) - \pi\| \leq (1-\varepsilon)^{rt} + (\alpha^{-(1-r)} A^r)^t \left(1 + \frac{b}{1-\lambda} + \mathbb{E}(V(\boldsymbol{X}_0))\right)$$

where $\boldsymbol{X}_0 \sim \nu_0$

$$\alpha^{-1} = \frac{1 + 2b + \lambda c}{1 + c} < 1, \quad A = 1 + 2(\lambda c + b)$$

for all initial distribution ν_0.

In fact, if κ is ϕ-irreducible, aperiodic with stationary measure π, the drift condition on any small set C, (not necessarily on $C = \{V(\boldsymbol{x}) \leq d\}$ for some d) is sufficient for geometric ergodicity (see [83, chapter 15] and [96]), hence the existence of a CLT for $\hat{q}_t$, provided that $\int [q(\boldsymbol{x})]^{2+\delta} \pi(d\boldsymbol{x}) < \infty$ for some $\delta > 0$, or if κ satisfies detailed balance with respect to π then $\int [q(\boldsymbol{x})]^2 \pi(d\boldsymbol{x}) < \infty$ guarantees CLT [16, 94]. A detailed survey on the various conditions under which CLT holds for $\hat{q}_t$ is given in [66].

This techniques has been applied for some Markov chain samplers targeting Bayesian posteriors such as the Gibbs samplers for Bayesian Lasso [68], Bayesian shrinkage models

[90], Bayesian penalized regression models [111], Bayesian quantile regression [67] and Bayesian general linear mixed models [97]. In a similar manner, [24, 32, 33, 61] study ergodicity of Metropolis-Hastings type algorithms.

An extension for time inhomogeneous Markov chains is also provided in [25] and they also consider the more general f-total variation distance. Moreover the notion of a generalized drift condition from which one can proceed similar derivations for the total variation distance bound is introduced in [115].

Using Theorem 5.1.1 to assess the convergence of a Markov chain has the advantage of being analytic in nature. However, many MCMC practitioners rely on simple 'convergence diagnostics' such as the autocorrelation plots of coordinate processes or checking whether parallel Markov chains approximately converge to the same distribution when initialized differently. (See for example [21, 38, 92, 79] and the popular software WinBUGS.) We suspect that this is because the assumptions Theorem 5.1.1 requires are too difficult to verify. Another issue we observe in our study is that the bound in Theorem 5.1.1 can suffer from numerical instability (see Section 5.5).

Many authors have also proposed sophisticated convergence diagnostics that use the simulation output. These approaches give justifiable diagnostics, and at the same time, they remain relatively simple to implement. Our contributions in this chapter fall under this framework. Existing works that also go under this framework are as follows.

The method proposed [8] is to segment the Markov chain into small batches and formulate kernel density approximations for each batch. From there, they can diagnose the convergence of the chain by estimating the Hellinger distances across each batches. Another notable work that relies on kernel density estimation is in [23] where they estimate the symmetric Kullback Leibler divergence of two parallel chains using kernel density approximations. From there, they formulate a hypothesis testing framework to answer the question on whether the two chains are sufficiently close. These kernel density approximation approaches are in contrast to our approach where we work directly with $\mathbb{E}[\kappa_t(\cdot \,|\, \boldsymbol{X}_1)]$ for $\boldsymbol{X}_1 \sim \nu$.

Approaches that do not rely on kernel density approximations are also proposed in for example [54]. They infer the convergence of the Markov chain by estimating the Kullback Leibler divergence between the Markov chain and a subsequence of the chain using numeric integration. Another example is in [110] where they estimate $\sup_{\boldsymbol{x}_1} \|\kappa_t(\cdot \,|\, \boldsymbol{x}_1) - \pi\|_{\mathrm{TV}}$ from the simulation output. Their estimation relies on the integral form of the total variation distance $\|\nu - \mu\|_{\mathrm{TV}} = \frac{1}{2}\int |\mu(\boldsymbol{x}) - \nu(\boldsymbol{x})|\, d\boldsymbol{x}$, and from there they estimate the integral by clustering the simulation output when parallel Markov chains simulated. Here, $\mu(\boldsymbol{x})$ and $\nu(\boldsymbol{x})$ are the density functions for measures μ and ν (with respect to Lebesgue measure) evaluated at $\boldsymbol{x} \in \mathbb{R}^d$. These approaches that estimate some integrals directly

are fundamentally different to our proposal. In this chapter, we derive analytic bounds on δ_t (and $\delta_t^{\text{reg-seq}}$), and the estimation comes in to play when we estimate some unknown constants in our bound.

Our proposed method is quite similar to the work in [63] where they also propose estimating an unknown constant in a bound they have derived using the simulation output. However their approach runs two parallel chains, and instead of considering regeneration, they consider the frequency with which the two chains couple. This coupling approach between two chains is extended to 'L-lag coupling' in [7]. This approach is in contrast to our approach where we propose running a single chain, and we consider its underlying regeneration times. Moreover, their bound is for the total variation distance between the joint distribution of sample paths and a product of the limiting distribution but we study the distributions $\mathbb{E}[\kappa_t(\cdot \mid \boldsymbol{X}_1)]$ and $\mathbb{E}[\kappa_{T_{N(t)}}(\cdot \mid \boldsymbol{X}_1)]$.

Estimating a burn-in period is a natural application of these bounds on the distance between the distribution of the Markov chain and its limiting distribution, whether it be analytic or estimated [64]. Typically, one can choose a burn-in period such that the Markov chain 'is initialized' sufficiently close to the limiting distribution using these bounds. However, other approaches and criteria for estimating a burn-in period have been proposed. For example [93] formulates the problem in terms of Bayesian posterior estimation and considers estimating a quantile of $h(\boldsymbol{X}_\infty)$, where $h : \mathbb{R}^d \to \mathbb{R}$ is a function of interest and $\boldsymbol{X}_\infty \sim \pi$, to some given precision. Other examples include the work in [100, 101], in which explicit bounds on the mean-squared error of $\hat{q}$ is derived for some class of h, and consider burn-in periods for which the mean-squared error is sufficiently small.

The other well studied problem is to quantify the variability of $\hat{q}_t$. A common approach is that when $\hat{q}_t$ admits a CLT $\sqrt{t}(\hat{q}_t - q) \xrightarrow{d} \mathsf{N}(0, \gamma_h^2)$ then γ_h^2 captures the asymptotic variability of $\hat{q}_t$. An estimate for $\gamma_h^2 = \int_{\mathbb{R}^d}(h(\boldsymbol{x}) - q)^2 \, \pi(d\boldsymbol{x})$ allows one to construct some asymptotic confidence interval for $\hat{q}_t$ and it is natural to stop the Markov chain simulation only when this interval is desirably small for example in [30, 64, 45, 75, 74, 73].

A popular estimator for γ_h^2 is the batch-means variance estimator, where the simulation assumes to terminates at $t = ab$ for some positive integers a and b, so that we can segment the sample path into a batches of length b. (See [75] and the references therein.) The estimator then proceeds with the premise that the averages within each individual batch are roughly independent. Formally, denoting

$$B_r = \frac{1}{b} \sum_{k=(r-1)b+1}^{rb} h(\boldsymbol{X}_k), \quad \text{for } r = 1, \ldots, a$$

the batch-means variance estimator is

$$\hat{\gamma}^2_{h,\mathrm{BM}} = \frac{b}{a-1}\sum_{r=1}^{a}(B_r - \hat{q}_t)^2.$$

Its consistency has been studied extensively for example [49] shows that if a is held constant, then $\hat{\gamma}^2_{h,\mathrm{BM}}$ is not even a weakly consistent estimator for γ^2_h as $t \to \infty$.

The first positive result on its consistency is given in [22], which is reformulated in [65] as follows.

Theorem 5.1.2. Let $\{\boldsymbol{X}_k, k \geq 1\}$ be an uniformly ergodic, positive Harris reccurent Markov chain wiht invariante distribution π. Suppose that $\mathbb{E}_\pi|h|^{2+\delta} < \infty$ for some $\delta > 0$, and (a) $a_t \to \infty$ as $t \to \infty$; (b) $b_t \to \infty$ and $b_t/t \to 0$ as $t \to \infty$; (c) $b_t^{-1}t^{1-2\alpha}\log t \to 0$ as $t \to \infty$, where $\alpha \in (0, \delta/(24+12\delta))$; and (d) there exists a constant $c \geq 1$ such that $\sum_t (b_t/t)^c < \infty$, then as $t \to \infty$, $\hat{\gamma}^2_{h,\mathrm{BM}} \to \gamma^2_h$ with probability 1.

The uniform ergodicity assumption is indeed quite strong and fails to hold in many practical Markov chain samplers, and to this end, an extension to geometrically ergodic Markov chain is also given in [65] as follows.

Theorem 5.1.3. Let $\{\boldsymbol{X}_k, k \geq 1\}$ be a geometrically ergodic, positive Harris recurrent Markov chain with invariant distribution π. Suppose that $\mathbb{E}_\pi|h|^{4+\delta} < \infty$ for some $\delta > 0$, and (a) $a_t \to \infty$ as $t \to \infty$; (b) $b_t \to \infty$ and $b_t/t \to 0$ as $t \to \infty$; (c) $b_t^{-1}t^{2\alpha}[\log t]^3 \to 0$ as $t \to \infty$, where $\alpha = 1/(2+\delta)$; and (d) there exists a constant $c \geq 1$ such that $\sum_t (b_t/t)^c < \infty$, then as $t \to \infty$, $\hat{\gamma}^2_{h,\mathrm{BM}} \to \gamma^2_h$ with probability 1.

The moment condition $\mathbb{E}_\pi|h|^{4+\delta}$ for some $\delta > 0$, is relaxed to $\mathbb{E}_\pi|h|^{2+\delta+\epsilon}$ for some $\delta, \epsilon > 0$ in [6].

The mean-squared-error consistency of the batch-means variance estimator is also studied in [31], and by minimizing the mean-squared-error of $\hat{\gamma}^2_{h,\mathrm{BM}}$, they show that the optimal number of batches should increase with order $t^{1/3}$. Finally, variations and extensions can be found in the literature. such as the overlapping batch-means discussed in [31], while [77] concerns an analogous batch-means asymptotic covariance estimator when h is vector-valued.

On the other hand, an alternative estimator when the regeneration times of the Markov chain is identifiable. Let $0 = T_0 < T_1 < T_2 < \ldots$ be the random regeneration times so that for all $r = 1, 2, \ldots$, $\{\{\boldsymbol{X}_k, T_{r-1} < k \leq T_r\}, M_r\}$ are iid where $M_r = T_r - T_{r-1}$. Denoting

the sum of $h(\boldsymbol{X}_k)$ over the r-th regenerative cycle as

$$H_r = h(\boldsymbol{X}_{T_{r-1}}) + h(\boldsymbol{X}_{T_{r-1}+1}) + h(\boldsymbol{X}_{T_{r-1}+2}) + \ldots + h(\boldsymbol{X}_{T_r-1}) = \sum_{k=T_{r-1}+1}^{T_r} h(\boldsymbol{X}_k),$$

then $\hat{q}_{T_n} = \frac{1}{T_n}\sum_{k=1}^{T_n} h(\boldsymbol{X}_k) = \frac{\sum_{r=1}^{n} H_r}{\sum_{r=1}^{n} M_r}$ is an alternative estimator for q. Here, one specifies number of regenerative cycles and consequently the total simulation time is random. When the Markov chain is geometrically ergodic and that $\mathbb{E}_\pi|h|^{2+p} < \infty$ for some $p > 0$, $\hat{q}_{T_n}$ exhibits a CLT $\sqrt{n\mathbb{E}M_1}(\hat{q}_{T_n} - q) \xrightarrow{\mathrm{d}} \mathsf{N}(0, \gamma_h^2)$ [6, 65]. Indeed $\hat{q}_{T_n}$ is asymptotically equivalent to $\hat{q}_t$, however the iid decomposition renders simple estimator

$$\hat{\gamma}_h^2 = \frac{\frac{1}{n}\sum(H_r - \hat{q}_{T_n} M_r)^2}{(\bar{M})^2} \tag{5.2}$$

and its consistency is established in [48, 57, 65]. Here $\bar{M}$ is the sample average of $M_1, \ldots, M_n$.

Note that these variance estimators only capture the asymptotic variability. To the best of our knowledge, the first non-asymptotic results for the MSE of $\hat{q}_{T_n}$ are established in [71, 72]. Moreover, [71] proposes a new estimator $\hat{q}_{T_{N(t)}}$, in which the practitioner specifies a deterministic t and upon reaching it, the simulation continues until the current regenerative cycle ends. They further suggest running parallel simulations and apply the 'median trick' to guarantee that realizations of this estimator is sufficiently concentrated around the true value with a given level of confidence. Formally, for some positive odd integer l, they propose running l parallel Markov chains so that there are l realizations $\hat{q}^1_{T_{N(t)}}, \ldots, \hat{q}^l_{T_{N(t)}}$. They then propose the estimator

$$\hat{q}_{T_{N(t)},l} = \text{med}\ (\hat{q}^1_{T_{N(t)}}, \ldots, \hat{q}^l_{T_{N(t)}})$$

and provide a formula to estimate (optimal) t and l for which

$$\mathbb{P}(|\hat{q}_{T_{N(t)},l} - q| > \epsilon) = \alpha$$

for any given $\epsilon > 0$ and α. Following this, estimators with non-asymptotic relative errors, instead of absolute errors, has also been studied in the literature [35].

In this chapter, we also recommend the estimator $\hat{q}_{T_{N(t)}}$ to practitioners whenever regeneration times can be identified. This is because the MSE of this estimator exhibits a non-asymptotic bound, whose constant depends on the mean tour length of the regeneration cycles. However, instead of suggesting running parallel chains and applying the 'median trick', we propose estimating the constants involved in the MSE directly from

the simulation output of a single MCMC run. We also demonstrate how we can quantify the asymptotic error of these estimations in a single MCMC run in Section 5.6. In this manner, whenever regeneration is easily identifiable, we can systematically address the convergence issues of the MCMC sampler in a way that is more accessible to practitioners, yet remains on safe theoretical footing.

We argue that our approach is not as restrictive as it may seem. Firstly, many practical Markov chains are indeed geometrically ergodic. Next the notion of split chains [4, 87] and retrospective identification of regeneration times [86] allow practitioners to systematically identify the regeneration times of the Markov chain. Moreover the notion of an 'artificial atom' is introduced in [14], equipping practitioners with additional tools for introducing regeneration times in their Markov chain samplers.

In summary, our contributions include an estimate for the total variation distance bound, an the extension of the MSE bound of [71] to the estimation of the total variation distance, a novel dimension-reduction visual diagnostic tool for the total variation convergence of an MCMC sampler, and an alternative sample-based estimate of the MSE bound in [71],

The rest of the chapter is structured as follows. In Section 5.2 we discuss our contribution introduced above. We then apply our proposed methods on some simple toy examples in Section 5.3. Following that, in Section 5.4 and 5.5 we apply our proposed methods to the Park & Casella Gibbs sampler [91] for the Bayesian Lasso model and an independence sampler for the Probit model [26]. We then give some concluding remarks.

5.2 Output diagnostics in regenerative Markov chains

We shall present our novel diagnostics for Markov chains in this section. We assume that the Markov chain has a stationary probability measure π and is geometrically ergodic. Moreover, we assume that the probability transition kernel κ satisfies the minorization condition, that is $\kappa(d\boldsymbol{y} \,|\, \boldsymbol{x}) \geq s(\boldsymbol{x})\nu(d\boldsymbol{y})$ for some function $s : \mathcal{X} \to [0, 1]$ and probability measure ν on $(\mathcal{X}, \mathscr{A})$. We denote $\{T_k, k \geq 0\}$ as the underlying (zero-delayed) regeneration times and $M_r = T_r - T_{r-1}$ as the r-th tour length.

5.2.1 Upper bounds on total variation distance

Since we assume that the Markov chain is a zero-delayed regenerative process, we require that $\boldsymbol{X}_1 \sim \nu$. Moreover, the marginalized t-step transition kernel is obtained by taking the expectation with respect to $\boldsymbol{X}_1$, namely, $\mathbb{E}[\kappa_t(A \,|\, \boldsymbol{X}_1)]$. In this way, a theoretically sound assessment of convergence, is to construct an estimate for

$$\delta_t := \|\mathbb{E}[\kappa_t(\cdot \,|\, \boldsymbol{X}_1)] - \pi\|_{\mathrm{TV}},$$

and examine how fast it decays with respect to t. Denoting $m_k = \mathbb{E}M_1^k$, our key insight is that the bias properties of regenerative estimators [46] allow us to bound the total variation distance, as follows.

Theorem 5.2.1 (Asymptotic total variation bound for MCMC). Let $\kappa_t(\cdot \,|\, \boldsymbol{X}_1)$ with $\boldsymbol{X}_1 \sim \nu$ be the t-step transition kernel of a geometrically ergodic Markov chain with limiting density π. Then, we have (for some constant $\varepsilon > 0$)

$$\delta_t := \|\mathbb{E}[\kappa_t(\cdot \,|\, \boldsymbol{X}_1)] - \pi\|_{\mathrm{TV}} \leq \frac{c_1}{t} + \mathscr{O}(\exp(-\varepsilon t)),$$

where $c_1 = \frac{m_2+m_1}{2m_1}$.

The proof is given in the Appendix A.1. The key feature of this bound is that the constant c_1 only depends on m_1 and m_2, which can be estimated from simulation using the iid realizations of $M_1, M_2, \ldots$ (for example $\hat{m}_k = \sum_{k=1}^{N(t)} M^k/N(t)$), and in this manner, we can approximately assess the convergence of the Markov chain, once we have a good estimate of this constant.

Note that we propose to estimate c_1, and unknown constants introduced in later sections, using iid realizations of $M_1, M_2, \ldots$. An obvious approach to quantifying the uncertainty in these constants is to run parallel chains to estimate confidence sets for these constants. Alternatively, we can exploit the fact the iid nature of $M_1, M_2, \ldots$, so that in a single run of the Markov chain, we can formulate partial M-estimators for these constants and approximate their corresponding asymptotic variance (see Section 5.6).

We also note that the bound on δ_t can severely overestimate the true TV distance, because the MCMC samplers that we consider converge geometrically fast, therefore any polynomially decaying estimator is suboptimal. If it is possible to precisely quantify the rate of geometric convergence, then that will be preferable. In reality, the rate of geometric decay of an MCMC sampler is typically unknown and difficult to estimate from simulation. In contrast, our bounds on δ_t (and $\delta_t^{\text{reg-seq}}$) decay only at a polynomial rates, but their rates of decay can be easily estimated from the simulation.

One may also consider analytical upper bounds that use the technique developed in [98, Theorem 12]. However, our simulation experience (shown later on) reveals that these bounds, in particular [68, Proposition 4], require large sample sizes to eventually overtake the simpler linear bound. Moreover, our simulation experience reveals to us that this analytic bound can exhibit numerical round-off issues.

Finally to ensure $\boldsymbol{X}_1 \sim \nu$ one can simulate the Markov chain starting with some arbitrary initial $\boldsymbol{w} \in \mathcal{X}$ and discard the samples until one observes the first instance of

regeneration as discussed in [86]. Our simulation experience reveals the cost for this is generally small when the associated parameters are appropriately tuned.

An advantage of being able to compute an upper bound for δ_t is that it motivates a natural burn-in estimator. Let us define the ϵ-*burn-in* of a Markov chain with transition kernel κ as the smallest t for which $\|\mathbb{E}[\kappa_t(\cdot \,|\, \boldsymbol{X}_1)] - \pi\|_{\mathrm{TV}} < \epsilon$, that is:

$$t_{\mathrm{b}} := \min\{t : \|\mathbb{E}[\kappa_t(\cdot \,|\, \boldsymbol{X}_1)] - \pi\|_{\mathrm{TV}} < \epsilon\}.$$

Hence, a key insight from the theorem above is that an asymptotic upper bound for the ϵ-burn-in, t_{b}, is $\lceil c_1/\epsilon \rceil$, It follows that the ϵ-burn-in is:

$$\left\lceil \frac{\sum_{k=1}^{N(t)} M_k^2 + \sum_{k=1}^{N(t)} M_k}{2\epsilon \sum_{k=1}^{N(t)} M_k} \right\rceil . \tag{5.3}$$

This estimator can admittedly be quite conservative as it relies on an upper bound of the total variation distance, not on the actual total variation distance.

Suppose now the sampling scheme is first budgeted for t-steps but is allowed to complete its current regenerative cycle, following that we resample from its history with equal probability. We shall denote the distribution of this by $\mathbb{E}[\kappa_{T_{N(t)}}(\cdot \,|\, \boldsymbol{X}_1)]$ as in (5.1), and in this manner, we can study $\delta_t^{\text{reg-seq}}$, the convergence rate of $\mathbb{E}[\kappa_{T_{N(t)}}(\cdot \,|\, \boldsymbol{X}_1)]$ to π. In such a simulation scheme we have the following two bounds regarding the total variation distance between the Markov chain and the limiting distribution. The first one is a non-asymptotic bound that decays at $\mathscr{O}(t^{-3/2})$ and the second one is an asymptotic bound that asymptotically decays at $\mathscr{O}(t^{-2})$. The idea and the proofs for these bounds closely follow the generalized splitting algorithm studied in [13], however in this thesis we are concerned with geometrically erdogic Markov chains.

Theorem 5.2.2 (Non-asymptotic TV bound for MCMC sampling until cycle ends). Let $\kappa_t(\cdot \,|\, \boldsymbol{X}_1)$ with $\boldsymbol{X}_1 \sim \nu$ be the t-step transition kernel of a geometrically ergodic Markov chain with limiting density π. For some $\omega > 0$ (typically unknown)

$$\delta_t^{\text{reg-seq}} := \|\mathbb{E}[\kappa_{T_{N(t)}}(\cdot \,|\, \boldsymbol{X}_1)] - \pi\|_{\mathrm{TV}} \leq c_1(t) t^{-3/2},$$

where $c_1(t) := \sqrt{(4/3) m_3 m_2 (m_1 + m_2/t)} m_1^{-3/2}$ is bounded uniformly in t.

Theorem 5.2.3 (Asymptotic TV bound for MCMC sampling until cycle ends). Let $\kappa_t(\cdot \mid \boldsymbol{X}_1)$ with $\boldsymbol{X}_1 \sim \nu$ be the t-step transition kernel of a geometrically ergodic Markov chain with limiting density π. For some $\omega > 0$ (typically unknown)

$$\delta_t^{\text{reg-seq}} := \|\mathbb{E}[\kappa_{T_{N(t)}}(\cdot \mid \boldsymbol{X}_1)] - \pi\|_{\text{TV}} \leq \frac{c_1}{t^2} + \frac{c_2(t)}{t^{5/2}} + \mathscr{O}(\exp(-\omega t))$$

where

$$c_1 := \frac{\mathbb{E}|M - 1 - 2r|M^2}{2m_1}, \quad r = \frac{m_2 + m_1}{2m_1}$$

and

$$c_2(t) := \frac{\sqrt{(6/5)(m_1 + m_2/t)m_2 m_5}}{m_1},$$

is bounded uniformly in t.

In this section we have provided simple quantitative estimates of the convergence of the MCMC sampler, in the next section we provide a simple qualitative convergence assessment via a single autocorrelation plot.

5.2.2 Global convergence diagnostic plot

As already mentioned, most users assess the performance of MCMC samplers by displaying a number of autocorrelation or trace plots of some functions of the output. Nonetheless, almost any projection will lose essential information about the process and thus convergence of the low-dimensional process is not a sufficient condition for the convergence of the high-dimensional process.

In this section we show that, in the limited number of cases where we can identify regenerative structure in the MCMC sampler, we need only monitor one single one-dimensional Markov chain. Specifically, we first introduce a dimension reduction technique which captures the total variation convergence of the underlying Markov chain. From there, we propose a novel visual aid for assessing this convergence.

This useful diagnostic visualization is based on a simple observation regarding the 'elapsed time process' denoted by $E(t) = t - T_{N(t)-1}$. Intuitively, the elapsed time process concerns the number of steps since last regeneration time. Our novel result is summarized in the following theorem and its proof is provided in Appendix A.4.

Theorem 5.2.4 (Elapsed-Time Convergence Diagnostic). Suppose $\boldsymbol{X}_1 \sim \nu$ and denote $E(\infty)$ as a stationary version of $E(t)$.

$$\delta_t := \|\mathbb{E}[\kappa_t(\cdot \mid \boldsymbol{X}_1)] - \pi\|_{\mathrm{TV}} \leq 2 \sup_A |\mathbb{P}[E(t) \in A] - \mathbb{P}[E(\infty) \in A]| \,,$$

In other words, twice the total variation error of $E(t)$ bounds the total variation error of $\boldsymbol{X}_t$.

This result implies that one only needs to monitor the convergence of the one-dimensional process $\{E(t), t \geq 0\}$. This is because Theorem 5.2.4 suggests that whenever the Elapsed time process converges rapidly, then so does $\mathbb{E}[\kappa_t(\cdot \mid \boldsymbol{X}_1)]$. To this end, we propose displaying the autocorrelation of the process $\{E(t), t \geq 0\}$ as a visual aid in assessing the convergence speed of the underlying Markov chain. We call this the *elapsed-time convergence diagnostic*, and of course, we expect a rapidly converging Markov chain exhibits a rapidly decaying elapsed-time convergence diagnostic plot.

5.2.3 Non-asymptotic variance upper bound

In completing our toolbox for MCMC diagnostics via regenerations, we consider the following estimator

$$\hat{q}_{N(t)} = \frac{1}{T_{N(t)}} \sum_{k=1}^{T_{N(t)}} h(\boldsymbol{X}_k).$$

The renewal theorem [3] guarantees $\hat{q}_{T_{N(t)}} \to q := \int h(\boldsymbol{x})\pi(d\boldsymbol{x})$ as $t \to \infty$. Moreover, we have the following non-asymptotic result from [71] regarding its MSE.

Theorem 5.2.5 (Upper Bound on MSE). Let $Z_r := H_r - q\, M_r$, then the MSE of $\hat{q}_{T_{N(t)}}$ satisfies:

$$\mathbb{E}(\hat{q}_{T_{N(t)}} - q)^2 \leq \frac{\mathbb{E}Z_1^2}{t m_1} + \frac{m_2 \mathbb{E}Z_1^2}{t^2 m_1^2}.$$

We also included a proof of this result in Appendix A.5. Note that this upper bound is asymptotically sharp in the sense that it is equivalent to the consistent estimator $\hat{\gamma}_h^2$ in (5.2) as $t \to \infty$. In addition, we again propose estimating the the constants involved in this non-asymptotic upper bound from the simulation output. For example, a consistent estimator of $\mathbb{E}Z^2$ is

$$\widehat{\mathbb{E}Z_1^2} = \frac{1}{N(t)} \sum_{r=1}^{N(t)} (H_r - \hat{q}_{T_{N(t)}} M_r)^2 \,. \tag{5.4}$$

5.3 Toy examples for illustration

We have proposed our novel MCMC diagnostics when regeneration times are identifiable. In this section we illustrate how Theorems 5.2.1, 5.2.2, 5.2.3, and 5.2.4 can be applied with two simple toy examples.

5.3.1 Independence sampler for univariate truncated normal

Let us consider the target density as $\pi(\boldsymbol{x}) = \frac{1}{2\sqrt{2\pi}} e^{-x^2/2}$ for $x \geq 0$, that is the density of a $\mathsf{N}(0,1)$ random variable constrained on $(0,\infty)$. We study independence samplers with proposal densities $g_\lambda(x) = \lambda e^{-\lambda x}$ for $x > 0$, that is the densities for $\mathsf{Exp}(\lambda)$ random variables. Formally, given λ, the probability transition kernel is

$$\begin{aligned}\kappa(dy \,|\, x) &= \min\left\{\frac{\pi(y)g_\lambda(x)}{\pi(x)g_\lambda(y)}, 1\right\} g_\lambda(y)\,dy + \delta_x(dy)\int 1 - \min\left\{\frac{\pi(u)g_\lambda(x)}{\pi(x)g_\lambda(u)}, 1\right\} g_\lambda(u)\,du \\ &= \min\left\{\frac{w(y;\lambda)}{w(x;\lambda)}, 1\right\} g_\lambda(y)\,dy + \delta_x(dy)\int 1 - \min\left\{\frac{w(u;\lambda)}{w(x;\lambda)}, 1\right\} g_\lambda(u)\,du,\end{aligned}$$

where $w(\cdot;\lambda) = \pi(\cdot)/g_\lambda(\cdot)$. Next, we recall the discussion on independence sampler in Section 4.2. Given the previous state x and current state y, the current state is the start of a new regenerative cycle if $U < r(y \,|\, x)$ where U is an independent $\mathsf{Unif}(0,1)$ draw, and

$$r(y \,|\, x) = \frac{\min\left\{\frac{w(y;\lambda)}{c}, 1\right\} \min\left\{\frac{c}{w(x;\lambda)}, 1\right\}}{\min\left\{\frac{w(y;\lambda)}{w(x;\lambda)}, 1\right\}}$$

for any $c > 0$.

We illustrate with two examples, first with a $\mathsf{Exp}(0.5)$ proposal, and second with a $\mathsf{Exp}(3.5)$ proposal. In both cases, we perform a pilot run to choose a c that approximately gives the most regenerations on average. Our simulation experience reveals to us that when $\lambda = 0.5$, we should choose $c \approx 0.4$ and when $\lambda = 3.5$, we should choose $c \approx 0.2$. The results illustrated in Figures 5.1, 5.2, 5.3, and 5.4.

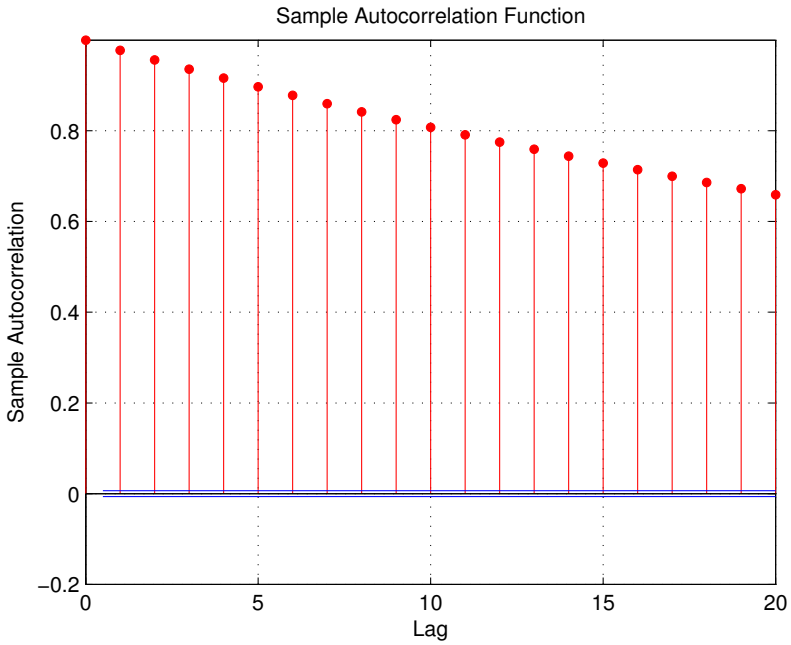

Figure 5.1: Sample autocorrelation function for the elapsed time process when $\lambda = 3.5$ as a demonstration of Theorem 5.2.4

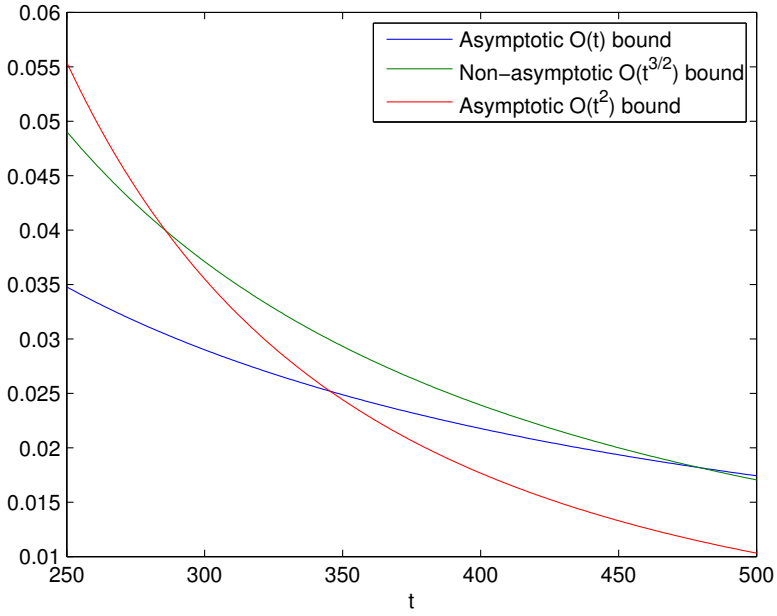

Figure 5.2: Estimated bounds as functions of t when $\lambda = 3.5$ as a demonstration of Theorems 5.2.1, 5.2.2,and 5.2.3.

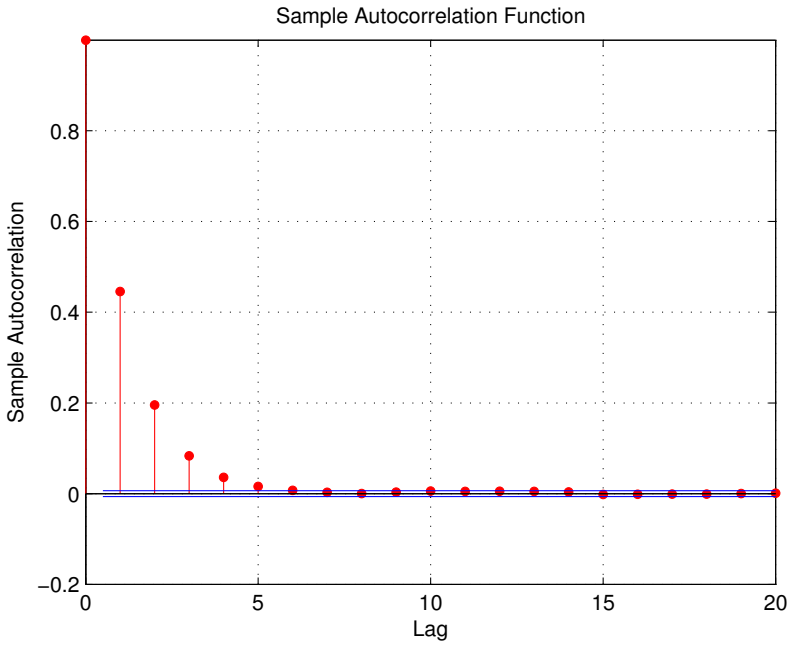

Figure 5.3: Sample autocorrelation function for the elapsed time process when $\lambda = 0.5$ as another demonstration of Theorem 5.2.4

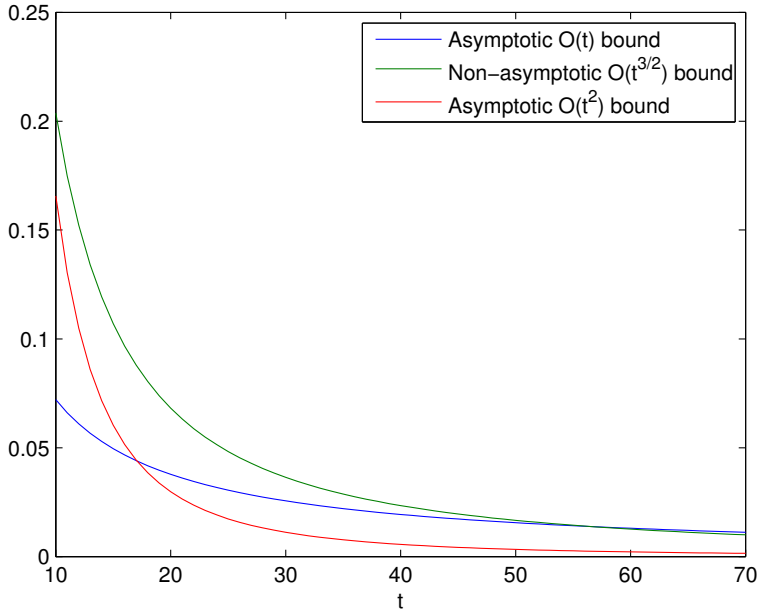

Figure 5.4: Estimated bounds as functions of t when $\lambda = 0.5$ as another demonstration of Theorems 5.2.1, 5.2.2,and 5.2.3.

Our studies reveal that $\mathsf{Exp}(0.5)$ is a better proposal in targeting π compared to $\mathsf{Exp}(3.5)$. This agrees with the theory of independence sampler as $\mathsf{Exp}(0.5)$ is a better approximation to $\mathsf{N}(0,1)$ constrained to $(0,+\infty)$.

5.3.2 Gibbs sampler for bivariate truncated normal

We now present another toy example to illustrate the diagnostic plot resulting from Theorem 5.2.4. (Estimates for the total variation distance bounds in Theorems 5.2.1, 5.2.2 and 5.2.3 are not shown for this example. Our simulation experience gives us similar results to Figures 5.2 and 5.4.)

Consider the bivariate random vector (X_1, X_2) with density

$$\pi(\boldsymbol{x}_1, \boldsymbol{x}_2) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left[x_1^2 + x_2^2 - 2\rho x_1 x_2\right]\right)$$

so that marginally $X_i \sim \mathsf{N}(0,1)$ for $i = 1, 2$ and that $\mathrm{Corr}(X_1, X_2) = \rho$. Observe that for $i \neq j$, the conditional densities $\pi(x_i \,|\, x_j)$ are the densities of $\mathsf{N}(\rho x_j, 1-\rho^2)$ random variables. In this manner, we can consider the Gibbs sampler with kernel

$$\kappa(\boldsymbol{y} \,|\, \boldsymbol{x}) = \pi(y_2 \,|\, y_1)\pi(y_1 \,|\, x_2),$$

which has limiting distribution π. Applying the usual trick described in Section 4.2 regarding Gibbs samplers, we may choose

$$s(x_1, x_2) = \varepsilon \exp\left(-\frac{1}{2(1-\rho^2)}\left[-2cx_2 + x_2^2\right]\right)$$

and

$$\nu(y_1, y_2) = \varepsilon^{-1}\pi(y_2 \,|\, y_1)\pi(y_1 \,|\, 0)\mathbb{I}\{y_1 \geq c\}$$

for some $c > 0$ normalizing constant for ε for ν. It follows that the probability transition kernel satisfies the minorization condition $\kappa(\boldsymbol{y} \,|\, \boldsymbol{x}) \geq s(\boldsymbol{x})\nu(\boldsymbol{y})$. In this manner, given the previous state $\boldsymbol{x}$ and the current state $\boldsymbol{y}$, the (retrospective) probability of the current state initializing a new regenerative cycle is

$$r(\boldsymbol{y} \,|\, \boldsymbol{x}) = \frac{s(\boldsymbol{x})\nu(\boldsymbol{y})}{\kappa(\boldsymbol{y} \,|\, \boldsymbol{x})} = \exp\left(-\frac{\rho x_2(y_1 - c)}{1-\rho^2}\right)\mathbb{I}\{y_1 \geq c\}.$$

It is well known the mixing property of this Gibbs sampler is poor for extreme values of ρ, that is if $|\rho|$ is close to 1. Figure 5.5 shows how our proposed graphical diagnostic, that is the autocorrelation plot of the elapsed time process, can capture this. In Figure 5.6 we show the autocorrelation plots for different functionals of the Markov chain compared to the autocorrelation plot of the elapsed time process when $\rho = -0.85$. We can see that the plot of the elpased time process gives the most conservative diagnostic regarding the mixing of the underlying Markov chain.

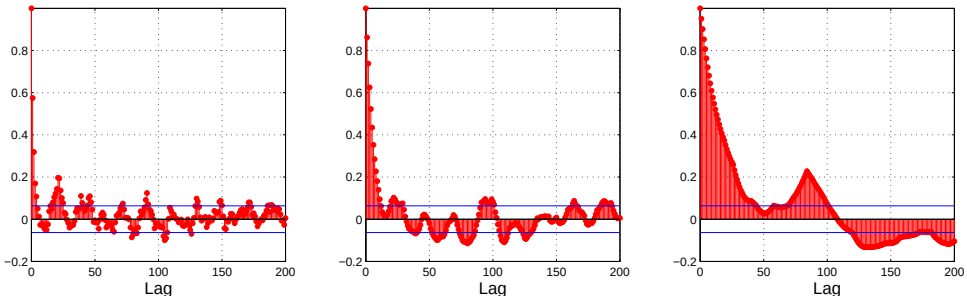

Figure 5.5: The sample autocorrelation functions of the elapsed time process for $\rho = 0.6, 0.9, 0.99$ (from left to right).

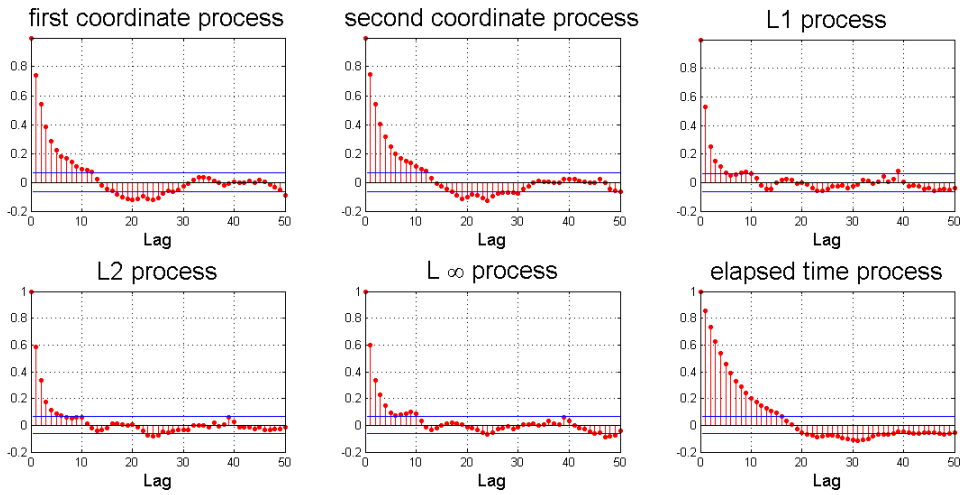

Figure 5.6: The sample autocorrelation functions for different functionals of the Markov chain.

5.4 Applications

In this section we establish the minorization condition for the Park & Casella Gibbs sampler [91] for the Bayesian Lasso model. The detail of this Gibbs sampler are described in Section 3.1 and Appendix A.6. The corresponding numerical results, when tested against real and synthetic datasets, are given in Section 5.5.

We also describe two more independence samplers, both of which use optimally tilted sequential proposal densities. In particular, the first independence sampler also targets the posterior density of the Bayesian Lasso model and the proposal density is the one in Section 3.3. The second one targets the Bayesian probit model, and the proposal density is the proposal density described in [11]. The minorization conditions for these independence samplers are straightforward application of the formulas described in Section 4.2 regarding independence samplers. The numerical results for these samplers are also given in Section 5.5.

These proposal densities are originally studied in the context of rejection samplers. We point out to readers in advance that in this paradigm, the Markov chains converge rapidly (see Section 5.5). In particular, our analyses reveal that this indpendence sampler for the Bayesian Lasso converges much faster than the original Gibbs sampler due to Park & Casella. In this way, we successfully demonstrate the value of these optimally tilted sequential proposals outside the framework of rejection samplings.

Moreover, the dataset we consider in the numerical experiment for the Bayesian probit model consists of 516 parameters. In this manner we demonstrate that our proposed MCMC diagnostics are not restricted for low dimensional problems.

5.4.1 Application to Park & Casella sampler

Given the centralized response variable $\boldsymbol{Y}$ and standardized model matrix X, the hierarchical formulation of Bayesian Lasso linear regression model is as follows (here $\boldsymbol{\beta}, \sigma$ are model parameters and λ is the Lasso regularization parameter):

$$\begin{aligned} p(\sigma^2) &\propto \sigma^{-2} \\ \beta_j \,|\, \sigma^2, \lambda &\overset{\text{iid}}{\sim} \mathsf{Lap}(\sigma/\lambda), \quad \text{for } j \in \{1, \ldots, p\} \\ \boldsymbol{Y} \,|\, \boldsymbol{\beta}, \lambda, \sigma^2 &\sim \mathsf{N}(\mathrm{X}\boldsymbol{\beta}, \sigma^2 \mathrm{I}). \end{aligned}$$

It follows that inference for the Bayesian Lasso linear regression requires one to take expectations with respect to the posterior distribution

$$\pi(\boldsymbol{\beta}, \sigma^2 \,|\, \lambda) = \frac{(\sigma^2)^{-n/2-p/2-1} (\lambda/2)^p \exp\left(-\frac{\|\boldsymbol{y} - \mathrm{X}\boldsymbol{\beta}\|_2^2}{2\sigma^2} - \frac{\lambda}{\sqrt{\sigma^2}} \|\boldsymbol{\beta}\|_1 \right)}{\ell(\lambda)}, \tag{5.5}$$

where $\ell(\lambda)$ is the marginal likelihood for λ. (The conditioning on $\boldsymbol{y}$ is dropped in the posterior π for the simplicity of notation.)

Recall (see Appendix A.6 for details or [91]) that the transition density for the Gibbs sampler of Park & Casella is

$$\kappa(\underbrace{\boldsymbol{\beta}_*, \sigma_*^2, \boldsymbol{\tau}_*}_{\boldsymbol{x}_{k+1}} \,|\, \underbrace{\boldsymbol{\beta}, \sigma^2, \boldsymbol{\tau}}_{\boldsymbol{x}_k}) = \pi(\sigma_*^2 \,|\, \boldsymbol{\beta}, \boldsymbol{\tau}) \pi(\tau_* \,|\, \boldsymbol{\beta}, \sigma_*^2) \pi(\boldsymbol{\beta}_* \,|\, \boldsymbol{\sigma}_*^2, \boldsymbol{\tau}_*).$$

Here $\pi(\boldsymbol{\beta} | \sigma^2, \boldsymbol{\tau}) = \mathsf{N}(\mathrm{A}_{\boldsymbol{\tau}} \mathrm{X}^\top \boldsymbol{y}, \sigma^2 \mathrm{A}_{\boldsymbol{\tau}})$, $\pi(\tau_j \,|\, \boldsymbol{\beta}, \sigma^2) = \mathsf{Wald}(\lambda^2, \sigma\lambda/|\beta_j|)$ (see, for example, [19]), and $\pi(\sigma \,|\, \boldsymbol{\beta}, \boldsymbol{\tau}) = \mathsf{InvGamma}((n-1)/2 + p/2, b(\boldsymbol{\beta}, \boldsymbol{\tau})/2)$, where $\mathrm{A}_{\boldsymbol{\tau}}^{-1} = \mathrm{X}^\top \mathrm{X} + \mathrm{diag}(\boldsymbol{\tau})$ and $b(\boldsymbol{\beta}, \tau) = \|\boldsymbol{y} - \mathrm{X}\boldsymbol{\beta}\|_2^2 + \boldsymbol{\beta}^\top \mathrm{diag}(\boldsymbol{\tau}) \boldsymbol{\beta}$. We have the following lemma and its proof is given in Appendix A.7.

Lemma 5.4.1 (Regenerative conditions for Park & Casella sampler). Let $\mathscr{D} = \mathbb{R}^p \times [l, u] \times [\boldsymbol{c}, \boldsymbol{d}]$, a subset of $\mathbb{R}^p \times \mathbb{R}_+ \times \mathbb{R}_+^p$ which is the state space on which $(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\tau})$ is defined. Define the probability measure:

$$\nu(\boldsymbol{\beta}_*, \sigma_*^2, \boldsymbol{\tau}_*) := \varepsilon^{-1} \kappa(\boldsymbol{\beta}_*, \sigma_*^2, \boldsymbol{\tau}_* | \tilde{\boldsymbol{\beta}}, \tilde{\sigma}^2, \tilde{\boldsymbol{\tau}}) \mathbb{I}\{(\boldsymbol{\beta}_*, \boldsymbol{\tau}_*, \sigma_*^2) \in \mathscr{D}\}, \tag{5.6}$$

where $(\tilde{\boldsymbol{\beta}}, \tilde{\sigma}^2, \tilde{\boldsymbol{\tau}}) \in \mathscr{D}$ is fixed and ε is the normalizing constant for ν. Further, define

$$\begin{aligned}\Upsilon(\boldsymbol{\beta}, \boldsymbol{\tau}; \boldsymbol{c}, \boldsymbol{d}) = \|\boldsymbol{y} - \mathrm{X}\tilde{\boldsymbol{\beta}}\|_2^2 - \|\boldsymbol{y} - \mathrm{X}\boldsymbol{\beta}\|_2^2 + \tilde{\boldsymbol{\beta}}^\top \mathrm{diag}(\tilde{\boldsymbol{\tau}})\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^\top \mathrm{diag}(\boldsymbol{\tau})\boldsymbol{\beta} \\ + \sum_{j \in \mathscr{J}} c_j^2(\tilde{\tau}_j - \tau_j) + \sum_{j \notin \mathscr{J}} d_j^2(\tilde{\tau}_j - \tau_j) + \boldsymbol{w}^\top(\mathrm{A}_{\tilde{\boldsymbol{\tau}}} - \mathrm{A}_{\boldsymbol{\tau}})\boldsymbol{w},\end{aligned}$$

where $\mathscr{J} := \{j : \tilde{\tau}_j - \tau_j \geq 0\}$, and

$$s(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\tau}) := \left(\frac{b(\boldsymbol{\beta}, \boldsymbol{\tau})}{b(\boldsymbol{\beta}, \boldsymbol{\tau})}\right)^a \left(\frac{\det(\mathrm{A}_{\tilde{\boldsymbol{\tau}}})}{\det(\mathrm{A}_{\boldsymbol{\tau}})}\right)^{n/2} \exp\left(\frac{\Upsilon(\boldsymbol{\beta}, \boldsymbol{\tau}; \boldsymbol{c}, \boldsymbol{d})_+}{2u} + \frac{\Upsilon(\boldsymbol{\beta}, \boldsymbol{\tau}; \boldsymbol{c}, \boldsymbol{d})_-}{2l}\right).$$

Here the notation $a_+ = \max\{a, 0\}$ and $a_- = \min\{a, 0\}$. Then, the measure ν and the function s satisfy the minorization condition:

$$\kappa(\boldsymbol{\beta}_*, \sigma_*^2, \boldsymbol{\tau}_* \,|\, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\tau}) \geq \nu(\boldsymbol{\beta}_*, \sigma_*^2, \boldsymbol{\tau}_*) s(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\tau}), \qquad \forall (\boldsymbol{\beta}, \sigma^2, \boldsymbol{\tau}).$$

Conditional on the simulated states $(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\tau})$ and $(\boldsymbol{\beta}_*, \sigma_*^2, \boldsymbol{\tau}_*)$, the probability that $(\boldsymbol{\beta}_*, \sigma_*^2, \boldsymbol{\tau}_*)$ is the start of a new regenerative cycle is:

$$r(\boldsymbol{\beta}_*, \sigma_*^2, \boldsymbol{\tau}_* \,|\, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\tau}) = \exp\left(\frac{\Upsilon(\boldsymbol{\beta}, \boldsymbol{\tau}; \boldsymbol{c}, \boldsymbol{d})_+}{2u} + \frac{\Upsilon(\boldsymbol{\beta}, \boldsymbol{\tau}; \boldsymbol{c}, \boldsymbol{d})_-}{2l} - \frac{\Upsilon(\boldsymbol{\beta}, \boldsymbol{\tau}; \boldsymbol{\beta}_*, \boldsymbol{\beta}_*)}{2\sigma_*^2}\right) \tag{5.7}$$

Note that to simulate $(\boldsymbol{\beta}_1, \sigma_1^2, \boldsymbol{\tau}_1) \sim \nu(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\tau})$, one only needs to simulate from $\kappa(\cdot|\tilde{\boldsymbol{\beta}}, \tilde{\sigma^2}, \tilde{\boldsymbol{\tau}})$ until a realization falls in $\mathscr{D}$. In practice, one needs to specify $(\tilde{\boldsymbol{\beta}}, \tilde{\sigma^2}, \tilde{\boldsymbol{\tau}})$, $[l, u]$, and $[\boldsymbol{c}, \boldsymbol{d}]$. Our simulation experience is that it pays to experiment with these parameters to increase the probability of a regeneration.

Finally geometric ergodicity of this Gibbs sampler is proven in [68]. Consequently we now have all the ingredients for identifying regeneration events during the course of running the Gibbs sampling of Park & Casella.

5.4.2 Independence sampler for the Bayesian Lasso

Again, dropping the conditional $\boldsymbol{y}$ in our notation, the posterior density of the Bayesian Lasso linear regression model is:

$$\pi(\boldsymbol{\beta}, \sigma) = \frac{\sigma^{-2} \times (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\|\boldsymbol{y} - \mathrm{X}\boldsymbol{\beta}\|_2^2\right) \times \frac{\lambda^d}{(2\sigma)^d} \exp\left(-\frac{\lambda}{\sigma}\|\boldsymbol{\beta}\|_1\right)}{\ell(\lambda \,|\, \boldsymbol{y})}.$$

Recall from Section 3.3 that under suitable bijective smooth coordinate transformations, π becomes a density $f(\boldsymbol{z}, r)$ whose expression looks like a product of laplace and normal laplace densities and in this manner we can use a sequential proposal

$$g(\boldsymbol{z}, r) := g_0(r) \prod_{k=1}^{d} g_k(z_k \,|\, r, z_1, \ldots, z_{k-1}),$$

where g_0 is a normal density truncated on the positive axis, and $\{g_k, k \geq 1\}$ are some laplace or normal laplace densities. In fact, the logarithm of the likelihood ratio is bounded:

$$\psi(\boldsymbol{z}, r) := \ln \frac{f(\boldsymbol{z}, r)}{g(\boldsymbol{z})} \leq \psi^*$$

for some constant ψ^*. Consequently, when g is a proposal density for π in an independence sampler, the Markov chain is uniformly ergodic as stated in Theorem 4.2.4.

Let us now consider g as a proposal density for independence sampler so that denoting, $\boldsymbol{x} := (\boldsymbol{z}, r)$ and $w(\boldsymbol{x}) = \exp(\psi(\boldsymbol{z}, r))$, the corresponding probability transition kernel is

$$\kappa(d\boldsymbol{y} \,|\, \boldsymbol{x}) = \alpha(\boldsymbol{y} \,|\, \boldsymbol{x}) g(\boldsymbol{y})\, d\boldsymbol{y} + (1 - \alpha^*(\boldsymbol{x}))\delta_{\boldsymbol{x}}(d\boldsymbol{y}),$$

where $\alpha(\boldsymbol{y} \,|\, \boldsymbol{x}) = \min\{w(\boldsymbol{y})/w(\boldsymbol{x}), 1\}$ and $\alpha^*(\boldsymbol{x}) = \int g(\boldsymbol{u}) \min\{w(\boldsymbol{u})/w(\boldsymbol{x}), 1\}\, d\boldsymbol{u}$. In other words, with probability $\alpha^*(\boldsymbol{x})$ we simulate from a density proportional to

$$g(\boldsymbol{y}) \min\{w(\boldsymbol{y})/w(\boldsymbol{x}), 1\},$$

and with probability $1 - \alpha^*(\boldsymbol{x})$, we remain in the same state $\boldsymbol{x}$. The kernel satisfies a minorization condition, and given the current state $\boldsymbol{y}$ and the previous state $\boldsymbol{x}$, the probability that $\boldsymbol{y}$ initiates a new regenerative cycle for an independent sampler can be is as follows. (See Section 4.2 regarding minorizations conditions for independence samplers.)

$$r(\boldsymbol{y} \,|\, \boldsymbol{x}) = \begin{cases} \frac{\min\{w(\boldsymbol{y})/c, 1\} \min\{c/w(\boldsymbol{x}), 1\}}{\min\{w(\boldsymbol{y})/w(\boldsymbol{x}), 1\}}, & \text{if } \boldsymbol{y} \neq \boldsymbol{x}, \\ 0, & \text{else.} \end{cases}$$

Of course, one can proceed with some running pilot runs, and from there, one can choose the c that empirically gives the most regenerations.

We shall point out in advance that our analysis based on regeneration times reveals to us that this algorithm has better mixing properties compared to the Gibbs sampler described in the previous section (i.e. Section 5.4.1). However, the construction of this proposal density do require $n > d$ (the number of observations to be larger than the number of model coefficients), on the other hand its competing Gibbs sampler actually remains to work even when $n \leq d$ (although it appears to us that the mixing of this Gibbs sampler is poor in such an extreme case). This is because the introduction of the auxiliary variable $\boldsymbol{\tau}$ guarantees the invertibility of the matrix $\mathrm{A}_{\boldsymbol{\tau}}^{-1} = \mathrm{X}^\top \mathrm{X} + \mathrm{diag}(\boldsymbol{\tau})$.

5.4.3 The Bayesian probit linear regression model

The Bayesian probit model (see for example [26]) can be summarized by the following hierarchy.

$$\begin{aligned} Y_i \,|\, Z_i &\overset{i.i.d}{\sim} \mathsf{Ber}(\mathbb{P}[Z_i \geq 0]) \quad \text{for } i = 1, \ldots, n \\ Z_i \,|\, \boldsymbol{\beta} &\overset{i.i.d}{\sim} \mathsf{N}(\boldsymbol{x}_i^\top \boldsymbol{\beta}, 1) \quad \text{for } i = 1, \ldots, n \\ \boldsymbol{\beta} &\sim \mathsf{N}(\mathbf{0}, \nu^2 \mathrm{I}). \end{aligned}$$

Here $Z_1, Z_2, \ldots$ are latent variables introduced to simplify the sampling scheme for the posterior distribution. It turns out that $\pi(\boldsymbol{\beta}, \boldsymbol{z} \,|\, \boldsymbol{y})$, the posterior distribution, can be written as $\pi(\boldsymbol{\beta}, \boldsymbol{z}) = \pi(\boldsymbol{z} \,|\, \boldsymbol{y}) \pi(\boldsymbol{\beta} \,|\, \boldsymbol{z}, \boldsymbol{y})$. Here $\pi(\boldsymbol{z} \,|\, \boldsymbol{y})$ is the density of $\mathsf{N}(\mathbf{0}, \mathrm{I} + \nu^2 \mathrm{X}\mathrm{X}^\top)$ constrained to $z_i > 0$ if $y_i = 1$, otherwise $z_i \leq 0$, and $\pi(\boldsymbol{\beta} \,|\, \boldsymbol{z}, \boldsymbol{y})$ is the density of $\mathsf{N}(\mathrm{C}\mathrm{X}^\top \boldsymbol{z}, \mathrm{C})$, where $\mathrm{C} = (\mathrm{X}^\top \mathrm{X} + \mathrm{I}/\nu^2)^{-1}$, see [9].

Of course sampling from $\pi(\boldsymbol{\beta} \,|\, \boldsymbol{z}, \boldsymbol{y})$ is routine, so the only challenge here is sampling from $\pi(\boldsymbol{z} \,|\, \boldsymbol{y})$. In this chapter we sample from $\pi(\boldsymbol{z} \,|\, \boldsymbol{y})$ using the independence Metropolis sampler with the optimal sequential importance sampling density given in [11, Equation 5], which again takes the form:

$$g(\boldsymbol{z}) := \prod_{k=1}^{d} g_k(z_k \,|\, z_1, \ldots, z_{k-1}),$$

but $\{g_k\}$ are truncated normal densities, and the logarithm of the likelihood ratio is bounded [11, Equation 6]:

$$\psi(\boldsymbol{z}) := \ln \frac{\pi(\boldsymbol{z} \,|\, \boldsymbol{y})}{g(\boldsymbol{z})} \leq \psi^*$$

for some constant ψ^*. This g is originally studied for rejection sampling targeting multivariate normal densities over linear constraints. We shall use g as the proposal density for an independence sampler in our numerical example. Consequently, the corresponding

minorization condition, and the retrospective probability for a regeneration follow exactly as described in Section 5.4.2.

5.5 Numerical Experiments

5.5.1 Park & Casella Gibbs sampler

In this section we consider two datasets that we shall model with the Bayesian Lasso regression. We identify the regenerative structure in the Gibbs sampler formulated in [91] and apply our diagnostics. The first dataset is a synthetic dataset that emulates a simple univariate problem. The second dataset is the 'diabetes dataset' of [27]. It consists of 10 predictor variables (age, sex, BMI, etc.) and a response variable which is a medical measurement for the level of diabetes for $n = 442$ patients.

5.5.1.1 Synthetic dataset

This synthetic dataset consists of $n = 5$ observations of pairs (x_i, y_i) generated in the following manner. For $i = 1, \ldots, 5$, simulate $x_i \sim \mathsf{N}(0, 1)$, then set $y_i = -0.05x_i + w_i$, where $w_i \overset{\text{iid}}{\sim} \mathsf{N}(0, 1)$.

The purpose of this simple example is to demonstrate the advantage of our estimator for (the bound on) δ_t. A geometric total variation distance bound between the t-step probability transition kernel of this Gibbs sampler and posterior density is derived in [68] using Theorem 5.1.1. That is, for any $r \in (0, 1)$,

$$\|\kappa_t(\cdot \mid \boldsymbol{X}_0) - \pi\| \leq (1 - \varepsilon)^{rt} + (\alpha^{-(1-r)} A^r)^t \left(1 + \frac{b}{1 - \lambda} + \mathbb{E}(V(\boldsymbol{X}_0))\right)$$

where $\boldsymbol{X}_0 \sim \nu_0$

$$\alpha^{-1} = \frac{1 + 2b + \lambda c}{1 + c} < 1, \quad A = 1 + 2(\lambda c + b)$$

for all initial distribution ν_0 and ε is a constant depends on the data. In this manner, it may seem that our estimation has less value when such geometrically decaying bounds can be derived analytically.

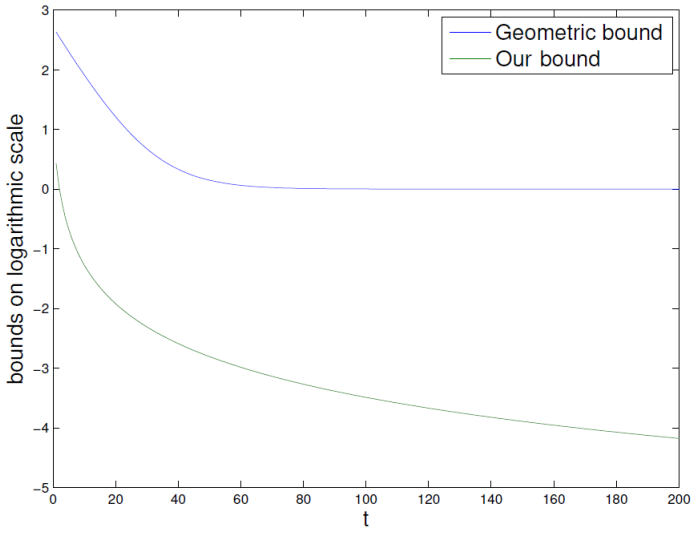

Figure 5.7: The graph compares the geometric bound derived in [68] to our bound given in Theorem 5.2.1 for the simple synthetic dataset. Here ε is approximately 0, resulting in the pathological behavior of the bound.

However, our numerical experiments reveal that these geometric bounds, in particular the bound given in [68], can be numerically unstable. Indeed, even for a simple univariate case, the parameters $n = 5$ and $y_k = -0.05x_k + w_k$ are chosen by trial and error, so that we can achieve a meaningful bound (in many of our trials, ε often gets rounded to 0, resulting in a bound that is always larger than 1). Further, we can see that the bound is overly conservative, it estimates that it takes about 1.3×10^{12} steps before the Markov chain is confidently within 0.01 total variation distance to the target density.

In experiments with real datasets (see next section), such geometric bounds for the Gibbs sampler [91] fail to return us meaningful results (due to rounding ε to 0) and we do not pursue further.

5.5.1.2 Diabetes dataset

Figure 5.8 is an illustration of Theorem 5.2.1 for this dataset. Setting $t = 10^4$, we simulate 200 independent parallel Markov chains. The regeneration cycles are identified in each chain so that we get 200 independent estimates for c_1. We then display the empirical median, and 0.95 confidence bound in the figure. Empirically, our estimate is $c_1 \in (4.096, 4.550)$ with 0.95 confidence. Therefore an approximate 0.01-burn-in period is $t_b \in (410, 455)$. Finally, Figure 5.9 is an illustration of our proposed visual diagnostics as a result of Theorem 5.2.4.

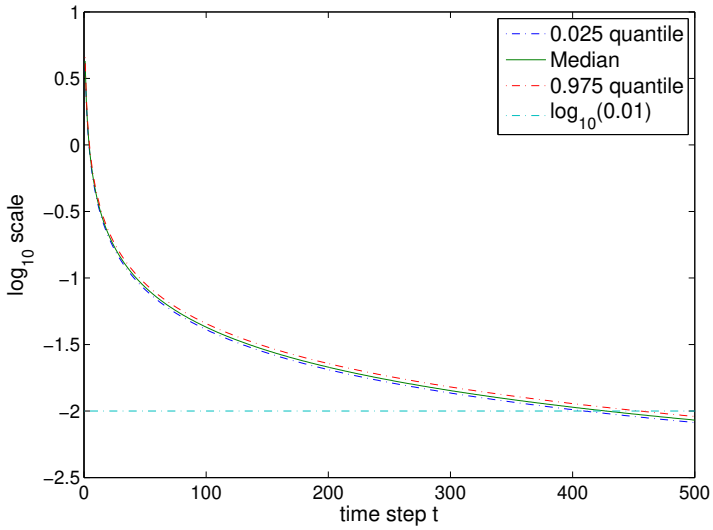

Figure 5.8: An estimate for δ_t for $t = 1, 2, \ldots, 500$ on a logarithmic scale.

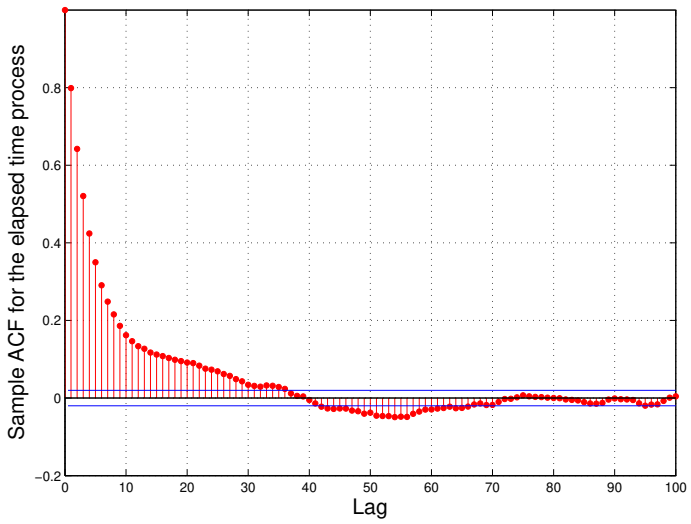

Figure 5.9: Autocorrelation plot of elapse time. The value for lag zero is one (not shown on the graph).

Finally Figure 5.10 compares the variance bound estimator given in Theorem 5.2.5 to the batch-means variance estimator for $t = 1, \ldots, 100$ Markov chain lengths. We also estimate the 'true' variance by running iid multiple Markov chains and use that as a benchmark. Our simulation experience suggests to us that the three always eventually coincide. However Figure 5.10 reveals that the batch-means variance estimator is less reliable as some variance estimates can under estimate while some may over estimate. Nevertheless, our simulation experience reveals to us that when chain length is about 10^5, all the variance estimators coincide.

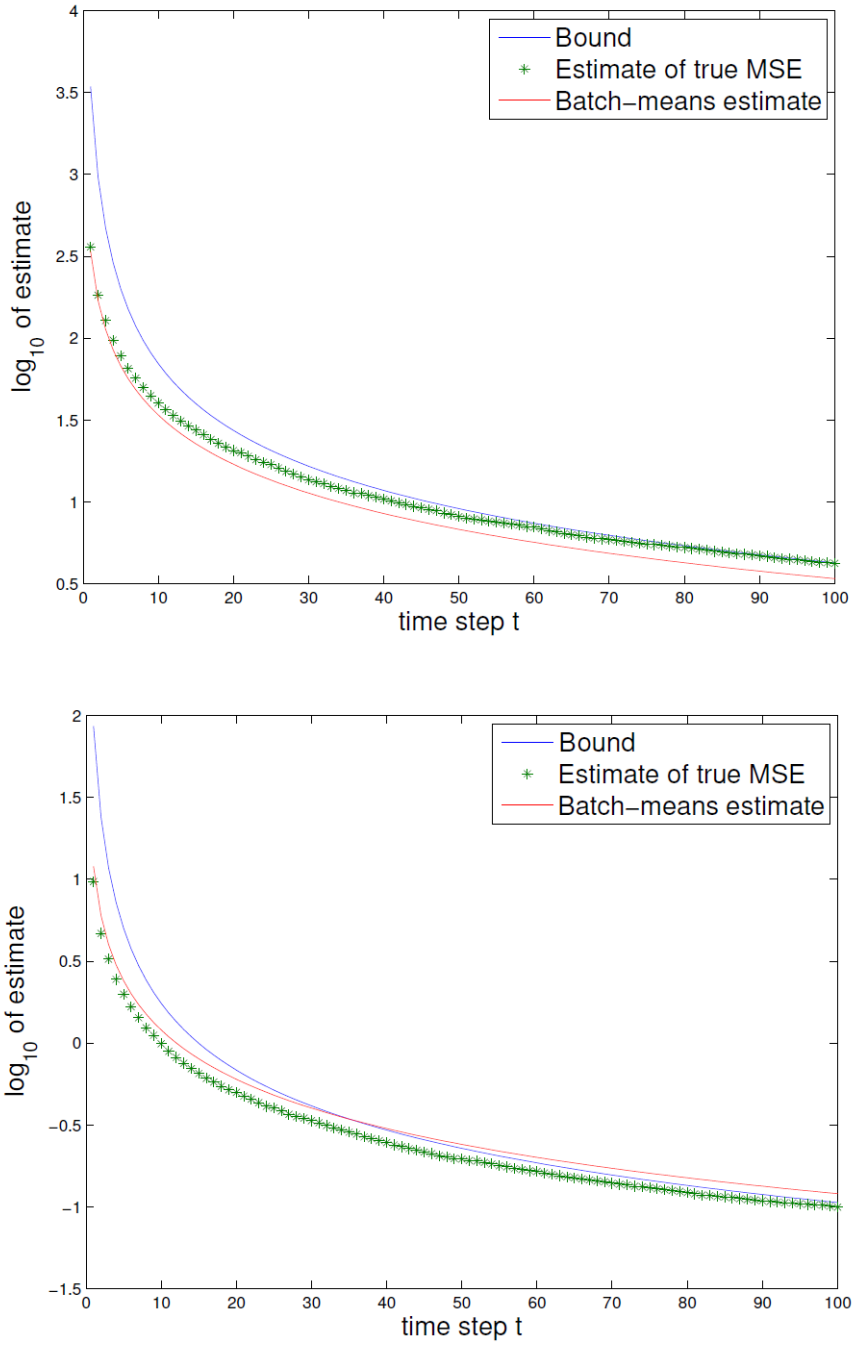

Figure 5.10: A comparison of variance estimates for estimating $\mathbb{E}\beta_5$ and $\mathbb{E}\beta_{10}$.

Note that our simulation experience reveals to us that the probability of regeneration derived here for the Park and Casella Gibbs sampler appears to to be sub-optimal. Upon testing the methodology for larger datasets, the number of regenerations observed becomes less and our diagnostics give very conservative results. For example, in a dataset with 14 variables (the Boston Housing Dataset of Section 5.5.2.2), the estimated number of burn-in required is about 1.1×10^5 before we can intialize the Markov chain within less than

0.01 TV distance to the target. Nevertheless, subsequent examples show that having a carefully tailored sampling scheme allow one to perform our proposed diagnostics in high dimensional settings.

5.5.2 Independence sampler with sequential proposal for the Bayesian Lasso

In this section we sample from the posterior density of the the Bayesian Lasso linear regression model using the independence sampler described in Section 5.4.2 instead of the Park & Casella Gibbs sampler. We consider the Diabetes Dataset and the Boston Housing Datasets [51].

5.5.2.1 Diabetes Dataset

Our simulation experience with this dataset confirms that this sampler can converge very fast – empirically we get the constant c_1 for Theorem 5.2.1 to be $c_1 \in (1.556, 1.625)$ with 95% confidence. In other words, the 0.01-burn-in this time approximately $(156, 163)$, this is drastically smaller than what the Gibbs sampler gives. Of course this difference can be a consequence of Lemma 5.4.1 being too loose. However, to our best effort (for example, our countless experiments on choosing $(\tilde{\boldsymbol{\beta}}, \tilde{\sigma^2}, \tilde{\boldsymbol{\tau}})$, and $\mathscr{D}$) we cannot achieve a less conservative diagnostics for that Gibbs sampler.

5.5.2.2 Boston Housing Dataset

Our simulation experience with this dataset again shows that this sampler can converge very fast – empirically we get the constant c_1 for Theorem 5.2.1 to be $c_1 \in (1.581, 1.649)$ with 95% confidence. In other words, the 0.01-burn-in this time approximately $(159, 165)$, which is drastically smaller than what the Gibbs sampler gives.

5.5.3 The Bayesian probit linear regression model

Here we consider the dataset described in [80]. It consists of 74 observations and 516 predictors for cancerous tissues, that is the response y_i takes the value 1 if the tissue is cancerous, and 0 otherwise. (Note the that a proper prior for $\boldsymbol{\beta}$, that is if $\nu^2 \neq 0$, give a non-degenerate posterior distribution.) We can see in this example that the convergence is very fast as the proposal density for the independent Metropolis sampler is carefully tailored (optimally tilted) for the problem. It is worthwhile noting that this is an example where regenerative structure can be frequently identified in a high dimensional setting.

Figure 5.11 again illustrates Theorem 5.2.1 by simulating 200 independent parallel Markov chains for this dataset (again we choose $t = 10^4$). Empirically, our estimate is $c_1 \in (1.978, 2.408)$ with 0.95 confidence, thus an approximate 0.01-burn-in period is $t_b \in (197, 240)$. Similarly, Figure 5.12 is an illustration of Theorem 5.2.4. The autocorrelation plot of the elapsed time process reveals that this Markov chain converges quickly.

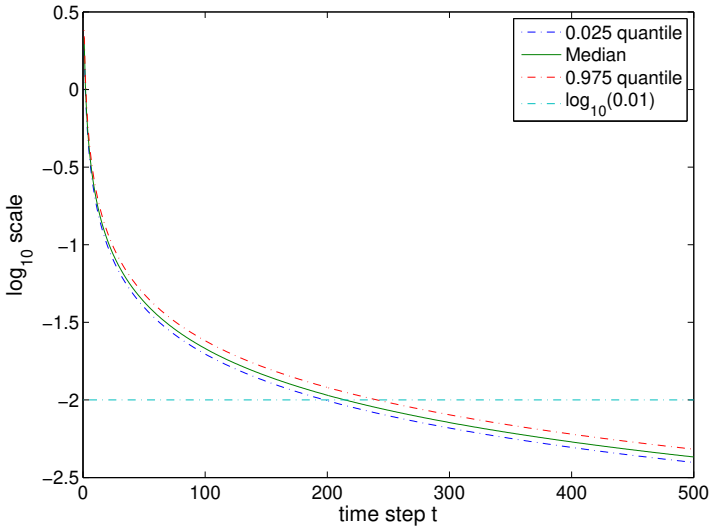

Figure 5.11: An estimate for δ_t for $t = 1, 2, \ldots, 100$ on a logarithmic scale.

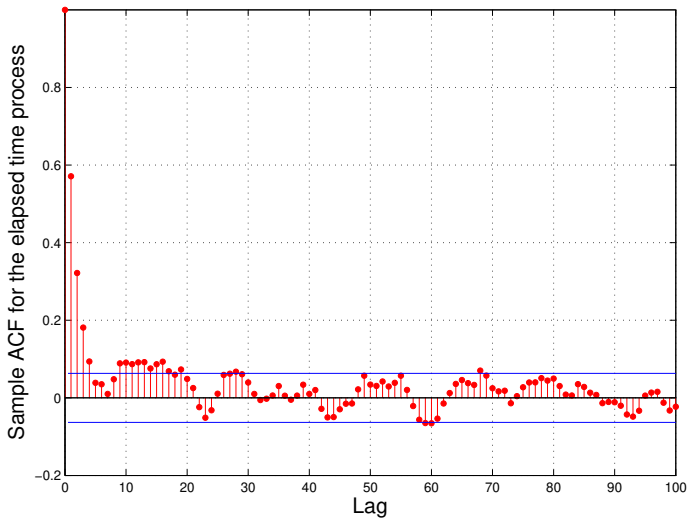

Figure 5.12: Autocorrelation plot of elapse time.

5.6 Inference for estimated constants via M-estimation

The illustrative examples in Sections 5.5.1.2 and 5.5.3 provide empirical 0.95 confidence bounds for the estimated constant c_1 of Theorem 5.2.1 by running 200 independent parallel Markov chains. This is actually quite wasteful, given that we know the regenerative cycle lengths $M_1, M_2, \ldots$, which are used to estimate c_1, are iid. In this section we demonstrate how one can formulate M-estimators for constants such as c_1 in Theorem 5.2.1. We begin by recalling the well-known asymptotic normality of M-estimators, when the number of sample is fixed (see [105] and the references therein):

Suppose $\boldsymbol{W}_k = (W_{k1}, W_{k2}, \ldots, W_{km}) \in \mathbb{R}^m$ are iid random vectors for $k = 1, \ldots, n$. Let $\varphi : \mathbb{R}^m \times \mathbb{R}^d \to \mathbb{R}^d$ and recall that the M-estimator for $\boldsymbol{\theta}^* \in \mathbb{R}^d$ is the solution $\hat{\boldsymbol{\theta}}_n$ to the equation

$$\sum_{k=1}^{n} \varphi(\boldsymbol{W}_k, \hat{\boldsymbol{\theta}}_n) = \mathbf{0},$$

where the true value $\boldsymbol{\theta}^*$ satisfies $\mathbb{E}[\varphi(\boldsymbol{W}_1, \boldsymbol{\theta}^*)] = \mathbf{0}$. Under some regularity conditions, we have $\hat{\boldsymbol{\theta}}_n \to \boldsymbol{\theta}^*$ in probability and that $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*)$ obeys some normal distribution as $n \to \infty$. (We refer to readers to [105] and the references therein for more details on M-estimators.)

Although the extension of this result to the case where we have a random number of samples appears to be part of the statistical folklore, we present here a formal proof, because we were unable to find a precise reference to this result.

Theorem. Let $\boldsymbol{W}_k = (W_{k1}, W_{k2}, \ldots, W_{km}) \in \mathbb{R}^m$ be iid random vectors for $k = 1, \ldots, n$ and $N(t)$ be a positive-integer-valued random variable such that $N(t) \to \infty$ a.s. and $N(t)/t \to b$ in probability as $t \to \infty$ for some finite constant b. Suppose $\varphi : \mathbb{R}^m \times \mathbb{R}^d \to \mathbb{R}^d$ is such that the solution $\hat{\boldsymbol{\theta}}_n$ to the equation $\sum_{k=1}^{n} \varphi(\boldsymbol{W}_k, \hat{\boldsymbol{\theta}}_n) = \mathbf{0}$ satisfies $\hat{\boldsymbol{\theta}}_n \to \boldsymbol{\theta}^*$ in probability, where $\boldsymbol{\theta}^*$ is the solution to $\mathbb{E}[\varphi(\boldsymbol{W}_1, \boldsymbol{\theta}^*)] = \mathbf{0}$, and that $\hat{\boldsymbol{\theta}}_n$ exhibits a CLT. Further, suppose that for each $\boldsymbol{w} \in \mathbb{R}^m$, $\varphi(\boldsymbol{w}, \cdot)$ is smooth in the second argument, $\mathbb{E}\varphi(\boldsymbol{W}, \boldsymbol{\theta})$ exists for all $\boldsymbol{\theta} \in \mathbb{R}^d$, and $\mathrm{A} := -\mathbb{E}[\partial_{\boldsymbol{\theta}} \varphi(\boldsymbol{W}_1, \boldsymbol{\theta}^*)]$ is invertible. Then the (regeneraitve-sequential) estimator $\hat{\boldsymbol{\theta}}_t^{\text{reg-seq}}$ defined as the solution to

$$\sum_{k=1}^{N(t)} \varphi(\boldsymbol{W}_k, \hat{\boldsymbol{\theta}}_t^{\text{reg-seq}}) = \mathbf{0}$$

satisfies

$$\sqrt{N(t)}(\hat{\boldsymbol{\theta}}_t^{\text{reg-seq}} - \boldsymbol{\theta}^*) \to \mathsf{N}(\mathbf{0}, \mathrm{A}^{-1}\mathrm{B}\mathrm{A}^{-\top})$$

as $t \to \infty$ where $\mathrm{B} = \mathbb{E}[\varphi(\boldsymbol{W}_1, \boldsymbol{\theta}^*)\varphi(\boldsymbol{W}_1, \boldsymbol{\theta}^*)^\top]$.

Proof. First, note that our assumption $\hat{\boldsymbol{\theta}}_n \to \boldsymbol{\theta}^*$ in probability as $n \to \infty$ guarantees $\hat{\boldsymbol{\theta}}^{\text{reg-seq}} \to \boldsymbol{\theta}^*$ in probability as $t \to \infty$. Observe that by the multivariate Taylor's theorem, there is a matrix A_t such that as $\hat{\boldsymbol{\theta}}_t^{\text{reg-seq}} \to \boldsymbol{\theta}^*$ in probability, $\mathrm{A}_t \to \mathrm{A}$ in probability and satisfies

$$\sum_{k=1}^{N(t)} \varphi(\boldsymbol{W}_k, \hat{\boldsymbol{\theta}}_t^{\text{reg-seq}}) = \sum_{k=1}^{N(t)} \varphi(\boldsymbol{W}_k, \boldsymbol{\theta}^*) - \mathrm{A}_t(\hat{\boldsymbol{\theta}}_t^{\text{reg-seq}} - \boldsymbol{\theta}^*) = \mathbf{0}.$$

Rearranging the above implies that for t sufficiently large,

$$\sqrt{N(t)}(\hat{\boldsymbol{\theta}}_t^{\text{reg-seq}} - \boldsymbol{\theta}^*) = \mathrm{A}_t^{-1}\sqrt{N(t)}\frac{1}{N(t)}\sum_{k=1}^{N(t)}\varphi(\boldsymbol{W}_k, \boldsymbol{\theta}^*).$$

The matrix inverse A_t^{-1} exists for t sufficiently large because matrix inversion is a continuous operation and $\mathrm{A}_t \to \mathrm{A}$ in probability as $t \to \infty$. Finally the assumption that $N(t)/t \to b$ for some finite constant t ensures that by [106, Colloray 1], $\sqrt{N(t)}\frac{1}{N(t)}\sum_{k=1}^{N(t)}\varphi(\boldsymbol{W}_k, \boldsymbol{\theta}^*)$ asymptotically obeys $\mathsf{N}(\mathbf{0}, \mathrm{B})$, and this completes the proof. □

With this result in hand, one can perform statistical inference for the estimated constants from the output of a single Markov chain run. Here an illustration to construct an M-estimator for $c_1 = \frac{m_2+m_1}{2m_1}$ of Theorem 5.2.1, and how we can estimate its asymptotic variability. Let

$$\varphi(x; \theta_1, \theta_2, \theta_3) = \varphi(x; \boldsymbol{\theta}) = \begin{pmatrix} x - \theta_1 \\ x^2 - \theta_2 \\ \theta_2 + \theta_1 - 2\theta_1\theta_3 \end{pmatrix}.$$

Denote $\boldsymbol{\theta}^* = (\theta_1^*, \theta_2^*, \theta_3^*)$ as the solution to the equation $\mathbb{E}[\varphi(M_1, \theta_1^*, \theta_2^*, \theta_3^*)]$ so that $\theta^* = (m_1, m_2, c_1)^\top$. We note that φ satisfies the conditions imposed [60, Theorem 2] so that the usual fixed sample size M-estimator $\hat{\boldsymbol{\theta}}_n \to \boldsymbol{\theta}^*$ in probability, so we can use our (regenerative-sequential) M-estimator for c_1 where

$$\hat{c}_t^{\text{reg-seq}} = \frac{\sum_{k=1}^{N(t)} M_k^2 + \sum_{k=1}^{N(t)} M_k}{2\sum_{k=1}^{N(t)} M_k}.$$

Renewal theorem guarantees that $N(t)/t \to 1/\mathbb{E}[M_1]$ in probability as $t \to \infty$, and the asymptotic variance of $\hat{c}_t^{\text{reg-seq}}$ is the $(3,3)$-th element of the matrix $\mathrm{ABA}^{-\top}$ divided by $N(t)$ i.e. the number of completed regenerative cycles where

$$\mathrm{A} = -\mathbb{E}\left[\partial_{\boldsymbol{\theta}}\varphi(M_1; \theta^*)\right] = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -1 + 2c_1 & -1 & 2m_1 \end{pmatrix}$$

is invertible and

$$\mathrm{B} = \mathbb{E}\left[\varphi(M_1; \boldsymbol{\theta}^*)\varphi(M_1; \boldsymbol{\theta}^*)^\top\right] = \begin{pmatrix} m_2 - m_1^2 & m_3 - m_1 m_2 & 0 \\ m_3 - m_1 m_2 & m_4 - m_2 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Again, in practice, we usually don't have formulas for m_r, so we substitute for them with consistent estimators for example $\hat{m}_k = \sum_{r=1}^{N(t)} M_r^k / N(t)$.

Consider again the example in Section 5.5.1.2. For $t = 10^4$, the empirical 0.95 confidence bound for c_1, from 200 independent parallel runs of Markov chain, is $(4.096, 4.550)$. In a single run of $t = 10^4$, using the strategy above, we get an asymptotic 0.95 confidence bound for c_1 of $(4.043, 4.523)$, which agrees with the "parallel-MCMC" one to two significant figures.

5.7 Concluding remarks for this chapter

In this chapter we have proposed novel MCMC diagnostics for geometrically ergodic Markov chains for which we can identify the underlying regenerative structure. In particular we derive two asymptotic and one non-asymptotic total variation distance bounds on a geometrically ergodic Markov chain. The constants involved in these bounds depend on moments of the regenerative cycle lengths. From there, we propose that in the cases where regeneration times are readily identifiable, we can use the simulation outputs to estimate these constants and quantify their asymptotic variances. In this manner we have output diagnostics that concern the total variation distance, and in this manner, it stands on better theoretical footing than the usual graphical analyses.

We have also introduced the notion of an elapsed process. We show the elapsed process gives a useful dimension reduction for assessing the total variation convergence of the underlying Markov chain. We propose a novel visual aid where we construct the autocorrelation plot for this one dimensional process. Our experiment shows that this plot seems to be able to distinguish between rapidly converging and slowly converging Markov chains.

In combining different ideas we have studied in this thesis, in the next chapter, we shall propose a novel regenerative Markov chain sampling, where we embed exact sampling in the event of a regeneration. In other words, the underlying Markov chain is regenerative, and in the event of a regeneration, the Markov chain has a certain probability of starting the new regenerative cycle exactly from the target (posterior) distribution.

Chapter 6

Reject-Regenerate Sampler

6.1 Introduction to this chapter

In Chapters 2 and 3 we have applied the exponential tilting technique to construct exact samplers from some Bayesian posterior densities resulting from practical data sets. However, as mentioned in Chapters 4 and 5, due to the curse of dimensionality no matter how careful one constructs a proposal density, rejection sampling will eventually become inefficient as the dimension of the problem grows.

Consider the situation where we have designed a sequential proposal density for efficient rejection sampling. We know that the rejection sampler will be efficient up to a certain dimension, which is typically unknown apriori. Beyond this unknown dimension, the rejection sampling will be inefficient and we will have to switch from exact rejection sampling to approximate MCMC sampling. This scenario has a number of undesirable features.

First, the user has to explicitly decide when a rejection sampler is inefficient. For example, should the cutoff for efficiency be an acceptance probability of 10^{-3} or 10^{-2}?

Second, the user has to run the rejection sampling algorithm to find out if it meets the efficiency criterion above. In the likely event that the rejection sampler does not meet the efficient criterion, this simulation effort has been effectively wasted, because the user now has to run a separate MCMC algorithm from scratch. The simulation effort from rejection sampling is not recycled by the MCMC sampler, but is simply used to make a dichotomous, all-or-nothing decision about the rejection sampler.

Given the above drawbacks of using rejection and MCMC sampling as two distinct algorithms, in this chapter we propose a single algorithm which combines the desirable features of both rejection and MCMC sampling and thus removes the need to make a choice between the two. We call this algorithm the *Reject-Regenerate* sampler.

The Reject-Regenerate sampler has the following desirable features. At a given step t, using an (exponentially tilted or otherwise) proposal density, the *Reject-Regenerate* sampler simulates a random variable $\boldsymbol{X}_t$. Then, with a certain probability the variable $\boldsymbol{X}_t$ is flagged as belonging to either one of these three states:

1. an draw within an Markov chain which initiates the next regenerative cycle;
2. an independent and exact/perfect draw from the target;
3. a regular draw within an MCMC run (which is neither exact, nor regenerative).

As a result of these features, the *Reject-Regenerate* sampler makes is unnecessary for the user to choose between rejection and MCMC sampling. If the rejection sampling is efficient, then most of the draws in the sequence $\{\boldsymbol{X}_t\}$ will be independent and exact draws from the target. However, if rejection sampling is not viable, then the sequence $\{\boldsymbol{X}_t\}$ will be interpreted as the output of an MCMC with the possibility of identifying regeneration cycles. In this way, the simulation effort in rejection sampling is recycled for MCMC sampling.

To be precise, in this chapter we propose a novel sampling scheme in which we identify the regeneration times of an usual independent sampler, and whenever regeneration occurs, it has a certain probability of achieving an independent exact draw from the target density.

We point out in advance that the proposal does not have to be derived from exponential tilting. However, unsurprisingly, numerical experiments reveal to us that a good approximation gives frequent regeneration, and in this manner we can readily apply the diagnostics we propose in Chapter 5.

We also point out in advance that our Reject-Regenerate sampler shares some common feature with Algorithm 1 in [56]. For example our Reject-Regenerate algorithm implicitly requires the whole state-space is a 'small set', and this is explicitly assumed in [56]. Further, the of goal Algorithm 1 in [56] is to ensure the initial draw of the simulated Markov chain comes from the target distribution, and this can be achieved by our Reject-Regenerate sampler too. The fundamental difference between the two algorithms lies in the construction. Algorithm 1 proposed in [56] utilizes an infinite mixture representation of the target distribution while our Reject-Regenerate sampler utilizes a four-component mixture representation of the transition kernel. Moreover, the objective of our Reject-Regenerate sampler is not to initialize the Markov chain with the target, rather it aims to provide an automated way for practitioners to switch between exact (rejection) sampling and a Markov chain sampling along with its underlying regeneration times.

6.2 The Reject-Regenerate sampler

Suppose that the target pdf is

$$\pi(\boldsymbol{x}) = \frac{p(\boldsymbol{x})}{\ell}$$

where $\ell = \int p(\boldsymbol{x})\, d\boldsymbol{x}$ is the normalizing constant and we have $p(\boldsymbol{x})$ available analytically. Our proposal pdf is $g(\boldsymbol{x})$ that satisfies

$$w(\boldsymbol{x}) := \frac{p(\boldsymbol{x})}{g(\boldsymbol{x}) \exp(\psi^*)} \leq 1.$$

where

$$\psi^* = \max_{\boldsymbol{x}} \psi(\boldsymbol{x}) = \max_{\boldsymbol{x}} \ln \frac{p(\boldsymbol{x})}{g(\boldsymbol{x})}.$$

Of course, in the case where g comes from a family of densities, indexed by some tilting parameter $\boldsymbol{\mu}$, in the spirit of optimal tilting, we choose $\psi^* = \min_{\boldsymbol{\mu}} \max_{\boldsymbol{x}} \psi(\boldsymbol{x}; \boldsymbol{\mu})$ and g is the corresponding optimal proposal.

Next denote

$$\min\{x, y\} := x \vee y, \quad \text{and} \quad w_\gamma(\boldsymbol{x}) := \min\{w(\boldsymbol{x})/\gamma, 1\}$$

for some $\gamma \in (0, 1]$. Now, suppose that we wish to simulate $\boldsymbol{X} \sim g$, conditional on $U \leq w_\gamma(\boldsymbol{X})$. The probability of this happening is

$$c_\gamma = \mathbb{E} w_\gamma(\boldsymbol{X}).$$

Note that rejection sampling corresponds to $\gamma = 1$ with acceptance probability $c_1 = \ell / \exp(\psi^*)$.

Recall that the probability transition kernel of an independence sampler with proposal g is

$$\kappa(d\boldsymbol{y} \,|\, \boldsymbol{x}) = \alpha(\boldsymbol{y} \,|\, \boldsymbol{x}) g(\boldsymbol{y}) d\boldsymbol{y} + (1 - \alpha^*(\boldsymbol{x})) \delta_{\boldsymbol{x}}(d\boldsymbol{y})$$

where

$$\alpha(\boldsymbol{y} \,|\, \boldsymbol{x}) = (1 \vee \tfrac{w(\boldsymbol{y})}{w(\boldsymbol{x})}), \quad \text{and} \quad \alpha^*(\boldsymbol{x}) = \int \alpha(\boldsymbol{u} \,|\, \boldsymbol{x}) g(\boldsymbol{u})\, d\boldsymbol{u}.$$

Given the current state of the Markov chain $\boldsymbol{x}$, the conventional implementation of the independence sampler is as follows. Independent draws $\boldsymbol{Y}' \sim g$, $U \sim \mathsf{Unif}(0, 1)$ are simulated and if $U < \frac{w(\boldsymbol{y}')}{w(\boldsymbol{x})}$, the next state of the chain $\boldsymbol{Y}$ is assigned $\boldsymbol{Y} \leftarrow \boldsymbol{Y}'$, otherwise $\boldsymbol{Y} \leftarrow \boldsymbol{x}$. This can be understood as the following probability transition kernel

$$\kappa(d\boldsymbol{y}, d\boldsymbol{y}', u \,|\, \boldsymbol{x}) = g(\boldsymbol{y}') \mathbb{I}_{u < \frac{w(\boldsymbol{y}')}{w(\boldsymbol{x})}} \delta_{\boldsymbol{y}'}(d\boldsymbol{y})\, d\boldsymbol{y}' + g(\boldsymbol{y}') \mathbb{I}_{u > \frac{w(\boldsymbol{y}')}{w(\boldsymbol{x})}} \delta_{\boldsymbol{x}}(d\boldsymbol{y})\, d\boldsymbol{y}'$$

which has marginal $\kappa(d\boldsymbol{y} \,|\, \boldsymbol{x})$ and is such that $\kappa(d\boldsymbol{y}' \,|\, \boldsymbol{x}) = g(d\boldsymbol{y}')$ and $\kappa(u \,|\, \boldsymbol{x}, \boldsymbol{y}') = 1$, for $u \in (0, 1)$ so that

$$\kappa(d\boldsymbol{y} \,|\, \boldsymbol{x}, \boldsymbol{y}', u) = \mathbb{I}_{u < \frac{w(\boldsymbol{y}')}{w(\boldsymbol{x})}} \delta_{\boldsymbol{y}'}(d\boldsymbol{y}) + \mathbb{I}_{u > \frac{w(\boldsymbol{y}')}{w(\boldsymbol{x})}} \delta_{\boldsymbol{x}}(d\boldsymbol{y}).$$

Next, define

$$g_\gamma(\boldsymbol{y}) := \frac{g(\boldsymbol{y})w_\gamma(\boldsymbol{y})}{c_\gamma}$$

and note that we have

$$1 \vee \frac{w(\boldsymbol{y})}{w(\boldsymbol{x})} \geq (1 \vee w(\boldsymbol{y})/\gamma) \times (1 \vee \gamma/w(\boldsymbol{x})) \geq (1 \vee w(\boldsymbol{y})/\gamma) \times \gamma,$$

and

$$\alpha^*(\boldsymbol{x}) \geq (1 \vee \gamma/w(\boldsymbol{x})) \times c_\gamma =: s_\gamma(\boldsymbol{x}).$$

It follows that we can decompose $\kappa(d\boldsymbol{y} \,|\, \boldsymbol{x})$ as a three-component mixture as follows.

$$\kappa(d\boldsymbol{y} \,|\, \boldsymbol{x}) = s_\gamma(\boldsymbol{x})g_\gamma(\boldsymbol{y})\, d\boldsymbol{y} + (\alpha^*(\boldsymbol{x}) - s_\gamma(\boldsymbol{x}))\frac{g(\boldsymbol{y})(1 \vee \frac{w(\boldsymbol{y})}{w(\boldsymbol{x})}) - g_\gamma(\boldsymbol{y})s_\gamma(\boldsymbol{x})}{\alpha(\boldsymbol{x}) - s_\gamma(\boldsymbol{x})}\, d\boldsymbol{y}$$
$$+ (1 - \alpha^*(\boldsymbol{x}))\, \delta_{\boldsymbol{x}}(d\boldsymbol{y}).$$

Regeneration happens whenever we simulate from the first component g_γ, again in practical implementation this is done retrospectively. Formally, let us define

$$r(\boldsymbol{y} \,|\, \boldsymbol{x}) = \frac{(1 \vee w(\boldsymbol{y})/\gamma) \times (1 \vee \gamma/w(\boldsymbol{x}))}{1 \vee \frac{w(\boldsymbol{y})}{w(\boldsymbol{x})}} \leq 1. \tag{6.1}$$

Given previous state $\boldsymbol{X}$ and current state $\boldsymbol{Y}$ of the Markov chain, where $\boldsymbol{Y} \neq \boldsymbol{X}$ (that is, a transition has happened), one simulates another independent $\boldsymbol{V} \sim \mathsf{Unif}(0,1)$ and decides that $\boldsymbol{Y}$ initiates a new regenerative cycle if $V < r(\boldsymbol{Y} \,|\, \boldsymbol{X})$. In other words, one retrospectively identifies $\boldsymbol{Y}$ as a draw from g_γ if $V < r(\boldsymbol{Y} \,|\, \boldsymbol{X})$. The following expression summarizes this independence sampler with regenerations.

$$\kappa(d\boldsymbol{y} \,|\, \boldsymbol{x}, \boldsymbol{y}', u, v) = \delta_{\boldsymbol{y}'}(d\boldsymbol{y})\mathbb{I}_{u < \frac{w(\boldsymbol{y}')}{w(\boldsymbol{x})}} \left[\mathbb{I}_{v < r(\boldsymbol{y}' \,|\, \boldsymbol{x})} + \mathbb{I}_{v > r(\boldsymbol{y}' \,|\, \boldsymbol{x})}\right] + \mathbb{I}_{u > \frac{w(\boldsymbol{y}')}{w(\boldsymbol{x})}} \delta_{\boldsymbol{x}}(d\boldsymbol{y}).$$

Finally, to get the exact sampling as a subset of regeneration, define

$$e(\boldsymbol{y}) = \frac{w(\boldsymbol{y})}{1 \vee (w(\boldsymbol{y})/\gamma)} \leq 1, \tag{6.2}$$

so that $c_\gamma \geq c_1$. Then,

$$g_\gamma(\boldsymbol{y}) = \frac{g(\boldsymbol{y})(1 \vee w(\boldsymbol{y})/\gamma)}{c_\gamma} = \frac{c_1}{c_\gamma}\frac{g(\boldsymbol{y})w(\boldsymbol{y})}{c_1} + \left(1 - \frac{c_1}{c_\gamma}\right)\frac{g(\boldsymbol{y})(1 \vee w(\boldsymbol{y})/\gamma) - g(\boldsymbol{y})w(\boldsymbol{y})}{c_\gamma - c_1}.$$

Notice that the first component of this mixture $g(\boldsymbol{y})w(\boldsymbol{y}) \propto \pi(\boldsymbol{y})$, thus we actually achieve an exact draw from π if a draw from g_γ actually comes from this component of the mixture.

Thus, simulation from this mixture can be accomplished by sampling from the joint:

$$g_\gamma(\boldsymbol{y}, v') = g_\gamma(\boldsymbol{y})\mathbb{I}_{\{v'<e(\boldsymbol{y})\}} + g_\gamma(\boldsymbol{y})\mathbb{I}_{\{v'>e(\boldsymbol{y})\}}.$$

In other words, simulate $\boldsymbol{Y} \sim g_\gamma(\boldsymbol{y})$ and $V' \sim \mathsf{Unif}(0,1)$ and then evaluate $\mathbb{I}_{\{V'<e(\boldsymbol{Y})\}}$ (to check if we sampled from the first component of this mixture).

Putting these observations together, we propose an algorithm where we simulate $\boldsymbol{Y} \sim g(\boldsymbol{y})$, independently $V, V', U \overset{\text{iid}}{\sim} \mathsf{Unif}(0,1)$, we consider the conditional probability measure $\kappa(d\boldsymbol{y} \,|\, \boldsymbol{x}, \boldsymbol{y}', u, v, v') =:$

$$\delta_{\boldsymbol{y}'}(d\boldsymbol{y})\mathbb{I}_{u<\frac{w(\boldsymbol{y}')}{w(\boldsymbol{x})}}\left[\mathbb{I}_{v<r(\boldsymbol{y}'\,|\,\boldsymbol{x}),v'<e(\boldsymbol{y}')} + \mathbb{I}_{v<r(\boldsymbol{y}'\,|\,\boldsymbol{x}),v'>e(\boldsymbol{y}')} + \mathbb{I}_{v>r(\boldsymbol{y}'\,|\,\boldsymbol{x})}\right] + \mathbb{I}_{u>\frac{w(\boldsymbol{y}')}{w(\boldsymbol{x})}}\delta_{\boldsymbol{x}}(d\boldsymbol{y}).$$

Moreover, the probability of exact sampling, conditional on $\boldsymbol{x}$ is:

$$s_\gamma(\boldsymbol{x}) \times \frac{c_1}{c_\gamma} = c_1 \times (1 \vee \gamma / w(\boldsymbol{x})).$$

The final algorithm is thus as follows. (Here $B = 1$ means regenerative draw and $B = 2$ means exact sampling draw.)

Algorithm 7 : MCMC with regeneration and exact sampling

Require: Current state of chain $(\boldsymbol{X}_n, B_n)$ and constant γ.

$B_{n+1} \leftarrow 0$

Simulate $\boldsymbol{Y} \sim g(\boldsymbol{y})$ and $U, V, V' \sim_{\text{iid}} \mathsf{Unif}(0,1)$, independently.

if $U \leq w(\boldsymbol{Y})/w(\boldsymbol{X}_n)$ **then**

  $\boldsymbol{X}_{n+1} \leftarrow \boldsymbol{Y}$

  if $V \leq r(\boldsymbol{Y} \,|\, \boldsymbol{X}_n)$ as in (6.1) **then**

    $B_{n+1} \leftarrow 1$

    if $V' \leq e(\boldsymbol{Y})$ as in (6.2) **then**

      $B_{n+1} \leftarrow 2$

else

  $\boldsymbol{X}_{n+1} \leftarrow \boldsymbol{X}_n$

return $(\boldsymbol{X}_{n+1}, B_{n+1})$ as the next state of the chain

Our numerical studies show that even when g renders a small probability retaining a perfect draw in rejection sampling, appropriate choices of γ can give a Markov chain that frequently regenerates. A potential application of this observation is that when π is a probability density defined on dimensions so big that even a mini-max exponentially tilted proposal density g is incompetent, we can still use it as a proposal density for an

independent Markov chain sampler which regenerates frequently so that its error analysis is remains relatively simple.

However, since we have nested exact sampling within regeneration events, despite having observed many regenerations, the probability of observing an exact sampling in the Markov chain does reduce slightly. This is illustrated in Figure 6.1.

An interesting case is where we choose $\gamma = 1$, then the algorithm corresponds to an independence sampler with regeneration, but regeneration corresponds to an exact draw from π. In other words, every regenerative cycle is initialized by an exact draw from the target. Simple analysis shows that in this setting exact component has a mixture weight equal to the probability of retaining a proposal in the rejection sampling context. This means that in this case (when $\gamma = 1$), the algorithm exploits the proposal g to its maximal efficiency in the sense of achieving an exact sample.

6.3 Numerical example

6.3.1 Toy example

Consider simulating from $f(x) = \frac{2}{\sqrt{2\pi}} e^{-x^2/2} \mathbb{I}\{x \geq 0\} = \frac{e^{-x^2/2}\mathbb{I}\{x \geq 0\}}{\ell}$, that is the standard normal distribution conditional on the positive axis. Suppose we use $\mathsf{Exp}(3)$ as the proposal density and set $c = \frac{e^9}{3}$ to ensure the the density of $\mathsf{Exp}(3)$, having scaled by c, bounds $e^{-x^2/2}$. Figure 6.1 estimates how the factor $k \in (0, 1)$ in $\gamma = kc$ affects the probability of achieving a sample from ν and from f within the Markov chain.

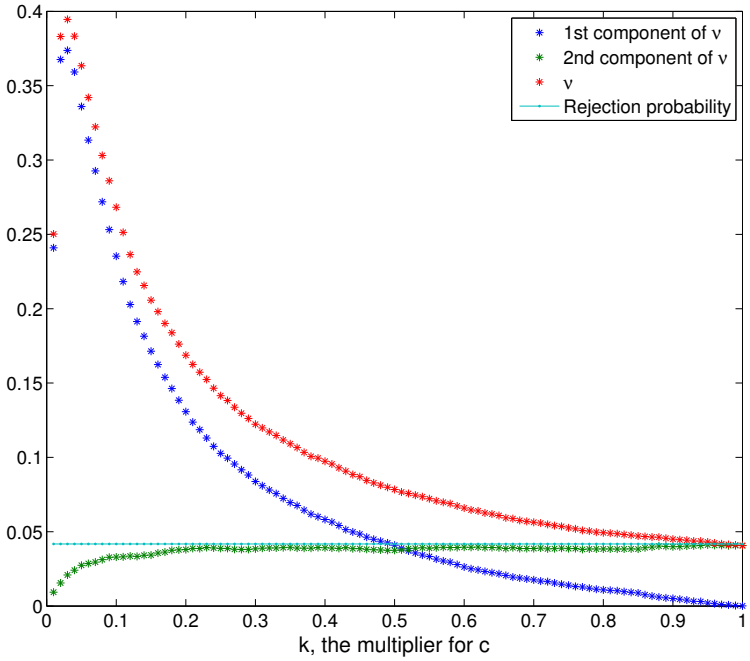

Figure 6.1: The green dots are the probabilities of having samples from f, the red dots are the probability of having samples from ν, and the blue dots are the probability of having samples from ν but not f. The horizontal line is $\frac{\sqrt{2\pi}}{2c} = 0.042$, which is the probability of retaining a draw in rejection sampling.

Of course the choice of $\mathsf{Exp}(3)$ as the proposal is sub-optimal, one can choose $\mu > 0$ for which $\mathsf{Exp}(\mu)$ is a more efficient proposal here.

6.3.2 Women wage dataset

When $\gamma = 0.5c$, the probability of regeneration is estimated to be 0.45 while the probability of an exact sample is estimated to be 0.29.

The table below summarizes the posterior distribution for the coefficients when we use the entire Markov chain.

	mean	0.25-quantile	0.975-quantile	st. deviation
kidslt6	-9.02×10^2	-1.12×10^3	-6.84×10^2	1.12×10^2
kidsge6	-1.63×10^1	-9.24×10^1	5.98×10^1	3.91×10^1
age	-5.47×10^1	-6.96×10^1	-4.03×10^1	7.47
edu	8.12×10^1	3.90×10^1	1.24×10^2	2.20×10^1
exper	1.33×10^2	9.87×10^1	1.67×10^2	1.75×10^1
nwifeinc	-8.98	-1.78×10^1	-1.86×10^{-1}	4.51
expersq	-1.89	-2.97	-8.22×10^{-1}	5.45×10^{-1}

The table below summarizes the posterior distribution for the coefficients when we only consider exact draws i.e. conditional on $B = 2$.

	mean	0.25-quantile	0.975-quantile	st. deviation
kidslt6	-9.01×10^2	-1.12×10^3	-6.85×10^2	1.12×10^2
kidsge6	-1.67×10^1	-9.13×10^1	6.00×10^1	3.93×10^1
age	-5.47×10^1	-6.98×10^1	-4.02×10^1	7.57
edu	8.11×10^1	3.92×10^1	1.24×10^2	2.19×10^1
exper	1.33×10^2	9.87×10^1	1.66×10^2	1.72×10^1
nwifeinc	-8.97	-1.79×10^1	-3.03×10^{-1}	4.56
expersq	-1.89	-2.91	-8.43×10^{-1}	5.30×10^{-1}

6.4 Concluding remarks for this chapter

In this chapter we have proposed a new sampler which we call the Reject-Regenerate sampler. The sampler is based of an usual independence sampler with the assumption that the proposal density bounds the target density in a certain way. The proposed algorithm identifies regeneration times within the Markov chain, and in the event of a regeneration, with some probability, the Markov chain achieves an exact draw from the target.

The validity of this sampler is done by careful analyses of an independence sampler by writing it as a nested mixture. To be specific, the transition part of the kernel is decomposed into a regenerative component and a non-regenerative component. Whenever a draw is made from the regenerative component, it initializes a new regenerative cycle. This is done by retrospectively checking, so that at no point of the algorithm do we need to simulate from some complicated mixture. We further decompose the regenerative component into a mixture that includes the target density. in this manner we have an independence sampler whose regeneration times can be identified, and whenever a new regenerative cycle is initialized, there is chance that the cycle starts with an exact draw from the target density.

Chapter 7

Concluding remarks for this thesis

This thesis is motivated by Bayesian posterior inference in which practitioners often need to integrate with respect to some intractable posterior probability distribution. Since these integrals do not have analytic forms, practitioners call for Monte Carlo methods.

In Chapters 2 and 3, we have shown that some Bayesian posterior inference, even when the dimension is large, exhibit efficient rejection samplers. This is done by constructing optimally tilted sequential proposal densities for these posterior densities. We have also tested these samplers on real datasets and they provide promising inferences. Rejection samplers give iid exact draws, and this way, their error analyses are simpler than their Markov chain counterparts.

Since these rejection samplers are bound to fail as the dimensions of the posterior densities increase, one eventually has to resort to MCMC samplers. Although MCMC samplers are feasible, their error analyses are a lot more difficult, and the MCMC community has devoted a lot of effort on this topic. Different analytic error bounds and precise statements regarding the validity of these bounds are available in the literature, however we observe that these error bounds do not receive as much attention as they deserve among practitioners. We believe this is mainly due to the difficulty in implementation.

In Chapter 5 we introduce novel output diagnostics for geometrically ergodic Markov chains whose regeneration times are identifiable. We argue that our methods stand on better theoretical footings compared to the conventional graphical diagnostics, and at the same time, are simple enough to implement. Our novel MCMC diagnostics also reveal that sequential proposal densities such as the ones studied in Chapters 2 and 3 render fast mixing independent samplers. Consequently, we demonstrate the values of these sequential proposals outside the framework of rejection sampling.

The final contribution is given in Chapter 6 where we propose our novel Reject-Regenerate algorithm. It is essentially an independence sampler, however we have introduced regeneration events in this sampler in such a way that when regeneration occurs, there is a chance that the new regenerative cycle is initialized perfectly from the target posterior density.

Appendix A

Appendix

A.1 Proof of theorem 5.2.1

The proof uses the following two lemmas.

Lemma A.1.1 (Uniform bias estimate). Suppose $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$ is a zero-delayed discrete regenerative process with regeneration times $0 = T_0 < T_1 < T_2 < \cdots$, where $T_n = M_1 + \cdots + M_n$, and stationary distribution $\mathbb{Q}$. Let $\mathbb{E}\exp(\varepsilon_1 M) < \infty$ for some $\varepsilon_1 > 0$ and let $\mathbb{Q}_t$ be the distribution of a state drawn at random from the whole history of the chain up until time t, that is, drawn at random from $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_t$. Then, we have for some $\varepsilon \in (0, \varepsilon_1]$

$$\sup_A |\mathbb{Q}_t[A] - \mathbb{Q}[A]| \leq \frac{\mathbb{E}M_1^2 + \mathbb{E}M_1}{2t\mathbb{E}M_1} + \mathscr{O}(\exp(-\varepsilon t)) \, .$$

Proof. The proof follows closely the ideas in [46]. Let $u(k) = \sum_{j=0}^{k} \mathbb{P}[T_j = k] = \mathbb{P}[\exists j : T_j = k]$ denote the renewal measure, and define the convolution operator $(a * b)(t) = \sum_{k=0}^{t} a(t-k)b(k)$ between two functions a and b. Further, define

$$e_A(t) := t(\mathbb{Q}[A] - \mathbb{Q}_t[A]) = \mathbb{E}\sum_{k=1}^{t} Z_k(A),$$

where $Z_k(A) = \mathbb{Q}[A] - \mathbb{I}\{\boldsymbol{X}_k \in A\}$. Wald's identity implies that

$$\mathbb{E}\sum_{k=1}^{M_1} Z_k(A) = \mathbb{Q}[A]\mathbb{E}M_1 - \mathbb{E}\sum_{k=1}^{M_1} \mathbb{I}\{\boldsymbol{X}_k \in A\} = 0.$$

Thus, we can then verify that e_A satisfies the renewal equation

$$e_A(t) = (v_A * u)(t),$$

where

$$\begin{aligned} v_A(t) &:= \mathbb{E}\left[\sum_{k=1}^{M_1} Z_k(A) - \sum_{k=1}^{t} Z_k(A); M_1 > t\right] \\ &= \mathbb{E}\left[\sum_{k=t+1}^{M_1} Z_k(A); M_1 > t\right] \end{aligned}$$

with

$$|v_A(t)| \leq \mathbb{E}\left[M_1 - t; M_1 > t\right]$$

Since $\mathbb{E}\exp(\varepsilon_1 M_1) < \infty$, then there exists some $\varepsilon_2 \in (0, \varepsilon_1]$ such that $\mathbb{E}M_1 \exp(\varepsilon_2 M_1) \leq \infty$, and therefore

$$\begin{aligned} |v_A(t)| \leq \mathbb{E}\left[M_1; M_1 \geq t\right] &\leq \frac{\mathbb{E}[M_1 \exp(\varepsilon_2 M_1); M_1 \geq t]}{\exp(\varepsilon_2 t)} \\ &\leq \frac{\mathbb{E}[M_1 \exp(\varepsilon_2 M_1)]}{\exp(\varepsilon_2 t)} = \mathscr{O}(\exp(-\varepsilon_2 t)) \, . \end{aligned}$$

An application of Theorem 2.1 on Page 196 in [3] yields for some $\varepsilon \in (0, \varepsilon_2]$:

$$e_A(t) = \frac{\sum_{k\geq 0} v_A(k)}{\mathbb{E}M_1} + \mathscr{O}(\exp(-\varepsilon t)), \quad t \uparrow \infty$$

uniformly in A. In other words,

$$\sup_A |e_A(t)| \leq \frac{\sum_{k\geq 0} \sup_A |v_A(k)|}{\mathbb{E}M} + \mathscr{O}(\exp(-\varepsilon t)), \quad t \uparrow \infty$$

Simplifying the upper bound $\sum_{k\geq 0} \sup_A |v_A(k)| \leq \sum_{k\geq 0} \mathbb{E}[M_1 - k; M_1 > k] = \frac{\mathbb{E}M_1^2 + \mathbb{E}M_1}{2}$ yields the desired result. □

Lemma A.1.2 (Time-average bound on total variation distance). Let $\boldsymbol{X}_1 \sim \nu$ and R be a $\mathbb{N}$-valued random variable that is not necessarily independent of $\boldsymbol{X}_1$. Denoting $\kappa_j(\cdot \,|\, \cdot)$ as the j-th step Markov transition kernel on $\mathcal{X} \subseteq \mathbb{R}^d$ and suppose that $\mathbb{E}[\kappa_j(\cdot \,|\, \boldsymbol{X}_1)]$ exhibits a density $f_j(\cdot)$ with respect to the Lebesgue measure on $\mathbb{R}^d$ for all $j = 0, 1, 2, \ldots$, where we define $\mathbb{E}[\kappa_0(\cdot \,|\, \boldsymbol{X}_1)] = \nu$. If $\kappa_j(\cdot \,|\, \cdot)$ has an invariant probability density π with respect to the Lebesgue measure, then for any positive integer t,

$$\|\mathbb{E}[\kappa_{t+R}(\cdot \,|\, \boldsymbol{X}_1)] - \pi\|_{\mathrm{TV}} \leq \|\mathbb{E}[\tfrac{1}{t+R}\textstyle\sum_{j=1}^{t+R} \kappa_j(\cdot \,|\, \boldsymbol{X}_1)] - \pi\|_{\mathrm{TV}},$$

where by Tonelli's theorem

$$\mathbb{E}[\kappa_{t+R}(\cdot \,|\, \boldsymbol{X}_1)] = \textstyle\sum_{r=0}^{\infty} \mathbb{P}[R = r]\mathbb{E}[\kappa_{t+r}(\cdot \,|\, \boldsymbol{X}_1)]$$

and

$$\mathbb{E}[\tfrac{1}{t+R}\textstyle\sum_{j=1}^{t+R} \kappa_j(\cdot \,|\, \boldsymbol{X}_1)] = \textstyle\sum_{r=0}^{\infty} \mathbb{P}[R = r] \times \tfrac{1}{t+r}\sum_{j=1}^{t+r} \mathbb{E}[\kappa_j(\cdot \,|\, \boldsymbol{X}_1)].$$

Proof. The idea of the proof is to construct a coupling algorithm between a state sampled from the history and the last state of the Markov chain.

We first note that for any measurable set A,

$$\sum_{r=0}^{n} \mathbb{P}[R = r]\mathbb{E}[\kappa_{t+r}(A \,|\, \boldsymbol{X}_1)] \quad \text{and} \quad \sum_{r=0}^{n} \mathbb{P}[R = r] \times \frac{1}{t+r}\sum_{j=1}^{t+r} \mathbb{E}[\kappa_j(A \,|\, \boldsymbol{X}_1)]$$

are bounded above by 1, non-decreasing in n, and exhibit densities with respect to the Lebesgue measure. So by Vitali-Hahn-Saks theorem, $\mathbb{E}[\kappa_{t+R}(\cdot \,|\, \boldsymbol{X}_1)]$ and $\mathbb{E}[\frac{1}{t+R}\sum_{j=1}^{t+R} \kappa_j(\cdot \,|\, \boldsymbol{X}_1)]$ are well-defined probability measures on $\mathcal{X}$. Moreover, their densities with respect to the Lebesgue measure are

$$\sum_{r=0}^{\infty} \mathbb{P}[R = r] f_{t+r}(\cdot) \quad \text{and} \quad \sum_{r=0}^{\infty} \mathbb{P}[R = r] \times \frac{1}{t+r}\sum_{j=1}^{t+r} f_j(\cdot).$$

Next, let

$$v_s(\cdot) = \frac{\mathbb{P}[R = s] \times \frac{1}{t+s}\sum_{j=1}^{t+s} f_j(\cdot)}{\sum_{r=0}^{\infty} \mathbb{P}[R = r] \times \frac{1}{t+r}\sum_{j=1}^{t+r} f_j(\cdot)}, \qquad \text{for } s = 0, 1, 2, \ldots,$$

and

$$w_k^{(s)}(\cdot) = \frac{\frac{1}{t+s} f_k(\cdot)}{\frac{1}{t+s}\sum_{j=1}^{t+s} f_j(\cdot)}.$$

Further, let $\boldsymbol{Z} \sim \mathbb{E}[\frac{1}{t+R}\sum_{j=1}^{t+R}\kappa_j(\cdot \,|\, \boldsymbol{X}_1)]$ and $\boldsymbol{Y} \sim \pi$ are maximally coupled so that

$$\mathbb{P}[\boldsymbol{Z} \neq \boldsymbol{Y}] = \|\mathbb{E}[\tfrac{1}{t+R}\textstyle\sum_{j=1}^{t+R}\kappa_j(\cdot \,|\, \boldsymbol{X}_1)] - \pi\|_{\mathrm{TV}}.$$

Suppose that the coupled pair $(\tilde{\boldsymbol{Z}}, \tilde{\boldsymbol{Y}})$ are sampled according to the following algorithm.

Algorithm 8 : Simulating $(\tilde{\boldsymbol{Z}}, \tilde{\boldsymbol{Y}})$

Require: $(\boldsymbol{Z}, \boldsymbol{Y})$, $v_r(\cdot)$, $w_j^{(r)}(\cdot)$, and $\kappa_j(\cdot \,|\, \cdot)$ for $j, r = 0, 1, 2, \ldots$.

Compute $v_r \leftarrow v_r(\boldsymbol{Z})$ for $r = 0, 1, 2, \ldots$.

Simulate $\rho \in \{0, 1, 2, \ldots\}$ with $\mathbb{P}[\rho = r] = v_r$.

Compute $w_j = w_j^{(r)}(\boldsymbol{Z})$

Simulate $\tau \in \{1, \ldots, t+\rho\}$ with $\mathbb{P}[\tau = j] = w_j$.

Simulate $\tilde{\boldsymbol{Z}} \sim \kappa_{t+\rho-\tau}(\cdot \,|\, \boldsymbol{Z})$.

if $\boldsymbol{Z} = \boldsymbol{Y}$, **then**

Set $\tilde{\boldsymbol{Y}} \leftarrow \tilde{\boldsymbol{Z}}$

else

Simulate $\tilde{\boldsymbol{Y}} \sim \kappa(\cdot \,|\, \boldsymbol{Y})$.

return $(\tilde{\boldsymbol{Z}}, \tilde{\boldsymbol{Y}})$.

Notice that marginally the unconditional probability law of $\tilde{\boldsymbol{Z}}$ is

$$\begin{aligned}
\int \sum_{r=0}^{\infty}\sum_{k=1}^{t+r} w_k^{(r)} v_r(\boldsymbol{z})\kappa_{t+r-k}(\cdot \,|\, \boldsymbol{z}) \times \mathbb{E}\frac{1}{t+R}\sum_{j=1}^{t+R} f_j(\boldsymbol{z})\, d\boldsymbol{z} & \\
&= \sum_{r=0}^{\infty}\mathbb{P}[R=r] \times \frac{1}{t+r}\sum_{k=1}^{t+r}\int \kappa_{t+r-k}(\cdot \,|\, \boldsymbol{z}) f_k(\boldsymbol{z})\, d\boldsymbol{z} \\
&= \sum_{r=0}^{\infty}\mathbb{P}[R=r] f_{t+r}(\cdot).
\end{aligned}$$

In other words $\tilde{\boldsymbol{Z}} \sim \mathbb{E}\kappa_{t+R}(\cdot \,|\, \boldsymbol{X}_1)$. It is also clear that $\tilde{\boldsymbol{Y}} \sim \pi(\cdot)$. Finally, we have that

$$\begin{aligned}
\mathbb{P}[\tilde{\boldsymbol{Z}} = \tilde{\boldsymbol{Y}}] &= \mathbb{P}[\tilde{\boldsymbol{Z}} = \tilde{\boldsymbol{Y}}, \boldsymbol{Z} = \boldsymbol{Y}] + \mathbb{P}[\tilde{\boldsymbol{Z}} = \tilde{\boldsymbol{Y}}, \boldsymbol{Z} \neq \boldsymbol{Y}] \\
&= \mathbb{P}[\boldsymbol{Z} = \boldsymbol{Y}] + \mathbb{P}[\tilde{\boldsymbol{Z}} = \tilde{\boldsymbol{Y}}, \boldsymbol{Z} \neq \boldsymbol{Y}] \\
&\geq \mathbb{P}[\boldsymbol{Z} = \boldsymbol{Y}]
\end{aligned}$$

so that

$$\mathbb{P}[\tilde{\boldsymbol{Z}} \neq \tilde{\boldsymbol{Y}}] \leq \|\mathbb{E}[\tfrac{1}{t+R}\textstyle\sum_{j=1}^{t+R}\kappa_j(\cdot \,|\, \boldsymbol{X}_1)] - \pi\|_{\mathrm{TV}}.$$

However, the pair $(\tilde{\boldsymbol{Z}}, \tilde{\boldsymbol{Y}})$ is not necessarily maximally coupled so that by the coupling inequality, we have

$$\|\mathbb{E}[\kappa_{t+R}(\cdot \,|\, \boldsymbol{X}_1)] - \pi\|_{\mathrm{TV}} \leq \|\mathbb{E}[\tfrac{1}{t+R}\sum_{j=1}^{t+R} \kappa_j(\cdot \,|\, \boldsymbol{X}_1)] - \pi\|_{\mathrm{TV}}.$$

□

With the two lemmas, we can now proceed to the proof of the theorem. Let $\mathbb{Q}_t[A]$ be the distribution of a state $\boldsymbol{X}$ picked at random from the Markov chain states: $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_t$. In other words,

$$\mathbb{Q}_t[A] := \frac{1}{t}\sum_{k=1}^{t} \mathbb{E}[\kappa_k(A \,|\, \boldsymbol{X}_1)],$$

where $\mathbb{E}[\kappa_k(A \,|\, \boldsymbol{X}_1)] = \mathbb{P}[\boldsymbol{X}_k \in A]$ for all $k \geq 0$. By assumption, the Markov chain is geometrically ergodic, that is, the distribution of the length M of a regenerative cycle of the chain is light-tailed. In other words, $\mathbb{E}\exp(\epsilon_1 M) < \infty$ for some $\epsilon_1 > 0$. The process $\boldsymbol{X}_1, \boldsymbol{X}_1, \ldots$ is also a zero-delayed regenerative process, because by assumption the initial $\boldsymbol{X}_1$ commences a new cycle. Therefore, the conditions of Lemma A.1.1 (see above) are satisfied and we have:

$$\|\mathbb{Q}_t - \pi\|_{\mathrm{TV}} \leq \frac{\mathbb{E}M^2 + \mathbb{E}M}{2t\mathbb{E}M} + \mathscr{O}(\exp(-\varepsilon t)) \tag{A.1}$$

for some $\epsilon \in (0, \epsilon_1]$. In addition, as a special case of Lemma A.1.2, if $\mathbb{P}[R = 0] = 1$, the lemma reduces to saying that the distribution of the final state of the Markov chain, $\boldsymbol{X}_t$, is closer to π than a state picked at random from the history of the chain up until time t: $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_t$. In other words,

$$\|\mathbb{E}[\kappa_t(\cdot \,|\, \boldsymbol{X}_1)] - \pi\|_{\mathrm{TV}} \leq \|\mathbb{Q}_t - \pi\|_{\mathrm{TV}} \ . \tag{A.2}$$

The result of the theorem then follows by combining (A.1)+(A.2). □

A.2 Proof of theorem 5.2.2

Again, suppose $\boldsymbol{X}_1, \boldsymbol{X}_1, \ldots$ is a zero-delayed discrete regenerative process with regeneration times $0 = T_0 < T_1 < T_2 < \cdots$, where $T_n = M_1 + \cdots + M_n$, and stationary distribution $\mathbb{Q}$. Let $\overline{\mathbb{Q}}_t$ be the distribution of a state drawn at random from the whole history of the chain up until time $T_{N(t)}$, that is, drawn at random from $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{T_{N(t)}}$. Again let $H_r(A) = \sum_{k=T_{r-1}}^{T_r} \mathbb{I}\{\boldsymbol{X}_k \in A\}$ for some $A \in \mathscr{A}$ so that Wald's identity and renewal theorem together imply

$$\mathbb{E}Z_r(A) := \mathbb{E}(H_r(A) - M_r\mathbb{Q}(A)) = 0.$$

In this way, we can write

$$\begin{aligned}
\overline{\mathbb{Q}}_t(A) - \mathbb{Q}(A) &= \mathbb{E}\frac{\sum_{r=1}^{N(t)} H_r(A) - M_r\mathbb{Q}(A)}{T_{N(t)}} \\
&= \mathbb{E}\frac{\sum_{r=1}^{N(t)} Z_r(A)}{T_{N(t)}} \\
&= \mathbb{E}\frac{\frac{1}{t}\sum_{r=1}^{N(t)} Z_r(A)}{1 + R(t)/t} \\
&= \mathbb{E}\left(\frac{\frac{1}{t}\sum_{r=1}^{N(t)} Z_r(A)}{1 + R(t)/t} - \frac{\sum_{r=1}^{N(t)} Z_r(A)}{t}\right) \\
&= \frac{1}{t}\mathbb{E}\frac{-R(t)}{1 + R(t)/t}\bar{Z}_t(A).
\end{aligned}$$

Where we have defined $\bar{Z}_t(A) = \frac{1}{t}\sum_{k=1}^{N(t)} Z_k(A)$ and have used the fact that $\mathbb{E}\sum_{r=1}^{N(t)} Z_r(A) = \mathbb{E}[N(t)]\mathbb{E}Z_r(A) = 0$ in the second last line. Then, using the fact that $\frac{1}{1+R(t)/t} \leq 1$, we obtain the uniform bound

$$\begin{aligned}
|\overline{\mathbb{Q}}_t(A) - \mathbb{Q}(A)| &= \frac{1}{t}\left|\mathbb{E}\frac{-R(t)}{1 + R(t)/t}\bar{Z}_t(A)\right| \\
&\leq \frac{1}{t}\left|R(t)\bar{Z} - t(A)\right| \\
&\leq \frac{\sqrt{\mathbb{E}R^2(t)\mathbb{E}\bar{Z}_t^2(A)}}{t} \\
&= \frac{\sqrt{\mathbb{E}R^2(t)}}{t}\sqrt{\mathbb{E}[N(t)]\mathbb{E}Z^2(A)/t^2} \\
&\leq \frac{\sqrt{\mathbb{E}[R^2(t)]}}{t^{3/2}}\sqrt{\frac{m_2}{m_1}(1 + \mathbb{E}R(t)/t)},
\end{aligned}$$

where the third last line uses Wald's identity and the last line uses the relation $\mathbb{E}R(t) = m_1\mathbb{E}[N(t)] - t$ and $Z^2(A) < M^2$. Finally, applying Lorden's inequality [78] gives the desired result. □

A.3 Proof of theorem 5.2.3

Note that if $m_{p+5} < \infty$ for some $p \geq 0$, then we have [47]

$$\mathbb{E}R(t) = r + o(1/t^{p+3})$$

where $r := \frac{m_2+m_1}{2m_1}$. Since $0 \le \frac{1}{1+x} - 1 + x \le x^2$ for $x \ge 0$, we have the error bound:

$$\begin{aligned}
\left|\overline{\mathbb{Q}}_t(A) - \mathbb{Q}(A)\right| &= \left|\mathbb{E}\left(\frac{1}{1+R(t)/t} - 1\right)\bar{Z}_t(A)\right| \\
&\le \frac{|\mathbb{E}R(t)\bar{Z}_t(A)|}{t} + \left|\mathbb{E}\left(\frac{1}{1+R(t)/t} - 1 + \frac{R(t)}{t}\right)\bar{Z}_t(A)\right| \\
&\le \frac{\mathbb{E}R(t)\bar{Z}_t(A)}{t} + \frac{\mathbb{E}R^2(t)|\bar{Z}_t(A)|}{t^2} \\
&\le \frac{|\mathbb{E}R(t)\sum_{r=1}^{N(t)} Z_r(A)|}{t^2} + \frac{\sqrt{\mathbb{E}[R^4(t)]\mathbb{E}[\bar{Z}_t^2(A)]}}{t^2}.
\end{aligned}$$

Since geometric ergodicity ensures $m_5 < \infty$, we apply Lordern's inequality [78] which gives us

$$\mathbb{E}[R^4(t)] \le \frac{6}{5}\frac{m_5}{m_1},$$

and

$$\mathbb{E}[\bar{Z}_t^2(A)] = \frac{\mathbb{E}[N(t)]}{t^2}\mathbb{E}[Z_1^2(A)] \le (1 + [R(t)]/t)m_2/t \le (1 + m_2/(m_1 t))m_2/t.$$

So we bound

$$\sqrt{\mathbb{E}[R^4(t)]\mathbb{E}[\bar{Z}_t^2(A)]} \le \frac{\sqrt{(6/5)(m_1 + m_2/t)m_2 m_5}}{m_1 t^{1/2}}.$$

For the first term, we verify that $e_A(t) := \mathbb{E}R(t)\sum_{r=1}^{N(t)} Z_r(A) < \infty$ satisfies the renewal equation $e_A(t) = (u * v_A)(t)$ with $v_A(t) := \mathbb{E}[R(t)Z_1(A)] = \mathbb{E}[(R(t) - r)Z_1(A)]$, see [5, Page 25]. We have

$$\begin{aligned}
|v_A(t)| &= |\mathbb{E}[(R(t) - r)Z_1(A); M_1 > t] + \mathbb{E}[(R(t) - r)Z_1(A); M_1 \le t]| \\
&\le \mathbb{E}[|M_1 - r|M_1; M_1 > t] + \mathbb{E}[|r(t - M_1) - r|M_1; M_1 \le t]
\end{aligned}$$

For the first term, we otain:

$$\mathbb{E}[|M_1 - r|M_1; M_1 > t] = \mathscr{O}(\mathbb{E}[M_1^{p+5}; M_1 > t]/t^{p+3}) = o(1/t^{p+3}).$$

For the second term,

$$\begin{aligned}
\mathbb{E}[|r(t - M_1) - r|M_1; M_1 \le t] &\le \mathbb{E}[|r(t - M_1) - r|M_1; M_1 \le t/2] + \mathbb{E}[|r(t - M_1) - r|M_1; M_1 \ge t/2] \\
&\le \sup_{s>t/2} |r(s) - r|\mathbb{E}M_1 + \sup_{s<t/2} |r(s) - r|\mathbb{E}[M_1; M > t/2] \\
&= o(1/t^{p+3}) + o(1/t^{p+4}).
\end{aligned}$$

Hence, we have the convergence uniformly in A:

$$\begin{aligned} e_A(t) &= \frac{\mathbb{E}(M_1 - 1 - 2r)M_1 Z_1(A)}{2m_1} + o(1/p^{p+2}) \\ &\leq \frac{\mathbb{E}|M_1 - 1 - 2r|M_1^2}{2m_1} + o(1/t^{p+2}). \end{aligned}$$

Putting it all together, we obtain

$$\delta_t^{\text{reg-seq}} \leq \sup_A |\overline{\mathbb{Q}}_t - \mathbb{Q}(A)| \leq \frac{\mathbb{E}|M_1 - 1 - 2r|M_1^2}{2m_1 t^2} + \frac{\sqrt{(6/5)(m_1 + m_2/t)m_2 m_5}}{m_1 t^{3/2}} + o(1/t^{p+4})$$

where the exponential convergence comes from the fact that $m_p < \infty$ for all $p > 0$. □

A.4 Proof of Theorem 5.2.4

Recall that $\boldsymbol{X}_0, \boldsymbol{X}_1, \ldots$ is a zero-delayed discrete regenerative process with first regenerative time $T_1 = T_0 + M_1 = 0 + M_1$. Define the function $g_A(t) = \mathbb{P}[\boldsymbol{X}_t \in A | M_1 > t]$ and note that $\mathbb{P}[\boldsymbol{X}_t \in A]$ satisfies the renewal equation:

$$\begin{aligned} \mathbb{P}[\boldsymbol{X}_t \in A] &= \sum_{k>t} \mathbb{P}[\boldsymbol{X}_t \in A, M_1 = k] + \sum_{k=1}^{t} \mathbb{P}[\boldsymbol{X}_t \in A, M_1 = k] \\ &= \mathbb{P}[\boldsymbol{X}_t \in A, M_1 > t] + \sum_{k=1}^{t} \mathbb{P}[\boldsymbol{X}_t \in A | M_1 = k]\mathbb{P}[M_1 = k] \\ &= g_A(t)\mathbb{P}[M_1 > t] + \sum_{k=1}^{t} \mathbb{P}[\boldsymbol{X}_{t-k} \in A]\mathbb{P}[M_1 = k] \ . \end{aligned}$$

By a similar argument, $\mathbb{E}g_A(E(t))$ also satisfies the renewal equation:

$$\mathbb{E}g_A(E(t)) = g_A(t)\mathbb{P}[M_1 > t] + \sum_{k=1}^{t} \mathbb{E}g_A(E(t-k))\mathbb{P}[M_1 = k] \ .$$

Note that the term $g_A(t)\mathbb{P}[M_1 > t]$ is common to both renewal equations. Since this term is continuous and bounded from above by the directly-Riemann-integrable function $\mathbb{P}[M_1 > t]$, the key renewal theorem (see [3, Proposition 1.3, Page 170]) implies that the two renewal equations above have the same unique solution. Hence,

$$\mathbb{P}[\boldsymbol{X}_t \in A] = \mathbb{E}g_A(E(t)) \ .$$

Using this identity, we have ($E(\infty)$ is a stationary version of $E(t)$):

$$\begin{aligned}\sup_A |\mathbb{P}[\boldsymbol{X}_t \in A] - \mathbb{Q}[A]| &= \sup_A |\mathbb{E}g_A(E(t)) - \mathbb{E}g_A(E(\infty))| \\ &\leq \sup_{g:\|g\|_\infty \leq 1} |\mathbb{E}g(E(t)) - \mathbb{E}g(E(\infty))| \\ &\leq 2\sup_A |\mathbb{P}[E(t) \in A] - \mathbb{P}[E(\infty) \in A]| \,,\end{aligned}$$

where the supremum in the second last line is over all bounded functions g with $\|g\|_\infty \leq 1$. The last line shows that the convergence rate of $E(t)$ determines the convergence rate of the Markov chain $\{\boldsymbol{X}_t, t \geq 0\}$. □

A.5 Proof of Theorem 5.2.5

Define $R(t) := T_{N(t)} - t$ (the remaining lifetime) and note that $\frac{1}{1+R(t)/t} \leq 1$ for all $t \geq 0$. In addition, we have Theorem 4.2.6: $\mathbb{E}R(t) \leq \frac{m_2}{m_1}$, which can be rewritten as ($\mathbb{E}T_{N(t)} = m_1\mathbb{E}N(t)$)

$$m_1\mathbb{E}N(t) \leq t + \frac{m_2}{m_1}.$$

To proceed, write

$$\begin{aligned}\hat{q}_{T_{N(t)}} - q &= \frac{\sum_{k=1}^{N(t)} H_k - M_k q}{T_{N(t)}} \\ &= \mathbb{E}\frac{\sum_{k=1}^{N(t)} Z_k}{T_{N(t)}} = \mathbb{E}\frac{\frac{1}{t}\sum_{k=1}^{N(t)} Z_k}{1 + R(t)/t}.\end{aligned}$$

Then, using the second part of Theorem 4.2.5, $\mathbb{E}(\sum_{k=1}^{N(t)} Z_k)^2 = \mathbb{E}N(t)\mathbb{E}Z_1^2$, we obtain

$$\begin{aligned}\mathbb{E}(\hat{q}_{T_{N(t)}} - q)^2 &= \frac{1}{t^2}\mathbb{E}\left[\frac{(\sum_{k=1}^{N(t)} Z_k)^2}{(1 + R(t)/t)^2}\right] \\ &\leq \frac{1}{t^2}\mathbb{E}\left(\sum_{k=1}^{N(t)} Z_k\right)^2 = \frac{\mathbb{E}N(t)\mathbb{E}Z_1^2}{t^2} \\ &\leq \mathbb{E}Z_1^2\left(\frac{1}{tm_1} + \frac{m_2}{t^2 m_1^2}\right).\end{aligned}$$

□

A.6 Background: Gibbs sampler for the Bayesian Lasso

The first key insight in [91] is that a $\mathsf{Laplace}(0, 1/\lambda)$ density is in fact a Gaussian-scale mixture [2]. In particular, for each β_j, $j \in \{1, \ldots, p\}$, we have the identity,

$$\frac{\lambda}{2\sqrt{\sigma^2}} \exp\left(-\frac{\lambda}{\sqrt{\sigma^2}}|\beta_j|\right) = \int_0^\infty \frac{1}{\sqrt{2\pi s_j}} \exp\left(-\frac{\beta_j^2}{2s_j}\right) \times \frac{\lambda^2}{2\sigma^2} \exp\left(-\frac{\lambda^2}{2\sigma^2} s_j\right) \; ds_j.$$

It follows form the change of variable $\tau_j = \sigma^2/s_j$,

$$\frac{\lambda}{2\sqrt{\sigma^2}} \exp\left(-\frac{\lambda}{\sqrt{\sigma^2}}|\beta_j|\right) = \int_0^\infty \frac{\lambda^2}{2\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\beta_j^2 \tau_j}{2\sigma^2}\right) \exp\left(-\frac{\lambda^2}{2\tau_j}\right) \tau_j^{-3/2} \; d\tau_j. \quad \text{(A.3)}$$

Hence if one considers sampling the triplet $(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\tau}) \in \mathbb{R}^p \times \mathbb{R}_+ \times \mathbb{R}_+^p$ from the joint density

$$\pi(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\tau}|\lambda) = \frac{(\sigma^2)^{-n/2-1} \exp\left(-\frac{1}{2\sigma^2}\|\boldsymbol{y} - \mathrm{X}\boldsymbol{\beta}\|_2^2\right) \prod_{j=1}^p \frac{\lambda^2}{2\sqrt{2\sigma^2}} \exp\left(-\frac{\beta_j^2 \tau_j}{2\sigma^2}\right) \exp\left(-\frac{\lambda^2}{2\tau_j}\right) \tau_j^{-3/2}}{\ell(\lambda)}, \quad \text{(A.4)}$$

the marginal samples $(\boldsymbol{\beta}, \sigma^2)$, from samples of the triplet $(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\tau})$, have the same distribution as (5.5). This is because (A.3) implies $\int_{\mathbb{R}_+^p} \pi(\boldsymbol{\beta}, , \sigma^2, \boldsymbol{\tau}|\lambda) \, d\boldsymbol{\tau} = \pi(\boldsymbol{\beta}, \sigma^2|\lambda)$.

The form of (A.4) suggests a natural (block) Gibbs sampler that cycles between the full conditional distributions $\pi(\boldsymbol{\beta} \,|\, \sigma^2, \boldsymbol{\tau}, \lambda)$, $\pi(\sigma^2| \,|\, \boldsymbol{\beta}, \boldsymbol{\tau}, \lambda)$ and $\pi(\boldsymbol{\tau} \,|\, \boldsymbol{\beta}, \sigma^2, \lambda)$. The second key insight in [91] is that $\pi(\boldsymbol{\tau} \,|\, \boldsymbol{\beta}, \sigma^2, \lambda)$ takes the product form

$$\pi(\boldsymbol{\tau} \,|\, \boldsymbol{\beta}, \lambda, \sigma^2) \propto \prod_{j=1}^p \exp\left(-\frac{\beta_j^2 \tau_j}{2\sigma^2}\right) \exp\left(-\frac{\lambda^2}{2\tau_j}\right) \tau_j^{-3/2}.$$

This means each τ_j are conditionally independent. Moreover, the conditional distribution of τ_j is $\mathsf{Wald}(\lambda', \mu_j')$ where $\lambda' = \lambda^2$ and $\mu_j' = \sqrt{\sigma^2}\lambda/|\beta_j|$ (see, for example, [19]). Finally, it is not hard to show that

$$\pi(\boldsymbol{\beta} \,|\, \boldsymbol{\tau}, \lambda, \sigma^2) \propto \exp\left(-\frac{1}{2\sigma^2} \left(\boldsymbol{\beta} - \mathrm{A}\mathrm{X}^\top \boldsymbol{y}\right)^\top \mathrm{A}^{-1} \left(\boldsymbol{\beta} - \mathrm{A}\mathrm{X}^\top \boldsymbol{y}\right)^\top\right),$$

where $\mathrm{A}^{-1} = \mathrm{X}^\top\mathrm{X} + \mathrm{diag}(\boldsymbol{\tau})$ is a symmetric invertible matrix, and

$$\pi(\sigma^2 \,|\, \boldsymbol{\beta}, \tau, \lambda) \propto (\sigma^2)^{-n/2-p/2-1} \exp\left(-\frac{1}{2\sigma^2} \left(\|\boldsymbol{y} - \mathrm{X}\boldsymbol{\beta}\|_2^2 + \boldsymbol{\beta}^\top \mathrm{diag}(\boldsymbol{\tau})\boldsymbol{\beta}\right)\right).$$

This means, $\boldsymbol{\beta}$ conditional on $(\sigma^2, \boldsymbol{\tau}, \lambda)$, is a p-dimensional Gaussian random variable with the mean vector $\mathrm{A}\mathrm{X}^\top\boldsymbol{y}$ and the covariance matrix $\sigma^2\mathrm{A}$ while the full conditional

distribution of σ^2 is $\mathsf{InvGamma}(a', b')$ where the shape parameter $a' = n/2 + p/2$ and the scale parameter $b' = \|\boldsymbol{y} - \mathrm{X}\boldsymbol{\beta}\|_2^2/2 + \boldsymbol{\beta}^\top \mathrm{diag}(\boldsymbol{\tau})\boldsymbol{\beta}/2$.

The shrinkage parameter can be chosen by maximizing the marginal likelihood $\ell(\lambda)$. This approach is known as the 'empirical Bayes', and this optimization program can be approximately solved using the EM algorithm in [15, Page 686, Appendix C].

A.7 Proof of Lemma 5.4.1

The full conditional distributions for the Bayesian Lasso posterior are $\pi(\boldsymbol{\beta}|\sigma^2, \boldsymbol{\tau}) = \mathsf{N}(\mathrm{A}\mathrm{X}^\top \boldsymbol{y}, \sigma^2 \mathrm{A})$ where $\mathrm{A}^{-1} = \mathrm{X}^\top\mathrm{X} + \mathrm{diag}(\boldsymbol{\tau})$, $\pi(\tau_j \,|\, \boldsymbol{\beta}, \sigma^2) = \mathsf{Wald}(\lambda^2, \sigma\lambda/|\beta_j|)$, and $\pi(\sigma \,|\, \boldsymbol{\beta}, \boldsymbol{\tau}) = \mathsf{InvGamma}((n-1)/2 + p/2, b(\boldsymbol{\beta}, \boldsymbol{\tau})/2)$ where $b(\boldsymbol{\beta}, \tau) = \|\boldsymbol{y} - \mathrm{X}\boldsymbol{\beta}\|_2^2 + \boldsymbol{\beta}^\top \mathrm{diag}(\boldsymbol{\tau})\boldsymbol{\beta}$.

Consider a block Gibbs sampler with transition density: (denote σ^2 by ξ)

$$\kappa(\boldsymbol{\beta}_*, \xi_*, \boldsymbol{\tau}_* \,|\, \boldsymbol{\beta}, \xi, \boldsymbol{\tau}) = \pi(\xi_* \,|\, \boldsymbol{\beta}, \boldsymbol{\tau})\pi(\boldsymbol{\beta}_*|\xi_*, \boldsymbol{\tau})\pi(\tau_* \,|\, \boldsymbol{\beta}_*, \xi_*).$$

We proceed with the usual calculation (see [86, 99]) for establishing a minorization condition for this transition density. We begin by observing that

$$\kappa(\boldsymbol{\beta}_*, \xi_*, \boldsymbol{\tau}_* \,|\, \boldsymbol{\beta}, \xi, \boldsymbol{\tau}) = \frac{\pi(\xi_* \,|\, \boldsymbol{\beta}, \boldsymbol{\tau})\pi(\boldsymbol{\beta}_*|\xi_*, \boldsymbol{\tau})}{\pi(\xi_*|\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\tau}})\pi(\boldsymbol{\beta}_*|\xi_*, \tilde{\boldsymbol{\tau}})}\kappa(\boldsymbol{\beta}_*, \xi_*, \boldsymbol{\tau}_*|\tilde{\boldsymbol{\beta}}, \tilde{\xi}, \tilde{\boldsymbol{\tau}}).$$

Choosing $\nu(\boldsymbol{\beta}_*, \xi_*, \boldsymbol{\tau}_*) \propto \kappa(\boldsymbol{\beta}_*, \xi_*, \boldsymbol{\tau}_*|\tilde{\boldsymbol{\beta}}, \tilde{\xi}, \tilde{\boldsymbol{\tau}})$ on $\mathscr{D} = \mathbb{R}^p \times [l, u] \times [\boldsymbol{c}, \boldsymbol{d}]$, a subset of $\mathbb{R}^p \times \mathbb{R}_+ \times \mathbb{R}_+^p$, where $(\tilde{\boldsymbol{\beta}}, \tilde{\xi}, \tilde{\boldsymbol{\tau}})$ are fixed, it remains for us to find a function s such that

$$\frac{\pi(\xi_* \,|\, \boldsymbol{\beta}, \boldsymbol{\tau})\pi(\boldsymbol{\beta}_*|\xi_*, \boldsymbol{\tau})}{\pi(\xi_*|\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\tau}})\pi(\boldsymbol{\beta}_*|\xi_*, \tilde{\boldsymbol{\tau}})} \geq s(\boldsymbol{\beta}, \xi, \boldsymbol{\tau})$$

whenever $\tau_{*,j}^2 \in [c_j^2, d_j^2]$ for all j and $\xi \in [l, u]$. Denoting $\boldsymbol{w} = \mathrm{X}^\top \boldsymbol{y}$ and $\mu_{\boldsymbol{\tau}} = \mathrm{A}_{\boldsymbol{\tau}}\mathrm{X}^\top \boldsymbol{y}$, we have

$$\begin{aligned}
\frac{\pi(\xi_* \,|\, \boldsymbol{\beta}, \boldsymbol{\tau})\pi(\boldsymbol{\beta}_*|\xi_*, \boldsymbol{\tau})}{\pi(\xi_*|\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\tau}})\pi(\boldsymbol{\beta}_*|\xi_*, \tilde{\boldsymbol{\tau}})} &= \left(\frac{b(\boldsymbol{\beta}, \boldsymbol{\tau})}{b(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\tau}})}\right)^a \exp\left(-\frac{1}{2\xi_*}\left(\|y - \mathrm{X}\boldsymbol{\beta}\|_2^2 - \|y - \mathrm{X}\tilde{\boldsymbol{\beta}}\|_2^2 + \boldsymbol{\beta}^\top \mathrm{diag}(\boldsymbol{\tau})\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}^\top \mathrm{diag}(\tilde{\boldsymbol{\tau}})\tilde{\boldsymbol{\beta}}\right)\right) \times \\
&\quad \left(\frac{\det(\mathrm{A}_{\tilde{\boldsymbol{\tau}}})}{\det(\mathrm{A}_{\boldsymbol{\tau}})}\right)^{n/2} \exp\left(-\frac{(\boldsymbol{\beta}_*^2 - \boldsymbol{\mu}_{\boldsymbol{\tau}})^\top \mathrm{A}_{\boldsymbol{\tau}}^{-1}(\boldsymbol{\beta}_*^2 - \boldsymbol{\mu}_{\boldsymbol{\tau}})}{2\xi_*} + \frac{(\boldsymbol{\beta}_*^2 - \boldsymbol{\mu}_{\tilde{\boldsymbol{\tau}}})^\top \mathrm{A}_{\tilde{\boldsymbol{\tau}}}^{-1}(\boldsymbol{\beta}_*^2 - \boldsymbol{\mu}_{\tilde{\boldsymbol{\tau}}})}{2\xi_*}\right) \\
&\geq \left(\frac{b(\boldsymbol{\beta}, \boldsymbol{\tau})}{b(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\tau}})}\right)^a \exp\left(-\frac{1}{2\xi_*}\left(\|y - \mathrm{X}\boldsymbol{\beta}\|_2^2 - \|y - \mathrm{X}\tilde{\boldsymbol{\beta}}\|_2^2 + \boldsymbol{\beta}^\top \mathrm{diag}(\boldsymbol{\tau})\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}^\top \mathrm{diag}(\tilde{\boldsymbol{\tau}})\tilde{\boldsymbol{\beta}}\right)\right) \times \\
&\quad \left(\frac{\det(\mathrm{A}_{\tilde{\boldsymbol{\tau}}})}{\det(\mathrm{A}_{\boldsymbol{\tau}})}\right)^{n/2} \exp\left(\frac{1}{2\xi_*}\left(\sum_{j \in \mathscr{J}} c_j^2(\tilde{\tau}_j - \tau_j) + \sum_{j \notin \mathscr{J}} d_j^2(\tilde{\tau}_j - \tau_j) + \boldsymbol{w}^\top(\mathrm{A}_{\tilde{\boldsymbol{\tau}}} - \mathrm{A}_{\boldsymbol{\tau}})\boldsymbol{w}\right)\right) \\
&\geq \left(\frac{b(\boldsymbol{\beta}, \boldsymbol{\tau})}{b(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\tau}})}\right)^a \left(\frac{\det(\mathrm{A}_{\tilde{\boldsymbol{\tau}}})}{\det(\mathrm{A}_{\boldsymbol{\tau}})}\right)^{n/2} \exp\left(\frac{\Upsilon(\boldsymbol{\beta}, \boldsymbol{\tau}; \boldsymbol{c}, \boldsymbol{d})_+}{2u} + \frac{\Upsilon(\boldsymbol{\beta}, \boldsymbol{\tau}; \boldsymbol{c}, \boldsymbol{d})_-}{2l}\right),
\end{aligned}$$

where

$$\Upsilon(\boldsymbol{\beta}, \boldsymbol{\tau}; \boldsymbol{c}, \boldsymbol{d}) = \|y - \mathrm{X}\tilde{\boldsymbol{\beta}}\|_2^2 - \|y - \mathrm{X}\boldsymbol{\beta}\|_2^2 + \tilde{\boldsymbol{\beta}}^\top \mathrm{diag}(\tilde{\boldsymbol{\tau}})\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^\top \mathrm{diag}(\boldsymbol{\tau})\boldsymbol{\beta} + \sum_{j \in \mathscr{J}} c_j^2(\tilde{\tau}_j - \tau_j) + \sum_{j \notin \mathscr{J}} d_j^2(\tilde{\tau}_j - \tau_j) + \boldsymbol{w}^\top(\mathrm{A}_{\tilde{\boldsymbol{\tau}}} - \mathrm{A}_{\boldsymbol{\tau}})\boldsymbol{w},$$

$\mathscr{J}\{j : \tilde{\tau}_j - \tau_j \geq 0\}$, and $\{\cdot\}_+ = \max\{0, \cdot\}$, and $\{\cdot\}_- = \min\{0, \cdot\}$. It follows that, $\vartheta := \frac{s\nu}{\kappa}$, the probability of regeneration takes the following expression for $(\boldsymbol{\beta}_*, \xi_*, \boldsymbol{\tau}_*) \in \mathscr{D}$.

$$\vartheta(\boldsymbol{\beta}_*, \xi_*, \boldsymbol{\tau}_* \,|\, \boldsymbol{\beta}, \xi, \boldsymbol{\tau}) = \exp\left(\frac{\Upsilon(\boldsymbol{\beta}, \boldsymbol{\tau}; \boldsymbol{c}, \boldsymbol{d})_+}{2u} + \frac{\Upsilon(\boldsymbol{\beta}, \boldsymbol{\tau}; \boldsymbol{c}, \boldsymbol{d})_-}{2l} - \frac{\Upsilon(\boldsymbol{\beta}, \boldsymbol{\tau}; \boldsymbol{\beta}_*, \boldsymbol{\beta}_*)}{2\xi_*}\right)$$

□

References

[1] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19:716 – 723, December 1974.

[2] D. F. Andrews and C. L. Mallows. Scale mixtures of normal distributions. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 99–102, 1974.

[3] Søren Asmussen. *Applied probability and queues*, volume 51. Springer Science & Business Media, 2008.

[4] Krishna B Athreya and Peter Ney. A new approach to the limit theory of recurrent Markov chains. *Transactions of the American Mathematical Society*, 245:493–501, 1978.

[5] Hernan P Awad and Peter W Glynn. On the theoretical comparison of low-bias steady-state estimators. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 17(1):4–es, 2007.

[6] Witold Bednorz and Krzysztof LatuszyŃski. A few remarks on "Fixed-width output analysis for Markov chain Monte Carlo" by jones et al. *Journal of the American Statistical Association*, 102(480):1485–1486, 2007.

[7] Niloy Biswas, Pierre E Jacob, and Paul Vanetti. Estimating convergence of markov chains with l-lag couplings. *arXiv preprint arXiv:1905.09971*, 2019.

[8] Edward L Boone, Jason RW Merrick, and Matthew J Krachey. A Hellinger distance approach to MCMC diagnostics. *Journal of Statistical Computation and Simulation*, 84(4):833–849, 2014.

[9] Zdravko Botev and Leo Belzile. *Truncated Multivariate Normal and Student Distributions*, 2019. R package version 2.0.

[10] Zdravko Botev, Yi-Lung Chen, Pierre L'Ecuyer, Shev MacNamara, and Dirk P Kroese. Exact posterior simulation from the linear LASSO regression. In *2018 Winter Simulation Conference (WSC)*, pages 1706–1717. IEEE, 2018.

[11] Zdravko I Botev. The normal law under linear restrictions: simulation and estimation via minimax tilting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(1):125–148, 2017.

[12] Zdravko I Botev and Pierre L'Ecuyer. Efficient probability estimation and simulation of the truncated multivariate student-t distribution. In *2015 Winter Simulation Conference (WSC)*, pages 380–391. IEEE, 2015.

[13] Zdravko I Botev and Pierre L'Ecuyer. Sampling conditionally on a rare event via generalized splitting. *INFORMS Journal on Computing*, 2020.

[14] Anthony E Brockwell and Joseph B Kadane. Identification of regeneration times in MCMC simulation, with application to adaptive schemes. *Journal of Computational and Graphical Statistics*, 14(2):436–458, 2005.

[15] George Casella. Empirical Bayes Gibbs sampling. *Biostatistics*, 2(4):485–500, 2001.

[16] Kung Sik Chan and Charles J. Geyer. Discussion: Markov chains for exploring posterior distributions. *Ann. Statist.*, 22(4):1747–1758, 12 1994. doi: 10.1214/aos/1176325754. URL `https://doi.org/10.1214/aos/1176325754`.

[17] Ming-Hui Chen and John J Deely. Bayesian analysis for a constrained linear multiple regression problem for predicting the new crop of apples. *Journal of Agricultural, Biological, and Environmental Statistics*, pages 467–489, 1996.

[18] Ming-Hui Chen, Joseph G Ibrahim, and Qi-Man Shao. Monte carlo methods in bayesian computation. page 6, 2000.

[19] R. S. Chhikara and J. L. Folks. *The Inverse Gaussian Distribution: Theory: Methodology, and Applications*, volume 95. Marcel Dekker, Inc., 1988.

[20] Siddhartha Chib. Bayes inference in the Tobit censored regression model. *Journal of Econometrics*, 51(1-2):79–99, 1992.

[21] Mary Kathryn Cowles and Bradley P Carlin. Markov chain Monte Carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association*, 91(434):883–904, 1996.

[22] Halim Damerdji. Strong consistency of the variance estimator in steady-state simulation output analysis. *Mathematics of Operations Research*, 19(2):494–512, 1994.

[23] Anand Dixit and Vivekananda Roy. Mcmc diagnostics for higher dimensions using kullback leibler divergence. *Journal of Statistical Computation and Simulation*, 87 (13):2622–2638, 2017.

[24] Randal Douc, Gersende Fort, Eric Moulines, Philippe Soulier, et al. Practical drift conditions for subgeometric rates of convergence. *The Annals of Applied Probability*, 14(3):1353–1377, 2004.

[25] Randal Douc, Eric Moulines, and Jeffrey S Rosenthal. Quantitative bounds on convergence of time-inhomogeneous Markov chains. *Annals of Applied Probability*, pages 1643–1665, 2004.

[26] Daniele Durante. Conjugate Bayes for probit regression via unified skew-normal distributions. *Biometrika*, 106(4):765–779, 2019.

[27] B. Efron, T. Hastie, I. Johnstone, and R Tibshirani. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.

[28] Ray C Fair. A note on the computation of the tobit estimator. *Econometrica: Journal of the Econometric Society*, pages 1723–1727, 1977.

[29] Ray C Fair. A theory of extramarital affairs. *Journal of Political Economy*, 86(1): 45–61, 1978.

[30] George Fishman. *Monte Carlo: concepts, algorithms, and applications*. Springer Science & Business Media, 2013.

[31] James M. Flegal and Galin L. Jones. Batch means and spectral variance estimators in Markov chain Monte Carlo. *The Annals of Statistics*, 38(2):1034–1070, 2010.

[32] Gersende Fort and Eric Moulines. V-subgeometric ergodicity for a Hastings–Metropolis algorithm. *Statistics & probability letters*, 49(4):401–410, 2000.

[33] Gersende Fort and Eric Moulines. Polynomial ergodicity of Markov transition kernels. *Stochastic Processes and their Applications*, 103(1):57–99, 2003.

[34] LLdiko E Frank and Jerome H Friedman. A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135, 1993.

[35] Lesław Gajek, Wojciech Niemiro, and Piotr Pokarowski. Optimal Monte Carlo integration with fixed relative precision. *Journal of Complexity*, 29(1):4–26, 2013.

[36] Christian E Galarza, Tsung-I Lin, Wan-Lun Wang, and Víctor H Lachos. On moments of folded and truncated multivariate student-t distributions based on recurrence relations. *Metrika*, 2021. doi: https://doi.org/10.1007/s00184-020-00802-1.

[37] Alan E Gelfand, Adrian FM Smith, and Tai-Ming Lee. Bayesian analysis of constrained parameter and truncated data problems using Gibbs sampling. *Journal of the American Statistical Association*, 87(418):523–532, 1992.

[38] Andrew Gelman, Donald B Rubin, et al. Inference from iterative simulation using multiple sequences. *Statistical science*, 7(4):457–472, 1992.

[39] Stuart Geman and Donald Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, PAMI(6):721–741, 1984.

[40] Alan Genz. Numerical computation of multivariate normal probabilities. *Journal of computational and graphical statistics*, 1(2):141–149, 1992.

[41] Alan Genz. Numerical computation of rectangular bivariate and trivariate normal and t probabilities. *Statistics and Computing*, 14(3):251–260, 2004.

[42] Alan Genz and Frank Bretz. Numerical computation of multivariate t-probabilities with application to power calculation of multiple contrasts. *Journal of Statistical Computation and Simulation*, 63(4):103–117, 1999.

[43] Alan Genz and Koon-Shlng Kwong. Numerical evaluation of singular multivariate normal distributions. *Journal of Statistical Computation and Simulation*, 68(1): 1–21, 2000.

[44] John Geweke. Efficient simulation from the multivariate normal and student-t distributions subject to linear constraints and the evaluation of constraint probabilities. In *Computing science and statistics: Proceedings of the 23rd symposium on the interface*, volume 571, page 578. Fairfax, Virginia: Interface Foundation of North America, Inc, 1991.

[45] Charles J Geyer. Practical Markov chain Monte Carlo. *Statistical science*, pages 473–483, 1992.

[46] Peter W. Glynn. Some topics in regenerative steady-state simulation. *Acta Applicandae Mathematica*, 34(1-2):225–236, 1994.

[47] Peter W Glynn. Simulation algorithms for regenerative processes. *Handbooks in Operations Research and Management Science*, 13:477–500, 2006.

[48] Peter W Glynn and Donald L Iglehart. Conditions for the applicability of the regenerative method. *Management Science*, 39(9):1108–1111, 1993.

[49] Peter W Glynn and Ward Whitt. Estimating the asymptotic variance with batch means. *Operations Research Letters*, 10(8):431–435, 1991.

[50] Chris Hans. Bayesian lasso regression. *Biometrika*, 96(4):835–845, 2009.

[51] David Harrison Jr and Daniel L Rubinfeld. Hedonic housing prices and the demand for clean air. *Journal of environmental economics and management*, 1978.

[52] Enkelejd Hashorva and Jürg Hüsler. On multivariate gaussian tails. *Annals of the Institute of Statistical Mathematics*, 55(3):507–522, 2003.

[53] W Keith Hastings. Monte Carlo sampling methods using markov chains and their applications. *Biometrika*, 75:97–109, 1970.

[54] Urban Hjorth and Anna Vadeby. Subsample distribution distance and mcmc convergence. *Scandinavian journal of statistics*, 32(2):313–326, 2005.

[55] Hsiu J Ho, Tsung-I Lin, Hsuan-Yu Chen, and Wan-Lun Wang. Some results on the truncated multivariate t distribution. *Journal of Statistical Planning and Inference*, 142(1):25–40, 2012.

[56] James P Hobert and Christian P Robert. A mixture representation of π with applications in markov chain monte carlo and perfect sampling. *The Annals of Applied Probability*, 14(3):1295–1305, 2004.

[57] James P Hobert, Galin L Jones, Brett Presnell, and Jeffrey S Rosenthal. On the applicability of regenerative simulation in Markov chain Monte Carlo. *Biometrika*, 89(4):731–743, 2002.

[58] Yosef Hochberg and Ajit C Tamhane. *Multiple comparison procedures*. John Wiley & Sons, Inc., 1987.

[59] Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometric*, 12(1):55–67, 1970.

[60] Peter J Huber et al. The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 221–233. University of California Press, 1967.

[61] Søren Fiig Jarner and Ernst Hansen. Geometric ergodicity of Metropolis algorithms. *Stochastic processes and their applications*, 85(2):341–361, 2000.

[62] Harold Jeffreys. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 186(1007):453–461, 1946.

[63] Valen E Johnson. A coupling-regeneration scheme for diagnosing convergence in Markov chain Monte Carlo algorithms. *Journal of the American Statistical Association*, 93(441):238–248, 1998.

[64] Galin L Jones and James P Hobert. Honest exploration of intractable probability distributions via Markov chain Monte Carlo. *Statistical Science*, pages 312–334, 2001.

[65] Galin L Jones, Murali Haran, Brian S Caffo, and Ronald Neath. Fixed-width output analysis for Markov chain Monte Carlo. *Journal of the American Statistical Association*, 101(476):1537–1547, 2006.

[66] Galin L Jones et al. On the Markov chain central limit theorem. *Probability surveys*, 1:299–320, 2004.

[67] Kshitij Khare and James P Hobert. Geometric ergodicity of the Gibbs sampler for Bayesian quantile regression. *Journal of Multivariate Analysis*, 112:108–116, 2012.

[68] Kshitij Khare, James P Hobert, et al. Geometric ergodicity of the Bayesian lasso. *Electronic Journal of Statistics*, 7:2150–2163, 2013.

[69] John P Klein and Melvin L Moeschberger. *Survival Analysis: Techniques for Censored and Truncated Data*. Springer Science, 2003.

[70] D. P. Kroese, T. Taimre, and Z. I. Botev. *Handbook of Monte Carlo methods*. John Wiley & Sons, New York, 2011.

[71] Krzysztof Łatuszyński, Błażej Miasojedow, and Wojciech Niemiro. Nonasymptotic bounds on the mean square error for MCMC estimates via renewal techniques. In *Monte Carlo and Quasi-Monte Carlo Methods 2010*, pages 539–555. Springer, 2012.

[72] Krzysztof Łatuszyński, Błażej Miasojedow, Wojciech Niemiro, et al. Nonasymptotic bounds on the estimation error of MCMC algorithms. *Bernoulli*, 19(5A):2033–2066, 2013.

[73] Averill M Law and John S Carson. A sequential procedure for determining the length of a steady-state simulation. *Operations Research*, 27(5):1011–1025, 1979.

[74] Averill M Law and W David Kelton. Confidence intervals for steady-state simulations ii: A survey of sequential procedures. *Management Science*, 28(5):550–562, 1982.

[75] Averill M Law and W David Kelton. Confidence intervals for steady-state simulations: I. a survey of fixed sample size procedures. *Operations Research*, 32(6): 1221–1239, 1984.

[76] Yifang Li and Sujit K Ghosh. Efficient sampling methods for truncated multivariate normal and student-t distributions subject to linear inequality constraints. *Journal of Statistical Theory and Practice*, 9(4):712–732, 2015.

[77] Ying Liu and James M Flegal. Weighted batch means estimators in Markov chain Monte Carlo. *Electronic Journal of Statistics*, 12(2):3397–3442, 2018.

[78] G. Lorden. On excess over the boundary. *The Annals of Mathematical Statistics*, 41(2):520–527, 1970.

[79] Himel Mallick and Nengjun Yi. A new Bayesian lasso. *Statistics and its interface*, 7(4):571–582, 2014.

[80] Ricardo Martinez, Richard Christen, Claude Pasquier, and Nicolas Pasquier. Exploratory analysis of cancer SAGE data. *Proceedings of the Discovery Challenge of the PKDD international conference on Principles of Knowledge Discovery in Databases*, 10 2005.

[81] Kerrie L Mengersen, Richard L Tweedie, et al. Rates of convergence of the Hastings and Metropolis algorithms. *The annals of Statistics*, 24(1):101–121, 1996.

[82] Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092, 1953.

[83] Sean P Meyn and Richard L Tweedie. *Markov chains and stochastic stability*. Springer Science & Business Media, 2012.

[84] John P Mills. Table of the ratio: area to bounding ordinate, for any portion of normal curve. *Biometrika*, pages 395–400, 1926.

[85] Thomas A Mroz. The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions. *Econometrica: Journal of the Econometric Society*, pages 765–799, 1987.

[86] Per Mykland, Luke Tierney, and Bin Yu. Regeneration in Markov chain samplers. *Journal of the American Statistical Association*, 90(429):233–241, 1995.

[87] Esa Nummelin. A splitting technique for Harris recurrent Markov chains. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 43(4):309–318, 1978.

[88] Esa Nummelin and Pekka Tuominen. Geometric ergodicity of harris recurrent marcov chains with applications to renewal theory. *Stochastic Processes and Their Applications*, 12(2):187–202, 1982.

[89] Ari Pakman and Liam Paninski. Exact Hamiltonian Monte Carlo for truncated multivariate Gaussians. *Journal of Computational and Graphical Statistics*, 23(2):518–542, 2014. doi: 10.1080/10618600.2013.788448. URL `https://doi.org/10.1080/10618600.2013.788448`.

[90] Subhadip Pal, Kshitij Khare, et al. Geometric ergodicity for Bayesian shrinkage models. *Electronic Journal of Statistics*, 8(1):604–645, 2014.

[91] Trevor Park and George Casella. The bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008.

[92] N. G. Polson, J. G. Scott, and J. Windle. The Bayesian bridge. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(4):713–733, 2014.

[93] Adrian E. Raftery and Steven Lewis. How many iterations in the Gibbs sampler? In *In Bayesian Statistics 4*, pages 763–773. Oxford University Press, 1992.

[94] Gareth Roberts, Jeffrey Rosenthal, et al. Geometric ergodicity and hybrid Markov chains. *Electronic Communications in Probability*, 2:13–25, 1997.

[95] Gareth O Roberts and Jeffrey S Rosenthal. Harris recurrence of metropolis-within-gibbs and trans-dimensional markov chains. *The Annals of Applied Probability*, pages 2123–2139, 2006.

[96] Gareth O Roberts, Jeffrey S Rosenthal, et al. General state space Markov chains and MCMC algorithms. *Probability surveys*, 1:20–71, 2004.

[97] Jorge Carlos Román and James P Hobert. Geometric ergodicity of Gibbs samplers for Bayesian general linear mixed models with proper priors. *Linear Algebra and its Applications*, 473:54–77, 2015.

[98] Jeffrey S Rosenthal. Minorization conditions and convergence rates for Markov chain Monte Carlo. *Journal of the American Statistical Association*, 90(430):558–566, 1995.

[99] Vivekananda Roy and James P Hobert. Convergence rates and asymptotic standard errors for Markov chain Monte Carlo algorithms for Bayesian probit regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(4):607–623, 2007.

[100] Daniel Rudolf. Explicit error bounds for lazy reversible Markov chain Monte Carlo. *Journal of Complexity*, 25(1):11–24, 2009.

[101] Daniel Rudolf and Nikolaus Schweizer. Error bounds of MCMC for functions with unbounded stationary variance. *Statistics & Probability Letters*, 99:6–12, 2015.

[102] Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6 (2):461–464, 1978.

[103] Hilary L Seal. Studies in the history of probability and statistics. xv: The historical development of the Gauss linear model. *Biometrika*, 54(1-2):1–24, 1967.

[104] Karl Sigman and Ronald W Wolff. A review of regenerative processes. *SIAM review*, 35(2):269–288, 1993.

[105] Leonard A Stefanski and Dennis D Boos. The calculus of M-estimation. *The American Statistician*, 56(1):29–38, 2002.

[106] Henry Teicher et al. On random sums of random vectors. *The Annals of Mathematical Statistics*, 36(5):1450–1458, 1965.

[107] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, 58(1):267–288, 1996.

[108] Robert Tibshirani. Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(3): 273–282, 2011.

[109] Luke Tierney. Markov chains for exploring posterior distributions. *Ann. Statist.*, 22(4):1701–1728, 12 1994. doi: 10.1214/aos/1176325750. URL `https://doi.org/10.1214/aos/1176325750`.

[110] Douglas VanDerwerken and Scott C Schmidler. Monitoring joint convergence of mcmc samplers. *Journal of Computational and Graphical Statistics*, 26(3):558–568, 2017.

[111] Dootika Vats et al. Geometric ergodicity of Gibbs samplers in Bayesian penalized regression models. *Electronic Journal of Statistics*, 11(2):4033–4064, 2017.

[112] Wim PM Vijverberg. Monte Carlo evaluation of multivariate normal probabilities. *Journal of econometrics*, 76(1-2):281–307, 1997.

[113] Abraham Wald. On cumulative sums of random variables. *The Annals of Mathematical Statistics*, 15(3):283–296, 1944.

[114] Abraham Wald. Some generalizations of the theory of cumulative sums of random variables. *The Annals of Mathematical Statistics*, 16(3):287–293, 1945.

[115] Jun Yang and Jeffrey S Rosenthal. Complexity results for MCMC derived from quantitative bounds. *arXiv preprint arXiv:1708.00829*, 2017.