Geometric Ergodicity of a Random-Walk Metorpolis Algorithm via Variable Transformation and Computer Aided Reasoning in Statistics

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Leif Thomas Johnson

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Dr. Charles J. Geyer, Adviser

July 2011

Acknowledgements

I would like to thank Glen Meeden, Galin Jones and Mats Heimdahl for serving on my committee. Galin deserves further thanks for all of the help and encouragement he offered, especially during my second year. Galin answered many questions and put up with many hours of my pestering, and gave me helpful feedback. I would be remiss not to thank Gabriel Chandler. Without a push from Gabe in this direction, I would never have considered studying statistics.

It is possibly impossible to thank Charlie Geyer enough for the help he has given me. Among other things, Charlie is always willing to look at any problem I was working on, and has seemingly infinite patience to help me come to an understanding with it. Charlie read countless drafts, and knew when to threaten to make me grade 5102 homework if I did not improve a proof.

While it is possibly impossible to thank Charlie enough, it is certainly impossible to thank my wife enough. Paloma has supported me through the entire graduate school process, from application to writing this dissertation. She tolerated my countless late nights and knew when to give me encouragement, and when to let me work through alone. I would not have been able to finish without her help.

Abstract

With the steady increase of affordable computing, more and more often analysts are turning to computationally intensive techniques like Markov chain Monte Carlo (MCMC). To properly quantify the quality of their MCMC estimates, analysts need to know how quickly the Markov chain converges. *Geometric ergodicity* is a very useful benchmark for how quickly a Markov chain converges. If a Markov chain is geometrically ergodic, there are easy to use consistent estimators of the Monte Carlo standard errors (MCSEs) for the MCMC estimates, and an easy to use Central Limit Theorem for the MCMC estimates. We provide a method for finding geometrically ergodic Markov chains for classes of target densities. This method uses variable transformation to induce a proxy target distribution, and a random-walk Metropolis algorithm for the proxy target distribution will be geometrically ergodic. Because the transformations we use are one-to-one, Markov chains generated for the proxy target density can be transformed back to a sample from the original target density without loss of inference. Furthermore, because the Markov chain for the proxy distribution is geometrically ergodic, the consistent MCSEs and the CLT apply to the sample on the scale of original target density. We provide example applications of this technique to multinomial logit regression and a multivariate T distribution.

Computer Aided Reasoning (CAR) uses a computer program to assist with mathematical reasoning. Proofs done in a *proof assistant program* are done formally, every step and inference is validated back to the axioms and rules of inference. Thus we have higher confidence in the correctness of a formally verified proof than one done with the traditional paper-and-pencil technique. Computers can track many more details than can be done by hand, so more complicated proofs with more cases and

details can be done in CAR than can be done by hand, and proof assistant programs can help fill in the gaps in a proof. We give a brief overview of the proof assistant program HOL Light, and use it to formally prove the Markov inequality with an expectation based approach.

Contents

List of Tables **vii**

List of Figures **viii**

List of Code Listings **ix**

1 Markov chain Monte Carlo **1**

1.1 Geometric Ergodicity 2

1.2 Random-Walk Metropolis Algorithm 7

1.3 Mean of an Exponential Distribution 9

2 Curvature and Variable Transformation **15**

2.1 Curvature 15

2.2 Variable Transformation 22

2.2.1 Variable Transformation in Gibbs Samplers 23

3 Change-of-Variable and Isotropy **27**

3.1 Multivariate Derivatives and the Chain Rule 27

3.2 Change-of-Variable 29

3.3 Isotropy 30

3.3.1 Multivariate Differentiability 35

3.4 Proof of Lemma 2 35

3.4.1 $\nabla h(\gamma)$ . . . 36
3.4.2 $\det\big(\nabla h(\gamma)\big)$ . . . 40
3.4.3 $\nabla \det\big(\nabla h(\gamma)\big)$ . . . 41

4 Variable Transformation in Random-Walk Metropolis 45
4.1 Exponentially Light to Super-Exponential . . . 45
4.1.1 Choice of f . . . 45
4.1.2 Transformation Theorem . . . 46
4.1.3 Curvature Condition . . . 49
4.2 Sub-Exponential to Exponentially Light . . . 58
4.2.1 Choice of f . . . 58
4.2.2 Transformation Theorem . . . 60
4.2.3 Curvature Conditions . . . 64
4.3 Discussion . . . 67
4.3.1 Implementation . . . 72
4.3.2 Different h . . . 75

5 Example Applications 80
5.1 Multinomial Logit Regression . . . 80
5.1.1 Multinomial Logit Normal Mixed Model . . . 83
5.1.2 Multinomial Logit with Conjugate Prior . . . 85
5.2 Multivariate T Distribution . . . 90

6 Computer Aided Reasoning in Statistics 92
6.1 Introduction . . . 92
6.2 HOL Light: Calculemus . . . 94
6.3 Notation . . . 99
6.4 Previous Work . . . 103

6.5 A Formal Proof of the Markov Inequality 106
6.6 Discussion 113

References **116**

List of Tables

1.1 Summary of independence Metropolis chains for Exp(1). Rows correspond to the different proposal distributions. 12

6.1 Number of the "Top 100 Theorems" formalized by each system (not an exhaustive list of systems) (Wiedijk, 2011). 97

List of Figures

1.1 Exponential densities: solid line Exp(1);dashed line Exp(0.5); dotted line Exp(3). 13

1.2 Histograms of $\bar{f}_{1000}$ for the mean of the Exp(1)distribution: top Exp(0.5) proposal; middle Exp(1) proposal; bottom Exp(3) proposal. 14

2.1 Contours of π from (2.9). 19

2.2 Centered vs non-centered parametrizations for a hierarchical model. . 25

4.1 Solid line: $\beta = h(\gamma)$ for f from(4.1) with $p = 3$. Dashed line: $\beta = \gamma$. . 47

4.2 Solid line: $\beta = h(\gamma)$ for f_1 from(4.20) with $b = 1$. Dashed line: $\beta = \gamma$. 61

4.3 Two different induced densities T_3 density, original modes of 10000 and 0. 71

4.4 Two different induced densities from transforming a standardnormal random variable. 73

6.1 Different tree representations of the vector $(2, 3, 5, 7)^T$ 102

List of Code Listings

4.1 R implementations of f and f_1 using $p = 3$ and $b = 1$. 75

4.2 R implementations of f^{-1} and f_1^{-1} using $p = 3$ and $b = 1$. 76

4.3 R implementations of h and $\log \det \nabla h$ 76

4.4 R implementation of $\log \pi_\gamma$ 77

6.1 Non pretty printed version of the HOL Light theorem , `|- !x y z. x + y + z = y + x + z.` 96

6.2 Vector traversal algorithm. 102

Chapter 1

Markov chain Monte Carlo

Markov chain Monte Carlo (MCMC) (Metropolis et al., 1953; Hastings, 1970; Geman and Geman, 1984; Gelfand and Smith, 1990) is a popular computational technique in statistics. A Markov chain that converges to a distribution of interest is used with Monte Carlo methods to estimate a quantity of interest, e.g. a mean of a posterior distribution in a Bayesian analysis. However, the Markov chain converges in the limit and any simulated Markov chain will have finite length. How quickly the Markov chain converges to a target distribution is important. If a Markov chain converges too slowly, MCMC estimates based on that Markov chain may not be trustworthy. If a Markov chain converges quickly enough ("quickly" is defined in Section 1.1), we can draw reliable inference from the MCMC estimates.

Showing that a Markov chain converges quickly to a target distribution is not simple. Nor is the convergence of a Markov chain something that should be taken for granted. Section 1.3 gives an example of a simple setting where small changes in the Markov chain have large impact on the convergence of the Markov chain.

Our goal is to establish that a Markov chain converges to the target distribution quickly enough that we can draw reliable inference from the MCMC estimates. In Chapter 4 we give a new method of ensuring quick convergence for a random-walk Metropolis algorithm for three classes of target densities on $\mathbb{R}^k$.

1.1 Geometric Ergodicity

A Markov chain is a stochastic process, $X_0, X_1, \ldots$ where the distribution of X_{j+1} is independent of $X_0, \ldots, X_{j-1}$ given the value of X_j. Each X_i takes values in the *state space*, $\mathcal{X}$ and we will use $\mathcal{B}$ to represent the Borel σ-algebra on $\mathcal{X}$. The distribution of X_{j+1} given the value of X_j is defined by the Markov transition kernel, P where $P(b, \cdot)$ is a probability measure and $P(\cdot, A)$ is a measurable function. The kernel, P defines the distribution of X_{j+1} given the value of X_j with the relationship

$$P(b, A) = \Pr(X_j \in A \,|\, X_{j-1} = b). \tag{1.1}$$

We say that π *is invariant for* P if

$$\pi(A) = \int_{\mathcal{X}} \pi(dx) P(x, A).$$

This condition may also be stated as P *preserves* π. A Markov chain is generated by picking X_0, or drawing X_0 from some distribution, ν then using P to draw X_{j+1} based on X_j. If P preserves π, the Markov chain is aperiodic, ϕ-irreducible for some measure ϕ on $(\mathcal{X}, \mathcal{B})$ and Harris recurrent, then the Markov chain is said to be ergodic. The aperiodic ergodic theorem (Meyn and Tweedie, 2009, Theorem 13.3.3) says that for any x

$$\|P^n(x, \cdot) - \pi(\cdot)\| \to 0, \qquad n \to \infty \tag{1.2}$$

where $P^n(x, \cdot)$ is the distribution of X_{j+n} given that $X_j = x$ and $\|\mu(\cdot) - \lambda(\cdot)\|$ is the total variation norm,

$$\|\mu(\cdot) - \lambda(\cdot)\| := \sup_{A \in \mathcal{B}} |\mu(A) - \lambda(A)| - \inf_{A \in \mathcal{B}} |\mu(A) - \lambda(A)|. \tag{1.3}$$

Suppose that f is a function such that $\mathrm{E}_\pi |f(X)| < \infty$, where $\mathrm{E}_\pi g(X)$ is the expectation of the function g with respect to the distribution π. If the Markov chain is ergodic there is a Strong Law of Large Numbers (SLLN) (Meyn and Tweedie, 2009, Theorem 17.1.7)

$$\bar{f}_n := \frac{1}{n}\sum_{i=1}^{n} f(X_i) \to \mathrm{E}_\pi f(X) \qquad \text{with probability 1 as } n \to \infty. \tag{1.4}$$

A Markov chain being aperiodic, ϕ-irreducible and Harris recurrent is quite a mouthful. Simply put, this means that: the Markov chain does not travel through a fixed cycle of sets; for any set A such that $\pi(A) > 0$, the Markov chain *can* visit A; and for any set A such that $\pi(A) > 0$, the the Markov chain *will* visit A infinitely often. For a more complete discussion of these conditions and the associated ergodicity results, see one of the sources Meyn and Tweedie (2009); Jones and Hobert (2001); Tierney (1994) or Geyer (1992).

For many probability distributions there are off-the-shelf ergodic samplers, such as random-walk Metropolis algorithms (Tierney, 1994) or Gibbs samplers[1] (Geman and Geman, 1984). However, ergodicity on its own is not strong enough to draw reliable inference about π. If $\bar{f}_n$ is an estimate of $\mathrm{E}_\pi f(X)$ based on an ergodic Markov chain with stationary distribution π, the SLLN (1.4) states that $\bar{f}_n$ converges to $\mathrm{E}_\pi f(X)$ with probability one[2]. But this convergence happens as n, the length of the Markov chain goes to infinity. In practice, we will always be drawing inference based on a finite length Markov chain, where the n in $\bar{f}_n$ has not gone to infinity. To draw valid

[1] This is not to say that these samplers are always ergodic. A random-walk Metropolis algorithm is not ergodic if the state space is not connected and the proposal distribution is short range; A Gibbs sampler for the state space $\left\{(x,y) \in \mathbb{R}^2 \middle| [x \le 1 \wedge y \le 1] \vee [x \ge 2 \wedge y \ge 2]\right\}$ is not irreducible and not ergodic.

[2] Convergence in total variation is not necessary for the convergence of $\bar{f}_n$. Let X be uniformly distributed on the set $\{0, \ldots, k-1\}$ for integer $k > 1$. Set $X_0 = 0$ and the transition kernel be defined by $P(x, \{y\}) = 1$ if $y = (x+1) \mod k$ and 0 otherwise. Hence $\bar{f}_n = n^{-1}\sum_{i=0}^{n} X_i \to \mathrm{E}(X)$, but the total variation norm does not go to zero.

inference about $\mathrm{E}_\pi f(X)$ we must be able to make statements about the distribution of $\bar{f}_n$ for finite n. For finite n, the distribution of $\bar{f}_n$ depends on "how quickly" the Markov chain converges to the invariant distribution.

We have referred several times to "how quickly" a Markov chain converges to the target, or invariant distribution. For a Markov chain with transition probability measure, P and invariant probability measure, π "how quickly" refers to how fast the total variation distance,

$$\|P^n(x,\cdot) - \pi(\cdot)\|$$

goes to zero. A Markov chain is *geometrically ergodic* if there exists a non-negative function M and constant $r < 1$ such that

$$\|P^n(x,\cdot) - \pi(\cdot)\| \leq M(x)r^n, \qquad x \in \mathcal{X}. \tag{1.5}$$

If M is bounded, then the Markov chain is *uniformly ergodic*.

Verifying (1.5) directly is too difficult to do directly. Fortunately, verifying (1.5) can be done by finding a drift condition and associated minorization condition.

If there exist some probability measure Q on $\mathcal{B}$ there set $C \in \mathcal{B}$, positive integer n_0 and an $\epsilon > 0$ such that

$$P^{n_0}(x, A) \geq \epsilon Q(A) \text{ for all } x \in C, A \in \mathcal{B} \tag{1.6}$$

then a *minorization condition* holds on the set C. In this case, C is called a *small set*.

We shall use $\mathbf{1}(\cdot)$ as the indicator function that maps true values to 1 and false values to 0. If there exists some function $V : \mathcal{X} \to [1, \infty)$, constants $0 < \gamma < 1$, and

$k < \infty$ such that for all $x \in \mathcal{X}$

$$\mathrm{E}\big[V(X_{i+1})|X_i = x\big] < \gamma V(x) + k\mathbf{1}(x \in C) \tag{1.7}$$

then a *drift condition* holds.

Geometric ergodicity of an aperiodic, ϕ-irreducible Markov chain is equivalent to the drift condition in (1.7) holding with an associated minorization condition (1.6) (Roberts and Rosenthal, 2004).

An alternative drift condition holds if there exist a non-negative function $W : \mathcal{X} \to \mathbb{R}^+$ that is unbounded off petite sets[3], and constants $0 < \lambda < 1$ and $b < \infty$ such that

$$\mathrm{E}\big[W(X_{i+1})|X_i = x\big] < \lambda W(x) + b. \tag{1.8}$$

It may be easier to verify (1.8) instead of (1.7). The two drift conditions in (1.8) and (1.7) are equivalent (Johnson et al., 2009; Jones and Hobert, 2004), so only one of the two needs to verified.

If a Markov chain is geometrically ergodic we can attempt to answer two very important questions:

(Q1) Has the Markov chain converged to the stationary distribution?

(Q2) Has a Markov chain been run long enough to get accurate MCMC estimates?

So if a Markov chain is geometrically ergodic, it converges "quickly enough" to draw reliable inference.

[3]Knowing the definition of a petite set will not aide this discussion. If it is appropriate to talk of $\|x\|$ going to ∞, as in the case of $\mathcal{X} = \mathbb{R}^k$, then it is sufficient that $W(x) \to \infty$ as $\|x\| \to \infty$. Those wishing to develop an understanding of petite sets should consult Meyn and Tweedie (2009, Chapters 5,6,8 and 9)

For a geometrically ergodic Markov chain, M and t in (1.5) provide an upper bound on the total variation distance between the Markov chain and the invariant distribution π. However M and t are typically unknown. The drift and minorization conditions used to establish geometric ergodicity can be used to provide an upper bound for this total variation distance (Rosenthal, 1995; Hobert and Robert, 2004) that can be used to answer (Q1).

For a function f such that $\mathrm{E}_\pi |f(X)|^{2+\delta} < \infty$ and for a geometrically ergodic Markov chain, in addition to the SLLN (1.4) there is a Central Limit Theorem (CLT) for $\bar{f}_n$,

$$\sqrt{n}\big(\bar{f}_n - \mathrm{E}_\pi \bar{f}\big) \xrightarrow{d} N(0, \sigma_f^2) \tag{1.9}$$

where $\sigma_f^2 = \mathrm{Var}_\pi\big(f(X_0)\big) + 2\sum_{i=1}^{\infty} \mathrm{Cov}_\pi\big(f(X_1), f(X_i)\big)$, which is finite[4][5] (Chan and Geyer, 2004). Geometric ergodicity of the Markov chain is not necessary for there to be a CLT for $\bar{f}_n$. Jarner and Roberts (2007) and Jarner and Roberts (2002) demonstrate a CLT for $\bar{f}_n$ based on a polynomially ergodic Markov chain, but that CLT is more complicated and harder to use.

On its own, this CLT does not assess the accuracy of the MCMC estimates, we also need an estimate of σ_f^2. Because Markov chains are almost always auto-correlated, using the naive estimator, $\hat{\sigma}_f^2 = \frac{1}{n-1}\sum_i \big(f(X_i) - \bar{f}_n\big)^2$ will not be a valid estimate of σ_f^2. Fortunately, there are relatively simple methods for finding a consistent $\hat{\sigma}_f^2$ when

[4]Chan and Geyer (2004) show that a geometrically ergodic Markov chain satisifies the conditions of Theorem 18.5.4 in Ibragimov and Linnik (1971), hence σ_f^2 is finite if $\mathrm{E}_\pi |f(X)|^{2+\delta}$ exists for some $\delta > 0$.

[5]A Markov chain is *reversible* with respect to π if

$$\int\int f(x,y)\pi(dx)P(x,dy) = \int\int f(y,x)\pi(dx)P(x,dy)$$

for any bounded measurable function f. If a geometrically ergodic Markov chain is *reverisble*, then the existence of a CLT for $\bar{f}_n$ only requires a second moment for f, i.e. $\mathrm{E}_\pi |f(X)|^2 < \infty$. The random-walk Metropolis algorithms is reversible, so when such a Markov chain is geometrically ergodic only a second moment is necessary for a CLT for $\bar{f}_n$.

using a geometrically ergodic Markov chain, such as overlapping batch means (Flegal and Jones, 2010) or consistent batch means (Jones et al., 2006). Once we have an estimate of σ_f^2, we estimate the Monte Carlo standard error (MCSE) of $\bar{f}_n$, $\hat{\sigma}_f/n$. Reporting MCSEs allows others to judge the accuracy of our estimates. A general overview of how to "estimate with confidence", or calculate and use a consistent $\hat{\sigma}_f^2$ to answer (Q2) in MCMC is available from Flegal and Jones (2011).

In Chapter 4, we advance a new technique for finding a geometrically ergodic Markov chain for a target distribution. This result does not directly involve finding drift or minorization conditions, it is based on the tail behavior of the target density, putting it in the "curvature family". This family (Mengersen and Tweedie, 1996; Roberts and Tweedie, 1996; Jarner and Hansen, 2000), discussed in Section 2.1, works by finding conditions that guarantee the existence of drift and minorization conditions, but the users of the theorem do not need to find the drift or minorization conditions themselves.

1.2 Random-Walk Metropolis Algorithm

For densities on a proper or improper subset $\mathbb{R}^k$, the random-walk Metropolis algorithm (Tierney, 1994) is one method of creating P, the transition probability measure. The random-walk Metropolis algorithm is a special case of the Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970).

The goal of the Metropolis-Hastings algorithm is to generate a Markov chain for h, the target density. Let h be a unnormalized density with respect to some positive measure μ, and for each x in the state space the proposal density, $q(x, \cdot)$ be a proper density with respect to μ. This q should be chosen such that it is easy to simulate draws from, e.g. the multivariate normal density. Let X_n be the current state of a Markov chain. The Metropolis-Hastings algorithm finds state X_{n+1} with the following

steps.

(i) Draw a *proposal*, Y_n from the distribution given by the density $q(X_n, \cdot)$.

(ii) Calculate the *acceptance probability*, $a(X_n, Y_n)$ where

$$r(x, y) = \frac{h(y)q(y, x)}{h(x)q(x, y)} \tag{1.10}$$

and

$$a(x, y) = \min\big(1, r(x, y)\big). \tag{1.11}$$

(iii) *Accept* the proposal with probability $a(X_n, Y_n)$ and *reject* the proposal with probability $1 - a(X_n, Y_n)$ — i.e. set $X_{n+1} = Y_n$ with probability $a(X_n, Y_n)$ and set $X_{n+1} = X_n$ with probability $1 - a(X_n, Y_n)$.

The *acceptance rate* refers to the proportion of proposed moves to Y_n that are accepted.

As long as $h(X_0) > 0$, then $h(X_i) > 0$ for all i with probability one. This can be shown by supposing that $h(X_n) > 0$ and $h(Y_n) = 0$. Then $r(X_n, Y_n) = 0$ as long as $h(X_n)q(X_n, Y_n) > 0$. Since Y_n was drawn from $q(X_n, \cdot)$, $h(X_n)q(X_n, Y_n) = 0$ with probability zero. Given that $r(X_n, Y_n) = 0$, the move to Y_n will be rejected with probability one. Hence with probability one, a chain started at X_0 such that $h(X_0) > 0$ will not move to a state Y_i with $h(Y_i) = 0$. Consequently, there is no risk with using a q that can propose moves outside of the support of h. Such moves will be rejected with probability one and the Markov chain will never leave the support of h.

The *Metropolis algorithm* is the Metropolis-Hastings algorithm with the restriction that q is symmetric, i.e. $q(x, y) = q(y, x)$. Hence the Metorpolis algorithm is a special

case of the Metropolis-Hastings algorithm. A *random-walk Metropolis algorithm* has the requirement that $q(x, y) = q(y - x) = q(x - y)$, the proposal density only depends on the difference between x and y.

For the Metropolis algorithm the acceptance probability simplifies to

$$a(x, y) = \min\left(1, \frac{h(y)}{h(x)}\right).$$

In a Metropolis algorithm, moves to a state with a higher density are *always* accepted, if Y_n is proposed and $h(Y_n) > h(X_n)$, then $a(X_n, Y_n) = 1$ and the move is accepted with probability one. Moves to a state with a lower density are *sometimes* accepted, if Y_n is proposed and $0 < h(Y_n) < h(X_n)$, the move is accepted with probability $h(Y_n)/h(X_n) < 1$. When X_n is in the tails of the target distribution, the value of this "sometimes" is directly related to how quickly h goes to zero. The implications of this behavior with regards to the geometric ergodicity of a random-walk Metropolis algorithm are discussed in Section 2.1.

A Markov chain generated using a random-walk Metropolis algorithm will be ergodic if either $q(x - y) > 0$ for all x and y, or if the set $\{x \mid h(x) > 0\}$ is open and connected and q is bounded away from zero in a neighborhood of zero(Tierney, 1994).

1.3 Mean of an Exponential Distribution

This section provides both a cautionary tale, and a motivating example. Through simulation, we will demonstrate Markov chains that even in the simple setting of an Exponential distribution, small changes in the Markov kernel can have large impact in the reliability of the MCMC estimates. Furthermore, we will demonstrate that a Markov kernel "seeming reasonable" is not a guarantee that the corresponding Markov chain will converge quickly enough to provide valid inference.

Let $X \sim \text{Exp}(\lambda)$, i.e. X follows a exponential distribution with rate parameter $\lambda > 0$. Even though we already that $\text{E}(X) = \lambda^{-1}$, we will estimate the mean of an Exp(1) distribution using MCMC and three different Markov chains.

The independence Metropolis algorithm is the Metropolis-Hastings algorithm with a proposal density that does not depend on the current state, $q(x, y) = q(y)$. The acceptance probability can be rewritten as

$$a(x, y) = \min\left(1, \frac{h(y)/q(y)}{h(x)/q(x)}\right).$$

We will generate Markov chains with the independence Metropolis algorithm using proposal densities corresponding to Exp(0.5), Exp(1) and Exp(3) distributions. These three densities are shown in Figure 1.1, the Exp(0.5) distribution has a heavier tail than the Exp(1) distribution, and the Exp(3) distribution has a much lighter tail than the Exp(1) distribution.

Using Exp(1) as the proposal distribution, $h(x) = q(x)$ for all x, so the acceptance probability is one and the Markov chain is an iid sample from the Exp(1) distribution. Because $h(x)/q(x)$ is bounded for all x, the Markov chain generated using Exp(0.5) as the proposal distribution will be uniformly (and hence geometrically) ergodic (Tierney, 1994, Corollary 4). The Markov chain generated using Exp(3) as the proposal density is not geometrically ergodic (Mengersen and Tweedie, 1996, Theorem 2.1). We know that $\text{E}(X^3)$ is finite for an exponential distribution. Thus a CLT holds for the MCMC estimates of the mean generated with the Exp(0.5) and Exp(1) proposals. A CLT does not hold for the MCMC estimates of the mean generated with the Exp(3) proposal (Roberts, 1999, Example 1), though a SLLN does since the Markov chain is ergodic.

For each proposal distribution, we ran 1000 Markov chains using the independence Metropolis algorithm and calculated the sample mean for each Markov chain. Each of

these Markov chains was 1000 iterations long, and had initial state $X_0 = 1$. Table 1.1 summarizes the simulations. Histograms of the $\bar{f}_{1000} = \frac{1}{1001}\sum_{i=0}^{1000} X_i$ are shown in Figure 1.2. The histograms for the Exp(0.5) and Exp(1) proposals look like normal histograms, centered at 1 with small standard deviations. The histogram for the Exp(3) proposal looks far from normal, with a spread that is enormous compared to the other two. Using a Kolmogorov-Smirnov test for normality supports this conclusion, giving p-values of 0.99, 0.70 and approximately 0 for the distribution of the standardized sample means from the Markov chains with the Exp(0.5), Exp(1) and Exp(3) proposal distributions respectively.

For each Markov chain we also calculated a 95% confidence interval for the mean. The form of these confidence intervals is

$$\bar{f}_n \pm t_* \frac{\hat{\sigma}_f}{\sqrt{n}}$$

where $\hat{\sigma}_f$ is calculated using the overlapping batch means method from Flegal and Jones (2010). I.e.

$$\hat{\sigma}^2_{OBM} = \frac{nb_n}{(n-b_n)(n-b_n+1)} \sum_{j=0}^{n-b_n} \left(\bar{X}_j(b_n) - \bar{X}_n\right)^2$$

where b_n is the batch size, $\bar{X}_n$ is the mean of the Markov chain and $\bar{X}_j(b_n) = b_n^{-1}\sum_{i=1}^{b_n} X_{j+i}$. The batch size was taken to be $b_n = \lfloor n^{1/2} \rfloor = 31$. For the Markov chains using the Exp(3) proposal, this does not produce asymptotically valid estimates of σ^2. We applied the OBM method and calculated the confidence intervals anyway. The observed coverage percentages are given in the "Coverage % (SE)" column of Table 1.1, the SEs of the observed coverage percentages are given in parenthesis. For the Markov chains using Exp(0.5) and Exp(1) proposals, the observed coverage was 93.4% and 95.7% respectively. These are close to the desired 95%, which is the

Table 1.1: Summary of independence Metropolis chains for Exp(1). Rows correspond to the different proposal distributions.

	Grand Mean (SE)	Mean MCSE (SE)	Coverage % (SE)	K-S p-value
Exp(0.5)	1.000 (0.046)	0.045 (0.005)	93.4 (0.785)	0.99
Exp(1)	1.000 (0.030)	0.031 (0.003)	95.7 (0.641)	0.70
Exp(3)	0.889 (0.297)	0.093 (0.044)	40.0 (1.549)	0.00

asymptotic coverage. Being this close to the specified CI level for such a short Markov chain (only 1000 steps) is a good sign. The Markov chain with the Exp(3) proposal has an observed coverage of 40.0%. This is not surprising because the Markov chain converges too slowly for the asymptotic results (Flegal and Jones, 2010). In Table 1.1 the column "Grand mean (SE)" gives the average and standard error of the $\bar{f}_n$ for each proposal distribution. The column "Mean MCSE (SE)" gives the average and standard error of $\hat{\sigma}_f^2$ for each proposal distribution. We can see that for the Markov chains with the Exp(0.5) and Exp(1) the mean estimated MCSEs (0.045 and 0.031, respectively) are close to the observed SEs (0.046 and 0.30, respectively).

In practice, inference will be drawn from only one Markov chain, and we won't know the true value being estimated. We would much rather run a Markov chain that will produce an estimate from a normal distribution like the top two plots in Figure 1.2 with asymptotically valid confidence intervals, and avoid using Markov chains that will produce an estimate from some other distribution without asymptotically valid confidence intervals, like in the bottom plot.

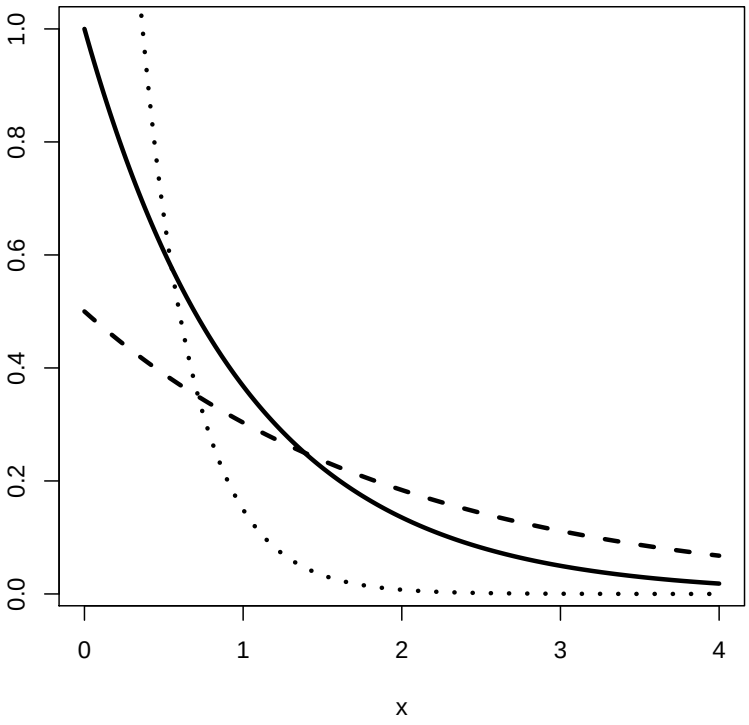


Figure 1.1: Exponential densities: solid line Exp(1);dashed line Exp(0.5); dotted line Exp(3).

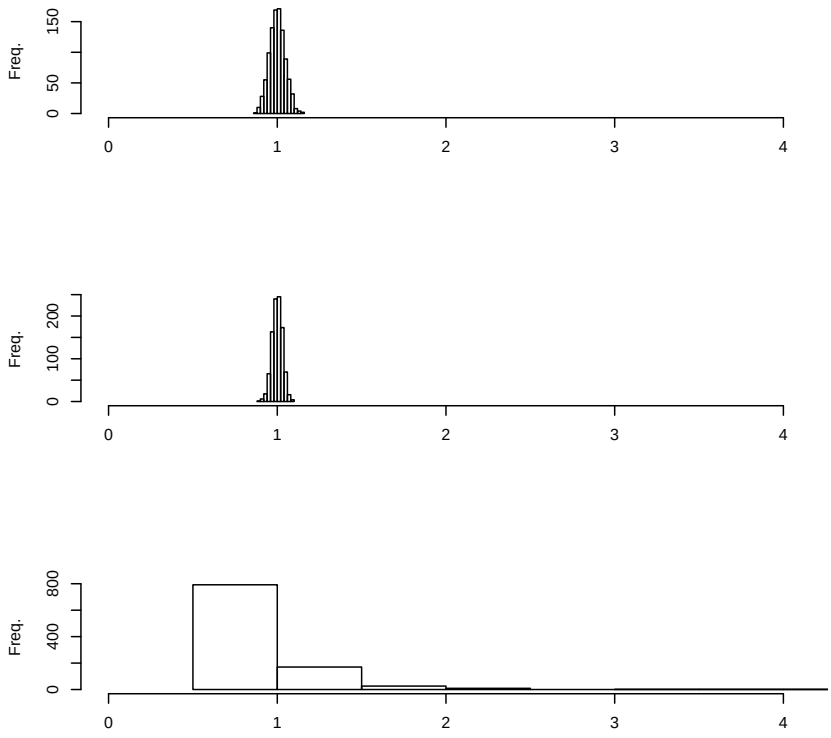


Figure 1.2: Histograms of $\bar{f}_{1000}$ for the mean of the Exp(1)distribution: top Exp(0.5) proposal; middle Exp(1) proposal; bottom Exp(3) proposal.

Chapter 2

Curvature and Variable Transformation

2.1 Curvature

Recall from (1.11) that in a random-walk Metropolis algorithm for a target density π, the acceptance probability for a proposed step from β_i to β' is

$$a(\beta_i, \beta') = \min\left(1, \frac{\pi(\beta')}{\pi(\beta_i)}\right).$$

Proposed steps "uphill", towards the mode of the distribution are *always* accepted. Proposed steps "downhill", into the tails of the distribution are *sometimes* accepted. If the tails of the target density are "light enough", the value of *sometimes* will decay quickly as the state moves into the tails of the distribution. Quick decay will stop the Markov chain from spending too much time in the tails of the target density and will imply a drift condition, and hence that the Markov chain is geometrically ergodic.

There are three important papers with results in this area, Mengersen and Tweedie (1996), Roberts and Tweedie (1996) and Jarner and Hansen (2000). Our results are based on the work by Jarner and Hansen (2000).

There are two important definitions, developed across these papers and summa-

rized in Jarner and Hansen (2000).

A density, π on $\mathbb{R}^k$ is *exponentially-light* if π is always positive, has continuous first derivatives and

$$\limsup_{\|x\|\to\infty} \frac{x}{\|x\|} \cdot \nabla \log \pi(x) < 0. \tag{2.1}$$

A density, π on $\mathbb{R}^k$ is *super-exponential* if π is always positive, has continuous first derivatives and

$$\limsup_{\|x\|\to\infty} \frac{x}{\|x\|} \cdot \nabla \log \pi(x) = -\infty. \tag{2.2}$$

Jarner and Hansen (2000) did not require a density to be always positive or have continuous first derivatives to be exponentially light. Their usage of the term was based only on the tail-weight of the density, in fact their term was "a density has exponentially light tails". We have altered this definition to make discussion of exponentially light versus super-exponential densities simpler to read. There is no theoretical impact from this change.

We are adding the definition *sub-exponential*, placed here to be near the definitions of exponentially-light and super-exponential.

A density, π on $\mathbb{R}^k$ is *sub-exponential* if π is always positive, has continuous first derivatives and

$$\limsup_{\|x\|\to\infty} \frac{x}{\|x\|} \cdot \nabla \log \pi(x) \leq 0. \tag{2.3}$$

Note that the lim sups in (2.2)–(2.3) do not need to be non-positive, the lim sup could be positive or even ∞ even when there is an exponential moment. Consider

the density

$$\pi(x) \propto e^{-|x|/10}(a + cos(x)), \qquad a \geq 1. \tag{2.4}$$

For any $a \geq 1$, this is a valid density with an exponential moment, that is

$$\int e^{s|x|}\pi(x)dx$$

exists for some $s > 0$. Even though this exponential moment exists the relevant lim sup (as in (2.3)) is,

$$\limsup_{\|x\|\to\infty} -\frac{x}{\|x\|} \cdot \nabla \log \pi(x) = \begin{cases} \inf & \text{if } a = 1 \\ \frac{\sqrt{1-a^{-2}}}{a-a^{-1}} - \frac{1}{10} & \text{if } a > 1 \end{cases}$$

which is positive for $a > \sqrt{101}$. Densities with positive or infinite lim sups (for the lim sup in (2.3)), such as π from (2.4), have non-monotonic tails. The tails of these densities go to zero (which is necessary to be integrable) but oscillate infinitely often. The theorems we give in Chapter 4 apply to densities that are *at least* sub-exponential, hence densities with non-monotonic tails do not receive further comment.

Exponentially-light and super-exponential are two important reference points for establishing convergence rates of random-walk Metropolis algorithms. e Mengersen and Tweedie (1996) established that for a symmetric density, π on $\mathbb{R}$ where

$$\lim_{x\to\infty} (d/dx) \log \pi(x)$$

exists, being exponentially-light is a necessary and sufficient condition for the geometric ergodicity of a random-walk Metropolis algorithm. The extension to non-symmetric densities is straightforward (Roberts and Tweedie, 1996; Jarner and Hansen,

2000).

A necessary condition for the geometric ergodicity of a random-walk Metropolis algorithm on $\mathbb{R}^k$ is the existence of an exponential moment (Jarner and Tweedie, 2003), i.e. for some $s > 0$

$$\int_{\mathbb{R}^k} e^{s\|x\|}\pi(x)dx < \infty.$$

Roberts and Tweedie (1996) provide a set of conditions on $\mathbb{R}^k$ that will imply the geometric ergodicity of a random-walk Metropolis algorithm. For a target density, π define $A(x) = \{y|\pi(y)/\pi(x) \geq 1\}$, $R(x) = A(x)^c$ and $R^r(x) = \{y|2x - y \in R(x)\}$ (the set $R(x)$ rotated 180° around x) and use Δ as the symmetric difference operator for two sets[1]. Then sufficient conditions for the geometric ergodicity of a random-walk Metropolis algorithm for π using proposal density $q(x, y) = q(x-y)$ with corresponding probability measure $Q(Z) = \int_Z q(z)dz$ are

$$\int \pi(x)^{\frac{1}{2}}dx < \infty, \tag{2.5}$$

$$\lim_{\|x\|\to\infty} Q[\{R^r(x) - x\}\Delta\{A(x) - x\}] \to 0, \tag{2.6}$$

$$\liminf_{\|x\|\to\infty} \int \left[1 - \left\{\frac{\pi(x+z)^{\frac{1}{2}}}{\pi(x)^{\frac{1}{2}}} \wedge 1\right\}^2\right] q(z)dz > 0, \tag{2.7}$$

$$\limsup_{\|x\|\to\infty} \int_{R^r(x)-x} \left[\left\{1 \wedge \frac{\pi(x)^{\frac{1}{2}}}{\pi(x+z)^{\frac{1}{2}}}\right\} - \frac{\pi(x-z)^{\frac{1}{2}}}{\pi(x)^{\frac{1}{2}}}\right] q(z)dz \leq 0. \tag{2.8}$$

Roberts and Tweedie (1996) applied these results to densities of the form

$$\pi(x) = h(x)e^{-p(x)},$$

where h and p are polynomials on $\mathbb{R}^k$, such that h is positive everywhere and p is of

[1]The symmetric difference of two sets, A and B is the set $(A \setminus B) \cup (B \setminus A)$.

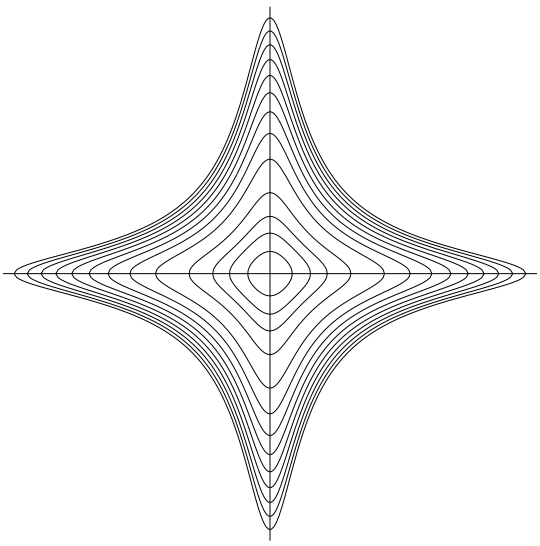


Figure 2.1: Contours of π from (2.9).

even order $d \geq 2$ satisfying

$$\lim_{\|x\| \to \infty} p_d(x) = \infty$$

where p_d is the d^{th} order term of p, showing that a random-walk Metropolis algorithm for such densities will be geometrically ergodic. However, Jarner and Hansen (2000) show that this is not necessarily true. If h is not a positive-definite polynomial, the random-walk Metropolis algorithm for π need not be geometrically ergodic.

An interesting example is the density

$$\pi(x, y) \propto e^{-(x^2 + x^2 y^2 + y^2)}. \tag{2.9}$$

As can be seen in Figure 2.1, the contour curves of π degenerate into spikes. Jarner and Hansen (2000) use this degeneration to show that a random-walk Metropolis algorithm for π in (2.9) is not geometrically ergodic.

Jarner and Hansen (2000) greatly extend and simplify the work of Roberts and

Tweedie (1996). They demonstrate that if a target density π is super-exponential, then there is a simple sufficient condition for the geometric ergodicity of a random-walk Metropolis algorithm for π, if the proposal density is identical in each coordinate, symmetric and bounded away from zero in a neighborhood around the origin.

In mathematical notation, these conditions for a density q are

$$q(x, y) = q(\|x - y\|) \qquad x, y \in \mathbb{R}^k \tag{2.10}$$

and there is some $\delta_q > 0$ and $\varepsilon_q > 0$ such that

$$\|x\| < \delta_q \Rightarrow q(x) > \varepsilon_q. \tag{2.11}$$

Theorem 1 (Jarner and Hansen Theorem 4.3)**.** *If π is super-exponential and satisfies*

$$\limsup_{\|x\| \to \infty} \frac{x}{\|x\|} \cdot \frac{\nabla \pi(x)}{\|\nabla \pi(x)\|} < 0. \tag{2.12}$$

then the random-walk Metropolis algorithm with q satisfying (2.11) *is geometrically ergodic.*

This condition is much simpler to use than the ones from Roberts and Tweedie (1996). However, like the results in Roberts and Tweedie (1996), the largest usage of these results in general discussion of geometric ergodicity rather than developing new algorithms or showing geometric ergodicity of a random-walk Metropolis algorithm for a target density of interest.

In several discussions of ergodicity of adaptive random-walk Metropolis algorithms, the proof of ergodicity uses some variation of: *Assume that the target density π is super-exponential, satisfies condition* (2.12) *and ...* (Atchadé and Rosenthal, 2005; Andrieu and Éric Moulines, 2006; Bai et al., 2010). That is, they assume that π is a density that satisfies the conditions of Theorem 1 so a random-walk Metropolis

algorithm with appropriate proposal density would already be geometrically ergodic. One such adaptive MCMC algorithm uses a proposal distribution $N(\cdot, \sigma^2 I_k)$, except σ^2 is automatically adjusted using the history of the chain to obtain a desirable asymptotic acceptance rate.

Examples from Jarner and Hansen (2000) are frequently used in other papers. Fort et al. (2003) show that the random-scan one-variable-at-a-time random-walk Metropolis algorithm is geometrically ergodic for some densities where the random-walk Metropolis algorithm is not. One specific example is the density (2.9).

There are very few places where the results from Jarner and Hansen (2000) have been used to show the geometric ergodicity of a random-walk Metropolis algorithm for a target density of interest. Both Johnson et al. (2009) and Christensen et al. (2001) apply Theorem 1 to the density for a logit-normal mixed model. The other is this thesis, in Section 5.1.1 Theorem 1 is applied to the density for a multinomial-logit-normal mixed model. Essentially a multivariate generalization of the logit-normal model. Christensen et al. (2001) apply results from Jarner and Hansen (2000) (but not Theorem 1) to Poisson-log-normal and exponential-log-normal models.

When multiplying densities (as in a mixed model) the super-exponential nature of a normal density will dominate the other density and tend to make an application of Theorem 1 straightforward. For any but the most pathological data models, $f(y|\mu)$,

$$\limsup_{\|\mu\|\to\infty} \frac{\mu}{\|\mu\|} \cdot \nabla \log f(y|\mu) = c < \infty.$$

If the prior, $f(\mu)$, is normal the posterior distribution will satisfy condition (2.2)

$$\begin{aligned}\limsup_{\|\mu\|\to\infty} \frac{\mu}{\|\mu\|}\nabla \log f(\mu|y) \leq & \\ \limsup_{\|\mu\|\to\infty} \frac{\mu}{\|\mu\|} \cdot \nabla \log f(y|\mu) + \limsup_{\|\mu\|\to\infty} \frac{\mu}{\|\mu\|} \cdot \nabla \log f(\mu) & \\ = -\infty &\end{aligned}$$

and therefore be super-exponential if $f(y|\mu)$ is always positive with continuous first derivatives. The curvature condition (2.12) will not automatically be satisfied and a different argument will be needed. Such as the Poisson-log-normal and exponential-log-normal models from Christensen et al. (2001).

2.2 Variable Transformation

Let Y be a random variable with distribution function F_Y, and h be a one-to-one differentiable function defined on $\mathcal{Y}$, the support of Y. Then h may be used to transform Y to a new variable, $X = h(Y)$. We can evaluate expectations with respect to X by evaluating expectations with respect to Y. For any measurable function f,

$$\mathrm{E}\big(f(X)\big) = \mathrm{E}\big(f(h(Y))\big).$$

This relationship is frequently used without acknowledgment, hence it is sometimes called *the law of the unconscious statistician* (Ross, 1988, Proposition 2.1), others (Casella and Berger, 2002, comments on Definition 2.2.1) "do not find this amusing". It's also true that

$$\mathrm{P}\big(X \in A\big) = \mathrm{P}\big(Y \in h^{-1}(A)\big).$$

This equivalence of probabilities gives a direct correspondence of samples from F_X to samples from F_Y.

The law of the unconscious statistician is the basis of the new MCMC results we are putting forward. Suppose we have a (possibly posterior) distribution, π_X we want to draw inference on. Furthermore, suppose π_X is analytically intractable; e.g. there is an unknown normalizing constant or high dimensionality. We can use MCMC techniques to draw inference on π_X. But not every Markov Chain sampling method works well for every distribution. In fact, some sampling methods perform well in some settings, but very poorly in others (Papaspiliopoulos and Roberts, 2007; Roberts and Sahu, 1997). For a one-to-one and differentiable h, setting $X = h(Y)$ induces a proxy distribution, π_Y that by the law of the unconscious statistician is sufficient for drawing inference about π_X. Applying h will transform a sample from π_Y to a sample from π_X. If h is well chosen, we can run a Markov chain for π_Y that performs well, even though a Markov chain for π_X does not.

2.2.1 Variable Transformation in Gibbs Samplers

Here we give a brief summary of some current work along these lines. Brief because it is not intended to be a complete review, just give a taste of the main results and demonstrate that our work is complementary.

Roberts and Sahu (1997) demonstrate that the correlation structure of a Gaussian distribution affects the convergence of a Gibbs sampler. They use simple variable transformations — permutations of vector order or location shifts of a vector — to change the distribution's correlation structure. Roberts and Sahu (1997) demonstrate that in Gaussian linear hierarchical models, these simple variable transformations can significantly improve the performance of a Gibbs sampler when combined with appropriate variable blocking and a correctly chosen update scheme.

In hierarchical models the data, Y depend on the parameters, Θ through un-

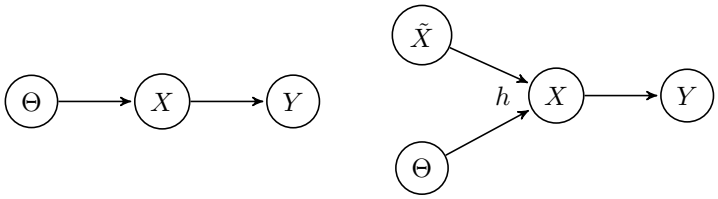
observed variables, X. MCMC is frequently used to draw samples about the posterior distribution, $\mathrm{P}(X, \Theta|Y)$. Papaspiliopoulos et al. (2007) summarize results, demonstrating that the performance of a Gibbs sampler can be very different for *centered* and *non-centered* parametrization. In the centered parametrization, Y depends on Θ through X. In the non-centered parametrization, X depends on $\widetilde{X}$ and Θ through the transformation h. This new variable $\widetilde{X}$ is *a priori* independent of Θ. The MCMC sampler is run on the posterior distribution, $\mathrm{P}(\widetilde{X}, \Theta|Y)$. Samples from this distribution can be easily transformed to samples from the posterior distribution $\mathrm{P}(X, \Theta|Y)$. Figure 2.2 illustrates the centered vs. non-centered parametrization. Papaspiliopoulos et al. (2007) show that the relative performance of a Gibbs sampler on a centered parametrization vs. a non-centered parametrization is not cut and dried. Depending on such factors as the size of the data relative to the size of the imputed variables. They give general guidelines for selecting which parametrization to use, and suggest some possible transformations, depending on the relationship between X and Θ. Taken directly from Papaspiliopoulos et al. (2007), these transformations are

- Location: if $X \sim F(\cdot - \Theta)$,
  then $h(\widetilde{X}, \theta) = \widetilde{X} + \theta$, $\widetilde{X} \sim F(\cdot)$.

- Scale: if $X \sim F(\cdot/\Theta)$,
  then $h(\widetilde{X}, \Theta) = \Theta\widetilde{X}$, $\widetilde{X} \sim F(\cdot)$.

- Inverse CDF: if $X \sim F_\Theta(\cdot)$ where F_Θ is a distribution function on $\mathbb{R}$,
  then $h(\widetilde{X}, \Theta) = F_\Theta^{-1}(\widetilde{X})$, $\widetilde{X} \sim \mathrm{Uniform}(0, 1)$.

Papaspiliopoulos and Roberts (2007) extend the work of Papaspiliopoulos et al. (2007) to Bayesian linear hierarchical models with Gaussian and non-Gaussian error distributions. Papaspiliopoulos and Roberts (2007) demonstrate that for both the centered and non-centered parametrizations, the convergence of a Gibbs sampler depends on the relative tail weights of the error distributions. In some circumstances,



(a) Centered. (b) Non-centered.

Figure 2.2: Centered vs non-centered parametrizations for a hierarchical model.

the Gibbs sampler for the centered parametrization may be uniformly ergodic, but the Gibbs sampler for the non-centered parametrization may be sub-geometric. The reverse can also happen. Their final recommendation is that a hybrid Gibbs sampler with a combination of centered and non-centered parametrizations may be advantageous. One parametrization may be useful for visiting the tails of the distribution, whereas the other may be useful for visiting the modal area.

There are two important reasons why the approach I am introducing is different from these methods:

1. The new approach uses a random-walk Metropolis sampler instead of a Gibbs sampler.

   Using a random-walk Metropolis sampler can be much easier than using a Gibbs sampler. The user does not need to figure out any conditional densities, if they can write down the log unnormalized density they can use a random-walk Metropolis sampler. Using a Gibbs sampler requires some paper-and-pencil work to find the conditional densities. This is impossible for many models.

2. These transformations are *within-model* transformations, the new method transforms the entire model. Hence our approach is different from the existing transformations for Gibbs samplers.

   I'm using *within-model* to mean the transformations alter just some of the pa-

rameters in a model. Such as in the location transformation above; replace $X\ F(\cdot - \Theta)$ with $X = \widetilde{X} + \Theta;\ \widetilde{X} \sim F(\cdot)$. My new approach requires no such selection or work. Analogous to the difference between a random-walk Metropolis sampler and a Gibbs sampler, if the user can write down the log unnormalized density they can use the transformation method without doing the paper-and-pencil work necessary for a *within-model* transformation method.

The transformations they study are "Gibbs-friendly" allowing Gibbs sampling of both models. No theorem shows that "Gibbs-friendly" transformations are generally useful in producing geometric ergodicity. The analysis is model by model.

That being said, there are some settings where the within-model methodology will have greater efficiency than the whole model methodology and in the end be worth the trouble — sometimes Gibbs is better than random-walk Metropolis. This new transformation method is different from the existing transformation methods for Gibbs samplers 2. Because these techniques can apply to different models or analyses, it seems likely that these two techniques can be complementary.

Chapter 3

Change-of-Variable and Isotropy

We are concerned with continuous random variables with support on $\mathbb{R}^k$, if a one-to-one and onto transformation is used, the density transformation formula is simple. Before we can state this formula , we need the ∇ ('del') operator and the multivariate chain rule.

In Section 3.3 we will introduce a class of transformations that work well with variable transformation. We also establish conditions on transformations in this class that will induce always-positive densities with continuous first derivatives when applied to random variables with always-positive densities with continuous first derivatives. These transformations will be used in Chapter 4 to induce super-exponential densities from exponentially light densities and sub-exponential densities.

3.1 Multivariate Derivatives and the Chain Rule

If f is a function, and x is in the domain of f, the typical understanding is that $\nabla f(x)$ is the matrix of first derivatives of f at x.

The operator ∇ takes functions from $\mathbb{R}^m$ to $\mathbb{R}^n$ as input and returns a function from $\mathbb{R}^m$ to $n \times m$ real-valued matrices. The function signature is

$$\nabla : \left(\mathbb{R}^m \mapsto \mathbb{R}^n\right) \to \left(\mathbb{R}^m \mapsto \mathbb{R}^m \times \mathbb{R}^n\right).$$

If f is a function from $\mathbb{R}^m$ to $\mathbb{R}^n$, then ∇f is a function from $\mathbb{R}^m$ to $\mathbb{R}^n \times \mathbb{R}^m$. And for $x \in \mathbb{R}^m$, evaluating $\big(\nabla f\big)(x)$ — usually written $\nabla f(x)$ — gives the matrix containing the first order partial derivatives of f at x[1]. Using $\nabla f(x)_{ij}$ to denote the entry in the i^{th} row and j^{th} column of $\nabla f(x)$, the matrix can be defined component-wise by

$$\nabla f(x)_{ij} = \frac{\partial f(x)_i}{\partial x_j}.$$

This can also be written as

$$\nabla f(x) = \begin{bmatrix} \frac{\partial f(x)_1}{\partial x_1} & \cdots & \frac{\partial f(x)_1}{\partial x_m} \\ \vdots & \ddots & \vdots \\ \frac{\partial f(x)_n}{\partial x_1} & \cdots & \frac{\partial f(x)_n}{\partial x_m} \end{bmatrix}.$$

The careful observer will note that for $f : \mathbb{R}^k \to \mathbb{R}$, this definition of ∇ means that $\nabla f(x)$ is a $1 \times k$ matrix, or a k-dimensional row-vector. Commonly $\nabla f(x)$ is treated as a column vector. The difference will not have much impact, except that transposes may seem to be in odd places. Lang (1993, XIII, §2) has a more detailed discussion of multivariate derivatives.

Now we can express the multivariate chain rule. Let f and g be functions,

$$f : \mathbb{R}^m \to \mathbb{R}^n$$
$$g : \mathbb{R}^l \to \mathbb{R}^m.$$

Then $f \circ g$ is a function from $\mathbb{R}^l$ to $\mathbb{R}^n$. The multivariate chain rule states

$$\nabla\big(f \circ g\big)(x) = \nabla(f)\big(g(x)\big)\ \nabla(g)(x)$$

which is a n by l matrix. This has the maximum number of parenthesis added for

[1]Here we are taking the existence of $\nabla f(x)$ for granted, but we do not do so in our proofs.

clarity. It is the same statement as

$$\nabla\big(f \circ g\big)\ (x) = \nabla f(g(x))\ \nabla g(x).$$

We will omit the parenthesis as much as possible, but include them when necessary to avoid ambiguity. Lang (1993, XIII, §3) has a more detailed discussion of the multivariate chain rule.

3.2 Change-of-Variable

Unless it is entirely clear from the context, densities will be labeled with the random variable they refer to with a subscript. E.g. if Y is a random variable, its density, f will be written as f_Y.

Let X be a continuous random variable on $\mathbb{R}^k$ with density f_X, and $h : \mathbb{R}^k \to \mathbb{R}^k$ be onto, one-to-one and differentiable. If $Y = h^{-1}(X)$ then f_Y is defined:

$$f_Y(y) = f_X\big(h(y)\big)\ |\nabla h(y)| \tag{3.1}$$

where $|\nabla h(y)|$ is the absolute value of the determinant of $\nabla h(y)$. It follows from (3.1) that

$$\log f_Y(y) = \log f_X\big(h(y)\big) + \log |\nabla h(y)| \tag{3.2}$$

$$\nabla \log f_Y(y) = \nabla \log f_X\big(h(y)\big)\ \nabla h(y) + \nabla \log |\nabla h(y)| \tag{3.3}$$

This definition is usually given the other way, i.e. $Y = h(X)$ so $X = h^{-1}(Y)$. The formulas are all the same, just with h and h^{-1} swapped. In our case h and h^{-1} are one-to-one, onto and invertible functions, so the transformation formulas are exactly the same if we replace h with h^{-1}. This difference is cosmetic, we are only using h

instead of h^{-1} to visually simplify the formulas.

3.3 Isotropy

In this section, h and h^{-1} will always be functions from $\mathbb{R}^k$ to $\mathbb{R}^k$, and f will always be a function from $\mathbb{R}$ to $\mathbb{R}$. The usual euclidean norm on $\mathbb{R}^k$ will be denoted as $\|\cdot\| = \sqrt{\sum_{i=1}^k x_i^2}$.

We say $h : \mathbb{R}^k \to \mathbb{R}^k$ is *isotropic* if it has the form

$$\begin{aligned} h(\gamma) &= f(\|\gamma\|)\frac{\gamma}{\|\gamma\|}, \qquad \gamma \neq 0 \\ h(0) &= 0 \end{aligned} \tag{3.4}$$

for some $f : \mathbb{R} \to \mathbb{R}$. We will refer to such constructions as *isotropic h defined with f*.

Let $h : \mathbb{R}^k \to \mathbb{R}^k$ be any function used for variable transformation of a variable on $\mathbb{R}^k$, h is not necessarily an isotropic function. Then a desirable set of conditions on $h : \mathbb{R}^k \to \mathbb{R}^k$ are:

$$h \text{ is one-to-one and onto;} \tag{3.5a}$$

$$\gamma \mapsto \nabla h(\gamma) \text{ is continuous;} \tag{3.5b}$$

$$\gamma \mapsto \det\big(\nabla h(\gamma)\big) \text{ is always positive and continuous;} \tag{3.5c}$$

$$\gamma \mapsto \nabla \det\big(\nabla h(\gamma)\big) \text{ is continuous.} \tag{3.5d}$$

The benefit of these conditions on h is given in the following lemma.

Lemma 1. *Let X be a $\mathbb{R}^k$ valued random variable with always positive density π_X that has continuous first derivatives and $h : \mathbb{R}^k \to \mathbb{R}^k$ satisfy conditions* (3.5a)–(3.5d)*. If $Y = h^{-1}(X)$, then the induced density π_Y will be always positive with continuous*

first derivatives.

Proof. Combining the fact that $\det\big(\nabla h(Y)\big)$ is always positive, with the density formula from (3.1), π_Y is given by

$$\pi_Y(y) = \pi_X\big(h(y)\big)\ \det\big(\nabla h(y)\big).$$

π_Y is the product of two always-positive functions, so π_Y is always positive. The first derivative of π_Y is

$$\nabla\pi_Y(y) = \det\big(\nabla h(y)\big)\nabla\pi_X\big(h(y)\big)\nabla h(y) + \pi_X\big(h(y)\big)\nabla\det\big(\nabla h(y)\big).$$

Condition (3.5d) states that $\nabla\det\big(\nabla h(y)\big)$ is continuous for all y, and since the π_X is continuous, this right-hand term is continuous. Since $\det\big(\nabla h(y)\big)$ is continuous by (3.5c) , showing that the left-hand term is continuous only requires showing that $\nabla\pi_X\big(h(y)\big)\nabla h(y)$ is continuous. This is a vector made by multiplying and adding continuous functions, so it must also be continuous. Hence $\nabla\pi_Y$ is continuous. □

Different fs are going to be used, one class of fs will be used to transform exponentially light densities to super-exponential densities, another class of fs will be used to transform sub-exponential densities to super-exponential. Here we give conditions on $f : \mathbb{R} \to \mathbb{R}$ that will cause isotropic h defined with f to satisfy conditions (3.5a)–(3.5d).

Lemma 2. *Let $h : \mathbb{R}^k \to \mathbb{R}^k$ be an isotropic function defined with $f : \mathbb{R} \to \mathbb{R}$ that*

satisfies the conditions:

$$f'(s) \text{ exists and is greater than zero for } s > 0; \tag{3.6a}$$

$$f(s) \to \infty \text{ as } s \to \infty; \tag{3.6b}$$

$$f''(s) \text{ exists and is continuous for } s > 0; \tag{3.6c}$$

$$f(s) \to 0 \text{ as } s \searrow 0 \text{ and } f(0) = 0; \tag{3.6d}$$

$$f'(0) \text{ exists and is greater than zero;} \tag{3.6e}$$

$$f''(0) \text{ exists and is equal to zero;} \tag{3.6f}$$

$$\lim_{s \searrow 0} \frac{f'(s) - f(s)/s}{s} = f''(0). \tag{3.6g}$$

Then the following are true

$$h \text{ satisfies conditions (3.5a)–(3.5d);} \tag{3.7a}$$

$$\text{for } \gamma \neq 0,\ \det\big(\nabla h(\gamma)\big) = f'(\|\gamma\|)\left(\frac{f(\|\gamma\|)}{\|\gamma\|}\right)^{k-1}, \tag{3.7b}$$
$$\text{and } \det\big(\nabla h(0)\big) = f'(0)^k;$$

$$\frac{\gamma}{\|\gamma\|} = \frac{h(\gamma)}{\|h(\gamma)\|} \text{ for } \gamma \neq 0; \tag{3.7c}$$

$$\nabla h(\gamma)\gamma = f'(\|\gamma\|)\gamma \qquad \text{for all } \gamma; \tag{3.7d}$$

$$\begin{aligned} &\text{if } \mathrm{I}_k \text{ is the } k \times k \text{ identity matrix, then } \nabla h(0) = f'(0)\,\mathrm{I}_k \text{ and} \\ &\text{for } \gamma \neq 0,\ \nabla h(\gamma) = \frac{f(\|\gamma\|)\,\mathrm{I}_k}{\|\gamma\|} + \left[f'(\|\gamma\|) - \frac{f(\|\gamma\|)}{\|\gamma\|}\right]\frac{\gamma\gamma^T}{\|\gamma\|^2}; \end{aligned} \tag{3.7e}$$

$$\text{for } \gamma \neq 0,\ \big[\nabla h(\gamma)\big]^2 = \frac{f(\|\gamma\|)^2}{\|\gamma\|^2}\,\mathrm{I}_k + \left[f'(\|\gamma\|)^2 - \frac{f(\|\gamma\|)^2}{\|\gamma\|^2}\right]\frac{\gamma\gamma^T}{\|\gamma\|^2}; \tag{3.7f}$$

$$\begin{aligned} &\text{for any } x \in \mathbb{R}^k \text{ and } \gamma \neq 0 \\ &x^T\nabla h(\gamma)^T\nabla h(\gamma)x = \\ &\qquad \frac{f(\|\gamma\|)^2}{\|\gamma\|^2}\|x\|^2 + \left[f'(\|\gamma\|)^2 - \frac{f(\|\gamma\|)^2}{\|\gamma\|^2}\right]\left(\frac{h(\gamma)\cdot x}{\|h(\gamma)\|}\right)^2. \end{aligned} \tag{3.7g}$$

Proof of this lemma is in Section 3.4.

Conditions (3.6a)–(3.6g) may seem intimidating and restrictive. However, only conditions and (3.6a)–(3.6c) apply when s is large. The rest of the conditions only restrict the behavior of f in a neighborhood of zero. Furthermore, these conditions are easily satisfied with a carefully chosen polynomial, e.g.

$$f(x) = ax^3 + bx, \qquad a \geq 0 \text{ and } b > 0$$

will satisfy conditions (3.6c)–(3.6g). Our end goal is to impact the tail behavior of a density. Thus we can simply choose a $g : \mathbb{R} \to \mathbb{R}$ satisfying (3.6a)–(3.6g) that has

the desired behavior for large s piecewise define an f that is a polynomial satisfying (3.6c)–(3.6g) near $s = 0$, and is equal to g for large s. This piecewise definition just needs to be done such that f'' is continuous and the intermediate pieces of f satisfy (3.6a)–(3.6g), which can typically be accomplished with carefully chosen cubic polynomials. Thus, the odd conditions in this list are not cause for concern.

An obvious consequence of Lemma 2 is isotropic h defined with f satisfying (3.6a)–(3.6g) will satisfy the conditions of Lemma 1. Hence this h will transform the random variable β with variables with always-positive density, π_β with continuous first derivatives to the random variable γ with always-positive density π_γ with continuous first derivatives.

The effort in developing the behavior of isotropic h is worth the trouble, as isotropic functions greatly simplify the coming tail behavior proofs. Early versions of this theory did not use isotropic functions, and the proofs in those cases are much more complicated (Johnson and Geyer, 2010).

Another useful property of isotropic h defined using f is the simple form of h^{-1}. If $h(\gamma) = \beta$, it follows that

$$\begin{aligned} h^{-1}(\beta) &= f^{-1}(\|\beta\|)\frac{\beta}{\|\beta\|}, \qquad \beta \neq 0 \\ h^{-1}(0) &= 0 \end{aligned} \tag{3.8}$$

The convenience in using an isotropy can be seen if we consider the dot product

$$\frac{\gamma}{\|\gamma\|} \cdot \nabla\Big(g\big(h(\gamma)\big)\Big).$$

By the multivariate chain rule, this dot product equals

$$\frac{\gamma}{\|\gamma\|} \cdot \nabla g\big(h(\gamma)\big)\nabla h(\gamma).$$

Because h is an isotropy, $\nabla h(\gamma)$ is symmetric, so this dot product equals

$$\frac{\nabla h(\gamma)\gamma}{\|\gamma\|} \cdot \nabla g\big(h(\gamma)\big)$$

and because h is isotropic, $\nabla h(\gamma)\gamma = f'(\|\gamma\|)\gamma$ and $\gamma/\|\gamma\| = h(\gamma)/\|h(\gamma)\|$. Therefore, this dot product is equal to

$$f'(\|\gamma\|)\frac{h(\gamma)}{\|h(\gamma)\|} \cdot \nabla g\big(h(\gamma)\big).$$

As will be seen in the coming chapters, this manipulation is very powerful and useful.

3.3.1 Multivariate Differentiability

Establishing the existence of a multivariate derivative can be non-trivial.

Rockafeller and Wets (1998, Ch. 7(D)) define the *difference quotient function* $\Delta_\tau g(\bar{x}) : \mathbb{R}^n \to \overline{\mathbb{R}}$ for functions $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ by

$$\Delta_\tau g(\bar{x})(w) := \frac{g(\bar{x} + \tau w) - g(\bar{x})}{\tau} \qquad \text{for } \tau \neq 0. \tag{3.9}$$

If as $\tau \searrow 0$ the functions $\Delta_\tau g(\bar{x})$ converge continuously to a function v then g is *semidifferentiable* at $\bar{x}$ (Rockafeller and Wets, 1998, Theorem 7.21). Furthermore if the semiderivative of g at $\bar{x}$ depends linearly on w, then g is differentiable at $\bar{x}$ (Rockafeller and Wets, 1998, Corollary 7.22).

3.4 Proof of Lemma 2

In this section, and the contained sub-sections, $h : \mathbb{R}^k \to \mathbb{R}^k$ shall always be an isotropic h defined with $f : \mathbb{R} \to \mathbb{R}$ that satisfies conditions (3.6a)–(3.6g). Note that it is entirely acceptable for f to not be defined for negative numbers. The proof

follows in exactly the same way, if $f'(0)$ and $f''(0)$ are interpreted as the right-hand derivatives of f and f' at 0, respectively.

First we will establish that (3.5a) and (3.7c) hold.

Note that it follows from (3.6a) and (3.6e) that f is differentiable for all non-negative x, and therefore continuous for all non-negative x. Combining this with (3.6d) means that h is continuous everywhere. It also follows from (3.6a) and (3.6e) that f is strictly increasing for non-negative x, so f is one-to-one which means that h must also be one-to-one. Since f is one-to-one, (3.6b) and the fact that

$$\|h(\gamma)\| = \frac{\|f(\|\gamma\|)\gamma\|}{\|\gamma\|} = f(\|\gamma\|),$$

imply that h is onto $\mathbb{R}^k$. Hence h is one-to-one and onto so h satisfies (3.5a).

Let $\gamma \neq 0$ and consider

$$\frac{h(\gamma)}{\|h(\gamma)\|} = \frac{f(\|\gamma\|)\gamma/\|\gamma\|}{f(\|\gamma\|)} = \frac{\gamma}{\|\gamma\|}$$

hence (3.7c) holds.

3.4.1 $\nabla h(\gamma)$

This sub-section will establish (3.7d)–(3.7g) and (3.5b).

For $\gamma \neq 0$ we have

$$\frac{\partial}{\partial \gamma_k}\left(\sum_{i=1}^{d}\gamma_i^2\right)^{1/2} = \frac{1}{2}\left(\sum_{i=1}^{d}\gamma_i^2\right)^{-1/2} 2\gamma_k = \frac{\gamma_k}{\|\gamma\|}$$

and

$$\nabla\|\gamma\| = \frac{\gamma^T}{\|\gamma\|}$$

hence, assuming that f is differentiable at $\|\gamma\| \neq 0$

$$\begin{aligned}\nabla h(\gamma) &= \frac{\gamma}{\|\gamma\|}\nabla f(\|\gamma\|) + \frac{f(\|\gamma\|)}{\|\gamma\|}\nabla\gamma + f(\|\gamma\|)\gamma\nabla\frac{1}{\|\gamma\|} \\ &= \frac{f'(\|\gamma\|)\gamma\gamma^T}{\|\gamma\|^2} + \frac{f(\|\gamma\|)\,\mathrm{I}_k}{\|\gamma\|} - \frac{f(\|\gamma\|)\gamma\gamma^T}{\|\gamma\|^3} \\ &= \frac{f(\|\gamma\|)\,\mathrm{I}_k}{\|\gamma\|} + \left[f'(\|\gamma\|) - \frac{f(\|\gamma\|)}{\|\gamma\|}\right]\frac{\gamma\gamma^T}{\|\gamma\|^2}\end{aligned} \tag{3.10}$$

where I_k is the k-dimensional identity matrix. Now set

$$\nabla h(0) = f'(0)\,\mathrm{I}_k$$

and consider the directional derivative in the direction γ

$$\begin{aligned}h'(0;\gamma) &= \lim_{s\downarrow 0}\frac{h(s\gamma) - h(0)}{s} \\ &= \lim_{s\downarrow 0}\frac{f(s\|\gamma\|)\gamma}{s\|\gamma\|} \\ &= f'(0)\gamma,\end{aligned} \tag{3.11}$$

where $f'(0)$ denotes the one-sided derivative of f at zero. Since these directional derivatives agree with

$$\nabla h(0) = f'(0)\,\mathrm{I}_k,$$

so (3.7e) holds because (3.6e) states that $f'(0)$ exists and is postitive.

It's clear from (3.7e) that $\nabla h(\gamma)$ is continuous at $\gamma \neq 0$, as it is the addition and multiplication of continuous non-zero functions. Thus to establish (3.5b), it is only necessary to show that

$$\lim_{\|\gamma\|\to 0}\nabla h(\gamma) = \nabla h(0). \tag{3.12}$$

It's clear that

$$\lim_{\|\gamma\|\to 0} \frac{f(\|\gamma\|)}{\|\gamma\|} = f'(0)$$

hence

$$\lim_{\|\gamma\|\to 0} \frac{f(\|\gamma\|)}{\|\gamma\|} \mathrm{I}_k = f'(0)\, \mathrm{I}_k\,. \tag{3.13}$$

Now consider the limit as $\|\gamma\|$ goes to zero of the right-hand component of $\nabla h(\gamma)$,

$$\lim_{\|\gamma\|\to 0} \left[f'(\|\gamma\|) - \frac{f(\|\gamma\|)}{\|\gamma\|} \right] \frac{\gamma\gamma^T}{\|\gamma\|^2}. \tag{3.14}$$

It follows from (3.6f) that $f'(x)$ is continuous at $x = 0$, so

$$\lim_{\|\gamma\|\to 0} f'(\|\gamma\|) - \frac{f(\|\gamma\|)}{\|\gamma\|} = 0. \tag{3.15}$$

Now consider the ij^{th} entry of the matrix $\gamma\gamma^T/\|\gamma\|^2$, equal to

$$\frac{\gamma_i\gamma_j}{\sum_l \gamma_l^2}.$$

By rearranging the sum in the denominator, and noting that for all γ_i and γ_j

$$|\gamma_i| \le \sqrt{\gamma_i^2 + \gamma_j^2} \text{ and } |\gamma_j| \le \sqrt{\gamma_i^2 + \gamma_j^2}$$

we can then bound the magnitude of the ij^{th} entry above by

$$\frac{\gamma_i^2 + \gamma_j^2}{\gamma_i^2 + \gamma_j^2 + \sum_{l\neq i,j} \gamma_l^2}.$$

which is finite and bounded above by one for all $\gamma \neq 0$. Hence the entries of the

matrix $\gamma\gamma^T/\|\gamma\|^2$ are between -1 and 1 as $\|\gamma\|$ goes to zero. So it follows from (3.15) that the limit in (3.14) is equal to $0\, I_k$, or the $k \times k$ matrix of all zeros. Therefore,

$$\lim_{\|\gamma\|\to 0} \nabla h(\gamma) = f'(0)\, \mathrm{I}_k + 0\, \mathrm{I}_k = \nabla h(0)$$

so (3.12) is true and so $\nabla h(\gamma)$ is continuous everywhere and (3.5b) holds.

Now consider $\nabla h(\gamma)\gamma$. First, let $\gamma = 0$, then

$$\nabla h(0)0 = f'(0)\, \mathrm{I}_k\, 0 = f'(0)0 = f'(\|\gamma\|)\gamma.$$

Now let $\gamma \neq 0$, then

$$\begin{aligned}
\nabla h(\gamma)\gamma &= \left[\frac{f(\|\gamma\|)\, \mathrm{I}_k}{\|\gamma\|} + \left[f'(\|\gamma\|) - \frac{f(\|\gamma\|)}{\|\gamma\|}\right]\frac{\gamma\gamma^T}{\|\gamma\|^2}\right]\gamma \\
&= \frac{f(\|\gamma\|)\gamma}{\|\gamma\|} + f'(\|\gamma\|)\gamma - \frac{f(\|\gamma\|)\gamma}{\|\gamma\|} \\
&= f'(\|\gamma\|)\gamma
\end{aligned}$$

so (3.7d) holds.

Next we will find the value of $\left[\nabla h(\gamma)\right]^2$.

$$\begin{aligned}
\left[\nabla h(\gamma)\right]^2 &= \nabla h(\gamma)\left(\frac{f(\|\gamma\|)}{\|\gamma\|}\, \mathrm{I}_k + \left[\frac{f'(\|\gamma\|)}{\|\gamma\|^2} - \frac{f(\|\gamma\|)}{\|\gamma\|^3}\right]\gamma\gamma^T\right) \\
&= \frac{f(\|\gamma\|)}{\|\gamma\|}\nabla h(\gamma) + \left[\frac{f'(\|\gamma\|)^2}{\|\gamma\|^2} - \frac{f(\|\gamma\|)f'(\|\gamma\|)}{\|\gamma\|^3}\right]\gamma\gamma^T.
\end{aligned} \tag{3.16}$$

If we expand $\frac{f(\|\gamma\|)}{\|\gamma\|}\nabla h(\gamma)$ using the definition of $\nabla h(\gamma)$ from (3.17) we can see that

$$\frac{f(\|\gamma\|)}{\|\gamma\|}\nabla h(\gamma) = \frac{f(\|\gamma\|)^2}{\|\gamma\|^2}\, \mathrm{I}_k + \left[\frac{f'(\|\gamma\|)f(\|\gamma\|)}{\|\gamma\|^3} - \frac{f(\|\gamma\|)^2}{\|\gamma\|^4}\right]\gamma\gamma^T$$

The last line of (3.16) contains both plus and minus $\frac{f(\|\gamma\|)f'(\|\gamma\|)}{\|\gamma\|^3}\gamma\gamma^T$. Canceling these terms, we have the equality

$$\big[\nabla h(\gamma)\big]^2 = \frac{f(\|\gamma\|)^2}{\|\gamma\|^2}\,\mathrm{I}_k + \left[f'(\|\gamma\|)^2 - \frac{f(\|\gamma\|)^2}{\|\gamma\|^2}\right]\frac{\gamma\gamma^T}{\|\gamma\|^2}.$$

so hence (3.7f) holds.

Let $\gamma \neq 0$, x be any vector and consider

$$\begin{aligned} x^T\nabla h(\gamma)^T\nabla h(\gamma)x &= x^T\big[\nabla h(\gamma)\big]^2 x \\ &= \frac{f(\|\gamma\|)^2}{\|\gamma\|^2}x^Tx + \left[f'(\|\gamma\|)^2 - \frac{f(\|\gamma\|)^2}{\|\gamma\|^2}\right]\frac{x^T\gamma\gamma^Tx}{\|\gamma\|^2} \\ &= \frac{f(\|\gamma\|)^2}{\|\gamma\|^2}\|x\|^2 + \left[f'(\|\gamma\|)^2 - \frac{f(\|\gamma\|)^2}{\|\gamma\|^2}\right]\left(\frac{x^T\gamma}{\|\gamma\|}\right)^2 \\ &= \frac{f(\|\gamma\|)^2}{\|\gamma\|^2}\|x\|^2 + \left[f'(\|\gamma\|)^2 - \frac{f(\|\gamma\|)^2}{\|\gamma\|^2}\right]\left(\frac{h(\gamma)\cdot x}{\|h(\gamma)\|}\right)^2. \end{aligned}$$

Therefore (3.7g) holds.

3.4.2 $\det\big(\nabla h(\gamma)\big)$

This sub-section will establish (3.7b) and (3.5c).

We now calculate $\det\big(\nabla h(\gamma)\big)$, where ∇h is given by (3.7e). First, (3.7d) states that

$$\nabla h(\gamma)\gamma = f'(\|\gamma\|)\gamma$$

Thus γ is an eigenvector of $\nabla h(\gamma)$ associated with the eigenvalue $f'(\|\gamma\|)$. For any

non-zero vector v orthogonal to γ,

$$\begin{aligned}\nabla h(\gamma)v &= \left[\frac{f(\|\gamma\|)\,\mathrm{I}_k}{\|\gamma\|} + \left[f'(\|\gamma\|) - \frac{f(\|\gamma\|)}{\|\gamma\|}\right]\frac{\gamma\gamma^T}{\|\gamma\|^2}\right]v \\ &= \frac{f(\|\gamma\|)}{\|\gamma\|}v.\end{aligned}$$

So v is an eigenvector of $\nabla h(\gamma)$ associated with the eigenvalue $f(\|\gamma\|)/\|\gamma\|$. The vectors v orthogonal to γ span an eigenspace of dimension $k-1$, so the multiplicity of the eigenvalue $f(\|\gamma\|)/\|\gamma\|$ is $k-1$. We also observe from (3.7e) that $\nabla h(\gamma)$ is a symmetric matrix, so its determinant is the product of its eigenvalues (Harville, 1997, Theorem 21.6.1). Thus,

$$\det\big(\nabla h(\gamma)\big) = f'(\|\gamma\|)\left(\frac{f(\|\gamma\|)}{\|\gamma\|}\right)^{k-1}, \qquad \gamma \neq 0.$$

It's clear from (3.7e) that

$$\det\big(\nabla h(0)\big) = \big(f'(0)\big)^k$$

so (3.7b) holds.

Since $f(0) = 0$, and f' is continuous at zero, it's clear that $\det\big(\nabla h(\gamma)\big)$ is continuous at $\gamma = 0$. It's clear that $\det\big(\nabla h(\gamma)\big)$ is continuous for $\gamma \neq 0$, since $f'(x)$ and $f(x)/x$ are continuous for $x > 0$. Finally, $f'(x)$ is positive for all non-negative x, and $f(x)/x$ is positive for $x > 0$ so $\det\big(\nabla h(\gamma)\big)$ is always positive so (3.5c) holds.

3.4.3 $\nabla \det\big(\nabla h(\gamma)\big)$

This sub-section will establish (3.5d)

First we will consider $\gamma \neq 0$. Using the quotient rule for differentiation we can see

that

$$\frac{\partial}{\partial \gamma_i}\frac{f(\|\gamma\|)}{\|\gamma\|} = \frac{\|\gamma\| f'(\|\gamma\|)\gamma_i/\|\gamma\| - f(\|\gamma\|)\gamma_i/\|\gamma\|}{\|\gamma\|^2}$$

so it follows that

$$\nabla \frac{f(\|\gamma\|)}{\|\gamma\|} = \left[\frac{f'(\|\gamma\|)}{\|\gamma\|} - \frac{f(\|\gamma\|)}{\|\gamma\|^2}\right]\frac{\gamma^T}{\|\gamma\|}.$$

Thus using the product rule for differentiation

$$\begin{aligned}\nabla \det\big(\nabla h(\gamma)\big) =& f''(\|\gamma\|)\left(\frac{f(\|\gamma\|)}{\|\gamma\|}\right)^{k-1}\frac{\gamma^T}{\|\gamma\|} \\ &+ (k-1)f'(\|\gamma\|)\left(\frac{f(\|\gamma\|)}{\|\gamma\|}\right)^{k-2}\left[\frac{f'(\|\gamma\|) - f(\|\gamma\|)/\|\gamma\|}{\|\gamma\|}\right]\frac{\gamma^T}{\|\gamma\|}\end{aligned} \tag{3.17}$$

By (3.6c) $f''(x)$ is continuous for $x > 0$, so this will be continuous for $\gamma \neq 0$.

Establishing the existence or value of $\nabla \det\big(\nabla h(\gamma)\big)$ at $\gamma = 0$ is not so obvious.

Our strategy is to show that $\Delta_\tau \det\big(\nabla h(0)\big)(w)$ converges continuously to the 0 vector as $\tau \searrow 0$ for all w. The convergence is continuous since $\det\big(\nabla h(\gamma)\big)$ only depends on the magnitude of γ, not the direction. The 0 vector is a linear function of w, so the $\det\big(\nabla h(\gamma)\big)$ is differentiable at 0, and this derivative is the 0 vector.

We want to establish that $\Delta_\tau \det\big(\nabla h(0)\big)(w)$ converges continuously to the zero vector. Let $w \neq 0$ be given and consider the following.

The key is to notice that $\det\big(\nabla h(\tau w)\big)$ depends only on the magnitude of τw, not the direction of w. To simplify this argument, we are going to take

$$g(x) = f'(x)\left(\frac{f(x)}{x}\right)^{k-1}$$

and set $g(0) = f'(0)^k$. Now if we take $s = \tau\|w\|$ it follows that

$$\lim_{\tau \searrow 0} \Delta_\tau \Big(\det\big(\nabla h(0)\big) \Big)(w) = \|w\| \lim_{s \searrow 0} \frac{g(s) - g(0)}{s}.$$

It has already been shown that g is continuous at 0, so both the numerator and denominator of this fraction go to zero as s does. So by L'Hôpital's rule, this right-hand limit will be equal to

$$\lim_{s \searrow 0} \frac{g'(s)}{1},$$

if this new limit exists.

Now,

$$\begin{aligned} g'(s) =& f''(s) \left(\frac{f(s)}{s} \right)^{k-1} \\ &+ (k-1) f'(s) \left(\frac{f(s)}{s} \right)^{k-2} \left[\frac{f'(s) - f(s)/s}{s} \right]. \end{aligned}$$

Since $f(0) = 0$ we can subtract $f(0)$ from each $f(s)$ while maintaining equality. Thus

$$\begin{aligned} g'(s) =& f''(s) \left(\frac{f(s) - f(0)}{s} \right)^{k-1} \\ &+ (k-1) f'(s) \left(\frac{f(s) - f(0)}{s} \right)^{k-2} \left[\frac{f'(s) - \big((f(s) - f(0)\big)/s}{s} \right]. \end{aligned}$$

So it follows from (3.6g) that

$$\begin{aligned} \lim_{s \searrow 0} g'(s) &= f''(0) f'(0)^{k-1} + (k-1) f'(0)^{k-1} f''(0) \\ &= k f''(0) f'(0)^{k-1} \\ &= 0 \end{aligned} \tag{3.18}$$

Since this limit is 0, it follows from (Rockafeller and Wets, 1998, Theorem 7.21) and (Rockafeller and Wets, 1998, Corollary 7.22) that $\nabla \det\big(\nabla h(0)\big)$ exists and is the zero vector.

Now we must show that the limit

$$\lim_{\|\gamma\| \to 0} \nabla \det\big(\nabla h(\gamma)\big)$$

exists and is the zero vector. From the form of $\nabla \det\big(\nabla h(\gamma)\big)$ in (3.17), we can see that this limit will be a product of a scalar and the vector $\gamma^T/\|\gamma\|$. We can rewrite $\nabla \det\big(\nabla h(\gamma)\big)$ as

$$g'(\|\gamma\|)\frac{\gamma^T}{\|\gamma\|}.$$

Since the components of $\gamma^T/\|\gamma\|$ are bounded between negative one and one as $\|\gamma\|$ goes to zero, we see that

$$\lim_{\|\gamma\| \to 0} \nabla \det\big(\nabla h(\gamma)\big) = \lim_{\|\gamma\| \to 0} g'(\|\gamma\|)\frac{\gamma^T}{\|\gamma\|}$$

which is equal to zero by (3.18). Therefore $\det\big(\nabla h(\gamma)\big)$ has continuous first derivatives and (3.5d) holds.

Chapter 4

Variable Transformation in Random-Walk Metropolis

4.1 Exponentially Light to Super-Exponential

We are going to use an isotropic function h such that if β is a random variable on $\mathbb{R}^k$ with exponentially light density π_β, the random variable defined by the transformation $\gamma = h^{-1}(\beta)$ will have a super-exponential density π_γ. A very simple f will suffice.

4.1.1 Choice of f

Let p be greater than two and define

$$f(x) = x^p + x. \tag{4.1}$$

So f has derivatives

$$f'(x) = px^{p-1} + 1 \tag{4.2}$$

$$f''(x) = p(p-1)x^{p-2}. \tag{4.3}$$

It's clear that f will satisfy conditions (3.6a)–(3.6g). We will use this f to define an isotropic h of the form (3.4), and set $\gamma = h^{-1}(\beta)$. An example of the relationship relationship between β and γ using this f with $p = 3$ is given in Figure 4.1.

4.1.2 Transformation Theorem

First we need a lemma.

Lemma 3. *For isotropic h with f defined as in* (4.1)

$$\lim_{\|\gamma\|\to\infty} \frac{\gamma}{\|\gamma\|} \cdot \nabla \log\det\big(\nabla h(\gamma)\big) = 0. \tag{4.4}$$

Proof. Recalling the value of $\det\big(\nabla h(\gamma)\big)$ from (3.7b) we can write out $\log\det\big(\nabla h(\gamma)\big)$ as

$$\log f'(\|\gamma\|) + \big(k-1\big)\bigg(\log f(\|\gamma\|) - \log\|\gamma\|\bigg).$$

Therefore $\nabla \log\det\big(\nabla h(\gamma)\big)$ is

$$\frac{f''(\|\gamma\|)\gamma}{f'(\|\gamma\|)\|\gamma\|} + \big(k-1\big)\bigg(\frac{f'(\|\gamma\|)\gamma}{f(\|\gamma\|)\|\gamma\|} - \frac{\gamma}{\|\gamma\|^2}\bigg) \tag{4.5}$$

and the dot product in (4.4) is equal to

$$\frac{f''(\|\gamma\|)}{f'(\|\gamma\|)} + \big(k-1\big)\bigg(\frac{f'(\|\gamma\|)}{f(\|\gamma\|)} - \frac{1}{\|\gamma\|}\bigg). \tag{4.6}$$

For $\|\gamma\|$ large, by the definitions of f' and f'' from (4.2) and (4.3), this dot product can be written as

$$\frac{p(p-1)\|\gamma\|^{p-2}}{p\|\gamma\|^{p-1}+1} + \big(k-1\big)\bigg(\frac{p\|\gamma\|^{p-1}+1}{\|\gamma\|^p + \|\gamma\|} - \frac{1}{\|\gamma\|}\bigg).$$

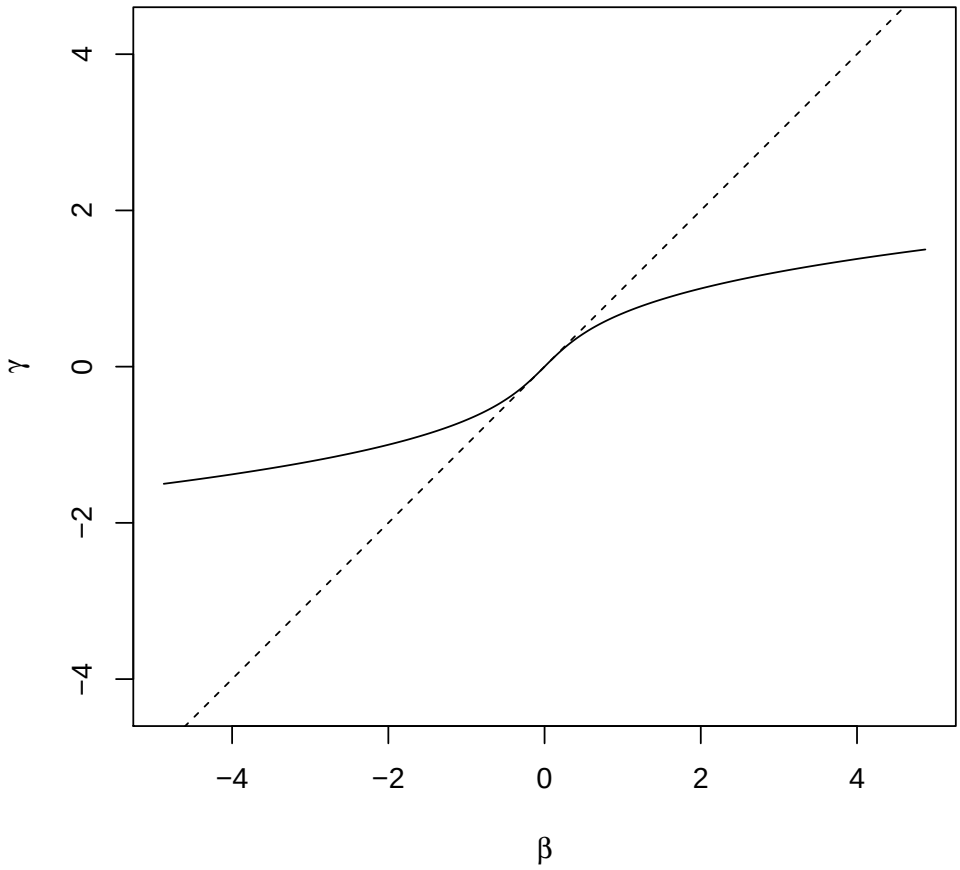


Figure 4.1: Solid line: $\beta = h(\gamma)$ for f from(4.1) with $p = 3$. Dashed line: $\beta = \gamma$.

Since $p > 2$, this clearly goes to zero as $\|\gamma\|$ goes to infinity. □

Theorem 2. *Let π_β be a density on $\mathbb{R}^k$ that is always positive and has continuous first derivatives and isotropic h defined with f from (4.1). Then π_γ, the density induced by the relationship $h(\gamma) = \beta$ is super-exponential.*

Proof. By Lemma 1, π_γ is always positive with continuous first derivatives.

Using the definition of $\nabla \log \pi_\gamma$ from (3.3) we can bound

$$\limsup_{\|\gamma\|\to\infty} \frac{\gamma}{\|\gamma\|} \cdot \nabla \log \pi_\gamma(\gamma)$$

above by the sum of

$$\limsup_{\|\gamma\|\to\infty} \frac{\gamma}{\|\gamma\|} \cdot \nabla \log \pi_\beta\big(h(\gamma)\big) \nabla h(\gamma) \tag{4.7}$$

and

$$\limsup_{\|\gamma\|\to\infty} \frac{\gamma}{\|\gamma\|} \cdot \nabla \log \det\big(\nabla h(\gamma)\big). \tag{4.8}$$

By Lemma 3, the lim sup in (4.8) is zero so we only need to show that the lim sup in (4.7) is $-\infty$.

Because $\nabla h(\gamma)$ is symmetric, the dot product in (4.7) can be rewritten as

$$\frac{\nabla h(\gamma)\gamma}{\|\gamma\|} \cdot \nabla \log \pi_\beta\big(h(\gamma)\big)$$

and by (3.7d) this is equal to

$$f'(\|\gamma\|)\frac{\gamma}{\|\gamma\|} \cdot \nabla \log \pi_\beta\big(h(\gamma)\big).$$

Since $\gamma/\|\gamma\| = h(\gamma)/\|h(\gamma)\|$ for $\gamma \neq 0$, for large $\|\gamma\|$ this dot product can be written

as

$$f'(\|\gamma\|)\frac{h(\gamma)}{\|h(\gamma)\|}\cdot\nabla\log\pi_\beta\big(h(\gamma)\big).$$

Since π_β has exponentially light tails, there is some $\varepsilon > 0$ such that for $\|\gamma\|$ large enough, this is bounded above by

$$-f'(\|\gamma\|)\varepsilon.$$

From the definition of f' in (4.2), it's clear that $f'(\|\gamma\|)$ goes to infinity as $\|\gamma\|$ does. So this upper bound goes to $-\infty$ and the lim sup in (4.7) is $-\infty$. □

Note that Theorem 2 does not preclude an original density that is super-exponential. Thus applying a transformation to a density that is already super-exponential, or super-exponential in some direction will result in a density that is also super-exponential, so this transformation is harmless when it comes to tail-weight.

4.1.3 Curvature Condition

Before we give sufficient conditions on π_β so that π_γ will satisfy the Jarner and Hansen (2000) curvature condition (2.12), we need some preliminary lemmas.

Lemma 4. *If β is a $\mathbb{R}^k$ valued random variable with exponentially light density π_β, and isotropic h is defined using f from* (4.1) *then*

$$\lim_{\|\gamma\|\to\infty}\|\nabla\log\pi_\beta\big(h(\gamma)\big)\nabla h(\gamma)\| = \infty. \tag{4.9}$$

Proof. The norm in (4.9) is non-negative. Thus the norm going to infinity is the same

as the square of the norm going to infinity. Hence (4.9) is equivalent to

$$\lim_{\|\gamma\|\to\infty} \nabla \log \pi_\beta\big(h(\gamma)\big)\big[\nabla h(\gamma)\big]^2 \nabla \log \pi_\beta\big(h(\gamma)\big)^T = \infty.$$

By equation (3.7g), the number inside of this limit is equal to

$$\frac{f(\|\gamma\|)^2}{\|\gamma\|^2}\|\nabla \log \pi_\beta\big(h(\gamma)\big)\|^2 + \left[f'(\|\gamma\|)^2 - \frac{f(\|\gamma\|)^2}{\|\gamma\|^2}\right]\left(\frac{h(\gamma)\cdot \log \pi_\beta\big(h(\gamma)\big)}{\|h(\gamma)\|}\right)^2. \quad (4.10)$$

Note that the left-hand term in this sum is non-negative, so it is sufficient to show that the right-hand term goes to infinity.

The definitions of f and f' from (4.1) and (4.2) let us see that

$$\begin{aligned}\frac{f'(\|\gamma\|)^2}{\|\gamma\|^2} - \frac{f(\|\gamma\|)^2}{\|\gamma\|^4} &= \frac{\big(p\|\gamma\|^{p-1}+1\big)^2}{\|\gamma\|^2} - \frac{\big(\|\gamma\|^p + \|\gamma\|\big)^2}{\|\gamma\|^4} \\ &= \frac{p^2\|\gamma\|^{2p-2} + 2p\|\gamma\|^{p-1} + 1}{\|\gamma\|^2} - \frac{\|\gamma\|^{2p} + 2\|\gamma\|^{p+1} + \|\gamma\|^2}{\|\gamma\|^4} \\ &= \frac{(p^2-1)\|\gamma\|^{2p-2} + 2(p-1)\|\gamma\|^{p-1}}{\|\gamma\|^2}\end{aligned}$$

This is positive because $p > 2$. Since π_β has exponentially light tails, there is a $\varepsilon > 0$ such that for large $\|\gamma\|$

$$\left(h(\gamma)\cdot \log \pi_\beta\big(h(\gamma)\big)\right)^2 > \|h(\gamma)\|^2\varepsilon.$$

Since $\gamma/\|\gamma\| = h(\gamma)/\|h(\gamma)\|$ this is equivalent to

$$\left(\gamma\cdot \log \pi_\beta\big(h(\gamma)\big)\right)^2 > \|\gamma\|^2\varepsilon.$$

So the right-hand term in (4.10) is bounded below by

$$(p^2-1)\|\gamma\|^{2p-2} + 2(p-1)\|\gamma\|^{p-1}\varepsilon$$

which goes to infinity as $\|\gamma\|$ does, so (4.9) holds. □

Lemma 5. *Let β be a $\mathbb{R}^k$ valued random variable with exponentially light density π_β, isotropic h be defined using f from (4.1). If $\gamma = h^{-1}(\beta)$ then the induced density π_γ satisfies*

$$\lim_{\|\gamma\|\to\infty} \frac{\|\nabla \log \pi_\gamma(\gamma)\|}{\|\nabla \log \pi_\beta\big(h(\gamma)\big)\nabla h(\gamma)\|} = 1 \tag{4.11}$$

Proof. By (3.3)

$$\nabla \log \pi_\gamma(\gamma) = \nabla \log \pi_\beta\big(h(\gamma)\big)\nabla h(\gamma) + \nabla \log \det\big(\nabla h(\gamma)\big)$$

so it suffices to show that

$$\frac{\|\nabla \log \det\big(\nabla h(\gamma)\big)\|}{\|\nabla \log \pi_\beta\big(h(\gamma)\big)\nabla h(\gamma)\|} \tag{4.12}$$

goes to zero as $\|\gamma\|$ goes to infinity. Lemma 4 states that this denominator goes to infinity, hence we only need to show that the numerator is bounded.

Recalling the definition of $\nabla \log \det\big(\nabla h(\gamma)\big)$ from (4.5) the numerator of this fraction is

$$\left\| \frac{f''(\|\gamma\|)\gamma}{f'(\|\gamma\|)\|\gamma\|} + (k-1)\left(\frac{f'(\|\gamma\|)\gamma}{f(\|\gamma\|)\|\gamma\|} - \frac{\gamma}{\|\gamma\|^2} \right) \right\|.$$

The triangle inequality bounds this above by the sum of the following three terms

$$\left\| \frac{f''(\|\gamma\|)\gamma}{f'(\|\gamma\|)\|\gamma\|} \right\|$$
$$(k-1)\left\| \frac{f'(\|\gamma\|)\gamma}{f(\|\gamma\|)\|\gamma\|} \right\|$$
$$(k-1)\left\| \frac{\gamma}{\|\gamma\|^2} \right\|$$

Using the definitions of f, f', and f'' from (4.1), (4.2) and (4.3) we can see that these three terms are respectively equal to

$$\frac{p(p-1)\|\gamma\|^{p-2}}{p\|\gamma\|^{p-1}+1}$$
$$(k-1)\frac{p\|\gamma\|^{p-1}+1}{\|\gamma\|^p+\|\gamma\|}$$
$$(k-1)\frac{1}{\|\gamma\|}$$

all of which are clearly bounded for large $\|\gamma\|$. So the numerator of (4.12) is bounded while the denominator goes to infinity, hence (4.11) is true. □

We are going to give two sufficient conditions for π_γ satisfying the Jarner and Hansen (2000) curvature condition. This next lemma will simplify the proofs that the sufficient conditions are strong enough.

Lemma 6. *Let β be an $\mathbb{R}^k$ valued random variable with exponentially light density π_β. If isotropic h is defined using f from (4.1), and $\gamma = h^{-1}(\beta)$, then the induced density π_γ has the property that*

$$\limsup_{\|\gamma\|\to\infty} \frac{\gamma}{\|\gamma\|} \cdot \frac{\nabla\pi_\gamma(\gamma)}{\|\nabla\pi_\gamma(\gamma)\|} \tag{4.13}$$

is bounded above by

$$\limsup_{\|\gamma\|\to\infty} f'(\|\gamma\|)\frac{\gamma}{\|\gamma\|}\cdot\frac{\nabla\log\pi_\beta\big(\nabla h(\gamma)\big)}{\|\nabla\log\pi_\beta\big(\nabla h(\gamma)\big)\nabla h(\gamma)\|} \tag{4.14}$$

Proof. By Theorem 2, π_γ is always positive, so the ratio

$$\nabla\pi_\gamma(\gamma)/\|\nabla\pi_\gamma(\gamma)\|$$

is equal to

$$\nabla\log\pi_\gamma(\gamma)/\|\nabla\log\pi_\gamma(\gamma)\|.$$

Thus if we multiply and divide by $\|\nabla\log\pi_\beta\big(h(\gamma)\big)\nabla h(\gamma)\|$, (4.13) is equal to

$$\limsup_{\|\gamma\|\to\infty}\frac{\gamma}{\|\gamma\|}\cdot\frac{\nabla\log\pi_\gamma(\gamma)}{\|\nabla\log\pi_\beta\big(h(\gamma)\big)\nabla h(\gamma)\|}\,\frac{\|\nabla\log\pi_\beta\big(h(\gamma)\big)\nabla h(\gamma)\|}{\|\nabla\log\pi_\gamma(\gamma)\|}$$

By Lemma 5, this right-hand fraction goes to one, so this lim sup is equal to

$$\limsup_{\|\gamma\|\to\infty}\frac{\gamma}{\|\gamma\|}\cdot\frac{\nabla\log\pi_\gamma(\gamma)}{\|\nabla\log\pi_\beta\big(h(\gamma)\big)\nabla h(\gamma)\|}.$$

If we expand $\nabla\log\pi_\gamma(\gamma)$ using (3.3), this lim sup is bounded above by the sum of

$$\limsup_{\|\gamma\|\to\infty}\frac{\gamma}{\|\gamma\|}\cdot\frac{\nabla\log\pi_\beta\big(h(\gamma)\big)\nabla h(\gamma)}{\|\nabla\log\pi_\beta\big(h(\gamma)\big)\nabla h(\gamma)\|}. \tag{4.15}$$

and

$$\limsup_{\|\gamma\|\to\infty}\frac{\gamma}{\|\gamma\|}\cdot\frac{\nabla\log\det\big(\nabla h(\gamma)\big)}{\|\nabla\log\pi_\beta\big(h(\gamma)\big)\nabla h(\gamma)\|}.$$

It follows from Lemmas 3 and 4 that this second lim sup is zero. Hence the lim sup

in (4.13) is bounded above by (4.15). Using the facts that $\nabla h(\gamma)$ is symmetric and $\nabla h(\gamma)\gamma = f'(\|\gamma\|)\gamma$ we can rewrite (4.15) as the lim sup in (4.14), which is the desired result. □

This next lemma does not seem very natural, but will be useful in later proofs.

Lemma 7. *Let $a(\gamma)$ and $b(\gamma)$ be functions such that both a and b are positive and bounded away from zero and infinity as $\|\gamma\|$ goes to infinity. Then for f as defined in* (4.1) *the fraction*

$$f'(\|\gamma\|)^2 \Big/ \left(\frac{f(\|\gamma\|)^2}{\|\gamma\|^2} a(\gamma) + \left[f'(\|\gamma\|)^2 - \frac{f(\|\gamma\|)^2}{\|\gamma\|^2} \right] b(\gamma) \right) \tag{4.16}$$

is positive and bounded away from zero as $\|\gamma\|$ goes to infinity.

Proof. The fraction in (4.16) is equal to the inverse of

$$\frac{f(\|\gamma\|)^2}{f'(\|\gamma\|)^2\|\gamma\|^2} a(\gamma) + \left[1 - \frac{f(\|\gamma\|)^2}{f'(\|\gamma\|)^2\|\gamma\|^2} \right] b(\gamma)$$

Since $a(\gamma)$ and $b(\gamma)$ are both positive and bounded away from zero and infinity for large $\|\gamma\|$, it is sufficient to show that

$$0 < \lim_{\|\gamma\| \to \infty} \frac{f(\|\gamma\|)^2}{f'(\|\gamma\|)^2\|\gamma\|^2} < 1. \tag{4.17}$$

Take $x = \|\gamma\|$, and recall the definitions of f and f' from (4.1) and (4.2). If follows that

$$\begin{aligned} \frac{f(\|\gamma\|)^2}{f'(\|\gamma\|)^2\|\gamma\|^2} &= \frac{(x^p + x)^2}{x^2(px^{p-1} + 1)^2} \\ &= \frac{(x^p + x)^2}{p^2x^{2p} + 2px^{p+1} + x^2} \end{aligned}$$

which is approximately equal to $x^{2p}/(p^2x^{2p})$ for large x, so the limit in (4.17) is p^{-2}, and since p is greater than 2 (4.17) holds. □

The first sufficient condition for the curvature of π_γ we will demonstrate is that if $\|\nabla \log \pi_\beta(\beta)\|$ is bounded above as $\|\beta\|$ goes to infinity, the induced density will satisfy the curvature condition in (2.12).

Theorem 3. *Suppose that β is an $\mathbb{R}^k$ valued random variable with exponentially light density π_β, and that $\|\nabla \log \pi_\beta(\beta)\|$ is bounded as $\|\beta\|$ goes to infinity. If $\beta = h(\gamma)$ for isotropic h defined using f from* (4.1)*, then the induced density, π_γ satisfies the curvature condition in* (2.12)*.*

Proof. By Lemma 6, it is only necessary to show that the lim sup in (4.14) is less than zero. This lim sup can be rearranged as

$$\limsup_{\|\gamma\|\to\infty} \frac{f'(\|\gamma\|)}{\|\nabla \log \pi_\beta\big(\nabla h(\gamma)\big)\nabla h(\gamma)\|} \frac{\gamma}{\|\gamma\|} \cdot \nabla \log \pi_\beta\big(\nabla h(\gamma)\big)$$

Since $\gamma/\|\gamma\|$ equals $h(\gamma)/\|h(\gamma)\|$ and π_β has exponentially light tails, there is a ε greater than zero such that this lim sup is bounded above by

$$\limsup_{\|\gamma\|\to\infty} -\frac{f'(\|\gamma\|)}{\|\nabla \log \pi_\beta\big(\nabla h(\gamma)\big)\nabla h(\gamma)\|}\varepsilon.$$

Since the fraction in this lim sup is non-negative — $f'(\|\gamma)\|)$ is greater than zero — showing that π_γ satisfies condition (2.12) only requires showing that the fraction

$$\frac{f'(\|\gamma\|)}{\|\nabla \log \pi_\beta\big(\nabla h(\gamma)\big)\nabla h(\gamma)\|}$$

is bounded away from zero for large $\|\gamma\|$. This is equivalent to the square of the same

fraction being bounded away from zero. This squared fraction can be written as

$$\frac{f'(\|\gamma\|)^2}{\nabla \log \pi_\beta\big(h(\gamma)\big)\nabla h(\gamma)\nabla h(\gamma)^T \nabla \log \pi_\beta\big(h(\gamma)\big)^T}. \tag{4.18}$$

Set

$$a(\gamma) = \|\nabla \log \pi_\beta\big(h(\gamma)\big)\|^2$$

and

$$b(\gamma) = \left(\frac{h(\gamma) \cdot \nabla \log \pi_\beta\big(h(\gamma)\big)}{\|h(\gamma)\|}\right)^2.$$

Then with the relationship in (3.7f), and the fact that $\gamma/\|\gamma\|$ equals $h(\gamma)/\|h(\gamma)\|$ the fraction in (4.18) is equal to the fraction in (4.16). The Cauchy-Schwarz inequality implies that

$$b(\gamma) \leq a(\gamma).$$

By assumption π_β is exponentially light so $b(\gamma)$ is bounded below. The Cauchy-Schwarz inequality bounds $b(\gamma)$ above by $a(\gamma)$, and $a(\gamma)$ is bounded above by assumption. Hence for some $\varepsilon > 0$ and $M > 0$

$$\varepsilon < b(\gamma) \leq a(\gamma) < M, \text{as } \|\gamma\| \to \infty.$$

By Lemma 7 the fraction in (4.18) is positive and bounded away from zero as $\|\gamma\|$ goes to infinity. □

Theorem 4. *Suppose that β is an $\mathbb{R}^k$ valued random variable with exponentially light density, π_β that satisfies the curvature condition in* (2.12)*. If $\beta = h(\gamma)$ for isotropic*

h defined using f from (4.1)*, then the induced density,* π_γ *satisfies the curvature condition in* (2.12)*.*

Proof. By Lemma 6, it is only necessary to show that the lim sup in (4.14) is less than zero. Using the fact that $\gamma/\|\gamma\| = h(\gamma)/\|h(\gamma)\|$, this lim sup can be rearranged as

$$\limsup_{\|\gamma\|\to\infty} \frac{\|\nabla \log \pi_\beta\big(h(\gamma)\big)\|\, f'(\|\gamma\|)}{\|\nabla \log \pi_\beta\big(h(\gamma)\big)\nabla h(\gamma)\|} \frac{h(\gamma)}{\|h(\gamma)\|} \cdot \frac{\nabla \log \pi_\beta\big(h(\gamma)\big)}{\|\nabla \log \pi_\beta\big(h(\gamma)\big)\|}$$

Because π_β satisfies the curvature condition in (2.12) there is a ε greater than zero such that this limsup is bounded above by

$$\limsup_{\|\gamma\|\to\infty} -\frac{\|\nabla \log \pi_\beta\big(h(\gamma)\big)\|\, f'(\|\gamma\|)}{\|\nabla \log \pi_\beta\big(h(\gamma)\big)\nabla h(\gamma)\|}\varepsilon \tag{4.19}$$

Showing that the fraction in this lim sup is positive and bounded away from zero for large $\|\gamma\|$ is sufficient to show that the lim sup is strictly less than zero. Because $f'(\|\gamma\|)$ is positive this fraction is non-negative, so it suffices to show that it is bounded away from zero. This is the same as the squared fraction being bounded away from zero. Set

$$a(\gamma) = \frac{\|\nabla \log \pi_\beta\big(h(\gamma)\big)\|^2}{\|\nabla \log \pi_\beta\big(h(\gamma)\big)\|^2} = 1$$

and

$$b(\gamma) = \left(\frac{\nabla \log \pi_\beta\big(h(\gamma)\big) \cdot h(\gamma)}{\|\nabla \log \pi_\beta\big(h(\gamma)\big)\|\, \|h(\gamma)\|}\right)^2.$$

Then the square of the fraction in (4.19) is equal to the fraction in (4.16), the $\|\nabla \log \pi_\beta\big(h(\gamma)\big)\|^2$ is moved to the denominator and is seen in $a(\gamma)$ and $b(\gamma)$. Since π_β is exponentially light, $\|\nabla \log \pi_\beta\big(h(\gamma)\big)\|$ is bounded away from zero for large $\|\gamma\|$

and we are not dividing by zero. The Cauchy-Schwarz inequality bounds $b(\gamma)$ above by one. Because π_β satisfies the curvature condition in (2.12), $b(\gamma)$ is bounded away from zero. Since $a(\gamma)$ and $b(\gamma)$ are bounded away from zero and infinity as $\|\gamma\|$ goes to infinity, by Lemma 7 the fraction in (4.19) is positive and bounded away from zero as $\|\gamma\|$ goes to infinity. □

4.2 Sub-Exponential to Exponentially Light

The goal is to use a variable transformation that will induce a super-exponential density, even though the original density was sub-exponential. Instead of directly inducing a super-exponential density, we will use a transformation that induces a exponentially light density, then use another transformation to induce a super-exponential density. E.g. β is a $\mathbb{R}^k$ valued random variable with sub-exponential density, π_β use h_1^{-1} such that $\gamma = h_1^{-1}(\beta)$ has exponentially light density π_γ. Then use h_2^{-1} to get $\eta = h_2^{-1}(\gamma)$ such that the induced density π_η is super-exponential. Thus the transformation $\eta = h_2^{-1}\big(h_1^{-1}(\beta)\big)$ will induce a density π_η that is super-exponential, even though the original density π_β is sub-exponential.

Section 4.1 already provides a suitable h_2^{-1}. This section provides h_1^{-1}.

4.2.1 Choice of f

We will use an isotropic h, setting $h(\gamma) = \beta$. Section 4.2.2 will show that a polynomial f (such as (4.1)) will not be strong enough. I.e. if π_β is not exponentially light, the induced density π_γ may not be exponentially light.

Let $b > 0$ and define

$$f_1(x) = \begin{cases} e^{bx} - \frac{e}{3} & \text{if } x > \frac{1}{b} \\ x^3 \frac{b^3 e}{6} + x \frac{be}{2} & \text{if } x \leq \frac{1}{b} \end{cases} \tag{4.20}$$

which has derivatives

$$f_1'(x) = \begin{cases} be^{bx} & \text{if } x > \frac{1}{b} \\ x^2 \frac{b^3 e}{2} + \frac{be}{2} & \text{if } x \le \frac{1}{b} \end{cases} \tag{4.21}$$

and

$$f_1''(x) = \begin{cases} b^2 e^{bx} & \text{if } x > \frac{1}{b} \\ x b^3 e & \text{if } x \le \frac{1}{b} \end{cases} \tag{4.22}$$

It is not hard to see that f_1, f_1' and f_1'' are all continuous at $1/b$; $f_1(0) = 0$; $f_1(x)$ goes to infinity as x does; $f_1'(0) = \frac{be}{2} > 0$; $f_1'(x)$ is positive for positive x; and $f''(0) = 0$. Hence f_1 satisfies conditions (3.6a)–(3.6g).

Lemma 8. *For isotropic h defined with f equal to f_1 from* (4.20)

$$\limsup_{\|\gamma\| \to \infty} \frac{\gamma}{\|\gamma\|} \cdot \nabla \log \det\bigl(\nabla h(\gamma)\bigr) = bk. \tag{4.23}$$

Proof. Since the lim sup is taken as $\|\gamma\|$ goes to infinity, we will only use the formulas for f_1 for large $\|\gamma\|$. Recall from (3.7b) that if $f = f_1$

$$\det\bigl(\nabla h(\gamma)\bigr) = f_1'(\|\gamma\|) \left(\frac{f_1(\|\gamma\|)}{\|\gamma\|} \right)^{k-1},$$

so

$$\nabla \log \det\bigl(\nabla h(\gamma)\bigr) = \frac{f_1''(\|\gamma\|)}{f_1'(\|\gamma\|)} \frac{\gamma^T}{\|\gamma\|} + (k-1) \left(\frac{f_1'(\|\gamma\|)}{f_1(\|\gamma\|)} \frac{\gamma^T}{\|\gamma\|} - \frac{\gamma^T}{\|\gamma\|^2} \right)$$

and the dot product in (4.23) is equal to (4.6). Clearly $1/\|\gamma\|$ goes to zero as $\|\gamma\|$

goes to infinity. Hence, taking $x = \|\gamma\|$ the lim sup in (8) is equal to

$$\lim_{x\to\infty}\left[\frac{f_1''(x)}{f_1'(x)} + (k-1)\frac{f_1'(x)}{f_1(x)}\right].$$

Rewriting using the definitions of f_1, f_1' and f_1'' for large x from (4.20), (4.21) and (4.22) this limit is equal to

$$\lim_{x\to\infty}\left[\frac{b^2e^{bx}}{be^{bx}} + (k-1)\frac{be^{bx}}{e^{bx}-e/3}\right]$$

which equals bk. □

4.2.2 Transformation Theorem

Let $\pi_{X,\alpha}$ denote a density on $\mathbb{R}^k$ that is always positive and has continuous first derivatives such that

$$\frac{\pi_{X,\alpha}(x)}{\|x\|^{-\alpha}} \to c \text{ as } \|x\| \to \infty \tag{4.24}$$

where c is some positive constant. Then for $\|x\|$ large enough,

$$\nabla \log \pi_{X,\alpha}(x) \approx -\alpha\frac{x^T}{\|x\|^2}. \tag{4.25}$$

It's clear that

$$\limsup_{\|x\|\to\infty} \frac{x}{\|x\|}\cdot\nabla\log\pi_{X,\alpha}(x) \;=\; \limsup_{\|x\|\to\infty} -\frac{\alpha}{\|x\|} = 0$$

so $\pi_{X,\alpha}$ is a sub-exponential density.

Let X be a random variable with a sub-exponential density, $\pi_{X,\alpha}$. And let $X = h(Y)$ for isotropic h. Before we show that h defined using $f = f_1$ from (4.20) will

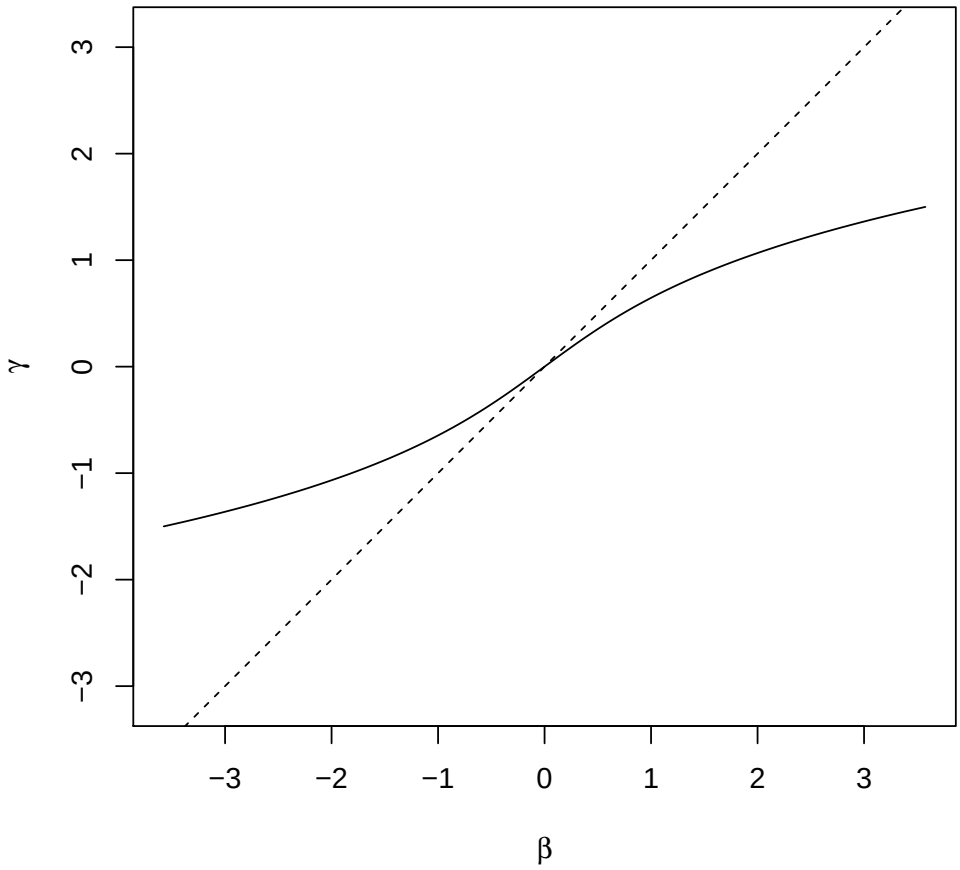


Figure 4.2: Solid line: $\beta = h(\gamma)$ for f_1 from(4.20) with $b = 1$. Dashed line: $\beta = \gamma$.

work, we will demonstrate that h defined using f from (4.1) may not be strong enough transformation to induce an exponentially light π_Y if $\pi_{X,\alpha}$ is sub-exponential. By Lemma 3, and the transformation density formula in (3.3)

$$\limsup_{\|y\|\to\infty} \frac{y}{\|y\|} \cdot \nabla \log \pi_Y(y) \;=\; \limsup_{\|y\|\to\infty} \frac{y}{\|y\|} \cdot \nabla \log \pi_{X,\alpha}\big(h(y)\big) \nabla h(y)$$

And because $\nabla h(y)$ is symmetric and $\nabla h(y) y = f'(\|y\|)y$. This right hand lim sup is equal to

$$\begin{aligned}
&\limsup_{\|x\|\to\infty} f'(\|x\|) \frac{h(x)}{\|h(x)\|} \cdot \nabla \log \pi_{X,\alpha}\big(h(x)\big) \\
&\qquad = \limsup_{\|x\|\to\infty} -\alpha f'(\|x\|) \frac{h(x)}{\|h(x)\|} \cdot \frac{h(x)}{\|h(x)\|^2} \\
&\qquad = \limsup_{\|x\|\to\infty} -\alpha \frac{f'(\|x\|)}{f(\|x\|)} \\
&\qquad = \limsup_{\|x\|\to\infty} -\alpha \frac{p\|x\|^{p-1} + 1}{\|x\|^p + \|x\|}
\end{aligned}$$

which is zero. Hence the induced density, π_Y will still be sub-exponential and not exponentially light.

Theorem 5. *Let β be a $\mathbb{R}^k$ valued random variable with sub-exponential density π_β, and there be some $\alpha > k$ and β_0 such that*

$$\frac{\beta}{\|\beta\|} \cdot \nabla \log \pi_\beta(\beta) \le -\frac{\alpha}{\|\beta\|}, \qquad \text{for } \|\beta\| > \|\beta_0\|. \tag{4.26}$$

If isotropic h is defined using f equal to f_1 from (4.20) and $\gamma = h^{-1}(\beta)$, then the induced density π_γ will be exponentially light.

Proof. By Lemma 1, π_γ is always positive with continuous first derivatives.

We need to show that (2.1) holds for π_γ. If we expand $\nabla \log \pi_\gamma(\gamma)$ using the

definition in (3.3), the lim sup in (2.1) is bounded above by the sum of

$$\limsup_{\|\gamma\|\to\infty} \frac{\gamma}{\|\gamma\|} \cdot \nabla \log \pi_\beta\big(h(\gamma)\big)\nabla h(\gamma) \tag{4.27}$$

and

$$\limsup_{\|\gamma\|\to\infty} \frac{\gamma}{\|\gamma\|} \cdot \nabla \log \det\big(h(\gamma)\big).$$

By Lemma 8 this second lim sup is equal to bk. Because $\nabla h(\gamma)$ is symmetric and $\nabla h(\gamma)\gamma = f_1'(\|\gamma\|)\gamma$ the lim sup in (4.27) is equal to

$$\limsup_{\|\gamma\|\to\infty} f_1'(\|\gamma\|)\frac{\gamma}{\|\gamma\|} \cdot \nabla \log \pi_\beta\big(h(\gamma)\big)$$

By assumption, and the fact that $\gamma/\|\gamma\| = h(\gamma)/\|h(\gamma)\|$, this lim sup is bounded above by

$$\limsup_{\|\gamma\|\to\infty} -\alpha\frac{f_1'(\|\gamma\|)}{\|h(\gamma)\|}$$

Recalling that $\|h(\gamma)\| = f_1(\|\gamma\|)$, and we can use the definitions of f_1 and f_1' for large γ from (4.20) and (4.21) and setting $y = \|\gamma\|$ this lim sup can be rewritten as

$$\limsup_{y\to\infty} -\alpha\frac{be^{by}}{e^{by} - e/3}$$

which is equal to $-b\alpha$. Therefore the lim sup in (2.1) is bounded above by $-b(\alpha - k)$ which is negative since α is greater than k, so π_γ is exponentially light. □

The key element of the proof of Theorem 5 was that the fraction $f'(y)/f(y)$ did not go to zero as y went to infinity. From this we can see that when dealing with

sub-exponential densities, using isotropic h the nature of

$$\lim_{y\to\infty} \frac{f'(y)}{f(y)}$$

will determine the result of the transformation. Suppose that π_β is a sub-exponential density that is not exponentially light. If this limit is zero then the induced density, π_γ will still not be exponentially light. If this limit is positive but not infinity, π_γ will be exponentially light. If this limit is equal to infinity, π_γ will be super-exponential.

The transformation based on f_1 is non-harmful in the sense that exponentially light or super-exponential densities, the induced density will be super-exponential. In a practical sense, the transformation based on f_1 may introduce numerical instabilities, as the floating point calculation e^x can overflow for relatively small values of x. E.g. e^{710} overflows. On the other hand, $e^{709} \approx 8.2e307$, or roughly 82 followed by 306 zeros; if the e^{709} is not far far out in the tails of the distribution of interest, the problem needs to be re-parametrized to be approachable by any computational method. However, it may be wise to make sure that the transformation is actually needed before using it.

4.2.3 Curvature Conditions

Theorem 6. *Let π_β be a sub-exponential density on $\mathbb{R}^k$, and there be an α_0 greater than k and $\beta_0 \in \mathbb{R}^k$ such that*

$$\|\nabla \log \pi_\beta(\beta)\| \leq \frac{\alpha_0}{\|\beta\|}, \qquad \textit{for } \|\beta\| > \|\beta_0\|. \tag{4.28}$$

If isotropic h is defined using f equal to f_1 from (4.20) and $\gamma = h^{-1}(\beta)$, then

$$\limsup_{\|\gamma\|\to\infty} \|\nabla \log \pi_\gamma(\gamma)\| < \infty. \tag{4.29}$$

Proof. By (3.3), the norm in (4.29) is equal to

$$\|\nabla \log \pi_\beta\big(h(\gamma)\big)\nabla h(\gamma) + \nabla \log \det\big(h(\gamma)\big)\|.$$

By the triangle inequality this is bounded above by the sum of

$$\|\nabla \log \pi_\beta\big(h(\gamma)\big)\nabla h(\gamma)\| \tag{4.30}$$

and

$$\|\nabla \log \det\big(h(\gamma)\big)\|. \tag{4.31}$$

If both (4.30) and (4.31) are bounded as $\|\gamma\|$ goes to infinity then (4.29) is true.

It follows from (4.5) that the square of (4.31) is equal to

$$\left[\frac{f_1''(\|\gamma\|)}{f_1'(\|\gamma\|)} + (k-1)\frac{f_1'(\|\gamma\|)}{f_1(\|\gamma\|)} - (k-1)\frac{1}{\|\gamma\|}\right]^2$$

Setting $y = \|\gamma\|$, and using the definitions of f_1, f_1' and f_1'' from (4.20), (4.21) and (4.22) this is equal to

$$\left[b + b(k-1)\frac{e^{by}}{e^{by} - e/3} - (k-1)\frac{1}{y}\right]^2$$

which clearly converges to $(bk)^2$ as y goes to infinity. So (4.31) is bounded as $\|\gamma\|$ goes to infinity.

Using the value of $\nabla h(\gamma)^2$ from (3.7f), and the fact that $\gamma/\|\gamma\|$ equals $h(\gamma)/\|h(\gamma)\|$ it follows that the square of (4.30) is equal to the sum of

$$\frac{f_1(\|\gamma\|)^2}{\|\gamma\|^2}\|\nabla \log \pi_\beta\big(h(\gamma)\big)\|^2 \tag{4.32}$$

and

$$f_1'(\|\gamma\|)^2 \left[1 - \frac{f_1(\|\gamma\|)^2}{\|\gamma\|^2 f_1'(\|\gamma\|)^2}\right] \left(\frac{h(\gamma) \cdot \nabla \log \pi_\beta\big(h(\gamma)\big)}{\|h(\gamma)\|}\right)^2 \tag{4.33}$$

Since for large x, $f_1(x) = e^{bx} - e/3$ and $f_1'(x) = be^{bx}$, $\|\gamma\|^2 f_1'(\|\gamma\|)^2$ is greater than $f_1(\|\gamma\|)^2$ for large $\|\gamma\|$. Then for large $\|\gamma\|$ the term

$$1 - \frac{f_1(\|\gamma\|)^2}{\|\gamma\|^2 f_1'(\|\gamma\|)^2}$$

is strictly between zero and one. Hence the middle term of (4.33) is positive for large $\|\gamma\|$. The other two terms in (4.33) are squares, so (4.33) is non-negative. Since (4.33) is non-negative, applying the Cauchy-Schwarz inequality to the right-most term, and bounding the middle term above by one yields the upper bound

$$f_1'(\|\gamma\|)^2 \, \|\nabla \log \pi_\beta\big(h(\gamma)\big)\|^2. \tag{4.34}$$

By the assumptions of this Theorem, for $\|h(\gamma)\| > \|\beta_0\|$,

$$\|\nabla \log \pi_\beta\big(h(\gamma)\big)\|^2 \leq \frac{\alpha_0^2}{\|h(\gamma)\|^2}$$

and since $\|h(\gamma)\| = f(\|\gamma\|)$ this upper bound is equal to

$$\alpha_0^2 \frac{1}{f_1(\|\gamma\|)^2}.$$

Combining this upper bound with (4.32) and (4.34) lets us bound $\|\nabla \log \pi_\gamma(\gamma)\|$ above by

$$\alpha_0^2 \frac{f_1(\|\gamma\|)^2}{f_1(\|\gamma\|)^2 \|\gamma\|^2} + \alpha_0^2 \frac{f_1'(\|\gamma\|)^2}{f_1(\|\gamma\|)^2}$$

which converges to α_0^2 as $\|\gamma\|$ goes to infinity, so (4.30) is bounded as $\|\gamma\|$ goes to infinity. □

Theorem 6 will apply to well behaved sub-exponential densities that are sub-exponential in every direction. If π is a sub-exponential but not exponentially light density, but there is some direction δ — taking $\delta^T\delta = 1$ — such that

$$\limsup_{s\to\infty} \frac{s\delta}{\|s\delta\|} \cdot \nabla \log \pi(s\delta) < 0$$

then Theorem 6 will not apply as the lower bound in (4.28) will not hold as $\|s\delta\|$ goes to infinity.

4.3 Discussion

Based on the results in Section 4.1 and Section 4.2 we can provide the following two theorems.

Theorem 7. *Let β be an $\mathbb{R}^k$ valued random variable with exponentially light density, π_β that satisfies one of the following two conditions is true*

(i) π_β satisfies the curvature condition (2.12)*.*

(ii) $\|\nabla \log \pi_\beta(\beta)\|$ is bounded as $\|\beta\|$ goes to infinity.

Then if $\gamma = h^{-1}(\beta)$ for isotropic h defined using f from (4.1)*, then a random-walk Metropolis algorithm for π_γ using proposal density, q that satisfies* (2.11) *will be geometrically ergodic.*

Proof. Apply Theorem 2, one of Theorems 4 or 3 and Theorem 1. □

The corresponding theorem for sub-exponential densities is:

Theorem 8. *Let β be an $\mathbb{R}^k$ valued random variable with sub-exponential density, π_β that satisfies conditions* (4.28) *and* (4.26) *for $\alpha_0 > \alpha > k$. If $\eta = h_2^{-1}(\beta)$ for isotropic h_2 defined using f_1 from* (4.20)*, and $\gamma = h_1^{-1}(\eta)$ for isotropic h_1 defined using f from* (4.1)*, then a random-walk Metropolis algorithm for π_γ using proposal density, q that satisfies* (2.11) *will be geometrically ergodic.*

Proof. By Theorems 5 and 6 π_η will satisfy the conditions of Theorem 7. □

These theorems state that the random-walk Metropolis generated Markov chain for π_γ is geometrically ergodic. The following theorem states that if h is used to transform the Markov chain for π_γ is transformed to a Markov chain for π_β, the Markov chain for π_β will converge at the same rate. Hence, since the Markov chain for π_γ is geometrically ergodic, the Markov chain for π_β is geometrically ergodic.

Theorem 9. *Let X be a random variable on the probability space $(\mathcal{X}, \mathcal{A}_x, \pi_X)$ and Y be the induced random variable on the probaility space $(\mathcal{Y}, A_Y, \pi_Y)$ by setting $Y = \varphi(X)$ for isomorphic[1] and measureable $\varphi : \mathcal{X} \to \mathcal{Y}$. Also let $\Phi = \{\Phi_0, \Phi_1, \dots\}$ be a Markov chain with transition probability measure P_X; and $\Phi_Y = \{\varphi(\Phi_0), \varphi(\Phi_1), \dots\}$ represent the Markov chain Φ transformed by φ, with corresponding transition probability measure P_Y. If P_X preserves π_X then P_Y preserves π_Y and*

$$\|\pi_Y(\cdot) - P_Y^n(y, \cdot)\|$$

converges to zero at the same rate as

$$\|\pi_X(\cdot) - P_X^n(\varphi^{-1}(y), \cdot)\|$$

Proof. Let $\varphi^{-1}(A)$ denote the inverse image of A under ϕ. Note that since $Y = \varphi(X)$

[1] $\varphi : \mathcal{X} \to \mathcal{Y}$ is an isomorphism if for all $y \in \mathcal{Y}$ and $x \in \mathcal{X}$, $\varphi(\varphi^{-1}(y)) = y$ and $\varphi^{-1}(\varphi(x)) = x$.

and φ is a measurable isomorphism, $A \in A_Y$, $\pi_Y(A) = \pi_X\big(\varphi^{-1}(A)\big)$ and $P_Y^n(y, A) = P_X^n\big(\varphi^{-1}(y), \varphi^{-1}(A)\big)$.

Since P_X preserves π_X, it follows that for all $A \in A_Y$

$$\int_{\mathcal{Y}} \pi_Y(dy) P_y(y, A) = \int_{\mathcal{X}} \pi_X(dx) P_X\big(x, \varphi^{-1}(A)\big) = \pi_X\big(\varphi^{-1}(A)\big) = \pi_Y(A),$$

so P_Y preserves π_Y.

Now consider $\|\pi_Y(\cdot) - P_Y^n(y, \cdot)\|$. Rewriting using the definition of the total variation norm (1.3) we can see that for any $y \in \mathcal{Y}$

$$\begin{aligned}
\|\pi_Y(\cdot) - P_Y^n(y, \cdot)\| &= \sup_{A \in A_Y} |\pi_Y(A) - P_Y^n(y, A)| - \inf_{A \in A_Y} |\pi_Y(A) - P_Y^n(y, A)| \\
&= \sup_{A \in A_Y} |\pi_X\big(\varphi^{-1}(A)\big) - P_X^n\big(\varphi^{-1}(y), \varphi^{-1}(A)\big)| \\
&\quad - \inf_{A \in A_Y} |\pi_X\big(\varphi^{-1}(A)\big) - P_X^n\big(\varphi^{-1}(y), \varphi^{-1}(A)\big)| \\
&= \sup_{A' \in A_X} |\pi_X(A') - P_X^n\big(\varphi^{-1}(y), A'\big)| \\
&\quad - \inf_{A' \in A_X} |\pi_X(A') - P_X^n\big(\varphi^{-1}(y), A'\big)| \\
&= \|\pi_X(\cdot) - P_X^n\big(\varphi^{-1}(y), \cdot\big)\|.
\end{aligned}$$

We can switch the sup and inf from being across $A \in A_Y$ to being across $A' \in A_X$ because φ is an isomorphism. It follows that the total variation distance $\|\pi_Y(\cdot) - P_Y^n(y, \cdot)\|$ goes to zero at the same rate as $\|\pi_X(\cdot) - P_X^n\big(\varphi^{-1}(y), \cdot\big)\|$. □

By the law of the unconscious statistician, in Theorems 7 and 8 the Markov chains generated from π_γ can be used to draw inference on π_β, by transforming the chain back to the β scale using h. Theorem 9 states that the Markov chain for π_β will be geometrically ergodic because the Markov chain for π_γ is. Furthermore, the transformations in these theorems do not seem like they will do any harm by either

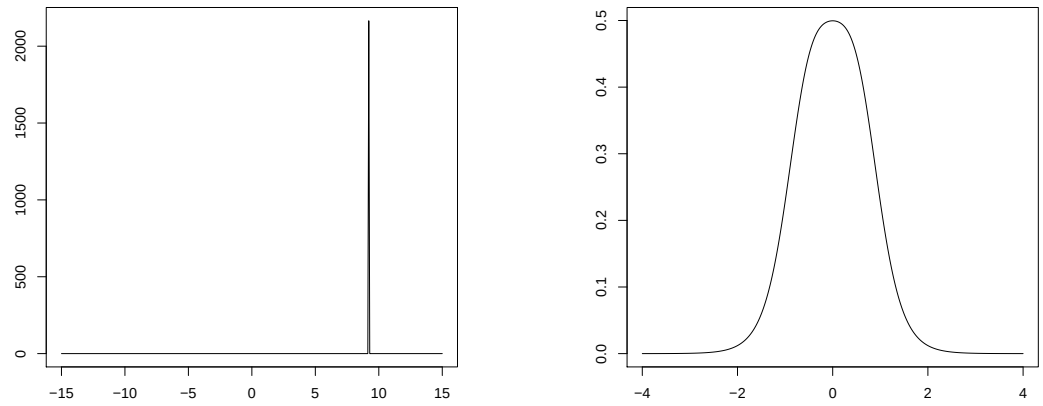
increasing the tail-weight of a density or ruining curvature.

In both the sub-exponential to exponentially light, and exponentially-light to super exponential cases, the transformation induces a density with lighter tails because

$$\frac{\partial\|\beta\|}{\partial\|\gamma\|} \to \infty, \qquad \text{as } \|\gamma\| \to \infty.$$

That is, as γ goes away from the origin, the rate at which β goes away from the origin increases. It's important that the transformations do this smoothly, so as to not ruin the curvature of the original density. Different rates are needed for the two cases, f_1 would also work for the exponentially light to super-exponential transformation but it's stronger than is needed.

If the mode of the original density is far from zero, the "width" of the mode will be decreased for the induced density, as the interval $(\beta, \beta + 1)$ corresponds to smaller and smaller intervals on the γ scale as $\|\beta\|$ increases. Figure 4.3 illustrates this point. Isotropic h with the exponential f_1 from (4.20) is used. Figure 4.3a is the induced density from transforming a $T_3 + 10000$ random variable, Figure 4.3b is the induced density from transforming a T_3 random variable where T_3 denotes a T random variable with 3 degrees of freedom. The mode of the density in Figure 4.3a is very narrow and tall, the mode of the density in Figure 4.3b is not. Both of these densities are exponentially light, so a random-walk Metropolis algorithm for each of the induced densities will be geometrically ergodic (Mengersen and Tweedie, 1996). Even though a random-walk Metropolis algorithm is geometrically ergodic in each case, there will be better performance for the density in Figure 4.3b. Since the induced mode is further from zero $f(\|\gamma\|)/\|\gamma\|$ is larger, so small changes in γ will correspond to large changes in β. There could be numerical stability issues. The risk of numerical instability is hard to quantify, as it will depend on how far the mode of the original density is away from zero as well as what estimates are being calculated.



(a) Induced density, original mode of 10000.

(b) Induced density, original mode of 0.

Figure 4.3: Two different induced densities T_3 density, original modes of 10000 and 0.

E.g. if we are calculating $\mathrm{E}_{\pi_\beta} g(\beta)$ for some g, any numerical instability will depend on g and how far the mode is from zero.

So re-centering the target distribution at the mode is advisable, if the mode can be calculated. Let $\tilde{\beta}$ be the mode of π_β, and set $\beta^* = \beta - \tilde{\beta}$. It's easy to see that the lim sups for the curvature condition (2.12) or the tail-weight (as in (2.2)) will have the same value for π_β and π_{β^*}. Hence re-centering a distribution at the mode will not have any ill-effects, and may reduce the risk of numerical instability.

Unfortunately, this transformation method may introduce multi-modality in the target distribution. Consider the univariate standard-normal density, f_Z. f_Z is already super-exponential so there would be no need to use the transformation, but this is just an example. Similar behavior would be seen with a T-density. We know that it has a unique mode at 0, and that

$$f_Z'(z) = -z f_Z(z)$$

Let $Z = h(X)$ for isotropic h defined using f from (4.1) with $p = 3$. For $x \neq 0$ the derivative of f_X is given by

$$f'_X(x) = f_Z\big(f(|x|)\frac{x}{|x|}\big)\Big[\text{sign}(x)f''(|x|) - f'(|x|)^2 f(|x|)\Big]$$

Since f_Z is always positive, this is zero if and only if

$$\text{sign}(x)f''(|x|) - f'(|x|)^2 f(|x|)$$

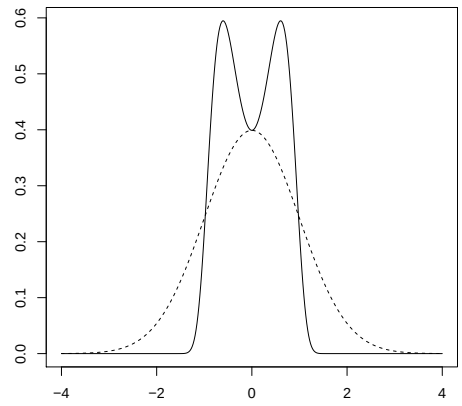
is zero. For our chosen f, this has three solutions, which leads to dual mode seen in Figure 4.4a. If we replace this f with

$$g(x) = \begin{cases} x & \text{if } x < R \\ (x-R)^3 + x & \text{if } x \geq R \end{cases}$$

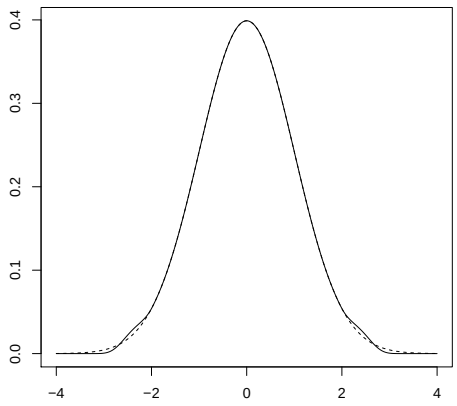
where $R \geq 0$, then for large enough R ($R = 2$ in this case) the induced density will only have one mode, as seen in Figure 4.4b. g satisfies conditions (3.6a)–(3.6g) so h is still an isotropy. There is no general way around the possibility of multi-modality, but it can likely be avoided if π_β is known. However, this multi-modality is not likely to be an issue, as the modes are likely to be very close to each other, or to be small modes out in the tails which will not slow convergence very much.

4.3.1 Implementation

One benefit of this transformation method is the ease of implementation. If there is an existing implementation of the random-walk Metropolis algorithm, the variable transformation can be done on top of it. To demonstrate this, we provide a complete *R* (R Development Core Team, 2011) implementation here. The `metrop` function from the `mcmc` *R*-package (Geyer, 2010) implements a random-walk Metropolis algorithm



(a) Solid line: f_X induced from f_Z using isotropic h defined with f from (4.1) with $p = 3$. Dashed line: f_Z.



(b) Solid line: f_X induced from f_Z using isotropic h defined with g, $R = 2$. Dashed line: f_Z.

Figure 4.4: Two different induced densities from transforming a standardnormal random variable.

with a $N(0, \sigma^2 \, \mathrm{I}_k)$ proposal density, σ^2 defaults to 1, but the caller can specify any positive value for σ^2. For any $\sigma^2 > 0$, this density satisfies condition (2.11), hence if the target density is super-exponential then the Markov chain will be geometrically ergodic by Theorem 1. The `metrop` function expects the log-density, which need not be normalized, so we will implement the log-density transformation.

Listing 4.1 gives R implementations of f from (4.1) using $p = 3$ and f_1 from (4.20) using $b = 1$. Listing 4.2 gives the corresponding implementations of f^{-1} and f_1^{-1}. Listing 4.3 implements isotropic h and $\log \det \nabla h$ for any f. And Listing 4.4 implements $\log \pi_\gamma$ for and f. Using this code, isotropic h defined using f from (4.1) with $p = 3$ is simply `h(f)`. For a vector `gamma`, the corresponding `beta` value could be found with the command:

```
h(f)(gamma)
```

By equation (3.8) h^{-1} is an isotropic function defined with f^{-1}. So we can find $h^{-1}(\beta)$ with the code

```
h(f.inv)(beta)
```

To use isotropic h with f_1 from (4.20), just replace `f` with `f1` in the preceding examples.

If the `ludens` implements the log-unnormalized density for π_β satisfying the conditions of Theorem 7, the transformation method can be used with `metrop` with the code:

```
out <- metrop(ludens.trans(ludens, f, df),
              initial=h(f.inv)(init.state), nbatch=10^4, outfun=h(f))
```

The Markov chain would be started at initial state $\gamma_0 = h^{-1}(\beta_0)$, where β_0 (represented by `init.state`) is the initial state on the β scale, and would be 10^4 steps long. `out$batch` is a $10^4 \times k$ matrix, each row is one of the states of the Markov chain on the original β scale. E.g.

```
out$batch[131,]
```

gives the value of β_{131} — vectors and matrices in R use one-based indexing and `out$batch` does not contain β_0. If π_β satisfies the conditions of Theorem 8, and `ludens` is an implementation of the log-unnormalized density for π_β, then the transformation method can be used with the code:

```
out <- metrop(ludens.trans( ludens.trans(ludens, f1, df1), f, df),
              initial=h(f.inv)( h(f1.inv)(init.state) ),
              nbatch=10^4,
              outfun=function(x) h(f1)( h(f)(x) ))
```

Code Listing 4.1: R implementations of f and f_1 using $p = 3$ and $b = 1$.

```
f <- function(x)
  x^3 + x

df <- function(x)
  3*x^2 + 1

f1 <- function(x)
  ifelse(x > 1,
         exp(x) - exp(1)/3,
         x^3 * exp(1)/6 + x * exp(1) / 2)

df1 <- function(x)
  ifelse(x > 1,
         exp(x),
         x^2 * exp(1)/2 + exp(1) / 2)
```

The Markov chain would be started at initial state $\gamma_0 = h_1^{-1}\big(h_2^{-1}(\beta_0)\big)$ and would be 10^4 steps long. Again, the contents of `out$batch` will be on the original β scale.

With either transformation, the user only needs to think about the starting value and the output on the original scale. If some function, `g` of the state was desired, only the `outfun` argument would need to be modified. E.g. replacing `h(f)` with `function(x) g(h(f)(x))` in the first example.

4.3.2 Different h

Our choice of h is certainly not the only one that will work. Earlier versions of this theory used non-isotropic h, define as component-wise transformations (Johnson and Geyer, 2010). These other transformations worked but did not have the property, that for some function $g : \mathbb{R}^k \to \mathbb{R}$

$$\nabla h(\gamma)\gamma = g(\gamma)\gamma$$

Code Listing 4.2: R implementations of f^{-1} and f_1^{-1} using $p = 3$ and $b = 1$.

```
f.inv <- function(x) {
  n <- (sqrt(3*(27*x^2+4)) - 9 * x)^(1/3)
  (2/3)^(1/3) / n - n / ( 2^(1/3) * 3^(2/3) )
}

f1.inv <- function(x) {
  poly.inv <- exp(1/3) * (sqrt(9*x^2+exp(2)) - 3 * x)^(-1/3)
  poly.inv <- poly.inv - 1/poly.inv
  ifelse(x > 2/3,
         log(x+exp(1)/3),
         poly.inv)
}
```

Code Listing 4.3: R implementations of h and $\log\det\nabla h$

```
h <- function(f) {
  function(x) {
    norm.x <- sqrt( sum(x*x) )
    if (norm.x == 0)
      x
    else
      f(norm.x) * x / norm.x
  }
}

log.det.dee.h <- function(f, df) {
  function(x) {
    norm.x <- sqrt( sum(x*x) )
    ifelse(norm.x == 0,
           length(x) * log(df(norm.x)),
           log(df(norm.x)) +
             (length(x) - 1)*(log(f(norm.x)) - log(norm.x)))
  }
}
```

Code Listing 4.4: R implementation of $\log \pi_\gamma$

```
ludens.trans <- function(ludens, f, df) {
  function(x)
    ludens( h(f)(x) ) + log.det.dee.h(f,df)(x)
}
```

held for all γ. Consequently, the proofs were more tedious than the proofs here.

For isotropic h, there are other possible fs than our f in (4.1) for an exponentially light to super-exponential transformation, and our f_1 in (4.20) for a sub-exponential to exponentially light transformation. We chose these particular f and f_1 as they are simple to calculate, satisfy the necessary conditions (3.6a)–(3.6g), and seem to be just-barely strong enough.

The transformations in Theorem 2 and Theorem 5 will not make densities "heavier". Theorem 2 only assumes that the original density is exponentially light, but does not preclude starting with a super-exponential density. Likewise, Theorem 5 only assumes that the original density is sub-exponential, but does not preclude starting with a exponentially light density.

I.e. suppose that π_β is an exponentially light density, and $\{\gamma_k\}_{k\geq 1}$ and $\{\eta_k\}_{k\geq 1}$ are both sequences with norms going to infinity such that

$$-\infty < \lim_{k\to\infty} \frac{\gamma_k \cdot \nabla \log \pi_\beta(\gamma_k)}{\|\gamma_k\|} < 0$$

and

$$\lim_{k\to\infty} \frac{\eta_k \cdot \nabla \log \pi_\beta(\eta_k)}{\|\eta_k\|} = -\infty.$$

By Theorem 2, the transformation from Theorem 2 will induce a super-exponential π_γ, but the transformation will be excessive in the case of the sequence $\{\eta_k\}_{k\geq 1}$. As

long as the transformation used is not too extreme, this should be non-harmful and safe to use. One can envision real-world situations where this conditions might arise. E.g. if we use $\mathbf{1}_i$ and $\mathbf{0}_j$ to represent the vectors of 1 repeated i times and 0 repeated j times respectively, we can then consider the random variable

$$X = \begin{pmatrix} U \\ V \end{pmatrix} \in \mathbb{R}^{p+q}$$

where U in $\mathbb{R}^p$ is a multivariate normal variable and V in $\mathbb{R}^q$ is the regression parameter for a multinomial logit regression model will have a density, f_X that is super-exponentially light in the direction $(\mathbf{1}_p, \mathbf{0}_q)^T$, but only exponentially light in the direction $(\mathbf{0}_p, \mathbf{1}_q)^T$. It's fair to consider the random variable in this vector order because of we can rewrite β as $\beta' = P_l\beta$ for some permutation matrix P_l, which will induce a density with the same curvature and tail weight[2]. One can envision using a the transformed variable,

$$Y = h'(X) = \Big(X_1, \ldots, X_p, h\big((X_{p+1}, \ldots, X_{p+q})\big)\Big)^T$$

for some appropriate — possibly isotropic – h. If h is properly chosen, it seems clear that we can induce a super-exponential f_Y, even though h' is not isotropic and would not work for Theorem 2. Theory could be developed along these lines, but it doesn't seem necessary at this time.

It seems plausible that a more exotic h without such a simple expression for $\det\big(\nabla h(\gamma)\big)$ may be desired. We only considered hs where $\det\big(\nabla h(\gamma)\big)$ was a product of k elements. Working with $\nabla\big[\det\big(\nabla h(\gamma)\big)\big]$ is painful for general ∇h, but there is

[2]Since P_l is a $k \times k$ permutation matrix, $P_l^T P_l = \mathrm{I}_k$ and $|\det(P_l)| = 1$. Applying the density and log-density transformation formulas in (3.1) and (3.2) will show that the lim sups in (2.12) and (2.2) are identical for π_β and $\pi_{\beta'}$ where $\beta' = P_l\beta$. Hence we can consider a vector in any order without changing the tail-curvature or tail-weight of the density.

a closed form solution for this expression. Computation is the real issue, calculating a determinant is computationally expensive. There is no determinant calculation algorithm with less than $o(k^2)$ complexity for a general $k \times k$ matrix. In the random-walk Metropolis algorithm for a target density π_γ, each step calculates the value of

$$\frac{\pi_\gamma(Y_{i+1})}{\pi_\gamma(X_i)}$$

where Y_{i+1} is the proposed state. Even if the value of $\pi(X_i)$ is stored from previous steps, the determinant of a $\nabla h(Y_{i+1})$ will be calculated at each step. Having a k on the order of a few thousand is a real possibility. So we avoided using a h without a simple form for $\det\big(\nabla h(\gamma)\big)$.

Chapter 5

Example Applications

5.1 Multinomial Logit Regression

Let $Y = (Y_1, \ldots, Y_m)^T$ be a vector of counts that follows a multinomial distribution with m categories, probabilities p, and count n. It is assumed that n is an integer strictly greater than one, and that

$$\sum_{i=1}^{m} p_i = 1$$

where each element of p is non-negative. Then Y has pmf

$$\mathrm{P}\big(Y = y|p, n\big) = \frac{n!}{y_1! \cdots y_m!} p_1^{y_1} \cdots p_m^{y_m}. \tag{5.1}$$

Equation 5.1 gives rise to the canonical parameter $\theta = (\theta_1, \ldots, \theta_m)^T$ where

$$\theta_j = \log(p_j) + c.$$

With the relation $e^c e^{\theta_j} = p_j$ we can rewrite (5.1) in canonical form

$$\mathrm{P}\big(Y = y \mid n, p\big) = \frac{n!}{\prod_j y_j!} e^{\langle y, \theta \rangle - nC(\theta)} \tag{5.2}$$

where

$$C(\theta) = \log\Big(\sum_i e^{\theta_i}\Big).$$

Instead of θ we are interested in some other parameter β, where θ is some additive combination of the elements of β. We express this relationship using a model matrix, M. Then we can express the relationship between θ and β

$$\theta = M\beta.$$

The model matrix, M, is a known $m \times k$ matrix — one row for each category, one column for each element of β.

For example, the elements of β (and hence columns of M) may correspond to one-way or two-way interactions between categorical variables or measurements of a quantitative variable.

In this report we will use $M_{j\cdot}$ to refer to the j^{th} row of M. Hence we have the relationship

$$\theta_j = M_{j\cdot}\beta.$$

We can rewrite (5.2) as

$$\mathrm{P}\big(Y = y|\beta, n\big) = \frac{n!}{\prod_j y_j!} e^{\langle y, M\beta\rangle - n\log\left(\sum_j e^{M_{j\cdot}\beta}\right)}. \tag{5.3}$$

This pmf arises from both contingency tables and multinomial logit regression. In contingency tables, there is some collection of categorical variables $C_1, \ldots, C_c$ and the observed vector Y is the count of the observations at each combination of the categorical variables. This has *exactly* the pmf in (5.3).

In multinomial logit regression Y is observed L times. These different observations will be denoted $Y^1, \ldots, Y^L$. Each Y^l has its own model matrix, M^l; count, n^l; and is assumed to be independent of the other observations. The general form of (5.3) arises from these conditions

$$\mathrm{P}\big(Y^1 = y^1; \cdots; Y^L = y^L | \beta, n^1, \cdots n^L\big) = \left[\prod_{l=1}^{L} \frac{n^l!}{\prod_j y_j^l!}\right] \exp\left[\sum_{l=1}^{L} \left\{ \langle y^l, M^l\beta\rangle - n^l \log\left(\sum_j e^{M^l_{j\cdot}\beta}\right)\right\}\right]. \quad (5.4)$$

Anything proved about (5.4) will also apply to (5.3), just set $L = 1$.

To simplify notation, we will let

$$f_Y(y|\beta, n) = \exp\left\{\sum_{l=1}^{L}\big\{\langle y^l, M^l\beta\rangle - n^l \log\big(\sum_j e^{M^l_{j\cdot}\beta}\big)\big\}\right\}. \quad (5.5)$$

This uses Y to represent $Y^1, \ldots, Y^L$; y to represent $y^1, \ldots, y^L$; and n to represent $n^1, \ldots, n^L$. This should be forgiven, enforcing the correct notation will only make the equations more complicated visually without adding any understanding.

The model is going to be applied in a Bayesian analysis. Section 5.1.1 will consider a normal prior for β; Section 5.1.2 will consider a conjugate prior for β. The goal of these two sections is the same, to find a geometrically ergodic Markov chain for the posterior density

$$\pi(\beta|Y, n) \propto f_Y(y|\beta, n)\pi(\beta) \quad (5.6)$$

where $\pi(\beta)$ is the prior on β.

The following Lemma is needed in both Section 5.1.1 and Section 5.1.2.

Lemma 9. *For counts y and total counts n. Using ∇_β to represent the derivative*

operator with respect to β, then

$$\|\nabla_\beta \log f_Y(y|\beta, n)\| \tag{5.7}$$

is bounded above as $\|\beta\|$ goes to infinity.

Proof. First note that

$$\nabla_\beta \log f_Y(y|\beta, n) = \sum_{l=1}^{L} \left\{ (M^l)^T y^l - n^l \sum_j (M^l_{j\cdot})^T \frac{e^{M^l_{j\cdot}\beta}}{\sum_i e^{M^l_{i\cdot}\beta}} \right\}.$$

Note that the fraction in this equation is between zero and one for all β. Since y^l, n^l, and M^l are all fixed, it follows that (5.7) is bounded. □

5.1.1 Multinomial Logit Normal Mixed Model

Let β have a multivariate normal prior, with positive-definite covariance matrix Σ. Then

$$\pi(\beta) \propto \exp\{-\beta^T \Sigma^{-1} \beta/2\}. \tag{5.8}$$

With the multivariate normal prior, $\pi(\beta|Y, n, \Sigma)$ is a multivariate generalization of the logit-normal model considered in Christensen et al. (2001) and Johnson et al. (2009). As in these papers, the proof strategy is to show that the density satisfies the conditions for Theorem 1 from Jarner and Hansen (2000), so a random-walk Metropolis algorithm (with appropriate proposal density) will be geometrically ergodic.

Theorem 10. *Let $\pi(\beta)$ be a multivariate normal density with positive-definite covariance matrix Σ as in* (5.8)*. Then the random-walk Metropolis algorithm for the posterior density* (5.6) *will be geometrically ergodic if the proposal density q satisfies condition* (2.11)*.*

Proof. We need to show that $\pi(\beta|y, n, \Sigma)$ is super-exponential and satisfies the curvature condition (2.12).

First we will show that the posterior in (5.6) is super-exponential. Since Σ is positive-definite, the density in (5.8) is super-exponential. Note that

$$\nabla \log \pi(\beta|y, n, \Sigma) = \nabla \log f_Y(y|\beta, n) + \nabla \log \pi(\beta). \tag{5.9}$$

Consequently to show that the posterior density is super-exponential, we only need to show that f_Y is always positive, has continuous first derivatives and that

$$\frac{\beta}{\|\beta\|} \cdot \nabla \log f_Y(y|\beta, n)$$

is bounded above by some real number as $\|\beta\|$ goes to infinity. Lemma 9 states that $\|\nabla \log f_Y(y|\beta, n)\|$ is bounded as $\|\beta\|$ goes to infinity, so this dot-product is bounded above. The form of f_Y given in (5.5) is a composition of functions with continuous first derivatives, so f_Y has continuous first derivatives. Note that $f_Y = \exp \circ\, g$ for some g. Since $\exp\{x\}$ is strictly positive, f_Y is always positive. Therefore $\pi(\beta|y, n, \Sigma)$ (5.6) is super-exponential.

Set

$$\begin{aligned}
n_1(\beta) &= \frac{\beta}{\|\beta\|} \cdot \frac{\nabla \log \pi(\beta)}{\|\nabla \log \pi(\beta)\|} \\
n_2(\beta) &= \frac{\|\nabla \log \pi(\beta)\|}{\|\nabla \log f_Y(y|\beta, n) + \nabla \log \pi(\beta)\|} \\
n_3(\beta) &= \frac{\beta}{\|\beta\|} \cdot \frac{\nabla \log f_Y(y|\beta, n)}{\|\nabla \log f_Y(y|\beta, n) + \nabla \log \pi(\beta)\|}
\end{aligned}$$

Then using the relationship in (5.9), the curvature condition (2.12) for the posterior

in (5.6) can be written

$$\limsup_{\|\beta\|\to\infty} n_1(\beta)n_2(\beta) + n_3(\beta) < 0.$$

Since $\pi(\beta)$ is a super-exponential density and $\|\nabla \log f_Y(y|\beta, n)\|$ is bounded (from Lemma 9) it follows that $n_3(\beta)$ goes to zero and $n_2(\beta)$ goes to one as $\|\beta\|$ goes to infinity. Therefore, this condition is the same as

$$\limsup_{\|\beta\|\to\infty} \frac{\beta}{\|\beta\|} \frac{\nabla \log \pi(\beta)}{\|\nabla \log \pi(\beta)\|} < 0.$$

This follows from Jarner and Hansen (2000)[Theorem 4.6]. Therefore (5.6) satisfies the curvature condition (2.12). □

Theorem 10 shows that a random-walk Metropolis algorithm for this posterior density is geometrically ergodic. The variable transformation is not needed in this case.

5.1.2 Multinomial Logit with Conjugate Prior

A conjugate prior consists of adding prior counts to each response cell. It is easier to specify these prior counts as prior probabilities and a single prior count. For each l in $1, \ldots, L$, we will use ξ^l for the prior probabilities, and ν^l as the prior count. Then the posterior density has the form

$$\pi(\beta|y, n, \xi, \nu) \propto \exp\Big\{ \sum_{l=1}^{L} \langle y^l + \xi^l \nu^l, M^l \beta \rangle - (n^l + \nu^l) \log\big(\sum_j e^{M_{j\cdot}\beta}\big) \Big\}. \qquad (5.10)$$

In the rest of this section, y^l will never be separated from $\xi^l\nu^l$, and n^l will never be separated from ν^l, so we will simplify (5.10) to

$$\pi(\beta|y,n) \propto \exp\Big\{\sum_{l=1}^{L}\langle y^l, M^l\beta\rangle - n^l\log\big(\sum_j e^{M_{j\cdot}\beta}\big)\Big\}. \tag{5.11}$$

Whenever we see y^l or n^l we can remember that it could be replaced with $y^l + \xi^l\nu^l$ or $n^l + \nu^l$ respectively.

The density in (5.11) is not super-exponential. This density is exactly equal to f_Y from (5.5). Hence by Lemma 9

$$\left|\frac{\beta}{\|\beta\|}\cdot\nabla\log\pi(\beta|y,n)\right|$$

is bounded as $\|\beta\|$ goes to infinity, so $\pi(\beta|y,n)$ can't be super-exponential.

Since $\pi(\beta|y,n)$ is not super-exponential, it may be a candidate for the transformation method. We need to show that $\pi(\beta|y,n)$ has exponentially light tails and satisfies either Theorem 4 or Theorem 3.

Before we can show this, we need a Theorem about convexity (Rockafeller and Wets, 1998, Theorem 2.14).

Theorem 11 (Higher-dimensional derivative tests)**.** *For a differentiable function f on an open convex set $O \subset \mathbb{R}^n$, each of the following conditions is both necessary and sufficient for f to be convex on O:*

(a) $\langle x_1 - x_0, \nabla f(x_1) - \nabla f(x_0)\rangle \geq 0$ for all x_0 and x_1 in O;

(b) $f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle$ for all x and y in O;

(c) $\nabla^2 f(x)$ is positive-semidefinite for all x in O (f twice differentiable).

For strict convexity, a necessary and sufficient condition is (a) holding with strict inequality when $x_0 \neq x_1$, or (b) holding with strict inequality when $x \neq y$. A condition

that is sufficient for strict convexity (but not necessary) is the positive definiteness of the Hessian matrix in (c) for all x in O.

We will prove that the density in (5.11) has exponentially light tails in stages. The following conditions will be useful

$$\text{The vector of all ones is not in the column space of } M^l \text{ for any } l. \tag{5.12}$$

$$0 < y_i^l \text{ for all } i \text{ and } l. \tag{5.13}$$

It is worth noting that conditions (5.12) and (5.12) imply that for all β and $\delta \in S_1 = \{\delta \mid \delta^T\delta = 1\}$

$$\pi(\beta) \neq \pi(\beta + s\delta), \qquad s > 0.$$

I.e. π *does not have a direction of constancy.*

Lemma 10. *If π is of the form in* (5.11) *and satisfies conditions* (5.12) *and* (5.13) *then there exists an $\epsilon > 0$ such that for all $\delta \in S_1$*

$$\langle \delta, \nabla \log \pi(\tilde{\beta} + \delta) \rangle \leq -\varepsilon$$

where $\tilde{\beta}$ is the posterior mode of π.

Proof. From (Geyer, 1990, Sec. 2.2) we know that π is log-concave, so $-\log \pi$ is convex. Furthermore, it follows from conditions (5.12) and (5.13) that $\tilde{\beta}$ is the unique posterior mode of π.

Let $\delta \in S_1$ be given and apply Theorem 11 (b) with $y = \tilde{\beta}$, $x = \tilde{\beta} + \delta$ and

$f = -\log \pi$ thus

$$\langle -\nabla \log \pi(\tilde{\beta} + \delta), \tilde{\beta} - (\tilde{\beta} + \delta) \rangle - \log \pi(\tilde{\beta} + \delta) \leq -\log \pi(\tilde{\beta})$$
$$\Leftrightarrow \langle \nabla \log \pi(\tilde{\beta} + \delta), \delta \rangle \leq \log \pi(\tilde{\beta} + \delta) - \log \pi(\tilde{\beta}). \quad (5.14)$$

Since S_1 is compact, and $\log \pi$ is a continuous function, there exists a $\delta' \in S_1$ such that $\log \pi(\tilde{\beta} + \delta') = \sup_{\delta \in S_1} \log \pi(\tilde{\beta})$. Since $\tilde{\beta}$ is the unique posterior mode, and π does not have a direction of constancy $\log \pi(\tilde{\beta} + \delta') < \log \pi(\tilde{\beta})$ and $\varepsilon = \log \pi(\tilde{\beta}) - \log \pi(\tilde{\beta} + \delta') > 0$. Hence $\log \pi(\tilde{\beta} + \delta) - \log \pi(\tilde{\beta}) \leq \log \pi(\tilde{\beta} + \delta') - \log \pi(\tilde{\beta}) = -\varepsilon < 0$. Combining this with (5.14) we get

$$\langle \delta, \nabla \log \pi(\tilde{\beta} + \delta) \rangle \leq -\varepsilon.$$

□

Theorem 12. *If π is of the form in* (5.11) *and satisfies conditions* (5.12) *and* (5.13)*, then π has exponentially light tails.*

Proof. Let $\tilde{\beta}$ represent the posterior mode of π. For $\|\beta\|$ large enough, we can replace β with $\tilde{\beta} + s\delta$ where s is greater than one and δ is in S_1. This makes our goal

$$\limsup_{s \to \infty} \sup_{\delta \in S_1} \frac{\tilde{\beta} + s\delta}{\|\tilde{\beta} + s\delta\|} \cdot \nabla \log \pi(\tilde{\beta} + s\delta) < 0.$$

This lim sup is bounded above by the sum of

$$\limsup_{s \to \infty} \sup_{\delta \in S_1} \frac{\tilde{\beta}}{\|\tilde{\beta} + s\delta\|} \cdot \nabla \log \pi(\tilde{\beta} + s\delta) \quad (5.15)$$

and

$$\limsup_{s\to\infty} \sup_{\delta\in S_1} \frac{s\delta}{\|\tilde{\beta} + s\delta\|} \cdot \nabla \log \pi(\tilde{\beta} + s\delta). \tag{5.16}$$

Note that by Lemma 9, the dot product in (5.15) is a fraction with a bounded numerator and an unbounded denominator. Hence this lim sup is zero.

It should be obvious that for all δ in S_1, the fraction

$$\frac{\|s\delta\|}{\|\tilde{\beta} + s\delta\|}$$

will go to one as s goes to infinity. If we replace the dot product with an inner product, the lim sup in (5.16) is equal to

$$\limsup_{s\to\infty} \sup_{\delta\in S_1} \langle \delta, \nabla \log \pi(\tilde{\beta} + s\delta) \rangle. \tag{5.17}$$

Again, we will use the fact that π is log-concave, so $-\log \pi$ is convex (Geyer, 1990, Sec. 2.2). Taking s greater than one, we can then apply Theorem 11 part (a) to $-\log \pi$ with $x = \tilde{\beta} + s\delta$ and $x_0 = \tilde{\beta} + \delta$. With some algebra, this yields the relationship

$$\langle \delta, \nabla \log \pi(\tilde{\beta} + s\delta) \rangle \leq \langle \delta, \nabla \log \pi(\tilde{\beta} + \delta) \rangle.$$

From Lemma 10, the right hand side of this inequality is bounded above by $-\varepsilon$ for some ε greater than zero. Therefore the lim sup in (5.16) is bounded above by

$$\limsup_{s\to\infty} \sup_{\delta\in S_1} -\varepsilon = -\varepsilon < 0$$

so π has exponentially light tails. □

It's clear that π in the form (5.11) is always positive and has continuous first

derivatives. If π satisfies conditions (5.12) and (5.13) then Theorem 12 shows that π is exponentially light, and Lemma 9 shows that $\|\nabla \log \pi(\beta)\|$ is bounded. Hence π satisfies the conditions of Theorems 2 and 3 so if $\gamma = h^{-1}(\beta)$ for isotropic h defined using f from (4.1), then a random-walk Metropolis algorithm for the induced π_γ will be geometrically ergodic and we can use this Markov chain to learn about π.

5.2 Multivariate T Distribution

It is well known (Mengersen and Tweedie, 1996; Jarner and Hansen, 2000; Jarner and Roberts, 2007) that a random-walk Metropolis algorithm for a T distribution is not geometrically ergodic. But our results show that a random-walk Metropolis algorithm density of the transformed T variable is.

The multivariate T density on $\mathbb{R}^k$ with v degrees of freedom is

$$f_T(t|\mu,\Sigma) = \frac{\Gamma\big[(v+k)/2\big]}{\Gamma\big[v/2\big]\,(v\pi)^{k/2}\,|\Sigma|^{1/2}} \left[1 + \frac{1}{v}(t-\mu)^T\Sigma^{-1}(t-\mu)\right]^{-\frac{v+k}{2}}. \tag{5.18}$$

So

$$\begin{aligned} \nabla \log f_T(t|\mu,\Sigma) &= -\frac{v+k}{2}\nabla \log\left[1 + \frac{1}{v}(t-\mu)^T\Sigma^{-1}(t-\mu)\right] \\ &= -(v+k)\frac{(t-\mu)^T\Sigma^{-1}}{v + (t-\mu)^T\Sigma^{-1}(t-\mu)}. \end{aligned} \tag{5.19}$$

Now consider the dot product

$$\begin{aligned} \frac{t}{\|t\|}\cdot\nabla \log f_T(t|\mu,\Sigma) &= \frac{t}{\|t\|}\cdot\frac{-(v+k)(t-\mu)^T\Sigma^{-1}}{v + (t-\mu)^T\Sigma^{-1}(t-\mu)} \\ &= \frac{-(v+k)}{\|t\|}\,\frac{t^T\Sigma^{-1}(t-\mu)}{v + (t-\mu)^T\Sigma^{-1}(t-\mu)} \\ &= \frac{-(v+k)}{\|t\|}\left[1 + \frac{v}{t^T\Sigma^{-1}(t-\mu)} - \frac{\mu^T\Sigma^{-1}(t-\mu)}{t^T\Sigma^{-1}(t-\mu)}\right]^{-1} \end{aligned} \tag{5.20}$$

Clearly, the left-hand term in this final equality goes to zero and the right-hand term goes to one as $\|t\|$ goes to infinity. Therefore

$$\limsup_{\|t\|\to\infty} \frac{t}{\|t\|} \cdot \nabla \log f_T(t|\mu, \Sigma) = 0.$$

It's easy to see that f_T is always positive and has continuous first derivatives, so f_T is a sub-exponential density.

It's clear from the form of $\frac{t}{\|t\|} \cdot \nabla \log f_T(t|\mu, \Sigma)$ in (5.20) that for $\alpha = k + v/2$ and $\alpha_0 = k + v + 1$ that for large enough $\|t\|$

$$\frac{t}{\|t\|} \cdot \nabla \log\big(\|t\|^{-\alpha_0}\big) \le \frac{t}{\|t\|} \cdot \nabla \log f_T(t|\mu, \Sigma) \le \frac{t}{\|t\|} \cdot \nabla \log\big(\|t\|^{-\alpha}\big).$$

Since v is positive, $\alpha = k + v/2$ is greater than k. Thus f_T and $\alpha = k + v/2$ satisfy the conditions of Theorem 5, so if $h_1(Y) = T$ for isotropic h_1 with f equal to f_1 from (4.20) then the induced density π_Y will be exponentially light. Furthermore, using $\alpha_0 = k + v_1$, f_T satisfies the conditions of Theorem 6 so $\|\nabla \log \pi_Y(y)\|$ is bounded as $\|y\|$ goes to infinity. Since π_Y is exponentially light, and $\|\nabla \log \pi_Y(y)\|$ is bounded, Theorems 2 and 3 state that if $h_2(X) = Y$ for isotropic h_2 defined using f from (4.1), π_X will be super-exponential and satisfy the curvature condition in (2.12) so a random-walk Metropolis algorithm (with appropriate proposal density, q) will be geometrically ergodic. Since both h_1 and h_2 are one-to-one and onto, a Markov chain on X can be used to learn about the distribution of T.

Chapter 6

Computer Aided Reasoning in Statistics

6.1 Introduction

Statisticians formulate and prove new theorems, such as the Cramér-Rao Inequality or the Delta method. Statisticians also apply theorems, methods and algorithms to real world problems, such as using Markov chain Monte Carlo to perform a Bayesian data analysis. Proving and formulating new theorems is done in a traditional paper-and-pencil mathematics approach. These days, almost every statistical analysis uses computer implementations of statistical theorems, methods and algorithms. There is little-to-none verification of these computer implementations.

A computer implementation of a statistical method is created by writing computer code. Assuming that the underlying algorithm is correct, the code is typically claimed to be correct if it "seems to do the right thing." Writing correct computer code is very difficult. Even individuals and organizations that are determined to have correct code encounter errors that seem small in hindsight, but create large problems. NASA lost the Mars Polar Lander ($110 million construction cost) when a software error incorrectly turned off the retro-rockets (JPL Special Review Board, 2000). Very little work has been done in this area in the context of statistics. Verification of computer

implementations in statistics remains ripe for further work.

Statisticians prove new theorems using traditional proofs, also called informal proofs or paper-and-pencil proofs. These are the kinds of proofs that we are used to seeing in textbooks and journal articles. These proofs face the dual burden of compelling belief and conveying understanding, two antagonistic goals (Harrison, 2008). Compelling belief requires taking care of all the fiddly bits, or low-level details which seldom aids understanding. Conveying understanding is accomplished by discussing the high level details. As a compromise, many proofs contain statements like "without loss of generality" or "by symmetry". Thus these paper-and-pencil proofs do not verify all steps, which is why they are considered to be informal proofs. Additionally, highly complex proofs are usually broken up into smaller proofs to make it easier to compel belief, conveying understanding takes a back seat and is done with a separate story.

A formal proof is a proof in which every step is checked all the way back to the formal axioms (Hales, 2008). Nothing is left out, no matter how intuitive. Thus there is less room for logical errors in a formal proof. No one wants to read such a proof, let alone write one by hand, which is why paper-and-pencil proofs are informal. Nor does formal proof help with highly complex proofs, as the proofs will be even longer when done formally.

Using a computer to do the formal proof, or computer aided reasoning, alleviates this burden. The proof is correct because it has been formally verified with a trusted system. Or put simply, **the proof is correct because the computer says so!** Additionally, computers are very good at book-keeping so a highly complex proof is much less of an issue for a computer than a person. For example, the proof of the four-color theorem requires handling thousands of cases, this can be done with a completely formal computerized proof (Gonthier, 2008).

6.2 HOL Light: Calculemus

Leibniz said a system for reasoning must contain a universal language and a calculus of reasoning. If two people disagreed, they need not argue, only write their claims into the system for reasoning and say "calculemus" (let us calculate) (Harrison, 2008). Mechanical theorem-proving systems are just that, a formalized universal language, and a calculus of reasoning. All you need to do is put your theorems in and 'turn the crank'.

Formal mathematics provides this formalized language and calculus of reasoning. Formal mathematics starts with a base, a select few axioms and rules of inference. A theorem is a conjecture or statement that has been demonstrated to be true by only using the axioms, rules of inference, and existing theorems. This demonstration is called a proof. It is simply done by turning the crank. This is too tedious to do by hand.

Computer proof assistant programs take formal mathematics and relieve us of the burden of turning the crank. These computer proof assistant programs turn proof into a mechanical process (Hales, 2008). Any theorem proved inside of a computer proof assistant program has been proved formally, as the program only accepts steps inside of the proof if they are valid by the axioms and rules of inference, or previously verified theorems. As well as working in a formal manner, computer proof assistant programs can help to fill in gaps in a proof. The size and type of the gaps depends on the proof assistant program in use.

There are many computer proof assistants available: HOL Light (Harrison, 2011); Isabelle (Paulson and Nipkow, 2009); Mizar (Mizar Project, 2009); ACL2 (Kaufmann and Moore, 2011); and others. We chose to use HOL Light for this work, because it is powerful, popular and theorems in HOL Light look close to theorems in mathematics. The name, "HOL Light" is short for "Higher order logic light". Higher order logic,

because that is what HOL Light implements. A key feature of higher order logic is the ability to quantify across functions, e.g.

$$\forall f : \mathbb{R} \to \mathbb{R}$$

is a valid quantification in higher order logic, but not in first order logic. "Light", not because HOL Light is low power, but because it has a small kernel and HOL Light needs to distinguish itself from other computer theorem provers associated with the name "HOL", e.g. HOL4 (HOL4 Project, 2011). In computerized theorem provers, the term "kernel" refers to the part of the system that must be accepted to be true. The rest of the system is proved true using the kernel. The kernel of HOL Light is less than 500 lines of code (Hales, 2008). For comparison, a complex piece of software like the Linux kernel has over 14.6 million lines of code. Verifying the correctness of 500 lines of code seems trivial in comparison, and there is work to provide as high a level of confidence in this kernel as possible (Harrison, 2006).

The power and popularity can be seen by the fact that HOL Light is the system at the top of the "Formalizing 100 Theorems" list (Wiedijk, 2011) with 76 of the 100 theorems proved true within HOL Light, displayed in Table 6.1. The closeness to traditional theorems can be seen by looking at a very trivial theorem:

$$\forall x, y, z \in \mathbb{R}, \qquad x + y + z = y + x + z.$$

Written in HOL Light the *pretty-printed*[1] theorem is

```
|- !x y z. x + y + z = y + x + z
```

Placed next to the traditional version, the translation can nearly be guessed. The

[1]Pretty-printing is printing something out in a visually pleasing way. Artistic license is granted so a pretty-printed version may not correspond exactly with the true version. Listing 6.1 shows how this theorem looks without pretty printing. Clearly, we'd prefer to use the pretty-printed version whenever possible.

Code Listing 6.1: Non pretty printed version of the HOL Light theorem , `|- !x y z. x + y + z = y + x + z.`

```
Comb (Const ("!", `:(real->bool)->bool`),
 Abs (Var ("x", `:real`),
  Comb (Const ("!", `:(real->bool)->bool`),
   Abs (Var ("y", `:real`),
    Comb (Const ("!", `:(real->bool)->bool`),
     Abs (Var ("z", `:real`),
      Comb
       (Comb (Const ("=", `:real->real->bool`),
         Comb
          (Comb (Const ("real_add", `:real->real->real`),
            Var ("x", `:real`)),
          Comb
           (Comb (Const ("real_add", `:real->real->real`),
             Var ("y", `:real`)),
            Var ("z", `:real`)))),
        Comb
         (Comb (Const ("real_add", `:real->real->real`), Var ("y", `:real`)),
         Comb
          (Comb (Const ("real_add", `:real->real->real`),
            Var ("x", `:real`)),
          Var ("z", `:real`)))))))))))
```

`!` symbol means "for all", and the `|-` symbol is the turnstile, usually read as "it is provable that". HOL Light hides the type information from view so there is no $\in \mathbb{R}$, but it is still there and can be displayed if desired[2].

We are not providing a tutorial on how use HOL Light, only the aspects needed for the current discussion will be explained. If you want to learn how to use HOL Light, some good starting places are the *AMS* articles (Hales, 2008; Wiedijk, 2008) or the HOL Light tutorial (Harrison, 2007).

When HOL Light is started, first the kernel is loaded. As previously mentioned, this kernel is the only part of HOL Light that must be believed or trusted *a priori*. The kernel of HOL Light is very small (only 500 lines) and contains, three axioms, ten rules of inference and the basic engine of HOL Light (Hales, 2008). Once this kernel has been loaded, any new theorem proved correct by HOL Light is formally verified by this kernel. Of course, nobody wants to start with just these axioms, rules of inference and basic engine. Once the kernel is loaded, HOL Light then loads a

[2]Displaying the type information is sometimes necessary. Sometimes to non-equivalent conditions will look the same when pretty-printed, because the difference is in the typing. Displaying the type information (or turning off the pretty printing) can help show the differences. If this happens, you are probably having a bad day of theorem proving.

Table 6.1: Number of the "Top 100 Theorems" formalized by each system (not an exhaustive list of systems) (Wiedijk, 2011).

System	# of 'top 100' theorems formalized
HOL Light	76
Mizar	51
Coq	49
C-CoRN	10
Isabelle	46
ProofPower	42
PVS	15
nqthm/ACL2	12
NuPRL/MetaPRL	8

large body of theory, and more advanced tools to use in developing theory[3]. All of this theory and these tools are formally verified by HOL Light, and hence formally correct. Verifying this theory and these tools every time HOL Light is started is a feature, not a bug! It helps keep the kernel smaller and easier to understand and verify. The downside is that starting HOL Light can take several minutes. This is a small price to pay to keep the kernel small.

Once HOL Light is started we can try to prove new theorems. A proof in HOL Light consists of the statement to be proved, and a list of instructions on how to prove the theorem. If the proof is successful, the statement becomes a theorem in HOL Light, and can be used to prove new theorems.

HOL Light proves theorems formally, and because it is a computer system it can handle complexity that is not possible in a paper-and-pencil proof. HOL Light is a *proof assistant*, so the instructions given to prove a statement do not need to be complete. HOL Light includes many tools for filling these gaps. But part of the power

[3]The theory includes (but is not limited to) some set theory, real analysis and number theory. The tools include more rules of inference, and *tactics* to try when proving new theorems. Basically, enough to get started in many applications. A brief overview is in the `VERYQUICKREFERENCE.txt` file in the HOL Light distribution. I keep a printed copy of this file on my desk when working in HOL Light.

of HOL Light comes from its extensibility.

HOL Light is implemented in OCaml (INRIA, 2010), so users can extend the core system using OCaml. OCaml functions can be written that return theorems! The OCaml functions must prove them inside of the HOL Light system, this extensibility does not provide an end-run around the formality of HOL Light.

Consider our earlier theorem,

$$\forall x, y, z \in \mathbb{R}, \qquad x + y + z = y + x + z.$$

This is clearly true, but we don't want to be burdened with verifying such trivial details just because we are working in a formal system. Fortunately, HOL Light includes an OCaml function, `REAL_ARITH` that is good at proving such statements about real numbers. `REAL_ARITH` takes statements as input, and if it can returns a theorem. If `REAL_ARITH` can't prove the theorem it throws an *exception*[4]. We can use `REAL_ARITH` at attempt to prove our conjecture:

```
# REAL_ARITH `!x y z:real. x + y + z = y + x + z ` ;;
val it : thm = |- !x y z. x + y + z = y + x + z
```

The `#` symbol is the OCaml prompt, and the `;;` ends the OCaml expression. The line beginning `val` gives the return value of the function, which is the desired theorem. To demonstrate that these functions can't cheat, we'll apply `REAL_ARITH` to the false statement $x = 2x$ for all real x:

```
# REAL_ARITH `!x:real. x = &2 * x` ;;
Exception: Failure "linear_ineqs: no contradiction".
```

Instead of returning a theorem, the function failed because it could not prove the statement. Note that the function failing does not mean that the statement is false

— this one is obviously false — just that the function could not prove the statement.

This extensibility is a powerful tool for managing complexity of proofs. It allows the use of statements equivalent to "it's clearly true that x + y + z = y + x + z" without sacrificing formality. It also allows arbitrarily complex proof searches to be written, which gives us hope that we may be able to use HOL Light to prove more complex theorems.

6.3 Notation

All the fun's in how you say a thing.

— Robert Frost

If "fun" is replaced with "work", this also holds for mathematics. If a theorem is expressed in the correct notation the proof may be obvious. A good example is the Cauchy-Schwarz inequality. It is expressed simply with inner products:

$$|\langle x, y\rangle|^2 \leq \langle x, x\rangle \cdot \langle y, y\rangle.$$

Proving the Cauchy-Schwarz inequality is simple if you remember that for all x and y, $\langle x - \lambda y, x - \lambda y\rangle$ is a quadratic polynomial in λ with at most one real root.

A counter example is the "multivariate" delta method, as stated by Casella and Berger (2002, p. 245):

Theorem (Multivariate Delta Method)**.** *Let* $\mathbf{X}_1, \ldots, \mathbf{X}_n$ *be a random sample with* $\mathrm{E}(X_{ij}) = \mu_i$ *and* $\mathrm{Cov}(X_{ik}, X_{jk}) = \sigma_{ij}$*. For a given function* g *with continuous*

[4]Throwing or raising an exception is a way for computer software to indicate an error of some kind has occurred. Thrown exceptions interrupt normal program flow, thrown exceptions are not returned. Thus, these tools (like `REAL_ARITH`) can only return theorems. The mechanics of exceptions are not important, we only need to know that we will see an error message instead of a theorem when exceptions are thrown in HOL Light.

first partial derivatives and a specific value of $\boldsymbol{\mu} = \big(\mu_1, \ldots, \mu_p\big)$ *for which* $\tau^2 = \sum\sum \sigma_{ij} \frac{\partial g(\boldsymbol{\mu})}{\partial \mu_i} \cdot \frac{\partial g(\boldsymbol{\mu})}{\partial \mu_j} > 0$,

$$\sqrt{n}\big[g(\bar{X}_1, \ldots, \bar{X}_s\big] \to \mathscr{N}(0, \tau^2) \text{ in distribution.}$$

The notation is distracting, and will probably lead to a confused proof. For contrast, consider the multivariate delta method as stated in Ferguson (2002, p. 45):

Theorem. *Let* $\mathbf{g}$ *be a mapping* $\mathbf{g} : \mathbb{R}^d \to \mathbb{R}^k$ *such that* $\dot{\mathbf{g}}(\mathbf{x})$ *is continuous in a neighborhood of* $\boldsymbol{\mu} \in \mathbb{R}^d$. *If* $\mathbf{X_n}$ *is a sequence of d-dimensional random vectors such that* $\sqrt{n}\big(\mathbf{X_n} - \boldsymbol{\mu}\big) \xrightarrow{\mathscr{L}} \mathbf{X}$, *then* $\sqrt{n}\big(\mathbf{g}(\mathbf{X_n}) - \mathbf{g}(\boldsymbol{\mu})\big) \xrightarrow{\mathscr{L}} \dot{\mathbf{g}}(\boldsymbol{\mu})\mathbf{X}$. *In particular, if* $\sqrt{n}(\mathbf{X_n} - \boldsymbol{\mu}) \xrightarrow{\mathscr{L}} \mathscr{N}\big(\mathbf{0}, \boldsymbol{\Sigma}\big)$ *where* $\boldsymbol{\Sigma}$ *is a* $d \times d$ *covariance matrix, then*

$$\sqrt{n}\big(\mathbf{g}(\mathbf{X_n}) - \mathbf{g}(\boldsymbol{\mu})\big) \xrightarrow{\mathscr{L}} \mathscr{N}\big(\mathbf{0}, \dot{\mathbf{g}}(\boldsymbol{\mu})\boldsymbol{\Sigma}\dot{\mathbf{g}}(\boldsymbol{\mu})^T\big).$$

The Ferguson version is better than the Casella and Berger version because $\tau^2 = \sum\sum \sigma_{ij} \frac{\partial g(\boldsymbol{\mu})}{\partial \mu_i} \cdot \frac{\partial g(\boldsymbol{\mu})}{\partial \mu_j}$ is replaced with $\dot{\boldsymbol{g}}(\boldsymbol{\mu})\boldsymbol{\Sigma}\dot{\boldsymbol{g}}(\boldsymbol{\mu})^T$. This replacement visually simplifies the equation, and does a better job of leading us down the path of the proof. On the one hand, such a small difference probably won't matter; replacing τ^2 is very easy to do. On the other hand, we are not really talking about people.

Theorems and proofs inside of computerized proof assistant programs work similarly. The theorem is expressed in some notation, and — hopefully — the proof follows from that. The notation in computer theorems is more important than in paper-and-pencil theorems because intuition is very hard to program. Experience teaches us when modifications — like replacing τ^2 with $\dot{\boldsymbol{g}}(\boldsymbol{\mu})\boldsymbol{\Sigma}\dot{\boldsymbol{g}}(\boldsymbol{\mu})^T$ — may be advantageous. These modifications are difficult for proof assistant programs to find; proof assistant programs only find the modifications they are programmed to look for. Furthermore a modification requires a mini-proof. e.g. replacing τ^2 with $\dot{\boldsymbol{g}}(\boldsymbol{\mu})\boldsymbol{\Sigma}\dot{\boldsymbol{g}}(\boldsymbol{\mu})^T$ requires prov-

ing equivalence of the two quantities. Proving equivalence of matrix multiplication to a double sum is easy, but other equivalences may not be so easy.

For those reasons, it will benefit us to write computer theorems in notation that is convenient for the proof assistant program.

"Convenient notation" depends on the proof assistant program in use. To demonstrate, let's consider vectors. What is a vector? Vectors are usually represented as columns of numbers. Honest reflection show that this is a mental construct and not a mathematical structure. ACL2 is implemented in *Lisp*, so the primary data structure is the *cons* pair. The natural vector representation is a *binary tree* with a *traversal rule*. We are taking binary tree as a tree structure consisting of a root node, nodes and leaves; each node has two children; leaves contain the values in the tree, and may contain `NIL` to indicate an empty node. A traversal rule is an algorithm for converting a binary tree into an ordered list. If f is a function that we want applied to a "vector" in order, we can express the traversal rule as a function, `traverse`, written in pseudo code in Listing 6.2. With this traversal rule, different binary trees can represent the same vector.

The `traverse` algorithm in Listing 6.2 will have the same application order — f(2); f(3); f(5); f(7) — for both Figure 6.1a and Figure 6.1b, but the tree in Figure 6.1a is visually more complicated. If we add 9 to the end of the vector, it's clear how the tree structure would change in Figure 6.1b, but not in Figure 6.1a. ACL2 is most effective when searching for proofs using induction. The structure in Figure 6.1b works very well with induction, the structure in Figure 6.1a does not. Hence when working with ACL2 it is advantageous to only consider vectors with structure like that in Figure 6.1b.

While we can see readily (or with some prompting) see the equivalence between representations — like the equivalence of τ^2 and $\dot{\boldsymbol{g}}(\boldsymbol{\mu})\boldsymbol{\Sigma}\dot{\boldsymbol{g}}(\boldsymbol{\mu})^T$, or the equivalence of the tree structures shown in Figure 6.1b and Figure 6.1a — it is harder for proof

Code Listing 6.2: Vector traversal algorithm.

```
traverse(f, node)
  if node is a number then
    evaluate f(node)
  else if node is an endpoint (NIL) then
    do nothing
  else
    traverse(f, node left child)
    traverse(f, node right child)
```

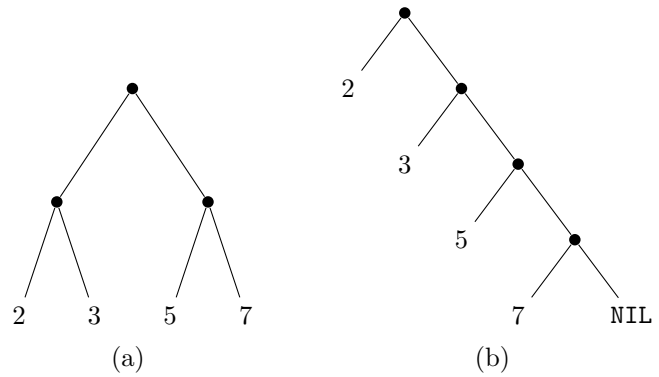


Figure 6.1: Different tree representations of the vector $(2, 3, 5, 7)^T$

assistant programs to spot these equivalences.

Computer proofs can be made easier by using notation that is convenient for the proof assistant program. The particular notation will depend on the particular program being used, e.g. in HOL Light, it is best to treat vectors as functions (Harrison, 2005). Some of this notation will come at the expense of human readability and loss of "closeness" to the paper-and-pencil theorem, but it will greatly ease computer proofs.

6.4 Previous Work

Early formalizations of measure theory and probability (Nęzusiak, 1989; Białas, 1990) did not proceed far enough for our purposes. There has been remarkable progress, beginning with John Harrison (Harrison, 1998) who made a formal construction of real numbers, real analysis and formalized the Kurzweil-Henstock gauge integral over real numbers (Bartle, 2001). Joe Hurd extended this work to measure theory and some basic probabilistic algorithms (Hurd, 2002). Hurd defined the spaces $\mathbb{B}$ and $\mathbb{B}^\infty$,

$$\mathbb{B} = \left\{ True, False \right\}$$

and

$$\mathbb{B}^\infty = \left\{ s \mid s : \mathbb{N} \to \mathbb{B} \right\} \tag{6.1}$$

and formalized a probability measure

$$\mathbb{P} : \mathcal{P}(\mathbb{B}^\infty) \to [0, 1] \tag{6.2}$$

such that if $i \in \mathbb{N}$ and $B_i = \{ x \in \mathbb{B}^\infty \mid x(i) = True \}$ then 6.2 satisfies

$$\mathbb{P}(B_i) = \tfrac{1}{2}.$$

Hurd was able to define several discrete distributions using $\mathbb{B}^\infty$ as the sample space, including Binomial$(n, \frac{1}{2})$, Uniform(n), Geometric$(\frac{1}{2})$, and Bernoulli(p) using the probability measure (6.2). This has been extended to Lebesgue-style integration (Richter, 2004).

Hasan and Tahar considerably extended Hurd's work. They formalized some Cumulative Distribution Function (CDF) properties (Hasan and Tahar, 2007c). Hasan and Tahar demonstrated a simple construction of a Uniform$(0, 1)$ random from $\mathbb{B}^\infty$. For $s \in \mathbb{B}^\infty$ we can construct X by

$$X = \lim_{n\to\infty} \sum_{k=0}^{n-1} \Big(\frac{1}{2}\Big)^{k+1} I\big(s(k)\big)$$

where $I(\cdot)$ is the indicator function, $I : \mathbb{B} \to \{0, 1\}$ defined by

$$I(t) = \begin{cases} 1 & \text{if } t = \text{True} \\ 0 & \text{otherwise} \end{cases}$$

Then X will have the desired CDF

$$Pr\big(X \leq x\big) = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{if } 0 \leq x < 1 \\ 1 & \text{if } 1 \leq x \end{cases}$$

Hasan represented a random variable, R as

$$R : \mathbb{B}^\infty \to \mathbb{R}.$$

Hasan defined the CDF of a random variable R, $F_R(x)$ in terms of (6.2)

$$F_R(x) = \mathbb{P}\big(\{s|R(s) \leq x\}\big)$$

and was able to verify CDF bounds, monotonicity, limits and the fact

$$\text{if } a < b, \text{ then } P(a < R \leq b) = F_R(b) - F_R(a).$$

Hasan and Tahar further extended this work by formalizing the Inverse Transform Method (ITM) for continuous random variables (Hasan and Tahar, 2007a). In short, if $F_Y(x)$ is a continuous CDF and $F_Y^*(x)$ is defined by

$$F_Y^*(y) = \begin{cases} \sup\{x \mid x \in \mathbb{R} \wedge F_Y(x) = 0\} & \text{if } y = 0 \\ F_Y^{-1}(y) & \text{if } 0 < y < 1 \\ \inf\{x \mid x \in \mathbb{R} \wedge F_Y(x) = 1\} & \text{if } y = 1 \end{cases}$$

Then if U follows a Uniform$(0, 1)$ distribution and we set $Y = F_Y^*(u)$ Y is a random variable with CDF $F_Y(x)$. Unfortunately Hasan and Tahar only used this method for random variables with a closed form CDF. Hasan and Tahar also verified some expectation and variance properties of discrete random variables (Hasan and Tahar, 2007b), including linearity of expectation and linearity of variance for independent random variables.

Finally, Mhamdi et al. (2010) build on the work of Hurd (2002) and Coble (2009) to prove the linearity and monotone convergence of Lebesgue integration. They use

these results to prove the Markov inequality, Chebyshev inequality and a version of the weak law of large numbers for random variables that have expectations calculated using Lebesgue integration.

While an impressive collection of results, these results do not go far enough. The expectation and variance results have only been verified for discrete random variables that take values in the non-negative integers and the ITM results have only been proved for continuous random variables that have a closed form CDFs.

Osman and Tahar said (Hasan and Tahar, 2007b)

> ... we found mechanical theorem-proving very efficient in book keeping. For example it is very common to get confused with different variables and mathematical notations and make human errors when working with large paper-pencil proofs, which leads to the loss of a lot of effort, whereas in the case of mechanical theorem provers such problems do not exist. ... Thus, it can be concluded that mechanical theorem-proving is a tedious but promising field, which can help mathematicians to cope with the explosion in mathematical knowledge and to save mathematical concepts from corruption.

6.5 A Formal Proof of the Markov Inequality

The Markov inequality states that for a non-negative random variable X and $a > 0$,

$$\mathrm{P}(X \geq a) \leq \frac{\mathrm{E}(X)}{a}. \tag{6.3}$$

Instead of using measure theory as a base, we use the expectation approach as outlined by Whittle (2000).

In HOL Light a set is a function from a type to boolean values. e.g. the set

$A = \{1, 2, \pi\}$ is represented as a function from $\mathbb{R}$ to true or false. A evaluates to true for 1, 2, or π. A evaluates to false for all other values. However, with a little "syntactic sugar", there is not much visual difference between this and the usual approach. e.g. HOL Light defines the `IN` operator. Let $W : B \to \mathbb{B}$ ($\mathbb{B}$ represents the boolean type) be a set on the generic type B. Then the `IN` operator is defined by the relationship

$$x \texttt{ IN } W \Leftrightarrow W(x)$$

that is, x `IN` W is true exactly when $W(x)$ evaluates to true. So the `IN` operator can be viewed as equivalent to the traditional $\in$.

We shall use Ω to denote the sample space. A random variable is a function from Ω to $\mathbb{R}$. Because sets are functions in HOL Light, it does not make sense to talk about a function

$$X : S \to \mathbb{R}$$

where $S : \Omega \to \mathbb{B}$ is a set. So Ω must be a type and not a set. It's fair to restrict consideration of X to the set S with conditions like

$$X(\omega) \geq 0, \qquad \text{for } \omega \in S$$

meaning that X is non-negative on the set S. As in the usual measure theoretic approach, the actual type Ω is not important for this formalization, just that it's there. HOL Light does not have any problems working with so-called *free* types, or types that are not specified.

We shall use 'E' as the expectation operator. E maps random variables to real

values. Using Ω as the sample space, the function signature of E is

$$\mathrm{E} : \big(\Omega \to \mathbb{R}\big) \to \mathbb{R}$$

Using this notation, the expectation of a random variable X can be written

$$\mathrm{E}\, X.$$

$\mathrm{E}\, X$ is equivalent to $\mathrm{E}(X)$. The parenthesis are optional in this case. To simplify the presentation of this theory, we will use the conventions that if X_1 and X_2 are random variables, and c_1 and c_2 are a real numbers then

$$\begin{aligned} c_1 X_1 &:= \omega \mapsto c_1 X_1(\omega) \\ c_1 X_1 + c_2 X_2 &:= \omega \mapsto c_1 X_1(\omega) + c_2 X_2(\omega) \\ |X_1| &:= \omega \mapsto |X_1(\omega)|. \end{aligned}$$

e.g. $c_1 X_1$ is taken to be the function that maps ω to $c_1 X_1(\omega)$. Thus if we write $\mathrm{E}(c_1 X_1 + c_2 X_2)$ what we really mean is

$$\mathrm{E}\big(\omega \mapsto c_1 X_1(\omega) + c_2 X_2(\omega)\big).$$

We can't get away with this in the HOL Light code without formally defining what all of the operations $+$, $-$, $*$ and $|\cdot|$ mean when used with functions from Ω to $\mathbb{R}$. Our HOL Light code uses the long-form of these compositions, but we will use the short form here.

In Whittle (2000, Chap. 2), a valid expectation operator, E must satisfy the following five axioms

A1 If $X \geq 0$ then $\mathrm{E}(X) \geq 0$.

A2 If c is a constant then $\mathrm{E}(cX) = c\,\mathrm{E}(X)$.

A3 $\mathrm{E}(X_1 + X_2) = \mathrm{E}(X_1) + \mathrm{E}(X_2)$.

A4 $\mathrm{E}(1) = 1$.

A5 If a sequence of r.v.s.$\{X_n\}$ increases monotonically to a point-wise limit X then

$$\mathrm{E}(X) = \lim \mathrm{E}(X_n).$$

Whittle includes the comment

> The equations in the axioms are all to be understood in the sense that, if the right-hand member is well defined, then the left-hand member is also well defined, and the two are equal. There are occasional failures. For example, suppose that $\mathrm{E}(X_1) = +\infty$ and $\mathrm{E}(X_2) = -\infty$. Then **A3** would give $\mathrm{E}(X_1 + X_2)$ the indeterminate value $+\infty - \infty$.

As we are working in a formal system, we cannot get away with such a statement. We need to include the expectation existing in the axioms. The usual way of writing that an expectation exists is

$$\mathrm{E}(|X|) < \infty.$$

We can't use this definition as HOL Light does not include ∞ as a number. The consensus is that including ∞ as a number in the traditional way probably won't work in a formal system, a formal system most likely needs to have all of the hyper-reals or none of them (Hurd, 2002). We avoid the issue of ∞ by saying that $\mathrm{E}(X)$ exists if there is a $c \in \mathbb{R}$ such that

$$\mathrm{E}(|X|) = c.$$

Equivalently we will write, expectation_exists(E, X) defined with the relationship

$$\text{expectation\_exists}(\mathrm{E}, X) \Leftrightarrow \exists c \in \mathbb{R} \text{ s.t. } \mathrm{E}(|X|) = c.$$

Our formalization of **A1**–**A3** pre-pends each of the axioms with the condition that the relevant expectation exists.

A1′ If $X(\omega) \geq 0$ for all ω and $\mathrm{E}(X)$ exists then $\mathrm{E}(X) \geq 0$.

A2′ If c is a constant and $\mathrm{E}(X)$ exists then $\mathrm{E}(cX)$ exists and $\mathrm{E}(cX) = c\,\mathrm{E}(X)$.

A3′ If $\mathrm{E}(X_1)$ exists and $\mathrm{E}(X_2)$ exists then $\mathrm{E}(X_1 + X_2)$ exists and $\mathrm{E}(X_1 + X_2) = \mathrm{E}(X_1) + \mathrm{E}(X_2)$.

A4′ $\mathrm{E}(1) = 1$.

The statement "$X(\omega) \geq 0$ for all ω" in **A1′** is perfectly valid. HOL Light infers the type of ω from the usage, in this case as input to X. Thus HOL Light interprets "$X(\omega) \geq 0$ for all ω" as "$X(\omega) \geq 0$ for all $\omega \in \Omega$" which is exactly what we mean. HOL Light would also make the same inference if we wrote $X(\omega) \geq 0$, but not if we wrote $X \geq 0$.

In **A4′**, E(1) means $\mathrm{E}(\omega \mapsto 1)$, where $\omega \mapsto 1$ maps all values to 1. There is no difference between **A4** and **A4′**. We do not use **A5** in our proof of the Markov inequality, so we do not formalize it. We say that E is a valid expectation operator if **A1′**–**A4′** hold. We will shorten this to

$$\text{expectation}(\mathrm{E}) \tag{6.4}$$

which is true if E satisfies **A1′**–**A4′** and false otherwise. We assume that expectation(E) holds in the rest of this section.

We prove the finite linearity of expectation. That is, for finite n, if $a_1, \ldots, a_n$ are real numbers and $\mathrm{E}(X_i)$ exists for random variables $X_1, \ldots X_n$, then $\mathrm{E}\big(\sum_{i=1}^n a_i X_i\big)$ exists and

$$\mathrm{E}\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i \,\mathrm{E}(X_i). \tag{6.5}$$

The proof is done by induction. The case $n = 1$ follows from **A2′**. Suppose that linearity holds for $n - 1 > 0$. Set $Y = a_1 X_1 + \cdots + a_{n-1} X_{n-1}$. Then

$$\mathrm{E}\left(\sum_{i=1}^{n} a_i X_i\right) = \mathrm{E}(Y + a_n X_n)$$

Applying (in this order) **A3′**, (6.5) (which holds for $n-1$) to Y, **A2′**, and simplifying yields (6.5) for any n.

We also prove that if X_1 and X_2 are random variables such that both $\mathrm{E}(X_1)$ and $\mathrm{E}(X_2)$ exist, and

$$X_1(\omega) \geq X_2(\omega), \qquad \text{for all } \omega.$$

then

$$\mathrm{E}(X_1) \geq \mathrm{E}(X_2). \tag{6.6}$$

This can be seen by setting the random variable $Y = X_1 - X_2$. We know that $\mathrm{E}(Y)$ exists by **A4′**, and by **A1′**

$$\mathrm{E}(Y) \geq 0.$$

By (6.5), we can replace $\mathrm{E}(Y)$, with $\mathrm{E}(X_1) - \mathrm{E}(X_2)$. Moving $\mathrm{E}(X_2)$ to the other side

of the inequality finishes the proof.

Probability is defined as the expectation of an indicator function. So indicator functions need to be defined as random variables, as that is what expectation operators take as as input. We define the indicator function, $\mathbf{1} : (\Omega \to \mathbb{B}) \to (\Omega \to \mathbb{R})$ by

$$\mathbf{1}(A)(\omega) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{if } \omega \notin A. \end{cases} \tag{6.7}$$

The case splitting is necessary as `true` is not one and `false` is not zero. Then for a set $B : \Omega \to \mathbb{B}$, we define $\mathrm{P}(B)$ by

$$\mathrm{P}(B) := \mathrm{E}\big(\mathbf{1}(B)\big). \tag{6.8}$$

Let X be a non-negative random variable, and $a > 0$, then

$$\frac{X(\omega)}{a} \geq \mathbf{1}\Big(\big\{w | X(\omega) \geq a\big\}\Big)(\omega), \qquad \text{for all } \omega.$$

That is, the random variable $\mathbf{1}\big(\{w|X(\omega) \geq a\}\big)$ is bounded above by the random variable $a^{-1}X$. Applying (6.6) and linearity of expectations gives the result

$$\mathrm{E}(X)/a \geq \mathrm{E}\big(\mathbf{1}(\{w|X(\omega) \geq a\})\big).$$

Some small algebraic manipulation and our definition of probability (6.8) shows that this is exactly the Markov inequality in 6.3.

The final Markov inequality theorem in HOL Light is (adapted to traditional

notation) is

$$
\begin{aligned}
\vdash \forall \mathrm{E}, X, a;& \\
&\text{expectation } (\mathrm{E}) \\
&\wedge\ a > 0 \\
&\wedge\ (\forall \omega, X(\omega) \geq 0) \\
&\wedge\ \text{expectation\_exists } (\mathrm{E}, X) \\
&\wedge\ \text{expectation\_exists } \Big(\mathrm{E}, S, \mathbf{1}\big(\{\omega | X(\omega) \geq a\}\big)\Big) \\
&\Rightarrow \mathrm{E}(\omega \mapsto X(\omega)/a) >= \mathrm{P}\big(\{\omega | X(\omega) \geq a\}\big)
\end{aligned}
\tag{6.9}
$$

This statement is visually quite different from the traditional Markov inequality, but it reads the same. It says that for any valid expectation operator, $\mathrm{E} : \big(\Omega \to \mathbb{R}\big) \to \mathbb{R}$ (recall, quantifying over functions is fine), non-negative random variable X, $a > 0$, if both $\mathrm{E}(X)$ and $\mathrm{P}(X \geq a)$ exist, then the Markov inequality holds. With some additional axioms (**A6′** on the next page), we should be able to drop the condition that $\mathrm{P}(X \geq a)$ exists.

6.6 Discussion

nothing is more mystifying than a proof of the obvious

— Paul Lockhart

The HOL Light code that does the work for the proof in Section 6.5 is 528 lines long. These 528 lines do not contain the paragraphs of explanation that we have here, so it's quite a bit of work for a result than can be explained in a few pages. But it is a *formal* proof, so no detail — no matter how obvious — is left unproven. As has been noted by others, having to prove everything, even the obvious details, makes the

process very tedious. However, the proof can be trusted because it is a formal proof.

This approach is different from the work in Section 6.4, as all of those results are built on specific expectation operators, such as discrete expectation or Lebesgue integration. Extending that work to arbitrary random variables will be more difficult, such as a mixture distribution that has discrete and continuous components. The expectation approach we use will not have this trouble.

However, we won't get much further with just Axioms **A1′**–**A4′**. A formal version of **A5** will be necessary for convergence in distribution, which is needed for the Central Limit Theorem. It is annoying that our Markov inequality conditions on the existence of a probability. We won't be able to get rid of this without more axioms, one candidate is

A6′ If the sample space is a metric space, then expectation_exists(E, X) for every bounded continuous random variable X.

In combination with **A5** (or the not-stated formal version, **A5′**), these axioms should be strong enough to show that $\mathrm{P}(B)$ exists for any $B = \{\omega | X(w) \leq y\}$ for any $y \in \mathbb{R}$. The proof would consist of using a sequence of bounded continuous functions that converge point-wise to $\mathbf{1}(B)$. Set The functions would be piecewise linear, with the pieces corresponding to steps in $\mathbf{1}(B)$ would converge to lines with infinite slopes. Thus **A5′** and **A6′** would show that $\mathrm{P}(B)$ exists. It seems clear that only one of **A5′** and **A6′** would be insufficient for this purpose. These $\mathrm{P}(B)$ would give the cumulative distribution function (CDF) for X. Countable unions, intersections and complements, combined with linearity of expectation would be sufficient to imply the exists for $\mathrm{P}(\{\omega | X(\omega) \in A\})$ where A is any open or closed interval. This line of reasoning could be extended to get to the probability of almost any set of interest. However, it would not be enough to get to the entire Borel σ-algebra on the reals, $\mathcal{B}(\mathbb{R})$. Showing the existence of $\mathrm{P}(B)$ for any $B \in \mathcal{B}(\mathbb{R})$ will require either an extension

theorem or trans-finite induction[5].

We believe that this approach will be useful in future work, as working at the expectation level provides a useful level of abstraction. Any operation we want to consider as expectation (e.g. discrete sums, Riemann or Lebesgue integration, or a mixture of sums and integrals) would only need to be shown to satisfy **A1′**–**A6′**. One goal is to look at drift and minorization conditions in MCMC, this will probably not be feasible until we have a larger collection of results to build from. A closer goal is to prove Central Limit Theorem for independent and identically distributed random variables. We believe that the expectation approach will make these goals easier to achieve than if we used a measure theory approach.

[5]There is more discussion about constructing Borel sets on uncountable Polish spaces in Jech (2003, Chap. 11)

References

Andrieu, C. and Éric Moulines (2006). On the ergodicity properties of some adaptive MCMC algorithms. *Ann. Appl. Probab.*, 16:1462–1505.

Atchadé, Y. F. and Rosenthal, J. S. (2005). On adaptive Markov chain Monte Carlo algorithms. *Bernoulli*, 11:815–828.

Bai, Y., Roberts, G. O., and Rosenthal, J. S. (2010). On the Containment Condition for Adaptive Markov Chain Monte Carlo Algorithms.

Bartle, R. G. (2001). *A Modern Theory of Integration*. American Mathematical Society, Providence, Rhode Island.

Białas, J. (1990). The σ-Additive Measure Theory. *Journal of Formalized Mathematics*, 2.

Casella, G. and Berger, R. L. (2002). *Statistical Inference*. Duxbury, Pacific Grove, CA, 2nd edition.

Chan, K. and Geyer, C. (2004). Comment on "Markov chains for exploring posterior distributions". *The Annals of Statistics*, pages 1747–1758.

Christensen, O., Møller, J., and Waagepetersen, R. (2001). Geometric Ergodicity of Metropolis-Hastings Algorithms for Conditional Simulation in Generalized Linear Mixed Models. *Methodology and Computing in Applied Probability*, 3:309–327.

Coble, A. (2009). *Anonymity, Information, and Machine-Assisted Proof*. PhD thesis, University of Cambridge, UK.

Ferguson, T. S. (2002). *A Course in Large Sample Theory*. Chapman & Hall/CRC, Boca Raton, FL.

Flegal, J. M. and Jones, G. L. (2010). Batch Means and Spectral Variance Estimation in Markov chain Monte Carlo. *The Annals of Statistics*, pages 1034–1070.

Flegal, J. M. and Jones, G. L. (2011). Implementing Markov chain Monte Carlo: Estimating with Confidence. In Brooks, S., Gelman, A., Jones, G., and Meng, X., editors, *Handbook of Markov Chain Monte Carlo*. CRC Press, London.

Fort, G., Moulines, E., Roberts, G., and Rosenthal, J. (2003). On the geometric ergodicity of hybrid samplers. *J. Appl. Probab.*, 40:123–146.

Gelfand, A. and Smith, A. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, pages 398–409.

Geman, S. and Geman, D. (1984). Stochastic relaxtion, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 721–741.

Geyer, C. J. (1990). *Likelihood and exponential families*. PhD thesis, University of Washington. `http://purl.umn.edu/56330`.

Geyer, C. J. (1992). Practical Markov chain Monte Carlo. *Statistical Science*, pages 473–511.

Geyer, C. J. (2010). *mcmc: Markov Chain Monte Carlo*. R package version 0.8.

Gonthier, G. (2008). Formal Proof — The Four Color Theorem. *Notices of the American Mathematical Society*, 55:1382–1393.

Hales, T. C. (2008). Formal Proof. *Notices of the American Mathematical Society*, 55:1370–1380.

Harrison, J. (1998). *Theorem Proving with the Real Numbers*. Springer-Verlag.

Harrison, J. (2005). A HOL Theory of Euclidean space. In Hurd, J. and Melham, T., editors, *Theorem Proving in Higher Order Logics 2005*, volume 3603 of *Lecture Notes in Computer Science*, pages 114–129, Oxford, UK. Springer-Verlag.

Harrison, J. (2006). Towards Self-verification of HOL Light. *Lecture Notes in Computer Science*, 4130:177–191.

Harrison, J. (2007). Hol Light Tutorial (for version 2.20). `http://www.cl.cam.ac.uk/~jrh13/hol-light/tutorial_220.pdf`.

Harrison, J. (2008). Formal Proof — Theory and Practice. *Notices of the American Mathematical Society*, 55:1395–1406.

Harrison, J. (2011). HOL Light. `http://www.cl.cam.ac.uk/~jrh13/hol-light/`.

Harville, D. A. (1997). *Matrix Algebra from a Statistician's Perspective*. Springer-Verlag, New York.

Hasan, O. and Tahar, S. (2007a). Formalization of Continuous Probability Distributions. In Pfenning, F., editor, *Automated Deduction – CADE-21*, volume 4603, pages 3–18. Springer Berlin / Heidelberg.

Hasan, O. and Tahar, S. (2007b). Verification of Expectation and Variance for Discrete Random Variables. Technical report, Concordia University, Montreal, Canada. `http://hvg.ece.concordia.ca/Publications/TECH_REP/FVEVDR_TR07`.

Hasan, O. and Tahar, S. (2007c). Verification of Probabilistic Properties in HOL Using the Cumulative Distribution Function. In Davies, J. and Gibbons, J., editors, *Integrated Formal Methods*, volume 4591, pages 333–352. Springer Berlin / Heidelberg.

Hastings, W. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, pages 97–109.

Hobert, J. and Robert, C. (2004). A mixture representation of π with applications in Markov chain Monte Carlo and perfect sampling. *The Annals of Applied Probability*, pages 1295–1305.

HOL4 Project (2011). Hol4 Home Page. `http://hol.sourceforge.net/`.

Hurd, J. (2002). *Formal Verification of Probabilistic Algorithms*. PhD thesis, University of Cambridge, Cambridge, UK.

Ibragimov, I. and Linnik, Y. V. (1971). *Independent and Stationary Sequences of Random Variables*. Wolters-Noordhoff, Groningen.

INRIA (2010). Objective Caml. `http://caml.inria.fr/ocaml/index.en.html`.

Jarner, S. and Hansen, E. (2000). Geometric ergodicity of Metropolis algorithms. *Stochastic Processes and their Applications*, 85:341–361.

Jarner, S. and Roberts, G. (2002). Polynomial convergence rates of Markov chains. *The Annals of Applied Probability*, pages 224–247.

Jarner, S. and Roberts, G. (2007). Convergence of Heavy-tailed Monte Carlo Markov Chain algorithms. *Scandinavian Journal of Statistics*, 34:781–815.

Jarner, S. F. and Tweedie, R. (2003). Necessary Conditions for Geometric and Polynomial Ergodicity of Random-Walk-Type Markov chains. *Bernoulli*, pages 559–578.

Jech, T. (2003). *Set Theory*. Springer, Berlin.

Johnson, A. A., Jones, G. L., and Neath, R. C. (2009). Component-wise Markov chain Monte Carlo.

Johnson, L. and Geyer, C. J. (2010). Geometric ergodicity of a random-walk Metropolis algorithm for a transformed density. Technical Report 680, School of Statistics, University of Minnesota. `http://purl.umn.edu/96959`.

Jones, G., Haran, M., Caffo, B., and Neath, R. (2006). Fixed-width output analysis for Markov chain Monte Carlo. *Journal of the American Statistical Association*, pages 1537–1547.

Jones, G. L. and Hobert, J. P. (2001). Honest Exploration of Intractable Probability Distributions via Markov chain Monte Carlo. *Statistical Science*, pages 312–334.

Jones, G. L. and Hobert, J. P. (2004). Sufficient burn-in for Gibbs samplers for a hierarchical random effects model. *The Annals of Statistics*, 32:784–817.

JPL Special Review Board (2000). Report on the Loss of the Mars Polar Lander and Deep Space 2 Missions. `http://spacese.spacegrant.org/Failure\%20Reports/MPL_failure_report.pdf`.

Kaufmann, M. and Moore, J. S. (2011). ACL2 version 4.2. `http://www.cs.utexas.edu/users/moore/acl2/`.

Lang, S. (1993). *Real and Functional Analysis*. Springer, New York, 3rd edition.

Mengersen, K. and Tweedie, R. (1996). Rates of convergence of the Hastings and Metropolis algorithms. *The Annals of Statistics*, 24:101–121.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., and Teller, E. (1953). Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, pages 1087–1092.

Meyn, S. and Tweedie, R. L. (2009). *Markov Chains and Stochastic Stability*. Cambridge University Press, Cambridge, 2nd edition.

Mhamdi, T., Hasan, O., and Tahar, S. (2010). On the Formalization of the Lebesgue Integration Theory in HOL. In Kaufmann, M. and Paulson, L., editors, *Interactive Theorem Proving*, volume 6172 of *Lecture Notes in Computer Science*, pages 387–402. Springer Berlin / Heidelberg.

Mizar Project (2009). Mizar Home Page. `http://mizar.org/`.

Nęzusiak, A. (1989). σ-Fields and Probability. *Journal of Formalized Mathematics*, 1.

Papaspiliopoulos, O. and Roberts, G. (2007). Stability of the Gibbs Sampler for Bayesian Hierarchical Models. *The Annals of Statistics*, 36:1–26.

Papaspiliopoulos, O., Roberts, G. O., and Sköld, M. (2007). A General Framework for the Parameterisation of Hierarchical Models. *Statist. Sci.*, 22:59–73.

Paulson, L. and Nipkow, T. (2009). Isabelle. `http://www.cl.cam.ac.uk/research/hvg/Isabelle/`.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Richter, S. (2004). Formalizing Integration Theory with an Application to Probabilistic Algorithms. In Slind, K., Bunker, A., and Gopalakrishnan, G., editors,

Theorem Proving in Higher Order Logics, volume 3223, pages 271–286. Springer Berlin / Heidelberg.

Roberts, G. (1999). A note on acceptance rate criteria for CLTs for Metropolis-Hastings algorithms. *Journal of Applied Probability*, pages 1210–1217.

Roberts, G. and Tweedie, R. (1996). Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika*, 83:95–110.

Roberts, G. O. and Rosenthal, J. S. (2004). General state space Markov chains and MCMC algorithms. *Probability Surveys*, 1:20–71.

Roberts, G. O. and Sahu, S. K. (1997). Updating Schemes, Correlation Structure, Blocking and Parameterization for the Gibbs Sampler. *J. R. Statis. Soc.* **B**, 59:291–317.

Rockafeller, R. and Wets, R.-B. (1998). *Variational Analysis*. Springer-Verlag, Berlin Heidelberg.

Rosenthal, J. S. (1995). Minorization Conditions and Convergence Rates for Markov Chain Monte Carlo. *Journal of the American Statistical Association*, pages 558–566.

Ross, S. (1988). *A First Course in Probability*. Macmillan, New York, NY, 3rd edition.

Tierney, L. (1994). Markov chains for Exploring Posterior Distributions. *The Annals of Statistics*, pages 1701–1728.

Whittle, P. (2000). *Probability via Expectation*. Springer, New York, 4th edition.

Wiedijk, F. (2008). Formal Proof — Getting Started. *Notices of the American Mathematical Society*, 55:1408–1414.

Wiedijk, F. (2011). Formalizing 100 Theorems. http://www.cs.ru.nl/~freek/100/.