Convergence Rate Bounds for Iterative Random Functions Using One-Shot Coupling

Sabrina Sixta and Jeffrey S. Rosenthal

December 6, 2021

Abstract

One-shot coupling is a method of bounding the convergence rate between two copies of a Markov chain in total variation distance. The method is divided into two parts: the contraction phase, when the chains converge in expected distance and the coalescing phase, which occurs at the last iteration, when there is an attempt to couple. The method closely resembles the common random number technique used for simulation. In this paper, we present a general theorem for finding the upper bound on the Markov chain convergence rate that uses the one-shot coupling method. Our theorem does not require the use of any exogenous variables like a drift function or minorization constant. We then apply the general theorem to two families of Markov chains: the random functional autoregressive process and the randomly scaled iterated random function. We provide multiple examples of how the theorem can be used on various models including ones in high dimensions. These examples illustrate how theorem's conditions can be verified in a straightforward way. The one-shot coupling method appears to generate tight geometric convergence rate bounds.

Contents

1 Introduction

The study of Markov chain convergence rates focuses on evaluating how fast a positive recurrent Markov chain converges to its stationary distribution. On one hand a great deal of progress has been made in bounding the convergence rate for Markov chains defined in discrete state spaces [39, 37, 35]. On the other hand, despite the major developments made in bounding Markov chains in continuous state space, many applications of continuous

state space Markov chains do not have established convergence rate bounds. For example, convergence rate bounds related to Markov chain Monte Carlo (MCMC) models are useful for deciding the size of the burn-in period [16, 11], but many applied MCMC models do not have known upper bounds on their convergence rate [11]. Users need to rely on ad-hoc convergence diagnostics (e.g. [10]), which offer no guarantees.

Methods using the drift and minorization conditions (eg. [38, 2]), which guarantee geometric ergodicity defined in Subsection 2.2, are the most studied techniques for bounding Markov chains in continuous state space [32, 16]. The minorization condition is satisfied for a Markov chain $\{X_n\}_{n\geq 1}$ under the following circumstances: there exists a small set K, a probability measure Q and a positive number $\epsilon > 0$ such that $P(\cdot|X_n = x) \geq \epsilon Q(\cdot)$ for $x \in K$. The drift condition is satisfied if there exists a positive function V, and constants $\alpha > 1$ such that $E[V(X_{n+1})|X_n = x] \leq V(x)/\alpha$ [25, 32]. Bounds generated using the drift and minorization conditions have been applied to a wide array of problems such as [36, 42, 16].

Despite the widespread use of bounds generated by the drift and minorization conditions, there are drawbacks. First, it can be a challenge to identify a small set K and drift function V [24]. Second, it is shown in [29] based on results from [19] that bounds that utilize the minorization condition do not scale well in high dimensions.

One-shot coupling can provide an upper bound on the convergence rate of a Markov chain while not needing to identify any exogenous sets or functions, and it scales well in high dimensions. The one-shot coupling technique first described in [34] works by first converging the expected distance between two copies of a Markov chain. At the last iteration, the probability of coupling is evaluated when the expected distance between the copies is small. This contrasts with the drift and minorization technique, which attempts to couple the two Markov chain copies every time they enter some fixed small set K.

The reason one-shot coupling scales well in high dimensions is that it focuses on the expected distance between two Markov chains, which parallels methods for finding the convergence rate of a Markov chain in Wasserstein distance [24]. Further, methods for finding Markov chain convergence rate bounds on the Wasserstein distance have been shown to scale well in high dimensions [8, 29]. One previously established solution for generating Markov chain convergence rate bounds in the total variation distance is to first establish a convergent rate bound in Wasserstein distance (see [28] and references therein) and then bound the total variation distance with the Wasserstein distance (see [24]). Wasserstein distance is often associated with optimal transport, which has been

shown to be related to the common random number technique (see Chapter 2 of [18] for details on the common random number technique and its relationship to optimal transport). Our approach is to show how methodology behind the common random number simulation technique can be used for constructing theoretical bounds in total variation distance.

The one-shot coupling method utilizes the common random number simulation technique to directly bound the total variation distance. This method has already been shown to scale well in certain high dimensional examples [34, 27] and will be shown in this paper to scale well in high dimension for the Bayesian regression Gibbs sampler (example 4.2) and the Bayesian location Gibbs sampler (example 4.3).

This paper formalizes a general theorem 3.1 for bounding the convergence rate of a Markov chain in total variation distance using the one-shot coupling method. This theorem is then used as the foundation for bounding the convergence rate for all of the examples in this paper, which can be partitioned into two families: the random functional autoregressive process and the autoregressive conditional heteroscedastic (ARCH) process.

Section 4 focuses on bounding the convergence rate in total variation for random functional autoregressive processes. That is, Markov chains, $\{X_n\}_{n\geq 1}$, which are of the following form for $g : \mathbb{R}^2 \to \mathbb{R}$

$$X_n = g(\theta_{1,n}, X_{n-1}) + \theta_{2,n} \tag{1}$$

where $(\theta_{1,n}, \theta_{2,n}) \in \mathbb{R}^2$ are random and $(\theta_{1,n}, \theta_{2,n}) \perp\!\!\!\perp (\theta_{1,m}, \theta_{2,m})$ when $n \neq m$. We introduce the Sideways Theorem 4.1 to bound the convergence rate of such Markov chains. Examples 4.2 and 4.3 are Gibbs samplers and show how the one-shot coupling method generates tight geometric convergence rate bounds in high dimensions. Subsection 4.6 extends the Sideways Theorem to autoregressive processes in $\mathbb{R}^d$ using Proposition 2.3, which provides an upper bound on the total variation distance for a Markov chain with independent coordinates when an upper bound on each coordinate is known.

Section 5 provides convergence rate bounds for various autoregressive conditional heteroscedastic (ARCH) processes, $\{X\}_{n\geq 1}$, which are of the following form for $f : \mathbb{R}^2 \to \mathbb{R}$,

$$X_n = f(\theta_{1,n}, X_{n-1})\theta_{2,n}$$

Where $(\theta_{1,n}, \theta_{2,n})$ are i.i.d. random variables with respect to n.

Proofs for the theorems presented in this paper are found in the appendix, Section 6. The code used to generate all of the tables and calculations can be found on github.com/sixter/OneShotCoupling.

2 Background and notation

Let $\{X_n\}_{n\geq 1}$ and $\{X'_n\}_{n\geq 1}$ be two copies of the Markov chain over the state space $\mathcal{X}$ and define $\mathcal{L}(X_n)$ to be the distribution of the random variable X_n.

2.1 Total variation distance

We are interested in measuring the distance between the distribution of ftwo Markov chains. To measure this we use the total variation metric.

Definition 2.1 (Total variation distance)**.** The total variation distance between the laws of two random variables, X and X', defined on the state space $\mathcal{X}$ is

$$\|\mathcal{L}(X) - \mathcal{L}(X')\| = \sup_{A\subseteq\mathcal{X}} |P(X \in A) - P(X' \in A)|$$

where $\mathcal{L}(X)$ represents the distribution of the random variable X and A is a measurable set.

For random variables, $X, X' \in \mathbb{R}$ with defined density functions $f_X, f_{X'}$,

$$\|\mathcal{L}(X) - \mathcal{L}(X')\| = \frac{1}{2}\int_{\mathbb{R}} |f_X(x) - f_{X'}(x)|\lambda(dx) \tag{2}$$

We define π to be the stationary distribution of the Markov chain. We want to measure the total variation distance of a Markov chain to its stationary distribution.

The following are properties of total variation, which will be used in conjunction with the One-Shot Coupling Theorem 3.1 to establish upper bounds on the convergence rate for the examples in this paper.

Proposition 2.1 states that the total variation between two random variables is equal to the total variation of any invertible transform of the same random variables. This proposition resembles Lemma 4.13 of [22].

Proposition 2.1. *Let $X, X' \in \mathcal{X}$ be two random variables and let $f : \mathcal{X} \to \mathcal{Y}$ be an invertible and measurable function. Then,*

$$\|\mathcal{L}(f(X)) - \mathcal{L}(f(X'))\| = \|\mathcal{L}(X) - \mathcal{L}(X')\| \tag{3}$$

The proof is in Section 6.1.

In general, for a measurable (not necessarily invertible) function f, $f^{-1}(f(\mathcal{B})) \subset \mathcal{B}$, so the third equality becomes $\leq$ and

$$\|\mathcal{L}(f(X)) - \mathcal{L}(f(X'))\| \leq \|\mathcal{L}(X) - \mathcal{L}(X')\|$$

Proposition 2.2 states that the total variation distance between two random variables is bounded above by the expected value of the conditional random variable.

Proposition 2.2. *Let X, X' be two random variables with corresponding σ-field $\mathcal{B}$ and $Y \in \mathcal{Y}$ be some related random variable. Then*

$$\|\mathcal{L}(X) - \mathcal{L}(X')\| \leq E\left[\|\mathcal{L}(X|Y) - \mathcal{L}(X'|Y)\|\right]$$

The proof is in Section 6.1.

Proposition 2.3 states that the convergence rate of a Markov chain in $\mathbb{R}^d$ with independent coordinates is d times the maximum coordinate-wise convergence rate. This means that the geometric convergence rate of the Markov chain is invariant to dimension when the coordinates are independent with a geometric convergence rate that is bounded above. This proposition is an application of inequality 1.2 of [31].

Proposition 2.3. *Let $\{\vec{X}_n\}_{n\geq 1} \in \mathbb{R}^d$ be a Markov chain such that each coordinate is independent of the other coordinates, $X_{i,n} \perp\!\!\!\perp X_{j,n}, i \neq j$. Further suppose that for two copies of the Markov chain $\{\vec{X}_n\}_{n\geq 1}$ and $\{\vec{X}'_n\}_{n\geq 1}$, $\max_{1\leq i\leq d}\|\mathcal{L}(X_{i,n}) - \mathcal{L}(X'_{i,n})\| \leq Ar^n$ for some $A \in \mathbb{R}_+$ and $r \in (0,1)$. Then,*

$$\|\mathcal{L}(\vec{X}_n) - \mathcal{L}(\vec{X}'_n)\| \leq dAr^n \tag{4}$$

The proof is in Section 6.1.

Proposition 4 of [29] proves that for any sequence of drift and minorization conditions, the geometric conver-

gence rate ρ established by the Rosenthal bound (Theorem 12 of [38]) will increase at an exponential rate for the autoregressive normal process in $\mathbb{R}^d$ as the dimension $d \to \infty$. This finding suggests that convergence bounds that use the drift and minorization condition do not scale well in dimension (see Lemma 3 and discussion in **??**). However, Proposition 2.3 shows that since each coordinate in this example is independent, the geometric convergence ρ rate is indeed invariant to dimension regardless of the bounding approach. Thus a drift and minorization bound, including the Rosenthal bound, can easily be applied the autoregressive normal process in $\mathbb{R}$ and then extended to $\mathbb{R}^d$ using Proposition 2.3. To see Proposition 2.3 applied to the autoregressive normal process in $\mathbb{R}^d$, see example 4.5.

2.2 Geometric ergodicity

In this paper, we establish convergence bounds for Markov chains that are geometrically ergodic. Geometric ergodicity is one of the conditions for guaranteeing the central limit theorem (CLT). See [9] or Section 5.2 of [32] for more details on how geometric ergodicity is related to the CLT.

Definition 2.2 (Geometric ergodicity)**.** Let X_n be a Markov chain with a stationary distribution π. The Markov chain is geometrically ergodic if there exists a $\rho < 1$ and a function $M(x) < \infty$, π-a.e. such that for $X_0 = x$,

$$\|\mathcal{L}(X_n) - \pi(\cdot)\| \leq M(x)\rho^n \tag{5}$$

The geometric rate of convergence for X_n is defined as $\rho^* = \inf\{\rho : \text{ equation 5 holds}\}$.

2.3 Coupling

To calculate an upper bound on the total variation, the coupling technique is used. The foundation of coupling is the coupling inequality followed by the Nummelin splitting technique. The former says that the total variation distance between two distributions is bounded above by the probability that they are unequal. The latter shows how the probability that they are unequal can be calculated.

Proposition 2.4 (Coupling inequality (see Section 6.1 of [37] for a proof.))**.** *The total variation of two random*

variables is bounded above by the probability that they are unequal.

$$\|\mathcal{L}(X) - \mathcal{L}(X')\| \leq P(X \neq X') \tag{6}$$

We are interested in calculating the probability that two random variables are unequal. The total variation measures the distance between two distributions, but is invariant to *how* these measures are jointly distributed. For example, let $X \sim N(0,1)$ and $X' \sim N(1,1)$ be two random variables. Regardless of whether X and X' were highly dependent, for example if $X = X' + 1$ or if X, X' were independent, their total variation distance would be the same. The Nummelin splitting technique makes use of this by constructing alternative random variables, Y and Y', such that the marginal distributions are the same $\mathcal{L}(X) = \mathcal{L}(Y)$, $\mathcal{L}(X') = \mathcal{L}(Y')$, and the probability that they are unequal is minimised. This technique was first shown in [26]. See [37] or [25] for an explanation. Finally note that the theory on maximal coupling guarantees that there exists alternative random variables Y, Y' as defined above, such that $\|\mathcal{L}(X) - \mathcal{L}(X')\| = P(Y \neq Y')$ [4].

Coupling techniques are widely used to calculate total variation upper bounds on Markov chains [35, 38, 37, 32, 36, 43]. This is because once a chain couples, in general it will remain equal forever [37], a consequence of the time invariant property of the transition distribution. The total variation at time n of two copies of a Markov chain, $\{X_n\}_{n\geq 1}$ and $\{X'_n\}_{n\geq 1}$, is bounded above by the probability that they have not coupled by time n. That is, for $T = \min\{k : X_k = X'_k\}$,

$$\|\mathcal{L}(X_n) - \mathcal{L}(X'_n)\| \leq P(T > n)$$

3 One-Shot Coupling

One-shot coupling is an alternative way of applying coupling methods to bound the total variation of two copies of a Markov chain, which was first introduced in [34]. One-shot coupling attempts to bring the two Markov chains close together, generally called the 'contracting' phase, and only tries to couple the chain at the last iteration, in the 'coalescing' phase. During the contracting phase the two copies of a Markov chain merge closer together, which is defined over some predefined metric.

To apply one-shot coupling, we define a Markov chain in terms of iterated random functions [6]. That is,

define a family of random functions $\{f_\theta : \theta \in \Theta\}$ such that θ is a random variable distributed according to μ and

$$X_n = f_{\theta_n}(X_{n-1})$$

The nth iteration of the Markov chain can be written in terms of $X_0 = x$ as follows,

$$X_n = (f_{\theta_n} \circ f_{\theta_{n-1}} \ldots \circ f_{\theta_1})(x) = f_{\theta_n}(f_{\theta_{n-1}}(\ldots f_{\theta_1}(x)\ldots))$$

To find an upper bound on the total variation distance between X_N and X'_N we do the following.

1. **Contracting phase:** For $n < N$, set $\theta_n = \theta'_n$ so that the two chains get 'closer' together.
2. **Coalescing phase:** For $n = N$, we specify $j \in \{1, \ldots, |\theta_n|\}$ and set $\theta_{i,n} = \theta'_{i,n}$ for all $i \neq j$. Assume that $j = 1$. We are then left with the random mappings $X_n = f_{(\theta_{1,n}, \theta_{-1,n})}(X_{n-1})$ and $X'_n = f_{(\theta_{1,n}, \theta'_{-1,n})}(X'_{n-1})$ where X_{n-1} and X'_{n-1} are close to each other in expectation. We apply coupling techniques to find the probability that they are equal.

The technique used in the contracting phase is also known as the common random number technique and is discussed in detail in Section 2.3.1 of [18]. The common random number technique is related to optimal transport so it is usually used to generate bounds on Wasserstein distances [18, 28, 12] (it is also used to generate bounds on other types of distances like Monge–Kantorovich or Prokhorov [18]).

The one-shot coupling method has been applied over a variety of specific examples, namely, a nested gamma model in [21], an image restoration model in [20], and a random walk on the unit sphere in [27]. The method is also used as motivation for a theorem that bounds total variation distance in terms of the Wasserstein distance [24]. This paper looks at exploring general methods for directly bounding the total variation distance using a contraction condition.

We begin with a general theorem to bound the total variation distance for a Markov chain using one-shot coupling.

Theorem 3.1 (One-Shot Coupling Theorem)**.** *Let $\{X_n\}_{n\geq 1}, \{X'_n\}_{n\geq 1}$ be two copies of a Markov chain such that $X_n = f_{\theta_n}(X_{n-1})$ and $X'_n = f_{\theta'_n}(X'_{n-1})$ where $(\theta_n, \theta'_n)_{n\geq 1}$ are independent random variables with respect to n and*

the marginal distribution of $\theta_n, \theta'_n \sim \mathcal{D}$, for some distribution $\mathcal{D}$. Suppose that the following two conditions hold for some non-negative integer n_0.

1. ***Contraction condition:*** *There exists a $D \in (0,1)$ such that for any $n \geq n_0$ when $\theta_{n+1} = \theta'_{n+1} \sim \mathcal{D}$*

$$E[|f_{\theta_{n+1}}(X_n) - f_{\theta_{n+1}}(X'_n)|] \leq DE[|X_n - X'_n|]$$

2. ***Coalescing condition:*** *There exists a $C > 0$ such that for any $n \geq n_0$*

$$||\mathcal{L}(X_{n+1}) - \mathcal{L}(X'_{n+1}))|| \leq CE[|X_n - X'_n|]$$

Then the total variation distance between the two Markov chains at iteration $n \geq n_0$ is

$$||\mathcal{L}(X_{n+1}) - \mathcal{L}(X'_{n+1})|| \leq CD^{n-n_0}E[|X_{n_0} - X'_{n_0}|]$$

Proof. Fix $n \geq n_0$. We are interested in finding an upper bound on $||\mathcal{L}(X_{n+1}) - \mathcal{L}(X'_{n+1})||$. To do so, we first generate alternative random variables, Y_n, Y'_n such that

1. for $0 \leq m \leq n_0$: $Y_m = X_m, Y'_m = X'_m$
2. for $n_0 \leq m < n$: $\theta_{m+1} = \theta'_{m+1} \sim \mathcal{D}$ and $Y_{m+1} = f_{\theta_{m+1}}(Y_m), Y'_{m+1} = f_{\theta_{m+1}}(Y'_m)$.
3. for $m = n$,: $\theta_m, \theta'_m \sim \mathcal{D}$ with an arbitrary joint distribution and $Y_{m+1} = f_{\theta_{m+1}}(Y_m), Y'_{m+1} = f_{\theta'_{m+1}}(Y'_m)$

By construction, $Y_m \stackrel{d}{=} X_m$ and $Y'_m \stackrel{d}{=} X'_m$ for $0 \leq m \leq n$.

Next we find an upper bound on the total variation distance between Y_{n+1}, Y'_{n+1}. By the contraction condition for $n_0 \leq m < n$, $E[|f_{\theta_{m+1}}(Y_m) - f_{\theta_{m+1}}(Y'_m)|] \leq DE[|Y_m - Y'_m|]$ and so,

$$\begin{aligned} E[|Y_n - Y'_n|] &= E[|f_{\theta_n}(Y_{n-1}) - f_{\theta_n}(Y'_{n-1})|] \\ &\leq DE[|Y_{n-1} - Y'_{n-1}|] \\ &\leq D^{n-n_0}E[|Y_{n_0} - Y'_{n_0}|] \end{aligned}$$

By the coalescing condition,

$$\begin{aligned}||\mathcal{L}(Y_{n+1}) - \mathcal{L}(Y'_{n+1}))|| &\leq CE[|Y_n - Y'_n|] \\ &\leq CD^{n-n_0}E[|Y_{n_0} - Y'_{n_0}|] \\ &= CD^{n-n_0}E[|X_{n_0} - X'_{n_0}|]\end{aligned}$$

Finally since $Y_n \stackrel{d}{=} X_n$ and $Y'_n \stackrel{d}{=} X'_n$,

$$||\mathcal{L}(X_{n+1}) - \mathcal{L}(X'_{n+1})||= ||\mathcal{L}(Y_{n+1}) - \mathcal{L}(Y'_{n+1})||\leq CD^{n-n_0}E[|X_{n_0} - X'_{n_0}|]$$

□

In most cases $n_0 = 0$. See the GARCH example 5.3 for an alternative case, $n_0 = 1$.

4 Random-functional autoregressive processes

The following section introduces a theorem to bound random-functional autoregressive processes, $\{X\}_{n\geq 1}$, which are of the following form for $g : \mathbb{R}^2 \to \mathbb{R}$

$$X_n = g(\theta_{1,n}, X_{n-1}) + \theta_{2,n} \tag{7}$$

where $(\theta_{1,n}, \theta_{2,n}) \in \mathbb{R}^2$ are random and $(\theta_{1,n}, \theta_{2,n}) \perp\!\!\!\perp (\theta_{1,m}, \theta_{2,m})$ when $n \neq m$.

The Sideways Theorem 4.1 provides an upper bound on the total variation distance for random-functional autoregressive processes. It is followed by a discussion on the existence of stationarity, which is not implied by the model and various examples.

Theorem 4.1 (Sideways Theorem)**.** *Let $X_n \in \mathbb{R}$ be a random-functional autoregressive model. That is, X_n is of the following form for $g : \mathbb{R}^2 \to \mathbb{R}$*

$$X_n = g(\theta_{1,n}, X_{n-1}) + \theta_{2,n} \tag{8}$$

where $(\theta_{1,n}, \theta_{2,n}) \in \mathbb{R}^2$ *and* $(\theta_{1,n}, \theta_{2,n}) \perp\!\!\!\perp (\theta_{1,m}, \theta_{2,m})$ *when* $n \neq m$*. Suppose that,*

1. ***Contraction condition:*** *There exists a* $D \in (0,1)$ *such that for* $n \geq 0$*,*

$$E[|g(\theta_{1,n+1}, X_n) - g(\theta_{1,n+1}, X'_n)|] \leq DE[|X_n - X'_n|]$$

2. ***Attributes of the conditional density*** $\theta_{2,n}|\theta_{1,n}$***:*** *The conditional density of* $\theta_{2,n}|\theta_{1,n}$

   (a) *is bounded above: There exists a* $K > 0$ *such that for all* $(\theta_{1,n}, \theta_{2,n}) \in \mathbb{R}^2$*, the conditional density function of* $\theta_{2,n}$ *is bounded above by* K*,* $f_{\theta_{2,n}}(\theta_{2,n}|\theta_{1,n}) \leq K$*.*

   (b) *has at most* M *local extrema points that are at most* $L > 0$ *distance apart: For any* $\theta_{1,n}$*, there are* M *local maximas and minimas (local extrema points) within the conditional density. The local extrema points are at most* L *distance apart.*

   (c) *is continuous for any* $\theta_{1,n}$

Then an upper bound on the geometric rate of convergence of the Markov chain is D *and the total variation distance between the two copies of the Markov chain,* X_n, X'_n*, is bounded above as follows,*

$$\|\mathcal{L}(X_{n+1}) - \mathcal{L}(X'_{n+1})\| \leq \left(\frac{K(M+1)}{2} + \frac{I_{M>1}}{L}\right) D^n E[|X_0 - X'_0|] \tag{9}$$

The attributes of the conditional density of $\theta_{2,n}|\theta_{1,n}$ serve to prove that the coalescing condition is satisfied. To prove the Sideways Theorem, we show that the contraction and coalescing conditions are satisfied and then apply the One-Shot Coupling Theorem 3.1.

Lemma 4.2 (Coalescing condition)**.** *If the density of* $\theta_{2,n}|\theta_{1,n}$ *for any* $\theta_{1,n}$ *is (1) bounded above, (2) has at most* M *local extrema points that are at most* L *distance apart, and (3) is continuous then for* $n \geq 0$*,*

$$||\mathcal{L}(X_{n+1}) - \mathcal{L}(X'_{n+1}))|| \leq CE[|X_n - X'_n|]$$

Where $C = \frac{K(M+1)}{2} + \frac{I_{M>1}}{L}$*. See Section 6.2.2 for a proof.*

Proof of Theorem 4.1. The following shows that the contraction condition holds for $D \in (0,1)$ and $n \geq 0$,

$$\begin{aligned} E[|f_{\theta_{n+1}}(X_n) - f_{\theta_{n+1}}(X'_n)|] &= E[|(g(\theta_{1,n+1}, X_n) + \theta_{2,n}) - (g(\theta_{1,n+1}, X'_n) + \theta_{2,n})|] \\ &= E[|g(\theta_{1,n+1}, X_n) - g(\theta_{1,n+1}, X'_n)|] \\ &\leq DE[|X_n - X'_n|] \qquad \text{by assumption 1} \end{aligned}$$

Lemma 4.2, which can be applied when condition 2 is satisfied (attributes of the conditional density of $\theta_{2,n}|\theta_{1,n}$), shows that the coalescing condition holds. According to the One-Shot Coupling Theorem 3.1, the total variation between two copies of the process can be bounded above using inequality 9. □

To use the Sideways Theorem 4.1 one must first find such a $D < 1$ that satisfies condition 1 and derive the conditional density of $\theta_{2,n}|\theta_{1,n}$. For many examples, conditions 2b and 2c of the Sideways Theorem 4.1 are easily verified if $\theta_{2,n}$ has a defined continuous density.

In [14], it is shown than when the function g is deterministic (g is a function of X_{n-1} only and not $\theta_{1,n}$) and given the same assumptions on $\theta_{2,n}$, the upper bound on the geometric rate of convergence is D (see corollary 8 and example 9 of [14]). This matches the results from our theorem.

Note that the Sideways Theorem 4.1 provides an upper bound on total variation distance, but does not imply the existence of a stationary distribution for the Markov chain. To develop the intuition for this, first note that convergence in total variation implies convergence in distribution [13]. Suppose that $\mathcal{L}(X_n), \mathcal{L}(X'_n)$ have distribution functions, F_n, F'_n, then by Helly's selection theorem (see Lemma 11.1.8 of [35], a right continuous function F exists such that $F_n \to F$ and $F'_n \to F$ pointwise. However, the function F may not necessarily be a distribution function. This is an illustration of why a stationary distribution may not exist.

A simple counter example would be the process $X_n = \frac{1}{2}X_{n-1} + n + Z_n, Z_n \sim N(0,1)$ where $g(\theta_{1,n}, X_n) = \frac{1}{2}X_{n-1} + n$ and $\theta_{2,n} = Z_n$. It is clear how the Sideways Theorem 4.1 could generate a geometric convergence bound over two iterations of the process if $E[X_0 - X'_0] < \infty$, but $X_n, X'_n \to \infty$ almost surely and so there is no stationary distribution.

4.1 An example of a non linear autoregressive process

Example 4.1 (Non linear autoregressive process)**.** This example is discussed in Section 4 of [28]. Let $\{X_n\}_{n\geq 1}$ be a Markov chain such that

$$X_{n+1} = \frac{1}{2}(X_n - \sin X_n) + W_n$$

where $\{W_n\}_{n\geq 1} \sim N(0,1)$ are i.i.d. random variables. In [28], it is assumed that $\{W_n\}_{n\geq 1}$ are i.i.d. random variables with mean 0 and variance 1.

For $g(x) = \frac{1}{2}(x - \sin(x))$, the derivative is $g'(x) = \frac{1}{2}(1 - \cos(x))$ and so $\sup_{x\in\mathbb{R}} g'(x) = 1$. This cannot be used. Instead, we find a value for D in terms of the second iteration. That is,

$$D^2 = \sup_{x,y} \frac{E[|X_{n+2} - X'_{n+2}|| \; X_n = x, X'_n = y]}{|x - y|}$$

Lemma 4.3. *The value of D as defined above can be written as*

$$D = \sup_{x,y} \frac{\sqrt{g(x,y)^2 + 4e^{-1/2} g(x,y) \sin f(x,y) \cos k(x,y) + 2\sin^2 f(x,y)(1 + e^{-2}(\cos^2 k(x,y) - \sin^2 k(x,y)))}}{2|x - y|}$$

where

$$f(x,y) = \frac{1}{4}(y - x + \sin x - \sin y) \qquad g(x,y) = \frac{1}{2}(x - y + \sin y - \sin x) \qquad k(x,y) = \frac{1}{4}(x + y - \sin y - \sin x)$$

The proof can be found in Section 6.3.1.

Using simulation, we can deduce that $D \approx 0.818$, which closely matches the geometric convergence rate found in [28] for the Wasserstein distance of 0.814.

Using the Sideways Theorem 4.1 notation, $K = \sqrt{\frac{2}{3\pi}}$ and $M = 1$. An upper bound on the total variation distance is

$$||\mathcal{L}(X_{n+1}) - \mathcal{L}(X'_{n+1})|| \leq \sqrt{\frac{2}{3\pi}} E[|X_0 - X'_0|] 0.669^{\lfloor N/2 \rfloor}$$

Thus if $X_0 = 1$ and $X'_0 = 2$, then after 21 iterations, the total variation between the two processes will be less than 0.01.

4.2 Random-coefficient autoregressive models

Corollary 1. *Let $X_n \in \mathbb{R}$ be a random-coefficient autoregressive model. That is, X_n is of the following form*

$$X_n = \theta_{1,n} X_{n-1} + \theta_{2,n}$$

where $(\theta_{1,n}, \theta_{2,n}) \perp\!\!\!\perp (\theta_{1,m}, \theta_{2,m})$ when $n \neq m$. If we replace condition 1 of the Sideways Theorem 4.1 with

1. *$E[|\theta_{1,n}|] < 1$*

Then equation 9 holds for $D = E[|\theta_{1,n}|]$.

Proof. If $E[|\theta_{1,n}|] < 1$ then set $D = E[|\theta_{1,n}|]$ and so the contraction condition in Theorem 4.1 holds,

$$E[|g(\theta_{1,n+1}, X_n) - g(\theta_{1,n+1}, X'_n)|] = E[|\theta_{1,n+1} X_n - \theta_{1,n+1} X'_n|] \leq D E[|X_n - X'_n|]$$

Since all of the conditions in Theorem 4.1 are satisfied, equation 9 holds. □

4.3 Bayesian regression Gibbs sampler

Example 4.2 (Bayesian regression Gibbs sampler)**.** Suppose we have the following observed data $Y \in \mathbb{R}^k$ and $X \in \mathbb{R}^{k \times p}$ where

$$Y|\beta, \sigma^2 \sim N_k(X\beta, \sigma^2 I_k)$$

for unknown parameters $\beta \in \mathbb{R}^p, \sigma^2 \in \mathbb{R}$. Suppose we apply the prior distributions on the unknown parameters,

- $\beta|\sigma^2 \sim N_p(0_p, \frac{\sigma^2}{\lambda} I_p)$, where $\lambda > 0$ is known.
- $\pi(\sigma^2) \propto 1/\sigma^2$

The Bayesian regression Gibbs sampler is based on the marginal posterior distributions of β_n, σ_n^2 and is defined as follows.

- $\beta_n|\sigma_{n-1}^2, Y \sim N_p(\tilde{\beta}, \sigma_{n-1}^2 A^{-1})$
- $\sigma_n^2|\beta_n, Y \sim \Gamma^{-1}\left(\frac{k+p}{2}, \frac{1}{2}\left[(\beta_n - \tilde{\beta})^T A(\beta_n - \tilde{\beta}) + C\right]\right)$. $\Gamma^{-1}(\alpha, \beta)$ represents the inverse gamma distribution.

Where $A = X^T X + \lambda I_p$ is positive semi-definite, $\tilde{\beta} = A^{-1} X^T Y$, and $C = Y^T (I_k - X A^{-1} X^T) Y$.

The following theorem gives an upper bound on the convergence rate of the Bayesian regression Gibbs sampler.

Theorem 4.4. *For two copies of the Bayesian regression Gibbs sampler, (β_n, σ_n) and (β'_n, σ'^2_n), defined in example 4.2,*

$$\|\mathcal{L}(\beta_{n+1}, \sigma_{n+1}) - \mathcal{L}(\beta'_{n+1}, \sigma'^2_{n+1})\| \leq KE[|\sigma_0^2 - \sigma_0'^2|] \left(\frac{p}{k+p-2}\right)^n \tag{10}$$

where $K = \frac{(C/2)^{\frac{k+2p}{2}}}{\Gamma(\frac{k+2p}{2})} \left(\frac{k+2p+2}{C}\right)^{\frac{k+2p}{2}+1} e^{-\frac{k+2p+2}{2}}$.

In Theorem 3.1 of [30] it was shown than for the equivalent example and some $0 < M_1 \leq M_2$, which are not specified,

$$M_1 \left(\frac{p}{k+p-2}\right)^n \leq \|\mathcal{L}(\beta_n, \sigma_n) - \pi\| \leq M_2 \left(\frac{p}{k+p-2}\right)^n$$

This means that the bound derived from the Sideways Theorem 1 is sharp up to a constant. The primary difference between Theorem 3.1 in [30] and the bound in Theorem 4.4 is that the latter provides explicit values for the constant, M_2 and so numerical upper bounds can be calculated.

Before proving the Theorem 4.4, we present some lemmas.

Lemma 4.5. *For the Bayesian regression Gibbs sampler, $\|\mathcal{L}(\beta_n, \sigma_n^2) - \mathcal{L}(\beta'_n, \sigma'^2_n)\| \leq \|\mathcal{L}(\sigma_n^2) - \mathcal{L}(\sigma'^2_n)\|$. The proof can be found in 6.3.2.*

Lemma 4.6 (Contraction condition)**.** *The Bayesian regression Gibbs sampler satisfies the contraction condition with $D = \left(\frac{p}{k+p-2}\right)$. The proof can be found in 6.3.2.*

Lemma 4.7 (Attributes of the conditional density $\theta_{2,n}|\theta_{1,n}$)**.** *For the Bayesian regression Gibbs sampler, $\theta_{2,n}|\theta_{1,n}$ has a continuous density, $M = 1$ and $K = \frac{(C/2)^{\frac{k+2p}{2}}}{\Gamma(\frac{k+2p}{2})} \left(\frac{k+2p+2}{C}\right)^{\frac{k+2p}{2}+1} e^{-\frac{k+2p+2}{2}}$. The proof can be found in 6.3.2.*

Given the above lemmas, the proof of Theorem 4.4 is straightforward when the Sideways Theorem is applied.

Proof of Theorem 4.4. Let $n \geq 0$.

$$\|\mathcal{L}(\beta_{n+1}, \sigma_{n+1}^2) - \mathcal{L}(\beta'_{n+1}, \sigma'^2_{n+1})\| \leq \|\mathcal{L}(\sigma_{n+1}^2) - \mathcal{L}(\sigma'^2_{n+1})\| \leq KE[|\sigma_0^2 - \sigma_0'^2|] \left(\frac{p}{k+p-2}\right)^n$$

where K is defined in Lemma 4.7. Lemma 4.5 implies the first inequality. The second inequality is a result of corollary 1, which is satisfied because of the contraction condition (Lemma 4.6) and the properties of the conditional density $\theta_{2,n}|\theta_{1,n}$ (Lemma 4.7).

□

Numerical Example 4.1 (Application of the Bayesian regression Gibbs sampler)**.** Suppose that we are interested in evaluating the delay in getting a PhD (Y), based on age, age squared, sex and whether the student has a child at home (X). For more information on this problem see [40, 41]. We want to find the upper bound on the total variation distance for a Bayesian regression Gibbs sampler on this model. In this case, there are 333 observed values ($k = 333$) and 4 covariates ($p = 4$). Using the notation from Theorem 4.4, $K = 0.0682$. Further suppose we are interested in evaluating the upper bound between two copies of the Markov chain X_n, X_n' such that $\sigma_0^2 = 1$ and $\sigma_0^{'2} = 1001$. Then,

$$\|\mathcal{L}(\beta_{n+1}, \sigma_{n+1}) - \mathcal{L}(\beta'_{n+1}, \sigma'_{n+1})\| \leq 68.16454\,(0.0119403)^n \tag{11}$$

After 3 iterations, total variation between the two chains will be less than 0.01.

4.4 Bayesian location model Gibbs sampler

Example 4.3 (Bayesian location model Gibbs sampler (found in Section 6 of [34]))**.** Suppose that we are given data points $Y_1, \ldots, Y_J \sim N(\mu, \tau^{-1})$ where μ, τ^{-1} are unknown and $J \geq 3$. Let μ, τ^{-1} have flat priors on $\mathbb{R}$ and $\mathbb{R}_+$. The Gibbs algorithm is based on the marginal posterior distributions of μ, τ^{-1} and the conditional distributions are defined as follows.

- $\mu_{n+1} = \bar{y} + Z_{n+1}/\sqrt{J\tau_n}$
- $\tau_{n+1}^{-1} = \frac{\frac{S}{2} + \frac{J}{2}(\bar{y} - \mu_{n+1})^2}{G_{n+1}}$

Where $Z_n \sim N(0,1)$ and $G_n \sim \Gamma(\frac{J+2}{2}, 1)$ are independent and $S = \sum_{i=1}^n (y_i - \bar{y})^2$. $\Gamma(\alpha, \beta)$ represents the gamma distribution, $\Gamma^{-1}(\alpha, \beta)$ represents the inverse gamma distribution.

The following theorem gives an upper bound on the convergence rate of the Bayesian location model Gibbs sampler.

Theorem 4.8. *For two copies of the Bayesian location model Gibbs sampler example 4.3,*

$$\|\mathcal{L}(\mu_{n+1}, \tau_{n+1}^{-1}) - \mathcal{L}(\mu'_{n+1}, \tau_{n+1}^{'-1})\| \leq KE[|\tau_0^{-1} - \tau_0^{'-1}|] \left(\frac{1}{J}\right)^n \tag{12}$$

where $K = \frac{(S/2)^{\frac{J-1}{2}}}{\Gamma(\frac{J-1}{2})} \left(\frac{S}{J+1}\right)^{-\frac{J-3}{2}} e^{-\frac{J+1}{2}}$.

This bound compares to the one derived in [34] which states that,

$$\|\mathcal{L}(\mu_n, \tau_n^{-1}) - \mathcal{L}(\mu'_n, \tau_n^{'-1})\| \leq \left(\frac{J}{2} + 1\right) E[|\tau_0^{-1} - \tau_0^{'-1}|] \left(\frac{1}{J}\right)^n$$

Both bounds return the same geometric rate of convergence. However, the magnitude of constant K is difficult to compare against $\left(\frac{J}{2} + 1\right)$ without knowing S. Note that the bound derived from corollary 1 is calculated in a systematic way.

Before proving Theorem 4.8, we present some lemmas.

Lemma 4.9. *For the Bayesian location model Gibbs sampler,* $\|\mathcal{L}(\mu_n, \tau_n^{-1}) - \mathcal{L}(\mu'_n, \tau_n^{'-1})\| \leq \|\mathcal{L}(\tau_n^{-1}) - \mathcal{L}(\tau_n^{'-1})\|$ *The proof can be found in 6.3.3.*

So, using the notation of corollary 1, $\theta_{1,n} = X_n Y_n$ and $\theta_{2,n} = Y_n$.

Lemma 4.10 (Contraction condition)**.** *The Bayesian location model Gibbs sampler satisfies the contraction condition with* $D = \frac{1}{J}$. *The proof can be found in 6.3.3.*

Lemma 4.11 (Attributes of the conditional density $\theta_{2,n}|\theta_{1,n}$)**.** *For the Bayesian location model Gibbs sampler,* $\theta_{2,n}|\theta_{1,n}$ *has a continuous density,* $M = 1$ *and*

$$K = \frac{(S/2)^{\frac{J-1}{2}}}{\Gamma(\frac{J-1}{2})} \left(\frac{S}{J+1}\right)^{-\frac{J-3}{2}} e^{-\frac{J+1}{2}} \tag{13}$$

The proof can be found in 6.3.3.

Given the above lemmas, the proof of Theorem 4.8 is straightforward when the Sideways Theorem is applied.

Proof of Theorem 4.8. Note that

$$\|\mathcal{L}(\mu_{n+1}, \tau_{n+1}^{-1}) - \mathcal{L}(\mu'_{n+1}, \tau_{n+1}^{-1'})\| \leq \|\mathcal{L}(\tau_{n+1}^{-1}) - \mathcal{L}(\tau_{n+1}^{-1'})\| \leq KE[|\sigma_0^2 - \sigma_0^{'2}|] \left(\frac{1}{J}\right)^n$$

where K is defined in Lemma 4.11. Lemma 4.9 implies the first inequality. The second inequality is a result of corollary 1, which is satisfied because of the contraction condition (Lemma 4.10) and the properties of the conditional density $\theta_{2,n}|\theta_{1,n}$ (Lemma 4.11).

□

Numerical Example 4.2 (Application of Bayesian location model Gibbs sampler)**.** Suppose that we are given the girth in inches of a sample of trees (see the `trees` dataset in `R`), $Y_1, \ldots, Y_{31} \sim N(\mu, \sigma^2)$, where μ, σ^2 are unknown. We want to find the upper bound on the total variation distance for the Gibbs sampler model applied to this problem. In this case the number of datapoints is 31 ($J = 31$) and using the notation from Theorem 4.8, $K = 13.74027$. Further, suppose that we are interested in evaluating the upper bound between two copies of the Markov chain, X_n, X'_n such that $\sigma_0^2 = 1$ and $\sigma_0^{'2} = 1001$. Using Theorem 4.8,

$$\|\mathcal{L}(\mu_{n+1}, \tau_{n+1}^{-1}) - \mathcal{L}(\mu'_{n+1}, \tau_{n+1}^{-1'})\| \leq 13740.27 \left(\frac{1}{31}\right)^n$$

After 6 iterations, the total variation between the two chains will be less than 0.01. This bound compares to the bound derived in [34], which states that $\|\mathcal{L}(\mu_n, \tau_n^{-1}) - \mathcal{L}(\mu'_n, \tau_n^{'-1})\| \leq 16500 \left(\frac{1}{31}\right)^n$.

4.5 Autoregressive normal process

Example 4.4 (Autoregressive normal process in $\mathbb{R}$)**.** Let $\{X_n\}_{n\geq 1} \in \mathbb{R}$ be an autoregressive normal process. Then for i.i.d. $Z_n \sim N(0, 1)$,

$$X_n = \frac{1}{2} X_{n+1} + \sqrt{\frac{3}{4}} Z_n$$

In this case $\theta_{1,n} = \frac{1}{2}$ and $\theta_{2,n} = \sqrt{\frac{3}{4}}Z_n$. $\theta_{2,n}$ has a continuous and uni-modal density function and $K = \sqrt{\frac{2}{3\pi}}$. By the Sideways Theorem 4.1,

$$\|\mathcal{L}(X_{n+1}) - \mathcal{L}(X'_{n+1})\| \leq \sqrt{\frac{2}{3\pi}} E[|X_0 - X'_0|] \left(\frac{1}{2}\right)^n \tag{14}$$

It is known that the geometric rate of convergence for the autoregressive normal process is $1/2$ [29], so once again, the Sideways Theorem 4.1 generates tight geometric convergence rates up to a constant.

When comparing the upper bound with the actual total variation distance, note that if $X_0 = x_0$ is known, $X_n \sim N(\frac{x_0}{2^n}, 1 - \frac{1}{4^n})$. Thus, the total variation distance between two copies of an autoregressive normal process X_n, X'_n where the initial values are known, $X_0 = x_0$ and $X'_0 = x'_0$, is as follows (see Section 2 of [34]),

$$\|\mathcal{L}(X_n) - \mathcal{L}(X'_n)\| = 1 - 2\Phi\left(-\frac{|x_0 - x'_0|}{2^{n+1}\sqrt{1 - \frac{1}{4^n}}}\right) \tag{15}$$

Figure 1 shows how the upper bound for the autoregressive normal process using equation 14 compares to the actual total variation distance when $x_0 = 0$ and $x'_0 = 1$. Both the total variation is less than 0.01 after 6 iteration and the upper bound on the total variation is less than 0.01 after 7 iterations.

In the following section we extend the above example to higher dimensions.

4.6 Processes in $\mathbb{R}^d$

Next we extend the autoregressive normal process as defined above to $\mathbb{R}^d$. To do so, we apply Proposition 2.3 to an autoregressive normal process in $\mathbb{R}^d$ with independent coordinates, example 4.5, and non-independent coordinates, example 4.6.

Example 4.5 (Autoregressive normal process in $\mathbb{R}^d$ with independent coordinates)**.** Let $\{\vec{X}_n\}_{n\geq 1} \in \mathbb{R}^d$ be an autoregressive normal process with independent coordinates. Then for i.i.d. $\vec{Z}_n \sim N(\vec{0}, I_d)$,

$$\vec{X}_n = \frac{1}{2}\vec{X}_{n-1} + \sqrt{\frac{3}{4}}\vec{Z}_n$$

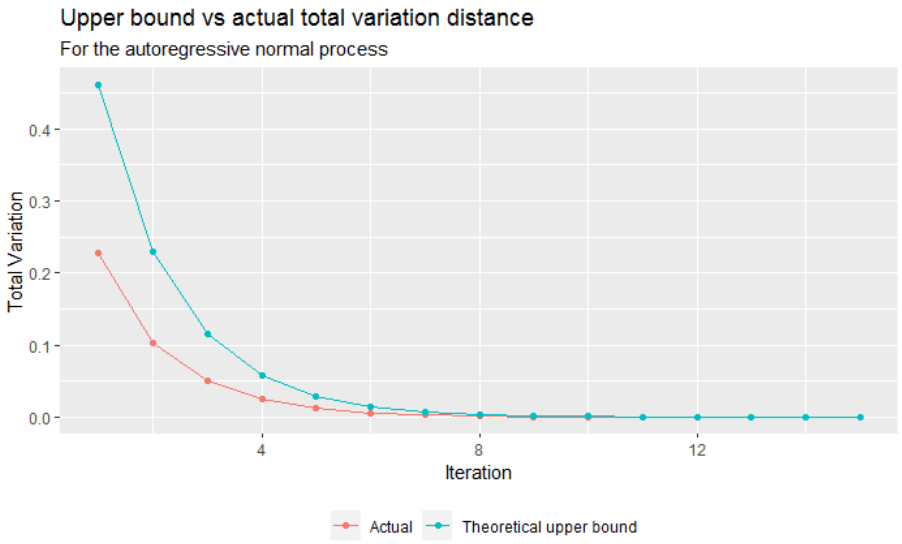

Figure 1: This figure compares the actual value of $\|\mathcal{L}(X_n) - \mathcal{L}(X'_n)\|$ against the upper bound derived from the Sideways Theorem 4.1, (equation 14) when X_n, X'_n are two copies of the autoregressive normal process (i.e. $X_n = \frac{1}{2}X_{n-1} + \sqrt{\frac{3}{4}}Z_n, Z_n \sim N(0,1)$) and $x_0 = 0$, $x'_0 = 1$.

And if $i \neq j$ then $Z_{i,n} \perp\!\!\!\perp Z_{j,n}$. Further, $X_{i,n} = \frac{1}{2}X_{i,n-1} + \sqrt{\frac{3}{4}}Z_{i,n}$ for $i \in \{1, \ldots, d\}$ and so by example 4.4,

$$\|\mathcal{L}(X_{i,n+1}) - \mathcal{L}(X'_{i,n+1})\| \leq \sqrt{\frac{2}{3\pi}} E[|X_{i,0} - X'_{i,0}|] \left(\frac{1}{2}\right)^n$$

Since each coordinate is independent and bounded above by the same value, Proposition 2.3 implies that

$$\|\mathcal{L}(\vec{X}_{n+1}) - \mathcal{L}(\vec{X}'_{n+1})\| \leq d\sqrt{\frac{2}{3\pi}} \sup_{0 \leq i \leq d} E[|X_{i,0} - X'_{i,0}|] \left(\frac{1}{2}\right)^n$$

Again, it is known that the geometric rate of convergence for the autoregressive normal process in $\mathbb{R}^d$ is $1/2$ [29]. Applying the Sideways Theorem 4.1 along with Proposition 2.3 not only returns the exact geometric rate of convergence of $1/2$, but also provides an upper bound on the total variation distance and does so in a computationally simple way.

Finally, to apply numbers to this example, suppose that $\vec{X}_n, \vec{X}'_n \in \mathbb{R}^{100}$ and the initial values of this process

are $\vec{X}_0 = (1, \ldots, 1)$ and $\vec{X}'_0 = (0, \ldots, 0)$. The total variation distance would be bounded above with $\|\mathcal{L}(\vec{X}_{n+1}) - \mathcal{L}(\vec{X}'_{n+1})\| \leq 100\sqrt{\frac{2}{3\pi}}\left(\frac{1}{2}\right)^n$. This means that at 14 iterations the total variation distance would be less than 0.01.

The following example is a more general version of the above where X_n is a general auto regressive normal process in $\mathbb{R}^d$

Example 4.6 (Autoregressive normal process in $\mathbb{R}^d$)**.** The random vector $\{\vec{X}_n\}_{n\geq 1} \in \mathbb{R}^d$ is an autoregressive normal process if for matrix A and random vector $\vec{W}_n \sim N(\vec{0}, \Sigma_d^2)$ (Σ_d^2 is a positive semi-definite matrix)

$$\vec{X}_n = A\vec{X}_{n-1} + \vec{W}_n$$

Theorem 4.12. *Suppose that A is a diagonalizable matrix. Then for two copies, $\vec{X}_n, \vec{X}'_n \in \mathbb{R}^d$, of the autoregressive normal process defined in example 4.6,*

$$\|\mathcal{L}(\vec{X}_n) - \mathcal{L}(\vec{X}'_n)\| \leq \sqrt{\frac{d}{2\pi}}\|\Sigma_d^{-1}\|_2 \cdot \|P\|_2\|P^{-1}\|_2 E[\|\vec{X}_0 - \vec{X}'_0\|_2] \max_{1\leq i\leq d}|\lambda_i|^n \tag{16}$$

where $A = PDP^{-1}$ with D as the corresponding diagonal matrix, λ_i is the ith eigenvalue of A and $\|\cdot\|_2$ denotes the Frobenius norm. The proof can be found in 6.3.4, which uses a modified version of the Sideways Theorem.

Numerical Example 4.3 (Application of the autoregressive normal process in $\mathbb{R}^d$)**.** To apply numbers to this example, suppose that $\vec{X}_n, \vec{X}'_n \in \mathbb{R}^{100}$ are two copies of the following process $\vec{X}_n = A\vec{X}_n + \vec{Z}_n, \vec{Z}_n \sim N(0, A)$ where

$$A = \begin{pmatrix} \frac{1}{2} & \frac{1}{8} & 0 & \cdots & 0 & 0 \\ \frac{1}{8} & \frac{1}{2} & \frac{1}{8} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \frac{1}{8} & \frac{1}{2} \end{pmatrix}$$

and the initial values of this process are $\vec{X}_0 = (1, \ldots, 1)$ and $\vec{X}'_0 = (0, \ldots, 0)$. The total variation distance would be bounded above with $\|\mathcal{L}(\vec{X}_n) - \mathcal{L}(\vec{X}'_n)\| \leq 98782.31\,(0.7498791)^n$. This means that after 56 iterations the total variation distance would be less than 0.01.

5 Randomly scaled iterated random functions

In this section we look at bounding the total variation norm between two copies of a process that is of the following form for $f : \mathbb{R}^2 \to \mathbb{R}$,

$$X_n = f(\theta_{1,n}, X_{n-1})\theta_{2,n}$$

Where $(\theta_{1,n}, \theta_{2,n})$ are i.i.d. random variables with respect to n.

5.1 Application of LARCH model

Example 5.1 (Linear ARCH process)**.** Let $\{X_n\}_{n\geq 1} \in \mathbb{R}$ be a linear ARCH process. Then for i.i.d. Z_n

$$X_n = (\beta_0 + \beta_1 X_{n-1})Z_n$$

See Section 7.3.3 of [7] for more details on this model.

The following theorem provides an upper bound on the convergence rate of two copies of a LARCH process.

Theorem 5.1. *Let $\{X_n\}_{n\geq 1} \in \mathbb{R}$ and $\{X'_n\}_{n\geq 1} \in \mathbb{R}$ be two copies of the linear ARCH process defined in example 5.1. Suppose that,*

- *$\beta_0, \beta_1 > 0$ and $Z_n > 0$ a.s.*
- *the density of $\log(Z_0)$ is bounded above, has at most M local maxima and minima and is continuous.*

Then, the process is geometrically ergodic if $\beta_1 E[|Z_0|] < 1$ and an upper bound on the total variation distance between the two processes is,

$$||\mathcal{L}(X_{n+1}) - \mathcal{L}(X'_{n+1})|| \leq \frac{\beta_1(M+1)}{2\beta_0} \sup_x e^x f_{Z_n}(e^x) D^n E[|X_0 - X'_0|] \tag{17}$$

Where $D = \beta_1 E[Z_0]$

Lemma 7.3.2 of [7] says that if $\beta_1 E[|Z_0|] < 1$, then a stationary distribution exists. This theorem makes an even stronger assertion that under some additional assumptions, the process will also be geometrically ergodic with geometric convergence rate $D = \beta_1 E[|Z_0|] < 1$.

To prove the theorem we first introduce two lemmas stating the contraction and coalescing condition.

Lemma 5.2 (Contraction condition)**.** *The LARCH process satisfies the contraction condition if* $D = \beta_1 E[Z_0] < 1$. *See Section 6.4.1 for a proof.*

Lemma 5.3 (Coalescing condition)**.** *Suppose that the density of* $\log(Z_0)$ *is bounded above, has at most* M *local maxima and minima and is continuous. Then the LARCH process satisfies the coalescing condition*

$$||\mathcal{L}(X_{n+1}) - \mathcal{L}(X'_{n+1})|| \leq CE[|X_n - X'_n|]$$

Where $n \geq 0$ *and* $C = \frac{\beta_1(M+1)}{2\beta_0} \sup_x e^x f_{Z_n}(e^x)$*, See Section 6.4.1 for a proof.*

Note that the density of $\log(Z_0)$ is $f_{\log(Z_0)}(x) = e^x f_{Z_0}(e^x)$.

Proof of Theorem 5.1. Suppose that the assumptions in Theorem 5.1 are satisfied. Then the LARCH model satisfies the contraction condition (Lemma 5.2) and the coalescing condition (Lemma 5.3). By the One-Shot Coupling Theorem 3.1, equation 17 holds. □

Numerical Example 5.1. We find the convergence rate of example 10.3.1 of [5], which is of the form,

$$X_n^2 = (1 + 0.5X_{n-1}^2)Z_n^2$$

Where $Z_n^2 \sim \chi^2(1)$. Further let $X_0 = 0.1$ and $X'_0 = 1.1$. The density of $\log(Z_n^2)$ is $f_{\log(Z_n^2)}(x) = (2\pi)^{-1/2} e^{(x-e^x)/2}$ and so, $\sup f_{\log(Z_n^2)}(x) = (2\pi)^{-1/2} e^{(0-e^0)/2} = \frac{1}{\sqrt{2\pi e}}$. The density of $\log(Z_n^2)$ is also unimodal, so $M = 1$. By Theorem 5.1 an upper bound on the total variation distance is

$$||\mathcal{L}(X_{n+1}) - \mathcal{L}(X'_{n+1})|| \leq \frac{1}{\sqrt{8\pi e}} 0.5^n \tag{18}$$

By 4 iterations the total variation distance will be less than 0.01. In comparison, figure 2 shows how the bound compares to a simulated estimate of the total variation distance for this process.

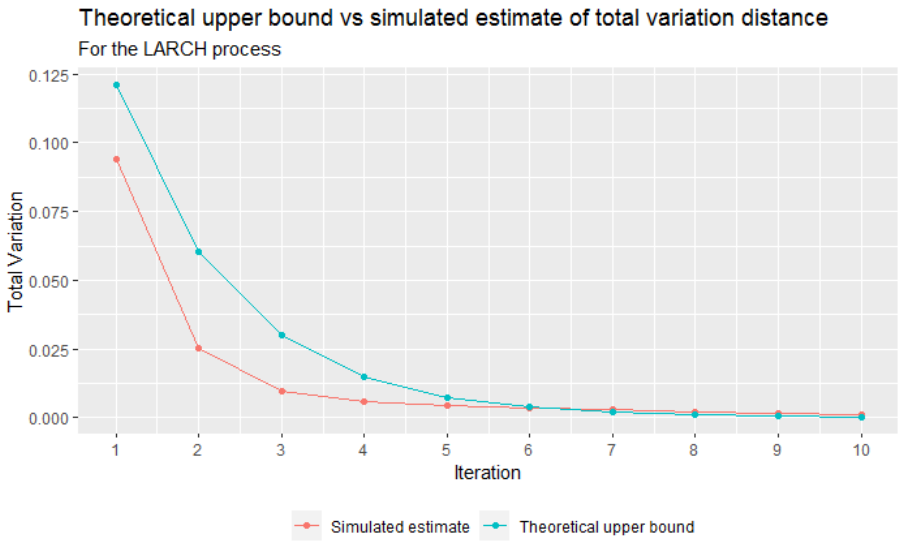

Figure 2: This figure compares a simulated approximation of $\|\mathcal{L}(X_n^2) - \mathcal{L}(X_n^{'2})\|$ against the upper bound using one-shot coupling, (equation 17). $X_n^2, X_n^{'2}$ are two copies of the LARCH process (i.e. $X_n^2 = (1 + 0.5X_{n-1}^2)Z_n^2$ and $Z_n \sim \chi^2(1)$) and $X_0^2 = 0.1, X_0^{'2} = 1.1$. To simulate total variation, 10 million simulations were run with bin length=0.01 for the estimated density function.

5.2 Application of Asymmetric ARCH model

Example 5.2 (Asymmetric ARCH process)**.** Let $\{X_n\}_n n \geq 1 \in \mathbb{R}$ be an asymmetric ARCH process. Then for i.i.d. Z_n

$$X_n = \sqrt{(aX_{n-1} + b)^2 + c^2} Z_n$$

Where $a > 0$. See exercise 4.1 of [7] for more details on this model.

The following theorem provides an upper bound on the convergence rate of two copies of an asymmetric ARCH process.

Theorem 5.4. *Let $\{X_n\}_{n\geq 1} \in \mathbb{R}$ and $\{X'_n\}_{n\geq 1} \in \mathbb{R}$ be two copies of the asymmetric ARCH process defined in example 5.2. Suppose further that the density of Z_n is centred at 0 and is monotonically decreasing around zero (i.e. $\pi(x) \geq \pi(y)$ if $|x| < |y|$.). Then, the process is geometrically ergodic if $|a|E[|Z_0|] < 1$ and an upper bound*

on the total variation distance between the two processes is

$$||\mathcal{L}(X_{n+1}) - \mathcal{L}(X'_{n+1})|| \leq \frac{|a|}{c} D^n E[|X_0 - X'_0|] \tag{19}$$

Where $D = |a|E[|Z_0|]$

Exercise 41 part 1 of [7] states that the process has a stationary solution if $D = |a|E[|Z_0|] < 1$. Theorem 5.4 shows that under certain additional assumptions on Z_n the process will also be geometrically ergodic with a specified quantitative bound.

To prove the theorem we first introduce two lemmas stating the contraction and coalescing condition.

Lemma 5.5 (Contraction condition)**.** *The asymmetric ARCH process satisfies the contraction condition if* $D = |a|E[|Z_0|] < 1$. *See Section 6.4.2 for a proof.*

Lemma 5.6 (Coalescing condition)**.** *Suppose that the density of* Z_n *is centred at 0 and is monotonically decreasing around zero. Then, the asymmetric ARCH process satisfies the coalescing condition*

$$||\mathcal{L}(X_{n+1}) - \mathcal{L}(X'_{n+1})|| \leq CE[|X_n - X'_n|]$$

where $n \geq 0$ *and* $C = \frac{|a|}{c}$. *See Section 6.4.2 for a proof.*

Proof of Theorem 5.4. Suppose that the assumptions in Theorem 5.4 are satisfied. Then the asymmetric ARCH model satisfies the contraction condition (Lemma 5.5) and the coalescing condition (Lemma 5.6). By the One-Shot Coupling Theorem 3.1, equation 19 holds. □

Numerical Example 5.2. Suppose $a = 0.5, b = 3, c = 5$, $Z_n \sim N(0,1)$ and $X_0 = 0, X'_0 = 5$. Then by Jensen's inequality, $D = 0.5E[|Z_0|] \leq 0.5E[Z_0^2]^{1/2} = 0.5$ and so by Theorem 5.4

$$||\mathcal{L}(X_n) - \mathcal{L}(X'_n)|| \leq \frac{0.5}{5} \times 5 \times 0.5^{n-1} = 0.5^n \tag{20}$$

So, by iteration $n = 7$, the total variation will be less than 0.01.

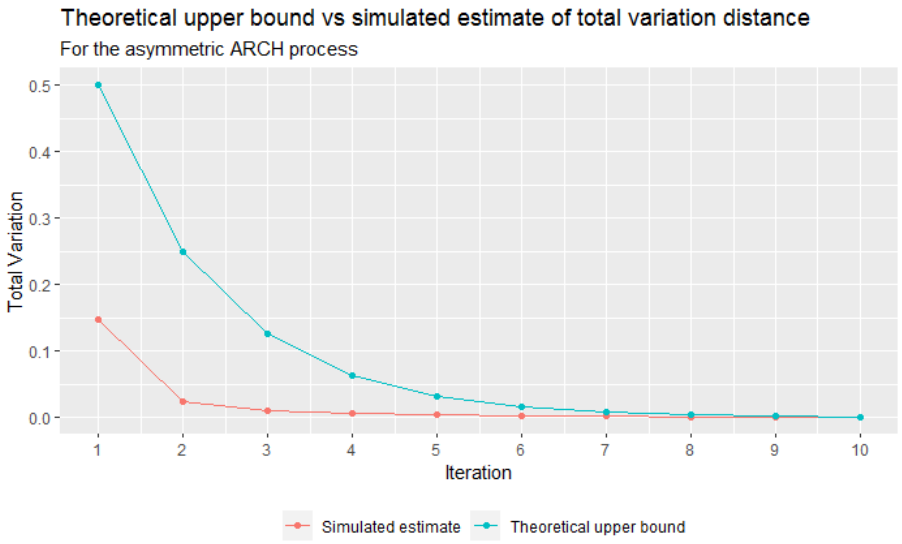

Figure 3: This figure compares a simulated approximation of $\|\mathcal{L}(X_n) - \mathcal{L}(X'_n)\|$ against the upper bound using one-shot coupling, (equation 20). X_n, X'_n are two copies of the asymmetric process (i.e. $X_n = \sqrt{(0.5X_{n-1}+3)^2 + 5^2}Z_n, Z_n \sim N(0,1)$) and $x_0 = 0$, $x'_0 = 5$. To simulate total variation, 10 million simulations were run with bin length=0.01 for the estimated density function.

In comparison, figure 4 shows how the bound compares to a simulated estimate of the total variation distance for this process.

5.3 Application of GARCH(1,1) model

Example 5.3 (GARCH(1,1) process)**.** Let $\{X_n\}_{n\geq 1} \in \mathbb{R}$ be a GARCH(1,1) process. Then for i.i.d. Z_n

$$X_n = \sigma_n Z_n$$

where,

$$\sigma_n^2 = \alpha^2 + \beta^2 X_{n-1}^2 + \gamma^2 \sigma_{n-1}^2$$

See Section 7.3.6 of [7] for more details on this model.

The following theorem provides an upper bound on the convergence rate of two copies of the GARCH(1,1)

process.

Theorem 5.7. *Let $\{X_n\}_{n\geq 1} \in \mathbb{R}$ and $\{X'_n\}_{n\geq 1} \in \mathbb{R}$ be two copies of the GARCH process defined in example 5.3. Suppose, that the density of Z_n is centered at 0 and is monotonically decreasing around zero. Then, the process is geometrically ergodic if $\beta^2 E[|Z_0^2|] + \gamma^2 < 1$. Further suppose that x_0, x'_0, σ_0^2, and $\sigma_0^{'2}$ are known. Then an upper bound on the total variation distance between the two processes is*

$$||\mathcal{L}(X_{n+1}) - \mathcal{L}(X'_{n+1})|| \leq \frac{D^n}{\alpha}\sqrt{\beta^2|x_0^2 - x_0^{'2}| + \gamma^2|\sigma_0^2 - \sigma_0^{'2}|} \tag{21}$$

Where $D = \sqrt{\beta^2 E[Z_0^2] + \gamma^2}$

To prove the theorem we first introduce three lemmas stating the contraction and coalescing conditions and and an upper bound on the total variation of the second iteration given the initial values.

Lemma 5.8 (Contraction condition)**.** *The GARCH(1,1) process satisfies the contraction condition if $D = \sqrt{\beta^2 E[Z_0^2] + \gamma^2} < 1$ See Section 6.4.3 for a proof.*

Lemma 5.9 (Coalescing condition)**.** *Suppose that the density of Z_n is centred at 0 and is monotonically decreasing around zero. Then the GARCH(1,1) process satisfies the coalescing condition,*

$$||\mathcal{L}(X_{n+1}) - \mathcal{L}(X'_{n+1})|| \leq \frac{D}{\alpha E[|Z_0|]} E[|X_n - X'_n|]$$

For $n \geq 1$, $D = \sqrt{\beta^2 E[Z_0^2] + \gamma^2}$. See Section 6.4.3 for a proof.

Lemma 5.10 (Initial condition)**.** *Suppose that we know $\sigma_0^2, \sigma_0^{'2}$ and X_0, X'_0, then*

$$E[|X_1 - X'_1|] \leq \sqrt{\beta^2|X_0^2 - X_0^{'2}| + \gamma^2|\sigma_0^2 - \sigma_0^{'2}|}E[|Z_0|]$$

See Section 6.4.3 for a proof.

Proof of Theorem 5.7. Suppose that the assumptions in Theorem 5.7 are satisfied and let $n \geq 1$. Then the GARCH(1,1) model satisfies the contraction condition (Lemma 5.8) and the coalescing condition (Lemma 5.9).

Thus by the One-Shot Coupling Theorem 3.1,

$$||\mathcal{L}(X_{n+1}) - \mathcal{L}(X'_{n+1})|| \leq \frac{D}{\alpha E[|Z_0|]} D^{n-1} E[|X_1 - X'_1|]$$

Further, by Lemma 5.10 when the initial values $\sigma_0^2, \sigma_0'^2, x_0, x'_0$ are known,

$$||\mathcal{L}(X_{n+1}) - \mathcal{L}(X'_{n+1})|| \leq \frac{D^n}{\alpha} \sqrt{\beta^2 |X_0^2 - X_0'^2| + \gamma^2 |\sigma_0^2 - \sigma_0'^2|}$$

where $D = \sqrt{\beta^2 E[Z_0^2] + \gamma^2}$ □

Numerical Example 5.3. In example 10.3.2 of [5] a GARCH(1,1) model is applied for the daily returns of the Dow Jones Industrial Index between between July 1997 and April 1999. Let

$$X_n = \sigma_n Z_n = \text{excess daily return of the Dow Jones Industrial Index at time } n$$

The following is the fitted GARCH volatility estimates when $Z_n \sim N(0,1)$,

$$\sigma_n^2 = 0.13000 + 0.1266 X_{n-1}^2 + 0.7922 \sigma_{n-1}^2$$

Suppose that we want to find the total variation of the fitted process with varying initial values representing two market states, $X_0 = 0.1, \sigma_0 = 0.01$ and $X'_0 = -0.1, \sigma'_0 = 0.1$ Then by Theorem 5.7,

$$||\mathcal{L}(X_{n+1}) - \mathcal{L}(X'_{n+1})|| \leq \sqrt{\frac{0.7922 |0.01^2 - 0.1^2|}{0.13}} D^n \approx 0.2456 D^n \tag{22}$$

Where $D = \sqrt{0.1266 + 0.7922} = \sqrt{0.9188}$

By iteration 77 the total variation distance between the two processes will be less than 0.01. In comparison, figure 4 shows how the bound compares to a simulated estimate of the total variation distance for this process. The actual total variation distance appears to be much smaller than the upper bound.

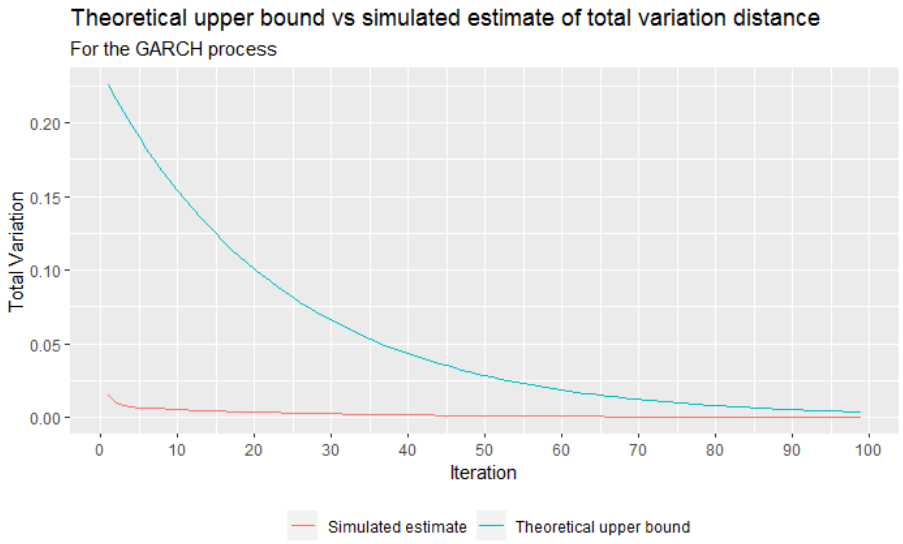

Figure 4: This figure compares a simulated approximation of $\|\mathcal{L}(X_n) - \mathcal{L}(X'_n)\|$ against the upper bound using one-shot coupling (equation 22). X_n, X'_n are two copies of the asymmetric process (i.e. $X_n = \sigma_n Z_n$ and $\sigma_n^2 = 0.13000 + 0.1266X_{n-1}^2 + 0.7922\sigma_{n-1}^2$ and $Z_n \sim N(0,1)$) and $X_0 = 0.1, \sigma_0 = 0.01$ and $X'_0 = -0.1, \sigma'_0 = 0.1$. To simulate total variation, 1 million simulations were run with bin length=0.01 for the estimated density function.

6 Appendix

6.1 Propositions related to the properties of total variation distance

Proof of Proposition 2.1. Let $\mathcal{A}$ be the sigma field of $\mathcal{X}$ and $\mathcal{B}$ be the sigma field of $\mathcal{Y}$.

First note that $f^{-1}(\mathcal{B}) = \{f^{-1}(B) : B \in \mathcal{B}\} = \mathcal{A}$:

- $f^{-1}(\mathcal{B}) \subset \mathcal{A}$**:** For $B \in \mathcal{B}$, $f^{-1}(B) \subset \mathcal{A}$ by measurability.
- $\mathcal{A} \subset f^{-1}(\mathcal{B})$**:** Let $A \in \mathcal{A}$. Then $f(A) \in \mathcal{B}$ and $f^{-1}(f(A)) \in f^{-1}(\mathcal{B})$ by definition. By invertibility, $f^{-1}(f(A)) = A$ and so $A \in f^{-1}(\mathcal{B})$.

The equality in equation 3 can then be proven as follows,

$$\begin{aligned}
\|\mathcal{L}(f(X)) - \mathcal{L}(f(X'))\| &= \sup_{B \in f(\mathcal{B})} |P(f(X) \in B) - P(f(X') \in B)| \\
&= \sup_{B \in f(\mathcal{B})} |P(X \in f^{-1}(B)) - P(X' \in f^{-1}(B))| \\
&= \sup_{A \in \mathcal{A}} |P(X \in A) - P(X' \in A)| && \text{Since } f^{-1}(\mathcal{B}) = \mathcal{A} \\
&= \|\mathcal{L}(X) - \mathcal{L}(X')\|
\end{aligned}$$

□

Proof of Proposition 2.2.

$$\begin{aligned}
\|\mathcal{L}(X) - \mathcal{L}(X')\| &= \sup_{A \in \mathcal{B}} |P(X \in A) - P(X' \in A)| \\
&= \sup_{A \in \mathcal{B}} |\int_{\mathcal{Y}} P(X \in A|y) - P(X' \in A|y)\mu(dy)| \\
&\leq \sup_{A \in \mathcal{B}} \int_{\mathcal{Y}} |P(X \in A|y) - P(X' \in A|y)|\mu(dy) && \text{by Jensen's inequality} \\
&\leq \int_{\mathcal{Y}} \sup_{A \in \mathcal{B}} |P(X \in A|y) - P(X' \in A|y)|\mu(dy) \\
&\leq E\left[\|\mathcal{L}(X|Y) - \mathcal{L}(X'|Y)\|\right]
\end{aligned}$$

□

Proof of Proposition 2.3. To prove this we use the concept of maximal coupling over the coordinates. By maximal coupling, for $i \in \{1, \ldots, d\}$ there exists random variables $X_{i,n}^M, X_{i,n}^{'M}$ such that $X_{i,n} \stackrel{d}{=} X_{i,n}^M$ and $X'_{i,n} \stackrel{d}{=} X_{i,n}^{'M}$ and

$$\|\mathcal{L}(X_{i,n}) - \mathcal{L}(X'_{i,n})\| = P(X_{i,n}^M \neq X_{i,n}^{'M})$$

(see Proposition 3g of [32] or Section 2 of [4]).

Further, there exists a unique product measure such that for any $A_1, \ldots A_d \in \mathcal{B}$, $P(\cap_{i=1}^d [X_{i,n}^M \in A_i]) =$

$\prod_{i=1}^{d} P(X_{i,n}^M \in A_i)$ (theorem 18.2 of [3]). For the unique product measure, the following equality holds,

$$P(\cap_{i=1}^{d} X_{i,n}^M \in A_i) = \prod_{i=1}^{d} P(X_{i,n}^M \in A_i) = \prod_{i=1}^{d} P(X_{i,n} \in A_i) = P(\cap_{i=1}^{d} X_{i,n} \in A_i)$$

And so by uniqueness, for $A \in \mathcal{B}^{\mathrm{d}}$, $P(X_n^M \in A) = P(X_n \in A)$. By definition this means that $\vec{X}_n \stackrel{d}{=} \vec{X}_n^M$, which implies that $(\vec{X}_n^M, \vec{X}_n^{\prime M}) \in \mathcal{C}(\vec{X}_n, \vec{X}_n')$, the set of all couplings of $\vec{X}_n, \vec{X}_n'$.

We now use $\vec{X}_n^M, \vec{X}_n^{\prime M}$ to prove equation 4.

$$\begin{aligned}
\|\mathcal{L}(\vec{X}_n) - \mathcal{L}(\vec{X}_n')\| &= \inf_{\vec{Y}, \vec{Y}' \in \mathcal{C}(\vec{X}_n, \vec{X}_n')} P(\vec{Y} \neq \vec{Y}') && \text{by equation 2.4 of [4]} \\
&\leq P(\vec{X}_n^M \neq \vec{X}_n^{\prime M}) \\
&= P(\cup_{i=1}^{d} [X_{i,n}^M \neq X_{i,n}^{\prime M}]) \\
&\leq \sum_{i=1}^{d} P(X_{i,n}^M \neq X_{i,n}^{\prime M}) && \text{by subadditivity} \\
&\leq dAr^n
\end{aligned}$$

□

6.2 Lemmas related to the Sideways Theorem

The following are lemmas and corresponding proofs and corollaries related to the Sideways Theorem (4.1).

6.2.1 Lemmas providing an upper bound on the integral difference between a function and a corresponding shift

The following lemmas are used in the proof of Lemma 4.2.

Lemma 6.1. *For any invertible, continuous function $f : \mathbb{R} \to \mathbb{R}$ where the codomain is $f(\mathbb{R}) = (a, b)$ and $\Delta > 0$,*

$$\int_{\mathbb{R}} |f(x + \Delta) - f(x)| dx = (b - a)\Delta$$

Proof. Since f is invertible and continuous, it is strictly monotone (Lemma 3.8 if [15]). Assume that f is strictly increasing. The integral can be written as follows,

$$\begin{aligned}
\int_{\mathbb{R}} |f(x+\Delta) - f(x)|dx &= \int_{\mathbb{R}} f(x+\Delta) - f(x)dx \\
&= \int_{\mathbb{R}} \int_a^b I_{f(x+\Delta)<y<f(x)} dydx \\
&= \int_{\mathbb{R}} \int_a^b I_{f^{-1}(y)-\Delta<x<f^{-1}(y)} dydx \\
&= \int_a^b \int_{\mathbb{R}} I_{f^{-1}(y)-\Delta<x<f^{-1}(y)} dxdy \qquad \text{by Fubini's theorem} \\
&= \int_a^b \Delta dy \\
&= (b-a)\Delta
\end{aligned}$$

If f is strictly decreasing apply the transform $h(x) = a + b - f(x)$. The function h is a strictly increasing invertible function with codomain (a, b) and so using the previous result for increasing functions,

$$\int_{\mathbb{R}} |f(x+\Delta) - f(x)|dx = \int_{\mathbb{R}} |h(x+\Delta) - h(x)|dx = (b-a)\Delta$$

□

Lemma 6.2. *Let $f : \mathbb{R} \to \mathbb{R}$ be a continuous function that is invertible over the set (c, d) and is a constant function over $(c, d)^C$. Further suppose that the codomain is $f(\mathbb{R}) = (a, b)$. Then for $\Delta > 0$ we get that*

$$\int_{\mathbb{R}} |f(x+\Delta) - f(x)|dx = (b-a)\Delta$$

Proof. Assume that f is an increasing function and so $f(c) = a$, $f(d) = b$ and $|f(x+\Delta) - f(x)| = f(x+\Delta) - f(x)$.

Let $0 < \epsilon < (c-d)/2$ and define

$$g_\epsilon(x) = \begin{cases} (f(c+\epsilon)-a)(1-e^{x-c-\epsilon})+a & \text{when } x \in (-\infty, c+\epsilon] \\ f(x) & \text{when } x \in (c+\epsilon, d-\epsilon] \\ (f(d-\epsilon)-b)(1-e^{d-\epsilon-x})+b & \text{when } x \in (d-\epsilon, \infty) \end{cases}$$

Note that $g_\epsilon(x)$ is continuous, invertible, an increasing function and the codomain is (a,b). By Lemma 6.1 for each $\epsilon > 0$

$$\int_{\mathbb{R}} g_\epsilon(x+\Delta) - g_\epsilon(x)dx = (b-a)\Delta$$

Further, for all $x \in \mathbb{R}$, $\lim_{\epsilon \to 0} g_\epsilon(x+\Delta) - g_\epsilon(x) = f(x+\Delta) - f(x)$ and so $g_\epsilon(x+\Delta) - g_\epsilon(x)$ converges pointwise to $f(x+\Delta) - f(x)$. Next, for $0 < \epsilon < (c-d)/2$, $|g_\epsilon(x+\Delta) - g_\epsilon(x)| < 2|b|$ and so the function $g_\epsilon(x+\Delta) - g_\epsilon(x)$ is uniformly bounded. The above statements allow us to apply the dominated convergence Theorem (theorem 16.5 of [3]) and so

$$\int_{\mathbb{R}} f(x+\Delta) - f(x)dx = \lim_{\epsilon \to 0} \int_{\mathbb{R}} g_\epsilon(x+\Delta) - g_\epsilon(x)dx = (b-a)\Delta$$

If f is strictly decreasing apply the transform $h(x) = a + b - f(x)$. The function h is a strictly increasing invertible function with codomain (a,b) and so using the previous result for increasing functions,

$$\int_{\mathbb{R}} |f(x+\Delta) - f(x)|dx = \int_{\mathbb{R}} |h(x+\Delta) - h(x)|dx = (b-a)\Delta$$

□

Lemma 6.3. *Let $f : \mathbb{R} \to \mathbb{R}$ be a continuous function with the following properties:*

- *the codomain is $(0, K)$*
- *$(m_1, m_2, \ldots, m_M)$ are the local maxima and minima points*
- *$\lim_{x \to \infty} f(x) = 0$ and $\lim_{x \to -\infty} f(x) = 0$*

Further suppose that $\Delta < \max_{i=2,\ldots,M}\{m_i - m_{i-1}\}$. *Then*

$$\int_{\mathbb{R}} |f(x-\Delta) - f(x)|dx \leq K(M+1)\Delta$$

Proof. Since $\Delta < \max_{i=2,\ldots,M}\{m_i - m_{i-1}\}$, we have that $m_1 - \Delta < m_1 < m_2 - \Delta < \ldots < m_M$. Let $I_1, \ldots, I_M$ be the intersection points or the points where $f(I_i) = f(I_i - \Delta)$.

Show that $m_i - \Delta < I_i < m_i$**:** Suppose that m_i is a local maximum point. Let $g(x) = f(x + \Delta)$. Within the interval $(m_i - \Delta, m_i)$, $f'(x) > 0$ and $g'(x) < 0$ by assumption. This implies that $f(m_i - \Delta) < f(m_i)$ and $g(m_i - \Delta) > g(m_i)$ by the mean value theorem. Further since $g(m_i - \Delta) = f(m_i)$ we have that $g(m_i - \Delta) > f(m_i - \Delta)$ and $g(m_i) < f(m_i)$.

Let $h(x) = g(x) - f(x)$. Then $h(m_i - \Delta) > 0$ and $h(m_i) < 0$ further h is a strictly decreasing function over $(m_i - \Delta, m_i)$ since $g, -f$ are strictly decreasing functions over the same interval. So by the intermediate value theorem, there exists an $\xi \in (m_i - \Delta, m_i)$ such that $h(\xi) = 0$ or $f(\xi) = g(\xi) = f(\xi + \Delta)$. Further by injectivity, ξ is unique. Let $I_i = \xi$. A similar proof can be given for when m_i is a local minimum.

Show that $\int_{I_i}^{I_{i+1}} |f(x+\Delta) - f(x)|dx \leq K\Delta$**:** Note first that $m_i - \Delta < I_i < m_i < m_{i+1} - \Delta < I_{i+1} < m_{i+1}$ further define

$$f_i(x) = \begin{cases} f(m_i) & \text{when } x \in (-\infty, m_i] \\ f(x) & \text{when } x \in (m_i, m_{i+1}] \\ f(m_{i+1}) & \text{when } x \in (m_{i+1}, \infty) \end{cases}$$

Note that over the interval $(m_i, m_{i+1}]$, the function f is either a strictly increasing or a strictly decreasing function.

$$
\begin{aligned}
&\int_{I_i}^{I_{i+1}} |f(x+\Delta) - f(x)|dx \\
&= \int_{I_i}^{m_i} |f(x+\Delta) - f(x)|dx + \int_{m_i}^{m_{i+1}-\Delta} |f(x+\Delta) - f(x)|dx + \int_{m_{i+1}-\Delta}^{I_{i+1}} |f(x+\Delta) - f(x)|dx \\
&\leq \int_{I_i}^{m_i} |f(x+\Delta) - f(m_i)|dx + \int_{m_i}^{m_{i+1}-\Delta} |f(x+\Delta) - f(x)|dx + \int_{m_{i+1}-\Delta}^{I_{i+1}} |f(m_{i+1}) - f(x)|dx \\
&= \int_{I_i}^{m_i} |f_i(x+\Delta) - f_i(x)|dx + \int_{m_i}^{m_{i+1}-\Delta} |f_i(x+\Delta) - f_i(x)|dx + \int_{m_{i+1}-\Delta}^{I_{i+1}} |f_i(x+\Delta) - f_i(x)|dx \\
&= \int_{I_i}^{I_{i+1}} |f_i(x+\Delta) - f_i(x)|dx \\
&\leq \int_{m_i-\Delta}^{m_{i+1}} |f_i(x+\Delta) - f_i(x)|dx \\
&= \int_{\mathbb{R}} |f_i(x+\Delta) - f_i(x)|dx \\
&= |f(m_i) - f(m_{i+1})|\Delta \leq K\Delta
\end{aligned}
$$

The last equality is a result of Lemma 6.2.

By similar reasoning it can be shown that

$$\int_{-\infty}^{I_1} |f(x+\Delta) - f(x)|dx \leq K\Delta \qquad \int_{I_M}^{\infty} |f(x+\Delta) - f(x)|dx \leq K\Delta$$

Finally note that the intersection points partition $\mathbb{R}$ into $M+1$ subsets and so

$$\int_{\mathbb{R}} |f(x-\Delta) - f(x)|dx \leq K(M+1)\Delta$$

□

6.2.2 Proof of Lemma 4.2

Lemma 4.2 represents the coalescing condition for the Sideways Theorem 4.1.

Proof of Lemma 4.2. Set $\theta_{1,n} = \theta'_{1,n}$. Define

$$\Delta = g(\theta_{1,n}, X_{n-1}) - g(\theta_{1,n}, X'_{n-1})$$

Let $f_{X_n}, f_{X'_n}$ be the density functions for X_n, X'_n, respectively and $f_{\theta_{2,n}}, f_{\theta_{2,n}+\Delta}$ be the density functions for $\theta_{2,n}, \theta_{2,n} + \Delta$.

Suppose that $\Delta, X_{n-1}, X'_{n-1} \in \mathbb{R}$ are known and so,

$$X_n = g(\theta_{1,n}, X_{n-1}) + \theta_{2,n} \implies \theta_{2,n} = X_n - g(\theta_{1,n}, X_{n-1})$$
$$X'_n = g(\theta_{1,n}, X'_{n-1}) + \theta'_{2,n} \implies \theta'_{2,n} - \Delta = X'_n - g(\theta_{1,n}, X_{n-1})$$

We know that $\theta_{2,n} \overset{d}{=} \theta'_{2,n}$ and in general Δ, $\theta_{1,n}$ are random variables, so

$$\|\mathcal{L}(X_n) - \mathcal{L}(X'_n)\| \leq E_{\theta_{1,n},\Delta}\left[\|\mathcal{L}(X_n|\theta_{1,n}, \Delta) - \mathcal{L}(X'_n|\theta_{1,n}, \Delta)\|\right] \qquad \text{by Proposition 2.2} \qquad (23)$$
$$= E_{\theta_{1,n},\Delta}\left[\|\mathcal{L}(\theta_{2,n}|\theta_{1,n}) - \mathcal{L}(\theta_{2,n} - \Delta|\theta_{1,n})\|\right] \qquad \text{by Proposition 2.1} \qquad (24)$$

By the assumptions in the theorem, the density of $\theta_{2,n}$ is continuous with M extrema points and has a codomain that is in $(0, K)$. Let $(m_1, m_2, \ldots, m_M)$ be the local extrema points where $m_i < m_j$ if $i < j$ and $L \leq \max_{2 \leq i \leq M}\{m_i - m_{i-1}\}$ be the maximum distance between two local extrema points. So, continuing from

the inequality 23 and by the definition of total variation, equation 2,

$$\begin{aligned}
\|\mathcal{L}(X_n) - \mathcal{L}(X'_n)\| &\leq E_{\theta_{1,n}}\left[E_\Delta\left[\frac{1}{2}\int_{\mathbb{R}}|f_{\theta_{2,n}}(x|\theta_{1,n}) - f_{\theta_{2,n}-\Delta}(x|\theta_{1,n})|dx\right]\right] \\
&= E_{\theta_{1,n}}\left[E_\Delta\left[\frac{1}{2}\int_{\mathbb{R}}|f_{\theta_{2,n}}(x|\theta_{1,n}) - f_{\theta_{2,n}}(x+\Delta|\theta_{1,n})|dx\right]\right] \\
&= E_{\theta_{1,n}}\left[E_\Delta\left[\frac{1}{2}\int_{\mathbb{R}}|f_{\theta_{2,n}}(x|\theta_{1,n}) - f_{\theta_{2,n}}(x+\Delta|\theta_{1,n})|dx I_{\Delta<L}\right]\right] + \\
&\qquad E_{\theta_{1,n}}\left[E_\Delta\left[\frac{1}{2}\int_{\mathbb{R}}|f_{\theta_{2,n}}(x|\theta_{1,n}) - f_{\theta_{2,n}}(x+\Delta|\theta_{1,n})|dx I_{\Delta>L}\right]\right] \\
&\leq E_{\theta_{1,n}}\left[E_\Delta\left[\frac{1}{2}\int_{\mathbb{R}}|f_{\theta_{2,n}}(x|\theta_{1,n}) - f_{\theta_{2,n}}(x+\Delta|\theta_{1,n})|dx\,\middle|\,|\Delta|<L\right]\right] + P_\Delta(|\Delta|>L) \\
&\leq \frac{1}{2}E_{\theta_{1,n}}\left[E_\Delta\left[K(M+1)|\Delta|\right]\right] + P_\Delta(|\Delta|>L) \qquad \text{by Lemma 6.3} \\
&\leq \frac{K(M+1)}{2}E_\Delta\left[|\Delta|\right] + \frac{E_\Delta[|\Delta|]}{L}
\end{aligned}$$

The coalescing condition is thus satisfied as follows with $C = \frac{K(M+1)}{2} + \frac{I_{M>1}}{L}$,

$$\begin{aligned}
||\mathcal{L}(X_{n+1}) - \mathcal{L}(X'_{n+1}))|| &\leq CE[|g(\theta_{1,n}, X_{n-1}) - g(\theta_{1,n}, X'_{n-1})|] \\
&= CE[|g(\theta_{1,n}, X_{n-1}) + \theta_{2,n} - (g(\theta_{1,n}, X'_{n-1}) + \theta_{2,n})|] \\
&= CE[|X_n - X'_n|]
\end{aligned}$$

□

6.3 Lemmas for random-functional autoregressive processes examples

6.3.1 Proof of Lemma 4.3

Proof of Lemma 4.3. First note that

$$
\begin{aligned}
&E[|X_{n+2} - X'_{n+2}|| \; X_n = x, X'_n = y] \\
&= E\left[\left| g\left(\frac{1}{2}(x - \sin x) + W_n\right) - g\left(\frac{1}{2}(y - \sin y) + W_n\right)\right|\right] \\
&= E\left[\left|\frac{1}{2}\left(\frac{1}{2}(x - \sin x) + W_n - \sin\left(\frac{1}{2}(x - \sin x) + W_n\right)\right) - \frac{1}{2}\left(\frac{1}{2}(y - \sin y) + W_n - \sin\left(\frac{1}{2}(y - \sin y) + W_n\right)\right)\right|\right] \\
&= \frac{1}{2}E\left[\left|\frac{1}{2}(x - y + \sin y - \sin x) + \sin\left(\frac{1}{2}(y - \sin y) + W_n\right) - \sin\left(\frac{1}{2}(x - \sin x) + W_n\right)\right|\right] \\
&= \frac{1}{2}E\left[|g(x,y) + G(x,y)|\right]
\end{aligned}
$$

Where $g(x,y) = \frac{1}{2}(x - y + \sin y - \sin x)$ and $G(x,y) = \sin\left(\frac{1}{2}(y - \sin y) + W_n\right) - \sin\left(\frac{1}{2}(x - \sin x) + W_n\right)$. By trigonometric identities, for $k(x,y) = \frac{x+y-\sin y-\sin x}{4}$ and $f(x,y) = \frac{y-x+\sin x-\sin y}{4}$.

$$
\begin{aligned}
G(x,y) &= 2\cos\left(\frac{x + y - \sin y - \sin x}{4} + W_n\right)\sin\left(\frac{y - x + \sin x - \sin y}{4}\right) \\
&= 2\cos\left(k(x,y) + W_n\right)\sin f(x,y) \\
&= 2\sin f(x,y)\left(\cos W_n \cos k(x,y) + \sin W_n \sin k(x,y)\right)
\end{aligned}
$$

And so for $k(x,y) = \frac{x+y-\sin y-\sin x}{4}$ and $f(x,y) = \frac{y-x+\sin x-\sin y}{4}$,

$$
\begin{aligned}
&E[|X_{n+2} - X'_{n+2}|| \; X_n = x, X'_n = y] \\
&= \frac{1}{2}E\left[|g(x,y) + 2\sin f(x,y)\left(\cos W_n \cos k(x,y) + \sin W_n \sin k(x,y)\right)|\right] \\
&\leq \frac{1}{2}\sqrt{E\left[\left(g(x,y) + 2\sin f(x,y)\left(\cos W_n \cos k(x,y) + \sin W_n \sin k(x,y)\right)\right)^2\right]} \\
&= \frac{1}{2}\sqrt{g(x,y)^2 + 4e^{-1/2}g(x,y)\sin f(x,y)\cos k(x,y) + 4\sin^2 f(x,y)E[\left(\cos W_n \cos k(x,y) + \sin W_n \sin k(x,y)\right)^2]} \\
&= \frac{1}{2}\sqrt{g(x,y)^2 + 4e^{-1/2}g(x,y)\sin f(x,y)\cos k(x,y) + 2\sin^2 f(x,y)(1 + e^{-2}(\cos^2 k(x,y) - \sin^2 k(x,y)))}
\end{aligned}
$$

□

6.3.2 Proof of lemmas used in Theorem 4.4

To prove the first part of this theorem, we apply the de-initialization technique which shows how the convergence rate of a Markov chain can be bounded above by the convergence rate of a more simpler Markov chain that includes sufficient information on the Markov chain of interest. The concept of de-initialization and a proposition that bounds total variation is provided below.

Definition 6.1 (De-initialisation)**.** Let $\{X_n\}_{n\geq 1}$ be a Markov chain. A Markov chain $\{Y_n\}_{n\geq 1}$ is a de-initialization of $\{X_n\}_{n\geq 1}$ if for each $n \geq 1$

$$\mathcal{L}(X_n|X_0, Y_n) = \mathcal{L}(X_n|Y_n)$$

Proposition 6.4 (Theorem 1 of [33])**.** *Let $\{Y_n\}_{n\geq 1}$ be a de-initialization of $\{X_n\}_{n\geq 1}$ then for any two initial distributions $X_0 \sim \mu$ and $X_0' \sim \mu'$,*

$$\|\mathcal{L}(X_n) - \mathcal{L}(X_n')\| \leq \|\mathcal{L}(Y_n) - \mathcal{L}(Y_n')\|$$

Proof of Lemma 4.5. Since $\beta_n = \tilde{\beta} + \sigma_n Z_n, Z_n \sim N_p(0, A^{-1})$ can be written as a random function of σ_n^2,

$$\mathcal{L}(\beta_n, \sigma_n^2|\beta_0, \sigma_0^2, \sigma_n^2) = \mathcal{L}(\beta_n, \sigma_n^2|\sigma_n^2)$$

and so σ_n^2 is a de-initialization of (β_n, σ_n^2). By Proposition 6.4,

$$\|\mathcal{L}(\beta_n, \sigma_n^2) - \mathcal{L}(\beta_n', \sigma_n^{'2})\| \leq \|\mathcal{L}(\sigma_n^2) - \mathcal{L}(\sigma_n^{'2})\|$$

We are thus interested in evaluating the convergence rate of σ_n^2 to bound the convergence rate of (β_n, σ_n^2).

To interpret this in another way, if σ_n^2 couples then the distribution of β_n is the same for both iterations, so it is automatically coupled. An alternative proof can be made using the results from [23]. □

Lemma 6.5. *For the Bayesian regression Gibbs sampler, $\sigma_n^2 = X_n Y_n \sigma_{n-1}^2 + Y_n$ where $X_n \sim \Gamma\left(\frac{p}{2}, \frac{C}{2}\right)$ and*

$Y_n \sim \Gamma^{-1}\left(\frac{k+p}{2}, \frac{C}{2}\right)$. *$\Gamma(\alpha, \beta)$ represents the gamma distribution and $\Gamma^{-1}(\alpha, \beta)$ represents the inverse gamma distribution.*

Proof. We can write σ_n^2 as an autoregressive process with independent $Z_n^2 \sim \chi^2(p)$ and $G_n \sim \Gamma(\frac{k+p}{2}, 1)$ as $\sigma_n^2 = \frac{Z_n^2}{C}\frac{C}{2G_n}\sigma_{n-1}^2 + \frac{C}{2G_n}$.

Let $X_n = \frac{Z_n^2}{C}$, $Y_n = \frac{C}{2G_n}$. We can rewrite $\sigma_n^2 = X_n Y_n \sigma_{n-1}^2 + Y_n$ where $X_n \sim \Gamma\left(\frac{p}{2}, \frac{C}{2}\right)$ and $Y_n \sim \Gamma^{-1}\left(\frac{k+p}{2}, \frac{C}{2}\right)$. Using the notation from the Sideways Theorem 4.1 $\theta_{1,n} = X_n Y_n$ and $\theta_{2,n} = Y_n$. □

Proof of Lemma 4.6. By Lemma 6.5, $\theta_{1,n} = X_n Y_n$ and so,

$$K = E[|\theta_{1,n}|] = E[X_n Y_n] = E[X_n]E[Y_n] = \frac{p}{C}\frac{C}{k+p-2} = \frac{p}{k+p-2}$$

□

Proof of Lemma 4.7. **Calculate the conditional density $\theta_{2,n}|\theta_{1,n}$** We remove the subscript n on the random variables. Let X, Y be as described in Lemma 6.5. Since the random variables are independent, the joint density is the product of the densities.

$$f_{X,Y}(x,y) = \frac{C/2}{\Gamma(p/2)}x^{p/2-1}e^{xC/2}\frac{C/2}{\Gamma((k+p)/2)}y^{-(k+p)/2-1}e^{-\frac{C/2}{y}} \tag{25}$$

Then $(\theta_1, \theta_2) = (XY, Y)$ is a transformation with the Jacobian $|J| = \theta_2^{-1}$ and the density written as follows,

$$\begin{aligned} f_{\theta_1,\theta_2}(\theta_1,\theta_2) &= f_{X,Y}\left(\frac{\theta_1}{\theta_2}, \theta_2\right)\theta_2^{-1} \\ &= \frac{C/2}{\Gamma(p/2)}\left(\frac{\theta_1}{\theta_2}\right)^{p/2-1} e^{-\frac{\theta_1}{\theta_2}C/2}\frac{C/2}{\Gamma((k+p)/2)}\theta_2^{-(k+p)/2-1}e^{-\frac{C/2}{\theta_2}}\theta_2^{-1} \end{aligned}$$

Next $f_{\theta_2|\theta_1}(\theta_2|\theta_1)$ is proportional to $f_{\theta_1,\theta_2}(\theta_1,\theta_2)$ and so we can derive the conditional density of θ_2 as follows,

$$f_{\theta_2|\theta_1}(\theta_2|\theta_1) \propto f_{\theta_1,\theta_2}(\theta_1,\theta_2) \tag{26}$$

$$\propto \theta_2^{1-p/2} e^{-\frac{1}{\theta_2}\theta_1 C/2} \theta_2^{-(k+p)/2-1} e^{-\frac{1}{\theta_2}C/2} \theta_2^{-1} \tag{27}$$

$$= \theta_2^{-(p/2+(k+p)/2)-1} e^{-\frac{1}{\theta_2}(\theta_1+1)C/2} \tag{28}$$

This is proportional to an inverse gamma distribution and so, $\theta_2|\theta_1 \sim \Gamma^{-1}\left(\frac{k+2p}{2}, (\theta_1+1)C/2\right)$. Since the conditional density is an inverse gamma distribution, the number of modes is $M = 1$ and the density function is continuous.

Calculate the maximum value of $f_{\theta_2|\theta_1}(\theta_2|\theta_1)$: Figure 5 shows how the maximum value of the density increases as the shape, $(\theta_1+1)C/2$ decreases when the rate, $\frac{k+2p}{2}$ is fixed. It can also be shown from equation 26 that the density function of $f_{\theta_2|\theta_1}(\theta_2|\theta_1)$ is maximized when $\theta_1 = 0$ since the normalizing constant will be the largest. This means that $f_{\theta_2|\theta_1}(\theta_2|\theta_1)$ reaches its maximum height when $\theta_1 = 0$ and so we find the value of $f_{\theta_2|\theta_1}(\theta_2|\theta_1)$ evaluated at $\theta_2 = \frac{C}{k+2p+2}$, the mode (Section 5.3 of [17]).

$$\begin{aligned}
K &= f_{\theta_2|\theta_1}\left(\frac{C}{k+2p+2}|\theta_1 = 0\right) \\
&= \frac{(C/2)^{\frac{k+2p}{2}}}{\Gamma(\frac{k+2p}{2})} y^{-\frac{k+2p}{2}-1} e^{-\frac{C/2}{y}}|_{y=\frac{C}{k+2p+2}} \\
&= \frac{(C/2)^{\frac{k+2p}{2}}}{\Gamma(\frac{k+2p}{2})} \left(\frac{C}{k+2p+2}\right)^{-\frac{k+2p}{2}-1} e^{-\frac{k+2p+2}{2}} \\
&= \frac{(C/2)^{\frac{k+2p}{2}}}{\Gamma(\frac{k+2p}{2})} \left(\frac{k+2p+2}{C}\right)^{\frac{k+2p}{2}+1} e^{-\frac{k+2p+2}{2}}
\end{aligned}$$

And so,

$$K = \frac{(C/2)^{\frac{k+2p}{2}}}{\Gamma(\frac{k+2p}{2})} \left(\frac{k+2p+2}{C}\right)^{\frac{k+2p}{2}+1} e^{-\frac{k+2p+2}{2}} \tag{29}$$

□

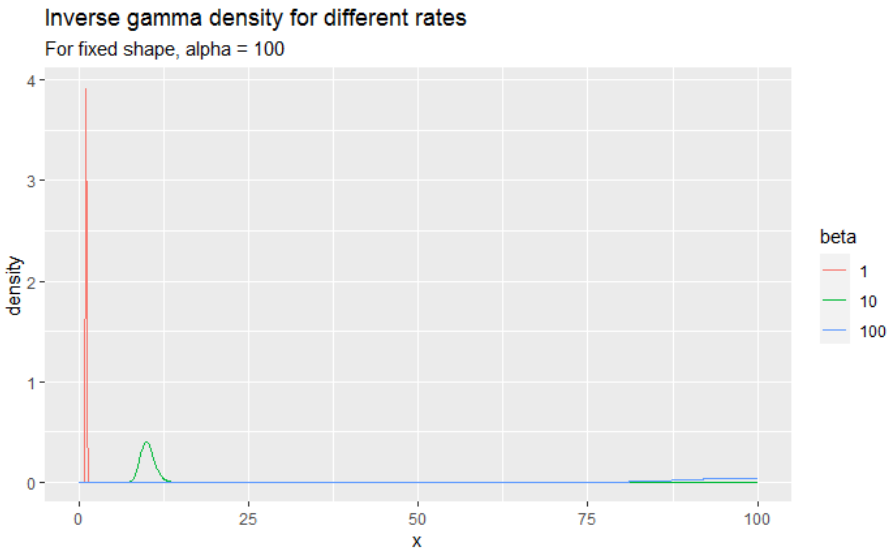

Figure 5: Inverse gamma density when $\alpha = 100$ and $\beta = 1, 10, 100$

6.3.3 Proof of lemmas used in Theorem 4.8

Proof of Lemma 4.9. Since (μ_n, τ_n^{-1}) can be written as a random function of τ_n^{-1},

$$\mathcal{L}(\mu_n, \tau_n^{-1} | \mu_0, \tau_0^{-1}, \tau_n^{-1}) = \mathcal{L}(\mu_n, \tau_n^{-1} | \tau_n^{-1})$$

and τ_n^{-1} is a de-initialization of (μ_n, τ_n^{-1}). Further, by Proposition 6.4,

$$\|\mathcal{L}(\mu_n, \tau_n^{-1}) - \mathcal{L}(\mu_n', \tau_n^{'-1})\| \leq \|\mathcal{L}(\tau_n^{-1}) - \mathcal{L}(\tau_n^{'-1})\|$$

To interpret this in another way, if τ_n couples then the distribution of μ_n is the same for both iterations, so it is automatically coupled. An alternative proof can be made using the results from [23]. □

Lemma 6.6. *For the Bayesian location model,* $\tau_n^{-1} = X_n Y_n \tau_{n-1}^{-1} + Y_n$ *where* $X_n \sim \Gamma\left(\frac{1}{2}, \frac{S}{2}\right)$ *and* $Y_n \sim \Gamma^{-1}\left(\frac{J+2}{2}, \frac{S}{2}\right)$

Proof of Lemma 6.6. The iteration τ_{n+1}^{-1} can be written as a function of its previous value, τ_n^{-1} since $\mu_{n+1} =$

$\bar{y} + Z_{n+1}/\sqrt{J\tau_n}$.

$$\tau_{n+1}^{-1} = \frac{Z_{n+1}^2}{S}\frac{S}{2G_{n+1}}\tau_n^{-1} + \frac{S}{2G_{n+1}} \tag{30}$$

Next we can rewrite, $\tau_n^{-1} = X_n Y_n \tau_{n-1}^{-1} + Y_n$ where $X_n = \frac{Z_{t+1}^2}{S} \sim \Gamma\left(\frac{1}{2}, \frac{S}{2}\right)$ and $Y_n = \frac{S}{2G_{t+1}} \sim \Gamma^{-1}\left(\frac{J+2}{2}, \frac{S}{2}\right)$. □

Proof of Lemma 4.10. By Lemma 6.6, $\theta_{1,n} = X_n Y_n$ and so by corollary 1

$$D = E[|\theta_{1,n}|] = E[X_n Y_n] = E[X_n]E[Y_n] = \frac{1}{S}\frac{S}{J} = \frac{1}{J}$$

□

Proof of Lemma 4.11. To find M, K and show that the conditional density is continuous, we (a) show that $\theta_2|\theta_1 \sim \Gamma^{-1}\left(\frac{J-1}{2}, (\theta_1+1)S/2\right)$, which directly implies that the conditional distribution is continuous and $M = 1$ and we (b) we find the value of K.

(a) Calculate the conditional density $\theta_{2,n}|\theta_{1,n}$ For simplicity, we remove the subscript n on the random variables. Let X, Y be as described in Lemma 6.6. Since the random variables are independent, the joint density is the product of the densities.

$$f_{X,Y}(x,y) = \frac{S/2}{\Gamma(1/2)}x^{1/2-1}e^{xS/2}\frac{S/2}{\Gamma((J+2)/2)}y^{-(J+2)/2-1}e^{-\frac{S/2}{y}} \tag{31}$$

Then $(\theta_1, \theta_2) = (XY, Y)$ is a transformation with the Jacobian $|J| = \theta_2^{-1}$ and the density written as follows,

$$\begin{aligned} f_{\theta_1,\theta_2}(\theta_1,\theta_2) &= f_{X,Y}\left(\frac{\theta_1}{\theta_2}, \theta_2\right)\theta_2^{-1} \\ &= \frac{S/2}{\Gamma(1/2)}\left(\frac{\theta_1}{\theta_2}\right)^{1/2-1} e^{-\frac{\theta_1}{\theta_2}S/2}\frac{S/2}{\Gamma((J+2)/2)}\theta_2^{-(J+2)/2-1}e^{-\frac{S/2}{\theta_2}}\theta_2^{-1} \end{aligned}$$

Next $f_{\theta_2|\theta_1}(\theta_2|\theta_1)$ is proportional to $f_{\theta_1,\theta_2}(\theta_1,\theta_2)$ and so we can derive the conditional density of θ_2 as follows,

$$f_{\theta_2|\theta_1}(\theta_2|\theta_1) \propto f_{\theta_1,\theta_2}(\theta_1,\theta_2) \tag{32}$$

$$\propto \theta_2^{1-1/2} e^{-\frac{1}{\theta_2}\theta_1 S/2} \theta_2^{-(J+2)/2-1} e^{-\frac{1}{\theta_2}S/2} \theta_2^{-1} \tag{33}$$

$$= \theta_2^{-(1/2+(J+2)/2)-1} e^{-\frac{1}{\theta_2}(\theta_1+1)S/2} \tag{34}$$

$$= \theta_2^{-(J-1)/2-1} e^{-\frac{1}{\theta_2}(\theta_1+1)S/2} \tag{35}$$

This is proportional to an inverse gamma distribution and so, $\theta_2|\theta_1 \sim \Gamma^{-1}\left(\frac{J-1}{2}, (\theta_1+1)S/2\right)$. We know that the inverse gamma distribution is continuous and unimodal, so $M = 1$.

(b) Calculate the maximum value of $f_{\theta_2|\theta_1}(\theta_2|\theta_1)$: Similar to figure 5 of example 4.2, $f_{\theta_2|\theta_1}(\theta_2|\theta_1)$ reaches its maximum height when $\theta_1 = 0$. It can also be shown from equation 32 that the density function of $f_{\theta_2|\theta_1}(\theta_2|\theta_1)$ is maximized when $\theta_1 = 0$ since the normalizing constant will be the largest. So the largest value of $f_{\theta_2|\theta_1}(\theta_2|\theta_1)$ will occur when $\theta_1 = 0$. To find the maximum conditional distribution, we find the value of $f_{\theta_2|\theta_1}(\theta_2|\theta_1 = 0)$ evaluated at $\theta_2 = \frac{S}{J+1}$, the mode (see Section 5.3 of [17]).

$$\begin{aligned} K &= f_{\theta_2|\theta_1}\left(\frac{S}{J+1}|\theta_1 = 0\right) \\ &= \frac{(S/2)^{\frac{J-1}{2}}}{\Gamma(\frac{J-1}{2})} y^{-\frac{J-1}{2}-1} e^{-\frac{S/2}{y}}|_{y=\frac{S}{J+1}} \\ &= \frac{(S/2)^{\frac{J-1}{2}}}{\Gamma(\frac{J-1}{2})} \left(\frac{S}{J+1}\right)^{-\frac{J-3}{2}} e^{-\frac{J+1}{2}} \end{aligned}$$

And so,

$$K = \frac{(S/2)^{\frac{J-1}{2}}}{\Gamma(\frac{J-1}{2})} \left(\frac{S}{J+1}\right)^{-\frac{J-3}{2}} e^{-\frac{J+1}{2}} \tag{36}$$

□

6.3.4 Proof of Theorem 4.12

Proof of Theorem 4.12. This example uses a modified version of the Sideways Theorem 4.1 to find an upper bound on the convergence rate. We will also use Proposition 2.1, which states that the total variation between two random variables is equal to the total variation of any invertible transformation of the same two random variables.

Let $\vec{X}_n, \vec{X}'_n \in \mathbb{R}^2$ be two copies of the autoregressive normal process as defined in example 4.6. Then for $\vec{Z}_n \sim N(\vec{0}, I_d)$,

$$\vec{X}_n = A\vec{X}_{n-1} + \Sigma_d \vec{Z}_n \qquad\qquad \vec{X}'_n = A\vec{X}'_{n-1} + \Sigma_d \vec{Z}'_n$$

We apply the one-shot coupling method to bound the total variation distance. For $n < N$ set $\vec{Z}_n = \vec{Z}'_n$.

Suppose X_0, X'_0 are known and define

$$\Delta = \|\Sigma_d^{-1} A^n (\vec{X}_0 - \vec{X}'_0)\|_2$$

Decompose $A = PDP^{-1}$ with D as the corresponding diagonal matrix, λ_i is the ith eigenvalue of A and $\|\cdot\|_2$ denotes the Frobenius norm. Then Δ is bounded above as follows,

$$\begin{aligned}
\Delta &= \|\Sigma_d^{-1} A^n (\vec{X}_0 - \vec{X}'_0)\|_2 \\
&= \|\Sigma_d^{-1} P D^n P^{-1} (\vec{X}_0 - \vec{X}'_0)\|_2 \\
&\leq \|\Sigma_d^{-1}\|_2 \cdot \|P\|_2 \|D^n\|_2 \|P^{-1}\|_2 \|\vec{X}_0 - \vec{X}'_0\|_2 && \text{by Lemma 1.2.7 of [1]} \\
&\leq \|\Sigma_d^{-1}\|_2 \cdot \|P\|_2 \|P^{-1}\|_2 \|\vec{X}_0 - \vec{X}'_0\|_2 \sqrt{\sum_{i=1}^{d} |\lambda_i|^{2n}} \\
&\leq \|\Sigma_d^{-1}\|_2 \cdot \|P\|_2 \|P^{-1}\|_2 \|\vec{X}_0 - \vec{X}'_0\|_2 \sqrt{d} \max_{1 \leq i \leq d} |\lambda_i|^n
\end{aligned}$$

For now assume that X_0, X'_0 are known and note that Σ_d^{-1} is an invertible transform. We bound the total variation distance as follows by applying two invertible transforms on the Markov chain and using the fact that

$\vec{Z}_m = \vec{Z}'_m, m < N.$

$$
\begin{aligned}
&\|\mathcal{L}(\vec{X}_N) - \mathcal{L}(\vec{X}'_N)\| \\
&\leq E_{\{\vec{Z}_m\}_{m<N}}\left[\|\mathcal{L}(\vec{X}_N) - \mathcal{L}(\vec{X}'_N)\|\right] && \text{by prop. 2.2} \\
&= E_{\{\vec{Z}_m\}_{m<N}}\left[\|\mathcal{L}(\Sigma_d^{-1}\vec{X}_N) - \mathcal{L}(\Sigma_d^{-1}\vec{X}'_N)\|\right] && \text{by prop. 2.1} \\
&= E_{\{\vec{Z}_m\}_{m<N}}\left[\|\mathcal{L}(\Sigma_d^{-1}A\vec{X}_{N-1} + \vec{Z}_N) - \mathcal{L}(\Sigma_d^{-1}A\vec{X}'_{N-1} + \vec{Z}'_N)\|\right] \\
&= E_{\{\vec{Z}_m\}_{m<N}}\left[\|\mathcal{L}(\Sigma_d^{-1}(A^N\vec{X}_0 + \sum_{m=1}^{N-1} A^{N-m}\vec{Z}_m) + \vec{Z}_N) - \mathcal{L}(\Sigma_d^{-1}(A^N\vec{X}'_0 + \sum_{m=1}^{N-1} A^{N-m}\vec{Z}_m) + \vec{Z}'_N)\|\right] \\
&= E_{\{\vec{Z}_m\}_{m<N}}\left[\|\mathcal{L}(\Sigma_d^{-1}A^N\vec{X}_0 + \vec{Z}_N) - \mathcal{L}(\Sigma_d^{-1}A^N\vec{X}'_0 + \vec{Z}'_N)\|\right] && \text{by prop. 2.1} \\
&= E_{\{\vec{Z}_m\}_{m<N}}\left[\|\mathcal{L}(\vec{Z}_N + \Sigma_d^{-1}A^N(\vec{X}_0 - \vec{X}'_0)) - \mathcal{L}(\vec{Z}'_N)\|\right] \\
&= \|\mathcal{L}(\vec{Z}_N + \Sigma_d^{-1}A^N(\vec{X}_0 - \vec{X}'_0)) - \mathcal{L}(\vec{Z}'_N)\|
\end{aligned}
$$

There exists a rotation matrix $R \in \mathbb{R}^{d\times d}$ such that

$$R[\Sigma_d^{-1}A(\vec{X}_n - \vec{X}'_n)] = (\|\Sigma_d^{-1}A(\vec{X}_n - \vec{X}'_n)\|_2, 0, \dots 0) = (\Delta, 0, \dots 0)$$

[1]. By properties of rotation, R is orthogonal, so $R^T = R^{-1}$ and $RZ_n \sim N(0, RI_dR^T) = N(0, I_d) \sim Z_n$. In other words, $RZ_n \overset{d}{=} Z_n \overset{d}{=} Z'_n$. Thus, continuing the above equality,

$$
\begin{aligned}
\|\mathcal{L}(\vec{X}_n) - \mathcal{L}(\vec{X}'_n)\| &\leq \|\mathcal{L}(\vec{Z}_n + \Sigma_d^{-1}A^n(\vec{X}_0 - \vec{X}'_0)) - \mathcal{L}(\vec{Z}'_n)\| \\
&= \|\mathcal{L}(R[\vec{Z}_n + \Sigma_d^{-1}A(\vec{X}_n - \vec{X}'_n)]) - \mathcal{L}(R\vec{Z}'_n)\| && \text{by prop. 2.1} \\
&= \|\mathcal{L}(\vec{Z}_n + (\Delta, 0, \dots 0)) - \mathcal{L}(\vec{Z}_n)\|
\end{aligned}
$$

Next, suppose that X_0, X_0' are unknown. Then, the inequality stated in equation 16 is shown as follows,

$$\begin{aligned}
\|\mathcal{L}(\vec{X}_n) - \mathcal{L}(\vec{X}_n')\| &\leq E_\Delta[\|\mathcal{L}(\vec{Z}_n + (\Delta, 0, \dots 0)) - \mathcal{L}(\vec{Z}_n)\|] && \text{by prop 2.2} \\
&= E_\Delta[\frac{1}{2}\int_{\mathbb{R}^d} \left| \frac{1}{(2\pi)^{d/2}} e^{-y_1^2/2 - \sum_{i=2}^d y_i^2/2} - \frac{1}{(2\pi)^{d/2}} e^{-(y_1-\Delta)^2/2 - \sum_{i=2}^d y_i^2/2} \right| d\vec{y}] \\
&= E_\Delta[\frac{1}{2}\int_{\mathbb{R}} \left| \frac{1}{\sqrt{2\pi}} e^{-y_1^2/2} - \frac{1}{\sqrt{2\pi}} e^{-(y_1-\Delta)^2/2} d \right| \vec{y}] \\
&= E_\Delta[\|\mathcal{L}(Z_{1,n} + \Delta) - \mathcal{L}(Z_{1,n})\|] \\
&\leq \frac{1}{\sqrt{2\pi}} E[\Delta] && \text{by Lemma 6.3} \\
&\leq \sqrt{\frac{d}{2\pi}} \|\Sigma_d^{-1}\|_2 \cdot \|P\|_2 \|P^{-1}\|_2 E[\|\vec{X}_0 - \vec{X}_0'\|_2] \max_{1 \leq i \leq d} |\lambda_i|^n
\end{aligned}$$

□

6.4 Lemmas for randomly scaled iterated random function examples

6.4.1 Proof of lemmas used in Theorem 5.1

Proof of Lemma 5.2. Let $\{X_n\}_{n\geq 1} \in \mathbb{R}$ and $\{X_n'\}_{n\geq 1} \in \mathbb{R}$ be two copies of the LARCH process. For fixed $n \geq 1$, let $Z_n = Z_n'$ and so,

$$\begin{aligned}
E[|X_n - X_n'|] &= E[|(\beta_0 + \beta_1 X_{n-1})Z_n - (\beta_0 + \beta_1 X_{n-1}')Z_n|] \\
&\leq \beta_1 E[|Z_n|] E[|X_{n-1} - X_{n-1}'|]
\end{aligned}$$

Since $Z_n \stackrel{d}{=} Z_0 > 0$ a.s., the geometric convergence rate is $D = \beta_1 E[Z_0]$. □

Proof of Lemma 5.3. For a fixed $n \geq 0$, suppose that Z_{n+1}, Z_{n+1}' are independent. By Proposition 2.2, the total variation distance between the two processes is bounded above by the expectation of the total variation.

$$||\mathcal{L}(X_{n+1}) - \mathcal{L}(X_{n+1}')|| \leq E[||\mathcal{L}((\beta_0 + \beta_1 X_n)Z_{n+1}) - \mathcal{L}((\beta_0 + \beta_1 X_n')Z_{n+1})||]$$

Note that Z_{n+1} and Z_{n+1}' are used interchangeably in the total variation distance since $Z_{n+1} \stackrel{d}{=} Z_{n+1}'$. Let

$Y_n = \beta_0 + \beta_1 X_n$, $Y'_n = \beta_0 + \beta_1 X'_n$, $\Delta = Y'_n - Y_n$, and $\Delta' = \frac{\Delta}{Y_n}$. WLOG $Y'_n > Y_n$ so that $\Delta, \Delta' > 0$. Then,

$$
\begin{aligned}
||\mathcal{L}(X_{n+1}) - \mathcal{L}(X'_{n+1})|| &\leq E[||\mathcal{L}(Y_n Z_{n+1}) - \mathcal{L}(Y'_n Z_{n+1})||] && \text{by Proposition 2.2} \\
&= E[||\mathcal{L}(Y_n Z_{n+1}) - \mathcal{L}((Y_n + \Delta) Z_{n+1})||] \\
&= E[||\mathcal{L}(Z_{n+1}) - \mathcal{L}((1 + \Delta') Z_{n+1})||] && \text{by Proposition 2.1} \\
&= E[||\mathcal{L}(\log(Z_{n+1})) - \mathcal{L}(\log(1 + \Delta') + \log(Z_{n+1}))||] && \text{by Proposition 2.1} \\
&\leq \frac{M+1}{2} \sup_x e^x f_{Z_n}(e^x) E[\log(1 + \Delta')] && \text{by lem 6.3. See prf of lem 4.2 for more details} \\
&\leq \frac{M+1}{2} \sup_x e^x f_{Z_n}(e^x) \frac{E[|\Delta|]}{\beta_0} && \text{by the mean value theorem} \\
&= \frac{M+1}{2} \sup_x e^x f_{Z_n}(e^x) \frac{\beta_1 E[|X_n - X'_n|]}{\beta_0}
\end{aligned}
$$

□

6.4.2 Proof of lemmas used in Theorem 5.4

Proof of Lemma 5.5. Let $\{X_n\}_{n \geq 1} \in \mathbb{R}$ and $\{X'_n\}_{n \geq 1} \in \mathbb{R}$ be two copies of the asymmetric ARCH process.

For a fixed $n \geq 1$, let $Z_n = Z'_n$ and so,

$$
\begin{aligned}
E[|X_n - X'_n|] &= E[|\sqrt{(aX_{n-1} + b)^2 + c^2} Z_n - \sqrt{(aX'_{n-1} + b)^2 + c^2} Z_n|] \\
&= |\sqrt{(aX_{n-1} + b)^2 + c^2} - \sqrt{(aX'_{n-1} + b)^2 + c^2}| E[|Z_n|]
\end{aligned}
$$

Note that the derivative of $f(x) = \sqrt{(ax + b)^2 + c^2}$ is

$$
|f'(x)| = |\frac{a(ax + b)}{\sqrt{(ax + b)^2 + c^2}}| \leq \frac{|a(ax + b)|}{\sqrt{(ax + b)^2}} = |a| \tag{37}
$$

and so,

$$
E[|X_n - X'_n|] \leq |a| E[|Z_n|] E[|X_{n-1} - X'_{n-1}|]
$$

Thus, the geometric convergence rate is $D = |a|E[|Z_0|]$. □

Proof of Lemma 5.6. Let $\{X_n\}_{n\geq 1} \in \mathbb{R}$ and $\{X'_n\}_{n\geq 1} \in \mathbb{R}$ be two copies of the asymmetric ARCH process.

For $n \geq 1$, Z_n, Z'_n are independent. By Proposition 2.2, the total variation distance between the two processes is bounded above by the expectation of the total variation with respect to $X_{n-1}, X'_{n-1}, Z_n, Z'_n$.

$$||\mathcal{L}(X_n) - \mathcal{L}(X'_n)|| \leq E[||\mathcal{L}(\sqrt{(aX_{n-1}+b)^2 + c^2}Z_n) - \mathcal{L}(\sqrt{(aX'_{n-1}+b)^2 + c^2}Z'_n)||]$$

Let $Y_{n-1} = \sqrt{(aX_{n-1}+b)^2 + c^2}$ and $Y'_{n-1} = \sqrt{(aX_{n-1}+b)^2 + c^2}$, $\Delta = Y'_{n-1} - Y_{n-1}$ and $\Delta' = \frac{\Delta}{Y_{n-1}}$. WLOG, $Y'_{n-1} < Y_{n-1}$, so $-1 < \Delta' < 0$, because $Y_{n-1}, Y'_{n-1} > 0$ and

$$\begin{aligned}
||\mathcal{L}(X_n) - \mathcal{L}(X'_n)|| &\leq E[||\mathcal{L}(Y_{n-1}Z_n) - \mathcal{L}(Y'_{n-1}Z_n)||] \\
&= E[||\mathcal{L}(Y_{n-1}Z_n) - \mathcal{L}((Y_{n-1}+\Delta)Z_n)||] && \text{by Proposition 2.1} \\
&= E[||\mathcal{L}(Z_n) - \mathcal{L}((1+\Delta')Z_n)||] && \text{by Proposition 2.1} \\
&\leq E[\sup_x 1 - \frac{\pi_{Z_n}(x)}{\pi_{(1+\Delta')Z_n}(x)}] && \text{by Lemma 6.16 of [22]}
\end{aligned}$$

Let the density of Z_n be $\pi_{Z_n}(x)$, then $\pi_{(1+\Delta')Z_n}(x) = \frac{1}{1+\Delta'}\pi_{Z_n}\left(\frac{x}{1+\Delta'}\right)$.

$$\begin{aligned}
||\mathcal{L}(X_n) - \mathcal{L}(X'_n)|| &\leq E[\sup_x 1 - (1+\Delta')\frac{\pi_{Z_n}(x)}{\pi_{Z_n}\left(\frac{x}{1+\Delta'}\right)}] \\
&\leq E[\sup_x 1 - (1+\Delta')] && \text{by assumption } \pi_{Z_n}(x) \geq \pi_{Z_n}\left(\frac{x}{1+\Delta'}\right) \\
&= E[\Delta'] \\
&\leq \frac{E[|Y_{n-1} - Y'_{n-1}|]}{c} && \text{since } Y_{n-1} \geq c \\
&\leq \frac{|a|}{c}E[|X_{n-1} - X'_{n-1}|] && \text{by equation 37}
\end{aligned}$$

□

6.4.3 Proof of lemmas used in Theorem 5.7

Proof of Lemma 5.8. Let $\{X_n\}_{n\geq 1} \in \mathbb{R}$ and $\{X'_n\}_{n\geq 1} \in \mathbb{R}$ be two copies of the GARCH process. For $n \geq 2$, let $Z_n = Z'_n$. First note that,

$$E[|X_n - X'_n|] = E[|\sigma_n Z_n - \sigma'_n Z_n|] = E[|\sigma_n - \sigma'_n||Z_n|] = E[|\sigma_n - \sigma'_n|]E[|Z_n|] \tag{38}$$

Next, we find an upper bound on $E[|\sigma_n - \sigma'_n|]$ by first noting that $\sigma_n^2 = \alpha^2 + (\beta^2 Z_{n-1}^2 + \gamma^2)\sigma_{n-1}^2$ by substitution.

$$\begin{aligned}
E[|\sigma_n - \sigma'_n|] &= E[|\sqrt{\alpha^2 + (\beta^2 Z_{n-1}^2 + \gamma^2)\sigma_{n-1}^2} - \sqrt{\alpha^2 + (\beta^2 Z_{n-1}^2 + \gamma^2)\sigma_{n-1}'^2}|] \\
&\leq E[\sqrt{\beta^2 Z_{n-1}^2 + \gamma^2}]E[|\sigma_{n-1} - \sigma'_{n-1}|] && \text{taking max of the derivative} \\
&= E[\sqrt{\beta^2 Z_{n-1}^2 + \gamma^2}]\frac{E[|X_{n-1} - X'_{n-1}|]}{E[|Z_{n-1}|]} && \text{by equation 38}
\end{aligned}$$

Finally, substituting $E[|\sigma_n - \sigma'_n|]$ into equation 38,

$$\begin{aligned}
E[|X_n - X'_n|] &\leq E[\sqrt{\beta^2 Z_{n-1}^2 + \gamma^2}]\frac{E[|X_{n-1} - X'_{n-1}|]}{E[|Z_{n-1}|]}E[|Z_n|] \\
&= E[\sqrt{\beta^2 Z_{n-1}^2 + \gamma^2}]E[|X_{n-1} - X'_{n-1}|] \\
&\leq \sqrt{\beta^2 E[Z_0^2] + \gamma^2}E[|X_{n-1} - X'_{n-1}|] && \text{by Jensen's inequality}
\end{aligned}$$

Thus, the geometric convergence rate is $D = \sqrt{\beta^2 E[Z_0^2] + \gamma^2}$. □

Proof of Lemma 5.9. Let $\{X_n\}_{n\geq 1} \in \mathbb{R}$ and $\{X'_n\}_{n\geq 1} \in \mathbb{R}$ be two copies of the GARCH process.

For $n \geq 2$, suppose that Z_n, Z'_n are independent. By Proposition 2.2, the total variation distance between the two processes is bounded above by the expectation of the total variation.

$$||\mathcal{L}(X_n) - \mathcal{L}(X'_n)|| \leq E[||\mathcal{L}(\sigma_n Z_n) - \mathcal{L}(\sigma'_n Z_n)||]$$

Let $\Delta = \sigma_n' - \sigma_n$ and $\Delta' = \frac{\Delta}{\sigma_n}$. WLOG, $\sigma_n' < \sigma_n$, so $\Delta, \Delta' < 0$ because $\sigma_n, \sigma_n' > 0$ and

$$\begin{aligned}
||\mathcal{L}(X_n) - \mathcal{L}(X_n')|| &= E[||\mathcal{L}(\sigma_n Z_n) - \mathcal{L}((\sigma_n + \Delta)Z_n)||] && \text{by Proposition 2.1} \\
&= E[||\mathcal{L}(Z_n) - \mathcal{L}((1+\Delta')Z_n)||] && \text{by Proposition 2.1} \\
&\leq E\left[\sup_x 1 - \frac{\pi_{Z_n}(x)}{\pi_{(1+\Delta')Z_n}(x)}\right] && \text{by Lemma 6.16 of [22]}
\end{aligned}$$

Let the density of Z_n be $\pi_{Z_n}(x)$, then $\pi_{(1+\Delta')Z_n}(x) = \frac{1}{1+\Delta'}\pi_{Z_n}\left(\frac{x}{1+\Delta'}\right)$.

$$\begin{aligned}
||\mathcal{L}(X_n) - \mathcal{L}(X_n')|| &\leq E\left[\sup_x 1 - (1+\Delta')\frac{\pi_{Z_n}(x)}{\pi_{Z_n}\left(\frac{x}{1+\Delta'}\right)}\right] \\
&\leq E[\sup_x 1 - (1+\Delta')] && \text{by assumption } \pi_{Z_n}(x) \geq \pi_{Z_n}\left(\frac{x}{1+\Delta'}\right) \\
&= E[\Delta'] \\
&\leq \frac{E[|\sigma_n' - \sigma_n|]}{\alpha} && \text{since } \sigma_n \geq \alpha \\
&\leq \frac{D}{\alpha E[|Z_{n-1}|]} E[|X_{n-1} - X_{n-1}'|] && \text{by equation in proof 6.4.3}
\end{aligned}$$

□

Proof of Lemma 5.10.

$$\begin{aligned}
E[|X_1 - X_1'|] &= |\sigma_1^2 - \sigma_1^{'2}|E[|Z_1|] && \text{by equation in proof 6.4.3} \\
&= |\sqrt{\alpha^2 + \beta^2 X_0^2 + \gamma^2 \sigma_0^2} - \sqrt{\alpha^2 + \beta^2 X_0^{'2} + \gamma^2 \sigma_0^{'2}}|E[|Z_1|] \\
&\leq \sqrt{|(\alpha^2 + \beta^2 X_0^2 + \gamma^2 \sigma_0^2) - (\alpha^2 + \beta^2 X_0^{'2} + \gamma^2 \sigma_0^{'2})|}E[|Z_1|] \\
& && \text{since } |\sqrt{x} - \sqrt{y}| = \sqrt{(\sqrt{x} - \sqrt{y})^2} = \sqrt{x + y - 2\sqrt{x}\sqrt{y}} \leq \sqrt{|x-y|} \\
&\leq \sqrt{\beta^2 |X_0^2 - X_0^{'2}| + \gamma^2 |\sigma_0^2 - \sigma_0^{'2}|}E[|Z_0|]
\end{aligned}$$

□

References

[1] Charu Aggarwal. *Linear Algebra and Optimization for Machine Learning: A Textbook*. Springer International Publishing, 2020. DOI: 10.1007/978-3-030-40344-7.

[2] Peter H. Baxendale. "Renewal theory and computable convergence rates for geometrically ergodic Markov chains". In: *The Annals of Applied Probability* 15.1B (2005), pp. 700–738. DOI: 10.1214/105051604000000710.

[3] Patrick Billingsley. *Probability and Measure: Anniversary Edition*. New York: Wiley Series in Probability and Statistics, 2012.

[4] Björn Böttcher. *Markovian Maximal Coupling of Markov Processes*. 2017. arXiv: 1710.09654 [math.PR].

[5] Peter J. Brockwell and Richard A. Davis. *Introduction to Time Series and Forecasting*. 2nd ed. Springer Texts in Statistics, 2002. DOI: 10.1007/978-3-319-29854-2.

[6] Persi Diaconis and David Freedman. "Iterated random functions". In: *SIAM Review* 41.1 (1999), pp. 45–76. DOI: 10.1137/S0036144598338446.

[7] Paul Doukhan. *Stochastic Models for Time Series*. 1st ed. Springer International Publishing, 2018. DOI: 10.1007/978-3-319-76938-7.

[8] Alain Durmus and Éric Moulines. "Quantitative bounds of convergence for geometrically ergodic Markov chain in the Wasserstein distance with application to the Metropolis Adjusted Langevin Algorithm". In: *Statistics and Computing* 25 (2015), pp. 5–19. DOI: 10.1007/s11222-014-9511-z.

[9] James M. Flegal, Murali Haran, and Galin L. Jones. "Markov Chain Monte Carlo: Can We Trust the Third Significant Figure?" In: *Statistical Science* 23.2 (2008), pp. 250–260. DOI: 10.1214/08-STS257.

[10] A. Gelman and D.B. Rubin. "Inference from Iterative Simulation using Multiple Sequences". In: *Statistical Science* 7.4 (1992), pp. 457–472. DOI: 10.1214/ss/1177011136.

[11] Charles J. Geyer. "Introduction to Markov Chain Monte Carlo". In: *Handbook of Markov Chain Monte Carlo*. New York: Chapman and Hall/CRC, 2011, pp. 1–46. DOI: 10.1201/b10905.

[12] Alison L. Gibbs. "Convergence in the Wasserstein Metric for Markov Chain Monte Carlo Algorithms with Applications to Image Restoration". In: *Stochastic Models* 20.4 (2004), pp. 473–492. DOI: 10.1081/STM-200033117.

[13] Alison L. Gibbs and Francis Edward Su. "On Choosing and Bounding Probability Metrics". In: *International Statistical Review / Revue Internationale de Statistique* 70.3 (2002), pp. 419–435. DOI: 10.2307/1403865.

[14] Denis Guibourg, Loïc Hervé, and James Ledoux. *Quasi-compactness of Markov kernels on weighted-supremum spaces and geometrical ergodicity*. 2012. arXiv: 1110.3240 [math.PR].

[15] Ernst Hairer and Gerhard Wanner. *Analysis by Its History*. New York: Springer-Verlag, 2008. DOI: 10.1007/978-0-387-77036-9.

[16] James P. Hobert and Galin L. Jones. "Honest Exploration of Intractable Probability Distributions via Markov Chain Monte Carlo". In: *Statistical Science* 16.4 (2001), pp. 312–334. DOI: 10.1214/ss/1015346317.

[17] Peter D. Hoff. *A First Course in Bayesian Statistical Methods*. New York: Springer, 2009. DOI: 10.1007/978-0-387-92407-6.

[18] Pierre E. Jacob. *Lecture notes for Couplings and Monte Carlo*. Available at https://sites.google.com/site/pierrejacob/cmclectures?authuser=0 (2021/09/17).

[19] Daniel Jerison. "The Drift and Minorization Method for Reversible Markov Chains". PhD thesis. Stanford University, 2016.

[20] Oliver Jovanovski. "Convergence bound in total variation for an image restoration model". In: *Statistics & Probability Letters* 90 (2014), pp. 11–16. DOI: 10.1016/j.spl.2014.03.007.

[21] Oliver Jovanovski and Neal Madras. "Convergence rates for a hierarchical Gibbs sampler". In: *Bernoulli* 1.23 (2013), pp. 603–625. DOI: 10.3150/15-BEJ758.

[22] David A. Levin, Yuval Peres, and Elizabeth L. Wilmer. *Markov Chains and Mixing Times*. 2nd ed. American Mathematical Society, 2017. DOI: 10.1090/mbk/107.

[23] Jun S. Liu, Wing Hung Wong, and Augustine Kong. "Covariance Structure of the Gibbs Sampler with Applications to the Comparisons of Estimators and Augmentation Schemes". In: *Biometrika* 81.1 (1994), pp. 27–40. DOI: 10.1093/biomet/81.1.27.

[24] Neal Madras and Denis Sezer. "Quantitative bounds for Markov chain convergence: Wasserstein and total variation distances". In: *Bernoulli* 16.3 (2010), pp. 882–908. DOI: 10.2307/25735016.

[25] Sean P. Meyn and Richard L. Tweedie. *Markov Chains and Stochastic Stability*. London: Springer-Verlag, 1993. DOI: 10.1007/978-1-4471-3267-7.

[26] E. Nummelin. "A splitting technique for Harris recurrent chains". In: *Z. Wahrscheinlichkeitstheorie und Verw. Geb.* 43 (1978), pp. 309–318. DOI: 10.1007/BF00534764.

[27] Natesh S. Pillai and Aaron Smith. "Kac's walk on n-sphere mixes in $n \log n$ steps". In: *The Annals of Applied Probability* 27.1 (2017), pp. 631–650. DOI: 10.1214/16-AAP1214.

[28] Qian Qin and James P. Hobert. *Geometric convergence bounds for Markov chains in Wasserstein distance based on generalized drift and contraction conditions*. 2021. arXiv: 1902.02964 [math.PR].

[29] Qian Qin and James P. Hobert. *Wasserstein-based methods for convergence complexity analysis of MCMC with applications*. 2020. arXiv: 1810.08826 [math.ST].

[30] Bala Rajaratnam and Doug Sparks. *MCMC-Based inference in the era of big data: a fundamental analysis of the convergence complexity of high-dimensional chains*. 2015. arXiv: 1508.00947 [math.ST].

[31] R.-D. Reiss. "Approximation of Product Measures with an Application to Order Statistics". In: *The Annals of Probability* 9.2 (1981), pp. 335–341. DOI: 10.1214/aop/1176994477.

[32] Gareth O. Roberts and Jeffrey S. Rosenthal. "General state space Markov chains and MCMC algorithms". In: *Probability Surveys* 1 (2004), pp. 20–71. DOI: 10.1214/154957804100000024.

[33] Gareth O. Roberts and Jeffrey S. Rosenthal. "Markov Chains and de-initializing processes". In: *Scandinavian Journal of Statistics* 28.3 (2001), pp. 489–504. DOI: 10.1111/1467-9469.00250.

[34] Gareth O. Roberts and Jeffrey S. Rosenthal. "One-shot coupling for certain stochastic recursive sequences". In: *Stochastic Processes and their Applications* 99 (2002), pp. 195–208. DOI: 10.1016/S0304-4149(02)00096-0.

[35] Jeffrey S. Rosenthal. *A First Look at Rigorous Probability Theory*. 2nd ed. World Scientific, 2016. DOI: 10.1142/6300.

[36] Jeffrey S. Rosenthal. "Analysis of the Gibbs sampler for a model related to James-Stein estimators". In: *Statistics and Computing* 6 (1996), pp. 269–275. DOI: 10.1007/BF00140871.

[37] Jeffrey S. Rosenthal. "Convergence Rates for Markov Chains". In: *SIAM Review* 37.3 (1995), pp. 387–405. DOI: 10.1137/1037083.

[38] Jeffrey S. Rosenthal. "Minorization Conditions and Convergence Rates for Markov Chain Monte Carlo". In: *Journal of the American Statistical Association* 90.430 (1995), pp. 558–566. DOI: 10.2307/2291067.

[39] Laurent Saloff-Coste. "Lectures on finite Markov chains". In: *Lectures on Probability Theory and Statistics: Ecole d'Eté de Probabilités de Saint-Flour XXVI-1996*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1997, pp. 301–413. DOI: 10.1007/BFb0092621.

[40] Rens van de Schoot et al. "What Took Them So Long? Explaining PhD Delays among Doctoral Candidates". In: *PLoS ONE* 8.7 (2013), e68839. DOI: 0.1371/journal.pone.0068839.

[41] Laurent Smeets and Rens van de Schoot. *R regression Bayesian (using brms)*. 2019. URL: www.rensvandeschoot.com/tutorials/r-linear-regression-bayesian-using-brms/ (visited on 06/03/2021).

[42] Aixin Tan, Galin L. Jones, and James P. Hobert. "On the Geometric Ergodicity of Two-Variable Gibbs Samplers". In: *Institute of Mathematical Statistics Collections* 10 (2013), pp. 25–42. DOI: 10.1214/12-IMSCOLL1002.

[43] Jun Yang and Jeffrey S. Rosenthal. *Complexity results for MCMC derived from quantitative bounds*. 2019. arXiv: 1708.00829 [stat.CO].