

# Gibbs Sampling, Conjugate Priors and Coupling

Persi Diaconis \*

Departments of Mathematics and Statistics  
Stanford University

Kshitij Khare

Department of Statistics  
University of Florida, Gainesville

Laurent Saloff-Coste<sup>†</sup>

Department of Mathematics  
Cornell University

## Abstract

We give a large family of simple examples where a sharp analysis of the Gibbs sampler can be proved by coupling. These examples involve standard statistical models – exponential families with conjugate priors or location families with natural priors. Our main approach uses a single eigenfunction (always explicitly available in the examples in question) and stochastic monotonicity. We give a satisfactory treatment of several examples that have defeated previous attempts at analysis.

## 1 Introduction

Ashok Maitra worked on the foundations of probability. While our main interactions with him were on the fine points of measure theory and extreme point representations of sets of measures, Ashok’s earliest work was on exponential families [4], a main topic of our present paper.

### 1.1 Background

We analyze a widely used method of simulation - the Gibbs sampler, also known as Glauber dynamics or the heat-bath algorithm. The Gibbs sampler is actively used in statistical physics, chemistry, biology and throughout Bayesian statistics. Briefly (more formal details are given

---

\*Research partially supported by NSF grant DMS-0505673.

†Research partially supported by NSF grant DMS-0603886.

AMS subject classification: 60J20, 60J10.

Keywords and phrases: Gibbs sampling, exponential families, stochastic monotonicity, coupling, location families.

below), given a multivariate probability density  $f(x_1, x_2, \dots, x_p)$  (often specified only up to a normalizing constant), the following Markov chain is run. From  $\underline{x} = (x_1, x_2, \dots, x_p)$ , change the coordinates one at a time; go to  $(x'_1, x_2, \dots, x_p)$ , then  $(x'_1, x'_2, \dots, x_p)$ , and so on until  $(x'_1, x'_2, \dots, x'_p) = \underline{x}'$ . Each change is made by sampling from the conditional distribution of the coordinate being changed with the other coordinates fixed. This is one step of the ‘systematic scan’ Gibbs sampler. Under repeated iteration (and regularity conditions), from any starting state  $\underline{x}$ , the distribution of iterations  $\underline{x}, \underline{x}', \underline{x}'', \dots$  converges to the stationary density  $f$ . Clear statements of regularity conditions are in [2, 41].

We are concerned with the rate of convergence to stationarity in various metrics. Given a starting state  $\underline{x}$  and  $\epsilon > 0$ , how many iterations are required so that the distribution of the  $\ell^{\text{th}}$  iterate is within distance  $\epsilon$  from the stationary distribution - and how does  $\ell$  depend on  $\underline{x}$  and  $\epsilon$ ?

There has been limited work on this problem beginning with Meyn and Tweedie [31], Rosenthal [34, 35, 36] and their collaborators. See Jones and Hobert [23, 24] for a readable overview and some clearly worked out examples. These authors use Harris recurrence techniques (minorization conditions on small sets and coupling) to get quantitative ‘honest bounds’. We tried these techniques out on a class of two-component examples introduced by Casella and George [7] and used in Liu [28]. These involved the Binomial  $(n, \theta)$  densities with a beta prior for  $\theta$ . For  $n = 100$ , with a uniform prior, simulations ‘showed’ that a few hundred steps of the Markov chain are necessary and sufficient to be very close to the stationary distribution. The best we could do by using careful estimates in the Harris recurrence approach is to show that  $10^{33}$  steps suffice to get within  $\frac{1}{100}$  (in total variation distance) of the stationary distribution (see [13]). However, we managed to explicitly diagonalize the Markov chain for this example, finding all the eigenvalues and eigenvectors (these turned out to be orthogonal polynomials). Using these, we proved that around 200 steps are necessary and sufficient to get within  $\frac{1}{100}$  (in total variation distance) of the stationary distribution. This story is carefully told in the companion paper [13] which has a lengthy review of the literature connected to the Gibbs sampler. We add that [13] is a discussion paper, and (despite serious attempts) none of the discussants were able to find useful improvements over the  $10^{33}$  bounds for the Beta-Binomial example using Harris recurrence techniques.

In [13], we found that explicit diagonalization can be carried out for a class of one-dimensional exponential families and conjugate priors - those with quadratic variance structure. However, there are many examples, even with two components, where explicit diagonalizations cannot be found. The present paper introduces some new techniques which help us to obtain sharp rates of convergence for examples which could not be handled previously. We turn to a more careful account.

## 1.2 The basic setup

Let  $(\mathcal{X}, \mathcal{F})$  and  $(\Theta, \mathcal{G})$  be measurable spaces equipped with  $\sigma$ -finite measures  $\mu$  and  $\nu$  respectively. Let  $\{f_\theta(x)\}_{\theta \in \Theta}$  be a family of probability densities on  $\mathcal{X}$  with respect to  $\mu$ . Let  $\pi(\theta)$  be

a probability density on  $\Theta$  with respect to  $\nu$ . These determine a joint density

$$f(x, \theta) = f_\theta(x)\pi(\theta) \text{ w.r.t. } \mu \times \nu. \quad (1)$$

The marginal density on  $\mathcal{X}$  is

$$m(x) = \int f_\theta(x)\pi(\theta)d\nu(\theta). \quad (2)$$

Throughout, we assume for simplicity that  $m(x) > 0$  for all  $x$ . The conditional densities are

$$f(x | \theta) = f_\theta(x) \text{ and } f(\theta | x) = \frac{f(x, \theta)}{m(x)} \quad (3)$$

Note that we use  $f(x | \theta)$  when  $\theta$  is thought of as a *random variable* and  $f_\theta(x)$  when  $\theta$  is thought of as a *fixed parameter*. The Gibbs sampler is an algorithm for drawing samples from  $f(x, \theta)$  when it is easy to sample from  $f(x | \theta)$  and  $f(\theta | x)$ . This is how it proceeds:

From  $(x, \theta)$

- Draw  $\theta'$  from  $f(\theta' | x)$
- Draw  $x'$  from  $f(x' | \theta')$ .

This defines a Markov chain with transition density

$$\tilde{K}(x, \theta ; x', \theta') = f(\theta' | x)f(x' | \theta'). \quad (4)$$

Under mild conditions specified in [2, 41] (always met in our examples), this Markov chain is ergodic (irreducible and aperiodic) with  $f(x, \theta)$  as its stationary density. In this paper we give families of examples where sharp analyses of rate of convergence to stationarity are possible. The examples include exponential families with conjugate priors or location families with natural priors. They include families where other techniques, such as spectral analysis or Harris recurrence, break down. An example of our results is as follows.

*Example (Geometric/Beta).*

Let

$$f_\theta(j) = \theta(1 - \theta)^j, \quad j = 0, 1, 2, \dots$$

$$\pi(\theta) = \text{Beta}(\alpha, \beta ; \theta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1 - \theta)^{\beta-1}, \quad \theta \in (0, 1), \quad \alpha, \beta > 0.$$

The joint density is

$$f(j, \theta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^\alpha(1 - \theta)^{\beta+j-1}.$$

The marginal (on  $j$ ) and conditional (of  $\theta$  given  $j$ ) densities are

$$m(j) = \frac{\Gamma(\alpha + \beta)\Gamma(\alpha + 1)\Gamma(\beta + j)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\alpha + \beta + j + 1)}, \quad f(\theta | j) = \text{Beta}(\alpha + 1, \beta + j; \theta).$$

The transition density of the  $j$ -chain is defined as follows.

$$k(j, j') = \int_0^1 f(\theta | j) f_\theta(j') d\theta.$$

This is a reversible chain with  $m(j)$  as the stationary density. As explained below in Section 4.1, the behavior of the bivariate chain is determined by the behavior of the  $x$ -chain (here the  $j$ -chain). Here, the  $j$ -chain has transition density

$$k(j, j') = \frac{\Gamma(\alpha + \beta + j + 1)\Gamma(\alpha + j + j')\Gamma(\beta + 2)}{\Gamma(\alpha + 1)\Gamma(\beta + j)\Gamma(\alpha + \beta + j + j' + 2)}.$$

**Theorem 1.1** *For the Geometric/Beta density with  $\alpha > 1$ ,  $\beta > 0$ , all  $j$  and  $l \geq 0$ ,*

$$\|k_j^l - m\|_{TV} \leq \left(j + \frac{\beta}{\alpha - 1}\right) \left(\frac{1}{\alpha}\right)^l. \quad (5)$$

The theorem shows that order  $\log_\alpha(j)$  steps suffice for convergence. For example, take  $\alpha = 2, \beta = 1$  (so  $\pi(\theta) = 2\theta, m(j) = \frac{4}{(j+1)(j+2)(j+3)}$ ) with starting values  $j = 100, \theta = \frac{1}{2}$ . Using (5), we get that *the total variation distance to stationarity is smaller than 0.01 for  $l = 14$* . The techniques developed in [13] to study chains of this type using orthogonal polynomials, do not apply to this particular example as all moments of  $m(j)$  except the mean are infinite. If we treat this example by using a version of Harris recurrence techniques from [5] specially tailored to our two component Gibbs samplers, then we obtain that *the total variation distance to stationarity is less than 0.01 after 1400 steps*.

*Remark* Laurent Miclo (private communication) has shown us how to use Cheegers inequality to prove that with  $\alpha = \beta = 1$  (where  $m(j)$  does not even have a mean) the chain has a spectral gap. Now, standard bounds show that order  $\log(j)$  steps suffice.

The two component examples treated here are simple examples for calibrating available tools. It is easy to sample from any of our  $f(x, \theta)$  distributions directly, sample  $\theta$  from  $\pi(\cdot)$  and then sample  $x$  from  $f(\cdot | \theta)$ . Note however that a host of widely used algorithms (auxilliary variables, data augmentation, slice sampler, ...) can be seen as two component Gibbs samplers - see [9] for extensive discussion. We hope that our new techniques will be useful here. Our ‘one eigenvector + monotonicity’ techniques have already been applied in a completely different setting of the ‘carries’ process (see [12]).

Our main theoretical results appear in Section 2. These include a pair of theorems that give explicit upper bounds for convergence in the presence of a single eigenfunction for stochastically montone Markov chains. For discrete spaces, bounds are in total variation. For continuous

spaces, bounds are in  $L^1$  Wasserstein distance. An extension of the second moment method and Wilson’s lemma for proving lower bounds is also given. Tools for proving stochastic monotonicity using total positivity are reviewed in Section 3.

Section 4 contains families of examples where the above machinery works. It gives a brief review of exponential families and conjugate priors, proving that (if the marginal has finite mean  $x^*$ ), the function  $x - x^*$  is always an eigenfunction. It is further shown that for one dimensional exponential families with any prior, the  $x$ -chain is stochastically monotone. The main results are compactly summarized in Corollary 4.1. In Section 4.2, several further examples are treated in detail including the interesting ‘sixth family of Morris’ (hyperbolic cosine density) where all previous techniques fall flat.

Section 5 has a host of location family examples where the techniques developed in this paper give sharp results. The main results are summarized in Corollary 5.1. This handles quite general densities, while our previous techniques could only handle exponential families.

Section 6 contains yet different arguments - direct couplings and strong stationary times. The examples include the well studied M/M/ $\infty$  queue and some multivariate examples, where orthogonal polynomial techniques could only handle very restricted starting states.

## 2 Monotone Markov Chains

Let  $\mathcal{X}$  be a subset of the real line  $\mathbb{R}$  with its Borel sets. Let  $K(x, dy)$  be a Markov kernel on  $\mathcal{X}$ . We say  $K$  is *stochastically monotone* if  $x, x' \in \mathcal{X}$ ,  $x \leq x'$ , then

$$K(x, (-\infty, y]) \geq K(x', (-\infty, y]) \text{ for all } y. \quad (6)$$

Monotone Markov chains have been thoroughly studied and applied. See [21, 30, 40] and the references there. They are currently in vogue because of ‘coupling from the past’. See [43] for extensive references on this subject. There is a standard coupling technique available for monotone Markov chains. Wilson [42] uses this coupling technique in the presence of an explicit eigenfunction to bound rates of convergence of stochastically monotone Markov chains on finite state spaces. In this section, this coupling argument is used to prove two general theorems about convergence to stationarity of ergodic, monotone Markov chains with stationary distribution  $\pi$  in the presence of an eigenfunction. Sections 4 and 5 give examples where the conditions are satisfied. Section 6 treats some of the examples by direct couplings.

### 2.1 Convergence of monotone chains: main statements

**Theorem 2.1** *Let  $(K, \pi)$  be an ergodic stochastically monotone Markov chain on  $\mathcal{X} \subseteq \mathbb{R}$ . Suppose that there exist  $\lambda \in (0, 1)$ ,  $\eta \in \mathbb{R}$ , and a monotone increasing function  $f$  such that*

$$Kf(x) = \int_{\mathcal{X}} f(y)K(x, dy) = \lambda f(x) + \eta, \quad \forall x \in \mathcal{X}, \quad (7)$$

and

$$c = \inf\{f(y) - f(x) \mid x, y \in \mathcal{X}, x < y\} > 0. \quad (8)$$

Then, for any starting state  $x$ ,

$$\|K_x^l - \pi\|_{TV} \leq c^{-1} \lambda^l \mathbf{E}|f(Z) - f(x)|, \text{ where } Z \sim \pi.$$

The proof is given below in Section 2.2. The next result replaces total variation by the  $L^1$  Wasserstein distance  $d_W$  defined by

$$\begin{aligned} d_W(\mu, \nu) &= \inf_{X \sim \mu, Y \sim \nu} \mathbf{E}|X - Y| \\ &= \sup\{|\mu(\phi) - \nu(\phi)| : \phi : \mathcal{X} \rightarrow \mathbb{R}, |\phi(x) - \phi(y)| \leq |x - y|\}. \end{aligned}$$

See, e.g., [18].

**Theorem 2.2** *Let  $(K, \pi)$  be an ergodic stochastically monotone Markov chain on  $\mathcal{X} \subseteq \mathbb{R}$ . Suppose that there exist  $\lambda \in (0, 1)$ ,  $\eta \in \mathbb{R}$  and a monotone increasing function  $f$  such that*

$$Kf = \lambda f + \eta, \quad (9)$$

and

$$c = \inf\left\{\frac{f(y) - f(x)}{y - x} \mid x, y \in \mathcal{X}, x < y\right\} > 0. \quad (10)$$

Then, for any starting state  $x$ ,

$$d_W(K_x^l, \pi) \leq c^{-1} \lambda^l \mathbf{E}|f(Z) - f(x)|, \text{ where } Z \sim \pi.$$

*Remarks 1.* Theorem 2.1 is used for chains on the integers, often with  $f(x) = x$ . It does not apply to continuous chains when the constant  $c$  vanishes. Theorem 2.2 does apply to both discrete and continuous chains.

2. By elementary manipulations with  $\bar{f} = \mathbf{E}_\pi(f)$ , the function  $f - \bar{f}$  with  $f$  as in (7) and (9) is an eigenfunction of the chain with eigenvalue  $\lambda$ , i.e.,  $K(f - \bar{f}) = \lambda(f - \bar{f})$ . It is instructive to compare the conclusions of the theorems with standard spectral bounds when  $\lambda$  is the second largest eigenvalue. For a Markov chain on the integers, the usual spectral bound is  $\|K_x^l - \pi\|_{TV} < \pi(x)^{-\frac{1}{2}} \lambda^l$ . For the Geometric/Beta  $x$ -chain of Theorem 1.1 with  $\alpha = 2$ , the spectral bound and the bound given by Theorem 2.1 are essentially the same. For the  $x$ -chain of the Poisson/Exponential family (c.f. Section 4.2), the stationary distribution is a geometric ( $\pi(x) = 2^{-x-1}$  for  $x = 0, 1, 2, \dots$ ),  $\lambda = 1/2$ . Theorem 2.1 gives an upper bound  $\|K_x^l - \pi\|_{TV} \leq (x+1)2^{-l}$ . The spectral bound  $\|K_x^l - \pi\|_{TV} \leq 2^{\frac{x+1-2l}{2}}$  is much weaker.

3. The bounds of theorems 2.1 and 2.2 are sharp surprisingly often – see the examples in Sections 4, 5 and 6. A natural example where they are slightly off is provided in Section 4.2.

4. The techniques of this section break down for the Geomtric/Beta example when  $\alpha = \beta = 1$ . Then,  $m(j) = \frac{1}{(j+1)(j+2)}$  fails to have a mean. The  $j$ -chain with transition kernel  $k(j, j') =$

$\frac{2(j+1)(j+2)}{(j+j'+1)(j+j'+2)(j+j'+3)}$  has generalized eigenfunction  $f(j) = j$  ( $\mathbf{E}[j' | j] = j + 1$ ), but we do not know how to use this. Preliminary computations show that the operator  $k$  on  $L^2(m)$  has continuous spectrum in  $[0, \frac{\pi}{8}]$ . We believe all of the Geometric/Beta chains have continuous spectrum. In contrast, all of the examples treated in [13] have an  $x$ -chain with a compact operator. For the Geometric/Uniform case, preliminary computations show that the  $x$ -chain has a spectral gap.:  $\text{spec}(k) \cap (-1, 1) \subseteq [-\beta^*, \beta^*]$  for some  $0 < \beta^* < 1$ . Now, the standard bound from Remark 2 gives  $\|K_x^\ell - m\|_{\text{TV}} \leq \sqrt{(x+1)(x+2)}(\beta^*)^\ell$ , so order  $\log x$  steps suffice.

## 2.2 Proof of Theorems 2.1 and 2.2

*Proof of Theorem 2.1* The proof begins by the standard route of finding a monotone realization of two copies of the Markov chain. The function  $f$  is then used to bound the coupling time. Finally, a coupling bound for two arbitrary starting states is turned into a bound on distance to stationarity.

Let  $F_x(y) = K(x, (-\infty, y])$ . Fix  $x \leq x'$  in  $\text{Support}(\pi)$ . Define a bivariate Markov chain  $\{R_n, S_n\}_{n=0}^\infty$  as follows: Set  $R_0 = x$ ,  $S_0 = x'$ . Let  $U_1, U_2, \dots$  be independent uniform random variables on  $(0, 1)$ . For  $i \geq 1$ , set

$$R_i = F_{R_{i-1}}^{-1}(U_i), S_i = F_{S_{i-1}}^{-1}(U_i) \text{ with } F_x^{-1}(u) = \inf \{y \in \text{Support}(\pi) \mid u \leq F_x(y)\}.$$

By construction, marginally  $R_i, S_i$  are both realizations of a Markov chain with kernel  $K$ .

Since  $K$  is stochastically monotone,  $z \leq z'$  entails  $F_z^{-1}(u) \leq F_{z'}^{-1}(u)$  for  $u$  in  $(0, 1)$ . Hence  $R_0 = x \leq x' = S_0$  entails  $R_1 = F_x^{-1}(U_1) \leq F_{x'}^{-1}(U_1) = S_1$ . Similarly  $R_n \leq S_n$  for all  $n$ . Further, the construction ensures that if  $R_{n_0} = S_{n_0}$  then  $R_n = S_n$  for all  $n \geq n_0$ . This completes the construction of the coupling.

We next bound the coupling time. For any  $n \geq 1$ ,

$$\begin{aligned} P(R_n \neq S_n \mid R_0 = x, S_0 = x') &= \mathbf{E}(\delta_{R_n \neq S_n} \mid R_0 = x, S_0 = x') \\ &\leq \mathbf{E} \left\{ \frac{f(S_n) - f(R_n)}{c} \mid R_0 = x, S_0 = x' \right\}. \end{aligned}$$

The last inequality uses  $S_n \geq R_n$ , the monotonicity of  $f$  and the hypothesis that  $f(y) - f(z) \geq c$  if  $y > z$ . Next, for any  $k$ , one easily checks that

$$\mathbf{E}[f(S_k) - f(R_k) \mid R_{k-1}, S_{k-1}] = \lambda(f(S_{k-1}) - f(R_{k-1})).$$

Hence, we obtain

$$\begin{aligned} P(R_n \neq S_n \mid R_0 = x, S_0 = x') &\leq \mathbf{E} \left[ \mathbf{E} \left[ \frac{f(S_n) - f(R_n)}{c} \mid R_{n-1}, S_{n-1} \right] \mid R_0 = x, S_0 = x' \right] \\ &= \frac{\lambda}{c} \mathbf{E}[f(S_{n-1}) - f(R_{n-1}) \mid R_0 = x, S_0 = x'] \\ &= \frac{\lambda^n}{c} (f(x') - f(x)). \end{aligned}$$

Recall that the total variation distance between two probability measures can be realized as

$$\|\mu - \nu\|_{\text{TV}} = \inf_{X \sim \mu, Y \sim \nu} P(X \neq Y).$$

For  $x \leq x'$ , it follows that

$$\|K_x^n - K_{x'}^n\|_{\text{TV}} \leq c^{-1}(f(x') - f(x))\lambda^n.$$

Thus, for all  $x, x'$

$$\|K_x^n - K_{x'}^n\|_{\text{TV}} \leq c^{-1}|f(x') - f(x)|\lambda^n.$$

Averaging over all  $x'$  yields

$$\begin{aligned} \|K_x^n - \pi\|_{\text{TV}} &\leq c^{-1}\lambda^n \int |f(x) - f(x')|\pi(dx') \\ &= c^{-1}\lambda^n \mathbf{E}|f(Z) - f(x)| \text{ where } Z \sim \pi. \end{aligned}$$

This completes the proof of Theorem 2.1. □

*Proof of Theorem 2.2* Arguing as in the proof of Theorem 2.1, for  $x < x'$ ,

$$\mathbf{E}[S_n - R_n \mid R_0 = x, S_0 = x'] \leq c^{-1}(f(x') - f(x))\lambda^n.$$

The coupling characterization of the Wasserstein distance and symmetry yield

$$d_W(K_x^n, K_{x'}^n) \leq c^{-1}|f(x) - f(x')|\lambda^n.$$

Convexity now yields ( $d_W$  is convex in each of its arguments)

$$d_W(K_x^n, \pi) \leq c^{-1} \int |f(x) - f(x')|\lambda^n \pi(dx') = c^{-1} \mathbf{E}_\pi |f(Z) - f(x)|\lambda^n, \text{ where } Z \sim \pi.$$

This finishes the proof of Theorem 2.2. □

### 2.3 Total variation lower bounds

Theorem 2.1 gives a total variation upper bound based on monotonicity and an eigenfunction. This section gives total variation lower bounds for some of our chains using an eigenfunction without requiring monotonicity. This theorem is based on the second moment method (see [38]), and is an extension of Wilson's lemma (see [42, Lemma 5] or [38, Theorem 4.13]).

**Theorem 2.3** Let  $K$  be an ergodic Markov kernel with stationary probability measure  $\pi$ . Let  $\lambda \in (0, 1)$  be an eigenvalue of  $K$  with associated real-valued eigenfunction  $\phi \in L^2(\pi)$ , such that  $\forall x \in \mathcal{X}$ ,

$$\int (\phi(y) - \phi(x))^2 K(x, dy) \leq (1 - \lambda)^2 \phi^2(x) + B\phi(x) + C,$$

for some  $B, C \geq 0$ . Let

$$T^* := \frac{4B}{\lambda(1 - \lambda)} + \sqrt{\frac{16B^2}{\lambda^2(1 - \lambda)^2} + \frac{8C}{1 - \lambda^2}}.$$

Then for  $t \leq \frac{\log |\phi(x)| + \log \epsilon - \log T^*}{-\log \lambda}$ ,

$$\|K_x^t - \pi\|_{TV} \geq 1 - \epsilon.$$

*Proof* Let  $\{X_t\}_{t \geq 0}$  be a Markov chain with kernel  $K$ . Let  $\mathbf{E}_x$  denote the expectation conditioned on  $X_0 = x$ . Without loss of generality we assume  $\phi(x) \geq 0$  or else we can repeat the whole argument with  $-\phi$  instead of  $\phi$ . Under the given hypothesis,

$$\begin{aligned} & \mathbf{E}_x [(\phi(X_{t+1}) - \phi(X_t))^2 \mid X_t] \leq (1 - \lambda)^2 \phi^2(X_t) + B\phi(X_t) + C \\ \Rightarrow & \mathbf{E}_x [\phi^2(X_{t+1}) \mid X_t] \leq 2\phi(X_t)\mathbf{E}_x[\phi(X_{t+1}) \mid X_t] + (1 - \lambda)^2 \phi^2(X_t) - \phi^2(X_t) + B\phi(X_t) + C \\ \Rightarrow & \mathbf{E}_x [\phi^2(X_{t+1})] \leq \lambda^2 \mathbf{E}_x [\phi^2(X_t)] + B\mathbf{E}_x[\phi(X_t)] + C \\ \Rightarrow & \mathbf{E}_x [\phi^2(X_{t+1})] \leq \lambda^2 \mathbf{E}_x [\phi^2(X_t)] + B\lambda^t \phi(x) + C. \end{aligned}$$

The above identity is true for all  $t \geq 0$ . Using this inductively, we get

$$\begin{aligned} \mathbf{E}_x [\phi^2(X_t)] & \leq \lambda^{2t} \phi^2(x) + B\phi(x) \left( \sum_{i=0}^{t-1} \lambda^{t-i-1} (\lambda^2)^i \right) + C \left( \sum_{i=0}^{t-1} (\lambda^2)^i \right) \\ \Rightarrow \mathbf{E}_x [\phi^2(X_t)] & \leq \lambda^{2t} \phi^2(x) + \frac{B\lambda^t}{\lambda(1 - \lambda)} \phi(x) + \frac{C}{1 - \lambda^2} \\ \Rightarrow \text{Var}_x(\phi(X_t)) & \leq \frac{B\lambda^t}{\lambda(1 - \lambda)} \phi(x) + \frac{C}{1 - \lambda^2} =: R_t \text{ (say)}. \end{aligned} \tag{11}$$

Note that since  $\pi K = \pi$ ,

$$\begin{aligned} \int \phi(x) \pi(dx) & = \int \mathbf{E}_x[\phi(X_1)] \pi(dx) \\ & = \lambda \int \phi(x) \pi(dx). \end{aligned}$$

Hence  $\int \phi(x)\pi(dx) = 0$ , as  $\lambda \in (0, 1)$ . Similarly,

$$\begin{aligned} \int \phi^2(x)\pi(dx) &= \int \mathbf{E}_x[\phi^2(X_t)]\pi(dx) \\ &\leq \lambda^{2t} \int \phi^2(x)\pi(dx) + \frac{B\lambda^t}{\lambda(1-\lambda)} \int \phi(x)\pi(dx) + \frac{C}{1-\lambda^2}. \end{aligned}$$

Hence  $\int \phi^2(x)\pi(dx) \leq \frac{C}{1-\lambda^2} \leq R_t \quad \forall t \geq 0$ . If  $t \leq \frac{\log \phi(x) + \log \epsilon - \log T^*}{-\log \lambda}$ , then

$$\begin{aligned} \lambda^t \phi(x) &\geq \frac{T^*}{\epsilon} \\ \Rightarrow \mathbf{E}_x[\phi(X_t)] &\geq \frac{T^*}{\epsilon} \\ \Rightarrow \mathbf{E}_x^2[\phi(X_t)] &\geq \frac{8}{\epsilon} \left( \frac{B}{\lambda(1-\lambda)} \mathbf{E}_x[\phi(X_t)] + \frac{C}{1-\lambda^2} \right). \end{aligned}$$

The previous assertion follows from the fact that the largest root of  $a^2 - \frac{8Ba}{\epsilon\lambda(1-\lambda)} - \frac{8C}{\epsilon(1-\lambda^2)}$  is less than  $\frac{T^*}{\epsilon}$ . Hence  $a^2 - \frac{8Ba}{\epsilon\lambda(1-\lambda)} - \frac{8C}{\epsilon(1-\lambda^2)} \geq 0$  for  $a \geq \frac{T^*}{\epsilon}$ . Using the definition of  $R_t$  in (11), we get,

$$\mathbf{E}_x[\phi(X_t)] \geq \sqrt{\frac{8R_t}{\epsilon}}.$$

Hence for  $t \leq \frac{\log \phi(x) + \log \epsilon - \log T^*}{-\log \lambda}$ , it follows by Chebyshev's inequality that

$$\begin{aligned} P_x \left( \phi(X_t) < \frac{1}{2} \sqrt{\frac{8R_t}{\epsilon}} \right) &\leq P_x \left( |\phi(X_t) - \mathbf{E}_x[\phi(X_t)]| > \frac{1}{2} \sqrt{\frac{8R_t}{\epsilon}} \right) \\ &\leq \frac{\epsilon}{2}. \end{aligned}$$

If  $Z \sim \pi$ , then

$$\begin{aligned} P_\pi \left( \phi(Z) > \frac{1}{2} \sqrt{\frac{8R_t}{\epsilon}} \right) &\leq \frac{\epsilon \mathbf{E}_\pi[\phi^2(Z)]}{2R_t} \\ &\leq \frac{\epsilon}{2}. \end{aligned}$$

Hence we get,

$$\begin{aligned} \|K_x^t - \pi\|_{\text{TV}} &= \sup_A |K_x^t(A) - \pi(A)| \\ &\geq \left| P_x \left( \phi(X_t) > \frac{1}{2} \sqrt{\frac{8R_t}{\epsilon}} \right) - P_\pi \left( \phi(Z) > \frac{1}{2} \sqrt{\frac{8R_t}{\epsilon}} \right) \right| \\ &\geq 1 - \frac{\epsilon}{2} - \frac{\epsilon}{2} \\ &= 1 - \epsilon. \end{aligned}$$

□

In Wilson's lemma, the required condition is  $\int (\phi(y) - \phi(x))^2 K(x, dy) \leq C$ . This assumption is stronger than ours. The Poisson/Exponential  $x$ -chain discussed in Section 4 is an example of a Markov chain where this assumption is not satisfied, but the weaker assumption in Theorem 2.3 is satisfied. These lower bounds are still not well understood: For example, we are unable to give a lower bound for the Geometric/Beta example using Theorem 2.3.

### 3 Total Positivity and Monotonicity

As in Section 1, let  $(\mathcal{X}, \mathcal{F})$  and  $(\Theta, \mathcal{G})$  be measurable spaces equipped with  $\sigma$ -finite measures  $\mu$  and  $\nu$  respectively. Let  $\{f_\theta(x)\}_{\theta \in \Theta}$  be a family of probability densities on  $\mathcal{X}$  with respect to  $\mu$ . Let  $\pi(\theta)$  be a probability density on  $\Theta$  with respect to  $\nu$ . The joint, marginal and conditional densities arising from this model are given by (1), (2) and (3) respectively. The marginal  $x$ -chain of the corresponding Gibbs sampler has density (w.r.t.  $\mu$ )

$$k(x, x') = \int f(\theta | x) f(x' | \theta) d\nu(\theta). \quad (12)$$

In this section, we use the properties of totally positive functions of order 2 to derive a useful condition for stochastic monotonicity of the  $x$ -chain.

**Definition 3.1** *Let  $\mathcal{X}, \mathcal{Y} \subseteq \mathbb{R}$ . A function  $L : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}^+$  is said to be **totally positive of order 2** ( $TP_2$ ) if*

$$L(x_1, y_1)L(x_2, y_2) \geq L(x_1, y_2)L(x_2, y_1) \text{ for all } x_1 < x_2, y_1 < y_2.$$

We state as a series of lemmas, some standard facts about  $TP_2$  functions.

**Lemma 3.1** *If  $L(x, y)$  is  $TP_2$  and  $f(x), g(y)$  are non-negative functions, then  $L(x, y)f(x)g(y)$  is  $TP_2$ .*

*Proof* This follows immediately from the definition of  $TP_2$  functions. □

**Lemma 3.2** *If  $K : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}^+$  and  $L : \mathcal{Y} \times \mathcal{Z} \rightarrow \mathbb{R}^+$  are  $TP_2$ , and  $\nu$  is a  $\sigma$ -finite measure on  $\mathcal{Y}$ , then  $M(x, z) = \int K(x, y)L(y, z)d\nu(y)$  is  $TP_2$ .*

*Proof* See Karlin [25, Lemma 3.1.1 (a), pg 99]. □

**Lemma 3.3** *Suppose  $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}^+$  is a Markov kernel. If  $k$  is  $TP_2$ , then the Markov chain corresponding to  $k$  is stochastically monotone.*

*Proof* See Karlin [25, Proposition 1.3.1, pg 22]. □

With these facts in mind, we now state and prove the main result.

**Theorem 3.1** *If  $f_\theta(x)$  is  $TP_2$  (as a function from  $\mathcal{X} \times \Theta$  to  $\mathbb{R}^+$ ), then the  $x$ -chain (12) is stochastically monotone for any choice of  $\pi$  (absolutely continuous with respect to  $\nu$ ).*

*Proof* By Lemma 3.1,  $f(\theta | x) = f_\theta(x)\pi(\theta)/m(x)$  is  $TP_2$ . By Lemma 3.2,

$$k(x, x') = \int f(\theta | x)f(x' | \theta)d\nu(\theta)$$

is  $TP_2$ . Since  $k$  is the transition density of the  $x$ -chain, the  $x$ -chain is stochastically monotone by Lemma 3.3. □

The theory of totally positive functions has applications in various areas of mathematics and statistics. A large collection of probability densities that arise in probability and statistics are totally positive (of all orders and hence in particular  $TP_2$ ). In addition to natural exponential families, the scaled Beta and Non-central t, Non-central chi-square, Non-central F and stable laws on  $(0, \infty)$  with index  $\frac{1}{k}$ ,  $k = 1, 2, 3, \dots$  are totally positive. These and further examples are derived in [25, pg 117–122]. This book contains further examples; for instance if  $X$  is a symmetric random variable with density function  $f$  such that  $f(x - y)$  is totally positive, then the density function of  $|X + u|$  is totally positive (with  $u$  as a parameter). See [25, pg 377–378] for more details. Other useful references about total positivity are [6, 39]. A different application of Theorems 2.1, 3.1 to bounding rates of convergence of Markov chains in ‘carries’ and card shuffling is in [11].

## 4 Exponential Family Examples

In this section we specialize to natural exponential families and conjugate priors. Very generally, these generate stochastically monotone Markov chains having  $x - x^*$  (where  $x^*$  is an appropriately chosen constant) as an eigenfunction.

### 4.1 Exponential Families

Many standard families of probability measures can be represented as exponential families after a reparametrization. For background, see [27] or the references in [13, Section 2]. Let  $\mu$  be a  $\sigma$ -finite measure on the Borel sets of  $\mathbb{R}$ . Let  $\Theta = \{\theta \in \mathbb{R} : \int e^{x\theta} \mu(dx) < \infty\}$ . We assume  $\Theta$  is non-empty and open. The reference measure  $\nu$  on  $\Theta$  is Lebesgue measure. Holder’s inequality shows that  $\Theta$  is convex. Let  $M(\theta) = \log \int e^{x\theta} \mu(dx)$  and define

$$f_\theta(x) = e^{\theta x - M(\theta)}. \tag{13}$$

For  $\theta \in \Theta$ , this is a family of probability densities with respect to  $\mu$ . Making allowable differentiations, we get,

$$\mathbf{E}_\theta(X) = \int x f_\theta(x) \mu(dx) = M'(\theta).$$

Fix  $n_0 > 0$  and  $x^*$  in the interior of the convex hull of the support of  $\mu$ . The family of conjugate priors is defined by

$$\pi(\theta) = z(x^*, n_0) e^{n_0 x^* \theta - n_0 M(\theta)}. \quad (14)$$

Here  $z(x^*, n_0)$  is a normalizing constant for the probability  $\pi(\theta)$  with respect to the Lebesgue measure  $d\theta$  on  $\Theta$ , shown to be positive and finite in [15].

These ingredients produce a bivariate density given by

$$f(x, \theta) = f_\theta(x) \pi(\theta) \text{ with respect to } \mu(dx) \times d\theta. \quad (15)$$

The Gibbs sampler analyzed here is based on the iterations: From  $(x, \theta)$ ,

- Draw  $\theta'$  from  $f(\theta' | x)$
- Draw  $x'$  from  $f(x' | \theta')$ .

Here, Bayes theorem shows that  $f(\theta | x)$  is in the conjugate family, with parameters  $n_0 + 1$  and  $\frac{n_0 x^* + x}{n_0 + 1}$ . The  $x$ -chain has transition density (with respect to  $\mu$ )

$$k(x, x') = \int f(\theta | x) f(x' | \theta) d\theta. \quad (16)$$

By elementary calculations, the  $x$ -chain has stationary density, the marginal density

$$m(x) = \int f_\theta(x) \pi(\theta) d\theta. \quad (17)$$

If the  $x$ -chain is close to stationarity, so is the bivariate chain (4). Indeed [13, Lemma 2.4] gives

$$\|k_x^\ell - m\|_{\text{TV}} \leq \|\tilde{K}_{x,\theta}^\ell - f\|_{\text{TV}} \leq \|k_x^{\ell-1} - m\|_{\text{TV}}, \quad \forall x \in \mathcal{X}, \theta \in \Theta.$$

For exponential families with conjugate priors, we now show that all of the hypotheses of Theorems 2.1 (integer case) or 2.2 (general case) hold.

**Proposition 4.1** *For the exponential family (13) with conjugate prior (14), the  $x$ -chain (16) admits  $x - x^*$  as an eigenfunction with eigenvalue  $\frac{1}{n_0 + 1}$ .*

*Proof* Let  $X_0 = x$  and  $X_1$  be the first two steps of a Markov chain governed by  $k$  of (16). Then,

$$\mathbf{E}(X_1 | X_0 = x) = \mathbf{E}(\mathbf{E}(X_1 | \theta) | X_0 = x) = \mathbf{E}(M'(\theta) | X_0 = x) = \frac{n_0 x^* + x}{n_0 + 1}. \quad (18)$$

The last equality follows from [15, Theorem 2] where it is shown to characterize conjugate priors for families with infinite support. The claim of the theorem is a simple rewriting of (18).  $\square$

*Remark* The proposition shows us that the parameter  $x^*$  is the mean of the marginal density  $m(x)$ . Since  $\frac{1}{n_0+1} < 1$ , and the constant function is an eigenfunction corresponding to the eigenvalue 1 of the Markov chain governed by  $k$ , it follows that,  $\int_{\mathcal{X}}(x - x^*)m(x)d\mu(x) = 0$ . Hence  $x^* = \int_{\mathcal{X}} xm(x)d\mu(x)$ .

*Example* Consider the geometric density on  $\mathcal{X} = \{0, 1, 2, \dots\}$  in the parametrization  $f_p(j) = p(1-p)^j$ . To write this as an exponential family, set  $f_\theta(j) = e^{j \log(1-p) + \log p}$ . Set  $\theta = \log(1-p)$  and  $M(\theta) = -\log(1 - e^\theta)$ . We recognize an exponential family on  $\mathcal{X}$  with  $\mu(j) \equiv 1$ ,  $\Theta = (-\infty, 0)$  and  $M(\theta) = -\log(1 - e^\theta)$ . The conjugate prior on  $\Theta$  has form

$$z(x^*, n_0)e^{n_0 x^* \theta - n_0 M(\theta)}, \quad n_0 > 0, \quad x^* \in (0, \infty).$$

Using the transformation  $p = 1 - e^{-\theta}$  from  $\Theta$  to  $(0,1)$ , we recognize that  $p$  has a Beta( $\alpha, \beta$ ) density with  $\alpha = n_0 + 1$ ,  $\beta = n_0 x^*$ . The restriction  $n_0 > 0$  is exactly what is needed so that the marginal density has a finite mean.

**Proposition 4.2** *The  $x$ -chain for a natural exponential family (13) is stochastically monotone. This remains true if any prior measure is used.*

*Proof* Following Section 3, it is enough to show that the family  $f_\theta(x)$  is totally positive of order 2. Suppose  $\theta_1, \theta_2 \in \Theta$ ;  $x_1, x_2 \in \text{Support}(\mu)$  have  $\theta_1 < \theta_2$ ,  $x_1 < x_2$ . Then since

$$f_{\theta_2}(x_1)f_{\theta_1}(x_2) \leq f_{\theta_1}(x_1)f_{\theta_2}(x_2) \iff e^{(\theta_1 - \theta_2)(x_1 - x_2)} \leq 1,$$

the family  $f_\theta(x)$  is  $TP_2$  by Theorem 3.1.  $\square$

Combining the above results, we obtain the following general result.

**Corollary 4.1** *For the exponential family (13) with conjugate prior (14) and marginal (17),*

(a) *If  $f_\theta(x)$  is supported on the positive integers, then, for any starting state  $x$  and all  $l \geq 0$ ,*

$$\|k_x^l - m\|_{TV} \leq \left(\frac{1}{n_0 + 1}\right)^l (|x| + |x^*|).$$

(b) With general support, for any starting state  $x$ , all  $l \geq 0$ , the Wasserstein distance satisfies

$$|x - x^*| \left( \frac{1}{n_0 + 1} \right)^l \leq d_W(k_x^l, m) \leq \left( \frac{1}{n_0 + 1} \right)^l (|x| + |x^*|).$$

*Proof* Part (a) follows immediately from Theorem 2.1 and Proposition 4.1. The upper bound for Part (b) follows immediately from Theorem 2.2 and Proposition 4.1. For proving the lower bound for Part (b) note that

$$\begin{aligned} & d_W(k_x^l, m) \\ &= \sup \left\{ \left| \int_{\mathcal{X}} \phi(y) k^l(x, y) dy - \int_{\mathcal{X}} \phi(y) m(y) dy \right| : |\phi(x') - \phi(y')| \leq |x' - y'| \forall x', y' \in \mathcal{X} \right\} \\ &\geq \left| \int_{\mathcal{X}} (y - x^*) k^l(x, y) dy - \int_{\mathcal{X}} (y - x^*) m(y) dy \right| \\ &= \left| (x - x^*) \left( \frac{1}{n_0 + 1} \right)^l - 0 \right| \\ &= |x - x^*| \left( \frac{1}{n_0 + 1} \right)^l. \end{aligned}$$

Here, we used Proposition 4.1 and the fact that  $x^*$  is the mean of the marginal density  $m$ .  $\square$

## 4.2 Examples

This section treats various examples in the exponential family setting, where a useful analysis was not available using previous techniques.

*Geometric/Beta* Theorem 1.1 treats this example. The translation in the natural parametrization is given above in Section 4.1 and Theorem 1.1 follows from Corollary 4.1. Here  $x - \beta/(\alpha - 1)$  is an eigenfunction with eigenvalue  $1/\alpha$ .

*Gamma/shape parameter* Consider the Gamma family

$$f_\theta(y) = \frac{y^{\theta-1} e^{-y}}{\Gamma(\theta)} = e^{\theta \log y - \log \Gamma(\theta)} \frac{e^{-y}}{y}, \quad 0 < \theta, y < \infty.$$

The conjugate prior for  $\theta$  has form  $\pi(\theta) = \frac{z(x^*, n_0) e^{n_0 x^* \theta}}{\Gamma(\theta)^{n_0}}$  for  $n_0 > 0$ ,  $x^* \in \mathbb{R}$ . From Proposition 4.2 and Theorem 2.2, for any starting state  $x$

$$d_W(k_x^l, m) \leq \left( \frac{1}{n_0 + 1} \right)^l (x + \mathbf{E}(Z)), \quad \text{where } Z \sim m.$$

This result holds even though the normalizing constant  $z(x^*, n_0)$  is not generally available. Note that the Gamma shape family is not one of the six families treated by Morris [32], [33] and its analysis was not previously available.

*Hyperbolic Cosine* This family was identified in Morris as the sixth family with variance a quadratic function of the mean. For the full parametrization and extensive references see [32, Section 2.4] or [20]. In the parametrization by the mean  $\theta$ , with shape parameter  $r = 1$ ,

$$f_{\theta}(x) = z^{-1} e^{x \tan^{-1} \theta} \beta \left( \frac{1}{2} + \frac{ix}{2}, \frac{1}{2} - \frac{ix}{2} \right) \quad -\infty < x < \infty, \quad (19)$$

with respect to Lebesgue measure. The normalizing constant is  $z = 2\pi(1 + \theta^2)^{r/2}$ . The Beta function  $\beta(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$  is real because  $\overline{\Gamma(a)} = \Gamma(\bar{a})$ . The conjugate prior is

$$\pi(\theta) = z^{-1}(\rho, \delta) \frac{e^{\rho\delta \tan^{-1} \theta}}{(1 + \theta^2)^{\frac{\rho}{2}}}, \quad -\infty < \theta, \delta < \infty, \rho \geq 1. \quad (20)$$

The normalizing constant is  $z^{-1}(\rho, \delta) = \frac{\Gamma(\frac{\rho}{2} - \rho\delta i)\Gamma(\frac{\rho}{2} + \rho\delta i)}{\Gamma(\frac{\rho}{2})\Gamma(\frac{\rho}{2} - \frac{1}{2})\sqrt{\pi}}$ . When  $\theta = 0$ , the resulting density corresponds to  $\frac{2}{\pi} \log |C|$ , where  $C$  is standard Cauchy.

From results in [32, Section 2.4], the marginal density  $m(x)$  has tails of the form  $\frac{c}{|x|^{\rho}}$ . Thus we must take  $\rho > 2$  in order to have  $x - x^*$  as an eigenfunction ( $\rho > 3$  is required for  $(x - x^*) \in L^2(m)$ ). From [32, Theorem 3.1], for  $\rho > 2$ ,  $x^*$  is finite. So  $x - x^*$  is an eigenfunction with eigenvalue  $\frac{1}{(1+\rho-2)}$ . This yields the following result.

**Theorem 4.1** *For the  $x$ -chain corresponding to the hyperbolic cosine density (19) with prior density (20), for any  $x, \delta, \rho \in \mathbb{R}$  and  $\rho > 2$ ,*

$$d_W(k_x^l, m) \leq \left( \frac{1}{\rho - 1} \right)^l \left( |x| + \frac{\rho|\delta|}{\rho - 2} \right).$$

We were unable to treat this example in [13] because only finitely many moments of the marginal density  $m$  exist.

*Poisson/Exponential* We now give an example where sharp bounds were obtained in [13, Section 4.2], and show that we can obtain the same bounds using techniques developed in this paper. Consider the Poisson distribution with a standard exponential prior. Here,

$$f_{\theta}(x) = \frac{e^{-\theta} \theta^x}{x!} \quad x = 0, 1, 2, \dots, \quad \pi(\theta) = e^{-\theta} \theta \in (0, \infty), \quad m(x) = \frac{1}{2^{x+1}}.$$

The  $x$ -chain has kernel  $k(x, y) = 2^{x+1} 3^{-x-y-1} \binom{x+y}{x}$ . The function  $x - 1$  is an eigenfunction of  $k$  with eigenvalue  $1/2$ . From Corollary 4.1, for any starting state  $x$ ,

$$\|k_x^l - m\|_{\text{TV}} \leq (x + 1)2^{-l}.$$

This is essentially the same as results derived using the complete diagonalization of  $k$  in [13, Section 4.2]. Those results show that

$$\|k_x^l - m\|_{\text{TV}} \leq 2^{-1-c} \text{ for } l = \log_2(x + 1) + c, \quad c > 0.$$

A **matching lower bound** showing that, starting from  $x$ ,  $\log_2 x$  steps are needed in total variation follows from Theorem 2.3 applied to the eigenfunction  $f(x) = x - 1$  with eigenvalue  $\lambda = 1/2$ . Elementary calculations show that

$$\sum_{y=0}^{\infty} (f(y) - f(x))^2 k(x, y) = \frac{f^2(x)}{4} + \frac{3}{4}f(x) + \frac{3}{2}.$$

Applying Theorem 2.3 with  $B = \frac{3}{4}$  and  $C = \frac{3}{2}$  gives  $\|k_x^\ell - m\|_{\text{TV}} \geq 1 - \epsilon$  if  $\ell \leq \log_2 |x - 1| + \log_2 \epsilon - \log_2 25$ .

*Beta/Binomial* We now give an example where the techniques developed in this paper lead to bounds that are slightly off. The usual binomial distribution  $\binom{n}{j} p^j (1-p)^{n-j}$  with Beta conjugate prior  $\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1} (1-p)^{\beta-1}$  is transformed to a natural exponential family by taking carrier measure  $\mu(j) = \binom{n}{j}$  on  $\{0, 1, 2, \dots, n-1, n\}$  and letting  $\theta = \log(\frac{p}{1-p})$ . Under this transformation, the conjugate prior (in form (14)) with parameters  $n_0, x^*$  corresponds to a Beta density with parameters  $\alpha = n_0 x^*, \beta = n_0(n - x^*)$ . For example, the uniform prior with  $\alpha = \beta = 1$  results from choosing  $n_0 = \frac{2}{n}, x^* = \frac{n}{2}$ . The  $x$ -chain is a Markov chain on  $\{0, 1, 2, \dots, n-1, n\}$  with transition density

$$k(x, x') = \frac{n+1}{2n+1} \frac{\binom{n}{x} \binom{n}{x'}}{\binom{2n}{x+x'}}, \quad \pi(x) = \frac{1}{n+1}.$$

For this choice, part (a) of Corollary 4.1 gives, for any starting state  $x$

$$\|k_x^l - m\|_{\text{TV}} \leq \left( \frac{n}{n+2} \right)^l \left( x + \frac{n}{2} \right).$$

This bound is slightly off and we discuss it below. A careful application of Theorem 2.1 with  $x = n$  gives a tighter bound  $\|k_n^l - \pi\|_{\text{TV}} \leq n \left(1 - \frac{2}{n+2}\right)^l$ . The standard spectral bound (using the largest eigenvalue  $\frac{n}{n+2}$ ) gives  $\|k_n^l - \pi\|_{\text{TV}} \leq \sqrt{n+1} \left(1 - \frac{2}{n+2}\right)^l$ . Both bounds yield that order  $n \log n$  steps suffice for convergence. The analysis in [13, Proposition 1.1] shows that order  $n$  steps are necessary and sufficient. Turning to numbers, when  $n = 100$ , Theorem 2.1 shows that  $l = 400$  steps suffice to get total variation distance lesser than  $\frac{1}{100}$ . The spectral bound shows that  $l = 350$  steps suffice. Both are practically useful and much much better than what Harris recurrence gives.

*Remark* For five of the six families with quadratic variance structure and their usual conjugate prior, order  $\log |x|$  steps are necessary and sufficient for convergence of the full Gibbs sampler starting at  $(x, \theta)$  (any  $\theta$ ). When comparable, the present approach matches the approach using the full spectrum. However, for continuous problems, the present approach only proves convergence in Wasserstein distance (while the chains converge in total variation). For the binomial family, the full diagonalization shows order  $n$  steps are necessary and sufficient. See Remark 3 in Section 2.1. The present analysis gives an upper bound of order  $n \log n$  for convergence.

## 5 Location Families

In this section, we treat a host of two-component Gibbs samplers arising out of location families. Some examples treated here were analyzed in [13] using explicit diagonalization and bounds were obtained from special starting points. But we show that even in these cases, the coupling techniques presented here can be used to obtain useful bounds from general starting points.

In this section,  $\mu$  either denotes the Lebesgue measure on  $\mathbb{R}$  or counting measure on the integers. We consider  $X = \theta + \varepsilon$  with  $\theta$  having density  $\pi(\theta)$  and  $\varepsilon$  having density  $g(x)$  (both with respect to  $\mu$ ). This can be written as

$$f_\theta(x) = g(x - \theta), \quad f(x, \theta) = g(x - \theta)\pi(\theta) \text{ (w.r.t. } \mu(dx) \times \mu(d\theta)\text{)}.$$

Hence,

$$m(x) = \int g(x - \theta)\pi(\theta)\mu(d\theta), \quad f(\theta | x) = \frac{g(x - \theta)\pi(\theta)}{m(x)}.$$

In [13] a family of ‘conjugate priors’ for  $g$  was suggested. Let  $g$  be the density of the sum of  $r$  independent and identically distributed copies of a random variable  $Z$ . Let  $\pi$  be the density of  $s$  copies of  $Z$ , by elementary manipulations, if  $Z$  has a finite mean,

$$\mathbf{E}(\theta | X) = \frac{s}{r + s}X.$$

Here  $s, r$  are positive integers. If  $Z$  is infinitely divisible,  $s, r$  may be any positive real numbers. For further details and examples, see [13, Section 2.3.3, Section 5].

The  $x$ -chain for the Gibbs sampler proceeds as follows:

- From  $x$  draw  $\theta$  from  $f(\theta | x)$ .
- Then set  $x' = \theta + \varepsilon'$  with  $\varepsilon'$  drawn from  $g$ .

The  $x$ -chain has stationary density  $m(x)$ , the convolution of  $\pi$  and  $g$  (thus  $m$  is the density of the sum of  $r + s$  independent copies of  $Z$ ). We now proceed to give conditions which guarantee that both conditions of Theorem 2.2 are valid.

**Proposition 5.1** *With notation as above, suppose that  $Z$  has finite mean  $z$ . Then the  $x$ -chain has eigenvector  $(x - (r + s)z)$  with eigenvalue  $\frac{s}{s+r}$ .*

*Proof* Let  $X_0 = x$  and  $X_1$  be two successive steps of the  $x$ -chain. Then,

$$\mathbf{E}(X_1 | X_0 = x) = \mathbf{E}(\mathbf{E}(X_1 | \theta) | X_0 = x) = \mathbf{E}(\theta + rz | X_0 = x) = \frac{s}{r + s}x + rz.$$

Use this to solve for  $d$  in

$$\mathbf{E}(X_1 - d | X_0 = x) = \frac{s}{s + r}(x - d).$$

Thus  $\frac{s}{s+r}x + rz - d = \frac{s}{s+r}x - \frac{ds}{s+r}$ , so  $rz = \frac{dr}{s+r}$  and  $d = (s+r)z$  as claimed.  $\square$

*Remark* If  $\mathbf{E}(X_1^2 | X_0 = x) = ax^2 + bx + c$  for some  $a, b, c$ , the density of  $Z$  belongs to one of the six exponential families treated by Morris [32, 33]. Then, the  $x$ -chain has a complete set of polynomial eigenfunctions; these cases are treated in [13]. Of course, there are many other exponential families.

**Proposition 5.2** *With notation as above, the  $x$ -chain is stochastically monotone if the density  $g$  is such that  $-\log(g)$  is convex.*

*Proof* The equivalence of total positivity of order two for the family  $\{g(x-\theta)\}$  and  $-\log(g(x))$  convex is standard fare; see Lehmann and Romano [27, pg 323]. The result now follows from the developments in Section 3.  $\square$

There is a large classical literature on such log concave densities. We gave examples at the end of Section 3 above. See [25, 27] for further examples and developments. For ease of reference we state the conclusions of this section.

**Corollary 5.1** *For the  $x$ -chain for the location family Gibbs sampler, with densities  $g, \pi$  based on  $r, s$  copies of the random variable  $Z$  respectively. Let  $z = \mathbf{E}(Z)$ . Suppose that  $-\log(g)$  is convex.*

(a) *If  $g$  is supported on the integers, then, for every  $x$  and  $l$ ,*

$$\|k_x^l - m\|_{TV} \leq (|x| + (r+s)|z|) \left(\frac{s}{s+r}\right)^l.$$

(b) *If  $g$  is absolutely continuous with respect to the Lebesgue measure on  $\mathbb{R}$ ,*

$$|x - (r+s)z| \left(\frac{s}{s+r}\right)^l \leq d_W(k_x^l, m) \leq (|x| + (r+s)|z|) \left(\frac{s}{s+r}\right)^l.$$

*Proof* Part (a) follows immediately from Theorem 2.1 and Proposition 5.1. The upper bound for Part (b) follows immediately from Theorem 2.2 and Proposition 5.1. The lower bound for Part (b) follows by a similar argument as in the proof of Corollary 4.1.  $\square$

Each of the six exponential families treated in [13, Section 5] is log-concave. We treat two cases.

*Example (Binomial)* For fixed  $p$ ,  $0 < p < 1$ , let  $\pi = \text{Bin}(n_1, p)$ ,  $g = \text{Bin}(n_2, p)$ . Then  $m = \text{Bin}(n_1 + n_2, p)$  and

$$f(\theta | x) = \frac{\binom{n_1}{\theta} \binom{n_2}{x-\theta}}{\binom{n_1+n_2}{x}}$$

is hypergeometric. The  $x$ -chain evolves as

$$X_{n+1} = S_{X_n} + \varepsilon_{n+1}$$

with  $S_{X_n}$  a hypergeometric with parameters  $n_1, n_2, X_n$  and  $\varepsilon_{n+1}$  drawn from  $\text{Bin}(n_2, p)$ . We verify that  $g$  is  $TP_2$  by checking

$$g(x' - \theta)g(x - \theta') \leq g(x - \theta)g(x' - \theta') \quad (21)$$

for integers  $x < x'$  and  $\theta < \theta'$ . Note that if  $x < \theta'$  then  $g(x - \theta') = 0$ . Similarly for  $x' - \theta > n$ . In these cases, (21) holds trivially. Hence assume  $\theta < \theta' \leq x < x'$  and  $x' - \theta \leq n$ . Then, after obvious simplifications, (21) is equivalent to

$$\begin{aligned} & (x - \theta)(x - \theta - 1) \dots (x - \theta' + 1)(n - (x' - \theta'))(n - (x' - \theta) + 1) \\ & \leq (x' - \theta)(x' - \theta - 1) \dots (x' - \theta' + 1)(n - (x - \theta'))(n - (x - \theta) + 1) \end{aligned}$$

This last inequality holds because  $x < x'$  and  $n - x' < n - x$ . This yields the following result.

**Theorem 5.1** *For all  $n_1, n_2 > 0$ ,  $0 < p < 1$ , the  $x$ -chain for the binomial location model satisfies*

$$\|k_x^l - m\|_{TV} \leq (x + (n_1 + n_2)p) \left( \frac{n_1}{n_1 + n_2} \right)^l. \quad (22)$$

*Remark* In [13, Section 5.1] this example was treated by a full diagonalization. For the case treated there, the starting state is  $x = 0$  and the results are essentially the same for the full range of choices of  $n_1, n_2, p$ . We note that the spectral approach requires bounding the orthogonal polynomials (here Krawtchouck polynomials) at the starting state  $x$ . This can be a difficult task for  $x \neq 0$ . Along these lines, consider the case where the  $x$ -chain is started at the center of the stationary density  $m$ . For simplicity, consider  $n_1 = n_2 = n$  and  $p = 1/2$ . Then the mean is  $n$  and Theorem 2.1 gives

$$\|k_x^l - m\|_{TV} \leq \mathbf{E}|Y - n| \left( \frac{1}{2} \right)^l, \text{ where } Y \sim \text{Bin}(2n, \frac{1}{2}).$$

Using deMoivres' formula for the mean absolute deviation [17],

$$\mathbf{E}|Y - n| = n \binom{2n}{n} \frac{1}{2^{2n}} \sim \sqrt{\frac{n}{\pi}}.$$

This gives a slight improvement over (22).

*Example (Hyperbolic)* This example was treated analytically in [13, Section 5.6] but was left unfinished because of the intractable nature of the eigenfunctions (Meixner - Pollaczek polynomials). We treat a special case which seems less foreign than the general case. Let  $\pi$  and  $g$  have the density of  $\frac{2}{\pi} \log |C|$  with  $C$  standard Cauchy. Thus from [13, Section 2]

$$\pi(x) = g(x) = \frac{1}{2 \cosh(\frac{\pi x}{2})} \text{ w.r.t. Lebesgue measure on } \mathbb{R}. \quad (23)$$

The marginal density is the density of  $\frac{2}{\pi} \log |C_1 C_2|$ , that is,

$$m(x) = \frac{x}{2 \sinh\left(\frac{\pi x}{2}\right)}. \quad (24)$$

By symmetry, the mean of  $m(x)$  is zero and  $x$  is an eigenfunction. We may easily verify that  $x_1 < x_2$ ,  $\theta_1 < \theta_2$  imply  $g(x_1 - \theta_2)g(x_2 - \theta_1) \leq g(x_1 - \theta_1)g(x_2 - \theta_2)$ . Indeed this is equivalent to  $(e^{(x_1 - \theta_1)} + e^{(\theta_1 - x_1)})(e^{(x_2 - \theta_2)} + e^{(\theta_2 - x_2)}) \leq (e^{(x_1 - \theta_2)} + e^{(\theta_2 - x_1)})(e^{(x_2 - \theta_1)} + e^{(\theta_1 - x_2)})$ , which is equivalent to  $(e^{x_2 - x_1} - e^{x_1 - x_2})(e^{\theta_2 - \theta_1} - e^{\theta_1 - \theta_2}) \geq 0$ , which of course is true. Using Corollary 5.1 along with the fact that the mean of  $m(x)$  is zero, gives us the following exact formula for the hyperbolic location model.

**Theorem 5.2** *The  $x$ -chain for the location family for the hyperbolic model, with  $n_1 = n_2 = 1$  satisfies, for any starting state  $x$  and all  $l \geq 1$ ,*

$$d_W(k_x^l, m) = |x|2^{-l}.$$

## 6 Further probabilistic Bounds

The theorems above make crucial use of stochastic monotonicity and the availability of an eigenfunction. In this section, more basic forms of the stochastic techniques of coupling and strong stationary times are used. The examples analyzed here include the well studied M/M/ $\infty$  queue, and some multivariate examples, where techniques using orthogonal polynomials provide bounds only for special starting points. All of the problems treated here are location models and we use the notation of Sections 4 and 5 without further comment.

*Example (Normal Location Model).* This is a location model with

$$g(x) = \frac{e^{-\frac{x^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}, \quad \pi(\theta) = \frac{e^{-\frac{(\theta-r)^2}{2\xi^2}}}{\sqrt{2\pi\xi^2}}, \quad (25)$$

where  $\sigma, r$  and  $\xi$  are parameters,  $\sigma, \xi > 0$ ,  $r \in \mathbb{R}$ . This leads to the marginal density

$$m(x) = \frac{e^{-\frac{(x-r)^2}{2(\sigma^2 + \xi^2)}}}{\sqrt{2\pi(\sigma^2 + \xi^2)}}. \quad (26)$$

The  $x$ -chain for the Gibbs sampler is a classical autoregressive process which may be represented as

$$X_{n+1} = aX_n + \varepsilon_{n+1} \text{ with } a = \frac{\xi^2}{\sigma^2 + \xi^2} \text{ and } \{\varepsilon_i\}_{i \geq 1} \text{ i. i. d. } N\left(\frac{\sigma^2 r}{\sigma^2 + \xi^2}, \sigma^2\right). \quad (27)$$

Consider the Markov chain in (27) with  $X_0 = x$ . Then

$$X_1 = ax + \varepsilon_1, \quad X_2 = a^2x + a\varepsilon_1 + \varepsilon_2, \quad \dots, \quad X_n = a^n x + a^{n-1}\varepsilon_1 + \dots + \varepsilon_n. \quad (28)$$

The stationary distribution may be represented as the infinite convolution

$$X_\infty = \varepsilon'_0 + a\varepsilon'_1 + a^2\varepsilon'_2 + \dots \quad (29)$$

for any independent sequence  $\{\varepsilon'_i\}_{i \geq 1}$  with common distribution  $N(\frac{\sigma^2 r}{\sigma^2 + \xi^2}, \sigma^2)$ . This yields the following result.

**Theorem 6.1** *For the  $x$ -chain for the Gibbs sampler location model (27), started at  $x$ ,*

$$d_W(k_x^l, m) \leq |x|a^l + \frac{a^l}{1-a} \left( \sigma + \frac{\sigma^2 r}{\sigma^2 + \xi^2} \right).$$

*Proof* To couple  $X_\ell$  and  $X_\infty$ , let  $(\varepsilon'_i)$  be a i.i.d.  $N(\frac{\sigma^2 r}{\sigma^2 + \xi^2}, \sigma^2)$  and, for a fixed  $l$ , set  $\varepsilon_i = \varepsilon'_{l-i}$ . Then use  $(\varepsilon_i)_1^l, (\varepsilon_i)_1^\infty$ , in (28), (29), respectively. This gives  $d_W(k_x^l, m) \leq \mathbf{E}|X_l - X_\infty|$ . To obtain the desired bound, use the simple fact that

$$\mathbf{E}|N| \leq (|\mu| + \sigma),$$

if  $N \sim N(\mu, \sigma^2)$ . □

*Remarks 1.* It can be checked that the  $x$ -chain for the Gibbs sampler location model (27) is stochastically monotone, and has  $f(x) = x - r$  as an eigenfunction corresponding to the eigenvalue  $a$ . Using this fact and proceeding along the same lines as the proof for Part (b) of Corollary 4.1 leads to the following bound.

$$|x - r|a^l \leq d_W(k_x^l, m) \leq (|x| + |r|)a^l.$$

2. The theorem gives essentially the same results as the detailed calculations of [13, Section 4.3]. Here we work in a different norm - the  $L^1$  Wasserstein distance. The coupling inherent in the proof is not the optimal coupling for the Wasserstein distances. For example, the optimal coupling for the  $L^2$  Wasserstein distance (indeed, for any convex distance) between two Gaussian measures (see [22]) on the line is achieved by the unique affine map determined by matching means and variances. With Frank Barth's help, we computed that the  $L^2$  Wasserstein distance to stationarity starting at  $x$  after  $n$  steps equals  $a^{2n}x^2 + (\sqrt{1 - a^{2n}} - 1)^2/(1 - a^2)$ . This differs infinitesimally from the bound above.

3. Note that for  $f(x) = x - r$ ,

$$\mathbf{E} [(f(X_{n+1}) - f(x))^2 | X_n = x] = (a - 1)^2 f^2(x) + \sigma^2.$$

Hence by Theorem 2.3 it follows that  $\|k_x^l - m\|_{\text{TV}} \geq 1 - \epsilon$  for  $\ell \leq \frac{\log|x-r| + \log \epsilon - \log \sqrt{8(\xi^2 + \sigma^2)}}{-\log a}$ . This provides a **matching lower bound** for the chi-square (and hence total variation) upper bound provided in [13, Theorem 4.3].

Figure 1

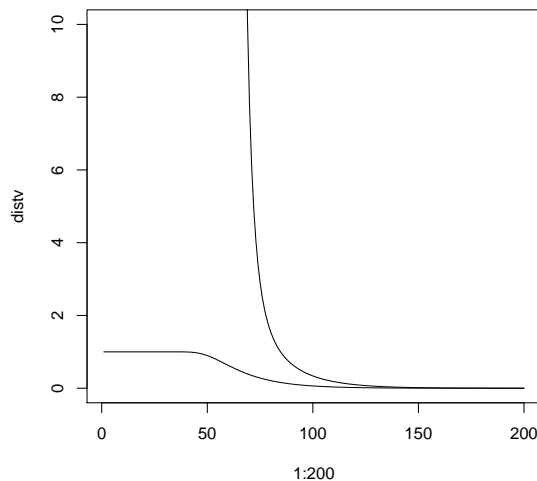


Figure 1: Plot of total variation distance as a function of the number of steps for a particular instance of the Markov chain considered in Theorem 6.1

5. From (28), the law of  $X_l$  is normal with mean  $a^l x + \frac{1-a^l}{1-a} \frac{\sigma^2 r}{\sigma^2 + \xi^2}$  and variance  $\frac{1-a^{2l}}{1-a^2} \sigma^2$ . The stationary distribution is normal with mean  $\frac{\sigma^2 r}{(1-a)(\sigma^2 + \xi^2)}$  and variance  $\frac{\sigma^2}{1-a^2}$ . Thus exact total variation distance calculations are also available in terms of the distance between two Gaussians. In a bit more detail, if  $X_1$  is  $N(\mu_1, \sigma_1^2)$  and  $X_2$  is  $N(\mu_2, \sigma_2^2)$ , the total variation distance between  $X_1$  and  $X_2$  is the same as the total variation distance between a standard normal variate  $Z$  and  $X$ , a  $N(\mu, \sigma^2)$  variate with  $\mu = \frac{(\mu_2 - \mu_1)}{\sigma_1}$ ,  $\sigma = \frac{\sigma_2}{\sigma_1}$ . The densities of  $Z$  and  $X$  cross at the two points  $x_{\pm} = \frac{\mu \pm \sqrt{\sigma^2 \mu^2 - (1 - \sigma^2) \sigma^2 \log \sigma^2}}{(1 - \sigma^2)}$ . Now,

$$\|X - Z\|_{\text{TV}} = |\Phi_{0,1}(x_+) - \Phi_{\mu,\sigma}(x_+)| + |\Phi_{0,1}(x_-) - \Phi_{\mu,\sigma}(x_-)|, \quad (30)$$

with  $\Phi_{\mu,\sigma}$  the cumulative distribution function of  $N(\mu, \sigma^2)$ . For the chain of Theorem 6.1, a plot of total variation distance as a function of  $l$  is shown in Figure 1 when  $r = 0$ ,  $\sigma^2 = \frac{1}{4}$ ,  $\xi^2 = 4$  for starting state  $x = 100$ . The same plot shows the exact chi-squared distance for these same parameters. The chi-squared distance is frequently used as an upper bound for the squared total variation distance. The figure shows this is a poor bound for  $l$  small and quite accurate for large  $l$ .

5. The same analysis obtains if  $x$  and  $\theta$  are  $d$ -dimensional. Of course, useful bounds on the vector norm in terms of the input parameters will be more difficult.

*Example (Gamma Location Model).* For  $0 < x, \theta < \infty$ ,  $0 < i_1, i_2, \sigma < \infty$ , let

$$g(x) = \frac{x^{i_1-1} e^{-\frac{x}{\sigma}}}{\sigma^{i_1} \Gamma(i_1)}, \quad \pi(\theta) = \frac{\theta^{i_2-1} e^{-\frac{\theta}{\sigma}}}{\sigma^{i_2} \Gamma(i_2)}. \quad (31)$$

The marginal density for the  $x$ -component of the Gibbs sampler is

$$m(x) = \frac{x^{i_1+i_2-1} e^{-\frac{x}{\sigma}}}{\sigma^{i_1+i_2} \Gamma(i_1+i_2)}. \quad (32)$$

Some elementary manipulations yield the following representation for the  $x$ -process.

$$X_{n+1} = A_{n+1}X_n + \varepsilon_{n+1}, \quad A_{n+1} \sim \text{Beta}(i_2, i_1), \quad \varepsilon_{n+1} \sim \text{Gamma}(i_1, \sigma), \quad (33)$$

where  $\{A_i\}_{1 \leq i < \infty}$  and  $\{\varepsilon_i\}_{1 \leq i < \infty}$  are all mutually independent. This leads to the following representation of the stationary distribution.

$$X_\infty = \varepsilon'_0 + A'_1 \varepsilon'_1 + A'_2 A'_1 \varepsilon'_2 + \dots \quad (34)$$

for some  $A'_i, \varepsilon'_i, 1 \leq i < \infty$ , which are all mutually independent and  $A'_i \sim \text{Beta}(i_2, i_1), \varepsilon'_i \sim \text{Gamma}(i_1, \sigma)$  for every  $1 \leq i < \infty$ . This yields an obvious coupling and gives the following result.

**Theorem 6.2** *For the  $x$ -chain (33) for the Gibbs sampler location model started at  $x$ ,*

$$|x - \sigma(i_1 + i_2)| \left( \frac{i_1}{i_1 + i_2} \right)^\ell \leq d_W(k_x^\ell, m) \leq (x + \sigma(i_1 + i_2)) \left( \frac{i_2}{i_1 + i_2} \right)^\ell.$$

*Proof* Let  $\{A'_i\}_{1 \leq i < \infty}, \{\varepsilon'_i\}_{1 \leq i < \infty}$  be mutually independent i.i.d. sequences as in (34). For a fixed  $l$ , let  $A_i = A'_{l-i}, \varepsilon_i = \varepsilon'_{l-i}$  and use these in (33). Then,

$$\begin{aligned} d_W(k_x^\ell, m) &\leq \mathbf{E}|X_l - X_\infty| \\ &= \mathbf{E}|A'_l A'_{l-1} \dots A'_1 x - \sum_{j=l}^{\infty} \left( \prod_{i=1}^j A'_i \right) \varepsilon'_j| \\ &\leq \left( \frac{i_2}{i_1 + i_2} \right)^\ell (x + \sigma(i_1 + i_2)). \end{aligned}$$

Since  $f(x) = x - \mathbf{E}X_\infty$  is an eigenfunction with eigenvalue  $\frac{i_2}{i_1+i_2}$ , it follows that

$$d_W(k_x^\ell, m) \geq |\mathbf{E}(X_\ell - X_\infty)| = |x - \sigma(i_1 + i_2)| \left( \frac{i_2}{i_1 + i_2} \right)^\ell.$$

□

*Remark* Lest the reader think that all of the classical families will yield to such a probabilistic approach, consider the case when  $g$  and  $\pi$  are geometric distributions of the form  $\theta(1-\theta)^j$  on  $\{0, 1, 2, \dots\}$ . The marginal  $m(j)$  is then negative binomial  $(2, \theta)$ . The conditional density  $f(\theta | x)$  is uniform on  $\{0, 1, 2, \dots, x-1, x\}$ . The  $x$ -chain can be represented as

$$X_{n+1} = [U_{n+1}(X_n + 1)] + \varepsilon_{n+1}$$

with  $U$  a (continuous) uniform variate on  $(0, 1)$  and  $\varepsilon$  a geometric ( $\theta$ ) variate. Here  $\lfloor x \rfloor$  is the largest integer smaller than  $x$ . The backward iteration does not appear simple to work with. This problem is solved by diagonalization in [13, Section 5.3] and monotonicity in Section 5 above.

*Example (Poisson Location Model)* For  $0 < r, s, \lambda < \infty$ , let

$$g(j) = \frac{e^{-r\lambda}(r\lambda)^j}{j!}, \quad \pi(j) = \frac{e^{-s\lambda}(s\lambda)^j}{j!}, \quad j = 0, 1, 2, \dots \quad (35)$$

The marginal density is

$$m(j) = \frac{e^{-(r+s)\lambda}((r+s)\lambda)^j}{j!}, \quad j = 0, 1, 2, \dots \quad (36)$$

The  $x$ -chain for the Gibbs sampler may be represented as

$$X_{n+1} = S_{X_n} + \varepsilon_{n+1} \quad \text{with} \quad S_x \sim \text{Bin}\left(x, \frac{s}{r+s}\right), \quad \varepsilon \sim \text{Poisson}(r\lambda), \quad (37)$$

and corresponds to the well studied M/M/ $\infty$  queue, as explained below.

**Lemma 6.1** *For the Poisson location chain (37), started at  $x$ ,*

$$X_n \sim P(r\lambda) * P(\rho r\lambda) * \dots * P(\rho^{n-1}r\lambda) * \text{Bin}(x, \rho^n)$$

and

$$X_\infty \sim P(r\lambda) * P(\rho r\lambda) * P(\rho^2 r\lambda) * \dots$$

where  $*$  stands for convolution,  $P(\lambda)$  stands for Poisson( $\lambda$ ), the variates are independent, and  $\rho = s/(s+r)$ . Thus  $X_\infty \sim P((r+s)\lambda)$ .

*Proof* The  $x$ -chain may be pictured as starting with  $x$  customers. Each time, a coin with probability of heads  $\rho = s/(s+r)$  is flipped for each current customer. Customers whose coin comes up heads disappear. Then Poisson( $r\lambda$ ) new customers are added. This is the classical M/M/ $\infty$  queue in discrete time. At stage  $n$ , the number of original customers remaining is Bin( $x, \rho^n$ ). The number of first stage customers remaining is Poisson( $\rho^{n-1}r\lambda$ ), and so on.  $\square$

As above, these considerations yield the following result.

**Theorem 6.3** *For the  $x$ -chain for the Gibbs sampler location model (37) started at  $x$ ,*

$$\|k_x^l - m\|_{TV} \leq d_W(k_x^l, m) \leq \left(\frac{s}{r+s}\right)^l (x + (r+s)\lambda).$$

*Proof* Because the variables are integer valued,  $\|k_x^l - m\|_{\text{TV}} \leq d_W(k_x^l, m)$ . Letting  $X_\ell$  and  $X_\infty$  be as in Lemma 6.1, we get  $d_W(k_x^l, m) \leq \mathbf{E}(|X_\ell - X_\infty|)$  and a simple computation yields

$$\mathbf{E}(|X_\ell - X_\infty|) \leq \left(\frac{s}{r+s}\right)^l (x + (r+s)\lambda),$$

proving the desired result.  $\square$

Hence  $\frac{\log x}{\log\left(\frac{r+s}{s}\right)}$  steps are sufficient for convergence. A **matching lower bound** can be obtained by noting that for the eigenfunction  $f(x) = x - (r+s)\lambda$  with eigenvalue  $\frac{s}{r+s}$ ,

$$\mathbf{E}[(f(X_{t+1}) - f(x))^2 \mid X_t = x] = \left(\frac{r}{r+s}\right)^2 f^2(x) + \frac{rsx}{(r+s)^2} + r\lambda.$$

By Theorem 2.3 it follows that  $\|k_x^\ell - m\|_{\text{TV}} \geq 1 - \epsilon$  if  $t \leq \frac{\log|x-(r+s)\lambda| + \log\epsilon - \log(4 + \sqrt{16 + 8(r+s)\lambda})}{\log\left(\frac{r+s}{s}\right)}$ .

*Example (Binomial Location Model)* The following example was treated in Section 5 using an eigenvector and monotonicity and in [13] using a complete diagonalization. The following direct coupling gives essentially the same results AND, as mentioned at the end, extends to the vector valued case. Let  $r, s$  be positive integers and fix  $p, 0 < p < 1$ . Set

$$g(x) = \binom{r}{x} p^x (1-p)^{r-x}, \quad \pi(\theta) = \binom{s}{\theta} p^\theta (1-p)^{s-\theta}. \quad (38)$$

The marginal density of the  $x$ -chain is

$$m(x) = \binom{r+s}{x} p^x (1-p)^{r+s-x}, \quad 0 \leq x \leq r+s. \quad (39)$$

The  $x$ -chain proceeds as follows: From  $X_n$  choose  $\theta_{n+1}$  from the hypergeometric distribution

$$f(\theta \mid X_n) = \frac{\binom{s}{\theta} \binom{r}{X_n - \theta}}{\binom{r+s}{X_n}}, \quad (X_n - r)_+ \leq \theta \leq \min(X_n, s).$$

Set

$$X_{n+1} = \theta_{n+1} + \varepsilon_{n+1} \text{ with } \varepsilon \sim \text{Bin}(r, p). \quad (40)$$

**Theorem 6.4** *For any  $r, s \geq 1$ , any starting state  $x$ , and all  $p$ , the  $x$ -chain (40) satisfies,*

$$\|k_x^l - m\|_{\text{TV}} \leq s \left(\frac{s}{r+s}\right)^{l-1} \text{ for all } l \geq 2.$$

*Proof* We construct a strong stationary time for the process lifted to binary vectors. See [1, 3, 10] for background on strong stationary times. Consider the following process on binary vectors of length  $r + s$ . Let  $X_n$  be the number of ones at time  $n$ . Let  $\theta_{n+1}$  be the number of ones in the first  $s$  coordinates after applying a random permutation to the vector. Finally, flip a  $p$ -coin for each of the last  $r$  coordinates and replace what was there by these outcomes. Evidently, the process  $X_0 = x, X_1, X_2, \dots$  is the same as (40). After one step, the last  $r$  coordinates have the correct stationary distribution. Let  $T$  be the first time after time 1 that all coordinates among the first  $s$  have been replaced by coordinates from  $(s + 1, s + 2, \dots, s + r)$ . This  $T$  is clearly a strong stationary time for the lifted process: If the coordinate process is  $Z_n = (Z_n^1, Z_n^2, \dots, Z_n^{r+s})$ ,

$$P(Z_n = z \mid T = t) = p^{|z|}(1 - p)^{r+s-|z|}, \quad |z| = z^1 + z^2 + \dots + z^{r+s}.$$

To bound  $T$ , let  $B_i$ ,  $1 \leq i \leq s$  be the event that all permutations up to and including time  $l$ , have kept coordinate  $i$  between 1 and  $s$ . Then

$$P(T \geq l + 1) = P(\cup_{i=1}^s B_i) \leq sP(B_1) = s \left( \frac{s}{r + s} \right)^{l-1}.$$

The desired result follows because, for any strong stationary time  $T$ ,

$$\|k_x^l - m\|_{\text{TV}} \leq P(T \geq l + 1).$$

See [1, 3, 10]. □

*Remarks 1.* Consider the case of  $r = 1$ . The first  $s$  coordinates evolve as follows: Choose a coordinate at random and replace it by a flip of a  $p$  coin; this is Glauber dynamics for the product measure. When  $p = 1/2$ , it becomes the Ehrenfest urn with holding  $1/2$ . The bound above shows that  $(s + 1) \log s$  steps suffice. This is the right order of magnitude, but off by a factor of  $1/2$ , see [8].

2. A similar argument can be carried through for the multivariate analog based on the Multinomial distributions

$$\binom{r}{x_1, x_2, \dots, x_k} p_1^{x_1} p_2^{x_2} \dots p_k^{x_k}, \quad \binom{s}{x_1, x_2, \dots, x_k} p_1^{x_1} p_2^{x_2} \dots p_k^{x_k}.$$

Lift to a process on vectors of length  $r + s$  with entries in  $\{1, 2, \dots, k\}$ . The same argument with  $X_n = (X_n^1, X_n^2, \dots, X_n^k)$  for  $X_n^i$  the number of coordinates taking value  $i$ , leads to exactly the same bound as in Theorem 6.4, uniformly in  $p_1, p_2, \dots, p_k$  and  $k$ . An exact analytic solution using multivariate orthogonal polynomials for the Multinomial is in [26], which provides bounds for special starting points. However, the analysis presented here works for general starting points.

*Acknowledgments* We thank Frank Barth and Peter Hall for their help with this paper.

## References

- [1] Aldous, D. and Diaconis, P. (1986). Shuffling cards and stopping times. *Amer. Math. Monthly* **93**, 333-348.
- [2] Athreya, K., Doss, H. and Sethuraman, J. (1996). On the convergence of the Markov chain simulation method, *Ann. Statist.* **24**, 89-100.
- [3] Aldous, D. and Diaconis, P. (1987). Strong Uniform times and Finite Random Walks, *Advances in Applied Math.* **8**, 69-97.
- [4] Barankin, E.W. and Maitra, A.P. (1963). Generalization of the Fisher-Darmois-Koopman-Pitman theorem on sufficient statistics, *Sankhya Ser. A* **25**, 217-244.
- [5] Berti, P., Consonni, G. and Pratelli, L. (2009). Discussion on the paper: Gibbs sampling , exponential families and orthogonal polynomials, *Statistical Science* **23**, 179-182.
- [6] Brown, L.D., Johnstone, I.M. and McGibbon, K.B. (1981). Variance diminishing transformations: A direct approach to total positivity and its statistical applications, *J. Amer. Statist. Assn.* **76**, 824-832.
- [7] Casella, G. and George, E. (1992). Explaining the Gibbs sampler, *American Statistician* **46**, 167-174.
- [8] Diaconis, P. (1988). *Group Representations in Probability and Statistics*, Institute of Mathematical Statistics - Monograph Series, Hayward, California.
- [9] Diaconis, P. and Andersen, H. (2007). Hit and run as a unifying device, *Journal de la Societ Franaise de Statistique* **148**, 5-28.
- [10] Diaconis, P. and Fill, J. (1990). Strong Stationary Times via a New Form of Duality, *Ann. Prob.* **16**, 1483-1522.
- [11] Diaconis, P. and Fulman, J. (2008). Carries, shuffling and an amazing matrix, *American Mathematical Monthly* **116**, 788-803.
- [12] Diaconis, P. and Fulman, J. (2009). Carries, shuffling, symmetric function, *Advances of Applied Mathematics* **43**, 176-196.
- [13] Diaconis, P., Khare, K. and Saloff-Coste L. (2008). Gibbs sampling, exponential families and orthogonal polynomials, *Statistical Science* **23**, 151-178.
- [14] Diaconis, P. and Saloff-Coste, L. (1993). Comparison Theorems for Reversible Markov Chains, *Ann. Appl. Prob.* **3**, 696-730.

- [15] Diaconis, P. and Ylvisaker, D. (1979). Conjugate priors for exponential families, *Ann. Stat.* **7**, 269-281.
- [16] Diaconis, P. and Ylvisaker, D. (1985). Quantifying Prior Opinion, In J.M. Bernardo, M.H. Degroot, D.V. Lindley, A.F.M. Smith, eds., *Bayesian Statistics 2: Proc. 2<sup>nd</sup> Valencia Int'l Meeting*, North Holland, Amstredam, 133-156.
- [17] Diaconis, P. and Zabell, S. (1991). Closed Form Summation for Classical Distributions: Variations on a theme of Demoiivre, *Statistical Sci.* **61**, 284-302.
- [18] Dudley, R.M. (1989). *Real Analysis and Probability*, Wadsworth, Belmont, CA.
- [19] Dyer, M., Goldberg, L., Jerrum, M. and Martin, R. (2005). Markov chain comparison, *Probability Surveys* **3**, 89-111.
- [20] Esch, D. (2003). The skew-t distribution: Properties and computations, Ph.D. Dissertation, Department of Statistics, Harvard University.
- [21] Fill, J. and Machida, M. (2001). Stochastic Monotonicity and Realizable Monotonicity, *Annals of Applied Probability* **29**, 938-978.
- [22] Givens, C. and Shortt, R. (1984) A class of Wasserstein metrics for probability distributions, *Michigan Math. J.* **31**, 231-240.
- [23] Jones, G.L. and Hobert, J.P. (2001). Honest exploration of intractable probability distributions via Markov chain Monte Carlo, *Statist. Sci.* **16**, 312-334.
- [24] Jones, G.L. and Hobert, J.P. (2004). Sufficient burn-in for Gibbs samplers for a hierarchical random effects model, *Annals of Statistics* **32**, 784-817.
- [25] Karlin, S. (1968). *Total Positivity*, Stanford University Press, Stanford.
- [26] Khare, K. and Zhou, H. (2009). Rates of convergence of some multivariate Markov chains with polynomial eigenfunctions, *Annals of Applied Probability* **19**, 737-777.
- [27] Lehmann, E. and Romano, J. (2005). *Testing Statistical Hypotheses*, Springer, New York.
- [28] Liu, J. (2001). *Monte Carlo Strategies in Scientific Computing*, Springer-Verlag, New York.
- [29] Liu, J., Wong, W. and Kong, A. (1995). Covariance structure and convergence rates of the Gibbs sampler with various scans, *Jour. Roy. Statist. Soc. B*, 157-169.
- [30] Lund, R.B. and Tweedie, R.L. (1996). Geometric Convergence Rates for Stochastically Ordered Markov Chains, *Mathematics of Operations Research*, **20**, 182-194.
- [31] Meyn, S.P. and Tweedie, R.L. (1993). *Markov chains and stochastic stability*, Springer-Verlag, London.

- [32] Morris, C. (1982). Natural exponential families with quadratic variance functions, *Ann. Statist.* **10**, 65-80.
- [33] Morris, C. (1983). Natural exponential families with quadratic variance functions: Statistical Theory, *Ann. Statist.* **11**, 515-589.
- [34] Rosenthal, J.S. (1995). Minorization conditions and convergence rates for Markov chain Monte Carlo, *Jour. Amer. Statist. Assoc.* **90**, 558-566.
- [35] Rosenthal, J.S. (1996). Analysis of the Gibbs sampler for a model related to James-Stein estimations, *Statist. Comput.* **6**, 269-275.
- [36] Rosenthal, J.S. (2002). Quantitative convergence rates of Markov chains: A simple account, *Electronic Communications in Probability* **7**, 123-128.
- [37] Roy, V. and Hobert, J.P. (2006). Convergence rates and asymptotic standard errors for MCMC algorithms for Bayesian probit regression, *Jour. Roy. Statist. Soc. B* **69**, 607-623.
- [38] Saloff-Coste, L. (2004). Total variation lower bounds for finite Markov chains: Wilson's lemma, *Random walks and geometry*, Walter de Gruyter GmbH and Co. KG, Berlin, 515-532.
- [39] Stanley, R. (1989). Unimodal and log-concave sequences in algebra, combinatorics and geometry, in *Graph Theory and Its Applications: East and West*, Ann. New York Acad. Sci. **576**, 500-535.
- [40] Stoyan, D. (1983). *Comparison Methods for Queues and other Stochastic Models*, John Wiley and Sons, New York.
- [41] Tierney, L. (1994). Markov chains for exploring posterior distributions (with discussion), *Ann. Statist.* **22**, 1701-1762.
- [42] Wilson, D.B. (2004). Mixing times of lozenge tiling and card shuffling Markov chains, *Annals of Applied Probability*, **14**, 274-325.
- [43] David Wilson's Website on Perfect Sampling "<http://research.microsoft.com/%20dbwilson/exact/>".