Gibbs Sampling, Conjugate Priors and Coupling

P. Diaconis *
Departments of Mathematics and Statistics
Stanford University

Kshitij Khare
Department of Statistics
Stanford University

L. Saloff-Coste†
Department of Mathematics
Cornell University

August 31, 2008

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. Many of them seem difficult to succesfully analyze using spectral or Harris recurrence techniques.

1 Introduction

The Gibbs sampler is an important tool in computational statistics. It gives a way to sample from a multivariate density $f(x_1, x_2, \cdots, x_p)$, perhaps known only up to a normalizing constant, using a sequence of one dimensional samples. From $\underline{x} = (x_1, x_2, \cdots, x_p)$ go to $(x_1', x_2, \cdots, x_p)$, then $(x_1', x_2', \cdots, x_p)$, and so on until $(x_1', x_2', \cdots, x_p') = \underline{x}'$, where at the i^{th} stage the coordinate is drawn from f with the other coordinates fixed. Iterating this gives a Markov chain $\underline{x}, \underline{x}', \underline{x}'', \cdots$ with f as stationary density under mild conditions [2, 36]. The running time analysis of the Gibbs sampler (how many steps should the chain be run to be close to stationarity) is an active area of research [13, 19, 20, 22, 24, 29, 30, 31, 32].

This paper considers two component examples of the form

$$f(x, \theta) = f_\theta(x)\pi(\theta) \tag{1}$$

*Research partially supported by NSF grant DMS-0505673.
†Research partially supported by NSF grant DMS 0603886.

with $f_\theta(x)$ a standard family of probability densities and $\pi(\theta)$ a prior density. For these, use of off the shelf tools can give unrealistic answers – in one example ($f_\theta(j) = \binom{n}{j}\theta^j(1-\theta)^{n-j}, \pi(\theta) =$ Uniform) the off the shelf tools involving Harris recurrence techniques suggest 10^{30} steps are needed for $n = 100$, while simulation and later theory showed a few hundred steps suffice. We previously developed theory that worked for a handful of cases (the six classical exponential families with conjugate priors) but then broke down.

In this paper we show that many problems (eg. essentially all one dimensional exponential families with conjugate priors) have a natural monotonicity property *and* one explicit eigenfunction which allows sharp analysis. Here is an example of our main results.

Example (Geometric/Beta) Let $\mathcal{X} = \{0, 1, \ldots\}$, $\Theta = (0, 1)$ and, for fixed $\alpha, \beta > 0$, set

$$f_\theta(j) = \theta(1-\theta)^j,\ j = 0, 1, 2, ...;\ \ \pi(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1},\ \theta \in (0, 1).$$

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 θ given j) densities are

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

The Gibbs sampler proceeds from (j, θ) as usual

- Pick θ' from $f(\theta \mid j)$.
- Pick j' from $f_{\theta'}(j')$.

This gives a Markov chain $\widetilde{K}(j, \theta; j', \theta')$ with stationary density $f(j, \theta)$. Measuring convergence in total variation distance (see below), we show

Theorem 1.1 *For the Geometric/Beta density with $\alpha > 1,\ \beta > 0$, all starting states (j, θ) and all ℓ,*

$$\|\widetilde{K}^\ell_{j,\theta} - f\|_{\text{TV}} \le \left(j + \frac{\beta}{\alpha-1}\right)\left(\frac{1}{\alpha}\right)^\ell. \tag{2}$$

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 $j = 100, \theta = \frac{1}{2}$. The right side of (2) is smaller than 0.01 for $\ell = 14$. Figure 1 shows a run of the j-component of the (j, θ) chain starting at $j = 100$.

Comment We were unable to analyze this example in [13] because the marginal density $m(j)$ has all moments beyond the mean infinite (we used orthogonal polynomials). We further treat this

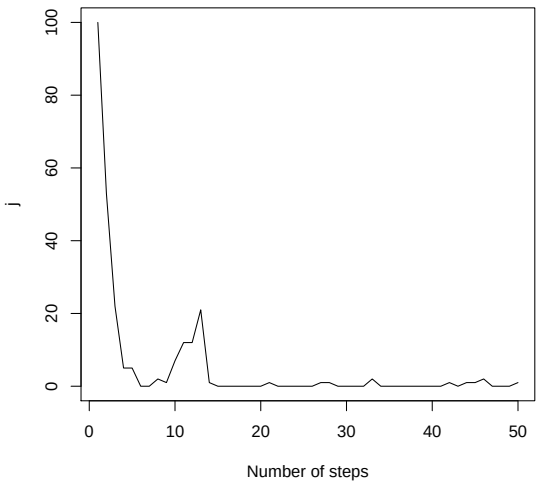


Figure 1: *Run of j-chain for Geometric/Beta starting at $j = 100$.*

example below using Harris recurrence techniques. These show the chain is close to stationarity after 1400 steps.

We conclude the section by giving needed background. Section 2 presents two new theorems for monotone Markov chains. It also gives a needed extension of the second moment method and Wilson's lemma for proving lower bounds. Section 3 introduces total positivity – our main tool for showing monotonicity. Section 4 treats general one dimensional exponential families with conjugate priors. Section 5 treats location families. In many cases we are able to establish sharp upper and lower bounds. In Section 6 similiar problems are treated directly (eg. without using monotonicity) via probabilistic techniques – strong stationary times and coupling. The examples include some queueing systems and some multivariate problems.

It is worth emphasizing that the present collection of examples are just illustrative. It is easy to sample from any of our $f(x, \theta)$ distributions directly (for example, sample θ from $\pi(\theta)$ and then sample x from $f_\theta(x)$). Further, we do not see how to extend present techniques to three or more component Gibbs samplers.

1.1 Notation and background

Let $(\mathcal{X}, \mathcal{F})$ and $(\Theta, \mathcal{G})$ be measurable spaces equipped with σ-finite measures μ and ν respectively. Let $\{f_\theta(x)\}_{\theta \in \Theta}$ be a family of probability densities on $\mathcal{X}$ with respect to μ. Let $\pi(\theta)$ be a probability density on Θ with respect to ν. These determine a joint density

$$f(x, \theta) = f_\theta(x)\pi(\theta) \text{ w.r.t. } \mu \times \nu. \tag{3}$$

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

$$m(x) = \int f_\theta(x)\pi(\theta)\nu(d\theta). \tag{4}$$

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

$$f(x \mid \theta) = f_\theta(x) \text{ and } f(\theta \mid x) = \frac{f(x,\theta)}{m(x)} \tag{5}$$

The Gibbs sampler is an algorithm for drawing samples from $f(x,\theta)$ when it is easy to sample from $f(x \mid \theta)$ and $f(\theta \mid x)$. This is how it proceeds:
From (x,θ)

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

This defines a Markov chain with transition density

$$\tilde{K}(x,\theta\ ;\ x',\theta') = f(\theta' \mid x)f(x' \mid \theta') \tag{6}$$

with respect to $\mu(dx') \times \nu(d\theta')$. Under mild conditions always met in our examples, this Markov chain is ergodic with stationary measure $f(x,\theta)\mu(dx)\nu(d\theta)$.

For two component chains, the 'x-chain' (from x draw θ' from $\pi(\theta' \mid x)$ and then x' from $f_{\theta'}(x')$) has transition kernel

$$k(x,x') = \int_\Theta \pi(\theta \mid x) f_\theta(x')\pi(d\theta) = \int_\Theta \frac{f_\theta(x)f_\theta(x')}{m(x)}\pi(d\theta). \tag{7}$$

Observe that $\int k(x,x')\mu(dx') = 1$ so that $k(x,x')$ is a probability density with respect to μ. Note further that $m(x)k(x,x') = m(x')k(x',x)$ so that the x-chain has $m(dx)$ as stationary distribution. The total variation distance between two densities f and g (with respect to a σ-finite measure μ) is defined as

$$\|f - g\|_{\text{TV}} = \frac{1}{2}\int |f(x) - g(x)|\mu(dx).$$

If the x-chain is close to stationarity, so is the bivariate chain (6). Indeed, [13, Lemma 2.4] gives

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

Figure 1.1 shows the total variation distance of the j-chain of the Geometric/Beta example as a function of the number of steps.
An elaborate developement of these ideas with many further details and references is in [13, Section 2].

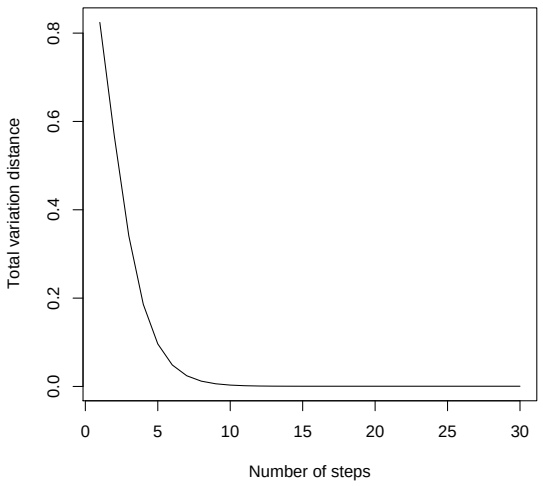


Figure 2: *Total variation distance for Geometric/Beta j-chain with* $\alpha = 2, \beta = 1$*, starting at* $j = 100$*.*

The most widely used technique for bounding rates of convergence for the Gibbs sampler are the Harris recurrence techniques of Meyn-Tweedie [26] and Rosenthal [29]. For a splendid introduction and literature review see [19]. We illustrate by using a version specially tailored to our two component Gibbs samplers [4].

Theorem 1.2 *[4, Proposition 3] Let* $k(x, y)$ *be the density of the x-chain (7). Suppose that k is ergodic with* $m(x)$ *as unique stationary distribution. Suppose*

$$\int k(x, x')\phi(x')\mu(dx') \leq a + b\phi(x) \quad \textit{for all } x \in \mathcal{X}.$$

For some measurable function $\phi : \mathcal{X} \to [0, \infty)$ *and constants* $a, b \in (0, 1)$*. Fix* $d > \frac{2a}{1-b}$*. Define* $A = \{x \in \mathcal{X} : \phi(x) \leq d\}$*. Suppose that* $\sup_{x \in A} m(x) < \infty, \inf_{A \times B} f(x, \theta) > 0$ *for some* $B \in \mathcal{G}$ *with* $\int_{A \times B} f_\theta(x)\pi(d\theta)\mu(dx) > 0$*. Then, for all* $r \in (0, 1)$ *and* $x \in \mathcal{X}$

$$\|k_x^\ell - m\|_{TV} \leq (1 - \epsilon)^{r\ell} + t^\ell \left(1 + \frac{a}{1-b} + \phi(x)\right)$$

with $\epsilon = \frac{\pi(B) \inf_{A \times B} f}{\sup_A \pi}$ *and* $t = \frac{(1+2a+2bd)^r (1+2a+bd)^{1-r}}{(1+d)^{1-r}}$.

To use this, suitable choices of ϕ, d, B must be found. This is a matter of art (and some computer experimentation) at this writing.

Example (Geometric/Beta) In the setting of the theorem above, $f_\theta(j) = \theta(1-\theta)^j, \pi(\theta) = 2\theta, m(j) = \frac{4}{(j+1)(j+2)(j+3)}$. The j-chain has density

$$k(j,j') = \frac{3(j+1)(j+2)(j+3)}{(j+j'+1)(j+j'+2)(j+j'+3)(j+j'+4)}.$$

For $B = [\theta_*, \theta^*], 0 < \theta_*, \theta^* < 1$, $A = \{0, 1, \cdots, d+1\}$, $\sup_x m(x) = m(0) = \frac{2}{3}$, $\inf_{A\times B} f(j,\theta) = 2\theta_*(1-\theta^*)^{d+1}$, $\pi(B) = \int_{\theta_*}^{\theta^*} 2\theta = (\theta^*)^2 - (\theta_*)^2$. The function $j-1$ is an eigenvector for the j-chain with eigenvalue $\frac{1}{2}$. This gives $a = 0, b = \frac{1}{2}$. Choosing $d = 1, \theta_* = \frac{1}{2}, \theta^* = \frac{3}{4}, r = \frac{1}{10}$, we find $\epsilon \geq \frac{3}{2}\left((\theta^*)^2 - (\theta_*)^2\right) 2\theta_*(1-\theta^*)^{d+1} = 0.0293$, $t = (1+d)^{2r-1}\left(1+\frac{d}{2}\right)^{1-r} = 0.8273$. This gives

$$\|k_j^\ell - m\|_{\text{TV}} \leq (0.9707)^{\frac{\ell}{10}} + (0.8273)^\ell j \quad \forall \ell \geq 1.$$

When $j = 100$, this is smaller than 0.01 for $\ell \geq 1400$, while our bounds show that 14 steps suffice.

Remark Despite the disparity, we regard this as quite a useful conclusion. For the seemingly similiar Binomial/Beta problem, the Harris recurrence bound gives 10^{30} as the required number of steps, while a more detailed analysis shows $\ell = 200$ steps suffice.

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

$$K(x, (-\infty, y]) \geq K(x', (-\infty, y]) \text{ for all } y. \tag{8}$$

Monotone Markov chains have been thoroughly studied and applied. See [17, 25, 35] and the references there. They are currently in vogue because of 'coupling from the past'. See [38] for extensive references on this subject.There is a standard coupling technique available for monotone Markov chains. Wilson [37] 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 π 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, π) 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 function f such that*

$$Kf = \lambda f + \eta \tag{9}$$

and

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

Then, for any starting state x,

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

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

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

See, e.g., [14].

Theorem 2.2 *Let (K, π) 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 function f such that*

$$Kf = \lambda f + \eta \tag{11}$$

and

$$c = \inf\left\{\frac{f(y) - f(x)}{y - x} \mid x, y \in \mathcal{X}, x < y\right\} > 0. \tag{12}$$

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 (9)-(11) is an eigenfunction of the chain with eigenvalue λ. It is instructive to compare the conclusions of the theorems with standard spectral bounds when λ is the second largest eigenvalue. For a Markov chain on the integers, the usual spectral bound is $\|K_x^l - \pi\|_{\text{TV}} < \pi(x)^{-\frac{1}{2}}\lambda^l$. For the Geometric/Beta x-chain 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 $x = 0, 1, 2, ...$), $\lambda = 1/2$. Theorem 2.1 gives an upper bound $\|K_x^l - \pi\|_{\text{TV}} \leq x2^{-l+1}$. The spectral bound $\|K_x^l - \pi\|_{\text{TV}} \leq 2^{x-l+2}$ 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. Here is a natural example where they are slightly off. Consider the x-chain

for the Beta/Binomial with a uniform prior (c.f. [13, Section 2]). This is a Markov chain on $\{0, 1, 2, ..., n-1, n\}$ with transition density

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

This is a monotone Markov chain with $kf = \left(1 - \frac{2}{n+2}\right) f$ for $f(x) = x - n/2$. As shown below in Section 4.2, Theorem 2.1 shows $\|k_n^l - \pi\|_{\text{TV}} \leq n\left(1 - \frac{2}{n+2}\right)^l$. The spectral bound shows $\|k_n^l - \pi\|_{\text{TV}} \leq \sqrt{n+1}\left(1 - \frac{2}{n+2}\right)^l$. 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.

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 $f(j) = j \quad (\mathbf{E}[j' \mid j] = j + 1)$, 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.: $spec(k) \cap (-1, 1) \subseteq [-\beta^*, \beta^*]$ 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 Support(π). Define a bivariate Markov chain $\{R_n, S_n\}_{n=0}^\infty$ as follows: Set $R_0 = x,\ S_0 = x'$. Let $U_1, U_2, ...$ 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

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

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 and is an extension of Wilson's lemma (see [37, Lemma 5] or [33, Theorem 4.13]).

Theorem 2.3 *Let K be an ergodic Markov kernel with stationary probability measure π. 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 ϕ. Under the given hypothesis,

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

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

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

Note that since $\pi K^t = \pi \ \ \forall t \geq 0$,

$$\begin{aligned}\int \phi(x)\pi(dx) &= \int \mathbf{E}_x[\phi(X_t)]\pi(dx) \\ &= \lambda^t \int \phi(x)\pi(dx) \to 0 \text{ as } t \to \infty.\end{aligned}$$

Hence $\int \phi(x)\pi(dx) = 0$. Similarily,

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

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

$$\begin{aligned}&\lambda^t\phi(x) \geq \frac{T^*}{\epsilon} \\ \Rightarrow\ &\mathbf{E}_x[\phi(X_t)] \geq \frac{T^*}{\epsilon} \\ \Rightarrow\ &\mathbf{E}_x^2[\phi(X_t)] \geq \frac{8}{\epsilon}\left(\frac{B}{\lambda(1-\lambda)}\mathbf{E}_x[\phi(X_t)] + \frac{C}{1-\lambda^2}\right) \\ &\left(\because \text{ The largest root of } a^2 - \frac{8Ba}{\epsilon\lambda(1-\lambda)} - \frac{8C}{\epsilon(1-\lambda^2)} \text{ is less than } \frac{T^*}{\epsilon}\right) \\ \Rightarrow\ &\mathbf{E}_x[\phi(X_t)] \geq \sqrt{\frac{8R_t}{\epsilon}}.\end{aligned}$$

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

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

Hence we get,

$$\begin{aligned} \|K_x^t - \pi\|_{\text{TV}} &= \sup_A \left|K_x^t(A) - \pi(A)\right| \\ &\geq 1 - \frac{\epsilon}{2} - \frac{\epsilon}{2} \\ &= 1 - \epsilon. \end{aligned}$$

□

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 σ-finite measures μ and ν respectively. Let $\{f_\theta(x)\}_{\theta\in\Theta}$ be a family of probability densities on $\mathcal{X}$ with respect to μ. Let $\pi(\theta)$ be a probability density on Θ with respect to ν. The joint, marginal and conditional densities arising from this model are given by (3), (4) and (5) respectively. The marginal x-chain of the corresponding Gibbs sampler has density (w.r.t. μ)

$$k(x,x') = \int f(\theta \mid x) f(x' \mid \theta) d\nu(\theta). \tag{13}$$

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 $L : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}^+$ 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 $K : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}^+$ and $L : \mathcal{Y}\times\mathcal{Z} \to \mathbb{R}^+$ are TP_2, and ν is a σ-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 [21, Lemma 3.1.1 (a), pg 99]. □

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

Proof See Karlin [21, 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* (13) *is stochastically monotone for any choice of* π*.*

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

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

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 $\frac{1}{k}$, $k = 1, 2, 3, ...$ are totally positive. These and further examples are derived in [21, 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 [21, pg 377–378] for more details. Other useful references about total positivity are [5, 34]. A different application of Theorems 2.1, 3.1 to bounding rates of convergence of Markov chains in 'carries' and card shuffling is in [10].

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 $x - x_0$ 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 [23] or the references in [13, Section 2]. Let μ be a σ-finite measure on the Borel sets of $\mathbb{R}$. Let $\Theta = \{\theta \in \mathbb{R} : \int e^{x\theta}\mu(dx) < \infty\}$. We assume Θ is non-empty and open. The reference measure ν on Θ is Lebesgue measure. Holder's inequality shows that Θ is convex. Let $M(\theta) = \log \int e^{x\theta}\mu(dx)$ and define

$$f_\theta(x) = e^{\theta x - M(\theta)}. \tag{14}$$

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

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

Fix $n_0 > 0$ and x_0 in the interior of the convex hull of the support of μ. The family of conjugate priors is defined by

$$\pi(\theta) = z(x_0, n_0)e^{n_0 x_0 \theta - M(\theta)}. \tag{15}$$

Here $z(x_0, n_0)$ is a normalizing constant for the probability $\pi(\theta)$ with respect to the Lebesgue measure $d\theta$ on Θ, shown to be positive and finite in [7].

These ingredients produce a bivariate density given by

$$f(x, \theta) = f_\theta(x)\pi(\theta) \text{ with respect to } \mu(dx) \times d\theta. \tag{16}$$

The Gibbs sampler is based on the iterations: From (x, θ),

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

Here, Bayes theorem shows that $f(\theta \mid x)$ is in the conjugate family, with parameters $n_0 + 1$ and $\frac{n_0 x_0 + x}{n_0 + 1}$. The x-chain has transition density (with respect to μ)

$$k(x, x') = \int f(\theta \mid x) f(x' \mid \theta) d\theta. \tag{17}$$

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

$$m(x) = \int f_\theta(x)\pi(\theta)d\theta. \tag{18}$$

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* (14) *with conjugate prior* (15)*, the x-chain* (17) *admits* $x - x_0$ *as an eigenfunction with eigenvalue* $\frac{n_0}{n_0+1}$.

Proof Let $X_0 = x$ and X_1 be the first two steps of a Markov chain governed by k of (17). Then,

$$\mathbf{E}(X_1 \mid X_0 = x) = \mathbf{E}(\mathbf{E}(X_1 \mid \theta) \mid X_0 = x) = \mathbf{E}(M'(\theta) \mid X_0 = x) = \frac{n_0 x_0 + x}{n_0 + 1}. \tag{19}$$

The last equality follows from [7, 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 (19). □

Remark While obvious in retrospect, the proposition shows us that the parameter x_0 is the mean of the marginal density $m(x)$.

Example Consider the geometric density on $\mathcal{X} = \{0, 1, 2, ...\}$ 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 Θ has form

$$z(x_0, n_0) e^{n_0 x_0 \theta - n_0 M(\theta)}, \quad n_0 > 0,\ x_0 \in (0, \infty).$$

Using the map $T(\theta) = 1 - e^{-\theta}$ from Θ to (0,1), we recognize a Beta$(\alpha, \beta\ ;\ p)$ density with $\alpha = n_0 + 1,\ \beta = n_0 x_0$. 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* (14) *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

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

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

Combining the above results, we obtain the following result.

Theorem 4.1 *For the exponential family* (14) *with conjugate prior* (15) *and marginal* (18)*,*

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

$$\|k_x^l - m\|_{\text{TV}} \le \left(\frac{n_0}{n_0 + 1}\right)^l (|x| + |x_0|).$$

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

$$|x - x_0| \left(\frac{n_0}{n_0 + 1}\right)^l \le d_W(k_x^l, m) \le \left(\frac{n_0}{n_0 + 1}\right)^l (|x| + |x_0|).$$

4.2 Examples

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 Theorem 4.1. Here $x-\beta/(\alpha-1)$ is an eigenfunction with eigenvalue $1/\alpha$.

Poisson/Exponential Consider the Poisson distribution with a standard exponential prior. This case was treated in [13, Section 4.2]. Here,

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

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,

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

This is essentially the same as results derived using the complete diagonalization of k. Those results show that

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

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 $\|k_x^\ell - m\|_{\text{TV}} \ge 1 - \epsilon$ if $\ell \le \log_2 |x - 1| + \log_2 \epsilon - \log_2 25$.

Beta/Binomial The usual binomial distribution $\binom{n}{j}p^j(1 - p)^{n-j}$ with Beta conjugate prior $\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1 - x)^{\beta-1}$ is transformed to a natural exponential family by taking carrier measure $\mu(j) = \binom{n}{j}$ on $\{0, 1, 2, ..., n - 1, n\}$ and letting $\theta = \log(\frac{p}{1-p})$. Under this transformation, the conjugate prior (in form (15)) with parameters $n_0,\ x_0$ corresponds to a Beta density with parameters $\alpha = n_0x_0,\ \beta = n_0(n - x_0)$. For example, the uniform prior with $\alpha = \beta = 1$ results from choosing $n_0 = \frac{2}{n}, x_0 = \frac{n}{2}$. For this choice, part (a) of Corollary 4.1 gives, for any starting state x

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

This is off by a factor of $\log n$ as discussed in Remark 3 following Theorem 2.2.

Gamma/shape parameter Consider the Gamma family

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

The conjugate prior for θ has form $\pi(\theta) = \frac{z(x_0,n_0)e^{n_0 x_0 \theta}}{\Gamma(\theta)^{n_0}}$ for $n_0 > 0,\ x_0 \in \mathbb{R}$. From Proposition 4.2 and Theorem 2.2, for any starting state x

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

This result holds even though the normalizing constant $z(x_0, n_0)$ is not generally available. Note that the Gamma shape family is not one of the six families treated by Morris [27], [28] 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 [27, Section 2.4] or [16]. In the parametrization by the mean θ, with shape parameter $r = 1$,

$$f_\theta(x) = z^{-1} e^{x \tan^{-1}\theta} \beta\left(\frac{1}{2} + \frac{ix}{2},\ \frac{1}{2} - \frac{ix}{2}\right)\ -\infty < x < \infty, \tag{20}$$

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

$$\pi(\theta) = z^{-1}(\rho,\delta)\frac{e^{\rho\delta\tan^{-1}\theta}}{(1+\theta^2)^{\frac{\rho}{2}}},\ -\infty < \theta,\ \delta < \infty,\ \rho \geq 1. \tag{21}$$

The normalizing constant is $z^{-1}(\rho,\delta) = \frac{\Gamma\left(\frac{\rho}{2} - \rho\delta i\right)\Gamma\left(\frac{\rho}{2} + \rho\delta i\right)}{\Gamma\left(\frac{\rho}{2}\right)\Gamma\left(\frac{\rho}{2} - \frac{1}{2}\right)\sqrt{\pi}}$. When $\theta = 0$, the resulting density corresponds to $\frac{2}{\pi}\log|C|$, where C is standard Cauchy.
From results in [27, 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 $x - x_0$ as an eigenfunction ($\rho > 3$ is required for $(x - x_0) \in L^2(m)$). From [27, Theorem 3.1], for $\rho > 2,\ x_0$ is finite. So $x - x_0$ is an eigenfunction with eigenvalue $\frac{1}{(1+\rho-2)}$. This yields the following result.

Theorem 4.2 *For the x-chain corresponding to the hyperbolic cosine density* (20) *with prior density* (21)*, 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.

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, θ) (any θ). 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 μ is the Lebesgue measure on $\mathbb{R}$ or counting measure on the integers. We consider $X = \theta + \varepsilon$ with θ having denstiy $\pi(\theta)$ and ε having density $g(x)$ (both with respect to μ). This can be written as

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

Hence,

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

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 π be the density of s copies of Z, by elementary manipulations, if Z has a finite mean,

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

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 θ from $f(\theta \mid x)$.
- Then set $x' = \theta + \varepsilon'$ with ε' drawn from g.

The x-chain has stationary density $m(x)$, the convolution of π and g (thus m is the density of $r+s$ copies of Z). We now proceed to give conditions which guarentee 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,

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

Use this to solve for d in

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

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. □

Remark If $\mathbf{E}(X_1^2 \mid X_0 = x) = ax^2 + bx + c$ for some a, b, c, the density of Z belongs to one of the six exponential families treated by Morris [27, 28]. Then, the x-chain has a complete set of polynomial eigenfunctions; these cases are treated in [13].

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 [23, pg 323]. The result now follows from the developements in Section 3. □

There is a large classical literature on such log concave densities. We give examples at the end of Section 3 above. See [21], [23] for further examples and developements. We content ourselves with a few examples below. 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, π 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,

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

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

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

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 $\pi = \mathrm{Bin}(n_1, p)$, $g = \mathrm{Bin}(n_2, p)$. Then $m = \mathrm{Bin}(n_1 + n_2, p)$ and

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

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 ε_{n+1} drawn form $\mathrm{Bin}(n_2, p)$. We verify that g is TP_2 by checking

$$g(x'-\theta)g(x-\theta') \le g(x-\theta)g(x'-\theta') \tag{22}$$

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, (22) holds trivially. Hence assume $\theta < \theta' \le x < x'$ and $x' - \theta \le n$. Then, after obvious simplifications, (22) is equivalent to

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

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*

$$\|k_x^l - m\|_{\mathrm{TV}} \le (x + (n_1+n_2)p)\left(\frac{n_1}{n_1+n_2}\right)^l. \tag{23}$$

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

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

Using deMoivres' formula for the mean absolute deviation [12],

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

This gives a slight improvement over (23).

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 π and g have the density of $\frac{2}{\pi}\log|C|$ with C standard Cauchy. Thus from [13, Section 2]

$$\pi(x) = g(x) = \frac{1}{2\cosh(\frac{\pi x}{2})} \text{ w.r.t. Lebesgue measure on } \mathbb{R}. \tag{24}$$

The marginal density is the density of $\frac{2}{\pi}\log|C_1C_2|$, that is,

$$m(x) = \frac{x}{2\sinh(\frac{\pi x}{2})}. \tag{25}$$

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 $g(x_1-\theta_2)g(x_2-\theta_1) \le g(x_1-\theta_1)g(x_2-\theta_2)$. Indeed this is equivalent to $(e^{(x_1-\theta_1)}+e^{(\theta_1-x_1)})(e^{(x_2-\theta_2)}+e^{(\theta_2-x_2)}) \le (e^{(x_1-\theta_2)}+e^{(\theta_2-x_1)})(e^{(x_2-\theta_1)}+e^{(\theta_1-x_2)})$, which is equivalent to $(e^{x_2-x_1}-e^{x_1-x_2})(e^{\theta_2-\theta_1}-e^{\theta_1-\theta_2}) \ge 0$, which of course is true. Thus, Corollary 5.1 yields the following result.

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 $l \ge 1$,*

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

6 Further probabilistic Bounds

The theorems above make crucial use of stochastic monotonicity and the availibility of an eigenfunction. In this section, more basic forms of the stochastic techniques of coupling and strong stationary times are used. 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

$$g(x) = \frac{e^{-\frac{x^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}, \qquad \pi(\theta) = \frac{e^{-\frac{(\theta-r)^2}{2\xi^2}}}{\sqrt{2\pi\xi^2}} \tag{26}$$

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

$$m(x) = \frac{e^{-\frac{(x-r)^2}{2(\sigma^2+\xi^2)}}}{\sqrt{2\pi(\sigma^2+\xi^2)}}. \tag{27}$$

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

$$X_{n+1} = aX_n + \varepsilon_{n+1} \text{ with } a = \frac{\xi^2}{\sigma^2+\xi^2} \text{ and } \{\varepsilon_i\}_{i\ge1} \text{ i. i. d. } \mathrm{N}\left(\frac{\sigma^2 r}{\sigma^2+\xi^2}, \sigma^2\right). \tag{28}$$

Consider the Markov chain in (28) with $X_0 = x$. Then

$$X_1 = ax + \varepsilon_1,\ X_2 = a^2x + a\varepsilon_1 + \varepsilon_2,\ ...,\ X_n = a^nx + a^{n-1}\varepsilon_1 + ... + \varepsilon_n. \tag{29}$$

The stationary distribution may be represented as the infinite convolution

$$X_\infty = \varepsilon_0' + a\varepsilon_1' + a^2\varepsilon_2' + ... \tag{30}$$

for any independent sequence $\{\varepsilon_i'\}_{i\ge1}$ with common distribution $\mathrm{N}(\frac{\sigma^2 r}{\sigma^2+\xi^2}, \sigma^2)$. This yields the following result.

Theorem 6.1 *For the x-chain for the Gibbs sampler location model* (28)*, started at* x*,*

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

Proof To couple X_ℓ and X_∞, let (ϵ_i') be a i.i.d. $\mathrm{N}(\frac{\sigma^2 r}{\sigma^2+\xi^2}, \sigma^2)$ and, for a fixed l, set $\epsilon_i = \epsilon_{l-i}'$. Then use $(\epsilon_i)_1^l$, $(\epsilon_i')_1^\infty$, in (29), (30), respectively. This gives $d_W(k_x^l, m) \le \mathbf{E}|X_l - X_\infty|$. To obtain the desired bound, use the simple fact that

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

if $N \sim \mathrm{N}(\mu, \sigma^2)$ with $\mu \ge 0$. □

Remarks 1. The theorem gives essentially the same results as the detailed calculations of [13, Section 4.3]. Here we work in a weaker norm. The coupling inherent in the proof is not the optimal coupling for the Wasserstein distance. For two Gaussian measures on the line the optimal coupling for the L^2 Wasserstein distance (indeed, for any convex distance) is achieved by the unique affine map determined by matching means and variances.. With Frank Barthes' 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. For more on L^2 Wasserstein for Gaussian vectors, see [18].

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

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

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

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

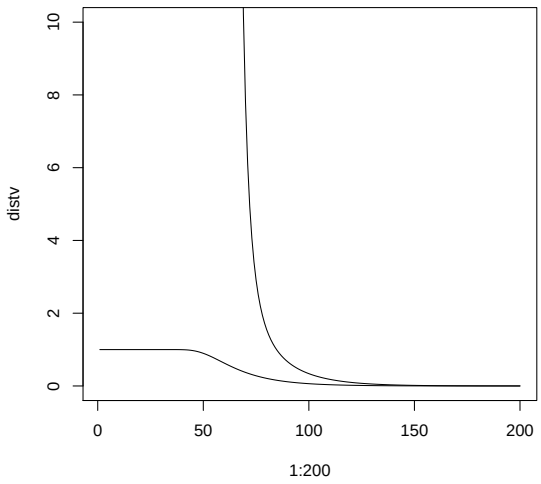
4. From (29), the law of X_l is normal with mean $a^l x + \frac{1-a^l}{1-a}\frac{\sigma^2 r}{\sigma^2+\xi^2}$ and variance $\frac{1-a^{2l}}{1-a^2}\sigma^2$. The stationary distribution is normal with mean $\frac{\sigma^2 r}{(1-a)(\sigma^2+\xi^2)}$ and variance $\frac{\sigma^2}{1-a^2}$. Thus exact total 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,

$$\|X - Z\|_{\mathrm{TV}} = |\Phi_{0,1}(x_+) - \Phi_{\mu,\sigma}(x_+)| + |\Phi_{0,1}(x_-) - \Phi_{\mu,\sigma}(x_-)| \tag{31}$$

with $\Phi_{\mu,\sigma}$ the cumulative distribution function of $N(\mu, \sigma^2)$. A plot of total variation distance as a function of l is shown in Figure 1 below 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.

Figure 1



4. Essentially the same analysis holds for any autoregressive process; ε does not have to be normal. Of course, these will not arise from the Gibbs sampler in any generality.

Example (Gamma Location Model). For $0 < x, \theta < \infty$, $0 < n_1, n_2, \sigma < \infty$, let

$$g(x) = \frac{x^{n_1-1}e^{-\frac{x}{\sigma}}}{\sigma^{n_1}\Gamma(n_1)}, \quad \pi(\theta) = \frac{\theta^{n_2-1}e^{-\frac{\theta}{\sigma}}}{\sigma^{n_2}\Gamma(n_2)}. \tag{32}$$

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

$$m(x) = \frac{x^{n_1+n_2-1}e^{-\frac{x}{\sigma}}}{\sigma^{n_1+n_2}\Gamma(n_1+n_2)}. \tag{33}$$

The classical relations between Beta and Gamma densities yield the following representation for the x-process.

$$X_{n+1} = A_{n+1}X_n + \varepsilon_{n+1}, \;\; A_{n+1} \sim \text{Beta}(n_1,\, n_2),\; \varepsilon_{n+1} \sim \text{Gamma}(n_1,\, \sigma),\; \text{independent.} \tag{34}$$

Further, the stationary distribution can be represented as

$$X_\infty = \varepsilon'_0 + A'_1\varepsilon'_1 + A'_2A'_1\varepsilon'_2 + ... \tag{35}$$

with A'_i, ε'_i independent and distributed as in (34). This yields an obvious coupling and gives the following result.

Theorem 6.2 *For the x-chain (34) for the Gibbs sampler location model started at x,*

$$\left|x - \frac{n_1(n_1+n_2)}{n_2}\right| \left(\frac{n_1}{n_1+n_2}\right)^{\ell} \le d_W(k_x^l, m) \le \left(\frac{n_1}{n_1+n_2}\right)^{l} \left(x + \frac{n_1(n_1+n_2)}{n_2}\right).$$

Proof Let A_i', ϵ_i' be independent i.i.d. sequences as in (35). For a fixed l, let $A_i = A_{l-i}', \epsilon_i = \epsilon_{l-i}'$ and use these in (34). Then,

$$\begin{aligned} d_W(k_x^l, m) &\le \mathbf{E}|X_l - X_\infty| = \mathbf{E}|A_l' A_{l-1}' ... A_1' x - \sum_{j=l}^{\infty} (\prod_{i=1}^{j} A_i') \varepsilon_j'| \\ &\le \left(\frac{n_1}{n_1+n_2}\right)^{l} \left(x + \frac{n_1(n_1+n_2)}{n_2}\right). \end{aligned}$$

Since $f(x) = x - \mathbf{E}X_\infty$ is an eigenfunction with eigenvalue $\frac{n_1}{n_1+n_2}$, it follows that

$$d_W(k_x^\ell, m) \ge |\mathbf{E}(X_\ell - X_\infty)| = \left|x - \frac{n_1(n_1+n_2)}{n_2}\right| \left(\frac{n_1}{n_1+n_2}\right)^{\ell}.$$

□

Remarks 1. The remarks following Theorem (6.1) hold with minor changes for this case as well.

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

$$X_{n+1} = \lfloor U_{n+1}(X_n + 1) \rfloor + \varepsilon_{n+1}$$

with U a (continuous) uniform variate on $(0, 1)$ and ε a geometric (θ) 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 $0 < r, s < \infty$, let

$$g(j) = \frac{e^{-r\lambda}(r\lambda)^j}{j!}, \quad \pi(j) = \frac{e^{-s\lambda}(s\lambda)^j}{j!}, \quad j = 0, 1, 2, ... \tag{36}$$

The marginal density is

$$m(j) = \frac{e^{-(r+s)\lambda}((r+s)\lambda)^j}{j!}, \quad j = 0, 1, 2, ... \tag{37}$$

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

$$X_{n+1} = S_{X_n} + \varepsilon_{n+1} \text{ with } S_x \sim \text{Bin}\left(x, \frac{s}{r+s}\right), \ \varepsilon \sim \text{Poisson}(r\lambda). \tag{38}$$

Lemma 6.1 *For the Poisson location chain* (38)*, started at* x*,*

$$X_n \sim \mathrm{P}(r\lambda) * \mathrm{P}(\rho r\lambda) * ... * \mathrm{P}\left(\rho^{n-1} r\lambda\right) * \mathrm{Bin}\left(x,\ \rho^n\right)$$

and

$$X_\infty \sim \mathrm{P}(r\lambda) * \mathrm{P}(\rho r\lambda) * \mathrm{P}\left(\rho^2 r\lambda\right) * ...$$

where $\mathrm{P}(\lambda)$ *stands for* $\mathrm{Poisson}(\lambda)$*, the variates are independent, and* $\rho = s/(s+r)$*.*

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 $\mathrm{Poisson}(r\lambda)$ 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 $\mathrm{Bin}\left(x,\ \rho^n\right)$. The number of first stage customers remaining is $\mathrm{Poisson}(\rho^{n-1} r\lambda)$, and so on. □

As above, these considerations yield the following result.

Theorem 6.3 *For the* x*-chain for the Gibbs sampler location model* (38) *started at* x*,*

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

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

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

proving the desired result. □

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 $\frac{s}{r+s}$,

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

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

Example (Binomial Location Model) Let $r,\ s$ be positive integers and fix $p,\ 0 < p < 1$. Set

$$g(x) = \binom{r}{x} p^x (1-p)^{r-x},\ \pi(\theta) = \binom{s}{\theta} p^\theta (1-p)^{s-\theta}. \tag{39}$$

The marginal density of the x-chain is

$$m(x) = \binom{r+s}{x} p^x (1-p)^{r+s-x},\ 0 \le x \le r+s. \tag{40}$$

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

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

Set

$$X_{n+1} = \theta_{n+1} + \varepsilon_{n+1} \text{ with } \varepsilon \sim \text{Bin}(r,\ p). \tag{41}$$

Theorem 6.4 *For any $r, s \ge 1$, any starting state x, and all p, the x-chain* (41) *satisfies,*

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

Proof We construct a strong stationary time for the process lifted to binary vectors. See [1, 3, 9] 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 θ_{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 $X_0 = x, X_1, X_2, ...$ is the same as (41). 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 $(s+1, s+2, ..., s+r)$. This T is clearly a strong stationary time for the lifted process: If the coordinate process is $Z_n = (Z_n^1, Z_n^2, ..., Z_n^{r+s})$,

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

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

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

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

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

See [1, 3, 9]. □

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 [6].

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

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

Lift to a process on vectors of length $r + s$ with entries in $\{1, 2, ..., k\}$. The same argument with $X_n = (X_n^1, X_n^2, ..., X_n^k)$ for X_n^i the number of coordinates taking value i, leads to exactly the same bound as in Theorem 6.4, uniformly in $p_1, p_2, ..., p_k$ and k. An exact analytic solution using multivariate orthogonal polynomials for the Multinomial is in [22].

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] Berti, P., Consonni, G. and Pratelli, L. (2008). Discussion on the paper: Gibbs sampling , exponential families and orthogonal polynomials. To appear in *Statistical Science*.

[5] 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.

[6] Diaconis, P. (1988). *Group Representations in Probability and Statistics*, Institute of Mathematical Statistics - Monograph Series, Hayward, California.

[7] Diaconis, P. and Ylvisaker, D. (1979). Conjugate priors for exponential families, *Ann. Stat.* **7**, 269–281.

[8] 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^{nd} Valencia Int'l Meeting, North Holland, Amstredam, 133–156.

[9] Diaconis, P. and Fill, J. (1990). Strong Stationary Times via a New Form of Duality, *Ann. Prob.* **16**, 1483–1522.

[10] Diaconis, P. and Fulman, J. (2008). Carries, shuffling and an amazing matrix. Preprint, Department of Statistics, Stanford University.

[11] Diaconis, P. and Saloff-Coste, L. (1993). Comparison Theorems for Reversible Markov Chains, *Ann. Appl. Prob.* **3**, 696–730.

[12] Diaconis, P. and Zabell, S. (1991). Closed Form Summation for Classical Distributions: Variations on a theme of Demoivre, *Statistical Sci.* **61**, 284–302.

[13] Diaconis, P., Khare, K. and Saloff-Coste L. (2008). Gibbs sampling, exponential families and orthogonal polynomials. *Statistical Science* **23**, 151-178.

[14] Dudley, R.M. (1989). *Real Analysis and Probability*, Wadsworth, Belmont, CA.

[15] Dyer, M., Goldberg, L., Jerrum, M. and Martin, R. (2005). Markov chain comparison, *Probability Surveys* **3**, 89–111.

[16] Esch, D. (2003). The skew-t distribution: Properties and computations, Ph.D. Dissertation, Department of Statistics, Harvard University.

[17] Fill, J. and Machida, M. (2001). Stochastic Monotonicity and Realizable Monotonicity, *Annals of Applied Probability* **29**, 938–978.

[18] Givens, C. and Shortt, R. (1984) A class of Wasserstein metrics for probability distributions. *Michigan Math. J.* **31**, 231–240.

[19] Jones, G.L. and Hobert, J.P. (2001). Honest exploration of intractable probability distributions via Markov chain Monte Carlo, *Statist. Sci.* **16**, 312-334.

[20] 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.

[21] Karlin, S. (1968). *Total Positivity*. Stanford University Press, Stanford.

[22] Khare, K. and Zhou, H. (2008). Rates of convergence of some multivariate Markov chains with polynomial eigenfunctions. Preprint, Department of Statistics, Stanford University.

[23] Lehmann, E. and Romano, J. (2005). *Testing Statistical Hypotheses*, Springer, New York.

[24] 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.

[25] Lund, R.B. and Tweedie, R.L. (1996). Geometric Convergence Rates for Stochastically Ordered Markov Chains, *Mathematics of Operations Research*, **20(1)**, 182–194.

[26] Meyn, S.P. and Tweedie, R.L. (1993). *Markov chains and stochastic stability*, Springer-Verlag, London.

[27] Morris, C. (1982). Natural exponential families with quadratic variance functions, *Ann. Statist.* **10**, 65–80.

[28] Morris, C. (1983). Natural exponential families with quadratic variance functions: Statistical Theory, *Ann. Statist.* **11**, 515–589.

[29] Rosenthal. J.S. (1995). Minorization conditions and convergence rates for Markov chain Monte Carlo, *Jour. Amer. Statist. Assoc.* **90**, 558-566.

[30] Rosenthal, J.S. (1996). Analysis of the Gibbs sampler for a model related to James-Stein estimations, *Statist. Comput.* **6**, 269-275.

[31] Rosenthal, J.S. (2002). Quantitative convergence rates of Markov chains: A simple account, *Electronic Communications in Probability* **7**, 123-128.

[32] 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.*

[33] 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.

[34] 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.

[35] Stoyan, D. (1983). *Comparison Methods for Queues and other Stochastic Models*, John Wiley and Sons, New York.

[36] Tierney, L. (1994). Markov chains for exploring posterior distributions (with discussion), *Ann. Statist.* **22**, 1701-1762.

[37] Wilson, D.B. (2004). Mixing times of lozenge tiling and card shuffling Markov chains, *Annals of Applied Probability*, **14(1)**, 274–325.

[38] David Wilson's Website on Perfect Sampling *"http://research.microsoft.com/ % dbwilson/exact/"*.