Quantitative bounds for Markov chain convergence: Wasserstein and total variation distances

NEAL MADRAS[1] and DENIZ SEZER[2]

[1]*Department of Mathematics and Statistics, York University, 4700 Keele Street, Toronto, ON M3J 1P3, Canada. E-mail: madras@mathstat.yorku.ca*
[2]*Department of Mathematics and Statistics, University of Calgary, 2500 University Drive, Calgary, AB T2N 1N4, Canada. E-mail: adsezer@math.ucalgary.ca*

We present a framework for obtaining explicit bounds on the rate of convergence to equilibrium of a Markov chain on a general state space, with respect to both total variation and Wasserstein distances. For Wasserstein bounds, our main tool is Steinsaltz's convergence theorem for locally contractive random dynamical systems. We describe practical methods for finding Steinsaltz's "drift functions" that prove local contractivity. We then use the idea of "one-shot coupling" to derive criteria that give bounds for total variation distances in terms of Wasserstein distances. Our methods are applied to two examples: a two-component Gibbs sampler for the Normal distribution and a random logistic dynamical system.

Keywords: convergence rate; coupling; Gibbs sampler; iterated random functions; local contractivity; logistic map; Markov chain; random dynamical system; total variation distance; Wasserstein distance

1. Introduction

In many theoretical or applied problems involving positive recurrent Markov chains, it is important to estimate the number of iterations until the distribution of the chain is "close" to its equilibrium distribution. Suppose we have a Markov chain with state space χ, initial state x, transition probability kernel P and limiting stationary distribution π. We would like a quantitative bound such as

$$d(P^n(x, \cdot), \pi(\cdot)) \leq g(x, n),$$

where d is a metric on the set of probability measures and $g(x, n)$ is a function that can be computed explicitly. For example, knowledge of such a function g can be valuable to Bayesian statisticians using Markov chain Monte Carlo (MCMC) approximations because it tells them how many MCMC steps will ensure a good approximation to the posterior distribution under consideration. An excellent survey on the theory of general state space Markov chains and MCMC is [19].

An important technical point is the specification of the metric d on the set of probability measures. Two common choices are the *total variation* (TV) *metric* (denoted d_{TV}) and the *Wasserstein metric* (denoted d_{W}); see Section 2 for definitions and basic properties of these two metrics.

There is a rich literature on Markov chain convergence in total variation distance. Many tools have been developed for convergence in TV, involving probabilistic methods (for example, coupling, strong uniform times; see [5,13,19] for reviews), analytic methods (spectral analysis, Fourier analysis, operator theory; see [5,21]) and geometric methods (path bounds, isoperimetry; see [13,21]). Much of the progress, and many of the sharpest results, have been for discrete state spaces [5,13,21], including spaces related to graphs, algebraic structures, or models from statistical physics. Some results extend to general state spaces, but some basic discrete properties and methods do not have convenient analogs in the general case. Continuous state spaces are of particular interest in Bayesian MCMC applications [10,19], but quantitative rigorous results about realistic examples are scarce.

Frequently, the desirable functions g to seek are of the form $g(x, n) = C(x)r^n$, where $C(x)$ and r can be computed explicitly. The existence of such a function for the TV metric is called *geometric ergodicity* and is known to hold under fairly general conditions (see, for example, [16,17]). Explicit identification of such functions can be an intricate task, however. A classical result in this context is due to Doeblin: if there exists a probability measure ν and $0 < \varepsilon < 1$ such that $P(x, dy) \geq \varepsilon \nu(dy)$ for every x, then $d_{\text{TV}}(P^n(x, \cdot), \pi) \leq (1 - \varepsilon)^n$. It is possible to get similar bounds using coupling when Doeblin's condition holds only on a subset K, if a "drift function" to K exists. More precisely, one needs (i) $P(x, dy) \geq \varepsilon \nu(dy)$ for all $x \in K$; (ii) a function $V > 1$ and a constant $\alpha > 1$ such that $E(V(Y_{n+1})|Y_n = y) < V(y)/\alpha$ for all $y \in K^c$. These conditions are called (i) *minorization* and (ii) *drift conditions* [16,20]. For practitioners who want to implement these conditions, the challenge is to identify such a set K and a drift function V that lead to tractable calculations and good results. See [11] for an impressive application of these conditions to a Bayesian random effects model. A good survey and another realistic application is in [14].

Coupling arguments for proving TV bounds typically use two coupled versions of a Markov chain that coalesce relatively quickly. This is often technically easier to do in discrete state spaces than in state spaces with no atoms. Minorization and drift conditions offer one solution to this difficulty: coalescence is facilitated when the coupled chains are simultaneously in the set K. However, in many situations, it may be hard to force coupled chains to coalesce, but it may be easier to force them to come (and stay) very close to each other. Closeness of two chains in the metric of the state space roughly corresponds to closeness of their distributions in the Wasserstein distance. For this reason, the Wasserstein distance can be a tractable alternative to the total variation distance for problems in continuous state spaces (see, for example, [8]). Although Wasserstein convergence can be weaker than TV convergence, we shall show that under certain conditions, bounds on the rate of Wasserstein convergence can be used to get bounds on the rate of TV convergence (see Section 4). Thus, proving Wasserstein convergence is sometimes a step toward proving TV convergence. Huber [12] also uses this general philosophy, employing rather different methods from ours.

A particularly successful framework for studying convergence in Wasserstein distance is random dynamical systems, or iterated function systems [6,22]. An iterated function system is a sequence of random maps of the form $F_n(x) = f_1 \circ f_2 \circ \cdots \circ f_n(x)$ or $\tilde{F}_n(x) = f_n \circ f_{n-1} \circ \cdots \circ f_1(x)$, where $f_1, f_2, \ldots$ are independent and identically distributed (i.i.d.) random maps. (Two examples are described later in this section.) The sequence $\{\tilde{F}_n(x) : n \geq 1\}$ is called the *forward sequence* and is a Markov chain. Many examples of Markov chains can be represented as forward iterates of i.i.d. random maps. $\{F_n(x) : n \geq 1\}$ is called the *backward* sequence and, under

certain conditions, it converges pointwise to a random variable, X_∞, independent of the starting point x. If X_∞ exists, in which case the system is called *attractive*, the distribution of X_∞ is also the stationary distribution π of the Markov chain $\tilde{F}_n(x)$. The rate at which $E[\rho(F_n(x), X_\infty)]$ converges to zero is an upper bound on the rate of convergence in distribution of the Markov chain $\tilde{F}_n(x)$ to π in Wasserstein distance. Indeed, since $F_n(x)$ has distribution $P^n(x, \cdot)$ (as does $\tilde{F}_n(x)$) and since $X_\infty \sim \pi$, we have

$$d_W(P^n(x, \cdot), \pi) \leq E[\rho(F_n(x), X_\infty)]. \tag{1}$$

One condition that guarantees attractivity is strong contractivity, that is, $E[\log \text{Lip} f] < 0$, where $\text{Lip} f$ is the Lipschitz constant of the (random) function f. This condition is a generalization of the stronger condition that there exists a constant $r \in (0, 1)$ such that $\rho(f(x), f(y)) \leq r\rho(x, y)$ for all x and y, with probability 1. (Gibbs [8] used a variation of this condition to get a bound for the Wasserstein distance of a Markov chain X_n to its stationary distribution using coupling. See also [6] for a related result.) However, applications frequently require weaker conditions. Steinsaltz [22] proves attractivity under a more general condition, called "local contractivity", which says that there exists a "drift function" $\phi : \mathcal{X} \mapsto [1, \infty)$ and a constant $r \in (0, 1)$ such that

$$G_n(x) := E[D_x F_n] \leq \phi(x) r^n,$$

where $D_x f := \limsup_{y \to x} \frac{\rho(f(x), f(y))}{\rho(x, y)}$. He proves that if local contractivity holds, then

$$E[\rho(F_n(x), X_\infty)] \leq C_x r^n \qquad \text{for every } n \geq 1,$$

where C_x is a number that can be computed explicitly; see Section 3.1 for further discussion. Steinsaltz's use of the term "drift" is analogous to, but different from, Rosenthal's use (which, in turn, is closely related to Foster–Lyapunov functions; see [7] for a review and references).

Like the minorization and drift conditions, the local contractivity condition requires preliminary work to obtain a drift function. The goal of the first part of this paper (Section 3) is to provide a systematic framework for doing this.

We developed our methods using two examples. The first is a simple Gibbs sampler chain for Bayesian estimation of the mean and variance of a Normal distribution. The second example is a randomized version of the classical logistic map from dynamical systems theory.

The paper is organized as follows. The remainder of this section is devoted to descriptions of our two main examples. Section 2 provides definitions and basic properties of the Wasserstein and total variation metrics. Section 3 examines the task of finding a drift function that produces quantitative bounds on Wasserstein convergence. Section 3.1 reviews the results of Steinsaltz [22] and Section 3.2 presents an approach to finding drift functions by looking for sub-eigenfunctions of a certain dominating operator. Section 3.3 then uses this approach to find drift functions for our Gibbs sampler example. Section 4 shows how bounds on the Wasserstein metric may be "upgraded" to bounds on the total variation metric in some situations. Section 4.1 reviews the idea of "one-shot coupling" [18] and presents our key technical result (Theorem 12). Sections 4.2 and 4.3 apply this result to our two examples.

Example 1 (Normal Gibbs sampler). A simple Bayesian estimation problem is the following. Consider a random sample of size J from the Normal distribution with mean θ and variance σ^2 (written $N(\theta, \sigma^2)$). We assume that θ and $S := \sigma^{-2}$ are themselves independent random variables from Normal and Gamma prior distributions respectively:

$$\theta \sim N(\xi, K^{-1}) \quad \text{and} \quad S := \sigma^{-2} \sim \Gamma(\alpha, \beta).$$

(Here, $\Gamma(\alpha, \beta)$ is the Gamma distribution with density $s^{\alpha-1}\beta^\alpha \exp(-\beta s)/\Gamma(\alpha)$.) Let $Y := Y_1, \ldots, Y_J$ be our random sample from $N(\theta, \sigma^2)$ (conditionally independent, given θ and σ). The joint posterior for θ and S given Y is

$$p(\theta, s|Y) \propto s^{\alpha-1+J/2} \exp\left[-\beta s - \frac{K(\theta - \xi)^2}{2} - \frac{s\sum(Y_j - \theta)^2}{2}\right] \tag{2}$$

(where $\sum$ is the sum over j from 1 to J). Besides positive values of K, we shall also consider the case $K = 0$. When $K = 0$, the prior for θ is not a probability distribution; however, the joint posterior *is* a probability distribution. (We can view $K = 0$ as the "flat prior" limit $K \to 0+$. The case $\beta = 0$ is similar.) The *Gibbs sampler* is the Markov chain (θ_t, S_t) defined recursively by drawing θ_t from its conditional distribution given Y and $S = S_{t-1}$, followed by drawing S_t from its conditional distribution given Y and $\theta = \theta_t$:

$$\theta_t \sim N\left(\frac{S_{t-1}\sum Y_j + K\xi}{S_{t-1}J + K}, \frac{1}{S_{t-1}J + K}\right),$$

$$S_t \sim \Gamma\left(\alpha + \frac{J}{2}, \beta + \frac{1}{2}\sum(Y_j - \theta_t)^2\right).$$

We can represent this procedure as follows:

$$\theta_t = \frac{Z_t}{\sqrt{S_{t-1}J + K}} + \frac{S_{t-1}\sum Y_j + K\xi}{S_{t-1}J + K}, \qquad \text{where } Z_t \sim N(0, 1), \tag{3}$$

$$S_t = \frac{G_t}{\beta + \frac{1}{2}\sum(Y_j - \theta_t)^2}, \qquad \text{where } G_t \sim \Gamma(\alpha + J/2, 1) \tag{4}$$

(and $\{Z_t\}$ and $\{G_t\}$ are independent i.i.d. sequences). Let

$$\bar{Y} = \frac{1}{J}\sum_{j=1}^{J} Y_j \quad \text{and} \quad \Sigma_0 = \beta + \frac{1}{2}\sum_{j=1}^{J}(Y_j - \bar{Y})^2$$

(we treat these as constants, since we always condition on Y). Since

$$\sum_{j=1}^{J}(Y_j - \theta)^2 = \sum_{j=1}^{J}(Y_j - \bar{Y})^2 + J(\bar{Y} - \theta)^2, \tag{5}$$

we can write equation (4) as

$$S_t = \frac{G_t}{\Sigma_0 + (J/2)(\bar{Y} - \theta_t)^2}. \tag{6}$$

Using equation (3), we can express (6) as a random dynamical system, as follows:

$$S_t = f_t(S_{t-1}), \qquad t = 1, 2, \ldots, \tag{7}$$

where $f_t : (0, \infty) \to (0, \infty)$ is the random function

$$f_t(s) = \frac{G_t}{\Sigma_0 + (J/2)(Z_t/\sqrt{sJ + K} + (\xi - \bar{Y})K/(sJ + K))^2} \tag{8}$$

with the random variables G_t and Z_t as above. The case $K = 0$ is of special interest (representing an improper prior for θ) and equation (8) specializes to

$$f_t(s) = \frac{G_t}{\Sigma_0 + Z_t^2/(2s)}. \tag{9}$$

We note that the posterior (2) is a proper probability distribution when $K = 0$, even though the prior is not (to see this, use (5) and integrate θ first).

Without loss of generality, we can assume that ξ is zero and that K is either 0 or 1. (Indeed, if $K > 0$, then we can let $\tilde{\theta} = (\theta - \xi)\sqrt{K}$, $\tilde{Y}_i = (Y_i - \xi)\sqrt{K}$, $\tilde{\sigma}^2 = K\sigma^2$ and $\tilde{\beta} = K\beta$; then $\tilde{Y}_i \sim N(\tilde{\theta}, \tilde{\sigma}^2)$, where $\tilde{\theta} \sim N(0, 1)$ and $\tilde{\sigma}^{-2} \sim \Gamma(\alpha, \tilde{\beta})$.) Accordingly, for our Markov chain $\{S_t\}$ with $K \in \{0, 1\}$, let P_K be the chain's transition probability kernel, let $p_K(\cdot, \cdot)$ be the density of P_K and let π_K be the stationary distribution.

We shall obtain quantitative bounds for the convergence of our Gibbs sampler chain P_K ($K \in \{0, 1\}$); see Propositions 11 and 14, and the discussions of numerical results following each. Roberts and Rosenthal [18] analyzed this chain with flat priors, that is, $K = \xi = \beta = 0$ and $\alpha = 1$. In particular, their results show that $\limsup_{n \to \infty} [d_{TV}(P_0^n(x, \cdot), \pi_0)]^{1/n} \leq 1/J$. This would equal our asymptotic rate if we could replace w by 1 in Proposition 14. The analysis of [18] uses the property that the recursion for $1/S_t$ is a linear function of $1/S_{t-1}$, which only holds when $K = 0$. Their approach cannot handle the case $K > 0$. Our method of Section 4 may be viewed as a more powerful (nonlinear) generalization of [18].

Example 2 (*Random logistic map*). We consider the i.i.d. random maps $f_1, f_2, \ldots$ on $[0, 1]$ defined by

$$f_i(x) = 4B_i x(1 - x),$$

where $B_1, B_2, \ldots$ are i.i.d. random variables having the Beta$(a + \frac{1}{2}, a - \frac{1}{2})$ distribution. Here, $a > \frac{1}{2}$ is a fixed number. It is known that the Beta(a, a) distribution is the unique stationary distribution for this iterated function system [3]. Our result for this example will provide bounds that are more qualitative than quantitative. Asymptotic convergence properties of this example have been studied in the literature. Steinsaltz [22] showed that the system is locally contractive if $a \geq 2$ and hence that the corresponding Markov chain converges to equilibrium exponentially

rapidly in the Wasserstein distance. Using the techniques of Section 4, we shall prove the following theorem.

Theorem 1. *Assume that $a > 1/2$ and let $x \in (0, 1)$. There then exists a constant $\tilde{C}_a$, depending only on a, such that*

$$d_{\mathrm{TV}}(\tilde{F}_n(x), \beta_{a,a}) \leq \tilde{C}_a [d_{\mathrm{W}}(\tilde{F}_{n-1}(x), \beta_{a,a})]^{a/(a+1)} \qquad \textit{for all } n \geq 1$$

(where $\beta_{a,a}$ is a random variable having the Beta(a, a) *distribution).*

Note that Theorem 1 does not assume local contractivity (indeed, local contractivity fails if $1/2 < a < 1$, by Corollary 3 of [23] and Theorem 1 of [22]).

Theorem 1 implies the following. Assume that the random logistic Markov chain $\{\tilde{F}_n(x) : n = 0, 1, \ldots\}$ converges to its equilibrium exponentially rapidly in Wasserstein distance, that is, that there exists a constant $\rho \in (0, 1)$ such that

$$\limsup_{n \to \infty} [d_{\mathrm{W}}(\tilde{F}_n(x), \beta_{a,a})]^{1/n} \leq \rho. \tag{10}$$

It then also converges exponentially rapidly in TV distance, perhaps at a modestly slower rate:

$$\limsup_{n \to \infty} [d_{\mathrm{TV}}(\tilde{F}_n(x), \beta_{a,a})]^{1/n} \leq \rho^{a/(a+1)} < 1.$$

Since the state space $(0, 1)$ has diameter 1, we trivially have $d_{\mathrm{W}}(\tilde{F}_n(x), \beta_{a,a}) \leq d_{\mathrm{TV}}(\tilde{F}_n(x), \beta_{a,a})$. Hence, we conclude that for $a > 1/2$, our random logistic Markov chain converges to the equilibrium exponentially rapidly in Wasserstein distance *if and only if* it converges exponentially rapidly in TV distance.

2. Wasserstein and total variation metrics

In this section, we review the definitions and some properties of two metrics on the space of probability measures: the Wasserstein metric and the total variation (TV) metric. For a broader review of metrics on probabilities, see [9].

Let (χ, ρ) be a complete separable metric space. Consider two probability measures, μ_1 and μ_2, on χ. Let Joint(μ_1, μ_2) denote the set of all probability measures M on $\chi \times \chi$ whose marginal distributions are μ_1 and μ_2, that is,

$$\mu_1(\mathrm{d}x) = \int_y M(\mathrm{d}x, \mathrm{d}y) \quad \text{and} \quad \mu_2(\mathrm{d}y) = \int_x M(\mathrm{d}x, \mathrm{d}y).$$

In other words, if two random variables X_1 and X_2 have distributions μ_1 and μ_2, respectively, then Joint(μ_1, μ_2) is the set of all "couplings" of X_1 and X_2.

The *Wasserstein distance* between μ_1 and μ_2, denoted $d_{\mathrm{W}}(\mu_1, \mu_2)$, is defined to be

$$d_{\mathrm{W}}(\mu_1, \mu_2) = \inf \left\{ \int_\chi \int_\chi \rho(x, y) M(\mathrm{d}x, \mathrm{d}y) : M \in \mathrm{Joint}(\mu_1, \mu_2) \right\}. \tag{11}$$

In other words, $d_W(\mu_1, \mu_2)$ is the infimum of $E(\rho(X_1, X_2))$ over all couplings of X_1 and X_2 (where $X_i \sim \mu_i$). It can be shown that there exists an M that attains the infimum (see, for example, Section 5.1 of [4]).

The *total variation (TV) distance* between μ_1 and μ_2, denoted $d_{TV}(\mu_1, \mu_2)$, is defined to be

$$d_{TV}(\mu_1, \mu_2) = \sup\{|\mu_1(A) - \mu_2(A)| : A \subset \chi\}. \tag{12}$$

This sup is attained by some set A (by the classical Hahn decomposition for the signed measure $\mu_1 - \mu_2$). An equivalent definition of d_{TV} is

$$d_{TV}(\mu_1, \mu_2) = \inf\{M(\{(x, y) : x \neq y\}) : M \in \text{Joint}(\mu_1, \mu_2)\}. \tag{13}$$

In other words, $d_{TV}(\mu_1, \mu_2)$ is the infimum of $\Pr\{X_1 \neq X_2\}$ over all couplings of X_1 and X_2 (where $X_i \sim \mu_i$). For convenience, we shall sometimes talk about the Wasserstein or TV distance between two random variables, which means the same thing as the Wasserstein or TV distance between their distributions.

The following is relatively well known (see, for example, Theorem 5.7 of [4] or Proposition 3 of [19]).

Proposition 2. *Assume that μ_1 and μ_2 are probability measures on χ, having density functions p_1 and ρ_2, respectively, with respect to a common reference measure λ. Then*

$$d_{TV}(\mu_1, \mu_2) = \frac{1}{2} \int_\chi |p_1(z) - p_2(z)| \lambda(dz) \tag{14}$$

$$= \int_{z : p_1(z) > p_2(z)} (p_1(z) - p_2(z)) \lambda(dz) \tag{15}$$

$$= 1 - \int_\chi \min\{p_1(z), p_2(z)\} \lambda(dz). \tag{16}$$

If the state space χ is bounded, then $d_W(\mu_1, \mu_2) \leq d_{TV}(\mu_1, \mu_2) \times [\sup\{\rho(x, y) : x, y \in \chi\}]$ and, in particular, TV convergence implies Wasserstein convergence. However, in general, neither convergence implies the other. For example, in $\mathbb{R}$, let μ_n be the two-point probability distribution that has $\mu_n(\{0\}) = 1 - n^{-1}$ and $\mu_n(\{n\}) = n^{-1}$. Then μ_n converges to the point mass at 0 in the TV metric, but not in Wasserstein. Also, let ν_n be the probability distribution on $[0, 1]$ with density $1 + \sin(2\pi n x)$; then ν_n converges to the uniform distribution on $[0, 1]$ in Wasserstein, but not in TV.

The following result will be very useful in Section 4.

Lemma 3. *Consider a deterministic measurable function $g : A \times B \to C$. Let W_1 and W_2 be two B-valued random variables and let U be an A-valued random variable that is independent of both W_i's. Define the C-valued random variables X_1 and X_2 by $X_i = g(U, W_i)$, $i = 1, 2$. Then*

$$d_{TV}(X_1, X_2) \leq d_{TV}(W_1, W_2).$$

Proof. Choose a joint distribution $M(\mathrm{d}w_1, \mathrm{d}w_2)$ of a random vector $(\tilde{W}_1, \tilde{W}_2)$ on $B \times B$ such that $\tilde{W}_i \stackrel{d}{=} W_i$ for $i = 1, 2$ and $M\{\tilde{W}_1 \neq \tilde{W}_2\} = d_{\mathrm{TV}}(W_1, W_2)$. Also, make $(\tilde{W}_1, \tilde{W}_2)$ independent of U and let $\tilde{X}_i = g(U, \tilde{W}_i)$. Then $\tilde{X}_i \stackrel{d}{=} X_i$ for $i = 1, 2$, so

$$d_{\mathrm{TV}}(X_1, X_2) \leq M\{\tilde{X}_1 \neq \tilde{X}_2\} \leq M\{\tilde{W}_1 \neq \tilde{W}_2\} = d_{\mathrm{TV}}(W_1, W_2). \qquad \square$$

3. Convergence in the Wasserstein metric

3.1. Local contractivity condition and a convergence theorem

Our main tool to obtain quantitative bounds for convergence in Wasserstein metric will be Steinsaltz's local contractivity convergence theorem [22]. Below, we review this result in a form convenient for us.

Definition 4. *An iterated function system is locally contractive if there exists a function $\phi : \mathcal{X} \mapsto [1, \infty)$ and $r \in (0, 1)$ such that*

$$G_n(x) := E[D_x F_n] \leq \phi(x) r^n \qquad \text{for all } n \geq 1,$$

where $D_x f := \limsup_{y \to x} \frac{\rho(f(x), f(y))}{\rho(x, y)}$. If this holds, then ϕ is called a drift function.

Theorem 5. *If an iterated function system is locally contractive with a drift function ϕ and if*

$$C_x := E\left[\rho(f(x), x) \sup_{0 \leq t \leq 1} \{\phi(x + t(f(x) - x))\}\right] < \infty,$$

then the system is attractive (in particular, $F_\infty(x)$ is independent of x) and

$$d_{\mathrm{W}}(F_n(x), F_\infty(x)) \leq E\rho(F_n(x), F_\infty(x)) \leq \frac{C_x r^n}{1 - r} \qquad \text{for every } x \in \chi.$$

Steinsaltz [22] also gives a sufficient condition, called the *growth condition*, for a function ϕ to be a drift function: a continuous function $\phi : \mathcal{X} \mapsto [1, \infty)$ is a drift function if $r < 1$, where

$$r := \sup_x E\left[\frac{\phi(f(x))}{\phi(x)} D_x f\right].$$

Here is a short argument (different from the original proof in [22]) to explain why. Let $\mathcal{L}$ be the positive linear operator which maps a generic function g to the function $\mathcal{L}(g)(x) = E[g(f(x))D_x f]$. Then $G_n(x) = \mathcal{L}^n(1)(x)$, with 1 here being the constant function equal to 1. Note that the growth condition is equivalent to $\mathcal{L}\phi \leq r\phi$. We will refer to any $\phi > 0$ satisfying $\mathcal{L}\phi \leq r\phi$ as an *r-sub-eigenfunction* for $\mathcal{L}$. Now, if $\phi \geq 1$ and ϕ is an r-sub-eigenfunction, then $G_n(x) = \mathcal{L}^n 1 \leq \mathcal{L}^n \phi \leq r^n \phi$ and hence ϕ is a drift function with rate r.

We note that Proposition 8 of [23] shows that the existence of a ϕ satisfying the growth condition is also necessary for local contractivity.

3.2. How to apply the local contractivity convergence theorem: Finding a drift function

Applying Steinsaltz's local contractivity convergence theorem to a specific problem would be easy if one knew how to write down a drift function. Here, we will propose two practical strategies that can help us to do this.

The first strategy is to find a linear operator $\tilde{\mathcal{L}}$ that dominates $\mathcal{L}$ and is simpler to manage. If ϕ is an r-sub-eigenfunction for $\tilde{\mathcal{L}}$, then it is an r-sub-eigenfunction for $\mathcal{L}$ as well.

One kind of operator that we can manage is defined as follows: let $\{A_i\}_{i=1}^n$ be a finite partition of the state space χ and let

$$\tilde{\mathcal{L}}\phi(x) = b(x) \sum_{i=1}^n 1_{A_i}(x) \int_\chi \phi(s)\mu_i(\mathrm{d}s), \tag{17}$$

where $b(x)$ is a positive function and each μ_i is a non-zero finite measure on χ.

Theorem 6. *Let $\tilde{\mathcal{L}}$ be an operator of the form* (17). *In order for $\tilde{\mathcal{L}}$ to have an r-sub-eigenfunction, it is necessary and sufficient that the matrix*

$$Q(i, j) = \int_{A_j} b(x)\mu_i(\mathrm{d}x)$$

has an r-sub-eigenvector $p = (p_1, p_2, \ldots, p_n)^t$, that is, $p_i > 0$ $\forall i$ and $Qp \le rp$. Moreover, if p is an r-sub-eigenvector for Q, then the function

$$\phi(x) = \sum_{j=1}^n p_j 1_{A_j}(x)b(x) \tag{18}$$

(and any positive multiple of it) is an r-sub-eigenfunction for $\tilde{\mathcal{L}}$.

Proof. If ϕ is an r-sub-eigenfunction of $\tilde{\mathcal{L}}$, then $b(x) \sum_{j=1}^n 1_{A_j}(x) \int \phi(\mathrm{d}c)\mu_j(c) \le r\phi(x)$, by definition of $\tilde{\mathcal{L}}$. Integrating both sides with respect to μ_i gives

$$\sum_{j=1}^n \int_{A_j} b(x)\mu_i(\mathrm{d}x) \int \phi(c)\mu_j(\mathrm{d}c) \le r \int \phi(x)\mu_i(\mathrm{d}x).$$

Therefore, the vector p defined by $p_i := \int \phi\mu_i$ is an r-sub-eigenvector for Q. Conversely, if p is an r-sub-eigenvector for Q and if ϕ is as defined in (18), then

$$\tilde{\mathcal{L}}\phi(x) = b(x) \sum_{i=1}^n 1_{A_i}(x) \sum_{j=1}^n p_j \int_{A_j} b(s)\mu_i(\mathrm{d}s) \le b(x) \sum_{i=1}^n 1_{A_i}(x)rp_i = r\phi(x).$$

Hence ϕ is an r-sub-eigenfunction and so is any positive multiple of it. □

For the case $n = 1$, Theorem 6 implies the following.

Corollary 7. *Assume that b is a positive function and μ is a finite measure such that $\mathcal{L}\phi(x) \leq b(x) \int_\chi \phi(s)\mu(ds)$ for every $x \in \chi$ and every positive ϕ. Let $r = \int b(s)\mu(ds)$. Then b is an r-sub-eigenfunction for $\mathcal{L}$.*

Note that for an r-sub-eigenfunction ϕ to be a drift function, it must be greater than 1. If ϕ is bounded away from 0, we can get a drift function simply by scaling ϕ. However, if ϕ is not bounded away from 0, we first need to truncate it, as in the following lemma.

Lemma 8. *Let ϕ be an r-sub-eigenfunction for $\mathcal{L}$. Let $\varepsilon > 0$ and define*

$$\phi_\varepsilon(x) = \frac{1}{\varepsilon}\max\{\phi(x), \varepsilon\}. \tag{19}$$

Define $A_0 := \sup_x E[\frac{D_x f}{\phi(x)}]$ and $r_\varepsilon := r + \varepsilon A_0$, and assume that $A_0 < \infty$. Then ϕ_ε is an r_ε-sub-eigenfunction for $\mathcal{L}$.

Proof. Since $\phi_\varepsilon(x) \geq 1$ for every x and $\frac{\phi_\varepsilon(f(x))}{\phi_\varepsilon(x)} \leq \frac{\phi(f(x))+\varepsilon}{\phi(x)}$, we have

$$E\left[\frac{\phi_\varepsilon(f(x))}{\phi_\varepsilon(x)}D_x f\right] \leq E\left[\frac{\phi(f(x))}{\phi(x)}D_x f\right] + \varepsilon E\left[\frac{D_x f}{\phi(x)}\right] \leq r + \varepsilon A_0. \tag{20}$$

$\square$

The second strategy is to switch to an easier operator, analogously to switching from one measure to another by the use of a Radon–Nikodym derivative.

Lemma 9. *Assume that a positive linear operator $\mathcal{L}_1$ has the integral representation $\mathcal{L}_1(\phi)(x) = \int \phi(y)K(x, dy)$ and let $\mathcal{L}_2(\phi)(x) = \frac{1}{h(x)}\int \phi(y)h(y)K(x, dy)$, where h is a strictly positive function. Then ϕ is an r-sub-eigenfunction for $\mathcal{L}_1$ if and only if $\frac{\phi}{h}$ is an r-sub-eigenfunction for $\mathcal{L}_2$.*

Proof. It is enough to prove one direction only. Let ϕ be an r-sub-eigenfunction for $\mathcal{L}_1$. Then

$$\mathcal{L}_2\left(\frac{\phi}{h}\right)(x) = \frac{1}{h(x)}\int \frac{\phi(y)}{h(y)}h(y)K(x, dy) \leq r\frac{\phi(x)}{h(x)}.$$

$\square$

In particular, this lemma tells us that if $r := \sup_x K(x, \chi) < 1$, then $1/h$ is an r-sub-eigenfunction for $\mathcal{L}_2$.

3.3. Example 1: Normal Gibbs sampler

We shall use the techniques of Section 3.2 to find drift functions for the Gibbs sampler example of Section 1. Recall that, without loss of generality, we assume that $K = 0$ or 1 and $\xi = 0$. The

following proposition gives three different drift functions that are valid under different conditions on the parameters and the data Y. It should be clear that other drift functions are possible; also, the bounds r_i can be tightened somewhat at the cost of additional effort and/or more complicated expressions. For numerical illustrations, see the remarks following the proof of Proposition 11.

Proposition 10. (i) *For given $K \geq 0$, let*

$$A := \frac{(\alpha + J/2)(|\bar{Y}|/\sqrt{K} + 1)(|\bar{Y}|/\sqrt{K} + 1/2)}{\Sigma_0^2} \quad and \quad r_1 := \frac{(|\bar{Y}|/\sqrt{K} + 1)(|\bar{Y}|/\sqrt{K} + 1/2)}{\alpha + J/2 - 1}.$$

If $r_1 < 1$, then for any ε such that $r_{1,\varepsilon} := r_1 + \varepsilon A < 1$, $\phi_{1,\varepsilon}(x) := \frac{1}{\varepsilon}\max(\varepsilon, \frac{1}{x^2})$ is a drift function with rate $r_{1,\varepsilon}$.

(ii) *Assume $K = 1$. Let $r_2 := (\alpha + \frac{J}{2})\frac{J^2}{\Sigma_0^2}(|\bar{Y}| + 1)(|\bar{Y}| + \frac{1}{2})$. If $r_2 < 1$, then $\phi_2(x) = 1$ is a drift function with rate r_2.*

(iii) *Assume $K = 1$. Define*

$$\hat{A} := \frac{(|\bar{Y}| + 1)(\alpha + J/2)J\sqrt{2\pi}}{2\Sigma_0^2}, \qquad b(x) := \frac{J}{\sqrt{2\pi}}\left(\frac{2|\bar{Y}|}{(xJ + 1)^{3/2}} + \frac{1}{xJ + 1}\right) \quad (21)$$

and

$$r_3 := \frac{1}{\sqrt{2\pi}}\left(4|\bar{Y}|\left(1 - \frac{1}{\sqrt{J(\alpha + J/2)/\Sigma_0 + 1}}\right) + \log\left(\frac{J(\alpha + J/2)}{\Sigma_0} + 1\right)\right).$$

If $r_3 < 1$, then for any ε such that $r_{3,\varepsilon} := r_3 + \varepsilon\hat{A} < 1$, the function $\phi_{3,\varepsilon}(x) = \frac{1}{\varepsilon}\max(\varepsilon, b(x))$ is a drift function with rate $r_{3,\varepsilon}$.

Proof. The idea of the proof is that for each case, we find a sub-eigenfunction ϕ for the operator $\mathcal{L}$ and, if necessary, we truncate ϕ, as in Lemma 8, to obtain a drift function.

Recall $\mathcal{L}(\phi)(x) = E[\phi(f(x))D_x f]$, where

$$f(x) = \frac{G}{\Sigma_0 + (J/2)(\bar{Y}K/(xJ + K) - Z/\sqrt{xJ + K})^2},$$

and G and Z are two independent random variables with $\Gamma(\alpha + J/2, 1)$ and $N(0, 1)$ distributions, respectively. We shall frequently use (without reference) the following two easy calculations for G and Z. First, the definition of the Gamma distribution implies that

$$E(G^p) = \frac{\Gamma(\alpha + J/2 + p)}{\Gamma(\alpha + J/2)} \qquad \text{for } p > -\left(\alpha + \frac{J}{2}\right). \quad (22)$$

Second, for all constants a, b, c, d, the Schwarz inequality and $E(Z^2) = 1$ imply

$$E(|a + bZ||c + dZ|) \leq \sqrt{a^2 + b^2}\sqrt{c^2 + d^2} \leq (|a| + |b|)(|c| + |d|). \quad (23)$$

(i) The local Lipschitz constant $D_x f$ is equal to the absolute value of the derivative f at x, so, by direct computation,

$$D_x f = \frac{GJ^2|\bar{Y}K/(xJ+K) - Z/\sqrt{xJ+K}||\bar{Y}K/(xJ+K)^2 - Z/(2(xJ+K)^{3/2})|}{(\Sigma_0 + (J/2)(\bar{Y}K/(xJ+K) - Z/\sqrt{xJ+K})^2)^2}. \quad (24)$$

Let k_x be the joint distribution of $f(x)$ and $\tilde{D}_x$, where

$$\tilde{D}_x := \frac{J^2|\bar{Y}K/(xJ+K) - Z/\sqrt{xJ+K}||\bar{Y}K/(xJ+K)^2 - Z/(2(xJ+K)^{3/2})|}{G}$$

and let $K_x(dc) = x^2(\int_{0<y<\infty} y k_x(dc, dy))$. Note that $f(x)^2 \tilde{D}_x = D_x f$. Therefore,

$$\mathcal{L}(\phi)(x) = E[\phi(f(x))D_x f] = E[\phi(f(x))f(x)^2 \tilde{D}_x] = \frac{1}{h(x)} \int \phi(c) h(c) K_x(dc),$$

where $h(c) = c^2$. Let $\mathcal{L}_1$ be the operator defined by $\mathcal{L}_1\phi(x) := \int \phi(c) K_x(dc)$ and let $\mathcal{L}_2 = \mathcal{L}$. By Lemma 9, we see that if ϕ is an r-sub-eigenfunction for $\mathcal{L}_1$ then $\frac{\phi}{h}$ is an r-sub-eigenfunction for $\mathcal{L}_2 = \mathcal{L}$. We find that

$$\sup_x \int_0^\infty K_x(dc) = \sup_x x^2 E[\tilde{D}_x]$$
$$\leq \sup_x \frac{x^2 J^2}{(xJ+K)^2} \frac{(|\bar{Y}|/\sqrt{K} + 1)(|\bar{Y}|/\sqrt{K} + 1/2)}{\alpha + J/2 - 1}$$
$$= r_1.$$

If $r_1 < 1$, then $\phi(x) = 1$ is an r_1-sub-eigenfunction for $\mathcal{L}_1$ and hence $\phi_1(x) = x^{-2}$ is an r_1-sub-eigenfunction for $\mathcal{L}$. Finally, note that for every $x > 0$,

$$E\left[\frac{D_x f}{\phi_2(x)}\right] = E\left[\frac{(xJ)^2}{(xJ+K)^2} \frac{G|\bar{Y}K/\sqrt{xJ+K} - Z||\bar{Y}K/\sqrt{xJ+K} - Z/2|}{(\Sigma_0 + (J/2)(\bar{Y}K/(xJ+K) - Z/\sqrt{xJ+K})^2)^2}\right] \leq A.$$

Hence, by Lemma 8, $\phi_{1,\varepsilon}$ is a drift function with growth rate less than $r_1 + \varepsilon A$.

(ii) When $K = 1$, $\sup_x E(D_x f) \leq r_2$. If $r_2 < 1$ and we let $\phi_2(x) = 1 \ \forall x$, then $\mathcal{L}\phi_2(x) = E(D_x f) \leq r_2 \phi_2(x)$ and thus $\phi_2(x)$ is a drift function with rate r_2.

(iii) We first derive a more explicit formula for $\mathcal{L}$ and then look for an operator $\tilde{\mathcal{L}}$ of the form (17) with $n = 1$ that dominates $\mathcal{L}$ (as in Corollary 7). Note that we can write

$$\mathcal{L}(\phi)(x) = \int_0^\infty \phi(c) \left(\int_{-\infty}^\infty \Delta_x(z, c) h_{Z, f(x)}(z, c) \, dz\right) dc,$$

where $h_{Z, f(x)}$ is the joint density of $(Z, f(x))$ and

$$\Delta_x(z, c) = \frac{cJ^2|\bar{Y}/(xJ+1) - z/\sqrt{xJ+1}||\bar{Y}/(xJ+1)^2 - z/(2(xJ+1)^{3/2})|}{\Sigma_0 + (J/2)(\bar{Y}/(xJ+1) - z/\sqrt{xJ+1})^2}$$

(observe that $\Delta_x(Z, f(x)) = D_x f$, by (24)). To simplify the formulae, let us put

$$A_x(z) = \frac{\bar{Y}}{(xJ+1)} - \frac{z}{\sqrt{xJ+1}}, \qquad B_x(z) = \left| \frac{\bar{Y}}{(xJ+1)^2} - \frac{z}{2(xJ+1)^{3/2}} \right|$$

and $u_x(z) = \Sigma_0 + \frac{J}{2} A_x(z)^2$.

To find $h_{Z, f(x)}$, we consider the mapping $T_x(z, g) = (z, g/u_x(z))$. Note that $T_x(Z, G) = (Z, f(x))$. $T_x(z, c)$ is one-to-one and $T_x^{-1}(z, c) = (z, c(u_x(z)))$. Let D be the Jacobian of T_x^{-1}. We have $h_{Z, f(x)}(z, c) = h_{Z, G}(T_x^{-1}(z, c)) |\det D|$ and $|\det D| = u_x(z)$; therefore,

$$h_{Z, f(x)}(z, c) = \frac{1}{\Gamma(\alpha + J/2)\sqrt{2\pi}} u_x(z) e^{-z^2/2} (c u_x(z))^{\alpha+J/2-1} e^{-c u_x(z)}.$$

Now,

$$\int_{-\infty}^{\infty} \Delta_x(z, c) h_{Z, f(x)}(z, c) \, dz$$

$$= \frac{cJ^2}{\Gamma(\alpha + J/2)\sqrt{2\pi}} \left(\int_{z \leq \bar{Y}/\sqrt{xJ+1}} e^{-z^2/2} A_x(z) B_x(z) (c u_x(z))^{\alpha+J/2-1} e^{-c u_x(z)} \, dz \right.$$

$$\left. - \int_{z > \bar{Y}/\sqrt{xJ+1}} e^{-z^2/2} A_x(z) B_x(z) (c u_x(z))^{\alpha+J/2-1} e^{-c u_x(z)} \, dz \right).$$

Substituting $u = c u_x(z)$ and noting that $du = -cJ \frac{1}{\sqrt{xJ+1}} A_x(z) \, dz$, we get

$$\int_{-\infty}^{\infty} \Delta_x(z, c) h_{Z, f(x)}(z, c) \, dz$$

$$= \int_{u \geq c\Sigma_0} \frac{J}{\Gamma(\alpha + J/2) 2\sqrt{2\pi}\sqrt{xJ+1}} u^{\alpha+J/2-1} e^{-u}$$

$$\times \left[e^{-(1/2)(xJ+1)((\bar{Y}/(xJ+1)) + \sqrt{(2/J)(u/c-\Sigma_0)})^2} \left| \frac{\bar{Y}}{xJ+1} - \sqrt{\frac{2}{J}\left(\frac{u}{c} - \Sigma_0\right)} \right| \right.$$

$$\left. + e^{-(1/2)(xJ+1)((\bar{Y}/(xJ+1)) - \sqrt{(2/J)(u/c-\Sigma_0)})^2} \left| \frac{\bar{Y}}{xJ+1} + \sqrt{\frac{2}{J}\left(\frac{u}{c} - \Sigma_0\right)} \right| \right] du.$$

Using the inequality $|t e^{-C(A+t)^2}| \leq |A| + \frac{1}{\sqrt{2C}}$ (where A and t are real and $C > 0$), we bound the term inside the brackets by $2(\frac{2|\bar{Y}|}{xJ+1} + \frac{1}{\sqrt{xJ+1}})$. Hence, $\mathcal{L}(\phi)(x) \leq b(x) \int_0^{\infty} \phi(c) \bar{H}(c\Sigma_0) \, dc$, where $b(x)$ is defined in (21) and $\bar{H}$ is one minus the c.d.f. of our gamma variable G, that is, $\bar{H}(x) = \Pr\{G > x\}$.

Next, we compute $r = \int_0^\infty b(c)\bar{H}(c\Sigma_0)\,dc$. Let g be the density of G. Note

$$\int_0^\infty \frac{1}{(cJ+1)^{3/2}}\bar{H}(c\Sigma_0)\,dc = \frac{2}{J}\int_0^\infty \left(1 - \frac{1}{\sqrt{xJ/\Sigma_0+1}}\right)g(x)\,dx$$

$$\leq \frac{2}{J}\left(1 - \frac{1}{\sqrt{\int_0^\infty (xJ/\Sigma_0+1)g(x)\,dx}}\right) \tag{25}$$

$$= \frac{2}{J}\left(1 - \frac{1}{\sqrt{J(\alpha+J/2)/\Sigma_0+1}}\right) \tag{26}$$

and

$$\int_0^\infty \frac{1}{cJ+1}\bar{H}(c\Sigma_0)\,dc = \frac{1}{J}\int_0^\infty \log\left(\frac{xJ}{\Sigma_0}+1\right)g(x)\,dx$$

$$\leq \frac{1}{J}\log\left(\int_0^\infty \left(\frac{xJ}{\Sigma_0}+1\right)g(x)\,dx\right) \tag{27}$$

$$= \frac{1}{J}\log\left(\frac{J(\alpha+J/2)}{\Sigma_0}+1\right), \tag{28}$$

where (25) and (27) follow from Jensen's inequality. Therefore, $r \leq r_3$. We conclude that ϕ_3 is an r_3-sub-eigenfunction.

Using (23), we have

$$E(D_x f) \leq \frac{(\alpha+J/2)J^2}{\Sigma_0^2}E\left(\left|\frac{\bar{Y}}{xJ+1} - \frac{Z}{\sqrt{xJ+1}}\right|\left|\frac{\bar{Y}}{(xJ+1)^2} - \frac{Z}{2(xJ+1)^{3/2}}\right|\right)$$

$$\leq \frac{(\alpha+J/2)J^2}{\Sigma_0^2(xJ+1)}\left(\frac{|\bar{Y}|}{\sqrt{xJ+1}}+1\right)\left(\frac{|\bar{Y}|}{(xJ+1)^{3/2}}+\frac{1}{2(xJ+1)}\right)$$

$$= \frac{(\alpha+J/2)J\sqrt{2\pi}}{2\Sigma_0^2(xJ+1)}\left(\frac{|\bar{Y}|}{\sqrt{xJ+1}}+1\right)b(x),$$

where b is defined in equation (21). Hence, $\sup_x E[D_x f/b(x)] \leq \hat{A}$. By Corollary 7 and Lemma 8, the function $\phi_{3,\varepsilon}$ is a drift function with growth rate less than $r_{3,\varepsilon}$. $\qquad\square$

Proposition 11. *Define r_i and $r_{i,\varepsilon}$ as in Proposition 10:*

(i) *Let $K \geq 0$ and assume that $\alpha + J/2 > 2$. If $r_{1,\varepsilon} < 1$, then for all $x > 0$ and all $n \geq 1$,*

$$d_W(P_K^n(x,\cdot),\pi_K) \leq \frac{\hat{C}_{1,\varepsilon,x}}{1-r_{1,\varepsilon}}r_{1,\varepsilon}^n,$$

where

$$\hat{C}_{1,\varepsilon,x} = \left(x + \frac{\alpha + J/2}{\Sigma_0}\right)\left(\max\left\{\frac{1}{\varepsilon x^2}, 1\right\}\right.$$

$$+ \left(\Sigma_0^2 + \frac{J\Sigma_0}{xJ + K}\left[\frac{(\bar{Y}K)^2}{xJ + K} + 1\right]\right.$$

$$+ \frac{J^2}{4(xJ + K)^2}\left[\frac{(\bar{Y}K)^4}{(xJ + K)^2} + \frac{6(\bar{Y}K)^2}{xJ + K} + 3\right]\right)$$

$$\left.\times \left(\varepsilon(\alpha + J/2 - 1)(\alpha + J/2 - 2)\right)^{-1}\right).$$

(ii) *Assume $K = 1$. If $r_2 < 1$, then for all $x > 0$ and all $n \geq 1$,*

$$d_{\mathrm{W}}(P_1^n(x, \cdot), \pi_1) \leq \frac{x + (\alpha + J/2)/\Sigma_0}{1 - r_2} r_2^n.$$

(iii) *Assume $K = 1$. If $r_{3,\varepsilon} < 1$, then for all $x > 0$ and all $n \geq 1$,*

$$d_{\mathrm{W}}(P_1^n(x, \cdot), \pi_1) \leq \frac{\hat{C}_{3,\varepsilon,x}}{1 - r_{3,\varepsilon}} r_{3,\varepsilon}^n,$$

where

$$\hat{C}_{3,\varepsilon,x} := \max\left\{1, \frac{J(2|\bar{Y}| + 1)}{\varepsilon\sqrt{2\pi}}\right\}\left(x + \frac{\alpha + J/2}{\Sigma_0}\right).$$

Proof. (i) If $r_{1,\varepsilon} < 1$, then $d_{\mathrm{W}}(P_K^n(x, \cdot), \pi_1) \leq \frac{C_{1,\varepsilon,x}}{1-r_{1,\varepsilon}} r_{1,\varepsilon}^n$, where

$$C_{1,\varepsilon,x} = E\left[|f(x) - x| \sup_{t \in [0,1]} \left\{\phi_{1,\varepsilon}\left(x + t\left(f(x) - x\right)\right)\right\}\right]$$

$$\leq E\left[(f(x) + x) \max\left\{\frac{1}{\varepsilon x^2}, \frac{1}{\varepsilon f(x)^2}, 1\right\}\right]$$

$$\leq E\left[(f(x) + x)\left(\max\left\{\frac{1}{\varepsilon x^2}, 1\right\} + \frac{1}{\varepsilon f(x)^2}\right)\right]$$

$$\leq E[f(x) + x]E\left[\max\left\{\frac{1}{\varepsilon x^2}, 1\right\} + \frac{1}{\varepsilon f(x)^2}\right],$$

the last line following from the FKG inequality (see, for example, Theorem 3.17 of [15]) since $1/f(x)^2$ is a decreasing function of the random variable $f(x)$. From

$$x + E(f(x)) \leq x + \frac{\alpha + J/2}{\Sigma_0} \qquad (29)$$

and (using equation (22) with $p = -2 > -(\alpha + J/2)$)

$$
\begin{aligned}
E(f(x)^{-2}) &= E\left[\left(\Sigma_0 + \frac{J}{2}\left(\frac{\bar{Y}K}{xJ+K} - \frac{Z}{\sqrt{xJ+K}}\right)^2\right)^2\right] E(G^{-2}) \\
&= \left(\Sigma_0^2 + J\Sigma_0 E\left[\left(\frac{\bar{Y}K}{xJ+K} - \frac{Z}{\sqrt{xJ+K}}\right)^2\right]\right. \\
&\quad + \frac{J^2}{4}E\left[\left(\frac{\bar{Y}K}{xJ+K} - \frac{Z}{\sqrt{xJ+K}}\right)^4\right]\right) \\
&\quad \Big/ \big((\alpha + J/2 - 1)(\alpha + J/2 - 2)\big),
\end{aligned}
$$

and calculation of the expectations in the brackets in the above expression, we find that $\hat{C}_{1,\varepsilon,x}$ is an upper bound for $C_{1,\varepsilon,x}$.

(ii) If $r_2 < 1$, then $\phi(x) = 1$ is a drift function with rate r_2. Hence, Theorem 5 implies that $d_W(P_1^n(x,\cdot), \pi_1) \leq \frac{C_{2,x}}{1-r_2}r_2^n$, and $C_{2,x} = E(|f(x) - x|) \leq x + \frac{\alpha + J/2}{\Sigma_0}$ by equation (29).

(iii) If $r_{3,\varepsilon} < 1$, then $d_W(P_K^n(x,\cdot), \pi_1) \leq \frac{C_{3,\varepsilon,x}}{1-r_{1,\varepsilon}}r_{1,\varepsilon}^n$ and $C_{3,\varepsilon,x} \leq E[f(x) + x]\sup_y(\phi_{3,\varepsilon}(y)) \leq \hat{C}_{3,\varepsilon,x}$ because of (29) and the fact that $\sup_y(\phi_{3,\varepsilon}(y)) = \max\{1, \frac{J(2|\bar{Y}|+1)}{\varepsilon\sqrt{2\pi}}\}$. $\qquad\square$

Remarks. (1) The criterion $r_2 < 1$ is essentially the condition that $\log\sup_x E(D_x f) < 0$. This is similar to the strong contractivity condition which says that $E(\log\sup_x D_x f) < 0$. Logically, neither condition implies the other. Each implies the weaker condition $\sup_{x,y} E(\log[\rho(f(x), f(y))/\rho(x,y)]) < 0$ used in [1] to prove attractivity (in a more restrictive setting).

(2) In the Bayesian model, as the number of observations J increases, $\bar{Y}$ and Σ_0/J both converge (to θ and σ^2, respectively). Therefore, for large J, we expect r_1 to be small, but r_2 and r_3 to be large.

(3) $(K = 1)$ To illustrate the calculations in the preceding propositions, we considered some cases with $5 \leq J \leq 10, \alpha = 1, 0.5 \leq \bar{Y} \leq 1.5$ and $5 \leq \Sigma_0 \leq 60$. As shown in Table 1, it is possible for any one of r_1, r_2 or r_3 to be less than the other two.

(a) In case A, we have $r_2 = 5/6$ and $\hat{C}_{2,x} = x + 0.1$. Hence, for $x = 1$, we have

$$d_W(P_1^n(1,\cdot), \pi_1) \leq 6.6 * (5/6)^n \qquad \text{for } n \geq 1 \text{ in case A.}$$

Table 1. Values of r_1, r_2 and r_3 in three cases of the Normal Gibbs sampler with $K = 1$. Observe that r_2 is best in case A, r_1 in case B and r_3 in case C. Numbers with "..." have had trailing digits truncated; other numbers are exact

Case	J	α	$\bar{Y}$	Σ_0	r_1	r_2	r_3
A	10	1	1.5	60	1	5/6	0.97...
B	5	1	0.5	5	0.6	5.25	1.02...
C	5	1	1	12	1.2	1.82...	0.9368...

Table 2. Values of expressions from Propositions 10(i) and 11(i) for the Normal Gibbs sampler with $K = 0$, for the cases given in Table 1. The values of ε were chosen somewhat arbitrarily. We use $x = 1$ in all cases. Numbers with "..." have had trailing digits truncated; other numbers are exact

Case	r_1	A	ε	$r_{1,\varepsilon}$	$\hat{C}_{1,\varepsilon,x}$	$d_{\mathrm{W}}(P_0^n(1,\cdot),\pi_0) \leq$
A	0.1	1/1200	1	0.1008...	202.4...	$226 * (0.101)^n$
B	0.2	0.07	0.5	0.235	31.28...	$40.9 * (0.235)^n$
C	0.2	0.012...	1	0.212...	55.28...	$70.3 * (0.213)^n$

In particular, $d_{\mathrm{W}}(P_1^n(1,\cdot),\pi_1) < 0.01$ for $n \geq 36$ in case A.

(b) For case B, we have $r_1 = 0.6$ and $A = 0.21$. We want to have $r_{1,\varepsilon} < 1$, where $r_{1,\varepsilon} = 0.6 + 0.21\varepsilon$. Suppose we choose $\varepsilon = 0.5$. Then $r_{1,\varepsilon} = 0.705$ and $\hat{C}_{1,\varepsilon,x} < (16 + \max\{1, 2x^{-2}\})(x + 0.7)$ for all $x > 0$. For $x = 1$, we obtain

$$d_{\mathrm{W}}(P_1^n(1,\cdot),\pi_1) \leq 104 * 0.705^n \qquad \text{for } n \geq 1 \text{ in case B.}$$

In particular, $d_{\mathrm{W}}(P_1^n(1,\cdot),\pi_1) < 0.01$ for $n \geq 27$ in case B.

(c) In case C, we have $r_3 < 0.9369$ and $\hat{A} < 0.305$. Choosing $\varepsilon = 0.01$ gives $r_{3,\varepsilon} < 0.94$ and $\hat{C}_{3,\varepsilon,x} < 599(x + 0.3)$. For $x = 1$, we obtain

$$d_{\mathrm{W}}(P_1^n(1,\cdot),\pi_1) \leq 12980 * 0.94^n \qquad \text{for } n \geq 1 \text{ in case C.}$$

Therefore, $d_{\mathrm{W}}(P_1^n(1,\cdot),\pi_1) < 0.01$ for $n \geq 228$ in case C.

(4) ($K = 0$) Consider the three cases of Table 1, but now using the prior distribution with $K = 0$. Table 2 gives the calculations of Propositions 10(i) and 11(i) (note that $r_1 = 1/[2\alpha + J - 2]$); the last column is the bound on the Wasserstein distance from equilibrium after n iterations, started from $x = 1$. We find that $d_{\mathrm{W}}(P_0^n(1,\cdot),\pi_0) < 0.01$ for $n \geq 5$ in case A and for $n \geq 6$ in cases B and C.

4. From Wasserstein distance to total variation distance

4.1. One-shot coupling

In this section, we present Theorem 12, our main tool for converting Wasserstein convergence rates to total variation convergence rates. Various methods of coupling have been used for proving convergence in TV distance [5,13,19]. Although not explicit in the final formulation, the idea behind this theorem is a certain kind of coupling method, called *one-shot coupling*, which has been successfully applied to iterated function systems by Roberts and Rosenthal [18] (see also [2,12]). We describe this method now.

We shall consider two copies of a Markov chain, running simultaneously. Let S_0 and $\tilde{S}_0$ be two initial values for this chain (possibly random with some joint distribution). Let $\{f_t\}$ be a

sequence of i.i.d. random maps that defines this Markov chain. Define

$$S_t = f_t(S_{t-1}) \quad \text{and} \quad \tilde{S}_t = f_t(\tilde{S}_{t-1}) \qquad \text{for } t = 1, \ldots, n-1.$$

That is, we use the same realization of the functions f_t on both copies of the chains, up to time $n-1$. Suppose, at time n, we can find two copies $\hat{f}_n$ and $\hat{\hat{f}}_n$ of f_n, that are independent from everything earlier (but not independent of each other), such that, with high probability, we have $\hat{f}_n(S_{n-1}) = \hat{\hat{f}}_n(\tilde{S}_{n-1})$. (The name "one-shot coupling" refers to the fact that we only try to coalesce the two copies of the chain at the single time n.) By the representation (13), this would imply that S_n and $\tilde{S}_n$ are close to each other in TV distance. Two conditions help us to find such $\hat{f}_n$ and $\hat{\hat{f}}_n$: first, S_{n-1} and $\tilde{S}_{n-1}$ need to be reasonably close; second, the density functions of the two random variables $f_t(x)$ and $f_t(y)$ need to have a large overlap when x and y are close. Theorem 12 is a precise refinement of this argument.

In what follows, let (χ, ρ) be a complete separable metric space and let P be a transition probability operator on the state space χ. Assume that P has a density p with respect to some reference measure λ (that is, $P(x, \mathrm{d}z) = p(x, z)\lambda(\mathrm{d}z)$). Let μ be any probability distribution on χ and let π be a stationary probability distribution for P.

Theorem 12. (a) *Assume that there is a constant A such that*

$$\int_\chi |p(x, z) - p(y, z)|\lambda(\mathrm{d}z) \le A\rho(x, y) \qquad \text{for all } x, y \in \chi. \tag{30}$$

Then

$$d_{\mathrm{TV}}(\mu P^n, \pi) \le \frac{A}{2} d_{\mathrm{W}}(\mu P^{n-1}, \pi) \qquad \text{for all } n \ge 1.$$

(b) *Assume the following conditions hold*:

(i) *there exists a function $h > 0$ on χ such that*

$$\int_\chi |p(x, z) - p(y, z)|\lambda(\mathrm{d}z) \le \frac{\rho(x, y)}{\max\{h(x), h(y)\}} \qquad \text{for all } x, y \in \chi; \tag{31}$$

(ii) *there exist positive constants B, q and ε_0 such that*

$$\pi\big(\{y : h(y) < \varepsilon\}\big) \le B\varepsilon^q \qquad \text{for all } \varepsilon \text{ in } (0, \varepsilon_0). \tag{32}$$

Let $\tilde{C} = (2q)^{-q/(1+q)} \max\{(q+1)B^{1/(1+q)}, (B^q \varepsilon_0)^{-1/(1+q)}\}$. Then

$$d_{\mathrm{TV}}(\mu P^n, \pi) \le \tilde{C}[d_{\mathrm{W}}(\mu P^{n-1}, \pi)]^{q/(1+q)} \qquad \text{for all } n \ge 1. \tag{33}$$

Remarks. (1) If we also know $\limsup_{n\to\infty}[d_{\mathrm{W}}(\mu P^n, \pi)]^{1/n} \le \rho < 1$, then the conditions of Theorem 12(b) imply that $\limsup_{n\to\infty}[d_{\mathrm{TV}}(\mu P^n, \pi)]^{1/n} \le \rho^{q/(1+q)}$.

(2) Observe that condition (30) should not be expected to hold uniformly for x and y near 0 in the random logistic model. Indeed, as x decreases to 0, the density of $f_t(x)$ becomes more and more peaked near 0. Essentially, this is because 0 is a fixed point of the continuous random function f_t. The same thing happens in the Gibbs sampler example when K is 0.

(3) Lemma 3 will be useful in obtaining bounds of the form (30) or (31).

Our first step in proving the above theorem is the following calculation.

Lemma 13. *Let η and ν be probability measures on χ. Let Ψ be a probability measure in* Joint(η, ν). *Then*

$$d_{\text{TV}}(\eta P, \nu P) \leq \frac{1}{2} \int_x \int_y \int_z |p(x, z) - p(y, z)| \lambda(dz) \Psi(dx, dy). \tag{34}$$

Proof. Since $(\eta P)(dz) = (\int_x \eta(dx) p(x, z)) \lambda(dz)$ and similarly for νP, we apply equation (14) to obtain

$$d_{\text{TV}}(\eta P, \nu P) = \frac{1}{2} \int_z \left| \int_x \eta(dx) p(x, z) - \int_y \nu(dy) p(y, z) \right| \lambda(dz)$$

$$= \frac{1}{2} \int_z \left| \int_x \int_y p(x, z) \Psi(dx, dy) - \int_x \int_y p(y, z) \Psi(dx, dy) \right| \lambda(dz)$$

$$\leq \frac{1}{2} \int \int \int |p(x, z) - p(y, z)| \lambda(dz) \Psi(dx, dy). \qquad \square$$

Proof of Theorem 12. We shall apply Lemma 13 with $\eta = \mu P^{n-1}$ and $\nu = \pi \ (= \pi P)$. Recall from Section 2 that there is a probability measure $\Psi \equiv \Psi_{\eta, \nu}$ in Joint(η, ν) such that $d_{\text{W}}(\eta, \nu) = \int_x \int_y \rho(x, y) \Psi(dx, dy)$. The proof of part (a) follows immediately.

For part (b), let $\varepsilon > 0$. Observe that the left-hand side of equation (31) is never greater than 2. Lemma 13 and the assumption (31) then imply that

$$d_{\text{TV}}(\eta P, \nu P) \leq I_A + I_B, \tag{35}$$

where

$$I_A = \frac{1}{2} \int \int_{\{x, y \, : \, \max\{h(x), h(y)\} \geq \varepsilon\}} \frac{\rho(x, y)}{\max\{h(x), h(y)\}} \Psi(dx, dy)$$

and

$$I_B = \int \int_{\{x, y \, : \, \max\{h(x), h(y)\} < \varepsilon\}} 1 \Psi(dx, dy).$$

Note that

$$I_A \leq \frac{1}{2} \int \int_{\{x, y \, : \, \max\{h(x), h(y)\} \geq \varepsilon\}} \frac{\rho(x, y)}{\varepsilon} \Psi(dx, dy) \leq \frac{d_{\text{W}}(\mu, \nu)}{2\varepsilon}$$

and $I_B \leq \pi(\{y : h(y) < \varepsilon\})$. Combining these bounds with the assumption (32) tells us that

$$d_{\mathrm{TV}}(\mu P^n, \pi) \leq \frac{d_{\mathrm{W}}(\mu P^{n-1}, \pi)}{2\varepsilon} + B\varepsilon^q \qquad \text{for all } \varepsilon \in (0, \varepsilon_0).$$

Let $A_n = d_{\mathrm{W}}(\mu P^{n-1}, \pi)$ and consider the function $G_n(\varepsilon) = A_n/(2\varepsilon) + B\varepsilon^q$. Simple calculus shows that G_n is minimized at $\varepsilon_n := (\frac{A_n}{2Bq})^{1/(1+q)}$ and the minimum value of the function is $G_n(\varepsilon_n) = C_{Bq} A_n^{q/(1+q)}$, where $C_{Bq} = (q+1)(Bq^{-q}2^{-q})^{1/(1+q)}$. Let $\alpha_0 = 2Bq\varepsilon_0^{1+q}$. If $A_n < \alpha_0$, then $\varepsilon_n < \varepsilon_0$, so $d_{\mathrm{TV}}(\mu P^n, \pi) \leq G_n(\varepsilon_n)$. If $A_n \geq \alpha_0$, then, trivially, $d_{\mathrm{TV}}(\mu P^n, \pi) \leq 1 \leq \alpha_0^{-q/(1+q)} A_n^{q/(1+q)}$. Thus equation (33) holds with $\tilde{C} = \max\{C_{Bq}, \alpha_0^{-q/(1+q)}\}$. $\qquad\square$

4.2. Example 1: Normal Gibbs sampler

We return to the Gibbs sampler example described in Section 1. Recall that we write P_K, p_K and π_K to denote the corresponding transition kernel, density and stationary distribution, where $K \in \{0, 1\}$, without loss of generality.

Proposition 14. *Let μ be an arbitrary initial probability distribution on $(0, \infty)$. Then*

$$d_{\mathrm{TV}}(\mu P_1^n, \pi_1) \leq \frac{J}{2}\left(1 + \frac{|\bar{Y}|}{\sqrt{2\pi}}\right) d_{\mathrm{W}}(\mu P_1^{n-1}, \pi_1) \qquad \text{for } n = 1, 2, \ldots \tag{36}$$

and

$$d_{\mathrm{TV}}(\mu P_0^n, \pi_0) \leq \tilde{C} d_{\mathrm{W}}(\mu P_0^{n-1}, \pi_0)^w \qquad \text{for } n = 1, 2, \ldots, \tag{37}$$

where

$$w = \frac{2\alpha + J - 1}{2\alpha + J + 1}$$

and

$$\tilde{C} = \left(\alpha + \frac{J+1}{2}\right) e^{(1-w)\Sigma_0} (2\alpha + J - 1)^{-w}.$$

Before proceeding, let us revisit the numerical examples of Table 1, as discussed in the remarks following Proposition 11.

(a) $(K = 1)$ If $d_{\mathrm{W}}(\mu P_1^n, \pi_1) \leq QS^n$ for some constants Q and S, then $d_{\mathrm{TV}}(\mu P_1^n, \pi_1) \leq \frac{J}{2}(1 + |\bar{Y}|/\sqrt{2\pi})(Q/S)S^n$. Thus, for the case where μ is the point mass at $x = 1$, we obtain the following upper bounds on $d_{\mathrm{TV}}(\mu P_1^n, \pi_1)$: $63.3(5/6)^n$ in case A, $443(0.705)^n$ in case B and $48{,}294(0.94)^n$ in case C. Hence, the total variation distance to equilibrium is less then 0.01 when $n \geq 49$ in case A, when $n \geq 31$ in case B and when $n \geq 249$ in case C.

(b) $(K = 0)$ We have $w = 11/13$ in case A and $w = 3/4$ in cases B and C. Numerical values for $\tilde{C}$ (rounded up) are 8722 in case A, 3.642 in case B and 20.96 in case C. If we know that $d_{\mathrm{W}}(\mu P_0^n, \pi_0) \leq QS^n$, then we obtain $d_{\mathrm{TV}}(\mu P_0^n, \pi_0) \leq \tilde{C}(Q/S)^w (S^w)^n$. Thus, for the case where μ is the point mass at $x = 1$, we obtain the following upper bounds on

$d_{\text{TV}}(\mu P_0^n, \pi_0)$: $5{,}958{,}000(0.144)^n$ in case A, $174.6(0.338)^n$ in case B and $1624(0.314)^n$ in case C. Therefore, $d_{\text{TV}}(P_0^n(1, \cdot), \pi_0) < 0.01$ for $n \geq 11$ in cases A and C, and for $n \geq 10$ in case B.

Logically, the proof of this proposition belongs at the end of this section since it relies on several lemmas that have not yet been proven. However, we shall present the proof now since it serves as a guide for what is to come.

Proof of Proposition 14. Equation (36) follows from Theorem 12(a) and Lemma 16 below. Equation (37) follows from Theorem 12(b) and Lemmas 17 and 18 below. In Theorem 12(b), we use $q = \alpha + (J - 1)/2$, $B = e^{\Sigma_0}$ and $\varepsilon_0 = 1$ (all courtesy of Lemma 18), and it is not hard to check that, in the definition of $\tilde{C}$, the first term inside the 'max' exceeds the second. $\qquad\square$

The proof of Lemma 18 relies on our knowledge of the explicit form of the equilibrium distribution (which is known in many MCMC problems). The proofs of Lemmas 16 and 17 rely heavily on Lemma 3, together with the following technical lemma.

Lemma 15. *Let Z be a standard Normal random variable*:

(a) *Let a and b be positive constants. Then $d_{\text{TV}}(\frac{Z}{\sqrt{a}}, \frac{Z}{\sqrt{b}}) \leq |a - b|/\max\{a, b\}$.*

(b) *Let t be a real constant. Then $d_{\text{TV}}(Z, Z + t) \leq |t|/\sqrt{2\pi}$.*

Proof. For positive x, let $\phi_x(\cdot)$ be the probability density function of $Z/\sqrt{x}$, that is, $\phi_x(t) = \sqrt{\frac{x}{2\pi}} e^{-xt^2/2}$ $(t \in \mathbb{R})$.

(a) Without loss of generality, assume that $0 < a < b$. Using equation (15) and $e^{-at^2/2} > e^{-bt^2/2}$, we obtain

$$
d_{\text{TV}}\left(\frac{Z}{\sqrt{a}}, \frac{Z}{\sqrt{b}}\right) = \int_{t : \phi_b(t) > \phi_a(t)} \left(\sqrt{\frac{b}{2\pi}} e^{-bt^2/2} - \sqrt{\frac{a}{2\pi}} e^{-at^2/2}\right) dt
$$

$$
< \int_{t : \phi_b(t) > \phi_a(t)} \left(\sqrt{\frac{b}{2\pi}} e^{-bt^2/2} - \sqrt{\frac{a}{2\pi}} e^{-bt^2/2}\right) dt
$$

$$
\leq (\sqrt{b} - \sqrt{a}) \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-bt^2/2} \, dt
$$

$$
= (\sqrt{b} - \sqrt{a}) \frac{1}{\sqrt{b}} \leq \frac{|b - a|}{b}.
$$

Since $b = \max\{a, b\}$, this proves part (a).

(b) Let $\phi = \phi_1$, the probability density function of Z. Then $\phi(\cdot - t)$ is the probability density function of $Z + t$. By symmetry, we can assume that $t > 0$. Observe that the function $\min\{\phi(u), \phi(u - t)\}$ equals $\phi(u)$ for $u \geq t/2$ and is symmetric (with respect to u) about $u = t/2$.

Using this observation with equation (16) shows that

$$d_{\mathrm{TV}}(Z, Z + t) = 1 - \int_{-\infty}^{\infty} \min\{\phi(u), \phi(u - t)\}\, du$$

$$= 1 - 2\int_{t/2}^{\infty} \phi(u)\, du = \int_{-t/2}^{t/2} \phi(u)\, du \leq \frac{t}{\sqrt{2\pi}},$$

where we have used the bound $\phi(u) \leq 1/\sqrt{2\pi}$ for all u. This proves part (b). $\qquad\square$

Lemma 16 ($K = 1$). *For all positive x and y,*

$$d_{\mathrm{TV}}(P_1(x, \cdot), P_1(y, \cdot)) \leq J|x - y|\big(1 + |\bar{Y}|/\sqrt{2\pi}\big).$$

Proof. For given $s > 0$, $p_1(s, \cdot)$ is the probability density function of (8) with $K = 1$ and $\xi = 0$. Therefore, Lemma 3 implies that

$$d_{\mathrm{TV}}(P_1(x, \cdot), P_1(y, \cdot)) \leq d_{\mathrm{TV}}\left(\frac{Z}{\sqrt{a}} - \frac{\bar{Y}}{a}, \frac{Z}{\sqrt{b}} - \frac{\bar{Y}}{b}\right),$$

where $a = xJ + 1$, $b = yJ + 1$ and $Z \sim N(0, 1)$. We then have

$$d_{\mathrm{TV}}(P_1(x, \cdot), P_1(y, \cdot)) = d_{\mathrm{TV}}\left(\frac{Z}{\sqrt{a}}, \frac{Z}{\sqrt{b}} + \bar{Y}\left[\frac{1}{a} - \frac{1}{b}\right]\right)$$

$$\leq d_{\mathrm{TV}}\left(\frac{Z}{\sqrt{a}}, \frac{Z}{\sqrt{b}}\right) + d_{\mathrm{TV}}\left(\frac{Z}{\sqrt{b}}, \frac{Z}{\sqrt{b}} + \bar{Y}\left[\frac{b - a}{ab}\right]\right)$$

$$= d_{\mathrm{TV}}\left(\frac{Z}{\sqrt{a}}, \frac{Z}{\sqrt{b}}\right) + d_{\mathrm{TV}}\left(Z, Z + \bar{Y}\left[\frac{b - a}{a\sqrt{b}}\right]\right)$$

$$\leq \frac{|b - a|}{\max\{a, b\}} + \frac{|\bar{Y}||b - a|}{\sqrt{2\pi}a\sqrt{b}} \qquad \text{(by Lemma 15).}$$

Finally, since $|a - b| = J|x - y|$ and $a, b \geq 1$, the lemma follows. $\qquad\square$

Lemma 17 ($K = 0$). *For all positive x and y,*

$$d_{\mathrm{TV}}(P_0(x, \cdot), P_0(y, \cdot)) = \frac{1}{2}\int_0^{\infty} |p_0(x, z) - p_0(y, z)|\, dz \leq \frac{|x - y|}{\max\{x, y\}}.$$

Proof. The equality in the lemma comes from equation (14). Recall from equation (9) that $p_0(x, \cdot)$ is the probability density function of $G/(\Sigma_0 + \frac{1}{2}[Z/\sqrt{x}]^2)$, where G has a particular Gamma distribution and Z has the standard Normal distribution. Therefore, Lemma 3 implies that $d_{\mathrm{TV}}(P_0(x, \cdot), P_0(y, \cdot)) \leq d_{\mathrm{TV}}(\frac{Z}{\sqrt{x}}, \frac{Z}{\sqrt{y}})$ and Lemma 15(a) completes the proof. $\qquad\square$

Lemma 18 ($K = 0$). $\pi_0([0, \varepsilon]) \leq e^{\Sigma_0}\varepsilon^{\alpha + (J-1)/2}$ *for all ε in $(0, 1]$.*

Proof. The density $\pi_0(s)$ is the integral over θ of the posterior density $p(\theta, s|Y)$, which is given by equation (2) with $K = 0$. Using equation (5), we see that

$$p(\theta, s|Y) = \frac{1}{\zeta} s^{\alpha-1+J/2} \exp[-s J(\bar{Y} - \theta)^2/2] e^{-s\Sigma_0} \qquad \text{for } s > 0 \text{ and } \theta \in \mathbb{R},$$

where $\zeta = \zeta(\alpha, J, \Sigma_0, Y)$ is the normalizing constant. Truncating the double integral that defines ζ shows that

$$\zeta \geq e^{-\Sigma_0} \int_0^1 \int_{-\infty}^{\infty} s^{\alpha-1+J/2} \exp[-s J(\bar{Y} - \theta)^2/2] \, d\theta \, ds.$$

Therefore, for ε in (0, 1],

$$\pi_0([0, \varepsilon]) \leq \frac{1}{\zeta} \int_0^\varepsilon \int_{-\infty}^{\infty} s^{\alpha-1+J/2} \exp[-s J(\bar{Y} - \theta)^2/2] \, d\theta \, ds$$

$$\leq e^{\Sigma_0} \frac{\int_0^\varepsilon s^{\alpha-3/2+J/2} \, ds}{\int_0^1 s^{\alpha-3/2+J/2} \, ds} = e^{\Sigma_0} \varepsilon^{\alpha-1/2+J/2},$$

using $\int_{-\infty}^{\infty} \exp[-s J(\bar{Y} - \theta)^2/2] \, d\theta = (2\pi J s)^{-1/2}$ in the second inequality. $\qquad \square$

Remark. Although we did not do it, one can compute ζ exactly when $K = 0$. In most practical MCMC applications, the normalizing constant is hard to evaluate or even estimate – which is one reason that people use MCMC instead of numerical analysis. In general, finding constants B and ε_0 for equation (32) can be hard. The above proof suggests one way to approach the challenge.

4.3. Example 2: Random logistic maps

Recall that we are considering i.i.d. random maps $f_1, f_2, \ldots$ on [0, 1] defined by

$$f_i(x) = 4 B_i x(1 - x),$$

where $B_i \sim \text{Beta}(a + \frac{1}{2}, a - \frac{1}{2})$ [$a > \frac{1}{2}$], and that the Beta(a, a) distribution is the unique stationary distribution for the iterated function system.

In this subsection, we prove Theorem 1. The proof of this theorem is similar to the proof of the '$K = 1$' part of Proposition 14.

We begin with some notation. Let $b(t)$ be the density of the B_i's, that is,

$$b(t) = \begin{cases} K_a t^{a-1/2} (1 - t)^{a-3/2} & \text{for } 0 \leq t \leq 1, \\ 0 & \text{otherwise,} \end{cases}$$

where $K_a = \Gamma(2a)/\Gamma(a + \frac{1}{2})\Gamma(a - \frac{1}{2})$. Let

$$Q(x) = 4x(1 - x) \qquad \text{for } 0 \leq x \leq 1.$$

Observe that $0 \leq Q(x) \leq 1$ for $0 \leq x \leq 1$. For a given $x \in (0, 1)$, let $b_x(\cdot)$ be the probability density function of $B_i Q(x)$, that is,

$$b_x(z) = \begin{cases} \dfrac{1}{Q(x)} b\left(\dfrac{z}{Q(x)}\right) & \text{for } 0 \leq z \leq Q(x), \\ 0 & \text{otherwise.} \end{cases}$$

Next, let $p(x, z)$ denote the transition density of the Markov chain corresponding to the iterated logistic maps. We then have

$$p(x, z) = b_x(z) \qquad \text{for } x, z \in [0, 1]. \tag{38}$$

Lemma 19. *For the iterated logistic maps with $a > 1/2$, we have*

$$\frac{1}{2} \int_0^1 |p(x, z) - p(y, z)| \, dz \leq \frac{8a|y - x|}{\max\{Q(x), Q(y)\}} \qquad \text{for } x, y \in (0, 1).$$

Proof. Without loss of generality, assume that $0 < Q(x) \leq Q(y)$. By equation (38), Proposition 2 and some calculation similar to that which was involved in the proof of Lemma 15, we have

$$\begin{aligned}
\frac{1}{2} \int_0^1 & |p(x, z) - p(y, z)| \, dz \\
&= \int_{\{z : b_x(z) > b_y(z)\}} \left(b_x(z) - b_y(z)\right) dz \\
&= \int_{\{z : b_x(z) > b_y(z)\}} K_a \left(\frac{z^{a-1/2}(Q(x) - z)^{a-3/2}}{Q(x)^{2a-1}} - \frac{z^{a-1/2}(Q(y) - z)^{a-3/2}}{Q(y)^{2a-1}}\right) dz \\
&< \int_{\{z : b_x(z) > b_y(z)\}} K_a z^{a-1/2} \left(\frac{(Q(x) - z)^{a-3/2}}{Q(x)^{2a-1}} - \frac{(Q(x) - z)^{a-3/2}}{Q(y)^{2a-1}}\right) dz \\
&\qquad (\text{since } Q(y) - z \geq Q(x) - z \geq 0) \\
&= \left(\frac{1}{Q(x)^{2a-1}} - \frac{1}{Q(y)^{2a-1}}\right) \int_{\{z : b_x(z) > b_y(z)\}} K_a z^{a-1/2} (Q(x) - z)^{a-3/2} \, dz \\
&\leq \left(1 - \left(\frac{Q(x)}{Q(y)}\right)^{2a-1}\right) \int_0^{Q(x)} \frac{K_a z^{a-1/2}(Q(x) - z)^{a-3/2}}{Q(x)^{2a-1}} \, dz \\
&= 1 - \left(\frac{Q(x)}{Q(y)}\right)^{2a-1}.
\end{aligned} \tag{39}$$

We now observe that for $p > 0$,

$$v^p - u^p \leq \max\{p, 1\} v^{p-1} |v - u| \qquad \text{for } v \geq u \geq 0 \tag{40}$$

(for $0 < p \leq 1$, this is simple algebra and for $p > 1$, this follows from applying the mean value theorem to the function $t \mapsto t^p$). Next, since $|Q'(x)| = |4 - 8x| \leq 4$, the mean value theorem implies that

$$|Q(y) - Q(x)| \leq 4|y - x| \qquad \text{for } x, y \in [0, 1]. \tag{41}$$

Finally, for $0 < Q(x) \leq Q(y)$, equations (39)–(41) imply that

$$\frac{1}{2} \int_0^1 |p(x, z) - p(y, z)| \, dz \leq \frac{Q(y)^{2a-1} - Q(x)^{2a-1}}{Q(y)^{2a-1}}$$
$$\leq \frac{\max\{(2a - 1), 1\}|Q(y) - Q(x)|}{Q(y)}$$
$$\leq \frac{[(2a - 1) + 1]4|y - x|}{Q(y)}.$$

This proves the lemma. $\qquad\square$

We can now apply Theorem 12(b) as follows. Let $\mu = \delta_x$ (point mass at x) and let π_a be the equilibrium $\beta_{a,a}$ distribution. Also, let λ be Lebesgue measure and let the function $h(\cdot)$ be $Q(\cdot)/(16a)$. Lemma 19 then proves condition (i) of Theorem 12(b). For condition (ii), we need to estimate $\pi_a(\{y \in [0, 1] : h(y) \leq \varepsilon\})$ for small positive ε. Let $A = 16a$. Observe that if $Q(y)/A \leq \varepsilon$ and $0 \leq y \leq 1/2$, then $A\varepsilon \geq 4y(1 - y) \geq 4y(1/2)$, so $y \leq A\varepsilon/2$. Similarly, if $Q(y)/A \leq \varepsilon$ and $1/2 \leq y \leq 1$, then $y \geq 1 - A\varepsilon/2$. Therefore, for $a \geq 1$ and $0 < \varepsilon \leq 1/A$, we have

$$\pi_a(\{y \in [0, 1] : h(y) \leq \varepsilon\})$$
$$= \pi_a([0, A\varepsilon/2]) + \pi_a([1 - A\varepsilon/2, 1])$$
$$= 2\pi_a([0, A\varepsilon/2]) \qquad \text{(since } \pi_a \text{ is symmetric about } 1/2\text{)}$$
$$= \tilde{K}_a \int_0^{A\varepsilon/2} t^{a-1}(1 - t)^{a-1} \, dt \qquad \left(\text{where } \tilde{K}_a = \Gamma(2a)/\Gamma(a)^2\right) \tag{42}$$
$$\leq \tilde{K}_a \int_0^{A\varepsilon/2} t^{a-1} \, dt \tag{43}$$
$$= \frac{\tilde{K}_a(8a\varepsilon)^a}{a}.$$

Therefore, equation (32) holds with $q = a$, $B = \tilde{K}_a 8^a a^{a-1}$ and $\varepsilon_0 = 1/(16a)$. For $1/2 < a < 1$, everything is the same except that we use the bound $(1 - t)^{a-1} \leq 2^{1-a}$ for $0 < t \leq 1/2$ in the integrand of (42), obtaining an extra multiplicative factor of 2^{1-a} in equation (43) and hence $B = 2\tilde{K}_a 4^a a^{a-1}$. We have thus shown that Theorem 1 follows from Theorem 12(b).

Acknowledgements

References

[1] Barnsley, M.F. and Elton, J.H. (1988). A new class of Markov processes for image encoding. *Adv. in Appl. Probab.* **20** 14–32. MR0932532

[2] Beskos, A. and Roberts, G.O. (2005). One-shot CFTP: Application to a class of truncated Gaussian densities. *Methodol. Comput. Appl. Probab.* **7** 407–437. MR2235153

[3] Chamayou, J.-F. and Letac, G. (1991). Explicit stationary distributions for compositions of random functions and random matrices. *J. Theoret. Probab.* **4** 3–36. MR1088391

[4] Chen, M.F. (1992). *From Markov Chains to Non-Equilibrium Particle Systems*. Singapore: World Scientific. MR2091955

[5] Diaconis, P. (1988). *Group Representations in Probability and Statistics. IMS Lecture Notes – Monograph Series* **11**. Hayward, CA: Institute of Mathematical Statistics. MR0964069

[6] Diaconis, P. and Freedman, D. (1999). Iterated random functions. *SIAM Rev.* **41** 45–76. MR1669737

[7] Fayolle, G., Malyshev, V.A. and Menshikov, M.V. (1995). *Topics in the Constructive Theory of Countable Markov Chains*. Cambridge: Cambridge Univ. Press. MR1331145

[8] Gibbs, A.L. (2004). Convergence in the Wasserstein metric for Markov chain Monte Carlo algorithms with applications to image restoration. *Stoch. Models* **20** 473–492. MR2094049

[9] Gibbs, A.L. and Su, F.E. (2002). On choosing and bounding probability metrics. *Internat. Statist. Rev.* **70** 419–435.

[10] Gilks, W.R., Richardson, S. and Spiegelhalter, D.J., eds. (1996). *Markov Chain Monte Carlo in Practice*. London: Chapman and Hall. MR1397966

[11] Hobert, J.P. and Geyer, C.J. (1998). Geometric ergodicity of Gibbs and block Gibbs samplers for a hierarchical random effects model. *J. Multivariate Anal.* **67** 414–430. MR1659196

[12] Huber, M. (2007). Perfect simulation for image restoration. *Stoch. Models* **23** 475–487. MR2341079

[13] Jerrum, M. (1998). Mathematical foundations of the Monte Carlo method. In *Probabilistic Methods for Algorithmic Discrete Mathematics* (M. Habib et al., eds.) 116–165. Berlin: Springer. MR1678570

[14] Jones, G.L. and Hobert, J.P. (2001). Honest exploration of intractable probability distributions via Markov chain Monte Carlo. *Statist. Sci.* **16** 312–334. MR1888447

[15] Madras, N. (2002). *Lectures on Monte Carlo Methods. Fields Institute Monographs* **16**. Providence, RI: Amer. Math. Soc. MR1870056

[16] Meyn, S.P. and Tweedie, R.L. (1993). *Markov Chains and Stochastic Stability*. London: Springer. MR1287609

[17] Roberts, G.O. and Rosenthal, J.S. (1997). Geometric ergodicity and hybrid Markov chains. *Electron. Comm. Probab.* **2** 13–25 (electronic). MR1448322

[18] Roberts, G.O. and Rosenthal, J.S. (2002). One-shot coupling for certain stochastic recursive sequences. *Stochastic Process. Appl.* **99** 195–208. MR1901153

[19] Roberts, G.O. and Rosenthal, J.S. (2004). General state space Markov chains and MCMC algorithms. *Probab. Surv.* **1** 20–71 (electronic). MR2095565

[20] Rosenthal, J.S. (1995). Minorization conditions and convergence rates for Markov chain Monte Carlo. *J. Amer. Stat. Assoc.* **90** 558–566. MR1340509

[21] Saloff-Coste, L. (1997). Lectures on finite Markov chains. In *Lectures on Probability Theory and Statistics (St. Flour, 1996). Lecture Notes in Math.* **1665** 301–413. Berlin: Springer. MR1490046

[22] Steinsaltz, D. (1999). Locally contractive iterated function systems. *Ann. Probab.* **27** 1952–1979. MR1742896

[23] Steinsaltz, D. (2001). Random logistic maps and Lyapunov exponents. *Indag. Math. (N.S.)* **12** 557–584. MR1908881