3 May 1998

ITERATED RANDOM FUNCTIONS

Persi Diaconis
Department of Mathematics & ORIE
Cornell University
Ithaca, NY 14853

David Freedman
Department of Statistics
University of California
Berkeley, CA 94720

Abstract. Iterated random functions are used to draw pictures or simulate large Ising models, among other applications. They offer a method for studying the steady state distribution of a Markov chain, and give useful bounds on rates of convergence in a variety of examples. The present paper surveys the field and presents some new examples. There is a simple unifying idea: the iterates of random Lipschitz functions converge if the functions are contracting on the average.

1. Introduction. The applied probability literature is nowadays quite daunting. Even relatively simple topics, like Markov chains, have generated enormous complexity. This paper describes a simple idea that helps to unify many arguments in Markov chains, simulation algorithms, control theory, queuing, and other branches of applied probability. The idea is that Markov chains can be constructed by iterating random functions on the state space S. More specifically, there is a family $\{f_\theta : \theta \in \Theta\}$ of functions that map S into itself, and a probability distribution μ on Θ. If the chain is at $x \in S$, it moves by choosing θ at random from μ, and going to $f_\theta(x)$. For now, μ does not depend on x.

The process can be written as $X_0 = x_0$, $X_1 = f_{\theta_1}(x_0)$, $X_2 = (f_{\theta_2} \circ f_{\theta_1})(x_0)$, ..., with $\circ$ for composition of functions. Inductively,

$$X_{n+1} = f_{\theta_{n+1}}(X_n), \tag{1.1}$$

where $\theta_1, \theta_2, \ldots$ are independent draws from μ. The Markov property is clear: given the present position of the chain, the conditional distribution of the future does not depend on the past.

Key words and phrases. Products of random matrices, iterated function systems, coupling from the past.

Typeset by $\mathcal{A}\mathcal{M}\mathcal{S}$-TEX

We are interested in situations where there is a stationary probability distribution π on S with

$$P\{X_n \in A\} \to \pi(A) \quad \text{as} \quad n \to \infty.$$

For example, suppose S is the real line $\mathbb{R}$, and there are just two functions,

$$f_+(x) = ax + 1 \quad \text{and} \quad f_-(x) = ax - 1,$$

where a is given and $0 < a < 1$. In present notation, $\Theta = \{+, -\}$; suppose $\mu(+) = \mu(-) = 1/2$. The process moves linearly,

$$X_{n+1} = aX_n + \xi_{n+1}, \tag{1.2}$$

where $\xi_n = \pm 1$ with probability 1/2. The stationary distribution has an explicit representation, as the law of

$$Y_\infty = \xi_1 + a\xi_2 + a^2\xi_3 + \cdots. \tag{1.3}$$

The random series on the right converges to a finite limit because $0 < a < 1$. Plainly, the distribution of Y_∞ is unchanged if Y_∞ is multiplied by a and then a new ξ is added: that is stationarity. The series representation (1.3) can therefore be used to study the stationary distribution; however, many mysteries remain, even for this simple case (Section 2.5).

There are a wealth of examples based on affine maps in d-dimensional Euclidean space. The basic chain is

$$X_{n+1} = A_{n+1}X_n + B_{n+1},$$

where the (A_n, B_n) are independent and identically distributed; A_n is a $d \times d$ matrix and B_n is $d \times 1$ vector. Section 2 surveys this area. Section 2.3 presents an interesting application for $d = 2$: with an appropriately chosen finite distribution for (A_n, B_n), the Markov chain can be used to draw pictures of fractal objects like ferns, clouds, or fire. Section 3 describes finite state spaces where the backward iterations can be explicitly tested to see if they have converged. The lead example is the "coupling from the past" algorithm of Propp and Wilson (1996, 1998), which allows simulation for previously intractable distributions, such as the Ising model on a large grid.

Section 4 gives examples from queuing theory. Section 5 introduces some rigor, and explains a unifying theme. Suppose that S is a complete separable metric space. Write ρ for the metric. Suppose that each f_θ is Lipschitz: for some K_θ and all $x, y \in S$,

$$\rho[f_\theta(x), f_\theta(y)] \le K_\theta \rho(x, y).$$

For $x_0 \in S$, define the "forward iteration" starting from $X_0 = x_0$ by

$$X_{n+1} = f_{\theta_{n+1}}(X_n) = (f_{\theta_{n+1}} \circ \cdots \circ f_{\theta_2} \circ f_{\theta_1})(x_0),$$

$\theta_1, \theta_2, \ldots$ being independent draws from a probability μ on Θ; this is just a rewrite of equation (1.1). Define the "backward iteration" as

$$Y_{n+1} = (f_{\theta_1} \circ f_{\theta_2} \circ \cdots \circ f_{\theta_{n+1}})(x_0). \tag{1.4}$$

Of course, Y_n has the same distribution as X_n for each n. However, the forward process $\{X_n : n = 0, 1, 2, \ldots\}$ has very different behavior from the backward process $\{Y_n : n = 0, 1, 2, \ldots\}$: the forward process moves ergodically through S, while the backward process converges to a limit. (Naturally, there are assumptions.) The next theorem, proved in Section 5.2, shows that if f_θ is contracting on average, then $\{X_n\}$ has a unique stationary distribution π. The "induced Markov chain" in the theorem is the forward process X_n. The kernel $P_n(x, dy)$ is the law of X_n given that $X_0 = x$, and the Prokhorov metric is used for the distance between two probabilities on S. This metric will be defined in Section 5.1; it is denoted "ρ", like the metric on S. (Section 5.1 also takes care of the measure-theoretic details.)

Theorem 1. *Let (S, ρ) be a complete separable metric space. Let $\{f_\theta : \theta \in \Theta\}$ be a family of Lipschitz functions on S, and let μ be a probability distribution on Θ. Suppose that $\int K_\theta \, \mu(d\theta) < \infty$, $\int \rho[f_\theta(x_0), x_0] \, \mu(d\theta) < \infty$ for some $x_0 \in S$, and $\int \log K_\theta \, \mu(d\theta) < 0$.*

(i) *The induced Markov chain has a unique stationary distribution π.*
(ii) *$\rho[P_n(x, \cdot), \pi] \le A_x r^n$ for constants A_x and r with $0 < A_x < \infty$ and $0 < r < 1$; this bound holds for all times n and all starting states x.*
(iii) *The constant r does not depend on n or x; the constant A_x does not depend on n, and $A_x < a + b\rho(x, x_0)$ where $0 < a, b < \infty$.*

The condition that $\int \log K_\theta \, \mu(d\theta) < 0$ makes $K_\theta < 1$ for typical θ, and formalizes the notion of "contracting on average". The key step in proving Theorem 1 is proving convergence of the backward iterations (1.4).

Proposition 1. *Under the regularity conditions of Theorem 1, the backward iterations converge almost surely to a limit, at an exponential rate. The limit has the unique stationary distribution π.*

(A sequence of random variables X_n converges "almost surely" if the exceptional set—where X_n fails to converge—has probability 0.)

The queuing-theory examples in Section 4 are interesting for several reasons: in particular, the backward iterations converge although the functions are not contracting on average. Section 6 has some examples that illustrate the theorem, and show why the regularity conditions are needed. Section 7 extends the theory to cover Dirichlet random measures, the states of the Markov chain being probabilities on some underlying space (like the real line). Closed-form expressions can sometimes be given for the distribution of the mean of a random pick from the Dirichlet; Section 7.3 has examples.

Previous surveys on iterated random functions include Chamayou and Letac (1991) as well as Letac (1986). The texts by Baccelli and Brémaud (1994), Brandt et al. (1990), and Duflo (1997) may all be seen as developments of the random iterations idea; Meyn and Tweedie (1993) frequently use random iterations to illustrate the general theory.

2. Affine functions. This paper got started when we were trying to understand a simple Markov chain on the unit interval, described in Section 2.1. Section 2.2 discusses some general theory for recursions in $\mathbb{R}^d$ of the form $X_{n+1} = A_{n+1}X_n + B_{n+1}$, where the $\{A_n\}$ are random matrices and $\{B_n\}$ are random vectors. (In strict mathematical terminology, the function $X \to AX + B$ is "affine" rather than linear when $B \neq 0$.) Under suitable regularity conditions, these matrix recursions are shown to have unique stationary distributions. With affine functions, the conditions are virtually necessary and sufficient. The theory is applied to draw fractal ferns (among other objects) in Section 2.3. Moments and tail probabilities of the stationary distributions are discussed in Section 2.4. Sections 2.5–6 are about the "fine structure": how smooth are the stationary distributions?

2.1. Motivating example. A simple example motivated our study—a Markov chain whose state space $S = (0, 1)$ is the open unit interval. If the chain is at x, it picks one of the two intervals $(0, x)$ or $(x, 1)$ with equal probability 1/2, and then moves to a random y in the chosen interval. The transition density is

$$k(x, y) = \frac{1}{2}\frac{1}{x}\, 1_{(0,x)}(y) \;+\; \frac{1}{2}\frac{1}{1-x}\, 1_{(x,1)}(y). \tag{2.1}$$

As usual, $1_A(y) = 1$ or 0, according as $y \in A$ or $y \notin A$. The first term in the sum corresponds to a leftward move from x; the second, to a rightward move.

Did this chain have a stationary distribution? If so, could the distribution be identified? Those were our two basic questions. After some initial floundering, we saw that the chain could be represented as the iteration of random functions

$$\phi_u(x) = ux, \quad \psi_u(x) = x + u(1 - x),$$

with u chosen uniformly on $(0, 1)$ and ϕ, ψ chosen with probability 1/2 each.

Theorem 1 shows there is a unique stationary distribution. We identified this distribution by guesswork, but there is a systematic method. Begin by assuming that the stationary distribution has a density $f(x)$. From (2.1),

$$f(y) = \int_0^1 k(x, y) f(x)\, dx = \frac{1}{2}\int_y^1 \frac{f(x)}{x}\, dx + \frac{1}{2}\int_0^y \frac{f(x)}{1-x}\, dx. \tag{2.2}$$

Differentiation gives

$$f'(y) = -\frac{1}{2}\frac{f(y)}{y} + \frac{1}{2}\frac{f(y)}{1-y} \quad \text{or} \quad \frac{f'(y)}{f(y)} = \frac{1}{2}\Big(-\frac{1}{y} + \frac{1}{1-y}\Big),$$

so

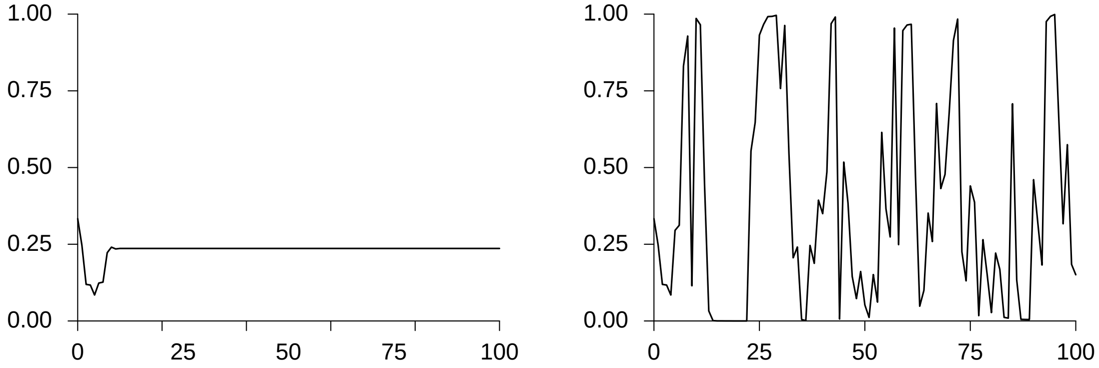
$$f(y) = \frac{1}{\pi\sqrt{y(1-y)}}. \tag{2.3}$$

This argument is heuristic, but it is easy to check that the "arcsine density" displayed in (2.3) satisfies equation (2.2)—and must therefore be stationary. The constant $\pi = 3.14\ldots$ makes $\int f(y)\,dy = 1$; the name comes about because

$$\int_0^z f(y)\,dy = \frac{2}{\pi}\arcsin\sqrt{z}.$$

Figure 1 illustrates the difference between the backward process (left hand panel, convergence) and the forward process (right hand panel, ergodic behavior). Position at time n is plotted against $n = 0, \ldots, 100$, with linear interpolation. Both processes start from $x_0 = 1/3$ and use the same random functions to move. The order in which the functions are composed is the only difference. In the left hand panel, the limit $0.236\ldots$ is random because it depends on the functions being iterated; but the limit does not depend on the starting point x_0.

Figure 1. The left hand panel shows convergence of the backward process; the right hand panel shows ergodic behavior by the forward process.

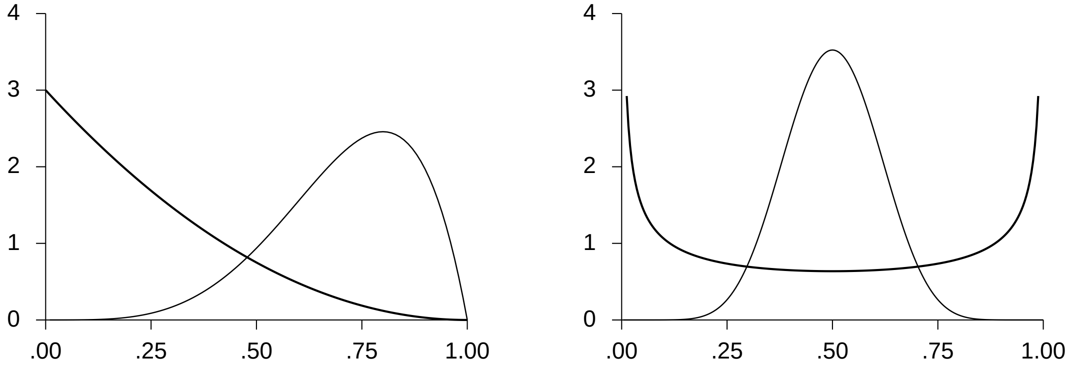


Remarks. Suppose $0 < p < 1$ and $q = 1 - p$. The same argument shows that choosing $(0, x)$ with probability p and $(x, 1)$ with probability q leads to a Beta(q, p) stationary distribution, with density $Cx^{q-1}(1-x)^{p-1}$ on $(0, 1)$. The normalizing constant is $C = \Gamma(q+p)/[\Gamma(q)\Gamma(p)]$, where Γ is Euler's gamma function. In our example, $q + p = 1$, so $\Gamma(q + p) = \Gamma(1) = 1$.

Although we will not pursue this idea, the probability p of moving to $(0, x)$ from x can even be allowed to depend on x. For example if $p(x) = x$, the stationary distribution is uniform. However, Theorem 1 is not in force when $p(x)$ depends on x. For instance, if $p(x) = 1 - x$, the process converges to 0 or 1 almost surely: if the starting state is x, the chance of converging to 1 is x. (The process is a martingale, and convergence follows from standard theorems.) Theorem 1 can be extended to cover μ that depend on x, but further conditions are needed.

Many of the constructions in this paper involve the Beta distribution. Figure 2 plots some of the densities. The stationary density (2.3) in our lead example is Beta($\frac{1}{2}, \frac{1}{2}$)—the bowl-shaped curve in the right hand panel; we return to this example in Section 6.3.

Figure 2. The Beta distribution. The left hand panel plots the Beta(1,3)-density (heavy line) and the Beta(5,2)-density (light line). The right hand panel plots the Beta($\frac{1}{2}, \frac{1}{2}$)-density (heavy line) and the Beta(10,10) density (light line).



2.2. Matrix recursions. Matrix recursions have been used in a host of modeling efforts; see, for instance, Priestley (1988). To define things in $\mathbb{R}^d$, let $X_0 = x_0 \in \mathbb{R}^d$, and

$$X_{n+1} = A_{n+1}X_n + B_{n+1} \text{ for } n = 0, 1, 2, \dots, \tag{2.4}$$

with (A_n, B_n) being i.i.d.; A_n is a $d \times d$ matrix and B_n is a $d \times 1$ vector: i.i.d. is the usual short-hand for "independent and identically distributed". Autoregressive processes like (2.4) will be discussed again in Section 6.1. Under suitable regularity conditions, the stationary distribution can be represented as the law of

$$B_1 + A_1B_2 + A_1A_2B_3 + A_1A_2A_3B_4 + \cdots . \tag{2.5}$$

Indeed, suppose this sum converges a.s. to a finite limit. The distribution is unchanged if a fresh (A, B) pair is chosen, the sum is multiplied by A, and then B is added: that is stationarity.

The notation may be a bit perplexing: A_n, B_n, A, B are all random rather than deterministic, and "a.s." is short-hand for "almost surely": the sum converges except for an event of probability 0. Conditions for convergence have been sharpened over the years; roughly, A_n must be a contraction "on average". Following work by Vervaat (1979) and Brandt (1986), definitive results were achieved by Bougerol and Picard (1992). To state the result, let $\| \;\|$ be a matrix norm on $\mathbb{R}^d$. Suppose that (A_n, B_n) are i.i.d. for $n = 1, 2, \dots,$ with

$$E\{\log^+ \|A_n\|\} < \infty, \quad E\{\log^+ \|B_n\|\} < \infty, \tag{2.6}$$

where $x^+ = x$ when $x > 0$ and $x^+ = 0$ when $x < 0$. A subspace L of $\mathbb{R}^d$ is "invariant" if $P\{X_1 \in L | X_0 = x\} = 1$ for all $x \in L$.

Theorem 2.1. *Assume (2.6) and define the Markov chain X_n by (2.4). Suppose that the only invariant subspace of $\mathbb{R}^d$ is $\mathbb{R}^d$ itself. The infinite random series*

$$\sum_{j=1}^{\infty}\Big(\prod_{i=1}^{j-1} A_i\Big)B_j \tag{2.7}$$

converges a.s. to a finite limit if and only if

$$\inf_{n>0}\frac{1}{n}E\{\log\|A_1\cdots A_n\|\} < 0. \tag{2.8}$$

If (2.8) holds, the distribution of (2.7) is the unique invariant distribution for the Markov chain X_n.

The moment assumptions in Theorem 2.1 cannot be essentially weakened; see Goldie and Maller (1997). Of course, the Markov chain (2.4) can be defined when A_n is expanding rather than contracting, but different normings are required for convergence. Anderson (1959) and Rachev-Samorodnitzky (1995) prove central limit theorems in the non-contractive case. On a lighter note, Embree and Trefethen (1998) use this machinery with $d = 2$ to study Fibonacci sequences with random signs and a damping parameter β, so $X_{n+1} = X_n \pm \beta X_{n-1}$.

2.3. Fractal images. This section shows how iterated random affine maps can be used to draw pictures in two dimensions. Fix $(a_1, b_1), \ldots, (a_k, b_k)$. Each a_i is a 2×2 contraction, while b_i is a 2×1 vector: $f_i(x) = a_i x + b_i$ is the associated affine map of the plane into itself, which is Lipschitz because a_i is a contraction. Fix positive weights $w_1, \ldots, w_k$, with $w_1 + \cdots + w_k = 1$. These ingredients specify a Markov chain $\{X_n\}$ moving through $\mathbb{R}^2$. Starting at x, the chain proceeds by choosing i at random with probability w_i and moving to $f_i(x)$.

Remarkably enough, given a target image, one can often solve for $\{a_i, b_i, w_i\}$ so that the collection of points $\{X_1, \ldots, X_N\}$ forms a reasonable likeness of the target, at least with high probability. The technique is based on work of Dubins and Freedman (1966), Hutchinson (1981), and Diaconis and Shahshahani (1986). It has been developed further by Barnsley and Elton (1988) as well as Barnsley (1993), and is now widely used.

We outline the procedure. Theorem 1 applies, so there is a unique stationary distribution, call it π. Let δ_x stand for point mass at x: that is, $\delta_x(A) = 1$ if $x \in A$ and $\delta_x(A) = 0$ if $x \notin A$. According to standard theorems, the empirical distribution of $\{X_1, \ldots, X_N\}$ converges to π:

$$\frac{1}{N}\sum_{i=1}^{N}\delta_{X_i} \to \pi.$$

Convergence is almost sure, in the weak-star topology. For any bounded continuous function f on $\mathbb{R}^2$,

$$\lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N} f(X_i) = \int_{\mathbb{R}^2} f\,d\pi \quad \text{with probability 1.}$$

See, for instance, Breiman (1960). In short, the pattern generated by the points $\{X_1, \ldots, X_N\}$ looks like π when N is large.

The parameters $\{a_i, b_i, w_i\}$ must be chosen so that π represents the target image. Here is one of the early algorithms. Suppose a picture is given as black and white points on an $m \times m$ grid. Corresponding to this picture there is a discrete probability measure ν on the plane, which assigns mass $1/b$ to each black point and mass 0 to each white point, b being the number of black points. We want the stationary π to approximate ν. Stationarity implies that for any bounded continuous function f on $\mathbb{R}^2$,

$$\sum_{i=1}^{k} w_i \int_{\mathbb{R}^2} f(a_i x + b_i)\, \pi(dx) = \int_{\mathbb{R}^2} f(x)\, \pi(dx). \tag{2.9}$$

The next idea is to replace $\int f d\pi$ on the right side of (2.9) by $\int f d\nu$:

$$\sum_{i=1}^{k} w_i \int_{\mathbb{R}^2} f(a_i x + b_i)\, \pi(dx) \doteq \int_{\mathbb{R}^2} f(x)\, \nu(dx). \tag{2.10}$$

For appropriate f's, we get a system of equations that can be solved—at least approximately—for $\{a_i, b_i, w_i\}$. For instance, take f to be linear or a low-order polynomial (and ignore complications due to unboundedness). In (2.10), the unknowns are the a_i, b_i, w_i. The equations are linear in the w's but nonlinear in the other unknowns. Exact solutions cannot be expected in general, because ν will be discrete while π will be continuous. Still, the program is carried out by Diaconis and Shahshahani (1986) and by many later authors; see Barnsley (1993) for a recent bibliography. Also see Fisher (1994).

Figure 3. A fern drawn by a Markov chain

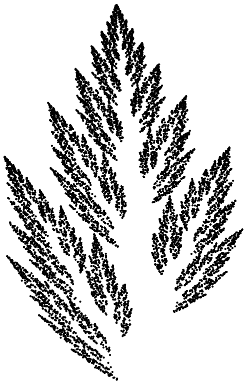


Figure 3 shows a picture of a fern. The parameters were suggested by Crownover (1995): $N = 10000$, $k = 2$, $w_1 = .2993$, $w_2 = .7007$, and

$$a_1 = \begin{pmatrix} +.4000 & -.3733 \\ +.0600 & +.6000 \end{pmatrix}, \quad b_1 = \begin{pmatrix} +.3533 \\ +.0000 \end{pmatrix},$$

$$a_2 = \begin{pmatrix} -.8000 & -.1867 \\ +.1371 & +.8000 \end{pmatrix}, \quad b_2 = \begin{pmatrix} +1.1000 \\ +0.1000 \end{pmatrix}.$$

2.4. Tail behavior. We turn now to the tail behavior of the stationary distribution. Some information can be gleaned from the moments, and invariance gives a recursion. We discuss (a bit informally) the case $d = 1$. Let (A_n, B_n) be i.i.d. pairs of real-valued random variables. Define the Markov chain $\{X_n\}$ by (2.4), and suppose the chain starts from its stationary distribution π. Write $\mathcal{L}(X)$ for the law of X. Then $\mathcal{L}(X_1) = \mathcal{L}(A_1X_0 + B_1)$, which implies $E(X_0) = E(X_1) = E(A_1)E(X_0) + E(B_1)$; so $E(X_0) = E(B_1)/[1 - E(A_1)]$. Similar expressions can be derived for higher moments and $d > 1$. See, for instance, Vervaat (1979) or Diaconis and Shashahani (1986); also see (6.4) below.

Moments may not exist, or may not capture relevant aspects of tail behavior. Under suitable regularity conditions, Kesten (1973) obtained estimates for the tail probabilities of the stationary π. For instance, when $d = 1$, he shows there is a positive real number κ such that $\pi(t, \infty) \approx C_+/t^\kappa$ and $\pi(-\infty, -t) \approx C_-/t^\kappa$ as $t \to \infty$. Goldie (1991) gives a different proof of Kesten's theorem and computes $C_\pm$; also see Babillot et al. (1997). Of course, there is still more to understand. For example, if A_n is uniform on $[0, 1]$, Z_n is independent Cauchy, and $B_n = (1 - A_n)Z_n$, the stationary distribution for $\{X_n\}$ is Cauchy. Thus, the conclusions of Kesten's theorem hold—although the assumptions do not. Section 7.3 contains other examples of this sort. It would be nice to have a theory that handles tail behavior in such examples.

2.5. Fine Structure. Even with an explicit representation for the stationary distribution, there are still many questions. Consider the chain described by equation (1.2). As in (1.3), the stationary distribution is the law of

$$Y_\infty = \xi_1 + a\xi_2 + a^2\xi_3 + \cdots,$$

the ξ_n being i.i.d. with $P(\xi_n = \pm 1) = 1/2$. We may ask about the "type" of π: is this measure discrete, continuous but singular, or absolutely continuous? (The terminology is reviewed below.) By the "law of pure types", mixtures cannot arise; and discrete measures can be ruled out too. See Jessen and Wintner (1935).

If $a = 1/2$, then π is just Lebesgue measure on $[-2, 2]$. If $0 < a < 1/2$, then π is singular. Indeed,

$$\xi_1 + a\xi_2 + \cdots + a^{N-1}\xi_N$$

takes on at most 2^N distinct values. For the remainder term,

$$0 < \sum_{j=N}^{\infty} a^j \xi_{j+1} < \frac{a^N}{1-a}.$$

Hence, π concentrates on a set of of intervals of total length $2^N a^N/(1 - a)$, which tends to 0 as N gets large—because $a < 1/2$.

It is natural to guess that π is absolutely continuous for $a > 1/2$. However, this is false. For example, if $a = (\sqrt{5} - 1)/2 = .618\ldots$, then π is singular: see

Erdös (1939, 1940). Which values of a give singular π's? This problem has been actively studied for 50 years, with no end in sight. See Garsia (1962) for a review of the classical work. There was a real breakthrough when Solomyak (1995) proved that π is absolutely continuous for almost all values of a in $[1/2, 1]$; also see Peres and Solomyak (1996, 1998).

2.6. Terminology. A "discrete" probability assigns measure 1 to a countable set of points, while a "continuous" probability assigns measure 0 to every point. A "singular" probability assigns measure 1 to a set of Lebesgue measure 0. By contrast, an "absolutely continuous" probability has a density with respect to Lebesgue measure. Textbook examples like the Binomial and Poisson distributions are discrete; the Normal, Cauchy, and Beta distributions are absolutely continuous. Ordering the rationals in $[0, 1]$ and putting mass $1/2^n$ on the nth rational gives you an interesting discrete probability. The uniform distribution on the Cantor set in $[0, 1]$ is continuous but singular.

3. The Propp-Wilson Algorithm.

This remarkable algorithm does exact Monte Carlo sampling from distributions on huge finite state spaces. Let S be the state space and let π be a probability on S. The objective is to make a random pick from π, on the computer. When S is large and π is complicated, the project can be quite difficult and the backward iteration is a valuable tool.

To begin with, there is a family of functions $\{f_\theta : \theta \in \Theta\}$ from S to S and a probability μ on Θ, so that π is the stationary distribution of the forward chain on S. In other words, for each $t \in S$,

$$\sum_{s \in S} \pi(s)\mu\{\theta : f_\theta(s) = t\} = \pi(t). \tag{3.1}$$

These functions will be constructed below. In some cases, the Metropolis algorithm is useful (Metropolis et al., 1953). In the present case, as will be seen, the Gibbs sampler is the construction to use. The probability μ on Θ will be called the "move measure": the chain moves by picking θ from μ and going from $s \in S$ to $f_\theta(s)$. If the construction is successful, the backward iterations

$$(f_{\theta_1} \circ f_{\theta_2} \circ \cdots \circ f_{\theta_n})(s) \tag{3.2}$$

will converge almost surely to a limiting random variable whose distribution is π. (A sequence in S converges if it is eventually constant, and $\theta_1, \theta_2, \ldots$ are independent draws from the move measure μ on Θ.)

Convergence is easier to check if there is monotonicity. Suppose S is a partially ordered set; write $s < t$ if s precedes t. Suppose too there is a smallest element 0 and a largest element 1. With partial orderings, the existence of a largest element is an additional assumption, even for a finite set; likewise for smallest. Finally, suppose that each f_θ is monotone: $s < t$ implies $f_\theta(s) \leq f_\theta(t)$. Now convergence is forced if, for some n,

$$(f_{\theta_1} \circ f_{\theta_2} \circ \cdots \circ f_{\theta_n})(0) = (f_{\theta_1} \circ f_{\theta_2} \circ \cdots \circ f_{\theta_n})(1). \tag{3.3}$$

This takes a moment to verify. Among other things, convergence would not be forced if we had equality on the forward iteration.

Propp and Wilson (1996, 1998) turn these observations into a practical algorithm for choosing a point at random from π. They make a sequence $\theta_1, \theta_2, \theta_3, \ldots$ of independent picks from the move measure μ in (3.1), and compute the backward iterations (3.2). At each stage, they check to see if (3.3) holds. If so, the common value—of the left side and the right side—is a pick from the exact stationary distribution π. The algorithm generates a random element of S whose distribution is the sought-for π itself, rather than an approximation to π; there is an explicit test for convergence; and in many situations, convergence takes place quite rapidly. These three features are what make the algorithm so remarkable.

By way of example, take the Ising model on an $n \times n$ grid; a reference is Kinderman and Snell (1980). The state space S consists of all functions s from $\{1, \ldots, n\} \times \{1, \ldots, n\}$ to $\{-1, +1\}$. The standard (barbaric) notation has $S = \{\pm 1\}^{[n]\times[n]}$. In the partial order, $s < t$ iff $s_{ij} \le t_{ij}$ for all positions (i, j) in the grid, and $s \neq t$. A boundary condition may be imposed, for instance, that $s = +1$ on the perimeter of the grid. The minimal state is -1 at all the unconstrained positions; the maximal state is $+1$ at all the unconstrained positions.

The probability distribution to be simulated is

$$\pi(s) = C_\beta e^{\beta H(s)}. \tag{3.4}$$

Here, β is a positive real number and C_β is a normalizing constant—which is quite hard to compute if n is large. In the exponent, $H(s)$ counts sign changes. Algebraically,

$$H(s) = \sum_{ij,k\ell} s_{ij} s_{k\ell}. \tag{3.5}$$

The indices i, j, k, ℓ run from 1 to n, and the position (i, j) must be adjacent to (k, ℓ): for instance, the position $(2, 2)$ is adjacent to $(2, 3)$ but not to $(3, 3)$.

A "single site heat bath"(a specialized version of the Gibbs sampler) is used to construct a chain with limiting distribution π. From state s, the chain moves by picking a site (i, j) on the grid

$$\{1, \ldots, n\} \times \{1, \ldots, n\}$$

and re-randomizing the value at (i, j). More specifically, let s_{ij+} agree with s at all sites other than (i, j); let $s_{ij+} = +1$ at (i, j). Likewise, s_{ij-} agrees with s at all sites other than (i, j), but $s_{ij-} = -1$ at (i, j). Let

$$\pi(+) = \frac{\exp[\beta H(s_{ij+})]}{\exp[\beta H(s_{ij+})] + \exp[\beta H(s_{ij-})]}$$

and $\pi(-) = 1 - \pi(+)$. The chance of moving to s_{ij+} from s is $\pi(+)$; the chance of moving to s_{ij-} is $\pi(-)$. In other words, the chance of re-randomizing to $+1$ at (i, j) is $\pi(+)$. This chance is computable because the ugly constant C_β has canceled out.

In principle, $\pi(+)$ and $\pi(-)$ depend on the site (i,j) and on values of s at sites other than (i,j); we write $\pi(\pm \mid i\,j\,s)$ when this matters. Of course, $\pi(+)$ is just the conditional π-probability that $s_{ij} = +$, given the values of s at all other sites. As it turns out, only the sites adjacent to (i,j) affect $\pi(+)$, because the values of s at more remote sites just cancel:

$$\pi(+ \mid i\,j\,s) = \frac{\exp\big(\beta \sum_{k\ell} s_{k\ell}\big)}{\exp\big(\beta \sum_{k\ell} s_{k\ell}\big) + \exp\big(-\beta \sum_{k\ell} s_{k\ell}\big)}\,. \tag{3.6}$$

The sum is over the sites (k,ℓ) adjacent to (i,j). Equation (3.6) is in essence the "Markov random field" property for the Ising model.

The single site heat bath can be cycled through sites (i,j) on the grid, or the site can be chosen at random. We follow the latter course, although the former is computationally more efficient. The algorithm is implemented using the backward iteration. The random functions are $f_\theta(s)$. Here, $s \in S$ is a state in the Ising model while $\theta = (i,j,u)$ consists of a position (i,j) in the grid and a real number u with $0 < u < 1$. The position is randomly chosen in the grid, and u is random over $(0,1)$. The function f is defined as follows: $s' = f_{iju}(s)$ agrees with s except at position (i,j). There, $s'_{ij} = +1$ if $u < \pi(+)$, and $s'_{ij} = -1$ otherwise.

Two things must be verified:

(i) π is stationary, and
(ii) f_θ is monotone.

Stationarity is obvious. For monotonicity, fix a site (i,j), two states s, t with $s \le t$, and $u \in (0,1)$. Clearly, $f_{iju}(s) \le f_{iju}(t)$ except perhaps at (i,j). At this special site, we must prove

$$\pi(+ \mid i\,j\,s) \le \pi(+ \mid i\,j\,t). \tag{3.7}$$

But the two conditional probabilities in (3.7) can be evaluated by (3.6), and

$$\sum_{k\ell} s_{k\ell} \le \sum_{k\ell} t_{k\ell}.$$

The condition $\beta > 0$ makes f_θ monotone increasing rather than monotone decreasing. The backward iteration completes after a finite, random number of steps, essentially by Theorem 1. Completion can be tested explicitly using (3.3). And the algorithm makes a random pick from π itself, rather than an approximation to π.

There are many variations on the Propp-Wilson algorithm, including some for point processes: see Møller (1998) or Häggström et al. (1998). A novel alternative is proposed by Fill (1998), who includes a survey of recent literature and a warning about biases due to aborted runs. There are no general bounds on the time to "coupling", which occurs when (3.3) is satisfied: chains starting from 0 and from 1, but using the same θ's, would have to agree from that time onwards. Experiments show that coupling generally takes place quite rapidly for the Ising model with β below a critical value, but quite slowly for larger β's. Propp and Wilson (1996) have algorithms that work reasonably well for all values of β—even above the critical value—and for grids up to size 2100 × 2100. For more discussion, and a comparison of the Metropolis algorithm with the Gibbs sampler, see Häggström and Nelander (1998).

Brown and Diaconis (1997) show that a host of Markov chains for shuffling and random tilings are monotone. These chains arise from hyperplane walks of Bidigare, Hanlon and Rockmore (1997). The analysis gives reasonably sharp bounds on time to coupling. Monotonicity techniques can be used for infinite state spaces too. For instance, such techniques have been developed by Borovkov (1984) and Borovkov and Foss (1992) to analyze complex queuing networks—our next topic.

4. Queuing theory. The existence of stationary distributions in queuing theory can often be proved using iterated random functions. There is an interesting twist, because the functions are generally not strict contractions, even on average. We give an example, and pointers to a voluminous literature. In one relatively simple model, the G/G/1 queue, customers arrive at a queue with i.i.d. interarrival times $U_1, U_2, \ldots$. The arrival times are the partial sums $0, U_1, U_1 + U_2, \ldots$. The jth customer has service time V_j; these too are i.i.d., and independent of the arrival times. Let W_j be the waiting time of the jth customer—the time before service starts. By definition, $W_0 = 0$. For $j > 0$, the W_j satisfy the recursion

$$W_{j+1} = (W_j + V_j - U_{j+1})^+. \tag{4.1}$$

Indeed, the jth customer arrives at time $T_j = U_1 + \cdots + U_j$ and waits time W_j, finishing service at time $T_j + W_j + V_j$. The $j+1$st customer arrives at time $T_j + U_{j+1}$. If $T_j + U_{j+1} > T_j + W_j + V_j$, then $W_{j+1} = 0$; otherwise, $W_{j+1} = W_j + V_j - U_{j+1}$.

The waiting-time process $\{ W_j : j = 0, 1, \ldots \}$ can therefore be generated by iterating the random functions

$$f_\theta(x) = (x + \theta)^+. \tag{4.2}$$

The parameter θ should be chosen at random from $\mu = \mathcal{L}(V_j - U_{j+1})$, which is a probability on the real line $\mathbb{R}$.

The function f_θ is a weak contraction but not a strict contraction: the Lipschitz constant is 1. Although Theorem 1 does not apply, the backward iteration still gives the stationary distribution. Indeed, the backward iteration starting from 0 can be written as

$$(f_{\theta_1} \circ \cdots \circ f_{\theta_n})(0) = \Big(\theta_1 + \big(\theta_2 + \cdots + (\theta_{n-1} + \theta_n^+)^+\big)^+\Big)^+. \tag{4.3}$$

Now there is a magical identity:

$$\Big(\theta_1 + \big(\theta_2 + \cdots + (\theta_{n-1} + \theta_n^+)^+\big)^+\Big)^+ = \max_{1 \le j \le n} (\theta_1 + \cdots + \theta_j)^+. \tag{4.4}$$

This identity holds for any real numbers $\theta_1, \ldots, \theta_n$. Feller (1971, p. 272) asks the reader to prove (4.4) by induction, and $n = 1$ is trivial. Separating the cases $y \le 0$

and $y > 0$, one checks that $(x + y^+)^+ = \max\{0, x, x + y\}$. That does $n = 2$. Now put θ_2 for x and θ_3 for y:

$$\begin{aligned}\big(\theta_1 + (\theta_2 + \theta_3^+)^+\big)^+ &= \big(\theta_1 + \max\{0, \theta_2, \theta_2 + \theta_3\}\big)^+ \\ &= \big(\max\{\theta_1, \theta_1 + \theta_2, \theta_1 + \theta_2 + \theta_3\}\big)^+ \\ &= \max\{0, \theta_1, \theta_1 + \theta_2, \theta_1 + \theta_2 + \theta_3\}.\end{aligned}$$

That does $n = 3$. And so forth. If the starting point is x rather than 0, you just need to replace θ_n in (4.4) by $\theta_n + x$.

In the queuing model, $\{U_j\}$ are i.i.d. by assumption, as are $\{V_j\}$; and the U's are independent of the V's. Set $X_j = V_j - U_{j+1}$ for $j = 1, 2, \ldots$. So the X_j are i.i.d. too. It is easily seen—given (4.3–4)—that the Markov chain $\{W_j : j = 0, 1, \ldots, \infty\}$ has for its stationary distribution the law of

$$\lim_{n\to\infty} \max_{1\le j\le n} (X_1 + \cdots + X_j)^+, \tag{4.5}$$

provided the limit is finite a.s.

Many authors now use the condition $E(X_1) < 0$ to insure convergence, via the strong law of large numbers: $X_1 + \cdots + X_j \approx jE(X_1) \to -\infty$ a.s., so the maximum of the partial sums is finite a.s. In a remarkable paper, Spitzer (1956) showed that no moment assumptions are needed.

Theorem 4.1. *Suppose the random variables $X_1, X_2, \ldots$ are i.i.d. The limit in (4.5) is finite a.s. if and only if*

$$\sum_{j=1}^{\infty} \frac{1}{j} P\{X_1 + \cdots + X_j > 0\} < \infty.$$

Under this condition, the limit in (4.5) has an infinitely divisible distribution with characteristic function

$$\prod_{j=1}^{\infty} \exp[\frac{1}{j}(\psi_j(t) - 1)],$$

where $\psi_j(t) = E\{\exp[it(X_1 + \cdots + X_j)^+]\}$ and $\exp x = e^x$.

The "G/G/1" in the G/G/1 queue stands for general arrival times, general service times, and one server: "general" means that $\mathcal{L}(U_j)$ and $\mathcal{L}(V_j)$ are not restricted to parametric families. The recent queuing literature contains many elaborations, including for instance queues with multiple servers and different disciplines; see Baccelli (1992) among others. There are surveys by Borovkov (1984) or Baccelli and Brémaud (1994). One remarkable achievement is the development of a sort of linear algebra for the real numbers under the operation $(x, y) \to \max\{x, y\}$ and $x \to x^+$. The book by Baccelli et al. (1992) gives many applications; queues are discussed in Chapter 7. The random-iterations idea helps to unify the arguments.

5. Rigor.

This section gives a more formal account of the basic setup; then Theorem 1 is proved in Section 5.2. The theorem and the main intermediate results are known: see Arnold and Crauel (1992), Barnsley and Elton (1988), Dubins and Freedman (1966), Duflo (1997), Elton (1990), or Hutchinson (1981). Even so, the self-contained proofs given here may be of some interest.

5.1. Background.

Let (S, ρ) be a complete, separable metric space. Then $f \in \mathrm{Lip}_K$ if f is a mapping of S into itself, with $\rho[f(x), f(y)] \le K\rho(x, y)$. The least such K is K_f. If f is constant, then $K_f = 0$. If $f \in \mathrm{Lip}_K$ for some $K < \infty$, then f is "Lipschitz"; otherwise, $K_f = \infty$. Of course, these definitions are relative to ρ. We pause for the measure theory. Let S_0 be a countable dense subset of S, and let $\overline{\mathcal{X}}$ be the set of all mappings from S_0 into S. We endow $\overline{\mathcal{X}}$ with the product topology and product σ-field. Plainly, $\overline{\mathcal{X}}$ is a complete separable metric space. Let $\mathcal{X}$ be the space of Lipschitz functions on S. The following lemma puts a measurable structure on $\mathcal{X}$.

Lemma 5.1.

(i) *$\mathcal{X}$ is a Borel subset of $\overline{\mathcal{X}}$.*
(ii) *$f \to K_f$ is a Borel function on $\mathcal{X}$.*
(iii) *$(f, s) \to f(s)$ is a Borel map from $\mathcal{X} \times S$ to S.*

Proof: For $f \in \overline{\mathcal{X}}$, let

$$L_f = \sup_{x \ne y \in S_0} \rho[f(x), f(y)]/\rho(x, y) \le \infty.$$

Plainly, $f \to L_f$ is a Borel function on $\overline{\mathcal{X}}$. If $L_f < \infty$ then f can be extended as a Lipschitz function to all of S with $K_f = L_f$. Conversely, if f is Lipschitz on S, its retraction to S_0 has $L_f = K_f$. Thus, the Lipschitz functions f on S can be identified as the functions f on S_0 with $L_f < \infty$, and $K_f = L_f$. This proves (i) and (ii).

For (iii), enumerate S_0 as $\{s_1, s_2, \dots\}$. Fix a positive integer n. Let $B_{n,1}$ be the set of points that are within $1/n$ of s_1. Let $B_{n,j+1}$ be the set of points that are within $1/n$ of s_{j+1}, but at a distance of $1/n$ or more from $s_1, \dots, s_j$. (In other words, take the balls of radius $1/n$ around the s_j and make them disjoint.) For each n, the $B_{n,j}$ are pairwise disjoint and

$$\bigcup_{j=1}^{\infty} B_{n,j} = S.$$

Given a mapping f of S into itself, let $f_n(s) = f(s_j)$ for $s \in B_{n,j}$. That is, f_n approximates f by $f(s_j)$ in the vicinity of s_j. The map $(f, s) \to f_n(s)$ is Borel from $\overline{\mathcal{X}} \times S$ to S. And on the set of Lipschitz f, this sequence of maps converges pointwise to the evaluation map. Q.E.D.

Remark. To make the connection with the setup of Section 1, if $\{f_\theta\}$ is a family of Lipschitz functions indexed by $\theta \in \Theta$, we require that the map $\theta \to f_\theta(x)$ be measurable for each $x \in S_0$. Then $\theta \to f_\theta$ is a measurable map from Θ to $\mathcal{X}$, and a measure on Θ induces a measure on $\mathcal{X}$. This section works directly with measures on $\mathcal{X}$.

The metric ρ induces a "Prokhorov metric" on probabilities, also denoted by ρ, as follows.

Definition 5.1. If P, Q are probabilities on S, then $\rho(P,Q)$ is the infimum of the $\delta > 0$ such that

$$P(C) < Q(C_\delta) + \delta \quad \text{and} \quad Q(C) < P(C_\delta) + \delta$$

for all compact $C \subset S$, where C_δ is the set of all points whose distance from C is less than δ.

Remarks.

(i) Plainly, $\rho(P,Q) \leq 1$.

(ii) Let ρ^* be as in Definition 5.1, with C ranging over all Borel sets. Plainly, $\rho^* < \delta$ entails $\rho \leq \delta$. That is, $\rho \leq \rho^*$. Conversely, suppose $\rho < \delta$. Fix a Borel set B and a small positive ε. Find a compact set $C \subset B$ with $P(B) < P(C) + \varepsilon$ and $Q(B) < Q(C) + \varepsilon$. Then

$$\begin{aligned} P(B) < P(C) + \varepsilon < Q(C_\delta) + \delta + \varepsilon \\ < Q(C_{\delta+\varepsilon}) + \delta + \varepsilon < Q(B_{\delta+\varepsilon}) + \delta + \varepsilon, \end{aligned}$$

and similarly for $Q(B)$. Thus, $\rho^* \leq \rho + \varepsilon$ and hence $\rho^* \leq \rho$. In short, $\rho^* = \rho$.

(iii) Dudley (1989) is a standard reference for results on the Prokhorov metric.

We need the definition of a random variable with an "algebraic tail". Basically, U has an algebraic tail if $\log(1 + U^+)$ has a Laplace transform in a neighborhood of 0, where $U^+ = \max\{0, U\}$ is the positive part of U. Of course, it is a matter of taste whether one uses $\log(1 + U^+)$ or $\log^+ U$.

Definition 5.2. A random variable U has an algebraic tail if there are positive, finite constants α, β such that $\text{Prob}\{U > u\} < \alpha/u^\beta$ for all $u > 0$. This condition has force only for large positive u; and we allow $\text{Prob}\{U = -\infty\} > 0$.

5.2. The Main Theorem. Fix a probability measure μ on $\mathcal{X}$. Assume that

$$f \to K_f \text{ has an algebraic tail relative to } \mu. \tag{5.1}$$

Fix a reference point $x_0 \in S$; assume too that

$$f \to \zeta(f) = \rho[f(x_0), x_0] \text{ has an algebraic tail relative to } \mu. \tag{5.2}$$

If, for instance, S is the line and the f's are linear, condition (5.1) constrains the slopes and then (5.2) constrains the intercepts. As will be seen later, any reference point in S may be used.

Consider a Markov chain moving around in S according to the following rule: starting from $x \in S$, the chain chooses $f \in \mathcal{X}$ at random from μ and goes to $f(x)$. We say that the chain "moves according to μ", or "μ is the move measure"; in Section 1, this Markov chain was called "the forward iteration".

Theorem 5.1. *Suppose μ is a probability on the Lipschitz functions. Suppose conditions (5.1) and (5.2) hold. Suppose further that*

$$\int_{\mathcal{X}} \log K_f \, \mu(df) < 0; \tag{5.3}$$

the integral may be $-\infty$. Consider a Markov chain on S that moves according to μ. Let $P_n(x, dy)$ be the law of the chain after n moves starting from x.

(i) *There is a unique invariant probability π.*

(ii) *There is a positive, finite constant A_x and an r with $0 < r < 1$ such that $\rho[P_n(x, \cdot), \pi] \leq A_x r^n$ for all $n = 1, 2, \ldots$ and $x \in S$.*

(iii) *The constant r does not depend on n or x; the constant A_x does not depend on n, and $A_x < a + b\rho(x, x_0)$ where $0 < a,\, b < \infty$.*

In (ii) and (iii), ρ is the Prokhorov metric (Definition 5.1). The argument for Theorem 5.1 can be sketched as follows. Although the forward process

$$X_n(x) = (f_n \circ f_{n-1} \circ \cdots \circ f_1)(x)$$

does not converge as $n \to \infty$, the backward process—with the composition in reverse order—does converge. Thus, we consider

$$Y_n(x) = (f_1 \circ f_2 \circ \cdots \circ f_n)(x). \tag{5.4}$$

The main step will be the following.

Proposition 5.1. *Assume (5.1–2–3). Define the backward process $\{Y_n(x)\}$ by (5.4). Then $Y_n(x)$ converges at a geometric rate as $n \to \infty$ to a random limit that does not depend on the starting point x.*

To realize the stationary process, let

$$\ldots, f_{-2}, f_{-1}, f_0, f_1, f_2, \ldots \tag{5.5}$$

be independent with common distribution μ, and let

$$W_m = f_m \circ f_{m-1} \circ f_{m-2} \circ \cdots, \tag{5.6}$$

where the composition "goes all the way". Rigor will come after some preliminary lemmas, and it will be seen that the process $\{W_m\}$ is stationary with the right transition law.

Lemma 5.2. *Let ξ_i be i.i.d random variables; $P\{\xi_i = -\infty\} > 0$ is allowed. Suppose there are positive, finite constants α, β such that $P\{\xi_i > v\} < \alpha e^{-\beta v}$ for all $v > 0$. Let ξ be distributed as ξ_i. Then*

(i) *$-\infty \leq E\{\xi\} < \infty$.*

(ii) *If c is a finite real number with $c > E\{\xi\}$, there are positive, finite constants A and r such that $r < 1$ and $P\{\xi_1 + \cdots + \xi_n > nc\} < Ar^n$ for all $n = 1, 2, \ldots$. The constants A and r depend on c and the law of ξ, not on n.*

Proof. *Case 1.* Suppose ξ is bounded below. Then (i) is immediate, with $-\infty < m < \infty$; (ii) is nearly standard, but we give the argument anyway. First, $E\{\exp(\lambda\xi)\} < \infty$ for $-\infty < \lambda < \beta$. Next, let $m = E\{\xi\}$. We claim that

$$E\{e^{\lambda\xi}\} = 1 + \lambda m + O(\lambda^2) \quad \text{as} \quad \lambda \to 0. \tag{5.7}$$

Indeed, fix γ with $0 < \gamma < \beta$; let $|t| < 1$ and $\lambda = t\gamma$. Then $|\lambda| < \gamma$, so

$$\frac{\gamma^2}{\lambda^2}|e^{\lambda\xi} - 1 - \lambda\xi| \leq e^{\gamma|\xi|} - 1 - \gamma|\xi|. \tag{5.8}$$

The right hand side of (5.8) has finite expected value, proving (5.7). As a result, there are positive constants λ_0 and d for which

$$E\{e^{\lambda\xi}\} \leq 1 + m\lambda + d\lambda^2 \leq e^{m\lambda + d\lambda^2}$$

provided $0 \leq \lambda \leq \lambda_0$. Let

$$r_{\lambda,c} = e^{-\lambda c}E(e^{\lambda\xi}).$$

By Markov's inequality,

$$P\{\xi_1 + \cdots + \xi_n > nc\} < r_{\lambda,c}^n. \tag{5.9}$$

If $0 \leq \lambda \leq \lambda_0$, we have a bound on $r_{\lambda,c}$. Set $\lambda = (c-m)/2d$ to complete the proof in Case 1, with $r = \exp[-(c-m)^2/4d]$. This is legitimate provided $m \leq c \leq c_0 = m + 2d\lambda_0$. Larger values of c may be replaced by c_0.

Case 2. Let ξ_i' be ξ_i truncated below at a constant that does not depend on i. Then $\sum_i \xi_i \leq \sum_i \xi_i'$. Case 1 applies to the truncated variables, whose mean will be less than c if the truncation point is sufficiently negative. Our idea of truncation can be defined by example: x truncated below at -17 equals x if $x \geq -17$, and -17 if $x \leq -17$. Q.E.D.

Let f_n be an i.i.d. sequence of picks from μ. Fix $x \in S$. Consider the forward process starting from x:

$$X_0(x) = x,\ X_1(x) = f_1(x),\ X_2(x) = (f_2 \circ f_1)(x),\ \ldots.$$

Lemma 5.3. $\rho[X_n(x), X_n(y)] \leq \left[\prod_{j=1}^n K_{f_j}\right]\rho(x,y)$.

Proof. This is obvious for $n = 0$ and $n = 1$. Now

$$\rho\big[f_{n+1}\big(X_n(x)\big), f_{n+1}\big(X_n(y)\big)\big] \leq K_{f_{n+1}}\rho[X_n(x), X_n(y)]. \qquad \text{Q.E.D.}$$

The next two lemmas will prove the uniqueness part of Theorem 5.1.

Lemma 5.4. *Suppose (5.1) and (5.3). If $\varepsilon > 0$ is sufficiently small, there are positive, finite constants A and r with $r < 1$ and*

$$P\Big\{\sum_{i=1}^{n} \log K_{f_i} > -n\varepsilon\Big\} < Ar^n$$

for all $n = 1, 2, \ldots$. The constants A and r depend on ε but not on n.

Proof. Apply Lemma 5.2 to the random variables $\xi_i = \log K_{f_i}$. Q.E.D.

Lemma 5.5. *Suppose (5.1) and (5.3). For sufficiently small positive ε: except for a set of $f_1, \ldots, f_n$ of probability less than Ar^n, $\rho[X_n(x), X_n(y)] \leq \exp(-n\varepsilon)\rho(x, y)$ for all $x, y \in S$. Again, A and r depend on ε but not on n.*

Proof. Use Lemmas 5.3 and 5.4. Q.E.D.

Corollary 5.1. *There is at most one invariant probability.*

Proof. Suppose π and π' were invariant. Choose x from π and x' from π', independently. Let $Y_n = X_n(x)$ and $Y_n' = X_n(x')$. Now $\rho(Y_n, Y_n') \leq \exp(-n\varepsilon)\rho(Y_0, Y_0')$ except for a set of exponentially small probability. So, the laws of Y_n and Y_n' merge; but the former is π and the latter is π'. Q.E.D.

The next lemma gives some results on variables with algebraic tails, leading to a proof that if (5.1) holds, and (5.2) holds for some particular x_0, then (5.2) holds for all $x_0 \in S$. The lemma and its corollary are only to assist the interpretation.

Lemma 5.6.

(i) *If U is non-negative and bounded above, then U has an algebraic tail.*
(ii) *If U has an algebraic tail and $c > 0$, then cU has an algebraic tail.*
(iii) *If U and V have algebraic tails, so does $U + V$; these random variables may be dependent. (In principle, there are two α's and two β's; it is convenient to use the larger α and the smaller β, if both of the latter are positive.)*

Proof. Claims (i) and (ii) are obvious. For claim (iii),

$$P\{U + V > t\} \leq P\{U > t/2\} + P\{V > t/2\}. \qquad \text{Q.E.D.}$$

Corollary 5.2. *Suppose condition (5.1) holds. If (5.2) holds for any particular $x_0 \in S$, then (5.2) holds for any $x_0 \in S$. In other words, there are finite positive constants α, β with $\mu\{\, f : \rho[f(x_0), x_0] > u \,\} < \alpha/u^\beta$ for all $u > 0$. The constant α may depend on x_0, but the shape parameter β does not.*

Proof. Use Lemma 5.6 and the triangle inequality. Q.E.D.

Lemma 5.7. *Let f and g be mappings of S into itself; let $x \in S$. Then*

$$\rho[(f \circ g)(x), x] \leq \rho[f(x), x] + K_f \rho[g(x), x].$$

Proof. By the triangle inequality,

$$\rho[(f \circ g)(x), x] \leq \rho[f(x), x] + \rho[(f \circ g)x, f(x)].$$

Now use the definition of K_f. Q.E.D.

Corollary 5.3. *Let $\{g_i\}$ be mappings of S into itself; let $x \in S$. Then*

$$\begin{aligned}\rho[(g_1 \circ g_2 \circ \cdots \circ g_m)(x), x] \leq & \rho[g_1(x), x] \\ & + K_{g_1}\rho[g_2(x), x] \\ & \quad + K_{g_1}K_{g_2}\rho[g_3(x), x] + \cdots \\ & \quad\quad + K_{g_1}K_{g_2}\cdots K_{g_{m-1}}\rho[g_m(x), x].\end{aligned}$$

Proof of Proposition 5.1. We assume conditions (5.1–3) and consider the behavior when $n \to \infty$ of the backward iterations $Y_n(x) = (f_1 \circ f_2 \circ \cdots \circ f_n)(x)$. Convergence of $Y_n(x)$ as $n \to \infty$ will follow from the Cauchy criterion. In view of Lemma 5.5, it is enough to consider $x = x_0$. As in Lemma 5.3,

$$\rho[Y_{n+m}(x), Y_n(x)] \leq K_{f_1}\cdots K_{f_n}\rho[(f_{n+1} \circ f_{n+2} \circ \cdots \circ f_{n+m})(x), x]. \tag{5.10}$$

We use Corollary 5.3 with f_{n+i} for g_i to bound the right hand side of (5.10), concluding that

$$\rho[Y_{n+m}(x), Y_n(x)] \leq \sum_{i=0}^{\infty}\left(\prod_{j=1}^{n+i} K_{f_j}\right)\rho[f_{n+i+1}(x), x]. \tag{5.11}$$

By Lemma 5.4, except for a set of probability $A'r^{n_0}$,

$$\prod_{j=1}^{n+i} K_{f_j} \leq e^{-(n+i)\varepsilon} \tag{5.12}$$

for all $n \geq n_0$ and all $i = 0, 1, \ldots$.

Next, condition (5.2) comes into play. Write $\zeta_j = \rho[f_j(x), x]$. By the Definition 5.2 of algebraic tails, there are positive finite constants α and β such that $P\{\zeta_j > s^j\} < \alpha/s^{\beta j}$. Choose $s > 1$ but so close to 1 that $se^{-\varepsilon} < 1$. Except for another set of exponentially small probability,

$$\zeta_{n+i+1} \leq s^{n+i+1} \tag{5.13}$$

for all $n \geq n_0$ and all $i = 0, 1, \ldots$. Now there are finite positive constants c_0, r_0, r_1, with $r_0 < 1$ and $r_1 < 1$, such that for all n_0, for all $n \geq n_0$, and all $m = 0, 1, \ldots,$

$$\rho[Y_{n+m}(x), Y_n(x)] \leq r_1^n, \tag{5.14}$$

except for a set of probability $c_0 r_0^{n_0}$. Thus, $Y_n(x)$ is Cauchy, and hence converges to a limit in S. We have already established that the limit does not depend on x; call the limit Y_∞. An exponential rate for the convergence of $Y_n(x)$ to Y_∞ follows by letting $m \to \infty$ in (5.14). Q.E.D.

Lemma 5.8. *Let X, X' be random mappings into S, with distributions λ, λ'. Suppose X, X' can be realized so that $P\{\rho(X, X') \geq \delta\} < \delta$. Then $\rho(\lambda, \lambda') \leq \delta$. (In the first instance, ρ is the metric on S; in the second, ρ is the induced Prokhorov metric on probabilities: see Definition 5.1.)*

Proof. Let C be a compact subset of S. Then $X \in C$ entails $X' \in C_\delta$, except for probability δ. Likewise, $X' \in C$ entails $X \in C_\delta$, except for probability δ. Q.E.D.

Remark. The converse to Lemma 5.8 is true too: one proof goes by discretization and the "Marriage Lemma". See Strassen (1965) or Dudley (1988, Chapter 11).

Proof of Theorem 5.1. There are only a few details to clean up. Recall the doubly-infinite sequence $\{f_i\}$ from (5.5). By Proposition 5.1, we can define W_m as follows:

$$W_m = \lim_{n \to \infty} (f_m \circ f_{m-1} \circ \cdots \circ f_{m-n})(x). \tag{5.15}$$

The limit does not depend on x. Proposition 5.1 applies, because—as before—

$$\mathcal{L}(f_m, f_{m-1}, \ldots) = \mathcal{L}(f_1, f_2, \ldots).$$

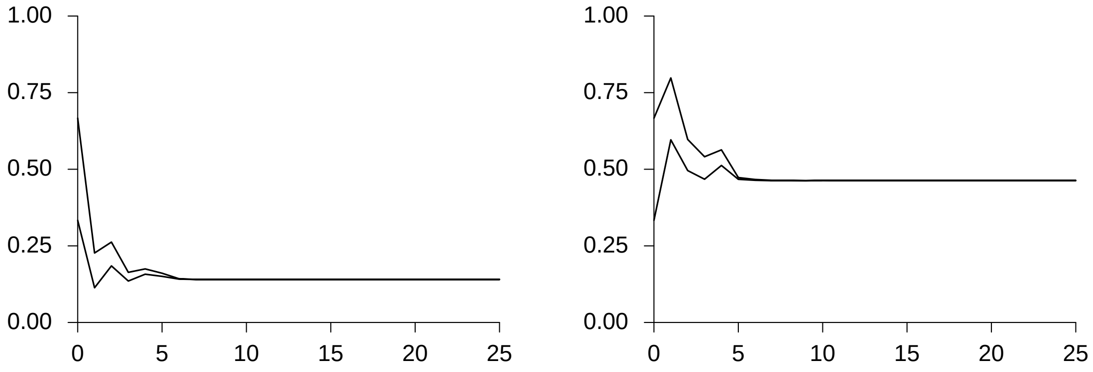
It is easy to verify that

$$W_m : m = \ldots,\ -2,\ -1,\ 0,\ 1,\ 2,\ \ldots \tag{5.16}$$

is stationary with the right transition probabilities. And Y_∞ is distributed like any of the W_m. Thus, the convergence assertion (ii) in Theorem 5.1 follows from Lemma 5.8 and Proposition 5.1. The argument is complete.

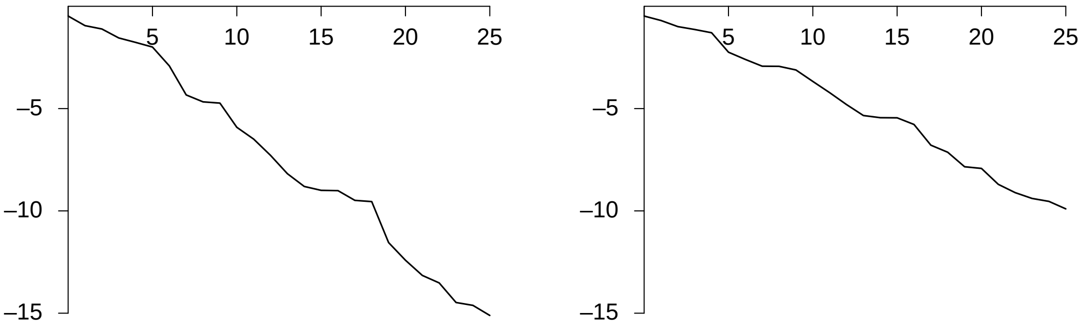
Proof of Theorem 1 and Proposition 1. These results are immediate from Proposition 5.1 and Theorem 5.1. Indeed, the moment conditions in Theorem 1 imply conditions (5.1–2–3); we stated Theorem 1 using the more restrictive conditions in order to postpone technicalities.

Figure 4. The backward iterations converge rapidly to a limit that is random but does not depend the starting state.



The essence of the thing is that the backward iterations converge at a geometric rate to a limit that depends on the functions being composed—but not on the starting point. Figure 4 illustrates the idea for the Markov chain discussed in Section 2.1. The left hand panel shows the backward iteration starting from $x_0 = 1/3$ or $x_0 = 2/3$. Exactly the same functions are used to generate the two paths; the only difference is the starting point. (Position at time n is plotted against $n = 0, 1, \dots, 25$, with linear interpolation.) The paths merge for all practical purposes around $n = 7$. The right hand panel shows the same thing, with a new lot of random functions. Convergence is even faster, but the limit is different—randomness in action. (By contrast, the forward iteration does not converge, but wanders around ergodically in the state space: Figure 1.) Figure 5 plots the logarithm (base 10) of the absolute difference between the paths in the corresponding panels of Figure 4. The linear decay on the log scale corresponds to exponential decay on the original scale. The difference in slopes between the two panels is due to the randomness in choice of functions; this difference wears off as the number of iterations goes up.

Figure 5. Logarithm to base 10 of the absolute difference between paths in the backward iteration.



Remarks.

(i) The notation in (5.15–16) may be a bit confusing:

$$\{W_n : n = 0, -1, -2, \dots\}$$

is not the backward process, and does not converge.

(ii) We use the algebraic tail condition to bound the probabilities of the exceptional sets in Proposition 5.1, that is, the sets where (5.12) and (5.13) fail. These probability bounds give the exponential rate of convergence in Theorem 5.1. With a little more effort, the optimal r can be computed explicitly, in terms of the mean and variance of $\log K_f$, and the shape parameter β in (5.2). If an exponential rate is not needed, it is enough to assume that $\log(1 + K_f)$ and $\log\big(1 + \rho[f(x_0), x_0]\big)$ are L_1.

(iii) Furstenberg (1963) uses the backward iteration to study products of random matrices. He considers the action of a matrix group on projective space and shows that there is a unique stationary distribution, which can be represented as a convergent backward iteration. Convergence is proved by martingale arguments. It seems worthwhile to study the domain of this method.

(iv) Let $(\mathcal{X}, \mathcal{B})$ be a measurable space and let $K(x, dy)$ be a Markov kernel on $(\mathcal{X}, \mathcal{B})$. When is there a family $\{f_\theta : \theta \in \Theta\}$ and a probability μ on Θ such that the Markov chain induced by these iterated random mappings has transitions $K(x, dy)$? This construction is always possible if $(\mathcal{X}, \mathcal{B})$ is "Polish", that is, a Borel subset of a complete separable metric space. See, for instance, Kifer (1986). The leading special case has $\mathcal{X} = [0, 1]$. Then Θ can also be taken as the unit interval, and μ as Lebesgue measure; $K(x, dy)$ can be described by its distribution function $F(x, y) = K(x, [0, y])$. Let $G(x, \cdot)$ be the inverse of $F(x, \cdot)$. If U is uniform, $G(x, U)$ is distributed as $K(x, dy)$. Finally, let $f_\theta(x) = G(x, \theta)$. Verification is routine, and the general case follows from the special case by standard tricks.

The question is more subtle—and the regularity conditions much more technical—if it is required that the $f_\theta(\cdot)$ be continuous. Blumenthal and Corson (1970) show that if $\mathcal{X}$ is a connected, locally connected, compact space, and $x \to K(x, \cdot)$ is continuous (weak star), and the support of $K(x, \cdot)$ is $\mathcal{X}$ for all x, then there is a probability measure on the Borel sets of the continuous functions from $\mathcal{X}$ to $\mathcal{X}$ which induces the kernel K. Quas (1991) gives sufficient conditions for representation by smooth functions when $\mathcal{X}$ is a smooth manifold. A survey of these and related results appears in Dubischar (1997).

6. More examples.

Autoregressive processes are an important feature of many statistical models, and can usefully be viewed as iterated random functions; the construction will be sketched here. We learned the trick from Anderson (1959), but he attributes it to Yule. Further examples and counterexamples to illustrate the theory are given in Section 6.2; Section 6.3 revisits the example discussed in Section 2.1.

6.1. Autoregressive processes.

Let $S = \mathbb{R}$, the real line. Let a be a real number with $0 < a < 1$ and let μ be a probability measure on $\mathbb{R}$. For present purposes, an autoregression is a Markov process on $\mathbb{R}$ with the following law of motion: starting from $x \in \mathbb{R}$, the chain picks ξ according to μ and moves to $ax + \xi$. Conditions (5.1) and (5.3) are obvious: if $f(x) = ax + \xi$, then $K_f = a$. For condition (5.2), we need to assume for instance that if ξ has distribution μ, there are positive, finite constants α, β with $P(|\xi| > u) < \alpha/u^\beta$ for all $u > 0$. If ξ_i are independent with common distribution μ, the forward process starting from x has $X_0(x) = x$,

$$X_1(x) = ax + \xi_1,\ X_2(x) = a^2x + a\xi_1 + \xi_2,\ X_3(x) = a^3x + a^2\xi_1 + a\xi_2 + \xi_3,$$

and so forth. This process converges in law, but does not converge almost surely: at stage n, new randomness is introduced by ξ_n. The backward process starting from x looks at first glance much the same: $Y_0(x) = x$,

$$Y_1(x) = ax + \xi_1,\ Y_2(x) = a^2x + \xi_1 + a\xi_2,\ Y_3(x) = a^3x + \xi_1 + a\xi_2 + a^2\xi_3,$$

and so forth. But this process converges a.s., because the new randomness introduced by ξ_n is damped by a^n. The stationary autoregressive process may be realized as

$$W_m = \xi_m + a\xi_{m-1} + a^2\xi_{m-2} + a^3\xi_{m-3} + \cdots .$$

Each W_m is obtained by doing the backward iteration on $\{\xi_m, \xi_{m-1}, \xi_{m-2}, \dots\}$. Equation (5.6) is the generalization. With the usual Euclidean distance, the constant A_x in Theorem 5.1 must depend on the starting state x. For a particularly brutal illustration, take $\xi_i \equiv 0$.

6.2. Without regularity conditions. This section gives some examples to indicate what can happen without our regularity conditions.

Example 6.1. This example shows that some sort of contracting property is needed to get a result like Theorem 5.1. Let $S = [0, 1]$. Arithmetic is to be done modulo 1: for instance, $2 \times .71 = .42$. Let

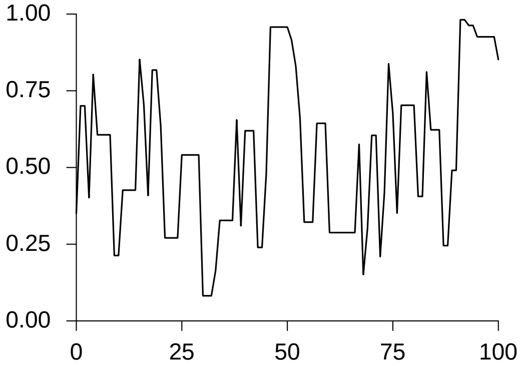
$$f(x) = x, \qquad g(x) = 2x \mod 1,$$

and $\mu\{f\} = \mu\{g\} = 1/2$. The forward and the backward process can both be represented as

$$X_n = 2^{\xi_1 + \cdots + \xi_n} x \mod 1,$$

the ξ_n being independent and taking values 0 or 1 with probability 1/2 each; x is the starting point. Clearly, the backward process converges only if the starting point is a binary rational. Furthermore, there are infinitely many distinct stationary probabilities: if $\zeta_1, \zeta_2, \dots$ is a stationary 0–1 valued process, then the law of $\sum_i \zeta_i/2^i$ is stationary for our chain. Since $K_f = 1$ and $K_g = 2$, condition (5.3) fails. Figure 6 plots X_n against n for $n = 0, \dots, 100$, with linear interpolation.

Figure 6. Iterated random functions on the unit interval. With probability 1/2, the chain stands pat; with probability 1/2, the chain moves from x to $2x$ modulo 1. The forward and backward process are the same, and do not converge.



Remark. Figure 6 involves on the order of 50 doublings, so numerical accuracy is needed to 50 binary digits, or 16 decimal places in x. That is about the limit of double-precision computer packages like MATLAB a PC. If, say, 1,000 iterations are wanted, accuracy to 150 decimal places would be needed. The work-around is easy. Code the states x as long strings of 0's or 1's, and do binary arithmetic. For plotting, convert to decimals: only the first 10 bits in X_n will matter.

Example 6.2. This example has a unique stationary distribution but the backward process does not converge. Let S be the integers mod N. Let

$$f(j) = j, \qquad g(j) = j + 1 \mod N,$$

with $\mu\{f\} = \mu\{g\} = 1/2$. The forward and the backward process can both be represented as

$$X_n = \xi_1 + \cdots + \xi_n + x \mod N,$$

the ξ_n being independent and taking values 0 or 1 with probability 1/2 each. Clearly, the backward process does not converge. On the other hand, the chain is aperiodic and irreducible, so there is a unique stationary distribution (the uniform), and there is an exponential rate of convergence. Let $\rho(i, j)$ be the least $k = 0, 1, \ldots$ such that $i + k = j$ or $j + k = i$. Then ρ is a metric: the distance between two points is the minimal number of steps it takes to get from one to the other, where steps can be taken in either direction. Relative to this metric, f and g are Lipschitz, with $K_f = K_g = 1$; condition (5.3) is violated.

The next example shows another sort of pathology when condition (5.3) holds but (5.1–2) fail.

Example 6.3. The state space S is $[0, \infty)$. Let the random variable ξ have a symmetric stable distribution with index $\alpha > 1$; see Samorodnitsky and Taqqu (1994) or Zolotarev (1986). Let μ be the law of $e^{\xi-1}$. Consider a Markov chain that moves from $x \in [0, \infty)$ by choosing K at random from μ and going to Kx. Then 0 is a fixed point and the unique stationary distribution concentrates at 0. If ξ_i are i.i.d. symmetric stable with index α, the forward and the backward process process can both be represented as

$$X_n = e^{\xi_1 + \cdots + \xi_n - n} x.$$

$X_n \to 0$ almost surely as $n \to \infty$, by the strong law of large numbers. On the other hand, the Prokhorov distance between $\mathcal{L}(X_n)$ and δ_0 is of order $1/n^{\alpha-1}$, by Lemmas 6.1 and 6.2 below. In particular, exponential rates of convergence do not obtain. Condition (5.3) holds: $\int \log K \, d\mu = -1$. However, (5.1) fails, and so does (5.2) for $x_0 \neq 0$.

Lemma 6.1. *Let δ_0 be point mass at 0, and let Φ be a continuous probability measure on $(0, \infty)$.*

(i) *There is a unique ε_0 with $0 < \varepsilon_0 < 1$ and $\Phi(\varepsilon_0, \infty) = \varepsilon_0$.*
(ii) *$\Phi(\varepsilon, \infty) < \varepsilon$ for $\varepsilon > \varepsilon_0$.*
(iii) *$\rho(\delta_0, \Phi) = \varepsilon_0$.*

Proof. Claims (i) and (ii) are easy to verify. For (iii), we need to compute the infimum of ε such that for all compact C,

$$\delta_0(C) < \Phi(C_\varepsilon) + \varepsilon \tag{6.1}$$

and

$$\Phi(C) < \delta_0(C_\varepsilon) + \varepsilon. \tag{6.2}$$

If $0 \notin C$, then (6.1) is vacuous. If $0 \in C$, then (6.1) is equivalent to $1 - \varepsilon < \Phi(C_\varepsilon)$. Furthermore, $0 \in C$ entails $[0, \varepsilon) \subset C_\varepsilon$. And $C_\varepsilon = [0, \varepsilon)$ when $C = \{0\}$. Thus, (6.1) for all compact C is equivalent to

$$\Phi(0, \varepsilon) > 1 - \varepsilon. \tag{6.3}$$

Likewise, if $0 \in C_\varepsilon$, then (6.2) is vacuous. If $0 \notin C_\varepsilon$ then (6.2) is equivalent to $\Phi(C) < \varepsilon$. But $0 \notin C_\varepsilon$ iff $C \subset [\varepsilon, \infty)$. Thus, (6.2) for all compact C is also equivalent to (6.3). Now (iii) follows from (ii). Q.E.D.

Lemma 6.2. *Let U be a symmetric stable random varible with index $\alpha > 1$. Let n be a large positive integer. The Prokhorov distance between δ_0 and the law of* $\exp(-n + n^{1/\alpha}U)$ *is of order $1/n^{\alpha-1}$.*

Proof. This follows from Lemma 6.1, since $P\{U > u\} \sim 1/u^\alpha$. Q.E.D.

Remark. Something can be done even when all the Lipschitz constants are 1, provided the functions are genuinely contracting on a recurrent set. For instance, Steinsaltz (1997, 1998) considers a Markov chain on $\mathbb{R}$ that moves by choosing one of the following two functions at random:

$$f_+(x) = \begin{cases} x+1 & \text{if } x \geq 0 \\ \frac{1}{2}x + 1 & \text{if } -2 \leq x \leq 0 \\ x+2 & \text{if } x \leq -2 \end{cases} \qquad f_-(x) = \begin{cases} x-1 & \text{if } x \leq 0 \\ \frac{1}{2}x - 1 & \text{if } 0 \leq x \leq 2 \\ x-2 & \text{if } x \geq 2. \end{cases}$$

These functions have Lipschitz constant 1. But, as a team, they are genuinely contracting on the interval $[-2, 2]$. This interval is recurrent. Indeed, from large negative x, the chain moves 2 units to the right and 1 unit to the left with equal probabilities; the reverse holds for large positive x. Thus, when the chain is near $\pm\infty$, it drifts back toward 0. Steinsaltz has some general theory, and other examples.

6.3. The Beta walk. The state space S is the closed unit interval [0,1]. Let Φ be a probability measure on S, and let $0 < p < 1$. Consider a chain with the following transition probabilities. Starting from $x \in [0, 1]$, the chain goes left with probability p and right with probability $1 - p$. To move, it picks u from Φ. If the move is to the left, the chain goes to ux; if to the right, it goes to $x + u(1 - x) = x + u - ux$. Call Φ the "moving measure". If Φ is Beta(α, α), call the chain a "Beta walk". The example in Section 2.1 was a Beta walk with $p = 1/2$ and $\alpha = 1/2$. We extend the terminology a little: Beta$(0,0)$ puts mass 1/2 at 0 and 1; Beta(∞,∞) puts mass 1 at 1/2.

These examples fit into the framework of Theorem 5.1: p and Φ probabilize the set of linear maps that shrink the unit interval—

- either toward 0, when the map sends x to ux,
- or toward 1, when the map sends x to $x + u - ux$.

All the Lipschitz constants are 1 or smaller. Conditions (5.1–2–3) are obvious, and there is exponential convergence to the unique stationary distribution. In the balance of this section, we prove the following theorem.

Theorem 6.1. *Suppose $S = [0,1]$, $p = 1/2$, and the move measure Φ is Beta(α,α). Let $\alpha' = \alpha/(\alpha+1)$; when $\alpha = \infty$, let $\alpha' = 1$. If α is 0, 1, or ∞, then the stationary distribution of the Beta walk is Beta(α',α'). For any other value of α, the stationary distribution is symmetric and has the same first three moments as Beta(α',α') but a different fourth moment: in particular, the stationary distribution is not Beta.*

Remarks. The second moment of Beta(a,a) is $(a+1)/(4a+2)$, which determines a; that is why agreement on 3 moments and discrepancy on the 4th shows the stationary measure not to be Beta. As will be seen, the discrepancy is remarkably small—on the order of 10^{-4} when $\alpha = 1/3$, and that is about as big as it gets.

The proof of the next lemma is omitted. The first term in the integral corresponds to a leftward move, taken with probability p; the second, to a rightward move; compare (2.1).

Lemma 6.3. *If the move measure Φ has density ϕ, and the starting state is chosen from a density ψ, the density of the position after one move is*

$$(T\psi)(y) = p\int_y^1 \frac{1}{x}\phi\Big(\frac{y}{x}\Big)\psi(x)\,dx + \overline{p}\int_0^y \frac{1}{1-x}\;\phi\Big(\frac{y-x}{1-x}\Big)\psi(x)\,dx.$$

The next result too is standard. Suppose X is Beta(a,b). Then

$$E\{X^n\} = \frac{\Gamma(a+n)}{\Gamma(a)}\;\frac{\Gamma(a+b)}{\Gamma(a+b+n)} = \frac{(a+n-1)\cdots(a+1)a}{(a+b+n-1)\cdots(a+b+1)(a+b)}\;.$$

The second equality follows from the recursion $\Gamma(x+1) = x\Gamma(x)$: there are n factors in the numerator and in the denominator.

Corollary 6.1. *If X is Beta(a,a), then*

$$E(X) = \frac{1}{2},\;\; E(X^2) = \frac{a+1}{4a+2},\;\; E(X^3) = \frac{a+2}{8a+4},\;\; E(X^4) = \frac{a+3}{2a+3}\,\frac{a+2}{8a+4}.$$

The proof of Theorem 6.1.

Case 1. Suppose $\alpha = 0$, so the move measure Φ puts mass 1/2 each at 0 and 1; this is the stationary measure too, with stationarity being achieved in one move. Since $\alpha' = 0$, the theorem holds.

Case 2. Suppose $\alpha = \infty$, so the move measure Φ concentrates on 1/2. Starting from x, the chain moves to $\frac{1}{2}x$ or $x + \frac{1}{2} - \frac{1}{2}x = \frac{1}{2} + \frac{1}{2}(1-x)$ with a 50–50 chance. Clearly, the uniform distribution is invariant, its image under the motion having mass $\frac{1}{2}$ uniformly distributed over $[0, \frac{1}{2}]$, and mass $\frac{1}{2}$ uniform on $[\frac{1}{2}, 1]$. Since $\alpha' = 1$ and Beta$(1,1)$ is uniform, the theorem holds.

Case 3. This was discussed in Section 2.1.

Case 4. Suppose the move measure Φ is Beta(α, α) with $0 < \alpha < 1$ or $1 < \alpha < \infty$. Recall that $\alpha' = \alpha/(\alpha+1)$; and let $U' \sim$ Beta(α', α'). By Corollary 6.1 and some tedious algebra,

$$E(U') = \frac{1}{2},\ E(U'^2) = \frac{2\alpha+1}{6\alpha+2},\ E(U'^3) = \frac{3\alpha+2}{12\alpha+4},\ E(U'^4) = \frac{4\alpha+3}{5\alpha+3}\,\frac{3\alpha+2}{12\alpha+4}.$$

We must now compute the first 4 moments of the stationary distribution; the latter exists by Theorem 5.1. Let U have the stationary distribution and let $V \sim$ Beta(α, α); make these two random variables be independent. As before, write $\mathcal{L}(Z)$ for the law of Z. Then

$$\mathcal{L}(U) = \frac{1}{2}\mathcal{L}(UV) + \frac{1}{2}\mathcal{L}(U+V-UV) = \frac{1}{2}\mathcal{L}(UV) + \frac{1}{2}\mathcal{L}(1-UV), \tag{6.4}$$

because $U + V - UV = 1 - (1-U)(1-V)$ and U, V are symmetric. In particular,

$$E(U^n) = \frac{1}{2}E(U^n)E(V^n) + \frac{1}{2}E[(1-UV)^n]. \tag{6.5}$$

$E(V^n)$ is given by Corollary 6.1, so equation (6.5) can be solved recursively for the moments of U, and $E(U^n) = E(U'^n)$ for $n = 1, 2, 3$. However,

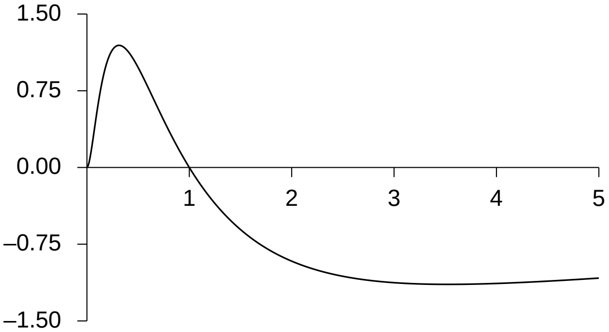
$$E(U^4) = \frac{1}{6}\,\frac{(2\alpha+3)(9\alpha^2+10\alpha+2)}{(3\alpha+1)(5\alpha^2+9\alpha+2)}.$$

Consequently,

$$E(U^4) - E(U'^4) = \frac{(1-\alpha)\alpha^2}{12(3\alpha+1)(5\alpha+3)(5\alpha^2+9\alpha+2)}. \tag{6.6}$$

(Again, unpleasant algebraic details are suppressed.) Figure 7 shows the graph of the right side of (6.6), plotted against α. As will be seen, the discrepancy is rather small.

Figure 7. Difference between 4th moment of stationary distribution and 4th moment of approximating Beta, scaled by 10^4 and plotted against α; symmetric chain, Beta(α, α) move distribution.



Remark. Theorem 6.1 is connected to results in Dubins and Freedman (1967). Consider generating a random distribution function by constructing its graph in the unit square. Draw a horizontal line through the square, cutting the vertical axis into a lower segment and an upper segment whose lengths stand in the ratio p to $1-p$. Pick a point at random on this line. That divides the square into four rectangles. Now repeat the construction in the lower left and upper right rectangles. (The description may be cumbersome, but the inductive step is easy.) The limiting monotone curve connecting all the chosen points is the graph of a random distribution function. The average of these distribution functions turns out to be absolutely continuous: let ϕ be its density. This density is, by construction, invariant under the following operation. Choose x at random uniformly on $[0,1]$; distribute mass p according to ϕ rescaled over $[0,x]$ and mass $1-p$ according to ϕ rescaled over $[x,1]$. If U is uniform and $X \sim \phi$, then

$$\mathcal{L}(X) = p\mathcal{L}(UX) + (1-p)\mathcal{L}(U + X - UX).$$

In short, ϕ is the stationary density for our Markov process. The equation in Lemma 6.3 is discussed in Section 9 of Dubins and Freedman (1967).

7. Dirichlet distributions. The Dirichlet distribution is the multidimensional analog of the more familiar Beta, and is often used in Bayesian nonparametric statistics. An early paper is Freedman (1963); also see Fabius (1964) or Ferguson (1973). Sections 7.1–2 sketch a construction of the Dirichlet. The setting is an infinite dimensional space, namely, the space of all probability measures on an underlying complete separable metric space. Section 7.3 discusses the law of the mean of F picked at random from a Dirichlet distribution, which can sometimes be computed in closed form. The setting is the real line.

7.1. Random measures. Let $(\mathcal{X}, \rho)$ be a complete separable metric space, for instance, the real line. Let $\mathcal{P}$ be the set of all probability measures on $\mathcal{X}$; p and q will be typical elements of $\mathcal{P}$, that is, typical probabilities on $\mathcal{X}$. We will be considering random probabilities P on $\mathcal{X}$: these are random objects with values in $\mathcal{P}$. The "law" of such an object is a probability on $\mathcal{P}$. Let α be a finite measure on $\mathcal{X}$. The "Dirichlet with base measure α", usually abbreviated as D_α, is the law of a certain random probability on $\mathcal{X}$. Thus, D_α is a probability on $\mathcal{P}$.

Here, we show how to construct D_α by modifying the argument for Theorem 5.1. The state space S for the Markov chain is $\mathcal{P}$. The variation distance between p and q is defined as

$$\|p - q\| = \sup_B |p(B) - q(B)|,$$

where B runs over all the Borel subsets of $\mathcal{X}$. The "parameter space" for the Lipschitz functions will be $\Theta = [0, 1] \times \mathcal{P}$. If $0 \leq u \leq 1$ and $p \in \mathcal{P}$, let $f_{u,p}$ map $\mathcal{P}$ into $\mathcal{P}$ by the rule

$$f_{u,p}(q) = uq + (1 - u)p.$$

It is easy to see that $f_{u,p}$ is an affine map of $\mathcal{P}$ into itself. Furthermore, this function is Lipschitz, with Lipschitz constant $K_{u,p} = u$.

If μ is any probability measure on the parameter space Θ, the Markov chain on $\mathcal{P}$ driven by μ has a unique stationary distribution. The Dirichlet will be obtained by specializing μ. Caution: the stationary distribution is a probability on $\mathcal{P}$, that is, a probability on the probabilities on $\mathcal{X}$; and there is a regularity condition, namely,

$$\mu\{(u, p) : u < 1\} > 0. \tag{7.1}$$

Recall that $\mathcal{L}$ stands for law. Then Q has the stationary distribution if

$$\mathcal{L}\big(UQ + (1 - U)P\big) = \mathcal{L}(Q), \tag{7.2}$$

where $\mathcal{L}(U, P) = \mu$ independent of Q. The stationary distribution may be represented by the backward iteration, as the law of the random probability

$$S_\infty = (1 - U_1)P_1 + U_1(1 - U_2)P_2 + U_1U_2(1 - U_3)P_3 + \cdots . \tag{7.3}$$

In (7.3), the (U_n, P_n) are independent, with common distribution μ; as will be seen in a moment, the sum converges almost surely. The limit is a random probability on $\mathcal{X}$ because each P_n is a random probability on $\mathcal{X}$, and the U_n are random elements of $[0, 1]$. Furthermore,

$$(1 - U_1) + U_1(1 - U_2) + U_1U_2(1 - U_3) + \cdots \tag{7.4}$$

telescopes to 1.

In variation distance, $\mathcal{P}$ is complete but not separable. Thus, Theorem 5.1 does not apply. Rather than deal with the measure-theoretic technicalities created by an inseparable space, we sketch a direct argument for convergence. First, we have to prove that the sum in (7.4) converges almost surely. Indeed, write T_n for the nth term. Then $E\{T_n\} = (1 - \phi)\phi^{n-1}$, where

$$\phi = E\{U_n\} < 1 \tag{7.5}$$

by (7.1). Thus $P\{T_n > \sqrt{\phi^{n-1}}\} < \sqrt{\phi^{n-1}}$, and $\sum_n \sqrt{\phi^{n-1}} < \infty$. An immediate consequence: with probability 1, the sum on the right in (7.3) is Cauchy and hence converges in variation norm (completeness). The law of S_∞ is easily seen to be stationary, using the criterion (7.2).

To get a geometric rate of convergence, suppose the chain starts from q. Let S_n be the sum of the first n terms in (7.3). After n moves starting from q, the backward process will be at $S_n + R_n$, where $R_n = U_1 U_2 \cdots U_n q$. By previous arguments, except for a set of geometrically small probability, $\|S_n - S_\infty\|$ and $\|R_n\|$ are geometrically small. We have proved the following result.

Theorem 7.1. *Suppose (7.1) holds. Consider the Markov chain on $\mathcal{P}$ driven by μ. Let $P_n(q, dp)$ be the law of the chain after n moves starting from q.*

(i) *There is a unique invariant probability π.*

(ii) *There is a positive, finite constant A and an r with $0 < r < 1$ such that $\rho[P_n(q,\cdot), \pi] \le Ar^n$ for all $n = 1, 2, \ldots$ and all $q \in \mathcal{P}$.*

In this theorem, ρ is the Prokhorov metric on probabilities on $\mathcal{P}$, constructed from the variation distance on $\mathcal{P}$, as in Definition 5.1. The constant A is universal, because variation distance is uniformly bounded. If condition (7.1) fails, the chain stagnates at the starting position q.

We now specialize μ to get the Dirichlet. Recall that α is a finite measure on $\mathcal{X}$. Let $\|\alpha\| = \alpha(\mathcal{X})$ be the total mass of α and let $\gamma = \alpha/\|\alpha\|$, which is a probability on $\mathcal{X}$. Let $\tilde{\gamma}$ be the image of γ under the map $x \to \delta_x$, with $\delta_x \in \mathcal{P}$ being point mass at $x \in \mathcal{X}$. Thus, $\tilde{\gamma}$ is a probability on $\mathcal{P}$, namely, the law of δ_x when $x \in \mathcal{X}$ is chosen at random from γ. (Caution: see Section 7.2 for measurability.) Finally, we set $\mu = \text{Beta}(\|\alpha\|, 1) \times \tilde{\gamma}$. In other words, μ is the law of (u, δ_x), where u is chosen from the Beta$(\|\alpha\|, 1)$ distribution and x is independently chosen from $\alpha/\|\alpha\|$. For this μ, the law of the random probability defined by (7.3) is Dirichlet, with base measure α.

Why does the construction give D_α? We sketch the argument for a leading special case, when $\mathcal{X} = \{0, 1, 2\}$; for details, see Sethuraman and Tiwari (1982). Let $\alpha_i = \alpha(i)$ for $i = 0, 1, 2$. Then $\|\alpha\| = \alpha_0 + \alpha_1 + \alpha_2$. All we need to check is stationarity. Let Q be a random pick from D_α . Condition (7.2) for stationarity is

$$\mathcal{L}(Q) = \mathcal{L}\big(UQ + (1-U)\delta_W\big), \tag{7.6}$$

where

(7.7a) $Q \sim D_\alpha$,

(7.7b) U is Beta$(\|\alpha\|, 1)$,

(7.7c) W is i with probability $\alpha_i/\|\alpha\|$, and

(7.7d) Q, U, W are independent.

Of course, $\{Q_0, Q_1\}$—the masses assigned by Q to 0 and 1—should be Dirichlet with parameters $\alpha_0, \alpha_1, \alpha_2$ by (7.7a). The density of a Dirichlet distribution with these parameters is

$$f(x, y) = Cx^{\alpha_0 - 1} y^{\alpha_1 - 1} (1 - x - y)^{\alpha_2 - 1}$$

for (x, y) with $x > 0$, $y > 0$, $x + y < 1$. The normalizing constant C makes $\int f = 1$; its numerical value will not matter here. Condition on W in (7.6) and use (7.7cd). Stationarity boils down to

$$T_0 + T_1 + T_2 = f(x, y), \tag{7.8}$$

where

$$T_0 = \frac{\alpha_0}{\|\alpha\|} \int \frac{1}{u^2} f\Big(\frac{x - 1 + u}{u}, \frac{y}{u}\Big) g(u)\, du \tag{7.9}$$

and g is the density of the random variable U in (7.6). By (7.7b), $g(u) = \|\alpha\| u^{\|\alpha\|-1}$. We deal with T_1 and T_2, below.

The next task is to determine the range of the integral in (7.9). There are several constraints on u. First is that

$$(x - 1 + u)/u > 0, \quad \text{or} \quad u > 1 - x. \tag{7.10}$$

Second, $(x - 1 + u)/u < 1$, which follows from $x < 1$. Third, $u > y$, which follows from (7.10), because $1 - x > y$. Fourth,

$$\frac{x - 1 + u}{u} + \frac{y}{u} < 1,$$

which follows from $x + y < 1$. Finally, $u < 1$. Thus, the integral in (7.9) goes from $1 - x$ to 1; there is quite a lot of cancellation of u's, and

$$\begin{aligned} T_0 &= C y^{\alpha_1 - 1} (1 - x - y)^{\alpha_2 - 1} \alpha_0 \int_{1-x}^{1} [u - (1 - x)]^{\alpha_0 - 1}\, du \\ &= C x^{\alpha_0} y^{\alpha_1 - 1} (1 - x - y)^{\alpha_2 - 1}. \end{aligned}$$

The terms T_1 and T_2 in (7.8) can be evaluated the same way:

$$T_1 = \frac{\alpha_1}{\|\alpha\|} \int_{1-y}^{1} \frac{1}{u^2} f\Big(\frac{x}{u}, \frac{y - 1 + u}{u}\Big) g(u)\, du = C x^{\alpha_0 - 1} y^{\alpha_1} (1 - x - y)^{\alpha_2 - 1};$$

$$T_2 = \frac{\alpha_2}{\|\alpha\|} \int_{x+y}^{1} \frac{1}{u^2} f\Big(\frac{x}{u}, \frac{y}{u}\Big) g(u)\, du = C x^{\alpha_0 - 1} y^{\alpha_1 - 1} (1 - x - y)^{\alpha_2}.$$

So

$$\begin{aligned} T_0 + T_1 + T_2 = C x^{\alpha_0} y^{\alpha_1 - 1} &(1 - x - y)^{\alpha_2 - 1} \\ &+ C x^{\alpha_0 - 1} y^{\alpha_1} (1 - x - y)^{\alpha_2 - 1} \\ &\quad + C x^{\alpha_0 - 1} y^{\alpha_1 - 1} (1 - x - y)^{\alpha_2} \\ = C x^{\alpha_0 - 1} y^{\alpha_1 - 1} &(1 - x - y)^{\alpha_2 - 1}, \end{aligned}$$

because $x + y + (1 - x - y) = 1$. This completes the proof of (7.6).

The same argument goes through for any finite $\mathcal{X}$. Then compact $\mathcal{X}$ can be handled by taking limits. Along the way, it helps to check that

$$\int_{\mathcal{P}} P\, D_\alpha(dP) = \alpha/\|\alpha\|. \tag{7.11}$$

A complete separable $\mathcal{X}$ can be embedded into a compact set, so the general case follows from the compact case; (7.11) shows that D_α sits on $\mathcal{X}$, as desired, rather than spilling over onto points added by compactification.

7.2. Measure-theoretic issues. Put the weak-star σ-field on $\mathcal{P}$: this is generated by the functions $p \to \int f\, dp$ as f ranges over the bounded continuous functions on $\mathcal{X}$. The variation norm is weak-star measurable, because

$$\|p - q\| = \sup_f \left| \int f\, dp - \int f\, dq \right| \tag{7.12}$$

as f ranges over the continuous functions on $\mathcal{X}$ with $0 \le f \le 1$. With a bit of effort, we can restrict f to a countable, dense set of continuous functions. Measurability of the norm is then clear. For example, if $\mathcal{X}$ is $[0,1]$, we can restrict f to the polynomials with rational coefficients.

Put the usual Borel σ-field on $[0,1]$. Then $(u, p, q) \to f_{u,p}(q)$ is jointly measurable, from $[0,1] \times \mathcal{P} \times \mathcal{P}$ to $\mathcal{P}$. Likewise, $(u,p) \to K_{u,p} = u$ is measurable. For each n, the map

$$(\theta_1, \theta_2, \ldots, \theta_n, q) \to \big(f_{\theta_1} \circ f_{\theta_2} \circ \cdots \circ f_{\theta_n}\big)(q)$$

is jointly measurable from $\Theta^n \times \mathcal{P}$ to $\mathcal{P}$. Finally, the map $x \to \delta_x$ is measurable from $\mathcal{X}$ to $\mathcal{P}$.

The "Borel" σ-field in $\mathcal{P}$ is generated by the open sets in the norm topology, and seems to fit better with variation distance. But there is a real problem: the map $x \to \delta_x$ is not measurable if we put the Borel σ-fields on $\mathcal{X}$ and $\mathcal{P}$. A reference is Dubins and Freedman (1964). We need the variation norm to get the Lipschitz property and the weak-star σ-field to handle measurability. In a complete separable metric space, all reasonable σ-fields coincide—ranging from the Borel σ-field to (for instance) the σ-field generated by the bounded, uniformly continuous functions. The space of probability measures is complete in the variation distance but not separable. That is the source of the measure-theoretic complications.

7.3. Random means. Let P be a random pick from D_α, as defined in Section 7.1 above. Let f be a measurable function on $\mathcal{X}$. Consider the random variable $\int_{\mathcal{X}} f\, dP$. (Of course, the random variable is defined only when the integral converges.) Feigen and Tweedie (1989) prove the following result.

Proposition 7.1. $\int_{\mathcal{X}} |f(x)|\, P(dx) < \infty$ *for* D_α*-almost all* P *if and only if*

$$\int_{\mathcal{X}} \log(1 + |f(x)|)\, \alpha(dx) < \infty.$$

We now specialize $\mathcal{X}$ to the real line $(-\infty, \infty)$, and $f(x)$ to x. Suppose

$$\int_{-\infty}^{\infty} \log(1 + |x|)\, \alpha(dx) < \infty. \tag{7.13}$$

Then

$$X(P) = \int_{-\infty}^{\infty} x\, dP, \qquad P \sim D_\alpha \tag{7.14}$$

is a random variable—being the mean of a P picked at random from D_α.

Formula (7.14) must be distinguished from (7.11). In (7.11), you pick P at random from D_α, and take the mean over all P's relative to D_α: for any measurable $A \subset \mathcal{X}$,

$$\int_{\mathcal{P}} P(A) D_\alpha(dP) = \alpha(A)/\|\alpha\|.$$

In (7.14), you pick P at random from D_α, and take the mean over all x's relative to P. That gives a random variable $X(P) = \int_{-\infty}^{\infty} x\, dP$.

In a number of cases, the distribution of X relative to D_α can be be computed explicitly, using the idea of iterated random functions. For instance, Cifarelli and Regazzini (1990) show that unless α is a point mass, $P \to \int x\, dP$ has an absolutely continuous distribution, and they give formulas for the density. Additional results are obtained by Diaconis and Kemperman (1996).

Example 7.1. Suppose α concentrates on two points, 0 and 1. Relative to D_α, $P \to X(P)$ has the Beta(α_0, α_1) distribution. This is immediate from the discussion in Section 7.1 above: after all, $X(P)$ is the mass P assigns to 1.

Example 7.2. If α is uniform on $[0, 1]$, then X has the density

$$\frac{e}{\pi} x^{-x}(1-x)^{-(1-x)} \sin(\pi x) \quad \text{for} \quad 0 < x < 1.$$

Example 7.3. If α is Cauchy then X also has the Cauchy distribution. See Yamamoto (1984). Of course, $\int x\, \alpha(dx)$ does not converge. On the other hand, (7.13) holds, so that for almost all P drawn from D_α, the integral in (7.14) does converge. Picks from D_α have a shorter tail than α.

Example 7.4. Let Z be Cauchy. If α is the law of $e^Z/(1 + e^Z)$, then X is uniform on $[0, 1]$.

For the mathematics behind examples (7.2–3–4), we refer to Diaconis and Kemperman (1996) where connections to the Markov moment problem and recent work of Kerov (1993) are explained. We conclude by showing how the law of X in (7.14) can be obtained as the stationary distribution under random iterated functions. This is fairly immediate on the basis of Section 7.1. The state space is the real line. From x, the chain moves to $Ux + (1-U)W$, where U is Beta$(\|\alpha\|, 1)$, and W is an independent pick from $\alpha/\|\alpha\|$. The limiting stationary distribution, which is $\mathcal{L}(X)$, is the distribution of

$$(1-U_1)W_1 + U_1(1-U_2)W_2 + U_1U_2(1-U_3)W_3 + \cdots,$$

where (U_i, W_i) are i.i.d. copies of (U, W): see (7.3).

Acknowledgment. We would like to thank Philippe Bougerol, Olle Häggström, and Yuval Peres for many useful suggestions. We also thank our very helpful editors, Dan Rockmore and Nick Trefethen.

Technical Report No. 511
Department of Statistics
University of California
Berkeley, CA 94720

References

Anderson, T. W. (1959), *On asymptotic distributions of estimates of parameters of stochastic difference equations*, Ann. Math. Statist. **30**, 676–87.

Arnold, L. (1998), *Random Dynamical Systems*, University of Bremen (to appear).

Arnold, L. and Crauel, H. (1992), *Iterated function systems and multiplicative ergodic theory*, Diffusion Theory and Related Problems in Analysis II (M. Pinsky and V. Wihstatz, ed.), Birkhauser, Boston, pp. 283–305.

Babillot, M., Bougerol, P., and Elie, L. (1997), *The random difference equation $X_n = A_n X_{n-1} + B_n$ in the critical case*, Ann. Prob. **25**, 478–93.

Baccelli, F. (1992), *Ergodic theory of stochastic Petri networks*, Ann. Prob. **20**, 375–396.

Baccelli, F. and Brémaud, P. (1994), *Elements of Queuing Theory*, Springer Verlag, New York.

Baccelli, F., Cohen, G., Olsder, G. J., and Quadrat, J. P. (1992), *Synchronization and Linearity*, Wiley, New York.

Barnsley, M. (1993), *Fractals Everywhere*, 2e, Academic Press.

Barnsley, M. and Elton, J. (1988), *A new class of Markov processes for image encoding*, Adv. Appl. Prob. **20**, 14–32.

Bidigare, P., Hanlon, P., and Rockmore, D. (1997), *A combinatorial description of the spectrum for the Tsetlin library and its generalization to hyperplane arrangements*, Duke Math Jour. (to appear).

Blumenthal, R. and Corson, H. (1970), *On continuous collections of measures*, Ann. Inst. Fourier Grenoble **20**, 193–199.

Blumenthal, R. and Corson, H. (1971), *On continuous collections of measures*, Proc. Sixth Berk. Symp. **2**, 33–40.

Borovkov, A. (1984), *Asymptotic Methods in Queuing Theory*, Wiley, New York.

Borovkov, A. and Foss, S. (1992), *Stochastically recursive sequences and their generalizations*, Siberian Adv. Math. **2**, 16–81.

Bougerol, P. and Picard, N. (1992), *Strict stationarity on generalized autoregressive processes*, Ann. Probab. **20**, 1714–1730.

Brandt, A. (1986), *The stochastic equation $Y_{n+1} = A_n Y_n + B_n$ with stationary coefficients*, Adv. Appl. Probab. **18**, 211–220.

Brandt, A., Franken, P., and Lisek, B. (1990), *Stationary Stochastic Models*, Wiley, New York.

Breiman, L. (1960), *The strong law of large numbers for a class of Markov chains*, Ann. Math. Statist. **31**, 801–3.

Breiman, L. (1968), *Probability*, Addison Wesley, New York.

Brown, K. and Diaconis, P. (1997), *Random walk and hyperplane arrangements*, Ann. Probab. (to appear).

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

Cifarelli, D. and Regazzini, E. (1990), *Distribution functions of means of a Dirichlet process*, Ann. Statist. **18**, 429–442.

Crownover, R. (1995), *Introduction to Fractals and Chaos*, Jones and Bartlett, Boston.

Diaconis, P. and Freedman, D. (1986), *On the consistency of Bayes estimates*, Ann. Statist. **14**, 1–67.

Diaconis, P. and Kemperman, J. (1996), *Some new tools for Dirichlet priors*, Bayesian Statistics 5 (J. Bernardo et al, eds.), Oxford University Press, pp. 97–106.

Diaconis, P. and Shahshahani, M. (1986), *Products of random matrices and computer image generation*, Contemp. Math. **50**, 173–82.

Dubins, L. and Freedman, D. (1966), *Invariant probabilities for certain Markov processes*, Ann. Math. Statist. **37**, 837–844.

Dubins, L. and Freedman, D. (1964), *Measurable sets of measures*, Pacific J. Math. **14**, 1211–22.

Dubins, L. and Freedman, D. (1967), *Random distribution functions*, Fifth Berkeley Symp. Math. Statist. Prob. **II part 1**, 183–214.

Dubischar, D. (1997), *The representation of Markov processes by random dynamical systems*, Technical Report 393, Institut für Dynamische Systeme, University of Bremen.

Dudley, R. (1989), *Real Analysis and Probability*, Wadsworth, Pacific Grove, Calif.

Duflo, M. (1997), *Random Iterative Models*, Springer Verlag, New York.

Elton, J. (1990), *A multiplicative ergodic theorem for Lipschitz maps*, Stochastic Proc. Appl. **34**, 39–47.

Embree, M. and Trefethen, L. N. (1998), *Surprising behavior of random Fibonacci sequences*, Technical report, Oxford University Computing Laboratory.

Erdös, P. (1939), *On a family of symmetric Bernoulli convolutions*, Amer. J. Math. **61**, 974–75.

Erdös, P. (1940), *On the smoothness properties of Bernoulli convolutions*, Amer. J. Math. **62**, 180–86.

Fabius, J. (1964), *Asymptotic behavior of Bayes estimates*, Ann. Math. Statist. **35**, 846–856.

Feigen, P. and Tweedie, E. (1989), *Linear functionals and Markov chains associated with the Dirichlet process*, Math. Proc. Camb. Phil. **105**, 579–585.

Feller, W. (1971), *An Introduction to Probability Theory and its Applications*, vol. II, 2nd ed., Wiley, New York.

Ferguson, T. (1973), *A Bayesian analysis of some nonparametric problems*, Ann. Statist. **1**, 209–230.

Ferguson, T., Phadia, E., and Wari, R. (1992), *Bayesian nonparametric inference*, Issues in Statistical Inference: Essays in Honor of D. Basu (M. Ghosh and P. K. Pathak, eds.), IMS Lecture Notes 17, Inst. Math. Statist., Hayward, Calif, pp. 127-50.

Fill, J. (1998), *An interruptible algorithm for perfect sampling via finite Markov chains*, Ann. Appl. Prob. (to appear).

Fisher, Y., ed. (1994), *Fractal Image Generation*, Springer, New York.

Freedman, D. (1963), *On the asymptotic behavior of Bayes estimates in the discrete case*, Ann. Math. Statist. **34**, 1386–1403.

Furstenberg, H. (1963), *Non-commuting random products*, Trans. Amer. Math. Soc. **108**, 377-428.

Garsia, A. (1962), *Arithmetic properties of Bernoulli convolutions*, Trans. Amer. Math. Soc. **102**, 409–32.

Goldie, C. (1991), *Implicit renewal theory and tails of solutions of random equations.*, Ann. Appl. Probab. **1**, 126–166.

Goldie, C. and Maller, R. (1997), *Stability of perpetuities*, Technical Report, Dept. of Mathematics, University of Western Australia.

Häggström, O. and Nelander, K. (1998), *On exact simulation of Markov random fields using couplings from the past*, Technical report, Chalmers Technical University, Göteberg, Sweden.

Häggström, O., van Lieshout, M.-C., and Mϕller, J. (1998), *Characterization results and Markov chain Monte Carlo algorithms including exact simulation for some spatial point processes*, Bernoulli Journal (to appear).

Hammersley, J. and Handscomb, D. (1964), *Monte Carlo Methods*, Chapman and Hall, London.

Hutchinson, J. (1981), *Fractals and self–similarity*, Indiana University Math. Jour. **30**, 713–747.

Jessen, A. and Wintner, A. (1935), *Distribution functions and the Riemann zeta function*, Trans. Amer. Math. Soc. **38**, 48–88.

Kerov, S. (1993), *Transition probabilities for continual Young diagrams and the Markov moment problem*, Technical Report, Institute for Electricity and Communications, St. Petersburg, 191665, Russia.

Kesten, H. (1973), *Random difference equations and renewal theory for products of random matrices*, Acta Math. **131**, 207–248.

Kifer, Y. (1986), *Ergodic Theory of Random Transformations*, Birkhauser, Boston.

Kinderman, R. and Snell, J. L. (1980)., *Markov Random Fields*, American Mathematical Society, Contemp. Math. vol. 1.

Letac, G. (1986), *A contraction principle for certain Markov chains and its applications*, Contemp. Math. **50**, 263–273.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., and Teller, E. (1953), *Equations of state calculations by fast computing machines*, Jour. Chem. Physics **21**, 1087–92.

Mϕller, J. (1998), *Markov Chain Monte Carlo and spatial point processes*, Stochastic geometry, Likelihood, and Computation (O. Barndorff-Nielsen, W. S. Kendall and M.-C. van Lieshout, eds.), Seminaire Européen de Statistique, Chapman and Hall, London. (to appear).

Meyn, S. P. and Tweedie, R. L. (1993), *Markov Chains and Stochastic Stability*, Springer, London.

Peres, Y. and Solomyak, B. (1996), *Absolute continuity of Bernoulli convolutions, a simple proof*, Math. Research Letters **3**, 231–239..

Peres, Y. and Solomyak, B. (1998), *Self-similar measures and intersections of Cantor sets*, Trans. Amer. Math. Soc. (to appear).

Priestley, M. (1988), *Non-Linear and Non-Stationary Time Series Analysis*, Academic Press, New York.

Propp, J. and Wilson, D. (1996), *Exact sampling with coupled Markov chains*, Random Structures and Algorithms **9**, 223–52.

Propp, J. and Wilson, D. (1998), *How to get a perfectly random sample from a generic Markov chain and generate a random spanning tree of a directed graph*, J. Algorithms (to appear).

Quas, A. (1991), *On representation of Markov chains by random smooth maps*, Bull. London Math. Soc. **23**, 487–492.

Rachev, S. and Samorodnitsky, G. (1995), *Limit laws for a stochastic process and random recursion arising in probabilistic modelling*, Adv. Appl. Probab. **27**, 185–202.

Samorodnitsky, G. and Taqqu, M. (1994)., *Stable Non-Gaussian Random Processes: Stochastic Models with Infinite Variance*, Chapman & Hall, New York.

Sethuraman, J. and Tiwari, R. (1982), *Convergence of Dirichlet measures and the interpretation of their parameters*, Statistical Decision Theory and Related Topics III (J. Berger and S. Gupta, eds.), Academic Press, New York, pp. 305–15.

Solomyak, B. (1995), *On the random series $\pm\lambda^i$ (an Erdös problem)*, Annals of Math. **242**, 611–625.

Spitzer, F. (1956), *A combinatorial lemma and its application to probability theory*, Trans. Amer. Math. Soc. **. 82**, 323-339.

Steinsaltz, D. (1997), *Zeno's walk: a random walk with refinements*, Probab. Th. Rel. Fields **107**, 99-121.

Steinsaltz, D. (1998), *Locally contractive iterated function systems*, Technical Report, Berlin University.

Strassen, V. (1965), *The existence of probability measures with given marginals*, Ann. Math. Statist. **36**, 423–38.

Vervaat, W. (1979), *On a stochastic difference equation and a representation of non-negative infinitely divisible random variables*, Adv. Appl. Probab. **11**, 750–783.

Yamato, H. (1984), *Characteristic functions of means of distributions chosen from a Dirichlet processes*, Ann. Probab. **12**, 262–267.

Zolotarev, V. (1986), *One-Dimensional Stable Distributions*, American Mathematical Society, Providence, R.I., Translations of mathematical monographs, vol. 65.