Achieving Limiting Distributions for Markov Chains Using Back Buttons

by

Andrey Feuerverger[1] and Jeffrey S. Rosenthal[1]

(April 22, 2002; last revised April 30, 2003.)

Abstract. As a simple model for browsing the World Wide Web, we consider Markov chains with the option of moving "back" to the previous state. We develop an algorithm which uses back buttons to achieve essentially any limiting distribution on the state space. This corresponds to spending the desired total fraction of time at each web page. On finite state spaces, our algorithm always succeeds. On infinite state spaces the situation is more complicated, and is related to both the tail behaviour of the distributions, and the properties of convolution equations.

1. Introduction.

We consider Markov chains with "back" buttons, where the chain sometimes retreats to its previous state instead of moving to the next state. This modified chain (or "backoff process") is intended to model a user browsing the World Wide Web, where the web pages are the states of the chain, and where the user sometimes presses the "back button".

Fagin et al. (2001) show that in this context, if the probability of going back is less than 1/2 and is independent of the walk's history, then the limiting distribution of the backoff process is the same as that of the original walk. However, if the probability of going back depends on the walk's history (e.g. through the current state), then the situation is more complicated. Fagin et al. (2001, Lemma 4.39) do indicate how to compute the limiting distribution in certain simple cases, however the resulting expressions are very complicated (requiring solving cubic equations with coefficients like 3936 and 3180, even for a very simple example on a state space of size 2).

In this paper, we take a different approach, and ask the following question. Suppose we are given fixed Markov chain transition probabilities, on a state space $\mathcal{X}$. Suppose we are also given a probability measure $\pi(\cdot)$ on $\mathcal{X}$, which represents our *desired* distribution of what

[1]Department of Statistics, University of Toronto, Toronto, Ontario, Canada M5S 3G3. Supported in part by NSERC of Canada. E-mail: `andrey@utstat.toronto.edu` and `jeff@math.toronto.edu`.

fraction of our time should be spent where. At each step, we may choose to (i) do nothing, (ii) go "back" one step, or (iii) go forward one step of the Markov chain. [Note that in (iii), the resulting jump is chosen randomly according to the underlying Markov chain, and we have no control over precisely where it jumps.] The question is, can we choose which of (i), (ii), and (iii) to perform at each step, in such a way that our long-term limiting distribution is precisely $\pi(\cdot)$?

From the point of view of browsing the Web, this question may be interpreted as follows. Suppose $\pi(j)$ represents your (objective or subjective) opinion of what fraction of your time you would like to spend at the Web site j. Then what back button scheme can you use when browsing the Web, so that the fraction of time you ultimately spend at web page j will indeed be equal to $\pi(j)$?

To address this, we shall introduce the BACKMET algorithm, which accomplishes our goal provided a certain non-negativity condition holds. Our algorithm is related to the *Metropolis-Hastings* algorithm, as we shall explain. However, unlike the Metropolis-Hastings algorithm, the BACKMET algorithm does not reject proposed moves until *after* the proposed state is visited (to more accurately model the web browsing context), and this difference is the essential difference between the BACKMET algorithm and previous MCMC algorithms.

We shall prove (Theorem 1) that the BACKMET algorithm suffices to achieve any limiting distribution, provided only that a certain condition (C1) is satisfied. We shall further prove (Theorem 6) that on a finite state space, this condition can always be satisfied.

On an infinite state space the situation is more complicated, depending on the tail behaviour of $\pi(\cdot)$ and of the transition probabilities $P(x,\cdot)$. However, in the case where the Markov chain is a random walk (corresponding to convolution of densities), we shall prove (Theorem 11) that BACKMET essentially always succeeds provided the target density is bounded away from zero on compact sets, and the target density tails are at least as heavy as the tails of the Markov chain jumps themselves. On the other hand, we shall also provide (Theorems 7 through 10) quite a number of necessary conditions which show that, for a fixed choice of parameters, the BACKMET algorithm on infinite state spaces will not always succeed.

Acknowledgements. We are very grateful to Allan Borodin and Jon Kleinberg for introducing us to this topic, and for helpful advice. We thank M. Kanter for referring us to the theorem of Strassen cited in the text. We thank the editors and referees for helpful comments.

2. The Metropolis-Hastings Algorithm.

To understand the BACKMET algorithm, we recall the *Metropolis-Hastings algorithm* (Metropolis et al., 1953; Hastings, 1970). This algorithm requires a fixed "target" density f, with respect to some reference measure $\nu(\cdot)$. It then proceeds for $n = 0, 1, 2, \ldots$ as follows. Given X_n, it chooses a proposal state Y_{n+1} according to some given proposal kernel $q(X_n, y)\nu(dy)$. It then *accepts* this proposal and sets $X_{n+1} = Y_{n+1}$ with probability $\min\left[1, \frac{f(X_n)\ q(X_n, X_{n-1})}{f(X_{n-1})\ q(X_{n-1}, X_n)}\right]$. Otherwise, it *rejects* this proposal and simply sets $X_{n+1} = X_n$.

It is well known (and straightforward to see) that the Metropolis-Hastings algorithm is *reversible* with respect to the distribution $f(x)\,\nu(dx)$, i.e. that

$$\int_A P(X_{n+1} \in B \,|\, X_n = x)\, f(x)\, \nu(dx) = \int_B P(X_{n+1} \in A \,|\, X_n = x)\, f(x)\, \nu(dx)\,, \qquad A, B \subseteq \mathcal{X}\,.$$

It then follows from the general theory of Markov chains (e.g. Meyn and Tweedie, 1993; Tierney, 1994) that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} P(X_i \in A) \;=\; \int_A f(x)\, \nu(dx)\,, \qquad A \subseteq \mathcal{X}\,.$$

That is, the limiting distribution of the X_i will be given by $f(x)\,\nu(dx)$. We remark that the Metropolis algorithm is considered to be one of the top ten computer algorithms of all time (Dongarra and Sullivan, 2000). It is one of a family of *Markov chain Monte Carlo* (MCMC) algorithms designed to achieve specified limiting distributions; see e.g. Tierney, 1994.

Since the Metropolis-Hastings algorithm achieves the desired limiting distribution $\pi(dx) = f(x)\,\nu(dx)$, by only moving forward one step (option (iii)) or doing nothing (option (i)), it appears that this algorithm solves our problem elegantly. However, this algorithm requires the ability to reject a proposed new state Y_{n+1}, without ever visiting it or counting it as a visited state. By contrast, when browsing the web, typically one needs to *visit* a site before one can choose to reject it by using the back button. Thus, we take the point of view that we seek an algorithm whose choices depend as little as possible on unobserved information, such as which state we are considering visiting next or what properties the future state may have. Nor do we wish our algorithm to require strong information about the underlying Markov chain, such as its own stationary distribution or limiting probabilities (if they even exist).

3. Formal Definitions.

We assume there are fixed underlying Markov chain transition probabilities $P(x, dy)$ on a state space $\mathcal{X}$. We also assume that there is a fixed probability distribution $\pi(\cdot)$ on $\mathcal{X}$, which represents the desired limiting distribution.

To avoid analytic technicalities, we assume that there is a reference probability measure $\nu(\cdot)$ on $\mathcal{X}$, such that the Markov chain transition probabilities $P(x, dy)$ and target distribution $\pi(dy)$ are absolutely continuous with respect to $\nu(\cdot)$, with $P(x, dy) = p(x, y)\nu(dy)$, and $\pi(dy) = f(y)\nu(dy)$ for some function $f : \mathcal{X} \to (0, \infty)$. (If $\mathcal{X}$ is *discrete* (i.e. finite or countable), then we can simply let ν be counting measure, with $\nu(\{x\}) = 1$ for all $x \in \mathcal{X}$. Similarly, if $\mathcal{X}$ is an open subset of $\mathbf{R}$ or $\mathbf{R}^d$, then $\nu(\cdot)$ can be Lebesgue measure, with $p(x, \cdot)$ and $f(\cdot)$ ordinary densities.)

To avoid the problem of certain states being unobtainable, we also assume that the Markov chain is *ϕ-irreducible*, meaning that there is a non-zero measure $\phi(\cdot)$ on $\mathcal{X}$, such that for any $x \in \mathcal{X}$ and any $A \subseteq \mathcal{X}$ with $\phi(A) > 0$, there is positive probability that the chain will eventually reach A when started at x. (If $\mathcal{X}$ is discrete, then it suffices that the chain be "irreducible" in the standard discrete sense.)

Our formal definitions are as follows. The backoff process keeps track of an entire *history stack* $H_n = \langle s_{n,1}, s_{n,2}, \ldots, s_{n,\ell(H_n)} \rangle$, of length $\ell(H_n)$, at each time n. Here $X_n \equiv s_{n,\ell(H_n)}$ represents the current state of the chain, while $s_{n,1}, s_{n,2}, \ldots, s_{n,\ell(H_n)-1}$ represent previously-visited states that the process can go "back" to.

Initially, $H_0 = \langle X_0 \rangle$ and $\ell(H_0) = 1$, where X_0 is some arbitrary initial state. Then, at times $n = 0, 1, 2, \ldots$, given the current history stack H_n, the user may choose one of three options for selecting the next history stack H_{n+1}:

(i) Do nothing. That is, $H_{n+1} = H_n$, and $\ell(H_{n+1}) = \ell(H_n)$.

(ii) Go "back" one step in the history stack. That is, if $H_n = \langle s_1, s_2, \ldots, s_{\ell-1}, s_\ell \rangle$, then $H_{n+1} = \langle s_1, s_2, \ldots, s_{\ell-1} \rangle$, and $\ell(H_{n+1}) = \ell(H_n) - 1$. [If $\ell(H_n) = 1$, then this option is not permitted.]

(iii) Go forward one step of the Markov chain. That is, if $H_n = \langle s_1, s_2, \ldots, s_\ell \rangle$, then $H_{n+1} = \langle s_1, s_2, \ldots, s_\ell, Z \rangle$, and $\ell(H_{n+1}) = \ell(H_n) + 1$, where Z is chosen randomly, independent of all previous random variables, such that $\mathbf{P}(Z \in A) = P(s_\ell, A)$, i.e. with Z having the probability distribution $P(s_\ell, \cdot)$ given by the underlying Markov chain transition probabilities.

The question we ask is, by a judicious choice of which of (i), (ii), or (iii) to perform at each step, can the user ensure that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} P(X_i \in A) \;=\; \pi(A)\,, \qquad A \subseteq \mathcal{X}\,,$$

i.e. that the limiting distribution is given by $\pi(\cdot)$. (Recall that here $X_i = s_{i,\ell(H_i)}$ is the top of the history stack at time i, i.e. is the position of the chain itself at time i.) As discussed above, we want the choice of which of (i), (ii), or (iii) to perform not to depend unduly on unobserved information, such as which state we are considering visiting next or what properties the future state may have.

Remark. In the back option (ii) above, it is tempting to instead let $H_{n+1} = \langle s_1, s_2, \ldots, s_{\ell-1}, s_\ell, s_{\ell-1}\rangle$. However, in this case, if the user used option (ii) twice in a row, they would then obtain $X_{n+2} = s_\ell$ instead of $X_{n+2} = s_{\ell-2}$, which is not appropriate. (On the other hand, the BACKMET algorithm as we shall define it never uses option (ii) twice in a row, so in fact it would not be affected by this change.)

4. The BACKMET Algorithm.

Let $p(x,y)$ and $f(y)$ be the transition and target densities as above, let $k \geq 2$ be an integer (to be a parameter in our algorithm), and let t_k be some density on $\mathcal{X}$ (so $t_k(x) \geq 0$ for $x \in \mathcal{X}$, with $\int_{\mathcal{X}} t_k(x)\,\nu(dx) = 1$), to be determined later (see Section 5). In terms of these quantities, we define

$$\alpha_k(x,y) \;=\; \min\left[1,\; \frac{t_k(y)\,p(y,x)}{t_k(x)\,p(x,y)}\right]\,, \tag{1}$$

with the convention that $\alpha_k(x,y) = 1$ if $t_k(x)\,p(x,y) = 0$.

We now define the BACKMET algorithm. Given the history H_n at time n, it proceeds as follows, for $n = 0, 1, 2, \ldots$.

(a) If $n \equiv 0 \pmod{k}$ (i.e. n is an integer multiple of k), then do a forward step of the Markov chain as in (iii) above.

(b) If $n \equiv 1 \pmod{k}$, then with probability $\alpha_k(X_n, X_{n-1})$ do nothing as in (i) above, otherwise with probability $1 - \alpha_k(X_n, X_{n-1})$ go back as in (ii) above.

(c) If $n \not\equiv 0 \pmod{k}$ and $n \not\equiv 1 \pmod{k}$, then do nothing as in (i) above.

In words, the BACKMET algorithm proceeds in blocks of k steps, as follows. For the first step it does one forward step of the Markov chain; for the second step it either does nothing or goes back one step; and for the following $k-2$ steps it does nothing. It then repeats this pattern during the next k steps. We shall justify this unusual pattern in the following sections.

It remains to specify the function t_k. Given any function $h : \mathcal{X} \to \mathbf{R}$, we define

$$(Qh)(y) \;=\; \int h(x)\, p(x,y)\, \nu(dx)\,. \tag{2}$$

(Thus $(Q^2h)(z) \;=\; \int\int h(x)\, p(x,y)\, p(y,z)\, \nu(dx)\, \nu(dy)$, etc.) In terms of this, we define the function $t_k : \mathcal{X} \to [0,\infty)$ by

$$t_k \;=\; \frac{k}{k-1}\Big(f - \frac{1}{k-1}Qf + \frac{1}{(k-1)^2}Q^2 f - \frac{1}{(k-1)^3}Q^3 f + \ldots\Big)\,. \tag{3}$$

Thus $\int_{x\in\mathcal{X}} t_k(x)\, \nu(dx) = 1$. We shall require the following condition:

(C1) The series (3) converges for all $x \in \mathcal{X}$, with $t_k(x) \geq 0$ for all $x \in \mathcal{X}$.

Of course, if $k \geq 3$ and f is bounded from above, then the convergence part of (C1) is automatic.

We can now state our main result (proved in Section 6), as follows.

Theorem 1. *Consider the BACKMET algorithm. Assume that (C1) holds for some k, and define t_k by (3). Then for π-a.e. starting point X_0,*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} P(X_i \in A) \;=\; \pi(A)\,,$$

i.e. the correct limiting distribution is achieved. Furthermore, if $\mathcal{X}$ is finite and $\nu(\cdot)$ is counting measure, then (C1) always holds for all sufficiently large k.

Remarks.

1. Note that $\frac{1}{n}\sum_{i=1}^{n} P(X_i \in A)$ is the expected fraction of the time that the chain spends in the subset A up to time n. Thus, Theorem 1 says that in the limit, the fraction of time spent in A is equal to $\pi(A)$, as desired.

2. If we assume the BACKMET algorithm is *aperiodic*, then we also have the slightly stronger conclusion that $\lim_{n\to\infty} P(X_n \in A) = \pi(A)$.

3. It is indeed possible, though rare, that the BACKMET algorithm will be periodic. This is true, for example, if P corresponds to simple symmetric random walk on the state space $\mathcal{X}$ consisting of M points arranged in a circle, and if $\pi(\cdot)$ is the uniform distribution on $\mathcal{X}$ (so that the Metropolis-Hastings algorithm will never reject).

4. Whether the BACKMET algorithm is periodic or not, a slight modification suffices to ensure that $\lim_{n \to \infty} P(X_n \in A) = \pi(A)$. Indeed, if the induced chain $\{X_{nk}\}$ has period r, then it suffices to choose Z uniformly from the set $\{0, 1, 2, \ldots, rk - 1\}$, and modify the algorithm by doing nothing for the first Z steps before running BACKMET as usual.

5. If we assume that the chain is "Harris recurrent", then the conclusion holds for all starting points X_0, not just for π-a.e. X_0. This is always the case if $\mathcal{X}$ is countable.

6. We stated earlier that we wanted an algorithm which did not use unobserved information. Now, the probability (1) of using the back button only depends on the two previously visited states x and y, which are both previously observed states. However, the BACKMET algorithm also requires the function t_k from (3), and the computation of this function requires global information. We view this as a limitation of our algorithm. However, in contrast to the Metropolis-Hastings algorithm, BACKMET only requires this single global function, computed just once at the beginning, and otherwise uses only observed information and evaluates t_k only at visited sites. Furthermore, even without computing t_k, the algorithm proves that there *exists* a natural way to browse the web achieving desired limiting probabilities, even though without knowing t_k explicitly the correct back probabilities (1) cannot be computed precisely.

7. Because of this need to visit states before rejecting them, the BACKMET algorithm is inherently less efficient than the Metropolis-Hastings algorithm, by a factor of approximately k. However, we will see in the next section that typically the value of k is quite moderate. In any case, the BACKMET algorithm still runs extremely quickly and converges quite rapidly. In particular, since $k - 2$ of every k steps involve doing nothing, they add very little in the way of computational burden. Thus, the BACKMET algorithm is certainly practical from a computational standpoint. (Of course, when a real person is browsing the web, they typically make their browsing choices much more slowly than a computer does – but that is a limitation of the human, not the algorithm.)

5. A Numerical Example.

To illustrate the use of the BACKMET algorithm, we construct an example as follows.

We let the state space be $\mathcal{X} = \{1, 2, \ldots, 1000\}$, consisting of 1000 states arranged in a circle (so that state 1 is adjacent to states 2 and 1000). We put an edge between every adjacent pair of states (to ensure connectivity), for a total of 1000 edges. We then add an additional 6000 edges uniformly at random, to create a messy, complicated underlying

Markov chain. (In particular, we do allow for multiple edges between the same pairs, or even self-loops.)

From the point of view of the World Wide Web, we may consider this to be creating an artificial web of 1000 pages with a very complicated link structure.

The underlying Markov chain then does a random walk on the graph $\mathcal{X}$. That is, from a state $x \in \mathcal{X}$, it chooses uniformly among all edges emanating from x, and moves to the other end of that edge. From the point of view of the Web, this involves choosing uniformly at random one of the out-links from the current web page.

For definiteness, we choose the target distribution $\pi(\cdot)$ so that state x has weight proportional to $10 + x$; thus, $f(x) = (x + 10) / \sum_{i=1}^{1000}(i + 10)$ (and $\nu(\cdot)$ is counting measure). In particular, state 1000 is nearly 100 times more desired than is state 1. Furthermore, we let $X_0 = 1$, corresponding to the "worst case" of starting at the lowest-ranked state.

We repeatedly ran the BACKMET algorithm on random such configurations. In each case, we found the minimal value of k which satisfied (C1), and then ran BACKMET with that value of k. We found that the values of k were typically around 50. It required less than a second on a modern PC to do a complete (linear) search for the minimal k, including computing the entire function t_k each time (and, of course, such computation only needs to be done once for each configuration). It then performed over 10^7 individual BACKMET iterations per second. Thus, computational speed was certainly no problem.

To investigate convergence of the probabilities to the target distribution, we obtained estimates of the total variation distance to stationarity of the probabilities in the j^{th} block of k BACKMET steps, i.e.

$$\begin{aligned} TV(j) &= \Big\| \frac{1}{k} \sum_{t=jk+1}^{jk+k} \mathbf{P}(X_t \in \cdot) - \pi(\cdot) \Big\|_{T.V.} \\ &= \sup_{A \subseteq \mathcal{X}} \Big| \frac{1}{k} \sum_{t=jk+1}^{jk+k} \mathbf{P}(X_t \in A) - \pi(A) \Big| \\ &= \frac{1}{2} \sum_{x \in \mathcal{X}} \Big| \frac{1}{k} \sum_{t=jk+1}^{jk+k} \mathbf{P}(X_t = x) - f(x) \Big| . \end{aligned} \tag{4}$$

Specifically, the value of $TV(j)$ was estimated by performing 10^8 multiple runs of the BACKMET algorithm on the same configuration, and counting the fraction of times that $X_t = x$, to accurately estimate $\mathbf{P}(X_t = x)$, and hence estimate $TV(j)$ via (4). [The C program used is available at `http://probability.ca/jeff/comp/backmet.c`.]

The results of runs on the first six different configurations tried are presented here.

Configuration	k	$TV(1)$	$TV(5)$	$TV(10)$	$TV(15)$	$TV(20)$
1.	58	0.988678	0.200154	0.032524	0.006556	0.001969
2.	59	0.982682	0.181410	0.038307	0.009954	0.003005
3.	64	0.981638	0.152516	0.022251	0.004447	0.001612
4.	51	0.983976	0.160724	0.027432	0.006117	0.002005
5.	42	0.992572	0.212237	0.035738	0.007781	0.002284
6.	78	0.978688	0.190675	0.036436	0.010157	0.003752

We see from these results that, even for these complicated graphs of size 1000, the required values of k are quite moderate. In addition, we see that the convergence to stationarity is quite rapid, with the total variation distance being far less than 1% after just 15–20 blocks of k iterations, or about 1000 total iterations of the BACKMET algorithm. (Furthermore, some of the remaining total variation distance is due to Monte Carlo error in the estimates of $P(X_t = x)$ as above.) Recalling that $(k-2)/k$ of the BACKMET steps involve doing nothing, and that over 10^7 individual BACKMET iterations can be performed per second, we see that the computational burden is quite minimal.

We conclude that for this example, the BACKMET algorithm runs extremely quickly, and converges in distribution quite rapidly, so that it is computationally very easy to run. Of course, a more complete investigation of BACKMET's performance would require running BACKMET on a large number of very different examples, to get a broader idea of its running and convergence properties. However, the above example suggests that usually BACKMET, like the Metropolis-Hastings algorithm, will run quickly and efficiently.

6. Theoretical Analysis.

The connection of the BACKMET algorithm to the Metropolis-Hastings algorithm is given by the following.

Proposition 2. *Consider the BACKMET algorithm, with $\alpha_k(x, y)$ as in (1). Let $W_n = X_{kn}$ (so that the process $\{W_n\}$ samples the process $\{X_i\}$ once every k^{th} step). Then $\{W_n\}$ follows a Metropolis-Hastings algorithm with target density function t_k. Hence, we have $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} P(W_i \in A) = \int_A t_k(x)\nu(dx)$.*

Proof. By construction of the BACKMET algorithm, we will have $W_{n+1} = X_{kn+1}$ with probability $\alpha_k(X_{kn+1}, W_n)$, while $W_{n+1} = W_n$ with probability $1 - \alpha_k(X_{kn+1}, W_n)$. It then follows from (1) that the chain $\{W_n\}$ does indeed follow a Metropolis-Hastings algorithm with target density t_k. The convergence results then follow, using ϕ-irreducibility, from general Markov chain theory (e.g. Meyn and Tweedie, 1993; Tierney, 1994). ∎

Now, under the BACKMET algorithm, we see that $X_{kn+i} = W_{n+1}$ as above for $i = 2, 3, \ldots, k$, i.e. for $k-1$ out of every k steps. Furthermore, X_{kn+k+1} is constructed by going forward from W_n by one step of the Markov chain. Hence, from the above Proposition, we immediately obtain the following.

Proposition 3. *The BACKMET algorithm satisfies that*

$$\begin{aligned}\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} P(X_i \in A) &= \frac{k-1}{k}\int_{x\in A} t_k(x)\nu(dx) + \frac{1}{k}\int_{y\in\mathcal{X}} t_k(y)P(y,A)\nu(dy) \\ &\equiv \frac{k-1}{k}T_k(A) + \frac{1}{k}(T_kP)(A)\,, \qquad A \subseteq \mathcal{X}\,,\end{aligned}$$

where $T_k(A) = \int_A t_k(x)\,\nu(dx)$, and $(T_kP)(A) = \int_{y\in\mathcal{X}} t_k(y)P(y,A)\nu(dy)$. That is, the limiting fraction of time that the backoff process spends in a subset A is given by a weighted average of $T_k(A)$ and $(T_kP)(A)$.

This Proposition tells us that, if we could choose the density t_k so that $\frac{k-1}{k}T_k(A) + \frac{1}{k}(T_kP)(A) = \pi(A)$ for all $A \subseteq \mathcal{X}$, then BACKMET would indeed converge to $\pi(\cdot)$. The question then becomes, how should the density t_k be chosen. We consider that next.

The following Lemma says that, if h is the density of μ, then Qh is the density of μP.

Lemma 4. *Let $h = \frac{d\mu}{d\nu}$ be a density function for a probability measure $\mu(\cdot)$ on $\mathcal{X}$ with respect to $\nu(\cdot)$. Then Qh is the density function for the probability measure μP, given by $(\mu P)(A) = \int_{x\in\mathcal{X}} P(x,A)\mu(dx)$.*

Proof. We compute that

$$\begin{aligned}\int_{y\in A}(Qh)(y)\,\nu(dy) &= \int_{y\in A}\int_{x\in\mathcal{X}} h(x)\,p(x,y)\,\nu(dx)\,\nu(dy) \\ &= \int_{y\in A}\int_{x\in\mathcal{X}} p(x,y)\,\mu(dx)\,\nu(dy) \\ &= \int_{x\in\mathcal{X}} p(x,A)\,\mu(dx) \;=\; (\mu P)(A)\,,\end{aligned}$$

as claimed. ∎

We then have

Lemma 5. *Assume (C1), define t_k by (3), and define the probability measure T_k by $T_k(A) = \int_A t_k(x)\, \nu(dx)$, so that $\frac{dT_k}{d\nu} = t_k$. Then $\frac{k-1}{k} T_k + \frac{1}{k} T_k P = \pi$.*

Proof. From the previous lemma, the density with respect to $\nu(\cdot)$ of $\frac{k-1}{k} T_k + \frac{1}{k} T_k P$ is given by

$$
\begin{aligned}
& \frac{k-1}{k} t_k + \frac{1}{k} Q t_k \\
= \ & \left(\frac{k-1}{k}\right) \frac{k}{k-1} \Big(f - \frac{1}{k-1} Q f + \frac{1}{(k-1)^2} Q^2 f - \frac{1}{(k-1)^3} Q^3 f + \ldots \Big) \\
& + \left(\frac{1}{k} Q\right) \frac{k}{k-1} \Big(f - \frac{1}{k-1} Q f + \frac{1}{(k-1)^2} Q^2 f - \frac{1}{(k-1)^3} Q^3 f + \ldots \Big) \\
= \ & \Big(f - \frac{1}{k-1} Q f + \frac{1}{(k-1)^2} Q^2 f - \frac{1}{(k-1)^3} Q^3 f + \ldots \Big) \\
& + \Big(\frac{1}{k-1} Q f - \frac{1}{(k-1)^2} Q^2 f + \frac{1}{(k-1)^3} Q^3 f - \ldots \Big) \\
= \ & f \, .
\end{aligned}
$$

Since f is the density with respect to $\nu(\cdot)$ of $\pi(\cdot)$, the result follows. ∎

Proof of Theorem 1.

We know from Proposition 3 that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} P(X_i \in A) \ = \ \frac{k-1}{k} T_k(A) + \frac{1}{k} (T_k P)(A) \, .$$

But the previous lemma says that

$$\pi(A) \ = \ \frac{k-1}{k} T_k(A) + \frac{1}{k} (T_k P)(A) \, .$$

The main result follows. The statement about finite $\mathcal{X}$ follows from Theorem 6 below. ∎

We conclude from Theorem 1 that, assuming (C1), the BACKMET algorithm always achieves the correct limiting distribution $\pi(\cdot)$. Hence, the one remaining question is, under what circumstances will (C1) hold? We consider that in the following sections.

7. The Existence of Solutions.

The above result leads to the following question. Given a probability distribution $\pi(\cdot)$ and a Markov chain $P(x, \cdot)$ on $\mathcal{X}$, under what circumstances can we write $\pi = \frac{k-1}{k} T_k + \frac{1}{k}(T_k P)$ for some distribution T_k? Alternatively, for which $\pi(\cdot)$ and P can we be sure that condition (C1) holds, so that Theorem 1 applies and the BACKMET algorithm guarantees that $\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} P(X_i \in A) = \pi(A)$ for all A?

Obviously this answer depends on $\pi(\cdot)$, P and $\mathcal{X}$. In the case where $\mathcal{X}$ is finite, the answer is particularly simple.

Theorem 6. *Suppose the state space $\mathcal{X}$ is finite, and $\nu(\cdot)$ is counting measure. Then there exists k large enough that (C1) holds.*

Proof. Recall that for $k > 2$, $\sum_{j=1}^{\infty} (k-1)^{-j} = 1 \,/\, (k-2)$. Furthermore, since $\nu(\cdot)$ is counting measure, we see from Lemma 4 that $|(Q^j f)(x)| \le 1$ for all j and x. Hence, if we choose k large enough that $\min_{x\in\mathcal{X}} f(x) > 1 \,/\, (k-2)$, then we see by inspection of (3) and the triangle inequality, that the series for t_k will converge and will be non-negative. ∎

Theorem 6 shows that on a finite state space, provided one is willing to spend *some* positive limiting fraction of time in each state, one can always achieve any desired limiting distribution by use of a BACKMET algorithm for sufficiently large k. Recalling that the BACKMET algorithm does nothing for $k-2$ of every k steps, this says that by adding in enough "do nothing" steps, we can always achieve the limiting distribution $\pi(\cdot)$.

Remark. Another interpretation in the finite-state-space case is as follows. Using matrix notation, we wish to solve for $tM = \pi$, where t and π are row vectors, and M is the matrix having entries $M_{ij} = \frac{k-1}{k}\delta_{ij} + \frac{1}{k} P(i,j)$. Hence, assuming M is invertible (which it often will be), a solution is $t = \pi M^{-1}$. The question then becomes whether or not πM^{-1} has all non-negative entries (which it often won't, but by Theorem 6 will for large enough k).

If $\mathcal{X}$ is *infinite*, even countably infinite, then the situation is more complicated. This is because even if $\pi(\{x\}) > 0$ for all $x \in \mathcal{X}$, we will still have $\inf_{x\in\mathcal{X}} \pi(\{x\}) = 0$, so that the non-negativity of t_k as defined in (3) is not guaranteed.

Numerical experiments suggest that, indeed, for moderate values of k and distributions $\pi(\cdot)$ chosen at random, typically t_k will *not* be non-negative, so that BACKMET will not always work.

To make further progress, we shall assume for the remainder of this paper that the Markov chain P corresponds to an actual random walk. That is, we shall assume that the state space has a group structure (e.g. is the real numbers, or the integers, or the integers mod n for some n, or a continuous circle), and that P represents a *convolution* (i.e. $(\rho P)(A) = (\mu * \rho)(A) \equiv \int_{x \in \mathcal{X}} \mu(A - x)\, \rho(dx)$ for all ρ for some fixed probability measure μ, where $A - x = \{y \in \mathcal{X};\ y = a - x$ for some $a \in A)$.

In this random walk case, we see that $\frac{k-1}{k} T + \frac{1}{k}(TP) = p * T$ where $p = \frac{k-1}{k}\delta_0 + \frac{1}{k}\mu$. So the question becomes, given the probability measures μ and π, does there exist a probability measure T such that $p * T = \pi$, where $p = \frac{k-1}{k}\delta_0 + \frac{1}{k}\mu$?

This is a type of *deconvolution* question (though "deconvolution" often implies a more statistical context). Much is known about such convolution and deconvolution questions, and we pursue them here. We begin with some background about convolution on different spaces.

8. Background about Convolution.

The class of all distributions on R, denoted here by $\mathcal{P}$, form an abelian (i.e. commutative) semigroup under convolution; the unit element is the distribution whose value is 0 with probability 1. Clearly no distribution in $\mathcal{P}$ has an inverse except for the *trivial* distributions, i.e. those distributions which are constant with probability 1. By a *prime* (or *indecomposable*) distribution is meant one that cannot be written as a convolution of two nontrivial distributions. The prime distributions on R are "quite abundant"; in particular, they are dense in $\mathcal{P}$ in the topology of weak convergence.

Convolution of distributions in $\mathcal{P}$ corresponds, of course, to multiplication of their Fourier transforms, i.e., of their characteristic functions. For further background, see for example Cuppens (1975), Feldman (1993), Lukacs (1970), Linnik and Ostrovskii (1977), and Ostrovskii (1977, 1987). For distributions on abelian groups other than R, the situation is summed up in Feldman (1993). Sherstnev (1999) discusses all factorizations of the uniform distribution for the finite abelian groups.

In addition to convolution on R, there are three other contexts which interest us here: the circle group consisting of the real numbers modulo 2π, the group of all integers, and the cyclic groups consisting of the integers modulo n. These are discussed further below.

8.1. The Continuous Case.

Consider probability densities $p(x)$, $q(x)$, and $r(x)$ on R which are related as $p * q = r$, with p and r being given, and q being sought. (Here $(p * q)(x) = \int_{y \in R} p(x - y)\, q(y) dy$ for $x \in R$.) Define the characteristic function

$$\hat{p}(t) = \int_{-\infty}^{\infty} p(x) e^{itx}\, dx$$

and likewise for $\hat{q}(t)$ and $\hat{r}(t)$. Then the relation $p * q = r$ is equivalent to $\hat{p}(t)\hat{q}(t) = \hat{r}(t)$ for all $t \in R$. Hence $\hat{q}(t) = \hat{r}(t)/\hat{p}(t)$ (provided that $\hat{p}(t) \neq 0$), and therefore by Fourier inversion,

$$q(x) \;=\; \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{\hat{r}(t)}{\hat{p}(t)} e^{-itx}\, dt\,. \tag{5}$$

Thus, given densities p and r, to determine whether or not there exists a density $q(x)$ such that $p * q = r$, we just perform the computation (5) and then check to see if the resulting $q(x)$ is nonnegative. (Note that the resulting $q(x)$ will necessarily be real valued and integrate to one.) Of course, it is a necessary condition that the modulus of $\hat{r}(t)/\hat{p}(t)$ should not exceed one. In the event that $\hat{p}(t)$ has an isolated zero, then for it to be possible for a solution to exist, $\hat{r}(t)$ must also be zero there, and it is also necessary that $\hat{r}(t)/\hat{p}(t)$ can be defined by continuity there. If $\hat{p}(t) = 0$ in an interval then we must also have $\hat{r}(t) = 0$ there, and if a solution exists it need no longer be unique. Finding solutions in such cases is not straightforward. Note that the computation shown in (5) assumes that $\hat{r}(t)/\hat{p}(t)$ is in $L^1(R)$. If it is not then a standard limiting variant of this equation must be used; see the general inversion formulas in many probability texts.

The approach described in the last paragraph uses Fourier transformation in an explicit way. Strassen (1963, Theorem 10) provides alternative conditions on p and r – not involving Fourier transforms – for the existence of a probability density q which solves $p * q = r$. Strassen's conditions are that for every bounded and continuous real function $h(x)$,

$$\int h(x) r(x)\, dx \leq \sup_y \int h(x + y) p(x)\, dx\,.$$

Unfortunately these conditions are not straightforward to verify, although they may sometimes be used in a simple way to show the nonexistence of solutions.

Now consider probability densities on the circle group $\mathcal{C}$ of real numbers modulo T, for some $T > 0$. A probability density is a nonnegative function $p(t)$ on $[0, T]$ which integrates to one on this interval. Convolution is defined circularly, similarly to before. Here the role of characteristic function is played by the sequence of Fourier coefficients

$$\hat{p}(j) = \frac{1}{T} \int_0^T p(t) e^{-2\pi i j t/T}\, dt\,, \qquad j = 0, \pm 1, \pm 2 \cdots,$$

while the inverse is

$$p(t) = \sum_{j=-\infty}^{\infty} \hat{p}(j) e^{2\pi i j t/T} .$$

This assumes that $\hat{p}(j)$ is in ℓ^1. The convolution $p * q = r$ corresponds to $\hat{p}(k)\hat{q}(k) = \hat{r}(k)$, $k = 0, \pm 1, \pm 2, \ldots$. The considerations in this case are similar to those above and are omitted.

8.2. The Discrete Case.

We now consider the case of probability distributions on the integers modulo n. Such a distribution p is specified by nonnegative probabilities $p(j)$, $j = 0, 1, \cdots, n-1$. Given two such probability distributions $p(j)$ and $q(j)$, their convolution $r = p * q$ on the integers modulo n is defined as $r(j)$ where $r(j) = \sum_{\ell=0}^{n-1} p(j-\ell)q(\ell)$ (here $j - \ell$ is interpreted modulo n to lie between 0 and $n-1$); this corresponds to summation (modulo n) of independent random variables on this group. The characteristic function corresponding to a distribution p on this group may be taken to be the discrete Fourier transform $\hat{p}(k)$, $k = 0, 1, \cdots, n-1$ where

$$\hat{p}(k) = \sum_{j=0}^{n-1} p(j) \exp\left\{ \frac{-2\pi i j k}{n} \right\} .$$

The inverse DFT is given by

$$p(j) = \frac{1}{n} \sum_{k=0}^{n-1} \hat{p}(k) \exp\left\{ \frac{2\pi i j k}{n} \right\}$$

for $j = 0, 1, \cdots, n-1$. The modulo n convolution equation $p * q = r$ is then equivalent to $\hat{p}(k)\hat{q}(k) = \hat{r}(k)$, $k = 0, 1, \cdots, n-1$. For given p and r, one may determine if there exists a distribution q satisfying $p * q = r$ by just computing

$$q(j) = \frac{1}{n} \sum_{k=0}^{n-1} \frac{\hat{r}(k)}{\hat{p}(k)} \exp\left\{ \frac{2\pi i j k}{n} \right\}$$

and checking if it is nonnegative. (As before, this function necessarily is real valued and sums to one.) If $\hat{p}(k)$ has zeros, then $\hat{r}(k)$ must also have zeros there, and determining solutions to $p * q = r$ must then be carried out otherwise. Strassen's Theorem applied to this case gives the condition

$$\sum h(k) r(k) \leq \sup_{\ell} \sum h(k+\ell) p(k)$$

for all real finite sequences $h(k)$.

We now consider the case of probability distributions on the set of all integers. Such a distribution p is specified by nonnegative probabilities $p(j)$, $j \in Z$, which sum to 1. Given two such probability distributions $p(j)$ and $q(j)$, their convolution $r = p * q$ on the integers is defined as $r(j)$ where $r(j) = (p*q)(j) = \sum_{\ell \in Z} p(j-\ell)q(\ell)$. The characteristic function corresponding to a distribution p on this group may be taken as $\hat{p}(t) = \sum_{j \in Z} p(j) \exp\{-ijt\}$, $t \in [-\pi, \pi]$, with inverse transform $p(j) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \hat{p}(t) \exp\{-ijt\}\, dt$, $j \in Z$. The considerations in this case are similar to those above and are omitted.

9. Necessary Conditions for Convolution Solutions.

In this section, we present a number of necessary conditions (some also sufficient) for the existence of non-negative solutions to convolution questions. From the point of view of the BACKMET algorithm, these conditions provide us with circumstances in which BACKMET *cannot* be applied.

We will be dealing, in each case, with probability densities p, q, and r defined on an abelian group, and with the convolution equation

$$p * q = r\,. \tag{6}$$

In each case, the probability densities themselves (under convolution) are an abelian semigroup with unit element. Also in each case, p and r will be given, and we are interested in whether or not there exists a probability q such that (6) holds.

From the point of view of the BACKMET algorithm, here p corresponds to the action of the random walk, q corresponds to the target density t_k as in the Metropolis-Hastings algorithm, and r corresponds to the density of the target distribution $\pi(\cdot)$. Further connections to BACKMET will be given in the next section.

Due to essential differences in the results for noncompact, compact, countable, and finite state spaces, it is necessary to provide separate statements of the results for each of the groups R, Z, $\mathcal{C}$, and $Z(\text{mod } n)$.

We begin with the state space R; analogous results apply to R^d.

Theorem 7. *For given probability densities r and p defined on R, the existence of a probability density q on R which is a solution of (6) is equivalent to either of the following two conditions:*
(a) There exists a density q on R such that $\hat{p}(t)\hat{q}(t) = \hat{r}(t)$ for all $t \in R$, where 'hat' refers to Fourier transformation on R, or equivalently, $\hat{r}(t)/\hat{p}(t)$ can be interpolated continuously (wherever undefined) to a function which is nonnegative definite on R.

(b) For every bounded and continuous real function h

$$\int_R h(x)\, r(x)\, dx \le \sup_y \int_R h(x+y)\, p(x)\, dx\,;$$

here we may restrict to functions such that $0 \le \inf h \le \sup h \le 1$.

In particular, equation (6) cannot be solved for a probability q on R if any of the following conditions holds:

(i) If $|\hat{r}(t)|$ exceeds $|\hat{p}(t)|$ for some $t \in R$; If $\hat{p}(t)$ has a zero at some t where $\hat{r}(t)$ does not; If p is smoother than r – more precisely, if $\hat{p}(t)$ tends to zero, as $|t| \to \infty$, at a rate faster than $\hat{r}(t)$. [Note that the degree of differentiability of a density function is related to the rate of decay of its characteristic function.]

(ii) If the density r has an essential zero (i.e. does not have full support) while the density p does not; If p and r are regarded as Lebesgue densities and $\sup r(x)$ essentially exceeds $\sup p(x)$; If the measure of the support of p exceeds that of r;

(iii) If the tails of r descend to zero more quickly than the tails of p;

(iv) If the variance (or scaling) of p exceeds that of r; If any absolute moment of order at least one about the mean of p exceeds the corresponding moment of r; If the entropy of p exceeds that of r (the entropy of a density f is defined by $-\int f \log f$);

(v) If r is a normal, Cauchy, Holtsmark, Levy-Smirnov or other stable law, and p is not a stable law with the same index and skewness parameters; If r is known to be a prime (i.e. indivisible) density on R, and p is any other nontrivial density; If all factors of r are known, and p is not among them.

Proof of Theorem 7. Assertion (a) is the standard Fourier result for convolutions. Assertion (b) is due to Strassen (1963, Theorem 10); permissibility of the restriction on h is elementary. The assertions in part (i) all follow from the fact the moduli of the Fourier transforms are bounded by 1. The middle assertion in part (ii) is proved via Strassen's inequality by allowing h to 'concentrate' near the essential supremum of r. The remaining two assertions of (ii) may be proved using the method of Loeve, Ch. V, Ex. 11. Part (iii) is established by straightforward analysis showing that the tails of a convolution are at least as heavy as those of its factors (and may also be established via Loeve, Ch. V, Ex. 11). Part (iv) is standard; see also Loeve, Ch. V, Ex. 3. The entropy result follows from the *entropy power inequality*, see e.g. Theorem 4 of Dembo and Cover (1991). Part (v) follows from basic results in Lukacs (1970, Chapters 5 and 6). ∎

We next consider the integers, Z; analogous results apply to Z^d.

Theorem 8. *For given probability sequences r and p defined on Z, the existence of a probability sequence q on Z which is a solution of (6) is equivalent to either of the following two conditions:*
(a) There exists a probability sequence q on Z such that $\hat{p}(t)\hat{q}(t) = \hat{r}(t)$ for all $t \in R \bmod 2\pi$, where 'hat' refers to Fourier transformation on Z, or equivalently, $\hat{r}(t)/\hat{p}(t)$ can be extended by continuous interpolation, wherever undefined, to a nonnegative definite function on $R \bmod 2\pi$.
(b) For every bounded sequence $h(k)$

$$\sum h(k)r(k) \leq \sup_{\ell} \sum h(k+\ell)p(k)\,;$$

here we may restrict to sequences such that $0 \leq \inf h \leq \sup h \leq 1$.

In particular, equation (6) cannot have a solution for a probability sequence q on Z if any of the following conditions holds:
(i) If $|\hat{r}(t)|$ exceeds $|\hat{p}(t)|$ for some $t \in T$; If $\hat{p}(t)$ has a zero for some $t \in T$ where $\hat{r}(t)$ does not;
(ii) If r has a zero (i.e. a point of nonsupport) and p does not; If p contains a sequence of consecutive nonzero values longer than any such sequence for r; If $\sup r(k)$ exceeds $\sup p(k)$; If any 'averaged' value of $r(k)$, based on uniformly bounded weights, exceeds the largest similarly weighted average among the arbitrarily translated versions of $p(x)$;
(iii) If the tails of r descend to zero more quickly than the tails of p;
(iv) If the variance (or scaling) of p exceeds that of r; If any absolute moment of order at least one about the mean of p exceeds the corresponding moment of r; If the entropy of p exceeds that of r.
(v) If r is a Poisson distribution, and p is not; If r is known to be a prime (i.e. indivisible) distribution on Z, and p is nontrivial; If all factors of r are known, and p is not among them.

Proof. Note that the last assertion of (ii) is just Strassen's condition. The proof is otherwise similar to that of Theorem 7 and is omitted. ∎

We next consider the circle group, $\mathcal{C}$; analogous results apply to other continuous compact abelian groups.

Theorem 9. *For given probability densities r and p defined on $\mathcal{C}$, the existence of a probability density q on $\mathcal{C}$ which is a solution of (6) is equivalent to either of the following two conditions:*

(a) There exists a density q on $\mathcal{C}$ such that $\hat{p}(k)\hat{q}(k) = \hat{r}(k)$ for all $k \in Z$, where 'hat' refers to Fourier transformation on $\mathcal{C}$, or equivalently, $\hat{r}(k)/\hat{p}(k)$ can be extended, where undefined, to a nonnegative definite function on Z.
(b) For every bounded function $h(t)$ on $\mathcal{C}$

$$\int_{\mathcal{C}} h(t)r(t)\,dt \leq \sup_u \int_{\mathcal{C}} h(t+u)p(t)\,dt\,;$$

here we may restrict to functions such that $0 \leq \inf h \leq \sup h \leq 1$.

In particular, equation (6) cannot have a solution for a probability q on $\mathcal{C}$ if any of the following conditions holds:
(i) If $|\hat{r}(k)|$ exceeds $|\hat{p}(k)|$ for some $k \in Z$; If $\hat{p}(k)$ is zero for some $k \in Z$ where $\hat{r}(k)$ is not; If p is smoother than r – more precisely, if $\hat{p}(k)$ tends to zero, as $|k| \to \infty$, at a rate faster than $\hat{r}(k)$.
(ii) If the density r has an essential zero and p does not; If the measure of any subset of the support of p exceeds that of every subset of the support of r; If $\sup r(t)$ exceeds $\sup p(t)$, or $\inf p(t)$ exceeds $\inf r(t)$;
(iii) If r is known to be a prime (i.e. indivisible) density on $\mathcal{C}$, and p is nontrivial; If all factors of r are known, and p is not among them.

Proof. The proof is as before except using results for circular convolution. The supremum and infimum conditions in (ii) follow from the Strassen inequalities. ∎

Finally, we consider the integers modulo n for some fixed n; analogous results apply to any finite abelian group.

Theorem 10. *For given finite probability sequences r and p defined on $Z \bmod n$, the existence of a probability sequence q on $Z \bmod n$ which is a solution of (6) is equivalent to either of the following two conditions:*
(a) There exists a distribution q on $Z \bmod n$ such that $\hat{p}(k)\hat{q}(k) = \hat{r}(k)$ for $k \in Z \bmod n$, where 'hat' refers to discrete Fourier transformation, or equivalently, $\hat{r}(k)/\hat{p}(k)$ can be extended, where undefined, to a nonnegative definite function on $Z \bmod n$.
(b) For every finite sequence $h(k)$ on $Z \bmod n$

$$\sum h(k)r(k) \leq \sup_\ell \sum h(k+\ell)p(k)\,;$$

we may restrict to sequences such that $0 \leq \min h \leq \max h \leq 1$.

In particular, equation (6) cannot have a solution for a probability q on $Z \bmod n$ if any of the following conditions hold:

(i) If $|\hat{r}(k)|$ exceeds $|\hat{p}(k)|$ for some $k \in Z$ mod n; If $\hat{p}(k)$ has a zero at some $k \in Z$ mod n where $\hat{r}(k)$ does not;

(ii) If r has a zero and p does not; If the length of a consecutive sequence of nonzero values of p exceeds any such of r; If $\sup r(k)$ exceeds $\sup p(k)$ or $\inf p(k)$ exceeds $\inf r(k)$; If any 'averaged' value of $r(k)$ exceeds the largest similarly weighted average among arbitrarily rotated versions of $p(x)$;

(iii) If r is known to be a prime (i.e. indivisible) distribution on Z mod n, and p is nontrivial; If all factors of r are known, and p is not among them.

Proof of Theorem 10. The last condition in (ii) is just Strassen's inequality. The remainder of the proof is similar to the above and is omitted. ∎

Remarks on Theorem 10:

(1) A complete solution is available for the uniform case when the $r(k)$ are all equal; see Sherstnev (1999) for all factorizations of r for this case.

(2) The problem of decomposition into prime factors for 'equally spaced' discrete distributions having equal probabilities is discussed in Lukacs (1970, Sect. 6.3).

10. Convolution with Perturbations of the Identity.

We now consider the effects in the BACKMET algorithm of allowing k to be large. Specifically, we consider the problem $p * q = r$ in the case when $p = \alpha\mu + (1-\alpha)\delta_0$, where δ_0 is the identity (i.e. the distribution assigning unit mass to zero), so that $p*q = \alpha\mu*q+(1-\alpha)q$. We shall allow $\alpha > 0$ to be as small as required. That is, given μ and r, we seek solutions q to the equation $\alpha\mu * q + (1-\alpha)q = r$, for $\alpha > 0$ sufficiently small.

We now present our main result for this problem. It says, essentially, that on the compact groups Z mod n or the circle group $\mathcal{C}$, a solution exists provided the target density r is bounded away from 0. On the unbounded groups Z or R, a solution exists provided r is bounded away from 0 on compact sets, and furthermore a tail condition (7) is satisfied, which says essentially that the tails of r must be at least as heavy as those of μ. A formal statement is as follows.

Theorem 11. *Suppose $p = \alpha\mu + (1-\alpha)\delta_0$, that r and μ are given, and that we seek solutions q of $p * q = r$ for $\alpha > 0$ sufficiently small. Then*

(a) On the finite state space Z mod n, if $r > 0$ then a solution for q is always possible provided α is small enough. (The positivity condition on r can be relaxed under associated conditions on μ.)

(b) On the compact state space $\mathcal{C}$, if $a < r < b$ for some $0 < a \leq b < \infty$ then a solution for q is always possible provided α is small enough. (The strict positivity of r can be somewhat relaxed via related conditions on μ.)
(c) On the state space R, if r is bounded away from 0 on all compact sets, and further, if the tails of r are at least as heavy as the tails of μ in the sense that there exists an $\epsilon > 0$ such that

$$r - \epsilon \cdot r * \mu \geq 0\,, \tag{7}$$

then a solution for q is always possible provided α is small enough. The condition (7) is met, for example, if additionally r is bounded above, and if for $|x|$ sufficiently large the densities r and μ have tails of the type $c_1 \cdot f(|x|) \leq r(x) \leq c_2 \cdot f(|x|)$ and $d_1 \cdot g(|x|) \leq \mu(x) \leq d_2 \cdot g(|x|)$ for some constants $0 < c_1 \leq c_2 < \infty$ and $0 < d_1 \leq d_2 < \infty$, and for some functions f, g which are monotone decreasing on R^+ and satisfy $\limsup_{x\to\infty} g(x)/f(x) < \infty$.
(d) On the countable state space Z, if $r > 0$ everywhere and the tail condition (7) holds, then a solution for q is always possible provided α is small enough. The tail condition (7) is satisfied under tail conditions on r and μ analogous to those given in part (c).

Proof of Theorem 11. Proceeding formally at first, write $(\alpha\mu + (1-\alpha)\delta_0) * q = r$ as

$$q = \frac{1}{1-\alpha}\left(\delta_0 + \frac{\alpha}{1-\alpha}\,\mu\right)^{-1} * r\,;$$

and (regarding δ_0 as an identity) expand the inverse term appearing in this expression as

$$\delta_0 - \left(\frac{\alpha}{1-\alpha}\right)\mu + \left(\frac{\alpha}{1-\alpha}\right)^2 \mu * \mu - \left(\frac{\alpha}{1-\alpha}\right)^3 \mu * \mu * \mu + \cdots$$

to obtain

$$q = \frac{1}{1-\alpha}\left\{r - \left(\frac{\alpha}{1-\alpha}\right) r * \mu + \left(\frac{\alpha}{1-\alpha}\right)^2 r * \mu * \mu - \left(\frac{\alpha}{1-\alpha}\right)^3 r * \mu * \mu * \mu + \cdots\right\}. \tag{8}$$

This solution is, of course, analogous to the series (3) used earlier, with $\alpha = 1/k$.

Next, observe that (8) is, in fact, a solution of $(\alpha\mu + (1-\alpha)\delta_0) * q = r$; thus it is enough only to establish that it is nonnegative. Note that the terms in (8) can be collected pairwise so that (8) consists of $r - \{\alpha/(1-\alpha)\}\, r * \mu$ plus a sum of further terms which, being convolutions of this first term, are each nonnegative whenever the first term is nonnegative. On the compact state spaces Z (mod n) and $\mathcal{C}$, the term $r - \{\alpha/(1-\alpha)\}\, r * \mu$ is clearly nonnegative, under the stated conditions, when α is sufficiently small. For state spaces R and Z the same also holds in view of the assumed condition (7); the example stated in terms of the tail functions f and g involves only routine analysis after noting that boundedness of

r implies boundedness of $r * \mu$. ∎

Remarks on Theorem 11:

(1) While the emphasis in Theorem 11 is on stating positive results — i.e. cases under which solutions are possible — we remark that the results of Theorems 7 through 10 provide a substantive basis for constructing many examples for which there are no solutions to the problem addressed in Theorem 11.

(2) The conditions indicated in the cases (c) and (d) of the Theorem are satisfied, for example, if $c_1 \cdot |x|^{-\beta} \leq r(x) \leq c_2 \cdot |x|^{-\beta}$ as $|x| \to \infty$ with a similar condition for μ, provided the index β corresponding to r is not less than that for μ. Similar conditions can also be stated for tails of exponential type, various combination or mixed tails, and for differing upper and lower tail types.

(3) Note that when condition (7) fails then the power series (8) will always violate non-negativity for α sufficiently small. Hence the sufficient (7) is also close to being necessary.

(4) For the continuous state spaces R and $\mathcal{C}$, since for any $\alpha > 0$ the tails of the Fourier transform $\hat{p}$ will be bounded away from zero, the rate of descent of $\hat{r}$ must necessarily be the same as that of the transform of the solution, $\hat{q}$.

(5) A problem similar to that of Theorem 11 was mentioned (though not analysed) as far back as 1938; see Problem 174 of Mauldin (1981).

Applying Theorem 11 to the BACKMET algorithm, we immediately conclude the following positive results. They generalise Theorem 6 to infinite state spaces, at least in the convolution case. Once again, we state the results specifically for the four abelian groups studied above, but analogous results apply to other situations.

Theorem 12. *Suppose the state space $\mathcal{X}$ is one of Z mod n, Z, R, or $\mathcal{C}$, and the underlying Markov chain P corresponds to a random walk (i.e. convolution). Assume the chain is ϕ-irreducible. Assume that $\pi(\cdot)$ has a density (with respect to counting measure for Z or Z mod n; or Lebesgue measure for R or $\mathcal{C}$) which is bounded away from 0, and is also bounded above (for $\mathcal{C}$) or satisfies the tail condition (7) (for Z or R). Then there exists k large enough, and t_k as given by (3), such that the BACKMET algorithm succeeds in ensuring that $\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} P(X_i \in A) = \pi(A)$ for all A. [See also the remarks following Theorem 1.]*

11. Summary.

We summarise our results as follows.

Our goal was to find an algorithm which uses the "back" button to achieve a prescribed limiting distribution $\pi(\cdot)$, when given Markov chain transition probabilities $P(x, \cdot)$, while using as little unobserved information as possible.

To accomplish this, we introduced the BACKMET algorithm. It requires a single global function t_k, but then only evaluates this function at previously observed states. We proved (Theorem 1) that this algorithm achieves the limiting distribution $\pi(\cdot)$ whenever the non-negativity condition (C1) holds.

We proved (Theorem 6) that condition (C1) always holds for *finite* state spaces, provided only that the target distribution $\pi(\cdot)$ is everywhere-positive.

We then considered the case where the Markov chain is a random walk (corresponding to convolution of densities), and proved (Theorem 11(b); Theorem 12) that more generally, condition (C1) always holds for random walks on *compact* spaces such as the circle group $\mathcal{C}$, provided only that the target distribution $\pi(\cdot)$ has density which is bounded away from 0 and also bounded above.

We also proved (Theorem 11(c),(d); Theorem 12) that condition (C1) always holds for random walks on unbounded spaces such as R and Z, provided the target density is bounded away from 0 on all compact sets, and also the tail condition (7) is satisfied (so that, roughly, the target density tails must be at least as heavy as the tails of the random walk jumps).

On the other hand, we have also provided (Theorems 7 through 10) quite a number of conditions which are *necessary* for the convolution equation to be solvable. This shows that, in general, for a given k and $\pi(\cdot)$ and p, the convolution equation will often not be solvable.

We feel that our results, taken together, give a fairly complete picture of when the BACKMET algorithm will or will not work. However, a number of questions remain, including what happens if the target density is not bounded away from 0 (even on compact sets), or if the state space is infinite and the Markov chain is not a random walk. It is possible that other algorithms, besides BACKMET, could be used in such cases. Similarly, the BACKMET algorithm is somewhat inefficient in that it often does not move, and one can ask whether it can be made more efficient, and if so, then how.

REFERENCES

Cuppens, R. (1975). Decomposition of Multivariate Probability. Academic Press, New York.

Dembo, A. and Cover T.M. (1991). Information theoretic inequalities. IEEE Transact. on Inform. Th. **37**, 1501–1518.

Dongarra, J. and Sullivan, F., eds. (2000), The Top 10 Algorithms. Computing in Science and Engineering **2(1)**.

Fagin, R., Karlin, A.R., Kleinberg, J., Raghavan, P., Rajagopalan, S., Rubinfeld, R., Sudan, M., and Tomkins, A. (2001), Random Walks with "Back Buttons". Ann. Appl. Probab. **11**, 810–862.

Hastings, W.K. (1970), Monte Carlo sampling methods using Markov chains and their applications. Biometrika **57**, 97–109.

Linnik, Yu.V. and Ostrovskii, I.V. (1977). The Decomposition of Random Variables and Random Vectors. American Mathematical Society.

Loeve, M. (1963). Probability Theory, 3rd ed. Van Norstrand.

Lukacs, E. (1970). Characteristic Functions. 2nd ed. Hafner, New York.

Mauldin, R.D., ed. (1981), The Scottish book: mathematics from the Scottish Café. Birkhäuser, Boston.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., and Teller, E. (1953), Equations of state calculations by fast computing machines. J. Chem. Phys. **21**, 1087–1091.

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

Ostrovskii, I.V. (1977). The arithmetic of probability distributions. J. Multiv. Anal. **7**, 475–490.

Ostrovskii, I.V. (1987). The arithmetic of probability distributions. Theory Probab. Appl. **31**, 1–24.

Sherstnev, V.I. (1999). A random variable uniformly distributed on a finite abelian group as a sum of independent summands. Th. Probab. and its Applic. SIAM Translation.

Strassen, V. (1963). The existence of probability measures with given marginals. Ann. Math. Statist. **36**, 423–439.

Tierney, L. (1994), Markov chains for exploring posterior distributions (with discussion). Ann. Stat. **22**, 1701–1762.