Norm Comparisons for Data Augmentation

by

James P. Hobert
Department of Statistics
University of Florida

and

Jeffrey S. Rosenthal
Department of Statistics
University of Toronto

Technical Report No. 0704 February 25, 2007

Norm Comparisons for Data Augmentation

by

James P. Hobert[1] and Jeffrey S. Rosenthal[2]

(February 2007)

1 Introduction

This short paper considers comparisons of different data augmentation algorithms in terms of their convergence and efficiency. It examines connections between the partial order $\preceq_1$ on Markov kernels, and inequalities of operator norms. It applies notions from Roberts and Rosenthal (2006) related to variance bounding Markov chains, together with L2 theory, to data augmentation algorithms (Tanner and Wong, 1987; Liu and Wu, 1999; Meng and van Dyk, 1999; Hobert and Marchev, 2006). In particular, our main result, Theorem 10, is a direct generalisation of one of the theorems in Hobert and Marchev (2006).

2 Background and Notation

Let $\pi(\cdot)$ be a probability measure on a measurable space $(\mathcal{X}, \mathcal{F})$. For measurable $f : \mathcal{X} \to \mathbf{R}$, write $\pi(f) = \int_{\mathcal{X}} f \, d\pi$. Let

$$L^2(\pi) = \{f : \mathcal{X} \to \mathbf{R} \ s.t. \ f \text{ measurable and } \pi(f^2) < \infty\},$$

$L^2_0(\pi) = \{f \in L^2(\pi) \ s.t. \ \pi(f) = 0\}$, and $L^2_{0,1}(\pi) = \{f \in L^2_0(\pi) \ s.t. \ \pi(f^2) = 1\}$. For $f, g \in L^2(\pi)$, write $\langle f, g\rangle = \int_{\mathcal{X}} f(x)\, g(x)\, \pi(dx)$, and $\|f\| = \sqrt{\langle f, f\rangle}$.

Let P be a Markov chain operator on $(\mathcal{X}, \mathcal{F})$. For a measure μ on $(\mathcal{X}, \mathcal{F})$, write μP for the measure on $(\mathcal{X}, \mathcal{F})$ defined by $(\mu P)(A) = \int_{\mathcal{X}} \mu(dy)\, P(y, A)$ for $A \in \mathcal{F}$. For a measurable function $f : \mathcal{X} \to \mathbf{R}$, write Pf for the measurable function defined by $(Pf)(x) = \int_{\mathcal{X}} f(y)\, P(x, dy)$ for $x \in \mathcal{X}$. Write $\|P\|$ for the norm of the operator P restricted to $L^2_0(\pi)$, i.e. $\|P\| = \sup\{\|Pf\| \ s.t. \ f \in L^2_{0,1}(\pi)\}$.

[1] Department of Statistics, University of Florida, P.O. Box 118545, Gainesville, FL 32611-8545, U.S.A. Internet: jhobert@stat.ufl.edu. Supported in part by NSF Grant DMS-05-03648.

[2] Department of Statistics, University of Toronto, Toronto, Ontario, Canada M5S 3G3. Internet: jeff@math.toronto.edu. Supported in part by NSERC of Canada.

The Markov chain operator P has stationary distribution $\pi(\cdot)$ if $\pi P = \pi$. P is *reversible* (with respect to $\pi(\cdot)$) if $\pi(dx)\, P(x, dy) = \pi(dy)\, P(y, dx)$ as measures on $\mathcal{X} \times \mathcal{X}$, or equivalently if P is a self-adjoint operator on $L^2(\pi)$. If P is reversible with respect to $\pi(\cdot)$, then P has stationary distribution $\pi(\cdot)$ (see e.g. Roberts and Rosenthal, 2004).

In terms of a Markov chain $\{X_n\}_{n=0}^{\infty}$ following the transitions P in stationarity, so $\mathcal{L}(X_n) = \pi(\cdot)$ and $\mathbf{P}[X_{n+1} \in A \,|\, X_n] = P(X_n, A)$ for all $A \in \mathcal{F}$ and all $n \in \mathbf{N}$, we have the interpretations $(Pf)(x) = \mathbf{E}[f(X_1) \,|\, X_0 = x]$, and $\langle f, g \rangle = \mathbf{E}[f(X_0)\, g(X_0)]$, and $\langle f, Pg \rangle = \mathbf{E}[f(X_0)\, (Pg)(X_0)] = \mathbf{E}[f(X_0)\, g(X_1)]$.

For a reversible Markov chain operator P on $L^2(\pi)$, write $\sigma(P)$ for the spectrum of P restricted to $L_0^2(\pi)$. Let $m_P = \inf \sigma(P)$, and $M_P = \sup \sigma(P)$. A reversible operator P is *positive* iff $m_P \geq 0$, i.e. if $\langle Pf, f \rangle \geq 0$ for all f. The following properties follow from basic operator theory (e.g. Rudin, 1991; Chan and Geyer, 1994).

Proposition 1. *Let P be a reversible Markov chain operator. Then*
(a) $\sigma(P) \subseteq [-1, 1]$, *i.e.* $-1 \leq m_P \leq M_P \leq 1$;
(b) $\|P\| = \max(-m_P,\, M_P)$, *so in particular* $M_P \leq \|P\|$;
(d) $m_P = \inf\{\langle Ph, h \rangle$ *s.t.* $h \in \mathcal{L}_{0,1}^2(\pi)\}$;
(e) $M_P = \sup\{\langle Ph, h \rangle$ *s.t.* $h \in \mathcal{L}_{0,1}^2(\pi)\}$;
(f) $\|P\| = \sup\{|\langle Ph, h \rangle|$ *s.t.* $h \in \mathcal{L}_{0,1}^2(\pi)\}$.

A Markov kernel P is *geometrically ergodic* if there is π-a.e. finite $M : \mathcal{X} \to [0, \infty]$ and $\rho < 1$ such that $|P^n(x, A) - \pi(A)| \leq M(x)\, \rho^n$ for all $n \in \mathbf{N}$, $x \in \mathcal{X}$, and $A \in \mathcal{F}$. From Roberts and Rosenthal (1997) and the above, we obtain:

Proposition 2. *Let P be a reversible Markov chain operator. Then the following are equivalent:*
(a) *P is geometrically ergodic;*
(b) $\|P\| < 1$;
(c) $m_P > -1$ *and* $M_P < 1$;
(d) $\sigma(P) \subseteq [-r, r]$ *for some* $r < 1$.

Remark. On a *finite* state space, $m_P = -1$ if and only if -1 is an eigenvalue, which occurs if and only if P is periodic (with even period). However, on an infinite state space, P could have spectrum converging to -1, and thus have $m_P = -1$, even if P is not periodic and does not have an eigenvalue equal to -1.

Given a Markov operator P and a measurable function $f : \mathcal{X} \to \mathbf{R}$, the corresponding *asymptotic variance* is given by $\mathrm{Var}(f, P) = \lim_{n\to\infty} n^{-1}\mathrm{Var}(\sum_{i=1}^{n} f(X_i))$, where again $\{X_n\}$ follows the Markov chain in stationarity. A Markov operator P satisfies a central limit theorem (CLT) for f if $n^{-1/2}\sum_{i=1}^{n}[f(X_i) - \pi(f)]$ converges weakly to $N(0, \sigma_f^2)$ for some $\sigma_f^2 < \infty$. Kipnis and Varadhan (1986) (see also Chan and Geyer, 1994) prove that if P is reversible, and $\mathrm{Var}(f, P) < \infty$, then P satisfies a CLT for f, and furthermore $\sigma_f^2 = \mathrm{Var}(f, P)$.

Roberts and Rosenthal (2006) define a Markov operator P to be *variance bounding* if $\sup\{\mathrm{Var}(f, P)\ s.t.\ f \in L^2_{0,1}(\pi)\} < \infty$, and prove the following:

Proposition 3. *Let P be a reversible Markov chain operator. Then the following are equivalent:*
(a) $\mathrm{Var}(f, P) < \infty$ *for all* $f \in L^2(\pi)$*;*
(b) P *is variance bounding;*
(c) $M_P < 1$.

In particular, comparing Propositions 2(c) and 3(c) shows that if P is geometrically ergodic then it is variance bounding.

3 Partial orderings

Let P and Q be Markov operators on $(\mathcal{X}, \mathcal{F})$, each having stationary distribution $\pi(\cdot)$. Write $P \succeq_1 Q$ if for all $f \in L^2(\pi)$ (or, equivalently, for all $f \in L^2_0(\pi)$), we have $\langle f, Pf\rangle \le \langle f, Qf\rangle$.

Peskun (1973), Tierney (1998), and Mira and Geyer (1999, Theorem 4.2), see also Mira (2001), prove that if P and Q are reversible, then $P \succeq_1 Q$ if and only if $\mathrm{Var}(P, f) \le \mathrm{Var}(Q, f)$ for all $f \in L^2(\pi)$. In particular, it follows that if $P \succeq_1 Q$ and Q is variance bounding, then P is variance bounding. However, the corresponding property for geometric ergodicity does not hold. That is, if $P \succeq_1 Q$ and Q is geometrically ergodic, it does not necessarily follow that P is also geometrically ergodic (Roberts and Rosenthal, 2006). This illustrates the potential conflict between small variance and rapid convergence (Mira, 2001; Rosenthal, 2003).

Concerning operator norms, we have the following.

Proposition 4. *If R and S are reversible, and $R \succeq_1 S$, then $\|R\| \le \max(-m_R,\ \|S\|)$.*

Proof. We have $\|R\| = \max(-m_R,\ M_R) \le \max(-m_R,\ M_S) \le \max(-m_R,\ \|S\|)$. ■

Corollary 5. *If R and S are reversible, and R is positive, and $R \succeq_1 S$, then $\|R\| \le \|S\|$.*

Proof. Since R is positive, $m_R \geq 0$, so $\max(-m_R,\ \|S\|) = \|S\|$. ∎

It then follows from Proposition 2 that:

Corollary 6. *If R and S are reversible, and R is positive, and $R \succeq_1 S$, and S is geometrically ergodic, then R is geometrically ergodic.*

4 Data Augmentation Algorithms

Consider now the case where the state space is a product space, $(\mathcal{X}, \mathcal{F}) \times (\mathcal{Y}, \mathcal{G})$. Let $\mu(\cdot)$ and $\nu(\cdot)$ be some σ-finite reference measures on $\mathcal{X}$ and $\mathcal{Y}$ respectively (e.g. Lebesgue measure of appropriate dimension), and let $\pi(\cdot)$ be a probability measure on $\mathcal{X} \times \mathcal{Y}$ having (unnormalised) density w with respect to $\mu \times \nu$:

$$\pi(A \times B) \;=\; \frac{\int_{y \in B} \int_{x \in A} w(x,y)\, \mu(dx)\, \nu(dy)}{\int_{y \in \mathcal{Y}} \int_{x \in \mathcal{X}} w(x,y)\, \mu(dx)\, \nu(dy)}\,.$$

Also, let π_x and π_y denote the marginal measures on $(\mathcal{X}, \mathcal{F})$ and $(\mathcal{Y}, \mathcal{G})$, respectively; e.g., $\pi_x(A) = \pi(A \times \mathcal{Y})$.

The data augmentation algorithm (Tanner and Wong, 1984) may be defined as follows. Let P_1 be the Markov operator on $\mathcal{X} \times \mathcal{Y}$ which leaves y fixed while updating x from the conditional density given by w, i.e.:

$$P_1((x,y),\ A \times \{y\}) \;=\; \frac{\int_{x \in A} w(x,y)\, \mu(dx)}{\int_{x \in \mathcal{X}} w(x,y)\, \mu(dx)}\,, \qquad A \in \mathcal{F}\,. \tag{1}$$

Similarly, define P_2 by:

$$P_2((x,y),\ \{x\} \times B) \;=\; \frac{\int_{y \in B} w(x,y)\, \nu(dy)}{\int_{y \in \mathcal{Y}} w(x,y)\, \nu(dy)}\,, \qquad B \in \mathcal{G}\,. \tag{2}$$

Then the traditional data augmentation algorithm corresponds to the operator $P = P_2\, P_1$, i.e. the Markov chain which updates first y (with P_1) and then x (with P_2). (This is the systematic scan version; the random scan version is $P = \frac{1}{2}(P_1 + P_2)$ though we do not consider that here.)

A data-augmentation algorithm Markov operator P on $(\mathcal{X}, \mathcal{F}) \times (\mathcal{Y}, \mathcal{G})$ then induces a corresponding restricted Markov operator $\widehat{P}$ on $(\mathcal{X}, \mathcal{F})$, by $\widehat{P}(x, A) = P((x,y),\ A \times \mathcal{Y})$, equivalent to performing P as usual but keeping track of only the x coordinate. It is well-known and easy to show that $\widehat{P}$ is reversible with respect to π_x. (In the language of Roberts and Rosenthal, 2001, the individual chain $\{Y_n\}$ and the pair chain $\{(X_n, Y_n)\}$ are *co-de-initialising.*)

Amit (1991) and Liu, Wong and Kong (1994, Lemma 3.2) prove the following:

Proposition 7. *Let $\{(X_n, Y_n)\}$ follow a systematic scan data augmentation algorithm P, and let $f \in L_0^2(\pi_x)$. Then $\langle f, \widehat{P} f\rangle = \mathrm{Var}_\pi[\mathbf{E}_\pi(f(X) \,|\, Y)] \geq 0$.*

Proposition 7 immediately implies:

Corollary 8. *A Markov chain operator $\widehat{P}$ corresponding to a systematic scan data augmentation algorithm is positive.*

Hobert and Marchev (2006), following Liu and Wu (1999) and Meng and van Dyk (1999), generalise the data augmentation algorithm as follows. Let R be any Markov chain operator on $(\mathcal{Y}, \mathcal{G})$ having π_y as a stationary distribution. Extend this trivially to $(\mathcal{X}, \mathcal{F}) \times (\mathcal{Y}, \mathcal{G})$ by $\overline{R} = I \times R$, i.e.

$$\overline{R}((x, y),\ \{x\} \times B) \ = \ R(y, B)\,.$$

Then define $P_R = P_1 \overline{R}\, P_2$; intuitively, P_R corresponds to first updating y with P_2, then updating y with R, and then updating x with P_1. Let $\widehat{P}_R$ be the corresponding restricted operator on $\mathcal{X}$ as above. It is clear that π_x is a stationary distribution for $\widehat{P}_R$.

Say that P_R *is a DA algorithm* if there is some other density function w^* on $\mathcal{X} \times \mathcal{Y}$, that also yields π_x as the x-marginal, such that if P_1^* and P_2^* are defined by (1) and (2) but with w^* in place of w, then $P_R = P_2^*\, P_1^*$, i.e. P_R is a traditional data augmentation algorithm based on the joint density w^*. In terms of this, Hobert and Marchev (2006, Theorem 3) prove:

Proposition 9. *Let R and S be two Markov operators on $(\mathcal{Y}, \mathcal{G})$ that are both reversible with respect to π_y, and let P_R, $\widehat{P}_R$, P_S and $\widehat{P}_S$ be as defined above. Then*
(a) *$\widehat{P}_R$ and $\widehat{P}_S$ are reversible with respect to π_x;*
(b) *if $R \succeq_1 S$ then $\widehat{P}_R \succeq_1 \widehat{P}_S$;*
(c) *if $R \succeq_1 S$, and if P_R and P_S are both DA algorithms, then $\|\widehat{P}_R\| \leq \|\widehat{P}_S\|$.*

In particular, Proposition 9(c) requires unnatural assumptions about P_R and P_S being DA algorithms, which are hard to verify and might well fail. Using the theory of the previous section, we are able to improve upon their result, as follows:

Theorem 10. *In Proposition 9, part (c) may be replaced by any of the following:*
(c′) if $R \succeq_1 S$, then $\|\widehat{P}_R\| \leq \max(-m_{\widehat{P}_R},\ \|\widehat{P}_S\|)$.
(c″) if $R \succeq_1 S$, and if $\widehat{P}_R$ is a positive operator, then $\|\widehat{P}_R\| \leq \|\widehat{P}_S\|$.
(c‴) if $R \succeq_1 S$, and if P_R is a DA algorithm, then $\|\widehat{P}_R\| \leq \|\widehat{P}_S\|$.

Proof. (c$'$) follows from combining Proposition 9(b) with Corollary 4. (c$''$) follows immediately from (c$'$) as in Corollary 5. (c$'''$) follows by combining (c$''$) with Corollary 8. ■

Comparing Theorem 10 with Proposition 2, we conclude:

Corollary 11. *If $R \succeq_1 S$, and $m_{\widehat{P}_R} > -1$, and $\widehat{P}_S$ is geometrically ergodic, then $\widehat{P}_R$ is geometrically ergodic.*

Now, if S is the identity operator I on $\mathcal{Y}$, then P_S corresponds to the traditional data augmentation algorithm; that is, $P_S = P$. Of course, $R \succeq_1 I$ for all R. Hence, Theorem 10 immediately implies:

Corollary 12. *Let R be a Markov operator on $(\mathcal{Y}, \mathcal{G})$ that is reversible with respect to π_y, and let P_R, $\widehat{P}_R$ and $\widehat{P}$ be as defined above. Then*
(a) *$\widehat{P}_R \succeq_1 \widehat{P}$;*
(b) *$\|\widehat{P}_R\| \le \max\left(-m_{\widehat{P}_R}, \|\widehat{P}\|\right)$;*
(c) *if $\widehat{P}_R$ is a positive operator, then $\|\widehat{P}_R\| \le \|\widehat{P}\|$;*
(d) *(Hobert and Marchev, 2006) if P_R is a DA algorithm, then $\|\widehat{P}_R\| \le \|\widehat{P}\|$.*

Remark. Corollary 12(d) essentially says that $\|P_1\, R\, P_2\| \le \|P_1\, P_2\|$. One might think this is "obvious", since $\|R\| \le 1$, and since $\|AB\| = \|BA\|$ for reversible A and B. However, it does not necessarily follow that $\|P_1\, R\, P_2\| \le \|R\|\, \|P_1\, P_2\|$ in general. For example, let

$$R = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \qquad P_1 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \qquad P_2 = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}.$$

Then $P_1 P_2 = 0$, but $P_1 R P_2 = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$ which has norm 1. ■

Hobert and Marchev leave as an open problem whether their additional assumption (that P_R and P_S are DA algorithms) is required to conclude that $\|\widehat{P}_R\| \le \|\widehat{P}_S\|$. Theorem 10(c$'''$) shows that at most half of their assumption, i.e. that just P_R is a DA algorithm, is required. But this still leaves the question of whether the result holds without any such assumption at all. In fact, it does not:

Example 13. Let $\mathcal{X} = \mathcal{Y} = \{0, 1\}$ and suppose that $\mathbf{P}(X = 0, Y = 0) = 1/4$, $\mathbf{P}(X = 0, Y = 1) = 3/8$, $\mathbf{P}(X = 1, Y = 0) = 1/4$ and $\mathbf{P}(X = 1, Y = 1) = 1/8$. Note that the marginal distribution of Y is uniform; i.e, $\mathbf{P}(Y = 0) = \mathbf{P}(Y = 1) = 1/2$. The marginal distribution of X is as follows: $\mathbf{P}(X=0)=5/8$ and $\mathbf{P}(X=1)=3/8$. Now define

$$R=\begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix} \quad \text{and} \quad S=\begin{pmatrix}1/2 & 1/2\\ 1/2 & 1/2\end{pmatrix}$$

and consider these to be Markov transition matrices on $\mathcal{Y}$. It's easy to see that R and S are both reversible with respect to the marginal distribution of Y. Moreover, $S-R$ has eigenvalues 0 and 1 so $R \succeq_1 S$. Note that a draw from S is equivalent to a draw from the marginal distribution of Y. It follows immediately that

$$\widehat{P}_S=\begin{pmatrix}5/8 & 3/8\\ 5/8 & 3/8\end{pmatrix}.$$

It's easy to show that

$$\widehat{P}_R=\begin{pmatrix}3/5 & 2/5\\ 2/3 & 1/3\end{pmatrix}.$$

Thus, $\widehat{P}_R$ and $\widehat{P}_S$ are both irreducible and aperiodic. Furthermore, $\widehat{P}_R$ has eigenvalues 1 and $-1/15$, so $\|\widehat{P}_R\|=1/15>\|\widehat{P}_S\|=0$.

Alternatively, if we instead take $\mathbf{P}(X=0,Y=0)=\mathbf{P}(X=1,Y=1)=1/2$, then $\widehat{P}_R$ is the same as R, so $\|\widehat{P}_R\|=1$ even though $R \succeq_1 S$ and $\|\widehat{P}_S\|=0$. This gives an even more "extreme" counter-example, but at the expense of making $\widehat{P}_R$ periodic. ■

5 Questions for Further Research

We close with a few brief questions for possible further research.

Is it possible to quantify the improvement of $\widehat{P}_R$ over $\widehat{P}_S$? For example, suppose $S-R-cI$ is positive for some $c>0$. What quantitative results does this imply about how much M_R is less than M_S, or $\mathrm{Var}(f,R)$ is less than $\mathrm{Var}(f,S)$, or $\|R\|$ is less than $\|S\|$?

Which of the results in this paper carry over to the non-reversible case? Or even to the case where $P=Q_1Q_2$ with each Q_i reversible? Various results about mixing of non-reversible operators are discussed in e.g. Mira and Geyer (1999), Fill (1991), and Dyer et al. (2006), but it is not clear how to apply them in the current context.

REFERENCES

Y. Amit (1991), On the rates of convergence of stochastic relaxation for Gaussian and Non-Gaussian distributions. J. Multivariate Analysis **38**, 89-99.

K.S. Chan and C.J. Geyer (1994), Discussion paper. Ann. Stat. **22**, 1747–1758.

M. Dyer, L.A. Goldberg, M. Jerrum, and R. Martin (2006), Markov chain comparison. Prob. Surveys **3**, 89–111.

J.A. Fill (1991), Eigenvalue bounds on convergence to stationarity for non-reversible Markov chains, with an application to the exclusion process. Ann. Appl. Prob. **1**, 62–87.

J.P. Hobert and D. Marchev (2006), A theoretical comparison of the data augmentation, marginal augmentation and PX-DA algorithms. Preprint. Available at:

`http://web.stat.ufl.edu/∼jhobert/`

C. Kipnis and S.R.S. Varadhan (1986), Central limit theorem for additive functionals of reversible Markov processes and applications to simple exclusions. Comm. Math. Phys. **104**, 1-19.

J.S. Liu, W. Wong, and A. Kong (1994), Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. Biometrika **81**, 27-40.

J.S. Liu and Y.N. Wu (1999), Parameter expansion for data augmentation. JASA **94**, 1264–1274.

A. Mira (2001), Ordering and improving the performance of Monte Carlo Markov chains. Stat. Sci. **16** (2001), 340–350.

A. Mira and C. Geyer (1999), Ordering Monte Carlo Markov chains. Technical Report No. 632, School of Statistics, University of Minnesota. Available at:

`http://eco.uninsubria.it/webdocenti/amira/papers.html`

X.L. Meng and D.A. van Dyk (1999), Seeking efficient data augmentation schemes via conditional and marginal augmentation. Biometrika **86**, 301–320.

P.H. Peskun (1973), Optimum Monte Carlo sampling using Markov chains. Biometrika **60**, 607–612.

G.O. Roberts and J.S. Rosenthal (1997), Geometric ergodicity and hybrid Markov chains. Electronic Comm. Prob. **2**, Paper no. 2, 13–25.

G.O. Roberts and J.S. Rosenthal (2001), Markov chains and de-initialising processes. Scandinavian Journal of Statistics **28**, 489–504.

G.O. Roberts and J.S. Rosenthal (2004), General state space Markov chains and MCMC algorithms. Prob. Surveys **1**, 20–71.

G.O. Roberts and J.S. Rosenthal (2006), Variance bounding Markov chains. Preprint. Available at: `http://probability.ca/jeff/research.html`

J.S. Rosenthal (2003), Asymptotic Variance and Convergence Rates of Nearly-Periodic MCMC Algorithms. J. Amer. Stat. Assoc. **98**, 169–177.

W. Rudin (1991), Functional Analysis, 2nd ed. McGraw-Hill, New York.

M.A. Tanner and W.H. Wong (1987), The calculation of posterior distributions by data augmentation (with discussion). J. Amer. Stat. Assoc. **82**, 528-550.

L. Tierney (1998), A note on Metropolis-Hastings kernels for general state spaces. Ann. Appl. Prob. **8**, 1–9.