Estimating L^1 Error of Kernel Estimator: Monitoring Convergence of Markov Samplers

Bin Yu*

Abstract

In many Markov chain Monte Carlo problems, the target density function is known up to a normalization constant. In this paper, we take advantage of this knowledge to facilitate the convergence diagnostic of a Markov sampler by estimating the L^1 error of a kernel estimator. Firstly, we propose an estimator of the normalization constant which is shown to be asymptotically normal under mixing and moment conditions. Secondly, the L^1 error of the kernel estimator is estimated using the normalization constant estimator, and the ratio of the estimated L^1 error to the true L^1 error is shown to converge to 1 in probability under similar conditions. Thirdly, we propose a sequential plot of the estimated L^1 error as a tool to monitor the convergence of the Markov sampler. Finally, a 2-dimensional bimodal example is given to illustrate the proposal, and two Markov samplers are compared in the example using the proposed diagnostic plot.

KEY WORDS: β-mixing; Diagnostic; Normalization constant.

*Bin Yu is Assistant Professor, Department of Statistics, University of California, Berkeley, CA 94720-3860. Research supported in part by the Junior Faculty Research Grant from University of California at Berkeley, grants DAAL03-91-G-007 and DAAH04-94-G-0232 from the Army Research Office, and grant DMS-9322817 from the National Science Foundation. The author is very grateful to Professors Peter Bickel and Andrew Gelman for many helpful discussions and their comments on the draft. Special thanks are due to Mr. Sam Buttrey for his help on simulation, to Professor Per Mykland and Mr. Karl Broman for commenting on the draft, and to two anonymous referees for their very helpful comments.

1 INTRODUCTION

Markov chain Monte Carlo (MCMC) methods have been used in both Bayesian and likelihood computations (Gelfand and Smith, 1990, Geyer and Thompson, 1992, Smith and Roberts, 1992, and Besag and Green, 1993). The MCMC method enables us to obtain (dependent) samples from a target density from which direct sampling is difficult. Quantities of interest of the target distribution, such as mean, variance, and tail probabilities, can then be approximated using the MCMC sample. Since the target distribution is the stationary distribution of the constructed Markov sampler, the success of the MCMC methods relies crucially on our ability to assess the convergence of the chain to its equilibrium.

The so-called convergence diagnostics problem has attracted attention from many authors. Although a priori bounds on the convergence rate exist (Rosenthal, 1991, 1993a, 1993b, and Mengersen and Tweedie, 1993), they are currently known only in some special cases. Convergence diagnostics based on a single run of MCMC or Markov samplers have been proposed using time series methods (Raftery and Lewis 1992). Gelman and Rubin (1992) proposed a multiple chain approach in the MCMC context, followed by Liu, Liu, and Rubin (1992) and Roberts (1992). In the context of Gibbs samplers, Ritter and Tanner (1992) and Cui, Tanner, Sinhua, and Hall (1992) suggested diagnostic statistics based on importance weights, using either multiple chains or a single chain.

In many MCMC problems, especially those using the Metropolis-Hastings algorithm, the target density is known up to a normalization constant. In this paper we take advantage of this knowledge to facilitate the convergence diagnostic of a Markov sampler. Based on a single MCMC sample aiming at a particular target density, we estimate the normalization constant using a kernel estimator and prove its asymptotic normality under mixing and moment conditions. Thus we have available two density estimators: one is the kernel estimator based on the MCMC sample and the other on the estimated normalization constant and the unnormalized target density. If the MCMC sample comes from around a local mode of the target density, the kernel estimator will approximate the conditional density around this mode. On the other hand, the other density estimator has the correct modes and so it will coincide well with the kernel estimator around the local modes visited by the Markov sampler, but differ greatly from the kernel estimator at unvisited modes, that is, their L^1 distance will be large if unvisited modes are major. This suggests that we look at the L^1 distance of these two estimators. The L^1 distance is favored over other distances such as L^2 because of its nice invariance property and its interpretability in terms of differences of probabilities (Devroye 1987, pp. x). Since the new density estimator is "parametric" (the normalization constant is a 1-dimensional "parameter"), it converges to the true target density at the $n^{-1/2}$ rate, provided that the chain mixes quickly. This rate outperforms the rate of any kernel estimator, hence the L^1 distance of the two density estimators estimates the L^1 error of the kernel estimator. It should be pointed out that the estimated L^1 error requires a multi-dimensional integration as does the numerical evaluation of the normalization constant. However, we do not need the estimated L^1 error to the precision needed for the normalization constant. When the Markov sampler is diagnosed as mixing well, at the price of one multi-dimensional integration for the estimated L^1 error, we can use the Markov sample to approximate quantities which require additional multi-dimensional integrations, such as mean, variance, and tail probabilities of the target density.

For 2-component Gibbs samplers, Zellner and Min (1992) proposed three 1-dimensional convergence diagnostic statistic, of which one is based the ratio of the target density values at two fixed points. This proposal is related to ours in the sense that it also uses the unnormalized density function, but only locally.

From a more theoretical point of view, density estimation in terms of L^1 norm in the iid case has been the focus of much research, for example, the books by Devroye (1987) and Devroye and Györfi (1985). It is worth noting, however, that any traditional method to estimate the L^1 error in the Markov sampler case that ignores the known unnormalized density form might give false impressions of convergence since they will be based on a single run one way or the other.

The paper is organized as follows. In Section 2, we investigate the statistical properties of the estimator of the normalization constant and the estimated L^1 error of the kernel estimator. Under geometric ergodicity of the Markov sampler, we prove the relative stability of the kernel estimator in the L^1 norm. Under an additional moment condition on the ratio of the kernel to the target density, we show that the normalization estimator is asymptotically normal and we obtain the first order expansion of the L^1 error of the kernel estimator. In section 3, we apply the results in section 2 to the convergence diagnostic problem of Markov samplers. A data-dependent way to choose the bandwidth for the kernel estimator is provided. As a diagnostic tool to monitor the convergence of the Markov sampler, we propose a sequential plot of the estimated L^1 errors for the Markov sampler. Different Markov samplers aiming at the same target density can be compared in terms of the estimated L^1 error sequential plot. Moreover, in Section 4 the proposed method is illustrated through simulations in a 2-dimensional bimodal example. Section 5 contains concluding remarks. All the proofs are deferred to the appendix.

2 ESTIMATING THE L^1 ERROR OF A KERNEL ESTIMATOR

In this section, based on the kernel estimator of the target density, estimators of the normalization constant and the L^1 error of the kernel estimator are proposed. Theoretical properties of these estimators are derived under mixing conditions on the Markov sampler and under a moment condition on the ratio between the kernel function and the unnormalized target density.

Assume that the d-dimensional target density $\pi(x)$ is known up to a normalization constant, i.e.,

$$\pi(x) = \theta g(x),$$

where $g(x)$ is known, non-negative and integrable, and $\theta := [\int g(x)dx]^{-1}$. Note that for simplicity we choose θ as the inverse of the normalization constant, and the integration domain in the definition of θ should be taken as the support of $g(x)$. Let $X_0, X_1, ..., X_n, ...$ be a sequence of observations from a Markov sampler in R^d with $\pi(x)$ as its stationary density with respect to the Lebesgue measure in R^d. Such a Markov sampler can be obtained using the Metropolis-Hasting algorithm (Metropolis, Rosenbluth, Rosenbluth, Teller, and Teller 1953, and Hastings 1970).

We further impose a geometric ergodicity condition on the Markov sampler as follows:

Assumption GE For some $0 < \rho < 1$, and $M(x)$ with $E_\pi M(x) < \infty$,

$$|P(X_n \in A | X_0 = x) - P_\pi(A)| \leq M(x)\rho^n \quad \text{for all measurable } A \text{ and } x \in R^d.$$

For sufficient conditions under which Assumption GE holds, see Rosenthal (1991, 1993a and 1993b) and Mengersen and Tweedie (1993).

Since the results in this paper are asymptotic in nature, under Assumption GE, we may assume from now on that X_0 has density π, i.e., $X_0, X_1, ...$ is stationary.

Another quantity which measures the mixing of the sequence is the β-mixing coefficient (Bradley

1986)

$$\beta(n) := \frac{1}{2}\sup\{\sum_i\sum_j |P(X_0 \in A_i, X_n \in B_j) - P(X_0 \in A_i)P(X_n \in B_j)| : \{A_i\} \text{ and } \{B_j\} \text{ are finite measurable partions of } R^d\}.$$

In the Markov case, the β-mixing coefficient can also be expressed as

$$\beta(n) = E[\sup_{A\subset R^d}|P(X_n \in A|X_0) - P(X_n \in A)|].$$

It is obvious from the second definition of $\beta(n)$ that, under Assumption GE,

$$\beta(n) \leq E_\pi M(x)\rho^n.$$

For a given 1-dimensional bounded symmetric kernel $K(\cdot)$ such that $\int_{R^d} K(|x|)dx = 1$, let $h(\cdot)$ be the d-dimensional kernel based on K (Silverman 1986):

$$h_\sigma(x) = \frac{1}{\sigma^d}K(\frac{|x|}{\sigma})$$

where bandwidth $\sigma > 0$ and $|\cdot|$ is the Euclidean norm in R^d. Then the kernel estimator with bandwidth b_n of $\pi(\cdot)$ can be defined as

$$\hat{\pi}_n(x) := \frac{1}{n}\sum_{i=1}^n h_{b_n}(x - X_i).$$

Technically speaking, θ is not an unknown parameter in the statistical sense, but rather a quantity unknown computationally. Intuitively, $\hat{\pi}_\sigma(x)/g(x) \approx \theta$ for any fixed point x, if $\hat{\pi}_\sigma$ estimates π well at x. However, this type of estimator estimates θ at a rate slower than $n^{-1/2}$. Using some insight from efficient estimation of smooth functionals in density estimation problems, Professor Peter Bickel suggested the following estimator of θ which will be shown to be asymptotically normal at the $n^{-1/2}$ rate:

Based on the kernel estimator $\hat{\pi}_\sigma$ with a fixed bandwidth σ, define

$$\begin{aligned}\hat{\theta}_\sigma &:= \frac{1}{n(n-1)}\sum_{i\neq j} h_\sigma(X_i - X_j)/g(X_j) \\ &= (n-1)^{-1}\sum_{j=1}^n \hat{\pi}_\sigma(X_j)/g(X_j) - (n(n-1))^{-1}\sum_{j=1} h_\sigma(0)/g(X_j).\end{aligned}$$

For a chosen measurable set $A \subseteq R^d$ with non-zero Lebesgue measure, denote

$$I_n = I_n(A) = \int_A |\hat{\pi}_n(x) - \pi(x)|dx,$$

and its estimator

$$\hat{I}_n = \hat{I}_n(A) = \int_A |\hat{\pi}_n(x) - \hat{\theta} f(x)|dx.$$

Lemma 1 Under Assumption GE, let $b_n = O(n^b)$ for some $b \in (0,1)$. Then for any $c > 1$, we have

$$|I_n(A) - EI_n(A)| = o((\log n)^{c/2}\ n^{-1/2}) \text{ a.s.}$$

Theorem 2 (Relative Stability) Suppose $\pi(x)$ has continuous second (partial) derivatives and $\int_{R^d} |s_i s_j| K(|s|) ds < \infty$ for $i, j = 1, 2, ..., d$. Then under Assumption GE and for $b_n = Bn^{-1/(d+4)}$ ($B > 0$), we have

$$\liminf_{n\to\infty}\ n^{2/(d+4)} EI_n(A) > 0, \quad \text{and}$$

$$\lim_{n\to\infty} I_n(A)/EI_n(A) = 1\ \ a.s.$$

Theorem 3 (Asymptotic Normality of the Normalization Constant Estimator) Assume that GE holds and that there exist $p > 2$ and $\sigma > 0$ such that for $dist(X, Y) = dist(X_0, X_j)$ ($j > 0$), or $dist(X) = dist(Y) = dist(X_0)$ and X, Y independent, we have

$$E|h_\sigma(X - Y)/g(X)|^p < \infty.$$

Then $\sqrt{n}(\hat{\theta}_\sigma - \theta)$ is asymptotically normal and so is $\sqrt{n}(\hat{\theta}_\sigma^{-1} - \theta^{-1})$.

Following directly from Theorem 2 and Theorem 3, we have

Theorem 4 (Relative Stability of the Estimated L^1 Error) Under the assumptions of Theorem 2 and Theorem 3, and for a non-zero measure compact set $A \subset R^d$, we have, as n tends to infinity,

$$\frac{\hat{I}_n(A)}{I_n(A)} = \frac{\int_A |\hat{\pi}_n(x) - \hat{\theta} g(x)|dx}{\int_A |\hat{\pi}_n(x) - \pi(x)|dx} \to 1 \quad \text{w. p. 1.}$$

Remarks: (i) The above results from Lemma 1 to Theorem 5 hold under the weaker condition $\beta(n) \le O(\rho_\beta^n)$ for some $\rho_\beta \in (0,1)$.

(ii) Because

$$\begin{aligned} |\hat{I}_n(A) - I(A)| &= |\int_A (|\hat{\pi}_n(x) - \hat{\theta} g(x)| - |\hat{\pi}_n(x) - \theta g(x)|) dx| \\ &\le \int_A (|\hat{\pi}_n(x) - \hat{\theta} g(x) - (\hat{\pi}_n(x) - \theta g(x))| dx \\ &= |\hat{\theta} - \theta| \int_A g(x) dx, \end{aligned}$$

and for any $\alpha < 1/2$, $n^{\alpha}(\hat{\theta} - \theta) \to 0$ with probability 1 by Theorem 3. Hence

$$|\hat{I}_n(A) - I(A)| = o_p(n^{-\alpha}),$$

for any $\alpha < 1/2$.

(iii) Note that in Assumption GE, if there is a positive constant such that $M(x) \leq \gamma < \infty$, Doob (1953, Lemma 7.1, pp. 222) gives an inequality on the covariance of functions based on two blocks of observations which are apart. In Rosenblatt (1970), the G_2 assumption is imposed to calculate the variance of $\hat{\pi}_n(x)$ and covariance of $\hat{\pi}_n(x)$ and $\hat{\pi}_n(y)$ $(x \neq y)$. He also gave a CLT for $\hat{\pi}_n(x)$. Our situation is a little different since the joint distributions might have atoms when the Metropolis-Hastings algorithms are used.

Let us now specify the G_2 assumption in Rosenblatt (1970). For any bounded measurable function h, let T be the transition probability operator of our stationary Markov sequence

$$(Th)(y) := E(h(X_1)|X_0 = y).$$

Let $||h||_2$ be the L^2 norm of h and the following be the modified L^2 norm of T^n

$$|T^n|_2 := \sup_h ||T^n h - Eh||_2 / ||h - Eh||_2.$$

T is said to satisfy G_2 if for some $m > 0$,

$$|T^m|_2 < 1.$$

Let $\alpha_k(x)$ be the atom mass of the kth order joint distribution, and $p_k(y|x)$ to be the continuous part of the conditional distribution

$$\alpha_k(x) = P(X_k = x | X_0 = x), \quad p_k(y|x)dy := P(X_k \in dy | X_0 = x) \quad \text{for } x \neq y.$$

Theorem 5 Suppose that π, α_k, p_k $(k \geq 1)$ and second (partial) derivatives of $\pi(x)$ are continuous, and that $\int |s_i s_j| K(|s|) ds < \infty$ for all $i, j = 1, ..., d$. Then under assumption G_2, and for a compact set A of positive Lebesgue measure and $b_n = B n^{-1/(d+4)}$ $(B > 0)$, $\sqrt{n b_n^d}(\hat{\pi}_n(x) - E\hat{\pi}_n(x))$ is asymptotically normal with mean zero and variance v_x, and as $n \to \infty$,

$$n^{2/(d+4)} E I_n(A) \to e(A),$$

where

$$
\begin{aligned}
v_x &= \lim_{n\to\infty}(nb_n^d)\mathrm{Var}(\hat{\pi}_n(x)) \\
&= \pi(x)\int K^2(|s|)ds(1+2\sum_{k=2}^{\infty}\alpha_k(x)) < \infty, \\
e(A) &= \int_A\int_{R^d}|x\sqrt{v_t}+\mathrm{bias}(t)|\phi(x)dxdt
\end{aligned}
$$

where ϕ is the standard normal density and $\mathrm{bias}(t) = 2^{-1}B^{(d+4)/2}\int s'\Delta\pi(t)sK(|s|)ds$.

Remarks: Similar results as presented in Theorems 2-5 can be shown in the case of Gibbs samplers (Gelfand and Smith, 1990). There the mixture estimator of the marginal density can be used, instead of $\hat{\theta}$, together with the known conditional density to form another estimator for the joint density besides the kernel estimator. Note that Yu (1993, Theorem 4.2, p. 725) showed that the convergence rate of the mixture estimator outperforms that of the kernel estimator. In the discrete case, a histogram estimator can be used in the place of the kernel estimator, and $\hat{\theta}$ can be similarly constructed based on the histogram estimator.

3 USING ESTIMATED L^1 ERROR TO MONITOR THE CONVERGENCE OF MARKOV SAMPLERS

In this section, we propose a convergence diagnostic statistic based on the estimated L^1 error of the kernel estimator studied in Section 2. This way, we use the information contained in the unnormalized target density $g(x)$ in addition to a single run of the Markov chain. If the Markov sampler mixes quickly, the results in Section 2 give the consistency of this diagnostic statistic. Many other diagnostic statistics proposed in the literature also possess this property. However, the estimated L^1 error statistic is motivated to capture the discrepancy between the Markov sample and the target density when the sampler does NOT mix quickly, or when the sampler is sticky. In other words, the estimated L^1 error statistic is also expected to have good "power" if we view the convergence diagnostic problem at a certain sample size as testing the null hypothesis that the Markov chain converges to the target density verses the alternative that it does not. See Section 4 for an approximate calculation of the estimated L^1 error in the bimodal case when the sampler fails to converge to the correct target density.

Since the consistency result (Theorem 3) holds sample-wise, we can monitor the convergence of a Markov sampler by sequentially plotting the estimated L^1 error $\hat{I}_n(A)$ against the the sample size. We choose A as large as the computing resource allows and try to include the modes in A if we know how. Since adding a few more observations does not change the kernel estimator very much, we only need to plot $\hat{I}_n(A)$ over regular intervals of size *nstep*, which can be decided based on the availability of the computing resources. In the worse case we take $nstep = n$, i.e., we only evaluate the estimated L^1 error at the end of pre-set iterations. Although we know the optimal rate to choose for b_n, the optimal multiplier B_{op} depends on $\pi(x)$ and thus has to be chosen based on data. A sensible choice can be made by modifying the recommendation of Silverman (1986) in the iid case. Denote the bandwidth recommended by Silverman as

$$b_n^{ind} = A(d,K)n^{-1/(d+4)}\sqrt{\sum_{i=1}^{d} s_i^2/d}$$

where s_i^2 are the component-wise sample variances, and $A(1,K) = 1.06$ and $A(2,K) = 0.96$ for the Gaussian kernel K. See Silverman (1986, p. 87) for other values of $A(d,K)$. In our case, the Markov sampler is often positively dependent; hence it calls for a wider bandwidth. An ad hoc rule is to choose the one in $\{jb_n^{ind} : j = 1, ..., J\}$ that gives the smallest corresponding estimated L^1 error. From the limited simulations reported in the next section, it seems adequate to choose $J = 7$. On the other hand, if one suspects that the Markov sampler is negatively dependent, choices such as b_n^{ind}/j may be considered.

If the Markov sampler mixes quickly, the L^1 error plot should decay at rate $n^{-2/(d+4)}$. Large values of $\hat{I}_n$ (> 0.3) indicate that the sampler has not produced a satisfactory sample from $\pi(x)$, due to the unstationarity of the sampler, the stickness of the sampler, or pure chance error. Markov samplers aiming at the same target density can be compared by comparing their L^1 error plots. We prefer Markov samplers with a low L^1 error (< 0.3) plot. However, false convergence signs can be seen from the above plots if both the Markov sampler and the region A miss the same major mode of $\pi(x)$.

When d is not small and we do not have good ideas about how to choose A, we are forced to either choose A large and therefore require much computing power to evaluate $\hat{I}_n(A)$, or alternatively, we can select as many points $x_1, x_2, ..., x_l$ as the computing power allows and evaluate

$$(\hat{\pi}_n(x_j) - \hat{\theta}g(x_j)),$$

and the standardized

$$z_j := \sqrt{nb_n^d}(\hat{\pi}_n(x_j) - \hat{\theta}g(x_j))/\sqrt{\hat{\theta}g(x_j)}.$$

Note that under assumption G_2, we can show as did Rosenblatt (1970) that the z_j's are asymptotically independent normal, but their asymptotic variances might differ. Individual values of the z_j's are informative: locations of z_j's with large values indicate positions of potential modes missed by the current run of the sampler. To avoid false alarms due to the random fluctuation in the kernel estimator, it is worth the effort to select more points around z_j's with large values and calculate the z values for these new points. Large values for the new z's would confirm that there is indeed a missing mode.

An alternative in dealing with the high dimensional situation is to combine our approach with the multiple-chain approach. We may take a 1-dimensional summary statistic, produce multiple runs of the sampler, and compare the L^1 errors between kernel estimators based on different runs of the 1-dimensional summary statistic. However, if all the runs miss the same region of the sample space projected to the direction of the 1-dimensional summary statistic or all the runs converge to an incorrect density, the pairwise L^1 errors will not reveal that. See Figure 1 for the sequential plot of the L^1 error between the kernel estimators based on two runs of a summary statistic in the Ising model as in Gelman and Rubin (1992). Only the last 1000 samples in their paper are used. The same bandwidth is used for both kernel estimators, and it is the bandwidth which minimizes the L^1 error among $\{jb_n^{ind}: \ j = 1, 2, ..., J\}$ where the combined sample variance from both runs is used in b_n^{ind}. The estimated L^1 error in Figure 1 stays high around 1, indicating that the two runs are not converging to the same target density.

4 EXAMPLES: BIMODAL TARGET DENSITIES

In this section, we investigate the performance of the proposed estimated L^1 error as a convergence diagnostic statistics in the case of bimodal target densities. First, we give an approximate "power" calculation when the Markov sampler fails to converge to the correct target density. Then a simulation example is carried out where the target density is the bimodal mixture of two 2-dimensional normals and two Metropolis Markov samplers are compared in terms of the proposed estimated L^1 error.

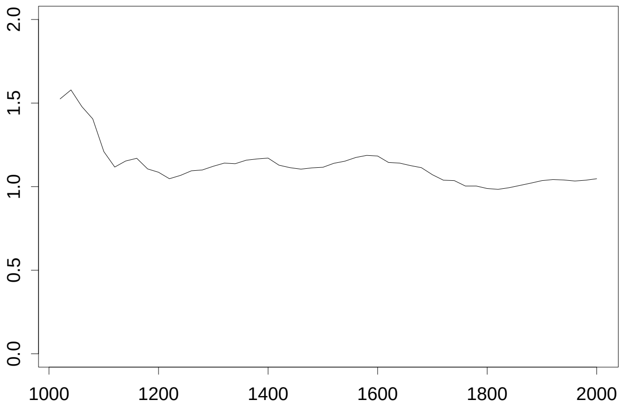


Figure 1: Sequential plot of estimated L^1 error of the kernel estimators based on the two runs of a 1-dimensional summary statistic in the Ising model– the same two runs as in Gelman and Rubin (1992).

Approximate "power" calculation

A bimodal $\pi(x)$ may be obtained by letting

$$\pi(x) = p_1\pi_1(x) + p_2\pi_2(x) = \theta(c_1\pi_1(x) + c_2\pi_2(x)) := \theta g(x),$$

where weights $p_1 + p_2 = 1$, and $\pi_1(x)$ and $\pi_2(x)$ are unimodal densities. If $\pi(x)$'s modes are well-separated, then we may express π as

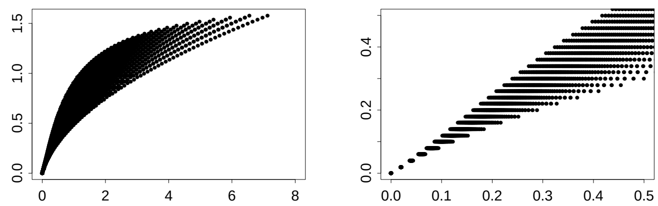
$$\pi(x) \approx p_1 I_{A_1}(x)\pi_1(x) + p_2 I_{A_2}(x)\pi_2(x),$$

where A_1 and A_2 are disjoint. If A is large enough to include both A_1 and A_2 and if for a fixed sample size n, our Markov sampler has been visiting the two modes with proportions q_1 and q_2 and if within each mode the sampler mixes well, then

$$\hat{\pi}_n(x) \approx q_1\pi_1(x) + q_2\pi_2(x), \text{ and } \hat{\theta} \approx c_1^{-1}q_1^2 + c_2^{-1}q_2^2.$$

Hence since $\int_{A_i} \pi_i(x)dx \approx 1$ for $i = 1, 2$,

$$\hat{I}_n(A) = \int |\hat{\pi}_n(x) - \hat{\theta}g(x)|dx$$



(a) (b)

Figure 2: Kernel L^1 error against estimated Kernel L^1 error in the bimodal case for the range (0,8) of estimated L1 error (a) and for the range (0,0.5) (b).

$$
\begin{aligned}
&\approx \int |q_1\pi_1(x) + q_2\pi_2(x) - (q_1^2/c_1 + q_2^2/c_2)g(x)|dx \\
&\approx \int_{A_1} |q_1 - (q_1^2/c_1 + q_2^2/c_2)c_1|\pi_1(x)dx + \int_{A_2} |q_2 - (q_1^2/c_1 + q_2^2/c_2)c_2|\pi_2(x)dx \\
&\approx |q_1 - (q_1^2/c_1 + q_2^2/c_2)c_1| + |q_2 - (q_1^2/c_1 + q_2^2/c_2)c_2| \\
&\approx |q_1 - q_1^2 - q_2^2 p_1/p_2| + |q_2 - q_2^2 - q_1^2 p_2/p_1|.
\end{aligned}
$$

Note that the true L^1 error of $\hat{\pi}(x)$ is

$$
\begin{aligned}
I_n(A) &= \int |\hat{\pi}_n(x) - \pi(x)|dx \\
&\approx \int |q_1\pi_1(x) + q_2\pi_2(x) - (p_1\pi_1(x) + p_2\pi_2(x))|dx \\
&\approx |q_1 - p_1| + |q_2 - p_2| \\
&= 2|q_1 - p_1|
\end{aligned}
$$

If the sampler mixes quickly, $p_1 \approx q_1$ and $p_2 \approx q_2$. Then $I_n(A) \approx 0$ and $\hat{I}_n(A) \approx 0$. If the sampler got stuck at the first mode, $q_1 \approx 1$ and $q_2 \approx 0$ and

$$\hat{I}_n(A) \approx p_2/p_1, \quad I_n(A) \approx 2(1 - p_1).$$

When p_2/p_1 is small, the second mode is a small bump, so as expected, not visiting mode 2 does not hurt much (both $\hat{I}_n$ and I_n are small). When the masses of the two modes are comparable, $p_2/p_1 \approx 1$ and $2(1 - p_1) \approx 1$, and we will see a big difference in the two estimates of $\pi(x)$. In Figure 2 (a), we plot corresponding values of $\hat{I}_n(A)$ and $I_n(A)$ calculated by the above formulas when (p_1, q_1) takes different values on $[0.1, 0.9]^2$. The plot takes a curved cone-shape starting at the

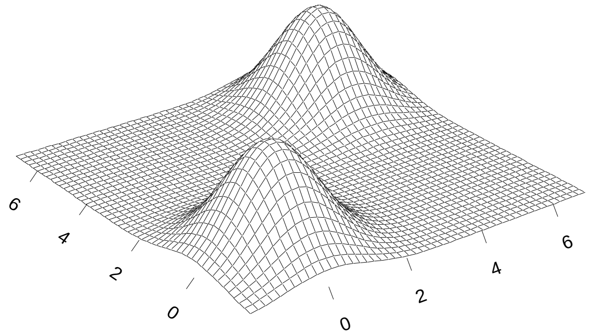


Figure 3: Target density in the simulations: mixture of two normals with identity covariance matrix and mean vectors (0,0) and (5,5).

origin. Apparently $\hat{I}_n$ can over-estimate I_n quite a lot, but fortunately, the over-estimation amount is an increasing function of I_n. For example, an observed value around 2 for $\hat{I}_n$ corresponds to the range of I_n from 0.7 to 1.2. From the diagnostic point of view, an L^1 error of size either 0.7 or 1.2 certainly means non-convergence. On the other hand, in Figure 2 (b) we see that an observed value around 0.3 for $\hat{I}_n$ corresponds to a much smaller range of I_n (from 0.2 to 0.35), and this range decreases to (0.1, 0.12) when $\hat{I}_n$ is around 0.1.

Mixture of two 2-dimensional normals: some simulation results

For the target density of a mixture of two 2-dimensional normals, we now compare two Metropolis samplers when two Gaussian jumping kernels are used. As shown in Figure 3, the target density is

$$\pi(x) = 0.5N(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}) + 0.5N(\begin{pmatrix} 5 \\ 5 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}).$$

Let

$$g(x) = 0.5\exp(-|x|^2/2) + 0.5\exp(-|x-5|^2/2), \quad \text{then } \theta = 1/(2\pi) = 0.1592.$$

Based on a given symmetric jumping kernel $q(x, dy)$, we use the Metropolis algorithm to simulate as follows n samples from a Markov chain with target density π:

Step 1: Take $x_0 = (0, 0)$.

Step 2: Given x_i, simulate a jump candidate y for the next step from $q(x_i, dy)$ and accept the jump y with probability $\min\{g(y)/g(x_i), 1\}$. Let

$$x_{i+1} = x_i + y \ \text{ if y is accepted; } x_{i+1} = x_i \text{ otherwise.}$$

Step 3: Stop when $i = n + 1$.

We will compare the following two jumping kernels

$$q_1(x, dy) = N(x, \begin{pmatrix} 1/2 & 0 \\ 0 & 1/2 \end{pmatrix})dy, \text{ and } q_2(x, dy) = N(x, \begin{pmatrix} 20.88 & 18 \\ 18 & 20.88 \end{pmatrix} dy.$$

According to Gelman, Roberts, and Gilks (1994), the second sampler is "optimal", and let us call the first sampler "sticky".

Simulations are carried out using a C-program. Both samplers were run for $n = 5000$ and the first 1000 were discarded. A Gussian kernel is used in the kernel estimator. The integration in the estimated L^1 error is evaluated on $A = [-2, 7] \times [-2, 7]$, with the number of grid points=50×50, $\sigma = 0.8$ is taken in $\hat{\theta} = \hat{\theta}_\sigma$, and $nstep = 100$. The evaluation of the estimated L^1 error started at the 1100th iteration of the remaining 4000 iterations, or the 2100th of the original 5000 iterations.

Figure 4 shows that the "sticky" sampler is trapped at the mode (0,0), but the optimal sampler visits the two modes with roughly equal amount of time. The kernel estimators in panels (c) and (d) correspond to the bandwidths selected from $\{jb_n^{ind} : \ j = 1, ..., 7\}$ which minimize the estimated L^1 error, based on the "sticky" and optimal samplers respectively. The kernel estimator from the "sticky" sampler estimates the density at the (0,0) mode well since all 4000 samples are used, while the kernel estimator from the optimal sampler splits the 4000 samples between two modes thus the estimated density surface is less smooth.

Using data from iterations 1001 to 5000, Figure 5 (a) shows that for the optimal sampler, the estimated L^1 error, which correspond to the selected bandwidth from $\{jb_n^{ind} : \ j = 1, ..., 7\}$, decreases from 0.72 at $n = 2100$ to 0.25 at $n = 5000$; while for the "sticky" sampler, the estimated L^1 error stays high around 1.1 or 1.2, because it missed entirely the mode at (5,5). Therefore the estimated L^1 error plot correctly diagnoses that the "sticky" sampler is sticky and the optimal sampler mixes well. Figure 5 (b) shows that $\hat{\theta}$ is very much more biased for the "sticky" sampler

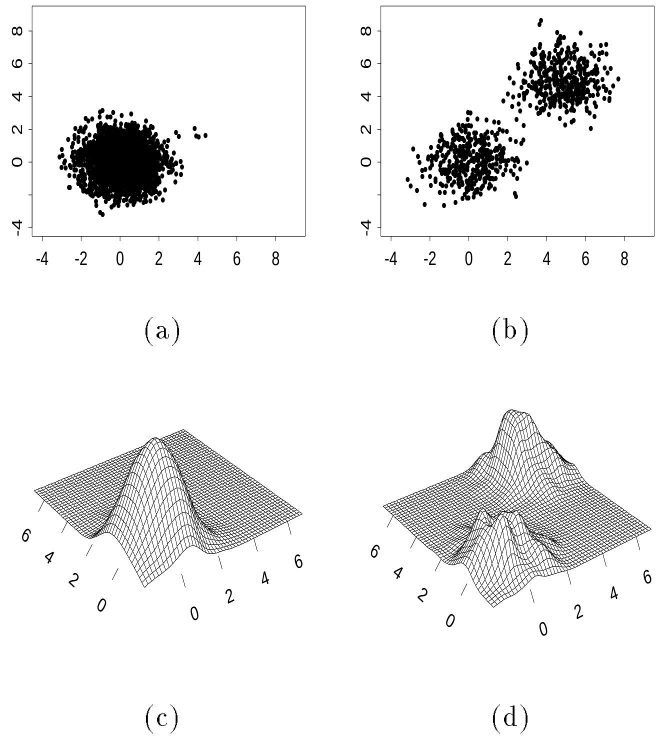
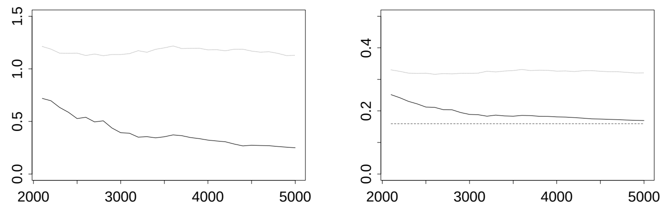


Figure 4: Scatter plots of the simulated Markov samples (n=1001,...,5000) for the "sticky" sampler (a) and the optimal sampler (b). Corresponding kernel estimates based on the "sticky" sampler (c) and the optimal sampler (d).

than for the optima sampler.

In Figure 6, the true L^1 errors at iterations $n = 2100, 2200, ..., 5000$ are plotted against the corresponding estimated L^1 errors for both the optimal and "sticky" samplers. Overall, the estimated L^1 error over-estimates the true L^1 error. However, for the group of points in the lower left corner from the optimal sampler, the true and estimated L^1 errors are close to each other at the end of the iteration since $\hat{\theta}$ estimates θ well – the over-estimation percentage is around 5% at $n = 5000$. For the group of points in the upper right corner from the "sticky" sampler, both true and estimated L^1 errors are high and the estimated L^1 errors over-estimate the true error by about 20% consistently. Figure 6 is consistent with what is in Figure 2 which is based on approximate calculations.



(a) (b)

Figure 5: (a): Sequential plot of estimated kernel L^1 errors for the "sticky" sampler (dotted line) and the optimal sampler (solid line), based on the simulated Markov samples (n=1001,...,5000). (b): Sequential plot of the estimated normalization constant for the true θ value (dashed line), the "sticky" sampler (dotted line), and the optimal sampler (solid line).

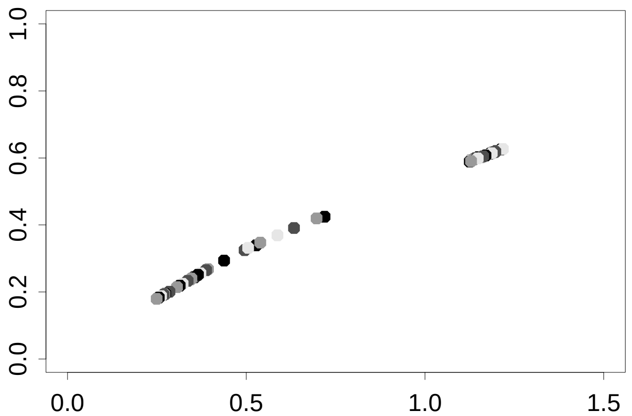


Figure 6: True L^1 error against estimated L^1 error at iterations $n = 2100, 2200, ..., 5000$ for the "sticky" sampler (the cluster in the upper right corner) and for the optimal sampler (the cluster in the lower left corner).

5 CONCLUDING REMARKS

We may regard the approach in this paper as introducing additional information in the unnormalized density to a single run Markov sampler. Since the estimated L^1 error does measure in a very direct way how close the sample is to the target density, the proposed approach can be used to compare different Markov samplers aiming at the same target density. For example, Metropolis-Hastings algorithms with different candidate kernels. When the numerical evaluation of the estimated L^1 error is not feasible, points should be chosen and the properly normalized differences of the two density estimators should be monitored sequentially. Alternatively, we may combine the proposed approach with the multiple-chain approach as discussed in Section 4. The advantages of our approach seems to be: (a) when the Markov sampler mixes quickly, the statistics we propose do stabilize sample-wise, while the multiple chain statistics in Roberts (1992) and Liu et al (1992) are proved to stabilize in expectations only; (b) since the estimated L^1 error is a meaningful measure comparable across different Markov samplers, it can also be used to compare different samplers. It is worth noting, however, that the approach here does not rely on the Markov property per se, but only the mixing property of the chain, while Roberts', Ritter and Tanner's, and Liu et al's proposals do take the Markov kernel into consideration.

Although the proposed diagnostic plot worked well in the example of the mixture of two 2-dimensional normals, more simulations are needed in general situations. The plot should also be compared with other diagnostic methods. Currently under investigation is the binned kernel estimator for moderately high dimensions ($d = 3, ...10$), and dimension reduction techniques such as projection pursuit for high dimensions.

APPENDIX: PROOFS

Proof of Lemma 1: We use a common blocking technique in the form used in Yu (1993) to translate the problem into that of iid blocks and then we employ the argument of Devroye (1988) for the iid case.

Divide the sequence $(X_1, X_2, ..., X_n)$ into μ_n blocks of size a_n and a remainder block. Let H_j, T_j and R_e denote the indices of the odd blocks, even blocks and the remainder block, $(\xi_1, ..., \xi_n)$ be a sequence of independent blocks with the same within-block distribution structure as the X-

sequence, i.e.

$$L(X_i, i \in H_j) = L(\xi_i, i \in H_j), \quad L(X_i, i \in T_j) = L(\xi_i, i \in T_j),$$

but with $(\xi_i, i \in H_j)$ independent for different j's.

For $j = 1, 2, ..., \mu_n$, let

$$Y_j = \sum_{i \in H_j} b_n^{-d}[K(|(x - X_i)/b_n|) - \pi(x)]$$

and

$$Z_j = \sum_{i \in H_j} b_n^{-d}[K(|(x - \xi_i)/b_n|) - EK(|(x - X_i)/b_n|)].$$

Then for any $\epsilon_n = o(1)$ and $b_n^{-d} a_n / n = o(\epsilon_n)$, by repeated use of the triangular inequality and for n large,

$$\begin{aligned} & P(|I_n - EI_n| > \epsilon_n) \\ \leq \quad & P(|\int |\sum Y_j| - E \int |\sum_j Y_j|| > \epsilon_n/2) \\ \leq \quad & P(|\int |\sum Z_j| - E \int |\sum_j Z_j|| > \epsilon_n/2) + \mu_n \beta(a_n). \end{aligned}$$

The last inequality holds by Lemma 3.1 in Yu (1993, p. 717).

Note further that using Devroye's notation for $W_{j,j} = (Z_j - \pi(x))/n$,

$$\begin{aligned} \int |W_{j,j} - EW_{j,j}| \quad & \leq \quad a_n n^{-1}(\int b_n^{-d}|K(|(x - X_1)/b_n|)|dx + E \int b_n^{-d}|K(|(x - X_1)/b_n|)|dx) \\ & \leq \quad 2a_n n^{-1} \int |K(|t|)|dt. \end{aligned}$$

Hence by the martingale inequality used in Devroye (Lemma 2, 1988),

$$P(|\int \sum_j Z_j - E \int \sum_j Z_j| > \epsilon_n/2) \leq 2 \exp(-\epsilon_n^2 \mu_n/(128[\int |K|]^2)).$$

It follows that for n large,

$$P(|I_n - EI_n| > \epsilon_n) \leq 2\text{exp}(-\epsilon_n^2 \mu_n/(128[\int |K|]^2)) + \mu_n \beta(a_n).$$

Under Assumption GE, $\beta(a_n) \leq O(\rho^{a_n})$. For any $\epsilon > 0$, let $\epsilon_n = (nb_n^d)^{-1/2}\epsilon$. If $b_n = n^{-b}$ for some $b \in (0, 1)$, we can choose $\mu_n = n/(\log n)^c$ for any $c > 1$, then $a_n \approx (\log n)^c$ and the following hold:

$$(b_n^d \mu_n)^{-1} = o(\epsilon_n), \sum_n \text{exp}(-\epsilon_n^2 \mu_n/(128[\int |K|]^2)) < \infty, \quad \sum_n \mu_n \beta(a_n) < \infty.$$

By Borel-Cantelli Lemma, $I_n - EI_n = o((\log n)^{c/2}/\sqrt{n})$ almost surely and for any $c > 1$. $\square$

Proof of Theorem 2: For the first statement, because for $b_n = Bn^{-1/(d+4)}$, $\sqrt{nb_n^d} = O(n^{2/(d+4)})$, it is enough to show that $\liminf \sqrt{nb_n^d}EI_n(A) > 0$.

Note that

$$\begin{aligned}
\text{bias}_n(t) &:= E\hat{\pi}_n(t) - \pi(t) \\
&= b_n^{-d}\int \pi(s)K(|(t-s)/b_n|)ds - \pi(t) \\
&= \int(\pi(t+sb_n) - \pi(t))K(|s|)ds \\
&= b_n\int s\cdot\nabla\pi(t)K(|s|)ds + 2^{-1}b_n^2\int s'\Delta\pi(t)sK(|s|)ds(1+o(1)) \\
&= 2^{-1}b_n^2\int s'\Delta\pi(t)sK(|s|)ds(1+o(1)).
\end{aligned}$$

The last equality holds because of the fact that $\int s_iK(|s|)ds = 0$, and by the continuity of the second derivatives of π, and by the dominated convergence theorem since $\int s_is_jK(|s|)ds < \infty$.

Recall that $\text{bias}(t) = 2^{-1}B^{(d+4)/2}\int s'\Delta\pi(t)sK(|s|)ds$, thus

$$\begin{aligned}
\liminf\sqrt{nb_n^d}EI_n &= \liminf\sqrt{nb_n^d}\int_A EI_n \\
&\geq \liminf\int_A E|\sqrt{nb_n^d}(\hat{\pi}_n(t) - \pi_n(t))|dt \\
&\geq \liminf\int_A |E(\sqrt{nb_n^d}(\hat{\pi}_n(t) - \pi(t))|dt \\
&= \liminf\int_A |\sqrt{nb_n^d}\text{bias}_n(t)|dt \\
&\geq \int_A \liminf\ |\sqrt{nb_n^d}\text{bias}_n(t)|dt \quad \text{(by Fatou's Lemma)} \\
&= \int_A \text{bias}(t)dt > 0.
\end{aligned}$$

Since $nb_n^d = O(n^{4/(d+4)}) = o(n)$, then for any positive constant c, $(\log n)^{c/2}/\sqrt{n} = o(1/\sqrt{nb_n^d})$. Hence

$$\begin{aligned}
I_n/EI_n &= (I_n - EI_n)/EI_n + 1 \\
&= \sqrt{nb_n^d}(I_n - EI_n)/[\sqrt{nb_n^d}EI_n] + 1 \\
&= o(\sqrt{nb_n^d}(\log n)^{c/2}/\sqrt{n})/[\sqrt{nb_n^d}EI_n] + 1\ a.s.\ \text{(by Lemma 1)} \\
&= 1 \text{ a.s.} \quad \square
\end{aligned}$$

Proof of Theorem 3: Straightforward from the CLT for U-statistics for β-mixing sequences in Yoshihara (1976) after noting that $H_\sigma(x-y) = h_\sigma(x-y)/g(x) + h_\sigma(x-y)/g(y)$ is a symmetric kernel and that the integrability condition is satisfied under our assumption, and that

$$Eh_\sigma(X_0 - X_j)/g(X_0) = Eh_\sigma(X_0 - X_j)/g(X_j) = \theta. \ \Box$$

Proof of Theorem 5:

Rosenblatt's (1970) proof is general enough to cover our case when the joint distributions have atoms, except that we need to modify the variance calculation to take into account the atoms. Note that

$$\begin{aligned}\mathrm{Var}(\hat{\pi}(x)) &= (nb_n^{2d})^{-1}[\mathrm{Var}(K(|(x-X_1)/b_n|) \\ &\quad +2\sum_{k=2}^{n}(1-(k-1)/n)\mathrm{Cov}(K(|(x-X_1)/b_n|), K(|(x-X_k)/b_n|))]\end{aligned}$$

$$\begin{aligned}&\mathrm{Cov}(K(|(x-X_1)/b_n|), K(|(x-X_k)/b_n|)) \\ = \ &E(K(|(x-X_1)/b_n|)K(|(x-X_k)/b_n|)) - (E(K(|(x-X_1)/b_n|))^2\end{aligned}$$

Letting $P_k(ds, dt)$ denote the joint distribution of X_1 and X_k under stationarity, then

$$\begin{aligned}&E(K(|(x-X_1)/b_n|)K(|(x-X_k)/b_n|)) \\ = \ &b_n^{-2d}\int\int K(|(x-s)/b_n|)K(|(x-t)/b_n|)P_k(ds, dt) \\ = \ &b_n^{-2d}\int \alpha_k(s)\pi(s)K^2(|(x-s)/b_n|)ds \\ &+b_n^{-2d}\int\int_{s\neq t}(1-\alpha_k(s))\pi(s)K(|(x-s)/b_n|)K(|(x-t)/b_n|)p_k(t|s)dsdt \\ = \ &b_n^{-d}\alpha_k(x)\pi(x)\int K^2(|s|)ds(1+o(1)).\end{aligned}$$

The last equality holds because π, α_k and p_k are continuous and by the dominated convergence theorem since $\int K^2(|s|)ds < \infty$. Moreover, under Assumption G_2, there is a $\rho_1 \in (0,1)$ such that

$$\begin{aligned}&\mathrm{Cov}(K(|(x-X_1)/b_n|), K(|(x-X_k)/b_n|)) \\ \leq \ &M\rho_1^k \mathrm{Var}(K(|(x-X_1)/b_n|)).\end{aligned}$$

Thus

$$nb_n^d \mathrm{Var}(\hat{\pi}_n(x)) \leq \pi(x)\int_{R^d} K^2(|s|)ds(1+2M\sum_{k=1}^{\infty}\rho_1^k).$$

Since $\pi(x)$ is continuous, by the dominated convergence theorem, as $n \to \infty$,

$$nb_n^d \text{var}(\hat{\pi}_n(x)) \to v_x.$$

We have shown in the proof of Theorem 2 that for $b_n = Bn^{-1/(d+4)}$, as $n \to \infty$,

$$\sqrt{nb_n^d} \text{bias}_n(x) = B^{2/(d+4)} \text{bias}_n(x)/b_n^2 \to \text{bias}(x).$$

Let

$$\begin{aligned} Y_n(x) &:= \sqrt{nb_n}|\hat{\pi}_n(x) - \pi(x)| \\ &= |\sqrt{nb_n^d}(\hat{\pi}_n(x) - E\hat{\pi}_n(x)) + \sqrt{nb_n^d}(E\hat{\pi}(x) - \pi(x))| \end{aligned}$$

Then

$$EY_n^2(x) = nb_n^d \text{Var}(\hat{\pi}(x)) + nb_n^d \text{bias}_n^2(x) \to v_x + \text{bias}^2(x).$$

Similarly as in Rosenblatt (1970),

$$Y_n(x) \to_D Y_0(x)$$

where $Y_x =_D |N(\text{bias}(x), v_x)|$. By Skorohod's representation theorem, there exists $\tilde{Y}_n(x) =_D Y_n(x)$ for $n = 0, 1, ...$ and $\tilde{Y}_n(x) \to \tilde{Y}_0(x) \quad a.s.$

Obviously, $\tilde{Y}$'s are non-negative r.v.'s and $E\tilde{Y}_n^2(x) \to E\tilde{Y}_0^2(x)$. It follows that

$$\begin{aligned} & \limsup \ E[\tilde{Y}_n(x) - \tilde{Y}_0(x)]^2 \\ \leq & \limsup \ E\tilde{Y}_n^2(x) + E\tilde{Y}_0^2(x) - 2 \liminf \ E[\tilde{Y}_n(x)\tilde{Y}_0(x)] \\ \leq & \ 2E\tilde{Y}_0^2(x) - 2E[\liminf \ \tilde{Y}_n(x)\tilde{Y}_0(x)] \ \text{(Fatou's Lemma)} \\ = & \ 0, \end{aligned}$$

which implies that

$$\lim_{n\to\infty} EY_n(x) = \lim_{n\to\infty} E\tilde{Y}_n(x) = E\tilde{Y}_0(x) = EY_0(x) =: e_x.$$

Since $EY_n(x)$ is uniformly bounded when x varies over a compact set A, by the dominated convergence theorem again,

$$EI_n(A) = \int_A EY_n(x)dx \to \int_A e_x dx = e(A). \ \square$$

References

[1] Besag, J. and Green P. J. (1993), "Spatial Statistics and Bayesian Computation," *Journal of the Royal Statistical Society*, Ser. B, 55, 25-37.

[2] Bradley, R. (1986), "Basic Properties of Strong Mixing Conditions," In *Dependence in Probability and Statistics: A Survey of Recent Results,* (E. Eberlein and M. S. Taqqu, eds.) 165-192. Birkhäuser, Boston.

[3] Cui, L., Tanner, M. A., Sinhua, B., and Hall, W. J. (1992), "Comment: Monitoring Convergence of the Gibbs Sampler: Further Experience with the Gibbs Stopper," *Statistical Science,* 7, 483-486.

[4] Devroye, L. (1988), "The Kernel Estimate Is Relatively Stable," *Probability Theory and Related Fields,* 77, 521-536.

[5] Devroye, L. and Györfi, L. (1985), *Nonparametric Density Estimation: The L^1 View*, New York: John Wiley & Sons.

[6] Doob, J. L. (1953), *Stochastic Processes,* New York: John Wiley & Sons.

[7] Gelfand, A. E. and Smith, A. F. M. (1990), "Sampling-based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association,* 85, 398-409.

[8] Gelman, A., Roberts, G., and Gilks, W. (1994), "Efficient Metropolis Jumping Rules," To appear in *Bayesian Statistics 5.*

[9] Gelman, A. and Rubin, D. B. (1992), "Inference from Iterative Simulation Using Multiple Sequences (with discussions)," *Statistical Science,* 7, 457-511.

[10] Geyer, C. J. and Thompson, E. A. (1993), "Constrained Monte Carlo Maximum Likelihood for Dependent Data" (with discussion), *Journal of the Royal Statistical Society,* Ser. B, 54, 657-699.

[11] Hastings, W. K. (1970), "Monte-Carlo Sampling Methods Using Markov Chains and Their Applications," *Biometrika*, 57, 97-109.

[12] Liu, C., Liu, J., and Rubin, D. B. (1992), "A Variational Control Variable for Assessing the Convergence of the Gibbs Sampler," In *1992 Proceedings of Statistical Computing Section of American Statistical Association,* 74-78.

[13] Mengersen, K. and Tweedie, R. (1993), "Rates of Convergence of the Hastings-Metropolis Algorithm," Preprint.

[14] Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953), "Equation of State Calculations by Fast Computing Machines," *Journal of Chemical Physics,* 21,1087-1092.

[15] Raftery, A. E. and Lewis, S. M. (1992), "Comment: One Long Run with Diagnostics: Implementation Strategies for Markov Chain Monte Carlo," *Statistical Science,* 7, 493-497.

[16] Ritter, C. and Tanner, M. A. (1992), "Facilitating the Gibbs Sampler: The Gibbs Stopper and the Griddy-Gibbs Sampler," *Journal of the American Statistical Association,* 87, 861-868.

[17] Roberts, G. O. (1992), "Convergence Diagnostics of the Gibbs Sampler," In *Bayesian Statistics*, J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith (eds.), 4, 775-782.

[18] Silverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis,* London: Chapman and Hall.

[19] Rosenblatt, M. (1970), "Density Estimates and Markov Sequences." In *Nonparametric techniques in statistical inferences,* M. L. Puri (eds.), Cambridge: Cambridge University Press.

[20] Rosenthal, J. (1991), "Rates of Convergence for Gibbs Sampling for Variance Component Models," Technical Report, Harvard University, Dept. of Mathematics.

[21] Rosenthal, J. (1993a), "Rates of Convergence for Data Augmentation on Finite Sample Spaces," *The Annals of Applied Probability,* 3, 819-839.

[22] Rosenthal, J. (1993b), "Minorization Conditions and Convergence Rates for Markov Chain Monte Carlo," Technical Report, University of Minnesota, School of Mathematics.

[23] Yoshihara, K. (1976). "Limiting Behavior of U-Statistics for Stationary, Absolutely Regular Processes," *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete,* 35, 237-252.

[24] Yu, B. (1993), "Density Estimation in the L^∞ Norm for Dependent Data with Applications to the Gibbs Sampler," *The Annals of Statistics*, 21, 711-735.

[25] Zellner, B. and Min, C. (1992), "Gibbs Sampler Convergence Criteria," Technical Report, Graduate School of Business, University of Chicago.