Outperforming the Gibbs sampler empirical estimator for nearest neighbor random fields

Priscilla E. Greenwood
University of British Columbia

Ian W. McKeague*
Florida State University

Wolfgang Wefelmeyer
University of Siegen

Abstract

Given a Markov chain sampling scheme, does the standard empirical estimator make best use of the data? We show that this is not so and construct better estimators. We restrict attention to nearest neighbor random fields and to Gibbs samplers with deterministic sweep, but our approach applies to any sampler that uses reversible variable-at-a-time updating with deterministic sweep. The structure of the transition distribution of the sampler is exploited to construct further empirical estimators that are combined with the standard empirical estimator to reduce asymptotic variance. The extra computational cost is negligible. When the random field is spatially homogeneous, symmetrizations of our estimator lead to further variance reduction. The performance of the estimators is evaluated in a simulation study of the Ising model.

1 Introduction

Suppose we want to calculate the expectation of a bounded function f under a distribution π on some space D. If D is of high dimension, or if π is defined indirectly, it may be difficult to calculate the expectation $\pi f = \int f(x)\,\pi(dx)$ analytically or even by numerical integration. The classical Monte Carlo method generates i.i.d. realizations $X^0, \ldots, X^n$ from π, and approximates πf by the *empirical estimator*

$$E_n^0 f = \frac{1}{n} \sum_{i=0}^{n-1} f(X^i).$$

*Research partially supported by NSF Grant ATM-9417528.

[1]All three authors were partially supported by NSERC, Canada.

[2]*AMS 1991 subject classifications.* Primary: 62M40, 65U05; secondary: 60J05, 62G20, 62M05

[3]*Key words and Phrases.* Markov chain Monte Carlo, Metropolis–Hastings algorithm, asymptotic relative efficiency, variance reduction, Ising model, parallel updating.

The estimator is strongly consistent and asymptotically normal. Often, however, this Monte Carlo method is difficult to implement. One reason is that high dimensional distributions are hard to simulate. Additional difficulties arise when π is defined indirectly, as in many Bayesian modeling situations, or only known up to a normalizing constant, as is usually the case for random fields.

The Markov chain Monte Carlo method (MCMC) generates a Markov chain $X^0, X^1, \ldots$, with π as invariant law. Again, the empirical estimator $E_n^0 f$ is used to approximate πf. If the chain is ergodic, the estimator is consistent; if the chain is geometrically ergodic, the estimator is asymptotically normal.

Over the last ten years, the special MCMC scheme known as the Gibbs sampler has become an important tool for estimating features in high dimensional distributions π. The method originated with the study of interacting particle systems, such as the Ising model in statistical physics, where it is known as the heat bath algorithm. The Gibbs sampler is also used in image analysis (Grenander, 1983, and Geman and Geman, 1984), Bayesian statistics (Smith and Roberts, 1993), spatial statistics (Besag and Green, 1993, and Graham, 1994), expert systems (Pearl, 1987, Spiegelhalter et al., 1993), incomplete data problems (Tanner and Wong, 1987), and hierarchical models (Gelfand et al., 1990).

There is a trade-off between speed of convergence of the Markov chain to stationarity and asymptotic variance of the empirical estimator. The asymptotic variance depends only on the stationary law of the chain. It is common to calculate the empirical estimator after a 'burn-in' has reached approximate stationarity, and one may switch at that point from a sampler with good rate to a sampler giving small variance. Speed of convergence of various MCMC schemes has been studied by Schervish and Carlin (1992), Chan (1993), Tierney (1994), Ingrassia (1994) and Athreya et al. (1995). For general Markov chains, see Meyn and Tweedie (1993). Some comparisons of the rates of different MCMC schemes may be found, e.g., in Frigessi et al. (1993) and Amit and Grenander (1993). Grenander (1993, Ch. 7) compares random and deterministic sweep strategies in terms of rates. He notes (p. 394) that estimator variance can be more relevant than convergence rate as an optimality criterion. The question of which Markov chain sampling scheme minimizes the asymptotic variance of the empirical estimator is studied by Peskun (1973), Frigessi et al. (1992) and Green and Han (1992), among others.

Here we consider a complementary question: Given a Markov chain sampling scheme, does the empirical estimator make best use of the sample? We will see that this is not so and will construct considerably better estimators in the case of the Gibbs sampler with deterministic sweep. Our approach will apply to any MCMC scheme with deterministic sweep and reversible local updating, in particular to local Metropolis–Hastings samplers with deterministic sweep.

Specifically, let $D = V^S$ with S a finite lattice and V a state space that may be discrete or continuous. The Gibbs sampler is described in terms of the one-dimensional conditional distributions $p_s(x_{-s}, dx_s)$ of $\pi(dx)$, where x_{-s} is obtained from x by omitting x_s. A deterministic sweep through the lattice is fixed by ordering the sites $s_1, \ldots, s_d$. The transition

distribution

$$Q(x, dy) = \prod_{j=1}^{d} p_{s_j}(y_{s_{<j}}, x_{s_{>j}}, dy_{s_j}),$$

with $s_{<j} = (s_1, \ldots, s_{j-1})$, has invariant law π. From an initial configuration X^0 the sampler generates a Markov chain $X^0, X^1, \ldots$ using the transition law Q. In the transition from X^i to X^{i+1} it updates, sequentially, the values at all sites $s_1, \ldots, s_d$. The sequence $X^0, X^1, \ldots$ can be viewed as a sequence of images or configurations on the lattice, i.e., an evolving random field.

The empirical estimator $E_n^0 f$ based on the Gibbs sampler is often considered as optimal when no information (apart from the simulated chain) about π is used. For classical i.i.d. Monte Carlo this is true—the empirical estimator has minimum asymptotic variance. This follows from a result of Levit (1974); see also the recent monograph of Bickel et al. (1993). A similar result for Markov chains is due to Penev (1991): If one uses only the information that the data comes from a Markov chain, and no model assumption about Q, then the empirical estimator has minimum asymptotic variance. This seems to support the popular impression that the empirical estimator makes best use of the data.

We will argue now that the specific structure of the transition distribution of the Gibbs sampler with deterministic sweep (that it is a composition of reversible variable-at-a-time updates) can easily be exploited to construct new 'empirical' estimators that can be combined with $E_n^0 f$ to produce considerably better estimators. The sampler updates the lattice site by site; view each update as a new image or configuration that differs form the previous one only at the updated site. The sampler generates, on its way from X^i to X^{i+1}, an intermediate chain of configurations, which we write as

$$X^{i.j} = (X^{i+1}_{s_{\leq j}}, X^i_{s_{>j}}), \quad j = 1, \ldots, d-1.$$

At 'time' $i.j$ only the sites $s_1, \ldots, s_j$ have been updated in the $(i+1)$-st pass. Interpolating the intermediate configurations into the chain $X^0, X^1, \ldots$, we obtain a *fine chain* $X^0, X^{0.1}, \ldots, X^{0.(d-1)}, X^1, X^{1.1}, \ldots$ indexed by the *fine time scale*. Under the stationary law, every intermediate configuration $X^{i.j}$ has distribution π. From each of the $d-1$ *interpolated chains* $X^{0.j}, X^{1.j}, \ldots$ we obtain a new 'empirical' estimator

$$E_n^j f = \frac{1}{n} \sum_{i=0}^{n-1} f(X^{i.j}), \quad j = 1, \ldots, d-1.$$

If the chain is ergodic, all the $E_n^j f$ are consistent; if the chain is geometrically ergodic, they are asymptotically normal.

Any convex combination of these empirical estimators is likely to result in an improved estimator unless f depends on only one component, say x_1. In practice one often uses the average

$$G_n^d f = \frac{1}{d} \sum_{j=0}^{d-1} E_n^j f.$$

This is just the empirical estimator on the fine chain and is analogous to the 'empirical estimator' used for *random* sweep. Geweke (1992) conjectures that $G_n^d f$ is efficient. Our simulations indicate that, at least in balanced situations, equal weights are close to optimal. We show, however, that equal weights are not strictly optimal, not even asymptotically. If one were to pursue the optimal linear combination, the weights would usually depend on π and would have to be estimated. The decrease in asymptotic variance would have to be balanced against the computational cost of estimating the optimal weights. It is therefore of interest to identify situations where one can do better than the single empirical estimator $E_n^0 f$, without estimating weights.

Consider the lattice with just two sites, $S = \{1, 2\}$. The fine chain is $X^0, X^{0.1}, X^1, X^{1.1}, \ldots$. We have noted that under the stationary law each configuration in the fine chain has distribution π. We show that when $d = 2$, the fine chain is time-reversible, in the sense that the stationary joint law of $X^0, X^{0.1}, \ldots, X^{n.1}$ is the same as $X^{n.1}, X^n, \ldots, X^0$. This means that under the stationary law, the empirical estimator $E_n^0 f$ has the same distribution as

$$E_n^1 f = \frac{1}{n} \sum_{i=0}^{n-1} f(X^{i.1}) = \frac{1}{n} \sum_{i=0}^{n-1} f(X_1^{i+1}, X_2^i).$$

Since the asymptotic variance does not depend on the initial distribution, the two estimators have the same asymptotic variance, and the best linear combination is

$$G_n f = \tfrac{1}{2}(E_n^0 f + E_n^1 f).$$

The estimator $G_n f$ is the empirical estimator computed from the fine chain. We show in Greenwood et al. (1995) that $G_n f$ cannot be further improved by any other estimator, except by using information about π, e.g., by using the one-dimensional conditional distributions $p_s(x_{-s}, dx_s)$ in a method akin to Rao–Blackwellization, see Smith and Roberts (1993, Section 4.2).

An important application of the two-step Gibbs sampler arises from a nearest neighbor random field on a finite square lattice. The nearest neighbor structure allows us to construct a two-step sampler by first updating all even sites and then all odd sites arranged in a checkerboard pattern. Even sites have only odd neighbors, so the conditional law of an odd site depends only on the values at even sites, and vice versa. The interpolated configurations $X^{i.1}$ are generated from the configurations X^i by updating just the even sites. The estimator $G_n f$ introduced above is the best linear combination of the two empiricals based on $X^0, X^1, \ldots$ and on $X^{0.1}, X^{1.1}, \ldots$. The checkerboard pattern has been widely used to perform Monte Carlo simulations on parallel computers, see, e.g., Heermann and Burkitt (1992).

Nearest neighbor models arise in condensed matter physics (e.g., Binder, 1992), lattice approximations to quantum fields (due to Guerra et al., 1975; also see Simon, 1974), lattice gases (Israel, 1979), and elsewhere. In such applications, homogeneities are likely to be present. When the random field is spatially homogeneous, i.e., invariant under translations, can we find a better estimator than $G_n f$? Suppose that π is invariant under a translation T

on the lattice S with periodic boundary conditions. Additional empirical estimators are

$$E_n^0 f \circ T = \frac{1}{n} \sum_{i=0}^{n-1} f(TX^i), \quad E_n^1 f \circ T = \frac{1}{n} \sum_{i=0}^{n-1} f(TX^{i.1}).$$

Now we ask whether equal weights give the optimal linear combination of the four empirical estimators $E_n^0 f, E_n^1 f, E_n^0 f \circ T, E_n^1 f \circ T$. To answer this question, we identify the field on V^S with a 'two-dimensional' field on $V^{S_e} \times V^{S_o}$ as before, where S_e and S_o are the sets of even and odd sites, respectively. Translations on V^S are of two types, those that take even into even and odd into odd sites, and those that take even into odd and odd into even sites. They are transformations from the 'two-dimensional' field $V^{S_e} \times V^{S_o}$ onto itself of the form $T(x_1, x_2) = (T_1 x_1, T_2 x_2)$ and $T(x_1, x_2) = (T_{21} x_2, T_{12} x_1)$. We call them parallel and transverse transformations.

We show in general that the average of $E_n^0 f, E_n^1 f, E_n^0 f \circ T, E_n^1 f \circ T$ is optimal. In fact, if π is invariant under a parallel or transverse transformation, the variances of all empirical estimators $E_n^0 f \circ T^j$ and $E_n^1 f \circ T^j$ involving arbitrary powers of T are equal. We describe various optimal linear combinations of them. Homogeneous nearest neighbor fields on a two-dimensional square lattice are invariant under the group generated by horizontal and vertical translations. We apply our results to write optimal estimators that combine all the translations or any subgroup of them.

The paper is organized as follows. Section 2 recalls the central limit theorem for empirical estimators on general Markov chains; Proposition 1 describes the best linear combination of the 'empirical' estimators when the transition distribution is invariant under some transformation. Section 3 considers the two-step Gibbs sampler and introduces the 'empirical' estimator $E_n^1 f$ based on the interpolated chain; Theorem 1 shows that the best linear combination of $E_n^0 f$ and $E_n^1 f$ is the average. Section 4 continues the study of the two-step Gibbs sampler when π is invariant under a parallel or transverse transformation T; Theorem 2 shows how best to combine 'empirical' estimators $E_n^0 f \circ T^j$ and $E_n^1 f \circ T^j$ involving powers of T; Theorem 3 extends this to deal with powers of two transformations. Applications to nearest neighbor fields are in Section 5.

2 Markov chains and invariance

In this section we consider a geometrically ergodic Markov chain and empirical estimators based on it. If there exist transformations that leave the transition distribution invariant, then we can construct additional 'empirical' estimators and find the best linear combination of these estimators. The use of group invariance is explored in the i.i.d. setting, e.g., by Bickel et al. (1993). Our results will be applied to the Gibbs sampler and a certain class of transformations in Sections 3 and 4.

Let $Q(x, dy)$ be a transition distribution with invariant distribution $\pi(dx)$ on a measurable space D. Fix an arbitrary initial distribution and let $X^0, \ldots, X^n$ be observations from the corresponding Markov chain. Assume that the Markov chain is geometrically ergodic,

i.e., ergodic (positive Harris recurrent) and there exists a positive constant $r < 1$ and a measurable function h on D with $\pi|h| < \infty$, such that $||Q^n(x,\cdot) - \pi|| \le h(x)r^n$ for all $x \in D$, where $||\cdot||$ denotes the total variation distance.

For a bounded measurable function $v(x,y)$ on $D \times D$, consider the expectation of v under the joint invariant distribution,

$$Ev(X^0, X^1) = \pi Qv = \iint \pi(dx)Q(x,dy)v(x,y).$$

The notations E for expectation, $\stackrel{\mathrm{d}}{=}$ for 'equal in law,' and $\sim$ for 'distributed as,' will *always* be with respect to the *stationary* law of the chain.

The Markov chain $\mathbf{X}^i = (X^i, X^{i+1})$ has invariant distribution πQ and is geometrically ergodic, which follows from the geometric ergodicity of X^i and since the n-step transition distribution of $\mathbf{X}^i$ is $P^n(\mathbf{x}, d\mathbf{y}) = Q^{n-1}(\mathbf{x}_2, d\mathbf{y}_1)Q(\mathbf{y}_1, d\mathbf{y}_2)$ for $\mathbf{x}, \mathbf{y} \in D \times D$. The empirical estimator

$$E_n v = \frac{1}{n} \sum_{i=0}^{n-1} v(X^i, X^{i+1})$$

is strongly consistent for πQv.

The next lemma follows from a suitable central limit theorem for Markov chains; apply Theorem 2 of Chan and Geyer (1994) to $\mathbf{X}^i$. Geometric ergodicity can be replaced by weaker conditions; see, e.g., Meyn and Tweedie (1993, Chapter 17) and the discussion in Tierney (1994).

Lemma 1 *Let the Markov chain be geometrically ergodic. Let $v_j(x,y)$, $j = 1, \ldots, m$, be bounded measurable functions on $D \times D$. Then the estimators $E_n v_j$, $j = 1, \ldots, m$, are jointly asymptotically normal with asymptotic covariances*

$$\begin{aligned}
\lim_{n\to\infty} n \operatorname{cov}(E_n v_j, E_n v_k) &= E\left(v_j(X^0,X^1) - Ev_j(X^0,X^1)\right) v_k(X^0,X^1) \\
&\quad + \sum_{r=2}^{\infty} \Big\{ E\Big(v_j(X^0,X^1) - Ev_j(X^0,X^1)\Big) v_k(X^{r-1},X^r) \\
&\qquad + E\Big(v_k(X^0,X^1) - Ev_k(X^0,X^1)\Big) v_j(X^{r-1},X^r) \Big\}.
\end{aligned}$$

We will often make use of the fact that the asymptotic covariance depends only on the law of the *stationary* chain.

Consider a bimeasurable transformation T on D that leaves $Q(x,dy)$ invariant in the sense that $Q(x,dy) = Q(Tx,Tdy)$. This forces π to be invariant under T. To see this, let B be measurable and write

$$\begin{aligned}
\pi(TB) &= \int \pi(dx)Q(x,TB) = \int \pi(Tdx)Q(Tx,TB) \\
&= \int \pi(Tdx)Q(x,B).
\end{aligned}$$

Hence $\pi(Tdx)$ is invariant under Q. Thus, since the invariant distribution is unique, $\pi(Tdx) = \pi(dx)$. Invariance of Q under T also implies that, for the stationary chain,

$$(X^0, X^1, \ldots, X^r) \stackrel{\mathrm{d}}{=} (TX^0, \ldots, TX^r), \tag{2.1}$$

e.g., for $r = 1$,

$$\pi(dx)Q(x, dy) = \pi(Tdx)Q(Tx, Tdy).$$

Of course, if the chain $X^0, X^1, \ldots$ is ergodic or geometrically ergodic, so is $TX^0, TX^1, \ldots$.

The following simple lemma describes the best weights for a linear combination of estimators in terms of their covariance matrix.

Lemma 2 *If X is an m-dimensional random vector with nonsingular covariance matrix Σ, then the variance of a linear combination $b'X$ is minimized over vectors b with $\sum_{j=1}^m b_j = 1$ by*

$$b = \Sigma^{-1}\mathbf{1}/\mathbf{1}'\Sigma^{-1}\mathbf{1}$$

with $\mathbf{1} = (1, \ldots, 1)'$. If Σ has equal row sums, then $b_j = 1/m$ for all j.

Note that circulant matrices have equal row sums.

Consider a bounded measurable function $f(x)$ on D. The corresponding empirical estimator is

$$E_n^0 f = \frac{1}{n}\sum_{i=0}^{n-1} f(X^i).$$

If the chain is ergodic, then $E_n^0 f$ is strongly consistent for πf. To any bimeasurable transformation T on D that leaves π invariant, there corresponds an 'empirical' estimator

$$E_n^0 f \circ T = \frac{1}{n}\sum_{i=0}^{n-1} f(TX^i).$$

It is consistent as before with $f(x)$ replaced by $f(Tx)$. The same is true for any power T^j of T. Suppose that Q is invariant under T. Under the stationary law, any convex combination of such 'empirical' estimators has smaller risk with respect to convex loss functions than the usual empirical estimator; use (2.1) and Brillinger (1963).

The following proposition answers the question of how best to use linear combinations of such estimators if the powers form a cyclic group.

Proposition 1 *Let the Markov chain be geometrically ergodic, with Q invariant under T. Suppose the asymptotic covariance matrix of $E_n^0 f \circ T^j, j = 0, \ldots, m-1$, is nonsingular, and $T^m = T^0$ for some $m \geq 2$. Then the best linear combination of $E_n^0 f \circ T^j$, $j = 0, \ldots, m-1$, in the sense of minimum asymptotic variance, is*

$$\overline{E}_n^0 f = \frac{1}{m}\sum_{j=0}^{m-1} E_n^0 f \circ T^j.$$

Proof For any j and k, the pair $E_n^0 f \circ T^k$, $E_n^0 f \circ T^j$ is the pair $E_n^0 f$, $E_n^0 f \circ T^{(j-k) \bmod m}$ evaluated with the chain $X^0, X^1, \ldots$ replaced by the chain $T^k X^0, T^k X^1, \ldots$. By (2.1) and Lemma 1, the asymptotic covariances of the two pairs above agree. Therefore, the asymptotic covariance matrix of $E_n^0 f \circ T^j$, $j = 0, \ldots, m-1$, is circulant, and the result follows by Lemma 2. □

Remark The asymptotic variance reduction of $\overline{E}_n^0 f$ is small if f is nearly invariant under T. In the extreme case, $f = f \circ T$, we have $\overline{E}_n^0 f = E_n^0 f$, and no improvement. On the other hand, even if we use only one power of T, say T itself ($m = 2$), the improvement may be dramatic if f is far from invariant under T. In the extreme case, if f is *anti-invariant*, $f - \pi f = -(f \circ T - \pi f \circ T)$, we have $\frac{1}{2}(E_n^0 f + E_n^0 f \circ T) = \pi f$. Then our estimator has asymptotic variance zero, and the relative efficiency of $E_n^0 f$ is zero. We shall discuss this further in reference to a specific example in Section 4.

3 Two-step Gibbs samplers

In this section we introduce an alternative 'empirical' estimator that exploits the structure of the transition distribution of two-step Gibbs samplers, and find the best linear combination of the usual empirical estimator and the new one. Versions of the two-step Gibbs sampler are the auxiliary variable method of Swendsen and Wang (1987), the data augmentation algorithm of Tanner and Wong (1987), and the successive substitution sampler of Gelfand and Smith (1990). A specific example of the two-step Gibbs sampler, to capture-recapture estimation, is given by George and Robert (1992).

Let $\pi(dx)$ be a probability measure on a product space $D = D_1 \times D_2$. It can be factored into marginal and conditional distributions in two ways:

$$\pi(dx) = m_1(dx_1) p_2(x_1, dx_2) = m_2(dx_2) p_1(x_2, dx_1).$$

The Gibbs sampler is defined as follows. At stage 0 pick $X^0 = (X_1^0, X_2^0)$ from some initial distribution on D. At stage i generate $X_1^i \sim p_1(X_2^{i-1}, dx_1)$ and then $X_2^i \sim p_2(X_1^i, dx_2)$. The sequence $X^0, X^1, \ldots$ is a Markov chain on D with transition distribution

$$Q(x, dy) = p_1(x_2, dy_1) p_2(y_1, dy_2).$$

The probability measure π is invariant under Q. Consider a bounded measurable function $f(x)$ on D. As in Section 2, if the chain is ergodic, the empirical estimator $E_n^0 f$ is strongly consistent for πf.

We view the sampler as an evolution of the configuration, updating the two sites one at a time. On its way from X^i to X^{i+1}, the sampler creates an intermediate configuration which, as in the Introduction, we denote by $X^{i.1} = (X_1^{i+1}, X_2^i)$. The fine chain $X^0, X^{0.1}, X^1, \ldots$ is again a Markov chain. This chain is not time-homogeneous, but has transition distributions that are periodic of order two, namely

$$\begin{aligned} Q_1(x, dy) &= p_1(x_2, dy_1) \varepsilon_{x_2}(dy_2), \\ Q_2(x, dy) &= p_2(x_1, dy_2) \varepsilon_{x_1}(dy_1), \end{aligned}$$

where ε_x is the point mass at x. The transition law of the original Gibbs sampler chain is $Q = Q_1 Q_2$. It is well known and easy to check that Q_1 and π satisfy the detailed balance equation

$$\pi(dx)Q_1(x,dy) \;=\; Q_1(y,dx)\pi(dy), \tag{3.1}$$

similarly for Q_2. That is, under the stationary law, X^0 and $X^{0.1}$ are reversible:

$$(X^0, X^{0.1}) \;\stackrel{\mathrm{d}}{=}\; (X^{0.1}, X^0). \tag{3.2}$$

In particular, $X^{0.1} \stackrel{\mathrm{d}}{=} X^0 \sim \pi$ and the 'new' empirical estimator

$$E_n^1 f = \frac{1}{n}\sum_{i=0}^{n-1} f(X^{i.1})$$

is strongly consistent for πf. Note that despite (3.2), the Gibbs sampler chain is not time-reversible unless the components of π are independent, because the transition law of the time-reversed chain (in which sites are updated in the opposite order) is $Q^* = Q_2 Q_1 \neq Q$.

Extending the proof of (3.2) inductively (this only works for two-step samplers), we obtain that the stationary fine chain is reversible:

$$(X^0, X^{0.1}, \ldots, X^{r.1}) \;\stackrel{\mathrm{d}}{=}\; (X^{r.1}, X^{r-1}, \ldots, X^0). \tag{3.3}$$

We can extract alternate components from (3.3) to obtain that the reversed interpolated chain has the same stationary law as the original Gibbs sampler:

$$(X^0, X^1, \ldots, X^r) \;\stackrel{\mathrm{d}}{=}\; (X^{r.1}, X^{(r-1).1}, \ldots, X^{0.1}). \tag{3.4}$$

In essence, this argument works because the updating $\ldots, Q_1, Q_2, Q_1, \ldots$ in reverse order is again $\ldots, Q_1, Q_2, Q_1, \ldots$. Note that the argument breaks down for $d > 2$ because none of the time-reversed interpolated chains (that update using a cyclic permutation of $Q_d, \ldots, Q_1$) have the transition distribution $Q = Q_1 \ldots Q_d$ of the original Gibbs sampler.

The result of this section is that the best linear combination of $E_n^0 f$ and $E_n^1 f$ in the sense of asymptotic variance has equal weights. We already know from (3.4) that $E_n^0 f$ and $E_n^1 f$ have equal variances under the stationary law.

Theorem 1 *If the Gibbs sampler is geometrically ergodic, then the best linear combination of $E_n^0 f$ and $E_n^1 f$ is*

$$G_n f = \tfrac{1}{2}(E_n^0 f + E_n^1 f).$$

Proof Trivial algebra or Lemma 2 shows that equal weights are optimal if $E_n^0 f$ and $E_n^1 f$ have equal asymptotic variances. First consider $E_n^0 f$. Apply Lemma 1 with both v_j and v_k equal to $v^0(x,y) = f(x)$. The asymptotic variance of $E_n^0 f$ is a sum of centering terms $(Ev^0(X^0, X^1))^2 = (\pi f)^2$ and of the terms

$$Ev^0(X^0, X^1)v^0(X^{r-1}, X^r) = Ef(X^0)f(X^{r-1}).$$

Similarly, with both v_j and v_k equal to $v^1(x,y) = f(y_1, x_2)$, the asymptotic variance of $E_n^1 f$ is a sum of centering terms $(Ev^1(X^0, X^1))^2 = (Ef(X^{0.1}))^2 = (\pi f)^2$ and of the terms

$$\begin{aligned} Ev^1(X^0, X^1)v^1(X^{r-1}, X^r) &= Ef(X^{0.1})f(X^{(r-1).1}) \\ &= Ef(X^{(r-1).1})f(X^{0.1}). \end{aligned}$$

By (3.4), the asymptotic variances of $E_n^0 f$ and $E_n^1 f$ are equal. □

Theorem 1 extends immediately to any two-step variable-at-a-time updating scheme for which each step satisfies detailed balance (e.g., local Metropolis–Hastings algorithms).

The estimator $G_n f$ suggested in Theorem 1 cannot be improved asymptotically, except by using information about π. This follows from an efficiency result for Gibbs samplers, based on a version of the Hájek–LeCam convolution theorem, which we prove in Greenwood et al. (1995). The asymptotic variance of $G_n f$ is $\frac{1}{2}\sigma^2(1+\rho)$, where σ^2 is the asymptotic variance of $E_n^0 f$ or $E_n^1 f$, and ρ is their asymptotic correlation coefficient. There is a reduced asymptotic variance for all $\rho < 1$; the reduction is 50% when $\rho = 0$.

Simulation Take π to be the uniform distribution on the triangle $\{x\colon x_1, x_2 > 0, x_1 + x_2 < 1\}$. The conditional law $p_1(x_1, dx_2)$ is uniform on the interval $(0, 1-x_1)$, similarly for p_2. Let $f(x)$ be the indicator of the smaller triangle $\{x\colon x_1, x_2 > 0, x_1 + x_2 < 2/3\}$ so that $\pi f = 4/9$. Based on 1000 runs of the Gibbs sampler with $n = 1000$, our estimator $G_n f$ gave a 19% reduction in variance over the empirical estimator $E_n^0 f$. The additional computation time for $G_n f$ was negligible. This example will be used for other simulations in Section 4.

Remark Theorem 1 does not generalize to d-step Gibbs samplers, for $d > 2$. Recall from the Introduction that the fine chain is $X^0, X^{0.1}, \ldots, X^{0.(d-1)}, X^1, X^{1.1}, \ldots$, and there are d consistent empirical estimators $E_n^j f$, $j = 0, \ldots, d-1$. To see that equal weights need not be optimal, take $d = 3$ and choose π with the first component independent of the other two. Then the updates at sites 2 and 3 arise from a two-step Gibbs sampler, and Theorem 1 is applicable provided f depends only on the last two components. Thus the best linear combination is $\frac{1}{2}(E_n^1 f + E_n^2 f)$, which differs from the average of the three:

$$\tfrac{1}{3}(E_n^0 f + E_n^1 f + E_n^2 f) = \tfrac{2}{3}E_n^1 f + \tfrac{1}{3}E_n^2 f.$$

This counter-example also shows that equal weights are not optimal even if π is exchangeable; take π to have i.i.d. components. Note that in this case equal weights are not optimal, even though the asymptotic variances are equal, since then the higher order terms in (2.1) vanish and $X^{0.j} \sim \pi$. Moreover, a continuity argument implies that suboptimality of equal weights occurs for general f and π, contrary to a remark of Geweke (1992) that equal weights are asymptotically efficient. To estimate the optimal weights, apply Lemma 2 in conjunction with a consistent estimator (based on the original output from the Gibbs sampler) of the asymptotic covariance matrix Σ of $E_n^0 f, \ldots, E_n^{d-1} f$. Consistent estimators of Σ are available from, e.g., Geyer (1992), who discusses the methods of batch means and window estimators. However, simple averages of more than just two of the $E_n^j f$ can perform well; see the simulation study in Section 5.

The equal asymptotic variance property of the empirical estimators $E_n^j f$ does not hold in general; we have constructed examples of d-step Gibbs samplers with $d > 2$, and functions f for which the asymptotic variances of the $E_n^j f$ do not coincide.

A d-step sampler can be treated as a two-step sampler by merging the first j components and also the last $d - j$ components, for some $1 \leq j < d$. Theorem 1 can be applied to the resulting two-step sampler provided $Q_1 Q_2 \dots Q_j$ and $Q_{j+1} \dots Q_d$ are reversible. This holds for the samplers of nearest neighbor random fields studied in Section 5.

4 Two-step Gibbs samplers and invariance

We continue to study the Gibbs sampler for a distribution $\pi(dx)$ on a product space $D = D_1 \times D_2$. If π has symmetries, can they be used to improve the estimator $G_n f$ that we introduced in Section 3? As in Section 2, we describe symmetries in terms of transformations that leave π invariant. Applications to Markov random fields on a lattice suggest two types of transformations which we call parallel and transverse. These give rise to further 'empirical' estimators that we combine with estimators arising from the interpolated chain.

We call a transformation T: $D_1 \times D_2 \to D_1 \times D_2$ *parallel* if it is a direct product $T(x_1, x_2) = (T_1 x_1, T_2 x_2)$; we call it *transverse* if $T(x_1, x_2) = (T_{21} x_2, T_{12} x_1)$. Note that the composition of two transverse transformations is parallel, and the composition of a parallel with a transverse is transverse.

First we treat parallel transformations. Suppose that T is parallel and leaves π invariant. We can write

$$\pi(dx) = m_1(dx_1) p_2(x_1, dx_2) = m_2(dx_2) p_1(x_2, dx_1)$$

and

$$\pi(Tdx) = m_1(T_1 dx_1) p_2(T_1 x_1, T_2 dx_2) = m_2(T_2 dx_2) p_1(T_2 x_2, T_1 dx_1).$$

Hence

$$m_1(dx_1) = m_1(T_1 dx_1), \quad m_2(dx_2) = m_2(T_2 dx_2), \tag{4.1}$$

and

$$p_1(x_2, dx_1) = p_1(T_2 x_2, T_1 dx_1), \quad p_2(x_1, dx_2) = p_2(T_1 x_1, T_2 dx_2). \tag{4.2}$$

Consider the transition distribution of the Gibbs sampler, $Q(x, dy) = p_1(x_2, dy_1) p_2(y_1, dy_2)$. By (4.2), the transition distribution is invariant in the sense of Section 2: $Q(x, dy) = Q(Tx, Tdy)$. The transition distributions from X^0 to $X^{0.1}$ and $X^{0.1}$ to X^1, respectively, are Q_1 and Q_2, defined in Section 3. By (4.2) the joint law of X^0 and $X^{0.1}$ is

$$\pi(dx) Q_1(x, dy) = \pi(Tdx) Q_1(Tx, Tdy).$$

Continue with the step from $X^{0.1}$ to X^1, and so on, to obtain

$$(X^0, X^{0.1}, \dots, X^r) \stackrel{\mathrm{d}}{=} (TX^0, TX^{0.1}, \dots, TX^r). \tag{4.3}$$

Now suppose that T is transverse and leaves π invariant. We can write

$$\pi(Tdx) = m_1(T_{21}dx_2)p_2(T_{21}x_2, T_{12}dx_1) = m_2(T_{12}dx_1)p_1(T_{12}x_1, T_{21}dx_2).$$

Comparing with the factorizations of $\pi(dx)$,

$$m_1(dx_1) = m_2(T_{12}dx_1), \quad m_2(dx_2) = m_1(T_{21}dx_2), \tag{4.4}$$

and

$$p_1(x_2, dx_1) = p_2(T_{21}x_2, T_{12}dx_1), \quad p_2(x_1, dx_2) = p_1(T_{12}x_1, T_{21}dx_2). \tag{4.5}$$

By (4.5) we obtain for the joint law of $X^0, X^{0.1}$:

$$\begin{aligned}
\pi(dx)Q_1(x, dy) &= m_1(dx_1)p_2(x_1, dx_2)p_1(x_2, dy_1)\varepsilon_{x_2}(dy_2) \\
&= m_2(T_{12}dx_1)p_1(T_{12}x_1, T_{21}dx_2)p_2(T_{21}x_2, T_{12}dy_1)\varepsilon_{T_{21}x_2}(T_{21}dy_2) \\
&= \varepsilon_{T_{21}y_2}(T_{21}dx_2)p_2(T_{21}y_2, T_{12}dx_1)p_1(T_{12}y_1, T_{21}dy_2)m_2(T_{12}dy_1) \\
&= Q_1(Ty, Tdx)\pi(Tdy).
\end{aligned} \tag{4.6}$$

Continue with the step from $X^{0.1}$ to X^1, and so on, to obtain that the transformed time-reversed fine chain has the same stationary law as the original fine chain:

$$(X^0, X^{0.1}, \ldots, X^r) \stackrel{\mathrm{d}}{=} (TX^r, TX^{(r-1).1}, \ldots, TX^0). \tag{4.7}$$

Suppose that π is invariant under a transformation T that is either parallel or transverse. Since both X^0 and $X^{0.1}$ have stationary distribution π, we have, under ergodicity of the original chain, two strongly consistent empirical estimators for πf:

$$\begin{aligned}
E_n^0 f \circ T &= \frac{1}{n}\sum_{i=0}^{n-1} f(TX^i), \\
E_n^1 f \circ T &= \frac{1}{n}\sum_{i=0}^{n-1} f(TX^{i.1}).
\end{aligned}$$

The same is true if we replace T by powers of T. We show now that the best linear combination of all these estimators is the average if the powers of T form a cyclic group.

Theorem 2 *Let the Gibbs sampler be geometrically ergodic. Let π be invariant under a parallel or transverse transformation T. Suppose the asymptotic covariance matrix of $E_n^0 f \circ T^j, E_n^1 f \circ T^j$, $j = 0, \ldots, m-1$, is nonsingular, and $T^m = T^0$. Then the empirical estimators $E_n^0 f \circ T^j, E_n^1 f \circ T^j$, $j = 0, \ldots, m-1$, have equal asymptotic variances, and the best linear combination is*

$$\overline{G}_n f = \frac{1}{m}\sum_{j=0}^{m-1} G_n f \circ T^j.$$

Proof Let T be parallel. The proof for T transverse is similar. Order the estimators in pairs $E_n^0 f \circ T^j, E_n^1 f \circ T^j$, $j = 0, \ldots, m-1$. The covariance matrix has $m \cdot m$ 2×2-submatrices. Our strategy will be to show that it is block-circulant of the form:

$$\begin{pmatrix} A_0 & A_1 & \ldots & A_2 & A_1 \\ A_1 & A_0 & \ldots & A_3 & A_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ A_2 & A_3 & \ldots & A_0 & A_1 \\ A_1 & A_2 & \ldots & A_1 & A_0 \end{pmatrix}$$

where $A_0, A_1 \ldots$ are circulant 2×2-submatrices. Such a matrix has equal row sums. The result will then follow from Lemma 2.

First we show that the submatrices in row 0 are circulant. The first submatrix on the main diagonal has equal elements on its diagonal by Theorem 1 and is clearly symmetric, hence circulant. Now consider the $(0,1)$ submatrix. We look first at its diagonal. By Lemma 1, applied with $v_j(x,y) = f(x)$ and $v_k(x,y) = f(Tx)$, the asymptotic covariance of $E_n^0 f$ and $E_n^0 f \circ T$ is, aside from centering terms, a sum of the terms

$$\begin{aligned} &Ev_j(X^0, X^1)v_k(X^{r-1}, X^r) + Ev_k(X^0, X^1)v_j(X^{r-1}, X^r) \\ &= Ef(X^0)f(TX^{r-1}) + Ef(TX^0)f(X^{r-1}). \end{aligned}$$

The corresponding terms in the asymptotic covariance of $E_n^1 f$ and $E_n^1 f \circ T$ are

$$\begin{aligned} &Ef(X^{0.1})f(TX^{(r-1).1}) + Ef(TX^{0.1})f(X^{(r-1).1}) \\ &= Ef(X^{(r-1).1})f(TX^{0.1}) + Ef(TX^{(r-1).1})f(X^{0.1}). \end{aligned}$$

By (3.3), the asymptotic covariances are equal.

We now look at the off-diagonal elements of the same $(0,1)$ submatrix. Similarly to the above, aside from centering, the asymptotic covariance of $E_n^1 f$ and $E_n^0 f \circ T$ is a sum of the terms

$$Ef(X^{0.1})f(TX^{r-1}) + Ef(TX^{0.1})f(X^{r-1}).$$

The corresponding terms in the asymptotic covariance of $E_n^0 f$ and $E_n^1 f \circ T$ are

$$\begin{aligned} &Ef(X^0)f(TX^{(r-1).1}) + Ef(TX^0)f(X^{(r-1).1}) \\ &= Ef(X^{(r-1).1})f(TX^0) + Ef(TX^{(r-1).1})f(X^0). \end{aligned}$$

By (3.3), the asymptotic covariances are equal.

The $(0,j)$ submatrix is the same as the $(0,1)$ submatrix with T replaced by T^j. Hence it is also circulant. By (4.3),

$$(X^0, X^{0.1}, \ldots, X^r) \stackrel{\mathrm{d}}{=} (T^k X^0, T^k X^{0.1}, \ldots, T^k X^r).$$

As in the proof of Proposition 1, one sees that the blocks on each diagonal are equal, and it follows that the covariance matrix is block-circulant. □

Note that equal asymptotic variances of the empirical estimators does not imply the optimality of equal weights per se; Lemma 2 shows that the best weights depend on the asymptotic covariances as well.

In the covariance matrix in the above proof, omit every second row and column. The resulting covariance matrix is circulant. Hence, if π is invariant under a transformation T that is either parallel or transverse, and $T^m = T^0$, and if the asymptotic covariance matrix of the estimators $E_n^0 f \circ T^j$, $j = 0, \ldots, m-1$ is nonsingular, then the best linear combination of them is the average, $\overline{E}_n^0 f$. Similarly, the best linear combination of $E_n^1 f \circ T^j$, $j = 0, \ldots, m-1$, is the average, $\overline{E}_n^1 f$.

If π is invariant under several transformations, we can do better than in Theorem 2. For simplicity, we consider only two transformations T and U and give conditions for the best linear combination of all the corresponding 'empirical' estimators to be the average.

Theorem 3 *Let the Gibbs sampler be geometrically ergodic. Let π be invariant under two commuting transformations T and U each of which is either parallel or transverse. Suppose the asymptotic covariance matrix of $E_n^0 f \circ U^{j_2} T^{j_1}$, $E_n^1 f \circ U^{j_2} T^{j_1}$, $j_1 = 0, \ldots, m_1 - 1$, $j_2 = 0, \ldots, m_2 - 1$, is nonsingular, and $T^{m_1} = T^0$, $U^{m_2} = U^0$. Then the empirical estimators $E_n^0 f \circ U^{j_2} T^{j_1}$, $E_n^1 f \circ U^{j_2} T^{j_1}$, $j_1 = 0, \ldots, m_1 - 1$, $j_2 = 0, \ldots, m_2 - 1$, have equal asymptotic variances, and the best linear combination is*

$$\frac{1}{m_1 m_2} \sum_{j_1=0}^{m_1-1} \sum_{j_2=0}^{m_2-1} G_n f \circ U^{j_2} T^{j_1}.$$

Proof Let T and U be parallel; the proof for other combinations of parallel and transverse transformations is similar. The covariance matrix has $m_2 \cdot m_2$ blocks each consisting of $m_1 \cdot m_1$ 2×2-submatrices. Our strategy will be to show that it is block-circulant of the form:

$$\left(\begin{array}{ccc|ccc|c|ccc}
A_1 & A_2 & & B_1 & B_2 & & & B_1 & B_2 & \\
A_2 & A_1 & \ddots & B_{m_1} & B_1 & \ddots & \ldots & B_{m_1} & B_1 & \ddots \\
 & \ddots & \ddots & & \ddots & \ddots & & & \ddots & \ddots \\
\hline
B_1 & B_{m_1} & & A_1 & A_2 & & & C_1 & C_2 & \\
B_2 & B_1 & \ddots & A_2 & A_1 & \ddots & \ldots & C_{m_1} & C_1 & \ddots \\
 & \ddots & \ddots & & \ddots & \ddots & & & \ddots & \ddots \\
\hline
 & \vdots & & & \vdots & & \ddots & & \vdots & \\
 & & & & & & & & &
\end{array}\right)$$

in which each block is a block-circulant matrix having circulant 2×2-submatrices denoted by $A_1, A_2, \ldots$ or $B_1, B_2, \ldots$ etc. Such a matrix has equal row sums. The result will then follow from Lemma 2.

Recall that an $m \times m$ matrix A is circulant if the elements in each diagonal are equal and $A_{1j} = A_{m-j,1}$ for $j = 1, \ldots, m$.

We prove first that the 2×2-submatrices are circulant. The (j_1, j_1') submatrix of the (j_2, j_2') block equals the (j_1, j_1') submatrix of the $(0,0)$ block, with $T^{j_1}, T^{j_1'}$ replaced by $T^{j_1}, U^{j_2'-j_2}T^{j_1'}$ and f replaced by $f \circ U^{j_2}$. The $(0,0)$ block is the covariance matrix of Theorem 2. In particular, its 2×2-submatrices are circulant.

Now we prove that the blocks in row 0 are circulant. Consider the $(0, j_2)$ block. We show that its $(j_1', j_1 + j_1')$ submatrix equals the $(0, j_1)$ submatrix. Since the submatrices are circulant, it suffices to compare the upper rows. The upper left elements are the asymptotic covariances of the pairs $E_n^0 f, E_n^0 f \circ U^{j_2}T^{j_1}$ and $E_n^0 f \circ T^{j_1'}, E_n^0 f \circ U^{j_2}T^{j_1+j_1'}$, respectively. By Lemma 1 the first covariance is, aside from centering terms, a sum of the terms

$$Ef(X^0)f(U^{j_2}T^{j_1}X^{r-1}) + Ef(U^{j_2}T^{j_1}X^0)f(X^{r-1}). \tag{4.8}$$

The corresponding term of the second covariance is

$$Ef(T^{j_1'}X^0)f(U^{j_2}T^{j_1+j_1'}X^{r-1}) + Ef(U^{j_2}T^{j_1+j_1'}X^0)f(T^{j_1'}X^{r-1}).$$

By (4.3) with $T = T^{j_1'}$, this term equals (4.8). The upper right elements are compared similarly.

We show that the $(m_1 - j_1, 0)$ submatrix of the $(0, j_2)$ block equals the $(0, j_1)$ submatrix. The upper left element of the $(m_1 - j_1, 0)$ submatrix is, aside from centering terms, a sum of terms

$$Ef(T^{m_1-j_1}X^0)f(U^{j_2}X^{r-1}) + Ef(U^{j_2}X^0)f(T^{m_1-j_1}X^{r-1}).$$

Use (4.3) with $T = T^{j_1}$ and $T^{m_1} = I$ to see that this term equals the corresponding term (4.8) of the $(0, j_1)$ submatrix. The upper right elements are compared similarly. This proves that the blocks are circulant.

Now we show that the block matrix is circulant. We show that the (j_2', j_2+j_2') block equals the $(0, j_2)$ block. Since the blocks are circulant and consist of circulant 2×2-submatrices, it suffices to compare the upper rows of the 2×2-submatrices in the upper rows of the two blocks. The upper left element of the $(0, j_1)$ submatrix of the $(j_2', j_2 + j_2')$ block is, aside from centering terms, a sum of terms

$$Ef(U^{j_2'}X^0)f(U^{j_2+j_2'}T^{j_1}X^{r-1}) + Ef(U^{j_2+j_2'}T^{j_1}X^0)f(U^{j_2'}X^{r-1}).$$

Use $U^{j_2'}T^{j_1} = T^{j_1}U^{j_2'}$ and (4.3) with $T = U^{j_2'}$ to see that this term equals the corresponding term (4.8) of the $(0, j_2)$ block. The upper right elements are compared similarly.

We show that the $(m_2 - j_2, 0)$ block equals the $(0, j_2)$ block. The upper left element of the $(0, j_1)$ submatrix of the $(m_2 - j_2, 0)$ block is, aside from centering terms, a sum of terms

$$Ef(U^{m_2-j_2}X^0)f(T^{j_1}X^{r-1}) + Ef(T^{j_1}X^0)f(U^{m_2-j_2}X^{r-1}).$$

Use (4.3), $U^{m_2} = I$, and $T^{j_1}U^{j_2} = U^{j_2}T^{j_1}$ to see that this term equals the corresponding term (4.8) of the $(0, j_2)$ block. The upper right elements are treated similarly. This shows that the covariance matrix is a circulant block matrix, and the proof is complete. □

Simulation We continue the simulation example from Section 3, with π the uniform distribution on the triangle $\{x: x_1, x_2 > 0, x_1 + x_2 < 1\}$. Note that π is exchangeable, i.e., invariant under $T(x_1, x_2) = (x_2, x_1)$. Let $f(x)$ be the indicator of the asymmetric triangle $\{x: x_1, x_2 > 0, 2x_2 < x_1\}$, so that $\pi f = 1/3$. We shall consider the symmetrized estimators $\overline{G}_n f$ and $\overline{E}_n^0 f$, with $m = 2$ and the transformations I, T. The simulations were based on 1000 runs of the Gibbs sampler with $n = 1000$. Compared to the empirical estimator $E_n^0 f$, the variance reductions of $G_n f$, $\overline{E}_n^0 f$, $\overline{G}_n f$ were 9%, 82%, 86%, respectively. In particular, compared to the symmetrized empirical estimator $\overline{E}_n^0 f$, the variance reduction of $\overline{G}_n f$ is 24%. The improvement through symmetrization by T is particularly impressive in this example, because T is close to being anti-invariant, in the sense of the Remark at the end of Section 2. If we had taken $f(x)$ to be the indicator of the triangle $\{x: x_1, x_2 > 0, x_2 < x_1\}$, then $\overline{E}_n^0 f = \overline{G}_n f = 1/2 = \pi f$, and the variances would be zero.

5 Nearest neighbor random fields

Consider a random field on the rectangular lattice

$$S = \{0, \ldots, k_1 - 1\} \times \{0, \ldots, k_2 - 1\},$$

where k_1 and k_2 are even. The lattice has $d = k_1 k_2$ sites. The configuration space is $D = V^S$, where V is a measurable state space. The random field is described by a probability measure $\pi(dy)$ on D. One can factor $\pi(dy)$ in d ways into a $(d-1)$-dimensional marginal and a one-dimensional conditional, as $\pi(dy) = m_s(dy_{-s}) q_s(y_{-s}, dy_s)$, where $y_{-s} = (y_r)_{r \in S \setminus \{s\}}$. We make the assumption that $q_s(y_{-s}, dy_s)$ depends on y_{-s} only through the values of y at the four nearest neighbors of the site $s = (s_1, s_2)$, i.e., $(s_1 \pm 1, s_2)$ and $(s_1, s_2 \pm 1)$ if they are in S. This is a nearest neighbor model with free boundary. Later we consider other types of boundary.

We call the site $s = (s_1, s_2)$ even or odd according to the parity of $s_1 + s_2$. The even and odd sites form a checkerboard pattern. Even sites have only odd neighbors, so the conditional law at an odd site depends only on the values of the field at even sites, and vice versa. To take advantage of the nearest neighbor structure, one fixes a sweep by numbering first the even and then the odd sites. The sampler first updates the even sites, using only the odd, and then vice versa. These two steps we think of as a two-step Gibbs sampler. To this sampler, we apply the results of Sections 3 and 4.

Label the sites $s_1, \ldots, s_d$. Write $y = (y_1, \ldots, y_d)$ for $(y_{s_1}, \ldots, y_{s_d})$, and $q_j(y_{-j}, dy_j)$ for $q_{s_j}(y_{-s_j}, dy_{s_j})$, where $y_{-j} = (y_{<j}, y_{>j})$ with $y_{<j} = (y_1, \ldots, y_{j-1})$ and $y_{>j} = (y_{j+1}, \ldots, y_d)$. The Gibbs sampler goes from y^0 to y^1 by the transition distribution

$$\begin{aligned} Q(y^0, dy^1) &= \prod_{j=1}^{d} q_j(y^1_{<j}, y^0_{>j}, dy^1_j) \\ &= \prod_{j=1}^{d/2} q_j(y^1_{<j}, y^0_{>j}, dy^1_j) \prod_{j=d/2+1}^{d} q_j(y^1_{<j}, y^0_{>j}, dy^1_j). \end{aligned} \tag{5.1}$$

The even sites are numbered $1, \ldots, d/2$, the odd sites $d/2+1, \ldots, d$. Set

$$x = (x_1, x_2), \quad x_1 = (y_1, \ldots, y_{d/2}), \quad x_2 = (y_{d/2+1}, \ldots, y_d).$$

The first product in (5.1) updates the even sites. Consider the transition from x^0 to x^1. Since for $j \leq d/2$ the transition distribution depends only on values at odd sites, which have not yet been updated, i.e.,

on x_2^0, the transition distribution for updating the entire set of even sites can be written

$$p_1(x_2^0, dx_1^1) = \prod_{j=1}^{d/2} q_j(y_{<j}^0, y_{>j}^0, dy_j^1),$$

which coincides with the conditional distribution on the even sites given the odd sites. This is the first step of the two-step sampler. In the second step we update the odd sites, the even sites having already been updated. The transition distribution for updating the set of odd sites can be written

$$p_2(x_1^1, dx_2^1) = \prod_{j=d/2+1}^{d} q_j(y_{<j}^1, y_{>j}^1, dy_j^1),$$

which coincides with the conditional distribution on the odd sites given the even sites. Now we have a two-step Gibbs sampler with transition distribution

$$Q(y^0, dy^1) = p_1(x_2^0, dx_1^1) p_2(x_1^1, dx_2^1)$$

and are in the setting of Sections 3 and 4 with $D = D_1 \times D_2$, where $D_1 = V^{S_e}$, $D_2 = V^{S_o}$. Here S_e and S_o are the sets of even and odd sites, respectively. The Gibbs sampler generates a Markov chain

$$(Y_s^i)_{s \in S} = (Y_j^i)_{j=1,\ldots,d} = (X_1^i, X_2^i) = X^i, \ i = 0, 1, \ldots$$

with some arbitrarily chosen initial value X^0. The interpolated chain, in the sense of Section 3, is

$$X^{i.1} = (X_1^{i+1}, X_2^i), \ i = 0, 1, \ldots.$$

Usually the simulations $X^0, X^1, \ldots$ are utilized for approximating the expectation of a bounded measurable function $f(y)$ on D through the empirical estimator $E_n^0 f$. In Section 3 we discussed the alternative empirical estimator $E_n^1 f$. By Theorem 1, if the Gibbs sampler is geometrically ergodic, the best linear combination of $E_n^0 f$ and $E_n^1 f$ is $G_n f = \frac{1}{2}(E_n^0 f + E_n^1 f)$. The latter holds, more generally, for any local Metropolis–Hastings sampler having a checkerboard sweep and local updates that depend only on a site and its nearest neighbors; all we need is reversibility of the composition of the local updates over S_e, and the same for S_o.

We now explore some of the homogeneities that a nearest neighbor random field might possess. For instance, π is spatially homogeneous if it is invariant under all translations

on the lattice, in which case the conditional distributions q_s are identical. Or π might be invariant under shifts in a certain direction or have periodicities. We exploit symmetries through corresponding transformations that are permutations of the sites. As described in Section 4, they generate additional 'empirical' estimators that we use to further improve $G_n f$. Define addition on S by $(s+t)_1 = s_1 + t_1 \bmod k_1$, $(s+t)_2 = s_2 + t_2 \bmod k_2$. For $t \in S$, the translation of S by t is defined as $T_t s = s - t$. This induces a translation on D, $(T_t x)_s = x_{T_t^{-1} s} = x_{s+t}$.

Horizontal translations. Think of the lattice as wrapped around a cylinder so that the vertical boundaries meet. The neighbors of each site $s = (s_1, s_2)$ along the vertical boundary now include $(s_1 \pm 1, s_2)$ with addition mod k_1.

A horizontal translation by an *even* number of sites is $T = T_{(p,0)}$, with p even. This translation takes even into even sites and odd into odd and is a parallel transformation in the sense of Section 4. Suppose that k_1 is a multiple of p, say $k_1 = mp$. Suppose that π is invariant under T. Then it is also invariant under powers $T^j = T_{(jp,0)}, j = 0, \ldots, m-1$. These transformations form a cyclic group. Theorem 2 implies that the empirical estimators

$$E_n^0 f \circ T_{(jp,0)},\ E_n^1 f \circ T_{(jp,0)}, \quad j = 1, \ldots, m-1,$$

have equal asymptotic variances, and the best linear combination is the average,

$$\overline{G}_n f = \frac{1}{m} \sum_{j=0}^{m-1} G_n f \circ T_{(jp,0)}.$$

A horizontal translation by an *odd* number of sites is $T = T_{(p,0)}$, with p odd. This translation takes even into odd sites and odd into even and is a transverse transformation. Even powers of T are parallel. Suppose that k_1 is a multiple of p, say $k_1 = mp$. Suppose that π is invariant under T. As above, the best linear combination of the corresponding empirical estimators is the average.

Horizontal and vertical translations. Think of the lattice as wrapped around a torus. The neighbors of each site $s = (s_1, s_2)$ along the boundaries now include $(s_1 \pm 1 \bmod k_1, s_2)$ and $(s_1, s_2 \pm 1 \bmod k_2)$, giving periodic boundary conditions. Suppose that $k_1 = m_1 p_1$ and $k_2 = m_2 p_2$. Suppose that π is invariant under both the horizontal translation $T = T_{(p_1,0)}$ and the vertical translation $T = T_{(0,p_2)}$. Theorem 3 implies that the empirical estimators

$$E_n^0 f \circ T_{(j_1 p_1, j_2 p_2)},\ E_n^1 f \circ T_{(j_1 p_1, j_2 p_2)}, \quad j_1 = 1, \ldots, m_1 - 1,\ j_2 = 1, \ldots, m_2 - 1$$

have equal asymptotic variances, and the best linear combination is the average,

$$\frac{1}{m_1 m_2} \sum_{j_1=0}^{m_1-1} \sum_{j_2=0}^{m_2-1} G_n f \circ T_{(j_1 p_1, j_2 p_2)}.$$

For $j = 0, \ldots, d-1$, let $E_n^j f$ be the empirical estimator of the Introduction, which is based on the configurations obtained after j of the sites have been updated in each sweep. Simple averages of some or all of the $E_n^j f$ may be used instead of $E_n^0 f$ or $G_n f$, for instance:

$$G_n^m f = \frac{1}{m} \sum_{j=0}^{m-1} E_n^{jd/m} f,$$

where d is divisible by m. However, except for $G_n^2 f = G_n f$, no optimality results are available for such estimators.

Ising model simulations Consider the classical two-dimensional Ising model used to study ferromagnetic materials; see, e.g., Kindermann and Snell (1980). In this case the state space is $V = \{+1, -1\}$, representing two spin orientations. Under the Gibbs distribution π, a configuration $y \in V^S$ has mass proportional to $\exp(-H(y))$, where the energy function H is given by $H(y) = -\beta \sum_{\langle s,t \rangle} y_s y_t$. Here β is the inverse temperature and the sum is over all sets $\langle s, t \rangle$ of neighbors s, t. We call a function *nearest neighbor* if it is of the form $f(y) = \sum_{\langle s,t \rangle} f_{st}(y_s, y_t)$. Estimating πf well means estimating πf_{st} well. But for functions f_{st} there are only two essentially different empirical estimators $E_n^j f_{st}$, namely $E_n f_{st}$ and $E_n^j f_{st}$ with j the number of the first of the two sites s, t in the sweep. We therefore expect that for nearest neighbor functions the estimators $G_n^m f$ are not better than $G_n f$, and in fact slightly worse because for different $\langle s, t \rangle$ the two essentially different empirical estimators do not arise equally often.

Specifically, let $f(y) = N^{-1} \sum_{\langle s,t \rangle} y_s y_t$ be the nearest neighbor correlation. Here N is the number of sets $\langle s, t \rangle$. Write $\rho = \pi f$. The results are given in Tables 1 to 3.

Table 1. Ising model, 4×4 lattice, free boundary. Expected nearest neighbor correlation ρ, $\sigma^2 = 10^4 \times$ variance of $E_n^0 f$, and variance reductions for $G_n f$, $G_n^4 f$, and $G_n^d f$. There were $n = 1000$ sweeps in each run. Each figure in the table was based on 1000 runs. A 'burn-in' of 1000 sweeps was used in each case to give approximate 'convergence to stationarity.'

β	0.05	0.1	0.2	0.3	0.4	0.5
ρ	0.05	0.10	0.21	0.33	0.47	0.62
σ^2	0.41	0.45	0.51	0.79	1.40	1.90
$G_n f$	49%	47%	32%	13%	2%	1%
$G_n^4 f$	47%	46%	29%	12%	1%	1%
$G_n^d f$	48%	45%	30%	12%	2%	1%

Table 2. Periodic boundary, 4×4 lattice.

β	0.05	0.1	0.2	0.3	0.4	0.5
ρ	0.05	0.10	0.23	0.42	0.69	0.88
σ^2	0.32	0.34	0.59	1.82	2.78	1.05
$G_n f$	49%	41%	10%	1%	0%	0%
$G_n^4 f$	47%	38%	9%	1%	0%	0%
$G_n^d f$	46%	36%	8%	1%	0%	0%

Table 3. Larger lattices, free boundary.

	$\beta = 0.1$			$\beta = 0.2$		
lattice	4×4	6×6	8×8	4×4	6×6	8×8
ρ	0.10	0.10	0.10	0.21	0.21	0.21
σ^2	0.45	0.18	0.09	0.51	0.23	0.12
$G_n f$	47%	44%	38%	32%	24%	22%
$G_n^4 f$	46%	42%	36%	29%	24%	22%
$G_n^d f$	45%	44%	37%	30%	24%	22%

The greatest improvements are obtained under moderate nearest neighbor dependence. The differences between Tables 1 and 2 are explained by the tendency of periodic boundary conditions to increase dependence, most markedly in small lattices. For functions f that are not invariant under translations, we find that symmetrizations of $G_n^m f$ can produce further variance reductions, but the extent of the reduction is highly dependent on the degree of asymmetry in f.

In general, $E_n^0 f$ and $E_n^{d/2} f$ are fairly strongly correlated, even for high temperatures and for nearest neighbor functions, because the subconfigurations on the odd sites used by the two estimators are *identical*. For nearest neighbor correlation, however, the two estimators are nearly independent at high temperatures. As pointed out after Theorem 1, the variance reduction is then close to 50% . The simulation results confirm this.

Next consider the Ising model with an external field. The Gibbs distribution now has energy function $H(y) = -\beta \sum_{\langle s,t \rangle} y_s y_t - h \sum_s y_s z_s$, where $z = (z_s)$ is an observed configuration representing an inhomogeneous external field, and h is the external field strength. Such an energy function arises in Bayesian image analysis as a posterior energy function, see Winkler (1995, p.31). In that case, channel noise modifies the unknown 'true image' by independently changing the color (± 1) of each pixel with probability p, to produce the observed image z. The external field strength is given by $h = (1/2)\log((1-p)/p)$. We used the same function f as before, so ρ is now the posterior-expected nearest neighbor correlation. We used two different observed images z, one having nearest neighbor correlation 0.33 and the other 0.083. The results are given in Tables 4 and 5.

Table 4. External field, observed image with nearest neighbor correlation $f(z) = 0.33$, free boundary, 4×4 lattice, $\beta = 0.1$.

p	0.05	0.1	0.2	0.3	0.4	0.5
ρ	0.30	0.27	0.20	0.15	0.11	0.10
σ^2	0.12	0.21	0.34	0.41	0.42	0.45
$G_n f$	6%	9%	21%	28%	41%	47%
$G_n^4 f$	6%	9%	20%	27%	39%	46%
$G_n^d f$	6%	9%	21%	26%	40%	45%

Table 5. External field, observed image with nearest neighbor correlation $f(z) = 0.083$, free boundary, 4×4 lattice, $\beta = 0.1$.

p	0.05	0.1	0.2	0.3	0.4	0.5
ρ	0.08	0.08	0.09	0.09	0.10	0.10
σ^2	0.06	0.11	0.24	0.34	0.42	0.45
$G_n f$	14%	17%	33%	45%	43%	47%
$G_n^4 f$	13%	15%	31%	42%	41%	46%
$G_n^d f$	13%	16%	32%	43%	41%	45%

The largest variance reductions are obtained when the channel noise is high (p close to 0.5), or equivalently, when the external field strength is small. This is explained by the posterior distribution becoming degenerate—concentrating its mass at the observed image—when the channel noise is low. This effect is less pronounced when the observed image has lower nearest neighbor correlation (compare Tables 4 and 5), which is consistent with our earlier remark that the greatest improvements are obtained under moderate nearest neighbor dependence.

As a final example consider the Metropolis sampler (see, e.g., Winkler, 1995, Ch. 8) with checkerboard sweep over the 4×4 lattice with free boundary and no external field. The proposal at each site is a spin-flip. We find a greater improvement than under the Gibbs sampler (compare Tables 1 and 6). Moreover, the variance of $G_n f$ is far less under the Metropolis sampler than under the Gibbs sampler, at all temperatures. This contrasts with the performance of the usual empirical estimator, which has smaller variance under the Metropolis sampler than under the Gibbs sampler only at *low* temperatures ($\beta \geq 0.3$).

Table 6. Metropolis sampler, no external field, free boundary, 4×4 lattice.

β	0.05	0.1	0.2	0.3	0.4	0.5
ρ	0.05	0.10	0.21	0.33	0.47	0.62
σ^2	1.08	0.60	0.51	0.61	0.94	1.31
$G_n f$	94%	80%	48%	22%	7%	4%
$G_n^4 f$	82%	70%	43%	19%	6%	3%
$G_n^d f$	81%	69%	41%	20%	6%	3%

Acknowledgment We thank the three referees for their very careful reading of the manuscript, which led to numerous improvements.

References

Amit, Y. and Grenander, U. (1991). Comparing sweep strategies for stochastic relaxation. *J. Multivariate Anal.* 37, 197–222.

Athreya, K. B., Doss, H. and Sethuraman, J. (1995). A proof of convergence of the Markov chain simulation method. *Ann. Statist.* In press.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *J. Roy. Statist. Soc. Ser. B* 36, 192–236.

Besag, J. and Green, P. J. (1993). Spatial statistics and Bayesian inference. *J. Roy. Statist. Soc. Ser. B* 55, 25–37.

Bickel, P. J. , Klaassen, C. A. J., Ritov, Y. and Wellner, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models.* Johns Hopkins University Press, Baltimore.

Binder, K. (ed.) (1992). *The Monte Carlo Method in Condensed Matter Physics.* Springer-Verlag, Berlin.

Brillinger, D. R. (1963). A note on re-use of samples. *Ann. Math. Statist.* 34, 341–343.

Chan, K. S. (1993). Asymptotic behavior of the Gibbs sampler. *J. Amer. Statist. Assoc.* 88, 320–326.

Chan, K. S. and Geyer, C. J. (1994). Contribution to the discussion of the paper by L. Tierney. *Ann. Statist.* 22, 1747–1758.

Frigessi, A., Hwang, C.-R., Sheu, S. J. and Di Stefano, P. (1993). Convergence rates of the Gibbs sampler, the Metropolis algorithm, and other single-site updating dynamics. *J. Roy. Statist. Soc. Ser. B* 55, 205–220.

Frigessi, A., Hwang, C.-R. and Younes, L. (1992). Optimal spectral structure of reversible stochastic matrices, Monte Carlo methods and the simulation of Markov random fields. *Ann. Appl. Probab.* 2, 610–628.

Gelfand, A. E., Hills, S. E., Racine-Poon, A. and Smith, A. F. M. (1990). Illustration of Bayesian inference in normal data models using Gibbs sampling. *J. Amer. Statist. Assoc.* 85, 972–985.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.* 85, 398–409.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.* 6, 721–741.

George, E. I. and Robert, C. P. (1992). Capture-recapture estimation via Gibbs sampling. *Biometrika* 79, 677–683.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments (with discussion). In: *Bayesian Statistics* 4 (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.), 169–193, Oxford University Press.

Geyer, C. J. (1992). Practical Markov chain Monte Carlo. *Statist. Science* 7, 473–483.

Graham, J. (1994). Monte Carlo Markov chain likelihood ratio test and Wald test for binary spatial lattice data. Preprint.

Green, P. J. and Han, X.-l. (1992). Metropolis methods, Gaussian proposals and antithetic variables. In: *Stochastic Models, Statistical Methods, and Algorithms in Image Analysis* (P. Barone, A. Frigessi and M. Piccioni, eds.), 142–164, Lecture Notes in Statistics 74, Springer-Verlag, Berlin.

Greenwood, P. E., McKeague, I. W. and Wefelmeyer, W. (1995). Information bounds for Gibbs samplers. In preparation.

Greenwood, P. E. and Wefelmeyer, W. (1990). Efficiency of estimators for partially specified filtered models. *Stochastic Process. Appl.* 36, 353–370.

Grenander, U. (1983). Tutorial in pattern theory. Lecture Notes. Division of Applied Mathematics, Brown University.

Grenander, U. (1993). *General Pattern Theory. A mathematical study of regular structures.* Clarendon Press, Oxford.

Guerra, F., Rosen, L. and Simon, B. (1975). The $P(\phi)_2$ Euclidean quantum field theory as classical statistical mechanics. *Ann. Math.* 101, 111–259.

Heermann, D. W. and Burkitt, A. N. (1992). Parallel algorithms for statistical physics problems. In: *The Monte Carlo Method in Condensed Matter Physics* (K. Binder, ed.), 53–74, Springer-Verlag, Berlin.

Ingrassia, S. (1994). On the rate of convergence of the Metropolis algorithm and Gibbs sampler by geometric bounds. *Ann. Appl. Probab.* 4, 347–389.

Israel, R. B. (1979). *Convexity in the Theory of Lattice Gases.* Princeton University Press.

Kindermann, R. and Snell, J. L. (1980). *Markov Random Fields and their Applications.* Amer. Math. Soc.

Levit, B. Ya. (1974). On optimality of some statistical estimates. In: *Proceedings of the Prague Symposium on Asymptotic Statistics* 2 (J. Hájek, ed.), 215–238, Charles University, Prague.

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

Pearl, J. (1987). Evidential reasoning using stochastic simulation. *Artificial Intelligence* 32, 245–257.

Penev, S. (1991). Efficient estimation of the stationary distribution for exponentially ergodic Markov chains. *J. Statist. Plann. Inference* 27, 105–123.

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

Schervish, M. J. and Carlin, B. P. (1992). On the convergence of successive substitution sampling. *J. Comput. Graph. Statist.* 1, 111–127.

Simon, B. (1974). *The $P(\phi)_2$ Euclidean (Quantum) Field Theory.* Princeton University Press.

Smith, A. F. M. and Roberts, G. O. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *J. Roy. Statist. Soc. Ser. B* 55, 3–23.

Spiegelhalter, D. J., Dawid, A. P., Lauritzen, S. L. and Cowell, R. G. (1993). Bayesian analysis in expert systems. *Statist. Science* 8, 219–283.

Swendsen, R. H. and Wang, J.-S. (1987). Nonuniversal critical dynamics in Monte Carlo simulations. *Phys. Rev. Lett.* 58, 86–88.

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *J. Amer. Statist. Assoc.* 82, 528–540.

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

Winkler, G. (1995). *Image Analysis, Random Fields and Dynamic Monte Carlo Methods.* Springer-Verlag, Berlin.