A New Proof of Peskun's Theorem Regarding the Asymptotic Variance of MCMC Estimators.

Radford M. Neal, University of Toronto

Peskun's theorem shows that the asymptotic variance of an MCMC estimator based on a reversible Markov chain will not increase if the matrix of transition probabilities for the chain is modified so as to increase the off-diagonal terms. I present a new proof of this result, which is more intuitive than Peskun's original proof, and which may provide hints for how to prove other results of this nature.

ISBA 2004, Viña Del Mar, Chile, May 2004

Asymptotic Variance of an MCMC Estimator

Let $X_1, X_2, \ldots$ be an irreducible, aperiodic Markov chain on a finite state space, $\mathcal{X}$, having $\pi(x)$ as its unique invariant distribution.

Let $f(x)$ be some function of state, whose expectation with respect to π is μ. The first n states from the Markov chain can be used to estimate μ, as follows:

$$\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} f(X_i)$$

The asymptotic variance of $\hat{\mu}$ is defined as follows:

$$V_\infty(\hat{\mu}) \;=\; \lim_{n\to\infty} n \, \mathrm{Var}(\hat{\mu}_n)$$

The asymptotic variance does not depend on the initial distribution for X_1. Also, the bias of the estimator will be of order $1/n$, regardless of initial distribution, so its asymptotic mean squared error will be equal to its asymptotic variance.

We would like to find a Markov chain for which V_∞ is as small as possible.

Peskun's Theorem (1973)

Let $X_1, X_2, \ldots$ and $X'_1, X'_2, \ldots$ be two irreducible, aperiodic Markov chains on the finite state space $\mathcal{X}$, both with $\pi(x)$ as their unique invariant distribution.

Let the transition probabilities for these chains be

$$T(i,j) = P(X_{t+1} = j \,|\, X_t = i), \qquad T'(i,j) = P(X'_{t+1} = j \,|\, X'_t = i)$$

Suppose these transition probabilities satisfy the following reversibility condition:

$$\pi(i)\, T(i,j) \;\;=\;\; \pi(j)\, T(j,i), \quad \text{for all } i,j \in \mathcal{X}$$

and similarly for T'.

Let $f(x)$ be some function of state, whose expectation with respect to π is μ. Consider the following two estimators for μ based on these two chains:

$$\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} f(X_i), \qquad \hat{\mu}'_n = \frac{1}{n} \sum_{i=1}^{n} f(X'_i)$$

If T and T' satisfy the following condition,

$$T'(i,j) \;\geq\; T(i,j), \quad \text{for all } i,j \in \mathcal{X} \text{ with } i \neq j$$

then the asymptotic variance of $\hat{\mu}'$ will be no greater than that of $\hat{\mu}$.

Why is Peskun's Theorem Interesting?

Peskun's Theorem tells us that changing a reversible Markov chain sampler to increase the probability of state changes can't be bad (asymptotically, at least). Indeed, we generally expect such a change to improve asymptotic variance.

If we can do this without increasing the computation time per transition, we'll have a better MCMC method. Conversely, going the other direction (increasing the probability of staying in the same state) can't help (asymptotically).

Some interesting consequences:

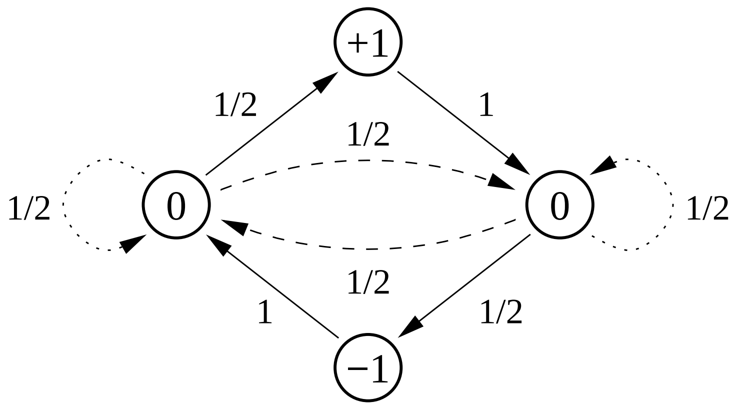
- The usual acceptance criterion for the Metropolis algorithm is optimal, since it maximizes the probability of acceptance within the class of valid criteria (Peskun, 1973; Tierney, 1998).
- Random scan Gibbs sampling can be improved by trying to avoid setting the component that is updated to the same value as it currently has (Liu, 1996).

Is Peskun's Theorem Obvious?

Since it seems inefficient to stay in the same place, Peskun's theorem might seem obvious. Two facts show that things are more subtle than this.

First, only the *asymptotic* variance is guaranteed not to increase if off-diagonal entries in the transition matrix are increased. The variance of an estimator based on finite number of iterations, started from π, may increase (Tierney, 1998).

Second, Peskun's theorem does not hold if the condition that the chains be reversible is omitted. Here's a counterexample using a non-reversible chain with four states:



$$\pi(x) \;=\; \begin{cases} 1/3 & \text{where } f(x) = 0 \\ 1/6 & \text{where } f(x) \neq 0 \end{cases}$$

$$\mu \;=\; 0$$

Values of $f(x)$ are shown in the circles. Solid arrows show values of both $T(i,j)$ and $T'(i,j)$; dotted arrows are for $T(i,j)$ only; dashed arrows are for $T'(i,j)$ only. The asymptotic variance is zero when using T, but not when using T'.

Outline of a New Proof

We can prove that the "new" chain based on T' has at least as small asymptotic variance as the "old" chain based on T as follows:

1. We reduce the problem to comparing asymptotic variances when T and T' differ only for transitions involving two states, A and B.

2. We see how simulations of the old and new chains differ only for certain "delta" transitions involving states A and B.

3. These delta transitions divide the Markov chain simulation into blocks of iterations, which start and end in either state A or state B. We can rewrite the old and new estimators, $\hat{\mu}$ and $\hat{\mu}'$, as weighted averages of block averages.

4. The only difference between the old and new chains is that in the new chain the sampling for "homogeneous" blocks (starting and ending in the same state) is *stratified* — there are the *same* number of blocks starting and ending with A as blocks starting and ending with B, whereas the split between these types is random in the old chain.

5. Finally, we see that this stratification will lower (or at least not increase) the asymptotic variance.

Looking at One Pair of States is Enough

Whenever $T'(i,j) \geq T(i,j)$ for all $i \neq j$, we can get to T' from T by a series of steps that each change transition probabilities for only a single pair of states.

For example:

$$T \;=\; \begin{bmatrix} 0.4 & 0.4 & 0.2 \\ 0.4 & 0.4 & 0.2 \\ 0.4 & 0.4 & 0.2 \end{bmatrix} \Rightarrow \begin{bmatrix} 0.3 & 0.5 & 0.2 \\ 0.5 & 0.3 & 0.2 \\ 0.4 & 0.4 & 0.2 \end{bmatrix} \Rightarrow \begin{bmatrix} 0.3 & 0.5 & 0.2 \\ 0.5 & 0.2 & 0.3 \\ 0.4 & 0.6 & 0.0 \end{bmatrix} \;=\; T'$$

So it's enough to prove Peskun's Theorem when T and T' differ for only two states, say A and B. The "old" transition probabilities, T, and "new" transition probabilities, T', are related as follows:

$$T'(i,j) \;=\; T(i,j), \quad \text{when } i \notin \{A,B\} \text{ or } j \notin \{A,B\}$$

$$\begin{aligned} T'(A,A) &= T(A,A) - \delta_A, & T'(A,B) &= T(A,B) + \delta_A \\ T'(B,A) &= T(B,A) + \delta_B, & T'(B,B) &= T(B,B) - \delta_B \end{aligned}$$

where δ_A and δ_B are positive.

Marking “Delta” Transitions

Transitions T and T' differ only if the current state is A or B, and then only with respect to how a probability mass of δ_A or δ_B is assigned to new states A or B. We can mark such “delta” transitions while simulating the Markov chain.

For each state, i, partition the interval $[0, 1)$ into intervals $[\ell(i, j), h(i, j))$ such that $h(i, j) - \ell(i, j) \; = \; T(i, j)$, and $\ell(A, A) = \ell(B, B) = 0$. We can use these intervals to simulate the old transitions T and the new transitions T' as follows:

Old transitions:

```
$U \sim$ Uniform$(0, 1)$
if $X_t = A$ and $U < \delta_A$ then
   $X_{t+1} = A$,  mark this transition
else if $X_t = B$ and $U < \delta_B$ then
   $X_{t+1} = B$,  mark this transition
else
   $X_{t+1} = j$ such that $U \in [\ell(X_t, j), h(X_t, j))$
```

New transitions:

```
$U \sim$ Uniform$(0, 1)$
if $X_t = A$ and $U < \delta_A$ then
   $X_{t+1} = B$,  mark this transition
else if $X_t = B$ and $U < \delta_B$ then
   $X_{t+1} = A$,  mark this transition
else
   $X_{t+1} = j$ such that $U \in [\ell(X_t, j), h(X_t, j))$
```

Clearly, T and T' differ only for the “delta” transitions marked above.

How Delta Transitions Define Blocks

We can use the markings of delta transitions to divide a simulation of one of these Markov chains into “blocks” of consecutive states, that both start and end with either state A or state B.

Since asymptotic variance doesn’t depend on the initial state distribution, let’s say $P(X_1 = A) = P(X_1 = B) = 1/2$, so that the chains will begin at the start of a block.

For the old chain, with transitions T, we might see blocks like this:

A B	B B	B B	B B	B A	A A	A A	A B	B A	A A	A

For the new chain, with transitions T', the blocks might look like this:

A B	A A	B B	A A	B A	B A	B B	A A	B A	B B	A

The difference is that in the old chain, the state stays the same when crossing a block boundary, whereas for the new chain, it changes from A to B or from B to A.

Notes: States A and B may also occur at places other than the start and end of a block. Blocks of length of one are possible, consisting of an A or B that both starts and ends the block.

Probabilities of the Four Types of Blocks

Blocks come in four types — AA, BB, AB, BA — based on start and end states. We show here that for both old and new chains, the probabilities of these types satisfy

$$P(AA) = P(BB) \ \text{ and } \ P(AB) = P(BA)$$

Proof: Since the new chain is reversible, $\pi(A)\,T'(A,B) \;=\; \pi(B)\,T'(B,A)$. By rewriting T', we get

$$\pi(A)\,(T(A,B) + \delta_A) \;=\; \pi(B)\,(T(B,A) + \delta_B)$$

The old chain is also reversible, with $\pi(A)\,T(A,B) \;=\; \pi(B)\,T(B,A)$, so we can conclude that

$$\pi(A)\,\delta_A \;=\; \pi(B)\,\delta_B$$

This lets us show that for a state, X_t, from the old chain (with t being large),

$$P(X_t \text{ starts block with } A) \;=\; P(X_{t-1} = A)\,P(\text{delta transition at } t{-}1) \;=\; \pi(A)\,\delta_A$$
$$P(X_t \text{ starts block with } B) \;=\; P(X_{t-1} = B)\,P(\text{delta transition at } t{-}1) \;=\; \pi(B)\,\delta_B$$

and hence $P(X_t \text{ starts block with } A) \;=\; P(X_t \text{ starts block with } B)$. In the same way, we see that $P(X_t \text{ ends block with } A) \;=\; P(X_t \text{ ends block with } B)$. It follows that

$$P(AA) + P(AB) \;=\; P(BB) + P(BA) \ \text{ and } \ P(AA) + P(BA) \;=\; P(BB) + P(AB)$$

so $P(AA) = P(BB)$ and $P(AB) = P(BA)$. We see this for the new chain similarly.

Contents of Blocks of Different Types

Although blocks of type AA and blocks of type BB are equally common, the distributions for their contents — and hence for their length and for the average value of $f(x)$ over states in the block — will generally be different.

In contrast, blocks of type AB and blocks of type BA have the *same* distribution of content — except that the BA blocks are the reversals of the AB blocks. This is a consequence of the chains being reversible.

For example: The probability of block AQB occurring at time t (with t large) is

$$\begin{aligned}
&P(X_t = A \text{ \& block starts})\, P(X_{t+1} = Q \,|\, X_t = A) P(X_{t+2} = B \text{ \& block ends} \,|\, X_{t+1} = Q) \\
&\quad = \;\; \pi(A)\, \delta_A\, T(A,Q)\, T(Q,B)\, \delta_B \;=\; \delta_A \delta_B\, \pi(A)\, T(A,Q)\, T(Q,B) \\
&\quad = \;\; \delta_A \delta_B\, T(Q,A)\, \pi(Q)\, T(Q,B) \;=\; \delta_A \delta_B\, T(Q,A)\, T(B,Q)\, \pi(B) \\
&\quad = \;\; \pi(B)\, \delta_B\, T(B,Q)\, T(Q,A)\, \delta_A
\end{aligned}$$

which is also the probability of block BQA occurring at time t.

This result is true for both the old chain (using T) and the new chain (using T').

Simulation Using Blocks

Rather than simulate the chains one state at a time, let's imagine simulating block by block. We'll need the probability that a block is "homogeneous" — ends with the same state it begins with — which is

$$\begin{aligned} P(\text{ends with } A \,|\, \text{starts with } A) \;&=\; \frac{P(AA)}{P(AA)+P(AB)} \;=\; \frac{P(BB)}{P(BB)+P(BA)} \\ &=\; P(\text{ends with } B \,|\, \text{starts with } B) \;=\; h \end{aligned}$$

We can simulate block transitions for the "old" and "new" chains as follows. We'll assume H is sampled the same for both chains, but block simulation is not coupled.

Old transitions:

```
$H \sim$ Bernoulli($h$)
if $H = 1$ then
   if previous block ended with $A$
      simulate an $AA$ block
   else
      simulate a $BB$ block
else
   if previous block ended with $A$
      simulate an $AB$ block
   else
      simulate an $AB$ block, then reverse it
```

New transitions:

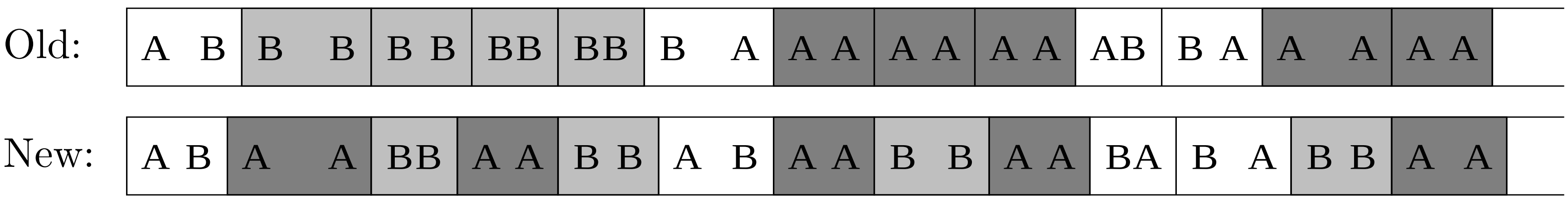
```
$H \sim$ Bernoulli($h$)
if $H = 1$ then
   if previous block ended with $A$
      simulate a $BB$ block
   else
      simulate an $AA$ block
else
   if previous block ended with $A$
      simulated an $AB$ block, then reverse it
   else
      simulate an $AB$ block
```

The Key Fact: In the New Chain Using T', Sampling for Homogeneous Blocks is Stratified

Comparing the simulations for the old and new chains, we see that they produce the *same* sequence of homogeneous/non-homogeneous blocks. However, for the new chain, the homogeneous blocks *alternate* between AA blocks and BB blocks.

This is true both when one homogeneous block follows another, and when any number of non-homogeneous blocks intervene. In the old chain, the type of homogeneous block changes only when an odd number of non-homogeneous blocks intervene.

This can be seen by example:



Old: A B | B B | B B | BB | BB | B A | A A | A A | A A | AB | B A | A A | A A

New: A B | A A | BB | A A | B B | A B | A A | B B | A A | BA | B A | B B | A A

Because AA blocks alternate with BB blocks in the new chain, the number of AA blocks will be equal to the number of BB blocks (plus or minus one). So sampling with the new chain is *stratified* in this respect. Furthermore, this is the *only* difference between the old and new chains. Intuitively, stratification should not increase asymptotic variance. We can show this formally using the following two lemmas.

Lemma 1: Asymptotic Variance for Block-by-Block Simulation is the Same as for State-by-State Simulation

First, we need a lemma showing that simulating a given number of blocks produces the same asymptotic variance as simulating a given number of transitions.

Lemma 1: *Let $X_1, X_2, \ldots$ be an irreducible, aperiodic Markov chain on a finite state space $\mathcal{X}$, with invariant distribution $\pi(x)$. Let $\mathcal{S}$ be some non-empty subset of $\mathcal{X}$, and let $f(x)$ be some function of state, whose expectation w.r.t. π is μ. Define*

$$N(k) = \min \left\{ n : \sum_{i=1}^{n} I_{\mathcal{S}}(X_i) = k \right\}$$

Consider the following two families of estimators:

$$\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} f(X_i), \qquad \tilde{\mu}_k = \frac{1}{N(k)} \sum_{i=1}^{N(k)} f(X_i)$$

The asymptotic variances of these estimators are the same:

$$\lim_{n\to\infty} n \operatorname{Var}(\hat{\mu}_n) \;=\; \lim_{n\to\infty} n \operatorname{Var}(\tilde{\mu}_{\lceil n\pi(\mathcal{S})\rceil})$$

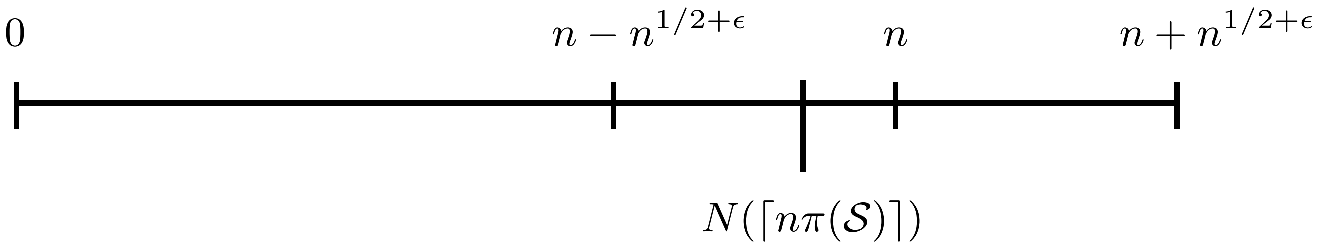
Note that asymptotically the expected value of $N(\lceil n\pi(\mathcal{S})\rceil)$ is n, so the right side above is a sensible asymptotic variance.

To apply this lemma to our problem, we can extend the state space so that it encodes whether we are at the end of a block, and then let $\mathcal{S}$ be the set of end-block states.

Proof of Lemma 1

We will see that as n increases, $n\text{Var}(\hat{\mu}_n)$ and $n\text{Var}(\tilde{\mu}_{\lceil n\pi(\mathcal{S})\rceil})$ both approach $(n + n^{1/2+\epsilon})\text{Var}(\hat{\mu}_{n+n^{1/2+\epsilon}})$, where ϵ is a positive constant to be set below.

Without loss of generality, suppose $\mu = 0$.



First, we note that $(n + n^{1/2+\epsilon})\,\hat{\mu}_{n+n^{1/2+\epsilon}} = n\hat{\mu}_n + n^{1/2+\epsilon}Z$, where Z is the average of $f(X_i)$ for i from $n+1$ to $n + n^{1/2+\epsilon}$. Dividing by $\sqrt{n + n^{1/2+\epsilon}}$, we get

$$\sqrt{n + n^{1/2+\epsilon}}\,\hat{\mu}_{n+n^{1/2+\epsilon}} = \sqrt{n/(n + n^{1/2+\epsilon})}\left[\sqrt{n}\hat{\mu}_n + n^{\epsilon}Z\right]$$

As n increases, the first factor on the right will go to one. By the CLT for Markov chains, $|Z|$ will be less than $(n^{1/2+\epsilon})^{-1/2+\epsilon} = n^{-1/4+\epsilon^2}$ with probability approaching one exponentially fast, so if ϵ is in $(0,\, (\sqrt{2}-1)/2)$, the term $n^{\epsilon}Z$ will go to zero. It follows that $n\text{Var}(\hat{\mu}_{n+n^{1/2+\epsilon}})$ will approach $n\text{Var}(\hat{\mu}_n)$. (Since $f(x)$ is bounded, an exponentially small probability of a large value for $|Z|$ cannot affect this limit.)

Proof of Lemma 1 (Continued)

The CLT also tells us that $N(\lceil n\pi(\mathcal{S})\rceil)$ will be in the interval $(n-n^{1/2+\epsilon},\ n+n^{1/2+\epsilon})$ with probability approaching one exponentially fast. If so, we can write

$$(n+n^{1/2+\epsilon})\,\hat{\mu}_{n+n^{1/2+\epsilon}} \;=\; N(\lceil n\pi(\mathcal{S})\rceil)\,\tilde{\mu}_{\lceil n\pi(\mathcal{S})\rceil} \;+\; (n+n^{1/2+\epsilon}-N(\lceil n\pi(\mathcal{S})\rceil))\,Y$$

where Y is the average of $f(X_i)$ for i from $N(\lceil n\pi(\mathcal{S})\rceil)+1$ to $n+n^{1/2+\epsilon}$. Dividing by $\sqrt{n+n^{1/2+\epsilon}}$, we get

$$\sqrt{n+n^{1/2+\epsilon}}\,\hat{\mu}_{n+n^{1/2+\epsilon}} = \frac{N(\lceil n\pi(\mathcal{S})\rceil)}{n\sqrt{1+n^{-1/2+\epsilon}}}\Big[\sqrt{n}\tilde{\mu}_{\lceil n\pi(\mathcal{S})\rceil} + (\sqrt{n}/N(\lceil n\pi(\mathcal{S})\rceil))\,KY\Big]$$

where $K = n+n^{1/2+\epsilon}-N(\lceil n\pi(\mathcal{S})\rceil)$ will be in $(0, 2n^{1/2+\epsilon})$ if $N(\lceil n\pi(\mathcal{S})\rceil)$ is in $(n-n^{1/2+\epsilon},\ n+n^{1/2+\epsilon})$. By the CLT for Markov chains, $|KY|$ will be less than $(2n^{1/2+\epsilon})^{1/2+\epsilon} = 2^{1/2+\epsilon}n^{1/4+\epsilon+\epsilon^2}$ with probability approaching one exponentially fast. Since $N(\lceil n\pi(\mathcal{S})\rceil)$ will approach n, we can see that $(n+n^{1/2+\epsilon})\mathrm{Var}(\hat{\mu}_{n+n^{1/2+\epsilon}})$ will approach $n\mathrm{Var}(\tilde{\mu}_{\lceil n\pi(\mathcal{S})\rceil})$.

Lemma 2: Partially Stratifying a Ratio Estimator Does Not Increase Asymptotic Variance

Second, we need a lemma showing that stratifying the number of AA and BB blocks won't increase the asymptotic variance.

Lemma 2: *Let $Z_1, Z_2, \ldots$ be an irreducible, aperiodic Markov chain with state space $\{0,1,2\}$, whose invariant distribution, ρ, satisfies $\rho(0) = \rho(1)$. Let Q_m for $m = 0,1,2$ be distributions for pairs $(H, L) \in \mathbb{R} \times \mathbb{R}^+$ having finite second moments. Conditional on $Z_1, Z_2, \ldots$, let (H_i, L_i) be drawn independently from Q_{Z_i}. Define*

$$Z'_i \;=\; \begin{cases} Z_i & \textit{if } Z_i = 2 \\ Z_k + \sum_{j=1}^{i-1} I_{\{0,1\}}(Z_j) \;\; \text{(modulo 2)} & \textit{if } Z_i \neq 2 \end{cases}$$

where $k = \min\{i : Z_i \neq 2\}$. Conditional on $Z_1, Z_2, \ldots$, let (H'_i, L'_i) be drawn independently from $Q_{Z'_i}$. Define two families of estimators as follows:

$$R_n \;=\; \sum_{i=1}^{n} H_i \,\Big/\, \sum_{i=1}^{n} L_i, \qquad R'_n \;=\; \sum_{i=1}^{n} H'_i \,\Big/\, \sum_{i=1}^{n} L'_i$$

Then the asymptotic variance of R' is no greater than that of R. In other words,

$$\lim_{n\to\infty} n\mathrm{Var}(R'_n) \;\le\; \lim_{n\to\infty} n\mathrm{Var}(R_n)$$

To apply this lemma to our problem, we let Q_0, Q_1, and Q_2 be the distributions of block lengths and sums of $f(X_i)$ for blocks of types AA, BB, and AB/BA.

Proof of Lemma 2

Let $N_{n,m} = \sum_{i=1}^{n} I_{\{m\}}(Z_i)$ and $N'_{n,m} = \sum_{i=1}^{n} I_{\{m\}}(Z'_i)$. Note that $E(N_{n,m}) = E(N'_{n,m})$ and $|N'_{n,1} - N'_{n,0}| \leq 1$, so the numbers of pairs from Q_0 and Q_1 are stratified in R'_n.

We can write

$$\mathrm{Var}(R_n) \ = \ \mathrm{Var}(E(R_n|N_n)) \ + \ E(\mathrm{Var}(R_n|N_n))$$

and similarly for R'_n.

Conditional on N_n, can write

$$\sum_{i=1}^{n} H_i \ = \ S_0 + S_1 + S_2, \qquad \sum_{i=1}^{n} L_i \ = \ T_0 + T_1 + T_2$$

where S_m and T_m are sums of $N_{n,m}$ values of H_i and L_i for which $Z_i = m$. By the CLT, the distributions of the pairs (S_m, T_m) will be asymptotically normal, with means, variances, and covariances that are linear functions of the $N_{n,m}$ values. The pair of sums, $(\sum H_i, \sum L_i)$, will therefore also be asymptotically normal, with mean, variances, and covariance that are linear functions of the $N_{n,m}$. We can also rewrite $\sum H'_i$ and $\sum L'_i$ in terms of S'_0, S'_1, S'_2 and T'_0, T'_1, T'_2, and proceed analogously.

The delta rule can be applied to show that if (X, Y) is asymptotically normal with mean (μ_x, μ_y), variances σ_x^2 and σ_y^2, and covariance γ_{xy}, then X/Y is asymptotically normal with mean $\mu_* = \mu_x/\mu_y$ and variance $(1/\mu_y^2)[\sigma_x^2 \ + \ \sigma_y^2\mu_*^2 \ - \ 2\gamma_{xy}\mu_*]$.

Proof of Lemma 2 (Continued)

We can now see that asymptotically $\text{Var}(R_n|N_n)$ and $\text{Var}(R'_n|N'_n)$ are the same linear functions of N_n and N'_n. It follows that $E(\text{Var}(R_n|N_n)) = E(\text{Var}(R'_n|N_n))$.

We can write $\text{Var}(E(R_n|N_n))$, and similarly $\text{Var}(E(R'_n|N'_n))$, as

$$\text{Var}(E(R_n|N_n)) \;=\; \text{Var}(E(E(R_n|N_n)|N_{n,2})) + E(\text{Var}(E(R_n|N_n)|N_{n,2}))$$

We see that $E(\text{Var}(E(R_n|N_n)|N_{n,2})) \geq E(\text{Var}(E(R'_n|N'_n)|N'_{n,2})) = 0$, since due to stratification, $N'_{n,0}$ and $N'_{n,1}$ are fixed given $N'_{n,2}$ and n, so $\text{Var}(E(R'_n|N'_n)|N'_{n,2}) = 0$.

Let $(\mu_{H,m},\, \mu_{L,m})$ be the mean of the distribution Q_m. Asymptotically, we can write

$$E(R_n|N_n) \;=\; \frac{\mu_{H,0}N_{n,0} + \mu_{H,1}N_{n,1} + \mu_{H,2}N_{n,2}}{\mu_{L,0}N_{n,0} + \mu_{L,1}N_{n,1} + \mu_{L,2}N_{n,2}}$$

and similarly for $E(R'_n|N'_n)$. By the CLT for Markov chains, $(N_{n,0}, N_{n,1}, N_{n,2})$ and $(N'_{n,0}, N'_{n,1}, N'_{n,2})$ asymptotically have (degenerate) multivariate normal distributions. Applying the delta rule, we can find that $E(E(R_n|N_n)|N_{n,2}) = E(E(R'_n|N'_n)|N'_{n,2})$, and hence their variances are asymptotically equal. It follows that asymptotically $\text{Var}(E(R_n|N_n)) \geq \text{Var}(E(R'_n|N'_n))$, and finally, that $\text{Var}(R_n)$ is asymptotically at least as large as $\text{Var}(R'_n)$.

Why is This New Proof Interesting?

- It gives new insight into why Peskun's theorem applies only to reversible chains.

  For a non-reversible chain, the AB blocks needn't be distributed in the same way as the reversals of BA blocks. Since the *old* chain is stratified with respect to AB and BA blocks, but the new chain is not, the asymptotic variance of the new chain might be be greater than that of the old chain, if AB and BA blocks differ.

- Lemmas 1 and 2 may be of wider interest. (Sufficiently so that they may have already been proved by someone else...?)

- **My main motivation...** I hope that the techniques in this proof — such as focusing on a minimal change in the transitions, and on how this divides the chain into blocks — may be useful in proving other results. In particular, I hope to be able to prove things about methods for modifying chains to be non-reversible, generalizing the ideas of Diaconis, Holmes, and Neal (2000).

Acknowledgements

This research was supported by the Natural Sciences and Engineering Research Council of Canada. Radford Neal holds a Canada Research Chair in Statistics and Machine Learning.

References

Diaconis, P., Holmes, S., and Neal, R. M. (2000) "Analysis of a non-reversible Markov chain sampler", *Annals of Applied Probability*, vol. 10, pp. 726-752.

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

Liu, J. S. (1996) "Peskun's theorem and a modified discrete-state Gibbs sampler", *Biometrika*, vol. 83, pp. 681-682.

Tierney, L. (1998) "A Note on Metropolis-Hastings kernels for general state spaces", *Annals of Applied Probability*, vol. 8, pp. 1-9.