Biometrika (2002), **89,** 4, *pp.* 745–754

Printed in Great Britain

Empirical supremum rejection sampling

By BRIAN S. CAFFO
Department of Biostatistics, Johns Hopkins University, Baltimore, Maryland 21205, U.S.A.
bcaffo@jhsph.edu

JAMES G. BOOTH
Department of Statistics, University of Florida, Gainesville, Florida 32611, U.S.A.
jbooth@stat.ufl.edu

and A. C. DAVISON
Institute of Mathematics, Swiss Federal Institute of Technology, 1015 Lausanne, Switzerland
anthony.davison@epfl.ch

Summary

Rejection sampling thins out samples from a candidate density from which it is easy to simulate, to obtain samples from a more awkward target density. A prerequisite is knowledge of the finite supremum of the ratio of the target and candidate densities. This severely restricts application of the method because it can be difficult to calculate the supremum. We use theoretical argument and numerical work to show that a practically perfect sample may be obtained by replacing the exact supremum with the maximum obtained from simulated candidates. We also provide diagnostics for failure of the method caused by a bad choice of candidate distribution. The implication is that essentially no theoretical work is required to apply rejection sampling in many practical cases.

Some key words: Accept-reject; Candidate distribution; Monte Carlo; Sample maximum; Super-efficient estimator.

1. Introduction

Rejection sampling is a way of generating a random sample from a target density f from which it is difficult to simulate, using a random sample from a more tractable candidate density g. A key requirement is that $C \equiv \sup_x f(x)/g(x)$ be finite. Let F and G denote the distribution functions corresponding to f and g, and let C_{UB} be an upper bound for C. Then the usual rejection sampling algorithm is given in Algorithm 1.

Algorithm 1 (*Standard rejection sampling*)
Step 1. *Generate* $X \sim G$ *and* $U \sim \mathrm{Un}(0, 1)$ *independently.*
Step 2. *Accept* X *if* $U \leqslant f(X)/\{C_{\mathrm{UB}} g(X)\}$.
Step 3. *If* X *is accepted, return* X; *otherwise go to Step* 1.

The value of C_{UB} is the average number of candidate variates required to obtain one target variate (Robert & Casella, 1999, Problem 2.36). Thus, the lower the upper bound, the more efficient is the method.

The choice of candidate for rejection sampling is often motivated by the ease with which

C can be calculated or tightly bounded, but this can give a value so large that rejection sampling is inefficient. By contrast more efficient candidates can yield ratios f/g that are complicated and even multimodal, making it harder to calculate C. In such cases we propose to estimate C using a sequence of lower bounds given by $\hat{C}$, the maximum ratio obtained from the simulated candidate variables. The new algorithm is given in Algorithm 2.

Algorithm 2 (*Empirical supremum rejection sampling*)
Step 1. *Initialise* $\hat{C}$.
Step 2. *Generate* $X \sim G$ *and* $U \sim \mathrm{Un}(0, 1)$ *independently.*
Step 3. *Accept* X *if* $U \leqslant f(X)/\{\hat{C}g(X)\}$.
Step 4. *Update* $\hat{C} = \max\{\hat{C}, f(X)/g(X)\}$.
Step 5. *If* X *is accepted, return* X; *otherwise go to Step* 2.

The sample maximum $\hat{C}$ can be a super-efficient estimator of C if $C < \infty$. We exploit its fast rate of convergence to show that rejection sampling using $\hat{C}$ accepts essentially the same sequence of candidates as would Algorithm 1. In fact, when f and g have discrete support the sequences of accepted values from the algorithms only differ for finitely many repetitions of the algorithm with probability one. This strong result does not quite hold in continuous cases, but we can relate the rate of convergence of the empirical supremum to the rate of convergence of the difference between averages computed using the two sequences. We show that the sequence from Algorithm 2 inherits the strong law of large numbers and central limit theorem obeyed by that from Algorithm 1. Thus, if the goal is to evaluate an expectation with respect to F, variates generated using Algorithm 2 may be treated as a random sample from F.

By the introduction of dependence among the accepted values, this work differs from two other variations on rejection sampling, namely adaptively improving the candidate as the algorithm progresses (Gilks & Wild, 1992; Wild & Gilks, 1993) and recycling the information contained in the uniform variates from Step 3 of Algorithm 1 (Casella & Robert, 1996, 1998). A third more closely related variation adjusts for the use of a possibly incorrect value of C with the independent Metropolis algorithm (Tierney, 1994).

In § 2 we formally develop the two rejection sampling algorithms, and in § 3 we state and prove the main results. In § 4 we discuss implementation and the possibility of improving on estimates of C. In § 5 we explore a simple example, while § 6 contains a brief discussion.

2. Empirical supremum rejection sampling

We evaluate the limiting behaviour of the accepted candidates from Algorithm 2 by comparing them with the accepted values from Algorithm 1 run with the same candidate and uniform variates.

Let $\{X_{ij}\}_{ij \in \mathbb{N} \times \mathbb{N}}$ be a doubly-indexed sequence of independent variates from G, mutually independent of the doubly-indexed sequence $\{U_{ij}\}_{ij \in \mathbb{N} \times \mathbb{N}}$ of independent $\mathrm{Un}(0, 1)$ variates. The subscript identifies the jth candidate and uniform variates used in generating the ith observation from the target distribution. Let

$$\tau_i = \min\left\{ j \in \mathbb{N} \,\middle|\, U_{ij} \leqslant \frac{f(X_{ij})}{Cg(X_{ij})} \right\}$$

and define $Y_i = X_{i\tau_i}$. Then the sequence $\{Y_i\}$ is generated according to Algorithm 1, with acceptance number τ_i.

We use similar notation to formalise Algorithm 2, but distinguish the acceptance number and accepted candidate with a tilde. Thus, the ith acceptance number is

$$\tilde{\tau}_i = \min\left\{ j \in \mathbb{N} \;\middle|\; U_{ij} \leqslant \frac{f(X_{ij})}{\hat{C}_i g(X_{ij})} \right\},$$

and the ith accepted candidate is $\tilde{Y}_i = X_{i\tilde{\tau}_i}$, where

$$\hat{C}_{i+1} = \max\left\{ \frac{f(X_{i1})}{g(X_{i1})}, \hat{C}_i \right\}. \tag{1}$$

In these formal descriptions of the algorithms we update (1) only once per accepted candidate, using the first candidate from the previous round, X_{i1}, in contrast to Algorithm 2, in which updating occurs with every candidate. This simplifies proofs because $\hat{C}_i$ is then the largest order statistic of exactly $i-1$ observations rather than of a random number of them. Also, $\hat{C}_i$ defined in this way is independent of X_{ij} for all j. In § 3 we show that our main results continue to hold with any scheme that implements larger lower bounds than those defined in (1). In particular, our results still hold if $\hat{C}_i$ is updated after every simulated candidate as in Algorithm 2.

The recursive definition of $\hat{C}_i$ requires an initial $\hat{C}_1$. As $C \geqslant 1$, we set $\hat{C}_1 = 1$ to prove the main theorem in § 3. In practice, where $\hat{C}_i$ is only evaluated up to a constant of proportionality one might simply set $\hat{C}_1 = f(X_{11})/g(X_{11})$, where X_{11} is the first candidate value generated.

The three main assumptions needed for the new algorithm are as follows, $\mathscr{X}_F$ and $\mathscr{X}_G$ denoting the supports of F and G:

Assumption 1. We require that $\mathscr{X}_F \subset \mathscr{X}_G$.

Assumption 2. We require that $C \equiv \sup_{x \in \mathscr{X}_F} f(x)/g(x) < \infty$.

Assumption 3. We require that $C = f(x_C)/g(x_C)$ for some $x_C \in \mathscr{X}_F$.

Assumptions 1 and 2 are required for Algorithm 1, but Assumption 3 is not. In most situations all three assumptions can be satisfied by choosing a candidate density with heavier tails than the target. In particular, they hold if f and g are bounded and g dominates f outside a compact subset of $\mathscr{X}_F$.

3. Convergence

The key quantity for comparing the sequences is $\mathrm{pr}(Y_i \neq \tilde{Y}_i)$, the probability that Algorithm 2 erroneously accepts a candidate that Algorithm 1 rejects. In the discrete case, $\sum_i \mathrm{pr}(Y_i \neq \tilde{Y}_i) < \infty$ with probability one, and hence the output from the two algorithms differs for only finitely many i. The sequence $\{\tilde{Y}_i\}$ therefore has the same limiting properties as $\{Y_i\}$.

Theorem 1. *If f is a density with respect to the counting measure then*

$$\mathrm{pr}(Y_i \neq \tilde{Y}_i \text{ infinitely often in } i) = 0.$$

Proof. Assumption 3 gives $C = f(x_C)/g(x_C)$ for some $x_C \in \mathscr{X}_F$. If $\gamma = \min\{i \in \mathbb{N} \mid X_{i1} = x_C\}$, then γ is geometric with success probability $g(x_C)$, where $g(x_C) > 0$ by the assumption that

$\mathscr{X}_F \subset \mathscr{X}_G$. As the algorithms are identical when $\hat{C}_i = C$, it follows that the event $\{Y_i \neq \tilde{Y}_i\}$ implies the event $\{\gamma \geqslant i\}$. Thus

$$\mathrm{pr}(Y_i \neq \tilde{Y}_i) \leqslant \mathrm{pr}(\gamma \geqslant i) = \{1 - g(x_C)\}^{i-1}$$

and hence $\sum_i \mathrm{pr}(Y_i \neq \tilde{Y}_i) < \infty$. □

One could argue that in practice Theorem 1 applies also to continuous cases, because C is typically evaluated only up to a given accuracy. For a fair comparison of the algorithms we should then use $\hat{C}_i$ to estimate C to within the same tolerance, and with probability one this must occur within finitely many iterations. More formally, however, in many continuous settings $\sum_i \mathrm{pr}(Y_i \neq \tilde{Y}_i)$ may not be finite because $(Y_i \neq \tilde{Y}_i)$ is precisely $O(i^{-1})$. This rate of convergence, which is needed to prove our main result stated in Theorem 2, only requires Assumptions 1, 2 and 3. In § 4, under the assumption that $\log(f/g)$ is smooth and unimodal, we motivate an even faster rate, namely $\mathrm{pr}(Y_i \neq \tilde{Y}_i) = O(i^{-2})$.

Let h be a real-valued F-measurable function, let $\mu_h = E\{h(Y_i)\}$ and suppose that $\sigma_h^2 = \mathrm{var}\{h(Y_i)\} < \infty$. Let $\bar{h}_n = n^{-1} \sum_{i=1}^n h(Y_i)$ denote the sample average obtained from the sequence $\{Y_i\}$, and let $\tilde{h}_n$ denote the sample average from the corresponding sequence $\{\tilde{Y}_i\}$. Almost sure convergence of $n^{\frac{1}{2}}(\bar{h}_n - \tilde{h}_n)$ to zero is sufficient to prove that $\{\tilde{h}_n\}$ inherits the strong law of large numbers and central limit theorem obeyed by $\{\bar{h}_n\}$.

Theorem 2. *If $E\{h(Y_i)^\delta\} < \infty$ for some $\delta > 2$ and Assumptions* 1, 2 *and* 3 *hold, then $\tilde{h}_n$ converges almost surely to μ_h and $n^{\frac{1}{2}}(\tilde{h}_n - \mu_h)$ converges in distribution to $N(0, \sigma_h^2)$ as $n \to \infty$.*

See the Appendix for the proof.

We end this section with a discussion of better estimates of C. Let $\{\hat{Y}_i\}$ be a sequence of accepted values from a variant of Algorithm 2 implementing larger lower bounds for C than $\hat{C}_i$. Then $\hat{Y}_i$ satisfies $\mathrm{pr}(Y_i \neq \hat{Y}_i) \leqslant \mathrm{pr}(Y_i \neq \tilde{Y}_i)$. The crucial quantity in Theorems 1 and 2 is the rate of convergence of $\mathrm{pr}(Y_i \neq \tilde{Y}_i)$. As this rate dominates $\mathrm{pr}(Y_i \neq \hat{Y}_i)$, the sequence $\{\hat{Y}_i\}$ inherits the same properties as $\tilde{Y}_i$. This is true in particular if the maximum is updated with every candidate rather than once for every accepted candidate.

4. Asymptotic behaviour of $\hat{C}_i$

When $\log(f/g)$ is smooth and unimodal at x_C, a more precise description of the asymptotic behaviour of $\hat{C}_i$ is possible. We can also derive a confidence interval estimate of C, resulting in an empirical supremum rejection sampling algorithm that produces exact independent identically distributed samples with a high probability. In addition, it is possible to diagnose a poor choice of candidate distribution leading to $C = \infty$.

Let $V \equiv v(X) = \log\{f(X)/g(X)\}$, where X is drawn from G. In many cases v will be smooth and unimodal, having a maximum at x_C, near which it is concave. If so, Taylor series expansion of v about x_c implies that for v close to $\log C$

$$\mathrm{pr}(V > v) \doteq \mathrm{pr}\{\log C - v > -\tfrac{1}{2}(x_C - X)^2 v''(x_C)\} \doteq a(\log C - v)^{\frac{1}{2}}, \tag{2}$$

where $a = 2g(x_C)[2/\{-v''(x_C)\}]^{\frac{1}{2}}$. Now let $V_k = v(X_{k1})$ for $k = 1, \ldots, i$, so that $\hat{C}_i$ is the maximum of $e^{V_1}, \ldots, e^{V_i}$. Then it follows from (2) that, as $i \to \infty$, $i^2 \log(C/\hat{C}_i)$ converges to a Weibull variable with shape $\frac{1}{2}$ and unknown scale parameter. The Weibull approxi-

mation suggests that

$$E\{\log(C/\hat{C}_i)\} \doteq E(C/\hat{C}_i - 1) = O(i^{-2})$$

and hence $\mathrm{pr}(Y_i \neq \tilde{Y}_i) = O(i^{-2})$ by Lemma A1.

We now extend these results to obtain an upper confidence limit for $\log C$. Equation (2) also implies that $\psi(V_k) \equiv -\{\log a + \frac{1}{2}\log(\log C - V_k)\}$ approximately has a standard exponential distribution for $k = 1, \ldots, i$. This approximation improves when considering the larger order statistics of the $\{V_k\}$. Let $V_{(k),i}$ be the kth order statistic of $V_1, \ldots, V_i$. Note that $\hat{C}_i = \exp(V_{(i),i})$ and $\hat{C}_{i-1} = \exp(V_{(i-1),i-1})$. Then $\psi(V_{(i-1),i}) - \psi(V_{(i),i})$ has an approximate standard exponential distribution. Let $e_{1-\alpha}$ denote the $1-\alpha$ quantile of the standard exponential distribution. Then

$$\alpha \doteq \mathrm{pr}\{\psi(V_{(i-1),i}) - \psi(V_{(i),i}) \geqslant e_{1-\alpha}\} = \mathrm{pr}\left\{\log C \leqslant \exp V_{(i),i} + \frac{V_{(i),i} - V_{(i-1),i}}{\exp(2e_{1-\alpha}) - 1}\right\},$$

giving an approximate level-α upper confidence bound for $\log C$, and hence for C.

This bound remains valid if $\hat{C}$ is updated with every candidate. A reviewer has pointed out that the bound must be updated in this manner, because an upper bound that is too large early in the simulations may otherwise produce an algorithm that will never accept a candidate and hence never improve the upper bound.

The previous arguments also provide diagnostics for assessing whether or not a chosen g provides a suitable candidate for f. If $C = \sup_x f(x)/g(x) = \infty$, then the upper tail behaviour of the random variable V_k will be quite different from that for finite C. Regardless of C, the large-sample joint distribution of the $\{V_k\}$ exceeding some threshold will be approximately that of independent generalised Pareto variables (Davison & Smith, 1990). This approximation assumes that the number of exceedances is small relative to the overall sample size and holds in wide generality (Pickands, 1975). The generalised Pareto distribution is

$$H(w) = \begin{cases} 1 - (1 + \kappa w/\sigma)_+^{-1/\kappa} & (\kappa \neq 0), \\ 1 - \exp(-w/\sigma) & (\kappa = 0), \end{cases} \tag{3}$$

where $\sigma > 0$. The parameter κ determines the shape of the upper tail, with $\kappa < 0$ giving a finite upper bound and therefore finite C. This suggests diagnosing an infinite C by testing the null hypothesis $\kappa = 0$ against the alternative $\kappa > 0$. The score statistic for $\kappa = 0$ under model (3) is equivalent to Greenwood's statistic, $G_s = \sum_{k=1}^{T} S_k^2$, where $S_k = U_{(k)} - U_{(k-1)}$, in which $U_{(1)}, \ldots, U_{(T-1)}$ are the order statistics of a random sample from the $\mathrm{Un}(0, 1)$ distribution and $U_{(0)} = 0$, $U_{(T)} = 1$. The distribution of G_s has been extensively tabulated (Burrows, 1979; Currie, 1981; Stephens, 1981); percentage points are also easily calculated by simulation.

In our context we suggest that the threshold be taken to be $V_{(i-T),i}$, for a moderate value of T, such as $T = 21$. The test is then applied to the spacings $S_k = S'_k / \sum_{k=1}^{T} S'_k$ of $S'_k = V_{(k),i} - V_{(k-1),i}$, for $k = i - T + 1, \ldots, i$. Values of G_s in the upper tail of its null distribution suggest that $\kappa > 0$ and thus that the candidate density has been badly chosen.

5. Example

Consider a candidate for a rejection sampler used to simulate from the conditional distribution of a random effect given observed data for a random intercept logistic/normal

Table 1. *Average and median proportion error rates for various candidate distributions and sample sizes, M, for $z = 10$, $n = 30$ and 1000 replications of M simulations from the new algorithm without and with upper confidence limit. A type A error occurs when a sampler incorrectly accepts a candidate it should have rejected and a type B error occurs when a sampler rejects a candidate it should have accepted*

		$M = 2$		$M = 5$		$M = 10$		$M = 100$	
		Error type		Error type		Error type		Error type	
		A	*B*	*A*	*B*	*A*	*B*	*A*	*B*
		t_3/Laplace candidate distribution							
ESUP	Mean	0·10	0	0·05	0	0·03	0	0·00	0
	Median	0	0	0	0	0	0	0	0
ESUP UCL	Mean	0·07	0·18	0·03	0·24	0·01	0·23	0·00	0·07
	Median	0	0	0	0·2	0	0·2	0	0·04
		$N(\alpha, \sigma^2)$ candidate distribution							
ESUP	Mean	0·77	0	0·49	0	0·33	0	0·07	0
	Median	1	0	0·4	0	0·3	0	0·06	0
ESUP UCL	Mean	0·49	0·01	0·20	0·01	0·10	0·01	0·01	0·01
	Median	0·5	0	0·2	0	0·1	0	0·01	0·01

ESUP, empirical supremum algorithm, run without the upper confidence limit.
ESUP UCL, empirical supremum algorithm, run with the upper confidence limit.

model, which specifies that, given p, $Z \sim \text{Bi}(n, p)$ with $\log\{p/(1-p)\} = Y$, where $Y \sim N(\alpha, \sigma^2)$. Such a model arises when obtaining shrinkage estimates for small area estimation (Agresti et al., 2000). Simulation of values of Y conditional on Z, α and σ allows one to replace intractable integrals with Monte Carlo estimates when finding marginal maximum likelihood estimates via the EM algorithm (Booth & Hobert, 1999). For simplicity we consider simulating from $Y|Z = z$ with $\alpha = 1$ and $\sigma = 0{\cdot}5$.

The marginal $N(\alpha, \sigma^2)$ distribution of the random effect is a poor candidate. Though it is feasible to derive a finite upper bound C, this choice of candidate distribution makes the algorithm increasingly inefficient as the sample size, n, increases; see B. Caffo's 2001 Ph.D. Thesis from the University of Florida. A better candidate distribution can be constructed by Laplace approximation (Tierney et al., 1989). Let μ and θ be the Laplace approximation to the mean and variance of $Y|Z = z$, $\alpha = 1$, $\sigma = 0{\cdot}5$; see Booth & Hobert (1999) for details. Shifting and scaling a t_3 distribution by μ and $\theta^{\frac{1}{2}}$ provides a better candidate that becomes increasingly efficient as $n \to \infty$. Furthermore, as Assumptions 1, 2 and 3 are met when using the marginal distribution of Y as a candidate, they are also met for the heavier tailed t_3 candidate distribution.

For $z = 10$ and $n = 30$, using a shifted and rescaled t_3 candidate distribution yields an acceptance rate of over 85%, as opposed to only 2% for the $N(\alpha, \sigma^2)$ candidate. However, a closed-form expression for C is unavailable for the t candidate distribution. We ran Algorithm 2 with the supremum updated with each simulated value for both candidate distributions. Knowledge of the exact value of C allowed us to determine whether a candidate was mistakenly accepted or rejected relative to Algorithm 1 for each candidate and uniform pair. Table 1 gives the average and median proportions of incorrect decisions in repeated sampling of 1000 repetitions of $M = 2$, 5, 10 and 100 candidate uniform pairs for Algorithm 2 with and without the upper confidence limit. Incorrect decisions are

stratified by whether a candidate was errantly accepted or errantly rejected, referred to as type A and B errors respectively. Although Algorithm 2 cannot make a type B error, it can when an upper confidence limit is used. For the accurate t_3/Laplace candidate, empirical supremum rejection sampling made few errant decisions even for very small sample sizes. For the less accurate $N(\alpha, \sigma^2)$ candidate, the algorithm with the upper confidence limit was preferable. For either candidate distribution, however, the new algorithm became essentially equivalent to the standard algorithm after just a few simulations.

6. Discussion

The main benefit of empirical supremum rejection sampling is that the choice of the candidate distribution need not be governed by the ability to calculate, bound or numerically solve for C, so useful candidate distributions that lead to complicated forms for f/g can be used. Furthermore, once a candidate distribution is chosen, one can diagnose whether or not $C = \infty$.

When rejection sampling is used to provide a large Monte Carlo sample, the new algorithm can be used in place of the standard algorithm with virtually no modification. This is not true, however, when rejection sampling is used to calculate only one sample point from the target distribution, for example when simulating a single value from an intractable full conditional distribution in a Gibbs sampler. It is sometimes possible to avoid Gibbs sampling entirely by sequential simulation, in which case empirical supremum rejection sampling can be used to simulate from the first distribution of the sequence, which is often intractable.

Acknowledgement

We thank the referee and associate editor for useful comments that improved the paper. This work was partially supported by the Swiss National Science Foundation and by the National Science Foundation.

Appendix

Technical details

Lemma A1. *Under Assumptions* 1 *and* 2, $\mathrm{pr}(Y_i \neq \tilde{Y}_i) \leqslant E(C/\hat{C}_i) - 1$.

Proof. That $\mathrm{pr}(Y_i \neq \tilde{Y}_i) \leqslant \mathrm{pr}(\tau_i \leqslant \tilde{\tau}_i)$ is clear as $Y_i \neq \tilde{Y}_i$ is equivalent to $X_{i\tau_i} \neq X_{i\tilde{\tau}_i}$. To prove that $\mathrm{pr}(\tau_i \leqslant \tilde{\tau}_i) \leqslant E(C/\hat{C}_i) - 1$, consider the events

$$A_{ij} = \left\{ U_{ij} \leqslant \frac{f(X_{ij})}{g(X_{ij})C} \right\}, \quad B_{ij} = \left\{ U_{ij} \leqslant \frac{f(X_{ij})}{g(X_{ij})\hat{C}_i} \right\}.$$

As $\hat{C}_i < C$ it follows that $A_{ij} \subset B_{ij}$.The probabilities of A_{ij} and B_{ij} are

$$\mathrm{pr}(A_{ij}) = 1/C,$$

$$\mathrm{pr}(B_{ij}) = E\left[\mathrm{pr}\left\{ U_{ij} \leqslant \frac{f(X_{ij})}{g(X_{ij})\hat{C}_i} \,\middle|\, X_{ij}, \hat{C}_i \right\} \right] = E\left[\min\left\{ \frac{f(X_{ij})}{g(X_{ij})\hat{C}_i}, 1 \right\} \right].$$

As $\hat{C}_i$ is independent of X_{ij} for all j, and the X_{ij} are independent and identically distributed, $\mathrm{pr}(B_{ij})$ is constant for all j.

The probability that Algorithm 2 accepts $X_{ij'}$ and Algorithm 1 accepts X_{ij} is

$$\begin{aligned}\mathrm{pr}(\tilde{\tau}_i = j', \tau_i = j) &= \prod_{l=1}^{j'-1} \{1 - \mathrm{pr}(B_{il})\} \, \mathrm{pr}(B_{ij'} \cap A_{ij'}^{\mathrm{c}}) \prod_{l=j'+1}^{j-1} \{1 - \mathrm{pr}(A_{il})\} \, \mathrm{pr}(A_{ij}) \\ &\leqslant \prod_{l=1}^{j'-1} \{1 - \mathrm{pr}(A_{il})\} \{\mathrm{pr}(B_{ij'}) - \mathrm{pr}(A_{ij'})\} \prod_{l=j'+1}^{j-1} \{1 - \mathrm{pr}(A_{il})\} \, \mathrm{pr}(A_{ij}) \\ &= \{\mathrm{pr}(B_{i1}) - C^{-1}\}(1 - C^{-1})^{j-2} C^{-1} \quad (j' < j).\end{aligned}$$

Summing over $j' < j$ gives $\mathrm{pr}(\tilde{\tau}_i < j, \tau_i = j) \leqslant (j-1)\{\mathrm{pr}(B_{i1}) - C^{-1}\}(1 - C^{-1})^{j-2} C^{-1}$, so

$$\begin{aligned}\mathrm{pr}(\tilde{\tau}_i < \tau_i) &\leqslant \sum_{j=1}^{\infty} \mathrm{pr}(\tilde{\tau}_i < j, \tau_i = j) \\ &= \{\mathrm{pr}(B_{i1}) - C^{-1}\} \sum_{j=1}^{\infty} (j-1)(1 - C^{-1})^{j-2} C^{-1} \\ &= \{\mathrm{pr}(B_{i1}) - C^{-1}\}(1 - C^{-1})^{-1}(C-1) = \{\mathrm{pr}(B_{i1}) - C^{-1}\} C.\end{aligned}$$

To complete the proof, we must show that $\mathrm{pr}(B_{i1}) \leqslant E(1/\hat{C}_i)$. However,

$$\mathrm{pr}(B_{i1}) = E\left[\min\left\{\frac{f(X_{i1})}{g(X_{i1})\hat{C}_i}, 1\right\}\right] \leqslant E\left\{\frac{f(X_{i1})}{g(X_{i1})\hat{C}_i}\right\} = E\left\{\frac{f(X_{i1})}{g(X_{i1})}\right\} E\left(\frac{1}{\hat{C}_i}\right) = E\left(\frac{1}{\hat{C}_i}\right);$$

recall that X_{i1} is independent of $\hat{C}_i$ by (1). □

LEMMA A2. *Let $\{Z_k\}$ be a sequence of independent random variables from a continuous density w, with associated distribution function W. Suppose that, for some b, $W(b) = 1$ and $W(b - \varepsilon) < 1$ for all $\varepsilon > 0$, where $w(b) > 0$. Let $Z_{(i)} = \max\{Z_k \mid k = 1, \ldots, i\}$. Then $b - E(Z_{(i)}) = O(i^{-1})$. Moreover, the same rate holds for sample minima with finite lower bounds.*

Proof. Suppose that $b = 1$ and $Z_k > 0$ for all $k \in \mathbb{N}$. Then

$$\begin{aligned}1 - E(Z_{(i)}) &= 1 - \int_0^1 \{1 - W(t)^i\} \, dt = \int_0^1 W(t)^i \, dt \\ &= \int_0^{1-\varepsilon} W(t)^i \, dt + \int_{1-\varepsilon}^1 W(t)^i \, dt \leqslant W(1-\varepsilon)^i(1-\varepsilon) + \int_{1-\varepsilon}^1 W(t)^i \, dt.\end{aligned}$$

As $W(1-\varepsilon)^i(1-\varepsilon) < O(i^{-1})$, we need only investigate $\int_{1-\varepsilon}^1 W(t)^i \, dt$.

By assumption, w is continuous from the left and $w(1) > 0$. Let k and ε be such that $w(x) > k$ for $1 - \varepsilon \leqslant x \leqslant 1$, and choose $0 < p < 1$ so that $k > p(1-\varepsilon)^{-1}$. Then

$$w(x) > k > p(1-\varepsilon)^{-1} \geqslant p(1-\varepsilon)^{p-1} \geqslant p x^{p-1},$$

for $1 - \varepsilon \leqslant x \leqslant 1$. Hence

$$\int_t^1 w(x) \, dx \geqslant \int_t^1 p x^{p-1} \, dx,$$

which implies that $W(1) - W(t) \geqslant 1 - t^p$. Thus, for $1 - \varepsilon \leqslant t \leqslant 1$, we have $W(t)^i \leqslant t^{ip}$. The result then follows from the fact that

$$\int_{1-\varepsilon}^1 t^{ip} \, dt = \frac{1}{ip+1} - \frac{(1-\varepsilon)^{ip+1}}{ip+1} = O(i^{-1}).$$

When the Z_i might be negative but have upper bound 1, let $Z'_{(i)} = \max(Z_{(i)}, 0)$. Then $Z'_{(i)}$ is the maximum of nonnegative random variables and hence the previous paragraph suggests that the rate of convergence of $E(Z'_{(i)})$ to 1 is $O(i^{-1})$. Thus $Z_{(i)}$ is tail equivalent to $Z'_{(i)}$, yielding the result for all independent identically distributed sequences bounded by 1.

Now assume that b is not necessarily 1. Then

$$Z_k = Z_k - b + 1 + (b-1) = Z_k^* + (b-1).$$

Hence the Z_k^* have upper bound 1 and hence

$$O(i^{-1}) = 1 - E(Z_{(i)}^*) = b - E(Z_{(i)}).$$

Finally, noting that the sample minimum is simply the negative of a sample maximum, we have the corresponding results for minima. □

LEMMA A3. *Under Assumptions* 1, 2 *and* 3, $\mathrm{pr}(Y_i \neq \tilde{Y}_i) = O(i^{-1})$.

Proof. By Lemma A1 we need only show that $E(C/\hat{C}_i) - 1 = O(i^{-1})$. Note that

$$C/\hat{C}_i = \min\{Cg(X_{k1})/f(X_{k1}) \mid k = 1, \ldots, i-1\}.$$

Thus $C/\hat{C}_i$ is the minimum of $i-1$ independent random variables bounded from below at 1. In the light of Lemma A2 we need only show that the density of the random variable $Z_k = Cg(X_{k1})/f(X_{k1})$ is strictly positive at 1. However the density of Z_k is strictly positive for every value $z = Cg(x)/f(x)$ for which $g(x) > 0$. In particular, Assumption 3 implies that the density of Z_k is strictly positive at $1 = Cg(x_C)/f(x_C)$. □

Proof of Theorem 2. We prove the strong law and central limit theorem simultaneously by showing that $n^{\frac{1}{2}}(\bar{h}_n - \tilde{h}_n)$ converges almost surely to zero. Lemma A3 implies that $\mathrm{pr}(Y_i \neq \tilde{Y}_i) = O(i^{-1})$, so $\sum_{i=1}^{\infty} \mathrm{pr}(Y_i \neq \tilde{Y}_i)^{\varepsilon}/i^{\frac{1}{2}} < \infty$ for any $\varepsilon > \frac{1}{2}$. If we let $\varepsilon = (\delta - 1)/\delta$, then $\varepsilon > \frac{1}{2}$. Assume for now that $E[\{h(Y_i) - h(\tilde{Y}_i)\}^{\delta}]$ is bounded in i. Then

$$\begin{aligned}\sum_{i=1}^{n} i^{-\frac{1}{2}} E\{|h(Y_i) - h(\tilde{Y}_i)|\} &= \sum_{i=1}^{n} i^{-\frac{1}{2}} E\{|h(Y_i) - h(\tilde{Y}_i)| I_{\{Y_i \neq \tilde{Y}_i\}}\} \\ &\leqslant \sum_{i=1}^{n} i^{-\frac{1}{2}} (E[\{h(Y_i) - h(\tilde{Y}_i)\}^{\delta}])^{1-\varepsilon} \mathrm{pr}(Y_i \neq \tilde{Y}_i)^{\varepsilon},\end{aligned}$$

where the inequality holds by the Hölder inequality. Thus

$$E\left\{\sum_{i=1}^{\infty} i^{-\frac{1}{2}} |h(Y_i) - h(\tilde{Y}_i)|\right\} < \infty,$$

and hence $\sum_{i=1}^{\infty} i^{-\frac{1}{2}} |h(Y_i) - h(\tilde{Y}_i)| < \infty$ almost surely. The claim then follows by Kronecker's lemma (Chow & Teicher, 1997, p. 114).

As $E\{h(Y_i)^{\delta}\}$ is constant, to show that $E[\{h(Y_i) - h(\tilde{Y}_i)\}^{\delta}]$ is bounded we need only show that $E\{h(\tilde{Y}_i)^{\delta}\}$ is bounded in i. Conditional on $\hat{C}_i$, the density of $\tilde{Y}_i$ is proportional to $\min(f, \hat{C}_i g)$ (Tierney, 1994). Therefore

$$\begin{aligned}E\{h(\tilde{Y}_i)^{\delta} \mid \hat{C}_i\} &= \int h(x)^{\delta} \min\{f(x), \hat{C}_i g(x)\}\, dx \Big/ \int \min\{f(x), \hat{C}_i g(x)\}\, dx \\ &\leqslant \int h(x)^{\delta} f(x)\, dx \Big/ \int \min\{f(x), \hat{C}_i g(x)\}\, dx \\ &= E\{h(Y_1)^{\delta}\} \Big/ \int \min\{f(x), \hat{C}_i g(x)\}\, dx \\ &\leqslant E\{h(Y_1)^{\delta}\} \Big/ \int \min\{f(x), g(x)\}\, dx,\end{aligned}$$

since $\hat{C}_i \geqslant 1$. Expectation over the distribution of $\hat{C}_i$ yields the result. □

REFERENCES

AGRESTI, A. A., BOOTH, J., HOBERT, J. & CAFFO, B. S. (2000). Random effects modeling of categorical resonse data. *Sociol. Methodol.* **30**, 27–80.

BOOTH, J. G. & HOBERT, J. P. (1999). Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *J. R. Statist. Soc.* B **61**, 265–85.

BURROWS, P. M. (1979). Selected percentage points of Greenwood's statistic. *J. R. Statist. Soc.* A **142**, 256–8.

CASELLA, G. & ROBERT, C. P. (1996). Rao–Blackwellisation of sampling schemes. *Biometrika* **83**, 81–94.

CASELLA, G. & ROBERT, C. P. (1998). Post-processing accept-reject samples: Recycling and rescaling. *J. Comp. Graph. Statist.* **7**, 139–57.

CHOW, Y. S. & TEICHER, H. (1997). *Probability Theory: Independence, Interchangeability, Martingales*, 3rd ed. New York: Springer-Verlag.

CURRIE, I. D. (1981). Further percentage points of Greenwood's statistic. *J. R. Statist. Soc.* A **144**, 360–3.

DAVISON, A. C. & SMITH, R. L. (1990). Models for exceedances over high thresholds (with Discussion). *J. R. Statist. Soc.* B **52**, 393–442.

GILKS, W. R. & WILD, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Appl. Statist.* **41**, 337–48.

PICKANDS, J. (1975). Statistical inference using extreme order statistics. *Ann. Statist.* **3**, 119–31.

ROBERT, C. P. & CASELLA, G. (1999). *Monte Carlo Statistical Methods.* New York: Springer.

STEPHENS, M. A. (1981). Further percentage points for Greenwood's statistic. *J. R. Statist. Soc.* A **144**, 364–6.

TIERNEY, L. (1994). Markov chains for exploring posterior distributions (with Discussion). *Ann. Statist.* **22**, 1701–28.

TIERNEY, L., KASS, R. E. & KADANE, J. B. (1989). Fully exponential Laplace approximations to expectations and variances of nonpositive functions. *J. Am. Statist. Assoc.* **84**, 710–6.

WILD, P. & GILKS, W. R. (1993). Algorithm AS 287: Adaptive rejection sampling from log-concave density functions. *Appl. Statist.* **42**, 701–8.

[*Received April* 2001. *Revised January* 2002]