A stochastic approximation method and its application to confidence intervals

By PAUL H. GARTHWAITE and M. C. JONES

Department of Statistics, The Open University, Milton Keynes, MK7 6AA, U.K.

p.h.garthwaite@open.ac.uk and m.c.jones@open.ac.uk

Summary

The oldest stochastic approximation method is the Robbins-Monro process. This estimates an unknown scalar parameter by stepping from one trial value for the parameter to another, adopting the last trial value as the estimate. More recent research suggests there are benefits from taking larger steps than with the Robbins-Monro process and then obtaining an estimate by averaging the later trial values. Work on the averaged estimator has made only general assumptions and here we consider a more explicit case that is of practical importance. Stronger asymptotic results are developed and simulations show they hold well for moderately long searches. The results motivate the development of a new method of searching for the endpoints of a confidence interval. This method performs decidedly better than a previously proposed method in terms of both the position of endpoints and the coverage of confidence intervals. The efficiency of the new method is typically well in excess of 90%.

Some key words: Confidence interval; Monte Carlo method; Robbins-Monro; Stochastic approximation.

1. Introduction

Stochastic approximation was introduced by Robbins and Monro (1951) for the following problem. Let $A(\theta)$ be a fixed but unknown monotonically increasing function and α a given known constant such that the equation $A(\theta) = \alpha$ has an unknown unique root $\theta = \theta^*$. It is assumed that, for each value of θ, there corresponds a random variable $I = I(\theta)$ whose expected value is $A(\theta)$. The task is to estimate θ^* by making successive observations of I at values $\theta_1, \theta_2, \ldots$, where the θ-values are chosen sequentially in accordance with some defined procedure and may depend on preceding observations.

In the Robbins-Monro process, θ_1 can be chosen arbitrarily while the sequence $\theta_i : i = 2, \ldots, n$ is generated by the recursion

$$\theta_{i+1} = \theta_i + a_i(\alpha - I_i) \tag{1}$$

where I_i is an observation of $I(\theta_i)$ and a_i is a sequence of positive numbers. Each θ_i is an estimate of θ^* and the procedure can be viewed as a search process that steps from one estimate of θ^* to another. In early work, the only estimator of θ^* that was considered was the last observation, θ_n. This estimator is consistent and asymptotically normally distributed provided the a_i and $A(\theta)$ satisfy mild conditions. In the standard Robbins-Monro process, $a_i = c/i$ for all i and then c controls the magnitude of steps. The optimal value of c is $1/g$ where

$$g = [dA(\theta)/\, d\theta]_{\theta=\theta^*}. \tag{2}$$

When $c = 1/g$, the asymptotic variance of the estimator θ_n equals

$$[\mathrm{var} I(\theta)]_{\theta=\theta^*}/(g^2 n). \tag{3}$$

We shall refer to the quantity in (3) as the *RM optimal variance*. If $I(\theta)$ has a normal distribution, then the RM optimal variance is a lower bound to the variance of any estimator of θ^* (Ruppert, 1991). Unfortunately, the optimal choice of c depends on the unknown function $A(.)$ and the asymptotic variance can increase substantially if c is poorly chosen. Polyak and Juditsky (1992) showed that the averaged sequence $\sum_{i=1}^{n} \theta_i/n$ is a more robust estimator of θ^*. Provided the coefficients a_i go to zero slower than $O(1/n)$, asymptotically this estimator is unbiased and achieves the RM optimal variance.

Work on the averaged estimator has made only general assumptions (Polyak and Juditsky, 1992; Kushner and Yang, 1993; Wang et al., 1997; Pelletier, 2000; Kushner and Yin, 2003, chapter 11). In this paper we consider a more explicit case that is of practical importance and derive stronger results. Specifically, we suppose that $I(\theta)$ is a Bernoulli variate with parameter $A(\theta)$ and that $\alpha \in (0,\, 1)$. A value θ^* is sought for which $\Pr(I(\theta^*) = 1) = \alpha$. The search follows the standard Robbins-Monro process with $a_i = c/i$ and we consider estimating θ^* from the final $100p\%$ of $\theta_1, \ldots, \theta_n$ $(0 < p \leq 1)$, either using an unweighted average,

$$\widehat{\theta}^*_u = \sum_{i=\lfloor (1-p)n \rfloor + 1}^{n} \theta_i/(pn), \tag{4}$$

or the weighted average,

$$\widehat{\theta}^*_w = \sum_{i=\lfloor (1-p)n \rfloor + 1}^{n} i\theta_i \Big/ \sum_{i=\lfloor (1-p)n \rfloor + 1}^{n} i. \tag{5}$$

The weights in the latter average are set proportional to i because, asymptotically, $\mathrm{var}(\theta_i) \propto 1/i$ in the Robbins-Monro process.

The asymptotic mean square errors, and hence efficiencies relative to the standard Robbins-Monro process, of the estimators are derived. Asymptotically, the estimators are unbiased and the variances depend upon the probability of interest (α), the number of steps in the search (n), the proportion of steps used in forming averages (p), and the constant that controls the magnitude of steps (c). For the unweighted average, the asymptotic variance decreases with p when c is large, equalling the RM optimal variance as $c \to \infty$ for $p = 1$. Simulation results are reported in which the asymptotic results hold well for $n = 500,000$ unless $p \to 1$ and c is very much larger than $1/g$.

Motivation for this work was to improve a method of forming confidence intervals that is based on the Robbins-Monro process (Garthwaite and Buckland, 1992). The method is well suited to forming confidence intervals from randomisation tests (Garthwaite, 1996) and a variety of one-parameter problems. It has been found to work well in practice (Carpenter, 1999; Matsui and Ohashi, 1999; Lee and Young, 2003). For example, Lee and Young (2003, p405) write, "The search . . . is efficiently implemented by the Robbins-Monro stochastic search algorithm, as detailed by Garthwaite and Buckland (1992), and recommended also by Carpenter (1999)". The efficiency of the method is typically about 80% relative to the RM optimal variance. It seems difficult to improve on this for searches of a moderate size, say 1000 to 5000 steps, but based on the stochastic process examined here, a new procedure for longer searches is proposed. Simulations are conducted for cases previously examined by Garthwaite and Buckland (1992) and results show that the new procedure performs markedly better in terms of the mean square error of both the position of endpoints and the coverage of confidence intervals. Efficiencies well in excess of 90% are then quite typical.

2. Asymptotic results for the stochastic approximation

2.1. *Theory*

In this subsection, we consider the Bernoulli case described in the introduction. We let $n \to \infty$ and present asymptotic formulae for the efficiencies of $\widehat{\theta}^*_u$ and $\widehat{\theta}^*_w$ (generically $\widehat{\theta}^*$) as estimators of θ^*, where

$$\text{efficiency}(\widehat{\theta}^*) = \frac{\text{RM optimal variance}}{\text{MSE}(\widehat{\theta}^*)} \times 100\% \tag{6}$$

and MSE stands for mean square error. We define the *steplength multiplier*, w, by setting the steplength constant $c = w/g$. The asymptotic variance for the final-point estimator associated with the standard Robbins-Monro process in this situation is M/n where $M \equiv M_0 w^2/(2w-1)$, $w > 1/2$ and $M_0 = \alpha(1-\alpha)/g^2$. The RM optimal variance is M_0/n and $w = 1$ corresponds to the optimal steplength constant.

Set $\lfloor (1-p)n \rfloor \simeq (1-p)n \equiv m$, taking m to be integer for appropriate large n; also set $N = n - m = pn$. Then, write

$$\widehat{\theta}^* = \frac{1}{N} \sum_{i=m+1}^{n} (C + Di)\theta_i$$

with $\sum_{i=m+1}^{n} (C+Di) = N$. When $C = 1, D = 0$, $\widehat{\theta}^* = \widehat{\theta}_u$ and when $C = 0, D = N/\sum_{i=m+1}^{n} i \simeq 2/\{(2-p)n\}$, $\widehat{\theta}^* = \widehat{\theta}_w$. Write $V_\ell = \ell^{-1}(\alpha - I_\ell)$ so that, for $k > m$, $\theta_k = \theta_m + c\sum_{j=m}^{k-1} V_j$. Then,

$$\widehat{\theta}^* - \theta^* \;=\; \theta_m - \theta^* + \frac{c}{N} \sum_{i=m+1}^{n} (C + Di) \sum_{j=m}^{i-1} V_j$$

so that

$$E\{(\widehat{\theta}^* - \theta^*)^2\} \;=\; E\{(\theta_m - \theta^*)^2\} + \frac{2c}{N} E(I) + \frac{c^2}{N^2} E(II)$$

where

$$I \;=\; (\theta_m - \theta^*) \frac{1}{2} \sum_{j=m}^{n} (n-j) \left\{2C + (n+j)D\right\} V_j,$$

$II \simeq T_0 + T_1 + T_2$,

$$\begin{aligned} T_0 &= \sum_{i=m+1}^{n} \sum_{k=m+1}^{n} \sum_{j=m}^{\min(i,k)-1} (C+Di)(C+Dk)V_j^2 \\ &\simeq \frac{1}{4} \sum_{j=m}^{n} (n-j)^2 \left\{2C + (n+j)D\right\}^2 V_j^2, \end{aligned}$$

$$\begin{aligned} T_1 &= 2 \sum_{i=m+1}^{n} \sum_{j=m}^{i-1} \sum_{k=m+1}^{n} \sum_{\ell=m}^{k-1} (C+Di)(C+Dk) V_j V_\ell \\ &\simeq \frac{1}{2} \sum_{\ell=m+1}^{n-2} \sum_{j=m}^{\ell-1} \Big[2C\left(2Cn + Dn^2\right)(n-\ell) + \left\{ D\left(2Cn + Dn^2\right) - 2C^2 \right\} (n^2 - \ell^2) \\ &\qquad - 2CD(n^3 - \ell^3) - \frac{1}{2} D^2 (n^4 - \ell^4) \Big] V_j V_\ell \end{aligned}$$

and

$$\begin{aligned} T_2 &= 2 \sum_{k=m+2}^{n} \sum_{i=m+1}^{k-1} \sum_{\ell=m+1}^{k-1} \sum_{j=m}^{\min(\ell,i)-1} (C+Di)(C+Dk) V_j V_\ell \\ &= T_1 + S \end{aligned}$$

where

$$S \simeq \frac{1}{2}\sum_{\ell=m+1}^{n-1}\sum_{j=m}^{\ell-1}\{2C(n-\ell)+D(n^2-\ell^2)\}\{2C(\ell-j)+D(\ell^2-j^2)\}V_jV_\ell.$$

Note that

$$E\{(\theta_m-\theta^*)^2\} \simeq M/\{(1-p)n\}. \tag{7}$$

Expectations of other relevant quantities depend on the results that

$$E((\theta_m-\theta^*)V_j) \simeq -Mgj^{-(w+1)}m^{w-1}, \tag{8}$$

$$E(V_j^2) \simeq -j^{-2}\alpha(1-\alpha) \tag{9}$$

and, for $\ell > j$,

$$E(V_jV_\ell) \simeq -\ell^{-(w+1)}(2w-1)^{-1}j^{w-2}w(1-w)\alpha(1-\alpha). \tag{10}$$

Expression (9) is rather immediate; expressions (8) and (10) are derived in the Appendix. Indeed, further manipulations concerning the desired expectations are cumbersome and so they are also confined to the Appendix.

However, the formidable calculations and formulae described in the Appendix reduce neatly to relatively straightforward formulae for the desired asymptotic efficiencies given (as percentages) as explicit functions of p and w as follows:

$$\text{efficiency}(\widehat{\theta}_u^*) = \begin{cases} \frac{50(w-1)(2w-1)p^2}{w\{(1-p)^w-1+wp\}} & \text{if } w \neq 1 \\ \frac{50p^2}{\{p+(1-p)\log(1-p)\}} & \text{if } w = 1, \end{cases} \tag{11}$$

$$\text{efficiency}(\widehat{\theta}_w^*) = \begin{cases} \frac{75(w-2)(w+1)(2w-1)p^2(2-p)^2}{2w^2\{3(1-p)^{w+1}-(w+1)(1-p)^3+w-2\}} & \text{if } w \neq 2 \\ \frac{675p^2(2-p)^2}{8[(1-p)^3\{3\log(1-p)-1\}+1]} & \text{if } w = 2. \end{cases} \tag{12}$$

The standard Robbins-Monro result can be verified in both cases: $\lim_{p\to 0}$ efficiency $(\widehat{\theta}^*) = 100(2w-1)/w^2\%$.

The results in equations (11) and (12) hold for the Bernoulli case considered here and are more explicit than others have derived for more general situations. Also, the conditions satisfied by the coefficients a_i are slightly weaker than usual.

We set a_i equal to c/i, so the a_i go to zero at rate $O(1/n)$, while previous convergence results for the averaged estimator require the a_i to go to zero slower than $O(1/n)$ (Kuschner and Yin, 2003, p. 373).

Contour plots of asymptotic efficiencies (11) and (12) are given in Figures 1 and 2, respectively. Concentrate first on the asymptotic efficiency of $\widehat{\theta}^*_u$. This is independent of p and takes the value 75% when $w = 2$. While some combinations of p and w result in poor efficiencies, it is extremely useful to find that high efficiencies can also be achieved away from Robbins-Monro optimality. Most importantly for practical purposes, there is a region of 90%+ efficiencies for values of p near 1 and large w. In fact, $\lim_{p\to 1} \text{efficiency}(\widehat{\theta}^*_u) = 100(2w-1)/(2w)\%$, $\lim_{w\to\infty} \text{efficiency}(\widehat{\theta}^*_u) = 100p\%$ and so $\lim_{p\to 1, w\to\infty} \text{efficiency}(\widehat{\theta}^*_u) = 100\%$. This is important because it suggests that by averaging a large proportion of values of θ_i, any of a wide variety of large values of w (and hence of c) can be employed to good effect: averaging compensates for choosing a large steplength constant, the latter being relatively easy to specify in practice.

* * * Figures 1 and 2 about here * * *

The corresponding contour plot for the weighted average is given in Figure 2, where $\lim_{p\to 1} \text{efficiency}(\widehat{\theta}^*_w) = 75(w+1)(2w-1)/(2w^2)\%$, $\lim_{w\to\infty} \text{efficiency}(\widehat{\theta}^*_w) = 75p(2-p)^2/\{(p^2-3p+3)\}\%$ and $\lim_{p\to 1, w\to\infty} \text{efficiency}(\widehat{\theta}^*_w) = 75\%$. Perhaps surprisingly, this does not show such excellent efficiencies to be possible for this estimator when $w >> 1$.

It turns out that virtually all these asymptotic results match with simulation results very well: see Section 2.2 to follow. Important implications of the theoretical work that are fully supported by the simulations are:

1. The unweighted average should be used as the estimator, rather than the weighted average.

2. Regardless of which average is used, it is best to form the average from as large a fraction of the data as other constraints permit.

3. For the unweighted average, the benefits in efficiency from averaging (rather than taking the last θ_i as the estimate) are only substantial when the steplength multiplier is above about five, and increase as the steplength multiplier increases, but may tail off when it becomes very large.

2.2. *Simulation tests*

Our aim is to develop a method of searching for the endpoints of a confidence interval. Exploratory simulations were run both to examine the accuracy of our asymptotic approximations and to guide development of the method. For the simulations $I(\theta)$ was chosen as

$$I(\theta) = \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases} \tag{13}$$

where X follows a distribution with mean θ. Four distributions for X were used: the normal distribution, the t distribution on 7 degrees of freedom (df), the logistic distribution and the Cauchy distribution. This choice gives one distribution with thin tails, two rather similar distributions with moderately thin tails and one with thick tails. In each case, standard forms of the distributions were used with variances that did not vary with θ. In line with our aim of forming confidence intervals, two values of α were used, $\alpha = 0.025$ and $\alpha = 0.005$. As the distributions are symmetric, it is sufficient to only examine searches for the lower confidence limit.

The early steps in a search are disproportionately large; for example they halve in magnitude between $i = 1$ and $i = 2$. This could make it difficult to return to the neighbourhood of θ^* if the initial steps send it far from that value. Also, poor starting points for searches may distort results. To avoid these difficulties, each simulation was started with I set equal to 5000 and the starting value, θ_{5000}, was obtained by generating 5000 observations from the distribution of X with the mean of that distribution set equal to 0. Then θ_{5000} was set equal to the $126th$ smallest observation for $\alpha = 0.025$ or the $26th$ smallest observation for $\alpha = 0.005$. These observations correspond to the 2.5% and 0.5% quantiles of the sampling distribution of the 5000 simulated observations, so they should be good starting points. The focus of interest is the effect on the estimator of θ^* from increasing c and from varying p, the proportion of values used to form the estimate of θ^*. The steplength multiplier, w, was given the value $1, 2, 5, 10$ or 20 and c was set equal to w/g.

Each simulation consisted of a search of 495,000 steps, from $i = 5000$ to $i = 499,999$. At step $i+1$ a value of X was generated from the chosen distribution with its mean set equal to θ_i. The value of X determined $I(\theta)$ through equation (13), which in turn determined θ_{i+1} through equation (1). At the end of a search, estimates were constructed using just the last θ_i ($\theta_{500,000}$), as for the standard Robbins-Monro process, or the last $100p\%$ of the θ_i. Values $p = 0.1(0.1)0.9, 0.95$ and 0.99 were examined. Both weighted and unweighted averages were calculated.

Simulations were replicated 5000 times for each distribution and each value of c and α. The MSE of the estimates was calculated for each set of simulations. These were compared with the RM optimal variance to give a simulated estimate of the efficiency (6). A selection of results are presented for $\alpha = 0.025$ in Table 1 and for $\alpha = 0.005$ in Table 2. Results for the t distribution and logistic distribution are not included; they were always in line with the results for the normal distribution.

* * * Tables 1 and 2 about here * * *

The results conform quite closely to asymptotic theory, especially for the weighted average (Table 2). Discrepancies from theory are greater for $\alpha = 0.005$

than $\alpha = 0.025$ and were greater for the Cauchy distribution than the normal, t or logistic distributions. The discrepancies also tended to increase as the sampling fraction approached 1 and as the steplength multiplier (w) increased above 10. In the main though, differences between the observed efficiency and the efficiency given by asymptotic theory were less than 2%.

One purpose of the simulations was to examine the accuracy of the asymptotic results. The other was to guide the development of a method for forming confidence intervals. In addition to fully backing up the asymptotic theory and points 1 to 3 of Section 2.1, the simulations further imply that starting searches at $i = 5000$ gives high efficiencies if searches begin from good starting points.

3. Confidence intervals

Garthwaite and Buckland (1992) give a method of forming confidence intervals that exploits the relationship between hypothesis tests and confidence intervals. Their method will be referred to as the GB method. It is based on the standard Robbins-Monro process and conducts a separate search for each endpoint of the interval. The method performs well and seems hard to improve upon when searches must be based on a modest number of steps (1000–5000). Here we present a procedure that improves upon the GB method when searches can be longer, as is typically feasible with modern computers.

We suppose that sample data $\mathbf{y}$ give a point estimate $\widehat{\theta}(\mathbf{y})$ of θ. In describing methods, we assume that the range of possible values for θ is $(-\infty, \infty)$. When the range of θ is bounded in one or both directions, a simple solution is to take a monotonic transformation of θ, such as $\xi = \ln(\theta)$ if $\theta > 0$. The search process is used to obtain the limits of a confidence interval for ξ and then the inverse-transformation is applied to these limits to obtain a confidence interval for θ.

Finding a decent starting value for a search is a tricky problem. The solution we adopt is to use the GB method for a moderate number of steps and use its final estimate as our starting point. We first describe the GB method before presenting the new method and giving simulation results that compare the two.

3.1. *Garthwaite-Buckland method*

Suppose the search is for the lower limit of an equal-tailed $100(1-2\alpha)\%$ confidence interval for θ. Then θ^* is the lower limit and θ_i is its estimate at step i. Slightly different situations are considered by Garthwaite and Buckland (1992) and Garthwaite (1996) but both fit into our framework.

- Garthwaite and Buckland (1992) suppose that the mechanism that produced the sample data could be simulated if the value of θ were known. A data set of form similar to $\mathbf{y}$ is simulated with θ set equal to θ_i. Let $\widehat{\theta}_i$ denote the estimate of θ given by the simulated data and put $x_i = \widehat{\theta}_i - \widehat{\theta}(\mathbf{y})$.

- Garthwaite (1996) addresses the task of forming confidence intervals that correspond to randomisation tests. At step i, a randomisation test of the hypothesis H_0: $\theta = \theta_i$ against H_1: $\theta > \theta_i$ is considered but only a single permutation of the data is generated, rather than the large number of permutations that would be generated for a complete randomisation test. Let $T(\theta_i)$ denote the value of the test statistic for this permutation and let $T^*(\theta_i)$ denote its value for the actual sample data. Put $x_i = T^*(\theta_i) - T(\theta_i)$ if $T^*(\theta)$ is an increasing function of θ, and $x_i = T(\theta_i) - T^*(\theta_i)$ if it is a decreasing function of θ.

In either case, I_i is defined in accordance with equation (13): $I_i = 1$ if $x_i \geq 0$; $I_i = 0$ if $x_i < 0$. Then θ_{i+1} is determined from equation (1). We have that $\Pr(X > 0 \,|\, \theta = \theta^*) = E[I(\theta^*)] = \alpha$, so the θ_i converge to the required confidence limit.

Implementation of the search process requires methods for (i) choosing a value for the steplength constant, and (ii) finding starting points for searches. The steplength constant, c, is estimated as $k\{\widehat{\theta}(\mathbf{y}) - \theta_i\}$ at step i, where

$$k = 2/\{z_\alpha (2\pi)^{-1/2} \exp(-z_\alpha^2/2)\} \tag{14}$$

and z_α is the upper $100\alpha\%$ point of the standard normal distribution. Underestimating c is a more serious fault than overestimating it and this choice gives c a value that is about *twice* its optimum when $\widehat{\theta}_i$ is approximately normally distributed.

The starting point for a search may be found by a variety of methods, perhaps using an analytic approximation or a bootstrap method with a modest number of resamples. The search is started with i set equal to m, where m is a moderate size so that early steps do not change dramatically in size. The GB method puts m equal to the smaller of $0.3(2-\alpha)/\alpha$ and 50. The progress of a search is monitored to check if its starting point was reasonable. If the difference $\widehat{\theta}(\mathbf{y}) - \theta_i$ either doubles or halves from its original value, $\widehat{\theta}(\mathbf{y}) - \theta_m$, then the search is restarted with the last θ_i taken as the starting value and with i reset to m.

As a further safeguard, the search is conducted twice, with half the total available steps taken in each half. The last θ_i from the first search is taken as the starting value for the second search. The average of the last θ_i in each of the two searches is taken as the estimate of θ^*.

3.2. *A new procedure*

The new procedure is a simple but effective extension of the GB method. For each confidence limit it conducts a separate search using the Robbins-Monro updating equation (cf. equation (1)),

$$\theta_{i+1} = \theta_i + a_i(\alpha - I_i) \qquad \text{for } i = m, \ldots, n \tag{15}$$

where $I_i = I(\theta_i)$ is given by equation (13). The procedure consists of three phases and a_i is chosen in different ways in each phase. In the first phase ia_i is comparatively small, during the second phase it is steadily increased, and during the third phase it is kept at the comparatively high level it reached at the end of the second phase. The final estimate of θ^* in one phase becomes the starting estimate of θ^* in the next phase. Further detail is as follows.

Phase 1 The GB method is followed exactly. Let $m^*/2$ denote the value of i after the last step of each half-search. This phase should be quite short and we suggest choosing m^* to be the smaller of 5000 and $n/20$, where n is the total number of steps. This generally provides a good starting point for the next phase of the search and conforms to the implications of the simulations.

Phase 2 $(i = m^*, \ldots, \nu m^*)$. During the second phase c is chosen in exactly the same way as in phase 1, but a_i is set equal to c/m^*, rather than c/i. In this phase i increases from m^* to νm^*. Thus for the last step of this phase, a_i equals $\nu c/i$, so the steplength multiplier has smoothly increased by a factor ν. With the GB method (phase 1) the steplength multiplier equals about 2 if X has a normal distribution. In a range of situations, a reasonable choice for ν is $\nu = 15$, so that the steplength multiplier is increased to about 30 for phase 3. The choice of ν is discussed further when simulation results are reported in Section 3.3.

Phase 3 $(i = \nu m^* + 1, \ldots, n)$. During this phase a_i is set equal to $\nu c/i$. Thus a_i is again proportional to i^{-1}, as in the standard RM process. This phase should be long; we found efficiency is particularly good for $n \geq 200\,000$, although much smaller values of n still give confidence intervals with good levels of accuracy.

After phase 3 has been completed, the unweighted average of the final $(n - n^*)$

values of the θ_i is taken as the estimate of the confidence limit:

$$\widehat{\theta}^*_u = \sum_{i=n^*+1}^{n} \theta_i/(n-n^*). \tag{16}$$

A weighted average could, again, be used instead,

$$\widehat{\theta}^*_w = \sum_{i=n^*+1}^{n} i\theta_i/ \sum_{i=n^*+1}^{n} i. \tag{17}$$

Turning to the choice of n^*, there are two relevant results from the work of Section 2. First, the benefit of estimating θ^* from the unweighted average, rather than the final θ_i, is only apparent when the steplength multiplier exceeds about five. Second, the proportion of values used to form the average, $p = (n - n^*)/n$, should be large. With our new procedure, a_i typically exceeds about four times its optimal value for the standard RM process when $i \geq 2m^*$ (the Cauchy distribution is an exception). That is, the steplength multiplier typically exceeds about four when $i \geq 2m^*$, which seems adequate, so we suggest setting n^* equal to $2m^*$.

We should mention that there are not exactly m^* steps in phase 1: searches start at $i = m$ rather than $i = 1$ and they restart if it becomes clear that a poor starting value was chosen. Hence, there are not exactly n steps in a search. It is also the case that some resource is used in choosing starting values, which should be taken into account when evaluating the efficiency of a search procedure, in principle. However, these factors are negligible in comparison with the length of our searches. For example, restarts increased the average number of steps by less than twenty for most distributions in simulations reported by Garthwaite and Buckland (1992). We will refer to n as the *nominal number of steps* in a search and treat it as the actual number of steps when determining efficiencies.

3.3. *Performance of the new procedure*

To examine the performance of the new procedure, simulations were conducted in which the limits of 95% and 99% equal-tailed confidence intervals were determined using $n = 500{,}000$ nominal steps for each limit. The following distributions were examined: inverse exponential (mean $= \theta$); standard exponential (mean $= 1/\theta$); gamma(r, θ) for $r = 2$, 5 and 50 (mean $= r\theta$); Cauchy; logistic; Student's t on 7 degrees of freedom; normal. For the last four of these distributions it was assumed that all parameters were known apart from the location parameter θ. For each distribution it was also supposed that an estimate of θ was based on a single observation. The maximum likelihood estimator was used for all distributions. These are the same distributions and conditions that were used by Garthwaite and Buckland (1992) to test the GB method. Here, however, in the cases of the gamma and exponential distributions we follow the suggestion of Garthwaite (1996) and first determine confidence limits for $\ln(\theta)$, from which we obtain confidence limits for θ by taking exponentials.

Three different methods were used to determine confidence limits. One was the GB method. The other two were based on the new procedure, one estimating θ^* from the unweighted average, $\widehat{\theta}^*_u$, while the other used the weighted average $\widehat{\theta}^*_w$. The settings used for the new procedure were $m^* = 5000$, $\nu = 15$ and $n^* = 10\,000$. Thus there were 5000 steps in phase 1, 70 000 in phase 2, 425 000 in phase 3 and θ_i values from 98% of the steps were used to form averages. For symmetric distributions, simulations were only conducted for the lower limit. For each distribution and confidence limit that was estimated, simulations were replicated 5000 times. Both the position of the confidence limit and the coverage of the tail area below the confidence limit were recorded for each simulation. The mean square error of the position was compared with the RM optimal variance to give an efficiency, using equation (6). Asymptotically, when the variance of the confidence limit estimator equals the RM optimal variance, $\alpha(1-\alpha)/n$ is the variance of the estimator of the tail area, approximately, where α is the intended tail area and n is the number of steps (Garthwaite and Buckland, 1992). Corresponding to equation (6) we define the efficiency of a tail area estimator as

$$\text{efficiency} \;=\frac{\alpha(1-\alpha)/n}{\text{MSE of tail area}}\times 100\%. \tag{18}$$

The efficiencies with which confidence limits and tail areas were estimated are presented in Table 3 for 2.5% and 97.5% confidence limits and in Table 4 for 0.5% and 99.5% confidence limits.

* * * Tables 3 and 4 about here * * *

The tables also show the steplength multiplier used by the GB method for each distribution and confidence limit. The GB method estimates the steplength constant as $k(\widehat{\theta}(\mathbf{y})-\theta^*)$, where k is defined in equation (14). The steplength multiplier is the ratio formed when this quantity is divided by $1/g$, the optimal steplength constant for the standard RM process. The value of g varies with the confidence limit and distribution and it is defined in equation (2). The procedures that use unweighted and weighted averages increase the steplength multiplier by a factor of ν during phase 2.

From the tables it is clear that the procedure with unweighted averages performed considerably better than both the other methods in the simulations. Its average efficiency was 97.8% for 2.5% and 97.5% confidence limits and, based on tail area, it was 91.8% for 0.5% and 99.5% confidence limits. Corresponding efficiencies were 77.7% and 76.6% for the procedure with weighted averages and 76.7% and 75.9% for the GB method, which are all much poorer. For 2.5% and 97.5% confidence limits, the efficiency of the unweighted average exceeded 90% for every distribution when judged by either the position of the confidence limit or the tail area (Table 3). In contrast, the efficiencies of the GB method and the procedure with the weighted average never exceeded 90% for the position or tail area of any 2.5% or 97.5% confidence limit. The dominance of the procedure with unweighted averages was less complete for the 0.5% and 99.5% confidence limits.

Specifically, it was poorer than the weighted average for the Cauchy distribution and it was poorer than the GB method in four other cases (Table 4).

Values of the steplength multiplier used by the GB method are also given in Tables 3 and 4. Values of less than one indicate that the steplength constant is underestimated, which is a much worse fault than overestimation. The tables show that underestimation almost never occurred. The exception is the Cauchy distribution, which has very heavy tails and these result in the steplength constant being underestimated substantially. For the Cauchy distribution the GB method had an efficiency of less than 30% for the 2.5% limit and less than 10% for the 0.5% limit. The performance of the GB method affects the efficiencies of the other two procedures as these procedures start phase 2 from estimates given by the GB procedure. Poor starting values have a greater affect on the unweighted average, as it does not down-weight early values. Hence it may not perform as well as the weighted average for distributions with very heavy tails (c.f. the 0.5% confidence limit for the Cauchy distribution).

Further simulations were conducted to examine the performance of the unweighted method for different numbers of steps (20 000, 100 000, 200 000 and 500 0000) and different values of the steplength multiplier (10, 20, 30 and 50). For searches of 20 000 steps, 1000 steps were taken in phase 1 and estimates from the first 2 000 steps were omitted when estimating θ^*. For the other searches, 5000 steps were taken in phase 1 and estimates from the first 10 000 steps were omitted when estimating θ^*. Hence p, the proportion of the θ_i used to form the average, took values 0.9, 0.9, 0.95 and 0.98 for the different numbers of steps. Both 95% and 99% confidence limits were determined for the nine distinct distributions considered in Tables 3 and 4. The efficiency in estimating each limit was determined on the basis of tail area using 5000 replicates. Averaging the efficiences for the upper and lower limits gives the efficiency of an estimated confidence interval in terms of its coverage. These latter efficiencies are quite similar for all distributions except for the Cauchy distribution, which can differ substantially. The average efficiency of the other eight distributions is given in Table 5 for each nominal number of steps and value of the steplength multiplier. Values for the Cauchy distribution are given separately in the table. Simulations were also conducted for the GB method for these same numbers of nominal steps. The resulting efficiencies are also recorded in Table 5.

* * * TABLE 5 ABOUT HERE * * *

The table shows that the new procedure performs well in searches that are reasonably long, say over 100 000 steps when, with the Cauchy distribution excluded, the average efficiency was nearly 90% or better. Asymptotic theory suggests that the steplength multiplier should be as large as possible, but the results also indicate that values above 50 can be undesirable for searches of 500 000 steps or less. A good choice for the steplength multiplier seems to be 20 to 30, although the optimal value seems to depend upon the number of steps in the search: the more steps, the larger the optimal value. It also depends on the confidence limit being sought: the more extreme the confidence level, the smaller the optimal value. At the same time, the average performance of the procedure is reasonably insensitive to the precise value chosen for the steplength multiplier.

Results for the Cauchy distribution suggest that efficiency for heavy-tailed distributions is improved substantially by increasing the number of steps in the search and using a large steplength multiplier. However, although the efficiency for the Cauchy distribution was as low as 40.4% for 100 000 steps, this implies the root mean square error in the coverage of the estimated 99% confidence intervals was only 0.0495% (i.e. the coverage of the intervals was 0.99 ± 0.000495). In many circumstances this would be an acceptable level of accuracy.

The efficiencies for the new procedure for searches of 20 000 steps were poorer than for longer searches. However, with a steplength multiplier of 10 or 20, the new procedure still performed better than the GB method, and by a comfortable margin for 95% intervals. For longer searches and for all values of the steplength multiplier that were examined, the new procedure had substantially higher average efficiencies than the GB method. For the Cauchy distribution, differences in efficiency between the two methods sometimes exceeded 60%.

4. Discussion

As the proposed method of searching for confidence limits is stochastic, it is prudent to monitor searches. A simple graphical check is to plot the estimate of the limit at the ith iteration, θ_i, against i. The estimates should oscillate close to a single value. Also, they should have started oscillating about that value before the n^* iteration, as the average of the θ_i for $i \geq n^*$ is taken as the estimate of the limit. Another simple check is to partition the iterations used to form the average into, say, ten blocks and, for each block, determine the proportion of iterations for which the estimate of the limit increased and the proportion for which it decreased. The former proportion should approximately equal α in searches for the lower limit while the latter should approximately equal α for the upper limit. These diagnostic checks have been used with the GB method and are illustrated in Garthwaite and Buckland (1992) and Garthwaite (1996).

In practice, searches will rarely fail to oscillate around the confidence limit that is being sought; the danger is that a search will become stuck at a value away from that limit. However, this rarely happens with the GB method and the asymptotic theory developed here has greatly increased the size of steps, reducing the risk of it happening still further. The new method is underpinned by theory and simulations have shown it works well; we believe it will prove useful in data analysis and also in simulation work, where it may be used as a tool to form confidence intervals from novel estimators of interest.

Appendix

Outline derivations of (11) and (12)

The following approximation will be used repeatedly in the manipulations that follow:

$$\lim_{n\to\infty}\sum_{i=m}^{n} i^r \simeq \begin{cases} (r+1)^{-1}n^{r+1}\{1-(1-p)^{r+1}\} & \text{if } r \neq -1 \\ -\log(1-p) & \text{if } r = -1; \end{cases}$$

of course, it underlies the approximate form for D corresponding to $\widehat{\theta}^*_w$ noted in Section 2.1. Actually, for brevity, we will not explicitly mention the logarithmic exceptions in this Appendix, noting that they arise where necessary as limiting cases of the main formulae. Also, note that $A(\theta_i) \simeq \alpha + (\theta_i - \theta^*)g$ where g is given by (2).

`Evaluating` $E(I)$.

First, for $j > m$,

$$E((\theta_m - \theta^*)V_j|\theta_m, ..., \theta_j) \simeq -j^{-1}(\theta_m - \theta^*)(\theta_j - \theta^*)g.$$

Second,

$$(\theta_m - \theta^*)(\theta_j - \theta^*) = (\theta_m - \theta^*)^2 + c(\theta_m - \theta^*)\sum_{i=m}^{j-1} V_i.$$

Write $P_j = E\{(\theta_m - \theta^*)(\theta_j - \theta^*)\}$. Then,

$$P_j \simeq P_m - w\sum_{i=m}^{j-1} i^{-1}P_i$$

so that, as is readily seen,

$$\begin{aligned} P_j &\simeq P_m \prod_{i=0}^{j-m-1}\left(1-(m+i)^{-1}w\right) \simeq P_m \exp\left(-w\sum_{i=m}^{j-1} i^{-1}\right) \\ &\simeq P_m \exp\left(-w\log\left(m^{-1}j\right)\right) \simeq Mj^{-w}m^{w-1} \end{aligned}$$

and hence (8). Thus,

$$E(I) \simeq -\frac{1}{2}Mm^{w-1}g\sum_{j=m}^{n}(n-j)\{2C+(n+j)D\}j^{-(w+1)}$$

$$
\begin{aligned}
&\simeq -\frac{1}{2}Mm^{w-1}g\sum_{j=m}^{n}\left\{n(2C+nD)j^{-(w+1)}-2Cj^{-w}-Dj^{-(w-1)}\right\}\\
&\simeq -\frac{Mg}{2(1-p)}\Big[(2C+nD)w^{-1}\left\{1-(1-p)^{w}\right\}\\
&\qquad -2C(1-p)(w-1)^{-1}\left\{1-(1-p)^{w-1}\right\}\\
&\qquad -nD(1-p)^{2}(w-2)^{-1}\left\{1-(1-p)^{w-2}\right\}\Big].
\end{aligned}
$$

In particular, in the unweighted case

$$
E(I_u) \simeq -Mg\left[\frac{(1-p)^{w}-1+wp}{(1-p)w(w-1)}\right] \tag{19}
$$

and in the weighted case

$$
E(I_w) \simeq -Mg\left[\frac{2\{(1-p)^{w}-1+wp\}-wp^{2}}{(1-p)(2-p)w(w-2)}\right]. \tag{20}
$$

`Evaluating` $E(T_0)$.

Utilising (9),

$$
\begin{aligned}
E(T_0) &\simeq \frac{1}{4}\alpha(1-\alpha)\sum_{j=m}^{n}(n-j)^{2}\left\{2C+(n+j)D\right\}^{2}j^{-2}\\
&\simeq \frac{1}{4}\alpha(1-\alpha)\sum_{j=m}^{n}\Big\{n^{2}(4C^{2}+4CDn+n^{2}D^{2})j^{-2}-4n(2C^{2}+CDn)j^{-1}\\
&\qquad +2(2C^{2}-2CDn-n^{2}D^{2})+4CDj+D^{2}j^{2}\Big\}\\
&\simeq \frac{n}{4}\alpha(1-\alpha)\sum_{j=m}^{n}\Big[(4C^{2}+4CDn+n^{2}D^{2})\{p/(1-p)\}\\
&\quad +4n(2C^{2}+CDn)\log(1-p)+2(2C^{2}-2CDn-n^{2}D^{2})p\\
&\quad +2CDn(2p-1)+n^{2}D^{2}\{p-p^{2}+(p^{3}/3)\}\Big].
\end{aligned}
$$

In the unweighted case, this reduces to

$$
E(T_{0,u}) \simeq \alpha(1-\alpha)n\left\{\frac{p(2-p)}{(1-p)}+2\log(1-p)\right\} \tag{21}
$$

while in the weighted case

$$
E(T_{0,w}) \simeq \alpha(1-\alpha)n\frac{p^{3}(4-p)}{3(1-p)(2-p)^{2}}. \tag{22}
$$

`Evaluating` $E(T_1)$.

First,

$$\begin{aligned} E(V_j V_{j+1}) &= (j+1)^{-1} E\{E(V_j(\alpha - I_{j+1})|\theta_1, ..., \theta_{j+1})\} \\ &\simeq (j+1)^{-1} g E\{E(V_j(\theta_{j+1} - \theta^*)|\theta_1, ..., \theta_{j+1})\} \\ &\simeq -(j+1)^{-1} g E(V_j(\theta_j - \theta^* + cV_j)). \end{aligned}$$

But $E(V_j^2) \simeq \alpha(1-\alpha) j^{-2}$ as above and

$$E(V_j(\theta_j - \theta^*)) \simeq -j^{-1} g E(\theta_j - \theta^*)^2 \simeq -gMj^{-2}.$$

Consequently,

$$E(V_j V_{j+1}) \simeq (2w-1)^{-1} j^{-3} w(1-w)\alpha(1-\alpha).$$

Similarly, for $\ell \geq j+2$,

$$\begin{aligned} E(V_j V_\ell) &= \ell^{-1} E\{E(V_j(\alpha - I_\ell)|\theta_1, ..., \theta_\ell)\} \\ &\simeq -g\ell^{-1} E\{E(V_j(\theta_\ell - \theta^*)|\theta_1, ..., \theta_\ell)\} \\ &\simeq -g\ell^{-1} E\{E(V_j(\theta_{\ell-1} - \theta^* + cV_{\ell-1})|\theta_1, ..., \theta_{\ell-1})\}) \\ &\simeq \ell^{-1}(\ell-1) E(V_j V_{\ell-1}) - \ell^{-1} w E(V_j V_{\ell-1}) \\ &= \{1 - \ell^{-1}(w+1)\} E(V_j V_{\ell-1}). \\ &\simeq \prod_{i=0}^{\ell-j} \left\{1 - (\ell-i)^{-1}(w+1)\right\} E(V_j V_{j+1}) \end{aligned}$$

and hence (10). We therefore have

$$\begin{aligned} E(T_1) \simeq & -\frac{1}{2} w^{-1} g^2 M \sum_{\ell=m+1}^{n-2} \Big[2C\left(2Cn + Dn^2\right)(n-\ell) + \Big\{ D\left(2Cn + Dn^2\right) - 2C^2 \Big\} (n^2 - \ell^2) \\ & \qquad - 2CD(n^3 - \ell^3) - \frac{1}{2} D^2 (n^4 - \ell^4) \Big] (\ell^{-2} - m^{w-1} \ell^{-(w+1)}) \\ = & -\frac{1}{2} w^{-1} g^2 M \sum_{\ell=m+1}^{n-2} \Big(\frac{1}{2} D^2 \ell^2 - 2CD\ell - \Big\{ D\left(2Cn + Dn^2\right) - 2C^2 \Big\} \end{aligned}$$

$$
\begin{aligned}
&\qquad - 2C\left(2Cn + Dn^2\right)\ell^{-1} + 2n^2(C + \frac{1}{2}Dn)^2\ell^{-2} \\
&- m^{w-1}\left[\frac{1}{2}D^2\ell^{-(w-3)} - 2CD\ell^{-(w-2)} - \left\{D\left(2Cn + Dn^2\right) - 2C^2\right\}\ell^{-(w-1)}\right. \\
&\qquad \left.\left. - 2C\left(2Cn + Dn^2\right)\ell^{-w} + 2n^2(C + \frac{1}{2}Dn)^2\ell^{-(w+1)}\right]\right) \\
= & -\frac{1}{2}w^{-1}g^2M\left(\frac{1}{2}D^2n^3\{p - p^2 + (p^3/3)\} - CDn^2(2p-1)\right. \\
&\qquad - \left\{D\left(2Cn + Dn^2\right) - 2C^2\right\}np \\
&\qquad + 2C\left(2Cn + Dn^2\right)\log(1-p) + 2n(C + \frac{1}{2}Dn)^2p/(1-p) \\
&- \left[\frac{1}{2}(w-4)^{-1}D^2n^3(1-p)^3\{1 - (1-p)^{w-4}\}\right. \\
&\qquad - 2(w-3)^{-1}CDn^2(1-p)^2\{1 - (1-p)^{w-3}\} \\
&\qquad - (w-2)^{-1}\left\{D\left(2Cn + Dn^2\right) - 2C^2\right\}n(1-p)\{1 - (1-p)^{w-2}\} \\
&\qquad - (w-1)^{-1}2C\left(2Cn + Dn^2\right)\{1 - (1-p)^{w-1}\} \\
&\qquad \left.\left. + 2nw^{-1}(C + \frac{1}{2}Dn)^2(1-p)^{-1}\{1 - (1-p)^w\}\right]\right)
\end{aligned}
$$

In the unweighted case, this reduces to

$$
\begin{aligned}
E(T_{1,u}) \simeq & -w^{-1}g^2Mn\left(p + 2\log(1-p) + \frac{p}{(1-p)} - \frac{1-p}{w-2}\left(1 - (1-p)^{w-2}\right)\right. \\
& \left. + \frac{2}{w-1}\left(1 - (1-p)^{w-1}\right) - \frac{1}{w(1-p)}\left(1 - (1-p)^w\right)\right). \qquad (23)
\end{aligned}
$$

And in the weighted case:

$$
\begin{aligned}
E(T_{1,w}) \simeq & -\frac{g^2Mn}{w(2-p)^2}\left(\frac{1}{3}p^3 - p^2 - p + \frac{p}{(1-p)} - \frac{(1-p)^3}{w-4}\left(1 - (1-p)^{w-4}\right)\right. \\
& \left. + \frac{2(1-p)}{w-2}\left(1 - (1-p)^{w-2}\right) - \frac{1}{w(1-p)}\left(1 - (1-p)^w\right)\right). \qquad (24)
\end{aligned}
$$

`Evaluating` $E(T_2)$.

We need to manipulate

$$
\begin{aligned}
E(S) \simeq & \frac{1}{2}w^{-1}(1-w)g^2M \times \\
& \sum_{\ell=m+1}^{n-1}\sum_{j=m}^{\ell-1}\{2C(n-\ell) + D(n^2 - \ell^2)\}\{2C(\ell - j) + D(\ell^2 - j^2)\}\ell^{-(w+1)}j^{w-2}.
\end{aligned}
$$

We start with the sum over j:

$$\begin{aligned}
R(\ell) &= \frac{1}{2}w^{-1}(1-w)g^2M\sum_{j=m}^{\ell-1}\{2C(\ell-j)+D(\ell^2-j^2)\}j^{w-2} \\
&\simeq \frac{1}{2}w^{-1}(1-w)g^2M\left[\left\{2C\ell+D\ell^2\right\}\frac{1}{(w-1)}\{\ell^{w-1}-m^{w-1}\}\right. \\
&- \left. 2C\frac{1}{w}\{\ell^w-m^w\}-D\frac{1}{(w+1)}\{\ell^{w+1}-m^{w+1}\}\right].
\end{aligned}$$

Multiply $R(\ell)$ by $\{2C(n-\ell)+D(n^2-\ell^2)\}\ell^{-w-1}$ and sum over ℓ from $m+1$ to $n-1$. Break this sum into three parts, $S=S_1+S_2+S_3$, and deal with each in turn. First,

$$\begin{aligned}
E(S_1) &= -\frac{1}{2}w^{-1}g^2M\sum_{\ell=m+1}^{n-1}\left(\ell^{-2}-m^{w-1}\ell^{-w-1}\right)\times \\
&\qquad \{2C(n-\ell)+D(n^2-\ell^2)\}(2C\ell+D\ell^2) \\
&= -\frac{1}{2}w^{-1}g^2M\sum_{\ell=m+1}^{n-1}\left(-D^2\ell^2-4CD\ell+\left\{D\left(2Cn+Dn^2\right)-4C^2\right\}\right. \\
&\qquad +2C\left(2Cn+Dn^2\right)\ell^{-1}-m^{w-1}\left[-D^2\ell^{-(w-3)}-4CD\ell^{-(w-2)}\right. \\
&\qquad \left.\left.+\left\{D\left(2Cn+Dn^2\right)-4C^2\right\}\ell^{-(w-1)}+2C\left(2Cn+Dn^2\right)\ell^{-w}\right]\right) \\
&= -\frac{1}{2}w^{-1}g^2M\left(-D^2n^3\{p-p^2+(p^3/3)\}-2CDn^2(2p-1)\right. \\
&\qquad +\left\{D\left(2Cn+Dn^2\right)-4C^2\right\}np-2C\left(2Cn+Dn^2\right)\log(1-p) \\
&\qquad -\left[-(w-4)^{-1}D^2n^3(1-p)^3\{1-(1-p)^{w-4}\}\right. \\
&\qquad -4(w-3)^{-1}CDn^2(1-p)^2\{1-(1-p)^{w-3}\} \\
&\qquad +(w-2)^{-1}\left\{D\left(2Cn+Dn^2\right)-4C^2\right\}n(1-p)\{1-(1-p)^{w-2}\} \\
&\qquad \left.\left.+(w-1)^{-1}2C\left(2Cn+Dn^2\right)\{1-(1-p)^{w-1}\}\right]\right).
\end{aligned}$$

In the unweighted case, this reduces to

$$\begin{aligned}
E(S_{1,u}) &\simeq -2w^{-1}g^2Mn\left(-p-\log(1-p)\right. \\
&+ \left.\frac{1-p}{w-2}\left(1-(1-p)^{w-2}\right)-\frac{1}{w-1}\left(1-(1-p)^{w-1}\right)\right) \qquad (25)
\end{aligned}$$

and in the weighted case to

$$E(S_{1,w}) \simeq \frac{-2w^{-1}g^2Mn}{(2-p)^2}\left(-\frac{1}{3}p^3+p^2\right.$$

$$+ \quad \frac{(1-p)^3}{w-4}\left(1-(1-p)^{w-4}\right) - \frac{(1-p)}{w-2}\left(1-(1-p)^{w-2}\right)\bigg). \quad (26)$$

Second,

$$\begin{aligned}
E(S_2) &= -w^{-2}(1-w)g^2MC\sum_{\ell=m+1}^{n-1}\left(\ell^{-1}-m^w\ell^{-w-1}\right)\{2C(n-\ell)+D(n^2-\ell^2)\} \\
&= -w^{-2}(1-w)g^2MC\sum_{\ell=m+1}^{n-1}\Big(-D\ell-2C+\left(2Cn+Dn^2\right)\ell^{-1} \\
&\qquad - m^w\left[-D\ell^{-(w-1)}-2C\ell^{-w}+\left(2Cn+Dn^2\right)\ell^{-(w+1)}\right]\Big) \\
&= -w^{-2}(1-w)g^2MC\Big(-\frac{1}{2}Dn^2(2p-1)-2Cnp-\left(2Cn+Dn^2\right)\log(1-p) \\
&\qquad -\Big[-(w-2)^{-1}Dn^2(1-p)^2\{1-(1-p)^{w-2}\} \\
&\qquad -2(w-1)^{-1}Cn(1-p)\{1-(1-p)^{w-1}\} \\
&\qquad +w^{-1}\left(2Cn+Dn^2\right)\{1-(1-p)^w\}\Big]\Big).
\end{aligned}$$

This gives

$$\begin{aligned}
E(S_{2,u}) &\simeq 2w^{-2}(1-w)g^2Mn\,(p+\log(1-p) \\
&- \frac{1-p}{w-1}(1-(1-p)^{w-1})+\frac{1}{w}(1-(1-p)^w)\Big)
\end{aligned} \quad (27)$$

and

$$E(S_{2,w}) \simeq 0. \quad (28)$$

And finally,

$$\begin{aligned}
E(S_3) &= -\frac{(1-w)g^2MD}{2w(w+1)}\sum_{\ell=m+1}^{n-1}\left(1-m^{w+1}\ell^{-w-1}\right)\{2C(n-\ell)+D(n^2-\ell^2)\} \\
&= -\frac{(1-w)g^2MD}{2w(w+1)}\sum_{\ell=m+1}^{n-1}\Big(-D\ell^2-2C\ell+\left(2Cn+Dn^2\right) \\
&\qquad - m^{w+1}\left[-D\ell^{-(w-1)}-2C\ell^{-w}+\left(2Cn+Dn^2\right)\ell^{-(w+1)}\right]\Big) \\
&= -\frac{(1-w)g^2MD}{2w(w+1)}\Big(-Dn^3\{p-p^2+(p^3/3)\}-Cn^2(2p-1)+\left(2Cn+Dn^2\right)np \\
&\qquad -\Big[-(w-2)^{-1}Dn^3(1-p)^3\{1-(1-p)^{w-2}\} \\
&\qquad -2(w-1)^{-1}Cn^2(1-p)^2\{1-(1-p)^{w-1}\} \\
&\qquad +w^{-1}\left(2Cn+Dn^2\right)n(1-p)\{1-(1-p)^w\}\Big]\Big)
\end{aligned}$$

Of course,

$$E(S_{3,u}) \simeq 0. \tag{29}$$

And in the weighted case:

$$\begin{aligned} E(S_{3,w}) &\simeq \frac{-2(1-w)g^2Mn}{(2-p)^2w(w+1)}\Big(-\frac{1}{3}p^3+p^2 \\ &+ \frac{(1-p)^3}{w-2}\left(1-(1-p)^{w-2}\right)-\frac{(1-p)}{w}\left(1-(1-p)^w\right)\Big). \end{aligned} \tag{30}$$

Adding together (7), $2c/pn$ times (19), and $(c/pn)^2$ times {(21) plus twice (23) plus (25) plus (27)} yields

$$\frac{2M\{(1-p)^w-1+wp\}}{w(w-1)p^2}$$

and hence (11); adding together (7), $2c/pn$ times (20), and $(c/pn)^2$ times {(22) plus twice (24) plus (26) plus (30)} yields

$$\frac{8M\{3(1-p)^{w+1}-(w+1)(1-p)^3+w-2\}}{3(w-2)(w+1)p^2(2-p)^2}$$

and hence (12).

References

Carpenter, J. (1999). Test inversion bootstrap confidence intervals. *J. R. Statist. Soc.* **B 61**, 159–72.

Garthwaite, P. H. (1996). Confidence intervals from randomization tests. *Biometrics* **52**, 1387–93.

Garthwaite, P. H. and Buckland, S. T. (1992). Generating Monte Carlo confidence intervals by the Robbins-Monro process. *Appl. Statist.* **41**, 159–71.

Lee, S. M. S and Young, G. A. (2003). Prepivoting by weighted bootstrap iteration. *Biometrika* **90**, 393–410.

Kushner, H. J. and Yang, J. (1993). Stochastic approximation with averaging of the iterates: optimal asymptotic rate of convergence for general processes. *Siam J. Control Optim.* **31**, 1045–62.

Kushner, H. J. and Yin, G. G. (2003). *Stochastic Approximation and Recursive Algorithms and Applications*. Second edition. New York: Springer.

Matsui, S. and Ohashi, Y. (1999). Analysis of recurrent events: application to a clinical trial of colony stimulating factor with the endpoint of febrile neutropenia. *Statist. Med.* **18**, 2409–20.

Pelletier, M. (2000). Asymptotic almost sure efficiency of averaged stochastic algorithms. *Siam J. Control Optim.* **39**, 49–72.

Polyak, B. T. and Juditsky, A. B. (1992). Acceleration of stochastic approximations by averaging. *Siam J. Control Optim.* **30**, 838–55.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *Ann. Math. Statist.* **22**, 400–7.

Ruppert, D. (1991). Stochastic approximation. In *Handbook of Sequential Analysis,* Eds. B. K. Ghosh and P. K. Sen, pp. 503–29. New York: Marcel Dekker.

Wang, I. J., Chong, E. R. P. and Kulkarni, S. R. (1997). Weighted averaging and stochastic approximation. *Math. Control Signals and Systems* **10**, 41–60.

Fig. 1. Asymptotic efficiency of $\widehat{\theta}_u^*$ relative to RM optimal variance, plotted as a function of p and w.

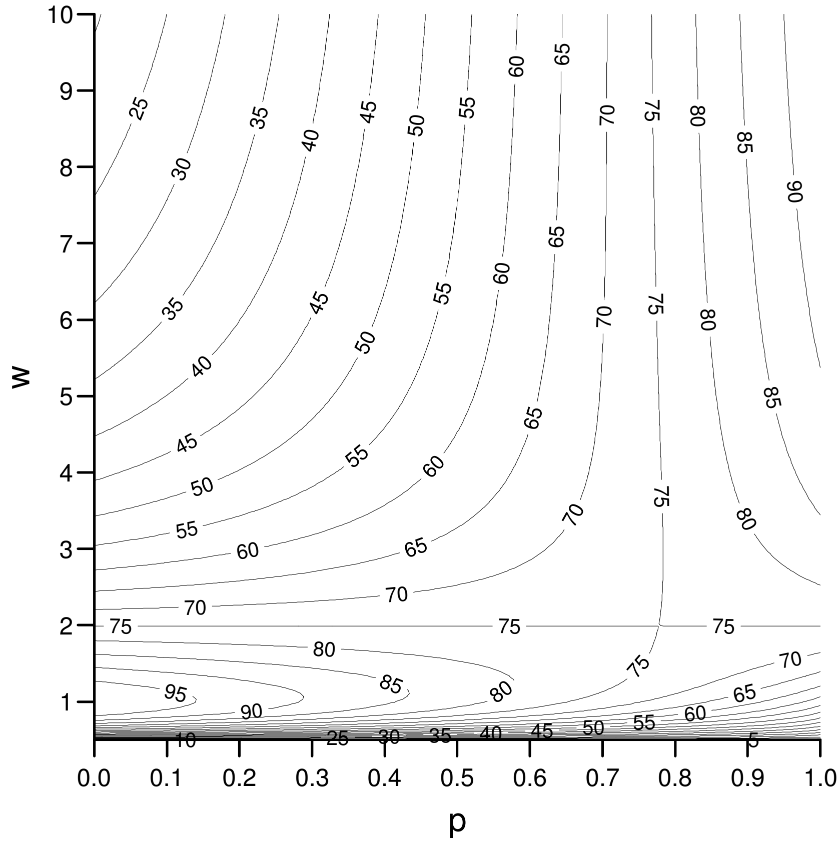

Fig. 2. Asymptotic efficiency of $\widehat{\theta}^*_w$ relative to RM optimal variance, plotted as a function of p and w.

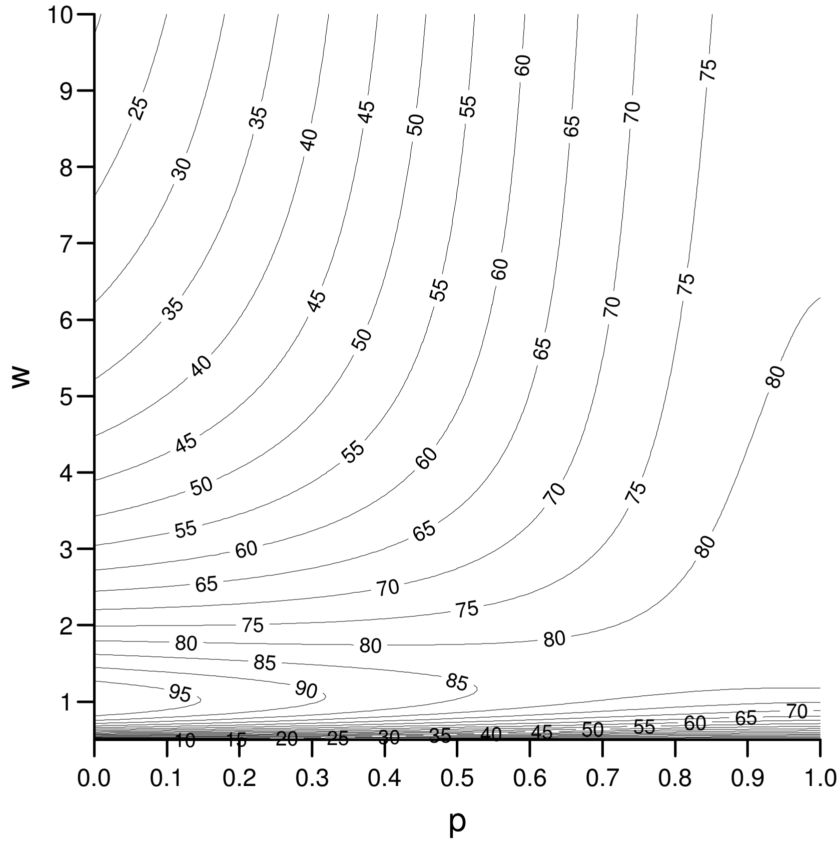

Table 1. *Efficiency of the unweighted average as an estimator and comparison with theoretical efficiency*

Steplength multiplier (w)	Sampling fraction (p)	$\alpha = 0.025$ Cauchy distrib.	$\alpha = 0.025$ Normal distrib.	$\alpha = 0.005$ Cauchy distrib.	$\alpha = 0.005$ Normal distrib.	Theoretical efficiency %
1	$\approx 0^*$	99.9	100.7	99.4	102.0	100
1	0.5	80.2	82.0	78.1	82.1	81.5
1	0.8	65.8	67.0	61.9	67.9	66.9
1	0.95	54.9	56.2	49.0	57.6	56.4
1	0.99	50.2	51.8	43.1	52.8	51.9
2	$\approx 0^*$	77.0	77.7	75.7	75.6	75.0
2	0.5	77.2	75.5	76.4	75.7	75.0
2	0.8	77.7	76.6	76.7	75.3	75.0
2	0.95	77.4	76.2	74.6	73.9	75.0
2	0.99	77.0	76.1	72.8	73.6	75.0
5	$\approx 0^*$	38.0	36.7	38.1	35.8	36.0
5	0.5	59.1	59.2	59.4	59.5	58.8
5	0.8	76.1	76.9	78.2	77.0	76.8
5	0.95	86.9	86.5	85.8	87.3	86.6
5	0.99	89.5	89.9	85.0	89.6	89.3
10	$\approx 0^*$	19.2	19.7	18.9	19.3	19.0
10	0.5	54.3	54.0	53.9	52.4	53.4
10	0.8	79.4	79.2	76.4	76.9	78.2
10	0.95	90.0	91.1	81.1	87.4	90.8
10	0.99	91.3	94.6	74.9	86.6	94.2
20	$\approx 0^*$	9.9	9.7	9.9	9.6	9.8
20	0.5	51.5	51.9	49.0	51.1	51.5
20	0.8	76.2	78.3	70.8	76.5	79.0
20	0.95	87.0	91.2	66.1	85.8	92.9
20	0.99	82.7	93.5	50.2	81.0	96.6

* ≈ 0 corresponds to using just the last observation as the estimate.

Table 2. *Efficiency of the weighted average as an estimator and comparison with theoretical efficiency*

Steplength multiplier (w)	Sampling fraction (p)	$\alpha = 0.025$ Cauchy distrib.	$\alpha = 0.025$ Normal distrib.	$\alpha = 0.005$ Cauchy distrib.	$\alpha = 0.005$ Normal distrib.	Theoretical efficiency %
1	$\approx 0^*$	101.5	99.7	96.0	99.8	100
1	0.5	86.3	85.0	79.6	84.2	84.4
1	0.8	78.8	77.3	71.4	76.7	77.1
1	0.95	76.8	75.1	68.6	74.6	75.2
1	0.99	76.6	74.9	68.3	74.4	75.0
2	$\approx 0^*$	76.9	74.6	74.3	75.3	75.0
2	0.5	79.6	78.4	77.2	77.4	77.2
2	0.8	83.6	83.4	81.6	82.0	81.6
2	0.95	86.0	85.7	84.3	84.4	84.1
2	0.99	86.3	86.0	84.5	84.7	84.4
5	$\approx 0^*$	37.0	35.9	37.2	35.3	36.0
5	0.5	60.5	60.3	59.9	60.4	59.5
5	0.8	77.4	76.5	76.5	76.3	75.9
5	0.95	82.5	81.7	80.9	81.0	80.6
5	0.99	82.8	82.1	81.2	81.4	81.0
10	$\approx 0^*$	19.0	18.9	19.2	19.4	19.0
10	0.5	54.8	53.4	53.5	53.4	53.2
10	0.8	74.9	73.9	72.2	74.0	73.0
10	0.95	79.6	78.9	76.8	78.7	78.0
10	0.99	79.9	79.3	76.9	79.0	78.4
20	$\approx 0^*$	10.0	9.8	9.8	9.8	9.8
20	0.5	51.0	52.4	49.1	51.5	50.6
20	0.8	71.8	73.2	66.5	71.2	71.4
20	0.95	76.3	78.5	68.9	75.9	76.4
20	0.99	76.5	78.9	68.6	76.1	76.8

* ≈ 0 corresponds to using just the last observation as the estimate.

Table 3. *Efficiency (%) in estimating 2.5% and 97.5% confidence interval limits and tail areas using three methods and steplength multipliers used by the GB method. Each search used 500 000 nominal steps.*

Distribution	Limit	Confidence limit			Tail area			Steplength
		Unweighted	Weighted	GB	Unweighted	Weighted	GB	multiplier
Inverse exp.	2.5%	100.4	77.3	73.6	100.5	77.3	73.6	2.1
Inverse exp.	97.5%	96.5	77.9	80.1	96.8	78.0	88.7	1.6
Standard exp.	2.5%	96.7	77.1	87.5	96.8	77.1	87.5	1.6
Standard exp.	97.5%	98.2	76.6	73.1	98.5	76.8	73.8	2.1
Gamma(2,θ)	2.5%	102.1	79.4	75.2	102.3	79.5	75.8	2.1
Gamma(5,θ)	2.5%	103.6	77.4	74.8	103.9	77.6	76.3	2.1
Gamma(50,θ)	2.5%	99.7	81.6	81.9	97.2	80.1	79.4	2.0
Gamma(2,θ)	97.5%	99.4	75.8	80.9	100.3	76.4	84.2	1.7
Gamma(5,θ)	97.5%	99.3	78.2	80.0	99.4	78.3	80.0	1.8
Gamma(50,θ)	97.5%	90.1	72.8	73.3	90.3	72.9	71.6	2.0
Cauchy	either	93.1	82.2	28.2	93.1	82.2	28.6	0.4
Logistic	either	98.6	77.6	87.7	98.7	77.6	87.8	1.6
t (7 d.f.)	either	96.6	76.5	89.4	96.9	76.6	89.5	1.5
Normal	either	94.4	77.1	76.5	94.5	77.1	76.5	2.0
Average		97.8	77.7	75.9	97.8	77.7	76.7	

Table 4. *Efficiency (%) in estimating 0.5% and 99.5% confidence interval limits and tail areas using three methods and steplength multipliers used by the GB method. Each search used 500 000 nominal steps.*

Distribution	Limit	Confidence limit			Tail area			Steplength
		Unweighted	Weighted	GB	Unweighted	Weighted	GB	multiplier
Inverse exp.	0.5%	90.5	75.4	68.3	91.4	75.9	68.2	2.3
Inverse exp.	99.5%	92.4	77.2	93.6	94.6	78.6	94.1	1.4
Standard exp.	0.5%	95.0	79.5	95.1	95.0	79.5	95.1	1.4
Standard exp.	99.5%	92.4	73.1	68.2	94.0	74.1	68.4	2.3
Gamma(2,θ)	0.5%	93.4	75.5	69.7	93.8	75.6	70.9	2.3
Gamma(5,θ)	0.5%	95.0	78.1	73.9	96.3	78.8	73.9	2.2
Gamma(50,θ)	0.5%	87.8	71.4	71.2	80.6	67.7	72.0	2.1
Gamma(2,θ)	99.5%	91.5	72.7	85.0	94.2	74.6	88.2	1.5
Gamma(5,θ)	99.5%	88.2	74.0	71.3	91.5	76.3	84.1	1.7
Gamma(50,θ)	99.5%	88.0	74.8	70.6	89.9	76.0	73.9	1.9
Cauchy	either	78.8	85.2	8.4	79.8	85.1	9.6	0.3
Logistic	either	95.2	78.5	95.3	96.3	79.1	95.5	1.4
t (7 d.f.)	either	92.6	75.8	93.2	93.5	76.4	93.6	1.3
Normal	either	92.5	73.9	74.4	93.8	74.7	74.5	2.0
Average		91.0	76.1	74.2	91.8	76.6	75.9	

Table 5

Table 5. *Average efficiency (%) of all distributions except the Cauchy distribution in the coverages of 95% and 99% confidence intervals for searches of varying numbers of nominal steps. Intervals are estimated by the unweighted average for different choices of the steplength multiplier in phase 3 and by the GB method. Values for the Cauchy distribution are given in brackets.*

Steplength multiplier	Nominal number of steps 20,000	100,000	200,000	500,000
95% confidence intervals				
New procedure				
10.0	88.3 (73.0)	90.1 (68.7)	93.1 (74.7)	94.7 (76.6)
20.0	90.0 (81.4)	92.4 (76.9)	96.4 (84.5)	97.0 (89.5)
30.0	91.3 (87.5)	93.7 (78.7)	96.7 (88.0)	97.9 (93.1)
50.0	91.8 (91.3)	93.1 (79.6)	95.5 (91.1)	98.1 (93.5)
GB method	80.2 (51.0)	80.9 (38.3)	80.8 (33.2)	81.2 (28.6)
99% confidence intervals				
New procedure				
10.0	84.4 (46.4)	88.0 (40.4)	91.1 (44.8)	93.4 (51.4)
20.0	81.5 (59.8)	89.4 (52.3)	91.2 (63.2)	94.4 (74.8)
30.0	78.2 (57.7)	88.5 (55.2)	90.5 (69.5)	93.0 (79.8)
50.0	78.6 (58.3)	89.0 (55.6)	88.2 (72.2)	88.6 (83.9)
GB method	81.1 (35.0)	82.2 (19.3)	81.8 (14.6)	82.3 (9.6)