2014

Importance accelerated Robbins-Monro recursion with applications to parametric confidence limits

Zdravjko I Botev
Chris Lloyd, *Melbourne Business School*

Electronic Journal of Statistics
Vol. 0 (0000)
ISSN: 1935-7524

Importance accelerated Robbins-Monro recursion with applications to parametric confidence limits

Zdravko I. Botev

University of New South Wales, Australia, e-mail: botev@unsw.edu.au

and

Chris J. Lloyd

University of Melbourne, Australia, e-mail: c.lloyd@mbs.edu

Abstract: Applying the standard stochastic approximation algorithm of Robbins and Monro (1951) to calculating confidence limits leads to poor efficiency and difficulties in estimating the appropriate governing constants as well as the standard error.

We suggest sampling instead from an alternative importance distribution and modifying the Robbins-Monro recursion accordingly. This can reduce the asymptotic variance by the usual importance sampling factor. It also allows the standard error and optimal step length to be estimated from the simulation. The methodology is applied to computing almost exact confidence limits in a generalised linear model.

MSC 2010 subject classifications: Primary 62F25; secondary 65C05.
Keywords and phrases: stochastic approximation, generalized linear model, confidence limits, profile upper limits, importance sampling.

Received January 0000.

1. INTRODUCTION

Robbins and Monro [1951] proposed a stochastic approximation scheme for solving equations of the form

$$M(\theta) \stackrel{\text{def}}{=} \mathbb{E}_\theta H(\mathbf{Y}) = \alpha \tag{1}$$

where $\mathbf{Y} \in \mathbb{R}^k$ and $\mathbb{E}_\theta$ means expectation with respect to a family of distributions $p(y;\theta)$ indexed by θ. Such schemes are of use when the function $M(\theta)$ is unknown or too complex to give explicitly. They showed that under weak regularity conditions the recursion

$$\theta_{n+1} = \theta_n - c_n(H(\mathbf{Y}_n) - \alpha) \tag{2}$$

converges in mean square to the solution θ^* of (1), where the $\mathbf{Y}_n$ are drawn from $p(y;\theta_n)$. The damping constants c_n must decrease fast enough that the series converges, but slow enough that the series is drawn towards the solution.

The mathematical conditions for this are that $\sum_n c_n^2 < \infty$ and $\sum_n c_n = \infty$. One simple family of such constants is $c_n = c/n^\gamma$ where $\gamma \in (0.5, 1]$.

Blum [1954] and Kallianpur [1954] showed almost sure convergence for $\gamma \in (2/3, 1]$. Chung [1954] showed that the rescaled errors $n^{\gamma/2}(\theta_n - \theta^*)$ converge in distribution to normal with mean zero if $\gamma \in (1/2, 1]$ and the even moments of $H(\mathbf{Y})$ are bounded in θ. This in turn recommends $\gamma = 1$ whence the asymptotic variance is given by

$$\frac{c^2 \mathrm{Var}_{\theta^*}(H(\mathbf{Y}))}{2cM'(\theta^*) - 1} \tag{3}$$

provided that $cM'(\theta^*) > 1/2$ and where $V_\theta(H(\mathbf{Y}))$ denotes variance with respect to $p(y;\theta)$. When c equals $c^* = 1/M'(\theta^*)$, (3) has its minimum value

$$V_{RM} = \mathrm{Var}_{\theta^*}(H(\mathbf{Y}))/[M'(\theta^*)]^2 \tag{4}$$

and is known as the Robbins-Monro (RM) variance bound since it sets the minimum variance of any estimator of θ^* based on simulating $H(\mathbf{Y})$, see Wetherill [1975]. Not surprisingly, both the variance and the optimal constant depend on the first derivative of $M(\theta)$ near the solution θ^*. Consequently, estimation of $M'(\theta^*)$ will be an important part of applying and evaluating the method. Note that the variance is more adversely affected when c is smaller than c^* than when it is larger. For instance, when $c = 0.5c^*$ the variance (3) is infinite while when $c = 2c^*$ it equals $4V_{RM}/3$.

The RM scheme has many applications. One statistical application, pioneered by Garthwaite and Buckland [1992], is to generate a $1-\alpha$ (upper) confidence limit for a parameter θ of a statistical model $p(y;\theta)$ for data Y. This requires solving the equation $\mathbb{P}(T(\mathbf{Y}) \le t;\theta) = \alpha$ for θ where $T(\mathbf{Y})$ is the maximum likelihood estimator (MLE) of θ with observed value t. The left hand side is seldom available in closed form. The RM scheme applies upon letting $H(\mathbf{Y})$ be $\mathbb{I}_{\{T(\mathbf{Y})\le t\}}$, where $\mathbb{I}_{\{A\}}$ is the indicator of the event A.

Even though it has never been pursued in the literature, this scheme can be naturally extended to models with nuisance parameters, see Section 4. Indeed, this application motivates a new accelerated version of the algorithm described in Section 2.

There are many difficulties that can arise with the RM algorithm, two of which are germane to the application just mentioned. The first is that even the minimised variance (4) can be large. For instance, if the estimator T is normally distributed, then $V_{RM} = \alpha(1-\alpha)/\phi^2(z_\alpha)$, where ϕ is the standard normal density and z_α is the α quantile of ϕ, which diverges rapidly when α is near 0. So it will be difficult to simulate a 99.5% limit using this algorithm. The problem with the large variance can sometimes be mitigated by running multiple instances of the RM algorithm O'Gorman [2014], but this may incur increased computational cost. The second problem is that the derivative $M'(\theta)$ is not easily estimated by this scheme. Consequently, the optimal constant c^* and the achieved variance (3) are difficult to obtain. The problem of obtaining the optimal constant c^* can be side-stepped by averaging iterates, as shown in the seminal paper of Polyak and Juditsky [1992]. This motivated Garthwaite and Jones [2009] to

define a more complicated scheme where c^* is estimated in three stages. Their numerical experiments show that the variance of the estimate generated by this stage-wise scheme is, at best, 10% larger than the RM bound. They also do not provide a standard error and recommend as many iterations as possible.

The idea of this paper is quite simple. We express $M(\theta)$ as the expectation of a different variable $H(\mathbf{Y}, \theta)$ with respect to a known θ-free distribution $p_S(\mathbf{y})$ as in Monte Carlo importance sampling, see Kroese et al. [2011]. An astute choice of p_S can lead to an asymptotic variance much smaller than (4), in practical cases 10–100 times smaller. Moreover, we can accurately estimate the optimal constant c^* and the achieved variance, unlike competing methods. The idea of combining the RM method with importance sampling has been considered in Jourdain and Lelong [2009], Lemaire and Pages [2010], Bardou et al. [2009]. However, the first two articles consider the RM procedure mainly as a tool for tuning the tilting parameter of an importance sampling density so that the variance of the estimator is minimal. Our setting and application are different — here the RM scheme estimates the quantity of interest directly and is not a tool for tuning the importance sampling density, which in our case is fixed. In contrast to Bardou et al. [2009], here we consider a more general change of measure, and instead of optimizing the tuning parameter of the importance sampling density, we focus on estimating the optimal damping sequence c_n.

The plan of the paper is as follows. Section 2 describes the new algorithm and states the theoretical convergence results. Section 3 illustrates the feasibility of the method on a toy one-parameter example. Section 4 illustrates how the method may be used to compute accurate confidence limits on a realistic example with many parameters.

2. Importance Accelerated Robbins-Monro

The general context is a parametric model $\tilde{p}(\mathbf{y}; \theta, \tilde{\psi})$ for data $\mathbf{Y}$ where $\theta \in \Theta \subseteq \mathbb{R}$ is a scalar parameter of interest and $\tilde{\psi} \in \Psi$ is a vector of nuisance parameters. Let $\tilde{H}(\mathbf{Y})$ be any statistic, though in later applications it will be the indicator function of a tail set. We seek the value θ^* of θ for which the mean of $\tilde{H}(\mathbf{Y})$ equals a known value. The problem is the free nuisance parameter $\tilde{\psi}$.

Let ψ_θ trace out a one dimensional continuous path in $\Psi \subseteq \mathbb{R}$ as θ varies over Θ. Replacing $\tilde{\psi}$ by ψ_θ reduces the parameter space to one dimension. Motivation for this will be given in Section 4, where ψ_θ will be the MLE of $\tilde{\psi}$ for fixed θ. We denote the restricted model $\tilde{p}(\mathbf{y}; \theta, \psi_\theta)$ by $p(\mathbf{y}; \theta)$ and consider solving

$$M(\theta) \stackrel{\text{def}}{=} \mathbb{E}_\theta \tilde{H}(\mathbf{Y}) = \alpha \tag{5}$$

where $\mathbb{E}_\theta$ denotes expectation with respect to the restricted one-parameter model.

2.1. Importance Accelerated Recursion

Choose any density $p_S(\mathbf{y})$ such that $p_S(\mathbf{y}) = 0 \Rightarrow p(\mathbf{y};\theta) = 0$ for all $\mathbf{y}$. Define $H(\mathbf{Y};\theta) \stackrel{\text{def}}{=} \tilde{H}(\mathbf{Y})w(\mathbf{Y};\theta)$ where $w(\mathbf{Y};\theta) = p(\mathbf{Y};\theta)/p_S(\mathbf{Y})$ are so-called importance weights. By the usual importance sampling argument $M(\theta) = \mathbb{E}_S H(\mathbf{Y};\theta)$, where $\mathbb{E}_S$ denotes expectation with respect to p_S. The key difference in this representation is that the variable $H(\mathbf{Y};\theta)$ now depends on θ rather than the distribution of $\mathbf{Y}$. The key advantage is that we have considerable freedom in how we select the importance distribution $p_S(\mathbf{y})$. In that regard, we use the following result.

Proposition 1. *Let $\mathbf{Y} \in \mathbb{R}^k$ have a known distribution $p_S(\mathbf{y})$ and suppose we wish to solve for θ*

$$M(\theta) = \int_{\mathbb{R}^k} H(\mathbf{y};\theta)p_S(\mathbf{y})\mathrm{d}\mathbf{y} = \mathbb{E}_S H(\mathbf{Y};\theta) = \alpha, \qquad \theta \in \Theta \subseteq \mathbb{R} \tag{6}$$

under the regularity conditions (i) to (vi) given in the supplementary appendix, which ensure that $M(\theta)$ has a unique root, is sufficiently smooth, and $H(\mathbf{Y};\theta)$ has bounded well-behaved second moment. Consider the recursion for $n = 1, 2, 3, \ldots$

$$\theta_{n+1} = \theta_n - \frac{c}{n}(H(\mathbf{Y}_n;\theta_n) - \alpha), \qquad \mathbf{Y}_1, \mathbf{Y}_2, \ldots \stackrel{\text{iid}}{\sim} p_S(\mathbf{y})\ . \tag{7}$$

Then, θ_n converges almost surely and in mean squared error to the unique solution θ^* of (6). Moreover, $n^{1/2}(\theta_n - \theta^*)$ converges in distribution to normal with mean zero and variance

$$\frac{c^2\, Var_S\{H(\mathbf{Y};\theta^*)\}}{2cM'(\theta^*) - 1}, \tag{8}$$

where V_S denotes variance with respect to p_S. The proof involves known results in stochastic approximation theory, but is nevertheless lengthy and thus delegated to the supplementary appendix.

The variance (8) has the same form as (3) except that $\mathrm{Var}_S\{H(\mathbf{Y};\theta^*)\}$ replaces $\mathrm{Var}_{\theta^*}\{\tilde{H}(\mathbf{Y})\}$. Consequently, the variance is minimised at the same value $c^* = 1/M'(\theta^*)$ as for the standard RM scheme. The minimised variance is

$$V_{IS} = \mathrm{Var}_S\{H(\mathbf{Y};\theta^*)\}/M'(\theta^*)^2 \tag{9}$$

which differs from (4) by the ratio $\mathrm{Var}_S\{H(\mathbf{Y};\theta^*)\}/\mathrm{Var}_{\theta^*}\{\tilde{H}(\mathbf{Y})\}$. With an astute choice of p_S, this variance ratio can be made much smaller than 1. The reciprocal of this ratio measures the relative efficiency compared to the standard scheme.

2.2. Estimating the Optimal Constant

For both traditional RM and the new algorithm, the optimal constant c^* depends on $M'(\theta)$. This is difficult to estimate within the standard RM scheme. In practice it can be crudely approximated on a case by case basis.

For instance, consider the Garthwaite and Buckland [1992] application for estimating a confidence limit of coverage $1-\alpha$ where $\tilde{H}(\mathbf{Y}) = \mathbb{I}_{\{T(\mathbf{Y})\geq t\}}$ and T is the MLE of θ with observed value t. Assuming T is normally distributed with mean θ and variance free of θ, it can be shown that ($\hat{\theta} = T(\mathbf{Y})$ being the MLE)

$$c^* = (\theta^* - \hat{\theta})/(z_\alpha \phi(z_\alpha)). \tag{10}$$

The only unknown θ^* is estimated at each step by θ_n. Because the variance is increased much more by under-estimating c^* than over-estimating it, Garthwaite and Buckland (GB) recommend doubling this value. In practice, this leads to variance about 20-30% higher than the RM bound in the problems they considered. Moreover, the RM variance itself is not estimated by the simulation.

The new proposal provides an estimator of $M'(\theta^*)$ and therefore of c^* without any normality assumption. The key point is that $\mathbb{E}_S H(\mathbf{Y};\theta) = M(\theta)$ for all values of θ and that perturbing θ provides information about the first derivative. For instance, for any small value δ (we used $\delta = 0.001$ in our experiments):

$$\mathbb{E}_S\left[H(\mathbf{Y};\theta+\delta) - H(\mathbf{Y};\theta-\delta)\right] = M(\theta+\delta) - M(\theta-\delta) \approx 2\delta M'(\theta) \tag{11}$$

and so $M'(\theta)$ can be estimated via the average of the iterates $H(\mathbf{Y}_n;\theta+\delta) - H(\mathbf{Y}_n;\theta-\delta)$, similar to Kiefer and Wolfowitz [1952]. Theoretically, the value δ is chosen to be as small as possible subject to numerical accuracy, and possibly decaying to zero at a rate $\mathcal{O}(n^{-s}), s \in (0, 1/2)$; see the supplementary appendix. In practice, it is taken to be a very small fraction of the standard error of $\hat{\theta}$. Each new estimate of $M'(\theta)$ can be used to generate a new estimate c_n^* of the optimal step length $c^* = 1/M'(\theta^*)$ at each step of the recursion.

To guard against divergence early in the recursion, one can start with c_n^* equal to a robust but suboptimal sequence, such as the above mentioned approximation of Garthwaite and Buckland and, after a suitable burn-in period, move smoothly to the more refined estimator based on (11) as the recursion progresses. A simple method is given in step 4 of the algorithm below where w is a "burn-in" parameter (we used $w = 50$ in our experiments) and the estimates c_n^* start moving from the robust to the more refined estimator after w steps.

Step 0. Initialise θ_1 and c_1^*.
Step 1. Generate $\mathbf{Y}_n$ from $p_S(\mathbf{y})$.
Step 2. Calculate $H(\mathbf{Y}_n, \theta_n) = \tilde{H}(\mathbf{Y}_n)p(\mathbf{Y}_n;\theta_n)/p_S(\mathbf{Y}_n)$.
Step 3. Calculate $\theta_{n+1} = \theta_n - c_n^*(H(\mathbf{Y}_n,\theta_n) - \alpha)/n$.
Step 4A. When $n < w$ calculate c_{n+1}^* using a robust estimate like GB.
Step 4B. When $n \geq w$, estimate $M'(\theta^*)$ by

$$m_n' = \tfrac{1}{(n-1)2\delta}\textstyle\sum_{i=1}^{n-1}\{H(\mathbf{Y}_i,\theta_i+\delta) - H(\mathbf{Y}_i,\theta_i-\delta)\} \tag{12}$$

and then the optimal step length c^* by

$$c_{n+1}^* = (wc_w^* + (n+1-w)/m_n')/(n+1) \tag{13}$$

Step 5. Cycle steps 1-4 until convergence of the sequence θ_n.

At step 4B, in equation (12) one can alternatively evaluate $H(\mathbf{Y}_i, \theta)$ at $\theta_n \pm \delta$ rather than at $\theta_i \pm \delta$ (which yields a computationally cheaper online estimator). At termination of the algorithm, the variance (9) can be easily estimated. If B is the index of the final iterate then the numerator is estimated by the sample variance of the iterates of $H(\mathbf{Y}_1, \theta_1), .., H(\mathbf{Y}_B, \theta_B)$ and the denominator by m'_B. The supplementary appendix discusses the theoretical convergence under certain constraints on the asymptotic behavior of m'_n.

2.3. Choice of the Function $\tilde{H}(\mathbf{Y})$

There is a slight ambiguity in the formulation of the basic equation (1) to be resolved. We can leave α on the right hand side, or we can subtract it from the function H. In the former case, the increment to θ_n at step 3 becomes $(\tilde{H}(\mathbf{Y}_n) - \alpha)\frac{p(\mathbf{Y}_n;\theta_n)}{p_S(\mathbf{Y}_n)}$ instead of $\tilde{H}(\mathbf{Y}_n)\frac{p(\mathbf{Y}_n;\theta_n)}{p_S(\mathbf{Y}_n)} - \alpha$ ignoring the common constants c_n^*/n. The asymptotics of these two schemes are identical and in our examples their performance could not be distinguished. The slope estimate in step 4B is identical under both formulations.

2.4. Choice of the Importance Distribution p_S

The ratio of the variance of the new algorithm to that of the standard RM algorithm is $\mathrm{Var}_S(H(\mathbf{Y};\theta^*))/\mathrm{Var}_{\theta^*}\{\tilde{H}(\mathbf{Y})\}$. Naturally, we want this factor to be as small as possible.

It is well known that the importance sampling density $p_S(\mathbf{y})$ that minimises the asymptotic variance (9) is proportional to the integrand $\tilde{H}(\mathbf{y})p(\mathbf{y};\theta)$, see, for example, [Kroese et al., 2011, p.364]. This choice will make the factor equal to zero, but since the normalizing constant of this distribution is the unknown $M(\theta)$, this is not much help. Nevertheless, matching the shape of p_S to the integrand $\tilde{H}(\mathbf{y})p(\mathbf{y};\theta)$ is the rough guiding principle of most methods for selecting p_S. The main restrictions are that it should be easy to generate from p_S and that $p_S(\mathbf{y})$ can be quickly calculated for any $\mathbf{y}$.

In higher dimensions finding an importance sampling distribution that matches the shape of the integrand can be quite difficult, and can easily lead to problems that are at least as hard as the original problem, see, for example, Kroese et al. [2011]. However, in our applications we will find that efficient choices of p_S are easily available. The reasons are first that θ is scalar and secondly that $H(y)$ will be a simple indicator function. The recommended choice will be thoroughly explained in our examples.

3. Simple One Parameter Example

We illustrate the method on a toy example, firstly to confirm the theory but also to clarify for the reader exactly how and why the method works.

Consider a single binomial observation $y = 60$ from $n = 100$ trials and an upper 99.5% limit for the success probability θ. The exact upper limit of

Clopper and Pearson [1934] is the unique solution for θ of $\mathbb{P}(Y \leq 60; \theta) = 0.005$ which yields $\theta^* = 0.7238$. We want to estimate this number from an efficient recursion.

3.1. Standard GB Algorithm

In our earlier notation for the Garthwaite and Buckland algorithm, let $M(\theta) = \mathbb{E}_\theta[\mathbb{I}_{Y\leq 60} - 0.005]$ where $\mathbb{E}_\theta$ denotes expectation with respect to the binomial probability function $p(y;\theta)$ with parameters $(100, \theta)$. We want to find the solution θ^* of $M(\theta) = 0$. Note that there is no nuisance parameter in this problem.

The function $M(\theta)$ is decreasing in θ and so the step length constant c needs to be negative, but for ease of presentation we change its sign and write the standard RM recursion as $\theta_{n+1} = \theta_n + c_n^*(\mathbb{I}_{Y_n\leq 60} - 0.005)/n$ where Y_n is generated from the binomial distribution with $\theta = \theta_n$. The optimal step length can be calculated numerically here and equals $c^* = 3.062$. However, this is not available in the GB algorithm, which instead uses the normal approximation given in (10), giving here 3.322. This value is estimated and doubled at each step via the equation

$$c_n^* = 2\frac{(\theta_n - 0.6)}{z_{.005}\phi(z_{.005})} = 53.70(\theta_n - 0.6) \tag{14}$$

One realisation of this algorithm is given as the dashed curve in the left panel of Figure 1. It is worth noting that for around 99.5% of the generated values $\mathbb{I}_{Y_n\leq 60} = 0$ and so the recursion almost always moves in a negative direction, offset by much rarer but larger positive movements.

3.2. Importance Accelerated Recursion

The importance accelerated algorithm involves always sampling from a selected distribution $p_S(y)$. So far we have said little about how to choose $p_S(y)$. There is a wide literature on this issue (see Kroese et al. [2011] and the references therein), but none of those complex methods are required here. A rather common prescription is to use the distribution that concentrates samples near the observed sample. This suggests using the binomial distribution with θ equal to the MLE, i.e. $p_S(y) = p(y;\theta_s)$ with $\theta_s = 0.6$. The new algorithm is $\theta_{n+1} = \theta_n - c_n^* H(Y_n, \theta_n)/n$ where $H(y_n, \theta_n) = (\mathbb{I}_{y_n\leq 60} - 0.005)\left(\frac{\theta_n}{0.6}\right)^{y_n}\left(\frac{1-\theta_n}{0.4}\right)^{100-y_n}$. Notice that the increments to θ_n are now negative/positive roughly half of the time because we are sampling at $\theta = 0.6$. The innovations can also take a smoother range of values depending on the value y_n, rather than the two values possible under the standard algorithm. So convergence will be more symmetric and smoother.

A realisation of 5000 iterates is displayed in the left panel of Figure 1 as a solid line. To focus attention on the key idea of sampling from an importance distribution, the same sequence of constants c_n^* is used in both cases, namely

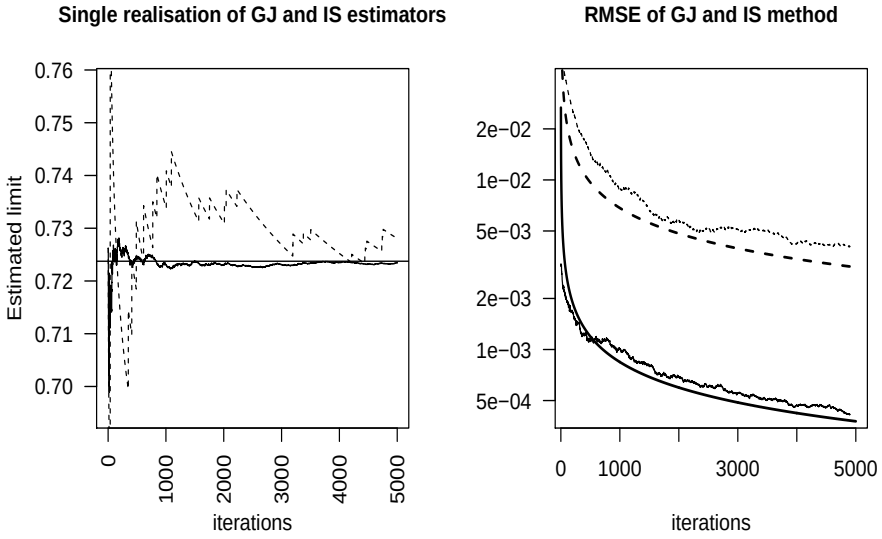

FIG 1. **Estimating a binomial confidence limit.** *Left.* A single realised path of the GB/RM algorithm (dashed) and new algorithm (solid). *Right.* RMSE at each iteration, across 1000 repetitions. Vertical scale is logarithmic. Upper curves are for GB/RM and lower are for new algorithm. The jagged curves are the observed RMSE and the smooth dashed curves are the optimal standard deviation from (4) and (9).

the sub-optimal Garthwaite and Buckland constants. Clearly the new method gets to the answer much more quickly for this realisation.

Both procedures were repeated 1000 times and the root mean square error (RMSE) calculated at each iteration. The right panel plots the RMSE against the iteration number (on a log-scale). The upper curves are for the standard GB method and the lower for the new method. The smooth dashed curves are the theoretical bounds given by (4) and (9) and the jagged curves are the empirical RMSE. The new method has error around 9.8 times smaller, which means the same accuracy as the GB method can be achieved with roughly 100 times less simulation. This is consistent with recommendations in Garthwaite and Jones [2009] that at least 200,000 simulations are required in practice using their non-accelerated method.

3.3. Estimating the Optimal Constant

For both algorithms, it is apparent in the plots that the observed RMSE is (around 20%) higher than the theoretical bound, because of the deliberate over-estimation of c^* involved in using (14). Figure 2 investigates the effectiveness of estimating $c^* = 3.062$ using (13) in step 4B, with burn-in parameter $w = 50$.

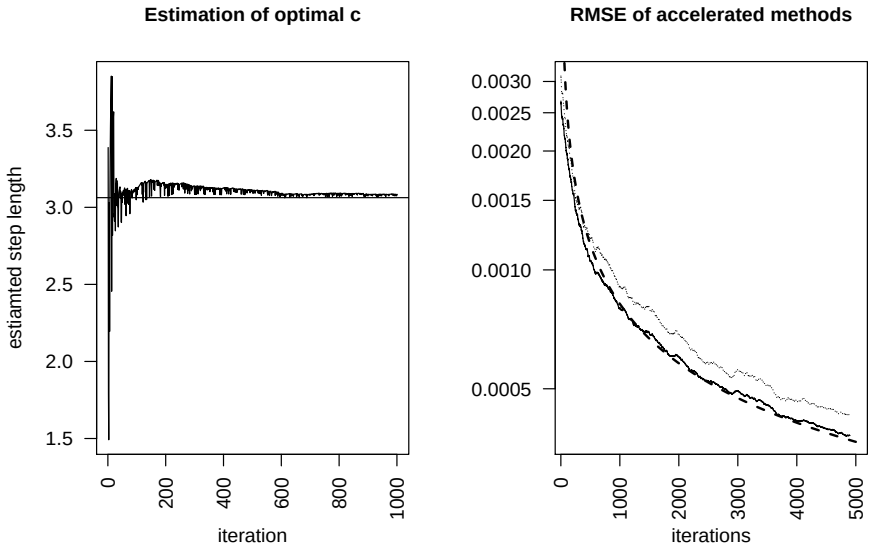

FIG 2. **Step length estimation.** *Left.* Realisation of estimate (13) of optimal step length $c^* = 3.062$ (horizontal line). *Right.* RMSE of new method using (13) (dashed) and (14) (dotted) to estimate c^*. Theoretical bound (8) is smooth dashed curve. Vertical scale is logarithmic.

The left panel shows the estimate of the optimal step length for the first 1000 iterations of a single run. Apparently, one arrives at close to the optimal value within a few hundred iterations. The right panel shows the RMSE of the final accelerated estimate, again estimated across 1000 repetitions. Our new method is the dashed line and achieves the bound (4) displayed as a smooth dashed line. The previous method with c^* estimated by (14) is the dotted line and is included for reference. Clearly, the major advance here is using the importance distribution to simulate Y_n, and the improved estimation of c^* is secondary. The other major advance is that the variability of θ_n as an estimator of θ^* can be estimated.

In this case, referring to (9), the sample variance of the $H(Y_n;\theta_n)$ turns out to equal 7.423×10^{-5} for the simulation presented in Figure 1, and the last value of the estimated slope is -0.329 (the true slope is -0.326) and so the variance (9) is estimated by 6.963×10^{-4}. With $n = 1000$ iterates, this implies a standard error of $\sqrt{0.0006963/1000} = 0.00083$ for the final estimate, which in this case was 0.7233 (the true value is $\theta^* = 0.7237$).

4. General Limits with Nuisance Parameters

The main motivation for the importance accelerated algorithm presented in this article is to compute highly accurate confidence limits for parametric models $\tilde{p}(\mathbf{y};\theta,\tilde{\psi})$ with nuisance parameters. To this end, it is necessary to give a very

short detour on the theory for the kind of limits we want to compute.

4.1. Profile Upper Limits

A very general method of generating an upper limit for θ starts with a statistical quantity $T(\mathbf{Y}, \theta)$ and equates tail probabilities to the target coverage error. Because the nuisance parameters $\tilde{\psi}$ are free, they are replaced by the (restricted) MLE $\hat{\psi}_\theta$ for known θ. The restricted MLE plays the role of ψ_θ in the earlier theory of section 2. We then solve

$$\mathbb{P}(T(\mathbf{Y}, \theta) \leq T(\mathbf{y}, \theta); \theta, \hat{\psi}_\theta) = \alpha, \tag{15}$$

see, for instance, [Davison and Hinkley, 1997, p.233]. One could further approximate the distribution of $T(\mathbf{Y}; \theta)$ by an asymptotic distribution but we want to use the exact distribution.

When the statistical quantity $T(\mathbf{Y}, \theta)$ is an approximate test statistic this is called the test inversion method. For discrete models, which will be the focus of this section, this is problematic because the left hand side of (15) is a discontinuous function of θ. Test inversion limits are not feasible for discrete data. Computability requires a choice of $T(\mathbf{Y}; \theta)$ that does not depend on θ.

An obvious choice is to let $T(\mathbf{Y}) = \hat{\Theta}$ be the MLE. The solution of (15) is then often known as the parametric bootstrap upper limit. For illustrating our new algorithm we will make a different choice. We take $T(\mathbf{Y})$ to itself be an approximate upper limit for θ. Theory described in Buehler [1957] and Kabaila and Lloyd [2001] indicates unequivocally that this is a better choice than the MLE. Moreover, the final limit is hardly affected by which of the standard approximate limits is used, see Kabaila and Lloyd [2002]. The resulting limits are called *profile* upper limits. If the reader finds this choice of $T(\mathbf{Y})$ problematic, one can use the MLE in what follows without any changes to the algorithm at all. The only difference is that the final computed upper limit returned by our new algorithm will tend to have poorer statistical properties (i.e. will tend to be larger) than the profile upper limit.

So our problem is to solve (15) where $T(\mathbf{Y}) = \hat{\theta}(\mathbf{Y}) + z_{1-\alpha}\hat{\sigma}(\mathbf{Y})$ in obvious notation, this being the most common approximate upper limit for θ. As in Section 2, denote $\tilde{p}(\mathbf{y}; \theta, \hat{\psi}_\theta)$ by $p(\mathbf{y}; \theta)$. In our accelerated algorithm we simulate data sets $\mathbf{Y}_n$ from the model $p(\mathbf{y}_n; \hat{\theta})$, i.e. from the fitted model. The innovation at step 2 of the general algorithm is

$$H(\mathbf{Y}_n, \theta_n) = \left(\mathbb{I}_{T(\mathbf{Y}_n) \leq t} - \alpha\right) \frac{p(\mathbf{Y}_n; \theta_n)}{p(\mathbf{Y}_n; \hat{\theta})}.$$

For each iterate, the model needs to be fitted twice, once to compute $T(\mathbf{y}_n)$ and again to find the restricted MLE $\hat{\psi}(\theta_n)$, suppressed in the notation $p(\mathbf{y}_n; \theta_n)$, using the original data set. The latter is easily computed in most packages, for instance using the offset command in R. We later point out several methods for reducing the number of model fits during the recursion.

4.2. Example

These data are from Gordon and Foss [1966] and have been use by many authors to illustrate accurate inference for discrete data, for instance Cox and Snell [1989] and Brazzale et al. [2007]. On 18 days, babies not crying at a specified time were chosen as subjects. One baby was randomly chosen for stimulation, which is the treatment. The response was whether or not the baby was crying at the end of a specified period. Each row of data in Table 1 is a 2×2 table measuring the effect of stimulation on crying.

TABLE 1
Crying babies data from Gordon and Foss (1966).

	Control babies		Treated babies	
Day	Not Crying	Crying	Not Crying	Crying
1	3	5	1	0
2	2	4	1	0
3	1	4	1	0
4	1	5	0	1
5	4	1	1	0
6	4	5	1	0
7	5	3	1	0
8	4	4	1	0
9	3	2	1	0
10	8	1	0	1
11	5	1	1	0
12	8	1	1	0
13	5	3	1	0
14	4	1	1	0
15	4	2	1	0
16	7	1	1	0
17	4	2	0	1
18	5	3	1	0

The data $\mathbf{Y}$ here consists of 36 binomial counts and can be modeled by a logistic regression. The model $\tilde{p}(\mathbf{y}; \theta, \tilde{\psi})$ is a generalised linear model, the interest parameter θ is the effect of the treatment on the log-odds of not crying and the nuisance parameters $\tilde{\psi}$ describe the effects of the dummy variables for each of the 18 days.

Alternative standard approximate methods give considerably different answers. A 95% upper limit for θ based on the Wald statistic is 2.631 and on the LR statistic is 2.785. To estimate the profile upper limit, we simulated 36 counts from the fitted model $B = 5000$ times. For each simulation, we needed to fit the model twice, one time to compute the Wald upper limit and a second time to compute the restricted MLE. The resulting estimate was 2.761 with simulation error 0.006.

A 95% *lower* limit for θ based on the Wald statistic is -0.460 and on the LR statistic is -0.252. The profile lower limit is estimated simply by replacing

the covariate indicating treatment by a covariate indicating non-treatment. Our estimate of this limit from $B = 5000$ iterations is -0.316 with simulation error 0.007.

4.3. Diagnostics

The importance accelerated method can break down for a poor choice of p_S. Recall the importance weights $w(\mathbf{y};\theta) = p(\mathbf{y};\theta^*)/p_S(\mathbf{y})$ and note that $\mathbb{E}_S\{w(\mathbf{Y};\theta^*)\} = 1$. Poor efficiency occurs when the numerator

$$\mathrm{Var}_S\{H(\mathbf{Y};\theta^*)\} = \mathrm{Var}_S\{(\mathbb{I}_{T(\mathbf{Y})\leq t} - \alpha)w(\mathbf{Y};\theta^*)\}$$

in (8) is large. So there will be a problem when the mean square weights $w^2(\mathbf{y};\theta^*)$ are large on that part of the sample space where $T(\mathbf{y}) \leq t$. Since the mean weights equal 1, this could only occur if the distribution of $w(\mathbf{Y};\theta)$ becomes highly skewed. This is characterized by the existence of very large weights with very small probability.

This situation is detected from the simulations in one of two ways. In the unlikely event that an improbable large weight appears in the sample, the sample variance will be appropriately large and estimated efficiency compared to standard RM will be less than 1. In the more likely case that the improbable large weights do not appear in the sample, the sample mean of the weights will be much less than the theoretical value of 1. So we diagnose estimation breakdown when either the efficiency or the mean sample weights are much less than 1.

For the previous example, the mean value of the observed weights was 0.99 and the relative efficiency of the new method was 9.79 compared to standard RM. This suggests that the choice of importance distribution has led to a reliable and highly efficient estimate.

With the current application to confidence limits in logistic regression, the only cases where breakdown occurred was when the ML estimator $\hat{\theta}$ was infinite. In this case, using $p_S(y) = p(y;\theta)$ with θ equal to a slight perturbation of the ML estimator was successful.

4.4. Computational Savings

The computation took just over 40 seconds using the algorithm as described so far but it can be speeded up significantly. Recall that the approximate upper limit $T(\mathbf{y}_n)$ has to be computed for each iteration which requires B fits of the model. These fits can be sped up by fitting the model to batches of simulated data simultaneously. Consider the first b simulated data sets $\mathbf{y}_1, ..., \mathbf{y}_b$. We fit a joint model to these data sets with b sets of parameters (θ_j, ψ_j) for $j = 1, ..., b$. For instance, in the previous example we define a dummy variable for each batch of 36 counts and interact this with the 19 parameters of the model.

The point is that this computation will often take far less than b times as long as a single fit. The optimal value of b depends on the package being used

but can be easily determined by experimentation. In the previous example, we found that $b = 5$ leads to a reduction in computation by a factor of 3.

The original algorithm also required computation of $\hat{\psi}_\theta(y_{\text{obs}})$ at $\theta = \theta_n$ which requires another B fits of a restricted model to the original data. This can be largely avoided also. One option is to use the fact that $\hat{\psi}_\theta(y_{\text{obs}})$ is a smooth function of θ from the first 100 or so iterations. One can then extrapolate or interpolate at subsequent values of θ using a polynomial spline. We have found that this approach was completely successful in the previous example and halved the total computing time. The same approach should work for any model where $\hat{\psi}_\theta$ is smooth in θ. A variant on this method is to combine the smooth extrapolation with an explicit calculation of $\hat{\psi}_\theta$ at $\theta = \theta_n$ say every 10-th iteration. If nothing else, this can be used to verify the extrapolation. In summary, these two computational devices reduced the computational cost by a factor of five (from 40 seconds to 8 seconds) in the previous example.

5. Conclusion

This paper presents a new accelerated Robbins-Monro algorithm that can reduce computation by an order of magnitude or more. The new computational approach is motivated by the problem of finding an accurate confidence limit. Previous research has only dealt with the standard RM scheme and with models without nuisance parameters. The proposed method opens the way to more general application of stochastic approximation to higher order frequentist inference.

The Robbins-Monro algorithm has been used to estimate an optimal importance sampling distribution from within a parametric family, see most recently Lemaire and Pages [2010]. However, inserting importance sampling into the Robbins-Monro algorithm itself appears to be new, to the statistical literature at least.

It has been known since Polyak and Juditsky [1992] and Kushner and Yang [1993] that the RM method bound can be achieved by averaging the iterates rather than just taking the last. Moreover, provided that the damping constants c_n decrease more slowly than $\mathcal{O}(1/n)$, the precision does not depend on these constants. Whether or not these ideas can be applied to the accelerated algorithm studied in this paper is an area for future research.

ACKNOWLEDGMENTS

Zdravko Botev acknowledges the support of the Australian Research Council under grant DE140100993.

Appendix A: Proof of Proposition 1

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $\mathbf{Y} : \Omega \to \mathbb{R}^k$ with pdf $p_S(\mathbf{y})$, and $p(\mathbf{y}; \theta)$ is another pdf with $\theta \in \Theta \subseteq \mathbb{R}$. Recall that we wish to find the root of

the function $M : \Theta \to \mathbb{R}$, where without loss of generality and for simplicity of notation we assume $\alpha = 0$. The non-adaptive Robbins-Monro recursion (that is, without any attempt to estimate c^*) is given by $\theta_{n+1} = \theta_n - c\, H(\mathbf{Y}_n; \theta_n)/n$. We derive the asymptotic distribution of θ_n under the following assumptions.

(i) Let $M(\theta^*) = 0$ and assume $\theta \neq \theta^* \Rightarrow (\theta - \theta^*)M(\theta) > 0$;
(ii) $M(\cdot)$ is continuously differentiable and $\sup_\theta |M'(\theta)| \leq K_1$;
(iii) For every $\delta > 0$ we have $\inf_{|\vartheta - \theta^*| \geq \delta} |M(\vartheta)| > 0$;
(iv) Constant c is such that for every $\delta > 0$ we have $\inf_{|\vartheta - \theta^*| < \delta} 2cM'(\vartheta) > 1$;
(v) Finite variance: for all $\vartheta \in \Theta$, we have $\mathbb{E}H^2(\mathbf{Y}; \vartheta) < \infty$;
(vi) For every $\theta_n \xrightarrow{\text{a.s.}} \theta^*$ and $\delta > 0$ we have $\lim_{n\uparrow\infty} \mathbb{E}H^2(\mathbf{Y}; \theta_n)\, \mathbb{I}_{\{H^2(\mathbf{Y};\theta_n) > n\delta\}} = 0$.

We now proceed to derive the asymptotic distribution using a result of Fabian [1968]. For convenience we restate a simpler one-dimensional version of his result using our notation. This result corresponds to the one in Fabian's paper given $\alpha = \beta = k = 1$ and regularity condition (2.2.3).

Theorem 1 (Fabian [1968])**.** *Let $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \cdots$ be a non-decreasing sequence of σ-fields. Let $U_n, V_n, T_n, \Gamma_n, \Phi_n \in \mathbb{R}$ be random variables for all n and $\sigma_n^2 \stackrel{def}{=} \mathbb{E}[V_n^2 \,|\, \mathcal{F}_n]$. Assume that:*

- $U_{n+1} = \left(1 - \dfrac{\Gamma_n}{n}\right) U_n + \dfrac{1}{n}\Phi_n V_n + \dfrac{1}{n^{3/2}} T_n$;
- $\Gamma_n, \Phi_{n-1}, V_{n-1}$ *are $\mathcal{F}_n$-measurable;*
- $\Gamma_n \xrightarrow{a.s.} \gamma > 1/2, \quad \Phi_n \xrightarrow{a.s.} \phi, \quad T_n \xrightarrow{a.s.} t \quad (\text{or } T_n \xrightarrow{L_2} t)$ *for $\gamma > 1/2$ and $\phi, t \in \mathbb{R}$;*
- $\mathbb{E}[V_n \,|\, \mathcal{F}_n] = 0, \quad \sigma_n \xrightarrow{a.s.} \sigma < \infty$ *and* $\lim_{n\uparrow\infty} \mathbb{E}[V_n^2 \mathbb{I}_{\{V_n^2 \geq n\delta\}}] = 0$ *for every $\delta > 0$.*

Then, $\sqrt{n}U_n$ converges in distribution to a normal with mean $\dfrac{2t}{2\gamma - 1}$ and variance $\dfrac{\phi^2\sigma^2}{2\gamma - 1}$.

To apply Fabian's result to the recursion $\theta_{n+1} = \theta_n - \frac{c}{n} H(\mathbf{Y}_n; \theta_n)$, let $\mathcal{F}_n = \sigma\{\theta_1, \mathbf{Y}_1, \ldots, \mathbf{Y}_{n-1}\}$ be the smallest σ-field with respect to which the indicated random variables are measurable. We have that $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \cdots$ is an increasing sequence of σ-fields. The recursion can be rewritten as follows:

$$\underbrace{\theta_{n+1} - \theta^*}_{U_{n+1}} = \left(1 - \frac{c\xi(\theta_n)}{n}\right) \underbrace{(\theta_n - \theta^*)}_{U_n} - \frac{1}{n} \underbrace{c\,[H(\mathbf{Y}_n, \theta_n) - M(\theta_n)]}_{V_n},$$

where $\xi(\theta) = (M(\theta) - M(\theta^*))/(\theta - \theta^*)$ if $\theta \neq \theta^*$ and $\xi(\theta) = M'(\theta^*)$, otherwise. We thus apply Theorem 1 using the substitutions (all identities hold in the almost sure sense):

$$\begin{aligned} U_n &= \theta_n - \theta^*, \qquad \Gamma_n = c\xi(\theta_n), \qquad \Phi_n = \phi = -1 \\ T_n &= t = 0, \qquad V_n = c(H(\mathbf{Y}_n, \theta_n) - M(\theta_n)) \end{aligned}$$

We now need to establish each of the dot points in Fabian's result:

1. It is clear from $\theta_{n+1} = \theta_n - \frac{c}{n} H(\mathbf{Y}_n; \theta_n)$ that since the sequence of random variables $\theta_2, \theta_3, \ldots, \theta_n$ is constructed from the sequence $\theta_1, \mathbf{Y}_1, \ldots, \mathbf{Y}_{n-1}$, then Γ_n and $V_{n-1} = c(H(\mathbf{Y}_{n-1}, \theta_{n-1}) - M(\theta_{n-1}))$ are both $\mathcal{F}_n$-measurable.
2. Under the assumptions (i),(ii), and (iii) we have that $\theta_n \overset{\text{a.s.}}{\to} \theta^*$ and in mean square sense. First, from $\theta_{n+1} = \theta_n - \frac{c}{n} H(\mathbf{Y}_n; \theta_n)$ we have

$$\theta_{n+1} - \theta_1 + c \sum_{k=1}^{n} \frac{M(\theta_k)}{k} = -\sum_{k=1}^{n} \frac{V_k}{k} \tag{16}$$

   so that $S_n \overset{\text{def}}{=} \sum_{k=1}^{n} V_k/k$ is a discrete-time martingale with respect to the filtration $\{\mathcal{F}_n, n \geq 1\}$. Further, the independence of the $\mathbf{Y}_1, \mathbf{Y}_2, \ldots,$ implies that $\mathbb{E} S_n^2 = \sum_{k=1}^{n} \mathbb{E}[V_k^2]/k^2$, which by (v) is uniformly bounded and finite. Hence, there exists a finite square integrable S such that $S_n \xrightarrow{\text{a.s.}} S$ and $S_n \xrightarrow{L_2} S$. Hence, the left-hand side of (16) converges a.s. and in mean square. Now, the argument of Blum [1954] is that if we assume θ_n does not converge to θ^*, we reach a contradiction. For example, assume that $\theta_n \xrightarrow{\text{a.s.}} \breve{\theta} \neq \theta^*$. Now, note that $\theta_{k+1} = \theta_k - \frac{c}{k} H(\mathbf{Y}_k; \theta_k)$ implies $\theta_{n+1} - \frac{1}{n}\sum_{k=1}^{n} \theta_k = -\frac{c}{n}\sum_{k=1}^{n} H(\mathbf{Y}_k; \theta_k)$ or

$$\tfrac{1}{n}\textstyle\sum_{k=1}^{n} \theta_k - \theta_{n+1} = \tfrac{1}{n}\sum_{k=1}^{n} V_k + \tfrac{c}{n}\sum_{k=1}^{n} M(\theta_k) \tag{17}$$

   Since, $\theta_n \xrightarrow{\text{a.s.}} \breve{\theta}$ and $M(\cdot)$ is continuous, Cesaro summation implies that $\frac{c}{n}\sum_{k=1}^{n} M(\theta_k) \xrightarrow{\text{a.s.}} cM(\breve{\theta})$ and $\frac{1}{n}\sum_{k=1}^{n} \theta_k \xrightarrow{\text{a.s.}} \breve{\theta}$. In addition, Kronecker's lemma (if $b_n \downarrow 0$ and $\sum_k a_k b_k < \infty$, then $b_n \sum_{i=1}^{n} a_i \to 0$) implies that $\frac{1}{n}\sum_{k=1}^{n} V_k \xrightarrow{\text{a.s.}} 0$, because $\sum_{k=1}^{\infty} V_k/k$ is finite. Therefore, taking limits on both sides of (17) yields the identity $M(\breve{\theta}) = 0$, whence we conclude that $\breve{\theta} = \theta^*$. Similarly, we can rule out all other possibilities like $\limsup_n \theta_n = \infty$ or $\theta^* < \liminf_n \theta_n$, and so on, except $\limsup_n \theta_n = \liminf_n \theta_n = \theta^*$.
3. Using point 2. above and the assumption (ii), it follows by the continuous mapping theorem that $\Gamma_n = c\xi(\theta_n) \overset{\text{a.s.}}{\to} cM'(\theta^*) = \gamma$, where from (iv) we have that $\gamma > 1/2$.
4. Using the fact that $\mathbf{Y}_1, \cdots, \mathbf{Y}_n$ are independent, we have with probability one

$$\begin{aligned}
\mathbb{E}[V_n \,|\, \mathcal{F}_n] &= c\mathbb{E}[H(\mathbf{Y}_n, \theta_n) - M(\theta_n) \,|\, \theta_n] = c\mathbb{E}[H(\mathbf{Y}, \theta_n) \,|\, \theta_n] - cM(\theta_n) = 0 \\
\sigma_n^2 &= \mathbb{E}[V_n^2 \,|\, \mathcal{F}_n] = c^2\mathbb{E}[(H(\mathbf{Y}_n, \theta_n) - M(\theta_n))^2 \,|\, \theta_n] \\
&= c^2\mathbb{E}H^2(\mathbf{Y}; \theta_n) - c^2 M^2(\theta_n) \overset{\text{a.s.}}{\to} c^2\mathbb{E}_S H^2(\mathbf{Y}; \theta^*) \overset{\text{def}}{=} \sigma^2,
\end{aligned}$$

   where we have used assumptions (v) and (vi).

Thus, all conditions of Theorem 1 are satisfied and we conclude that $\sqrt{n}(\theta_n - \theta^*)$ converges in distribution to normal with mean zero and variance given by $c^2 \mathrm{Var}_S(H(\mathbf{Y}; \theta^*))/(2cM'(\theta^*) - 1)$.

A.1. Importance accelerated Robbins-Monro scheme

Regarding the asymptotic properties of the adaptive version, where we attempt to estimate the optimal value of the derivative $M'(\theta^*)$ online, we need to control the asymptotic behavior of m'_n so that: (a) it ultimately converges to $M'(\theta^*)$ as $n \uparrow \infty$; (b) it does not become too small or too large too rapidly, compromising the RM recursion. To achieve (a) we let $\delta = \delta_n$, where $\delta_n = \mathcal{O}(n^{-s}),\ s \in (0, 1/2)$ goes to zero, but not too fast. Regarding objective (b), we consider the clipped estimator ($c_1 \gg c_2 > 0$):

$$m_n \stackrel{\text{def}}{=} \begin{cases} \text{sign}(m'_n)c_2 n^{-1/2+\epsilon}, & \text{if } |m'_n| < c_2 n^{-1/2+\epsilon} \\ \text{sign}(m'_n)c_1, & \text{if } |m'_n| > c_1 \\ m'_n, & \text{otherwise} \end{cases}, \tag{18}$$

where we recall that $m'_n = \frac{1}{n-1}\sum_{k=1}^{n-1} Z_k$ with $Z_k \stackrel{\text{def}}{=} \frac{H(\mathbf{Y}_k;\theta_k+\delta_k)-H(\mathbf{Y}_k;\theta_k-\delta_k)}{2\delta_k}$. Note that m'_n is deliberately a sum up to $n-1$, so that it and m_n are $\mathcal{F}_n$ measurable.

We now can consider the asymptotic behavior of the adaptive Robbins-Monro scheme: $\theta_{n+1} = \theta_n - \frac{1}{n\, m_n} H(\mathbf{Y}_n, \theta_n)$. This can again be recast into a Fabian recursion:

$$\theta_{n+1} - \theta^* = \left(1 - \frac{\xi(\theta_n)/m_n}{n}\right)(\theta_n - \theta^*) + \frac{1}{n}\frac{-1}{m_n}(H(\mathbf{Y}_n;\theta_n) - M(\theta_n)) \,. \tag{19}$$

via the substitution $U_n = \theta_n - \theta^*$, $V_n = H(\mathbf{Y}_n;\theta_n) - M(\theta_n)$, $\Phi_n = -1/m'_n$, $\Gamma_n = \xi(\theta_n)/m'_n$, $T_n = t = 0$.

Assuming that $m_n \xrightarrow{\text{a.s.}} M'(\theta^*)$ and $\theta_n \xrightarrow{\text{a.s.}} \theta^*$, it follows that $\Phi_n \xrightarrow{\text{a.s.}} -1/M'(\theta^*)$ and $\Gamma_n \xrightarrow{\text{a.s.}} 1$, and by Fabian's theorem the asymptotic distribution of $\sqrt{n}(\theta_n - \theta^*)$ is zero-mean normal with variance $\text{Var}_S(H(\mathbf{Y};\theta^*))/[M'(\theta^*)]^2$. Thus, the only difficulty is showing that $\theta_n \xrightarrow{\text{a.s.}} \theta^*$ and $m_n \xrightarrow{\text{a.s.}} M'(\theta^*)$.

Proof of $\theta_n \xrightarrow{\textbf{a.s.}} \theta^*$ and $m_n \xrightarrow{\textbf{a.s.}} M'(\theta^*)$. First, note that from (i) we have $\mathbb{E}[Z_k \,|\, \mathcal{F}_k] = \frac{M(\theta_k+\delta_k)-M(\theta_k-\delta_k)}{2\delta_k} > 0$, almost surely. By repeating the martingale arguments in the previous section, one can easily show that $\sum_{k=1}^n \frac{Z_k - \mathbb{E}[Z_k \,|\, \mathcal{F}_k]}{k}$ with variance of order $\mathcal{O}(\sum_k (k\delta_k)^{-2}) = \mathcal{O}(\sum_k k^{-2(1-s)})$ converges to a square integrable random variable for $s < 1/2$. Hence, by Kronecker's lemma we have $\frac{1}{n}\sum_{k=1}^n (Z_k - \mathbb{E}[Z_k \,|\, \mathcal{F}_k]) \xrightarrow{\text{a.s.}} 0$. Now, since $\left|\mathbb{E}[Z_k \,|\, \mathcal{F}_k]\right| \le |M'(\theta_k)| + \mathcal{O}(k^{-2s})$ with $s > 0$, it follows by Kronecker's lemma that $\frac{1}{n}\sum_{k=1}^n \mathbb{E}[Z_k \,|\, \mathcal{F}_k]$ converges almost surely. In other words, $m'_n \xrightarrow{\text{a.s.}} \lim_{n\uparrow\infty} \frac{1}{n}\sum_{k=1}^n M'(\theta_k)$. Note that if $\theta_k \xrightarrow{\text{a.s.}} \theta^*$, then Cesaro summation implies that $\frac{1}{n}\sum_{k=1}^n M'(\theta_k) \xrightarrow{\text{a.s.}} M'(\theta^*)$ and hence $m'_n \xrightarrow{\text{a.s.}} M'(\theta^*)$. Thus, it remains to show that $\theta_k \xrightarrow{\text{a.s.}} \theta^*$.

As in the non-adaptive case, $\mathbb{E}[V_k/m_k \,|\, \mathcal{F}_k] = 0$ (almost surely) and thus $S_n = \sum_{k=1}^n \frac{V_k}{k m_k}$ is a martingale with respect to $\{\mathcal{F}_n, n = 1, 2, \ldots\}$ with second moment $\mathbb{E}S_n^2 = \sum_{k=1}^n \frac{\mathbb{E}(V_k/m_k)^2}{k^2}$, which is bounded by (v) and the clipping from below in (18). Hence, S_n converges almost surely and in mean square sense.

Given this, repeating the argument by contradiction as in the non-adaptive case we can show that $\mathbb{P}(\liminf \theta_n = -\infty) = \mathbb{P}(\limsup \theta_n = \infty) = 0$.

Next, recall Lemma 1 of Venter [1967], which states that if $\zeta_{n+1} = (1 - a_n)\zeta_n - b_n$ and $a_n \downarrow 0$ (from above), $\sum_n a_n = \infty, \sum_n b_n < \infty$, then $\zeta_n \to 0$. We can apply this lemma here in the almost sure sense on the recursion (19) with the substitution $\zeta_n = U_n, a_n = \xi(\theta_n)/(nm_n)$ and $b_n = V_n/(nm_n)$. Here, $\sum_n b_n < \infty$ almost surely, and since θ_n cannot converge to $\pm\infty$, we have $a_n > 0$ and $a_n \downarrow 0$ with $\sum_n a_n = \infty$, establishing that $U_n \xrightarrow{\text{a.s.}} 0$ or $\theta_k \xrightarrow{\text{a.s.}} \theta^*$.

References

O. Bardou, N. Frikha, and G. Páges. Computing VaR and CVaR using stochastic approximation and adaptive unconstrained importance sampling. *Monte Carlo Methods and Applications*, 15(3):173–210, 2009.

J. R. Blum. Multidimensional stochastic approximation methods. *The Annals of Mathematical Statistics*, 25(4):737–744, 1954.

A. R. Brazzale, A. C. Davison, and N. Reid. *Applied asymptotics: case studies in small-sample statistics*, volume 23. Cambridge University Press, 2007.

R. J. Buehler. Confidence intervals for the product of two binomial parameters. *Journal of the American Statistical Association*, 52(280):482–493, 1957.

K. L. Chung. On a stochastic approximation method. *The Annals of Mathematical Statistics*, 25(3):463–483, 1954.

C. J. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934.

D. R. Cox and E. J. Snell. *Analysis of binary data*, volume 32. CRC Press, 1989.

A. C. Davison and D. V. Hinkley. *Bootstrap methods and their application*. Cambridge University Press, 1997.

V. Fabian. On asymptotic normality in stochastic approximation. *The Annals of Mathematical Statistics*, 39(4):1327–1332, 1968.

P. H. Garthwaite and S. T. Buckland. Generating Monte Carlo confidence intervals by the Robbins-Monro process. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 41(1):159–171, 1992.

P. H. Garthwaite and M. C. Jones. A stochastic approximation method and its application to confidence intervals. *Journal of Computational and Graphical Statistics*, 18(1):184–200, 2009.

T. Gordon and B. M. Foss. The role of stimulation in the delay of onset of crying in new-born infants. *J. Exp. Psychol.*, 16:79–81, 1966.

B. Jourdain and J. Lelong. Robust adaptive importance sampling for normal random vectors. *The Annals of Applied Probability*, 19(5):1687–1718, 2009.

P. Kabaila and C. J. Lloyd. The importance of the designated statistic on Buehler upper limits on a system failure probability. *Technometrics*, 44(4), 2002.

P. V. Kabaila and C. J. Lloyd. Profile upper confidence limits from discrete data. *Austral.NZ J. Statist.*, 42:67–80, 2001.

G. Kallianpur. A note on the Robbins-Monro stochastic approximation method. *The Annals of Mathematical Statistics*, 25(2):386–388, 1954.

J. Kiefer and J. Wolfowitz. Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics*, 23(3):462–466, 1952.

D. P. Kroese, T. Taimre, and Z. I. Botev. *Handbook of Monte Carlo Methods*, volume 706. John Wiley & Sons, 2011.

H. J. Kushner and J. Yang. Stochastic approximation with averaging of the iterates: Optimal asymptotic rate of convergence for general processes. *SIAM Journal on Control and Optimization*, 31(4):1045–1062, 1993.

V. Lemaire and G. Pages. Unconstrained recursive importance sampling. *The Annals of Applied Probability*, 20(3):1029–1067, 2010.

T. W. O'Gorman. Regaining confidence in confidence intervals for the mean treatment effect. *Statistics in medicine*, 33(22):3859–3868, 2014.

B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.

H. Robbins and S. Monro. A stochastic approximation method. *The Annals of Mathematical statistics*, 22(3):400–407, 1951.

J.H. Venter. An extension of the robbins-monro procedure. *The Annals of Mathematical Statistics*, pages 181–190, 1967.

G. B. Wetherill. *Sequential methods in statistics*. Chapman and Hall, second edition, 1975.