ADAPTIVE DESIGNS FOR STOCHASTIC ROOT-FINDING

V. Roshan Joseph[1], Yubin Tian[2], and C. F. Jeff Wu[1]

[1] *Georgia Institute of Technology and* [2] *Beijing Institute of Technology*

Abstract: The Robbins-Monro procedure (1951) for stochastic root-finding is a nonparametric approach. Wu (1985, 1986) has shown that the convergence of the sequential procedure can be greatly improved if we know the distribution of the response. Wu's approach assumes a parametric model and therefore its convergence rate slows down when the assumed model is different from the true model. This article proposes a new approach that is robust to the model assumptions. The approach gives more importance to observations closer to the root, which improves the fit to the true model around the root and makes the convergence faster. Simulation study shows that the new approach gives a superior performance over the existing methods.

Key words and phrases: Gaussian process, Robbins-Monro procedure, Sequential design, Stochastic approximation.

1. Introduction

Finding the root of a function is arguably the oldest and the most important problem in numerical mathematics. An interesting situation occurs when we do not know this function and can only observe the values of it with some error. This problem has numerous applications in science and engineering. For example, a control engineer will be interested to find the value of a control variable for maintaining some system response at a target value. The exact relationship between the control variable and the response may be unknown, but the response can be observed with some measurement noise. The problem becomes very complicated when the true relationship is highly nonlinear and the measurements are extremely noisy. Some other applications

of stochastic root-finding include the quantile estimation problem in bio-assay experiments (Finney (1978)), quality and reliability improvement (Joseph and Wu (2002)), sensitivity experiments (Neyer (1994)), and adaptive control and signal processing (Chen (2002), Kushner and Yin (1997), Benveniste, Métivier, and Priouret (1990)). A recent account of this subject is given by Spall (2003).

The problem can be formally stated as follows. Suppose we want to find the root (θ) of an unknown function $M(x)$. The experimenter can observe a random variable (Y) whose mean is $M(x)$. Thus, one can try to find the root numerically by observing Y's at some values of x. There are two ways to conduct the experiment, a sequential design (adaptive design) or a fixed design (non-adaptive design). In a fixed design the design points are chosen prior to the experiment, whereas in a sequential design they are chosen sequentially, i.e., x_{n+1} will be chosen based on $x_1, x_2, \cdots, x_n$ and $Y_1, Y_2, \cdots, Y_n$. Most often (particularly in nonlinear systems) the "optimal" x values depend on the distribution of Y, but very little is known about it before the experiment. Therefore a nonadaptive design can exhibit poor optimality properties, whereas a sequential design approach enables one to optimally select the design points. Therefore a sequential design is expected to outperform a fixed design.

One sequential design strategy known as stochastic approximation is to choose $x_1, x_2, \cdots$ such that $x_n \to \theta$ in probability. In a seminal paper, Robbins and Monro (1951) proposed the following method, which closely resembles the Newton-Raphson method for nonlinear root-finding. Start at some x_1 that is believed to be close to the root θ. Then generate the other design points sequentially using the following scheme:

$$x_{n+1} = x_n - a_n y_n, \tag{1.1}$$

where $\{a_n\}$ is a sequence of pre-specified constants. Assume that $M(x)$ is nondecreasing and the slope $\dot{M}(\theta) > 0$. Robbins & Monro proved that if the $\{a_n\}$ satisfies the conditions: $a_n > 0$, $\sum_{n=1}^{\infty} a_n = \infty$, and $\sum_{n=1}^{\infty} a_n^2 < \infty$, then $x_n \to \theta$,

in probability, as $n \to \infty$. For example $a_n = c/n$, where c is a positive constant, satisfies the above conditions. Based on the results of Chung (1954), Hodges and Lehmann (1956), and Sacks (1958), the procedure is fully asymptotically efficient with $a_n = 1/\{n\dot{M}(\theta)\}$. This clearly shows the difference between deterministic root-finding and stochastic root-finding problems. In the former, a constant sequence $a_n = 1/\dot{M}(\theta)$ would work, but in the latter, a decreasing sequence of constants at some particular rate is necessary to ensure the desired convergence. For practical implementation of the Robbins-Monro procedure some prior value of the slope is required. If a good prior value is not available, then the slope is estimated by using the least squares estimate $\sum(x_i - \bar{x}_n)y_i / \sum(x_i - \bar{x}_n)^2$. This is known as adaptive Robbins-Monro procedure, which under some truncation rule has the same asymptotic optimality properties as that of the Robbins-Monro procedure (see Anbar (1978), Lai and Robbins (1979) for details). Lai (2003) gives a recent review of this subject.

The Robbins-Monro procedure is a nonparametric procedure in the sense that the x_n converges to θ irrespective of the distribution of Y. Wu (1985, 1986) observed that the experimenters often know the distribution (such as normal or binomial) and therefore more efficient sequential procedures can be developed. The basic idea in Wu's approach is to approximate $M(x)$ by a parametric function $F(x|\gamma)$. Then, after observing the data $(x_1, y_1), \cdots, (x_n, y_n)$, the sequential procedure is to select x_{n+1} such that $F(x_{n+1}|\hat{\gamma}_n) = 0$, where $\hat{\gamma}_n$ is the maximum likelihood estimate (MLE) of γ. Ying and Wu (1997) showed that $x_n \to \theta$ almost surely irrespective of the functional form of $M(x)$. Wu (1985) has demonstrated in the case of binary data that the MLE-based sequential procedure performs much better than the Robbins-Monro procedure because of its efficient use of the complete set of data. This was also confirmed by Young and Easterling (1994) through extensive simulations. However, the MLE-based approach may loose its efficiency if F is not a good approximation to M. In this work we propose an adaptive design procedure based on a flexible Bayesian modeling, whose

performance is more robust to the deviations of F from M.

The article is organized as follows. In Section 2, assuming normal distribution for Y, we propose a modeling approach that takes into account of the uncertainties in the parametric part of the model. In Section 3 the issues related to estimation are considered. Due to some estimation problems, a fully Bayesian approach is proposed in Section 4. Extensions of the proposed approach to nonnormal distributions are considered in Section 5. The performance of the proposed approach is compared with the existing methods through simulations in Section 6 and the convergence is studied in Section 7. Some concluding remarks and future research directions are given in Section 8.

2. Modeling

Assume that Y follows a normal distribution. Extensions to other distributions will be considered in a later section. Let $Y = M(x) + e$, where $e \sim N(0, \sigma^2)$ and the function $M(x)$ is unknown but is assumed to be increasing in x. In Wu's MLE-based approach $M(x)$ is approximated by $\beta(x - \theta)$. With the above choice for the mean, Wu's approach reduces to the well-known iterated least squares procedure (Lai and Robbins, 1982). The true $M(x)$ can be nonlinear, in which case, the MLE-based approach may loose its efficiency. This is because the MLE approach assumes all the observations to be from the model $Y = \beta(x - \theta) + e$ and therefore gives equal weights to all observations. This can slow down the convergence of the MLE based approach. We propose a more flexible modeling that takes this uncertainty into account.

We assume $M(x)$ to be a random function with mean $\beta(x - \theta)$. This can be formulated using a Bayesian approach by putting a prior on $M(x)$. One approach to introduce randomness in the function is to let $M(x) = (\beta + \epsilon(x))(x - \theta)$, where $\epsilon(x)$ is a realization from a Gaussian process (GP). Such stochastic processes are widely used for modeling deterministic functions in computer experiments (Santner, Williams, and

Notz (2003)). Thus we have the model,

$$Y = (\beta + \epsilon(x))(x - \theta) + e, \;\; e \sim N(0, \sigma^2), \;\; \epsilon(x) \sim GP(0, \tau^2 R), \tag{2.1}$$

where the covariance function is defined as $cov(\epsilon(x_i), \epsilon(x_j)) = \tau^2 R(x_i, x_j)$. There are several choices for the correlation function R. The most popular one in computer experiments is the exponential correlation function given by $R_{ij} = R(x_i, x_j) = \exp(-\lambda |x_i - x_j|^p)$, where $\lambda > 0$ and $0 < p \leq 2$.

Note that $var\{M(x)\} = \tau^2 (x - \theta)^2$. Hence as $x \to \theta$, $var\{M(x)\} \to 0$. This is an important feature in our modeling. As the points get closer to θ, the variance approaches 0, and therefore in the estimation *more importance is given to the recent observations*. We also consider a special case of the Gaussian process, where the correlation between any two points is equal to 0. This leads to an independent process, which is easier to handle than a dependent process. Thus the model is given by

$$Y = (\beta + \epsilon(x))(x - \theta) + e, \;\; e \sim N(0, \sigma^2), \;\; \epsilon(x) \sim N(0, \tau^2), \tag{2.2}$$

and $cov(\epsilon(x_i), \epsilon(x_j)) = 0$ for $x_i \neq x_j$. To distinguish from (2.1), we will call (2.2) as independent error model and (2.1) as dependent error model.

3. Estimation

Suppose we have observed the data $(x_1, y_1), \cdots, (x_n, y_n)$. Let

$$\begin{aligned} y &= (y_1, \cdots, y_n)', \; \epsilon = (\epsilon(x_1), \cdots, \epsilon(x_n))', \; X = \begin{bmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_n \end{bmatrix}', \; \eta = \beta \begin{pmatrix} -\theta \\ 1 \end{pmatrix}, \\ T(\theta) &= diag\{x_1 - \theta, \cdots, x_n - \theta\}, \;\; \text{and} \;\; R = (R_{ij})_{n \times n}. \end{aligned}$$

The x's are generated sequentially, but fortunately the likelihood is not affected by the sequential design. Therefore we can obtain the likelihood as though the data are generated from a fixed design. Thus, the joint (or hierarchical) likelihood is given by

$$L_{joint} = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\{\frac{-1}{2\sigma^2}(y - X\eta - T(\theta)\epsilon)'(y - X\eta - T(\theta)\epsilon)\} \frac{\exp\{\frac{-1}{2\tau^2}\epsilon' R^{-1}\epsilon\}}{(2\pi\tau^2)^{n/2}|R|^{1/2}}. \tag{3.1}$$

For the moment assume that β, τ^2 and the parameters in the correlation function (λ and p) are known. Then, we can estimate θ, and $\epsilon = (\epsilon_1, \cdots, \epsilon_n)'$ by maximizing (3.1) and our sequential procedure will be to set x_{n+1} at the current estimate of θ.

Note that we do not require the values of $\epsilon_1, \cdots, \epsilon_n$ for the sequential procedure. Their presence makes the inference difficult and therefore we can treat them as nuisance parameters. It is well known that when the dimension of the nuisance parameters increases with n, the MLE's can become inconsistent. Hence it is desirable to eliminate the nuisance parameters in our problem. There are several approaches to tackle nuisance parameters (Severini (2000)), of which the integrated likelihood seems to be the most suitable for the present problem.

Integrating out ϵ from (3.1) we get (the proportionality constant is omitted)

$$L = \frac{1}{|\sigma^2 I + \tau^2 R(\theta)|^{1/2}} \exp[-\frac{1}{2}(y - X\eta)'\{\sigma^2 I + \tau^2 R(\theta)\}^{-1}(y - X\eta)],$$

where $R(\theta) = T(\theta)RT(\theta)$. Thus the MLE of θ can be obtained by minimizing

$$-2\log L = \log|\sigma^2 I + \tau^2 R(\theta)| + (y - X\eta)'\{\sigma^2 I + \tau^2 R(\theta)\}^{-1}(y - X\eta), \tag{3.2}$$

and our sequential procedure becomes

$$x_{n+1} = \hat{\theta}_n = arg\min_{\theta} -2\log L. \tag{3.3}$$

For the independent error model in (2.2), the objective function in the above minimization simplifies to

$$\sum_{i=1}^{n} \log\{\sigma^2 + \tau^2(x_i - \theta)^2\} + \sum_{i=1}^{n} \frac{\{y_i - \beta(x_i - \theta)\}^2}{\sigma^2 + \tau^2(x_i - \theta)^2}.$$

This can be compared with Wu's MLE-based approach. In his approach the MLE is obtained by minimizing $\sum_{i=1}^{n}\{y_i - \beta(x_i - \theta)\}^2$. Different from this, our approach uses weights equal to $\{\sigma^2 + \tau^2(x_i - \theta)^2\}^{-1}$ in the objective function. The weights increase

as x_i gets closer to θ giving more importance to observations closer to θ. This property makes the estimation in our approach more robust to the model misspecifications.

The minimization of (3.2) is complicated because of multiple local minima. It can be seen in the following extreme case. All the proofs are given in the Appendix.

PROPOSITION 1 *When $\sigma^2 = 0$, the function in (3.2) has at least $n + 1$ local minima with respect to θ.*

For example, when $n = 100$ we are faced with the minimization of a function with at least 101 local minima. Thus we have converted the simple problem of finding the root of a function to a very complex optimization problem! This method is therefore useful only when the cost of actually obtaining a new y is much higher than the computational cost, which is the case in most practical situations involving physical experiments. The optimization can be simplified as follows. Order the x's as $x_{(1)} < x_{(2)} < \cdots < x_{(n)}$. As shown in the proof of Proposition 1 that for the case of $\sigma^2 = 0$, $L = 0$ at all the design points and it has at least one local maximum in each of the intervals $(-\infty, x_{(1)}), (x_{(1)}, x_{(2)}), \cdots, (x_{(n)}, \infty)$. Finding the maximum in each of these intervals is easier and then one could get the global maximum. Because $-2 \log L$ is continuous in σ^2 a similar algorithm will work well even for the case of $\sigma^2 > 0$. The optimization can be further simplified by searching for the global minimum of $-2 \log L$ only in the intervals around x_n.

4. A Fully Bayesian Approach

So far we have assumed that β, τ^2, and the parameters in the correlation function to be known. In practice these values are not known. We may try to estimate these parameters also from the data. Suppose we use the Gaussian correlation function given by $R(x_i, x_j) = exp(-\lambda |x_i - x_j|^2)$, which gives sample paths that are infinitely differentiable. This is a good choice when $M(x)$ is very smooth. Thus we can minimize (3.2) with respect to the parameters θ, β, τ, and λ. It is reasonable to assume that σ is

known. If unknown, it can be easily estimated by collecting a sample of observations at any fixed x. The sequential procedure remains the same as $x_{n+1} = \hat{\theta}_n$.

It is well known that the data generated by stochastic approximation methods do not give much information about the slope parameter β. See, for example, Lai and Robbins (1979). Estimation of the correlation parameters is even more difficult. When the data do not give much information about the parameters it is important to use the prior information that we have about the parameters. Thus, using a fully Bayesian approach, many of the finite sample estimation problems can be mitigated.

Assume $\theta \sim N(x_1, \sigma_\theta^2), \beta \sim N(\beta_0, \sigma_\beta^2), \tau \sim Unif(\tau_l, \tau_u)$, and $\lambda \sim Unif(\lambda_l, \lambda_u)$. Other prior distributions may also be used. The posterior distribution (after integrating out ϵ's) is

$$f(\theta, \beta, \tau, \lambda | y) \propto \frac{e^{-\frac{1}{2}(y-X\eta)'\{\sigma^2 I + \tau^2 R(\theta)\}^{-1}(y-X\eta)}}{|\sigma^2 I + \tau^2 R(\theta)|^{1/2}} \, e^{\frac{-(\theta - x_1)^2}{2\sigma_\theta^2}} \, e^{\frac{-(\beta-\beta_0)^2}{2\sigma_\beta^2}} \, 1_{[\tau_l, \tau_u]}(\tau) 1_{[\lambda_l, \lambda_u]}(\lambda).$$

Finding the posterior mean of the parameters is difficult, whereas the maximum-a-posteriori (MAP) estimators can be easily computed. We can obtain the MAP estimators by minimizing

$$\log|\sigma^2 I + \tau^2 R(\theta)| + (y - X\eta)'\{\sigma^2 I + \tau^2 R(\theta)\}^{-1}(y - X\eta) + \frac{(\theta - x_1)^2}{\sigma_\theta^2} + \frac{(\beta - \beta_0)^2}{\sigma_\beta^2}, \quad (4.1)$$

with respect to θ, β, τ, and λ subject to the conditions $\tau_l \le \tau \le \tau_u$ and $\lambda_l \le \lambda \le \lambda_u$. Note that in the case of an independent error model, there is no λ in the objective function.

Consider the following special cases:

1. $\tau = 0, \sigma_\beta = 0$: The sequential procedure based on (4.1) becomes

$$x_{n+1} = x_n - \frac{1}{(n + \frac{\sigma^2}{\beta_0^2 \sigma_\theta^2})\beta_0} y_n, \quad (4.2)$$

which is the same as the Robbins-Monro procedure in (1.1).

2. $\tau = 0$: The MAP estimates of θ and β can be obtained by minimizing

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}\{y_i - \beta(x_i - \theta)\}^2 + \frac{(\theta - x_1)^2}{\sigma_\theta^2} + \frac{(\beta - \beta_0)^2}{\sigma_\beta^2}. \tag{4.3}$$

We will call the resulting sequential procedure as Wu's MAP procedure because it reduces to Wu's (1986) MLE approach when $\sigma_\theta = \infty$ and $\sigma_\beta = \infty$.

Thus, the Robbins-Monro procedure and Wu's procedure are special cases of the proposed sequential procedure. Moreover, these special cases are obtained by putting some extreme values for the parameters, such as $\tau = 0$ and/or $\sigma_\beta = 0$, which may not be realistic. Therefore, by choosing more realistic values for these parameters in the proposed procedure, we can expect to see some improvement over these two existing procedures.

5. Non-Normal Distributions

The underlying distribution of the observations can be different from normal. For example, an explosive designer may be interested in finding the level of shock necessary to make 99.99% of the explosives fire (Neyer (1994)), in which case the data are binary and a Bernoulli distribution should be used. The Robbins-Monro procedure does not assume any distributions for Y and therefore it can be applied irrespective of the underlying distributions. Although the Robbins-Monro procedure, in this sense, is a nonparametric method, its efficiency can be greatly improved if we know the true distribution (see Joseph (2004) for the case of binary data). Wu (1985, 1986) has extended the MLE approach to generalized linear models, which is a very general and versatile approach. As described in Section 1, Wu assumes a parametric model for $M(x)$, say $F(x|\gamma)$, and uses $F(x|\hat{\gamma}_n)$ in place of $M(x)$ to determine the root. Ying and Wu (1997) showed that Wu's MLE-based sequential design generates points that converge to θ irrespective of the parametric function F. Although this is asymptotically valid, in finite samples the results can be seriously

affected by an improper choice of F. We can extend the approach in Section 2 to model the uncertainties in F and thereby developing a sequential design that is more robust to model uncertainties.

Suppose Y has some distribution with mean $M(x)$. We want to find θ such that $M(\theta) = \alpha$. Choose a monotonic function g such that the range of $g\{M(x)\}$ is in $(-\infty, \infty)$. Let $g\{M(x)\} = g(\alpha) + (\beta + \epsilon(x))(x - \theta)$, where $\epsilon(x) \sim GP(0, \tau^2 R)$. Now we can write down the posterior distribution, obtain the MAP estimate of θ, and get the sequential design. For example, consider the binary data. Here g could be logit or probit. Make the assumptions as in Section 4, then the posterior distribution becomes

$$\prod_{i=1}^{n}\{M(x_i)\}^{y_i}\{1 - M(x_i)\}^{1-y_i} \ \frac{\exp\{\frac{-1}{2\tau^2}\epsilon' R^{-1}\epsilon\}}{\tau^n |R|^{1/2}} \ e^{\frac{-(\theta - x_1)^2}{2\sigma_\theta^2}} \ e^{\frac{-(\beta-\beta_0)^2}{2\sigma_\beta^2}} \ 1_{[\tau_l,\tau_u]}(\tau) 1_{[\lambda_l,\lambda_u]}(\lambda),$$

where $M(x_i) = g^{-1}\{g(\alpha) + (\beta + \epsilon(x_i))(x_i - \theta)\}$. If $\hat{\theta}_n$ is the MAP estimate of θ, then the sequential design is $x_{n+1} = \hat{\theta}_n$. In general it is difficult to eliminate the nuisance parameters ϵ's as done in the case of normal distributions. Overall, the estimation problem in non-normal distributions is much more complex and we will leave the details as a topic for future research.

6. Simulations

In this section we will investigate the performance of the proposed procedure in (4.1) using simulations. It will be compared with the existing procedures such as the Robbins-Monro (RM) procedure in (4.2) and Wu's MAP procedure in (4.3).

Consider a nonlinear function $M(x) = e^x + 2x - 5$, whose root is 1.0587. Suppose $\sigma = 0.5$ and we start at $x_1 = 3$. To use the procedures in (4.1), (4.2), and (4.3), we need to select the necessary prior parameters. Let $\sigma_\theta = 1, \beta_0 = 6, \sigma_\beta = 0.25\beta_0, \tau_l = 0, \tau_u = 10\sigma, \lambda_l = 0$, and $\lambda_u = 100$. Let $n = 10$, which means the best estimate of the root is x_{11}. Then 100 simulations were performed on the four procedures: proposed procedure based on the dependent error model, proposed procedure based on the independent error model, Wu's MAP, and the RM procedure. The recursions for a

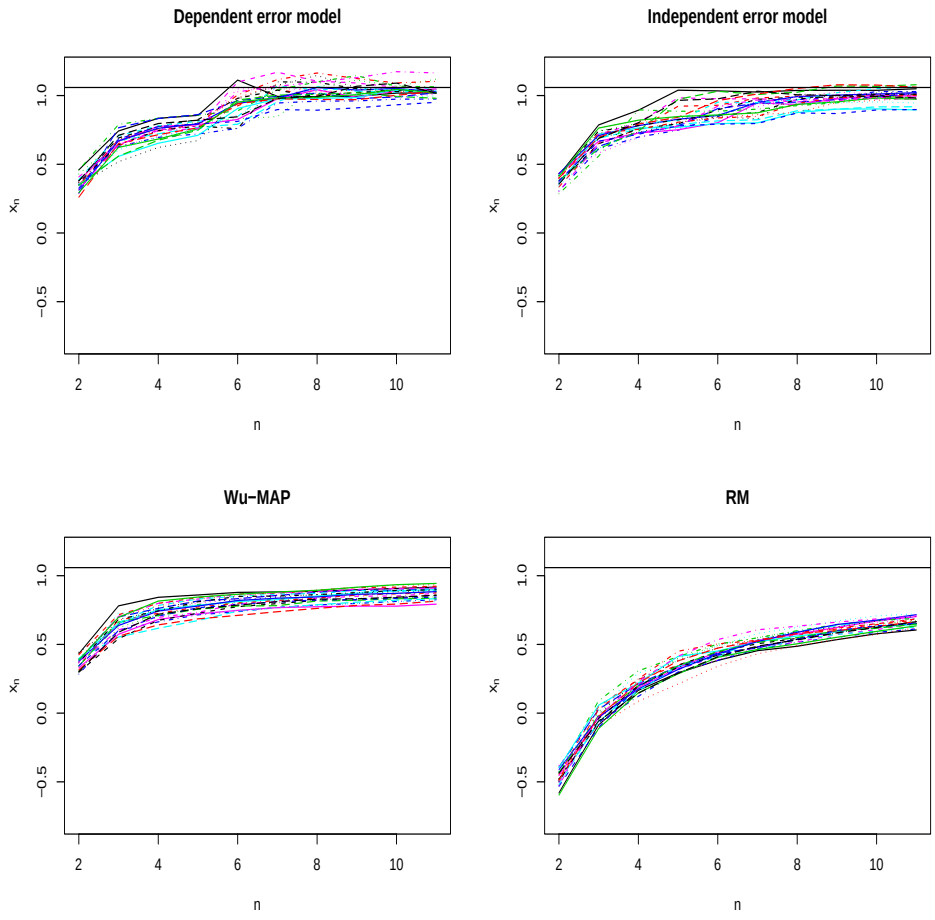

Figure 6.1: Simulation study. Recursions from x_2 to x_{11} for $M(x) = e^x + 2x - 5$.

Table 6.1: Test functions, Prior specifications, and MSE of x_{11}

		Prior		MSE			
$M(x)$	σ	x_1	β_0	Dep.	Indep.	Wu	RM
$e^x + 2x - 5$	0.5	3	6	.0026	.0029	.0271	.1527
$x^2 - 2$	0.05	2	2	.00004	.00004	.00031	.00030
$-0.4 + x + 0.2\sin(5x)$	0.05	-1	0.5	.0002	.0002	.0012	.0002
$e^{2x}/(1 + e^{2x}) - 0.9$	0.04	0	0.2	.0008	.0011	.0258	.0756

few of those simulations are shown in Figure 6.1. We see that both the proposed procedures outperform Wu's MAP procedure and the RM procedure. Note that the starting point $x_1 = 3$ is far away from the root $\theta = 1.0587$. Because Wu's MAP procedure gives equal weights to all observations, the convergence is very slow. The x_2 and x_3 of the dependent and independent error models are very similar to that of Wu's MAP. But because less weights are given to observations far from θ, the new procedures quickly "forget" about the starting point and converge to θ at a much faster rate. Three more functions were selected for simulations. The functions and the prior parameter values x_1 and β_0 are shown in Table 6.1. The other prior parameters are kept the same as before. The mean squared error (MSE) of x_{11} with respect to θ is computed from the simulations and are given in Table 6.1. We see that the two proposed procedures have smaller MSE values and thus perform better than the existing methods.

It is surprising that the performance of the independent error model is comparable to the more complicated dependent error model. Naturally one would expect the dependent error model to perform better, which is not seen here. Thus we conclude that using a dependent process for the error does not significantly improve the performance of the procedure. This phenomenon can be explained as follows. First, the

most important property underlying the performance of the new procedure is that the variance decrease as x converges to θ, which is shared by both the procedures. Second, stochastic approximation procedures produce very little information for estimating slope and correlation parameters, and therefore little is gained by using a dependent process. Thus, based on the simulation study, we recommend using the independent error model because of its simplicity.

We also need to check the sensitivity of the proposed procedure with respect to the prior specification. Each of the prior parameters is varied one at a time and the simulations are repeated. The MSE values for the function $M(x) = e^x + 2x - 5$ are plotted in Figure 6.2. We can see that the two proposed procedures are robust to the prior specification. One of the critical parameters is the starting point. We can see that the proposed procedures perform very well when x_1 is far away from the root ($\theta = 1.0587$). When x_1 is close to θ, the MSE values are very small, and therefore practically these procedures are not different. They become significantly different when x_1 is away from θ and in those cases the proposed procedures clearly produce superior performance. The slope parameter β_0 has a significant effect on both Wu's MAP and the RM procedures, whereas it does not affect the proposed procedures. The same conclusion can be drawn with respect to σ_θ and σ_β/β_0.

The prior specification is always the most difficult thing to do in any Bayesian procedures. We provide the following guidelines based on our experience. The starting point and the slope parameter β_0 should be chosen based on the prior knowledge. The specification of the other parameters seems to be less critical. The choice $\sigma_\beta = 0.25\beta_0$ seems to be reasonable. The parameter τ_u should be selected based on the knowledge of the function. If the function is expected to be highly nonlinear, then a large value should be chosen. Because the weights used in the procedure are inversely proportional to $\sigma^2/\tau^2 + (x_i - \theta)^2$, it is the ratio τ/σ that matters. The choice $\tau_u = 10\sigma$ worked well in the simulation study. One nice feature of the proposed procedures is that the

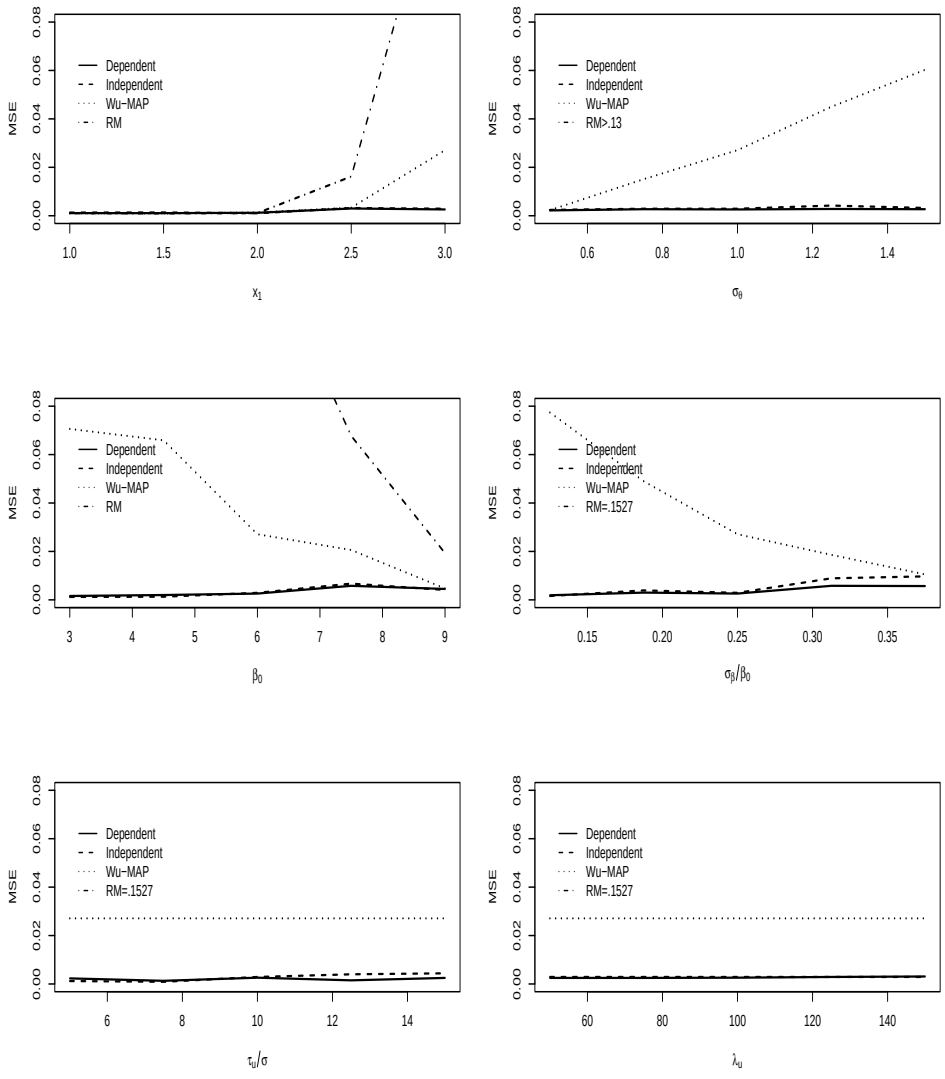

Figure 6.2: Simulation study. Mean squared error of the estimator of θ against the prior parameters, for the function $M(x) = e^x + 2x - 5$.

performance is not very sensitive to the prior specification. A reasonable prior should result in a good performance.

7. Convergence

In this section we study the convergence of the proposed sequential procedure. To make the mathematics tractable, we only study the independent error model. This will get rid of the covariance terms in the $R(\theta)$ matrix, thus simplifying the calculations. Moreover, the simulations in the previous section have indicated that the performance of the proposed procedure with independent process is as good as that of the dependent process. Therefore, consider the independent error model (2.2) with $R_{ij} = 0$ for $i \neq j$ and $R_{ii} = R(x_i, \theta, \lambda)$ (here we consider a more general form for the correlation function by allowing R_{ii} to depend on θ and λ).

The conditional density of Y_n given $y_1, \ldots, y_{n-1}$ is

$$f_{Y_n}(y_n | y_1 \ldots, y_{n-1})$$

$$= \frac{1}{\sqrt{2\pi}} |\sigma^2 + \tau^2 R_{nn}(x_n - \theta)^2|^{-1/2} exp\{-\frac{[y_n - \beta(x_n - \theta)]^2}{2[\sigma^2 + \tau^2 R_{nn}(x_n - \theta)^2]}\}. \tag{7.1}$$

Let the parameter be $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3, \theta_4) = (\beta, \gamma, \tau, \lambda)$, where $\gamma = \beta\theta$. Let the MLE based on $y_1, \ldots, y_n$ be $\hat{\boldsymbol{\theta}} = (\hat{\theta}_{n,1}, \hat{\theta}_{n,2}, \hat{\theta}_{n,3}, \hat{\theta}_{n,4}) = (\hat{\beta}_n, \hat{\gamma}_n, \hat{\tau}_n, \hat{\lambda}_n)$. To prove the consistency of $(\hat{\beta}_n, \hat{\gamma}_n)$, we extend the result of Datta (1997) in the following lemma. In order to state the lemma, define the following notations. Let $\Theta \subset \mathbb{R}^m$ be a bounded parameter space. Assume that the i^{th} experiment E_i is determined by the former $i-1$ observations $Y_1, \ldots, Y_{i-1}$ and that for $\boldsymbol{\theta} \in \Theta$, the i^{th} observation Y_i, given $Y_1, \ldots, Y_{i-1}$, has a density $f(y, e_i, \boldsymbol{\theta})$ with respect to some σ finite measure μ. Further, let the true value of the parameter be $\boldsymbol{\theta}_0$ and $P_{\boldsymbol{\theta}_0}$ denote the probability distribution governing $Y_1, Y_2, \ldots$ when $\boldsymbol{\theta} = \boldsymbol{\theta}_0$.

Lemma 1. Suppose the following three conditions hold:

(i) Given $\varepsilon > 0$, there exist $\eta(\varepsilon) > 0$ such that

$$P_{\boldsymbol{\theta}_0}\{\inf_{\boldsymbol{\theta}\in N^c_{l,\varepsilon}} n^{-1}\sum_{i=1}^{n}(k(E_i,\boldsymbol{\theta}_0)-k(E_i,\boldsymbol{\theta}))>\eta(\varepsilon)\}\to 1,\ \ as\ \ n\to\infty, \tag{7.2}$$

where $k(E_i,\boldsymbol{\theta}) = \int(logf(y,E_i,\boldsymbol{\theta}))f(y,E_i,\boldsymbol{\theta}_0)d\mu$ and $N^c_{l,\varepsilon}(\boldsymbol{\theta}_0) = \{\boldsymbol{\theta} = (\theta_1,\ldots,\theta_m) : (\theta_1-\theta_{0,1})^2+\ldots+(\theta_l-\theta_{0,l})^2>\varepsilon\}$.

$$(ii)\ \limsup_{n\to\infty}\sup_{e_1,\ldots,e_n} n^{-1}\sum_{i=1}^{n}\int(\|f(y,e_i,\cdot)\|-M)_+f(y,e_i,\boldsymbol{\theta}_0)d\mu\to 0,\ \text{as } M\to\infty, \tag{7.3}$$

where $\|f(y,e_i,\cdot)\|$ is the sup norm about $\boldsymbol{\theta}$ for given y and e_i, and x_+=max(x,0) for $x\in R$.

$$(iii)\ \limsup_{n\to\infty}\sup_{e_1,\ldots,e_n,\boldsymbol{\theta}\in\Theta} n^{-1}\sum_{i=1}^{n}\int\sup_{\tilde{\boldsymbol{\theta}}\in N_\rho(\boldsymbol{\theta})}(|\ logf(y,e_i,\tilde{\boldsymbol{\theta}})-logf(y,e_i,\boldsymbol{\theta})\ |)f(y,e_i,\boldsymbol{\theta}_0)d\mu$$
$$\to 0,\ \text{as } \rho\to 0, \tag{7.4}$$

where $N_\rho(\boldsymbol{\theta}) = \{\tilde{\boldsymbol{\theta}} = (\tilde{\theta}_1,\ldots,\tilde{\theta}_m) : (\tilde{\theta}_1-\theta_1)^2+\ldots+(\tilde{\theta}_m-\theta_m)^2\le\rho\}$.
Then the component $(\hat{\theta}_{n,1},\ldots,\hat{\theta}_{n,l})$ of $(\hat{\theta}_{n,1},\ldots,\hat{\theta}_{n,m})$ which maximizes $\sum_{i=1}^n logf(y_i,e_i,\theta)$, i.e., the MLE $(\hat{\theta}_{n,1},\ldots,\hat{\theta}_{n,l})$ is consistent for $(\theta_{0,1},\ldots,\theta_{0,l})$ under $P_{\boldsymbol{\theta}_0}$, as $n\to\infty$.

In our problem, $E_i = x_i, i\ge 1$. Let $B_1 = \{(u,v) : u = 1, |v| < \delta_1 < 1\}$, $B_2 = \{(u,v) : u\ne 1, u>\delta_2>0, v=0\}$, and $B_3 = \{(u,v) : u\ne 1, v\ne 0, u>\delta_{31}> 0, h(u,v)>\delta_{32}>0\}$, where

$$h(u,v) = \frac{4v^2}{4v^2+(|1-u|\pm\sqrt{(1-u)^2+4v^2})^2}\{u-\frac{2v^2}{1-u}+\frac{2v^2+1-u}{1-u}\cdot\frac{(|1-u|\pm\sqrt{(1-u)^2+4v^2})^2}{4v^2}\} \tag{7.5}$$

We now show that under some conditions the MLE $(\hat{\beta}_n,\hat{\gamma}_n)$ is consistent.

Theorem 1. Assume that x take values in a bounded subset $\mathcal{E}$ of $\mathbb{R}^1$ and the parameter space Θ is a bounded subset of $\mathbb{R}^4$, for which $0<d<\beta$. Assume also that for all $x\in\mathcal{E}$ and $\boldsymbol{\theta}\in\Theta$, $R(x,\theta,\lambda)(x-\theta)^2$, $\frac{d}{d\theta}[R(x,\theta,\lambda)(x-\theta)^2]$ and $\frac{d}{d\lambda}[R(x,\theta,\lambda)(x-\theta)^2]$ have upper bounds. Then we have:

(1) If the experiments $x_1, \ldots, x_i, \ldots$ satisfy

$$P_{\boldsymbol{\theta}_0}\{(n^{-1}\sum_{i=1}^{n} x_i^2, n^{-1}\sum_{i=1}^{n} x_i) \in B_1 \cup B_2 \cup B_3\} \to 1 \ \ as \ \ n \to \infty,$$

condition (i) of Lemma 1 holds with $l = 2$ and $m = 4$.

(2) Conditions (ii) and (iii) of Lemma 1 hold.

By applying Lemma 1, we prove the consistency of the MLE $(\hat{\beta}_n, \hat{\gamma}_n)$ under the assumptions of Theorem 1. Therefore, $\hat{\theta}_n = \hat{\gamma}_n/\hat{\beta}_n$ is also consistent. For the MAP estimator of (β, γ), we add a term $f_0(y, \theta) = \pi(\boldsymbol{\theta})g(y)$ for $i = 0$ in Lemma 1 and Theorem 1, where $\pi(\boldsymbol{\theta})$ is the prior density for $\boldsymbol{\theta}$ and $g(y)$ is a positive and integral function about the σ finite measure μ. Since $\pi(\boldsymbol{\theta})$ in Section 4 is bounded for $\boldsymbol{\theta} \in \Theta$, conditions $(i) \sim (iii)$ of Lemma 1 are all valid under the assumptions of Theorem 1. Then the MAP estimator for (β, γ), i.e., the component $(\hat{\theta}_{n,1}, \hat{\theta}_{n,2})$ of $(\hat{\theta}_{n,1}, \ldots, \hat{\theta}_{n,4})$ which maximizes $\sum_{i=0}^{n} logf(y_i, e_i, \theta)$, is consistent.

The assumptions of the theorem are mild. We can make some truncation on $x_j, j \geq 1$, so that $(n^{-1}\sum_{i=1}^{n} x_i^2, n^{-1}\sum_{i=1}^{n} x_i) \in B_1 \cup B_2 \cup B_3$ always hold. Then for a proper function $R(x, \theta, \lambda)$ the assumptions of Theorem 1 are satisfied.

8. Conclusions

Wu's MLE approach to stochastic root-finding has a drawback that, if the assumed parametric model is different from the true model, then the convergence of the procedure becomes slow. In this article we propose a new adaptive design to overcome this problem. This adaptive design automatically gives more weight to the observations closer to the root and therefore gives a better local fit to the true model around the root which makes the procedure converge faster irrespective of the model assumption. Two versions of the proposed approach namely, dependent error model and independent error model are discussed. Their superior performance over the Robbins-Monro procedure and Wu's MAP procedure is demonstrated through simulations.

The convergence for the sequential procedure is proved under some regularity

conditions. Simulations clearly show that the procedure is promising and can be considered for adoption in practice. Extensions of the approach to non-normal distributions are also discussed, although more work is needed for their practical implementation. This paper deals with only univariate functions. The Gaussian process modeling is known to perform well in higher dimensions and therefore the extension of this methodology to multivariate case will be a worthwhile topic for future research. Applications to stochastic optimization is also an interesting topic for research.

Acknowledgements

The research of Joseph and Wu was supported by the U.S. Army Research Laboratory and the U.S. Army Research Office under contract number W911NF-05-1-0264. The authors thank two referees for their valuable comments and suggestions.

Appendix

Proof of Proposition 1:

Let

$$a(t) = (y - X\eta)' R^{-1}(t)(y - X\eta) = \sum_{i=1}^{n}\sum_{j=1}^{n} \bar{r}_{ij} \frac{\{y_i - \beta(x_i - t)\}}{(x_i - t)} \frac{\{y_j - \beta(x_j - t)\}}{(x_j - t)},$$

where $\bar{r}_{ij} = (R^{-1})_{ij}$. We have that

$$L = \frac{1}{\tau^n |R|^{1/2}} \frac{1}{\prod_{i=1}^{n} |x_i - t|} \exp\{-\frac{a(t)}{2\tau^2}\}.$$

We have that $x_k \neq \theta$, otherwise the optimization is not necessary, and hence $y_k \neq 0$ for all $k = 1, \cdots, n$. Also since R is positive definite, $a(t) > 0$ for all t. Taking appropriate limits, we obtain $L = 0$ for $t \in \{x_1, \cdots, x_n, -\infty, \infty\}$. Also $L > 0$ for $t \notin \{x_1, \cdots, x_n, -\infty, \infty\}$ and L is a continuous function in t. Thus the result follows from Rolle's theorem. ♢

Proof of Lemma 1:

Let

$$Z_i(\theta) = logf(Y_i, E_i, \theta), i \geq 1, D_n(\theta) = n^{-1}\sum_{i=1}^{n} Z_i(\theta),$$

and

$$\tilde{D}_n(\theta) = n^{-1}\sum_{i=1}^{n}\int [logf(y, E_i, \theta)]f(y, E_i, \theta_0)d\mu = n^{-1}\sum_{i=1}^{n} k(E_i, \theta).$$

It is easy to see that the conditions of the L_1 law of large numbers (Datta 1997, Theorem 2.1) follow from (2) and (3) of Lemma 1. Therefore, by the same theorem,

$$\sup_{\theta} |D_n(\theta) - \tilde{D}_n(\theta)| \to 0$$

in P_{θ_0} probability.

For $\varepsilon > 0$, we have

$n^{-1}\sum_{i=1}^{n} k(E_i, \theta_0) - k(E_i, \hat{\theta}_n)$

$= \tilde{D}_n(\theta_0) - D_n(\theta_0) + D_n(\theta_0) - D_n(\hat{\theta}_n) + D_n(\hat{\theta}_n) - \tilde{D}_n(\hat{\theta}_n)$

$\leq \tilde{D}_n(\theta_0) - D_n(\theta_0) + D_n(\hat{\theta}_n) - \tilde{D}_n(\hat{\theta}_n) \leq 2\sup_{\theta} |D_n(\theta) - \tilde{D}_n(\theta)| < \eta(\varepsilon)$

with P_{θ_0} probability tending to one, as $n \to \infty$. Then, by condition (1) of Lemma 1, with probability P_{θ_0} tending to one, $\hat{\theta}_n$ is not in $N^c_{l,\varepsilon}(\theta_0)$, i.e., $(\hat{\theta}_{n,1} - \theta_{0,1})^2 + \ldots + (\hat{\theta}_{n,l} - \theta_{0,l})^2 \leq \varepsilon\}$.

Since $\varepsilon > 0$ is arbitrary, we obtain that $(\hat{\theta}_{n,1}, \ldots, \hat{\theta}_{n,l})$ is consistent. $\diamondsuit$

Proof of Theorem 1:

Let a be an upper bound for $R(x, \theta, \lambda)(x - \theta)^2$. Let x_i be the design point determined by the former $i - 1$ observations $y_1, \ldots, y_{i-1}$. Let $\mu = \beta x_i - \gamma$ and $\mu_0 = \beta_0 x_i - \gamma_0$. From the conditional density $f_{Y_i}(y_i|y_1 \ldots, y_{i-1})$, we have

$$k(x_i, \boldsymbol{\theta}) = \int (logf(y, x_i, \boldsymbol{\theta}))f(y, x_i, \boldsymbol{\theta}_0)dy$$

$$= -\frac{1}{2}[log(\sigma^2 + \tau^2 R_{ii}(\theta, \lambda)(x_i - \theta)^2] - \frac{1}{2}log(2\pi)$$

$$-\frac{\sigma^2+\tau_0^2R_{ii}(\theta_0,\lambda_0)(x_i-\theta_0)^2}{2[\sigma^2+\tau^2R_{ii}(\theta,\lambda)(x_i-\theta)^2]}-\frac{(\mu_0-\mu)^2}{2[\sigma^2+\tau^2R_{ii}(\theta,\lambda)(x_i-\theta)^2]}, \tag{A.1}$$

and

$$k(x_i,\boldsymbol{\theta}_0)=-\frac{1}{2}[log(\sigma^2+\tau_0^2R_{ii}(\theta_0,\lambda_0)(x_i-\theta_0)^2]-\frac{1}{2}log(2\pi)-\frac{1}{2}. \tag{A.2}$$

Then for $l=2$, $\boldsymbol{\theta}\in N^c_{2,\varepsilon}(\boldsymbol{\theta}_0)$ and by the fact that $log(x)+\frac{1}{x}$ has a minimum at $x=1$, we have from (A.1) and (A.2) that

$$\begin{aligned}
&n^{-1}\sum_{i=1}^n k(x_i,\boldsymbol{\theta}_0)-k(x_i,\boldsymbol{\theta})\\
&=n^{-1}\sum_{i=1}^n\{\frac{1}{2}[log\frac{\sigma^2+\tau^2R_{ii}(\theta,\lambda)(x_i-\theta)^2}{\sigma^2+\tau_0^2R_{ii}(\theta_0,\lambda_0)(x_i-\theta_0)^2}+\frac{\sigma^2+\tau_0^2R_{ii}(\theta_0,\lambda_0)(x_i-\theta_0)^2}{\sigma^2+\tau^2R_{ii}(\theta,\lambda)(x_i-\theta)^2}]\\
&\qquad\qquad -\frac{1}{2}+\frac{(\mu_0-\mu)^2}{2[\sigma^2+\tau^2R_{ii}(\theta,\lambda)(x_i-\theta)^2]}\}\\
&\geq n^{-1}\sum_{i=1}^n\frac{[(\beta-\beta_0)x_i-(\gamma-\gamma_0)]^2}{2(\sigma^2+\tau^2a)}\geq[2(\sigma^2+\tau^2a)]^{-1}\varepsilon^2n^{-1}\sum_{i=1}^n(x_i\cos\alpha-\sin\alpha)^2.
\end{aligned} \tag{A.3}$$

Now, consider $n^{-1}\sum_{i=1}^n(x_i\cos\alpha-\sin\alpha)^2$. When $n^{-1}\sum_{i=1}^n x_i^2=1$, its global minimum is $1-|n^{-1}\sum_{i=1}^n x_i|$; when $n^{-1}\sum_{i=1}^n x_i^2\neq 1, n^{-1}\sum_{i=1}^n x_i=0$, the minimum is $n^{-1}\sum_{i=1}^n x_i^2$ or 1; when $n^{-1}\sum_{i=1}^n x_i^2\neq 1, n^{-1}\sum_{i=1}^n x_i\neq 0$, the minimum is $h(n^{-1}\sum_{i=1}^n x_i^2, n^{-1}\sum_{i=1}^n x_i)$ or $n^{-1}\sum_{i=1}^n x_i^2$, where the function $h(u,v)$ is defined in (7.5). Then, under the conditions of this theorem, we have from (A.3) that there exist a positive constant δ such that

$$P_{\boldsymbol{\theta}_0}(\inf_{\boldsymbol{\theta}\in N^c_{2,\varepsilon}(\boldsymbol{\theta}_0)} n^{-1}\sum_{i=1}^n k(x_i,\boldsymbol{\theta}_0)-k(x_i,\boldsymbol{\theta})>\delta[2(\sigma^2+\tau^2a)]^{-1}\varepsilon^2)\to 1, as\ \ n\to\infty.$$

From the conditional density we also have

$$\begin{aligned}
&\int(\|f(y,x_i,\cdot)\|)-M)_+f(y,x_i,\boldsymbol{\theta}_0)dy\\
&=\frac{1}{2}\int(\|log(2\pi)+log(\sigma^2+\tau^2R_{ii}(\theta,\lambda)(x_i-\theta)^2)+\frac{(u-(\beta-\beta_0)x_i+(\gamma-\gamma_0))^2}{\sigma^2+\tau^2R_{ii}(\theta,\lambda)(x_i-\theta)^2}\|-2M)_+
\end{aligned}$$

$$\cdot\frac{exp\{-\frac{1}{2}u^2\cdot(\sigma^2+\tau_0^2R_{ii}(\theta_0,\lambda_0)(x_i-\theta_0)^2)^{-1}}{\sqrt{2\pi(\sigma^2+\tau_0^2R_{ii}(\theta_0,\lambda_0)(x_i-\theta_0)^2)}}du$$

$$\leq\frac{1}{2}\int(\|log(2\pi)+log(\sigma^2+c_1a)+\frac{(|u|+c_2)^2}{\sigma^2}\|)-2M)_+\cdot\frac{exp\{-\frac{1}{2}u^2\cdot(\sigma^2+\tau_0^2a)^{-1}}{\sqrt{2\pi\sigma^2}}du, \tag{A.4}$$

where c_1 and c_2 are positive constants. Then condition (ii) of Lemma holds from (A.4).

Additionally, for

$$g(\boldsymbol{\theta})=g(\theta_1,\theta_2,\theta_3,\theta_4)=g(\beta,\gamma,\tau,\lambda)$$
$$=log(\sigma^2+\tau^2R_{ii}(\theta,\lambda)(x_i-\theta)^2)+\frac{(u-\beta x_i+\gamma+\beta_0x_i-\gamma_0)^2}{\sigma^2+\tau^2R_{ii}(\theta,\lambda)(x_i-\theta)^2},$$

it is easy to see that there exist positive constants $a_{ij}, i=1,2,3$, and $j=1,2,3,4$, such that $\forall x_i\in\mathcal{E}$ and $\forall\boldsymbol{\theta}\in\Theta$,

$$|\frac{\partial}{\partial\theta_i}g(\boldsymbol{\theta})|\leq a_{i1}|u|^2+a_{i2}|u|+a_{i3}, i=1,\ldots,4. \tag{A.5}$$

Then we have

$$\int\sup_{\tilde{\boldsymbol{\theta}}\in N_\rho(\boldsymbol{\theta})}(|\,logf(y,x_i,\tilde{\boldsymbol{\theta}})-logf(y,x_i,\boldsymbol{\theta})\,|)f(y,x_i,\boldsymbol{\theta}_0)dy$$

$$=\int\sup_{\tilde{\boldsymbol{\theta}}\in N_\rho(\boldsymbol{\theta})}|\sum_{i=1}^4[\frac{\partial}{\partial\theta_i}g(\boldsymbol{\theta}+\xi(\tilde{\boldsymbol{\theta}}-\boldsymbol{\theta})](\tilde{\theta}_i-\theta_i)|\cdot\frac{exp\{-\frac{1}{2}u^2\cdot(\sigma^2+\tau_0^2R_{ii}(\theta_0,\lambda_0)(x_i-\theta_0)^2)^{-1}\}}{\sqrt{2\pi(\sigma^2+\tau_0^2R_{ii}(\theta_0,\lambda_0)(x_i-\theta_0)^2}}du$$

$$\leq\rho\sum_{i=1}^4\int(a_{i1}|u|^2+a_{i2}|u|+a_{i3})\cdot\frac{exp\{-\frac{1}{2}u^2\cdot(\sigma^2+\tau_0^2a)^{-1}}{\sqrt{2\pi\sigma^2}}du, \tag{A.6}$$

where $\tilde{\boldsymbol{\theta}}=(\tilde{\theta}_1,\tilde{\theta}_2,\tilde{\theta}_3,\tilde{\theta}_4)$ and $|\xi|\leq1$. From (A.6), we get condition (iii) of Lemma 1.

♢

References

Anbar, D. (1978). A stochastic Newton-Raphson method. *J. Statist. Planning and Inference* **2**, 153-163.

Benvensite, A., Métivier, M., and Priouret, P. (1990). *Adaptive Algorithms and Stochastic Approximations*. Berlin: Springer-Verlag.

Chen, H. F. (2002). *Stochastic Approximation and Its Applications*. Dordrecht: Kluwer Academic Publishers.

Chung, K. L. (1954). On a stochastic approximation method. *Ann. Math. Statist.* **25**, 463-483.

Datta, S. (1997). A uniform L_1 law of large numbers for functions on a totally bounded metric space. *Sankhya: Series A* **59**, 167-174.

Finney, D. J. (1978). *Statistical Methods in Biological Assay*. London: Griffin.

Hodges, J. L. and Lehmann, E. L. (1956). Two approximations to the Robbins-Monro process. *Proc. Third Berkeley Symp.* **1**, Ed. J. Neyman , 39-55. Berkeley, CA: University of California.

Joseph, V. R. and Wu, C. F. J. (2002). Operating window experiments: a novel approach to quality improvement. *J. Qual. Technol.* **34**, 345-354.

Joseph, V. R. (2004). Efficient Robbins-Monro procedure for binary data. *Biometrika* **91**, 461-470

Kushner, H. J. and Yin, G. G. (1997). *Stochastic Approximation Algorithms and Applications*. New York: Springer.

Lai, T. L. (2003). Stochastic approximation. *Ann. Statist.* **31**, 391-406.

Lai, T. L. and Robbins, H. (1979). Adaptive design and stochastic approximation. *Ann. Statist.* **7**, 1196-1221.

Lai, T. L. and Robbins, H. (1982). Iterated least squares in multi-period control. *Advances in Applied Mathematics* **3**, 50-73.

Neyer, B. T. (1994). D-optimality-based sensitivity test. *Technometrics* **36**, 61-70.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *Ann. Math. Statist.* **29**, 373-405.

Sacks, J. (1958). Asymptotic distribution of stochastic approximation procedures. *Ann. Math. Statist.* **29**, 373-405.

Santner, T. J., Williams, B. J., and Notz, W. I. (2003). *Design and Analysis of Computer Experiments*. New York: Springer.

Severini, T. A. (2000). *Likelihood Methods in Statistics*. New York: Oxford.

Spall, J. C. (2003). *Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control*. New Jersey: Wiley.

Wu, C. F. J. (1985). Efficient sequential designs with binary data. *J. Am. Statist. Assoc.* **80**, 974-984.

Wu, C. F. J. (1986). Maximum likelihood recursion and stochastic approximation in sequential designs. *Statistical Procedures and Related Topics* (J. Van Ryzin, ed.), IMS Monograph Series **8**, 298-313.

Ying, Z. and Wu, C. F. J. (1997). An asymptotic theory of sequential designs based on maximum likelihood recursions. *Statistica Sinica* **7**, 75-91.

Young, L. J. and Easterling, R. G. (1994). Estimation of extreme quantiles based on sensitivity tests: a comparative study. *Technometrics* **36**, 48-60.

School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332-0205, USA

Department of Mathematics, Beijing Institute of Technology, Beijing 100081, PRC

E-mail: (roshan@isye.gatech.edu, tyubin@2911.net, jeffwu@isye.gatech.edu)