

# Adaptive Designs for Stochastic Root-Finding

V. Roshan Joseph

School of Industrial and Systems Engineering

Georgia Institute of Technology

Atlanta, GA 30332-0205, USA

roshan@isye.gatech.edu

## SUMMARY

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 very different from the true model. This article proposes a new approach that is robust to the model assumptions. The approach utilizes a pinned Gaussian process that 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.

*Some key words:* 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 lots of 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 subject to large noise. 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 ( $\theta$ ) 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 (nonadaptive 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  $x_1, x_2, \dots, x_n$  and  $Y_1, Y_2, \dots, Y_n$ . Most often (particularly in nonlinear

systems) the “optimal”  $x$ -values depend on the distribution of  $Y$  and 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 out perform a fixed design.

One sequential design strategy known as stochastic approximation is to choose  $x_1, x_2, \dots$  such that  $x_n \rightarrow \theta$  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  $\theta$ . Then generate the other design points sequentially using the following scheme:

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

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  $x_n \rightarrow \theta$ , in probability, as  $n \rightarrow \infty$ . 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  $\theta$  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  $(x_1, y_1), \dots, (x_n, y_n)$ , 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  $\gamma$ . Wu (1985) and Ying and Wu (1997) showed that the  $x_n \rightarrow \theta$  almost surely irrespective of the functional form of  $M(x)$ . Wu (1985) has demonstrated in the case of binary data that these maximum likelihood based sequential procedures performs much better than the Robbins-Monro procedure because of its efficient use of the complete set of data. See also the simulation results by Young and Easterling (1994). However, the MLE based approach may lose its efficiency if  $F$  is not a good approximation to  $M$ . In this work we propose a flexible modeling using Bayesian methods that is robust to the deviations of  $F$  from  $M$ .

The article is organized as follows. In Section 2, assuming normal distribution, we propose two different modeling approaches 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. Ex-

tensions 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. Some concluding remarks and future research directions are given in Section 7.

## 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. If that happens, the MLE-based approach may lose 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 may not be a good choice as this model holds only locally around  $x = \theta$ . This can slow down the convergence rate 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 nicely by using a Bayesian approach by putting a prior on  $M(x)$ . The randomness can be introduced by putting an additive error  $M(x) = \beta(x - \theta) + \epsilon$  or by putting an error in the slope  $M(x) = (\beta + \epsilon)(x - \theta)$ , where  $\epsilon \sim N(0, \tau^2)$ . If  $M(x)$  is assumed to be a continuous function of  $x$ , then  $\epsilon$ 's at different values of  $x$  has to be *correlated* with correlation approaching one as

the points get closer. This can be achieved by treating  $\epsilon(x)$  as a realization from a Gaussian process with continuous sample paths. These infinite dimensional processes have been used in Bayesian regression by Blight and Ott (1975) and are closely related to the kriging method that is very popular in spatial statistics and computer experiments (see, for example, Santer, Williams, and Notz, 2003).

## 2.1 Additive Error Model

We have the following hierarchical model,

$$Y = M(x) + e, \quad e \sim N(0, \sigma^2), \quad M(x) = \beta(x - \theta) + \epsilon(x),$$

where  $\epsilon(x)$  follows a Gaussian process with mean 0 and covariance function  $\tau^2 r$ . There are several choices for  $r$ . The exponential correlation function is given by  $r(x_i, x_j) = \exp(-\lambda|x_i - x_j|^p)$ , where  $\lambda \geq 0$  and  $0 < p \leq 2$ . Assuming a solution exists, we should have  $M(\theta) = 0$ . Therefore we need  $\epsilon(\theta) = 0$ . To achieve this let  $\epsilon(x) = \delta(x) - \delta(\theta)$ , where  $\delta(x)$  follows a Gaussian process with mean 0 and covariance function  $\tau^2 r/2$ . Then  $\epsilon(x)$  follows a *pinned Gaussian process* with mean 0 and covariance function

$$\text{cov}\{\epsilon(x_i), \epsilon(x_j)\} = \frac{\tau^2}{2} \{1 - r(x_i, \theta) - r(x_j, \theta) + r(x_i, x_j)\}. \quad (2)$$

The variance is given by  $\text{Var}\{\epsilon(x)\} = \tau^2 \{1 - r(x, \theta)\}$ . Note that as  $x \rightarrow \theta$ ,  $\text{Var}\{\epsilon(x)\} \rightarrow 0$ . This is an important feature in our modeling. As the points get closer to  $\theta$ , the variance approaches 0, and therefore in the estimation *more importance is given to the recent observations*. Succinctly, the model can be written as

$$Y = \beta(x - \theta) + \epsilon(x) + e, \quad e \sim N(0, \sigma^2), \quad \epsilon(x) \sim GP\{0, \tau^2 R(\theta)\}, \quad (3)$$

where  $R_{ij}(\theta) = \frac{1}{2}\{1 - r(x_i, \theta) - r(x_j, \theta) + r(x_i, x_j)\}$ .

## 2.2 Slope Error Model

Using Taylor's theorem, the  $\epsilon(x)$  in (3) can be expanded as  $\epsilon(x) = \dot{\epsilon}(x^*)(x - \theta)$ , where  $x^*$  is an intermediate value. Thus the model in (3) becomes  $Y = \{\beta + \dot{\epsilon}(x^*)\}(x - \theta) + e$ . This suggests the following model,

$$Y = \{\beta + \epsilon(x)\}(x - \theta) + e, \quad e \sim N(0, \sigma^2), \quad \epsilon(x) \sim GP(0, \tau^2 R), \quad (4)$$

where  $R_{ij} = r(x_i, x_j)$ . Note that the  $\epsilon(x)$  in (3) and (4) are different. Both are used for modeling the nonlinearity in  $M(x)$ , but in different ways. For the  $\epsilon(x)$  it is possible to impose a covariance function as in (2), in which case we can interpret  $\beta$  as  $\dot{M}(\theta)$ .

## 3. ESTIMATION

First consider the additive error model. Suppose we have observed the data  $(x_1, y_1), \dots, (x_n, y_n)$ .

Let

$$X = \begin{bmatrix} 1 & x_1 \\ \cdot & \cdot \\ 1 & x_n \end{bmatrix}, \quad y = (y_1, \dots, y_n)', \quad \epsilon = (\epsilon(x_1), \dots, \epsilon(x_n))', \quad \eta = \beta \begin{pmatrix} -\theta \\ 1 \end{pmatrix}, \quad R(\theta) = (R_{ij}(\theta))_{n \times n}.$$

The  $x$ 's are generated sequentially. So let  $x_k = x_k(x_1, \dots, x_{k-1}, y_1, \dots, y_{k-1})$ . Therefore although  $x_1$  is a constant  $x_2, \dots, x_n$  are random variables because of their dependence on data. But fortunately the likelihood is not affected by the sequential design. Therefore by assuming the validity of strong likelihood principle we can ignore the sequential nature of

the experiment and obtain the MLE's as though the data is generated from a fixed design. It is pointed out that although the MLE's remain the same their sampling distributions are affected by the sequential design (Woodroffe, 1991). Thus the joint (or hierarchical) likelihood for the additive error model is given by

$$L_{joint} = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{\frac{-1}{2\sigma^2}(y - X\eta - I\epsilon)'(y - X\eta - I\epsilon)\right\} \frac{1}{(2\pi\tau^2)^{n/2}|R(\theta)|^{1/2}} \exp\left\{\frac{-1}{2\tau^2}\epsilon'R^{-1}(\theta)\epsilon\right\}. \quad (5)$$

For the moment assume that  $\beta$ ,  $\tau^2$  and the parameters in the correlation function ( $\lambda$  and  $p$ ) are known. Then we can estimate  $\theta$ , and  $\epsilon = (\epsilon_1, \dots, \epsilon_n)'$  by maximizing (5). Then our sequential procedure will be to set  $x_{n+1}$  at the current estimate of  $\theta$ .

Note that we do not require the values of  $\epsilon_1, \dots, \epsilon_n$  for the sequential procedure. Therefore we can treat them as nuisance parameters. Their presence makes the inference difficult. 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 (see Severini, 2000), of which the integrated likelihood seems to be the most suitable for the present problem.

Integrating out  $\epsilon$  from (5) we get (the proportionality constant is omitted)

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

Thus the MLE of  $\theta$  can be obtained by minimising

$$-2 \log L = \log |\sigma^2 I + \tau^2 R(\theta)| + (y - X\eta)' \{\sigma^2 I + \tau^2 R(\theta)\}^{-1} (y - X\eta). \quad (6)$$

Thus our sequential procedure becomes

$$x_{n+1} = \hat{\theta}_n = \arg \min_{\theta} -2 \log L. \quad (7)$$

Similar calculations can be done for the slope error model (4) to obtain the sequential procedure. Let  $T(\theta) = \text{diag}\{x_1 - \theta, \dots, x_n - \theta\}$  and  $R = (R_{ij})_{n \times n}$ . Then the MLE of  $\theta$  can be obtained by minimising  $-2 \log L$  in (6) with  $R(\theta) = T(\theta)RT(\theta)$ .

The minimisation of (6) is complicated because of multiple local minima. It can be seen in the following extreme case.

**PROPOSITION 1** *When  $\sigma^2 = 0$ , the function in (6) has at least  $n + 1$  local minima with respect to  $\theta$ .*

*Proof:* Consider the slope error model. 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

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

We have that  $x_k \neq \theta$ , otherwise the optimization is not necessary, and hence  $y_k \neq 0$  for all  $k = 1, \dots, n$ . Also since  $R$  is positive definite,  $a(t) > 0$  for all  $t$ . Taking appropriate limits, we obtain  $L = 0$  for  $t \in \{x_1, \dots, x_n, -\infty, \infty\}$ . But  $L > 0$  for  $t \notin \{x_1, \dots, x_n, -\infty, \infty\}$  and  $L$  is a continuous function in  $t$ . Thus the result follows from Rolle's theorem. The proof for the additive error model can be done similarly using an eigen value decomposition of  $R(t)$ .  $\diamond$

For example, when  $n = 100$  we are faced with the minimisation 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  $x_{(1)} < x_{(2)} < \dots < x_{(n)}$ . As shown in the proof of proposition 1 that for the case  $\sigma^2 = 0$ ,  $L = 0$  at all the design points and it has at least one local maximum in each of the intervals  $(-\infty, x_{(1)})$ ,  $(x_{(1)}, x_{(2)})$ ,  $\dots$ ,  $(x_{(n)}, \infty)$ . 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  $\sigma^2$  a similar algorithm will work well even for the case  $\sigma^2 > 0$ . The optimization can be further simplified by searching for the global minimum of  $-2 \log L$  only in the intervals around  $x_n$ .

One could argue that there is some loss of information caused by eliminating the nuisance parameters  $\epsilon_1, \dots, \epsilon_n$  from the model and therefore the sequential procedure based on (5) should work better than the one in (7). Clearly this is the case when  $\sigma^2 = 0$ , because the sequential procedure based on (5) will use an interpolating spline which will give a much better fit to the true model than a straight line fit. But when  $\sigma^2 > 0$  and large, an interpolating spline will be a poor fit to the true model and therefore the procedure in (7) could outperform it. Also as  $n \rightarrow \infty$ ,  $R(\theta)$  becomes ill-conditioned and therefore the procedure in (7) is a more numerically stable algorithm than the one based on (5).

## 4. A FULLY BAYESIAN APPROACH

So far we have assumed that  $\beta$ ,  $\tau^2$ , and the parameters in the correlation function to be known. In real 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 minimise (6) with respect to the parameters  $\theta, \beta, \tau$ , and  $\lambda$ . It is reasonable to assume that  $\sigma$  is known. 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 a lot of information about the slope parameter  $\beta$ . The information in the least squares estimator of  $\beta$  is proportional to  $\sum_{i=1}^n (x_i - \theta)^2$ , which increases very slowly with  $n$  because the  $x$ 's lie close to  $\theta$ . For example, if we use the Robbins-Monro procedure in (1), then the least squares estimate of the slope converges to  $\dot{M}(\theta)$  at the rate of  $(\log n)^{\frac{1}{2}}$  (See, Lai and Robbins, 1979). The consistency of  $\hat{\beta}$  in Wu's MLE procedure was proved by Frees and Ruppert (1990), but the estimates can be biased in small samples. See Coad and Woodroffe (1998). In our procedure less weights are given to the observations far away from  $\theta$ . This makes the estimation of  $\beta$  much harder and the procedure may not even produce consistent estimates. Although we do not need a consistent estimate of  $\beta$  for the stochastic approximation method to work, inaccurate estimates can slow down the convergence rate. Some of these estimation problems in finite samples can be mitigated by adopting a fully Bayesian approach by putting prior distributions for all the unknown parameters.

Assume

$$\theta \sim N(x_1, \sigma_\theta^2), \quad \beta \sim N(\beta_0, \sigma_\beta^2), \quad \tau \sim \text{Uniform}(\tau_l, \tau_u), \quad \lambda \sim \text{Uniform}(\lambda_l, \lambda_u).$$

Other prior distributions may also be used. The posterior distribution (after integrating out  $\epsilon$ 's) is

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

Finding the posterior mean of the parameters is very hard. The maximum-a-posteriori (MAP) estimators are much easier to compute. We can obtain the MAP estimators by minimising

$$\log |\sigma^2 I + \tau^2 R(\theta)| + (y - X\eta)'\{\sigma^2 I + \tau^2 R(\theta)\}^{-1}(y - X\eta) + \frac{(\theta - x_1)^2}{\sigma_\theta^2} + \frac{(\beta - \beta_0)^2}{\sigma_\beta^2}, \quad (8)$$

with respect to  $\theta, \beta, \tau$ , and  $\lambda$  subject to the conditions  $\tau_l \leq \tau \leq \tau_u$  and  $\lambda_l \leq \lambda \leq \lambda_u$ .

Consider the following special cases.

1.  $\tau = 0, \sigma_\beta = 0$ : The sequential procedure based on (8) becomes

$$x_{n+1} = x_n - \frac{1}{(n + \frac{\sigma^2}{\beta_0^2 \sigma_\theta^2})\beta_0} y_n, \quad (9)$$

which is the same as the Robbins-Monro procedure in (1).

2.  $\tau = 0$ : The MAP estimates of  $\theta$  and  $\beta$  can be obtained by minimising

$$\frac{1}{\sigma^2} \sum_{i=1}^n \{y_i - \beta(x_i - \theta)\}^2 + \frac{(\theta - x_1)^2}{\sigma_\theta^2} + \frac{(\beta - \beta_0)^2}{\sigma_\beta^2}. \quad (10)$$

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  $\sigma_\beta = \infty$ .

## 5. NON-NORMAL DISTRIBUTIONS

The underlying distribution of the observations can be different from normal. For example, a quality engineer will be interested to find the level of the stack force necessary to make 0.1% of multifeeds in the paper feeder of a copier machine (Joseph and Wu, 2002), in which case the the data are binary and we have to use a Bernoulli distribution. The Robbins-Monro procedure does not assume any distributions for  $Y$  and therefore it can be applied irrespective of the underlying distributions. Although Robbins-Monro procedure, in this sense, is a nonparametric method, its efficiency can be greatly improved if we know the true distribution (See, Joseph, 2003, for the case of binary data). Wu (1985, 1986) has extended the MLE approach to generalized linear models also, 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  $\theta$  irrespective of the parametric function  $F$ . Although this is the case, 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  $\theta$  such that  $M(\theta) = \alpha$ . Choose a monotonic function  $g$  such that the range of  $g \circ M$  is in  $(-\infty, \infty)$ . Consider the additive error model as in Section 2.1. Let  $g\{M(x)\} = g(\alpha) + \beta(x - \theta) + \epsilon(x)$ , where  $\epsilon(x) \sim GP\{0, \tau^2 R(\theta)\}$ . Now we can write down the posterior distribution, obtain the MAP

estimate of  $\theta$ , 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

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

where  $M(x_i) = g^{-1}\{g(\alpha) + \beta(x_i - \theta) + \epsilon(x_i)\}$ . If  $\hat{\theta}_n$  is the MAP estimate of  $\theta$ , then the sequential design is  $x_{n+1} = \hat{\theta}_n$ .

In general it is difficult to eliminate the nuisance parameters  $\epsilon$ 's as done in the case of normal distributions. Some simplifications can be done for the exponential family of distributions using generalized linear mixed models (GLMM). Although there is a huge literature on GLMM ( See, for example, McCulloch and Searle, 2001), the available techniques are not directly applicable to our case because of the dependence of variance on  $\theta$ . Overall, the estimation problem in non-normal distributions is much more complex and we will leave this as a topic for future research.

## 6. SIMULATIONS

In this section we will investigate the convergence properties of the proposed method in (8) using simulations. It will be compared with the existing procedures such as the Robbins-Monro procedure in (9) and Wu's MAP procedure in (10).

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 methods in (8), (9), and (10), we need to select the necessary prior parameters. Let  $\sigma_\theta = 1, \beta_0 = 6, \sigma_\beta = 1, \tau_l = 0, \tau_u = 5, \lambda_l = 0$ , and  $\lambda_u = 100$ .

$\tau_u$  for the slope error model is taken as half of that of the additive error model. A smaller value of  $\tau_u$  should be used for the slope error model to avoid negative values of the slope (assuming the function to be increasing). Let  $n = 10$ , which means, the best estimate of the root is  $x_{11}$ . 100 simulations were performed on the four methods: the additive error model, slope error model, Wu's MAP, and the Robbins-Monro procedure. The recursions for a few of those simulations are shown in Figure 1. We see that both the additive error model and the slope error model outperform the Wu's MAP procedure and the Robbins-Monro 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 additive and slope error models are very similar to that of Wu's MAP. But because less weights are given to observations far from  $\theta$ , the procedure quickly "forgets" about the starting point and converges to  $\theta$  at a much faster rate. Three more functions were selected for simulations. The functions and the prior parameter values are shown in Table 1. The mean squared error (MSE) of  $x_{11}$  with respect to  $\theta$  is computed from the simulations and are given in Table 2. We see that the two proposed methods have smaller MSE values and therefore they perform better than the existing methods.

Among the prior parameters, the starting point is the most critical one determining the performance of the proposed method. The MSE values for the four methods with different starting values are plotted for the function  $M(x) = e^x + 2x - 5$  and are given in Figure 2. We see that the two proposed methods perform very well when  $x_1$  is far away from the root ( $\theta = 1.0587$ ). When  $x_1$  is close to  $\theta$ , the MSE values are very small, and therefore practically these methods are not different. They become significantly different when  $x_1$  is away from  $\theta$

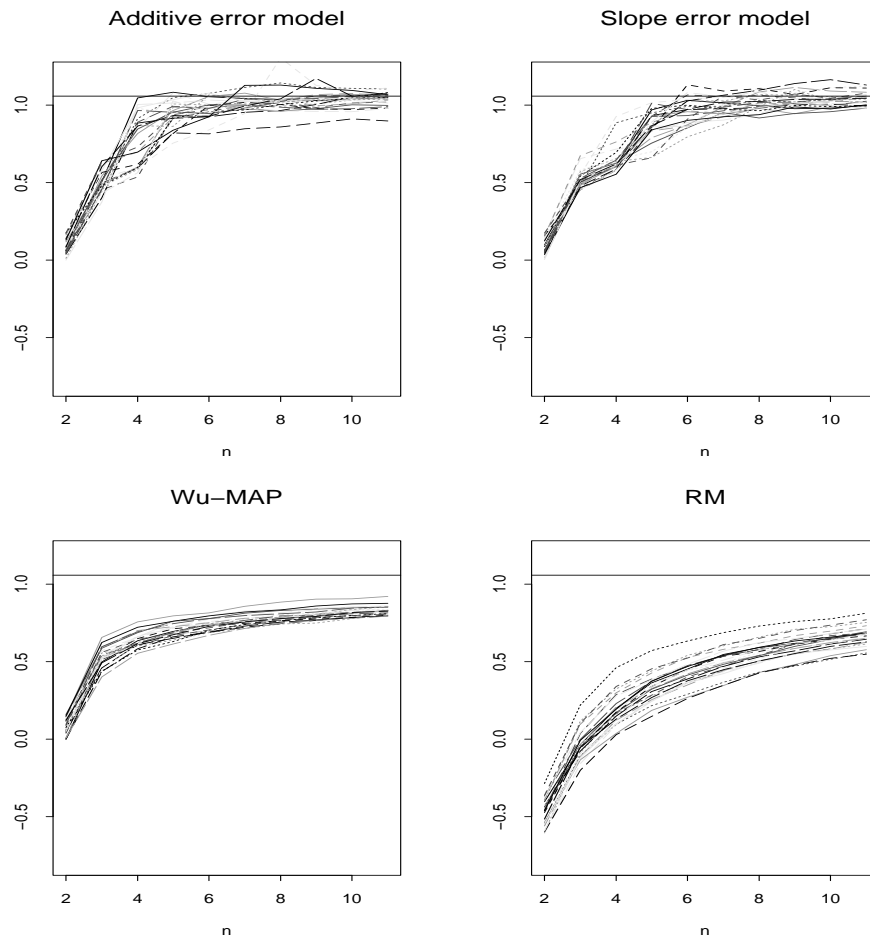


Figure 1: Simulation study. Recursions from  $x_2$  to  $x_{11}$  for  $M(x) = e^x + 2x - 5$ .

and in those cases the proposed methods clearly produce superior performance.

Table 1: Functions and prior parameters for the simulations

$M(x)$	$\sigma$	$x_1$	$\sigma_\theta$	$\beta_0$	$\sigma_\beta$	$\tau_l$	$\tau_u(\text{add.})$	$\lambda_l$	$\lambda_u$
$e^x + 2x - 5$	0.5	3	1	6	1	0	5	0	100
$x^2 - 2$	0.05	2	1	2	0.5	0	2	0	100
$-0.4 + x + 0.2 \sin(5x)$	0.05	-1	1	0.5	0.1	0	2	0	100
$e^{2x}/(1 + e^{2x}) - 0.9$	0.01	0	1	0.2	0.05	0	1	0	100

Table 2: MSE of  $x_{11}$  from the simulations

$M(x)$	Additive	Slope	Wu	RM
$e^x + 2x - 5$	.0015	.0026	.0565	.1545
$x^2 - 2$	.00008	.00007	.00031	.00030
$-0.4 + x + 0.2 \sin(5x)$	.00017	.00020	.00065	.00026
$e^{2x}/(1 + e^{2x}) - 0.9$	.0008	.0006	.0253	.0714

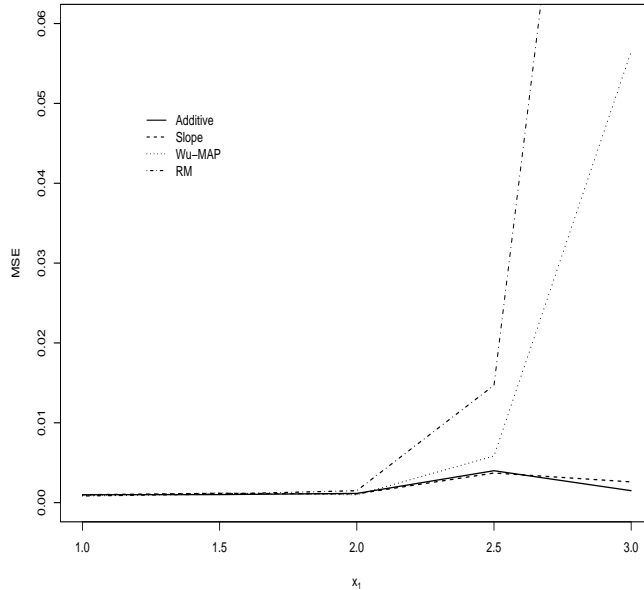


Figure 2: Simulation study. Mean squared error of the estimator of  $\theta$  against the starting point  $x_1$ , for the function  $M(x) = e^x + 2x - 5$ .

## 7. CONCLUSIONS

Wu’s MLE approach to stochastic root-finding has a drawback that if the assumed parametric model is very different from the true model, then the convergence of the procedure becomes very slow. In this article we proposed a new adaptive design using a pinned Gaussian process that overcomes this problem. This adaptive design automatically gives more importance 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 to converge faster irrespective of the assumed model. Two versions of the proposed approach namely, additive error model and slope error model are discussed in the paper. Their superior performance over the Robbins-Monro procedure and Wu’s MAP procedure is demonstrated through simulations.

The convergence proof for the sequential procedure is anticipated to be very complicated because of its complexity. Simulations clearly show that the procedure is promising and worth considering for future research. Extensions of the approach to non-normal distributions are also discussed, although a lot more work is necessary 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 the above methodology to multivariate case will be a worthwhile topic for future research. Applications to stochastic optimization is also an interesting topic for research.

## ACKNOWLEDGEMENT

I am thankful to Professor C. F. Jeff Wu for the valuable comments and suggestions.

## 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.
- Blight, B. J. N. and Ott, L. (1975). "A Bayesian approach to model inadequacy for polynomial regression," *Biometrika* **62**, 79-88.
- 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.
- Coad, D. S. and Woodroffe, M. (1998). "Approximate bias calculations for sequentially designed experiments," *Sequential Analysis* **17**, 1-31.
- Finney, D. J. (1978). *Statistical Methods in Biological Assay*, London: Griffin.
- Frees, E. W. and Ruppert, D. (1990). "Estimation following a sequentially designed experiment," *J. Am. Statist. Assoc.* **85**, 1123-1129.
- Hodges, J. L. & 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*, to appear.
- 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.
- Ljung, L. (1999). *System Identification-Theory for the User*, New Jersey: Prentice Hall.
- McCulloch, C. E. and Searle, S. R. (2001). *Generalized, Linear and Mixed Models*. Wiley, NY.
- 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.
- Santer, 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.
- Woodroffe, M. (1991). "Corrected confidence levels for adaptive nonlinear regression," *Am. J. Math. and Management Sciences* **11**, 79-93.
- 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.