2009 American Control Conference
Hyatt Regency Riverfront, St. Louis, MO, USA
June 10-12, 2009

Stopping Small-Sample Stochastic Approximation

David W. Hutchison and James C. Spall

Abstract— **The practical application of stochastic approximation methods requires a reliable means to stop the iterative process when the estimate is close to the optimizer or when further improvement in the estimate is doubtful. Conventional ideas on stopping stochastic approximation algorithms employ criteria based on a proxy distribution — usually the asymptotic distribution. Yet difficulties may arise when applying such distributions to small (finite) samples. We propose an approach that uses the distribution of a statistically similar process called a surrogate for the proxy distribution rather than the asymptotic distribution. Under certain conditions, surrogate-based probability calculations are close to the actual probabilities. The question of how surrogate processes may be developed is also addressed. Two example applications are given.**

I. MOTIVATION

Consider the problem of searching for the optimum of a function whose form is unknown. Stochastic approximation is an iterative procedure suitable for these types of problems, using noisy observations to estimate the root of an unknown function, and, when so structured, enabling the identification of optima. For applications we want to terminate the sequence of estimates when requisite accuracy has been obtained. Therein lies the problem: it is a non-trivial issue to know how accurate the approximation is at any stage.

The asymptotic performance of stochastic approximation has been well-studied, but few general results are known for small-sample situations. The need for a stopping rule for stochastic approximation was recognized in 1952 by Kiefer and Wolfowitz [10]. Chow and Robbins [2] developed a method to sequentially determine bounds on the mean of a continuous random variable with unknown variance.

Since this initial work much of the effort in stopping stochastic approximation has been on estimating the parameters of the asymptotic distribution in order to apply the Chow-Robbins criterion. We consider an aternative method based on an approximate finite-sample distribution.

II. STOCHASTIC APPROXIMATION

A. Problem Statement

Consider a general function $L\colon \mathbb{R}^p \to \mathbb{R}$ defined for $\theta \in \Theta \subseteq \mathbb{R}^p, p > 0$. Our interest is the minimization problem:

$$\underset{\theta \,\in\, \Theta}{\arg\min}\, L(\theta). \tag{1}$$

We assume $L(\theta)$ is bounded from below. The exact form of $L(\theta)$ is not known, and whatever observations we have of the function are obscured by noise. We assume the existence of the gradient of L, $g\colon \mathbb{R}^p \to \mathbb{R}^p$, and that $L(\theta)$ has a unique minimizer denoted by θ^*.

Let $\hat{\theta}_k$ be an estimate for θ^* at iteration k, a_k a step size at time k, and $G_k(\hat{\theta}_k) \in \mathbb{R}^p$ some information related to the gradient of the process, also at time k. We choose an initial estimate $\hat{\theta}_0$ and update the estimates following the scheme

$$\hat{\theta}_{k+1} = \hat{\theta}_k - a_k G_k(\hat{\theta}_k), \quad k = 0, 1, 2, \cdots . \tag{2}$$

We denote noisy observations of the gradient by $Y(\theta)$, and model these observations by

$$Y(\theta) = g(\theta) + e(\theta), \tag{3}$$

where e is a random vector. If we can assume errors with mean zero, then $E[Y(\theta)] = g(\theta)$. Robbins and Monro [13] studied the problem of finding the roots of an unknown function $g(\theta)$ based on noisy observations of $g(\hat{\theta}_k)$. If g is the gradient of L, the loss function in (1), then we can solve the minimization problem. After setting $G_k(\hat{\theta}_k) = Y_k(\hat{\theta}_k)$ the iteration formula for stochastic root-finding is

$$\hat{\theta}_{k+1} = \hat{\theta}_k - a_k Y_k(\hat{\theta}_k). \tag{4}$$

The convergence of a stochastic approximation algorithm requires that conditions be placed on the objective function, the step size sequence, and the bias and variance of the observed or estimated gradient. See, for example, Spall [18, p. 105–107]. With these conditions established, and with $\hat{\theta}_k$ generated according to (2), one can prove that $\hat{\theta}_k \xrightarrow{\text{a.s.}} \theta^*$ as $k \to \infty$ (see Nevel'son and Has'minskiĭ [12]). A general discussion of the stochastic approximation method may be found in Spall [18, Chap. 4].

To obtain a limiting distribution that is not degenerate we scale the error $\hat{\theta}_k - \theta^*$. If the step size function takes the form $a_k = a/(k+1)^\beta$ for $\frac{1}{2} < \beta \leq 1$ (and satisfies certain regularity conditions), one can show that the distribution of the scaled error is asymptotically normal:

$$k^{\beta/2}(\hat{\theta}_k - \theta^*) \xrightarrow{d} N_p(0, \Sigma^*),$$

where Σ^* is a covariance matrix determined by the sequence a_k and by the Hessian of $L(\theta)$, the underlying function (see [12, Thm. 5.1 p. 140] and [18, section 4.4, p. 112 ff]). One informal but natural interpretation of this fact is that $\hat{\theta}_k$ is approximately multivariate normal with mean θ^* and covariance Σ^*/k^β. We denote this distribution by $\hat{F}_k$, that is, $\hat{F}_k \equiv N_p(\theta^*, \Sigma^*/k^\beta)$.

D. W. Hutchison is with Department of Applied Mathematics and Statistics, The Johns Hopkins University, Baltimore, MD 21218 DWHutchison@jhu.edu

J. C. Spall is with the Johns Hopkins University Applied Physics Laboratory, Laurel, MD 20723, USA James.Spall@jhuapl.edu

978-1-4244-4524-0/09/$25.00 ©2009 AACC

III. The Stopping Problem

Direct calculation of stopping times requires knowledge of the joint probability distribution functions of the stochastic approximation process, F_k. This is rarely possible in practice because the true distribution functions (or even their forms) are generally not known.

One solution is to use $\hat{F}_k$ in lieu of F_k. Since the form of the distribution $\hat{F}_k$ is known, this is a well-defined problem. We refer to this method as the asymptotic method or the asymptotic proxy, the latter term reflecting the way it uses (known-form) $\hat{F}_k$ as a "proxy" for F_k.

Our approach to the estimation of the F_k is different. Though non-traditional, it seems better in the context of "stopping" to develop a proxy distribution that attempts to approximate F_k directly, rather than one that approximates the limiting distribution $\hat{F}_k$. The desired result is a proxy that gives acceptable performance even when the sample size is small. This is the idea behind the use of "surrogate processes" to find a proxy for F_k, and the main focus of this section is a discussion of these processes.

A. Surrogate Processes

The concept of surrogate processes is to develop a simple parameterized version of the original process whose behavior is statistically indistinguishable from that of the original for some positive value of the parameter, but is statistically determined when the parameter is zero.

The applicability of this method has been shown for parameter estimation in maximum likelihood estimation problems, among others (see Spall [17]). Spall's formulation sought an estimate $\hat{\theta}$ of an input vector θ from a set of data whose distribution depended on θ and a known scalar η. When the sample is small, it is difficult to say much about the probabilities of $\hat{\theta}$ because the form of the distribution is unknown. An indirect approach is to construct a parameterized sequence producing statistically similar data and resulting in an estimate $\tilde{\theta}$, the probabilities of which are calculable, and then to look for conditions where the probabilities of $\tilde{\theta}$ are close to those of $\hat{\theta}$ irrespective of the sample size. We apply the same principle to the sequence of estimates from a stochastic approximation process.

Let $V_k = \{e_0, e_1, \ldots, e_k\}$ denote the noise arising in the measurements of the gradient through time k. Let θ be a point in $\mathbb{R}^p$, and let $T(\theta, e_k, k)$ be a transformation function describing a single step of the stochastic approximation process at time k. We assume that T is continuously differentiable with respect to θ.[1] The estimate from the algorithm at time k is denoted by $\hat{\theta}_k$. The next estimate expressed in terms on the transformation T is then

$$\begin{aligned}\hat{\theta}_{k+1} &= T(\hat{\theta}_k, e_k, k)\\ &= T(\cdots T(T(\hat{\theta}_0, e_0, 0),\ e_1, 1)\cdots,\ e_k, k)\\ &\equiv T_k(\hat{\theta}_0, V_k).\end{aligned} \tag{5a}$$

The notation T_k is used to represent the generating formula for the general stochastic approximation process. When the stochastic sequence is that given by equation (4), for example, then the transformation T is a nonlinear operator with $T(\theta, e_k, k) = \theta - a_k g(\theta) - a_k e_k$, and T_k is the k-fold composition of T with itself.

We parameterize the transformation in (5a) with $\eta \in \mathbb{R}$:

$$\hat{\theta}_{k+1} = T_k(\hat{\theta}_0, V_k; \eta). \tag{5b}$$

We select parameter η such that $\eta = 0$ produces a sequence of estimates that behave in a manner "statistically similar" to estimates of the original process, though whose properties are known, or knowable with tolerable effort, for each k.

We denote the surrogate process by $\{\tilde{\theta}_k, k \geq 1\}$ where the estimate $\tilde{\theta}_k$ is generated by $\tilde{\theta}_{k+1} = T_k(\hat{\theta}_0, V_k; 0)$. The distribution function of $\hat{\theta}_k$ is denoted by F_k, and the distribution function of $\tilde{\theta}_k$ is given by $\tilde{F}_k$.

The sequence of $\tilde{\theta}_k$ need not converge to θ^*. This poses no problem for a stopping proxy based on the shape of the distribution, since location is not a factor in the computation (stopping occurs when the *dispersion* of the proxy is small).

Let $h = [h_1 \ h_2 \ \ldots \ h_p]^{\mathrm{T}}$ be a vector of scalar perturbations, $0 < h_j < \infty$, for $j = 1, \ldots, p$, and let $\theta = [t_1 \ t_2 \ \ldots \ t_p]^{\mathrm{T}}$ be any parameter vector in $\mathbb{R}^p$. The set $S_h(\theta) = [t_1 - h_1,\ t_1 + h_1] \times \cdots \times [t_p - h_p,\ t_p + h_p]$ is a symmetric hyper-rectangular region centered at θ.

Let $\hat{t}_{kj}$ be the components of $\hat{\theta}_k$ and t_j^* the components of θ^*. The probability $P(\hat{\theta}_k \in S_h(\theta^*))$ should be interpreted

$$P(\hat{\theta}_k \in S_h(\theta^*)) = P(t_j^* - h_j \leq \hat{t}_{kj} \leq t_j^* + h_j,\ \forall j).$$

We note that the original stochastic approximation process, $\hat{\theta}_k = T_k(\hat{\theta}_0, V_k; \eta)$, and the surrogate, $\tilde{\theta}_k = T_k(\hat{\theta}_0, V_k; 0)$, both involve the same noise sequence, and therefore $\tilde{\theta}_k$ is defined on the same filtered probability space as $\hat{\theta}_k$.

B. Theoretical Constructs

There are several conditions that must be imposed to justify the use of surrogate processes — additional to those required for convergence of the stochastic approximation.

We require the distribution of the noise terms in V_k to be continuous in a neighborhood of $S_h(\theta^*)$. As the noise is frequently observable, this condition may be verifiable based on the data.

We also require that the transformation $T_k(\hat{\theta}_0, V_k; \eta)$ be differentiable in η This allows the formation of the Taylor expansion of the components of $\hat{\theta}_k$. We use a linear approximation of $\hat{t}_{kj}$ near $\eta = 0$ to simplify an expression relating the components of $\hat{\theta}_k$ and $\tilde{\theta}_k$.

Additionally, conditions on the region $S_h(\theta^*)$ are required to ensure that certain probabilities will exist near the boundary of the hyper-rectangle enclosing the minimizer θ^*.

Finally, we impose a "fidelity" condition. Clearly, the question of how well a parameterized process represents the original process is central to whether or not its distribution can be used as a proxy for the distribution of the original process. We require a measure of how good a particular parameterization is likely to be. This is accomplished with

[1] This is true for the usual Robbins-Monro formulation as in (4) if the gradient $g(\theta)$ is continuously differentiable, a common assumption.

conditions that relate the distributions of $\hat{\theta}_k$ and $\check{\theta}_k$. The condition is used to bound complicated probabilities that arise in the proof by much simpler ones. This is a technical condition that is difficult to check *a priori*, but it could be tested empirically during algorithm execution.

For a more detailed discussion of the these conditions, see [7, Sec. 3.3].

The following theorem is only for the special case of a hyper-rectangle $S_h(\theta^*)$. The result extends to general regions using an integrability theorem.

Theorem 1. *Let N be a strictly positive integer, and let $\{\hat{\theta}_k, k \geq 1\}$ be a stochastic approximation process defined by (5b) satisfying the conditions for convergence. Let $\{\tilde{\theta}_k, k \geq 1\}$ be a surrogate process defined by (5b) with $\eta = 0$. Let F_k be the distribution of $\hat{\theta}_k$ and $\tilde{F}_k$ be the distribution of $\tilde{\theta}_k$. Finally, let $S_h(\theta^*)$ be a symmetric hyper-rectangular region centered at θ^*. Then for all $k \leq N$,*

$$\left| P(\hat{\theta}_k \in S_h(\theta^*)) - P(\tilde{\theta}_k \in S_h(\theta^*)) \right| = O(\eta). \tag{6}$$

A general outline of the proof is given here; for a complete proof of the theorem see [7, Sec. 3.3]. For clarity, we use $\hat{E}$ and $\tilde{E}$ to denote the events $\{\hat{\theta}_k \in S_h(\theta^*)\}$ and $\{\check{\theta}_k \in S_h(\theta^*)\}$, respectively. We form the absolute difference between the probabilities of these events and express the difference as a sum of probabilities using the set identities $\hat{E} = (\hat{E} \cap \tilde{E}^{\mathrm{c}}) \cup (\hat{E} \cap \tilde{E})$ and $\tilde{E} = (\tilde{E} \cap \hat{E}^{\mathrm{c}}) \cup (\hat{E} \cap \tilde{E})$. We then relate and bound terms in the sum of probabilities using the fidelity property. The component-wise Taylor expansion of $\hat{\theta}_k(\eta)$ in terms of the parameter η is used to further bound the sums of probabilities to an order depending on η. Using a result of Spall [17, Lemma 2], the probability equation is simplified. The probabilities are shown to be $O(\eta)$ by computing the probability density functions and bounding the integrals of the densities with the mean value theorem.

IV. DEVELOPING A SURROGATE

We consider a common case in which the noise is independent and identically distributed. If the conditions for convergence are satisfied and if the noise distribution is from a family that is closed under addition and scalar multiplication, then one possible approach to determining a suitable surrogate process is through linearization.

Consider the sequence given by (4). The generating transformation for the $k+1$st estimate of the stochastic approximation is

$$\hat{\theta}_{k+1} = T_k(\hat{\theta}_0, V_k) = \hat{\theta}_k - a_k g(\hat{\theta}_k) - a_k e_k. \tag{7}$$

The Jacobian of $g(\theta)$ is assumed to exist. Let $\varphi_k \in \mathbb{R}^p$. Using $R_k \equiv R(\varphi_k, h_k)$ for notational convenience, the first order Taylor expansion of the gradient about the point φ_k is

$$g(\theta) = g(\varphi_k) + H(\varphi_k)(\theta - \varphi_k) + R_k \tag{8}$$

where R_k is a remainder term. The notation $H(\varphi_k)$ is shorthand for $\nabla g(\theta)\big|_{\theta=\varphi_k}$, the Jacobian of g evaluated at φ_k. Since g is the gradient of L, the Jacobian of g is the Hessian of L, and we use the symbol H to denote this Hessian/Jacobian matrix.

Substituting (8) into (7) gives:

$$\begin{aligned} T_k(\hat{\theta}_0, V_k) &= \hat{\theta}_k - a_k g(\varphi_k) \\ &\quad - a_k H(\varphi_k)(\hat{\theta}_k - \varphi_k) - a_k e_k - a_k R_k. \end{aligned} \tag{9}$$

A natural parameterization of (9), then, is to place a factor η on the remainder term $a_k R_k$:

$$\begin{aligned} T_k(\hat{\theta}_k, V_k; \eta) &= \hat{\theta}_k - a_k g(\varphi_k) \\ &\quad - a_k H(\varphi_k)(\hat{\theta}_k - \varphi_k) - a_k e_k - a_k \eta R_k \end{aligned} \tag{10}$$

When $\eta = 1$, the remainder term is included in the generating transformation, and (10) is the same as (9). When $\eta = 0$ we have a simplified process that omits the remainder term. This is our surrogate process which we denote by $\tilde{\theta}_k$:

$$\begin{aligned} \tilde{\theta}_{k+1} &\equiv T_k(\hat{\theta}_0, V_k; 0) \\ &= \tilde{\theta}_k - a_k g(\varphi_k) - a_k H(\varphi_k)(\tilde{\theta}_k - \varphi_k) - a_k e_k \end{aligned} \tag{11}$$

Of course, the parameterization in (10) must be checked against the conditions of Theorem 1 to confirm that the resulting distribution will be that of a valid proxy.

With $\tilde{\theta}_k$ computed according to (11), if the e_k are distributed multivariate normal, then all of the $\tilde{\theta}_k$ are multivariate normal, so the surrogate is a normal process.

V. APPLICATIONS

To illustrate the method we choose two examples: an optimization that seeks the minimum of a simple function, and a more practical simulation optimization problem.

A. The Experimental Approach

We use stochastic approximation to generate a fixed-length sequence of estimates. The sequence is then passed through a stopping algorithm to compute when stopping should have occurred. This constitutes one trial. Multiple trials are conducted and the Monte Carlo nature of the experiment enables us to estimate the distribution of the stopping times and the distance from the final estimate to the true minimum.

B. Application 1: A Function with Known θ^*

The test function we chose was the two dimensional instance of function 25 from the Moré et al. suite of optimization problems [11, Sect. 3] — the so-called variably-dimensioned function.

Let $\theta = [t_1\ t_2]^{\mathrm{T}} \in \mathbb{R}^2$ and $L\colon \mathbb{R}^2 \to \mathbb{R}$; then this function is defined as

$$\begin{aligned} L(\theta) &= (t_1 - 1)^2 + (t_2 - 1)^2 \\ &\quad + (t_1 + 2t_2 - 3)^2(1 + (t_1 + 2t_2 - 3)^2). \end{aligned}$$

By inspection, this function has a unique global minimizer at $\theta^* = [1\ 1]^{\mathrm{T}}$ with $L(\theta^*) = 0$. Moré uses $\hat{\theta}_0 = [\frac{1}{2}\ 0]^{\mathrm{T}}$ for his testing, and we use the same initial point for our analysis.

It is obvious from this form that H is positive semi-definite, and therefore L is strictly convex.

For this example, we suppose the loss function L and gradient g are not known, but that we are able to provide inputs θ and observe noisy gradient values. We assume the components of the noise e_k are i.i.d. $N(0, 10^2)$, resulting in a high level of noise in observations near θ^* (compared to the function values the noise obscures). This situation corresponds to one where measurement errors are independent of the gradient observations.

For a sequence of inputs $\{\hat{\theta}_k\}$ we have a sequence of observations $\{Y_k\}$ artificially returned by $Y_k(\hat{\theta}_k) = g(\hat{\theta}_k) + e_k$ where the random errors are generated by a pseudo-random number generator on a computer.

The distribution function F_k is unknown, and all but impossible to calculate. However, it can be approximated using a Monte Carlo experiment. A single trial consists of using the Robbins-Monro stochastic approximation algorithm (4) to approximate the minimum of the variably-dimensioned function in two dimensions given noisy observations of the gradient. The sample path thus generated represents the trial.

The process is stopped deterministically after 5,000 steps. The purpose is to generate a vector of estimates that are tested sequentially with a stopping rule to determine whether the stochastic approximation would have stopped before $k = 5,000$ or not. The complete Monte Carlo experiment consists of 10,000 such sample paths.

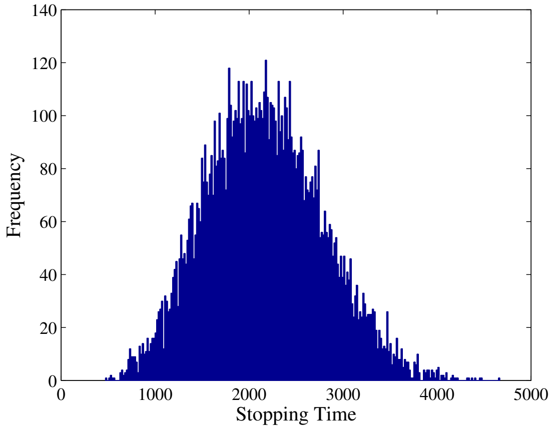
The stopping time $\kappa(\delta, \alpha)$ is based on an accuracy tolerance δ and level of significance α. Suppose $\mathcal{C}(\alpha)$ is a $1 - \alpha$ confidence region. The stopping time $\kappa(\delta, \alpha)$ is found by choosing the first k such that $\operatorname{diam}(\mathcal{C}(\alpha)) \leq \delta$. If $\mathcal{C}_{\text{EM}}$ is an empirical confidence region, we can compute an empirical stopping time $\kappa_{\text{EM}}(\delta, \alpha)$.

The example uses the step size sequence determined by the relation $a_k = a/(k+1)^{0.501}$. The empirical distribution becomes more peaked with increasing k, and while the dispersion is relatively large initially, there is rapid concentration of the distribution once the mean moves into the vicinity of the minimum. The empirical stopping time found by the first passage rule is $\kappa_{\text{EM}}(0.25, 0.10) = 3359$.

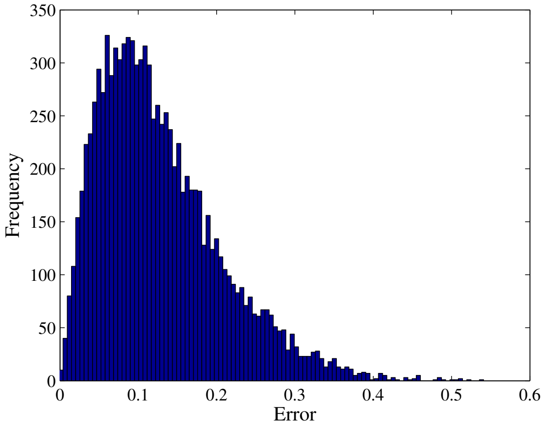
A histogram of stopping times based on a surrogate distribution is shown in Figure 1a. The mean stopping time is $\bar{\kappa}(0.25, 0.10) = 2160$, which is smaller than but comparable to the predicted value of 3359. Figure 1b shows the distribution of the error in the stopped process. The average error is 0.1276.

The step size sequence satisfies the conditions for convergence, and the use of an asymptotic proxy is valid. The average stopping time obtained when using the asymptotic proxy was 22.45. The median stopping time was 21. The average error was 0.1785. Refer to Figure 2.

Based on these results, we claim for each trial that there is a 0.90 probability that the final estimate $\hat{\theta}_\kappa$ is within 0.25 of the optimal value θ^*. Since the true minimum is known, we calculated the actual errors to find that our claim of accuracy was correct in 9,163 of 10,000 cases, so the empirically-determined probability is actually greater than predicted. The stopping algorithm is conservative in this case.



(a) Empirical distribution of the stopping times for the spectral stopping criterion using a linearized gradient surrogate process (10,000 observations).



(b) Distribution of the error $\|\hat{\theta}_\kappa - \theta^*\|$, where κ is the stopping time determined as in Figure 1a (10,000 observations).

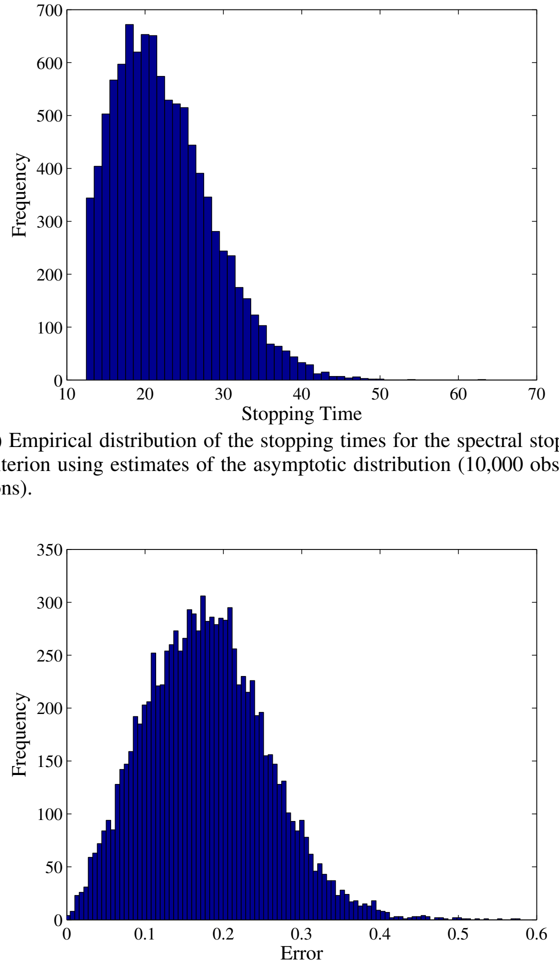
Fig. 1: Stopping with the proxy from a surrogate distribution.

More importantly, it seems evident that, at least in this example, the asymptotic distribution is not a good proxy for the distribution of $\hat{\theta}_k$.

C. Application 2: Simulation Optimization

Long-term trends in air travel have added increasing numbers of travelers and flights to an already congested air travel system. The result has been an inevitable rise in air traffic delay. The costs and causes of air traffic delay have been documented in previous studies [3], along with strategies to reduce the cost of controllable delays [5].

Prominent delay control measures are gate holding policies, metering aircraft through a control point, vectoring, and others. These strategies reflect control decisions that are typically made at the individual flight level, often just hours (sometimes minutes) in advance of execution, and are based on projected flows [4]. Deciding on the control measures to apply to a set of flights for any given day is a difficult nonlinear optimization problem that can be tackled using simulation optimization.



(a) Empirical distribution of the stopping times for the spectral stopping criterion using estimates of the asymptotic distribution (10,000 observations).

(b) Distribution of the error $\|\hat{\theta}_\kappa - \theta^*\|$, where κ is the stopping time determined as in Figure 2a (10,000 observations).

Fig. 2: Stopping with the asymptotic distribution proxy.

Simulation optimization is a category of search methods where the "measurements" are outputs from a simulation model (see Spall [18, Chap. 14–15] or Gosavi [6, Chap. 5] for more details on simulation optimization).

Earlier studies showed that reductions in the cost of delay can be obtained by using a simulation optimization procedure to process delay cost measurements. For details see [8], [9].

1) Problem Formulation: For our study we used the SIMMOD simulation [1] to model the flow of 84 flights (departures and arrivals) on a network of four airports.

We structure the air flow to create considerable delay in the system. The intent is to determine the effectiveness of gate holds in reducing delay and costs. Gate holds occur when a flight is delayed departing its gate. The decision parameters in our formulation are actual (versus scheduled) aircraft departure times, and we formulate the problem to optimize these continuous-valued decision parameters.

TABLE I: System delays (minutes) averaged across all trials.

Type Delay	Initial Values	Final Values	Change	Percent
Gate	0	1114.3	+1114.3	NA
Ground	8.2	10.0	+1.8	+21.9%
Air	1925.1	1313.7	-611.4	-31.8%
Weighted	6083.3	5272.7	-810.6	-13.3%

Suppose $\theta \in \Theta \subset \mathbb{R}^p$ ($p = 84$, the number of flights) is a vector of controllable system parameters (in our case the departure times for each flight). Let $L(\theta)$ be a loss function, which is the sum of all delays in the system (gate, ground, and air), weighted by their relative costs [3].[2] We observe output $y(\theta)$ from the simulation, where $E[y(\theta)] = L(\theta)$. Our objective is to minimize $L(\theta)$ subject to relevant constraints:

$$\min_{\theta \in \Theta} L(\theta) = \min_{\theta \in \Theta} E[y(\theta)]. \tag{12}$$

We kept all inputs fixed except for aircraft departure times. This ensured that each simulation run began with the same starting circumstances except for intentional changes to the departure times.

SIMMOD is a terminating simulation: there is an event that terminates the run, and it always runs to completion. We ignored all outputs except the minutes of delay. Delay factors in the loss function specify the cost of ground and air delays relative to gate delays were taken from Geisinger [3].

The loss function, then, is the total weighted delay time, computed as follows:

$$\begin{aligned} L(\theta) = \text{gate delay}(\theta) + \text{ground delay}(\theta) \times 2.34 \\ + \text{air delay}(\theta) \times 3.15 \end{aligned} \tag{13}$$

Only the noisy loss function is observable and there are a variety of stochastic approximation methods based on these measurements. In this setting we use the particularly efficient method of simultaneous perturbation stochastic approximation (SPSA) developed by Spall [14], [15]. The reader is referred to Spall [16], [18, Chap. 7] for a general discussion of SPSA, and to Hutchison [8], [9] for details in applying SPSA to the constrained aircraft delay problem.

2) Results: We ran a number of Monte Carlo trials, each trial consisting of one million iterations (i.e., each trial was a sample path of one million points). The large number of iterations was needed to compensate for the high dimension and the magnitude of noise in the simulation.

The relevant outputs from the simulation were collected and saved as a data stream, as in the known function example.

Initial and final delay figures (averaged across all trials) are given in Table I. The average weighted delay before optimization was 6083. The average stopping time based on the spectral stopping criterion was $\kappa = 660,012$, resulting in a (smaller) average weighted delay of 5273. In particular, expensive air delay was reduced from 1925 minutes to 1314 minutes, a reduction of 31.8%.

[2] One normally thinks of delays as resulting from random components of the air traffic system: weather, head or tail winds, aircraft load, or any number of other random effects. In our congested system, delays are also the result of inefficient scheduling. The objective is to decrease the overall cost of these scheduling delays in the face of noise introduced by other random effects by increasing inexpensive delays (gate hold times) in the hope of reducing the more expensive delays (ground or air delays).

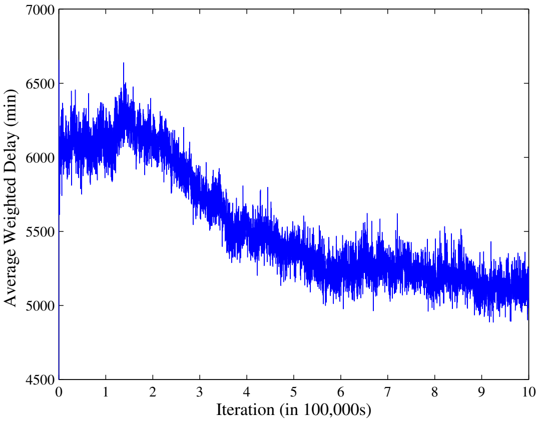


Fig. 3: Plot of the loss function (weighted delay time) at each iteration of the process.

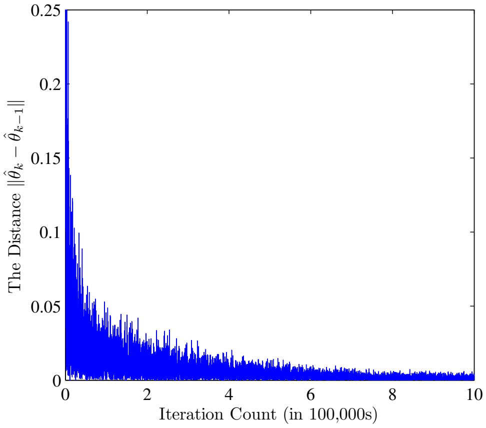


Fig. 4: Distance between the estimates $\hat{\theta}_k$ and $\hat{\theta}_{k-1}$ for one trial of the simulation optimization.

While ground delays have increased significantly (21.9%), the additional 1.8 minutes of average ground delay represents only 0.08% of overall system costs. The initial and final numbers for ground delays are very small compared to other delays in the system, and time lost on the ground has a minor impact on system-wide costs resulting from delay.

Flights were delayed at the gate an average of about 15.6 minutes, though in individual trials some flights were delayed the maximum time of one hour. No flight was consistently held that long, but on average six flights departed 55 minutes or more after their scheduled time. Additionally, 28 flights (on average) delayed their departure five minutes or less.

The progress of the simulation optimization can be observed in a plot of the average weighted aircraft delay by iteration for all one million iterations (Figure 3). The plot shows only every tenth point and is averaged over all trials.

The true error $\|\theta^* - \hat{\theta}_k\|$ is unobtainable. However, we can measure the distance $\|\hat{\theta}_k - \hat{\theta}_{k-1}\|$ as $\hat{\theta}_k$ moves away from $\hat{\theta}_0$. Figure 4 shows a graph of this distance for one trial. We expect $\|\hat{\theta}_k - \hat{\theta}_{k-1}\|$ to get small. When the estimate is in the vicinity of θ^* (or if the process is running out of steam), we anticipate that the iteration-to-iteration change in the estimate should be small. When we see such behavior, it is not proof that $\hat{\theta}_k$ is near θ^* (it could be on a functional plateau or near a local minimum, for example), but it is necessary behavior for it to be near the minimum.

References

[1] ATAC Corporation, Sunnyvale, CA. *SIMMOD Reference Manual*, 1995.

[2] Y. S. Chow and H. Robbins. On the asymptotic theory of fixed-width sequential confidence intervals for the mean. *Annals of Mathematical Statistics*, 36(2):457–462, April 1965.

[3] K. Geisinger. Airline delay 1976–1986: Based upon the Standardized Delay Reporting System. Report FAA-APO-88-13, Office of Aviation Policy and Plans, Federal Aviation Administration, Washington, DC, March 1989.

[4] E. P. Gilbo. Optimization of air traffic management strategies at airports with uncertainty in airport capacity. In M. Papageorgiou and A. Pouliezos, editors, *Proceedings of the 8th IFAC/IFIP/IFORS Symposium*, pages 35–40, Chania, Greece, 16–18 June 1997.

[5] E. P. Gilbo. Optimizing airport capacity utilization in air traffic flow management subject to constraints at arrival and departure fixes. *IEEE Transactions on Control Systems Technology*, 5(5):490–503, September 1997.

[6] A. Gosavi. *Simulation-Based Optimization: Parametric Optimization Techniques and Reinforcement Learning*. Kluwer Academic, Boston, 2003.

[7] D. W. Hutchison. *Stopping Times and Confidence Bounds for Small-sample Stochastic Approximation Algorithms*. Ph.d. dissertation, The Johns Hopkins University, Baltimore, MD, May 2009.

[8] D. W. Hutchison and S. D. Hill. Simulation optimization of airline delay with constraints. In B. A. Peters, J. S. Smith, D. J. Medeiros, and M. W. Rohrer, editors, *Proceedings of the 2001 Winter Simulation Conference*, pages 1017–1022, Arlington, VA, 9–12 December 2001.

[9] D. W. Hutchison and S. D. Hill. Simulation optimization of airline delay with constraints and multiple objectives. In *Proceedings of the Fourth International Symposium on Uncertainty Modeling and Analysis*, pages 417–422, College Park, MD, 21–24 September 2003.

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

[11] J. J. Moré, B. S. Garbow, and K. E. Hillström. Testing unconstrained optimization software. *ACM Transactions on Mathematical Software*, 7(1):17–41, 1981.

[12] M. B. Nevel'son and R. Z. Has'minskiĭ. *Stochastic Approximation and Recursive Estimation*. Number 47 in Translations of Mathematical Monographs. American Mathematical Society, Providence, RI, 1973.

[13] H. Robbins and S. Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22(3):400–407, September 1951.

[14] J. C. Spall. A stochastic approximation algorithm for large-dimensional systems in the Kiefer-Wolfowitz setting. In *Proceedings of the IEEE Conference on Decision and Control*, pages 1544–1548, Austin, TX, 7–9 December 1988.

[15] J. C. Spall. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Transactions on Automatic Control*, 37(3):332–341, March 1992.

[16] J. C. Spall. An overview of the simultaneous perturbation method for efficient optimization. *Johns Hopkins APL Technical Digest*, 19(4):482–492, October–December 1998.

[17] J. C. Spall. Uncertainty bounds in parameter estimation with limited data. In M. Dror, P. L'Ecuyer, and F. Szidarovszky, editors, *Modeling Uncertainty: An Examination of Stochastic Theory, Methods, and Applications*, volume 46 of *International Series In Operations Research and Management*, pages 685–709. Kluwer Academic Publishers, Boston, 2002.

[18] J. C. Spall. *Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control.* John Wiley and Sons, Inc., Hoboken, NJ, 2003.