Proceedings of the
44th IEEE Conference on Decision and Control, and
the European Control Conference 2005
Seville, Spain, December 12-15, 2005

A Method for Stopping Nonconvergent Stochastic Approximation Processes

David W. Hutchison and James C. Spall

Abstract— **When a stochastic approximation process satisfies the conditions for convergence there are well-established methods to terminate the iterative process in a manner that allows approximate statistics to be calculated on the final result. Many of these use the asymptotic properties of convergent stochastic approximation. However, such methods converge slowly due to step size restrictions, so in practical application it is common to use a step size that violates the conditions for convergence in order to obtain an answer more quickly. Constant gain stochastic approximation is a special case of this practice. In these cases stopping rules based on asymptotic methods are no longer analytically supportable, and other techniques must be found. This paper presents one such method based on the use of a surrogate process to calculate the stopping condition. A discussion of this approach to stopping stochastic approximation is offered in the context of a simple example, including some empirical results.**

I. INTRODUCTION

Consider a general, real-valued function $L : \mathbb{R}^p \to \mathbb{R}$ defined for $\theta \in \mathbb{R}^p$. We are interested in the following minimization problem:

$$\min_{\theta} L(\theta). \tag{1}$$

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, $g : \mathbb{R}^p \to \mathbb{R}^p$, and that $L(\theta)$ has a unique minimum denoted by θ^*. We do not require an explicit mathematical expression for $L(\theta)$, and in the most general case we may only be able to obtain a sequence of noisy measurements (evaluations) of L or g (perhaps both).

In this paper we examine the "noisy gradient" case where, perhaps in addition to noisy observations of $L(\theta)$, we also have noisy observations of the gradient of $L(\theta)$. We denote the sequence of gradient observations by $\{Y_m(\theta)\}$, and model these observations by

$$Y_m(\theta) = g(\theta) + e_m(\theta),$$

where $\{e_m\}$ a sequence of random vectors. If we can assume the observations are unbiased, then the errors have mean zero, and $E[Y_m(\theta)] = g(\theta)$. In accordance with Robbins-Monro [1], problem (1) is solved as a root-finding problem using the noisy observations $Y_m(\theta),\ m = 1, 2, ..., n$.

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

J. C. Spall is on the Principal Professional Staff at The Johns Hopkins University Applied Physics Laboratory, Laurel, MD 20723 James.Spall@jhuapl.edu

Let $\hat{\theta}_k$ be an estimate for the optimal value θ^* at iteration k and a_k a step size. We choose an initial estimate $\hat{\theta}_0$ and update it with the following scheme [1]:

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

A general discussion of the stochastic approximation method may be found in Spall [2, sections 4.1 and 4.2]. The asymptotic properties of $\hat{\theta}_k$ are well-known [2, sections 4.3 and 4.4], and under appropriate conditions the sequence of iterates generated by (2) converges to θ^* almost surely. See, for example, [3, Thm. 2.1, p. 95], [4, Thm. 4.2, p. 88], or [2, section 4.3]. Additionally, when properly scaled, the error between $\hat{\theta}_k$ and θ^* is normally distributed with mean zero as k gets large (see [4, Thm. 5.1, p. 140] and [2, section 4.4]). We denote the distribution of $\hat{\theta}_k$ by F_k, and the asymptotic distribution by F^*.

Almost all that is known about stochastic approximation comes from limit theorems. The earliest analytical results were by Robbins and Monro. They proved convergence in quadratic mean of $\hat{\theta}_k$ to θ^* for their algorithm under mild conditions, thereby implying convergence in probability. A slight tightening of these conditions enabled Blum [5] to prove almost sure convergence (see also Gladyshev [6] or Kushner and Clark [7]). Subsequent results have proved asymptotic normality of the iterate $\hat{\theta}_k$ [6], [8], [9], [10], a theoretical rate of convergence of $O(k^{-1/2})$ [11], [12], convergence probability bounds [13], and conditions that are necessary and sufficient for convergence [7], [14].

The conditions required for convergence are stringent, particularly those on the sequence of step sizes. From the conditions in Spall we see that we must have $a_k > 0$, $\sum a_k = \infty$, and $\sum a_k^2 < \infty$. The sequence generated by $a_k = a/k$ for some small a satisfies the convergence requirements and leads to the best theoretical rate of convergence. But this sequence decreases rapidly, and the stochastic approximation converges slowly.

It is common for practitioners of stochastic approximation to use a larger step size that moves the iteration faster. For example, the sequence generated by $a_k = a/k^\alpha$ gives greater freedom in specifying the step size sequence. When $1/2 < \alpha \leq 1$ the step size sequence satisfies the conditions for convergence. However, when $0 \leq \alpha \leq 1/2$ the approximation process is no longer guaranteed to converge, but in a practical application where a finite number of steps will be taken, this is not restrictive.

Some applications work well with a constant step size ($\alpha = 0$), such as applications in adaptive tracking or control. Other applications like neural networks may also use a

0-7803-9568-9/05/$20.00 ©2005 IEEE

constant step size. The approximation process is known to not converge in this case, but the choice works well for small samples and is easy, and therefore is acceptable (even preferable) for many applications.

This practice is problematic, however, for stopping rules that use an estimate of the asymptotic distribution found by sequential estimation. Constant step size algorithms violate the conditions for convergence, and generally do not converge. Additionally, the asymptotic distribution of the standardized iterate is generally not multivariate normal [15]. In this case using the asymptotic distribution F^* as an estimator for F_k in stopping calculations may make little sense.

II. STOPPING RULES

The need for a stopping rule for stochastic approximation was recognized by Kiefer and Wolfowitz [16]. Computer implementations of stochastic approximation algorithms require a stopping rule that terminates the computation after a finite number of steps. Of course, the rule should relate to the quality of the solution found.

We distinguish between a stopping criterion and a stopping rule. A stopping criterion is an optimization objective stated mathematically, the satisfaction of which means that we have obtained the goal of our optimization. A stopping rule is part of an optimization algorithm that contains one or more stopping criteria, and perhaps other criteria as well, by which the algorithm is stopped. A suitable stopping rule should do the following;

1) Stop when the error is small enough.
2) Optionally, stop when improvements to the current solution become small.
3) Stop when the iteration budget has been exhausted.

The first instance that any of these conditions are satisfied should cause the algorithm to stop. The first condition represents a stopping criterion and should contain a mathematical interpretation of what constitutes "small enough." The second condition, if used, is a heuristic criterion that we hope will stop the algorithm when it is close to the optimal point in the event we are unable to calculate the error. The third condition is a fail-safe, and thus has no relation to stopping criteria; it is included to prevent the algorithm from running forever in the event the first two conditions fail to terminate the algorithm.

The best case is that some proximity criterion can be calculated or estimated and we stop when it becomes small, for example, $\|\hat{\theta}_k - \theta^*\| < \delta$ for some suitable norm. Since the estimates $\hat{\theta}_k$ are random, the preferred approach is to stop when the probability is high that the tolerance condition has been met. Let δ be a small positive number. We look for conditions like

$$P_{\hat{\theta}_k}\left(\| \hat{\theta}_k - \theta^* \| < \delta \;\middle|\; \hat{\theta}_0\right) \geq 1 - \alpha \tag{3}$$

(see also [17], [18]). Probability conditions involving $|L(\hat{\theta}_k) - L(\theta^*)| < \delta$ are also possible. Given an $\alpha \in [0, 1]$ and $\delta > 0$, we stop at time equal to the smallest k such that condition (3) is true. Stopping criteria of this type agree with asymptotic theory in that $k \to \infty$ as $\delta \to 0$.

Unfortunately, proximity criteria are hard to estimate directly since θ^* is unknown, and in the general case, there is no assured way to stop the algorithm when the error is small. The task in developing a good stopping rule, then, is not necessarily to find a good stopping criterion, but rather, to find a good heuristic criterion.

Stopping heuristics principally aim at stopping the algorithm when further expected improvements are unlikely or have fallen below some threshold. There are two general approaches to implementing this philosophy. The first is to use a criterion based on estimate-to-estimate changes, and variations on this approach (contraction criteria). The second approach is to look at the distribution of the estimates themselves (distributional criteria), checking to see if the distribution is concentrating about its mean, and stopping when the degree of concentration meets some threshold. We favor the second approach due to the difficulty of the calculations in the first.

There are many schemes that lead to valid stopping rules based on distributional criteria. We look at a common rule introduced by Yin [19], [20], [21] and an alternative based on the distribution F_k.

1) Small-Volume Stopping Heuristic: The small-volume stopping heuristic uses the asymptotic properties of converging stochastic approximation. When the step size sequence is determined by $a_k = a/k^\alpha$, the quantity $k^{\alpha/2}(\hat{\theta}_k - \theta^*)$ is asymptotically multivariate normal with mean zero and covariance Σ^*. If we estimate Σ^* at time step k by $\hat{\Sigma}_k^*$, where $\hat{\Sigma}_k^* \to \Sigma^*$, then we can approximate the distribution of $k^{\alpha/2}(\hat{\theta}_k - \theta^*)$ by $N(0, \hat{\Sigma}_k^*)$. The confidence region for such a random variable is an ellipsoid defined by

$$E_k(c) = \{\theta : k^\alpha(\hat{\theta}_k - \theta)^{\mathrm{T}}(\hat{\Sigma}_k^*)^{-1}(\hat{\theta}_k - \theta) < c\},$$

for some $c > 0$.

Let χ_p^2 denote the chi-square distribution with p degrees of freedom, and let $\chi_p^2(\alpha)$ be the $1 - \alpha$ percentile of χ_p^2. Then, in the limit, the random quantity in E_k is the sum of centered, normalized normally distributed random variables, that is, it is distributed χ_p^2. We set c_α equal to $\chi_p^2(\alpha)$ to obtain

$$\lim_{k\to\infty} P\left(\theta \in E_k(c_\alpha)\right) = 1 - \alpha. \tag{4}$$

Yin's approach is to compute the volume of E_k, denoted by $V(E_k)$, and stop at a time κ equal to the first k such that $V(E_k) < \delta^p$ for some $\delta > 0$:

$$\kappa = \inf\left\{k \geq k_{tp} : V(E_k(c_\alpha)) \leq \delta^p\right\}. \tag{5}$$

(We have modified Yin's condition somewhat by including k_{tp}, which denotes the "turning point" of the stochastic approximation. The confidence ellipsoids E_k do not necessarily decrease monotonically in size with increasing k. Depending on the covariance of the initial estimate $\hat{\theta}_0$, $V(E_k)$ typically increases sharply during the early iterations after which it peaks and begins a slow, steady decline. The point corresponding to k_{tp} represents the peak.)

In effect, we stop the iteration once the size of the $1-\alpha$ confidence region for the normalized error falls below a threshold level determined by volume. When $V(E_k)$ is small, we would expect $\hat{\theta}_k$ to be close to θ^*. Even when this is not true, a small volume may indicate that there is little chance further iteration will result in measurable improvement, and stopping the algorithm is still desired.

Glynn and Whitt [22, p. 184–185] note that the stopping times in (5) may terminate the process too early if the estimator $\hat{\theta}_k$ is badly behaved for small k. Additionally, they feel it is desirable that the stopping time agree with asymptotic theory, and so we should expect $\kappa \to \infty$ as δ decreases to zero. They modify the condition slightly to obtain this behavior. Let $c(k)$ be a strictly positive function that decreases monotonically to zero as $k \to \infty$ and satisfies $c(k) = o(k^{-1/2})$. They define the small-volume stopping times as

$$\kappa = \inf\left\{k \geq k_{tp} : V(E_k(c_\alpha)) + c(k) \leq \delta^p\right\}. \qquad (6)$$

In small samples the confidence region may not be an ellipsoid, but the same procedure applies, although computing the volume of the region $E_k(c_\alpha)$ may be more difficult. Often the sample covariance of the estimate $\hat{\theta}_k$ produces an ellipsoidal region that approximates the actual region well, though this requires a Monte Carlo approach.

A drawback to this stopping rule is that it provides no insight on how to choose δ, and no understanding of how small the errors are at termination. With this criterion, the process stops when the confidence region is small, which does not necessarily mean when $\hat{\theta}_k$ is close to θ^*. When $E_k(c_\alpha)$ is small it means only that the variability in the normalized error is small. Since the average error is presumably centered at zero, this is equivalent to saying that the error itself is small. Whether or not this statement is valid depends on one's belief that the normalized error is exhibiting asymptotic behavior. Moreover, it is possible that the confidence region could be long and narrow. Thus even when the volume of the region is small and centered at zero, long extremities of the region could allow $\hat{\theta}_k$ to be arbitrarily far from θ^* in some dimensions.

2) Spectral Stopping Heuristic: Suppose the confidence region E_k is enclosed in a ball, which we denote by $B(E_k)$. If the estimates are normal, then the $1-\alpha$ confidence region $E_k(c_\alpha)$ is an ellipsoid with axes $u_m\sqrt{\lambda_m \chi_p^2(\alpha)}$, where u_m is a normalized eigenvector of Σ_k and λ_m is the corresponding eigenvalue, for $m = 1, \ldots, p$. Since E_k is a nonsingular covariance matrix, none of the λ_m are zero.

The radius δ of the ball must therefore be greater than or equal to the longest axis of the ellipsoid, which is related to the largest eigenvalue of the covariance matrix Σ_k. This leads to the specification of a stopping criteria based on the covariance of the iterate $\hat{\theta}_k$. The spectral radius of Σ_k is defined as $\rho(\Sigma_k) = \max_m\{|\lambda_m|\}$. We define the spectral radius stopping rule as follows:

$$\kappa = \min\left\{k : \rho(\Sigma_k) \leq \frac{\delta^2}{\chi_p^2(\alpha)}\right\}. \qquad (7)$$

In other words, stop when the largest eigenvalue of Σ_k is less than $\delta^2/\chi_p^2(\alpha)$.

Again, when the confidence region is not ellipsoidal the procedure works just as well by using as E_k an ellipsoid that encloses the true confidence region. This ellipsoid is often adequately approximated by using the sample covariance of the estimate $\hat{\theta}_k$.

The spectral stopping heuristic is superior to the small-volume stopping heuristic in two important ways. First, when the algorithm is terminated with the spectral stopping heuristic, there is a greater than $1-\alpha$ probability that the iterate is contained in the ball $B(E_k)$. If θ^* is also contained in the ball, then $\|\hat{\theta}_\kappa - \theta^*\| < 2\delta$ with probability greater than $1-\alpha$. Thus the choice of δ is related directly to the desired proximity tolerance.

The second advantage is that it does not rely on asymptotic theory, making the approach suitable for the small sample case, as well as for cases where the conditions for asymptotic convergence are not satisfied. When the stochastic approximation process is nonconvergent, the analytical justification for using the small-volume heuristic, which relies on asymptotic theory, is lost. The spectral stopping heuristic uses the distribution of $\hat{\theta}_k$ which always exists. While the distribution F_k is usually unknown and difficult to find, if it cannot be determined directly, an estimator may be used in the calculation. When the conditions for asymptotic normality hold, the asymptotic distribution can be used as the estimator, and it is computed in the same manner as in the small-volume heuristic. If the conditions do not hold, one may be able to approximate F_k by the use of surrogate processes (see Hutchison and Spall [23]).

III. EXAMPLE

We illustrate these ideas by using a function from the Moré et al. [28] suite of optimization problems, the so-called variably-dimensioned function in two dimensions, to generate stochastic for tests of the stopping rules. Let $\theta = [t_1 \ t_2]^{\mathrm{T}} \in \mathbb{R}^2$ and $L_{VD} : \mathbb{R}^2 \to \mathbb{R}$. Then the variably dimensioned function is defined as

$$L_{VD}(\theta) = (t_1 - 1)^2 + (t_2 - 1)^2 + (t_1 + 2t_2 - 3)^2(1 + (t_1 + 2t_2 - 3)^2).$$

The gradient is

$$g_{VD}(\theta) = \begin{bmatrix} 4t_1 + 4t_2 + 4(t_1 + 2t_2 - 3)^3 - 8 \\ 4t_1 + 10t_2 + 8(t_1 + 2t_2 - 3)^3 - 14 \end{bmatrix}.$$

This function satisfies the conditions for convergence of (2), and there is a unique global minimum located at the point $\theta^* = [1 \ 1]^{\mathrm{T}}$.

For the purposes of this example, we suppose the form of the loss function L_{VD} is not known, but we are able to provide inputs θ and observe the noisy gradient. We assume the components of the noise e_k are independent and normally distributed with mean zero and variance ${\sigma_k}^2$. Now model the stochastic approximation process as follows: for a sequence of inputs $\{\hat{\theta}_k\}$ we have a sequence of observations $\{Y_k\}$

generated by $Y_k(\hat{\theta}_k) = g_{VD}(\hat{\theta}_k) + e_k$. Let $\hat{\theta}_k = [\hat{t}_{1k}\ \hat{t}_{2k}]^{\mathrm{T}}$. Using the usual Robbins-Monro iteration the process is

$$\begin{aligned}\hat{\theta}_{k+1} = &\quad \hat{\theta}_k - a_k g_{VD}(\hat{\theta}_k) - a_k e_k \\ = &\quad \begin{bmatrix} \hat{t}_{1k} \\ \hat{t}_{2k} \end{bmatrix} \\ &-a_k \begin{bmatrix} 4\hat{t}_{1k} + 4\hat{t}_{2k} + 4(\hat{t}_{1k} + 2\hat{t}_{2k} - 3)^3 - 8 \\ 4\hat{t}_{1k} + 10\hat{t}_{2k} + 8(\hat{t}_{1k} + 2\hat{t}_{2k} - 3)^3 - 14 \end{bmatrix} \\ &-a_k e_k.\end{aligned}$$

To compute a proxy for F_k we use a surrogate process based on linearization of the gradient using an estimated Jacobian as in [23]. Let $\tilde{\theta}_k = [\tilde{t}_{1k}\ \tilde{t}_{2k}]^{\mathrm{T}}$. The idealized process linearized about $[1\ 1]^{\mathrm{T}}$ is

$$\tilde{\theta}_{k+1} = \begin{bmatrix} \tilde{t}_{1k} \\ \tilde{t}_{2k} \end{bmatrix} - a_k \begin{bmatrix} 4\tilde{t}_{1k} + 4\tilde{t}_{2k} - 8 \\ 4\tilde{t}_{1k} + 10\tilde{t}_{2k} - 14 \end{bmatrix} - a_k e_k.$$

We denote the distribution function of $\tilde{\theta}_k$ by D_k.

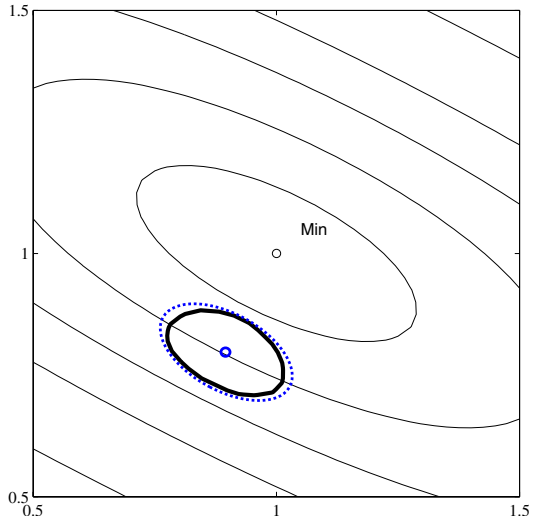
For this example, the surrogate process $\tilde{\theta}_k$ is the sum of scaled bivariate normal random variables for every k. Therefore D_k is bivariate normal and can be calculated exactly.

The true distribution function F_k is unknown and impossible to calculate for large k. However, it can be approximated using a Monte Carlo experiment. We used Robbins-Monro algorithm with step size $a_k = a/k^{\beta}$. An initial estimate of $\hat{\theta}_0 = [\frac{1}{2}, 0]^{\mathrm{T}}$ was used in all cases, and noise was as described with $\sigma_k = 10$. The algorithm was run for 10,000 steps to generate a stream of data. This data stream was then used by each stopping procedure to calculate a stopping time based on a tolerance $\delta = 0.1$. This process was repeated 10,000 times, generating 10,000 stopping times.

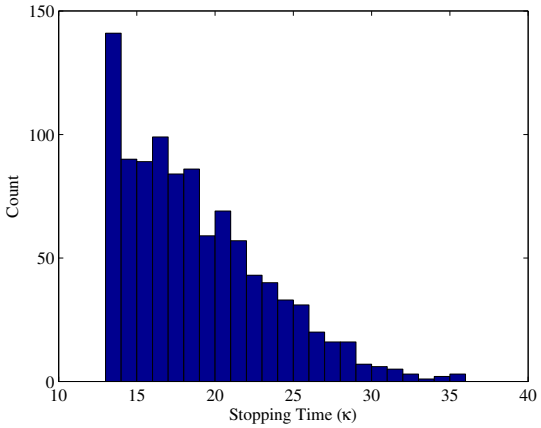
Figure 1 shows the case for $a_k = a/k$, the asymptotically optimal step size. Figure 1a is a plot of the 95% confidence regions at iteration $k = 500$ overlayed on a contour plot of the variably-dimensioned function. The heavy solid line is an approximation to the true confidence region of $\hat{\theta}_{500}$ as determined by 10,000 Monte Carlo trials. The thin solid line (small circle) is the confidence region determined using F^*, and the dotted line is the confidence region based on D_k. Each of the two computed regions has been artificially centered on the Monte Carlo approximation for purposes of comparison.

When the procedure was stopped using the small-volume heuristic using an estimate for the asymptotic distribution as a proxy for the probability calculations, the rapid decrease in the step size caused the confidence region to decrease rapidly as well. As a result, the small-volume heuristic stopped early in every case. Moreover, for any iterate, the asymptotic distribution proved to be a poor proxy for the actual distribution of the iterate.

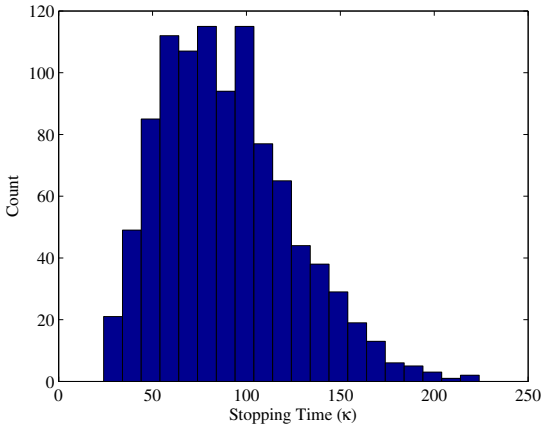
Procedures stopped using the spectral heuristic with a surrogate distribution as a proxy generally obtained better results when measured by the average distance of the stopped



(a) 95% confidence regions after 500 iterations.



(b) Histogram of the stopping times based on the small-volume heuristic.



(c) Histogram of stopping times based on the spectral heuristic.

Fig. 1. A comparison of results from the test case with an asymptotically optimal step size of $a_k = a/k$.

process from θ^*, though the difference was not statistically significant. The reason for this lack of significance is that, even though the process ran longer with the spectral rule, the step size was so small that the estimate $\hat{\theta}_k$ hardly moved.

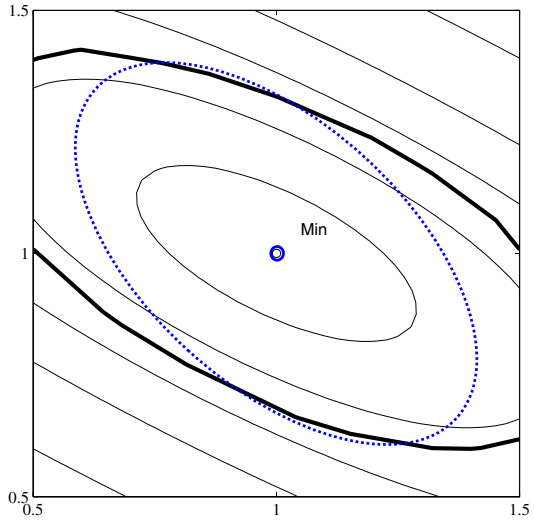
Figures 1b and 1c tell the story in terms of the distribution of the stopping times for each heuristic. For the small-volume case the average stopping time was $\kappa = 19.3$, while for the spectral case the average was $\kappa = 89.8$ (the turning point was $k_{tp} = 12$ for this problem).

We note as an aside that in none of the empirical trials did stopping based on δ imply that $\|\hat{\theta}_k - \theta^*\| < \delta$. In both the small-volume and spectral cases the average distance of the stopped process from θ^* was much greater than the chosen δ. This is not a problem with the stopping calculation, but rather with the basic assumption that θ^* is contained within the $1-\alpha$ confidence region. We can see here that the assumption does not hold, but in the general case, the truth of the assumption can never be known. This serves to point out that it is risky to draw conclusions about accuracy of the final iterate based on tolerance used for stopping, even though that is the intent.

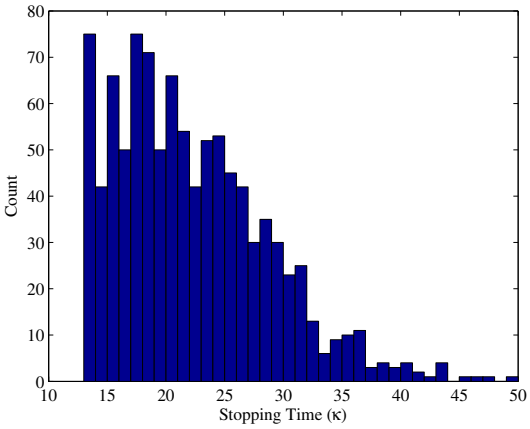
Figure 2 shows the case for $a_k = a/k^{1/4}$. This step size does not guarantee convergence, and the use of the small-volume heuristic is not analytically valid (though we used it anyway to observe its performance). Figure 2a is a plot of the 95% confidence regions at iteration $k = 500$ overlayed on a contour plot of the variably-dimensioned function. The confidence regions are as described for Figure 1.

Note that using a more slowly decreasing step size resulted in a mean $\bar{\theta}_\kappa$ that is closer to the optimal point θ^* than did the optimal step size, but also resulted in much greater dispersion in the individual $\hat{\theta}_\kappa$. One could compensate for this behavior by beginning with a smaller step size or reducing it during the iteration (see Kushner and Huang [29] for some possible reduction schemes). As a result, more iterations are needed to force the confidence region smaller. With a constant step size, increasing the number of iterations does not produce a noticeably better result. Based on empirical results, the shape of the distribution of the final iterate changes little with a deterministic $\kappa \in \{100, 200, 500, 1000, 5000, 10000\}$. This behavior should be expected since with a constant step size the only mechanism to reduce variability is the gradient function.

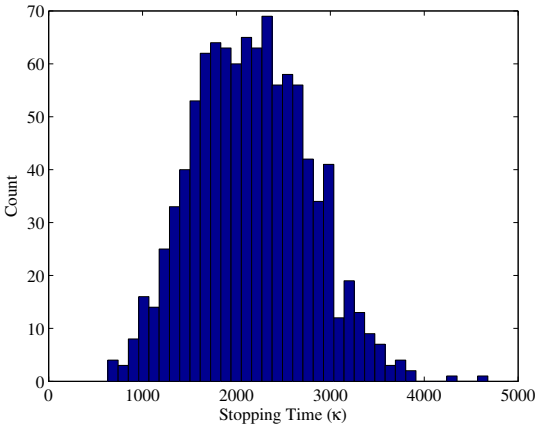
We see from Figures 2b and 2c that the small-volume heuristic continues to stop the process very early (mean stopping time is $\kappa = 23.1$). The average stopping time for the spectral heuristic was $\kappa = 2160.4$. The spectral stopping heuristic also stops early, but this may be a consequence of the poor quality of the surrogate D_k as a proxy for F_k (as evidenced by Figure 1a). The distribution D_k was calculated using a linearized gradient, linearizing about the initial point $\hat{\theta}_0$. However, after even just a few hundred iterations the estimate has moved far from the initial point, and the linearization is no longer a good one. A possible solution to this problem is to simply re-estimate the Jacobian and linearize about the current point.



(a) 95% confidence regions after 500 iterations.



(b) Histogram of the stopping times based on the small-volume heuristic.



(c) Histogram of stopping times based on the spectral heuristic.

Fig. 2. A comparison of results from the test case with a step size of $a_k = a/k^{1/4}$.

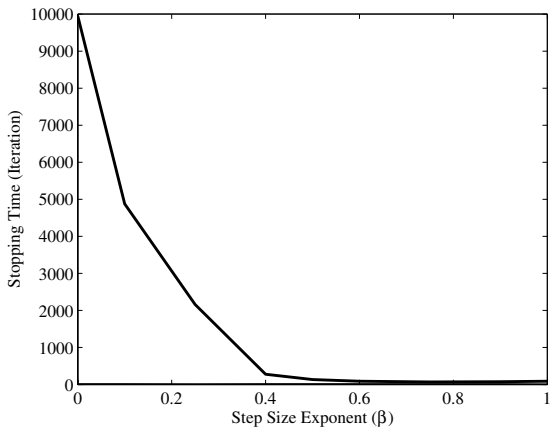


Fig. 3. Mean stopping time plotted against the exponent β in the step size $a_k = a/k^\beta$. The case for $\beta = 1$ corresponds to the asymptotically optimal step size; $\beta = 0$ corresponds to constant step size. The mean stopping times for both the small-volume and spectral heuristics are shown, but the line for the small-volume stopping times is indistinguishable from the horizontal axis at the scale shown.

IV. CONCLUSIONS

Stopping rules based on probabilistic conditions require a distribution for computation. Since actual distributions are impossibly complex, asymptotic theory can provide a proxy distribution to use instead. However, stopping rules tied to asymptotic behavior have limitations. When asymptotic theory does not apply, as in the use of step sizes that violate convergence conditions, other proxies must be sought. Surrogate processes seek to estimate the actual distribution at each iterate, and may produce a suitable proxy distribution.

Even when the use of asymptotic distributions is appropriate, the results can be poor. Figure 3 shows how the average stopping time for each heuristic changes as the step size changes from optimal to constant (by changing β). Among other things, this change results in increased variability of the estimates $\hat{\theta}_k$. The mean spectral stopping time behaves in the manner expected, increasing as the variability increases. The average of the small-volume stopping time, however, increases little (from 19.3 to 23.1). This indicates a peculiar insensitivity to the algorithm, which augurs poorly for the usefulness of that heuristic.

The stopping problem becomes more complicated when constant step sizes are used since the confidence region decreases extremely slowly in size, slowly enough, in fact, that stopping with the methodologies as described in this paper may not apply. Whether or not this is the case is highly problem-dependent.

REFERENCES

[1] H. Robbins and S. Monro, "A stochastic approximation method." *Annals of Mathematical Statistics*, vol. 22, no. 3, pp. 400–407, 1951.

[2] J. C. Spall, *Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control*. New York: Wiley, 2003.

[3] H. J. Kushner and G. G. Yin, *Stochastic Approximation Algorithms and Applications*. New York: Springer-Verlag, 1997.

[4] M. B. Nevel'son and R. Z. Has'minskiĭ, *Stochastic Approximation and Recursive Estimation*, ser. Translations of Mathematical Monographs. Providence, RI: American Mathematical Society, 1973, no. 47.

[5] J. R. Blum, "Multidimensional stochastic approximation methods." *Annals of Mathematical Statistics*, vol. 25, no. 4, pp. 737–744, 1954.

[6] E. Gladyshev, "On stochastic approximation." *Theory of Probability and Its Applications*, vol. 10, pp. 275–278, 1965.

[7] H. J. Kushner and D. S. Clark, *Stochastic Approximation Methods for Constrained and Unconstrained Systems*, ser. Series in Applied Mathematical Sciences. New York: Springer-Verlag, 1978, vol. 26.

[8] K. L. Chung, "On a stochastic approximation method." *Annals of Mathematical Statistics*, vol. 25, pp. 463–483, 1954.

[9] V. Fabian, "On asymptotic normality in stochastic approximation." *Annals of Mathematical Statistics*, vol. 39, no. 4, pp. 1327–1332, 1968.

[10] J. Sacks, "Asymptotic distribution of stochastic approximation procedures." *Annals of Mathematical Statistics*, vol. 29, no. 2, pp. 373–405, 1958.

[11] H.-F. Chen, "Convergence rate of stochastic approximation algorithms in the degenerate case." *SIAM Journal on Control and Optimization*, vol. 36, no. 1, pp. 100–114, 1998.

[12] L. Ljung, "Strong convergence of a stochastic approximation algorithm," *The Annals of Statistics*, vol. 6, no. 3, pp. 680–696, 1978.

[13] L. D. Davisson, "Convergence probability bounds for stochastic approximation." *IEEE Transactions on Information Theory*, vol. 16, no. 6, pp. 680–685, 1970.

[14] I.-J. Wang, E. K. P. Chong, and S. R. Kulkarni, "Equivalent necessary and sufficient conditions on noise sequences for stochastic approximation algorithms." *Advanced Applied Probability*, vol. 28, pp. 784–801, 1996.

[15] G. C. Pflug, "Stochastic minimization with constant stepsize: Asymptotic laws." *SIAM Journal on Control and Optimization*, vol. 24, pp. 655–666, 1986.

[16] J. Kiefer and J. Wolfowitz, "Stochastic estimation of the maximum of a regression function." *Annals of Mathematical Statistics*, vol. 23, no. 3, pp. 462–466, 1952.

[17] G. C. Pflug, "Stepsize rules, stopping times, and their implementation in stochastic quasigradient algorithms." in *Numerical Techniques for Stochastic Optimization*, Y. Ermoliev and R. J.-B. Wets, Eds. New York: Springer-Verlag, 1988, pp. 353–372.

[18] ——, *Optimization of Stochastic Models: The Interface between Simulation and Optimization*. Boston: Kluwer Academic Publishers, 1996.

[19] G. Yin, "A stopped stochastic approximation algorithm." *Systems and Control Letters*, vol. 11, pp. 107–115, 1988.

[20] ——, "Stopping times for stochastic approximation." in *Modern Optimal Control: A Conference in Honor of Solomon Lefschetz and Joseph P. LaSalle*, ser. Lecture Notes in Pure and Applied Mathematics, E. O. Roxin, Ed. New York: M. Dekker, 1989, no. 119, pp. 409–420.

[21] ——, "A stopping rule for the Robbins-Monro method." *Journal of Optimization Theory and Applications*, vol. 67, no. 1, pp. 151–173, October 1990.

[22] P. W. Glynn and W. Whitt, "The asymptotic validity of sequential stopping rules for stochastic simulations." *The Annals of Applied Probability*, vol. 2, no. 1, pp. 180–198, February 1992.

[23] D. W. Hutchison and J. C. Spall, "Stochastic approximation in finite samples using surrogate processes." in *Proceedings of the 43rd IEEE Conference on Decision and Control*, Atlantis, Paradise Island, Bahamas, December 14–17, 2004, pp. 4157–4162.

[24] D. L. Burkholder, "On a class of stochastic approximation processes." *Annals of Mathematical Statistics*, vol. 27, pp. 1044–1059, 1956.

[25] Y. S. Chow and H. Robbins, "On the asymptotic theory of fixed-width sequential confidence intervals for the mean." *Annals of Mathematical Statistics*, vol. 36, no. 2, pp. 457–462, 1965.

[26] R. L. Sielken, Jr., "Some stopping times for stochastic approximation procedures." *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, vol. 27, pp. 79–86, 1973.

[27] D. F. Stroup and H. I. Braun, "On a new stopping rule for stochastic approximation." *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, vol. 60, pp. 535–554, 1982.

[28] J. J. Moré, B. S. Garbow, and K. E. Hillström, "Testing unconstrained optimization software." *ACM Transactions on Mathematical Software*, vol. 7, no. 1, pp. 17–41, 1981.

[29] H. J. Kushner and H. Huang, "Asymptotic properties of stochastic approximations with constant coefficients." *SIAM Journal on Control and Optimization*, vol. 19, no. 1, pp. 87–105, 1981.