ON THE DETERMINATION OF THE STEP SIZE IN STOCHASTIC QUASIGRADIENT METHODS

Georg Ch. Pflug

May 1983
CP-83-25

INTERNATIONAL INSTITUTE FOR APPLIED SYSTEMS ANALYSIS
A-2361 Laxenburg, Austria

PREFACE

The Adaptation and Optimization project at IIASA is largely concerned with the development of algorithmic procedures for stochastic programming problems. In this paper, Professor Georg Pflug of the University of Giessen considers existing methods of controlling the step size in algorithms based on stochastic quasi-gradient techniques, and presents a new, adaptive step-size rule that leads to more rapid convergence of the associated algorithm.

ANDRZEJ WIERZBICKI
Chairman
System and Decision Sciences

ABSTRACT

For algorithms of the Robbins-Monro type, the best choice (from the asymptotic point of view) for the step-size constants a_n is known to be a/n. From the practical point of view, however, adaptive step-size rules seem more likely to produce quick convergence. In this paper a new adaptive rule for controlling the step size is presented and its behavior is studied.

ON THE DETERMINATION OF THE STEP SIZE IN STOCHASTIC QUASIGRADIENT METHODS

Georg Ch. Pflug

1. INTRODUCTION AND HISTORY OF THE PROBLEM

We consider the problem of unconstrained minimization of a function:

$$h(x) = \min!$$

$$x \in \mathbb{R}^k$$

by a stochastic quasigradient method. This implies the use of a steepest-descent (gradient) algorithm for which only statistical estimates of the gradients but not their exact values are available. In particular it is assumed that at each point x and for every $\varepsilon > 0$ we can observe a vector-valued random variable $Y_{x,\varepsilon}$ such that its expectation $E(Y_{x,\varepsilon})$ satisfies

$$|E(Y_{x,\varepsilon}) - \nabla h(x)| < \varepsilon \quad .$$

Sometimes there is even an unbiased estimate Y_x of the gradient, i.e.,

$$E(Y_x) = \nabla h(x) \quad .$$

The unknown minimum point $x_0 = \operatorname{argmin} h$ is estimated by a recursive

sequence $\{X_n\}$ of the form

$$X_{n+1} = X_n - a_n Y_n \quad , \tag{1}$$

where Y_n is a sequence of stochastic quasigradients, i.e., the conditional expectation of Y_n given the history of the approximation process satisfies

$$\lim_{n \to 0} |E(Y_n | X_1, \ldots, X_n) - \nabla h(X_n)| = 0 \quad .$$

The values a_n are the step-size constants and X_1 is an arbitrary starting value.

Univariate recursions of the form (1) were considered for the first time in a pioneering paper by H. Robbins and S. Monro in 1951. These authors examine the problem of recursively estimating the root of an unknown regression function $R(\cdot)$. In the minimization case this amounts to assuming that one can obtain an unbiased estimate of $h'(\cdot)$. If, however, only an unbiased estimate of $h(\cdot)$ [not of $h'(\cdot)$] is available, then $h'(\cdot)$ has to be approximated by numerical differentiation. The corresponding procedure was developed by J. Kiefer and J. Wolfowitz in 1952.

These two methods were generalized to the multidimensional case by Blum (1954). Sacks (1958) proved the asymptotic normality of the properly normalized process X_n in the Robbins-Monro (RM) case. The Kiefer-Wolfowitz (KW) situation is a bit more complicated, since in this case two speeds of approximation influence the asymptotic behavior: the deterministic speed of the approximation of $\nabla h(\cdot)$ by finite differences and the stochastic convergence rate derived from the Central Limit Theorem. It was shown by Fabian (1967) that the rate of convergence can be increased considerably by using higher-order numerical approximations of the gradient. Fabian (1968) also gave a very general result concerning the asymptotic normality of recursive schemes, including the RM and KW processes.

The asymptotic distribution depends on (i) the local properties of $h(\cdot)$ at the minimum point $x_0 = \operatorname{argmin} h(\cdot)$ (or, more precisely, on the Hessian $\nabla^2 h(\theta)$, if this exists); (ii) the covariance

structure of Y_x; (iii) the step-size constants a_n; and (iv) the way in which $\nabla h(\cdot)$ is approximated numerically. It is, however, independent of the starting value X_1. In particular, there is -- from the asymptotic point of view -- a best choice for the constants a_n, namely

$$a_n = \frac{a}{n} \quad .$$

This choice maximizes the convergence rate. Moreover, in the univariate case there is even an optimal choice of the constant a, namely $a = 1/h''(x_0)$, which minimizes the asymptotic variance. However, if only asymptotic convergence is required then the conditions

$$a_n \geq 0 \; ; \quad \sum a_n = \infty \; ; \quad \sum a_n^2 < \infty \tag{2}$$

are sufficient.

The asymptotic approach is really rather unsatisfactory for practical applications. Due to the fact that the asymptotic distribution of X_n is independent of the starting value X_1, the asymptotically optimal choice of the a_n is very bad for finite samples, especially if $|X_1 - x_0|$ is large. This is illustrated by the following example.

1.1. Example. Let $h(x) = |x - x_0|$. We consider, for simplicity, only the deterministic gradient algorithm

$$X_{n+1} = X_n - \frac{a}{n} \operatorname{sgn} (X_n - x_0) \quad .$$

Let N be the first index for which $|X_n - x_0| \leq \varepsilon$. Then N depends exponentially on $|X_1 - x_0|$! Thus we pay for a bad choice of starting value by incurring an exponentially increasing number of necessary steps. This disadvantage disappears if we consider only the asymptotic distribution.

In practice it is preferable to choose the step length a_n such that it depends on the (unknown) distance $|X_n - x_0|$. If $|X_n - x_0|$ is very large the procedure should make large corrections;

the step length should be decreased only when $|X_n - x_0|$ becomes smaller. On the other hand, it is clear that an adaptive choice of the a_n entails greater mathematical difficulty since in such a case the a_n can no longer be treated as constants, but become random variables $a_n = a_n(x_1,\ldots,x_n)$. We should emphasize the fact that methods based on the adaptive choice of step length a_n are quite different from random search techniques. In the adaptive choice approach a_n is a function of the (random) history of the process, whereas in random search methods the a_n are random variables which are independent of the past, but whose distributional parameters may depend on past events.

A first step toward the use of adaptively chosen step lengths in the RM case can be found in a paper by H. Kesten (1958). He proposed to take any deterministic sequence α_n satisfying (2) and set

$$a_1 = \alpha_1$$

$$a_n = \begin{cases} a_{n-1} = \alpha_m \quad \text{(say)} & \text{if} \quad \operatorname{sgn} Y_{n-1} = \operatorname{sgn} Y_n \\ \alpha_{m+1} & \text{otherwise} \end{cases} .$$

Kesten shows that the convergence properties hold in this case, but he was unable to give a mathematical argument to justify his procedure.

A further contribution was made by V. Fabian (1960), who proposes a random linear search after the stochastic gradient has been evaluated. He takes additional random observations of $h(X_n + j\alpha_n Y_n)$, say $V_{n,j}$; $j \geq 1$, and chooses $a_n = j \cdot \alpha_n$ where j is the largest integer such that $V_{n,1} \geq V_{n,2} \geq \ldots \geq V_{n,j}$. With this choice it is also possible to derive the a.s. convergence properties.

A different method of controlling the step size was proposed by Yu. Ermoliev et al. (1981). They assume that an unbiased estimate Z_n of the objective function value $h(X_n)$ is available and define (for $k \in \mathbb{N}$)

$$E_n = \frac{1}{k} \sum_{j=n-k+1}^{n} Z_j \quad .$$

Then a_n is chosen according to the rule

$$a_{n+1} = \begin{cases} a_n/2 & \text{if } |E_{n-s} - E_n| \leq \delta \\ a_n & \text{otherwise} \end{cases} ,$$

where k, s and δ are appropriately chosen constants. This step-size rule is quite plausible since a_n is decreased as soon as it is evident that the mean improvement in the value of the objective function is too small. However, it is also unsatisfactory, for the following reasons:

(i) The procedure cannot distinguish between two different situations: random fluctuations around the minimum point x_0, and small gradients combined with large variances far away from x_0. In the second case the procedure will, with high probability, continue to reduce the step size.

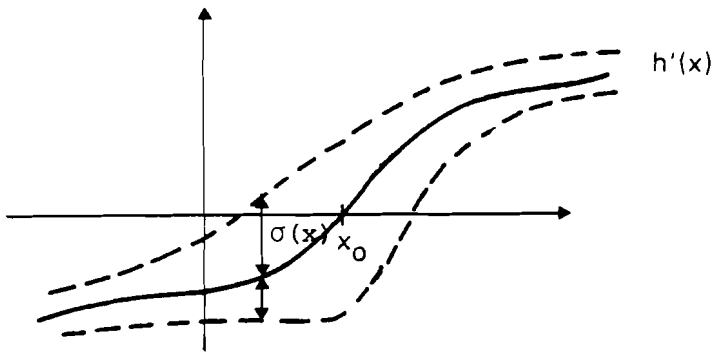
(ii) Divergence caused by overshooting will not be detected.

(iii) An additional estimate of the value of the objective function must be provided.

A new method for controlling the step size is proposed in Section 4 of this paper. However, we shall begin by considering some instructive examples.

2. EXAMPLES

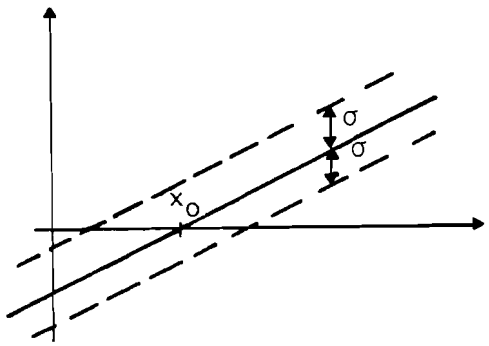
A graphical representation is often quite useful in describing univariate problems. Assume that Y_x is an unbiased estimate of the derivative $h'(x)$ of the objective function $h(x)$. Let $\sigma^2(x) = \mathrm{Var}(Y_x)$. The following diagram shows a possible behavior of $E(Y_x) = h'(x)$ (full line) and the functions $h'(x) \pm \sigma(x)$ (dashed lines).



2.1 Example (Univariate quadratic problem). Let the univariate objective function take the form

$$h(x) = \frac{\alpha}{2}(x - x_0)^2$$

and suppose that the stochastic gradients Y_n have expectations αX_n and variances σ^2. The situation can be described diagrammatically as follows:



To obtain a better understanding of the influence of the choice of step-size constants a_n, we shall for the moment take them to be constant, $a_n \equiv a$. Then, introducing the error variables $Z_n = Y_n - h'(X_n)$, procedure (1) takes the form

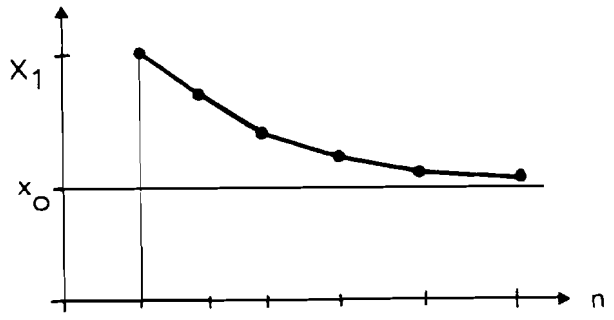
$$X_{n+1} = X_n - a\alpha(X_n - x_0) + aZ_n$$

or, equivalently, with $c = a\alpha$,

$$(X_n - x_0) = (X_1 - x_0)(1 - c)^{n-1} + a \sum_{i=1}^{h-1} (1 - c)^{i-1} Z_{n-i} \quad .$$

X_n is the superposition of a nonrandom drift

$$x_0 + (1 - c)^{n-1}(X_1 - x_0)$$



and the zero-mean stochastic process

$$a \sum_{i=1}^{n-1} (1 - c)^{i-1} Z_{n-i} \quad .$$

The above can be approximated by the stationary process

$$U_n = a \sum_{i=1}^{\infty} (1 - c)^{i-1} Z_{n-i} \quad .$$

U_n is an AR(1) process, since it is a stationary solution of the stochastic difference equation

$$U_{n+1} = U_n(1 - c) + aZ_n$$

with moments

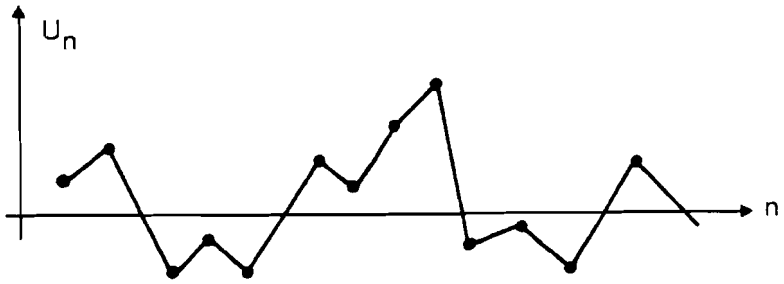
$$E(U_n) = 0$$

$$\mathrm{Var}(U_n) = \frac{\sigma^2 a^2}{2c - c^2}$$

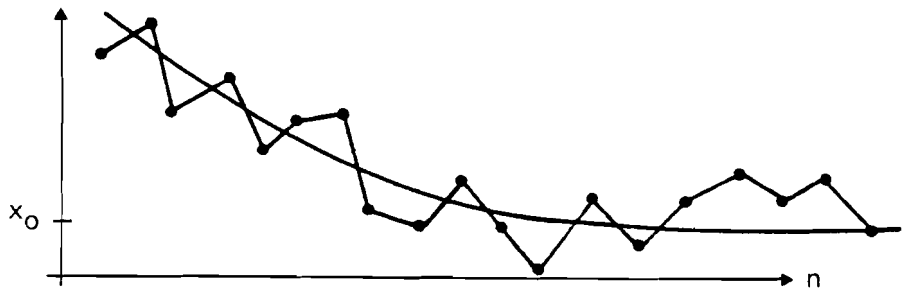
$$\text{Cov}\ (U_n, U_{n-s}) = \frac{\sigma^2 a^2 (1-c)^s}{2c - c^2}$$

$$\text{Corr}\ (U_n, U_{n-s}) = (1-c)^s \quad .$$

Taking a trajectory from this process:



we obtain a typical picture of the process X_n by superposition.



Analogously, the gradient process

$$Y_n = \alpha(X_n - x_0) + Z_n$$

$$= \alpha(1-c)^{n-1}(X_1 - x_0) + c\sum_{i=1}^{n-1}(1-c)^{i-1} Z_{n-i} + Z_n$$

can be approximated by the superposition of a deterministic component

$$\alpha(1-c)^{n-1}(X_1 - x_0)$$

and the zero-mean stationary process

$$V_n = c \sum_{i=1}^{\infty} (1-c)^{i-1} Z_{n-i} + Z_n \quad .$$

V_n is an ARMA(1,1) process since

$$V_{n+1} - (1-c)V_n = (2c-1)Z_n + Z_{n+1} \quad .$$

The moments are

$$E(V_n) = 0$$

$$\mathrm{Var}(V_n) = \frac{\sigma^2}{1-(c/2)}$$

$$\mathrm{Cov}(V_n, V_{n-s}) = \sigma^2 c(1-c)^{s-1} + (1-c)^s \frac{\sigma^2 c}{2-c} \quad \text{for} \quad s \geq 1 \quad .$$

Note that if $\sigma^2 = 0$ then $X_n \to x_0$ and $Y_n \to 0$ for any c such that $0 < c < 1$. Hence there is no need for a reduction of the step size in the deterministic situation unless $a \geq \frac{1}{\alpha}$. If, however, $\sigma^2 > 0$, then

$$\lim_{n\to\infty} \mathrm{Var}\ (X_n - x_0) = \frac{\sigma^2 a^2}{2c - c^2}$$

$$\lim_{n\to\infty} \mathrm{Var}\ (Y_n) = \sigma^2 + \frac{\sigma^2 c}{2-c}$$

and the process X_n will oscillate around the solution x_0 unless we reduce the step size. The asymptotic variance decreases as $a \to 0$, but on the other hand a small value of a results in slow convergence of the deterministic part. What we can learn from this example is that *the step size should be reduced if it is evident that the deterministic drift has fallen to zero and the fluctuation of* X_n *is due only to the random element (the stationary process* U_n*).*

2.2 Example (Multidimensional quadratic problem). Let the objective function be of the form

$$h(x) = \frac{1}{2} x' A x \quad ,$$

where A is a positive definite matrix. Without loss of generality, x_0 is assumed to be zero. The error variables Z_n are independent and identically distributed with expectation 0 and covariance matrix C. Again we let the a_n remain constant and equal to a. The process (1) takes the form

$$X_{n+1} = (I - aA)X_n - aZ_n$$

or, equivalently,

$$X_n = (I - aA)^{n-1}X_1 + a\sum_{i=1}^{n-1}(I - aA)^{i-1}Z_{n-i} \quad .$$

Once again, X_n can be approximated by the sum of a deterministic drift and the following vector-valued AR(1) process:

$$U_n = a\sum_{i=1}^{\infty}(I - aA)^{i-1}Z_{n-i} \quad .$$

This process is well-defined if a is smaller than the inverse of the largest eigenvalue of A. The gradient process

$$Y_n = AX_n + Z_n$$

can be rewritten as

$$Y_n = A(I - aA)^{n-1}X_1 + aA\sum_{i=1}^{n-1}(I - aA)^{i-1}Z_{n-i} + Z_n \quad .$$

The stochastic part of this process can be approximated by the stationary vector-valued ARMA(1,1) process

$$V_n = aA\sum_{i=1}^{\infty}(I - aA)^{i-1}Z_{n-i} + Z_n \quad , \qquad (3)$$

which fulfills the difference equation

$$V_{n+1} - (I - aA)V_n = (2aA - I)Z_n + Z_{n+1} \quad .$$

Clearly (3) determines the covariance structure of the process. Instead of considering the autocovariance matrices $E(V_n V'_{n-s})$ we calculate the following two numbers:

$$E(\|V_n\|^2) = \operatorname{tr}\ (C(a^2A^2 \sum_{i=0}^{\infty} (I - aA)^{2i} + I))$$

$$E(V'_{n+1}V_n) = \operatorname{tr}(C(a^2A^2 \sum_{i=1}^{\infty} (I - aA)^{2i-1} + aA)) \quad ,$$

where tr (B) denotes the trace of matrix B. Use of the formula

$$\sum_{i=0}^{\infty} B^i = (I - B)^{-1} \quad ,$$

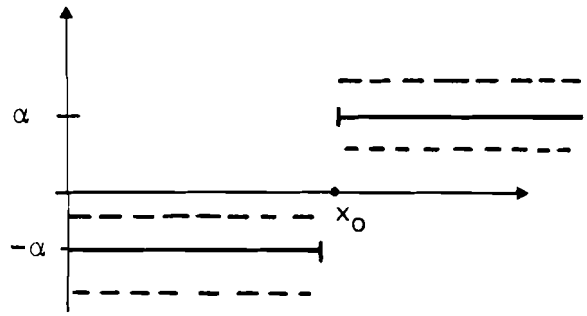
which is valid for positive definite matrices B with all eigenvalues less than unity, leads to the simplifications

$$E(\|V_n\|^2) = \operatorname{tr}\ (C(I - \tfrac{a}{2}A)^{-1}) \tag{4}$$

$$E(V'_{n+1}V_n) = \operatorname{tr}\ (C\,a\,A(2I - (2I - aA)^{-1})) \quad . \tag{5}$$

As in the univariate case, the approximation process X_n converges (for fixed a) only if C = 0, i.e., if the procedure is a deterministic one.

2.3 Example (Nonsmooth univariate case). In this example we consider the objective function $h(x) = \alpha|x - x_0|$ and assume that the error variables Z_n are again independent and identically distributed with expectation zero and variance σ^2. Furthermore, we assume that the distribution of the Z_n is symmetrical around zero and possesses finite moments of any order. Since $h'(x) = \alpha \operatorname{sgn}(x - x_0)$ if $x \neq x_0$, the problem may be represented graphically as follows:



Without loss of generality, we can assume that $x_0 = 0$. The stochastic approximation process (1) is then given by

$$X_{n+1} = X_n - a\alpha \text{ sgn } (X_n) + aZ_n \quad . \tag{6}$$

To which limiting distribution does this recursion converge, if any? Or, equivalently, what are the stationary distributions of the Markovian process (6)? Let G be the c.d.f. of the variables Z_n. A stationary distribution F must clearly fulfill

$$F = \tilde{F} * G \quad , \tag{7}$$

where

$$\tilde{F}(u) = \begin{cases} F(u-c) & u < -c \\ F(u+c) - F(0) + F(u-c) & -c \leq u \leq c \\ F(u+c) & u \geq c \end{cases}$$

and $c = a\alpha$.

This convolution equation is best handled by considering the Fourier transforms. Let X be distributed according to F and let

$$\psi_1(t) = E(e^{itX} 1_{\{x>0\}})$$

$$\psi_2(t) = E(e^{itX} 1_{\{x<0\}}) \quad .$$

Then (7) can be rewritten as

$$(\psi_1(t)e^{-ict} + \psi_2(t)e^{ict})\phi(at) = \psi_1(t) + \psi_2(t) \quad , \tag{8}$$

where $\phi(\cdot)$ is the characteristic function of the Z_n. We assume that $\phi(\cdot)$ does not vanish anywhere, i.e., Z_n is non-lattice. Any solution of (7) must be symmetric, i.e., such that

$$\psi_2(t) = \psi_1(-t)$$

$$\psi_1(0) = 1/2 \quad .$$

The functional equation (8) can then be written as

$$\log \psi_1(t) - \log \psi_1(-t) = \log (\phi^{-1}(at) - e^{ict}) - \log (e^{-ict} - \phi^{-1}(at)) \quad . \tag{9}$$

Since, from the moment conditions on Z_n,

$$\phi(t) = 1 - \frac{1}{2} t^2\sigma^2 + O(t^2) \quad ,$$

taking the derivative of equation (9) at the point 0 leads to

$$4\ \psi_1'(0) = \frac{-1}{ic} (a^2\sigma^2 + c^2) \quad .$$

Hence $E(X\ 1_{\{x>0\}}) = \dfrac{a\sigma^2 + c^2}{4c}$ and therefore

$$E(|X|) = \frac{c^2 + a^2\sigma^2}{2c} \quad . \tag{10}$$

On taking higher derivatives we see that $E(|X|^k)$ is uniquely determined for odd k. (For even k the kth derivative vanishes on both sides.) Let $2\beta_k = E(|X|^{2k+1})$; $k \geq 0$. Thus

$$\int_0^\infty x^{2k+1}\, dF(x) = \beta_k \quad .$$

Or, by introducing the distribution function

$$H(x) = \frac{\int_0^x u\, d\, F(u)}{\beta_1}$$

we obtain

$$\frac{\beta_k}{\beta_0} = \int_0^\infty x^k \, dH(\sqrt{x}) \quad . \tag{11}$$

We see that $H(\cdot)$ and consequently $F(\cdot)$ is uniquely determined by the sequence $\{\beta_k\}$ *if the corresponding moment problem (11) has a unique solution.*

However, the author was unable to solve (10) explicitly even for a normal error distribution. It also seems to be difficult to determine the even moments, especially the variance of the symmetric solution.

Nevertheless, we can still take the first absolute moment as a measure of dispersion. From (10) it can be seen that in this case X_n does not converge to zero unless $a \to 0$ *even when* $\sigma^2 = 0$. This is the important difference between examples 2.1 and 2.3. The asymptotic dispersion (10) can be viewed as a superposition of a "deterministic part" $c/2$ and a stochastic part $a^2\sigma^2/2c$.

3. DETERMINISTIC STEEPEST-DESCENT METHODS

In this section we study step-size rules for deterministic steepest-descent methods. Let $h(\cdot)$ be a quasiconvex, continuous function defined on $\mathbb{R}^k$. This means that the sets

$$S(x) = \{y \mid h(y) \leq h(x)\}$$

are closed, convex sets. We assume that h is continuously differentiable for $x \neq x_0$ with gradient $\nabla h(x) \neq 0$ for $x \neq x_0$ and that $S(x_0) = \{x_0\}$, i.e., $x_0 = \operatorname{argmin} h(\cdot)$. An algorithm of the form

$$x_{n+1} = x_n - a_n \nabla h(x_n) \tag{12}$$

is known as a *steepest-descent algorithm*. In mathematical programming the step-size constants a_n are usually determined from

$$h(x_n - a_n \nabla h(x_n)) = \inf_a (x_n - a \nabla h(x_n)) \quad , \tag{13}$$

i.e., they are found by a line search. However, this type of procedure cannot be used in stochastic gradient methods since for these problems only a stochastic estimate of the optimal a would be available. Such an estimate would require additional observations as well as contradicting the basic philosophy of stochastic approximation: Not to waste too much time trying to get a better estimation of the next step when the current point is still a long way from the solution.

Let us therefore concentrate on those step-size rules which depend only on n (the number of the step) and the history $(x_1,\ldots,x_{n-1})$ of the iteration process, and which do not require any additional evaluation of the objective function.

One important subclass of these rules is formed by sequences a_n which depend only on n. The corresponding convergence properties are given by the following theorem.

3.1 Theorem. Let the function h be defined as above and suppose that for every $\varepsilon > 0$

$$\inf_{\|x - x_0\| \geq \varepsilon} \frac{\langle x - x_0 , \nabla h(x) \rangle}{\|\nabla h(x)\|^2} \geq \eta_1(\varepsilon) > 0$$

$$\inf_{\|x - x_0\| \geq \varepsilon} \langle x - x_0 , \nabla h(x) \rangle \geq \eta_2(\varepsilon) > 0 \quad .$$

If

$$a_n \geq 0 \quad ; \quad a_n \to 0 \quad ; \quad \sum a_n = \infty \tag{14}$$

then the iteration $\{x_n\}$ given by (12) converges to x_0.

Proof. Without loss of generality we can assume that $x_0 = 0$. If $\|x_n\| \geq \varepsilon$ then

$$\|x_{n+1}\|^2 = \|x_n\|^2 - 2a_n \langle x_n , \nabla h(x_n) \rangle + a_n^2 \|\nabla h(x_n)\|^2$$

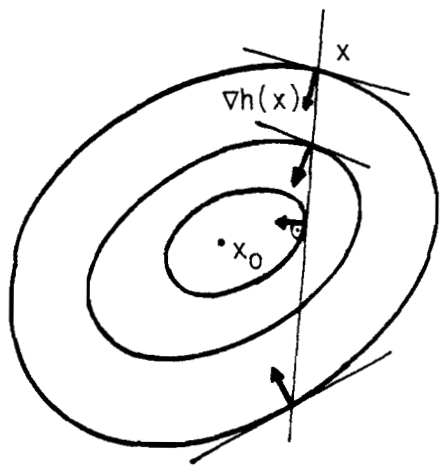
$$\leq \|x_n\|^2 - 2a_n \eta_2(\varepsilon) \left(1 - \frac{a_n}{\eta_1(\varepsilon)}\right) \quad .$$

Thus, for large n, $\|x_n\|^2 \geq \varepsilon$ implies that $\|x_{n+1}\|^2 \leq \|x_n\|^2 - 2a_nK$, where K is a constant depending only on ε. If, however, $\|x_n\| \leq \varepsilon$ then, by the continuity of $\nabla h(\cdot)$, $\|\nabla h(x_n)\|$ is bounded and hence $\|x_{n+1}\| \leq 2\varepsilon$ for large n. From $\sum a_n = \infty$ we can conclude that $\limsup \|x_n\| \leq \varepsilon$. Since ε was arbitrary the theorem is proven.

Before trying to construct a more adaptive step-size rule we first draw attention to the following lemma.

3.2 Lemma. Let h be convex and twice-differentiable. The function $a \mapsto \langle \nabla h(x) , \nabla h(x-a\nabla h(x)) \rangle$ is monotonically decreasing and vanishes if and only if a is the solution of (13).

Proof. The assertions follow easily from simple calculus. The situation can be illustrated by the following figure.



Since $\langle \nabla h(x) , \nabla h(x - a\nabla h(x)) \rangle < 0$ implies that a is larger than the optimal a given by (12), we are led to the following heuristic step-size rule:

$$a_{n+1} = \begin{cases} a_n & \text{if } \langle \nabla h(x_n), \nabla h(x_{n+1}) \rangle > 0 \\ a_n/2 & \text{if } \langle \nabla h(x_n), \nabla h(x_{n+1}) \rangle \leq 0 \end{cases} . \tag{15}$$

The decrease by a factor of 1/2 is somewhat arbitrary; any factor q $(0 < q < 1)$ could be taken.

With this rule we can state and prove a convergence theorem.

3.3 Theorem. Let h be defined as in Theorem 3.1. In addition, we assume that it is inf-compact (i.e., the sets S(x) are compact) and $\|\nabla h(x)\| \leq K\|x - x_0\|$. The iteration

$$x_{n+1} = x_n - a_n \nabla h(x_n) \quad ,$$

where a_n is given by (15), converges to x_0 for every starting value $(x_1, a_1 > 0)$.

Proof. We show first that, for $a > 0$, $\langle \nabla h(x), \nabla h(x - a\nabla h(x)) \rangle > 0$ implies that $h(x) > h(x - a\nabla h(x))$. By virtue of the quasiconvexity of h

$$S(x) \subseteq \{y \mid \langle y - x, \nabla h(x) \rangle \leq 0\} \quad .$$

Let $z = x - a\nabla h(x)$. Suppose that $h(z) \geq h(x)$, i.e., $x \in S(z)$. Then $0 \geq \langle z - x, \nabla h(z) \rangle = -a \langle \nabla h(x), \nabla h(z) \rangle > 0$ and the theorem is proven by contradiction.

Now consider the sequence a_n. If $\sum a_n < \infty$ then x_n converges. Let the limit be y. If $\|\nabla h(y)\| > 0$ then $\langle \nabla h(x), \nabla h(x - a\nabla h(x)) \rangle > 0$ for small a in a neighborhood of y. Thus y can be the limit only if $y = x_0$. If $a_n \to 0$ but $\sum a_n = \infty$ then x_n converges to x_0 by Theorem 3.1. If a_n does not converge to zero, then there is an index $N \in \mathbb{N}$ such that $a_n \equiv a$ for $n \geq N$ and $\langle \nabla h(x_n), \nabla h(x_{n+1}) \rangle > 0$. Hence $h(x_n)$ is decreasing for $n \geq N$. The sequence $\{x_n\}$ has a cluster point y since h is inf-compact. Let $z = y - a\nabla h(y)$. Then, by continuity, $h(z) = h(y)$ and $\langle \nabla h(z), \nabla h(x) \rangle \geq 0$. This implies that either $x = z$ or $\nabla h(x) = 0$, but in any case $x = x_0$.

We finish this section by looking at two examples.

3.4 Example. Let $h(x) = \|x\|$. Then $\nabla h(x) = \frac{x}{\|x\|}$ if $x \neq 0$. Since $\frac{\langle x, \nabla h(x) \rangle}{\|\nabla h(x)\|^2} = \|x\|$ and $\langle x, \nabla h(x) \rangle = \|x\|$ the assumptions of Theorems 3.1 and 3.3 are satisfied. The recursion (12) then takes

the form

$$x_{n+1} = x_n\left(1 - \frac{a_n}{\|x_n\|}\right) \quad .$$

Or, writing $x_n = v_n x_1$,

$$v_{n+1} = v_n - \frac{a_n}{\|x_1\|}\,\mathrm{sgn}\,(x_n) \quad .$$

If $\{a_n\}$ satisfies (14) and $a_{n+1} \leq a_n \leq 2a_{n+1}$ then

$$\|x_n\| = O(a_n) \quad .$$

If $\{a_n\}$ is determined by (15) then

$$\|x_n\| = O(2^{-n})$$

and we see that (15) is much better than (14) in this case.

3.5 Example. Let $h(x) = \frac{1}{2}x'Ax$, where A is a positive definite matrix. Then $\nabla h(x) = Ax$ and

$$\frac{\langle x, \nabla h(x)\rangle}{\|\nabla h(x)\|^2} = \frac{x'Ax}{x'A^2x} \geq \frac{m}{M^2} \quad ,$$

where M and m are the largest and smallest eigenvalues of A, respectively. Similarly,

$$\langle x, \nabla h(x)\rangle = x'Ax \geq m\|x\|^2$$

and thus the assumptions of Theorems 3.1 and 3.3 are satisfied. Choosing the constants a_n according to (14) leads to

$$x_n = \Pi(I - a_n A)x_1 \quad .$$

This implies

$$\|x_{n+1}\| = O\left(\prod_{j=1}^{n}(1-a_j m)\right) \quad .$$

Choosing, for instance, $a_j = a/j$ we obtain

$$\|x_n\| = o(n^{-am}) \quad .$$

For this example, rule (15) can be written as

$$a_{n+1} = \begin{cases} a_n & \text{if } x_n'A^2x_n - a_n x_n'A^3x_n > 0 \\ a_n/2 & \text{otherwise} \quad . \end{cases}$$

It is evident that the constants a_n can never fall below $m^2/2M^3$. Thus, if the objective function is quadratic then the constants determined by rule (15) do not converge to zero and the rate of convergence of the iteration is at least

$$\|x_n\| = 0\left(\left(1 - \frac{m^3}{2M^3}\right)^n\right) \quad .$$

Again, rule (15) is superior to (14).

4. A STOCHASTIC STEP-SIZE RULE

A stochastic version of rule (15) is presented in this section. We once again consider the approximation process (1)

$$X_{n+1} = X_n - a_n Y_n \quad ,$$

where

$$E(Y_n | X_1, \ldots, X_n) = \nabla h(X_n) \quad .$$

It would be possible to approximate $\langle \nabla h(X_n), \nabla h(X_{n+1}) \rangle$ by $Y_n'Y_{n+1}$. However, it would be incorrect to compare this quantity with zero; we should rather look at the expectation of this value for the stationary distribution of (1). Since this distribution depends on $h(\cdot)$ we have to make some additional assumptions.

We assume that $h(\cdot)$ is quadratic, i.e., $h(x) = \frac{1}{2} x'Ax$, since this is the most important case, and also that the covariance

matrix C of the stochastic gradients Y_n is independent of X_n. By (5) the expectation of $Y_n' Y_{n+1}$ under the stationary distribution is

$$E(V_{n+1}' V_n) = \mathrm{tr}\ (C\, a A (2I - (2I - aA)^{-1})) \quad .$$

To simplify the rule we replace $aA(2I - (2I - aA)^{-1}) = aA\left(2I - \frac{1}{2} \sum_{i=0}^{\infty} (aA)^i\right)$ by $\frac{3}{2}$ aA, neglecting terms of higher order in A.

The quantity tr(aCA) can be estimated by taking a random direction D_n at X_n and estimating the gradient at $X_n + aD_n$. To be more explicit, let Y_n^1 and Y_n^2 be two independent estimates of $\nabla h(X_n)$. Let $D_n = \frac{1}{2}(Y_n^1 - Y_n^2)$. Then $E(D_n | X_1, \ldots, X_n) = 0$ and $\mathrm{Cov}\ (D_n) = \frac{1}{2} C$. Let $\tilde{X}_{n+1} = X_n + aD_n$ and $\tilde{Y}_{n+1}$ be an estimate of $\nabla h(\tilde{X}_{n+1})$, i.e., $\tilde{Y}_{n+1} = A\tilde{X}_{n+1} + \tilde{Z}_{n+1}$. Then

$$\begin{aligned} E(D_n' \tilde{Y}_{n+1}) &= E(D_n' (A(X_n + aD_n) + \tilde{Z}_{n+1})) \\ &= aE(D_n' A D_n) = \frac{a}{2} \mathrm{tr}(AC) \quad . \end{aligned}$$

A more parsimonious use of the random variables can be achieved by setting $Y_n = \frac{1}{2}(Y_n^1 + Y_n^2)$, which has the advantage of reducing the covariance matrix by a factor of 1/2. The step-size reduction is then based on the comparison

$$\frac{1}{k} \sum_{n=1}^{k} Y_{n+1}' Y_n < \frac{3}{2} E(D_n' \tilde{Y}_{n+1}) \quad .$$

This method is summarized in algorithm 4.1. The notation $Y_n := Y(X_n)$ is used to indicate an independent function call of the gradient estimate. In particular, $Y^1 := Y(X_n)$; $Y^2 := Y(X_n)$ does not mean that $Y^1 = Y^2$.

4.1 Algorithm

Step 1. Choose starting values X_1, a

Step 2. Set n := 0; N := 0. Go to Step 4

Step 3. Set $n := n+1$

Observe $Y_n^1 = Y(X_n)$; $Y_n^2 = Y(X_n)$

Set $Y_n := \frac{1}{2}(Y_n^1 + Y_n^2)$

$D_n := \frac{1}{2}(Y_n^1 - Y_n^2)$

$X_{n+1} := X_n - aY_n$

$\tilde{X}_{n+1} := X_n - aD_n$

$\tilde{Y}_{n+1} := Y(\tilde{X}_{n+1})$

$\delta_n := D_n' \tilde{Y}_{n+1}$

Step 4. Perform Step 3 twice

Step 5. If

$$\frac{1}{n-N} \sum_{k=N}^{n-1} Y_{k+1}' Y_k \leq \frac{3}{2} \frac{1}{n-N+1} \sum_{k=N}^{n} \delta_k$$

then set $a := a/2$; $N := n$; stop, if $a \leq \varepsilon$; perform Step 3 twice and return to Step 5

Otherwise perform Step 3 once and then return to Step 5

It is important to notice that if the procedure is deterministic $D_n = 0$ and $\delta_n = 0$. Therefore algorithm 4.1 is very close to rule (15) except that the algorithm uses the arithmetic mean of the inner products $Y_{k+1}' Y_k$.

In order to reduce this difference we could use a variant of the algorithm which employs a sequential t-test instead of a simple comparison of mean values.* This algorithm is presented below in more formal notation, which omits iteration indices.

4.2 Algorithm

Step 1. Choose starting values X, a

Step 2. Set $n := 0$; $k := 1$; $\gamma = 0$; $\eta = 1$

*For the theory of sequential tests see Govindarajulu (1975).

Step 3. Set $n := n+1$; $k := k+1$

Observe $Y^1 := Y(X)$; $Y^2 := Y(X)$

Set $V := Y$

$$Y := \frac{1}{2}(Y^1 + Y^2)$$

$$D := \frac{1}{2}(Y^1 - Y^2)$$

$$X := X - aY$$

$$\tilde{X} := X - aD$$

$$\tilde{Y} := Y(\tilde{X})$$

$$\delta := (\frac{n-1}{n})\delta + (\frac{1}{n}) \cdot D'\tilde{Y}$$

$$\gamma := \begin{cases} (\frac{k-2}{k-1})\gamma + \frac{1}{k-1} Y'V & \text{if } k \neq 1 \\ \gamma & \text{if } k = 1 \end{cases}$$

$$\eta := \begin{cases} (\frac{k-2}{k-1})\eta + \frac{1}{k-1} (Y'V)^2 & \text{if } k \neq 1 \\ \eta & \text{if } k = 1 \end{cases}$$

If $\gamma \leq \frac{3}{2}\delta - \frac{t}{k}(\eta-\gamma^2)$

then set $a := a/2$; $k := 0$; $\delta := \delta/2$;

stop if $a \leq \varepsilon$; go to Step 3

If $|\gamma - \frac{3}{2}\delta| < \frac{t}{k}(\eta-\gamma^2)$

then go to Step 3

If $\gamma \geq \frac{3}{2}\delta + \frac{t}{k}(\eta-\gamma^2)$

then set $k := 0$; $n := 0$; go to Step 3

The constant t represents the upper α-fractile of a standard normal distribution and should be set to a value between 1 and 3. It is shown below that this algorithm results in a convergent iteration process.

4.3 Theorem. Let $h(x) = \frac{1}{2} x'Ax$ and let the covariance matrix of the gradient estimations Y_n be constant. Then the recursive sequence (1) with step sizes given by algorithm 4.1 or 4.2 converges a.s. to zero.

Proof. We must consider two different cases. If $a_n \to a > 0$ then the distribution of the Y_n approaches the stationary distribution. Hence, by ergodicity,

$$\frac{1}{n} \sum Y'_{k+1} Y_k \to E(Y'_{n+1} Y_n) = \operatorname{tr} (CaA(2I - (2I - aA)^{-1}))$$

$$< \frac{3}{2} \operatorname{tr} (aCA) = E (D_n' \tilde{Y}_{n+1}) \quad .$$

Hence, with probability 1, there must be an index N such that a_N has to be reduced. Hence $a_n \not\to 0$ is impossible.

If $a_n \to 0$ then $\frac{1}{n} \sum Y'_{k+1} Y_k \to 0$ and hence $E(Y'_{k+1} Y_k) = 0$. This however implies that $X_n \to 0$.

REFERENCES

Blum, J.R. (1954). Multidimensional stochastic approximation method. Ann. Math. Statist. 25, 737-744.

Ermoliev, Yu., G. Leonardi and J. Vira (1981). The Stochastic Quasigradient Method Applied to a Facility Location Model. Working Paper WP-81-14, International Institute for Applied Systems Analysis, Laxenburg, Austria.

Fabian, V. (1960). Stochastic approximation methods. Czech. Math. J. 10, 123-159.

Fabian, V. (1967). Stochastic approximation of minima with improved asymptotic speed. Ann. Math. Statist. 38, 191-200.

Fabian, V. (1968). On asymptotic normality in stochastic approximation. Ann. Math. Statist. 39, 1327-1332.

Govindarajulu, Z. (1975). Sequential Statistical Procedures. Academic Press, New York.

Kesten, H. (1958). Accelerated stochastic approximation. Ann. Math. Statist. 29, 41-59.

Kiefer, H. and J. Wolfowitz (1952). Stochastic estimation of the maximum of a regression function. Ann. Math. Statist. 23, 462-466.

Robbins, H. and S. Monro (1951). A stochastic approximation method. Ann. Math. Statist. 22, 400-407.

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