ON STEEPEST DESCENT

by

A. A. Goldstein
University of Washington

Mathematical Note No. 394
Mathematics Research Laboratory
BOEING SCIENTIFIC RESEARCH LABORATORIES
February 1965
ABSTRACT

This paper continues studies initiated in [1] concerning iterative methods of driving the gradient to 0. In [1], we were concerned with functions which were twice differentiable; in this paper only first derivatives will be assumed. Other results included are a fixed point theorem for "gradient" operators and a simple proof of the classical method of steepest descent.
Let $H$ be a Hilbert space, $x_0$ an arbitrary point of $H$ and $f$ a functional on $H$. Let $S$ denote the level set of $f$ at $x_0$, viz: $S = \{ x \in H : f(x) \leq f(x_0) \}$. Let $f'(x, \cdot)$ denote the Fréchet derivative of $f$ at $x$, and $Vf(x)$ its representor in $H$. Set $\Delta(x, \rho) = f(x) - f(x - \rho \varphi(x))$, $g(x, \rho) = \Delta(x, \rho) / |Vf(x), \varphi(x)|\rho$ and fix $\delta$, $0 < \delta \leq 1/2$. Here $\varphi$ denotes a bounded map from $S$ to $H$, satisfying $[Vf(x), \varphi(x)] \geq 0$, such that given $\epsilon > 0$ there exists $\delta > 0$ for which $[Vf(x), \varphi(x)] < \delta$ implies $||Vf(x)|| < \epsilon$.

For example, see Remark 4, below.

**Theorem 1.** Assume that on $S$ $f$ is Fréchet differentiable and bounded below, and that $Vf$ is uniformly continuous on $S$. Set $x_{k+1} = x_k$ when $[Vf(x_k), \varphi(x_k)] = 0$. Otherwise choose $\rho_k$ so that $\delta \leq g(x_k, \rho_k) \leq 1 - \delta$ when $g(x_k, 1) < \delta$, or $\rho_k = 1$ when $g(x_k, 1) \geq \delta$, and set $x_{k+1} = x_k - \rho_k \varphi(x_k)$. Then:

(a) $Vf(x_k)$ converges to 0 while $f(x_k)$ converges downward to a limit $L$. 
(b) If the sequence \{x_k\} has cluster points then every cluster point satisfies \(\nabla f(z) = 0\). If \(\varphi = \nabla f\) has finitely many zeros on \(S\) and \(S\) is compact then \{x_k\} converges.

(c) If \(S\) and \(f\) are convex, and \(\varphi(x) = \nabla f(x)\), then
\[
L = \inf\{f(x) : x \in H\}.
\]
If \{x_k\} has weak cluster points then all such minimize \(f\).

Proof. (a) We have that 
\[
\Delta(p,x) = p[\nabla f(\xi), \varphi(x)] = p[\nabla f(\xi), \varphi(x)] + p[\nabla f(\xi) - \nabla f(x), \varphi(x)] \text{ where } \xi \text{ lies on the open line segment joining } x \text{ and } x - p\varphi(x).
\]
Thus \(g(p,x) = 1 + \frac{[\nabla f(\xi) - \nabla f(x), \varphi(x)]}{[\nabla f(x), \varphi(x)]}\). Since \(\nabla f\) is continuous at \(x\) and \(||\xi - x|| \leq p||\varphi(x)||\), \(g(0,x) = 1\). If \(g(l|x) < \delta\), then \(g\) being continuous takes on all values between 1 and \(\delta\); there exists therefore numbers \(p > 0\) so that \(\delta \leq g(p,x) \leq 1 - \delta\). Set \(\nabla f(x_k) = \nabla f_k, \varphi(x_k) = \varphi_k\), and assume \(\nabla f_k \neq 0\), and \(x_k \in S\). Thus 
\[
\Delta(x_k,p_k) = p_k g(x_k,p_k)[\nabla f_k,\varphi_k] \leq p_k \delta[\nabla f_k,\varphi_k] > 0.
\]
Thus \(f(x_{k+1}) < f(x_k)\) and \(x_{k+1} \in S\). Assume \(\nabla f_k,\varphi_k \neq 0\) This implies that there exists a subsequence \{x_k\} and a number \(c > 0\) such that \([\nabla f_k,\varphi_k] \geq c\).

It follows moreover that \(\{p_k\}\) is bounded away from 0. For if not take a thinner subsequence \(\{p_k\}\), if necessary, such that \(p_k \to 0\). Since \(\varphi_k\) is bounded on \(S\), \(||x_k - \xi_k|| \to 0\). By the uniform continuity of \(\nabla f\) on \(S\), it follows that
\[
([\nabla f(\xi_k) - \nabla f(x_k), \varphi(x_k)]) \to 0,
\]
and therefore \(g(p_k,x_k) \to 1\), contradicting that \(g(p_k,x_k) \leq 1 - \delta\) for all \(k\). There exists
therefore a number \( q > 0 \) such that \( p_k \geq q \). Hence
\( \Delta(x_k, p_k) \geq q \delta \), from which we may contradict the hypothesis that
\( f \) is bounded below. Thus \( [\nabla f, x_k^p] \to 0 \), showing (a).

(b) Let \( z \) be a cluster point of the sequence \( \{x_k\} \). Then
\( [\nabla f(z), \varphi(z)] = 0 \), whence \( \nabla f(z) = 0 \). Thus the number of roots
of \( \nabla f(z) \) is \( \geq \) the number of cluster points of \( \{x_k\} \). Therefore
if \( \nabla f \) has a unique root on \( S \), then \( \{x_k\} \) converges to it. If
the roots of \( \nabla f \) are finite in number we may suppose the cluster
points of \( \{x_k\} \) to be thus also. Let \( x_1 \) be a cluster point and
let \( x_2 \) be a closest neighboring cluster point to \( x_1 \). Set
\( \epsilon = ||x_1 - x_2|| \) and let \( N(x_1) \) denote spheres of radius \( \epsilon/3 \)
centered at \( x_1 \). \( H \cup N(x_1) \) contains a finite number of points
at most, say \( m \). Since \( ||x_{k+1} - x_k|| \) converges to \( 0 \), we can
choose \( k \) so that \( x_k \in N(x_1) \) and \( ||x_{k+1} - x_k|| < \epsilon/3m \). But
this implies that \( H \cup N(x_1) \) contains more than \( m \) points, which
is a contradiction. This contradiction persists unless we suppose
that \( \{x_k\} \) has a unique cluster point.

(c) We need the inequality \( f(y) \geq f(x) + [\nabla f(x), y-x] \) which
is valid for all \( x \) and \( y \) in \( S \). By the convexity of \( f, \)
\( [f(x + t(y - x)) - f(x)]/t \leq f(y) - f(x) \) for \( t \in [0,1] \). By the
differentiability of \( f, \lim_{t \to 0}[f(x + t(y - x)) - f(x)]/t = [\nabla f(x), y-x] \).
The remainder of the proof is as in [1].

Remarks

1. The hypotheses for the first part of (b) can be changed to:
\( \nabla f \) and \( \varphi \) continuous on \( S \), and \( [\varphi(x), f(x)] \geq 0 \) for all \( x \in S \),
\[(\varphi(x), f(x)) = 0 \text{ only if } \nabla f(x) = 0.\]

2. In (b) if the sequence \(\{x_k\}\) has an isolated cluster point then it converges.

3. The inequality in the proof of (c) may sometimes be used in (b) for terminating computations. We have
\[f(x_k) > f(z) \geq f(x_k) + [\nabla f(x_k, z - x_k)] \geq f(x_k) - \|\nabla f_k\|D,\]
where \(D\) is the diameter of \(S\).

4. An example for \(\varphi\) is the following. Let \(H = \mathbb{F}_n\) and let \(\delta_i, (1 \leq i \leq n)\), denote the rows of the identity matrix. Choose \(l_o\) so that
\[\|\nabla f(x)\| \geq (1 - \rho)\|\nabla f(x)\|, \quad (1 \leq i \leq n).\]
Set
\[\varphi(x) = \|\nabla f(x)\|\delta_i,\]
then
\[\|\nabla f(x)\| \geq \|\nabla f(x)\|^{2/(1 - \rho)}.\]

**Corollary.** Let \(Q\) be a uniformly continuous non-linear operator on \(H\) which is the gradient of a convex "potential" \(\varphi\). Assume \(\varphi\) satisfies
\[\varphi/\|x\|^2 \leq \rho/\|x\|^2/2\]
with \(\rho < 1\). Then \(Q\) has a constructible fixed point.

**Proof.** Define \(f(x) = ([x, x]/2\rho) - \varphi(x)/\sigma, \sigma > 0\). Thus
\[f(x) \geq (2\rho)^{-1} \|x\|^2 (1 - \rho)\]
and \(f\) is bounded below. Clearly, for every set \(x_o\), the set \(S = \{x \in H : f(x) \leq f(x_o)\}\) is closed convex and bounded, and \(\nabla f\) is uniformly continuous. Applying (c) we construct a point for which \(\nabla f(z) = 0\). Thus \(z\) is a fixed point of \(Q\). Q.E.D. Observe that if \(H\) is finite dimensional we need only assume \(Q\) continuous.

Because the method of steepest descent is usually more laborious than the algorithm above, it is not as practical. We shall discuss it below, however, because the proofs require slightly different
techniques.

For simplicity we restrict our attention to gradient directions and to sequences with cluster points.

**Theorem 2 (Steepest Descent).** Assume $f$ continuous on $S$.
Choose $p > 0$ to minimize $f(x_k - p
abla f(x_k))$; set $x_{k+1} = x_k - p_k
abla f(x_k)$, and assume that $\{x_k\}$ has cluster points. Then (b) of Theorem 1 may be asserted.

**Proof.** Let $z$ be a cluster point of $x^n$, clearly $f(x^n)$ converges downward to $f(z)$. We shall show that $\nabla f(z) = 0$. For suppose not and let $\rho_o > 0$ minimize $f(z - \rho_o \nabla f(z))$. Thus $f(z) = f(z - \rho_o \nabla f(z)) + \eta$ for some $\eta > 0$. But for every $x_n$, $f(x_n - \rho_o \nabla f(x_n))$

$$= f(z - \rho_o \nabla f(z) + (x_n - z) + \rho_o (\nabla f(z) - f(x_n))) = f(z - \rho_o \nabla f(z))$$

$$+ [\nabla f(\xi_n'), x_n - z + \rho_o (\nabla f(z), - f(x_n))],$$

with $\xi_n'$ in "between" $(z - \rho_o \nabla f(z)$ and $x_n - \rho_o \nabla f(x_n)$. Let $\{x_n\}$ be a subsequence converging to $z$. Then $\nabla f(\xi_n')$ converges to $\nabla f(z - \rho_o \nabla f(z))$ and $x_n - z + \rho_o(\nabla f(z) - \nabla f(x_n))$ converges to $0$. Hence for $n$ sufficiently large $f(x_n - \rho_o \nabla f(x_n)) \leq f(z - \rho_o \nabla f(z)) + \eta/2 = f(z) - \eta/2$. But $f(z) < f(x_n - \rho_o \nabla f(x_n)) \leq f(x_n - \rho_o \nabla f(x_n)) \leq f(z) - \eta/2$, a contradiction; hence $\nabla f(z) = 0$.

A modified Steepest Descent due to H. Curry is more difficult to prove. Curry's proof is geometric, intricate, and much abbreviated. We prove his assertion below.

**Theorem 3 (Curry [2]).** Assume $\nabla f$ continuous on $S$ and choose $\rho_k$ to be first zero of $\frac{d}{dp}(f(x - p
abla f(x)))$ on the half-ray $\{x - p f(x): p \geq 0\}$. Assume that the sequence $\{x_k\}$ has cluster points. Then (b) of Theorem 1 can be asserted.
Proof. As above, set \( \Delta f(x,p) = f(x) - f(x - pVf(x)) \). Then \( \frac{\Delta f}{\partial p} = -[\nabla f(u), \frac{du}{dp}] = [\nabla f(x - pVf(x))Vf(x)] \), and \( \rho g(x,p) = f(x,p)/||Vf(x)||^2 \). Let \( \rho \) be the least positive root of \( (\rho g(x,p))' = [\nabla f(x - pVf(x))Vf(x)]/||Vf(x)||^2 \). Thus
\[
\rho g(x,p) = \int_0^\rho \frac{[\nabla f(x) - pVf(x), Vf(x)]}{||Vf(x)||^2} \, dp.
\]
Assume now \( Vf(x) \) does not converge to 0. Then for some cluster point \( z \) of \( \{x_k\} \), \( Vf(z) \neq 0 \). For each \( x \) consider the above integrand as a function of \( p \). The integrand is 1 at \( p = 0 \) and 0 at \( p = \rho \). We shall choose \( \rho_0 \) so that the integrand \( I(x,\rho) \) is \( \geq \frac{1}{2} \). Take a subsequence \( \{x_k\} \) which converges to \( z \). This sequence with its limit point is a compact subset of \( S \), so that \( Vf \) is uniformly continuous on this compact set, which we call \( K \). There exists therefore positive numbers \( q \) and \( r \) such that for all \( K \)
\[
q < ||\nabla f(x_k)|| < r.
\]
By the uniform continuity of \( Vf \) on \( K \), given \( e = q/2 \) there exists \( \delta \) such that \( ||Vf(x - \rho Vf(x)) - Vf(x)|| < q/2 < ||Vf(x)||/2 \) whenever \( \rho ||Vf(x)|| < \delta \), for all \( x \in K \). Choose \( \rho_0 = \delta/r \). Then if \( \rho \leq \rho_0 \), \( \rho ||Vf(x)|| < \delta \). Thus if \( \rho \leq \rho_0 \) and \( x \in K \), \( ||Vf(x - \rho Vf(x)) - Vf(x)|| < ||Vf(x)||/2 \). Thus
\[
\int_0^\rho \frac{[\nabla f(x) - \rho Vf(x), Vf(x)]}{||Vf(x)||^2} \, dp = \int_0^\rho \frac{\rho g(x,\rho)}{||Vf(x)||^2} \, dp.
\]
Which proves that if \( x \in K \)
\[
\frac{\rho g(\rho)}{||Vf(x)||^2} > \frac{1}{2} \]
and \( \rho \leq \rho_0 \), \( I(x,\rho) > \frac{1}{2} \) and therefore \( \rho_0 < \rho \). It follows that \( \rho g(\rho) > \int_0^{\rho_0} \frac{\rho g(\rho)}{||Vf(x)||^2} \, dp = \frac{\rho_0}{2} g_0 \) and thus \( \Delta f(x,\rho_0) \geq \rho_0/2 \), contradicting that \( f \) is bounded below on \( K \).
ACKNOWLEDGEMENT

I am indebted to Prof. Alexander Ostrowski for suggesting to me that \( \{x_k\} \) in Theorem (1) converges if \( Wf \) has finitely many zeros.

REFERENCES
