D1-82-0413

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.

BLANK PAGE

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 $\nabla f(x)$ its representer in H. Set $\Delta(x,\rho) = f(x) - f(x - \rho\varphi(x))$, $g(x,\rho) = \Delta(x,\rho)/[\nabla f(x), \varphi(x)]\rho$ and fix σ, $0 < \sigma \leq 1/2$. Here φ denotes a bounded map from S to H, satisfying $[\nabla f(x), \varphi(x)] \geq 0$, such that given $\epsilon > 0$ there exists $\delta > 0$ for which $[\nabla f(x), \varphi(x)] < \delta$ implies $\|\nabla f(x)\| < \epsilon$. For example, see Remark 4, below.

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

(a) $\nabla f(x_k)$ 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 $\Delta(\rho,x) = \rho[\nabla f(\xi), \varphi(x)] = \rho[\nabla f(x), \varphi(x)] + \rho[\nabla f(\xi) - \nabla f(x), \varphi(x)]$ where ξ lies on the open line segment joining x and $x - \rho\varphi(x)$. Thus $g(\rho,x) = 1 + [\nabla f(\xi) - \nabla f(x), \varphi(x)]/[\nabla f(x), \varphi(x)]$. Since ∇f is continuous at x and $\|\xi - x\| \leq \rho\|\varphi(x)\|$, $g(0,x) = 1$. If $g(1,x) < \delta$, then g being continuous takes on all values between 1 and δ; there exists therefore numbers $\rho > 0$ so that $\delta \leq g(\rho,x) \leq 1 - \delta$. 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 $\Delta(x_k,\rho_k) = \rho_k g(x_k,\rho_k)[\nabla f_k,\varphi_k] \geq \rho_k \delta[\nabla f_k,\varphi_k] > 0$. Thus $f(x_{k+1}) < f(x_k)$ and $x_{k+1} \in S$. Assume $[\nabla f_k,\varphi_k] \not\to 0$. This implies that there exists a subsequence $\{x_k\}$ and a number $\epsilon > 0$ such that $[\nabla f_k,\varphi_k] \geq \epsilon$. It follows moreover that $\{\rho_k\}$ is bounded away from 0. For if not take a thinner subsequence $\{\rho_k\}$, if necessary, such that $\{\rho_k\} \to 0$. Since φ_k is bounded on S, $\{\|x_k - \xi_k\|\} \to 0$. By the uniform continuity of ∇f on S, it follows that $\{[\nabla f(\xi_k) - \nabla f(x_k), \varphi(x_k)]\} \to 0$, and therefore $g(\rho_k,x_k) \to 1$, contradicting that $g(\rho_k,x_k) \leq 1 - \delta$ for all k. There exists

therefore a number $q > 0$ such that $\rho_k \geq q$. Hence $\Delta(x_k,\rho_k) \geq q\delta\epsilon$, from which we may contradict the hypothesis that f is bounded below. Thus $[\nabla f_k,\varphi_k] \to 0$, 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 ∇f has a unique root on S, then $\{x_k\}$ converges to it. If the roots of ∇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 $N(x_i)$ denote spheres of radius $\epsilon/3$ centered at x_i. $H \sim \cup N(x_i)$ 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 $H \sim \cup N(x_i)$ 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: ∇f and φ continuous on S, and $[\varphi(x), f(x)] \geq 0$ for all $x \in S$,

$[\varphi(x), f(x)] = 0$ 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 $f(x_k) > f(z) \geq f(x_k) + [\nabla f_k, z-x_k] \geq f(x_k) - \|\nabla f_k\| D$, where D is the diameter of S.

4. An example for φ is the following. Let $H = E_n$ and let δ_i, $(1 \leq i \leq n)$, denote the rows of the identity matrix. Choose i_o so that $[\nabla f(x), \delta_{i_o}] \geq [\nabla f(x), \delta_i]\ \forall i$, $(1 \leq i \leq n)$. Set $\varphi(x) = \|\nabla f(x)\| \delta_{i_o}$, then $[\nabla f(x), \varphi(x)] \geq \|\nabla f(x)\|^2/\sqrt{n}$.

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

Proof. Define $f(x) = ([x,x]/2\sigma) - \varphi(x)/\sigma$, $\sigma > 0$. Thus $f(x) \geq (2\sigma)^{-1}\|x\|^2(1 - \rho)$ 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 ∇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 $\rho > 0$ to minimize $f(x_k - \rho\nabla f(x_k))$; set $x_{k+1} = x_k - \rho_k\nabla f(x_k)$, 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 $f(z - \rho\nabla f(z))$. 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 ξ_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 $f(z) < f(x_n - \rho_n\nabla f(x_n)) \leq f(x_n - \rho_o\nabla f(x_n)) \leq f(z) - \eta/2$, 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 ∇f continuous on S and choose ρ_k to be first zero of $\frac{d}{dp}(f(x - \rho\nabla f(x)))$ on the half-ray $\{x - \rho f(x) : \rho \geq 0\}$. Assume that the sequence $\{x_k\}$ has cluster points. Then (b) of Theorem 1 can be asserted.

Proof. As above, set $\Delta f(x,\rho) = f(x) - f(x - \rho\nabla f(x))$. Then $\frac{d\Delta f}{d\rho} = -[\nabla f(u), \frac{du}{d\rho}] = [\nabla f(x - \rho\nabla f(x), \nabla f(x)]$, and $\rho g(x,\rho) = f(x,\rho)/\|\nabla f(x)\|^2$. Let $\hat{\rho}$ be the least positive root of $(\rho g(x,\rho))' = [\nabla f(x - \rho\nabla f(x)), \nabla f(x)]/\|\nabla f(x)\|^2$. Thus $\hat{\rho} g(x,\hat{\rho}) = \int_0^{\hat{\rho}} \frac{[\nabla f(x) - \rho f(x), \nabla f(x)]}{\|f(x)\|^2} d\rho$. Assume now $\nabla f(x_k)$ does not converge to 0. Then for some cluster point z of $\{x_k\}$, $\nabla f(z) \neq 0$. For each x consider the above integrand as a function of ρ. The integrand is 1 at $\rho = 0$ and 0 at $\rho = \hat{\rho}$. We shall choose ρ_o so that the integrand $I(x,\rho)$ is $\geq 1/2$. Take a subsequence $\{x_k\}$ which converges to z. This sequence with its limit point is a compact subset of S, so that ∇f is uniformly continuous on this compactum, 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 ∇f on K, given $\epsilon = q/2$ there exists δ such that $\|\nabla f(x - \rho\nabla f(x)) - \nabla f(x)\| < q/2 < \|\nabla f(x)\|/2$ whenever $\rho\|\nabla f(x)\| < \delta$, for all $x \in K$. Choose $\rho_o = \delta/r$. Then if $\rho \leq \rho_o$, $\rho\|\nabla f(x)\| < \delta$. Thus if $\rho \leq p_o$ and $x \in K$, $\|\nabla f(x - \rho\nabla f(x) - \nabla f(x)\| < \|\nabla f(x)\|/2$. Thus

$$\tfrac{1}{2} > \frac{1}{\|\nabla f(x)\|}\|\nabla f(x - \rho\nabla f(x)) - \nabla f(x)\| \geq \frac{1}{\|\nabla f(x)\|}\left|[\nabla f(x - \rho\nabla f(x)) - \nabla f(x), \frac{\nabla f(x)}{\|\nabla f(x)\|}]\right|$$

$$= \left|[\frac{\nabla f(x - \rho\nabla f(x))}{\|\nabla f(x)\|}, \frac{\nabla f(x)}{\|\nabla f(x)\|}] - 1\right|.$$ Which proves that if $x \in K$ and $\rho \leq \rho_o$, $I(x,\rho) > \frac{1}{2}$ and therefore $\rho_o < \hat{\rho}$. It follows that $\hat{\rho} g(\hat{\rho}) > \int_0^{\rho_o} \frac{1}{2} d\rho = \frac{1}{2}\rho_o$ and thus $\Delta f(x_k, \rho_k) \geq q\rho_o/2$, 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 ∇f has finitely many zeros.

REFERENCES

[1] A. A. Goldstein. Minimizing functionals on Hilbert Space. Computer Methods in Optimization Problems, Academic Press, N.Y. 1964, pp. 159-165.

[2] Curry, H. The method of steepest descent for non-linear minimization problems. Quarterly of Applied Math. 2, 258, 1944.