IMPRACTICAL

PACIFIC JOURNAL OF MATHEMATICS
Vol. 16, No. 1, 1966

MINIMIZATION OF FUNCTIONS HAVING LIPSCHITZ CONTINUOUS FIRST PARTIAL DERIVATIVES

LARRY ARMIJO

A general convergence theorem for the gradient method is proved under hypotheses which are given below. It is then shown that the usual steepest descent and modified steepest descent algorithms converge under the some hypotheses. The modified steepest descent algorithm allows for the possibility of variable stepsize.

For a comparison of our results with results previously obtained, the reader is referred to the discussion at the end of this paper.

Principal conditions. Let f be a real-valued function defined and continuous everywhere on E^n (real Euclidean n-space) and bounded below E^n. For fixed $x_0 \in E^n$ define $S(x_0) = \{x : f(x) \leqq f(x_0)\}$. The function f satisfies: condition I if there exists a *unique* point $x^* \in E^n$ such that $f(x^*) = \inf_{x \in E^n} f(x)$; Condition II at x_0 if $f \in C^1$ on $S(x_0)$ and $\nabla f(x) = 0$ for $x \in S(x_0)$ if and only if $x = x^*$; Condition III at x_0 if $f \in C^1$ on $S(x_0)$ and ∇f is Lipschitz continuous on $S(x_0)$, i.e., there exists a Lipschitz constant $K > 0$ such that $|\nabla f(y) - \nabla f(x)| \leqq K|y - x|$ for every pair $x, y \in S(x_0)$; Condition IV at x_0 if $f \in C^1$ on $S(x_0)$ and if $r > 0$ implies that $m(r) > 0$ where $m(r) = \inf_{x \in S_r(x_0)} |\nabla f(x)|$, $S_r(x_0) = S_r \cap S(x_0)$, $S_r = \{x : |x - x^*| \geqq r\}$, and x^* is any point for which $f(x^*) = \inf_{x \in E^n} f(x)$. (If $S_r(x_0)$ is void, we define $m(r) = \infty$.)

It follows immediately from the definitions of Conditions I through IV that Condition IV implies Conditions I and II, and if $S(x_0)$ is bounded, then Condition IV is equivalent to Conditions I and II.

2. The convergence theorem. In the convergence theorem and its corollaries, we will assume that f is a real-valued function defined and continuous everywhere on E^n, bounded below on E^n, and that Conditions III and IV hold at x_0.

THEOREM. *If $0 < \delta \leqq 1/4K$, then for any $x \in S(x_0)$, the set*

$$S^*(x, \delta) = \{x_\lambda : x_\lambda = x - \lambda \nabla f(x),\ \lambda > 0,\ f(x_\lambda) - f(x) \leqq -\delta|\nabla f(x)|^2\} \tag{1}$$

Received January 30, 1964. The research for this paper was supported in part by General Dynamics/Astronautics, San Diego, California, Rice University, Houston, Texas, and the Martin Company, Denver, Colorado. The author is currently employed by the National Engineering Science Company, Houston, Texas.

PRACTICAL

Lecture 5

HW 1 : Due in 2 days.

Last time

- More convexity
- Characterization smooth convex functions
- Subgradients

Today

- Subdifferential calculus
- What's to come?
- Gradient Descent

Subdifferential calculus.

Proposition: Subdifferential calculus

Suppose that $f_1, f_2 : \mathbb{R}^d \to \mathbb{R}$ are convex functions. Then the following holds

1. (Sums) $\partial(f_1 + f_2)(x) = \partial f_1(x) + \partial f_2(x)$.
2. (Chain rule) If $A : \mathbb{R}^n \to \mathbb{R}^d$ linear

$$\partial(f_1 \circ A)(x) = A^T \partial f_1(Ax).$$

3. (Scalings)

$$\partial(\alpha f_1)(x) = \alpha \partial f_1(x).$$

4. (Max) For all x, define $M(x) = \{ i \mid f_i(x) = \max\{f_1(x), f_2(x)\}\}$.

$$\partial \max\{f_1, f_2\}(x) = \operatorname{conv}\{ g \in \partial f_i \mid i \in M(x) \}.$$

convex hull

5. (Smooth functions) Assume that f_i is diff at x.

$$\partial f_i(x) = \{ \nabla f(x) \}.$$

This one you should prove.

We will not prove this result, as we need additional machinary from convex geometry. But you are free to use it.

What's next? Algorithms!

We will cover Smooth first

3 to 4 lectures.

- Gradient Descent
- Descent Lemma
- Stepsizes / Lineseach
- Nonconvex smooth opt guarantees
- Better guarantees for convex
- Complexity Lower Bounds
- Acceleration

Gradient Descent

Bread & Butter of opt. theory.

Gradient Descent (GD) updates

$$x_{k+1} \leftarrow x_k - \alpha_k \nabla f(x_k) \qquad (☺)$$

Follow descent direction!

Another view of GD

$$x_{k+1} = \min_x \left\{ \overbrace{f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{2\alpha_k}\|x - x_k\|^2}^{h_k} \right\}. \qquad (♡)$$

Why are (☺) and (♡) the same?

The loss function is convex

$$\nabla h_k(x_{k+1}) = 0 = \nabla f(x_k) + \frac{1}{\alpha_k}(x_{k+1} - x_k)$$

$$\Updownarrow$$

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k)$$

Intuition

This will be a recurrent theme in algorithm design.

Descent Lemma ← Bread & Butter of opt. theory.

Lemma: For any f with L-Lipschitz gradient, and $k \geq 0$

$$f(x_{k+1}) \leq f(x_k) - \left(\alpha_k - \frac{L\alpha_k^2}{2}\right) \|\nabla f(x_k)\|^2$$

$\dashv$

Consequences

1. Decrease when $\left(\alpha_k - \frac{L\alpha_k^2}{2}\right) > 0$

$$\Updownarrow$$

$$\alpha_k < \frac{2}{L}$$

2. Best decrease when $\alpha_k = \frac{1}{L}$.

of $-\frac{1}{2L} \|\nabla f(\bar{x}_k)\|_2^2$.

Proof: We use the Taylor approximation bound

$$\left| f(\bar{x}_{k+1}) - \left(f(\bar{x}_k) + \langle \nabla f(\bar{x}_k), \bar{x}_{k+1} - \bar{x}_k \rangle \right| \leq \frac{L}{2} \|x_{k+1} - x_k\|^2$$

Subtituting

$$f(x_{k+1}) - f(x_k) + \alpha_k \|\nabla f(\bar{x}_k)\|^2 \leq \frac{L\alpha_k^2}{2} \|\nabla f(\bar{x}_k)\|^2$$

Rearranging

$$\Rightarrow f(\bar{x}_{k+1}) \leq f(\bar{x}_k) - \left(\alpha_k - \frac{L\alpha_k^2}{2}\right)\|\nabla f(\bar{x}_k)\|^2.$$

$\square$

How to pick stepsizes?

Natural idea

According to DL, we should pick $\alpha_k = 1/L \Rightarrow \frac{1}{2L}\|\nabla f(x_k)\|^2$ descent.

The problem is that we don't know L a priori! **IMPRACTICAL**

Exact lineseach

We know we have descent if we follow $-\nabla f(x_k)$. Let's pick the best descent:

1D problem

$$\alpha_k = \underset{\alpha \in \mathbb{R}}{\operatorname{argmin}}\ f(x_k - \alpha \nabla f(x_k))$$

It outperforms $\alpha_k = 1/L$ since

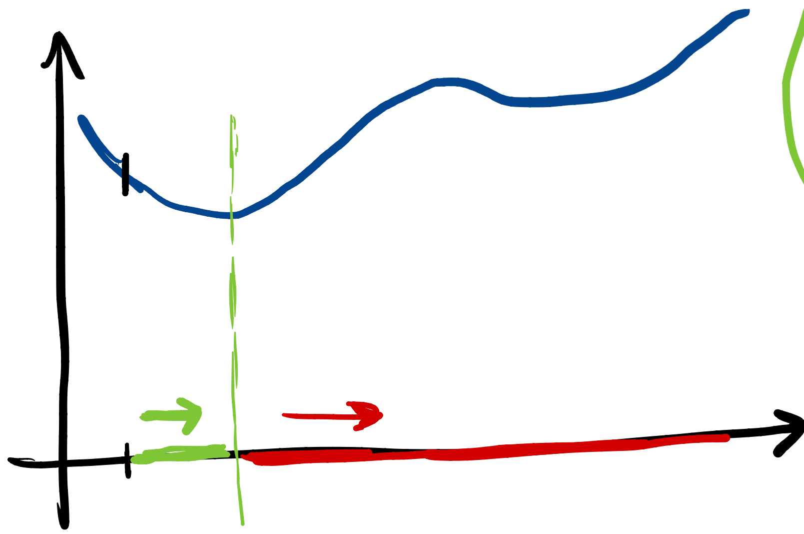
$$f(x_{k+1}) \leq f(x_k - \alpha \nabla f(x_k)) \ \forall \alpha$$

$$\leq f\left(x_k - \frac{1}{L}\nabla f(x_k)\right).$$

IMPRACTICAL It requires solving an optimization problem at each iter!

Backtracking Linesearch

Idea: How about we try smaller stepsizes until we see sufficient descent?



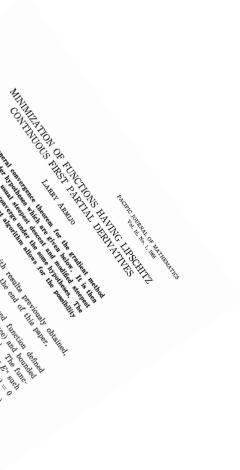
What is sufficient? (2)

How do we make them small? (1)

(1) Decrease exponentially fast. Pick $a \in \mathbb{R}^d$ and $\tau \in (0,1)$ and try

$$\alpha_k = a\tau^n \quad \text{for } n = 1, 2, \ldots$$

(2) To measure descent we use



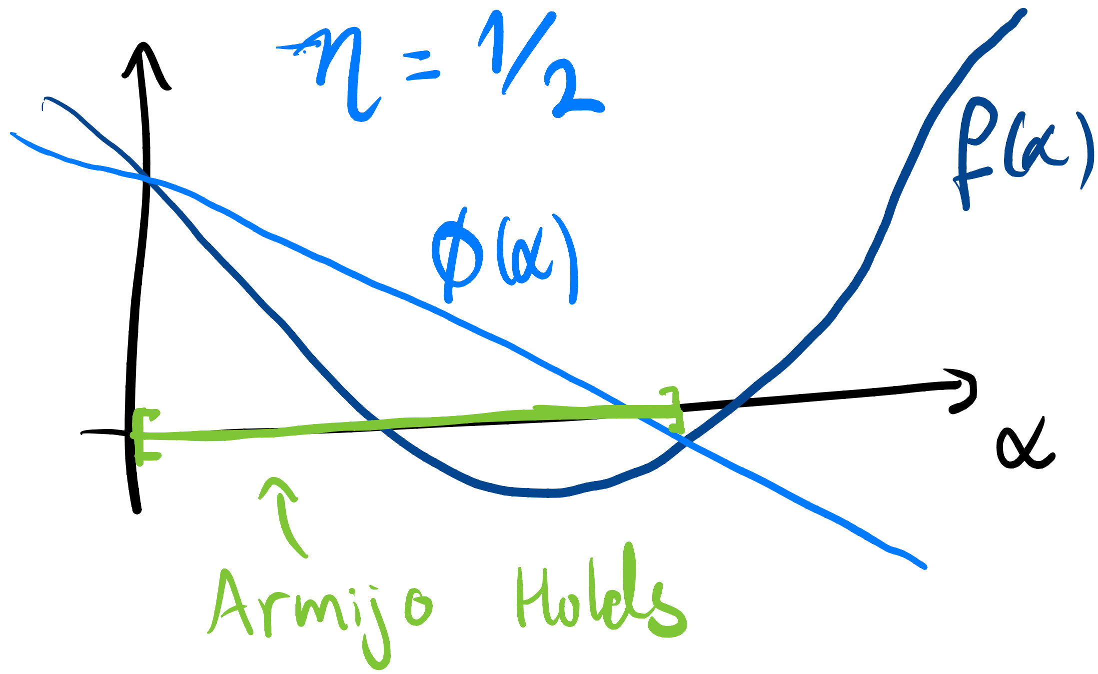
Could not find a picture

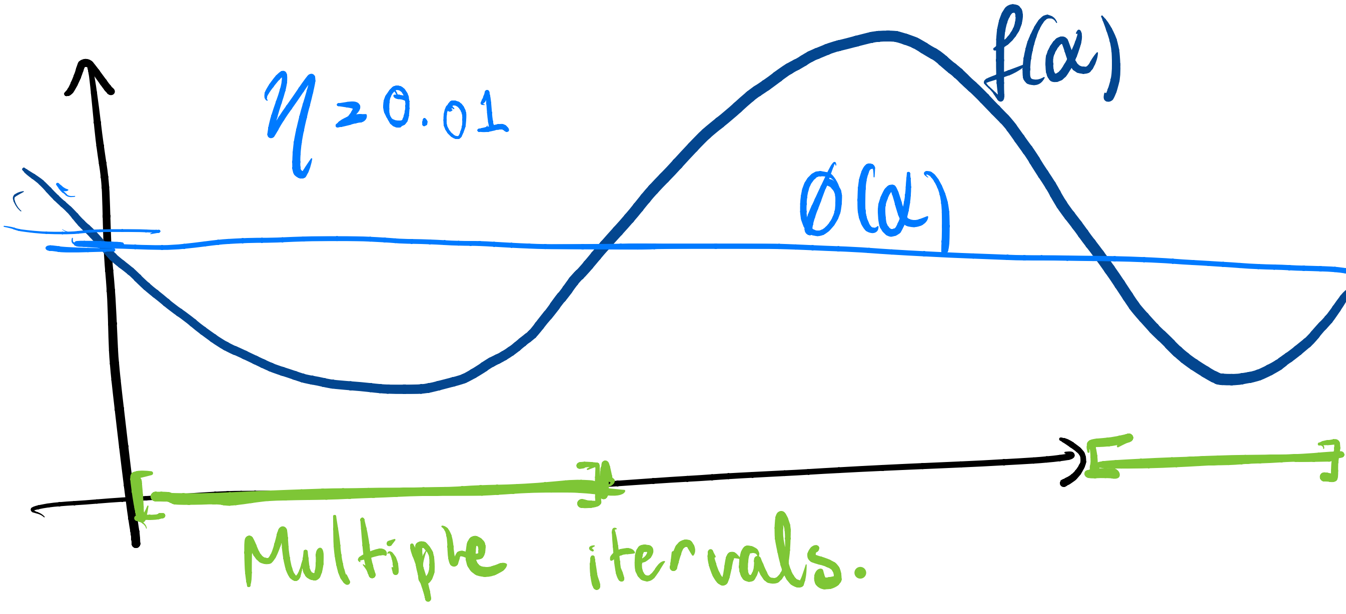
The so-called **Armijo Condition:**

Pick $\eta \in (0,1)$, declare sufficient descent when

$$f(x_k - \alpha \nabla f(x_k)) \leq \underbrace{f(x_k) - \eta \alpha \|\nabla f(x_k)\|^2}_{\phi(\alpha)} \quad (\star)$$

Intuition





The full backtracking algorithm

Pick

$$\alpha_k = \sup_n \left\{ a\tau^n \;\middle|\; (\star) \text{ holds with } \alpha = a\tau^n \right\}$$

Lemma The Armijo condition holds for

$$\alpha \in \left[0, \frac{2(1-\eta)}{L}\right]$$

Proof: By the DL

$$f(x_k - \alpha \nabla f(x_k)) \leq f(x_k) - \left(\alpha - \frac{L\alpha^2}{2}\right)\|\nabla f\|^2$$

$$\overset{?}{\leq} f(x_k) - \eta\alpha \|\nabla f(x_k)\|^2$$

would hold if $\left(\alpha - \frac{L\alpha^2}{2}\right) \geq \eta\alpha$

$$\Leftrightarrow \quad \alpha \leq \frac{2(1-\eta)}{L}.$$

$\blacksquare$

Consequence PRACTICAL

1. Backtracking only require

$\left\lceil \log_{1/\tau}\left(\frac{aL}{2(1-\eta)}\right)\right\rceil$ steps to stop.

Check this!

Armijos original choice

If we take $\eta = \tau = 1/2$

$a = 1$

and $L \leq 10^6$ ← Function is very unstable

$\Rightarrow$ 20 steps are enough.

2. Note that $\alpha_k \geq \min\left\{a, \frac{2\tau(1-\eta)}{L}\right\}$.

Then

$$f(x_{k+1}) \leq f(x_k) - \eta\,\alpha_k \|\nabla f(x_k)\|^2$$

$$\leq f(x_k) - \eta \min\left\{a, \frac{2\tau(1-\eta)}{L}\right\} \|\nabla\|^2$$

Thus, if $a \geq \frac{1}{L}$ and $\eta = \tau = \frac{1}{2}$

Reasonable.

$a \geq 1 \geq \frac{1}{L}$ if $L \geq 1$.

$$\leq f(x_k) - \frac{1}{2} \min\left\{\frac{1}{L}, \frac{1}{2L}\right\} \|\nabla f(x_k)\|^2$$

$$= f(x_k) - \frac{1}{4L} \|\nabla f(x_k)\|^2$$

Only lost constant fraction.