Lecture 5 Scribe? 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 $f, h : \mathbb{R}^d \to \mathbb{R}$ are convex functions. Then the following holds

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

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

3. (Scalings)

$$\partial(\alpha f)(x) = \alpha\, \partial f(x).$$

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

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

Follow descent direction!

Another view of GD

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

Why are $(\smile)$ and $(\heartsuit)$ 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 $\leftarrow$ 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

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

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

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

It outperforms $\alpha_k = 1/L$ since

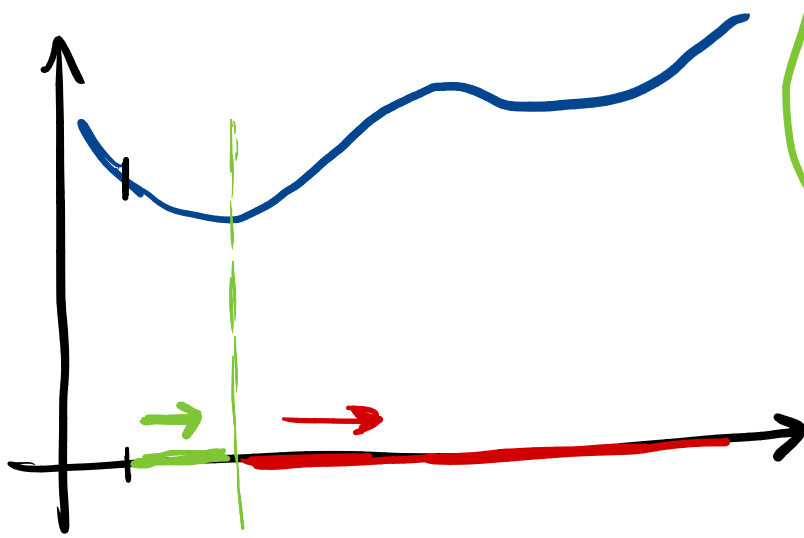
$$f(x_{k+1}) \leq f(x_k - \alpha \nabla f(x_k)) \quad \forall \alpha$$

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

IMPRACTICAL It requires solving an optimization problem at each iter!

Backtracking Linesearch

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



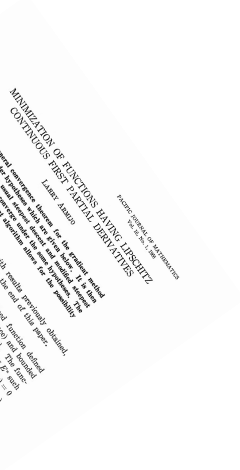
What is sufficient? (2)

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


MINIMIZATION OF FUNCTIONS HAVING LIPSCHITZ CONTINUOUS FIRST PARTIAL DERIVATIVES

LARRY ARMIJO

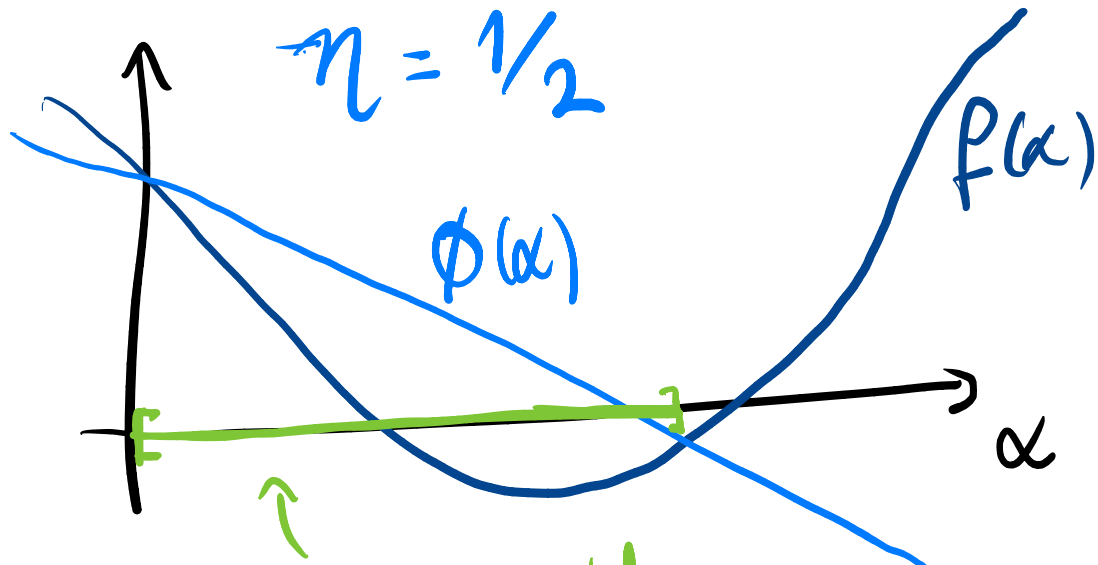
Could not find a picture

The so-called Armijo Condition:

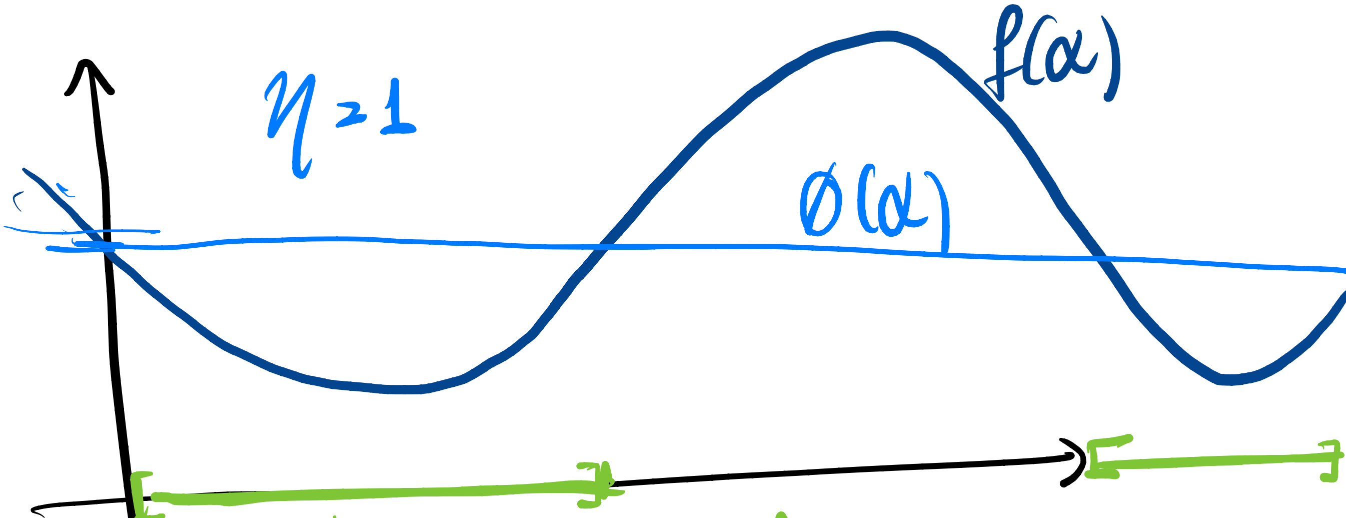
Pick $\eta \in (0,1)$, declare sufficient descent when

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

Intuition



Armijo Holds



Multiple itervals.

The full backtracking algorithm

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

Lemma The Armijo condition holds for
$$\alpha \in \left[0, \frac{2(1-\eta)}{L}\right]$$

Proof: By the DL

$$f(x_k - \alpha \nabla f(x_k)) \le f(x_k) - \left(\alpha - \frac{L\alpha^2}{2}\right)\|\nabla f\|^2$$
$$\overset{?}{\le} f(x_k) - \eta\alpha \|\nabla f(x_k)\|^2$$

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

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

$\blacksquare$

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!

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

Armijos original choice

$a = 1$

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.