IE510Applied Nonlinear Programming

Lecture 2a: Gradient Methods I: Introduction

Ruoyu Sun

Jan, 2018

Outline

Last Time Questions

- **Q1:** Why should we study optimality condition first?
  What other benefits?
- **Q2:** We learned optimality conditions for unconstrained problems last time.
- **Q3:** Have we learned any sufficient condition for global optimality last time?

  Yes. Convex + $\nabla f(x)=0$.
- **Q4:** Judge: If there is a unique stationary point, then it is the global-min or global-max.

  Wrong. Counterexample: x^3 at $x=0$.

  Unique stationary pt, but not global-min or global-max.

Last Time summary

- Necessary condition

$$\nabla f(x^*) = 0, \quad \text{(first-order condition)},$$
$$\nabla^2 f(x^*) \succeq 0, \quad \text{(second-order condition)}.$$

- Sufficient condition

$$\nabla f(x^*) = 0, \quad \text{(first-order condition)},$$
$$\nabla^2 f(x^*) \succ 0, \quad \text{(second-order condition)}.$$

- How to use optimality condition to directly find (local) minima 3-step method.
- Existence of global-min: ① compact domain or level set; ② coersive
- Convexity leads to "every local-min is global-min"

Today

- Introduction of Gradient Descent method
- After today's course, you will be able to
  - explain to your high-school nephew what is gradient descent (GD)

    Advanced: have 3 real world examples of GD
  - list different forms of iterative descent methods

    Advanced: understand pros and cons of each form
  - Advanced: derive GD from three different perspectives

Outline

Motivation

Figure: Hill Climbing[1]

- Which route is fast for climbing? (if foggy)
- **Greedy approach** (conceptual)
  - climb 1m along various directions;
  - pick the one with ____ highest elevation
    (biggest increase in height)

Gradient Descent

- Issue: cannot check all directions
- One solution: instead of 1m, if climb ϵ meter along each direction, the best direction is approximately ___gradient
- When consider minimization, pick negative gradient

$$\mathbf{x} \leftarrow \mathbf{x} -$$

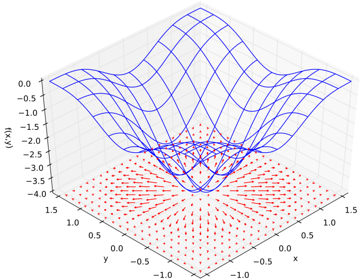


Figure: The gradients of a function (Wikipedia: Gradient)

Gradient Descent

- Issue: cannot check all directions
- One solution: instead of 1m, if climb ϵ meter along each direction, the best direction is approximately ____
- When consider minimization, pick negative gradient

$$\mathbf{x} \leftarrow \mathbf{x} - \alpha \nabla f(x)$$

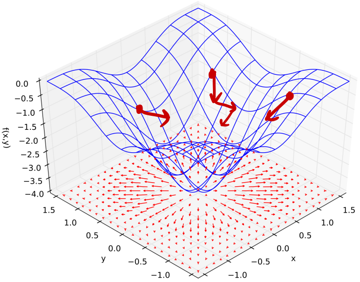


Figure: The gradients of a function (Wikipedia: Gradient)

Example: Basketball

p is position, x is the force. $\min(p-ax)^2$, if $ax-p=0$, on target

(of basket) ax is the position of your ball

Arbitrary strength x^0, the ball falls at position ax^0,

error ax^0-p.

If $ax^0-p>0$, too far away, reduce strength x.

- - - <0, too close, increase strength x.

$$x' = x^0 - \underbrace{(ax^0-p)}_{\text{gradient}} \begin{cases} < x^0, & \text{if } ax^0-p>0. \\ > x^0, & \text{o.w} \end{cases}$$

Analysis of GD: Fixed Point

- Think: ~~How~~ What to analyze (an algorithm)?
  - Step 2: Does it converge ?
  - Step 1: Converge to what ?
- " Converge to what": "fixed point" analysis.
  When the algorithm converges to x^∞, what is x^∞?

$$x^{k+1} = x^k - \alpha \nabla f(x^k),$$

becomes $x^\infty = x^\infty - \alpha \nabla f(x^\infty)$,

i.e. $\cancel{\alpha} \nabla f(x^\infty) = 0$,

- **Remark**: GD tries to find global min, but GD actually finds stationary point

Gradient Descent is a Descent Method

- To prove convergence, need a basic understanding
- If $\nabla f(\mathbf{x}) = 0$, then $\mathbf{x}$ is a stationary point; done
- If $\nabla f(\mathbf{x}) \neq 0$, then $-\nabla f(x)$ is a descent direction: there is an interval $(0, \delta)$ of stepsizes such that

$$f(x - \alpha \nabla f(x)) < f(\mathbf{x}), \ \forall\ \alpha \in (0, \delta).$$

- Proof by Taylor expansion: $x_\alpha - x = -\alpha \nabla f(x)$

$$f(\underset{\substack{\| \\ x - \alpha\nabla f(x)}}{x^\alpha}) = f(x) + \underbrace{f(x)'(x_\alpha - x)}_{-\alpha\|\nabla f(x)\|^2} + o\underbrace{(\|x_\alpha - x\|^2)}_{\alpha^2\|\nabla f(x)\|^2}$$

$$\leq f(x) - \frac{\alpha}{2}\|\nabla f(x)\|^2, \text{ when } \alpha \text{ small enough},$$

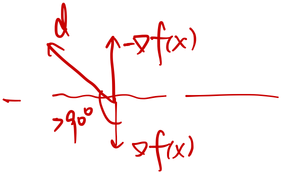
Iterative Descent Method

- More generally, if a given direction $\mathbf{d}$ that is with obtuse angle with $\nabla f(\mathbf{x})$

  $\langle \nabla f(x), d \rangle < 0$

  there is an interval $(0, \delta)$ of stepsizes such that

$$f(\mathbf{x} + \alpha \mathbf{d}) < f(\mathbf{x}), \ \forall \ \alpha \in (0, \delta).$$



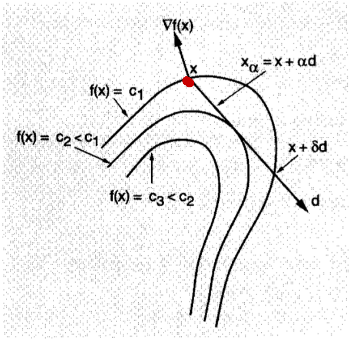


Figure: descent method (textbook Fig 1.2.3)

Iterative Descent Methods

$$\mathbf{x}^{r+1} = \mathbf{x}^r + \alpha_r \mathbf{d}^r, \; r = 0, 1, \cdots$$

where, if $\nabla f(\mathbf{x}^r) \neq 0$, the direction $\mathbf{d}^r$ satisfies $\nabla f(\mathbf{x}^r)\mathbf{d}^r < 0$, and α^r is a positive stepsize

- General Case: Gradient methods

$$\mathbf{x}^{r+1} = \mathbf{x}^r - \alpha_r \mathbf{D}^r \nabla f(\mathbf{x}^r), \; r = 0, 1, \cdots$$

 where $\mathbf{D}^r$ is a positive definite matrix called scaling matrix

- Special case I: Steepest descent (a.k.a. GD in non-optimization world) $D^r = I$;

$$\mathbf{x}^{r+1} = \mathbf{x}^r - \alpha_r \nabla f(\mathbf{x}^r), \; r = 0, 1, \cdots$$

- Special case II: Newton's method $D^r = \nabla^2 f(x^r)$

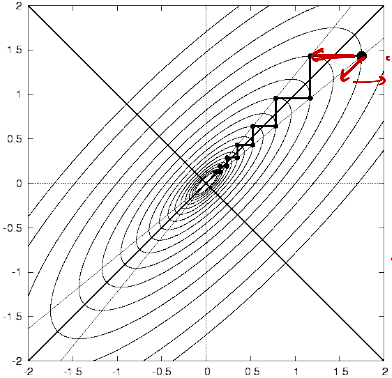
$$\mathbf{x}^{r+1} = \mathbf{x}^r - \alpha_r \left(\nabla^2 f(\mathbf{x}^r)\right)^{-1} \nabla f(\mathbf{x}^r), \; r = 0, 1, \cdots$$

Discussion

Figure: The Steepest Descent (ERASIP: DSPHumour)

Discussion

- In practice steepest descent may have slow convergence
  - Practical performance?
  - **Exercise**: implement the steepest descent for a 2-D convex quadratic problem. Show the convergence plot on the contour of the function



"globally" good direction mismatch

zigzag behavior

Figure: The Steepest Descent in Practice (Komarix.org)

Newton's Method

- Newton's method: generally fast convergence

1. It treats the objective (locally) as a quadratic problem around $\mathbf{x}^r$

$$f(\mathbf{x}) \approx f(\mathbf{x}^r) + \langle \nabla f(\mathbf{x}^r), \mathbf{x} - \mathbf{x}^r \rangle + \frac{1}{2}(\mathbf{x} - \mathbf{x}^r)^T \nabla^2 f(\mathbf{x}^r)(\mathbf{x} - \mathbf{x}^r)$$

$$\nabla f(x^r) + \nabla^2 f(x^r)(x - x^r) = 0$$

Minimizer of RHS: $x^r + \nabla^2 f(x^r)^{-1} \nabla f(x^r) = x^{r+1}$.

2. Pros: how many iterations does it take for Newton method to minimize a quadratic function f? 1

Cons 0: not directly apply to nonconvex problems

3. Cons 1: difficult to make it numerically stable
4. Cons 2: each iteration is time-consuming

Choice of Stepsize

- **Constant Stepsize**:

$$\alpha_r = \alpha$$

Comment: practically used often, but what's the constant? $\frac{1}{L}$.

- **Minimization Rule**: Pick α_r such that

$$\alpha_r = \arg\min_{\alpha \geq 0} f(\mathbf{x}^r + \alpha \mathbf{d}^r)$$

Comment: maximum reduction, but hard to compute (or time-consuming)

- **Limited Minimization Rule**: Pick α_r such that

$$\alpha_r = \arg\min_{\alpha \in [0,\, s]} f(\mathbf{x}^r + \alpha \mathbf{d}^r)$$

region $[0, s]$

Choice of Stepsize (Cont.)

- **Diminishing Stepsize**: useful in practice, but cannot diminish too fast? Why??

1000 miles

$x^* \longrightarrow x^0$

$$\alpha_r \to 0, \quad \boxed{\sum_{r=1}^{\infty} \alpha_r = \infty}$$

$\sum_{r=1}^{\infty} \alpha_r = 100$ miles, impossible to reach x^*.

Example: $\alpha_r = \frac{1}{r}$, $\cancel{\alpha_r = \frac{1}{2^r}}$, $\cancel{\alpha_r = \frac{1}{r^2}}$, $\alpha_r = \frac{1}{r^{0.8}}$.

- **Armijo rule**: Let $\sigma \in (0, \frac{1}{2})$. Fix s as a constant, and $0 < \beta < 1$ as a constant, Keep shrinking α by $s, \beta s, \beta^2 s, \cdots$ until the following is satisfied

in hill climbing: 1m, 0.5m, 0.25m, 0.125m, ...

$$f(\mathbf{x}^r + \alpha \mathbf{d}^r) - f(\mathbf{x}^r) \le \sigma\alpha\langle \nabla f(\mathbf{x}^r), \mathbf{d}^r \rangle$$

Comment: achieves descent, but need to test many times

good balance between efficiency & time-complexity

The Armijo Rule

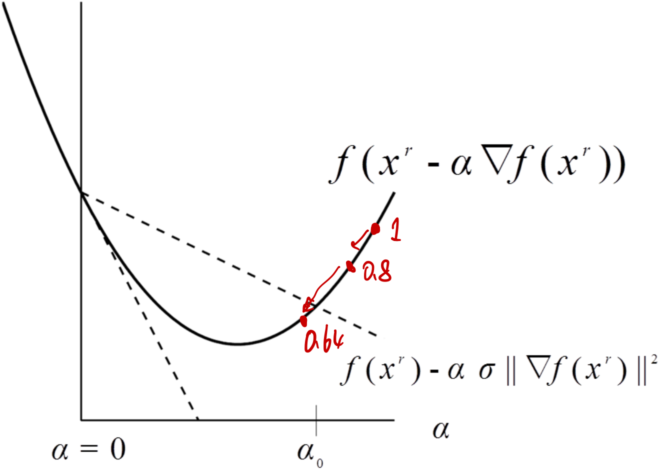


Figure: The Armijo Stepsize Selection

The Overall Strategy

- No matter what strategy we choose, there should be sufficient descent in the objective at each step
- The objective function $f(\mathbf{x})$ serves as a "potential" to guide the optimization process

  such a potential not always easy to find.
- These methods are called "descent" methods, for precisely this reason
- Basically a "good" stepsize and a "good" direction is all that is required to find the (local) optimal solutions
- Next time: theoretical analysis of descent methods

2nd Interpretation: quadratic approximation

- Recall: Newton method does quadratic approximation around $\mathbf{x}^r$

$$f(\mathbf{x}) \approx f(\mathbf{x}^r) + \langle \nabla f(\mathbf{x}^r), \mathbf{x} - \mathbf{x}^r \rangle + \frac{1}{2}(\mathbf{x} - \mathbf{x}^r)^T \nabla^2 f(x^r) (\mathbf{x} - \mathbf{x}^r)$$

GD: take gradient of RHS,

$0 = \nabla f(x^r) + L(x - x^r)$

- GD also does quadratic approximation around $\mathbf{x}^r$ $\Rightarrow -\frac{1}{L}\nabla f(x^r) + x^r = x$

$$f(\mathbf{x}) \approx f(\mathbf{x}^r) + \langle \nabla f(\mathbf{x}^r), \mathbf{x} - \mathbf{x}^r \rangle + \frac{1}{2}(\mathbf{x} - \mathbf{x}^r)^T L I (\mathbf{x} - \mathbf{x}^r)$$

Minimize RHS to get $x^{r+1} = x^r - \frac{1}{L}\nabla f(x^r)$.

- Universal idea: reducing to easier problem
- Generalization: successive convex approximation

3rd Interpretation: fixed point algorithm

- Recall: A stationary point x^* satisfies

$$\nabla f(x^*) = 0, \quad \text{(first-order condition)},$$

  Solving this equation is Step 1 of "Algorithm 1" last time

- A simple way to derive GD: let $x = x + \beta \nabla f(x)$, and make it

$$x^{r+1} = x^r + \beta \nabla f(x^r)$$

  Example? $x^5 + x + 1 = 0$. Solve it by $x^{r+1} = x^r + \beta((x^r)^5 + x^r + 1)$.

- Recall Step 3: among all candidates, find the best one
  - What if infinitely many?
  - Even if finite, how to find all? **Multiple initial points**.
  - In practice, try your best... no guarantee

- **Finding stationary points** is a **major task** of the course

3rd Interpretation: fixed point algorithm

- Recall: A stationary point x^* satisfies

$$\nabla f(x^*) = 0, \quad \text{(first-order condition)},$$

  Solving this equation is Step 1 of "Algorithm 1" last time

- A simple way to derive GD: let $x = x + \beta \nabla f(x)$, and make it

$$x \quad = x \quad + \beta \nabla f(x \quad)$$

  Example? $x^5 + x + 1 = 0$. Solve it by

- Recall Step 3: among all candidates, find the best one
  - What if infinitely many?
  - Even if finite, how to find all? **Multiple initial points**.
  - In practice, try your best... no guarantee

- **Finding stationary points** is a **major task** of the course