Optimization Background (a)

Mingyi Hong

University Of Minnesota

Outline

- Basic Concepts in Nonlinear Optimization
- How to Analyze Algorithm Convergence
- Basic Concepts on Graph Theory
- Main reference: D. P. Bertsekas "Nonlinear Programming", Version 2 or 3
- Also refer to L. Bottou, F. E. Curtis and J. Nocedal, "Optimization Methods for Large-Scale Machine Learning", SIAM Review.

Differentiable unconstrained minimization

$$\begin{array}{ll} \text{minimize}_x & f(x) \\ \text{subject to} & x \in \mathbb{R}^n \end{array}$$

- **Objective function** $f : \mathbb{R}^n \to \mathbb{R}$ is a continuous function
- **Optimization variable** $x \in \mathbb{R}^n$

Differentiable unconstrained minimization

$$\begin{array}{ll} \text{minimize}_x & f(x) \\ \text{subject to} & x \in \mathbb{R}^n \end{array}$$

- **Objective function** $f : \mathbb{R}^n \to \mathbb{R}$ is a continuous function
- **Optimization variable** $x \in \mathbb{R}^n$
- **Unconstrained local minimum** x^*: $\exists \epsilon > 0$ s.t. $f(x) \geq f(x^*)$, for all $\|x - x^*\| \leq \epsilon$; i.e., x^* is the best in a small enough neighborhood
- **Unconstrained global minimum** $\hat{x}$: $f(x) \geq f(\hat{x})$ for all $x \in \mathbb{R}^n$

Differentiable unconstrained minimization

$$\begin{array}{ll} \text{minimize}_x & f(x) \\ \text{subject to} & x \in \mathbb{R}^n \end{array}$$

- **Objective function** $f : \mathbb{R}^n \to \mathbb{R}$ is a continuous function
- **Optimization variable** $x \in \mathbb{R}^n$
- **Unconstrained local minimum** x^*: $\exists \epsilon > 0$ s.t. $f(x) \geq f(x^*)$, for all $\|x - x^*\| \leq \epsilon$; i.e., x^* is the best in a small enough neighborhood
- **Unconstrained global minimum** $\hat{x}$: $f(x) \geq f(\hat{x})$ for all $x \in \mathbb{R}^n$
- **Graphically...**

Existence of Optimal Solution

- Consider the following problem

$$\inf_{x \in \mathbb{R}} \exp^{-|x|} = ? \tag{2.1}$$

- Is the minimum attained?

Existence of Optimal Solution

- Consider the following problem

$$\inf_{x \in \mathbb{R}} \exp^{-|x|} = ? \tag{2.1}$$

- Is the minimum attained?
- **Bolzano-Weierstrass Theorem** Every continuous function f attains its infimum over a compact set X. That is, there exists an $x^* \in X$ such that

$$f(x^*) = \inf_{x \in X} f(x) \tag{2.2}$$

Existence of Optimal Solution

- Consider the following problem

$$\inf_{x\in\mathbb{R}} \exp^{-|x|} = ? \tag{2.1}$$

- Is the minimum attained?
- **Bolzano-Weierstrass Theorem** Every continuous function f attains its infimum over a compact set X. That is, there exists an $x^* \in X$ such that

$$f(x^*) = \inf_{x\in X} f(x) \tag{2.2}$$

- Alternatively, if the level set (for some x^0)

$$\{x \mid f(x) \leq f(x^0)\} \tag{2.3}$$

of a continuous function f is compact, then the global min of

$$\min \; f(x), \quad \text{subject to } x \in \mathbb{R}^n \tag{2.4}$$

is attained

Checkable Conditions for Local Min

- Consider a **twice continuously differentiable** function f
- Given a point x, how to decide whether it is a local/global min

Checkable Conditions for Local Min

- Consider a **twice continuously differentiable** function f
- Given a point x, how to decide whether it is a local/global min
- First answer: check $f(x) \geq f(x^*)$ for all $x \in \mathbb{R}^n$

Checkable Conditions for Local Min

- Consider a **twice continuously differentiable** function f

- Given a point x, how to decide whether it is a local/global min

- First answer: check $f(x) \geq f(x^*)$ for all $x \in \mathbb{R}^n$

- Good enough?

Checkable Conditions for Local Min

- Consider a **twice continuously differentiable** function f
- Given a point x, how to decide whether it is a local/global min
- First answer: check $f(x) \geq f(x^*)$ for all $x \in \mathbb{R}^n$
- Good enough?
- We need easily checkable conditions

Checkable Conditions for Local Min

- Consider a **twice continuously differentiable** function f
- Given a point x, how to decide whether it is a local/global min
- First answer: check $f(x) \geq f(x^*)$ for all $x \in \mathbb{R}^n$
- Good enough?
- We need easily checkable conditions
- **Idea.** Use Taylor expansion to analyze local behavior around x

Checkable Conditions for Local Min

- We have the following sufficient conditions [makes sense?]

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

- Together they are "sufficient" for local min

Why Optimality Conditions?

- Optimality conditions are useful because:
  - provide guarantees for a candidate solution to be optimal (sufficient condition)
  - indicate when a point is NOT optimal (necessary condition)
- Guide the design of algorithm
  - Algorithms should look for points achieving the optimality conditions
  - Algorithm should stop when the optimality condition is approximately satisfied

Quadratic Problems

$$\begin{array}{rl} \text{minimize} & f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{Q}\mathbf{x} + \mathbf{b}^T\mathbf{x} \\ \text{subject to} & \mathbf{x} \in \mathbb{R}^n \end{array}$$

- Necessary condition for optimality

$$\nabla f(\mathbf{x}) = \mathbf{Q}\mathbf{x} + \mathbf{b} = 0, \quad \nabla^2 f(\mathbf{x}) = \mathbf{Q} \succeq 0$$

Convex Functions

- A continuous function $f : \mathbb{R}^n \mapsto \mathbb{R}$ is called *convex* if for all $x, y \in \mathbb{R}^n$ and for all $\lambda \in [0, 1]$, we have

$$f[\lambda x + (1 - \lambda) y] \leq \lambda f(x) + (1 - \lambda) f(y).$$

- A continuous function $f : \mathbb{R}^n \mapsto \mathbb{R}$ is called *concave* if for all $x, y \in \mathbb{R}^n$ and for all $\lambda \in [0, 1]$, we have

$$f[\lambda x + (1 - \lambda) y] \geq \lambda f(x) + (1 - \lambda) f(y).$$

Convex Sets

- A set $S \subseteq \mathbb{R}^n$ is convex if for any $x, y \in S$ and any $\lambda \in [0, 1]$, we have
$$\lambda x + (1 - \lambda) y \in S$$

- There are convex sets and non-convex sets
- There is no such thing as a "concave set"

Properties

- If $f(x)$ is a convex function, then $-f(x)$ is a concave function
- If $f_1(x), f_2(x)$ are both convex functions, then $g(x) = f_1(x) + f_2(x)$ are convex as well (prove?)
- If $f_1(x), f_2(x)$ are both convex functions, and $a \geq 0,\ b \geq 0$, then

$$g(x) = a \times f_1(x) + b \times f_2(x)$$

  are convex as well

Properties (Continued)

- Can we use other alternative, and perhaps simpler, ways to characterize the convexity/concavity?
- Yes we can!
- Given a smooth function with scalar variable; It is convex (resp. concave) if and only if its second-order derivative is nonnegative ≥ 0 (resp. nonpositive ≤ 0)
- For vector problems, the above condition becomes its Hessian matrix is positive semidefinite $\nabla^2 f(\mathbf{x}) \succeq 0, \ \forall \ \mathbf{x}$
- Go back to the quadratic problems?

Properties (continued)

- Generally speaking, for the following types of unconstrained problems

$$\min \; f(x), \quad \max \; f(x) \tag{2.5}$$

we have the following understanding:

	convex function	concave function
max	hard	easy
min	easy	hard

Properties

- Let's pick one problem and see why

Properties

- Let's pick one problem and see why
- Consider minimizing a convex function

Properties

- Let's pick one problem and see why
- Consider minimizing a convex function
- **Claim** any local minimum is global minimum

Properties

- Let's pick one problem and see why
- Consider minimizing a convex function
- **Claim** any local minimum is global minimum
- Proven using definition of convexity
  - Suppose $\bar{x}$ is local but not global minimum
  - Then there exists $f(x) < f(\bar{x})$
  - Due to convexity, for any $c \in (0, 1)$

$$\begin{aligned} f[c\bar{x} + (1-c)x] &\leq cf(\bar{x}) + (1-c)f(x) \\ &< cf(\bar{x}) + (1-c)f(\bar{x}) \\ &= f(\bar{x}) \end{aligned}$$

Properties

- Let's pick one problem and see why
- Consider minimizing a convex function
- **Claim** any local minimum is global minimum
- Proven using definition of convexity
  - Suppose $\bar{x}$ is local but not global minimum
  - Then there exists $f(x) < f(\bar{x})$
  - Due to convexity, for any $c \in (0, 1)$

$$
\begin{aligned}
f[c\bar{x} + (1-c)x] &\leq cf(\bar{x}) + (1-c)f(x) \\
&< cf(\bar{x}) + (1-c)f(\bar{x}) \\
&= f(\bar{x})
\end{aligned}
$$

- Contradiction to $\bar{x}$ being local optimal (why?)

Gradient Descent Methods

The Gradient Descent Algorithm

- We will start with a family of classical method: Gradient Descent

The Gradient Descent Algorithm

- We will start with a family of classical method: Gradient Descent
- If $\nabla f(\mathbf{x}) = 0$, then $\mathbf{x}$ is a candidate solution (satisfying first-order sufficient condition); Done

The Gradient Descent Algorithm

- We will start with a family of classical method: Gradient Descent
- If $\nabla f(\mathbf{x}) = 0$, then $\mathbf{x}$ is a candidate solution (satisfying first-order sufficient condition); Done
- If $\nabla f(\mathbf{x}) \neq 0$, there is an interval $(0, \delta)$ of stepsizes such that

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

The Gradient Descent Algorithm

- We will start with a family of classical method: Gradient Descent
- If $\nabla f(\mathbf{x}) = 0$, then $\mathbf{x}$ is a candidate solution (satisfying first-order sufficient condition); Done
- If $\nabla f(\mathbf{x}) \neq 0$, there is an interval $(0, \delta)$ of stepsizes such that

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

- Show this using Mean Value Theorem?

The Gradient Descent Algorithm

- We will start with a family of classical method: Gradient Descent
- If $\nabla f(\mathbf{x}) = 0$, then $\mathbf{x}$ is a candidate solution (satisfying first-order sufficient condition); Done
- If $\nabla f(\mathbf{x}) \neq 0$, there is an interval $(0, \delta)$ of stepsizes such that

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

- Show this using Mean Value Theorem?
- More generally, if a given direction $\mathbf{d}$ that is with obtuse angle with $\nabla f(\mathbf{x})$

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

there is an interval $(0, \delta)$ of stepsizes such that [try to prove]

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

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 descent 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

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

- Special case II: Newton's method

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

Choice of Stepsizes

- **Constant Stepsize**:

$$\alpha_r = \alpha$$

Comment: practically used often, but what's the constant?

- **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 solving the optimization problem may be expensive

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

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

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
- 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 topic: theoretical analysis of descent methods

Convergence Rate Analysis (Overview)

- Analyze convergence behavior, focused on the "Steepest gradient descent" method

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

- **Question**: When does the algorithm converge, to what solution?

- **Question**: How fast does the algorithm converge?

Convergence of Iterative Methods

- **Convergence to stationary solutions**
  - Sanity check
  - Minimal requirement of any reasonable algorithm
  - Does not give global efficiency of the algorithm
  - Linear rate/Supperlinear rate/Sublinear rate

Convergence of Iterative Methods

- **Convergence to stationary solutions**
  - Sanity check
  - Minimal requirement of any reasonable algorithm
  - Does not give global efficiency of the algorithm
  - Linear rate/Supperlinear rate/Sublinear rate
- **Iteration complexity analysis (convergence rate)**
  - Measures the number of iterations required to get an optimal solution (e.g., $f(\mathbf{x}^r) - f(\mathbf{x}^*) \leq \epsilon$)
  - Current analysis is all for the worst case
  - Gives global behavior of the algorithm

The Analysis of GD Method

- Suppose exists a constant L, which bounds the maximum eigenvalue of Hessian matrix of f:

$$\nabla^2 f(\mathbf{x}) \preceq L\mathbf{I},$$

where $\mathbf{I}$ is an identity matrix

- This implies that the curvature of the function is bounded
- Then we have

$$\begin{aligned} f(\mathbf{x}) &\overset{(i)}{=} f(\mathbf{y}) + \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + (\mathbf{x} - \mathbf{y})^T \nabla^2 f(\tilde{\mathbf{x}})(\mathbf{x} - \mathbf{y}) \\ &\leq f(\mathbf{y}) + \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + \frac{L}{2}\|\mathbf{x} - \mathbf{y}\|^2 := u(\mathbf{x}; \mathbf{y}), \ \forall \ \mathbf{x}, \mathbf{y} \end{aligned}$$

where the first step is due to mean value theorem; the second step is due to the boundedness of Hessian

The Descent Lemma

- The same result, but stated in a slightly different way
- **Key Lemma**: The Descent Lemma

  Suppose $f : \mathbb{R}^n \to \mathbb{R}$ has Lipschitz gradient

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|, \ \forall \ \mathbf{x}, \mathbf{y} \in \mathbb{R}^n$$

  Then we have

$$f(\mathbf{x}) \leq f(\mathbf{y}) + \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + \frac{L}{2}\|\mathbf{x} - \mathbf{y}\|^2 := u(\mathbf{x}; \mathbf{y}), \ \forall \ \mathbf{x}, \mathbf{y}$$

- Read Prop. A. 24 of Bertsekas for proof

Apply the Descent Lemma

- **Example**: for a quadratic problem with $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x}$, what is L? can you verify that the descent lemma is true?
- Replace $\mathbf{y}$ by $\mathbf{x}^r$ in the descent lemma

$$f(\mathbf{x}) \le f(\mathbf{x}^r) + \langle \nabla f(\mathbf{x}^r), \mathbf{x} - \mathbf{x}^r \rangle + \frac{L}{2}\|\mathbf{x} - \mathbf{x}^r\|^2 := u(\mathbf{x}; \mathbf{x}^r), \ \forall \ \mathbf{x}$$

- Minimize the r.h.s. with respect to $\mathbf{x}$, and let $\mathbf{x}^* = \mathbf{x}^{r+1}$:

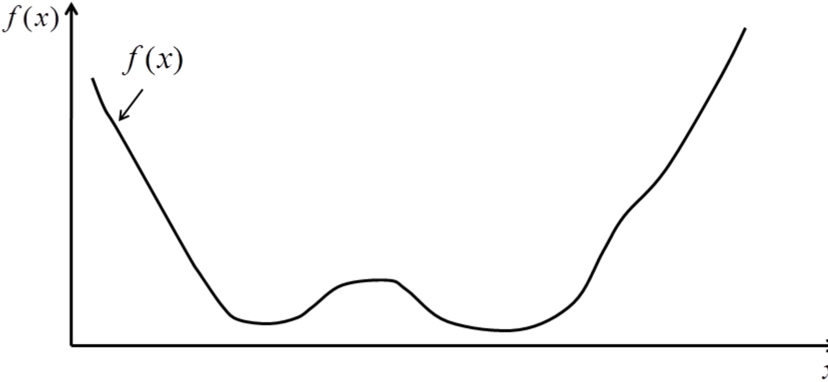
$$\mathbf{x}^{r+1} = \mathbf{x}^r - \frac{1}{L}\nabla f(\mathbf{x}^r)$$

- **Claim** such $\mathbf{x}^{r+1}$ always decreases the objective! Why?

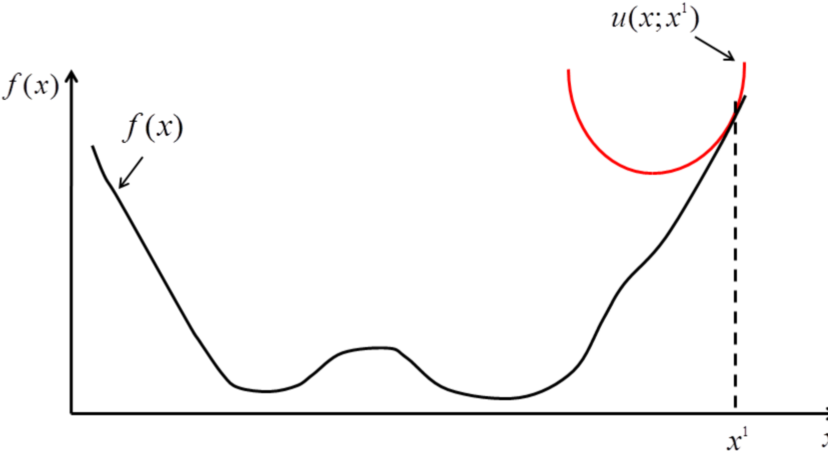
- By how much? Plug the expression $\mathbf{x}^{r+1}$ into the descent:

$$f(\mathbf{x}^{r+1}) \le f(\mathbf{x}^r) - \frac{1}{2L}\|\nabla f(\mathbf{x}^r)\|^2$$

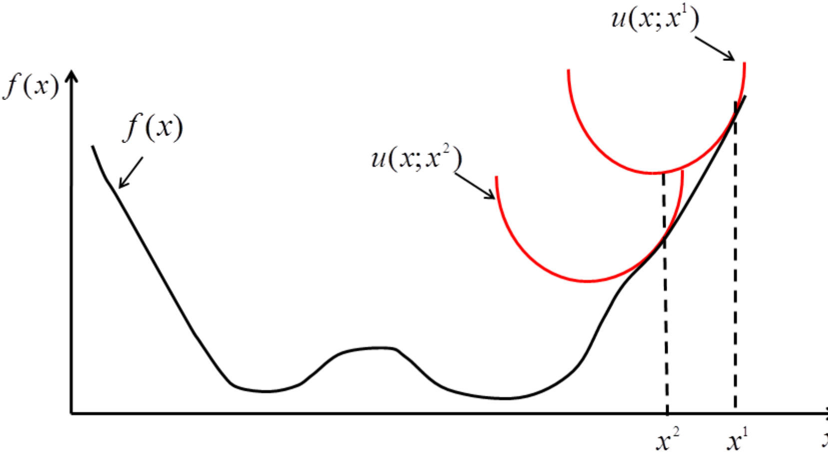
One Missing Piece



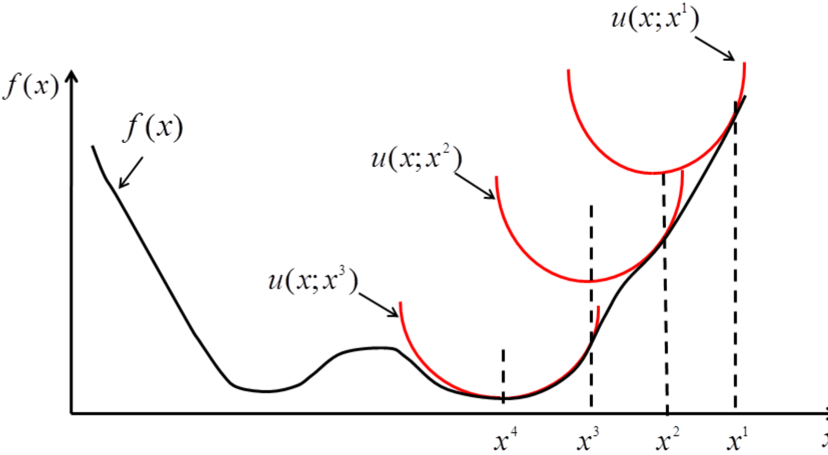
One Missing Piece



One Missing Piece



One Missing Piece



Convergence Analysis

- **Convergence**: Must at least converge to first order optimality ($\nabla f(\mathbf{x}) = 0$)
  1. Basic requirement
  2. Not necessarily convergent to global, or local, optimal
  3. This only says that the algorithm is reasonable

- Provide an analysis of GD for constant stepsize rule, for any reasonable direction (not necessarily $-\nabla f(\mathbf{x})$)

A General Analysis for Convergence

- Prove convergence to points that satisfy the first order optimality
  - we call them stationary solutions
- The direction $\mathbf{d}^r$ cannot be orthogonal to $\nabla f(\mathbf{x}^r)$ [figure]

A General Analysis for Convergence

- Prove convergence to points that satisfy the first order optimality – we call them stationary solutions
- The direction $\mathbf{d}^r$ cannot be orthogonal to $\nabla f(\mathbf{x}^r)$ [figure]
- **Gradient related condition**: For any sequence $\{\mathbf{x}^r\}$ that converges to a nonstationary point, the corresponding direction $\{\mathbf{d}^r\}$ is bounded and satisfies

$$\lim_{r\to\infty} \langle \nabla f(\mathbf{x}^r), \mathbf{d}^r \rangle < 0$$

A General Analysis for Convergence

- Prove convergence to points that satisfy the first order optimality – we call them stationary solutions
- The direction $\mathbf{d}^r$ cannot be orthogonal to $\nabla f(\mathbf{x}^r)$ [figure]
- **Gradient related condition**: For any sequence $\{\mathbf{x}^r\}$ that converges to a nonstationary point, the corresponding direction $\{\mathbf{d}^r\}$ is bounded and satisfies

$$\lim_{r\to\infty} \langle \nabla f(\mathbf{x}^r), \mathbf{d}^r \rangle < 0$$

- Is this condition satisfied for $\mathbf{d}^r = -\mathbf{D}^r \nabla f(\mathbf{x}^r)$ with $\mathbf{D}^r \succ 0$?

A General Analysis of Convergence

- Here we state a few assumptions about the algorithm/problem
- Let $\mathbf{x}^r$ be a sequence generated by a gradient method

$$\mathbf{x}^{r+1} = \mathbf{x}^r + \alpha_r \mathbf{d}^r$$

- $\mathbf{d}^r$ is gradient related
- Assume the following Lipschitz continuous condition is satisfied

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \le L\|\mathbf{x} - \mathbf{y}\|, \; \forall \; \mathbf{x}, \mathbf{y} \in \mathbb{R}^n$$

- Note, the problem is not necessarily second-order differentiable

A General Analysis of Convergence (Assumptions)

- Assume either one of the following choices of stepsize
  1. There exits a scalar ϵ such that for all r
  $$\epsilon < \alpha_r \leq -\frac{(2-\epsilon)\langle \nabla f(\mathbf{x}^r), \mathbf{d}^r \rangle}{L\|\mathbf{d}^r\|^2}$$
  2. $\alpha_r \to 0$, and $\sum_{r=1}^{\infty} \alpha_r = \infty$ (i.e., $\alpha_r = \frac{1}{r}$)

A General Analysis of Convergence (Assumptions)

- Assume either one of the following choices of stepsize
  1. There exits a scalar ϵ such that for all r
  $$\epsilon < \alpha_r \le -\frac{(2-\epsilon)\langle \nabla f(\mathbf{x}^r), \mathbf{d}^r\rangle}{L\|\mathbf{d}^r\|^2}$$
  2. $\alpha_r \to 0$, and $\sum_{r=1}^{\infty} \alpha_r = \infty$ (i.e., $\alpha_r = \frac{1}{r}$)

- **Claim**: $\nabla f(\mathbf{x}^r) \to 0$, or $f(\mathbf{x}^r) \to -\infty$

A General Analysis of Convergence (Assumptions)

- Assume either one of the following choices of stepsize
  1. There exits a scalar ϵ such that for all r
  $$\epsilon < \alpha_r \leq -\frac{(2-\epsilon)\langle \nabla f(\mathbf{x}^r), \mathbf{d}^r \rangle}{L\|\mathbf{d}^r\|^2}$$
  2. $\alpha_r \to 0$, and $\sum_{r=1}^{\infty} \alpha_r = \infty$ (i.e., $\alpha_r = \frac{1}{r}$)

- **Claim**: $\nabla f(\mathbf{x}^r) \to 0$, or $f(\mathbf{x}^r) \to -\infty$

- If $\mathbf{d}^r = -\nabla f(\mathbf{x}^r)$, then the first condition becomes
$$\epsilon < \alpha_r \leq \frac{(2-\epsilon)}{L}$$
Therefore, we can pick, for example, $\alpha_r = \frac{1}{L}$

Convergence Analysis

- Given $\mathbf{x}^r$ and the descent direction $\mathbf{d}^r$, the Lipschitz assumption implies (cf. the descent Lemma)

$$f(\mathbf{x}^r + \alpha_r \mathbf{d}^r) - f(\mathbf{x}^r) \leq \alpha_r \langle \nabla f(\mathbf{x}^r), \mathbf{d}^r \rangle + \frac{L}{2} \alpha_r^2 \|\mathbf{d}^r\|^2$$

- Plugin the upper-bound for α_r:

$$f(x^r + \alpha_r \mathbf{d}^r) - f(\mathbf{x}^r) \leq \underbrace{-\epsilon(2-\epsilon)}_{<0} \frac{(\langle \nabla f(\mathbf{x}^r), \mathbf{d}^r \rangle)^2}{2L\|\mathbf{d}^r\|^2}$$

- Clearly, the objective is always decreasing
- Question: Where have we used the "gradient related" condition?

Convergence Analysis (cont.)

- Assume $\bar{x}$ is a finite nonstationary limit point.
- Then $f(\mathbf{x}^r) \downarrow f(\bar{\mathbf{x}})$
- Then $\langle \nabla f(\mathbf{x}^r), \mathbf{d}^r \rangle \to 0$, because otherwise $f(\bar{\mathbf{x}}) \to -\infty$

Convergence Analysis (cont.)

- Assume $\bar{x}$ is a finite nonstationary limit point.
- Then $f(\mathbf{x}^r) \downarrow f(\bar{\mathbf{x}})$
- Then $\langle \nabla f(\mathbf{x}^r), \mathbf{d}^r \rangle \to 0$, because otherwise $f(\bar{\mathbf{x}}) \to -\infty$
- Is this possible? No, the gradient related condition asserts that $\langle \nabla f(\mathbf{x}^r), \mathbf{d}^r \rangle < 0$, and the condition on α_r asserts that $\alpha_r > 0$
- We arrived at a contradiction – the claim is proved
- How about the diminishing stepsize? Same analysis, but much more involved

Diminishing Stepsizes

Proposition 1.2.4: (Convergence for a Diminishing Stepsize) Let $\{x^k\}$ be a sequence generated by a gradient method $x^{k+1} = x^k + \alpha^k d^k$. Assume that for some constant $L > 0$, we have

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|, \qquad \forall\, x, y \in \Re^n, \tag{1.26}$$

and that there exist positive scalars c_1, c_2 such that for all k we have

$$c_1\|\nabla f(x^k)\|^2 \leq -\nabla f(x^k)' d^k, \qquad \|d^k\|^2 \leq c_2\|\nabla f(x^k)\|^2. \tag{1.27}$$

Suppose also that

$$\alpha^k \to 0, \qquad \sum_{k=0}^{\infty} \alpha^k = \infty.$$

Then either $f(x^k) \to -\infty$ or else $\{f(x^k)\}$ converges to a finite value and $\nabla f(x^k) \to 0$. Furthermore, every limit point of $\{x^k\}$ is a stationary point of f.

Convergence Rate Analysis

- Define an ϵ optimal solution as $\{\mathbf{x}_\epsilon := f(\mathbf{x}^r) - f^* \leq \epsilon\}$
- **Convergence Rate**:
  1. Measures the number of iterations required to get an ϵ optimal solution
  2. Gives global behavior of the algorithm
  3. A popular and important measure for evaluating algorithms in big data related applications
  4. Question: What determines the convergence rate?

Convergence Rate Analysis

- We focus on a family of special functions, and show that gradient descent methods is able to converge linearly
- This means some measure of optimality shrinks by a constant factor at each iteration
- For example, define the error as $e(\mathbf{x}) := f(\mathbf{x}) - f(\mathbf{x}^*) \geq 0$
- Then,e.g., $e(\mathbf{x}^{r+1}) \leq 0.1 \times e(\mathbf{x}^r)$

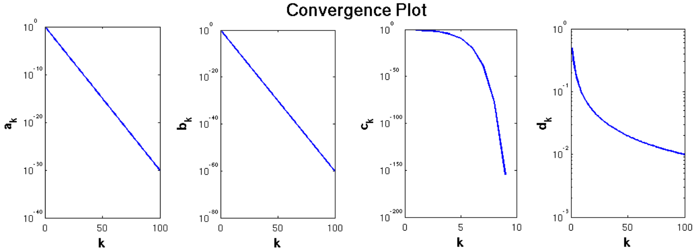


Figure 3.1: Illustration of Convergence Speed. [1] Linear rate (slower); [2] Linear rate (faster);[3] superlinear rate; [4] sublinear rate. y-axis: log of the error, x-axis: iteration number.(Wikipedia: Rate of Convergence)

Convergence Rate Analysis

- Linear convergence means that there exists $\beta \in (0,1)$

$$\lim_{r\to\infty} \sup \frac{e(\mathbf{x}^{r+1})}{e(\mathbf{x}^r)} < \beta$$

Convergence Rate Analysis

- Linear convergence means that there exists $\beta \in (0,1)$

$$\lim_{r\to\infty} \sup \frac{e(\mathbf{x}^{r+1})}{e(\mathbf{x}^r)} < \beta$$

- If the above is true for all r, it means $e(\mathbf{x}^{r+1}) \le (\beta)^r e(\mathbf{x}^0)$, or (note $\beta < 1$, so $\ln(\beta) < 0$)

$$\underbrace{\ln(e(\mathbf{x}^{r+1}))}_{\text{log of error}} \le \underbrace{r\ln(\beta)}_{\text{“linear” in iteration \#}} + \ln(e(\mathbf{x}^0))$$

- Log of error a linear function in # iteration r (with negative slope)!

Convergence Rate Analysis

- Linear convergence means that there exists $\beta \in (0,1)$

$$\lim_{r\to\infty} \sup \frac{e(\mathbf{x}^{r+1})}{e(\mathbf{x}^r)} < \beta$$

- If the above is true for all r, it means $e(\mathbf{x}^{r+1}) \le (\beta)^r e(\mathbf{x}^0)$, or (note $\beta < 1$, so $\ln(\beta) < 0$)

$$\underbrace{\ln(e(\mathbf{x}^{r+1}))}_{\text{log of error}} \le \underbrace{r\ln(\beta)}_{\text{“linear” in iteration \#}} + \ln(e(\mathbf{x}^0))$$

- Log of error a linear function in # iteration r (with negative slope)!
- Superliner convergence means

$$\lim_{r\to\infty} \sup \frac{e(\mathbf{x}^{r+1})}{e^p(\mathbf{x}^r)} < \beta$$

for some constant $p > 1$

Convergence Rate Analysis

- Why "$\limsup$" is needed?
- This says that as long as in the limit, the linear convergence behavior occurs, we call the algorithm "linearly convergent"
- But we will analyze an algorithm that is much "stronger", in the sense it satisfies (for all iterations)

$$\frac{e(\mathbf{x}^{r+1})}{e(\mathbf{x}^r)} < \beta, \ \forall\, r$$

with some $\beta \in (0, 1)$.

Strongly Convex Functions

- Recall that f is continuously differentiable, f is convex iff

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle.$$

- If f is twice continuously differentiable, then

$$f \text{ is convex} \Leftrightarrow \nabla^2 f(\mathbf{x}) \succeq 0, \text{ for all } \mathbf{x}$$

- **A New Notion**: f is strongly convex iff exists $\sigma > 0$

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{\sigma}{2}\|\mathbf{x} - \mathbf{y}\|^2.$$

- The entire function has "enough curvature" [graphically]

Strongly Convex Functions

- A function is "strongly convex" if (intuitively)
  1. It is convex
  2. It has no "flat" regions

- If f is twice continuously differentiable, then exists $\sigma > 0$

$$f \text{ is strongly convex} \Leftrightarrow \nabla^2 f(\mathbf{x}) \succeq \sigma \mathbf{I}, \text{ for all } \mathbf{x}$$

- Note, for two matrices, the notation $A \succeq B$ means $A - B \succeq 0$. That is, $A - B$ is a positive semidefinite matrix

- **Example**: A quadratic function $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x}$ with strictly positive definite $\mathbf{A}$, i.e., $\mathbf{A} \succeq \sigma\mathbf{I}$; Here σ is the smallest eigenvalue for $\mathbf{A}$

GD for SC Function: Step 1

- Let's begin the analysis of gradient descent for strongly convex problems
- **Goal** To see what does convergence speed depends on

GD for SC Function: Step 1

- Let's begin the analysis of gradient descent for strongly convex problems
- **Goal** To see what does convergence speed depends on
- For simplicity, consider the case where $\mathbf{d}^r = -\nabla f(\mathbf{x}^r)$
- Then $\alpha_r = \frac{1}{L}$ and

$$f(\mathbf{x}^{r+1}) = f(\mathbf{x}^r + \alpha_r \mathbf{d}^r) \leq f(\mathbf{x}^r) - \frac{1}{2L}\|\nabla f(\mathbf{x}^r)\|^2$$

GD for SC Function: Step 1

- Let's begin the analysis of gradient descent for strongly convex problems

- **Goal** To see what does convergence speed depends on

- For simplicity, consider the case where $\mathbf{d}^r = -\nabla f(\mathbf{x}^r)$

- Then $\alpha_r = \frac{1}{L}$ and

$$f(\mathbf{x}^{r+1}) = f(\mathbf{x}^r + \alpha_r \mathbf{d}^r) \le f(\mathbf{x}^r) - \frac{1}{2L}\|\nabla f(\mathbf{x}^r)\|^2$$

- This shows that the after one round of algorithm, the objective function achieves "sufficient descent"

- The amount of the descent can be measured by the size of the gradient

GD for SC Function: Step 2

- Using the strong convexity assumption, we have

$$f(\mathbf{x}^*) \geq f(\mathbf{x}^r) + \langle \nabla f(\mathbf{x}^r), \mathbf{x}^* - \mathbf{x}^r \rangle + \frac{\sigma}{2}\|\mathbf{x}^r - \mathbf{x}^*\|^2 \qquad (3.6)$$

GD for SC Function: Step 2

- Using the strong convexity assumption, we have

$$f(\mathbf{x}^*) \geq f(\mathbf{x}^r) + \langle \nabla f(\mathbf{x}^r), \mathbf{x}^* - \mathbf{x}^r \rangle + \frac{\sigma}{2}\|\mathbf{x}^r - \mathbf{x}^*\|^2 \qquad (3.6)$$

- Let us view the right hand side function as a function of $\mathbf{x}^*$

$$g(x^*) = f(\mathbf{x}^r) + \langle \nabla f(\mathbf{x}^r), \mathbf{x}^* - \mathbf{x}^r \rangle + \frac{\sigma}{2}\|\mathbf{x}^r - \mathbf{x}^*\|^2 \qquad (3.7)$$

GD for SC Function: Step 2

- Using the strong convexity assumption, we have

$$f(\mathbf{x}^*) \geq f(\mathbf{x}^r) + \langle \nabla f(\mathbf{x}^r), \mathbf{x}^* - \mathbf{x}^r \rangle + \frac{\sigma}{2}\|\mathbf{x}^r - \mathbf{x}^*\|^2 \tag{3.6}$$

- Let us view the right hand side function as a function of $\mathbf{x}^*$

$$g(x^*) = f(\mathbf{x}^r) + \langle \nabla f(\mathbf{x}^r), \mathbf{x}^* - \mathbf{x}^r \rangle + \frac{\sigma}{2}\|\mathbf{x}^r - \mathbf{x}^*\|^2 \tag{3.7}$$

- Minimizing the right hand side over $\mathbf{x}^*$ we obtain (optimal solution is $\mathbf{x}^* = \mathbf{x}^r - \frac{1}{\sigma}\nabla f(\mathbf{x}^r)$)

$$\begin{aligned} f(\mathbf{x}^*) &\geq f(\mathbf{x}^r) + \langle \nabla f(\mathbf{x}^r), \mathbf{x}^* - \mathbf{x}^r \rangle + \frac{\sigma}{2}\|\mathbf{x}^r - \mathbf{x}^*\|^2 \\ &\geq f(\mathbf{x}^r) - \frac{1}{2\sigma}\|\nabla f(\mathbf{x}^r)\|^2 \end{aligned}$$

The current obj value is not too far away from the goal

GD for SC Function: Step 3

- Step 1 tells us how much descent we have at each step
- Step 2 tells us how close we are to the global min
- Combing the previous two steps, we have

$$
\begin{aligned}
f(\mathbf{x}^{r+1}) - f(x^*) &\overset{\text{Step 1}}{\leq} f(\mathbf{x}^r) - f(\mathbf{x}^*) - \frac{1}{2L}\|\nabla f(\mathbf{x}^r)\|^2 \\
&\overset{\text{Step 2}}{\leq} f(\mathbf{x}^r) - f(\mathbf{x}^*) - \frac{\sigma}{L}(f(\mathbf{x}^r) - f(\mathbf{x}^*))
\end{aligned}
$$

- Rearranging terms, use the definition $e(\mathbf{x}) := f(\mathbf{x}) - f(\mathbf{x}^*)$:

$$
e(\mathbf{x}^{r+1}) \leq (1 - \frac{\sigma}{L})e(\mathbf{x}^r) := \beta e(\mathbf{x}^r)
$$

GD for SC Functions

- Linear convergence, with constant $\beta = (1 - \frac{\sigma}{L}) \in (0, 1)$ (why?)
- L/σ is so-called the condition number of f
  - σ: the smallest eigenvalue of the Hessian of f (recall our quadratic problem)
  - L: the largest eigenvalue of the Hessian of f (recall our quadratic problem)
- Large condition number implies large β
- L/σ big: ill-conditioned (slow convergence for GD)
- L/σ small: well-conditioned (fast convergence for GD)

Summary of the Proof

- **Summary of proof**: Two steps:
  1. S1: Sufficient Descent (the decrease achieved after each iteration):
  $$f(\mathbf{x}^{r+1}) - f(\mathbf{x}^r) \le \ldots$$
  2. S2: Estimating cost-to-go (how far away we are from the optimal):
  $$f(\mathbf{x}^r) - f(\mathbf{x}^*) \le \ldots$$

GD for SC Function

- Large condition number means the problem is badly scaled, slow convergence of the algorithm.

GD for SC Function

- Large condition number means the problem is badly scaled, slow convergence of the algorithm.
- How many iterations are needed for $e(\mathbf{x}^r)$ to reach below ϵ?

GD for SC Function

- Large condition number means the problem is badly scaled, slow convergence of the algorithm.
- How many iterations are needed for $e(\mathbf{x}^r)$ to reach below ϵ?
- Suppose $e(\mathbf{x}^0) = D_0$, then $e(\mathbf{x}^r) = \beta^r D_0 \leq \epsilon$, so we require:

$$r \geq -\ln(D_0/\epsilon)/\ln(\beta) = \ln\left(\frac{D_0}{\epsilon}\right)/\ln(\frac{1}{\beta})$$

- As long as the total # of iteration is larger than the right hand side above, we are guaranteed to reach an ϵ optimal solution

GD for SC Function

- Large condition number means the problem is badly scaled, slow convergence of the algorithm.
- How many iterations are needed for $e(\mathbf{x}^r)$ to reach below ϵ?
- Suppose $e(\mathbf{x}^0) = D_0$, then $e(\mathbf{x}^r) = \beta^r D_0 \le \epsilon$, so we require:

$$r \ge -\ln(D_0/\epsilon)/\ln(\beta) = \ln\left(\frac{D_0}{\epsilon}\right)/\ln(\frac{1}{\beta})$$

- As long as the total # of iteration is larger than the right hand side above, we are guaranteed to reach an ϵ optimal solution
- Larger the β, the large the number of iterations required!
- The number of iterations scale with error by: $\ln(1/\epsilon)$

GD for SC Function

- Large condition number means the problem is badly scaled, slow convergence of the algorithm.
- How many iterations are needed for $e(\mathbf{x}^r)$ to reach below ϵ?
- Suppose $e(\mathbf{x}^0) = D_0$, then $e(\mathbf{x}^r) = \beta^r D_0 \le \epsilon$, so we require:

$$r \ge -\ln(D_0/\epsilon)/\ln(\beta) = \ln\left(\frac{D_0}{\epsilon}\right)/\ln(\frac{1}{\beta})$$

- As long as the total # of iteration is larger than the right hand side above, we are guaranteed to reach an ϵ optimal solution
- Larger the β, the large the number of iterations required!
- The number of iterations scale with error by: $\ln(1/\epsilon)$
- **Remark:** For regular convex problems, sublinear convergence rate $r \ge 1/(\epsilon)$ (proof omitted)

Other First-Order Methods

Other 1st-Order Methods: Incremental Gradient

- Let's consider the least square problem (with m data points)

$$\min \quad f(\mathbf{x}) = \frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2 = \frac{1}{2}\sum_{i=1}^{m}\|\mathbf{A}^i\mathbf{x} - \mathbf{b}^i\|^2 := \frac{1}{2}\sum_{i=1}^{m} g_i(\mathbf{x})$$

  $\mathbf{A}^i$, $\mathbf{b}^i$ represents the ith row of $\mathbf{A}$ and $\mathbf{b}$, or the ith piece of data

- The gradient method needs **all data** (or all $g_i's$) in each iteration

Other 1st-Order Methods: Incremental Gradient

- Let's consider the least square problem (with m data points)

$$\min \quad f(\mathbf{x}) = \frac{1}{2}\|\mathbf{A}\mathbf{x}-\mathbf{b}\|^2 = \frac{1}{2}\sum_{i=1}^{m}\|\mathbf{A}^i\mathbf{x}-\mathbf{b}^i\|^2 := \frac{1}{2}\sum_{i=1}^{m} g_i(\mathbf{x})$$

  $\mathbf{A}^i$, $\mathbf{b}^i$ represents the ith row of $\mathbf{A}$ and $\mathbf{b}$, or the ith piece of data

- The gradient method needs **all data** (or all $g_i's$) in each iteration

- Consider the following incremental method: At iteration $r+1$

Let $\quad \psi_0 = \mathbf{x}^r$

Inner Loop $\quad \psi_i = \psi_{i-1} - \underbrace{\alpha_r \nabla g_i(\psi_{i-1})}_{\text{a single data point}}, \; i = 1, \cdots, m$

Update the variable $\quad \mathbf{x}^{r+1} = \psi_m$

- Total m inner loops; What is the advantage of incrementalism?

View as Gradient Method with Errors

- View the incremental method as gradient method with errors

$$\mathbf{x}^{r+1} = \underbrace{\mathbf{x}^r - \alpha_r \sum_{i=1}^{m} \nabla g_i(\mathbf{x}^r)}_{\text{The usual gradient step}} \underbrace{+ \alpha_r \sum_{i=1}^{m} \left(\nabla g_i(\mathbf{x}^r) - \nabla g_i(\psi_{i-1})\right)}_{\text{The error term}}$$

- Error term proportional to stepsize α_r

View as Gradient Method with Errors

- View the incremental method as gradient method with errors

$$\mathbf{x}^{r+1} = \underbrace{\mathbf{x}^r - \alpha_r \sum_{i=1}^{m} \nabla g_i(\mathbf{x}^r)}_{\text{The usual gradient step}} \underbrace{+\alpha_r \sum_{i=1}^{m} \left(\nabla g_i(\mathbf{x}^r) - \nabla g_i(\psi_{i-1})\right)}_{\text{The error term}}$$

- Error term proportional to stepsize α_r
- Convergence or a diminishing stepsize: square summable, infinite travel

$$\alpha_r \to 0, \quad \underbrace{\sum_r \alpha_r = \infty}_{\text{infinite travel}}, \quad \underbrace{\sum_r \alpha_r^2 < \infty}_{\text{square summable}} \tag{4.1}$$

- Convergence to a neighborhood of $\mathbf{x}^*$ for a constant stepsize

Comments

- Incremental type of algorithm is an old algorithm; see the following for a survey
  "Incremental Gradient, Subgradient, and Proximal Methods for Convex Optimization: A Survey", D. P. Bertsekas 2010.

- Recently it has attracted significant attention in ML and optimization communities

- Significant progress has been made to develop variants of incremental algorithm that achieves linear convergence

- Closely related to the stochastic gradient descent (SGD) algorithm, decentralized algorithm, etc.