Optimization Theory and Applications

Kun Zhu (zhukun@nuaa.edu.cn)

November 19, 2018

Introduction

- We consider a class of search methods for real-valued functions in $\mathbb{R}^n$, which use the gradient of the given function
- **Definition**: In optimization, ***gradient method*** is an algorithm to solve problems of the form

$$\min_{\mathbf{x}\in\mathbb{R}^n} \quad f(\mathbf{x})$$

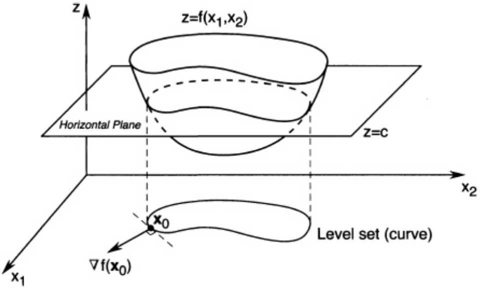
  with the search directions defined by the gradient of the function at the current point

Introduction

- Examples of gradient methods
  - Gradient descent
    - Fixed step-size descent
    - Steepest descent
    - Stochastic gradient descent
  - Conjugate gradient

Introduction

- **Remind**: a level set of a function $f : \mathbb{R}^n \to \mathbb{R}$ is the set of points $\mathbf{x}$ satisfying $f(\mathbf{x}) = c$ for some constant c



- $\mathbf{d} = \nabla f(\mathbf{x})$ points the direction of maximum rate of increase
- The direction $\mathbf{d} = -\nabla f(\mathbf{x})$ points the ***direction of maximum rate of decrease***

Introduction

- **Proof**:
  - Recall that the rate of increase of f at $\mathbf{x}$: $\langle \nabla f(\mathbf{x}), \mathbf{d} \rangle$, $\|\mathbf{d}\| = 1$
  - Recall Cauchy-Schwarz inequality: for any two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$

    $$|\langle \mathbf{x}, \mathbf{y} \rangle| \leq \|\mathbf{x}\| \|\mathbf{y}\|$$

    with the equality holds if and only if $\mathbf{x} = \alpha \mathbf{y}$
  - Based on Cauchy-Schawarz inequality, we have

    $$\langle \nabla f(\mathbf{x}), \mathbf{d} \rangle \leq \|\nabla f(\mathbf{x})\| \|\mathbf{d}\| = \|\nabla f(\mathbf{x})\|$$
  - The equality holds when $\mathbf{d} = \frac{\nabla f(\mathbf{x})}{\|\nabla f(\mathbf{x})\|}$

Introduction

- Main idea of the ***gradient descent algorithm***:

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \alpha_k \nabla f(\mathbf{x}^{(k)})$$

  - α_k is the ***step size***, and $\nabla f(\mathbf{x}^{(k)})$ points the ***direction***

- The gradient varies during the search proceeds, and tends to 0 as we approach the minimizer

Introduction

- For sufficiently small step size, the gradient algorithm has ***descent property***
- **Proposition**: Suppose $\nabla f(\mathbf{x}^{(\mathbf{k})}) \neq 0$, there exists $\bar{\alpha} > 0$ such that for all $\alpha_k \in (0, \bar{\alpha})$, we have

$$f(\mathbf{x}^{(\mathbf{k+1})}) < f(\mathbf{x}^{(\mathbf{k})})$$

Introduction

- **Proof**:
  - Consider $\phi(\alpha) = f(\mathbf{x}^{(k)} - \alpha \nabla f(\mathbf{x}^{(k)}))$
  - By chain rule, we have
    $$\phi'(0) = -\|\nabla f(\mathbf{x}^{(k)})\|^2 < 0$$
  - Hence, there exists $\bar{\alpha} > 0$ such that for all $\alpha_k \in (0, \bar{\alpha})$, we have
    $$\phi(\alpha_k) < \phi(0)$$
  - Rewriting, we obtain
    $$f(\mathbf{x}^{(k+1)}) < f(\mathbf{x}^{(k)})$$

Introduction

- The impact of choices for α_k
  - If α_k too small, we need to iterate many times to get to the solution
  - If α_k too large, algorithm may zig-zag around the solution (overshoot)
- Step size α_k can be chosen in many different ways
  - We can either fix $\alpha_k = \alpha$ for all k, or let α_k vary from iteration to iteration
  - Aggressive scheme:

$$\alpha_k = \arg\min_{\alpha \geq 0} f(\mathbf{x}^{(\mathbf{k})} - \alpha \nabla f(\mathbf{x}^{(k)}))$$

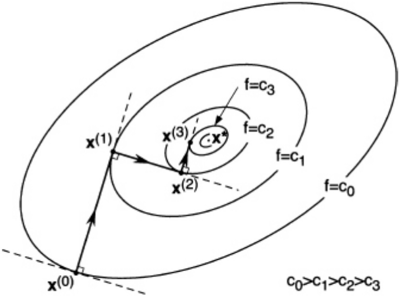
Method of Steepest Descent

- The method of ***steepest descent***:
  - It is a gradient method
  - The step size is chosen to achieve the maximum amount of decrease of the objective function at each individual step

$$\alpha_k = \arg\min_{\alpha \geq 0} f(\mathbf{x}^{(\mathbf{k})} - \alpha \nabla f(\mathbf{x}^{(\mathbf{k})}))$$

Method of Steepest Descent

- The procedures of the steepest descent method:
  - At each step, starting from $\mathbf{x}^{(\mathbf{k})}$
  - We conduct a ***line search*** in the direction $-\nabla f(\mathbf{x}^{(\mathbf{k})})$ until a minimizer $\mathbf{x}^{(\mathbf{k+1})}$ is found
- The following figure shows a typical sequence resulting from the steepest descent method



Method of Steepest Descent

- The steepest descent algorithm moves in orthogonal steps, which is proved in the following proposition
- ***Proposition:*** If $\{\mathbf{x}^{(\mathbf{k})}\}_{k=0}^{\infty}$ is a steepest descent sequences for a given function $f : \mathbb{R}^n \rightarrow \mathbb{R}$, then for each k the vector $\mathbf{x}^{(\mathbf{k+1})} - \mathbf{x}^{(\mathbf{k})}$ is orthogonal to the vector $\mathbf{x}^{(\mathbf{k+2})} - \mathbf{x}^{(\mathbf{k+1})}$

Method of Steepest Descent

- Property of Steepest Descent Method
- ***Proposition:*** If $\{\mathbf{x}^{(\mathbf{k})}\}_{k=0}^{\infty}$ is the steepest descent sequence for $f : \mathbb{R}^n \to \mathbb{R}$ and if $\nabla f(\mathbf{x}^{(\mathbf{k})}) \neq 0$, then $f(\mathbf{x}^{(\mathbf{k+1})}) < f(\mathbf{x}^{(\mathbf{k})})$
  - For each new point generated, the corresponding function value decreases
  - The steepest descent method possesses the ***descent property***

Method of Steepest Descent

- **Proof**:
  - Given $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \alpha \nabla f(\mathbf{x}^{(k)})$, $\alpha_k \geq 0$ is the local minimizer of
  $$\phi_k(\alpha) = f(\mathbf{x}^{(k)} - \alpha \nabla f(\mathbf{x}^{(k)}))$$
  - Therefore, for all $\alpha \geq 0$, we have
  $$\phi_k(\alpha_k) \leq \phi_k(\alpha)$$
  - By chain rule, we have
  $$\phi'(0) = -\|\nabla f(\mathbf{x}^{(k)})\|^2 < 0$$
  - Hence, there exists $\bar{\alpha} > 0$ such that for all $\alpha \in (0, \bar{\alpha}]$, we have
  $$\phi_k(\alpha) < \phi_k(0)$$
  - Accordingly, we have
  $$f(\mathbf{x}^{(k+1)}) = \phi_k(\alpha_k) \leq \phi_k(\bar{\alpha}) < \phi_k(0) = f(\mathbf{x}^{(k)})$$
- Question: What if $\nabla f(\mathbf{x}^{(\mathbf{k})}) = 0$?

Method of Steepest Descent

- If $\nabla f(\mathbf{x}^{(\mathbf{k})}) = 0$, then the point $\mathbf{x}^{(\mathbf{k})}$ satisfies the FONC
- In this case, $\mathbf{x}^{(\mathbf{k+1})} = \mathbf{x}^{(\mathbf{k})}$, which can be used as the basis for a ***stopping criterion*** for the algorithm
- Question: is it a good choice to set $\nabla f(\mathbf{x}^{(\mathbf{k})}) = 0$ as the stopping criterion?

Method of Steepest Descent

- Design of practical stopping criterion
  - Check the norm of the gradient $\|\nabla f(\mathbf{x}^{(k)})\| < \varepsilon$
  - Check the absolute difference between objective function values of every two successive iterations $|f(\mathbf{x}^{(k+1)}) - f(\mathbf{x}^{(k)})| < \varepsilon$
  - Check the norm of the difference between two successive points $\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\| < \varepsilon$
  - Check the “relative” values of the quantities

$$\frac{|f(\mathbf{x}^{(k+1)}) - f(\mathbf{x}^{(k)})|}{|f(\mathbf{x}^{(k)})|} < \varepsilon$$

$$\frac{\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\|}{\|\mathbf{x}^{(k)}\|} < \varepsilon$$

Method of Steepest Descent

- **Note:** The "relative" stopping criteria are preferable to the "absolute" criteria because the relative criteria are "***scale-independent***"
  - Scaling the objective function does not change the satisfaction of the criterion $\frac{|f(\mathbf{x}^{(k+1)})-f(\mathbf{x}^{(k)})|}{|f(\mathbf{x}^{(k)})|} < \varepsilon$
  - Scaling the decision does not change the satisfaction of the criterion $\frac{\|\mathbf{x}^{(k+1)}-\mathbf{x}^{(k)}\|}{\|\mathbf{x}^{(k)}\|} < \varepsilon$
- To avoid dividing by very small numbers, we can modify the stopping criteria as follows:

$$\frac{|f(\mathbf{x}^{(k+1)})-f(\mathbf{x}^{(k)})|}{\max\{1,|f(\mathbf{x}^{(k)})|\}} < \varepsilon$$

$$\frac{\|\mathbf{x}^{(k+1)}-\mathbf{x}^{(k)}\|}{\max\{1,\|\mathbf{x}^{(k)}\|\}} < \varepsilon$$

Method of Steepest Descent

- **Example:** We use the method of steepest descent to find the minimizer of

$$f(x_1, x_2, x_3) = (x_1 - 4)^4 + (x_2 - 3)^2 + 4(x_3 + 5)^4$$

The initial point is $\mathbf{x}^{(0)} = [4, 2, -1]^T$, and perform three iterations

Method of Steepest Descent

- Iteration 1
  - Step 1: Compute the gradient of $f(\mathbf{x})$

    $\nabla f(\mathbf{x}) = [4(x_1 - 4)^3, 2(x_2 - 3), 16(x_3 + 5)^3]^T$

    $\nabla f(\mathbf{x}^{(0)}) = [0, -2, 1024]^T$

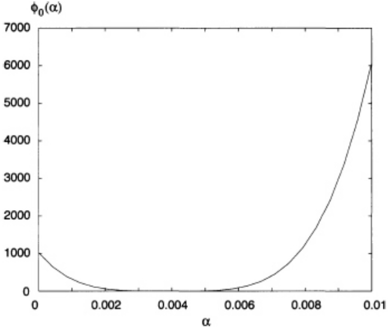
  - Step 2: Select step size α_0

$$\begin{aligned}\alpha_0 &= \arg\min_{\alpha \geq 0} f(\mathbf{x}^{(0)} - \alpha \nabla f(\mathbf{x}^{(0)})) \\ &= \arg\min_{\alpha \geq 0}(0 + (2 + 2\alpha - 3)^2 + 4(-1 - 1024\alpha + 5)^4) \\ &= \arg\min_{\alpha \geq 0} \phi_0(\alpha)\end{aligned}$$

  - Remind: $\phi_k(\alpha) = f(\mathbf{x}^{(0)} - \alpha \nabla f(\mathbf{x}^{(0)}))$

Method of Steepest Descent

- Using the secant method we obtain $\alpha_0 = 3.967 \times 10^{-3}$



- Step 3: Determine the next point $\mathbf{x}^{(1)}$

$$\mathbf{x}^{(1)} = \mathbf{x}^{(0)} - \alpha_0 \nabla f(\mathbf{x}^{(0)}) = [4, 2.008, -5.062]^T$$

Method of Steepest Descent

- Iteration 2
  - Step 1: Compute the gradient

    $\nabla f(\mathbf{x}^{(1)}) = [0, -1.984, -0.003875]^T$

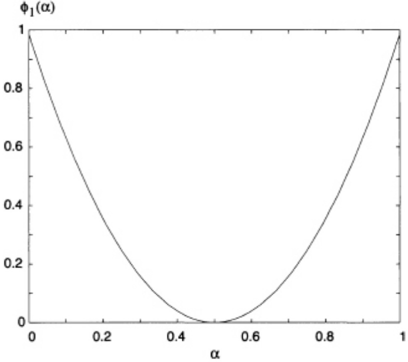
  - Step 2: Select step size α_1

    $\alpha_1 = \arg\min_{\alpha \geq 0}(0 + (2.008 + 1.984\alpha - 3)^2 + 4(-5.062 + 0.003875\alpha + 5)^4)$

    $= \arg\min_{\alpha \geq 0} \phi_1(\alpha)$

Method of Steepest Descent

- Using the secant method we obtain $\alpha_1 = 0.5$



- Step 3: Determine the next point $\mathbf{x}^{(2)}$

$$\mathbf{x}^{(2)} = \mathbf{x}^{(1)} - \alpha_1 \nabla f(\mathbf{x}^{(1)}) = [4, 3, -5.060]^T$$

Method of Steepest Descent

- Iteration 3
  - Step 1: Compute the gradient

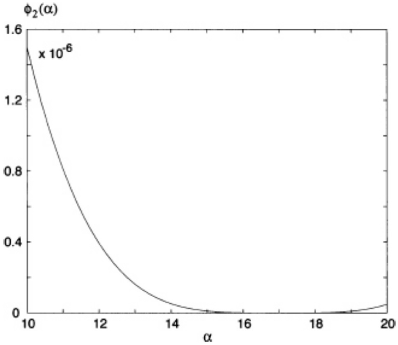
    $\nabla f(\mathbf{x}^{(2)}) = [0, 0, -0.003525]^T$

  - Step 2: Select step size α_1

    $$\begin{aligned}\alpha_2 &= \arg\min_{\alpha \geq 0}(0 + 0 + 4(-5.06 + 0.003525\alpha + 5)^4) \\ &= \arg\min_{\alpha \geq 0} \phi_2(\alpha)\end{aligned}$$

Method of Steepest Descent

- Using the secant method we obtain $\alpha_2 = 16.29$



- Step 3: Determine the next point $\mathbf{x}^{(2)}$

$$\mathbf{x}^{(3)} = \mathbf{x}^{(2)} - \alpha_1 \nabla f(\mathbf{x}^{(2)}) = [4, 3, -5.002]^T$$

Analysis of Optimization Algorithms

- Rely heavily on mathematical tools
- Analysis provides insight into:
  - Range of applicability of an algorithm
  - Appropriate choice of algorithm for a given problem
  - Qualitative behavior of an algorithm
- We must be able to answer:
  - Does the method work?
  - When does it work?
  - How well does it work?
- Not good enough to superficially use commercial optimization software package

Analysis of Optimization Algorithms

- Several characterizations of performance:
  - ***Globally convergent***: start from any initial point, the algorithm converges to a ”solution”
    - Usually, by ”solution” we mean a point satisfying the FONC
  - **Locally convergent**: starting from an initial point that is close enough to a solution, the algorithm converges to the solution
  - **Rate of convergence**: how fast an algorithm converges

Analysis of Gradient Methods

- We analyze gradient algorithms applied to quadratics only:

  $$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T Q\mathbf{x} - \mathbf{b}^T\mathbf{x}$$

  where $Q \in \mathbb{R}^{n\times n}$, $Q = Q^T > 0$, $\mathbf{b} \in \mathbb{R}^n$, and $\mathbf{x} \in \mathbb{R}^n$

- We restrict our attention to quadratics because:
  - Simplifies analysis
  - Local behavior near solution (Global convergence for quadratics tells us something about local convergence in more general functions)

Analysis of Gradient Methods

- Note that there is no loss of generality in assuming Q to be a symmetric matrix
- For example, for the quadratic form $\mathbf{x}^T A \mathbf{x}$, where $A \neq A^T$, we can transform it to a symmetric form

$$(\mathbf{x}^T A \mathbf{x})^T = \mathbf{x}^T A^T \mathbf{x} = \mathbf{x}^T A \mathbf{x}$$

- Accordingly,

$$\begin{aligned}\mathbf{x}^T A \mathbf{x} &= \tfrac{1}{2}\mathbf{x}^T A \mathbf{x} + \tfrac{1}{2}\mathbf{x}^T A^T \mathbf{x} \\ &= \tfrac{1}{2}\mathbf{x}^T (A + A^T)\mathbf{x} \\ &= \tfrac{1}{2}\mathbf{x}^T Q \mathbf{x}\end{aligned}$$

where $Q = A + A^T$

Steepest Descent Applied to Quadratics

- Consider $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T Q\mathbf{x} - \mathbf{b}^T\mathbf{x}$
- We have the gradient $\nabla f(\mathbf{x}) = Q\mathbf{x} - \mathbf{b}$ and the Hessian $F(\mathbf{x}) = Q$
- For simplicity, write $\mathbf{g}^{(k)} = \nabla f(\mathbf{x}^{(k)})$
- Let $\phi_k(\alpha) = f(\mathbf{x}^{(k)} - \alpha\mathbf{g}^{(k)})$ which is quadratic:

$$\phi_k(\alpha) = \frac{1}{2}(\mathbf{x}^{(k)} - \alpha\mathbf{g}^{(k)})^T Q(\mathbf{x}^{(k)} - \alpha\mathbf{g}^{(k)}) - (\mathbf{x}^{(k)} - \alpha\mathbf{g}^{(k)})^T\mathbf{b}$$

$$\phi_k(\alpha) = \left(\frac{1}{2}\mathbf{g}^{(k)T} Q\mathbf{g}^{(k)}\right)\alpha^2 - \left(\mathbf{g}^{(k)T}\mathbf{g}^{(k)}\right)\alpha + \text{C}$$

Steepest Descent Applied to Quadratics

- The steepest descent algorithm for the quadratic function can be represented as

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \alpha_k \mathbf{g}^{(k)}$$

where

$$\begin{aligned} \alpha_k &= \arg\min_{\alpha \geq 0} \phi(\alpha) = f(\mathbf{x}^{(k)} - \alpha \mathbf{g}^{(k)}) \\ &= \arg\min \left(\tfrac{1}{2}\mathbf{g}^{(k)T} Q \mathbf{g}^{(k)}\right) \alpha^2 - \left(\mathbf{g}^{(k)T}\mathbf{g}^{(k)}\right) \alpha + C \end{aligned}$$

Steepest Descent Applied to Quadratics

- In this quadratic case, we can find a closed-form solution for α_k
- We apply the FONC to $\phi_k(\alpha)$ to obtain

$$\phi_k'(\alpha) = \mathbf{g}^{(k)T} Q \mathbf{g}^{(k)} \alpha - (\mathbf{g}^{(k)T} \mathbf{g}^{(k)}) = 0$$

- Accordingly, we get

$$\alpha_k = \frac{\mathbf{g}^{(k)T} \mathbf{g}^{(k)}}{\mathbf{g}^{(k)T} Q \mathbf{g}^{(k)}}$$

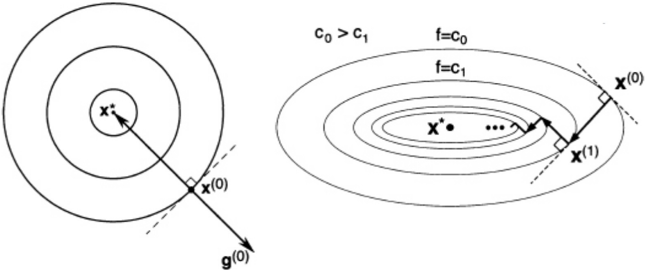
- The algorithm of steepest descent for the quadratic:

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \frac{\mathbf{g}^{(k)T} \mathbf{g}^{(k)}}{\mathbf{g}^{(k)T} Q \mathbf{g}^{(k)}} \mathbf{g}^{(k)}$$

Steepest Descent Applied to Quadratics

- **Example**: application of steepest descent for the following objective functions

$f(x_1, x_2) = x_1^2 + x_2^2$ and $f(x_1, x_2) = \frac{x_1^2}{5} + x_2^2$



- The impact of selecting different initial points

Convergence Analysis of Gradient Methods

- For gradient method, the convergence depends on the property of the objective function f and the choice of step size α
- We analyze the convergence of gradient methods for quadratics
  - For method of steepest descent
  - For gradient methods with fixed step size

Convergence Analysis of Gradient Methods

- **Theorem**: For quadratic functions, in the steepest descent algorithm, we have $\mathbf{x}^{(k)} \to \mathbf{x}^*$ for any $\mathbf{x}^{(0)}$

Convergence Analysis of Gradient Methods

- We investigate the convergence of the fixed step size algorithm

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \alpha \mathbf{g}^{(k)}$$

- This algorithm is of interest due to its simplicity which does not require a line search at each step
- The convergence depends on the choice of α
- **Theorem**: For the fixed step size gradient algorithm, $\mathbf{x}^{(k)} \to \mathbf{x}^*$ for any $\mathbf{x}^{(0)}$, if and only if

$$0 < \alpha < \frac{2}{\lambda_{max}(Q)}$$

Convergence Analysis of Gradient Methods

- **Example**: Let the function f be given by

$$f(\mathbf{x}) = \mathbf{x}^T \begin{bmatrix} 4 & 2\sqrt{2} \\ 0 & 5 \end{bmatrix} \mathbf{x} + \mathbf{x}^T \begin{bmatrix} 3 \\ 6 \end{bmatrix} + 24$$

We wish to find the minimizer of f using a fixed step size gradient algorithm

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

where α is a fixed step size. Check the condition for α to guarantee the convergence of the algorithm

Convergence Analysis of Gradient Methods

- We first symmetrize the matrix in the quadratic term of f to get

$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T \left[\begin{array}{cc} 8 & 2\sqrt{2} \\ 2\sqrt{2} & 10 \end{array} \right] \mathbf{x} + \mathbf{x}^T \left[\begin{array}{c} 3 \\ 6 \end{array} \right] + 24$$

- The eigenvalues of the matrix in the quadratic term are 6 and 12
- Accordingly, the algorithm converges to the minimizer for all $\mathbf{x}^{(0)}$ if and only if α lies in the range $0 < \alpha < 2/12$

Convergence Analysis of Gradient Methods

- Summary of convergence property of gradient methods
  - If the objective function f is convex, the step size is chosen via a line-search that satisfies the Wolfe condition, then the corresponding gradient method is globally convergent
  - For an objective function with quadratic form $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T Q\mathbf{x} - \mathbf{b}^T\mathbf{x}$, where Q is positive definite, the steepest descent method is globally convergent
  - For an objective function with quadratic form $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T Q\mathbf{x} - \mathbf{b}^T\mathbf{x}$, where Q is positive definite, the fixed step size descent method is globally convergent if $0 < \alpha < \frac{2}{\lambda_{max}(Q)}$

Analysis of Gradient Methods

- **Rate of convergence**: the speed at which a convergent sequence approaches its limit
  - Determines how fast the algorithm converges to a solution point
- The ***order of convergence*** of a sequence is a measure of its rate of convergence
  - The higher the order, the faster the rate of convergence

Analysis of Gradient Methods

- Given a sequence $\{\mathbf{x}^{(k)}\}$ that converges to $\mathbf{x}^*$. That is, $\lim_{k\to\infty} \|\mathbf{x}^{(k)} - \mathbf{x}^*\| = 0$, we say that the ***order of convergence*** is p, where $p \in \mathbb{R}$, if

  $$\lim_{k\to\infty} \frac{\|\mathbf{x}^{(k+1)} - \mathbf{x}^*\|}{\|\mathbf{x}^{(k)} - \mathbf{x}^*\|^p} = \mu \text{ where } \mu \in (0, \infty)$$

  If for all $p > 0$

  $$\lim_{k\to\infty} \frac{\|\mathbf{x}^{(k+1)} - \mathbf{x}^*\|}{\|\mathbf{x}^{(k)} - \mathbf{x}^*\|^p} = 0$$

  then we say the order of convergence is ∞
- Question: what if $\mu = 0$

Analysis of Gradient Methods

- If $p = 1$ (first-order convergence) and $\lim_{k\to\infty} \frac{\|\mathbf{x}^{(k+1)} - \mathbf{x}^*\|}{\|\mathbf{x}^{(k)} - \mathbf{x}^*\|^p} = 1$, the convergence is ***sublinear***
- If $p = 1$ and $\lim_{k\to\infty} \frac{\|\mathbf{x}^{(k+1)} - \mathbf{x}^*\|}{\|\mathbf{x}^{(k)} - \mathbf{x}^*\|^p} < 1$, the convergence is ***linear***
- If $p > 1$, the convergence is ***superlinear***
- If $p = 2$ (second-order convergence), the convergence is ***quadratic***
- If $p = 3$ (third-order convergence), the convergence is ***cubic***

Analysis of Gradient Methods

- ***Example***: Given $x^{(k)} = 1/k$ and thus $x^{(k)} \to 0$, analyze the order of convergence

Analysis of Gradient Methods

- ***Example***: Given $x^{(k)} = 1/k$ and thus $x^{(k)} \to 0$
- Then

$$\frac{|x^{(k+1)}|}{|x^{(k)}|^p} = \frac{1/(k+1)}{1/k^p} = \frac{k^p}{k+1}$$

- If $p > 1$, it grows to ∞
- If $p < 1$, it converges to 0
- If $p = 1$, it converges to 1
- Hence, the order of convergence is 1

Analysis of Gradient Methods

- Given $x^{(k)} = \gamma^k$ where $0 < \gamma < 1$, thus $x^{(k)} \to 0$
- Then

$$\frac{|x^{(k+1)}|}{|x^{(k)}|^p} = \frac{\gamma^{k+1}}{(\gamma^k)^p} = \gamma^{k+1-kp} = \gamma^{k(1-p)+1}$$

- If $p > 1$, it grows to ∞
- If $p < 1$, it converges to 0
- If $p = 1$, it converges to γ
- Hence, the order of convergence is 1

Analysis of Gradient Methods

- ***Example***: Given $x^{(k)} = \gamma^{q^k}$ where $q > 1$, $0 < \gamma < 1$, thus $x^{(k)} \to 0$
- Then
$$\frac{|x^{(k+1)}|}{|x^{(k)}|^p} = \frac{\gamma^{q^{k+1}}}{(\gamma^{(q^k)})^p} = \gamma^{q^{k+1}-pq^k} = \gamma^{(q-p)q^k}$$
- If $p < q$, it converges to 0
- If $p > q$, it grows to ∞
- If $p = q$, it converges to 1
- Hence, the order of convergence is q

Analysis of Gradient Methods

- ***Example***: Given $x^{(k)} = 1$ for all k, thus $x^{(k)} \to 1$
- Then
$$\frac{|x^{(k+1)} - 1|}{|x^{(k)} - 1|^p} = \frac{0}{0^p} = 0$$
for all p
- Hence, the order of convergence is ∞

Analysis of Gradient Methods

- The order of convergence can be interpreted using the notion of the order symbol O
- $g(h) = O(h)$ means that there exists a constant c such that $|g(h)| \leq c|h|$ for sufficiently small h
- The order of convergence is ***at least*** p if

$$\|\mathbf{x}^{(k+1)} - \mathbf{x}^*\| = O(\|\mathbf{x}^{(k)} - \mathbf{x}^*\|^p)$$

- $g(h) = \Omega(h)$ means that there exists a constant c such that $|g(h)| \geq c|h|$ for sufficiently small h
- The order of convergence is ***at most*** p if

$$\|\mathbf{x}^{(k+1)} - \mathbf{x}^*\| = \Omega(\|\mathbf{x}^{(k)} - \mathbf{x}^*\|^p)$$

Analysis of Gradient Methods

- ***Example***: Suppose we are given a scalar sequence $\{x^{(k)}\}$ that converges with order of convergence p and satisfies

  $$\lim_{k\to\infty} \frac{|x^{k+1}-2|}{|x^k-2|^3} = 0$$

- The limit of $\{x^{(k)}\}$ is 2
- $|x^{(k+1)} - 2| = O|x^{(k)} - 2|^3$
- Hence, we conclude that $p \geq 3$

Analysis of Gradient Methods

- ***Example***: Consider the problem of finding a minimizer of the function $f : \mathbb{R} \to \mathbb{R}$ given by
  $$f(x) = x^2 - \frac{x^3}{3}$$
  Suppose that we use the algorithm $x^{(k+1)} = x^{(k)} - \alpha f'(x^{(k)})$ with step size $\alpha = 1/2$ and initial condition $x^{(0)} = 1$
- We first show the algorithm converges to a local minimizer of f
  - $f'(x) = 2x - x^2$
  - $x^{(k+1)} = x^{(k)} - \alpha f'(x^{(k)}) = \frac{1}{2}(x^{(k)})^2$
  - with $x^{(0)} = 1$, we can have $x^{(k)} = (1/2)^{2k-1}$
  - The algorithm converges to 0
- Next, we find the order of convergence. Note that $\frac{|x^{(k+1)}|}{|x^{(k)}|^2} = \frac{1}{2}$. Therefore, the order of convergence is 2

Analysis of Gradient Methods

- **Theorem**: The steepest descent algorithm has order of convergence of 1 in the worst case

Analysis of Gradient Methods

- Summary of Gradient method
- Pros:
  - Basis of many iterative algorithms
  - Simple and reliable
- Cons:
  - Slow convergence (most gradient descent methods possess the worst-case linear convergence property)
  - Convergence rate depends critically on the condition number
  - Zigzaging when approaching the minimizer
  - Need to find the optimal step-size