PRO 5970 Métodos de Otimização Não Linear

Celma de Oliveira Ribeiro

2023

Departmento de Engenharia de Produção
Universidade de São Paulo

Unconstrained Optimization

The problem of interest

Given $f : \mathbb{R}^n \to \mathbb{R}$ find

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

Fundamentals

- Recognizing a local minimum
- Gradient-based algorithms
  - Line search methods
  - Trust region
- Derivative free algorithms

Unconstrained Optimization

The problem of interest

Given $f : \mathbb{R}^n \to \mathbb{R}$ find

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

How to recognize a local minimum?

Necessary conditions and sufficient conditions

Unconstrained Optimization

What are necessary and sufficient conditions for a local minimum?

- Necessary conditions: Conditions satisfied by every local minimum
- Sufficient conditions: Conditions which guarantee a local minimum

Easy to characterize a local minimum if f is sufficiently smooth

Unconstrained Optimization - using derivatives

A naive example

Consider $f : \mathbb{R} \to \mathbb{R}$, $f(x) = (x-2)^2$

- Find a stationary point $f^{'}(x^*) = 0$
- Identify if it is a minimum

Finding the real roots of $g(x) = f^{'}(x)$ may be difficult. Examples: find the minimum of $f(x) = x^2 - e^x$

Unconstrained Optimization - using derivatives

First order necessary condition

If x^* is a local minimizer and f is continuously differentiable in a open neighborhood of x^*, then $\nabla f(x^*) = 0$

Unconstrained Optimization - using derivatives

First order necessary condition

If x^* is a local minimizer and f is continuously differentiable in a open neighborhood of x^*, then $\nabla f(x^*) = 0$

Second order necessary condition

If x^* is a local minimizer and $\nabla^2 f$ exists and is continuously in a open neighborhood of x^*, then $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*)$ is positive semidefinite.

Unconstrained Optimization - using derivatives

First order necessary condition

If x^* is a local minimizer and f is continuously differentiable in a open neighborhood of x^*, then $\nabla f(x^*) = 0$

Second order necessary condition

If x^* is a local minimizer and $\nabla^2 f$ exists and is continuously in a open neighborhood of x^*, then $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*)$ is positive semidefinite.

Second order sufficient condition

Suppose that $\nabla^2 f(x^*)$ is continuous in an open neighborhood of x^* If the following two conditions are satisfied, then x^* is a strict local minimizer of f:

- $\nabla f(x^*) = 0$
- $\nabla^2 f(x^*)$ is positive semidefinite.

Algorithms for unconstrained optimization

How do we compute the optimal (local or global) solutions?

- Analytically: only possible for some simple problems (e.g. univariate minimization)
- Numerically: required for most engineering optimization problems (too large and complex to solve analytically)

Numerical optimization algorithms used to solve these problems

Algorithms for unconstrained optimization

Goals

- Robust: low failure rate, convergence conditions are met
- Fast: convergence in a few iterations and low cost per iteration
- Feasible: reasonable memory requirements

Algorithm design involves tradeoffs to achieve these goals (e.g. using high-order information may lead to fewer iterations, but each iteration becomes more expensive)

Algorithms are iterative in nature

Categorization

- Gradient-based v. Derivative-free
- Global v. local

Gradient-based algorithms tend to be local, while derivative-free algorithms tend to be global

Characteristics

- Commom feature: they start from an initial guess $x_0 \in \mathbb{R}^n$ and generate a sequence of iterates $\{x_k\}_k$ such that $x_k \to x^*$, $x^* \in \arg\min_{x \in \mathbb{R}^n} f(x)$

Characteristics

- Commom feature: they start from an initial guess $x_0 \in \mathbb{R}^n$ and generate a sequence of iterates $\{x_k\}_k$ such that $x_k \to x^*$, $x^* \in \arg\min_{x \in \mathbb{R}^n} f(x)$
- Usually make progress towards solving the problem in every iteration, that is, $f(x_{k+1}) < f(x_k)$ (descent methods).

Characteristics

- Commom feature: they start from an initial guess $x_0 \in \mathbb{R}^n$ and generate a sequence of iterates $\{x_k\}_k$ such that $x_k \to x^*$, $x^* \in \arg\min_{x \in \mathbb{R}^n} f(x)$
- Usually make progress towards solving the problem in every iteration, that is, $f(x_{k+1}) < f(x_k)$ (descent methods).
- In practice x^* cannot be obtained precisely. The process stops when x_k is sufficiently close to x^* .

Algorithms for unconstrained optimizationn

Characteristics

- Commom feature: they start from an initial guess $x_0 \in \mathbb{R}^n$ and generate a sequence of iterates $\{x_k\}_k$ such that $x_k \to x^*$, $x^* \in \arg\min_{x \in \mathbb{R}^n} f(x)$
- Usually make progress towards solving the problem in every iteration, that is, $f(x_{k+1}) < f(x_k)$ (descent methods).
- In practice x^* cannot be obtained precisely. The process stops when x_k is sufficiently close to x^* .
- Optimality conditions can serve as a stopping criterion when they are satisfied to within a predetermined error tolerance.

Characteristics

- Commom feature: they start from an initial guess $x_0 \in \mathbb{R}^n$ and generate a sequence of iterates $\{x_k\}_k$ such that $x_k \to x^*$, $x^* \in \arg\min_{x \in \mathbb{R}^n} f(x)$
- Usually make progress towards solving the problem in every iteration, that is, $f(x_{k+1}) < f(x_k)$ (descent methods).
- In practice x^* cannot be obtained precisely. The process stops when x_k is sufficiently close to x^* .
- Optimality conditions can serve as a stopping criterion when they are satisfied to within a predetermined error tolerance.
- It is important that $\{x_k\}_k$ converges to x^* at a rapid rate

faster convergence $\leftrightarrow$ higher computational cost per iteration.

Line search

The algorithm chooses a direction p_k and searches along this direction from the current point x_k for a new iterate with a lower function value

Algorithms for unconstrained optimization

Line search

The algorithm chooses a direction p_k and searches along this direction from the current point x_k for a new iterate with a lower function value

Trust region

The information gathered about f is used to construct a model function m_k whose behavior near the current point x_k is similar to that of the actual objective function f.

As m_k may not be a good approximation of f far from x_k, we restrict the search for a minimizer of m_k to some region around x_k

$$min_p m_k (x_k + p)$$

where $x_k + p$ lies inside the trust region

Descent Directions

Definition Let $\hat{x} \in \mathbb{R}^n$ and $f : \mathbb{R}^n \to \mathbb{R}$. The vetor $p \in \mathbb{R}^n$is called a **descent direction** for f at $\hat{x}$ if $\exists \bar{\alpha} \in \mathbb{R}_{++}$ such that

$$0 < \alpha \leq \bar{\alpha} \Rightarrow f(\hat{x} + \alpha p) < f(\hat{x})$$

If there is a descent direction $\hat{x}$ is not a local minimizer.

Descent Directions

Let $f : \mathbb{R}^n \to \mathbb{R}$ be partially differentiable with continuous partial derivatives and let $\hat{x} \in \mathbb{R}^n$, $p \in \mathbb{R}^n$. Suppose $p^{'} \nabla f(\hat{x}) < 0$. Then p is a descent direction for f at $\hat{x}$

Obvious choice: $p = -\nabla f(\hat{x})$

Usually considers normalized direction $\frac{-\nabla f(\hat{x})}{||\nabla f(\hat{x})||}$

Examples

(1) Consider $f(x) = (x_1 - 1)^2 + (x_2 - 3)^2 \ \forall x \in \mathbb{R}^2$

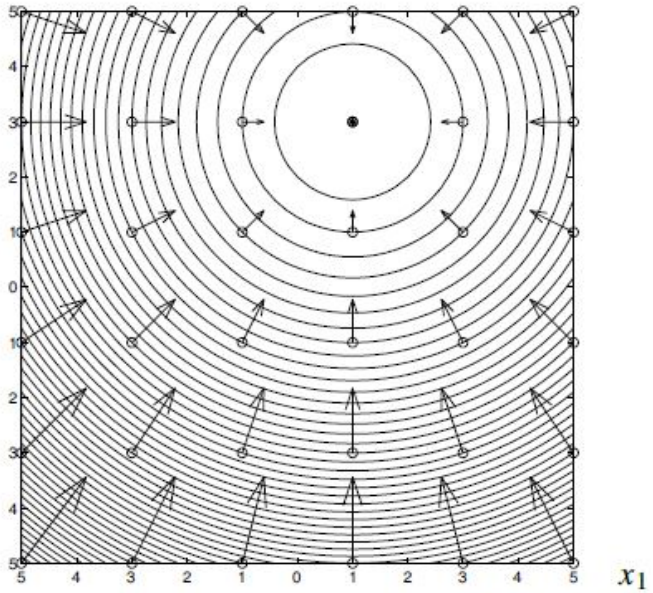
Show that $\begin{bmatrix} 1 \\ 3 \end{bmatrix}$ is a minimizer (use information about the Hessian). Can you find a descent direction at this point?

(2) Consider $f(x) = (x_1 - 1)^2 + (x_2 - 3)^2 - 1.8\,(x_1 - 1)\,(x_2 - 3)$

- Find $\nabla f(x)$ and $\nabla^2 f(x)$
- Write $f(x)$ as $f(x) = \frac{1}{2} x^t Q x + c^t x$ (Hint: $c = \begin{bmatrix} 3.4 \\ -4.2 \end{bmatrix}$)
- Beginning with $x^{(0)} = \begin{bmatrix} 3 \\ -5 \end{bmatrix}$ find the next step of the steepest descent method $\left(x^{(1)}\right)$

Example

Level curves ad descent directions of $f(x) = (x_1 - 1)^2 + (x_2 - 3)^2 \ \forall x \in \mathbb{R}^2$



Point $\begin{bmatrix} 1 \\ 3 \end{bmatrix}$ is indicated as $\bullet$

Generic Line Search Method

Given x_0, set k := 0.

Until x_k has converged,

1 Calculate a search direction p_k from x_k, ensuring that this is a descent direction

2 Calculate a suitable step-length $\lambda_k > 0$ so that

$$f(x_k + \lambda_k d_k) < f(x_k)$$

The computation of λ_k is called line search, (usually an inner iterative loop).

3 Set $x_{k+1} \leftarrow x_k + \lambda_k d_k$

4 k $\leftarrow$ k + 1.

Go to Step 1.

Methods differ according to steps (1) and (2) .

Steepest descent algorithm

Given x_0, set k := 0.

1. $d_k = -\nabla f(x_k)$. If $\|d_k\| \leq \epsilon$,then stop.
2. Solve $min_\lambda f(x_k + \lambda d_k)$, obtaining the step-length λ_k, perhaps chosen by an exact or inexact line-search.
3. Set $x_{k+1} \leftarrow x_k + \lambda_k d_k$
4. k $\leftarrow$ k + 1.

Go to Step 1.

Example - steepest descent

$$\nabla f(x^{(0)}) = \begin{bmatrix} 18.4 \\ -19.6 \end{bmatrix}$$

Example - steepest descent

$$\nabla f(x^{(0)}) = \begin{bmatrix} 18.4 \\ -19.6 \end{bmatrix}$$

$$x^{(1)} = x^{(0)} + \alpha^{(0)} \nabla f(x^{(0)}) = \begin{bmatrix} 3 \\ -5 \end{bmatrix} + \alpha^{(0)} \begin{bmatrix} 18.4 \\ -19.6 \end{bmatrix}$$

Example - steepest descent

$$\nabla f(x^{(0)}) = \begin{bmatrix} 18.4 \\ -19.6 \end{bmatrix}$$

$$x^{(1)} = x^{(0)} + \alpha^{(0)} \nabla f(x^{(0)}) = \begin{bmatrix} 3 \\ -5 \end{bmatrix} + \alpha^{(0)} \begin{bmatrix} 18.4 \\ -19.6 \end{bmatrix}$$

Obtain $\alpha^{(0)}$ through the minimization of $f\left(\alpha^{(0)}\right)$, $x^{(1)} \approx \begin{bmatrix} -1.8467 \\ 0.1628 \end{bmatrix}$

Example - steepest descent

$$\nabla f(x^{(0)}) = \begin{bmatrix} 18.4 \\ -19.6 \end{bmatrix}$$

$$x^{(1)} = x^{(0)} + \alpha^{(0)} \nabla f(x^{(0)}) = \begin{bmatrix} 3 \\ -5 \end{bmatrix} + \alpha^{(0)} \begin{bmatrix} 18.4 \\ -19.6 \end{bmatrix}$$

Obtain $\alpha^{(0)}$ through the minimization of $f\left(\alpha^{(0)}\right)$, $x^{(1)} \approx \begin{bmatrix} -1.8467 \\ 0.1628 \end{bmatrix}$

Far from optimal solution! After two iterations you get close to the optimal solution

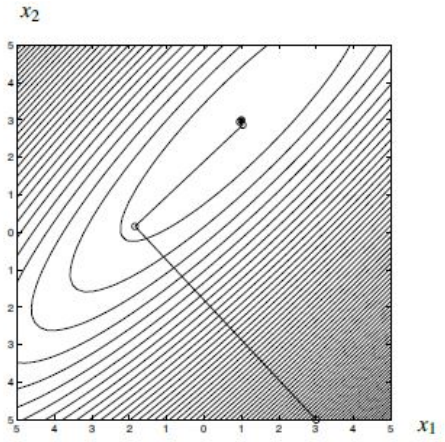
Example - steepest descent

$$\nabla f(x^{(0)}) = \begin{bmatrix} 18.4 \\ -19.6 \end{bmatrix}$$

$$x^{(1)} = x^{(0)} + \alpha^{(0)} \nabla f(x^{(0)}) = \begin{bmatrix} 3 \\ -5 \end{bmatrix} + \alpha^{(0)} \begin{bmatrix} 18.4 \\ -19.6 \end{bmatrix}$$

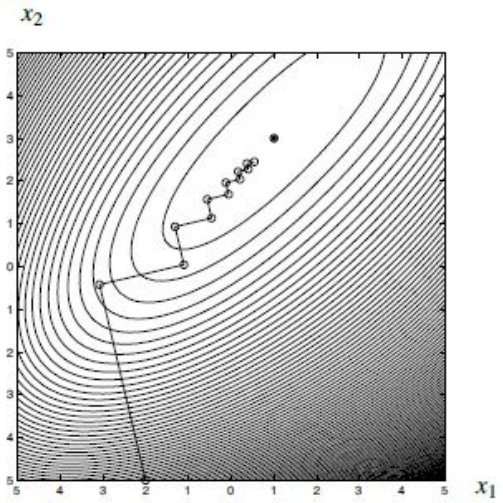
Obtain $\alpha^{(0)}$ through the minimization of $f\left(\alpha^{(0)}\right)$, $x^{(1)} \approx \begin{bmatrix} -1.8467 \\ 0.1628 \end{bmatrix}$

Far from optimal solution! After two iterations you get close to the optimal solution



Example - Steepest descent

Starting point $x^{(0)} = \begin{bmatrix} -2 \\ -5 \end{bmatrix}$



The steepest descent

The convergence depends on the objective function:

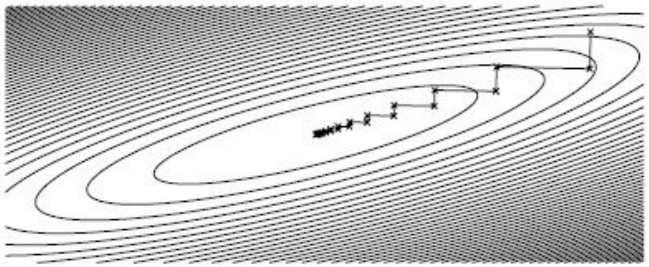


Figure 1: Slow convergence

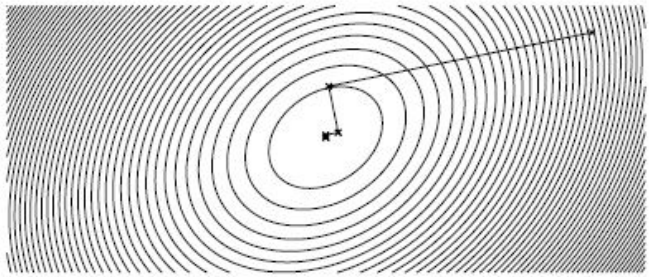


Figure 2: Fast convergence

Bissection

Consider $f(x) : \mathbb{R} \to \mathbb{R}$ convex and differentiable.

Let $\left[a^1, b^1\right]$ be an uncertainty interval

Let l be a given precision level, and n the lowest integer such that $\left(\frac{1}{2}\right)^n \leq \frac{l}{b^1 - a^1}$

$k \leftarrow 1$

1. Consider $\left[a^k, b^k\right]$ Calcule $\lambda^k = \frac{a^k + b^k}{2}$ e $f'\left(\lambda^k\right)$
2. If $f'\left(\lambda^k\right) = 0$,

Bissection

Consider $f(x) : \mathbb{R} \to \mathbb{R}$ convex and differentiable.

Let $\left[a^1, b^1\right]$ be an uncertainty interval

Let l be a given precision level, and n the lowest integer such that $\left(\frac{1}{2}\right)^n \leq \frac{l}{b^1 - a^1}$

$k \leftarrow 1$

1. Consider $\left[a^k, b^k\right]$ Calcule $\lambda^k = \frac{a^k + b^k}{2}$ e $f'\left(\lambda^k\right)$
2. If $f'\left(\lambda^k\right) = 0$, stop
3. If $f'\left(\lambda^k\right) > 0$,

Bissection

Consider $f(x) : \mathbb{R} \to \mathbb{R}$ convex and differentiable.

Let $\left[a^1, b^1\right]$ be an uncertainty interval

Let l be a given precision level, and n the lowest integer such that $\left(\frac{1}{2}\right)^n \leq \frac{l}{b^1 - a^1}$

$k \leftarrow 1$

1. Consider $\left[a^k, b^k\right]$ Calcule $\lambda^k = \frac{a^k + b^k}{2}$ e $f'\left(\lambda^k\right)$
2. If $f'\left(\lambda^k\right) = 0$, stop
3. If $f'\left(\lambda^k\right) > 0$,$a^{k+1} \leftarrow a^k$ e $b^{k+1} \leftarrow \lambda^k$ Go to step 5
4. If $f'\left(\lambda^k\right) < 0$,

Bissection

Consider $f(x) : \mathbb{R} \to \mathbb{R}$ convex and differentiable.

Let $\left[a^1, b^1\right]$ be an uncertainty interval

Let l be a given precision level, and n the lowest integer such that $\left(\frac{1}{2}\right)^n \leq \frac{l}{b^1 - a^1}$

$k \leftarrow 1$

1. Consider $\left[a^k, b^k\right]$ Calcule $\lambda^k = \frac{a^k + b^k}{2}$ e $f'\left(\lambda^k\right)$
2. If $f'\left(\lambda^k\right) = 0$, stop
3. If $f'\left(\lambda^k\right) > 0$,$a^{k+1} \leftarrow a^k$ e $b^{k+1} \leftarrow \lambda^k$ Go to step 5
4. If $f'\left(\lambda^k\right) < 0$,$a^{k+1} \leftarrow \lambda^k$ e $b^{k+1} \leftarrow b^k$ Go to step 5
5. If $k = n$ stop, the minimum in in $\left[a^k, b^k\right]$. Otherwise $k \leftarrow k + 1$ Go to step 1

Bissection

Example

Min $f(\lambda) = \lambda^2 + 2\lambda$ on $[-3, 6]$, com $l = 0.2$

Bissection

Example

Min $f(\lambda) = \lambda^2 + 2\lambda$ on $[-3, 6]$, com $l = 0.2$

$n = 6$

k	a_k	b_k	λ_k	$f'(\lambda_k)$

Bissection

Example

Min $f(\lambda) = \lambda^2 + 2\lambda$ on $[-3, 6]$, com $l = 0.2$

$n = 6$

k	a_k	b_k	λ_k	$f'(\lambda_k)$
1	-3.0000	6.0000	1.5000	5.0000

Bissection

Example

Min $f(\lambda) = \lambda^2 + 2\lambda$ on $[-3, 6]$, com $l = 0.2$

$n = 6$

k	a_k	b_k	λ_k	$f'(\lambda_k)$
1	-3.0000	6.0000	1.5000	5.0000
2	-3.0000	1.5000	-0.7500	0.5000

Bissection

Example

Min $f(\lambda) = \lambda^2 + 2\lambda$ on $[-3, 6]$, com $l = 0.2$

$n = 6$

k	a_k	b_k	λ_k	$f'(\lambda_k)$
1	-3.0000	6.0000	1.5000	5.0000
2	-3.0000	1.5000	-0.7500	0.5000
3	-3.0000	-0.7500	-1.8750	-1.7500

Bissection

Example

Min $f(\lambda) = \lambda^2 + 2\lambda$ on $[-3, 6]$, com $l = 0.2$

$n = 6$

k	a_k	b_k	λ_k	$f'(\lambda_k)$
1	-3.0000	6.0000	1.5000	5.0000
2	-3.0000	1.5000	-0.7500	0.5000
3	-3.0000	-0.7500	-1.8750	-1.7500
4	-1.8750	-0.7500	-1.3125	-0.6250
5	-1.3125	-0.7500	-1.0313	-0.0625
6	-1.0313	-0.7500	-0.8907	0.2186
7	-1.0313	-0.8907		

Solution -0.961

Unimodal functions

Let $\phi : \mathbb{R} \to \mathbb{R}$ and $x^* \in [a, b]$ a minimum of ϕ,

The function ϕ is said to be unimodal on $[a, b]$ if for

$$a \leq x_1 < x_2 \leq b$$

- $x_2 < x^* \Rightarrow \phi(x_1) > \phi(x_2)$
- $x_1 > x^* \Rightarrow \phi(x_2) > \phi(x_1)$

Dichotomous Search Algorithm

- Inicialization step
  Choose $\epsilon > 0$, $l > 0$ and an initial interval of uncertainty, $[a_1, b_1]$
- while $(b_k - a_k) > l$
  - $\lambda_k = \frac{a_k + b_k}{2} - \epsilon \qquad \mu_k = \frac{a_k + b_k}{2} + \epsilon$
  - if $\phi(\lambda_k) < \phi(\mu_k)$
  - then
  - $\quad a_{k+1} = a_k$ and $b_{k+1} = \mu_k$
  - else
  - $\quad a_{k+1} = \lambda_k$ and $b_{k+1} = b_k$
  - endif
- $k = k + 1$;
- end while
- Output: $x^* = \frac{a_k + b_k}{2}$

Unconstrained Optimization

Example

Find the minimum of $f(x) = \frac{1}{4}x^4 - \frac{5}{3}x^3 - 6x^2 + 19x - 7$

(unimodular in $[-4, 4]$)

k	a_k	b_k	$b_k - a_k$
0	-4	0	4
1	-4	-1.98	2.02
2	-3.0001	-1.98	1.0201
3	-3.0001	-2.4849	0.5152
10	-2.5669	-2.5626	0.0043
20	-2.5652	-2.5652	$4.65e^{-6}$
23	-2.5652	-2.5652	$5.99e^{-7}$

$$x^* = -2.5652 \ , \ f(x^*) = -56.2626$$

Taylor's theorem

Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable and that $p \in \mathbb{R}^n$. Then we have

$$f(x + p) = f(x) + \nabla f(x + tp)^T p$$

for com $t \in (0, 1)$

Theorem Convergence

Assume f twice continuously differentiable on an open interval (a,b) and that there exists $x^* \in (a, b)$ with $f^{'}(x^*) \neq 0$ Define Newton's method by the sequence

$$x_{k+1} = x_k - \frac{f(x_k)}{f^{'}(x_k)}, k = 1, 2, 3, ...$$

Assume also that x_k converges to x^* as $k \to \infty$. Then, for k sufficiently large, $\|x_{k+1} - x^*\| \leq M\|x_k - x^*\|^2$ if $M > \frac{f^{''}(x^*)}{2f^{'}(x^*)}$

Thus, x_k converges to x^* quadratically .

Line search with derivarives - Newton's Method

The method is based on the quadratic approximation of the function ϕ at a given point λ_k

$$\phi(\lambda) \approx q(\lambda) = \underbrace{\phi(\lambda_k) + \phi^{'}(\lambda_k)(\lambda - \lambda_k) + \frac{1}{2}\phi^{''}(\lambda_k)(\lambda - \lambda_k)^2}_{\text{derivative equals to zero}}$$

Main step

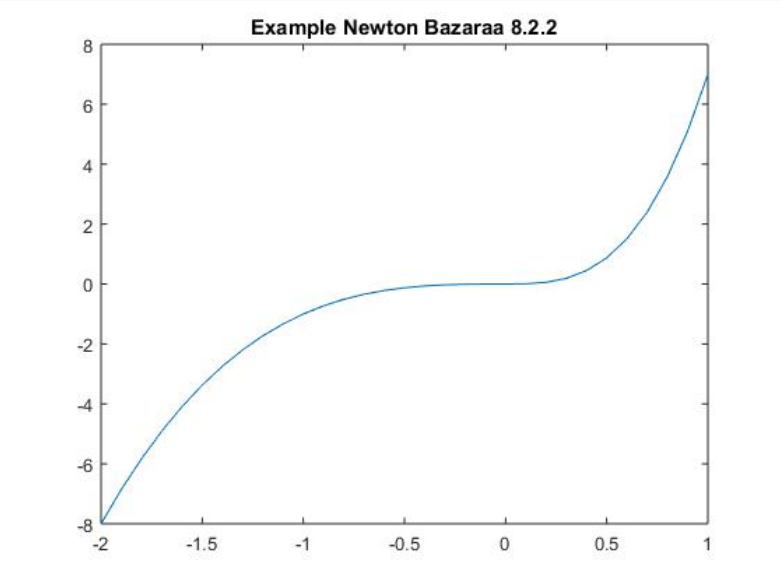
$$\lambda_{k+1} = \lambda_k - \frac{\phi^{'}(\lambda_k)}{\phi^{''}(\lambda_k)}$$

Unconstrained Optimization

- Newton's method presents quadratic convergence when the initial point is close to the optimal solution.
- Numerical difficulties occur when the second derivative is close to zero.
- If a poor starting point is chosen the method may fail to converge or diverge.

Example 1

$$\phi(\lambda) = \begin{cases} 4\lambda^3 - 3\lambda^4 & \text{if} \lambda \geq 0 \\ 4\lambda^3 + 3\lambda^4 & \text{if} \lambda < 0 \end{cases}$$



Consider

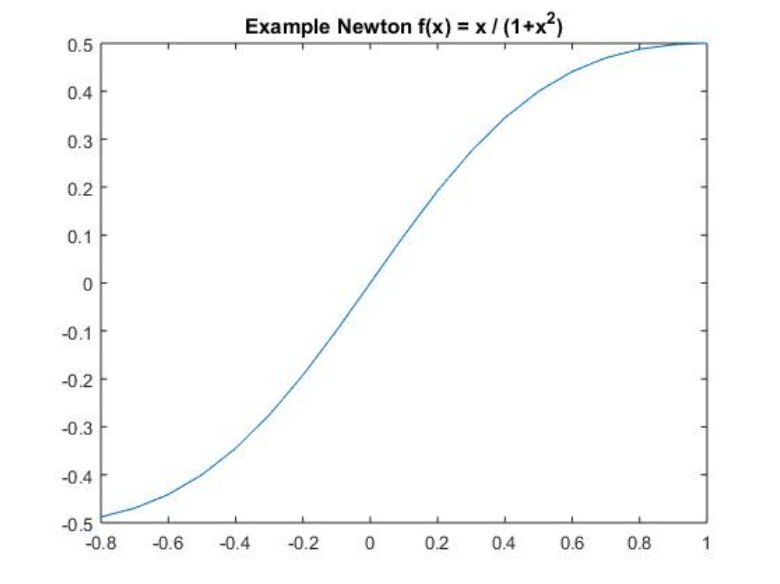
a) $\lambda_0 = 0.40$

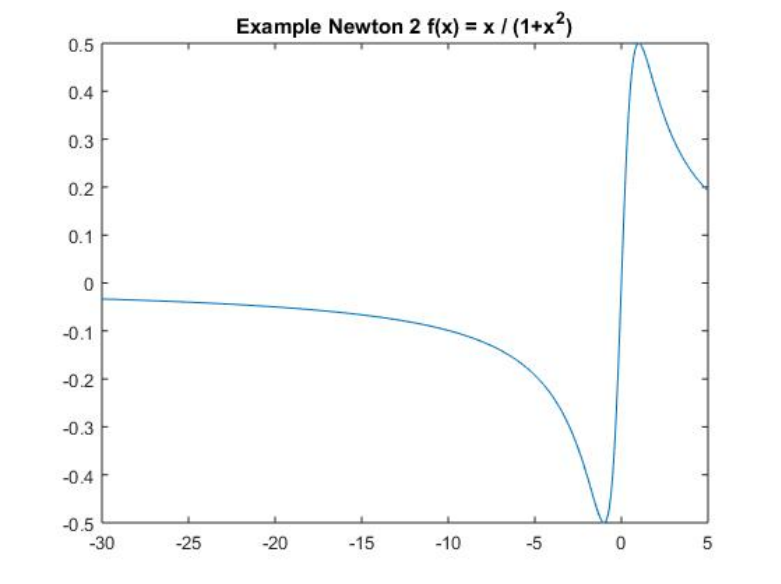
b) $\lambda_0 = 0.60$

Example 2

Find a zero of $f(x) = \frac{x}{1+x^2}$ In this case the sequence is given as

$$\lambda_{k+1} = \lambda_k - \frac{f(\lambda_k)}{f'(\lambda_k)}$$





Consider

a) $x_0 = 0.5$ and $\epsilon = 10^{-4}$

b) $x_0 = 0.75$ and $\epsilon = 10^{-4}$