Iterative Algorithms in Optimization, Variational Analysis and Fixed Point Theory

Unit 01: Introduction. Fundamentals of iterative algorithms.

university of groningen

About the course

Description

This course provides the theoretical bases for the numerical approximation of solutions to optimization problems, variational inequalities and fixed point problems by means of iterative algorithms.

Assessment

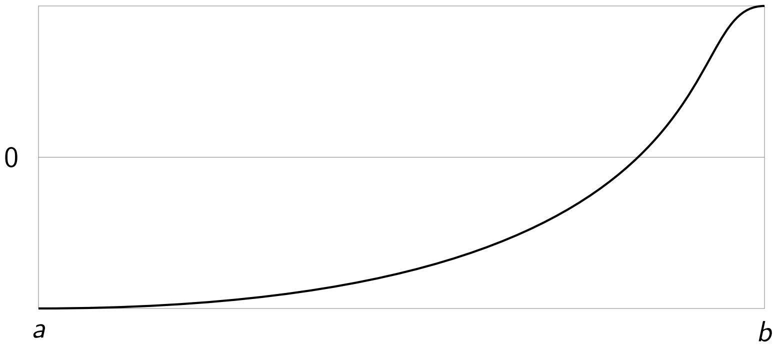
Computational exercises 20%, Homework assignments 50%, Exam 30%.

Dynamics and Schedule

16 sessions: Lectures, exercises. Mo, 9-11. Tu, 9-11. We, 17-19.

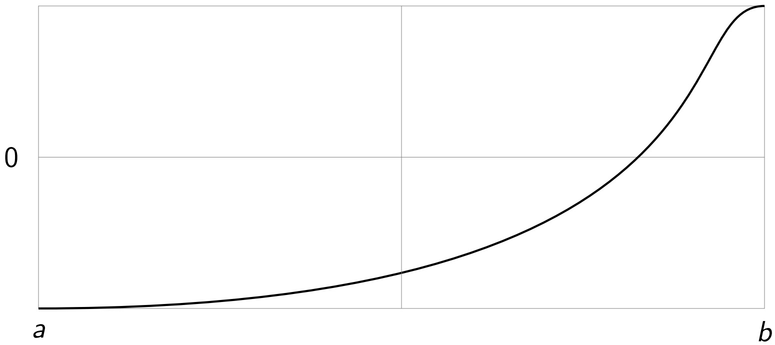
Introduction to iterative algorithms

Example: Bisection method to solve $g(x) = 0$



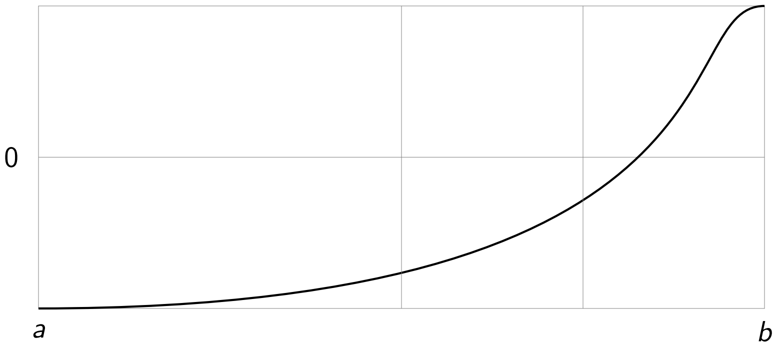
Introduction to iterative algorithms

Example: Bisection method to solve $g(x) = 0$



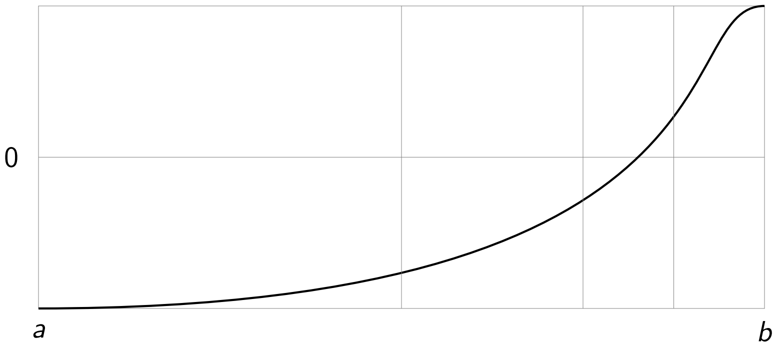
Introduction to iterative algorithms

Example: Bisection method to solve $g(x) = 0$



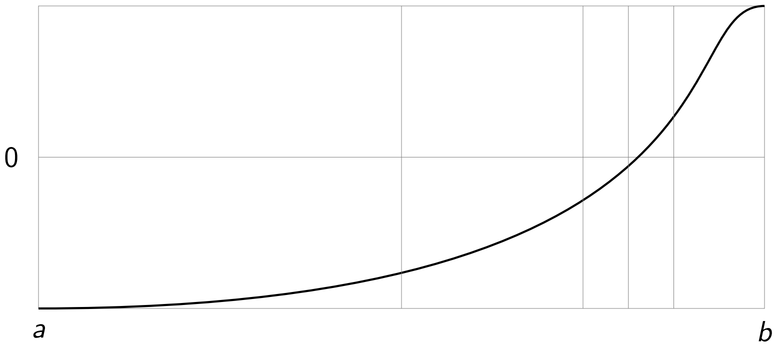
Introduction to iterative algorithms

Example: Bisection method to solve $g(x) = 0$



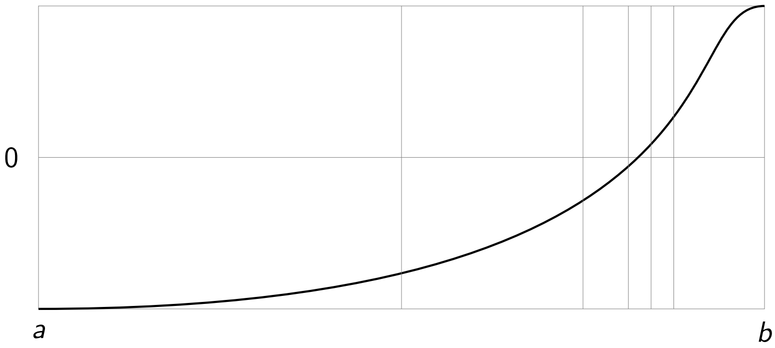
Introduction to iterative algorithms

Example: Bisection method to solve $g(x) = 0$



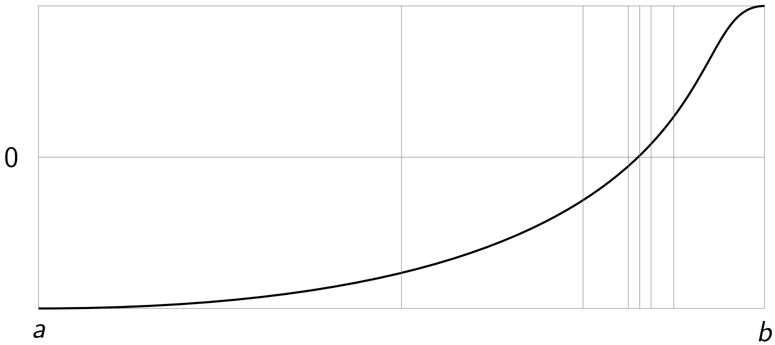
Introduction to iterative algorithms

Example: Bisection method to solve $g(x) = 0$



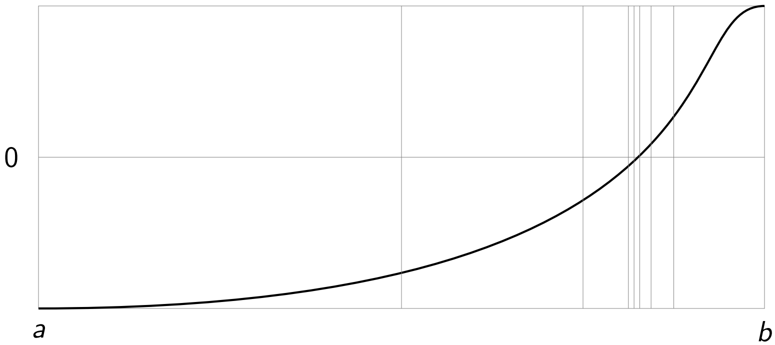
Introduction to iterative algorithms

Example: Bisection method to solve $g(x) = 0$



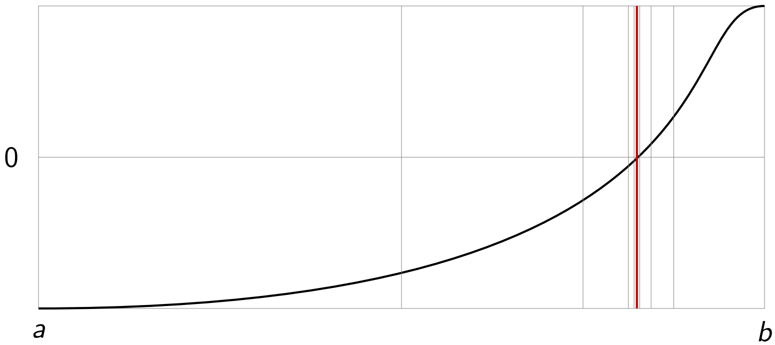
Introduction to iterative algorithms

Example: Bisection method to solve $g(x) = 0$



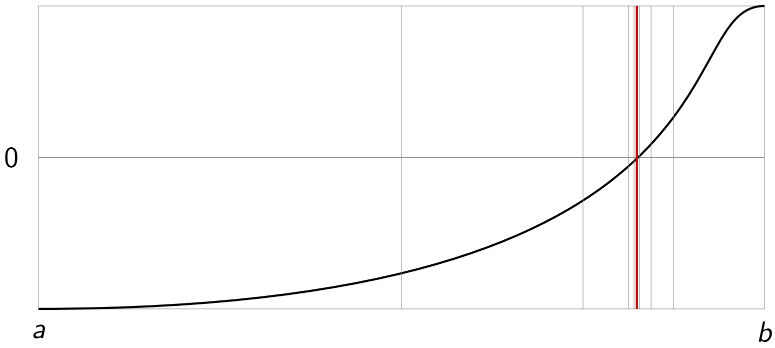
Introduction to iterative algorithms

Example: Bisection method to solve $g(x) = 0$



Introduction to iterative algorithms

Example: Bisection method to solve $g(x) = 0$



After k iterations, the distance to a solution is $|x_k - \hat{x}| \leq \dfrac{b-a}{2^k}$.

Problems to be solved

We shall approximate points in a set S of solutions in different contexts:

- $S = \text{Argmin}(f)$. Minimizers of an extended real-valued function $f : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$, especially in the convex case.

Problems to be solved

We shall approximate points in a set S of solutions in different contexts:

- $S = \mathrm{Argmin}(f)$. Minimizers of an extended real-valued function $f : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$, especially in the convex case.

- $S = \mathrm{Zer}(A)$. Zeroes of a set-valued operator $A : \mathbb{R}^N \to 2^{\mathbb{R}^N}$, especially in the monotone case.

Problems to be solved

We shall approximate points in a set S of solutions in different contexts:

- $S = \mathrm{Argmin}(f)$. Minimizers of an extended real-valued function $f : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$, especially in the convex case.

- $S = \mathrm{Zer}(A)$. Zeroes of a set-valued operator $A : \mathbb{R}^N \to 2^{\mathbb{R}^N}$, especially in the monotone case.

- $S = \mathrm{Fix}(T)$. Fixed points of a function $T : \mathbb{R}^N \to \mathbb{R}^N$, especially in the nonexpansive case.

Quality of approximate solutions

A merit function $\Phi : \mathbb{R}^N \to \mathbb{R}_+$ for a problem with solution set S satisfies:

- $\Phi(z) = 0$ if $z \in S$, and $\Phi(z) > 0$ otherwise.
- If $\Phi(z) < \Phi(w)$, then z is a better approximate solution than w.

Quality of approximate solutions

A merit function $\Phi : \mathbb{R}^N \to \mathbb{R}_+$ for a problem with solution set S satisfies:

- $\Phi(z) = 0$ if $z \in S$, and $\Phi(z) > 0$ otherwise.
- If $\Phi(z) < \Phi(w)$, then z is a better approximate solution than w.

Examples

- $\Phi(z) = f(z) - \inf(f)$ for $S = \mathrm{Argmin}(f)$.

Quality of approximate solutions

A merit function $\Phi : \mathbb{R}^N \to \mathbb{R}_+$ for a problem with solution set S satisfies:

- $\Phi(z) = 0$ if $z \in S$, and $\Phi(z) > 0$ otherwise.
- If $\Phi(z) < \Phi(w)$, then z is a better approximate solution than w.

Examples

- $\Phi(z) = f(z) - \inf(f)$ for $S = \operatorname{Argmin}(f)$.

  Also $\Phi(z) = \|\nabla f(z)\|$, if f is convex and differentiable.

Quality of approximate solutions

A merit function $\Phi : \mathbb{R}^N \to \mathbb{R}_+$ for a problem with solution set S satisfies:

- $\Phi(z) = 0$ if $z \in S$, and $\Phi(z) > 0$ otherwise.
- If $\Phi(z) < \Phi(w)$, then z is a better approximate solution than w.

Examples

- $\Phi(z) = f(z) - \inf(f)$ for $S = \operatorname{Argmin}(f)$.
  Also $\Phi(z) = \|\nabla f(z)\|$, if f is convex and differentiable.
- $\Phi(z) = \inf\{\|v\| : v \in Az\}$ for $S = \operatorname{Zer}(A)$.

Quality of approximate solutions

A merit function $\Phi : \mathbb{R}^N \to \mathbb{R}_+$ for a problem with solution set S satisfies:

- $\Phi(z) = 0$ if $z \in S$, and $\Phi(z) > 0$ otherwise.
- If $\Phi(z) < \Phi(w)$, then z is a better approximate solution than w.

Examples

- $\Phi(z) = f(z) - \inf(f)$ for $S = \mathrm{Argmin}(f)$.
  Also $\Phi(z) = \|\nabla f(z)\|$, if f is convex and differentiable.
- $\Phi(z) = \inf\{\|v\| : v \in Az\}$ for $S = \mathrm{Zer}(A)$.
- $\Phi(z) = \|Tz - z\|$ or $\Phi(z) = \|Tz - z\|^2$ for $S = \mathrm{Fix}(T)$.

Quality of approximate solutions

A merit function $\Phi : \mathbb{R}^N \to \mathbb{R}_+$ for a problem with solution set S satisfies:

- $\Phi(z) = 0$ if $z \in S$, and $\Phi(z) > 0$ otherwise.
- If $\Phi(z) < \Phi(w)$, then z is a better approximate solution than w.

Examples

- $\Phi(z) = f(z) - \inf(f)$ for $S = \mathrm{Argmin}(f)$.
  Also $\Phi(z) = \|\nabla f(z)\|$, if f is convex and differentiable.
- $\Phi(z) = \inf\{\|v\| : v \in Az\}$ for $S = \mathrm{Zer}(A)$.
- $\Phi(z) = \|Tz - z\|$ or $\Phi(z) = \|Tz - z\|^2$ for $S = \mathrm{Fix}(T)$.
- $\Phi(z) = \mathrm{dist}(z, S)$ in all cases ($\Phi(z) = \|z - \hat{x}\|$ if $S = \{\hat{x}\}$).

Introduction to iterative algorithms

An iterative algorithm is a procedure that computes a sequence (x_k) of points in $\mathbb{R}^N$ that approximate a solution to a problem.

Introduction to iterative algorithms

An iterative algorithm is a procedure that computes a sequence (x_k) of points in $\mathbb{R}^N$ that approximate a solution to a problem. It requires:

- An initial guess x_0.

Introduction to iterative algorithms

An iterative algorithm is a procedure that computes a sequence (x_k) of points in $\mathbb{R}^N$ that approximate a solution to a problem. It requires:

- An initial guess x_0.
- A sequence (u_k) in $\mathbb{R}^M$ of parameters.

Introduction to iterative algorithms

An iterative algorithm is a procedure that computes a sequence (x_k) of points in $\mathbb{R}^N$ that approximate a solution to a problem. It requires:

- An initial guess x_0.
- A sequence (u_k) in $\mathbb{R}^M$ of parameters.
- An operator $F : \mathbb{R}^N \times \mathbb{R}^M \to \mathbb{R}^N$ used to compute x_{k+1}, given x_k:

$$x_{k+1} = F(u_k, x_k).$$

Introduction to iterative algorithms

An iterative algorithm is a procedure that computes a sequence (x_k) of points in $\mathbb{R}^N$ that approximate a solution to a problem. It requires:

- An initial guess x_0.
- A sequence (u_k) in $\mathbb{R}^M$ of parameters.
- An operator $F : \mathbb{R}^N \times \mathbb{R}^M \to \mathbb{R}^N$ used to compute x_{k+1}, given x_k:

$$x_{k+1} = F(u_k, x_k).$$

- A stopping rule, related to a (normalized) merit function, that is activated when the approximation is sufficiently good.

Introduction to iterative algorithms

An iterative algorithm is a procedure that computes a sequence (x_k) of points in $\mathbb{R}^N$ that approximate a solution to a problem. It requires:

- An initial guess x_0.
- A sequence (u_k) in $\mathbb{R}^M$ of parameters.
- An operator $F : \mathbb{R}^N \times \mathbb{R}^M \to \mathbb{R}^N$ used to compute x_{k+1}, given x_k:

$$x_{k+1} = F(u_k, x_k).$$

- A stopping rule, related to a (normalized) merit function, that is activated when the approximation is sufficiently good.

Important questions: convergence and complexity.

Convergence

A sequence (x_k) in $\mathbb{R}^N$ converges to a limit $\hat{x} \in \mathbb{R}^N$ as $k \to \infty$ if, for every $\varepsilon > 0$, there is $K \in \mathbb{N}$ such that

$$\|x_k - \hat{x}\| < \varepsilon$$

for every $k \geq K$.

Convergence

A sequence (x_k) in $\mathbb{R}^N$ converges to a limit $\hat{x} \in \mathbb{R}^N$ as $k \to \infty$ if, for every $\varepsilon > 0$, there is $K \in \mathbb{N}$ such that

$$\|x_k - \hat{x}\| < \varepsilon$$

for every $k \geq K$.

Proposition

The following statements about a sequence (x_k) in $\mathbb{R}^N$ are equivalent:

i) *it is convergent;*

i) *it has Cauchy's property; and*

i) *it is bounded and has at most one limit point.*

Keynote examples

Theorem (Banach-Picard Iterations and the Fixed Point Theorem)

Let $q \in (0,1)$, and let $T : \mathbb{R}^N \to \mathbb{R}^N$ be $q-$Lipschitz. Then, T has a unique fixed point $\hat{x}$. Moreover, for each $x_0 \in \mathbb{R}^N$, the sequence (x_k), defined by iterating $x_{k+1} = Tx_k$, converges to $\hat{x}$.

Keynote examples

Theorem (Banach-Picard Iterations and the Fixed Point Theorem)

Let $q \in (0,1)$, *and let* $T : \mathbb{R}^N \to \mathbb{R}^N$ *be* $q-$*Lipschitz. Then,* T *has a unique fixed point* $\hat{x}$. *Moreover, for each* $x_0 \in \mathbb{R}^N$, *the sequence* (x_k), *defined by iterating* $x_{k+1} = Tx_k$, *converges to* $\hat{x}$.

Theorem (Krasnoselskii-Mann Iterations)

Let $\gamma \in (0,1)$ *and let* $T : \mathbb{R}^N \to \mathbb{R}^N$ *be nonexpansive (*$1-$*Lipschitz), with* $\mathrm{Fix}(T) \neq \emptyset$. *For each* $x_0 \in \mathbb{R}^N$, *the sequence* (x_k), *defined by iterating* $x_{k+1} = \gamma Tx_k + (1-\gamma)x_k$, *converges to a fixed point of* T.

Useful shortcuts

Lemma (Opial)

Let Ω be a nonempty subset of $\mathbb{R}^N$, and let (x_k) be a sequence in $\mathbb{R}^N$. Suppose that:

i) *For every $y \in \Omega$, $\lim\limits_{k\to\infty} \|x_k - y\|$ exists; and*

ii) *Every limit point of (x_k) belongs to Ω.*

Then, (x_k) converges, as $k \to \infty$, to a point in Ω.

Useful shortcuts

Lemma (Opial)

Let Ω be a nonempty subset of $\mathbb{R}^N$, and let (x_k) be a sequence in $\mathbb{R}^N$. Suppose that:

i) *For every $y \in \Omega$, $\lim_{k\to\infty} \|x_k - y\|$ exists; and*

ii) *Every limit point of (x_k) belongs to Ω.*

Then, (x_k) converges, as $k \to \infty$, to a point in Ω.

Remark (Fejér)

If $\|x_{k+1} - y\| \leq \|x_k - y\|$ for all $k \geq 0$ and $y \in \Omega$, then $\lim_{k\to\infty} \|x_k - y\|$ exists for every $y \in \Omega$.

Convergence rates

Let (ϕ_k) be nonnegative, and let (ρ_k) be positive, with $\lim_{k\to\infty} \rho_k = 0$

We shall use the following notation:

- $\phi_k = \mathcal{O}(\rho_k)$ means that $\phi_k \to 0$ at least as fast as ρ_k:

$$C := \sup_{k\geq 0} \left[\frac{\phi_k}{\rho_k}\right] < +\infty,$$

Convergence rates

Let (ϕ_k) be nonnegative, and let (ρ_k) be positive, with $\lim_{k\to\infty} \rho_k = 0$

We shall use the following notation:

- $\phi_k = \mathcal{O}(\rho_k)$ means that $\phi_k \to 0$ at least as fast as ρ_k:

$$C := \sup_{k\geq 0} \left[\frac{\phi_k}{\rho_k}\right] < +\infty,$$

- $\phi_k = o(\rho_k)$ means that $\phi_k \to 0$ strictly faster than ρ_k:

$$\lim_{k\to 0} \left[\frac{\phi_k}{\rho_k}\right] = 0.$$

Convergence rates

Let (ϕ_k) be nonnegative, and let (ρ_k) be positive, with $\lim_{k\to\infty} \rho_k = 0$

We shall use the following notation:

- $\phi_k = \mathcal{O}(\rho_k)$ means that $\phi_k \to 0$ at least as fast as ρ_k:

$$C := \sup_{k\geq 0} \left[\frac{\phi_k}{\rho_k}\right] < +\infty,$$

- $\phi_k = o(\rho_k)$ means that $\phi_k \to 0$ strictly faster than ρ_k:

$$\lim_{k\to 0} \left[\frac{\phi_k}{\rho_k}\right] = 0.$$

Clearly, $\phi_k = o(\rho_k)$ implies $\phi_k = \mathcal{O}(\rho_k)$, but the constant C is important!

- If $\phi_k = \mathcal{O}(q^k)$ for some $q \in (0, 1)$, we say ϕ_k converges linearly to zero.

Convergence rates

Some terminology

- If $\phi_k = \mathcal{O}(q^k)$ for some $q \in (0,1)$, we say ϕ_k converges linearly to zero.
  - It is the case for Banach-Picard iterations.

Convergence rates

Some terminology

- If $\phi_k = \mathcal{O}(q^k)$ for some $q \in (0,1)$, we say ϕ_k converges linearly to zero.
  - It is the case for Banach-Picard iterations.
- If $\phi_k \neq \mathcal{O}(q^k)$ for every $q \in (0,1)$, ϕ_k converges sublinearly.
  - En many frequent cases, we still have $\phi_k = \mathcal{O}(k^{-p})$ for some $p > 0$.

Convergence rates

Some terminology

- If $\phi_k = \mathcal{O}(q^k)$ for some $q \in (0,1)$, we say ϕ_k converges linearly to zero.
  - It is the case for Banach-Picard iterations.
- If $\phi_k \neq \mathcal{O}(q^k)$ for every $q \in (0,1)$, ϕ_k converges sublinearly.
  - En many frequent cases, we still have $\phi_k = \mathcal{O}(k^{-p})$ for some $p > 0$.
- Finally, if $\phi_k = o(q^k)$ for all $q \in (0,1)$, then ϕ_k converges superlinearly.

Convergence rates

Some terminology

- If $\phi_k = \mathcal{O}(q^k)$ for some $q \in (0,1)$, we say ϕ_k converges linearly to zero.
  - It is the case for Banach-Picard iterations.
- If $\phi_k \neq \mathcal{O}(q^k)$ for every $q \in (0,1)$, ϕ_k converges sublinearly.
  - En many frequent cases, we still have $\phi_k = \mathcal{O}(k^{-p})$ for some $p > 0$.
- Finally, if $\phi_k = o(q^k)$ for all $q \in (0,1)$, then ϕ_k converges superlinearly.
  - It is the case for Newton's method, under favorable conditions.

Consider a set $\{(u_k, v_k)\}_{k=1}^{K}$ of samples corresponding to $v \sim mu + b$, where m and b are unknown.

Consider a set $\{(u_k, v_k)\}_{k=1}^{K}$ of samples corresponding to $v \sim mu + b$, where m and b are unknown. Define

$$\Psi(m, b) = \frac{1}{2} \sum_{k=1}^{K} (v_k - mu_k - b)^2.$$

We look for the values of m and b that minimize Ψ.

Practical aspects

Linear regression

Consider a set $\left\{(u_k, v_k)\right\}_{k=1}^{K}$ of samples corresponding to $v \sim mu + b$, where m and b are unknown. Define

$$\Psi(m, b) = \frac{1}{2}\sum_{k=1}^{K}(v_k - mu_k - b)^2.$$

We look for the values of m and b that minimize Ψ. The relation $\nabla\Psi(\hat{m}, \hat{b}) = 0$ gives a system of two equations with two unknowns:

$$\left\{\begin{array}{rcl} \hat{m}\sum u_k^2 + \hat{b}\sum u_k & = & \sum u_k v_k \\ \hat{m}\sum u_k + \hat{b}K & = & \sum v_k \end{array}\right.$$

Practical aspects

Linear regression

Consider a set $\left\{(u_k, v_k)\right\}_{k=1}^{K}$ of samples corresponding to $v \sim mu + b$, where m and b are unknown. Define

$$\Psi(m, b) = \frac{1}{2}\sum_{k=1}^{K}(v_k - mu_k - b)^2.$$

We look for the values of m and b that minimize Ψ. The relation $\nabla\Psi(\hat{m}, \hat{b}) = 0$ gives a system of two equations with two unknowns:

$$\left\{\begin{array}{rcl} \hat{m}\sum u_k^2 + \hat{b}\sum u_k & = & \sum u_k v_k \\ \hat{m}\sum u_k + \hat{b}K & = & \sum v_k \end{array}\right.$$

If we suspect that $\Phi(x_k)$ converges linearly to zero, we can use this with $u_k = k$, and $v_k = \ln\left(\Phi(x_k)\right)$.

Complexity

Consider a problem with merit function Φ and a sequence (x_k) produced by an algorithm

Suppose that for every $\varepsilon > 0$, which represents a tolerance or precision level, there is a number k_ε such that

$$\Phi(x_{k_\varepsilon}) < \varepsilon.$$

A complexity bound is either a map $\varepsilon \mapsto k_\varepsilon$, or an inequality that maps ε to a set of valid k_ε's.

Complexity

Consider a problem with merit function Φ and a sequence (x_k) produced by an algorithm

Suppose that for every $\varepsilon > 0$, which represents a tolerance or precision level, there is a number k_ε such that

$$\Phi(x_{k_\varepsilon}) < \varepsilon.$$

A complexity bound is either a map $\varepsilon \mapsto k_\varepsilon$, or an inequality that maps ε to a set of valid k_ε's.

Complexity bounds from convergence rates

- From $\Phi(x_k) \leq C/k^p$, we get $k > \sqrt[p]{C/\varepsilon}$.

Complexity

Consider a problem with merit function Φ and a sequence (x_k) produced by an algorithm

Suppose that for every $\varepsilon > 0$, which represents a tolerance or precision level, there is a number k_ε such that

$$\Phi(x_{k_\varepsilon}) < \varepsilon.$$

A complexity bound is either a map $\varepsilon \mapsto k_\varepsilon$, or an inequality that maps ε to a set of valid k_ε's.

Complexity bounds from convergence rates

- From $\Phi(x_k) \leq C/k^p$, we get $k > \sqrt[p]{C/\varepsilon}$.
- From $\Phi(x_k) \leq Ae^{-Bk}$, we get $k > \ln\left(\sqrt[B]{A/\varepsilon}\right)$.

Examples

Banach-Picard Iterations

We have

$$\|Tx_k - x_k\| \leq \|Tx_0 - x_0\| \; q^k.$$

Examples

Banach-Picard Iterations

We have

$$\|Tx_k - x_k\| \leq \|Tx_0 - x_0\| \; q^k.$$

Krasnoselskii-Mann Iterations

We can prove that

$$\|Tx_k - x_k\|^2 \leq \frac{\mathrm{dist}(x_0, S)^2}{\gamma(1-\gamma)k} = \mathcal{O}\left(\frac{1}{k}\right) \quad \text{and} \quad \|Tx_k - x_k\|^2 = o\left(\frac{1}{k}\right).$$

Exercises

1. Let (a_k), (b_k) and (c_k) be nonnegative sequences, and suppose
$$a_{k+1} + b_k \leq a_k + c_k$$
for all $k \geq 0$. Assume $C := \sum c_k < +\infty$. Show that
$$\lim_{k\to\infty} a_k \text{ exists,} \qquad \text{and} \qquad \sum b_k \leq a_0 + C < +\infty.$$

2. Let (d_k) be nonnegative, with $D := \sum d_k < +\infty$, and write $\delta_k = \min\{d_0, \ldots, d_k\}$. Prove that
$$(k+1)\delta_k \leq D \qquad \text{and} \qquad \lim_{k\to\infty} (k+1)\delta_k = 0.$$

Exercises, continued

3. Let T be nonexpansive, with $\mathrm{Fix}(T) \neq \emptyset$. Apply Krasnoselskii-Mann iterations
$$x_{k+1} = \gamma T x_k + (1-\gamma)x_k.$$
   1. Show that $\|x_{k+1} - \hat{x}\|^2 + \gamma(1-\gamma)\|Tx_k - x_k\|^2 \leq \|x_k - \hat{x}\|^2$ for every $\hat{x} \in \mathrm{Fix}(T)$. Conclude that (x_k) converges to a fixed point of T.
   2. Verify that $k \mapsto \|Tx_k - x_k\|^2$ is nonincreasing.
   3. Prove that the following convergence rates hold:
$$\|Tx_k - x_k\|^2 \leq \frac{\mathrm{dist}(x_0, S)^2}{\gamma(1-\gamma)k} = \mathcal{O}\left(\frac{1}{k}\right) \quad \text{and} \quad \|Tx_k - x_k\|^2 = o\left(\frac{1}{k}\right).$$