Introduction to Optimization

Lecture 01: Motivation and examples. Optimization problems and their solutions.

university of groningen

Nothing takes place in the world whose meaning is not that of some maximum or minimum.

Leonhard Euler (1707 - 1783)

Nothing takes place in the world whose meaning is not that of some maximum or minimum.

Leonhard Euler (1707 - 1783)

Principle of Least Time

The path taken by a ray between two given points is the path that can be traveled in the least time.

Nothing takes place in the world whose meaning is not that of some maximum or minimum.

Leonhard Euler (1707 - 1783)

Principle of Least Time

The path taken by a ray between two given points is the path that can be traveled in the least time.

Law of Refraction

For a given pair of media, the ratio of the sines of the angles of incidence and refraction equals the ratio of the refractive indices of the two media.

Material-efficient cans

Material-efficient cans

Some terminology

An optimization problem is usually written as

$$\min_{x \in C} f(x),$$

where

- f is the objective function.
- C is the feasible set, or the set of constraints. Points in C are feasible.

Some terminology

An optimization problem is usually written as

$$\min_{x \in C} f(x),$$

where

- f is the objective function.
- C is the feasible set, or the set of constraints. Points in C are feasible.

The problem is constrained if $C \subsetneq \mathbb{R}^N$, and unconstrained otherwise.

Some terminology

An optimization problem is usually written as

$$\min_{x \in C} f(x),$$

where

- f is the objective function.
- C is the feasible set, or the set of constraints. Points in C are feasible.

The problem is constrained if $C \subsetneq \mathbb{R}^N$, and unconstrained otherwise.

A point $\hat{x} \in C$ is a global minimizer if $f(\hat{x}) \leq f(x)$ for all $x \in C$. It is a local minimizer if $f(\hat{x}) \leq f(x)$ for all $x \in C$ close to $\hat{x}$.

Some terminology

An optimization problem is usually written as

$$\min_{x \in C} f(x),$$

where

- f is the objective function.
- C is the feasible set, or the set of constraints. Points in C are feasible.

The problem is constrained if $C \subsetneq \mathbb{R}^N$, and unconstrained otherwise.

A point $\hat{x} \in C$ is a global minimizer if $f(\hat{x}) \leq f(x)$ for all $x \in C$. It is a local minimizer if $f(\hat{x}) \leq f(x)$ for all $x \in C$ close to $\hat{x}$.

The minimizer is unique or strict if the corresponding inequality is strict.

Course description

Introductory course on optimization, where the fundamental concepts, techniques and tools are presented, discussed and put into practice by means of modelling and analysis.

Course description

Introductory course on optimization, where the fundamental concepts, techniques and tools are presented, discussed and put into practice by means of modelling and analysis.

Assessment:

- 40% Written exam
- 15% Computational Exercises
- 45% Homework Assignments

Learning outcomes

1. Identify and state unconstrained and constrained optimization problems, and assess whether they have solutions

Learning outcomes

1. Identify and state unconstrained and constrained optimization problems, and assess whether they have solutions
2. Characterize and calculate solutions to optimization problems by means of optimality conditions

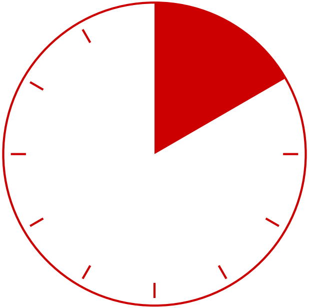
Learning outcomes

1. Identify and state unconstrained and constrained optimization problems, and assess whether they have solutions
2. Characterize and calculate solutions to optimization problems by means of optimality conditions
3. Prove the convergence of optimization algorithms and establish their convergence rates

Learning outcomes

1. Identify and state unconstrained and constrained optimization problems, and assess whether they have solutions
2. Characterize and calculate solutions to optimization problems by means of optimality conditions
3. Prove the convergence of optimization algorithms and establish their convergence rates
4. Implement optimization algorithms to approximate solutions numerically

Break



Motivation from data analysis

We begin with a data set

$$\mathcal{D} = \{(a_j, y_j) \in \mathbb{R}^K \times \mathbb{R}^M : j = 1, 2, \dots, J\},$$

where the a_j's are features and the y_j's are observations or labels.

Motivation from data analysis

We begin with a data set

$$\mathcal{D} = \{(a_j, y_j) \in \mathbb{R}^K \times \mathbb{R}^M : j = 1, 2, \dots, J\},$$

where the a_j's are features and the y_j's are observations or labels.

The aim is to discover or learn a function $\phi : \mathbb{R}^K \to \mathbb{R}^M$ such that

$$\phi(a_j) \simeq y_j$$

for, let us say, many j's.

The optimization point of view

Typically, ϕ belongs to a family of functions and, within that family, it is defined by some unknown parameter $x \in \mathbb{R}^N$.

The optimization point of view

Typically, ϕ belongs to a family of functions and, within that family, it is defined by some unknown parameter $x \in \mathbb{R}^N$.

We define a loss function

$$\mathcal{L}_{\mathcal{D}}(x) = \frac{1}{J} \sum_{j=1}^{J} \ell(a_j, y_j, x),$$

in such a way that the parameters most consistent with the data set are the ones that give the lowest values of $\mathcal{L}_{\mathcal{D}}$.

Regularization

The term regularization refers loosely to changes in the problem that either simplify the resolution or induce properties in the solutions.

Regularization

The term regularization refers loosely to changes in the problem that either simplify the resolution or induce properties in the solutions.

Examples

- To find the minimum of a nonsmooth function, one may replace the function by a smooth one that is similar, in some sense.

Regularization

The term regularization refers loosely to changes in the problem that either simplify the resolution or induce properties in the solutions.

Examples

- To find the minimum of a nonsmooth function, one may replace the function by a smooth one that is similar, in some sense.
- Approximate solutions with fewer nonzero entries may be preferred, for storage or explainability reasons (feature selection).

Regularization

The term regularization refers loosely to changes in the problem that either simplify the resolution or induce properties in the solutions.

Examples

- To find the minimum of a nonsmooth function, one may replace the function by a smooth one that is similar, in some sense.
- Approximate solutions with fewer nonzero entries may be preferred, for storage or explainability reasons (feature selection).

Regularization also helps reduce overfitting.

Least squares

If $\phi : \mathbb{R}^N \to \mathbb{R}$ is linear, namely $y = \phi(a) = x^* a$, we find x my

$$\min_{x \in \mathbb{R}^N} \frac{1}{2J} \sum_{j=1}^{J} (x^* a_j - y_j)^2 = \min_{x \in \mathbb{R}^N} \frac{1}{2J} \|Ax - y\|^2.$$

Least squares

If $\phi : \mathbb{R}^N \to \mathbb{R}$ is linear, namely $y = \phi(a) = x^* a$, we find x my

$$\min_{x \in \mathbb{R}^N} \frac{1}{2J} \sum_{j=1}^{J} (x^* a_j - y_j)^2 = \min_{x \in \mathbb{R}^N} \frac{1}{2J} \|Ax - y\|^2.$$

The minimum is attained at $x = (A^* A)^{-1} A^* y$, if $y \in \operatorname{ran}(A)$.

Least squares

If $\phi : \mathbb{R}^N \to \mathbb{R}$ is linear, namely $y = \phi(a) = x^* a$, we find x my

$$\min_{x \in \mathbb{R}^N} \frac{1}{2J} \sum_{j=1}^{J} (x^* a_j - y_j)^2 = \min_{x \in \mathbb{R}^N} \frac{1}{2J} \|Ax - y\|^2.$$

The minimum is attained at $x = (A^* A)^{-1} A^* y$, if $y \in \operatorname{ran}(A)$.

Exercise

Is the minimum attained if $y \notin \operatorname{ran}(A)$? If so, where?

Least squares

If $\phi : \mathbb{R}^N \to \mathbb{R}$ is linear, namely $y = \phi(a) = x^* a$, we find x my

$$\min_{x \in \mathbb{R}^N} \frac{1}{2J} \sum_{j=1}^{J} (x^* a_j - y_j)^2 = \min_{x \in \mathbb{R}^N} \frac{1}{2J} \|Ax - y\|^2.$$

The minimum is attained at $x = (A^* A)^{-1} A^* y$, if $y \in \mathrm{ran}(A)$.

Exercise

Is the minimum attained if $y \notin \mathrm{ran}(A)$? If so, where?

Regularization: Ridge, LASSO, TV...

Support vector machines

Classification problem: Individual j is labeled $y_j \in \{-1, 1\}$.

Support vector machines

Classification problem: Individual j is labeled $y_j \in \{-1, 1\}$. A vector a_j, with features of j, is known.

Support vector machines

Classification problem: Individual j is labeled $y_j \in \{-1, 1\}$. A vector a_j, with features of j, is known. We wish to predict the label of a new individual.

Support vector machines

Classification problem: Individual j is labeled $y_j \in \{-1, 1\}$. A vector a_j, with features of j, is known. We wish to predict the label of a new individual.

A support vector machine finds a hyperplane that separates the points labeled 1 from those labeled -1.

Support vector machines

Classification problem: Individual j is labeled $y_j \in \{-1, 1\}$. A vector a_j, with features of j, is known. We wish to predict the label of a new individual.

A support vector machine finds a hyperplane that separates the points labeled 1 from those labeled -1.

If this is possible, we may want to maximize the distance to the closest point on each side to obtain a more robust classification.

Support vector machines

Classification problem: Individual j is labeled $y_j \in \{-1, 1\}$. A vector a_j, with features of j, is known. We wish to predict the label of a new individual.

A support vector machine finds a hyperplane that separates the points labeled 1 from those labeled -1. ✍

If this is possible, we may want to maximize the distance to the closest point on each side to obtain a more robust classification.

Nonlinear transformations of the space are possible. ✍

Other examples

- Matrix factorization
- Matrix completion (a.k.a. the *Netflix Prize* Problem) ✍
- Logistic regression
- Image processing: deblurring, in-painting
  - Topic of the computational project
- Face, voice and pattern recognition
- (Deep) learning ✍

What these problems have in common

- They can be formulated as minimizing functions of one or more real variables (sometimes in a huge number).

What these problems have in common

- They can be formulated as minimizing functions of one or more real variables (sometimes in a huge number).
- The functions involved are continuous. They are either smooth, or they have simple or structured forms of nonsmoothness.

What these problems have in common

- They can be formulated as minimizing functions of one or more real variables (sometimes in a huge number).
- The functions involved are continuous. They are either smooth, or they have simple or structured forms of nonsmoothness.
- The function to be minimized is often a sum of simple functions that depend either on few data points or involve few variables.

What these problems have in common

- They can be formulated as minimizing functions of one or more real variables (sometimes in a huge number).
- The functions involved are continuous. They are either smooth, or they have simple or structured forms of nonsmoothness.
- The function to be minimized is often a sum of simple functions that depend either on few data points or involve few variables.
- Usually, the nonsmooth parts can be separated from the rest (additive structure).