Chapter I. General Optimization

Lecture 1: The world of Nonlinear Optimization

Yurii Nesterov
EPLouvain (UCL)

Outline

- About this course
- General formulation of the problem
- Important examples
- Black Box and Iterative Methods
- Analytical and Arithmetical Complexity
- Uniform Grid Method
- Lower complexity bounds
- Lower bounds for Global Optimization
- Rules of the Game

About this course

What we are going to do:

- Give a description of the modern optimization theory.
- Take into account the historical aspect in order to understand the logic of the development.
- Develop a "computational sense", which helps us to understand what we *can* and what we *cannot* expect from a numerical method.

Why that is important:

- Optimization formulations are very popular among practitioners.
- Optimization Theory is simple and easy to learn.
- Optimization is an excellent example of a *comprehensive* theory.

General formulation of the problem

Let x be an n-dimensional real vector: $x = (x^{(1)}, \ldots, x^{(n)}) \in \mathbb{R}^n$,

S be a subset of $\mathbb{R}^n$: $S \subseteq \mathbb{R}^n$,

$f_0(x) \ldots f_m(x)$ are some real-valued function of x.

Problem formulation:

$$\begin{array}{lll} \min\limits_{x \in S} & f_0(x) & (\equiv -\max(-f_0(x)) \\ \text{s.t.:} & f_j(x) \begin{pmatrix} = \\ \le \end{pmatrix} 0, \; j = 1 \ldots m. & \end{array}$$

Terminology:

- $f_0(x)$ is the *objective function*,
- $f(x) = (f_1(x), \ldots, f_m(x))$ are *functional constraints*,
- S is the *basic feasible set*.
- Q is the *feasible set*: $Q = \{x \in S \mid \; f_j(x) \le 0, \; j = 1 \ldots m\}$.

Types of minimization problems

- *Constrained problems*: $Q \subset \mathbb{R}^n$,
- *Smooth problems*: all $f_j(x)$ are differentiable,
- *Nonsmooth problems*: there is a nondifferentiable component $f_k(x)$,
- *Linearly constrained problems*: all functional constraints are linear:

$$f_j(x) = \sum_{i=1}^{n} a_j^{(i)} x^{(i)} + b_j \equiv \underbrace{\langle a_j, x \rangle}_{\text{inner product}} + b_j, \; j = 1 \ldots m,$$

and S is a polyhedron.

If $f_0(x)$ is also linear, then this is a *Linear Programming Problem*.

If $f_0(x)$ is quadratic, then this is a *Quadratic Programming Problem*.

Some terminology

Feasibility:

- Problem is *feasible* if $Q \neq \emptyset$.
- Problem is *strictly feasible* if $\exists x \in \mathrm{int}\, Q$ such that $f_j(x) < 0$ for all inequality constraints and $f_j(x) = 0$ for all equality constraints.

Extremum:

- x^* is an optimal *global solution* to the problem if $f_0(x^*) \leq f_0(x)$ for all $x \in Q$ (*global minimum*).

  Then $f_0(x^*)$ is called the *optimal value* of the problem.
- x^* is an optimal *local solution* to the problem if $f_0(x^*) \leq f_0(x)$ for all $x \in \mathrm{int}\, \bar{Q} \subset Q$ (*local minimum*).

Example of optimization problem, I

Let $x^{(1)} \dots x^{(n)}$ be our *design variables*.

Then we can fix some *characteristics* of our decision $f_0(x), \dots, f_m(x)$. That could be:

- The price of the project,
- Amount of required resources,
- Reliability of the system,

and many others.

We fix the most important characteristics, $f_0(x)$, as our *objective*.

For all others we impose some bounds: $a_j \leq f_j(x) \leq b_j$.

Thus, we come to the problem:

$$\min_{x \in S} \left\{ f_0(x) : \; a_j \leq f_j(x) \leq b_j, \; j = 1 \dots m \right\},$$

where S stands for the *structural* constraints (e.g., positivity of some variables).

Example of optimization problem, II

Let our initial problem be as follows:

$$\text{Find } x \in \mathbb{R}^n : \; f_1(x) = a_1, \ldots, f_m(x) = a_m. \tag{1}$$

Then we can consider the problem: $\min\limits_{x} \sum\limits_{j=1}^{m} (f_j(x) - a_j)^2$

(may be with some additional constraints on x).

Note:

The problem (1) is almost *universal*. It covers:

- Ordinary differential equations
- Partial differential equations
- Problems, arising in Game Theory

and many other fields.

Example of the problem, III.

Let our decision variable $x^{(1)} \dots x^{(n)}$ be *integer*.

This can be described by the constraint: $\sin(\pi x^{(i)}) = 0, \quad i = 1 \dots n.$

Thus, we could treat also the *Integer Programming* Problems:

$$
\begin{aligned}
& \min \ f_0(x), \\
\text{s.t.:} \quad & a_j \leq f_j(x) \leq b_j, \ j = 1 \dots m, \\
& x \in S, \\
& \sin(\pi x^{(i)}) = 0, \ i = 1 \dots n.
\end{aligned}
$$

Conclusions

1955:

Nonlinear Optimization is very important. It covers almost all fields of Numerical Analysis.

1975:

In general, optimization problems are UNSOLVABLE.

Performance of Numerical Method

$$\text{Numerical Method } \mathcal{M} \quad \Longleftrightarrow \quad \text{Problem } \mathcal{P}$$

What we can say about the performance of $\mathcal{M}$?

Observations:

1. Best $\mathcal{M}$ for a *single* problem $\mathcal{P}$ is a silly notion.

   (All methods are worse than the trivial one returning the solution of $\mathcal{P}$ all the time.)
2. Therefore we need:
   - Description of a *class* of problems $\mathcal{F} \supset \mathcal{P}$.
   - Description of an *oracle* $\mathcal{O}$, which provides $\mathcal{M}$ by some information about $\mathcal{P}$.

Class and *Oracle* compose the MODEL of our problem. (Not unique !)

We can define the *performance* of $\mathcal{M}$ on $(\mathcal{F}, \mathcal{O})$ as its performance on the WORST $\mathcal{P}_w$ from $\mathcal{F}$.

(This $\mathcal{P}_w$ can be bad only for our $\mathcal{M}$!)

Some definitions

Performance of $\mathcal{M}$ on $\mathcal{P}$ is

The total amount of *Computational Efforts*
which is required by method $\mathcal{M}$
in order to *Solve the Problem* $\mathcal{P}$

To Solve the Problem could mean:

1. Find the *exact* solution.

   (Impossible to find in finite time even for the simplest nonlinear problems.)

2. Find an *approximate* solution with a small accuracy $\epsilon > 0$.

   (For that apply an *iterative process*.)

General Iterative Scheme

Input:

- Starting point x_0.
- Accuracy $\epsilon > 0$.

Initialization. Set $k = 0$, $I_{-1} = \emptyset$, where

- k is the iteration counter.
- I_k is the *informational set* accumulated after k iterations.

Main Loop

1. Call the oracle $\mathcal{O}$ at x_k.
2. Update the informational set: $I_k = I_{k-1} \bigcup (x_k, \mathcal{O}(x_k))$.
3. Apply rules of method $\mathcal{M}$ to I_k and form the new test point x_{k+1}.
4. Check the stopping criterion. If **yes** then form an output $\bar{x}$.
   Otherwise set $k = k + 1$ and go to 1.

Computational Efforts

1. **Analytical complexity:**

   The number of CALLS OF ORACLE, *which is required to solve the problem $\mathcal{P}$ up to accuracy ϵ.*

2. **Arithmetical complexity:**

   The total number of ARITHMETIC OPERATIONS *(including the work of oracle and of the method) , which is required to solve the problem $\mathcal{P}$ up to accuracy ϵ.*

Note: The meaning of words *up to accuracy* $\epsilon > 0$ must be *exact*.

Black Box Concept

1. The only information available from the oracle is its answer. No intermediate results are available.

2. The oracle is local*:*

A small variation of the problem far enough from the test point x does not change the answer at x.

Note: This concept is extremely popular, but it is not the end of the story. We will see this later.

Examples of the oracle

1. **Zero-order oracle**
   - *Input:* test point x.
   - *Output:* value $f(x)$.
2. **First-order oracle**
   - *Input:* test point x.
   - *Output:* value $f(x)$ and gradient $f'(x) \in \mathbb{R}^n$.
3. **Second-order oracle**
   - *Input:* test point x.
   - *Output:* value $f(x)$, gradient $f'(x) \in \mathbb{R}^n$, and Hessian $f''(x) \in \mathbb{S}^n$.

Uniform Grid Method

Problem Formulation: $\min_{x \in \mathbb{B}_n} f(x)$, where $\mathbb{B}_n$ is an n-dimensional box in $\mathbb{R}^n$:

$$\mathbb{B}_n = \left\{ x \in \mathbb{R}^n : \ 0 \le x^{(i)} \le 1, \ i = 1, \ldots, n \right\}.$$

Assumption ($\equiv$ Problem Class)

The objective function $f(\cdot)$ is *Lipschitz continuous* on $\mathbb{B}_n$:

$$| f(x) - f(y) | \le L \| x - y \|_\infty, \quad \forall x, y \in \mathbb{B}_n,$$

with some constant L (*Lipschitz constant*).

Here $\| \cdot \|_\infty$ is the *infinity-norm* on $\mathbb{R}^n$:

$$\|x\|_\infty = \max_{1 \le i \le n} |x^{(i)}|.$$

Scheme of method $\mathcal{UG}(p)$

($p \geq 1$ is an integer input parameter)

1. Form p^n points $x_{(i_1,i_2,\dots,i_n)} = \left(\frac{1}{2p} + \frac{i_1}{p}, \dots, \frac{1}{2p} + \frac{i_n}{p}\right)$,

where $i_1 = 0, \dots, p-1, \dots, i_n = 0, \dots, p-1$.

2. Among all points $x_{(\dots)}$ find the point $\bar{x}$ with the minimal value of the objective function.

3. Return the pair $(\bar{x}, f(\bar{x}))$ as the result.

Note: 1. This method can be treated as an iterative process with no influence of the accumulated information on the sequence of test points.

2. This is a zero-order method.

Theorem 1.1

Let f^* be the global optimal value of our problem. Then

$$f(\bar{x}) - f^* \quad \leq \quad \tfrac{L}{2p}.$$

Proof. 1. It is clear that $\mathbb{B}_n = \bigcup\limits_{i=(i_1,\ldots,i_n)} \mathcal{B}_\infty\left(x_{(i)}, \tfrac{1}{2p}\right)$,

where $\mathcal{B}_\infty(x, r) = \{y \in \mathbb{R}^n : \ \|y - x\|_\infty \leq r\}$.

2. Let x^* be the global minimum of our problem. Then there exists an index $i^* = (i_1^*, \ldots, i_n^*)$ such that

$$\|x^* - x_{(i^*)}\|_\infty \quad \leq \quad \tfrac{1}{2p}.$$

3. Therefore, $f(x_{(i^*)}) - f(x^*) \leq \frac{L}{2p}$. Note that $f(\bar{x}) \leq f(x_{(i^*)})$. $\square$

Approximate solution

$$\text{Find } \bar{x} \in \mathbb{B}_n : f(\bar{x}) - f^* \le \epsilon.$$

Corollary 1.1. *The analytical complexity of method $\mathcal{UG}$ is as follows:*

$$\mathcal{A}(\mathcal{G}) \quad = \quad \left(\left\lfloor \tfrac{L}{2\epsilon} \right\rfloor + 1 \right)^n,$$

where $\lfloor a \rfloor$ is the integer part of a.

Proof. Indeed, let us take $p = \left\lfloor \frac{L}{2\epsilon} \right\rfloor + 1$. Then

$$p \quad \ge \quad \tfrac{L}{2\epsilon},$$

and therefore, in view of T.1.1, $f(\bar{x}) - f^* \le \frac{L}{2p} \le \epsilon$. $\square$

Questions:

- How good is this estimate?
- How good is this method?

Lower complexity bounds

1. Are based on the *Black Box* concept.

2. Can be derived for a specific class of problems $\mathcal{F}$ equipped by an oracle $\mathcal{O}$.

3. Are valid for *any* iterative scheme.

4. Provide us with a lower bound for the *analytical complexity* of the class $\mathcal{F}$.

5. Use the idea of *resisting* oracle.

Resisting Oracle

1. It is trying to create a *worst* problem for each particular method.

2. It starts from an "empty" function and it tries to answer each call of the method in the worst possible way.

3. However, the answers must be *compatible* with

- Previous answers,
- Description of the problem class.

Note that:

- After the termination of the method, it is possible to *reconstruct* the created problem.
- If we run the method on this problem, it will reproduce the same sequence of test points.

Lower bounds for Global Optimization

Problem Formulation: $\min\limits_{x \in \mathbb{B}_n} f(x)$.

Problem Class: Objective function $f(\cdot)$ is *Lipschitz continuous* on $\mathbb{B}_n$.

Approximate solution: Find $\bar{x} \in \mathbb{B}_n : f(\bar{x}) - f^* \leq \epsilon$.

Theorem 1.2. *Analytical complexity of this class for* 0*-order methods is at least* $\left\lfloor \frac{L}{2\epsilon} \right\rfloor^n$ calls of oracle.

Proof

Assume that for any problem of our class, method $\mathcal{M}$ needs no more than $N < p^n$ calls of oracle in order to find ϵ-solution, where

$$p = \left\lfloor \frac{L}{2\epsilon} \right\rfloor \ (\geq 1),$$

Let us apply this method to the following resisting oracle:

It reports that $f(x) = 0$ at any test point x.

Therefore this method can find only $\bar{x} \in \mathbb{B}_n : \quad f(\bar{x}) = 0.$

Since $N < p^n$, there exists $i^* = (i_1, \ldots, i_n)$, such that in the box $\mathcal{B}_\infty(x_{(i^*)}, \frac{1}{2p})$ there was no test points.

Consider the function $\bar{f}(x) = \min\left\{0, L\left(\|x - x_{(i^*)}\|_\infty - \frac{1}{2p}\right)\right\}$. Note that

$$\bar{f}^* = -\frac{L}{2p} \leq -\epsilon.$$

On the other hand,

- $\bar{f}$ is Lipschitz continuous in $\|\cdot\|_\infty$ with constant L.
- $\bar{f}(x) = 0$ outside the box $\mathcal{B}_\infty(x_{(i^*)}, \frac{1}{2p})$.

$\square$

What we can say now?

	$\mathcal{UG}$	Lower bound
Complexity Estimate	$\left(\left\lfloor \frac{L}{2\epsilon} \right\rfloor + 1\right)^n$	$\left\lfloor \frac{L}{2\epsilon} \right\rfloor^n$

Thus, Uniform Grid Method is *optimal* on our problem class.

What does it mean: unsolvable?

Lower complexity bound: $\left(\frac{L}{2\epsilon}\right)^n$.

Example: $L = 2$, $n = 10$ (very small size), $\epsilon = 0.01$ (1% accuracy).

Lower bound:	10^{20} calls of oracle,
Complexity of the oracle:	n a.o.,
Total complexity:	10^{21} a.o.,
Sun Station (best for 1996):	10^6 flops per second,
Total time:	10^{15} seconds,
1 year:	less than $3.2 \cdot 10^7$ sec.
We need:	32 000 000 years.

Note: $n \to n + 1 \quad \Rightarrow \quad$ Multiply complexity by 100.

But: $\epsilon \to 2\epsilon \quad \Rightarrow \quad$ Divide complexity by 1000.

Thus, for $\epsilon = 8\%$ we need two weeks.

Why this works in another fields?

Example: Integration.

Problem: Compute the integral $\mathcal{I} = \int\limits_0^1 f(x)dx$.

Discrete Sum: $S_N = \frac{1}{N} \sum\limits_{i=1}^{n} f(x_i), \quad x_i = \frac{i}{N},\ i = 1, \dots, N.$

Result: If $f(\cdot)$ is Lipschitz continuous then

$$N = L/\epsilon \quad \Rightarrow \quad | \, \mathcal{I} - S_N \, | \leq \epsilon.$$

This approach is standard. Why?

Because of *dimension* !

Integration:	$1 - 3$
Optimization:	$1 - 10^6$.

What is the next?

Reasons to stop:

- We have already proved everything.
- This problem is too difficult to solve.
- We cannot wait for 32 000 000 years. Forget it.

Reasons to continue:

- We need to solve these problems.
- We know that people have already solved many of them. They are satisfied by the results.
- May be we want too much?

Rules of the Game

Primary:

- Description of goals.
- Description of problem class.
- Description of oracle.

Secondary:

- Desired properties of the methods.

Global Optimization (Lecture 1)

Goals: Find a global minimum.

Problem Class: Continuous functions.

Oracle: $0-1-2$ order black box.

Desired properties: Convergence to a global minimum.

Features:

- This game is too short.
- We always loose it.

Problem Sizes: Sometimes people report on solving problems with several thousands of variables. No guarantee for success even for very small problems.

History:

- Starts from 1955.
- Several local peaks of interest (simulated annealing, neural networks, genetic algorithms).

Nonlinear Optimization (Lectures 2, 3)

Goals: Find a local minimum.

Problem Class: Differentiable functions.

Oracle: $1-2$ order black box.

Desired properties: Convergence to a local minimum. Fast local convergence.

Features:

- Variability of approaches.
- Most widespread software.
- The goal is not always acceptable.

Problem Sizes: upto 1000 variables.

History:

- Starts from 1955.
- Peak period: 1965 – 1975.
- Theoretical activity now is rather low.

Convex Optimization (Lectures 4 – 10)

Goals: Find a global minimum.

Problem Class: Convex sets and functions.

Oracle: 1st and 2nd order Black Box.

Desired properties: Convergence to a global minimum. Rate of convergence can be depenent on dimension.

Features:

- Very reach and interesting theory.
- Comprehensive complexity theory.
- Efficient practical methods.
- The problem class is sometimes restrictive.

Problem Sizes: upto 1000 variables.

History:

- Starts from 1970.
- Peak period: 1975 – 1985, 2005 - now.
- Theoretical activity now is high.

Structural Optimization (Lectures 11 – 16)

Goals: Find a global minimum.

Problem Class: Convex sets and functions.

Oracle: 1st and 2nd order oracle which is not a Black Box.

Desired properties: Fast convergence to a global minimum. Rate of convergence depends on the structure of the problem.

Features:

- New and perspective theory.
- Avoid the black box concept.
- Problem class is the same as in Convex Programming.

Problem Sizes: Sometimes up to 10 000 000 variables. Recent extention on Huge-Scale Problems ($n >> 10^6$).

History:

- Starts from 1984.
- Peak period: 1990 – 2005.
- Very high theoretical activity right now.